CSCE 970 Lecture 7: Parameter Learning

Stephen D. Scott

Introduction

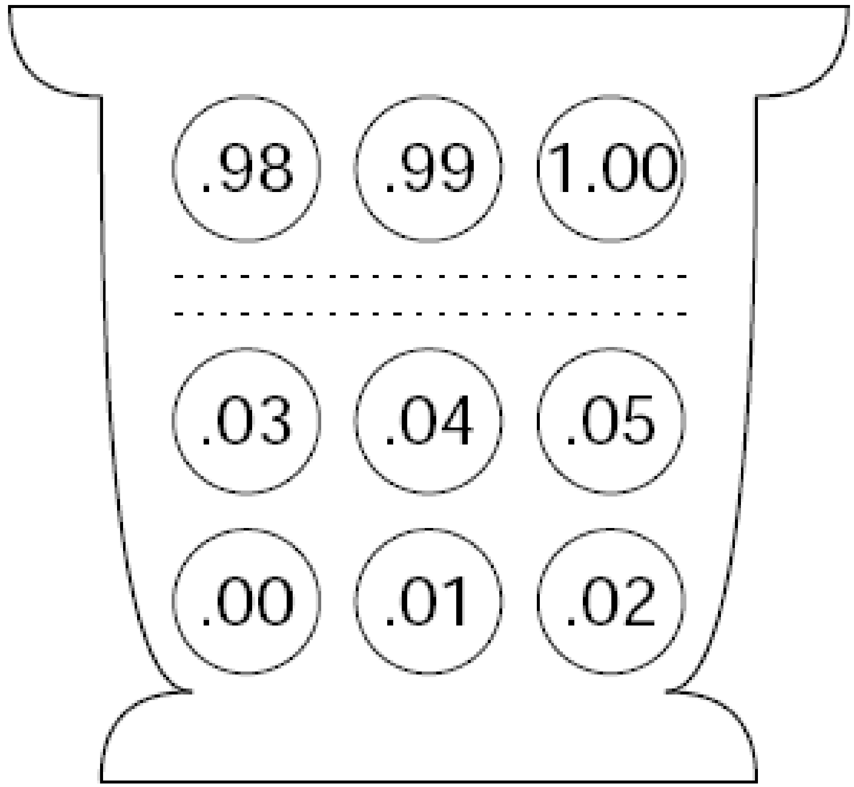
- Now we'll discuss how to parameterize a Bayes net
- Assume that the structure is given
- Start by representing prior beliefs, then incorporate results from data

Outline

- Learning a single parameter
  - Uniform prior belief
  - Beta distributions
  - Learning a relative frequency
- Beta distributions with nonintegral parameters
- Learning parameters in a Bayes net
  - Urn examples
  - Equivalent sample size
- Learning with missing data items

Learning a Single Parameter

All Relative Frequencies Equally Probable



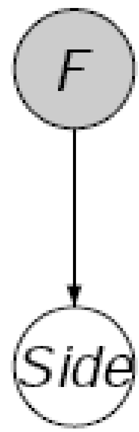
- Assume urn with 101 coins, each with different probability f of heads
- If we choose a specific coin f from the urn and flip it,

$$P(Side = heads \mid f) = f$$

Learning a Single Parameter

All Relative Frequencies Equally Probable (cont'd)

$$P(f) = 1/101 \qquad .00 \leq f \leq 1.00$$



$$P(Side = heads|f) = f$$

- If we choose the coin from the urn uniformly at random, then can represent with an augmented Bayes net
- Shaded node represents belief about a relative frequency

Learning a Single Parameter

All Relative Frequencies Equally Probable (cont'd)

$$
\begin{aligned}
P(Side = heads) &= \sum_{f=0.0}^{1.0} P(Side = heads \mid f)P(f) = \sum_{f=0.0}^{1.0} f/101 \\
&= \left(\frac{1}{(100)(101)}\right) \sum_{f=0}^{100} f \\
&= \left(\frac{1}{(100)(101)}\right) \left(\frac{(100)(101)}{2}\right) = 1/2
\end{aligned}
$$

Get same result if a continuous set of coins

Learning a Single Parameter

All Relative Frequencies Not Equally Probable

- Don't necessarily expect all coins to be equally likely

- E.g. may believe that coins more likely with $P(Side = heads) \approx 0.5$

- Further, need to characterize the strength of this belief with some measure of concentration (i.e. lack of variance)

- Will use the beta distribution

Learning a Single Parameter

All Relative Frequencies Not Equally Probable

Beta Distribution

- The beta distribution has parameters a and b and is denoted $beta(f; a, b)$

- Think of a and b as frequency counts in a pseudosample (for a prior) or in a real sample (based on training data)
  - a is the number of times coin came up heads, b tails

- If $N = a + b$, beta's probability density function is:

$$\rho(f) = \frac{\Gamma(N)}{\Gamma(a)\Gamma(b)} f^{a-1}(1-f)^{b-1}$$

  where

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$$

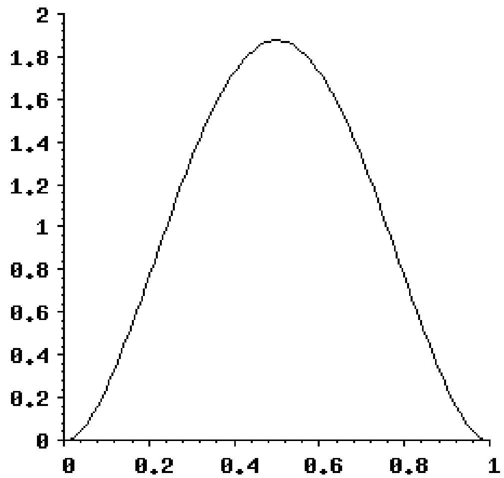
  is generalization of factorial

- Special case of Dirichlet distribution (Defn 6.4, p. 307)

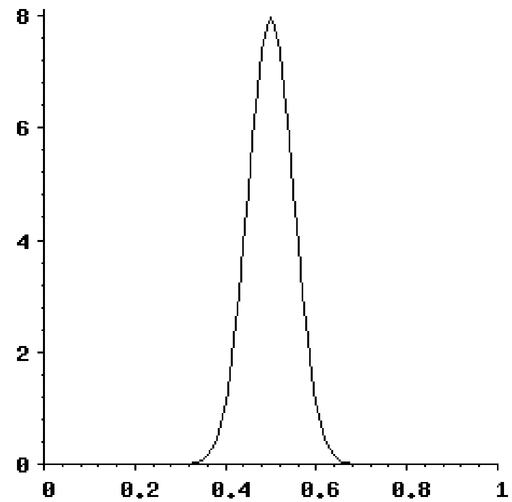
Learning a Single Parameter

All Relative Frequencies Not Equally Probable

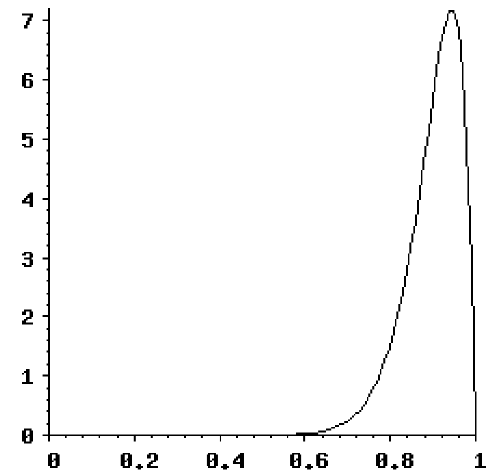
Beta Distribution (cont'd)



$beta(f; 3, 3)$



$beta(f; 50, 50)$



$beta(f; 18, 2)$

- Concentration of mass is at $E(F) = P(heads) = a/(a + b)$
- The larger N is, the more concentrated the pdf is (i.e. less variance)
- Thus relative values of a and b can represent prior beliefs, and $N = a + b$ represents strength of prior
- What does $beta(f; 1, 1)$ look like?

