Journal of Machine Learning Research 14 (2013) 111-152 Submitted 10/11; Revised 8/12; Published 1/13

Pairwise Likelihood Ratios for Estimation of Non-Gaussian Structural Equation Models

Aapo Hyvärinen AAPO.HYVARINEN@HELSINKI.FI
Dept of Computer Science and HIIT
Dept of Mathematics and Statistics
University of Helsinki
Helsinki, Finland

Stephen M. Smith STEVE@FMRIB.OX.AC.UK
FMRIB (Oxford University Centre for Functional MRI of the Brain)
Nuffield Dept of Clinical Neurosciences
University of Oxford
Oxford, UK

Editor: Peter Spirtes

Abstract

We present new measures of the causal direction, or direction of effect, between two non-Gaussian random variables. They are based on the likelihood ratio under the linear non-Gaussian acyclic model (LiNGAM). We also develop simple first-order approximations of the likelihood ratio and analyze them based on related cumulant-based measures, which can be shown to find the correct causal directions. We show how to apply these measures to estimate LiNGAM for more than two variables, and even in the case of more variables than observations. We further extend the method to cyclic and nonlinear models. The proposed framework is statistically at least as good as existing ones in the cases of few data points or noisy data, and it is computationally and conceptually very simple. Results on simulated fMRI data indicate that the method may be useful in neuroimaging where the number of time points is typically quite small.

Keywords: structural equation model, Bayesian network, non-Gaussianity, causality, independent component analysis

1. Introduction

Estimating structural equation models (SEMs), or linear Bayesian networks is a challenging problem with many applications in bioinformatics, neuroinformatics, and econometrics. If the data is Gaussian, the problem is fundamentally ill-posed. Recently, it has been shown that using the non-Gaussianity of the data, such models can be identifiable (Shimizu et al., 2006). This led to the Linear Non-Gaussian Acyclic Model, or LiNGAM.

The original method for estimating LiNGAM was based on first applying independent component analysis (ICA) to the data and then deducing the network connections from the results of ICA. However, we believe that it may be possible to develop better methods for estimating LiNGAM directly, without resorting to ICA algorithms.

A framework called DirectLiNGAM was, in fact, proposed by Shimizu et al. (2011) to provide an alternative to the ICA-based estimation. DirectLiNGAM was shown to give promising results especially in the case where the number of observed data points is small compared to the dimension of the data. It can also have algorithmic advantages because it does not need gradient-based iterative methods. An essential ingredient in DirectLiNGAM is a measure of the causal direction between two variables.

An alternative approach to estimating SEMs is to first estimate which variables have connections and then estimate the direction of the connection. While a rigorous justification for such an approach may be missing, this is intuitively appealing especially in the case where the amount of data is limited. Determining the directions of the connections can be performed by considering each connection separately, which requires, again, analysis of the causal direction between two variables. Such an approach was found to work best by Smith et al. (2011) which considered causal analysis of simulated functional magnetic resonance imaging (fMRI) data, where the number of time points is typically small. A closely related approach was proposed by Hoyer et al. (2008), in which the PC algorithm was used to estimate the existence of connections, followed by a scoring of directions by an approximate likelihood of the LiNGAM model; see also Ramsey et al. (2011).

Thus, we see that measuring pairwise causal directions is a central problem in the theory of LiNGAM and related models. In fact, analyzing the causal direction between two non-Gaussian random variables (with no time structure) is an important problem in its own right, and was considered in the literature before the advent of LiNGAM (Dodge and Rousson, 2001).

In this paper, we develop new measures of causal direction between two non-Gaussian random variables, and apply them to the estimation of LiNGAM. The approach uses the ratio of the likelihoods of the models corresponding to the two directions of causal influence. A likelihood ratio is likely to provide a statistically powerful method because of the general optimality properties of likelihood. We further propose first-order approximations of the likelihood ratio which are easy to compute and have simple intuitive interpretations. They are also closely related to higher-order cumulants and may be more resistant to noise. The framework is also simple to extend to cyclic or nonlinear models.

The paper is structured as follows. The measures of causal directions are derived in Section 2. In Section 3 we show how to apply them to estimating the model with more than two variables. The extension to cyclic models is proposed in Section 4 and an extension to a nonlinear model in Section 5. We report simulations with comparisons to other methods in Section 6, experiments on simulated brain imaging data in Section 7, and results on a publicly available benchmark data set in Section 8. Section 9 concludes the paper. Preliminary results were published by Hyvärinen (2010).

2. Finding Causal Direction Between Two Variables

In this section, we present our main contribution: new measures of causal direction between two random variables.

This section is structured as follows: We first define the problem in Section 2.1. We derive the likelihood ratio in Section 2.2. We propose a general-purpose approximation for the likelihood ratio in Section 2.3. The connection to mutual information is explained in Section 2.4. We derive a particularly simple approximation for the likelihood ratio in Section 2.5, and propose an instance for the case of sparse, symmetric densities. A theoretical analysis of the approximation based on cumulants is given in Section 2.6. We give intuitive interpretations of the approximations in Sec-

tion 2.7, and discuss their noise-tolerance in Section 2.8. Finally, we show how to use the likelihood ratio approximations in the case of skewed variables in Section 2.9.

For the benefit of the reader, we have further created Table 3 in the Conclusion on page 150 that lists the main new measures proposed in this paper.

2.1 Problem Definition

Denote the two observed random variables by x and y. Assume they are non-Gaussian, as well as standardized to zero mean and unit variance. Our goal is to distinguish between two causal models. The first one we denote by $x \to y$ and define as

$$y = \rho x + d$$

where the disturbance d is independent of x, and the regression coefficient is denoted by ρ. The second model is denoted by $y \to x$ and defined as

$$x = \rho y + e$$

where the disturbance e is independent of y. The parameter ρ is the same in the two models because it is equal to the correlation coefficient. Note that these models belong to the LiNGAM family (Shimizu et al., 2006) with two variables. In the following, we assume that x, y follow one of these two models.

Note that in contrast to Dodge and Rousson (2001) or Dodge and Yadegari (2010), we do not assume that d or e are normal, or have zero cumulants. We make no assumptions on their distributions. It is not even necessary to assume that they are non-Gaussian; it is enough that x and y are non-Gaussian. (This is related to the identifiability theorem in ICA which says that one of the latent variables can be non-Gaussian, see Comon, 1994).

2.2 Likelihood Ratio

An attractive way of deciding between the two models is to compute their likelihoods and their ratio. Consider a sample $(x_1, y_1), \ldots, (x_T, y_T)$ of data. The likelihood of the LiNGAM in which $x \to y$ was given by Hyvärinen et al. (2010) as

$$\log L(x \to y) = \left[\sum_t G_x(x_t) + G_d\left(\frac{y_t - \rho x_t}{\sqrt{1-\rho^2}}\right) \right] - T \log(1-\rho^2)$$

where $G_x(u) = \log p_x(u)$, and G_d is the standardized log-pdf of the residual when regressing y on x. The last term here is a normalization term due to the use of standardized log-pdf G_d. From this we can compute the likelihood ratio, which we normalize by $\frac{1}{T}$ for convenience:

$$\begin{aligned} R = \frac{1}{T}\log L(x \to y) - \frac{1}{T}\log L(y \to x) \\ = \frac{1}{T}\sum_t G_x(x_t) + G_d\left(\frac{y_t - \rho x_t}{\sqrt{1-\rho^2}}\right) - G_y(y_t) - G_e\left(\frac{x_t - \rho y_t}{\sqrt{1-\rho^2}}\right). \quad (1) \end{aligned}$$

We can thus compute R and decide based on it what the causal direction is. If R is positive, we conclude $x \to y$, and if it is negative, we conclude $y \to x$.

To use (1) in practice, we need to choose the G's and estimate ρ. The statistically optimal way of estimating ρ would be to maximize the likelihood, but in practice it may be better to estimate it simply by the conventional least-squares solution to the linear regression problem. Nevertheless, maximization of likelihood might be more robust against outliers, because log-likelihood functions often grow more slowly than the squaring function when moving away from the origin.

Choosing the four log-pdf's G_x, G_y, G_d, G_e could, in principle, be done by modelling the relevant log-pdf's by parametric (Karvanen and Koivunen, 2002) or non-parametric (Pham and Garrat, 1997) methods, which will be discussed in more detail below. However, for small sample sizes such modelling can be very difficult. In the following, we provide various parametric approximations.

2.3 Maximum Entropy Approximations of Likelihood Ratio

The likelihood ratio has a simple information-theoretic interpretation which also means we can use well-known entropy approximations for its practical computation in the case where we do not want to postulate functional forms for the G's.

If we take the asymptotic limit of the likelihood ratio, we obtain

$$R \longrightarrow -H(x) - H(\hat{d}/\sigma_d) + H(y) + H(\hat{e}/\sigma_e) \tag{2}$$

where we denote differential entropy by H, the estimated residuals by $\hat{d} = y - \rho x, \hat{e} = x - \rho y$, and the variances of the estimated residuals by σ_d^2, σ_e^2.

Thus, we can approximate the likelihood ratio using any general, possibly non-parametric, approximations of differential entropy. For example, we can use the maximum entropy approximations by Hyvärinen (1998) which are computationally simple. In fact, we only need to approximate one-dimensional differential entropies, which is much simpler than approximating two-dimensional entropies.

One version of the approximations by Hyvärinen (1998) is given by

$$\hat{H}(u) = H(\nu) - k_1[E\{\log\cosh u\} - \gamma]^2 - k_2[E\{u\exp(-u^2/2)\}]^2 \tag{3}$$

where $H(\nu) = \frac{1}{2}(1 + \log 2\pi)$ is the entropy of the standardized Gaussian distribution, and the other constants are numerically evaluated as

$$\begin{aligned} k_1 &\approx 79.047, \\ k_2 &\approx 7.4129, \\ \gamma &\approx 0.37457. \end{aligned}$$

This approximation is valid for standardized variables; in fact, all the variables in (2) are standardized. The intuitive idea in this approximation is that since the Gaussian distribution has maximum entropy among all distributions of unit variance, entropy can be approximated by a measure of non-Gaussianity which is subtracted from $H(\nu)$. The sum of the second and third terms on the right-hand side of (3) is a measure of non-Gaussianity (ignoring their negative signs) since the terms are the squared differences of certain statistics from the corresponding values obtained for a Gaussian distribution. In fact, γ is defined as the expectation of $\log\cosh$ for a standardized Gaussian distribution, so the second term on the right-hand side is zero for a Gaussian distribution, just like the skewness-like statistic measured by the last term.

The expression in (2) also readily gives a simple intuitive interpretation of the estimation of causal direction. The (negative) entropies can all be interpreted as measures of non-Gaussianity, since the variables are standardized. Thus, in (2) we essentially compute the sum of the non-Gaussianities of the explaining variable and the resulting residual, and compare them for the two directions. The directions which leads to maximum non-Gaussianity is chosen.[1]

2.4 Connection to Mutual Information

It is also possible to give an information-theoretic interpretation which connects the likelihood ratios to independence measures.

A widely-used independence measure is mutual information, defined for two variables x, y as

$$I(x,y) = H(x) + H(y) - H(x,y)$$

where H denotes differential entropy. For a linear transformation

$$\begin{pmatrix} u \\ v \end{pmatrix} = \mathbf{A} \begin{pmatrix} x \\ y \end{pmatrix},$$

we have the entropy transformation formula

$$H(u,v) = H(x,y) + \log|\det \mathbf{A}|.$$

On the other hand, the transformation from x, y to x, d has unit determinant, since

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ a & 1 \end{pmatrix} \begin{pmatrix} x \\ d \end{pmatrix}.$$

Thus, we have

$$H(x,d) = H(x,y)$$

and likewise for $H(y,e)$. We can now consider the mutual information of the regressors and the residuals in the two models, and in particular, compute the difference of the mutual informations to see which one is smaller. In fact, the difference of the mutual informations is asymptotically equal to the likelihood ratio R since

$$\begin{aligned} I(x,d) - I(y,e) &= H(x) + H(d) - H(x,d) - (H(y) + H(e) - H(y,e)) \\ &= H(x) + H(d) - H(y) - H(e) = H(x) + H(\frac{d}{\sigma_d}) - H(y) - H(\frac{e}{\sigma_e}) - \log\sigma_d + \log\sigma_e \\ &= H(x) + H(\frac{d}{\sigma_d}) - H(y) - H(\frac{e}{\sigma_e}) \end{aligned}$$

where the joint entropies $H(x,e)$ and $H(y,d)$ as well as the variances of the residuals (which are equal) cancel. Thus, our criterion is equivalent to evaluating the independence of x vs. d and y vs. e using mutual information, and choosing the direction in which the regressor is more independent of the residual.

Again, these developments show the important practical advantage that we only need to evaluate one-dimensional entropies, although the definition of mutual information contains a two-dimensional entropy.

1. Note that this is not the same as the simple heuristic approach in which we only compute the non-Gaussianities of the actual variables x, y and assume that direction must be from the more non-Gaussian variable to the less non-Gaussian one.

2.5 First-Order Approximation of Likelihood Ratio

Next we develop some simple approximations of the likelihood ratio. Our goal is to find causality measures which are simpler (conceptually and possibly also computationally) than the likelihood ratio or its general approximation given above.

Let us make a first-order approximation

$$G(\frac{y-\rho x}{\sqrt{1-\rho^2}}) = G(y) - \rho\, x\, g(y) + O(\rho^2)$$

where g is the derivative of G, and likewise for the regression in the other direction. Then, we get the approximation $\tilde{R}$:

$$R \approx \tilde{R} = \frac{1}{T}\sum_t G(x_t) + G(y_t) - \rho x_t g(y_t) - G(y_t) - G(x_t) + \rho y_t g(x_t) = \frac{\rho}{T}\sum_t -x_t g(y_t) + g(x_t) y_t .$$

Pham and Garrat (1997) proposed a method for estimating the derivatives of log-pdf's of random variables. Their method could be directly used for estimating g. However, since our main goal here is to find methods which work for small sample sizes, we try to avoid such estimation of the g's which has potentially a very large number of parameters. Instead, here we assume that we have some prior knowledge on the distributions of the variables in the model. In fact, a result well-known in the theory of ICA is that it does not matter very much how we choose the log-pdf's in the model as long as they are roughly of the right kind (Hyvärinen et al., 2001). This claim is partly justified by the cumulant-based analysis and simulations below.

In particular, very good empirical results are usually obtained by modelling any sparse (i.e., super-Gaussian, or positively kurtotic), symmetric densities by either the logistic density

$$G(u) = -2\log\cosh(\frac{\pi}{2\sqrt{3}}u) + \text{const.} \tag{4}$$

or the Laplacian density

$$G(u) = -\sqrt{2}|u| + \text{const.}$$

where the additive constants are immaterial. The Laplacian density is not very often used in ICA because its derivative is discontinuous at zero which leads to problems in maximization of the ICA likelihood. However, here we do not have such a problem so we can use the Laplacian density as well.

Thus, if we approximate all the log-pdf's by (4), we get the "non-linear correlation"

$$\tilde{R}_{\text{sparse}} = \rho\hat{E}\{x\tanh(y) - \tanh(x)\, y\} \tag{5}$$

where we have omitted the constant $\frac{\pi}{2\sqrt{3}}$ which is close to one, as well as a multiplicative scaling constant. Here, $\hat{E}$ means the sample average. This is the quantity we would use to determine the causal direction. Under $x \to y$, this is positive, and under $y \to x$, it is negative.

2.6 Cumulant-Based Approach

To get further insight into the likelihood ratio approximation in (5), we consider a cumulant-based approach which can be analyzed exactly. The theory of ICA has shown that cumulant-based approaches can shed light into the convergence properties of likelihood-based approaches. However,

cumulant-based methods tend to be very sensitive to outliers, so their utility is mainly in the theoretical analysis; for analysing real data, the measure in (5) is preferred.

Here, an approach based on fourth-order cumulants is possible by defining

$$\tilde{R}_{c4}(x,y) = \rho\hat{E}\{x^3y - xy^3\} \tag{6}$$

where the idea is that the third-order monomial analyzes the main nonlinearity in the nonlinear correlation. In fact, we can approximate tanh by a Taylor expansion

$$\tanh(u) = u - \frac{1}{3}u^3 + O(u^5). \tag{7}$$

Then, first-order terms are immaterial because they produce terms like $\hat{E}\{xy - xy\}$ which cancel out, and the third-order terms can be assumed to determine the qualitative behaviour of the nonlinearity.

