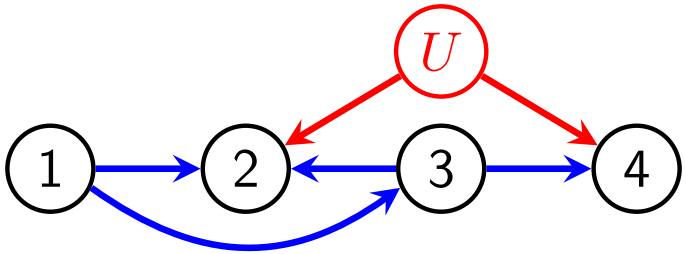
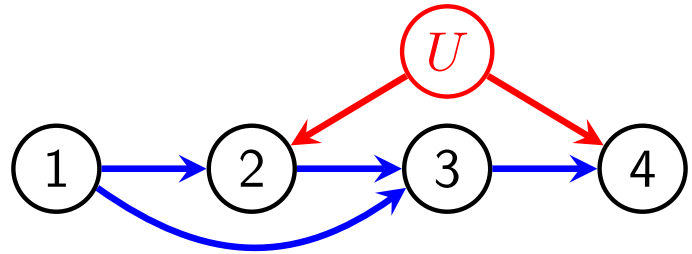
Geometry of Graphical Model Selection

Robin Evans, University of Oxford

ICMS Workshop
7th April 2017

Some Graphical Models





Model on left satisfies $X_1 \perp\!\!\!\perp X_4 \mid X_3$.

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

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

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

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?

Statistics for the Lazy

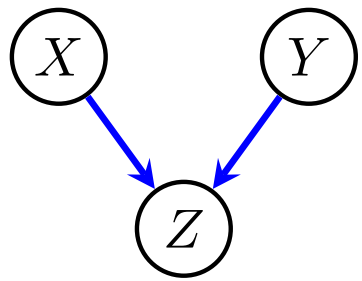
High-level view:

See, e.g., Uhler et al. (2013) for how this applies to graphical models.

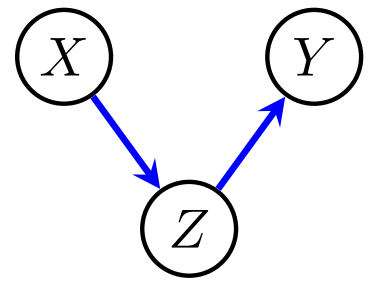
Directed Acyclic Graphs

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

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



$$\rho_{xy} = 0$$



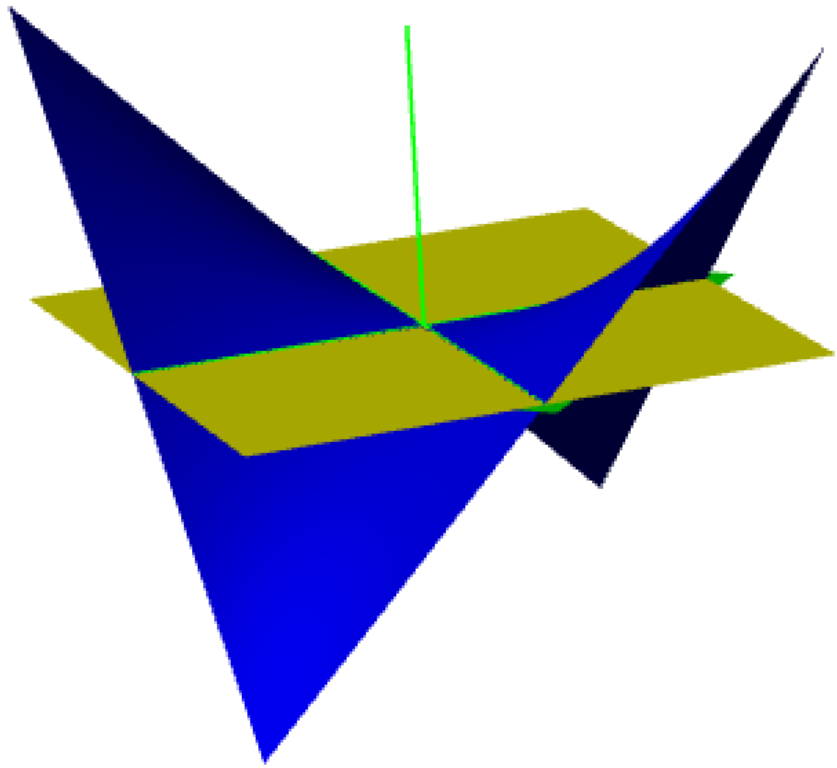
$$\rho_{xy \cdot z} = 0$$

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.

