# Introduction to categorical data and conditional independence

**MSc Further Statistical Methods, Lecture 1**
**Hilary Term 2007**

Steffen Lauritzen, University of Oxford; January 17, 2007

## Categorical Data

Examples of categorical variables

- *Sex*: Male, Female;
- *Colour of Hair*: Blond, Red, Neutral, Dark;
- *Degree of Satisfaction with work*: Low, Medium, High
- *Yearly income*: Below 10,000, 10,001-20,000, 20,001-40,000, above 40,000;

Some are *nominal*, others *ordinal*. They have different number of *states.*

## Contingency Table

Data often presented in the form of a *contingency table* or *cross-classification*:

| | Sex | |
|---|---|---|
| Admitted | Male | Female |
| Yes | 1198 | 557 |
| No | 1493 | 1278 |

This is a *two-way table* (or two-way classification) with categorical variables $A$: Admitted? and $S$: Sex. In this case it is a *$2 \times 2$-table*.

The numerical entries are *cell counts* $n_{ij}$, the number of cases in the category $A = i$ and $S = j$. The *total number of cases* is $n = \sum_{ij} n_{ij}$.

## Data in list form

Data can also appear in the form of a *list of cases:*

| case | Admitted | Sex |
|---|---|---|
| 1 | Yes | Male |
| 2 | Yes | Female |
| 3 | No | Male |
| 4 | Yes | Male |
| ⋮ | ⋮ | ⋮ |

The contingency table is then formed from the list of cases by counting the number of cases in each cell of the table.

## Multinomial sampling model

The standard sampling model for data of this form specifies that cases are independent and $p_{ij} = P(A = i, S = j)$ is the probability that a given case belongs to cell $ij$.

The cell counts then follow a *multinomial distribution*

$$P(N_{ij} = n_{ij}, i = 1, \ldots I, j = 1, \ldots J) = \frac{n!}{\prod_{ij} n_{ij}!} \prod_{ij} p_{ij}^{n_{ij}}.$$

The *expected cell counts* are

$$m_{ij} = \mathbf{E}(N_{ij}) = np_{ij}.$$

Other sampling schemes *fixes certain marginal totals* or have a *Poisson total $N$*, leading to cell counts being independent Poisson.

## Hypothesis of independence

A typical hypothesis of interest is that of *independence* between the two variables, i.e. that

$$p_{ij} = P(A = i, S = j) = P(A = i)P(S = j) = p_{i+}p_{+_j},$$

where

$$p_{i+} = P(A = i) = \sum_j p_{ij}, \quad p_{+j} = P(S = j) = \sum_i p_{ij}$$

are the *marginal probabilities*.

## Likelihood ratio test

Without assuming independence, the MLE of the cell probabilities and expected cell counts are

$$\hat{p}_{ij} = n_{ij}/n, \quad \hat{m}_{ij} = n\hat{p}_{ij} = n_{ij}.$$

Similarly, assuming independence, the MLE becomes

$$\hat{\hat{p}}_{ij} = n_{i+}n_{+j}/n^2, \quad \hat{\hat{m}}_{ij} = n\hat{\hat{p}}_{ij} = n_{i+}n_{+j}/n,$$

where

$$n_{i+} = \sum_j n_{ij}, \quad n_{+j} = \sum_i n_{ij}$$

are the *marginal counts*. Hence we get

$$\begin{aligned} G^2 &= -2\log\Lambda = -2\log\frac{L(\hat{\hat{p}})}{L(\hat{p})} \\ &= 2\sum_{ij} n_{ij}\log\frac{\hat{p}_{ij}}{\hat{\hat{p}}_{ij}} = 2\sum_{ij} n_{ij}\log\frac{\hat{m}_{ij}}{\hat{\hat{m}}_{ij}} \\ &= 2\sum_{ij} n_{ij}\log\frac{n_{ij}}{\hat{\hat{m}}_{ij}} = 2\sum \mathsf{OBS}\log\frac{\mathsf{OBS}}{\mathsf{EXP}}, \end{aligned}$$

Here OBS refers to the *observed cell counts* and EXP to the *expected cell counts* under the hypothesis.

It can be shown that for large cell counts, *$G^2$ is approximately $\chi^2$-distributed with degrees of freedom equal to $(I-1)(J-1)$* which is equal to 1 in this case.

## Pearson's $\chi^2$ statistic

An alternative to the LRT statistic or *deviance* $G^2$, one can use the statistic

$$\chi^2 = \sum \frac{(\text{OBS} - \text{EXP})^2}{\text{EXP}},$$

which is an approximation to the deviance and also has approximately the same distribution, under the null hypothesis, for large cell counts.

For the approximations to be valid, it is a *common rule of thumb for both $G^2$ and $\chi^2$ that the expected cell counts $\hat{\hat{m}}_{ij}$ must be larger than 5.*

This condition is often *not* satisfied, in particular in multi-way tables with many variables.

## Sparse tables

Data on oral lesions by region in India:

| | Kerala | Gujarat | Andhra |
|---|---|---|---|
| Labial Mucosa | 0 | 1 | 0 |
| Buccal Mucosa | 8 | 1 | 8 |
| Commisure | 0 | 1 | 0 |
| Gingiva | 0 | 0 | 1 |
| Hard Palate | 0 | 1 | 0 |
| Soft palate | 0 | 1 | 0 |
| Tongue | 0 | 1 | 1 |
| Floor of Mouth | 1 | 0 | 1 |
| Alveolar Ridge | 1 | 0 | 1 |

## Exact testing methods

In sparse tables such as the data on oral lesions, asymptotic results can be very misleading.

Instead one can exploit that, under the hypothesis of independence, *the marginals are sufficient* and the conditional distribution of the counts $\{N_{ij}\}$ is:

$$P\{(n_{ij}) \,|\, (n_{i+}), (n_{+j})\} = \frac{\prod_{i=1}^{I} n_{i+}! \prod_{j=1}^{J} n_{+j}!}{n! \prod_{i=1}^{I} \prod_{j=1}^{J} n_{ij}!}. \quad (1)$$

*Fisher's exact test* rejects for small values of the *observed value* of $P\{(n_{ij}) \,|\, (n_{i+}), (n_{+j})\}$ and evaluates the $p$-value in this distribution as well.

## Monte-Carlo testing

In principle, exact testing requires enumeration of all possible tables with a given margin.

However, there is an *efficient algorithm* due to Patefield (1981) which generates samples $\{\tilde{n}_{ij}\}_k, k = 1, \dots K$ from the distribution (1).

By choosing $K$ large, the correct $p$-value *for any test statistic $T$* can be calculated to any degree of accuracy as

$$\tilde{p} = \frac{|\{k : \tilde{t}_k \geq t_{\text{obs}}\}|}{K},$$

where $\tilde{t}_k$ is calculated from the table $\{\tilde{n}_{ij}\}_k$.

This may well be preferable to using asymptotic results.

## Three-way tables

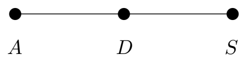
Admissions to Berkeley by department

| 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 $A \perp\!\!\!\perp S \,|\, D$ 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

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

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 and Log-Linear Models

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

Steffen Lauritzen, University of Oxford; January 18, 2007

## 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...

## 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.

## Florida murderers

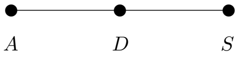
Sentences in 4863 murder cases in Florida over the six years 1973-78

| | 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.

## 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

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

## 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.

## Marginal and conditional independence

Note that there the two conditions

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

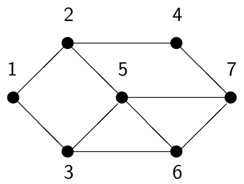
are very different and will typically not both hold unless we either have $A \perp\!\!\!\perp (D, S)$ or $(A, D) \perp\!\!\!\perp S$, i.e. if one of the variables are completely independent of both of the others.

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.

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

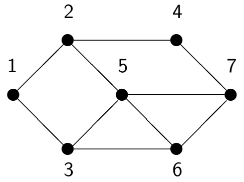
Intuitively:

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

Conditional independence can be expressed through Factorization of probabilities:

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

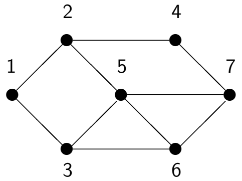
## Factorization



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

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

## 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 $B$*.

## 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.

## 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

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

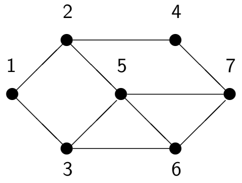
where $\psi_a(x)$ depends on $x$ through $x_a = (x_v)_{v \in a}$ only.

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.*

*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.

## 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 $2 \perp\!\!\!\perp 6 \mid \{3, 4, 5\}$.

## 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 0th 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

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

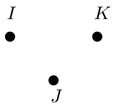
*Then any probability distribution which factorizes w.r.t. $\mathcal{A}$ also satisfies the global Markov property w.r.t. $G(\mathcal{A})$.*

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

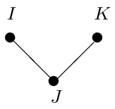


Thus it corresponds to *complete independence*.

## Conditional independence

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

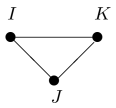
It has generating class $\mathcal{A} = \{\{I, J\}, \{J, K\}\}$ and dependence graph



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

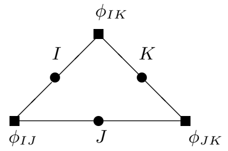
## No interaction of second order

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.

