CAM: Causal Additive Models, high-dimensional order search and penalized regression

Peter Bühlmann    Jonas Peters    Jan Ernest
Seminar for Statistics
ETH Zürich
{buhlmann, peters, ernest}@stat.math.ethz.ch
October 8, 2013

Abstract

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 non-regularized (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 (Pearl 2000 [2000], Spirtes et al. [2000], cf.) and the sheer size of the space about 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 (Hastie et al., 2009; Bühlmann and van de Geer, 2011) 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, i.e., there are no hidden variables, and that the causal influence diagram doesn’t allow for directed cycles. Generalizations to include hidden variables and directed cycles are briefly discussed in Section 7. To formalize a model, one can use the concepts of graphical modeling (Lauritzen, 1996) or structural equation models (Pearl, 2000). The approaches are equivalent in the nonparametric or multivariate Gaussian case, but this is not true anymore when placing additional restrictions which can be very useful (Shimizu et al., 2006; Peters et al., 2013; Peters and Bühlmann, 2013a). 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, i.e., 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 equivalent search for computing the BIC-regularized maximum likelihood estimator (Chickering, 2002) or the PC-algorithm using multiple conditional independence testing (Spirtes et al., 2000). However, for the latter, the constraint of additive structural equations cannot (be easily) observed, and maximum likelihood estimation among all (e.g. linear Gaussian) DAG models is computationally challenging and statistical guarantees for high-dimensional cases are only available under rather strong assumptions (van de Geer and Bühlmann, 2013).

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 Teyssier and Koller (2005) and Schmidt et al.
If the order among the variables would be known, the problem boils down to variable selection in multivariate (potentially non-linear) 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 (Yuan and Lin, 2006; Ravikumar et al., 2009; Meier et al., 2009). 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. Empirical results in Section 6 support that we can do much more accurate estimation than for non-identifiable models such as the popular linear Gaussian structural equation model.

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. Identifiability results for this model class has been given by Mooij et al. (2009) and Peters et al. (2013). The approach by Mooij et al. (2009) 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 presence of many variables. An extension of our model additive structural equation model with Gaussian errors to the case with a nonparametric specification of the error distribution is presented in Nowzohour and Bühlmann (2013), 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, Voorman et al. (2013) 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 non-regularized (restricted) maximum likelihood. The former is a well-understood subject thanks to the broad literature in sparse regression and related techniques (Tibshirani, 1996; Meinshausen and Bühlmann, 2006; Yuan and Lin, 2006; Zhao and Yu, 2006; Zou, 2006; Wainwright, 2009, cf.). Regarding the latter issue about order selection, a recent analysis in van de Geer (2013) extends our low-dimensional consistency result for the (non-restricted) 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).

2 Additive structural equation models

Consider the general structural equation model (SEM):

$$X_j = f_j(X_{pa_D(j)}, \varepsilon_j), \quad \varepsilon_1, \ldots, \varepsilon_p \text{ (mutually) independent},$$

where $pa_D(j)$ denotes the set of parents for node $j$ in DAG $D$ and $f_j$ is a function from $\mathbb{R}^{|pa_D(j)|+1} \to \mathbb{R}$. 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 $\varepsilon_j$ ($j = 1, \ldots, p$). The model 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 potentially misspecified additive SEM with Gaussian errors:

$$X_j = \sum_{k \in pa_D(j)} f_{j,k}(X_k) + \varepsilon_j,$$

$$\varepsilon_1, \ldots, \varepsilon_p \text{ independent with } \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2) \ (j = 1, \ldots, p),$$

$$\mathbb{E}[f_{j,k}(X_k)] = 0 \text{ for all } j, k,$$

(1)

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

$$X_j = \sum_{k \in pa_D(j)} \beta_{j,k} X_k + \varepsilon_j,$$

$$\varepsilon_1, \ldots, \varepsilon_p \text{ independent with } \varepsilon_j \sim \mathcal{N}(0, \sigma_j^2) \ (j = 1, \ldots, p).$$

(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 non-linear, the DAG $D$ is identifiable from the distribution $P$ of $X_1, \ldots, X_p$ (Peters et al., 2013). On the other hand, for a general SEM or for a linear Gaussian SEM as in (2) one can only identify the Markov equivalence class of the DAG $D$, assuming causal minimality; an exception arises when assuming same error variances $\sigma_j \equiv \sigma^2$ for all $j$ in (2) which again implies identifiability of the DAG $D$ from $P$ (Peters and Bühmann, 2013a). In the sequel, we consider the fully identifiable case of model (1).

### 2.1 The likelihood

We slightly re-write model (1) as

$$X_j = \sum_{k \in \text{pa}_D(j)} f_{j,k}(X_k) + \varepsilon_j = \sum_{k \neq j} f_{j,k}(X_k) + \varepsilon_j \quad (j = 1, \ldots, p),$$

$f_{j,k}(\cdot) \not\equiv 0$ if and only if there is a directed edge $k \rightarrow j$ in $D$,

$$E[f_{j,k}(X_k)] = 0 \quad \text{for all } j, k,$$

$\varepsilon_1, \ldots, \varepsilon_p$ independent and $\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$. (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 $P$ (Peters et al., 2013). Denote by $\theta$ 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 $\theta^0$ (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 parameter(s).

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

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

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

$$\sigma_j^2 = E[(X_j - \sum_{k \neq j} f_{j,k}(X_k))^2],$$
and the expected negative log-likelihood is:

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

2.2 The function class

We assume that the functions in model (1) or (3) are from a class of smooth functions

\[ F = \{ f : \mathbb{R} \to \mathbb{R}, f \in C^\alpha \}, \]

where \( C^\alpha \) denotes the space of \( \alpha \)-times differentiable functions.

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

\[ F_n = \{ f = \sum_{r=1}^{a_n} \alpha_r b_r(\cdot), f \in F \}. \tag{4} \]

Furthermore, the space of additive functions is denoted by

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

\[ F_n^{\oplus \ell} = \{ f : \mathbb{R}^\ell \to \mathbb{R}; f(x) = \sum_{k=1}^{\ell} f_k(x_k), f_k \in F_n \}, \]

where \( \ell = 2, \ldots, p \). Clearly \( F_n^{\oplus \ell} \subseteq F^{\oplus \ell} \). For \( f \in F^{\oplus \ell} \) we denote by \( f_k \) its \( k \)th additive function.

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 \( \pi \) on \( \{1, \ldots, p\} \), define:

\[ X^\pi, \quad X_j^\pi = X_{\pi(j)}. \]

A true ordering or permutation \( \pi^0 \) allows for a lower-triangular (or autoregressive) representation of the model in (3):

\[ X_j^{\pi^0} = \sum_{k=1}^{j-1} f_{j,k}^{\pi^0}(X_k^{\pi^0}) + \varepsilon_j^{\pi^0} (j = 1, \ldots, p), \tag{5} \]
where \( f_{\pi,j,k}(\cdot) = f_{\pi_0(j),\pi_0(k)}(\cdot) \) and \( \varepsilon_{\pi,0}^\pi = \varepsilon_{\pi_0(j)} \), i.e., with permuted indices in terms of the original quantities in (3). Note that \( \pi(1) \) has no parents, called the source node. Such a true order or permutation is typically not unique: it is an element of

\[
\Pi^0 = \{ \pi^0; \text{[5] holds} \}.
\]

If all functions \( f_{j,k}(\cdot) \) are nonlinear, the true DAG \( D^0 \) is identifiable from the distribution (Peters et al., 2013), and \( \Pi^0 \) consists of all permutations and corresponding orderings of the variables which are consistent with the arrow directions in the true DAG \( D^0 \).

