Parametrizations of Discrete Graphical Models

Robin J. Evans
`www.stat.washington.edu/∼rje42`

9th December 2010

Outline

Acknowledgements

This work joint with Thomas Richardson.

Thanks to the other committee members for their time and advice: Adrian Dobra, Brian Flaherty, Peter Hoff, Steffen Lauritzen and James Robins.

Thanks also to Tamás Rudas for discussions and his helpful comments.

Set Up

Random variables $(X_i)_{i=1}^n$ taking values in $\times_{i=1}^n(\mathfrak{X}_i)$.

Finite discrete space, so write $\mathfrak{X}_v = \{0, 1, \ldots, |\mathfrak{X}_v| - 1\}$.

Positive probability measure P.

Notational shortcuts (example with $n = 3$):

$$p_{010} \equiv P(X_1 = 0, X_2 = 1, X_3 = 0)$$
$$p_{0\cdot 0} \equiv P(X_1 = 0, X_3 = 0).$$

The graph vertex i used synonymously with random variable X_i.

Graphical Models

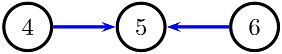
Intuitive visual representation of conditional independences.

Relationship between 1 and 3 is entirely mediated by 2.

$$1 \perp\!\!\!\perp 3 \mid 2.$$

4 and 6 both affect 5, but no direct relationship. Marginally 4 and 6 independent:

$$4 \perp\!\!\!\perp 6.$$



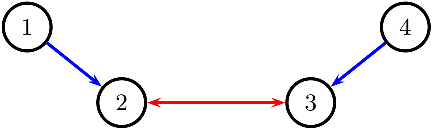
However, *conditionally* on 5, they may become dependent.

$$4 \not\perp\!\!\!\perp 6 \mid 5.$$

Why use graphical models?

- Parsimony: saturated model has $2^n - 1$ parameters in binary case.
- Tractable model search space.
- Efficient inference.
- Intuition.
- Powerful language for reading off conditional independence.
- Causal interpretation.

Motivating Example



This graph represents the independences $1 \perp\!\!\!\perp 3, 4$ and $4 \perp\!\!\!\perp 1, 2$.

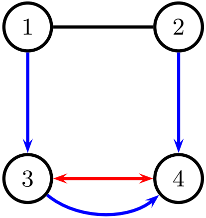
Parameters of Richardson (2009):

$$P(X_1 = 0) \quad P(X_4 = 0) \quad P(X_2 = 0 \mid X_1) \quad P(X_3 = 0 \mid X_4)$$
$$P(X_2 = 0,\ X_3 = 0 \mid X_1, X_4).$$

Alternatively, could choose conditional odds ratio between X_2 and X_3 for last parameter (noticed by Madigan for AMP Chain Graphs).

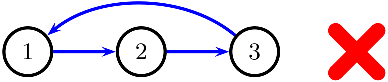
Variation independence means that prior specification, parameter interpretation, MCMC, regression modelling all become easier.

Euphonious Graphs

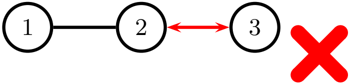


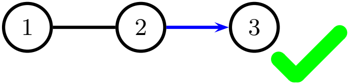
We work with mixed graphs, which have 3 types of edges.

No directed cycles:



No arrow head adjacent to an undirected edge:



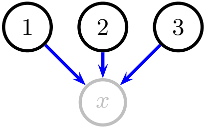


We call this class *Mixed Euphonious Graphs* (MEGs). Can be thought of as an undirected graph with conditional ADMG.

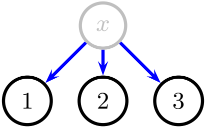
MEGs include all undirected graphs, DAGs, bidirected graphs, ancestral graphs and ADMGs. They are equivalent to summary graphs (Wermuth and Cox, 2000).