Our main results of the cumulant-based approach is the following:

Theorem 1 *If the causal direction is $x \to y$, we have*

$$\tilde{R}_{c4} = kurt(x)(\rho^2 - \rho^4) \tag{8}$$

where $kurt(x) = E\{x^4\} - 3$ is the kurtosis of x. If the causal direction is the opposite, we have

$$\tilde{R}_{c4} = kurt(y)(\rho^4 - \rho^2). \tag{9}$$

Proof Consider the fourth-order cumulant

$$C(x,y) = \text{cum}(x,x,x,y) = E\{x^3y\} - 3E\{xy\}$$

where we assume the two variables are standardized. We have $\text{kurt}(x) = C(x,x) = \text{cum}(x,x,x,x)$. The nonlinear correlation can be expressed using this cumulant as

$$\tilde{R}_{c4} = \rho[C(x,y) - C(y,x)]$$

since the linear correlation terms cancel out. We use next two well-known properties of cumulants. First, the linearity property says that for any two random variables v, w and constants a, b we have

$$\text{cum}(v,v,v,av+bw) = a\,\text{cum}(v,v,v,v) + b\,\text{cum}(v,v,v,w)$$

and second, $\text{cum}(v,w,x,y) = 0$ if any of the variables v,w,x,y is statistically independent of the others. Thus, assuming the causal direction is $x \to y$, that is, $y = \rho x + d$ with x and d independent, we have

$$\begin{aligned}\tilde{R}_{c4} &= \rho[\text{cum}(x,x,x,\rho x+d) - \text{cum}(x,\rho x+d,\rho x+d,\rho x+d)] \\ &= \rho[\rho\text{cum}(x,x,x,x) + \text{cum}(x,x,x,d) \\ &\quad - \rho^3\text{cum}(x,x,x,x) - 3\rho^2\text{cum}(x,x,x,d) - 3\rho\text{cum}(x,x,d,d) - \text{cum}(x,d,d,d) \\ &= \rho[\rho\text{kurt}(x) - \rho^3\text{kurt}(x)] = \text{kurt}(x)[\rho^2 - \rho^4]\end{aligned}$$

which proves (8). The proof of (9) is completely symmetric: exchanging the roles of x and y will simply change the sign of the nonlinear correlation, and the kurtosis will be taken of y. ■

The regression coefficient ρ is always smaller than one in absolute value, and thus $\rho^2 - \rho^4 > 0$. Assuming that the relevant kurtosis is positive, which is very often the case for real data, the sign of $\tilde{R}_{c4}$ can be used to determine the causal direction in the same way as in the case of the likelihood approximation $\tilde{R}$ in (5). Thus, the cumulant-based approach allowed us to prove rigorously that a nonlinear correlation of the form (6) can be used to infer the causal direction, since it takes opposite signs under the two models. Note that this nonlinear correlation has exactly the same algebraic form as the likelihood ratio approximation (5); only the nonlinear scalar function is different. In particular, this analysis shows that the exact form of the nonlinearity is not important: the cubic nonlinearity is valid for all distributions of positive kurtosis.

If the relevant kurtosis is negative, a simple change of sign is needed. In general, we should thus multiply $\tilde{R}_{c4}$ by the sign of the kurtosis to obtain

$$\tilde{R}'_{c4}(x,y) = \text{sign}(\text{kurt}(x))\rho\hat{E}\{x^3y - xy^3\}.$$

Here, we get the complication that we have to choose whether we use the sign of the kurtosis of x or y. Usually, however, the signs would be the same, and we might have prior information on their sign, which is in most applications positive.[2]

Related cumulant-based measure were proposed by Dodge and Rousson (2001) and Dodge and Yadegari (2010). Their fourth-order measures used the ratio of marginal kurtoses, as opposed to the cross-cumulants we use here. They further assumed the disturbances to be Gaussian (or at least to have zero cumulants), which makes their measures less general than ours. In fact, their method relies on the fact that kurtosis is decreased by adding a Gaussian disturbance, but if the disturbance is much more kurtotic than the regressor, the opposite can be the case.

2.7 Intuitive Interpretations

Next, we provide some intuitive interpretations of the obtained first-order approximations of the likelihood ratio.

2.7.1 GRAPHICAL INTERPRETATION

The cumulants and nonlinear correlations have a simple intuitive intepretation. Let us consider the cumulant first. The expectations $E\{x^3y\}$ or $E\{xy^3\}$ are basically measuring points where both x and y have large values, but in contrast to ordinary correlation, they are strongly emphasizing large values of the variable which is raised to the third power.

Assume the data follows $x \to y$, and that both variables are sparse (super-Gaussian). Then, both variables simultaneously have large values mainly in the cases where x takes a large value, making y large as well. Now, due to regression towards the mean, that is, $|\rho| < 1$, the value of x is typically larger than the value of y. Thus, $E\{x^3y\} > E\{xy^3\}$. This is why $E\{x^3y\} - E\{xy^3\} > 0$ under $x \to y$. The idea is illustrated in Figure 1.

2. In the general case where the (real) kurtoses of x and y are allowed to have different signs, we need to compute two quantities: $\tilde{R}'_{c4}(x,y) = \text{sign}(\text{kurt}(x))\rho\hat{E}\{x^3y - xy^3\}$ and $\tilde{R}'_{c4}(y,x) = \text{sign}(\text{kurt}(y))\rho\hat{E}\{y^3x - yx^3\}$. According to the analysis above, the former quantity is positive if $x \to y$, and the latter quantity is positive if $y \to x$. However, in practice, this does not lead to a simple decision rule since due to finite sample size, or violations of the model, it could be that both of these quantities are positive, or none of them. In such cases, the decision rule should be defined so as to indicate that the causal direction could not be decided.

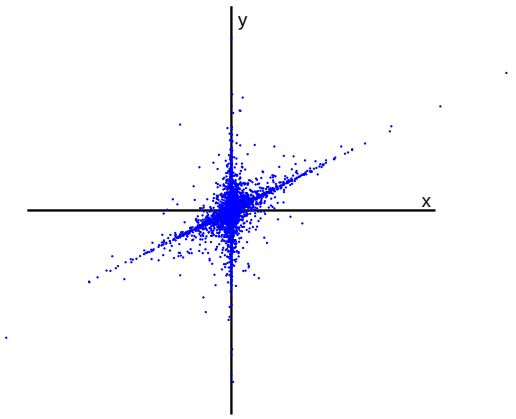

Figure 1: Intuitive illustration of the nonlinear correlations. Here, $x \to y$ and the variables are very sparse. The nonlinear correlation $E\{x^3y\}$ is larger than $E\{xy^3\}$ because when both variables are simultaneously large (the "arm" of the distribution on the right and the left), x attains larger values than y due to regression towards the mean.

This interpretation is valid for the tanh-based nonlinear correlation as well, because we can use the function $h(u) = u - \tanh(u)$ instead of tanh to measure the same correlations but with opposite sign. In fact, we have

$$\tilde{R}_{\text{sparse}} = \rho\hat{E}\{h(x)y - xh(y)\}$$

because the linear terms cancel each other. The function h is a soft thresholding function, and thus has the same effect of emphasizing large values as the third power. Thus the same logic applies for h and the third power.

2.7.2 INTERPRETATION AS IMPLICATION

Even if the data is not assumed to follow any particular model, the nonlinear correlation could be interpreted as a logical implication. In general, if the existence of event A implies the existence of event B, but there is no implication in the other direction, a causal influence from A to B might be inferred. Since $A \Rightarrow B$ is equivalent to $\neg B \Rightarrow \neg A$, there has to be some clear distinction between the events and their negations for this interpretation to be meaningful. We assume here that the events are rare, that is, have small probabilities.

Now, let us consider the events A, defined as "x takes a very large value" and B, defined as "y takes a relatively large value of the same sign as x". Notice that because the regression coefficient is smaller than one, we cannot require y to take particularly large values. It is assumed here that the thresholds for deciding when a value is large are chosen so that both of these events are rare.

To investigate implication, we can consider how to refute it. To refute $A \Rightarrow B$, we should consider cases where x takes a very large value but y takes a value of the opposite sign. This can be measured by $Ex^3(-y)$ where x^3 looks at large values of x and the minus sign changes this into a measure of how much large values of x coexist with y's of opposite sign.

Thus, $Ex^3y - Exy^3$ can be seen as measuring of how much evidence we have to refute $y \Rightarrow x$ (latter term) minus the evidence to refute $x \Rightarrow y$ (negative of first term). If it is large, we accept the implication $x \Rightarrow y$ together with its causal interpretation.

It might be argued that the connection between causality and implication could also plausibly be defined in the opposite direction: If A implies B as defined above, then B causes A. However, we shall now argue that the interpretation we gave above follows naturally from the definition of a SEM with two variables. Assume $x \to y$ and $\rho > 0$. If x is very large, y is likely to be large and of the same sign, since it is not very probable that d would cancel out the effect of ax. Thus, we have $A \Rightarrow B$ when x causes y under the SEM framework.

2.8 Noise-Tolerance of the Nonlinear Correlations

An interesting point to note is that the cumulant in (6) is, in principle, immune to additive measurement noise. Assume that instead of the real x, y, we observe noisy versions $\tilde{x} = x + n_1$ and $\tilde{y} = y + n_2$ where the noise variables are independent of each other and x and y. By the basic properties of cumulants (see proof of Theorem 1), the nonlinear correlations are not affected by the noise at all in the limit of infinite sample size. Thus, our method in not biased by noise. This is in stark contrast to ICA algorithms which are strongly affected by additive noise; thus ICA-based LiNGAM (Shimizu et al., 2006) would not yield consistent estimators in the presence of noise.

To be more precise, we have

$$\begin{aligned} E\{\tilde{x}^3\tilde{y}\} - E\{\tilde{x}\tilde{y}^3\} &= \mathrm{cum}(\tilde{x},\tilde{x},\tilde{x},\tilde{y}) - \mathrm{cum}(\tilde{x},\tilde{y},\tilde{y},\tilde{y}) \\ &= \mathrm{cum}(x,x,x,y) - \mathrm{cum}(x,y,y,y) = E\{x^3y\} - E\{xy^3\} \end{aligned}$$

due to the independence of n_1 and n_2 of the other variables and each other.

On the other hand, the estimation of ρ is strongly affected by the noise. This implies that $\tilde{R}_{c4}$ is not immune to noise. Nevertheless, measurement noise would only decrease the absolute value of ρ and not change its sign. Thus, the sign of $\tilde{R}_{c4}$ is not affected by additive measurement noise in the limit of infinite sample. This applies for both Gaussian and non-Gaussian noise.

The fact that the ρ is only a multiplicative scaling in the nonlinear correlations (6) or (5) must be contrasted with its role in the likelihood ratio (1) where its effect is more complicated. Thus, when ρ is underestimated due to measurement noise, it may have a stronger effect on the likelihood ratio, while its effect on the nonlinear correlations is likely to be weaker. While this logic is quite approximative, simulations below seem to support it.

On the other hand, the standardization of the variables is also affected by noise, in particular if the noise variances are not equal. As long as the noise variances are equal, the error in standardization will affect the measures by a multiplicative constant only, effectively making the cumulants smaller. Thus, the noise-tolerance of the cumulants may be useful in practice only if the variances of the noise variables are equal.

2.9 Skewed Variables

Above, the likelihood ratio approximations and cumulants were developed for sparse, typically symmetrically-distributed variables. Here, we consider the extension to skewed variables. Again, the underlying motivations is that if we know the distributions are skewed, we can use this prior knowledge to obtain particularly simple measures of causal direction. The cumulant-based analysis

is mainly for theoretical interest due to the sensitivity of cumulants to outliers; we provide a more robust nonlinearity for analysing real data.

2.9.1 CUMULANT-BASED APPROACH

The cumulant-based approach allows for a very simple extension of the framework to skewed variables. As a simple analogue to (6), we can define a third-order cumulant-based statistic as follows

$$\tilde{R}_{c3}(x,y) = \rho\hat{E}\{x^2y - xy^2\}. \tag{10}$$

The justification for this definition is in the following theorem, which is the analogue of Theorem 1:

Theorem 2 *If the causal direction is $x \to y$, we have*

$$\tilde{R}_{c3} = skew(x)(\rho^2 - \rho^3) \tag{11}$$

and if the causal direction is the opposite, we have

$$\tilde{R}_{c3} = skew(y)(\rho^3 - \rho^2). \tag{12}$$

Proof Consider the third-order cumulant

$$C(x,y) = \mathrm{cum}(x,x,y) = Ex^2y$$

where we assume the two variables are standardized. We have $\mathrm{skew}(x) = C(x,x) = \mathrm{cum}(x,x,x)$. The nonlinear correlation can be expressed using this cumulant as

$$\tilde{R}_{c3} = \rho[C(x,y) - C(y,x)].$$

Assuming the causal direction is $x \to y$, we have

$$\begin{aligned}\tilde{R}_{c3} &= \rho[\mathrm{cum}(x,x,\rho x+d) - \mathrm{cum}(x,\rho x+d,\rho x+d)] \\ &= \rho[\rho\,\mathrm{cum}(x,x,x) + \mathrm{cum}(x,x,d) - \rho^2\mathrm{cum}(x,x,x) - 2\rho\,\mathrm{cum}(x,x,d) - \mathrm{cum}(x,d,d)] \\ &= \rho[\rho\,\mathrm{skew}(x) - \rho^2\mathrm{skew}(x)] = \mathrm{skew}(x)[\rho^2 - \rho^3]\end{aligned}$$

which proves (11). The proof of (12) is again completely symmetric. ■

To use the measure (10) in practice, we have to take into account the fact that we cannot assume, in general, the skewnesses of the variables to have some particular sign. In some applications this is possible: For example, in resting-state fMRI data it might be safe to assume that the skewnesses are all positive because it is much more common that the signals obtain large values due to activation than due to inhibition (however, this point needs to be confirmed by empirical investigations of fMRI data).

In the general case, we propose that before computing these nonlinear correlations, the signs of the variables are first chosen so that the skewnesses are all positive. This can be accomplished simply by multiplying the variables by the signs of their skewnesses to get a new variable x^*

$$x^* = \mathrm{sign}(\mathrm{skew}(x))\,x \tag{13}$$

and the same for y (this transformation has to be done before computing ρ). Now, we have a situation similar to the previous measures: Under $x \to y$, $\tilde{R}'_{c3}(x,y) > 0$. This is because again, $|\rho| < 1$, and therefore $\rho^2 - \rho^3 > 0$ regardless of the sign of the coefficient. Likewise, for $y \to x$, $\tilde{R}'_{c3}(y,x) < 0$.

Our measure is related to the directionality measure proposed by Dodge and Rousson (2001), which in our notation would be:

$$\tilde{R}_{DR}(x,y) = [\hat{E}\{x^2 y\}]^2 - [E\{xy^2\}]^2 \tag{14}$$

which has the advantage of of being particularly simple, and does not require the skewnesses to be of any particular sign. However, our measure is more closely related to likelihood ratios which may give it some advantage in terms of statistical performance, as will be seen in the simulations below.

2.9.2 ROBUST, LIKELIHOOD-BASED APPROACH

The skewed case might also be approached by defining a skewed log-pdf and using the methods in previous sections. Unfortunately, in the theory of ICA, general-purpose skewed densities can hardly be found, and thus it is not clear how to define such densities and how generally they would be applicable. Nevertheless, a likelihood-based approach is likely to be more robust against outliers than the cumulant-based one (unless the model pdf has very light tails) which is why we develop one here.

We propose the following nonlinearity:

$$g_{\text{skew}}(x) = \log\cosh(\max(x,0)) \tag{15}$$

which can be justified as follows. Consider the following family of pdf's, defined using the derivative of the log-pdf

$$(\log p)'(x) = g_{\text{skew}}(x) - \beta x - \alpha \tag{16}$$

where β and α are parameters. Let us take α and β so that we get a standardized pdf with zero mean and unit variance. Numerical calculations show that this is obtained by values which are approximately $\alpha_0 = 0.188$ and $\beta_0 = 1.32$. The ensuing pdf is illustrated in Figure 2.

