Causal Discovery from Experiments

Frederick Eberhardt

Joint work with Antti Hyttinen, Patrik Hoyer and Richard Scheines
Bayes network learning

- Interpretation of result: \( x \perp y \mid z \) and \( w \perp z \mid \{x, y\} \)
- Useful for:
  - Fewer parameters to represent joint distribution
  - Efficient inference algorithms
Limits of learning from passive observation

• ‘Markov equivalence class’: Many different DAGs represent same independencies, and yield identical scores!

• Several assumptions required:
  - Directed acyclic graph
  - Causal Markov condition
  - Causal faithfulness condition
  - No confounding hidden variables (‘no latents’)

• Even with all these assumptions satisfied, no uniformly consistent structure learning: For any given sample size, no guarantee that we are close to correct solution in terms of causal effects!
Two options of how to proceed

- Make stronger assumptions
  - sparsity of structure
  - form of functional relations
  - time order
  - ...

- Perform experiments
  - randomized controlled trials
  - natural experiments / instrumental variables
  - ...

Experiments!

- If we want to know what happens if we manipulate $z$, the best strategy is to manipulate $z$ and look at the consequences!

- Often, one does not have a particular intervention in mind, rather the goal is to understand the system (learn the full causal interaction graph) to the extent that one can, to at least qualitatively, predict the response to any manipulations.
Types of experiments:

- Hard (‘surgical’) or soft interventions
- Multiple interventions in a single experiment, or only one intervention per experiment allowed?
- Fixed sequence of experiments, or adaptive to the results of earlier experiments?
- Output of experiments
  - conditional independencies only
  - correlations between the different variables
  - the full manipulated distribution
  - ...

![Graph](attachment:image.png)
Assumptions about the model space

• Just as in the purely observational data situation, we need to decide which assumptions we are willing to make...

- Directed acyclic graph (DAG)? or potentially cyclic graphs?
- Causal Markov
- Causal faithfulness?
- No latents?
- Linear model?
Research questions

- Pick any combination of types of experiments and assumptions, and answer the following:
  - How many experiments are needed (worst case) to guarantee we know the full structure?
  - Which experiments need to be run? In the adaptive case, how are the experiments selected based on previous results?
  - How is the full graph inferred from the results of the experiments?

- Related work includes: (Cooper & Yoo; Eaton & Murphy; Nyberg & Korb; Schmidt & Murphy; Tong & Koller)
Example... (#1)

Assume: Markov, faithfulness, no latents, acyclic, linear
Experiments: surgical, fixed, independence tests, single

⇒ $N - 1$ experiments sufficient and in the worst case necessary, for graphs over $N > 2$ variables

Strategy: Each experiment intervenes on a different variable. One variable unintervened, no fully observational case required.

Example:
Example... (#2)

Assume: Markov, faithfulness, no latents, acyclic, linear
Experiments: surgical, fixed, independence tests, single

⇒ No sequence of experiments is sufficient to discover the worst case causal graphs over $N$ variables

Example: (the two graphs below are Markov equivalent observationally and for any single-variable interventions)

(But a double intervention on \( \{x, y\} \) would distinguish them)
Summary of experimental strategies

- using independence tests only, for $N$ variables

<table>
<thead>
<tr>
<th>Interventions per experiment</th>
<th>Strength of interventions</th>
<th>Number of experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No latents</td>
</tr>
<tr>
<td>Single</td>
<td>Surgical</td>
<td>$N - 1$</td>
</tr>
<tr>
<td>Multiple</td>
<td>Surgical</td>
<td>$\lceil \log_2(N) \rceil + 1$</td>
</tr>
</tbody>
</table>
Using experiments in linear models

- Example:
  (assume: linearity, faithfulness, acyclic, no latents)

\[
\begin{align*}
  x & := e_x \\
  y & := px + e_y \\
  z & := qx + ry + e_z
\end{align*}
\]

- General form:

\[
x := Bx + e
\]

where \( x \) is the vector of observed variables, and the latents are represented by non-zero off-diagonal entries in the covariance matrix \( \Sigma_e = E\{ee^T\} \)
Model: \[ x := Bx + e \quad \Sigma_e = E\{ee^T\} \]

‘Surgical’ interventions are obtained by, for each intervened variable, setting the corresponding row of \( B \) to zero and the corresponding row and column of \( \Sigma_e \) to zero (except for the diagonal element, which is set equal to one).

Note #1: Latent variables do not change these experimental effects in any way.

