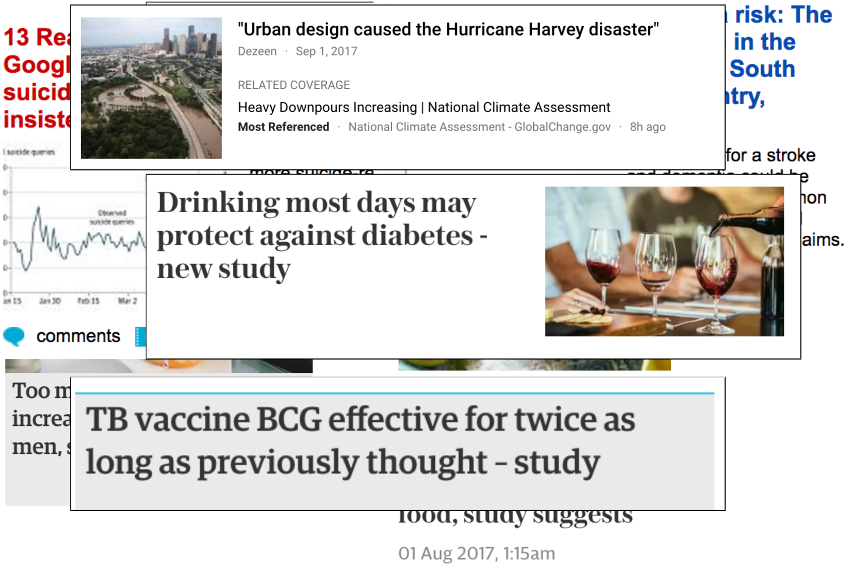
Model Selection and Local Geometry

Robin Evans, University of Oxford

CRM Montréal
20th June 2018

Causal Claims are Ubiquitous



13 Rea
Googl
suicid
insiste

comments

more suicide re

risk: The
in the
South
try,

for a stroke
and dementia could be
non
aims.

"Urban design caused the Hurricane Harvey disaster"
Dezeen · Sep 1, 2017

RELATED COVERAGE

Heavy Downpours Increasing | National Climate Assessment
Most Referenced · National Climate Assessment - GlobalChange.gov · 8h ago

Drinking most days may protect against diabetes - new study

Too m
increa
men, s

TB vaccine BCG effective for twice as long as previously thought – study

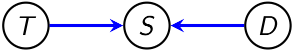
food, study suggests

01 Aug 2017, 1:15am

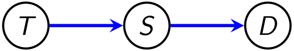
Distinguishing Between Causal Models

Observational data is cheap and readily available. Using it to rule out some causal models could save a lot of time and effort.

Can it be done?



$$p(t, s, d) = p(t)\, p(d)\, p(s \mid t, d)$$
$$T \perp\!\!\!\perp D$$



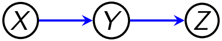
$$p(t, s, d) = p(t)\, p(s \mid t)\, p(d \mid s)$$
$$T \perp\!\!\!\perp D \mid S$$

Not always... but sometimes!

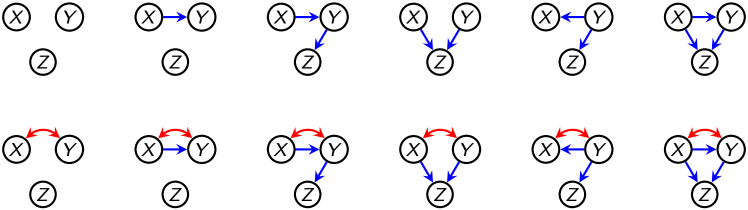
This is the basis of some causal search algorithms (e.g. PC, FCI).

The Holy Grail: Structure Learning

Given a distribution P from true model (or rather data from P)...



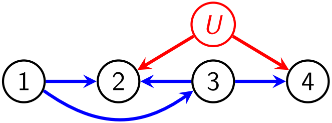
...and a set of possible causal models...

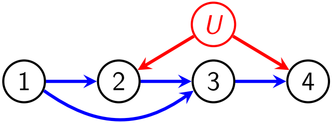


...return list of models which are compatible with data. [Some models are not observationally distinguishable.]

Question for today: is this feasible?

An Example





Model on left satisfies $X_1 \perp\!\!\!\perp X_4 \mid X_3$, in other words:

$$\sum_{x_2} p(x_4 \mid x_1, x_2, x_3) \cdot p(x_2 \mid x_1, x_3) \qquad \text{is independent of } x_1.$$

Model on right satisfies the **Verma constraint**:

$$\sum_{x_2} p(x_4 \mid x_1, x_2, x_3) \cdot p(x_2 \mid x_1) \qquad \text{is independent of } x_1.$$

Hence, the two models can be distinguished, and direction of the $2 - 3$ edge identified.