Further numerical calculations show that the higher-order cumulants of the standardized pdf are both positive: Skewness is approximately 0.37 and kurtosis 0.47.

Now, we can add any linear function and/or constant to $(\log p)'$ without changing the value of the approximative likelihood ratio in (5). In particular, using the true derivative of log-pdf in (16) is equivalent to using the algebraically simpler g_{skew}.

Thus, we obtain the following approximation for the likelihood ratio:

$$\tilde{R}_{skrb}(x,y) = \rho\hat{E}\{g_{\text{skew}}(x)y - xg_{\text{skew}}(y)\} \tag{17}$$

with g_{skew} defined in (15). Again, this applies for positively skewed variables only. If the skewnesses are not known a priori, they can be made positive by (13).

3. Estimating a Network with More Than Two Variables

In this section, we consider the general case of more than two variables. We present two approaches: First, we use the pairwise analysis in a DirectLiNGAM framework, and second, we present a two-stage method where the possible connections in a sparse graph are first pruned using covariance information.

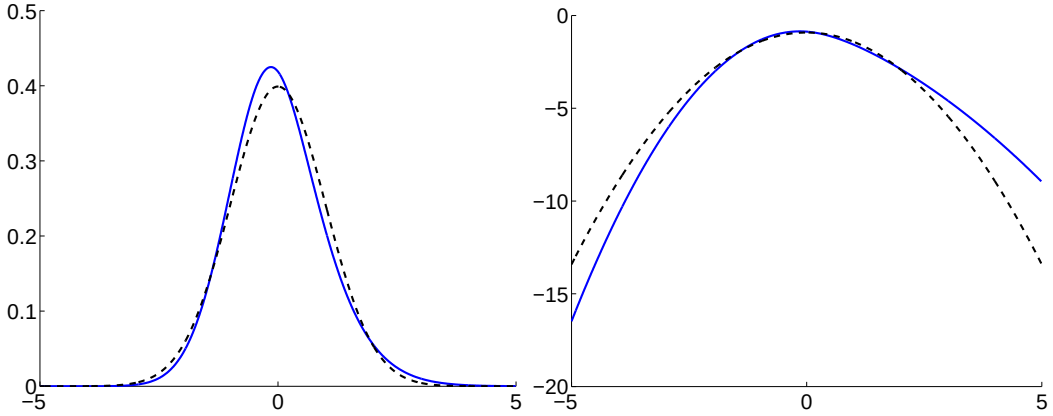

Figure 2: The pdf for robust modelling of skewed densities. Left: the pdf corresponding to the derivative of log-pdf in (16) is plotted (solid curve) with α and β chosen so that the density is standardized. For comparison, the Gaussian density of the same mean and variance is plotted as well (dashed). Right: the logarithms of the same density functions.

3.1 Model Definition

Denote by $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ the vector of observed variables. The linear non-Gaussian acyclic model (LiNGAM) proposed by Shimizu et al. (2006) can be expressed as

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}$$

where $\mathbf{e}$ is the vector of disturbances, and $\mathbf{B}$ is the matrix that describes the influences of the x_i on each other; the diagonal of $\mathbf{B}$ is defined to be zero.

It was shown by Shimizu et al. (2006) that the model is identifiable under the following assumptions: a) the e_i are non-Gaussian, b) the e_i are mutually independent, and c) the matrix $\mathbf{B}$ corresponds to a directed acyclic graph (DAG). It is well-known that the DAG property is equivalent to an existence of an ordering of the variables x_i (not necessarily unique) in which there are only connections "forward" in the ordering; if the variables are re-ordered according to the causal ordering, the matrix $\mathbf{B}$ has all zeros above the diagonal.

3.2 Using Pairwise Measures in the DirectLiNGAM Framework

The first way to use the pairwise analysis developed above to estimate LiNGAM which has more than two variables is to use the DirectLiNGAM framework (Shimizu et al., 2011).

3.2.1 FINDING ROOT OF GRAPH

In the DirectLiNGAM approach, we first compute the likelihood ratios of all different pairs of variables, and store the log-likelihood ratio for x_i and x_j as the (i, j)-th entry of a matrix $\mathbf{M}$. Alternatively, we can use the likelihood ratio approximations which can be all subsumed under the algebraic form

$$\mathbf{M} = \mathbf{C} \odot E\{\mathbf{x}g(\mathbf{x})^T - g(\mathbf{x})\mathbf{x}^T\} \tag{18}$$

where $\odot$ is element-wise multiplication. The nonlinearity g is typically chosen so that it is $g(u) = \tanh(u)$ for symmetric sparse data and $g(u) = -u^2$ or the function in (15) for skewed data. $\mathbf{C}$ is the covariance matrix of the data; since the data is assumed standardized $\mathbf{C}$ equals the matrix of correlation coefficients.

Now, for the variables x_i which have no parents, all entries in the i-th row of $\mathbf{M}$ are non-negative, neglecting random errors. (Note that there is no reason why there would be only one such "root" variable.) This was shown to be exactly true for the cumulant-based approaches $g(u) = -u^3$ and $g(u) = -u^2$ (assuming that the kurtoses or skewnesses, respectively, are positive) and is true as a first-order approximation based on (7) for $g(u) = \tanh(u)$. The reverse also holds if we assume faithfulness.[3]

Thus, we first find the row, say with index i^*, which is most likely to have all non-negative entries (the actual estimation procedure is considered below). Then, we regress ("deflate") the variable x_{i^*} out of all the other variables (Shimizu et al., 2011). We iterate this procedure by computing $\mathbf{M}$ again for the deflated $\mathbf{x}$. By locating the row which is most likely to have only non-negative entries in the newly computed $\mathbf{M}$, we thus find a variable which has no parents except for possibly the first variable found in the previous step. Repeating this, we find variables which are next in the partial order given by the DAG. Thus in the end we have the causal ordering of the variables.

After such estimation of the causal ordering, estimating the coefficients b_{ij} is easy by just ordinary least-squares estimation (Shimizu et al., 2006).

Alternatively, we could use a simple approximation which is very simple and computationally efficient. Instead of carrying out deflation by regression as described above, we simply remove the entries of the rows and columns corresponding to the already "found" variables in the matrix $\mathbf{M}$, and iterate the procedure. Thus, we obtain the causal ordering directly from a single matrix of nonlinear correlations, without any deflation. This is an approximation with no rigorous justification (because when removing the root we should also remove its effect on all the entries of $\mathbf{M}$) and it is likely to be inconsistent. However, in simulations reported below it works quite well. It has the benefit of being computationally extremely simple, and it gives a simple conceptual link between causal ordering and the nonlinear correlations and cumulants.

3.2.2 AGGREGATING PAIRWISE MEASURES

To use the method just described we have to solve the problem of aggregating the pairwise measures. We need to find the row which is most likely to be all non-negative up to random errors. Obviously, we could just take the sums of the entries in each row and locate the maximum sum but this is not likely to be optimal. So, we next develop a more principled way of aggregation.

Consider the $m_{ij}, j = 1,\dots,n$ for a fixed i, which are the estimates of pairwise likelihood ratios or some approximations. Assume they are independent and have Gaussian distributions $N(\mu_{ij}, \sigma^2)$, where the variances are assumed to be equal for simplicity. The variance σ^2 is the estimation error due to finite sample, and the μ_{ij} are the true values. The posterior of μ_{ij} given m_{ij} is then Gaussian

3. For a variable x_0 with no parents, any other variable is of the form $x_j = ax_0 + d$ where a expresses the total effect coming from x_0, and d is a sum of the inputs from other external influences, which are, by definition, independent of x_0. Thus, the pairwise model holds with a d independent of x_0 and the pairwise measure is non-negative. On the other hand, consider x_i which does have parents. Now, go backwards in the graph until you find a node x_0 which has no parents (in a DAG, such a variable is guaranteed to exist). According to the logic just given, we have $x_i = ax_0 + d$, again with an independent d. By faithfulness, $a \neq 0$. Since changing the direction simply changes the sign of our measures, there will be a negative entry in the i-th row, and it has to be non-zero.

with mean m_{ij} and variance σ. Thus, the posterior log-probability that all of the $\mu_{ij}, j = 1,\ldots,n$ are positive can be calculated as

$$\log \prod_j P(\mu_{ij} > 0 | m_{ij}) = \log \prod_j P(\frac{\mu_{ij} - m_{ij}}{\sigma} > -\frac{m_{ij}}{\sigma} | m_{ij}) = \sum_j \log \Phi(\frac{m_{ij}}{\sigma}) \tag{19}$$

where Φ is the cumulative distribution function of the standardized Gaussian distribution. Estimating σ is possible but we prefer to assume it is very small and make the following approximation:

$$\log \Phi(\frac{m_{ij}}{\sigma}) \approx -\frac{1}{2\sigma^2} \min(0, m_{ij})^2$$

which can be seen to be quite accurate by a simple numerical comparison, and avoids numerical problems in computing the logarithm of Φ for large negative values. Now, σ is simply a multiplicative scaling constant which can be ignored when comparing estimates of the log-probabilities in (19).

Thus, we propose the following way of aggregating the pairwise likelihood ratios. Compute for each row of $\mathbf{M}$

$$m_i = -\sum_j \min(0, [\mathbf{M}]_{ij})^2$$

which, intuitively speaking, punishes violations of the positivity. The index i^* with maximum m_i is thus taken as the estimate of a variable with no parents, that is, a first variable in the causal ordering.

3.3 Two-Stage Approach to Estimating a Sparse Model

If the matrix $\mathbf{B}$ is known to be sparse, we can use a two-stage method in which we first estimate the connections in an undirected sense, and then find their directions using our pairwise method. This two-stage method is interesting from the viewpoint of clearly dividing the estimating problem into two parts.

We first find undirected connections by using any known method for estimating a Gaussian undirected model (Spirtes et al., 1993). In the simplest case, this can be based on the inverse covariance matrix, or the precision matrix. As is well-known in the theory of Gaussian graphical models, there is an intimate connection between the non-zero entries in the precision matrix and the existence of connections in the SEM—although the connection is not quite simple, especially for directed graphs. In contrast, the direction of a connection cannot be easily determined from the covariances, and is often unidentifiable, which was of course the original motivation for introducing non-Gaussian models (Shimizu et al., 2006). Nevertheless, as a first approximation, we can prune the set of candidate connections using the inverse covariance matrix, and apply our pairwise analysis only on those connections which this covariance-based analysis indicates to be present.

In an estimated inverse covariance matrix, there are of course no exact zeros. Thus, we use bootstrapping to test if each entry is non-zero. That is, we draw bootstrap samples of the data, and compute the inverse covariance for each such sample. The ratio of the mean and the standard deviation of the bootstrap estimates of any given entry is then compared with the relevant quantile of a standardized Gaussian distribution.[4] The test is made separately for each non-diagonal entry of the inverse covariance matrix.

4. In the simulations below, we also tried methods for sparse estimation of the inverse covariance matrix. However, we found that this simple testing procedure works by far the best. The sparse inverse estimation methods are, in fact, developed for the case of a very large number of variables, and thus may not be useful in our simulations where we typically have 5-10 variables only.

Depending on the goal of the analysis, it may or may not be necessary to do corrections for multiple testing. If we do such corrections, we can actually claim that the connections found are statistically significant. However, this is obtained at the cost of a large number of false negatives. On the other hand, if we simply consider the existence of the connections as another set of parameters to estimate, it may be more advantageous not to make such corrections to reduce the overall error rate. In fact, a false negative (setting an existing connection to zero) could be considered quite a serious error in this context, so we prefer to use a rather large α. In the simulations below, we thus set the false positive rate $\alpha = 0.01$ with no correction for multiple testing. Such corrections will of course be needed if our aim were to claim that a particular connection exists, but if our goal here is mainly the inference of the causal ordering, some false positives should not matter since they are likely to correspond to small values of the coefficients anyway.

Then, for each of those significantly non-zero connections, we determine the direction of causality using our pairwise tests. There is no need to do any kind of deflation anymore. If we want to convert the obtained estimates into a total ordering of the variables, we input those connections which were not pruned to the ordering method presented by Shimizu et al. (2006).

4. Estimating Cyclic Models

An important generalization of the DAG framework would be to estimate cyclic models. Here, we assume the following well-known generative model for the data. First, the external influences arrive in the system at time $t = 0$

$$\mathbf{x}(0) = \mathbf{e}$$

where $\mathbf{x}(t)$ is value a hypothetical dynamic system at time point t. Then, at subsequent time steps, the external influence is completemented by feedback as

$$\mathbf{x}(t+1) = \mathbf{e} + \mathbf{B}\mathbf{x}(t)$$

where the matrix $\mathbf{B}$ has zero diagonal, which means we do not allow self-loops. Assuming that $\mathbf{B}$ is stable in the sense that its largest eigenvalue is smaller than one in absolute value, we have in the limit

$$\mathbf{x} = \sum_{k \geq 0} \mathbf{B}^k \mathbf{e} = (\mathbf{I} - \mathbf{B})^{-1} \mathbf{e}$$

and thus

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e} \tag{20}$$

where $\mathbf{B}$ is now allowed to be cyclic. This gives a simple interpretation of a model of the form (20) in the case where $\mathbf{B}$ is allowed to be cyclic. As above, the e_i are assumed independent and non-Gaussian.

In fact, estimation of such a model by ICA is possible if $\mathbf{B}$ is small enough, namely if all its entries are smaller than one in absolute value. Then, it is possible to estimate the model even by ICA, since after estimating ICA, we can find the right permutation of the components based on putting the largest entries of each row in the diagonal. Thus, the model is identifiable under these assumptions. This is shown in detail in the following Theorem:

Theorem 3 *Assume that the data follows the cyclic LiNGAM model in (21) with no self-loops. Assume further that all the entries in the matrix* $\mathbf{B}$ *have absolute values smaller than one, and that*

the dominant eigenvalue of $\mathbf{B}$ *is smaller than one in absolute value. Then, the model is uniquely identifiable, that is, the matrix* $\mathbf{B}$ *can be estimated from the data without any ambiguity.*

Proof: The data actually follows the ICA model as

$$\mathbf{x} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{e}. \tag{21}$$

The ICA model is known to be identifiable up to a) the ordering of the components and b) a scalar multiplier for each of the components (Comon, 1994). The unidentifiability of the scalar multiplier disappears here because by definition, the diagonal of the inverse of the mixing matrix has all ones due to the diagonal matrix in (21). Thus, it was shown by Shimizu et al. (2006) that this implies the idenitifiability of the LiNGAM model if we can solve the indeterminacy of the permutation. Acyclicity was used for this purpose by Shimizu et al. (2006). Here, we use the assumption of absolute values smaller than one. In fact, consider the estimate of the inverse of the mixing matrix. Normalize it by dividing each row by its maximum element. Then, it equals $\mathbf{I} - \mathbf{B}$ up to a random permutation of the rows. Due to our assumption of $\mathbf{B}$, all non-diagonal entries in this matrix are smaller than one in absolute value. Thus, the original (correct) permutation of the rows can be found by locating on each row i the unique entry which is equal to one. Denoting its column index by $j(i)$, the original matrix is given by permuting the rows of the matrix to the ordering given by $j(i)$, that is, the ordering which puts the ones in the diagonal. □

This also suggests that we can estimate the model using the sparse graphs idea above. We prune the inverse covariance matrix to find where there are (probably) connections, and then find the directions of the connections using our pairwise measures. Using pairwise connections makes sense if we further assume that there are no pairwise loops, that is, connections $x_i \to x_j$ and $x_j \to x_i$ are not both non-zero. The main justification for this approach is that since the connections are weak, one can assume that the cyclicity has little effect on local pairwise measures. However, an exact convergence of such a method to the right parameter values does not seem possible to show in general.

5. Estimation in Case of Nonlinear Relations

In this Section, we generalize our method to a nonlinear model.

5.1 Definition of Nonlinear Model

Another interesting extension of the linear causal models is obtained by considering nonlinearities instead of non-Gaussianities (Hoyer et al., 2009). We define the two models as follows. The first one, $x \to y$, is given by

$$y = f(x) + d$$

where f is a nonlinear function, not necessarily invertible or even differentiable. The disturbance d is again independent of x. Both x and y are standardized to unit variance. The second model is denoted by $y \to x$ and defined as

$$x = g(y) + e$$

where g is another nonlinear function, and e is a disturbance variable. Other approaches to inferring the causal direction with nonlinear relations were introduced by Zhang and Hyvärinen (2009), Daniušis et al. (2010) and Mooij et al. (2010).

