Constraints on marginalized DAGs and their uses.

Robin J. Evans
www.stats.ox.ac.uk/∼evans

Department of Statistics, University of Oxford

Algebraic Statistics Workshop, NIMS
17th July 2014

Outline

Correlation does not imply causation

Correlation does not imply causation

Find a Job | M&S Wine | Feedback   Wednesday, N

MailOnline

Home | News | U.S. | Sport | TV&Showbiz | Femail | Health | Science | Money | Video | Cof

Health Home | Health Directory | Health Boards | Diets | MyDish Recipe Finder

How a short nap can raise the risk of diabetes: Study finds people who have a siesta are more likely to have high blood pressure and high cholesterol

- **Napping for more than 30 minutes at a time can raise the risk of diabetes, according to a new study**
- **It can also increase likelihood of high blood pressure and high cholesterol**

By PAT HAGAN

PUBLISHED: 01:04, 21 September 2013 | **UPDATED:** 10:34, 21 September 2013

598 shares   102 View comments

They were much favoured by Margaret Thatcher, Albert Einstein and Winston Churchill.

But while afternoon naps may revitalise tired brains, they can also increase the risk of diabetes, according to new research.

Correlation does not imply causation

Find a Job | M&S Wine | Feedback Wednesday, N

MailOnline

Home | News | U.S. | Sport | TV&Showbiz | Fem

Health Home | Health Directory | Health Boards | Diets | M

How a short nap can rais diabetes: Study finds pe siesta are more likely to pressure and high chole

- **Napping for more than 30 minutes at a tim according to a new study**
- **It can also increase likelihood of high bloo**

By PAT HAGAN

PUBLISHED: 01:04, 21 September 2013 | **UPDATED:** 10:34, 21 Septemb

598 shares

They were much favoured by Margaret Thatcher, Alber

But while afternoon naps may revitalise tired brains, the according to new research.

ELSEVIER

Sleep Medicine

Volume 14, Issue 10, October 2013, Pages 950–954

Original Article

Longer habitual afternoon napping is associated with a higher risk for impaired fasting plasma glucose and diabetes mellitus in older adults: results from the Dongfeng–Tongji cohort of retired workers

Weimin Fang[a, b], Zhongliang Li[a], Li Wu[a], Zhongqiang Cao[a], Yuan Liang[a, c], Handong Yang[d], Youjie Wang[a, b], Tangchun Wu[a]

[a] Ministry of Education Key Laboratory of Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, China

[b] Department of Maternal and Child Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, China

[c] Department of Social Medicine, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, China

[d] Dongfeng General Hospital, Dongfeng Motor Corporation and Hubei University of Medicine, China

Abstract

Objectives

Afternoon napping is a common habit in China. We used data obtained from the Dongfeng–Tongji cohort to examine if duration of habitual afternoon napping was associated with risks for impaired fasting plasma glucose (IFG) and diabetes mellitus (DM) in a Chinese elderly population.

Methods

Correlation does not imply causation

Find a Job | M&S Wine | Feedback Wednesday, N

MailOnline

Home | News | U.S. | Sport | TV&Showbiz | Fem

Health Home | Health Directory | Health Boards | Diets | M

How a short nap can rais diabetes: Study finds pe siesta are more likely to pressure and high chole

- Napping for more than 30 minutes at a tim according to a new study
- It can also increase likelihood of high bloo

By PAT HAGAN

But while afternoon naps may revitalise tired brains, the according to new research.

Sleep Medicine

Volume 14, Issue 10, October 2013, Pages 950–954

ELSEVIER

Original Article

Longer habitual afternoon napping is associated with a higher risk for impaired fasting plasma glucose and diabetes mellitus in older adults: results from the Dongfeng–Tongji cohort of retired workers

Weimin Fang[a, b], Zhongliang Li[a], Li Wu[a], Zhongqiang Cao[a], Yuan Liang[a, c], Handong Yang[d], Youjie Wang[a, b], Tangchun Wu[a]

[a] Ministry of Education Key Laboratory of Environment and Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, China

[b] Department of Maternal and Child Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, China

> "Dr Matthew Hobbs, head of research for Diabetes UK, said there was no proof that napping **actually caused** diabetes."

Abstract

Objectives

Afternoon napping is a common habit in China. We used data obtained from the Dongfeng–Tongji cohort to examine if duration of habitual afternoon napping was associated with risks for impaired fasting plasma glucose (IFG) and diabetes mellitus (DM) in a Chinese elderly population.

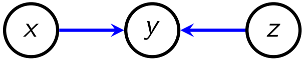
Methods

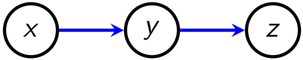
Distinguishing Between Causal Models

But can we still tell what causes what from observational data?

Distinguishing Between Causal Models

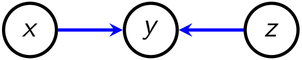
But can we still tell what causes what from observational data?

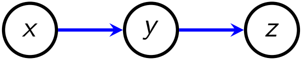




Distinguishing Between Causal Models

But can we still tell what causes what from observational data?



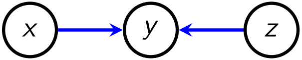


$$X \perp\!\!\!\perp Z$$

$$p(x, z) = p(x)p(z)$$

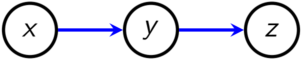
Distinguishing Between Causal Models

But can we still tell what causes what from observational data?



$X \perp\!\!\!\perp Z$

$p(x, z) = p(x)p(z)$

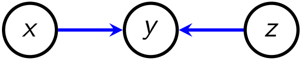


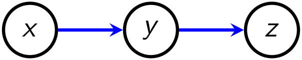
$X \perp\!\!\!\perp Z \mid Y$

$p(y)p(x, y, z) = p(x, y)p(y, z)$

Distinguishing Between Causal Models

But can we still tell what causes what from observational data?





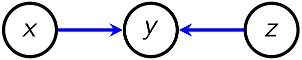
$$X \perp\!\!\!\perp Z$$
$$p(x, z) = p(x)p(z)$$

$$X \perp\!\!\!\perp Z \mid Y$$
$$p(y)p(x, y, z) = p(x, y)p(y, z)$$

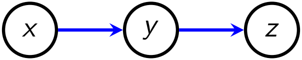
Maybe!

Distinguishing Between Causal Models

But can we still tell what causes what from observational data?



$$X \perp\!\!\!\perp Z$$
$$p(x,z) = p(x)p(z)$$



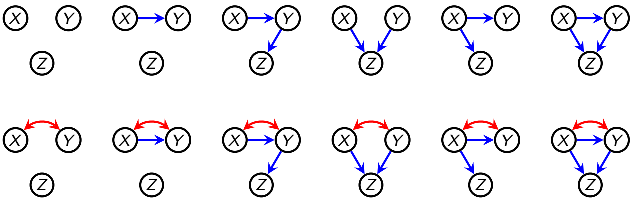
$$X \perp\!\!\!\perp Z \mid Y$$
$$p(y)p(x,y,z) = p(x,y)p(y,z)$$

Maybe!

In order to do this well, we need to understand in what ways causal models will be **observationally** different.

Structure Learning

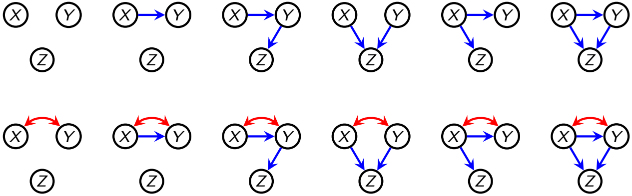
Given a distribution P (or rather data from P) and a set of possible causal models...



...return list of models which are compatible with data.

Structure Learning

Given a distribution P (or rather data from P) and a set of possible causal models...



...return list of models which are compatible with data.

We can do this by testing whether constraints implied by the model(s) are satisfied by P. e.g. PC, FCI algorithms.

To do this we need to know what the constraints are (the focus of this talk).

Outline

1. Introduction
2. Conditional Independence and Algebraic Models
3. DAGs
4. Margins of DAG Models
5. Ordinary Markov Model
6. Verma Constraints
7. Results
8. Inequalities
9. Summary

Outline

Models for Contingency Tables

Take finite discrete random variables $X_V = (X_1, \ldots, X_n)$.

Models for Contingency Tables

Take finite discrete random variables $X_V = (X_1, \ldots, X_n)$.

For $x_V = (x_1, \ldots, x_n)$, joint distribution is parameterized by

$$p(x_V) = p(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n).$$

We can consider a statistical model defined by polynomial constraints in the indeterminates $p(x_1, \ldots, x_n)$. We always assume

$$\sum_{x_V} p(x_V) = 1, \qquad\qquad p(x_V) > 0 \qquad \forall x_V.$$

Margins

For $M \subseteq V$, the **marginal distribution** over X_M is

$$p(x_M) = \sum_{x_{V \setminus M}} p(x_V) = \sum_{x_{V \setminus M}} p(x_M, x_{V \setminus M}).$$

Margins

For $M \subseteq V$, the **marginal distribution** over X_M is

$$p(x_M) = \sum_{x_{V \setminus M}} p(x_V) = \sum_{x_{V \setminus M}} p(x_M, x_{V \setminus M}).$$

A **conditional distribution** of X_A given X_B is

$$p(x_A \,|\, x_B) = \frac{p(x_A, x_B)}{p(x_B)}.$$

Margins

For $M \subseteq V$, the **marginal distribution** over X_M is

$$p(x_M) = \sum_{x_{V\setminus M}} p(x_V) = \sum_{x_{V\setminus M}} p(x_M, x_{V\setminus M}).$$

A **conditional distribution** of X_A given X_B is

$$p(x_A \,|\, x_B) = \frac{p(x_A, x_B)}{p(x_B)}.$$

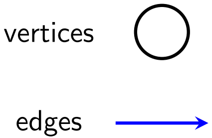
A **conditional independence** statement $X_A \perp\!\!\!\perp X_B \,|\, X_C$ assumes that $p(x_A \,|\, x_B, x_C) = p(x_A \,|\, x_C)$, or equivalently

$$p(x_A, x_B, x_C) \cdot p(x_C) - p(x_A, x_C) \cdot p(x_B, x_C) = 0$$

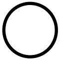
for all x_A, x_B, x_C.

Outline

Directed Acyclic Graphs

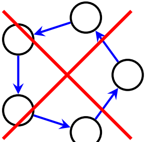


Directed Acyclic Graphs

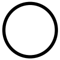
vertices 

edges 

no directed cycles

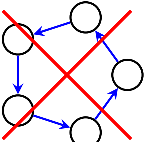


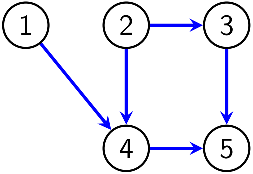
Directed Acyclic Graphs

vertices 

edges 

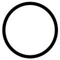
no directed cycles





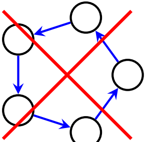
directed acyclic graph (DAG), $\mathcal{G}$

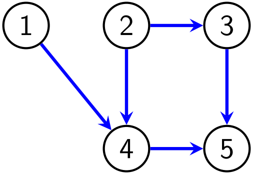
Directed Acyclic Graphs

vertices 

edges 

no directed cycles





directed acyclic graph (DAG), $\mathcal{G}$

If $w \to v$ then w is a **parent** of v: $\mathrm{pa}_{\mathcal{G}}(4) = \{1, 2\}$.

If $w \to \cdots \to v$ then w is a **ancestor** of v: $\mathrm{an}_{\mathcal{G}}(5) = \{1, 2, 3, 4, 5\}$.

An **ancestral set** contains all its own ancestors.

DAG Models

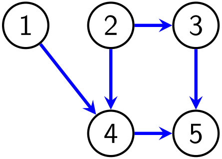
vertex		random variable
	$\iff$	X_a

DAG Models

vertex $\Longleftrightarrow$ random variable

X_a

graph $\mathcal{G}$ $\Longleftrightarrow$ model $\mathcal{M}$



$$\mathcal{M}(\mathcal{G}) = \{P \text{ satisfying } (*)\}$$

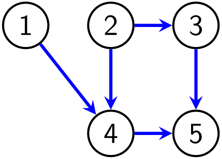
$$p(x_V) = \prod_{i \in V} p(x_i \mid x_{\mathrm{pa}(i)}). \qquad (*)$$

DAG Models

vertex $\iff$ random variable

X_a

graph $\mathcal{G}$ $\iff$ model $\mathcal{M}$



$\mathcal{M}(\mathcal{G}) = \{P \text{ satisfying } (*)\}$

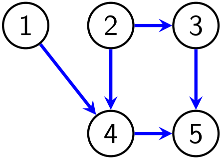
$$p(x_V) = \prod_{i \in V} p(x_i \mid x_{\text{pa}(i)}). \qquad (*)$$

So in example above:

$$p(x_V) = p(x_1) \cdot p(x_2) \cdot p(x_3 \mid x_2) \cdot p(x_4 \mid x_1, x_2) \cdot p(x_5 \mid x_3, x_4)$$

Algebraic Models

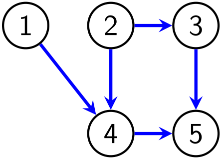
Can also define model as a list of conditional independences:



pick an topological ordering of the graph: $1, 2, 3, 4, 5$.

Algebraic Models

Can also define model as a list of conditional independences:



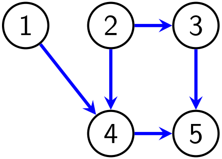
pick an topological ordering of the graph: $1, 2, 3, 4, 5$.

Can *always* factorize a joint distribution as:

$$p(x_V) = p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_1, x_2) \cdot p(x_4 \mid x_1, x_2, x_3) \cdot p(x_5 \mid x_1, x_2, x_3, x_4).$$

Algebraic Models

Can also define model as a list of conditional independences:



pick an topological ordering of the graph: $1, 2, 3, 4, 5$.

