Data Mining 2013
Bayesian Networks (1)

Ad Feelders

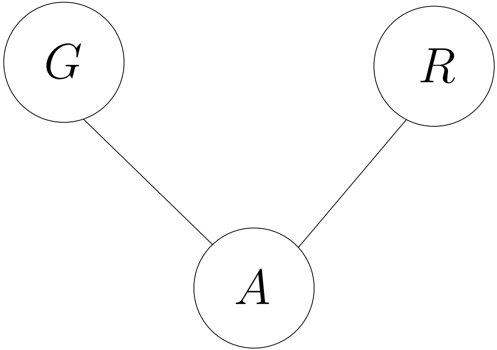
Universiteit Utrecht

October 17, 2013

Do you like noodles?

Race	Gender	Do you like noodles? Yes	No
Black	Male	32	86
	Female	35	121
White	Male	61	73
	Female	42	70

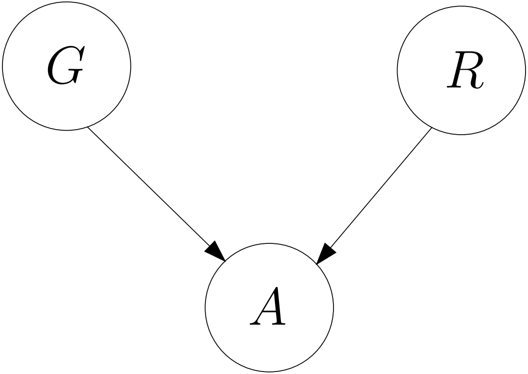
Do you like noodles? Undirected



$$G \perp\!\!\!\perp R \mid A$$

Strange: Gender and Race are prior to Answer, but this model says they are independent *given* Answer!

Do you like noodles? Directed



$$G \perp\!\!\!\perp R$$

Gender and Race are marginally independent (but *dependent* given Answer).

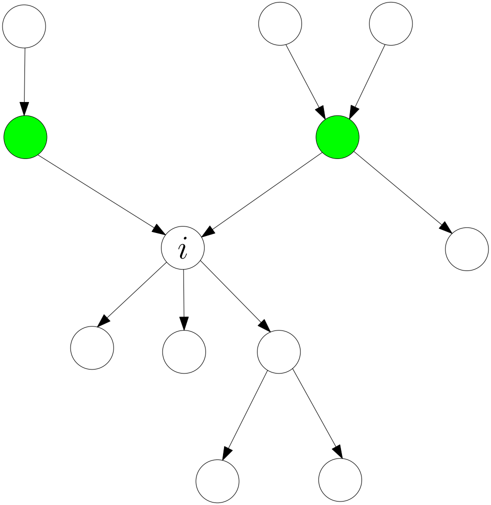
Directed Independence Graphs

$G = (K, E)$, K is a set of vertices and E is a set of edges with *ordered* pairs of vertices.

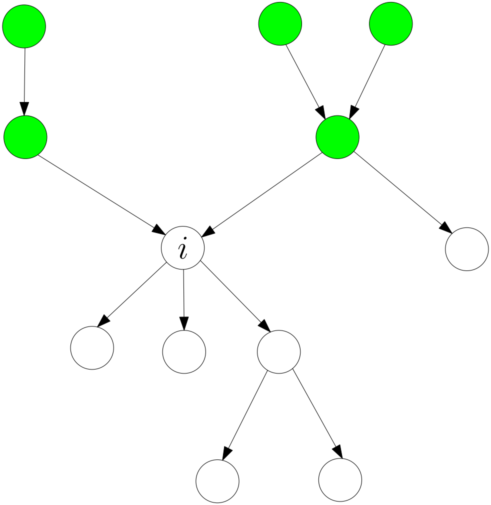
- No directed cycles (DAG)
- parent/child
- ancestor/descendant
- ancestral set

Because G is a DAG, there exists a *complete ordering* of the vertices that is respected in the graph (edges point from lower ordered to higher ordered nodes).