5.2 Likelihood Ratio for Nonlinear Model

The likelihood of the model $x \to y$ can be obtained as the sum of the log-prior of the variable x and the log-likelihood of the residual:

$$\log p(x,y) = \log p_x(x) + \log p_d(y - f(x)) = G_x(x) + G_d(\frac{\hat{d}}{\sigma_d}) - \log \sigma_d$$

where we denote, like above, the variance of the standardized residual by σ_d^2, the log-pdf of the standardized residual by G_d, and the log-pdf of x by G_x. Thus, like in the linear case, we obtain

$$R \quad = \quad \left[\frac{1}{T}\sum_t G_x(x_t) + G_d(\frac{y_t - f(x_t)}{\sigma_d}) - G_y(y_t) - G_e(\frac{x_t - g(y_t)}{\sigma_e})\right] \quad - \quad \log \sigma_d \quad + \quad \log \sigma_e. \quad (22)$$

An important difference to the linear case is that the variances of the residuals need not be equal, $\sigma_d \neq \sigma_e$, so they do not cancel. In an information-theoretic formulation, we obtain asymptotically

$$R \longrightarrow -H(x) - H(\hat{d}/\sigma_d) + H(y) + H(\hat{e}/\sigma_e) - \log \sigma_d + \log \sigma_e. \quad (23)$$

We can approximate R using the same maximum entropy approximations (Hyvärinen, 1998) as in the linear case in Section 2.3. The only difference is that we need to add the log-variances of the residuals to the expression. Thus, an important advantage of our approach is that we do not need any measures of independence per se; estimation of one-dimensional differential entropies is sufficient.

On the other hand, it may be advantageous to adapt the approximation to the nonlinear case. First, it does not seem useful to consider the prior non-Gaussianities of the variables, since a nonlinear mixing can change non-Gaussianities in completely unpredictable ways. This is unlike in the case of ICA, where a linear mixing decreases non-Gaussianity. Second, we can assume that the residuals tend to be sparse, and model them as Laplacian. This has the further advantage of making the measure more robust to outliers.

Now, for a Laplacian variable, the scale parameter σ is most naturally estimated as the mean absolute deviation (MAD), which is the maximum likelihood estimate. If we plug this estimate in the likelihood ratio, and omit the priors on x and y, we have

$$R = \left[\frac{1}{T}\sum_t -\frac{\sqrt{2}|y_t - f(x_t)|}{\hat{\sigma}_d}) + \frac{\sqrt{2}|x_t - g(y_t)|}{\hat{\sigma}_e}\right] - \log \hat{\sigma}_d + \log \hat{\sigma}_e$$

$$= -\sqrt{2}\frac{\hat{\sigma}_d}{\hat{\sigma}_d} + \sqrt{2}\frac{\hat{\sigma}_e}{\hat{\sigma}_e} - \log \hat{\sigma}_d + \log \hat{\sigma}_e$$

which gives finally the following objective

$$\tilde{R}_{\text{mad}} = -\log \hat{E}\{|\hat{d}|\} + \log \hat{E}\{|\hat{e}|\} \quad (24)$$

where $\hat{E}$ denotes the sample average, and thus $\hat{E}\{|.|\}$ denotes the MAD. In other words, we have an objective which simply compares the mean absolute deviations in the two cases.

The likelihood ratio depends on the estimated nonlinearities f, g. The estimation of f and g can be done with classic least-squares estimation methods independently of any developments in this paper. A large number of non-parametric methods have been developed in the literature, see, for example, Hoyer et al. (2009) for an example.

5.3 Connection to Independence-Based Nonlinear Methods

In fact, our method has a close connection to the independence-based method by Hoyer et al. (2009), generalizing the connection shown in Section 2.4. Using basic information-theoretic properties, we have under $x \to y$

$$H(x,y) = H(x) + H(y|x) = H(x) + H(y - f(x)|x) = H(x) + H(d|x) = H(x,d)$$

and likewise, this is equal to for $H(y,e)$. Now, just like in the linear case, we can consider the difference between the mutual informations of the regressors and residuals in the two directions, and obtain

$$I(x,d) - I(y,e) = H(x) + H(d) - H(y) - H(e) = H(x) + H(\frac{d}{\sigma_d}) - H(y) - H(\frac{e}{\sigma_e}) + \log\sigma_d - \log\sigma_e$$

where two terms equal to $h(x,y)$ cancel. Here, we see that asymptotically, our objective derived from the likelihood ratio is equal to the difference of the two mutual informations (with sign reversed). Its sign tells which mutual information is larger, and in particular, in which direction the residual of the regression is more independent. Thus, using the likelihood ratio is equivalent to using mutual information as independence measure in the methods by Hoyer et al. (2009).

The developments given above thus show that when comparing independencies of the residuals like Hoyer et al. (2009), it is not necessary to explicitly estimate mutual information; estimation of one-dimensional entropies leads to an equivalent result.

6. Simulations

We conducted simulations comparing the different methods proposed in this paper, as well as previously proposed LiNGAM estimation methods. In all the simulations, we emphasize difficult conditions. In most of the simulations, this means the case where the number of observations is small; the exception being the simulations with added measurement noise. We also take weakly non-Gaussian disturbances according to the logistic distribution in Equation (4), with the same aim of simulating difficult conditions.

The methods were compared with three previously published methods:

- LiNGAM estimated using ICA, as proposed in the original paper introducing LiNGAM by Shimizu et al. (2006).[5]
- DirectLiNGAM, specifically the kernel-based version proposed by Shimizu et al. (2011).[6]
- In case of skewed data, we used the measure proposed by Dodge and Rousson (2001), given in Equation (14).

The LiNGAM methods were implemented using the software found on the authors' web sites.

We computed different performance indices for the methods. For acyclic models, we computed

5. Since basic FastICA, which is an integral part of the method, has convergence problems with the basic tanh nonlinearity in the case of a small sample size, we used the stabilized version by Hyvärinen (1999) obtained in the standard FastICA package by the option "stabilize". For skewed data, we used the skewness as a measure of non-Gaussianity.
6. We did not include the earliest version of DirectLiNGAM proposed by Shimizu et al. (2009) in the comparison because in later simulations by Sogawa et al. (2010); Hyvärinen (2010), its performance was found clearly inferior to that of the kernel-based version of DirectLiNGAM.

1. The Spearman rank-correlation coefficient between the causal ordering given by the method and the true ordering.

2. The percentage of connections for which a method correctly estimated the direction, considering only connections existing in the data-generating process. Here, the point was to look at the abilities of the methods to find the directions locally, and thus the global ordering given by the method was *not* used (except for DirectLiNGAM which essentially only computes a global ordering and derives local ordering from that). For the ICA-based LiNGAM, we computed the measure $\text{sign}(|b_{ij}| - |b_{ji}|)$ and used it in the same way as the signs of the pairwise measures.

3. The percentage of data sets for which a method correctly estimated the first variable in the causal ordering, that is, the variable with no parents.

For cyclic models, the comparison was based on the second measure only, since the other two are not well-defined. Furthermore, we computed the CPU time needed for the computations.

Unless otherwise mentioned, the connection matrices were generated completely randomly, giving a fully connected DAG. The non-zero coefficients in the acyclic **B** had a uniform distribution in the union of the intervals $[-0.6,-0.2]$ and $[0.2,0.6]$.

6.1 Simulation 1: Sparse Influences

In the first, basic simulation, sample size and data dimension were varied so that there were in total four different scenarios:

1. $n = 5$, $T = 100$, fully connected DAG

2. $n = 2$, $T = 100$, fully connected DAG

3. $n = 5$, $T = 200$, fully connected DAG

4. $n = 5$, $T = 400$, fully connected DAG

The disturbances had logistic distributions, with standard deviations equal to one. 2,000 data sets were generated for each scenario; however, for DirectLiNGAM and ICA-LiNGAM only 1,000 were used due to excessive computational demands.

To estimate the model, we used the following methods proposed above. First, the maximum entropy approximation to the likelihood ratio in (3) was used in DirectLiNGAM with deflation. Second, the LR first-order approximation matrix (18) was used in DirectLiNGAM with the nonlinearity $g(u) = \tanh(u)$ and with deflation. Third, the nonlinear correlations in (18) were used to estimate the causal ordering without any deflation, simply by locating the minimum of the row sums of that matrix, removing the corresponding rows and columns, and so on, as described at the end of section 3.

See Figure 3 for the results. Typically, the tanh-based likelihood ratio approximation ("tanh") with deflation was the best. The method without deflation ("nodf") gives, by definition, the same result for the total causal directions correct and first variable found, but looking at the rank-correlations, we see that it is typically the second best. The maximum entropy approximation is usually the third best. ICA-based LiNGAM is usually fourth but when there is more data, it can have very good

performance. The (kernel-based) DirectLiNGAM ("kdir") is typically last, although not necessarily worse than ICA-based LiNGAM.

Regarding computational load, the methods proposed here are one to two orders of magnitude faster than the others.

6.2 Simulation 2: Sparse Influences with Noise

In the second simulation, we tested the noise-tolerance of the algorithms. The data dimension was varied from $n = 2$ to $n = 8$ and fully connected DAGs were used as above. The sample size was set to $T = 10,000$, which means we are now analyzing the statistical consistency[7] of the method only and neglecting random effects by taking a very large sample size. The noise was Gaussian and had unit variance. The performance indices and algorithms are as in the first simulation. The results are shown in Figure 4. We can see that the tanh-based approximation is clearly the best, as predicted by our cumulant-based analysis. ICA-based LiNGAM, the maximum entropy approximations, and especially kernel-based DirectLiNGAM seem to be more sensitive to noise.

6.3 Simulation 3: Skewed Influences

In the third simulation, we tested the performance of the methods with skewed data. We used the nonlinear correlation based on the third order cumulant ("skew"), introduced in Section 2.9, as well as the robust measure in Equation (15), denoted by "skw2".

We used two different skewed distributions for the disturbances. In both cases, the data was obtained from a Gaussian mixture. One of the Gaussian distributions in the mixture had zero mean and unit variance, while the other had mean equal to three and unit variance. The two distributions we generated were distinguished by the amount of data points drawn from the two Gaussians. In the first case ("pdf 1"), the "outlying" distribution with mean three generated 20% of the data, while in the second case ("pdf 2"), it generated only 5%. Thus, pdf 2 was quite sparse whereas pdf 1 was not. We would then expect sparsity-based methods to work well with pdf 2 but not very well with pdf 1. The data dimension were to $n = 2, n = 5$ and sample sizes $T = 100, 200$, respectively. DAGs were generated to be fully connected.

The results are shown in Figure 5. We see that all the methods have very similar performance, except the Dodge-Rousson measure which was somewhat worse. However, the computational loads are very different, our two likelihood ratio approximations being faster than the earlier LiNGAM methods by at least an order of magnitude.

6.4 Simulation 4: Skewed Influences with Noise

We further conducted a simulation with observational noise added to the skewed data. Again, we fixed the sample size to $T = 10,000$ and the noise variance to two (larger than above since these methods seem to be more tolerant to Gaussian noise), while the dimension and the skew data distribution were varied. We used only the skewed and sparse pdf 2. The results are in Figure 6. Here, we start seeing clear differences in the statistical performances of the methods. In line with our theoretical analysis, the skewness cumulant-based method is the most resistant to noise. The robust skewed LR approximation in Section 2.9.2 is second.

7. That is, convergence in the limit of infinite sample.

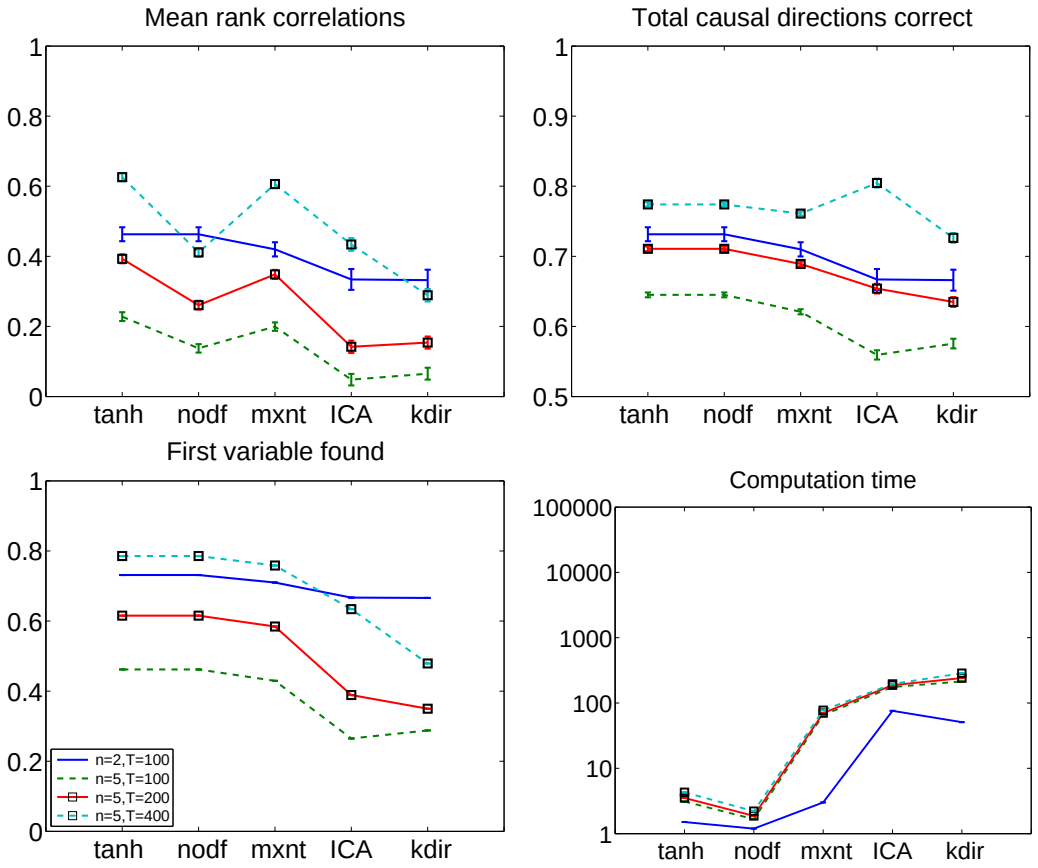

Figure 3: Simulation 1. Results of basic simulation with sparse, non-skewed data without noise. Top left: Mean of rank-correlation coefficients between the estimated causal ordering and the true ordering. The error bars are standard errors of the mean. Top right: The proportion of (really existing) connections for which the method estimated the direction correctly (chance level is 50%). Bottom left: The proportion of data sets for which the method estimated the first variable in the causal ordering correctly, that is, the variable with no parents. Bottom right: Computation times of one run of the different algorithms in milliseconds; note the logarithmic scale. Different colours are different data-generating scenarios. The algorithms used are as follows:
"tanh": LR approximations in (18) based on tanh nonlinearity, combined with deflation in DirectLiNGAM;
"nodf": no deflation in likelihood ratio approximations, that is, ordering based on the LR approximation matrix in (18) without any recomputation of the matrix;
"mxnt": maximum entropy approximation in (3) for likelihood ratios;
"ICA": LiNGAM estimated by ICA;
"kdir": kernel-based DirectLiNGAM.