**Remark 1.** For the Gaussian linear SEM (2), \( \Pi^0 \) consists of all permutations since we can represent every Gaussian distribution as in (5) for every order \( \pi \). However, 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 non-zero coefficients (assuming faithfulness of the true distribution).

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

\[
\theta^{\pi,0} = \arg\min_{\theta} \mathbb{E}_{\theta^0}[ -\log( p_{\pi,\theta}(X)) ],
\]

where the density \( p_{\pi,\theta} \) is of the form:

\[
\log(p_{\pi,\theta}(x)) = \log(p_{\theta}(x^\pi)) = \sum_{j=1}^{p} \log \left( \frac{1}{\sigma_j^\pi} \varphi \left( \frac{x_j^\pi - \sum_{k=1}^{j-1} f_{j,k}^\pi(x_k^\pi)}{\sigma_j^\pi} \right) \right).
\]

(Note that the definition of \( \theta^{\pi,0} \) can be used for any permutation \( \pi \): 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:

\[
\{ f_{j,k}^{\pi,0} \}_{k=1,...,j-1} = \arg\min_{\{g_{j,k}\}_{k=1,...,j-1}} \mathbb{E}_{\theta^0}[(X_j^\pi - \sum_{k=1}^{j-1} g_{j,k}(X_k^\pi))^2].
\]

We do not enforce that \( \mathbb{E}[ f_{j,k}^{\pi,0}(X_k^\pi) ] = 0 \) since the minimization below does not depend whether we shift constants between different functions \( f_{j,k}^{\pi,0}(\cdot) \).
for different indices $k$. It holds that:

$$(\sigma_j^{\pi,0})^2 = \arg\min_{\sigma^2} \left( \log(\sigma) + \frac{1}{2\sigma^2} E_{q^0}[ (X_j^\pi - \sum_{k=1}^{j-1} f_{j,k}^{\pi,0}(X_k^\pi))^2 ] \right)$$

$$= E_{q^0}[ (X_j^\pi - \sum_{k=1}^{j-1} f_{j,k}^{\pi,0}(X_k^\pi))^2 ].$$

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

$$E_{q^0}[ -\log(p_{q^0,\theta}(X))] = \sum_{j=1}^{p} \log(\sigma_j^{\pi,0}) + C, \quad C = p \log(2\pi)^{1/2} + p/2.$$  

Since $\Pi^0$ contains all permutations such that the model (3) can be represented as in (5), the permutations $\pi \notin \Pi^0$ lead to a larger expected negative log-likelihood (larger KL-divergence between the true and best projected distribution):

$$\min_{\pi \notin \Pi^0} p^{-1} (E_{q^0}[ -\log(p_{q^0,\theta}(X))] - E_{q^0}[ -\log(p_{q^0}(X))]) \geq \xi_p,$$

where $\xi_p > 0$. The number $\xi_p$ 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 (6) can be expressed as:

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

and $\xi_p > 0$ if all functions $f_{j,k}^{0}$ are nonlinear [Peters et al., 2013].

### 2.4 Maximum likelihood estimation for order: low-dimensional setting

We assume having $n$ i.i.d. realizations $X^{(1)}, ..., X^{(n)}$ from model (3). For a $n \times 1$ vector $x = (x^{(1)}, ..., x^{(n)})^T$, 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 \( X^{(1)}, \ldots, X^{(n)} \). We consider the unpenalized maximum likelihood estimator:

\[
\hat{f}_j = \arg\min_{g_j \in \mathcal{F}_n^{\otimes j-1}} \| X_j^\pi - \sum_{k=1}^{j-1} g_{j,k}(X_k^\pi) \|_2^2, \\
(\hat{\sigma}_j^\pi)^2 = \| X_j^\pi - \sum_{k=1}^{j-1} \hat{f}_{j,k}(X_k^\pi) \|_2^2.
\]

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

\[
\hat{\pi} \in \arg\min_{\pi} \sum_{j=1}^p \log(\hat{\sigma}_j^\pi). \tag{8}
\]

Estimation of \( \hat{f}_j^\pi \) is based on \( \mathcal{F}_n \) with pre-specified basis functions \( b_r(\cdot) \) with \( r = 1, \ldots, a_n \). The basis functions could depend on the predictor variable, e.g. 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 much smaller number such as \( a_n = O(1) \) might be sufficient for estimation of the true underlying order.

2.5 Penalized 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 super-DAG (super-graph) \( D^\hat{\pi} \) of the underlying DAG \( D^0 \), where the parents of a node \( \hat{\pi}(j) \) are defined as \( \text{pa}_{D^\hat{\pi}}(\hat{\pi}(j)) = \{\hat{\pi}(k); \ k < j\} \) 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 efficiency for estimating the intervention distribution when it is based on the true DAG \( D^0 \) or a not too large super-graph \( \hat{D}^\pi \supseteq D^0 \). The task of estimating such a super-graph \( \hat{D}^\pi \supseteq D^0 \) is conceptually straightforward: once we have an estimator \( \hat{D}^\pi \) that consistently estimates a super-DAG of \( D^0 \) from (3) we can use model selection or a penalized multivariate (auto-) regression technique in the model representation.
For additive model fitting, we can either use hypothesis testing for additive models (Marra and Wood, 2011) or the Group Lasso (Ravikumar et al., 2009), or its improved version with a sparsity-smoothness penalty proposed in Meier et al. (2009). All the techniques mentioned above perform variable selection, where we denote by 

\[ \hat{D}_\pi = \{(\hat{\pi}(k), \hat{\pi}(j)); \hat{f}_{j,k}^\pi \neq 0\}, \]

the selected variables or features with indices in the original order (we obtain estimates \( \hat{f}_{j,k}^\pi \)) in the representation \((5)\) with correspondence to the indices \(\hat{\pi}(k), \hat{\pi}(j)\) in the original order); we identify these selected features in \(\hat{D}_\pi\) as the edge set of a DAG. For example with the Group Lasso, assuming some condition avoiding near collinearity of functions, i.e., a compatibility condition for the Group Lasso (Bühlmann and van de Geer, 2011, Ch.5.6, Th.8.2) and that the \(\ell_2\)-norms of the non-zero functions are sufficiently large, we obtain the screening property (since we implicitly assume that \(\hat{\pi} = \pi^0\) with high probability): with high probability and asymptotically tending to one,

\[ \hat{D}_\pi \supseteq D^0 = \{(k,j); f_{j,k}^0 \neq 0\} \] (9)

saying that all relevant variables are selected. Similarly with hypotheses testing, assuming that the non-zero \(f_{j,k}^0\) have sufficiently large \(\ell_2\)-norms, we also obtain that \((9)\) holds with high probability.

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

### 2.6 Consistent estimation of causal effects

The property in \((9)\) has an important implication for causal inference: all estimated causal effects and estimated intervention distributions based on the estimated DAG \(\hat{D}^\pi\) are consistent. In fact, using the do-calculus (Pearl, 2000, cf. (3.10)), we have for the single intervention (at variable \(X_k\)) distribution for \(X_j\), for all \(j \neq k\):

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

where \(p_D(\cdot|\text{do}(\cdot))\) denotes the intervention density based on a DAG \(D\).

