COMP538: Introduction to Bayesian Networks

Lecture 7: Parameter Learning with Incomplete Data

Nevin L. Zhang
lzhang@cse.ust.hk

Department of Computer Science and Engineering
Hong Kong University of Science and Technology

Fall 2008

Objective

- Objective: Parameter learning with incomplete data.
- Reading: Zhang and Guo (2007), Chapter 7
- Reference: Heckerman (1996) (first half), Cowell *et al* (1999, Chapter 9)

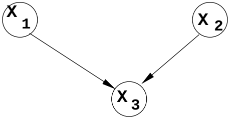
Outline

Missing Data

- Real-world data usually contains missing entries.
- We need to deal with **incomplete data sets** that looks like the following:

X_1	X_2	X_3	X_1	X_2	X_3
1	1	1	2	1	1
?	1	2	2	1	2
1	?	?	2	?	1
2	1	1	?	2	?

where ? indicates missing values.



Missing at Random

- To deal with missing values, we need to make the **missing at random (MAR)** assumption:
  - Actual value of X and the event X-is-missing are conditionally independent given other observed variables.

$$P(X|X\text{-is-mising, other observed variables}) = P(X|\text{other observed variables})$$

  - Given all the observed variables, the fact that X's value is missing gives us no additional information about the value.

Missing at Random

- The assumption is sometimes not true.
  - A patient record contains no value for "chest X-ray result" suggests that the doctor did not think chest X-ray test is necessary;
  - The result would be negative even if performed.
- However, it can be made true by introducing, when necessary, an auxiliary binary variable $Observed{-}X$.
  - $Observed{-}X$ is always observed, taking value "yes" when X is observed and "no" otherwise.
  - We now have

$$\begin{aligned} & P(X|X\text{-is-mising}, Observed{-}x, \text{other observed variables}) \\ = \quad & P(X|Observed{-}X, \text{other observed variables}) \end{aligned}$$

Outline

Basic Idea of EM

- One algorithm for finding MLE: The **expectation-maximization (EM)** algorithm.
- Developed in the Statistics community (Dempster *et al.* 1977). Adapted for Bayesian networks by Lauritzen (1994).
- It is an iterative algorithm.
  - Starts with an initial estimation θ^0.
  - At each iteration t,
    - **Expectation**: Complete the data set based on θ^t.
    - **Maximization**: Re-estimate parameters using the completed data set, obtaining θ^{t+1}.

The expectation step

How to complete data?

- θ^t is given. So, there is a joint distribution $P(.|\theta^t)$ over all variables.
- Consider an incomplete data case $D_3 = (1, ?, ?)$.
  - EM computes $P(X_2, X_3|X_1 = 1, \theta^t)$.
  - Suppose

$$P(X_2 = 1, X_3 = 1|X_1 = 1, \theta^t) = 1/4, P(X_2 = 1, X_3 = 2|X_1 = 1, \theta^t) = 1/4$$

$$P(X_2 = 2, X_3 = 1|X_1 = 1, \theta^t) = 1/4, P(X_2 = 2, X_3 = 2|X_1 = 1, \theta^t) = 1/4$$

  - EM splits D_3 into the following four **partial** data cases:

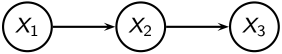
$$(1, ?, ?) \Rightarrow (1, 1, 1)[1/4], (1, 1, 2)[1/4], (1, 2, 1)[1/4], (1, 2, 2)[1/4]$$

    Each of them is counted as one fourth of a data case.
  - Note: The MAR assumption is implicitly used.

The maximization step

- After data completion, we get a data set with complete data cases.
  - Some of the data cases are partial data cases.
- EM re-estimates the parameters using the complete data set.
  - Partial data cases are counted according to their associated weights.

