Further Statistical Methods

Contingency Tables, Hilary Term, 2015

Marco Scutari

scutari@stats.ox.ac.uk
Department of Statistics
University of Oxford

July 27, 2015

Course Information

Lectures

Week 5: Monday 11am, Thursday 11am, Friday 10am

No Practical!

Reference Books (further references in the next slide)

AA Agresti A (2013). Categorical Data Analysis. Wiley, 3rd edition.

DE Edwards D (2000). Introduction to Graphical Modelling. Springer, 2nd edition.

SF Fienberg SE (2007). The Analysis of Cross-Classified Categorical Data. Springer, 2nd edition.

JP Pearl J (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann.

Other Useful Books on Contingency Tables

- Agresti A (2010). Analysis of Ordinal Categorical Data. Wiley, 2nd edition.
- Bishop YMM, Fienberg SE, Holland PW (2007). Discrete Multivariate Analysis: Theory and Practice. Springer.
- Koller D, Friedman N (2009). Probabilistic Graphical Models. MIT ress.
- Lauritzen S (1996). Graphical Models. Oxford University Press.
- Pesarin F, Salmaso L (2010). Permutation Tests for Complex Data: Theory, Applications and Software. Wiley.
- Whittaker J (1990). Graphical Models in Applied Multivariate Statistics. Wiley.

Overview

1. Models and Probability Distributions

   [AA 2, 3 & 8; DE 2]

2. Hypothesis Testing

   [AA; DE 5; SF 3.8]

3. Graphical Models

   [JP 3; DE 2]

Models and Probability Distributions

What is a Contingency Table?

A contingency table is a tabular representation of the absolute frequencies of 2 or more discrete variables, encoding their joint distribution in a set of cells corresponding to the combinations (configurations) of their values. Each of those discrete variables can be:

1. a categorical random variable, defined on an unordered set of values (i.e. the `level()`s of the `factor`);
2. an ordinal random variable, defined on an ordered set of values (e.g. small/large; $0 - 10, 11 - 20, > 20$).

The main difference is that in the latter case the CDF is defined, as is the concept of trend.

A natural counterpart to a contingency table is a probability table, which has the same layout but has probabilities instead of frequencies in the cells. So, in the case of a two-dimensional table,

$$\{n_{ij}\} \quad \text{in the contingency table and} \quad \{\pi_{ij}\} \quad \text{in the probability table.}$$

A Two-Dimensional Contingency Table: Seat-Belts

This example from [AA 3] shows fatality results for children under 18 who were passengers in car accidents in Florida in 2008, according to whether the child was wearing a seat belt.

	Injury Outcome		
Seat Belt Use	Fatal	Nonfatal	Total
No	54	10325	10379
Yes	25	51790	51815
Total	79	62115	62194

The data is observational (it does not arise from a designed experiment) so none of the totals are fixed.

A Two-Dimensional Contingency Table (R Code)

```
> # data in table form.
> belt = matrix(c(54, 25, 10325, 51790), nrow = 2,
+          dimnames = list(Seatbelt = c("No", "Yes"),
+                               Injury = c("Fatal", "Nonfatal")))
> belt = as.table(belt)
> belt
        Injury
Seatbelt Fatal Nonfatal
     No     54    10325
     Yes    25    51790
> # data in data frame form.
> as.data.frame(belt)
  Seatbelt   Injury  Freq
1       No    Fatal    54
2      Yes    Fatal    25
3       No Nonfatal 10325
4      Yes Nonfatal 51790
```

A Three-Dimensional Contingency Table: Lizards

This small data set is from [SF] and is also used extensively in [DE].

Species	Perch Diameter	Perch Height	
		> 4.75ft	$\leqslant 4.75$ft
Anolis	$\leqslant 4$in	32	86
	> 4in	11	35
Distichus	$\leqslant 4$in	61	73
	> 4in	41	70

For a sample of 409 lizards, the following variables were recorded:

- the species, which can be either "Sagrei" or "Distichus";
- the height of the branch they were perched on, discretised in two categories `narrow` ($\leqslant 4$in) and `wide` (> 4in);
- the diameter of that same branch, discretised in two categories `high` (> 4.75ft) and `low` ($\leqslant 4.75$ft).

A Three-Dimensional Contingency Table

```
> lizards = read.table("lizards.txt", header = TRUE)
> head(lizards)
  Species Diameter Height
1  Sagrei   narrow    low
2  Sagrei   narrow    low
3  Sagrei   narrow    low
4  Sagrei   narrow    low
5  Sagrei   narrow    low
> table(lizards)
, , Height = high

           Diameter
Species     narrow wide
  Distichus     73   70
  Sagrei        86   35

, , Height = low

           Diameter
Species     narrow wide
  Distichus     61   41
  Sagrei        32   11
```

```
> table(lizards[, c("Diameter", "Height", "Species")])
, , Species = Distichus

        Height
Diameter high low
  narrow   73  61
  wide     70  41

, , Species = Sagrei

        Height
Diameter high low
  narrow   86  32
  wide     35  11
```

A Contingency Table with Ordinal Variables: Income

This small example from [DE 5] describes a survey on job satisfaction as a function of income in the United States. The sample size can considered to be fixed, as the number of questionnaires is fixed in advance.

Income ($)	Job Satisfaction			
	Very Dissatisfied	Little Dissatisfied	Moderately Satisfied	Very Satisfied
< 6000	20	24	80	82
$6000 - 15000$	22	38	104	125
$15000 - 25000$	13	28	81	113
> 25000	7	18	54	92

```
> job = read.table("job.satisfaction.txt", header = TRUE)
> job$Income = ordered(job$Income,
+   levels = c("< 6000", "6000-15000", "15000-25000", "> 25000"))
> job$Satisfaction = ordered(job$Satisfaction,
+   levels = c("Very Dissatisfied", "Little Dissatisfied",
+              "Moderately Satisfied", "Very Satisfied"))
```

Notation for Cells and Totals

Standard notation is:

- n_{ijk} with $i = 1, \ldots, R$, $j = 1, \ldots, C$ and $k = 1, \ldots, L$ is the cell on the ith row, jth column and kth level (of the third variable).
- Row, columns and level totals (marginals) are

$$n_{i++} = \sum_{j=1}^{C} \sum_{k=1}^{L} n_{ijk}, \quad n_{+j+} = \sum_{i=1}^{R} \sum_{k=1}^{L} n_{ijk}, \quad n_{++k} = \sum_{i=1}^{R} \sum_{j=1}^{C} n_{ijk}. \qquad (1)$$

- Totals for the sub-tables defined by one of the variables are

$$n_{i+k} = \sum_{j=1}^{C} n_{ijk}, \qquad \text{and} \qquad n_{j+k} = \sum_{i=1}^{R} n_{ijk}. \qquad (2)$$

- n is the sample size, i.e. the overall total of the table; in two-dimensional tables it is also denoted as n_{+++}.

The notation for the probabilities follows the same scheme (e.g. π_{i++} is the probability associated with n_{i++}).

margin.table(): Totals and Marginals

We can compute marginals with `margin.table()`, which has a `margin` argument to specify which dimensions of the table to retain. For a single variable, it produces n_{i++}, n_{+j+} and n_{++k}.

```
> margin.table(table(lizards), margin = 1)
Species
   Sagrei Distichus
      164       245
```

For two variables, it produces n_{ij+}, n_{i+k} and n_{+jk}.

```
> margin.table(table(lizards), margin = 2:3)
        Height
Diameter high low
  narrow  159  93
  wide    105  52
```

Combining `margin.table()` with subsetting we can produce all sub-tables and marginals.

expand.dft(): Expanding a Contingency Table

It is sometimes convenient to expand a contingency table into a data frame with one row for each observation; many functions in R can handle the latter but not the former.

```
> library(vcdExtra)
> lizards.df = expand.dft(table(lizards))
> head(lizards.df)
  Species Diameter Height
1  Sagrei   narrow   high
2  Sagrei   narrow   high
3  Sagrei   narrow   high
4  Sagrei   narrow   high
5  Sagrei   narrow   high
6  Sagrei   narrow   high
> str(lizards.df)
’data.frame':   409 obs. of  3 variables:
 $ Species : Factor w/ 2 levels "Distichus","Sagrei": 2 2 2 2 2 2 2 2 ...
 $ Diameter: Factor w/ 2 levels "narrow","wide": 1 1 1 1 1 1 1 1 ...
 $ Height  : Factor w/ 2 levels "high","low": 1 1 1 1 1 1 1 1 ...
```

Probabilistic Assumptions for Contingency Tables

The right distribution for the frequencies in a contingency tables depends on the underlying sampling distribution.

- Poisson sampling treats the counts n_{ijk} as independent Poissons with parameters μ_{ijk}, which means that the overall total n is not considered fixed.
- Multinomial sampling treats counts n_{ijk} as the outcomes of a multinomial with probabilities π_{ijk} that sum up one. n is considered fixed.
- Independent multinomial sampling one or more sets of marginal counts are fixed, and each of the resulting sub-tables (e.g. n_{ijk} for fixed k) has an independent multinomial distribution with probabilities $\pi_{ij|k}$ such that $\sum_{ij} \pi_{ij|k} = 1$. As a side effect, n is also fixed as a result.

The most common assumption is by far multinomial sampling.

Different Sampling Schemes: Are They Related?