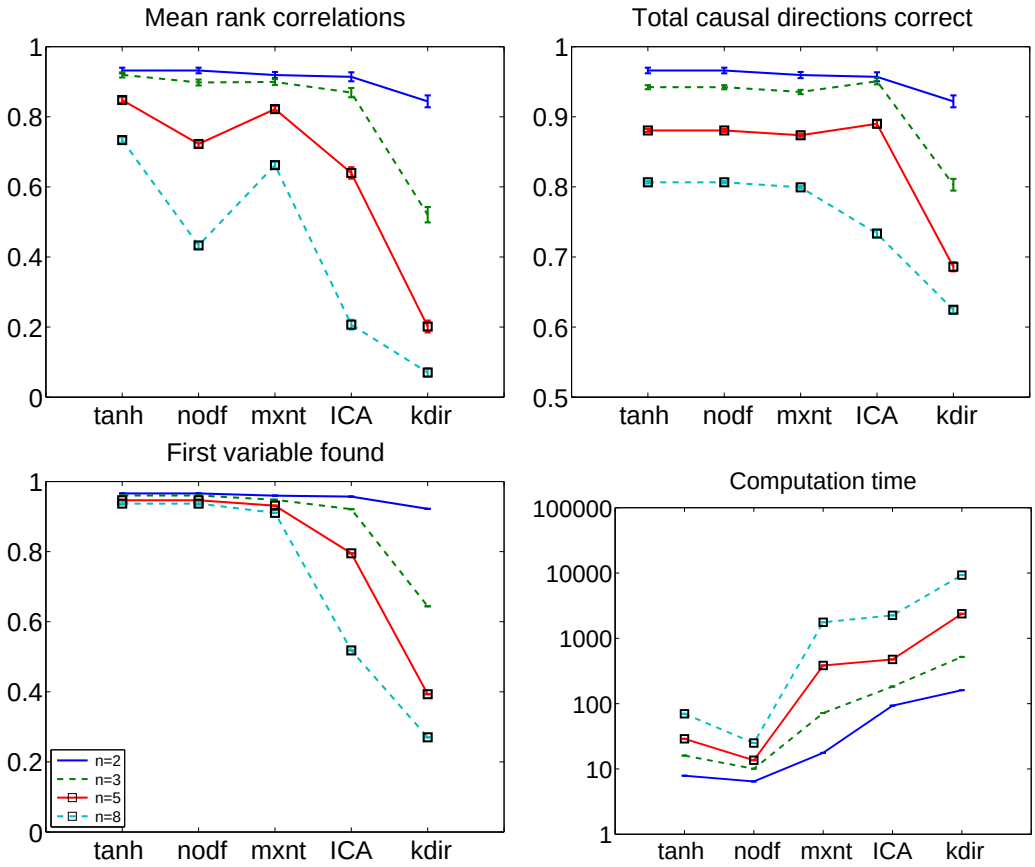

Figure 4: Simulation 2, with noise. Legend as in Figure 3, and with $T = 10,000$. The noise standard deviations were all equal to one.

6.5 Simulations 5 and 6: Two-Stage Approach and Sparse Graphs

Next we investigated the utility of the two-stage approach of Section 3.3. We generated sparse graphs only. The graphs were based on a simple "serial" structure $x_1 \to x_2 \to \ldots \to x_n$ with a random connection strength in the same range as above. We further added 0, 1, or 2 connections in random locations in the graph (preserving the DAG structure), the number of connections having equal probabilities for the three values. The data sizes were 500, 900, 900, 900 and the number of variables 5, 9, 15, 20, respectively. We used higher dimensions than above because otherwise the networks could not be very sparse. In the testing for the existence of connections, we set the false-positive rate to $\alpha = 0.01$ without correction for multiple testing, as motivated above.

In Simulation 5, we used sparse, non-skewed (logistic) influences, and in Simulation 6, skewed influences as in Simulation 3. To add more realism to the simulations, we also added noise to the data. The noise standard deviations were 0.2 in Simulation 5 and 0.6 in Simulation 6.

The results for Simulation 5 are shown in Figure 7. We can see that the two-stage method has a performance which compares quite favourably with the other methods: ICA-LiNGAM and DirectLiNGAM perform quite badly with these combinations of sample size and dimension. Note that

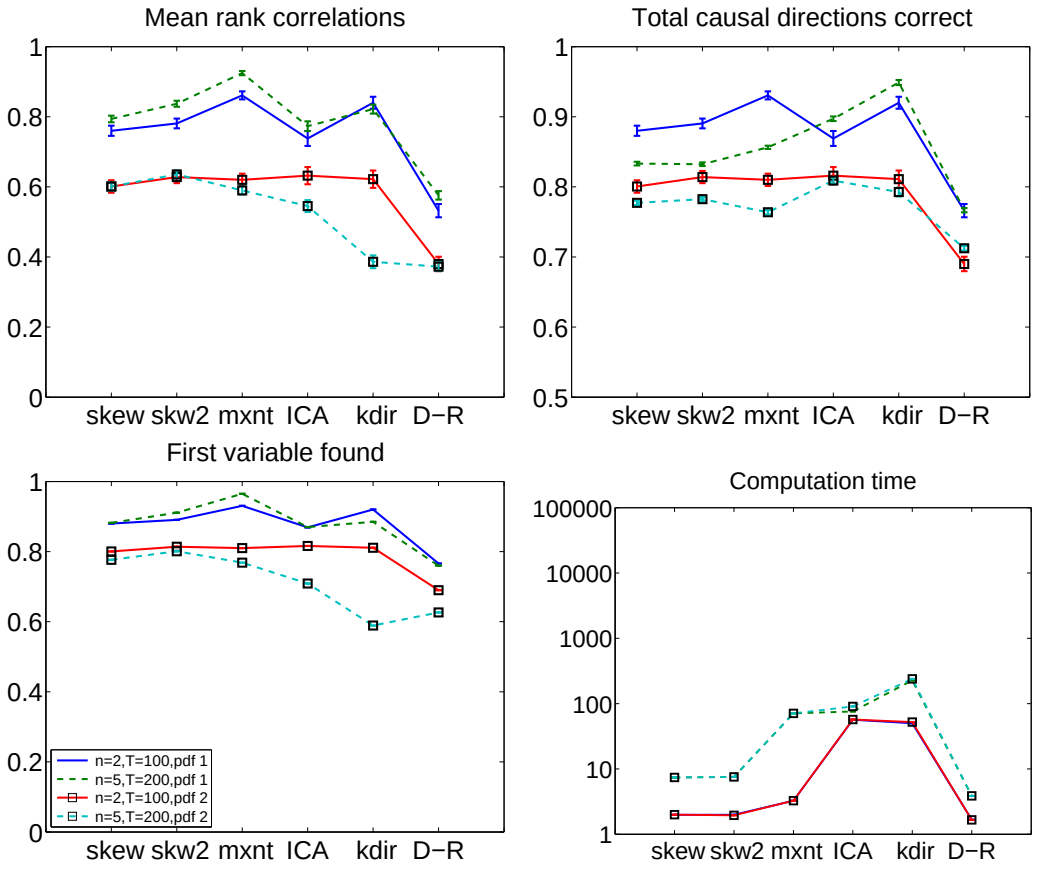

Figure 5: Simulation 3, with skewed data. Legend as in Figure 3, with the following new algorithms:
"skew": cumulant-based LR approximation in (10), combined with deflation in DirectLiNGAM;
"skw2": the robust LR approximation proposed in Section 2.9.2; and
"D-R": the measure by Dodge and Rousson (2001).

for "total causal directions correct", the two-stage method has, by definition, the same performance as "tanh" and "nodf". In fact, if our interest in only to discover the directions without bothering to estimate which variables are connected, or we are given perfect prior knowledge on which variables are connected, there is in fact no need to do the pruning in the first stage of the method.

So, the utility of the new method ("icth") is mainly seen in the mean rank correlations plot: There is a modest improvement. The point here is that knowledge of which variables are connected improves the estimation of the causal ordering (DAG structure) by showing which directionalities should be used when pooling their information together, and which directionalities should be discarded (because the variables are not connected at all).

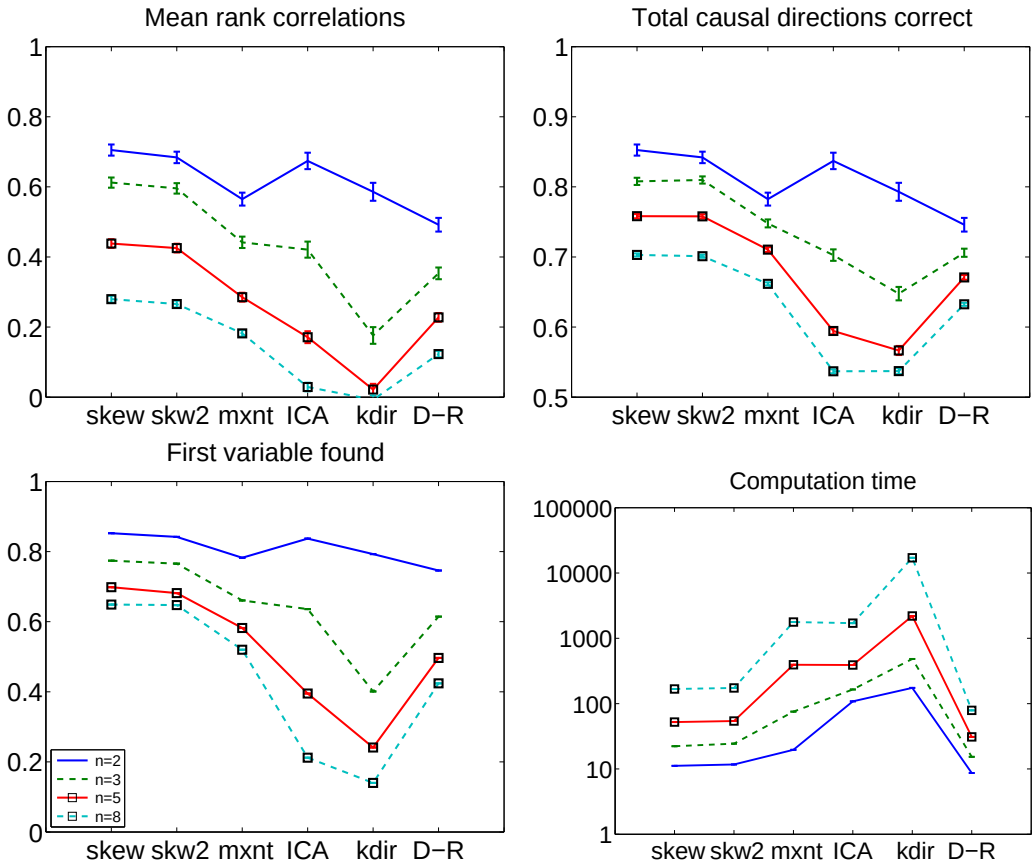

Figure 6: Simulation 4, with skewed data with noise. Legend as in Figure 5.

Interestingly, all the methods proposed in this paper are clearly superior to the methods proposed earlier (ICA-based LiNGAM and kernel-based DirectLiNGAM). Thus, the main utility of the present framework may indeed be in estimating directionality in sparse networks.

We carried out the same simulation for skewed influences using the skewed pdf 1. Results are in Figure 8. When looking at methods using the same causality measure ("skew" vs. "icsk", and "skw2" vs. "ics2"), we see that the pruning methods are better in terms of the mean rank correlations. However, the maximum entropy method without pruning is actually the best.

6.6 Overview of Simulations 1–6

To provide a succinct overview of the simulations reported above, we averaged the performance indices over the different scenarios (taking into account only scenarios in which the algorithm took part). Furthermore, we divided the simulations into three groups: basic data (simulations 1 and 2), skewed data (simulations 3 and 4) and sparse connections either with sparse data (simulation 5) or skewed data (simulation 6). We further averaged the performance indices inside these groups.

The results are shown in Figure 9.

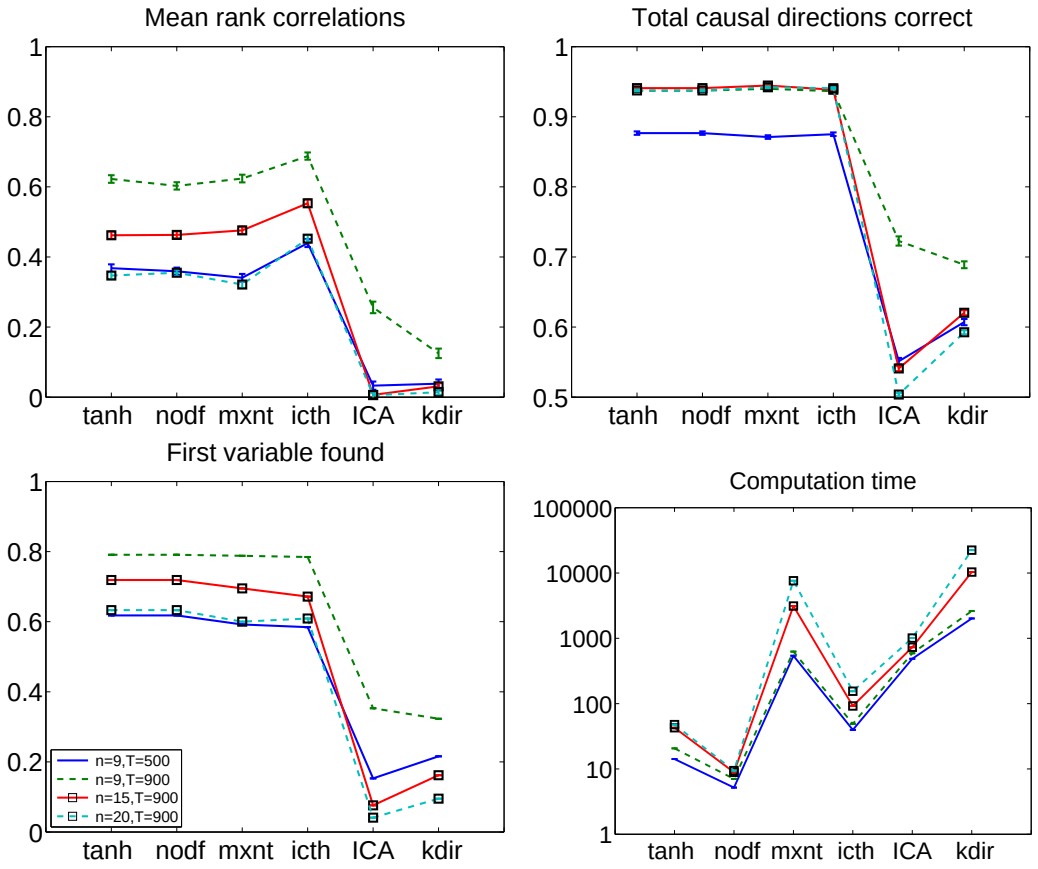

Figure 7: Simulation 5, with the two-stage pruning method and only sparse graphs. Legend as in Figure 3, but now including the new algorithm "icth" which prunes the graph based on inverse covariance and then estimates the directions using the same method as "tanh". (Note that only "icth" uses information on the pruned inverse covariance, other methods are as in Simulation 1.)

6.7 Simulation 7: More Variables than Observations

Next, we considered the case where there are more variables than observations, or at least the number of variables is equal to the number of observations. We considered four scenarios, with n ranging from 100 to 200 and T ranging from 100 to 400. In preliminary simulations, it turned out that the problem was too difficult for logistic disturbances, so we used Laplacian disturbances here.

We only attempted to estimate the first two variables and not the whole causal ordering. The very first variables in the causal ordering can be considered to be the exogenous ones and thus finding them is of special interest (Sogawa et al., 2011). We only used three of the new proposed methods because none of implementations of the existing LiNGAM methods was such that it could readily be used for this case.

