Estimating a Causal Order among Groups of Variables in Linear Models

Doris Entner and Patrik O. Hoyer

HIIT & Department of Computer Science, University of Helsinki, Finland

Abstract. The machine learning community has recently devoted much attention to the problem of inferring causal relationships from statistical data. Most of this work has focused on uncovering connections among scalar random variables. We generalize existing methods to apply to collections of multi-dimensional random *vectors*, focusing on techniques applicable to linear models. The performance of the resulting algorithms is evaluated and compared in simulations, which show that our methods can, in many cases, provide useful information on causal relationships even for relatively small sample sizes.

1 Introduction

Many techniques have recently been developed for inferring causal relationships from data over a set of random variables [1,2,3,4,5,6,7,8,9]. While most of this work has focused on uncovering connections among scalar random variables, in many actual cases each of the variables of interest may consist of multiple related, but distinct, measurements. For instance, in fMRI data analysis one is often interested in the functional connectivity among brain *regions*, and for each such region-of-interest one has data measured from a set of multiple voxels. Typically, in these cases, some aggregate of each area is computed, after which the standard approaches are directly applicable. However, it can be shown that not only may information be lost when computing aggregates, but the outputs of such methods may not even be correct in the large sample limit.

A simple example illustrating one of the problems inherent with working with aggregates is the following. Consider three sets of variables with causal connections $\mathcal{X} \to \mathcal{Y} \to \mathcal{Z}$, i.e. the variables in $\mathcal{X}$ may influence the variables in $\mathcal{Y}$, but not directly the variables in $\mathcal{Z}$, and the variables in $\mathcal{Y}$ may influence the variables in $\mathcal{Z}$. In this case, each variable $x \in \mathcal{X}$ is independent of each $z \in \mathcal{Z}$ *conditional on the full set of mediating variables* $\mathcal{Y}$. However, when replacing the variables of each group with their respective mean value (the typical aggregate used), denoted by $\bar{x}$, $\bar{y}$, and $\bar{z}$, in general we obtain $\bar{x} \not\perp\!\!\!\perp \bar{z} \mid \bar{y}$ [1,10]. Thus, it is important to develop methods for causal discovery that exploit the full information available, as opposed to only aggregates of the data.

Towards this end, in this paper we extend two existing approaches [5,6] designed for causal discovery among scalar random variables to the case of random vectors (i.e. groups of variables), both exploiting any kind of non-Gaussianity

present in the data. We also extend a recent method [8] for inferring the causal relationship among *two* arbitrarily distributed multi-dimensional variables to an arbitrary number of such variables. After describing the resulting algorithms, we evaluate and compare their performance in numerical simulations.

2 Model and Problem Statement

Throughout the paper, we will use the term 'group' to denote a set of underlying variables all belonging to the same (multi-dimensional) random vector representing a single concept (e.g. one region in fMRI analysis). We use the term 'variable' to represent a single scalar random variable belonging to one of the groups. Thus, for $g = 1, \ldots, G$, let $\mathcal{X}_g$ denote group g, and let the random vector $\boldsymbol{x}_g = (x_1^{(g)}, \ldots, x_{n_g}^{(g)})^T$ collect the n_g random variables belonging to group g. We assume that the *groups* $\mathcal{X}_g$ can be arranged in a causal order $K = (k_1, \ldots, k_G)$, such that the causal relationships among the groups can be represented by a directed acyclic graph. The data generating process is assumed to be a set of linear equations, given by

$$\boldsymbol{x}_{k_i} = \sum_{j < i} \mathbf{B}_{k_i, k_j} \boldsymbol{x}_{k_j} + \boldsymbol{e}_{k_i}, \ \ i = 1, \ldots, G, \tag{1}$$