Poisson sampling is simply

$$n_{ijk} \sim Pois(\mu_{ijk}) \qquad \text{independently for all } i, j, k. \qquad (3)$$

From probability theory then we know that

$$n = \sum_{ijk} n_{ijk} \sim \sum_{ijk} Pois(\mu_{ijk}) = Pois\left(\sum_{ijk} \mu_{ijk}\right) \qquad (4)$$

so if we let $\pi_{ijk} = \mu_{ijk} / \sum_{ijk} \mu_{ijk}$ we have

$$n_{ijk} \mid n \sim Bi(n, \pi_{ijk}) \qquad \text{and} \qquad \{n_{ijk} \mid n\} \sim Mul(n, \{\pi_{ijk}\}) \qquad (5)$$

which is multinomial sampling. Moving to independent multinomial sampling involves a rescaling of the probabilities to re-normalise them:

$$\{n_{ijk} \mid n_k\} \sim Mul(n_k, \{\pi_{ij|k}\}) \quad \text{with} \quad \pi_{ij|k} = \pi_{ijk} / \sum_{ij} \pi_{ijk}. \qquad (6)$$

Roles of the Variables in a Contingency Table

Modelling a contingency table differs substantially depending on which roles we assign to the variables, which in turn depends on the aim of the analysis.

- If there is one clear variable of interest, we can put that into a GLM and use all the all other variables as regressors encoded as dummy variables. The resulting GLM will then be Binomial or Multinomial depending on how many levels the response has.
- We may interested in explaining the cell counts as a function of all the variables, resulting in a Poisson GLM.
- We may also be interested in which variables are conditionally or marginally dependent on each other; conditional independence tests and graphical models are best for that.

Contingency Table as a Binomial GLM

```
> summary(glm(Species ~ Diameter + Height, data = lizards,
+   family = binomial))
[...]

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2390  -0.9326  -0.6609   1.1170   1.8048

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.1437     0.1503   0.956 0.338972
Diameterwide  -0.8029     0.2198  -3.652 0.000260 ***
Heightlow     -0.7511     0.2242  -3.350 0.000807 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 550.85  on 408  degrees of freedom
Residual deviance: 526.57  on 406  degrees of freedom
AIC: 532.57
```

Contingency Table as a Poisson GLM

```
> summary(glm(Freq ~ Seatbelt + Injury, data = as.data.frame(belt),
+   family = poisson))
[...]

Deviance Residuals:
      1        2        3        4
 8.4053  -5.7648  -0.4012   0.1794

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.57897    0.11286   22.85   <2e-16 ***
SeatbeltYes     1.60790    0.01075  149.52   <2e-16 ***
InjuryNonfatal  6.66729    0.11257   59.23   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 115243.62  on 3  degrees of freedom
Residual deviance:    104.07  on 1  degrees of freedom
AIC: 144.74
```

GLMs for Multinomial Responses

Multinomial GLMs are an extension of Binomial GLMs. After choosing one level of the response as the baseline (say, y_R), the model fits the following set of simultaneous equations:

$$
\begin{aligned}
&\operatorname{logit}(Y_i = y_r \mid Y_i \in \{y_r, y_R\}) = \\
&\qquad = \log\left[\frac{\mathrm{P}(Y_i = y_r \mid Y_i \in \{y_r, y_R\})}{1 - \mathrm{P}(Y_i = y_r \mid Y_i \in \{y_r, y_R\})}\right] = \\
&\qquad = \beta_{0(r)} + x_{i1}\beta_{1(r)} + \ldots + x_{ip}\beta_{p(r)} \qquad \text{for } r = 1, \ldots, R.
\end{aligned} \tag{7}
$$

Only $R - 1$ simultaneous equations are needed, the logit values for other pairs of levels (y_r, y_s) can be derived as

$$
\begin{aligned}
&\operatorname{logit}(Y_i = y_r \mid Y_i \in \{y_r, y_s\}) = \\
&\quad = \operatorname{logit}(Y_i = y_r \mid Y_i \in \{y_r, y_R\}) - \operatorname{logit}(Y_i = y_s \mid Y_i \in \{y_s, y_R\})
\end{aligned} \tag{8}
$$

and from there the parameters of the regression models.

Contingency Tables as Multinomial GLMs

```
> library(nnet)
> summary(multinom(Satisfaction ~ Income, data = job))

Call:
multinom(formula = Satisfaction ~ Income, data = job)

Coefficients:
                     (Intercept) Income.L Income.Q Income.C
Little Dissatisfied         0.61     0.56   -0.094    0.022
Moderately Satisfied        1.70     0.50    0.023   -0.038
Very Satisfied              1.97     0.88    0.044   -0.025

Std. Errors:
                     (Intercept) Income.L Income.Q Income.C
Little Dissatisfied         0.17     0.37     0.34     0.31
Moderately Satisfied        0.15     0.33     0.30     0.28
Very Satisfied              0.15     0.32     0.30     0.27

Residual Deviance: 2085
AIC: 2109
```

From Multinomial to Ordered Responses

If the response is an ordinal random variable, then the multinomial GLM changes cumulative logit, i.e. a logit link on the cumulative distribution function

$$\begin{aligned}\operatorname{logit}(Y_i \leqslant y_r) = \log\left(\frac{\mathrm{P}(Y_i \leqslant y_r)}{1 - \mathrm{P}(Y_i \leqslant y_r)}\right) = \\ = \log\left(\frac{F_{Y_i}(y_r)}{1 - F_{Y_i}(y_r)}\right) = \beta_{0(r)} + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p \qquad (9)\end{aligned}$$

with a different intercept for each level but the same regression coefficients across levels. Intercepts $\beta_{0(r)}$ are constrained to be increasing in r so that $\mathrm{P}(Y_i \leqslant y_r \mid \mathbf{X})$ increases in r for any fixed set of explanatory variables $\mathbf{X}$.

This is called a cumulative or proportional odds (ratio) model.

Contingency Tables as Ordinal Regressions

```
> library(MASS)
> summary(polr(Satisfaction ~ Income, data = job))

Re-fitting to get Hessian

Call:
polr(formula = Satisfaction ~ Income, data = job)

Coefficients:
           Value Std. Error t value
Income.L  0.4163      0.136   3.052
Income.Q  0.0538      0.128   0.422
Income.C -0.0150      0.119  -0.126

Intercepts:
                                         Value   Std. Error t value
Very Dissatisfied|Little Dissatisfied     -2.641   0.133    -19.881
Little Dissatisfied|Moderately Satisfied  -1.490   0.087    -17.173
Moderately Satisfied|Very Satisfied        0.151   0.068      2.215

Residual Deviance: 2087.63
AIC: 2099.63
```

Estimating Parameters in Contingency Tables

Under the multinomial sampling assumption, estimating the parameters of a contingency table means estimating the probabilities π_{ijk} associated with the cells.

- The usual frequentist estimator is the relative frequency

$$\hat{\pi}_{ijk} = \frac{n_{ijk}}{n} \tag{10}$$

  which is also the maximum likelihood estimator.
- Some careful considerations are required when dealing with sparse tables, i.e. tables with low counts and/or many zero cells.
- Bayesian posterior estimators are constructed from the Dirichlet conjugate prior for the multinomial distribution,

$$Dir(\{\alpha_{ijk}\}) \quad \text{and} \quad Mul(n, \{\pi_{ijk}\}) \Rightarrow Dir(\{\alpha_{ijk} + n\pi_{ijk}\}). \tag{11}$$

prop.table(): Computing Cell Probabilities

The frequentist estimator $\hat{\pi}_{ijk}$, the $\hat{\pi}_{ij|k}$, and marginal probabilities like $\hat{\pi}_{i++}$ can all be computed with `prop.table()`. The syntax is similar to that of `margin.table()`, and the two functions can be combined.

```
> prop.table(table(lizards))
, , Height = high

           Diameter
Species      narrow  wide
  Sagrei      0.210 0.086
  Distichus   0.178 0.171

, , Height = low

           Diameter
Species      narrow  wide
  Sagrei      0.078 0.027
  Distichus   0.149 0.100
```

```
> prop.table(table(lizards), margin = 3)
, , Height = high

           Diameter
Species      narrow  wide
  Sagrei      0.326 0.133
  Distichus   0.277 0.265

, , Height = low

           Diameter
Species      narrow  wide
  Sagrei      0.221 0.076
  Distichus   0.421 0.283
```

Sparse Contingency Tables: Small Cell Counts

The frequentist estimator $\hat{\pi}_{ijk}$ is problematic for sparse contingency tables, that is, when n is not large compared to the number of cells $R \times C \times L$ because:

- a number of cell are bound to have zero counts $(n_{ijk} = 0)$, and we do not know whether it is impossible to observe that configuration of the variables or it is just rare enough that we do not have it in the sample;
- some estimated probabilities will be either $\hat{\pi}_{ijk} = 0$ or $\hat{\pi}_{ijk} = 1$, which places them right at the boundary of their domain and thus breaks the assumptions of most asymptotic results.

In such cases we have three options:

- applying a continuity correction to the n_{ijk} or collapsing levels;
- using a Bayesian posterior approach to move the $\hat{\pi}_{ijk}$ away from zero and one;
- use a shrinkage approach to do essentially the same thing but in a non-Bayesian way.

Continuity Corrections

In some situations the easiest solution to small n_{ijk} is to collapse levels for one or more variables, e.g. merging adjacent age brackets. The new cell counts are larger as a result, which improves the properties of the contingency table because the number of cell is reduced at the same time.

