Statistics, 2016
Vol. 50, No. 3, 471–485, http://dx.doi.org/10.1080/02331888.2015.1060237

Taylor & Francis
Taylor & Francis Group

Score-based causal learning in additive noise models

Christopher Nowzohour* and Peter Bühlmann

Seminar für Statistik, ETH Zürich, Zürich, Switzerland

(*Received 14 April 2014; accepted 1 June 2015*)

Downloaded by [ETH Zurich] at 08:49 06 June 2016

Given data sampled from a number of variables, one is often interested in the underlying causal relationships in the form of a directed acyclic graph. In the general case, without interventions on some of the variables it is only possible to identify the graph up to its Markov equivalence class. However, in some situations one can find the true causal graph just from observational data, for example, in structural equation models with additive noise and nonlinear edge functions. Most current methods for achieving this rely on nonparametric independence tests. One of the problems there is that the null hypothesis is independence, which is what one would like to get evidence for. We take a different approach in our work by using a penalized likelihood as a score for model selection. This is practically feasible in many settings and has the advantage of yielding a natural ranking of the candidate models. When making smoothness assumptions on the probability density space, we prove consistency of the penalized maximum likelihood estimator. We also present empirical results for simulated scenarios and real two-dimensional data sets (cause–effect pairs) where we obtain similar results as other state-of-the-art methods.

Keywords: causal inference; structure learning; hidden variables; latent variables; path diagrams; structural equation models

1. Introduction

Statistical causal inference is an important but relatively new field. Traditionally, most statistical statements and assertions are associational (X and Y are correlated), rather than causal (changes in X cause changes in Y). While the former are statements about the joint distribution, the latter are about the underlying causal mechanisms. In practice, the relevant question often is whether variable X has a causal effect[1] on variable Y, possibly mediated by some other variables $Z_1, \ldots, Z_d$ in the causal network. In general, the only way to completely identify the causal model is by performing experiments (interventions). However, it is often possible to at least narrow down the space of candidate models by using only observational data.[2,3] There are many situations where one is dependent on purely observational data – either because performing experiments is infeasible (e.g. astronomical data), unethical (e.g. clinical cancer studies), or both (e.g. economical data). Some real-life examples include identifying gene expression networks [4,5] and analysing fMRI data from the human brain.[6]

When modelling causal networks between some given variables, structural equation models (SEMs) are used frequently, where each variable is expressed as a function of some other variables (its causes) as well as some noise. Thus, the model is determined by the cause–effect structure (in the form of a directed graph over the variables), the functional dependencies, and

*Corresponding author. Email: nowzohour@stat.math.ethz.ch

© 2015 Taylor & Francis

the joint distribution of the noise terms. Assumptions typically made include that the underlying causal model is acyclic (i.e. there are no feedback loops) and that the noise terms are independent (i.e. there are no unobserved variables). We furthermore assume that the noise is additive, that is, the effect variable minus some noise term is a deterministic function of the cause variables. Although quite restrictive, this is a common assumption in many other settings (e.g. regression) and allows straightforward estimation. The standard case then is to parameterize the model by making the functional dependencies linear and the noise Gaussian.[2] In this case the space of candidate models (in the form of directed acyclic graphs) clusters in equivalence classes, which prohibit full identification – every model in a given equivalence class can induce the same joint distribution over the variables. In a sense, this is quite exceptional, however. It has been shown that as soon as one departs from the linearity or the Gaussianity assumptions the model becomes fully identifiable.[3] [8–12] We are thus interested in the nonparametric case, where either the functional dependencies are nonlinear or the noise terms are non-Gaussian (or both). An inference procedure for this case based on nonparametric independence tests has been suggested by Mooij et al.[13] Their method is using the fact that when fitting the wrong model the noise terms will not be independent. There are a few problems with this approach, however. First, the null hypothesis of the tests employed is independence, which is what one would like to show, and statistical hypothesis testing only allows to reject such hypotheses. Second, because of the many tests involved there is a multiple testing problem. Third, nonparametric independence testing among many variables is statistically hard, and the tests tend to be computationally intensive.

We take a different approach in the form of a score-based method, which is consistent, fast, and easily adaptable to greedy methods for large problems. Score-based methods are widely used for fitting Gaussian SEMs [14] or discrete Bayesian networks.[15] Maximum a posteriori estimation was used in the setting of nonlinear models with Gaussian noise by Imoto et al.[16] Two other score-based methods have recently been proposed: for the parametric setting of Gaussian and linear models with same error variances [9] and for linear models with non-Gaussian noise.[17] Most closely related to this paper is an approach from Bühlmann et al.[18] They consider a semi-parametric SEM with additive, nonlinear functions in the parental variables and additive Gaussian noise, and they prove consistency and present an algorithm for cases with potentially many variables. In contrast, we consider here a model with a nonparametric specification of the error distribution (while the focus is on cases with few variables only). Thus, our model is more general but harder to estimate from data. We propose a penalized maximum likelihood method and prove its asymptotic consistency for finding the true underlying graph provided some technical assumptions about the class of probability densities hold. Our nonparametric setting also includes the well-known LiNGAM model [11] as a special case, and thus we provide here a score-based approach for LiNGAM. Independent work by Kpotufe et al. [19] considers a similar problem as ours: however, while they only treat the case with two variables, we allow for more realistic multivariate settings.

This paper is organized as follows: In Section 2 we review the basic notation and definitions we will use later on before describing our method. In Section 3 we present our main theorem and the assumptions for proving consistency in the large sample limit. In Section 4 we discuss simulation results showing that the method works in practice under controlled conditions. In Section 5, we test our method on some real-world data sets and compare it to other causal inference methods.

2. The method

Suppose data are sampled from real-valued random variables $X_1, \ldots, X_d$, which have some causal structure. We are interested in finding this causal structure (in the form of a directed acyclic graph) just by using observational data. Before we describe our method and the assumptions it

Downloaded by [ETH Zurich] at 08:49 06 June 2016

rests on, we will give definitions of some of the basic terms used in this paper (some of which can be found in for example, Lauritzen,[20] Pearl,[1] Triebel.[21]

2.1. *Notation and definitions*

Given a set of vertices $\mathcal{V} = \{1, \ldots, d\}$ and edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$, we define the d-dimensional *graph* G as the ordered pair $(\mathcal{V}, \mathcal{E})$. If $\mathcal{E}$ is asymmetric, G is called a *directed graph*. Given two vertices $\alpha, \beta \in \mathcal{V}$, a *directed path* of length n from α to β is a sequence of vertices $\alpha = v_0, \ldots, v_n = \beta$, s.t. $(v_i, v_{i+1}) \in \mathcal{E}$ for all $i = 0, \ldots, n-1$. If G is directed and for all $v \in \mathcal{V}$ there is no path of length $n \geq 1$ from v to itself, then G is called a *directed acyclic graph (DAG)*. If $\mathcal{V}' \subseteq \mathcal{V}$ and $\mathcal{E}' \subseteq \mathcal{E}|_{\mathcal{V}' \times \mathcal{V}'}$, then $G' = (\mathcal{V}', \mathcal{E}')$ is called a *subgraph* of G, and we write $G' \subseteq G$. If $\mathcal{E}' \subset \mathcal{E}|_{\mathcal{V}' \times \mathcal{V}'}$, we call G' a *proper* subgraph of G and write $G' \subset G$. In a graph G, we define the *parents* of a vertex v as the set $\mathrm{pa}_G(v) := \{u \in \mathcal{V} : (u, v) \in \mathcal{E}\}$. The structural Hamming distance (SHD) between two graphs G, G' is defined as the number of single edge operations (edge additions, deletions, reversals) necessary to transform G into G'.