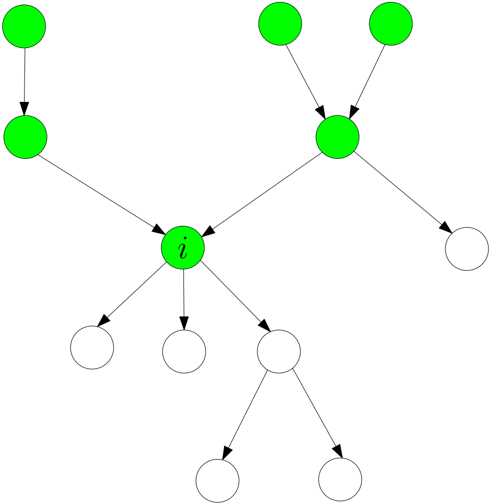
Parents Of Node i: pa(i)



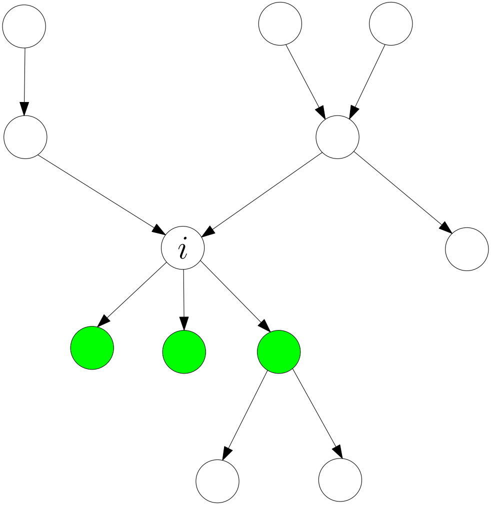
Ancestors Of Node i: an(i)



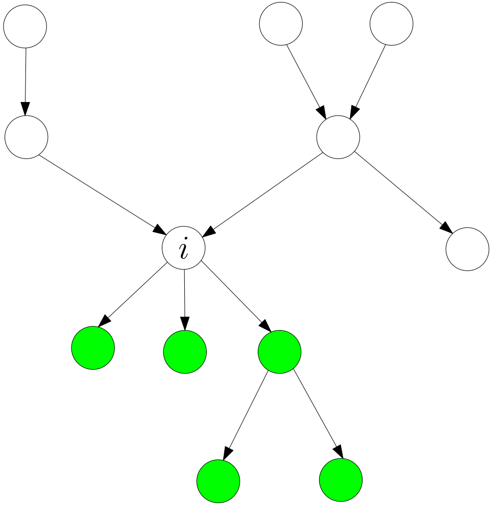
Ancestral Set Of Node i: $\text{an}^+(i)$



Children Of Node i: ch(i)



Descendants Of Node i: de(i)



Construction of DAG

Suppose that *prior* knowledge tells us the variables can be labeled $X_1, X_2, \ldots, X_k$ such that X_i is prior to X_{i+1}.
(for example: causal or temporal ordering)

Corresponding to this ordering we can use the product rule to factorize the joint distribution of $X_1, X_2, \ldots, X_k$ as

$$P(X) = P(X_1)P(X_2 \mid X_1) \cdots P(X_k \mid X_{k-1}, X_{k-2}, \ldots, X_1)$$

This is an identity of probability theory, no independence assumption have been made yet!

Constructing a DAG from pairwise independencies

In constructing a DAG, an arrow is drawn from i to j, where $i < j$, unless $P(X_j \mid X_{j-1}, \ldots, X_1)$ does not depend on X_i, in other words, unless

$$j \perp\!\!\!\perp i \mid \{1, \ldots, j\} \setminus \{i, j\}$$

More loosely

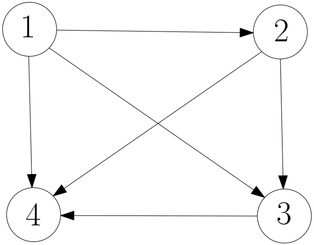
$$j \perp\!\!\!\perp i \mid \text{prior variables}$$

Compare this to pairwise independence

$$j \perp\!\!\!\perp i \mid \text{rest}$$

in undirected independence graphs.

Construction Of DAG

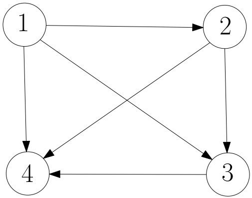


$$P(X) = P(X_4|X_1, X_2, X_3)P(X_3|X_1, X_2)P(X_2|X_1)P(X_1)$$

Suppose the following independencies are given:

1. $X_1 \perp\!\!\!\perp X_2$
2. $X_4 \perp\!\!\!\perp X_3|(X_1, X_2)$
3. $X_1 \perp\!\!\!\perp X_3|X_2$

Construction Of DAG

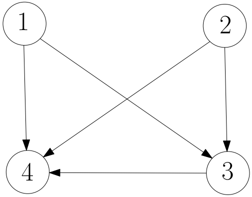


$$P(X) = P(X_4|X_1, X_2, X_3)P(X_3|X_1, X_2)\underbrace{P(X_2|X_1)}_{P(X_2)}P(X_1)$$

1. If $X_1 \perp\!\!\!\perp X_2$, then $P(X_2|X_1) = P(X_2)$.

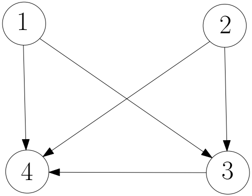
The edge $1 \rightarrow 2$ is removed.