In the case of 2×2 tables, cell counts can be corrected by adding or subtracting $1/2$ to the cell counts based on the consideration that for large expected counts ($n\pi_{ijk} \gg 0$ or $\mu_{ijk} \gg 0$)

$$\mathrm{P}(X \leqslant x) = \mathrm{P}(X < x + 1) \simeq \mathrm{P}(X \leqslant x + 1/2) \qquad (12)$$

for both the Poisson and Binomial distributions. This idea is called Yates correction in the context of independence tests, but it has quite a few limitations so other solutions are preferred in modern practice.

The Dirichlet-Multinomial Posterior Distribution

The Dirichlet prior for the π_{ijk} is

$$f(\{\pi_{ijk}\}; \{\alpha_{ijk}\}) = \frac{\Gamma(\sum_{ijk} \alpha_{ijk})}{\prod_{ijk} \Gamma(\alpha_{ijk})} \prod\nolimits_{ijk} \pi_{ijk}^{\alpha_{ijk}-1}, \qquad \alpha_{ijk} > 0, \pi_{ijk} \in (0,1), \sum\nolimits_{ijk} \pi_{ijk} = 1; \tag{13}$$

and the multinomial density is

$$f(\{n_{ijk}\}; n, \{\pi_{ijk}\}) = \frac{n!}{\prod_{ijk} n_{ijk}!} \prod\nolimits_{ijk} \pi_{ijk}^{n_{ijk}}, \qquad \pi_{ijk} \in [0,1], n_{ijk} \in \mathbb{N}^{+}, \sum\nolimits_{ijk} \pi_{ijk} = 1; \tag{14}$$

so the Dirichlet posterior is

$$f(\{\pi_{ijk}\}; \{n\pi_{ijk} + \alpha_{ijk}\}) = \frac{\Gamma(n + \sum_{ijk} \alpha_{ijk})}{\prod_{ijk} \Gamma(n\pi_{ijk} + \alpha_{ijk})} \prod\nolimits_{ijk} \pi_{ijk}^{n\pi_{ijk} + \alpha_{ijk} - 1}. \tag{15}$$

Posterior for the Independent Multinomial Sampling

In the case of independent multinomial sampling, we typically have a collection

$$f_1(\{\pi_{ij|1}\}; \{\alpha_{ij|1}\}), \ldots, f_L(\{\pi_{ij|L}\}; \{\alpha_{ij|L}\}) \tag{16}$$

of $k = 1, \ldots, L$ independent priors that result in L independent Dirichlet posteriors

$$f_k(\{\pi_{ij|k}\}; \{n\pi_{ij|k} + \alpha_{ij|k}\}) = \frac{\Gamma(n_{++k} + \sum_{ijk} \alpha_{ij|k})}{\prod_{ijk} \Gamma(n\pi_{ij|k} + \alpha_{ij|k})} \prod_{ijk} \pi_{ijk}^{n\pi_{ijk} + \alpha_{ijk} - 1}, \tag{17}$$

which are then combined to give the overall posterior for the contingency table:

$$f(\{\pi_{ij|k}\}; \{n\pi_{ij|k} + \alpha_{ij|k}\}) = \prod_{k=1}^{L} f_k(\{\pi_{ij|k}\}; \{n\pi_{ij|k} + \alpha_{ij|k}\}). \tag{18}$$

What is the Interpretation of the Dirichlet Conjugate?

The Dirichlet distribution can be viewed as a generalisation of the multinomial distribution. If we take a $Dir(\{\alpha_{ijk}\})$,

$$f(\{\pi_{ijk}\};\{\alpha_{ijk}\}) = \frac{\Gamma(\sum_{ijk}\alpha_{ijk})}{\prod_{ijk}\Gamma(\alpha_{ijk})}\prod_{ijk}\pi_{ijk}^{\alpha_{ijk}-1} \tag{19}$$

and we say $\alpha_{ijk} - 1 = m_{ijk}$ with $m_{ijk} \in \mathbb{N}$ and $\sum_{ijk}\alpha_{ijk} = m + RCL$, then we have

$$f(\{\pi_{ijk}\};\{m_{ijk}\}) = \frac{\Gamma(\sum_{ijk}[m_{ijk}+1])}{\prod_{ijk}\Gamma(m_{ijk}+1)}\prod_{ijk}\pi_{ijk}^{m_{ijk}} \tag{20}$$

$$= \frac{(m + RCL - 1)!}{\prod_{ijk} m_{ijk}!}\prod_{ijk}\pi_{ijk}^{m_{ijk}}. \tag{21}$$

In other words, we can think of the Dirichlet as a multinomial encoding an imaginary sample with size $\sum_{ijk}\alpha_{ijk}$ supporting the prior.

Parameters in the Prior and the Posterior

Coherently with this interpretations, the estimated probability for each cell in the prior is

$$\tau_{ijk} = \frac{\alpha_{ijk}}{N} \qquad \text{with} \qquad N = \sum\nolimits_{ijk} \alpha_{ijk} \tag{22}$$

and the corresponding estimate in the posterior is

$$\tilde{\pi}_{ijk} = \frac{\alpha_{ijk} + n\pi_{ijk}}{n + \sum_{ijk} \alpha_{ijk}}, \tag{23}$$

which can be rewritten as a convex combination of the prior and the observed cell probabilities

$$\frac{\alpha_{ijk} + n\pi_{ijk}}{n + \sum_{ijk} \alpha_{ijk}} = \frac{N\tau_{ijk} + n\pi_{ijk}}{n + N} = \frac{N}{N+n}\tau_{ijk} + \frac{n}{N+n}\pi_{ijk}. \tag{24}$$

The Imaginary Sample Size

The quantity $N = \sum_{ijk} \alpha_{ijk}$ is called the imaginary sample size, and controls the "weight" of the prior compared to the observed data:

- if $N \gg n$ then the prior dominates the likelihood;
- if $n \gg N$ then the likelihood dominates the prior.

We prefer the latter because when we are using a simple prior, such as the uniform

$$\alpha_{ijk} = \frac{N}{RCL} \qquad \text{for all } i, j \text{ and } k \tag{25}$$

the ratio N/n acts as a smoothing or regularisation parameter for the posterior.

Note that the uniform prior is called the non-informative prior, and indeed we know from information theory is has the highest possible entropy. This does not mean that it is completely uninformative!

Effect of the Imaginary Sample Size

For example, with the non-informative prior, we can see how posterior probabilities $\tilde{\pi}_{ijk}$ get closer to $\tau_{ijk} = 1/RCL = 0.125$ with $RCL = 8$ if we increase N from 5 to 50.

```
> N = 5
> prop.table(table(lizards) + N/8)
, , Height = high

           Diameter
Species      narrow  wide
  Sagrei      0.209 0.086
  Distichus   0.178 0.171

, , Height = low

           Diameter
Species      narrow  wide
  Sagrei      0.079 0.028
  Distichus   0.149 0.101
```

```
> N = 50
> prop.table(table(lizards) + N/8)
, , Height = high

            Diameter
Species       narrow  wide
  Sagrei       0.201 0.090
  Distichus    0.173 0.166

, , Height = low

            Diameter
 Species       narrow  wide
   Sagrei       0.083 0.038
   Distichus    0.147 0.103
```

Shrinkage: the James-Stein Estimator

A shrinkage estimator $\tilde{\pi}_{ijk}$ is defined as the convex combination of the observed distribution and a target distribution τ_{ijk}, which in the case of contingency tables means

$$\tilde{\pi}_{ijk} = \lambda \tau_{ijk} + (1 - \lambda)\hat{\pi}_{ijk}, \qquad \lambda \in [0, 1] \tag{26}$$

as for Bayesian posterior estimator where

$$\hat{\lambda} = \frac{N}{N + n}. \tag{27}$$

A closed-form estimate for the shrinkage coefficient λ is

$$\hat{\lambda} = \frac{1 - \sum_{ijk} \hat{\pi}^2_{ijk}}{(n - 1) \sum_{ijk} (\tau_{ijk} - \hat{\pi}_{ijk})^2} \tag{28}$$

as derived in Hausser & Strimmer (JMLR 10:1469–1484, 2009) from James & Stein (1961) and Ledoit & Wolf (2003).

Shrinkage Estimators and Bayesian Posteriors

It is clear from the respective definitions that there is a one-to-one correspondence between shrinkage and posterior estimators:

- the target distribution plays the role of the prior;
- and the shrinkage coefficient is determined by the sample size and the imaginary sample size.

Both have a few properties in common:

- they provide regularised estimates for small samples;
- as $n \to \infty$ they converge to the maximum likelihood estimates, i.e. $\tilde{\pi}_{ijk} \to \hat{\pi}_{ijk}$;
- for small n they smooth estimated probabilities and provide non-zero estimated probabilities for cells with zero counts, i.e. $\tilde{\pi}_{ijk} > 0$.

The shrinkage estimator is an empirical Bayes estimator, whereas the posterior estimator is a full Bayesian estimator.

