Biometrika (2023), **110**, 4, *pp.* 1117–1124 https://doi.org/10.1093/biomet/asad008
Advance Access publication 8 February 2023

Ancestor regression in linear structural equation models

By C. SCHULTHEISS and P. BÜHLMANN

Seminar for Statistics, ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland
christoph.schultheiss@stat.math.ethz.ch buehlmann@stat.math.ethz.ch

Summary

We present a new method for causal discovery in linear structural equation models. We propose a simple technique based on statistical testing in linear models that can distinguish between ancestors and non-ancestors of any given variable. Naturally, this approach can then be extended to estimating the causal order among all variables. We provide explicit error control for false causal discovery, at least asymptotically. This holds true even under Gaussianity, where other methods fail due to non-identifiable structures. These Type I error guarantees come at the cost of reduced power. Additionally, we provide an asymptotically valid goodness-of-fit p-value for assessing whether multivariate data stem from a linear structural equation model.

Some key words: Causal inference; LiNGAM; Structural equation model.

1. Introduction

We propose a very simple yet effective method for inferring the ancestor variables in a linear structural equation model from observational data. Consider a response variable of interest Y and covariates X in a linear structural equation model. The procedure is as follows. For a nonlinear function $f(\cdot)$, such as $f(Y) = Y^3$, run a least squares regression of $f(Y)$ versus Y and all covariates X; the p-value corresponding to the kth covariate X_k measures the significance of X_k being an ancestor variable of Y, and it provides Type I error control. We refer to this method as ancestor regression. Its power, i.e., Type II error, depends on the nature of the underlying data-generating probability distribution. Obviously, the proposed method is extremely simple and easy to use; yet it can deal with the difficult problem of finding the causal order among random variables. In particular, the proposed method does not require any new software and is computationally very efficient.

Structure search methods based on observational data for the graphical structure in linear structural equation models have been developed extensively for various settings, including the Markov equivalence class in linear Gaussian structural equation models (Spirtes et al., 2001, § 5.4; Chickering, 2002), the single identifiable directed acyclic graph in non-Gaussian linear structural equation models (Shimizu et al., 2006; Gnecco et al., 2021) and models with equal error variances (Peters & Bühlmann, 2014). None of these methods comes with p-values and Type I error control. Moreover, for the identifiable cases, the corresponding algorithms require certain assumptions such as non-Gaussian errors. In particular, the method of Shimizu et al. (2006) and extensions thereof are not consistent when there are at least two normally distributed additive error terms involved such that false causal claims cannot be avoided even in the large-sample limit. If the errors are just slightly non-Gaussian, the method requires a very large number samples to achieve favourable behaviour. In contrast, our procedure does not rely on any condition apart from linearity but automatically exploits whether the structure is identifiable or not. In the latter case, we miss out on some causal relationships, but our Type I error control retains the same asymptotic guarantees. The price paid for these guarantees is reduced empirical power relative to competing methods, with the reduction being substantial in some cases.

Regarding notation, we use upper-case letters to denote random variables, e.g., X and Y, and use lower-case letters to denote independent and identically distributed copies of a random variable, e.g., x. If $X \in \mathbb{R}^p$, then $x \in \mathbb{R}^{n \times p}$. With a slight abuse of notation, x can denote either the copies or the realizations thereof. We write x_j for the jth column of a matrix x and $x_{i,j}$ for the element in row i and column j. With $\leftarrow$ we emphasize that an equality between random variables is induced by a causal mechanism. All proofs are given in the Supplementary Material.

2. Ancestor regression

2.1. *Model and method*

Let $X \in \mathbb{R}^p$ be governed by the linear structural equation model

$$X_j \leftarrow \Psi_j + \sum_{k \in \text{PA}(j)} \theta_{j,k} X_k \quad (j = 1, \ldots, p), \tag{1}$$

where $\Psi_1, \ldots, \Psi_p$ are independent and centred random variables. We assume that $0 < \text{var}(\Psi_j) = \sigma_j^2 < \infty$ for all j such that the covariance matrix of X exists and has full rank. We use $\text{PA}(j)$, $\text{CH}(j)$, $\text{AN}(j)$ and $\text{DE}(j)$ to denote j's parents, children, ancestors and descendants, respectively. Further, we assume that there exists a directed acyclic graph representing this structure.

