# Angles and Model Selection

Robin Evans, University of Oxford

Technical University Munich
24th October 2019

# Causal Claims are Ubiquitous

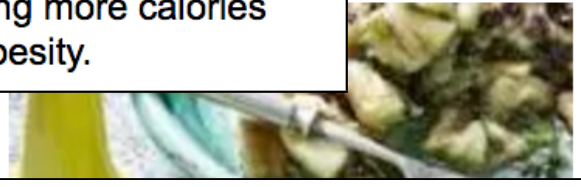
**A varied diet may NOT be healthy: Eating different types of food causes people to take in MORE calories that could lead to weight gain and health problems**

A new advisory published by the American Heart Association has warned Americans that eating a diet with many food options can actually lead to you consuming more calories and to obesity.

Cause, Not
Health

T

Too much sug
increase depr
men, study su



**Drinking most days may protect against diabetes - new study**

can afford quality
food, study suggests

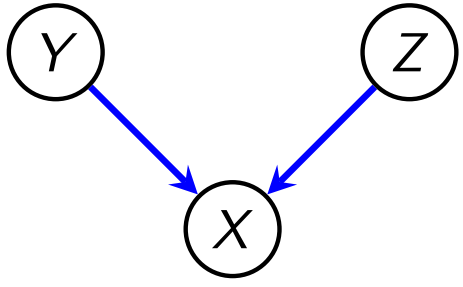
01 Aug 2017, 1:15am

# Distinguishing Between Causal Models

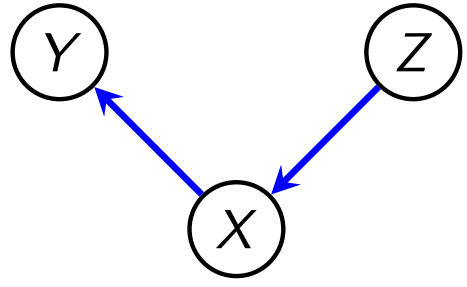
Causality is best inferred from experiments.
But doing experiments is hard (expensive, impractical, unethical...)

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?




$Z \perp\!\!\!\perp Y$



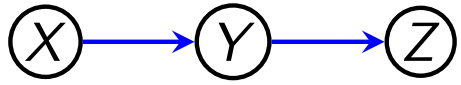

$Z \perp\!\!\!\perp Y \mid X$

Not always... but sometimes!

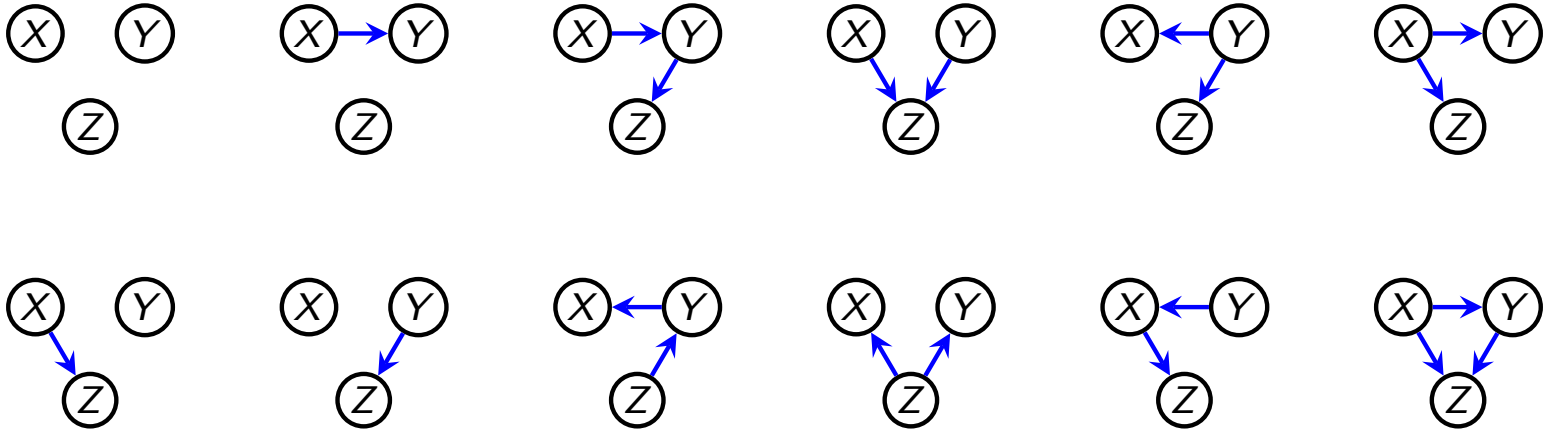
This is the basis of some causal search algorithms (e.g. PC, FCI).

# 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.]

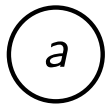
**Question for today: is this feasible? How easy/difficult is it?**

# Outline

# Undirected Gaussian Graphical Models

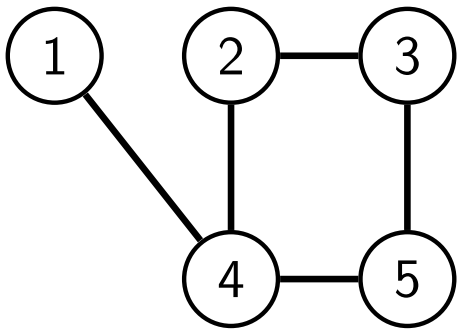
Suppose we have data $X_V = (X_1, X_2, \ldots, X_p)^T \sim N_p(0, K^{-1})$.