Tables with Structure: Some Special Cases

There are specific situations in which the structure of the data is best modelled by some special model tailored to the underlying sampling mechanism. An example is the Bradley-Terry model for pairwise comparisons that result in a preference for one level over the other. It is based on symmetric logit functions for each pair (i, j),

$$\log\left(\frac{\pi_{ij}}{\pi_{ji}}\right) = \beta_i - \beta_j, \tag{29}$$

and level i "wins" over level j if $\beta_i > \beta_j$. The estimated probability for this event is

$$\hat{\pi}_{ij} = \frac{e^{\beta_i}}{e^{\beta_i} + e^{\beta_j}} \tag{30}$$

with a confidence interval based on the covariance matrix of the maximum likelihood estimates through

$$\mathrm{VAR}(\hat{\beta}_i - \hat{\beta}_j) = \mathrm{VAR}(\hat{\beta}_i) + \mathrm{VAR}(\hat{\beta}_j) - 2\,\mathrm{COV}(\hat{\beta}_i, \hat{\beta}_j). \tag{31}$$

Hypothesis Testing

Common Hypotheses of Interest

A large part of the analysis of contingency table is testing different kinds of hypotheses, which involve several different tests and testing frameworks. When we have a variable of interest we can treat as a response, we use GLMs and deviance testing to compare the relevant nested models. Otherwise, we use different statistics to test the following hypotheses:

- whether two variables are marginally or conditionally independent;
- whether one ordinal variable show a trend (increasing or decreasing) as a function of a second ordinal variable;
- whether one or more categorical variables have the same distribution for all the levels of a separate set of variables (a homogeneity test);
- testing paired observations for a statistically significant difference between the two measures.

Parametric, Nonparametric and Semiparametric Tests

In order to determine a threshold α for the type I error, we need to provide a null distribution; there is more than way to do that for each test. Depending on how we do that, we classify tests as follows.

- Parametric tests: the full distribution is completely specified by the null hypothesis. They can be:
    - asymptotic tests (e.g. χ^2 log-likelihood ratio tests);
    - exact tests (e.g. F tests in linear models).
- Nonparametric tests: no distributional assumption is made, and an empirical null distribution is built using either bootstrap resampling or permutations.
- Semiparametric tests: the null distribution is specified up to one or more parameters, which are estimated from the empirical null distribution through bootstrap resampling or permutations.

Pros and Cons of Different Types of Tests

- Parametric tests can be biased when assumptions are violated or sample size is not large enough for the test statistic to converge to the asymptotic distribution. Permutation tests are always unbiased by construction, so they always reject the null hypothesis $\alpha \times 100\%$ of the time.
- Nonparametric tests are slower than parametric tests due to the need of generating the permutations or the bootstrap samples and to evaluate the test statistic on each of them.
- However, it is always possible to define a nonparametric test, even when a closed-form null distribution is not available or unfeasible to compute.
- Semiparametric tests are a compromise that requires much less resampling (typically $10\times$ less for the same precision) while still being reasonably robust.
- Nonparametric tests condition on the observed data set, whereas parametric tests are defined on the general population the sample is drawn from. This affects the interpretation of inference results.
- If a test statistic is consistent, the behaviour of all classes of tests is the same in the limit of the sample size.

Marginal Independence: Pearson's X^2 Test

The basic (unconditional) form of Pearson's X^2 statistic is

$$\mathrm{X}^2(X, Y) = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}, \quad \text{where} \quad \hat{\mu}_{ij} = \frac{n_{i+} n_{+j}}{n}. \tag{32}$$

It tests the independence hypothesis

$$H_0 : X \perp\!\!\!\perp_P Y \qquad \text{versus} \qquad H_1 : X \not\perp\!\!\!\perp_P Y \tag{33}$$

and as a parametric test it is asymptotically distributed under the null as a $\chi^2_{(R-1)(C-1)}$. The degrees of freedom are computed as the difference between the number of free parameters in the observed table $(R \times C - 1)$ and the number of free parameters under the null $(R - 1 + C - 1)$. From the definition, the sufficient statistic under the null are the marginal counts $\{n_{i+}\}$ and $\{n_{+j}\}$.

Degrees of Freedom and Sparse Contingency Tables

In sparse contingency tables, some of the n_{ij} may be zero, as well as some of the $\{n_{i+}\}$ and $\{n_{+j}\}$. Some n_{ij} may be zero because the underlying π_{ij} is small compared to the sample size and that configuration of variables has not been observed; we call this a sampling zero. On the other hand, it may be that $\pi_{ij} = 0$ so it is impossible to observe configuration of variables; we call the cell a structural zero and the contingency table an incomplete table.
In the general case, the adjusted degrees of freedom for the χ^2 are

$$\nu = (T_e - z_e) - (T_p - z_p) \tag{34}$$

where (from [DF 3.8]):

- T_e is the total number of cells;
- T_p is the number of parameters fitted;
- z_e is the number of cells with $\hat{\pi}_{ij} = 0$ (i.e. the sampling zeros);
- z_p is the number of parameters $\hat{\pi}_{ij}$ cannot be estimated (because either $\hat{\pi}_{i+} = 0$ or $\hat{\pi}_{+j} = 0$ or both, i.e. the structural zeros).

Other Options for Pearson's X^2

It is almost impossible to correctly adjust the degrees of freedom because empty cells are very easily counted more than once in z_e and z_p if they appear in patterns (which is most of the time); and in any case the convergence to the χ^2 distribution is problematic.
Continuity correction is only available for 2×2 tables in the form of Yates' correction,

$$\mathrm{X}^2(X, Y) = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{\left(|n_{ij} - \hat{\mu}_{ij}| - 1/2\right)^2}{\hat{\mu}_{ij}}. \tag{35}$$

This leaves:

- shrinkage tests, in which $\tilde{\pi}_{ij} \neq 0$ by construction (but are otherwise identical to the asymptotic χ^2 test);
- permutation tests, which does not use the χ^2 distribution;
- semiparametric tests, in which the degrees of freedom are computed from the data.

Pearson X^2 as a Permutation Test

Under H_0 the sufficient statistics for the multinomial model are the marginal counts $\{n_{i+}\}$ and $\{n_{+j}\}$ for X and Y, because their joint distribution is the product of the marginals of $\mathcal{T} = (X, Y)$. Thus, permutations of the data that result in the same $\{n_{i+}\}$ and $\{n_{+j}\}$ result in contingency tables $\mathcal{T}^* = (X^*, Y^*)$ that have

$$\mathrm{P}(\mathcal{T}^* \mid H_0) = \frac{\prod_i n_{i+}! \prod_j n_{+j}!}{n! \prod_{ij} \hat{\mu}_{ij}!} = \mathrm{P}(\mathcal{T} \mid H_0) \tag{36}$$

because their probability depends on the data only through the sufficient statistics (by definition). So $\mathcal{T}^*$ have a uniform probability distribution under H_0, which means we can generate $\mathcal{T}^*$ from H_0 to build a reference null distribution for the X^2 test statistic.

Note that inference is then conditional on the sample, not on the general population, because we condition on $\{n_{i+}\}$ and $\{n_{+j}\}$.

A Monte Carlo Implementation of Permutation X^2

A practical, Monte Carlo implementation of Pearson's X^2 as a permutation test then is as follows:

1. Compute the marginals $\{n_{i+}\}$ and $\{n_{+j}\}$ from $\mathcal{T} = (X, Y)$.
2. Compute the value of Pearson's X^2 for $\mathcal{T}$, i.e. $\mathrm{X}^2(\mathcal{T})$.
3. Generate a large enough number B of random contingency tables $\mathcal{T}^* = (X^*, Y^*)$ with fixed marginals $\{n_{i+}\}$ and $\{n_{+j}\}$ by permuting the data.
4. Estimate the empirical distribution of Pearson's X^2 under H_0 as $\{\mathrm{X}^2(\mathcal{T}^*_1), \ldots, \mathrm{X}^2(\mathcal{T}^*_B)\}$.
5. Compute the p-value for the test statistic as

$$\mathrm{P}(\mathrm{X}^2(\mathcal{T}^*) \geqslant \mathrm{X}^2(\mathcal{T})) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}(\mathrm{X}^2(\mathcal{T}^*_b) \geqslant \mathrm{X}^2(\mathcal{T})) \qquad (37)$$

using the right tail of the empirical distribution under H_0.

chisq.test(): Asymptotic and Permutation X^2

- Asymptotic χ^2 test, without Yates' correction.

```
> chisq.test(belt, correct = FALSE)

        Pearson's Chi-squared test

X-squared = 151.8729, df = 1, p-value < 2.2e-16
```

- Asymptotic χ^2 test, with Yates' correction.

```
> chisq.test(belt, correct = TRUE)

        Pearson's Chi-squared test with Yates' continuity correction

X-squared = 148.1748, df = 1, p-value < 2.2e-16
```

- Monte Carlo permutation test, with $B = 5000$ permutations.

```
> chisq.test(belt, simulate.p.value = TRUE, B = 5000)

        Pearson's Chi-squared test with simulated p-value (based on 5000
        replicates)

X-squared = 151.8729, df = NA, p-value = 2e-04
```

Packages for Hypothesis Testing: coin and bnlearn

Two other packages that implement permutation tests are bnlearn and coin; bnlearn implements all of parametric, semiparametric and nonparametric tests, coin just nonparametric tests. Both packages provide both marginal and conditional tests.

```
> library(vcdExtra)
> belt.df = expand.dft(belt)
> library(bnlearn)
> ci.test("Seatbelt", "Injury", data = belt.df, test = "x2")

        Pearson's X^2

data:  Seatbelt ~ Injury
x2 = 151.8729, df = 1, p-value < 2.2e-16
alternative hypothesis: true value is greater than 0
> ci.test("Seatbelt", "Injury", data = belt.df, test = "mc-x2")

        Pearson's X^2 (MC)

data:  Seatbelt ~ Injury
mc-x2 = 151.8729, Monte Carlo samples = 5000, p-value < 2.2e-16
alternative hypothesis: true value is greater than 0
```

From Marginal to Conditional Independence Tests

In contingency tables with more than two dimensions, we may also want to test the more general hypothesis of conditional independence:

$$H_0 : X \perp\!\!\!\perp_P Y \mid \mathbf{Z} \qquad \text{versus} \qquad H_1 : X \not\perp\!\!\!\perp_P Y \mid \mathbf{Z} \tag{38}$$

where $\mathbf{Z}$ is a set of variables that does not include either X or Y. If $\mathbf{Z} = \{\varnothing\}$, we are back testing marginal independence.