Let X_j with $j \in \{1, \ldots, p\}$ be a variable of interest, which in § 1 was the response Y. Consider a nonlinear function $f(\cdot)$. The following result describes the population property of ancestor regression, with general function $f(\cdot)$.

THEOREM 1. *Assume that the data X follow model (1). Consider the ordinary least squares regression $f(X_j)$ versus X. Denote the corresponding ordinary least squares parameter by $\beta^{f,j} := E(XX^{\mathrm{T}})^{-1}E\{Xf(X_j)\}$ and assume that it exists. Then*

$$\beta_k^{f,j} = 0 \quad \forall k \notin \{\text{AN}(j) \cup j\}.$$

Importantly, X_j itself must also be included in the set of predictors. The beauty of Theorem 1 lies in the fact that no assumptions on the distribution of the Ψ_l or the size of the $\theta_{l,k}$, apart from existence of the moments, need be imposed for any l and $k \in \{1, \ldots, p\}$. This allows one to control against the false discovery of ancestor variables.

Typically, $\beta_k^{f,j} \neq 0$ holds for an ancestor since a nonlinear function of that ancestor cannot be completely regressed out by the other regressors using only linear terms. For ancestors that are much further upstream, this effect could become vanishingly small. However, this is not a big issue because when fitting a linear model using the detected ancestors, those indirect ancestors are assigned a direct causal effect of 0 anyway.

Based on Theorem 1, we suggest testing for $\beta_k^{f,j} \neq 0$ to detect some or even all ancestors of X_j. Doing so for all k requires nothing more than fitting a multiple linear model and using its corresponding z-tests for individual covariates.

Let $x \in \mathbb{R}^{n \times p}$ be n independent and identically distributed copies from (1). Define the quantities

$$\hat{\beta}^{f,j} := (x^{\mathrm{T}}x)^{-1}x^{\mathrm{T}}f(x_j), \quad \hat{\sigma}^2 := \frac{\|f(x_j) - x\hat{\beta}^{f,j}\|_2^2}{n-p}, \quad \hat{\text{var}}(\hat{\beta}_k^{f,j}) = (x^{\mathrm{T}}x)_{k,k}^{-1}\hat{\sigma}^2, \tag{2}$$

where $f(\cdot)$ is understood to be applied elementwise in $f(x_j)$.

THEOREM 2. *Assume that the data X follow model (1), $E\{f(X_j)^2\} < \infty$, $E(X_k^4) < \infty$ for all k, and $\beta^{f,j}$ exists. Let x be n independent and identically distributed copies thereof. Using the definitions from*

(2), *we have*

$$\hat{\beta}_k^{f,j} = \beta_k^{f,j} + o_{\mathrm{p}}(1), \quad \hat{\mathrm{var}}(\hat{\beta}_k^{f,j}) = O_{\mathrm{p}}\left(\frac{1}{n}\right),$$

$$z_k^j := \frac{\hat{\beta}_k^{f,j}}{\{\hat{\mathrm{var}}(\hat{\beta}_k^{f,j})\}^{1/2}} \xrightarrow{\mathrm{d}} N(0,1) \quad \forall k \notin \{\mathrm{AN}(j) \cup j\}.$$

Because of this limiting distribution, we suggest testing the null hypothesis $H_{0,k\to j} : k \notin \mathrm{AN}(j)$ with the p-value

$$p_k^j = 2\{1 - \Phi(|z_k^j|)\}, \tag{3}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

For ancestors for which $\beta_k^{f,j} \neq 0$, the absolute z-statistic increases as $\sqrt{n}$. In typical set-ups, one can thus detect all ancestors. Having found all ancestors, one could infer the parents with an ordinary least squares regression of X_j versus $X_{\mathrm{AN}(j)}$, using the t-test for assigning the significance of being a parental variable. Such a procedure may have poor error control for small samples as it requires full power in the first step to detect all ancestors; we provide error control only for the estimated ancestral set.