Can *always* factorize a joint distribution as:

$$p(x_V) = p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_1, x_2) \cdot p(x_4 \mid x_1, x_2, x_3) \cdot p(x_5 \mid x_1, x_2, x_3, x_4).$$

So by identifying this with $(*)$, see the model is the same as setting

$$p(x_i \mid x_1, x_2, \ldots, x_{i-1}) = p(x_i \mid x_{\text{pa}(i)}), \qquad \text{for each } i.$$

Algebraic Models

Thus $\mathcal{M}(\mathcal{G})$ is precisely distributions such that:

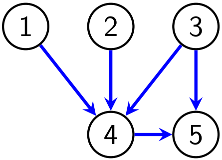
$$X_i \perp\!\!\!\perp X_{[i-1]\setminus \mathrm{pa}(i)} \,|\, X_{\mathrm{pa}(i)}, \qquad\qquad i \in V.$$

Algebraic Models

Thus $\mathcal{M}(\mathcal{G})$ is precisely distributions such that:

$$X_i \perp\!\!\!\perp X_{[i-1]\setminus \mathrm{pa}(i)} \,|\, X_{\mathrm{pa}(i)}, \qquad\qquad i \in V.$$

Example:



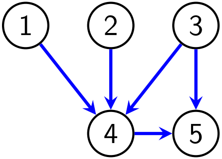
$$\begin{aligned} &X_2 \perp\!\!\!\perp X_1 \\ &X_3 \perp\!\!\!\perp X_1 \,|\, X_2 \\ &X_4 \perp\!\!\!\perp X_3 \,|\, X_1, X_2 \\ &X_5 \perp\!\!\!\perp X_1, X_2 \,|\, X_3, X_4. \end{aligned}$$

Algebraic Models

Thus $\mathcal{M}(\mathcal{G})$ is precisely distributions such that:

$$X_i \perp\!\!\!\perp X_{[i-1]\setminus \mathrm{pa}(i)} \mid X_{\mathrm{pa}(i)}, \qquad\qquad i \in V.$$

Example:



$$\begin{aligned} &X_2 \perp\!\!\!\perp X_1 \\ &X_3 \perp\!\!\!\perp X_1 \mid X_2 \\ &X_4 \perp\!\!\!\perp X_3 \mid X_1, X_2 \\ &X_5 \perp\!\!\!\perp X_1, X_2 \mid X_3, X_4. \end{aligned}$$

So for discrete variables this is an algebraic model.

Structural Equation Model View

There is a second way to think about DAG models.

A distribution $P \in \mathcal{M}(\mathcal{G})$ iff[a] there exist functions f_i and independent variables E_i such that recursively setting

$$X_i = f_i(X_{\text{pa}(i)}, E_i)$$

gives X_V the distribution P.

[a]This only makes sense if P has a density.

Structural Equation Model View

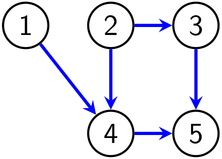
There is a second way to think about DAG models.

A distribution $P \in \mathcal{M}(\mathcal{G})$ iff[a] there exist functions f_i and independent variables E_i such that recursively setting

$$X_i = f_i(X_{\mathrm{pa}(i)}, E_i)$$

gives X_V the distribution P.

[a] This only makes sense if P has a density.



$$\begin{aligned}
X_1 &= f_1(E_1) \\
X_2 &= f_2(E_2) \\
X_3 &= f_3(X_2, E_3) \\
X_4 &= f_4(X_1, X_2, E_4) \\
X_5 &= f_5(X_3, X_4, E_5).
\end{aligned}$$

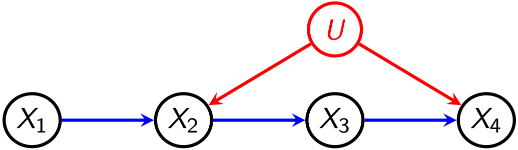
Reasons to Like DAG Models

- Induced constraints are all conditional independences: (reasonably) intuitive and simple to interpret;
- causal interpretation;
- modular structure is useful computationally and statistically;
- curved exponential families, known dimension;
- **algebraic model** for discrete variables.

Outline

Marginalization

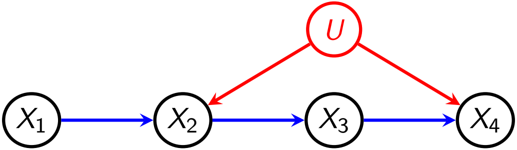
Sometimes we cannot observe all the variables. Consider:



with U unobserved.

Marginalization

Sometimes we cannot observe all the variables. Consider:



with U unobserved. This is a model defined (implicitly) by an integral:

$$p(x_1, x_2, x_3, x_4) = \int p(u)\, p(x_1)\, p(x_2 \mid x_1, u)\, p(x_3 \mid x_2)\, p(x_4 \mid x_3, u)\, du$$

We do **not** assume U is discrete, since we cannot observe it.

Marginalization

What we consider is **not** a latent variable model in the usual sense. **No state-space is assumed** for hidden variables (though uniform on $(0,1)$ is sufficient).

$$p(x_1, x_2, x_3, x_4) = \int p(u)\, p(x_1)\, p(x_2 \mid x_1, u)\, p(x_3 \mid x_2)\, p(x_4 \mid x_3, u)\, du$$

Marginalization

What we consider is **not** a latent variable model in the usual sense. **No state-space is assumed** for hidden variables (though uniform on $(0, 1)$ is sufficient).

$$p(x_1, x_2, x_3, x_4) = \int p(u)\, p(x_1)\, p(x_2 \mid x_1, u)\, p(x_3 \mid x_2)\, p(x_4 \mid x_3, u)\, du$$

But:

- cannot directly test membership of the model;
- model is complicated (as we shall see);
- not even clear it is a (semi-)algebraic model.

Marginalization

What we consider is **not** a latent variable model in the usual sense. **No state-space is assumed** for hidden variables (though uniform on $(0, 1)$ is sufficient).

$$p(x_1, x_2, x_3, x_4) = \int p(u)\, p(x_1)\, p(x_2 \mid x_1, u)\, p(x_3 \mid x_2)\, p(x_4 \mid x_3, u)\, du$$

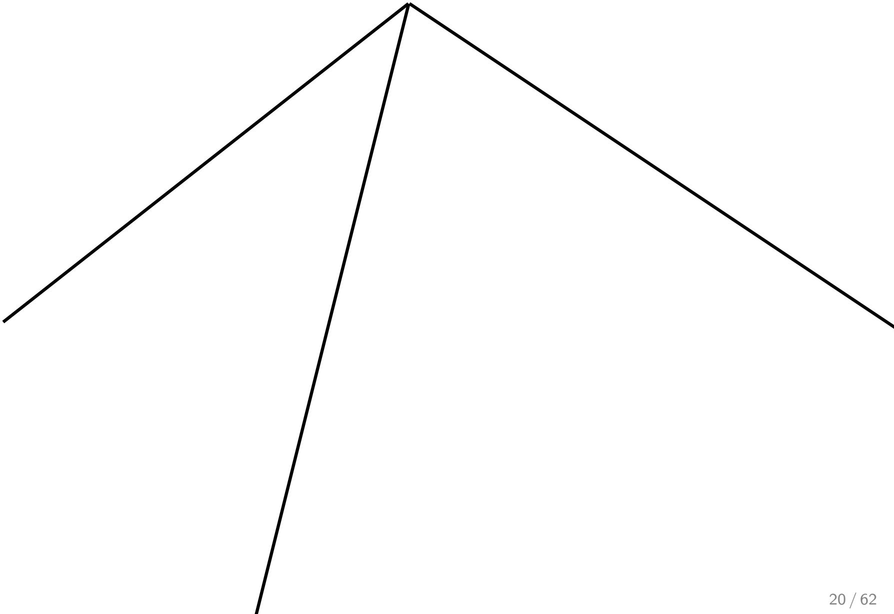
But:

- cannot directly test membership of the model;
- model is complicated (as we shall see);
- not even clear it is a (semi-)algebraic model.

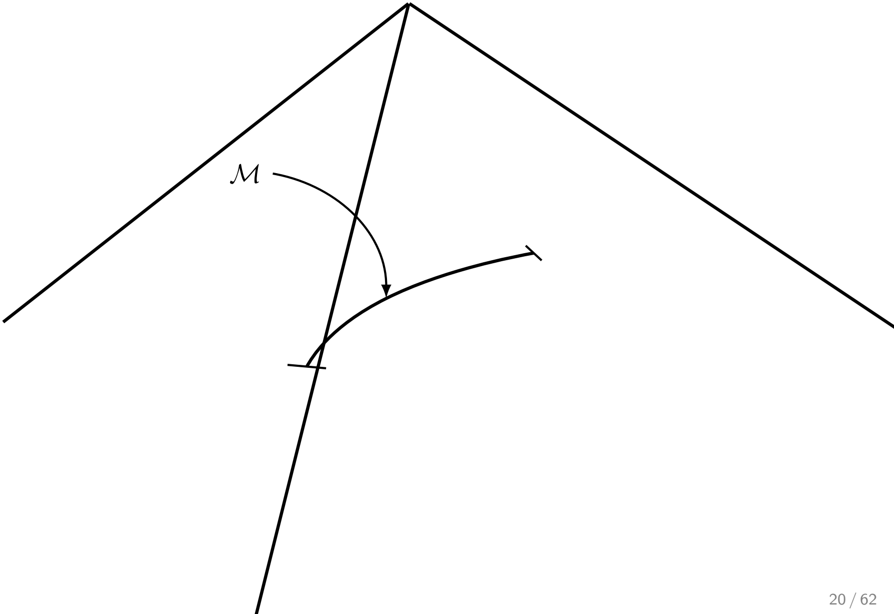
We aim to study the set of distributions constructed in this way.

Strategy: find some constraints satisfied by these models, define a new larger model using these constraints, and study that.

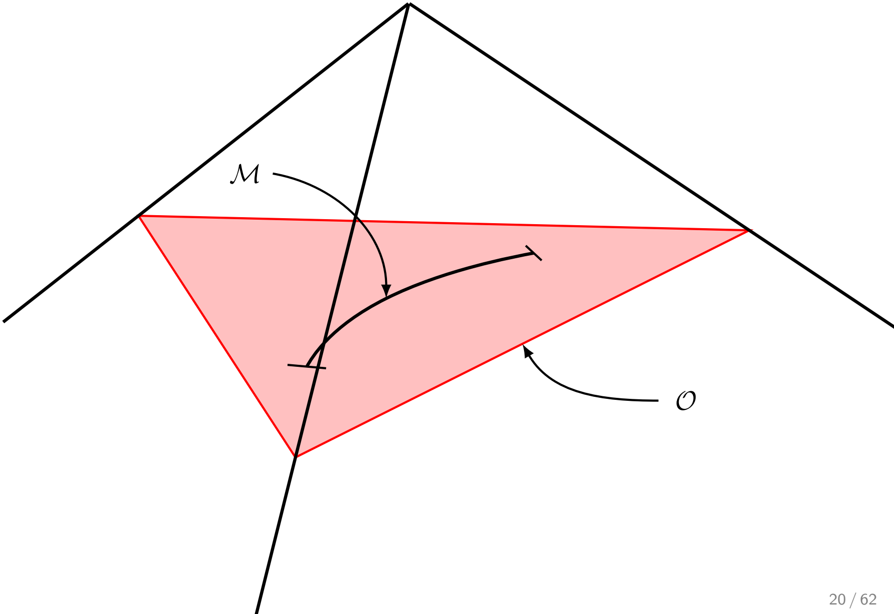
Getting the Picture



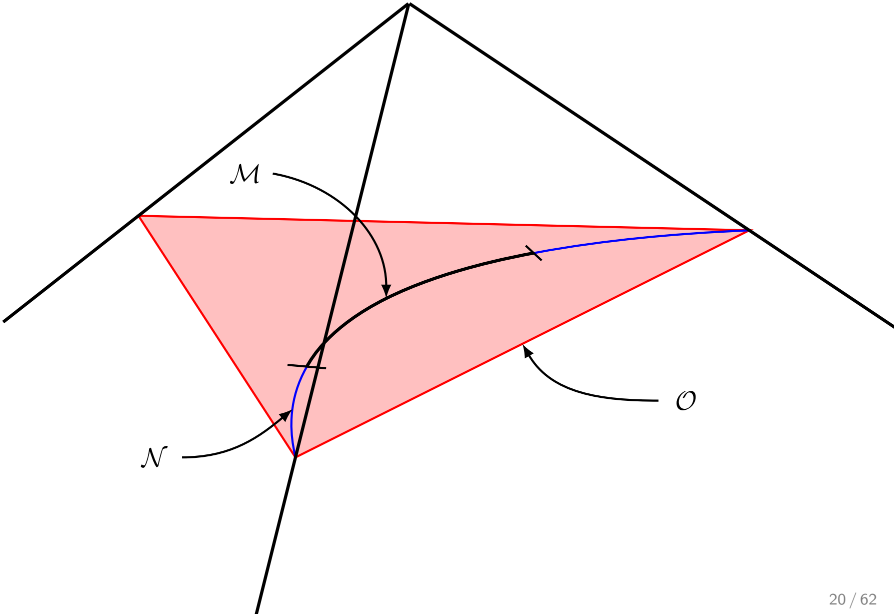
Getting the Picture



Getting the Picture



Getting the Picture



Latent Variable Models

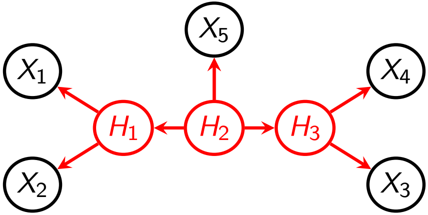
Traditional latent variable models would assume that the hidden variables are discrete with some fixed number of states.

