SC6/SM9 Graphical Models

Michaelmas Term, 2016

Robin Evans

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

November 24, 2016

Course Website

The class site is at

`http://www.stats.ox.ac.uk/~evans/gms/`

You'll find

- lecture notes;
- slides;
- problem sheets;
- data sets.

Course Information

There will be four problem sheets and four associated classes.

Part C students, your classes are weeks 3, 5, 7 and HT0. Signal your preference on the sign-up sheet for:

- Thursdays 3pm–4:30pm;
- Fridays 1pm–2:30pm.

Hand in work by Tuesday, 5pm.

MSc students, class times are:

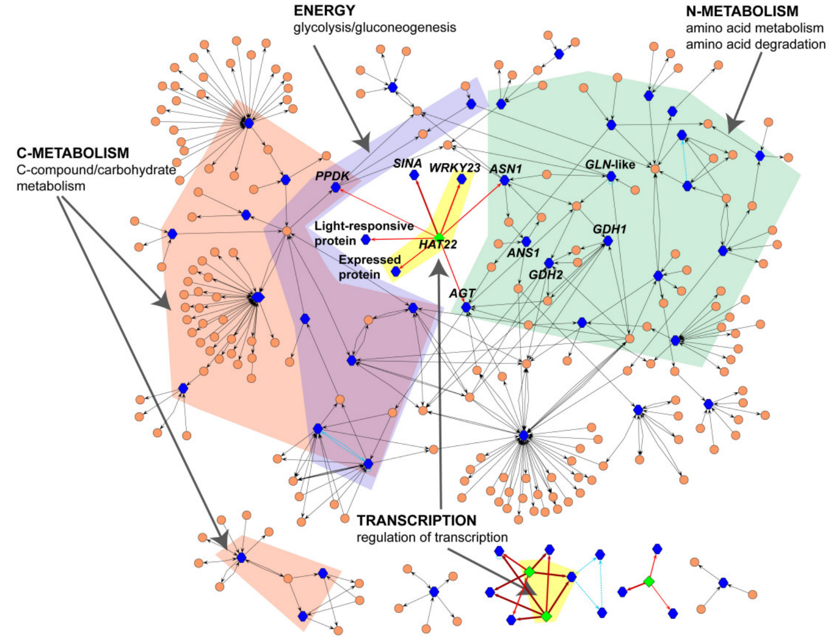
Sheet	Day	Time	Place
1	Wednesday Week 4	11:00	LG.01
2	Tuesday Week 5	12:00	LG.01
3	Tuesday Week 7	12:00	LG.01
4	Thursday Week 8	11:00	LG.01

Books

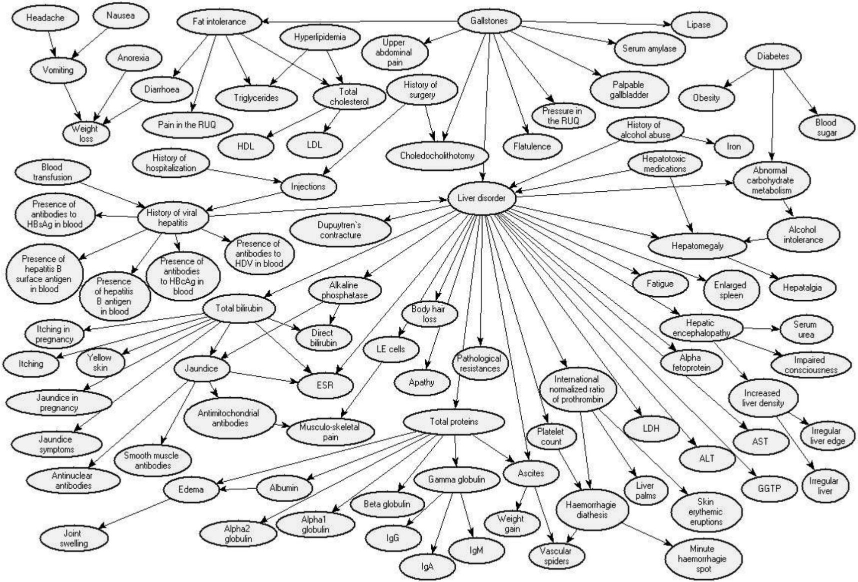
These books might be useful.

- Lauritzen (1996). *Graphical Models*, OUP.
- Wainwright and Jordan (2008). *Graphical Models, Exponential Families, and Variational Inference*. (Available online).
- Pearl (2009). *Causality*, (3rd edition), Cambridge.
- Koller and Friedman (2009), *Probabilistic Graphical Models: Principles and Techniques*, MIT Press.

Gene Regulatory Networks



Medical Diagnosis



Main Issues

There are two main problems with large data sets that we will consider in this course:

- statistical;
  we need to predict outcomes from scenarios that have never been observed (i.e., we need a model).
- computational:
  - we can't store probabilities for all combinations of variables;
  - even if we could, we can't sum/integrate them to find a marginal or conditional probability:

$$P(X = x) = \sum_{\boldsymbol{y}} P(X = x, \boldsymbol{Y} = \boldsymbol{y}).$$

The solution is to impose nonparametric structure, in the form of conditional independences.

Conditional Independence

Simpson's Paradox

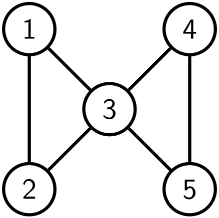
Death Penalty?	Defendant's Race	
	White	Black
Yes	53	15
No	430	176

Simpson's Paradox

Victim's Race	Death Penalty?	Defendant's Race: White	Defendant's Race: Black
White	Yes	53	11
	No	414	37
Black	Yes	0	4
	No	16	139

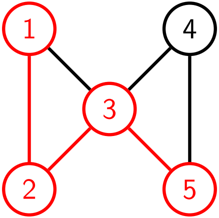
Undirected Graphical Models

Undirected Graphs



$$V = \{1, 2, 3, 4, 5\}$$
$$E = \{\{1, 2\}, \{1, 3\}, \{2, 3\}, \{3, 4\}, \{3, 5\}, \{4, 5\}\}.$$

Paths

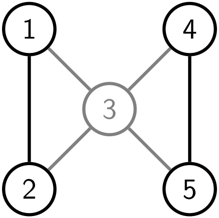


Paths:

$$\pi_1 : 1 - 2 - 3 - 5$$
$$\pi_2 : 3$$

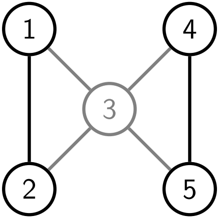
Note that paths may consist of one vertex and no edges.

Induced Subgraph



The **induced subgraph** $G_{\{1,2,4,5\}}$ drops any edges that involve $\{3\}$.