vertex $\iff$ random variable




$X_a$

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




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

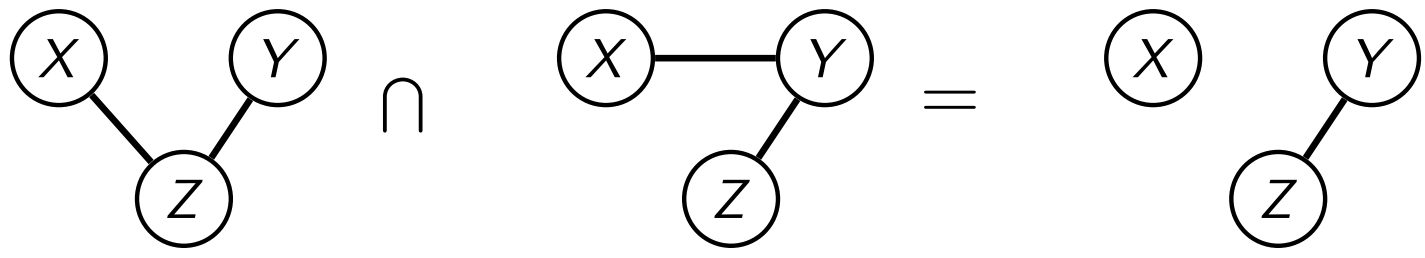
If $i$ and $j$ are not joined by an edge, then $k_{ij} = 0$:

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

# Undirected 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):

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

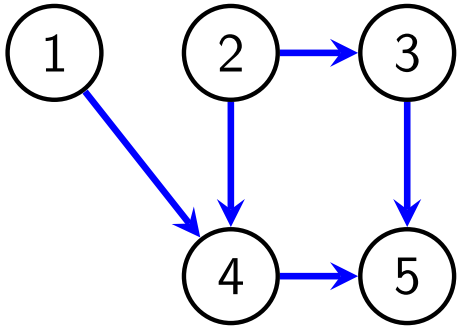
Convexity doesn't always mean a problem is easy, but...

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 $\ell_1$-regularized Gaussian MLE problems.*

# Directed Graphical Models

graph $\mathcal{G}$




$\Longleftrightarrow$

model $\mathcal{M}$

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

If $i$ and $j$ are not joined by an edge, then for a certain set $C \subseteq V \setminus \{i, j\}$ we have

$$X_i \perp\!\!\!\perp X_j \mid X_C. \qquad (\dagger)$$

[Note: can always pick parents of either $i$ or $j$.]

# 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.

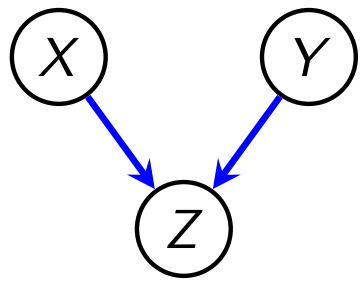
In this talk:

- We show that it is 'statistically' difficult to perform this model selection.
- We also show that, even far from $\Sigma = I$ it may be surprisingly difficult to distinguish between models.

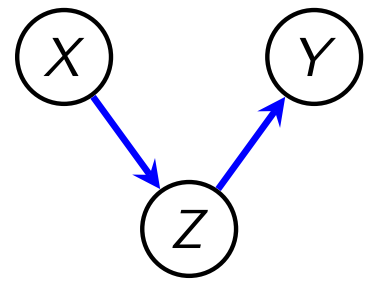
# Directed Acyclic Graphs

Selection in the class of discrete 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?




$$\rho_{xy} = 0$$



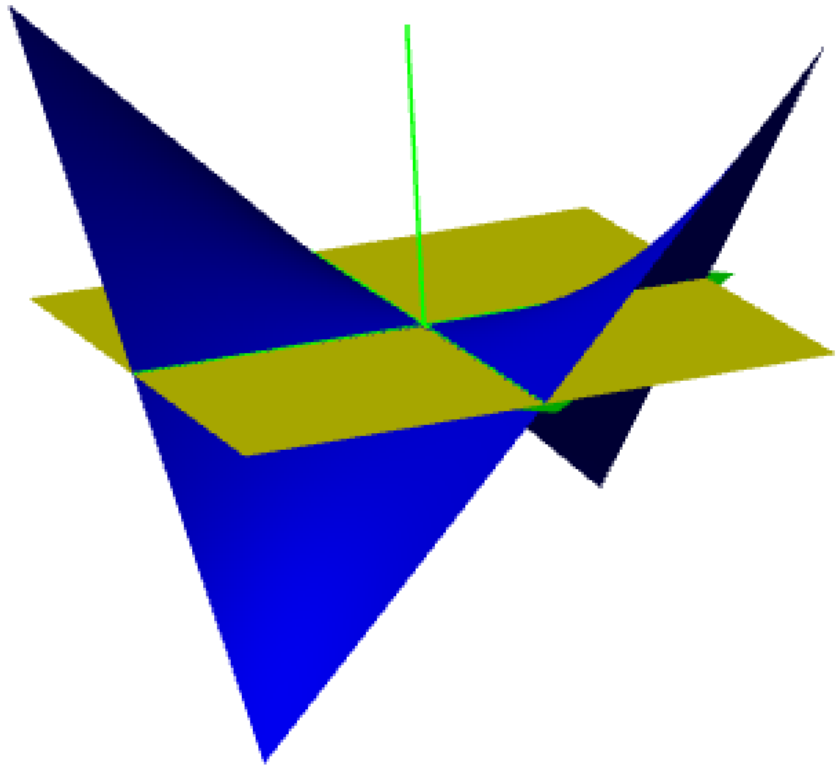

$$\rho_{xy \cdot z} = 0$$

But we have

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

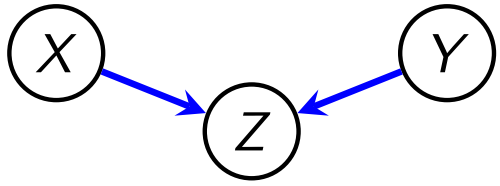
so—if one of $\rho_{xz}$ or $\rho_{zy}$ is small—the models will be very similar.

