Author’s Accepted Manuscript version of the book chapter:

Rosenström, T & García Velázquez, R 2020, Distribution-Based Causal Inference: A Review and Practical Guidance for Epidemiologists. in W Wiedermann, D Kim, EA Sungur & A von Eye (eds), Direction Dependence in Statistical Modeling: Methods of Analysis. John Wiley & Sons Ltd. , Hoboken, NJ, pp. 267-294.

Distribution-based causal inference: a review and practical guidance for epidemiologists

Tom Rosenström,[1,2,*] Regina García-Velázquez[2]

[1]Department of Mental Disorders, Norwegian Institute of Public Health, Oslo Norway

[2]Department of Psychology and Logopedics, University of Helsinki, Helsinki, Finland

*Correspondence to Tom Rosenström (tom.rosenstrom@helsinki.fi)

17th Mar 2018

An invited book chapter to an edited volume on causal discovery methods (Eds. Wolfgang Wiedermann, Alexander von Eye)

1. Introduction

During training, many empirical researchers have likely heard phrases along the lines of “causality cannot be inferred from cross-sectional data”, wherein “inferring causality” refers to distinguishing a cause from its consequence. This statement is actually wrong. It probably reflects the huge role that the correlation coefficient has played in empirical research. One cannot infer causation from a product-moment, or Pearson's, correlation coefficient, without supplementing it with other knowledge. More recent research, however, has derived several other statistics that are able to infer direction of causation in a cross-sectional setting (Wiedermann and von Eye 2016). So far these have seen relatively little use in epidemiologic research, even though inferring causation (i.e., etiology) is a central topic in epidemiology. This may reflect, in part, a healthy streak of conservatism at the face of novel methods. Eventually, however, too much conservatism may frustrate scientific progress, because many questions of epidemiology do not lend themselves well for experimentation. Neglecting possibilities available for observational data is a luxury we cannot always afford.

This chapter aims to familiarize researchers in epidemiology and related fields with the topic of distribution-based causal inference methods, and to discuss how to build trust in the results from such methods. We review, replicate, and extend some of the few studies that have used the methods in real-world epidemiological issues where the ground truth was not known *a priori* (Rosenström et al. 2012; Helajärvi et al. 2014). In particular, we concentrate on simple cases of distribution-based causal inference applied to survey data and linear models. These types of data and models permeate much research in epidemiology, including our example case of research on causality between sleep problems and other depressive symptoms, introduced below.

2. Direction of dependence in linear regression

While correlation does not imply causation, Dodge and Rousson (2000) were perhaps first to derive "other expressions of the correlation coefficient" that do make it possible to infer which among two *skewed* variables, X and Y, causes the other, or in other words, which variable is the proverbial 'cart' and which variable the 'horse'. Specifically, it is possible to distinguish between two models, one with Y as the dependent variable and the other with X as the dependent, or 'causally descendent', variable. This amounts to distinguishing between two systems of equations:

$$\begin{cases} Y = \mu_Y + \beta_x X + \epsilon_Y \\ X = \mu_X + \epsilon_X \end{cases} \tag{1a}$$

and

$$\begin{cases} Y = \mu_Y + \epsilon_Y \\ X = \mu_X + \beta_y Y + \epsilon_X\text{'} \end{cases} \tag{1b}$$

where β, μ_Y, and μ_X are constants (i.e., regression coefficient, or slope, and means, or intercepts) and the "residual" variables ϵ_X and ϵ_Y are independent of each other and also independent of the predictor. Dodge and Rousson observed that, under Eq. 1, a following relation holds asymptotically for skewed variables:

$$\begin{cases} X \text{ causes } Y \text{ if } T(X,Y) > 0 \\ Y \text{ causes } X \text{ if } T(X,Y) < 0\text{'} \end{cases}$$

where the test statistic $T(X,Y) := M(X,Y)_{21} - M(X,Y)_{12}$ is based on sample versions of the difference in squared and centralized third cumulants, defined as $M(X,Y)_{ij} = \{E[(X-\mu_X)^i(Y-\mu_Y)^j]/({\sigma_X}^i{\sigma_Y}^j)\}^2$. Here E is the expectation operator and σ_X refers to variance of the variable X.

As shown by Dodge and Rousson (2000), as well as others with slightly different formulations (Hyvärinen and Smith 2013), sufficient conditions for causal inference are that one and only one of the two linear models hold (1a or 1b) and that at least one of the variables has a skewed distribution. We do not reproduce the analytic proofs, but instead offer the following intuition and then proceed to more general estimators.

Many frequently studied variables in epidemiology have “right-skewed” population distributions (e.g., alcoholic drinks per week, number of children, and depressive symptoms). Right-skewness means that a variable gets ‘small’ values frequently and ‘high’ values only rarely in comparison to a normal (a.k.a., Gaussian) distribution (a case of left-skewed variable can be transformed to right-skewed without loss of information by multiplying by minus one). Then if a variable Y is a sum of two independent right-skewed variables, βX and ϵ_Y, it gets high values whenever either one of the two variables gets a high value, which is necessarily more often than for βX and ϵ_Y on average.[1] Thus, the dependent (causally descendent) variable is less skewed than the independent (causally antecedent) variable and they show a characteristic pattern in bivariate scatter plots (non-symmetry over permutation of axes; cf. Figure 1).[2] A similar signal is absent when the data is generated from the same linear model operating on two normally distributed variables (lower-right panel of Figure 1). Any weighted sum of normally distributed (Gaussian) variables is also a Gaussian variable, implying no skewness or excess kurtosis, and no

[1] Probability distribution of a sum of two independent random values is a convolution of the two original distributions (Klenke, 2008). “Convolution” operation is a concept of mathematical (functional) analysis, and formalizes a sort of “smearing” of two distributions to a new one.

[2] If ϵ_Y is normally distributed and the causal antecedent X is not, it can be shown analytically that $\gamma_Y = \gamma_X Cor(X,Y)^3 < \gamma_X$, where γ_X is “skewness coefficient” of X (Dodge and Rousson, 2000).

information beyond correlations in any of the higher moments that characterize bivariate probability distributions (Hyvärinen, Karhunen, and Oja 2001; Klenke 2008).

[FIGURE 1 HERE]

However, it has later turned out that the measures $M(X,Y)_{21}$ and $M(X,Y)_{12}$ can also be based on kurtosis instead of skewness of the distribution, and even more generally, on *any* type of deviation from normal distribution (Dodge and Yadegari 2010; Hyvärinen and Smith 2013; Shimizu et al. 2006; Wiedermann 2018). According to the Central Limit Theorem (CLT) of probability theory, sums of *independent* random variables of almost any probability distribution tend towards a normal distribution (Hyvärinen, Karhunen, and Oja 2001; Klenke 2008). More precisely, this is Lindeberg's version of CLT, according to the Finnish mathematician Jarl Waldemar Lindeberg (1876-1932), which only requires that the random variables have finite variance and that their sequence satisfies a certain (i.e., "Lindeberg's") regularity condition (e.g., Klenke 2008). There is necessarily more summation in the causal descendent than in the antecedent in linear model, because the descendent is a weighted sum of the antecedent plus the residual variable. Thus, one can infer the causal antecedent as being the variable that leads to least Gaussian distributions for antecedent and the residual (one of which can even be Gaussian; Hyvärinen and Smith 2013). Unless, of course, both already are Gaussian, or the linear models does not apply. That is, necessary conditions for pairwise distribution-based causal inference are that (*i*) at least one of the variables must have non-Gaussian distribution, that (*ii*) the linear model applies (lest the causal descendent is some other function than the weighted sum required by CLT), and that (*iii*) the residual variable must be independent of the causal antecedent (again, required to apply CLT to the components βX and ϵ_Y of the descendent variable Y). The assumption *iii* is already present in *ii*, but worth highlighting separately for its important role.

When explicitly stated, the assumption *iii* also suggests an algorithm to evaluate direction of causation between two variables. Assuming the variables are non-Gaussian and one of the above two linear systems of equations holds, 1a or 1b (i.e., assuming conditions *i* and *ii*), then a result known as Darmois-Skitovich theorem implies that the causally antecedent variable is the variable that is independent of its residual when regressed onto the other variable (Shimizu et al. 2011). That is, we can define the above $M(X,Y)_{12}$ to be $\hat{MI}\ (X,\epsilon_Y)$, an estimate of mutual information between X and residual of Y when regressed on X, or $\epsilon_Y = Y - \frac{Cov(X,Y)}{Var(X)}X$, where $Cov(\cdot,\cdot)$ and $Var(\cdot)$ are covariance and variance operators, respectively (Shimizu et al. 2011). Analogously, $M(X,Y)_{21}$ is defined to be $\hat{MI}\ (Y,\epsilon_X)$. Then, the statistic $T(X,Y)$ from above will become an estimate of causal direction based on non-Gaussianity and least mutual information between a predictor and its residual. While there are many other estimators for distribution-based causal inference (Hyvärinen and Smith 2013), here we will concentrate on this "DirectLiNGAM" estimator, which uses a kernel-based estimate of mutual information and has been found useful in previous empirical studies and simulations (Shimizu et al. 2011; Rosenström et al. 2012; Helajärvi et al. 2014). The name refers to a "direct" algorithm for estimating Linear Non-Gaussian Acyclic Models (i.e., LiNGAMs) as opposed to the earlier iterative algorithm (Shimizu et al. 2006; Shimizu et al. 2011).

