SC6/SM9 Graphical Models

Michaelmas Term, 2024

Robin Evans

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

Course Websites

The class site is at

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

The Canvas site is

`https://canvas.ox.ac.uk/courses/274755/`

You'll find

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

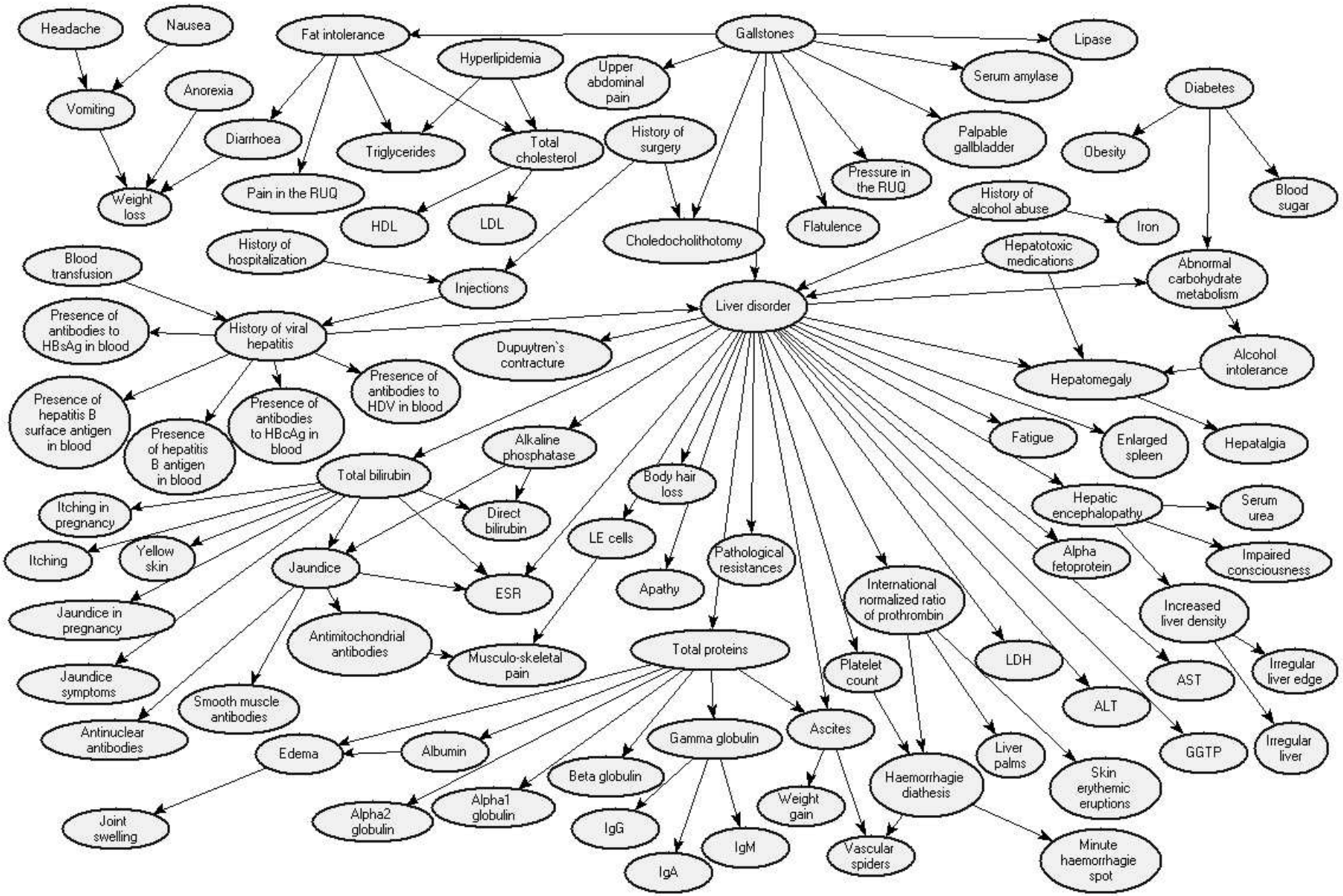
There will be four problem sheets and four associated classes.

Books

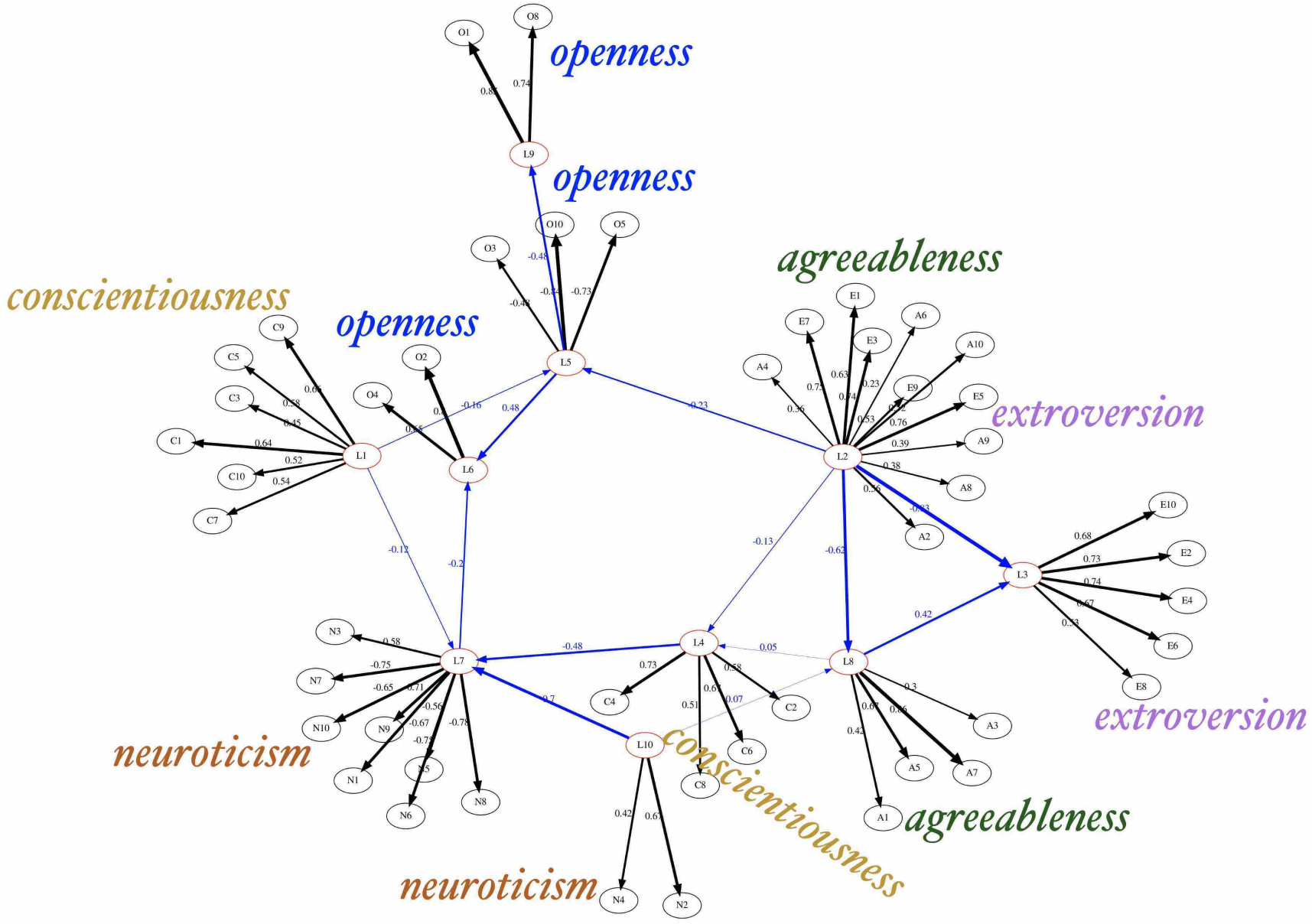
These books might be useful.

- 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.
- Agresti (2002). *Categorical Data Analysis*, (2nd edition), John Wiley & Sons.

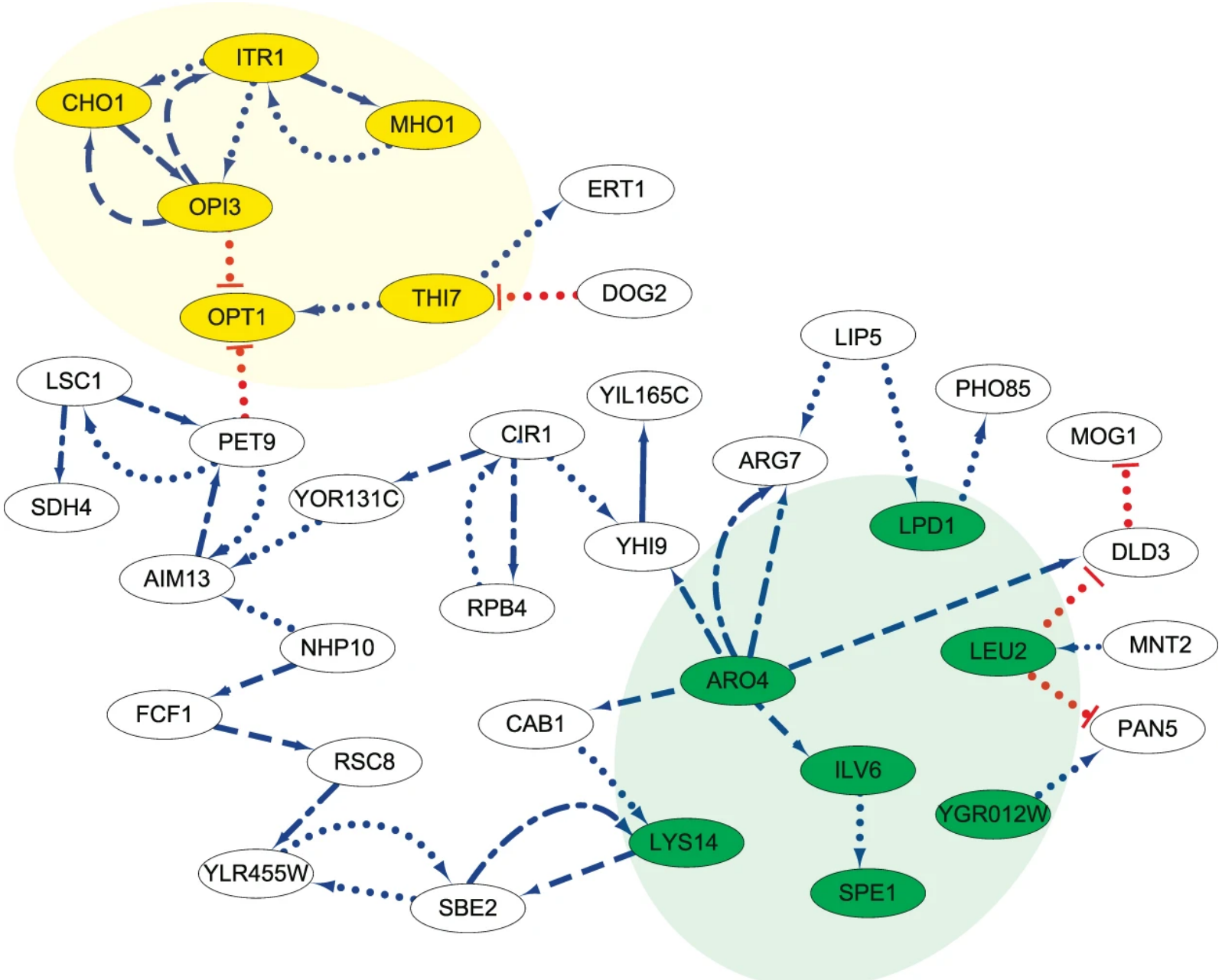
Medical Diagnosis



Psychometrics



Gene Regulatory Networks



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:

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

Our solution will be 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	Black
White	Yes	53	11
	No	414	37
Black	Yes	0	4
	No	16	139

Morals

Let:

- D be an indicator that the death penalty was imposed;
- V be an indicator for the race of the victim;
- R be an indicator for the race of the defendant.

By changing the numbers only very slightly, it is easy to obtain either:

$$D \perp\!\!\!\perp R \qquad \text{and} \qquad D \not\perp\!\!\!\perp R \mid V.$$

Similarly, one can generate tables such that

$$D \not\perp\!\!\!\perp R \qquad \text{and} \qquad D \perp\!\!\!\perp R \mid V.$$

Exponential Families

Contingency Tables: Some Notation

We will consider multivariate systems of vectors $X_V \equiv (X_v : v \in V)$ for some set $V = \{1, \ldots, p\}$.

Write $X_A \equiv (X_v : v \in A)$ for any $A \subseteq V$.

We assume that each $X_v \in \{1, \ldots, d_v\}$ (usually $d_v = 2$).

If we have n i.i.d. observations write

$$X_V^{(i)} \equiv (X_1^{(i)}, \ldots, X_p^{(i)})^T, \qquad i = 1, \ldots, n.$$

Contingency Tables: Some Notation

We typically summarize categorical data by counts:

aspirin	heart attack
Y	N
Y	Y
N	N
N	N
Y	N
⋮	⋮

	heart attack	
	Y	N
no aspirin	28	656
aspirin	18	658

Write

$$n(x_V) = \sum_{i=1}^{n} \mathbb{1}\{X_1^{(i)} = x_1, \ldots, X_p^{(i)} = x_p\}$$

A **marginal table** only counts some of the variables.

$$n(x_A) = \sum_{i=1}^{n} \mathbb{1}\{X_a^{(i)} = x_a : a \in A\} = \sum_{x_{V \setminus A}} n(x_A, x_{V \setminus A}).$$

Marginal Table

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

If we sum out the Victim's race...

