Identifiability of Gaussian structural equation models with equal error variances

Jonas Peters*
Seminar for Statistics
ETH Zurich
Switzerland

Peter Bühlmann*
Seminar for Statistics
ETH Zurich
Switzerland

May 1, 2014

Abstract

We consider structural equation models in which variables can be written as a function of their parents and noise terms, which are assumed to be jointly independent. Corresponding to each structural equation model, there is a directed acyclic graph describing the relationships between the variables. In Gaussian structural equation models with linear functions, the graph can be identified from the joint distribution only up to Markov equivalence classes, assuming faithfulness. In this work, we prove full identifiability if all noise variables have the same variances: the directed acyclic graph can be recovered from the joint Gaussian distribution. Our result has direct implications for causal inference: if the data follow a Gaussian structural equation model with equal error variances and assuming that all variables are observed, the causal structure can be inferred from observational data only. We propose a statistical method and an algorithm that exploit our theoretical findings.

1 Introduction

1.1 Graphical and structural equation models

For random variables $X_1, \ldots, X_p$, we define a graphical model as a pair $\{\mathcal{G}, \mathcal{L}(\mathbf{X})\}$, where $\mathcal{L}(\mathbf{X}) = \mathcal{L}(X_1, \ldots, X_p)$ is a joint probability distribution that is Markov with respect to a directed acyclic graph $\mathcal{G}$ [Lauritzen, 1996, Chapter 3.2]. Structural equation models, also referred to as a functional models, are related to graphical models. They are specified by a collection $\mathcal{S} = \{S_1, \ldots, S_p\}$ of p equations

$$S_j : \quad X_j = f_j(X_{\mathbf{PA}_j}, N_j) \quad (j = 1, \ldots, p) \tag{1}$$

and a joint distribution $\mathcal{L}(\mathbf{N}) = \mathcal{L}(N_1, \ldots, N_p)$ of the noise variables. Here, $\mathbf{PA}_j \subset \{1, \ldots, p\} \setminus \{j\}$ denotes the parents of j. We require the noise terms to be jointly independent, so $\mathcal{L}(\mathbf{N})$ is a product distribution. The graph $\mathcal{G}$ of a structural equation model is obtained by drawing directed edges from each variable X_k, $k \in \mathbf{PA}_j$, occurring on the right-hand side of equation (1) to X_j. The graph $\mathcal{G}$ is required to be acyclic. Furthermore, given a structural equation model, the joint distribution $\mathcal{L}(\mathbf{X})$ is fully determined and $\mathcal{L}(\mathbf{X})$ is Markov with respect to the graph $\mathcal{G}$ [Pearl, 2009, Theorem 1.4.1].

1.2 Identifiability from the distribution

We address the following problem. Given the joint distribution $\mathcal{L}(\mathbf{X}) = \mathcal{L}(X_1, \ldots, X_p)$ from a graphical model or from a structural equation model with directed acyclic graph $\mathcal{G}_0$, can we recover the graph $\mathcal{G}_0$? By first considering graphical models one can easily see that the answer is negative: the joint distribution

*{peters, buhlmann}@stat.math.ethz.ch

$\mathcal{L}(\mathbf{X})$ is Markov with respect to different directed acyclic graphs, e.g., to all fully connected directed acyclic graphs. Thus, there are many possible graphical models $\{\mathcal{G}, \mathcal{L}(\mathbf{X})\}$ for the same distribution $\mathcal{L}(\mathbf{X})$. Similarly, there are structural equation models with different structures that could have generated the distribution $\mathcal{L}(\mathbf{X})$. By making additional assumptions one obtains restricted graphical models and restricted structural equation models for which the graph is identifiable from the joint distribution. It is precisely here that the difference between graphical and functional models becomes apparent.

Given a graphical model, the distribution $\mathcal{L}(\mathbf{X})$ is faithful with respect to the directed acyclic graph $\mathcal{G}_0$ if each conditional independence found in $\mathcal{L}(\mathbf{X})$ is implied by the Markov condition. If faithfulness holds, one can obtain the Markov equivalence graph of the true directed acyclic graph $\mathcal{G}_0$ [Spirtes et al., 2000]. But the Markov equivalence class may still be large [cf. Andersson et al., 1997] and the directed acyclic graph $\mathcal{G}_0$ is not identifiable. Furthermore, faithfulness in its full generality cannot be tested from data [Zhang and Spirtes, 2008]. Since both the Markov condition and faithfulness only restrict the conditional independences in the joint distribution, it is not surprising that two graphs entailing the same conditional independences cannot be distinguished.

Structural equation models enable us to exploit a different type of restriction. First, a general Gaussian structural equation model is equivalent to a Gaussian graphical model $\{\mathcal{G}_0, \mathcal{L}(\mathbf{X})\}$, so the structure $\mathcal{G}_0$ is not identifiable from $\mathcal{L}(\mathbf{X})$. Recently, however, it has been shown that this case is exceptional: (i) if we consider linear functions and non-Gaussian noise, one can identify the underlying directed acyclic graph $\mathcal{G}_0$ [Shimizu et al., 2006]; (ii) if one restricts the functions to be additive in the noise component and excludes the linear Gaussian case, as well as a few other pathological function-noise combinations, one can show that $\mathcal{G}_0$ is identifiable from $\mathcal{L}(\mathbf{X})$ [Hoyer et al., 2009, Peters et al., 2011]. In this work, we prove that there is a third way to deviate from the general linear Gaussian case: (iii) Gaussian structural equation models where all functions are linear, but the normally distributed noise variables have equal variances σ^2, are again identifiable. The identifiability results (i) and (ii) require a condition called causal minimality. In its original form, Zhang and Spirtes [2008] define causal minimality as follows: for the true causal graph $\mathcal{G}_0$, $\mathcal{L}(\mathbf{X})$ is not Markov to any proper subgraph of $\mathcal{G}_0$. Causal minimality is therefore a weak form of faithfulness. Remark 3 shows that for proving (iii) we assume causal minimality.

