Discovering Cyclic Causal Models by Independent Components Analysis

Gustavo Lacerda
Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Peter Spirtes
Joseph Ramsey
Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA 15213

Patrik O. Hoyer
Dept. of Computer Science
University of Helsinki
Helsinki, Finland

Abstract

We generalize Shimizu et al's (2006) ICA-based approach for discovering linear non-Gaussian acyclic (LiNGAM) Structural Equation Models (SEMs) from causally sufficient, continuous-valued observational data. By relaxing the assumption that the generating SEM's graph is acyclic, we solve the more general problem of linear non-Gaussian (LiNG) SEM discovery. LiNG discovery algorithms output a set of distribution-equivalent SEMs that, in the large sample limit, correctly represent the population distribution. We also give sufficient conditions under which only one of the distribution-equivalent output SEMs is "stable", and apply a LiNG discovery algorithm to simulated data.

1 Linear SEMs

Linear structural equation models (SEMs) are statistical causal models widely used in the natural and social sciences (including econometrics, political science, sociology, and biology) [1].

The variables in a linear SEM can be divided into two sets, the error terms (typically unobserved), and the substantive variables. Corresponding to each substantive variable $\mathbf{x}_i$ is a linear equation with $\mathbf{x}_i$ on the left-hand-side, and the direct causes of $\mathbf{x}_i$ plus the corresponding error term on the right-hand-side.

Each SEM with jointly independent error terms can be associated with a directed graph (abbreviated as DG) that represents the causal structure of the model and the form of the linear equations. The vertices of the graph are the substantive variables, and there is a directed edge from $\mathbf{x}_i$ to $\mathbf{x}_j$ just when the linear coefficient of $\mathbf{x}_i$ in the structural equation for $\mathbf{x}_j$ is non-zero. [1]

1.1 The model and an illustration

Let $\mathbf{x}$ be the random vector of substantive variables, $\mathbf{e}$ be the vector of error terms, and B be the matrix of linear coefficients for the substantive variables. Then equation 1 describes the linear SEM model:

$$\mathbf{x} = B\mathbf{x} + \mathbf{e} \tag{1}$$

For example, consider the model defined by:

$$\begin{aligned} \mathbf{x}_1 &= \mathbf{e}_1 \\ \mathbf{x}_2 &= 1.2\mathbf{x}_1 - 0.3\mathbf{x}_4 + \mathbf{e}_2 \\ \mathbf{x}_3 &= 2\mathbf{x}_2 + \mathbf{e}_3 \\ \mathbf{x}_4 &= -\mathbf{x}_3 + \mathbf{e}_4 \\ \mathbf{x}_5 &= 3\mathbf{x}_2 + \mathbf{e}_5 \end{aligned} \tag{2}$$

Note that the coefficient of each variable on the left-hand-side of the equation is 1.

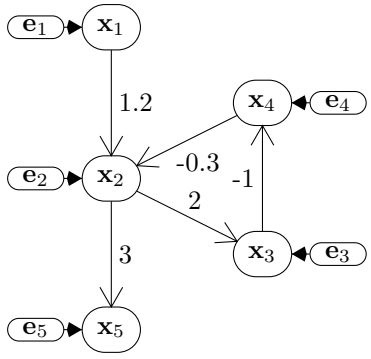

Fig. 1: Example 1

$\mathbf{x}$ can also be expressed directly as a linear combination of the error terms, as long as $I - B$ is invertible. Solving for $\mathbf{x}$ in Eq. 1 gives $\mathbf{x} = (I - B)^{-1}\mathbf{e}$. If we

[1] SEMs with acyclic graphs are called "recursive" or "acyclic", and SEMs with cyclic graphs are called "non-recursive" or "cyclic" [13].

let $A = (I - B)^{-1}$, then $\mathbf{x} = A\mathbf{e}$. A is called the reduced form matrix (in the terminology of Independent Components Analysis (see 3.1), it is called the "mixing matrix").

The distributions over the error terms in a SEM, together with the linear equations, entail a joint distribution over the substantive/print. variables. These probability distributions can be interpreted in terms of physical processes, as shown next.

1.2 Physical interpretation of linear SEMs

One interpretation of the equations is that they describe relationships among a set of variables that are in equilibrium ([14]). Under this interpretation, we will refer to the equations as "simultaneous" equations.