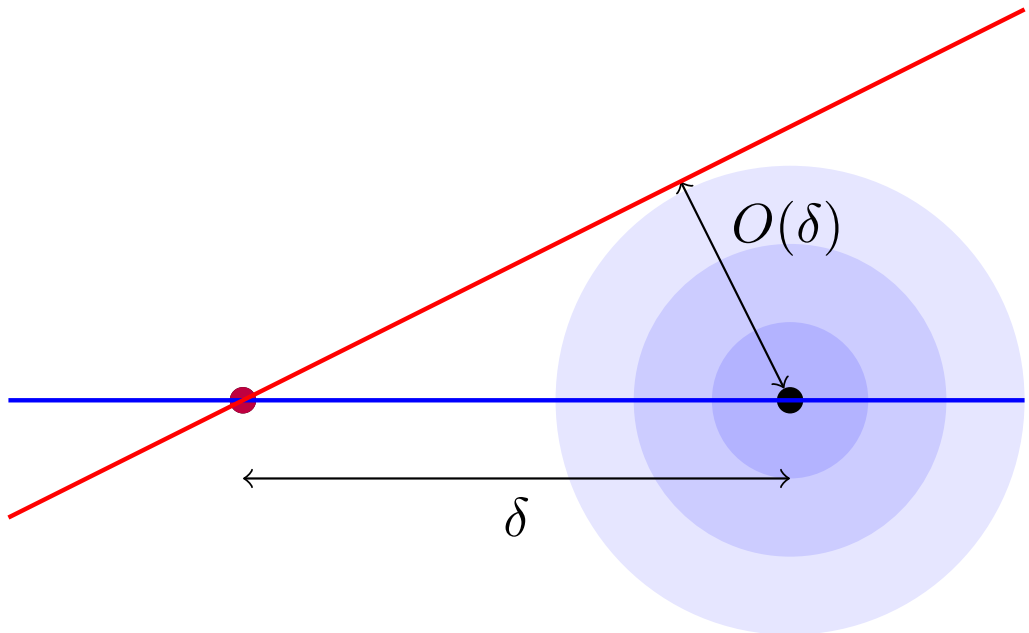
Marginal and Conditional Independence



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

A Picture

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

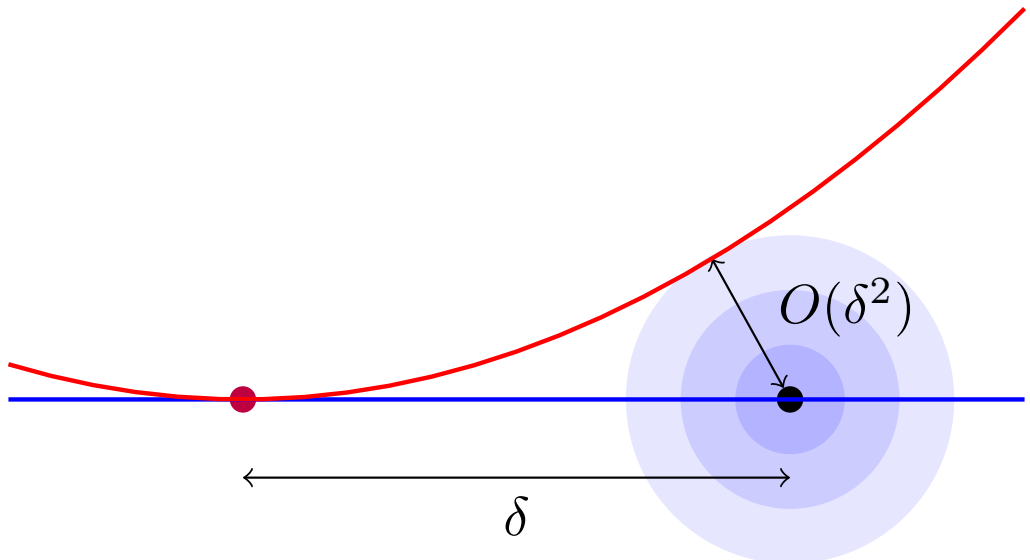


We intuitively expect to have power to test against alternatives long as our effect sizes are of order $n^{-1/2}$.