Learning a Single Parameter

All Relative Frequencies Not Equally Probable

Updating the Beta Distribution

- Say we're representing our prior as $beta(f; a, b)$ and then we see a data set with s heads and t tails
- Then the updated beta distribution that reflects the data $\mathbf{d}$ has a pdf

$$\rho(f \mid \mathbf{d}) = beta(f; a+s, b+t)$$

- I.e. we just add the data counts to the pseudocounts to reparameterize the beta distribution
- Further, the probability of seeing the data is

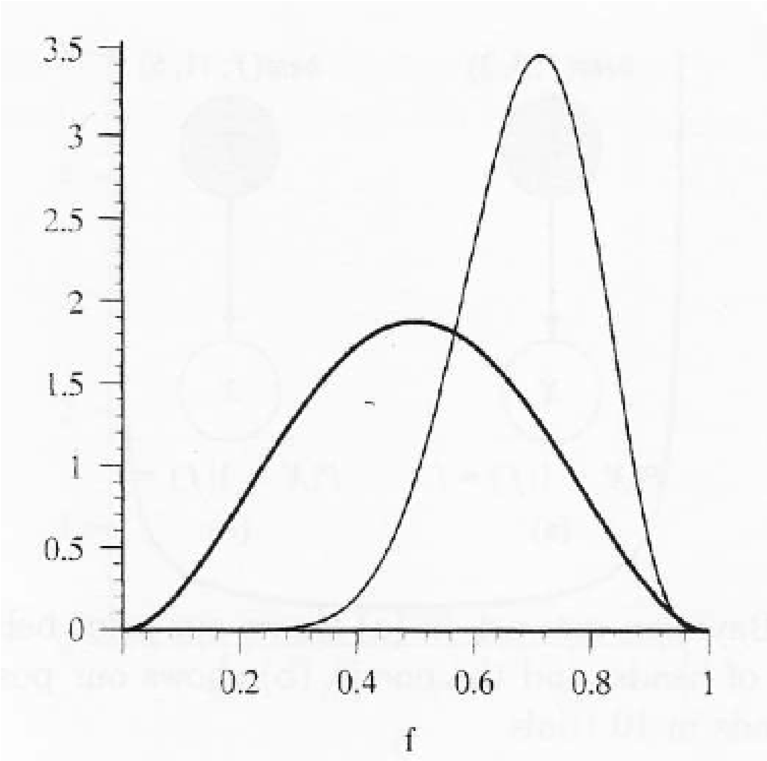
$$P(\mathbf{d}) = \frac{\Gamma(N)}{\Gamma(N+M)} \frac{\Gamma(a+s)\Gamma(b+t)}{\Gamma(a)\Gamma(b)} \quad ,$$

where $N = a + b$ and $M = s + t$

Learning a Single Parameter

All Relative Frequencies Not Equally Probable

Updating the Beta Distribution (example)



Bold curve is $beta(f;3,3)$ and light curve is $beta(f;11,5)$, after seeing data $\mathbf{d} = \{1,1,2,1,1,1,1,1,2,1\}$

Learning a Single Parameter
The Meaning of Beta Parameters

- If $a = b = 1$, then we assume nothing about what value is more likely, and let the data override our uninformed prior

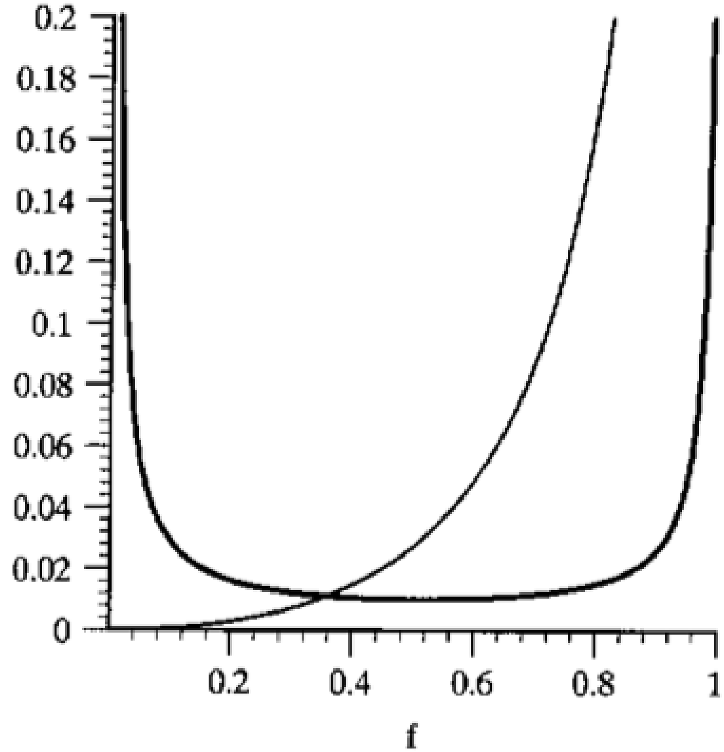
- If $a, b > 1$, then we believe that the distribution centers on $a/(a+b)$, and the strength of this belief is related to the magnitudes of the values

- If $a, b < 1$, then we believe that one of the two values (heads, tails) dominates the other, but we don't know which one
  - E.g. if $a = b = 0.1$ then our prior on heads is $0.1/0.2 = 1/2$, but if heads comes up after one coin toss, then posterior is $1.1/1.2 = 0.917$