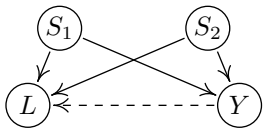
It may come as a surprise that for a class of Gaussian structural equation models the underlying directed acyclic graph is identifiable. The assumption of equal error variances seems natural for applications with variables from a similar domain and is commonly used in time series models.

1.3 Causal interpretation

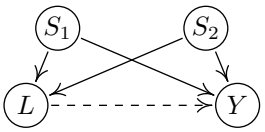
Our result has implications for causal inference. If $\mathcal{G}_0$ is interpreted as the causal graph of the data generating process for $X_1, \ldots, X_p$, the problem considered here is to infer the causal structure from the joint distribution. This is particularly interesting when the causal graph is of interest but interventional experiments are too expensive, unethical or even impossible to perform. In the causal setting, our result reads as follows. If the observational data are generated by a Gaussian structural equation model that represents the causal relationships and has equal error variances, then the causal graph is identifiable from the joint distribution. Despite the potentially important application in causal inference, we present the main statement and its proof without causal terminology; in particular, equations (1) and (2) can be interpreted as holding in distribution.

2 Identifiability for Gaussian models with equal error variances

We first introduce some notation. The index set $\mathbf{J} = \{1, \ldots, p\}$ corresponds to a set of vertices in a graph. Associated with $j \in \mathbf{J}$ are random variables X_j from $\mathbf{X} = (X_1, \ldots, X_p)$. Given a directed acyclic graph $\mathcal{G}$, we denote the parents of a node j by $\mathbf{PA}_j^{\mathcal{G}}$, the children by $\mathbf{CH}_j^{\mathcal{G}}$, the descendants by $\mathbf{DE}_j^{\mathcal{G}}$ and the non-descendants by $\mathbf{ND}_j^{\mathcal{G}}$.



Graph $\mathcal{G}$



Graph $\mathcal{G}'$

Figure 1: The situation dealt with in the second part of case (ii) of the proof of Theorem 1, with $\mathbf{S} = \{S_1, S_2\}$ and $\mathbf{D} = \emptyset$. It contains the proof's main argument.

We consider a structural equation model with directed acyclic graph $\mathcal{G}_0$ of the form

$$X_j = \sum_{k \in \mathbf{PA}_j^{\mathcal{G}_0}} \beta_{jk} X_k + N_j \quad (j = 1, \ldots, p)\,, \tag{2}$$

where all N_j are independent and identically distributed according to $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 > 0$. Additionally, for each $j \in \{1, \ldots, p\}$ we require $\beta_{jk} \neq 0$ for all $k \in \mathbf{PA}_j^{\mathcal{G}_0}$.

Theorem 1 *Let $\mathcal{L}(\mathbf{X})$ be generated from model (2). Then $\mathcal{G}_0$ is identifiable from $\mathcal{L}(\mathbf{X})$ and the coefficients β_{jk} can be reconstructed for all j and $k \in \mathbf{PA}_j^{\mathcal{G}_0}$.*

Problem 2 *The idea of the proof is to assume that there are two structural equation models with distinct graphs $\mathcal{G}$ and $\mathcal{G}'$ that lead to the same joint distribution. We exploit the Markov condition and causal minimality, see Remark 3, in order to find variables L and Y that have the same set of parents $\mathbf{S} = \{S_1, S_2\}$ in both graphs, but reversed edges between each other in $\mathcal{G}$ and $\mathcal{G}'$, as shown in Fig. 1. Defining $L^* = L\,|_{\,\mathbf{S}=s}$ for some value $s \in \mathbb{R}^2$, we can use the equal error variances to show that L^* has different variances in both graphs. This leads to a contradiction.*

Problem 3 *Theorem 1 assumes that the coefficients $\beta_{jk} \neq 0$ do not vanish for any $k \in \mathbf{PA}_j^{\mathcal{G}_0}$. Lemma 8 below and Proposition 2 in Peters et al. [2011] show that this condition implies causal minimality. From our point of view, causal minimality is a natural condition and in accordance with the intuitive understanding of a causal influence between variables.*

Problem 4 *Theorem 1 can be generalized to the case where the error covariance matrix has the form $\mathrm{Cov}(N_1, \ldots, N_p) = \sigma^2 \mathrm{diag}(\alpha_1, \ldots, \alpha_p)$ with pre-specified $\alpha_1, \ldots, \alpha_p$ and unknown σ^2.*