We note that the screening property \((9)\) also holds when replacing \(\hat{D}^\pi\) with the full DAG induced by \(\hat{\pi}\), denoted by \(D^\pi\). Thus, the feature selection step in Section 2.5 is not needed to achieve consistent estimation of causal effects. However, a smaller DAG \(D^0 \subseteq \hat{D}^\pi \subseteq D^\pi\) typically leads to better (more efficient) estimates of the interventional distributions than the full DAG \(D^\pi\).
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 (8). Such a restriction makes the computation more tractable, and it is also 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 from [Meinshausen and Bühlmann (2006)] 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 [Ravikumar et al., 2009], ideally with a sparsity-smoothness penalty ([Meier et al., 2009]; see also [Voorman et al., 2013]). This provides us with a set of variables

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

which denotes the selected variables in the estimated conditional expectation

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

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,

$$\hat{A}_j = \{k; \ k \neq j, \ \hat{h}_{j,k} \neq 0\}.$$

We define an bidirected graph $\hat{G}$ with an edge $k \rightarrow j$ if and only if $k \in \hat{A}_j$. We emphasize that the edge set in $\hat{G}$ is different from $\hat{D}^{\pi_0}$ in Section 2.5 because for the former, the additive regression is against all other variables; similarly, the functions $\hat{h}_{j,k}(\cdot)$ are different from $\hat{f}_{j,k}^\pi(\cdot)$ in Section 2.4.

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

3.2 Restricted maximum likelihood estimator

We restrict the space of permutations in the definition of (8) such that they are “compatible” with the neighborhood selection sets $A_j$. Note that
for the estimator $\hat{\sigma}^\pi$ in (8), 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 $R_{\pi,j} = \{\pi(j - 1), \ldots, \pi(1)\} \cap \hat{\pi}(j)$. 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

$$\hat{f}_{j}^{\pi,R} = \text{argmin}_{g_j \in F_{\hat{\pi},j-1}} \|X_j^\pi - \sum_{k \in R_{\pi(j)}} g_{j,k}(X_k^\pi)\|_2^2(n),$$

and the restricted maximum likelihood estimator is

$$\hat{\pi} \in \text{argmin}_\pi \sum_{j=1}^p \log(\hat{\sigma}_{j}^{\pi,R}). \quad (10)$$

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 (10) is substantially easier than for the unrestricted MLE (8) if $\max_j |\hat{A}_j|$ is small (which is ensured if the true neighborhoods are sparse). That is, define the equivalence relation

$$\pi \sim \pi' \text{ if and only if } R_{\pi,\pi - 1(j)} = R_{\pi',\pi' - 1(j)} \forall j.$$  

Then, the set of all permutations can be partitioned in equivalence classes $\cup_r R_r$ and the minimization in (10) can be restricted to single representatives of each equivalence class $R_r$. The equivalence relation above might be more intuitive when using the notion of a restricted DAG $D_{\pi,\text{restr}}^\pi$ whose parental set for node $\pi(j)$ equals $\text{pa}_{D_{\pi,\text{restr}}} (\pi(j)) = R_{\pi,j}$. We then have that

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

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 (8). 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 \( \alpha \)-times differentiable, with \( \alpha \geq 1 \), whose derivatives up to order \( \alpha \) 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):

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

(ii)

\[
\mathbb{E}|X_j|^4 < \infty, \quad j = 1, \ldots, p
\]

\[
\sup_{f \in \mathcal{F}} \mathbb{E}|f(X_j)|^4 < \infty, \quad j = 1, \ldots, p.
\]

**(A3)** For the error variances we have \( \sigma_{\pi,0}^2 > 0 \) for all \( j = 1, \ldots, p \) and all \( \pi \).

**(A4)** The true functions \( f_{j,k}^0 \) can be approximated on a compact space \( \mathcal{C} \subset \mathbb{R} \): for all \( k \in \text{pa}_{\mathcal{D}_0}(j) \), \( j = 1, \ldots, p \),

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

where

\[
f_{n:j}^0 = \arg\min_{g_j \in \mathcal{F}_{\alpha}^{\hat{x}^{n:j}}} \mathbb{E}[g_{j,k}(X_k) - \sum_{k \in \text{pa}_{\mathcal{D}_0}(j)} g_{j,k}(X_k^0)]^2.
\]

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 (6). Then we have:

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

A proof is given in the Appendix. As mentioned before, the assumption \( \xi_p > 0 \) holds if all functions \( f_{j,k}^0 \) are nonlinear (Peters et al., 2013). 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.

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 (7) an assumption: its justification, however, is somewhat less backed up because the identifiability results from Peters et al. (2013) do not carry over immediately. The latter say that the set of correct orderings \( \Pi^0 \) can be identified from the distribution of \( X_1, \ldots, X_p \), and we require here that identifiability is given by a function of all the error variances \( \sigma^2_j \) \((j = 1, \ldots, p)\), as for the case of Gaussian errors; 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 (7).

Furthermore, assume that the number of basis functions \( a_n \) for functions in \( F_n \) is small such that assumption (A4) does not hold, e.g. \( a_n = O(1) \). We denote by

\[
(\sigma_j^{\pi,0,a_n})^2 = \min_{g_j \in F_n^{j-1}} \mathbb{E}_{\theta_0} [ (X_{\pi_j}^\pi - \sum_{k=1}^{j-1} g_{j,k}(X_{\pi_k}^\pi))^2 ],
\]

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

\[
\xi^a_p = \min_{\pi \notin \Pi^0, \pi^0 \in \Pi^0} p^{-1} \sum_{j=1}^p (\log(\sigma_j^{\pi,0,a_n}) - \log(\sigma_j^{\pi^0,0,a_n})). \quad (11)
\]

Requiring

\[
\liminf_{n \to \infty} \xi^a_p > 0
\]

is still reasonable: if (7) with \( \xi_p > 0 \) holds because of non-linearity of the additive functions (Peters et al., 2013), and see the interpretation above.
for non-Gaussian errors, we believe that it typically also holds for the best projected additive functions in $\mathcal{F}_n^{\otimes p-1}$ as long as some non-linear structure is included when using $a_n$ basis functions; here, the best projected additive function for the $j$th variable $X^\pi_j$ is defined as $f_n^\pi_j = \arg\min_{g_j \in \mathcal{F}_n^{\otimes j-1}} \mathbb{E}[(X^\pi_j - \sum_{k=1}^{j-1} g_j,k(X^\pi_k))^2]$. We also note that for $a_n \to \infty$, even when diverging very slowly, and assuming (A4) we have that $\xi_{p_{\infty}}^a \to \xi_{p_{\infty}}$ and thus $\lim \inf_{n \to \infty} \xi_{p_{\infty}}^a > 0$. In general, the choice of the number of basis functions $a_n$ is a trade-off between identifiability (due to non-linearity) and estimation accuracy: for $a_n$ small we might have a smaller value in (11), i.e. it might be that $\xi_{p_{\infty}}^{a_{\infty}} \leq \xi_{p_{\infty}}^{a_n}$ for $a_n \leq a'_{\infty}$, 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 $\epsilon_1, \ldots, \epsilon_p$ having $\mathbb{E}[\epsilon_j] = 0$ ($j = 1, \ldots, p$). Assume either of the following:

1. (A1)-(A4) hold, and $\xi_{p_{\infty}} > 0$ in formula (7).

2. (A1)-(A3) hold, and $\lim \inf_{n \to \infty} \xi_{a_{\infty}}^{a_n} > 0$ in formula (11).

Then,

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

A proof is given in the Appendix.

4.2 Restricted MLE for sparse high-dimensional setting

We consider here the restricted MLE in (10) and show that it can cope with high-dimensional settings where $p \gg n$.

The model in (1) is changing 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 $n$. 