The choice of $f(\cdot)$ has an impact on the constant in the growth of z_k^j for ancestors. If the Ψ_l are symmetric, any even function yields $\beta_k^{f,j} = 0$ for all k. Therefore, odd functions should be used. In our simulations and the real-data analysis, we use $f(X_j) = X_j^3$ as it is the simplest odd function, which induces only slightly higher moments than linear functions. This choice leads to empirically competitive performance relative to other candidates in our simulations.

2.2. *Adversarial set-ups*

There are cases where $\beta_k^{f,j} \neq 0$ does not hold for some ancestors, leading to reduced power of the method. We provide necessary and sufficient conditions for this to hold and present some examples. We first define the j-restricted Markov boundary of k as

$$\mathrm{MA}^{\to j}(k) := \left[\mathrm{PA}(k) \cup \mathrm{CH}(k) \cup \bigcup_{l \in \mathrm{CH}(k)} \{\mathrm{PA}(l) \setminus k\}\right] \cap \{\mathrm{AN}(j) \cup j\}.$$

It contains all the variables in the Markov boundary of k which are ancestors of j or j itself. For example, if $k \in \mathrm{AN}(j)$, then all its parents are in the restricted Markov boundary, but not necessarily all its children.

THEOREM 3. *Let $k \in \mathrm{AN}(j)$. Then*

$$\beta_k^{f,j} = 0 \ \ \forall f(\cdot) \quad \textit{if and only if} \quad E(X_k \mid X_j) = E(X_{\mathrm{MA}\to j(k)}^{\mathrm{T}} \gamma^{j,k} \mid X_j),$$

where $\gamma^{j,k}$ is the least squares parameter for regressing X_k versus $X_{\mathrm{MA}\to j(k)}$. In particular,

$$\beta_k^{f,j} = 0 \ \ \forall f(\cdot) \quad \textit{if} \quad E(X_k \mid X_{\mathrm{MA}\to j(k)}) = X_{\mathrm{MA}\to j(k)}^{\mathrm{T}} \gamma^{j,k}.$$

Intuitively speaking, if the conditional expectation of X_k given the j-restricted Markov boundary is linear, k could also be a child of all these variables; hence it is not detectable as an ancestor of j. In the following, we present two examples that fulfil the conditions of Theorem 3. These are the only examples that we know of.

Example 1 (*Gaussian*). It is well known in causal discovery for linear structural equation models that Gaussian error terms lead to nonidentifiability. Define $\text{CH}^{\to j}(k) := [\text{CH}(k) \cap \{\text{AN}(j) \cup j\}]$, i.e., the children of k through which a directed path from k to j begins.

PROPOSITION 1. *Assume that the data X follow model* (1)*. Let $k \in \text{AN}(j)$ with $\Psi_k \sim N(0, \sigma_k^2)$. Then*

$$\beta_k^{f,j} = 0 \ \ \forall f(\cdot) \quad \text{if} \quad \Psi_l \sim N(0, \sigma_l^2) \ \ \forall\, l \in \text{CH}^{\to j}(k).$$

Under the additional assumptions of Theorem 2*,*

$$z_k^j := \frac{\hat{\beta}_k^{f,j}}{\{\hat{\text{var}}(\hat{\beta}_k^{f,j})\}^{1/2}} \xrightarrow{\text{d}} N(0,1).$$

Therefore, if every directed path from k to j starts with an edge for which the nodes on both ends have Gaussian noise terms, we have no power to detect this ancestor relationship. However, nor do we detect the opposite direction as guaranteed by Theorem 1, and thus control against false positives is guaranteed.

