Learning causal models from interventions: inference and active learning

Alain Hauser
Seminar für Statistik, ETH Zürich
February 1, 2013, Eindhoven

Joint work with Peter Bühlmann
Estimating DAG models

Concepts:

- Markov property
- Faithfulness
- Markov equivalence
- Interventions
Estimating DAG models

Concepts:
- Markov property
- Faithfulness
- Markov equivalence
- Interventions

Learning objective:
- DAG structure (⇝ model selection)
- Markov factors
- Causal effects (⇝ prediction)
Causal model: example

Random variables:

$X_1$: landlord’s presence
$X_2$: hagelslag
$X_3$: vanillavla
$X_4$: Jonas’ mood
Causal model: example

Random variables:

\( X_1 \): landlord’s presence
\( X_2 \): hagelslag
\( X_3 \): vanillavla
\( X_4 \): Jonas’ mood

Directed acyclic graph (DAG) of causal dependencies:
Causal model: example

Random variables:

$X_1$: landlord’s presence
$X_2$: hagelslag
$X_3$: vanillavla
$X_4$: Jonas’ mood

Directed acyclic graph (DAG) of causal dependencies:

Factorization of density:

$$f(x) = f(x_1)f(x_2|x_1)f(x_3|x_1)f(x_4|x_2, x_3)$$

$f$ has **Markov property** of $D$

Alain Hauser (ETH Zürich) Learning causal models from intervention Eindhoven 2013 3 / 24
Intervention: example

Random variables:

$X_1$: landlord’s presence
$X_2$: hagelslag
$X_3$: vanillavla
$X_4$: Jonas’ mood

True DAG $D$

Observational density: $f(x) = f(x_1)f(x_2|x_1)f(x_3|x_1)f(x_4|x_2, x_3)$
**Intervention: example**

Random variables:

$X_1$: landlord’s presence  
$X_2$: hagelslag  
$X_3$: vanillavla  
$X_4$: Jonas’ mood

Intervention at $X_2$: providing hagelslag

Observational density: $f(x) = f(x_1)f(x_2|x_1)f(x_3|x_1)f(x_4|x_2, x_3)$
Intervention: example

Random variables:

$X_1$: landlord’s presence
$X_2$: hagelslag
$X_3$: vanillavla
$X_4$: Jonas’ mood

Intervention DAG: $D(\{2\})$

Observational density: $f(x) = f(x_1)f(x_2|x_1)f(x_3|x_1)f(x_4|x_2, x_3)$

Interventional density: $f(x|\text{do}(X_2 = U)) = f(x_1)f'(x_2)f(x_3|x_1)f(x_4|x_2, x_3)$
Markov equivalence

A probability density in general obeys the Markov properties of several DAGs; those DAGs are called Markov equivalent. 

\[ \sim \text{ limited identifiability} \] under observational data

On the other hand, intervention effects do depend on the DAG.

\[ \Rightarrow \text{ improved identifiability of causal models under interventional data} \]
Markov equivalence

A probability density in general obeys the Markov properties of several DAGs; those DAGs are called **Markov equivalent** ⇝ **limited identifiability** under observational data

On the other hand, intervention effects **do** depend on the DAG ⇝ **improved identifiability** of causal models under interventional data
Interventional Markov equivalence

Assume experiment in which **different** interventions at targets $I_1, I_2, \ldots$ are performed, summerized as **family of targets** $\mathcal{I} = \{I_1, I_2, \ldots\}$.

Note: observational case corresponds to special family $\mathcal{I} = \{\emptyset\}$

**Definition (\(\mathcal{I}\)-Markov equivalence)**

Given a family of targets $\mathcal{I}$, two DAGs $D_1$ and $D_2$ are called **\(\mathcal{I}\)-Markov equivalent** if they produce the same class of tuples of interventional densities.
Interventional Markov equivalence

Assume experiment in which **different** interventions at targets $I_1, I_2, \ldots$ are performed, summerized as **family of targets** $I = \{I_1, I_2, \ldots\}$. Note: observational case corresponds to special family $I = \{\emptyset\}$

**Definition (\(I\)-Markov equivalence)**

Given a family of targets $I$, two DAGs $D_1$ and $D_2$ are called **\(I\)-Markov equivalent** if they produce the same class of tuples of interventional densities.