Advantages: semi-algebraic model after eliminating variables is semi-algebraic, and can fit with (e.g.) EM algorithm.

Latent Variable Models

Traditional latent variable models would assume that the hidden variables are discrete with some fixed number of states.

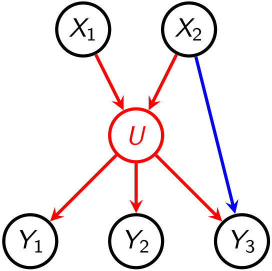
Advantages: semi-algebraic model after eliminating variables is semi-algebraic, and can fit with (e.g.) EM algorithm.



But: latent variables lead to singularities and nasty statistical properties (see e.g. Drton, Sturmfels and Sullivant, 2009)

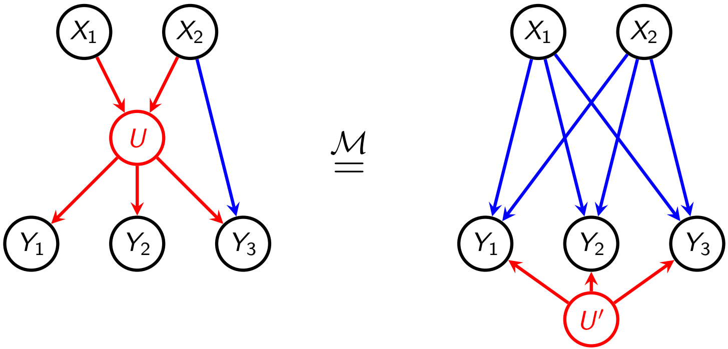
Simplifications

Simplification 1. WLOG latents vertices have no parents.



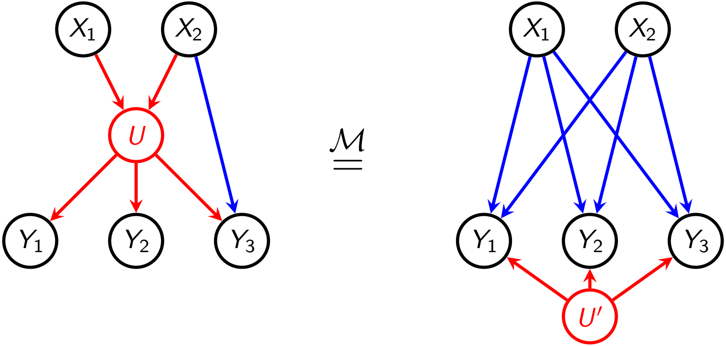
Simplifications

Simplification 1. WLOG latents vertices have no parents.



Simplifications

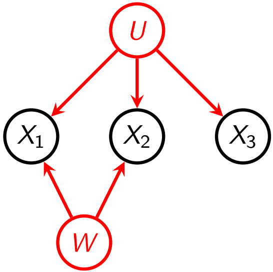
Simplification 1. WLOG latents vertices have no parents.



(Of course, this is not true if we assume a specific state-space: e.g. phylogenetic model)

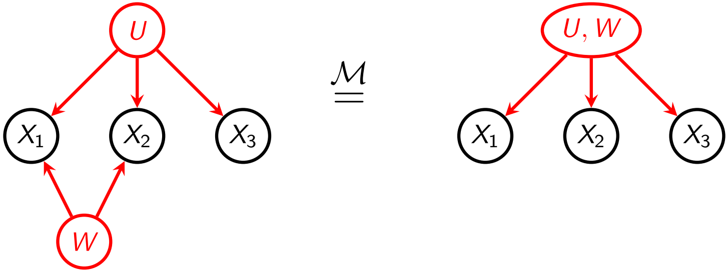
Simplifications

Simplification 2. If U, W are latent with $\mathrm{ch}_{\mathcal{G}}(W) \subseteq \mathrm{ch}_{\mathcal{G}}(U)$, then we don't need W.



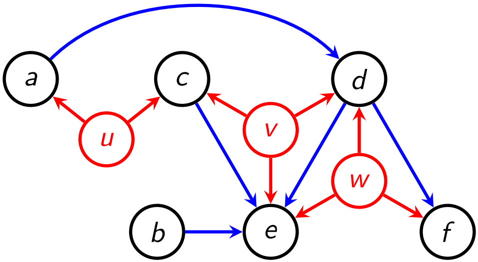
Simplifications

Simplification 2. If U, W are latent with $\mathrm{ch}_{\mathcal{G}}(W) \subseteq \mathrm{ch}_{\mathcal{G}}(U)$, then we don't need W.



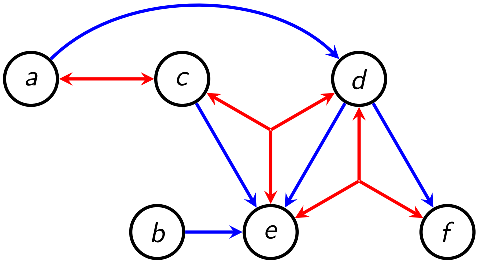
mDAGs

So we only need to consider models like this:



mDAGs

So we only need to consider models like this:

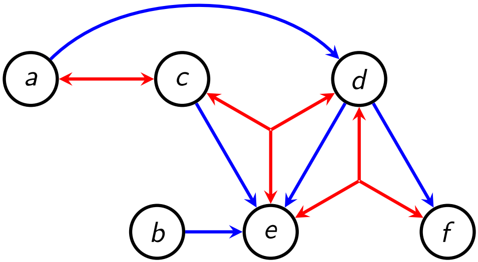


...which we represent with a hyper-graph called an **mDAG**.

The red edges $\longleftrightarrow$ are called **bidirected**.

mDAGs

So we only need to consider models like this:



...which we represent with a hyper-graph called an **mDAG**.

The red edges $\leftrightarrow$ are called **bidirected**.

We want the set of distributions that can be obtained by the latent variable; this is the **complete model** $\mathcal{M}(\mathcal{G})$ for mDAG $\mathcal{G}$.

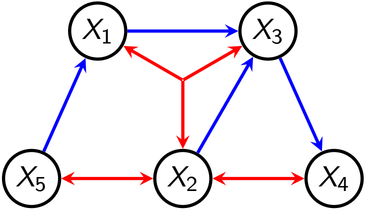
Geared Graphs

Call an mDAG **geared** if its bidirected edges satisfy the running intersection property.

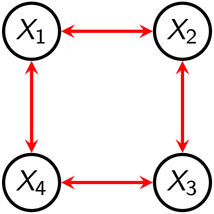
Geared Graphs

Call an mDAG **geared** if its bidirected edges satisfy the running intersection property. Examples:

geared

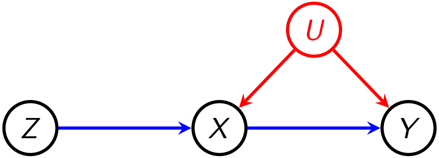


not geared



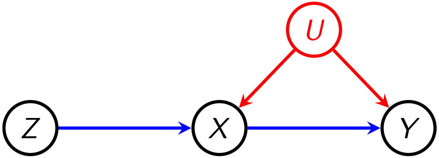
Functional Dependences

Consider the situation below.



Functional Dependences

Consider the situation below.



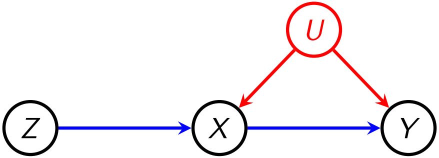
Recall the structural equation view: for some 'error' variables E_x, E_y:

$$X = f_X(Z, U, E_x) \qquad Y = f_Y(X, U, E_y).$$

Without loss of generality, can assume $U' = (U, E_x, E_y)$, so all additional randomness is contained in U'.

Functional Dependences

Consider the situation below.



Recall the structural equation view: for some 'error' variables E_x, E_y:

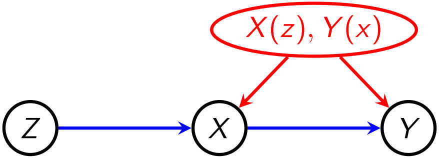
$$X = f_X(Z, U, E_x) \qquad\qquad Y = f_Y(X, U, E_y).$$

Without loss of generality, can assume $U' = (U, E_x, E_y)$, so all additional randomness is contained in U'.

U' 'tells' X and Y what to do given their other parents.

Functional Dependences

Consider the situation below.



Recall the structural equation view: for some 'error' variables E_x, E_y:

$$X = f_X(Z, U, E_x) \qquad\qquad Y = f_Y(X, U, E_y).$$

Without loss of generality, can assume $U' = (U, E_x, E_y)$, so all additional randomness is contained in U'.

U' 'tells' X and Y what to do given their other parents.

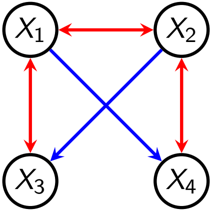
Set $U = (X(z), Y(x))$, drawn from **finite** set of functions.

Geared Graphs

If a graph is geared we can iterate this process to show that a finite state-space is sufficient:

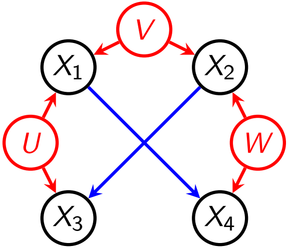
Geared Graphs

If a graph is geared we can iterate this process to show that a finite state-space is sufficient:



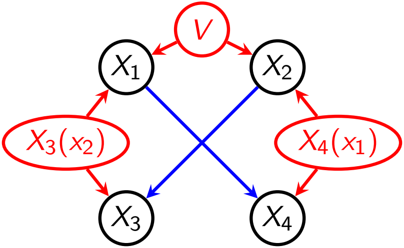
Geared Graphs

If a graph is geared we can iterate this process to show that a finite state-space is sufficient:



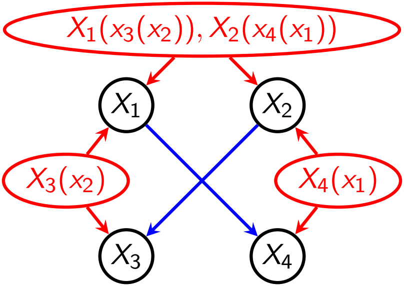
Geared Graphs

If a graph is geared we can iterate this process to show that a finite state-space is sufficient:



Geared Graphs

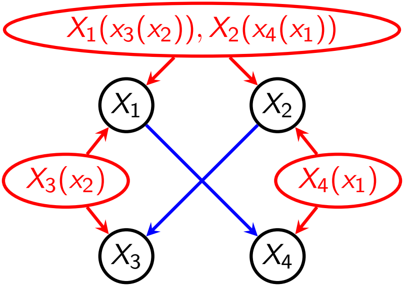
If a graph is geared we can iterate this process to show that a finite state-space is sufficient:



This shows that geared graphs do represent semi-algebraic models.

Geared Graphs

If a graph is geared we can iterate this process to show that a finite state-space is sufficient:

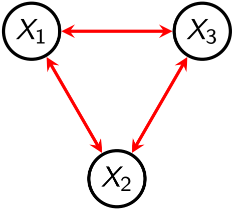


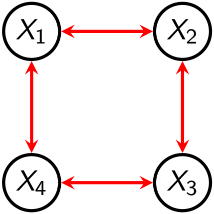
This shows that geared graphs do represent semi-algebraic models.

This representation turns out to be important in proving completeness of constraints.

Non-Geared Graphs

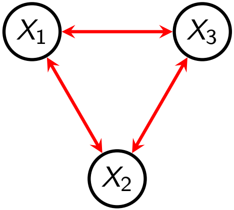
With a graph which is not geared, we cannot do this.

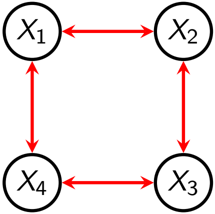




Non-Geared Graphs

With a graph which is not geared, we cannot do this.



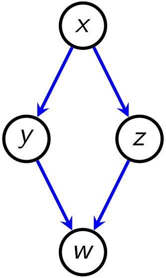


Open Problem: These models may or may not be semi-algebraic.

Outline

Ancestral Sets

Recall an **ancestral set** contains its own ancestors, e.g. $\{x, y, z\}$.

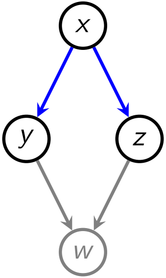


Marginalize w:

$$p(x, y, z) = \sum_{\mathbf{w}} p(x)\, p(y \mid x)\, p(z \mid x)\, p(\mathbf{w} \mid y, z)$$

Ancestral Sets

Recall an **ancestral set** contains its own ancestors, e.g. $\{x, y, z\}$.



Marginalize w:

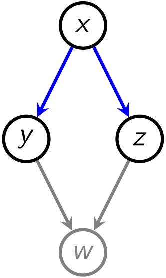
$$p(x, y, z) = \sum_{\mathbf{w}} p(x)\, p(y \mid x)\, p(z \mid x)\, p(\mathbf{w} \mid y, z)$$

$$= \quad p(x)\, p(y \mid x)\, p(z \mid x)$$

Obeys graphical model with w removed.

Ancestral Sets

Recall an **ancestral set** contains its own ancestors, e.g. $\{x, y, z\}$.



Marginalize w:

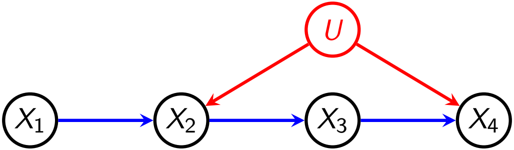
$$p(x, y, z) = \sum_{\mathbf{w}} p(x)\, p(y \mid x)\, p(z \mid x)\, p(\mathbf{w} \mid y, z)$$

$$= \quad p(x)\, p(y \mid x)\, p(z \mid x)$$

Obeys graphical model with w removed.