Example 2 (*Special constellation of distributions and coefficients*). A pathological case occurs when the error term of a child, say l, has the same distribution as the inherited contribution from the error term of the parent, say k. Then k is not detectable as l's ancestor. Likewise, it is not detected as an ancestor of any of l's descendants j to which all directed paths from k start with the edge $k \to l$.

PROPOSITION 2. *Assume that the data X follow model* (1)*. Let $k \in \text{AN}(j)$ and $\text{CH}^{\to j}(k) = \{l\}$. Then*

$$\beta_k^{f,j} = 0 \ \ \forall f(\cdot) \quad \text{if} \quad \Psi_l \overset{\text{d}}{=} \theta_{l,k}\Psi_k.$$

For the variables discussed here, the limiting Gaussian distribution as stated in Theorem 2 is not guaranteed even though $\beta_k^{f,j} = 0$; see also the proof in the Supplementary Material.

2.3. *Simulation example*

We study ancestor regression in a small simulation example. We generate data from a linear structural equation model with six variables. The causal order is fixed as X_1 to X_6. Otherwise, the structure is randomized and changes in each simulation run: X_k is a parent of X_l for $k < l$ with probability 0.4 such that there is an average of six parental relationships. The edge weights are sampled uniformly and the Ψ_k are assigned by permuting a fixed set of six error distributions. The full data-generating process is described in the Supplementary Material.

We aim to find the ancestors of X_4, which can be any subset of $\{X_1, X_2, X_3\}$. We create 1000 different set-ups and test each on samples of sizes varying from 10^2 to 10^6. For the nonlinear function we use $f(X_j) = X_j^3$. By z-statistic we mean z_k^4 as defined in Theorem 2. We calculate p-values according to (3) and apply a Bonferroni–Holm correction to them, without cutting off at 1 for the sake of visualization.

In Fig. 1 we see the desired $\sqrt{n}$ growth of the absolute z-statistic for the ancestors, while for the non-ancestors their sample averages are close to the theoretical mean under the asymptotic null distribution. Indirect ancestors are harder to detect than parents. Although the null distribution is only asymptotically achieved, the Type I familywise error rate is controlled for every sample size, supporting our method's main benefit, namely robustness against false causal discovery.