- If $a < 1$ and $b > 1$, then we believe that "heads" is uncommon

Learning a Single Parameter

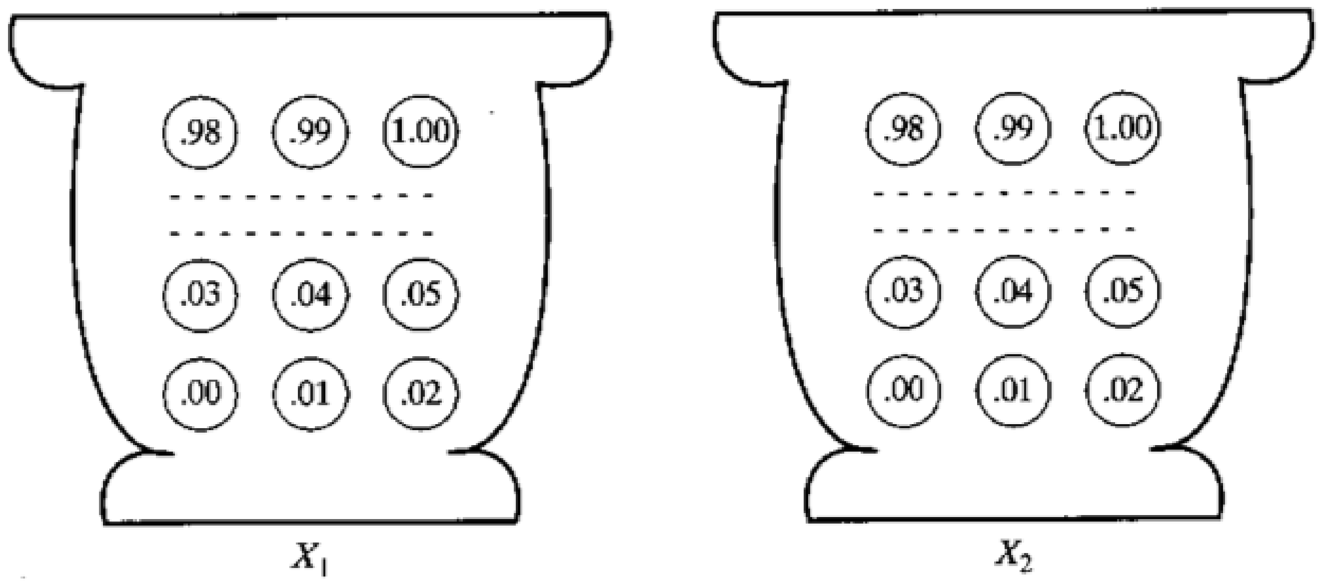
$a, b < 1$



U-shaped curve is $beta(f; 1/360, 19/360)$, other curve is $beta(f; 3 + 1/360, 19/360)$, after seeing three "heads," and probability of next one being heads is $(3 + 1/360)/(3 + 20/360) = 0.983$

Learning Parameters in a Bayes Net

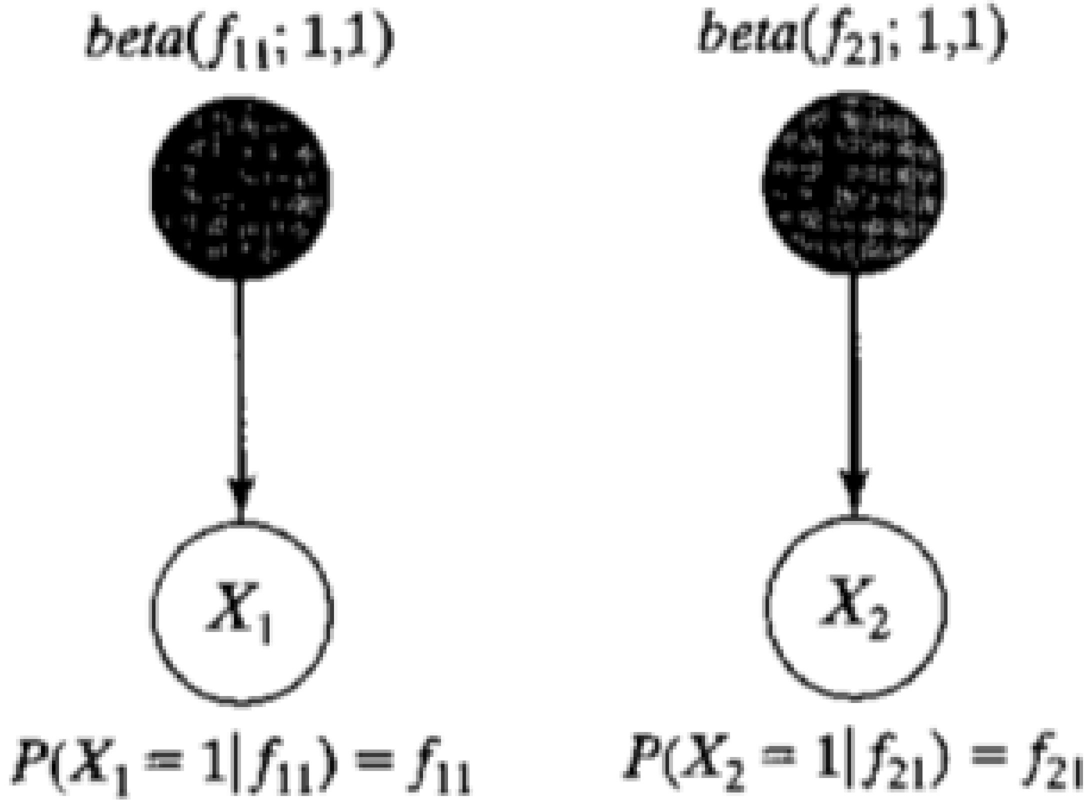
Example: Two Independent Urns



Experiment: Independently draw a coin from each urn X_1 and X_2, and repeatedly flip them

Learning Parameters in a Bayes Net

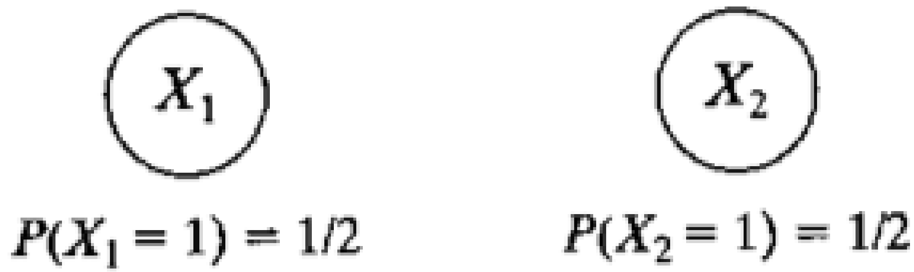
Example: Two Independent Urns (cont'd)



If prior on each urn is uniform ($beta(f_{i1}; 1, 1)$), then get above augmented Bayes net

Learning Parameters in a Bayes Net

Example: Two Independent Urns (cont'd)



Marginalizing and noting independence of coins yields the above embedded Bayes net with joint distribution ("1" = "heads"):

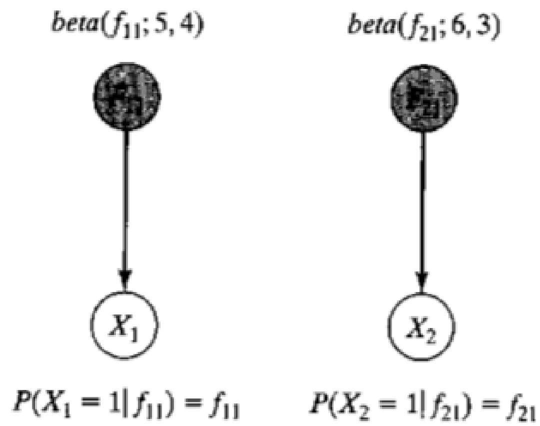
$$\begin{aligned}
P(X_1 = 1, X_2 = 1) &= P(X_1 = 1)P(X_2 = 1) = (1/2)(1/2) = 1/4 \\
P(X_1 = 1, X_2 = 2) &= P(X_1 = 1)P(X_2 = 2) = (1/2)(1/2) = 1/4 \\
P(X_1 = 2, X_2 = 1) &= P(X_1 = 2)P(X_2 = 1) = (1/2)(1/2) = 1/4 \\
P(X_1 = 2, X_2 = 2) &= P(X_1 = 2)P(X_2 = 2) = (1/2)(1/2) = 1/4
\end{aligned}$$

Learning Parameters in a Bayes Net

Example: Two Independent Urns (cont'd)