3 Penalized maximum likelihood estimator

Consider data which are independent and identically distributed realizations of $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$ from model (2) with true coefficients β_{jk}^0. The representation in vector form is $\mathbf{X} = B\mathbf{X} + \mathbf{N}$, where B is the $p \times p$ matrix with entries $B_{jk} = \beta_{jk}$. To make the manuscript easier to read we write B or β whenever we think of a matrix or a vector of parameters, respectively. As estimator for the coefficients $B^0 = (\beta_{jk}^0)_{j,k}$ and the error variance σ^2, we consider

$$\{\hat{\beta}(\lambda), \hat{\sigma}^2(\lambda)\} = \underset{\beta \in \mathcal{B}, \sigma^2 \in \mathbb{R}^+}{\operatorname{argmin}} \; -\ell(\beta, \sigma^2; \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}) + \lambda \|\beta\|_0\,, \tag{3}$$

where

$$-\ell(\beta, \sigma^2; \mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}) = \frac{np}{2} \log(2\pi\sigma^2) + \frac{n}{2\sigma^2} \mathrm{tr}\{(I - B)^T (I - B) \hat{\Sigma}\}\,,$$

with sample covariance matrix $\hat{\Sigma}$, is the negative log-likelihood assuming equal error variances σ^2 and $\|\beta\|_0 = |\{j, k : \beta_{jk} \neq 0\}|$. Furthermore, $\mathcal{B} = \{B \in \mathbb{R}^{p \times p} : \mathrm{Adj}(B) \text{ has only zero eigenvalues}\}$ contains only those coefficient matrices whose corresponding graphs do not have cycles [Cvetković et al., 1995, p.81]. Here, $\mathrm{Adj}(B)_{jk} = 1_{\beta_{jk} \neq 0}$ is the adjacency matrix. Minimizing over all $\beta \in \mathcal{B}$ includes optimizing

over all directed acyclic graphs, see Section 4. The induced directed acyclic graph from $\hat{\beta}(\lambda)$ is denoted by $\hat{\mathcal{G}}$. For $\lambda = \log(n)/2$ the objective function in equation (3) is the BIC score.

The convergence rate and consistency of the penalized maximum likelihood estimator for the true coefficients β^0_{jk} and the true structure $\mathcal{G}_0$ follow from an analysis in van de Geer and Bühlmann [2013, Theorem 5.1], under regularity conditions. More precisely, for $\lambda_n = \log(n)/2$ we have

$$\sum_{j,k=1}^{p} \{\hat{\beta}_{jk}(\lambda_n) - \beta^0_{jk}\}^2 = O_P\{\log(n) n^{-1}\} \qquad (n \to \infty)\,,$$
$$\mathrm{pr}(\hat{\mathcal{G}}_n = \mathcal{G}_0) \to 1 \qquad (n \to \infty)\,.$$

The results in van de Geer and Bühlmann [2013, Section 5] also cover the high-dimensional sparse setting where $p = p_n = O\{n/\log(n)\}$.

One could use a combination of the PC-algorithm and minimization of the penalized likelihood in equation (3): the former, which is computationally very efficient, could be used for estimating the Markov equivalence class and the latter for orienting remaining undirected edges. A related approach has been suggested by Tillman et al. [2010]. For consistency in the first step one necessarily requires a version of the strong faithfulness assumption, which can be very restrictive [Uhler et al., 2013]. Penalized maximum likelihood estimation does not need such an assumption [van de Geer and Bühlmann, 2013] but pays a price in terms of computational complexity.

4 Greedy search algorithm

Because the optimization in equation (3) is over the space of all directed acyclic graphs, the estimator is hard to compute. Already for $p = 20$, there are 2.3×10^{72} directed acyclic graphs [OEIS Foundation Inc., 2011], which makes an exhaustive search infeasible. Instead, we propose a greedy procedure that we call greedy directed acyclic graph search with equal error variance. At each iteration t we are given a directed acyclic graph $\mathcal{G}_t$ and move to the neighbouring directed acyclic graph with the largest drop in the BIC score. If all neighbours have a higher BIC score in equation (3) than $\mathcal{G}_t$, the algorithm terminates. Here, we say that two directed acyclic graphs are neighbours if they can be transformed into each other by one edge addition, removal or reversal. Chickering [2002] proposes a similar search strategy but with the search done in the space of Markov equivalence classes rather than over directed acyclic graphs.

In order to shorten the runtime, we randomly search through neighbouring directed acyclic graphs until we find a directed acyclic graph with a better score than $\mathcal{G}_t$ and use this directed acyclic graph for $\mathcal{G}_{t+1}$. We consider at least k neighbours; if there are several directed acyclic graphs among the first k with better scores than $\mathcal{G}_t$, we take the best one. The whole procedure further improves if we increase the probability of changing edges pointing into nodes whose residuals have a high variance. This modification and the score function are the only parts of the algorithm that make use of the equal error variances. Additionally, we restart the method five times starting from a random sparse graph with $k = p, k = 2p, k = 3p, k = 5p$ and $k = 300$. This choice is ad hoc but works well in practice, as it decreases the risk of getting stuck in a local optimum. R code for this method is available as Supplementary Material.

5 Experiments

5.1 Existing methods

We compare our method against the PC-algorithm [Spirtes et al., 2000] and greedy equivalence search [Chickering, 2002]. The latter approximates the BIC-regularized maximum likelihood estimator for non-restricted Gaussian structural equation models. Both methods can only recover the Markov equivalence class, see Section 1.2, and therefore leave some arrows undirected. The Markov equivalence class can be represented by a completed partially directed acyclic graph. In the experiments, we report the structural Hamming distance between the true and estimated partially directed acyclic graphs; this assigns a distance of two for each pair of reversed edges, for example, $\to$ in the true and $\leftarrow$ in the estimated graph; all other edge mistakes count as one.