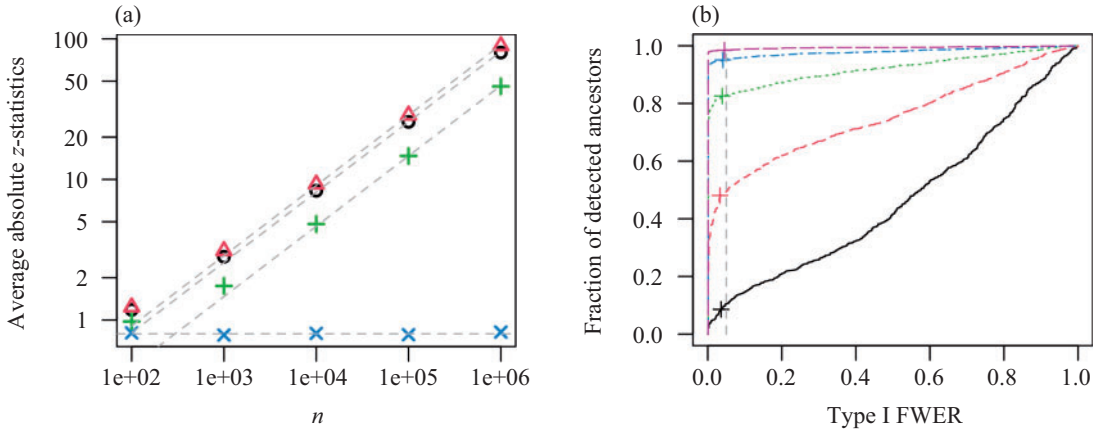

Fig. 1. Detection of the ancestors of X_4 in a linear structural equation model with six variables, where the results are based on 1000 simulation runs: (a) average absolute z-statistic for all ancestors (black circles), parents (red triangles), non-parental ancestors (green plus signs) and non-ancestors (blue crosses) for different sample sizes n, where the dashed lines correspond to $\sqrt{n}$ growth fitted to perfectly match at $n = 10^5$ and the horizontal dashed line corresponds to $(2/\pi)^{1/2}$, i.e., the first absolute moment of the asymptotic null distribution, a standard Gaussian; (b) fraction of simulation runs with at least one false causal detection, i.e., the Type I familywise error rate (FWER), versus the fraction of detected ancestors for sample sizes 10^2 (black solid), 10^3 (red dashed), 10^4 (green dotted), 10^5 (blue dot-dashed) and 10^6 (pink long-dashed), where the curves use the level α of the test as an implicit parameter, with the plus symbols corresponding to a nominal α of 5% and the grey vertical dashed line to the actual 5%.

3. Ancestor detection in networks: nodewise and recursive

3.1. *Algorithm and goodness-of-fit test*

In the previous section, we assumed that there is a response variable X_j that is of special interest. This is not always the case. Instead, one might be interested in inferring the full set of causal connections between the variables. Naturally, our ancestor detection technique can be extended to that problem by applying it nodewise. We suggest the procedure sketched below. The detailed algorithm can be found in the Supplementary Material. Notably, the algorithm is invariant with respect to the ordering of the variables.

First, the set of ancestors is defined based on the significant p-values, after multiplicity correction over all $p(p-1)$ z-tests, of ancestor regression. Any correction controlling the Type I familywise error rate is applicable, and here we use the Bonferroni–Holm correction. Next, further ancestral relationships are constructed recursively by adding the estimated ancestors of every estimated ancestor; this recursive construction facilitates the detection of all ancestors. The procedure cannot increase the Type I familywise error rate over just using the significant p-values, because a false causal discovery can be propagated only if it existed in the first place.

Since there is no guarantee that the recursive construction will not create directed cycles, such that variables are claimed to be their own ancestors, we need to address this concern. If such cycles are found, the significance level is gradually reduced until no more directed cycles are output. This means that the output becomes somewhat independent of the significance level; for example, in a case with two variables and with $p_1^2 = 10^{-6}$ and $p_2^1 = 10^{-3}$ as in (3), we would never claim that $X_2 \to X_1$ no matter how large α is chosen. We denote the estimated set of ancestors for X_j by $\hat{\text{AN}}(j)$. Notably, the algorithm determines a causal order between the variables but does not always lead to a unique parental graph. For instance, if $\hat{\text{AN}}(3) = \{1, 2\}$ and $\hat{\text{AN}}(2) = \{1\}$, X_1 may be a causal parent of X_3, but its effect could also be fully mediated by X_2.

One can regard the greatest significance level such that no loops are created as a p-value for the null hypothesis that the modelling assumption (1) holds. We denote this level, which is a further output of our algorithm, by $\hat{\alpha}$. Thus we have a goodness-of-fit test for our modelling assumption with an asymptotically valid p-value: a small realized $\hat{\alpha}$ would provide evidence against the linear structural equation model in (1). If such evidence exists, it is advisable to take the outcome of ancestor regression, or any other causal discovery method that relies on linear structural equation models, with