However, **empirically** this seems to be difficult to do correctly (Shpitser et al., 2013). Why?

Outline

1. Introduction

2. Graphical Models

3. The Problem

4. Tangent Cones and *k*-Equivalence

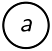
5. Overlap

6. Methods

Undirected Gaussian Graphical Models

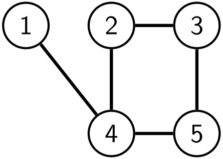
Suppose we have data $X_V = (X_1, X_2, \dots, X_p)^T \sim N_p(0, K^{-1})$.

vertex $\Longleftrightarrow$ random variable



$\Longleftrightarrow$ X_a

graph $\mathcal{G}$



$\Longleftrightarrow$

model $\mathcal{M}$

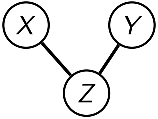
$\mathcal{M}(\mathcal{G}) = \{K \text{ satisfying } (*)\}$

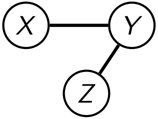
If i and j are not joined by an edge, then $k_{ij} = 0$:

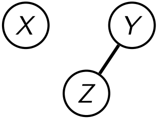
$$X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}} \qquad (*)$$

Undirected Gaussian Graphical Models

So in an **undirected Gaussian graphical model** represents zeroes in a concentration matrix by missing edges in an undirected graph:





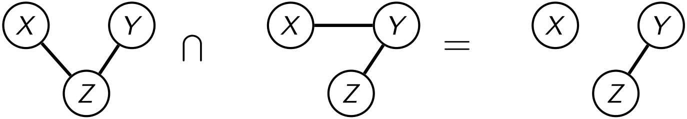


$$\begin{pmatrix} k_{xx} & 0 & k_{xz} \\ 0 & k_{yy} & k_{yz} \\ k_{xz} & k_{yz} & k_{zz} \end{pmatrix} \qquad \begin{pmatrix} k_{xx} & k_{xy} & 0 \\ k_{xy} & k_{yy} & k_{yz} \\ 0 & k_{yz} & k_{zz} \end{pmatrix} \qquad \begin{pmatrix} k_{xx} & 0 & 0 \\ 0 & k_{yy} & k_{yz} \\ 0 & k_{yz} & k_{zz} \end{pmatrix}$$

Undirected Graphs

Undirected graphical models have a lot of nice properties:

- Exponential family of models;
- convex log-likelihood function, relevant submodels all convex (linear subspaces);
- closed under intersection;



As a consequence, model selection in this class is highly feasible, even when $p \gg n$.

Graphical Lasso

For example, the graphical Lasso and several other methods can be used to perform automatic model selection via a convex optimization (Meinshausen and Bühlmann, 2006; Friedman et al., 2008):

$$\text{minimize}_{K \succ 0} \qquad -\log\det K + \operatorname{tr}(KS) + \lambda \sum_{i<j} |k_{ij}|.$$

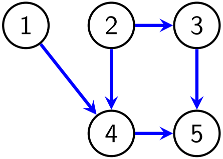
Convexity doesn't always mean a problem is easy, but...

From Hsieh et al. (2013):

> *State-of-the-art methods thus do not scale to problems with more than 20,000 variables. In this paper, we develop an algorithm . . . which can solve 1 million dimensional* ℓ_1*-regularized Gaussian MLE problems.*

Directed Graphical Models

graph $\mathcal{G}$



$\iff$

model $\mathcal{M}$

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

We do not allow directed cycles: $v \to \cdots \to v$.

If $i \to j$ say i is a **parent** of j. Denote

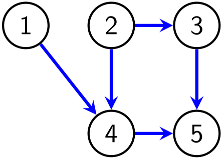
$$\mathrm{pa}_{\mathcal{G}}(j) = \{i : i \to j \text{ in } \mathcal{G}\}.$$

If i and j are not joined by an edge, and introducing $i \to j$ does not create a directed cycle, then

$$X_i \perp\!\!\!\perp X_j \mid X_{\mathrm{pa}_{\mathcal{G}}(j)} \qquad (\dagger)$$

Algebraic Models

Example:



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

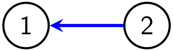
For Gaussian models, $X_i \perp\!\!\!\perp X_j \mid X_C$ means

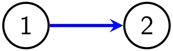
$$\begin{aligned}
&\rho_{ij\cdot C} \equiv \mathrm{Cor}(X_i, X_j \mid X_C) = 0\\
\iff \quad &\sigma_{ij} - \Sigma_{iC}(\Sigma_{CC})^{-1}\Sigma_{Cj} = 0.
\end{aligned}$$

These are polynomial constraints, so this is an **algebraic model**.