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

Contingency Tables

The death penalty data is on the class website.

```
> getwd()

[1] "/Users/robin/Dropbox/Teaching/Graphical Models/Datasets"

> deathpen <- read.table("deathpen.txt", header=TRUE)
> deathpen

  DeathPen Defendant Victim freq
1      Yes     White  White   53
2       No     White  White  414
3      Yes     Black  White   11
4       No     Black  White   37
5      Yes     White  Black    0
6       No     White  Black   16
7      Yes     Black  Black    4
8       No     Black  Black  139
```

Contingency Tables

We can fit models on it in R:

```
> summary(glm(freq ~ Victim*Defendant + Victim*DeathPen,
+                 family=poisson, data=deathpen))
```

```
Coefficients:
                            Estimate Std. Error
(Intercept)                  4.93737    0.08459
VictimWhite                 -1.19886    0.16812
DefendantWhite              -2.19026    0.26362
DeathPenYes                 -3.65713    0.50641
VictimWhite:DefendantWhite   4.46538    0.30408
VictimWhite:DeathPenYes      1.70455    0.52373

Residual deviance:    5.394  on 2  degrees of freedom
```

(So $p \approx 0.07$ in hypothesis test of model fit.)

Contingency Tables

If we fit the marginal table over the races of Victim and Defendant, the parameters involving 'Defendant' are the same.

```
> summary(glm(freq ~ Victim*Defendant,
+                 family=poisson, data=deathpen))
```

```
Coefficients:
                          Estimate Std. Error
(Intercept)                4.26970    0.08362
VictimWhite               -1.09164    0.16681
DefendantWhite            -2.19026    0.26360
VictimWhite:DefendantWhite  4.46538    0.30407
```

Contingency Tables

We can also check that the subsets of $S = \{\texttt{Victim}\}$ are given by the other condition we had:

$$\lambda_W = \lambda_W^{AS} + \lambda_W^{BS} - \lambda_W^{S}.$$

```
> out1 <- glm(freq ~ Victim*Defendant, family=poisson,
+             data=deathpen)$coef[1:2]
> out2 <- glm(freq ~ Victim*DeathPen, family=poisson,
+             data=deathpen)$coef[1:2]
> out <- glm(freq ~ Victim, family=poisson,
+            data=deathpen)$coef[1:2]
>
> out1 + out2 - out

(Intercept) VictimWhite
   4.937366   -1.198864
```

Indeed these match the coefficients from the larger model.

Poisson-Multinomial Equivalence

The following distributions are equivalent.

1. Independent Poisson random variables:

$$X_i \sim \text{Poisson}(\mu_i) \quad \text{for } i = 1, \ldots, k.$$

2. One Poisson random variable $N \sim \text{Poisson}(\mu)$ where $\mu = \sum_i \mu_i$; and a multinomial

$$(X_1, \ldots, X_k)^T | \{N = n\} \sim \text{Multinom}(n, (\pi_1, \ldots, \pi_k)^T),$$

where $\pi_i = \mu_i / \mu$.

Poisson-Multinomial Equivalence

We can see this by comparing the likelihoods.

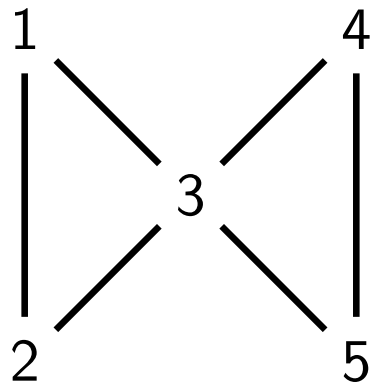
The Poisson likelihood is

$$
\begin{aligned}
&L(\mu_1, \ldots, \mu_k; x_1, \ldots, x_k) \\
&= \prod_{i=1}^{k} e^{-\mu_i} \mu_i^{x_i} = \prod_{i=1}^{k} e^{-\mu \pi_i} \mu^{x_i} \pi_i^{x_i} \\
&= \mu^{\sum_i x_i} e^{-\mu \sum_i \pi_i} \prod_{i=1}^{k} \pi_i^{x_i} \\
&= \mu^n e^{-\mu} \prod_{i=1}^{k} \pi_i^{x_i} \\
&= L(\mu; n) \cdot L(\pi_1, \ldots, \pi_k; x_1, \ldots, x_k \mid n).
\end{aligned}
$$

Hence the distributions are equivalent.

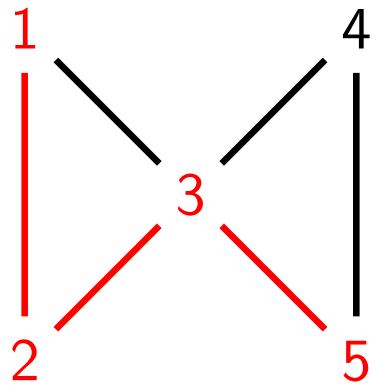
Undirected Graphical Models

Undirected Graphs



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

Paths



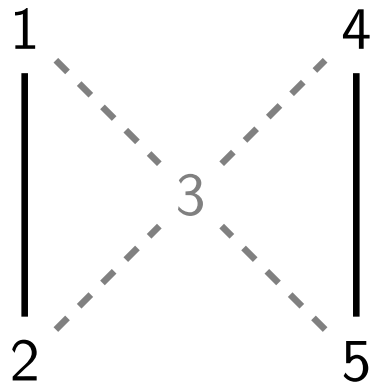
Paths:

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

$$\pi_2 : 3$$

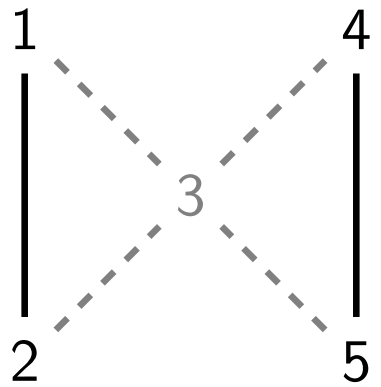
Note that paths may consist of one vertex and no edges. In this case it is a path of 'length 0'.

Induced Subgraph



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

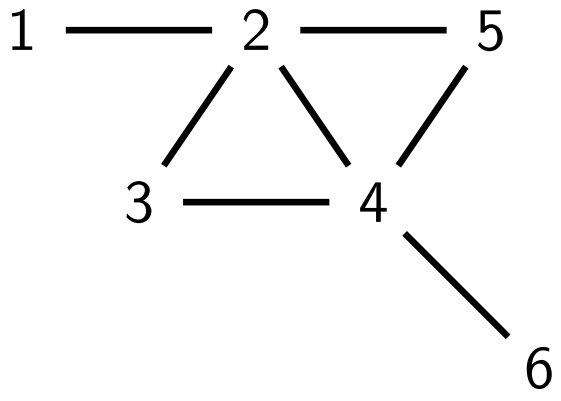
Separation



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

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

Cliques and Running Intersection



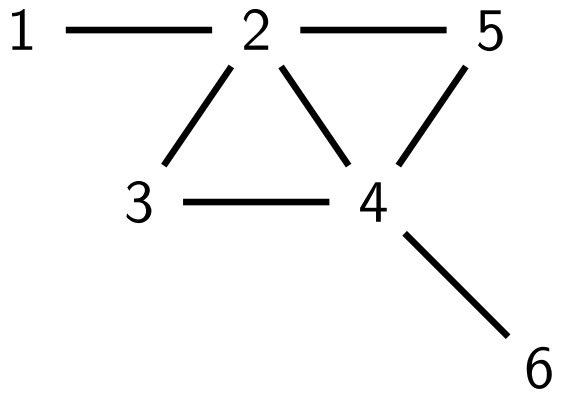
Cliques:

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

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:

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

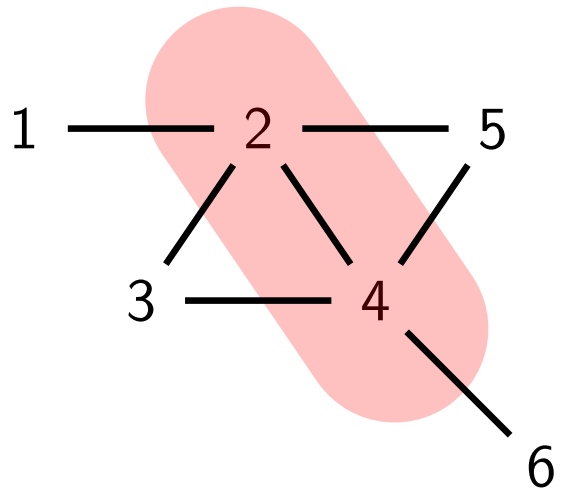
Any ordering works in this case as long $\{1,2\}$ and $\{4,6\}$ aren't the first two entries.

Estimation

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

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

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

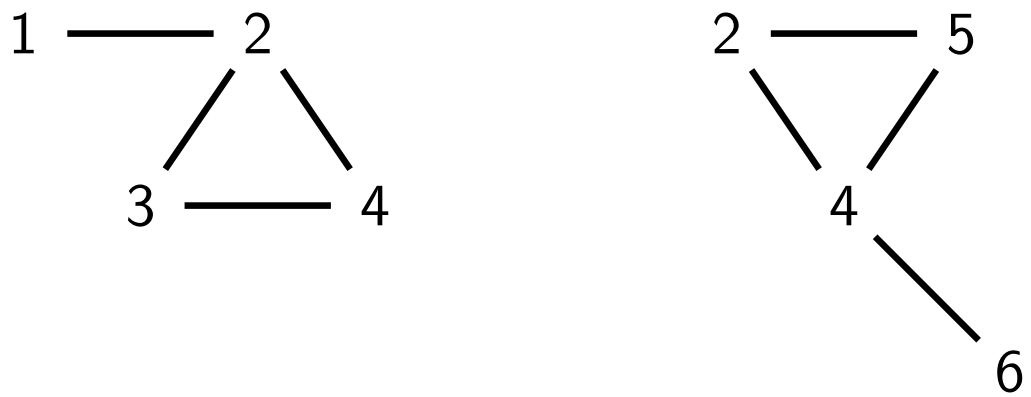


Estimation

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

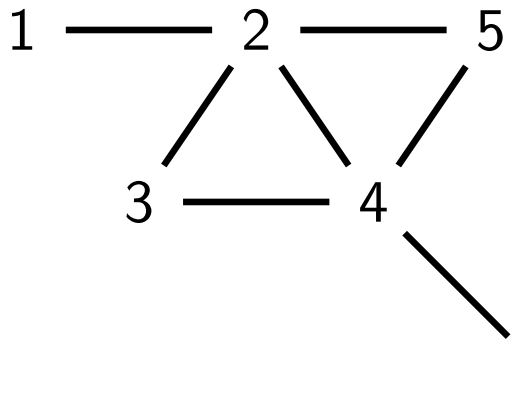
$$X_1, X_3 \perp\!\!\!\perp X_5, X_6 \mid X_2, X_4$$

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



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:

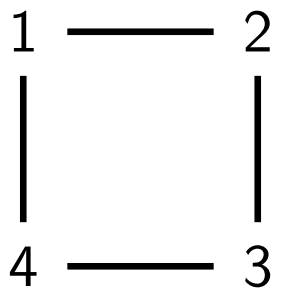
$$p(x_{123456}) \cdot p(x_{24}) \cdot p(x_2) \cdot p(x_4) = p(x_{12}) \cdot p(x_{234}) \cdot p(x_{245}) \cdot p(x_{46}).$$

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

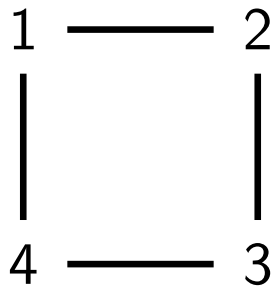
$$p(x_{1234}) = \psi_{12}(x_{12}) \cdot \psi_{23}(x_{23}) \cdot \psi_{34}(x_{34}) \cdot \psi_{14}(x_{14}),$$

the ψs are constrained by the requirement that

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

There 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 factors:

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

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

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 $p(x_{C_i}) \approx q(x_{C_i})$.
end function

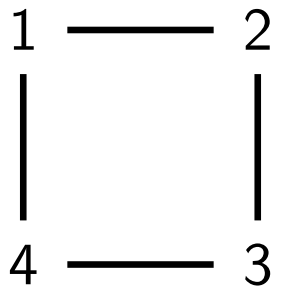
If any distribution satisfying $p(x_{C_i}) = q(x_{C_i})$ for each $i = 1, \ldots, k$ exists, then the algorithm converges to the **unique distribution** with those margins and which is Markov with respect to the graph with cliques $C_1, \ldots, C_k$.

Some Data

		$X_2 = 0$		$X_2 = 1$	
		$X_1 = 0$	1	0	1
$X_4 = 0$	$X_3 = 0$	9	9	0	8
	1	6	4	4	3
$X_4 = 1$	0	22	0	2	6
	1	5	3	10	5

Margins

Suppose we want to fit the 4-cycle model:



The relevant margins are:

$n(x_{12})$	$X_2 = 0$	1
$X_1 = 0$	42	16
1	16	22

$n(x_{23})$	$X_3 = 0$	1
$X_2 = 0$	40	18
1	16	22

$n(x_{34})$	$X_4 = 0$	1
$X_3 = 0$	26	30
1	17	23

$n(x_{14})$	$X_4 = 0$	1
$X_1 = 0$	19	39
1	24	14

Start with a Uniform Table

$p^{(0)}$		$X_2 = 0$		$X_2 = 1$	
		$X_1 = 0$	1	0	1
$X_4 = 0$	$X_3 = 0$	6	6	6	6
	1	6	6	6	6
$X_4 = 1$	0	6	6	6	6
	1	6	6	6	6

Set margin X_1, X_2 to correct value

$p^{(1)}$		$X_2 = 0$		$X_2 = 1$	
		$X_1 = 0$	1	0	1
$X_4 = 0$	$X_3 = 0$	10.5	4	4	5.5
	1	10.5	4	4	5.5
$X_4 = 1$	0	10.5	4	4	5.5
	1	10.5	4	4	5.5

Replace

$$p^{(1)}(x_1, x_2, x_3, x_4) = p^{(0)}(x_1, x_2, x_3, x_4) \cdot \frac{n(x_1, x_2)/n}{p^{(0)}(x_1, x_2)}$$

Set Margin X_2, X_3 to Correct Value

$p^{(2)}$		$X_2 = 0$		$X_2 = 1$	
		$X_1 = 0$	1	0	1
$X_4 = 0$	$X_3 = 0$	14.48	5.52	3.37	4.63
	1	6.52	2.48	4.63	6.37
$X_4 = 1$	0	14.48	5.52	3.37	4.63
	1	6.52	2.48	4.63	6.37

Replace

$$p^{(2)}(x_1, x_2, x_3, x_4) = p^{(1)}(x_1, x_2, x_3, x_4) \cdot \frac{n(x_2, x_3)/n}{p^{(1)}(x_2, x_3)}$$

Set Margin X_3, X_4 to Correct Value

$p^{(3)}$		$X_2 = 0$		$X_2 = 1$	
		$X_1 = 0$	1	0	1
$X_4 = 0$	$X_3 = 0$	13.45	5.12	3.13	4.3
	1	5.54	2.11	3.94	5.41
$X_4 = 1$	0	15.52	5.91	3.61	4.96
	1	7.49	2.86	5.33	7.32

Replace

$$p^{(3)}(x_1, x_2, x_3, x_4) = p^{(2)}(x_1, x_2, x_3, x_4) \cdot \frac{n(x_3, x_4)/n}{p^{(2)}(x_3, x_4)}$$

Set Margin X_1, X_4 to Correct Value

$p^{(4)}$		$X_2 = 0$		$X_2 = 1$	
		$X_1 = 0$	1	0	1
$X_4 = 0$	$X_3 = 0$	9.81	7.26	2.28	6.09
	1	4.04	2.99	2.87	7.67
$X_4 = 1$	0	18.94	3.93	4.41	3.3
	1	9.15	1.9	6.5	4.87

Replace

$$p^{(4)}(x_1, x_2, x_3, x_4) = p^{(3)}(x_1, x_2, x_3, x_4) \cdot \frac{n(x_1, x_4)/n}{p^{(3)}(x_1, x_4)}$$

Notice that sum of first column is now 41.94.