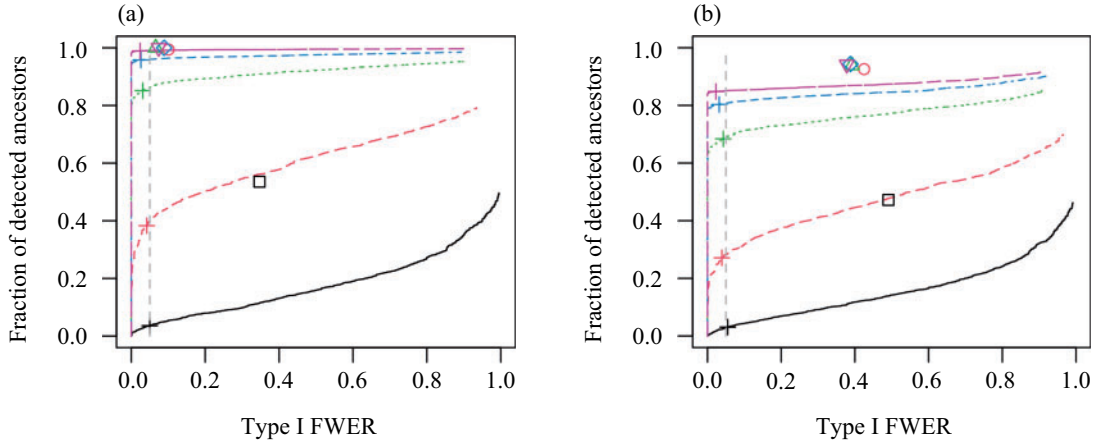

Fig. 2. Nodewise ancestor detection in a linear structural equation model with six variables. The results are based on 1000 simulation runs. Each simulation run uses two different data-generating processes: (a) exactly one error term follows a Gaussian distribution; (b) exactly two error terms follow a Gaussian distribution. Each panel plots the familywise error rate (FWER) of false causal detection versus the fraction of detected ancestors. The curves use the level of the test α as the implicit parameter, with the plus symbols corresponding to a nominal α of 5% and the grey vertical dashed line to the actual 5%. The other symbols represent the performance of the LiNGAM algorithm. We consider different sample sizes: 10^2 (black solid line and square), 10^3 (red dashed line and circle), 10^4 (green dotted line and upward-pointing triangle), 10^5 (blue dot-dashed line and diamond) and 10^6 (pink long-dashed line and downward-pointing triangle).

caution. We make use of this p-value in the data analysis in § 4. The next corollary summarizes the properties of our algorithm.

COROLLARY 1. *Assume that the conditions of Theorem 2 hold for all* $j \in \{1, \ldots, p\}$. *Let* $\hat{\text{AN}}(j)$ *for all* $j \in \{1, \ldots, p\}$ *and* $\hat{\alpha}$ *be the output of the nodewise ancestor regression algorithm with significance level* α *and Bonferroni–Holm correction. Then*

$$\lim_{n\to\infty} \text{pr}\{\exists j, k \neq j : k \notin \text{AN}(j),\ k \in \hat{\text{AN}}(j)\} \leqslant \alpha, \qquad \lim_{n\to\infty} \text{pr}(\hat{\alpha} \leqslant \alpha') \leqslant \alpha' \ \ \forall \alpha' \in (0, \alpha).$$

3.2. *Simulation example*

We extend the simulation from § 2.3 to estimation of the ancestors of each variable using the algorithm described in § 3.1. We compare our method with LiNGAM (Shimizu et al., 2006) using the implementation provided in the `R` (R Development Core Team, 2023) package `pcalg` (Kalisch et al., 2012). For every simulation run, we use two slightly different data-generating processes. In the first, only one of the Ψ_k follows a Gaussian distribution; in the second, two of the error terms follow a Gaussian distribution, and an edge between their respective nodes is always present. As LiNGAM provides an estimated set of parents, we additionally apply our recursive algorithm to its output to get an estimated set of ancestors, which enables comparison with our method.

The results are shown in Fig. 2. For the model with only one Gaussian error variable, we can reliably detect almost all ancestors without any false causal claims for sufficiently large sample sizes. The few exceptions can be explained by some set-ups being very close to the nonidentifiable case discussed in Proposition 2. Not all curves reach a power of 1 even when the significance level is taken to be arbitrarily large. This can be explained by the possible insensitivity to significance level outlined in § 3.1.