5.2 Random graphs

For varying n and p we compare the three methods. For a given value p, we randomly choose an ordering of the variables with respect to the uniform distribution and include each of the $p(p-1)/2$ possible edges with a probability of p_{edge}. All noise variances are set to 1 since scaling all noise variables with a common factor yields exactly the same estimates $\hat{\beta}$ and $\hat{\mathcal{G}}$. The coefficients β^0_{jk} are uniformly chosen from $[-1, -0.1] \cup [0.1, 1]$. We consider a sparse setting with $p_{\text{edge}} = 3/(2p-2)$, which results in an expected number of $3p/4$ edges, and a dense setting with $p_{\text{edge}} = 0.3$. Table 5.2 shows the average structural Hamming distance to the true directed acyclic graph and to the true completed partially directed acyclic graph over 100 simulations for the sparse setting. Except for $p = 40$ and $n = 100$, the graphs estimated by the proposed method are closer to the true directed acyclic graph than the resulting graphs from state of the art methods, who can only recover the true Markov equivalence class; greedy directed acyclic graph search also performs better when comparing the distance to the true completed partially directed acyclic graph. Table 5.2 shows the analogous results for the dense setting, in which the improvement with greedy directed acyclic graph search with equal error variances is even larger.

		$n = 100$			$n = 500$			$n = 1000$		
p		GDS_{EEV}	PC	GES	GDS_{EEV}	PC	GES	GDS_{EEV}	PC	GES
5	DAG	1.5	3.9	3.6	0.5	2.9	2.8	0.4	3.0	2.5
	CPDAG	1.5	2.9	2.3	0.5	1.4	1.2	0.3	1.0	0.7
20	DAG	12.2	14.1	18.0	4.5	11.1	10.3	2.7	10.1	8.7
	CPDAG	13.9	10.9	17.0	5.2	7.7	7.6	3.0	6.9	5.6
40	DAG	44.7	29.6	53.0	15.7	22.6	26.1	10.7	20.1	21.9
	CPDAG	50.0	24.4	53.1	18.9	15.9	23.4	13.4	13.3	17.5

Table 1: Structural Hamming distance between estimated and true directed acyclic graph and estimated and true Markov equivalence class, for sparse graphs with p nodes and sample size n. DAG, directed acyclic graph; CPDAG, completed partially directed acyclic graph; GDS_{EEV}, greedy directed acyclic graph search with equal error variances; PC, PC-algorithm; GES, greedy equivalence search.

		$n = 100$			$n = 500$			$n = 1000$		
p		GDS_{EEV}	PC	GES	GDS_{EEV}	PC	GES	GDS_{EEV}	PC	GES
5	DAG	1.2	2.9	3.0	0.6	2.4	2.2	0.3	2.1	2.1
	CPDAG	1.3	2.1	1.9	0.5	1.2	0.7	0.2	0.8	0.5
20	DAG	30.0	56.6	63.9	12.5	55.7	66.3	8.2	57.6	69.1
	CPDAG	31.0	56.1	63.2	13.1	55.5	66.2	8.8	57.5	68.5
40	DAG	216.1	242.8	323.1	185.2	247.2	430.4	172.0	248.9	470.6
	CPDAG	217.1	242.4	323.0	185.7	247.0	430.1	172.2	248.5	470.4

Table 2: Structural Hamming distance between estimated and true directed acyclic graph and estimated and true Markov equivalence class, for dense graphs with p nodes and sample size n. DAG, directed acyclic graph; CPDAG, completed partially directed acyclic graph; GDS_{EEV}, greedy directed acyclic graph search with equal error variances; PC, PC-algorithm; GES, greedy equivalence search.

As a proof of concept, we also simulate data with $n = 500$ from a non-faithful distribution: $X_1 = N_1$, $X_2 = -X_1 + N_2$ and $X_3 = X_1 + X_2 + N_3$. As stated by the theory, the PC-algorithm and greedy equivalent search fail here: in all 100 experiments, they output $X_1 \to X_2 \leftarrow X_3$, which is not the correct Markov equivalence class. Greedy directed acyclic graph search always identified the correct directed acyclic graph.

5.3 Deviation from equal error variances

When the data are generated by a Gaussian structural equation model with different error variances, the method is not guaranteed to find the correct directed acyclic graph or the correct Markov equivalence