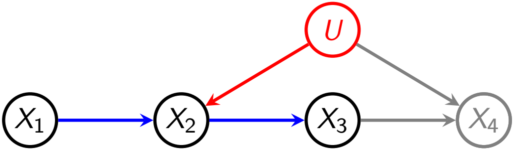
Models 'closed' under marginalization of vertices with no children.

Ancestral Sets



$$p(x_1, x_2, x_3, x_4)$$
$$= \int p(u)\, p(x_1)\, p(x_2 \mid x_1, u)\, p(x_3 \mid x_2)\, p(x_4 \mid x_3, u)\, du$$

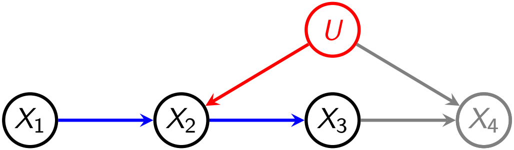
Ancestral Sets



$$p(x_1, x_2, x_3)$$

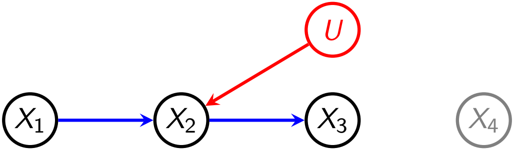
$$= \sum_{\mathbf{x_4}} \int p(u)\, p(x_1)\, p(x_2 \mid x_1, u)\, p(x_3 \mid x_2)\, p(\mathbf{x_4} \mid x_3, u)\, du$$

Ancestral Sets



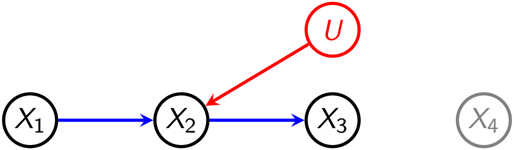
$$p(x_1, x_2, x_3)$$
$$= \sum_{\mathbf{x_4}} \int p(u)\, p(x_1)\, p(x_2 \mid x_1, u)\, p(x_3 \mid x_2)\, p(\mathbf{x_4} \mid x_3, u)\, du$$
$$= \int p(u)\, p(x_1)\, p(x_2 \mid x_1, u)\, p(x_3 \mid x_2) \sum_{\mathbf{x_4}} p(\mathbf{x_4} \mid x_3, u)\, du$$

Ancestral Sets



$$
\begin{aligned}
& p(x_1, x_2, x_3) \\
&= \sum_{\mathbf{x_4}} \int p(u)\, p(x_1)\, p(x_2 \mid x_1, u)\, p(x_3 \mid x_2)\, p(\mathbf{x_4} \mid x_3, u)\, du \\
&= \int p(u)\, p(x_1)\, p(x_2 \mid x_1, u)\, p(x_3 \mid x_2) \sum_{\mathbf{x_4}} p(\mathbf{x_4} \mid x_3, u)\, du \\
&= \int p(\mathbf{u})\, p(x_1)\, p(x_2 \mid x_1, \mathbf{u})\, p(x_3 \mid x_2)\, d\mathbf{u}
\end{aligned}
$$

Ancestral Sets

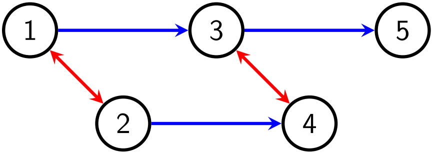


$$
\begin{aligned}
& p(x_1, x_2, x_3) \\
&= \sum_{\mathbf{x_4}} \int p(u)\, p(x_1)\, p(x_2 \mid x_1, u)\, p(x_3 \mid x_2)\, p(\mathbf{x_4} \mid x_3, u)\, du \\
&= \int p(u)\, p(x_1)\, p(x_2 \mid x_1, u)\, p(x_3 \mid x_2) \sum_{\mathbf{x_4}} p(\mathbf{x_4} \mid x_3, u)\, du \\
&= \int p(\mathbf{u})\, p(x_1)\, p(x_2 \mid x_1, \mathbf{u})\, p(x_3 \mid x_2)\, d\mathbf{u} \\
&= p(x_1)\, p(x_3 \mid x_2) \int p(\mathbf{u})\, p(x_2 \mid x_1, \mathbf{u})\, d\mathbf{u}
\end{aligned}
$$

gives $X_1 \perp\!\!\!\perp X_3 \mid X_2$.

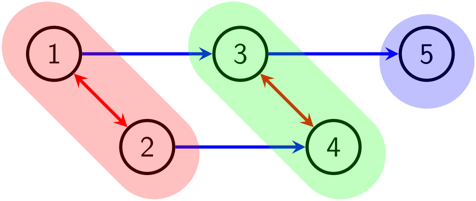
Districts

Define a **district** in an mDAG to be maximal sets connected by latent variables:



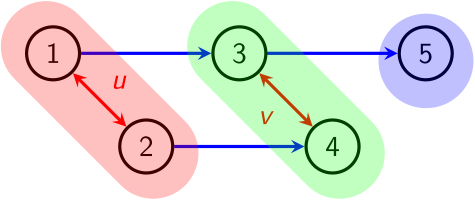
Districts

Define a **district** in an mDAG to be maximal sets connected by latent variables:



Districts

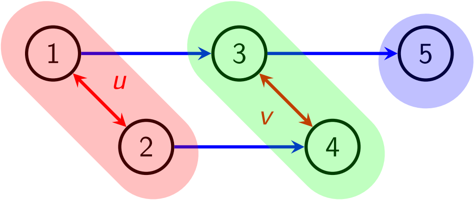
Define a **district** in an mDAG to be maximal sets connected by latent variables:



$$\int p(u)\, p(x_1 \mid u)\, p(x_2 \mid u) \quad p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v) \quad p(x_5 \mid x_3)\; du\, dv$$

Districts

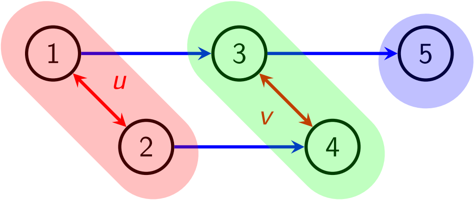
Define a **district** in an mDAG to be maximal sets connected by latent variables:



$$\int p(u)\, p(x_1 \mid u)\, p(x_2 \mid u)\; p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\; p(x_5 \mid x_3)\, du\, dv$$

Districts

Define a **district** in an mDAG to be maximal sets connected by latent variables:

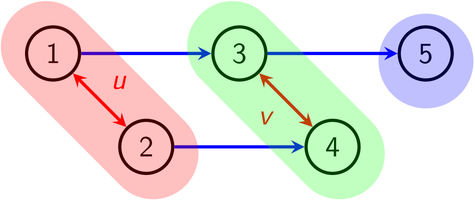


$$\int p(u)\, p(x_1 \mid u)\, p(x_2 \mid u)\; p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\; p(x_5 \mid x_3)\; du\, dv$$

$$= \int p(u)\, p(x_1 \mid u)\, p(x_2 \mid u)\, du \int p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\, dv\; p(x_5 \mid x_3)$$

Districts

Define a **district** in an mDAG to be maximal sets connected by latent variables:



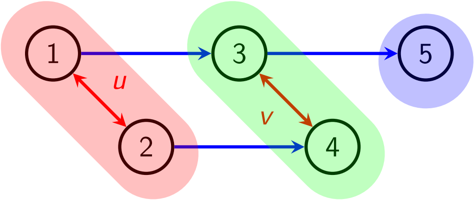
$$\int p(u)\, p(x_1 \mid u)\, p(x_2 \mid u)\; p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\; p(x_5 \mid x_3)\, du\, dv$$

$$= \int p(u)\, p(x_1 \mid u)\, p(x_2 \mid u)\, du \int p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\, dv\; p(x_5 \mid x_3)$$

$$= q(x_1, x_2) \cdot q(x_3, x_4 \mid x_1, x_2) \cdot q(x_5 \mid x_3).$$

Districts

Define a **district** in an mDAG to be maximal sets connected by latent variables:



$$\int p(u)\, p(x_1 \mid u)\, p(x_2 \mid u)\; p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\; p(x_5 \mid x_3)\, du\, dv$$

$$= \int p(u)\, p(x_1 \mid u)\, p(x_2 \mid u)\, du \int p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\, dv\; p(x_5 \mid x_3)$$

$$= q(x_1, x_2) \cdot q(x_3, x_4 \mid x_1, x_2) \cdot q(x_5 \mid x_3).$$

$$= \prod_i q_{D_i}(x_{D_i} \mid x_{\mathrm{pa}(D_i)})$$

Axiomatic Approach

Define $\mathcal{O}(\mathcal{G})$ as set of P satisfying:

1. **Ancestrality:** $P \in \mathcal{O}(\mathcal{G})$ only if

$$\sum_{x_w} p(x_V) \in \mathcal{O}(\mathcal{G}_{-w})$$

for each childless w.

Axiomatic Approach

Define $\mathcal{O}(\mathcal{G})$ as set of P satisfying:

1. **Ancestrality:** $P \in \mathcal{O}(\mathcal{G})$ only if

$$\sum_{x_w} p(x_V) \in \mathcal{O}(\mathcal{G}_{-w})$$

   for each childless w.

2. **Factorization into districts:** $P \in \mathcal{O}(\mathcal{G})$ only if

$$p(x_V) = \prod_D q_D(x_D \,|\, x_{\mathrm{pa}(D)})$$

   for districts D and some functions q_D.

Axiomatic Approach

Define $\mathcal{O}(\mathcal{G})$ as set of P satisfying:

1. **Ancestrality:** $P \in \mathcal{O}(\mathcal{G})$ only if

$$\sum_{x_w} p(x_V) \in \mathcal{O}(\mathcal{G}_{-w})$$

   for each childless w.

2. **Factorization into districts:** $P \in \mathcal{O}(\mathcal{G})$ only if

$$p(x_V) = \prod_D q_D(x_D \,|\, x_{\mathrm{pa}(D)})$$

   for districts D and some functions q_D.

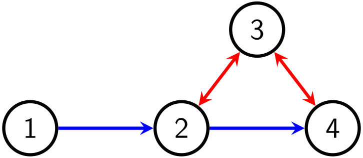
Call this the **ordinary Markov model** (OMM).

Properties of the OMM

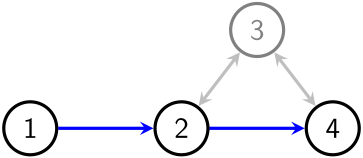
First described by Richardson (2003, 2009); factorization and parametrizations in Evans and Richardson (2013, 2014).

- Strict superset of latent variable model;
- equivalent to taking all the conditional independences from the original model which only involve 'visible' variables;
- therefore algebraic (quadratic constraints in the probabilities);
- has parametrization, so irreducible variety;
- curved exponential families.

Example

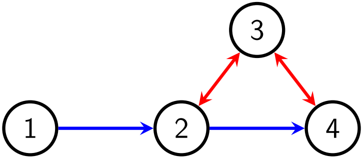


Example



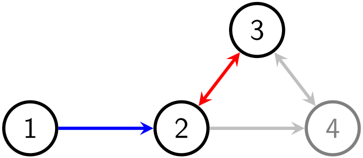
So $X_1 \perp\!\!\!\perp X_4 \mid X_2$

Example



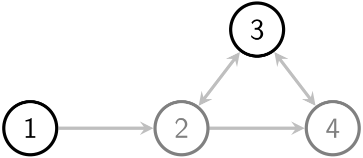
So $X_1 \perp\!\!\!\perp X_4 \mid X_2$

Example



So $X_1 \perp\!\!\!\perp X_4 \mid X_2$

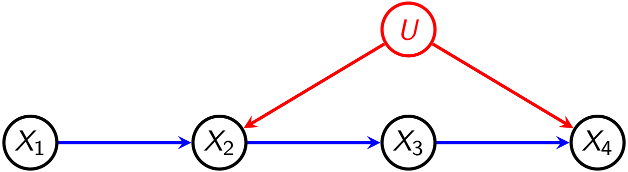
Example



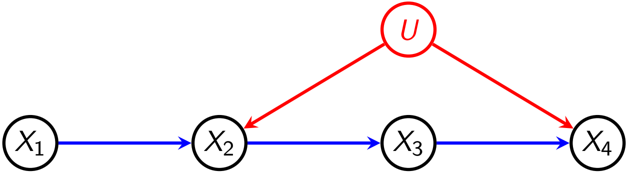
So $X_1 \perp\!\!\!\perp X_4 \mid X_2$ and $X_1 \perp\!\!\!\perp X_3$.

Outline

A Deficiency

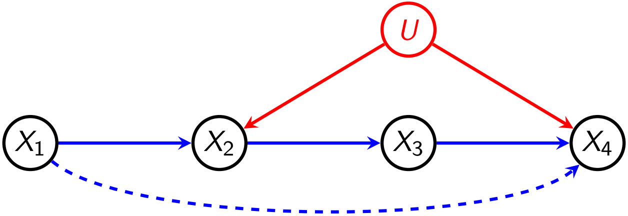


A Deficiency



If U is latent, OMM gives only $X_3 \perp\!\!\!\perp X_1 \mid X_2$.

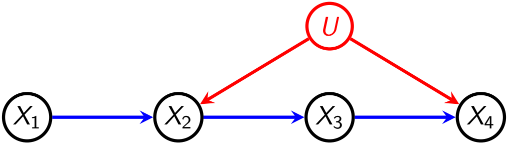
A Deficiency



If U is latent, OMM gives only $X_3 \perp\!\!\!\perp X_1 \mid X_2$.