Set margin X_1, X_2 to correct value again

$p^{(5)}$		$X_2 = 0$		$X_2 = 1$	
		$X_1 = 0$	1	0	1
$X_4 = 0$	$X_3 = 0$	9.82	7.27	2.28	6.1
	1	4.02	2.97	2.86	7.63
$X_4 = 1$	0	18.87	3.92	4.39	3.29
	1	9.18	1.91	6.52	4.89

Replace

$$p^{(5)}(x_1, x_2, x_3, x_4) = p^{(4)}(x_1, x_2, x_3, x_4) \cdot \frac{n(x_1, x_2)/n}{p^{(4)}(x_1, x_2)}$$

Eventually...

$\hat{p}$		$X_2 = 0$		$X_2 = 1$	
		$X_1 = 0$	1	0	1
$X_4 = 0$	$X_3 = 0$	10.07	7.41	2.29	6.23
	1	3.87	2.85	2.77	7.51
$X_4 = 1$	0	18.7	3.83	4.26	3.22
	1	9.36	1.91	6.68	5.04

Waiting for this process to converge leads to the MLE.

$$p^{(5)}(x_1, x_2, x_3, x_4) = p^{(4)}(x_1, x_2, x_3, x_4) \cdot \frac{n(x_1, x_2)/n}{p^{(4)}(x_1, x_2)}$$

Notice that sum of first column is now ...

Gaussian Graphical Models

Multivariate Data

```
> library(ggm)
> data(marks)
> dim(marks)

[1] 88  5

> head(marks, 8)

  mechanics vectors algebra analysis statistics
1        77      82      67       67         81
2        63      78      80       70         81
3        75      73      71       66         81
4        55      72      63       70         68
5        63      63      65       70         63
6        53      61      72       64         73
7        51      67      65       65         68
8        59      70      68       62         56
```

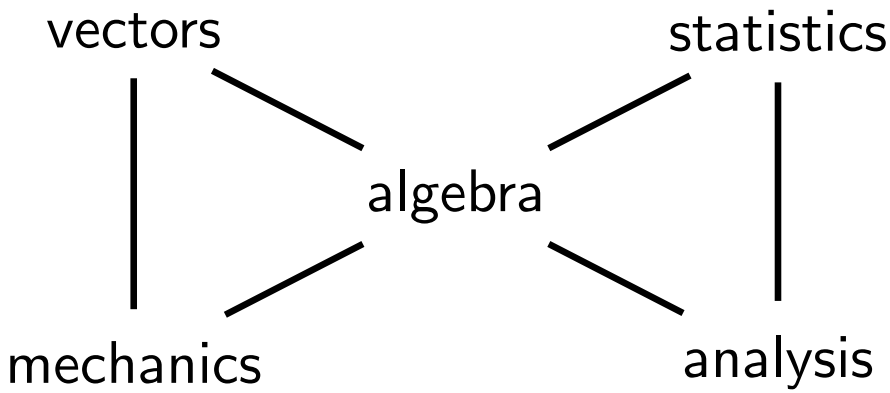
Multivariate Data

```
> sapply(marks, mean)

 mechanics    vectors    algebra   analysis statistics
      39.0       50.6       50.6       46.7       42.3

> cor(marks)

           mechanics vectors algebra analysis statistics
mechanics      1.000   0.553   0.547    0.409      0.389
vectors        0.553   1.000   0.610    0.485      0.436
algebra        0.547   0.610   1.000    0.711      0.665
analysis       0.409   0.485   0.711    1.000      0.607
statistics     0.389   0.436   0.665    0.607      1.000
```

Multivariate Data

```
> conc <- solve(cov(marks))  # concentration matrix
> round(1000*conc, 2)

           mechanics vectors algebra analysis statistics
mechanics       5.24   -2.44   -2.74     0.01      -0.14
vectors        -2.44   10.43   -4.71    -0.79      -0.17
algebra        -2.74   -4.71   26.95    -7.05      -4.70
analysis        0.01   -0.79   -7.05     9.88      -2.02
statistics     -0.14   -0.17   -4.70    -2.02       6.45
```

Undirected Graphs



	mech	vecs	alg	anlys	stats
mechanics	5.24	-2.43	-2.72	0.01	-0.15
vectors	-2.43	10.42	-4.72	-0.79	-0.16
algebra	-2.72	-4.72	26.94	-7.05	-4.70
analysis	0.01	-0.79	-7.05	9.88	-2.02
statistics	-0.15	-0.16	-4.70	-2.02	6.45

The Multivariate Gaussian Distribution

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

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

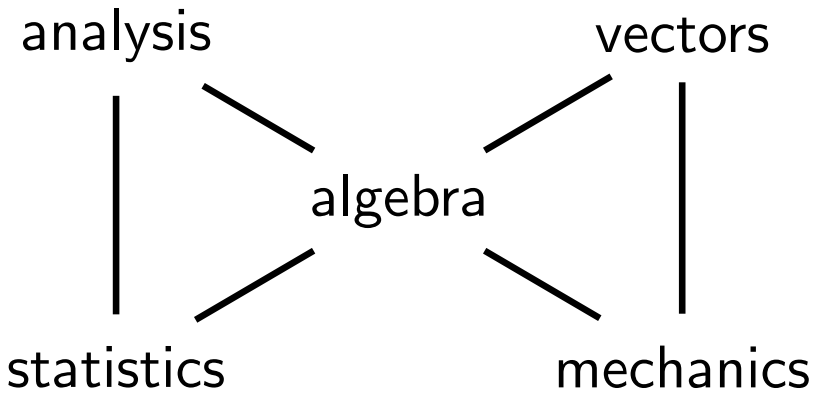
The log-likelihood for Σ is

$$l(\Sigma) = -\frac{n}{2} \log |\Sigma| - \frac{n}{2} \operatorname{tr}(S\Sigma^{-1})$$

where S is the sample covariance matrix, and this is maximized by choosing $\hat{\Sigma} = S$.

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.:

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

Equating terms gives:

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

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 Theorem 4.24 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

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

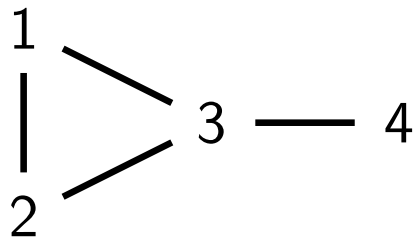
Example

```
> round(W, 3)           # sample covariance

       [,1]   [,2]   [,3]   [,4]
[1,]  1.158 -0.374 -0.242  0.036
[2,] -0.374  1.099  0.227 -0.065
[3,] -0.242  0.227  1.169 -0.378
[4,]  0.036 -0.065 -0.378  1.085

> round(solve(W), 3)    # sample concentration

      [,1]   [,2]   [,3]  [,4]
[1,] 0.995  0.308  0.160 0.040
[2,] 0.308  1.044 -0.138 0.004
[3,] 0.160 -0.138  1.026 0.344
[4,] 0.040  0.004  0.344 1.040
```

Note that these are fairly close to the true values.

Example

Fit the model with decomposition $(\{1,2\},\{3\},\{4\})$:




```
> K_hat = matrix(0, 4, 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,] 0.993  0.308  0.146 0.000
[2,] 0.308  1.044 -0.139 0.000
[3,] 0.146 -0.139  1.021 0.336
[4,] 0.000  0.000  0.336 1.038
```

Note this is close to the true concentration matrix.

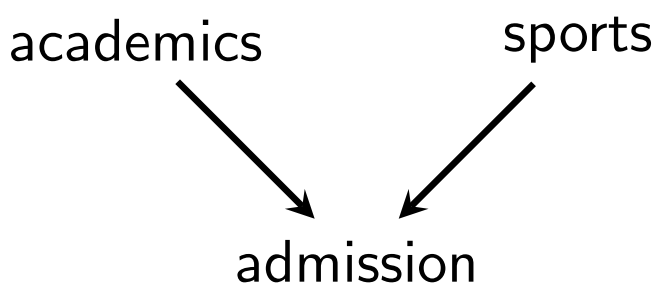
Directed Graphical Models

Marginal Independences

The graphs considered so far are all **undirected**.

Undirected graphs are very powerful, but they are also restrictive, in the sense that they cannot represent a **marginal** independence.

This rules out, for example, regression type models, where we might assume that some of the inputs are marginally independent.



A graph representing admission to the Holly-League School Yarvard.

Directed Graphs

Directed graphs give each edge an orientation.

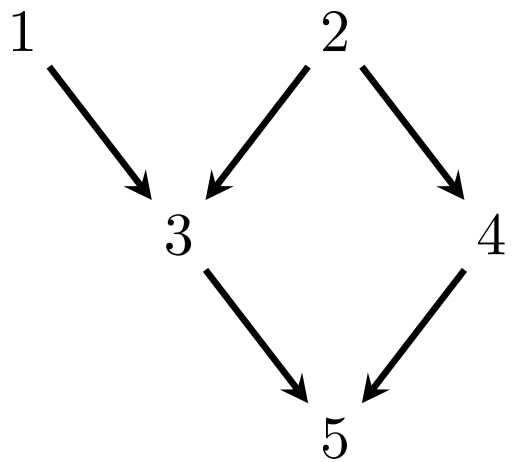
A **directed graph** $\mathcal{G}$ is a pair (V, D), where

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

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

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

If $i \to j$ or $i \leftarrow j$ we say i and j are **adjacent** and write $i \sim j$.



Acyclicity

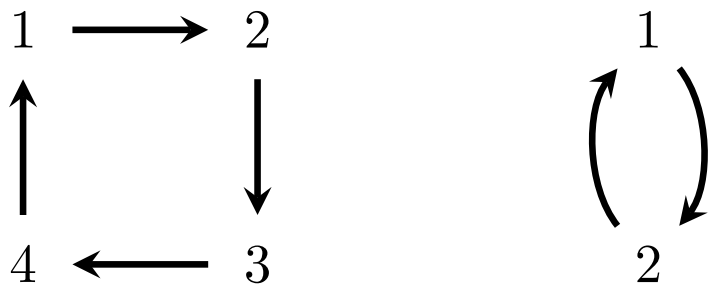
Paths are sequences of adjacent vertices, without repetition:

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

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

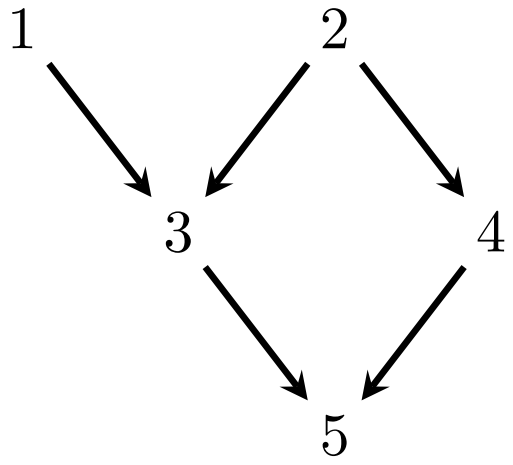
$$
\begin{array}{rl}
i \to j & \begin{cases} i \in \mathrm{pa}_{\mathcal{G}}(j) & i \text{ is a parent of } j \\ j \in \mathrm{ch}_{\mathcal{G}}(i) & j \text{ is a child of } i \end{cases} \\
a \to \cdots \to b \text{ or } a = b & \begin{cases} a \in \mathrm{an}_{\mathcal{G}}(b) & a \text{ is an ancestor of } b \\ b \in \mathrm{de}_{\mathcal{G}}(a) & b \text{ is a descendant of } a \end{cases}
\end{array}
$$

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



$$\mathrm{pa}_{\mathcal{G}}(3) = \{1, 2\}$$
$$\mathrm{ch}_{\mathcal{G}}(5) = \emptyset$$

$$\mathrm{an}_{\mathcal{G}}(4) = \{2, 4\}$$
$$\mathrm{de}_{\mathcal{G}}(1) = \{1, 3, 5\}$$
$$\mathrm{nd}_{\mathcal{G}}(1) = \{2, 4\}.$$

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:

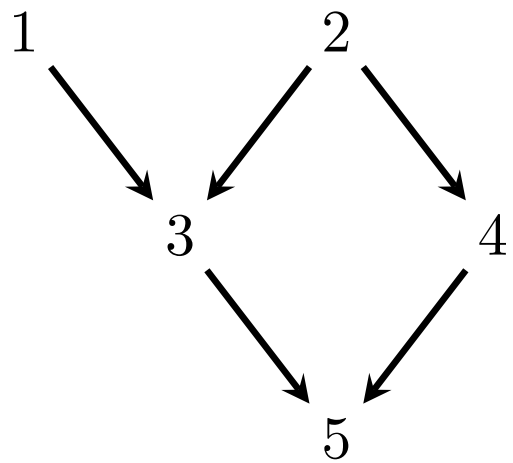
$1, 2, 3, 4, 5$

$1, 2, 4, 3, 5$

$2, 1, 3, 4, 5$

$2, 1, 4, 3, 5$

$2, 4, 1, 3, 5$



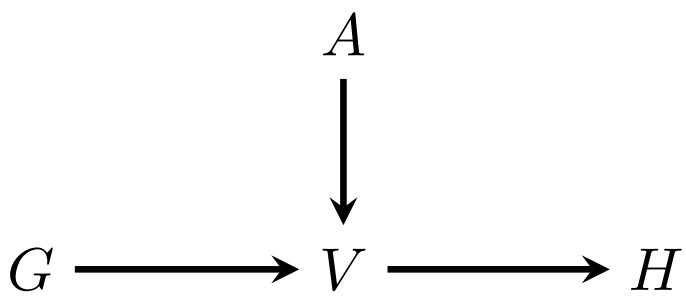
Parameter Estimation

G : group assigned to patient;

A : patient's age in years;

V : whether patient received flu vaccine;

H : patient hospitalized with respiratory problems;

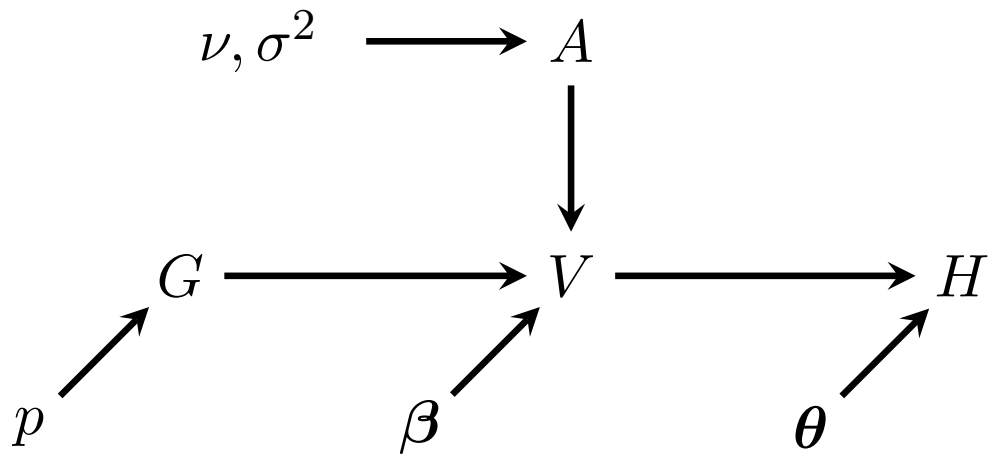


Parameter Estimation

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

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

Assuming independent priors:

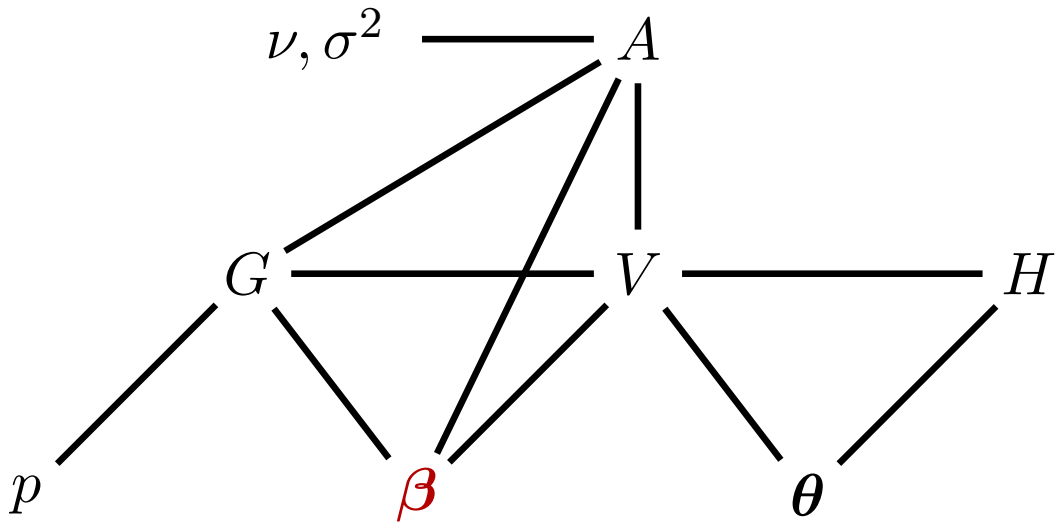


Bayesian Inference

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

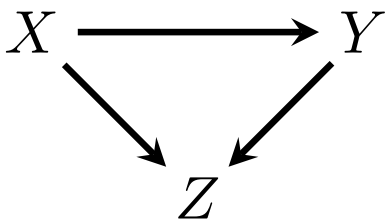


Markov Equivalence

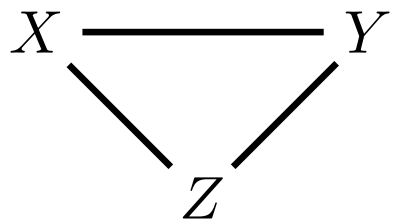
All undirected graphs induce distinct models.

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

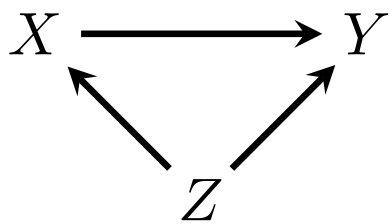
The same is not true for directed graphs:



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

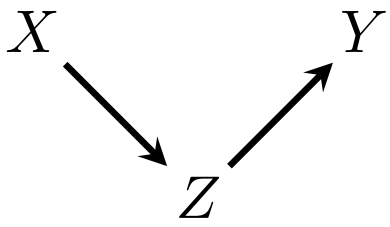


$$\psi_{XYZ}(x, y, z)$$



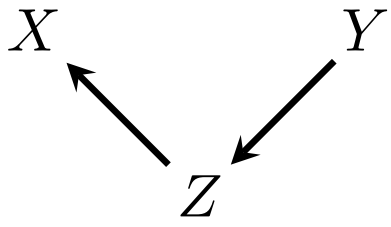
$$p(z) \cdot p(x \mid z) \cdot p(y \mid x, z)$$

Markov Equivalence



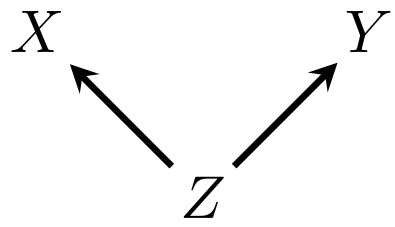
$$p(x) \cdot p(z \mid x) \cdot p(y \mid z)$$

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



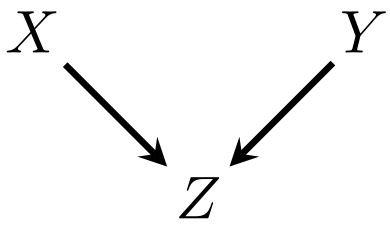
$$p(y) \cdot p(z \mid y) \cdot p(x \mid z)$$

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



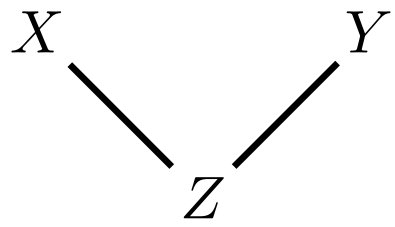
$$p(z) \cdot p(x \mid z) \cdot p(y \mid z)$$

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



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

$$X \perp\!\!\!\perp Y$$

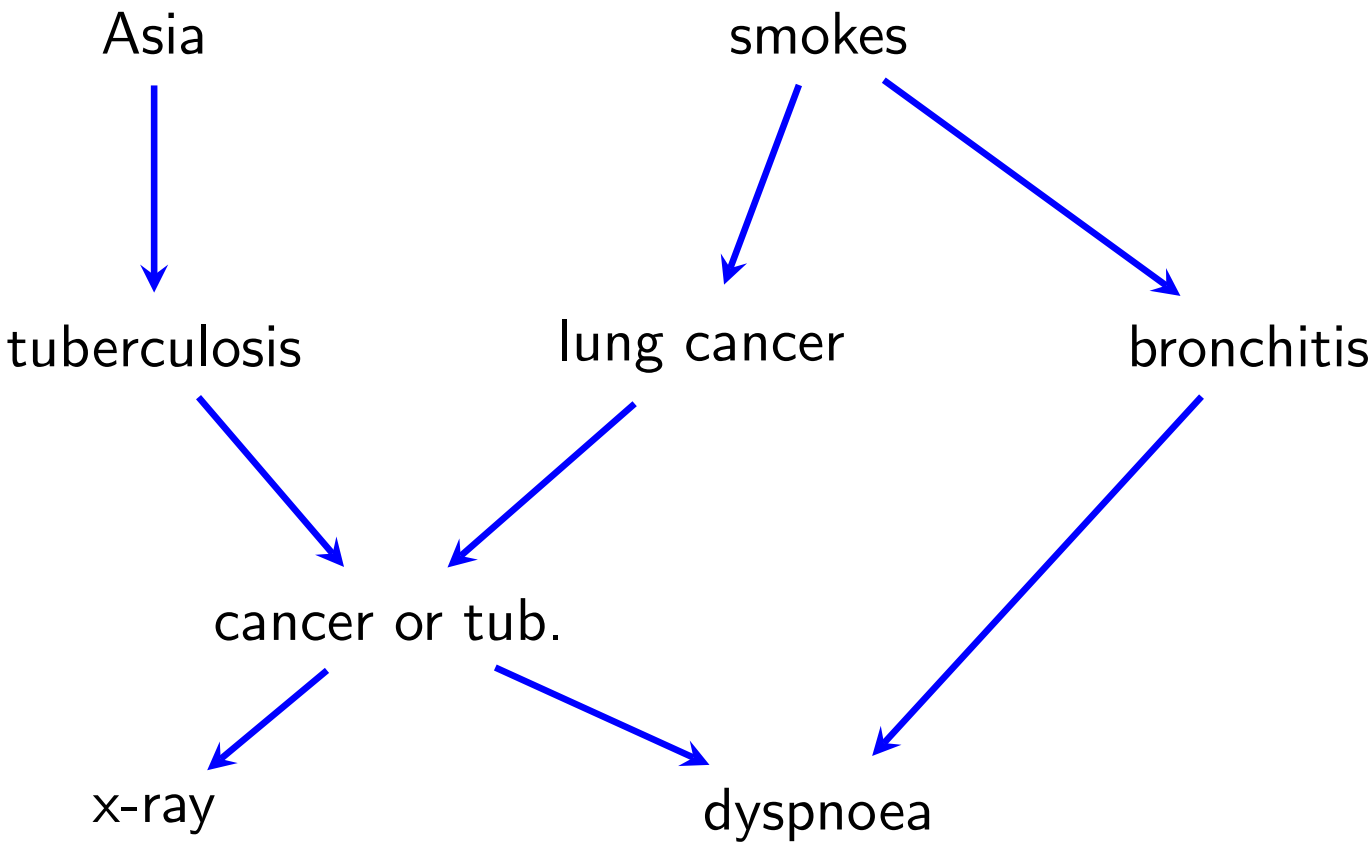


$$\psi_{XZ}(x, z) \cdot \psi_{YZ}(y, z)$$

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

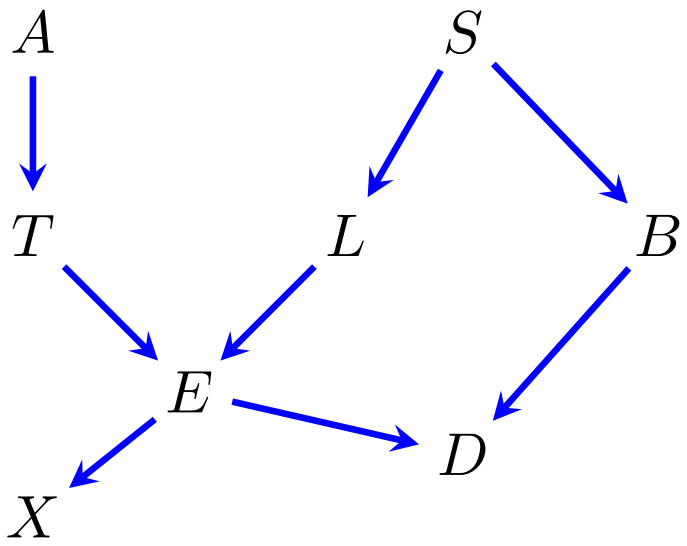
Expert Systems

Expert Systems



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

Variables



A has the patient recently visited southern Asia?

S does the patient smoke?

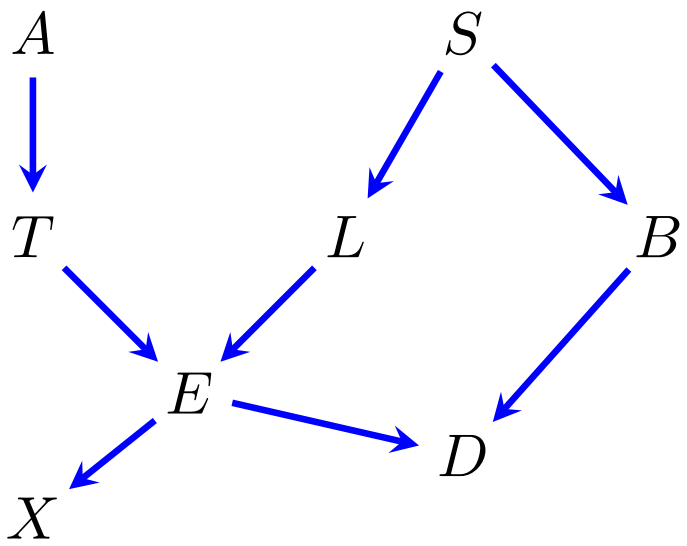
T,L,B tuberculosis, lung cancer, bronchitis.

E logical: tuberculosis OR lung cancer.

X shadow on chest X-ray?

D does the patient have dyspnoea?