**Theorem (Hauser and Bühlmann, 2012a)**

Two DAGs $D_1$ and $D_2$ are **\(I\)-Markov equivalent** if and only if

1. $D_1$ and $D_2$ have the same skeleton and the same v-structures,
2. and $D_1^{(I)}$ and $D_2^{(I)}$ have the same skeleton for all $I \in I$.

Note: Theorem of Verma and Pearl (1990) for $I = \{\emptyset\}$. 

---

**Alain Hauser (ETH Zürich)**

Learning causal models from intervention

Eindhoven 2013
Interventional essential graph

Definition

Let $\mathcal{I}$ be a family of targets. The $\mathcal{I}$-essential graph of some DAG $D$ is defined as $\mathcal{E}_\mathcal{I}(D) := \bigcup_{D' \sim_\mathcal{I} D} D'$.

In words: $\mathcal{E}_\mathcal{I}(D)$ is a partially directed graph

- having the same skeleton as $D$
- with a directed edge where the corresponding arrows of all DAGs $\mathcal{I}$-equivalent to $D$ have the same orientation
- with an undirected edge where the orientation of the corresponding arrow is not common to all DAGs $\mathcal{I}$-equivalent to $D$
Interventional essential graph

**Definition**

Let $\mathcal{I}$ be a family of targets. The **$\mathcal{I}$-essential graph** of some DAG $D$ is defined as $\mathcal{E}_\mathcal{I}(D) := \bigcup_{D' \sim_D \mathcal{I}} D'$.

In words: $\mathcal{E}_\mathcal{I}(D)$ is a partially directed graph

- having the same skeleton as $D$
- with a **directed edge** where the corresponding arrows of all DAGs $\mathcal{I}$-equivalent to $D$ have the same orientation
- with an **undirected edge** where the orientation of the corresponding arrow is *not* common to all DAGs $\mathcal{I}$-equivalent to $D$

An $\mathcal{I}$-essential graph is a unique representation of an $\mathcal{I}$-Markov equivalence class.
Characterization of $\mathcal{I}$-essential graphs

**Theorem (Hauser and Bühlmann, 2012a)**

A graph $G$ is the $\mathcal{I}$-essential graph of a DAG $D$ if and only if

1. $G$ is a chain graph;
2. each chain component of $G$ is chordal;
3. $a \rightarrow b \leftarrow c$ is no induced subgraph of $G$;
4. $G$ has no line $a \rightarrow b$ for which there exists some $I \in \mathcal{I}$ such that $|I \cap \{a, b\}| = 1$;
5. every arrow $a \rightarrow b \in G$ is strongly $\mathcal{I}$-protected.

Reproduces a result of Andersson et al. (1997) for the observational case $\mathcal{I} = \{\emptyset\}$. 
Characterization of $\mathcal{I}$-essential graphs

**Theorem (Hauser and Bühlmann, 2012a)**

A graph $G$ is the $\mathcal{I}$-essential graph of a DAG $D$ if and only if

1. $G$ is a chain graph;
2. each chain component of $G$ is chordal;
3. $a \rightarrow b \leftarrow c$ is no induced subgraph of $G$;
4. $G$ has no line $a \rightarrow b$ for which there exists some $I \in \mathcal{I}$ such that $\vert I \cap \{a, b\} \vert = 1$;
5. every arrow $a \rightarrow b \in G$ is strongly $\mathcal{I}$-protected.

Reproduces a result of Andersson et al. (1997) for the observational case $\mathcal{I} = \{\emptyset\}$. 