Mutual information is a measure for degree of dependence between two random variables. It tells how much information entropy in one variable can be obtained through the other variable. Theoretically, mutual information is defined as an expected difference between true bivariate entropy and entropy of an ‘independence distribution’ (i.e., product of marginal distributions): $E[log(p(X,Y)) - log(p(X)p(Y))]$, where $p(\cdot,\cdot)$ and $p(\cdot)$ are the bivariate and marginal probability density functions, respectively, and E is the expectation operation with respect to the bivariate distribution. Thus, mutual information is quantified as departure from bivariate independence. The equation is noteworthy, because often one is interested in what happens in terms of the (population) distribution with respect to which the expectation is taken, rather than what happens for each and every observation per se. In other words, we have no reason to expect that a deviant minority with opposite causal direction would ruin our inferences about dominant population-level direction of causation. This is an important advantage, for example, in our target research problem on causal direction between sleep problems and other depressive symptoms, as there likely are sub-populations that exhibit rather different causal processes in comparison to most cases of depression (e.g., brain trauma patients).

In what follows, we first (in section 3) give a review of previous empirical work and (section 4) introduce a practical research problem in epidemiology, which both represents a novel replication effort and is used as an example data throughout the rest of the chapter. Then, (section 5) we discuss strategies to evaluate the assumptions necessary for distribution-based causal inference, (section 6) analyze the example data and (section 7 & 8) discuss strategies to study robustness and statistical power of distribution-based causal inferences, and (section 9) strategies for, as well as importance of, “triangulation” with multiple methods that are non-overlapping in their assumptions. Finally, in section 10, we conclude the chapter with comments on both present content and other causal-inference methodologies.

3. Previous epidemiologic applications of distribution-based causal inference

Psychiatric epidemiology deals with complex disorders whose etiology is not yet understood to large extent. The classic ways of thinking in psychiatric epidemiology have been challenged by the recent, influential network theory (Cramer et al. 2010; Borsboom 2017). Whereas the traditional diagnostic practice perceives symptoms as passive reflections of an underlying psychiatric root cause, the network theory recognizes the symptoms as causally active entities that can 'cause' each other and thereby give rise to syndromes. A classic example in network theory has suggested that sleep problems can gradually give rise to a full-blown depressive syndrome, for example, through inducing fatigue and concentration problems, which then lead to performance issues in daily life, and ultimately to all other depressive symptoms (Cramer et al. 2010; Borsboom 2017). This is a completely hypothetical example, however, and the direction of causation between sleep and the average of other depressive symptoms (a proxy of the syndrome) remains an open question.

Rosenström et al. (2012) discuss about the multiple difficulties involving the study of causation between human sleep characteristics and depressive disorders. In such topics, it is nearly impossible to design a definitive study. It is not ethically acceptable to experimentally induce depression, because of the involved human suffering and the high risk of suicide in major depressive disorder. Due to possible lagged effects, temporal order of events may not directly inform about the causal order (the horse might as well push the cart as pull it). Furthermore, sleep problems and other depressive symptoms are relatively common phenomena in the population, and it is probably possible to find strong individual cases to argue the causation both ways—as epidemiologists, we were interested in the direction that dominates on average in the population. Rosenström et al. (2012) used the above-introduced distribution-based causality statistic as a research tool fit to assessing population averages. As so often occurs in practice, however, it did not provide a unique answer: sleep problems was estimated to cause depression in most cases, but also opposite findings were obtained. Here, we return to the topic in our running example, using yet another classic real-world dataset, as well as computer simulations.

Another previous application of distribution-based causal inference in epidemiology was a study on direction of causation between average television viewing time and obesity (Helajärvi et al. 2014). Overweight, obesity, insulin resistance, and diabetes have been major public health concerns in the Western world lately. Also the time spent in "sedentary behaviors" has increased in comparison to past times of manual labor. Especially watching television has been recognized as an activity with uncharacteristically low waking-time metabolic rates in evolutionary terms. The relative time spent watching television has been suggested to cause weight gain, but the proposition has been challenged by a reverse causation hypothesis, according to which obese and overweight people may find physical activity less appealing than lean people and therefore spend more time in a substitute activity of watching television. Helajärvi et al. (2014) used distribution-based causal inference to show that high television watching times are more likely to cause changes in weight than the other way around.

That distribution-based causal inference is possible to begin with often comes as a surprise for epidemiology researchers who are well-aware of the fact that an analogous correlation-based causal inference is not possible. Therefore, 'toy examples' using real datasets where the direction of causation is obvious have been necessary to demonstrate that the method works. In this category, Shimizu et al. (2011) have shown that the method correctly infers father's education and occupation as causes of his son's education and occupation rather than other way around. Similarly, Rosenström et al. (2012) showed that the method suggests parents' socioeconomic status as a cause of their children's socioeconomic status rather than the other way around. While computer simulations and mathematical analyses are the primary tools to study how well statistical methods function, toy examples with real data are also indispensable in building trust on 'black-box' methods that reveal very little about the true mechanism behind the inferred causal effect. The target method of this chapter has so far withstood the test. In the later sections, we discuss about another method that has not always withstood similar tests.

4. A running example: Re-visiting the case of sleep problems and depression

Whereas Rosenström et al. (2012) studied epidemiologic, Finland-based "Young Finns" and USA-based "Wisconsin Longitudinal Study" datasets, here we use data from the Swedish Adoption/Twin Study on Aging (SATSA) that were available to us through the Inter-University Consortium on Political and Social Research (Pedersen 2015). Specifically, these data include 1439 observations both on a depression score and on average hours slept per night (1326 complete and 1325 valid observations; one person reported no sleep at all). Sleeping hours was a self-reported quantity, whereas the depression score we used was an average of non-sleep-related depressive symptoms assessed by the Center for Epidemiologic Studies Depression (CES-D) scale (Radloff 1977). The symptom statuses were reported as amount of symptom presence during the past week (0 = "never/almost never", 1 = "Rather seldom/never", 2 = "Quite often", and 3 = “Always/almost always”). Figure 2 illustrates the data. Altogether 598 of the subjects were men (41.56%). Average age of participants was 63.42 years at the time of data collection in 1990 (s.d. 13.01 years, range from 32 to 95). SATSA is a twin and adoption study and these participants constitute 134 pairs of monozygotic (i.e., “identical”) twins reared apart, 184 monozygotic twin pairs reared together, 345 dizygotic (“fraternal”) twin pairs reared apart, and 286 dizygotic twin pairs reared together. This feature of the data will be useful in the causal triangulation section of this chapter.

[FIGURE 2 HERE]

5. Evaluating the assumptions in practical work

The assumptions of the DirectLiNGAM approach to distribution-based causal inference (in a bivariate case) are:

i) Linear model: one and only one of the two systems of equations in Eq. 1 hold.

ii) Non-Gaussian continuous variable: The variables have a continuous distribution and at least one of the two independent terms (predictor and residual) has some other distribution than the Normal distribution.

iii) Independence: the predictor (causally antecedent) variable in Eq. 1 is statistically independent of the residual.

The following sub-sections discuss strategies for evaluating whether these necessary preconditions of distribution-based causality hold in practice. Applications of the strategies are provided for the example case of sleep problems and depression.

5.1. Testing linearity

It can be very difficult to know whether a given dataset reflects an essentially linear data-generating process. In practice, one is typically willing to accept a linear approximation if both visual inspection and polynomial regression coefficients thus indicate. That is the approach we take here, as well.

To illustrate using our running example, we found a significant regression coefficient for the quadratic effect as well as the linear, when regressing the depression variable of SATSA data on the standardized (z-score transformed to 0 mean and variance of 1) sleep-hours variable and its square ($\beta_{quadratic} = 0.062$, p < 0.001), not just for the linear slope ($\beta_{linear} =$ -0.146, p < 0.001). Standardization of variables is an important part of the polynomial regression method, because polynomials of unstandardized variables can be close to multicollinearity. Usually, it is a good idea to examine some of the higher-order polynomials as well (e.g., cubic transformation of X, or X^3), but typically these explain progressively less variance in noisy epidemiologic data compared to the lower-order polynomials (i.e., 1, X, and X^2).

To understand the nonlinearity we detected, we examined the panel “c” of the figure that illustrates the SATSA data from our running example (Figure 2). It shows a scatterplot of the sleep-hours and depression variables, revealing that both much less or much more sleep in comparison to the population average hours appears to be associated with high values of CES-D (i.e., with depression). This is an observation we can readily understand. Typically, researchers consider both insomnia (too little sleep) and hypersomnia (too much sleep) as a symptom of depression, and in fact, diagnostic definitions of depression do not differentiate between insomnia and hypersomnia. Therefore, we considered absolute deviation from the population-average hours slept per night as our new, continuous "sleep problems" variable in the analyses that follow (cf. panel d in Figure 2). With this transformation, we both understood what natural phenomenon our transformed variable stands for and were able to remove obvious nonlinearities in the data ($\beta_{quadratic} = 0.012$, p = 0.231 for the new sleep deviation variable).

It is generally not advisable to use arbitrary, uninterpretable transformations to linearize data before applying methods for distribution-based causal inference. This is because nonlinear transformations alter substantive meaning of variables, as well as their distributions. One might lose track of what phenomenon is being modeled, and at the same time, manipulate the inferred direction of causation. Thus, some substantive understanding is desirable prior to application of variable transformations in this context. However, the transformation need not be quite as straightforward as in our running example here. For example, Rosenström et al. (2012) discuss more advanced ways to re-interpret nonlinear psychometric data using Item Response Theory models. Similarly, one cannot remove seemingly 'outlier' observations to make the data more linear prior to application of distribution-based causal inference, because that alters the distributions in question towards something else than the distributions reflecting the natural data-generating process under investigation (one should of course remove very clear recording errors, etc.; for example, we verified that an individual who appeared to report zero hours of sleep throughout year had no consequences for our analyses). In our running example, the sleep measure derived as absolute deviation from population-mean hours slept per night is a substantively meaningful variable in that it quantifies both hyper- and insomnia, and it fulfills the assumption of linearity with respect to the depression score. Taking absolute values increased skewness, however. Therefore, we also performed sensitivity analyses conducting DirectLiNGAM on separated datasets with hours slept below the mean of 7.21 (N= 620) and above it (N=706).