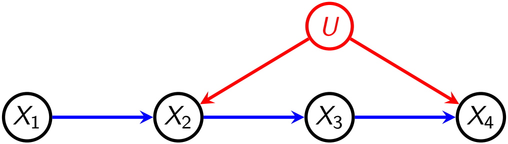
But if we add an arrow $X_1 \to X_4$, we still have $X_3 \perp\!\!\!\perp X_1 \mid X_2$.
So can we detect that $X_1 \not\to X_4$?

The Verma Constraint



$$p(x_1, x_2, x_3, x_4) = \int p(\mathbf{u})\, p(x_1)\, p(x_2 \mid x_1, \mathbf{u})\, p(x_3 \mid x_2)\, p(x_4 \mid x_3, \mathbf{u})\, d\mathbf{u}$$

The Verma Constraint



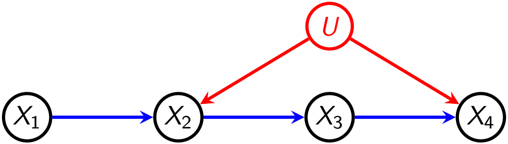
$$p(x_1, x_2, x_3, x_4) = \int p(\mathbf{u})\, p(x_1)\, p(x_2 \mid x_1, \mathbf{u})\, p(x_3 \mid x_2)\, p(x_4 \mid x_3, \mathbf{u})\, d\mathbf{u}$$

$$= p(x_1)\, p(x_3 \mid x_2) \int p(\mathbf{u})\, p(x_2 \mid x_1, \mathbf{u})\, p(x_4 \mid x_3, \mathbf{u})\, d\mathbf{u}$$

$$= p(x_1)\, p(x_3 \mid x_2)\, q(x_2, x_4 \mid x_1, x_3).$$

(This is our district factorization.)

The Verma Constraint



$$
\begin{aligned}
p(x_1, x_2, x_3, x_4) &= \int p(\mathbf{u})\, p(x_1)\, p(x_2 \mid x_1, \mathbf{u})\, p(x_3 \mid x_2)\, p(x_4 \mid x_3, \mathbf{u})\, d\mathbf{u} \\
&= p(x_1)\, p(x_3 \mid x_2) \int p(\mathbf{u})\, p(x_2 \mid x_1, \mathbf{u})\, p(x_4 \mid x_3, \mathbf{u})\, d\mathbf{u} \\
&= p(x_1)\, p(x_3 \mid x_2)\, q(x_2, x_4 \mid x_1, x_3).
\end{aligned}
$$

(This is our district factorization.) But note that

$$
\begin{aligned}
\sum_{x_2} q(x_2, x_4 \mid x_1, x_3) &= \sum_{\mathbf{x_2}} \int p(u)\, p(\mathbf{x_2} \mid x_1, u)\, p(x_4 \mid x_3, u)\, du \\
&= p(x_4 \mid x_3)
\end{aligned}
$$

is independent of x_1, precisely because $X_1 \not\to X_4$.

Verma Constraints are Polynomials

This is the **Verma constraint** (Pearl and Verma, 1990):

$$\sum_{x_2} \frac{p(x_1, x_2, x_3, x_4)p(x_2)}{p(x_1) \cdot p(x_2, x_3)} = \sum_{x_2} \frac{p(x_1', x_2, x_3, x_4)p(x_2)}{p(x_1') \cdot p(x_2, x_3)}$$

Verma Constraints are Polynomials

This is the **Verma constraint** (Pearl and Verma, 1990):

$$\sum_{x_2} \frac{p(x_1, x_2, x_3, x_4)p(x_2)}{p(x_1) \cdot p(x_2, x_3)} = \sum_{x_2} \frac{p(x_1', x_2, x_3, x_4)p(x_2)}{p(x_1') \cdot p(x_2, x_3)}$$

Gives degree-4 polynomial (662 terms) in binary case.
(if $X_3 \not\perp\!\!\!\perp X_1 \,|\, X_2$ get degree 6 polynomial with 480 terms)

Verma Constraints are Polynomials

This is the **Verma constraint** (Pearl and Verma, 1990):

$$\sum_{x_2} \frac{p(x_1, x_2, x_3, x_4) p(x_2)}{p(x_1) \cdot p(x_2, x_3)} = \sum_{x_2} \frac{p(x_1', x_2, x_3, x_4) p(x_2)}{p(x_1') \cdot p(x_2, x_3)}$$

Gives degree-4 polynomial (662 terms) in binary case.
(if $X_3 \not\perp\!\!\!\perp X_1 \mid X_2$ get degree 6 polynomial with 480 terms)

Note degree increases with number of states of X_1 and X_2.
Generally:

$$|\mathfrak{X}_1| + |\mathfrak{X}_2| \qquad \text{(or } |\mathfrak{X}_1|(1 + |\mathfrak{X}_2|)\text{)}$$

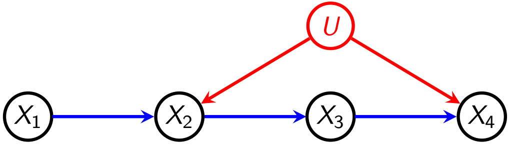
Reflects difficulty of estimating $p(x_1)$ and $p(x_3 \mid x_1, x_2)$ and dividing out by them(?)

Subgraphs

$q(x_2, x_4 \mid x_1, x_3)$ behaves as a density in which $X_1 \perp\!\!\!\perp X_4 \mid X_3$, though this does not hold under p.

Subgraphs

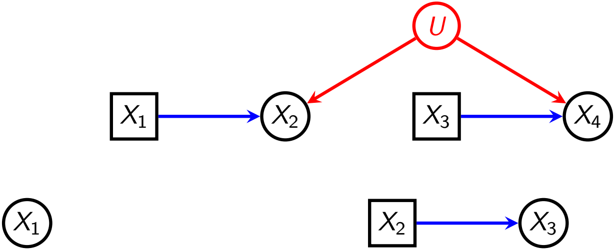
$q(x_2, x_4 \mid x_1, x_3)$ behaves as a density in which $X_1 \perp\!\!\!\perp X_4 \mid X_3$, though this does not hold under p.



$$p(x_1, x_2, x_3, x_4) = p(x_1)\, p(x_3 \mid x_2) \int p(u)\, p(x_2 \mid x_1, u)\, p(x_4 \mid x_3, u)\, du$$

Subgraphs

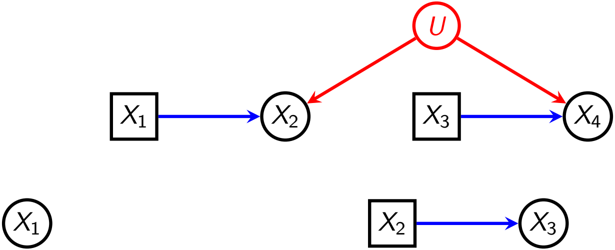
$q(x_2, x_4 \mid x_1, x_3)$ behaves as a density in which $X_1 \perp\!\!\!\perp X_4 \mid X_3$, though this does not hold under p.



$$p(x_1, x_2, x_3, x_4) = p(x_1)\, p(x_3 \mid x_2) \int p(u)\, p(x_2 \mid x_1, u)\, p(x_4 \mid x_3, u)\, du$$

Subgraphs

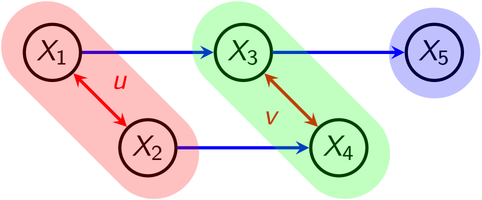
$q(x_2, x_4 \mid x_1, x_3)$ behaves as a density in which $X_1 \perp\!\!\!\perp X_4 \mid X_3$, though this does not hold under p.



$$\frac{p(x_1, x_2, x_3, x_4)}{p(x_1) \cdot p(x_3 \mid x_2)} = \int p(u)\, p(x_2 \mid x_1, u)\, p(x_4 \mid x_3, u)\, du$$

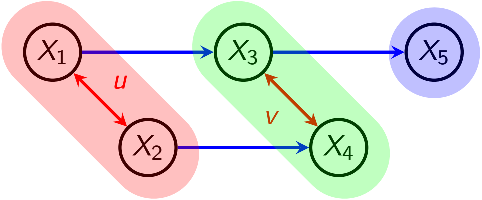
So each factor of the distribution q_D corresponds to a 'piece' of the graph $\mathcal{G}[D]$.

Districts



$$\int p(x_1 \mid u)\, p(x_2 \mid u)\; p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\; p(x_5 \mid x_3)\; du\, dv$$

Districts

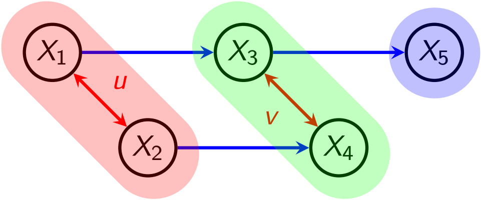


$$\int p(x_1 \mid u)\, p(x_2 \mid u)\; p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\; p(x_5 \mid x_3)\, du\, dv$$

$$= \int p(x_1 \mid u)\, p(x_2 \mid u)\, du \cdot \int p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\, dv \cdot p(x_5 \mid x_3)$$

$$= q(x_1, x_2) \cdot q(x_3, x_4 \mid x_1, x_2) \cdot q(x_5 \mid x_3).$$

Districts



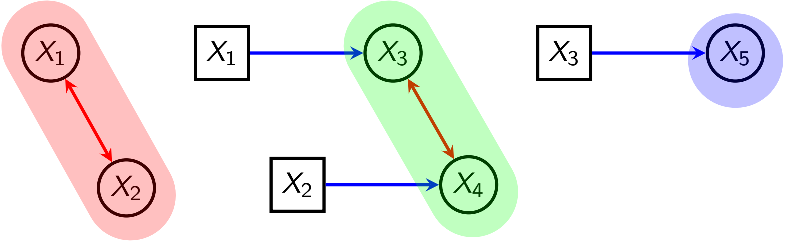
$$\int p(x_1 \mid u)\, p(x_2 \mid u)\; p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\; p(x_5 \mid x_3)\, du\, dv$$

$$= \int p(x_1 \mid u)\, p(x_2 \mid u)\, du \cdot \int p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\, dv \cdot p(x_5 \mid x_3)$$

$$= q(x_1, x_2) \cdot q(x_3, x_4 \mid x_1, x_2) \cdot q(x_5 \mid x_3).$$

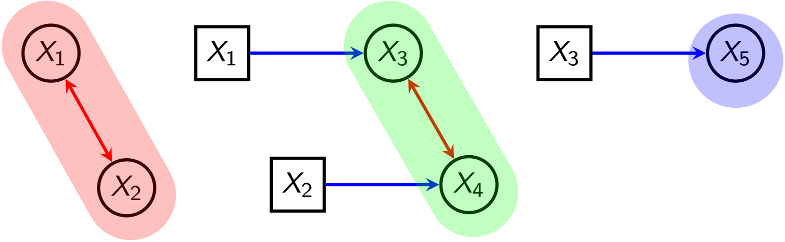
The form of each q is important.

Districts



$$\int p(u)\, p(x_1 \mid u)\, p(x_2 \mid u)\; p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\; p(x_5 \mid x_3)\, du\, dv$$

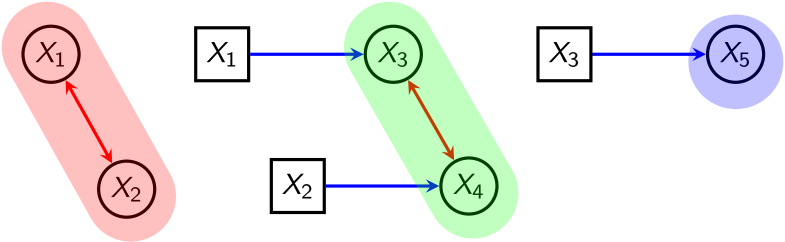
Districts



$$\int p(u)\, p(x_1 \mid u)\, p(x_2 \mid u)\;\; p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\;\; p(x_5 \mid x_3)\; du\, dv$$

$$= \int p(u)\, p(x_1 \mid u)\, p(x_2 \mid u)\; du \int p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\; dv\;\; p(x_5 \mid x_3)$$

Districts



$$\int p(u)\, p(x_1 \mid u)\, p(x_2 \mid u) \;\; p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v) \;\; p(x_5 \mid x_3)\; du\, dv$$

$$= \int p(u)\, p(x_1 \mid u)\, p(x_2 \mid u)\, du \int p(v)\, p(x_3 \mid x_1, v)\, p(x_4 \mid x_2, v)\, dv \;\; p(x_5 \mid x_3)$$

$$= q(x_1, x_2) \cdot q(x_3, x_4 \mid x_1, x_2) \cdot q(x_5 \mid x_3).$$

$$= \prod_i q_{D_i}(x_{D_i} \mid x_{\mathrm{pa}(D_i)})$$

Each q_D piece should come from the model based on district subgraph and its parents ($\mathcal{G}[D]$).

Axiomatic Approach II

Define $\mathcal{N}(\mathcal{G})$ as a model satisfying:

1. **Ancestrality** $P \in \mathcal{N}(\mathcal{G})$ only if

$$\sum_{x_w} p(x_V) \in \mathcal{N}(\mathcal{G}_{-w})$$

   for each childless w.

Axiomatic Approach II

Define $\mathcal{N}(\mathcal{G})$ as a model satisfying:

1. **Ancestrality** $P \in \mathcal{N}(\mathcal{G})$ only if

$$\sum_{x_w} p(x_V) \in \mathcal{N}(\mathcal{G}_{-w})$$

   for each childless w.

2. **Factorization into districts** $P \in \mathcal{N}(\mathcal{G})$ only if

$$p(x_V) = \prod_D q_D(x_D \,|\, x_{\mathrm{pa}(D)})$$

   for districts D, where $q_D \in \mathcal{N}(\mathcal{G}[D])$.

