JMLR: Workshop and Conference Proceedings 13: 1-16
2nd Asian Conference on Machine Learning (ACML2010), Tokyo, Japan, Nov. 8–10, 2010.

Pairwise Measures of Causal Direction in Linear Non-Gaussian Acyclic Models

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

Editor: Masashi Sugiyama and Qiang Yang

Abstract

We present new measures of the causal direction 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 and analyze them based on related cumulant-based measures. The cumulant-based measures can be shown to give the right causal directions, and they are statistically consistent even in the presence of measurement noise. We further show how to apply these measures to estimate LiNGAM for more than two variables, and even in the case of more variables than observations. 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.

Keywords: Structural equation models, Bayesian networks, non-gaussianity, cumulants, 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 was proposed that using the non-gaussianity of the data, such models could be identifiable (Shimizu et al., 2006).

The original method for estimating LiNGAM was based on first doing independent component analysis (ICA) for the data and then deducing the network connections from the results of ICA. However, it seems that it should 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. (2009) 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 measure of the causal direction between two variables. Various methods were compared in (Sogawa et al., 2010b).

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 there is not enough data. 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. (2010) which considered causal analysis of simulated functional magnetic resonance imaging data, where the number of data points is typically small.

Thus, we see that measuring pairwise causal directions is a central problem in the theory of LiNGAM and related models. In fact, analyzing causal relations between two variables is an important problem in its own right, and was considered in the literature before the advent of LiNGAM (Dodge and Rousson, 2001; Shimizu and Kano, 2008).

In this paper, we develop new measures of causal direction between two variables, and apply them to the estimation of LiNGAM. The measures are presented in Section 2. In Section 3 we show how to apply them to estimating the model with more than two variables. Simulations with comparisons to other methods are reported in Section 4 and Section 5 concludes the paper.

2. Finding Causal Direction Between Two Variables

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

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

$$y = \rho x + d \tag{1}$$

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

$$x = \rho y + e \tag{2}$$

where the disturbance e is independent of y. The parameter ρ is the same in the two models because it is essentially 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.

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 $x \to y$ is given by Hyvärinen et al. (2010) as

$$\log L(x \to y) = \sum_t G_x(x_t) + G_d(\frac{y_t - \rho x_t}{\sqrt{1-\rho^2}}) - \log(1-\rho^2) \tag{3}$$

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

$$R = \frac{1}{T}\log L(x \to y) - \frac{1}{T}\log L(y \to x)$$
$$= \frac{1}{T}\sum_t G_x(x_t) + G_d(\frac{y_t - \rho x_t}{\sqrt{1-\rho^2}}) - G_y(y_t) - G_e(\frac{x_t - \rho y_t}{\sqrt{1-\rho^2}}) \quad (4)$$

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 (4) 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 much more robust against outliers.

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. However, for small sample sizes such modelling can be very difficult. Fortunately, 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). In particular, very good empirical results are usually obtained by modelling any sparse, symmetric densities by either the logistic density

$$G(u) = -2\log\cosh(\frac{\pi}{2\sqrt{3}}u) + \text{const.} \quad (5)$$

or the Laplacian density

$$G(u) = -\sqrt{2}|u| + \text{const.} \quad (6)$$

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.

2.3 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. We will also see later that such measures can have statistical advantages as well.

Let us make a first-order approximation

$$G(\frac{y - \rho x}{\sqrt{1-\rho^2}}) = G(y) - \rho\, x\, g(y) + o(\rho^2) \quad (7)$$

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

$$R \approx \tilde{R} = \frac{1}{T}\sum_t G(x_t) + G(y_t) - \rho x_t g(y_t) - G(y_t) - G(x_t) + \rho y_t g(x_t)$$
$$= \frac{\rho}{T}\sum_t -x_t g(y_t) + g(x_t) y_t \quad (8)$$

For example, if we approximate all the log-pdf's by (5), we get the "non-linear correlation"

$$\tilde{R} = \rho \hat{E}\{x \tanh(y) - \tanh(x)\, y\} \quad (9)$$

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.

