Data Mining
Graphical Models for Discrete Data
Undirected Graphs (2)

Ad Feelders

Universiteit Utrecht

Overview of Coming Lectures

- Introduction
- Independence and Conditional Independence
- Graphical Representation of Conditional Independence
- ☞ Log-linear Models
  - Hierarchical
  - Graphical
  - Decomposable
- Maximum Likelihood Estimation
- Model Testing
- Model Selection

2×2 Table

The probability function P_{12} of bivariate Bernoulli random vector (X_1, X_2) is determined by

$$P(x_1, x_2) = p(x_1, x_2)$$

where $p(x_1, x_2)$ is the table of probabilities:

$p(x_1, x_2)$	$x_2 = 0$	$x_2 = 1$	Total
$x_1 = 0$	$p(0,0)$	$p(0,1)$	$p_1(0)$
$x_1 = 1$	$p(1,0)$	$p(1,1)$	$p_1(1)$
Total	$p_2(0)$	$p_2(1)$	1

Probability function for 2×2 Table

Again we can write this as a single formula:

$$P(x_1, x_2) = p(0,0)^{(1-x_1)(1-x_2)} p(0,1)^{(1-x_1)x_2} p(1,0)^{x_1(1-x_2)} p(1,1)^{x_1 x_2}$$

Taking logarithms and collecting terms in x_1, x_2, and $x_1 x_2$ gives:

$$\begin{aligned} \log P(x_1, x_2) &= \log p(0,0) + \log\left(\frac{p(1,0)}{p(0,0)}\right) x_1 + \\ &\quad \log\left(\frac{p(0,1)}{p(0,0)}\right) x_2 + \log\left(\frac{p(1,1)p(0,0)}{p(0,1)p(1,0)}\right) x_1 x_2 \end{aligned}$$

Verify this using elementary properties of logarithms:

1. $\log a^b = b \log a$,
2. $\log \frac{a}{b} = \log a - \log b$, and
3. $\log ab = \log a + \log b$.

Log-linear expansion

Reparameterizing the right hand side leads to the so-called *log-linear expansion*

$$\log P(x_1, x_2) = u_\emptyset + u_1 x_1 + u_2 x_2 + u_{12} x_1 x_2$$

The coefficients, $u_\emptyset, u_1, u_2, u_{12}$ are known as the u-terms.

For example, the coefficient of the product $x_1 x_2$,

$$u_{12} = \log \left(\frac{p(1,1)p(0,0)}{p(0,1)p(1,0)} \right) = \log \text{cpr}(X_1, X_2)$$

is the logarithm of the cross product ratio of X_1 and X_2.

Independence and u-terms

Claim:

$$X_1 \perp\!\!\!\perp X_2 \Leftrightarrow u_{12} = 0$$

Proof: the factorisation criterion states that $X_1 \perp\!\!\!\perp X_2$ iff there exist two functions g and h such that

$$\log P(x_1, x_2) = g(x_1) + h(x_2) \text{ for all } (x_1, x_2)$$

If $u_{12} = 0$, we get

$$\log P(x_1, x_2) = u_\emptyset + u_1 x_1 + u_2 x_2,$$

so

$$g(x_1) = u_\emptyset + u_1 x_1 \qquad h(x_2) = u_2 x_2$$

suffices. If $u_{12} \neq 0$, no such decomposition is possible.

Three Dimensional Bernoulli

The joint distribution of three binary variables can be written:

$$P(x_1, x_2, x_3) = p(0,0,0)^{(1-x_1)(1-x_2)(1-x_3)} \cdots p(1,1,1)^{x_1 x_2 x_3}$$

Log-linear expansion

$$\begin{aligned} \log P(x_1, x_2, x_3) &= u_\emptyset + u_1 x_1 + u_2 x_2 + u_3 x_3 + u_{12} x_1 x_2 + \\ & \quad u_{13} x_1 x_3 + u_{23} x_2 x_3 + u_{123} x_1 x_2 x_3, \end{aligned}$$

with

$$u_{123} = \log \left(\frac{\text{cpr}(X_2, X_3 | X_1 = 1)}{\text{cpr}(X_2, X_3 | X_1 = 0)} \right)$$

Independence and the u-terms

Observation:

$$X_2 \perp\!\!\!\perp X_3 | X_1 \Leftrightarrow u_{23} = 0 \text{ and } u_{123} = 0$$

Proof: use factorisation criterion.

$X_2 \perp\!\!\!\perp X_3 | X_1 \Leftrightarrow$ there are functions $g(x_1, x_2)$ and $h(x_1, x_3)$ such that

$$\log P(x_1, x_2, x_3) = g(x_1, x_2) + h(x_1, x_3)$$

This is only possible when $u_{23} = 0$ (so the term $x_2 x_3$ drops out), and $u_{123} = 0$ (so the term $x_1 x_2 x_3$ drops out).

Log-linear expansion: non-binary variables

variables with more than 2 lables

For a 2×2 table the log-linear expansion is given by:

$$\log P(x_1, x_2) = u_\emptyset + u_1 x_1 + u_2 x_2 + u_{12} x_1 x_2$$

constants

for $x \in \{0, 1\}^2$, where $u_\emptyset, u_1, u_2$ and u_{12} are *constants*.

What if the x_i have more than two levels? In that case the u terms become *functions* of x rather than *constants*:

$$\log P(x_1, x_2) = u_\emptyset + u_1(x_1) + u_2(x_2) + u_{12}(x_1, x_2)$$

not constants

$x \in \{0, 1, 2\}$ $P(0), P(1), P(2)$

$P(x) = P(1)^{\delta_{x=1}} P(2)^{\delta_{x=2}} P(0)^{(1-\delta_{x=1}-\delta_{x=2})}$ δ: indicator function

$\log P(x) = \delta_{x=1} \log P(1) + \delta_{x=2} \log P(2) + (1 - \delta_{x=1} - \delta_{x=2}) \log P(0)$

$= \delta_{x=1} \log \frac{P(1)}{P(0)} + \delta_{x=2} \log \frac{P(2)}{P(0)} + \log P(0)$

Log-linear expansion: non-binary variables

Suppose $x \in \{0, 1, 2\}$. We can write

$$P(x) = p(1)^{\delta_{x=1}} p(2)^{\delta_{x=2}} p(0)^{(1-\delta_{x=1}-\delta_{x=2})},$$

where δ_A is the indicator function, that is,

$$\delta_A = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

Taking logarithms left and right, we get

$$\begin{aligned} \log P(x) &= \delta_{x=1} \log p(1) + \delta_{x=2} \log p(2) + (1 - \delta_{x=1} - \delta_{x=2}) \log p(0) \\ &= \delta_{x=1} \log p(1) + \delta_{x=2} \log p(2) + \log p(0) - \delta_{x=1} \log p(0) - \delta_{x=2} \log p(0) \\ &= \log p(0) + \log \frac{p(1)}{p(0)} \delta_{x=1} + \log \frac{p(2)}{p(0)} \delta_{x=2} \\ &= u_\emptyset + u(x) \end{aligned}$$

Log-linear expansion: non-binary variables

Where $u_\emptyset = \log p(0)$ and

$$u(x) = \begin{cases} \log \frac{p(1)}{p(0)} & \text{if } x = 1 \\ \log \frac{p(2)}{p(0)} & \text{if } x = 2 \\ 0 & \text{if } x = 0 \end{cases}$$

Similar rules apply to the case of multiple non-binary variables.

Why the log-linear representation?

Why do we use the log-linear representation of the probability table?

1. We are interested in expressing conditional independence constraints.
2. There is a straightforward correspondence between such constraints being satisfied, and the elimination of certain collections of u-terms from the log-linear expansion.
3. This correspondence is established by applying the factorisation criterion: $X \perp\!\!\!\perp Y \mid Z$ if and only if there exist functions g and h such that

$$\log P(x, y, z) = g(x, z) + h(y, z)$$

Log-linear expansion: general

Let $X = (X_1, X_2, \ldots, X_k)$ be a vector of discrete random variables, and let $K = \{1, 2, \ldots, k\}$ denote the set of indices (coordinates) of X.

The log-linear expansion of the probability distribution $P_K(X)$ is

$$\log P_K(x) = \sum_{a \subseteq K} u_a(x_a)$$

where the sum is taken over all possible subsets a of $K = \{1, 2, \ldots, k\}$.

Log-linear expansion: general

The log-linear expansion of the probability distribution $P_K(X)$ is

$u_a(x_a) = 0$ if $x_i = 0$ for any $i \in a$

ex. $u_{12} x_1 x_2$ if x_1 or $x_2 = 0$. $u_{12}(x_a) = 0$

$$\log P_K(x) = \sum_{a \subseteq K} u_a(x_a)$$

- We get a (set of) u-term(s) for each subset of the variables.
- We code the values of X_i as $\{0, 1, \ldots, d_i - 1\}$, where d_i is size of the domain of X_i.
- Set $u_a(x_a) = 0$ whenever $x_i = 0$ for any X_i with $i \in a$ to eliminate redundant parameters.
- This is analogous to the case where X is binary.
- There are as many u-terms in the full log-linear expansion as there are cells in the contingency table.

Log-linear expansion: example

Suppose we have 3 variables, X_1, X_2, and X_3 with 3 possible values each. Then the contingency table contains $3^3 = 27$ cells, one for each possible combination of values. The are also 27 u-terms in the full log-linear expansion:

$3^3 = 27$

$u_a(x_a)$	# u-terms	calculation
$u_\emptyset$	1	
$u_1(x_1)$	2	$(3-1)$ $X_1 = 0$
$u_2(x_2)$	2	$(3-1)$
$u_3(x_3)$	2	$(3-1)$
$u_{12}(x_1, x_2)$	4	$(3-1)(3-1)$
$u_{13}(x_1, x_3)$	4	$(3-1)(3-1)$
$u_{23}(x_2, x_3)$	4	$(3-1)(3-1)$
$u_{123}(x_1, x_2, x_3)$	8	$(3-1)(3-1)(3-1)$
Total	27	

-Independence and the u-terms

Let

- $X = (X_1, X_2, \ldots, X_k)$
- $a, b, c \subseteq \{1, 2, \ldots, k\}$
- a, b, c disjoint and $a \cup b \cup c = \{1, 2, \ldots, k\}$

Independence and the u-terms

We have

$$X_b \perp\!\!\!\perp X_c \mid X_a$$

if and only if all u-terms in the log-linear expansion with coordinates in both b and c, are equal to zero.