# Measures of association. Ordinal variables. Symmetric Tables

**MSc Further Statistical Methods, Lecture 3**
**Hilary Term 2007**

Steffen Lauritzen, University of Oxford; January 24, 2007

## Measures of association

If (conditional) independence among a pair of variables does not hold, it becomes of interest to *quantify* and *describe* the dependence.

When variables are nominal, there is no direct analogue of covariance or correlation and one must use other measures of association.

We consider the *relative risk* and the *odds-ratio*.

For ordinal variables there are analogues of the correlation coefficient. We shall consider *Kruskal's $\gamma$-coefficient*.

## Relative risk

Consider $2 \times 2$-table with probabilities

|  | B |  |
|---|---|---|
| A | 1 | 2 |
| 1 | $p_{11}$ | $p_{12}$ |
| 2 | $p_{21}$ | $p_{22}$ |

The *relative risk* ($\rho = RR$) compares $P(A=1 \,|\, B=1) = p_{1\,|\,1} = p_{11}/(p_{11}+p_{21})$ with $P(A=1 \,|\, B=2) = p_{1\,|\,2} = p_{12}/(p_{12}+p_{22})$:

$$\rho = \frac{p_{11}}{p_{12}} \frac{p_{12}+p_{22}}{p_{11}+p_{21}}.$$

## Example

The empirical counterpart of the relative risk is

$$\hat{\rho} = \frac{n_{11}}{n_{12}} \frac{n_{12}+n_{22}}{n_{11}+n_{21}}$$

|  | Sex |  |
|---|---|---|
| Admitted | Male | Female |
| Yes | 1198 | 557 |
| No | 1493 | 1278 |

Here

$$\hat{\rho} = \frac{1198}{557} \frac{557+1278}{1198+1493} = 1.47$$

so it appears that chances for a male to be admitted is about 47% higher than those for females.

## Odds–ratio

The relative risk is an asymmetric measure of association between $A$ and $B$. This may sometimes be inconvenient, so an alternative is the *odds-ratio* $\theta$.

The (conditional) *odds* for $A=1$ given $B=1$ are

$$\omega(A=1 \,|\, B=1) = \omega_{11} = \frac{P(A=1 \,|\, B=1)}{P(A=2 \,|\, B=1)} = \frac{p_{11}}{p_{21}}$$

and similarly for $B=2$. The odds-ratio is thus

$$\theta = \frac{\omega_{11}}{\omega_{12}} = \frac{(p_{11}/p_{21})}{p_{12}/p_{22}} = \frac{p_{11}p_{22}}{p_{12}p_{21}},$$

which is fully symmetric in $A$ and $B$ and in the labels $1$ and $2$. Thus it does not change if we relabel the variables or its states.

The odds-ratio is also known as the *cross-product ratio* and its empirical counterpart is

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}},$$

which for the admission example gives

$$\hat{\theta} = \frac{1198 \times 1278}{557 \times 1493} = 1.84.$$

One can easily show that

$$A \perp\!\!\!\perp B \iff \theta = 1$$

and a value of $\theta$ greater than one corresponds to positive association (as in the admission example) whereas $\theta < 1$ corresponds to negative association.

## Conditional odds-ratios

More generally, if $A$ and $B$ have more than two states, the odds-ratio is defined for two pairs of states $(i, i^*)$ and $(j, j^*)$ as

$$\theta_{ii^*jj^*} = \frac{p_{ij}p_{i^*j^*}}{p_{ij^*}p_{i^*j}}$$

and *$A \perp\!\!\!\perp B$ if and only if all such ratios are equal to one*.

Conditioning on the values of a third variable $C = k$ we similarly have conditional independence $A \perp\!\!\!\perp B \,|\, C$ if and only if

$$\theta_{ii^*jj^*\,|\,k} = \frac{p_{ijk}p_{i^*j^*k}}{p_{ij^*k}p_{i^*jk}} = 1$$

for all combinations of the indices.

## No second-order interaction

*If the distribution satisfies the restriction of a log-linear model with no second-order interaction,* i.e. if

$$p_{ijk} = a_{ij}b_{jk}c_{ik}$$

then

$$\theta_{ii^*jj^*\,|\,k} = \frac{a_{ij}b_{jk}c_{ik}a_{i^*j^*}b_{j^*k}c_{i^*k}}{a_{ij^*}b_{j^*k}c_{ik}a_{i^*j}b_{jk}c_{i^*k}} = \frac{a_{ij}a_{ij^*}}{a_{ij^*}a_{i^*j}}$$

so *the conditional odds-ratio is constant in $k$.*

*This does not imply absence of a Simpson paradox!* For the marginal distribution of $I, J$ is

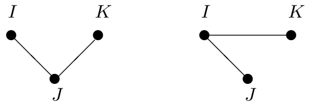
$$p_{ij+} = a_{ij} \sum_k b_{jk}c_{ik} = a_{ij}\tilde{b}_{ij}.$$

For the $IJ$ odds-ratio to be the same in the marginal table as in the condition it must additionally hold that $\tilde{b}$ satisfies

$$\tilde{b}_{ij} = \alpha_i \beta_j.$$

This holds if either $I \perp\!\!\!\perp K \mid J$ or $J \perp\!\!\!\perp K \mid I$.

Thus, *a Simpson paradox concerning association between $I$ and $J$ is avoided if one of the following graphical models hold,* and typically not otherwise.



## Example

| | Overall | Department | | | | | |
|---|---|---|---|---|---|---|---|
| | | I | II | III | IV | V | VI |
| odds-ratio | 1.84 | 0.35 | 0.8 | 1.13 | 0.92 | 1.22 | 0.83 |

The empirical odds-ratios for the admission data indicate a strong example of Simpson's paradox.

For department I, Sex and admission is strongly negatively associated. For other departments the association is moderate and of changing sign.

But overall, the association is strong and positive!

## Two ordinal variables

| | Job satisfaction | | | |
|---|---|---|---|---|
| Income | Very diss. | Little diss. | Mod. sat. | Very sat. |
| $< 15,000$ | 1 | 3 | 10 | 6 |
| $15,000$–$25,000$ | 2 | 3 | 10 | 7 |
| $25,000$–$40,000$ | 1 | 6 | 14 | 12 |
| $> 40,000$ | 0 | 1 | 9 | 11 |

For ordinal variables we consider concordant and discordant pairs: A pair $(i_1, j_1), (i_2, j_2)$ is *concordant*

$$i_1 < i_2 \text{ and } j_1 < j_2$$

it is *discordant* if it is the other way around

$$i_1 < i_2 \text{ and } j_1 > j_2,$$

and otherwise it is *tied*.

## Kruskal's gamma

Kruskal's $\gamma$-coefficient is defined as

$$\gamma = \frac{p_c - p_d}{p_c + p_d},$$

where $p_c$ and $p_d$ are the probability that a random pair of individuals is a concordant or discordant pair.

Clearly, $-1 \leq \gamma \leq 1$ and $\gamma = 0$ for independent variables, so $\gamma$ is an analogue of the correlation.

As for the correlation, *the variables can be dependent and still have $\gamma = 0$.*

Also *$\gamma = 1$ if and only $p_{ij} = 0$ for $j < i$* and similarly for $\gamma = -1$.

The empirical analogue of Kruskal's $\gamma$ is

$$\hat{\gamma} = \frac{n_c - n_d}{n_c + n_d} = \frac{1331 - 841}{1331 + 841} = 0.221$$

in the example. So there is a mild (but significant) positive relation between income and job satisfaction.

A test using $|\hat{\gamma}|$ as test statistic can be made using Monte-Carlo $p$-values (not implemented in MIM).

MIM features a variety of alternative test statistic for exploiting ordinality.

These include the Wilcoxon statistic, the Kruskal–Wallis statistic and the Jonckheere–Terpstra statistic. See Edwards (2002), Chapter 5 for detailed description of these.

## Wilcoxon test

| | | | Response | | |
|---|---|---|---|---|---|
| Centre | Status | Treatment | Poor | Moderate | Excellent |
| 1 | 1 | Active | 3 | 20 | 5 |
| | | Placebo | 11 | 14 | 8 |
| | 2 | Active | 3 | 14 | 12 |
| | | Placebo | 6 | 13 | 5 |
| 2 | 1 | Active | 12 | 12 | 0 |
| | | Placebo | 11 | 10 | 0 |
| | 2 | active | 3 | 9 | 4 |
| | | Placebo | 6 | 9 | 3 |

Multicentre analgesic trial. Here are four variables $C$: Centre, $S$: Status, $T$: Treatment, and $R$: Response.

*Wilcoxon test-statistic* compares distribution of *ranks* between two distributions. Ranks are well-defined for ordinal data.

## Several categories

| | Response | | |
|---|---|---|---|
| Drug regimen | None | Partial | Complete |
| 1 | 2 | 0 | 0 |
| 2 | 1 | 1 | 0 |
| 3 | 3 | 0 | 0 |
| 4 | 2 | 2 | 0 |
| 5 | 1 | 1 | 4 |

Two variables $D$: Drug regimen, $R$: response. The *Kruskal-Wallis* test statistic measure deviations from independence in direction of at *least one distribution stochastically larger* than the others.

Kruskal-Wallis test specializes to Wilcoxon test for binary variables.