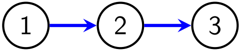
Markov Equivalence

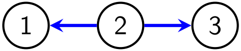
Sometimes two graphs imply the same set of independences: these are said to be **Markov equivalent**.

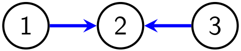




Two directed acyclic graphs are Markov equivalent if and only if they have the same **skeleton**, and the same **unshielded colliders**: $\rightarrow\leftarrow$







Directed Acyclic Graphs

Selection in the class of discrete Directed Acyclic Graphs is known to be NP Complete, i.e. ‘computationally difficult’ (Chickering, 1996).

Guarantees are hard: Cussens uses integer programming to find optimal discrete BNs for moderate ($\approx$50 variables).

Various attempts to develop a ‘directed graphical lasso’ have been made:

- Shojaie and Michailidis (2010) and Ni et al. (2015) assume a known causal ordering—reduces to edges being present or missing;
- Fu and Zhou (2013), Gu et al. (2014), Aragam and Zhou (2015) provide a procedure that is non-convex.

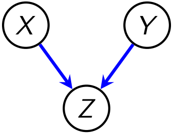
In this talk:

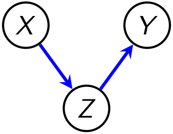
- We show that it is not possible to develop such a convex, ‘lasso-like’ procedure to select directed graphical models.
- In fact we will show that (for similar reasons) it is also ‘statistically’ difficult to perform this model selection.

Directed Acyclic Graphs

Selection in the class of Directed Acyclic Graphs is known to be NP Complete, i.e. 'computationally difficult' (Chickering, 1996).

I claim it can also be 'statistically' difficult. E.g.: how do we distinguish these two Gaussian graphical models?





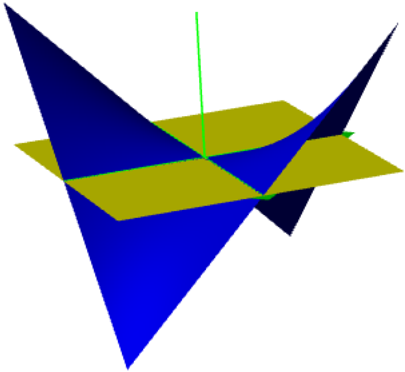
$\rho_{xy} = 0$ $\rho_{xy \cdot z} = 0$

But we have

$$\rho_{xy \cdot z} = 0 \qquad \Longleftrightarrow \qquad \rho_{xy} - \rho_{xz} \cdot \rho_{zy} = 0$$

so—if one of ρ_{xz} or ρ_{zy} is small—the models will be very similar.

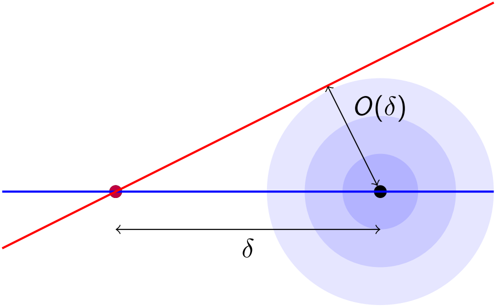
Marginal and Conditional Independence



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

A Picture

Suppose we have two sub-models (red and blue).

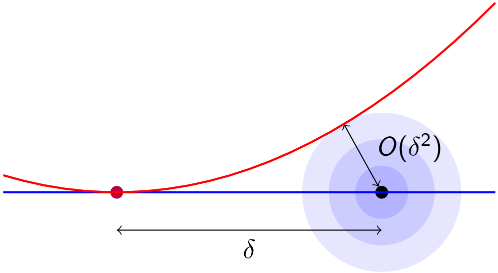


We intuitively expect to have power to test against alternatives long as our effect sizes are of order $n^{-1/2}$.

This applies to testing against the smaller intersection model and also against the red model.

A Slightly Different Picture

Suppose we have two sub-models with the *same tangent space*:



This time we still need $\delta \sim n^{-1/2}$ to obtain constant power against the intersection model, but $\delta \sim n^{-1/4}$ to have constant power against the red model!

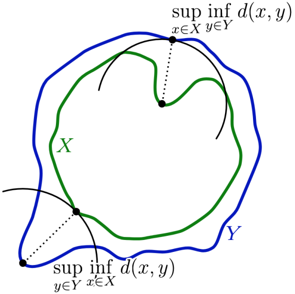
Hausdorff Distance

Hausdorff distance is a ‘maximin’ version of distance.