Conditional independence tests, when all conditioning variables are discrete, are constructed from marginal tests by conditioning on all the configurations of the variables in $\mathbf{Z}$:

$$\mathrm{P}(X, Y \mid \mathbf{Z}) = \sum\nolimits_{\mathbf{z}} \mathrm{P}(X, Y \mid \mathbf{Z} = \mathbf{z}) \tag{39}$$

This works due to the law of total probability, and reduces a conditional test to a collection of marginal tests defined on the two-dimensional contingency tables corresponding to the $\mathbf{z}$. This also means that in practice all conditional tests are effectively in 3 dimensions, because all the variables in $\mathbf{Z}$ are collapsed in a single discrete variable whose levels are the configurations $\mathbf{z}$.

Null Distribution of Conditional Independence Tests

The null distribution for a conditional (parametric or semiparametric) test also stems from the law of total probability. The configurations $\mathbf{z}$ of $\mathbf{Z}$ define a partition of the probability space, so

$$(X, Y \mid \mathbf{Z} = \mathbf{z}_i) \perp\!\!\!\perp_P (X, Y \mid \mathbf{Z} = \mathbf{z}_j) \qquad \text{for } i \neq j. \tag{40}$$

For each configuration, we know that the marginal test for the corresponding sub-table has (say) a $\chi^2_{(R-1)(C-1)}$ assuming we do not have problems with zero cell counts. Then, the distribution of the conditional test is the sum of the i.i.d distributions of the marginal tests. If there are L configurations of $\mathbf{Z}$, then

$$\sum\nolimits_{\mathbf{z}} \chi^2_{(R-1)(C-1)} \sim \chi^2_{(R-1)(C-1)L}. \tag{41}$$

Conditional Pearson's X^2 Test

The conditional version of Pearson's X^2 test therefore is

$$\mathrm{X}^2(X, Y \mid \mathbf{Z}) = \sum_{k=1}^{L} \mathrm{X}^2(X, Y \mid \mathbf{Z} = \mathbf{z}_k) =$$

$$= \sum_{k=1}^{L} \left[\sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(n_{ijk} - \hat{\mu}_{ijk})^2}{\hat{\mu}_{ijk}} \right], \text{ where } \hat{\mu}_{ijk} = \frac{n_{i+k} n_{+jk}}{n_{++k}}. \quad (42)$$

and under the null has the asymptotic distribution

$$\mathrm{X}^2(X, Y \mid \mathbf{Z}) \sim \chi^2_{(R-1)(C-1)L}. \quad (43)$$

Note that once we condition on $\mathbf{Z}$, its value is assumed to be known; hence it is irrelevant whether the variables in $\mathbf{Z}$ are categorical or ordinal.

Conditional Pearson's X^2 in bnlearn

The syntax is the same as before, but in addition to "Diameter" (x argument) and "Height" (y argument) we also specify the conditioning variable(s) "Species" (z argument).

```
> ci.test("Diameter", "Height", "Species", data = lizards, test = "x2")

        Pearson's X^2

data:  Diameter ~ Height | Species
x2 = 2.0256, df = 2, p-value = 0.3632
alternative hypothesis: true value is greater than 0
```

As an alternative, we can perform the corresponding permutation test.

```
> ci.test("Diameter", "Height", "Species", data = lizards, test = "mc-x2")

        Pearson's X^2 (MC)

data:  Diameter ~ Height | Species
mc-x2 = 2.0174, Monte Carlo samples = 5000, p-value = 0.3722
alternative hypothesis: true value is greater than 0
```

How Are Permutations Done in Conditional Tests?

In the presence of a set of conditioning variables $\mathbf{Z}$, the conditional test is constructed as a collection of marginal tests. As a result, the sufficient statistics under the null hypothesis are the sufficient statistics for each of the sub-tables the marginal test statistics are computed on.

Therefore to permute the data and obtain the empirical null distribution:

1. We fix the marginal counts $\{n_{i+k}\}$ and $\{n_{+jk}\}$ (and thus the n_{++k} subtotal) for all the L configurations.
2. For each configuration in turn, we permute the corresponding sub-table to get $\mathcal{T}^*_{b(k)}$, $b = 1, \ldots, B$ and $k = 1, \ldots, L$;
3. We construct the overall permuted table as $\mathcal{T}^*_b = \{\mathcal{T}^*_{b(k)}\}$.
4. We compute $\mathrm{X}^2(\mathcal{T}^*_b)$ a large number B of times to obtain the empirical null distribution.

The G^2 Test

Another independence test is the G^2 test, which a log-likelihood ratio test (in statistics) and a mutual information test (in computer science and information theory). The marginal test is

$$\mathrm{G}^2(X,Y) = -2n \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{n_{ij}}{n} \log \frac{n_{i+}n_{+j}}{n_{ij}n} = -2n \sum_{i=1}^{R} \sum_{j=1}^{C} \pi_{ij} \log \frac{\pi_{i+}\pi_{+j}}{\pi_{ij}} \quad (44)$$

with distribution $\chi^2_{(R-1)(C-1)}$; and the conditional test is

$$\mathrm{G}^2(X,Y \mid \mathbf{Z}) = -2n \sum_{k=1}^{L} \left[\sum_{i=1}^{R} \sum_{j=1}^{C} \frac{n_{ijk}}{n} \log \frac{n_{ijk}n_{++k}}{n_{i+k}n_{+jk}} \right] \quad (45)$$

with distribution $\chi^2_{(R-1)(C-1)L}$. Note that as $n \to \infty$, $\mid \mathrm{X}^2 - \mathrm{G}^2 \mid \to 0$ in probability. X^2 converges to the asymptotic χ^2 distribution more quickly.

The nonparametric G^2 test is computed in the same way as for the X^2 test.

The Relationship Between the X^2 and G^2 Tests

X^2 and G^2 can approximate each other using the fact that

$$\log x \leqslant x - 1, \qquad x > 0$$

so we can bound the logarithm as follows:

$$\frac{a-b}{a} \leqslant \log\left(\frac{a}{b}\right) \leqslant \frac{a-b}{b}.$$

where the equality hold if and only if $a = b$. Under the null this is the case for G^2 so we can approximate the logarithm with the mean of its bounds:

$$\log\left(\frac{a}{b}\right) \simeq \frac{1}{2}\left(\frac{a-b}{a} + \frac{a-b}{b}\right) = \frac{a^2-b^2}{2ab}$$

which means that

$$\mathrm{G}^2(X,Y) \propto -\sum_{i=1}^{R}\sum_{j=1}^{L} \pi_{ij} \log\frac{\pi_i\pi_j}{\pi_{ij}} \simeq -\sum_{i=1}^{R}\sum_{j=1}^{L} \pi_{ij} \frac{\pi_{ij}^2 - (\pi_i\pi_j)^2}{2\pi_{ij}\pi_i\pi_j}$$

$$\simeq \sum_{i=1}^{R}\sum_{j=1}^{L} \frac{(\pi_{ij} - \pi_i\pi_j)^2}{\pi_i\pi_j} = \mathrm{X}^2(X,Y).$$

A Compromise: Semiparametric Tests

The semiparametric versions of G^2 and Pearson's X^2 uses the asymptotic $\chi^2_{(R-1)(C-1)L}$ distribution but estimates the degrees of freedom from the data as

$$df = \frac{1}{B}\sum_{b=1}^{B} \mathrm{X}^2(\mathcal{T}_b^*) \qquad \text{or} \qquad df = \frac{1}{B}\sum_{b=1}^{B} \mathrm{G}^2(\mathcal{T}_b^*) \tag{46}$$

because the degrees of freedom are the expectation of the χ^2 distribution and therefore can be approximated by the mean of the test statistics obtained from the permutations.

This is a much easier estimation problem than that of a nonparametric test, because we are computing a point estimate of the mean instead of an empirical estimate of the whole distribution. Fewer permutations are required, and the degrees of freedom are self-adjusting in the presence of zero cell counts.

