Statistical test for consistent estimation of causal effects in linear non-Gaussian models

Doris Entner*, **Patrik O. Hoyer***, **Peter Spirtes****
*HIIT and Department of Computer Science, University of Helsinki
**Department of Philosophy, Carnegie Mellon University

Abstract

In many fields of science researchers are faced with the problem of estimating causal effects from non-experimental data. A key issue is to avoid inconsistent estimators due to confounding by measured or unmeasured covariates, a problem commonly solved by 'adjusting for' a subset of the observed variables. When the data generating process can be represented by a directed acyclic graph, and this graph structure is known, there exist simple graphical procedures for determining which subset of covariates should be adjusted for to obtain consistent estimators of the causal effects. However, when the graph is *not* known no general and complete procedures for this task are available. In this paper we introduce such a method for linear non-Gaussian models, requiring only partial knowledge about the temporal ordering of the variables: We provide a simple statistical test for inferring whether an estimator of a causal effect is consistent when controlling for a subset of measured covariates, and we present heuristics to search for such a set. We show empirically that this statistical test identifies consistent vs inconsistent estimates, and that the search heuristics outperform the naïve approach of adjusting for all observed covariates.

1 Introduction

Researchers in a variety of fields, e.g. epidemiology, econometrics or psychology, are interested in causal relationships between quantities of interest. The preferred approach to inferring such relationships is based on randomized experiments: To estimate the causal effect of some 'treatment' variable x on an 'outcome' variable y, the treatment is randomized so that any statistically significant correlation between the treatment and the outcome is necessarily due to a causal effect.[1] However, it is not always possible to carry out such randomized experiments, so methods for inferring causal effects from non-experimental ('passive observational') data are of great interest. For instance, in epidemiology one might be interested in estimating the causal effect of some risk factor x on some health indicator y, and the researcher can measure but not influence the exposure of individuals to the risk factor.

Typically, in addition to the treatment and outcome variables, data is also available on a set $\mathcal{W}$ of related variables (termed 'covariates'). In the above example these might include age, gender, or indicators of the general health of the patient. Some of these variables may be *confounders*, i.e. affecting both the treatment and the outcome in such a way that a naïve estimator of the causal effect is biased and inconsistent. Hence, we may need to *control for* (i.e. *adjust for*) a subset of the covariates to obtain a consistent and unbiased estimator of the causal effect (Spirtes et al., 1998; Greenland et al., 1999; Spirtes et al., 2000; Pearl, 2009a). Here, it is important to distinguish such statistical control from experimental control. In the latter the researcher actively determines the exposure, for example by assigning patients a certain medicine. In statistical control (the subject of this paper), the control is passive; the causal effect is estimated from the observed joint distribution by conditioning on (i.e. taking into account, or 'controlling for') the appropriate covariates (Pearl, 2009a). For discrete variables this involves performing a stratified analysis and then averaging the results, while in the case of continuous variables this typically involves including some of the covariates in the relevant regression (see Section 2 for details). A

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

[1] Note that we do not consider the problem of 'selection bias' in this paper. That is, we assume that the values of the variables do not influence whether a data vector is included in the dataset or not.

central question in observational studies is thus how to select a variable subset $\mathcal{Z} \subseteq \mathcal{W}$ to adjust for. Because some confounders may not even have been measured (i.e. are outside the set of covariates $\mathcal{W}$), it is even possible that there is no set $\mathcal{Z}$ that yields a consistent estimator of the causal effect when adjusted for.

In general, even when a full temporal ordering of the observed variables is known, without domain-specific background knowledge one cannot infer whether adjusting for a given set of covariates yields a consistent estimator of the causal effect of the treatment on the outcome, as demonstrated in Section 2.1. In this paper, however, we show that in linear models with non-Gaussian variables, one can devise a statistical test for precisely this purpose, assuming only limited knowledge about the time ordering of the variables. (Both linearity and partial knowledge about the time ordering are rather common assumptions in many fields.) We also provide heuristics using this test to search for a variable subset, so as to obtain a consistent estimator of the causal effect, and demonstrate the performance of the test and the search procedure empirically.

Our results parallel recent work in the 'causal discovery' literature showing that non-Gaussianity can aid in structure learning of graphical models (Shimizu et al., 2006; Hoyer et al., 2008). While such methods aim at providing preliminary hypotheses of the causal structure among a set of observed variables, our procedure targets a more restricted setting in which a specific causal effect is estimated, with some (limited) background knowledge of temporal ordering. This allows us to identify causal effects without the restrictive assumption of no hidden variables (Shimizu et al., 2006), yet still avoiding the statistically and computationally difficult task of learning arbitrary graph structures in the presence of latent variables (Hoyer et al., 2008).

2 Model and Method

2.1 A Simple Example

We begin by considering the two linear acyclic causal models (recursive structural equation models) of Figure 1. In both models, the observed variables consist of the treatment x, the outcome y, and a single covariate w. The model in **(b)** additionally contains two hidden variables u_1 and u_2. The causal relationships between the variables are represented by a directed acyclic graph over the full variable set, and the relationships between the variables are given in the linear equations next to the graphs. The disturbance terms e in the model are mutually independent.

