Marginal Log-Linear Parameters
Lessons for the Continuous Case

Robin Evans, University of Oxford

Challenges for Categorical Data Analysis Workshop, LSE

1st November 2024

DEPARTMENT OF
STATISTICS

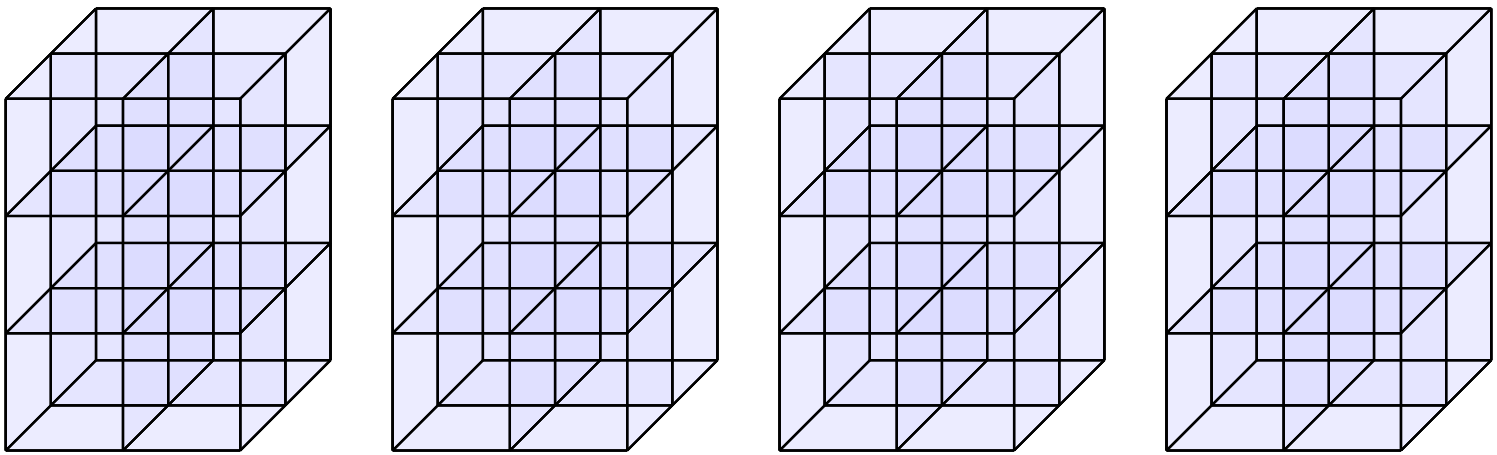
Outline

Contingency tables

Let $X_V = (X_v \in \mathfrak{X}_v : v \in V)$ take values in finite sets $\mathfrak{X}_V = \times_{v \in V} \mathfrak{X}_v$.

Suppose $V = \{1, 2, 3, 4\}$ with dimensions 2, 3, 2 and 4, and that we are given instances of X_V as a $2 \times 3 \times 2 \times 4$ contingency table.

We assume that the data are i.i.d. from a mass function $p(x_V)$.



How should we parameterize such a model?

With the mass function $p(x_V)$?

Log-linear parameters

For $A \subseteq V$, let x_A denote the subvector of x_V with entries in A.

The log-linear parameters associated with p are defined by

$$\log p(x_V) = \sum_{A \subseteq V} \lambda_A(x_A) \qquad \forall x_V \in \mathfrak{X}_V,$$

with appropriate identifiability constraints on the λ_A parameters.
(For example, $\lambda_A(x_A) = 0$ if $x_a = 0$ for any $a \in A$.)

If $\mathfrak{X}_v = \{0, 1\}$ for each $v \in V$, we can use an inverse Möbius to get:

$$\lambda_A = \lambda_A(1_A) = \sum_{x_A \in \mathfrak{X}_A} (-1)^{|A| - \sum |x_A|} \log p(x_A, 0_{V \setminus A}).$$

[Here $\sum |x_A|$ is the sum of the 1s in the vector x_A.]

Log-linear parameters

For example, if $V = \{1, 2\}$ we have:

$$\lambda_{\{1,2\}} = \log \frac{p_{00}p_{11}}{p_{10}p_{01}},$$

the log odds ratio between X_1 and X_2. [Here $p_{ab} = P(X_1 = a, X_2 = b)$.]

Similarly, If $V = \{1, 2, 3\}$ then:

$$\lambda_{\{1,2\}} = \log \frac{p_{000}p_{110}}{p_{100}p_{010}},$$

the conditional log odds ratio between X_1 and X_2 given $X_3 = 0$.

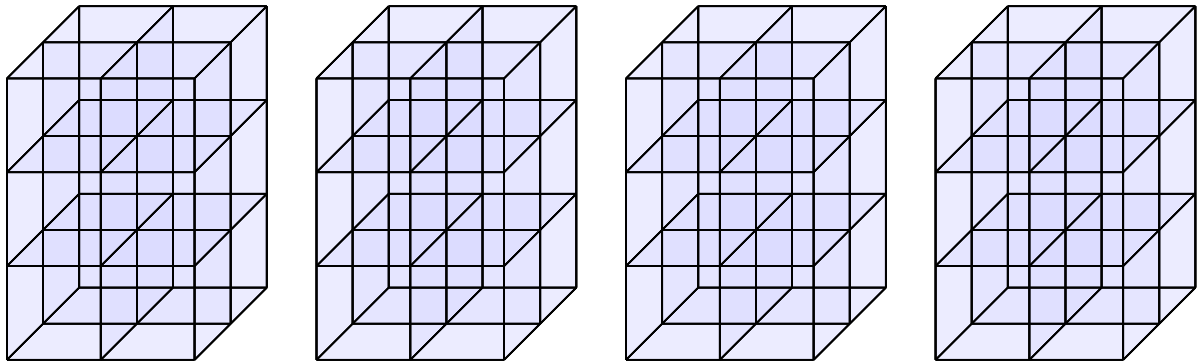
Log-linear parameters include:

- (conditional) odds ratios (as a linear transformation);
- three-way interactions.

The collection of log-linear parameters λ_A for $\emptyset \neq A \subseteq V$ constitutes a smooth parameterization of the set of positive distributions over $\mathfrak{X}_V$.