As a cautionary note, even if one can statistically assess whether Y *could* be nonlinear in X or *vice versa*, undetected complex relationships between the variables typically cannot be fully ruled out by means of empirical analysis. Whether the assumption *i* (and *iii*) is reasonable must be assessed also in light of substantive understanding. To illustrate, the first panel of Figure 3 shows a sample of apparently stochastic data which can be modeled using a linear model with a statistically significant slope and which shows no quadratic effect, but which has, in fact, been derived from a deterministic nonlinear system.

[FIGURE 3 HERE]

5.2. Testing non-Normality

Non-normality can be verified by rejecting a hypothesis of normal distribution using, for example, Lilliefors' test, which is an extension of Kolmogorov-Smirnov test (Lilliefors 1967). The test is readily available in statistical programs, but very sensitive to deviations from normality. As distribution-based causal inference relies on information in the higher, non-Gaussian, moments of statistical distributions, the statistical power of the method depends on the magnitude of the higher moments and is likely to be much lower than the power of the Lilliefors' test. Therefore, the Lilliefors' test and a visual inspection of histograms suit well for establishing the necessary condition *ii* (non-Gaussianity), but they may not be sufficient to ensure good statistical power for causal inference. Statistical power can be assessed by simulation, as further illustrated in the section 8. In our running example, Lilliefors' test rejected a null hypothesis of normal distribution for both sleep deviations ($D = 0.163$, $p < 0.001$) and depression ($D = 0.127$, $p < 0.001$). The sample skewness of the standardized sleep deviation and depression variables was 2.58 and 1.03. The estimates of excess kurtosis were 13.39 and 0.78, respectively.

5.3. Testing independence

If the LiNGAM model holds, and if X causes Y, then X should be statistically independent of the residual $\epsilon_Y = Y - \hat{\beta} X$, where $\hat{\beta}$ is the ordinary least squares regression coefficient. By definition, X is uncorrelated with the least squares residual, but it should also be fully independent in the sense that $E[f(X)g(Y - \hat{\beta} X)] = E[f(X)]E[g(Y - \hat{\beta} X)]$ holds for all (absolute integrable) functions f and g (Hyvärinen, Karhunen, and Oja 2001; Klenke 2008). Studying independence of arbitrary distributions is a difficult task, but several general methods do exist (Hoeffding 1948; Kallenberg and Ledwina 1999; Einmahl and McKeague 2003; Gretton and Györfi 2010). However, the assumption *iii* is not strictly necessary in the sense that distribution-based causal inference may work despite confounding (Rosenström et al. 2012) and algorithms designed for estimation in presence of confounding exist (Shimizu and Bollen 2014). As discussed above, the DirectLiNGAM test statistic compares *expected* values instead of testing strict hypotheses. The best course of action in practice may be to test whether the independence between predictor and estimated residual variable holds fully, and if not, use sensitivity analyses and triangulation (see below) instead of totally abandoning distribution-based causal inference.

In our running example, we observed that there was no *linear* dependence between the depression score and its residual when regressed on sleep deviations ($p = 0.996$), but a clear *statistical* dependence when assessed with tests such as Kallenberg 's and Ledwina's V test ($p <$ 0.001), Hoeffding's test ($p = 0.005$), and empirical likelihood test ($p < 0.001$), all of which are sensitive to dependencies beyond simple linear relationships (Hoeffding 1948; Kallenberg and Ledwina 1999; Einmahl and McKeague 2003). Similarly, by definition, there was no Pearson's product-moment correlation between sleep deviation and its residual when regressed on depression scores ($p = 0.984$), but there was a clear dependence when assessed using the above nonlinear measures (all $p < 0.001$). That is, whereas the residual and the predictor are *uncorrelated* with each other by definition of Ordinary Least Square regression, they typically are *not necessarily independent* of each other. The residual that is least dependent on the associated predictor may be indicative of causation.

6. Distribution-based causality estimates for the running example

We report (standardized) DirectLiNGAM and skewness- and kurtosis-based estimators for pairwise causal direction between depression score and sleep deviations, as in our previous work (Rosenström et al. 2012). The latter two estimates may sometimes reveal specific distributional properties most important for the general DirectLiNGAM estimate. The kurtosis-based estimator was previously called "tanh-based" because it is specifically based on hyperbolic-tangent approximation to likelihood ratio. In SATSA data, however, we observed that all the three estimators indicated absolute sleep deviations being a cause of other depressive symptoms rather than the other way around (Table 1). That is, whichever non-Gaussian moments of the respective distributions we looked at, they indicated sleep deviations as being a cause of depressive symptoms more likely, or more strongly (i.e., in expected value), than the other way around. The results of the sensitivity analyses we performed supported the same direction of dependence for both hypersomnic and insomnic sleep deviations, with the exception of the DirectLiNGAM estimate in the hypersomnia subsample (Table 1). However, we did not further interpret the single deviant result as a key LiNGAM assumption failed in that case: the linear correlation coefficient between depression score and oversleep did not statistically differ from zero (*r* = 0.03, *p* = 0.482). The linear association between insomnia and depression score was statistically significant (*r* = 0.27, *p* < 0.001).

[TABLE 1 HERE]

7. Conducting sensitivity analyses

7.1. Convergent evidence from multiple estimators

In the running example, we observed a certain type of indication for robustness, because different estimators, using different types of deviation from Gaussian distributions, converged in their estimates of causal direction (Table 1). That is, the estimated causal direction was not sensitive to specific distributional property beyond the necessary requirement of non-Gaussian distribution. Such robustness property does not necessarily hold (e.g., Rosenström et al., 2012), and establishing it can be comforting and evoke trust. It only indicates robustness with respect to distributional characteristics, however, not with respect to model assumptions. Epidemiologic triangulation is a process to establish robustness over model assumptions and modeling approaches, and it will be discussed in the section 9.

7.2. Simulation-based analysis of robustness to latent confounding

As discussed above, the requirement of perfect independence between residual and predictor variable may be unnecessary for causal inference, as well as overly restrictive. When relaxing this assumption, it may be desirable to gather some insight on possible biases that could result. For that and related purposes, one can conduct brief simulation studies to investigate the extent to which the method is sensitive to the simulated conditions. Unobserved confounding variables are a typical reason for failures of independence between the residual and the predictor variable in regression models. Complete confounding implies that there is no direct effect between X and Y (see Figure 4). In other words, experimental manipulations of X have no effect on Y despite their association with each other, unless also the 'true' causes of both the variables (variable Z in Figure 4) are manipulated. In practice, the possibility of confounding is difficult to definitively test in observation data. However, through simulation we can have a clue of the extent to which unobserved confounders may bias our causal inferences: by generating data which are as similar as possible to the observed data and by manipulating the degree of confounding in it.

[FIGURE 4]

For example, we generated a large number of simulated datasets with the same characteristics as our real-world data from the running example, and examined both possible directions of dependence, i.e. X and Y being the cause (i.e., sleep deviation causing other depression symptoms and *vice versa*). The datasets were manipulated so that they contained different degrees of 'unobserved' confounding in X and Y. Then, each one of the datasets was analyzed and the output saved. Finally, we investigated how robust DirectLiNGAM was to different degrees of confounding by obtaining the proportion of success in correctly picking the causally antecedent variable of a given simulation condition.

In what follows, we will walk the reader through the simulation step by step, displaying pseudocode and comments on it. By “pseudocode” we mean an informal code that is not based on a concrete programming language. Instead, it is a general-purpose text that allows one to understand and implement the simulation using whichever programming language that best serves the case.

7.2.1. Obtain data-based parameters

We estimated regression models in both directions on the standardized variables, and saved the regression residuals ($\hat{\epsilon}_Y$ and $\hat{\epsilon}_X$).. The slope coefficient was the same in both models as a consequence of the standardization, regardless of which variable was set as predictor or outcome ($\hat{\beta} = 0.148$) Distributions of the predictor and the residual in the simulation were approximated by a bootstrap distributions of their empirical distribution (Efron & Tibshirani, 1993).

7.2.2. Define parameters and simulation conditions

Once we had the parameters to simulate data akin to our running example, we defined the conditions for our simulation experiment. Our purpose was to check how sensitive DirectLiNGAM is to latent confounders by varying the degree of confounding. The parameter λ quantified the amount of variance due to the latent confounder Z (cf. Fig. 4). In total, we had four simulation settings coming from two times two conditions: conditions A and B relate to the direction of dependency tested, and conditions 1 and 2 define alternative distributions for the latent confounder Z.

In condition A, the regression model emulated the situation in which the predictor was distributed as the sleep variable in SATSA data, with also other parameters being as in the empiric model (slope and residuals). In condition B, the assumed predictor was distributed as the CESD depression score. Conditions 1 and 2 were included because the distribution of the latent confounder Z is a potential factor affecting the statistical power inferred from our simulation. We addressed this aspect by switching the distribution of Z, so that in the condition 1 it had the same distribution as the antecedent variable, and in the condition 2 it had a different distribution (i.e., bootstrap distribution of the descendent variable).

Because larger sample sizes (N) improve statistical power, we ran the sensitivity analyses using several sample sizes: 200, 500, 1000, 1325 (i.e., the sample size of our running example), and 5000. Thus, we could investigate whether DirectLiNGAM is more or less sensitive to latent confounding depending on sample size. DirectLiNGAM was computed for all combinations of N values and values of λ, totaling 30 parameter combinations for each simulation setting A1, A2, B1, and B2.Finally, one has to set the number of replications (R) the simulation will be run (a single run generates one dataset and the corresponding directLiNGAM estimate). We chose R=10 000. The larger the R, the more precise information we have on the unavoidable effects of sampling variance.