We are able to control the familywise error rate even for small samples using a nominal size of $\alpha = 5\%$, supporting our theoretical results. This is not the case for LiNGAM, which is designed such that it must always determine a causal order based on the underlying independent component analysis (Hyvarinen, 1999), even when not enough information is available. Therefore, no Type I error guarantees can be provided. The power of LiNGAM approaches 1 much faster than ancestor regression; and if one allows for a bit more liberal Type I error, LiNGAM appears preferable in the model with one Gaussian noise term. The picture changes when one looks at slight violations of the LiNGAM assumption, i.e., another Gaussian error term. In this case, LiNGAM is still more powerful but does

Table 1. *Analysis of the dataset in Sachs et al. (2005): the second column reports the raw p-value from ancestor regression, p_k^j, associated with the respective edge and the third column the raw p-value from the subsequent linear model fit; the rows are ordered from low to high according to the p-value from ancestor regression.*

Causal effect	Ancestor regression	Linear regression	SC	MH
PIP3 → PIP2	3.3×10^{-39}	5.5×10^{-43}	⟶	⟶
PIP3 → PLCg	6.7×10^{-39}	1.4×10^{-36}	⟶	⇢
PKA → Erk	2.9×10^{-26}	7.2×10^{-2}	⟶	⇢
JNK → p38	6.6×10^{-20}	2.4×10^{-19}	—	—
PKA → Akt	7.2×10^{-20}	9.4×10^{-4}	⟶	⟶
JNK → PKC	1.2×10^{-16}	5.1×10^{-88}	⟵	⟵
RAF → MEK	5.4×10^{-15}	0	⟶	⟵
PKC → p38	3.1×10^{-13}	0	⟶	⟶
Akt → Erk	7.6×10^{-7}	0	—	⟶

SC, presents the conclusions of the consensus network in Sachs et al. (2005); MH, shows the conclusions of the method from Mooij & Heskes (2013); ⟶, the edge is present; ⇢, there exists a directed path with the same orientation but no edge; ⟵, the edge is reversed; —, there is no directed path

not control the error at all. Regardless of the sample size, a wrong causal claim is made in around 40% of the set-ups. Ancestor regression is more robust with respect to this deviation, as the Type I error guarantees do not require non-Gaussian error terms. For the unidentifiable edges, it avoids making any decision and can control the error rate at any desired level at the price of some reduction in power. In this simulation, Proposition 1 applies to around 14% of the ancestral connections.

The Supplementary Material reports additional simulation results for settings varying between non-Gaussian and Gaussian scenarios. When close to the fully Gaussian case, despite satisfying the LiNGAM assumption (Shimizu et al., 2006) in population, this clearly worsens the performance of LiNGAM for finite sample sizes.

4. Real-data example

We analyse the flow cytometry dataset in Sachs et al. (2005), which contains cytometry measurements of 11 phosphorylated proteins and phospholipids. The data are from various experimental conditions, some of which are interventional environments. The authors provide a ground truth on how these quantities affect each other, the so-called consensus network. The dataset has been further analysed in various follow-up papers; see, for example, Mooij & Heskes (2013) and Taeb et al. (2022). Following these works, we consider data from eight different environments, seven of which are interventional. The sample size per environment ranges from 707 to 913.

For each environment individually, we estimate the ancestral relationships using our recursive algorithm sketched in § 3.1 with nonlinear function $f(X_j) = X_j^3$ and $\alpha = 0.05$. The goodness-of-fit p-value $\hat{\alpha}$ per environment, corrected for the number of environments, ranges from 0.14 to 3×10^{-12}. All but one of the p-values are lower than 0.04, indicating that for these environments the data do not follow model (1). The deviation can be in terms of hidden variables, nonlinear effects, or noise that is not additive. While the aforementioned results and other published findings usually yield one graph harmonized over different environments, the fact that our results are highly varying across environments suggests questioning the standard autonomy assumption in causality (Aldrich, 1989) that an intervention does not change the underlying graph, except for edges that point into the intervened node.