Conditional Probability Tables



We have our factorization:

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

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

Probabilities

$p(a) =$

yes	no
0.01	0.99

$p(s) =$

yes	no
0.5	0.5

$p(t \mid a) =$

A	yes	no
yes	0.05	0.95
no	0.01	0.99

$p(\ell \mid s) =$

S	yes	no
yes	0.1	0.9
no	0.01	0.99

$p(b \mid s) =$

S	yes	no
yes	0.6	0.4
no	0.3	0.7

$p(x \mid e) =$

E	yes	no
yes	0.98	0.02
no	0.05	0.95

$p(d \mid b, e) =$

B	E	yes	no
yes	yes	0.9	0.1
	no	0.8	0.2
no	yes	0.7	0.3
	no	0.1	0.9

Factorizations

$$p(\ell \mid x, d, a, s) = \frac{p(\ell, x, d \mid a, s)}{\sum_{\ell'} p(\ell', x, d \mid a, s)}$$

From the graph $p(\ell, x, d \mid a, s)$ is

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

By this method there are up to 5×256 multiplications and $256 - 32 = 224$ additions.

This amounts to a complexity of around 1504 arithmetic operations.

Factorizations

But this is the same as:

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

Each large bracket requires 16 multiplications and 8 additions, and gives a vector of length 8.

Then the outer sum has 64 entries, so at most 128 multiplications and 32 additions.

This totals 208 arithmetic operations.

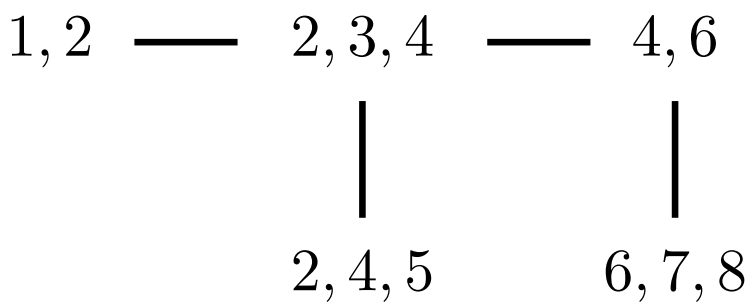
Junction Trees

Junction Trees

A **junction tree**:

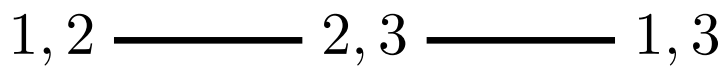
- is a (connected) undirected graph without cycles (a tree);
- has vertices C_i that consist of **subsets** of a set V;
- satisfies the property that if $C_i \cap C_j = S$ then every vertex on the (unique) path from C_i to C_j contains S.

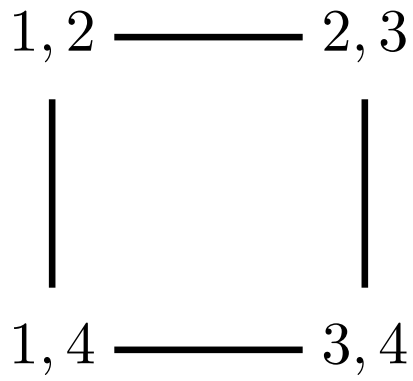
Example.



Junction Trees

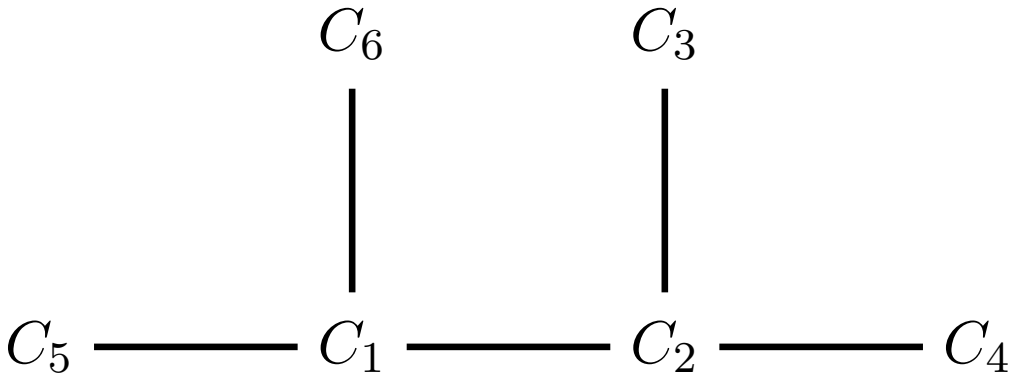
The following graphs are **not** junction trees:





Junction Trees

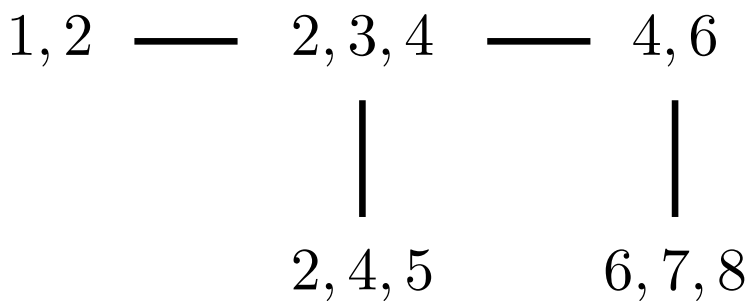
Junction trees can be constructed directly from sets of cliques satisfying running intersection.



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

Example: Junction Trees and RIP

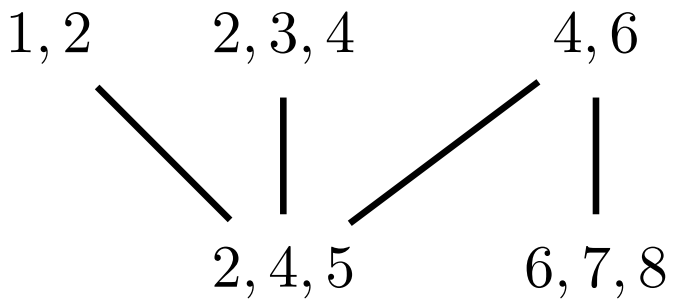
Given sets $\{1, 2\}$, $\{2, 3, 4\}$, $\{2, 4, 5\}$, $\{4, 6\}$, $\{6, 7, 8\}$, we can build this tree:



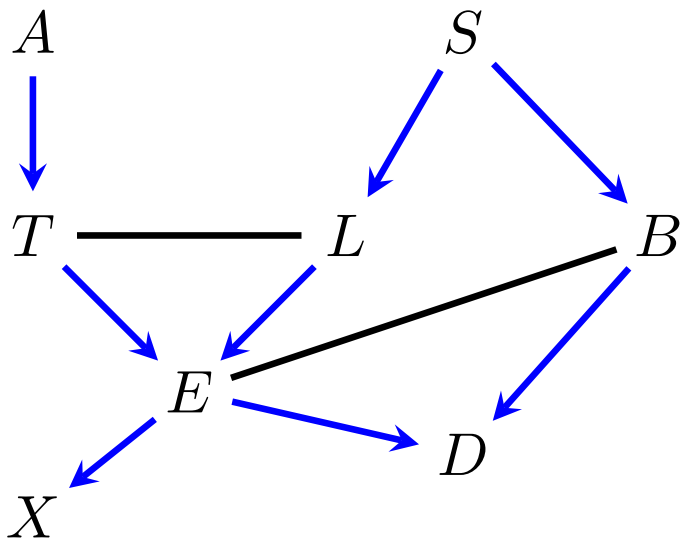
Example: Junction Trees and RIP

Equally, we could use a different ordering:

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



Forming A Junction Tree



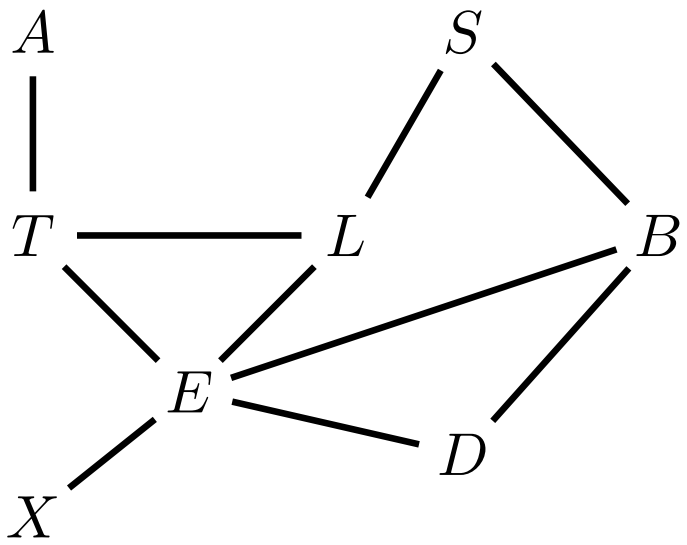
Steps to Forming a Junction Tree:

Moralize

Drop directions

Triangulate (add edges to get a decomposable graph)

Forming A Junction Tree



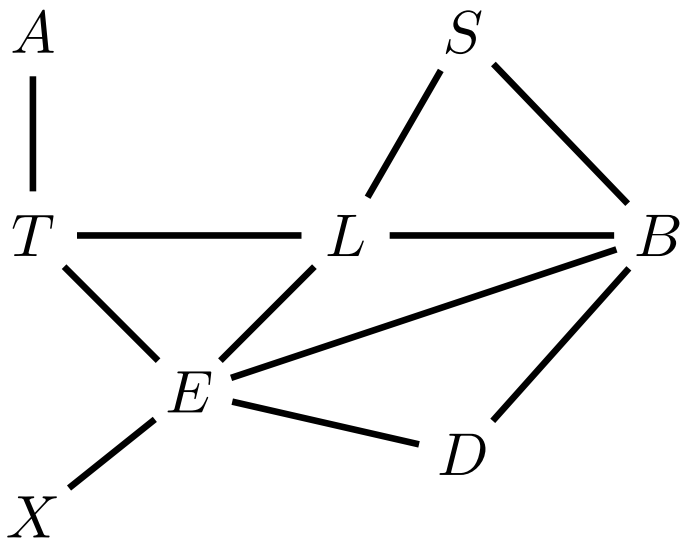
Steps to Forming a Junction Tree:

Moralize

Drop directions

Triangulate (add edges to get a decomposable graph)

Forming A Junction Tree



Steps to Forming a Junction Tree:

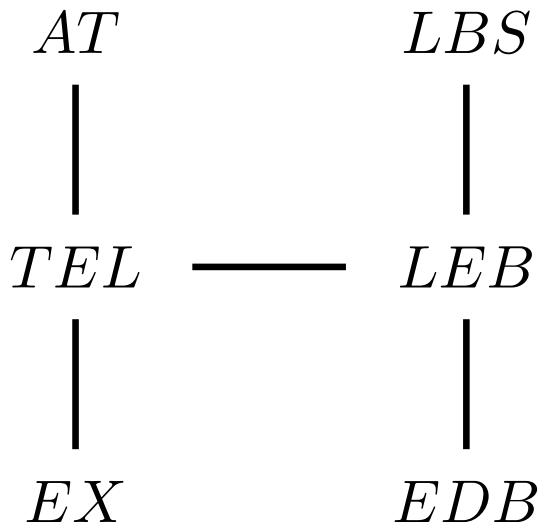
Moralize

Drop directions

Triangulate (add edges to get a decomposable graph)

Forming A Junction Tree

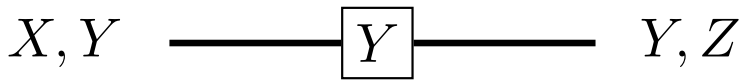
Finally, form the tree of cliques.



Message Passing

Updating / Message Passing

Suppose we have two vertices and one separator set.