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

A Slightly Different Picture

Suppose we have two slightly different sub-models:



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!

Tangent Cones

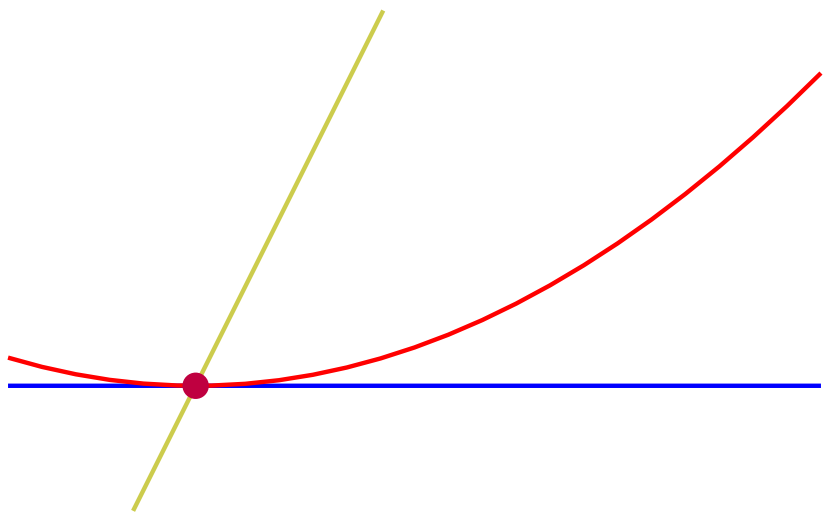
Definition

Let $\Theta \subseteq \mathbb{R}^d$ be a parameter space containing θ_0. The **tangent cone** of Θ (at θ_0), $\mathrm{TC}_{\theta_0}(\Theta)$ is the set of vectors of the form

$$\lim_n \alpha_n(\theta_n - \theta_0),$$

for sequences $\theta_n \to \theta_0$.

For regular (differentiable) models this a vector space (the **tangent space**) is just the derivative of Θ at θ_0.



Overlap

Definition

Say that two models Θ_1 and Θ_2 **overlap** if there is a point $\theta \in \Theta_1 \cap \Theta_2$ such that $\mathrm{TC}_\theta(\Theta_1) = \mathrm{TC}_\theta(\Theta_2)$.

Example. Two directed Gaussian graphical models overlap at any diagonal Σ if they have the same skeleton.

Further, if they have different skeletons then they overlap almost nowhere.

Gaussian Graphical Models

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

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

For $X \perp\!\!\!\perp Y$, we can have any η, ε, and as they $\to 0$ we see that the tangent cone is

$$\mathrm{TC}_I(X \perp\!\!\!\perp Y) = \langle \delta_{13} + \delta_{31}, \delta_{23} + \delta_{32} \rangle.$$

where δ_{ij} is matrix with (i,j)th entry 1 and otherwise 0.

The model $X \perp\!\!\!\perp Y \mid Z$ is similar but we need $\rho_{xy} = \rho_{xz}\rho_{yz} = \varepsilon\eta$. However in the limit we still get

$$\mathrm{TC}_I(X \perp\!\!\!\perp Y \mid Z) = \langle \delta_{13} + \delta_{31}, \delta_{23} + \delta_{32} \rangle.$$

Gaussian Graphical Models

For convenience write

$$\Lambda_{ij} = \{\alpha(\delta_{ij} + \delta_{ji}), \alpha \in \mathbb{R}\}.$$

Consider a class of Gaussian graphical models that may be defined by a single independence $X_i \perp\!\!\!\perp X_j \mid X_{S_{ij}}$ whenever i and j are not adjacent in $\mathcal{G}$.

Examples. Maximal ancestral graphs, directed acyclic graphs, LWF chain graphs, MR chain graphs...