Definitions

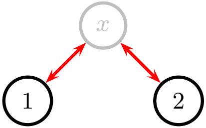
parents $\mathrm{pa}_{\mathcal{G}}(x)$



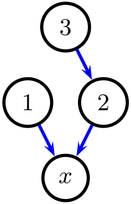
children $\mathrm{ch}_{\mathcal{G}}(x)$



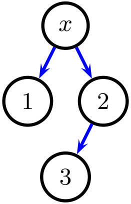
spouses $\mathrm{sp}_{\mathcal{G}}(x)$



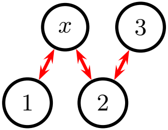
ancestors $\mathrm{an}_{\mathcal{G}}(x)$



descendants $\mathrm{de}_{\mathcal{G}}(x)$



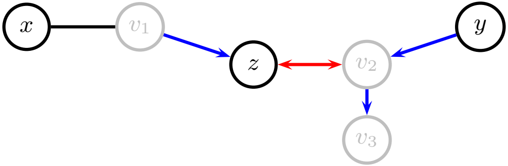
district $\mathrm{dis}_{\mathcal{G}}(x)$



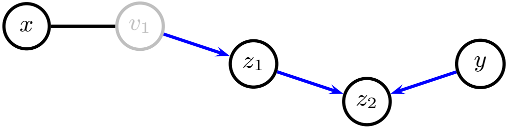
m-separation

Two vertices x and y are m-separated by a set Z if all paths from x to y are blocked by Z.

Either: at least one collider is not conditioned upon, and nor are any of its descendants:



Or: at least one non-collider is conditioned upon:

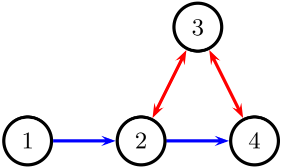


m-separation extends to sets X and Y if every $x \in X$ and $y \in Y$ are m-separated.

Global Markov Property

Let P be a distribution over the vertices of $\mathcal{G}$. The global Markov property (GMP) for MEGs states that

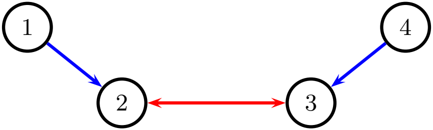
$$X \text{ m-separated from } Y \text{ by } Z \quad \Longrightarrow \quad X \perp\!\!\!\perp Y \mid Z \; [P]$$



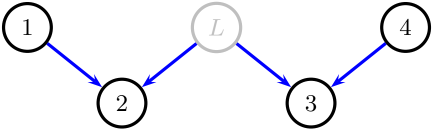
Here $1 \perp\!\!\!\perp 4 \mid 2$ and $1 \perp\!\!\!\perp 3$.

This global Markov property generalizes those of DAGs, undirected graphs and bidirected graphs.

What's wrong with a DAG?



The bidirected edge 'represents' the presence of a latent variable. So why not use a DAG with latent variables?



How do we model the latent variable?
May be an abstract concept.

Latent variable models are not curved exponential families.

Ordinary DAGs not closed under marginalization.

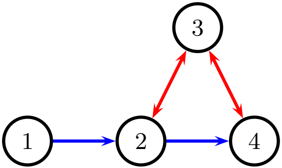
Parametrizing MEGs

Richardson (2009) gives a parametrization of discrete distributions obeying the global Markov property for an ADMG. This extends easily to MEGs.

Define a *head* H to be any set of vertices which is

(i) connected by $\leftrightarrow$-arrows in $\mathrm{an}_{\mathcal{G}}\, H$;

(ii) *barren*: no element of H is an ancestor of any other.



H	1	2	3	23	4	34
T	$\emptyset$	1	$\emptyset$	1	2	12

For a head H, the corresponding *tail* is the set of ancestors which are connected to H by paths of colliders in $\mathrm{an}_{\mathcal{G}}\, H$. The tail is the Markov blanket for H in $\mathrm{an}_{\mathcal{G}}\, H$.

Generalized Möbius Parameters

The distributions obeying the GMP are parametrized by the probabilities

$$P(X_H = \boldsymbol{i}_H \mid X_T = \boldsymbol{i}_T),$$

where H is a head and T its tail.