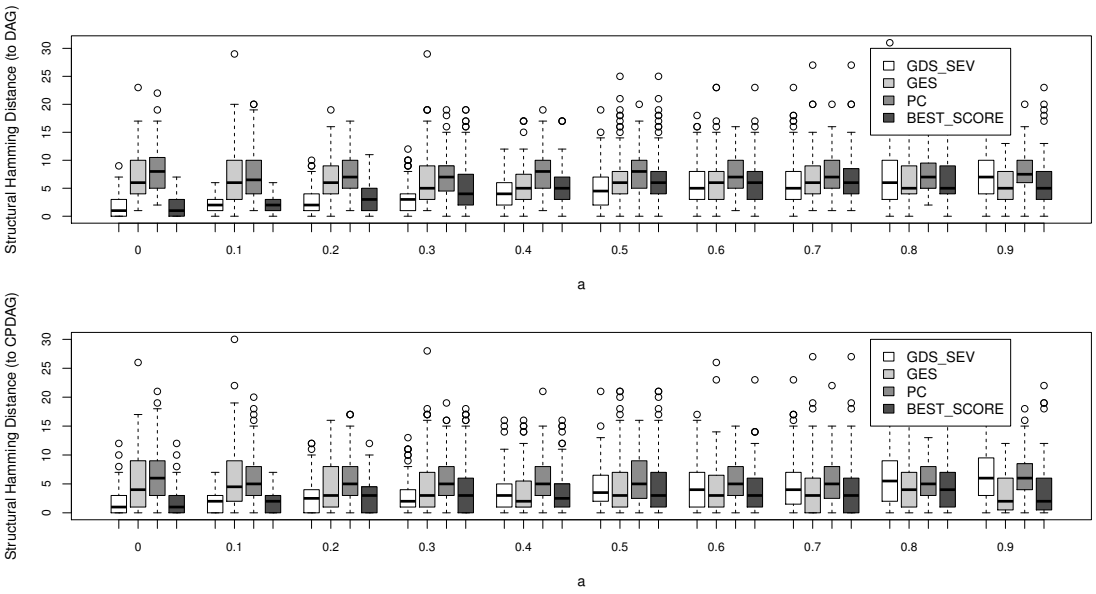


Figure 2: Box plots for the structural Hamming distance of greedy directed acyclic graph search (white), greedy equivalence search (light grey), PC-algorithm (grey) and a best-score method (dark grey) to the true directed acyclic graph, DAG, (top) and to the true partially directed acyclic graph, CPDAG, (bottom). The graph shows various values of a measuring perturbation a of equal error variances; only $a = 0$ corresponds to equal error variances.

class. When the true data generating process follows such a Gaussian structural equation model with different variances, we can always represent it as a model with equal error variances if we apply a fine-tuned rescaling of the variables $X_i \mapsto a_i X_i$ with a_i equal to the inverse of the standard deviation of the error in the ith structural equation. Of course, such a rescaling is only possible when knowing the error variances, hence the word fine-tuned. In the hypothetical case where the data would be scaled with such a deceptive fine-tuned standardization, the graph identified by our method would belong to the correct Markov equivalence class. We emphasize, however, that this is for an artificial scenario which is different from having raw data from a Gaussian structural equation model with different error variances. An important question is how sensitive our method is to deviations from the assumption of equal error variances. We investigate this empirically. For $p = 10$ and $n = 500$, we sample the noise variances uniformly from $[1 - a, 1 + a]$ and vary a between 0 and 0.9. Theorem 1 establishes identifiability of the graph only for $a = 0$. As before, the coefficients β^0_{jk} are uniformly chosen from $[-1, -0.1] \cup [0.1, 1]$. The parameter p_{edge} is chosen to be $2/(p-1)$, on average resulting in p edges; this is in between the sparse and the dense setting. Figure 5.3 shows that the performance of greedy directed acyclic graph search is relatively robust as the parameter a changes. Even for large values of a, the method does not perform worse than the PC-algorithm. The best-score method reports the result of greedy directed acyclic graph search or greedy equivalence search depending on which method obtained the better score. Greedy directed acyclic graph search was chosen in 100%, 100%, 88%, 36%, 7%, 1%, 2%, 0%, 0% and 0% of the cases, for a ranging between 0 and 0.9, respectively.

5.4 Real data

We now apply the greedy equivalence search and greedy directed acyclic graph search to seven data sets containing microarray data, described by Dettling and Bühlmann [2003] and Bühlmann et al. [2013], and compare their BIC scores. When greedy equivalence search obtains the better score, this indicates that the assumption of equal error variances is not justified. In Figure 5.3 we have seen that even then it might

sometimes be useful to look at the greedy directed acyclic graph search solution. If, on the other hand, greedy directed acyclic graph search obtains a better score than greedy equivalence search, we prefer the solution obtained by greedy directed acyclic graph search, which furthermore is a graph rather than a Markov equivalence class. To avoid a high-dimensional setting with $p > n$, we always chose the $0.8n$ genes with the highest variance. Table 5.4 shows that in two out of the seven data sets, greedy directed acyclic graph search obtained a better score than greedy equivalence search. For the Colon example,

	Prostate	Lymphoma	Riboflavin	Leukemia	Brain	Cancer	Colon
GES	4095	4560	2711	5456	1411	5891	3224
GDS_{EEV}	6057	5404	3236	5481	1343	6288	3201

Table 3: BIC scores of greedy equivalent search and greedy directed acyclic graph search on different type of microarray data; smaller is better. GES, greedy equivalence search; GDS_{EEV}, greedy directed acyclic graph search with equal error variances.

greedy directed acyclic graph search proposes a directed acyclic graph with 192 edges, greedy equivalence search a graph with 217 edges. There are 91 edges in both solutions, 61 with the same orientation. The graphs therefore differ on roughly half of the edges.

Acknowledgement

We thank R. Tanase for fruitful discussions. The research leading to these results received funding from the European Union's Seventh Framework Programme.

Appendix

Some lemmata

In the following two sections we consider different subsets of the set of variables $\mathbf{X}$: to simplify notation we do not distinguish between indices and variables, since the context should clarify the meaning. This way, we can also speak of the parents $\mathbf{PA}_B^{\mathcal{G}}$ of a variable $B \in \mathbf{X}$. We also consider sets of variables $\mathbf{S} \subset \mathbf{X}$ to be a single multivariate variable.