# Marginal and Conditional Independence



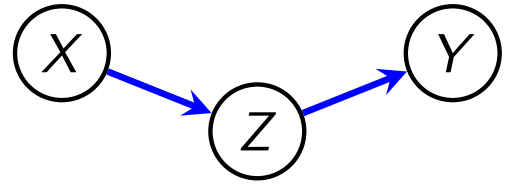
$X \perp\!\!\!\perp Y \mid Z$ $X \perp\!\!\!\perp Y$

# 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 $\eta, \varepsilon$, and need $\rho_{xy} = 0$.

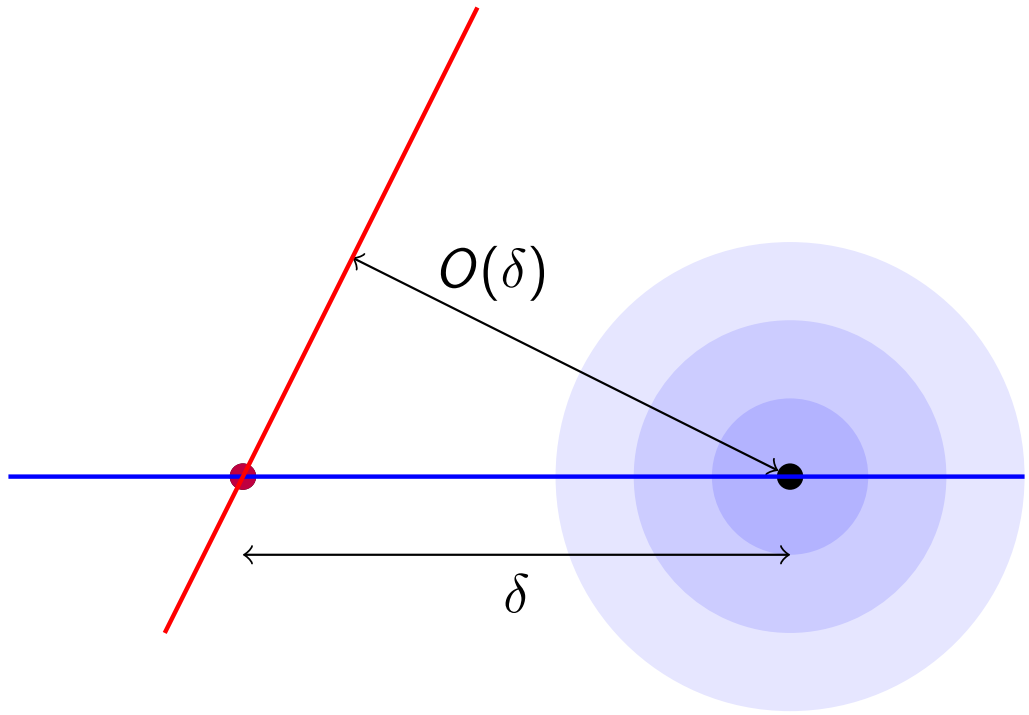
The model $X \perp\!\!\!\perp Y \mid Z$ is similar but we need $\rho_{xy} = \varepsilon\eta$.

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 graphical models with the same skeleton**.

# A Picture

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

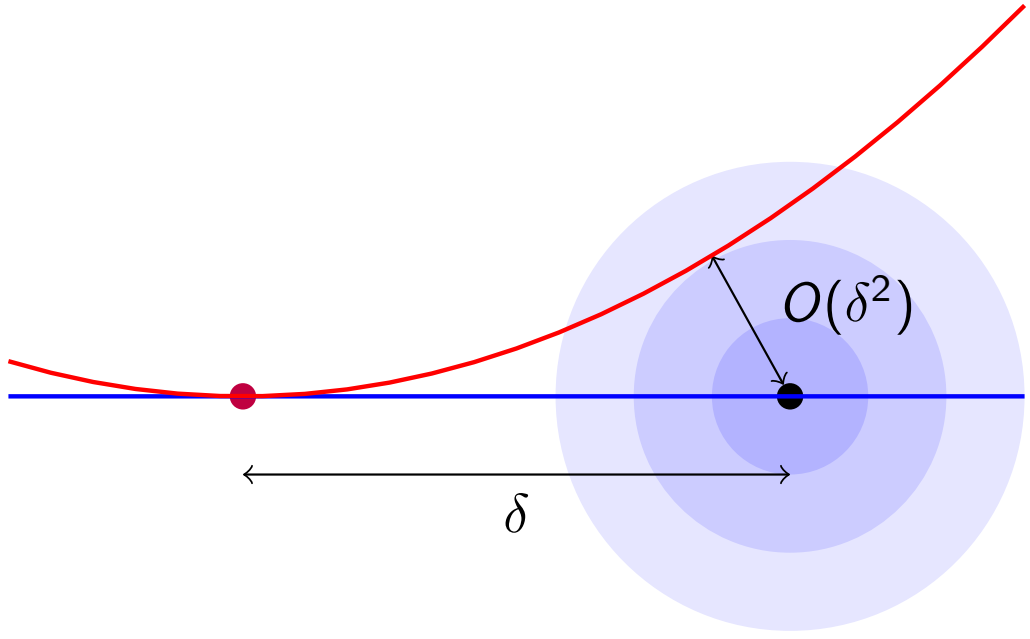



We intuitively expect to have power to test against alternatives long as our effect sizes are of order $n^{-1/2}$.