The results are shown in Figure 10. While the performance of the methods is not very good, it is very much above chance level (which would be 0.01 or less for finding the first variable). It is

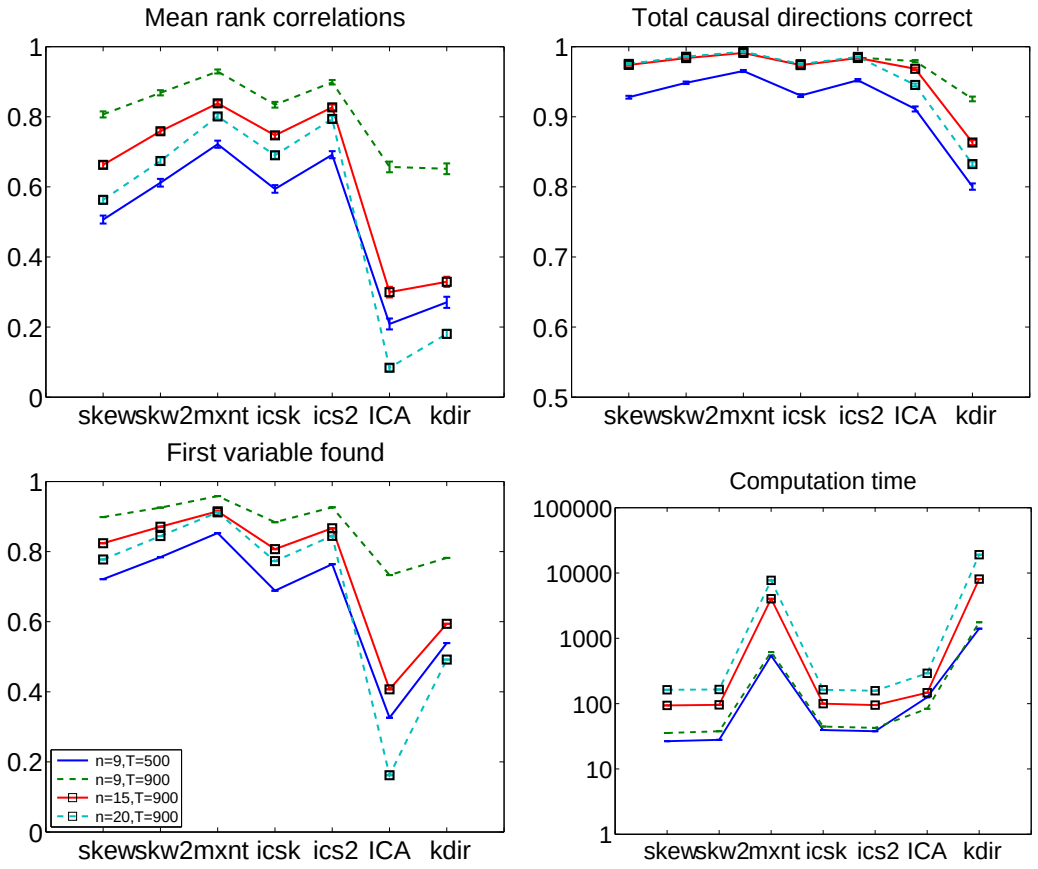

Figure 8: Simulation 6, with skewed data, the two-stage pruning method and only sparse graphs. Legend as in Figure 5, but now including the new algorithm "icsk" which prunes the graph based on inverse covariance and estimates the directions based on the skewness cumulant, and "ics2" which uses the robust skewness measure.

interesting to note that here the first-order approximation of likelihood is more than 100 times faster than the maximum entropy approximation.

6.8 Simulation 8: Cyclic Graphs

To test the new framework in the case of cyclic graphs, we created cyclic graphs by a simple ring structure: $x_1 \to x_2, \ldots, x_n \to x_1$. Further connections (0, 1, or 2) were added in random locations as in Simulation 5 above. Such data were created according to the generating model in Section 4. We further added noise with standard deviation 0.2. The dimensions of the data and the sample sizes were as in Simulations 5 and 6. The influences had logistic distributions.

The only methods we compared were ICA-based LiNGAM and our two-stage pruning methods, since the DirectLiNGAM methods cannot be used in the cyclic case. The results are shown in Figure 11. The first observation is that both methods performed relatively well, obtaining 70%-90% percent of the directions right. Our new method is slightly better than ICA-based LiNGAM.

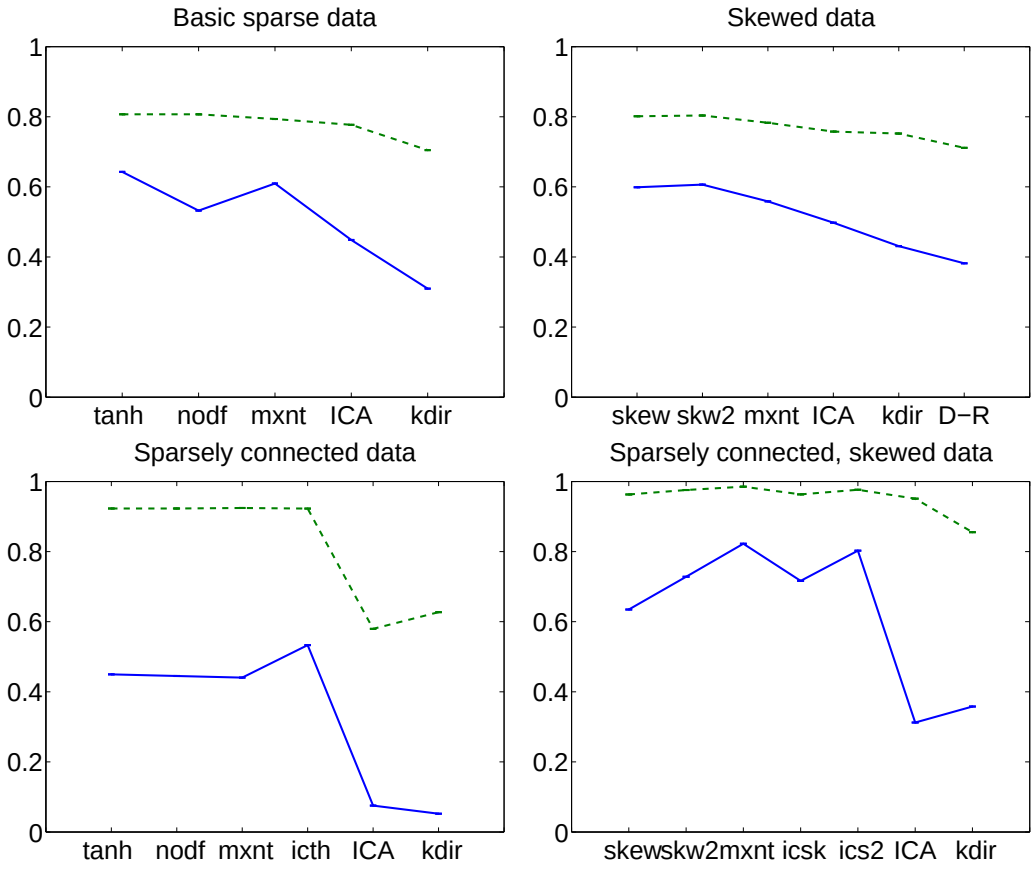

Figure 9: Overview of Simulations 1–6. Median correlations (blue, solid) and average directions correct (green, dashed) are plotted averaged over different scenarios and similar simulations.

It should be emphasized here that our method assumes that there are no self-loops, so there is no indeterminacy in the results, as shown in Section 4.

6.9 Simulation 9: Nonlinear Relations

Finally, we performed simulations on the nonlinear model. We generated data from a model

$$x_2 = \alpha \text{sign}(x_1)|x_1|^\gamma + d \tag{25}$$

where both x_1 and d were standardized Gaussian. The exponents γ were given values 0.5 and 2, and the parameter α was randomly drawn between 0.5 and 1.5. The sample sizes were either $T = 200$ or $T = 500$.

We then fitted the nonlinearity of the same functional form (25), that is, using the parameters α and γ, to the data with a least-squares fit, and estimated the causal direction using the criterion in (22), or the criterion in (24). (Thus, we did not use a nonparametric model of the nonlinearity. See Section 8 for estimation with non-parametric nonlinearities.) For comparison, we used the methods

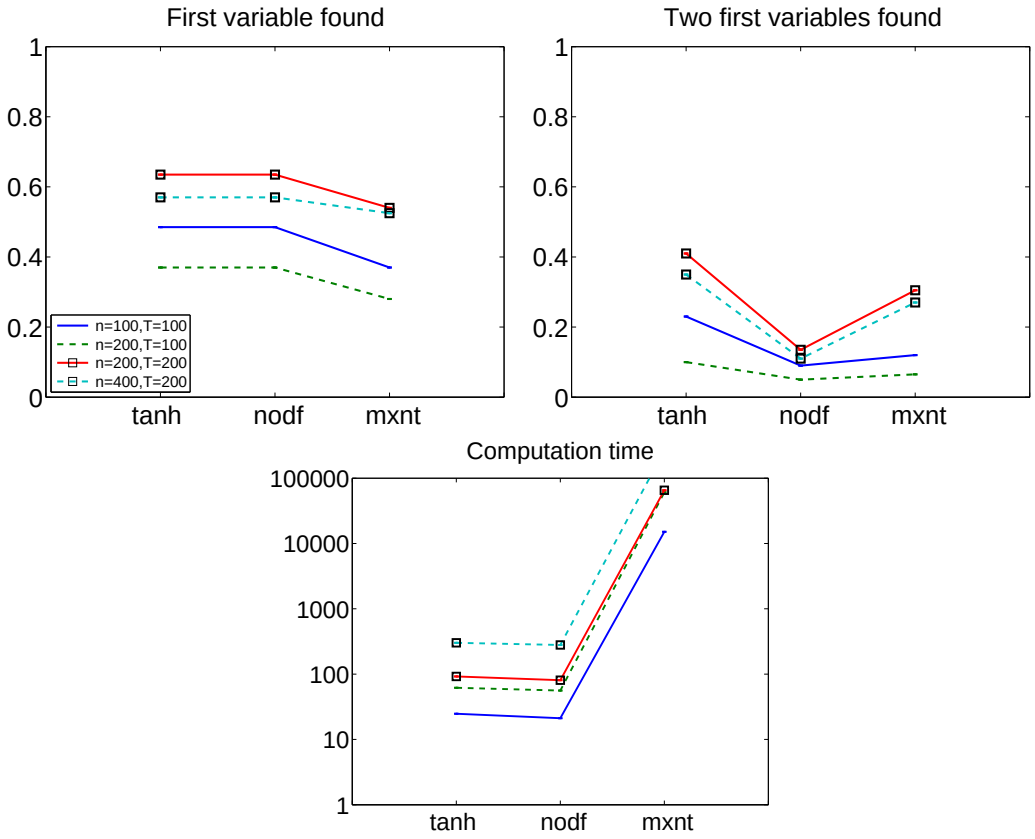

Figure 10: Simulation 7, with more variables than observations. Legend as in Figure 3. Rank correlations and causal directions correct are omitted because we only computed the first two variables for lack of computation time.

tanh and maxent introduced above in a purely linear way (i.e., *not* fitting the nonlinear function above, but just a linear function exactly as in previous simulations), to see if linear methods are able to cope with this data.

Furthermore, we used the criterion of the original method by Hoyer et al. (2009), based on the HSIC independence test by Gretton et al. (2008) of x (resp. y) and the residual in the regression of y on x (resp. of x on y). This was implemented by code provided by A. Gretton,[8] using the default setting for the kernel width.

The results are shown in Figure 12. Our likelihood ratio methods both performed relatively well, although the independence-based method by Hoyer et al. (2009) was arguably better than our maximum entropy method. However, the HSIC-based method was 10-100 times slower due to the use of kernel methods. The linear methods did not perform well at all.

8. Downloaded from `http://www.gatsby.ucl.ac.uk/~gretton/indepTestFiles/indep.htm`.

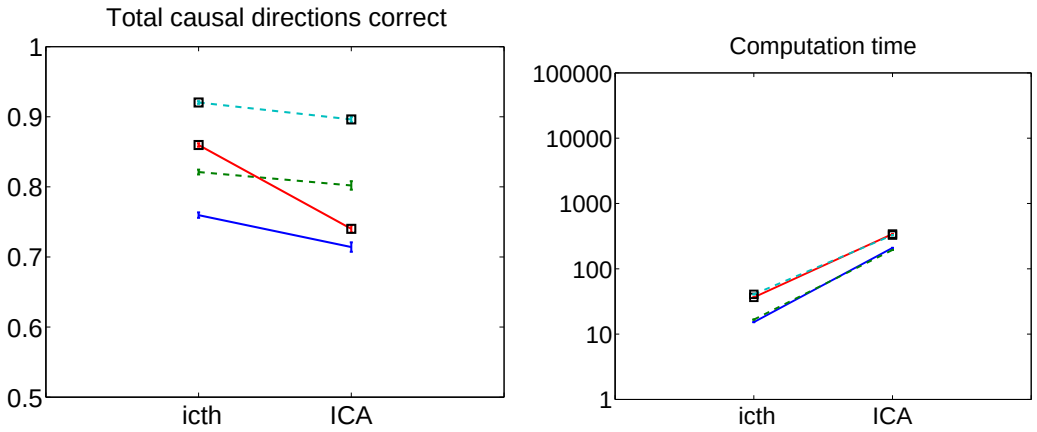

Figure 11: Simulation 8, with cyclic sparse graphs. Legend (sample sizes and dimensions) as in Figure 7.

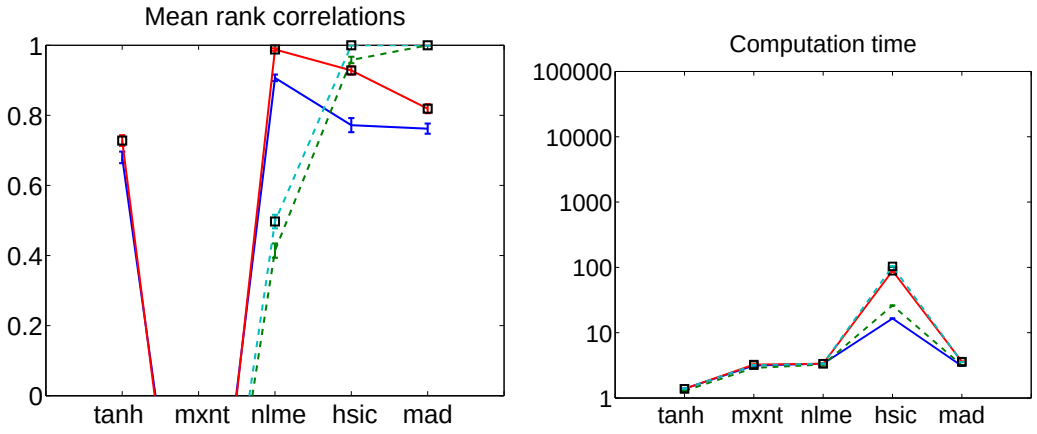

Figure 12: Simulation 9, with nonlinear model. The new algorithms are "nlme", the proposed likelihood ratio method extended to the nonlinear case using maximum entropy approximation in (22); "mad", a simplified and robustified approximation of the likelihood ratio in (24); "hsic", the original nonlinear method using independence (Hoyer et al., 2009). Blue: $\gamma = 0.5, T = 200$, Green: $\gamma = 2, T = 200$, Red: $\gamma = 0.5, T = 500$, Cyan: $\gamma = 2, T = 500$.

6.10 Simulation 10: Misspecified Disturbances

We further performed simulations in which the model is misspecified. First, we considered Simulation 1 with the following change: the disturbances generating the data had Laplacian distributions. Everything else was identical to Simulation 1, including the assumed log-pdf's and nonlinearities. Thus, the distribution of the disturbances was not exactly known, and was misspecified in the estimation. We also added the basic skewness method in the set of algorithms.

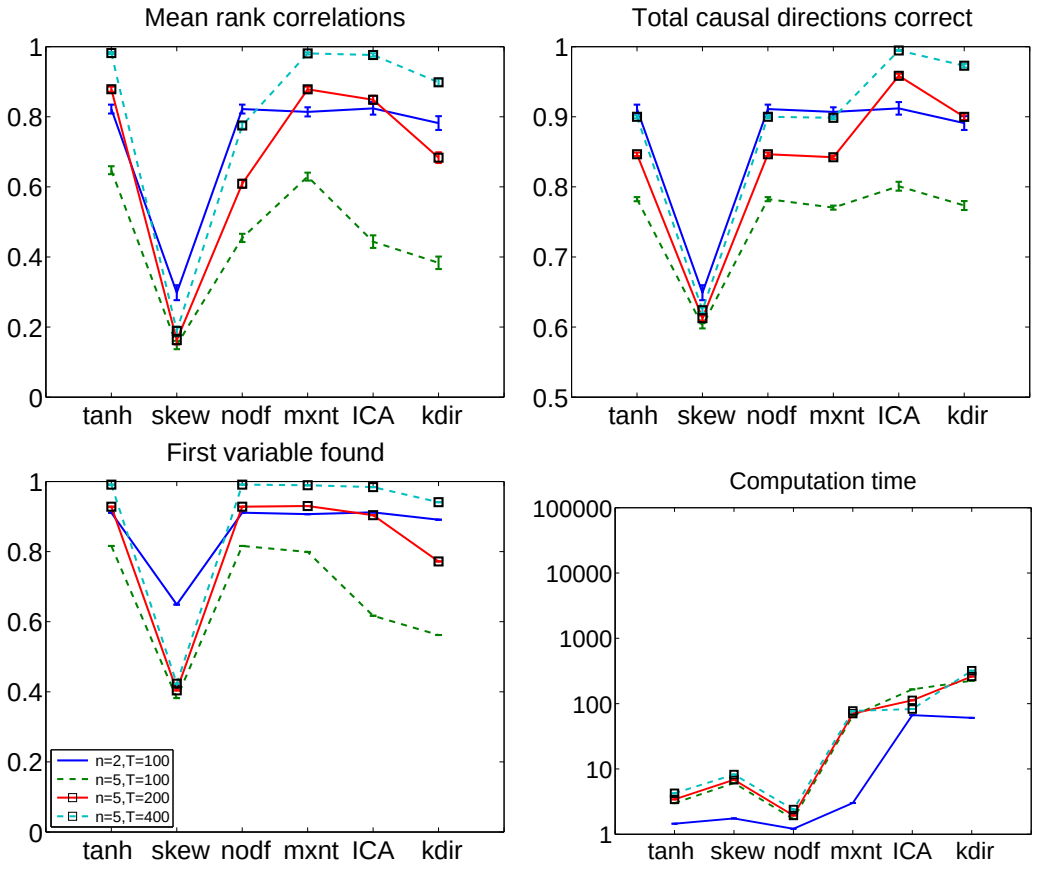

Figure 13: Simulation 10. Like Simulation 1 but with Laplacian disturbances used in generating the data, and the "skew" method added.