A joint density p over $X_1, \ldots, X_d$ is *Markov* with respect to a DAG D, if it factorizes along D:

$$p(x_1, \ldots, x_d) = \prod_{k=1}^{d} p(x_k | \{x_l\}_{l \in \mathrm{pa}_D(k)}). \tag{1}$$

A DAG D is *causally minimal* with respect to a joint density p, if $\nexists D' \subset D$ s.t. p is Markov with respect to D'.

An *SEM* $M = \{f_k, p_{\epsilon_k}\}_{k=1,\ldots,d}$ is a set of functions f_k and densities p_{ϵ_k}, specifying each variable X_k as a function of some of the other variables and a noise term ϵ_k (independent of the other noise terms) with density p_{ϵ_k}. The model M induces a DAG D, where a directed edge (k, l) is added if the function for X_l directly depends on X_k. We will assume in this paper, that M is *recursive*, i.e. its graph D is actually a DAG. We can write the model equations as

$$X_k = f_k(\{X_l\}_{l \in \mathrm{pa}_D(k)}), \epsilon_k), \quad k = 1, \ldots d.$$

If the functions are additive in the noise, i.e. if

$$X_k = f_k(\{X_l\}_{l \in \mathrm{pa}_D(k)})) + \epsilon_k, \quad k = 1, \ldots, d, \tag{2}$$

the model is called an *additive noise model (ANM)*. We call $\mathcal{M} := (\mathcal{F}, \mathcal{P}^\epsilon)$ a *functional model class*[4] of dimension d if $\mathcal{F} \subset C^0(\mathbb{R}^{d-1})$ is a class of functions containing the possible edge functions f_k and $\mathcal{P}^\epsilon$ is a class of univariate probability densities containing the possible error densities p_{ϵ_k}.

The joint density of an ANM is of the form (1) and thus Markov to its DAG D. Vice versa we say that D induces a class of joint densities $\mathcal{P}$ on $X_1, \ldots, X_d$ from a functional model class $\mathcal{M}$, where

$$\mathcal{P} = \left\{ \prod_{k=1}^{d} p_k(x_k - f_k(\{x_l\}_{l \in \mathrm{pa}_D(k)})) : f_k \in \mathcal{F}, p_k \in \mathcal{P}^\epsilon \right\}. \tag{3}$$

Thus, $\mathcal{P}$ contains all joint densities that can be generated by ANMs from class $\mathcal{M}$ with DAG D. The class $\mathcal{M}$ is said to be *identifiable*, if the intersection of any two density classes $\mathcal{P}^1, \mathcal{P}^2$ induced by distinct graphs D_1, D_2 only contains densities for which there exists a unique graph that is causally minimal. We assume throughout the paper that the data-generating process is an ANM with associated causally minimal DAG D_0 with induced density class $\mathcal{P}^0$ and true joint

Downloaded by [ETH Zurich] at 08:49 06 June 2016

density $p^0 \in \mathcal{P}^0$. Causal minimality here essentially means that every edge in D creates a dependency in the joint distribution (i.e. there is an edge from X_l to X_k only if f_k is not constant in x_l).

For the density class, we often consider the weighted *Sobolev space* of functions $W_r^s(\mathbb{R}^n, \langle\cdot\rangle^\beta)$ which is defined as follows:

$$W_r^s(\mathbb{R}^n, \langle\cdot\rangle^\beta) := \{f \in L^r(\mathbb{R}^n) : D^\alpha(f \cdot \langle\cdot\rangle^\beta) \in L^r(\mathbb{R}^n)\ \forall |\alpha| \le s\},$$

where $\langle x\rangle^\beta = (1 + \|x\|^2)^{\beta/2}$ is a polynomial weighting function parameterized by $\beta \in \mathbb{R}$, D^α is the partial derivative operator according to the multi-index α, and r, s are integers at least 1. Note that for $\beta = 0$ this is the usual Sobolev space, while for $\beta > 0$ this is more restrictive (as the tails get bigger weights), and for $\beta < 0$ it is less restrictive. We will mostly be interested in the $\beta < 0$ case.

2.2. *Penalized maximum likelihood estimation*

We now describe our method to learn the true causal structure from data. Suppose we measure d variables, and we have n i.i.d. samples $\{x_k^j\}$ with $j = 1, \ldots, n$ and $k = 1, \ldots, d$. Let $D_1, \ldots, D_N$ be the candidate DAGs under consideration[5] and $\mathcal{P}^1, \ldots, \mathcal{P}^N$ their induced density classes for some model class $\mathcal{M}$. If $\mathcal{M}$ is identifiable, we aim to infer the true DAG D_0 by finding the density class $\mathcal{P}^0$ that contains the true joint density p^0 (if there is more than one such class, we choose the one corresponding to the smallest graph). Of course, we do not know p^0 – instead we estimate it by computing 'best representatives' $\hat{p}_n^i$ from each class $\mathcal{P}^i$. These are chosen via nonparametric maximum likelihood:

$$\hat{p}_n^i = \arg\max_{p \in \mathcal{P}^i} \sum_{j=1}^n \log p(x_1^j, \ldots, x_d^j).$$

Then, each model is scored with a penalized log-likelihood:

$$S_n^i = \frac{1}{n} \sum_{j=1}^n \log \hat{p}_n^i(x_1^j, \ldots, x_d^j) - \#(\text{edges})_i \cdot a_n, \tag{4}$$

where a_n controls the strength of the penalty. Taking the maximum over these scores, we get the estimator

$$\hat{D}_n = D_{\hat{I}_n}, \quad \text{where } \hat{I}_n = \arg\max_{i=1,\ldots,N} S_n^i.$$

Hence, the estimated DAG is $D_{\hat{I}}$. We will show in Section 3 that this procedure is consistent for a_n proportional to $1/\log n$ and that therefore $\hat{D}_n = D_0$ in the large sample limit.

The question arises how to find the maximum likelihood estimators $\hat{p}_n^i$ in each class in this nonparametric setting. We present here an exemplary procedure that has proved useful in practice. To estimate the edge functions of the SEM, we employ a nonparametric regression method. The error densities are then inferred from the residuals using a density estimation method. The estimated joint density is finally given by the product of the residual densities, in accordance with Equation (3).

This gives the following three-step procedure for each DAG D_i:

(1) For each node k estimate the residuals $\hat{\epsilon}_k$ by nonparametrically regressing X_k on $\{X_l\}_{l \in \text{pa}_{D_i(k)}}$. If $\text{pa}_{D_i(k)} = \varnothing$, set $\hat{\epsilon}_k = x_k$.
(2) For each node k estimate the residual densities $\hat{p}_{\epsilon_k}$ from the estimated residuals $\hat{\epsilon}_k$.