We will assume that the underlying dynamical equations relate the values of a substantive variable at time t to the values of other substantive variables at time $t-1$, and its respective constant error term. For example, the underlying dynamical equations for the SEM in Example 1 are:

$$\begin{aligned}
\mathbf{x}_1[t] &= \mathbf{e}_1 \\
\mathbf{x}_2[t] &= 1.2\mathbf{x}_1[t-1] - 0.3\mathbf{x}_4[t-1] + \mathbf{e}_2 \\
\mathbf{x}_3[t] &= 2\mathbf{x}_2[t-1] + \mathbf{e}_3 \\
\mathbf{x}_4[t] &= -\mathbf{x}_3[t-1] + \mathbf{e}_4 \\
\mathbf{x}_5[t] &= 3\mathbf{x}_2[t-1] + \mathbf{e}_5
\end{aligned} \tag{3}$$

Note that the error terms are not indexed by time because they are constant over time (although they may vary from individual to individual in the population). The physical interpretation of this is that the error terms are changing much more slowly than the state of the system (i.e. the values of the substantive variables). This is a slight variation of the interpretation given in [4].

In this interpretation of SEMs, the simultaneous equations are also "structural" in the sense that the effect of intervening in the system to set $\mathbf{x}_i$ to a constant c can be modeled by replacing the simultaneous equation for $\mathbf{x}_i$ by a new equation $\mathbf{x}_i = c$ (See [13], [12], [8], [7]).

The dynamical equations lead to the corresponding "simultaneous" equations describing the equilibrium state only if the value of each substantive variable $\mathbf{x}_i[t]$ converges to a constant c_i as $t \to \infty$. If, for example, the product of the coefficients in the cycle (henceforth called the "cycle-product") containing $\mathbf{x}_2$, $\mathbf{x}_3$, and $\mathbf{x}_4$ has an absolute value greater than 1, this convergence will not occur and the SEM is thus said to be "unstable" (see 5.2).

In both the acyclic and cyclic cases, the dynamical equations are deterministic, and the distribution of $\mathbf{x}$ results from different units in the population having different values for the error terms $\mathbf{e}$.

Note that in the dynamical equations, we have (often unrealistically) assumed that $\mathbf{x}_i[t]$ does not depend upon $\mathbf{x}_i[t-1]$ – we will say such a SEM has no "self-loops". If such a dependency does occur, e.g. $\mathbf{x}_3[t] = 0.5\mathbf{x}_3[t-1] + \mathbf{x}_2[t-1] + \mathbf{e}'_3$, in equilibrium this turns into $\mathbf{x}_3 = 0.5\mathbf{x}_3 + \mathbf{x}_2 + \mathbf{e}'_3$, which we can then renormalize as $\mathbf{x}_3 = 2\mathbf{x}_2 + \mathbf{e}_3$. Hence from the simultaneous equation, we cannot determine the coefficient of $\mathbf{x}_3[t-1]$ in the equation for $\mathbf{x}_3[t]$. However, this inability to estimate the coefficient of $\mathbf{x}_3[t-1]$ does not lead to problems in estimating the simultaneous equations, and does not lead to problems in predicting the effects of manipulating $\mathbf{x}_3$ to a constant except in two circumstances. First, it is possible that manipulating $\mathbf{x}_3$ can turn a stable system into an unstable system (by breaking loops that otherwise keep the system stable). Second, if the coefficient of $\mathbf{x}_3[t-1]$ is exactly 1, it is impossible to renormalize the equation, because after subtraction, the coefficient of $\mathbf{x}_3[t]$ on the left-hand-side of the equation is zero. If the coefficient of $\mathbf{x}_3[t-1]$ is exactly one, this does not imply that there is no set of simultaneous linear equations describing the equilibrium, but it does imply that the form of the simultaneous equations does not match the form of the underlying dynamical equations.

2 The problem and its history

2.1 The problem of DG causal discovery

Using the interpretation from 1.2, we can frame the problem as follows: find the set of all SEMs such that each SEM in the set describes the equilibrium population distribution of a set of variables, under the assumption that there is such a SEM.

2.2 Richardson's Cyclic Causal Discovery (CCD) Algorithm

While many algorithms have been suggested for discovering (equivalence classes of) DAGs from data, for general DGs only one provably correct algorithm is known, namely Richardson's Cyclic Causal Discovery (CCD) algorithm.