## Two ordinal variables

| | Job satisfaction | | | |
|---|---|---|---|---|
| Income | Very diss. | Little diss. | Mod. sat. | Very sat. |
| $< 15,000$ | 1 | 3 | 10 | 6 |
| $15,000$–$25,000$ | 2 | 3 | 10 | 7 |
| $25,000$–$40,000$ | 1 | 6 | 14 | 12 |
| $> 40,000$ | 0 | 1 | 9 | 11 |

Two ordinal variables: $J$: Job satisfaction, $I$: Income. *Jonckheere-Terpstra* test measures deviations from independence in direction of *all distributions being stochastically ordered.*

The Jonckheere–Terpstra test specializes to the Wilcoxon test if one of the two ordinal variables are binary.

## Square tables

In some cases, the variables $A$ and $B$ represent 'the same thing' and quite different hypotheses become relevant, for example that of *marginal homogeneity*

$$p_{i+} = p_{+i}.$$

|  | After |  |  |
|---|---|---|---|
| Before | Approve | Disapprove | Total |
| Approve | 794 | 150 | 944 |
| Disapprove | 86 | 570 | 656 |
| Total | 880 | 720 | 1600 |

Attitude towards UK prime minister. Opinion poll data from Agresti, Ch. 10.

A panel of 1600 persons were asked at two points in time whether they approved of the policy of the current PM. The interesting question is whether the opinion has changed. If it has not, we say there is *marginal homogeneity*

$$p_{i+} = p_{+i}, \text{ for all } i. \tag{1}$$

In $2 \times 2$ case this is equivalent to having $\delta = 0$ where

$$\begin{aligned} \delta &= p_{1+} - p_{+1} \\ &= (p_{11} + p_{12}) - (p_{11} + p_{21}) = p_{12} - p_{21} \end{aligned}$$

so

$$p_{1+} = p_{+2} \iff p_{12} = p_{21},$$

i.e. *marginal homogeneity is equivalent to symmetry,* where the hypothesis of symmetry is given as

$$p_{ij} = p_{ji}. \tag{2}$$

The empirical counterpart of $\delta$ is

$$\hat{\delta} = \frac{n_{12} - n_{21}}{n}.$$

Under the assumption of homogeneity, the variance of $\hat{\delta}$ can be calculated as

$$\mathbf{V}(n\hat{\delta}) = 2np_{12} = 2np_{21} = 2np.$$

Under the hypothesis

$$\hat{p} = \frac{n_{12} + n_{21}}{2n},$$

so

$$\chi^2 = \frac{n\hat{\delta}^2}{2n\hat{p}} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

is for large $n$ approximately $\chi^2$ distributed with 1 degree of freedom.

In the example, we get

$$\chi^2 = \frac{(86 - 150)^2}{86 + 150} = 17.4$$

which is highly significant.

## More than two states

The test for symmetry of $A$ and $B$ as expressed in (2) generalizes immediately to several states as

$$\chi^2 = \sum_i \sum_{j>i} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}}$$

which is approximately $\chi^2$ distributed with $I(I-1)/2$ degrees of freedom.

Clearly, *marginal symmetry implies marginal homogeneity*.

However, *the converse is false in the multi-state case*.

Testing for marginal homogeneity is more complicated then, see Agresti, Ch. 10.

# Missing Data and the EM algorithm

**MSc Further Statistical Methods**
**Lecture 4 and 5**
**Hilary Term 2007**

Steffen Lauritzen, University of Oxford; January 31, 2007

## Missing data problems

| case | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| $1$ | $a_1$ | $b_1$ | $*$ | $d_1$ | $e_1$ | * |
| $2$ | $a_2$ | $*$ | $c_2$ | $d_2$ | $e_2$ | $*$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $n$ | $a_n$ | $b_n$ | $c_n$ | $*$ | $*$ | $*$ |

$*$ or *NA* denotes values that are *missing*, i.e. non-observed.

## Examples of missingness

- non-reply in surveys;
- non-reply for specific questions: "missing" $\sim$ don't know, essentially an additional state for the variable in question
- recording error
- variable out of range
- just not recorded (e.g. too expensive)

Different types of missingness demand different treatment.

## Notation for missingness

Data matrix $Y$, *missing data matrix* $M = \{M_{ij}\}$:

$$M_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is missing} \\ 0 & \text{if } Y_{ij} \text{ is observed.} \end{cases}$$

Convenient to introduce the notation $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, where $Y_{\text{mis}}$ are conceptual and denote the data that were not observed.

This notation follows Little and Rubin (2002).

## Patterns of missingness

Little and Rubin (2002) classify these into the following *techincal* categories.

We shall illustrate with a case of cross-classification of Sex, Race, Admission and Department, $S, R, A, D$.

*Univariate:* $M_{ij} = 0$ unless $j = j^*$, e.g. an unmeasured response. Example: $R$ unobserved for some, but data otherwise complete.

*Multivariate:* $M_{ij} = 0$ unless $j \in J \subset V$, as above, just with multivariate response, e.g. in surveys. Example: For some subjects, both $R$ and $S$ unobserved.

*Monotone:* There is an ordering of $V$ so $M_{ik} = 0$ implies $M_{ij} = 0$ for $j < k$, e.g. drop-out in longitudinal studies. Example: For some, $A$ is unobserved, others neither $A$ nor $R$, but data otherwise complete.

*Disjoint:* Two subsets of variables never observed together. Controversial. Appears in Rubin's causal model. Example: $S$ and $R$ never both observed.

*General:* none of the above. Haphazardly scattered missing values. Example: $R$ unobserved for some, $A$ unobserved for others, $S, D$ for some.

*Latent:* A certain variable is never observed. Maybe it is even unobservable. Example: $S$ never observed, but believed to be important for explaining the data.

## Methods for dealing with missing data

*Complete case analysis:* analyse only cases where all variables are observed. Can be adequate if most cases are present, but will generally give serious biases in the analysis. In survey's, for example, this corresponds to making inference about the population of responders, not the full population;

*Weighting methods.* For example, if a population total $\mu = \mathbf{E}(Y)$ should be estimated and unit $i$ has been selected with probability $\pi_i$ a standard method is the *Horwitz–Thompson estimator*

$$\hat{\mu} = \frac{\sum \frac{Y_i}{\pi_i}}{\sum \frac{1}{\pi_i}}.$$

To correct for non-response, one could let $\rho_i$ be the response-probability, estimate this in some way as $\hat{\rho}_i$ and then let

$$\tilde{\mu} = \frac{\sum \frac{Y_i}{\pi_i \hat{\rho}_i}}{\sum \frac{1}{\pi_i \hat{\rho}_i}}.$$

*Imputation methods:* Find ways of estimating the values of the unobserved values as $\hat{Y}_{\text{mis}}$, then proceed as if there were complete data. Without care, this can give misleading results, in particular because the "sample size" can be grossly overestimated.

*Model-based likelihood methods: Model the missing data mechanism and then proceed to make a proper likelihood-based analysis,* either via the method of maximum-likelihood or using Bayesian methods. This appears to be the most sensible way.

Typically this approach was not computationally feasible in the past, but modern algorithms and computers have changed things completely. Ironically, the efficient algorithms are indeed based upon imputation of missing values, but with proper corrections resulting.

## Mechanisms of missingness

The data are *missing completely at random*, MCAR, if

$$f(M \mid Y, \theta) = f(M \mid \theta), \text{ i.e. } M \perp\!\!\!\perp Y \mid \theta.$$

Heuristically, the values of $Y$ have themselves no influence on the missingness. Example is recording error, latent variables, and variables that are missing *by design* (e.g. measuring certain values only for the first $m$ out of $n$ cases). Beware: it may be counterintuitive that *missing by design is MCAR*.

The data are *missing at random*, MAR, if

$$f(M \mid Y, \theta) = f(M \mid Y_{\text{obs}}, \theta), \text{ i.e. } M \perp\!\!\!\perp Y_{\text{mis}} \mid (Y_{\text{obs}}, \theta).$$

Heuristically, only the observed values of $Y$ have influence on the missingness. By design, e.g. if individuals with certain characteristics of $Y_{\text{obs}}$ are not included in part of study (where $Y_{\text{mis}}$ is measured).

The data are *not missing at random*, NMAR, in all other cases.

For example, if certain values of $Y$ cannot be recorded when they are out of range, e.g. in survival analysis.

The classifications above of the mechanism of missingness lead again to increasingly complex analyses.

It is not clear than the notion MCAR is helpful, but MAR is. Note that *if data are MCAR, they are also MAR.*

## Likelihood-based methods

The most convincing treatment of missing data problems seems to be via modelling the missing data mechanism, i.e. *by considering the missing data matrix $M$ as an explicit part of the data.*

The likelihood function then takes the form

$$\begin{aligned} L(\theta \mid M, y_{\text{obs}}) &\propto \int f(M, y_{\text{obs}}, y_{\text{mis}} \mid \theta)\, dy_{\text{mis}} \\ &= \int C_{\text{mis}}(\theta \mid M, y_{\text{obs}}, y_{\text{mis}}) f(y_{\text{obs}}, y_{\text{mis}} \mid \theta)\, dy_{\text{mis}}, \quad (1) \end{aligned}$$

where the factor $C_{\text{mis}}(\theta \mid M, y) = f(M \mid y_{\text{obs}}, y_{\text{mis}}, \theta)$ is based on an explicit model for the missing data mechanism.