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

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

$\psi_Y(y)$

$y = 0$	1
1	1

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

y \ $z = 0$	1
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

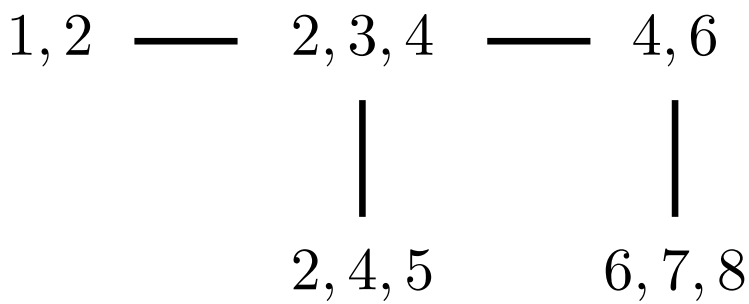
Pass message from Y, Z to X, Y. We set

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

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

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

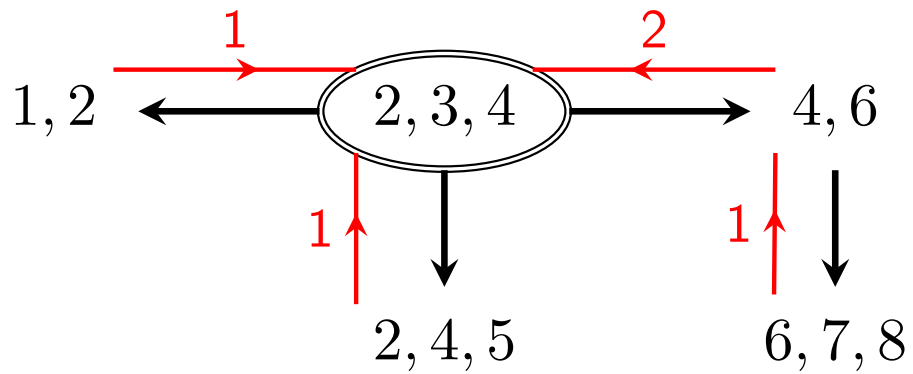
Rooting



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

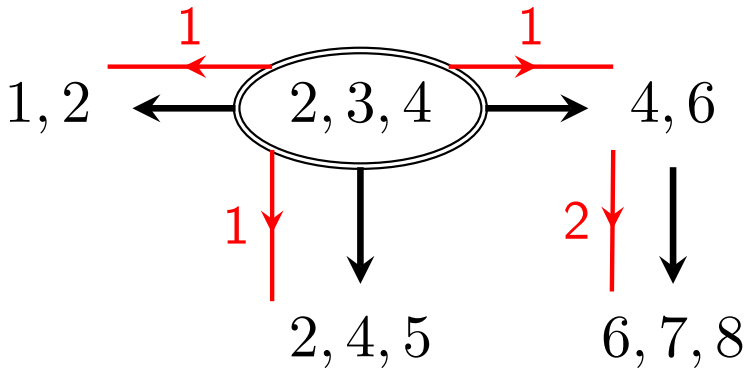
Collection and Distribution

function 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



Collection and Distribution

function 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



Initialization

$p(a) =$

yes	no
0.01	0.99

$p(s) =$

yes	no
0.5	0.5

$p(t \mid a) =$

A	yes	no
yes	0.05	0.95
no	0.01	0.99

$p(\ell \mid s) =$

S	yes	no
yes	0.1	0.9
no	0.01	0.99

$p(b \mid s) =$

S	yes	no
yes	0.6	0.4
no	0.3	0.7

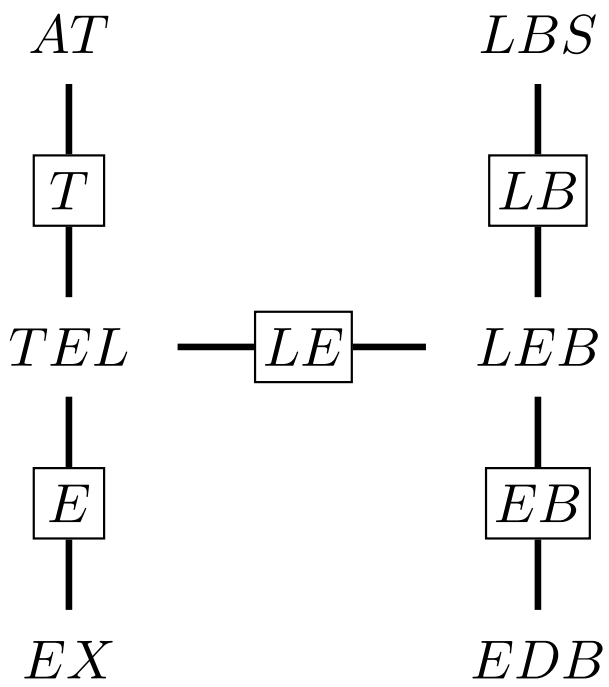
$p(x \mid e) =$

E	yes	no
yes	0.98	0.02
no	0.05	0.95

$p(d \mid b, e) =$

B	E	yes	no
yes	yes	0.9	0.1
	no	0.8	0.2
no	yes	0.7	0.3
	no	0.1	0.9

Initialization

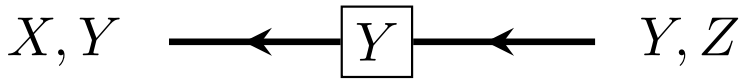


Can set, for example:

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

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

$$\psi_Y(y) = \sum_z \psi_{YZ}(y, z) = (0.6, 0.4);$$

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

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

From the Chest Clinic Network

Marginal Probability Tables:

$E \setminus X$	yes	no
yes	0.06	0
no	0.05	0.89

$A \setminus T$	yes	no
yes	0	0.01
no	0.01	0.98

B	yes		no	
$L \setminus S$	yes	no	yes	no
yes	0.03	0	0.02	0
no	0.27	0.15	0.18	0.35

E	yes		no	
$L \setminus B$	yes	no	yes	no
yes	0.03	0.02	0	0
no	0	0.01	0.41	0.52

E	yes		no	
$T \setminus L$	yes	no	yes	no
yes	0	0	0.01	0
no	0.05	0	0	0.94

E	yes		no	
$B \setminus D$	yes	no	yes	no
yes	0.03	0	0.02	0.01
no	0.33	0.08	0.05	0.47

From the Chest Clinic Network

Suppose now that we have a shadow on the chest X-ray:

$E \setminus X$	yes	no
yes	0.58	-
no	0.42	-

$A \setminus T$	yes	no
yes	0	0.01
no	0.09	0.9

B	yes		no	
$L \setminus S$	yes	no	yes	no
yes	0.27	0.01	0.18	0.03
no	0.15	0.08	0.1	0.19

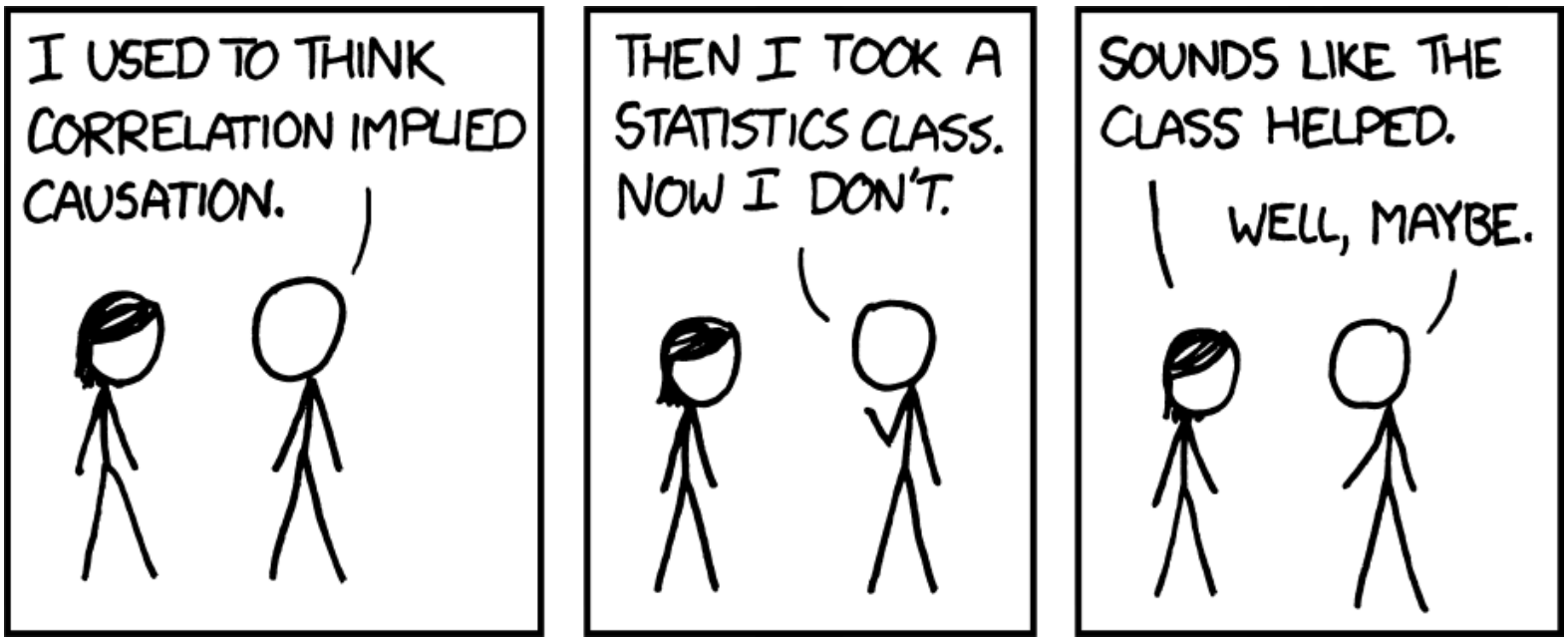
E	yes		no	
$L \setminus B$	yes	no	yes	no
yes	0.28	0.21	0	0
no	0.04	0.05	0.19	0.24

E	yes		no	
$T \setminus L$	yes	no	yes	no
yes	0.01	0	0.09	0
no	0.48	0	0	0.42

E	yes		no	
$B \setminus D$	yes	no	yes	no
yes	0.29	0.03	0.18	0.08
no	0.15	0.04	0.02	0.21

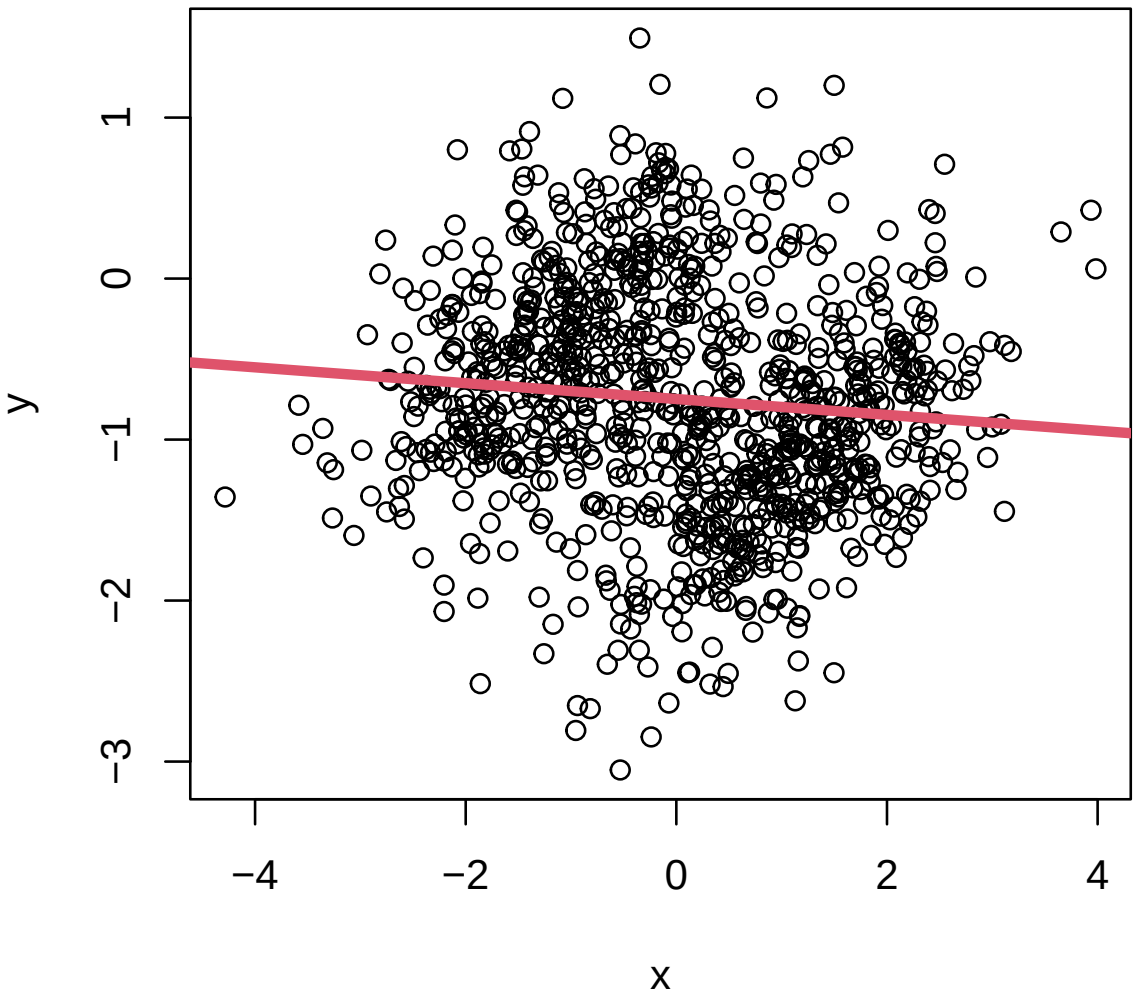
Causal Inference

Correlation

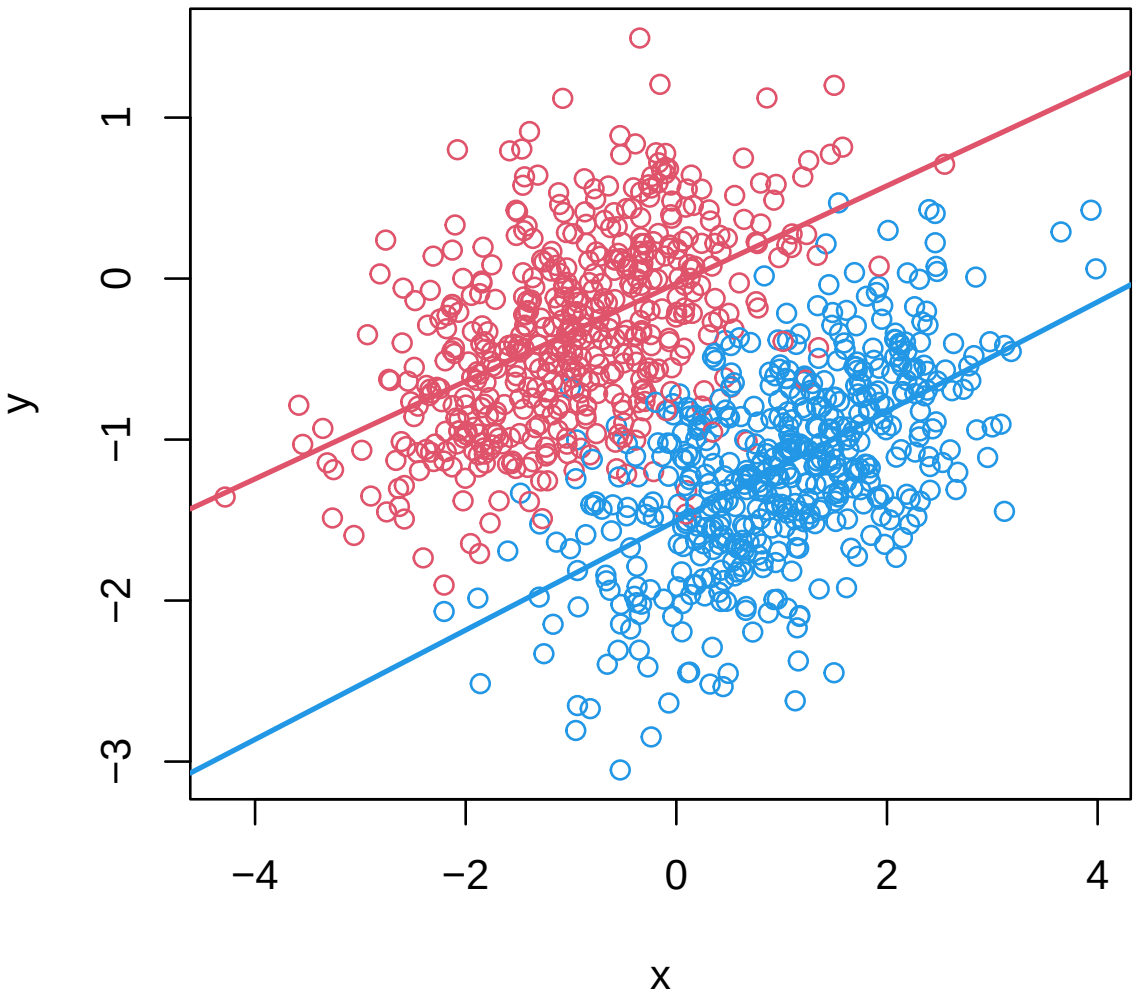


Credit: xkcd.com/552

Controlling for Covariates



Controlling for Covariates



Causation

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

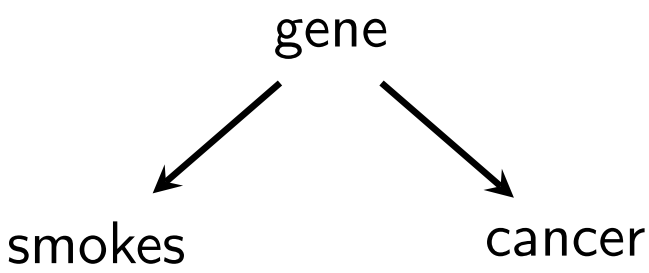
BUT: how do we know that this is a causal relationship? And what do we mean by that?