The results are in Figure 13. We see that the performance of most methods is actually better. This was expected in light of the theory of ICA, where it is well-known that if the actual data is more non-Gaussian than assumed in the estimation method, this is not a problem for most methods, and only increases the performance of the method compared to the case of less non-Gaussian data. In fact, the reason why we used the logistic distribution in generating data in many of the simulations above was in order to make the problem more difficult. On the other hand, the reason for using the logistic distribution in the algorithms is that it is widely used in ICA and has been empirically found to work well, partly due to the fact that its log-pdf is smooth, unlike many other super-Gaussian log-pdf's including the Laplacian.

Of course, if the non-Gaussianity is completely misspecified in the estimation method, estimation with fixed nonlinearities will inevitably fail. This is why the skewness method was hardly above chance level.

6.11 Simulation 11: Latent Variables

We conducted a further simulation to gain some insight into the robustness of the different methods to the existence of latent variables. We first created data $\mathbf{x}_0$ as in Simulation 1, with $n = 4, T = 500$. Then, we added a latent variable to the data as

$$\mathbf{x} = \mathbf{x}_0 + \alpha \mathbf{b} \tilde{s}$$

where $\tilde{s}$ is a latent variable with a standardized logistic distribution, $\mathbf{b}$ is a weight vector with elements drawn from a standardized Gaussian distribution, and α is the general strength of the latent variable, which took the values $[0, 0.25, 0.5, 1]$ in the different scenarios. (The value of $\alpha = 0$ effectively means no latent variables and is provided for comparison.) The latent variable $\tilde{s}$ violates the assumption of LiNGAM of having only one (independent) external input for each variable x_i.

The results are in Figure 14. Basically, we see that the latent variable deteriorates the performance of all the algorithms quite uniformly. It does not seem that any of the algorithms would be more resistant, or more sensitive, to latent variables than the others.

Recently, the framework presented here was generalized to a model including Gaussian latent variables by Chen and Chan (2012).

7. Experiments on Simulated fMRI Data

Since causal discovery experiments on real data are very difficult to validate, we use here brain imaging data which has been simulated using state-of-the-art biophysical models (Smith et al., 2011).

7.1 Simulation of fMRI Data

The simulations are described in detail by Smith et al. (2011); here we give a short summary. Networks of varied complexity were used to simulate fMRI timeseries. The simulations were based upon the dynamic causal modelling (DCM) forward model (Friston et al., 2003). DCM uses the nonlinear balloon model (Buxton et al., 1998) for the vascular dynamics, that is, the connection between the neural activities and the measured signal, sitting above a simple neural network model of the neural dynamics. Estimating causality from fMRI data is particularly challenging as the signal-to-noise ratio is relatively poor, fMRI timeseries are fairly Gaussian, and the number of timepoints is generally in the low hundreds.

We defined a number of nodes, which corresponded to brain regions. First, we generated the external inputs to the nodes, u_i, which are not quite the same as the external influences in the SEM, although related. They were binary (activity is "up" or "down") and generated using a Poisson process that controls the likelihood of switching the state. Neural noise of standard deviation 1/20 of the difference in height between the two states was added. The mean durations of the states were 2.5s (up) and 10s (down), with the asymmetry representing longer average "rest" than "firing" durations.

The neural activities z_i were then simulated using the DCM neural network model, as defined by

$$\dot{\mathbf{z}} = \sigma \mathbf{A} \mathbf{z} + \mathbf{M} \mathbf{u}$$

where $\mathbf{A}$ is the matrix defining network dynamics and $\mathbf{M}$ contains the weights controlling how the external inputs feed into the network (often just the identity matrix). The off-diagonal terms

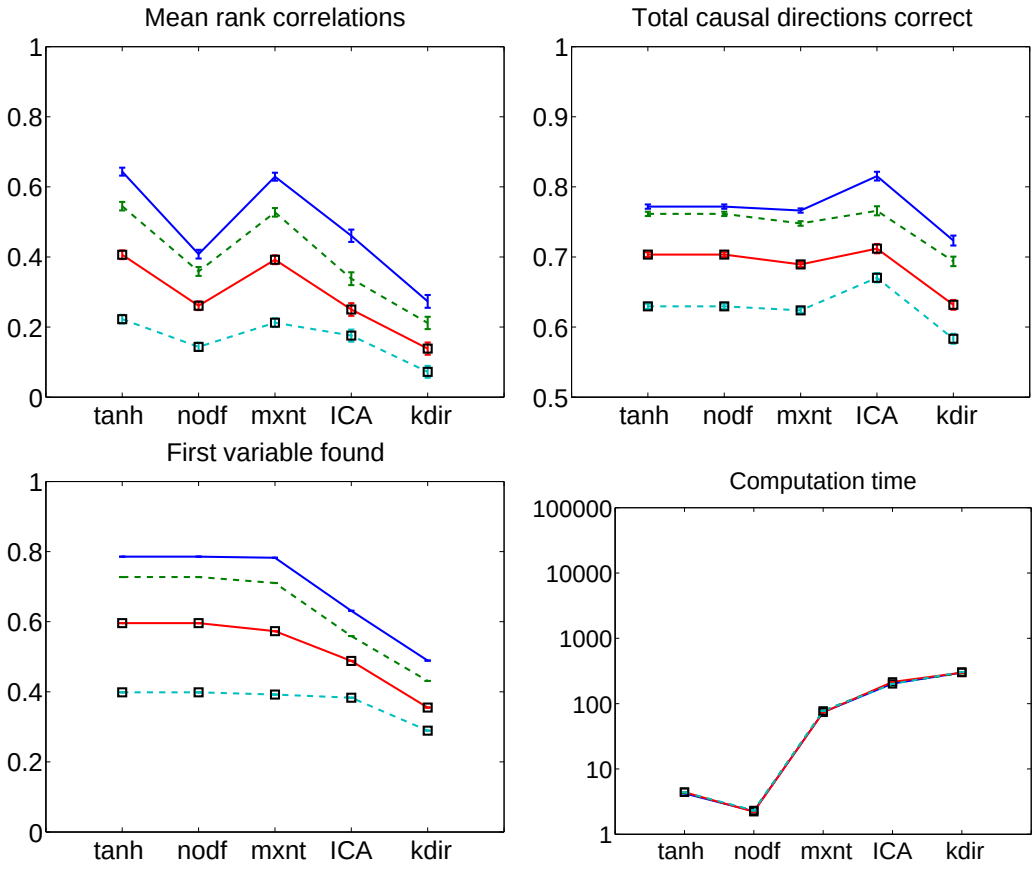

Figure 14: Simulation 11. Like Simulation 1, with $n = 4, T = 500$, but with a latent variable added. The four scenarios (curves) correspond to different strengths of the latent variable, starting with zero strength in blue curve.

in $\mathbf{A}$ determine the network connections between nodes, and the diagonal elements are all set to -1, to model within-node temporal decay; thus σ controls both the within-node (neural) temporal inertia/smoothing and the neural lag between nodes.

A central problem in fMRI is that the measured signal does not directly correspond to $\mathbf{z}$. To simulate this, each node's neural timeseries z_i was fed through the nonlinear balloon model for vascular dynamics responding to changing neural demand. The balloon model parameters were in general set according to the prior means in DCM. However, it is known that the haemodynamic processes vary across brain areas and subjects, resulting in different lags between the neural processes and the BOLD data, with variations of up to at least 1s (Handwerker et al., 2004; Chang et al., 2008). We therefore added randomness into the balloon model parameters at each node, resulting in variations in HRF (haemodynamic response function) delay of standard deviation 0.5s. Finally, thermal white (measurement) noise of standard deviation 0.1–1% (of mean signal level) was added.

Thus, we obtained the measured fMRI signals. They were sampled with a sampling interval of 3s (in most simulations), corresponding to a typical time of repetition (TR) in brain imaging literature.

The simulations comprised 50 separate realisations (or "subjects"), all using the same simulation parameters, except for having independently generated external inputs and different HRF parameters at each node (as described above); furthermore, the connection strengths were slightly perturbated for each subject. Each "subject's" data was a 10-minute fMRI session (200 timepoints) in many of the simulations.

For a summary of the specifications for the 28 simulations see Table 1.

Sim	n	length (mins)	TR (s)	noise (%)	HRF std (s)	other factors
1	5	10	3.00	1.0	0.5	
2	10	10	3.00	1.0	0.5	
3	15	10	3.00	1.0	0.5	
4	50	10	3.00	1.0	0.5	
5	5	60	3.00	1.0	0.5	
6	10	60	3.00	1.0	0.5	
7	5	250	3.00	1.0	0.5	
8	5	10	3.00	1.0	0.5	shared inputs
9	5	250	3.00	1.0	0.5	shared inputs
10	5	10	3.00	1.0	0.5	global mean confound
11	10	10	3.00	1.0	0.5	timeseries mixed with each other
12	10	10	3.00	1.0	0.5	new random timeseries mixed in
13	5	10	3.00	1.0	0.5	backwards connections
14	5	10	3.00	1.0	0.5	cyclic connections
15	5	10	3.00	0.1	0.5	stronger connections
16	5	10	3.00	1.0	0.5	more connections
17	10	10	3.00	0.1	0.5	
18	5	10	3.00	1.0	0.0	
19	5	10	0.25	0.1	0.5	neural lag=100ms
20	5	10	0.25	0.1	0.0	neural lag=100ms
21	5	10	3.00	1.0	0.5	2-group test
22	5	10	3.00	0.1	0.5	nonstationary connection strengths
23	5	10	3.00	0.1	0.5	stationary connection strengths
24	5	10	3.00	0.1	0.5	only one strong external input
25	5	5	3.00	1.0	0.5	
26	5	2.5	3.00	1.0	0.5	
27	5	2.5	3.00	0.1	0.5	
28	5	5	3.00	0.1	0.5	

Table 1: Summary of the 28 fMRI simulations' specifications (from Smith et al., 2011)

7.2 Estimation Methods for Simulated fMRI Data

We used pairwise measures with three different nonlinearities: tanh, skewness, and the robust measure of skewness. We did not estimate the existence of connections at all since that is not the main topic of the paper: We only looked at the estimated directionalities for those connections which really existed in the simulated data.

Since the skewness of the data was mainly positive, we used this prior information of positive skewness skewness-based measures. In other words, we skipped the skewness correction in (13).

For comparison, we used two methods by Patel et al. (2006) which were the most successful of the many methods tested by Smith et al. (2011), as well as basic ICA-based LiNGAM (Shimizu et al., 2006) which was applied on the whole data (not pairwise).

7.3 Results on Simulated fMRI Data

The goal is thus to recover the directionalities defined by the non-zero entries of the neural dynamics matrix **A** by estimating the directionalities given by **B** in our SEM. We evaluated the results using the same measures as Smith et al. (2011) to allow for direct comparison.

Our evaluations looked at the distribution of correctly estimated connections over the 50 simulated subjects. We concentrate here on evaluating methods for single subject (single session) data sets, and only utilise multiple subjects' data sets in order to characterise variability of results across multiple random instantiations of the same underlying network simulation. This is in contrast to the approach by Ramsey et al. (2011) who estimated the network over random subsets of 10 subjects, which is an easier task, at least if the subjects are not very different.

The raw connection strengths b_{ij} were converted into z-scores in order to make the plots more qualitatively interpretable, as the connection strengths are then more comparable across the different methods. The conversion from raw connection strengths to z-scores was achieved by using a null distribution of connection strengths, obtained by feeding in truly null timeseries data into each of the estimation methods. The null data was created by testing for connections between timeseries from *different* subjects' data sets, which have no causal connections between them (i.e., we randomly shuffled the subject labels for each node in the network). See Smith et al. (2011) for details. To specifically look at estimated directionalities, we use the higher of the two directions' measures to be the estimated connection strength.

The results are shown in Figs 15-16. The distributions are over all 50 simulated "subjects" and over all correct network edges; higher is better. Note, however, that this plot does not take into account the false positives, that is, the values estimated in the network matrix that should be empty, and concentrates exclusively on the estimation of causal directions. The plots, known as "violin plots", are simply (vertically-oriented) smoothed histograms, reflected in the vertical axis for better visualisation.

We see that the pairwise methods perform much better than Patel's measures or ICA-based LiNGAM on all the simulations. (The comparison to ICA-based LiNGAM may not be entirely fair since it estimates more than just directionalities.) In fact, our methods perform extremely well in most simulations. In all the simulations, the pairwise measures are the best, although in two cases the performances of all methods are so close to chance level that any comparison is difficult. The results are not very good in the following cases:

- Simulation 13 which has backwards connections (i.e., both $x \to y$ and $y \to x$) which is not surprising since it is against the basic philosophy of our modelling. However, the performance is

Caption for Figs. 15 and 16 on pages 147 and 148: The z-scores of the different measures used to determine the directionality, computed over subjects and connections, are shown as violin plots (i.e., histograms rotated to be horizontal and made symmetric). If the directions are found completely correctly, the violin plots are concentrated at the top. The blue dots show the the percentage of correctly estimated directions. First, we have three pairwise methods, and for comparison, two methods by Patel, as well as ICA-based LiNGAM. Each panel is one simulation.

---

clearly better than chance, which shows that our method is able to find the dominant direction some of the time.

- Simulation 22 which has nonstationary connection strengths, which violates another basic assumption of the model.
- Simulation 24 in which one of the inputs is strongly dominant. This is presumably because the effective signal-to-noise ratio is too poor for many of the connections.
- Simulations 25-27 in which the number of data points is smaller (recording length is shorter), performance being close to chance level for all methods.

Among the pairwise measures, there is no clear winner. However, the robust skewness measure is most often the best (in those cases where a clear difference can be seen), and never much worse then the other two.

8. Nonlinear Causal Discovery on Real Data

Finally, we applied our nonlinear methods on the Tübingen-UCI cause-effect data set[9] which consists of real measurements in which the true direction of causation in known. We used a total of 81 data sets, consisting of the subset of those data that had exactly two variables. In addition to our two new nonlinear methods, we applied the original HSIC-based methods by Hoyer et al. (2009), as well as the linear likelihood ratio with maximum negentropy approximation of Section 2.3. The relations in this data set are often quite nonlinear, and the linear methods are hardly above chance level (results not shown except for one method below), so we concentrate on the nonlinear methods here.

The nonlinear regression was performed by first fitting a least-squares regression curve using a Gaussian process as implemented in the `fit_gp` package by J. Mooij, based on code by C. E. Rasmussen and H. Nickisch. We used only the first 1,000 data points due to the excessive computational complexity of HSIC.

The results are shown in Table 2. The linear method, as well as our basic nonlinear method using maximum entropy were hardly better than chance. The method by Hoyer et al. (2009) was close to 62%. On the other hand, our simplest approximation using mean absolute deviation was 69% correct.