All G^2 Tests

```
> ci.test("Diameter", "Height", "Species", data = lizards, test = "mi")

        Mutual Information (disc.)

data:  Diameter ~ Height | Species
mi = 2.0256, df = 2, p-value = 0.3632
alternative hypothesis: true value is greater than 0

> ci.test("Diameter", "Height", "Species", data = lizards, test = "sp-mi")

        Mutual Information (disc., semipar.)

data:  Diameter ~ Height | Species
sp-mi = 2.0256, df = 1.974, Monte Carlo samples = 100, p-value = 0.3576
alternative hypothesis: true value is greater than 0

> ci.test("Diameter", "Height", "Species", data = lizards, test = "mc-mi")

        Mutual Information (disc., MC)

data:  Diameter ~ Height | Species
mc-mi = 2.0256, Monte Carlo samples = 5000, p-value = 0.3666
alternative hypothesis: true value is greater than 0
```

Fisher's (Exact) Test (Marginal Only)

Fisher's exact test is the probability of getting a particular $\mathcal{T}$ under H_0, that is, the $\mathrm{P}(\mathcal{T} \mid H_0)$ we used in (36) in defining the permutation tests:

$$\mathrm{P}(\mathcal{T}) = \frac{\prod_i n_{i+}! \prod_j n_{+j}!}{n! \prod_{ij} n_{ij}!} \tag{47}$$

which is an extension of the hypergeometric distribution. So the p-value of the test is the proportion of contingency tables $\mathcal{T}^*$ such that $\mathrm{P}(\mathcal{T}) < \mathrm{P}(\mathcal{T}^*)$.

This test would be the best possible test except:

- it is not computationally feasible to use it on tables with very large R or C because there are too many possible tables to enumerate;
- as a conditional test, it is not computationally feasible even for moderate R and C because L increases very quickly with the number of conditioning variables.

In practice it is most often computed as a permutation test or using its asymptotic distribution $\chi^2_{(R-1)(C-1)L}$.

Fisher's test: fisher.test()

```
> fisher.test(belt)

        Fisher's Exact Test for Count Data

data:  belt
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  6.623513 18.173941
sample estimates:
odds ratio
  10.83069
> fisher.test(job)
Error in fisher.test(job) : FEXACT error 501. [...] The algorithm cannot
proceed. Reduce the workspace size or use another algorithm.
> fisher.test(job, simulate.p.value = TRUE, B = 5000)

        Fisher's Exact Test for Count Data with simulated p-value (based
        on 5000 replicates)

data:  job
p-value = 0.2326
alternative hypothesis: two.sided
```

Testing Ordinal Variables for Trend: Jonckheere-Terpstra

The test statistic is defined for X and Y ordinal as

$$\mathrm{JT}(X, Y \mid \mathbf{Z}) = \sum_{k=1}^{L} \sum_{i=2}^{R} \sum_{j=1}^{i-1} \left[\sum_{s=1}^{C} w_{ijsk} n_{isk} - \frac{n_{i+k}(n_{i+k}+1)}{2} \right] \qquad (48)$$

where the w_{ijsk} are Wilcoxon scores, defined as

$$w_{ijsk} = \sum_{t=1}^{s-1} \left[n_{itk} + n_{jtk} + \frac{n_{isk} + n_{jsk} + 1}{2} \right], \qquad (49)$$

and has an asymptotic normal distribution with mean and variance defined in [DE 5]. The null hypothesis is that of homogeneity; if we denote with $F_{i,k}(y) = \mathrm{P}(Y < y \mid X = i, \mathbf{Z} = k)$, then

$$H_0 : F_{1,k}(y) = F_{2,k}(y) = \ldots = F_{T,k}(y) \qquad \text{for } \forall y \text{ and } \forall k. \qquad (50)$$

The alternative hypothesis $H_1 = H_{1,1} \cup H_{1,2}$ is that of stochastic ordering, either increasing or decreasing:

$$H_{1,1} : F_{i,k}(y) \geqslant F_{j,k}(y) \qquad \text{with } i < j \text{ for } \forall y \text{ and } \forall k \qquad (51)$$

$$H_{1,2} : F_{i,k}(y) \leqslant F_{j,k}(y) \qquad \text{with } i < j \text{ for } \forall y \text{ and } \forall k. \qquad (52)$$

ci.test() and Jonckheere-Terpstra

ci.test() in bnlearn provides an implementation of this test, the other being in the ciTest_ordinal() in the gRim package.

```
> ci.test(job, test = "jt")

        Jonckheere-Terpstra

data:  Income ~ Satisfaction
jt = 3.053, p-value = 0.002266
alternative hypothesis: true value is not equal to 0
```

Remember that conditioning variables can be either ordinal or categorical (or a mixture of the two): the construction of the conditional test is exactly the same, as well as the null and the alternative distribution.

The test is two-tailed, with the left tail corresponding to a decreasing trend and the right tail corresponding to an increasing trend.

McNemar's Test for Paired Variables

McNemar's test is a statistical test used on paired categorical data (for paired ordinal data, different tests for rank agreement are used) in which each variables has 2 levels. The null hypothesis is marginal homogeneity, that is, that the marginal distributions of X and Y are the same:

$$H_0 : \pi_{1+} = \pi_{+1} \qquad \text{versus} \qquad H_1 : \pi_{1+} \neq \pi_{+1}. \tag{53}$$

This is the same as testing the off-diagonal elements, because

$$\pi_{1+} - \pi_{+1} = \pi_{11} + \pi_{12} - \pi_{11} - \pi_{21} = \pi_{12} - \pi_{21}. \tag{54}$$

which means that we accept the null hypothesis when there is a low number of discordant cells (i.e. cells not on the diagonal of the table). The test statistic is

$$\mathrm{MN}(X, Y) = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \sim \chi^2_1. \tag{55}$$

`mcnemar.test()` and `mh_test()`

```
> mcnemar.test(belt, correct = TRUE)

        McNemar's Chi-squared test with continuity correction

data:  belt
McNemar's chi-squared = 10248.25, df = 1, p-value < 2.2e-16
> mcnemar.test(belt, correct = FALSE)

        McNemar's Chi-squared test

data:  belt
McNemar's chi-squared = 10250.24, df = 1, p-value < 2.2e-16
```

`coin` has an implementation of McNemar's test as a permutation test in the `mh_test()` function, which can also perform conditional tests.

Graphical Models

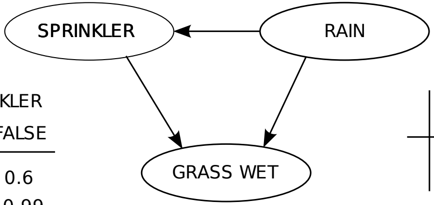
Graphical Models

Graphical models are defined by:

- a network structure, $\mathcal{G} = (\mathbf{V}, E)$, either an undirected graph (Markov networks, gene association networks, correlation networks, etc.) or a directed graph (Bayesian networks). Each node $v_i \in \mathbf{V}$ corresponds to a random variable X_i;
- a global probability distribution, $\mathbf{X}$, which can be factorised into a small set of local probability distributions according to the edges $e_{ij} \in E$ present in the graph.

This combination allows a compact representation of the joint distribution of large numbers of random variables and simplifies inference on the resulting parameter space.

A Simple Bayesian Network: Watson's Lawn



RAIN	SPRINKLER	
	TRUE	FALSE
FALSE	0.4	0.6
TRUE	0.01	0.99

RAIN	
TRUE	FALSE
0.2	0.8

SPRINKLER	RAIN	GRASS WET	
		TRUE	FALSE
FALSE	FALSE	0.0	1.0
FALSE	TRUE	0.8	0.2
TRUE	FALSE	0.9	0.1
TRUE	TRUE	0.99	0.01

Graphical Separation and Independence

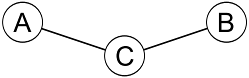
The main role of the graph structure is to express the conditional independence relationships among the variables in the model, thus specifying the factorisation of the global distribution. Different classes of graphs express these relationships with different semantics, which have in common the principle that graphical separation of two (sets of) nodes implies the conditional independence of the corresponding (sets of) random variables.

For networks considered here, separation is defined as:

- (u-)separation in Markov networks;
- d-separation in Bayesian networks.

Graphical Separation

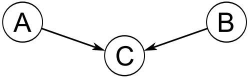
separation (undirected graphs)



$$\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$$

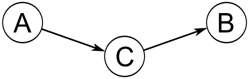
$$\mathrm{P}(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \mathrm{P}(\mathbf{A} \mid \mathbf{C})\, \mathrm{P}(\mathbf{B} \mid \mathbf{C})\, \mathrm{P}(\mathbf{C})$$

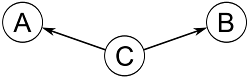
d-separation (directed acyclic graphs)



$$\mathbf{A} \not\!\perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$$

$$\mathrm{P}(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \mathrm{P}(\mathbf{C} \mid \mathbf{A}, \mathbf{B})\, \mathrm{P}(\mathbf{A})\, \mathrm{P}(\mathbf{B})$$





$$\mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$$

$$\begin{aligned} &\mathrm{P}(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \\ &\quad = \mathrm{P}(\mathbf{B} \mid \mathbf{C})\, \mathrm{P}(\mathbf{C} \mid \mathbf{A})\, \mathrm{P}(\mathbf{A}) \\ &\quad = \mathrm{P}(\mathbf{A} \mid \mathbf{C})\, \mathrm{P}(\mathbf{B} \mid \mathbf{C})\, \mathrm{P}(\mathbf{C}) \end{aligned}$$

Maps and Independence

A graph $\mathcal{G}$ is a dependency map (or D-map) of the probabilistic dependence structure P of $\mathbf{X}$ if there is a one-to-one correspondence between the random variables in $\mathbf{X}$ and the nodes $\mathbf{V}$ of $\mathcal{G}$, such that for all disjoint subsets $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ of $\mathbf{X}$ we have

$$\mathbf{A} \perp\!\!\!\perp_P \mathbf{B} \mid \mathbf{C} \Longrightarrow \mathbf{A} \perp\!\!\!\perp_G \mathbf{B} \mid \mathbf{C}. \tag{56}$$

Similarly, $\mathcal{G}$ is an independency map (or I-map) of P if

$$\mathbf{A} \perp\!\!\!\perp_P \mathbf{B} \mid \mathbf{C} \Longleftarrow \mathbf{A} \perp\!\!\!\perp_G \mathbf{B} \mid \mathbf{C}. \tag{57}$$

$\mathcal{G}$ is said to be a perfect map of P if it is both a D-map and an I-map, that is

$$\mathbf{A} \perp\!\!\!\perp_P \mathbf{B} \mid \mathbf{C} \Longleftrightarrow \mathbf{A} \perp\!\!\!\perp_G \mathbf{B} \mid \mathbf{C}, \tag{58}$$

and in this case P is said to be isomorphic to $\mathcal{G}$.