15
In contrast to the situation from Section 4.1, we make assumptions which exclude concurvity (or collinearity) among the additive functions. 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:

\[
E_{\text{add}}[X_j|X_{\{−j\}}] = \sum_{k \in \{−j\}} h^*_jk(X_k),
\]

\( \{h^*_jk; k \in \{−j\}\} = \text{argmin}_{\{h_j \in \mathcal{F}^{p−1}; E[h_j,k(X_k)]=0\}} E[(X_j - \sum_{k \in \{−j\}} h_jk(X_k))^2]. \)

In general, some variables are irrelevant: we denote the set of relevant variables by \( A_j \):

\[ A_j \subseteq \{1, \ldots, p\} \setminus j \text{ is the smallest set such that } E_{\text{add}}[X_j|X_{\{−j\}}] = E_{\text{add}}[X_j|X_{A_j}]. \]

We assume the following:

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

\[ E_{\text{add}}[(X_j - E_{\text{add}}[X_j|X_{A_j\setminus k}])|X_k] \neq 0. \]

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

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

A proof is given in the Appendix. Lemma 1 justifies, for the population case, to pursue preliminary neighborhood selection followed by restricted maximum likelihood estimation: because \( \text{pa}(j) \subseteq A_j \), the restriction in the maximum likelihood estimator is appropriate and a true permutation in \( \Pi^0 \) satisfies the restriction \( R \) (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:

- (i) \( \hat{A}_j \supseteq A_j \ (j = 1, \ldots, p) \),
- (ii) \( \max_{j=1,\ldots,p} |\hat{A}_j| \leq M < \infty \).

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 group compatibility condition (for the basis vectors) and a beta-min condition about the minimal size of the \( \ell_2 \)-norm of the coefficients for the basis functions of the active variables in \( A_j \).
see Bühlmann and van de Geer (2011, Ch.5.6,Th.8.2). Assumption (B2,ii) can be ensured by assuming $\max_j |A_j| \leq M_1 < \infty$ and e.g. group restricted eigenvalue assumptions for the design matrix (with the given basis); see Zhang and Huang (2008) and van de Geer et al. (2011) 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 $|B| \leq M$, with $M$ as in (B2), denote by $h_{j,g}^B = (X_j - \sum_{k \in B} g_k(X_k))^2$. For some $0 < K < \infty$, it holds that

$$\max_{j=1,\ldots,p} \max_{B \subseteq \{1,\ldots,p\} \setminus j, |B| \leq M} \sup_{g \in F \ominus \{j\}} \rho_K(h_{j,g}^B) \leq D_1 < \infty,$$

where

$$\rho_K^2(h_{j,g}^B) = 2K^2\mathbb{E}_{g_0}\left[\exp(\|h_{j,g}^B(X)\|/K) - 1 - |h_{j,g}^B(X)|/K\right].$$

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

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

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)

$$\max_j \mathbb{E}|X_j|^4 \leq D_3 < \infty,$$

$$\max_j \sup_{f \in F} \mathbb{E}|f(X_j)|^4 \leq D_4 < \infty.$$

(B4) For the error variances we have $\min_{\pi} \min_j \sigma_j^{\pi,0} \geq L > 0$.

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 $\xi_p$ in (7):
\[
\min(\sqrt{\log(p)/n}, \mathbb{E}[(f_{j,k}^0(X_k) - f_{n;j,k}(X_k))^2]) = o(\xi_p).
\]

2. (A1), (A4) and (B1)-(B4) hold, and for $\xi_{ap}$ in (11):
\[
\min(\sqrt{\log(p)/n}, \mathbb{E}[(f_{j,k}^0(X_k) - f_{n;j,k}(X_k))^2]) = o(\xi_{ap}).
\]

Then, for the restricted maximum likelihood estimator in (10):
\[
\mathbb{P}[\hat{\pi} \in \Pi^0] \to 1 \ (n \to \infty).
\]

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

(B_{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
\[
f_{j,k}^0(x) = \sum_{r=1}^{\infty} a_{j,k,r}^0 b_{j,k,r}(x)
\]
with smoothness condition:
\[
\sum_{r=k}^{\infty} |\alpha_{j,k,r}^0| \leq Ck^{-\beta}.
\]

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

5 Computation

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 consists of two corresponding parts: IncEdge is a greedy procedure providing an estimate $\hat{\pi}$ for Equation (8) and Prune performs the feature selection. Section 3.1 discusses the benefits of performing a preliminary neighborhood selection before estimating the causal order. We call the corresponding part of our algorithm.
Figure 1: Step PNS. For each variable the set of possible parents is reduced. This leads to a considerable computational gain of the whole procedure.

PNS. The combination PNS + IncEdge provides an estimate for equation (10).

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

5.1 Preliminary Neighborhood Selection: EstCIG

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 using the R-function \texttt{gamboost} from the package \texttt{mboost} (Hothorn et al., 2010). We select the ten variables as the ones which have been picked most during 100 iterations of the boosting method; hereby, we only consider variables that have been picked at least twice during the iterations. The sets $\hat{A}_j$ obtained by this procedure estimate the Markov blanket of node $X_j$ in the underlying DAG. The new graph $\hat{G}$ contains the sets $\hat{A}_j$ as parental sets for $X_j$. Figure 1 summarizes this step. Importantly, we do not disregard true parents if the sample size is large enough (Section 4.2, Lemma 1).

5.2 Estimating the correct order by greedy search: IncEdge

Let us first consider the situation without PNS. Searching over all permutations $\pi$ for finding $\hat{\pi}$ in (8) 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 $(k, j)$ between nodes $k$ and $j$ that corresponds to the largest gain in log-likelihood. We therefore
Figure 2: Step **IncEdge**. At each iteration the edge leading to the largest decrease of the negative log-likelihood is included.

Compute the score function in (8), with $\pi$ corresponding to the current DAG,

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

and construct a matrix, whose entry $(k,j)$ specifies by how much this score is reduced after adding the edge $k \rightarrow j$ and therefore allowing a non-constant function $f_{j,k}$ (see Figure 2). For implementation, we use gam (based on penalized regression splines with ten basis functions) 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 $j$th 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 1000 nodes (see Section 6.1.3).

### 5.3 Pruning of the DAG by feature selection: **Prune**

Section 2.5 describes the penalized regression that corresponds to pruning the DAG that has been estimated by Step **IncEdge**, see Figure 3. We implement this task by applying gam from the R-package mgcv and take the
Figure 3: Step Prune. For each node, variable selection techniques are exploited to remove non-relevant edges.

covariates with reported $p$-value lower or equal to 0.001, independently of the sample size. Many other choices would be possible, too.

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 PNS removes additional “superfluous” edges. The structural Hamming distance to the true graph, however, may reduce significantly after performing Prune, see Section 6.1.2.

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). In Section 6.1.4 we further check the robustness of our method against model misspecification, i.e. in the case of non-Gaussian noise or non-additive 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 $\pi^0$ and connect each pair of nodes with a probability $p_{\text{conn}}$. 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 bandwidth of one and add Gaussian noise with variance uniformly sampled between 0.2 and 0.4. All nodes without parents have a variance between 1 and 2. The experiments are based on 100 repetitions if the description does not say differently.
6.1.1 Structural Intervention Distance

As a performance measure, we consider the recently proposed structural intervention distance (SID), see Peters and Bühlmann (2013b). The SID is well suited for quantifying the correctness of an order among variables, mainly in terms of inferring causal effects afterwards. 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. Section 6.1.2 shows that the SID is almost unaffected by the feature selection step with penalized regression and hence, regularization for feature selection is only needed for selecting a right-sized model.