An Example



	X_1	X_2	X_3
$\mathbf{D}_1$	1	1	1
$\mathbf{D}_2$	2	2	2
$\mathbf{D}_3$	1	-	1
$\mathbf{D}_4$	2	-	2

- Choose θ^0:

$P(X_1)$

X_1	1	2
	1/2	1/2

$P(X_2|X_1)$

X_1 \ X_2	1	2
1	2/3	1/3
2	1/3	2/3

$P(X_3|X_2)$

X_2 \ X_3	1	2
1	2/3	1/3
2	1/3	2/3

An Example

- Because

$$P(X_2{=}1|\mathbf{D}_3, \theta^0) = 4/5 \quad P(X_2{=}2|\mathbf{D}_3, \theta^0) = 1/5$$

- $\mathbf{D}_3$ is split into: $\mathbf{D}_{3.1}{=}(1,1,1)[4/5]\mathbf{D}_{3.2}{=}(1,2,1)[1/5]$.
- Similarly, $\mathbf{D}_4$ is also split into two partial data cases.

- The completed data:

	X_1	X_2	X_3	weights
$\mathbf{D}_1$	1	1	1	1
$\mathbf{D}_2$	2	2	2	1
$\mathbf{D}_{3.1}$	1	1	1	4/5
$\mathbf{D}_{3.2}$	1	2	1	1/5
$\mathbf{D}_{4.1}$	2	1	1	1/5
$\mathbf{D}_{4.2}$	2	2	2	4/5

An Example

- The completed data:

	X_1	X_2	X_3	weights
$\mathbf{D}_1$	1	1	1	1
$\mathbf{D}_2$	2	2	2	1
$\mathbf{D}_{3.1}$	1	1	1	4/5
$\mathbf{D}_{3.2}$	1	2	1	1/5
$\mathbf{D}_{4.1}$	2	1	1	1/5
$\mathbf{D}_{4.2}$	2	2	2	4/5

- Calculate θ^1:

$P(X_1)$

X_1	1	2
	1/2	1/2

$P(X_2|X_1)$

X_1 \ X_2	1	2
1	9/10	1/10
2	1/10	9/10

$P(X_3|X_2)$

X_2 \ X_3	1	2
1	9/10	1/10
2	1/10	9/10

Exercise: Repeat the process for two more steps.

Review of the Complete-Data Case

- In MLE, we maximize the loglikelihood $l(\theta|\mathbf{D})$.
- In the case of complete data,
  - We have
$$l(\theta|\mathbf{D}) = \sum_{l} logP(D_l|\theta)$$
  - Estimation is done in two steps:
    1. Compute the loglikelihood
$$l(\theta|\mathbf{D}) = \sum_{i,k} \sum_{j} m_{ijk} log\theta_{ijk}$$
       Or equivalently, the sufficient statistics $m_{ijk} = \sum_l \chi(i, j, k : D_l)$
    2. Calculate estimate:
$$\theta^*_{ijk} = \frac{m_{ijk}}{\sum_j m_{ijk}}$$

Expected Loglikelihood

- Now consider the case incomplete data:
  - Suppose value of a variable X_l is missing from D_l.
  - In the expectation step, the case is completed and split into several partial cases:

$$(X_l{=}1, D_l)[P(X_1{=}1|D_l, \theta^t)], \qquad (X_l{=}2, D_l)[P(X_1{=}2|D_l, \theta^t)]$$

  - Correspondingly, in the loglikelihood function, we have

$$P(X_l{=}1|D_l, \theta^t) log P(X_l{=}1, D_l|\theta) + P(X_l{=}2|D_l, \theta^t) log P(X_l{=}2, D_l|\theta)$$

  - In general, we have the so-called **expected loglikelihood**:

$$l(\theta|\mathbf{D}, \theta^t) = \sum_l \sum_{\mathbf{x}_l \in \Omega_{\mathbf{X}_l}} P(\mathbf{X}_l{=}\mathbf{x}_l|D_l, \theta^t) log P(D_l, \mathbf{X}_l{=}\mathbf{x}_l|\theta)$$

    where $\mathbf{X}_l$ is in both face because there could be more that one missing values.

EM in terms Expected Loglikelihood

Formally, the next estimate θ^{t+1} is obtained from the current one θ^t in two steps:

1. The E-step computes the current expected loglikelihood function, now denoted by $Q(\theta|\theta^t)$ for simplicity, of θ given data $\mathbf{D}$, i.e.

$$Q(\theta|\theta^t) = \sum_{l} \sum_{\mathbf{x}_l \in \Omega_{\mathbf{X}_l}} P(\mathbf{X}_l{=}\mathbf{x}_l|D_l, \theta^t) log P(D_l, \mathbf{X}_l{=}\mathbf{x}_l|\theta),$$

   where $\mathbf{X}_l$ is the set of variables whose values a7re missing from data case D_l.

