Journal of Machine Learning Research 13 (2012) 3387-3439 Submitted 10/11; Revised 5/12; Published 11/12

Learning Linear Cyclic Causal Models with Latent Variables

Antti Hyttinen ANTTI.HYTTINEN@HELSINKI.FI
Helsinki Institute for Information Technology
Department of Computer Science
P.O. Box 68 (Gustaf Hällströmin katu 2b)
FI-00014 University of Helsinki
Finland

Frederick Eberhardt FDE@CMU.EDU
Department of Philosophy
Baker Hall 135
Carnegie Mellon University
Pittsburgh, PA 15213-3890, USA

Patrik O. Hoyer PATRIK.HOYER@HELSINKI.FI
Helsinki Institute for Information Technology
Department of Computer Science
P.O. Box 68 (Gustaf Hällströmin katu 2b)
FI-00014 University of Helsinki
Finland

Editor: Christopher Meek

Abstract

Identifying cause-effect relationships between variables of interest is a central problem in science. Given a set of experiments we describe a procedure that identifies linear models that may contain cycles and latent variables. We provide a detailed description of the model family, full proofs of the necessary and sufficient conditions for identifiability, a search algorithm that is complete, and a discussion of what can be done when the identifiability conditions are not satisfied. The algorithm is comprehensively tested in simulations, comparing it to competing algorithms in the literature. Furthermore, we adapt the procedure to the problem of cellular network inference, applying it to the biologically realistic data of the DREAM challenges. The paper provides a full theoretical foundation for the causal discovery procedure first presented by Eberhardt et al. (2010) and Hyttinen et al. (2010).

Keywords: causality, graphical models, randomized experiments, structural equation models, latent variables, latent confounders, cycles

1. Introduction

Inferring causal relationships from data is of fundamental importance in many areas of science. One cannot claim to have fully grasped a complex system unless one has a detailed understanding of how the different components of the system affect each other, and one cannot predict how the system will respond to some targeted intervention without such an understanding. It is well known that a statistical dependence between two measured quantities leaves the causal relation underdetermined—in addition to a causal effect from one variable to another (in either or both directions), the dependence might be due to a common cause (a confounder) of the two.

In light of this underdetermination, *randomized experiments* have become the gold standard of causal discovery. In a randomized experiment, the values of some variable x_i are assigned at random by the experimenter and, consequently, in such an experiment any correlation between x_i and another measured variable x_j can uniquely be attributed to a causal effect of x_i on x_j, since any incoming causal effect on x_i (from x_j, a common cause, or otherwise) would be 'broken' by the randomization. Since their introduction by Fisher (1935), randomized experiments now constitute an important cornerstone of experimental design.

Since the 1980s causal graphical models based on directed graphs have been developed to systematically represent causal systems (Glymour et al., 1987; Verma and Pearl, 1988). In this approach, causal relations among a set of variables $\mathcal{V}$ are represented by a set of directed edges $\mathcal{D} \subseteq (\mathcal{V} \times \mathcal{V})$ connecting nodes in a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{D})$, where a directed edge from node x_i to node x_j in the graph represents the *direct* causal effect of x_i on x_j (relative to the set of variables $\mathcal{V}$). The causal relationships in such a model are defined in terms of stochastic functional relationships (or alternatively conditional probability distributions) that specify how the value of each variable is influenced by the values of its direct causes in the graph. In such a model, randomizing a variable x_i is tantamount to removing all arrows pointing *into* that variable, and replacing the functional relationship (or conditional probability distribution) with the distribution specified in the experiment. The resulting truncated model captures the fact that the value of the variable in question is no longer influenced by its normal causes but instead is determined explicitly by the experimenter. Together, the graph structure and the parameters defining the stochastic functional relationships thus determine the joint probability distribution over the full variable set under any experimental conditions.

The question that interests us here is how, and under what conditions, we can learn (i.e., infer from data) the structure and parameters of such causal models. The answer to this question depends largely on what assumptions we are willing to make about the underlying models and what tools of investigation we consider. For instance, some causal discovery methods require assuming that the causal structure is *acyclic* (has no directed cycles), while others require *causal sufficiency*, that is, that there are no unmeasured common causes affecting the measured variables. Many algorithms provide provably consistent estimates only under the assumption of *faithfulness*, which requires that the structure of the graph uniquely determines the set of (conditional) independencies that hold between the variables. For some methods the functional form of the relationships has to take a certain predetermined form (e.g., linearity). Under various combinations of the above assumptions, it is possible to consistently infer (at least partial information concerning) the causal relationships underlying the observed data from non-experimental ('passive observational') data (Richardson, 1996; Spirtes et al., 2000; Pearl, 2000; Chickering, 2002a,b; Shimizu et al., 2006).

In many cases, researchers may not be willing to make some of the assumptions mentioned above, or they may want to guarantee that the full structure of the model is inferred (as opposed to only inferring an equivalence class of possible models, a common result of many discovery methods). A natural step is thus to use the power of randomized experiments. The question then becomes: Under what assumptions on the model and for what sets of experiments can one guarantee consistent learning of the underlying causal structure. Here, almost all of the existing literature has focused on the acyclic case (Cooper and Yoo, 1999; Tong and Koller, 2001; Murphy, 2001;

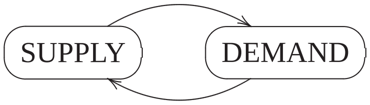


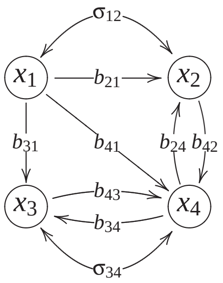
Figure 1: Classic supply-demand model.

Eberhardt et al., 2005; Meganck et al., 2005; Nyberg and Korb, 2006; Eberhardt and Scheines, 2007; Eaton and Murphy, 2007).

The acyclicity assumption, common to most discovery algorithms, permits a straightforward interpretation of the causal model and is appropriate in some circumstances. But in many cases the assumption is clearly ill-suited. For example, in the classic demand-supply model (Figure 1) demand has an effect on supply and vice versa. Intuitively, the true causal structure is acyclic over time since a cause always precedes its effect: Demand of the previous time step affects supply of the next time step. However, while the causally relevant time steps occur at the order of days or weeks, the measures of demand and supply are typically cumulative averages over much longer intervals, obscuring the faster interactions. A similar situation occurs in many biological systems, where the interactions occur on a much faster time-scale than the measurements. In these cases a cyclic model provides the natural representation, and one needs to make use of causal discovery procedures that do not rely on acyclicity (Richardson, 1996; Schmidt and Murphy, 2009; Itani et al., 2008).

In this contribution we consider the problem of learning the structure and parameters of linear cyclic causal models from equilibrium data. We derive a necessary and sufficient condition for identifiability based on second-order statistics, and present a consistent learning algorithm. Our results and learning method *do not* rely on causal sufficiency (the absence of hidden confounding), *nor* do they require faithfulness, that is, that the independencies in the data are fully determined by the graph structure. To our knowledge these results are the first under assumptions that are this weak. Given that the model space is very general (essentially only requiring linearity), randomized experiments are needed to obtain identification. While for certain kinds of experimental data it is easy to identify the full causal structure, we show that significant savings either in the number of experiments or in the number of randomized variables per experiment can be achieved. All-in-all, the present paper provides the full theoretical backbone and thorough empirical investigation of the inference method that we presented in preliminary and abbreviated form in Eberhardt et al. (2010) and Hyttinen et al. (2010). It establishes a concise theory for learning linear cyclic models with latent variables.

We start in Section 2 by introducing the model and its assumptions, how the model is to be interpreted, and how experimental interventions are represented. In Section 3 we derive conditions (on the set of randomized experiments to be performed) that are necessary and sufficient for model identification. These results provide the foundation for the correct and complete learning method presented in Section 4. This section also discusses the underdetermination which results if the identifiability conditions are not met. Section 5 presents empirical results based on thorough simulations, comparing the performance of our procedure to existing methods. Finally, we adapt the procedure to the problem of cellular network inference, and apply it to the biologically realistic *in silico* data of the DREAM challenges in Section 6. Some extensions and conclusions are given in Sections 7 and 8.



$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ b_{21} & 0 & 0 & b_{24} \\ b_{31} & 0 & 0 & b_{34} \\ b_{41} & b_{42} & b_{43} & 0 \end{bmatrix} \quad \mathbf{\Sigma_e} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & \sigma_{34} \\ 0 & 0 & \sigma_{34} & \sigma_4^2 \end{bmatrix}$$

Figure 2: An example of a linear cyclic model with latent variables. A non-zero coefficient b_{21} is represented in the graph by the arc $x_1 \rightarrow x_2$. Similarly, the non-zero covariance between disturbances e_1 and e_2 is represented by the arc $x_1 \leftrightarrow x_2$. In the graph the disturbance term for each individual variable has been omitted for clarity. Note that a pair of opposing directed edges, such as $x_3 \rightarrow x_4$ and $x_3 \leftarrow x_4$, represents reciprocal causation (feedback relationship) between the variables, whereas a double-headed arrow, such as $x_3 \leftrightarrow x_4$, represents confounding.

2. Model

We start by presenting the basic interpretation of the cyclic model in the passive observational (Section 2.1) and experimental settings (Section 2.2). We establish canonical forms for both the model and the experiments to simplify the presentation of the subsequent theory. We then discuss different stability assumptions to ensure the presence of model equilibria, and show how they relate to the model interpretation and model marginalization (Section 2.3).

2.1 Linear Cyclic Model with Latent Variables

Following the framework presented in Bollen (1989), we consider a general linear structural equation model (SEM) with correlated errors as our underlying data generating model. In such a model the value of each observed variable $x_j \in \mathcal{V}$ $(j = 1, \ldots, n)$ is determined by a linear combination of the values of its causal parents $x_i \in \mathrm{pa}(x_j)$ and an additive disturbance ('noise') term e_j:

$$x_j \quad := \sum_{x_i \in \mathrm{pa}(x_j)} b_{ji} x_i + e_j.$$

Representing all the observed variables as a vector $\mathbf{x}$ and the corresponding disturbances as a vector $\mathbf{e}$, these structural equations can be represented by a single matrix equation

$$\mathbf{x} \quad := \quad \mathbf{Bx} + \mathbf{e}, \tag{1}$$

where $\mathbf{B}$ is the $(n \times n)$-matrix of coefficients b_{ji}. A graphical representation of such a causal model is given by representing any non-zero causal effect b_{ji} by an edge $x_i \rightarrow x_j$ in the corresponding graph. An example graph and matrix $\mathbf{B}$ are shown in Figure 2.

The set of equations is said to be *recursive* or *acyclic* if the graph describing the causal relations has no directed cycles, or (equivalently) if there exists a causal order of the variables for which the corresponding matrix $\mathbf{B}$ is lower triangular. When the graph contains directed cycles (feedback loops), such as for the model of Figure 2, then the model is said to be *non-recursive* or *cyclic*. In this paper we do *not* assume a priori that the underlying model is acyclic. In other words, our model family allows for *both* cyclic and acyclic cases.

While in a 'fully observed' SEM the disturbance terms e_i would be assumed to be independent of each other, we allow for unobserved confounding by modeling arbitrary correlations among the disturbances $e_1, ..., e_n$. Specifically, denote by $\boldsymbol{\mu}_\mathbf{e}$ and $\boldsymbol{\Sigma}_\mathbf{e}$ the mean vector and the variance-covariance matrix (respectively) of the disturbance vector $\mathbf{e}$. The diagonal elements of $\boldsymbol{\Sigma}_\mathbf{e}$ represent the variances of the disturbances, while the off-diagonal entries represent the covariances. In the corresponding graph a non-zero covariance between e_i and e_j is represented by the double-headed arc $x_i \leftrightarrow x_j$. Notice that in this implicit representation, a latent variable that confounds three observed variables is represented by three (pairwise) covariances. To keep the notation as simple as possible, we will adopt the assumption standard in the literature that the disturbances have zero mean, that is, $\boldsymbol{\mu}_\mathbf{e} = \mathbf{0}$. In Appendix A we show that it is usually possible to transform the observed data to a form consistent with this assumption. We are thus ready to define the underlying data-generating model:

Definition 1 (Linear Cyclic Model with Latent Variables) *A linear cyclic model with latent variables $\mathcal{M} = (\mathbf{B}, \boldsymbol{\Sigma}_\mathbf{e})$, is a structural equation model over a set of observed variables $x_1, \cdots, x_n \in \mathcal{V}$ of the form of Equation 1, where the disturbance vector $\mathbf{e}$ has mean $\boldsymbol{\mu}_\mathbf{e} = \mathbf{0}$ and an arbitrary symmetric positive-definite variance-covariance matrix $\boldsymbol{\Sigma}_\mathbf{e}$.*

In order to give a fully generative explanation of the relationship between the model parameters and the data, additional constraints on $\mathbf{B}$ are needed. Typically, a cyclic model is used to represent a causal process that is collapsed over the time dimension and where it is assumed that the data sample is taken after the causal process has 'settled down'. The traditional interpretation of non-recursive SEMs assumes that the disturbances represent background conditions that do not change until the system has reached equilibrium and measurements are taken. So for a given set of initial values for the variables $\mathbf{x}(0)$, a data vector is generated by drawing one vector of disturbances $\mathbf{e}$ from the error distribution and iterating the system

$$\mathbf{x}(t) \quad := \quad \mathbf{B}\mathbf{x}(t-1) + \mathbf{e} \tag{2}$$

by adding in the constant (with respect to time) $\mathbf{e}$ at every time step until convergence. At time t the vector $\mathbf{x}$ thus has the value

$$\mathbf{x}(t) \quad := \quad (\mathbf{B})^t \mathbf{x}(0) + \sum_{i=0}^{t-1} (\mathbf{B})^i \mathbf{e}.$$

For $\mathbf{x}(t)$ to converge to an equilibrium, the geometric sequence $(\mathbf{B}^i)_{i=0\ldots t}$ and the geometric series $\sum_{i=0}^{t-1} \mathbf{B}^i$ must converge as $t \to \infty$. For arbitrary $\mathbf{x}(0)$ and arbitrary $\mathbf{e}$, a necessary and sufficient condition for this is that the eigenvalues λ_k of $\mathbf{B}$ satisfy $\forall k : |\lambda_k| < 1$ (Fisher, 1970). In that case $(\mathbf{B})^t \to 0$ and $\sum_{i=0}^{t-1} \mathbf{B}^i \to (\mathbf{I} - \mathbf{B})^{-1}$ as $t \to \infty$, so $\mathbf{x}(t)$ converges to

$$\mathbf{x} \quad = \quad (\mathbf{I} - \mathbf{B})^{-1} \mathbf{e},$$

where $(\mathbf{I} - \mathbf{B})$ is guaranteed to be invertible given the above restriction on the eigenvalues. Notice that the observed value $\mathbf{x}$ at equilibrium is independent of the starting point $\mathbf{x}(0)$, and completely determined by $\mathbf{B}$ and $\mathbf{e}$. Multiple samples of $\mathbf{x}$ are obtained by repeating this equilibrating process for different samples of $\mathbf{e}$. Hence, for $\mathcal{M} = (\mathbf{B}, \mathbf{\Sigma_e})$ the variance-covariance matrix over the observed variables is

$$\mathbf{C_x} = E\{\mathbf{x}\mathbf{x}^T\} = (\mathbf{I}-\mathbf{B})^{-1}E\{\mathbf{e}\mathbf{e}^T\}(\mathbf{I}-\mathbf{B})^{-T} = (\mathbf{I}-\mathbf{B})^{-1}\mathbf{\Sigma_e}(\mathbf{I}-\mathbf{B})^{-T}. \tag{3}$$

The equilibrium we describe here corresponds to what Lauritzen and Richardson (2002) called a *deterministic* equilibrium, since the equilibrium value of $\mathbf{x}(t)$ is fully determined given a sample of the disturbances $\mathbf{e}$. Such an equilibrium stands in contrast to a *stochastic* equilibrium, resulting from a model in which the disturbance term is sampled anew at each time step in the equilibrating process. We briefly return to consider such models in Section 7. We note that if the model happens to be acyclic (i.e., has no feedback loops), the interpretation in terms of a deterministic equilibrium coincides with the standard recursive SEM interpretation, with no adjustments needed.

It is to be expected that in many systems the value of a given variable x_i at time t has a non-zero effect on the value of the same variable at time $t+1$. (For instance, such systems are obtained when approximating a linear differential equation with a difference equation.) In such a case the coefficient b_{ii} (a diagonal element of $\mathbf{B}$) is by definition non-zero, and the model is said to exhibit a 'self-loop' (a directed edge from a node to itself in the graph corresponding to the model). As will be discussed in Section 2.3, such self-loops are inherently unidentifiable from equilibrium data, so there is a need to define a standardized model which abstracts away non-identifiable parameters. For this purpose we introduce the following definition.

Definition 2 (Canonical Model) *A linear cyclic model with latent variables* $(\mathbf{B}, \mathbf{\Sigma_e})$ *is said to be a* canonical model *if it does not contain self-loops (i.e., the diagonal of* $\mathbf{B}$ *is zero).*

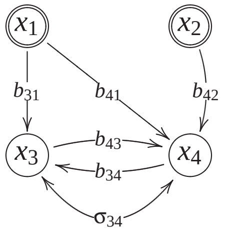
We will show in Section 2.3 how one can obtain the canonical model that yields in all experiments the same observations at equilibrium as an arbitrary (i.e., including self-loops) linear cyclic model with latent variables.

2.2 Experiments

As noted in the introduction, one of the aims of inferring causal models is the ability to predict how a system will react when it is subject to intervention. One key feature of linear cyclic models with latent variables is that they naturally integrate the representation of experimental manipulations, as discussed in this subsection.

We characterize an experiment $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ as a partition of the observed variables $\mathcal{V}$ (i.e., $\mathcal{J}_k \cup \mathcal{U}_k = \mathcal{V}$ and $\mathcal{J}_k \cap \mathcal{U}_k = \emptyset$) into a set $\mathcal{J}_k$ of intervened variables and a set $\mathcal{U}_k$ of passively observed variables. Note that in this representation, a passive observational data set is a 'null-experiment' in which $\mathcal{J}_k = \emptyset$ and $\mathcal{U}_k = \mathcal{V}$. Following the standard view (Spirtes et al., 2000; Pearl, 2000), we consider in this paper randomized "surgical" interventions that break all incoming causal influences to the intervened variables by setting the intervened variables to values determined by an exogenous intervention distribution with mean $\boldsymbol{\mu}_{\mathbf{c}}^k$ and covariance $\text{cov}(\mathbf{c}) = \mathbf{\Sigma}_{\mathbf{c}}^k$. In the graph of the underlying model, this corresponds to cutting all edges into the intervened nodes; see Figure 3 for an example.

To simplify notation, we denote by $\mathbf{J}_k$ and $\mathbf{U}_k$ two $(n \times n)$ diagonal 'indicator matrices', where $(\mathbf{J}_k)_{ii} = 1$ if and only if $x_i \in \mathcal{J}_k$, all other entries of $\mathbf{J}_k$ are zero, and $\mathbf{U}_k = \mathbf{I} - \mathbf{J}_k$. The vector $\mathbf{c}$ represents the values of the intervened variables determined by the intervention distribution, and



$$\mathbf{U}_k\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ b_{31} & 0 & 0 & b_{34} \\ b_{41} & b_{42} & b_{43} & 0 \end{bmatrix} \quad \mathbf{U}_k\boldsymbol{\Sigma}_\mathbf{e}\mathbf{U}_k + \mathbf{J}_k = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & \sigma_{34} \\ 0 & 0 & \sigma_{34} & \sigma_4^2 \end{bmatrix}$$

Figure 3: Manipulated model corresponding to an intervention on variables x_1 and x_2 in the model of Figure 2, that is, the result of an experiment $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ with $\mathcal{J}_k = \{x_1, x_2\}$ and $\mathcal{U}_k = \{x_3, x_4\}$.

is zero otherwise. The behavior of the model in an experiment $\mathcal{E}_k$ is then given by the structural equations

$$\mathbf{x} := \mathbf{U}_k\mathbf{B}\mathbf{x} + \mathbf{U}_k\mathbf{e} + \mathbf{c}. \tag{4}$$

For an intervened variable $x_j \in \mathcal{J}_k$, the manipulated model in Equation 4 replaces the original equation $x_j := \sum_{i \in \mathrm{pa}(j)} b_{ji}x_i + e_j$ with the equation $x_j := c_j$, while the equations for passively observed variables $x_u \in \mathcal{U}_k$ remain unchanged.

Here the intervention vector $\mathbf{c}$ is constant throughout the equilibrating process, holding the intervened variables fixed at the values sampled from the intervention distribution. A different approach could consider interventions that only "shock" the system initially, and then allow the intervened variables to fluctuate. This would require a different representation and analysis from the one we provide here.

As in the passive observational setting discussed in Section 2.1, we have to ensure that the time series representation of the *experimental* setting

$$\mathbf{x}(t) := \mathbf{U}_k\mathbf{B}\mathbf{x}(t-1) + \mathbf{U}_k\mathbf{e} + \mathbf{c}$$

is guaranteed to converge to an equilibrium as $t \to \infty$, where both $\mathbf{c}$ and $\mathbf{e}$ are time-invariant. We do so by extending the assumption that guarantees convergence in the passive observational setting to *all* experimental settings.

Definition 3 (Asymptotic Stability) *A linear cyclic model with latent variables* $(\mathbf{B}, \boldsymbol{\Sigma}_\mathbf{e})$ *is* asymptotically stable *if and only if for every possible experiment* $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$*, the eigenvalues* λ_i *of the matrix* $\mathbf{U}_k\mathbf{B}$ *satisfy* $\forall i : |\lambda_i| < 1$.

Asymptotic stability implies that in an experiment $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ the samples we obtain at equilibrium are given by $\mathbf{x} = (\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}(\mathbf{U}_k\mathbf{e} + \mathbf{c})$. Note that the passive observational case is included in terms of the null-experiment where $\mathcal{J}_k$ is empty. In practice, the assumption of asymptotic stability implies that the system under investigation will not break down or explode under *any* intervention, so the equilibrium distributions are well defined for all circumstances. Obviously, this will not be true for many real feedback systems, and in fact the assumption can be weakened for our purposes. However, as we discuss in more detail in Section 2.3, the assumption of an underlying generating model that satisfies asymptotic stability simplifies the interpretation of our results. For an acyclic model $(\mathbf{B}, \mathbf{\Sigma_e})$ all eigenvalues of all matrices $\mathbf{U}_k\mathbf{B}$ are zero, so the stability condition is in this case trivially fulfilled.

In general, experiments can take many forms: Apart from varying several rather than just one variable at the same time, the interventions on the variables can be independent from one another, or correlated, with different means and variances for each intervened variable. To simplify notation for the remainder of this paper, we will adopt a standardized notion of an experiment:

Definition 4 (Canonical Experiment) *An experiment $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ is said to be a* canonical experiment *if the intervened variables in $\mathcal{J}_k$ are randomized surgically and uncorrelated with the disturbances and with each other, with zero mean and unit variance.*

This notational simplification makes the partition into intervened and passively observed variables the only parameter specifying an experiment, and allows us to derive the theory purely in terms of the covariance matrices $\mathbf{C}^k_{\mathbf{x}}$ of an experiment. The following lemma shows that we can make the assumption of uncorrelated components of $\mathbf{c}$ without loss of generality. First, however, we need one additional piece of notation: For any $(n \times n)$-matrix $\mathbf{A}$, we denote by $\mathbf{A}_{\mathcal{S}_r\mathcal{S}_c}$ the block of $\mathbf{A}$ that remains after deleting the rows corresponding to variables in $\mathcal{V} \setminus \mathcal{S}_r$ and columns corresponding to variables in $\mathcal{V} \setminus \mathcal{S}_c$, keeping the order of the remaining rows and columns unchanged.

Lemma 5 (Correlated Experiment) *If in an experiment $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$, where intervention variables $\mathbf{c}$ are randomized[1] independently of the disturbances $\mathbf{e}$ such that $E(\mathbf{c}) = \boldsymbol{\mu}^k_{\mathbf{c}}$ and $cov(\mathbf{c}) = \mathbf{\Sigma}^k_{\mathbf{c}}$, a linear cyclic model with latent variables $(\mathbf{B}, \mathbf{\Sigma_e})$ produces mean $\tilde{\boldsymbol{\mu}}^k_{\mathbf{x}}$ and covariance matrix $\tilde{\mathbf{C}}^k_{\mathbf{x}}$, then in a canonical experiment where intervention variables $\mathbf{c}$ are randomized independently of $\mathbf{e}$ with $E(\mathbf{c}) = \mathbf{0}$ and $cov(\mathbf{c}) = \mathbf{J}_k$, the model produces observations with mean and covariance given by*

$$\begin{aligned} \boldsymbol{\mu}^k_{\mathbf{x}} &= \mathbf{0}, & (5)\\ \mathbf{C}^k_{\mathbf{x}} &= \tilde{\mathbf{C}}^k_{\mathbf{x}} - \tilde{\mathbf{T}}^k_{\mathbf{x}}(\tilde{\mathbf{C}}^k_{\mathbf{x}})_{\mathcal{J}_k\mathcal{J}_k}(\tilde{\mathbf{T}}^k_{\mathbf{x}})^T + \tilde{\mathbf{T}}^k_{\mathbf{x}}(\tilde{\mathbf{T}}^k_{\mathbf{x}})^T, & (6) \end{aligned}$$

where $\tilde{\mathbf{T}}^k_{\mathbf{x}} = (\tilde{\mathbf{C}}^k_{\mathbf{x}})_{\mathcal{V}\mathcal{J}_k}((\tilde{\mathbf{C}}^k_{\mathbf{x}})_{\mathcal{J}_k\mathcal{J}_k})^{-1}$.

Proof To improve readability, proofs for all lemmas and theorems in this paper are deferred to the appendix.

The lemma shows that whenever in an actual experiment the values given to the intervened variables are not mutually uncorrelated, we can easily convert the estimated mean and covariance matrix to a standardized form that would have been found, had the interventions been uncorrelated with zero mean and unit variance.[2] The substantive assumption is that the values of the intervened

[1] Randomization implies here that the covariance matrix of the intervention variables $\text{cov}(\mathbf{c}_{\mathcal{J}_k}) = (\mathbf{\Sigma}^k_{\mathbf{c}})_{\mathcal{J}_k\mathcal{J}_k}$ is symmetric positive-definite.

[2] The lemma should come as no surprise to readers familiar with multiple linear regression: The $[\bullet, j]$-entries of the matrix $\mathbf{T}^k_{\mathbf{x}}$ are the regression coefficients when x_j is regressed over the intervened variables. The regressors do not have to be uncorrelated to obtain unbiased estimates of the coefficients.

variables (the components of $\mathbf{c}$) are uncorrelated with the disturbances (the components of $\mathbf{e}$). This excludes so-called 'conditional interventions' where the values of the intervened variables depend on particular observations of other (passively observed) variables in the system. We take this to be an acceptably weak restriction.

Mirroring the derivation in Section 2.1, in a *canonical* experiment $\mathcal{E}_k$ the mean and covariance are given by:

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{x}}^k &= \mathbf{0}, & (7)\\ \mathbf{C}_{\mathbf{x}}^k &= (\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}(\mathbf{J}_k+\mathbf{U}_k\boldsymbol{\Sigma}_{\mathbf{e}}\mathbf{U}_k)(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-T}. & (8)\end{aligned}$$

We can now focus on analyzing the covariance matrix obtained from a *canonical* experiment $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ on a *canonical* model $(\mathbf{B}, \boldsymbol{\Sigma}_{\mathbf{e}})$. For notational simplicity we assume without loss of generality that variables $x_1, \cdots, x_j \in \mathcal{J}_k$ are intervened on and variables $x_{j+1}, \cdots, x_n \in \mathcal{U}_k$ are passively observed. The covariance matrix for this experiment then has the block form

$$\mathbf{C}_{\mathbf{x}}^k = \begin{bmatrix} \mathbf{I} & (\mathbf{T}_{\mathbf{x}}^k)^T \\ \mathbf{T}_{\mathbf{x}}^k & (\mathbf{C}_{\mathbf{x}}^k)_{\mathcal{U}_k\mathcal{U}_k} \end{bmatrix}, \qquad (9)$$

where

$$\begin{aligned} \mathbf{T}_{\mathbf{x}}^k &= (\mathbf{I}-\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k})^{-1}\mathbf{B}_{\mathcal{U}_k\mathcal{J}_k},\\ (\mathbf{C}_{\mathbf{x}}^k)_{\mathcal{U}_k\mathcal{U}_k} &= (\mathbf{I}-\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k})^{-1}\left(\mathbf{B}_{\mathcal{U}_k\mathcal{J}_k}(\mathbf{B}_{\mathcal{U}_k\mathcal{J}_k})^T+(\boldsymbol{\Sigma}_{\mathbf{e}})_{\mathcal{U}_k\mathcal{U}_k}\right)(\mathbf{I}-\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k})^{-T}.\end{aligned}$$

The upper left hand block is the identity matrix $\mathbf{I}$, since in a canonical experiment the intervened variables are randomized independently with unit variance. We will consider the more complicated lower right hand block of covariances between the passively observed variables in Section 3.2. The lower left hand block $\mathbf{T}_x^k$ consists of covariances that represent the so-called *experimental effects* of the intervened $x_i \in \mathcal{J}_k$ on the passively observed $x_u \in \mathcal{U}_k$. An experimental effect $t(x_i \rightsquigarrow x_u || \mathcal{J}_k)$ is the overall causal effect of a variable x_i on a variable x_u in the experiment $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$; it corresponds to the coefficient of x_i when x_u is regressed on the set of intervened variables in this experiment. If only variable x_i is intervened on in the experiment, then the experimental effect $t(x_i \rightsquigarrow x_u || \{x_i\})$ is standardly called the *total effect* and denoted simply as $t(x_i \rightsquigarrow x_u)$. If all observed variables except for x_u are intervened on, then an experimental effect is called a *direct effect*: $t(x_i \rightsquigarrow x_u || \mathcal{V} \setminus \{x_u\}) = b(x_i \to x_u) = (\mathbf{B})_{ui} = b_{ui}$.