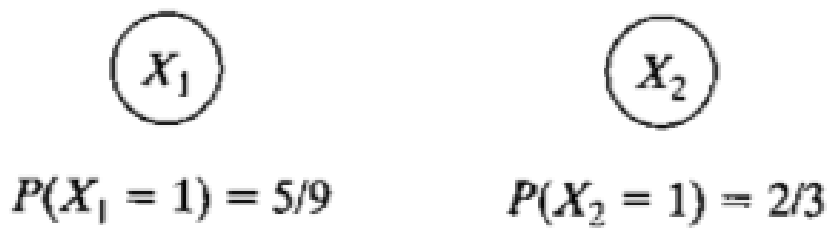
- Now sample one coin from each urn and toss each one 7 times
- End up with a set of pairs of outcomes, each of the form (X_1, X_2): $\mathbf{d} = \{(1,1),(1,1),(1,1),(1,2),(2,1),(2,1),(2,2)\}$
- I.e. coin X_1 got $s_{11} = 4$ heads and $t_{11} = 3$ tails and coin X_2 got $s_{21} = 5$ heads and $t_{21} = 2$ tails
- Thus

$$
\begin{aligned}
\rho(f_{11} \mid \mathbf{d}) &= beta(f_{11}; a_{11} + s_{11}, b_{11} + t_{11}) = beta(f_{11}; 5, 4) \\
\rho(f_{21} \mid \mathbf{d}) &= beta(f_{21}; a_{21} + s_{21}, b_{21} + t_{21}) = beta(f_{21}; 6, 3)
\end{aligned}
$$



Learning Parameters in a Bayes Net

Example: Two Independent Urns (cont'd)

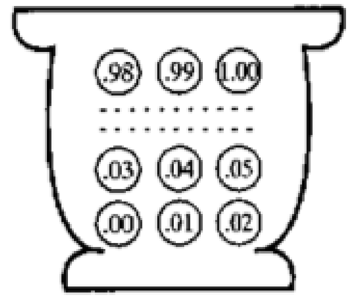


Marginalizing yields the above embedded Bayes net with joint distribution:

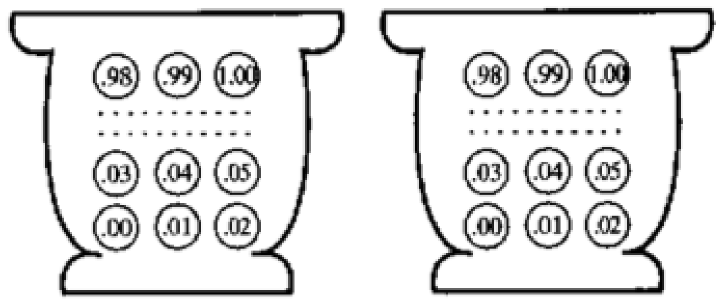
$$
\begin{aligned}
P(X_1 = 1, X_2 = 1) &= P(X_1 = 1)P(X_2 = 1) = (5/9)(2/3) = 10/27 \\
P(X_1 = 1, X_2 = 2) &= P(X_1 = 1)P(X_2 = 2) = (5/9)(1/3) = 5/27 \\
P(X_1 = 2, X_2 = 1) &= P(X_1 = 2)P(X_2 = 1) = (4/9)(2/3) = 8/27 \\
P(X_1 = 2, X_2 = 2) &= P(X_1 = 2)P(X_2 = 2) = (4/9)(1/3) = 4/27
\end{aligned}
$$

Learning Parameters in a Bayes Net

Example: Three Dependent Urns



X_1



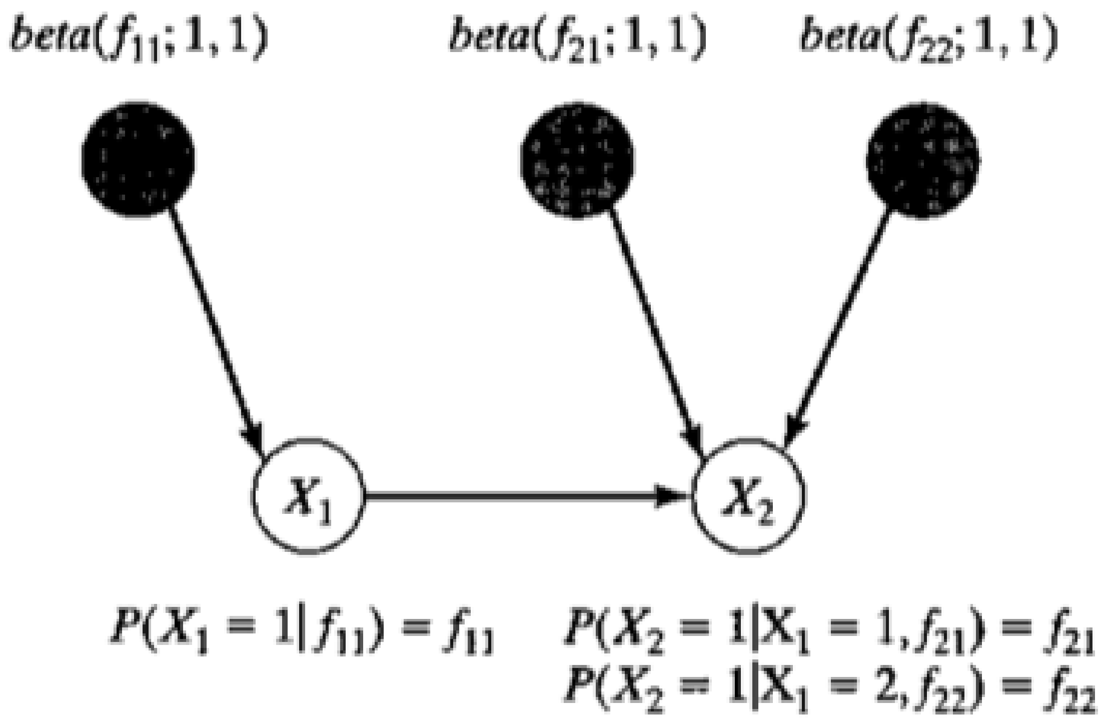
$X_2|X_1 = 1$ $\qquad$ $X_2|X_1 = 2$

Experiment: Independently draw a coin from each urn X_1, $X_2 \mid X_1 = 1$, and $X_2 \mid X_1 = 2$, then repeatedly flip X_1's coin

- If X_1 flip is heads, flip coin from urn $X_2 \mid X_1 = 1$
- If X_1 flip is tails, flip coin from urn $X_2 \mid X_1 = 2$

Learning Parameters in a Bayes Net

Example: Three Dependent Urns (cont'd)



If prior on each urn is uniform ($beta(f_{ij}; 1, 1)$), then get above augmented Bayes net

Learning Parameters in a Bayes Net

Example: Three Dependent Urns (cont'd)

$P(X_1 = 1) = 1/2$ $\quad$ $P(X_2 = 1 | X_1 = 1) = 1/2$

$P(X_2 = 1 | X_1 = 2) = 1/2$

Marginalizing yields the above embedded Bayes net with joint distribution:

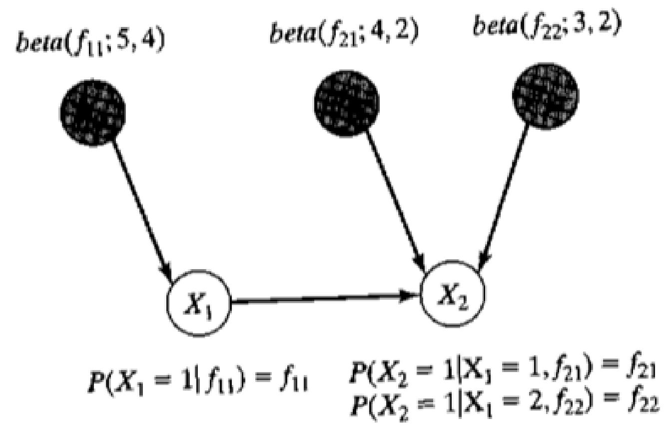
$$
\begin{aligned}
P(X_1 = 1, X_2 = 1) &= P(X_2 = 1 \mid X_1 = 1)P(X_1 = 1) = (1/2)(1/2) = 1/4 \\
P(X_1 = 1, X_2 = 2) &= P(X_2 = 2 \mid X_1 = 1)P(X_1 = 1) = (1/2)(1/2) = 1/4 \\
P(X_1 = 2, X_2 = 1) &= P(X_2 = 1 \mid X_1 = 2)P(X_1 = 2) = (1/2)(1/2) = 1/4 \\
P(X_1 = 2, X_2 = 2) &= P(X_2 = 2 \mid X_1 = 2)P(X_1 = 2) = (1/2)(1/2) = 1/4
\end{aligned}
$$

Learning Parameters in a Bayes Net

Example: Three Dependent Urns (cont'd)

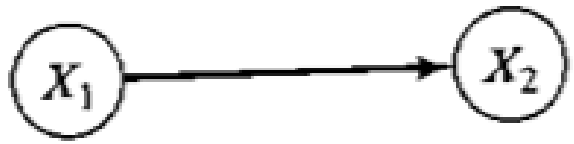
- Now continue experiment until you get a set of 7 pairs of outcomes, each of the form (X_1, X_2):
  $\mathbf{d} = \{(1,1),(1,1),(1,1),(1,2),(2,1),(2,1),(2,2)\}$
- I.e. coin X_1 got $s_{11} = 4$ heads and $t_{11} = 3$ tails, coin X_2 got $s_{21} = 3$ heads when X_1 was heads and $t_{21} = 1$ tail when X_1 was heads, and coin X_2 got $s_{22} = 2$ heads when X_1 was tails and $t_{22} = 1$ tail when X_1 was tails
- Thus

$$
\begin{aligned}
\rho(f_{11} \mid \mathbf{d}) &= beta(f_{11}; a_{11} + s_{11}, b_{11} + t_{11}) = beta(f_{11}; 5, 4) \\
\rho(f_{21} \mid \mathbf{d}) &= beta(f_{21}; a_{21} + s_{21}, b_{21} + t_{21}) = beta(f_{21}; 4, 2) \\
\rho(f_{22} \mid \mathbf{d}) &= beta(f_{22}; a_{22} + s_{22}, b_{22} + t_{22}) = beta(f_{21}; 3, 2)
\end{aligned}
$$



Learning Parameters in a Bayes Net

Example: Three Dependent Urns (cont'd)



$P(X_1 = 1) = 5/9$

$P(X_2 = 1|X_1 = 1) = 2/3$
$P(X_2 = 1|X_1 = 2) = 3/5$

Marginalizing yields the above embedded Bayes net with joint distribution:

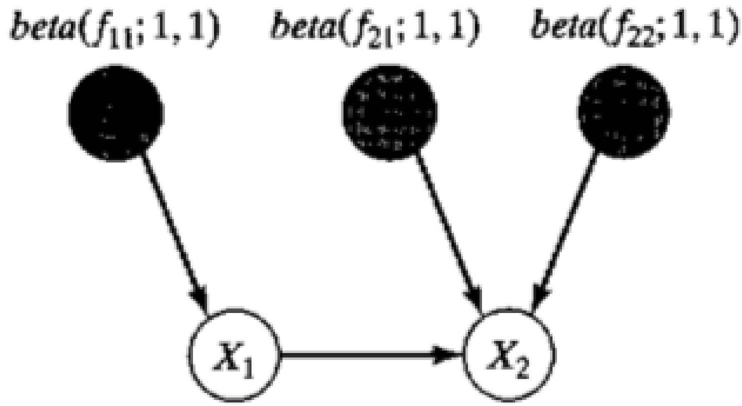
$$
\begin{aligned}
P(X_1 = 1, X_2 = 1) &= P(X_2 = 1 \mid X_1 = 1)P(X_1 = 1) = (2/3)(5/9) = 10/27 \\
P(X_1 = 1, X_2 = 2) &= P(X_2 = 2 \mid X_1 = 1)P(X_1 = 1) = (1/3)(5/9) = 5/27 \\
P(X_1 = 2, X_2 = 1) &= P(X_2 = 1 \mid X_1 = 2)P(X_1 = 2) = (3/5)(4/9) = 12/45 \\
P(X_1 = 2, X_2 = 2) &= P(X_2 = 2 \mid X_1 = 2)P(X_1 = 2) = (2/5)(4/9) = 8/45
\end{aligned}
$$

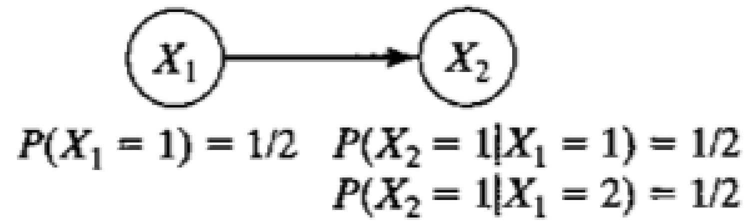
Learning Parameters in a Bayes Net

- When all the data are completely specified, the algorithm for parameterizing the network is very simple
  - Define the prior and initialize the parameters of each node's conditional probability table with that prior (in the form of pseudocounts)
  - When a fully-specified example is presented, update the counts by matching the attribute values to the appropriate row in each CPT
  - To compute a conditional probability, simply normalize each count table

Prior Equivalent Sample Size

The Problem





Given the above Bayes net and the following data set
$\mathbf{d} = \{(1, 2), (1, 1), (2, 1), (2, 2), (2, 1), (2, 1), (1, 2), (2, 2)\}$,
what is $P(X_2 = 1)$?

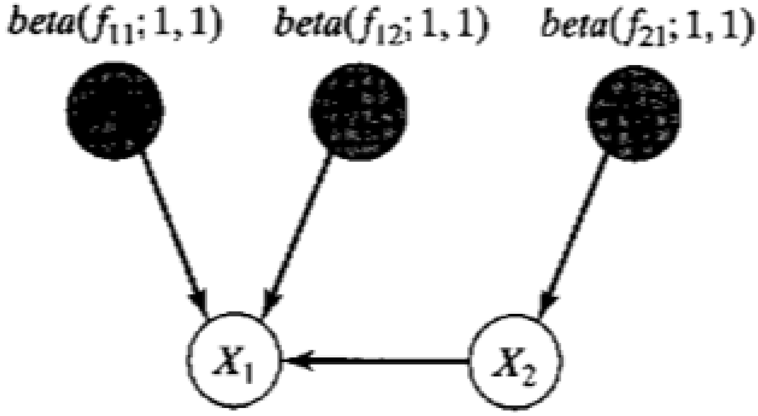
Prior Equivalent Sample Size

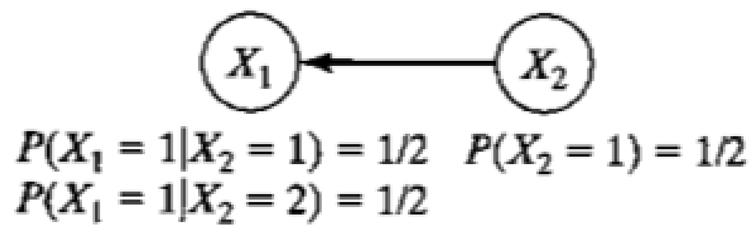
The Problem (cont'd)