Downloaded by [ETH Zurich] at 08:49 06 June 2016

(3) Compute the penalized likelihood score

$$S_n^i = \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{d}\log \hat{p}_{\epsilon_k}(\hat{\epsilon}_k^j) - \#(\text{edges})_i \cdot a_n.$$

Of course, an exhaustive search over all DAGs is only feasible for small values of d, since the number of DAGs grows super-exponentially with the number of vertices[6] and nonparametric regression in d dimensions is ill-posed in general without making structural constraints, due to the curse of dimensionality.[7] The methods used in steps 1 and 2 should be chosen depending on the model class $\mathcal{M}$. Examples are (generalized) additive model (GAM) regression for step 1 and kernel density estimation for step 2.

As an illustration we look at the two-dimensional case, where there are only two variables X_1 and X_2. There are three DAGs inducing the following models:

$$\begin{aligned}
D_1 :& X_1 \longrightarrow X_2\\
& X_1 = \epsilon_1\\
& X_2 = f(X_1) + \epsilon_2\\
& p_1(x_1, x_2) = p_{X_1}(x) \cdot p_{X_2|X_1}(x_2|x_1) = p_{\epsilon_1}(x_1) \cdot p_{\epsilon_2}(x_2 - f(x_1))\\
D_2 :& X_1 \longleftarrow X_2\\
& X_1 = g(X_2) + \epsilon_1\\
& X_2 = \epsilon_2\\
& p_2(x_1, x_2) = p_{X_1|X_2}(x_1|x_2) \cdot p_{X_2}(x_2) = p_{\epsilon_1}(x_1 - g(x_2)) \cdot p_{\epsilon_2}(x_2)\\
D_3 :& X_1 \perp\!\!\!\perp X_2\\
& X_1 = \epsilon_1\\
& X_2 = \epsilon_2\\
& p_3(x_1, x_2) = p_{X_1}(x_1) \cdot p_{X_2}(x_2) = p_{\epsilon_1}(x_1) \cdot p_{\epsilon_2}(x_2).
\end{aligned}$$

We do steps 1, 2, and 3 as described above and choose the model with the highest (log-)likelihood penalized likelihood score.

Comparing this score-based approach with independence-test-based methods, the main difference occurs at step 2, where we estimate the residual densities instead of testing their independence. In terms of complexity, we swap one d-dimensional independence test against d univariate density estimations. Simulations show that this is faster by a factor on the order of 100 with current implementations. However, even though we do not test residual independence directly, it is still the discriminatory property by which to identify the true model. By constructing the densities according to Equation (3), we enforce the error terms to be independent in the estimated joint density. If they are not actually, the considered model will obtain a poor score. Thus, we are searching for the best fitting densities where the errors are independent.

3. Theoretical results

We now show that our method is consistent, that is, that it will identify the true underlying DAG given enough samples. In the following $\mathcal{P}_D$ denotes the induced density class of DAG D. We make the following assumptions:

Downloaded by [ETH Zurich] at 08:49 06 June 2016

(AS1) Identifiability: The data $\{x_k^j\}_{\substack{k=1,\dots,d \\ j=1,\dots,n}}$ are i.i.d. realizations (over $j = 1, \dots, n$) of an identifiable SEM with induced d-dimensional DAG D_0. In particular, the SEM can be the ANM (2) with nonlinear edge functions f_k or non-Gaussian noise variables[8] ϵ_k for all $k = 1, \dots, d$.[10, Lemma 1] There are no hidden variables, that is, the noise terms are jointly independent.

(AS2) Causal minimality: There is no proper subgraph D' of D_0, s.t. p^0 is Markov with respect to D'.

(AS3) Smoothness of log-densities: For all DAGs D the log-densities of $\mathcal{P}_D$ (restricted to their respective support) are elements of a bounded weighted Sobolev space. That is, $\exists r \geq 1$, $s > d$, $\beta < 0$, $C > 0$ s.t.

$$\sum_{|\alpha| \leq s} \|D^\alpha (\langle \cdot \rangle^\beta \cdot \mathbf{1}\{p > 0\} \cdot \log p)\|_r < C \quad \forall p \in \mathcal{P}_D,$$

where $\| \cdot \|_r$ is the usual L^r-norm.

(AS4) Moment condition for densities: For all DAGs D, we have

$$\exists \gamma > s - d/r \quad \text{s.t.} \quad \|p \cdot \langle \cdot \rangle^{\gamma - \beta}\|_r < \infty \quad \forall p \in \mathcal{P}_D,$$

where r, s, d, and β are determined by (AS3).

(AS5) Uniformly bounded variance of log-densities: For all DAGs D, we have

$$\forall p^0 \in \mathcal{P}_D \text{ there exists } K > 0 \quad \text{s.t.} \quad \sup_{p \in \mathcal{P}_D} \operatorname{var}_{p^0}(\log p(X_1, \dots, X_d)) < K.$$

(AS6) Closedness of density classes: For all DAGs D, the induced density class $\mathcal{P}_D$ is a closed set, with the topology given by the Kullback–Leibler (KL) divergence $D_{\mathrm{KL}}(p(\mathbf{x}) \| q(\mathbf{x})) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} \, \mathrm{d}\mathbf{x}$.

The first two assumptions concern the general model set-up and ensure identifiability (i.e. non-overlapping induced density classes). (AS1) requires the data to come from an identifiable ANM due to nonlinearity or non-Gaussianity, as in [8]. (AS2) ensures there are no 'superfluous' edges in the true DAG, that is, the true model is the most parsimonious fitting the data.

The last four assumptions are technical and used to prove consistency of the penalized maximum likelihood estimator. (AS3) essentially requires the log-densities to be smooth. (AS4) requires the densities to have some (at least fractional) finite moments. (AS5) requires the log-densities, for every underlying density p^0, to have uniformly bounded second moments. Finally, (AS6) guarantees the existence of the maximizers of the likelihood and the negative information entropy in each class. Furthermore, it is needed to ensure the true density p^0 has positive KL distance from all wrong density classes. Note that the latter statement alone would suffice to show consistency, since all statements can be written in terms of the supremums of likelihood and negative entropy, instead of their actual maximizers. However, for better comprehensibility we chose the present formulation with the slightly stronger assumption.

Making these assumptions, the penalized maximum likelihood estimator is consistent. We show this by proving that the probability of the true model obtaining a smaller score than any other model vanishes in the large sample limit.

Downloaded by [ETH Zurich] at 08:49 06 June 2016

THEOREM 1 *Assume* (AS1)–(AS6). *Let* S_n^i *be the penalized likelihood score of DAG* D_i, *given by*

$$S_n^i = \frac{1}{n}\sum_{j=1}^{n} \log \hat{p}_n^i(x_1^j, \ldots, x_d^j) - \#(\text{edges})_i \cdot a_n,$$

where $\#(\text{edges})_i$ *is the number of edges in DAG* D_i, *and* $a_n = 1/\log n$. *Denote by* i_0 *the index of the true DAG* $D_0 = D_{i_0}$. *Then, we have*

$$P(S_n^{i_0} \leq S_n^i) \to 0 \quad \text{as } n \to \infty \quad \forall i \neq i_0.$$

The proof relies on entropy methods and is presented in the appendix. In practice the $1/\log n$ penalty rate might be too large. We used $a_n = 1/\sqrt{n}$ for some simulations in Section 4 (where the noise is Gaussian), which lead to reasonably good performance for finite sample size $n = 300$. Moreover, under stronger assumptions, we have the following result.

Remark 1 When replacing (AS5) with the stronger assumption of sub-Exponential tails of $\log p(X_1, \ldots, X_d)$, we can improve the penalty rate a_n in Theorem 1 from $1/\log n$ to $cn^{-1/(2+d/s)}$, for some $c > 0$ sufficiently large.