Independence and the u-terms: Example

Let $X = (X_1, \ldots, X_5)$, and $a = \{1, 3\}, b = \{4\}, c = \{2, 5\}$. According to the factorization criterion, the conditional independence

$$X_4 \perp\!\!\!\perp (X_2, X_5) \mid (X_1, X_3)$$

holds if, and only if, there are functions g and h such that

$$\log P(x_1, \ldots, x_5) = g(x_1, x_3, x_4) + h(x_1, x_2, x_3, x_5)$$

For this to be possible, the u-terms that contain elements from both the sets $\{4\}$ and $\{2, 5\}$ have to be zero. So all u-terms $u_{24}, u_{45}, u_{124}, u_{145}, \ldots, u_{12345}$ have to be zero.

Independence and the u-terms

Proof: If for all $s \subseteq K$ with $u_s(x_s) \neq 0$, we have

1. $s \subseteq a \cup b$, or
2. $s \subseteq a \cup c$

(i.e. s does not contain coordinates from *both* b *and* c), then

$$\log P_K(x) = \sum_{s \subseteq a \cup b} u_s(x_s) + \sum_{s \subseteq a \cup c} u_s(x_s) - \sum_{s \subseteq a} u_s(x_s)$$

counted twice

But this function is of the form $g(x_a, x_b) + h(x_a, x_c)$ and hence $X_b \perp\!\!\!\perp X_c \mid X_a$ by the factorisation criterion.

Note: we subtract $\sum_{s \subseteq a} u_s(x_s)$ because subsets of a were added twice.

Overview of Coming Lectures

- Introduction
- Independence and Conditional Independence
- Graphical Representation of Conditional Independence
- Log-linear Models (theory)
  - ☞ Hierarchical
    - Graphical
    - Decomposable
- Maximum Likelihood Estimation
- Model Testing
- Model Selection

Hierarchical Log-Linear Models

certain u-term include $\Rightarrow$ all subsets of that is included.

ex: $u_{123}(x_1, x_2, x_3) = 0 \Rightarrow$ $u_{12} \neq 0$, $u_{13} \neq 0$, $u_{23} \neq 0$

- In most applications, it does not make sense to include the three-way association u_{123} unless the two-way associations u_{12}, u_{13} and u_{23} are all present as well.
- A log-linear model is said to be *hierarchical* if the presence of a term implies that all lower-order terms are also present. That is, if $u_A(x_A)$ is present, then for all $a \subseteq A$, $u_a(x_a)$ must be present as well.
- Hence, a hierarchical model is uniquely identified by listing its highest order interaction terms.

For example, in the model with 3 binary variables we have

$$u_{123} = \log\left(\frac{\text{cpr}(X_2, X_3 \mid X_1 = 1)}{\text{cpr}(X_2, X_3 \mid X_1 = 0)}\right)$$

$$u_{23} = \log \text{cpr}(X_2, X_3 \mid X_1 = 0)$$

So what does it mean if we set $u_{23} = 0$ and $u_{123} \neq 0$?

if $u_{23} = 0 \Rightarrow X_2 \perp\!\!\!\perp X_3 \mid X_1$

if $u_{123} \neq 0$ $\because u_{23} = 0 \Rightarrow \log(\text{cpr}(X_2, X_3 \mid X_1 = 1) \Rightarrow X_2 \not\perp\!\!\!\perp X_3 \mid X_1$

Hierarchical Models for three dimensions

(excluded)

Model	Omitted	Interpretation
123	none	saturated
12,13,23	u_{123}	homogeneous association
12,13	u_{123}, u_{23}	$X_2 \perp\!\!\!\perp X_3 \mid X_1$
12,23	u_{123}, u_{13}	$X_1 \perp\!\!\!\perp X_3 \mid X_2$
13,23	u_{123}, u_{12}	$X_1 \perp\!\!\!\perp X_2 \mid X_3$
12,3	u_{123}, u_{13}, u_{23}	$(X_1, X_2) \perp\!\!\!\perp X_3$
13,2	u_{123}, u_{12}, u_{23}	$(X_1, X_3) \perp\!\!\!\perp X_2$
23,1	u_{123}, u_{12}, u_{13}	$(X_2, X_3) \perp\!\!\!\perp X_1$
1,2,3	$u_{123}, u_{12}, u_{13}, u_{23}$	mutual independence

Overview of Coming Lectures

- Introduction
- Independence and Conditional Independence
- Graphical Representation of Conditional Independence
- Log-linear Models
  - Hierarchical
  - ☞ Graphical
  - Decomposable
- Maximum Likelihood Estimation
- Model Testing
- Model Selection

Graphical Log-Linear Model

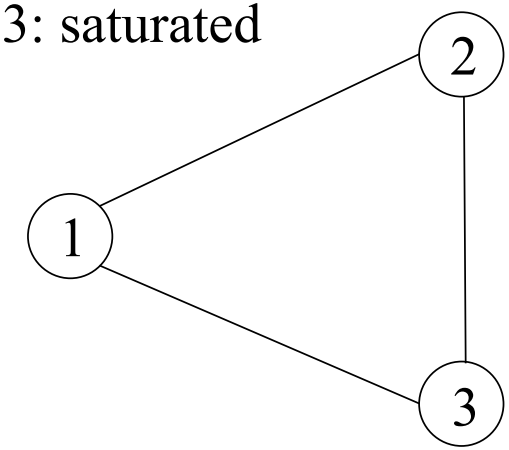
Given its independence graph $G = (K, E)$, the log-linear model for the random vector X is a *graphical model* for X if the distribution of X is *arbitrary* apart from constraints of the form that for all pairs of coordinates not in the edge set E, the u-terms containing the selected coordinates are equal to zero.

All constraints of a graphical model can be read from the independence graph.

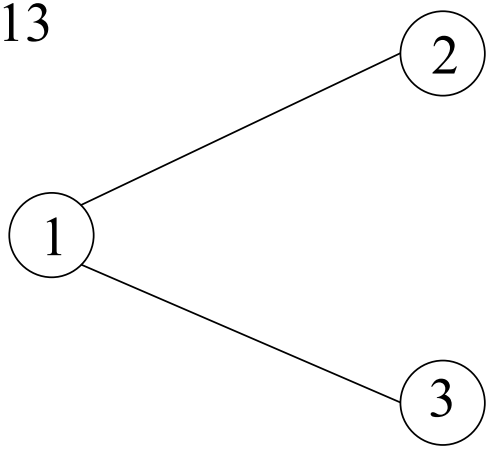
A graphical model is a hierarchical model in which the highest order interaction terms correspond to the cliques in the graph.

Some hierarchical models and their independence graphs

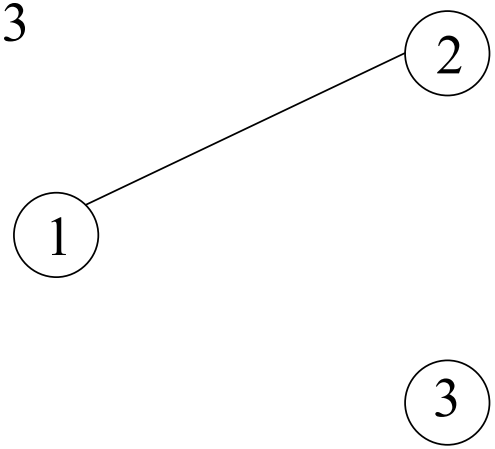
123: saturated



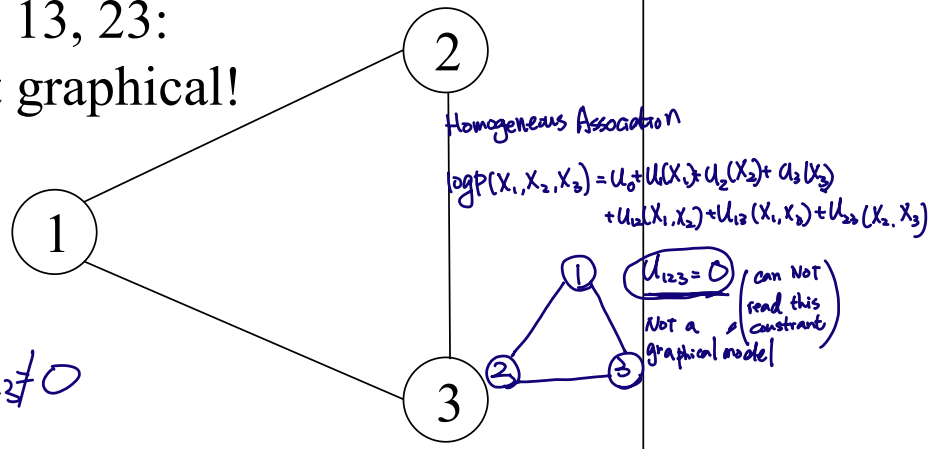
12,13



12,3



12, 13, 23:
not graphical!



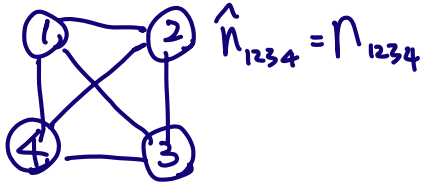
Overview of Coming Lectures

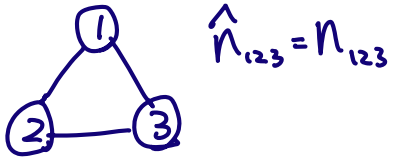
- Introduction
- Independence and Conditional Independence
- Graphical Representation of Conditional Independence
- Log-linear Models
  - Hierarchical
  - Graphical
  - Decomposable
- ☞ Maximum Likelihood Estimation
- Model Testing
- Model Selection

Estimating Models from Data: Maximum Likelihood

- The Maximum Likelihood (ML) estimator of graphical log-linear model M returns estimates of the cell probabilities that maximize the probability of the observed data, subject to the constraint that the conditional independencies of M are satisfied by the estimates.
- ML estimator of graphical log-linear model M satisfies the likelihood equations

$$\hat{n}_a^M = N\hat{P}_a^M = n_a$$



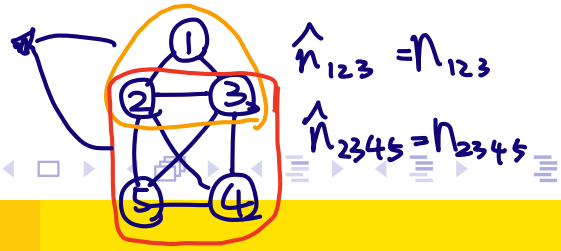


whenever the subset of vertices a in the graph form a clique (maximal complete subgraph).