Graphical models are formally defined as I-maps under the respective definitions of graphical separation, but sometimes we assume they are perfect maps for particular algorithms in model estimation and inference.

Factorisation into Local Distributions

The most important consequence of defining graphical models as I-maps is the factorisation of the global distribution into local distributions:

- in Markov networks, local distributions are associated with the cliques $\mathbf{C}_i$ (maximal subsets of nodes in which each element is adjacent to all the others) in the graph,

$$\mathrm{P}(\mathbf{X}) = \prod_{i=1}^{k} \psi_i(\mathbf{C}_i), \tag{59}$$

  and the ψ_k functions are called potentials.
- in Bayesian networks, each local distribution is associated with a single node X_i and depends only on the joint distribution of its parents Π_{X_i}:

$$\mathrm{P}(\mathbf{X}) = \prod_{i=1}^{p} \mathrm{P}(X_i \mid \Pi_{X_i}). \tag{60}$$

A Note About Potentials

Potentials are non-negative functions representing the relative mass of probability of each clique C_i. They are proper probability or density functions only when the graph is decomposable or triangulated, that is when it contains no induced cycles other than triangles. With any other type of graph inference becomes very hard, if possible at all, because $\psi_1, \psi_2, \ldots, \psi_k$ have no direct statistical interpretation.

In this case the global distribution factorises again according to the chain rule and can be written as

$$\mathrm{P}(\mathbf{X}) = \frac{\prod_{i=1}^{k} \mathrm{P}(\mathbf{C}_i)}{\prod_{i=1}^{k} \mathrm{P}(\mathbf{S}_i)} \tag{61}$$

where $\mathbf{S}_i$ are the nodes of $\mathbf{C}_i$ which are also part of any other clique up to $\mathbf{C}_{i-1}$.

Neighbourhoods and Markov Blankets

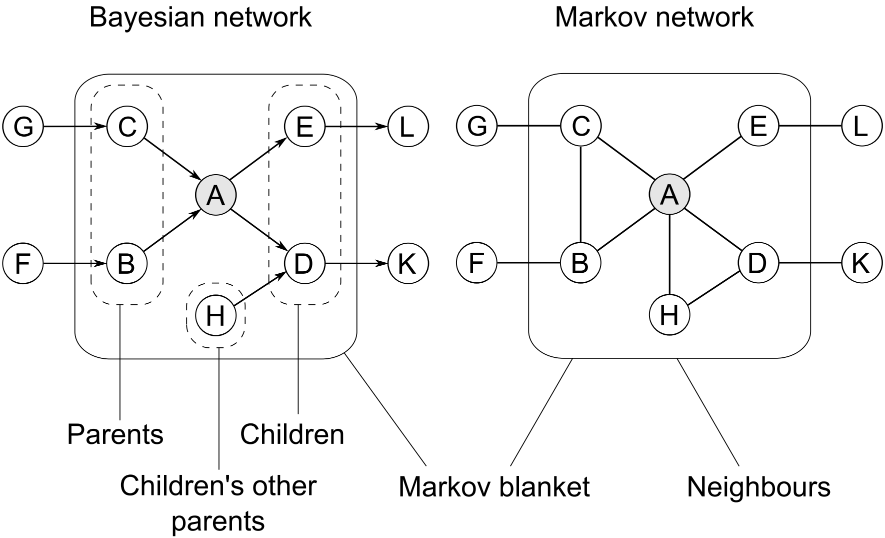
Furthermore, for each node X_i two sets are defined:

- the neighbourhood, the set of nodes that are adjacent to X_i. These nodes cannot be made independent from X_i.

- the Markov blanket, the set of nodes that completely separates X_i from the rest of the graph. Generally speaking, it is the set of nodes that includes all the knowledge needed to do inference on X_i, from estimation to hypothesis testing to prediction, because all the other nodes are conditionally independent from X_i given its Markov blanket.

These sets are related in Markov and Bayesian networks; in particular, Markov blankets can be shown to be the same using a moral graph.

Neighbourhoods and Markov Blankets



Markov networks vs Bayesian networks

Markov networks and Bayesian networks do not appear to be closely related, as they are so different in construction and interpretation.

- There are indeed dependency models that have an undirected perfect map but not a directed acyclic one, and vice versa.
- However, it can be shown that every dependency structure that can be expressed by a decomposable graph can be modelled both by a Markov network and a Bayesian network.
- It can also be shown that every dependency model expressible by an undirected graph is also expressible by a directed acyclic graph, with the addition of some auxiliary nodes.

These two results indicate that there is a significant overlap between Markov and Bayesian networks, and that in many cases both can be used to the same effect.

Probability Distributions: Discrete and Continuous

Data used in graphical modelling should respect the following assumptions:

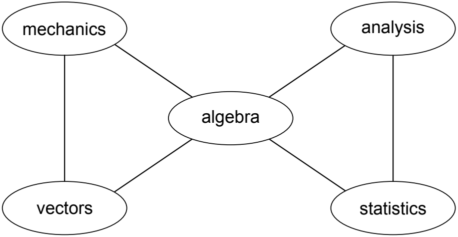
- if all the variables X_i are discrete, both the global and the local distributions are assumed to be multinomial. Local distributions are described using conditional probability tables;
- if all the variables X_i are continuous, the global distribution is assumed to be a multivariate Gaussian distribution, and the local distributions are univariate or multivariate Gaussian distributions. Local distributions are described using partial correlation coefficients;
- if both continuous and discrete variables are present, we can assume a mixture or conditional Gaussian distribution, discretise continuous attributes or use a nonparametric approach.

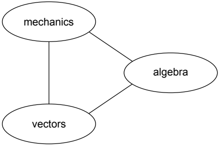
Other Distributional Assumptions

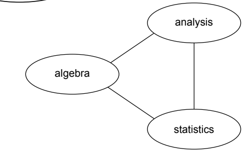
Other fundamental distributional assumptions are:

- observations must be independent. If some form of temporal or spatial dependence is present, it must be specifically accounted for in the definition of the network (as in *dynamic Bayesian networks*);
- if the model will be used as a causal graphical model, that is, to infer cause-effect relationship from experimental or (more frequently) observational data, there must be no latent or hidden variables that influence the dependence structure of the model;
- all the relationships between the variables in the network must be conditional independencies, because they are by definition the only ones that can be expressed by graphical models.

A Gaussian Markov Network (MARKS)



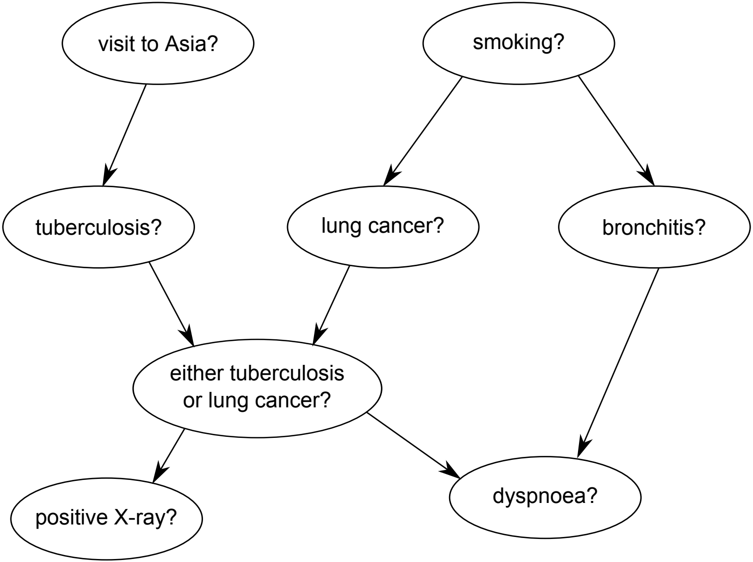




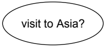
$$\begin{bmatrix} 1.000 & 0.332 & 0.235 \\ 0.332 & 1.000 & 0.327 \\ 0.235 & 0.327 & 1.000 \end{bmatrix}$$

$$\begin{bmatrix} 1.000 & 0.451 & 0.364 \\ 0.451 & 1.000 & 0.256 \\ 0.364 & 0.256 & 1.000 \end{bmatrix}$$

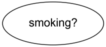
A Discrete Bayesian Network (ASIA)



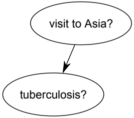
A Discrete Bayesian Network (ASIA)



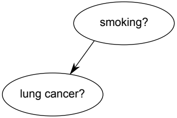
yes	no
0.01	0.99



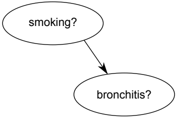
yes	no
0.50	0.50



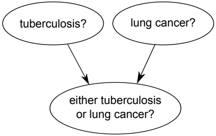
	yes	no
yes	0.05	0.95
no	0.01	0.99



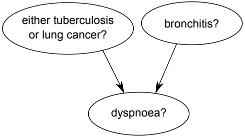
	yes	no
yes	0.10	0.90
no	0.01	0.99



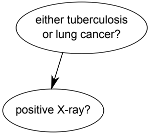
	yes	no
yes	0.60	0.40
no	0.30	0.70



	yes	no
yes:yes	1	0
yes:no	1	0
no:yes	1	0
no:no	0	1



	yes	no
yes:yes	0.90	0.10
yes:no	0.70	0.30
no:yes	0.80	0.20
no:no	0.10	0.90



	yes	no
yes	0.98	0.02
no	0.05	0.95

Learning a Graphical Model

Model selection and estimation are collectively known as learning, and are usually performed as a two-step process:

1. structure learning, learning the graph structure from the data.
2. parameter learning, learning the local distributions implied by the graph structure learned in the previous step.