Interventional Markov equivalence: example

Observational Markov equivalence class of $D$ with corresponding essential graph
Interventional Markov equivalence class of $D$ for family of targets $\mathcal{I} = \{\emptyset, \{2\}\}$. Corresponds to an experiment which measures

- observational data ($I = \emptyset$)
- interventional data from an intervention at $X_2$ ($I = \{2\}$)
Learning causal models

- Interventional data coming from different interventions: independent — but not identically distributed!
Learning causal models

- Interventional data coming from different interventions: independent — but not identically distributed!
- Nevertheless:
  - Gaussian or discrete causal models: analytical parameter estimation via MLE for given DAG structure
Learning causal models

- Intervenotional data coming from different interventions: independent — but not identically distributed!
- Nevertheless:
  - Gaussian or discrete causal models: analytical parameter estimation via MLE for given DAG structure
  - For fix $p$, optimization of the BIC leads to consistent model selection in the limit $n \to \infty$ (Hauser and Bühlmann, 2013)
Problem: **model selection** is computationally intrinsically hard (NP-hard; Chickering, 1996)
Learning causal models

- Problem: **model selection** is computationally intrinsically hard (NP-hard; Chickering, 1996)
- Replacing $\ell_0$ by $\ell_1$ regularization does not help; reason: **DAG constraint** (non-convex constraint!)
Learning causal models

- **Problem:** model selection is computationally intrinsically hard (NP-hard; Chickering, 1996)
- Replacing $\ell_0$ by $\ell_1$ regularization does not help; reason: DAG constraint (non-convex constraint!)
- **Solution:** causal inference via greedy algorithm on space of $\mathcal{I}$-essential graphs $\rightsquigarrow$ Greedy Interventional Equivalence Search (GIES): natural generalization of the Greedy Equivalence Search (GES) algorithm of Chickering (2002) to interventional data
GIES: example step

- Main idea of GIES: greedy optimization of BIC by traversing space of $I$-essential graphs
- Small steps: proceed from one $I$-essential graph to a neighbor
- Search directions: **forward** (adding edges), **backward** (removing edges), **turning** (reversing edges)
GIES: example step

- Main idea of GIES: greedy optimization of BIC by traversing space of $\mathcal{I}$-essential graphs
- Small steps: proceed from one $\mathcal{I}$-essential graph to a neighbor
- Search directions: **forward** (adding edges), **backward** (removing edges), **turning** (reversing edges)

Possible forward step:

1

2 3

4

G

D

D′

representative

\textbf{neighbor}

\textbf{\textit{ess. graph}}
GIES: example step

- Main idea of GIES: greedy optimization of BIC by traversing space of $\mathcal{I}$-essential graphs
- Small steps: proceed from one $\mathcal{I}$-essential graph to a neighbor
- Search directions: forward (adding edges), backward (removing edges), turning (reversing edges)

Possible forward step:
GIES: example step

- Main idea of GIES: greedy optimization of BIC by traversing space of $\mathcal{I}$-essential graphs
- Small steps: proceed from one $\mathcal{I}$-essential graph to a neighbor
- Search directions: **forward** (adding edges), **backward** (removing edges), **turning** (reversing edges)

Possible forward step:
GIES: example step

- Main idea of GIES: greedy optimization of BIC by traversing space of $\mathcal{I}$-essential graphs
- Small steps: proceed from one $\mathcal{I}$-essential graph to a neighbor
- Search directions: forward (adding edges), backward (removing edges), turning (reversing edges)

Possible forward step:

\[
\begin{align*}
G & \quad \text{representative} \\
D & \quad \text{neighbor} \\
D' & \quad \text{ess. graph} \\
G' & \quad \text{ess. graph}
\end{align*}
\]
Search space: DAGs vs. essential graphs

Neglecting (interventional) Markov equivalence narrows search space
Search space: DAGs vs. essential graphs

Neglecting (interventional) Markov equivalence narrows search space
Search space: DAGs vs. essential graphs

Neglecting (interventional) Markov equivalence narrows search space.
Search space: DAGs vs. essential graphs

Neglecting (interventional) Markov equivalence narrows search space

\[ G \rightarrow D \rightarrow D' \rightarrow G' \]

\[ G \rightarrow D' \rightarrow G' \]