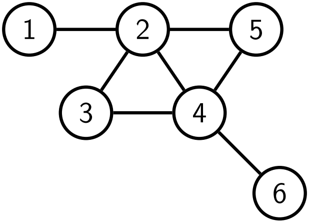
Separation



All paths between $\{1, 2\}$ and $\{5\}$ pass through $\{3\}$.

Hence $\{1, 2\}$ and $\{5\}$ are **separated** by $\{3\}$.

Cliques and Running Intersection



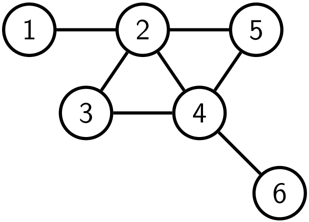
Cliques:

$$\{1,2\} \qquad \{2,3,4\} \qquad \{2,4,5\} \qquad \{4,6\}.$$

Separator sets:

$$\emptyset \qquad \{2\} \qquad \{2,4\} \qquad \{4\}.$$

Cliques and Running Intersection



A different ordering of the cliques:

$$\{2,3,4\} \qquad \{2,4,5\} \qquad \{4,6\} \qquad \{1,2\}.$$

Separator sets:

$$\emptyset \qquad \{2,4\} \qquad \{4\} \qquad \{2\}.$$

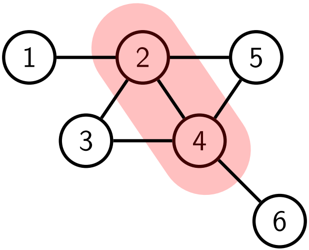
Any ordering works in this case as long $\{1,2\}$ and $\{4,6\}$ aren't the first two entries.

Estimation

Given a decomposition of the graph, we have an associated conditional independence: e.g. $(\{1,3\},\{2,4\},\{5,6\})$ suggests

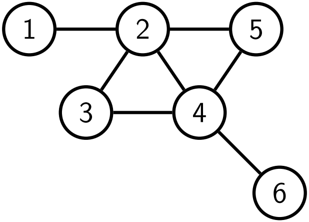
$$X_1, X_3 \perp\!\!\!\perp X_5, X_6 \mid X_2, X_4$$

$$p(x_{123456}) \cdot p(x_{24}) = p(x_{1234}) \cdot p(x_{2456}).$$



And $p(x_{1234})$ and $p(x_{2456})$ are Markov with respect to $\mathcal{G}_{1234}$ and $\mathcal{G}_{2456}$ respectively.

Estimation



Repeating this process on each subgraph we obtain:

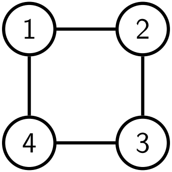
$$p(x_{123456}) \cdot p(x_{24}) \cdot p(x_2) \cdot p(x_4) = p(x_{12}) \cdot p(x_{234}) \cdot p(x_{245}) \cdot p(x_{46}).$$

i.e.

$$p(x_{123456}) = \frac{p(x_{12}) \cdot p(x_{234}) \cdot p(x_{245}) \cdot p(x_{46})}{p(x_{24}) \cdot p(x_2) \cdot p(x_4)}.$$

Non-Decomposable Graphs

But can't we do this for any factorization?



No! Although

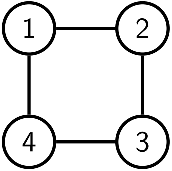
$$p(x_{1234}) = \psi_{12}(x_{12}) \cdot \psi_{23}(x_{23}) \cdot \psi_{34}(x_{34}) \cdot \psi_{14}(x_{14}),$$

the ψs are constrained by the requirement that

$$\sum_{x_{1234}} p(x_{1234}) = 1.$$

These is no nice representation of the ψ_Cs in terms of p.

Non-Decomposable Graphs



If we 'decompose' without a complete separator set then we introduce constraints between the separate terms:

$$p(x_{1234}) = p(x_1 \mid x_2, x_4) \cdot p(x_3 \mid x_2, x_4),$$

but how to ensure that $X_2 \perp\!\!\!\perp X_4 \mid X_1, X_3$?

Gaussian Graphical Models

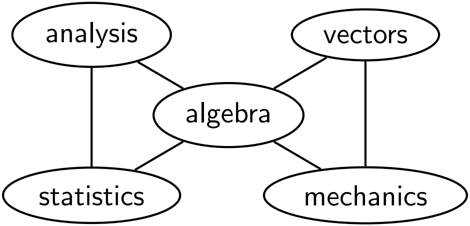
The Multivariate Gaussian Distribution

Let $X_V \sim N_p(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{p \times p}$ is a non-singular symmetric matrix.

$$\log p(x_V; \Sigma) = -\frac{1}{2} \log |\Sigma| - \frac{1}{2} x_V^T \Sigma^{-1} x_V + \text{const.}$$

Gaussian Graphical Models

We have $X_a \perp\!\!\!\perp X_b \mid X_{V \setminus \{a,b\}}$ if and only if $k_{ab} = 0$.



	mechanics	vectors	algebra	analysis	statistics
mechanics	k_{11}	k_{12}	k_{13}	0	0
vectors		k_{22}	k_{23}	0	0
algebra			k_{33}	k_{34}	k_{35}
analysis				k_{44}	k_{45}
statistics					k_{55}

Likelihood

From Lemma 4.23, we have

$$\log p(x_V) + \log p(x_S) = \log p(x_A, x_S) + \log p(x_B, x_S).$$

This becomes

$$x_V^T \Sigma^{-1} x_V + x_S^T (\Sigma_{SS})^{-1} x_S - x_{AS}^T (\Sigma_{AS,AS})^{-1} x_{AS} - x_{SB}^T (\Sigma_{SB,SB})^{-1} x_{SB} = 0$$

But can rewrite each term in the form $x_V^T M x_V$, e.g.:

$$x_{AS}^T (\Sigma_{AS,AS})^{-1} x_{AS} = x_V^T \begin{pmatrix} \multicolumn{2}{c}{(\Sigma_{AS,AS})^{-1}} & 0 \\ & & 0 \\ 0 & 0 & 0 \end{pmatrix} x_V$$

Equating terms gives:

$$\Sigma^{-1} = \begin{pmatrix} \multicolumn{2}{c}{(\Sigma_{AS,AS})^{-1}} & 0 \\ & & 0 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & \multicolumn{2}{c}{(\Sigma_{SB,SB})^{-1}} \\ 0 & & \end{pmatrix} - \begin{pmatrix} 0 & 0 & 0 \\ 0 & (\Sigma_{SS})^{-1} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Maximum Likelihood Estimation

Iterating this process with a decomposable graph shows that:

$$\Sigma^{-1} = \sum_{i=1}^{k} \left\{ (\Sigma_{C_i,C_i})^{-1} \right\}_{C_i,C_i} - \sum_{i=1}^{k} \left\{ (\Sigma_{S_i,S_i})^{-1} \right\}_{S_i,S_i}.$$