with $\mathbf{B}_{k_i, k_j}$ arbitrary (real) matrices of dimension $n_{k_i} \times n_{k_j}$, containing the direct effects from group $\mathcal{X}_{k_j}$ to group $\mathcal{X}_{k_i}$. The vectors of disturbance terms $\boldsymbol{e}_{k_i}$ are assumed to be zero mean, and mutually independent over groups, i.e. $\boldsymbol{e}_{k_i} \perp\!\!\!\perp \boldsymbol{e}_{k_j}, \ i \neq j$, but are allowed to be dependent within each group. If we arrange the groups in a causal order K and define $\boldsymbol{x} = (\boldsymbol{x}_{k_1}, \ldots, \boldsymbol{x}_{k_G})$ and $\boldsymbol{e} = (\boldsymbol{e}_{k_1}, \ldots, \boldsymbol{e}_{k_G})$, we can rewrite Equation (1) in matrix form as $\boldsymbol{x} = \mathbf{B}\boldsymbol{x} + \boldsymbol{e}$ with $\mathbf{B}$ a lower block triangular matrix. The model reduces to standard LiNGAM (Linear Non-Gaussian Acyclic Model, [4,5]) when $\forall g : n_g = 1$ and all disturbances $\boldsymbol{e}$ are non-Gaussian. It also includes the model of [6] when $G = 2$, $n_1 = n_2 = 1$ and the disturbances are non-Gaussian. Finally, it contains as a special case the noisy model of [8] when $G = 2$ (but with no restriction on the n_g and $\boldsymbol{e}$).

We assume that all variables in $\boldsymbol{x}$ are observed, and that the grouping of these variables is known. Given merely observations of $\boldsymbol{x}$ generated by Model (1) (i.e. $\mathbf{B}$ and $\boldsymbol{e}$ are unknown), we want to estimate the unknown causal order K among these groups. We denote the data matrix of observations over the variables $\boldsymbol{x}$ as $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_G)^T$, where each column corresponds to one observation and each row to one variable. The observations are grouped according to the G groups, arranged in a random order, such that the first n_1 rows correspond to group $\mathcal{X}_1$, the following n_2 rows to group $\mathcal{X}_2$, and so on.

We note that our model family is equivalent to that given by [7]. The main difference between our approach and theirs is that they do not assume to know which variable belongs to what group, which results in algorithms exponential in the number of involved variables, whereas our algorithm explicitly builds upon such knowledge, allowing to construct computationally and statistically more efficient algorithms, polynomial in the number of groups.

3   Method

The overall algorithm for finding a causal order among the groups follows the approach introduced in [5]. We first search for an exogenous group (Section 3.1), and then 'regress out' the effect of this group on all other groups (Section 3.2). We iterate this process to generate a full causal order over the G groups.

3.1   Finding an Exogenous Group

We generalize three existing methods to search for an exogenous group, formally defined as below. Note that since the connections among the groups are assumed to be acyclic, there always exists at least one such exogenous ('source') group.

Definition 1 *A group $\mathcal{X}_j$ is exogenous if for $\boldsymbol{x}_j$ all matrices $\mathbf{B}_{j,i}$ of Equation (1) are zero.*

GroupDirectLiNGAM   As our first approach, we generalize the idea of DirectLiNGAM [5] to find an exogenous *variable*, to finding an exogenous *group*. The following lemma, which corresponds directly to Lemma 1 in [5], states a criterion to find an exogenous group using regressions and independence tests.

Lemma 1. *Let $\boldsymbol{x}$ follow Model (1) with non-Gaussian disturbance terms $\boldsymbol{e}$. Let $\boldsymbol{r}_i^{(j)} := \boldsymbol{x}_i - \mathbf{C}\boldsymbol{x}_j$ be the residuals when regressing $\boldsymbol{x}_i$ on $\boldsymbol{x}_j$ using ordinary least squares (OLS). A group $\mathcal{X}_j$ is exogenous if and only if $\boldsymbol{x}_j \perp\!\!\!\perp \boldsymbol{r}_i^{(j)}$ for all $i \neq j$.*

The proof of this lemma, and the proof of Lemma 2 in Section 3.2 are left to the online appendix at http://www.cs.helsinki.fi/u/entner/GroupCausalOrder/