2. The M-step computes the next estimate θ^{t+1} by maximizing the current expected loglikelihood:

$$Q(\theta^{t+1}|\theta^t) \geq Q(\theta|\theta^t) \text{ for all } \theta.$$

   Or

$$\theta^{t+1} = \arg\max_{\theta} Q(\theta|\theta^t)$$

Characteristic Function

- Consider a specific value $\mathbf{x}_l$ for $\mathbf{X}_l$. Define

$$\chi(i, j, k : D_l, \mathbf{X}_l{=}\mathbf{x}_l) = \begin{cases} 1 & \text{if } X_i = j \text{ and } pa(X_i) = k \text{ are in } (D_l, \mathbf{X}_l{=}\mathbf{x}_l) \\ 0 & \text{otherwise} \end{cases}$$

Computation in the E-step

- Then

$$
\begin{aligned}
Q(\theta|\theta^t) &= \sum_l \sum_{\mathbf{x}_l \in \Omega_{\mathbf{X}_l}} P(X_l{=}\mathbf{x}_l|D_l,\theta^t) log P(D_l, \mathbf{X}_l{=}\mathbf{x}_l|\theta) \\
&= \sum_l \sum_{\mathbf{x}_l \in \Omega_{\mathbf{X}_l}} P(\mathbf{X}_l{=}\mathbf{x}_l|D_l,\theta^t) \sum_{i,j,k} \chi(i,j,k : D_l, \mathbf{X}_l{=}\mathbf{x}_l) log\theta_{ijk} \\
&= \sum_{i,j,k} \sum_l \sum_{\mathbf{x}_l \in \Omega_{\mathbf{X}_l}} P(\mathbf{X}_l{=}\mathbf{x}_l|D_l,\theta^t) \chi(i,j,k : D_l, \mathbf{X}_l{=}\mathbf{x}_l) log\theta_{ijk} \\
&= \sum_{i,j,k} m_{ijk}^t log\theta_{ijk} \\
&= \sum_{i,k} \sum_j m_{ijk}^t log\theta_{ijk}
\end{aligned}
$$

where the **sufficient statistics** m_{ijk}^t are given by

$$
m_{ijk}^t = \sum_l \sum_{\mathbf{x}_l \in \Omega_{\mathbf{X}_l}} P(\mathbf{X}_l{=}\mathbf{x}_l|D_l,\theta^t) \chi(i,j,k : D_l, X_l{=}\mathbf{x}_l)
$$

Computation in the M-Step

- Maximizing θ (Corollary 1.1), we get

$$\theta_{ijk}^{t+1} = \frac{m_{ijk}^t}{\sum_j m_{ijk}^t} \text{ for all } i, j, \text{ and } k.$$

- Interpretation:
  - $m_{ijk}^t = \sum_l \sum_{\mathbf{x}_l \in \Omega_{\mathbf{X}_l}} P(\mathbf{X}_l{=}\mathbf{x}_l | D_l, \theta^t)\chi(i, j, k : D_l, \mathbf{X}_l{=}\mathbf{x}_l)$ is
    - The number of cases where $X_i{=}j$ and $pa(X_i){=}k$ in the **completed** data set.
    - Or expected number of cases where $X_i{=}j$ and $pa(X_i){=}k$
  - Hence,

$$\begin{aligned}\theta_{ijk}^{t+1} &= \frac{\text{number of cases where } X_i = j \text{ and } pa(X_i){=}k \text{ in the completed data set}}{\text{number of cases where } pa(X_i){=}k \text{ in the completed data set}} \\ &= \frac{\text{expected number of cases where } X_i{=}j \text{ and } pa(X_i){=}k}{\text{expected number of cases where } pa(X_i){=}k}\end{aligned}$$

Sufficient Statistics Rewritten

Simplifying notation:

$$
\begin{aligned}
m^t_{ijk} &= \sum_l \sum_{\mathbf{x}_l \in \Omega_{\mathbf{X}_l}} P(\mathbf{X}_l{=}\mathbf{x}_l | D_l, \theta^t)\chi(i,j,k : D_l, \mathbf{X}_l{=}\mathbf{x}_l) \\
&= \sum_l \sum_{\mathbf{X}_l} P(\mathbf{X}_l | D_l, \theta^t)\chi(i,j,k : D_l, \mathbf{X}_l)
\end{aligned}
$$