In the binary case we write

$$q_{H|T}^{\boldsymbol{i}_T} \equiv P(X_H = 0 \mid X_T = \boldsymbol{i}_T)$$

for the *generalized Möbius parameters*. (Ordinary Möbius parameters $q_A = P(X_A = 0)$)

Mixed Graph	**DAG**	**Bidirected**
m-separation	d-separation	(m-separation)
head	single vertex $\{v\}$	connected set
tail	parents $\mathrm{pa}(v)$	always empty
gen. Möbius params	$P(X_v \mid X_{\mathrm{pa}(v)})$	Möbius parameters

Disadvantages of Generalized Möbius Parameters

Variation dependence. Causes problems with fitting. Frechét bounds give some constraints:

$$\max\{0, q_1 + q_2 - 1\} \leq q_{12} \leq \min\{q_1, q_2\}.$$

Structure of graph creates less obvious inequalities.

Prior selection requires extra thought.

Correlation. Similarly, can interfere with MCMC procedures and likelihood fitting.

No obvious way to create **Parsimonious Submodels**.

Intuition. Related to variation dependence; rather subjective claim.

Marginal Log-Linear Parameters

For $L \subseteq M \subseteq V$ and $\boldsymbol{i}_L \in \mathfrak{X}_L$, define

$$\lambda_L^M(\boldsymbol{i}_L) = \frac{1}{|\mathfrak{X}_M|} \sum_{\boldsymbol{j}_M \in \mathfrak{X}_M} \log p_{\boldsymbol{j}_M} \prod_{v \in L} \left(|\mathfrak{X}_v| \mathbb{I}_{\{i_v = j_v\}} - 1 \right).$$

Some examples in the binary case:

$$\lambda_1^1(0) = \frac{1}{2} \log \frac{p_{0\cdot\cdot}}{p_{1\cdot\cdot}} \qquad \lambda_{123}^{123}(0,0,0) = \frac{1}{8} \log \frac{p_{000}\, p_{110}\, p_{101}\, p_{011}}{p_{100}\, p_{010}\, p_{001}\, p_{111}}$$

$$\lambda_1^{12}(0) = \frac{1}{4} \log \frac{p_{00\cdot}\, p_{01\cdot}}{p_{10\cdot}\, p_{11\cdot}} \qquad \lambda_{12}^{12}(0,0) = \frac{1}{4} \log \frac{p_{00\cdot}\, p_{11\cdot}}{p_{10\cdot}\, p_{01\cdot}}$$

And the trinary:

$$\lambda_1^1(0) = \frac{1}{3} \log \frac{p_{0\cdot\cdot}^2}{p_{1\cdot\cdot}\, p_{2\cdot\cdot}} \qquad \lambda_{12}^{12}(0,0) = \frac{1}{9} \log \frac{p_{00\cdot}^4\, p_{11\cdot}\, p_{12\cdot}\, p_{21\cdot}\, p_{22\cdot}}{p_{10\cdot}^2\, p_{01\cdot}^2\, p_{20\cdot}^2\, p_{02\cdot}^2}$$

Parametrizations

Bergsma and Rudas (2002) introduce a class of parameters for discrete probability distributions.

Take a set of *margins*

$$\mathbb{M} = \{M_1, \dots, M_k\}, \qquad M_i \subseteq V \text{ for each } i$$

ordered so that $M_j \nsubseteq M_i$ for $j > i$, and $M_k = V$. Let

$$\mathbb{L}_i = \mathscr{P}(M_i) \setminus (\mathbb{L}_1 \cup \cdots \cup \mathbb{L}_{i-1}).$$

This collection of margins $\mathbb{M}$ and sets of *effects* $\mathbb{L}_i$ is a *complete* and *hierarchical* parametrization.