Axiomatic Approach II

Define $\mathcal{N}(\mathcal{G})$ as a model satisfying:

1. **Ancestrality** $P \in \mathcal{N}(\mathcal{G})$ only if

$$\sum_{x_w} p(x_V) \in \mathcal{N}(\mathcal{G}_{-w})$$

   for each childless w.

2. **Factorization into districts** $P \in \mathcal{N}(\mathcal{G})$ only if

$$p(x_V) = \prod_D q_D(x_D \mid x_{\mathrm{pa}(D)})$$

   for districts D, where $q_D \in \mathcal{N}(\mathcal{G}[D])$.

Note that one can iterate between 1 and 2.

Axiomatic Approach II

Define $\mathcal{N}(\mathcal{G})$ as a model satisfying:

1. **Ancestrality** $P \in \mathcal{N}(\mathcal{G})$ only if

$$\sum_{x_w} p(x_V) \in \mathcal{N}(\mathcal{G}_{-w})$$

   for each childless w.

2. **Factorization into districts** $P \in \mathcal{N}(\mathcal{G})$ only if

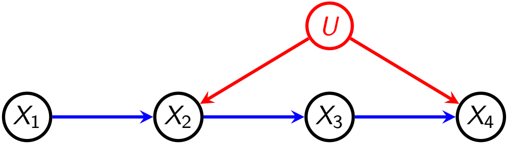
$$p(x_V) = \prod_{D} q_D(x_D \mid x_{\text{pa}(D)})$$

   for districts D, where $q_D \in \mathcal{N}(\mathcal{G}[D])$.

Note that one can iterate between 1 and 2.

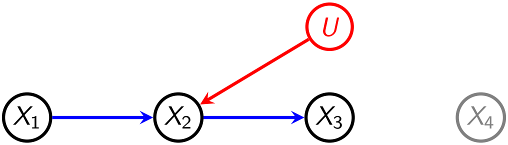
Call this the **nested Markov model** (NMM).

Verma Example



X_4 childless,

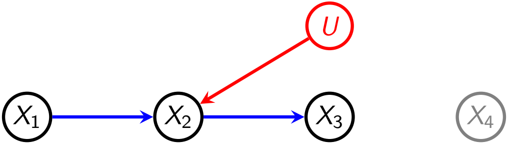
Verma Example



X_4 childless, so if $P \in \mathcal{N}(\mathcal{G})$, then

$$p(x_1, x_2, x_3) = p(x_1) \cdot \left(\int p(u) \cdot p(x_2 \mid x_1, u)\, du \right) \cdot p(x_3 \mid x_2)$$

Verma Example

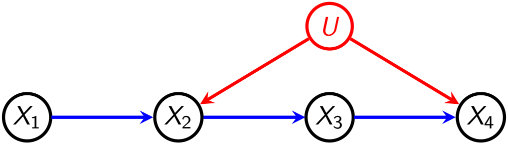


X_4 childless, so if $P \in \mathcal{N}(\mathcal{G})$, then

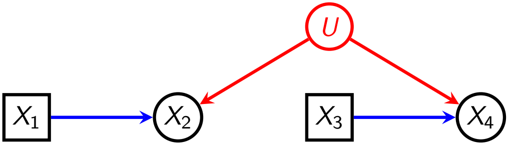
$$p(x_1, x_2, x_3) = p(x_1) \cdot \left(\int p(u) \cdot p(x_2 \,|\, x_1, u)\, du \right) \cdot p(x_3 \,|\, x_2)$$
$$= p(x_1) \cdot p(x_2 \,|\, x_1) \cdot p(x_3 \,|\, x_2),$$

and therefore $X_1 \perp\!\!\!\perp X_3 \,|\, X_2$.

Verma Example

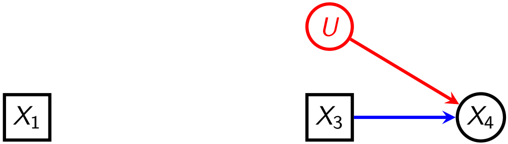


Verma Example



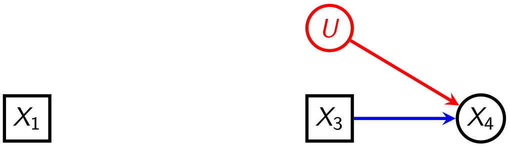
Can consider the district $\{2, 4\}$ and distribution q_{24}...

Verma Example



Can consider the district $\{2, 4\}$ and distribution q_{24}...
and then marginalize X_2.

Verma Example



Can consider the district $\{2, 4\}$ and distribution q_{24}...
and then marginalize X_2.

We see that $X_1 \perp\!\!\!\perp X_3, X_4\,[q_{24}]$.

Properties of the Nested Markov Model

- $$\mathcal{M}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G}) \subseteq \mathcal{O}(\mathcal{G}),$$
  i.e. the constraints given by nested Markov property are 'correct';

Properties of the Nested Markov Model

- $$\mathcal{M}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G}) \subseteq \mathcal{O}(\mathcal{G}),$$

  i.e. the constraints given by nested Markov property are 'correct';
- in general $\mathcal{M}(\mathcal{G}) \subsetneq \mathcal{N}(\mathcal{G})$, because of inequality constraints;

Properties of the Nested Markov Model

- $$\mathcal{M}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G}) \subseteq \mathcal{O}(\mathcal{G}),$$

  i.e. the constraints given by nested Markov property are 'correct';
- in general $\mathcal{M}(\mathcal{G}) \subsetneq \mathcal{N}(\mathcal{G})$, because of inequality constraints;
- constraints are generalization of conditional independence;

Properties of the Nested Markov Model

- $$\mathcal{M}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G}) \subseteq \mathcal{O}(\mathcal{G}),$$

  i.e. the constraints given by nested Markov property are 'correct';
- in general $\mathcal{M}(\mathcal{G}) \subsetneq \mathcal{N}(\mathcal{G})$, because of inequality constraints;
- constraints are generalization of conditional independence;
- curved exponential families (discrete case).

Properties of the Nested Markov Model

- $$\mathcal{M}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G}) \subseteq \mathcal{O}(\mathcal{G}),$$

  i.e. the constraints given by nested Markov property are 'correct';
- in general $\mathcal{M}(\mathcal{G}) \subsetneq \mathcal{N}(\mathcal{G})$, because of inequality constraints;
- constraints are generalization of conditional independence;
- curved exponential families (discrete case).

Properties of the Nested Markov Model

- $$\mathcal{M}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G}) \subseteq \mathcal{O}(\mathcal{G}),$$
  i.e. the constraints given by nested Markov property are 'correct';
- in general $\mathcal{M}(\mathcal{G}) \subsetneq \mathcal{N}(\mathcal{G})$, because of inequality constraints;
- constraints are generalization of conditional independence;
- curved exponential families (discrete case).

Theory of nested Markov model is well developed:

- global, local, factorization and moralization based Markov properties;

Properties of the Nested Markov Model

- $$\mathcal{M}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G}) \subseteq \mathcal{O}(\mathcal{G}),$$
  i.e. the constraints given by nested Markov property are 'correct';
- in general $\mathcal{M}(\mathcal{G}) \subsetneq \mathcal{N}(\mathcal{G})$, because of inequality constraints;
- constraints are generalization of conditional independence;
- curved exponential families (discrete case).

Theory of nested Markov model is well developed:

- global, local, factorization and moralization based Markov properties;
- parametrization in discrete case (Shpitser et al, 2012);

Properties of the Nested Markov Model

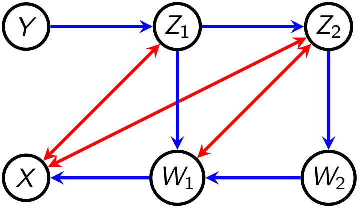
- $$\mathcal{M}(\mathcal{G}) \subseteq \mathcal{N}(\mathcal{G}) \subseteq \mathcal{O}(\mathcal{G}),$$
  i.e. the constraints given by nested Markov property are 'correct';
- in general $\mathcal{M}(\mathcal{G}) \subsetneq \mathcal{N}(\mathcal{G})$, because of inequality constraints;
- constraints are generalization of conditional independence;
- curved exponential families (discrete case).

Theory of nested Markov model is well developed:

- global, local, factorization and moralization based Markov properties;
- parametrization in discrete case (Shpitser et al, 2012);
- fitting and search methods (Shpitser et al, 2013).

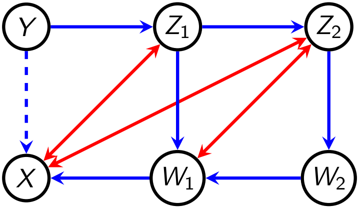
Example

In the below example, X and Y are not adjacent: is there a constraint implied?



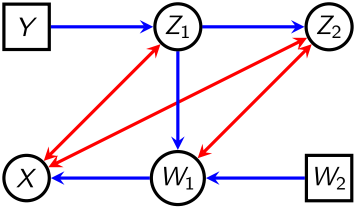
Example

In the below example, X and Y are not adjacent: is there a constraint implied?



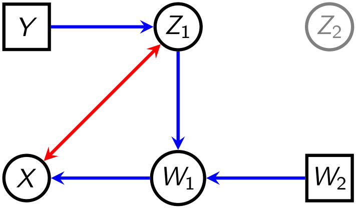
Example

In the below example, X and Y are not adjacent: is there a constraint implied?



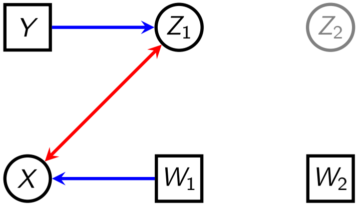
Example

In the below example, X and Y are not adjacent: is there a constraint implied?



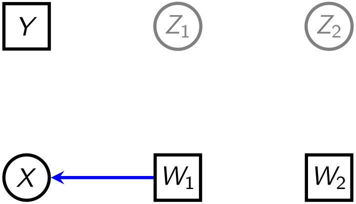
Example

In the below example, X and Y are not adjacent: is there a constraint implied?



Example

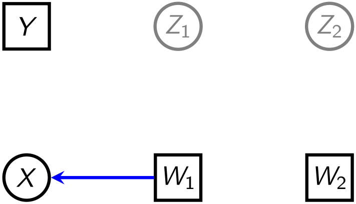
In the below example, X and Y are not adjacent: is there a constraint implied?



So $X \perp\!\!\!\perp Y \mid W_1$ in a twice re-weighted distribution P^{**}.

Example

In the below example, X and Y are not adjacent: is there a constraint implied?



So $X \perp\!\!\!\perp Y \mid W_1$ in a twice re-weighted distribution P^{**}.

So can distinguish between these two structures...
...but this is a degree-12 polynomial!

Outline

Main Result

How do we know there isn't **another** 'axiom' we could use?

Main Result

How do we know there isn't **another** 'axiom' we could use?

Theorem (Evans)

For any discrete DAG model, the nested and complete Markov models are algebraically equivalent (i.e. same dimension):

$$\overline{\mathcal{N}(\mathcal{G})} = \overline{\mathcal{M}(\mathcal{G})}.$$

where $\bar{\mathcal{S}}$ is the Zariski closure of $\mathcal{S}$.

Main Result

How do we know there isn't **another** 'axiom' we could use?

Theorem (Evans)

For any discrete DAG model, the nested and complete Markov models are algebraically equivalent (i.e. same dimension):

$$\overline{\mathcal{N}(\mathcal{G})} = \overline{\mathcal{M}(\mathcal{G})}.$$

where $\bar{\mathcal{S}}$ is the Zariski closure of $\mathcal{S}$.

In addition:

Theorem (Evans and Richardson)

Nested models are curved exponential families.

Main Result

How do we know there isn't **another** 'axiom' we could use?

Theorem (Evans)

For any discrete DAG model, the nested and complete Markov models are algebraically equivalent (i.e. same dimension):

$$\overline{\mathcal{N}(\mathcal{G})} = \overline{\mathcal{M}(\mathcal{G})}.$$

where $\bar{\mathcal{S}}$ is the Zariski closure of $\mathcal{S}$.