2.4 Cumulant-based approach

To get further insight into the likelihood ratio approximation in (9), 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. Here, an approach based on fourth-order cumulants is possible by defining

$$\tilde{R}_{c4}(x, y) = \rho \hat{E}\{x^3 y - x y^3\} \quad (10)$$

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

$$\tanh(u) = u - \frac{1}{3}u^3 + o(u^3) \quad (11)$$

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*

$$\tilde{R}_{c4} = \mathit{kurt}(x)(\rho^2 - \rho^4) \quad (12)$$

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

$$\tilde{R}_{c4} = \mathit{kurt}(y)(\rho^4 - \rho^2). \quad (13)$$

Proof Consider the fourth-order cumulant

$$C(x, y) = \mathrm{cum}(x, x, x, y) = Ex^3 y - 3Exy \quad (14)$$

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

$$\tilde{R}_{c4} = \rho[C(x,y) - C(y,x)] \tag{15}$$

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

$$\mathrm{cum}(v,v,v,av+bw) = a\,\mathrm{cum}(v,v,v,v) + b\,\mathrm{cum}(v,v,v,w) \tag{16}$$

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

$$\begin{aligned}\tilde{R}_{c4} &= \rho[\mathrm{cum}(x,x,x,\rho x+d) - \mathrm{cum}(x,\rho x+d,\rho x+d,\rho x+d)] \\ &= \rho[\rho\mathrm{cum}(x,x,x,x) + \mathrm{cum}(x,x,x,d) \\ &\quad - \rho^3\mathrm{cum}(x,x,x,x) - 3\rho^2\mathrm{cum}(x,x,x,d) - 3\rho\mathrm{cum}(x,x,d,d) - \mathrm{cum}(x,d,d,d) \\ &= \rho[\rho\mathrm{kurt}(x) - \rho^3\mathrm{kurt}(x)] = \mathrm{kurt}(x)[\rho^2 - \rho^4]\end{aligned} \tag{17}$$

which proves (12). The proof of (13) 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 (9). Thus, the cumulant-based approach allowed us to prove rigorously that a nonlinear correlation of the form (10) 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 (9), only the nonlinear scalar function is different.

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

$$\tilde{R}'_{c4}(x,y) = \mathrm{sign}(\mathrm{kurt}(x))\rho\hat{E}\{x^3y - xy^3\} \tag{18}$$

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.

2.5 Intuitive 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.

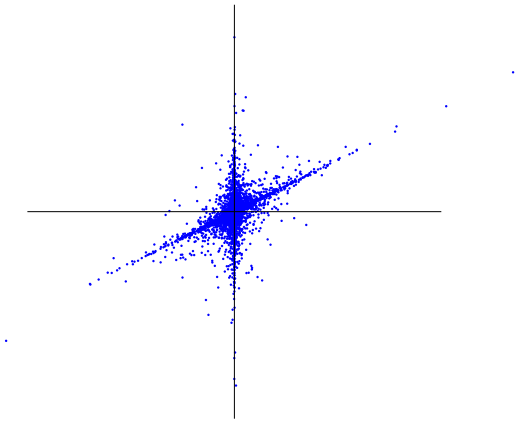

Figure 1: Intuitive illustration of the nonlinear correlations. Here, $x \to y$ and the variables are very sparse. The horizontal axis is x and the vertical axis is y. The nonlinear correlation $E\{x^3y\}$ is larger than $E\{xy^3\}$ because when both variables are simultaneously large (the "arm" of the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

Assume the data follows $x \to y$, and that both variables are sparse. 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, i.e. $|\rho| < 1$, 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 Fig. 1.

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

$$\tilde{R} = \rho \hat{E}\{h(x)y - x\,h(y)\} \tag{19}$$

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.

2.6 Noise-tolerance of the nonlinear correlations

An interesting point to note is that the cumulant in (10) 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. 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.

However, 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 (10) or (9) must be contrasted with its role in the likelihood ratio (4) 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 speculative, simulations below seem to support it.

2.7 Skewed variables

The cumulant-based approach also allows for a very simple extension of the framework to skewed variables. As a simple analogue to (10), we can define a third-order cumulant-based statistic as follows