We are interested in finding the causal effect α of the treatment x on the outcome y. Formally, this effect is defined as the rate of change in the expected value of y when *setting* x (in an experiment) to a certain value, i.e. $\frac{\partial}{\partial x} E[y|do(x)]$, which in the case of the considered linear generating models coincides with the corresponding edge coefficient α (Pearl, 2009a, Ch. 5.4). We can obtain an unbiased and consistent estimator[2] of this edge coefficient from *observational* data by estimating a regression of y on x and a set $\mathcal{Z}$ (using ordinary least squares, OLS), where the set $\mathcal{Z}$ fulfills the so called *back-door criterion* with respect to the ordered pair (x, y), i.e. $\mathcal{Z}$ does not contain any descendants of x, and $\mathcal{Z}$ blocks (d-separates) every *back-door path* from x to y, that is every path between x and y that contains an arrow into x ("$x \leftarrow$"). Any such set $\mathcal{Z}$ is termed *admissible* (Pearl, 2009a, Ch. 3.3, Ch. 5.3).

First, consider the model in **(a)**. It is well understood that, if $\beta \neq 0$ and $\gamma \neq 0$, regressing y on x while disregarding w leads an inconsistent estimate a of the true causal effect α. On the other hand, 'controlling' for w by including it in the regression, as in $y = ax + bw + r_y$, will lead for this model to a consistent estimate a of α. Thus, in this case it is crucial to include w in the regression. However, in model **(b)** controlling for w has the exact opposite effect: the regression $y = ax + r_y$ yields a consistent estimate a of α, while including w in the regression results in an inconsistent estimate. For details and further discussion see the Supplementary Material and (Spirtes et al., 1998; Greenland et al., 1999; Spirtes, 2000; Pearl, 2009b), respectively.

When domain-specific background knowledge allows us to uniquely choose one of the two models we can decide whether to include w in the regression. However, in many cases one does not *a priori* have such information available. Instead, one often only knows a (partial) temporal ordering. In such cases it would be useful if one could use the data and this ordering to infer which model fits the data. Unfortunately, the two models are equivalent in the set of covariance matrices over the observed variables x, y, and w (see the Supplementary Material for a proof of this fact), which implies that for Gaussian data the two models cannot be distinguished based on the data. Even if we knew that w precedes x which in turn precedes y we could not distinguish the two models for Gaussian data. Thus, knowing a temporal ordering of the vari-

[2] An estimator $\hat{\theta}$ of a parameter θ is *unbiased* if $E(\hat{\theta}) = \theta$, i.e. the expected value of the estimator is the true parameter. An estimator $\hat{\theta}$ is a *consistent* estimator of θ if it converges in probability to θ ($\theta_k \xrightarrow{P} \theta$), i.e. for every $\varepsilon > 0$: $P(|\hat{\theta}_k - \theta| > \varepsilon) \to 0$ as $k \to \infty$, where $\hat{\theta}_k$ is an estimate of θ using k samples, see for example (Wasserman, 2004). In the rest of the paper we will focus on consistency, but we emphasize that in our models the estimators are either both consistent and unbiased, or both inconsistent and biased, which derives from the properties of the OLS estimator.

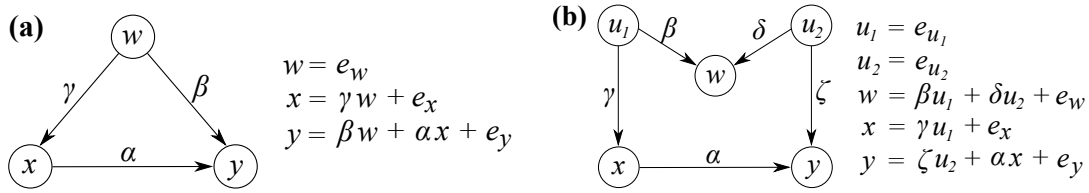

Figure 1: Two example models, with w, x, and y observed, and u_1 and u_2 hidden variables. For instance, let in **(a)** x = education, y = income, w = intelligence, and in **(b)** x = cholesterol, y = blood pressure, w = weight, u_1 = diet, u_2 = exercise. In case the two graphs accurately represent the causal relationships between the variables, adjusting for w in **(a)** yields a consistent estimate of the causal effect of x on y, whereas in **(b)** this results in an inconsistent estimate of α. To understand the latter statement, observe that weight is associated with cholesterol and blood pressure: obesity may result from a high caloric diet and minimal exercise, and hence obese individuals may show high values for both cholesterol and blood pressure, even in case the true causal effect α of the former on the later was 0. Conversely, the naïve estimate of the causal effect of x on y (i.e. not conditioning on w), yields in model **(a)** an inconsistent estimate a of α, and in model **(b)** a consistent estimate (Pearl, 2009a).

ables is in general *not* sufficient to decide whether to adjust for a specific covariate.

Nevertheless, if the variables are *non-Gaussian* it *is* possible to decide whether w should be included in the regression, without knowing a priori which is the true graph, but instead only knowing the time ordering of the observed variables. To see this, consider regressing y on x (disregarding w) using data drawn from model **(a)**. When we estimate the regression $y = ax + r_y$ using OLS, it is well known that the residual r_y is necessarily *linearly uncorrelated* with x (regardless of the confounding). However, when the disturbances are non-Gaussian, the residual r_y is nevertheless *statistically dependent* on x. This can be seen by expressing x and r_y in terms of the disturbances, $x = \gamma e_w + e_x$ and $r_y = y - ax = ((\alpha - a)\gamma + \beta)e_w + (\alpha - a)e_x + e_y$, and noting that whenever the confounding leads to an inconsistent estimator, i.e. $(a - \alpha) \xrightarrow{P} 0$, the disturbance e_x has a non-vanishing influence on both x and r_y. By the Darmois-Skitovitch Theorem (see Appendix) x and r_y are thus necessarily statistically dependent if e_x is non-Gaussian. If, on the other hand, w is included in the regression r_y is asymptotically independent of x, by the assumption of independent disturbances. Similarly, one can show that for model **(b)** the residual r_y is asymptotically independent of x if and only if w is *not* included in the regression. For details see the Supplementary Material.