- Wait a minute...We started with a uniform prior over both X_1 and X_2, saw the same number of "1"s as "2"s for X_2 in $\mathbf{d}$, and yet the marginal for X_2 is not $1/2$?!?!?!?!?
- The problem is that there are two parents for X_2 versus one for X_1:
  - X_1's prior of $beta(f_{11}; 1, 1)$ implies that in our prior, X_1 took the value 1 once in two trials
  - On the other hand, X_2's prior of two beta distributions implies that X_2 took the value 1 twice in four trials

Prior Equivalent Sample Size

Another Problem





Given the above Bayes net and the same data set
$\mathbf{d} = \{(1,2),(1,1),(2,1),(2,2),(2,1),(2,1),(1,2),(2,2)\}$,
what is $P(X_2 = 1)$?

Prior Equivalent Sample Size

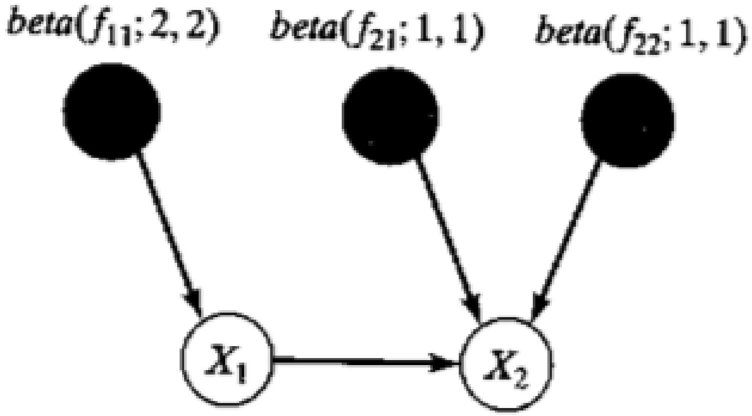
Another Problem (cont'd)

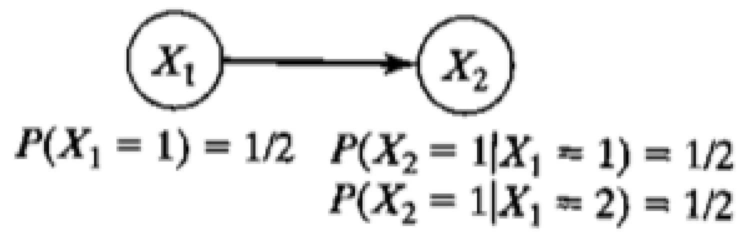
- Wait a minute...Now we have an embedded BN that's Markov equivalent to the previous one, but we get a different marginal?

- How do we fix this?

Prior Equivalent Sample Size

Equivalent Sample Size





- Changing X_1's prior to $beta(f_{11}; 2, 2)$ retains the prior probability of $1/2$ over X_1's values, but match its pseudosample size to that of X_2
- Given the above Bayes net and the same data set $\mathbf{d} = \{(1,2),(1,1),(2,1),(2,2),(2,1),(2,1),(1,2),(2,2)\}$, what is $P(X_2 = 1)$?
- Similar result if we double X_2's sample size in $X_2 \rightarrow X_1$ network

Prior Equivalent Sample Size

- Consider a binomial augmented Bayes net with densities $beta(f_{ij}; a_{ij}, b_{ij})$ for all i and j

- If there is some N such that for all i and j,

$$N_{ij} = a_{ij} + b_{ij} = P(\mathbf{pa}_{ij})N \quad ,$$

  then the network has equivalent sample size N

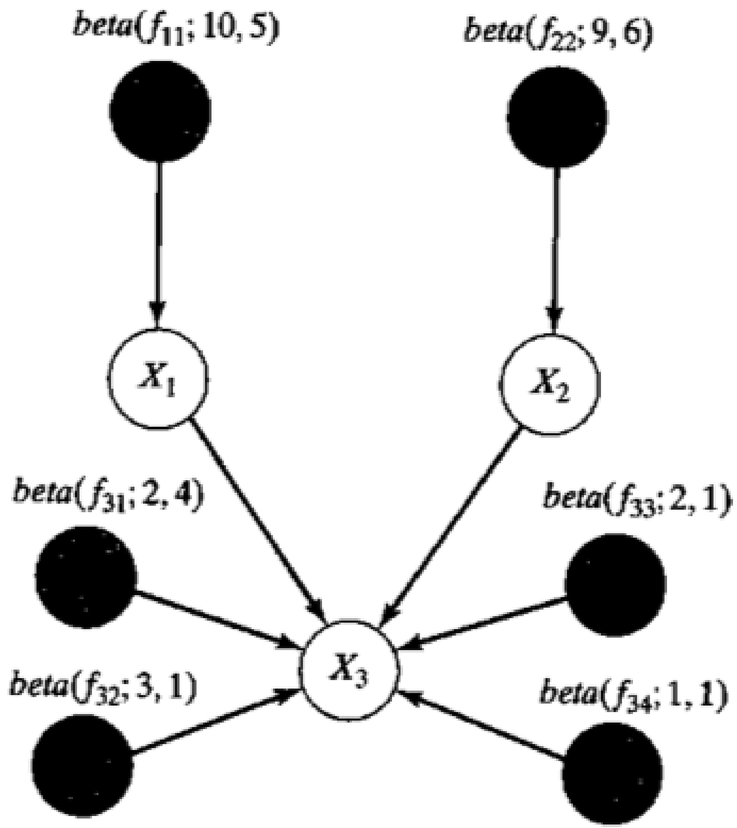
- $\mathbf{pa}_{ij}$ is an instantiation of the parents $\mathbf{NA}_i$ of node X_i

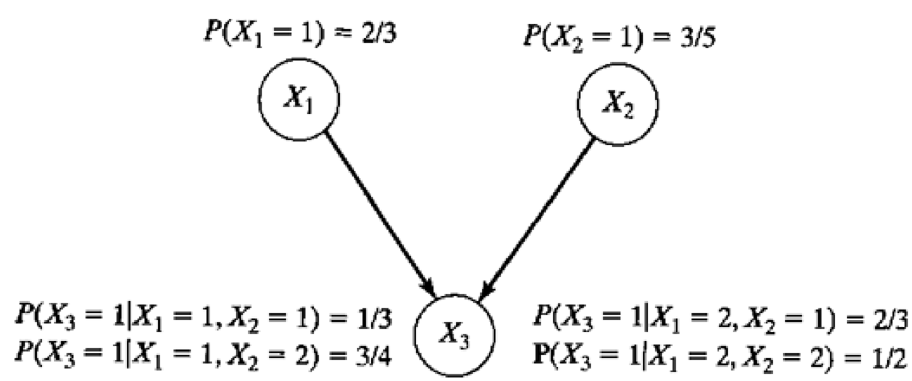
- If the network has an equivalent sample size N, then for each node X_i, $i \in \{1, \ldots, n\}$ (q_i is number of instantiations of X_i's parents),

$$\sum_{j=1}^{q_i} N_{ij} = \sum_{j=1}^{q_i} N \cdot P(\mathbf{pa}_{ij}) = N$$

Prior Equivalent Sample Size

Example ($N = 15$)