## Ignoring the missing data mechanism

The likelihood function *ignoring the missing data mechanism* is

$$L_{\text{ign}}(\theta \mid y_{\text{obs}}) \propto f(y_{\text{obs}} \mid \theta) = \int f(y_{\text{obs}}, y_{\text{mis}} \mid \theta)\, dy_{\text{mis}}. \quad (2)$$

When is $L \propto L_{\text{ign}}$ so the missing data mechanism can be ignored for further analysis? *This is true if:*

1. The data are *MAR*;
2. The parameters $\eta$ governing the missingness are *separate* from parameters of interest $\psi$ i.e. the parameters vary in a product region, so that information about the value of one does not restrict the other.

## Ignorable missingness

If data are MAR and the missingness parameter is separate from the parameter of interest, we have $\theta = (\eta, \psi)$ and

$$C_{\text{mis}}(\theta) = f(M \mid y_{\text{obs}}, y_{\text{mis}}, \eta) = f(M \mid y_{\text{obs}}, \eta)$$

Hence, the correction factor $C_{\text{mis}}$ is constant (1) and can be taken outside in the integral so that

$$L(\theta \mid M, y_{\text{obs}}) \propto C_{\text{mis}}(\eta) L_{\text{ign}}(\theta \mid y_{\text{obs}})$$

and since

$$f(y_{\text{obs}}, y_{\text{mis}} \mid \theta) = f(y_{\text{obs}}, y_{\text{mis}} \mid \psi)$$

we get

$$L(\theta \mid M, y_{\text{obs}}) \propto C_{\text{mis}}(\eta) L_{\text{ign}}(\psi \mid y_{\text{obs}}),$$

which shows that the missingness mechanism can be ignored when concerned with likelihood inference about $\psi$.

For a Bayesian analysis the parameters must in addition be *independent w.r.t. the prior:*

$$f(\eta, \psi) = f(\eta) f(\psi).$$

If the data are *NMAR* or the parameters are *not separate,* then *the missing data mechanism cannot be ignored.*

Care must then be taken to model the mechanism $f(M \mid y_{\text{obs}}, y_{\text{mis}}, \theta)$ and the corresponding likelihood term must be properly included in the analysis.

Note: *$Y_{\text{mis}}$ is MAR if data is $(M, Y)$,* i.e. if $M$ is considered part of the data, since then $M \perp\!\!\!\perp Y_{\text{mis}} \mid (M, Y_{\text{obs}}, \theta)$.

## The EM algorithm

The EM algorithm is an alternative to Newton–Raphson or the method of scoring for computing MLE in cases where the complications in calculating the MLE are due to *incomplete observation* and data are *MAR*, missing at random, with *separate parameters* for observation and the missing data mechanism, so the missing data mechanism can be ignored.

Data $(X, Y)$ are the *complete data* whereas only *incomplete data* $Y = y$ are observed. (Rubin uses $Y = Y_{\text{obs}}$ and $X = Y_{\text{mis}}$).

The *complete data log-likelihood* is:

$$l(\theta) = \log L(\theta; x, y) = \log f(x, y; \theta).$$

The *marginal log-likelihood* or *incomplete data log-likelihood* is based on $y$ alone and is equal to

$$l_y(\theta) = \log L(\theta; y) = \log f(y;\theta).$$

We wish to maximize $l_y$ in $\theta$ but $l_y$ is typically quite unpleasant:

$$l_y(\theta) = \log \int f(x, y; \theta)\, dx.$$

The EM algorithm is a method of maximizing the latter iteratively and alternates between two steps, one known as the *E-step* and one as the *M-step*, to be detailed below.

We let $\theta^*$ be and arbitrary but fixed value, typically the value of $\theta$ at the current iteration.

The *E-step calculates the expected complete data log-likelihood ratio* $q(\theta \mid \theta^*)$:

$$\begin{aligned} q(\theta \mid \theta^*) &= \mathbf{E}_{\theta^*}\left[\log \frac{f(X, y; \theta)}{f(X, y; \theta^*)} \mid Y = y\right] \\ &= \int \log \frac{f(x, y; \theta)}{f(x, y; \theta^*)} f(x \mid y; \theta^*)\, dx. \end{aligned}$$

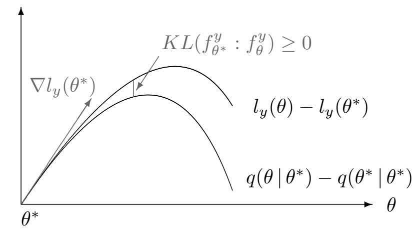
The *M-step maximizes $q(\theta \mid \theta^*)$ in $\theta$ for for fixed $\theta^*$*, i.e. calculates

$$\theta^{**} = \arg\max_\theta q(\theta \mid \theta^*).$$

*After an E-step and subsequent M-step, the likelihood function has never decreased.*

The picture on the next overhead should show it all.

## Expected and complete data likelihood


$KL(f^y_{\theta^*} : f^y_\theta) \geq 0$
$\nabla l_y(\theta^*)$
$l_y(\theta) - l_y(\theta^*)$
$q(\theta \mid \theta^*) - q(\theta^* \mid \theta^*)$
$\theta^*$
$\theta$

$$l_y(\theta) - l_y(\theta^*) = q(\theta \mid \theta^*) + KL(f^y_{\theta^*} : f^y_\theta)$$

$$\nabla l_y(\theta^*) = \frac{\partial}{\partial\theta} l_y(\theta)\Big|_{\theta=\theta^*} = \frac{\partial}{\partial\theta} q(\theta \mid \theta^*)\Big|_{\theta=\theta^*}.$$

## Kullback-Leibler divergence

The *KL divergence* between $f$ and $g$ is

$$KL(f : g) = \int f(x) \log \frac{f(x)}{g(x)}\, dx.$$

Also known as *relative entropy* of $g$ with respect to $f$.

Since $-\log x$ is a convex function, Jensen's inequality gives

*$KL(f : g) \geq 0$ and $KL(f : g) = 0$ if and only if $f = g$*, since

$$KL(f : g) = \int f(x) \log \frac{f(x)}{g(x)}\, dx \geq -\log \int f(x) \frac{g(x)}{f(x)}\, dx = 0,$$

so KL divergence defines an (asymmetric) distance measure between probability distributions.

## Expected and marginal log-likelihood

Since $f(x \mid y; \theta) = f\{(x, y); \theta\}/f(y; \theta)$ we have

$$\begin{aligned} q(\theta \mid \theta^*) &= \int \log \frac{f(y;\theta) f(x \mid y; \theta)}{f(y;\theta^*) f(x \mid y; \theta^*)} f(x \mid y; \theta^*)\, dx \\ &= \log f(y;\theta) - \log f(y;\theta^*) \\ &\quad + \int \log \frac{f(x \mid y; \theta)}{f(x \mid y; \theta^*)} f(x \mid y; \theta^*)\, dx \\ &= l_y(\theta) - l_y(\theta^*) - KL(f^y_{\theta^*} : f^y_\theta). \end{aligned}$$

Since the KL-divergence is minimized for $\theta = \theta^*$, differentiation of the above expression yields

$$\frac{\partial}{\partial\theta} q(\theta \mid \theta^*)\Big|_{\theta=\theta^*} = \frac{\partial}{\partial\theta} l_y(\theta)\Big|_{\theta=\theta^*}.$$

Let now $\theta_0 = \theta^*$ and define the iteration

$$\theta_{n+1} = \arg\max_\theta q(\theta \mid \theta_n).$$

Then

$$\begin{aligned} l_y(\theta_{n+1}) &= l_y(\theta_n) + q(\theta_{n+1} \mid \theta_n) + KL(f^y_{\theta_{n+1}} : f^y_{\theta_n}) \\ &\geq l_y(\theta_n) + 0 + 0. \end{aligned}$$

So the log-likelihood never decreases after a combined E-step and M-step.

*It follows that any limit point must be a saddle point or a local maximum of the likelihood function.*

## Mixtures

Consider a sample $Y = (Y_1, \ldots, Y_n)$ from individual densities

$$f(y; \alpha, \mu) = \{\alpha\phi(y - \mu) + (1 - \alpha)\phi(y)\}$$

where $\phi$ is the normal density

$$\phi(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

and $\alpha$ and $\mu$ are both unknown, $0 < \alpha < 1$.

This corresponds to a fraction $\alpha$ of the observations being contaminated, or originating from a different population.

## Incomplete observation

The likelihood function becomes

$$L_y(\alpha, \mu) = \prod_i \{\alpha\phi(y_i - \mu) + (1 - \alpha)\phi(y_i)\}$$

is quite unpleasant, although both Newton–Raphson and the method of scoring can be used.

*But suppose we knew which observations came from which population?*

In other words, let $X = (X_1, \ldots, X_n)$ be i.i.d. with $P(X_i = 1) = \alpha$ and suppose that the conditional distribution of $Y_i$ given $X_i = 1$ was $\mathcal{N}(\mu, 1)$ whereas given $X_i = 0$ it was $\mathcal{N}(0, 1)$, i.e. that $X_i$ was indicating whether $Y_i$ was contaminated or not.