Given two sets A, B the **Hausdorff distance** between A and B is

$$\begin{aligned} d_H(A,B) &= \max\left\{ \sup_{a\in A}\inf_{b\in B}\|a-b\|,\ \sup_{b\in B}\inf_{a\in A}\|a-b\| \right\} \\ &= \max\left\{ \sup_{a\in A} d(a,B),\ \sup_{b\in B} d(b,A) \right\} \end{aligned}$$

Examples



$\sup_{x\in X}\inf_{y\in Y} d(x,y)$
X
Y
$\sup_{y\in Y}\inf_{x\in X} d(x,y)$

k-equivalence

k-equivalence at θ amounts to the Hausdorff distance shrinking faster than ε^k in an ε-ball.

Definition (Ferraroti et al., 2002)

We say Θ_1 and Θ_2 are k-**equivalent** at $\theta \in \Theta_1 \cap \Theta_2$ if

$$d_H(\Theta_1 \cap N_\varepsilon(\theta),\ \Theta_2 \cap N_\varepsilon(\theta)) = o(\varepsilon^k).$$

They are k-**near-equivalent** if

$$d_H(\Theta_1 \cap N_\varepsilon(\theta),\ \Theta_2 \cap N_\varepsilon(\theta)) = O(\varepsilon^k).$$

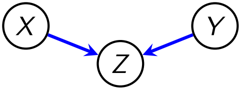
Examples.

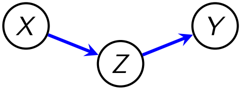
Intersecting $\implies$ 1-near-equivalent.

Same tangent cone $\iff$ 1-equivalent.

For regular models
k-equivalence $\implies$ $(k+1)$-near-equivalence. $(k \in \mathbb{N})$

Gaussian Graphical Models





$$X \perp\!\!\!\perp Y$$

$$\begin{pmatrix} 1 & 0 & \eta \\ & 1 & \varepsilon \\ & & 1 \end{pmatrix}$$

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

$$\begin{pmatrix} 1 & \varepsilon\eta & \eta \\ & 1 & \varepsilon \\ & & 1 \end{pmatrix}$$

For $X \perp\!\!\!\perp Y$, we can have any small η, ε, and need $\rho_{xy} = 0$.

The model $X \perp\!\!\!\perp Y \mid Z$ is similar but we need $\rho_{xy} = \varepsilon\eta$.

This is clearly only $O(\varepsilon\eta)$ from the $X \perp\!\!\!\perp Y$ model, so we have 2-near-equivalence at the identity matrix.

This extends to any two Gaussian models with the same skeleton.

Time Series

Time series models may also be 2-near-equivalent:

An MA(1) and AR(1) model have respective correlation matrices:

$$\begin{pmatrix} 1 & \rho & 0 & 0 & \cdots \\ \rho & 1 & \rho & 0 & \cdots \\ 0 & \rho & 1 & \rho & \\ \vdots & & & \ddots & \end{pmatrix} \qquad \begin{pmatrix} 1 & \theta & \theta^2 & \theta^3 & \cdots \\ \theta & 1 & \theta & \theta^2 & \cdots \\ \theta^2 & \theta & 1 & \theta & \\ \vdots & & & \ddots & \end{pmatrix}$$

So for small θ or ρ these may be hard to distinguish.

Statistical Consequences of k-(near-)equivalence

Suppose that models $\Theta_1, \Theta_2 \subseteq \Theta$ are k-near-equivalent at θ_0.

Consider a sequence of local ‘alternatives’ in Θ_1 of the form

$$\theta_n = \theta_0 + \delta n^{-\gamma} + o(n^{-\gamma});$$

then:

- we have limiting power to distinguish Θ_1 from $\Theta_1 \cap \Theta_2$ only if $\gamma \leq 1/2$ (i.e. the usual parametric rate);
- we have limiting power to distinguish Θ_1 from Θ_2 only if $\gamma \leq 1/(2k)$.

So if effect size is halved, we need 4^k times as much data to be sure we pick Θ_1 over Θ_2!

Submodels

Suppose that we have two models $\mathcal{M}_1, \mathcal{M}_2$.

Many classes of model (e.g. undirected graphs) are closed under intersection, so there is some nice submodel $\mathcal{M}_{12} = \mathcal{M}_1 \cap \mathcal{M}_2$.

However, suppose that this intersection is not so simple, but contains several distinct submodels...

Theorem

Suppose we have submodels $\mathcal{N}_1, \ldots, \mathcal{N}_k$ such that

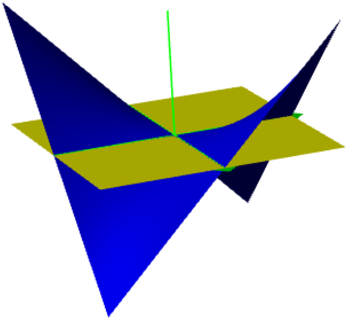
$$\mathcal{N}_i \cap \mathcal{M}_1 = \mathcal{N}_i \cap \mathcal{M}_2, \qquad \textit{for each } i = 1, \ldots, k,$$

and the spaces $\mathrm{TC}_\theta(\mathcal{N}_i)^\perp$ are all linearly independent.