Theorem

Whenever $\mathcal{G}$ and $\mathcal{H}$ have the same skeleton, the associated Gaussian graphical models overlap.

The tangent space at any diagonal covariance matrix is

$$\mathrm{TC}_I(\mathcal{G}) \equiv \bigoplus_{i \sim j} \Lambda_{ij}.$$

Statistical Consequences of Overlap

Suppose that models $\Theta_1, \Theta_2 \subseteq \Theta$ overlap (and are regular) at θ_0.

Consider a sequence of local 'alternatives' in Θ_1 of the form

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

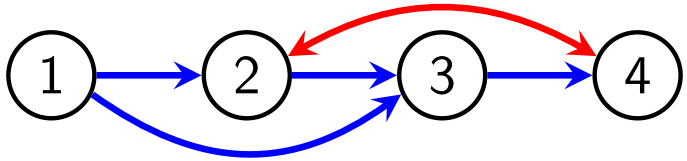
then:

- 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 $\gamma \leq 1/4$.

So if effect size is halved, we need 16 times as much data to be sure we pick Θ_1 over Θ_2!

This helps to explain the problems with nested models.

Gaussian Verma Constraint



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

$$\begin{aligned} f(R) &= \rho_{14} - \rho_{14}\rho_{12}^2 - \rho_{14}\rho_{23}^2 + \rho_{14}\rho_{12}\rho_{13}\rho_{23} \\ &\qquad - \rho_{13}\rho_{34} + \rho_{13}\rho_{23}\rho_{24} + \rho_{12}^2\rho_{13}\rho_{34} - \rho_{12}\rho_{13}^2\rho_{24} \\ &= \rho_{14} - \rho_{13}\rho_{34} - \rho_{14}\rho_{12}^2 - \rho_{14}\rho_{23}^2 + \rho_{13}\rho_{23}\rho_{24} + O(\varepsilon^4) \\ &= \rho_{14} - \rho_{13}\rho_{34} + O(\varepsilon^3) \\ &= \rho_{14} + O(\varepsilon^2). \end{aligned}$$

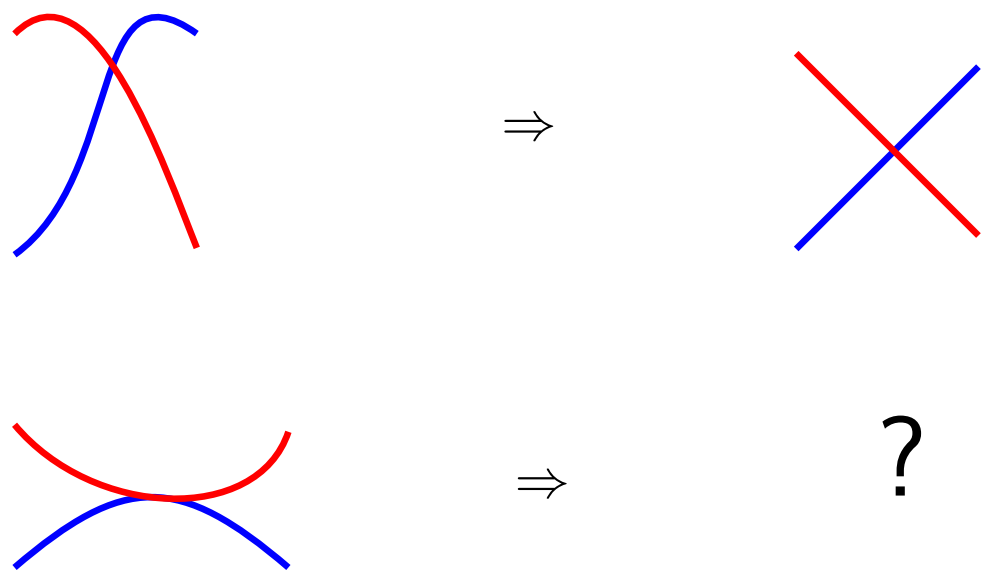
Model is not only locally linearly equivalent to the model of $X_1 \perp\!\!\!\perp X_4$, but also *quadratically* equivalent to the model $X_1 \perp\!\!\!\perp X_4 \mid X_3$.

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

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.