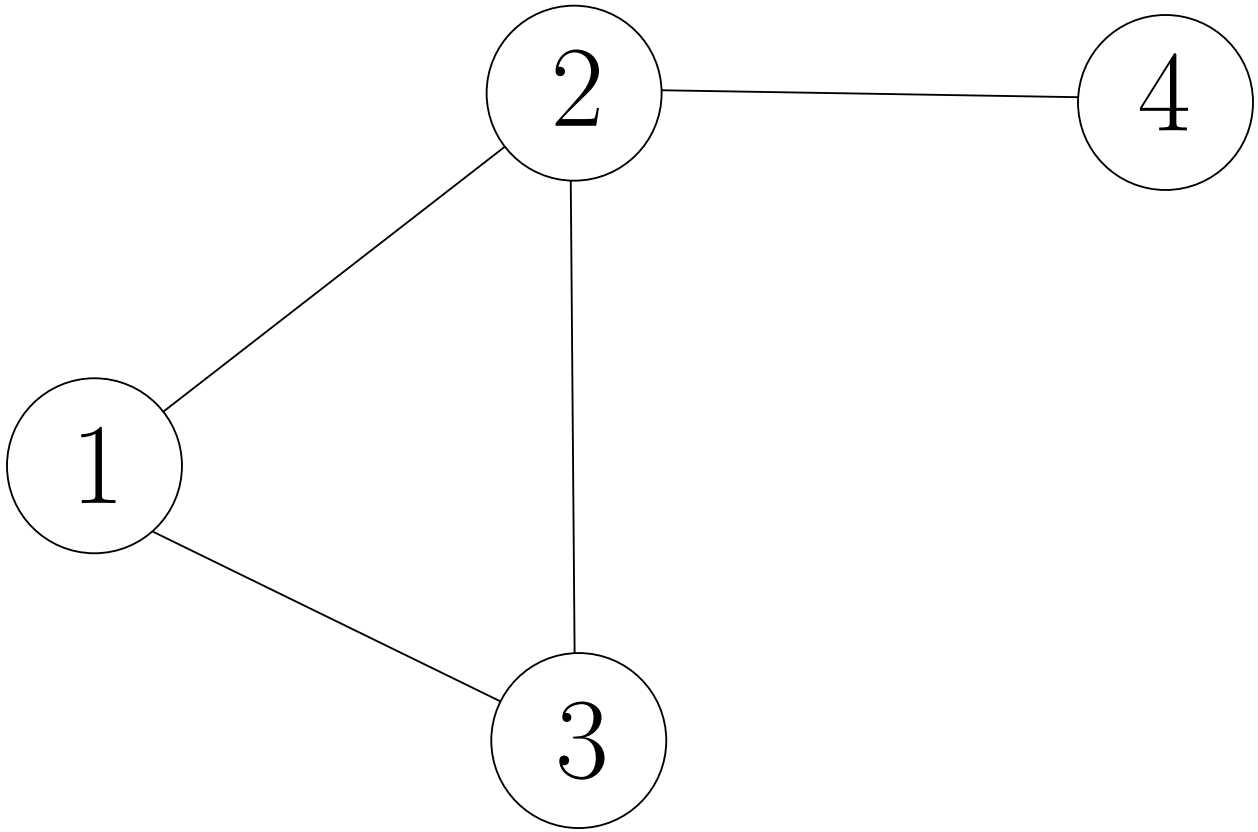
- Slogan: Observed = Fitted for every marginal table corresponding to a complete subgraph.
- The same likelihood equations hold for all hierarchical models, where the margins a correspond to the highest order interaction terms in the model.

CLIQUE = maximal complete subgraph.

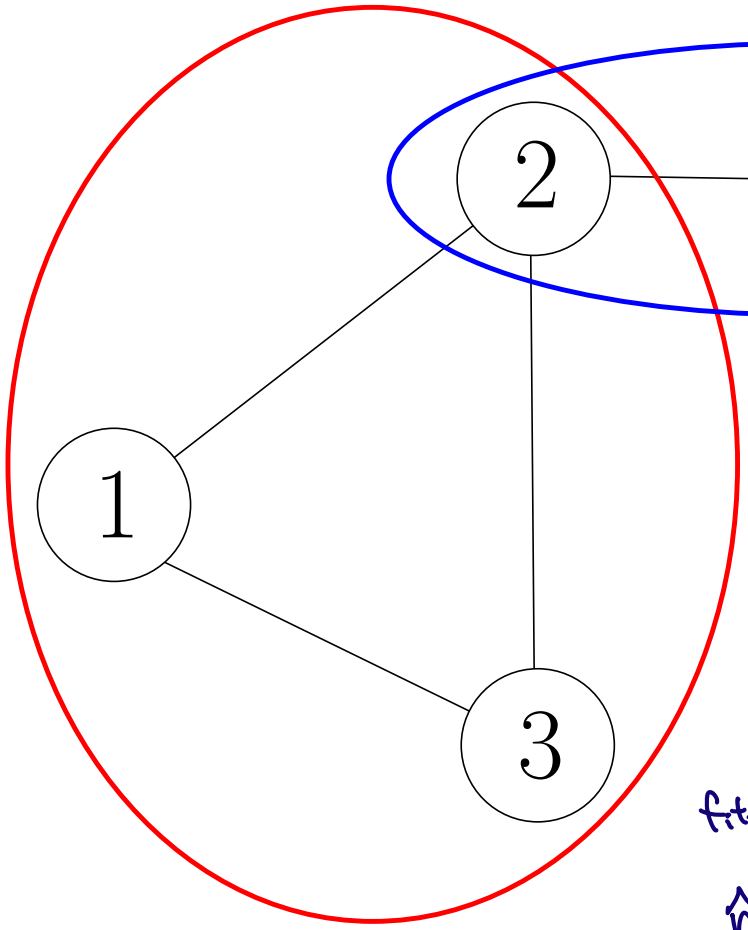
all pairs of nodes are connected by an edge



Determine the cliques



Observed=Fitted for margins corresponding to cliques



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

$P(X, Y \mid Z) = P(X \mid Z) P(Y \mid Z)$ if $X \perp\!\!\!\perp Y \mid Z$

$$\hat{P}(X_1, X_2, X_3, X_4) = \hat{P}(X_1, X_3, X_4 \mid X_2) \hat{P}(X_2)$$
$$= \hat{P}(X_1, X_3 \mid X_2) \hat{P}(X_4 \mid X_2) \hat{P}(X_2)$$
$$= \frac{\hat{P}(X_1, X_2, X_3)}{\hat{P}(X_2)} \hat{P}(X_2, X_4)$$
$$= \frac{\hat{P}(X_1, X_2, X_3) \hat{P}(X_2, X_4)}{\hat{P}(X_2)}$$

fitted count

$$\hat{n}(X_1, X_2, X_3, X_4) = \frac{\hat{n}(X_1, X_2, X_3) \hat{n}(X_2, X_4)}{\hat{n}(X_2)}$$

$$\hat{n}(x_1, x_2, x_3) = n(x_1, x_2, x_3)$$

$$= \frac{n(X_1, X_2, X_3) \hat{n}(X_2, X_4)}{n(X_2)}$$

$n(X_2)$ also ↳ subset of CLIQUE

$$\hat{n}(x_2, x_4) = n(x_2, x_4)$$

-Observed=Fitted for margins corresponding to cliques

We can see as follows why this has to be the case:

1. If there are no constraints to fit an observed table of counts, then the parameter estimates that yield fitted counts equal to the observed counts maximize the likelihood function. For example, the saturated model will yield fitted counts identical to the observed counts.
2. By definition, a graphical model is arbitrary (has no constraints) except for the constraints that can be read from the independence graph.
3. Suppose a forms a clique in the independence graph. Now consider the partitioning $X = (X_a, X_b)$ where b contains all variables not in a.
   We can write (product rule):

$$P(X) = P(X_a)P(X_b \mid X_a)$$

   Since $P(X_a)$ is not constrained by the model (complete graph), all model constraints apply only to $P(X_b \mid X_a)$. Therefore, the maximum likelihood estimates will yield $\hat{n}_a = n_a$.

Maximum Likelihood Estimation: Example

$$
\begin{aligned}
\hat{P}(x_1, x_2, x_3, x_4) &= \hat{P}(x_1, x_3, x_4 | x_2)\hat{P}(x_2) && \text{(product rule)} \\
&= \hat{P}(x_1, x_3 | x_2)\hat{P}(x_4 | x_2)\hat{P}(x_2) && (X_4 \perp\!\!\!\perp (X_1, X_3) \mid X_2) \\
&= \hat{P}(x_1, x_3 | x_2)\hat{P}(x_2, x_4) && \text{(product rule)} \\
&= \frac{\hat{P}(x_1, x_2, x_3)\hat{P}(x_2, x_4)}{\hat{P}(x_2)} && \text{(product rule)}
\end{aligned}
$$

In terms of counts we have (multiply by N on left and N^2/N on right):

$$
\begin{aligned}
\hat{n}(x_1, x_2, x_3, x_4) &= \frac{\hat{n}(x_1, x_2, x_3)\hat{n}(x_2, x_4)}{\hat{n}(x_2)} \\
&= \frac{n(x_1, x_2, x_3)n(x_2, x_4)}{n(x_2)} && \text{(fitted = observed for complete subgraph)}
\end{aligned}
$$

In this case we have a closed form solution for the maximum likelihood fitted counts.

Note that if fitted = observed for margin a, then it follows that observed = fitted for every subset of a as well.