Margin of interest

Log-linear (LL) parameters are nice description and can model sparse data very efficiently; however, they do not allow us to easily control **marginal** structure. Every parameter is a function of **joint** probabilities.



Suppose that the entries in our table are sex, race, religiosity, and income quartile.

We might wish to set that sex and race are marginally independent; or that sex is independent of religiosity conditional upon race.

These **cannot** be (straightforwardly!) enforced using LL parameters.

Marginal log-linear parameters

Bergsma and Rudas (2002) introduced **marginal** log-linear (MLL) parameters.

These are log-linear parameters defined within a margin of V.
Includes the multivariate logistic models (Glonek and McCullagh, 1997) and ordinary log-linear parameters as special cases.

Denote log-linear parameter for **effect** A in margin $M \supseteq A$ by $\lambda_A^M(x_A)$.

Example

Suppose we have (X_1, X_2, X_3). We could parameterize these using (e.g.)

$\{1,2\}$	$\lambda_1^{12},\ \lambda_2^{12},\ \lambda_{12}^{12}$	$p(x_1, x_2)$
$\{2,3\}$	$\lambda_3^{23},\ \lambda_{23}^{23}$	$p(x_3 \mid x_2)$
$\{1,2,3\}$	$\lambda_{13}^{123},\ \lambda_{123}^{123}$	$\phi_{13\|2}(x_{13} \mid x_2)$

Bergsma and Rudas show that it is **necessary** to have exactly one parameter for each effect for a smooth parameterization.

Conditional independence

With log-linear parameters we can enforce conditional independences of the form $X_i \perp\!\!\!\perp X_j \mid X_{V\setminus\{i,j\}}$ by setting

$$\lambda_{ijC} = 0 \qquad \forall C \subseteq V \setminus \{i,j\}.$$

Similarly, with MLL parameters we can impose a general conditional independence of the form $X_i \perp\!\!\!\perp X_j \mid X_K$ by letting $M = K \cup \{i,j\}$ and choosing

$$\lambda^M_{ijC} = 0 \qquad \forall C \subseteq K.$$

Inspired by this, Rudas et al. (2010) define the set

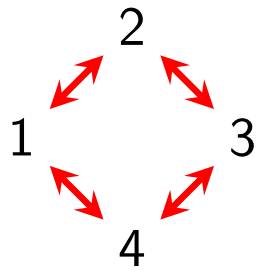
$$\mathbb{D}(i,j \mid K) = \{\{i,j\} \cup C : C \subseteq K\}.$$

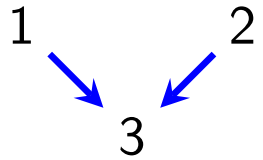
(See also Forcina et al., 2010)

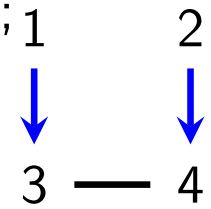
Conditional independence

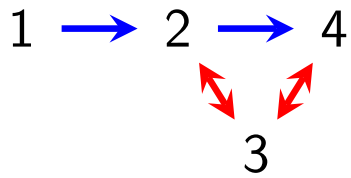
A (discrete) conditional independence model is smooth if it can be written as a collection of equality constraints among a smooth MLL parameterization.

- undirected graphs (ordinary LLPs);
- bidirected graphs (multivariate logistic parameters);
- directed acyclic graphs;
- Lauritzen-Wermuth-Frydenberg chain graphs (Lauritzen, 1996);
- maximal ancestral graphs (Evans and Richardson, 2013).
- ...

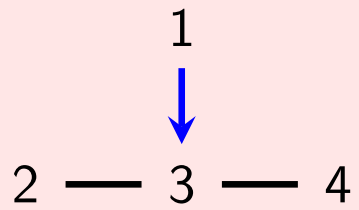




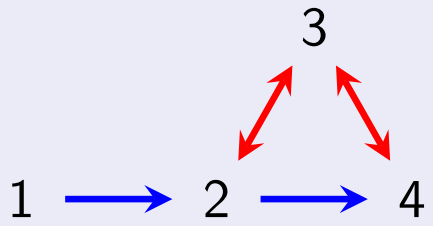




But not AMP chain graphs (Drton, 2009).



Conditional independence—example



This graph imposes that $X_1 \perp\!\!\!\perp X_3$ and $X_1 \perp\!\!\!\perp X_4 \mid X_2$.

Choose the margins $\{1,3\}$ and $\{1,2,4\}$ and the effects:

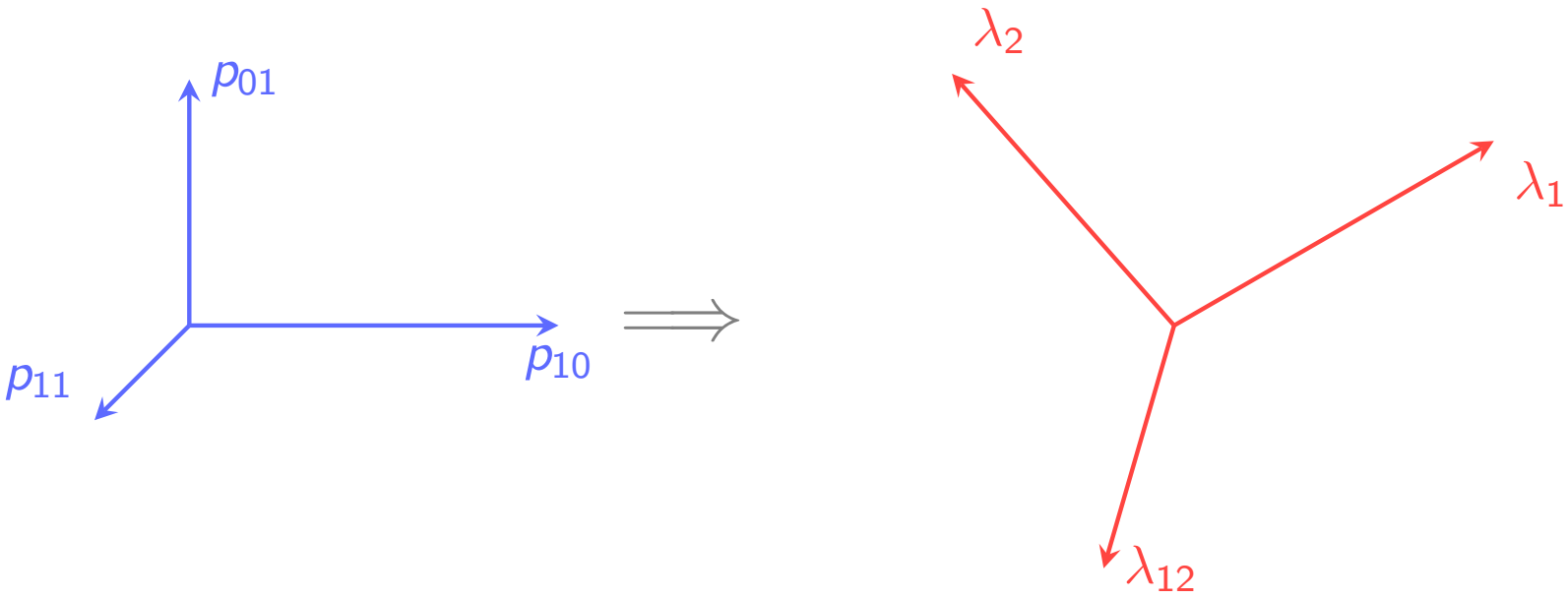
M	$\mathbb{L}$
$\{1,3\}$	$\{1\},\{3\},\{1,3\}$
$\{1,2,4\}$	$\{2\},\{1,2\},\{4\},\{1,4\},\{2,4\},\{1,2,4\}$
$\{1,2,3,4\}$	$\{2,3\},\{1,2,3\},\{3,4\},\{1,3,4\},\{2,3,4\},\{1,2,3,4\}$

This is a hierarchical parameterization, so smooth.

Set $\lambda_{13}^{13} = \lambda_{14}^{124} = \lambda_{124}^{124} = 0$ to enforce model.

Tangent spaces

The tangent space of the model can be in any co-ordinate space...



..but the log-linear parameterization gives much more useful directions!

Conditional independence models

We can think about the tangent space of a discrete model in terms of (marginal) log-linear parameters.

Let p_0 be the uniform distribution, so $\lambda_L^M = 0$ for all $L \subseteq M \subseteq V$.

Define Λ_A as the vector space of perturbations to p_0 that:

- modifies λ_A by ε;
- keeps $\lambda_L = o(\varepsilon)$ for $L \neq A$.

Then we can consider the tangent space of a model at this point:

$$\mathrm{TC}(p_0) = \bigoplus_{\emptyset \neq A \subseteq V} \Lambda_A.$$

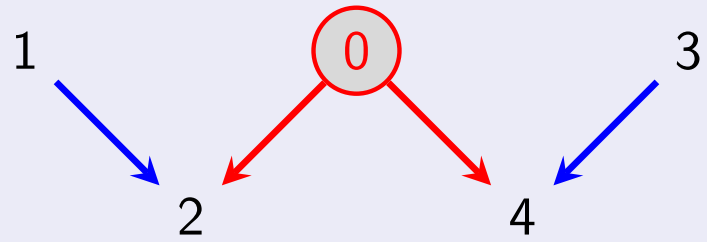
Then the tangent space of a model in which $X_i \perp\!\!\!\perp X_j \mid X_C$ is restricted to

$$\mathrm{TC}(p_0) = \bigoplus_{\emptyset \neq A \notin \mathbb{D}(i,j|C)} \Lambda_A.$$

Maximal ancestral models

As noted, we can parameterize maximal ancestral graph (MAG) models using marginal log-linear parameters.

A (directed) **maximal ancestral graph** (MAG) model is just a collection of conditional independences that can be represented by a DAG with hidden variables (Richardson and Spirtes, 2002).



This MAG implies the independences

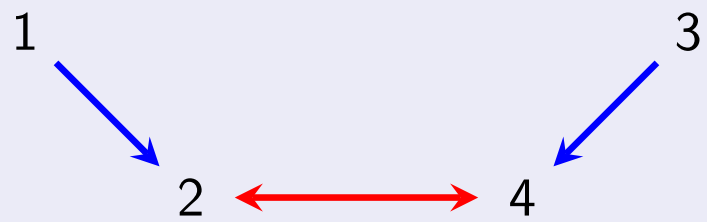
$$X_1 \perp\!\!\!\perp X_3, X_4 \qquad\qquad X_3 \perp\!\!\!\perp X_2 \mid X_1,$$

which cannot be faithfully represented by any DAG.

Maximal ancestral models

As noted, we can parameterize maximal ancestral graph (MAG) models using marginal log-linear parameters.

A (directed) **maximal ancestral graph** (MAG) model is just a collection of conditional independences that can be represented by a DAG with hidden variables (Richardson and Spirtes, 2002).



This MAG implies the independences

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

which cannot be faithfully represented by any DAG.

Markov Equivalence

In Hu and Evans (2020), we gave a criterion for two MAGs to be Markov equivalent (i.e. same m-separations) based on collections of subsets.

The **parameterizing sets** for a MAG $\mathcal{G}$ are

$$\mathcal{S}(\mathcal{G}) = \{H \cup A : H \in \mathcal{H}(\mathcal{G}),\, A \subseteq \text{tail}_{\mathcal{G}}(H)\},$$

where $\mathcal{H}(\mathcal{G})$ is the collection of **heads** in $\mathcal{G}$.

Given a vertex v in a head H, if we condition on $X_{H\setminus\{v\}}$, then the distribution cannot be m-separated from any $t \in H \cup \text{tail}_{\mathcal{G}}(H)$.

As an analogy, for DAGs heads = vertices and tails = parent sets.

Theorem (Hu and Evans, 2020)

Two MAGs $\mathcal{G}$ and $\mathcal{G}'$ are Markov equivalent if and only if $\mathcal{S}(\mathcal{G}) = \mathcal{S}(\mathcal{G}')$.

Parameterizing set

There is an interesting duality between the parameterizing set and the 'constrained set'.