For maximum likelihood estimation, using Lemma 4.23 we have

$$\begin{aligned}
\hat{\Sigma}^{-1} &= \sum_{i=1}^{k} \left\{ (\hat{\Sigma}_{C_i,C_i})^{-1} \right\}_{C_i,C_i} - \sum_{i=1}^{k} \left\{ (\hat{\Sigma}_{S_i,S_i})^{-1} \right\}_{S_i,S_i} \\
&= \sum_{i=1}^{k} \left\{ (W_{C_i,C_i})^{-1} \right\}_{C_i,C_i} - \sum_{i=1}^{k} \left\{ (W_{S_i,S_i})^{-1} \right\}_{S_i,S_i}
\end{aligned}$$

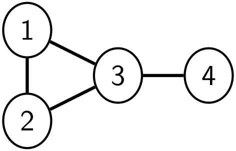
where $W_{CC} = \frac{1}{n} \sum_i X_C^{(i)} X_C^{(i)T}$ is the sample covariance matrix.

Example

```
> true_inv          # true concentration matrix

     [,1] [,2] [,3] [,4]
[1,]  1.0  0.3  0.2  0.0
[2,]  0.3  1.0 -0.1  0.0
[3,]  0.2 -0.1  1.0  0.3
[4,]  0.0  0.0  0.3  1.0

> solve(true_inv)  # Sigma

      [,1]   [,2]  [,3]   [,4]
[1,]  1.17 -0.382 -0.30  0.090
[2,] -0.38  1.136  0.21 -0.063
[3,] -0.30  0.209  1.19 -0.356
[4,]  0.09 -0.063 -0.36  1.107

> dat <- rmvnorm(1000, mean=rep(0,4), sigma = solve(true_inv))
> W <- cov(dat)     # sample covariance
```

Example

Fit the model in which there is a decomposition $(\{1,2\},\{3\},\{4\})$:



```
> K_hat[1:3, 1:3] = solve(W[1:3, 1:3])
> K_hat[3:4, 3:4] = K_hat[3:4, 3:4] + solve(W[3:4, 3:4])
> K_hat[3, 3] = K_hat[3, 3] - 1/W[3, 3]
> K_hat

     [,1]  [,2]  [,3] [,4]
[1,] 0.97  0.28  0.17 0.00
[2,] 0.28  0.97 -0.12 0.00
[3,] 0.17 -0.12  1.08 0.31
[4,] 0.00  0.00  0.31 1.00
```

Note this is close to the true concentration matrix.

Directed Graphical Models

Directed Graphs

We have so far used **undirected graphs**.

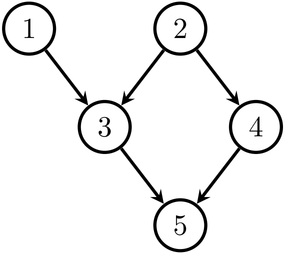
In **directed graphs** each edge has an orientation.

A directed graph $\mathcal{G}$ is a pair (V, D), where

- V is a set of vertices;
- D is a set of ordered pairs of vertices (i, j) such that $i, j \in V$ and $i \neq j$.

If $(i, j) \in D$ we write $i \to j$.

$V = \{1, 2, 3, 4, 5\}$
$D = \{(1, 3), (2, 3), (2, 4), (3, 5), (4, 5)\}$



Acyclicity

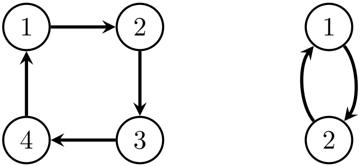
Paths are sequences of adjacent vertices, without repetition:

$$1 \to 3 \leftarrow 2 \to 4 \to 5 \qquad\qquad 1 \to 3 \to 5.$$

The path is **directed** if all the arrows point away from the start.

(A path of length 0 is just a single vertex.)

A **directed cycle** is a directed path from i to $j \neq i$, together with $j \to i$.



Graphs that contain no directed cycles are called **acyclic**. or more specifically, **directed acyclic graphs** (DAGs).

All the directed graphs we consider are acyclic.

Happy Families

$$i \to j \quad \begin{cases} i \in \mathrm{pa}_{\mathcal{G}}(j) & i \text{ is a parent of } j \\ j \in \mathrm{ch}_{\mathcal{G}}(i) & j \text{ is a child of } i \end{cases}$$

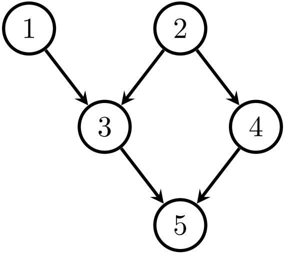
$$\begin{array}{c} a \to \cdots \to b \\ \text{or } a = b \end{array} \quad \begin{cases} a \in \mathrm{an}_{\mathcal{G}}(b) & a \text{ is an ancestor of } b \\ b \in \mathrm{de}_{\mathcal{G}}(a) & b \text{ is a descendant of } a \end{cases}$$

If $w \notin \mathrm{de}_{\mathcal{G}}(v)$ then w is a **non-descendant** of v:

$$\mathrm{nd}_{\mathcal{G}}(v) = V \setminus \mathrm{de}_{\mathcal{G}}(v).$$

(Notice that no v is a non-descendant of itself).

Examples



$$\mathrm{pa}_{\mathcal{G}}(3) = \{1, 2\} \qquad \mathrm{an}_{\mathcal{G}}(4) = \{2, 4\}$$

$$\mathrm{ch}_{\mathcal{G}}(5) = \emptyset \qquad \mathrm{de}_{\mathcal{G}}(1) = \{1, 3, 5\}$$

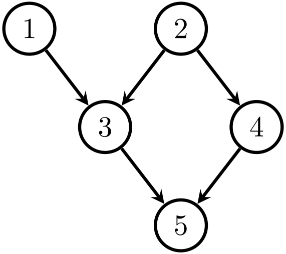
$$\mathrm{nd}_{\mathcal{G}}(1) = \{2, 4\}.$$

Topological Orderings

If the graph is acyclic, we can find a **topological ordering**: i.e. one in which no vertex comes before any of its parents. (Proof: induction)

Topological orderings:

$$1, 2, 3, 4, 5$$
$$1, 2, 4, 3, 5$$
$$2, 1, 3, 4, 5$$
$$2, 1, 4, 3, 5$$
$$2, 4, 1, 3, 5$$



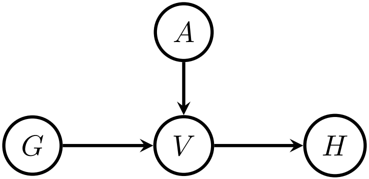
Parameter Estimation

G : group assigned to patient;