*Then $\mathcal{M}_1$ and $\mathcal{M}_2$ are k-**near-equivalent** at any $\theta \in \mathcal{M}_1 \cap \mathcal{M}_2 \cap \mathcal{N}_1 \cap \cdots \cap \mathcal{N}_k$.*

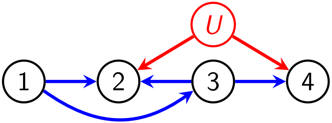
Marginal and Conditional Independence

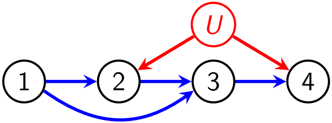
$X \perp\!\!\!\perp Y \mid Z$ $X \perp\!\!\!\perp Y$



These models coincide if $X \perp\!\!\!\perp Z$ or $Y \perp\!\!\!\perp Z$ (the axes).

Nested Models





Recall the constraints distinguishing these models:

$$\sum_{x_2} p(x_4 \mid x_1, x_2, x_3) \cdot p(x_2 \mid x_1, x_3) \qquad \text{is independent of } x_1$$

$$\sum_{x_2} p(x_4 \mid x_1, x_2, x_3) \cdot p(x_2 \mid x_1) \qquad \text{is independent of } x_1.$$

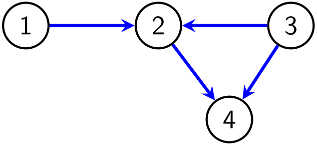
Note, the two models will become equivalent if **either**

- $X_2 \perp\!\!\!\perp X_3 \mid X_1$, **or**
- $X_4 \perp\!\!\!\perp X_2 \mid X_1, X_3$.

Hence the Theorem is satisfied with $k = 2$.

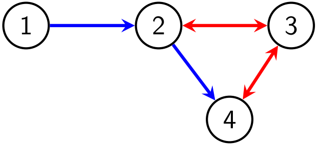
Discriminating Paths

In fact things can get much worse.



$X_1 \perp\!\!\!\perp X_3$

$X_4 \perp\!\!\!\perp X_1 \mid X_2, X_3$



$X_1 \perp\!\!\!\perp X_3$

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

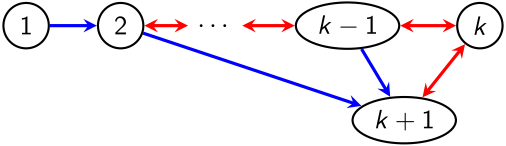
These graphs become Markov equivalent if **either**:

- $X_1 \perp\!\!\!\perp X_2$ (so $\rho_{12} = 0$);
- $X_2 \perp\!\!\!\perp X_3$ (so $\rho_{23} = 0$);
- $X_3 \perp\!\!\!\perp X_4 \mid X_1, X_2$ (so $\rho_{34 \cdot 12} = 0$).

So the theorem is satisfied with $k = 3$.

Discriminating Paths

This can be generalized into a **discriminating path** of arbitrary length.



In principle, one can distinguish:

$$
\begin{array}{ll}
\leftrightarrow k \leftrightarrow & X_1 \perp\!\!\!\perp X_{k+1} \mid X_2, \dots, X_{k-1} \\
\leftarrow k \rightarrow & X_1 \perp\!\!\!\perp X_{k+1} \mid X_2, \dots, X_{k-1}, X_k.
\end{array}
$$

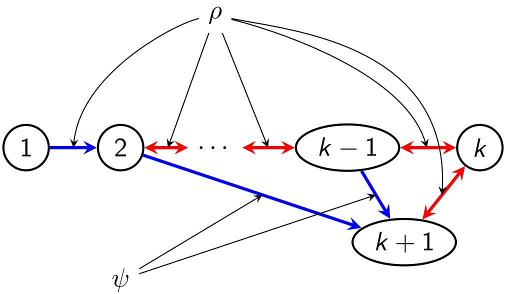
But: these graphs become Markov equivalent if **any** of:

- $X_i \perp\!\!\!\perp X_{i+1}$ for any $i = 1, \dots, k-1$;
- $X_{k+1} \perp\!\!\!\perp X_k \mid X_1, \dots, X_{k-1}$.

These are k distinct submodels, so the two models are k-near-equivalent.