Suppose a MAG $\mathcal{G}$ contains m-separations $a_i \perp_m b_i \mid C_i$ for $i \in I$.

Then

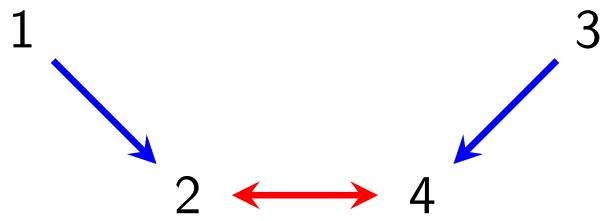
$$\mathcal{P}(V) \setminus (\mathcal{S}(\mathcal{G}) \cup \{\emptyset\}) = \bigcup_{i \in I} \mathbb{D}(a_i, b_i \mid C_i).$$

- So in other words, the parameterizing set reflects the sets that are **not** constrained by a conditional independence.
- In addition, the parameterizing set is **precisely** the collection of effects in an MLL parameterization of the same model.

Parameterizing Sets Example

Consider this MAG, which implies

$X_1 \perp\!\!\!\perp X_3, X_4 \quad \text{and} \quad X_3 \perp\!\!\!\perp X_2 \mid X_1;$



head	tail	parameterizing sets
$\{1\}$	$\emptyset$	$\{1\}$
$\{2\}$	$\{1\}$	$\{2\}, \{1,2\}$
$\{3\}$	$\emptyset$	$\{3\}$
$\{4\}$	$\{3\}$	$\{4\}, \{3,4\}$
$\{2,4\}$	$\{1,3\}$	$\{2,4\}, \{1,2,4\}, \{2,3,4\}, \{1,2,3,4\}$

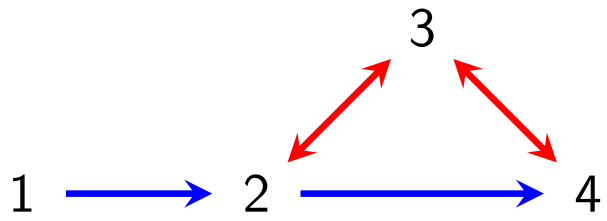
Parameterizing set is missing only subsets for:

I	$X_1 \perp\!\!\!\perp X_3, X_4$	$X_3 \perp\!\!\!\perp X_2 \mid X_1$
$\mathbb{D}(I)$	$\{1,3\}, \{1,4\}, \{1,3,4\}$	$\{2,3\}, \{1,2,3\}.$

Parameterizing Sets Example

Consider this MAG, which implies

$X_3 \perp\!\!\!\perp X_1$ and $X_4 \perp\!\!\!\perp X_1 \mid X_2$:



head	tail	parameterizing sets
$\{1\}$	$\emptyset$	$\{1\}$
$\{2\}$	$\{1\}$	$\{2\}, \{1,2\}$
$\{3\}$	$\emptyset$	$\{3\}$
$\{2,3\}$	$\{1\}$	$\{2,3\}, \{1,2,3\}$
$\{4\}$	$\{2\}$	$\{4\}, \{2,4\}$
$\{3,4\}$	$\{1,2\}$	$\{3,4\}, \{1,3,4\}, \{2,3,4\}, \{1,2,3,4\}$

Parameterizing set is missing only subsets $\{1,3\}$, $\{1,4\}$ and $\{1,2,4\}$.

Conditional independence models

Note that the Markov equivalence result is entirely nonparametric: it is independent of the character of the random variables.

One can show that:

- all missing sets are due to m-separations in the graph;
- the sets that are present (and hence those that are absent) characterize the Markov equivalence class.

The upshot is that the parameterizing set (and therefore the tangent space in the discrete case) is a signature for every Markov equivalence class of a MAG model.

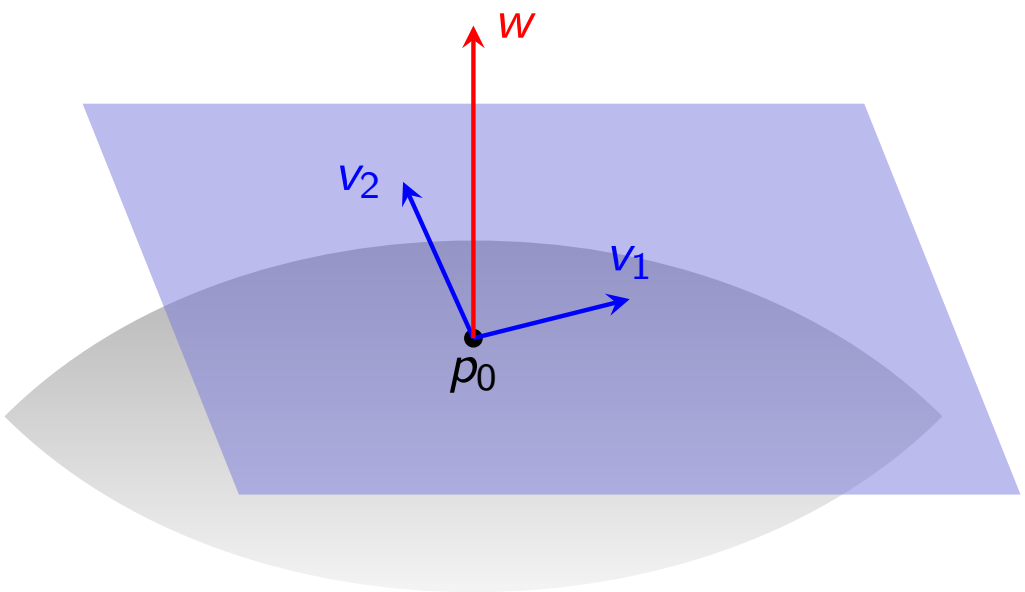
$$\mathrm{TC}(p_0) = \bigoplus_{A \in \mathcal{S}(\mathcal{G})} \Lambda_A.$$

We conjecture something similar characterization of the tangent space is true in the general case.

Tangent spaces

Go to a book on semi-parametric statistics (e.g. Tsiatis, 2006), and it might say something like:

The **tangent space** of a nonparametric model at a distribution P is the set of functions that have expectation zero under P.