- Let $\mathbf{Y}_l$ be the set of variables observed in D_l.
- We have: $P(\mathbf{X}_l | D_l, \theta^t) = \sum_{Y_l} P(\mathbf{X}_l, \mathbf{Y}_l | D_l, \theta^t)$
- Hence

$$
m^t_{ijk} = \sum_l \sum_{\mathbf{X}_l} \sum_{Y_l} P(\mathbf{X}_l, \mathbf{Y}_l | D_l, \theta^t)\chi(i,j,k : D_l, \mathbf{X}_l)
$$

Sufficient Statistics Rewritten

$$
\begin{aligned}
m^t_{ijk} &= \sum_{l} \sum_{\mathbf{X}_l} \sum_{Y_l} P(\mathbf{X}_l, \mathbf{Y}_l | D_l, \theta^t) \chi(i, j, k : D_l, \mathbf{X}_l) \\
&= \sum_{l} \sum_{\mathbf{Y}_l, \mathbf{X}_l \text{ s.t. } X_i = j, pa(X_i) = k \text{ in } (D_l, \mathbf{X}_l)} P(\mathbf{X}_l, \mathbf{Y}_l | D_l, \theta^t) \\
&= \sum_{l} \sum_{\mathbf{Y}_l, \mathbf{X}_l \text{ s.t. } X_i = j, pa(X_i) = k \text{ in } (\mathbf{Y}_l, \mathbf{X}_l)} P(\mathbf{X}_l, \mathbf{Y}_l | D_l, \theta^t) \\
&\qquad \text{because } P(\mathbf{X}_l, \mathbf{Y}_l | D_l, \theta^t) = 0 \text{ if } \mathbf{Y}_l \neq D_l \\
&= \sum_{l} P(X_i = j, pa(X_i) = k | D_l, \theta^t) \\
&\qquad \text{analogy: } \textstyle\sum_{A,B,C:B=1} P(A, B, C) = P(B=1)
\end{aligned}
$$

One EM-step

$\texttt{EM}-\texttt{Step}(\mathbf{D}, \theta^t)$:

- E-step:
  - Compute $P(X_i, pa(X_i)|D_l, \theta^t)$ for all X_i and D_l.
  - Compute the sufficient statistics $m^t_{ijk} = \sum_l P(X_i{=}j, pa(X_i){=}k|D_l, \theta^t)$ for all i, j, k.
- M-step: Compute

$$\theta^{t+1}_{ijk} = \frac{m^t_{ijk}}{\sum_j m^t_{ijk}}$$

  for all i, j, and k.

- Return θ^{t+1}.

Questions:

- The first step of E-step is standard Bayesian network inference. Which inference algorithm to use, VE or CTP?
- What is the problem if we implement EM-step using the basic idea directly?

The EM algorithm

$\texttt{EM}(\mathbf{D})$:

- Randomly pick θ^0.
- For $t = 0$ to termination
  - $\theta^{t+1} = \texttt{EM}-\texttt{STEP}(\mathbf{D}, \theta^t)$

When should we terminate?

Loglikelihood and Expected Loglikelihood

$$
\begin{aligned}
l(\theta|\mathbf{D}) &= \sum_{l} logP(D_l|\theta) \\
&= \sum_{l}\sum_{\mathbf{X}_l} P(\mathbf{X}_l|D_l,\theta^t) logP(D_l|\theta) \\
&= \sum_{l}\sum_{\mathbf{X}_l} P(\mathbf{X}_l|D_l,\theta^t) log\frac{P(D_l,\mathbf{X}_l|\theta)}{P(\mathbf{X}_l|D_l,\theta)} \\
&= \sum_{l}\sum_{\mathbf{X}_l} P(\mathbf{X}_l|D_l,\theta^t) logP(D_l,\mathbf{X}_l|\theta) - \sum_{l}\sum_{\mathbf{X}_l} P(\mathbf{X}_l|D_l,\theta^t) logP(\mathbf{X}_l|D_l,\theta) \\
&= Q(\theta|\theta^t) - \sum_{l}\sum_{\mathbf{X}_l} P(\mathbf{X}_l|D_l,\theta^t) logP(\mathbf{X}_l|D_l,\theta).
\end{aligned}
$$