A : patient's age in years;

V : whether patient received flu vaccine;

H : patient hospitalized with respiratory problems;

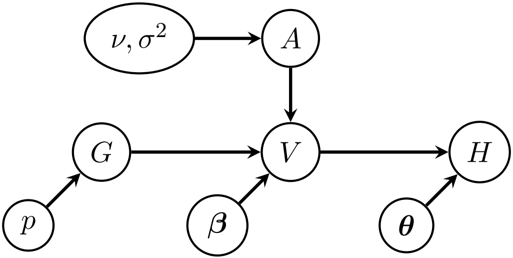


Parameter Estimation

We can model the data (G_i, A_i, V_i, H_i) as

$$\begin{aligned}
\text{group} &: G_i \sim \text{Bernoulli}(p); \\
\text{age} &: A_i \sim N(\nu, \sigma^2); \\
\text{vaccine} &: V_i \mid A_i, G_i \sim \text{Bernoulli}(\mu_i) \text{ where} \\
&\qquad \log \mu_i = \beta_0 + \beta_1 A_i + \beta_2 G_i. \\
\text{hospital} &: H_i \mid V_i \sim \text{Bernoulli}(\exp(\theta_0 + \theta_1 V_i)).
\end{aligned}$$

Assuming independent priors:

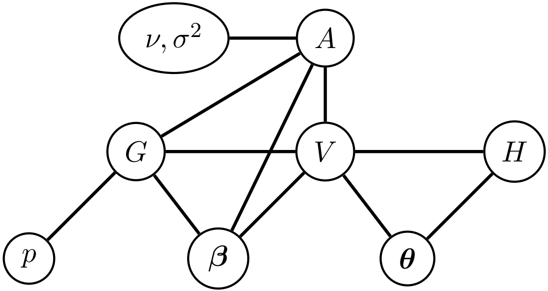


Bayesian Inference

From our argument, we have

$$\begin{aligned}\pi(\boldsymbol{\beta} \mid G, A, V, H) &= \pi(\boldsymbol{\beta} \mid G, A, V) \\ &\propto p(V \mid A, G, \boldsymbol{\beta}) \cdot \pi(\boldsymbol{\beta}).\end{aligned}$$

Looking at the moral graph we see

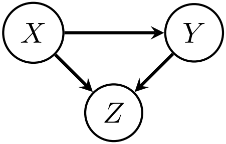


Markov Equivalence

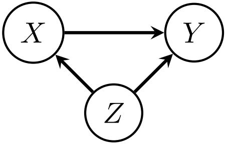
All undirected graphs induce distinct models.

$$v \not\sim w \qquad \iff \qquad X_v \perp\!\!\!\perp X_w \mid X_{V \setminus \{v,w\}} \text{ implied}$$

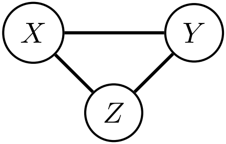
The same is not true for directed graphs:



$p(x) \cdot p(y \mid x) \cdot p(z \mid x, y)$

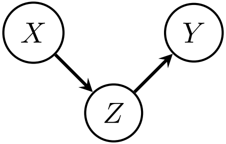


$p(z) \cdot p(x \mid z) \cdot p(y \mid x, z)$

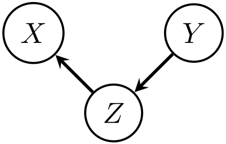


$\psi_{XYZ}(x, y, z)$

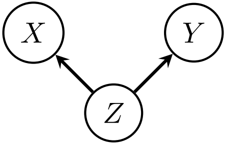
Markov Equivalence



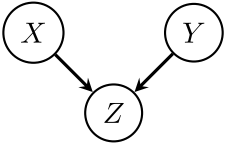
$$p(x) \cdot p(z \mid x) \cdot p(y \mid z)$$
$$X \perp\!\!\!\perp Y \mid Z$$



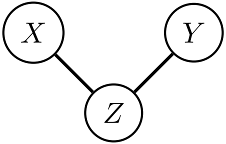
$$p(y) \cdot p(z \mid y) \cdot p(x \mid z)$$
$$X \perp\!\!\!\perp Y \mid Z$$



$$p(z) \cdot p(x \mid z) \cdot p(y \mid z)$$
$$X \perp\!\!\!\perp Y \mid Z$$



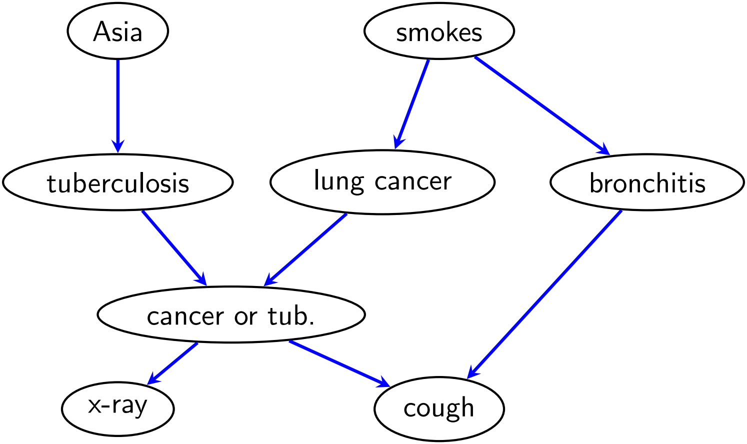
$$p(x) \cdot p(y) \cdot p(z \mid x, y)$$
$$X \perp\!\!\!\perp Y$$



$$\psi_{XZ}(x, z) \cdot \psi_{YZ}(y, z)$$
$$X \perp\!\!\!\perp Y \mid Z$$

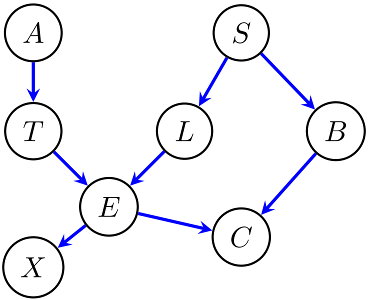
Expert Systems

Expert Systems



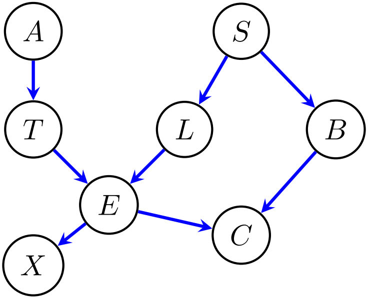
The ‘Chest Clinic’ network, a fictitious diagnostic model.

Variables



- A has the patient recently visited Asia?
- S does the patient smoke?
- T,C,B Tuberculosis, lung cancer, bronchitis.
- E logical: Tuberculosis OR lung cancer.
- X shadow on chest X-ray?
- C does the patient have a persistent cough?