The following four statements are all plausible and their proofs mostly involve technicalities. The reader may skip to the next section and use the lemmata whenever needed.

Lemma 5 *Let* $(A_1, \ldots, A_m) \sim \mathcal{N}\{(\mu_1, \ldots, \mu_m)^T, \Sigma\}$ *with strictly positive definite* Σ *and define* $A_1^* = A_{1 \mid (A_2,\ldots,A_m)=(a_2,\ldots,a_m)}$*, in distribution. Then* $\text{var}(A_1^*) \leq \text{var}(A_1)$ *for all* $(a_2, \ldots, a_m) \in \mathbb{R}^{m-1}$.

We use the notation of conditional variables rather than conditional distributions to improve readability.
Proof. Let us decompose Σ into

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \Sigma_{12}^T \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix}$$

with an $(m-1) \times 1$ vector Σ_{12}. Since Σ_{22}^{-1} is positive definite, $\text{var}(A_1^*) = \sigma_1^2 - \Sigma_{12}^T \Sigma_{22}^{-1} \Sigma_{12} \leq \sigma_1^2$. □

Lemma 6 *[Peters et al., 2011] Let* Y, N, Q *and* R *be random variables taking values in* $\mathcal{Y}, \mathcal{N}, \mathcal{Q}$ *and* $\mathcal{R}$*, respectively, whose joint distribution is absolutely continuous with respect to some product measure; we denote the densities by* $p_{Y,Q,R,N}(y, q, r, n)$*. The variables* Q *and* R *can be multivariate. Let* $f : \mathcal{Y} \times \mathcal{Q} \times \mathcal{N} \to \mathbb{R}$ *be a measurable function. If* $N \perp\!\!\!\perp (Y, Q, R)$ *then for all* $q \in \mathcal{Q}, r \in \mathcal{R}$ *with* $p_{Q,R}(q, r) > 0$*:* $f(Y, Q, N) \,|_{Q=q,R=r} = f(Y \,|_{Q=q,R=r}, q, N)$*, in distribution.*

Lemma 7 (Peters et al. [2011]) *Let* $\mathcal{L}(\mathbf{X})$ *be generated by a structural equation model as in* (2) *with corresponding directed acyclic graph* $\mathcal{G}$ *and consider a variable* $X \in \mathbf{X}$*. If* $\mathbf{S} \subseteq \mathbf{ND}_X^{\mathcal{G}}$ *then* $N_X \perp\!\!\!\perp \mathbf{S}$.

Figure 3: Nodes adjacent to L in $\mathcal{G}$ and $\mathcal{G}'$

Lemma 8 *Let $\mathcal{L}(\mathbf{X})$ be generated from a structural equation model as in (2) with directed acyclic graph $\mathcal{G}$. Consider a variable $B \in \mathbf{X}$ and one of its parents $A \in \mathbf{PA}_B^{\mathcal{G}}$. For all sets $\mathbf{S}$ with $\mathbf{PA}_B^{\mathcal{G}} \setminus \{A\} \subseteq \mathbf{S} \subseteq \mathbf{ND}_B^{\mathcal{G}} \setminus \{A\}$ we have $B \not\perp\!\!\!\perp A \mid \mathbf{S}$.*

Proof. Define $Q = \mathbf{PA}_B^{\mathcal{G}} \setminus \{A\}$ such that we have $\mathbf{S} = (Q, R)$ for some R. Using Lemma 6 we obtain:

$$B|_{Q=q,R=r} = f(q) + \beta A|_{Q=q,R=r} + N_B \,,$$

in distribution, with $N_B \perp\!\!\!\perp A|_{Q=q,R=r}$. But since $\beta \neq 0$, $A|_{Q=q,R=r} \not\perp\!\!\!\perp B|_{Q=q,R=r}$. □

Proof of Theorem 1.

If we assumed faithfulness, we could recover the correct Markov equivalence class, which itself implies the existence of an L and Y shown in Remark 2 [Chickering, 1995, Theorem 2]. Since we are not assuming faithfulness, proving existence of a situation similar to that in Fig. 1 requires more work. This part of the proof, due to not assuming faithfulness, is taken from Peters et al. [2011] and remains almost the same. The difference to Peters et al. [2011] is that we can prove causal minimality and need not assume it. New are also Lemmata 5 and 8, as well as the proof's main argument given in the second part of case (ii).

Proof. We assume that there are two structural equation models as in equation (2) that both induce $\mathcal{L}(\mathbf{X})$, one with graph $\mathcal{G}$, the other with graph $\mathcal{G}'$. We will show that $\mathcal{G} = \mathcal{G}'$. Since directed acyclic graphs do not contain any cycles, we always find nodes that have no descendants. To see this start a directed path at some node; after at most $\#\mathbf{X} - 1$ steps we reach a node without a child. Eliminating such a node from the graph leads to a directed acyclic graph, again; we can discard further nodes without children in the new graph. We repeat this process for all nodes that have no children in both $\mathcal{G}$ and $\mathcal{G}'$ and have the same parents in both graphs. If we end up with no nodes left, the two graphs are identical and the result is proved. Otherwise, we end up with a smaller set of variables that we again call $\mathbf{X}$, two smaller graphs that we again call $\mathcal{G}$ and $\mathcal{G}'$ and a node L that has no children in $\mathcal{G}$ and either $\mathbf{PA}_L^{\mathcal{G}} \neq \mathbf{PA}_L^{\mathcal{G}'}$ or $\mathbf{CH}_L^{\mathcal{G}'} \neq \emptyset$. We will show that this leads to a contradiction. Importantly, because of the Markov property of the distribution with respect to $\mathcal{G}$, all other nodes are independent of L given $\mathbf{PA}_L^{\mathcal{G}}$:

$$L \perp\!\!\!\perp \mathbf{X} \setminus (\mathbf{PA}_L^{\mathcal{G}} \cup \{L\}) \;\mid\; \mathbf{PA}_L^{\mathcal{G}} \,. \tag{4}$$