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 \texttt{IncEdge} is used, the SHD is usually large because the output is a fully connected graph. Only after the Step \texttt{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 consumption time is reduced significantly if we include the preliminary neighborhood selection \texttt{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 distribution. We compare the performance of our method to greedy equivalence search (GES) (Chickering, 2002), the PC algorithm (Spirtes et al., 2000), the conservative PC algorithm (CPC) (Ramsey et al., 2006), LiNGAM (Shimizu et al., 2006) and regression with subsequent independence tests (RESIT) (Mooij et al., 2009; Peters et al., 2013). 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 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...
Figure 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 $p$.

size $p = 10$ and $p = 100$, whereas in both cases, the sample size is $n = 200$. RESIT is not applicable for graphs with $p = 100$. Figure 5 shows that the proposed method outperforms the other approaches both in terms of SID and SHD.

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

6.1.4 Robustness against model misspecification

This work focuses on the additive model (1). 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 (Peters et al., 2013, cf.). The score functions (8) and (10) and their corresponding optimization problems, however, depend on the model assumptions.

As a first experiment we examine deviations from the Gaussian noise assumption by setting $\varepsilon_j = \text{sign}(N_j)|N_j|^\gamma$ with $N_j \sim N(0, \sigma^2_j)$ for different exponents $0.1 \leq \gamma \leq 4$. Only $\gamma = 1$ corresponds to normally distributed noise. Figure 6 shows the change in SID when varying $\gamma$.

As a second experiment, we examine deviations from additivity by sim-
Figure 5: SHD (left) and SID (right) for $p = 10$ (top) and $p = 100$ (bottom).

Figure 6: SID for $p = 25$ and $n = 300$ in the case of misspecified models. The plot shows the SID for deviations of the noise from a normal distribution (only $\gamma = 1$ corresponds to Gaussian noise).
Figure 7: SID for $p = 25$ and $n = 300$ in the case of misspecified models. The top and lower panel show the SID for deviations from additivity for sparse and non-sparse truths respectively (only $\omega = 1$ corresponds to a fully additive model).

Calculating from the model

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

for different values of $\omega \in [0, 1]$ and Gaussian noise. Note that $\omega = 1$ corresponds to the fully additive model [3], whereas for $\omega = 0$, the value of $X_j$ is given as a non-additive function of all its parents. Figure 7 shows the result for a sparse truth with expected number of $p$ edges (top) and a non-sparse 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 non-additive functions. If the underlying truth is non-sparse, 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).
Figure 8: Gene expressions in isoprenoid pathways. The twenty best scoring edges provided by the method CAM.

6.2 Real data

We apply our methodology to microarray data described by Wille et al. (2004). The authors concentrate on 39 genes (118 observed samples) on two isoprenoid pathways in Arabidopsis thaliana. The dashed edges in Figures 8 and 9 indicate the causal direction within each pathway. While Wille et al. (2004) apply graphical Gaussian models in order to estimate the underlying interaction network by an undirected model, 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 8 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 (Meinshausen and Bühmann, 2010) 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 (Meinshausen and Bühmann, 2010). These edges are shown in Figure 9. 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 → CMK does not appear in Figure 8 since it was
Figure 9: Gene expressions in isoprenoid pathways. Edges estimated by stability selection: all directions are in correspondence with the direction of the pathways.

ranked as the 22nd best scoring edge.

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 [Peters et al. 2013]. 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 non-regularized (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 than for non-identifiable models such as the popular linear Gaussian structural equation model.

7.1 Extensions

The estimation principle of first pursuing order search based on non-regularized maximum likelihood and then using penalized regression for feature selection works with other structural equation models where the underlying DAG is identifiable from the observational distribution. Closely related examples include non-linearly transformed additive structural equation models (Zhang and Hyvärinen 2009) or Gaussian structural equation models with same error variances (Peters and Bühlmann 2013).

If the DAG $D$ is non-identifiable from the distribution $P$, the methodology needs to be adapted; see also Remark 1 considering the linear Gaussian SEM. The true orders $\Pi^0$ 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 e.g. the $\ell_0$-penalty (Chickering 2002; van de Geer and Bühlmann 2013).

Finally, it would be interesting to extend (sparse) permutation search to (non-identifiable) models with hidden variables (Spirtes et al. 2000; Pearl 2000; Colombo et al. 2012) or with graph structures allowing for cycles (Spirtes 1995; Richardson 1996; Mooij et al. 2011).

8 Appendix

8.1 Proof of Theorem 1

We proceed by showing that:

$$\sum_{j=1}^{p} \log(\hat{\sigma}_{\pi}^{j}) \geq \sum_{j=1}^{p} \log(\sigma_{\pi,0}^{j}) - \Delta_n \forall \pi \notin \Pi^0, \quad (12)$$

$$\sum_{j=1}^{p} \log(\hat{\sigma}_{\pi,0}^{j}) \leq \sum_{j=1}^{p} \log(\sigma_{j}^{0}) + \Delta_n \forall \pi^0 \in \Pi^0, \quad (13)$$
where $\Delta_n/p = o_P(\xi_p)$. Using (6) and (12)-(13) we obtain: for $n$ sufficiently large,

$$\sum_{j=1}^p \log(\hat{\sigma}_j^2) > \sum_{j=1}^p \log(\hat{\sigma}_j^2) \forall \pi \notin \Pi^0, \forall \pi^0 \in \Pi^0,$$

and thus, for $\hat{\pi}$ being a minimizer of the negative log-likelihood:

$$\mathbb{P}[\hat{\pi} \in \text{argmin}_\pi \sum_{j=1}^p \log(\hat{\sigma}_j^2) \subseteq \Pi^0] \to 1 \ (n \to \infty).$$

### 8.1.1 Proof of (12)

For $\pi \notin \Pi^0$ we have:

$$(\sigma_j^{\pi,0})^2 = \mathbb{E}[(X_j^\pi - \sum_{k=1}^{j-1} f_{j,k}^{\pi,0}(X_k^\pi))^2] = \min_{g \in \mathcal{F}^{j-1}} \mathbb{E}[(X_j^\pi - \sum_{k=1}^{j-1} g_k(X_k^\pi))^2]$$

$$\leq \min_{g \in \mathcal{F}^{j-1}} \mathbb{E}[(X_j^\pi - \sum_{k=1}^{j-1} g_k(X_k^\pi))^2]$$

$$= \min_{g \in \mathcal{F}^{j-1}} (\|X_j^\pi - \sum_{k=1}^{j-1} g_k(X_k^\pi)\|^2_2/n - (P_n - P)(h_{j,g}^\pi))$$

$$\leq \min_{g \in \mathcal{F}^{j-1}} \|X_j^\pi - \sum_{k=1}^{j-1} g_k(X_k^\pi)\|^2_2/n + \sup_{g \in \mathcal{F}^{j-1}} |(P_n - P)(h_{j,g}^\pi)|$$

$$= (\hat{\sigma}_j^2) + \sup_{g \in \mathcal{F}^{j-1}} |(P_n - P)(h_{j,g}^\pi)|$$

$$\leq (\hat{\sigma}_j^2) + \Delta_{n,j}^\pi,$$

where $h_{j,g}^\pi(X) = (X_j^\pi - \sum_{k=1}^{j-1} g_k(X_k^\pi))^2$ and $\Delta_{n,j}^\pi = \sup_{g \in \mathcal{F}^{j-1}} |(P_n - P)(h_{j,g}^\pi)|$. We will show in Lemma 2 stated below that

assuming (A1)-(A2): $\max_{\pi \notin \Pi^0} \Delta_{n,j}^\pi = o_P(1).$ (14)

Then, using Taylor expansion and (A3), this then implies (12).