2.2 Problem Definition

We generalize the above example problem as follows. We consider the task of estimating the causal effect of one continuous-valued scalar random variable (x) on another (y), from samples of these two variables and a set of observed covariates ($\mathcal{W}$). We assume that the covariates $\mathcal{W}$ are known not to be caused by the treatment x, typically because they precede the treatment, and it is also known that the outcome y does not affect any of the other variables in the model.

The data-generating model is assumed to be the following. The full set of random variables is given by $\mathcal{V} = \{x, y\} \cup \mathcal{W} \cup \mathcal{U}$, where x, y, and $\mathcal{W}$ are as above and $\mathcal{U}$ is a (possibly empty) set of latent (unobserved) variables. The variables in $\mathcal{V}$ can be represented by a directed acyclic graph (DAG), and the relationships among the variables are linear. With a vector $\boldsymbol{v}$ containing the variables in $\mathcal{V}$, the model is given by

$$\boldsymbol{v} := \mathbf{B}\boldsymbol{v} + \boldsymbol{e} \tag{1}$$

where the coefficient matrix $\mathbf{B}$ (containing the causal direct effects) is strictly lower triangular when the variables are arranged in a causal ordering. The vector $\boldsymbol{e}$ of (unobserved) disturbance variables consists of mutually independent zero-mean random variables (which is not a strong assumption because latent variables are allowed). Here, we do not place any assumptions on the type of distributions of the disturbances $\boldsymbol{e}$, but we will see that our method will only give informative results when the disturbances are non-Gaussian.

Given this model, the question is now whether the causal effect of x on y can be consistently estimated from data over the observed variables $\{x, y\} \cup \mathcal{W}$. Depending on the sets $\mathcal{W}$ and $\mathcal{U}$, and on the DAG connecting these variables with each other and with x and y, this may or may not be achieved by including a set $\mathcal{Z} \subseteq \mathcal{W}$ in the regression of y on x, as in $y = ax + \boldsymbol{c}^T\boldsymbol{z} + r_y$. When this regression is estimated using OLS, the regression coefficient a is a consistent estimate of the true causal effect α of x on y (i.e. the coefficient of x in the equation for y in Equation (1)) if the set $\mathcal{Z}$ is admissible (as stated in Section 2.1).

If the underlying DAG is known, there exist algorithms for deciding if an admissible set $\mathcal{Z}$ exists, and if so

Algorithm 1 (Statistical test for consistency)

Given a dataset over the observed variables $\{x, y\} \cup \mathcal{W}$ and a set $\mathcal{Z} \subseteq \mathcal{W}$ of covariates to adjust for.

Estimate the following two regressions using OLS:

$$x = \boldsymbol{b}^T \boldsymbol{z} + r_x \quad (2)$$

$$y = ax + \boldsymbol{c}^T \boldsymbol{z} + r_y. \quad (3)$$

Perform a statistical test for Gaussianity of the residual r_x, proceed only if Gaussianity is rejected.

Perform a statistical test for independence of r_x and r_y, interpret the result as follows (Theorem 1):

If independence is *not rejected* at threshold p_t:
The estimated effect a is inferred to be consistent
If independence is *rejected* at threshold p_t:
The estimated effect a is inferred to be inconsistent

finding a minimal such set (Tian et al., 1998). Here, we show that in the above model family, one can devise a statistical test for consistency of the estimator of the causal effect of x on y, with only the knowledge that the covariates in $\mathcal{W}$ precede the treatment x which in turn precedes the outcome y.

2.3 Statistical Test for Consistency

To test if an arbitrary adjustment set $\mathcal{Z} \subseteq \mathcal{W}$ yields a consistent estimator of the causal effect of x on y, we regress x on $\boldsymbol{z}$, and y on the combination of x and $\boldsymbol{z}$, to obtain the residuals r_x and r_y, respectively. After ensuring that the residual r_x is non-Gaussian, we test for statistical dependence between r_x and r_y.[3] This procedure is formalized in Algorithm 1. In the remainder of this section we give the conditions under which this method is guaranteed to asymptotically correctly identify when a set $\mathcal{Z}$ yields a consistent estimator, and consider issues occurring for finite sample sizes.

We first note that the residuals r_x and r_y are by construction linearly uncorrelated, since, when using OLS, all regressors are uncorrelated with the residual:

$$\begin{aligned}\operatorname{cov}(r_x, r_y) &= E(r_x r_y) = E((x - \boldsymbol{b}^T \boldsymbol{z})\, r_y) \\ &= E(x\, r_y) - \boldsymbol{b}^T E(\boldsymbol{z}\, r_y) = 0,\end{aligned}$$

where both expectations are zero because r_y is derived from a regression on both x and $\boldsymbol{z}$. Because zero linear correlation is equivalent with statistical independence for Gaussian variables, it is easy to see that our method can only yield substantive results for non-Gaussian variables.

[3]Note that it is not sufficient to test dependence of x and r_y, since they may be dependent even though the estimator is consistent (see the Supplementary Material).