Construction Of DAG



$$P(X) = P(X_4|X_1, X_2, X_3)P(X_3|X_1, X_2)P(X_2)P(X_1)$$

Construction Of DAG

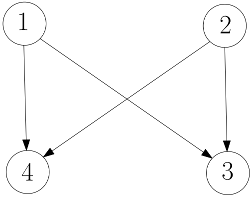


$$P(X) = \underbrace{P(X_4|X_1, X_2, X_3)}_{P(X_4|X_1,X_2)} P(X_3|X_1, X_2)P(X_2)P(X_1)$$

2. If $X_4 \perp\!\!\!\perp X_3|(X_1, X_2)$, then $P(X_4|X_1, X_2, X_3) = P(X_4|X_1, X_2)$.

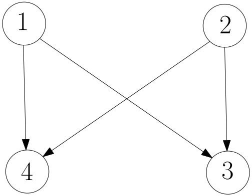
The edge $3 \rightarrow 4$ is removed.

Construction Of DAG



$$P(X) = P(X_4|X_1, X_2)P(X_3|X_1, X_2)P(X_2)P(X_1)$$

Construction Of DAG

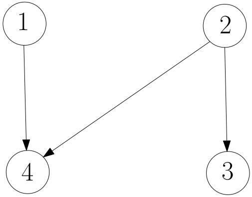


$$P(X) = P(X_4|X_1, X_2) \underbrace{P(X_3|X_1, X_2)}_{P(X_3|X_2)} P(X_2)P(X_1)$$

3. If $X_1 \perp\!\!\!\perp X_3|X_2$, then $P(X_3|X_1, X_2) = P(X_3|X_2)$

The edge $1 \rightarrow 3$ is removed.

Construction Of DAG



$$P(X) = P(X_4|X_1, X_2)P(X_3|X_2)P(X_2)P(X_1)$$

Joint density of Bayesian Network

We can write the joint density more elegantly as

$$P(X_1, \ldots, X_k) = \prod_{i=1}^{k} P(X_i \mid X_{pa(i)})$$

Independence Properties of DAGs: Moral Graph

Can we infer other/stronger independence statements from the directed graph like we did using separation in the undirected graphical models?

- d-separation (Pearl)
- make moral graph and use separation

Given a DAG $G = (K, E)$ we construct the moral graph G^m by marrying parents, and deleting directions, that is,

1. For each $i \in K$, we connect all vertices in $\text{pa}(i)$ with undirected edges.
2. We replace all directed edges in E with undirected ones.

Independence Properties of DAGs: Moral Graph

The directed independence graph G possesses the conditional independence properties of its associated moral graph G^m. Why? We have the factorisation:

$$P(X) = \prod_{i=1}^{k} P(X_i \mid X_{pa(i)})$$
$$= \prod_{i=1}^{k} g_i(X_i, X_{pa(i)})$$

by setting $g_i(X_i, X_{pa(i)}) = P(X_i \mid X_{pa(i)})$.

Independence Properties of DAGs: Moral Graph

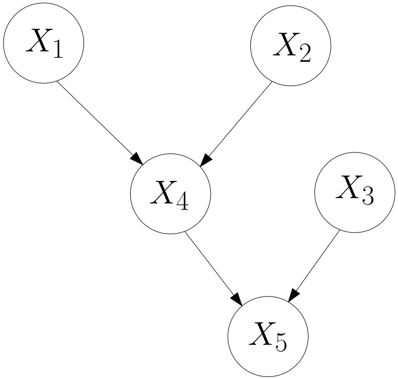
$$P(X) = \prod_{i=1}^{k} g_i(X_i, X_{pa(i)}) \qquad (1)$$

We thus have an expansion for the joint probability distribution in terms of functions $g(X_a)$ for $a = \{i\} \cup pa(i)$. Recall that $X \perp\!\!\!\perp Y \mid Z$ if and only if there exist functions g and h such that

$$P(x, y, z) = g(x, z)h(y, z)$$

By application of the factorisation criterion to the expansion (1), we can deduce all pairwise conditional independence statements of the form $i \perp\!\!\!\perp j \mid rest$.