4. Numerical results

In this section we present simulation results to show that our method works under controlled conditions. In each case, the data-generating process is an ANM with acyclic graph structure. We first reproduce some results from an earlier paper by Hoyer et al.,[8] where the model involves just two variables and is parameterized by two parameters, controlling linearity and Gaussianity, respectively. Then, we extend this setup to a slightly more general class of models. Finally, we look at cases with more than two variables.

In our implementation, we use GAM regression [22] or local polynomial regression (LOESS, see Cleveland [23]) for step 1 and logspline density estimation (see Kooperberg et al. [24]) or kernel density estimation for step 2. For models with more than two variables, penalization becomes important. We used a factor of $a_n = 1/\sqrt{n}$ instead of the very severe $1/\log n$. This can be justified since in the relevant simulations the noise is Gaussian and the log-densities can be assumed to be sub-Exponential. In this case, the faster rate can be used (see Remark 1). All computations were carried out in the statistical computing language `R` (using packages `mgcv` and `logspline`) and the code is available on request from the authors.

4.1. *Identifiability depending on Linearity and Gaussianity*

Hoyer et al. [8] illustrate their method with a two-dimensional ANM of the form

$$X_1 = \epsilon_1,$$
$$X_2 = X_1 + bX_1^3 + \epsilon_2,$$

with the parameter b ranging from -1 to 1, thus controlling the linearity of the model. The noise terms ϵ_1, ϵ_2 are transformed Normal random variables:

$$\epsilon_k = \text{sgn}(\nu_k) \cdot |\nu_k|^q, \quad \nu_k \overset{\text{iid}}{\sim} \mathcal{N}(0, 1),$$

where the parameter q ranges from 0.5 to 2 and thus controls Gaussianity. The true direction $M_1 : X_1 \to X_2$ cannot be identified with traditional methods (e.g. the PC algorithm), since the

Downloaded by [ETH Zurich] at 08:49 06 June 2016

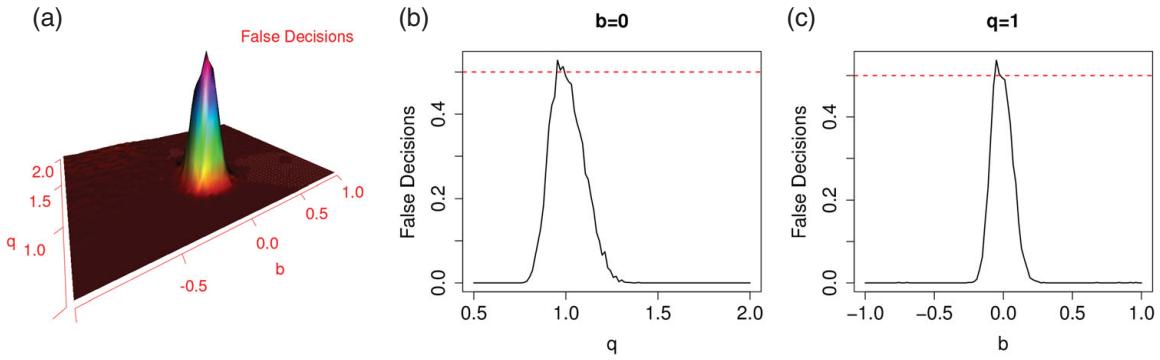

Figure 1. False decision rates for a two-dimensional ANM with two parameters b and q, controlling linearity and Gaussianity ($n = 300$). For $b = 0$ the model is linear, for $q = 1$ the noise is Gaussian: (a) Full $b \times q$ grid. (b) b fixed. (c) q fixed.

Downloaded by [ETH Zurich] at 08:49 06 June 2016

backwards model $M_2 : X_1 \leftarrow X_2$ entails precisely the same conditional independence relations (none) and thus belongs to the same Markov equivalence class. If $b = 0$ and $q = 1$ there exists a backwards model entailing the same joint density. As soon as we move away from this point, however, the model becomes identifiable.[8] We confirm this numerically, showing our method performs as expected in this setting.

We discretize the parameter space $(b, q) \in [-1, 1] \times [0.5, 2]$, and for each grid point we repeat the simulation 1000 times, with $n = 300$ samples per trial. We then count the number of times the backwards model gets wrongly chosen by the method,[9] and this false decision rate serves as our measure of quality of the method. As can be seen in Figure 1, the false decision rate peaks around $(b, q) = (0, 1)$ with around 50% wrong decisions, corresponding to random guessing. Away from this region it quickly drops to zero. In this setting the regressions were done using LOESS and the density estimations using logsplines.

4.2. *Random edge functions*

We now generalize the set-up of the scenario from Section 4.1 in allowing a bigger function class for the edge function. Specifically, we randomly generate functions by sampling a random path from a Wiener process and smoothing it with cubic splines.[10] To measure their nonlinearity, we use the normalized L^2-difference between the function and its best linear approximation on the interval $[-1, 1]$, as described in [25]. A number of randomly generated functions with different nonlinearity values are shown in Figure 2. We again choose a uniform grid of nonlinearity values (in the interval $[0, 0.4]$) and, for each grid point, generate 100 random functions. With each function we perform 100 simulations and average the results. The noise is standard Gaussian in this setting. In Figure 2, we see the results for a small sample ($n = 300$) and a large sample ($n = 1500$) case. The findings are analogous to the simple cubic model – the false decision rate decreases with nonlinearity of the edge function and sample size. Again, the regressions were done using LOESS and the density estimations using logsplines.

4.3. *Larger networks and thresholding*

In a practical situation, the reliability of any method invariably depends on whether its assumptions are met, as well as some other factors. In our case this would include the nonlinearity of the edge functions, the non-Gaussianity of the noise, the sample size, and the number of nodes. It would be desirable to have some criterion indicating there is insufficient information to make a decision. While this is hard to make concrete, a good first heuristic seems to be the separation of the best-scoring model from the rest. We concretely look at the ratio of the smallest (Δ_1) and the

largest (Δ_2) score difference (see Figure 3(b)). If this is smaller than some threshold t, we make no decision (no selection of a model).

The effect of this can be seen in Figure 3(a). Starting from a full DAG with three nodes as the ground truth, we randomly generate 100 different sets of nonlinear[11] edge functions, and for each set of edge functions we generate 100 data sets with standard Gaussian noise of sample size $n = 300$. With each data set, we run an exhaustive search over all 25 candidate models and, if making a decision after thresholding, compute the SHD between the best-scoring DAG and the ground truth. Comparing the thresholds $t = 0$ and $t = 0.01$, the false decision rate falls from 3.9% to 2.4% while in 3.1% of the cases no decision is made.

We also look at two simulation settings suggested in [10], where the graph consists of four nodes and the edge functions are nonlinear but parametrized by four and five parameters, respectively. In both cases, `nonlinear1` and `nonlinear2`, 100 sets of parameters are drawn from a uniform distribution and then data (with a sample size of $n = 400$) is generated. Our method identifies the correct DAG in 96/97 out of the 100 cases for `nonlinear1/2` (in the other cases, there is one additional edge). This certainly improves upon the results reported in [10] (86 correct decision in both cases).