The covariance between two variables can be computed by so called 'trek-rules'. Some form of these rules dates back to the method of path analysis in Wright (1934). In our case, these trek-rules imply that the experimental effect $t(x_i \rightsquigarrow x_u || \mathcal{J}_k)$ can be expressed as the sum of contributions by all directed paths starting at x_i and ending in x_u in the manipulated graph, denoted by the set $\mathcal{P}(x_i \rightsquigarrow x_u || \mathcal{J}_k)$. The contribution of each path $p \in \mathcal{P}(x_i \rightsquigarrow x_u || \mathcal{J}_k)$ is determined by the product of the coefficients b_{ml} associated with the edges $x_l \to x_m$ on the path, as formalized by the following formula

$$t(x_i \rightsquigarrow x_u || \mathcal{J}_k) = \sum_{p \in \mathcal{P}(x_i \rightsquigarrow x_u || \mathcal{J}_k)} \prod_{(x_l \to x_m) \in p} b_{ml},$$

where the product is taken over all edges $x_l \to x_m$ on the path p. The full derivation of this formula is presented in Appendix C (see also Equation 12a in Mason, 1956).

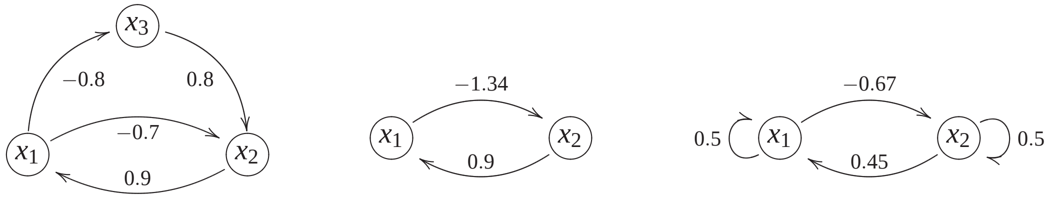


Figure 4: Left: The original asymptotically stable model. Center: The marginalized model that is only weakly stable. Right: A marginalized model with self cycles that is asymptotically stable.

If the model includes cycles, there will be an infinite number of directed paths from one variable to the other. In the example model of Figure 3, the experimental effects can be calculated using the trek-rules as follows:

$$t(x_1 \rightsquigarrow x_3 || \{x_1, x_2\}) = (b_{31} + b_{41}b_{34})(1 + b_{43}b_{34} + (b_{43}b_{34})^2 + \cdots) = \frac{b_{31} + b_{41}b_{34}}{1 - b_{43}b_{34}}, \tag{10}$$

$$t(x_1 \rightsquigarrow x_4 || \{x_1, x_2\}) = (b_{41} + b_{31}b_{43})(1 + b_{43}b_{34} + (b_{43}b_{34})^2 + \cdots) = \frac{b_{41} + b_{31}b_{43}}{1 - b_{43}b_{34}}. \tag{11}$$

The convergence of the geometric series is guaranteed by the assumption of asymptotic stability for the experiment $\mathcal{J}_k = \{x_1, x_2\}$, which ensures that the (only) non-zero eigenvalue $\lambda = b_{43}b_{34}$ satisfies $|\lambda| < 1$.

Note that the experimental effects are unaffected by the latent confounding. Since the interventions break any incoming arrows on the intervened variables, this independence also follows directly from the graphical d-separation criterion extended to cyclic graphs (Spirtes, 1995): In Figure 3, variables x_1 and x_3 are not d-connected by any of the undirected paths through the double headed arrows.

2.3 Marginalization

One of the key features of linear structural equation models with correlated errors is that the model family is closed under marginalization. That is, if instead of the original variable set $\mathcal{V}$ we only have access to a subset $\tilde{\mathcal{V}} \subset \mathcal{V}$ of variables, then if the original model $(\mathbf{B}, \mathbf{\Sigma_e})$ is in the model family, then the marginalized model $(\tilde{\mathbf{B}}, \tilde{\mathbf{\Sigma}}_\mathbf{e})$ over $\tilde{\mathcal{V}}$ is in the family, too. Any directed paths through marginalized variables are transformed into directed edges in $\tilde{\mathbf{B}}$, and any confounding effect of the marginalized variables is integrated into the covariance matrix $\tilde{\mathbf{\Sigma}}_\mathbf{e}$ of the disturbances.

For example, in Figure 4 on the left we show the graph structure and the edge coefficients of an asymptotically stable model $(\mathbf{B}, \mathbf{\Sigma_e})$ over the variables $\mathcal{V} = \{x_1, x_2, x_3\}$. For the purpose of argument, assume that variable x_3 is not observed. We thus want to describe a marginalized model $(\tilde{\mathbf{B}}, \tilde{\mathbf{\Sigma}}_\mathbf{e})$ over just the variables $\tilde{\mathcal{V}} = \{x_1, x_2\}$. Critically, the two models should produce the same observations with respect to the variables x_1 and x_2 in both the passive observational setting and in any experiment intervening on $\{x_1\}$, $\{x_2\}$, or $\{x_1, x_2\}$. In other words, the marginalized model should be such that any observations on $\tilde{\mathcal{V}}$ coincides with those obtained from the original model in all experiments that can be performed in both. Thus, in the experiment intervening on x_1, the experimental effect $t(x_1 \rightsquigarrow x_2 || \{x_1\}) = -0.7 - 0.8 \cdot 0.8 = -1.34$ of the original model should equal the corresponding experimental effect of the marginalized model. If we do not want to add any additional self-cycles, the only possibility is to set $\tilde{b}_{21} = -1.34$. Similarly, we set $\tilde{b}_{12} = t(x_2 \rightsquigarrow x_1 || \{x_2\}) = 0.9$. This gives the model of Figure 4 (center).

Note, however, that while the original model was asymptotically stable (as can easily be seen by computing the eigenvalues of $\mathbf{B}$), the marginalized canonical model is *not* asymptotically stable, as $\tilde{\mathbf{B}}$ has an eigenvalue that is larger than 1 in absolute value. We thus see that when relevant variables are not included in the analysis, asymptotic stability may not hold under marginalization. Fortunately, it turns out that for our purposes of identification a much weaker assumption is sufficient. We term this assumption *weak stability*:

Definition 6 (Weak Stability) *A linear cyclic causal model with latent variables* $(\mathbf{B}, \mathbf{\Sigma_e})$ *is* weakly stable *if and only if for every experiment* $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$, *the matrix* $\mathbf{I} - \mathbf{U}_k\mathbf{B}$ *is invertible.*

Note that the invertibility of matrix $\mathbf{I} - \mathbf{U}_k\mathbf{B}$ is equivalent to matrix $\mathbf{U}_k\mathbf{B}$ not having any eigenvalues equal to exactly 1. (Complex-valued eigenvalues with modulus 1 are allowed as long as the eigenvalue in question is not exactly $1 + 0i$.) Any asymptotically stable model is therefore by definition also weakly stable.

We noted earlier that asymptotic stability is an unnecessarily strong assumption for our context. In fact, weak stability is all that is *mathematically* required for all the theory presented in this article. However, while mathematically expedient, weak stability alone can lead to interpretational ambiguities: Under the time series interpretation of a cyclic model that we presented in Equation 2, a weakly stable model that is not asymptotically stable will fail to have an equilibrium distribution for one or more experiments. While Figure 4 illustrates that asymptotic stability may be lost when marginalizing hidden variables, one cannot in general know whether a learned model that is not asymptotically stable for some experiments corresponds to such an unproblematic case, or whether the underlying system truly is unstable under those experiments.

For the remainder of this article, to ensure a consistent interpretation of any learned model, we assume that there is a true *underlying* asymptotically stable data generating model, possibly including hidden variables—thereby guaranteeing well-defined equilibrium distributions for all experiments. The interpretation of any learned weakly stable model $(\mathbf{B}, \mathbf{\Sigma_e})$ is then only that the distribution *over the observed variables* produced at equilibrium by the true underlying asymptotically stable model has mean and covariance as described by Equations 7 and 8.[3] All equations derived for asymptotically stable models carry over to weakly stable models.[4] In the following two Lemmas, we give the details of how the canonical model over the observed variables is related to the original linear cyclic model in the case of hidden variables and self-cycles (respectively).

The marginalized model of any given linear structural equation model with latent variables can be obtained with the help of the following Lemma.

Lemma 7 (Marginalization) *Let* $(\mathbf{B}, \mathbf{\Sigma_e})$ *be a weakly stable linear cyclic model over the variables* $\mathcal{V}$, *with latent variables. Let* $\mathcal{M} \subset \mathcal{V}$ *denote the set of marginalized variables. Then the marginal-*

[3] Alternatively, one could avoid making this assumption of asymptotic stability of the underlying model, but in that case the predictions of the outcomes of experiments must be conditional on the experiments in question resulting in equilibrium distributions.

[4] The sums of divergent geometric series can be evaluated by essentially extending the summing formula $\sum_{i=0}^{\infty} b^i = \frac{1}{1-b}$ to apply also when $b > 1$ (Hardy, 1949).

ized model $(\tilde{\mathbf{B}}, \tilde{\mathbf{\Sigma}}_{\mathbf{e}})$ over variables $\tilde{\mathcal{V}} = \mathcal{V} \setminus \mathcal{M}$ defined by

$$
\begin{aligned}
\tilde{\mathbf{B}} &= \mathbf{B}_{\tilde{\mathcal{V}}\tilde{\mathcal{V}}} + \mathbf{B}_{\tilde{\mathcal{V}}\mathcal{M}}(\mathbf{I} - \mathbf{B}_{\mathcal{M}\mathcal{M}})^{-1}\mathbf{B}_{\mathcal{M}\tilde{\mathcal{V}}}, \\
\tilde{\mathbf{\Sigma}}_{\mathbf{e}} &= (\mathbf{I} - \tilde{\mathbf{B}})\left[(\mathbf{I} - \mathbf{B})^{-1}\mathbf{\Sigma}_{\mathbf{e}}(\mathbf{I} - \mathbf{B})^{-T}\right]_{\tilde{\mathcal{V}}\tilde{\mathcal{V}}}(\mathbf{I} - \tilde{\mathbf{B}})^{T}
\end{aligned}
$$

is also a weakly stable linear cyclic causal model with latent variables. The marginalized covariance matrix of the original model and the covariance matrix of the marginalized model are equal in any experiments where any subset of the variables in $\tilde{\mathcal{V}}$ are intervened on.

The expressions for $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{\Sigma}}_{\mathbf{e}}$ have simple intuitive explanations. First, the coefficient matrix $\tilde{\mathbf{B}}$ of the marginalized model is given by the existing coefficients between the variables in $\tilde{\mathcal{V}}$ in the original model plus any paths in the original model from variables in $\tilde{\mathcal{V}}$ through variables in $\mathcal{M}$ and back to variables in $\tilde{\mathcal{V}}$. Second, the disturbance covariance matrix $\tilde{\mathbf{\Sigma}}_{\mathbf{e}}$ for the marginalized model is obtained by taking the observed covariances over the variables in $\tilde{\mathcal{V}}$ and accounting for the causal effects among the variables in $\tilde{\mathcal{V}}$, so as to ensure that the resulting covariances in the marginal model equal those of the original model in any experiment.

In addition to marginalizing unobserved variables, we may be interested in deriving the canonical model (i.e., without self-loops) from an arbitrary linear cyclic model with self-loops. This is possible with the following lemma.

Lemma 8 (Self Cycles) *Let $\mathbf{U}_i$ be an $(n \times n)$-matrix that is all zero except for the element $(\mathbf{U}_i)_{ii} = 1$. For a weakly stable model $(\mathbf{B}, \mathbf{\Sigma}_{\mathbf{e}})$ containing a self-loop for variable x_i with coefficient b_{ii}, we can define a model without that self-loop given by*

$$
\begin{aligned}
\tilde{\mathbf{B}} &= \mathbf{B} - \frac{b_{ii}}{1 - b_{ii}}\mathbf{U}_i(\mathbf{I} - \mathbf{B}), \\
\tilde{\mathbf{\Sigma}}_{\mathbf{e}} &= (\mathbf{I} + \frac{b_{ii}}{1 - b_{ii}}\mathbf{U}_i)\mathbf{\Sigma}_{\mathbf{e}}(\mathbf{I} + \frac{b_{ii}}{1 - b_{ii}}\mathbf{U}_i)^{T}.
\end{aligned}
$$

The resulting model $(\tilde{\mathbf{B}}, \tilde{\mathbf{\Sigma}}_{\mathbf{e}})$ is also weakly stable and yields the same observations at equilibrium in all experiments.

Figure 5 shows explicitly the relation of edge strengths in the two models of the lemma. Since we are only rescaling some of the coefficients, the graph structure of the model stays intact, except for the deleted self-loop. The structure of the covariance matrix $\mathbf{\Sigma}_{\mathbf{e}}$ also remains unchanged, with only the ith row and the ith column rescaled. For a model $(\mathbf{B}, \mathbf{\Sigma}_{\mathbf{e}})$ with several self-loops we can apply Lemma 8 repeatedly to obtain a model without any self-loops, which is equivalent to the original model in the sense that it yields the same equilibrium data as the original model for all experiments.

Note that, as with marginalization, the standardization by removal of self-cycles may produce a canonical model that is only weakly stable, and not asymptotically stable, even if the original model was asymptotically stable.

Ultimately, self-loops affect the speed and path to convergence to the equilibrium, but not the equilibrium itself. Our approach will not yield any insight on self-loops, because we do not address the causal process in a time series. However, the indeterminacy regarding self-loops also means that any predictions at equilibrium are not affected by the learned model being represented in canonical form, that is, without the possibly existing self-loops. So, although self-loops are not strictly forbidden for the data generating model, we can present the theory in the following sections entirely in terms of models without them.

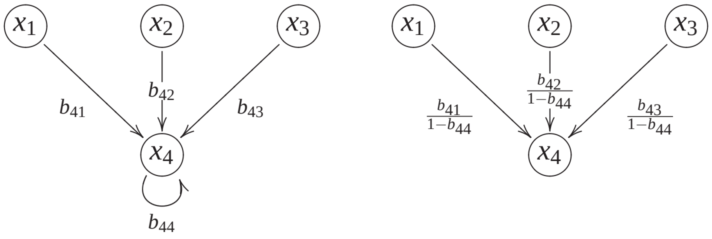

x_1
x_2
x_3
b_{41}
b_{42}
b_{43}
x_4
b_{44}
x_1
x_2
x_3
$\frac{b_{41}}{1-b_{44}}$
$\frac{b_{42}}{1-b_{44}}$
$\frac{b_{43}}{1-b_{44}}$
x_4

Figure 5: Perturbation of coefficients from a model with self-loops (on the left) to a model without self-loops (on the right). The two models are indistinguishable from equilibrium data.

3. Identifiability

The full characterization of the model under passive observational and experimental circumstances now allows us to specify conditions (on the set of experiments) that are sufficient (Section 3.1) and necessary (Section 3.2) to identify the model parameters. Throughout, for purposes of full identification (uniqueness of the solution) and notational simplicity, we assume that in each experiment we observe the covariance matrix in the infinite sample limit as described by Equation 8, and that both the underlying model and all experiments are canonical. For reasons discussed in the previous section we also assume that there is an underlying generating model that is asymptotically stable, even though the marginalized parts of the model we observe may only be weakly stable. Readers who are primarily interested in the learning algorithm we have developed can skip to Section 4 and return to the identifiability conditions of this section when required.

3.1 Sufficiency

Going back to our four variable example in Figure 3, in which x_1 and x_2 are subject to intervention, we already derived in Equations 10 and 11 the experimental effects $t(x_1 \rightsquigarrow x_3 || \{x_1,x_2\})$ and $t(x_1 \rightsquigarrow x_4 || \{x_1,x_2\})$ using the trek-rules. Taken together, these equations imply the following

$$\begin{aligned} t(x_1 \rightsquigarrow x_3 || \{x_1,x_2\}) &= b_{31} + t(x_1 \rightsquigarrow x_4 || \{x_1,x_2\}) b_{34} \\ &= t(x_1 \rightsquigarrow x_3 || \{x_1,x_2,x_4\}) + t(x_1 \rightsquigarrow x_4 || \{x_1,x_2\}) t(x_4 \rightsquigarrow x_3 || \{x_1,x_2,x_4\}). \end{aligned} \tag{12}$$

Note that Equation 12 relates the experimental effects of intervening on $\{x_1,x_2\}$ to the experimental effects of intervening on $\{x_1,x_2,x_4\}$. It shows that the experimental effect $t(x_1 \rightsquigarrow x_3 || \{x_1,x_2\})$ can be calculated by separating the single path *not* going through x_4 (with contribution b_{31}) from the remaining paths that all go through x_4. The last edge on these paths is always $x_4 \to x_3$. The total contribution of the paths through x_4 is therefore the product $t(x_1 \rightsquigarrow x_4 || \{x_1,x_2\}) b_{34}$.

Equation 12 illustrates two separate but related approaches to identifying the full model parameters from a set of measured experimental effects: On the one hand, it provides an example of how experimental effects from one set of experiments can be used to identify experimental effects of a novel experiment (not in the existing set). Thus, if we had a set of experiments that allowed us to infer all the experimental effects of all the experiments that intervene on all but one variable, then we would have determined all the direct effects and would thereby have identified the **B**-matrix. On the other hand, Equation 12 shows how the measured experimental effects can be used to construct *linear* constraints on the (unknown) direct effects b_{ji}. Thus, if we had a set of experiments that supplies constraints that would be sufficient for us to solve for all the direct effects, then we would again be able to identify the **B**-matrix. In either case, the question crucial for identifiability is: Which sets of experiments produce experimental effects that are sufficient to identify the model? Unsurprisingly, the answer is the same for both cases. For reasons of simplicity, we present the identifiability proof in this section in terms of the first approach. We use the second approach, involving a system of linear constraints, for the learning algorithm in Section 4.

The example in Equation 12 can be generalized in the following way: As stated earlier, for an asymptotically stable model, the experimental effect $t(x_i \rightsquigarrow x_u||\mathcal{J}_k)$ of $x_i \in \mathcal{J}_k$ on $x_u \in \mathcal{U}_k$ in experiment $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ is the sum-product of coefficients on all directed paths from x_i to x_u. We can calculate the sum-product in two parts with respect to an observed variable $x_j \in \mathcal{U}_k$. First we consider all the paths that *do not* go through x_j. The sum-product of all those paths is equal to the experimental effect $t(x_i \rightsquigarrow x_u||\mathcal{J}_k \cup \{x_j\})$, since all paths through x_j are intercepted by additionally intervening on x_j. Second, the remaining paths are all of the form $x_i \rightsquigarrow \tilde{x}_j \rightsquigarrow x_u$, where $\tilde{x}_j$ is the last occurrence of x_j on the path (recall that paths may contain cycles, so there may be multiple occurrences of x_j on the path). The sum-product of coefficients on all subpaths $x_i \rightsquigarrow \tilde{x}_j$ is given by $t(x_i \rightsquigarrow x_j||\mathcal{J}_k)$ and the sum-product of coefficients on all subpaths $\tilde{x}_j \rightsquigarrow x_u$ is $t(x_j \rightsquigarrow x_u||\mathcal{J}_k \cup \{x_j\})$. Taking all combinations of subpaths $x_i \rightsquigarrow \tilde{x}_j$ and $\tilde{x}_j \rightsquigarrow x_u$, we obtain the contribution of all the paths through x_j as the product $t(x_i \rightsquigarrow x_j||\mathcal{J}_k)t(x_j \rightsquigarrow x_u||\mathcal{J}_k \cup \{x_j\})$. We thus obtain

$$t(x_i \rightsquigarrow x_u||\mathcal{J}_k) \quad = \quad t(x_i \rightsquigarrow x_u||\mathcal{J}_k \cup \{x_j\}) + t(x_i \rightsquigarrow x_j||\mathcal{J}_k)t(x_j \rightsquigarrow x_u||\mathcal{J}_k \cup \{x_j\}). \tag{13}$$

This equation is derived formally in Appendix F, where it is also shown that it holds for all weakly stable models (not only asymptotically stable models).

We now show that equations of the above type from two different experiments can be combined to determine the experimental effects of a novel third experiment. Consider for example the model in Figure 2 over variables $\mathcal{V} = \{x_1, x_2, x_3, x_4\}$. Say, we have conducted two single-intervention experiments $\mathcal{E}_1 = (\mathcal{J}_1, \mathcal{U}_1) = (\{x_1\}, \{x_2, x_3, x_4\})$ and $\mathcal{E}_2 = (\{x_2\}, \{x_1, x_3, x_4\})$. By making the following substitutions in Equation 13 for each experiment, respectively,

$$\begin{array}{cc} \mathcal{J}_k := \mathcal{J}_1 = \{x_1\} & \mathcal{J}_k := \mathcal{J}_2 = \{x_2\} \\ x_i := x_1 & x_i := x_2 \\ x_j := x_2 & x_j := x_1 \\ x_u := x_3 & x_u := x_3 \end{array}$$

we get two equations relating the experimental effects in the original two experiments to some experimental effects of the *union* experiment $\mathcal{E}_3 = (\{x_1, x_2\}, \{x_3, x_4\})$ (we denote it as the "union" experiment because $\mathcal{J}_3 = \mathcal{J}_1 \cup \mathcal{J}_2$):

$$\begin{bmatrix} 1 & t(x_1 \rightsquigarrow x_2||\{x_1\}) \\ t(x_2 \rightsquigarrow x_1)||\{x_2\}) & 1 \end{bmatrix} \begin{bmatrix} t(x_1 \rightsquigarrow x_3||\{x_1, x_2\}) \\ t(x_2 \rightsquigarrow x_3||\{x_1, x_2\}) \end{bmatrix} \quad = \quad \begin{bmatrix} t(x_1 \rightsquigarrow x_3||\{x_1\}) \\ t(x_2 \rightsquigarrow x_3||\{x_2\}) \end{bmatrix}.$$

In the above equation, the quantities in the matrix on the left, and the elements of the vector on the right-hand-side, are experimental effects that are available from the experimental data. The unknown quantities are in the vector on the left-hand-side. Now, if the matrix on the left is invertible, we can directly solve for the experimental effects of the third experiment just from the experimental effects in the first two. (Similar equations hold for other experimental effects as well). The following lemma shows that the matrix is invertible when the weak stability condition holds, and that in general, from experimental effects observed in two experiments, we can always estimate the experimental effects in their union and in their intersection experiments.

Lemma 9 (Union/Intersection Experiment) *For a weakly stable canonical model the experimental effects in two experiments $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ and $\mathcal{E}_l = (\mathcal{J}_l, \mathcal{U}_l)$ determine the experimental effects in*

- *the* union *experiment $\mathcal{E}_{k \cup l} = (\mathcal{J}_k \cup \mathcal{J}_l,\ \mathcal{U}_k \cap \mathcal{U}_l)$, and*
- *the* intersection *experiment $\mathcal{E}_{k \cap l} = (\mathcal{J}_k \cap \mathcal{J}_l,\ \mathcal{U}_k \cup \mathcal{U}_l)$.*

Since there are no experimental effects in experiments intervening on $\emptyset$ or $\mathcal{V}$, the experimental effects are considered to be determined trivially in those cases. In the case of union experiments, also the full covariance matrix $\mathbf{C}_{\mathbf{x}}^{k \cup l}$ of the experiment can be determined. For intersection experiments, $\mathbf{C}_{\mathbf{x}}^{k \cap l}$ can be fully determined if passive observational data is available (see Appendix J).

In a canonical model the coefficients $b(\bullet \rightarrow x_u)$ on the arcs into variable x_u (the direct effects of the other variables on that variable) are equal to the experimental effects when intervening on everything except x_u, that is, $b(\bullet \rightarrow x_u) = t(\bullet \rightsquigarrow x_u || \mathcal{V} \setminus \{x_u\})$. So in order to determine particular direct effects, it is sufficient to ensure that a given set of experiments provides the basis to apply Lemma 9 repeatedly so as to obtain the experimental effects of the experiments that intervene on all but one variable. In our example with four variables, we can first use Lemma 9 to calculate the experimental effects when intervening on $\{x_1\} \cup \{x_2\} = \{x_1, x_2\}$ (as suggested above), and given a further experiment that intervenes only on x_4, we can then determine the experimental effects of an experiment intervening on $\{x_1, x_2\} \cup \{x_4\} = \{x_1, x_2, x_4\}$. The experimental effects we obtain constitute the direct effects $b(\bullet \rightarrow x_3)$. Hence, if single-intervention experiments are available for each variable it is easy to see that all direct effects of the model are identified using the lemma.

What then is the general condition on the set of experiments such that we can derive all possible direct effects by iteratively applying Lemma 9? It turns out that we can determine all direct effects if the following *pair condition* is satisfied for all ordered pairs of variables.

Definition 10 (Pair Condition) *A set of experiments $\{\mathcal{E}_k\}_{k=1,\ldots,K}$ satisfies the pair condition for an ordered pair of variables $(x_i, x_u) \in \mathcal{V} \times \mathcal{V}$ (with $x_i \neq x_u$) whenever there is an experiment $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ in $\{\mathcal{E}_k\}_{k=1,\ldots,K}$ such that $x_i \in \mathcal{J}_k$ (x_i is intervened on) and $x_u \in \mathcal{U}_k$ (x_u is passively observed).*

It is not difficult to see that the pair condition holding for all ordered pairs of variables is sufficient to identify $\mathbf{B}$. Consider one variable x_u. From a set of experiments satisfying the pair condition for all ordered pairs, we can find for all $x_i \neq x_u$ an experiment satisfying the pair condition for the pair (x_i, x_u). We refer to such an experiment as $\tilde{\mathcal{E}}_i = (\tilde{\mathcal{J}}_i, \tilde{\mathcal{U}}_i)$ in the following. Now, by iteratively using Lemma 9, we can determine the experimental effects in the union experiment $\tilde{\mathcal{E}}_\cup = (\tilde{\mathcal{J}}_\cup, \tilde{\mathcal{U}}_\cup)$ of experiments $\{\tilde{\mathcal{E}}_i\}_{i \neq u}$, where variables in set $\tilde{\mathcal{J}}_\cup = \bigcup_{i \neq u} \tilde{\mathcal{J}}_i$ are intervened on. Each x_i was intervened on at least in one experiment, thus $\forall i \neq u : x_i \in \tilde{\mathcal{J}}_\cup$. Variable x_u was passively observed in each experiment, thus $x_u \notin \tilde{\mathcal{J}}_\cup$. The experimental effects of this union experiment intervening on $\tilde{\mathcal{J}}_\cup = \mathcal{V} \setminus \{x_u\}$ are thus the direct effects $b(\bullet \rightarrow x_u)$. Repeating the same procedure for each $x_u \in \mathcal{V}$ allows us to identify all direct effects.

Thus, if the pair condition is satisfied for all ordered pairs, we can determine all elements of $\mathbf{B}$, and only the covariance matrix $\mathbf{\Sigma_e}$ of the disturbances remains to be determined. The passive observational data covariance matrix $\mathbf{C}_\mathbf{x}^0$ can be estimated from a null-experiment $\mathcal{E}_0 = (\emptyset, \mathcal{V})$. Given $\mathbf{B}$ and $\mathbf{C}_\mathbf{x}^0$ we can solve for $\mathbf{\Sigma_e}$ using Equation 3:

$$\mathbf{\Sigma_e} \quad = \quad (\mathbf{I}-\mathbf{B})\mathbf{C}_\mathbf{x}^0(\mathbf{I}-\mathbf{B})^T. \tag{14}$$

If there is no null-experiment, then the block $(\mathbf{\Sigma_e})_{\mathcal{U}_k,\mathcal{U}_k}$ of the covariance matrix can instead be determined from the covariance matrix in any experiment $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ using Equation 8:

$$(\mathbf{\Sigma_e})_{\mathcal{U}_k\mathcal{U}_k} \quad = \quad [(\mathbf{I}-\mathbf{U}_k\mathbf{B})\mathbf{C}_\mathbf{x}^k(\mathbf{I}-\mathbf{U}_k\mathbf{B})^T]_{\mathcal{U}_k\mathcal{U}_k}. \tag{15}$$

Consequently, given $\mathbf{B}$, we can determine $(\mathbf{\Sigma_e})_{ij} = \sigma_{ij}$ if the following covariance condition is met.

Definition 11 (Covariance Condition) *A set of experiments $\{\mathcal{E}_k\}_{k=1,\ldots,K}$ satisfies the covariance condition for an unordered pair of variables $\{x_i, x_j\} \subseteq \mathcal{V}$ whenever there is an experiment $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ in $(\mathcal{E}_k)_{k=1,\ldots,K}$ such that $x_i \in \mathcal{U}_k$ and $x_j \in \mathcal{U}_k$, that is, both variables are passively observed.*

Similarly to the pair condition, if we know $\mathbf{B}$, and if the covariance condition is satisfied for all pairs of variables, we can identify all covariances in $\mathbf{\Sigma_e}$. Notice that the variances $(\mathbf{\Sigma_e})_{ii}$ can be determined since the assumption includes that each variable x_i must be passively observed at least in one experiment.

Putting the results together we get a sufficient identifiability condition for a canonical model:

Theorem 12 (Identifiability–Sufficiency) *Given canonical experiments $\{\mathcal{E}_k\}_{k=1,\ldots,K}$ a weakly stable canonical model $(\mathbf{B}, \mathbf{\Sigma_e})$ over the variables $\mathcal{V}$ is identifiable if the set of experiments satisfies the pair condition for each ordered pair of variables $(x_i, x_j) \in \mathcal{V}\times\mathcal{V}$ (with $x_i \neq x_j$) and the covariance condition for each unordered pair of variables $\{x_i, x_j\} \subseteq \mathcal{V}$.*

The identifiability condition is satisfied for our four-variable case in Figure 2 by, for example, a set of experiments intervening on $\{x_1,x_2\}, \{x_2,x_4\}, \{x_1,x_4\}$ and $\{x_3\}$. Obviously, a full set of single-intervention experiments or a full set of all-but-one experiments together with a passive observational data set would also do. We return to this issue in Section 4.2.