Different pieces of the same story. Example for $V = \{1, 2, 3\}$:

M	$\mathbb{L}$
$\{1\}$	$\{1\}$
$\{1,2\}$	$\{2\}$, $\{1,2\}$
$\{3\}$	$\{3\}$
$\{1,3\}$	$\{1,3\}$
$\{1,2,3\}$	$\{2,3\}$, $\{1,2,3\}$

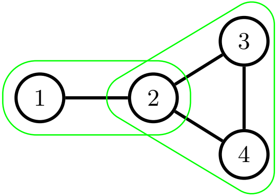
Special Cases

The usual log-linear parameters are of the form

$$\lambda_L^V \qquad L \subseteq V.$$

To parametrize distributions over an undirected graph $\mathcal{G}$ with complete subsets $\mathcal{C}(\mathcal{G})$, we can use

$$\lambda_L^V \qquad L \in \mathcal{C}(\mathcal{G}).$$



The multivariate logistic parameters of Glonek and McCullagh (1995) are

$$\lambda_L^L \qquad L \subseteq V.$$

Conditional Independence

Clearly $\lambda_{12}^{12} = 0$ if and only if $1 \perp\!\!\!\perp 2$.

Theorem (Rudas et al., 2010, Lemma 1)

Let $\mathbb{L} = \mathscr{P}(A \cup B \cup C) \setminus (\mathscr{P}(A \cup C) \cup \mathscr{P}(B \cup C))$. Then

$$A \perp\!\!\!\perp B \mid C \quad \iff \quad \lambda_D^{ABC} = 0 \text{ for all } D \in \mathbb{L}.$$

So $\lambda_L^M = 0$ if *any* two proper subsets of L are independent conditional on the rest of M.

Example: for $1 \perp\!\!\!\perp 3, 4 \mid 2$, we need

$$\lambda_{13}^{1234} = \lambda_{14}^{1234} = \lambda_{134}^{1234} = \lambda_{123}^{1234} = \lambda_{124}^{1234} = \lambda_{1234}^{1234} = 0.$$

The Ingenuous Parametrization

Intuition: given lower dimensional margins, λ_A^{HT} for $H \subseteq A \subseteq H \cup T$ parametrizes $H|T$.

Example:

$$\lambda_1^{12}(0) + \lambda_{12}^{12}(0,0) = \frac{1}{2} \log \frac{p_{00\cdot}}{p_{10\cdot}}$$

$$\lambda_1^{12}(0) - \lambda_{12}^{12}(0,0) = \frac{1}{2} \log \frac{p_{01\cdot}}{p_{11\cdot}}$$

For a MEG $\mathcal{G}$, call the heads $H_1, H_2, \dots$ and their tails $T_1, T_2, \dots$. Then set

$$M_i = H_i \cup T_i \qquad \mathbb{L}_i = \{A \mid H_i \subseteq A \subseteq H_i \cup T_i\}.$$

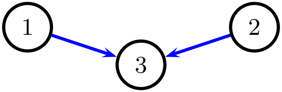
Call this the *ingenuous parametrization*.

Theorem (Evans, 2010)

The ingenuous parameters for a MEG $\mathcal{G}$ parametrize all distributions obeying the global Markov property with respect to $\mathcal{G}$.

Completion

The previous result does not guarantee us a 'nice' parametrization in the Bergsma and Rudas framework.



M	$\mathbb{L}$
$\{1\}$	$\{1\}$
$\{2\}$	$\{2\}$
$\{1, 2, 3\}$	$\{3\}$, $\{1, 3\}$, $\{2, 3\}$, $\{1, 2, 3\}$

This is not a complete parametrization. We could add in the λ_{12}^{123}, but this parameter makes no sense in the context of the model.

Instead, $\lambda_{12}^{12} = 0$ under this model.

Incomplete Parametrizations

Main results from Bergsma and Rudas (2002) rely on hierarchical and complete parametrization. We can *complete* by adding in missing effects.

Clearly additional parameters determined by the ingenuous parameters and consequences of model.

But complicated functional dependence is not useful. Call a completion *sound* if it is hierarchical and additional parameters are identically zero under the model.