Downloaded by [ETH Zurich] at 08:49 06 June 2016

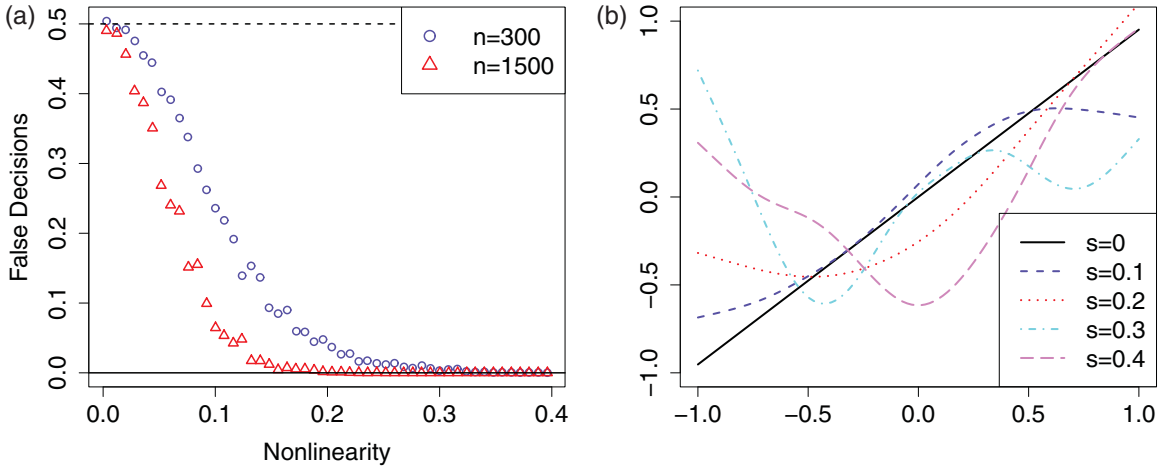

Figure 2. (a) False decision rates with randomly sampled edge functions and Gaussian noise decreases with nonlinearity of the functions. (b) Examples of randomly generated functions, where parameter s controls nonlinearity.

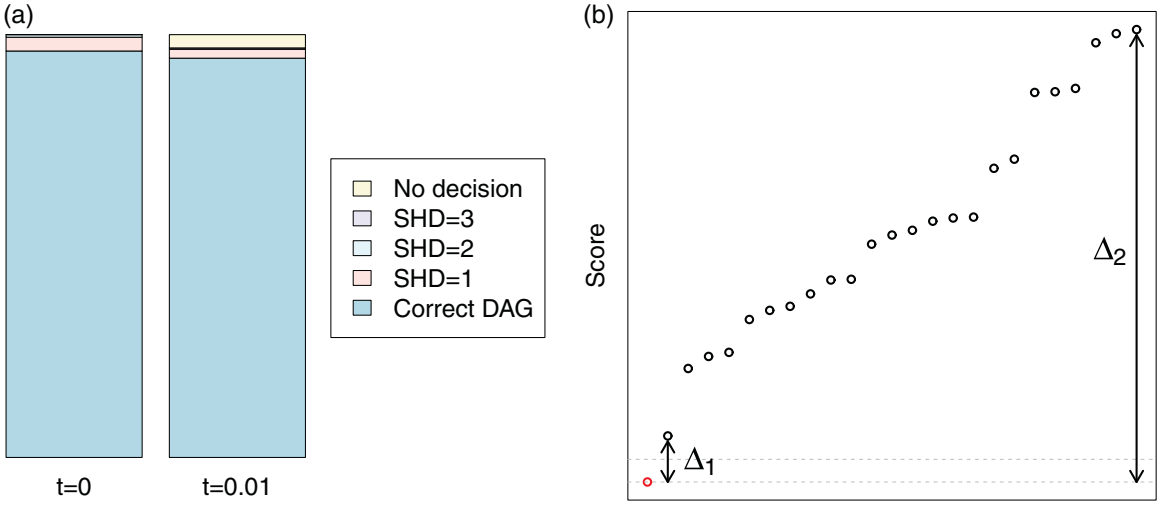

Figure 3. (a) SHD between the best-scoring DAG and the ground truth for a 3-node simulation with ($t = 0.01$) and without ($t = 0$) thresholding. (b) Illustration of thresholding for a single simulation run. Let $s_1, \ldots, s_D$ be the (increasingly) ordered scores. Then $\Delta_1 = s_1/s_2$ and $\Delta_2 = s_1/s_N$.

Table 1. Success rates of different causal inference methods on cause–effect pairs at a decision rate of 100%.

Method	SCL	AN	Lingam	PNL	IGCI	GPI
Accuracy	66%	63%	58%	68%	75%	70%

Notes: SCL, Score-based Causal Learning (our method); AN, Additive Noise with independence testing; PNL. Post-Nonlinear; IGCI, Information-Geometric Causal Inference; GPI, Gaussian Process Inference.
All values except SCL taken from [28]. All data sets were subsampled three times (if $n > 500$), and the results were averaged.

In all of these multivariate settings, we used GAM for regression and logsplines for density estimation.

5. Real data

To determine the performance on real-world data sets, we apply our method to so-called cause–effect pairs. These are bivariate data sets where the true causal direction is known. An example would be the altitude and the average temperature of weather stations. Mooji and Janzing [26] describe eight such pairs and compare several methods that were submitted as part of the Causality Pot-Luck Challenge. Our method identifies seven out of the eight pairs correctly,[12] thus beating all other compared methods except,[27] who take into account post-nonlinear additive noise.

We next consider the extended collection of cause–effect pairs, which can be found at http://webdav.tuebingen.mpg.de/cause-effect. This currently comprises 86 data sets, 81 of which are bivariate. Using our method on these 81 bivariate data sets, we identify the true model in 66% of the cases.[13] In [28] a subset of these data sets were used to compare various causal inference methods. Running our method on those data sets, it compares well with the other methods (see Table 1), being slightly better than independence testing (AN) and outperforming the Lingam method.

In both of these settings, we used LOESS and kernel density estimation.

6. Conclusions

We presented a new fully nonparametric likelihood score-based method for causal inference in nonlinear or non-Gaussian ANMs. We proved consistency of the penalized maximum likelihood estimator for finding the correct model. We showed via simulation studies that our method works well in practice when the ground truth is an ANM with sufficiently nonlinear edge functions or non-Gaussian error terms. Our method compares favourably to other causal inference procedures on both simulated and real-world data.

As a major open challenge, the current approach of exhaustively searching through the whole model space becomes computationally infeasible for more than a handful of variables. Since our method is score-based and the scoring criterion is local (i.e. decomposable), it is straightforward to implement a greedy algorithm although there will be no guarantee for finding a global optimum.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes

1. X has a causal effect on Y if manipulating X changes the distribution of Y, see Pearl.[1]
2. In fact, this is how SEMs where first introduced and continue to be used today.[7]

Downloaded by [ETH Zurich] at 08:49 06 June 2016

3. Except for a set of degenerate cases of measure zero.
4. Here, we implicitly assume that the model has additive noise.
5. For example, all DAGs with d nodes.
6. The first few values of the number of DAGs $N(d)$ with d nodes are $N(2) = 3$, $N(3) = 25$, $N(4) = 543$, $N(5) = 29{,}281$, $N(6) = 3{,}781{,}503$, for example.
7. The latter problem can be dealt with in certain cases, for example, additive models, where the edge functions are additive in the parental variables.
8. Excluding a set of exceptions of measure zero.[8, Theorem 1].
9. That is, when the likelihood score of the backwards model is lower than that of the forwards model.
10. A Wiener path (random normal increments) is sampled on a 1000 point grid spanning $[-1, 1]$ and the resulting vector rescaled to an interval of length 2 and consequently smoothed using cubic splines. The resulting functions are linear outside $[-1, 1]$ and nonlinear inside.
11. With nonlinearity values in $[0.39, 0.4]$.
12. This corresponds to a p-value of .0352 under the random guessing null hypothesis.
13. This corresponds to a p-value of .005 under the random guessing null hypothesis.