3.2 Necessity

To show that the conditions of Theorem 12 are not only sufficient but in fact also necessary for identifiability, we consider what happens when the pair condition or the covariance condition is not satisfied for some variable pair. Since the covariance condition only ensures the identifiability of $\mathbf{\Sigma_e}$ when $\mathbf{B}$ is already identified, we start with the more fundamental *pair condition*.

Consider the two models in Figure 6. The models differ in their parameters, and even in their structure, yet produce the same observations in all experiments that do not satisfy the pair condition for the (ordered) pair (x_2, x_4). That is, for any experiment (including a passive observation), for which it is *not the case* that $x_2 \in \mathcal{J}_k$ and $x_4 \in \mathcal{U}_k$, the two models are indistinguishable, despite the fact that for an experiment that *satisfies* the pair condition for (x_2, x_4), the two models will in general have different experimental effects (due to the difference in the direct effect b_{42}). Since the effect due to b_{42} cannot be isolated in the left model without satisfying the pair condition for the pair

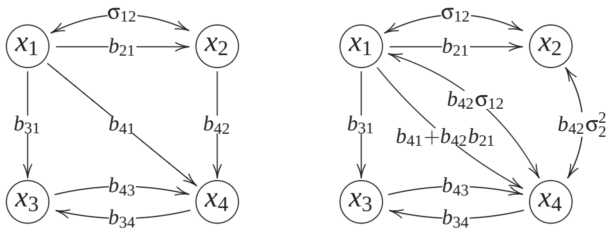


Figure 6: Underdetermination of the model. On the left: the data generating model $(\mathbf{B}, \mathbf{\Sigma_e})$. On the right: a model $(\tilde{\mathbf{B}}, \tilde{\mathbf{\Sigma}}_\mathbf{e})$ producing the same observations in all experiments not satisfying the pair condition for the ordered pair (x_2, x_4).

(x_2, x_4), its effect can be accounted for elsewhere in the right model, for example, the effect of the missing path $x_1 \to x_2 \to x_4$ is accounted for in the model on the right by the perturbed coefficient $b_{41} + b_{42}b_{21}$ on the arc $x_1 \to x_4$.

The $\tilde{\mathbf{B}}$-matrix for the model on the right was constructed from the one on the left by perturbing the coefficient b_{42} corresponding to the pair (x_2, x_4), for which the pair condition is not satisfied. The perturbation corresponds to setting $\delta := -b_{42}$ in the following lemma.

Lemma 13 (Perturbation of B) *Let* $\mathbf{B}$ *be the coefficient matrix of a weakly stable canonical model over* $\mathcal{V}$ *and let* $\{\mathcal{E}_k\}_{k=1,\ldots,K}$ *be a set of experiments on* $\mathbf{B}$ *that does not satisfy the pair condition for some pair* (x_i, x_j)*. Denote the sets* $\mathcal{K} = \mathcal{V} \setminus \{x_i, x_j\}$ *and* $\mathcal{L} = \{x_i, x_j\}$*. Then a model with coefficient matrix* $\tilde{\mathbf{B}}$ *defined by*

$$\tilde{\mathbf{B}}_{\mathcal{K}\mathcal{V}} = \mathbf{B}_{\mathcal{K}\mathcal{V}}, \quad \tilde{\mathbf{B}}_{\mathcal{L}\mathcal{L}} = \begin{bmatrix} 0 & b_{ij} \\ b_{ji}+\delta & 0 \end{bmatrix}, \quad \tilde{\mathbf{B}}_{\mathcal{L}\mathcal{K}} = (\mathbf{I} - \tilde{\mathbf{B}}_{\mathcal{L}\mathcal{L}})(\mathbf{I} - \mathbf{B}_{\mathcal{L}\mathcal{L}})^{-1}\mathbf{B}_{\mathcal{L}\mathcal{K}}$$

will produce the same experimental effects *as* $\mathbf{B}$ *for* any *experiment that does not satisfy the pair condition for the pair* (x_i, x_j)*. The free parameter* δ *must be chosen such that* $\tilde{\mathbf{B}}$ *is weakly stable.*

Lemma 13 shows that if the pair condition is not satisfied for the pair (x_i, x_j), then b_{ji} cannot be identified on the basis of the measured experimental effects. As in our example, it is generally the case that for $\delta \neq 0$ the models $\mathbf{B}$ and $\tilde{\mathbf{B}}$ will produce different experimental effects in any experiment that satisfies the pair condition for the pair (x_i, x_j). The choice of δ is not crucial, since most choices will produce a weakly stable perturbed model.

To see the effect of the perturbation more clearly, we can write it explicitly as follows:

$$\begin{aligned} \forall l \neq j,\, \forall k : \tilde{b}_{lk} &= b_{lk}, && \text{(no changes to any edges that do not end in } x_j\text{)} \\ \tilde{b}_{ji} &= b_{ji} + \delta, && \text{(perturb the direct effect of } x_i \text{ on } x_j \text{ by } \delta\text{)} \\ \tilde{b}_{jj} &= 0, && \text{(no self-loop at } x_j\text{)} \\ \forall k \notin \{i, j\} :\ \tilde{b}_{jk} &= b_{jk} - \delta \frac{b_{ik} + b_{ij}b_{jk}}{1 - b_{ij}b_{ji}}. && \text{(needed adjustments to incoming arcs to } x_j\text{)} \end{aligned}$$

The above form makes it clear that if the pair condition is not satisfied for the pair (x_i, x_j), in general all coefficients on the jth row of $\mathbf{B}$ may be unidentified as well. Hence, to guarantee the identifiability of coefficient b_{ji} we must have the pair condition satisfied for all pairs $(\bullet, x_j)$. In Figure 6 the coefficient b_{42} is unidentified because the pair condition for the pair (x_2, x_4) is not satisfied. But as a result, b_{41} is also unidentified. Nevertheless, in this particular example, the coefficient b_{43} happens to be identified, because of the structure of the graph.

If the pair condition is not satisfied for several pairs, then Lemma 13 can be applied iteratively for each missing pair to arrive at a model with different coefficients, that produces the same experimental effects as the original for all experiments not satisfying the pairs in question. Each missing pair adds an additional degree of freedom to the system.

We emphasize that Lemma 13 only ensures that the *experimental effects* of the original and perturbed model are the same. However, the following lemma shows that the covariance matrix of disturbances can always be perturbed such that the two models become completely indistinguishable for any experiment that does not satisfy the pair condition for some pair (x_i, x_j), as was the case in Figure 6.

Lemma 14 (Perturbation of $\mathbf{\Sigma_e}$) *Let the true model generating the data be $(\mathbf{B}, \mathbf{\Sigma_e})$. For each of the experiments $\{\mathcal{E}_k\}_{k=1,\ldots,K}$, let the obtained data covariance matrix be $\mathbf{C}_\mathbf{x}^k$. If there exists a coefficient matrix $\tilde{\mathbf{B}} \neq \mathbf{B}$ such that for all $\{\mathcal{E}_k\}_{k=1,\ldots,K}$ and all $x_i \in \mathcal{J}_k$ and $x_j \in \mathcal{U}_k$ it produces the same experimental effects $t(x_i \rightsquigarrow x_j \,||\, \mathcal{J}_k)$, then the model $(\tilde{\mathbf{B}}, \tilde{\mathbf{\Sigma}}_\mathbf{e})$ with $\tilde{\mathbf{\Sigma}}_\mathbf{e} = (\mathbf{I} - \tilde{\mathbf{B}})(\mathbf{I} - \mathbf{B})^{-1}\mathbf{\Sigma_e}(\mathbf{I} - \mathbf{B})^{-T}(\mathbf{I} - \tilde{\mathbf{B}})^T$ produces data covariance matrices $\tilde{\mathbf{C}}_\mathbf{x}^k = \mathbf{C}_\mathbf{x}^k$ for all $k = 1, \ldots, K$.*

Lemma 14, in combination with Lemma 13, shows that for identifiability the pair condition must be satisfied for all pairs. If the pair condition is not satisfied for some pair, then an alternative model (distinct from the true underlying model) can be constructed (using the two lemmas) which produces the exact same covariance matrices $\mathbf{C}_\mathbf{x}^k$ for all the available experiments. In Figure 6, the effect of the missing link $x_2 \to x_4$ is imitated by the additional covariance $b_{42}\sigma_2^2$ between e_2 and e_4 and by the covariance $b_{42}\sigma_{12}$ between e_1 and e_4.

The result implies that identifying the coefficient matrix $\mathbf{B}$ exclusively on the basis of constraints based on *experimental effects* already fully exploits the information summarized by the second order statistics. The covariances between the passively observed variables (corresponding to the lower right hand block in Equation 9) do not provide any further information. We thus obtain the result:

Theorem 15 (Completeness) *Given the covariance matrices in a set of experiments $\{\mathcal{E}_k\}_{k=1,\ldots,K}$ over the variables in $\mathcal{V}$, all coefficients $b(x_i \to x_j)$ of a weakly stable canonical model are identified if and only if the pair condition is satisfied for all ordered pairs of variables with respect to these experiments.*

Intuitively, the covariances between the passively observed variables do not help in identifying the coefficients $\mathbf{B}$ because they also depend on the unknowns $\mathbf{\Sigma_e}$, and the additional unknowns swamp the gains of the additional covariance measures.

If $\mathbf{B}$ is known or the pair condition is satisfied for all pairs, but the *covariance condition* is not satisfied for a pair $\{x_i, x_j\}$, then in general the covariance σ_{ij} cannot be identified: In all the manipulated graphs of the experiments the arc $x_i \leftrightarrow x_j$ is cut, and thus σ_{ij} does not affect the data in any way. It follows that the covariance condition is necessary as well. However, unlike for the pair condition, not satisfying the covariance condition for some pair does not affect the identifiability of any of the other covariances.

We can now summarize the previous results in the form of a sufficient and necessary identifiability condition for the full model. Theorem 12 states that satisfying the pair condition and covariance condition for all pairs is sufficient for model identifiability. Theorem 15 shows that the coefficients cannot be identified if the pair condition is not satisfied for all pairs of variables, and in the previous paragraph we showed that satisfying the covariance condition for all pairs is necessary to identify all covariances and variances of the disturbances. This yields the following main result.

Corollary 16 (Model Identifiability) *The parameters of a weakly stable canonical model* $(\mathbf{B},\mathbf{\Sigma_e})$ *over the variables in* $\mathcal{V}$ *can be identified if and only if the set of experiments* $\{\mathcal{E}_k\}_{k=1,\ldots,K}$ *satisfies the pair condition for all ordered pairs* $(x_i,x_j) \in \mathcal{V}\times\mathcal{V}$ *(such that* $x_i \neq x_j$*) and the covariance condition for all unordered pairs* $\{x_i,x_j\} \subseteq \mathcal{V}$*.*

Finally, note that all of our identifiability results and our learning algorithm (Section 4) are solely based on second-order statistics of the data and the stated model space assumptions. No additional background knowledge is included. When the data are multivariate Gaussian, these statistics exhaust the information available, and hence our identifiability conditions are (at least) in this case necessary.

4. Learning Method

In this section, we present an algorithm, termed LLC, for inferring a linear cyclic model with latent variables, provided finite sample data from a set of experiments over the given variable set. Although Lemma 9 (Union/Intersection Experiment) naturally suggests a procedure for model discovery given a set of canonical experiments that satisfy the conditions of Corollary 16 (Model Identifiability), we will pursue a slightly different route in this section. It allows us to not only identify the model when possible, but can also provide a more intuitive representation of the (common) situation when the true model is either over- or underdetermined by the given set of experiments. As before, we will continue to assume that we are considering a set of canonical experiments on a weakly stable canonical model (Definitions 2, 4 and 6). From the discussion in Section 2 it should now be clear that this assumption can be made essentially without loss of generality: Any asymptotically stable model can be converted into a weakly stable canonical model and any experiment can be redescribed as a canonical experiment, as long as the interventions in the original experiment were independent of the disturbances. As presented here, the basic LLC algorithm provides only estimates of the values of all the edge coefficients in $\mathbf{B}$, as well as estimates of the variances and covariances among the disturbances in $\mathbf{\Sigma_e}$. We later discuss how to obtain error estimates for the parameters and how to adapt the basic algorithm to different learning tasks such as structure discovery.

4.1 LLC Algorithm

To illustrate the derivation of the algorithm, we again start with Equation 12, which was derived from the experiment that intervenes on x_1 and x_2 in Figure 3,

$$t(x_1 \rightsquigarrow x_3||\{x_1,x_2\}) \quad = \quad b_{31} + t(x_1 \rightsquigarrow x_4||\{x_1,x_2\})b_{34}.$$

This provides a linear constraint of the measured experimental effects $t(x_1 \rightsquigarrow x_j||\{x_1,x_2\})$ on the unknown direct effects b_{31} and b_{34} into x_3. In general, the experimental effects observed in an experiment $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ can be used to provide linear constraints on the unknown direct effects that, like Equation 12, have the form

$$t(x_i \rightsquigarrow x_u||\mathcal{J}_k) \quad = \quad b_{ui} + \sum_{x_j \in \mathcal{U}_k \setminus \{x_u\}} t(x_i \rightsquigarrow x_j||\mathcal{J}_k)b_{uj}, \qquad (16)$$

where $x_i \in \mathcal{J}_k$ and $x_j, x_u \in \mathcal{U}_k$. Analogously to the equations in Section 3.1, for asymptotically stable models Equation 16 is also naturally interpretable in terms of the sum of paths connecting the variables: The experimental effect of x_i on x_u is a sum of the direct effect of x_i on x_u and the effect of each path from x_i to any other $x_j \in \mathcal{U} \setminus \{x_u\}$, multiplied by the direct connection from that x_j to x_u. (Alternatively, one can also see how Equation 16 is reached by iteratively applying Equation 13.)

Since the covariance matrix $\mathbf{C}_{\mathbf{x}}^k$ of an experiment $\mathcal{E}_k$ contains the experimental effects for all pairs (x_i, x_j) with $x_i \in \mathcal{J}_k$ and $x_j \in \mathcal{U}_k$, each experiment generates $m_k = |\mathcal{J}_k| \times |\mathcal{U}_k|$ constraints of the form of Equation 16. For a set of experiments $\{\mathcal{E}_k\}_{k=1,\dots,K}$ we can represent the constraints as a system of equations linear in the (n^2-n) unknown coefficients b_{ji} in $\mathbf{B}$. (Recall that $b_{ii} = 0$ for all i in canonical models.) We thus have a matrix equation

$$\mathbf{Tb} \quad = \quad \mathbf{t}, \tag{17}$$

where $\mathbf{T}$ is a $((\sum_{k=1}^K m_k) \times (n^2-n))$-matrix of (measured) experimental effects, $\mathbf{b}$ is the (n^2-n)-vector of unknown b_{ji} and $\mathbf{t}$ is a $(\sum_{k=1}^K m_k)$-ary vector corresponding to the (measured) experimental effects on the left-hand side of Equation 16.

Provided that matrix $\mathbf{T}$ has full column rank, we can solve this system of equations for $\mathbf{b}$ and rearrange $\mathbf{b}$ into $\mathbf{B}$ (including the diagonal of zeros). Since any one constraint (e.g., Equation 16) only includes unknowns of the type $b_{u\bullet}$, corresponding to edge-coefficients for edges into some node $x_u \in \mathcal{U}_k$, we can rearrange the equations such that the system of equations can be presented in the following form

$$\begin{bmatrix} \mathbf{T}_{11} & & & \\ & \mathbf{T}_{22} & & \\ & & \ddots & \\ & & & \mathbf{T}_{nn} \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_n \end{bmatrix} = \begin{bmatrix} \mathbf{t}_1 \\ \mathbf{t}_2 \\ \vdots \\ \mathbf{t}_n \end{bmatrix}, \tag{18}$$

where $\mathbf{T}$ is a block diagonal matrix with all entries outside the blocks equal to zero. Instead of solving the equation system in Equation 17 with (n^2-n) unknowns, Equation 18 allows us to separate the system into n blocks each constraining direct effects $b_{u\bullet}$ into a different x_u. We can thus separately solve n equation systems $\mathbf{T}_{uu}\mathbf{b}_u = \mathbf{t}_u$ with $(n-1)$ unknowns in each. The matrix $\mathbf{T}$ has full column rank if and only if all $\mathbf{T}_{uu}$ have full column rank as well.

For example, in the case of the experiment intervening on $\mathcal{J}_k = \{x_1, x_2\}$ of the 4-variable model in Figure 3, we obtain the following experimental covariance matrix:

$$\mathbf{C}_{\mathbf{x}}^k = \begin{bmatrix} 1 & 0 & t(x_1 \rightsquigarrow x_3 || \{x_1,x_2\}) & t(x_1 \rightsquigarrow x_4 || \{x_1,x_2\}) \\ 0 & 1 & t(x_2 \rightsquigarrow x_3 || \{x_1,x_2\}) & t(x_2 \rightsquigarrow x_4 || \{x_1,x_2\}) \\ t(x_1 \rightsquigarrow x_3 || \{x_1,x_2\}) & t(x_2 \rightsquigarrow x_3 || \{x_1,x_2\}) & \mathrm{var}_k(x_3) & \mathrm{cov}_k(x_3,x_4) \\ t(x_1 \rightsquigarrow x_4 || \{x_1,x_2\}) & t(x_2 \rightsquigarrow x_4 || \{x_1,x_2\}) & \mathrm{cov}_k(x_3,x_4) & \mathrm{var}_k(x_4) \end{bmatrix}.$$

This covariance matrix allows us to construct the following four linear constraints on the unknown b's:

$$t(x_1 \rightsquigarrow x_3 || \{x_1,x_2\}) \quad = \quad b_{31} + t(x_1 \rightsquigarrow x_4 || \{x_1,x_2\}) b_{34}, \tag{19}$$

$$t(x_1 \rightsquigarrow x_4 || \{x_1,x_2\}) \quad = \quad b_{41} + t(x_1 \rightsquigarrow x_3 || \{x_1,x_2\}) b_{43}, \tag{20}$$

$$t(x_2 \rightsquigarrow x_3 || \{x_1,x_2\}) \quad = \quad b_{32} + t(x_2 \rightsquigarrow x_4 || \{x_1,x_2\}) b_{34}, \tag{21}$$

$$t(x_2 \rightsquigarrow x_4 || \{x_1,x_2\}) \quad = \quad b_{42} + t(x_2 \rightsquigarrow x_3 || \{x_1,x_2\}) b_{43}. \tag{22}$$

If we have a further experiment $\mathcal{E}_l = (\mathcal{J}_l, \mathcal{U}_l)$ with $\mathcal{J}_l = \{x_4\}$ then we obtain the following three additional constraints:

$$\begin{aligned} t(x_4 \rightsquigarrow x_1 \| \{x_4\}) &= b_{14} + t(x_4 \rightsquigarrow x_2 \| \{x_4\}) b_{12} + t(x_4 \rightsquigarrow x_3 \| \{x_4\}) b_{13}, &(23)\\ t(x_4 \rightsquigarrow x_2 \| \{x_4\}) &= b_{24} + t(x_4 \rightsquigarrow x_1 \| \{x_4\}) b_{21} + t(x_4 \rightsquigarrow x_3 \| \{x_4\}) b_{23}, &(24)\\ t(x_4 \rightsquigarrow x_3 \| \{x_4\}) &= b_{34} + t(x_4 \rightsquigarrow x_1 \| \{x_4\}) b_{31} + t(x_4 \rightsquigarrow x_2 \| \{x_4\}) b_{32}. &(25) \end{aligned}$$

Converting the Equations 19-25 to the form of the Equation 18, we see that Equations 19, 21 and 25 become part of $\mathbf{T}_{33}$, while Equations 20 and 22 become part of $\mathbf{T}_{44}$, and the remaining Equations 23 and 24 become part of $\mathbf{T}_{11}$ and $\mathbf{T}_{22}$, respectively. We will focus on $\mathbf{T}_{33}$ consisting of Equations 19, 21 and 25:

$$\begin{aligned} \mathbf{T}_{33}\mathbf{b}_3 &= \begin{bmatrix} 1 & 0 & t(x_1 \rightsquigarrow x_4 \| \{x_1, x_2\}) \\ 0 & 1 & t(x_2 \rightsquigarrow x_4 \| \{x_1, x_2\}) \\ t(x_4 \rightsquigarrow x_1 \| \{x_4\}) & t(x_4 \rightsquigarrow x_2 \| \{x_4\}) & 1 \end{bmatrix} \begin{bmatrix} b_{31} \\ b_{32} \\ b_{34} \end{bmatrix} \\ &= \begin{bmatrix} t(x_1 \rightsquigarrow x_3 \| \{x_1, x_2\}) \\ t(x_2 \rightsquigarrow x_3 \| \{x_1, x_2\}) \\ t(x_4 \rightsquigarrow x_3 \| \{x_4\}) \end{bmatrix} = \mathbf{t}_3. \end{aligned}$$

Given Lemma 9 (Union/Intersection Experiment) it should now be clear that the experimental effects of experiments $\mathcal{E}_k$ and $\mathcal{E}_l$ are sufficient to determine the experimental effects of an experiment intervening on $\mathcal{J} = \mathcal{V} \setminus \{x_3\}$, which would directly specify the values for b_{31}, b_{32} and b_{34}. Unsurprisingly, the matrix $\mathbf{T}_{33}$ is invertible and the coefficients b_{31}, b_{32} and b_{34} can be solved also from the above equation system. In Appendix K we show formally that when the pair condition is satisfied for *all* ordered pairs, then $\mathbf{T}$ has full column rank.

Once we have obtained $\mathbf{B}$ using the above method, the covariance matrix $\mathbf{\Sigma_e}$ can be obtained easily using Equation 14 if a null-experiment $\mathcal{E}_0 = (\emptyset, \mathcal{V})$ is available, or else using Equation 15 in the more general case where only the covariance condition is satisfied for all pairs.

Until now, we have described the algorithm in terms of the covariances and the experimental effects 'observed' in a given experiment. In practice, of course, we only have finite sample data, and the above quantities must be *estimated* from the data, and the estimated covariances and experimental effects do not precisely equal their true underlying values. This naturally has practical ramifications that we describe in the context of the algorithm below.

The LLC algorithm (Algorithm 1), for models that are *linear*, may have *latent* variables and may contain *cycles*, gathers the ideas described so far in this section. It omits all but the most rudimentary handling of the inevitable sampling variability in the estimates. The algorithm minimizes the sum of squared errors in the available linear constraints by solving the equation system using the Moore-Penrose pseudo-inverse. Thus, whenever the linear constraints derived from different experiments are partly conflicting, the algorithm will find a compromise that comes as close as possible to satisfying all the available constraints. Similarly, to improve the statistical estimation of $\mathbf{\Sigma_e}$, we average over all the instances when a particular pair of variables was passively observed. When the covariance condition is not satisfied for a particular pair, then the covariance of the disturbances for that pair remains undefined.

There are several standard modifications that can be made to this basic algorithm in light of statistical variability of the finite sample data. Whenever the sample size differs substantially between experiments, a re-weighting of the constraint equations according to the sample size of the experiment they were obtained from, favors the more precise constraints. Simple bootstrapping of the observed samples in each experiment separately, can be used to obtain rough estimates of error for the *identified* parameters. In Section 6.2 we calculate a Z-score from these error estimates, which in turn is used for structure discovery. Finally, some form of regularization can help to avoid overfitting (see Sections 6.2 and 6.3). Although we have presented the LLC algorithm here in its stripped down form to illustrate its main contribution, the code implementation[5] provides various options for using these additional features.

When the pair condition is not satisfied for all ordered pairs, then **T** does not provide a sufficient set of constraints and the model is underdetermined.[6] Nevertheless, some inferences about the model are still possible. We discuss the details in the following section on underdetermination. For now, note that the algorithm also outputs a list of pairs that satisfy the pair condition, and a list of pairs that satisfy the covariance condition. We will show that these can be used to characterize the underdetermination.

We thus have an algorithm that fully exploits the set of available experiments: When the model identifiability conditions are satisfied it returns an estimate of the true model, when the system is overdetermined it finds a compromise to the available constraints, and when the model is underdetermined we show in the next section what can and cannot be recovered, and how one may proceed in such circumstances.

4.2 Underdetermination

Even when the set of experiments does not satisfy the pair condition for all ordered pairs of variables, the LLC algorithm will nevertheless return a model with estimates for all the coefficients. If there were no sampling errors, one could then check the null-space of the **T**-matrix to identify which entries of **B** are actually underdetermined: An element of **B** is determined if and only if it is orthogonal to the null-space of **T**. In some cases one may find that specific coefficients are determined due to particular values of other coefficients even though that was not clear from the satisfied pair conditions. The coefficient b_{43} in the example in Figure 6 (see the discussion following Lemma 13) is a case in point.

In practice, however, using the null-space to identify the remaining underdetermination can be misleading. The constraints in **T** are based on estimates and so its null-space may not correctly identify which coefficients are determined. One can take a more conservative approach and treat any b_{jk} as undetermined for all k whenever there exists an i such that the pair condition is not fulfilled for the ordered pair (x_i, x_j). This follows from the fact that perturbing the model according to Lemma 13 (Perturbation of **B**) with respect to pair (x_i, x_j), may change all coefficients of the form $b_{j\bullet}$, while leaving the observed experimental effects unchanged. Similarly, the fifth step of the algorithm implements a conservative condition for the identifiability of the covariance matrix: covariance σ_{ij} can be treated as determined if the covariance condition is satisfied for the pair $\{x_i, x_j\}$ *and* the direct effects $\mathbf{B}_{\{x_i,x_j\},\mathcal{V}}$ are determined. Depending on which parameters are identified, Lemma 9 (Union/Intersection Experiment) can be used to make consistent predictions of the

[5] Code implementing the learning algorithm is available at `http://www.cs.helsinki.fi/u/ajhyttin/exp/`.

[6] Because of statistical variability, **T** may well have full rank even in this case, but some of the dimensions it spans only represent errors in the estimates rather than information about the coefficients. See Section 4.2 for details.

Algorithm 1 LLC algorithm

```
1. Input data from a set of experiments $\{\mathcal{E}_k\}_{k=1,\dots,K}$. Initialize matrix $\mathbf{T}$ and vector $\mathbf{t}$ as empty.

2. Using $\{\mathcal{E}_k\}_{k=1,\dots,K}$, determine which ordered pairs of variables satisfy the pair condition and
   which pairs of variables satisfy the covariance condition.

3. For each experiment $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$:

    (a) Estimate the covariance matrix $\mathbf{C}_{\mathbf{x}}^k$.
    (b) From the estimated covariance matrix, extract the experimental effects
        $t(x_i \rightsquigarrow x_u || \mathcal{J}_k)$ for all $(x_i, x_u) \in \mathcal{J}_k \times \mathcal{U}_k$.
    (c) For each pair $(x_i, x_u) \in \mathcal{J}_k \times \mathcal{U}_k$ add an equation

            $$b_{ui} + \sum_{x_j \in \mathcal{U}_k \setminus \{x_u\}} t(x_i \rightsquigarrow x_j || \mathcal{J}_k) b_{uj} \quad = \quad t(x_i \rightsquigarrow x_u || \mathcal{J}_k)$$

        into the system $\mathbf{T}\mathbf{b} = \mathbf{t}$.

4. Solve the equations by $\mathbf{b} = \mathbf{T}^{\dagger}\mathbf{t}$, where $\mathbf{T}^{\dagger}$ is the Moore-Penrose pseudo-inverse of $\mathbf{T}$, and
   rearrange $\mathbf{b}$ to get $\mathbf{B}$.

5. For any pair $\{x_i, x_j\} \subseteq \mathcal{V}$ calculate the covariance of the disturbances as a mean of the co-
   variances estimated in those experiments $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ where $\{x_i, x_j\} \subseteq \mathcal{U}_k$, by

        $$(\mathbf{\Sigma_e})_{ij} \quad = \quad \mathrm{mean}(\{((\mathbf{I} - \mathbf{U}_k\mathbf{B})\mathbf{C}_{\mathbf{x}}^k(\mathbf{I} - \mathbf{U}_k\mathbf{B})^T)_{ij} \,|\, \{x_i, x_j\} \subseteq \mathcal{U}_k\}),$$

   including variances when $x_i = x_j$. (The mean is undefined for a particular pair if the covari-
   ance condition is not satisfied for that pair.)

6. Output the estimated model $(\mathbf{B}, \mathbf{\Sigma_e})$, a list of ordered pairs of variables for which the pair
   condition is not satisfied, and a list of pairs of variables for which the covariance condition is
   not satisfied.
```

experimental effects or the entire covariance matrix for union- or intersection[7] experiments of the available experiments even if the set of experiments does not satisfy the identifiability conditions.

Instead of characterizing the underdetermination, one may consider how to satisfy the model identifiability conditions. There are two general approaches one could pursue. One approach is to strengthen the underlying assumptions, the other to perform additional experiments. Taking the first approach, the additional assumptions may be domain specific or domain general. In econometrics it is common to include background knowledge of the domain that excludes the presence of certain edges, that is, certain edge coefficients are known to be zero. *Faithfulness*, on the other hand, is an assumption we did not make, but that is widely used in causal discovery algorithms (Spirtes et al., 2000). For the linear models we consider here, the assumption of faithfulness requires that a zero-

[7] We note that to fully determine the covariance matrix $\mathbf{C}_{\mathbf{x}}^{k \cap l}$ in an *intersection* experiment, one may require additional passive observational data. See the discussion following Lemma 9.

covariance between two variables entails the absence of a causal connection between the variables. While reasonable for many circumstances, there are well-known cases where faithfulness is not satisfied. For example, if two or more paths between two variables cancel each other out exactly, then one would find a zero-covariance between the variables despite the fact that the variables are (multiply!) causally connected. Moreover, if the data is noisy, a close to unfaithful causal relation may not be distinguishable from an unfaithful one unless a large amount of data or particular experiments are available. Nevertheless, if faithfulness is judged to be a reasonable assumption, then it can provide additional constraints. We have discussed the integration of faithfulness and background knowledge into the current framework in Hyttinen et al. (2010). It remains, however, an open task to develop a procedure for linear cyclic models with latent variables that is *complete* with regard to the additional inferences one can draw on the basis of faithfulness.

If one is able to perform additional experiments, an obvious strategy is to select the next experiment such that it maximizes the number of additional pair conditions that are satisfied. If experiments that intervene on multiple variables simultaneously are taken into consideration, a brute force search for such a best experiment will be exponential in the number of variables. In that case one may consider more efficient selection strategies or heuristics. In most cases any additional experiment will also repeat tests for pairs for which the pair condition is already satisfied. When included in Equation 18, constraints derived from such tests can make the inference more reliable, so one may deliberately select experiments to include particular repeats.

A selection of experiments that is greedy with respect to the satisfaction of additional pair conditions will not necessarily result in the minimum number of experiments overall. For example, if one has six variables $x_1,\ldots,x_6$, and no pair condition has been satisfied so far, that is, no experiment has been performed, then a greedy strategy may recommend a sequence of five intervention sets to fulfill the pair condition for all pairs:

$$\mathcal{J}_1=\{x_1,x_2,x_3\},\mathcal{J}_2=\{x_4,x_5,x_6\},\mathcal{J}_3=\{x_1,x_4\},\mathcal{J}_4=\{x_2,x_5\},\mathcal{J}_5=\{x_3,x_6\}.$$

However, the following four intervention sets are sufficient to satisfy the pair condition for all pairs, but would not be selected by any procedure that is greedy in this respect:

$$\mathcal{J}_1=\{x_1,x_2,x_3\},\mathcal{J}_2=\{x_3,x_4,x_5\},\mathcal{J}_3=\{x_5,x_6,x_1\},\mathcal{J}_4=\{x_2,x_4,x_6\}.$$

The optimal selection of experiments (given possible background knowledge) is closely related to the theory in combinatorics of finding so-called ‘minimal completely separating systems’ for directed graphs (see Hyttinen et al., 2012 and Spencer, 1970 for some relevant results). A full discussion here is beyond the scope of this paper.

From a statistical perspective we have found that intervening on more variables simultaneously leads to a higher accuracy of the estimates even if the total sample size across all experiments is maintained constant (Eberhardt et al., 2010). That is, for two sets of experiments that each satisfy the pair condition for all pairs of variables (e.g., the set of four experiments on six variables above versus a set of six experiments each intervening on a single variable), the sequence of experiments intervening on multiple variables simultaneously will provide a better estimate of the underlying model even if the total sample size is the same.

5. Simulations

We compared the performance of the LLC-algorithm against well-known learning algorithms able to exploit experimental data. Since there is no competing procedure that applies directly to the search space including cyclic *and* latent variable models, we chose for our comparison two procedures that could easily be adapted to the experimental setting and that would provide a good contrast to illustrate the performance of LLC under different model space assumptions. As baseline we used the learning procedure by Geiger and Heckerman (1994) for acyclic Bayesian networks with linear Gaussian conditional probability distributions, referred to as GH. Experimental data is incorporated into the calculation of the local scores in GH using the technique described by Cooper and Yoo (1999). Given that GH assumes acyclicity and causal sufficiency (the absence of latent confounding), it provides a useful basis to assess the increased difficulty of the task when these assumptions are dropped. We also compare to an algorithm for learning Directed Cyclic Graphical models (DCG, Schmidt and Murphy, 2009), designed for discrete cyclic causal models without latent confounding. In this model, the passively observed distribution is represented as a globally normalized product of potentials

$$P(x_1,\ldots,x_n)=\frac{1}{Z}\prod_{i=1}^{n}\phi(x_i;x_{\mathrm{pa}(i)}),$$

where Z is a global normalizing constant. By using unnormalized potentials instead of normalized conditional probability distributions, cycles are allowed in the graph structure. Experimental data is then modeled by simply dropping the potentials corresponding to manipulated variables from the expression, resulting in a manipulated distribution, such as, for example,

$$P(x_2,\ldots,x_n||x_1)=\frac{1}{Z'}\prod_{i=2}^{n}\phi(x_i;x_{\mathrm{pa}(i)}),$$

with a new normalizing constant Z'. Schmidt and Murphy (2009) use potentials of the form $\phi(x_i;x_{\mathrm{pa}(i)})=\exp(b_i(x_i)+\sum_{j\in\mathrm{pa}(i)}w_{ij}(x_i,x_j))$ to model discrete data and learn the model by maximizing the penalized likelihood function using numerical optimization techniques. To fit this approach we discretized the continuous data (at the very end of the data-generating process) to binary data using 0 as threshold value. While the DCG model may be useful in analyzing cyclic systems under intervention, one should note that the underlying causal generative process is not very clear. Certainly, our data generating processes do not in general yield distributions that fit the model family of DCG.

At first glance, it would appear natural to consider two further procedures for comparison: the Cyclic Causal Discovery algorithm (CCD, Richardson, 1996) that allows for cycles but not latent variables, and the Fast Causal Inference algorithm (FCI, Spirtes et al., 2000) that allows for latents but not for cycles. Both are based on conditional independence tests and return equivalence classes of causal models. However, while background knowledge can be integrated into both procedures to learn from a single experimental data set, it is not clear how (possibly conflicting) results from different experiments should be combined. Identifying the appropriate combining procedure for these algorithms would thus require a separate analysis. The approach by Claassen and Heskes (2010) provides some steps in this direction with regard to FCI, but their framework does not quite fit our context since in their framework the interventions are not targeted at particular variables. We considered a comparison with the recent proposal by Itani et al. (2008), but as of this writing no fully automated procedure was available to the present authors.

To compare the LLC- with the GH- and DCG-algorithms we considered models under five different conditions:

1. linear acyclic models without latent variables,
2. linear cyclic models without latent variables,
3. linear acyclic models with latent variables,
4. linear cyclic models with latent variables, and
5. non-linear acyclic models without latent variables.

For each condition we randomly generated 20 causal models with 10 observed variables each. In the underlying graphs each node had 0-3 parents. In models with latent variables, there were 5 additional latent variables, exogenous to the 10 observed variables. The structural equations were of the form

$$x_j \quad := \quad \sum_{i \in \mathrm{pa}(j)} (b_{ji} x_i + a_{ji} x_i^2) + e_j,$$

where[8] $e_j \sim N(0, \sigma_j^2)$, $b_{ji} \sim \pm\mathrm{Unif}(0.2, 0.8)$ and $a_{ji} = 0$ except for the fifth condition with non-linear models where $a_{ji} \sim \mathrm{Unif}(-0.2, 0.2)$. For the second and fourth condition we sampled until we obtained models that contained at least one cycle. From each model we collected samples in the passive observational setting (null experiment) and in ten additional experiments, each intervening on a single (but different) variable. The intervened variables were always randomized to a normal distribution with zero mean and unit variance. The total number of samples (1,000 to 100,000) were divided evenly among the 11 different experiments, so that adjustments to account for the fact that one experiment may provide more accurate estimates than another were unnecessary. Note that the described set of experiments satisfies the identifiability condition for the LLC-method in Theorem 12 (Identifiability–Sufficiency).

There are a variety of ways to assess the output of the algorithms. Given that every test condition violates at least one of the assumptions of one of the algorithms being tested, we decided against a direct comparison of the quantitative output of each procedure. Instead we used the same qualitative measure that is applied in the cellular network inference challenge that we consider as a case study in Section 6. Following Stolovitzky et al. (2009), the simulations were designed such that each method was required to output a list of all possible edges among the observed variables, sorted in decreasing order of confidence that an edge is in the true graph. To this end, we adapted the three algorithms in the following way. For LLC, the edges were simply ranked from highest to lowest according to the absolute value of their learned coefficients in **B**. Although the magnitude of a coefficient does not directly represent the confidence in the presence of the edge, we found empirically that it worked quite well in the simulations. (See Section 6 for an alternative approach based on resampling.) For GH, we calculated the marginal edge probabilities over all DAG structures (with an in-degree bound of 3) using the dynamic programming algorithm of Koivisto and Sood (2004), thus obtaining a score for the confidence in each possible edge. Given that DCG uses binary variables, each edge is associated with four weights: $w_{ij}(0,0)$, $w_{ij}(0,1)$, $w_{ij}(1,0)$ and $w_{ij}(1,1)$. Since the weights were penalized (with regularization parameter λ), an edge $x_j \rightarrow x_i$ is absent whenever the four associated weights are zero. Following Schmidt and Murphy (2009), we used the L^2-norm of the weights for each edge to determine its strength and hence its rank. As with LLC, this seemed to work well to generate the order.

[8] Although the disturbances e_j are uncorrelated in the data generating model, the disturbances of the learned model are in fact correlated when some of the original variables are considered unobserved.

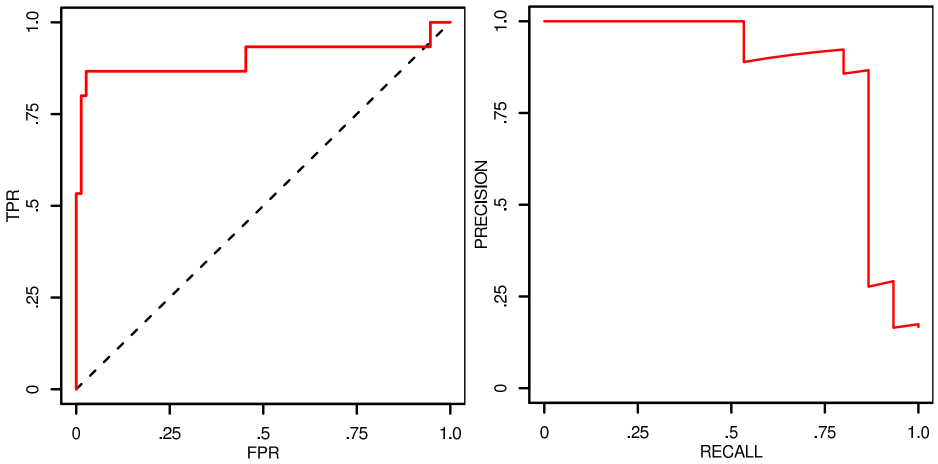


Figure 7: Examples of ROC- (left) and PR-curves (right) of the output of LLC run on 1,000 samples evenly divided over 11 experiments on a linear acyclic model without latents (condition 1).

Given the ordered lists of all possible edges, we can obtain a binary prediction for the presence or absence of an individual edge by simply defining a threshold above which edges would be predicted to be present. These binary predictions can then be compared with the ground truth of the underlying model. However, since the selection of the threshold is to some extent arbitrary (and requires domain specific knowledge of the general sparsity of the generating models), we follow the common approach of reporting Receiver Operating Characteristic (ROC) curves and Precision Recall (PR) curves, and areas under these curves, as explained below. This evaluation of the simulations is also consistent with the evaluation of the case study in Section 6.

A ROC-curve (Figure 7, left) is drawn by plotting the true positive rate (TPR) against the false positive rate (FPR) for different values of the threshold score, where

$$\begin{aligned} \text{TPR} &= \frac{\text{\# edges correctly predicted to be present}}{\text{\# edges in generating model}}, \\ \text{FPR} &= \frac{\text{\# edges incorrectly predicted to be present}}{\text{\# edges not in generating model}}. \end{aligned}$$

The ROC-curve for a powerful classification method should reach close to the top left corner (perfect classification) for some threshold value of the score, while classifying at random would result in the dashed curve in Figure 7. The area under the ROC-curve (AUROC) is often used as a simple one-figure score to assess the power of a classification algorithm. When discovering causal edges in our setting, the AUROC-value specifies the probability that a random edge present in the true model will obtain a higher score than a random absent edge. The AUROC-value usually ranges from 0.5 (random classification) to 1.0 (perfect classification).

Another measure of the quality of search algorithms examines the trade-off between Precision and Recall on a PR-curve (Figure 7, right), where

$$\begin{aligned}\text{Precision} &= \frac{\text{\# edges correctly predicted to be present}}{\text{\# edges predicted to be present}},\\ \text{Recall} &= \frac{\text{\# edges correctly predicted to be present}}{\text{\# edges in generating model}}.\end{aligned}$$

A perfect classification algorithm should have a precision of 1 for all recall values. The area under the PR-curve (AUPR) specifies the average precision over different threshold values of the score, and can range from 0.0 to 1.0 (perfect classification).

Figure 8 shows the results of our simulations. For DCG we ran the algorithm with several regularization parameter values ($\lambda = 2^8, 2^7, \ldots, 2^{-7}, 2^{-8}$), and always report the best AUROC- and AUPR-score. LLC and GH are run without any further tuning. In the first condition (linear acyclic models without latents), all methods seem to learn the correct causal structure as the sample size increases. For small sample sizes the GH approach benefits from the use of Bayesian priors. Such priors could also be added to the LLC-algorithm, if better performance is needed for very low sample sizes. In the other conditions GH does not achieve good results even with large sample sizes. The performance of GH actually tends to get worse with increasing sample size because the method starts adding incorrect edges to account for measured correlations that cannot be fit otherwise, since the generating model is not included in the restricted model class GH uses. In contrast, LLC suffers at low sample sizes at least in part because of the larger model class it considers. In the second (cyclic models without latents), third (acyclic models with latents) and fourth condition (cyclic models with latent variables), both LLC and DCG find quite good estimates of the causal structure, when sufficient samples are available. Some inaccuracies of the DCG-method are due to the discretization of the data. The performance of DCG in the presence of latent confounding is surprisingly good given that the DCG model does not represent latent variables explicitly. The result may also suggest that the dependencies among the observed variables that were due to latent confounding may have been weak compared to the dependencies due to the causal relationships among the observed variables. For the non-linear data condition, the only discrete (and therefore non-linear) method DCG achieves the best results.

Without further adjustments GH and DCG cannot be scaled to larger sample sizes or a large number of variables (n). The super-exponential growth of the number of DAGs currently limits the GH approach to not more than 30-50 variables. Additionally, the calculation of local scores can be time consuming. On the other hand, DCG requires a numerical optimization over $n+4n(n-1)$ parameters, which is also infeasible for large n.

In its most basic form (i.e., Algorithm 1), the LLC algorithm only requires the straightforward estimation of the covariance matrices and a calculation of a pseudo-inverse for n matrices with a dimensionality of $(n-1)\times(n-1)$ each. Such a procedure, as used in our simulations, can thus scale to a relatively high (e.g., $n = 100$) number of variables. However, as we see in the next section, it may be useful to add regularization to the basic procedure, and one may have to resort to resampling approaches to obtain estimates of the errors in the coefficients, needed to infer which edges are present and which are absent. Such adaptations and extensions of the basic method can, of course, add significantly to the complexity of the method, but may also pay off in terms of a higher accuracy on small sample sizes.

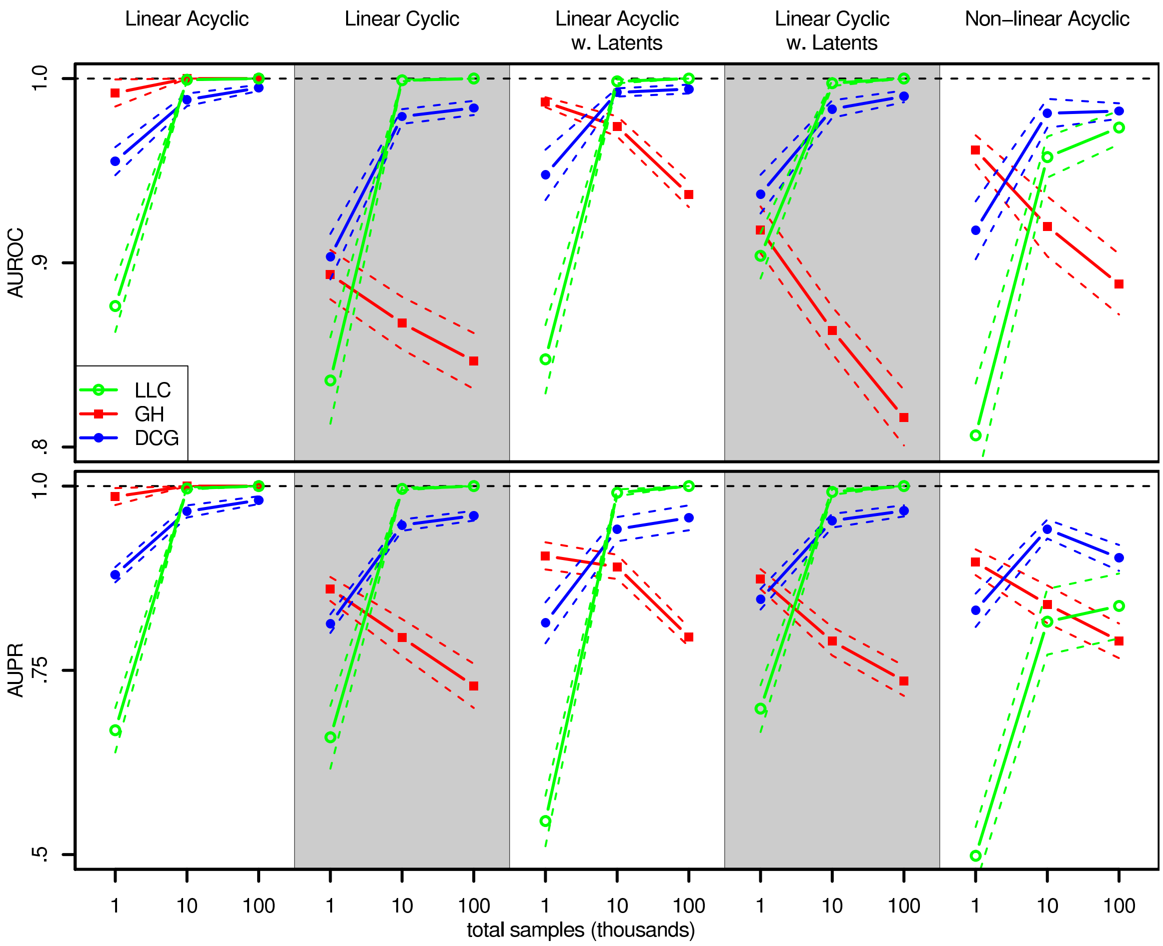


Figure 8: Simulation results: AUROC (top) and AUPR (bottom) values for the LLC-, GH- and DCG-algorithms in the five model conditions (columns, see main text for details) for a total sample size of 1,000-100,000 (x-axis) evenly divided over a passive observation and 10 single intervention experiments. Each point on the solid lines is an average over 20 models with 10 observed variables each, the dashed lines indicate the standard deviation of this average. The light gray shading in this and subsequent figures is used solely for visual distinction.

6. Case Study: DREAM Challenge Data

DREAM (Dialogue for Reverse Engineering Assessments and Methods) is a yearly held challenge for the fair evaluation of strengths and weaknesses of cellular network inference procedures. In this section, we describe how we applied an adapted version of the LLC-method to the *in silico* network challenges of DREAM 3 and DREAM 4, conducted in 2008 and 2009, respectively. The network sizes of the 25 individual models, divided into 5 sub-challenges, ranged from 10 to 100 nodes.

The participants were asked to learn the directed graph structure of a gene regulatory network in different types of cells, from experimental data. Data was *in silico*, or simulated, in order to

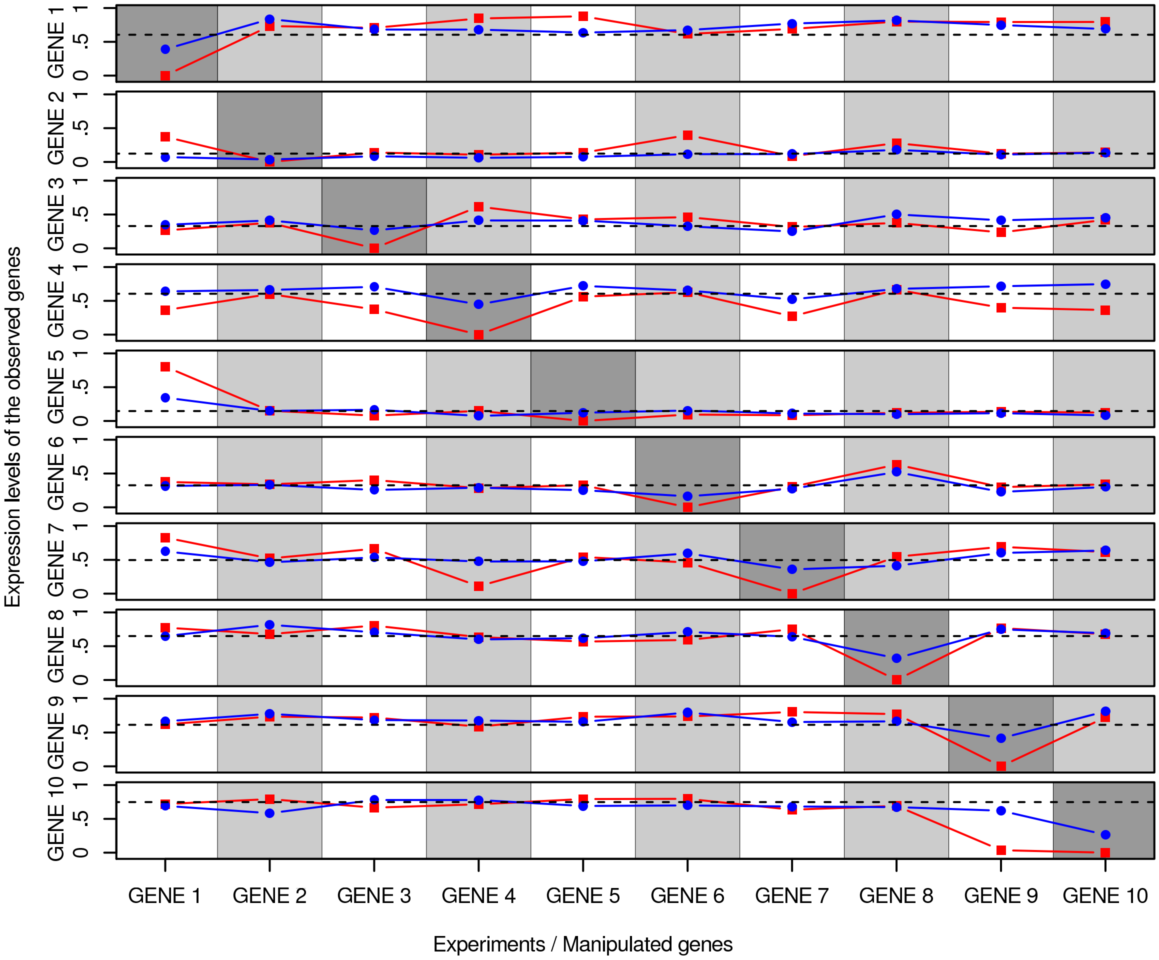


Figure 9: An example of the data provided for one of the 10 variable DREAM network inference challenge. Each row shows the steady state expression levels for each of the 10 genes when the gene indicated on the columns is knocked down (●) or knocked out (■). For each gene, the dashed line indicates the passively observed value. The dark gray shading highlights the diagonal elements, marking the measured levels when intervening on the respective gene. From the 10th row we see that the expression level of the 10th gene responds strongly only to the manipulation of the 9th gene or the 10th gene itself.

have access to the ground truth network structures. The data generating models were designed to be biologically plausible (Marbach et al., 2009) in order to achieve a realistic performance assessment of the network learning algorithms. The networks were based on modules extracted from known biological interaction networks, preserving functional and structural properties of the original networks. Data was then generated simulating a biologically plausible dynamical process and adding noise (Prill et al., 2010).

The data provided to the participants included two measures of the steady states of gene expression levels (the levels converge to these values over time) as mRNA concentrations, in several different conditions. One data set is visualized in Figure 9. Given that only two data vectors were provided for each condition, GH and DCG, tested in Section 5, are not directly applicable. The challenges also provided several time series of how the modeled cell recovers from a perturbation back to its equilibrium state. We do not include the time series data in our analysis, since LLC (or the other procedures we considered) cannot straightforwardly exploit this data. Each team was supposed to output a confidence measure or a score for their belief in the existence of each possible edge in the model. The performance of the learning algorithms was compared using AUROC and AUPR scores for a single data set (Stolovitzky et al., 2009), in the same manner as explained in Section 5. Finally, in each sub-challenge of 5 models, the competing teams were compared using a total score averaging the individual network scores over all 5 networks.

Below, we discuss how we adapted LLC so that we could apply it to these challenges, and compare the results we obtained with the scores achieved by the teams that participated in the original challenge.

6.1 Estimating the Total Effects

When gene i is knocked down or knocked out, we can treat the result in our framework as an outcome of an experiment where variable x_i is intervened on. However, the DREAM data provides only the steady state values of the expression levels, and not the full covariance matrices. We can still find the total effects in the experiments by the following approach. First, we treat the steady state values as the expected values of the variables under the different interventions (or passive observation), rather than as individual samples. Second, the passive observational steady state values are deducted from all the interventional steady state values such that we can assume $E(\mathbf{x}) = \mathbf{0}$ and thus $E(\mathbf{e}) = \mathbf{0}$. Recall that the total effect $t(x_i \rightsquigarrow x_j)$ is just the regression coefficient of x_i when x_j is regressed over the only manipulated variable x_i. Thus, the expected or steady state value of x_j when x_i is manipulated to a value $x_i^{i,ko}$ (knocked out) is simply $t(x_i \rightsquigarrow x_j) \cdot x_i^{i,ko}$. Similar reasoning applies when x_i is manipulated to a value $x_i^{i,kd}$, and so we can estimate $t(x_i \rightsquigarrow x_j)$ by the least squares solution of the equation system:

$$\begin{aligned} t(x_i \rightsquigarrow x_j) \cdot x_i^{i,ko} &= x_j^{i,ko}, \\ t(x_i \rightsquigarrow x_j) \cdot x_i^{i,kd} &= x_j^{i,kd}. \end{aligned}$$

Given that the data set satisfies the pair condition for all ordered pairs, the DREAM experiments fulfill the requirements given in Section 3 for model identifiability and all total effects $t(x_i \rightsquigarrow x_j)$ can be estimated for all pairs (x_i, x_j).

6.2 Network Inference

Given the estimated total effects, we could directly apply the LLC algorithm to estimate the direct effects matrix $\mathbf{B}$. However, we found that to obtain strong results we had to adapt the algorithm in the following way.

First, unlike in the simulations in Section 5, we found that here the absolute value of a coefficient b_{ji} does not provide a good confidence measure for the existence of the edge $x_i \to x_j$, since it does not consider the possibly large variance of the estimate for b_{ji} in any way. As direct re-sampling approaches are not possible with the available data, we created K noisy data sets by adding noise from a normal distribution with variance $\sigma^2 = 0.1$ to each raw data point. We then estimated the total effects as explained above.

Second, to estimate the direct effects $\mathbf{B}$ we solved the LLC equation system in Equation 17 using an L^1-norm penalization with weight $\lambda = 0.1$. An estimate of the direct effects $\mathbf{B}$ (vectorized as $\mathbf{b}$) from the noisy data set is thus calculated by

$$\min_{\mathbf{b}} \|\mathbf{Tb} - \mathbf{t}\|_{L^2}^2 + \lambda \|\mathbf{b}\|_{L^1}.$$

As explained in Section 4 the estimation can be done by n separate minimization problems. Note that the L^1-norm penalization can be thought of as a prior for sparse structures, in a way somewhat similar to the use of a faithfulness assumption.

Finally, we calculate the Z-scores for each link b_{ji} by

$$Z_{ji} = \text{mean}(\{b_{ji}^k\}_{k=1}^K)/\text{std}(\{b_{ji}^k\}_{k=1}^K).$$

The higher the Z-score the more confident we are of the existence of the edge. Using Z-scores allows for a high score for a small coefficient as long as its estimated variance is small as well.

Figure 10 summarizes the results. The first observation is that the DREAM 4 challenges were more competitive than the DREAM 3 challenges as the variation of the results for the 10 best teams is lower. Our overall ranks in the five challenges are 3rd, 9th, 3rd, 2nd and 10th among the approximately 30 teams that participated in the actual challenges. There is no clear difference in evaluation with either score metric. We take these results to be encouraging, especially since—unlike many other candidates—we did not use the available time series data. How to exploit the time series data remains an open question. The noise in the data, not having access to a sufficient number of samples and the possible non-linearity of the causal relations constitute additional sources of errors.

6.3 Prediction Accuracy

In addition to structure discovery, another important aspect of causal modeling is prediction under previously unseen experimental conditions. Thus, DREAM 4 featured a bonus round for predicting the steady state values of the gene expression levels in novel experimental settings. The data were the same as for the structure discovery challenges. For the five 10-variable models, the teams were asked to predict all steady state expression levels in 5 situations where always a pair of genes is knocked out. For the five 100-variable models predictions were requested for 20 double knockout settings each.

The knocked out values of variables x_i and x_j are defined by the data as $x_i^{i,ko}$ and $x_j^{j,ko}$. We can estimate the values of the variables x_u such that $u \neq i, j$ using the interpretation of the experimental effects as regression coefficients:

$$x_u^{i,j,ko} \quad = \quad t(x_i \rightsquigarrow x_u || \{x_i, x_j\}) \cdot x_i^{i,ko} + t(x_j \rightsquigarrow x_u || \{x_i, x_j\}) \cdot x_j^{j,ko}.$$

Since we can estimate $t(x_i \rightsquigarrow x_k || \{x_i\})$ and $t(x_j \rightsquigarrow x_k || \{x_j\})$ as described in the previous section, we can also estimate the quantities $t(x_i \rightsquigarrow x_k || \{x_i, x_j\})$ and $t(x_j \rightsquigarrow x_k || \{x_i, x_j\})$ using Lemma 9 (Union/Intersection). We solve the linear equation group (Equation 35 in Appendix G) for the experimental effects using an L_2 prior with regularization parameter λ. In other words, we assume that the data generating model is a linear cyclic model with latent variables and we predict the steady

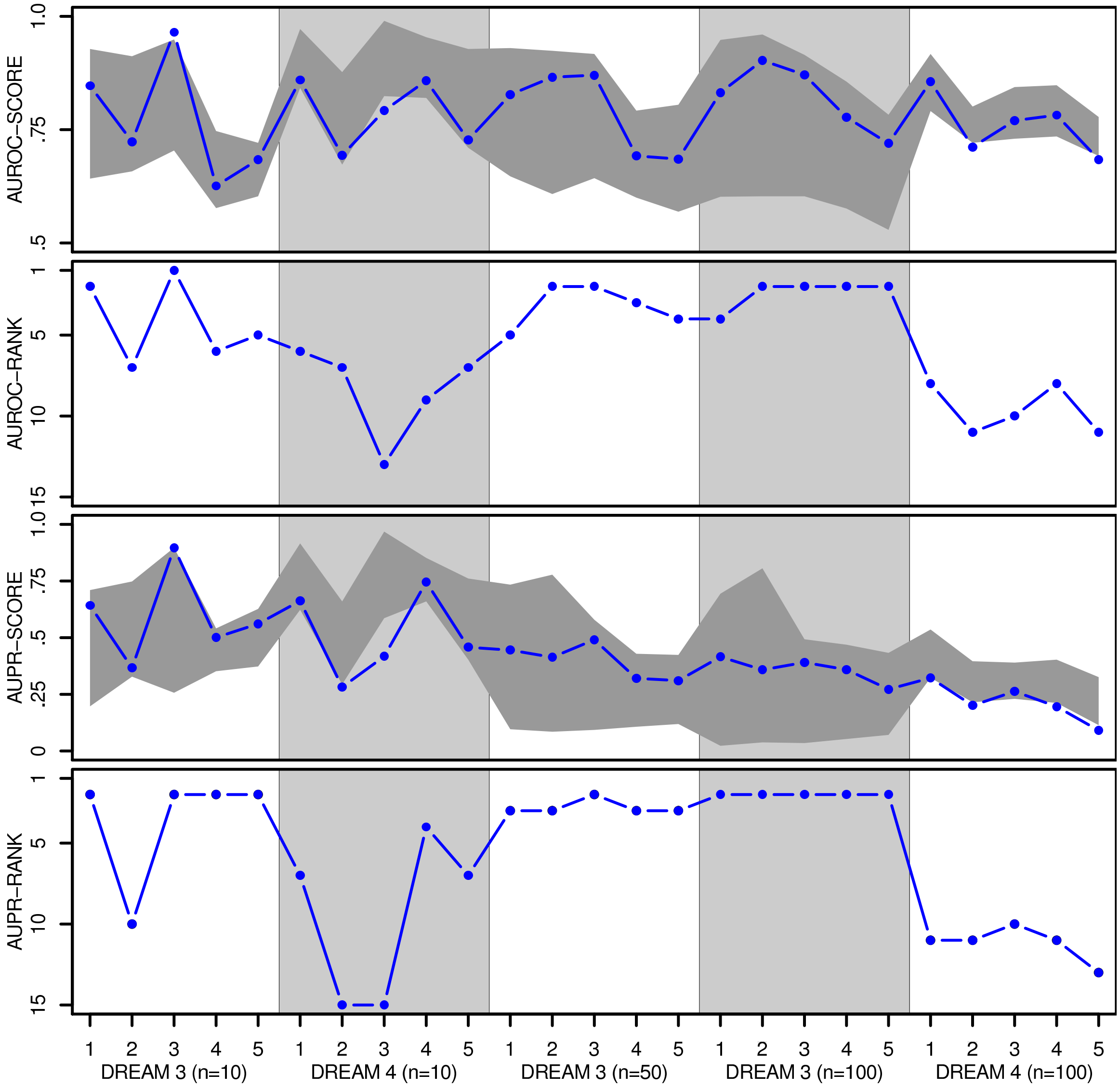


Figure 10: Summary of the results for the DREAM *in silico* network inference challenges: The AUROC- and AUPR-scores (first and third row) and the corresponding rank among the competitors, for each of the DREAM 3 and DREAM 4 challenges. The top of the dark gray area shows the best results among the competing teams for each individual data set, while the bottom always shows the 10th best result. Overall there were about 30 competitors in each of the challenges.

state values of the specific combined (double) knockout experiment on the basis of the relevant single knockout experimental data provided. (The double knockout effects are identified based on the single knockout experimental data by Lemma 9.) In this way, in each individual prediction task we disregard the data that is irrelevant to this specific prediction, and only use the data that is actually

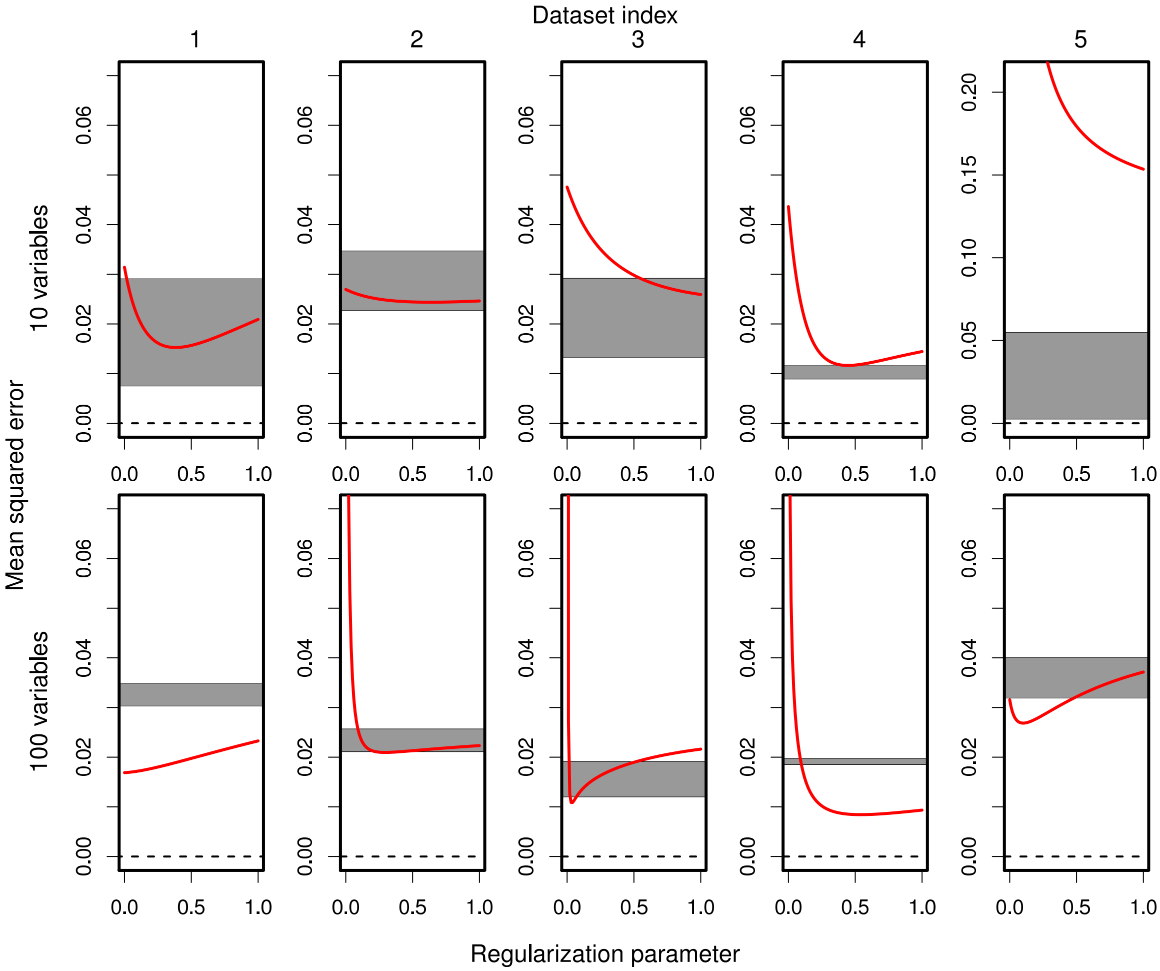


Figure 11: Predictive performance: Mean squared errors of predictions in double intervention experiments on five 10-variable models (top) and 100-variable models (bottom) plotted as a function of the regularization parameter. The red line shows the prediction errors for our procedure. The bottom of the dark gray area shows the best result among the competing teams for each individual data set, while the top always shows the third best result.

relevant. In practice, this means that the predictions are more robust to any slight violations of the modeling assumptions not crucial to the prediction task at hand.

Figure 11 assesses the quality of the predictions. The predictions are compared using the mean squared error from the ground truth, that is, the average sum of squared errors over the variables and over the different predictions requested. For the 10-variable models the results of our procedure are competitive with those of the seven participating teams. For the 100-variable models our procedure achieves in aggregate the best predictions among the five participating teams for a range of the regularization parameter values.

7. Extensions

We have presented and developed the theory in this paper in terms of the standard interpretation of linear non-recursive structural equation models, in which the vector of disturbances $\mathbf{e}$ is held constant throughout the equilibrating process. Following Lauritzen and Richardson (2002) we refer to this most common interpretation of cyclic models as the *deterministic equilibrium* interpretation, since the value of the observed variables $\mathbf{x}$ at equilibrium is a deterministic function of the disturbances $\mathbf{e}$. In this model, as defined in Section 2.1, different observed vectors $\mathbf{x}$ arise solely from different outside influences $\mathbf{e}$, yielding a covariance matrix $\mathbf{C}_\mathbf{x}^k$ for each experiment $\mathcal{E}_k$. In this section we discuss some preliminary ideas for extending the theory to other related linear cyclic models.

In Section 6 we have already seen an application of the method to data in which there is only a single passive-observational data vector $\mathbf{x}_0$ and two experimental data vectors $\mathbf{x}_k^{kd}, \mathbf{x}_k^{ko}$ (corresponding to gene knockdown and knockout experiments, respectively) for each experiment *intervening on a single variable at a time*. In this case, to make the LLC method applicable, one essentially must assume that there is a single (constant) disturbance vector $\mathbf{e}$ that does not change between the different experimental conditions, so that the experimental effects are given by the change in values (from the passive observational to the experimental data) of the non-intervened variables divided by the corresponding change in value of the intervened variable. Under this assumption, the theory presented in this paper is directly applicable to estimate the direct effects among the variables from the experimental effects.

If, however, one wants to apply the full machinery provided in this paper to data of the above kind, but in which each experiment intervenes on *multiple* variables simultaneously, it is not sufficient to obtain just one or two experimental data vectors $\mathbf{x}_k$. Rather, in general multiple data vectors may be needed to be able to disentangle the effects of each of the intervened-upon variables on the non-intervened ones. The details of the required experimental protocols, as well as sufficient and necessary identifiability conditions, are however left for future work.

A different extension considers models in which the observed data vectors arise from an equilibrium reached by a process with *stochastic* dynamics. Specifically, consider a time-series process

$$\mathbf{x}(t) \quad := \quad \mathbf{B}\mathbf{x}(t-1) + \mathbf{e}(t),$$

where $\mathbf{e}(t)$ is sampled anew at each time step t, always from the same distribution with mean $\boldsymbol{\mu}_\mathbf{e} = \mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}_\mathbf{e}$. All the variables in $\mathbf{x}$ are updated simultaneously given their values of the previous time step and the new disturbance term $\mathbf{e}(t)$.[9] Obviously, this system no longer has a deterministic equilibrium, but for an asymptotically stable model $(\mathbf{B}, \boldsymbol{\Sigma}_\mathbf{e})$ the process converges to an equilibrium in which a sample vector $\mathbf{x}(t = \infty)$ is drawn from

$$\begin{aligned} \boldsymbol{\mu}_\mathbf{x} &= \mathbf{0}, \\ \mathbf{C}_\mathbf{x} &= \lim_{t\to\infty} \sum_{i=1}^{t} \mathbf{B}^{t-i} \boldsymbol{\Sigma}_\mathbf{e} (\mathbf{B}^T)^{t-i}. \end{aligned}$$

As in the deterministic model, the observed vector $\mathbf{x}$ drawn at equilibrium is independent of the initial values at the start of the process. Different observed data vectors $\mathbf{x}$ would be obtained by running multiple parallel chains. Interventions could be modeled as setting a given variable to a value

[9] We note that this model differs from Lauritzen and Richardson (2002)'s stochastic equilibrium model, discussed in Sections 6 and 7 of their paper. They consider a sequential update of the variables in a particular order.

drawn from some distribution, and then keeping that variable constant throughout the equilibrating process.

In such a model the covariances between the intervened and non-intervened variables correspond to experimental effects, mirroring the deterministic case. Hence the theory presented in this paper could be used to estimate the direct effects matrix $\mathbf{B}$. Given the direct effects, and given a passive-observational covariance matrix $\mathbf{C_x}$, one could estimate $\mathbf{\Sigma_e}$ using the relation

$$\mathbf{\Sigma_e} \quad = \quad \mathbf{C_x} - \mathbf{B}\mathbf{C_x}\mathbf{B}^T.$$

Note, however, that the expression for the covariance among the non-intervened variables is *not* directly parallel to the deterministic case, so some of the theory presented in this paper would need to be adapted if this particular model were of primary interest.

In all the models discussed so far, we have been assuming that interventions take full control of the intervened variable by making it independent of its normal causes. This representation of an intervention is consistent with interventions in randomized controlled trials or in cases where a variable is "clamped" to a particular value. However, interventions needn't be "surgical" in this sense, but could instead only add an additional influence to the intervened variable without breaking the relations between the intervened variable and its causal parents. Such interventions are sometimes referred to as "soft" interventions. In linear models they are formally equivalent to instrumental variables, which are known to be useful for causal discovery. In our model a soft intervention is simply represented by an added influence that does not affect the coefficient matrix $\mathbf{B}$, nor the disturbance term $\mathbf{e}$. That is, the matrix $\mathbf{U}_k$ is deleted in both instances from Equation 4, but the influence $\mathbf{c}$ is still added. Assuming that the influence of the soft interventions on the intervened variables is known, that is, that $\mathbf{c}$ is measured, and that multiple simultaneous soft interventions are performed independently, it can be shown that one can still determine the experimental effects of the intervened variables. The entire machinery described here thus transfers with only some very minor adjustments. Given that soft interventions can be combined independently of one another, very efficient experimental protocols can be developed. In Eberhardt et al. (2010) we found that even from a statistical perspective, soft interventions appear to require the overall least number of samples for causal discovery.

Lastly, it is worth noting that the LLC-Algorithm presented here uses the measured experimental effects $t(x_i \rightsquigarrow x_u || \mathcal{J})$ to linearly constrain the unknown *direct* effects b_{ji} of $\mathbf{B}$. There may be circumstances in which it might be beneficial to instead use the experimental effects to linearly constrain the *total* effects $t(x_i \rightsquigarrow x_u)$.[10] In fact, such a representation was originally developed in Eberhardt et al. (2010). Given an experiment $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$, the linear constraint of the measured experimental effects on the unknown total effects $t(x_i \rightsquigarrow x_u)$ is then given by

$$t(x_i \rightsquigarrow x_u) \quad = \quad t(x_i \rightsquigarrow x_u || \mathcal{J}_k) + \sum_{x_j \in \mathcal{J}_k \setminus \{x_i\}} t(x_i \rightsquigarrow x_j) t(x_j \rightsquigarrow x_u || \mathcal{J}_k).$$

The constraint has a similar form to the constraint on direct effects in Equation 16, but combines a different set of experimental effects. Such a representation of the constraints in terms of total effects forms the basis for an algorithm analogous to LLC to identify the total effects. Once all the total effects are determined, one can, if needed, easily infer the direct effects (see Eberhardt et al., 2010).

[10] Recall that the total effect corresponds to the experimental effect in the single-intervention experiment where only the cause is subject to intervention, that is, $t(x_i \rightsquigarrow x_u) = t(x_i \rightsquigarrow x_u || \{x_i\})$.

8. Conclusion

We have described a procedure that uses data from a set of experiments to identify linear causal models that may contain cycles and latent variables. While assuming linearity is a significant restriction, we are not aware of any other procedure that works with assumptions that are as weak in all other regards. Given this model space, we have shown how important the satisfaction of the pair condition and the covariance condition is for identifiability. Additionally, we have noted that when the identifiability conditions are not satisfied, the underdetermination of the model is generally fairly local.

Despite our analysis in terms of *canonical* models and sets of *canonical* experiments, we have indicated that these are in fact only very weak conditions: Any data from a non-conditional surgical experiment can be turned into data from a corresponding canonical one (if the experiment was not canonical to start with), and almost any linear cyclic model with latent variables can be represented by a canonical model that is completely equivalent with respect to the available data and any novel predictions produced. Thus, our procedure can handle a quite general model family and experimental setup.

We have shown that the LLC algorithm performs quite well in comparison with algorithms designed for solving similar inference problems. Moreover, within the DREAM challenges, we have a good comparison of how our algorithm (suitably adapted to the problem) performs for realistic data. It is competitive across all challenges despite the linearity assumption.

In Section 7 we have suggested how our model and search procedure can be generalized to models with stochastic dynamics; in Eberhardt et al. (2010) we also considered experiments with so-called “soft” interventions. An open question remains: What are the minimal conditions a model must satisfy such that a search procedure based on experiments that satisfy the pair condition for all ordered pairs of variables is sufficient for model identifiability? In Hyttinen et al. (2011) we showed that this condition is necessary and sufficient for identifiability in discrete acyclic models with a noisy-or parametrization. It is not known to what extent the condition generalizes to other model families.

Acknowledgments

The authors would like to thank the anonymous reviewers for extremely detailed and helpful comments. A.H. & P.O.H. were supported by the Academy of Finland and by University of Helsinki Research Funds. F.E. was supported by a grant from the James S. McDonnell Foundation on ‘Experimental Planning and the Unification of Causal Knowledge’.

Appendix A. Centering the Data

Here we show how to center the data, so that it can be modeled with a linear cyclic model with latent variables that assumes a zero mean for the disturbances. We also consider how to translate the predictions of the model to predictions for the actual data generating process. Throughout, we assume that in each experiment we observe the mean and covariance matrix in the infinite sample limit.

Let the true data generating model be a linear cyclic model with latent variables $(\mathbf{B}, \mathbf{\Sigma}_\mathbf{e}, \boldsymbol{\mu}_\mathbf{e})$ where $\boldsymbol{\mu}_\mathbf{e} \neq \mathbf{0}$. Say, we have observed passive observational data with mean $\boldsymbol{\mu}_\mathbf{x}^0$. In an arbitrary experiment $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ the data generating model produces data with the following mean and covariance matrix:

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{x}}^k &= (\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}(\mathbf{U}_k\boldsymbol{\mu}_{\mathbf{e}} + \mathbf{J}_k\boldsymbol{\mu}_{\mathbf{c}}^k), \\ \mathbf{C}_{\mathbf{x}}^k &= (\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}(\boldsymbol{\Sigma}_{\mathbf{c}}^k + \mathbf{U}_k\boldsymbol{\Sigma}_{\mathbf{e}}\mathbf{U}_k)(\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-T}. \end{aligned} \tag{26}$$

If we first center all data vectors by

$$\bar{\mathbf{x}} = \mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}^0, \tag{27}$$

then the centered data has mean $\bar{\boldsymbol{\mu}}_{\mathbf{x}}^k = \boldsymbol{\mu}_{\mathbf{x}}^k - \boldsymbol{\mu}_{\mathbf{x}}^0$ and unaltered covariance matrix $\bar{\mathbf{C}}_{\mathbf{x}}^k = \mathbf{C}_{\mathbf{x}}^k$. The centering of Equation 27 implies that instead of randomizing the intervened variables in $\mathcal{J}_k$ with mean $(\boldsymbol{\mu}_{\mathbf{c}}^k)_{\mathcal{J}_k}$ and covariance $(\boldsymbol{\Sigma}_{\mathbf{c}}^k)_{\mathcal{J}_k\mathcal{J}_k}$, the centered variables are considered to be randomized with mean $(\bar{\boldsymbol{\mu}}_{\mathbf{c}}^k)_{\mathcal{J}_k} = (\boldsymbol{\mu}_{\mathbf{c}}^k - \boldsymbol{\mu}_{\mathbf{x}}^0)_{\mathcal{J}_k}$ and covariance $(\bar{\boldsymbol{\Sigma}}_{\mathbf{c}}^k)_{\mathcal{J}_k\mathcal{J}_k} = (\boldsymbol{\Sigma}_{\mathbf{c}}^k)_{\mathcal{J}_k\mathcal{J}_k}$. The subsequent equations show that the corresponding model with zero mean disturbance $(\tilde{\mathbf{B}}, \tilde{\boldsymbol{\Sigma}}_{\mathbf{e}}, \tilde{\boldsymbol{\mu}}_{\mathbf{e}})$, where $\tilde{\mathbf{B}} = \mathbf{B}$, $\tilde{\boldsymbol{\Sigma}}_{\mathbf{e}} = \boldsymbol{\Sigma}_{\mathbf{e}}$ and $\tilde{\boldsymbol{\mu}}_{\mathbf{e}} = \mathbf{0}_n$, generates the centered data when the intervened variables are randomized with mean $(\bar{\boldsymbol{\mu}}_{\mathbf{c}}^k)_{\mathcal{J}_k}$ and covariance $\bar{\mathbf{C}}_{\mathbf{x}}^k$:

$$\begin{aligned} \tilde{\mathbf{C}}_{\mathbf{x}}^k &= (\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})^{-1}(\bar{\boldsymbol{\Sigma}}_{\mathbf{c}}^k + \mathbf{U}_k\tilde{\boldsymbol{\Sigma}}_{\mathbf{e}}\mathbf{U}_k)(\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})^{-T} \\ &= (\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}(\boldsymbol{\Sigma}_{\mathbf{c}}^k + \mathbf{U}_k\boldsymbol{\Sigma}_{\mathbf{e}}\mathbf{U}_k)(\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-T} = \mathbf{C}_{\mathbf{x}}^k = \bar{\mathbf{C}}_{\mathbf{x}}^k, \\ \tilde{\boldsymbol{\mu}}_{\mathbf{x}}^k &= (\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})^{-1}(\mathbf{U}_k\tilde{\boldsymbol{\mu}}_{\mathbf{e}} + \mathbf{J}_k\bar{\boldsymbol{\mu}}_{\mathbf{c}}^k) \\ &= (\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k(\boldsymbol{\mu}_{\mathbf{c}}^k - \boldsymbol{\mu}_{\mathbf{x}}^0) \qquad || \boldsymbol{\mu}_{\mathbf{x}}^0 = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\mu}_{\mathbf{e}} \\ &= (\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k\boldsymbol{\mu}_{\mathbf{c}}^k - (\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\mu}_{\mathbf{e}} + (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\mu}_{\mathbf{e}} - \boldsymbol{\mu}_{\mathbf{x}}^0 \\ &= (\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k\boldsymbol{\mu}_{\mathbf{c}}^k + (\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}(-\mathbf{J}_k + \mathbf{I} - \mathbf{U}_k\mathbf{B})(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\mu}_{\mathbf{e}} - \boldsymbol{\mu}_{\mathbf{x}}^0 \\ &= (\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k\boldsymbol{\mu}_{\mathbf{c}}^k + (\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}(-(\mathbf{I} - \mathbf{U}_k) + \mathbf{I} - \mathbf{U}_k\mathbf{B})(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\mu}_{\mathbf{e}} - \boldsymbol{\mu}_{\mathbf{x}}^0 \\ &= (\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}(\mathbf{U}_k\boldsymbol{\mu}_{\mathbf{e}} + \mathbf{J}_k\boldsymbol{\mu}_{\mathbf{c}}^k) - \boldsymbol{\mu}_{\mathbf{x}}^0 = \boldsymbol{\mu}_{\mathbf{x}}^k - \boldsymbol{\mu}_{\mathbf{x}}^0 = \bar{\boldsymbol{\mu}}_{\mathbf{x}}^k. \end{aligned}$$

Thus, the centering is innocuous with regard to $\mathbf{B}$ and $\boldsymbol{\Sigma}_{\mathbf{e}}$. The identities show also how to translate the predictions of the zero-mean disturbance model in some novel experiment $\mathcal{E}_k$ to the predictions of the actual data generating model: $\boldsymbol{\mu}_{\mathbf{x}}^k = \boldsymbol{\mu}_{\mathbf{x}}^0 + \tilde{\boldsymbol{\mu}}_{\mathbf{x}}^k$ and $\mathbf{C}_{\mathbf{x}}^k = \tilde{\mathbf{C}}_{\mathbf{x}}^k$.

In the unlikely case that passive observational data is not available, we can simply center the data vectors $\mathbf{x}$ observed in experiment $\mathcal{E}_k$:

$$\bar{\mathbf{x}} := \mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}^k.$$

This essentially corresponds to just ignoring the observed mean in each experiment. The theory in the paper can be used to estimate the direct effects matrix $\mathbf{B}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{e}}$, as the data covariance matrices are independent of the mean of the disturbances. This is sufficient for structure discovery, but if we want to achieve consistent predictions of the observed mean in novel experimental setups, $\boldsymbol{\mu}_{\mathbf{e}}$ also needs to be estimated. In experiment $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$, the data has mean $\boldsymbol{\mu}_{\mathbf{x}}^k$, so $(\boldsymbol{\mu}_{\mathbf{e}})_{\mathcal{U}_k}$ can be estimated by $((\mathbf{I} - \mathbf{U}_k\mathbf{B})\boldsymbol{\mu}_{\mathbf{x}}^k)_{\mathcal{U}_k}$, since

$$\begin{aligned} ((\mathbf{I} - \mathbf{U}_k\mathbf{B})\boldsymbol{\mu}_{\mathbf{x}}^k)_{\mathcal{U}_k} &= ((\mathbf{I} - \mathbf{U}_k\mathbf{B})(\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}(\mathbf{U}_k\boldsymbol{\mu}_{\mathbf{e}} + \mathbf{J}_k\boldsymbol{\mu}_{\mathbf{c}}^k))_{\mathcal{U}_k} \\ &= (\mathbf{U}_k\boldsymbol{\mu}_{\mathbf{e}} + \mathbf{J}_k\boldsymbol{\mu}_{\mathbf{c}}^k)_{\mathcal{U}_k} = (\boldsymbol{\mu}_{\mathbf{e}})_{\mathcal{U}_k}. \end{aligned}$$

Thus, if each variable x_i is observed unmanipulated in some experiment *and* $\mathbf{B}$ is identified, then the whole vector $\boldsymbol{\mu}_{\mathbf{e}}$ can be estimated. The predicted mean $\boldsymbol{\mu}_{\mathbf{x}}^k$ for an arbitrary novel experiment $\mathcal{E}_k$ can then be obtained using Equation 26. See Appendices B and J for additional discussion on predicting means.

Appendix B. Proof of Lemma 5 (Correlated Experiment)

In a correlated experiment $\mathcal{E}_k$, where $\mathbf{c}$ is randomized with mean $\boldsymbol{\mu}_{\mathbf{c}}^k$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{c}}^k$ such that $(\boldsymbol{\Sigma}_{\mathbf{c}}^k)_{\mathcal{J}_k\mathcal{J}_k}$ is symmetric positive-definite, the model $(\mathbf{B},\boldsymbol{\Sigma}_{\mathbf{e}})$ produces the following observations:

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}_{\mathbf{x}}^k &= (\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}\boldsymbol{\mu}_{\mathbf{c}}^k,\\
\tilde{\mathbf{C}}_{\mathbf{x}}^k &= (\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}(\boldsymbol{\Sigma}_{\mathbf{c}}^k+\mathbf{U}_k\boldsymbol{\Sigma}_{\mathbf{e}}\mathbf{U}_k)(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-T}\\
&= \begin{bmatrix} (\boldsymbol{\Sigma}_{\mathbf{c}}^k)_{\mathcal{J}_k\mathcal{J}_k} & (\boldsymbol{\Sigma}_{\mathbf{c}}^k)_{\mathcal{J}_k\mathcal{J}_k}\mathbf{B}^T_{\mathcal{U}_k\mathcal{J}_k}(\mathbf{I}-\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k})^{-T} \\ (\mathbf{I}-\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k})^{-1}\mathbf{B}_{\mathcal{U}_k\mathcal{J}_k}(\boldsymbol{\Sigma}_{\mathbf{c}}^k)_{\mathcal{J}_k\mathcal{J}_k} & * \end{bmatrix},\\
* &= (\mathbf{I}-\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k})^{-1}((\boldsymbol{\Sigma}_{\mathbf{e}})_{\mathcal{U}_k,\mathcal{U}_k}+\mathbf{B}_{\mathcal{U}_k\mathcal{J}_k}(\boldsymbol{\Sigma}_{\mathbf{c}}^k)_{\mathcal{J}_k\mathcal{J}_k}\mathbf{B}^T_{\mathcal{U}_k\mathcal{J}_k})(\mathbf{I}-\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k})^{-T}.
\end{aligned}$$

Then, matrix $\tilde{\mathbf{T}}_{\mathbf{x}}^k$, defined in the lemma in terms of the observed covariance matrix $\tilde{\mathbf{C}}_{\mathbf{x}}^k$, can be expressed solely in terms of the model parameters $\mathbf{B}$:

$$\begin{aligned}
\tilde{\mathbf{T}}_{\mathbf{x}}^k &= (\tilde{\mathbf{C}}_{\mathbf{x}}^k)_{\mathcal{V}\mathcal{J}_k}((\tilde{\mathbf{C}}_{\mathbf{x}}^k)_{\mathcal{J}_k\mathcal{J}_k})^{-1} = \begin{bmatrix} (\boldsymbol{\Sigma}_{\mathbf{c}}^k)_{\mathcal{J}_k\mathcal{J}_k} \\ (\mathbf{I}-\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k})^{-1}\mathbf{B}_{\mathcal{U}_k\mathcal{J}_k}(\boldsymbol{\Sigma}_{\mathbf{c}}^k)_{\mathcal{J}_k\mathcal{J}_k} \end{bmatrix}\begin{bmatrix} (\boldsymbol{\Sigma}_{\mathbf{c}}^k)_{\mathcal{J}_k\mathcal{J}_k} \end{bmatrix}^{-1}\\
&= \begin{bmatrix} \mathbf{I} \\ (\mathbf{I}-\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k})^{-1}\mathbf{B}_{\mathcal{U}_k\mathcal{J}_k} \end{bmatrix} = ((\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1})_{\mathcal{V}\mathcal{J}_k},
\end{aligned}$$

where matrix $(\tilde{\mathbf{C}}_{\mathbf{x}}^k)_{\mathcal{J}_k\mathcal{J}_k} = (\boldsymbol{\Sigma}_{\mathbf{c}}^k)_{\mathcal{J}_k\mathcal{J}_k}$ is invertible, since it is a positive-definite matrix. The following identities apply:

$$\begin{aligned}
\tilde{\mathbf{T}}_{\mathbf{x}}^k(\tilde{\mathbf{T}}_{\mathbf{x}}^k)^T &= ((\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1})_{\mathcal{V}\mathcal{J}_k}(((\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1})_{\mathcal{V}\mathcal{J}_k})^T\\
&= (\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-T},\\
\tilde{\mathbf{T}}_{\mathbf{x}}^k(\tilde{\mathbf{C}}_{\mathbf{x}}^k)_{\mathcal{J}_k\mathcal{J}_k}(\tilde{\mathbf{T}}_{\mathbf{x}}^k)^T &= ((\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1})_{\mathcal{V}\mathcal{J}_k}(\boldsymbol{\Sigma}_{\mathbf{c}}^k)_{\mathcal{J}_k\mathcal{J}_k}(((\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1})_{\mathcal{V}\mathcal{J}_k})^T\\
&= (\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}\boldsymbol{\Sigma}_{\mathbf{c}}^k(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-T}.
\end{aligned}$$

Now from Equations 5 and 6 we can calculate the statistics of the experiment if the intervened variables had been randomized with zero mean and unit variance (appearing in Equations 7 and 8):

$$\begin{aligned}
\boldsymbol{\mu}_{\mathbf{x}}^k &= \mathbf{0},\\
\mathbf{C}_{\mathbf{x}}^k &= \tilde{\mathbf{C}}_{\mathbf{x}}^k - \tilde{\mathbf{T}}_{\mathbf{x}}^k(\tilde{\mathbf{C}}_{\mathbf{x}}^k)_{\mathcal{J}_k\mathcal{J}_k}(\tilde{\mathbf{T}}_{\mathbf{x}}^k)^T + \tilde{\mathbf{T}}_{\mathbf{x}}^k(\tilde{\mathbf{T}}_{\mathbf{x}}^k)^T\\
&= (\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}(\boldsymbol{\Sigma}_{\mathbf{c}}^k+\mathbf{U}_k\boldsymbol{\Sigma}_{\mathbf{e}}\mathbf{U}_k-\boldsymbol{\Sigma}_{\mathbf{c}}^k+\mathbf{J}_k)(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-T}\\
&= (\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}(\mathbf{U}_k\boldsymbol{\Sigma}_{\mathbf{e}}\mathbf{U}_k+\mathbf{J}_k)(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-T}.
\end{aligned}$$

Notice that the formulas in the lemma can also be used to transform the predictions $\mathbf{0}$ and $\mathbf{C}_{\mathbf{x}}^k$ in a canonical experiment to predictions $\tilde{\boldsymbol{\mu}}_{\mathbf{x}}^k$ and $\tilde{\mathbf{C}}_{\mathbf{x}}^k$ in a non-canonical experiment, where $\mathbf{c}$ is randomized with mean $\boldsymbol{\mu}_{\mathbf{c}}^k$ and covariance $\boldsymbol{\Sigma}_{\mathbf{c}}^k$:

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}_{\mathbf{x}}^k &= \tilde{\mathbf{T}}_{\mathbf{x}}^k\boldsymbol{\mu}_{\mathbf{c}}^k,\\
\tilde{\mathbf{C}}_{\mathbf{x}}^k &= \mathbf{C}_{\mathbf{x}}^k + \tilde{\mathbf{T}}_{\mathbf{x}}^k(\boldsymbol{\Sigma}_{\mathbf{c}}^k)_{\mathcal{J}_k\mathcal{J}_k}(\tilde{\mathbf{T}}_{\mathbf{x}}^k)^T - \tilde{\mathbf{T}}_{\mathbf{x}}^k(\tilde{\mathbf{T}}_{\mathbf{x}}^k)^T,
\end{aligned}$$

where $\tilde{\mathbf{T}}^k_{\mathbf{x}} = (\mathbf{C}^k_{\mathbf{x}})_{\mathcal{V}\mathcal{J}_k}$.

Appendix C. Derivation of the Trek Rule for Asymptotically Stable Models

From the definition of asymptotic stability it follows that the eigenvalues of $\mathbf{U}_k\mathbf{B}$ are all less than one in absolute value. As the eigenvalues of matrix $\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k}$ are equal to those of $\mathbf{U}_k\mathbf{B}$, matrix $(\mathbf{I}-\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k})^{-1}$ can be written as the following geometric series:

$$(\mathbf{I}-\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k})^{-1} = \mathbf{I}+\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k}+\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k}\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k}+\cdots.$$

Now, the experimental effect $t(x_i \rightsquigarrow x_u||\mathcal{J}_k)$ can be expressed as the sum-product implied by the trek rules:

$$\begin{aligned}
t(x_i \rightsquigarrow x_u||\mathcal{J}_k) &= (\mathbf{T}^k_{\mathbf{x}})_{\{x_u\}\{x_i\}}\\
&= ((\mathbf{I}-\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k})^{-1}\mathbf{B}_{\mathcal{U}_k\mathcal{J}_k})_{\{x_u\}\{x_i\}}\\
&= ((\mathbf{I}+\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k}+\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k}\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k}+\cdots)\mathbf{B}_{\mathcal{U}_k\mathcal{J}_k})_{\{x_u\}\{x_i\}}\\
&= b_{ui}+\sum_{j\in\mathcal{U}_k} b_{uj}b_{ji}+\sum_{j\in\mathcal{U}_k}\sum_{l\in\mathcal{U}_k} b_{uj}b_{jl}b_{li}+\cdots\\
&= \sum_{p\in\mathcal{P}(x_i \rightsquigarrow x_u||\mathcal{J}_k)}\;\prod_{(x_l\to x_m)\in p} b_{ml}.
\end{aligned}$$

Appendix D. Proof of Lemma 7 (Marginalization)

In the following, note that the experiment of the marginalized model $\tilde{\mathcal{E}}_k = (\tilde{\mathcal{J}}_k, \tilde{\mathcal{U}}_k)$ and the corresponding experiment of the full model $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ satisfy $\mathcal{J}_k = \tilde{\mathcal{J}}_k$ and $\mathcal{U}_k = \tilde{\mathcal{U}}_k \cup \mathcal{M}$. Without loss of generality the variables are labeled such that $x_1,\ldots,x_i \in \tilde{\mathcal{J}}_k$, $x_{i+1},\ldots,x_j \in \tilde{\mathcal{U}}_k$ and $x_{j+1},\ldots,x_n \in \mathcal{M}$ to allow for easy block matrix manipulation.

D.1 Weak Stability

We show that if the full model $(\mathbf{B},\mathbf{\Sigma_e})$ is weakly stable then the marginalized model $(\tilde{\mathbf{B}},\tilde{\mathbf{\Sigma}}_{\mathbf{e}})$ is also weakly stable. Make the counter-assumption that $(\tilde{\mathbf{B}},\tilde{\mathbf{\Sigma}}_{\mathbf{e}})$ is weakly unstable, thus there exists an experiment $\tilde{\mathcal{E}}_k$ such that $(\mathbf{I}-\tilde{\mathbf{U}}_k\tilde{\mathbf{B}})$ is singular, or equivalently matrix $\tilde{\mathbf{U}}_k\tilde{\mathbf{B}}$ has a unit eigenvalue: $\exists\tilde{\mathbf{v}} \neq \mathbf{0}$ such that $\tilde{\mathbf{U}}_k\tilde{\mathbf{B}}\tilde{\mathbf{v}} = \tilde{\mathbf{v}}$. The following shows that then $\mathbf{U}_k\mathbf{B}$ also has a unit eigenvalue corresponding to the eigenvector $\mathbf{v}$ defined below:[11]

[11] Invertibility of $(\mathbf{I}-\mathbf{B}_{\mathcal{M}\mathcal{M}})$ follows from the weak stability of $(\mathbf{B},\mathbf{\Sigma_e})$ in experiment $(\tilde{\mathcal{V}},\mathcal{M})$.

$$
\begin{aligned}
\mathbf{U}_k\mathbf{B}\mathbf{v} &= \begin{bmatrix} \tilde{\mathbf{U}}_k & \\ & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{B}_{\tilde{\mathcal{V}}\tilde{\mathcal{V}}} & \mathbf{B}_{\tilde{\mathcal{V}}\mathcal{M}} \\ \mathbf{B}_{\mathcal{M}\tilde{\mathcal{V}}} & \mathbf{B}_{\mathcal{M}\mathcal{M}} \end{bmatrix} \mathbf{v} \qquad || \mathbf{v} = \begin{bmatrix} \tilde{\mathbf{v}} \\ (\mathbf{I}-\mathbf{B}_{\mathcal{M}\mathcal{M}})^{-1}\mathbf{B}_{\mathcal{M}\tilde{\mathcal{V}}}\tilde{\mathbf{v}} \end{bmatrix} \\
&= \begin{bmatrix} \tilde{\mathbf{U}}_k\mathbf{B}_{\tilde{\mathcal{V}}\tilde{\mathcal{V}}} & \tilde{\mathbf{U}}_k\mathbf{B}_{\tilde{\mathcal{V}}\mathcal{M}} \\ \mathbf{I}\cdot\mathbf{B}_{\mathcal{M}\tilde{\mathcal{V}}} & \mathbf{I}\cdot\mathbf{B}_{\mathcal{M}\mathcal{M}} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{v}} \\ (\mathbf{I}-\mathbf{B}_{\mathcal{M}\mathcal{M}})^{-1}\mathbf{B}_{\mathcal{M}\tilde{\mathcal{V}}}\tilde{\mathbf{v}} \end{bmatrix} \\
&= \begin{bmatrix} \tilde{\mathbf{U}}_k\mathbf{B}_{\tilde{\mathcal{V}}\tilde{\mathcal{V}}}\tilde{\mathbf{v}} + \tilde{\mathbf{U}}_k\mathbf{B}_{\tilde{\mathcal{V}}\mathcal{M}}(\mathbf{I}-\mathbf{B}_{\mathcal{M}\mathcal{M}})^{-1}\mathbf{B}_{\mathcal{M}\tilde{\mathcal{V}}}\tilde{\mathbf{v}} \\ \mathbf{B}_{\mathcal{M}\tilde{\mathcal{V}}}\tilde{\mathbf{v}} + \mathbf{B}_{\mathcal{M}\mathcal{M}}(\mathbf{I}-\mathbf{B}_{\mathcal{M}\mathcal{M}})^{-1}\mathbf{B}_{\mathcal{M}\tilde{\mathcal{V}}}\tilde{\mathbf{v}} \end{bmatrix} \\
&= \begin{bmatrix} \tilde{\mathbf{U}}_k(\mathbf{B}_{\tilde{\mathcal{V}}\tilde{\mathcal{V}}} + \mathbf{B}_{\tilde{\mathcal{V}}\mathcal{M}}(\mathbf{I}-\mathbf{B}_{\mathcal{M}\mathcal{M}})^{-1}\mathbf{B}_{\mathcal{M}\tilde{\mathcal{V}}})\tilde{\mathbf{v}} \\ (\mathbf{I}-\mathbf{B}_{\mathcal{M}\mathcal{M}})(\mathbf{I}-\mathbf{B}_{\mathcal{M}\mathcal{M}})^{-1}\mathbf{B}_{\mathcal{M}\tilde{\mathcal{V}}}\tilde{\mathbf{v}} + \mathbf{B}_{\mathcal{M}\mathcal{M}}(\mathbf{I}-\mathbf{B}_{\mathcal{M}\mathcal{M}})^{-1}\mathbf{B}_{\mathcal{M}\tilde{\mathcal{V}}}\tilde{\mathbf{v}} \end{bmatrix} \quad ||\text{Def. of } \tilde{\mathbf{B}} \\
&= \begin{bmatrix} \tilde{\mathbf{U}}_k\tilde{\mathbf{B}}\tilde{\mathbf{v}} \\ (\mathbf{I}-\mathbf{B}_{\mathcal{M}\mathcal{M}}+\mathbf{B}_{\mathcal{M}\mathcal{M}})(\mathbf{I}-\mathbf{B}_{\mathcal{M}\mathcal{M}})^{-1}\mathbf{B}_{\mathcal{M}\tilde{\mathcal{V}}}\tilde{\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{v}} \\ (\mathbf{I}-\mathbf{B}_{\mathcal{M}\mathcal{M}})^{-1}\mathbf{B}_{\mathcal{M}\tilde{\mathcal{V}}}\tilde{\mathbf{v}} \end{bmatrix} = \mathbf{v}.
\end{aligned}
$$

Thus, $(\mathbf{I}-\mathbf{U}_k\mathbf{B})$ is singular and the full model $(\mathbf{B},\mathbf{\Sigma_e})$ is not weakly stable. Because this is contradictory to the assumptions, $(\tilde{\mathbf{B}},\tilde{\mathbf{\Sigma}}_\mathbf{e})$ must be weakly stable.

D.2 Equal Covariance Matrices

We need to show that in experiment $\mathcal{E}_k$ the covariance matrix $(\mathbf{C}^k_\mathbf{x})_{\tilde{\mathcal{V}}\tilde{\mathcal{V}}}$ produced by $(\mathbf{B},\mathbf{\Sigma_e})$ is equal to the covariance matrix $\tilde{\mathbf{C}}^k_\mathbf{x}$ produced by $(\tilde{\mathbf{B}},\tilde{\mathbf{\Sigma}}_\mathbf{e})$. This requires us first to derive the following identities:

$$
\begin{aligned}
(\mathbf{I}-\tilde{\mathbf{B}}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{U}}_k})^{-1} &= (\mathbf{I}-\mathbf{B}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{U}}_k}-\mathbf{B}_{\tilde{\mathcal{U}}_k\mathcal{M}}(\mathbf{I}-\mathbf{B}_{\mathcal{M}\mathcal{M}})^{-1}\mathbf{B}_{\mathcal{M}\tilde{\mathcal{U}}_k})^{-1}, && (28)\\
\tilde{\mathbf{B}}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{J}}_k} &= \mathbf{B}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{J}}_k}+\mathbf{B}_{\tilde{\mathcal{U}}_k\mathcal{M}}(\mathbf{I}-\mathbf{B}_{\mathcal{M}\mathcal{M}})^{-1}\mathbf{B}_{\mathcal{M}\tilde{\mathcal{J}}_k}, && (29)\\
(\mathbf{I}-\tilde{\mathbf{B}}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{U}}_k})^{-1}\tilde{\mathbf{B}}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{J}}_k} &= ((\mathbf{I}-\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k})^{-1}\mathbf{B}_{\mathcal{U}_k\mathcal{J}_k})_{\tilde{\mathcal{U}}_k\tilde{\mathcal{J}}_k}, && (30)\\
((\mathbf{I}-\tilde{\mathbf{U}}_k\tilde{\mathbf{B}})^{-1})_{\tilde{\mathcal{V}}\tilde{\mathcal{J}}_k} &= ((\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1})_{\tilde{\mathcal{V}}\tilde{\mathcal{J}}_k}. && (31)
\end{aligned}
$$

The goal is to derive Equation 31, which means that both models produce the same experimental effects from $x_i \in \tilde{\mathcal{J}}_k$ to $x_u \in \tilde{\mathcal{U}}_k$.

Equations 28 and 29 follow directly from the marginalized model definition in Lemma 7. To show Equation 30, we invert the matrix $(\mathbf{I}-\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k})$ in blocks (the unneeded blocks on rows corresponding to the marginalized variables are replaced with a '$\cdot$'-symbol):

$$
\begin{aligned}
(\mathbf{I}-\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k})^{-1} &= \begin{bmatrix} \mathbf{I}-\mathbf{B}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{U}}_k} & -\mathbf{B}_{\tilde{\mathcal{U}}_k\mathcal{M}} \\ -\mathbf{B}_{\mathcal{M}\tilde{\mathcal{U}}_k} & \mathbf{I}-\mathbf{B}_{\mathcal{M}\mathcal{M}} \end{bmatrix}^{-1} \quad ||\text{block matrix inversion \& Eq. 28} \\
&= \begin{bmatrix} (\mathbf{I}-\tilde{\mathbf{B}}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{U}}_k})^{-1} & (\mathbf{I}-\tilde{\mathbf{B}}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{U}}_k})^{-1}\mathbf{B}_{\tilde{\mathcal{U}}_k\mathcal{M}}(\mathbf{I}-\mathbf{B}_{\mathcal{M}\mathcal{M}})^{-1} \\ \cdot & \cdot \end{bmatrix}.
\end{aligned}
$$

Then, we can verify Equation 30:

$$\begin{aligned}
& ((\mathbf{I}-\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k})^{-1}\mathbf{B}_{\mathcal{U}_k\mathcal{J}_k})_{\tilde{\mathcal{U}}_k\tilde{\mathcal{J}}_k} \\
= & \left(\begin{bmatrix} (\mathbf{I}-\tilde{\mathbf{B}}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{U}}_k})^{-1} & (\mathbf{I}-\tilde{\mathbf{B}}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{U}}_k})^{-1}\mathbf{B}_{\tilde{\mathcal{U}}_k\mathcal{M}}(\mathbf{I}-\mathbf{B}_{\mathcal{M}\mathcal{M}})^{-1} \\ \cdot & \cdot \end{bmatrix}\begin{bmatrix} \mathbf{B}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{J}}_k} \\ \mathbf{B}_{\mathcal{M}\tilde{\mathcal{J}}_k} \end{bmatrix}\right)_{\tilde{\mathcal{U}}_k\tilde{\mathcal{J}}_k} \\
= & \begin{bmatrix} (\mathbf{I}-\tilde{\mathbf{B}}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{U}}_k})^{-1}(\mathbf{B}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{J}}_k}+\mathbf{B}_{\tilde{\mathcal{U}}_k\mathcal{M}}(\mathbf{I}-\mathbf{B}_{\mathcal{M}\mathcal{M}})^{-1}\mathbf{B}_{\mathcal{M}\tilde{\mathcal{J}}_k}) \\ \cdot \end{bmatrix}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{J}}_k} \qquad ||\text{Eq. 29} \\
= & \begin{bmatrix} (\mathbf{I}-\tilde{\mathbf{B}}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{U}}_k})^{-1}\tilde{\mathbf{B}}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{J}}_k} \\ \cdot \end{bmatrix}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{J}}_k} = (\mathbf{I}-\tilde{\mathbf{B}}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{U}}_k})^{-1}\tilde{\mathbf{B}}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{J}}_k}.
\end{aligned}$$

Equation 31 follows quite directly from Equation 30:

$$\begin{aligned}
((\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1})_{\tilde{\mathcal{V}}\tilde{\mathcal{J}}_k} &= (((\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1})_{\mathcal{V}\mathcal{J}_k})_{\tilde{\mathcal{V}}\tilde{\mathcal{J}}_k} = \begin{bmatrix} \mathbf{I} \\ (\mathbf{I}-\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k})^{-1}\mathbf{B}_{\mathcal{U}_k\mathcal{J}_k} \end{bmatrix}_{\tilde{\mathcal{V}}\tilde{\mathcal{J}}_k} \\
&= \begin{bmatrix} \mathbf{I} \\ ((\mathbf{I}-\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k})^{-1}\mathbf{B}_{\mathcal{U}_k\mathcal{J}_k})_{\tilde{\mathcal{U}}_k\tilde{\mathcal{J}}_k} \end{bmatrix} \qquad ||\text{Eq. 30} \\
&= \begin{bmatrix} \mathbf{I} \\ (\mathbf{I}-\tilde{\mathbf{B}}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{U}}_k})^{-1}\tilde{\mathbf{B}}_{\tilde{\mathcal{U}}_k\tilde{\mathcal{J}}_k} \end{bmatrix} = ((\mathbf{I}-\tilde{\mathbf{U}}_k\tilde{\mathbf{B}})^{-1})_{\tilde{\mathcal{V}}\tilde{\mathcal{J}}_k}.
\end{aligned}$$

Next, we use matrix $\tilde{\mathbf{v}} = [\mathbf{I}_{j\times j}\ \mathbf{0}_{j\times(n-j)}]$ to avoid the complicated block matrix notation. Multiplication from the left by $\tilde{\mathbf{v}}$ just selects the rows corresponding to variables in $\tilde{\mathcal{V}}$, multiplication from the right by $\tilde{\mathbf{v}}^T$ selects the columns corresponding to variables in $\tilde{\mathcal{V}}$. We prove the following identities.:

$$\begin{aligned}
(\mathbf{I}-\tilde{\mathbf{U}}_k\tilde{\mathbf{B}})^{-1}\tilde{\mathbf{J}}_k &= \tilde{\mathbf{v}}(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k\tilde{\mathbf{v}}^T, \qquad (32) \\
(\mathbf{I}-\tilde{\mathbf{U}}_k\tilde{\mathbf{B}})^{-1}\tilde{\mathbf{U}}_k(\mathbf{I}-\tilde{\mathbf{B}})\tilde{\mathbf{v}} &= \tilde{\mathbf{v}}(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}\mathbf{U}_k(\mathbf{I}-\mathbf{B}). \qquad (33)
\end{aligned}$$

Equation 32 just restates Equation 31 using matrix $\tilde{\mathbf{v}}$. Equation 33 is verified by the following derivation:

$$\begin{aligned}
& (\mathbf{I}-\tilde{\mathbf{U}}_k\tilde{\mathbf{B}})^{-1}\tilde{\mathbf{U}}_k(\mathbf{I}-\tilde{\mathbf{B}})\tilde{\mathbf{v}}-\tilde{\mathbf{v}}(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}\mathbf{U}_k(\mathbf{I}-\mathbf{B}) \quad ||\mathbf{U}_k=\mathbf{I}-\mathbf{J}_k,\tilde{\mathbf{U}}_k=\mathbf{I}-\tilde{\mathbf{J}}_k \\
= & (\mathbf{I}-\tilde{\mathbf{U}}_k\tilde{\mathbf{B}})^{-1}(\mathbf{I}-\tilde{\mathbf{U}}_k\tilde{\mathbf{B}}-\tilde{\mathbf{J}}_k)\tilde{\mathbf{v}}-\tilde{\mathbf{v}}(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}(\mathbf{I}-\mathbf{U}_k\mathbf{B}-\mathbf{J}_k) \\
= & \tilde{\mathbf{v}}-(\mathbf{I}-\tilde{\mathbf{U}}_k\tilde{\mathbf{B}})^{-1}\tilde{\mathbf{J}}_k\tilde{\mathbf{v}}-\tilde{\mathbf{v}}+\tilde{\mathbf{v}}(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k \ ||\text{Eq. 32} \\
= & -\tilde{\mathbf{v}}(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k\tilde{\mathbf{v}}^T\tilde{\mathbf{v}}+\tilde{\mathbf{v}}(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k \quad ||\mathbf{J}_k\tilde{\mathbf{v}}^T\tilde{\mathbf{v}}=\mathbf{J}_k \\
= & \mathbf{0}.
\end{aligned}$$

Finally, we can show that the covariance matrix $\tilde{\mathbf{C}}^k_{\mathbf{x}}$ of the marginalized model matches the marginalized covariance matrix $(\mathbf{C}^k_{\mathbf{x}})_{\tilde{\mathcal{V}}\tilde{\mathcal{V}}}$ of the original model:

$$\begin{aligned}
\tilde{\mathbf{C}}^k_{\mathbf{x}} &= (\mathbf{I}-\tilde{\mathbf{U}}_k\tilde{\mathbf{B}})^{-1}(\tilde{\mathbf{J}}_k+\tilde{\mathbf{U}}_k\tilde{\mathbf{\Sigma}}_{\mathbf{e}}\tilde{\mathbf{U}}_k)(\mathbf{I}-\tilde{\mathbf{U}}_k\tilde{\mathbf{B}})^{-T} \quad ||\text{definition of } \tilde{\mathbf{\Sigma}}_{\mathbf{e}}\\
&= (\mathbf{I}-\tilde{\mathbf{U}}_k\tilde{\mathbf{B}})^{-1}(\tilde{\mathbf{J}}_k+\tilde{\mathbf{U}}_k(\mathbf{I}-\tilde{\mathbf{B}})\tilde{\mathbf{v}}(\mathbf{I}-\mathbf{B})^{-1}\mathbf{\Sigma}_{\mathbf{e}}(\mathbf{I}-\mathbf{B})^{-T}\tilde{\mathbf{v}}^T(\mathbf{I}-\tilde{\mathbf{B}})^T\tilde{\mathbf{U}}_k)(\mathbf{I}-\tilde{\mathbf{U}}_k\tilde{\mathbf{B}})^{-T}\\
&= (\mathbf{I}-\tilde{\mathbf{U}}_k\tilde{\mathbf{B}})^{-1}\tilde{\mathbf{J}}_k(\mathbf{I}-\tilde{\mathbf{U}}_k\tilde{\mathbf{B}})^{-T}+\\
&\quad (\mathbf{I}-\tilde{\mathbf{U}}_k\tilde{\mathbf{B}})^{-1}\tilde{\mathbf{U}}_k(\mathbf{I}-\tilde{\mathbf{B}})\tilde{\mathbf{v}}(\mathbf{I}-\mathbf{B})^{-1}\mathbf{\Sigma}_{\mathbf{e}}(\mathbf{I}-\mathbf{B})^{-T}\tilde{\mathbf{v}}^T(\mathbf{I}-\tilde{\mathbf{B}})^T\tilde{\mathbf{U}}_k(\mathbf{I}-\tilde{\mathbf{U}}_k\tilde{\mathbf{B}})^{-T} \ ||\text{Eq. 33}\\
&= (\mathbf{I}-\tilde{\mathbf{U}}_k\tilde{\mathbf{B}})^{-1}\tilde{\mathbf{J}}_k\tilde{\mathbf{J}}_k(\mathbf{I}-\tilde{\mathbf{U}}_k\tilde{\mathbf{B}})^{-T}+\\
&\quad \tilde{\mathbf{v}}(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}\mathbf{U}_k(\mathbf{I}-\mathbf{B})(\mathbf{I}-\mathbf{B})^{-1}\mathbf{\Sigma}_{\mathbf{e}}(\mathbf{I}-\mathbf{B})^{-T}(\mathbf{I}-\mathbf{B})^T\mathbf{U}_k(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-T}\tilde{\mathbf{v}}^T \ ||\text{Eq. 32}\\
&= \tilde{\mathbf{v}}(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-T}\tilde{\mathbf{v}}^T+\tilde{\mathbf{v}}(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}\mathbf{U}_k\mathbf{\Sigma}_{\mathbf{e}}\mathbf{U}_k(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-T}\tilde{\mathbf{v}}^T\\
&= \tilde{\mathbf{v}}(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}(\mathbf{J}_k+\mathbf{U}_k\mathbf{\Sigma}_{\mathbf{e}}\mathbf{U}_k)(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-T}\tilde{\mathbf{v}}^T=(\mathbf{C}^k_{\mathbf{x}})_{\tilde{\mathcal{V}}\tilde{\mathcal{V}}}.
\end{aligned}$$

Appendix E. Proof of Lemma 8 (Self Cycles)

Again, we first show weak stability and then confirm that the covariance matrices are equal.

E.1 Weak Stability

First, we show that the model $(\tilde{\mathbf{B}},\tilde{\mathbf{\Sigma}}_{\mathbf{e}})$ without the self-loop is weakly stable, if the model $(\mathbf{B},\mathbf{\Sigma}_{\mathbf{e}})$ with the self-loop is weakly stable. Notice that the weak stability of $(\mathbf{B},\mathbf{\Sigma}_{\mathbf{e}})$ in experiment $(\mathcal{V}\setminus\{x_i\},\{x_i\})$ implies that $b_{ii}\neq 1$. So, assume that $(\mathbf{B},\mathbf{\Sigma}_{\mathbf{e}})$ is weakly stable. Make the counter-assumption that $(\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})$ is not invertible in some experiment $\mathcal{E}_k$, then $\exists\mathbf{v}\neq\mathbf{0}$ such that $\mathbf{U}_k\tilde{\mathbf{B}}\mathbf{v}=\mathbf{v}$. Matrix $\mathbf{B}$ can be written as a function of matrix $\tilde{\mathbf{B}}$ by inverting the definition of $\tilde{\mathbf{B}}$ in the lemma:

$$\mathbf{B} \quad = \quad (\mathbf{I}-b_{ii}\mathbf{U}_i)\tilde{\mathbf{B}}+b_{ii}\mathbf{U}_i.$$

If $x_i\in\mathcal{J}_k$ we have that $\mathbf{U}_k\mathbf{U}_i=\mathbf{0}_{n\times n}$, then

$$\mathbf{U}_k\mathbf{B} \quad = \quad \mathbf{U}_k(\mathbf{I}-b_{ii}\mathbf{U}_i)\tilde{\mathbf{B}}+b_{ii}\mathbf{U}_k\mathbf{U}_i=\mathbf{U}_k\tilde{\mathbf{B}}$$

and $\mathbf{U}_k\mathbf{B}\mathbf{v}=\mathbf{U}_k\tilde{\mathbf{B}}\mathbf{v}=\mathbf{v}$. Alternatively if $x_i\in\mathcal{U}_k$, we have that $\mathbf{U}_k\mathbf{U}_i=\mathbf{U}_i$, then

$$\begin{aligned}
\mathbf{U}_k\mathbf{B}\mathbf{v} &= \mathbf{U}_k(\mathbf{I}-b_{ii}\mathbf{U}_i)\tilde{\mathbf{B}}\mathbf{v}+b_{ii}\mathbf{U}_k\mathbf{U}_i\mathbf{v} \quad ||\text{Multiplication of diagonal matrices commutes}\\
&= (\mathbf{I}-b_{ii}\mathbf{U}_i)\mathbf{U}_k\tilde{\mathbf{B}}\mathbf{v}+b_{ii}\mathbf{U}_k\mathbf{U}_i\mathbf{v}\\
&= (\mathbf{I}-b_{ii}\mathbf{U}_i)\mathbf{v}+b_{ii}\mathbf{U}_i\mathbf{v}=\mathbf{v}.
\end{aligned}$$

In both cases matrix $\mathbf{U}_k\mathbf{B}$ has a unit eigenvalue, and thus $\mathbf{I}-\mathbf{U}_k\mathbf{B}$ is singular. This is contradictory to the assumption that the model $(\mathbf{B},\mathbf{\Sigma}_{\mathbf{e}})$ is weakly stable, and so the model $(\tilde{\mathbf{B}},\tilde{\mathbf{\Sigma}}_{\mathbf{e}})$ must be weakly stable.

E.2 Equal Covariance Matrices

Then we show that in an arbitrary experiment $\mathcal{E}_k$ the two models produce data with the same covariance matrices. First, if variable $x_i\in\mathcal{J}_k$, then $\mathbf{U}_k\mathbf{U}_i=\mathbf{0}_{n\times n}$, $\mathbf{U}_k\mathbf{B}=\mathbf{U}_k\tilde{\mathbf{B}}$ (as shown above) and

$$\mathbf{U}_k\tilde{\mathbf{\Sigma}}_{\mathbf{e}}\mathbf{U}_k \quad = \quad \mathbf{U}_k(\mathbf{I}+\frac{b_{ii}}{1-b_{ii}}\mathbf{U}_i)\mathbf{\Sigma}_{\mathbf{e}}(\mathbf{I}+\frac{b_{ii}}{1-b_{ii}}\mathbf{U}_i)^T\mathbf{U}_k=\mathbf{U}_k\mathbf{\Sigma}_{\mathbf{e}}\mathbf{U}_k.$$

The covariance matrices are trivially equal:

$$\begin{aligned}
\tilde{\mathbf{C}}_{\mathbf{x}}^k &= (\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-1}(\mathbf{J}_k+\mathbf{U}_k\tilde{\mathbf{\Sigma}}_{\mathbf{e}}\mathbf{U}_k)(\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-T}\\
&= (\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}(\mathbf{J}_k+\mathbf{U}_k\mathbf{\Sigma}_{\mathbf{e}}\mathbf{U}_k)(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-T}=\mathbf{C}_{\mathbf{x}}^k.
\end{aligned}$$

Alternatively, if variable $x_i \in \mathcal{U}_k$, then $\mathbf{U}_k\mathbf{U}_i = \mathbf{U}_i$, and because

$$\begin{aligned}
(\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1} &= (\mathbf{I}-\mathbf{U}_k\mathbf{B}+\frac{b_{ii}}{1-b_{ii}}\mathbf{U}_k\mathbf{U}_i(\mathbf{I}-\mathbf{B}))(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}\\
&= \mathbf{I}+\frac{b_{ii}}{1-b_{ii}}\mathbf{U}_i(\mathbf{I}-\mathbf{U}_k\mathbf{B}-\mathbf{J}_k\mathbf{B})(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1} \quad ||\mathbf{U}_i\mathbf{J}_k=\mathbf{0}_{n\times n}\\
&= \mathbf{I}+\frac{b_{ii}}{1-b_{ii}}\mathbf{U}_i,
\end{aligned}$$

the covariance matrices are also equal:

$$\begin{aligned}
\tilde{\mathbf{C}}_{\mathbf{x}}^k &= (\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-1}(\mathbf{J}_k+\mathbf{U}_k\tilde{\mathbf{\Sigma}}_{\mathbf{e}}\mathbf{U}_k)(\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-T} \quad ||\text{definition of } \tilde{\mathbf{\Sigma}}_{\mathbf{e}}\\
&= (\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-1}(\mathbf{J}_k+\mathbf{U}_k(\mathbf{I}+\frac{b_{ii}}{1-b_{ii}}\mathbf{U}_i)\mathbf{\Sigma}_{\mathbf{e}}(\mathbf{I}+\frac{b_{ii}}{1-b_{ii}}\mathbf{U}_i)^T\mathbf{U}_k)(\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-T}\\
&= (\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-1}((\mathbf{I}+\frac{b_{ii}}{1-b_{ii}}\mathbf{U}_i)\mathbf{J}_k(\mathbf{I}+\frac{b_{ii}}{1-b_{ii}}\mathbf{U}_i)^T \quad ||\text{Multip. of diag. mat. commutes}\\
&\quad +\mathbf{U}_k(\mathbf{I}+\frac{b_{ii}}{1-b_{ii}}\mathbf{U}_i)\mathbf{\Sigma}_{\mathbf{e}}(\mathbf{I}+\frac{b_{ii}}{1-b_{ii}}\mathbf{U}_i)^T\mathbf{U}_k)(\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-T}\\
&= (\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-1}(\mathbf{I}+\frac{b_{ii}}{1-b_{ii}}\mathbf{U}_i)(\mathbf{J}_k+\mathbf{U}_k\mathbf{\Sigma}_{\mathbf{e}}\mathbf{U}_k)(\mathbf{I}+\frac{b_{ii}}{1-b_{ii}}\mathbf{U}_i)^T(\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-T} \quad ||\text{id. above}\\
&= (\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-1}(\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}(\mathbf{J}_k+\mathbf{U}_k\mathbf{\Sigma}_{\mathbf{e}}\mathbf{U}_k)\\
&\quad \cdot(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-T}(\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^T(\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-T}\\
&= (\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}(\mathbf{J}_k+\mathbf{U}_k\mathbf{\Sigma}_{\mathbf{e}}\mathbf{U}_k)(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-T}=\mathbf{C}_{\mathbf{x}}^k.
\end{aligned}$$

Appendix F. Derivation of Equation 13

Lemma 7 (Marginalization) showed that weak stability and experimental effects from an intervened variable $x_i \in \mathcal{J}_k$ to an observed variable $x_u \in \mathcal{U}_k$ are preserved (as part of the covariance matrix) when some variables in $\mathcal{U}_k$ are marginalized. Then, it is sufficient to show that Equation 13 applies in a weakly stable model where variables $\mathcal{U}_k \setminus \{x_j, x_u\}$ are marginalized. Lemma 8 (Self cycles) allows us to assume without loss of generality that there are no self-loops in this model.

Examine experiment $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ where $\mathcal{U}_k = \{x_j, x_u\}$ in the marginalized model $(\mathbf{B}, \mathbf{\Sigma}_{\mathbf{e}})$. The experimental effects in the experiment intervening on $\mathcal{J}_k \cup \{x_j\}$ are just the direct effects $t(x_i \rightsquigarrow x_u || \mathcal{J}_k \cup \{x_j\}) = b_{ui}$ and $t(x_j \rightsquigarrow x_u || \mathcal{J}_k \cup \{x_j\}) = b_{uj}$. The remaining experimental effects $t(x_i \rightsquigarrow x_u || \mathcal{J}_k)$ and $t(x_i \rightsquigarrow x_j || \mathcal{J}_k)$ appear in the matrix $((\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1})_{\mathcal{U}_k\mathcal{J}_k}$:

$$\begin{aligned}
((\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1})_{\mathcal{U}_k\mathcal{J}_k} &= (\mathbf{I}-\mathbf{B}_{\mathcal{U}_k\mathcal{U}_k})^{-1}\mathbf{B}_{\mathcal{U}_k\mathcal{J}_k} = \begin{bmatrix} 1 & -b_{ju} \\ -b_{uj} & 1 \end{bmatrix}^{-1} \begin{bmatrix} \cdots & b_{ji} & \cdots \\ \cdots & b_{ui} & \cdots \end{bmatrix}\\
&= \frac{1}{1-b_{uj}b_{ju}} \begin{bmatrix} 1 & b_{ju} \\ b_{uj} & 1 \end{bmatrix} \begin{bmatrix} \cdots & b_{ji} & \cdots \\ \cdots & b_{ui} & \cdots \end{bmatrix} = \begin{bmatrix} \cdots & \frac{b_{ji}+b_{ju}b_{ui}}{1-b_{uj}b_{ju}} & \cdots \\ \cdots & \frac{b_{ui}+b_{uj}b_{ji}}{1-b_{uj}b_{ju}} & \cdots \end{bmatrix}.
\end{aligned}$$

Now Equation 13 can be verified:

$$
\begin{aligned}
t(x_i \rightsquigarrow x_u || \mathcal{J}_k \cup \{x_j\}) + t(x_i \rightsquigarrow x_j || \mathcal{J}_k) t(x_j \rightsquigarrow x_u || \mathcal{J}_k \cup \{x_j\}) &= b_{ui} + \frac{b_{ji} + b_{ju} b_{ui}}{1 - b_{uj} b_{ju}} b_{uj} \\
= \frac{b_{ui} - b_{uj} b_{ju} b_{ui} + b_{uj} b_{ji} + b_{uj} b_{ju} b_{ui}}{1 - b_{uj} b_{ju}} = \frac{b_{ui} + b_{uj} b_{ji}}{1 - b_{uj} b_{ju}} &= t(x_i \rightsquigarrow x_u || \mathcal{J}_k).
\end{aligned}
$$

Appendix G. Proof of Lemma 9 (Union/Intersection Experiment)

In this proof, we first derive a linear equation system on the unknown experimental effects and then show that it has a unique solution under weak stability.

G.1 Generalizations of Equation 13

Equation 13 can be generalized to relate some experimental effects in $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ to some experimental effects in $\mathcal{E}_{k \cup l} = (\mathcal{J}_k \cup \mathcal{J}_l, \mathcal{U}_k \cap \mathcal{U}_l)$ by applying Equation 13 iteratively:

$$
t(x_i \rightsquigarrow x_u || \mathcal{J}_k) \quad = \quad t(x_i \rightsquigarrow x_u || \mathcal{J}_k \cup \mathcal{J}_l) + \sum_{x_j \in \mathcal{J}_l \setminus \mathcal{J}_k} t(x_i \rightsquigarrow x_j || \mathcal{J}_k) t(x_j \rightsquigarrow x_u || \mathcal{J}_k \cup \mathcal{J}_l). \tag{34}
$$

Here $x_i \in \mathcal{J}_k$, $x_u \in \mathcal{U}_k \cap \mathcal{U}_l$. Another way of writing the generalization relates some experimental effects in $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ to experimental effects in $\mathcal{E}_{k \cap l} = (\mathcal{J}_k \cap \mathcal{J}_l, \mathcal{U}_k \cup \mathcal{U}_l)$:

$$
t(x_i \rightsquigarrow x_u || \mathcal{J}_k \cap \mathcal{J}_l) \quad = \quad t(x_i \rightsquigarrow x_u || \mathcal{J}_k) + \sum_{x_j \in \mathcal{J}_k \setminus \mathcal{J}_l} t(x_i \rightsquigarrow x_j || \mathcal{J}_k \cap \mathcal{J}_l) t(x_j \rightsquigarrow x_u || \mathcal{J}_k).
$$

Here $x_i \in \mathcal{J}_k \cap \mathcal{J}_l$, $x_u \in \mathcal{U}_k$.

G.2 Equations for the Experimental Effects in the Union Experiment

First, partition $\mathcal{V}$ into the following disjoint sets: $\mathcal{I} = \mathcal{J}_k \cap \mathcal{J}_l$ (intervened in both experiments), $\mathcal{K} = \mathcal{J}_k \setminus \mathcal{J}_l$ (intervened only in $\mathcal{E}_k$), $\mathcal{L} = \mathcal{J}_l \setminus \mathcal{J}_k$ (intervened only in $\mathcal{E}_l$) and $\mathcal{O} = \mathcal{U}_k \cap \mathcal{U}_l$ (passively observed in both experiments). For each pair (x_k, x_u) with $x_k \in \mathcal{K}$ and $x_u \in \mathcal{O}$ we can form an equation of the form of Equation 34 using experimental effects from experiment $\mathcal{E}_k$:

$$
t(x_k \rightsquigarrow x_u || \mathcal{J}_k \cup \mathcal{J}_l) + \sum_{x_j \in \mathcal{L}} t(x_k \rightsquigarrow x_j || \mathcal{J}_k) t(x_j \rightsquigarrow x_u || \mathcal{J}_k \cup \mathcal{J}_l) \quad = \quad t(x_k \rightsquigarrow x_u || \mathcal{J}_k).
$$

Equations for all such pairs can be represented neatly by block matrices:

$$
(\mathbf{T}_{\mathbf{x}}^{k \cup l})_{\mathcal{O}\mathcal{K}} + (\mathbf{T}_{\mathbf{x}}^{k \cup l})_{\mathcal{O}\mathcal{L}} (\mathbf{T}_{\mathbf{x}}^{k})_{\mathcal{L}\mathcal{K}} \quad = \quad (\mathbf{T}_{\mathbf{x}}^{k})_{\mathcal{O}\mathcal{K}}.
$$

Similarly, equations can be formed for all pairs (x_k, x_u) with $x_k \in \mathcal{L}$ and $x_u \in \mathcal{O}$ using experimental effects from experiment $\mathcal{E}_l$. For pairs (x_k, x_u) with $x_k \in \mathcal{I}$ and $x_u \in \mathcal{O}$, equations could be formed using the experimental effects from either experiments, but it turns out that only equations using the experimental effects of experiment $\mathcal{E}_k$ are needed. The equations form the following system:

$$
\left[(\mathbf{T}_{\mathbf{x}}^{k \cup l})_{\mathcal{O}\mathcal{I}} \; (\mathbf{T}_{\mathbf{x}}^{k \cup l})_{\mathcal{O}\mathcal{K}} \; (\mathbf{T}_{\mathbf{x}}^{k \cup l})_{\mathcal{O}\mathcal{L}} \right] \underbrace{\begin{bmatrix} \mathbf{I}_{|\mathcal{I}|} & & \\ & \mathbf{I}_{|\mathcal{K}|} & (\mathbf{T}_{\mathbf{x}}^{l})_{\mathcal{K}\mathcal{L}} \\ (\mathbf{T}_{\mathbf{x}}^{k})_{\mathcal{L}\mathcal{I}} & (\mathbf{T}_{\mathbf{x}}^{k})_{\mathcal{L}\mathcal{K}} & \mathbf{I}_{|\mathcal{L}|} \end{bmatrix}}_{\mathbf{Q}} = \left[(\mathbf{T}_{\mathbf{x}}^{k})_{\mathcal{O}\mathcal{I}} \; (\mathbf{T}_{\mathbf{x}}^{k})_{\mathcal{O}\mathcal{K}} \; (\mathbf{T}_{\mathbf{x}}^{l})_{\mathcal{O}\mathcal{L}} \right]. \tag{35}
$$

G.3 Invertibility

Now, we know the matrix on the right and matrix $\mathbf{Q}$, and we would like to solve for the matrix on the left by multiplying from the right by $\mathbf{Q}^{-1}$. Thus, we need to show that $\mathbf{Q}$ is invertible. Since the variables in $\mathcal{O}$ do not appear in matrix $\mathbf{Q}$ in any way, consider a marginalized model $(\tilde{\mathbf{B}}, \tilde{\mathbf{\Sigma}}_\mathbf{e})$ over $\tilde{\mathcal{V}} = \mathcal{V} \setminus \mathcal{O}$, where variables $\mathcal{O}$ are marginalized. The marginalized experiments corresponding to experiments $\mathcal{E}_k$ and $\mathcal{E}_l$ are $\tilde{\mathcal{E}}_k = (\mathcal{I} \cup \mathcal{K}, \mathcal{L})$ and $\tilde{\mathcal{E}}_l = (\mathcal{I} \cup \mathcal{L}, \mathcal{K})$ respectively. If $(\mathbf{B}, \mathbf{\Sigma}_\mathbf{e})$ is weakly stable as we assume, also $(\tilde{\mathbf{B}}, \tilde{\mathbf{\Sigma}}_\mathbf{e})$ is weakly stable by Lemma 7 (Marginalization). All the experimental effects in $\mathbf{Q}$ are preserved in the marginalization. The blocks can be now expressed using Equation 9:

$$\begin{aligned}
(\mathbf{T}^k_\mathbf{x})_{\mathcal{L}\mathcal{I}} = (\tilde{\mathbf{T}}^k_\mathbf{x})_{\mathcal{L}\mathcal{I}} &= ((\mathbf{I} - \tilde{\mathbf{B}}_{\mathcal{L}\mathcal{L}})^{-1}\tilde{\mathbf{B}}_{\mathcal{L},\mathcal{I}\cup\mathcal{K}})_{\mathcal{L}\mathcal{I}} = (\mathbf{I} - \tilde{\mathbf{B}}_{\mathcal{L}\mathcal{L}})^{-1}\tilde{\mathbf{B}}_{\mathcal{L}\mathcal{I}},\\
(\mathbf{T}^k_\mathbf{x})_{\mathcal{L}\mathcal{K}} = (\tilde{\mathbf{T}}^k_\mathbf{x})_{\mathcal{L}\mathcal{K}} &= ((\mathbf{I} - \tilde{\mathbf{B}}_{\mathcal{L}\mathcal{L}})^{-1}\tilde{\mathbf{B}}_{\mathcal{L},\mathcal{I}\cup\mathcal{K}})_{\mathcal{L}\mathcal{K}} = (\mathbf{I} - \tilde{\mathbf{B}}_{\mathcal{L}\mathcal{L}})^{-1}\tilde{\mathbf{B}}_{\mathcal{L}\mathcal{K}},\\
(\mathbf{T}^l_\mathbf{x})_{\mathcal{K}\mathcal{L}} = (\tilde{\mathbf{T}}^l_\mathbf{x})_{\mathcal{K}\mathcal{L}} &= ((\mathbf{I} - \tilde{\mathbf{B}}_{\mathcal{K}\mathcal{K}})^{-1}\tilde{\mathbf{B}}_{\mathcal{K},\mathcal{I}\cup\mathcal{L}})_{\mathcal{K}\mathcal{L}} = (\mathbf{I} - \tilde{\mathbf{B}}_{\mathcal{K}\mathcal{K}})^{-1}\tilde{\mathbf{B}}_{\mathcal{K}\mathcal{L}}.
\end{aligned}$$

The matrices inverted in the expressions are invertible, because the marginalized model is weakly stable. Now $\mathbf{Q}$ can be written as a product of 3 simple square matrices:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{I}_{|\mathcal{I}|} & & \\ & \mathbf{I}_{|\mathcal{K}|} & (\mathbf{I} - \tilde{\mathbf{B}}_{\mathcal{K}\mathcal{K}})^{-1}\tilde{\mathbf{B}}_{\mathcal{K}\mathcal{L}} \\ (\mathbf{I} - \tilde{\mathbf{B}}_{\mathcal{L}\mathcal{L}})^{-1}\tilde{\mathbf{B}}_{\mathcal{L}\mathcal{I}} & (\mathbf{I} - \tilde{\mathbf{B}}_{\mathcal{L}\mathcal{L}})^{-1}\tilde{\mathbf{B}}_{\mathcal{L}\mathcal{K}} & \mathbf{I}_{|\mathcal{L}|} \end{bmatrix} =$$

$$\begin{bmatrix} \mathbf{I}_{|\mathcal{I}|} & & \\ & -(\mathbf{I} - \tilde{\mathbf{B}}_{\mathcal{K}\mathcal{K}})^{-1} & \\ & & (\mathbf{I} - \tilde{\mathbf{B}}_{\mathcal{L}\mathcal{L}})^{-1} \end{bmatrix} \left[\begin{array}{c|cc} \mathbf{I}_{|\mathcal{I}|} & & \\ \hline & \mathbf{I} - \tilde{\mathbf{B}}_{\mathcal{K}\mathcal{K}} & -\tilde{\mathbf{B}}_{\mathcal{K}\mathcal{L}} \\ \tilde{\mathbf{B}}_{\mathcal{L}\mathcal{I}} & -\tilde{\mathbf{B}}_{\mathcal{L}\mathcal{K}} & \mathbf{I} - \tilde{\mathbf{B}}_{\mathcal{L}\mathcal{L}} \end{array}\right] \begin{bmatrix} \mathbf{I}_{|\mathcal{I}|} & & \\ & -\mathbf{I}_{|\mathcal{K}|} & \\ & & \mathbf{I}_{|\mathcal{L}|} \end{bmatrix}.$$

The matrices on the left and on the right are invertible as block diagonal matrices with invertible blocks. Consider the middle matrix in the blocks indicated by the lines. Because the upper right-hand block is just zeros, the matrix is invertible if the two diagonal blocks are invertible. The lower right-hand block is invertible since the marginalized model is weakly stable in the experiment $(\mathcal{I}, \mathcal{K} \cup \mathcal{L})$. As a product of 3 invertible matrices matrix $\mathbf{Q}$ is invertible. Note that the factorization is valid also in the case where $\mathcal{I} = \emptyset$.

G.4 Matrix Equations for the Experimental Effects

The derivation of the equations and proof of invertibility for the intersection experiment proceeds very similarly. Here the formulas for solving the experimental effects in the union and intersection experiment are presented for completeness:

$$\begin{aligned}
\left[(\mathbf{T}^{k\cup l}_\mathbf{x})_{\mathcal{O}\mathcal{I}}\, (\mathbf{T}^{k\cup l}_\mathbf{x})_{\mathcal{O}\mathcal{K}}\, (\mathbf{T}^{k\cup l}_\mathbf{x})_{\mathcal{O}\mathcal{L}}\right] &= \left[(\mathbf{T}^{k}_\mathbf{x})_{\mathcal{O}\mathcal{I}}\, (\mathbf{T}^{k}_\mathbf{x})_{\mathcal{O}\mathcal{K}}\, (\mathbf{T}^{l}_\mathbf{x})_{\mathcal{O}\mathcal{L}}\right] \begin{bmatrix} \mathbf{I} & & \\ & \mathbf{I} & (\mathbf{T}^l_\mathbf{x})_{\mathcal{K}\mathcal{L}} \\ (\mathbf{T}^k_\mathbf{x})_{\mathcal{L}\mathcal{I}} & (\mathbf{T}^k_\mathbf{x})_{\mathcal{L}\mathcal{K}} & \mathbf{I} \end{bmatrix}^{-1},\\
\begin{bmatrix} (\mathbf{T}^{k\cap l}_\mathbf{x})_{\mathcal{K}\mathcal{I}} \\ (\mathbf{T}^{k\cap l}_\mathbf{x})_{\mathcal{L}\mathcal{I}} \\ (\mathbf{T}^{k\cap l}_\mathbf{x})_{\mathcal{O}\mathcal{I}} \end{bmatrix} &= \begin{bmatrix} \mathbf{I} & -(\mathbf{T}^l_\mathbf{x})_{\mathcal{K}\mathcal{L}} & \\ -(\mathbf{T}^k_\mathbf{x})_{\mathcal{L}\mathcal{K}} & \mathbf{I} & \\ -(\mathbf{T}^k_\mathbf{x})_{\mathcal{O}\mathcal{K}} & & \mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} (\mathbf{T}^l_\mathbf{x})_{\mathcal{K}\mathcal{I}} \\ (\mathbf{T}^k_\mathbf{x})_{\mathcal{L}\mathcal{I}} \\ (\mathbf{T}^k_\mathbf{x})_{\mathcal{O}\mathcal{I}} \end{bmatrix}.
\end{aligned}$$

See Appendix J on how to determine the full covariance matrices in the union and intersection experiments.

Appendix H. Proof of Lemma 13 (Perturbation of B)

Experiments $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ with $x_i \in \mathcal{J}_k$, $x_j \in \mathcal{U}_k$ do not have to be considered as the pair condition is not satisfied for the pair (x_i, x_j). Consider then experiments $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ with $x_j \in \mathcal{J}_k$. As explained in the text after Lemma 13, $\mathbf{B}$ and $\tilde{\mathbf{B}}$ differ only on the j:th row. Then, if $x_j \in \mathcal{J}_k$, we have that $\mathbf{U}_k\tilde{\mathbf{B}} = \mathbf{U}_k\mathbf{B}$ and the experimental effects must be equal.

That leaves us with experiments $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$ with $x_i \in \mathcal{U}_k$ and $x_j \in \mathcal{U}_k$. In the special case of experiment $\mathcal{E}_{k'} = (\mathcal{K}, \mathcal{L}) = (\mathcal{V} \setminus \{x_i, x_j\}, \{x_i, x_j\})$, the experimental effects are the same by the definition of the alternative coefficient matrix $\tilde{\mathbf{B}}$:

$$\tilde{\mathbf{T}}_{\mathbf{x}}^{k'} = (\mathbf{I} - \tilde{\mathbf{B}}_{\mathcal{L}\mathcal{L}})^{-1}\tilde{\mathbf{B}}_{\mathcal{L}\mathcal{K}} = (\mathbf{I} - \tilde{\mathbf{B}}_{\mathcal{L}\mathcal{L}})^{-1}(\mathbf{I} - \tilde{\mathbf{B}}_{\mathcal{L}\mathcal{L}})(\mathbf{I} - \mathbf{B}_{\mathcal{L}\mathcal{L}})^{-1}\mathbf{B}_{\mathcal{L}\mathcal{K}} = (\mathbf{I} - \mathbf{B}_{\mathcal{L}\mathcal{L}})^{-1}\mathbf{B}_{\mathcal{L}\mathcal{K}} = \mathbf{T}_{\mathbf{x}}^{k'}.$$

Otherwise the intervention set $\mathcal{J}_k$ has a presentation $\mathcal{J}_k = \mathcal{K} \cap (\mathcal{J}_k \cup \mathcal{L})$. We just noted that the experimental effects are the same in experiment $(\mathcal{K}, \mathcal{L})$. Earlier we showed that experimental effects are equal when x_j is intervened on, this holds in particular for experiment $(\mathcal{J}_k \cup \mathcal{L}, \mathcal{U}_k \setminus \mathcal{L})$. By Lemma 9 (Union/Intersection Experiment) the effects of an intersection experiment $\mathcal{E}_k$ are defined by the experimental effects of the two original experiments, so the experimental effects must be equal in experiment $\mathcal{E}_k$.

Appendix I. Proof of Lemma 14 (Perturbation of Σ_e)

Take any experiment $\mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k)$. The two models $(\mathbf{B}, \mathbf{\Sigma_e})$ and $(\tilde{\mathbf{B}}, \tilde{\mathbf{\Sigma}}_\mathbf{e})$ produce the same experimental effects. Then, we can prove the following identities:

$$\begin{aligned}
\mathbf{U}_k(\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})^{-1}\mathbf{J}_k &= \mathbf{U}_k(\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k, && (36)\\
(\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})^{-1}\mathbf{J}_k &= (\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k, && (37)\\
(\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})^{-1}\mathbf{J}_k(\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})^{-T} &= (\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k(\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-T}, && (38)\\
(\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})^{-1}\mathbf{U}_k(\mathbf{I} - \tilde{\mathbf{B}}) &= (\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}\mathbf{U}_k(\mathbf{I} - \mathbf{B}). && (39)
\end{aligned}$$

Equation 36 follows directly from the fact that the experimental effects of the two models are the same in experiment $\mathcal{E}_k$. Equation 37 is proven by the following:

$$\begin{aligned}
& (\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})^{-1}\mathbf{J}_k \quad ||\mathbf{U}_k + \mathbf{J}_k = \mathbf{I}\\
= \; & \mathbf{U}_k(\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})^{-1}\mathbf{J}_k + \mathbf{J}_k(\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})^{-1}\mathbf{J}_k \quad ||\mathbf{J}_k(\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}}) = \mathbf{J}_k\\
= \; & \mathbf{U}_k(\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})^{-1}\mathbf{J}_k + \mathbf{J}_k(\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})(\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})^{-1}\mathbf{J}_k\\
= \; & \mathbf{U}_k(\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})^{-1}\mathbf{J}_k + \mathbf{J}_k \quad ||\text{Eq. 36}\\
= \; & \mathbf{U}_k(\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k + \mathbf{J}_k = (\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k.
\end{aligned}$$

Equation 38 follows from Equation 37:

$$\begin{aligned}
& (\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})^{-1}\mathbf{J}_k(\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})^{-T} \quad ||\mathbf{J}_k\mathbf{J}_k = \mathbf{J}_k, \mathbf{J}_k = \mathbf{J}_k^T\\
= \; & (\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})^{-1}\mathbf{J}_k\mathbf{J}_k^T(\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})^{-T}\\
= \; & (\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})^{-1}\mathbf{J}_k((\mathbf{I} - \mathbf{U}_k\tilde{\mathbf{B}})^{-1}\mathbf{J}_k)^T \quad ||\text{Eq. 37}\\
= \; & (\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k((\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k)^T\\
= \; & (\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k(\mathbf{I} - \mathbf{U}_k\mathbf{B})^{-T}.
\end{aligned}$$

Equation 39 is proven by the following:

$$\begin{aligned}(\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-1}\mathbf{U}_k(\mathbf{I}-\tilde{\mathbf{B}}) &= (\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-1}(\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}}-\mathbf{J}_k)\\ &= \mathbf{I}-(\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-1}\mathbf{J}_k \quad ||\text{Eq. 37}\\ &= \mathbf{I}-(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k = (\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}\mathbf{U}_k(\mathbf{I}-\mathbf{B}).\end{aligned}$$

Finally, the covariance matrices produced by the two models can be shown to be equal:

$$\begin{aligned}\tilde{\mathbf{C}}_{\mathbf{x}}^k &= (\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-1}(\mathbf{J}_k+\mathbf{U}_k\tilde{\mathbf{\Sigma}}_{\mathbf{e}}\mathbf{U}_k)(\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-T} \quad ||\text{Definition of } \tilde{\mathbf{\Sigma}}_{\mathbf{e}}\\ &= (\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-1}\mathbf{J}_k(\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-T}+ \quad ||\text{Eq. 38 and 39}\\ &\quad (\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-1}\mathbf{U}_k(\mathbf{I}-\tilde{\mathbf{B}})(\mathbf{I}-\mathbf{B})^{-1}\mathbf{\Sigma}_{\mathbf{e}}(\mathbf{I}-\mathbf{B})^{-T}(\mathbf{I}-\tilde{\mathbf{B}})^T\mathbf{U}_k(\mathbf{I}-\mathbf{U}_k\tilde{\mathbf{B}})^{-T}\\ &= (\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}\mathbf{J}_k(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-T}+\\ &\quad (\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}\mathbf{U}_k(\mathbf{I}-\mathbf{B})(\mathbf{I}-\mathbf{B})^{-1}\mathbf{\Sigma}_{\mathbf{e}}(\mathbf{I}-\mathbf{B})^{-T}(\mathbf{I}-\mathbf{B})^T\mathbf{U}_k(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-T}\\ &= (\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}(\mathbf{J}_k+\mathbf{U}_k\mathbf{\Sigma}_{\mathbf{e}}\mathbf{U}_k)(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-T} = \mathbf{C}_{\mathbf{x}}^k.\end{aligned}$$

Appendix J. Covariance Matrices of Union and Intersection Experiments

Even if the set of experiments does not allow for the identification of the full model, consistent predictions are still possible in some unseen experimental settings assuming the data generating model is a linear cyclic model with latent variables. Lemma 9 already showed that the experimental effects can be predicted in the union and intersection experiments of any two already conducted experiments. In the following we extend this result to the prediction of the entire covariance matrices.

Let the data generating model be $(\mathbf{B},\mathbf{\Sigma}_{\mathbf{e}})$. Say we have conducted experiment $\mathcal{E}_k$ observing covariance matrix $\mathbf{C}_{\mathbf{x}}^k$ and experiment $\mathcal{E}_l$ observing covariance matrix $\mathbf{C}_{\mathbf{x}}^l$. By solving Equation 17 using the pseudoinverse we can find a matrix $\tilde{\mathbf{B}}$ that produces the same experimental effects in the two experiments. Now define

$$\begin{aligned}\tilde{\mathbf{M}}_1 &:= (\mathbf{I}-\mathbf{U}_{k\cup l}\tilde{\mathbf{B}})^{-1}\mathbf{J}_{k\cup l}(\mathbf{I}-\mathbf{U}_{k\cup l}\tilde{\mathbf{B}})^{-T},\\ \tilde{\mathbf{M}}_2 &:= (\mathbf{I}-\mathbf{U}_{k\cup l}\tilde{\mathbf{B}})^{-1}\mathbf{U}_{k\cup l}(\mathbf{I}-\tilde{\mathbf{B}}).\end{aligned}$$

using the estimate $\tilde{\mathbf{B}}$. Now, we can show that matrix $\tilde{\mathbf{M}}_1+\tilde{\mathbf{M}}_2\mathbf{C}_{\mathbf{x}}^k\tilde{\mathbf{M}}_2^T$ is equal to the covariance matrix $\mathbf{C}_{\mathbf{x}}^{k\cup l}$ that the true data generating model would produce in experiment $\mathcal{E}_{k\cup l}=(\mathcal{J}_{k\cup l},\mathcal{U}_{k\cup l})=(\mathcal{J}_k\cup\mathcal{J}_l,\mathcal{U}_k\cap\mathcal{U}_l)$:

$$\begin{aligned}&\tilde{\mathbf{M}}_1+\tilde{\mathbf{M}}_2\mathbf{C}_{\mathbf{x}}^k\tilde{\mathbf{M}}_2^T\\ =\;&(\mathbf{I}-\mathbf{U}_{k\cup l}\tilde{\mathbf{B}})^{-1}\mathbf{J}_{k\cup l}(\mathbf{I}-\mathbf{U}_{k\cup l}\tilde{\mathbf{B}})^{-T} \quad ||\text{Eq. 38 and 39}\\ &+(\mathbf{I}-\mathbf{U}_{k\cup l}\tilde{\mathbf{B}})^{-1}\mathbf{U}_{k\cup l}(\mathbf{I}-\tilde{\mathbf{B}})\mathbf{C}_{\mathbf{x}}^k(\mathbf{I}-\tilde{\mathbf{B}})^T\mathbf{U}_{k\cup l}(\mathbf{I}-\mathbf{U}_{k\cup l}\tilde{\mathbf{B}})^{-T}\\ =\;&(\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-1}\mathbf{J}_{k\cup l}(\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-T} \quad ||\text{Eq. 8}\\ &+(\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-1}\mathbf{U}_{k\cup l}(\mathbf{I}-\mathbf{B})\mathbf{C}_{\mathbf{x}}^k(\mathbf{I}-\mathbf{B})^T\mathbf{U}_{k\cup l}(\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-T}\\ =\;&(\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-1}\mathbf{J}_{k\cup l}(\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-T}+(\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-1}\mathbf{U}_{k\cup l}(\mathbf{I}-\mathbf{B})(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}\\ &\cdot(\mathbf{J}_k+\mathbf{U}_k\mathbf{\Sigma}_{\mathbf{e}}\mathbf{U}_k)(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-T}(\mathbf{I}-\mathbf{B})^T\mathbf{U}_{k\cup l}(\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-T} \quad ||\mathbf{U}_{k\cup l}=\mathbf{U}_l\mathbf{U}_k\mathbf{U}_k\\ =\;&(\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-1}\mathbf{J}_{k\cup l}(\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-T}+(\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-1}\mathbf{U}_l\mathbf{U}_k\mathbf{U}_k(\mathbf{I}-\mathbf{B})(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}\\ &\cdot(\mathbf{J}_k+\mathbf{U}_k\mathbf{\Sigma}_{\mathbf{e}}\mathbf{U}_k)(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-T}(\mathbf{I}-\mathbf{B})^T\mathbf{U}_k\mathbf{U}_k\mathbf{U}_l(\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-T} \quad ||\mathbf{U}_k=\mathbf{I}-\mathbf{J}_k\end{aligned}$$

$$\begin{aligned}
&= (\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-1}\mathbf{J}_{k\cup l}(\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-T}+(\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-1}\mathbf{U}_l\mathbf{U}_k((\mathbf{I}-\mathbf{U}_k\mathbf{B})-\mathbf{J}_k)(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-1}\\
&\quad\cdot(\mathbf{J}_k+\mathbf{U}_k\mathbf{\Sigma}_\mathbf{e}\mathbf{U}_k)(\mathbf{I}-\mathbf{U}_k\mathbf{B})^{-T}((\mathbf{I}-\mathbf{U}_k\mathbf{B})-\mathbf{J}_k)^T\mathbf{U}_k\mathbf{U}_l(\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-T}\quad||\mathbf{U}_k\mathbf{J}_k=\mathbf{0}_{n\times n}\\
&= (\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-1}\mathbf{J}_{k\cup l}(\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-T}\\
&\quad+(\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-1}\mathbf{U}_l\mathbf{U}_k(\mathbf{J}_k+\mathbf{U}_k\mathbf{\Sigma}_\mathbf{e}\mathbf{U}_k)\mathbf{U}_k\mathbf{U}_l(\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-T}\quad||\mathbf{U}_l\mathbf{U}_k\mathbf{U}_k=\mathbf{U}_{k\cup l}\\
&= (\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-1}(\mathbf{J}_{k\cup l}+\mathbf{U}_{k\cup l}\mathbf{\Sigma}_\mathbf{e}\mathbf{U}_{k\cup l})(\mathbf{I}-\mathbf{U}_{k\cup l}\mathbf{B})^{-T}=\mathbf{C}_\mathbf{x}^{k\cup l}.
\end{aligned}$$

To predict the whole covariance matrix in the intersection experiment, we need the passive observational data covariance matrix $\mathbf{C}_\mathbf{x}^0$ in addition to the observations in experiments $\mathcal{E}_k$ and $\mathcal{E}_l$. Now, define matrices

$$\begin{aligned}
\tilde{\mathbf{M}}_3 &:= (\mathbf{I}-\mathbf{U}_{k\cap l}\tilde{\mathbf{B}})^{-1}\mathbf{J}_{k\cap l}(\mathbf{I}-\mathbf{U}_{k\cap l}\tilde{\mathbf{B}})^{-T},\\
\tilde{\mathbf{M}}_4 &:= (\mathbf{I}-\mathbf{U}_{k\cap l}\tilde{\mathbf{B}})^{-1}\mathbf{U}_{k\cap l}(\mathbf{I}-\tilde{\mathbf{B}}).
\end{aligned}$$

Then, we can show that $\tilde{\mathbf{M}}_3+\tilde{\mathbf{M}}_4\mathbf{C}_\mathbf{x}^0\tilde{\mathbf{M}}_4^T$ is equal to the covariance matrix $\mathbf{C}_\mathbf{x}^{k\cap l}$ that the data generating model would produce in experiment $\mathcal{E}_{k\cap l}=(\mathcal{J}_{k\cap l},\mathcal{U}_{k\cap l})=(\mathcal{J}_k\cap\mathcal{J}_l,\mathcal{U}_k\cup\mathcal{U}_l)$:

$$\begin{aligned}
&\tilde{\mathbf{M}}_3+\tilde{\mathbf{M}}_4\mathbf{C}_\mathbf{x}^0\tilde{\mathbf{M}}_4^T\\
&= (\mathbf{I}-\mathbf{U}_{k\cap l}\tilde{\mathbf{B}})^{-1}\mathbf{J}_{k\cap l}(\mathbf{I}-\mathbf{U}_{k\cap l}\tilde{\mathbf{B}})^{-T}\quad||\text{Eq. 38 and 39}\\
&\quad+(\mathbf{I}-\mathbf{U}_{k\cap l}\tilde{\mathbf{B}})^{-1}\mathbf{U}_{k\cap l}(\mathbf{I}-\tilde{\mathbf{B}})\mathbf{C}_\mathbf{x}^0(\mathbf{I}-\tilde{\mathbf{B}})^T\mathbf{U}_{k\cap l}(\mathbf{I}-\mathbf{U}_{k\cap l}\tilde{\mathbf{B}})^{-T}\\
&= (\mathbf{I}-\mathbf{U}_{k\cap l}\mathbf{B})^{-1}\mathbf{J}_{k\cap l}(\mathbf{I}-\mathbf{U}_{k\cap l}\mathbf{B})^{-T}\quad||\text{Eq. 3}\\
&\quad+(\mathbf{I}-\mathbf{U}_{k\cap l}\mathbf{B})^{-1}\mathbf{U}_{k\cap l}(\mathbf{I}-\mathbf{B})\mathbf{C}_\mathbf{x}^0(\mathbf{I}-\mathbf{B})^T\mathbf{U}_{k\cap l}(\mathbf{I}-\mathbf{U}_{k\cap l}\mathbf{B})^{-T}\\
&= (\mathbf{I}-\mathbf{U}_{k\cap l}\mathbf{B})^{-1}\mathbf{J}_{k\cap l}(\mathbf{I}-\mathbf{U}_{k\cap l}\mathbf{B})^{-T}\\
&\quad+(\mathbf{I}-\mathbf{U}_{k\cap l}\mathbf{B})^{-1}\mathbf{U}_{k\cap l}(\mathbf{I}-\mathbf{B})(\mathbf{I}-\mathbf{B})^{-1}\mathbf{\Sigma}_\mathbf{e}(\mathbf{I}-\mathbf{B})^{-T}(\mathbf{I}-\mathbf{B})^T\mathbf{U}_{k\cap l}(\mathbf{I}-\mathbf{U}_{k\cap l}\mathbf{B})^{-T}\\
&= (\mathbf{I}-\mathbf{U}_{k\cap l}\mathbf{B})^{-1}(\mathbf{J}_{k\cap l}+\mathbf{U}_{k\cap l}\mathbf{\Sigma}_\mathbf{e}\mathbf{U}_{k\cap l})(\mathbf{I}-\mathbf{U}_{k\cap l}\mathbf{B})^{-T}=\mathbf{C}_\mathbf{x}^{k\cap l}.
\end{aligned}$$

The above formulas for the prediction of covariance matrices can be used iteratively to find consistent estimates for the covariance matrices in different experiments, as long as the intervention set of the experiment can be reached by taking successive unions and intersections from the intervention sets of the actually conducted experiments.[12]

Appendix K. LLC Algorithm

We show here that matrix $\mathbf{T}$ of the LLC learning method is full column rank if the pair condition is satisfied for all pairs. This implies that the coefficients or direct effects are fully identified.

First we show that the equations of the type of Equation 16 obtained in the union experiment $\mathcal{E}_{k\cup l}$ are merely linear combinations of equations obtained in experiment $\mathcal{E}_k$ and $\mathcal{E}_l$. This is a rather direct consequence of Lemma 9 and its proof in Appendix G. In an arbitrary experiment $\mathcal{E}_k$, equations for all pairs (x_i,x_u) with $x_i\in\mathcal{J}_k$ and $x_u\in\mathcal{U}_k$, can be represented neatly in matrix notation:

$$\begin{aligned}
\mathbf{B}_{\{x_u\}\mathcal{J}_k}+\mathbf{B}_{\{x_u\}(\mathcal{U}_k\setminus\{x_u\})}(\mathbf{T}_\mathbf{x}^k)_{(\mathcal{U}_k\setminus\{x_u\})\mathcal{J}_k} &= (\mathbf{T}_\mathbf{x}^k)_{\{x_u\}\mathcal{J}_k}\quad\Leftrightarrow\\
(\mathbf{B}_{\{x_u\}\mathcal{J}_k})^T+((\mathbf{T}_\mathbf{x}^k)_{(\mathcal{U}_k\setminus\{x_u\})\mathcal{J}_k})^T(\mathbf{B}_{\{x_u\}(\mathcal{U}_k\setminus\{x_u\})})^T &= ((\mathbf{T}_\mathbf{x}^k)_{\{x_u\}\mathcal{J}_k})^T.
\end{aligned}$$

[12] Note that if $\boldsymbol{\mu}_\mathbf{e}\neq\mathbf{0}$, $\tilde{\mathbf{M}}_2\boldsymbol{\mu}_\mathbf{x}^k$ and $\tilde{\mathbf{M}}_4\boldsymbol{\mu}_\mathbf{x}^0$ provide estimates for the observed means in the union and intersection experiments.

Now, partition $\mathcal{V}$ similarly as in Appendix G. Consider an arbitrary $x_u \in \mathcal{O}$ (observed in both experiments). Define $\tilde{\mathcal{O}} = \mathcal{O} \setminus \{x_u\}$. Equations corresponding to pairs (x_i, x_u) with $x_i \in \mathcal{I} \cup \mathcal{K}$ obtained in experiment $\mathcal{E}_k$ and equations corresponding to pairs (x_j, x_u) with $x_j \in \mathcal{L}$ obtained in experiment $\mathcal{E}_l$ can be collected into a single system constraining coefficients $b_{u\bullet}$:

$$\left[\begin{array}{ccc|c} \mathbf{I} & & ((\mathbf{T}_{\mathbf{x}}^k)_{\mathcal{L}\mathcal{I}})^T & ((\mathbf{T}_{\mathbf{x}}^k)_{\tilde{\mathcal{O}}\mathcal{I}})^T \\ & \mathbf{I} & ((\mathbf{T}_{\mathbf{x}}^k)_{\mathcal{L}\mathcal{K}})^T & ((\mathbf{T}_{\mathbf{x}}^k)_{\tilde{\mathcal{O}}\mathcal{K}})^T \\ & ((\mathbf{T}_{\mathbf{x}}^l)_{\mathcal{K}\mathcal{L}})^T & \mathbf{I} & ((\mathbf{T}_{\mathbf{x}}^l)_{\tilde{\mathcal{O}}\mathcal{L}})^T \end{array}\right] \left[\begin{array}{c} (\mathbf{B}_{\{x_u\}\mathcal{I}})^T \\ (\mathbf{B}_{\{x_u\}\mathcal{K}})^T \\ (\mathbf{B}_{\{x_u\}\mathcal{L}})^T \\ (\mathbf{B}_{\{x_u\}\tilde{\mathcal{O}}})^T \end{array}\right] = \left[\begin{array}{c} ((\mathbf{T}_{\mathbf{x}}^k)_{\{x_u\}\mathcal{I}})^T \\ ((\mathbf{T}_{\mathbf{x}}^k)_{\{x_u\}\mathcal{K}})^T \\ ((\mathbf{T}_{\mathbf{x}}^l)_{\{x_u\}\mathcal{L}})^T \end{array}\right]. \tag{40}$$

Notice, that the left-hand block of the matrix on the left is just the transpose of the $\mathbf{Q}$ matrix introduced in Appendix G. As $\mathbf{Q}$ was shown to be invertible under the assumption that the data generating model is weakly stable, we can multiply the equation group by $\mathbf{Q}^{-T}$ from the left. As blocks of Equation 35 in Appendix G we get the following identities:

$$\begin{aligned} \mathbf{Q}^{-T}\left[\begin{array}{c} ((\mathbf{T}_{\mathbf{x}}^k)_{\{x_u\}\mathcal{I}})^T \\ ((\mathbf{T}_{\mathbf{x}}^k)_{\{x_u\}\mathcal{K}})^T \\ ((\mathbf{T}_{\mathbf{x}}^l)_{\{x_u\}\mathcal{L}})^T \end{array}\right] &= \left[\begin{array}{c} ((\mathbf{T}_{\mathbf{x}}^{k\cup l})_{\{x_u\}\mathcal{I}})^T \\ ((\mathbf{T}_{\mathbf{x}}^{k\cup l})_{\{x_u\}\mathcal{K}})^T \\ ((\mathbf{T}_{\mathbf{x}}^{k\cup l})_{\{x_u\}\mathcal{L}})^T \end{array}\right], \\ \mathbf{Q}^{-T}\left[\begin{array}{c} ((\mathbf{T}_{\mathbf{x}}^k)_{\tilde{\mathcal{O}}\mathcal{I}})^T \\ ((\mathbf{T}_{\mathbf{x}}^k)_{\tilde{\mathcal{O}}\mathcal{K}})^T \\ ((\mathbf{T}_{\mathbf{x}}^l)_{\tilde{\mathcal{O}}\mathcal{L}})^T \end{array}\right] &= \left[\begin{array}{c} ((\mathbf{T}_{\mathbf{x}}^{k\cup l})_{\tilde{\mathcal{O}}\mathcal{I}})^T \\ ((\mathbf{T}_{\mathbf{x}}^{k\cup l})_{\tilde{\mathcal{O}}\mathcal{K}})^T \\ ((\mathbf{T}_{\mathbf{x}}^{k\cup l})_{\tilde{\mathcal{O}}\mathcal{L}})^T \end{array}\right]. \end{aligned}$$

Thus, multiplying the Equation 40 from the left by $\mathbf{Q}^{-T}$ produces the following equation system:

$$\begin{aligned} \left[\begin{array}{ccc|c} \mathbf{I} & & & ((\mathbf{T}_{\mathbf{x}}^{k\cup l})_{\tilde{\mathcal{O}}\mathcal{I}})^T \\ & \mathbf{I} & & ((\mathbf{T}_{\mathbf{x}}^{k\cup l})_{\tilde{\mathcal{O}}\mathcal{K}})^T \\ & & \mathbf{I} & ((\mathbf{T}_{\mathbf{x}}^{k\cup l})_{\tilde{\mathcal{O}}\mathcal{L}})^T \end{array}\right] \left[\begin{array}{c} (\mathbf{B}_{\{x_u\}\mathcal{I}})^T \\ (\mathbf{B}_{\{x_u\}\mathcal{K}})^T \\ (\mathbf{B}_{\{x_u\}\mathcal{L}})^T \\ (\mathbf{B}_{\{x_u\}\tilde{\mathcal{O}}})^T \end{array}\right] &= \left[\begin{array}{c} ((\mathbf{T}_{\mathbf{x}}^{k\cup l})_{\{x_u\}\mathcal{I}})^T \\ ((\mathbf{T}_{\mathbf{x}}^{k\cup l})_{\{x_u\}\mathcal{K}})^T \\ ((\mathbf{T}_{\mathbf{x}}^{k\cup l})_{\{x_u\}\mathcal{L}})^T \end{array}\right] \\ \Leftrightarrow \quad \left[\begin{array}{c} (\mathbf{B}_{\{x_u\}\mathcal{I}})^T + ((\mathbf{T}_{\mathbf{x}}^{k\cup l})_{\tilde{\mathcal{O}}\mathcal{I}})^T(\mathbf{B}_{\{x_u\}\tilde{\mathcal{O}}})^T \\ (\mathbf{B}_{\{x_u\}\mathcal{K}})^T + ((\mathbf{T}_{\mathbf{x}}^{k\cup l})_{\tilde{\mathcal{O}}\mathcal{K}})^T(\mathbf{B}_{\{x_u\}\tilde{\mathcal{O}}})^T \\ (\mathbf{B}_{\{x_u\}\mathcal{L}})^T + ((\mathbf{T}_{\mathbf{x}}^{k\cup l})_{\tilde{\mathcal{O}}\mathcal{L}})^T(\mathbf{B}_{\{x_u\}\tilde{\mathcal{O}}})^T \end{array}\right] &= \left[\begin{array}{c} ((\mathbf{T}_{\mathbf{x}}^{k\cup l})_{\{x_u\}\mathcal{I}})^T \\ ((\mathbf{T}_{\mathbf{x}}^{k\cup l})_{\{x_u\}\mathcal{K}})^T \\ ((\mathbf{T}_{\mathbf{x}}^{k\cup l})_{\{x_u\}\mathcal{L}})^T \end{array}\right]. \end{aligned}$$

For the union experiment $\mathcal{E}_{k\cup l} = (\mathcal{J}_{k\cup l}, \mathcal{U}_{k\cup l})$ we have that $\mathcal{I} \cup \mathcal{K} \cup \mathcal{L} = \mathcal{J}_{k\cup l}$ and $\tilde{\mathcal{O}} = \mathcal{U}_{k\cup l} \setminus \{x_u\}$. The equation system can be written in in the following simple form:

$$(\mathbf{B}_{\{x_u\}\mathcal{J}_{k\cup l}})^T + ((\mathbf{T}_{\mathbf{x}}^{k\cup l})_{(\mathcal{U}_{k\cup l}\setminus\{x_u\})\mathcal{J}_{k\cup l}})^T(\mathbf{B}_{\{x_u\}(\mathcal{U}_{k\cup l}\setminus\{x_u\})})^T = ((\mathbf{T}_{\mathbf{x}}^{k\cup l})_{\{x_u\}\mathcal{J}_{k\cup l}})^T.$$

These are all of the equations from experiment $\mathcal{E}_{k\cup l}$ constraining coefficients $b_{u\bullet}$. As we considered arbitrary $x_u \in \mathcal{O}$, the same procedure can be repeated for each $x_u \in \mathcal{O}$. This exhausts all equations obtained in the union experiment. All of the equations obtained in experiment $\mathcal{E}_{k\cup l}$ are thus linear combinations of some of the equations obtained in the original two experiments $\mathcal{E}_k$ and $\mathcal{E}_l$.

Finally, matrix $\mathbf{T}$ can be verified to have full column rank as follows. Matrix $\mathbf{T}$ being full column rank is equivalent to system $\mathbf{Tb} = \mathbf{t}$ having at most a unique solution. The original equation system $\mathbf{Tb} = \mathbf{t}$ consists of all the equations (like Equation 16) gathered in experiments $\{\mathcal{E}_k\}_{k=1,\ldots,K}$. We can always add equations that would be obtained in the union experiment $\mathcal{E}_{k\cup l}$ of two experiments $\mathcal{E}_k$

and $\mathcal{E}_l$ whose equations are already in the system, without further restricting the possible solutions of the system. This is because the added equations are merely linear combinations of some of the equations already in the system. If the pair condition is satisfied for all pairs, by adding always equations from the union experiments of two experiments, whose equations are already in the system, we are eventually able to add equations for experiments intervening on sets $\mathcal{V} \setminus \{x_u\}$, for all variables $x_u \in \mathcal{V}$ (this follows the rationale discussed after Definition 10). These equations specify the direct effects $\mathbf{b}$ directly and uniquely. Since the solution space was not restricted throughout the procedure of adding new equations, we can deduce that the original system had at most a unique solution, which implies that the original matrix $\mathbf{T}$ has full column rank.

References

K. A. Bollen. *Structural Equations with Latent Variables*. John Wiley & Sons, 1989.

D. M. Chickering. Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research*, 2:445–498, 2002a.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002b.

T. Claassen and T. Heskes. Causal discovery in multiple models from different experiments. In *Advances in Neural Information Processing Systems 23*, pages 415–423, 2010.

G. Cooper and C. Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 116–125, 1999.

D. Eaton and K. Murphy. Exact Bayesian structure learning from uncertain interventions. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, 2007.

F. Eberhardt and R. Scheines. Interventions and causal inference. *Philosophy of Science*, 74:5: 981–995, 2007.

F. Eberhardt, C. Glymour, and R. Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *Proceedings of the Twenty-First Conference Conference on Uncertainty in Artificial Intelligence*, pages 178–184, 2005.

F. Eberhardt, P. O. Hoyer, and R. Scheines. Combining experiments to discover linear cyclic models with latent variables. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010.

F. M. Fisher. A correspondence principle for simultaneous equation models. *Econometrica*, 38(1): pp. 73–92, 1970.

R. A. Fisher. *The Design of Experiments*. Hafner, 1935.

D. Geiger and D. Heckerman. Learning Gaussian networks. Technical Report MSR-TR-94-10, Microsoft Research, 1994.

C. Glymour, R. Scheines, P. Spirtes, and K. Kelly. *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science and Statistical Modeling*. Academic Press, 1987.

G. H. Hardy. *Divergent Series*. Oxford: Clarendon Press, 1949.

A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Causal discovery for linear cyclic models with latent variables. In *Proceedings of the Fifth European Workshop on Probabilistic Graphical Models*, 2010.

A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Noisy-or models with latent confounding. In *Proceedings of the Twenty-Seventh Conference Conference on Uncertainty in Artificial Intelligence*, 2011.

A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Experiment selection for causal discovery. Submitted, 2012.

S. Itani, M. Ohannessian, K. Sachs, G. P. Nolan, and M. A. Dahleh. Structure learning in causal cyclic networks. In *JMLR Workshop & Conference Proceedings*, volume 6, pages 165–176, 2008.

M. Koivisto and K. Sood. Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research*, 5:549–573, 2004.

S. L. Lauritzen and T. S. Richardson. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64:321–348, 2002.

D. Marbach, T. Schaffter, C. Mattiussi, and D. Floreano. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of Computational Biology*, 16(2):229–239, 2009.

S. J. Mason. Feedback theory : further properties of signal flow graphs. Technical Report 303, Research Laboratory of Electronics, Massachusetts Institute of Technology, 1956.

S. Meganck, B. Manderick, and P. Leray. A decision theoretic approach to learning Bayesian networks. Technical report, Vrije Universiteit Brussels, 2005.

K. P. Murphy. Active learning of causal Bayes net structure. Technical report, U.C. Berkeley, 2001.

E. Nyberg and K. Korb. Informative interventions. In *Causality and Probability in the Sciences*. College Publications, London, 2006.

J. Pearl. *Causality*. Oxford University Press, 2000.

R. J. Prill, D. Marbach, J. Saez-Rodriguez, P. K. Sorger, L. G. Alexopoulos, X. Xue, N. D. Clarke, G. Altan-Bonnet, and G. Stolovitzky. Towards a rigorous assessment of systems biology models: The DREAM 3 challenges. *PLoS ONE*, 5(2):e9202, 2010.

T. S. Richardson. *Feedback Models: Interpretation and Discovery*. PhD thesis, Carnegie Mellon, 1996.

M. Schmidt and K. Murphy. Modeling discrete interventional data using directed cyclic graphical models. In *Proceedings of the Twenty-Fifth Conference Conference on Uncertainty in Artificial Intelligence*, 2009.

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

J. Spencer. Minimal completely separating systems. *Journal of Combinatorial Theory*, 8(4):446 – 447, 1970.

P. Spirtes. Directed cyclic graphical representation of feedback models. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, 1995.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. MIT Press, 2nd edition, 2000.

G. Stolovitzky, R. J. Prill, and A. Califano. Lessons from the DREAM 2 challenges. *Annals of the New York Academy of Sciences*, 1158(1):159–195, 2009.

S. Tong and D. Koller. Active learning for structure in Bayesian networks. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 863–869, 2001.

T. Verma and J. Pearl. Causal networks: Semantics and expressiveness. In *Proceedings of the Fourth Conference Conference on Uncertainty in Artificial Intelligence*, pages 352–359, 1988.

S. Wright. The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):pp. 161–215, 1934.