# Graphical and Log-Linear Models

## MSc Further Statistical Methods, Lecture 2
## Hilary Term 2007

Steffen Lauritzen, University of Oxford; January 18, 2007

# Three-way tables

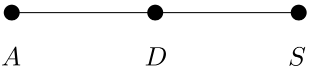
Admissions to Berkeley by department

| Department | Sex | Whether admitted | |
|---|---|---|---|
| | | Yes | No |
| I | Male | 512 | 313 |
| | Female | 89 | 19 |
| II | Male | 353 | 207 |
| | Female | 17 | 8 |
| III | Male | 120 | 205 |
| | Female | 202 | 391 |
| IV | Male | 138 | 279 |
| | Female | 131 | 244 |
| V | Male | 53 | 138 |
| | Female | 94 | 299 |
| VI | Male | 22 | 351 |
| | Female | 24 | 317 |

Here are three variables $A$: Admitted?, $S$: Sex, and $D$: Department.

# Conditional independence

For three variables it is of interest to see whether independence holds for fixed value of one of them, e.g. *is the admission independent of sex for every department separately?* We denote this as $A \perp\!\!\!\perp S \mid D$ and graphically as




Algebraically, this corresponds to the relations

$$p_{ijk} = p_{i+\,|\,k} p_{+j\,|\,k} p_{++k} = \frac{p_{i+k} p_{+jk}}{p_{++k}}.$$

# Marginal and conditional independence

Note that there the two conditions

$$A \perp\!\!\!\perp S, \quad A \perp\!\!\!\perp S \mid D$$

are very different and will typically not both hold unless we either have $A \perp\!\!\!\perp (D, S)$ or $(A, D) \perp\!\!\!\perp S$, i.e. if one of the variables are completely independent of both of the others.

This fact is a simple form of what is known as *Yule–Simpson paradox.*

It can be much worse than this:

A *positive conditional association can turn into a negative marginal association* and vice-versa.

# Admissions revisited

Admissions to Berkeley

| Sex | Whether admitted | |
|---|---|---|
| | Yes | No |
| Male | 1198 | 1493 |
| Female | 557 | 1278 |

Note this marginal table shows much lower admission rates for females.

Considering the departments separately, there is only a difference for department I, and it is the other way around...

# Florida murderers

Sentences in 4863 murder cases in Florida over the six years 1973-78

| | Sentence | |
|---|---|---|
| Murderer | Death | Other |
| Black | 59 | 2547 |
| White | 72 | 2185 |

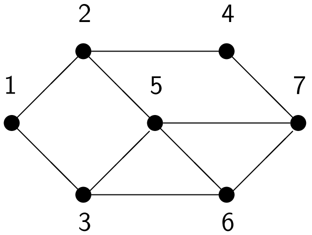
The table shows a greater proportion of white murderers receiving death sentence than black (3.2% vs. 2.3%), although the difference is not big, the picture seems clear.

# Controlling for colour of victim

| | | Sentence | |
|---|---|---|---|
| Victim | Murderer | Death | Other |
| Black | Black | 11 | 2309 |
| | White | 0 | 111 |
| White | Black | 48 | 238 |
| | White | 72 | 2074 |

Now the table for given colour of victim shows a very different picture. In particular, note that 111 white murderers killed black victims and none were sentenced to death.

# Graphical models




For several variables, complex systems of conditional independence can be described by undirected graphs.

Then a set of variables $A$ is conditionally independent of set $B$, given the values of a set of variables $C$ if $C$ *separates* $A$ *from* $B$.

# Conditional independence

Random variables $X$ and $Y$ are *conditionally independent* given the random variable $Z$ if

$$\mathcal{L}(X \mid Y, Z) = \mathcal{L}(X \mid Z).$$

We then write $X \perp\!\!\!\perp Y \mid Z$

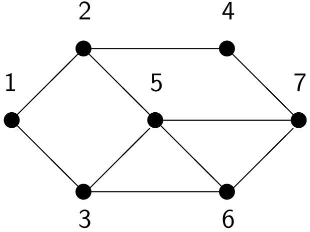
Intuitively:

Knowing $Z$ renders $Y$ *irrelevant* for predicting $X$.