7.2.3. Define the simulation model

The structural model underlying the data-generating process of Y in accordance to LiNGAM assumptions is the linear model (e.g., $Y = \beta_x X + e_y$). Because our purpose was to introduce and investigate latent confounding, we had to generate the confounder Z and use it when generating Y. The data-generating linear model in this simulation is therefore: $Y_{\mathrm{sim}} = \lambda(\beta_Z Z) + (1 -$

$\lambda)(\beta_x X) + e_y$, where λ controls the degree of confounding. The simulated predictor variable X_{sim} is then a weighted sum of "unconfounded" X and a regression on Z, that is, $X_{\text{sim}} = \lambda(\beta_Z Z + e_X) + (1 - \lambda)X$. The distribution of Z cannot be identified from empirical data and was therefore set to a specific distribution under two conditions: in condition 1, Z was bootstrapped from the same distribution as X (e.g. both X and Z were independently bootstrapped from the depression score variable), and in condition 2, Z was bootstrapped from the distribution of the other variable of the pair (e.g. when X was bootstrapped from the depression score, Z was bootstrapped from the sleep variable). Table 2 shows the respective roles of the empirical bootstrap distributions (SATSA variables) in the data generating process of the four simulation settings. The residual distributions were also bootstrapped from the empirical distribution. β_Z was set as equal to β_X. The next lines show a brief sketch of the data-generating procedure.

[TABLE 2 HERE]

The pseudocode for data simulation of simulation setting A1:

```
R = 10000
lambda = vector(0, .2, .4, .6, .8, 1)
N = vector(200, 500, 1000, sample_size_of (SATSA), 5000)
residuals = residual_cesd
predictor = sleep deviations
confounder = sleep deviations
output = initialize_array(rows = R, cols = length(lambda), dim = length(N))
x_PD = parametric_estimate(X)
for each (n in N) do {
  for each (j in lambda) do {
    for each (i in 1 to R) do {
            % generate data
            e = pick_random_with_replacement(residuals), n))
            e2 = pick_random_with_replacement(residuals), n))
            e = e - mean(e)
            e2 = e2 - mean(e2)
            x = pick_random_with_replacement(predictor), n))
            z = pick_random_with_replacement(confounder), n))
            y = ((1-j)*betax*x + e + j*betaz*z)
            x = ((1-j)*x + j*(betaz*z + e2))
            % generate output
            output[i,index_of(j),index_of(n)] = DirectLiNGAM(x, y)
      } endfor
   } endfor
} endfor
```

The "for" statement is used when repeating the same action across all values of a given vector, or from index 1 to another integer. In this case, we repeated the simulation R=10 000 times per value of lambda (λ = .0, .2, .4, .6, .8, 1) and per sample size (N= 200, 500, 1000, 1325, 5000). This sums up to 6*5=30 experimental conditions, each of which had R=10 000 replications.[3]

[3] In addition to the pseudo-code shown here, the actual Octave/Matlab code for the simulations can be found from the web page: http://www.iki.fi/tom.rosenstrom

7.2.4. Run simulation and interpret results

As the output of our simulation experiment, we computed the success rate of DirectLiNGAM in picking up the correct causal direction over all R replications, and plotted the success rate with respect to different degrees of confounding (λ) and for different sample sizes (N). Because we have ourselves generated the data, we know the ground truth behind it: (1) that X causes Y rather than *vice versa* and (2) the degree there is a common variable causing them both. Knowing the ground truth makes it possible to estimate how successful the method is *despite* latent confounding. The analysis revealed that the distribution of the confounder Z had a negligible effect on the results, as can be noted comparing simulation settings A1 to A2 and B1 to B2 (Figure 5). Furthermore, switching the causal roles of the original cause and residual distributions in the simulation did not have a mentionable effect on the estimation success (Figure 5; a very small bias may be present at the 80% confounding, which could be further investigated in the future).

The success rate in causal estimation remained higher than 90% when introducing up to 40% of latent confounding in samples equal or bigger than N=1000 (Figure 5). As expected, when latent confounding was 80–100%, DirectLiNGAM estimates were nearly random, meaning that the method would pick up either X or Y as being the cause with almost the same probability. When the sample size was N=5000, DirectLiNGAM remained robust even up to 60% of latent confounding. In summary, the algorithm may be able to tolerate a considerable amount of latent confounding (violation of assumption *iii*) without noticeable performance loss in causal inferences.

[FIGURE 5]

8. Simulation-based analysis of statistical power

In the above section 7, we illustrated how to benefit from computer simulation in sensitivity analyses. There, the aim was to simulate controlled experiments that closely resemble the data at hand to investigate consequences of potential partially inaccurate model assumptions (i.e., degree of bias in the final estimate given a known degree of latent confounding). The conclusions one can draw from such sensitivity analyses are context specific by design. However, simulation studies are also helpful in collecting more general knowledge on algorithmic performance under different conditions. Here, we strive to provide the reader with intuition on statistical power of the DirectLiNGAM algorithm in estimation of pairwise directional dependence. Assumption of a non-Gaussian distribution is a *necessary* precondition for the kind of methods discussed here, but it does not automatically provide *sufficient* statistical power. This section tries to provide the reader with a rough intuition on how 'big' deviation from a Gaussian distribution is sufficient for a good statistical performance of the causal estimation algorithm.

A "deviation" from Gaussian distribution is commonly quantified using skewness, excess kurtosis, or differential entropy. A Gaussian distribution has a zero skewness and excess kurtosis, and other things being equal, the greater the absolute value of these statistics the less the evaluated distribution resembles a Gaussian distribution. Gaussian distribution is also the distribution of random movement and errors of measurement: on average, observations of a variable with a Gaussian distribution provide the least information imaginable for a continuously distributed variable with a given variance (alternatively, they are the least 'surprising' events; Cover and Thomas, 2006). Therefore, Gaussian distribution with variance σ has the maximal entropy (i.e., $\log(2\pi e\sigma^2)/2$), and the lower the entropy the more a distribution "deviates" from a Gaussian one. Thus, also information entropy can quantify deviations from normality for given variance. We arranged a simulation protocol to assess DirectLiNGAM estimation success under deviations of different magnitude, and to answer the question "*how* non-Gaussian variables one needs for causal inference".

Instead of distributions inferred from data, we used Log-normal distributions for the antecedent and residual variables (i.e., for X and e_y), which is a distribution for the exponent of a Gaussian variable (i.e., its logarithm would have a Gaussian distribution). Skewness, excess kurtosis, and entropy of a Log-normal variable are simple functions of mean and variance (σ) after log-transformation (i.e., for the generating Gaussian variable). We manipulated these parameters and calculated the estimation success by generating log-normally distributed predictors and residuals with varying scale parameters (eight conditions ranging from σ=0.05 to 0.75; β was adjusted to hold $Cor(X,Y)$ at a constant 0.4), with the following sample size conditions: N=100, 200, 500, 1000, 5000. Estimation success was computed as an average over R=10,000 replications of each condition. The resulting Figure 6 provides the reader with intuition on how statistical power of DirectLiNGAM responds to changes in these commonly used quantitative characterizations of statistical distributions. In general terms, the larger the sample the smaller the departure from Gaussianity that is sufficient to reach correct detection of causality. Samples of size 200 or less require clear deviations (entropy difference of .09, skewness= 1.53, or excess kurtosis= 2.35), while samples of size 500 and larger achieve statistical power above 95% already when showing only small departures from normality (entropy difference of .02, skewness= .46, or excess kurtosis= .37). The largest sample-size condition (N=5000) reached power above 95% with minimal deviations from normality.

While simulations are a good way to answer questions like "*what if* model assumptions are violated or conditions not ideal", triangulation, discussed below, is a technique that may be able to answer questions like "*are* difficult-to-test assumptions violated in the data at hand".

9. Triangulating causal inferences

Epidemiologists frequently deal with issues of life and death, literally. They also have many historical examples on "spurious", or misleading, findings. Thus, both the ramifications of false inferences and the previous experience warrant a cautious attitude towards translation of epidemiologic practice to public health policy. At the same time, doing so is an important part of evidence-based medicine. To cope with these conflicting demands, epidemiologists have introduced the idea of causal triangulation in etiologic epidemiology (Lawlor, Tilling, and Davey Smith 2017). Triangulation differs from generic attempts to show robustness across several estimators by aiming to show robustness across several estimators that have *different* key assumptions with respect to each other.

Typically, all causal inference techniques involve some assumptions that are difficult to test for, but necessary preconditions for applying the technique. However, it is often possible to find methods that make use of entirely different, or even 'opposite' types of information, to derive their inference on causality. For example, we could seek for a cross-sectional observational technique that does not rely on non-Gaussianity of the data when deriving otherwise similar statements on direction of causation between two variables? One such technique would be Direction of Causation (DoC) models studied in behavior genetics (Heath et al. 1993; Duffy and Martin 1994). DoC models assume that *normally distributed* variables for distinct genetic and environmental influences give rise to the observed (phenotypic) variables and correlations. Typical applications use twin models, where biological knowledge on monozygotic twins' 100% genetic similarity and dizygotic twins' 50% average genetic similarity is used to partition observed variance into additive genetic (A) influences, shared environmental influences common to both twins (C), and non-shared environmental influences unique to each twin (E; see, e.g., Neale and Cardon 1992).

We consider structural models nested within the path diagram in Figure 7, which describes a set of possible causal relationships (arrows) and correlations (arcs) between observed (boxes) and unobserved latent (circles) variables. All the paths are not identified at the same time, but we can test a direct-effect model with no latent correlations (no arcs; "reciprocal causation model") against the full correlational model it nests within (no arrows, but arcs; "no phenotypic causation model"). If the reciprocal causation model is not rejected, we can test if another one of the direct effects could be set to zero, indicating that sleep deviations cause depression, or *vice versa*.

To intuitively understand how the DoC approach infers causal directions, consider a case where similarity between twins on a causally antecedent trait is explained by genes and the similarity on the causally descendent trait is explained by the shared environment of the twins. Then twins reared apart would show the same (genetic) cross-trait similarity as twins reared together. With the opposite causation, only twins reared together would show cross-trait correlations because of their shared environment. In practice, one does not necessarily need data on different rearing statuses. For successful causal inference, however, one needs to be able to estimate three or more sources of familial similarity and their composition in the studied phenotypic/trait variables must differ across the variables (Heath et al. 1993).