Note #2: No acyclicity assumption and no faithfulness assumption.
Linear constraints on direct effects

Consider:

Experiment intervening on all $x_j \in J$
Linear constraints on direct effects

Consider:

Experiment intervening on all $x_j \in J$
Linear constraints on direct effects

Consider:

Experiment intervening on all $x_j \in J$

The experimental effect of $x_j$ on $x_u$ can be written as a linear constraint on the direct effects into $x_u$:

$$t(x_j \rightsquigarrow x_u | J) = b(x_j \rightarrow x_u) + \sum_{x_i \in U} t(x_j \rightsquigarrow x_i | J) \times b(x_i \rightarrow x_u)$$
Linear constraints on direct effects

Consider:

Experiment intervening on all $x_j \in J$

The experimental effect of $x_j$ on $x_u$ can be written as a linear constraint on the direct effects into $x_u$:

$$t(x_j \leadsto x_u || J) = b(x_j \rightarrow x_u) + \sum_{x_i \in U} t(x_j \leadsto x_i || J) \times b(x_i \rightarrow x_u)$$
Linear constraints on direct effects

Consider:

Experiment intervening on all $x_j \in J$

The experimental effect of $x_j$ on $x_u$ can be written as a linear constraint on the direct effects into $x_u$:

$$t(x_j \leadsto x_u|J) = b(x_j \rightarrow x_u) + \sum_{x_i \in U} t(x_j \leadsto x_i|J) \times b(x_i \rightarrow x_u)$$
Linear constraints on direct effects

Consider:

Experiment intervening on all \( x_j \in J \)

The experimental effect of \( x_j \) on \( x_u \) can be written as a linear constraint on the direct effects into \( x_u \):

\[
t(x_j \leadsto x_u \| J) = b(x_j \to x_u) + \sum_{x_i \in U} t(x_j \leadsto x_i \| J) \times b(x_i \to x_u)
\]
\[ t(x \sim z | x) = b(x \to z) + t(x \sim y | x) \times b(y \to z) \]

\[ pr + q = b(x \to z) + p \times b(y \to z) \]

Matrix equation:

\[ h = Hb \]

- direct effects (unknown)
- experimental effects (measured)
Identifiability

The procedure is guaranteed to identify the generating model if and only if for each ordered pair \((x_i, x_j)\) there exists at least one experiment in which \(x_i\) is intervened on and \(x_j\) is not.

• this ‘pair condition’ is a necessary and sufficient condition to identify the causal structure among the observed variables \((B)\)

• using the observational covariance matrix \(C_x\) we get the error covariance matrix \(\Sigma_e\) by:

\[
C_x = (I - B)^{-1}\Sigma_e(I - B)^{-T} \\
\Rightarrow \\
\Sigma_e = (I - B)C_x(I - B)^T
\]
Sets of experiments satisfying the pair condition

- $N$ experiments intervening on all but one variable
- $N$ experiments intervening on a different single variable each time
- More efficient schemes:
  - e.g. for six variables $x_1, \ldots, x_6$, four experiments intervening on three variables each suffice:
    \[
    \{x_1, x_2, x_3\} \quad \{x_3, x_4, x_5\} \\
    \{x_5, x_6, x_1\} \quad \{x_2, x_4, x_6\}
    \]
# Summary of experimental strategies

- using independence tests vs. linearity

<table>
<thead>
<tr>
<th>Linearity</th>
<th>Interventions per experiment</th>
<th>Strength of interventions</th>
<th># of experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>No latents</td>
</tr>
<tr>
<td>No</td>
<td>Single</td>
<td>Surgical</td>
<td>$N - 1$</td>
</tr>
<tr>
<td>No</td>
<td>Multiple</td>
<td>Surgical</td>
<td>$[\log_2(N)]+1$</td>
</tr>
<tr>
<td>Yes</td>
<td>Single</td>
<td>Surgical</td>
<td>$N - 1$</td>
</tr>
<tr>
<td>Yes</td>
<td>Multiple</td>
<td>Surgical</td>
<td>$[\log_2(N)]+1$</td>
</tr>
</tbody>
</table>
Outline of Extensions

- Noisy-OR parametrization for binary variables
- Incorporating the assumption of faithfulness
- Application to DREAM network inference challenge
- Open problems and future directions
Noisy-OR parametrization for binary variables

• Q: Any similar procedures available for discrete-valued variables? Let’s start by considering binary variables.

• It is easy to show that, for unrestricted parametrizations, one cannot identify the full model from experiments intervening on only a single variable at a time.

• We consider a model based on the ‘noisy-OR’ parametrization

\[ X_j := \bigvee_{i \in \text{Pa}(j)} (B_{ij} \land X_i) \lor E_j \]

where the ‘leak’ variables can be mutually dependent, i.e. with arbitrary \( P(E_1, \ldots, E_n) \)
• Main result:
The noisy-OR model with latent confounding is identifiable from a set of experiments under the same conditions as the linear model, i.e. a necessary and sufficient condition on the experiments is that for each ordered variable pair \((x_i, x_j)\), there must exist an experiment in which \(x_i\) is intervened and \(x_j\) is non-intervened.