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!

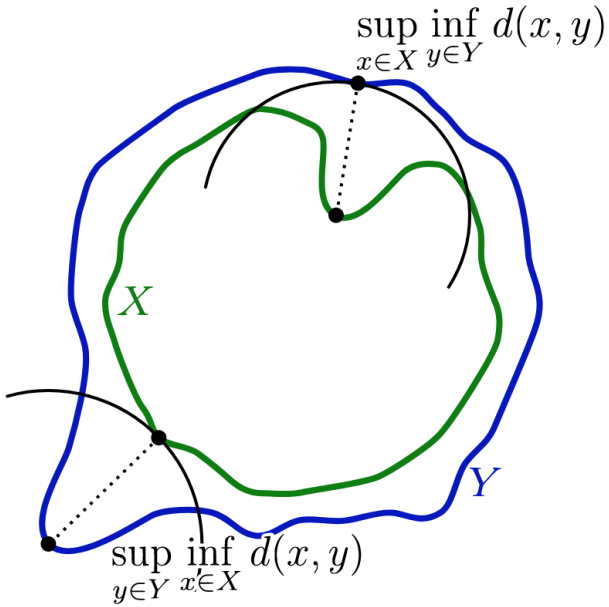
# Hausdorff Distance

Hausdorff distance is a 'maximin' version of distance.

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

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

**Examples**



# $k$-equivalence

$k$-equivalence at $\theta$ amounts to the Hausdorff distance shrinking faster than $\varepsilon^k$ in an $\varepsilon$-ball.

## Definition (Ferraroti et al., 2002)

We say $\Theta_1$ and $\Theta_2$ 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. $(k \in \mathbb{N})$

# Statistical Consequences of $k$-(near-)equivalence

Suppose that regular models $\Theta_1, \Theta_2 \subseteq \Theta$ are $k$-near-equivalent at $\theta_0$.

Consider a sequence of local 'alternatives' in $\Theta_1$ of the form

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

then:

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

# Submodels

Many classes of model (e.g. undirected graphs) are closed under intersection, so there is some nice submodel $\mathcal{M}_{1,2} = \mathcal{M}_1 \cap \mathcal{M}_2$.

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

## Theorem

*Let $\mathcal{M}_1, \mathcal{M}_2$ be **algebraic models**, regular at $\theta$. Suppose we have algebraic models $\mathcal{N}_1, \ldots, \mathcal{N}_k$ (also regular at $\theta$) such that*

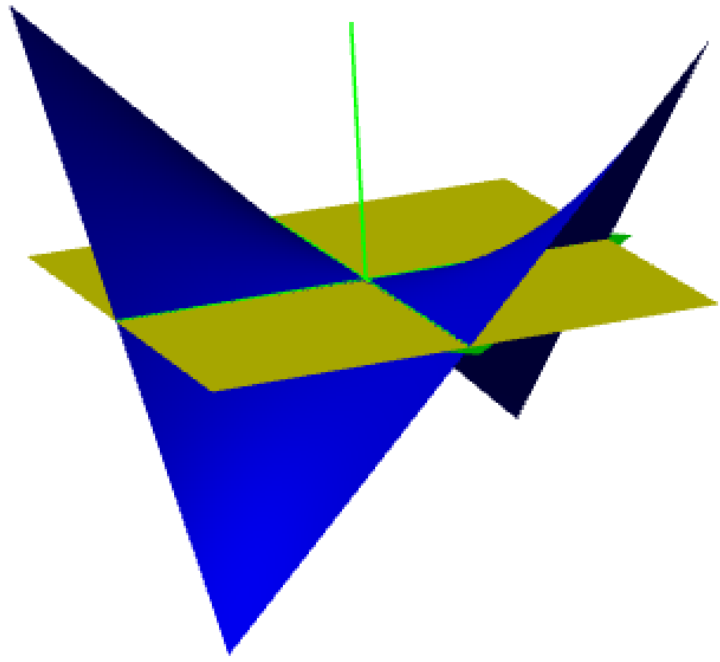
$$\mathcal{N}_i \cap \mathcal{M}_1 = \mathcal{N}_i \cap \mathcal{M}_2, \qquad \text{for each } i = 1, \ldots, k,$$

*and the spaces $\mathsf{TS}_\theta(\mathcal{N}_i)^\perp$ all have linearly independent bases.*

*Then $\mathcal{M}_1$ and $\mathcal{M}_2$ are $k$-**near-equivalent** at $\theta$.*

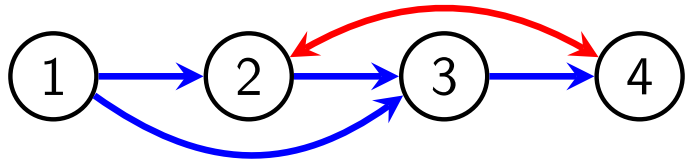
# Marginal and Conditional Independence

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



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

# Gaussian Verma Constraint




From Drton, Sullivant and Sturmfels (2009), the *Verma constraint* for a Gaussian version of this model is given by zeroes of

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

This collapses to $X_1 \perp\!\!\!\perp X_4 \mid X_3$ if any of

$$\rho_{13} = 0 \qquad \rho_{24\cdot 3} = 0 \qquad \rho_{23\cdot 1} = 0.$$

Hence theorem satisfied with $k = 3$.