### 8.1.2 Proof of (13)

For $\pi \in \Pi^0$ we argue the other way round. Denote by

$$f_{n,j}^{\pi,0} = \arg\min_{g \in \mathcal{F}^{j-1}} \mathbb{E}[(X_j^\pi - \sum_{k=1}^{j-1} g_k(X_k^\pi))^2].$$

29
Then:

$$
(\hat{\sigma}^\pi_j)^2 = \min_{g \in \mathcal{F}_{n}^{j-1}} \|X^\pi_j - \sum_{k=1}^{j-1} g_k(X^\pi_k)\|^2/n
$$

$$
\leq \mathbb{E}[(X^\pi_j - \sum_{k=1}^{j-1} f_{n;j,k}^\pi(X^\pi_k))^2] + \sup_{g \in \mathcal{F}_{n}^{j-1}} |(P_n - P)(h^\pi_{j,g})|
$$

$$
\leq \mathbb{E}[(X^\pi_j - \sum_{k=1}^{j-1} f_{j,k}^\pi(X^\pi_k))^2] + \gamma_{n,j}^\pi + \sup_{g \in \mathcal{F}_{n}^{j-1}} |(P_n - P)(h^\pi_{j,g})|
$$

$$
= (\sigma_0^\pi_j)^2 + \gamma_{n,j}^\pi + \Delta_{n,j}^\pi,
$$

$$
\gamma_{n,j}^\pi = \mathbb{E}[(X^\pi_j - \sum_{k=1}^{j-1} f_{n;j,k}^\pi(X^\pi_k))^2] - \mathbb{E}[(X^\pi_j - \sum_{k=1}^{j-1} f_{j,k}^\pi(X^\pi_k))^2].
$$

We will show below in Section 8.1.3 that \( \max_{\pi \in \Pi^0, j=1,\ldots,p} |\gamma_{n,j}^\pi| = o(1). \) Therefore,

$$
(\hat{\sigma}^\pi_j)^2 \leq (\sigma_0^\pi_j)^2 + \gamma_{n,j}^\pi + \Delta_{n,j}^\pi,
$$

$$
\max_{\pi \in \Pi^0, j=1,\ldots,p} |\gamma_{n,j}^\pi| = o(1),
$$

assuming (A1)-(A2) (see Lemma 2): \( \max_{\pi \in \Pi^0, j=1,\ldots,p} \Delta_{n,j}^\pi = o_P(1). \)

Then, using Taylor expansion and (A3), this then implies (13).

### 8.1.3 Bound for \( \gamma_{n,j}^\pi \)

We have that

$$
\gamma_{n,j}^\pi \leq \sum_{k=1}^{j-1} (f_{j,k}^\pi(X^\pi_k) - f_{n;j,k}^\pi(X^\pi_k))^2 \right)^{1/2}
$$

$$
\cdot \mathbb{E}[(2X^\pi_j - \sum_{k=1}^{j-1} f_{j,k}^\pi(X^\pi_k) - \sum_{k=1}^{j-1} f_{n;j,k}^\pi(X^\pi_k))^2]^{1/2}.
$$

We know that for \( \pi \in \Pi^0 \), the summation can be restricted to the indices \( \text{pa}_{D^0}(j) \). By (A2)(ii), the term \( \mathbb{E}[(2X^\pi_j - \sum_{k=1}^{j-1} f_{j,k}^\pi(X^\pi_k) - \sum_{k=1}^{j-1} f_{n;j,k}^\pi(X^\pi_k))^2] \) is bounded, and we only need to deal with the error

$$
\mathbb{E}[(\sum_{k \in \text{pa}_{D^0}(j)} f_{j,k}^0(X_k) - f_{n;j,k}^0(X_k))^2]^{1/2}.
$$

(15)
Bound for \([15]\):
The true functions \(f_{j,k}^0\) can be approximated on a compact space \(C \subset \mathbb{R}\):
\[
\mathbb{E}[(f_{j,k}^0(X_k) - f_{n,j,k}^0(X_k))^2 I(X_k \in C)] = o(1).
\]
due to assumption (A4). Outside the compact space, we use Hölder’s inequality: for any \(\kappa > 0\)
\[
\mathbb{E}[(f_{j,k}^0(X_k) - f_{n,j,k}^0(X_k))^2 I(X_k \notin C)] \leq \mathbb{E}[(f_{j,k}^0(X_k) - f_{n,j,k}^0(X_k))^{2(1+\kappa)}]^{1/(1+\kappa)} \mathbb{P}[X_k \notin C]^{\kappa/(1+\kappa)} = o(1)
\]
invoking again (A2)(ii). This ensures that
\[
\gamma_{n,j}^\pi = o(1) \text{ for all } \pi, j
\]
and since \(p < \infty\) is fixed, we also have a uniform bound.

8.1.4 Uniform convergence result

Lemma 2. Denote by
\[
\Delta_{n,j}^\pi = \sup_{g \in \mathcal{F}^{\otimes j-1}} |(P_n - P)(h_{j,g}^\pi)| \text{ where } h_{j,g}^\pi(X) = (X_j^\pi - \sum_{k=1}^{j-1} g_k(X_k^\pi))^2.
\]
Assume (A1)-(A2). Then,
\[
\max_{\pi,j=1,\ldots,p} \Delta_{n,j}^\pi = o_P(1) \text{ as } n \to \infty.
\]

Proof: Since \(p\) is fixed, it suffices to prove that \(\Delta_{n,j}^\pi = o_P(1)\) for all \(\pi, j\).
Denote by \(\mathcal{H}_B(u, \mathcal{L}_j^\pi, L_2(P))\) the logarithm of the bracketing number with \(L_2(P)\)-norm for the function class
\[
\mathcal{L}_j^\pi = \{h_{j,g}^\pi; g \in \mathcal{F}^{\otimes j-1}\}.
\]
We then have that
\[
\sup_{g \in \mathcal{F}^{\otimes j-1}} n^{-1} \sum_{i=1}^n (h_{j,g}^\pi(X_i) - \mathbb{E}[h_{j,g}^\pi(X)]) = o_P(\int_0^R \mathcal{H}_B^{1/2}(u, \mathcal{L}_j^\pi, L_2(P)) du).
\]
(16)