Then the marginal distribution of $Y$ is precisely the mixture distribution and the 'complete data likelihood' is

$$\begin{aligned} L_{x,y}(\alpha,\mu) &= \prod_i \alpha^{x_i}\phi(y_i-\mu)^{x_i}(1-\alpha)^{1-x_i}\phi(y_i)^{1-x_i} \\ &\propto \alpha^{\sum x_i}(1-\alpha)^{n-\sum x_i}\prod_i \phi(y_i-\mu)^{x_i} \end{aligned}$$

so taking logarithms we get (ignoring a constant) that

$$\begin{aligned} l_{x,y}(\alpha,\mu) &= \sum x_i \log\alpha + \left(n-\sum x_i\right)\log(1-\alpha) \\ &\quad -\sum_i x_i(y_i-\mu)^2/2. \end{aligned}$$

If we did not know how to maximize this explicitly, differentiation easily leads to:

$$\hat{\alpha} = \sum x_i/n, \quad \hat{\mu} = \sum x_i y_i / \sum x_i.$$

Thus, when complete data are available the frequency of contaminated observations is estimated by the observed frequency and the mean $\mu$ of these is estimated by the average among the contaminated observations.

## E-step and M-step

By taking expectations, we get the E-step as

$$\begin{aligned} q(\alpha,\mu\,|\,\alpha^*,\mu^*) &= \mathbf{E}_{\alpha^*,\mu^*}\{l_{X,y}(\alpha,\mu)\,|\,Y=y\} \\ &= \sum x_i^* \log\alpha + \left(n-\sum x_i^*\right)\log(1-\alpha) \\ &\quad -\sum_i x_i^*(y_i-\mu)^2/2 \end{aligned}$$

where

$$x_i^* = \mathbf{E}_{\alpha^*,\mu^*}(X_i\,|\,Y_i=y_i) = P_{\alpha^*,\mu^*}(X_i=1\,|\,Y_i=y_i).$$

Since this has the same form as the complete data likelihood, just with $x_i^*$ replacing $x_i$, the M-step simply becomes

$$\alpha^{**} = \sum x_i^*/n, \quad \mu^{**} = \sum x_i^* y_i / \sum x_i^*,$$

i.e. here the mean of the contaminated observations is estimated by a weighted average of all the observations, the weight being proportional to the probability that this observation is contaminated. In effect, $x_i^*$ act as *imputed values* of $x_i$.

The imputed values $x_i^*$ needed in the E-step are calculated as follows:

$$\begin{aligned} x_i^* &= \mathbf{E}(X_i\,|\,Y_i=y_i) = P(X_i=1\,|\,Y_i=y_i) \\ &= \frac{\alpha^*\phi(y_i-\mu^*)}{\alpha^*\phi(y_i-\mu^*)+(1-\alpha^*)\phi(y_i)}. \end{aligned}$$

## Incomplete two-way tables

As another example, let us consider a 2×-table with $n_1 = \{n^1_{ij}\}$ complete observations of two binary variables $I$ and $J$, $n^2 = \{n_{i+}$ observations where only $I$ was observed, and $n^3 = \{n_{+j}$ observations where only $J$ was observed, and let us assume that the mechanism of missingness can be ignored.

The complete data log-likelihood is

$$\log L(p) = \sum_{ij}(n^1_{ij}+n^2_{ij}+n^3_{ij})\log p_{ij}$$

and the E-step needs

$$n^*_{ij} = n^1_{ij} + n^{2*}_{ij} + n^{3*}_{ij}$$

where

$$n^{2*}_{ij} = \mathbf{E}(N^2_{ij}\,|\,p, n^2_{i+}) = p_{j\,|\,i}n^2_{i+}$$

and

$$n^{3*}_{ij} = \mathbf{E}(N^3_{ij}\,|\,p, n^3_{+j}) = p_{i\,|\,j}n^2_{+j}.$$

We thus get

$$n^{2*}_{ij} = \frac{p_{ij}}{p_{i0}+p_{i1}}n^2_{i+}, \quad n^{3*}_{ij} = \frac{p_{ij}}{p_{0j}+p_{1j}}n^3_{+j}. \qquad (3)$$

The M-step now maximizes $\log L(p) = \sum_{ij} n^*_{ij}\log p_{ij}$ by letting

$$p_{ij} = (n^1_{ij} + n^{2*}_{ij} + n^{3*}_{ij})/n \qquad (4)$$

where $n$ is the total number of observations.

The EM algorithm alternates between (3) and (4) until convergence.

# Latent Variable Models and Factor Analysis

## MSc Further Statistical Methods
## Lectures 6 and 7
## Hilary Term 2007

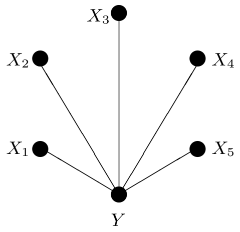
Steffen Lauritzen, University of Oxford; February 8, 2007

## Basic idea

Latent variable models attempt to explain complex relations between several variables by simple relations between the variables and an underlying unobservable, i.e. *latent* structure.

Formally we have a collection $X = (X_1, \ldots, X_p)$ of *manifest* variables which can be observed, and a collection $Y = (Y_1, \ldots, Y_q)$ of *latent* variables which are unobservable and 'explain' the dependence relationships between the manifest variables.

Here 'explaining' means that the *manifest variables are assumed to be conditionally independent given the latent variables*, corresponding e.g. to the following graph:



Here $Y$ is the latent variable(s) and there are 5 manifest variables $X_1, \ldots, X_5$.

For the model to be useful, *$q$ must be much smaller than $p$*.

Data available will be repeated observations of the vector $X = (X_1, \ldots, X_p)$ of manifest variables.

Latent variable models are typically classified according to the following scheme:

| | Manifest variable | |
|---|---|---|
| Latent variable | Metrical | Categorical |
| Metrical | Factor analysis | Latent trait analysis |
| Categorical | Latent profile analysis | Latent class analysis |

Other terminologies are used, e.g. *discrete factor analysis* for latent trait analysis.

Categorical variables can either be ordinal or nominal, and metrical variables can either be discrete or continuous.

## An example

A classical latent trait model is behind intelligence testing.

The intelligence of any individual is assumed to be a latent variable $Y$ measured on a continuous scale.

An intelligence test is made using a battery of $p$ tasks, and an individual scores $X_i = 1$ if the individual solves task $i$ and $0$ otherwise.

The test is now applied to a number of individuals to establish and estimate the parameters in the model.

Subsequently the test battery will be used to estimate the intelligence of a given individual by using

$$\mathbf{E}(Y \mid X_1 = x_1, \ldots, X_p = x_p)$$

as the estimate of intelligence for a given individual with score results $x = (x_1, \ldots, x_p)$.

Typical models will now have the intelligence distributed as

$$Y \sim \mathcal{N}(\mu, \sigma^2)$$

and the manifest variables as

$$\pi_i(y) = P(X_i = 1 \mid Y = y) = \frac{e^{\alpha_i + \beta_i y}}{1 + e^{\alpha_i + \beta_i y}}$$

corresponding to

$$\operatorname{logit}\{\pi_i(y)\} = \alpha_i + \beta_i y,$$

i.e. the response for each item being a logistic regression on the latent intelligence.

This model has too many parameters so we need to standardise and choose e.g. $\mu = 0$ and $\sigma^2 = 1$ to have a chance of estimating $\alpha_i$ and $\beta_i$.

We may increase the dimensionality of this model by assuming $Y$ and $\beta_i$ are $q$-dimensional and have

$$Y \sim \mathcal{N}_q(0, I), \quad \operatorname{logit}\{\pi_i(y)\} = \alpha_i + \beta_i^\top y.$$

This model is known as the *logit/normit model*.

Estimation is typically done by the EM-algorithm. The E-step involves numerical integration and the M-step needs in principle iterative methods as well.

See Bartholomew and Knott (1999), pp. 80–83 for details.

## Estimation in latent variable models

Historically, algorithms for maximizing the likelihood function have been developed separately for each specific model.

Generally, estimation problems can be very difficult and there are problems with uniqueness of estimates.

The difficulties show in particular if sample sizes are small and $p$ is not large relatively to $q$.

There are also severe problems with the asymptotic distribution of likelihood ratio tests.

*Latent variable models are perfectly suitable for the EM algorithm as $Y$ is MCAR.*

However, the general 'well-established' knowledge is that the EM algorithm is too slow.

Typicallly, the EM algorithm quickly gets close to the MLE, but then slows down. This suggests a hybrid approach to be suitable, where the EM algorithm is applied initially to get good starting values, then special algorithms for the final convergence.

MIM implements a version of the EM-algorithm which is applicable for *latent class analysis, latent profile analysis,* and *factor analysis,* but *not* latent trait analysis.

## The linear normal factor model

The $p$ *manifest* variables $X^\top = (X_1, \dots, X_p)$ are linearly related to the $q$ *latent* variables $Y^\top = (Y_1, \dots, Y_q)$ as

$$X = \mu + \Lambda Y + U, \qquad (1)$$

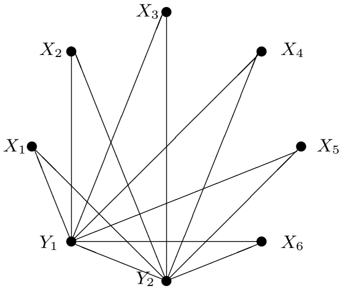
where $Y$ and $U$ are independent and follow multivariate normal distributions

$$Y \sim \mathcal{N}_q(0, I), \quad U \sim \mathcal{N}_p(0, \Psi),$$

where $\Psi$ is a *diagonal* matrix, i.e. the indidividual error terms $U_i$ are assumed independent.