CCD outputs a "partial ancestral graph" (PAG) that represents both a set of directed graphs that entail the same set of zero partial correlations for all values of the linear coefficients, and features common to those directed graphs (such as ancestor relations). The algorithm performs a series of statistical tests of zero partial correlations to construct the PAG. The set of

zero partial correlations that is entailed by a linear SEM depends only upon the linear coefficients, and not upon the distribution of the error terms. Under some assumptions[2], in the large sample limit, CCD outputs a PAG that represents the true graph.

There are a number of limitations to this algorithm. First, the set of DGs contained in a PAG can be large, and while they all entail the same zero partial correlations (viz., those judged to hold in the population), they need not entail the same joint distribution or even the same covariances. Hence in some cases, the set represented by the PAG will include cyclic graphs that do not fit the data well. Therefore, even assuming that the errors are all Gaussian, it is possible in theory to reduce the size of the set of graphs output by CCD, although in practice this can be intractable. For details on the algorithm, see [9].

3 Shimizu et al's approach for discovering LiNGAM SEMs

The "LiNGAM algorithm" [11], which uses Independent Components Analysis (ICA), reliably discovers a unique correct LiNGAM SEM, under the following assumptions about the data: the structural equations of the generating process are linear and can be represented by an acyclic graph; the error terms have non-zero variance; the samples are independent and identically distributed; no more than one error term is Gaussian; and the error terms are jointly independent.[3]

3.1 Independent Components Analysis (ICA)

Independent components analysis ([3], [5]) is a statistical technique used for estimating the mixing matrix A in equations of the form $\mathbf{x} = A\mathbf{e}$ ($\mathbf{e}$ is often called "sources" and written $\mathbf{s}$), where $\mathbf{x}$ is observed and $\mathbf{e}$ and $\mathbf{A}$ are not ([3], [5]).

ICA algorithms find the invertible linear transformation $W = A^{-1}$ of the data X that makes the error distributions corresponding to the implied samples E of $\mathbf{e}$ maximally non-Gaussian (and thus, maximally independent). The matrix A can be identified up to scaling and permutation as long as the observed distribution is a linear, invertible mixture of independent components, at most one of which is Gaussian [3]. There are efficient algorithms for estimating A [5].

If we run an ICA algorithm on data generated by a linear SEM, the matrix W obtained will be a row-scaled, row-permuted version of $I - B$, where B is the coefficient matrix of the true model (this is a consequence of the derivation in 1.1). We are now left with the problem of finding the proper permutation and scale for the W matrix so that it equals $I - B$.

3.2 The LiNGAM discovery algorithm

Fig. 2(a) represents the W matrix output by a run of ICA, after removing the edges whose coefficients are statistically indistinguishable from zero[4]:

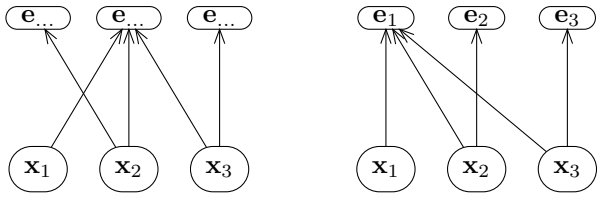

Fig. 2: (a) the raw W matrix output by ICA on a SEM whose graph is $\mathbf{x}_2 \rightarrow \mathbf{x}_1 \leftarrow \mathbf{x}_3$ (b) the corresponding $\widetilde{W}$ matrix, obtained by permuting W

Since the order of the error terms given by ICA is arbitrary, the algorithm needs to correctly match each error term $\mathbf{e}_i$ to its respective substantive variable $\mathbf{x}_i$. This means finding the correct permutation of the rows of the W matrix. We know that the W corresponding to the correct model cannot have a zero in the diagonal because this would imply that an error term has zero variance, which is a violation of LiNGAM's assumptions. We call such permutations "inadmissible".

Since, by assumption, the data was generated by a DAG, there is exactly one row-permutation of W that is admissible. To visualize this, this constraint says that there is exactly one way to reorder the error terms so that every $\mathbf{e}_i$ is the target of a vertical arrow.[5]for

In this example, swapping the first and second error term is the only permutation that produces an admissible matrix, as seen in Fig. 2(b).