Theorem 1. *Given the model of Section 2.2 and using the procedure described in Algorithm 1 with a fixed conditioning set $\mathcal{Z}$, the following statements hold:*

(a) Under the assumption that the disturbance term *of x, denoted by e_x, has a non-Gaussian distribution, if the residuals r_x and r_y are asymptotically mutually independent, then a is a consistent estimator of the true causal effect α.*

(b) Under the assumption that the distribution over the variables in $\mathcal{V}$ is linearly faithful to the generating DAG (Spirtes et al., 2000, p. 47), if the residual *r_x is non-Gaussian, and r_x and r_y are asymptotically mutually independent, then a is a consistent estimator of the true causal effect α.*

(c) If the residuals r_x and r_y are asymptotically statistically dependent, then the set $\mathcal{Z}$ is not admissible.

The proof is given in the appendix. We now discuss the assumptions and implications of the theorem.

From parts (a) or (b), we can conclude that, under the given assumptions, if the residuals are independent then the estimator is consistent. In practice, of course, one can never fully confirm independence, but only hope for not rejecting it at a predefined threshold p_t (as in Algorithm 1). Choosing this threshold is crucial; if it is too high, the type I error rate ("false positives", where a consistent estimate is believed to be inconsistent), which is directly controlled for by this threshold, will be large. On the other hand, if the threshold is set too low, the type II error rate ("false negatives", where a truly inconsistent estimate is judged to be consistent) will typically become large. This implies a trade-off between the number of estimates being judged consistent and the errors being made in these estimates.

The main drawback of part (a) of Theorem 1 is that the assumption of non-Gaussianity of the *disturbance* variable e_x is not testable. To avoid this, part (b) replaces this assumption with the requirement of a non-Gaussian *residual* r_x (which can be tested using any standard test for normality). However, this comes at a cost, as we then need to assume linear faithfulness (Spirtes et al., 2000, p. 47), an assumption that any zero partial correlation in the observed distribution corresponds to d-separation in the generating graph. This assumption is necessary as there exist models with linearly unfaithful parameter values for which the estimator is inconsistent even though r_x is non-Gaussian, and r_x and r_y are independent (see the Supplementary Material). However, these linear unfaithful parameterizations are of measure zero among all parameterizations (Spirtes et al., 2000, p. 66).

Using part (c) of Theorem 1 we can conclude that if the residuals r_x and r_y are dependent the set $\mathcal{Z}$ is not

admissible. While there exist graphs with parameter settings for which the corresponding estimator of the causal effect is nonetheless consistent, such parameter values are special cases of measure zero. Hence, for all practical purposes, a dependence between r_x and r_y indicates that the estimator ought not to be trusted.

3 Heuristics for Search

The procedure given in Algorithm 1 provides a method to test whether adjusting for a given set $\mathcal{Z}$ yields a consistent estimator of the causal effect. However, there are $2^{|\mathcal{W}|}$ subsets of $\mathcal{W}$, so testing all subsets may be computationally intractable. We here introduce simple strategies to search for such a set $\mathcal{Z}$ in quadratic time with respect to $|\mathcal{W}|$, the number of covariates.

We assume that we have a test for statistical dependence that returns a p-value under the null hypothesis that r_x and r_y are independent. In our implementation, we use the Hilbert Schmidt independence criterion (HSIC, Gretton et al., 2008), which is a kernel-based method that is guaranteed to asymptotically detect *any* form of dependence, and, as an alternative, a non-linear correlation test which is computationally more efficient but only detects certain forms of dependencies between the variables.

The first search heuristic is forward selection. Starting from the empty set, $\mathcal{Z} = \emptyset$, for each cardinality $m \geq 1$ of $\mathcal{Z}$ we augment the conditioning set of cardinality $m-1$ by adding a single covariate not yet in the set, and among these possibilities choose the one yielding the highest p-value for independence of r_x and r_y, among those with a sufficiently non-Gaussian residual r_x. If all potential sets $\mathcal{Z}$ yield a Gaussian residual, we select the set yielding the least Gaussian r_x. The procedure returns the set $\mathcal{Z}$ yielding the highest p-value for independence. Pseudocode is given in Algorithm 2.

A similar but alternative strategy is to perform backward elimination. We start from the full covariate set, $\mathcal{Z} = \mathcal{W}$, and for each cardinality $m < |\mathcal{W}|$ of $\mathcal{Z}$ we greedily remove variables from the 'best' set of cardinality $m+1$ in a similar fashion as in the forward selection approach. The pseudocode is similar to Algorithm 2, and hence left to the Supplementary Material.

4 Simulations

We first evaluate the performance of Algorithm 1 by simulating data from 5000 randomly generated models as in Equation (1) (the disturbances are generally non-Gaussian, but are allowed to be close to Gaussian) with 5 observed variables (i.e. $|\mathcal{W}| = 5$) and 3 hidden ones (i.e. $|\mathcal{U}| = 3$), and estimate the causal effect of x on y when adjusting for a subset $\mathcal{Z} \subseteq \mathcal{W}$, drawn uniformly at random from among all possible subsets. Figure 2 shows the p-value of the independence test (non-linear correlation) of the residuals r_x and r_y versus the error in the estimate, scaled according to the standard deviation of the true effects. (Note that only the roughly 3000 subsets yielding a statistically significantly non-Gaussian residual r_x are shown.) Each shaded square in the plot indicates the proportion of estimates having the corresponding error when normalizing each column (i.e. fixed interval of p-values) such that the field with the highest mass is shaded in black. The plots show that, for each of the three sample sizes, the larger the p-value of the independence test, the smaller the average error in the estimated causal effects. For sufficiently large p-values (e.g. greater than 0.4), the errors are (rather) small for all sample sizes. Plots for models with 10 covariates and 5 hidden variables show the same trend, but are a bit more scattered. For Matlab code to reproduce this and all the following results see `www.cs.helsinki.fi/u/entner/ConsistencyTest/`.