Conditional independence can be expressed through Factorization of probabilities:

$$\begin{aligned} X \perp\!\!\!\perp Y \mid Z \quad &\iff \quad p_{xyz}p_{++z} = p_{x+z}p_{+yz} \\ &\iff \quad \exists a, b : p_{xyz} = a_{xz}b_{yz}. \end{aligned}$$

# Graphical models




For several variables, complex systems of conditional independence can be described by undirected graphs.

A set of variables $A$ is conditionally independent of set $B$, given the values of a set of variables $C$ if *$C$ separates $A$ from $B$*.

# Global Markov property and factorization

Formally we say for a given graph $\mathcal{G}$ that a distribution obeys *the global Markov property* (G) if

$$S \text{ separates } A \text{ from } B \text{ implies } A \perp\!\!\!\perp B \mid S.$$

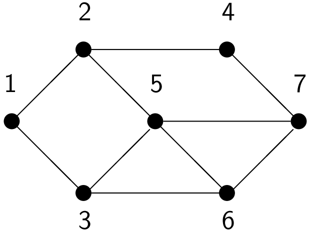
A distribution *factorizes* w.r.t. $\mathcal{G}$ if

$$p(x) = \prod_{a \text{ complete}} \psi_a(x)$$

where $\psi_a(x)$ depends on $x$ through $x_a = (x_v)_{v \in a}$ only.

It can be shown that *a positive probability distribution is globally Markov w.r.t. a graph if and only if it factorizes as above.*

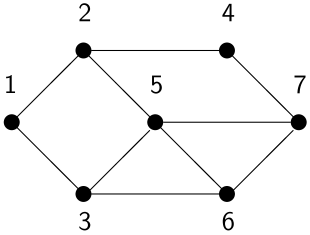
# Global Markov property




To find conditional independence relations, one should look for separating sets, such as $\{2,3\}$, $\{4,5,6\}$, or $\{2,5,6\}$

For example, it follows that $1 \perp\!\!\!\perp 7 \mid \{2,5,6\}$ and $2 \perp\!\!\!\perp 6 \mid \{3,4,5\}$.

# Factorization




A probability distribution factorizes w.r.t. this graph iff it can be written in the form

$$
\begin{aligned}
p(x) \quad &= \quad \psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)\psi_{24}(x_2, x_4)\psi_{25}(x_2, x_5) \\
&\quad\quad \times\psi_{47}(x_4, x_7)\psi_{356}(x_3, x_5, x_6)\psi_{567}(x_5, x_6, x_7)
\end{aligned}
$$

# Log–linear models

$\mathcal{A} = \{a_1, \ldots, a_K\}$ denotes a set of (pairwise incomparable) subsets of $a_i \subseteq V$.

A probability distribution $p$ (or function) *factorizes* w.r.t. $\mathcal{A}$ if it can be written as a product of terms where each only depend on variables in the same subset of $\mathcal{A}$, i.e. as

$$p(x) = \prod_{a \in \mathcal{A}} \psi_a(x)$$

where $\psi_a(x)$ depends on $x$ through $x_a = (x_v)_{v \in a}$ only.

The set of distributions which factorize w.r.t. $\mathcal{A}$ is the *log–linear model* generated by $\mathcal{A}$.

$\mathcal{A}$ is the *generating class* of the log–linear model.

*If the distribution factorizes without being everywhere positive, it will also satisfy all the Markov properties,* but not the other way around.

Formally, we define *the graphical model with graph* $G = (V, E)$ to be the log-linear model with $\mathcal{A} = \mathcal{C}$, where $\mathcal{C}$ are the *cliques* (i.e. maximal complete subsets) of the graph.