The latent variables $Y_j$ are the *factors* and $\Lambda$ the matrix of *factor loadings*.

## Dependence graph of LNF model



Graph only displays conditional independences. In addition, $Y_1 \perp\!\!\!\perp Y_2$.

## Linear factor analysis

The *idea* of the LNF model is to describe the variation in $X$ by variation in a latent $Y$ plus noise, where the number of factors $q$ is considerably smaller than $p$.

The *problem* is now to determine the smallest $q$ for which the model is adequate, estimate the factor loadings and the error variances.

The marginal distribution of the observed $X$ is

$$X \sim \mathcal{N}_p(\mu, \Sigma), \quad \Sigma = \Lambda\Lambda^\top + \Psi.$$

The factor loadings $\Lambda$ cannot be determined uniquely. For example, if $O$ is an orthogonal $q \times q$-matrix and we let $\tilde{Y} = OY$ and $\tilde{\Lambda} = \Lambda O^\top$ we have

$$\tilde{\Lambda}\tilde{Y} = \Lambda O^\top OY = \Lambda Y$$

and thus

$$X = \mu + \Lambda Y + U = X + \mu + \tilde{\Lambda}\tilde{Y} + U.$$

Since also $\tilde{Y} \sim \mathcal{N}_q(0, I)$ and

$$\tilde{\Lambda}\tilde{\Lambda}^\top = \Lambda O^\top O \Lambda^\top = \Lambda\Lambda^\top,$$

$\Lambda$ and $\tilde{\Lambda}$ specify same distribution of the observable $X$.

Hence $\Lambda$ is only identifiable modulo orthogonal equivalence.

## Maximum likelihood estimation

Let

$$S = \frac{1}{N}\sum_{n=1}^{N}(X_n - \bar{X})(X_n - \bar{X})^\top$$

be the empirical covariance matrix. The likelihood function after maximizing in $\mu$ to obtain $\hat{\mu} = \bar{X}$ is

$$\log L(\Sigma) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log\det(\Sigma) - \frac{n}{2}\operatorname{tr}(\Sigma^{-1}S).$$

Maximizing this under the constraint $\Sigma = \Lambda\Lambda^\top + \Psi$ can be quite tricky.

After some (complex) manipulation, the likelihood equations can be collected in two separate equations. One is the obvious equation

$$\Psi = \operatorname{diag}(S - \Lambda\Lambda^\top) \qquad (2)$$

which gives $\Psi$ in terms of $S$ and $\Lambda$.

To express $\Lambda$ in terms of $S$ and $\psi$ is more complex. Introduce

$$S^* = \Psi^{-1/2} S \Psi^{-1/2}, \quad \Lambda^* = \Psi^{-1/2}\Lambda.$$

Then the MLE of $\Lambda^*$ can be determined by the following two criteria:

1. The columns of $\Lambda^* = (\lambda_1^* : \cdots : \lambda_q^*)$ are eigenvectors of the $q$ largest eigenvalues of $S^*$.
2. If $\Gamma$ is a diagonal matrix with $\Gamma_{ii}$ being the eigenvalue associated with $\lambda_i^*$, then

$$\Gamma_{ii} > 1, \quad S^*\Lambda^* = \Lambda^*\Gamma. \qquad (3)$$

A classic algorithm begins with an initial value of $\Psi$, finds the eigenvectors $e_i^*$ corresponding to the $q$ largest eigenvalues of $S^*$, lets $\lambda_i^* = \theta_i e_i^*$ and solves for $\theta_i$ in (3). When $\Lambda^*$ and thereby $\Lambda$ has been determined in this way, a new value for $\Psi$ is calculated using (2).

The algorithm can get severe problems if at some point the constraints $\psi_{ii} > 0$ and $\Gamma_{ii} > 1$ are violated.

The EM algorithm is a viable alternative which may not be sufficiently well exploited. See B & K(1999), §3.6 for details of this.

## Choice of the number of factors

Under regularity conditions, the *deviance*

$$\begin{aligned} D &= -2\{\log L(H_0) - \log L(H_1)\} \\ &= n\{\operatorname{tr}(\hat{\Sigma}^{-1}S) - \log\det(\hat{\Sigma}^{-1}S) - p\} \end{aligned}$$

has an approximate $\chi^2$-distribution with $\nu$ degrees of freedom where

$$\nu = \frac{1}{2}\{(p-q)^2 - (p+q)\}.$$

One can now either choose $q$ as small as possible with the deviance being non-significant, or one can minimze AIC or BIC where

$$AIC = D + 2\nu, \quad BIC = D + \nu \log N.$$

## Interpretation

To interpret the results of a factor analysis, it is customary to look at the *communality* $c_i$ of the manifest variable $X_i$

$$c_i = \frac{\mathbf{V}(X_i) - \mathbf{V}(U_i)}{\mathbf{V}(X_i)} = 1 - \frac{\psi_{ii}}{\psi_{ii} + \sum_{j=1}^{q} \lambda_{ij}^2}$$

which is the proportion of the variation in $X_i$ explained by the latent factors. Each factor $Y_j$ contributes

$$\frac{\lambda_{ij}}{\psi_{ii} + \sum_{j=1}^{q} \lambda_{ij}^2}$$

to this explanation.

Typically the variables $X$ are standardized so that they add to 1 and have unit variance, corresponding to considering just the empirical correlation matrix $C$ instead of $S$.

Then

$$\psi_{ii} + \sum_{j=1}^{q} \lambda_{ij}^2 = 1$$

so that $c_i = 1 - \psi_{ii}$ and $\lambda_{ij}^2$ is the proportion of $\mathbf{V}(X_i)$ explained by $Y_j$.

## Orthogonal rotation

Since $Y$ is only defined up to an orthogonal rotation, we can choose a rotation ourselves which seems more readily interpretable, for example one that 'partitions' the latent variables into groups of variables that mostly depend on specific factors, known as a *varimax* rotation

A little more dubious rotation relaxes the demand of orthogonality and allows skew coordinate systems and other variances than 1 on the latent factors, corresponding to possible dependence among the factors. Such rotations are *oblique*.

## Example

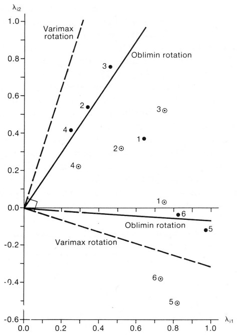
This example is taken from Bartholomew (1987) and is concerned with 6 different scores in intelligent tests. The $p = 6$ manifest variables are

1. Spearman's G-score
2. Picture completion test
3. Block Design
4. Mazes
5. Reading comprehension
6. Vocabulary

A 1-factor model gives a deviance of 75.56 with 9 degrees of freedom and is clearly inadequate.

A 2-factor model gives a deviance of 6.07 with 4 degrees of freedom and appears appropriate.

The loadings of each of the 6 variables can be displayed as black dots in the following diagram



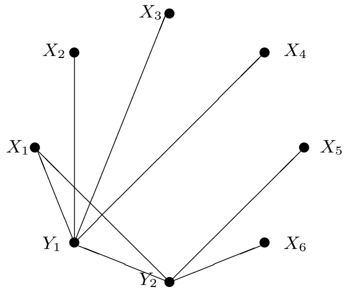
This diagram also shows axes corresponding to varimax and oblique rotations

It is tempting to conclude that 2, 3 and 4 seem to be measuring the same thing, whereas 5 and 6 are measuring something else. The G-score measures a combination of the two.

The axes of the oblique rotation represent the corresponding "dimensions of intelligence".

Or is it all imagination?

## Dependence graph of simplified model



$Y_1$ and $Y_2$ are no longer independent.

# Multilevel Analysis

**Further Statistical Methods, Lecture 8**
**Hilary Term 2007**

Steffen Lauritzen, University of Oxford; February 15, 2007

## Multilevel observations

Multilevel analysis is concerned with observations with a *nested* structure.

For a two-level analysis we typically think of *individuals* within *groups*. The individual level is in general called *level one*, the group level *level two*.

An example of observations of this type can for example be performance measures for *pupils* of a specific age-group within *classes*.

The levels could be nested yet another time as e.g. classes within *schools*. And further, the schools could be grouped according to *regions* within *countries*, etc. although at the top-level there might well be problems of compatibility of performance measures.

For simplicity we will only consider two levels, pupils within classes.

## An example

As our basic example we will consider a Dutch study comprising $N = 131$ classes, each of sizes between 4 and 35, with a total of $M = 2287$ pupils.

The performance measure of interest is the score on a language test, and explanatory variables include class sizes and the IQ of individual pupils.

We let $Y_{ij}, j = 1, \ldots N, i = 1, \ldots n_j$ be the score for pupil $i$ in class $j$ and study the dependence of this response on covariates such as the IQ $x_{ij}$ of the pupil and the size $z_j$ of the class.

$x_{ij}$ are *level one covariates* and $z_j$ *level two covariates*.

## A simple regression model

A first attempt could be to let

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_j + R_{ij}$$

with $R_{ij}$ independent and distributed as $\mathcal{N}(0, \sigma^2)$.

This is a standard linear regression model which only has an indirect multilevel character.

*The model ignores that pupils in the same class will tend to have more similar scores than those in different classes,* even when the covariates are taken into account.