After the algorithm finds the correct permutation, it finds the correct scaling, i.e. "normalizing" W by dividing each row by its diagonal element, so that the diagonal of the output matrix is all 1s (i.e. the coef-

[2]The assumptions are: linearity of the equations, the existence of a unique reduced form, that there are no zero partial correlations in the population that are not entailed for all values of the free parameters of the true graph, and that the error terms are uncorrelated

[3]The error terms are typically not jointly independent if the set of variables is not "causally sufficient", i.e. if there is an unobserved direct common cause of two or more of the observed variables.

[4]This is a simplification. Shimizu's paper actually finds an ordering in a non-local way, based on solving the linear assignment problem. See the next section.

[5]Another consequence of acyclicity is that there will be no right-pointing arrows in this representation, provided that the $\mathbf{x}$s are listed in an order that is compatible with the DAG.

ficient of each error term is 1, as specified in Section 1).

Bringing it all together, the algorithm computes B by using $B = I - W'$, where $W' = normalize(\widetilde{W})$, $\widetilde{W} = RowPermute(W)$ and $W = ICA(X)$.

Besides the fact that it determines the direction of every causal arrow, another advantage of LiNGAM over conditional-independence-based methods ([12]) is that it does not require the faithfulness assumption.

For more details on the LiNGAM approach, see [11].

4 Discovering LiNG SEMs

The assumptions of the family of LiNG discovery algorithms described below (abbreviated as "LiNG-D") are the same as the LiNGAM assumptions minus the assumption that the SEM is acyclic. In this more general case, as in the acyclic case, candidate models are generated by finding all admissible matches of the error terms ($\mathbf{e}_i$'s) to the observed variables ($\mathbf{x}_i$'s). In other words, each candidate corresponds to a row-permutation of the W matrix that has a zeroless diagonal.

As in LiNGAM, the output is the set of admissible models. In LiNGAM, this set always contains a single model, because of the acyclicity assumption. If the true model has cycles, however, more than one model will be admissible.

The remainder of this section addresses the problem of finding the admissible models, given that the zeros obtained by running ICA on finite samples are not exact.

4.1 Local algorithms

Local algorithms generate candidate models by locally testing which entries of W are zero, and finding all admissible permutations based on that. More formally, we call an algorithm "local" if, for each entry $w_{i,j}$ of W, it makes a decision about whether $w_{i,j}$ is zero using only $w_{i,j}$, and runs a combinatorial algorithm to find the row-permutations of W in which the diagonal has no zeros.

4.1.1 Deciding which entries are zero

There are several methods for deciding which small entries to set to zero:

- **Thresholding**: the simplest method for estimating which elements of W are zero is to simply choose a threshold value, and set every coefficient smaller than the threshold (in absolute value) to zero. This method fails to account for the fact that different coefficients may have different spreads, and will miss all coefficients smaller than the threshold.

- **Test the non-zero hypothesis by bootstrap sampling**: another method for estimating which elements of W are zero is to do bootstrap sampling. Bootstrap samples are created by resampling with replacement from the original data. Then ICA is run on each bootstrap sample, and each coefficient $w_{i,j}$ is calculated for each bootstrap sample. This leads to a real-valued distribution for each coefficient.[6] Then, for each one, a quantile test (a non-parametric one, if enough resamples are used) is performed in order to decide whether 0 is an outlier. If it isn't, the coefficient is pruned.

4.1.2 Constrained n-Rooks: the problem and an algorithm

Once it is decided which entries are zero, the algorithm searches for every row-permutation of W that has a zeroless diagonal. Each such row-permutation corresponds to a placement of n rooks onto the non-zero entries on an $n \times n$ chessboard such that no two rooks threaten each other. Then the rows can be permuted so that all the rooks end up on the diagonal, thus ensuring that the diagonal has no zeros.

To solve this problem, we use a simple depth-first search that prunes search paths that have nowhere to place the next rook. In the worst case, every permutation is admissible, and the search takes $O(n!)$.