We thus need to evaluate the log bracketing number \(\mathcal{H}_B^{1/2}(u, \mathcal{L}_j^\pi, L_2(P))\) for fixed \(j\) and \(\pi\). We write for the function \(h_{j,g}^\pi(x) = (x_j^\pi - \sum_{k=1}^{j-1} g_k(x_k^\pi))^2 =: \tilde{g}_L^2(x)\). If we have brackets \(\tilde{g}_L(\cdot) \leq \tilde{g}(\cdot) \leq \tilde{g}_U(\cdot)\), we have also the brackets \(\tilde{g}_L^2(\cdot) \leq \tilde{g}_L^2(\cdot) \leq \tilde{g}_U^2(\cdot)\). For the \(L_2(P)\) distance we further note that
\[
\|\tilde{g}_L^2 - \tilde{g}_U^2\|_{L_2(P)} \leq \|\tilde{g}_L - \tilde{g}_U\|_{L_4(P)} \|\tilde{g}_L + \tilde{g}_U\|_{L_4(P)}.
\]
Because $\|\tilde{g}_L + \tilde{g}_U\|_{L_4(P)} \leq C_4 < \infty$ (where $C_4$ does not depend on $\pi$ and $j$, see (A2)(ii)), we can use bracketing with respect to the $L_4(P)$-norm for the non-squared function $\tilde{g}(x) = (x_\pi - \sum_{k=1}^{j-1} g_k(x_\pi^k))$. We have for fixed $j$ and $\pi$:

$$H_B(u, L_\pi^j, L_2(P)) \leq C_5 H_B(u, F^{\oplus j-1} \circ \pi, L_4(P))$$

where the constant $C_5$ does not depend on $\pi$ and $j$, and $F^{\oplus j-1} \circ \pi$ contains functions of the form

$$f \circ \pi : \mathbb{R}^p \to \mathbb{R}, \quad x \mapsto \sum_{k=1}^{j-1} f_k(x_\pi^k)$$

for $f \in F^{\oplus j-1}$. Hence,

$$\max_{\pi, j} H_B(u, L_\pi^j, L_2(P)) \leq C_5 \max_{\pi, j} H_B(u, F^{\oplus j-1} \circ \pi, L_4(P)) \leq C_5 \max_{\pi} H_B(u, F^{\oplus p-1} \circ \pi, L_4(P)).$$

Furthermore, since we can cover $F^{\oplus p-1}$ by at most $(p-1)$ times as many brackets with distance $u/(p-1)$ in comparison to covering $F$ with brackets with distance $u$:

$$\max_{\pi} H_B(u, F^{\oplus p-1} \circ \pi, L_4(P)) \leq \max_{j=1,\ldots,p} (p-1) H_B(u/(p-1), F, L_4(P_j)),$$

where $P_j$ is the marginal distribution of $X_j$. The bracketing number for $F$ is bounded by assumption (B3,ii): for some positive constant $C_6 < \infty$:

$$\max_{j=1,\ldots,p} H_B(u, F, L_4(P_j)) \leq C_6 u^{-V} \max_{j=1,\ldots,p} (\sum_{m=1}^{\infty} (M_m^2 \mathbb{P}[X_j \in I_m])^{V/(V+4)}(V+4)^{V/4},$$

with $V = 1/\alpha$, see van der Vaart (1998, Ex.19.9). Putting the last three bounds together we arrive at: for some $C_7 < \infty$,

$$\max_{\pi, j} H_B^{1/2}(u, L_\pi^j, L_2(P)) \leq C_7 p^{1/2+V/2} u^{-V/2} \max_{j=1,\ldots,p} (\sum_{m=1}^{\infty} (M_m^2 \mathbb{P}[X_j \in I_m])^{V/(V+4)}(V+4)^{V/4}/8.$$ 

Therefore, due to assumption (A2)(i): for some $C_8 < \infty$,

$$\max_{\pi, j} \int_0^R H_B^{1/2}(u, L_\pi^j, L_2(P)) du \leq C_8 p^{1/2+V/2} < \infty.$$
Thus, since $V = 1/\alpha$, the bound in (16) becomes:

$$\max_{\pi,j} \Delta_{n,j}^\pi = \max_{\pi,j} \sup_{g \in \mathcal{F} \oplus j} n^{-1} \sum_{i=1}^{n} (h_{j,g}^\pi(X_i) - \mathbb{E}[h_{j,g}^\pi(X)]) = o_p(1).$$

This completes the proof. \hfill \square

8.2 Proof of Lemma 1

Write

$$\mathbb{E}_{\text{add}}[X_j|\{X_k; k \neq j\}] = \sum_{k \in A_j} h_{jk}^*(X_k),$$

and suppose that $\text{pa}(j) \not\subseteq A_j$. Let $k^* \in \text{pa}(j)$ and $k^* \not\in A_j$. Then, due to (B1), $X_{k^*}$ would still have an influence on $X_j$ after having subtracted the additive effect from $X_{A_j}$, and hence $\mathbb{E}_{\text{add}}[X_j|\{X_k; k \neq j\}]$ must contain $X_{k^*}$ as well, i.e., $k^* \in A_j$. But this is a contradiction and therefore, it must hold that $\text{pa}(j) \subseteq A_j$. \hfill \square

8.3 Proof of Theorem 2

The proof under assumption 1. in the theorem follows exactly along the same lines as for Theorem 1. With assumption 2., we replace $\sigma_{\pi,0}^j$ in (12)-(13) by $\sigma_{\pi,0,a_n}^j$.

8.4 Proof of Theorem 3

Instead of the difference

$$|\hat{\sigma}_j^2 - (\sigma_{\pi,0})^2|$$

as in in the proofs of (12) and (13), we need to analyze:

$$\max_j \max_{\pi} |\hat{\sigma}_{\pi,R}^j - (\sigma_{\pi,0})^2|$$

$$\leq \max_j \max_{B \subseteq \{1,\ldots,p\}\setminus j; |B| \leq M<\infty} |\hat{\sigma}_B^j - (\sigma_{B,0})^2|.$$
Furthermore, the bound in for $\gamma_{n,j}$ in Section 8.1.3 can be taken as there, by noting that $p_{D}\in\mathcal{P}(j)$ is a set of bounded cardinality (due to assumption (B2)).

### 8.4.1 Uniform convergence result

Using the proof of (12) and (13), we only need to replace Lemma 2 by the following.

**Lemma 3.** Denote by $\Delta_{n,j} = \sup_{g \in \mathcal{F} \ni |B|} |(P_{n} - P)h_{j,g}|$ where $h_{j,g}(X) = (X - \sum_{k \in B} g_{k}(X_{k}))^{2}$. Then:

$$\max_{j} \max_{|B| \leq M} \Delta_{n,j} = O_{P}(\sqrt{\log(p_{n})/n}),$$

as $p_{n} \to \infty$ and $n \to \infty$.

Proof: We largely follow the proof of Lemma 2 but for the sake of completeness, we provide the details here. We invoke a uniform inequality from van de Geer (2000, Th.5.11): due to assumption (B3,i) we have:

$$P\left[ \sup_{g \in \mathcal{F} \ni |B|} n^{-1/2} \sum_{i=1}^{n} (h^{B}_{j,g}(X_{i}) - \mathbb{E}[h^{B}_{j,g}(X)]) \geq a \right] \leq C_{1} \exp \left( -\frac{a^{2}}{2C_{2}R^{2}} \right),$$

$$a \geq C_{3} \int_{0}^{R} \mathcal{H}_{B}^{1/2}(u, \mathcal{L}^{B}_{j}, P) du \vee R.$$

Here $\mathcal{H}_{B}(u, \mathcal{L}^{B}_{j}, P)$ is the logarithm of the bracketing number with $L_{2}(P)$-norm for the function class

$$\mathcal{L}^{B}_{j} = \{ h^{B}_{j,g}, g \in \mathcal{F} \ni |B| \}. $$

Therefore, by using the union bound over $B$ with $|B| \leq M$ (the cardinality of such $B$'s is bounded by $2^{M} < \infty$) $j \in \{1, \ldots, p\}$:

$$\max_{j=1, \ldots, p} \max_{B, |B| \leq M} \sup_{g \in \mathcal{F} \ni |B|} n^{-1} \sum_{i=1}^{n} (h^{B}_{j,g}(X_{i}) - \mathbb{E}[h^{B}_{j,g}(X)])$$

$$= O_{P}(\log(p)^{1/2}n^{-1/2} \max_{j} \int_{0}^{R} \mathcal{H}_{B}^{1/2}(u, \mathcal{L}^{B}_{j}, P) du). \quad (17)$$