To apply Lemma 1 in practice, we need to test for (in)dependence between two vectors of random variables, and combine the results of several such independence tests. Assuming that the test returns p-values p_{ji} under the null hypothesis of $\boldsymbol{x}_j \perp\!\!\!\perp \boldsymbol{r}_i^{(j)}$, $i \neq j$, we can get a measure of how exogenous group $\mathcal{X}_j$ is by combining these p-values using Fisher's method [11]. This means that we select as the exogenous group the one minimizing

$$\mu^{(j)} = -\sum\nolimits_{i \neq j} \log(p_{ji}). \tag{2}$$

To obtain the p-values p_{ji} we can test for joint dependence of the two vectors $\boldsymbol{x}_j$ and $\boldsymbol{r}_i^{(j)}$ using the Hilbert Schmidt Independence Criterion (HSIC, [12]), which, however, requires many samples to detect dependencies for high dimensional vectors. Alternatively, we can perform pairwise tests of each variable in $\boldsymbol{x}_j$ against each variable in $\boldsymbol{r}_i^{(j)}$ using nonlinear correlations, and combine the resulting $n_j \times n_i$ p-values appropriately. Details are left to the online appendix.

Pairwise Measure   Our second approach is based on modifying the pairwise measure [6] designed for inferring the causal relationship between two linearly related non-Gaussian scalar random variables x and y. If the true underlying causal direction is from x to y, the model is defined as $y = \rho x + e_y$ with $x \perp\!\!\!\perp e_y$.

As pointed out in Section 2, this is just a special case of our more general model. The (normalized) ratio of the log likelihoods for the two possible causal models is given by $R(x, y) = (\log L(x \to y) - \log L(y \to x)) / m$, where m is the sample size and L the likelihood of the specified direction, under some suitable assumption on the distributions of the disturbances. If the true underlying causal direction is $x \to y$, then $R(x, y) > 0$ in the large sample limit. Symmetrically, if $x \leftarrow y$, then $R(x, y) < 0$ in the limit.

To use the ratio $R(\cdot, \cdot)$ to find an exogenous group $\mathcal{X}_j$, the naïve approach is to calculate $R(x_k^{(j)}, x_l^{(i)})$ for each pair with $x_k^{(j)} \in \mathcal{X}_j$, $k = 1, \ldots, n_j$, $x_l^{(i)} \in \mathcal{X}_i$, $l = 1, \ldots, n_i$, $i \neq j$, and combine these measures. However, even if $\mathcal{X}_j$ is exogenous, these pairs do not necessarily meet the model assumption because of the dependent error terms within each group, and hence there is no guarantee for correctness even in the large sample limit. This approach, termed the *Naïve Pairwise Measure*, may however have a statistical advantage for small sample sizes (see Section 4).

To obtain a consistent method (simply termed *Pairwise Measure* in Section 4), we replace the second variable of the pairs $(x_k^{(j)}, x_l^{(i)})$ with a quantity which guarantees that the model assumption is met if $\mathcal{X}_j$ is exogenous: We first estimate the regression model $x_l^{(i)} = \sum_{\tilde{k}=1}^{n_j} \hat{b}_{l\tilde{k}} x_{\tilde{k}}^{(j)} + r_{l,(i)}$. If $\mathcal{X}_j$ is exogenous then the regression coefficients $\hat{b}_{l\tilde{k}}$ are consistent estimators of the true total effects (when marginalizing out any intermediate groups). Hence, defining $z_{k,l}^{(i)} := x_l^{(i)} - \sum_{\tilde{k}=1; \tilde{k} \neq k}^{n_j} \hat{b}_{l\tilde{k}} x_{\tilde{k}}^{(j)} = \hat{b}_{lk} x_k^{(j)} + r_{l,(i)}$ yields a pair $(x_k^{(j)}, z_{k,l}^{(i)})$ meeting the model assumption of [6] if $\mathcal{X}_j$ is exogenous. Thus, in this case, $R(x_k^{(j)}, z_{k,l}^{(i)}) > 0$, in the limit, for all k, l, and $i \neq j$. On the contrary, if $\mathcal{X}_j$ is not exogenous the measure can take either sign, and simulations show that it is unlikely to always obtain a positive one. A way to combine the ratios is suggested in [6], which can be modified for the group case as