Tangent spaces

Suppose that $V = \{1, \ldots, k\} := [k]$.

This can be decomposed into a sequence of conditional spaces:

$$\mathrm{TC}(p_0) = \Lambda_1(p_0) \oplus \Lambda_{2|1}(p_0) \oplus \cdots \oplus \Lambda_{k|[k-1]}(p_0),$$

where $\Lambda_{i|[i-1]} = \{h : \mathbb{E}[h(X_1, \ldots, X_i) \mid X_1 = x_1, \ldots, X_{i-1} = x_{i-1}] = 0\}$.

But we can go further! We can write

$$\Lambda_{k|[k-1]} = \bigoplus_{C \subset [k-1]} \Lambda_{C \cup \{k\}},$$

where $\Lambda_A = \{h : \mathbb{E}[h(X_A) \mid X_{A \setminus \{a\}} = x_{A \setminus \{a\}}] = 0, \quad \forall a \in A, x_{A \setminus \{a\}}\}$.

Conditional distributions and sets

Inspired by this derivation, we choose to associate a conditional distribution of the form $X_i \mid X_B$ with the collection of sets

$$\mathbb{D}(\{i\} \mid B) = \{\{i\} \cup A : A \subseteq B\}.$$

For example, the conditional distribution $P(X \mid Y, Z)$ can be associated with

$$\mathbb{D}(\{X\} \mid \{Y, Z\}) = \{\{X\}, \{X, Y\}, \{X, Z\}, \{X, Y, Z\}\}.$$

We can deduce from this that (for example), $P(X_i, X_j \mid X_B)$ should be associated with

$$\begin{aligned}\mathbb{D}(\{i,j\} \mid B) &:= \mathbb{D}(\{i\} \mid \{j\} \cup B) \cup \mathbb{D}(\{i\} \mid B) \\ &= \{C : C \subseteq B \cup \{i,j\} \text{ and } C \cap \{i,j\} \neq \emptyset\}.\end{aligned}$$

Alignment of conditionals

Graham et al. (2024) consider data fusion based on sources with *aligned conditionals*.

That is, we are interested in a **target** distribution Q, and there are **sources** $P(\cdot \mid S = s)$ for $s = 1, \ldots, K$ such that for each s:

$$Q(X_j \mid X_A) = P(X_j \mid X_A, S = s).$$

Their Example 3 concerns a prospective study on a different population and a case-control study. In other words:

$$P(Y \mid L, A, S = 1) = Q(Y \mid L, A)$$
$$P(L, A \mid Y, S = 2) = Q(L, A \mid Y).$$

Two natural questions are:

- Under what circumstances do we obtain constraints on P?
- What set of alignments is sufficient to recover Q?

Alignment of conditionals

The tangent space representation helps us to answer these questions.

There will be an equality constraint only if we have two conditionals $i \mid B$ and $j \mid C$ such that: (i) $i = j$ or (ii) $i \in C$ and $j \in B$.

That is, if there is no set common to $\mathbb{D}(i \mid B)$ and $\mathbb{D}(j \mid C)$, there is no equality constraint.

Theorem

There is no equality constraint on P if there is no intersection between any of the sets $\mathbb{D}(i \mid B)$ for aligned conditionals $P(X_i \mid X_B)$.

Conjecture

If there is an intersection between any of the sets $\mathbb{D}(i \mid B)$ for aligned conditionals $P(X_i \mid X_B)$, then at least for some state-spaces, there is an equality constraint.

Alignment: example

Example 3, scenario (iii.a) assumes Q contains causally sufficient covariates L, treatment A and outcome Y. We have data from:

- a prospective study on a population with different distribution of L, A;
- a case-control study on the target Q.

So we have:

$$P(Y \mid L, A, S = 1) = Q(Y \mid L, A)$$
$$P(L, A \mid Y, S = 2) = Q(L, A \mid Y).$$

We can see that sources 1 and 2 respectively give:

$$\mathbb{D}(\{Y\} \mid \{L, A\}) = \{ \quad \{Y\}, \quad \{L, Y\}, \{A, Y\}, \{L, A, Y\}\}$$

$$\mathbb{D}(\{L, A\} \mid \{Y\}) = \{\{L\}, \{A\}, \{L, A\}, \{L, Y\}, \{A, Y\}, \{L, A, Y\}\}$$

giving an intersection of $\{L, Y\}, \{A, Y\}, \{L, A, Y\}$.

Constraints

We know that

$$Q(L, A) \cdot Q(Y \mid A, L) = Q(Y) \cdot Q(A, L \mid Y)$$
$$Q(L, A) \cdot P(Y \mid A, L, S = 1) = Q(Y) \cdot P(A, L \mid Y, S = 2)$$
$$\frac{P(Y \mid A, L, S = 1)}{P(A, L \mid Y, S = 2)} = \frac{Q(Y)}{Q(L, A)}.$$

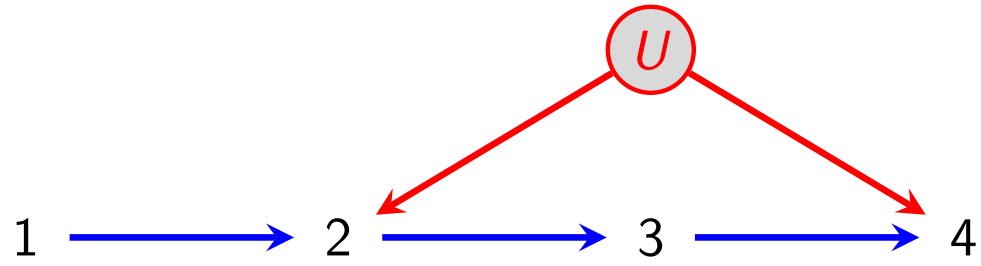
In other words, these conditionals are the same up to a product of functions of Y and functions of L, A.

This is because $\mathrm{OR}(\{Y\}, \{L, A\})$ is contained in both conditionals!

If only have controls ($Y = 0$) from our case-control study, there is no dependence information contained in $Q(A, L \mid Y = 0)$, so no constraint.