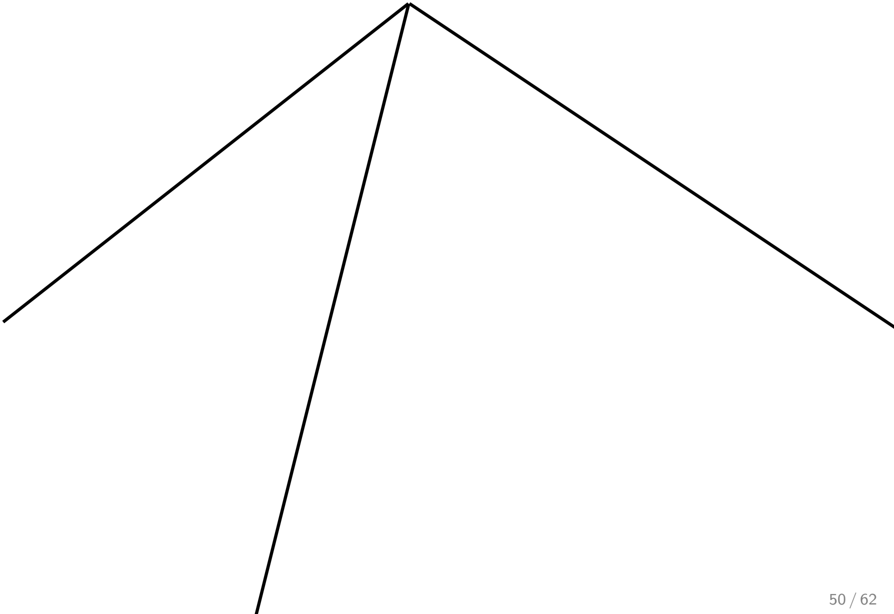
In addition:

Theorem (Evans and Richardson)

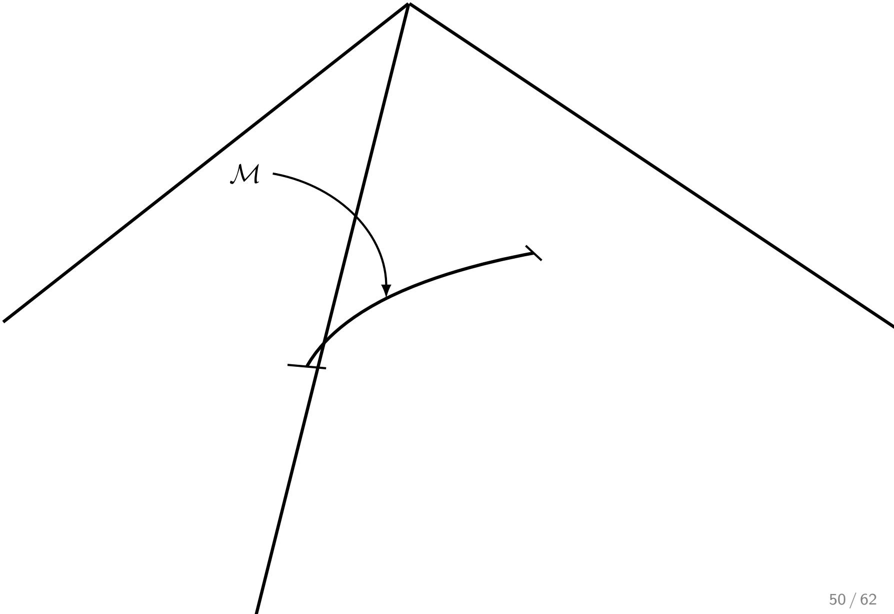
Nested models are curved exponential families.

This has very nice statistical implications.

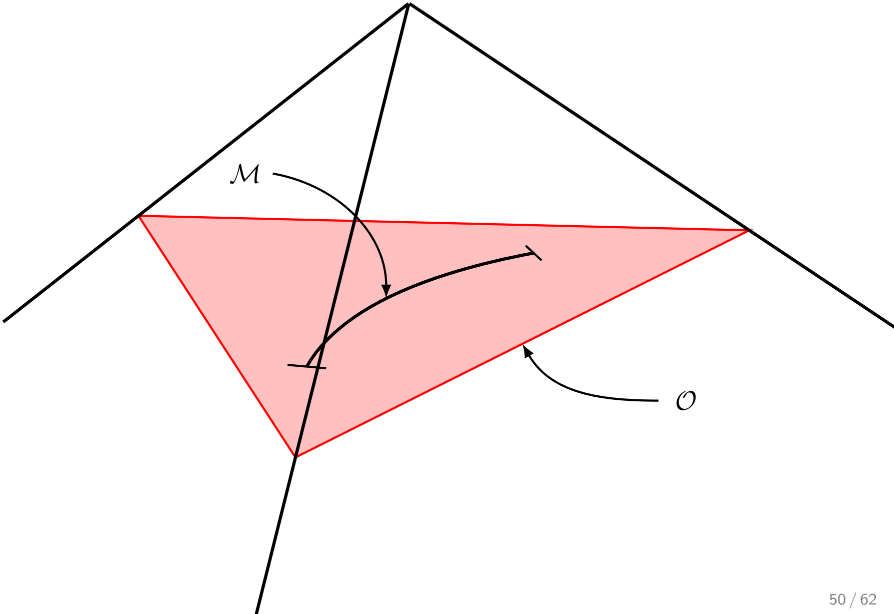
Getting the Picture



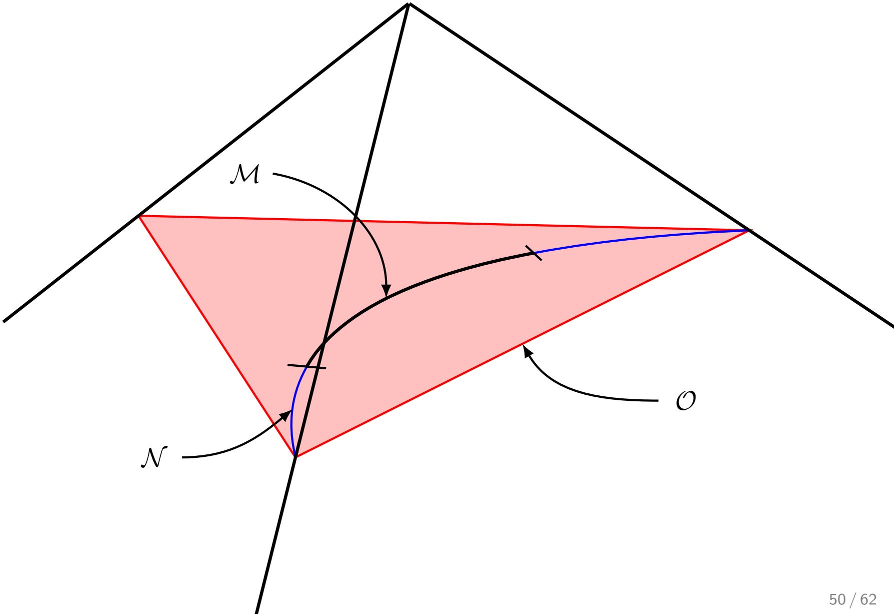
Getting the Picture



Getting the Picture

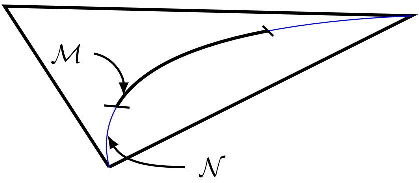


Getting the Picture



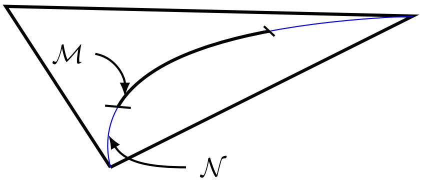
Proof idea for main result

- The nested model can be defined parametrically;



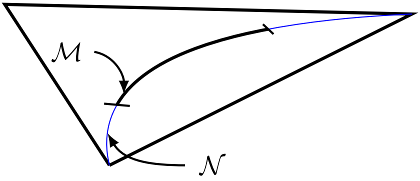
Proof idea for main result

- The nested model can be defined parametrically;
- therefore its Zariski closure is an irreducible variety;



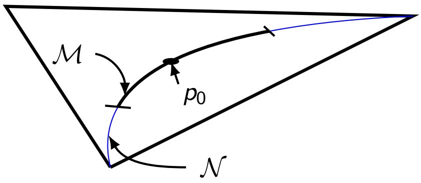
Proof idea for main result

- The nested model can be defined parametrically;
- therefore its Zariski closure is an irreducible variety;
- hence if, in a neighbourhood of a single point, the nested and complete models are the same dimension, then they have the same Zariski closure;



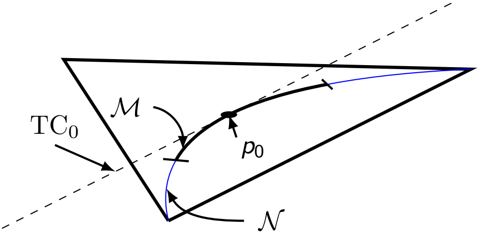
Proof idea for main result

- The nested model can be defined parametrically;
- therefore its Zariski closure is an irreducible variety;
- hence if, in a neighbourhood of a single point, the nested and complete models are the same dimension, then they have the same Zariski closure;
- the uniform distribution (complete independence, all states equally likely) is contained in any mDAG model;



Proof idea for main result

- The nested model can be defined parametrically;
- therefore its Zariski closure is an irreducible variety;
- hence if, in a neighbourhood of a single point, the nested and complete models are the same dimension, then they have the same Zariski closure;
- the uniform distribution (complete independence, all states equally likely) is contained in any mDAG model;
- we can perturb the relationship between latent and observed variables to 'move' $\mathcal{M}$ in any direction within the tangent space of $\mathcal{N}$.



Proof Outline

Can use log-linear parameters:

$$\log p(x_V) = \sum_{A \subseteq V} \lambda_A(x_A).$$

Uniform distribution has $\lambda_A = 0$ for all $A \neq \emptyset$.

Proof Outline

Can use log-linear parameters:

$$\log p(x_V) = \sum_{A \subseteq V} \lambda_A(x_A).$$

Uniform distribution has $\lambda_A = 0$ for all $A \neq \emptyset$.

If $X_A \perp\!\!\!\perp X_B \,|\, X_C$, then $\lambda_D(x_D) \approx 0$ for D such that

$$D \subseteq A \cup B \cup C, \quad D \cap A \neq \emptyset, \quad D \cap B \neq \emptyset.$$

Lemma

If $X_A \perp\!\!\!\perp X_B \,|\, X_C$ under $\mathcal{M}$, then $\Lambda_D \perp TC_0(\mathcal{M})$ for D as above.

Proof Outline

Can use log-linear parameters:

$$\log p(x_V) = \sum_{A \subseteq V} \lambda_A(x_A).$$

Uniform distribution has $\lambda_A = 0$ for all $A \neq \emptyset$.

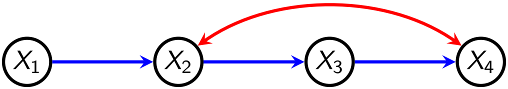
If $X_A \perp\!\!\!\perp X_B \,|\, X_C$, then $\lambda_D(x_D) \approx 0$ for D such that

$$D \subseteq A \cup B \cup C, \quad D \cap A \neq \emptyset, \quad D \cap B \neq \emptyset.$$

Lemma

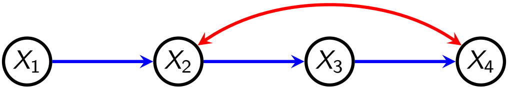
If $X_A \perp\!\!\!\perp X_B \,|\, X_C$ under $\mathcal{M}$, then $\Lambda_D \perp TC_0(\mathcal{M})$ for D as above.
In fact, this is true even for a dormant independence.

Verma Example



We have $X_1 \perp\!\!\!\perp X_3 \mid X_2$ and (after a re-weighting) $X_1 \perp\!\!\!\perp X_4 \mid X_3$.

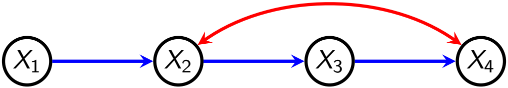
Verma Example



We have $X_1 \perp\!\!\!\perp X_3 \,|\, X_2$ and (after a re-weighting) $X_1 \perp\!\!\!\perp X_4 \,|\, X_3$.

Hence $\Lambda_{13} + \Lambda_{123} + \Lambda_{14} + \Lambda_{134} \perp TC_0(\mathcal{M})$.

Verma Example

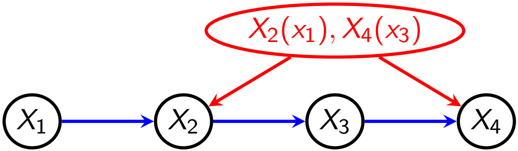


We have $X_1 \perp\!\!\!\perp X_3 \mid X_2$ and (after a re-weighting) $X_1 \perp\!\!\!\perp X_4 \mid X_3$.

Hence $\Lambda_{13} + \Lambda_{123} + \Lambda_{14} + \Lambda_{134} \perp TC_0(\mathcal{M})$.

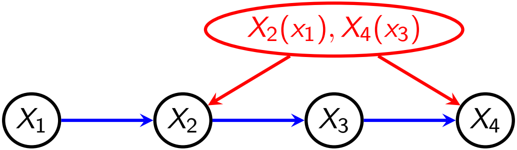
So: need to show all the *other* spaces λ_A are inside the tangent cone.

Verma Example



Perturbing	controls
X_1	Λ_1
$X_3 \mid X_2$	$\Lambda_3 + \Lambda_{23}$
$X_2(x_1)$	$\Lambda_2 + \Lambda_{12}$
$X_4(x_3)$	$\Lambda_4 + \Lambda_{34}$
$X_2(x_1), X_4(x_3)$ *jointly*	$\Lambda_{24} + \Lambda_{124} + \Lambda_{234} + \Lambda_{1234}$

Verma Example

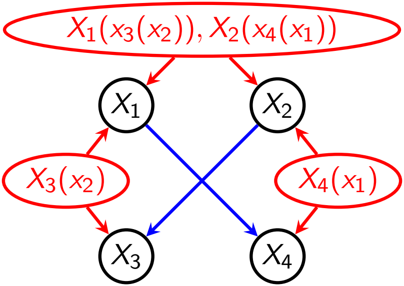


Perturbing	controls
X_1	Λ_1
$X_3 \mid X_2$	$\Lambda_3 + \Lambda_{23}$
$X_2(x_1)$	$\Lambda_2 + \Lambda_{12}$
$X_4(x_3)$	$\Lambda_4 + \Lambda_{34}$
$X_2(x_1), X_4(x_3)$ *jointly*	$\Lambda_{24} + \Lambda_{124} + \Lambda_{234} + \Lambda_{1234}$

$\Lambda_{13}, \Lambda_{123}, \Lambda_{14}, \Lambda_{134}$ are constrained, so that's all of them!