- ML Estimation: Numeric Example

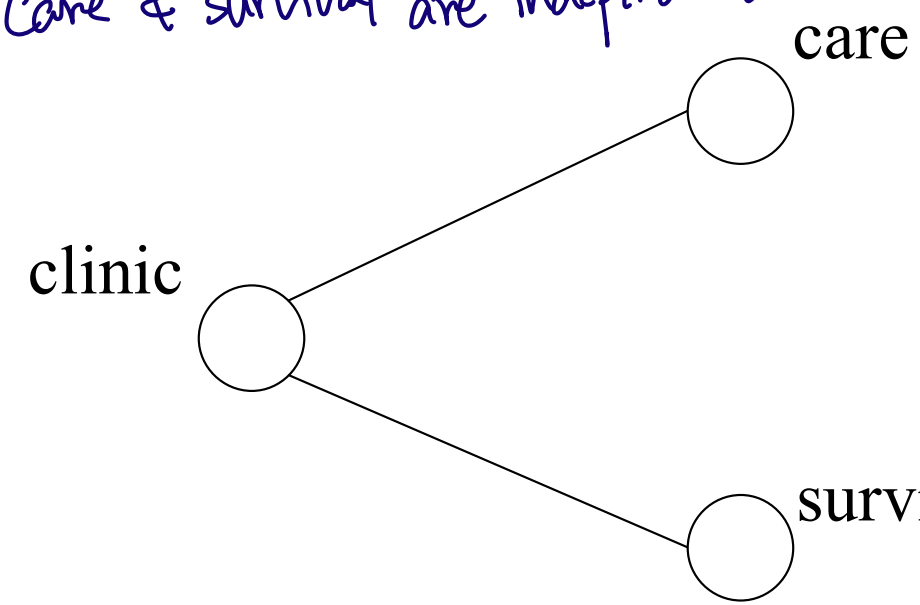
$n(cl, ca, s)$		survival	
clinic	care	no	yes
clinic 1	less	3	176
	more	4	293
clinic 2	less	17	197
	more	2	23

caring rate ↓ ↓

care.

	less	more
clinic 1	179	297
clinic 2	214	25

Within each clinic, Care & survival are independent



survival	no	yes	
care: less	20	373	393
care: more	6	316	322
	26	689	

$$CPR(X_1, X_2) = \frac{P(0,0)P(1,1)}{P(0,1)P(1,0)}$$

$$CPR(\text{care}, \text{survival}) = \frac{n(\text{less}, \text{no})\, n(\text{more}, \text{yes})}{n(\text{less}, \text{yes})\, n(\text{more}, \text{no})}$$

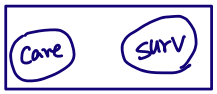
$$= \frac{20 \times 316}{373 \times 6} = 2.82 > 1$$

(positive association)

Fit indepence mode

	no	yes	
less	14	373	393
more	12	310	322
	26	689	715

$\frac{393 \times 26}{715}$



Margin Constraints and Sufficient Statistics

The margin constraints are:

1. $\hat{n}(\text{clinic}, \text{care}) = n(\text{clinic}, \text{care})$ ∴ Cliques
2. $\hat{n}(\text{clinic}, \text{survival}) = n(\text{clinic}, \text{survival})$

$n(cl, ca)$	care	
clinic	less	more
clinic 1	179	297
clinic 2	214	25

$n(cl, s)$	survival	
clinic	no	yes
clinic 1	7	469
clinic 2	19	220

Computation of fitted values

The maximum likelihood fitted counts are given by:

$$\hat{n}(\text{clinic}, \text{care}, \text{survival}) = \frac{n(\text{clinic}, \text{care}) n(\text{clinic}, \text{survival})}{n(\text{clinic})}$$

For example:

$$\hat{n}(\text{clinic 1}, \text{less}, \text{yes}) = \frac{n(\text{clinic 1}, \text{less}) n(\text{clinic 1}, \text{yes})}{n(\text{clinic 1})} = \frac{179 \times 469}{476} = 176.37$$

Observed counts and fitted counts

$n(cl, ca, s)$		survival	
clinic	care	no	yes
clinic 1	less	3	176
	more	4	293
clinic 2	less	17	197
	more	2	23

fited count

$\hat{n}(cl, ca, s)$		survival	
clinic	care	no	yes
clinic 1	less	2.63	176.37
	more	4.37	292.63
clinic 2	less	17.01	196.99
	more	1.99	23.01

Model seems to fit very well!

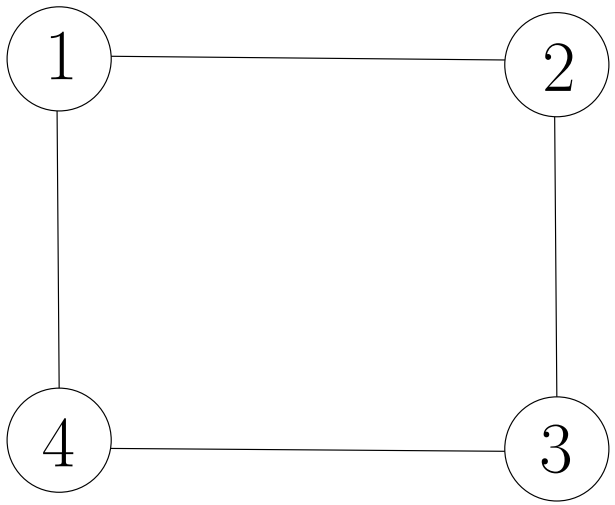
Overview of Coming Lectures

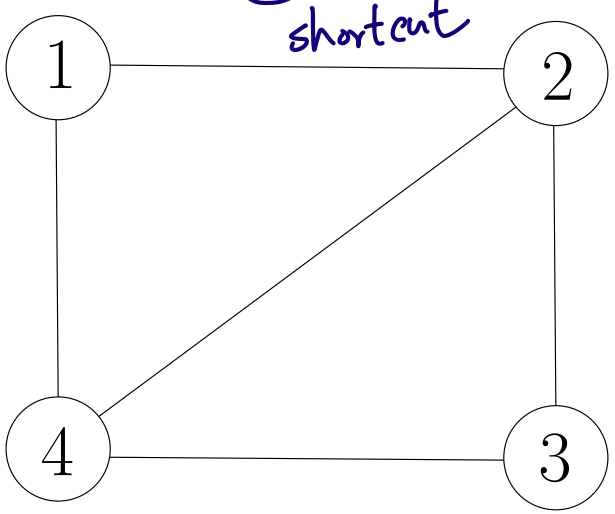
- Introduction
- Independence and Conditional Independence
- Graphical Representation of Conditional Independence
- Log-linear Models
  - Hierarchical
  - Graphical
  - ☞ Decomposable (subsets of the model)
- Maximum Likelihood Estimation
- Model Testing
- Model Selection

Decomposable Graphical Models

- Decomposable graphical models have explicit formulas for the maximum likelihood estimates.
- Decomposable models have *triangulated* independence graphs: they have no *chordless cycles* of length greater than three.
- A cycle is chordless if only the *successive* pairs of vertices in the cycle are connected by an edge. (There is no "shortcut" in the cycle).

Example





- The left graph is *not* decomposable because it contains the chordless cycle $1 - 2 - 3 - 4 - 1$.
- The graph on the right *is* decomposable.
  The cycle $1 - 2 - 3 - 4 - 1$ is no longer chordless because 2 and 4 are adjacent in the graph but not successive in the cycle.

Maximum Likelihood Estimation for Decomposable Models

An ordering $C_1, \ldots, C_m$ of the cliques of the graph has the running intersection property (RIP) iff

$$C_j \cap (C_1 \cup \ldots \cup C_{j-1}) \subseteq C_i,$$

for some $i < j$, and for all $j = 2, \ldots, m$.

We define the corresponding separator sets

$$S_j = C_j \cap (C_1 \cup \ldots \cup C_{j-1}),$$

with $S_1 = \emptyset$.

ML Estimation for Decomposable Models

Let $C_1, \ldots, C_m$ be a RIP ordering of the cliques, with separator sets $S_1, \ldots, S_m$.

The maximum likelihood fitted counts are given by

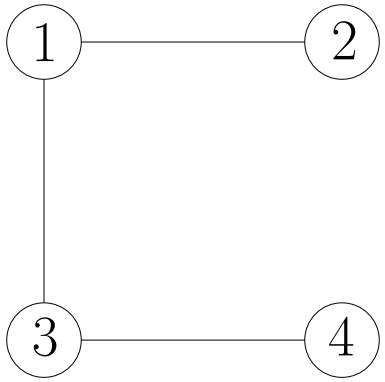
$$\hat{n}(x) = \frac{\prod_{j=1}^{m} n(x_{C_j})}{\prod_{j=2}^{m} n(x_{S_j})}$$

where $n(x_\emptyset) = N$.

Likewise, the maximum likelihood fitted probabilities are given by

$$\hat{P}(x) = \frac{\prod_{j=1}^{m} n(x_{C_j})}{\prod_{j=1}^{m} n(x_{S_j})} = \frac{\prod_{j=1}^{m} n(x_{C_j})}{N \prod_{j=2}^{m} n(x_{S_j})}$$

Example

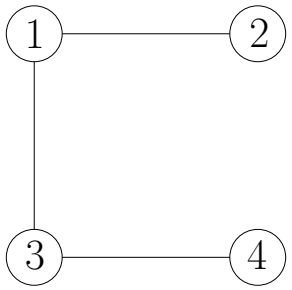


The clique ordering $C_1 = \{1, 2\}, C_2 = \{3, 4\}, C_3 = \{1, 3\}$ does not have the running intersection property and does therefore not produce correct estimates.

j	C_j	S_j	RIP?
1	$\{1, 2\}$	$\emptyset$	–
2	$\{3, 4\}$	$\emptyset$	✓
3	$\{1, 3\}$	$\{1, 3\}$	✗

Not subset of

Example



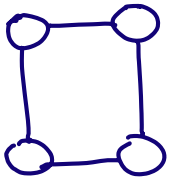
The clique ordering $C_1 = \{1, 2\}, C_2 = \{1, 3\}, C_3 = \{3, 4\}$ does have the running intersection property:

j	C_j	S_j	RIP?
1	$\{1, 2\}$	$\emptyset$	–
2	$\{1, 3\}$	$\{1\}$	✓
3	$\{3, 4\}$	$\{3\}$	✓

The corresponding ML fitted counts are:

$$\hat{n}(x_1, x_2, x_3, x_4) = \frac{n(x_1, x_2)n(x_1, x_3)n(x_3, x_4)}{n(x_1)n(x_3)}$$

Iterative Proportional Fitting (IPF)



- If the model is not decomposable, there is no closed form solution (formula) for the maximum likelihood estimates.
- We need iterative algorithms to compute the fitted counts/probabilities.
- Iterative Proportional Fitting (IPF) is such an algorithm.

IPF: Example

Fit independence model to

$n(x_1, x_2)$		$x_2 = 0$	$x_2 = 1$	$n_1(x_1)$
	$x_1 = 0$	20	10	30
	$x_1 = 1$	40	30	70
	$n_2(x_2)$	60	40	100

The margin constraints for the independence model are

1. $\hat{n}_1(x_1) = n_1(x_1)$, and
2. $\hat{n}_2(x_2) = n_2(x_2)$

IPF makes the fitted counts agree with the observed counts on each margin in turn.