Nested Markov models

Marginal models are not defined purely by conditional independence:

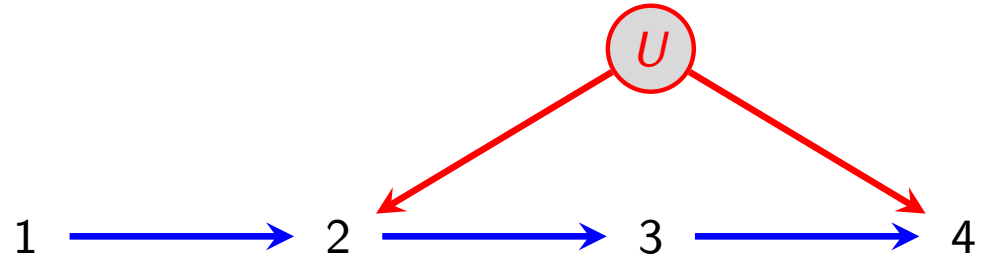


This is a model defined (implicitly) by an integral:

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

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

The Verma Constraint



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

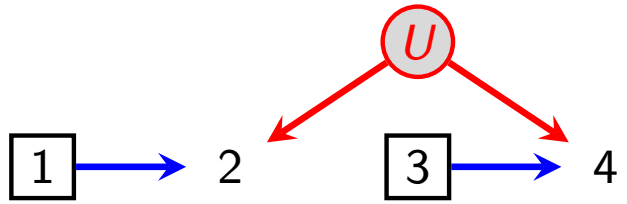
But note that

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

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

Nested Markov model

In other words, we find that $X_1 \perp\!\!\!\perp X_4 \mid X_3$ **after** we have 'fixed' (intervened on) X_3.



Equality constraints of the kind on the previous slide are called **nested constraints** (or Verma constraints, or dormant independences).

We describe the set of distributions restricted in this way as the **nested Markov model** for $\mathcal{G}$, or $\mathcal{N}(\mathcal{G})$.

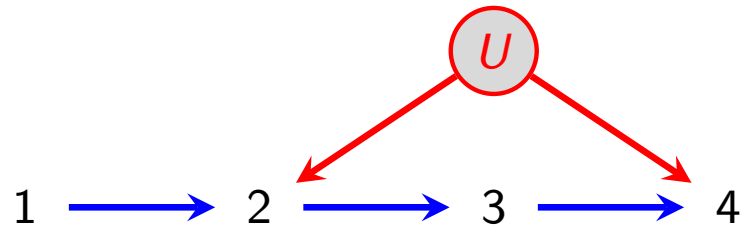
Importantly, at the uniform distribution (or anywhere such that $X_2 \perp\!\!\!\perp X_3$), the **directions restricted** by the nested constraint are the same as those restricted by a model in which the **ordinary** conditional independence $X_1 \perp\!\!\!\perp X_4 \mid X_3$ holds.

Consequently, we can show that $\mathrm{TC}_{p_0}(\mathcal{N}(\mathcal{G})) = \sum_{A \in \mathcal{S}(\mathcal{G})} \Lambda_A.$

Marginal model

We define the **marginal model** as the set of distributions that can be realised over the observed variables for arbitrary latent variables.

Then our proof requires us to show that we can move in any direction within $\mathrm{TC}_{p_0}(\mathcal{N}(\mathcal{G}))$ in the marginal model.