1. Representative
2. Neighbor
Search space: DAGs vs. essential graphs

Neglecting (interventional) Markov equivalence narrows search space
Simulation study: structure learning

SHD between true DAG and estimated $\mathcal{I}$-essential graph ($n = 1000, p = 20$). For oracle estimates: number of non-orientable edges
Simulation study: structure learning

SHD between estimated and true $\mathcal{I}$-essential graphs ($p = 20$). Upper part: observational data; lower part: $k = 12$ intervention targets of size 4.
DREAM4 *in silico* network challenge

- **Goal**: learn structure of gene regulatory network, predict intervention effects
- **Data**: realistic *in silico* steady-state and time series data, observational and interventional data points
- **Our proceeding**: *cross-validation* of gene expression levels under interventions.
- **Compare CV-values to those of algorithms ignoring interventional nature of data**

Competition website:
DREAM4 challenge: results

\[ \Delta \text{MSE} := \text{MSE of competitor} - \text{MSE of GIES} \]

Conclusions:
- slight advantage over competing methods
- estimation sensitive to model misspecification: acyclicity and normality assumptions violated
Active learning

- Up to now: given list of interventions; learning causal models from corresponding data
Active learning

- Up to now: given list of interventions; learning causal models from corresponding data
- Natural question: given intervention experiments performed so far, what is best next intervention?
Active learning: settings

- Which single-vertex intervention yields maximal number of orientable edges? \(\leadsto\) algorithm OptSingle

- Which set of (multi-vertex) interventions guarantees full identifiability with a minimum number of experiments? \(\leadsto\) algorithm OptUnb

Proof of optimality proves conjecture of Eberhardt (2008) concerning number of interventions necessary and sufficient for full identifiability (Hauser and Bühlmann, 2012b)
Active learning: settings

- Which single-vertex intervention yields maximal number of orientable edges? \(\leadsto\) algorithm OptSingle
- Which set of (multi-vertex) interventions guarantees full identifiability with a minimum number of experiments? \(\leadsto\) algorithm OptUnb

Proof of optimality proves conjecture of Eberhardt (2008) concerning number of interventions necessary and sufficient for full identifiability (Hauser and Bühlmann, 2012b)
Number of intervention steps needed for full identifiability of DAGs, measured in targets ($T$) or intervened variables ($V$). Thin lines: Kaplan-Meier estimates; colored bands: 95% confidence region.
Conclusions

- Causal models not fully identifiable from observational data
- Interventional data improves identifiability; identifiability specified and characterized by $\mathcal{I}$-essential graphs
Conclusions

- Causal models not fully identifiable from observational data
- Intervenational data improves identifiability; identifiability specified and characterized by $\mathcal{I}$-essential graphs
- Gaussian causal models: consistent model selection through maximization of BIC; computationally feasible with greedy learning
Conclusions

- Causal models not fully identifiable from observational data
- Intervenational data improves identifiability; identifiability specified and characterized by $\mathcal{I}$-essential graphs
- Gaussian causal models: consistent model selection through maximization of BIC; computationally feasible with greedy learning
- Neglection of interventional Markov equivalence leads to worse structure learning
Conclusions

- Causal models not fully identifiable from observational data
- Interventional data improves identifiability; identifiability specified and characterized by $\mathcal{I}$-essential graphs
- Gaussian causal models: consistent model selection through maximization of BIC; computationally feasible with greedy learning
- Neglection of interventional Markov equivalence leads to worse structure learning
- Active learning strategy leads to significantly faster identification of causal models than randomly chosen interventions
Future work

- Identifiability results for “non-perfect” interventions
Future work

- Identifiability results for “non-perfect” interventions
- More experiments needed: discrete data, high-dimensional data
Future work

- Identifiability results for “non-perfect” interventions
- More experiments needed: discrete data, high-dimensional data
- Accounting for hidden variables, confounders, etc.
Future work

- Identifiability results for “non-perfect” interventions
- More experiments needed: discrete data, high-dimensional data
- Accounting for hidden variables, confounders, etc.

Merci viu mau!
Thank you!
References