Moralisation: Example



Moralisation: Example

This graph corresponds to the factorisation

$$\begin{aligned} P(X) &= P(X_1)P(X_2)P(X_3)P(X_4|X_1,X_2)P(X_5|X_3,X_4) \\ &= g_1(X_1)g_2(X_2)g_3(X_3)g_4(X_1,X_2,X_4)g_5(X_3,X_4,X_5) \end{aligned}$$

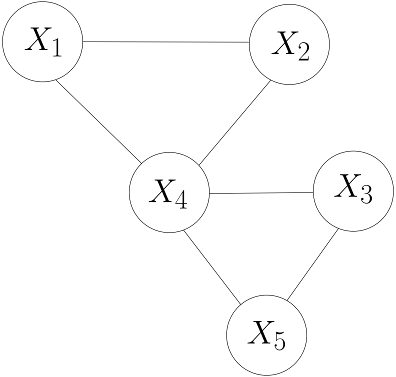
Log version:

$$\log P(X) = g_1^*(X_1) + g_2^*(X_2) + g_3^*(X_3) + g_4^*(X_1,X_2,X_4) + g_5^*(X_3,X_4,X_5),$$

where $g_i^*(X) = \log g_i(X)$.
We can read off the pairwise independencies:

$2 \perp\!\!\!\perp 3 \mid \textit{rest}$
$1 \perp\!\!\!\perp 3 \mid \textit{rest}$
$1 \perp\!\!\!\perp 5 \mid \textit{rest}$
$2 \perp\!\!\!\perp 5 \mid \textit{rest}$

Moralisation: Example



$\{i\} \cup pa(i)$ becomes a complete subgraph in the moral graph (by marrying all unmarried parents).

Moralisation Continued

Warning: the complete moral graph can obscure independencies!

To verify

$$i \perp\!\!\!\perp j \mid S$$

construct the moral graph on

$$A = \text{an}^+(\{i, j\} \cup S),$$

that is i, j, S and all their ancestors.

Moralisation Continued

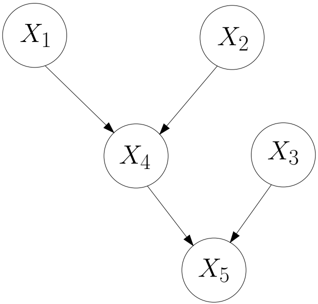
Since for $i \in A$, $pa(i) \in A$, we know that the joint distribution of X_A is given by

$$P(X_A) = \prod_{i \in A} P(X_i \mid X_{pa(i)})$$

which corresponds to the subgraph G_A of G.

1. This is a product of factors $P(X_i|X_{pa(i)})$, involving the variables $X_{\{i\} \cup pa(i)}$ only.
2. So it factorizes according to G_A^m, and thus the independence properties for undirected graphs apply.
3. So, if S separates i and j in G_A^m, then $i \perp\!\!\!\perp j \mid S$.

Moralisation Continued: example



Are X_3 and X_4 independent?

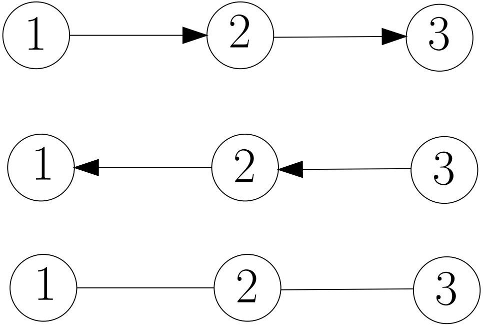
Are X_1 and X_3 independent?

Are X_3 and X_4 independent given X_5?

Equivalence

When no marrying of parents is required (there are no "v-structures"), then the independence properties of the directed graph are identical to those of its undirected version.

These three graphs express the same independence properties:



Learning Bayesian Networks: Overview

1. structure known, complete data
2. structure known, incomplete data
3. structure unknown, complete data
4. structure unknown, incomplete data: beyond the scope ...

Maximum Likelihood Estimation

Find value of unknown parameter(s) that maximize the probability of the observed data.

n independent observations on binary variable $X \in \{1, 2\}$. We observe $n(1)$ outcomes $X = 1$ and $n(2) = n - n(1)$ outcomes $X = 2$.
What is the maximum likelihood estimate of $p(1)$?
The likelihood function (probability of the data) is given by:

$$L = p(1)^{n(1)}(1 - p(1))^{n-n(1)}$$

Taking the log we get

$$\mathcal{L} = n(1) \log p(1) + (n - n(1)) \log(1 - p(1))$$

Maximum Likelihood Estimation

Take derivative with respect to $p(1)$, equate to zero, and solve for $p(1)$.

$$\frac{d\mathcal{L}}{dp(1)} = \frac{n(1)}{p(1)} - \frac{n - n(1)}{1 - p(1)} = 0,$$

since $\frac{d \log x}{dx} = \frac{1}{x}$ (where log is the natural logarithm).