Conditional Probability Tables



We have our factorization:

$$p(a, s, t, l, b, e, x, c) = p(a) \cdot p(s) \cdot p(t \mid a) \cdot p(l \mid s) \cdot p(b \mid s) \cdot \\ \cdot p(e \mid t, l) \cdot p(x \mid e) \cdot p(c \mid e, b).$$

Assume that we are given each of these factors. How could we calculate $p(l \mid x, c, a, s)$?

Probabilities

$$p(a) = \begin{array}{cc} \text{yes} & \text{no} \\ \hline 0.01 & 0.99 \end{array}$$

$$p(s) = \begin{array}{cc} \text{yes} & \text{no} \\ \hline 0.5 & 0.5 \end{array}$$

$$p(t \mid a) = \begin{array}{c||cc} A & \text{yes} & \text{no} \\ \hline \text{yes} & 0.05 & 0.95 \\ \text{no} & 0.01 & 0.99 \end{array}$$

$$p(l \mid s) = \begin{array}{c||cc} S & \text{yes} & \text{no} \\ \hline \text{yes} & 0.1 & 0.9 \\ \text{no} & 0.01 & 0.99 \end{array}$$

$$p(b \mid s) = \begin{array}{c||cc} S & \text{yes} & \text{no} \\ \hline \text{yes} & 0.6 & 0.4 \\ \text{no} & 0.3 & 0.7 \end{array}$$

$$p(x \mid e) = \begin{array}{c||cc} E & \text{yes} & \text{no} \\ \hline \text{yes} & 0.98 & 0.02 \\ \text{no} & 0.05 & 0.95 \end{array}$$

$$p(c \mid b, e) = \begin{array}{cc||cc} B & E & \text{yes} & \text{no} \\ \hline \text{yes} & \text{yes} & 0.9 & 0.1 \\ & \text{no} & 0.8 & 0.2 \\ \text{no} & \text{yes} & 0.7 & 0.3 \\ & \text{no} & 0.1 & 0.9 \end{array}$$

Factorizations

$$p(l \mid x, c, a, s) = \frac{p(l, x, c \mid a, s)}{\sum_l p(l, x, c \mid a, s)}$$

From the graph $p(l, x, c \mid a, s)$ is

$$\sum_{t,e,b} p(t \mid a) \cdot p(l \mid s) \cdot p(b \mid s) \cdot p(e \mid t, l) \cdot p(x \mid e) \cdot p(c \mid e, b).$$

But this is:

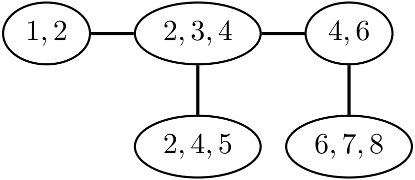
$$p(l \mid s) \sum_e p(x \mid e) \left(\sum_b p(b \mid s) \cdot p(c \mid e, b) \right) \left(\sum_t p(t \mid a) \cdot p(e \mid t, l) \right).$$

Junction Trees

A **junction tree**:

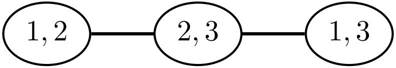
- is a connected undirected graph without cycles (a tree);
- has vertices C_i that consist of **subsets** of a set V;
- satsifies the property that if $C_i \cap C_j = S$ then every vertex on the (unique) path from C_i to C_j contains S.

Example.



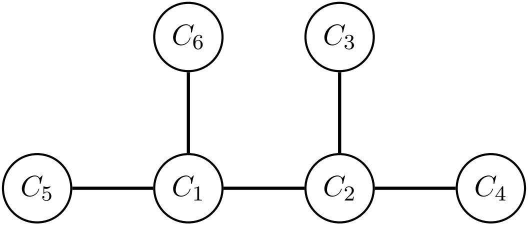
Junction Trees

The following graph is **not** a junction tree:



Junction Trees

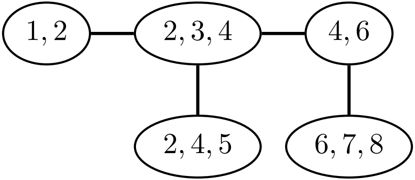
Junction trees can be constructed directly from sets of cliques satisfying running intersection.



$$C_i \cap \bigcup_{j<i} C_j = C_i \cap C_{\sigma(i)}.$$

Example: Junction Trees and RIP

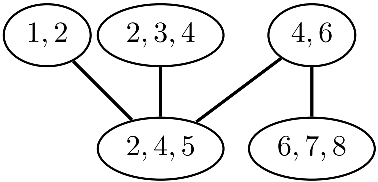
Given sets $\{1,2\}$, $\{2,3,4\}$, $\{2,4,5\}$, $\{4,6\}$, $\{6,7,8\}$, we can build this tree:



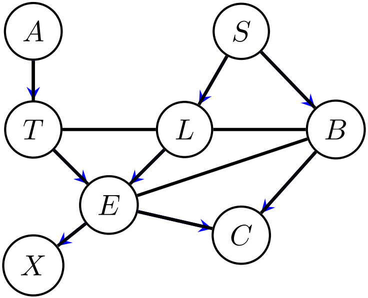
Example: Junction Trees and RIP

Equally, we could use a different ordering:

$$\{6, 7, 8\}, \{4, 6\}, \{2, 4, 5\}, \{1, 2\}, \{2, 3, 4\}.$$



Forming A Junction Tree



Steps to Forming a Junction Tree:

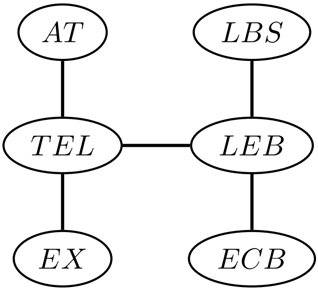
Moralize

Drop directions

Triangulate (add edges to get a decomposable graph)

Forming A Junction Tree

Finally, form the tree of cliques.