4.2 Non-local algorithms

In the LiNGAM (acyclic) approach, it is straightforward to use non-local methods, by solving the linear assignment problem (i.e. finding the best match between the $\mathbf{e}_i$s and $\mathbf{x}_i$s). For example, the Hungarian algorithm [6] can be used to find the single best row-permutation of W, by minimizing a loss function that heavily penalizes entries in the diagonal that are close to zero (such as $x \rightarrow |1/x|$) [11]. For general LiNG discovery, however, algorithms that find the best linear assignment do not suffice, since there may be multiple admissible permutations.

One idea is to use a k-th best assignment algorithm [2] to find all of the permutations of W that have a

[6] One needs to be careful when doing this, since each run of ICA may return a W in a different row-permutation. This means that we first need to row-permute each bootstrap W to match with the original W.

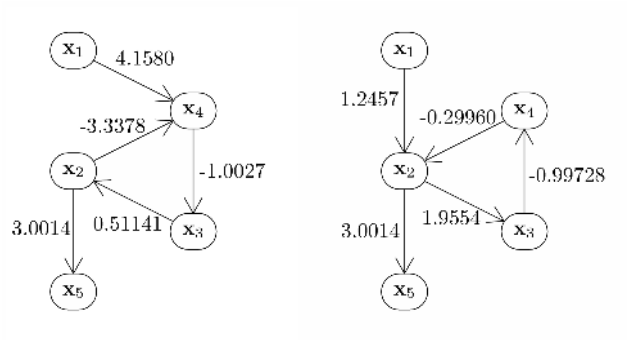

Fig. 3: The output of LiNG-D: Candidate #1 and Candidate #2

zeroless diagonal (i.e. the k permutations with the least penalty on the diagonal), for increasing k.

With enough data, all permutations corresponding to inadmissible models will score poorly, and there should be a clear separation between admissible and inadmissible models.

4.3 Sample run

We generated 15000 sample points using Model 1 and error terms distributed according to a symmetric Gaussian-squared distribution[7].

Fig. 3 shows the output of the local thresholding algorithm with the cut-off set to 0.05.

For the sake of reproducibility, our code with instructions is available from: `www.phil.cmu.edu/~tetrad/cd2008.html` .

By assuming that the true model is stable, one would select candidate #2. Since our simulation used a stable model, this is indeed the correct answer (see Fig. 1). In general, however, there may be multiple stable models, and one cannot reliably select the correct one.

Since ICA cannot give us coefficients from a node to itself, we will be unsure about the existence or non-existence of self-loops. Since self-loops can affect the stability of a model in either direction, whenever we use the stability criterion, we must assume the absence of self-loops.

[7]The distribution was created by sampling from the standard Gaussian(0,1) and squaring it. If the value sampled was negative, it was made negative again.

5 Theory

5.1 Notions of DG equivalence

There are a number of different senses in which the directed graphs associated with SEMs can be "equivalent" or "indistinguishable" given observational data, assuming linearity and no dependence between error terms.

- DGs G_1 and G_2 are *zero partial correlation equivalent* if and only if the set of zero partial correlations entailed for all values of the free parameters (non-zero linear coefficients, distribution of the error terms) of a linear SEM with DG G_1 is the same as the set of zero partial correlations entailed for all values of the free parameters of a linear SEM with G_2. For linear models, this is the same as *d-separation equivalence* [12]

- DGs G_1 and G_2 are *covariance equivalent* if and only if for every set of parameter values for the free parameters of a linear SEM with DG G_1, there is a set of parameter values for the free parameters of a linear SEM with DG G_2 such that the two SEMs entail the same covariance matrix over the substantive variables, and vice-versa.

- DGs G_1 and G_2 are *distribution equivalent* if and only if for every set of parameter values for the free parameters of a linear SEM with DG G_1, there is a set of parameter values for the free parameters of a linear SEM with DG G_2 such that the two SEMs entail the same distribution over the substantive variables, and vice-versa.

It follows from well-known theorems about the Gaussian case [12], and some trivial consequences of known results about the non-Gaussian case [11], that the following relationships exist among the different senses of equivalence for acyclic SEMs: If all of the error terms are assumed to be Gaussian, distribution equivalence is equivalent to covariance equivalence, which in turn is equivalent to d-separation equivalence. If not all of the error terms are assumed to be Gaussian, then distribution equivalence entails (but is not entailed by) covariance equivalence, which entails (but is not entailed by) d-separation equivalence.