Lack of Convexity

Example. For usual undirected graphical models, one can solve the convex program:

$$\text{minimize}_K \qquad \log\det K + \text{tr}(KS) + \lambda \sum_{i,j} |k_{ij}|.$$

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

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

[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.

Time Series

As a non-graphical example, time series models also experience overlap:

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.

Discrete Models

For discrete models it is more helpful to work with a log-linear parameterization, e.g.:

$$\log P(X_V = x_V) = \sum_{A \subseteq V} (-1)^{\|x_A\|} \lambda_A.$$

We can also define **marginal** log-linear parameters in the same way with reference to a particular margin:

$$\log P(X_M = x_M) = \sum_{A \subseteq M} (-1)^{\|x_A\|} \lambda_A^M.$$

Then, starting at a uniform distribution $\boldsymbol{\lambda} = \mathbf{0}$, we will write the vector space spanned by λ_A as Λ_A.

If one has a model in which contains $\lambda_A = \varepsilon > 0$ with all other $\lambda_B = o(\varepsilon)$, then Λ_A is contained in the tangent cone of the model.

Importantly, all marginal parameters with the same effect A have the same derivative at $\boldsymbol{\lambda} = \mathbf{0}$.

Discrete Models

Proposition

Let λ_A^M, λ_A^L be marginal log-linear parameters. Then within an ε neighbourhood of the independence model,

$$\lambda_A^M = \lambda_A^L + O(\varepsilon^2).$$

As a consequence of this, the parameters give the same tangent space on the independence model.

Proof.

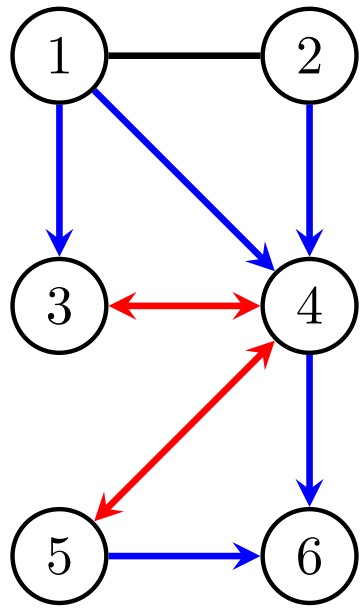
By adapting a proof from Evans (2015), one has

$$\lambda_A^M = \lambda_A^L + f(\lambda_m^M, \cdots, \lambda_M^M),$$

for a smooth function f which is zero whenever all but one of the arguments is zero. □

Discrete Models

One can define **ancestral graph models** using zeroes of marginal log-linear parameters (Evans and Richardson, 2014).



These generalize DAGs, undirected models, marginal independence models.

Discrete Ancestral Graphs

Proposition

For any two discrete ancestral graphs, either the models are identical or they do not overlap.

Proof.

If the models are distinct then either (WLOG):

- $i \sim j$ in $\mathcal{G}$ but not $\mathcal{H}$;
  In this case models with $\lambda_{ij} = \varepsilon$ and all other log-linear parameters zero are in $\mathcal{G}$ but not $\mathcal{H}$.
- $i - k - j$ a v-structure in $\mathcal{H}$ but not in $\mathcal{G}$;
  Then models with $\lambda_{ijk} = \varepsilon$ and all other log-linear parameters zero are in $\mathcal{G}$ but not $\mathcal{H}$.
- inducing path from i to j in $\mathcal{H}$ but not in $\mathcal{G}$.
  Similar to v-structure proof.

□

Imsets

Somewhat related to the previous proof, we can define a **characteristic imset** (Studený et al., 2010) for ancestral graphs as follows:

$$k_{\mathcal{G}} \equiv \sum_{H \in H(\mathcal{G})} \sum_{S \subseteq T} \delta_{H \cup T}.$$

Here $H(\mathcal{G})$ is the collection of 'heads' (and complete sets in undirected part).

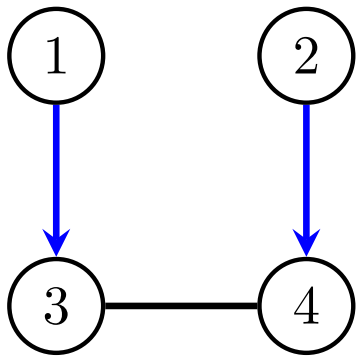
[Approximately, heads are bidirected-connected sets and tails are their parents.]