Next, we evaluate the performance of the forward selection procedure (Algorithm 2) and the backward elimination algorithm of Section 3, as well as a brute force search, i.e applying Algorithm 1 to all subsets of covariates and picking the one yielding the highest p-value for the independence test of r_x and r_y among those with a non-Gaussian residual r_x. We simulate data from 100 randomly generated models as in Equation (1), and apply these three search strategies with both HSIC and non-linear correlation as independence tests, and estimate the causal effect of x on y using the returned set $\mathcal{Z}$. As suggested by Figure 2, the threshold for the p-value of the independence test in Algorithm 1 is set to $p_t = 0.4$ in all simulations. Furthermore, for the purpose of comparison, we also calculate

Algorithm 2 (Forward Selection)

Let $\mathcal{Z}_0 := \emptyset$ and $m := 0$
If the residual r_x from Algorithm 1 is Gaussian
  set $p_0 :=$ NaN,
Else obtain a p-value p_0 from the independence test of the residuals r_x and r_y from Algorithm 1
Repeat while $m < |\mathcal{W}|$
  $m := m + 1$
  For every set $\mathcal{Z} = \mathcal{Z}_{m-1} \cup \{w\}$, $w \in \mathcal{W} \setminus \mathcal{Z}_{m-1}$, test whether r_x is Gaussian
  For every such $\mathcal{Z}$ with non-Gaussian r_x get a p-value $p_{\mathcal{Z}}$ from the independence test of r_x and r_y
  Set $p_m := \max\{p_{\mathcal{Z}}\}$ and let $\mathcal{Z}_m$ be the corresponding set $\mathcal{Z}$
  If all r_x were Gaussian, set $p_m :=$ NaN and let $\mathcal{Z}_m$ be the set $\mathcal{Z}$ yielding the least Gaussian r_x
Return $\max_m\{p_m\}$ and the corresponding set $\mathcal{Z}_m$

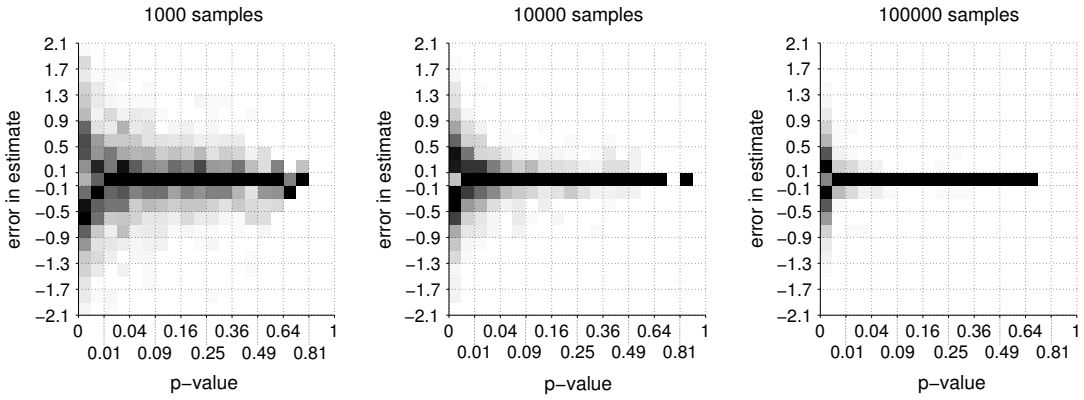

Figure 2: For 5000 models with $|\mathcal{W}| = 5$ covariates and $|\mathcal{U}| = 3$ hidden variables we estimate the causal effect of x on y when adjusting for a randomly selected subset $\mathcal{Z}$ of the covariates (only estimates were Gaussianity of the residual r_x was rejected are displayed). The p-value of the independence test of the residuals (note the non-uniform scale) is plotted versus the standardized error in the estimate, when using non-linear correlation as independence test and 1000, 10000 and 100000 samples (from left to right). For each interval of p-values (columns in the plot) the counts are normalized such that the square with the highest count is shaded in black.

the estimate obtained by including none ($\mathcal{Z} = \emptyset$) and all ($\mathcal{Z} = \mathcal{W}$) of the covariates in the regression. The latter is a common choice in practice when it is known that the covariates precede x (no 'selection bias' can be introduced by conditioning), and the relationships are linear (little loss in the accuracy of the estimates). We also compare with LiNGAM (Shimizu et al., 2006), a method which learns the whole graph structure under the assumption that all relevant variables are observed. An extension of this method (lvLiNGAM, Hoyer et al., 2008), which allows for latent variables, is not applicable due to its computational complexity. We measure the correctness of the procedures by calculating the root mean squared error of those estimates deemed consistent (normalized by the standard deviation of the true effects). The results are shown in Figure 3. (Note that for the models with $|\mathcal{W}| = 10$ and $|\mathcal{U}| = 5$ the brute force approach with HSIC was computationally too expensive with 10000 samples.) For the smallest sample size of 100 the search heuristics using the statistical test introduced in Algorithm 1 do not seem to bring any improvements over simply including all or none of the covariates in the conditioning set. However, with growing sample size our novel methods generally outperform the control methods. Surprisingly, in the case with 5 covariates and 3 hidden variables, including none of the covariates performs slightly better than when including all. This is contrary to Greenland's (2003) argument that the bias in estimate introduced by activating a path when conditioning on some variable is typically less than the bias eliminated by deactivating a path by conditioning. The choice of the independence tests does not seem to affect the error much. However, it does affect the number of estimates deemed consistent. On average, using HSIC results in about twice as many estimates compared with using the non-linear correlation test. With HSIC the proportion of estimates judged consistent in the smaller models ($|\mathcal{W}| = 5$, $|\mathcal{U}| = 3$) varies from about 50% for the 100 sample case to about 20% for the 10000 sample case. For the larger models ($|\mathcal{W}| = 10$, $|\mathcal{U}| = 5$) the corresponding numbers are 75% and 35%. In total, in about 25% of the cases there truly existed an admissible set, but confounding may have been weak such that some estimates were only slightly distorted.