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

# Angle of Surfaces

What happens if we instead consider the angle between these two surfaces?

We can simulate 'typical' covariances by sampling uniformly from the space of correlation matrices (Joe, 2006).

Unfortunately, this leads to matrices with small eigenvalues.

We get around this by sampling using a beta distribution with $a = b = 2$ (and scaling suitably).

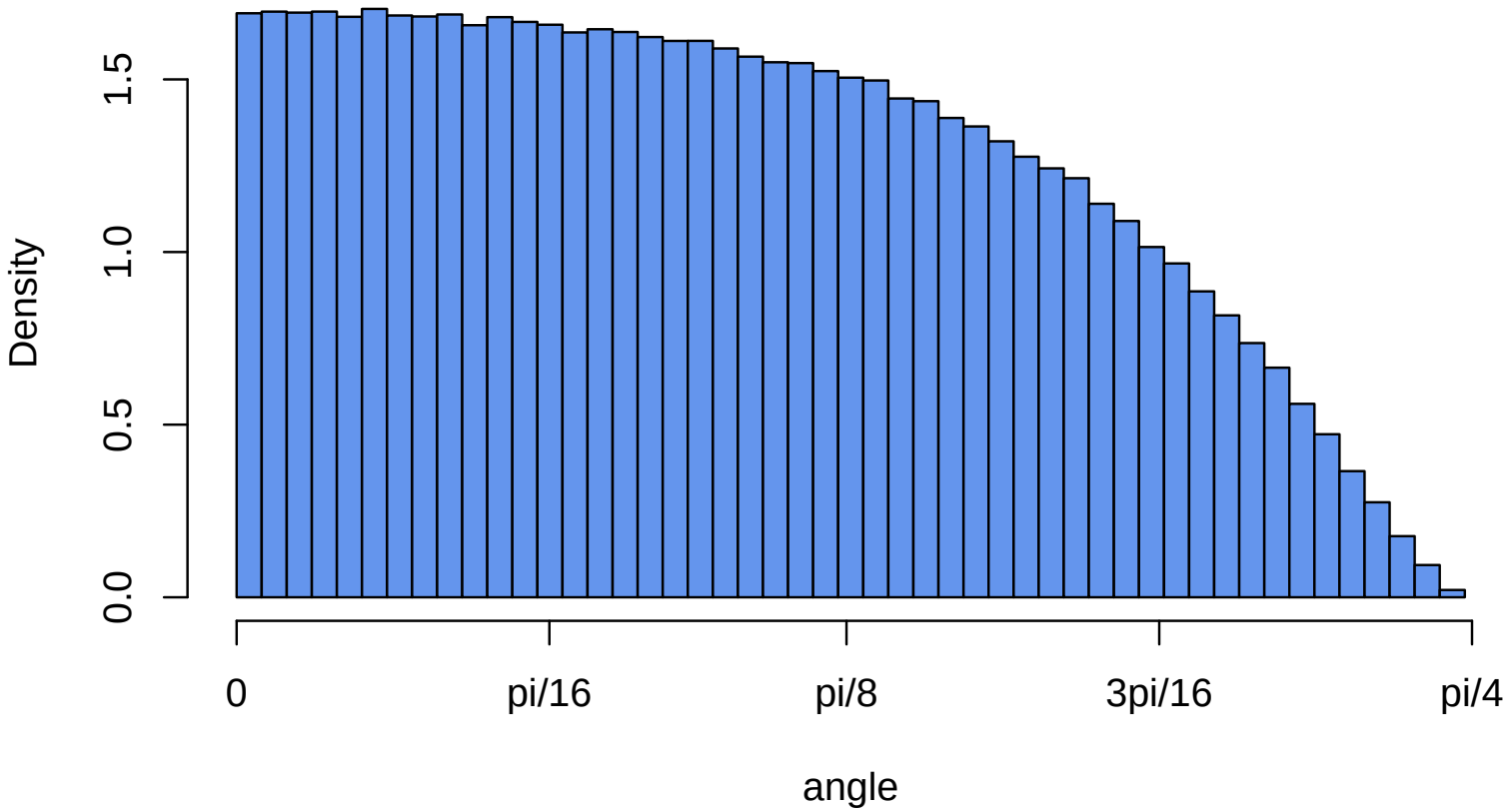
We can set $\rho_{12} = 0$

$p = 3$: need either $\rho_{13} = 0$ or $\rho_{23} = 0$, so pick the former arbitrarily.

$p = 4$: can use the Jacobian to fix $\rho_{12 \cdot 34} = 0$ (but keep $\rho_{12} = 0$).