$$\mu^{(j)} = \frac{1}{n_j \sum_{i \neq j} n_i} \sum_{k=1}^{n_j} \sum_{i \neq j} \sum_{l=1}^{n_i} \min\{0, R(x_k^{(j)}, z_{k,l}^{(i)})\}^2. \qquad (3)$$

That is, we penalize each negative value according to its squared magnitude and adjust for the group sizes. We select the group minimizing this measure as the exogenous one.

Trace Method Our third method for finding an exogenous group is based on the approach of [8,9], termed the Trace Method, designed to infer the causal order among two groups of variables $\mathcal{X}$ and $\mathcal{Y}$ with n_x and n_y variables, respectively. If the underlying true causality is given by $\mathcal{X} \to \mathcal{Y}$, the model is defined as $\boldsymbol{y} = \mathbf{B}\boldsymbol{x} + \boldsymbol{e}$, where the connection matrix $\mathbf{B}$ is chosen independently of the covariance matrix of the regressors $\boldsymbol{\Sigma} := \mathrm{cov}(\boldsymbol{x}, \boldsymbol{x})$, and the disturbances $\boldsymbol{e}$ are independent of $\boldsymbol{x}$. Note that this method is based purely on second-order statistics and does not make any assumptions about the distribution of the error terms $\boldsymbol{e}$, as opposed to the previous two approaches where we needed non-Gaussianity. The measure to infer the causal direction defined in [8] is given by

$$\Delta_{\mathcal{X}\to\mathcal{Y}} := \log\left(tr(\hat{\mathbf{B}}\hat{\boldsymbol{\Sigma}}\hat{\mathbf{B}}^T)/n_y\right) - \log\left(tr(\hat{\boldsymbol{\Sigma}})/n_x\right) - \log\left(tr(\hat{\mathbf{B}}\hat{\mathbf{B}}^T)/n_y\right) \quad (4)$$

where $tr(\cdot)$ denotes the trace of a matrix, $\hat{\boldsymbol{\Sigma}}$ an estimate of the covariance matrix of $\boldsymbol{x}$, and $\hat{\mathbf{B}}$ the OLS estimate of the connection matrix from $\boldsymbol{x}$ to $\boldsymbol{y}$. The measure for the backward direction $\Delta_{\mathcal{Y}\to\mathcal{X}}$ is calculated similarly by exchanging $\hat{\mathbf{B}}$ with the OLS estimate of the connection matrix from $\boldsymbol{y}$ to $\boldsymbol{x}$ and $\hat{\boldsymbol{\Sigma}}$ with the estimated covariance matrix of $\boldsymbol{y}$. If the correct direction is given by $\mathcal{X}\to\mathcal{Y}$, Janzing et al. [8] (i) conclude that $\Delta_{\mathcal{X}\to\mathcal{Y}} \approx 0$, (ii) show for the special case of $\mathbf{B}$ being an orthogonal matrix and the covariance matrix of $\boldsymbol{e}$ being $\lambda\mathbf{I}$, that $\Delta_{\mathcal{Y}\to\mathcal{X}} < 0$, and (iii) show for the noise free case that $\Delta_{\mathcal{Y}\to\mathcal{X}} \geq 0$. Hence, the underlying direction is inferred to be the one yielding Δ closer to zero [8]. In particular, if $|\Delta_{\mathcal{X}\to\mathcal{Y}}|\,/\,|\Delta_{\mathcal{Y}\to\mathcal{X}}| < 1$, then the direction is judged to be $\mathcal{X}\to\mathcal{Y}$.