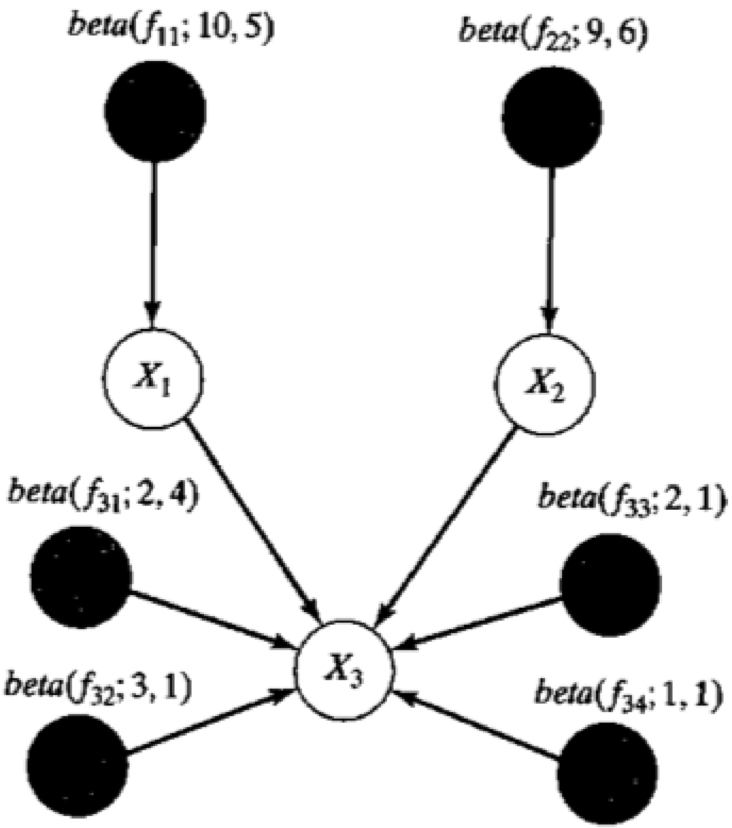


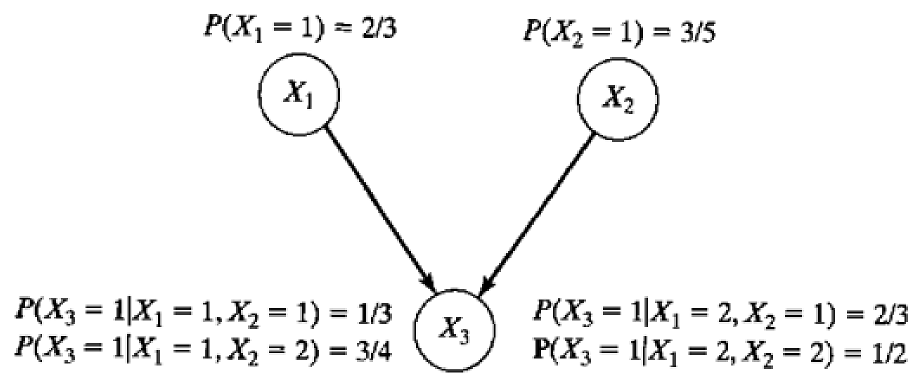
$$a_{11} + b_{11} = 10 + 5 = 15 \qquad N \cdot P(\mathbf{pa}_{11}) = 15 \cdot 1 = 15$$

$(\mathbf{pa}_{11} = \emptyset)$

Prior Equivalent Sample Size

Example ($N = 15$) (cont'd)



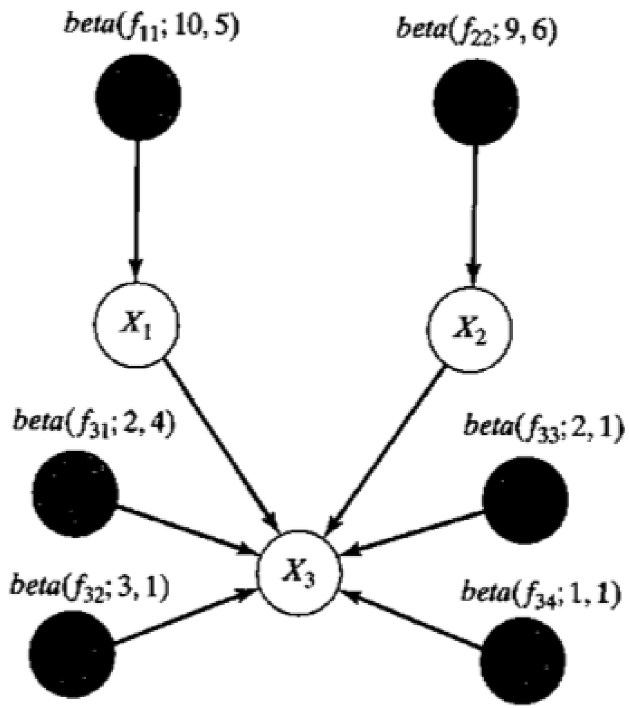


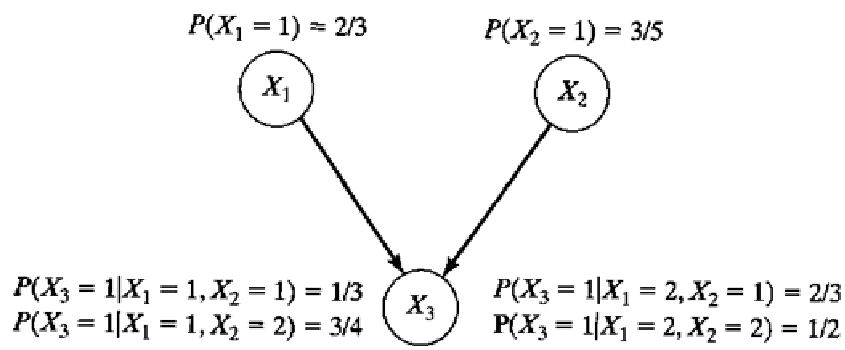
$$a_{22} + b_{22} = 9 + 6 = \color{red}{15} \qquad N \cdot P(\mathbf{pa}_{22}) = 15 \cdot 1 = \color{red}{15}$$

$(\mathbf{pa}_{22} = \emptyset)$

Prior Equivalent Sample Size

Example ($N = 15$) (cont'd)





$a_{31}+b_{31} = 2+4 = 6$ $\quad N \cdot P(\mathbf{pa}_{31}) = 15 \cdot P(X_1 = 1, X_2 = 1) = 15(2/3)(3/5) = 6$