Solving for $p(1)$, we get

$$p(1) = \frac{n(1)}{n},$$

i.e., the fraction of one's in the sample!

ML Estimation of Multinomial Distribution

Estimate the probabilities $p(1), p(2), \ldots, p(J)$ of getting outcomes $1, 2, \ldots, J$. If in n trials, we observe $n(1)$ outcomes of 1, $n(2)$ of 2, ..., $n(J)$ of J, then the obvious guess is to estimate

$$p(j) = \frac{n(j)}{n}, \qquad j = 1, \ldots, J$$

This is also the maximum likelihood estimate.

BN-Factorisation

For a given BN-DAG, the joint distribution factorises according to

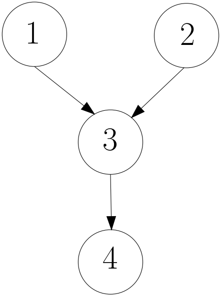
$$\prod_{i=1}^{k} p(X_i \mid X_{pa(i)})$$

So to specify the distribution we have to estimate the parameters

$$p(X_i \mid X_{pa(i)}) \qquad\qquad i = 1, 2, \ldots, k$$

The conditional distribution of each variable given its parents.

Example BN and Factorisation



$$P(X_1, X_2, X_3, X_4) = p_1(X_1)p_2(X_2)p_{3|12}(X_3|X_1, X_2)p_{4|3}(X_4|X_3)$$

Example BN: Parameters

$$P(X_1, X_2, X_3, X_4) = p_1(X_1)p_2(X_2)p_{3|12}(X_3|X_1, X_2)p_{4|3}(X_4|X_3)$$

Now we have to estimate the following parameters (X_4 ternary, rest binary):

$$\begin{array}{ll} p_1(1) & p_1(2) = 1 - p_1(1) \end{array}$$

$$\begin{array}{ll} p_2(1) & p_2(2) = 1 - p_2(1) \end{array}$$

$$\begin{array}{ll}
p_{3|1,2}(1|1,1) & p_{3|1,2}(2|1,1) = 1 - p_{3|1,2}(1|1,1) \\
p_{3|1,2}(1|1,2) & p_{3|1,2}(2|1,2) = 1 - p_{3|1,2}(1|1,2) \\
p_{3|1,2}(1|2,1) & p_{3|1,2}(2|2,1) = 1 - p_{3|1,2}(1|2,1) \\
p_{3|1,2}(1|2,2) & p_{3|1,2}(2|2,2) = 1 - p_{3|1,2}(1|2,2)
\end{array}$$

$$\begin{array}{lll}
p_{4|3}(1|1) & p_{4|3}(2|1) & p_{4|3}(3|1) = 1 - p_{4|3}(1|1) - p_{4|3}(2|1) \\
p_{4|3}(1|2) & p_{4|3}(2|2) & p_{4|3}(3|2) = 1 - p_{4|3}(1|2) - p_{4|3}(2|2)
\end{array}$$

Example Data Set

obs	X_1	X_2	X_3	X_4
1	1	1	1	1
2	1	1	1	1
3	1	1	2	1
4	1	2	2	1
5	1	2	2	2
6	2	1	1	2
7	2	1	2	3
8	2	1	2	3
9	2	2	2	3
10	2	2	1	3

Maximum Likelihood Estimation

obs	X_1	X_2	X_3	X_4
1	1	1	1	1
2	1	1	1	1
3	1	1	2	1
4	1	2	2	1
5	1	2	2	2
6	2	1	1	2
7	2	1	2	3
8	2	1	2	3
9	2	2	2	3
10	2	2	1	3

$$\hat{p}_1(1) = \frac{n(x_1 = 1)}{n} = \frac{5}{10} = \frac{1}{2}$$

Maximum Likelihood Estimation

obs	X_1	X_2	X_3	X_4
1	1	1	1	1
2	1	1	1	1
3	1	1	2	1
4	1	2	2	1
5	1	2	2	2
6	2	1	1	2
7	2	1	2	3
8	2	1	2	3
9	2	2	2	3
10	2	2	1	3

$$\hat{p}_2(1) = \frac{n(x_2 = 1)}{n} = \frac{6}{10}$$

Maximum Likelihood Estimation

obs	X_1	X_2	X_3	X_4
1	1	1	1	1
2	1	1	1	1
3	1	1	2	1
4	1	2	2	1
5	1	2	2	2
6	2	1	1	2
7	2	1	2	3
8	2	1	2	3
9	2	2	2	3
10	2	2	1	3

$$\hat{p}_{3|1,2}(1|1,1) = \frac{n(x_1 = 1, x_2 = 1, x_3 = 1)}{n(x_1 = 1, x_2 = 1)} = \frac{2}{3}$$