We suggest using the Trace Method to find an exogenous group $\mathcal{X}_j$ among G groups in the following way. For each j, we calculate the measures $\Delta_{\mathcal{X}_j\to\mathcal{X}_i}$ and $\Delta_{\mathcal{X}_i\to\mathcal{X}_j}$, for all $i \neq j$, and infer as exogenous group the one minimizing

$$\mu^{(j)} = \sum_{i\neq j} \left(\Delta_{\mathcal{X}_j\to\mathcal{X}_i}\,/\,\Delta_{\mathcal{X}_i\to\mathcal{X}_j}\right)^2. \quad (5)$$

3.2   Estimating a Causal Order

Following the approach of [5], after finding an exogenous group we 'regress out' the effect of this group on all other groups. Since the resulting data set follows again the model in Equation (1) having the same causal order as the original groups, we can search for the next group in the causal order in this reduced data set. This is formally stated in the following lemma, which corresponds to the combination of Lemma 2 and Corollary 1 in [5].

Lemma 2. *Let $\boldsymbol{x}$ follow Model (1), and assume that the group $\mathcal{X}_j$ is exogenous. Let $\boldsymbol{r}_i^{(j)} := \boldsymbol{x}_i - \mathbf{C}\boldsymbol{x}_j$ be the residuals when regressing $\boldsymbol{x}_i$ on $\boldsymbol{x}_j$ using OLS, for $i = 1,\ldots,G$, $i \neq j$, and denote by $\boldsymbol{r}^{(j)}$ the column vector concatenating all these residuals. Then $\boldsymbol{r}^{(j)} = \mathbf{B}^{(j)}\boldsymbol{r}^{(j)} + \boldsymbol{e}^{(j)}$ follows Model (1). Furthermore, the residuals in $\boldsymbol{r}_i^{(j)}$ follow the same causal order as the original groups $\boldsymbol{x}_i$, $i \neq j$.*

Using Lemma 2, and the methods of Section 3.1, we can formalize the approach to find a causal order among the groups as shown in Algorithm 1.

3.3   Handling Large Variable Sets with Few Observations

The OLS estimation used in Algorithm 1 requires an estimate of the inverse covariance matrix which can lead to unreliable results in the case of low sample size. One approach to solving this problem is to use regularization. For the L^2-regularized estimate of the connection matrix we obtain $\hat{\mathbf{C}}_{i,j} = \mathbf{X}_i\mathbf{X}_j^T\,(\mathbf{X}_j^T\mathbf{X}_j + \lambda\mathbf{I})^{-1} = \mathrm{cov}(\mathbf{X}_i,\mathbf{X}_j)\,m\,(m\,\mathrm{cov}(\mathbf{X}_j,\mathbf{X}_j) + \lambda\mathbf{I})^{-1}$, with m the sample size and λ the regularization parameter, see for example [13]. In particular, this provides a regularized estimate of the covariance matrix.

Algorithm 1 (Estimating a Causal Order among Groups)

Input: Data matrix $\mathbf{X}$ generated by Model (1), arranged in a random causal order

  Initialize the causal order $K := [\,]$.

  repeat

    Find an exogenous group $\mathcal{X}_j$ from $\mathbf{X}$ using one of the approaches in Section 3.1.

    Append j to K.

    Replace the data matrix $\mathbf{X}$ with the matrix $\mathbf{R}^{(j)}$ concatenating all residuals $\mathbf{R}_i^{(j)}$, $i \neq j$, from the regressions of $\boldsymbol{x}_i$ on $\boldsymbol{x}_j$ using OLS:

$$\mathbf{X}_i = \mathbf{C}_{i,j}\mathbf{X}_j + \mathbf{R}_i^{(j)} \;\; \text{with } \mathbf{C}_{i,j} = \operatorname{cov}(\mathbf{X}_i, \mathbf{X}_j)\operatorname{cov}(\mathbf{X}_j, \mathbf{X}_j)^{-1}.$$

  until $G - 1$ group indices are appended to K

  Append the remaining group index to K.