Downloaded by [ETH Zurich] at 08:49 06 June 2016

References

[1] Pearl J. Causality. New York: Cambridge University Press; 2000.
[2] Spirtes P, Glymour C, Scheines R. Causation, prediction, and search. New York: Springer Verlag; 1993.
[3] Verma TS, Pearl J, Equivalence and synthesis of causal models. Paper presented at the 6th Conference on Uncertainty in Artificial Intelligence (UAI). 1991 July 27–29; Cambridge, MA, USA. p. 220–227.
[4] Statnikov A, Henaff M, Lytkin NI, Aliferis CF. New methods for separating causes from effects in genomics data. BMC Genomics. 2012;13 (Suppl. 8):S22.
[5] Stekhoven DJ, Moraes I, Sveinbjornsson G, Hennig L, Maathuis MH, Bühlmann P. Causal stability ranking. Bioinformatics. 2012;28:2819–2823.
[6] Ramsey JD, Hansen SJ, Hanson C, Halchenko YO, Poldrack RA, Glymour C. Six problems for causal inference from fMRI. NeuroImage. 2010;49:1545–1558.
[7] Bollen KA. Structural equations with latent variables. New York: Wiley; 1989.
[8] Hoyer PO, Janzing D, Mooji J, Peters J, Schölkopf B, Nonlinear causal discovery with additive noise models. In: Koller D, Schuurmans D, Bengio Y, Bottou L, editors. Advances in neural information processing systems (NIPS) Vol. 21, 2009. p. 689–696.
[9] Peters J, Bühlmann P. Identifiability of Gaussian structural equation models with equal error variances. Biometrika. 2014;101:219–228.
[10] Peters J, Mooji J, Janzing D, Schölkopf B, Identifiability of causal graphs using functional models. Paper presented at the 27th Conference on Uncertainty in Artificial Intelligence (UAI). 2011 July 14–17; Barcelona, Spain. p. 589–598.
[11] Shimizu S, Hoyer PO, Hyvärinen A, Kerminen A. A linear non-Gaussian acyclic model for causal discovery. J Mach Learn Res. 2006;7:2003–2030.
[12] Zhang K, Hyvärinen A, On the identifiability of the post-nonlinear causal model. Paper presented at the 25th Conference on Uncertainty in Artificial Intelligence (UAI). 2009 June 18–21; Montreal, QC, Canada. p. 647–655.
[13] Mooij J, Janzing D, Peters J, Schölkopf B, Regression by dependence minimization and its application to causal inference. Proceedings of the 26th International Conference on Machine Learning (ICML), 2009 June 14–18; Montreal, Canada. p. 745–752.
[14] Chickering DM. Optimal structure identification with greedy search. J Mach Learn Res. 2002;3:507–554.
[15] Koller D, Friedman N. Probabilistic graphical models. Cambridge: The MIT Press; 2009.
[16] Imoto S, Goto T, Miyano S, Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. In: Pacific Symposium on Biocomputing. 2002 January 3–7; Lihue, Hawaii. p. 175–186.
[17] Hyvärinen A, Smith SM. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. J Mach Learn Res. 2013;14:111–152.
[18] Bühlmann P, Peters J, Ernest J. CAM: Causal additive models, high-dimensional order search and penalized regression. Ann Stat. 2014;42:2526–2556.
[19] Kpotufe S, Sgouritsa E, Janzig D, Schölkopf B, Consistency of causal inference under the additive noise model. Proceedings of The 31st International Conference on Machine Learning (ICML). 2014 June 21–26; Beijing, China. p. 478–486.
[20] Lauritzen SL. Graphical models. Oxford: Oxford University Press; 1996.
[21] Triebel H. Theory of function spaces. New York: Springer; 1983.
[22] Hastie T, Tibshirani R. Generalized additive models. Stat Sci. 1986;1:297–318.
[23] Cleveland WS. Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc. 1979;74: 829–836.

[24] Kooperberg C, Stone CJ. A study of logspline density estimation. Comput Stat Data Anal. 1991;12: 327–347.
[25] Emancipator K, Kroll MH. A quantitative measure of nonlinearity. Clin. Chem. 1993;39:766–772.
[26] Mooji J, Janzing D, Distinguishing between cause and effect. In: Journal of Machine Learning Workshop and Conference Proceedings, 2010;6:147–156.
[27] Zhang K, Hyvärinen A, Distinguishing causes from effects using nonlinear acyclic causal models. In: Journal of Machine Learning Workshop and Conference Proceedings 2010;6:157–164.
[28] Janzing D, Mooji J, Zhang K, Lemeire J, Zscheischler J, Daniusis P, Steudel B, Schölkopf B. Information-geometric approach to inferring causal directions. Artificial Intelligence. 2012;182–183:1–31.
[29] van de Geer S. Empirical processes in M-Estimation. Cambridge (UK): Cambridge University Press; 2000.
[30] van der Vaart AW, Wellner JA. Weak convergence and empirical processes: with applications to statistics. New York: Springer; 1996.
[31] Bühlmann P, van de Geer S. Statistics for high-dimensional data. Berlin: Springer; 2011.
[32] Nickl R, Pötscher BM. Bracketing metric entropy rates and empirical central limit theorems for function classes of besov- and sobolev-type. J Theor Probab. 2007;20:177–199.

Downloaded by [ETH Zurich] at 08:49 06 June 2016

Appendix: Consistency Proof

The proof heavily relies on entropy methods and empirical process theory. For a good overview of the necessary material, we refer to [29,30]. For an overview of Sobolov and related function spaces, we refer to [21].

Throughout this section, we will adopt the following notation for taking expectations of some random variable f with respect to a distribution Q (following van de Geer [29]):

$$Qf := \int f \, \mathrm{d}Q.$$

In particular, this means we will write expectations and means as

$$Pf = \mathbb{E}[f(X)],$$

$$P_n f = \frac{1}{n} \sum_{j=1}^{n} f(X^j),$$

where P is the true distribution with density $p^0, f : \mathbb{R}^d \to \mathbb{R}$ is some function, X is a vector of random variables (one corresponding to each node) with distribution P, $\{X^j\}_{j=1,\dots,n}$ are independent copies of X, and P_n is the empirical distribution (placing weight $1/n$ on each X^j).

With this notation, we can write the maximum likelihood estimator $\hat{p}_n^i$ and the entropy minimizer p^i in class $\mathcal{P}_i$ (which exist by assumption (AS6) but need not be unique) as

$$\hat{p}_n^i = \arg\max_{p \in \mathcal{P}^i} P_n \log p, \tag{A1}$$

$$p^i = \arg\max_{p \in \mathcal{P}^i} P \log p. \tag{A2}$$

Note that the true density p^0 minimizes the information entropy over the complete density space $\bigcup_{i=1}^N \mathcal{P}^i$ since the Kullback–Leiber divergence $P \log(p^0/p)$ is positive for all densities $p \neq p^0$.