Iterative Proportional Fitting

We begin with a table $\hat{n}^{(0)}$ of uniform counts (left)

$\hat{n}^{(0)}$	0	1	
0	1	1	2
1	1	1	2
	2	2	

	0	1	
0			30
1			70

First step: fit to row margin

$$\hat{n}(x_1, x_2)^{(1)} = n_1(x_1) \times \hat{P}(x_2 \mid x_1)^{(0)} = n_1(x_1) \times \frac{\hat{n}(x_1, x_2)^{(0)}}{\hat{n}_1(x_1)^{(0)}}$$

We compute (row 1):

$$\hat{n}(0,0)^{(1)} = 30 \times \frac{1}{2} = 15 \qquad \hat{n}(0,1)^{(1)} = 30 \times \frac{1}{2} = 15$$

Iterative Proportional Fitting

We begin with a table $\hat{n}^{(0)}$ of uniform counts (left)

$\hat{n}^{(0)}$	0	1	
0	1	1	2
1	1	1	2
	2	2	

	0	1	
0	15	15	30
1			70

First step: fit to row margin

$$\hat{n}(x_1, x_2)^{(1)} = n_1(x_1) \times \hat{P}(x_2 \mid x_1)^{(0)} = n_1(x_1) \times \frac{\hat{n}(x_1, x_2)^{(0)}}{\hat{n}_1(x_1)^{(0)}}$$

We compute (row 1):

$$\hat{n}(0,0)^{(1)} = 30 \times \frac{1}{2} = 15 \qquad \hat{n}(0,1)^{(1)} = 30 \times \frac{1}{2} = 15$$

Iterative Proportional Fitting

$\hat{n}^{(0)}$	0	1	
0	1	1	2
1	1	1	2
	2	2	

	0	1	
0	15	15	30
1			70

First step continued (row 2):

$$\hat{n}(1,0)^{(1)} = 70 \times \frac{1}{2} = 35 \qquad \hat{n}(1,1)^{(1)} = 70 \times \frac{1}{2} = 35$$

which yields $\hat{n}^{(1)}$:

$\hat{n}^{(1)}$	0	1	
0	15	15	30
1	35	35	70
	50	50	

Iterative Proportional Fitting

$\hat{n}^{(1)}$	0	1	
0	15	15	30
1	35	35	35
	50	50	

	0	1
0		
1		
	60	40

Second step: fit to column margin

$$\hat{n}(x_1, x_2)^{(2)} = n_2(x_2) \times \hat{P}(x_1 \mid x_2)^{(1)} = n_2(x_2) \times \frac{\hat{n}(x_1, x_2)^{(1)}}{\hat{n}_2(x_2)^{(1)}}$$

Which gives (first column):

$$\hat{n}(0,0)^{(2)} = 60 \times \frac{15}{50} = 18 \qquad \hat{n}(1,0)^{(2)} = 60 \times \frac{35}{50} = 42$$

	0	1
0	18	
1	42	
	60	40

IPF

Likewise for the second column:

$$\hat{n}(0,1)^{(2)} = 40 \times \frac{15}{50} = 12 \qquad \hat{n}(1,1)^{(2)} = 40 \times \frac{35}{50} = 28$$

This yields $\hat{n}^{(2)}$:

$\hat{n}^{(2)}$	0	1	
0	18	12	30
1	42	28	70
	60	40	

Notice that the row totals are still 30 and 70, so we have simultaneously satisfied the conditions

$$\hat{n}_1(x_1) = n_1(x_1) \text{ and } \hat{n}_2(x_2) = n_2(x_2)$$

so the algorithm has converged.

IPF: General Algorithm Sketch

Say we have m margins $\{a_1, a_2, \ldots, a_m\}$ to be fitted $(\cup_i a_i = K)$.

We have to find a table $\hat{n}(x)$ that agrees with the observed table $n(x)$ on the m margins corresponding to the subsets a_i.

The algorithm cycles through the list of subsets

$$a = a_i, \qquad\qquad i = 1, 2, \ldots, m$$

fitting $\hat{n}(x)$ to each margin in turn.

This is repeated until convergence is reached, i.e. all margin constraints are (approximately) satisfied simultaneously.

IPF updating rule

To fit to the margin a, the observed count $n_a(x_a)$ on x_a is distributed over the cells $x = (x_a, x_b)$ of the full table according to

$$\hat{n}_{ab}(x_a, x_b)^{(t+1)} = n_a(x_a)\hat{P}(x_b|x_a)^{(t)}$$

where b is the complement of a, and

$$\hat{P}(x_b|x_a)^{(t)} = \frac{\hat{n}_{ab}(x_a, x_b)^{(t)}}{\hat{n}_a(x_a)^{(t)}},$$

is the current estimate of the conditional probability $P(X_b = x_b|X_a = x_a)$.

IPF Pseudocode

Algorithm 1 IPF($n(x)$, $\mathcal{A}$)

$t \leftarrow 0$
for all values x of X **do**
  $\hat{n}(x)^{(t)} \leftarrow 1$
end for
repeat
  for all margins $a \in \mathcal{A}$ **do**
    $b \leftarrow K \setminus a$
    for all values x_a of X_a **do**
      for all values x_b of X_b **do**
        $\hat{n}_{ab}(x_a, x_b)^{(t+1)} \leftarrow n_a(x_a) \left(\frac{\hat{n}_{ab}(x_a, x_b)^{(t)}}{\hat{n}_a(x_a)^{(t)}} \right)$
      end for
    end for
    $t \leftarrow t + 1$
  end for
until convergence

Some properties of IPF

- For the algorithm to produce the correct result, the initial solution must satisfy the constraints of the model to be fitted. A uniform table of counts is a safe choice since it sets all u-terms except $u_\emptyset$ to zero.
- If the cliques of a decomposable model are presented in RIP order to IPF, then the algorithm will converge in one iteration (one cycle through all cliques).
- Otherwise IPF will converge in two iterations on decomposable models.

Overview of Coming Lectures

- Introduction
- Independence and Conditional Independence
- Graphical Representation of Conditional Independence
- Log-linear Models
  - Hierarchical
  - Graphical
  - Decomposable
- Maximum Likelihood Estimation
- ☞ Model Testing
- Model Selection

The Log-Likelihood score of a model

The likelihood score of a model M is

$$L^M = \prod_x \hat{P}^M(x)^{n(x)},$$

where $\hat{P}^M(x)$ is the fitted probability of cell x according to model M.

Hence, the likelihood score of model M is the probability of the observed data using the fitted cell probabilities according to model M.

Likewise, the log-likelihood score of a model M is

$$\mathcal{L}^M = \sum_x n(x) \log \hat{P}^M(x)$$

Example

Suppose we have data

$n(x)$	$x_2 = 0$	$x_2 = 1$	$n(x_1)$
$x_1 = 0$	30	10	40
$x_1 = 1$	30	30	60
$n(x_2)$	60	40	100

The independence model gives probability estimates:
$\hat{P}(0,0) = 0.24$, $\hat{P}(0,1) = 0.16$, $\hat{P}(1,0) = 0.36$, $\hat{P}(1,1) = 0.24$.

The probability of the observed data according to this model is

$$0.24^{30} \times 0.16^{10} \times 0.36^{30} \times 0.24^{30}$$

This is the likelihood score of the model given the data.
The corresponding log-likelihood score is

$$\mathcal{L} = 30 \log 0.24 + 10 \log 0.16 + 30 \log 0.36 + 30 \log 0.24 \approx -134.6$$

Example (continued)

$n(x)$	$x_2 = 0$	$x_2 = 1$	$n(x_1)$
$x_1 = 0$	30	10	40
$x_1 = 1$	30	30	60
$n(x_2)$	60	40	100

The saturated model gives probability estimates:
$\hat{P}(0,0) = 0.3$, $\hat{P}(0,1) = 0.1$, $\hat{P}(1,0) = 0.3$, $\hat{P}(1,1) = 0.3$.

$\frac{30}{100}$ $\frac{10}{100}$

The probability of the observed data according to this model is

$$0.3^{30} \times 0.1^{10} \times 0.3^{30} \times 0.3^{30}$$

The corresponding log-likelihood score is

$$\mathcal{L} = 30 \log 0.3 + 10 \log 0.1 + 30 \log 0.3 + 30 \log 0.3 \approx -131.4$$

higher (Compare to other model)

Of course this is better than the independence model.

can never do better than saturated model

Model Deviance

Since for the saturated (unconstrained) model

$$\hat{P}(x) = \frac{n(x)}{N},$$

the log-likelihood score of the saturated model is

$$\mathcal{L}^{\text{sat}} = \sum_x n(x) \log \frac{n(x)}{N}$$

The deviance of a fitted model compares the log-likelihood score of the fitted model to that of the saturated model.

difference

The larger the model deviance, the poorer the fit.

↑ worst prediction

Model Deviance

Deviance of M is 2 (log-likelihood of the saturated model $-$ log-likelihood of M):

$$\begin{aligned}
\text{dev}(M) &= 2(\mathcal{L}^{\text{sat}} - \mathcal{L}^{M}) \\
&= 2\left(\sum_x n(x)\log\frac{n(x)}{N} - \sum_x n(x)\log\hat{P}^{M}(x)\right) \\
&= 2\left(\sum_x n(x)\left(\log\frac{n(x)}{N} - \log\hat{P}^{M}(x)\right)\right) \\
&= 2\sum_x n(x)\log\frac{n(x)}{N\hat{P}^{M}(x)}
\end{aligned}$$

which can be summarised by the *slogan*

$$2\sum_{\text{cells}} \text{observed} \times \log\frac{\text{observed}}{\text{fitted}}$$

Deviance difference

Let $M_0 \subseteq M_1$, that is M_0 is the simpler model
(the u-terms present in M_0 are a subset of the u-terms present in M_1).

better

The *deviance difference* between M_0 and M_1 is

$$\text{dev}(M_0) - \text{dev}(M_1) = -2\mathcal{L}^{M_0} + 2\mathcal{L}^{M_1} = 2(\mathcal{L}^{M_1} - \mathcal{L}^{M_0})$$

For large N we have that:

$$2(\mathcal{L}^{M_1} - \mathcal{L}^{M_0}) \approx_{M_0} \chi^2_\nu$$

χ^2_ν: chi-square distribution with ν degrees of freedom.
ν: number of *additional* restrictions (zero u-terms) of M_0 compared to M_1.

We reject the null hypothesis that M_0 is the true model when

$$2(\mathcal{L}^{M_1} - \mathcal{L}^{M_0}) > \chi^2_{\nu;\alpha},$$

where α is the significance level of the test, and $P(X^2 > \chi^2_{\nu;\alpha}) = \alpha$, that is, $\chi^2_{\nu;\alpha}$ is the *critical value*.

The test is called a likelihood ratio test because

$$\log \frac{L^{M_1}}{L^{M_0}} = \log L^{M_1} - \log L^{M_0} = \mathcal{L}^{M_1} - \mathcal{L}^{M_0}$$

Model Testing: example

Does

$$\text{survival} \perp\!\!\!\perp \text{care} \mid \text{clinic}$$

give a good fit of the observed table? Test against the saturated model.