Maximum Likelihood Estimation

obs	X_1	X_2	X_3	X_4
1	1	1	1	1
2	1	1	1	1
3	1	1	2	1
4	1	2	2	1
5	1	2	2	2
6	2	1	1	2
7	2	1	2	3
8	2	1	2	3
9	2	2	2	3
10	2	2	1	3

$$\hat{p}_{3|1,2}(1|1,1) = \frac{n(x_1 = 1, x_2 = 1, x_3 = 1)}{n(x_1 = 1, x_2 = 1)} = \frac{2}{3}$$

ML Estimation of BN

The joint probability for n independent observations is

$$
\begin{aligned}
P(X^{(1)}, \ldots, X^{(n)}) &= \prod_{j=1}^{n} P(X^{(j)}) \\
&= \prod_{j=1}^{n} \prod_{i=1}^{k} p(X_i^{(j)} \mid X_{pa(i)}^{(j)})
\end{aligned}
$$

ML Estimation of BN

The likelihood function is thus given by

$$L = \prod_{i=1}^{k} \prod_{x_i, x_{pa(i)}} p(x_i \mid x_{pa(i)})^{n(x_i, x_{pa(i)})}$$

where $n(x_i, x_{pa(i)})$ is a count of the number of records with $X_i = x_i$, and $X_{pa(i)} = x_{pa(i)}$.

ML Estimation of BN

Taking the log of the likelihood, we get

$$\mathcal{L} = \sum_{i=1}^{k} \sum_{x_i, x_{pa(i)}} n(x_i, x_{pa(i)}) \log p(x_i \mid x_{pa(i)})$$

This is a collection of independent multinomial estimation problems.

The maximum likelihood estimate of $p(x_i \mid x_{pa(i)})$ is:

$$\hat{p}(x_i \mid x_{pa(i)}) = \frac{n(x_i, x_{pa(i)})}{n(x_{pa(i)})}$$

$n(x_i, x_{pa(i)})$: number of records in data with $X_i = x_i$ and $X_{pa(i)} = x_{pa(i)}$.

$n(x_{pa(i)})$: number of records in data with $X_{pa(i)} = x_{pa(i)}$.

Data Set and Likelihood

$$P(X_1, X_2, X_3, X_4) = p_1(X_1)p_2(X_2)p_{3|12}(X_3|X_1, X_2)p_{4|3}(X_4|X_3)$$

obs	X_1	X_2	X_3	X_4
1	1	1	1	1
2	1	1	1	1
3	1	1	2	1
4	1	2	2	1
5	1	2	2	2
6	2	1	1	2
7	2	1	2	3
8	2	1	2	3
9	2	2	2	3
10	2	2	1	3

$p_1(1)p_2(1)p_{3|12}(1|1,1)p_{4|3}(1|1)$

Data Set and Likelihood

$$P(X_1, X_2, X_3, X_4) = p_1(X_1)p_2(X_2)p_{3|12}(X_3|X_1, X_2)p_{4|3}(X_4|X_3)$$

obs	X_1	X_2	X_3	X_4
1	1	1	1	1
2	1	1	1	1
3	1	1	2	1
4	1	2	2	1
5	1	2	2	2
6	2	1	1	2
7	2	1	2	3
8	2	1	2	3
9	2	2	2	3
10	2	2	1	3

$p_1(1)p_2(1)p_{3|12}(1|1,1)p_{4|3}(1|1)$

$p_1(1)p_2(1)p_{3|12}(1|1,1)p_{4|3}(1|1)$

Data Set and Likelihood

$$P(X_1, X_2, X_3, X_4) = p_1(X_1)p_2(X_2)p_{3|12}(X_3|X_1, X_2)p_{4|3}(X_4|X_3)$$

obs	X_1	X_2	X_3	X_4	
1	1	1	1	1	$p_1(1)p_2(1)p_{3\|12}(1\|1, 1)p_{4\|3}(1\|1)$
2	1	1	1	1	$p_1(1)p_2(1)p_{3\|12}(1\|1, 1)p_{4\|3}(1\|1)$
3	1	1	2	1	$p_1(1)p_2(1)p_{3\|12}(2\|1, 1)p_{4\|3}(1\|2)$
4	1	2	2	1	
5	1	2	2	2	
6	2	1	1	2	
7	2	1	2	3	
8	2	1	2	3	
9	2	2	2	3	
10	2	2	1	3	