EM and Loglikelihood

- Hence we have

$$
\begin{aligned}
l(\theta^t|\mathbf{D}) &= Q(\theta^t|\theta^t) - \sum_l \sum_{\mathbf{X}_l} P(\mathbf{X}_l|D_l,\theta^t) log P(\mathbf{X}_l|D_l,\theta^t) \\
&\leq Q(\theta^{t+1}|\theta^t) - \sum_l \sum_{\mathbf{X}_l} P(\mathbf{X}_l|D_l,\theta^t) log P(\mathbf{X}_l|D_l,\theta^t) \\
&\leq Q(\theta^{t+1}|\theta^t) - \sum_l \sum_{\mathbf{X}_l} P(\mathbf{X}_l|D_l,\theta^t) log P(\mathbf{X}_l|D_l,\theta^{t+1}) \\
&= l(\theta^{t+1}|\mathbf{D})
\end{aligned}
$$

  where

  - the first inequality is due to the definition of θ^{t+1}, and
  - the second inequality is due to Corollary 1.1.

- So, $l(\theta^t|\mathbf{D})$ monotonically increases with t.
- On the other hand, $l(\theta^t|\mathbf{D})$ is upper bounded by 0.
- Hence EM converges.

Complete Statement of the EM algorithm

$\texttt{EM}(\mathbf{D})$:

- Randomly pick θ^0.
- For $t = 0$ to ∞
  - $\theta^{t+1} = \texttt{EM}-\texttt{STEP}(\mathbf{D}, \theta^t)$
  - If $l(\theta^{t+1}|\mathbf{D}) \leq l(\theta^t|\mathbf{D}) + \epsilon$, return θ^{t+1}

What does EM converge to?

- (Mclachlan and Krishnan 1997) [1] If

$$\theta^{t+1} = \arg\max_{\theta} Q(\theta|\theta^t) = \theta^t$$

then

$$\frac{\partial l(\theta|\mathbf{D})}{\partial\theta}|_{\theta=\theta^t} = 0$$

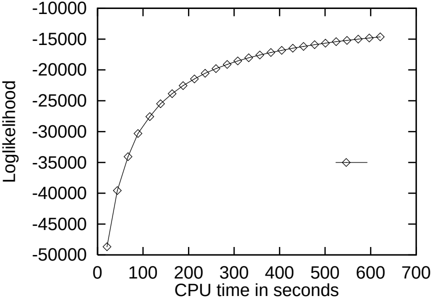
- EM converges to

global maxima, local maxima, or saddle points.

[1] McLachlan, G.J. and Krishnan, T. (1997). *The EM algorithm and extensions*. Wiley Interscience.

Empirical Experience with EM

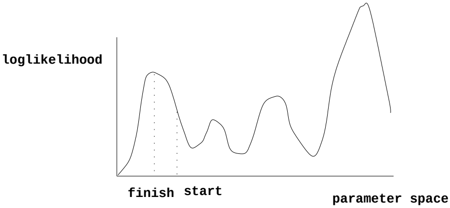
- Usually fast ,especially at first few iterations.



- Rate of convergence: The more missing data, the slower the convergence.

Local Maxima

- There is no guarantee that EM converge to the global optimum.
- It might be stacked at local maxima.



- Solution:
  - Multiple random restart.
  - Simulated annealing.

Outline

The Case of Complete Data

- θ: random variable.
- Prior $p(\theta)$: product Dirichlet distribution

$$p(\theta) = \prod_{i,k} p(\theta_{i.k}) \propto \prod_{i,k} \prod_{j} \theta_{ijk}^{\alpha_{ijk}-1}$$

- Posterior $p(\theta|\mathbf{D})$: also product Dirichlet distribution

$$p(\theta|\mathbf{D}) \propto \prod_{i,k} \prod_{j} \theta_{ijk}^{m_{ijk}+\alpha_{ijk}-1}$$