$$\tilde{R}_{c3}(x,y) = \rho\hat{E}\{x^2y - xy^2\} \tag{20}$$

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*

$$\tilde{R}_{c3} = skew(x)(\rho^2 - \rho^3) \tag{21}$$

and if the causal direction is the opposite, we have

$$\tilde{R}_{c3} = skew(y)(\rho^3 - \rho^2). \tag{22}$$

Proof Consider the third-order cumulant

$$C(x,y) = \mathrm{cum}(x,x,y) = Ex^2y \tag{23}$$

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

$$\tilde{R}_{c3} = C(x,y) - C(y,x) \tag{24}$$

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

$$\begin{aligned}\tilde{R}_{c3} &= \rho[\mathrm{cum}(x,x,\rho x+d) - \mathrm{cum}(x,\rho x+d,\rho x+d)] \\ &= \rho[\rho\,\mathrm{cum}(x,x,x) + \mathrm{cum}(x,x,d) - \rho^2\mathrm{cum}(x,x,x) - 2\rho\,\mathrm{cum}(x,x,d) - \mathrm{cum}(x,d,d)] \\ &= \rho[\rho\,\mathrm{skew}(x) - \rho^2\mathrm{skew}(x)] = \mathrm{skew}(x)[\rho^2 - \rho^3]\end{aligned} \tag{25}$$

which proves (21). The proof of (22) is again completely symmetric. ■

To use the measure (20) in practice, we have to take into account the fact that we cannot usually assume the skewnesses of the variables to have some particular sign. To tackle this, 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^*

$$x^* = \mathrm{sign}(\mathrm{skew}(x))\,x \tag{26}$$

and the same for y. 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 $y \to x$, $\tilde{R}'_{c3}(y, x) < 0$.

The skewed case might also be approached by defining a skewed log-pdf and using the methods in previous sections. However, 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. In fact, in our simulations reported below, the skewness cumulant seems to work surprisingly well and it may not be necessary to consider likelihood-based skewness measures. Nevertheless, a likelihood-based approach is likely to be more robust against outliers than the cumulant-based one.

3. Estimating a Network With More Than Two Variables

In this section, we consider the general case of more than two variables.

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

$$\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e} \tag{27}$$

$\mathbf{e}$ 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 ordered according to the causal ordering the matrix $\mathbf{B}$ has all zeros above the diagonal.

3.2 Using pairwise measures in general LiNGAM estimation

We can use the pairwise analysis developed above to estimate LiNGAM which has more than two variables using the DirectLiNGAM framework (Shimizu et al., 2009). 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

$$\mathbf{M} = \mathbf{C} \odot E\{\mathbf{x}g(\mathbf{x})^T - g(\mathbf{x})\mathbf{x}^T\} \tag{28}$$

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 $g(u) = -u^2$ 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 a variable x_i which has no parents, all entries in the i-th row of $\mathbf{M}$ are positive, neglecting random errors. (Note that there is no reason why such first variable would be

unique.) This was shown to be exactly true for the cumulant-based approaches $g(u) = -u^3$ and $g(u) = -u^2$ and is true as a first-order approximation for $g(u) = \tanh(u)$.

Thus, we first find the row, say with index i^*, which is most likely to have all positive 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., 2009). 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 positive 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 doing to 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 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.

3.3 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 positive 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. Next we develop a more principled way of aggregation.

Consider the $m_{ij}, j = 1, \ldots, n$ for a fixed i, which are the estimates of pairwise likelihood ratios or some approximations. Assume they are independent and have gaussian distributions $N(\mu_{ij}, \sigma^2)$. 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 with mean m_{ij} and variance σ^2. Thus, the posterior log-probability that all of the $\mu_{ij}, j = 1, \ldots, n$ are positive can be calculated as

$$\log \prod_j P(\mu_{ij} > 0 | m_{ij}) = \log \prod_j P(\frac{\mu_{ij} - m_{ij}}{\sigma} > -\frac{m_{ij}}{\sigma} | m_{ij}) = \log \prod_j \Phi(\frac{m_{ij}}{\sigma}) \\ = \sum_j \log \Phi(\frac{m_{ij}}{\sigma}) \quad (29)$$

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:

$$\log \Phi(\frac{m_{ij}}{\sigma}) \approx -\frac{1}{2\sigma^2} \min(0, m_{ij})^2 \quad (30)$$

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 (29).

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

$$m_i = -\sum_j \min(0, [\mathbf{M}]_{ij})^2 \tag{31}$$

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, i.e. the first variable in the causal ordering.

4. Simulations

We made simulations comparing the different methods proposed in this paper, as well as previously proposed LiNGAM estimation methods. In the first set of simulations, basic settings were used. In the second, we investigated the noise-tolerance of the methods. In the third, we considered skewed distributions. Finally in the fourth simulation, we considered the case of more variables than observations (Sogawa et al., 2010a). In most of the simulations, we emphasize the case where the number of observations is small.