• In addition, we provide concrete algorithms for learning the model from data.

• Note: The model can also handle cases where turning a variable OFF is a cause for another variable (but note that this is not the same as ‘preventive’ causes)
Faithfulness

- The linear and noisy-or results presented so far did not assume faithfulness
- However, this led to demanding identifiability conditions. To what extent can we identify sparse models with fewer experiments if faithfulness is assumed?
- Example:

\[
\begin{align*}
x_1 & \indep x_3 \parallel x_1 \\
x_1 & \indep x_3 \mid x_2 \parallel x_1
\end{align*}
\]
‘Skeleton’ constraints:

1. If, in any experiment, two non-intervened variables \( x_i \) and \( x_j \) are marginally or conditionally independent (with any conditioning set), then \( b(x_j \rightarrow x_i) = b(x_i \rightarrow x_j) = \Sigma_{e[i,j]} = 0 \)

2. If, in any experiment, an intervened variable \( x_i \) is marginally or conditionally independent (with any conditioning set) of a non-intervened variable \( x_j \), then \( b(x_i \rightarrow x_j) = 0 \)

\[
\begin{align*}
\mathcal{X}_i \perp \perp \mathcal{X}_j \mid \ldots \mid \ldots & \quad \mathcal{X}_i \perp \perp \mathcal{X}_j \| \mathcal{X}_i \\
\quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \qa
• ‘Orientation’ constraints:

1. If, in an experiment where $x_i$ is intervened and $x_j$ and $x_k$ are non-intervened, we observe a non-zero experimental effect of $x_i$ on $x_j$ but a zero experimental effect of $x_i$ on $x_k$, we can infer that $b(x_j \rightarrow x_k) = 0$

2. If, in an experiment where $x_i$ is intervened and $x_j$ and $x_k$ are non-intervened, we observe a non-zero experimental effect of $x_i$ on $x_j$ and furthermore $x_i$ is conditionally independent of $x_k$ given $x_j$, then we infer that $b(x_k \rightarrow x_j) = 0$ and $\Sigma_e[j, k] = 0$
• The given skeleton and orientation rules are sound but they are not complete, i.e. they do not exhaust the set of inferences that faithfulness allows

• The general problem of designing a complete set of rules is still an open problem

• Simulations show that relying on faithfulness may be risky, because in limited sample sizes it can be hard to detect weak experimental effects.
DREAM data

- ‘Dialogue for Reverse Engineering Assessments and Methods’, a yearly held challenge for cellular network inference methods
- *In silico* data, i.e. simulated but realistic
- Networks of 10-100 genes, steady-state expression levels in experiments where a single gene is ‘knocked-down’ or ‘knocked-out’ at a time
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. 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.
Network inference results

Approx. 30 teams competed; the gray area indicates 1st – 10th best results in each case
The linear model is very competitive on this task, in particular for the large (100 node) networks.
Open problems & current research

1. Overlapping variables with experimental datasets

- Each experiment is a partition of the variables into 3 sets, i.e. \( \mathcal{E}_k = (\mathcal{J}_k, \mathcal{U}_k, \mathcal{L}_k) \), where \( \mathcal{J}_k \) denotes the intervened variables, \( \mathcal{U}_k \) the non-intervened but observed variables, and \( \mathcal{L}_k \) any non-intervened unobserved variables in this experiment.

- The ‘pair condition’ is satisfied for pair \( (x_i, x_j) \) if there exists an experiment \( \mathcal{E}_k \) such that \( x_i \in \mathcal{J}_k \) and \( x_j \in \mathcal{U}_k \).

- Identifiability:
  - ‘Worst-case necessity’: To guarantee identifiability, the pair condition must be satisfied for all ordered pairs
  - Sufficiency requires (slightly) stronger assumptions
2. General conditions on parametrizations

- Under what parametrizations is it possible to learn the full causal model from experiments that satisfy the pair condition for all ordered pairs of variables?
  - Linear model for continuous variables
  - Noisy-OR model for binary variables
  - ... (?)

Q: What are the general conditions?

3. Complete set of inference rules based on faithfulness?

- How to make the inference as reliable as possible, with a limited sample size
Thank you!

- A. Hyttinen, F. Eberhardt, and P. O. Hoyer, “Learning linear cyclic causal models with latent variables” (submitted)


- A. Hyttinen, F. Eberhardt, and P. O. Hoyer, “Causal discovery for linear cyclic models with latent variables”, PGM 2010

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

- F. Eberhardt, C. Glymour & R. Scheines, “On the Number of Experiments Sufficient and in the Worst Case Necessary to Identify All Causal Relations Among N Variables”, UAI 2005