5 Car Mileage Data

We apply the brute force search using Algorithm 1 to a data set[4] containing 82 observations of the following variables, arranged in a plausible causal order: cab space (in cubic feet), vehicle weight (in 100 lb), engine horsepower, top speed (in mph), and average miles per gallon. All variables are log-transformed to better fulfill the linearity assumption. We assign $\mathcal{W}$, x and y in any possible combination consistent with the above order, and obtain the following results. The method judges the causal effect of horse power on the average miles per gallon to be consistent when $\mathcal{Z} = \emptyset$ (with a p-value of 0.225 for the independence test of the residuals) and $\mathcal{Z} = \{\text{cab space}\}$ (p-value of 0.575), and estimates in both cases a strength of about -0.66. Furthermore, the effect of top speed on miles per gallon is deemed consistent and of strength -2.1 when conditioning on $\mathcal{Z} = \emptyset$ (p-value of 0.175) or $\mathcal{Z} = \{\text{cab space}\}$ (p-value of 0.425). Both these relations seem reasonable as does (at least) the sign of the estimated effect. All other effects which lead to non-Gaussian residu-

[4]The data are available at `http://www.stat.cmu.edu/~larry/all-of-statistics/=data/carmileage.dat`.

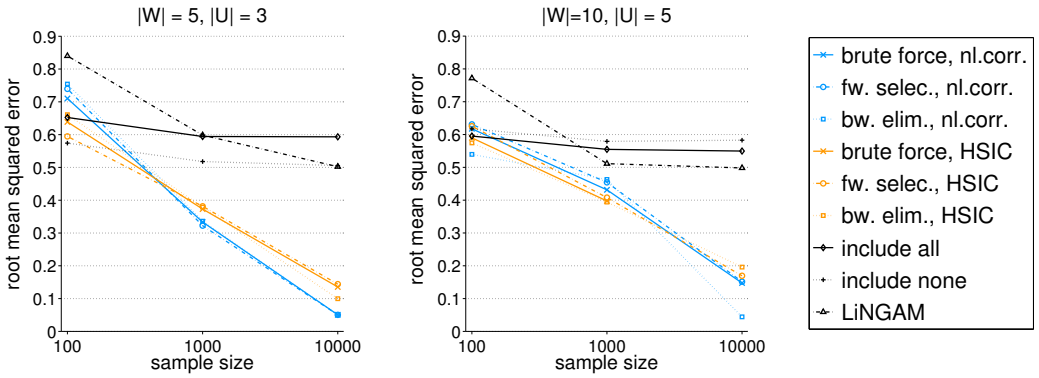

Figure 3: The sample size (100, 1000 and 10000) is plotted versus the root mean squared error of the estimates of the causal effect of x on y, which were deemed consistent, from 100 models with 5 covariates and 3 hidden variables (left figure) as well as with 10 covariates and 5 hidden variables (right figure) using different algorithms (brute force approach, forward selection and backward elimination introduced in Section 3, including all and none of the variables in the conditioning set, and LiNGAM (Shimizu et al., 2006)) and two independence tests (nonlinear correlation, blue lines, and HSIC, orange lines).

als r_x are deemed inconsistent, some of which would be meaningful as well, such as a positive causal effect of horse power on top speed. This may be due to an error in the magnitude of the effect, or simply a wrong decision by the algorithm. Furthermore, the method correctly rejects the only estimate having the intuitively wrong sign (the weight of the car having a positive effect on top speed), and a number of estimates of strength zero, such as cab space on top speed.

6 Conclusions

When seeking to derive causal conclusions from passive observational data, one of the main problems is the possibility of an inconsistent estimator due to confounding by observed or unobserved covariates. In general, detailed domain knowledge of the causal structure is needed to select a suitable covariate set which yields consistent estimators of the desired causal effects when adjusted for. We have shown that, in the restricted space of linear models, when the data are non-Gaussian, one can use statistical tests to judge whether adjusting for a given variable set is appropriate, and such tests can be used to search for a set likely to yield a consistent estimator.

Acknowledgments

DE and POH were supported by the Academy of Finland project #1125272.

Appendix: Proof of Theorem 1

For reasons of space, we here give a relatively condensed proof of Theorem 1. A more detailed version can be found in the Supplementary Material. In the proof we will make heavy use of the 'reduced form' representation of the model in Equation (1), given by

$$\boldsymbol{v} = \mathbf{A}\boldsymbol{e} \tag{4}$$

with $\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}$ containing the total effects of the disturbances $\boldsymbol{e}$ on the variables $\boldsymbol{v}$. That is, each variable $v \in \mathcal{V}$ can be written as a linear combination of the disturbance terms: $v = \mathbf{A}_v\boldsymbol{e}$ with $\mathbf{A}_v$ being the corresponding row of matrix $\mathbf{A}$. We further use the terminology of *active* and *blocked* paths as defined in Pearl (2009a), and restated in the Supplementary Material. We also rely on a property about dependence and independence of two sums of independent random variables, known as the Darmois-Skitovitch Theorem:

Darmois-Skitovitch Theorem (Darmois, 1953; Skitovitch, 1953). Let $e_1, \ldots, e_n$ be independent random variables ($n \geq 2$), $s_1 = \beta_1 e_1 + \ldots + \beta_n e_n$ and $s_2 = \gamma_1 e_1 + \ldots + \gamma_n e_n$ with constants β_i, γ_i, $i = 1, \ldots, n$. If s_1 and s_2 are independent, then those e_j which influence both sums s_1 and s_2 (i.e. $\beta_j\gamma_j \neq 0$) are Gaussian.