Simulation

Take the discriminating path model:



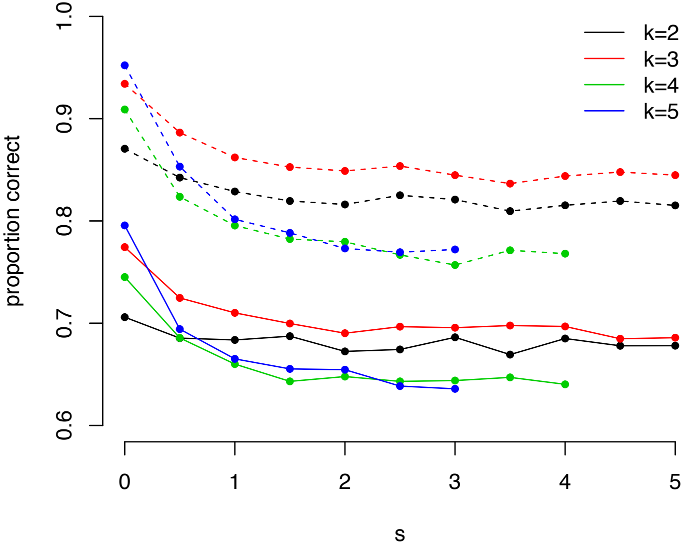
We generate data from the relevant Gaussian conditional independence model.

Fit the two models, and pick one with the smaller deviance.

We fix $\psi = 0.5$, let $\rho \to 0$, and see what sample size is required to maintain power.

Our results predict we will need $n \sim \rho^{-2k}$.

Discriminating Paths



effect size $\rho_s = 0.4 \times 2^{-s}$, sample size $n = n_{\text{init}} \times 2^{2sk}$.

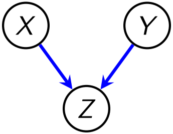
Required Sample Sizes

Sample sizes used for solid lines at $s = 1$ and $s = 2$.

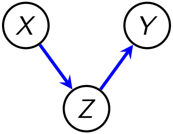
k	$\rho = 0.2$	$\rho = 0.1$
2	512	8,192
3	16 000	1 024 000
4	204 800	52 428 800
5	5.1 million	5.24 billion

Discrete DAG Models

For discrete, fully observed models, the situation is slightly different.



$X \perp\!\!\!\perp Y$



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

These models correspond to zero log-linear parameters

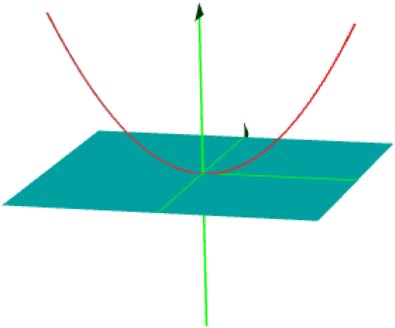
$$\lambda_{XY}^{XY} = 0 \qquad\qquad \lambda_{XY}^{XYZ} = \lambda_{XYZ}^{XYZ} = 0,$$

and clearly have different dimensions.

Even though λ_{XY}^{XY} and λ_{XY}^{XYZ} are ‘similar’ in the same manner as before, we have an extra parameter to play with.

Sketch

Qualitatively, the two discrete models look a bit like this:



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

Discrete Directed Graphs

Proposition

For any two discrete DAGs, either the models are identical or they are not 1-equivalent*.

*Actually, set of points at which they are 1-equivalent for any sensible polynomial submodel is measure zero.

In fact this result extends to ancestral graph models (Richardson and Spirtes, 2002), but **not** nested models.

Statistically we have a reprieve: there is always at least one parameter that we can use to distinguish between any two models.

Overlap

However, models that are not 1-equivalent can still be problematic.

Definition

Say that two models Θ_1, Θ_2 **overlap** at $\theta \in \Theta_1 \cap \Theta_2$ if

$$\mathrm{TC}_\theta(\Theta_1 \cap \Theta_2) \subset \mathrm{TC}_\theta(\Theta_1) \cap \mathrm{TC}_0(\Theta_2).$$

So in other words, there are directions of approaching θ in each model separately, but not in the intersection.

Overlap is weaker than 1-equivalence:

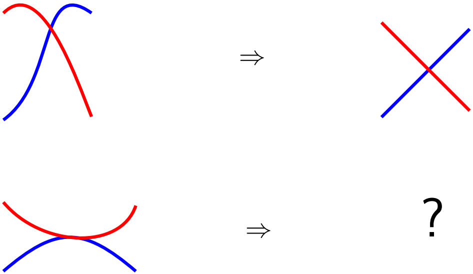
Proposition

If two regular algebraic models are 1-equivalent at θ, then either they are identical in a neighbourhood of θ, or the models overlap.

Computational Consequences of Overlap

Theorem

Suppose that models $\Theta_1, \Theta_2 \subseteq \Theta$ overlap (and are regular) at θ_0. Then there is no smooth reparameterization of Θ such that Θ_1 and Θ_2 are both convex.



This means that we can't adapt methods like the Lasso without making the problem non-convex (or using a more drastic relaxation).