4.1 Basic simulations

The connection matrices were either generated completely randomly, giving fully connected DAG, or using a simple "serial" structure $x_1 \to x_2 \to \ldots \to x_n$ with a random connection strength. In either case, the non-zero coefficients in the acyclic $\mathbf{B}$ had a uniform distribution in the union of the intervals $[-0.8, -0.2]$ and $[0.2, 0.8]$.

Sample size and data dimension were varied so that there were in total six 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. $n = 5$, $T = 200$, "serial" DAG
5. $n = 5$, $T = 800$, fully connected DAG
6. $n = 5$, $T = 800$, "serial" DAG

The disturbances had Laplacian distributions, with standard deviations drawn from the same distribution as the non-zero coefficients in $\mathbf{B}$. 1,000 data sets were generated of each scenario.

To estimate the model, we used three methods proposed above. First, the original pairwise likelihood ratio (LR) in (4) was used in the DirectLiNGAM framework, using the true Laplacian log-pdf for the disturbances, and including deflation (regressing the found

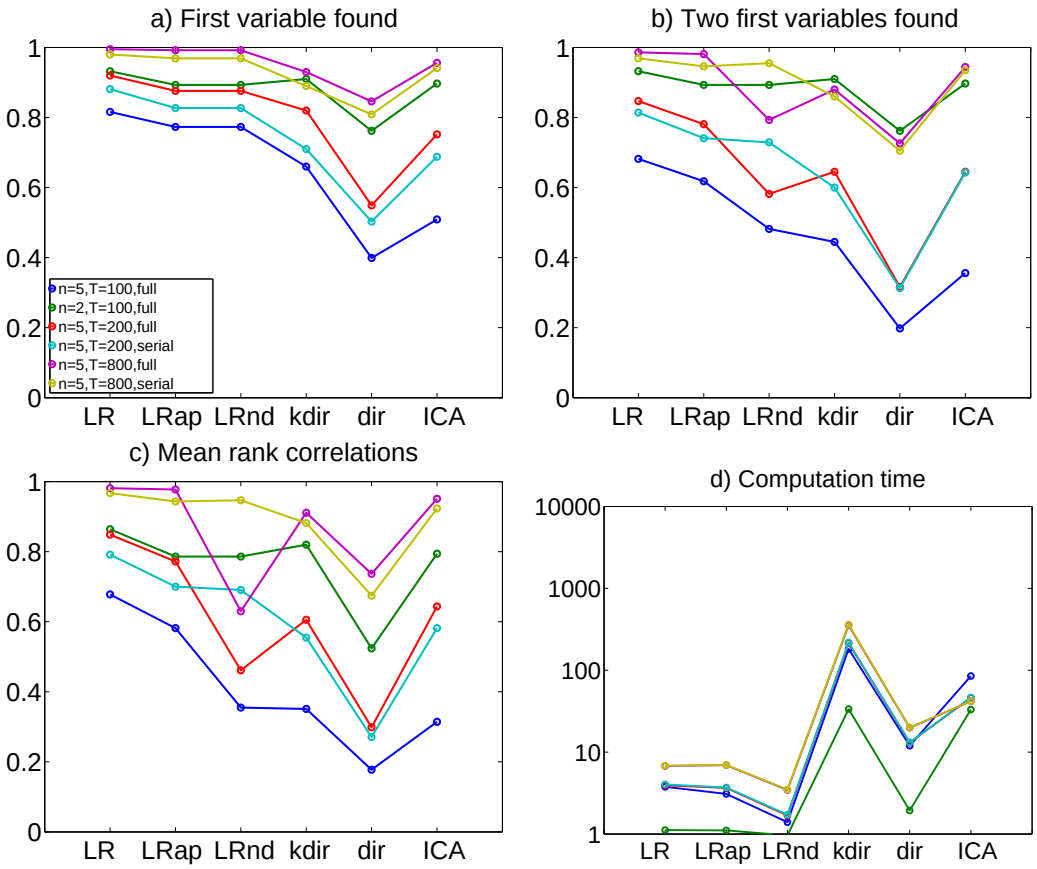

Figure 2: Results of basic simulations. a) The proportion of data sets for which the method estimated the first variable in the causal ordering correctly, i.e. the variable with no parents. b) The proportion of data sets for which the method estimated the first two variables correctly. c) Mean of rank-correlation coefficients between the estimated causal ordering and the true ordering. d) Computation times of one run of the different algorithms in milliseconds. Notice logarithmic scale. Different colours are different data-generating scenarios. The new algorithms used are: "LR": the true likelihood ratios as in (4) combined with deflation in DirectLiNGAM. "LRap": LR approximations in (28) based on tanh nonlinearity, combined with deflation in DirectLiNGAM. "LRnd": no deflation in likelihood ratio approximations, i.e. ordering based on the LR approximation matrix in (28) without any recomputation of the matrix. For comparison we used the previously proposed methods: "kdir": KernelDirectLiNGAM; "dir": original DirectLiNGAM; "ICA": LiNGAM estimated by ICA.