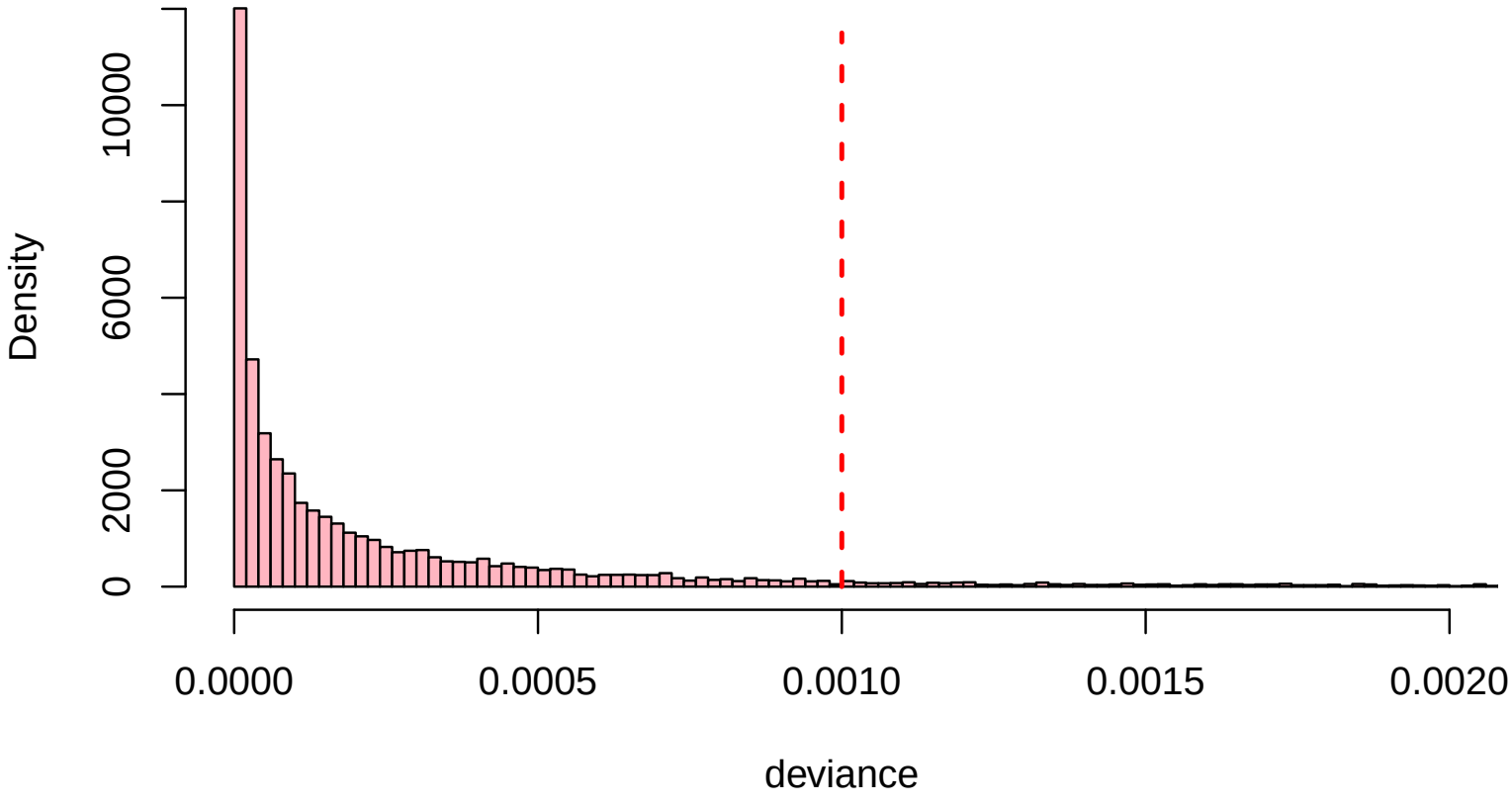
# Angles ($p = 3$)

For the 3-variate case, this gives the set of principal angles shown below:



# Angles ($p = 3$)

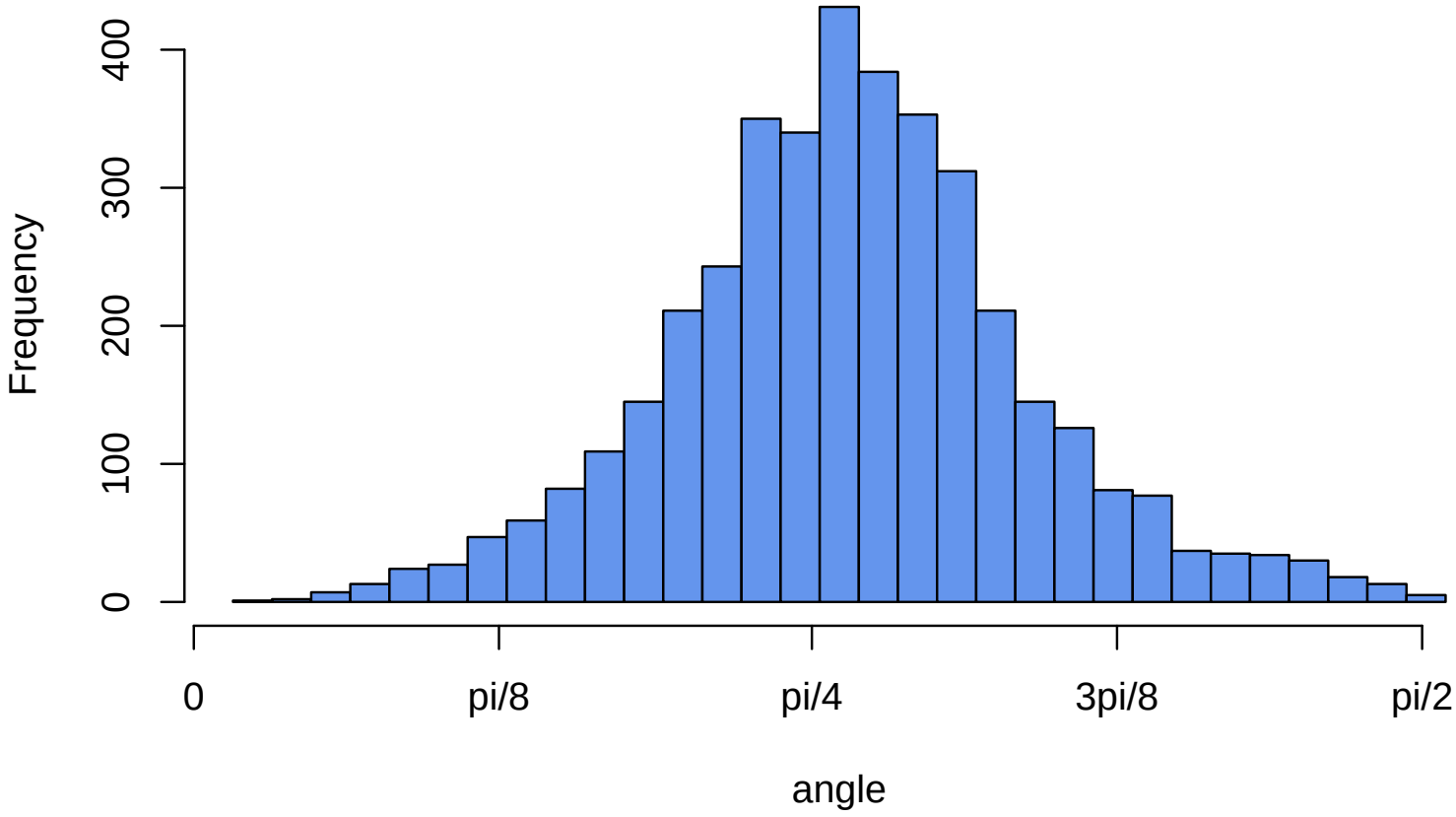
We shift using the Jacobian of $\rho_{12 \cdot 3}$, but keep within the model $\rho_{12} = 0$. The changes in deviance are shown on this histogram.