Lack of Convexity

Example. For usual undirected Gaussian graphical models, one can solve use the graphical Lasso, which solves the convex program:

$$\text{minimize}_{K \succ 0} \qquad -\log\det K + \operatorname{tr}(KS) + \lambda \sum_{i<j} |k_{ij}|.$$

Example. For graphical models of marginal independence, the parameter spaces are defined by constraints of the form $\{\rho_{ij} = 0 \text{ whenever } i \not\sim j\}$.

The likelihood **not** convex in terms of covariance, but one can instead solve a problem like

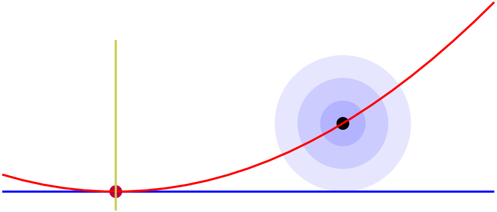
$$\text{minimize}_{\Sigma \succ 0} \qquad \|\Sigma - S\|^2 + \lambda \sum_{i,j} |\sigma_{ij}|$$

[Less efficient, but consistent for model selection and estimation has $n^{1/2}$-rate.]

This approach **cannot** be taken for models with overlap, because the angle between the models is always zero.

Towards Methods

An idea: can we **use** the fact that other marginal log-linear parameters are 'close', to deduce the correct log-linear representation?



If we 'blur' our likelihood by the right amount, we could obtain the correct sparsity level.

Then:

- learn the tangent space model;
- use that with earlier result to reconstruct the DAG.

Penalised Selection

Consider the usual Lasso approach:

$$\arg\min_{\boldsymbol{\lambda}} \left\{ -l(\boldsymbol{x}, \boldsymbol{\lambda}) + \nu_n \sum_{A \subseteq V} |\lambda_A| \right\}$$

if $\nu_n \sim n^\gamma$ for $\frac{1}{2} \leq \gamma < 1$ then the maxima $\hat{\boldsymbol{\lambda}}^n$ are consistent for model selection.

Theorem

Let

$$\boldsymbol{\lambda}^n = \mathbf{0} + \boldsymbol{\lambda} n^{-c} + o(n^{-c}).$$

be a sequence of points inside the DAG model for $\mathcal{G}$.
If $\frac{1}{4} < c < \frac{1}{2}$, the lasso will be consistent for the 'representation' of $\mathcal{G}$.

Asymptotic regime may not be realistic, but one can specify a sparsity level to choose penalization level in practice.

Summary

- Model selection in some classes of graphical models is harder than in others; this is at least partly explained by the local geometry of the model classes.
- Most Gaussian graphical models with the same skeleton are at least '2-near-equivalent', and are therefore statistically hard to distinguish.
- Discrete directed acyclic graph models are not 1-equivalent, but do 'overlap': this leads to computational problems.
- In particular, no 'directed graphical lasso' can exist.
- New methods could be created to use this information about the model geometry.

Thank you!

References I

Aragam and Zhou. Concave penalized estimation of sparse Gaussian Bayesian networks. Journal of Machine Learning Research, 16:2273-2328, 2015.

Bergsma and Rudas. Marginal log-linear parameters, *Ann. Statist.*, 2002.

Chickering. Learning Bayesian networks is NP-complete, *Learning from data.* Springer New York, 121-130, 1996.

Evans. Model selection and local geometry. *arXiv:1801.08364*, 2018.

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

Ferrarotti, Fortuna, and Wilson. Local approximation of semialgebraic sets. *Annali della Scuola Normale Superiore di Pisa*, 1:1-11, 2002.

Fu and Zhou. Learning sparse causal Gaussian networks with experimental intervention: regularization and coordinate descent. *JASA*, 108(501):288-300, 2013

Gu, Fu and Zhou. Adaptive penalized estimation of directed acyclic graphs from categorical data. *arXiv:1403.2310*, 2014.

Hsieh et al. BIG & QUIC: Sparse inverse covariance estimation for a million variables. *NIPS*, 2013.

Meinshausen and Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 1436–1462, 2006.

References II

Ni, Stingo and Baladandayuthapani. Bayesian nonlinear model selection for gene regulatory networks. *Biometrics*, 71(3):585-595, 2015

Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect, *Math. Modelling*, 1986.

Shojaie and Michailidis. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519-538, 2010.

Uhler, Raskutti, Bühlmann, Yu. Geometry of the faithfulness assumption in causal inference, *Annals of Statistics*, 2013.

Zwiernik, Uhler and Richards. Maximum likelihood estimation for linear Gaussian covariance models. *JRSS-B*, 2016.