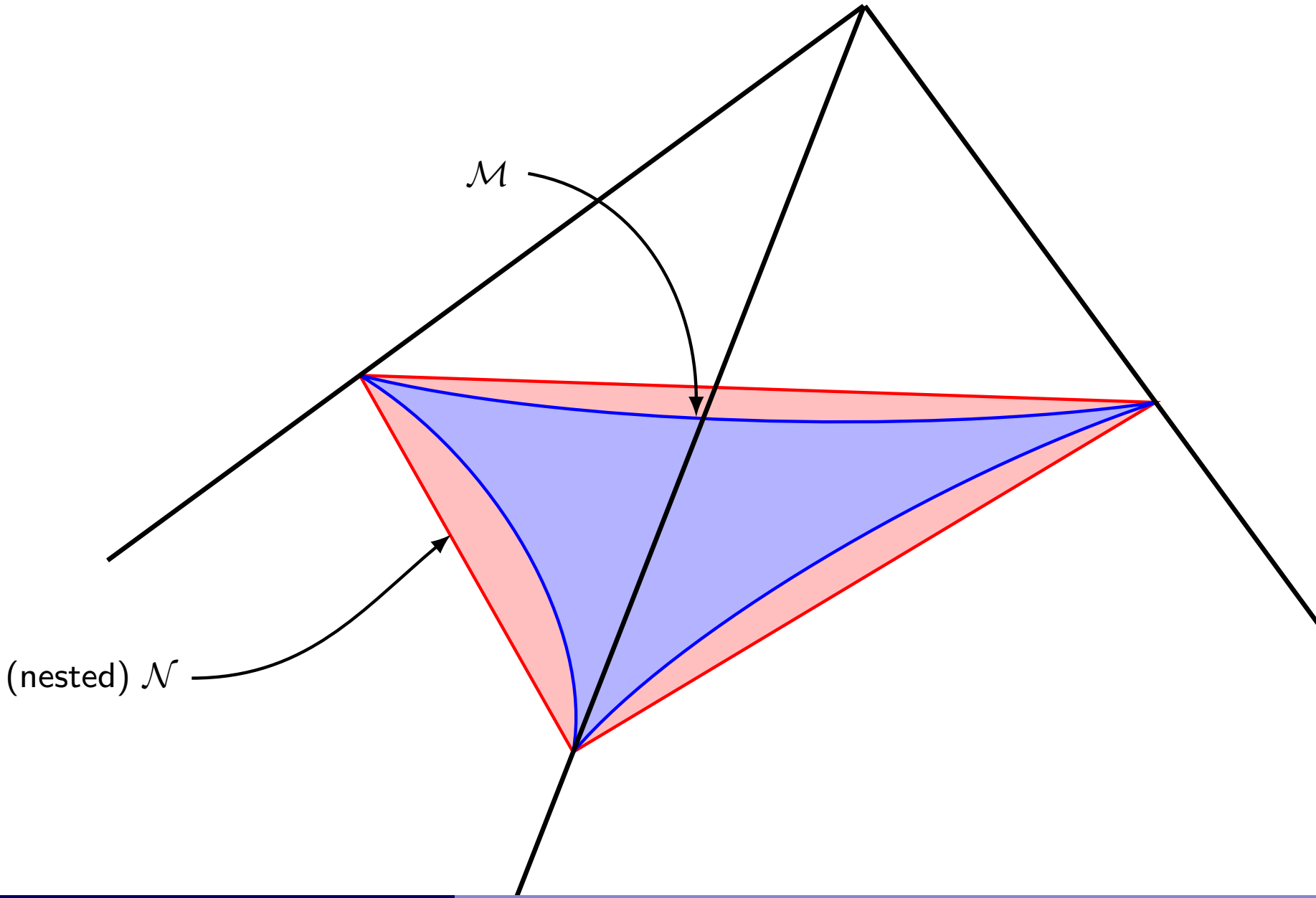


To do this, we construct very specific distributions. For example, to show that we can move in λ_{124}, we shift:

$$p(u) \cdot p(x_2 \mid x_1, u) \cdot p(x_4 \mid u) \qquad \text{[note } \textbf{not } p(x_4 \mid x_3, u)\text{]}$$

in a co-ordinated way.

Getting the picture



Main result

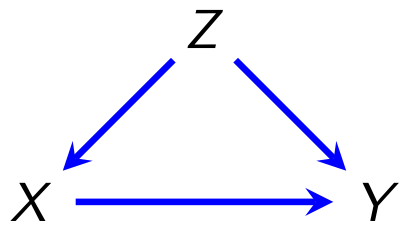
The tangent spaces of the **nested Markov model** ($\mathcal{N}(\mathcal{G})$) and the **marginal model** ($\mathcal{M}(\mathcal{G})$) are the same at the uniform distribution p_0:

$$\mathrm{TC}_{p_0}(\mathcal{N}(\mathcal{G})) = \mathrm{TC}_{p_0}(\mathcal{M}(\mathcal{G})).$$

Hence the **dimension** of the two models is the same.

This result holds in the discrete case, but does it hold in general?

Causal simulation



The **frugal parameterization** (Evans and Didelez, 2024) of the causal system above is:

$$p(z, x) \qquad p(y \mid do(x)) \qquad \phi^*_{ZY|X}(z, y \mid x),$$

where $\phi^*_{ZY|X}$ parameterizes the conditional dependence between Z and Y given X. This is typically a copula if one of Y or Z is discrete.

This parameterization corresponds to log-linear 'effects' as:

$p(z, x)$	$\{Z\}, \{X\}, \{Z, X\}$
$p(y \mid do(x))$	$\{Y\}, \{X, Y\}$
$\phi^*_{ZY|X}(z, y \mid x)$	$\{Z, Y\}, \{Z, X, Y\}.$

Odds ratio for general distributions

The **odds ratio** for generic distributions with density p is defined as

$$\mathrm{OR}(x, y) = \frac{p(x, y) \cdot p(x^*, y^*)}{p(x^*, y) \cdot p(x, y^*)},$$

for some arbitrary baseline values x^*, y^* provided that $p(x, y^*) > 0$ and $p(x^*, y) > 0$ almost surely.

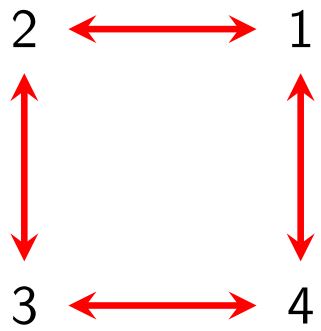
Chen (2007) shows that a general likelihood for random variables X and Y can be written as

$$p(x, y) = \frac{p(x \mid y^*)\,\mathrm{OR}(x, y) p(y \mid x^*)}{\int p(x \mid y^*)\,\mathrm{OR}(x, y) p(y \mid x^*)\, d\mu(x, y)}. \qquad (*)$$

Marginal independence models

We can use this to parameterize the bidirected four cycle graph in the case of a general distribution:

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



Shown by Lupparelli et al. (2009) that setting the marginal log-linear parameters

$$\lambda^{13}_{13} = \lambda^{24}_{24} = 0$$

and completing with parameters from $V = \{1, 2, 3, 4\}$ is a smooth and variation independent parameterization.

Marginal independence for general distributions

A similar approach can be taken in the general case. We work with the quantities

$$p(x_{13}) \qquad p(x_{24}) \qquad \mathrm{OR}(x_{13}, x_{24}),$$

enforcing that the first two distributions factorize into independent pieces.

We also have a relationship between $p(x_{13} \mid x_{24}^*)$ and $p(x_{13})$ which uses $\mathrm{OR}(x_{13}, x_{24})$ and $p(x_{24} \mid x_{13}^*)$.

This enables us to set up an integral equation that we conjecture will always converge to likelihood in $(*)$.

Recap

- Marginal log-linear parameters are a flexible way to model multivariate discrete data.
- The effects in a collection of marginal log-linear parameters are related to tangent spaces of the corresponding models.
- This makes them relevant to many other areas, at least conceptually (and at least for me!):
  - demonstrating equivalence of different models;
  - Markov equivalence of MAGs;
  - model selection methods;
  - constraints on conditional distributions
  - causal parameterization and simulation;
  - general likelihoods.
- I am convinced that better understanding of how (marginal) log-linear parameters can be extended to the general nonparametric case, will help to open up exciting new frontiers in multivariate statistics!

Thank you!

References I

Bergsma, W.P. and Rudas, T. Marginal models for categorical data. *Annals of Statistics*, 30(1), pp.140-159, 2002.

Chen, H.Y. A semiparametric odds ratio model for measuring association. *Biometrics*, 63(2), pp 413–421, 2007.

Evans, R.J. Margins of discrete Bayesian networks, *Annals of Statistics*, 46(6A), 2018.

Forcina, A., Lupparelli, M., and Marchetti, G.M. Marginal parameterizations of discrete models defined by a set of conditional independencies, *Journal of Multivariate Analysis*, 101(10) pp 2519-2527, 2010.