Lemma

The ingenuous parametrization always has a sound completion.

Each MEG model defines a curved exponential family.

Asymptotics are regular: χ^2-tests have expected behaviour.

Variation Independence

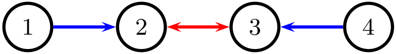
Parameters $\theta_1, \dots, \theta_k$ taking values in $\Theta_1, \dots, \Theta_k$ are *variation independent* if $(\theta_1, \dots, \theta_k)$ takes any value in $\Theta_1 \times \cdots \times \Theta_k$.

Bergsma and Rudas (2002) show that any complete and hierarchical MLL parametrization is variation independent if and only if it satisfies a condition called *ordered decomposability*.

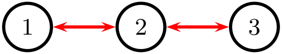
Theorem (Evans, 2010)

The ingenuous parametrization for a MEG $\mathcal{G}$ is variation independent if and only if $\mathcal{G}$ has no heads of size greater than or equal to 3.

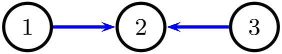
Note that this includes all DAGs and, for example,



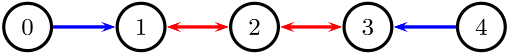
Cases We Miss



Cases We Miss

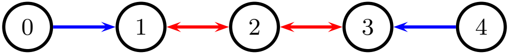


Cases We Miss



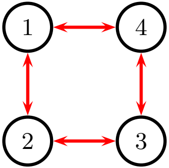
Bidirected 5-chain has no known variation independent parametrization.

Cases We Miss



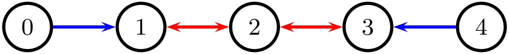
Bidirected 5-chain has no known variation independent parametrization.

The bidirected 4-cycle *does* have a variation independent parametrization.



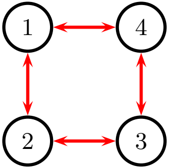
This model corresponds to $1 \perp\!\!\!\perp 3$ and $2 \perp\!\!\!\perp 4$.

Cases We Miss



Bidirected 5-chain has no known variation independent parametrization.

The bidirected 4-cycle *does* have a variation independent parametrization.



This model corresponds to $1 \perp\!\!\!\perp 3$ and $2 \perp\!\!\!\perp 4$.

Take margins $\{1,3\}$, $\{2,4\}$ and $\{1,2,3,4\}$, and set $\lambda_{13}^{13} = \lambda_{24}^{24} = 0$.

This 'disconnected sets' approach does not seem to generalize.

Evaluating Probabilities

In general it is hard to recover probabilities from marginal log-linear parameters.

Can use Iterative Proportional Fitting (IPF). May be slow to evaluate the likelihood many times (e.g. MCMC).

Try to solve directly (binary case):

$$\lambda_1^1(0) = \frac{1}{2} \log \frac{p_{0\cdot\cdot}}{1 - p_{0\cdot\cdot}}.$$

$$\lambda_2^{12}(0) + \lambda_{12}^{12}(0,0) = \frac{1}{2} \log \frac{p_{00\cdot}}{p_{01\cdot}} = \frac{1}{2} \log \frac{p_{00\cdot}}{p_{0\cdot\cdot} - p_{00\cdot}}.$$

This can be continued for higher order margins! Does not seem to generalize easily to non-binary variables.

Based on approach of Qaqish and Ivanova (2006) for multivariate logistic parameters.

Detecting Invalid Parameters

In higher dimensional cases we may have to avoid variation dependence ($q_A = P(X_A = 0)$):

$$
\begin{aligned}
&\exp(8\lambda^{123}_{123}) = \frac{p_{000}\, p_{110}\, p_{101}\, p_{011}}{p_{100}\, p_{010}\, p_{001}\, p_{111}} \\
&= \frac{q_{123}(q_3 - q_{13} - q_{23} + q_{123})(q_2 - q_{12} - q_{23} + q_{123})(q_1 - q_{12} - q_{13} + q_{123})}{(q_{23} - q_{123})(q_{23} - q_{123})(q_{23} - q_{123})(1 - q_1 - q_2 - q_3 + q_{12} + q_{13} + q_{23} - q_{123})} \\
&= \frac{\prod_i (q_{123} - u_i)}{\prod_i (l_i - q_{123})} \equiv f(q_{123})
\end{aligned}
$$

Probabilities must be positive, so

$$\max_i u_i \leq q_{123} \leq \min_i l_i.$$

These are the Frechét bounds. Provided $\max_i u_i < \min_i l_i$, we have a unique solution.