To make the arguments easier to understand, we introduce the following notation, see also Fig. 3. We partition $\mathcal{G}$-parents of L into $\mathbf{Y}, \mathbf{Z}$ and $\mathbf{W}$. Here, $\mathbf{Z}$ are also $\mathcal{G}'$-parents of L, $\mathbf{Y}$ are $\mathcal{G}'$-children of L and $\mathbf{W}$ are not adjacent to L in $\mathcal{G}'$. Let $\mathbf{D}$ be the $\mathcal{G}'$-parents of L that are not adjacent to L in $\mathcal{G}$ and by $\mathbf{E}$ the $\mathcal{G}'$-children of L that are not adjacent to L in $\mathcal{G}$. Thus: $\mathbf{PA}_L^{\mathcal{G}} = \mathbf{Y} \cup \mathbf{Z} \cup \mathbf{W}$, $\mathbf{CH}_L^{\mathcal{G}} = \emptyset$, $\mathbf{PA}_L^{\mathcal{G}'} = \mathbf{Z} \cup \mathbf{D}$, $\mathbf{CH}_L^{\mathcal{G}'} = \mathbf{Y} \cup \mathbf{E}$. Consider $\mathbf{T} = \mathbf{W} \cup \mathbf{Y}$. We distinguish two cases.

Case (i): $\mathbf{T} = \emptyset$. Then there must be a node $D \in \mathbf{D}$ or a node $E \in \mathbf{E}$, otherwise L would have been discarded. If there is a $D \in \mathbf{D}$ then (4) implies $L \perp\!\!\!\perp D \mid \mathbf{S}$ for $\mathbf{S} = \mathbf{Z} \cup \mathbf{D} \setminus \{D\}$, which contradicts Lemma 8 applied to $\mathcal{G}'$. If $\mathbf{D} = \emptyset$ and there is $E \in \mathbf{E}$ then $E \perp\!\!\!\perp L \mid \mathbf{S}$ holds for $\mathbf{S} = \mathbf{Z} \cup \mathbf{PA}_E^{\mathcal{G}'} \setminus \{L\}$, which also contradicts Lemma 8; to avoid cycles it is necessary that $\mathbf{Z} \subseteq \mathbf{ND}_E^{\mathcal{G}'}$.

Case (ii): $\mathbf{T} \neq \emptyset$. Then $\mathbf{T}$ contains a $\mathcal{G}'$-youngest node with the property that there is no directed $\mathcal{G}'$-path from this node to any other node in $\mathbf{T}$. This node may not be unique.

Suppose that $W \in \mathbf{W}$ is such a youngest node. Consider the directed acyclic graph $\tilde{\mathcal{G}}'$ that equals $\mathcal{G}'$ with additional edges $Y \to W$ and $W' \to W$ for all $Y \in \mathbf{Y}$ and $W' \in \mathbf{W} \setminus \{W\}$. In $\tilde{\mathcal{G}}'$, L and W are not

adjacent. Thus we find a set $\tilde{\mathbf{S}}$ such that $\tilde{\mathbf{S}}$ d-separates L and W in $\tilde{\mathcal{G}}'$; indeed, one can take $\tilde{\mathbf{S}} = \mathbf{PA}_L^{\tilde{\mathcal{G}}'}$ if $W \notin \mathbf{DE}_L^{\tilde{\mathcal{G}}'}$ and $\tilde{\mathbf{S}} = \mathbf{PA}_W^{\tilde{\mathcal{G}}'}$ if $L \notin \mathbf{DE}_W^{\tilde{\mathcal{G}}'}$. Then $\mathbf{S} = \tilde{\mathbf{S}} \cup \{\mathbf{Y}, \mathbf{Z}, \mathbf{W} \setminus \{W\}\}$ d-separates L and W in $\tilde{\mathcal{G}}'$.

We now prove this claim. All $Y \in \mathbf{Y}$ are already in $\tilde{\mathbf{S}}$ in order to block $L \to Y \to W$. Suppose there is a $\tilde{\mathcal{G}}'$-path that is blocked by $\tilde{\mathbf{S}}$ and unblocked if we add Z and W' nodes to $\tilde{\mathbf{S}}$. How can we unblock a path by including more nodes? The path $L \cdots V_1 \cdots U_1 \cdots W$, see Fig. 4, must contain a collider V_1 that is an ancestor of a Z with $V_1, \ldots, V_m, Z \notin \tilde{\mathbf{S}}$ and corresponding nodes U_i for a W' node. Choose V_1 and U_1 on the given path so close to each other such that there is no such collider in between. If there is no V_1, choose U_1 close to L, if there is no U_1, choose V_1 close to W. Now the path $L \leftarrow Z \cdots V_1 \cdots U_1 \cdots W' \to W$ is unblocked given $\tilde{\mathbf{S}}$, which contradicts the fact that $\tilde{\mathbf{S}}$ d-separates L and W. This ends the claim's proof.