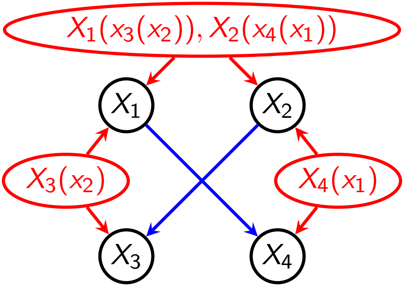
Geared Graphs

Back to our harder example:



Geared Graphs

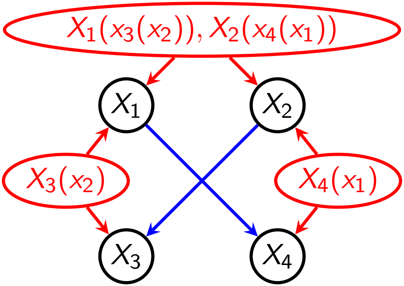
Back to our harder example:



Perturbing	controls
$X_3(x_2)$	$\Lambda_3 + \Lambda_{23}$
$X_4(x_1)$	$\Lambda_4 + \Lambda_{14}$

Geared Graphs

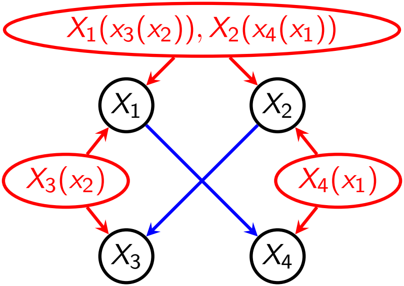
Back to our harder example:



Perturbing	controls
$X_3(x_2)$	$\Lambda_3 + \Lambda_{23}$
$X_4(x_1)$	$\Lambda_4 + \Lambda_{14}$
$X_1(x_3(x_2))$	$\Lambda_1 + \Lambda_{13} + \Lambda_{123}$
$X_2(x_4(x_1))$	$\Lambda_2 + \Lambda_{24} + \Lambda_{124}$

Geared Graphs

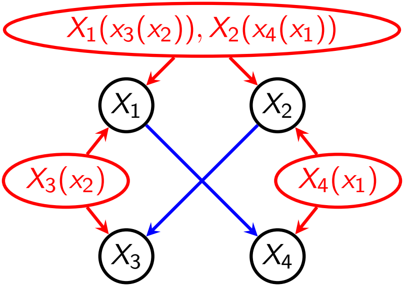
Back to our harder example:



Perturbing	controls
$X_3(x_2)$	$\Lambda_3 + \Lambda_{23}$
$X_4(x_1)$	$\Lambda_4 + \Lambda_{14}$
$X_1(x_3(x_2))$	$\Lambda_1 + \Lambda_{13} + \Lambda_{123}$
$X_2(x_4(x_1))$	$\Lambda_2 + \Lambda_{24} + \Lambda_{124}$
$X_2(x_4(x_1)), X_1(x_3(x_2))$ *jointly*	$\Lambda_{12} + \Lambda_{124} + \Lambda_{123} + \Lambda_{1234} +$

Geared Graphs

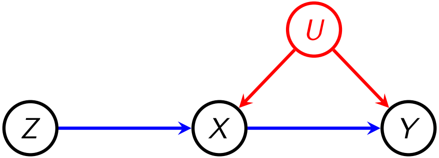
Back to our harder example:



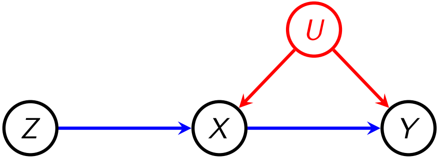
Perturbing	controls
$X_3(x_2)$	$\Lambda_3 + \Lambda_{23}$
$X_4(x_1)$	$\Lambda_4 + \Lambda_{14}$
$X_1(x_3(x_2))$	$\Lambda_1 + \Lambda_{13} + \Lambda_{123}$
$X_2(x_4(x_1))$	$\Lambda_2 + \Lambda_{24} + \Lambda_{124}$
$X_2(x_4(x_1)), X_1(x_3(x_2))$ *jointly*	$\Lambda_{12} + \Lambda_{124} + \Lambda_{123} + \Lambda_{1234} +$
	$+ \Lambda_{134} + \Lambda_{234} + \Lambda_{34}$

Outline

Inequality Results

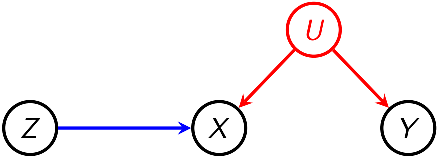


Inequality Results



$$p(x, y \mid z) = \int p(u)\, p(x \mid z, u) \cdot p(y \mid x, u)\, du$$

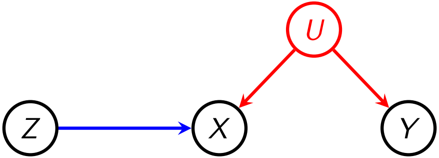
Inequality Results



$$p(x, y \mid z) = \int p(u)\, p(x \mid z, u) \cdot p(y \mid x, u)\, du$$

Let $p^*(x, y \mid z) \equiv \int p(u)\, p(x \mid z, u) \cdot p(y \mid x = 0, u)\, du$

Inequality Results



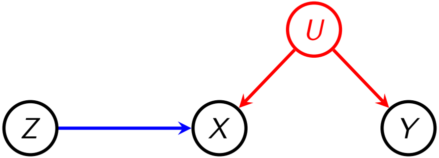
$$p(x, y \mid z) = \int p(u)\, p(x \mid z, u) \cdot p(y \mid x, u)\, du$$

$$\text{Let } p^*(x, y \mid z) \equiv \int p(u)\, p(x \mid z, u) \cdot p(y \mid x = 0, u)\, du$$

Can't observe p^* but:

- **Compatibility:** $p(0, y \mid z) = p^*(0, y \mid z)$ for each z, y; and
- **Independence:** $Y \perp\!\!\!\perp Z$ under p^*.

Inequality Results



$$p(x, y \mid z) = \int p(u)\, p(x \mid z, u) \cdot p(y \mid x, u)\, du$$

$$\text{Let } p^*(x, y \mid z) \equiv \int p(u)\, p(x \mid z, u) \cdot p(y \mid x = 0, u)\, du$$

Can't observe p^* but:

- **Compatibility:** $p(0, y \mid z) = p^*(0, y \mid z)$ for each z, y; and
- **Independence:** $Y \perp\!\!\!\perp Z$ under p^*.

This 'compatibility' requirement turns out to place an inequality restriction on p: $$\max_x \sum_y \max_z p(x, y \mid z) \leq 1.$$

Inequality Results

Generalizing this argument, we find a rich theory of results on inequalities (Evans, 2012).

Inequality Results

Generalizing this argument, we find a rich theory of results on inequalities (Evans, 2012).

However these results are **not exhaustive**!
Finding **all** inequality constraints in marginal models is probably an NP hard problem.

Inequality Results

Generalizing this argument, we find a rich theory of results on inequalities (Evans, 2012).

However these results are **not exhaustive**!
Finding **all** inequality constraints in marginal models is probably an NP hard problem.

Additionally:

- fitting models with inequality constraints is not trivial;
- the usual asymptotic results do not necessarily apply.

Inequality Results

Generalizing this argument, we find a rich theory of results on inequalities (Evans, 2012).

However these results are **not exhaustive**!
Finding **all** inequality constraints in marginal models is probably an NP hard problem.

Additionally:

- fitting models with inequality constraints is not trivial;
- the usual asymptotic results do not necessarily apply.

Maybe the nested model is a good compromise!

Outline

Summary

We have seen that:

- we can provide graphical derivations of constraints on DAG models; this leads to:
  - **(i)** the ordinary Markov model (conditional independences);
  - **(ii)** the nested Markov model (higher order polynomial constraints);
  - **(iii)** some inequalities.

Summary

We have seen that:

- we can provide graphical derivations of constraints on DAG models; this leads to:
  - **(i)** the ordinary Markov model (conditional independences);
  - **(ii)** the nested Markov model (higher order polynomial constraints);
  - **(iii)** some inequalities.
- the nested Markov model is 'complete' for algebraic constraints;

Summary

We have seen that:

- we can provide graphical derivations of constraints on DAG models; this leads to:
  - **(i)** the ordinary Markov model (conditional independences);
  - **(ii)** the nested Markov model (higher order polynomial constraints);
  - **(iii)** some inequalities.
- the nested Markov model is 'complete' for algebraic constraints;
- statistical and practical properties generally better than latent variable models;

Summary

We have seen that:

- we can provide graphical derivations of constraints on DAG models; this leads to:
  - **(i)** the ordinary Markov model (conditional independences);
  - **(ii)** the nested Markov model (higher order polynomial constraints);
  - **(iii)** some inequalities.
- the nested Markov model is ‘complete’ for algebraic constraints;
- statistical and practical properties generally better than latent variable models;
- we can also give graphical derivations for some inequalities.

Algebraic Questions

Are the complete models always semi-algebraic?

Algebraic Questions

Are the complete models always semi-algebraic?

Are polynomials of higher order harder to learn in finite samples?
Is so, can we give a careful explanation of why?

Algebraic Questions

Are the complete models always semi-algebraic?

Are polynomials of higher order harder to learn in finite samples? Is so, can we give a careful explanation of why?

Can we give a full characterization of when two complete models are the same?

Algebraic Questions

Are the complete models always semi-algebraic?

Are polynomials of higher order harder to learn in finite samples? Is so, can we give a careful explanation of why?

Can we give a full characterization of when two complete models are the same?

We've dealt with marginalization, but what about conditioning?

Thank you!

References

Evans. Graphical methods for inequality constraints in marginalized DAGs, *MLSP*, 2012.

Evans Margins of directed graphical models. Draft, 2014.

Evans and Richardson. Marginal log-linear parameters for graphical Markov models. *JRSS-B*, 2013.

Evans and Richardson. Markovian acyclic directed mixed graphs for discrete data. *Ann. Statist.*, (in press) 2014.

Pearl. On the testability of causal models with latent and instrumental variables, *UAI-95*, 1995.

Richardson. Markov properties for acyclic directed mixed graphs, *SJS*, 2003.

Richardson. A factorization criterion for acyclic directed mixed graphs, *UAI*, 2009.

Shpitser et al. Nested Markov Properties for Acyclic Directed Mixed Graphs. *UAI*, 2012.

Shpitser et al. Sparse nested Markov models with log-linear parameters. *UAI*, 2013.

Verma and Pearl. Equivalence and synthesis of causal models, *UAI-90*, 1990.

d-Separation

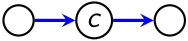
A **path** is a sequence of edges in the graph; vertices may not be repeated.

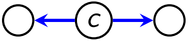
d-Separation

A **path** is a sequence of edges in the graph; vertices may not be repeated.

A path from a to b is **blocked** by $C \subseteq V \setminus \{a, b\}$ if either

(i) any non-collider is in C:



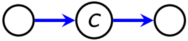


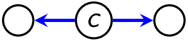
d-Separation

A **path** is a sequence of edges in the graph; vertices may not be repeated.

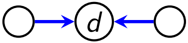
A path from a to b is **blocked** by $C \subseteq V \setminus \{a, b\}$ if either

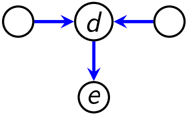
(i) any non-collider is in C:





(ii) or any collider is not in C, nor has descendants in C:



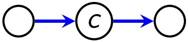


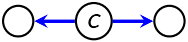
d-Separation

A **path** is a sequence of edges in the graph; vertices may not be repeated.

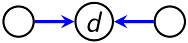
A path from a to b is **blocked** by $C \subseteq V \setminus \{a, b\}$ if either

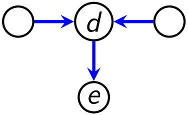
(i) any non-collider is in C:





(ii) or any collider is not in C, nor has descendants in C:





Two vertices a and b are **d-separated** given $C \subseteq V \setminus \{a, b\}$ if **all** paths are blocked.

Parameterizations

The nested and ordinary Markov models are also defined by

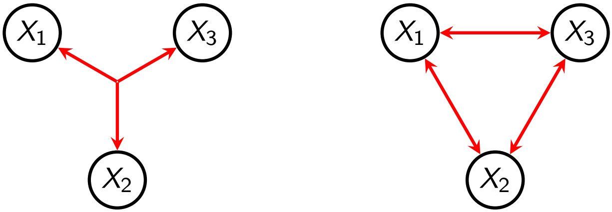
$$P(X_V = x_V) = \sum_{O \subseteq C \subseteq V} (-1)^{|C \setminus O|} \prod_{H \in [C]_{\mathcal{G}}} q_H(x_T).$$

for some pairs of sets (H, T), and partitioning function $[\cdot]_{\mathcal{G}}$. (See Evans and Richardson, 2014, for details)

Note the form is the same for the ordinary and nested models, but the partitioning function differs (as does the interpretation of the parameters q).

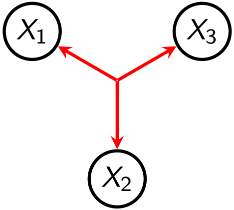
ADMGs are not sufficient

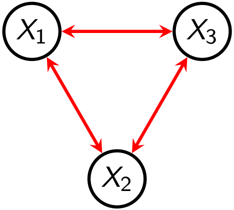
In general we need to distinguish between $\{1,2,3\}$ and $\{1,2\}$, $\{1,3\}$, $\{2,3\}$.



ADMGs are not sufficient

In general we need to distinguish between $\{1,2,3\}$ and $\{1,2\}$, $\{1,3\}$, $\{2,3\}$.





The model on the right is not saturated. Still true if we dichotomize.

ADMGs are not sufficient

Lemma

Let $\mathcal{F}$, $\mathcal{G}$, $\mathcal{H}$ be mutually independent σ-algebrae (so that $\mathcal{F} \perp\!\!\!\perp \mathcal{G} \vee \mathcal{H}$ and so on), and let X, Y and Z be random variables such that

(i) X is $\mathcal{F} \vee \mathcal{G}$-measureable;

(ii) Y is $\mathcal{G} \vee \mathcal{H}$-measureable;

(iii) Z is $\mathcal{F} \vee \mathcal{H}$-measureable.

Then $P(X = Y = Z) > 1 - \epsilon$ implies

$$\operatorname{Var} X < 3\epsilon.$$