- Prediction:

$$P(D_{m+1}|\mathbf{D}) = P(X_1, X_2, \ldots, X_n|\mathbf{D}) = \prod_{i} P(X_i|pa(X_i), \mathbf{D})$$

where

$$P(X_i{=}j|pa(X_i){=}k, \mathbf{D}) = \frac{m_{ijk}+\alpha_{ijk}}{m_{i*k}+\alpha_{i*k}}$$

Absorbing one Data Case

- Product Dirichlet density:

$$p(\theta) = \prod_{i,k} p(\theta_{i.k}) \propto \prod_{i,k} \prod_{j} \theta_{ijk}^{\alpha_{ijk}-1}$$

- Denote it by $\kappa(\theta|\alpha)$,where α stands for the vector of all α_{ijk}.
- Consider one incomplete case D_1 . Let $\mathbf{X}_1$ be the set of variables unobserved in D_1.
- We have

$$\begin{aligned} p(\theta|D_1) &\propto p(\theta)P(D_1|\theta) \\ &= p(\theta) \sum_{\mathbf{x}_1 \in \Omega_{\mathbf{X}_1}} P(D_1, \mathbf{X}_1 = \mathbf{x}_1|\theta) \\ &= \sum_{\mathbf{x}_1 \in \Omega_{\mathbf{X}_1}} p(\theta)P(D_1, \mathbf{X}_1 = \mathbf{x}_1|\theta) \qquad (1) \end{aligned}$$

Absorbing one Data Case

- Because $(D_1, \mathbf{X}_1{=}\mathbf{x}_1)$ is a complete case, each term $p(\theta)P(D_1, \mathbf{X}_1 = \mathbf{x}_1|\theta)$ corresponds to a product Dirichlet density $\kappa(\theta|\alpha_{\mathbf{x}_1})$.
- So the posterior distribution $P(\theta|D_1)$ is a **mixture of product Dirichlet densities**:

$$p(\theta|D_1) = \sum_{\mathbf{x}_1 \in \Omega_{\mathbf{X}_1}} w_{\mathbf{x}_1} \kappa(\theta|\alpha_{\mathbf{x}_1}).$$

  - This does not factorize.
  - Parameter independence (both global and local independence) no longer true for posterior $p(\theta|D_1)$.

Absorbing one Data Case

- Now if the prior were a mixture of N product Dirichlet densities

$$p(\theta) = \sum_{n=1}^{N} w_n \kappa(\theta|\alpha_n)$$

- Then posterior would be a mixture of $N * |\Omega_{\mathbf{X}_1}|$ product Dirichlet densities.
- The number of product Dirichlet density increases quickly as we absorb more and more cases.
  - If we start with a product Dirichelet prior, after absorbing n cases we get a mixture of this many product Dirichelet densities:

$$|\Omega_{\mathbf{X}_1}| * |\Omega_{\mathbf{X}_2}| * \ldots * |\Omega_{\mathbf{X}_n}|$$

- Conclusion: Approximation is necessary.

Fractional Updating

- Assume that
  - We have absorbed $\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_l$.
  - We have obtained an approximation of $p(\theta|\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_l)$,
  - which is a product Dirichlet distribution with hyperparameters:

$$\alpha^l = \{\alpha^l_{ijk} | i=1, \cdots, n; j=1, \cdots, q_i; k=1, \cdots, r_i\}$$

Fractional Updating

- Now consider absorbing the next data case $\mathbf{D}_{l+1}$ and approximating $p(\theta|\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_l, \mathbf{D}_{l+1})$
- Based on the above approximation, compute $P(\mathbf{D}_{l+1}|\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_l)$.
- It can be represented using a Bayesian network $\mathcal{N}^l{=}(S, \theta^l)$, where

$$\theta^l_{ijk} = \frac{\alpha^l_{ijk}}{\sum_{k=1}^{r_i} \alpha^l_{ijk}} \qquad (2)$$

- Denote $P(\mathbf{D}_{l+1}|\mathbf{D}_1, \cdots, \mathbf{D}_l)$ by P^l

Fractional Updating

- Let $\mathbf{X}_{l+1}$ be the set of variables whose values are missing from $\mathbf{D}_{l+1}$.
- The probability of $\mathbf{X}_{l+1}$ taking a particular value $\mathbf{x}_{l+1}$ is

$$P^l(\mathbf{X}_{l+1}{=}\mathbf{x}_{l+1})$$

- Completing $\mathbf{D}_{l+1}$, we get

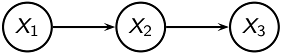
$$(\mathbf{D}_{l+1}, \mathbf{X}_{l+1}{=}\mathbf{x}_{l+1}) \quad [P^l(\mathbf{X}_{l+1}{=}\mathbf{x}_{l+1})]$$