[FIGURE 7]

In our running example, we had data on monozygotic and dizogotic twins reared together and reared apart, which we modeled using DoC models. We direct the reader to behavior genetics literature for more details (Neale and Cardon 1992; Heath et al. 1993), and simply provide the results here. For the DoC method, we used log-transformations to make the variables closer to being normally distributed. The reciprocal causation model was not rejected in a likelihood-ratio test ($\chi^2 = 0.72, d.f. = 1, p = 0.396$), allowing us to test unidirectional causal hypotheses. However, both a DoC model with sleep as a cause for depression ($\chi^2 = 3.55, d.f. = 1, p = 0.059$) and a DoC model with depression as a cause for sleep ($\chi^2 = 3.27, d.f. = 1, p = 0.071$) were close to being rejected, though not quite statistically significant. In terms of Bayesian Information Criterion (lower values indicate better fit), the unidirectional causal models were practically indistinguishable from each other (-13479.2 and -13479.4, respectively), but not from the reciprocal causation model (-13475.9).

Before we rush to conclude that we have a case of reciprocal causation, however, we must address the limitations of the DoC method in this specific case. First, we did not have the minimum of three biometric sources of variance required for detecting reciprocal causation, because neither of our variables had a statistically significant contribution from shared environmental influences (Table 3). Under these conditions, we could not have rejected the hypothesis of reciprocal causation, even if it were false. Second, the inheritance pattern of depression and sleep variables had remarkably similar composition, meaning that we had little power to distinguish between directions of causation (Heath et al. 1993; Duffy and Martin 1994). Third, although we cannot directly assess the degree of third-variable confounding, a considerable extent is expected in this case and not well-handled by the DoC model (Heath et al. 1993). Fourth, measurement error can bias causal inferences based on DoC models, and therefore efforts to minimize it would be a desirable part of a DoC analysis (Heath et al. 1993). Altogether, promising as it was, the DoC modeling approach provided little causal information in this case. However, it is to be expected in causal triangulation that some approaches turn out more informative than other approaches, and that confidence can be built only gradually.

[TABLE 3]

One could continue the process of causal triangulation using, for example, instrumental variable regression method for causal inference, which is yet another method based on different assumptions than distribution- and DoC-based causal inference (Heath et al. 1993). In instrumental variable regression, one needs an auxiliary variable that is a known cause of a target variable in causal inference and known to affect the other target variable only through the first target variable. For example, genes controlling the circadian clock might serve as an instrument that is causal for sleep deviations, and for depression only through their effect on sleep (Lawlor et al. 2008). Of course, that would be an assumption, and the clock genes might also have other unknown effects on the brain. All in all, every causal-inference method at researcher's disposal is an asset in causal triangulation, as any single method is unlikely to be decisive.

10. Conclusion

In this chapter, we briefly reviewed past epidemiologic studies using distribution-based causal inference (DirectLiNGAM in particular), discussed the novel method from the viewpoint of more established epidemiologic research, and replicated previous findings on causality between sleep problems and other depressive symptoms (running example). At the population level, sleep problems were more likely to cause at least mild forms of depressive symptoms than the other way around—a conclusion that may well differ in severely symptomatic clinical samples (Rosenström et al. 2012). In addition, we showed how to conduct simulation-based sensitivity analyses and how to study statistical power of the algorithm in different settings. Finally, we discussed use of DirectLiNGAM as a part of a general process of causal triangulation in etiologic epidemiology. To provide an example of alternative causal inference technique with very different assumptions to DirectLiNGAM, we applied Direction of Causation (DoC) models from behavior genetics. These turned out uninformative in our running example, but nevertheless served to illustrate the general process of triangulation.

DoC models applied herein also illustrated that DirectLiNGAM is, in fact, a rather robust technique for causal inference. Population samples that reflect natural data-generating processes are often available, whereas it can be quite difficult to satisfy the assumptions of DoC or instrumental-variable methods (Heath et al. 1993; Duffy and Martin 1994; Lawlor et al. 2008). In addition to the above-discussed assumptions, validity of DoC models also depends on the validity of the applied inheritance model; for example, the classic ACE model we used requires a number of conditions to hold, such as non-assortative mating and equal environments for identical twins and other siblings (Neale and Cardon 1992). Furthermore, the inheritance patterns that DoC methods use often are functions of changing environmental conditions (Heath et al. 1985), which could sometimes lead to surprises in DoC modeling. For example, a colleague once described a lack of trust toward DoC methodology due to having found that recent observations on a variable had 'caused' historical observations in the same variable according to his DoC application, thus reversing the 'arrow of time' (personal communication). So far DirectLiNGAM has not led to comparable spurious findings. It shows remarkably good robustness properties and statistical power, while making much less stringent assumptions than many alternative methods. However, more research is needed on possible biases of distribution-based causal inference methods in various real-world research problems.

With complex constructs, such as psychiatric disorders, the assumption of no confounding due to third, unobserved variables may not be very realistic. Here and previously, we noted that the DirectLiNGAM approach can be quite robust against confounding. There are also later extensions of the method specifically developed to handle unobserved confounding (Shimizu and Bollen 2014). Distribution-based causal inference techniques have also been developed for time-series analysis, for some nonlinear models, and for other special cases (Hyvärinen et al. 2010; Wiedermann and von Eye 2016). On the methodological side, the field is developing rapidly, whereas within epidemiology, it has yet to demonstrate its value. Distribution-based causal inference methods have essential similarities with a statistical signal-processing technique known as independent component analysis, which has generated much interest and many applications (Hyvärinen, Karhunen, and Oja 2001; Shimizu et al. 2006). Time will show whether these methods find their place in the standard toolkit of epidemiologists as well.

References

Borsboom, D. (2017). A Network Theory of Mental Disorders. *World Psychiatry* 16 (1): 5–13. doi:10.1002/wps.20375.

Cover, T. A. & Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley & Sons, Inc., Hoboken.

Cramer, A. O. J., Waldorp, L. J., van der Maas, H. L. J., & Borsboom, D. (2010). Comorbidity: A Network Perspective. *Behavioral and Brain Sciences* 33: 137–93. doi:10.1017/S0140525X09991567.

Dodge, Y. & Rousson, V. (2000). Direction Dependence in a Regression Line. *Communications in Statistics - Theory and Methods* 29 (9-10): 1957–72. doi:10.1080/03610920008832589.

Dodge, Y., & Rousson, V. (2001). On asymmetric properties of the correlation coefficient in the regression setting. *The American Statistician*, *55*(1), 51-54. doi: 10.1198/000313001300339932

Dodge, Y., & Yadegari, I. (2010). On direction of dependence. *Metrika*, *72*(1), 139-150. doi: 10.1007/s00184-009-0273-0. Duffy, D. L. & Martin, N. G. (1994). Inferring the Direction of Causation in Cross-Sectional Twin Data: Theoretical and Empirical Considerations. *Genetic Epidemiology* 11 (6): 483–502. doi:10.1002/gepi.1370110606.

Einmahl, J. H. J. & McKeague, I. W. (2003). Empirical Likelihood Based Hypothesis Testing. *Bernoulli* 9 (2): 267–90. http://www.jstor.org/stable/3318940.

Efron, B. & Tibshirani, R.J. (1993). *An Introduction to Bootstrapping*. Chapman & Hall, New York.

García-Velázquez, R., Jokela, M., & Rosenström, T. (2017). Symptom Severity and Disability in Psychiatric Disorders: The U.S. Collaborative Psychiatric Epidemiology Survey. *Journal of Affective Disorders* 222: 204–10. doi:10.1016/j.jad.2017.07.015.

Gretton, A. & Györfi, L. (2010). Consistent Nonparametric Tests of Independence. *The Journal of Machine Learning Research* 11: 1391–1423.

Heath, A. C., Berg, K., Eaves, L. J., Solaas, M. H., Corey, L. A., Sundet, J., Magnus, P., and Nance, W. E. (1985). Education Policy and the Heritability of Educational Attainment. *Nature* 314 (6013): 734–36.

Heath, A. C., Kessler, R. C., Neale, M. C., Hewitt, J. K., Eaves, L. J., & Kendler, K. S. (1993). Testing Hypotheses About Direction of Causation Using Cross-Sectional Family Data. *Behavior Genetics* 23 (1): 29–50.

Helajärvi, H., Rosenström, T., Pahkala, K., Kähönen, M., Lehtimäki, T., Heinonen, O. J., Oikonen, M., Tammelin, T., Viikari, J. S. A., & Raitakari, O. T. (2014). Exploring Causality Between TV Viewing and Weight Change in Young and Middle-Aged Adults. the Cardiovascular Risk in Young Finns Study. *PLoS ONE* 9 (7): e101860. doi:10.1371/journal.pone.0101860.

Hoeffding, W. (1948). A Non-Parametric Test of Independence. *The Annals of Mathematical Statistics* 19 (4): 546–57. http://www.jstor.org/stable/2236021.

Hyvärinen, A. & Smith, S. M. (2013). Pairwise Likelihood Ratios for Estimation of Non-Gaussian Structural Equation Models. *Journal of Machine Learning Research* 14: 111–52. http://jmlr.csail.mit.edu/papers/v14/hyvarinen13a.html.

Hyvärinen, A., J. Karhunen, and E. Oja. 2001. *Independent Component Analysis.* Wiley, New York.

Hyvärinen, A., Zhang, K., Shimizu, S., & Hoyer, P.O. (2010). Estimation of a Structural Vector Autoregression Model Using Non-Gaussianity. *Journal of Machine Learning Research* 11: 1709–31. http://dl.acm.org/citation.cfm?id=1756006.1859907.

Kallenberg, W. C. M. & Ledwina, T. (1999). Data-Driven Rank Tests for Independence. *Journal of the American Statistical Association* 94 (445): 285–301. doi:10.1080/01621459.1999.10473844.

Klenke, A. 2008. *Probability Theory: A Comprehensive Course*. London, England: Springer-Verlag.