We made moves in correlation space of magnitude 0.001, and selected $n = 1,000$ as the sample size.

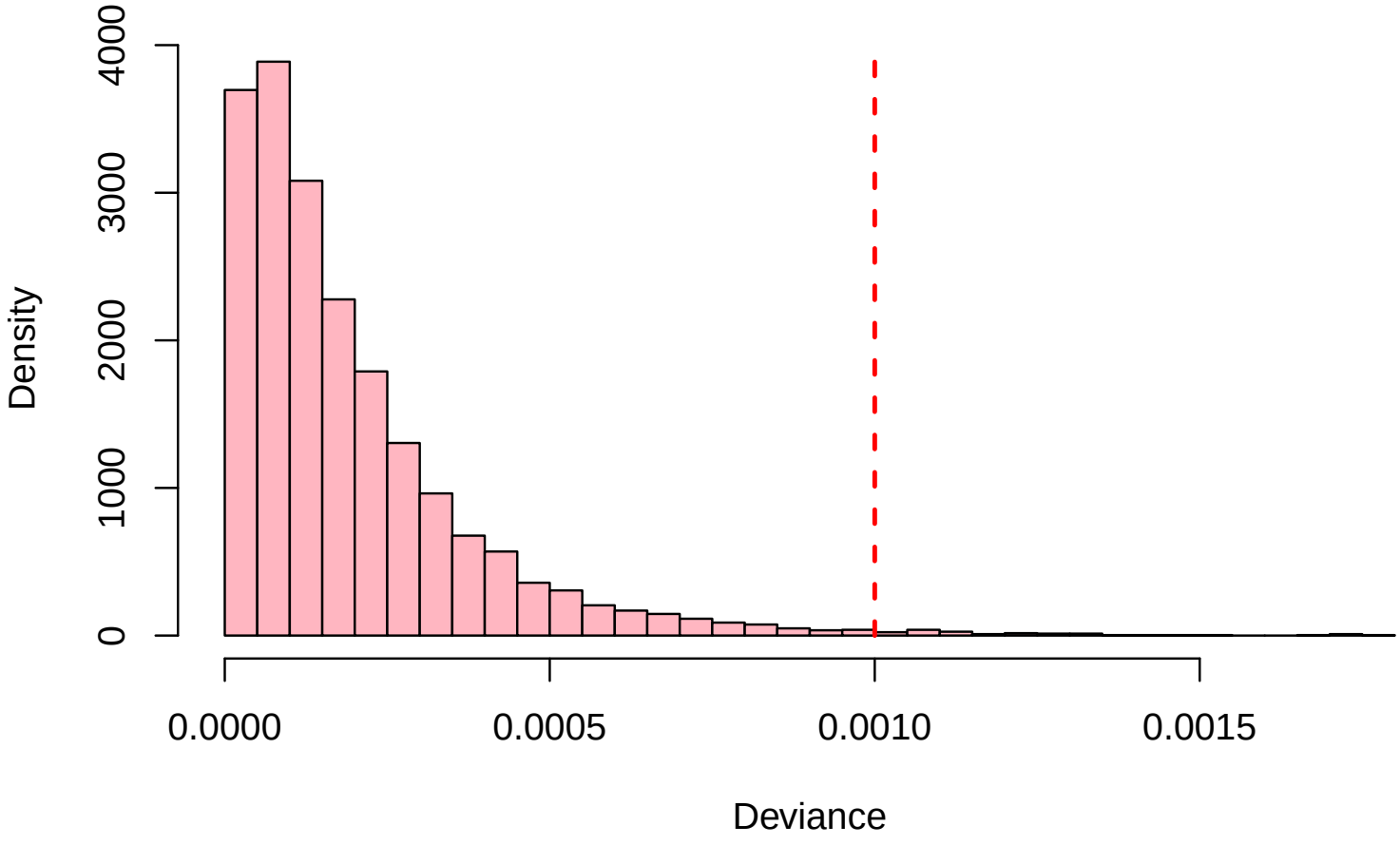
# Angles ($p = 4$)

For $|V| = 4$ we choose $a = b = 10$ to get more plausible covariance matrices.