Henceforth, we focus on the environment with the highest $\hat{\alpha}$ which seems to be most conformable with a linear structural equation model. The dataset contains 723 observations. For each node, we fit a linear model using the claimed set of ancestors as predictors to see which ancestors could be direct parents. We summarize our findings in Table 1. Most ancestors show an indication of being direct

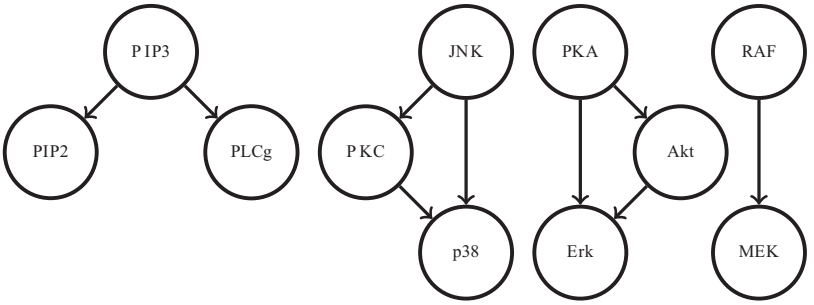

Fig. 3. Ancestral relations obtained with ancestor regression.

parents. However, as laid out in § 2.1, we do not have Type I error guarantees for finding parents in cases where some ancestors are missing.

For comparison, we show the conclusions drawn by the consensus network and Mooij & Heskes (2013) for these edges. The results of our method are in agreement with at least one of these works, except for the two edges coming from JNK. One of the indirect paths of Mooij & Heskes (2013) corresponds to the highest p-value in the linear model fit, which is a further point of agreement. The ancestral graph output by our method, shown in Fig. 3, consists of four disconnected components. When considering these components individually, we observe that the part containing JNK, where we have somewhat unexpected findings, has the strongest indication of violating the model assumptions in terms of the goodness-of-fit p-value $\hat{\alpha}$.

ACKNOWLEDGEMENT

The project leading to this application received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (No 786461).

SUPPLEMENTARY MATERIAL

The Supplementary Material contains proofs and additional simulation results.

REFERENCES

ALDRICH, J. (1989). Autonomy. *Oxford Econ. Papers* **41**, 15–34.
CHICKERING, D. M. (2002). Optimal structure identification with greedy search. *J. Mach. Learn. Res.* **3**, 507–54.
GNECCO, N., MEINSHAUSEN, N., PETERS, J. & ENGELKE, S. (2021). Causal discovery in heavy-tailed models. *Ann. Statist.* **49**, 1755–78.
HYVARINEN, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Networks* **10**, 626–34.
KALISCH, M., MÄCHLER, M., COLOMBO, D., MAATHUIS, M. H. & BÜHLMANN, P. (2012). Causal inference using graphical models with the R package pcalg. *J. Statist. Software* **47**, 1–26.
MOOIJ, J. M. & HESKES, T. (2013). Cyclic causal discovery from continuous equilibrium data. In *Proc. 29th Conf. Uncertainty in Artificial Intelligence (UAI'13)*. Arlington, Virginia: AUAI Press, pp. 431–9.
PETERS, J. & BÜHLMANN, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika* **101**, 219–28.
R DEVELOPMENT CORE TEAM (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.
SACHS, K., PEREZ, O., PE'ER, D., LAUFFENBURGER, D. A. & NOLAN, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**, 523–9.
SHIMIZU, S., HOYER, P. O., HYVÄRINEN, A., KERMINEN, A. & JORDAN, M. (2006). A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.* **7**, 2003–30.
SPIRTES, P., GLYMOUR, C. N. & SCHEINES, R. (2001). *Causation, Prediction, and Search*. New York: Academic Press.
TAEB, A., GAMELLA, J. L., HEINZE-DEML, C. & BÜHLMANN, P. (2022). Perturbations and causality in Gaussian latent variable models. *arXiv:* 2101.06950v3.

[*Received on* 18 *May* 2022. *Editorial decision on* 6 *February* 2023]