So for example, given Gaussian error terms, $A \leftarrow B$ and $A \rightarrow B$ are zero partial correlation equivalent, covariance equivalent, and distribution equivalent. But given non-Gaussian error terms, $A \leftarrow B$ and $A \rightarrow B$ are zero-partial-correlation equivalent and covariance equivalent, but not distribution equivalent. So for Gaussian errors and this pair of DGs, no algorithm

that relies only on observational data can reliably select a unique acyclic graph that fits the population distribution as the correct causal graph without making further assumptions; but for all (or all except one) non-Gaussian errors there will always be a unique acyclic graph that fits the population distribution.

While there are theorems about the case of cyclic SEMs and Gaussian errors, we are not aware of any such theorems about cyclic SEMs with non-Gaussian errors with respect to distribution equivalence. In the case of cyclic SEMs with all Gaussian errors, distribution equivalence is equivalent to covariance equivalence, which entails (but is not entailed by) d-separation equivalence. In the case of cyclic SEMs in which some of the error terms may be non-Gaussian, distribution equivalence entails (but is not entailed by) covariance equivalence, which in turn entails (buedu.cmu.cmu.tetrad.search.t is not entailed by) d-separation equivalence. However, given all (or all but one) non-Gaussian error terms, the important difference between acyclic SEMs and cyclic SEMs is that no two different acyclic SEMs are distribution equivalent, but there are different cyclic SEMs that are distribution equivalent.

Hence, no algorithm that relies only on observational data can reliably select a unique cyclic graph that fits the data as the correct causal graph without making further assumptions. For example, the two cyclic graphs in Fig. 3 are distribution equivalent.

5.2 The output of LiNG-D is as fine as possible

The following theorem states that any two SEMs output by LiNG-D entail the same distribution.

Theorem 1 *The output of LiNG-D is the set of all SEMs that represent the observed distribution.*

Proof: The weight matrix output by ICA is determined only up to scaling and row permutation. Intuitively, then, permuting the error terms does not change the mixture. Now, more formally:

Let M_1 and M_2 be candidate models output by LiNG-D. Then W_1 and W_2 are row-permutations of W:

$W_1 = P_1W$, $W_2 = P_2W$

Likewise, for the error terms: $E_1 = P_1E$, $E_2 = P_2E$

Then the list of samples X implied by M_1 is $A_1E_1 = (W_1)^{-1}E_1 = (P_1W)^{-1}(P_1E) = W^{-1}{P_1}^{-1}P_1E = W^{-1}E$.

By a similar argument, the list of samples X implied by M_2 is also $W^{-1}E$. Therefore, any two SEM models output by LiNG-D entail the same distribution.

Now, it remains to be shown that the output of LiNG-D is exhaustive.

Suppose that there is a SEM S that represents the same distribution as some T, which is output by LiNG-D. Then the reduced-form coefficient matrices for S and T, A_S and A_T, are the same up to column-permutation and scaling. Hence, $I - B_S$ and $I - B_T$ are also the same up to scaling and row-permutation (by $I - B = A^{-1}$). By the definition of SEM, neither $I - B_T$ nor $I - B_S$ has zeros on the diagonal. Since $I - B_T$ is a scaled row-permutation of W that has no zeros on the diagonal, so is $I - B_S$. Thus S is also output by LiNG-D.

QED.

In general, each candidate model $B' = I - W'$ has the structure of a row-permutation of W. The structures can be generated by analyzing what happens when we permute the rows of W. Remember that edges in B (and thus W) are read column to row. Thus, row-permutations of a model change the positions of the arrow-heads (targets), but not the arrow-tails (sources). When the graph is a simple cycle, the only other graph is the one obtained by reversing the direction of the loop ([9]). Richardson proved that this operation preserves the set of entailed zero partial correlations, but did not consider distribution equivalence [9].

5.3 Adding the assumption of stability

There are cases where the set of equations in a cyclic SEM has a solution, but no interpretation as a dynamical system reaching equilibrium. Such systems are known as "unstable". In dynamical systems, "stable" models are ones in which the effects of one-time noise dissipate. For example, a model that has a single cycle whose cycle-product (product of coefficients of edges in the cycle) is ≥ 1 is unstable, while one that has a single cycle whose cycle-product is between -1 and 1 is stable. On the other hand, if a positive feedback loop of cycle-product 2 is counteracted by a negative loop with cycle-product -1.5, then the model is stable, because the effective cycle-product is 0.5.