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., & Smith, D. G. (2008). Mendelian Randomization: Using Genes as Instruments for Making Causal Inferences in Epidemiology. *Statistics in Medicine* 27 (8): 1133–63. doi:10.1002/sim.3034.

Lawlor, D. A., Tilling, K., & Smith, G. D. (2017). Triangulation in Aetiological Epidemiology. *International Journal of Epidemiology* 45 (6): 1866–86. doi:10.1093/ije/dyw314.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association* 62 (318): 399–402.

Neale, M. C. & Cardon, L. R. (1992). *Methodology for Genetic Studies of Twins and Families*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Pedersen, N. L. (2015). Swedish Adoption/Twin Study on Aging (SATSA), 1984, 1987, 1990, 1993, 2004, 2007, and 2010. Inter-university Consortium for Political; Social Research (ICPSR) [distributor]. http://doi.org/10.3886/ICPSR03843.v2.

Radloff, L. S. (1977). The CES-D Scale: A Self Report Depression Scale for Research in the General Population. *Applied Psychological Measurement* 1 (3): 385–401.

Rosenström, T., Jokela, M., Puttonen, S., Hintsanen, M., Pulkki-Raback, L., Viikari, J.S., Raitakari, O.T., & Keltikangas-Järvinen, L. (2012). Pairwise Measures of Causal Direction in the

Epidemiology of Sleep Problems and Depression. *PLoS ONE* 7 (11): e50841. doi:10.1371/journal.pone.0050841.

Shimizu, S. & Bollen, K. (2014). Bayesian Estimation of Causal Direction in Acyclic Structural Equation Models with Individual-Specific Confounder Variables and Non-Gaussian Distributions. *Journal of Machine Learning Research* 15 (1): 2629–52.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. (2006). A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research* 7: 2003–30.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., & Bollen, K. (2011). DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model. *Journal of Machine Learning Research* 12: 1225–48.

Wiedermann, W. (2018). A note on fourth moment-based direction dependence measures when regression errors are non-normal. *Communications in Statistics: Theory and Methods*, (in press). doi: 10.1080/03610926.2017.1388403.

Wiedermann, W., & von Eye, A. (2016). *Statistics and Causality: Methods for Applied and Empirical Research*. Hoboken, US-NJ: Wiley.

Figure captions

Figure 1: Illustration of skewness-based causal signal. Histogram of a Log-normal distributed variable X, and a variable that is a weighted sum of X and another similarly distributed "residual" variable. In the lower-left panel, a scatter plot of X and the weighted-sum variable (the outcome; cf. Y in text) are shown, whereas the lower-right panel shows similarly treated Gaussian variables.

Figure 2: Illustrating the SATSA data. Whereas the histograms (a and b) show unstandardized data, data was standardized for bivariate analyses for cross-study comparability, as a wide range of alternative assessment tools exists. Both bipolar (c; direction of deviance matters) and absolute/unipolar (d; both directions equally 'bad') sleep deviations were studied. LOESS (local regression) lines in c and d panels show how use of absolute values linearizes the association.

Figure 3: Illustrating lurking nonlinearities. The first panel shows a scatterplot of data points that an epidemiologist could legitimately approach via linear regression model. The second panel shows trajectories of the nonlinear Rössler system that was used to generate the data points by taking every 100th iteration from a numeric iteration of the system of differential equations. Rössler's classic parameter values were used (a = b = 0.2 and c = 5.7; see, e.g., Wikipedia page for Rössler attractor). Human eye and brain are an exceptionally good pattern-detection device, and the reader may see the concentric pattern that hints about the underlying non-randomness even in the left-most panel. However, even a minor degree of measurement noise would destroy the appearance.

*Figure 4: Illustration of the assumed model in DirectLiNGAM (*a*; one variable causes the other) and a fully confounded model (*b*; neither* X *nor* Y *is causal despite their correlation). In our simulation protocol, the degree of confounding is manipulated so that* Y *is a weighted sum of situations* (a) *and* (b) *in a gradient from 0 to 100%, where* λ=0 *would correspond to situation* (a) *and* λ=1 *would correspond to situation* (b). *The simulation was tailored to inform about the power of DirectLiNGAM to detect the right causal direction specifically in our running example.*

Figure 5: Results for simulation-based sensitivity analyses of DirectLiNGAM in the investigated conditions of latent confounding. Different panels correspond to different assigned distributions for the simulated cause, residual, and confounder variables (assigned distributions in Table 2).

Figure 6: Results for the power analysis of DirectLiNGAM in terms of skewness, excess kurtosis, and entropy difference (to a Gaussian variable of equal variance).

Figure 7: Path diagram for Direction of Causation (DoC) models in the running example. By constraining different paths, DoC models study whether family data is best explained by simple correlations between the two observed variables' (boxes) genetic (A) and shared (C) and non-shared (E) environmental influences ('spurious' association), by regression of one variable on the other (causation), or by regression of both variables on each other (reciprocal causation).

Tables

Table 1. Estimates of causal direction for depression and sleep variables of the running example

Method	Depression as cause %	Sleep as cause %	Statistic	Lower CI	Upper CI	Hypersomnia subsample, sleep as cause %	Insomnia subsample, sleep as cause %
DirectLiNGAM	0.20	99.80	-0.0461	-0.0964	-0.0115	40.1	99.65
Skew-based	0.10	99.90	-0.0438	-0.0920	-0.0094	89.8	99.35
Kurtosis-based	0.25	99.75	-0.0034	-0.0066	-0.0009	89.9	83.85

Note: results are shown for 2000 bootstrap resamples. Percent selected as cause is shown for each estimator, as well as T(depression, sleep) statistic and its 95 percent bootstrap percentile confidence intervals (CI). The two last columns replicate the analysis in those who sleep more than average ('Hypersomnia subsample') and in those who sleep less than population average hours per night ('Insomnia subsample').

Table 2. Role of the variables on generating the data of the four simulation settings.

Simulation setting	Role	SATSA variable
A1	Predictor	Sleep deviation
	Confounder	Sleep deviation
	Residuals	$\hat{e}_{CESD}$
A2	Predictor	Sleep deviation
	Confounder	CESD score
	Residuals	$\hat{e}_{CESD}$
B1	Predictor	CESD score
	Confounder	CESD score
	Residuals	$\hat{e}_{sleep}$
B2	Predictor	CESD score
	Confounder	Sleep deviation
	Residuals	$\hat{e}_{sleep}$

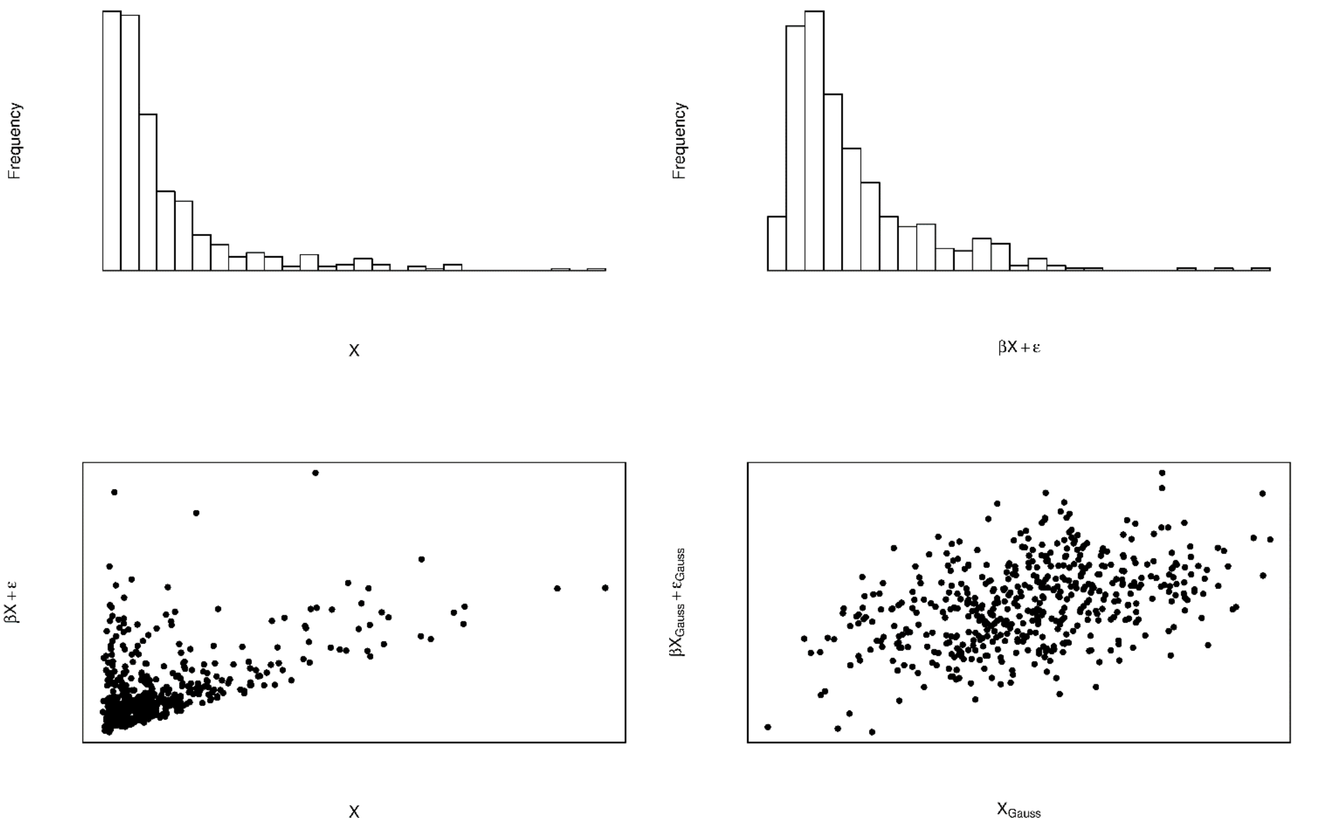
Note: The variables in the simulation settings were bootstrapped from the standardized SATSA variables and regression residuals ê.