variables out of the remaining ones). Second, the LR first-order approximation matrix (28) was used in DirectLiNGAM with the nonlinearity $g(u) = \tanh(u)$ and with deflation. Third, the nonlinear correlations in (28) 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 3.2.

The methods were compared with three previously published methods: LiNGAM estimated using ICA (Shimizu et al., 2006), the original DirectLiNGAM (Shimizu et al., 2009), and KernelDirectLiNGAM (Sogawa et al., 2010b). In the case of KernelDirectLiNGAM, only 200 datasets were used to keep the computation time reasonable. These methods were implemented using the software found on the authors' web sites.

We computed three different performance indices for the methods. First, the percentage of data sets for which a method correctly estimated the first variable in the causal ordering, i.e. the variable with no parents. Second, the number of data sets in which a method correctly estimated the first two variables. Finally, we computed the Spearman rank-correlation coefficient for the causal ordering given by the method and the true ordering.

See Figure 2 for the results. Typically, the likelihood ratio (LR) was the best with respect to any of the three performance indices, and the likelihood ratio approximation (LRap) was the second. KernelDirectLiNGAM (kdir) is typically third, although for larger sample sizes, ICA-based LINGAM (ICA) may be better. What may be surprising is how well the likelihood ratio approximation without deflation (LRnd) works, although it is based on the analysis of a single matrix in (28). Regarding computational load, the methods proposed here are one or two orders of magnitude faster than statistically main contenders, KernelDirectLiNGAM and ICA-based LINGAM.

4.2 Simulations with noisy data

In a second set of experiments, we tested the noise-tolerance of the algorithms. The data dimension was set to $n = 5$ and fully connected DAG's were used. The sample size to $T = 10,000$, which means we are now analyzing the consistency of the method only and neglecting random effects by taking a very large sample size. The performance indices and algorithms are as in the first simulation. The results are shown in Fig. 3. We can see that the proposed methods are all clearly better than ICA-based LiNGAM which is not very noise-resistant. KernelDirectLiNGAM does seem to be more sensitive to noise as well. In line with our theoretical analysis, the method based on nonlinear correlations is now better than the method using the true likelihood ratio.

4.3 Simulations with skewed data

In the third set of experiments, we tested the performance of the methods with skewed data. We included now the nonlinear correlation using the third order cumulant, introduced in Section 2.7, among the algorithms. We used two different skewed distributions for the disturbances with the aim of imitating distributions found in fMRI data. 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