In other cases, we have chosen invalid parameter values.

Summary

We have:

- provided a new parametrization of discrete models based on MEGs;
- linked MEGs (and hence ADMGs) to Bergsma and Rudas' framework;
- shown precisely when it is variation independent;
- given a method for non-iterative likelihood evaluation in binary case;
- seen how to use this method to see whether parameters are valid.

Further Work

Investigate practical ways of using this parametrization for parsimonious and penalized modelling.

Considerations of Markov equivalence.

Implement tools for

(i) Bayesian analysis of these models;

(ii) regression models.

Fuller characterization of variation independence for graphical models.

Chain and lattice models under stationarity.

Apply parametrization to non-Markov graphical models.

Thank you!

Ordered Decomposability

Incomparable subsets $M_1, \ldots, M_k$ of V are *decomposable* if for each $i = 3, \ldots, k$, there exists $j_i < i$ such that

$$\left(\bigcup_{l=1}^{i-1} M_l \right) \cap M_i = M_{j_i} \cap M_i.$$

This is the *running intersection property*.

(Possibly comparable) subsets $M_1, \ldots, M_k$ of V are *ordered decomposable* if they are hierarchical and for each $i = 3, \ldots, k$, the inclusion maximal elements of $\{M_1, \ldots, M_i\}$ are decomposable.

A parametrization $\mathbb{M}$ and $\mathbb{L}_1, \ldots, \mathbb{L}_k$ is ordered decomposable if there is an ordering on the margins $\mathbb{M}$ which is both hierarchical and ordered decomposable.

Fitting with Generalized Möbius Parameters

Let $D_1, \dots D_d$ be districts of $\mathcal{G}$, and $O_i = \{v : i_v = 0\} \cap D_i$. Then (Evans and Richardson, 2010)

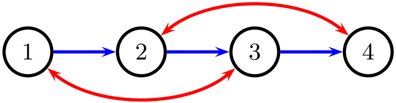
$$P\left(X_V = \boldsymbol{i}_V\right) = \prod_{i=1}^{d} \sum_{C : O_i \subseteq C \subseteq D_i} (-1)^{|C \setminus O_i|} \prod_{H \in [C]_{\mathcal{G}}} q_{H|T}^{(\boldsymbol{i}_T)}.$$

Thus $\boldsymbol{p} = M \exp(P \log \boldsymbol{q})$ for some matrices M and P.

Parameters are variation dependent, so we must take care when performing a search.

Fitting is performed by considering each vertex in turn, and ensuring that probabilities stay positive.

Verma Constraints



There is no edge between 1 and 4.

1 and 4 cannot be m-separated by any conditioning set.

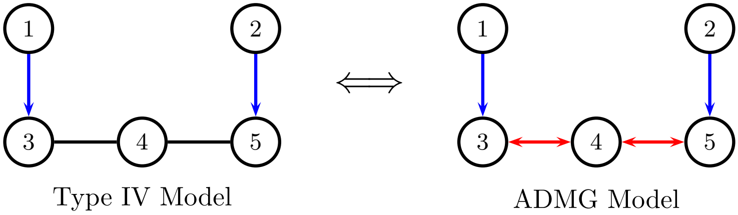
There *is* a non-parametric constraint:

$$\frac{\partial}{\partial x_1} \sum_{x_2} p(x_4 \mid x_1, x_2, x_3) \cdot p(x_2 \mid x_1) = 0.$$

This corresponds to $1 \perp\!\!\!\perp 4$ after intervention on 3.