Data Set and Likelihood

$$P(X_1, X_2, X_3, X_4) = p_1(X_1)p_2(X_2)p_{3|12}(X_3|X_1, X_2)p_{4|3}(X_4|X_3)$$

obs	X_1	X_2	X_3	X_4	
1	1	1	1	1	$p_1(1)p_2(1)p_{3\|12}(1\|1,1)p_{4\|3}(1\|1)$
2	1	1	1	1	$p_1(1)p_2(1)p_{3\|12}(1\|1,1)p_{4\|3}(1\|1)$
3	1	1	2	1	$p_1(1)p_2(1)p_{3\|12}(2\|1,1)p_{4\|3}(1\|2)$
4	1	2	2	1	$p_1(1)p_2(2)p_{3\|12}(2\|1,2)p_{4\|3}(1\|2)$
5	1	2	2	2	
6	2	1	1	2	
7	2	1	2	3	
8	2	1	2	3	
9	2	2	2	3	
10	2	2	1	3	

Data Set and Likelihood

$$P(X_1, X_2, X_3, X_4) = p_1(X_1)p_2(X_2)p_{3|12}(X_3|X_1, X_2)p_{4|3}(X_4|X_3)$$

obs	X_1	X_2	X_3	X_4	
1	1	1	1	1	$p_1(1)p_2(1)p_{3|12}(1|1,1)p_{4|3}(1|1)$
2	1	1	1	1	$p_1(1)p_2(1)p_{3|12}(1|1,1)p_{4|3}(1|1)$
3	1	1	2	1	$p_1(1)p_2(1)p_{3|12}(2|1,1)p_{4|3}(1|2)$
4	1	2	2	1	$p_1(1)p_2(2)p_{3|12}(2|1,2)p_{4|3}(1|2)$
5	1	2	2	2	$p_1(1)p_2(2)p_{3|12}(2|1,2)p_{4|3}(2|2)$
6	2	1	1	2	
7	2	1	2	3	
8	2	1	2	3	
9	2	2	2	3	
10	2	2	1	3	

Data Set and Likelihood

$$P(X_1, X_2, X_3, X_4) = p_1(X_1)p_2(X_2)p_{3|12}(X_3|X_1, X_2)p_{4|3}(X_4|X_3)$$

obs	X_1	X_2	X_3	X_4	
1	1	1	1	1	$p_1(1)p_2(1)p_{3\|12}(1\|1,1)p_{4\|3}(1\|1)$
2	1	1	1	1	$p_1(1)p_2(1)p_{3\|12}(1\|1,1)p_{4\|3}(1\|1)$
3	1	1	2	1	$p_1(1)p_2(1)p_{3\|12}(2\|1,1)p_{4\|3}(1\|2)$
4	1	2	2	1	$p_1(1)p_2(2)p_{3\|12}(2\|1,2)p_{4\|3}(1\|2)$
5	1	2	2	2	$p_1(1)p_2(2)p_{3\|12}(2\|1,2)p_{4\|3}(2\|2)$
6	2	1	1	2	$p_1(2)p_2(1)p_{3\|12}(1\|2,1)p_{4\|3}(2\|1)$
7	2	1	2	3	$p_1(2)p_2(1)p_{3\|12}(2\|2,1)p_{4\|3}(3\|2)$
8	2	1	2	3	$p_1(2)p_2(1)p_{3\|12}(2\|2,1)p_{4\|3}(3\|2)$
9	2	2	2	3	$p_1(2)p_2(2)p_{3\|12}(2\|2,2)p_{4\|3}(3\|2)$
10	2	2	1	3	$p_1(2)p_2(2)p_{3\|12}(1\|2,2)p_{4\|3}(3\|1)$

The Likelihood Function

Estimate 10 probabilities in total.