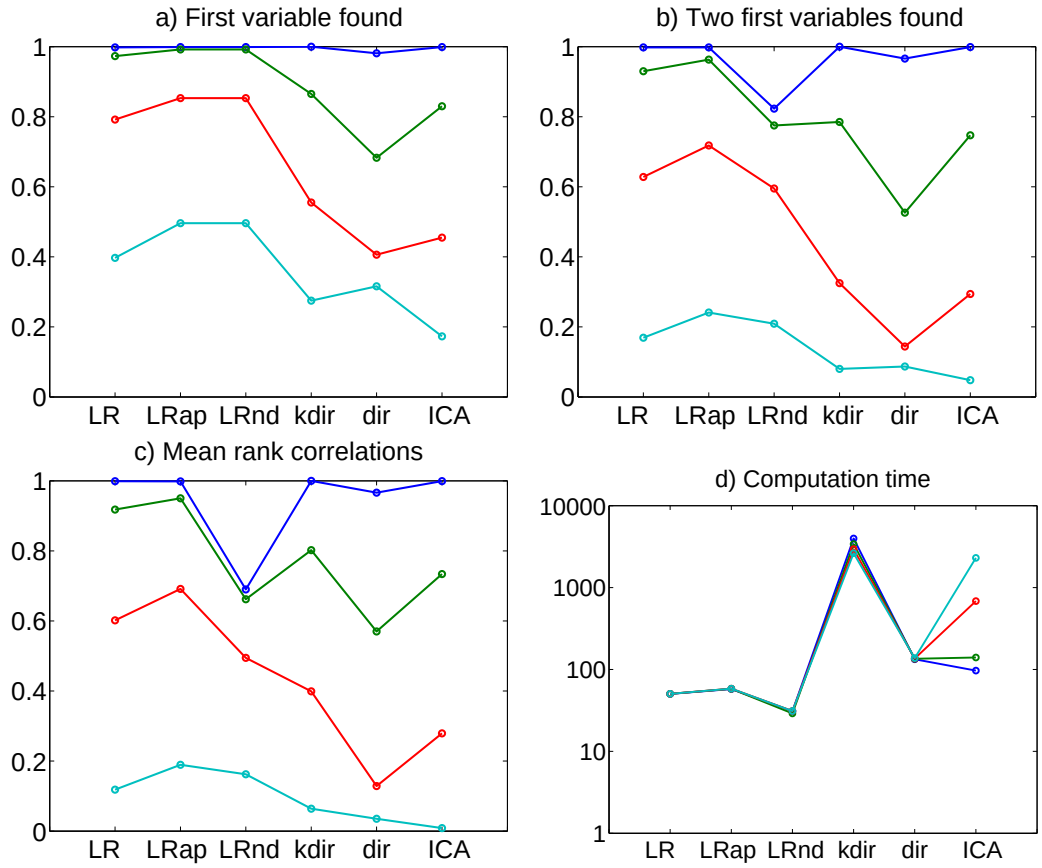

Figure 3: Simulations with noise. Legend as in Fig. 2, and with $n = 5, T = 10,000$. The noise was gaussian and white, with standard deviations taking the values 0, 0.1, 0.5, and 1.

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 $n = 5, n = 10$ and sample sizes $T = 200, 500$, respectively. DAG's were generated to be fully connected.

The results are shown in Fig. 4. We see that most methods do not work very well especially in the case of the non-sparse pdf (pdf 1), but what may be surprising is the poor performance of ICA-based LiNGAM even for the sparse pdf (pdf 2). The skew cumulant-based method and KernelDirectLiNGAM work quite well and the difference in statistical performance is small; the cumulant-based method is slightly better in the case of the sparse skewed data (pdf 2). Our cumulant-based method is also computationally one or two orders of magnitude faster than KernelDirectLiNGAM.