Compute the deviance

u(surv, care) = 0

u(clinc, care, surv) = 0

$$2 \sum_{\text{cells}} \text{observed} \times \log \frac{\text{observed}}{\text{fitted}} \approx 0.082$$

$$\chi^2_{2;0.05} \approx 6$$

So we "accept" the model.

Note: since M_1 is the saturated model, the deviance difference between M_0 and M_1 is just the deviance of M_0.

Fitted Counts and Observed Counts

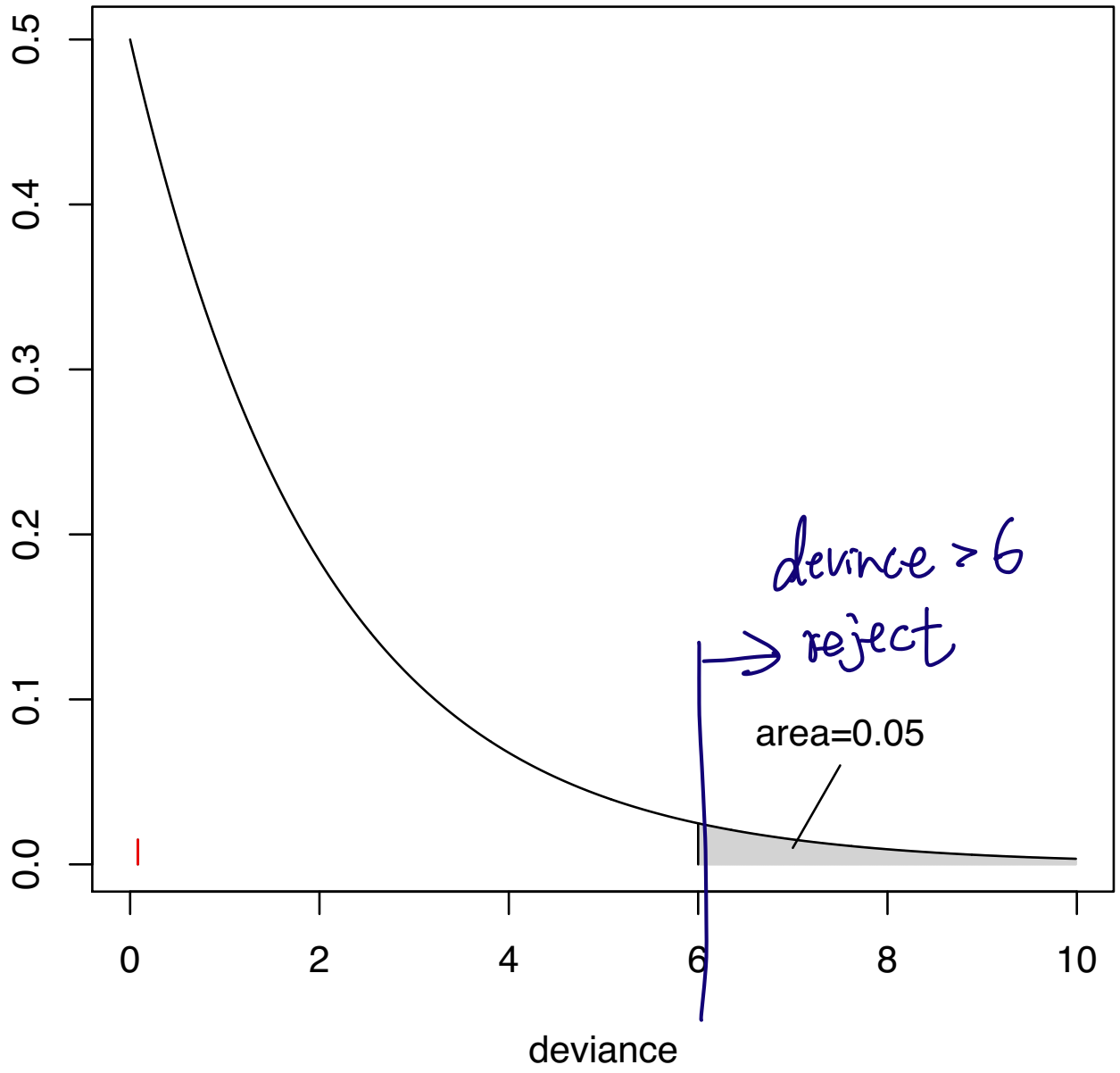
natural logarithm

Dev = $2[3 \log\frac{3}{2.63} + 23\log\frac{23}{23.0}] = 0.082$

$\hat{n}$(clinic, care, survival)		survival	
clinic	care	no	yes
clinic 1	less	2.63	176.37
	more	4.37	292.63
clinic 2	less	17.01	196.99
	more	1.99	23.01

n(clinic, care, survival)		survival	
clinic	care	no	yes
clinic 1	less	3	176
	more	4	293
clinic 2	less	17	197
	more	2	23

Test of survival ⫫ care|clinic; χ^2_2 distribution.



Model Testing: example

Does the mutual independence model give a good fit of the observed table? Test against the saturated model.

u(surv, care) = 0

u(clinc, care, surv) = 0

u(clinc, care) = 0

u(clinc, surv) = 0

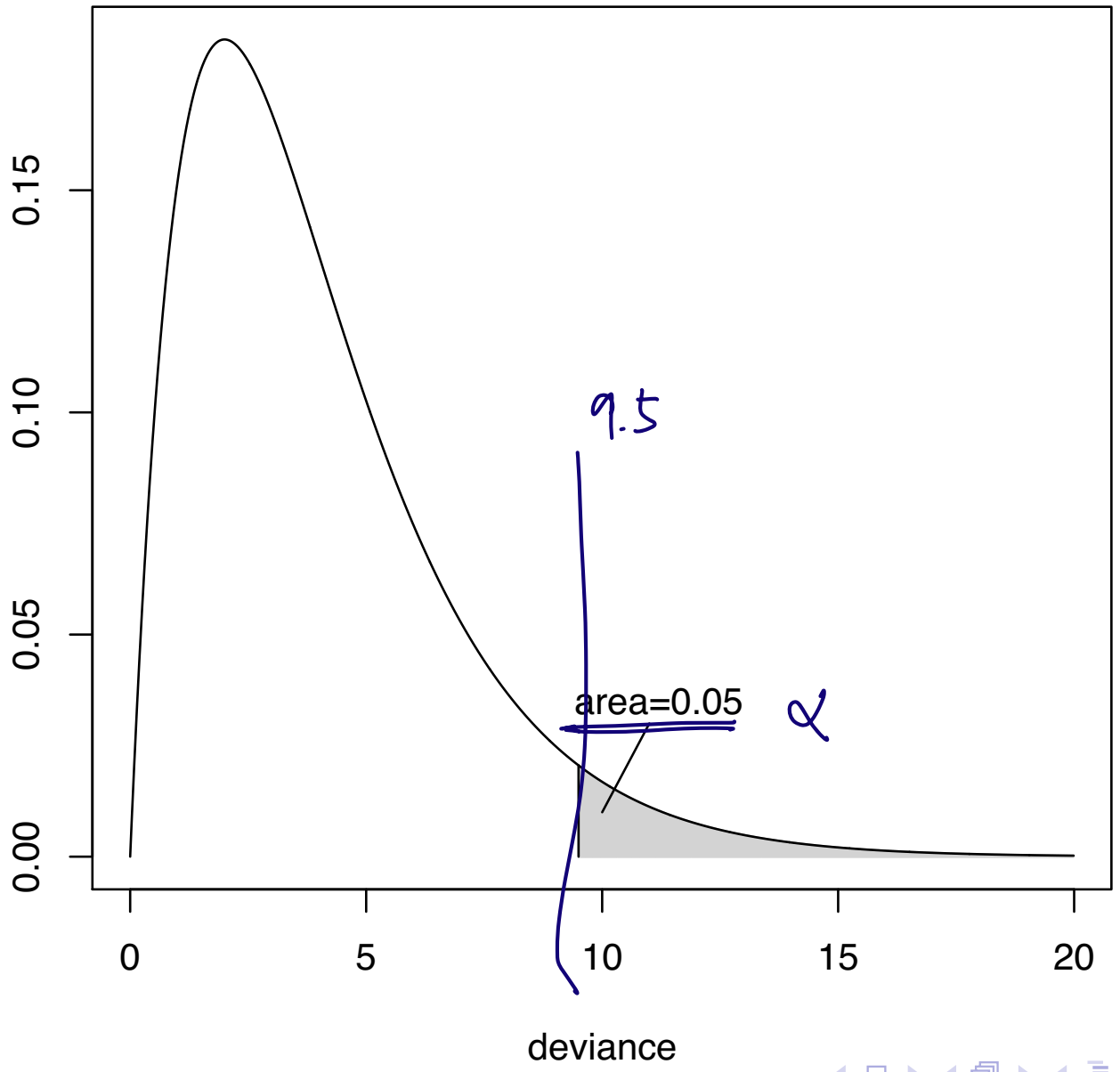
Compute the deviance

$$2 \sum_{\text{cells}} \text{observed} \times \log \frac{\text{observed}}{\text{fitted}} \approx 211$$

$$\chi^2_{4;0.05} \approx 9.5$$

So we reject the mutual independence model.

Test of Independence Model; χ^2_4 distribution.



Overview of Coming Lectures

- Introduction
- Independence and Conditional Independence
- Graphical Representation of Conditional Independence
- Log-linear Models
  - Hierarchical
  - Graphical
  - Decomposable
- Maximum Likelihood Estimation
- Model Testing
- ☞ Model Selection

Model Selection

The Problem: find a good model for a high-dimensional table when little prior knowledge is available.

Solution: Search the space of possible models.
Two approaches:

- Use significance testing
- Use a quality function

Quality Functions: AIC and BIC

Two components:

- the lack-of-fit of the model
- complexity of the model

Akaike's Information Criterion assigns quality to model M as follows

$$\mathrm{AIC}(M) = \mathrm{dev}(M) + 2\,\dim(M)$$

where $\dim(M)$ is the number of parameters (the number of u-terms) of the model.

Bayesian Information Criterion assigns quality to model M as follows

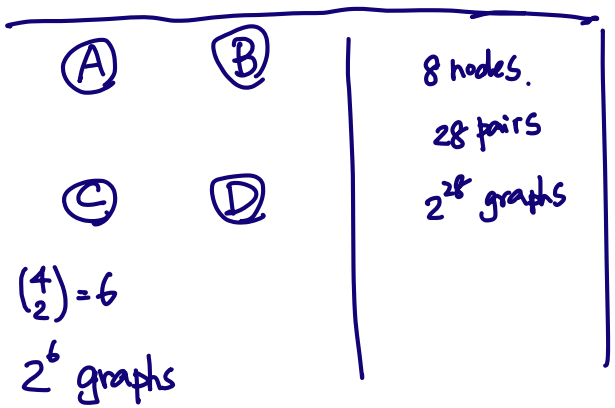
$$\mathrm{BIC}(M) = \mathrm{dev}(M) + \log(N)\,\dim(M)$$

Search

Exhaustive search is usually not feasible.