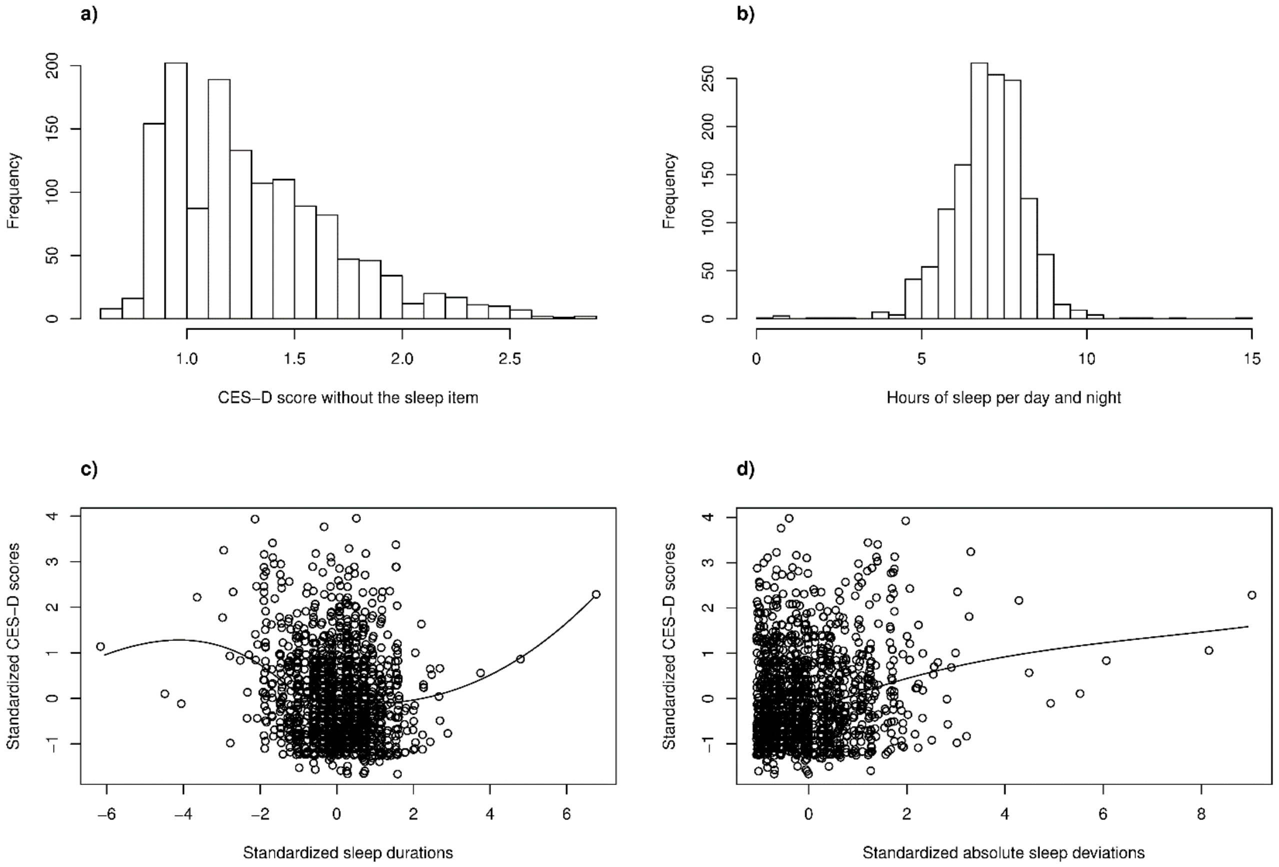
Table 3. Estimated biometric sources of variance for depression and sleep variables of the running example

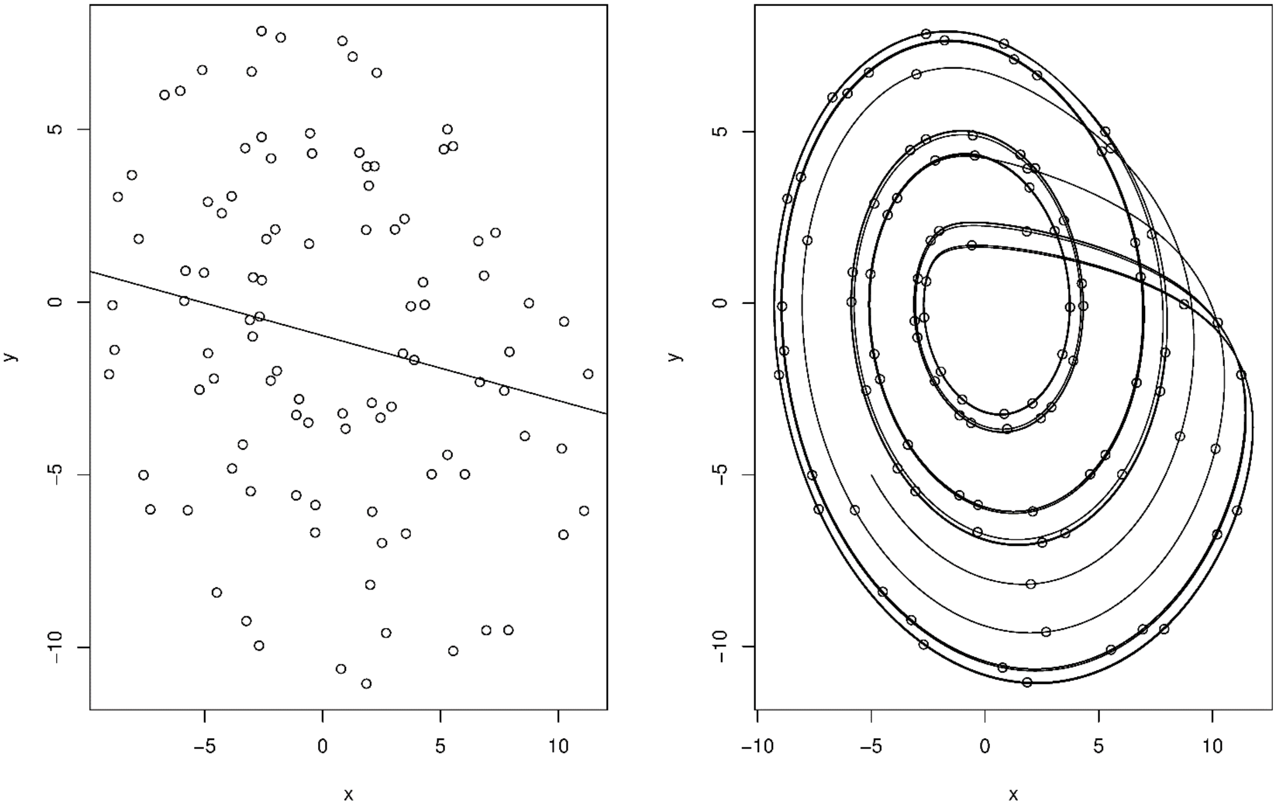
Variable: component	Estimate	Lower CI	Upper CI
Depression score: A	0.216	0.014	0.400
Depression score: C	0.113	-0.071	0.299
Depression score: E	0.671	0.554	0.801
Sleep deviations: A	0.276	0.083	0.446
Sleep deviations: C	-0.001	-0.171	0.171
Sleep deviations: E	0.725	0.597	0.863

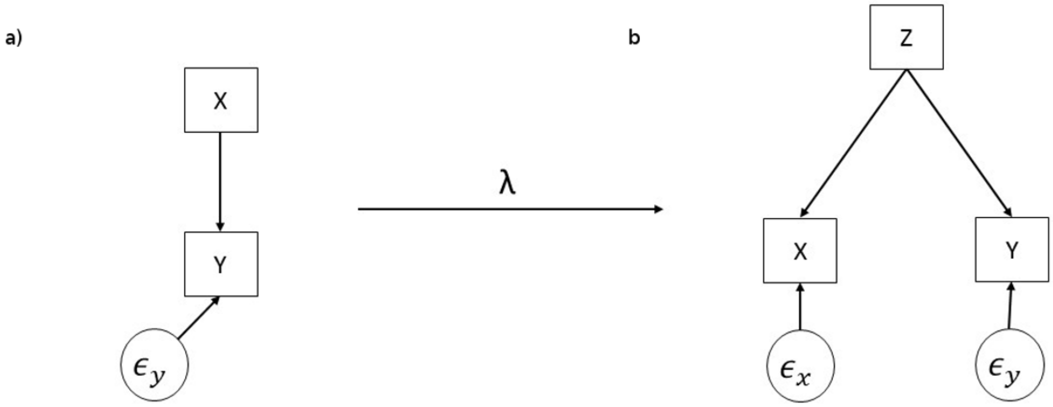
Note: "A" refers to additive genetic influences, "C" to shared environmental influences of the twins, "E" to non-shared environmental influences unique to only one member of each twin pair, and "CI" to 95% likelihood-profile confidence intervals.