Initialization

$$p(a) = \begin{array}{cc} \text{yes} & \text{no} \\ \hline 0.01 & 0.99 \end{array}$$

$$p(s) = \begin{array}{cc} \text{yes} & \text{no} \\ \hline 0.5 & 0.5 \end{array}$$

$$p(t \mid a) = \begin{array}{c||cc} A & \text{yes} & \text{no} \\ \hline \text{yes} & 0.05 & 0.95 \\ \text{no} & 0.01 & 0.99 \end{array}$$

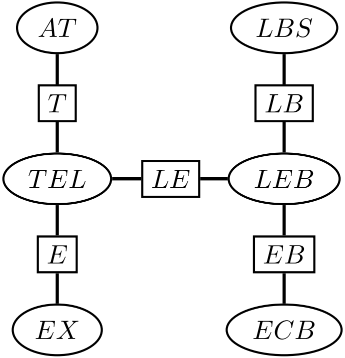
$$p(l \mid s) = \begin{array}{c||cc} S & \text{yes} & \text{no} \\ \hline \text{yes} & 0.1 & 0.9 \\ \text{no} & 0.01 & 0.99 \end{array}$$

$$p(b \mid s) = \begin{array}{c||cc} S & \text{yes} & \text{no} \\ \hline \text{yes} & 0.6 & 0.4 \\ \text{no} & 0.3 & 0.7 \end{array}$$

$$p(x \mid e) = \begin{array}{c||cc} E & \text{yes} & \text{no} \\ \hline \text{yes} & 0.98 & 0.02 \\ \text{no} & 0.05 & 0.95 \end{array}$$

$$p(c \mid b, e) = \begin{array}{cc||cc} B & E & \text{yes} & \text{no} \\ \hline \text{yes} & \text{yes} & 0.9 & 0.1 \\ & \text{no} & 0.8 & 0.2 \\ \text{no} & \text{yes} & 0.7 & 0.3 \\ & \text{no} & 0.1 & 0.9 \end{array}$$

Initialization

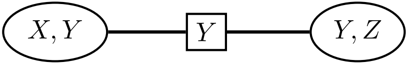


Can set, for example:

$$\psi_{AT}(a,t) = p(a) \cdot p(t \mid a) \qquad \psi_{LBS}(l,b,s) = p(s) \cdot p(l \mid s) \cdot p(b \mid s)$$
$$\psi_{TEL}(t,e,l) = p(e \mid t,l) \qquad \psi_{ELB}(e,l,b) = 1$$
$$\psi_{EX}(e,x) = p(x \mid e) \qquad \psi_{ECB}(e,c,b) = p(c \mid e,b).$$

Updating / Message Passing

Suppose we have two vertices and one separator set.



$\psi_{XY}(x, y)$

		$y = 0$	1
x	0	0.3	0.9
	1	0.7	0.1

$\psi_Y(y)$

$y = 0$	1
1	1

$\psi_{YZ}(y, z)$

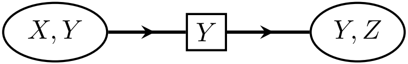
		$z = 0$	1
y	0	0.3	0.1
	1	0.2	0.4

Initialize with

$$\psi_{XY}(x, y) = p(x \mid y) \qquad \psi_{YZ}(y, z) = p(z \mid y) \cdot p(y) \qquad \psi_Y(y) = 1.$$

Updating / Message Passing

Suppose we have two vertices and one separator set.



$\psi_{XY}(x,y)$

		$y=0$	1
x	0	0.3	0.9
	1	0.7	0.1

$\psi_Y(y)$

$y=0$	1
1	1

$\psi_{YZ}(y,z)$

		$z=0$	1
y	0	0.3	0.1
	1	0.2	0.4

Pass message from X, Y to Y, Z. We set

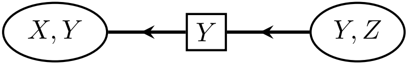
$$\psi'_Y(y) = \sum_x \psi_{XY}(x,y) = (1,1);$$

$$\psi'_{YZ}(y,z) = \frac{\psi'_Y(y)}{\psi_Y(y)} \psi_{YZ}(y,z) = \psi_{YZ}(y,z).$$

So in this case nothing changes.

Updating / Message Passing

Suppose we have two vertices and one separator set.



$\psi_{XY}(x, y)$

		$y = 0$	1
x	0	0.3	0.9
	1	0.7	0.1

$\psi'_Y(y)$

$y = 0$	1
1	1

$\psi'_{YZ}(y, z)$

		$z = 0$	1
y	0	0.3	0.1
	1	0.2	0.4

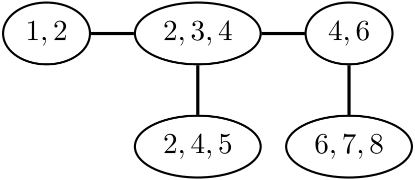
Pass message from Y, Z to X, Y. We set

$$\psi''_Y(y) = \sum_x \psi_{YZ}(y, z) = (0.4, 0.6);$$

$$\psi'_{XY}(x, y) = \frac{\psi''_Y(y)}{\psi'_Y(y)} \psi_{XY}(x, y) = \begin{matrix} 0.12 & 0.54 \\ 0.28 & 0.06 \end{matrix}.$$

And now we note that $\psi'_{XY}(x, y) = p(x, y)$ as intended.

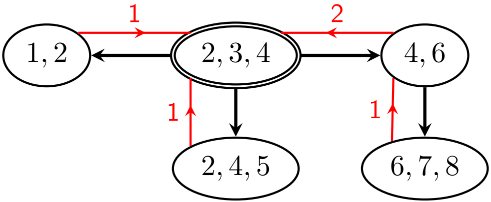
Rooting



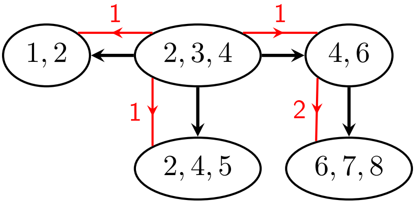
Given a tree, we can pick any vertex as a 'root', and direct all edges away from it.

Collection and Distribution

function $\textsc{Collect}$(rooted tree $\mathcal{T}$, potentials ψ_t)
  let $1 < \ldots < k$ be a topological ordering of $\mathcal{T}$
  for t in $k, \ldots, 2$ **do**
    send message from ψ_t to $\psi_{\sigma(t)}$;
  end for
  return updated potentials ψ_t
end function

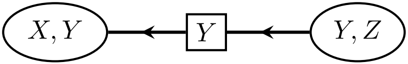


function $\textsc{Distribute}$(rooted tree $\mathcal{T}$, potentials ψ_t)
    let $1 < \ldots < k$ be a topological ordering of $\mathcal{T}$
    for t in $2, \ldots, k$ **do**
        send message from $\psi_{\sigma(t)}$ to ψ_t;
    end for
    return updated potentials ψ_t
end function



Evidence

Now, suppose we want to calculate $p(x \mid z = 0)$.



$\psi_{XY}(x, y)$

		$y = 0$	1
x	0	0.12	0.54
	1	0.28	0.06

$\psi_Y(y)$

$y = 0$	1
0.4	0.6

$\psi_{YZ}(y, z)$

		$z = 0$	1
y	0	0.6	0
	1	0.4	0

Replace $\psi_{YZ}(y, z)$ with $p(y \mid z = 0)$.