One of the building blocks of the proof of Theorem 1 is a uniform law of large numbers (ULLN) for the classes of log-densities:

$$\sup_{p \in \mathcal{P}^i} |(P_n - P) \log p| \xrightarrow{P} 0 \quad \text{as } n \to \infty \quad \forall i.$$

To show this, an entropy argument is used. We first define the bracketing entropy of a function space. Let $\mathcal{G}$ be a set of functions from $\mathbb{R}^d$ to $\mathbb{R}$. Two functions $g^L, g^U : \mathbb{R}^d \to \mathbb{R}$ (not necessarily in $\mathcal{G}$) form an ϵ-bracket for some $g \in \mathcal{G}$, if $g^L \leq g \leq g^U$ and $\|g^L - g^U\|_{1,\mu} < \epsilon$, where $\|\cdot\|_{1,\mu}$ is the weighted L^1-norm, that is, $\|f\|_{1,\mu} = \int |f(x)\mu(x)| \, \mathrm{d}x$. Suppose $\{g_i^L, g_i^U\}_{i=1,\dots,N_{[]}}$ is the smallest set s.t. $\forall g \in \mathcal{G}$ $\exists i$ s.t. g_i^L, g_i^U form an ϵ-bracket for g, where $N_{[]}$ denotes the number of such pairs. Then, $H_{[]}(\epsilon, \mathcal{G}, \|\cdot\|_{1,\mu}) := \log N_{[]}$ is called the bracketing entropy of $\mathcal{G}$.

The following result connects bracketing entropy $H_{[]}(\epsilon, \mathcal{G}, \|\cdot\|_{1,p^0})$ with respect to the L^1-norm weighted with the true density p^0 and the uniform convergence of the empirical process $(P_n - P)g$. Note that here and throughout this section we use the notation ' $a(\epsilon) \lesssim b(\epsilon)$' as shorthand for ' $a(\epsilon) \leq cb(\epsilon)$ $\forall \epsilon > 0$ for some constant c not depending on ϵ'.

Downloaded by [ETH Zurich] at 08:49 06 June 2016

LEMMA A.1 *Suppose that:*

(i) $\exists 0 \leq \alpha < 1$ *s.t.* $H_{[]}(\epsilon, \mathcal{G}, \|\cdot\|_{1,p^0}) \lesssim \epsilon^{-\alpha} \quad \forall \epsilon > 0$ *and*
(ii) $\exists K$ *s.t.* $var(g(X_1, \ldots, X_d)) < K \quad \forall g \in \mathcal{G}$

Then, $\mathcal{G}$ *satisfies the ULLN:*

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}} |(P_n - P)g| > \delta_n\right) \to 0 \quad \text{as} \quad n \to \infty,$$

where $\delta_n = c/\log n$ *for some* $c > 0$.

Proof We first show that it suffices to look at the supremum over the bracketing functions. Let $g \in \mathcal{G}$ and g_i^L, g_i^U be its δ_n-brackets. We then have

$$\begin{aligned}(P_n - P)g &< (P_n - P)g_i^U + \delta_n,\\ \text{and} \quad &> (P_n - P)g_i^L - \delta_n.\end{aligned}$$

So we have

$$|(P_n - P)g| < \max_{i=1,\ldots,N_{[]}} (|(P_n - P)g_i^L|, |(P_n - P)g_i^U|) + \delta_n$$

and hence

$$\sup_{g\in\mathcal{G}} |(P_n - P)g| < \max_{g\in\{g_i^L, g_i^U\}_i} |(P_n - P)g| + \delta_n.$$

Now

$$\begin{aligned}\mathbb{P}\left(\sup_{g\in\mathcal{G}} |(P_n - P)g| > 2\delta_n\right) &\leq \mathbb{P}\left(\max_{g\in\{g_i^L, g_i^U\}_i} |(P_n - P)g| > \delta_n\right)\\ &\leq 2N_{[]}(\delta_n) \max_{g\in\{g_i^L, g_i^U\}_i} \mathbb{P}(|(P_n - P)g| > \delta_n)\\ &\lesssim \exp(\delta_n^{-\alpha}) \frac{K^2}{n\delta_n^2}, \qquad \text{(A3)}\end{aligned}$$

where the last line follows from Chebyshev's inequality. Substituting for δ_n gives

$$\mathbb{P}(\ldots) \lesssim \log^2 n \cdot \exp(c^{-\alpha} \log^\alpha n - \log n) \longrightarrow 0 \quad \text{as} \quad n \to \infty.$$

■

Note that if we replace condition (ii) with the assumption that $g(X_1, \ldots, X_d)$ are sub-Exponential (as in Remark 1), we apply the sub-Exponential tail bound (see, e.g. [31, Lemma 14.9]) instead of Chebyshev's inequality and obtain $\exp(\delta_n^{-\alpha} - n\delta_n^2/const.)$ instead of Equation (A3), which converges to zero for $\delta_n = cn^{-1/(2+\alpha)}$, for $c > 0$ sufficiently large.

Lemma A.1 shows that a sufficient condition for the ULLN is finite bracketing entropy. To this end, we make use of the following result:

LEMMA A.2 (Nickl and Pötscher [32, Theorem 1]) *Suppose* $\mathcal{G}$ *is a (non-empty) bounded subset of the weighted Sobolev space* $W_p^s(\mathbb{R}^d, \langle x\rangle^\beta)$ *for some* $\beta < 0$. *Suppose* $\exists \gamma > s - d/p > 0$ *s.t. the moment condition*

$$\|\langle\cdot\rangle^{\gamma-\beta}\|_{1,\mu} = \|\mu(x)\langle x\rangle^{\gamma-\beta}\|_1 < \infty$$

holds for some Borel measure μ *on* $\mathbb{R}^d$. *Then*

$$H_{[]}(\epsilon, \mathcal{G}, \|\cdot\|_{1,\mu}) \lesssim \epsilon^{-d/s}.$$

The relevant sets of functions $\mathcal{G}$ in this context are the log-densities of each class, that is, $\{\mathbf{1}\{p > 0\} \log p | p \in \mathcal{P}^i\}$, with the relevant Borel measure μ being the true density p^0.

Essentially, the idea of the proof of Theorem 1 is to show that the maximum log-likelihood in each induced density class converges to the minimal entropy. For non-overlapping models (e.g. $X_1 \to X_2$ and $X_1 \leftarrow X_2$), the minimal entropy will be different in each class (with the minimum occurring in the true model class), and the likelihood will eventually

pick up on this difference. Since the penalty term vanishes asymptotically, an ever so small difference in entropy will differentiate the true model class from the others. For overlapping (e.g. hierarchical) models, the minimal entropy can occur in more than one class. In this case, the penalty term picks out the most parsimonious model (which is the true model according to the Causal Minimality assumption). Note that the penalty $1/\log n$ is quite large compared with for example, the BIC penalty ($\log n/n$). This is due to the slow convergence of maximum likelihood to minimal entropy (Lemmas A.3 and A.1). If the penalty vanishes too quickly, it will be drowned out by the noise in the likelihood and have no effect. The convergence can be improved (and thus the penalty relaxed) when making stronger assumptions on the distributions, for example, sub-Gaussian tails.

The following lemma shows convergence of maximum log-likelihood to minimal entropy in each class, given that a ULLN holds.