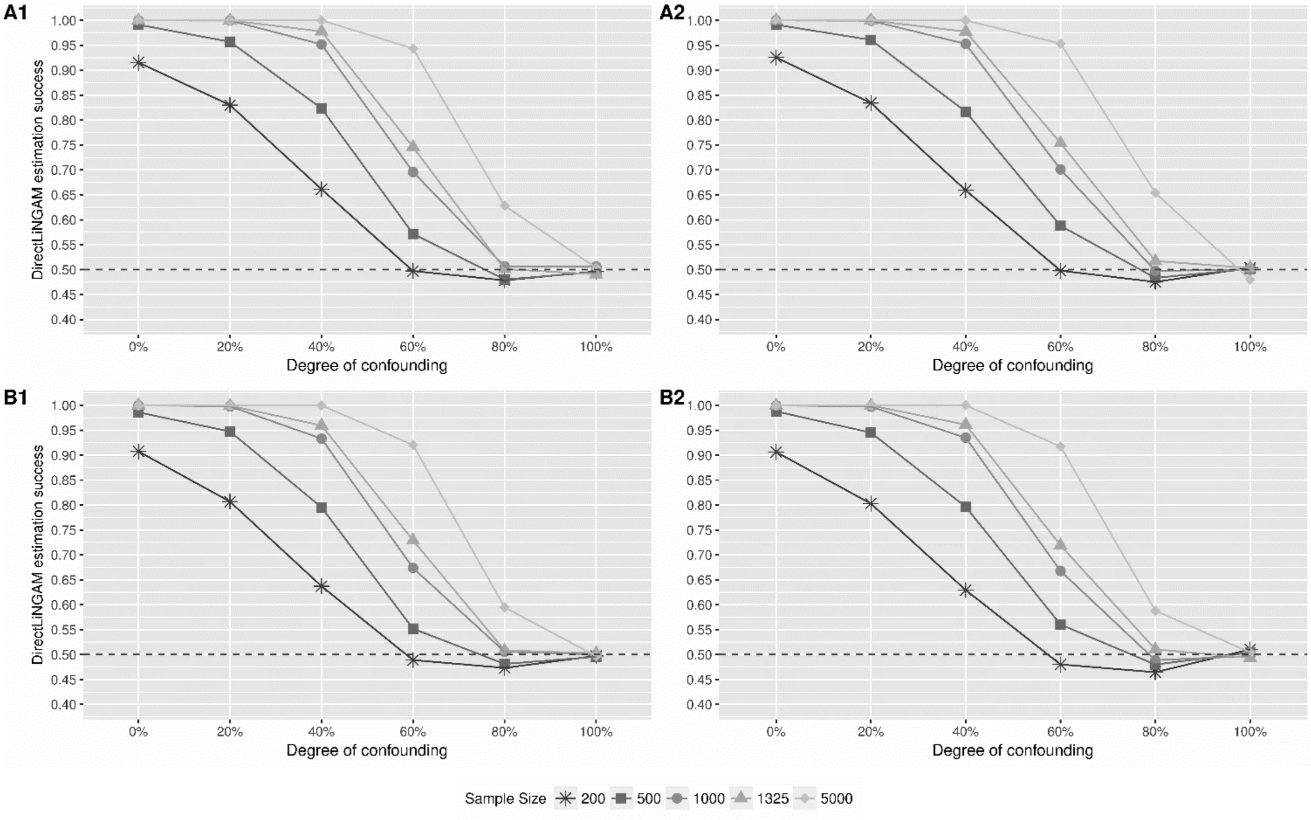
Figures

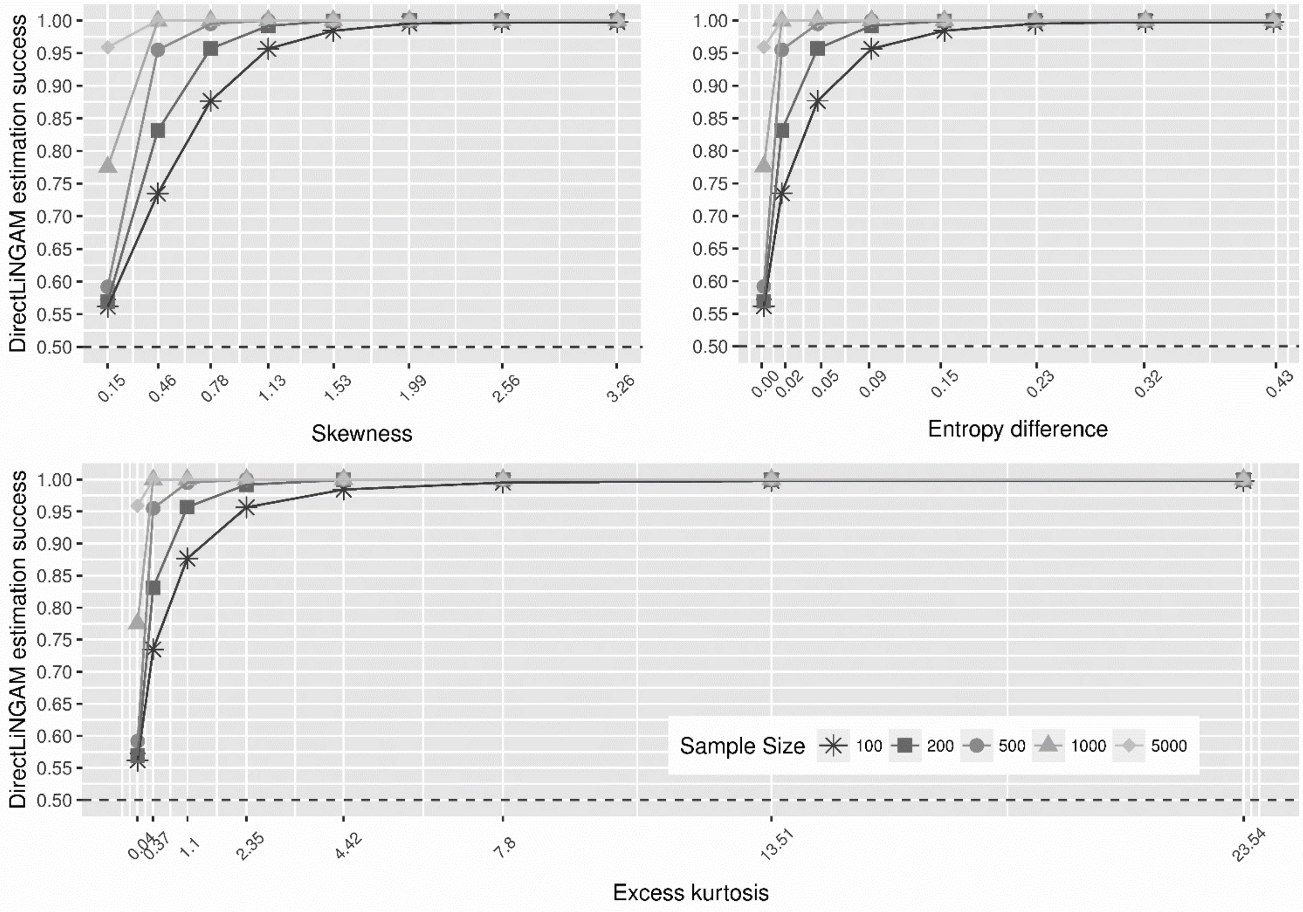

Frequency
X
Frequency
$\beta X + \varepsilon$
$\beta X + \varepsilon$
X
$\beta X_{Gauss} + \varepsilon_{Gauss}$
X_{Gauss}

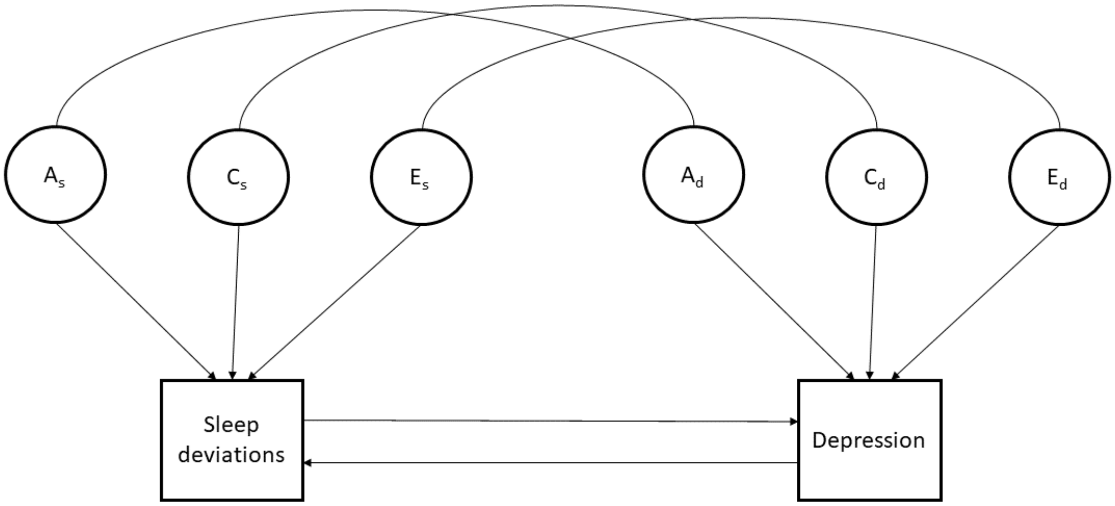

a)
Frequency
0
50
100
150
200
1.0
1.5
2.0
2.5
CES−D score without the sleep item
b)
Frequency
0
50
100
150
200
250
0
5
10
15
Hours of sleep per day and night
c)
Standardized CES−D scores
−1
0
1
2
3
4
−6
−4
−2
0
2
4
6
Standardized sleep durations
d)
Standardized CES−D scores
−1
0
1
2
3
4
0
2
4
6
8
Standardized absolute sleep deviations


y
5
0
−5
−10
−5
0
5
10
x
y
5
0
−5
−10
−10
−5
0
5
10
x


a)
X
Y
ϵ_y
λ
b
Z
X
Y
ϵ_x
ϵ_y


A1
A2
B1
B2
DirectLiNGAM estimation success
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0%
20%
40%
60%
80%
100%
Degree of confounding
Sample Size
200
500
1000
1325
5000


DirectLiNGAM estimation success
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.15
0.46
0.78
1.13
1.53
1.99
2.56
3.26
Skewness
0.00
0.02
0.05
0.09
0.15
0.23
0.32
0.43
Entropy difference
0.04
0.37
1.1
2.35
4.42
7.8
13.51
23.54
Excess kurtosis
Sample Size
100
200
500
1000
5000


A_s
C_s
E_s
A_d
C_d
E_d
Sleep deviations
Depression