A general way to express stability is $\lim_{t\to\infty} B^t = 0$, which is mathematically equivalent to: for all eigenvalues e of B, $|e| < 1$, in which $|z|$ means the modulus of z (eigenvalues can be complex-valued). This eigenvalues criterion is easy to compute.

As discussed previously, it is impossible to measure stability from ICA's output without assuming the absence of self-loops. Therefore, in this section, it is assumed that the true model has no self-loops.

It is often the case that some of the SEMs output by LiNG-D are unstable. In many domains, however, it is common to assume that the true model is stable.

In this section, we will prove that if the SEM generating the population distribution has a graph in which the cycles are disjoint, then among the candidate SEMs output by LiNG-D, at most one will be stable.

Theorem 2 *SEMs in the form of a simple cycle with a cycle-product π such that $|\pi| \geq 1$ are unstable.*

Proof: Let k be the length of the cycle. Then $B^k = \pi I$. Then for all integers i, $B^{ik} = \pi^i I$. So if $|\pi| \geq 1$, the entries of B^{ik} do not get smaller than the entries of B as i increases. Thus, B^t will not converge to 0 as $t \to \infty$.

Theorem 3 *If a SEM M has a graph in the form of a simple cycle C, then there is at most one more SEM M' with simple cycle C' in its distribution-equivalence-class, and the following holds about their cycle-products: $\pi_{M'} = 1/\pi_M$.*

Proof: Without loss of generality, we let M's coefficient matrix have the form:

$$B_M = \begin{bmatrix} 0 & \dots & 0 & b_{k,1} \\ b_{1,2} & 0 & \dots & 0 \\ 0 & b_{2,3} & \ddots & 0 \\ 0 & 0 & \ddots & 0 \end{bmatrix}$$

Note that the cycle-product $\pi_M = b_{k,1} \prod_{i=0}^{k-1} b_{i,i+1}$.

$$W_M = I - B_M = \begin{bmatrix} 1 & 0 & \dots & -b_{k,1} \\ -b_{1,2} & 1 & \dots & 0 \\ 0 & -b_{2,3} & \ddots & 0 \\ 0 & 0 & \ddots & 1 \end{bmatrix}$$

The only other admissible row-permutation of W is the one in which the first row gets pushed to the bottom:

$$RowPermute(W_M) = \begin{bmatrix} -b_{1,2} & 1 & \dots & 0 \\ 0 & -b_{2,3} & \ddots & 0 \\ 0 & 0 & \ddots & 1 \\ 1 & 0 & \dots & -b_{k,1} \end{bmatrix}$$

Normalizing the diagonal to be all 1s, we get $W_{M'}$. Computing $B_{M'} = I - W_{M'}$, one can see that the cycle-product $\pi_{M'} = \frac{1}{b_{k,1}} \prod_{i=0}^{k-1} \frac{1}{b_{i,i+1}} = 1/\pi_M$.

Corollary 1: *If a SEM M has a graph in the form of a simple cycle C, then at most one SEM in the output of LiNG-D is stable.*

Theorem 4 *If a SEM has a graph such that all cycles are disjoint (in the sense that they do not share a node), then at most one SEM in the output of LiNG-D is stable.*

Proof:

By specification, each SEM output by LiNG-D corresponds to a row-permutation of W such that the diagonal contains only non-zeroes. Suppose that there is a row-permutation of W such that the diagonal contains only non-zeroes and corresponds to a stable model; let this permutation of W be called $I - B$.

Claim 1: Every admissible permutation of the rows of $I - B$ is a composition of (0 or more) reversals of cycles in the graph, i.e. permutations that cannot represented as a composition of cycle reversals put a zero in the diagonal.

Let P be an admissible permutation that maps row i onto row j. Then the jth entry in row i of $I - B$ is non-zero, since the diagonal of the permuted matrix has no zeros. Hence, $\mathbf{x}_j$ is a parent of $\mathbf{x}_i$. Thus every admissible permutation-cycle corresponds to a cycle C in the graph, and has the effect of reversing C.

Since every permutation can be represented as a sequence of permutation-cycles that don't intersect each other, it follows that P is a composition of (0 or more) reversals of graph-cycles.

Thus P corresponds to a reversal of a cycle in the graph corresponding to $I - B$.