This is a *very serious mistake* if the variations in score at group level are not fully explained by the covariates.

## Introducing random effects

For a moment, ignore the covariates $x_{ij}$ and $z_j$ and consider instead the model

$$Y_{ij} = \beta_0 + U_j + R_{ij}$$

where $U_j \sim \mathcal{N}(0, \tau^2)$. This model then has

$$\mathbf{V}(Y_{ij}) = \sigma^2 + \tau^2, \quad \mathrm{Cov}(Y_{ij}, Y_{i'j}) = \tau^2, \quad \mathrm{Cov}(Y_{ij}, Y_{i'j'}) = 0$$

so that scores of pupils within the same class are correlated. The correlation is

$$\rho = \frac{\tau^2}{\sigma^2 + \tau^2}$$

and is known as the *intraclass correlation coefficient.*

This type of model is also known as a *random effects model* since one could think of $\beta_j = \beta_0 + U_j$ as a group effect, in this case modelled as a random effect. Adding back the covariates leads to

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_j + U_j + R_{ij}.$$

It can give a better overview to introduce an intermediate variable describing the total class effect

$$M_j = \beta_0 + \beta_2 z_j + U_j; \quad Y_{ij} = M_j + \beta_1 x_{ij} + R_{ij}$$

where $M_j$ now become missing data, or rather latent variables.

## Estimation of parameters

The maximum likelihood (ML) estimates of the parameters can be obtained using the EM algorithm, treating $M_j$ as missing variables.

For 'complete data', with $M_j$ observed, the estimation problem splits into two simple linear regression problems

1. Estimating $(\beta_0, \beta_2, \tau^2)$ by regressing $M_j$ on $z_j$;
2. Estimating $\beta_1, \sigma^2$ by regressing $Y_{ij} - M_j$ on $x_{ij}$

Unfortunately the ML estimates of the variance components $(\sigma^2, \tau^2)$ can be very biased, as these do not take into account the loss in degrees of freedom due to the estimation of regression coefficients.

Instead a method known as *residual maximum likelihood* or REML is often used.

This involves (in principle) the following steps

1. Calculate initial estimates of regression coefficients using OLS, ignoring the multi-level structure;
2. Form residuals
$$\hat{r}_{ij} = y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 x_{ij} - \hat{\beta}_2 z_j.$$
3. These residuals $\hat{R}$ follow a multivariate normal distribution with mean $0$ and a covariance matrix $\Sigma(\sigma^2, \tau^2)$;
4. The REML estimates of $(\sigma^2, \tau^2)$ are the maximum likelihood estimates *based on the residuals*$\hat{R}$.
5. Revised estimates of the regression parameters are then calculated using appropriate weighted least squares.

An algorithm of EM type exists for calculating the REML estimates, but this and other methods have also been implemented in generally available software.

## Estimating random effects

It could be of independent interest, for example when making performance ranking, to estimate the level two effects which are not explained by covariates, i.e.

$$\beta_j = \beta_0 + U_j.$$

This can be done by calculating

$$\hat{\beta}_j = \hat{\beta}_0 + \hat{\mathbf{E}}(U_j \mid Y),$$

i.e. the estimated conditional expectation given the observed data.

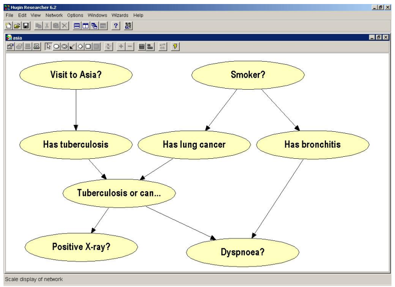
## A Bayesian alternative

An alternative method of analysis is to specify prior distributions of the unknown parameters.

The resulting model is then a *Bayesian hierarchical model.*

It has a simple representation as a Bayesian graphical model and WinBUGS provides the necessary software for estimating all relevant effects using Markov chain Monte-Carlo methods (MCMC).

## Example of a directed graphical model



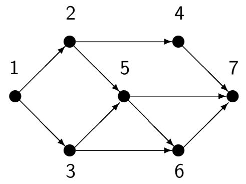
## Directed graphical models

A probability distribution *factorizes* w.r.t. a *directed acyclic graph* (DAG) $\mathcal{D}$ if it has density or probability mass function $f$ of the form

$$f(x) = \prod_{v \in V} f(x_v \mid x_{\mathrm{pa}(v)}),$$

i.e. into a product of the conditional distributions of each node given its parents.

## Example of DAG factorization



The above graph corresponds to the factorization

$$\begin{aligned} f(x) &= f(x_1) f(x_2 \mid x_1) f(x_3 \mid x_1) f(x_4 \mid x_2) \\ &\times f(x_5 \mid x_2, x_3) f(x_6 \mid x_3, x_5) f(x_7 \mid x_4, x_5, x_6). \end{aligned}$$

## Including parameters and observations

Directed graphical models become particularly useful when *parameters are explicitly included* in the graph.

The factorization can then be written as

$$f(x \mid \theta) = \prod_{v \in V} f(x_v \mid x_{\mathrm{pa}(v), \theta}).$$

Each conditional distribution may only depend of part of the parameter, the ‘parameter parents’.

To be able to describe complex observational patterns, we would wish to represent repeated structures. This can be done through *plates* as in WinBUGS.

## Warnings

*Beware that prior distributions can be influential.*

Note in particular that the parameters mean different things when covariates are centered in different ways, yielding different models with default prior specifications:

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2), \quad \alpha \sim N(0, 100), \quad \beta \sim N(0, 100)$$

is very different from

$$Y_i \sim N(\alpha + \beta x_i^*, \sigma^2), \quad \alpha \sim N(0, 100), \quad \beta \sim N(0, 100),$$

where $x_i^* = x_i - \bar{x}$. Without the prior specifications, the models would be equivalent, only the interpretation of $\alpha$ would be different.

WinBUGS makes inference on the parameters by MCMC computation. It is easy to specify a very complex model in WinBUGS. However, the results of the MCMC computation may then be very unreliable.

Additional comment:

Snijders and Bosker (1999) write that BUGS needs balanced data, i.e. equal group sizes, to be applied.

This is not correct, on the contrary, BUGS was developed to allow very unbalanced designs indeed.

# Longitudinal data

**Further Statistical Methods**
**Lecture 9**
**Hilary Term 2007**

Steffen Lauritzen, University of Oxford; February 20, 2007

## Longitudinal data

Longitudinal data can be seen as a specific type of multi-level data, where the level one units refer to observations over *time* of the value of specific quantities, taken on the same level two unit.

Typically level two units are here *individuals* $i = 1, \dots, N$. For each of them we have observations $Y_{ij}, j = 1, \dots, n_i$ taken at *times* $t_1, \dots, t_{n_i}$.

Models for longitudinal data differ from general multilevel data partly by almost always using *time as a covariate*, but specifically by using *time in the dependence structure* between measurements taken on the same units.

## Covariates for longitudinal data

As in the multilevel data we may have covariates $x_{ij} = (x_{ij1}, \dots x_{ijk})^\top$ and $z_i = (z_{i1} \dots, z_{il})^\top$ at both levels.

But for longitudinal data $x_{ij}$ typically include time or functions of time, such as e.g.

$$x_{ij1} = 1, \quad x_{ij2} = t_{ij}, \quad x_{ij3} = t_{ij}^2$$

corresponding to a quadratic trend, or

$$x_{ij1} = 1, \quad x_{ij2} = \cos(2\pi f t_{ij}), \quad x_{ij3} = \sin(2\pi f t_{ij})$$

corresponding to a periodic trend with period $\lambda = 1/f$, etc.

## A general linear model

The general linear model for longitudinal data is then given as

$$Y_{ij} = \alpha^\top z_j + \beta^\top x_{ij} + \epsilon_{ij},$$

where the errors $\epsilon_{ij}$ are multivariate Gaussian and *correlated* as

$$\mathrm{Cov}(\epsilon_{ij}, \epsilon_{i'j'}) = v_{ii'jj'}$$

where

$$v_{ii'jj'} = \begin{cases} c(t_{ij}, t_{ij'}) & \text{if } i = i' \\ 0 & \text{otherwise,} \end{cases}$$

for some *covariance model* determined by the function $c$. The models thus allow for correlation between observations from the same individual but assume independence between individuals.

## Correlation models

A flexible class of covariance models has three components:

$$c(t_{ij}, t_{ij'}) = \nu^2 + \sigma^2 \rho(t_{ij} - t_{ij'}) + \tau^2 \delta_{jj'},$$

where $\delta_{jj'}$ is $1$ for $j = j'$ and $0$ otherwise.

The first component $\nu^2$ reflects the intrinsic correlation between measurements taken on the same individual, as in the multilevel case.

The second component $\sigma^2$ describes a (stationary) serial correlation as known from time series analysis.

The final component $\tau^2$ corresponds to an instantaneous noise term.