Pass message from Y, Z to X, Y. We set

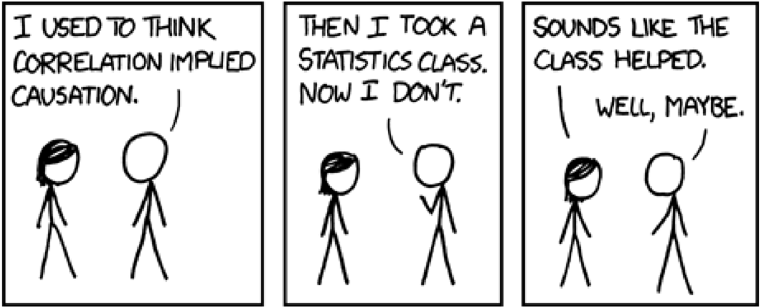
$$\psi_Y(y) = \sum_x \psi_{YZ}(y, z) = (0.60.4);$$

$$\psi'_{XY}(x, y) = \frac{\psi''_Y(y)}{\psi'_Y(y)} \psi_{XY}(x, y) = \begin{array}{cc} 0.18 & 0.36 \\ 0.42 & 0.04 \end{array}.$$

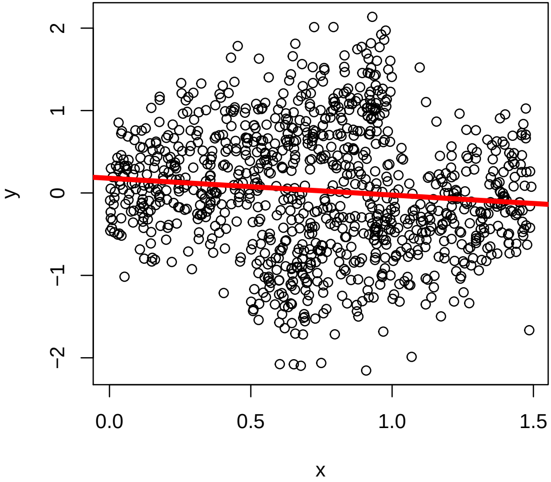
And now calculate $\sum_y \psi_{XY}(x, y) = (0.54, 0.46)$.

Causal Inference

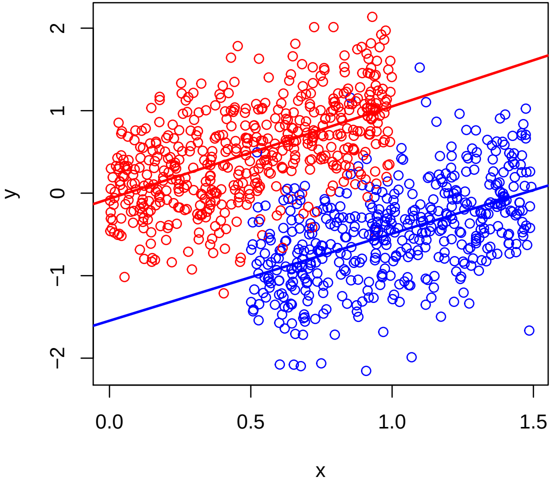
Correlation



Controlling for Covariates

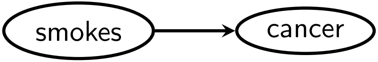


Controlling for Covariates



Causation

Example. Smoking is strongly predictive of lung cancer. So maybe smoking causes lung cancer to develop.

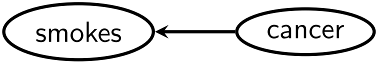


BUT: how do we know that this is a causal relationship? And what do we mean by that?

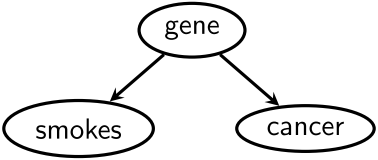
The central question is: "if we stop people from smoking, will they be less likely to get lung cancer?"

That is: does this 'intervention' on one variable change the distribution of another variable?

Alternative Explanations



Reverse Causation. Lung cancer causes smoking: people with (undiagnosed) lung cancer smoke to soothe irritation in the lungs.



Confounding / Common Cause. There is a gene that makes people likely to smoke, and also more likely to get lung cancer.

Example

Suppose we take 32 men and 32 women, ask them whether they smoke and check for lung damage.

	women			men	
	not smoke	smoke		not smoke	smoke
no damage	21	6		6	6
damage	3	2		2	18

Marginally, there is clearly a strong relationship between smoking and damage

	not smoke	smoke
no damage	27	12
damage	5	20

$$P(D = 1 \mid S = 1) = \frac{5}{8} \qquad P(D = 1 \mid S = 0) = \frac{5}{32}.$$

Example

This might suggest that if we had prevented them all from smoking, only $\frac{5}{32} \times 64 = 10$ would have had damage, whereas if we had made them all smoke, $\frac{5}{8} \times 64 = 40$ would have damage.

But: both smoking and damage are also correlated with gender, so this effect may be inaccurate. If we repeat this separately for men and women:

no-one smoking:

$$\frac{3}{21+3} \times 32 + \frac{2}{6+2} \times 32 = 12$$

everyone smoking

$$\frac{2}{6+2} \times 32 + \frac{18}{18+6} \times 32 = 32.$$

Compare these to 10 and 40.

‘do’ notation

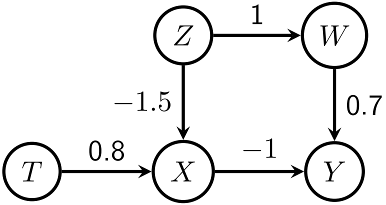
In this example there is a difference between predicting damage when we ‘observe’ that someone smokes . . .

$$P(D = 1 \mid S = 1) = \frac{5}{8},$$

. . . and prediciting damage when we intervene to make someone smoke:

$$P(D = 1 \mid do(S = 1)) = \frac{32}{64} = \frac{1}{2}.$$

Gaussian




```
> set.seed(513)
> n <- 1e3
> Z <- rnorm(n)
> T <- rnorm(n)
> W <- Z + rnorm(n)
> X <- 0.8*T - 1.5*Z + rnorm(n)
> Y <- 0.7*W      - X + rnorm(n)
```

Back-Door Paths

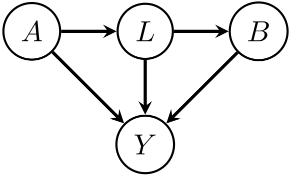
```
> summary(lm(Y ~ X))$coefficients[,1:2]

            Estimate Std. Error
(Intercept)    0.035       0.04
X             -1.285       0.02

> summary(lm(Y ~ X + Z))$coefficients[,1:2]

            Estimate Std. Error
(Intercept)    0.043      0.038
X             -1.024      0.032
Z              0.645      0.062

> summary(lm(Y ~ X + W))$coefficients[,1:2]

            Estimate Std. Error
(Intercept)    0.029      0.031
X             -1.011      0.019
W              0.668      0.027
```