Tangent Cones

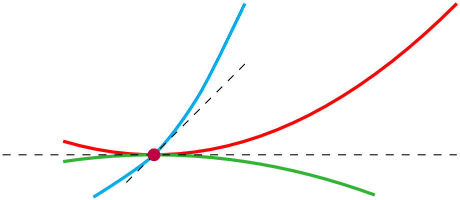
Definition

The **tangent cone** of Θ (at θ), is the set of vectors $\mathrm{TC}_\theta(\Theta)$ of the form

$$\lim_n \alpha_n(\theta_n - \theta),$$

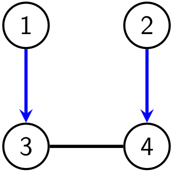
for sequences $\theta_n \to \theta$.

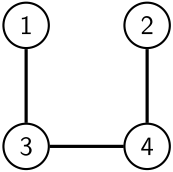
For regular models this a vector space (the **tangent space**), the derivative of Θ at θ.



Chain Graphs

For LWF chain graphs, distinct models may may be k-near-equivalent for arbitrarily large k.





$$X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3$$
$$X_2 \perp\!\!\!\perp X_3 \mid X_1, X_4$$
$$X_1 \perp\!\!\!\perp X_2$$

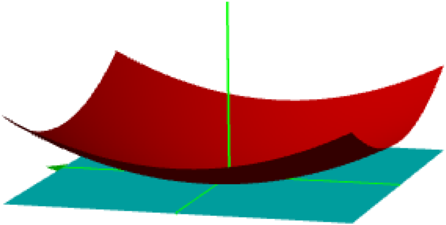
$$X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3$$
$$X_2 \perp\!\!\!\perp X_3 \mid X_1, X_4$$
$$X_1 \perp\!\!\!\perp X_2 \mid X_3, X_4$$

Their shared tangent cones are $\Lambda_{13} \oplus \Lambda_{34} \oplus \Lambda_{24}$.

These models are identical whenever any of $X_1 \perp\!\!\!\perp X_3$, $X_3 \perp\!\!\!\perp X_4$, or $X_2 \perp\!\!\!\perp X_4$ holds.

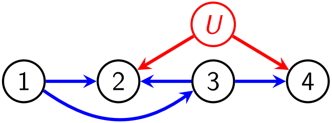
Other Kinds of Overlap

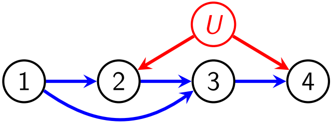
Note it is not necessary for two models to share submodels in order to have k-equivalence for any $k \geq 1$.



Discrete Verma Constraint

Consider the two models:





The are defined by the constraints:

$$\sum_{x_2} p(x_4 \mid x_1, x_2, x_3) \cdot p(x_2 \mid x_1, x_3) \qquad \text{is independent of } x_1;$$

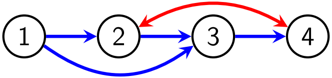
$$\sum_{x_2} p(x_4 \mid x_1, x_2, x_3) \cdot p(x_2 \mid x_1) \qquad \text{is independent of } x_1.$$

Though distinct, these constraints become identical if either:

$$X_2 \perp\!\!\!\perp X_3 \mid X_1 \qquad\qquad X_4 \perp\!\!\!\perp X_2 \mid X_1, X_3.$$

This satisfies the theorem, so the models are 2-near-equivalent.

Gaussian Verma Constraint



From Drton, Sullivant and Sturmfels (2009), the *Verma constraint* for a Gaussian model on four variables is given by zeroes of fourth order polynomial on correlations:

$$\begin{aligned} f(R) &= \rho_{14} - \rho_{14}\rho_{12}^2 - \rho_{14}\rho_{23}^2 + \rho_{14}\rho_{12}\rho_{13}\rho_{23} \\ &\qquad - \rho_{13}\rho_{34} + \rho_{13}\rho_{23}\rho_{24} + \rho_{12}^2\rho_{13}\rho_{34} - \rho_{12}\rho_{13}^2\rho_{24} \\ &= (\rho_{14} - \rho_{13}\rho_{34})(1 - \rho_{12}^2 - \rho_{23}^2 + \rho_{23}\rho_{12}\rho_{13}) + \cdots \\ &\qquad - \rho_{13}(\rho_{34}\rho_{23} - \rho_{24})(\rho_{23} - \rho_{12}\rho_{13}) \\ &= \rho_{14} - \rho_{13}\rho_{34} + O(\varepsilon^3) \\ &= \rho_{14} + O(\varepsilon^2). \end{aligned}$$

Model is not only locally linearly equivalent to the model of $X_1 \perp\!\!\!\perp X_4$, but also *quadratically* equivalent to the model $X_1 \perp\!\!\!\perp X_4 \mid X_3$.

In this case we would generally need effect sizes $\sim n^{-1/6}$(!)