Claim 2: Every row-permutation of $I - B$ that reverses a cycle corresponds to an unstable model.

By Corollary 1, every cycle-reversal leads to an unstable cycle, and hence, since none of the cycles touch, an unstable model.

From Claims 1 and 2, it follows that $I - B$ is the only stable SEM that is output by LiNG-D.

QED.

This condition is sufficient, but not necessary. It is easy to come up with SEMs where we have exactly one stable SEM in the distribution-equivalence class, despite intersecting cycles.

5.4 Correctness of LiNG-D

Theorem 5 *If the simultaneous equations are linear and can be represented by a directed graph; the error terms have non-zero variance; the samples are inde-*

pendently and identically distributed; no more than one error term is Gaussian; and the error terms are jointly independent, then in the large sample limit, LiNG-D outputs a distribution-equivalence class of SEMs that describe the population distribution.

Proof: ICA gives pointwise consistent estimates of A and W under the assumptions listed [3]. This entails that there are pointwise consistent tests of whether an entry in the W matrix is zero, and hence by definition in the large sample limit, the limit of both type I and type II errors of tests of zero coefficients are zero. Given the correct zeroes in the W matrix, the output of the local version of the LiNG-D algorithm is correct in the sense that the simultaneous equation describes the population distribution.

6 Discussion

We have presented an approach for discovering general LiNG SEMs that improves upon the state-of-the-art by narrowing the output to the distribution-equivalence-class of SEMs and by relaxing the faithfulness assumption. We have also shown that stability can be a powerful constraint, sometimes narrowing the candidates to a single SEM. There are a number of open questions that remain for future research.

- The LiNG-D algorithm generates all admissible permutations. The time-complexity of n-Rooks is high when the correct model has many cycles. Is there an algorithm to efficiently search for the stable models, without going through all candidates? In the case where the cycles are disjoint, it is possible to just find the correct permutation for each cycle independently, but no such trick is known in general.

- How can prior information or a prior distribution be incorporated into the algorithm?

- Can the algorithm be modified to allow the assumption of causal sufficiency assumption to be relaxed?

- Can the algorithm be modified to allow for mixtures of non-Gaussian and Gaussian (or almost Gaussian) error terms?

Acknowledgements

The authors wish to thank Anupam Gupta, Michael Dinitz and Cosma Shalizi.

References

[1] K. Bollen (1989) - *Structural Equations with Latent Variables*, John Wiley & Sons, New York.

[2] C. R. Chegireddy, H. W. Hamacher (1987) - Algorithms for finding K-best perfect matchings *Discrete Applied Mathematics*, **18**:155-165.

[3] P. Comon (1994) - Independent component analysis - a new concept? *Signal Processing*, **36**:287-314.

[4] F. Fisher (1970) - A correspondence principle for simultaneous equation models. *Econometrica,* **38**(**1**)*:73-92.*

[5] A. Hyvärinen, J. Karhunen, E. Oja (2001) - *Independent Component Analysis.* Wiley Interscience.

[6] H. W. Kuhn (1955), The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly*, **2**:83-97, 1955.

[7] J. Pearl (1995) - Causal diagrams for empirical research. *Biometrika,* **82***: 669-709.*

[8] J. Pearl (2000) - *Causality: Models, Reasoning, and Inference.* Cambridge University Press.

[9] T. Richardson (1996) - A Polynomial-Time Algorithm for Deciding Markov Equivalence of Directed Cyclic Graphical Models. *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*, Portland, Oregon, 462-469, Morgan Kauffman.

[10] T. Richardson (1996) - A Discovery Algorithm for Directed Cyclic Graphs. *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*, Portland, Oregon, 454-461, Morgan Kauffman.

[11] S. Shimizu, P. Hoyer, A. Hyvärinen, A. Kerminen (2006) - A linear, non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* **7**:2003-2030.

[12] P. Spirtes, C. Glymour, R. Scheines (1993) - *Causation, Prediction, Search.* Springer-Verlag Lecture Notes in Statistics 81.

[13] R. Strotz and H. Wold (1960) - Recursive versus nonrecursive systems: an attempt at synthesis. *Econometrica* **28**:417-427.

[14] G. Wyatt (2004) - *Macroeconomic Models in a Causal Framework*, Exempla Books.