LEMMA A.3 *Suppose that a ULLN for the classes* $\log \mathcal{P}^i$ *holds with convergence rate* δ_n*, that is,*

$$P\left(\sup_{p\in\mathcal{P}^i} |(P_n - P)(\mathbf{1}\{p > 0\}\log p)| > \delta_n\right) \to 0 \quad \text{as} \quad n \to \infty.$$

Then

$$P(|P_n \log \hat{p}_n^i - P \log p^i| > \delta_n) \to 0 \quad \text{as} \quad n \to \infty.$$

Proof By the definition of the MLE Equation (A1), we have

$$P_n \log \hat{p}_n^i \geq P_n \log p^i = P \log p^i + (P_n - P)\log p^i,$$

i.e.

$$P_n \log \hat{p}_n^i - P \log p^i \geq (P_n - P)\log p^i. \tag{A4}$$

Let $\tilde{\mathcal{P}}_n^i$ be the restriction of $\mathcal{P}^i$ to densities whose support contains the data, i.e. $\tilde{\mathcal{P}}_n^i = \{p \in \mathcal{P}^i | \text{supp}(p) \supseteq \{X^1, \ldots, X^n\}\}$. Note that the maximum log-likelihood as well as minimum entropy are the same over $\mathcal{P}^i$ and $\tilde{\mathcal{P}}_n^i$, since densities with support not including the data will yield values of $-\infty$. So we also have:

$$\begin{aligned} P_n \log \hat{p}_n^i &= \max_{p\in\mathcal{P}^i} P_n \log p = \max_{p\in\tilde{\mathcal{P}}_n^i} P_n \log p \\ &= \max_{p\in\tilde{\mathcal{P}}_n^i} (P \log p + (P_n - P)\log p) \\ &\leq P \log p^i + \sup_{p\in\tilde{\mathcal{P}}_n^i} (P_n - P)\log p, \end{aligned}$$

i.e.

$$P_n \log \hat{p}_n^i - P \log p^i \leq \sup_{p\in\tilde{\mathcal{P}}_n^i} (P_n - P)\log p.$$

This together with Equation (A4) yields:

$$\begin{aligned} |P_n \log \hat{p}_n^i - P \log p^i| &\leq \max\left(|(P_n - P)\log p^i|, \sup_{p\in\tilde{\mathcal{P}}_n^i} (P_n - P)\log p\right) \\ &\leq \max\left(|(P_n - P)\log p^i|, \sup_{p\in\tilde{\mathcal{P}}_n^i} |(P_n - P)|\log p\right) \\ &\leq \sup_{p\in\tilde{\mathcal{P}}^i} |(P_n - P)\log p| \\ &\leq \sup_{p\in\mathcal{P}^i} |(P_n - P)(\mathbf{1}\{p > 0\}\log p)|. \end{aligned}$$

We thus have

$$P(|P_n \log \hat{p}_n^i - P \log p^i| > \delta_n) \leq P\left(\sup_{p\in\mathcal{P}^i} |(P_n - P)(\mathbf{1}\{p > 0\}\log p)| > \delta_n\right),$$

which converges to zero as $n \to \infty$ by assumption. ■

Downloaded by [ETH Zurich] at 08:49 06 June 2016

Finally, before proving Theorem 1, we show the following useful lemma.

LEMMA A.4 *Let $a, b, a', b' \in \mathbb{R}$ and $\epsilon > 0$. If one of the following holds:*

(1) $a - b > \epsilon$ *and* $a' - b' \leq 0$
(2) $a - b < \epsilon$ *and* $a' - b' \geq 2\epsilon$

we have $|a - a'| > \epsilon/2$ *or* $|b - b'| > \epsilon/2$.

Proof Assume (i). Then, we have

$$\epsilon = \epsilon - 0 \leq a - b + b' - a' = |a - a' - (b - b')| \leq |a - a'| + |b - b'|,$$

and the result follows. Similarly for (ii):

$$\epsilon = 2\epsilon - \epsilon \leq a' - b' + b - a = |a' - a - (b' - b)| \leq |a' - a| + |b' - b|.$$

■

We can now prove the main theorem.

Proof of Theorem 1 We will make repeated use of Lemma A.3. For that matter, note that assumptions (AS3)–(AS5), together with Lemmas A.1 and A.2 (taking $\mu = p^0$) satisfy the sufficient conditions. (AS6) ensures the existence of $\hat{p}_n^i, p^i$ as defined in Equations (A1) and (A2).

Let $i \neq i_0$. We differentiate two cases: (i) where $\mathcal{P}^i$ includes the true density p^0 and (ii) where it does not. Let $\delta_n = (\#(\text{edges})_i - \#(\text{edges})_{i_0}) \cdot 1/\log n$ denote the difference of the penalties in the two scores.

Case (i). $p^0 \in \mathcal{P}^i$, which implies $p^i = p^0$. Assumptions (AS1) and (AS2) together with Theorem 2 in [10] guarantee identifiability of the true graph. In particular, this means that in this case $\mathcal{P}^i$ must correspond to a graph containing the true graph. Hence, $\#(\text{edges})_i > \#(\text{edges})_{i_0}$, i.e. $\delta_n > 0$. We then have

$$\begin{aligned}
P(S_n^{i_0} \leq S_n^i) &\leq P\left(P_n \log \hat{p}_n^i - P_n \log \hat{p}_n^{i_0} > \frac{\delta_n}{2}\right) \\
&\leq P\left(|P_n \log \hat{p}_n^{i_0} - P \log p^0| > \frac{\delta_n}{4} \vee |P_n \log \hat{p}_n^i - P \log p^i| > \frac{\delta_n}{4}\right) \\
&\leq P\left(|P_n \log \hat{p}_n^{i_0} - P \log p^0| > \frac{\delta_n}{4}\right) + P\left(|P_n \log \hat{p}_n^i - P \log p^i| > \frac{\delta_n}{4}\right) \to 0
\end{aligned}$$

as $n \to \infty$, where the second line follows from $p^i = p^0$ and Lemma A.4 (first case), and the convergence in the last line follows from Lemma A.3.

Case (ii). $p^0 \notin \mathcal{P}^i$, which implies $P \log p^0 > P \log p^i$. Hence, $\exists \delta > 0$ s.t. $P \log p^0 > P \log p^i + 4\delta$. Let $N > 0$ s.t. $\#(\text{edges})_{i_0} \cdot 1/\log n < \delta \forall n \geq N$. Then, we have

$$\begin{aligned}
P(S_n^{i_0} \leq S_n^i) &= P(P_n \log \hat{p}_n^{i_0} - P_n \log \hat{p}_n^i \leq -\delta_n) \\
&\leq P(P_n \log \hat{p}_n^{i_0} - P_n \log \hat{p}_n^i < \delta) \\
&\leq P(|P_n \log \hat{p}_n^{i_0} - P \log p^0| > \delta \vee |P_n \log \hat{p}_n^i - P \log p^i| > \delta) \\
&\leq P(|P_n \log \hat{p}_n^{i_0} - P \log p^0| > \delta) + P(|P_n \log \hat{p}_n^i - P \log p^i| > \delta) \to 0
\end{aligned}$$

as $n \to \infty$, where the third line follows from Lemma A.4 (second case), and the convergence in the last line follows again from Lemma A.3. ■

Downloaded by [ETH Zurich] at 08:49 06 June 2016