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?

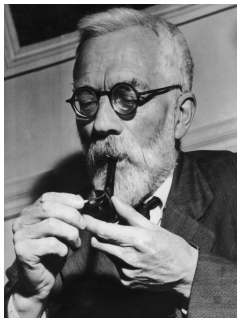
Alternative Explanations

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



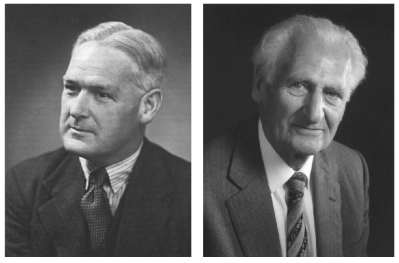
Confounding / Common Cause. There is a gene that makes people likely to smoke, and also more likely to get lung cancer.

Historical Causal Arguments



Ronald Fisher (who was a heavy smoker) disputed the idea that observational data could be used to **prove** that smoking is a cause of lung cancer. He offered other explanations (see previous slide!).

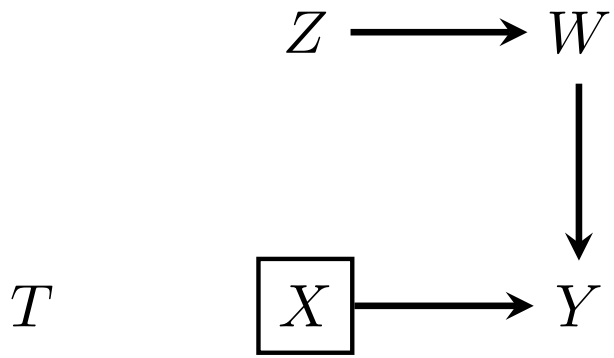
The great epidemiologists Austin Bradford Hill and Richard Doll published huge amounts of data noting the very strong associations.



Jerome Cornfield, an American biostatistician developed the *Cornfield inequality*, which notes that any confounding factor that could explain the association would have to be at least nine times more prevalent in smokers than nonsmokers.

Causal Models

A DAG model can also encode causal information:



If we intervene to experiment on X, just delete incoming edges.

In distribution, just delete factor corresponding to X:

$$p(t, z, w, x, y) = p(t) \cdot p(z) \cdot p(w \mid z) \cdot p(x \mid t, z) \cdot p(y \mid w, x).$$
$$p(t, z, w, y \mid do(x)) = p(t) \cdot p(z) \cdot p(w \mid z) \quad \times \quad p(y \mid w, x).$$

All other factors are preserved.

do-Calculus

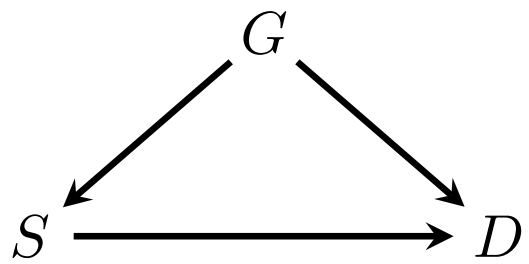
Note that (generally) $p(y \mid do(x)) \neq p(y \mid x)$ and $p(y \mid do(x)) \neq p(y)$.

It is **neither** a conditional **nor** an ordinary marginal distribution.

Causal Discovery is hard!

Determining which of the three explanations is correct is generally very hard, though methods do exist for distinguishing between such models.

Consider the following causal model, which we will assume is correct:



Here G is gender, S is smoking, and D is an indicator of lung damage.

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 estimate 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:

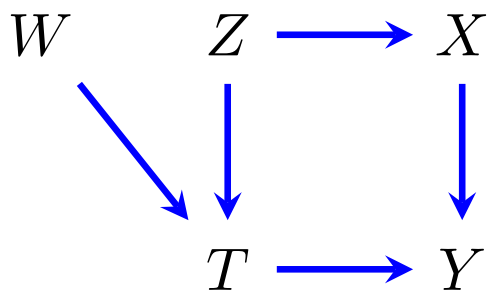
$$P(D = 1 \mid do(S = 1)) = \frac{32}{64} = \frac{1}{2}.$$

Adjustment

Causal and Non-Causal Paths

A directed path from T to Y is said to be **causal** for $T \to Y$.

Any other path is said to be **non-causal**.



Adjustment Using Parents

Note that we have

$$p(w, z, x, y \mid do(t)) = \frac{p(w, z, t, x, y)}{p(t \mid w, z)}.$$

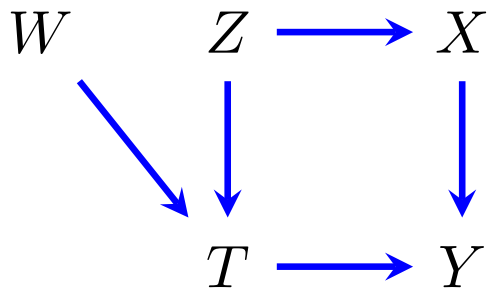
Hence, to obtain (e.g.) $p(y \mid do(t))$ we just marginalize:

$$\begin{aligned}\sum_{w,z,x} p(y, w, z, x \mid do(t)) &= \sum_{w,z,x} \frac{p(y, w, z, t, x)}{p(t \mid w, z)} \\ &= \sum_{w,z,x} p(w, z) \cdot p(x, y \mid t, z, w) \\ &= \sum_{w,z} p(w, z) \cdot p(y \mid t, z, w).\end{aligned}$$

In this case we call $\{W, Z\}$ an **adjustment set**.

The set of parents of a variable is **always** a valid adjustment set.

Back-Door Paths



A **back-door** path from T to Y starts with an arrowhead at T.

Example. $T \leftarrow Z \rightarrow X \rightarrow Y$.

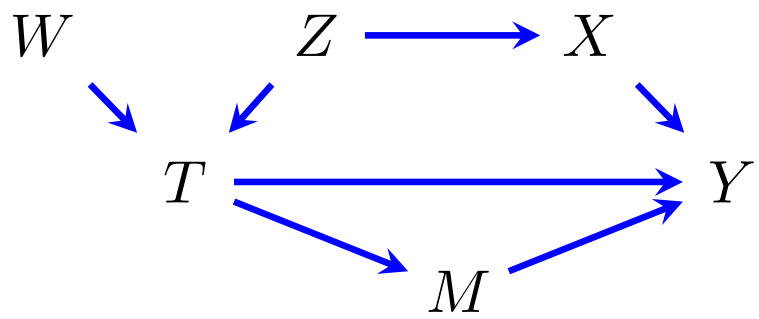
To estimate $p(y \mid do(t))$ we must block all back-door paths **without** blocking any causal ones, nor inducing any selection bias.

Back-Door Criterion

Definition

A **back-door adjustment set** for the pair (T, Y) is one which:

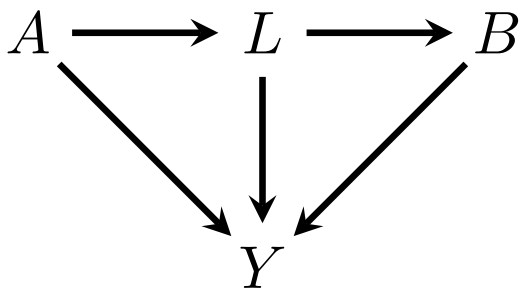
- blocks all back-door paths from T to Y;
- does not contain any descendants of T.



Examples:

$\{Z\}$, $\{X\}$, $\{Z, X\}$
$\{W, Z\}$, $\{W, X\}$, $\{W, Z, X\}$.

Example: HIV Treatment



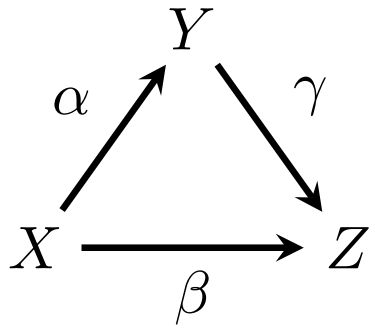
A treatment with AZT (an HIV drug);
L opportunisitic infection;
B treatment with antibiotics;
Y survival at 5 years.

$$p(a, \ell, b, y) = p(a) \cdot p(\ell \mid a) \cdot p(b \mid \ell) \cdot p(y \mid a, \ell, b)$$
$$p(\ell, y \mid do(a, b)) = p(\ell \mid a) \cdot p(y \mid a, \ell, b)$$
$$p(y \mid do(a, b)) = \sum_{\ell} p(\ell \mid a) \cdot p(y \mid a, \ell, b).$$

Structural Equation Models

Covariance Matrices

Let $\mathcal{G}$ be a DAG with variables V.



$$X = \varepsilon_x \qquad Y = \alpha X + \varepsilon_y \qquad Z = \beta X + \gamma Y + \varepsilon_z.$$

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ \alpha & 0 & 0 \\ \beta & \gamma & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} + \begin{pmatrix} \varepsilon_x \\ \varepsilon_y \\ \varepsilon_z \end{pmatrix}.$$

Covariance Matrices

Rearranging:

$$\begin{pmatrix} 1 & 0 & 0 \\ -\alpha & 1 & 0 \\ -\beta & -\gamma & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} \varepsilon_x \\ \varepsilon_y \\ \varepsilon_z \end{pmatrix}.$$

Now, you can check that:

$$(I - B)^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -\alpha & 1 & 0 \\ -\beta & -\gamma & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ \alpha & 1 & 0 \\ \beta + \alpha\gamma & \gamma & 1 \end{pmatrix},$$

so (recalling that $D = I$)

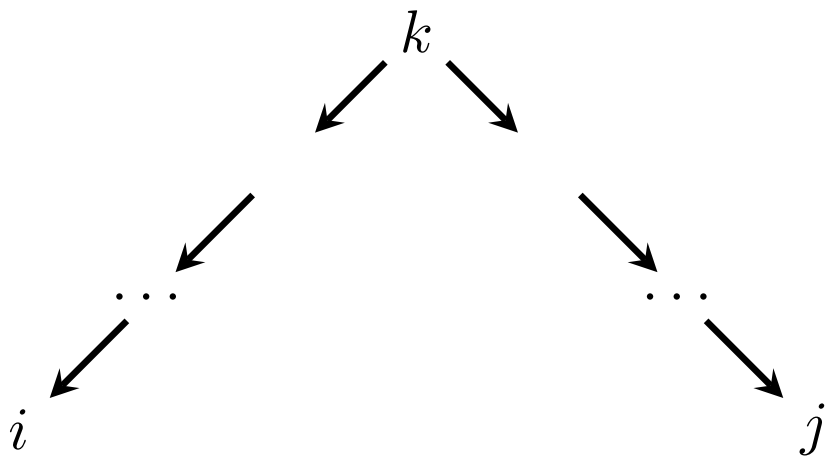
$$\begin{aligned} \Sigma &= (I - B)^{-1}(I - B)^{-T} \\ &= \begin{pmatrix} 1 & \alpha & \beta + \alpha\gamma \\ \alpha & 1 + \alpha^2 & \alpha\beta + \gamma + \alpha^2\gamma \\ \beta + \alpha\gamma & \alpha\beta + \gamma + \alpha^2\gamma & 1 + \gamma^2 + \beta^2 + 2\alpha\beta\gamma + \alpha^2\gamma^2 \end{pmatrix}. \end{aligned}$$

Treks

Let $\mathcal{G}$ be a DAG with variables V.

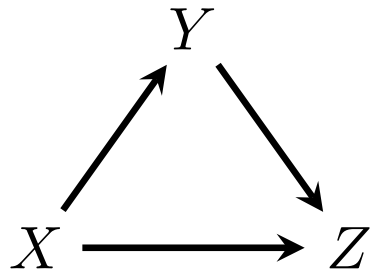
A **trek** from i to j with source k is a pair (π_l, π_r) of directed paths.

- π_l (the **left side**) is directed from k to i;
- π_r (the **right side**) is directed from k to j.



Trek Examples

Consider this DAG:



The treks from Z to Z are:

$$Z$$
$$Z \leftarrow X \rightarrow Z$$
$$Z \leftarrow X \rightarrow Y \rightarrow Z$$
$$Z \leftarrow Y \rightarrow Z$$
$$Z \leftarrow Y \leftarrow X \rightarrow Z$$
$$Z \leftarrow Y \leftarrow X \rightarrow Y \rightarrow Z.$$

Note that:

- A vertex may be in both the left and right sides.
- We may have $i = k$ or $j = k$ or both.