Another approach is to apply the methods of Section 3.1 for finding an exogenous group to N data sets, each of which consists of G groups formed by taking subsets of the variables of the corresponding original groups. We then calculate measures $\mu_n^{(j)}$, $j = 1, \ldots, G$, $n = 1, \ldots, N$, as in Equations (2), (3) or (5), for each such data set separately, and pick the group $\mathcal{X}_{j^*}$ which minimizes the sum over these sets to be an exogenous one, i.e.

$$j^* = \arg\min_j \; \sum_{1 \leq n \leq N} \mu_n^{(j)} \tag{6}$$

where $\mu_n^{(j)}$ is the measure of group j in the nth data set. We then can proceed as in Algorithm 1 to find the whole causal order.

Note that the same approach can be used when multiple data sets are available, which are assumed to have the same causal order among the groups but possibly different parameter values. An example for such a scenario is given by fMRI data from several individuals. An equivalent of Equation (6) was suggested in [14] for the single variable case with multiple data sets.

4   Simulations

Together, the methods of Section 3 provide a diverse toolbox for inferring the model of Section 2. Here, we provide simulations to evaluate the performance of the variants of Algorithm 1, and compare it to a few ad hoc methods. Matlab code is available at http://www.cs.helsinki.fi/u/entner/GroupCausalOrder/

We generate models following Equation (1) by randomly creating the connection matrices $\mathbf{B}_{k_i,k_j}$, $i > j$ with, on average, $s\%$ of the entries being nonzero and additionally ensure that at least one entry is nonzero, to ensure a complete graph over the groups. To obtain the disturbance terms $\boldsymbol{e}_{k_i}$ for each group, we linearly mix random samples from various independent non-Gaussian variables as to obtain dependent error terms within each group. Finally, we generate the sample matrix $\mathbf{X}$ and randomly block-permute the rows (groups) to hide the generating causal order from the inference algorithms.
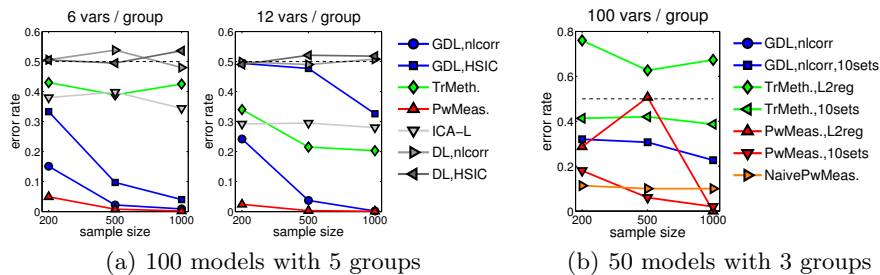
Fig. 1. Sample size (x-axis) against error rate (y-axis) for various model sizes and algorithms, as indicated in the legends (abbreviations: GDL = GroupDirectLiNGAM; nlcorr, HSIC: nonlinear correlation or HSIC as independence test; TrMeth. = Trace Method; PwMeas. = Pairwise Measure; ICA-L = modified ICA-LiNGAM approach; DL = DirectLiNGAM on the mean-variables; 10sets = Equation (6) on $N = 10$ data sets; L2reg = L^2-regularization for covariance matrix). The dashed black line indicates the number of mistakes made when randomly guessing an order.

We compare the variants of Algorithm 1 to two ad hoc methods. The first one is a modified ICA-based LiNGAM approach [4] where instead of searching for a permutation yielding a lower triangular connection matrix **B** (i.e. finding a causal order among the *variables*), we search for a block permutation yielding a lower block triangular matrix **B** (i.e. finding a causal order among the *groups*). Secondly, we compare our approach to DirectLiNGAM [5], when replacing each group by the mean of all its variables.[1]

We measure the performance of the methods by computing the error rates for predicting whether $\mathcal{X}_i$ is prior to $\mathcal{X}_j$, for all pairs (i, j), $i < j$.