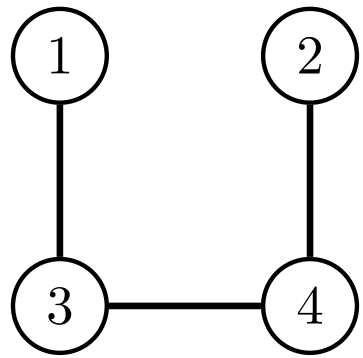
Theorem

For MAGs $\mathcal{G}$ and $\mathcal{H}$, we have $k_{\mathcal{G}} = k_{\mathcal{H}}$ if and only if $\mathcal{G}$ and $\mathcal{H}$ are Markov equivalent graphs.

Chain Graphs

LWF chain graphs do not satisfy the same property, and distinct models may overlap.





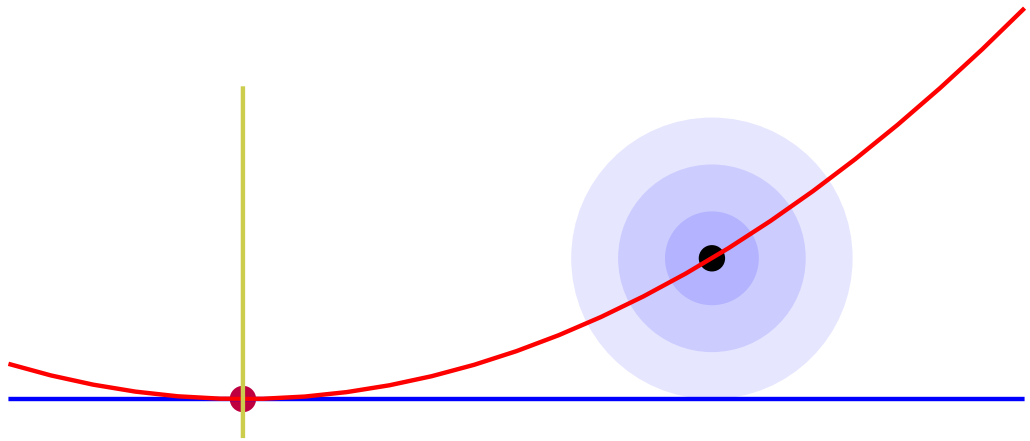
$$X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3$$
$$X_2 \perp\!\!\!\perp X_3 \mid X_1, X_4$$
$$X_1 \perp\!\!\!\perp X_2$$

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

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

Towards Methods

An idea: can we **use** the fact that other marginal log-linear parameters are 'close', to deduce the correct imset 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 previous Theorem to reconstruct the MAG equivalence class (using essentially the same algorithm as FCI).

Penalised Selection

Consider the usual Lasso approach:

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

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 MAG model for $\mathcal{G}$.
Then if $\frac{1}{4} < c < \frac{1}{2}$, the lasso will be consistent for the imset representation of $\mathcal{G}$.

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

Classes of Models

Class	Difficulty	Reference
undirected	fast	Meinshausen and Bühlmann (2006)
bidirected	fast	Zwiernik et al. (2016)
directed	hard	Chickering (1996)
ancestral	⋮	
nested	harder?	Shpitser et al. (2013)

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.
- This is manifested in the tangent cones of the models.
- This perspective can be used to learn about what makes models similar / different.

Thank you!

References

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

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

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

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

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.

Studený, Hemmecke and Lindner. Characteristic imset: a simple algebraic representative of a Bayesian network structure. *PGM*. 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.

Heads and Tails

Let $\mathcal{G}$ be an ADMG with vertices V. Say that $H \subseteq V$ is a **head** if there is some set S of the form:

$$S \equiv \mathrm{dis}_{\mathcal{G}_{\mathrm{an}(H)}}(\mathrm{an}_{\mathcal{G}}(H))$$

such that H is the set of nodes in S that does not have any descendants in S.

The **tail** of H is the set $T \equiv (S \setminus H) \cup \mathrm{pa}_{\mathcal{G}}(S)$.

See Evans and Richardson (2013) for full details.