Glonek, G.F. and McCullagh, P. Multivariate logistic models. *Journal of the Royal Statistical Society: Series B*, 57(3), pp 533–546, 1995.

Hu, Z. and Evans, R.J. Faster algorithms for Markov equivalence, *UAI-20*, 2020.

Hu, Z. and Evans, R.J. Towards standard imsets for maximal ancestral graphs. *Bernoulli*, 2024.

References II

Lupparelli, M., Marchetti, G.M. and Bergsma, W.P. Parameterizations and Fitting of Bi-directed Graph Models to Categorical Data. *Scandinavian Journal of Statistics*, 36: 559–576, 2009.

Rudas, T., Bergsma, W. and Németh, R. Marginal conditional independence models with application to graphical modeling. *Biometrika*, 2010.

Studený, M. *Probabilistic Conditional Independence Structures*, Springer, 2005.

Marginal log-linear parameters

Consider a collection of pairs $\mathcal{L} = \{(M, L) : L \subseteq M \subseteq V\}$. Let

$$\text{mar}(\mathcal{L}) = \{M : (M, L) \in \mathcal{L}\}$$
$$\text{eff}(\mathcal{L}) = \{L : (M, L) \in \mathcal{L}\}.$$

A parameterization is said to be **complete** if every $L \subseteq V$ is represented exactly once in the collection $\mathcal{L}$.

The parameterization is said to be **hierarchical** if we can order the elements of $\text{mar}(\mathcal{L})$ as $M_1, \ldots, M_k$ so that:

- $M_i \not\subseteq M_j$ for $i > j$;
- up to the jth margin the parameterization is complete for each $j \leq k$.

Imsets

An **i**nteger-valued **m**ulti-**set** (imset) is an algebraic way to represent conditional independence introduced by Milan Studený (e.g. Studený, 2005).

It is a vector with entries indexed by subsets of V and integer values;

$$\text{let} \quad \delta_A(B) = \begin{cases} 1 & \text{if } A = B \\ 0 & \text{otherwise.} \end{cases}$$

The imset $u_{\langle A,B|C\rangle} = \delta_C - \delta_{A\cup C} - \delta_{B\cup C} + \delta_{A\cup B\cup C}$ represents the conditional independence $X_A \perp\!\!\!\perp X_B \mid X_C$, in the sense that

$$\langle \mathrm{H}, u_{\langle A,B|C\rangle}\rangle := \mathrm{H}(X_C) - \mathrm{H}(X_A, X_C) - \mathrm{H}(X_B, X_C) + \mathrm{H}(X_A, X_B, X_C) = 0$$

if and only if $X_A \perp\!\!\!\perp X_B \mid X_C$ under p, where $\mathrm{H}(\cdot)$ is the entropy operator:

$$\mathrm{H}(X_A) = -\int_{\mathfrak{X}_A} p(x_A) \log p(x_A)\, dx_A.$$

Structural/characteristic imsets

An imset like $u_{\langle A,B|C\rangle}$ is called **structural**.

If we perform a Möbius transform it becomes a **characteristic** imset, denoted by $c_{\langle A,B|C\rangle}$. Then one can show that

$$c_{\langle A,B|C\rangle}(S) = \begin{cases} 0 & \text{if } S \in \mathbb{D}(A,B \mid C) \\ 1 & \text{otherwise.} \end{cases}$$

So, if a conditional independence model can be represented by cond. independences $X_A \perp\!\!\!\perp X_B \mid X_C$ for which the sets $\mathbb{D}(A,B \mid C)$ are disjoint, then

$$c_{\mathcal{G}}(S) = \begin{cases} 0 & \text{if } S \in \mathbb{D}(A,B \mid C) \text{ for any } \langle A,B \mid C\rangle \in \mathcal{I}(\mathcal{G}) \\ 1 & \text{otherwise.} \end{cases}$$

Further, the model can be scored by using the corresponding **structural** imset, and indeed

$$\langle \mathrm{H}, u_{\mathcal{G}} \rangle \approx -\ell(p; X_1, \ldots, X_n).$$

Decomposition Proof

Given $h(x_V)$, let $t(x_V) = h(x_V) \cdot p(x_V)$, so if $h \in \Lambda_{k|[k-1]}$:

$$\int_{\mathfrak{X}_k} t(x_{[k]})\, dx_k = p_0(x_{[k-1]}) \int_{\mathfrak{X}_k} p_0(x_k \mid x_{[k-1]}) h(x_{[k]})\, dx_k = 0.$$

Then write

$$t^{(k-1)}(x_{[k-2]}, x_k) = \int_{\mathfrak{X}_{k-1}} t(x_{[k]})\, dx_{k-1},$$

so that

$$t(x_{[k]}) = \underbrace{\left\{ t(x_{[k]}) - t^{(k-1)}(x_{[k-2]}, x_k) \right\}}_{\in \mathsf{T}_{k-1,k|[k-2]}} + \underbrace{t^{(k-1)}(x_{[k-2]}, x_k)}_{\mathsf{T}_{k|[k-2]}}.$$

Now one can check that $\mathsf{T}_{k|[k-1]} = \mathsf{T}_{k-1,k|[k-2]} \oplus \mathsf{T}_{k-1|[k-2]}$ and that these spaces are orthogonal.

Subset spaces

By a recursion:

$$\mathsf{T}_{k|[k-1]} = \bigoplus_{C \subset [k-1]} \mathsf{T}_{C \cup \{k\}},$$

so we can decompose into separate subset spaces, such that each $t \in \mathsf{T}_A$ is a function only of x_A and where

$$\int_{\mathfrak{X}_a} t(x_{A\setminus\{a\}}, y_a)\, dy_a = 0 \qquad \forall a \in A, x_{A\setminus\{a\}} \in \mathfrak{X}_{A\setminus\{a\}}.$$

These correspond to Λ_A where $h \in \Lambda_A$ if $\mathbb{E}_p[h(x_A) \mid X_{A\setminus\{a\}} = x_{A\setminus\{a\}}] = 0$ for all $a \in A$ and $x_{A\setminus\{a\}} \in \mathfrak{X}_{A\setminus\{a\}}$.