# Example

Consider a three way contingency table, where e.g. $m_{ijk}$ denotes the mean of the counts $N_{ijk}$ in the cell $(i, j, k)$ which has then been expanded as e.g.

$$\log m_{ijk} = \alpha_i + \beta_j + \gamma_k \tag{1}$$

or

$$\log m_{ijk} = \alpha_{ij} + \beta_{jk} \tag{2}$$

or

$$\log m_{ijk} = \alpha_{ij} + \beta_{jk} + \gamma_{ik}, \tag{3}$$

or (with redundancy)

$$\log m_{ijk} = \gamma + \delta_i + \phi_j + \eta_k + \alpha_{ij} + \beta_{jk} + \gamma_{ik},$$

The additive terms in the expansion are known as *interaction terms of order* $|a| - 1$ or $|a|$*-factor interactions*.

Interaction terms of $0$th order are called *main effects*.

# Dependence graph of log–linear model

For any generating class $\mathcal{A}$ we can construct the dependence graph of the corresponding log–linear model.

This is determined by the relation

$$\alpha \sim \beta \iff \exists a \in \mathcal{A} : \alpha, \beta \in a.$$

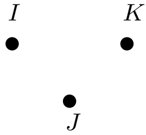
*Then any probability distribution which factorizes w.r.t. $\mathcal{A}$ also satisfies the global Markov property w.r.t. $G(\mathcal{A})$.*

This is by default the graph displayed in MIM.

# Independence

The log–linear model specified by (1) is known as the *main effects model.*

It has generating class consisting of singletons only $\mathcal{A} = \{\{I\}, \{J\}, \{K\}\}$. It has dependence graph

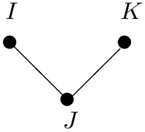



Thus it corresponds to *complete independence*.

# Conditional independence

The log–linear model specified by (2) has no interaction between $I$ and $K$.

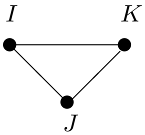
It has generating class $\mathcal{A} = \{\{I, J\}, \{J, K\}\}$ and dependence graph




Thus it corresponds to the *conditional independence* $I \perp\!\!\!\perp K \mid J$.

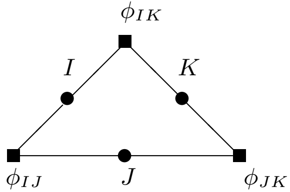
# No interaction of second order

The log–linear model specified by (3) has no second-order interaction. It has generating class $\mathcal{A} = \{\{I, J\}, \{J, K\}, \{I, K\}\}$ and its dependence graph




is the complete graph. Thus it has no conditional independence interpretation.

# Interaction graphs




The *interaction graph* of $\mathcal{A}$ is the graph with vertices $V \cup \mathcal{A}$ and edges define by

$$\alpha \sim a \iff \alpha \in a.$$

Using this graph all log–linear models admit a simple visual representation. Can be requested in MIM.

# Likelihood function

The likelihood function for an unknown $p$ can be expressed as

$$L(p) = \prod_{\nu=1}^{n} p(x^\nu) = \prod_{x \in \mathcal{X}} p(x)^{n(x)}.$$

In contingency table form the data follow a multinomial distribution

$$P\{N(x) = n(x), x \in \mathcal{X}\} = \frac{n!}{\prod_{x \in X} n(x)!} \prod_{x \in \mathcal{X}} p(x)^{n(x)}$$

but this only affects the likelihood function by a constant factor.

It can be shown that in log-linear models, *the likelihood function has at most one maximum*. When zero-values are allowed, it *always* has one.

MIM uses an algorithm for fitting known as *Iterative Proportional Fitting* which, if properly implemented, also works in the case where probabilities are allowed to be zero (sparse tables).

Also implemented e.g. in *R* in `loglin` with front end `loglm` in MASS.

An alternative is to "pretend" that counts are independent and Poisson distributed and use `glm`. However, the algorithm used there does *not* work when estimated cell probabilities are zero.