We thus need to evaluate the log bracketing number $\mathcal{H}_{B}^{1/2}(u, \mathcal{L}^{B}_{j}, P)$ for fixed $B$ and $j$. We write for the function $h^{B}_{j,g}(x) = (x - \sum_{k \in B} g_{k}(x_{k}))^{2} =: \ldots$
If we have brackets \( \tilde{g}_L(\cdot) \leq \tilde{g}(\cdot) \leq \tilde{g}_U(\cdot) \), we have also the brackets 
\[
\tilde{g}_L^2(\cdot) \leq \tilde{g}^2(\cdot) \leq \tilde{g}_U^2(\cdot).
\]
For the \( L_2(P) \) distance we further note that
\[
\|\tilde{g}_L^2 - \tilde{g}_U^2\|_{L_2(P)} \leq \|\tilde{g}_L - \tilde{g}_U\|_{L_4(P)} \|\tilde{g}_L + \tilde{g}_U\|_{L_4(P)}.
\]
Because \( \|\tilde{g}_L + \tilde{g}_U\|_{L_4(P)} \leq C_4 < \infty \) (where \( C_4 \) does not depend on \( B \) and \( j \), see (B3,iii)), we can use bracketing with respect to the \( L_4(P) \)-norm for the non-squared function \( \tilde{g}(x) = (x_j - \sum_{k \in B} g_k(x_k)) \). We have for fixed \( j \) and \( B \):
\[
\mathcal{H}_B(u, \mathcal{L}_j^B, L_2(P)) \leq C_5 \mathcal{H}_B(u, \mathcal{F}^{\leq |B|}, B, L_4(P))
\]
where the constant \( C_5 \) does not depend on \( B \) and \( j \). Hence,
\[
\max_{j,B} \mathcal{H}_B(u, \mathcal{L}_j^B, P) \leq C_5 \max_B \mathcal{H}_B(u, \mathcal{F}^{\leq |B|}, L_4(P)) \leq C_5 \mathcal{H}_B(u, \mathcal{F}^{\leq M}, L_4(P)).
\]
Furthermore, since we can cover \( \mathcal{F}^{\leq M} \) by at most \( M \) times as many brackets with distance \( u/M \) in comparison to covering \( \mathcal{F} \) with brackets with distance \( u \):
\[
\mathcal{H}_B(u, \mathcal{F}^{\leq M}, L_4(P)) \leq \max_{j=1,\ldots,p} M \mathcal{H}_B(u/M, \mathcal{F}, L_4(P_j)),
\]
where \( P_j \) is the marginal distribution of \( X_j \). The bracketing number for \( \mathcal{F} \) is bounded by assumption (B3,ii): for some positive constant \( C_6 < \infty \),
\[
\max_{j=1,\ldots,p} \mathcal{H}_B(u, \mathcal{F}, L_4(P_j)) \leq C_6 u^{-V} \max_j \left( \sum_{m=1}^{\infty} (M_m^2 \mathbb{P}[X_j \in I_m])^{V/(V+4)}(V+4)/4 \right),
\]
with \( V = 1/\alpha \), see van der Vaart (1998) Ex.19.9. Putting the last three bounds together we arrive at: for some \( C_7 < \infty \),
\[
\max_{j,B} \mathcal{H}_B^{1/2}(u, \mathcal{L}_j^B, P) \leq C_7 u^{-V/2} \max_j \left( \sum_{m=1}^{\infty} (M_m^2 \mathbb{P}[X_j \in I_m])^{V/(V+4)}(V+4)/8 \right).
\]
Therefore, due to assumption (B3,ii): for some \( C_8 < \infty \),
\[
\max_j \max_B \int_0^R \mathcal{H}_B^{1/2}(u, \mathcal{L}_j^B, P) du \leq C_8.
\]
Thus, since \( V = 1/\alpha \), the bound in (17) becomes:
\[
\max_{j=1,\ldots,p} \max_{B,|B| \leq M} \sup_{g \in \mathcal{F}^{\leq |B|}} n^{-1} \sum_{i=1}^{n} \left( h^{B,j,g}_i(X_i) - \mathbb{E}[h^{B,j,g}(X)] \right) = O_p(\log(p)^{1/2}n^{-1/2}).
\]
Therefore:

\[
\max_j \max_{B, |B| \leq M} \sup_{g \in \mathcal{F}^{|B|}} \Delta^B_{n,j} = \max_j \max_{B, |B| \leq M} \sup_{g \in \mathcal{F}^{|B|}} \left| n^{-1} \sum_{i=1}^n h^B_{j,g}(X_i) - \mathbb{E}[h^B_{j,g}(X)] \right| = O_P(\log(p)^{1/2}n^{-1/2}).
\]

This completes the proof. \qed

References

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Verlag.

Chickering, D. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554.

Colombo, D., Maathuis, M., Kalisch, M., and Richardson, T. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics*, 40:294–321.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning; Data Mining, Inference and Prediction*. Springer, New York, second edition.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2010). Model-based boosting 2.0. *Journal of Machine Learning Research*, 11:2109–2113.

Lauritzen, S. (1996). *Graphical Models*. Oxford University Press.

Marra, G. and Wood, S. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, 55:2372–2387.

Meier, L., van de Geer, S., and Bühlmann, P. (2009). High-dimensional additive modeling. *Annals of Statistics*, 37:3779–3821.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society, Series B*, 72:417–473.
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, 24th Annual Conference on Neural Information Processing Systems (NIPS 2011)*.

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)*, pages 745–752.

Nowzohour, C. and Bühlmann, P. (2013). Score-based causal learning in additive noise models. In preparation.

Pearl, J. (2000). *Causality: models, reasoning and inference*. Cambridge Univ. Press.

Peters, J. and Bühlmann, P. (2013a). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika (accepted)*.

Peters, J. and Bühlmann, P. (2013b). Structural intervention distance (SID) for evaluating causal graphs. arXiv:1306.1043.

Peters, J., Mooij, J., Janzing, D., and Schölkopf, B. (2013). Causal discovery with continuous additive noise models. arXiv:1309.6779.

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)*, pages 401–408.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society Series B*, 71:1009–1030.

Richardson, T. (1996). A discovery algorithm for directed cyclic graphs. In *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 454–461.

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*, volume 22, page 1278. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Shimizu, S., Hoyer, P., Hyvärinen, A., and Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030.
Spirtes, P. (1995). Directed cyclic graphical representations of feedback models. In Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI), pages 491–499.

Spirtes, P., Glymour, C., and Scheines, R. (2000). Causation, Prediction, and Search. MIT Press, second edition.

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), pages 584–590.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B, 58:267–288.

van de Geer, S. (2000). Empirical Processes in M-Estimation. Cambridge University Press.

van de Geer, S. (2013). On the uniform convergence of empirical norms and inner products, with application to causal inference. Preprint.

van de Geer, S. and Bühlmann, P. (2013). $\ell_0$-penalized maximum likelihood for sparse directed acyclic graphs. Annals of Statistics, 41:536–567.

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). Electronic Journal of Statistics, 5:688–749.

van der Vaart, A. (1998). Asymptotic Statistics. Cambridge University Press.

Voorman, A., Shojaie, A., and Witten, D. (2013). Graph estimation with joint additive models. Biometrika. To appear.

Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (Lasso). IEEE Transactions on Information Theory, 55:2183–2202.

Wille, A., Zimmermann, P., Vranova, E., Frholz, 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 Biology, 5(11):R92.
Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 69:49–67.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36:1567–1594.

Zhang, K. and Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 647–655.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.