Contribution of observation 1 to the likelihood:

$$L(1,1,1,1) = p_1(1)p_2(1)p_{3|1,2}(1|1,1)p_{4|3}(1|1)$$

Contribution of observation 3:

$$\begin{aligned} L(1,1,2,1) &= p_1(1)p_2(1)p_{3|1,2}(2|1,1)p_{4|3}(1|2) \\ &= p_1(1)p_2(1)(1 - p_{3|1,2}(1|1,1))p_{4|3}(1|2) \end{aligned}$$

Joint contribution of observation 1 and 3 is:

$$p_1(1)^2 p_2(1)^2 p_{3|1,2}(1|1,1)(1 - p_{3|1,2}(1|1,1))p_{4|3}(1|1)p_{4|3}(1|2)$$

For all observations

Likelihood function for all observations together:

$$
\begin{aligned}
L(\mathcal{D}) \quad = \quad & p_1(1)^5(1-p_1(1))^5 p_2(1)^6(1-p_2(1))^4 p_{3|1,2}(1|1,1)^2(1-p_{3|1,2}(1|1,1)) \\
& (1-p_{3|1,2}(1|1,2))^2 p_{3|1,2}(1|2,1)(1-p_{3|1,2}(1|2,1))^2 p_{3|1,2}(1|2,2) \\
& (1-p_{3|1,2}(1|2,2)) p_{4|3}(1|1)^2 p_{4|3}(2|1)(1-p_{4|3}(1|1)-p_{4|3}(2|1)) \\
& p_{4|3}(1|2)^2 p_{4|3}(2|2)(1-p_{4|3}(1|2)-p_{4|3}(2|2))^3
\end{aligned}
$$

Or in log form

$$
\begin{aligned}
\mathcal{L}(\mathcal{D}) \quad = \quad & 5\log p_1(1) + 5\log(1-p_1(1)) + 6\log p_2(1) + 4\log(1-p_2(1)) \\
& +2\log p_{3|1,2}(1|1,1) + \log(1-p_{3|1,2}(1|1,1)) \\
& +2\log(1-p_{3|1,2}(1|1,2)) + \log p_{3|1,2}(1|2,1) + 2\log(1-p_{3|1,2}(1|2,1)) \\
& +\log p_{3|1,2}(1|2,2) + \log(1-p_{3|1,2}(1|2,2)) \\
& +2\log p_{4|3}(1|1) + \log p_{4|3}(2|1) + \log(1-p_{4|3}(1|1)-p_{4|3}(2|1)) \\
& +2\log p_{4|3}(1|2) + \log p_{4|3}(2|2) + 3\log(1-p_{4|3}(1|2)-p_{4|3}(2|2))
\end{aligned}
$$

ML estimate of $p_{3|1,2}(1|1,1)$

Take partial derivative of $\mathcal{L}$ wrt $p = p_{3|1,2}(1|1,1)$:

$$\mathcal{L} = \ldots + 2\log p + \log(1-p) + \ldots$$
$$\frac{\partial \mathcal{L}}{\partial p} = \frac{2}{p} - \frac{1}{1-p}$$

Equate to zero and solve for p: $p = \frac{2}{3}$

For all observations

Likelihood function for all observations together:

$$
\begin{aligned}
L(\mathcal{D}) \quad = \quad & p_1(1)^5(1-p_1(1))^5 p_2(1)^6(1-p_2(1))^4 p_{3|1,2}(1|1,1)^2(1-p_{3|1,2}(1|1,1)) \\
& (1-p_{3|1,2}(1|1,2))^2 p_{3|1,2}(1|2,1)(1-p_{3|1,2}(1|2,1))^2 p_{3|1,2}(1|2,2) \\
& (1-p_{3|1,2}(1|2,2)) p_{4|3}(1|1)^2 p_{4|3}(2|1)(1-p_{4|3}(1|1)-p_{4|3}(2|1)) \\
& p_{4|3}(1|2)^2 p_{4|3}(2|2)(1-p_{4|3}(1|2)-p_{4|3}(2|2))^3
\end{aligned}
$$

Or in log form

$$
\begin{aligned}
\mathcal{L}(\mathcal{D}) \quad = \quad & 5\log p_1(1) + 5\log(1-p_1(1)) + 6\log p_2(1) + 4\log(1-p_2(1)) \\
& +2\log p_{3|1,2}(1|1,1) + \log(1-p_{3|1,2}(1|1,1)) \\
& +2\log(1-p_{3|1,2}(1|1,2)) + \log p_{3|1,2}(1|2,1) + 2\log(1-p_{3|1,2}(1|2,1)) \\
& +\log p_{3|1,2}(1|2,2) + \log(1-p_{3|1,2}(1|2,2)) \\
& +2\log p_{4|3}(1|1) + \log p_{4|3}(2|1) + \log(1-p_{4|3}(1|1)-p_{4|3}(2|1)) \\
& +2\log p_{4|3}(1|2) + \log p_{4|3}(2|2) + 3\log(1-p_{4|3}(1|2)-p_{4|3}(2|2))
\end{aligned}
$$