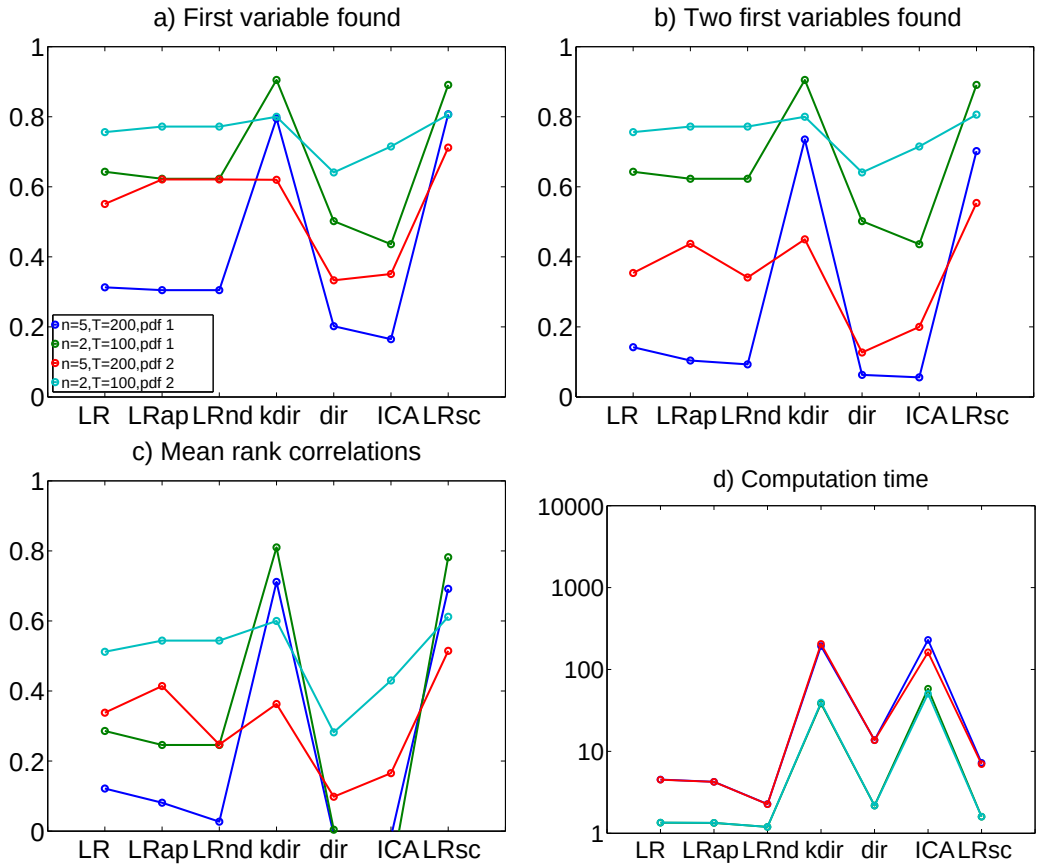

Figure 4: Simulations with skewed data. Legend as in Fig. 2, with the additional algorithm "LRsc" using the skewness cumulant.

4.4 Simulations with more variables than observations

Finally, we considered the case where there are more variables than observations. We considered two scenarios, $n = 200$, $T = 100$ and $n = 500$, $T = 200$. We only attempted to estimated 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., 2010a). We only used our 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 Fig. 5. 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).

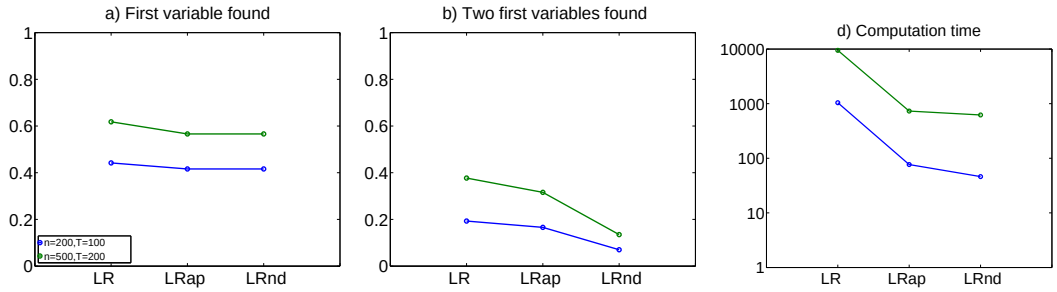

Figure 5: Simulations with more variables than observations. Legend as in Fig. 2. Rank correlations are omitted because we only computed the first two variables for lack of computation time.

5. Conclusion

We proposed very simple measures of the pairwise causal direction based on likelihood ratio tests and their approximations. The pairwise measures can also be used to estimate the whole Bayesian network in the DirectLiNGAM framework. We also found that ordering the variables based on a single nonlinear correlation matrix gives surprisingly good results.

We also proposed a cumulant-based version of the nonlinear correlations. It could actually be shown that the cumulant gives the right 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.

The proposed measures seem to be mainly useful in the case where the number of data points is small compared to the dimension of the data, or the data is noisy. The importance of estimating causal networks with few data points has been recently highlighted by Smith et al. (2010) in the context of brain imaging. In such a case, the statistical performance of our methods is clearly superior to ICA-based LiNGAM and, to a lesser extent, KernelDirectLiNGAM. The new methods are also computationally much faster than KernelDirectLiNGAM. This indicates that when estimating the LiNGAM model, it may be best to choose a suitable algorithm depending on data dimension, sample size, and noise level.

Acknowledgments

I'm grateful to Steve Smith and Christian Beckmann for interesting discussions, as well as to Shohei Shimizu and Patrik Hoyer for comments on the manuscript. This work was supported by the Finnish Centre-of-Excellence in Algorithmic Data Analysis of the Academy of Finland.

References

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

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.

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.

S. Shimizu and Y. Kano. Use of non-normality in structural equation modeling: Application to direction of causation. *Journal of Statistical Planning and Inference*, 138:3483–3491, 2008.

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)*, page 506513, Montréal, Canada, 2009.

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. 2010. Submitted manuscript.

Y. Sogawa, S. Shimizu, A. Hyvärinen, T. Washio, T. Shimamura, and S. Imoto. Discovery of exogenous variables in data with more variables than observations. In *Proc. Int. Conf. on Artificial Neural Networks (ICANN2010)*, Thessaloniki, Greece, 2010a. In press.

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, 2010b.