A straightforward approach is local search with hill climbing:

1. pick some initial model
2. consider the quality of all neighbors of the current model
3. if they all have lower quality, stop and return the current model.
4. otherwise move to the neighbor with highest quality and return to 2.



Start from random model

local optimization

Example: Decomposable Graphical Models

why not cross validation?

Hill climbing local search with decomposable graphical models and AIC scoring.

1. pick an initial model, e.g. the empty graph
2. neighbors
   - add an edge
   - delete an edge

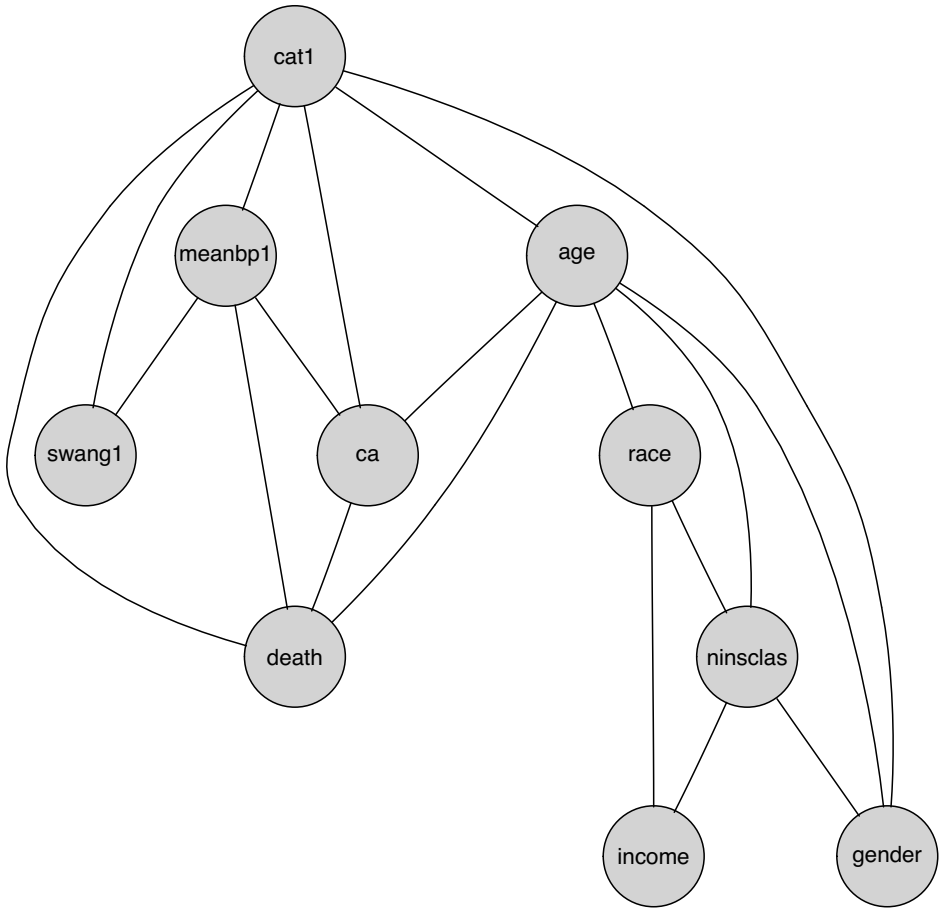
   (without creating a chordless cycle of length > 3)
3. if all neighbors have higher AIC, stop and return the current model.
4. otherwise move to the neighbor with lowest AIC and return to 2.

Local search in gRim

```
# fit initial model (empty graph)
> rhc.init <- dmod(~.^1,data=rhc.dat)
# display some info about this model
> summary(rhc.init)
is graphical=TRUE; is decomposable=TRUE
generators (glist):
   :"cat1"
   :"death"
   :"swang1"
   :"gender"
   :"race"
   :"ninsclas"
   :"income"
   :"ca"
   :"age"
   :"meanbp1"
```

Result of Search (AIC, decomposable)

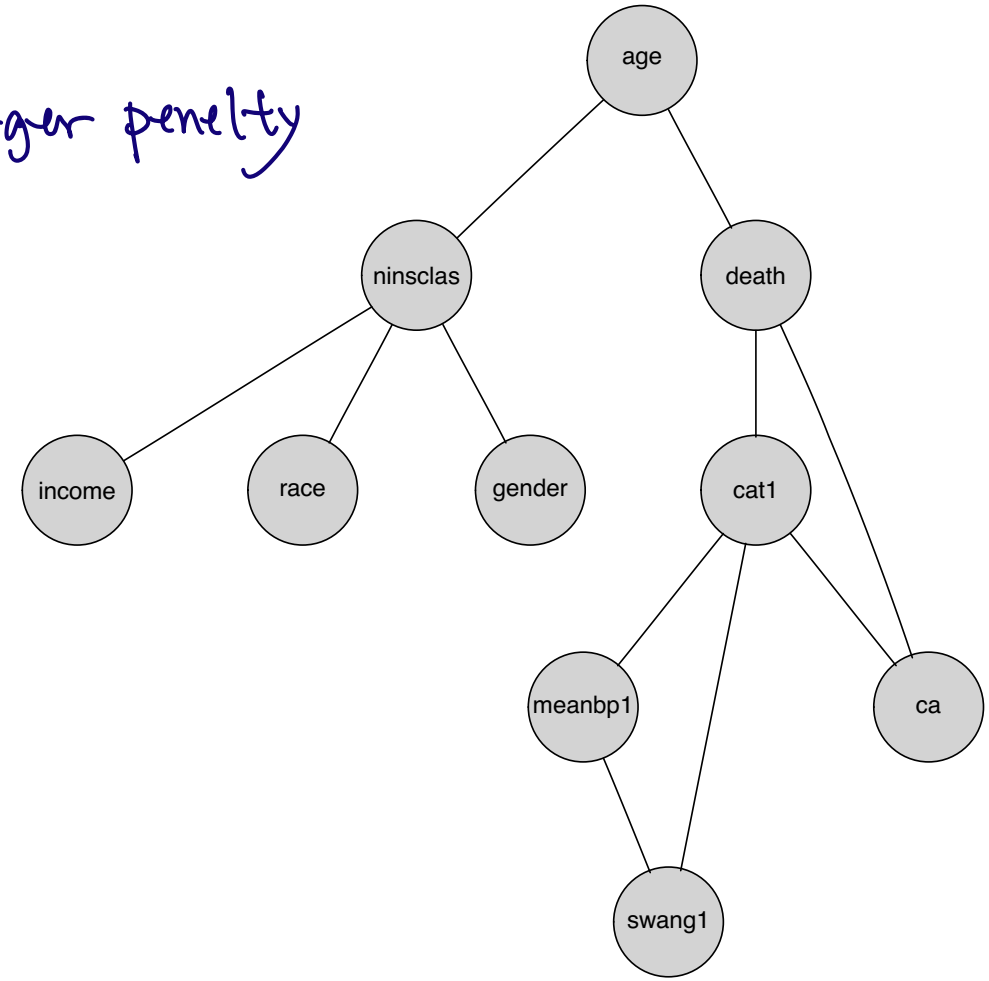
```
# perform stepwise search of decomposable models using AIC (only add edges)
> rhc.step1 <- stepwise(rhc.init,direction="forward")
> plot(rhc.step1)
```



Result of Search (BIC, decomposable)

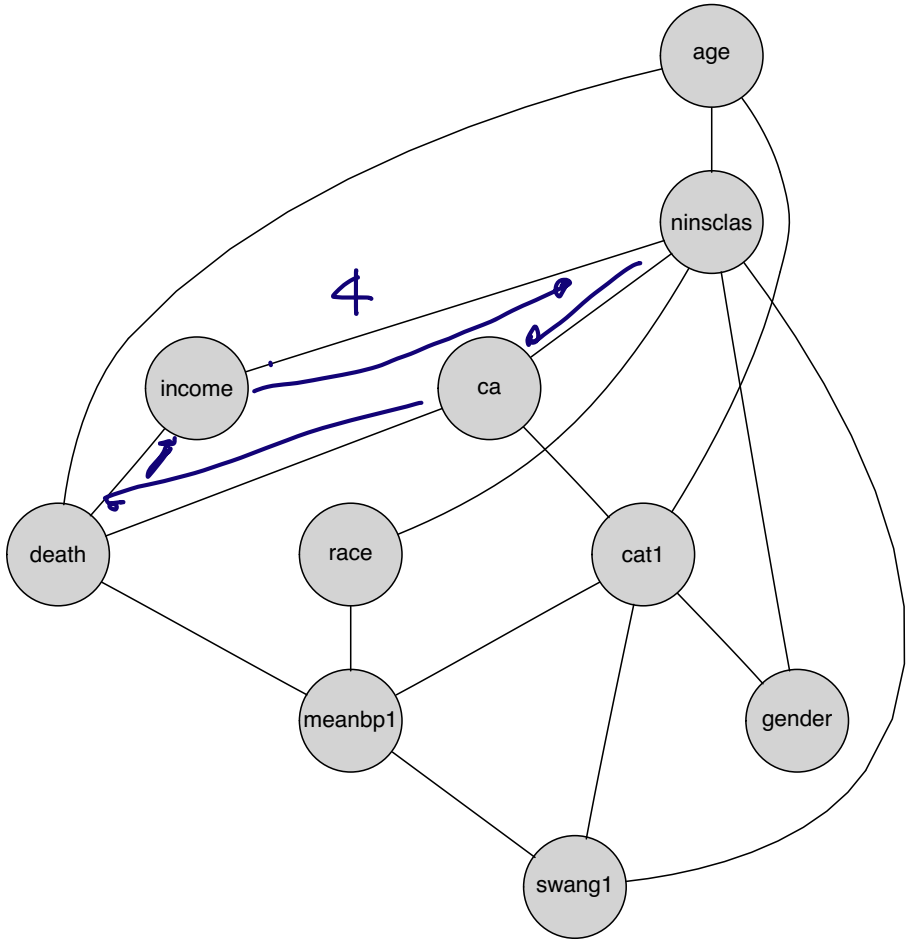
```
> rhc.step2 <- stepwise(rhc.init,direction="forward",k=log(nrow(rhc.dat)))
> plot(rhc.step2)
```