Proof of Theorem 1 (a). We show that if the estimator is inconsistent then the residuals are dependent. Let $\boldsymbol{v} = (v_1, \ldots, v_{n-2}, x, y)^T$ and $\boldsymbol{e} = (e_{v_1}, \ldots, e_{v_{n-2}}, e_x, e_y)^T$ such that the matrix $\mathbf{A}$ (Equation (4)) is lower triangular with a unit diagonal (possible by the acyclicity and partial temporal ordering assumption). It follows that $x = \mathbf{A}_x\boldsymbol{e}$ with $\mathbf{A}_x = (A_{x,1}, \ldots, A_{x,n-2}, 1, 0)$, and $y = \mathbf{A}_y\boldsymbol{e}$, with $\mathbf{A}_y = (A_{y,1}, \ldots, A_{y,n-2}, \alpha, 1)$, with $A_{x,i}, A_{y,i} \in \mathbb{R}$, $i = 1, \ldots, n-2$, and α, the coefficient for e_x in y, being the true causal effect of x on y. Furthermore, for all other variables v_i, $i = 1, \ldots, n-2$, the representation

$v_i = \mathbf{A}_{v_i}\boldsymbol{e}$ is such that the coefficients for e_x and e_y are zero (because of the lower triangularity of $\mathbf{A}$). Writing the residuals of the regressions in terms of the disturbances yields $r_x = x - \boldsymbol{b}^T\boldsymbol{z} = (\mathbf{A}_x - \sum_{z\in\mathcal{Z}} b_z\mathbf{A}_z)\,\boldsymbol{e} = (\ldots, 1, 0)\,\boldsymbol{e}$ and $r_y = y - ax - \boldsymbol{c}^T\boldsymbol{z} = (\mathbf{A}_y - a\mathbf{A}_x - \sum_{z\in\mathcal{Z}} c_z\mathbf{A}_z)\,\boldsymbol{e} = (\ldots, \alpha - a, 1)\,\boldsymbol{e}$ (where the dots indicate the entries of the disturbances other than e_x and e_y). Given the premise that the estimator is inconsistent (i.e. $(a-\alpha) \not\xrightarrow{P} 0$) the disturbance e_x has a non-vanishing coefficient in the representation of *both* residuals r_x and r_y. By the assumed non-Gaussianity of e_x the Darmois-Skitovitch theorem ensures that r_x and r_y are asymptotically statistically dependent. □

For the proofs of Theorem 1 (b) and (c) we make use of the following Lemma.

Lemma 1. *We are given a set of variables $\mathcal{V} = v \cup \mathcal{V}'$, where v is a single variable and $\mathcal{V}'$ a non-empty set of variables not including v, following the model in Equation* (1)*, and we assume that the distribution over these variables is linearly faithful to the generating DAG. Regressing v on a set $\mathcal{Z}' \subseteq \mathcal{V}'$ not containing any descendants of v yields*

$$r_v = v - \sum_{z\in\mathcal{Z}'} \hat{c}_z z = \mathbf{A}_v\boldsymbol{e} - \sum_{z\in\mathcal{Z}'} \hat{c}_z\mathbf{A}_z\boldsymbol{e} = \boldsymbol{d}_v\boldsymbol{e}$$
$$= (d_{v,1}, \ldots, d_{v,w}, \ldots, d_{v,n})(e_1, \ldots, e_w, \ldots, e_n)^T$$

with $\mathbf{A}_v$ and $\mathbf{A}_z$, $z \in \mathcal{Z}'$, the corresponding rows of the matrix $\mathbf{A}$ in Equation (4) *and $\boldsymbol{d}_v = \mathbf{A}_v - \sum_{z\in\mathcal{Z}'} \hat{c}_z\mathbf{A}_z$. When estimating the regression using OLS, for $w \in \mathcal{V}'$ holds that $d_{v,w} \not\xrightarrow{P} 0$ (with $d_{v,w}$ the coefficient of e_w in $\boldsymbol{d}_v$) if and only if*

1. *for $w \in \mathcal{Z}'$ there is an active back-door path (not blocked by $\mathcal{Z}' \setminus \{w\}$) from w to v pointing into v,*
2. *for $w \in \mathcal{V}' \setminus \mathcal{Z}'$ there is (i) a directed active path from w to v (not blocked by $\mathcal{Z}'$) or (ii) a directed active path from w to some $z \in \mathcal{Z}'$ (not blocked by $\mathcal{Z}' \setminus \{z\}$) for which there is an active back-door path to v (not blocked by $\mathcal{Z}' \setminus \{z\}$) pointing into v.*

Proof of Lemma 1. Note that points 1 and 2, respectively, are equivalent to e_w being d-connected to v given $\mathcal{Z}'$, by a path pointing into v, which follows straight from the definition of d-separation. Furthermore, any active back-door path is pointing into v, since any other back-door path includes at least one collider at some descendant of v, and such variables are not in the conditioning set (by assumption).