The set $\mathbf{S}$ d-separates L and W also in $\mathcal{G}'$ because $\mathcal{G}'$ contains less paths. We have $L \perp\!\!\!\perp W \mid \mathbf{S}$ which contradicts Lemma 8 applied to $\mathcal{G}$. Summarizing, $W \in \mathbf{W}$ cannot be the $\mathcal{G}'$-youngest node.

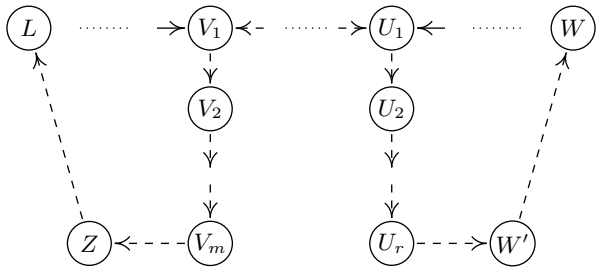


Figure 4: Assume the path $L \cdots V_1 \cdots U_1 \cdots W$ is blocked by $\tilde{\mathbf{S}}$, but unblocked if we include Z and W'. Then the dashed path is unblocked given $\tilde{\mathbf{S}}$.

Therefore, the $\mathcal{G}'$-youngest node in $\mathbf{T}$ must be some $Y \in \mathbf{Y}$. It holds that

$$\sigma^2_{\mathcal{G}} = \sigma^2_{\mathcal{G}'} = \min_{X \in \mathbf{X}} \mathrm{var}(X) = \sigma^2 \,. \tag{5}$$

We define $\mathbf{S} = \mathbf{PA}_L^{\mathcal{G}} \setminus \{Y\} \cup \mathbf{D}$. Clearly, $\mathbf{S} \subseteq \mathbf{ND}_L^{\mathcal{G}}$ since L does not have any descendants in $\mathcal{G}$. Define $Q = \mathbf{PA}_L^{\mathcal{G}} \setminus \{Y\}$ and take any $s = (q, d)$. Define $L^* = L \,|_{\mathbf{S}=s}$, in distribution, and $Y^* = Y \,|_{\mathbf{S}=s}$, in distribution. Then, from $\mathcal{G}$ and using Lemma 6 we find $L^* = f_L(q, Y^*) + N_L = f(q) + \beta \cdot Y^* + N_L$, in distribution, with $N_L \perp\!\!\!\perp Y \,|_{\mathbf{S}=s}$. The independence holds because $\mathbf{S} \subseteq \mathbf{ND}_L^{\mathcal{G}}$. Then, we have

$$\mathrm{var}(L^*) = \beta^2 \mathrm{var}(Y^*) + \sigma^2 > \sigma^2 \,. \tag{6}$$

Since $\mathbf{PA}_L^{\mathcal{G}'} \subseteq \mathbf{S}$ we find from $\mathcal{G}'$ and Lemma 5 that

$$\mathrm{var}(L^*) \leq \sigma^2 \,. \tag{7}$$

since $\det\{\mathrm{cov}(\mathbf{X})\} \neq 0$. Equations (6) and (7) contradict each other.

To prove Remark 4, replace $\mathrm{var}(X)$ by $\mathrm{var}(X)/\alpha_X$ in (5) and σ^2 by $\sigma^2 \alpha_X$ in (6) and (7). □

References

S. A. Andersson, D. Madigan, and M. D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *Annals of Statistics*, 25:505–541, 1997.

P. Bühlmann, M. Kalisch, and L. Meier. High-dimensional statistics with a view towards applications in biology. *Annual Review of Statistics and its Applications (to appear)*, 2013.

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

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

D.M. Cvetković, M. Doob, and H. Sachs. *Spectra of Graphs: Theory and Application*. Barth, Heidelberg, third and enlarged edition, 1995.

M. Dettling and P. Bühlmann. Boosting for tumor classification with gene expression data. *Bioinformatics*, 19(9):1061–1069, 2003.

P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21 (NIPS 2008)*, pages 689–696, Red Hook, New York, 2009. Curran Associates, Inc.

S. Lauritzen. *Graphical Models*. Oxford University Press, New York, 1996.

OEIS Foundation Inc. The on-line encyclopedia of integer sequences. `http://oeis.org/A003024`, 2011.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2nd edition, 2009.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Identifiability of causal graphs using functional models. In *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 589–598, Corvallis, Oregon, 2011. AUAI Press.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, Massachusetts, 2nd edition, 2000.

R. Tillman, A. Gretton, and P. Spirtes. Nonlinear directed acyclic structure learning with weakly additive noise models. In *Advances in Neural Information Processing Systems 22 (NIPS 2009)*, pages 1847–1855, Red Hook, New York, 2010. Curran Associates, Inc.

C. Uhler, G. Raskutti, P. Bühlmann, and B. Yu. Geometry of the faithfulness assumption in causal inference. *Annals of Statistics*, 41(2):436–463, 2013.

S. van de Geer and P. Bühlmann. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2):536–567, 2013.

J. Zhang and P. Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18:239–271, 2008.