## The variogram

The *variogram* for a stochastic process $X(t)$ is the function

$$\gamma(u) = \frac{1}{2}\mathbf{E}\left[\{X(t) - X(t-u)\}^2\right], \quad u \geq 0.$$

For the error process with three components just defined we get

$$\gamma(u) = \tau^2 + \sigma^2\{1 - \rho(u)\}, \text{ for } u > 0.$$

Choosing $\rho$ so that $\rho(0) = 1, \lim_{t\to\infty} \rho(t) = 0$ yields

$$\gamma(0) = \tau^2, \quad \lim_{t\to\infty} \gamma(u) = \sigma^2 + \tau^2 \tag{1}$$

whereas the process variance is

$$\mathbf{V}\{Y(t_{ij})\} = c(t_{ij}, t_{ij}) = \nu^2 + \sigma^2 + \tau^2, \tag{2}$$

as reflected in the following diagram, taken from Diggle et al. (2002).

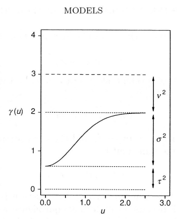


**Fig. 5.4.** The variogram for a model with a random intercept, serial correlation, and measurement error.

## Sample variogram

To identify reasonable suggestions for the covariance structure, *residuals* $r_{ij}$ from a least squares fit of the parameters are calculated and the *sample variogram* is based on a curve through points $(u_{ijk}, v_{ijk})$, where

$$u_{ijk} = t_{ij} - t_{ik}, \quad v_{ijk} = \frac{1}{2}(r_{ij} - r_{ik})^2$$

or rather averages of $v_{ijk}$ for indices corresponding to identical time differences $u$.

Such a sample variogram gives a first idea of the importance of the three components of variance using (1) and (2) and some idea of the shape of the serial correlation function $\rho$.

An example of a sample variogram, taken from Diggle et al. (2002) is seen below. Note that there are few large time differences, so the variogram becomes noisy for large lags, here around lag 10.

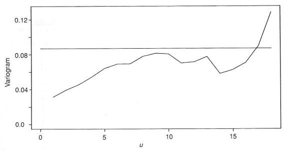


**Fig. 3.16.** Sample variogram of milk protein residuals. Horizontal line estimates process variance.

In this case there is essentially no within pig correlation.

## Choice of correlation function

Generally the time series are often many but short, so there is little information about the shape of the serial correlation function and one is forced to rather ad hoc choices.

The serial correlation function must be positive definite to ensure matrices of the form $m_{rs} = \rho(t_r - t_s)$ are positive definite, for all choices of $k$ and $t_1, \dots, t_k$.

Typical choices which satisfy these restrictions are

$$\rho_1(t) = e^{-\theta|t|}, \quad \rho_2(t) = e^{-\theta t^2/2},$$

known as the *exponential* and *Gaussian* correlation model.

It can be difficult to distinguish these from the sample variogram.

## Estimation of parameters

In principle this is done in the same way as in other multi-level models, using *residual maximum likelihood* (REML).

Straight maximum likelihood yields strongly biased estimates of the variance parameters and should be avoided.

Routines for calculating the REML estimates are available in many forms of software.

They can be calculated using the following steps:

1. Calculate estimates $(\tilde{\alpha}, \tilde{\beta})$ of the linear parameters by ordinary least squares (OLS), ignoring the correlation;

2. Calculate the residuals

$$r_{ij} = y_{ij} - \tilde{\alpha}^\top z_j - \tilde{\beta}^\top x_{ij}$$

from the OLS analysis;

3. The vector $R$ of residuals is $\mathcal{N}(0, W)$ where the covariance matrix $W$ has the form

$$W = \nu^2 A + \sigma^2 B(\theta) + \tau^2 C$$

where $A, B, C$ are known matrices, $B$ possibly depending on $\theta$;

4. Calculate the MLE of $(\nu^2, \sigma^2, \tau^2, \theta)$ *based on the likelihood for the residuals*;

5. Calculate the final estimates $(\hat{\alpha}, \hat{\beta})$ using *weighted least squares* (WLS) with weights determined by the given covariance model and its estimated parameters.

# Alternative Methods and Models for Longitudinal Data

**Further Statistical Methods, Lecture 10**
**HT 2007**

Steffen Lauritzen, University of Oxford; February 21, 2007

## Types of longitudinal data

There are many cases where the 'standard model' from last lecture is inadequate, i.e. when the data are not well described as the sum of three components: a general trend, a (stationary) component with serial correlation, and random noise.

This is for example true for such cases as

- *Biokinetics:* A substance is introduced into a person and the concentration level of one or more components is measured at selected time intervals over a period.

  The 'substance' can e.g. be one or more specific drugs or types of food.

  The purpose of such analysis may be to understand the *shape of the curve*, to get a grip of the *duration* of a transient phenomenon, or e.g. the variation in the *maximally achieved value.*
- *Cucumber plants* are grown in greenhouses. One would like to know how different watering/fertilization/treatment schemes affect the growth. Cucumbers are picked daily from each plant and recorded.

  Cucumbers have a season. It takes a while before they develop, then they give a lot of cucumbers for a while, and then stop. The farmer would like to have a lot of cucumbers when others don't, so the price is high.
- *Event history data* follow individuals over time and record when events happen.
- *Flowers* under different conditions. They develop buds, the buds become flowers, and then die. Different treatments make the plants develop differently.

  Plants that have lots of buds and some flowers are selling best.

  This can be seen as a type of event history data.
- *Panel data* follow a group of individuals (panel members) over time. From time to time the members are filling questionnaires, for example on their political or consumer preferences.
- *Growth models.* It is not always reasonable to assume this to be trend plus stationary error. Typically growth can be high in some periods and low in others, with some random variation.
- *Speech analysis.* Frequency properties of speech is recorded at dense discrete time points (millisecond intervals). One is interested in describing the behaviour as different phonemes are pronounced, e.g. for automatic speech recognition and -understanding.

## Descriptive methods

Transform an observed curve to a some *features*, e.g.

- The area $A$ under the curve, representing the total amount of something;
- The maximal value $M$ reached of the curve;
- The total duration $D$ of a signal, i.e. the time spent above a certain level.
- A set of Fourier- or wavelet coefficients $F$;
- etc...

Now use your favourite (multivariate) technique to analyse (part of) the vector $A, M, D, F$.

## Differential equations

If the phenomenon observed is well understood, there might be a relevant differential equation explaining the main features of the observations.

An example from insulin kinetics postulates the following relation between the plasma glucose concentration $G(t)$, insulin concentration $I(t)$, and the insulin's effect on the net glucose disappearance $X(t)$:

$$\begin{aligned}
\dot{G}(t) &= -p_1\{G(t) - G_b\} - X(t)G(t), \quad G(0) = 0, \\
\dot{X}(t) &= -p_2 X(t) + p_3\{I(t) - I_b\}, \quad X(0) = 0, \\
\dot{I}(t) &= -n\{I(t) - I_b\} + \gamma\{G(t) - h\}^+ t, \quad I(0) = 0.
\end{aligned}$$

This is known as *Bergman's minimal model*.

The parameters are *individual* and to be determined from observations. The important quantities are

- Insulin sensitivity: $S_I = p_1/p_2$;
- Glucose effectiveness: $S_G = p_1$;
- Pancreatic responsiveness: $(\phi_1, \phi_2)$ where $\phi_1 = (I_{\max} - I_b)/\{n(G_0 - G_b)\}$, $\phi_2 = \gamma \times 10^4$.

This is generally difficult, as only $G(t), I(t)$ can be observed, and only at discrete time points. Using graphical models and MCMC in the right way, it is possible.

This general area is known as PK/PD for pharmaco-kinetics/-dynamics.

## Dynamic models

These models, also known as *state-space models* (SSM) are similar in spirit to differential equation models.

Typically they have two levels, but sometimes more. One level describes the development of an unobserved (hidden) *state* $X_t$, typically using a Markov model with e.g.

$$\mathcal{L}(X_{t+1} \mid X_s = x_s, s \leq t, \theta) \sim \mathcal{N}\{A_t(\theta)x_t, \sigma_t^2(\theta)\}$$

and an *observational model* for $Y_t$ with

$$\mathcal{L}(Y_t \mid X, \eta) = \mathcal{N}\{B_t(\eta)x_t, \tau^2(\eta)\},$$

where $Y_t, t = 1, \ldots, T$ are observed.

Parameters are then estimated by using a variant of the EM algorithm. The E-step can be performed elegantly using a recursive algorithm known as the *Kalman Filter*.

MCMC is also a viable alternative and a hot research topic is that of *particle filters* which can be seen as MCMC variants of the Kalman filter.

Generalisations include replacing each of the models above with *generalised linear models*.

For example, in the cucumber example it is natural to consider Poisson model for the observed number of cucumbers on a plant.

In speech analysis, $Y$ is typically a *feature vector* of the signal and the state space equation should depend on what the individual is saying. Hence another level is typically introduced with $Z_t$ discrete taking values in possible *phonemes* and and following a Markov model so that

$$P(Z_{t+1} = z_{t+1} \mid Z_s = z_s, s \leq t) = q(z_{t+1} \mid z_t, \theta),$$

and

$$\mathcal{L}(X_{t+1} \mid (X_t = x_s, Z_t = z_s), s \leq t, \theta) \sim \mathcal{N}\{A_t(\theta, z_t)x_t, \sigma_t^2(\theta, z_t)\}.$$

and

$$\mathcal{L}(Y_t \mid X, \eta) = \mathcal{N}\{B_t(\eta)x_t, \tau^2(\eta)\},$$

where $Y_t, t = 1, \ldots, T$ are observed.

Such models are *switching state space* models (SSM).

If the middle level is missing, it is also called a *hidden Markov model* (HMM).