We first show that if point 1 or 2 holds, then $d_{v,w} \not\xrightarrow{P} 0$. If $w \in \mathcal{Z}'$ with $w = e_w$ point 1 cannot hold. For any other disturbance variable e_w, if point 1 or 2 is fulfilled, we know that e_w is not d-separated from v given $\mathcal{Z}'$, which together with the faithfulness assumption implies that the partial correlation of e_w and v given $\mathcal{Z}'$ is non-vanishing. Hence, in the regression $v = \sum_{z\in\mathcal{Z}'} \hat{c}_z z + r_v$, the coefficient for e_w in the representation of r_v must be non-vanishing, i.e. $d_{v,w} \not\xrightarrow{P} 0$.

Next we show that if points 1 and 2 are violated, then $d_{v,w} \xrightarrow{P} 0$. First, for any $w \in \mathcal{Z}'$ which has no parents, i.e. $w = e_w$, point 1 cannot hold. In this case, since in the regression of v on the variables in $\mathcal{Z}'$ the regressors are uncorrelated with the residual r_v, we obtain $0 = cov(w, r_v) = cov(e_w, r_v) = E(e_w\,\boldsymbol{d}_v\boldsymbol{e}) = \sum_{i=1}^n d_{v,i}\,E(e_w\,e_i)$ and $E(e_w\,e_i) \xrightarrow{P} 0$ for $w \neq i$ (since all e_i are independent). Thus, $\sum_{i=1}^n d_{v,i}\,E(e_w\,e_i) \xrightarrow{P} d_{v,w}V(e_w) \xrightarrow{P} 0$, implying that $d_{v,w} \xrightarrow{P} 0$. For any other disturbance variable e_w, $w \in \mathcal{V}'$, we know that the negation of points 1 and 2 imply that e_w is d-separated from v given $\mathcal{Z}'$, which in the linear model family implies that the partial correlation of e_w and v given $\mathcal{Z}'$ is vanishing. Thus, in the regression $v = \sum_{z\in\mathcal{Z}'} \tilde{c}_z z + be_w + \tilde{r}_v$ the regression coefficient $b \xrightarrow{P} 0$, and hence for the regression $v = \sum_{z\in\mathcal{Z}'} \hat{c}_z z + r_v$, we get that for all $z \in \mathcal{Z}'$ the coefficients $\tilde{c}_z$ and $\hat{c}_z$ converge to the same value and hence also $\tilde{r}_v$ and r_v. Because $0 = cov(e_w, \tilde{r}_v) = cov(e_w, r_v)$ we can, as above, conclude that $d_{v,w} \xrightarrow{P} 0$. Note that the linear faithfulness assumption was *not* used to prove this direction. □

Proof of Theorem 1 (b). We show that an inconsistent estimator implies dependent residuals. Since r_x is non-Gaussian, when expressing r_x in terms of the disturbances ($r_x = \boldsymbol{d}_x\boldsymbol{e}$) the coefficient of at least one non-Gaussian disturbance e_w has to be non-vanishing. By Lemma 1 follows that there exists an active path from w to x (not blocked by $\mathcal{Z}$) pointing into x of a type as in point 1 or 2. The inconsistent estimator of the effect from x on y implies that there is an active back-door path from x to y (again not blocked by $\mathcal{Z}$). Connecting these two active paths yields an active path from w to y when conditioning on $\{x\} \cup \mathcal{Z}$, which has the form of point 1 or 2 of Lemma 1. Thus, by the lemma the effect of e_w on y is also non-vanishing (here the linear faithfulness assumption is needed). From the non-Gaussianity of e_w and the Darmois-Skitovitch theorem then follows that r_x and r_y are dependent. □

Proof of Theorem 1 (c). Expressing the residuals in terms of the disturbances we obtain $r_x = \boldsymbol{d}_x\boldsymbol{e}$ and $r_y = \boldsymbol{d}_y\boldsymbol{e}$. Since r_x and r_y are dependent there exists a $w \in \{x\} \cup \mathcal{W} \cup \mathcal{U}$ whose error term e_w has a non-vanishing coefficient in this representation for both r_x and r_y. By Lemma 1 we thus know that there exists some kind of active paths (as in point 1 or 2 of the lemma) from w to x (conditional on $\mathcal{Z}$) and from w to y (conditional on $\{x\} \cup \mathcal{Z}$), and concatenating these paths yields an active back-door path from x to y. □

References

Darmois, G. (1953). Analyse générale des liaisons stochastiques. *Revue de l'Institut International de Statistique*, 21:2–8.

Greenland, S. (2003). Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology*, 14(3):300–306.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*.

Hoyer, P. O., Shimizu, S., Kerminen, A. J., and Palviainen, M. (2008). Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49:362–378.

Pearl, J. (2009a). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition.

Pearl, J. (2009b). Myth, confusion, and science in causal analysis. Technical Report R-348, University of California, Los Angeles, CA.

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

Skitovitch, W. (1953). On a property of the normal distribution. *Doklady Akademii Nauk SSSR*, 89:217–219.

Spirtes, P. (2000). The limits of causal discovery from observational data. *American Economic Association, Boston, MA*.

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

Spirtes, P., Richardson, T., Meek, C., Scheines, R., and Glymour, C. (1998). Using path diagrams as a structural equation modeling tool. *Sociological Methods & Research*, 27(2):182–225.

Tian, J., Paz, A., and Pearl, J. (1998). Finding minimal separating sets. Technical Report R-254, Computer Science Department, University of California, Los Angeles, CA.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer.