# Causal discovery for linear cyclic models with latent variables

Antti Hyttinen[1], Frederick Eberhardt[2], and Patrik O. Hoyer[1,3]
[1] HIIT / Dept. of Computer Science, University of Helsinki, Finland
[2] Dept. of Philosophy, Washington University in St Louis, MO, USA
[3] CSAIL, Massachusetts Institute of Technology, Cambridge, MA, USA

**Supplementary material**

*Remark.* At the end of section 2 we mention the existence of special circumstances when it is possible to identify some direct effects $b(\bullet \to x_j)$ even when it is not the case that all pair conditions of the form $(\bullet, x_j)$ are satisfied. Here is an example. Consider a model for $\mathbf{V} = \{x_1, x_2, x_3\}$ where a single experiment with $\mathbf{J}_m = \{x_1\}$ has been performed. The pair condition is satisfied by this experiment for the pairs $(x_1, x_2)$ and $(x_1, x_3)$. One of the equations the experiment yields is $t(x_1 \rightsquigarrow x_3 || \mathbf{J}_m) = t(x_1 \rightsquigarrow x_2 || \mathbf{J}_m)\, b(x_2 \to x_3) + b(x_1 \to x_3)$. If it now turns out that the measured experimental effect $t(x_1 \rightsquigarrow x_2 || \mathbf{J}_m) = 0$, then the direct effect $b(x_1 \to x_3)$ is determined although the pair condition for the pair $(x_2, x_3)$ is not satisfied.

**Lemma 1.** *Let the true model generating the data be* $(\mathbf{B}, \mathbf{\Sigma_e})$. *For each of the experiments* $(\mathcal{E}_m)_{m=1,...,M}$ *the obtained data covariance matrix is* $\mathbf{\Sigma}^m_{\mathbf{x}}$. *If there is a direct effects matrix* $\widehat{\mathbf{B}} \neq \mathbf{B}$ *such that for all* $(\mathcal{E}_m)_{m=1,...,M}$ *and all* $x_i \in \mathbf{J}_m$ *and* $x_j \in \mathbf{U}_m$ *it produces the same experimental effects* $t(x_i \rightsquigarrow x_j \,||\, \mathbf{J}_m)$, *then the model* $(\widehat{\mathbf{B}}, \widehat{\mathbf{\Sigma}}_{\mathbf{e}})$ *with* $\widehat{\mathbf{\Sigma}}_{\mathbf{e}} = (\mathbf{I} - \widehat{\mathbf{B}})(\mathbf{I} - \mathbf{B})^{-1}\mathbf{\Sigma_e}(\mathbf{I} - \mathbf{B})^{-T}(\mathbf{I} - \widehat{\mathbf{B}})^T$ *has data covariance matrices* $\widehat{\mathbf{\Sigma}}^m_{\mathbf{x}} = \mathbf{\Sigma}^m_{\mathbf{x}}$ *for all* $m = 1, ..., M$.

*Proof.* Consider an experiment $\mathcal{E}_m = (\mathbf{J}_m, \mathbf{U}_m)$, i.e. an experiment in which the variables in $\mathbf{J}_m$ are intervened on. Additionally, we let $\mathbf{J}_m$ denote a diagonal matrix where element $\mathbf{J}_m[i, i] = 1$ if variable $x_i$ is intervened on, and $\mathbf{J}_m[i, i] = 0$ if $x_i$ is not intervened on. (Context will make it unambiguous whether it refers to the set or the matrix.) Also using $\mathbf{U}_k = \mathbf{I} - \mathbf{J}_k$, the matrices index the rows and columns corresponding to intervened or observed variables respectively. By these definitions we can write the manipulated model for this experiment as follows (see the definition of the manipulated model in Section 2 in the paper):

$$\begin{aligned} \mathbf{B}^m &= \mathbf{U}_m\mathbf{B} \\ \mathbf{\Sigma}^m_{\mathbf{e}} &= \mathbf{J}_m + \mathbf{U}_m\mathbf{\Sigma_e}\mathbf{U}_m. \end{aligned}$$

Solving eq. (1) in the paper for $\mathbf{x}$, for the manipulated model, we obtain $\mathbf{x} = (\mathbf{I} - \mathbf{B}^m)^{-1}\mathbf{e}$ with the above covariance matrix for $\mathbf{e}$. Hence the covariance matrix of the observed data in this experiment can be presented as

$$\mathbf{\Sigma}^m_{\mathbf{x}} = E\{\mathbf{x}\mathbf{x}^T\} = (\mathbf{I} - \mathbf{U}_m\mathbf{B})^{-1}\mathbf{\Sigma}^m_{\mathbf{e}}(\mathbf{I} - \mathbf{U}_m\mathbf{B})^{-T} = (\mathbf{I} - \mathbf{U}_m\mathbf{B})^{-1}(\mathbf{J}_m + \mathbf{U}_m\mathbf{\Sigma_e}\mathbf{U}_m)(\mathbf{I} - \mathbf{U}_m\mathbf{B})^{-T}$$

Because by the definition of the experiment all intervened variables are marginally independent with unit variance, they have no experimental effects on each other. That is, $\mathbf{J}_m(\mathbf{I} - \mathbf{U}_m\widehat{\mathbf{B}})^{-1} = \mathbf{J}_m(\mathbf{I} - \mathbf{U}_m\widehat{\mathbf{B}})(\mathbf{I} - \mathbf{U}_m\widehat{\mathbf{B}})^{-1} = \mathbf{J}_m$. Similarly, $\mathbf{J}_m(\mathbf{I} - \mathbf{U}_m\mathbf{B})^{-1} = \mathbf{J}_m$. The experimental effects part of the covariance matrix is thus

$$\begin{aligned} \mathbf{U}_m\mathbf{\Sigma}^m_{\mathbf{x}}\mathbf{J}_m &= \mathbf{U}_m(\mathbf{I} - \mathbf{U}_m\mathbf{B})^{-1}(\mathbf{J}_m + \mathbf{U}_m\mathbf{\Sigma_e}\mathbf{U}_m)(\mathbf{I} - \mathbf{U}_m\mathbf{B})^{-T}\mathbf{J}_m \\ &= \mathbf{U}_m(\mathbf{I} - \mathbf{U}_m\mathbf{B})^{-1}(\mathbf{J}_m + \mathbf{U}_m\mathbf{\Sigma_e}\mathbf{U}_m)\mathbf{J}_m \\ &= \mathbf{U}_m(\mathbf{I} - \mathbf{U}_m\mathbf{B})^{-1}\mathbf{J}_m \end{aligned}$$

By the antecendent of the Lemma the experimental effects of the two models are equal in experiment $\mathcal{E}_m$:

$$\mathbf{U}_m(\mathbf{I} - \mathbf{U}_m\widehat{\mathbf{B}})^{-1}\mathbf{J}_m = \mathbf{U}_m(\mathbf{I} - \mathbf{U}_m\mathbf{B})^{-1}\mathbf{J}_m$$

Combining the above two results it directly follows that also

$$(\mathbf{I} - \mathbf{U}_m\widehat{\mathbf{B}})^{-1}\mathbf{J}_m = (\mathbf{I} - \mathbf{U}_m\mathbf{B})^{-1}\mathbf{J}_m$$

Now

$$\begin{aligned}
(\mathbf{I} - \mathbf{U}_m\widehat{\mathbf{B}})^{-1}\mathbf{U}_m(\mathbf{I} - \widehat{\mathbf{B}}) &= (\mathbf{I} - \mathbf{U}_m\widehat{\mathbf{B}})^{-1}(\mathbf{I} - \mathbf{U}_m\widehat{\mathbf{B}} - \mathbf{J}_m) \\
&= \mathbf{I} - (\mathbf{I} - \mathbf{U}_m\widehat{\mathbf{B}})^{-1}\mathbf{J}_m \\
&= \mathbf{I} - (\mathbf{I} - \mathbf{U}_m\mathbf{B})^{-1}\mathbf{J}_m \\
&= (\mathbf{I} - \mathbf{U}_m\mathbf{B})^{-1}(\mathbf{I} - \mathbf{U}_m\mathbf{B} - \mathbf{J}_m) \\
&= (\mathbf{I} - \mathbf{U}_m\mathbf{B})^{-1}\mathbf{U}_m(\mathbf{I} - \mathbf{B})
\end{aligned}$$

and the data covariance matrices are equal:

$$\begin{aligned}
\widehat{\mathbf{\Sigma}}^m_{\mathbf{x}} &= (\mathbf{I} - \mathbf{U}_m\widehat{\mathbf{B}})^{-1}(\mathbf{J}_m + \mathbf{U}_m\widehat{\mathbf{\Sigma}}_{\mathbf{e}}\mathbf{U}_m)(\mathbf{I} - \mathbf{U}_m\widehat{\mathbf{B}})^{-T} \\
&= (\mathbf{I} - \mathbf{U}_m\widehat{\mathbf{B}})^{-1}\mathbf{J}_m(\mathbf{I} - \mathbf{U}_m\widehat{\mathbf{B}})^{-T} \\
&\quad +(\mathbf{I} - \mathbf{U}_m\widehat{\mathbf{B}})^{-1}\mathbf{U}_m(\mathbf{I} - \widehat{\mathbf{B}})(\mathbf{I} - \mathbf{B})^{-1}\mathbf{\Sigma}_{\mathbf{e}}(\mathbf{I} - \mathbf{B})^{-T}(\mathbf{I} - \widehat{\mathbf{B}})^T\mathbf{U}_m(\mathbf{I} - \mathbf{U}_m\widehat{\mathbf{B}})^{-T} \\
&= (\mathbf{I} - \mathbf{U}_m\mathbf{B})^{-1}\mathbf{J}_m(\mathbf{I} - \mathbf{U}_m\mathbf{B})^{-T} \\
&\quad +(\mathbf{I} - \mathbf{U}_m\mathbf{B})^{-1}\mathbf{U}_m(\mathbf{I} - \mathbf{B})(\mathbf{I} - \mathbf{B})^{-1}\mathbf{\Sigma}_{\mathbf{e}}(\mathbf{I} - \mathbf{B})^{-T}(\mathbf{I} - \mathbf{B})^T\mathbf{U}_m(\mathbf{I} - \mathbf{U}_m\mathbf{B})^{-T} \\
&= (\mathbf{I} - \mathbf{U}_m\mathbf{B})^{-1}(\mathbf{J}_m + \mathbf{U}_m\mathbf{\Sigma}_{\mathbf{e}}\mathbf{U}_m)(\mathbf{I} - \mathbf{U}_m\mathbf{B})^{-T} = \mathbf{\Sigma}^m_{\mathbf{x}}
\end{aligned}$$

□

**Theorem 1** (Completeness Theorem)**.** *Given the data covariance matrices from a sequence of experiments* $(\mathcal{E}_m)_{m=1,\ldots,M}$ *over the variables in* $\mathbf{V}$*, all direct effects* $b(x_i \rightarrow x_j)$ *are identified if and only if the pair condition is satisfied for all ordered pairs of variables w.r.t. these experiments.*[1]

*Proof.* This follows directly from the combination of Lemma 1 (above) and Theorem 1 in (Eberhardt et al., 2010) as follows. First, if the pair condition is satisfied for all pairs the sufficiency part of Theorem 1 of Eberhardt et al (2010) applies and $\mathbf{B}$ is identified. Second, when there is one or more pairs for which the pair condition is not satisfied, by the necessity part of their result there exists some $\widehat{\mathbf{B}}$, not equal to the true $\mathbf{B}$ consistent with all the experimental effects from the performed experiments. By Lemma 1 of the present paper, there exists values for $\widehat{\mathbf{\Sigma}}_{\mathbf{e}}$ such that the full covariance matrices (in all experiments) match those produced by the the original model. Hence based on the covariance matrices of the observed data we cannot distinguish between the models $(\mathbf{B}, \mathbf{\Sigma}_{\mathbf{e}})$ and $(\widehat{\mathbf{B}}, \widehat{\mathbf{\Sigma}}_{\mathbf{e}})$. Thus, $\mathbf{B}$ is not identified in this case. □

**Theorem 2** (Model Identifiability Theorem)**.** *Given a sequence of experiments* $(\mathcal{E}_m)_{m=1,\ldots,M}$ *over the variables in* $\mathbf{V}$ *the model* $(\mathbf{B}, \mathbf{\Sigma}_{\mathbf{e}})$ *is fully identified if and only if for each ordered pair of variables* $(x_i, x_j)$ *there is an experiment* $\mathcal{E}_b = (\mathbf{J}_b, \mathbf{U}_b)$ *with* $x_i \in \mathbf{J}_b$ *and* $x_j \in \mathbf{U}_b$ *and another experiment* $\mathcal{E}_e = (\mathbf{J}_e, \mathbf{U}_e)$ *with* $x_i, x_j \in \mathbf{U}_e$*.*

*Proof.* Assume first that for each ordered pair of variables $(x_i, x_j)$ there is an experiment $\mathcal{E}_b = (\mathbf{J}_b, \mathbf{U}_b)$ with $x_i \in \mathbf{J}_b$ and $x_j \in \mathbf{U}_b$. By Theorem 1 in Eberhardt et al (2010) the total effects $t(\bullet\rightsquigarrow\bullet)$ are identified. The direct effects can be computed simply by $\mathbf{B} = \mathbf{I} - \mathbf{D}\mathbf{T}^{-1}$, where $\mathbf{D}$ rescales the columns of $\mathbf{T}^{-1}$. Now for the pair $x_i, x_j$ there is an experiment $\mathcal{E}_e = (\mathbf{J}_e, \mathbf{U}_e)$ with $x_i, x_j \in \mathbf{U}_e$. The covariance matrix entries can be obtained

[1] Note the inevitable limitation of identifiability with regard to self-loops discussed in (Eberhardt et al., 2010).

from the formula $\mathbf{\Sigma}_{\mathbf{x}}^{e} = (\mathbf{I} - \mathbf{B})^{-1}(\mathbf{J}_e + \mathbf{U}_e\mathbf{\Sigma}_{\mathbf{e}}\mathbf{U}_e)(\mathbf{I} - \mathbf{B})^{-T}$ by rearranging $\mathbf{U}_e\mathbf{\Sigma}_{\mathbf{e}}\mathbf{U}_e = (\mathbf{I} - \mathbf{B})\mathbf{\Sigma}_{\mathbf{x}}^{e}(\mathbf{I} - \mathbf{B})^{T} - \mathbf{J}_e$. Since multiplication by $\mathbf{U}_e$ only puts the columns and rows corresponding to intervened variables to zero, we can read off the elements $\mathbf{\Sigma}_{\mathbf{e}}[i, i], \mathbf{\Sigma}_{\mathbf{e}}[j, j], \mathbf{\Sigma}_{\mathbf{e}}[i, j]$. Repeating this for each pair the whole matrix $\mathbf{\Sigma}_{\mathbf{e}}$ is determined.

If for some pair $(x_i, x_j)$ there is no experiment $\mathcal{E}_b$ in which $x_i \in \mathbf{J}_b$ and $x_j \in \mathbf{U}_b$, by the Completeness Theorem (above) $\mathbf{B}$ is not identified. If for some pair $(x_i, x_j)$ there is no experiment $\mathcal{E}_e = (\mathbf{J}_e, \mathbf{U}_e)$ with $x_i, x_j \in \mathbf{U}_e$, then the confounding edge is cut in every experiment, and so the confounding covariance parameter for this pair does not in any way affect the observed data and thus remains undetermined. □

## References

F. Eberhardt, P. O. Hoyer, and R. Scheines. 2010. Combining experiments to discover linear cyclic models with latent variables. In *AISTATS 2010*.