Treks

Let Σ be Markov with respect to a DAG $\mathcal{G}$, so that

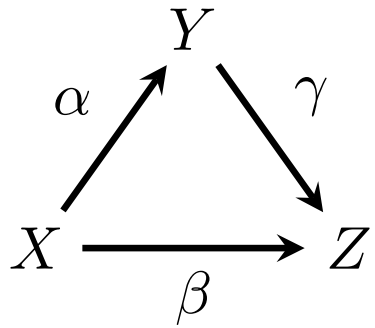
$$\Sigma = (I - B)^{-1} D (I - B)^{-T}.$$

Let $\tau = (\pi_l, \pi_r)$ be a trek with source k. The **trek covariance** associated with τ is:

$$c(\tau) = d_{kk} \left(\prod_{(i \to j) \in \pi_l} b_{ji} \right) \left(\prod_{(i \to j) \in \pi_r} b_{ji} \right).$$

Trek Covariance Examples

Consider this DAG:

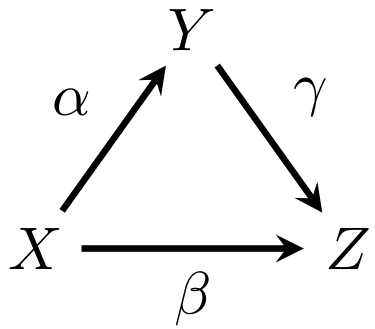


Trek covariances include:

$$c(Z) = 1 \qquad c(Z \leftarrow X) = \beta$$
$$c(Z \leftarrow X \rightarrow Y \rightarrow Z) = \beta \cdot \alpha \cdot \gamma \qquad c(Y \rightarrow Z) = \gamma.$$

Note that an empty product is 1 by convention.

Covariance Matrices



$$Z \qquad\qquad Z \leftarrow Y \rightarrow Z$$
$$Z \leftarrow X \rightarrow Z \qquad\qquad Z \leftarrow Y \leftarrow X \rightarrow Z$$
$$Z \leftarrow X \rightarrow Y \rightarrow Z \qquad\qquad Z \leftarrow Y \leftarrow X \rightarrow Y \rightarrow Z.$$

Recall that

$$\sigma_{zz} = 1 + \gamma^2 + \beta^2 + 2\alpha\beta\gamma + \alpha^2\gamma^2.$$

The Trek Rule

Theorem (8.29, The Trek Rule)

Let $\mathcal{G}$ be a DAG and let X_V be Gaussian and Markov with respect to $\mathcal{G}$. Then

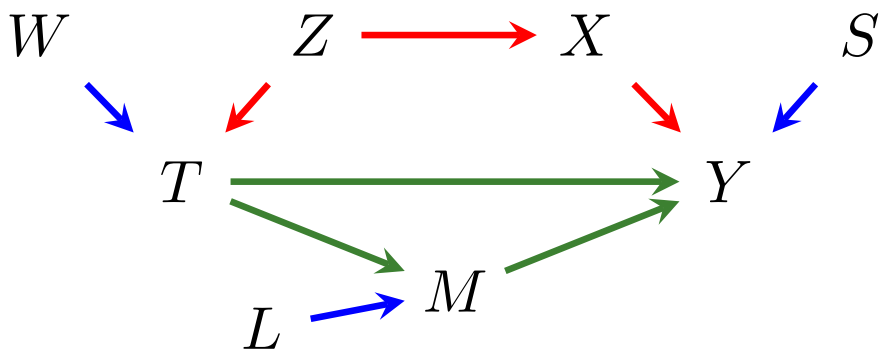
$$\sigma_{ij} = \sum_{\tau \in \mathcal{T}_{ij}} c(\tau),$$

where $\mathcal{T}_{ij}$ is the set of treks from i to j.

That is, the covariance between each X_i and X_j is the sum of the trek covariances over all treks between i and j.

Optimal Adjustment

Adjustment Sets



In this graph we:

- must leave causal paths open, so do **not** adjust for M (or T or Y);
- need to block back-door path, so must adjust for Z, X or both; (3)
- can decide whether to adjust for any of W, L, S. (8 possibilities)

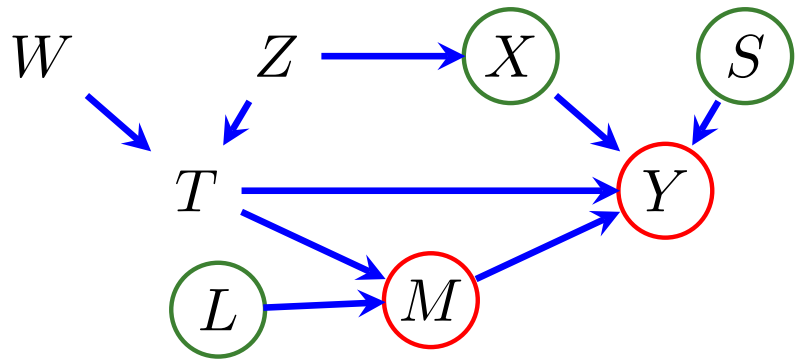
In total there are $3 \times 8 = 24$ possible adjustment sets.

Efficient Adjustment

Indeed, Rotnitzky and Smucler (2020) show that the most efficient adjustment set to use is:

$$\mathrm{pa}_{\mathcal{G}}(\mathrm{cn}_{\mathcal{G}}(T \to Y)) \setminus (\mathrm{cn}_{\mathcal{G}}(T \to Y) \cup \{T\}),$$

where $\mathrm{cn}_{\mathcal{G}}(T \to Y)$ is everything on a causal (i.e. directed) path from T to Y, excluding T itself.

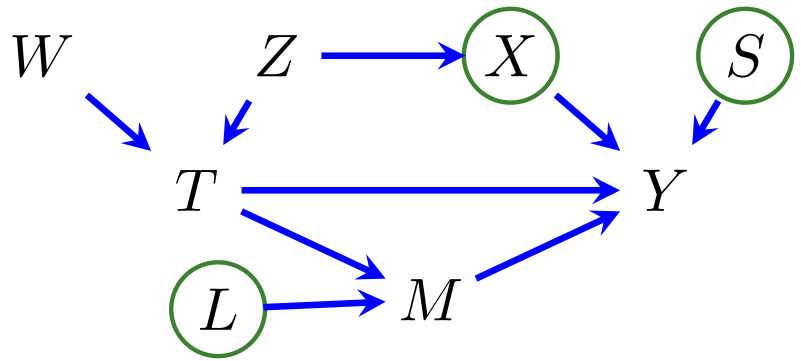


In our case $\mathrm{cn}_{\mathcal{G}}(T \to Y) = \{M, Y\}$.

And the optimal adjustment set is then parents of this set **not** on the causal path; i.e. $\boldsymbol{C} = \{X, L, S\}$.

Intuition behind Efficient Adjustment

Notice that we adjust for some unnecessary variables (L and S), even though these are not actually confounders.



Notice also that we **do not** control for instruments. (i.e. variables affecting only treatment).

In theory conditioning on an instrument will **increase** the variance in the estimate, because it **reduces** variance in X. In practice, conditioning on an instrument will also induce bias.

Intuition behind Efficient Adjustment

Think of effect estimation as a regression.

```
> T <- rnorm(100, sd=1)
> Y <- T + rnorm(100, sd=1)
> summary(lm(Y ~ T))$coef[,1:2]

            Estimate Std. Error
(Intercept)    -0.10      0.098
T               0.95      0.107

> T <- rnorm(100, sd=0.1)
> Y <- T + rnorm(100, sd=1)
> summary(lm(Y ~ T))$coef[,1:2]

            Estimate Std. Error
(Intercept)    -0.03       0.11
T               0.51       1.10
```

Reducing the variation in T **increases** the standard error.

Intuition behind Efficient Adjustment

Think of effect estimation as a regression.

```
> T <- rnorm(100, sd=1)
> Y <- T + rnorm(100, sd=1)
> summary(lm(Y ~ T))$coef[,1:2]

            Estimate Std. Error
(Intercept)    -0.10      0.098
T               0.95      0.107

> T <- rnorm(100, sd=1)
> Y <- T + rnorm(100, sd=0.1)
> summary(lm(Y ~ T))$coef[,1:2]

            Estimate Std. Error
(Intercept)   -0.003      0.011
T              0.995      0.011
```

However, reducing the variation in Y **decreases** the standard error.

More about Efficient Adjustment

The key quantity is:

$$\frac{\text{residual variance in } Y}{\text{residual variance in } T}.$$

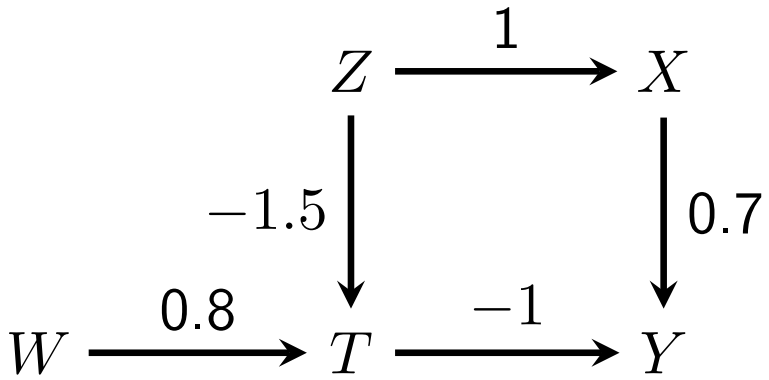
We want the top to be small and the bottom to be large for good precision.

The result giving $O_{\mathcal{G}}(T \to Y)$ was first proved in the multivariate Gaussian case by Henckel et al. (2019).

It was extended to the general case by Rotnitzky and Smulcer.

It has been extended further to models that also have hidden variables (we do not discuss these today, see Smucler et al., 2020).

Linear Gaussian Causal Models




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

Back-Door Paths

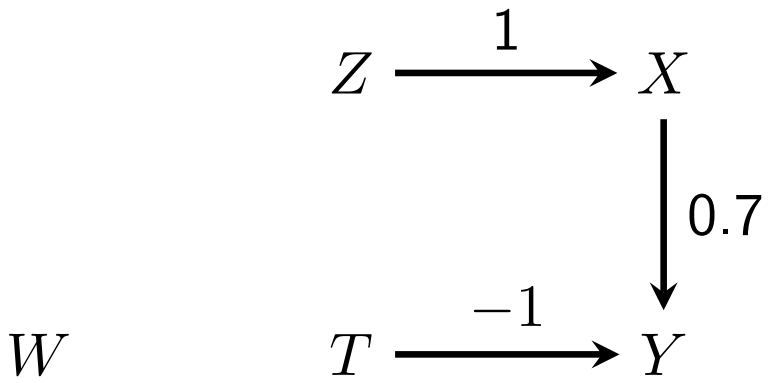
```
> summary(lm(Y ~ T))$coefficients[,1:2]

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

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

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

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

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

Instruments

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

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

            Estimate Std. Error
(Intercept)    0.044      0.038
T             -1.009      0.039
Z              0.665      0.070
W             -0.030      0.048

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

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

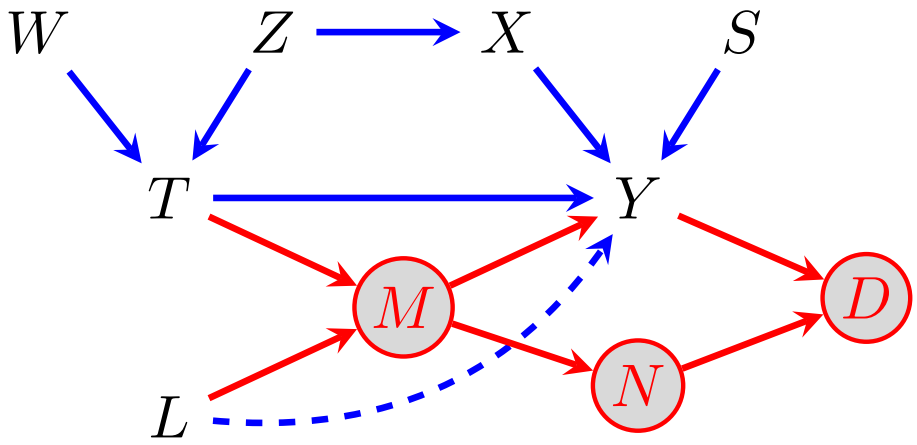
Simulating Intervention




```
> W <- rnorm(n)
> Z <- rnorm(n)
> X <- Z + rnorm(n)
> T <- rnorm(n, sd=sd(T))   # set T independently
> Y <- 0.7*X - T + rnorm(n)
> summary(lm(Y ~ T))$coefficients[,1:2]

            Estimate Std. Error
(Intercept) -0.00078      0.046
T           -1.03058      0.023
```

Forbidden Projection

Forbidden Projection



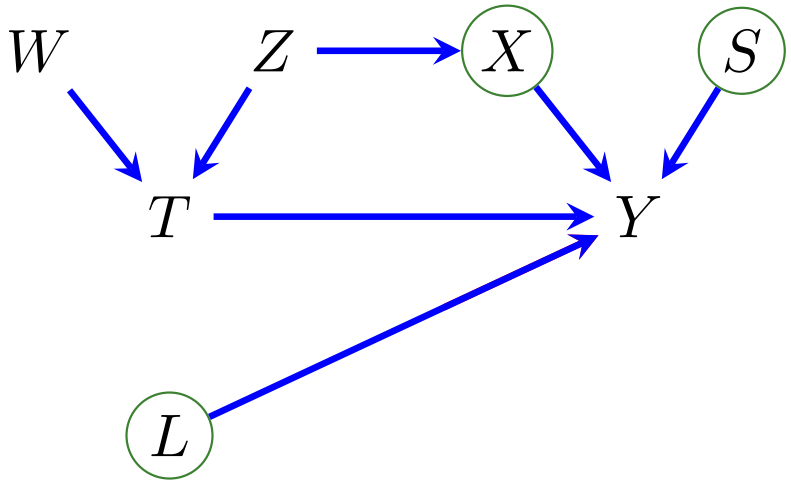
We have:

$$\mathrm{forb}_{\mathcal{G}}(T \to Y) = \{M, Y, N, D\},$$

so we must project out M, N and D.

Now note that $\mathrm{pa}_{\tilde{\mathcal{G}}}(Y) \setminus \{T\} = \{X, L, S\} = O_{\mathcal{G}}(T \to Y)$.

Forbidden Projection



We have:

$$\mathrm{forb}_{\mathcal{G}}(T \to Y) = \{M, Y, N, D\},$$

so we must project out M, N and D.

Now note that $\mathrm{pa}_{\tilde{\mathcal{G}}}(Y) \setminus \{T\} = \{X, L, S\} = O_{\mathcal{G}}(T \to Y)$.

Applied Example

The following DAG is taken from Arif and MacNeil (2022).