Results for simulated data of sample size 200, 500 and 1000 generated from 100 random models having 5 groups with either 6 or 12 variables each, and $s = 10\%$, are shown in Figure 1 **(a)**. As expected, most methods based on Algorithm 1 improve their performance with increasing sample size. The only exception is the Trace Method on the smaller models; to be fair the method was not really designed for so few dimensions. Overall, the best performing method is the Pairwise Measure, closely followed by GroupDirectLiNGAM for the larger sample sizes. The ad hoc methods using DirectLiNGAM on the mean perform about as well as guessing an order (indicated by the dashed black line), whereas the modified ICA-LiNGAM approach performs better than guessing. However, it does not seem to converge for growing sample size, probably due to the dependent errors within each group, which is a violation of the ICA model assumption.

We next replace each group by a subset of its variables of size $m = 1, \ldots, n_g$, and apply Algorithm 1 to these subgroups. As expected, the larger m is, the less ordering mistakes are made. Details can be found in the online appendix.

---

[1] We do not compare our results to methods such as PC [1] or GES [3], as they cannot distinguish between Markov-equivalent graphs. Hence, in these simulations, they cannot provide any conclusions about the ordering among the groups since we generate complete graphs over the groups to ensure a total causal order.

Finally, we test the strategies described in Section 3.3 for handling low sample sizes in high dimensions on 50 models with 3 groups of 100 variables each, using 200, 500 and 1000 samples, and $s = 5\%$. For L^2-regularization, we choose the parameter λ using 10-fold cross validation on the covariance matrix. When taking subgroups, we use $N = 10$ data sets, and each subgroup containing ten variables. The error rates are shown in Figure 1 **(b)** (we only show the L^2-regularized results if they were better than without regularization). Unreliable estimates of the covariance matrix seem to affect especially the Trace Method, and the Pairwise Measure on the smaller sample sizes. On the smallest sample, using subsets seems to be advantageous for most methods, however, the best performing approach is the Naïve Pairwise Measure, which, however, does not seem to converge to be consistent, where as GroupDirectLiNGAM and the Pairwise Measure are.

In general, the simulations show that the introduced method often correctly identifies the true causal order, and clearly outperforms the simple ad hoc approaches. It is left to future work to study the performance in cases of model violations as well as to apply the method to real world data.

References

1. P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.
2. J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.
3. D.M. Chickering and C. Meek. Finding optimal bayesian networks. In *UAI*, 2002.
4. S. Shimizu, P.O. Hoyer, A. Hyvärinen, and A.J. Kerminen. A linear non-gaussian acyclic model for causal discovery. *JMLR*, 7:2003–2030, 2006.
5. 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. *JMLR*, 12:1225–1248, 2011.
6. A. Hyvärinen. Pairwise measures of causal directions in linear non-gaussian acyclic models. In *JMLR W.&C.P.*, volume 13, pages 1–16, 2010.
7. Y. Kawahara, K. Bollen, S. Shimizu, and T. Washio. GroupLiNGAM: Linear non-gaussian acyclic models for sets of variables. *arXiv*, 1006.5041v1, June 2010.
8. D. Janzing, P.O. Hoyer, and B. Schölkopf. Telling cause from effect based on high-dimensional observations. In *ICML*, 2010.
9. J. Zscheischler, D. Janzing, and K. Zhang. Testing whether linear equations are causal: A free probability theory approach. In *UAI*, 2011.
10. R. Scheines and P. Spirtes. Causal structure search: Philosophical foundations and problems, 2008. NIPS 2008 Workshop: Causality: objectives and assessment.
11. R.A. Fisher. *Statistical Methods for Research Workers*. London: Oliver and Boyd, 11th edition, 1950.
12. A. Gretton, K. Fukumizu, C.H. Teo, L. Song, B. Schölkopf, and A.J. Smola. A kernel statistical test of independence. In *NIPS*, 2008.
13. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, 2nd edition, 2008.
14. S. Shimizu. Joint estimation of linear non-gaussian acyclic models. *Neurocomputing*, 81:104–107, 2012.