Type IV Models

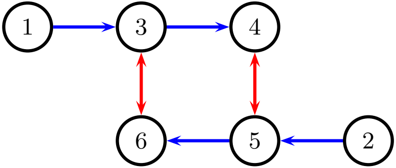
Block recursive Markov models (type IV chain graph models) are a special case of ADMGs.



Rudas et al. (2010) parametrize Type IV models. Their approach coincides with ours for the special case of DAGs, but not in general.

ADMGs are a bigger class

ADMGs are a strictly larger class than Type IV models.

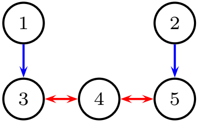


This graph corresponds to the factorization

$$p(x_{123456}) = p(x_1)\, p(x_2)\, p(x_{36} \mid x_{15})\, p(x_{45} \mid x_{23}).$$

Note that we cannot order the heads $\{3, 6\}$ and $\{4, 5\}$ so that tails always precede heads. There is no obvious generalization of well-ordering.

Different Margins



M (RBN)	M (ing.)	$\mathbb{L}$
1	1	$\{1\}$
2	2	$\{2\}$
$1,2,3$	$1,3$	$\{3\}$,$\{1,3\}$
$1,2,4$	4	$\{4\}$
$1,2,3,4$	$1,3,4$	$\{3,4\}$, $\{1,3,4\}$
$1,2,5$	$2,5$	$\{5\}$, $\{2,5\}$
$1,2,4,5$	$2,4,5$	$\{4,5\}$, $\{2,4,5\}$
$1,2,3,4,5$	$1,2,3,4,5$	$\{3,4,5\}$, $\{1,3,4,5\}$, $\{2,3,4,5\}$, $\{1,2,3,4,5\}$

However these parameters are equal under the model! So variation independence properties are the same for both.

The ingenuous margins would seem to be the smallest possible choices of M.

A General Result on Equality of Parameters

Lemma

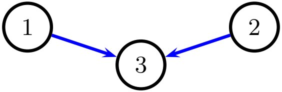
Suppose that $X_A \perp\!\!\!\perp X_B \mid X_C$, and A is non-empty. Then for any $D \subseteq C$,

$$\lambda^{ABC}_{AD} = \lambda^{AC}_{AD}.$$

This shows, for example, that the Rudas et al. (2006) parameters are equal to the Rudas et al. (2010) parameters for DAGs.

I have not yet found a good example where this lemma is *necessary* for proving variation independence.

Smaller Margins



Rudas et al (2006) parameter: $\lambda_2^{12}(0) = \frac{1}{4} \log \frac{p_{00\cdot}\, p_{10\cdot}}{p_{01\cdot}\, p_{11\cdot}}$.

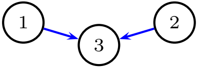
Ingenuous and Rudas et al (2010): $\lambda_2^{2}(0) = \frac{1}{2} \log \frac{p_{\cdot 0 \cdot}}{p_{\cdot 1 \cdot}}$.

The two parameters are the same under the model.

We get the MLE of λ_2^2 (and λ_2^{12}) just by plugging in empirical probabilities $\hat{p}$ to λ_2^2:

$$\hat{\lambda}_2^{12} = \hat{\lambda}_2^{2} = \frac{1}{2} \log \frac{\hat{p}_{\cdot 0 \cdot}}{\hat{p}_{\cdot 1 \cdot}}.$$

Smaller Margins (2)



$$\hat{\lambda}_2^{12} = \frac{1}{4} \log \frac{\hat{p}_{00\cdot}\,\hat{p}_{10\cdot}}{\hat{p}_{01\cdot}\,\hat{p}_{11\cdot}}$$

is asymptotically unbiased for λ_2^{12}, but less stable under empirical distribution. The smallest of the probabilities controls the stability.

Sample size 1000, $p_{0\cdot\cdot} = 0.1$ and $p_{\cdot 0\cdot} = 0.4$