This work-flow is implicitly Bayesian; given a data set $\mathcal{D}$ and if we denote the parameters of the global distribution as $\mathbf{X}$ with Θ, we have

$$\underbrace{\mathrm{P}(\mathcal{M} \mid \mathcal{D})}_{learning} = \underbrace{\mathrm{P}(\mathcal{G} \mid \mathcal{D})}_{structure\,learning} \cdot \underbrace{\mathrm{P}(\Theta \mid \mathcal{G}, \mathcal{D})}_{parameter\,learning} \tag{62}$$

and structure learning is done in practise as

$$\mathrm{P}(\mathcal{G} \mid \mathcal{D}) \propto \mathrm{P}(\mathcal{G})\,\mathrm{P}(\mathcal{D} \mid \mathcal{G}) = \mathrm{P}(\mathcal{G}) \int \mathrm{P}(\mathcal{D} \mid \mathcal{G}, \Theta)\,\mathrm{P}(\Theta \mid \mathcal{G}) d\Theta. \tag{63}$$

Local Distributions: Divide and Conquer

Most tasks related to both learning and inference are NP-hard (they cannot be solved in polynomial time in the number of variables). They are still feasible thanks to the decomposition of $\mathbf{X}$ into the local distributions; under some assumptions (parameter independence) there is never the need to manipulate more than one of them at a time.

In Bayesian networks, for example, structure learning boils down to

$$\mathrm{P}(\mathcal{D} \mid \mathcal{G}) = \int \prod \left[\mathrm{P}(X_i \mid \Pi_{X_i}, \Theta_{X_i})\,\mathrm{P}(\Theta_{X_i} \mid \Pi_{X_i})\right] d\Theta \tag{64}$$

$$= \prod \left[\int \mathrm{P}(X_i \mid \Pi_{X_i}, \Theta_{X_i})\,\mathrm{P}(\Theta_{X_i} \mid \Pi_{X_i}) d\Theta_{X_i}\right] \tag{65}$$

and parameter learning boils down to

$$\mathrm{P}(\Theta \mid \mathcal{G}, \mathcal{D}) = \prod \mathrm{P}(\Theta_{X_i} \mid \Pi_{X_i}, \mathcal{D}). \tag{66}$$

Constraint-, Score-based and Hybrid Structure Learning

Despite the (sometimes confusing) variety of theoretical backgrounds and terminology they can all be traced to only three approaches:

- constraint-based algorithms: they use statistical tests to learn conditional independence relationships (called *constraints* in this setting) from the data and assume that the graph underlying the probability distribution is a perfect map to determine the correct network structure.
- score-based algorithms: each candidate network is assigned a score reflecting its goodness of fit, which is then taken as an objective function to maximise.
- hybrid algorithms: conditional independence tests are used to learn at least part of the conditional independence relationships from the data, thus restricting the search space for a subsequent score-based search. The latter determines which edges are actually present in the graph and, in the case of Bayesian networks, their direction.

Structure Learning, Contingency Tables and GLMs

The lack of arc between two variables X and Y in a graphical model implies they are independent given other variables $\mathbf{Z}$. Any such statement can be rephrased as:

- A nested GLM model testing in which Y is the response and

  $$H_0 : \beta_X = 0 \qquad \text{versus} \qquad H_1 : \beta_X \neq 0. \tag{67}$$

  If H_0 is rejected we assume the graph is a perfect map and we include an arc between X and Y; or we remove it instead if we accept H_0. Or we can use AIC/BIC to see which model fits the data best.

- A conditional independence test non explicitly associated with a regression model, such as Pearson's X^2.

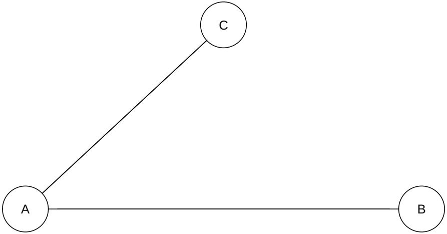
The difference between the two interpretations is often blurred (e.g. model selection of any kind is equivalent to testing partial correlations, Pearson's X^2 is the sum of Pearson's residuals in a Poisson GLM model, etc.).

Log-Linear GLMs Represented as Markov Networks

The equivalence between GLMs and graphical models goes beyond 2nd-order interactions (which are represented by a single arc): higher interactions terms are expressed by groups of arcs, and thus by groups of interaction terms. If we assume Poisson sampling and we build the corresponding log-linear model, say

$$\log(\pi_{ijk}) = \mu + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} \tag{68}$$

in which the u terms represents the contrasts for each variable and interaction and the respective regression coefficients. The corresponding graph is:

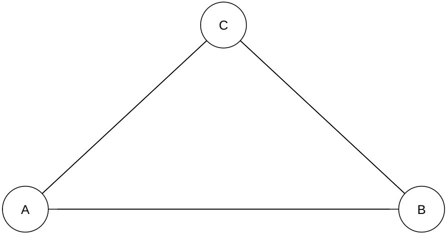


Log-Linear GLMs Represented as Markov Networks

The saturated model would then be

$$\log(\pi_{ijk}) = \mu + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC} + u_{ijk}^{ABC} \qquad (69)$$

which has graph:



Removing the arc between B and C implies $u_{jk}^{BC} = 0$, and in turn we set $u_{ijk}^{ABC} = 0$ to keep the hierarchical structure of the interactions. Thus we obtain the model in the previous slide.

Bayesian Networks as Hierarchical Regressions

In Bayesian networks, each node is modelled with a univariate distribution which is conditional on its parents. Therefore, it is natural to use the graph to represent a hierarchical regression model.

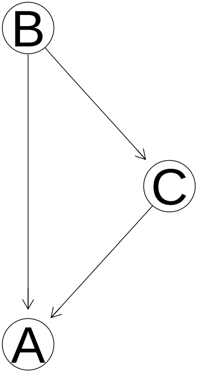
So, for the graph on the right, we have the following GLMs for the nodes.

$$g(A) = \mu_A + u_B + u_C \tag{70}$$

$$g(B) = \mu_B \tag{71}$$

$$g(C) = \mu_C + u_B \tag{72}$$

In this representation each node has its own independent error term, making the regressions independent of each other.



Likelihood, Bayesian and Shrinkage Parameter Learning

Once the structure of the model is known, the problem of estimating the parameters of the global distribution can be solved by estimating the parameters of the local distributions, one at a time.

Three common choices are:

- maximum likelihood estimators: just the usual empirical estimators. Often described as either maximum entropy or minimum divergence estimators in information-theoretic literature.
- Bayesian posterior estimators: posterior estimators, based on conjugate priors to keep computations fast, simple and in closed form.
- shrinkage estimators: regularised estimators based either on James-Stein or Bayesian shrinkage results, or regularised regression models.

Inference on Graphical Models

Inference on Bayesian networks usually consists of conditional probability (CPQ) or maximum a posteriori (MAP) queries.

Conditional probability queries are concerned with the distribution of a subset of variables $\mathbf{Q} = \{X_{j_1}, \ldots, X_{j_l}\}$ given some evidence $\mathbf{E}$ on another set $X_{i_1}, \ldots, X_{i_k}$ of variables in $\mathbf{X}$:

$$CPQ(\mathbf{Q} \mid \mathbf{E}, \mathcal{M}) = \mathrm{P}(\mathbf{Q} \mid \mathbf{E}, \mathcal{G}, \Theta) = \mathrm{P}(X_{j_1}, \ldots, X_{j_l} \mid \mathbf{E}, \mathcal{G}, \Theta). \quad (73)$$

Maximum a posteriori queries are concerned with finding the configuration $\mathbf{q}^*$ of the variables in $\mathbf{Q}$ that has the highest posterior probability:

$$MAP(\mathbf{Q} \mid \mathbf{E}, \mathcal{M}) = \mathbf{q}^* = \operatorname*{argmax}_{\mathbf{q}} \mathrm{P}(\mathbf{Q} = \mathbf{q} \mid \mathbf{E}, \mathcal{G}, \Theta). \quad (74)$$

Both can be computed exactly in a few cases (discrete models, mostly), but are commonly computed using importance sampling or MCMC.

That's It!