Presumably, one reason for the weak performance of our nonlinear method using maximum entropy approximations was that many of the data sets have strong outliers. The MAD-based objective in Equation (24) is quite robust against them (although the nonlinear regression method was not

9. Data set can be found at `http://webdav.tuebingen.mpg.de/cause-effect/`.

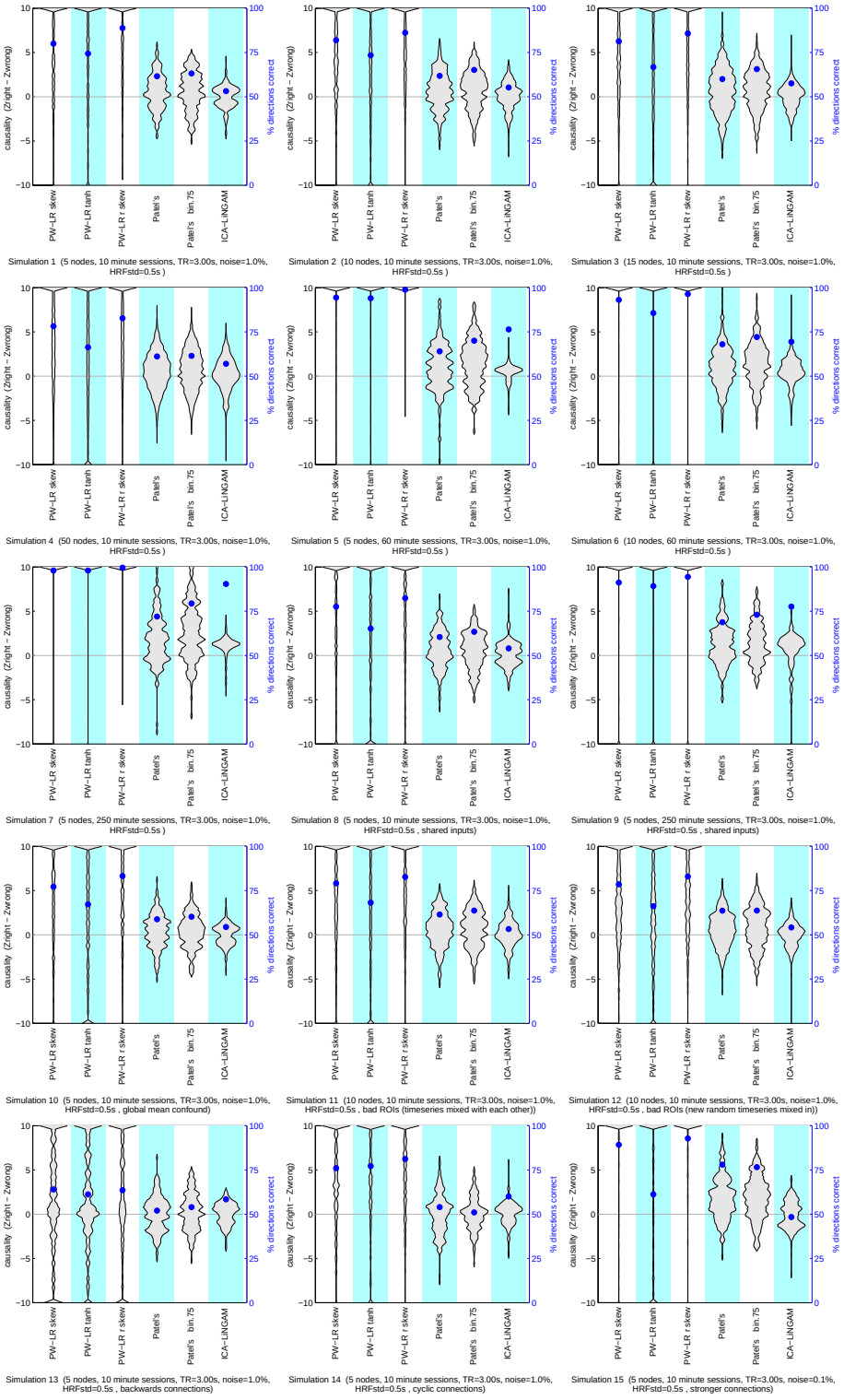

Figure 15: Results on simulated fMRI data, first half. See page 146 for caption.

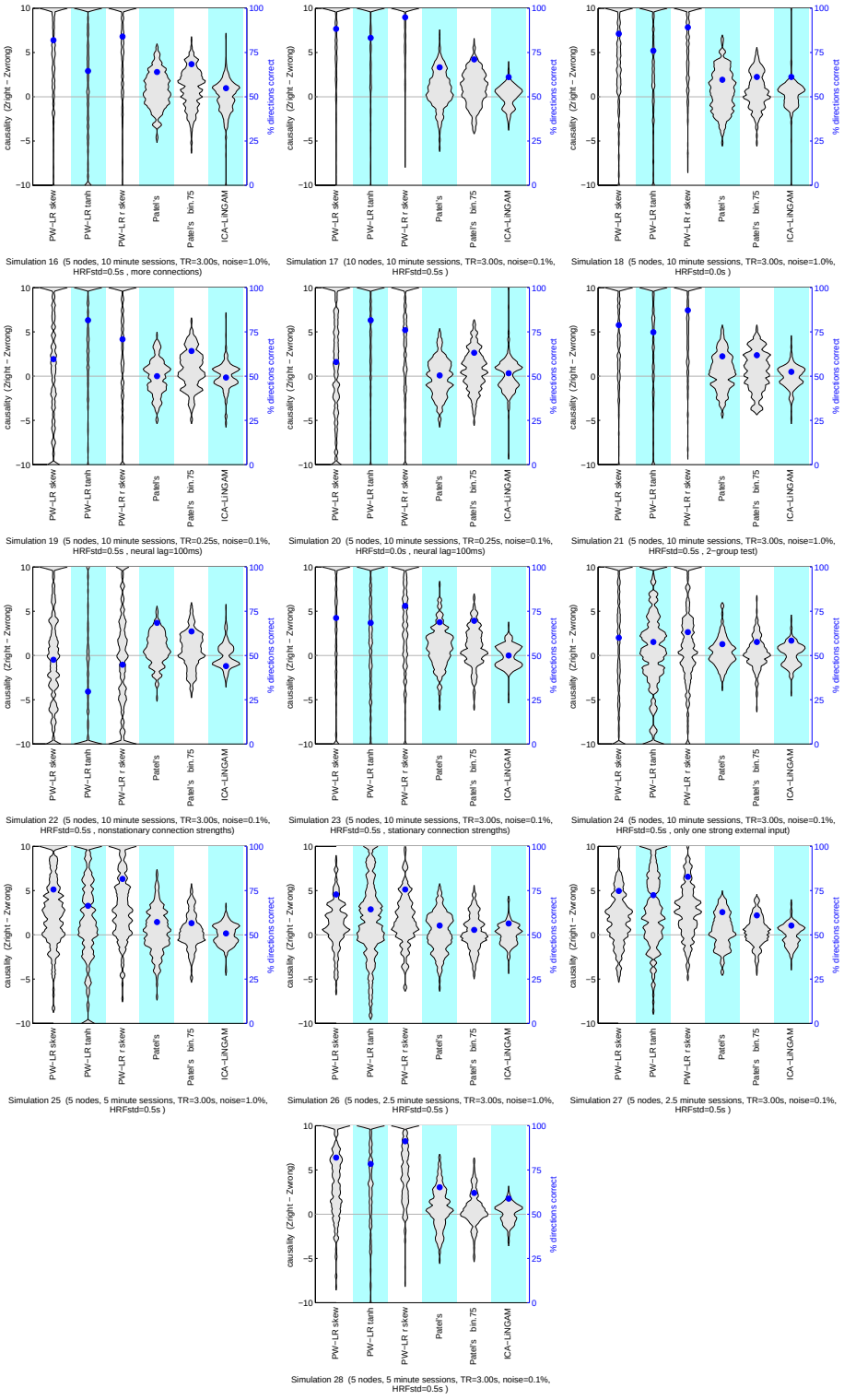

Figure 16: Results on simulated fMRI data, second half. See page 146 for caption.

Method	% correct
mxnt	50.6
hsic	61.7
nlme	53.1
mad	69.1

Table 2: Nonlinear models applied on the Tübingen-UCI data set. The algorithms are "nlme", the nonlinear likelihood ratio with maximum entropy approximation; "mad", the approximation using mean absolute deviations. For comparison: "hsic", the original nonlinear method using the HSIC measure of independence (Hoyer et al., 2009); and "mxnt", the linear method using the maximum entropy approximation.

made robust). The maximum entropy approximation might be greatly improved if the estimation of the variances used in the maximum entropy objective were made robust against outliers. Furthermore, both of our methods might be improved if the nonlinear fitting used a robust criterion instead of least squares.

9. Conclusion

We proposed very simple measures of the pairwise causal direction based on likelihood ratio tests and their approximations. We started with general measures based on entropy approximation which can accommodate different kinds of distributions, in Equations (2) and (3). Assuming that prior knowledge is available, we can develop more specific methods; for sparse variables we propose (5) and for skewed variables (17). We further showed how the measures can be extended to cyclic and nonlinear models. The different measures are recapitulated in Table 3.

We also showed how the pairwise measures can be used to estimate the whole Bayesian network in two ways. This is possible either in the DirectLiNGAM framework, or by a two-stage method based on first estimating the existence of the connections and then orienting them using the pairwise measures.

We also proposed a cumulant-based version of the nonlinear correlations. It was shown that the cumulant gives the correct pairwise direction. This shows the utility of using cumulants in theoretical analysis, and gives an intuitive interpretation of a new kind of cumulant. The cumulant-based analysis also indicated the noise-robustness of the nonlinear correlation methods, which was confirmed in the simulations. However, in practice the cumulant-based methods may suffer from sensitivity to outliers and thus their utility may be mainly in theoretical analysis.

The proposed measures seem to be particularly useful in the case where the number of data points is small compared to the dimension of the data, or the data is noisy. In such a case, the statistical performance of our methods is clearly superior to ICA-based LiNGAM and, to a lesser extent, DirectLiNGAM. The new methods are also computationally much faster than DirectLiNGAM. The importance of estimating causal networks with few data points has been recently highlighted by Smith et al. (2011) in the context of brain imaging. In fact, applied to the simulations by Smith et al. (2011), the new pairwise measures were clearly better then the methods originally tested.

Proposed measures for linear acyclic model (LiNGAM)			
Assumptions on non-Gaussianity	None	Sparse	Skewed
Equation for main new pairwise measure	(2) with (3)	(5)	(17)
Equation for previous measure by Dodge and Rousson	—	—	(14)

Proposed measures for extensions of LiNGAM		
Extension type	Cyclic case	Nonlinear case
Equation for new pairwise measure	Any LiNGAM measure	(23) with (3); or (24)

Table 3: The pairwise measures proposed in this paper recapitulated. The new cumulant-based approximations (Equations 6 and 10) have been omitted since they are mainly for theoretical analysis and not for practical use. Some of the measures by Dodge and Rousson (2001); Dodge and Yadegari (2010) would otherwise fit the "sparse" category but they assume the disturbances to be Gaussian and are thus less general. In the cyclic case, no new pairwise measures were introduced and it was merely proposed that the LiNGAM measures can be directly used even in the cyclic case. Of the measures highlighted above, the sparse-LiNGAM measure in Equation (5) was proposed in an earlier report on this work (Hyvärinen, 2010) while all others are new.

Thus, when estimating the LiNGAM model, it may be important to choose a suitable algorithm depending on data dimension, sample size, noise level, the distributions of the external influences, and other relevant factors.

Basic code for the pairwise measures is distributed on the Internet.[10]

Acknowledgments

We are grateful to Shohei Shimizu and Patrik Hoyer for deep and insightful comments on the manuscript, as well as to Christian Beckmann and Mark Woolrich for interesting discussions. We would also like to thank an anonymous referee for improving the derivation in Section 5.3. This work was supported by Academy of Finland, Computational Science Program and the Finnish Centre-of-Excellence in Algorithmic Data Analysis.

References

R.B. Buxton, E.C. Wong, and L.R. Frank. Dynamics of blood flow and oxygenation changes during brain activati on: the balloon model. *Magnetic Resonance in Medicine*, 39:855–864, 1998.

C. Chang, M.E. Thomason, and G.H. Glover. Mapping and correction of vascular hemodynamic latency in the BOLD signal. *NeuroImage*, 43:90–102, 2008.

10. Code can be found at `http://www.cs.helsinki.fi/u/ahyvarin/code/pwcausal/`.

Z. Chen and L. Chan. Causal discovery for linear non-gaussian acyclic models in the presence of latent gaussian confounders. In *Proc. Int. Conf. on Latent Variable Analysis and Signal Separation*, pages 17–24, 2012.

P. Comon. Independent component analysis—a new concept? *Signal Processing*, 36:287–314, 1994.

P. Daniušis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. In *Proc. 26th Conference on Uncertainty in Artificial Intelligence (UAI2010)*, 2010.

Y. Dodge and V. Rousson. On asymmetric properties of the correlation coefficient in the regression setting. *The American Statistician*, 55:51–54, 2001.

Y. Dodge and I. Yadegari. On direction of dependence. *Metrika*, 72:139–150, 2010.

K. J. Friston, L. Harrison, and W. Penny. Dynamic causal modelling. *NeuroImage*, 19(4):1273–1302, 2003.

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

D.A. Handwerker, J.M. Ollinger, and M. D'Esposito. Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage*, 21:1639–1651, 2004.

P. O. Hoyer, A. Hyvärinen, R. Scheines, P. Spirtes, J. Ramsey, G. Lacerda, and S. Shimizu. Causal discovery of linear acyclic models with arbitrary distributions. In *Proc. 24th Conf. on Uncertainty in Artificial Intelligence (UAI2008)*, pages 282–289, Helsinki, Finland, 2008.

P. O. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21, pages 689–696. MIT Press, 2009.

A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. In *Advances in Neural Information Processing Systems*, volume 10, pages 273–279. MIT Press, 1998.

A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

A. Hyvärinen. Pairwise measures of causal direction in linear non-gaussian acyclic models. In *Proc. Asian Conf. on Machine Learning, JMLR W&CP*, volume 13, pages 1–16, Tokyo, Japan, 2010.

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

A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *J. of Machine Learning Research*, 11:1709–1731, 2010.

J. Karvanen and V. Koivunen. Blind separation methods based on pearson system and its extensions. *Signal Processing*, 82(4):663–573, 2002.

J. M. Mooij, O. Stegle, D. Janzing, K. Zhang, and B. Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23 (NIPS*2010)*, pages 1687–1695, 2010.

R.S. Patel, F.D. Bowman, and J.K. Rilling. A Bayesian approach to determining connectivity of the human brain˙ *Human Brain Mapping*, 27:267–276, 2006.

D.-T. Pham and P. Garrat. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Trans. on Signal Processing*, 45(7):1712–1725, 1997.

J. D. Ramsey, S. J. Hanson, and C. Glymour. Multi-subject search correctly identifies causal connections and most causal directions in the DCM models of the Smith et al. simulation study. *NeuroImage*, 58(3):838–848, 2011.

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

S. Shimizu, A. Hyvärinen, Y. Kawahara, and T. Washio. A direct method for estimating a causal ordering in a linear non-gaussian acyclic model. In *Proc. 25th Conference on Uncertainty in Artificial Intelligence (UAI2009)*, pages 506–513, Montréal, Canada, 2009.

S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen. DirectLiNGAM: A direct method for learning a linear non-gaussian structural equation model. *J. of Machine Learning Research*, 12:1225–1248, 2011.

S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich. Network modelling methods for fMRI. *NeuroImage*, 54:875–891, 2011.

Y. Sogawa, S. Shimizu, Y. Kawahara, and T. Washio. An experimental comparison of linear non-gaussian causal discovery methods and their variants. In *Proc. Int. Joint Conf. on Neural Networks (IJCNN2010)*, Barcelona, Spain, 2010.

Y. Sogawa, S. Shimizu, A. Hyvärinen, T. Washio, T. Shimamura, and S. Imoto. Estimating exogenous variables in data with more variables than observations. *Neural Networks*, 24(8):875–880, 2011.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer-Verlag, 1993.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proc. 25th Conference on Uncertainty in Artificial Intelligence (UAI2009)*, pages 647–655, Montréal, Canada, 2009.