Instruments

Adding in unnecessary variables to the regression generally increases the variance.

```
> summary(lm(Y ~ X + W + T))$coefficients[,1:2]

             Estimate Std. Error
(Intercept)     0.029      0.031
X              -1.006      0.022
W               0.671      0.027
T              -0.018      0.036

> summary(lm(Y ~ X + W + Z))$coefficients[,1:2]

             Estimate Std. Error
(Intercept)     0.028      0.031
X              -1.026      0.026
W               0.682      0.031
Z              -0.053      0.061
```

Example: HIV Treatment



A treatment with AZT (an HIV drug);
L opportunisitic infection;
B treatment with antibiotics;
Y survival at 5 years.

$$
\begin{aligned}
p(a, l, b, y) &= p(a) \cdot p(l \mid a) \cdot p(b \mid l) \cdot p(y \mid a, l, b) \\
p(l, y \mid do(a, b)) &= p(l \mid a) \cdot p(y \mid a, l, b) \\
p(y \mid do(a, b)) &= \sum_l p(l \mid a) \cdot p(y \mid a, l, b).
\end{aligned}
$$

Iterative Proportional Fitting

The Iterative Proportional Fitting Algorithm

function IPF(collection of margins $q(x_{C_i})$)
  set $p(x_V)$ to uniform distribution;
  while $\max_i \max_{x_{C_i}} |p(x_{C_i}) - q(x_{C_i})| >$ tol **do**
    for i in $1, \ldots, k$ **do**
      update $p(x_V)$ to $p(x_{V \setminus C_i} \mid x_{C_i}) \cdot q(x_{C_i})$;
    end for
  end while
  return distribution p with margins $p(x_{C_i}) = q(x_{C_i})$.
end function

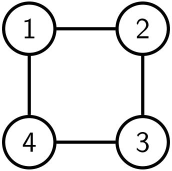
If any distribution satisfying $p(x_{C_i}) = q(x_{C_i})$ exists, then the algorithm converges to the **unique distribution** with those margins and which is Markov with respect to the graph with cliques $C_1, \ldots, C_k$.

Some Data

		$X_2 = 0$		$X_2 = 1$	
		$X_1 = 0$	1	0	1
$X_4 = 0$	$X_3 = 0$	5	10	18	1
	1	0	3	4	0
$X_4 = 1$	0	24	0	9	3
	1	1	2	2	7

Margins

Suppose we want to fit the 4-cycle model:



The relevant margins are:

$n(x_{12})$	$X_2 = 0$	1
$X_1 = 0$	30	33
1	15	11

$n(x_{23})$	$X_3 = 0$	1
$X_2 = 0$	39	6
1	31	13

$n(x_{34})$	$X_4 = 0$	1
$X_3 = 0$	34	36
1	7	12

$n(x_{14})$	$X_4 = 0$	1
$X_1 = 0$	27	36
1	14	12

Start with a Uniform Table

		$X_2 = 0$		$X_2 = 1$	
		$X_1 = 0$	1	0	1
$X_4 = 0$	$X_3 = 0$	5.56	5.56	5.56	5.56
	1	5.56	5.56	5.56	5.56
$X_4 = 1$	0	5.56	5.56	5.56	5.56
	1	5.56	5.56	5.56	5.56

Set Margin X_1, X_2 to Correct Value

		$X_2 = 0$		$X_2 = 1$	
		$X_1 = 0$	1	0	1
$X_4 = 0$	$X_3 = 0$	7.5	3.75	8.25	2.75
	1	7.5	3.75	8.25	2.75
$X_4 = 1$	0	7.5	3.75	8.25	2.75
	1	7.5	3.75	8.25	2.75

Replace

$$p^{(i+1)}(x_1, x_2, x_3, x_4) = p^{(i)}(x_1, x_2, x_3, x_4) \cdot \frac{n(x_1, x_2)}{p^{(i)}(x_1, x_2)}$$

Set Margin X_2, X_3 to Correct Value

		$X_2 = 0$		$X_2 = 1$	
		$X_1 = 0$	1	0	1
$X_4 = 0$	$X_3 = 0$	13	6.5	11.62	3.88
	1	2	1	4.88	1.62
$X_4 = 1$	0	13	6.5	11.62	3.88
	1	2	1	4.88	1.62

Replace

$$p^{(i+1)}(x_1, x_2, x_3, x_4) = p^{(i)}(x_1, x_2, x_3, x_4) \cdot \frac{n(x_2, x_3)}{p^{(i)}(x_2, x_3)}$$

Set Margin X_3, X_4 to Correct Value

		$X_2 = 0$		$X_2 = 1$	
		$X_1 = 0$	1	0	1
$X_4 = 0$	$X_3 = 0$	12.63	6.31	11.29	3.76
	1	1.47	0.74	3.59	1.2
$X_4 = 1$	0	13.37	6.69	11.96	3.99
	1	2.53	1.26	6.16	2.05

Replace

$$p^{(i+1)}(x_1, x_2, x_3, x_4) = p^{(i)}(x_1, x_2, x_3, x_4) \cdot \frac{n(x_3, x_4)}{p^{(i)}(x_3, x_4)}$$

Set Margin X_1, X_4 to Correct Value

		$X_2 = 0$		$X_2 = 1$	
		$X_1 = 0$	1	0	1
$X_4 = 0$	$X_3 = 0$	11.76	7.36	10.52	4.39
	1	1.37	0.86	3.35	1.4
$X_4 = 1$	0	14.15	5.74	12.66	3.42
	1	2.67	1.08	6.52	1.76

Replace

$$p^{(i+1)}(x_1, x_2, x_3, x_4) = p^{(i)}(x_1, x_2, x_3, x_4) \cdot \frac{n(x_1, x_4)}{p^{(i)}(x_1, x_4)}$$

Notice that sum of first column is now 29.96.

Set Margin X_1, X_2 to Correct Value

		$X_2 = 0$		$X_2 = 1$	
		$X_1 = 0$	1	0	1
$X_4 = 0$	$X_3 = 0$	11.78	7.37	10.53	4.39
	1	1.37	0.86	3.34	1.39
$X_4 = 1$	0	14.14	5.73	12.64	3.42
	1	2.68	1.09	6.54	1.77

Eventually:

Waiting for this process to converge leads to the MLE:

		$X_2 = 0$		$X_2 = 1$	
		$X_1 = 0$	1	0	1
$X_4 = 0$	$X_3 = 0$	11.76	7.33	10.5	4.4
	1	1.38	0.86	3.35	1.4
$X_4 = 1$	0	14.18	5.72	12.66	3.44
	1	2.68	1.08	6.48	1.76