# Angles ($p = 4$)

This gives the following power on the deviance scale:




For a movement of 0.001 in correlation space, with $n = 1,000$.

# Moral

More or less wherever the models $\rho_{12} = 0$ and $\rho_{12.3} = 0$ intersect, it's hard to tell which one is right.

The same story holds for $\rho_{12} = 0$ and $\rho_{12.34} = 0$.

So be careful!

# 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.
- Even relatively far from diagonal $\Sigma$, it can be tricky to tell which of several models is most likely.

# Thank you!

# References


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.

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*, 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.

Joe. Generating random correlation matrices based on partial correlations, *Journal of Multivariate Analysis*, 2006.

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

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

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

# Tangent Cones

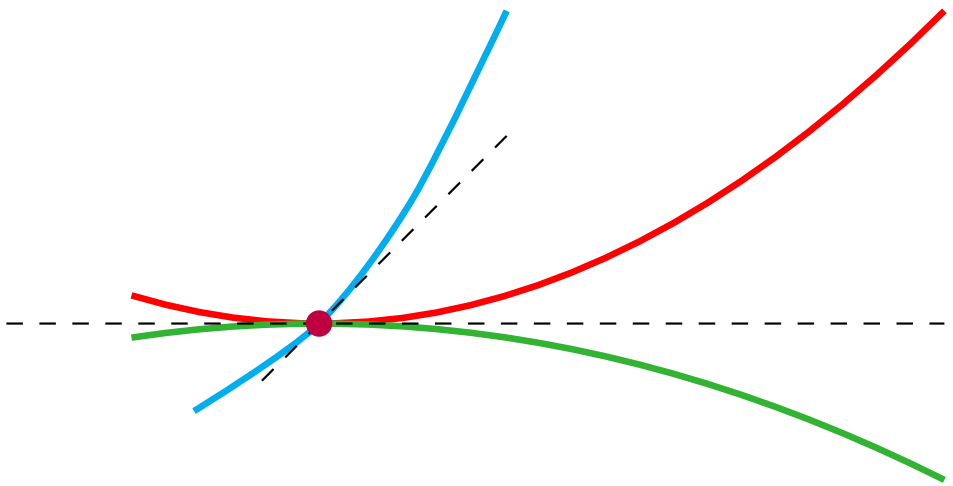
## Definition

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

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

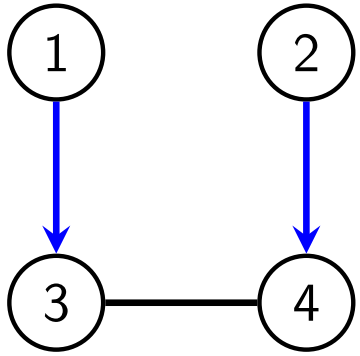
for sequences $\theta_n \to \theta$.

For regular models this a vector space (the **tangent space**), the derivative of $\Theta$ at $\theta$.

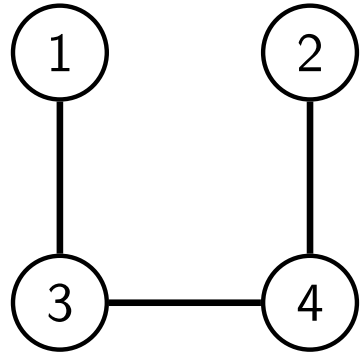


# Chain Graphs

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




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



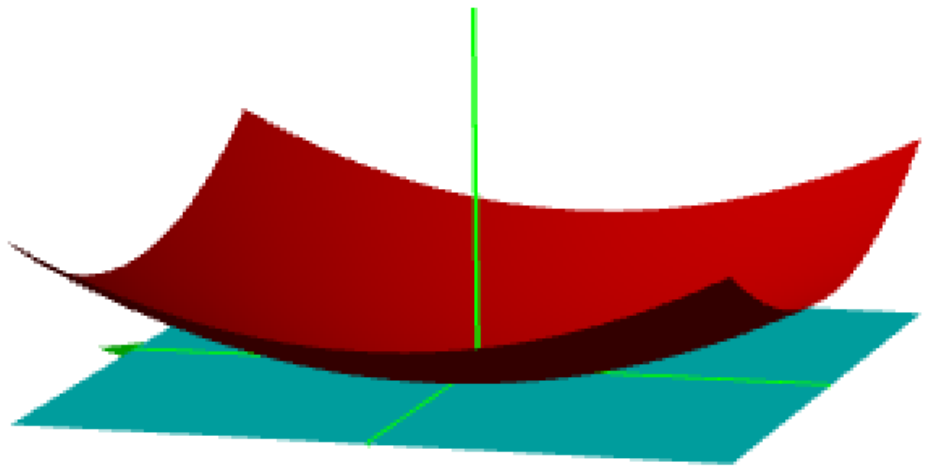

$$X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3$$
$$X_2 \perp\!\!\!\perp X_3 \mid X_1, X_4$$
$$X_1 \perp\!\!\!\perp X_2 \mid X_3, X_4$$

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 $k \geq 1$.



# 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 $k = 2$.

# 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 $\theta$ or $\rho$ these may be hard to distinguish.