∴ BIC gives higher penalty



Result of Search (BIC, unrestricted)

```
> rhc.step3 <- stepwise(rhc.init,direction="forward",k=log(nrow(rhc.dat)),
                        type="unrestricted")
> plot(rhc.step3)
```



Trace of Search (AIC, unrestricted)

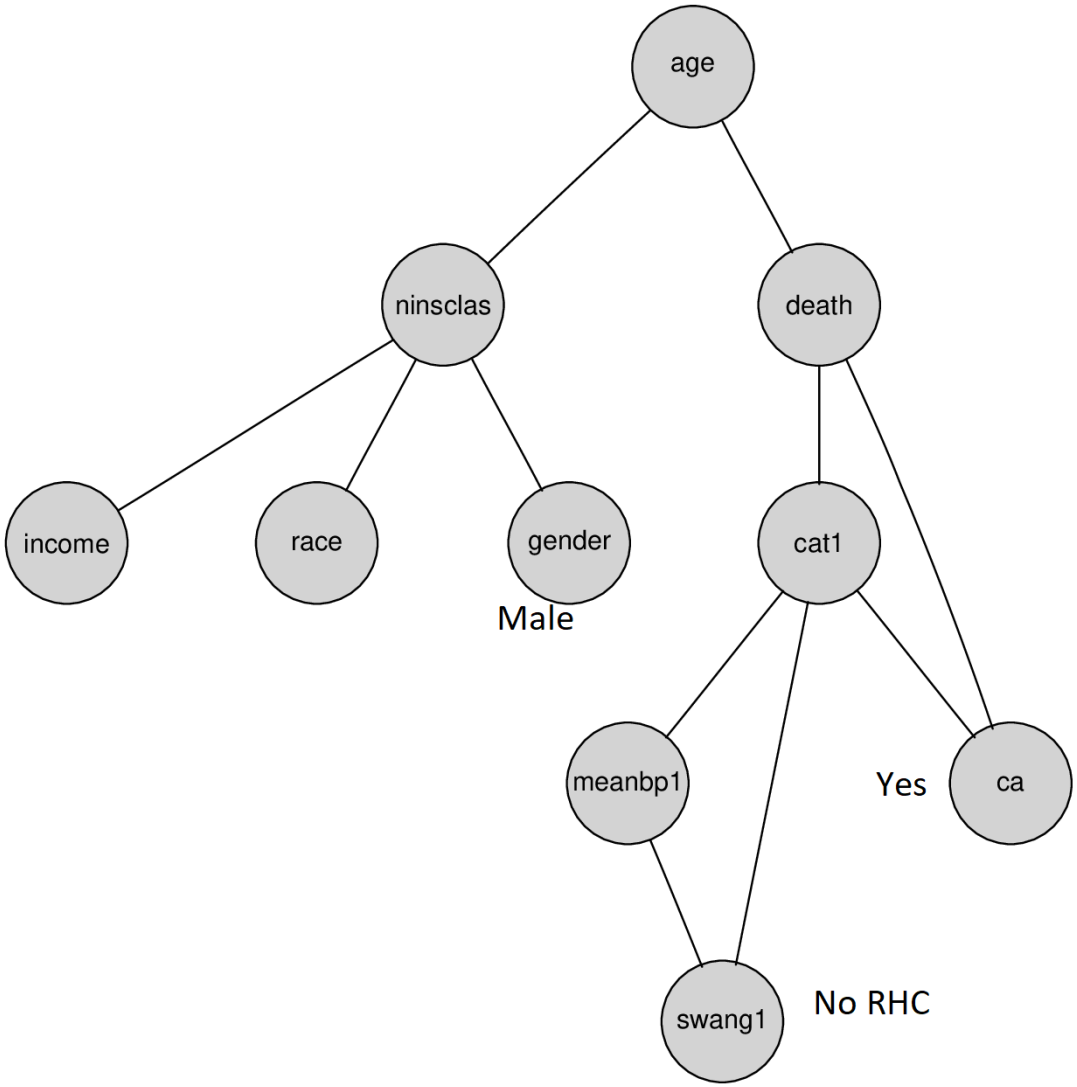
```
> rhc.step4 <- stepwise(rhc.init,direction="both",
                        type="unrestricted",details=1)
STEPWISE:
 criterion: aic ( k = 2 )
 direction: both
 type     : unrestricted
 search   : all
 steps    : 1000
. FORWARD: type=unrestricted search=all, criterion=aic(2.00), alpha=0.00
. Initial model: is graphical=TRUE is decomposable=TRUE
  change.AIC -3061.4959 Edge added: ninsclas age
  change.AIC -1685.0212 Edge added: cat1 ca
  change.AIC -1347.0960 Edge added: income ninsclas
  change.AIC -420.0886 Edge added: swang1 cat1
  change.AIC -306.3228 Edge added: race ninsclas
  change.AIC -285.8656 Edge added: age cat1
  change.AIC -253.7602 Edge added: death ca
  change.AIC -497.6596 Edge added: age death
  ...
```

Model Use: Inference and Prediction with gRain

```
# prepare selected model for use in gRain
> rhc.mod2 <- grain(as(rhc.step2,"graphNEL"),rhc.dat)
# perform inference on "death" from evidence on "gender","ca", and "swang1"
> predict(rhc.mod2, c("death"), c("gender","ca","swang1"),
  data.frame(gender="Male",ca="Yes",swang1="No RHC"), type = "dist")
$pred
$pred$death
             No       Yes
[1,] 0.2235305 0.7764695

# change cancer to "Metastatic"
> predict(rhc.mod2, c("death"), c("gender","ca","swang1"),
  data.frame(gender="Male",ca="Metastatic",swang1="No RHC"), type = "dist")
$pred
$pred$death
              No       Yes
[1,] 0.09421555 0.9057845
```

Evidence entered



Model Use: Inference and Prediction with gRain

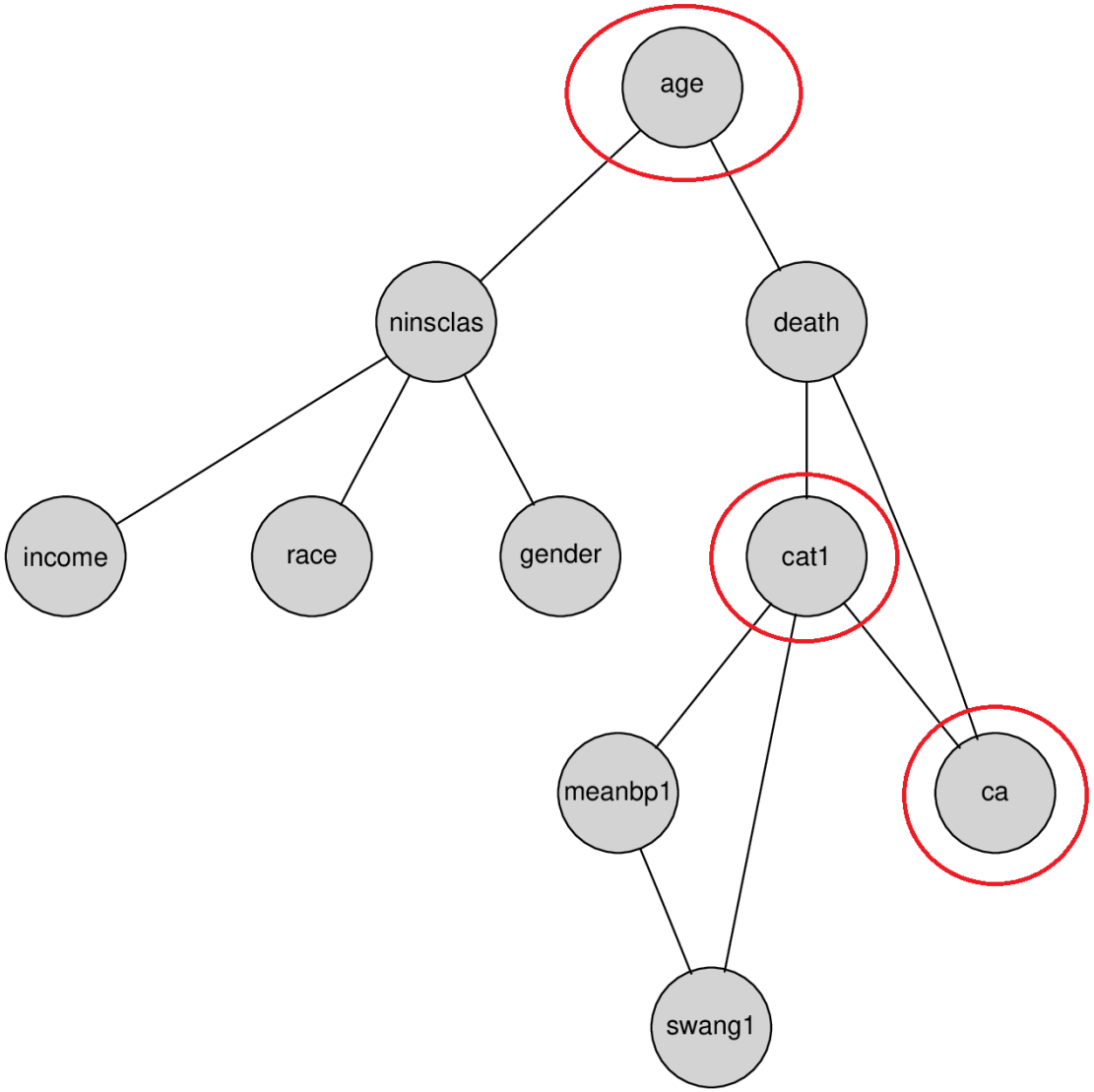
```
# predict death (in-sample) from its Markov blanket
> death.pred <- predict(rhc.mod2, c("death"), c("ca","cat1","age"),
                        rhc.dat, type = "class")
> table(rhc.dat$death,death.pred$pred$death)

       No  Yes
  No   730 1283
  Yes  524 3198

> (730+3198)/nrow(rhc.dat)
[1] 0.6849172

# model is a little better than just predicting the majority class
> summary(rhc.dat$death)
  No  Yes
2013 3722
> 3722/nrow(rhc.dat)
[1] 0.6489974
```

Markov Blanket of Death



Literature

- Y. Bishop, S.E. Fienberg, P.W. Holland, Discrete Multivariate Analysis, MIT Press, 1975.
- J. Whittaker, Graphical Models in Applied Multivariate Statistics, Wiley, 1990.
- D. Edwards, Introduction to Graphical Modelling (2nd edition), Springer, 2000.
- S. Højsgaard, D. Edwards and S. Lauritzen, Graphical Models with R, Springer, 2012.