- Updating the estimation using the completed data, we get a Dirichlet distribution whose hyperparameters are as follows:

$$\alpha_{ijk}^{l+1} = \alpha_{ijk}^{l} + P(X_i{=}k, \pi(X_i){=}j|\mathbf{D}_1, \cdots, \mathbf{D}_{l+1}) \tag{3}$$

  This is the approximation of $p(\theta|\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_l, \mathbf{D}_{l+1})$ given by fractional updating.

An Example



	X_1	X_2	X_3
$\mathbf{D}_1$	1	1	1
$\mathbf{D}_2$	2	2	2
$\mathbf{D}_3$	1	-	1
$\mathbf{D}_4$	2	-	2

- Prior $p(\theta)$: product Dirichlet density with hyperparameters α^0 given by

α^0_{1jk}

k \ j	1	2
1	2	2

α^0_{2jk}

k \ j	1	2
1	1	1
2	1	1

α^0_{3jk}

k \ j	1	2
1	1	1
2	1	1

An Example

- D_1 is complete.
- $p(\theta|D_1)$ is product Dirichlet density with hyperparameters α^1 given by

α^1_{1jk}

k \ j	1	2
1	3	2

α^1_{2jk}

k \ j	1	2
1	2	1
2	1	1

α^1_{3jk}

k \ j	1	2
1	2	1
2	1	1

- D_2 is also complete.
- $p(\theta|D_1, D_2)$ is product Dirichlet density with hyperparameters α^2 given by

α^2_{1jk}

k \ j	1	2
1	3	3

α^2_{2jk}

k \ j	1	2
1	2	1
2	1	2

α^2_{3jk}

k \ j	1	2
1	2	1
2	1	2

An Example

- $D_3 = (1, -, 1)$ is not complete.
- We need to complete the data case. This is a prediction task, i.e. predicting parts of D_3.
- Consider $P(D_3|D_1, D_2)$.
  - It can be presented by a Bayesian network with parameters given by:

$P(X_1|\theta^2)$

X_1	1	2
	3/6	3/6

$P(X_2|X_1, \theta^2)$

X_1 \ X_2	1	2
1	$\frac{2}{3}$	$\frac{1}{3}$
2	$\frac{1}{3}$	$\frac{2}{3}$

$P(X_3|X_2, \theta^2)$

X_2 \ X_3	1	2
1	$\frac{2}{3}$	$\frac{1}{3}$
2	$\frac{1}{3}$	$\frac{2}{3}$

- In this network, we have

$$P(X_2{=}1|\mathbf{D}_3, \theta^2){=}\frac{4}{5} \qquad P(X_2{=}2|\mathbf{D}_3, \theta^2){=}\frac{1}{5}$$

An Example

- Hence, D_3 is split into two fractional samples:

$$\mathbf{D}_{3.1}{=}(1,1,1)[\frac{4}{5}], \quad \mathbf{D}_{3.2}{=}(1,2,1)[\frac{1}{5}]$$

- Updating $p(\theta|D_1, D_2)$ using those two samples, we get $p(\theta|D_1, D_2, D_3)$
- $p(\theta|D_1, D_2, D_3)$ is product Dirichlet density with hyperparameters α^3 given by

α^3_{1jk}

$k \backslash j$	1	2
1	4	3

α^3_{2jk}

$k \backslash j$	1	2
1	$\frac{14}{5}$	$\frac{6}{5}$
2	1	2

α^3_{3jk}

$k \backslash j$	1	2
1	$\frac{14}{5}$	1
2	$\frac{6}{5}$	2

- Exercise: Complete the example by absorbing D_4.

Notes

- Complexity of fractional updating: exponential in the number of variables whose values are missing.
- Due to approximation, the order of absorbing cases influences the final result.
- For more sophisticated approximations, see Spiegelhalter and Lauritzen (1990) and Cowell *et al* (1999, Chapter 9).