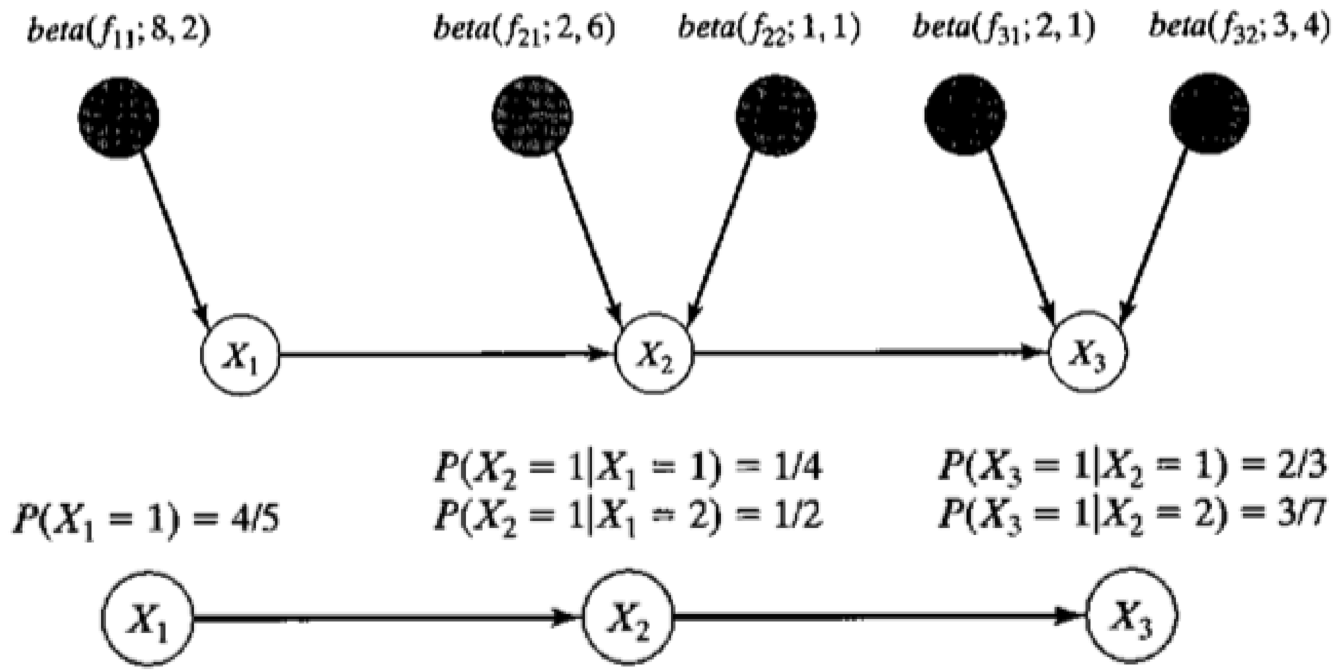
$a_{32}+b_{32} = 3+1 = 4$ $\quad N \cdot P(\mathbf{pa}_{32}) = 15 \cdot P(X_1 = 1, X_2 = 2) = 15(2/3)(2/5) = 4$

$a_{33}+b_{33} = 2+1 = 3$ $\quad N \cdot P(\mathbf{pa}_{33}) = 15 \cdot P(X_1 = 2, X_2 = 1) = 15(1/3)(3/5) = 3$

$a_{34}+b_{34} = 1+1 = 2$ $\quad N \cdot P(\mathbf{pa}_{34}) = 15 \cdot P(X_1 = 2, X_2 = 2) = 15(1/3)(2/5) = 2$

Prior Equivalent Sample Size

Group Exercise



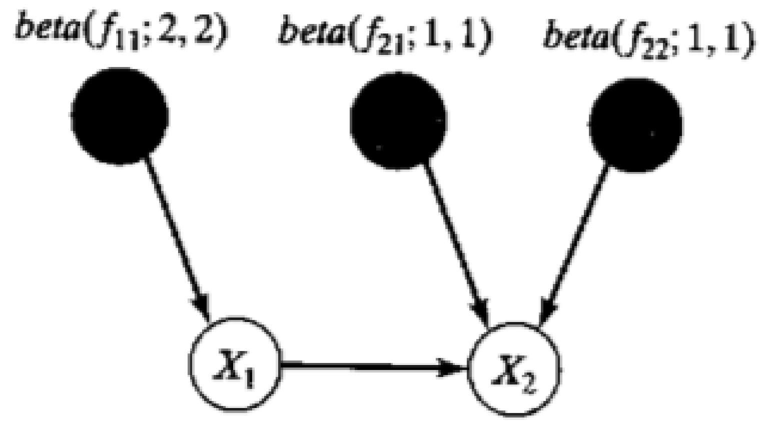
Does the above network have an equivalent sample size?

Prior Equivalent Sample Size

Creating a Network with an Equivalent Sample Size

Can get a uniform prior with pseudosample size N by setting

$$a_{ij} = b_{ij} = N/(2q_i)$$



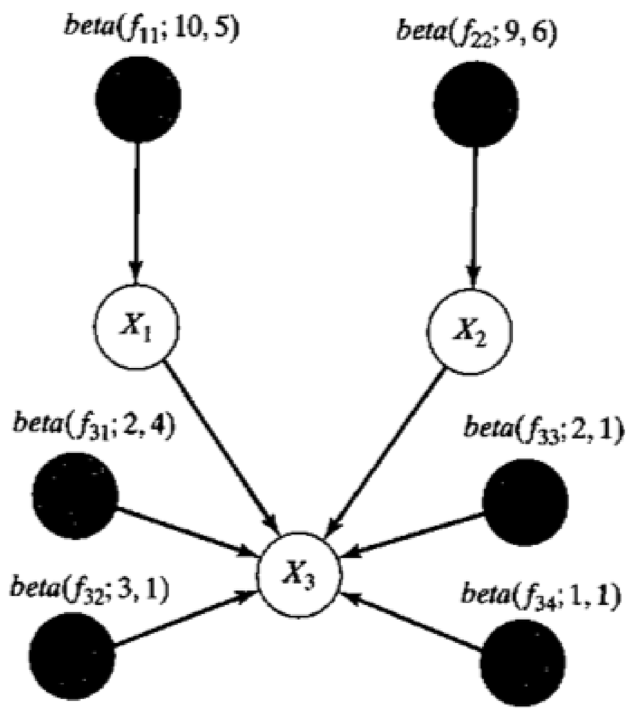
$q_1 = 1$ since $\mathrm{pa}_1 = \emptyset$, $q_2 = 2$ since $\mathrm{pa}_2 = \{\{1\}, \{2\}\}$; $N = 4$

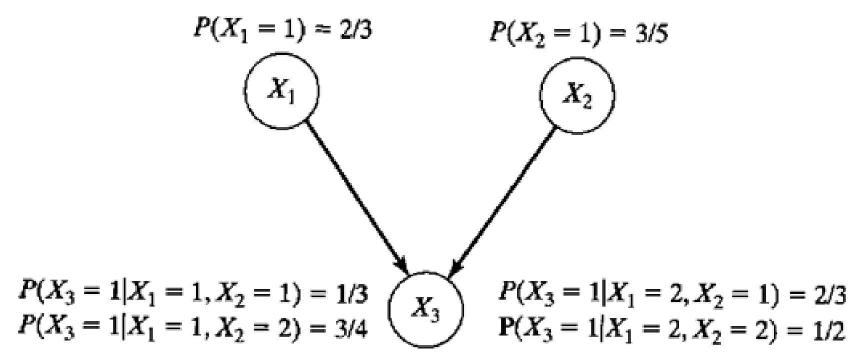
Prior Equivalent Sample Size

Creating a Network with an Equivalent Sample Size (cont'd)

Can get a nonuniform prior with pseudosample size N by setting

$$\begin{aligned} a_{ij} &= P(X_i = 1 \mid \mathbf{pa}_{ij}) P(\mathbf{pa}_{ij}) N \\ b_{ij} &= P(X_i = 2 \mid \mathbf{pa}_{ij}) P(\mathbf{pa}_{ij}) N \end{aligned}$$



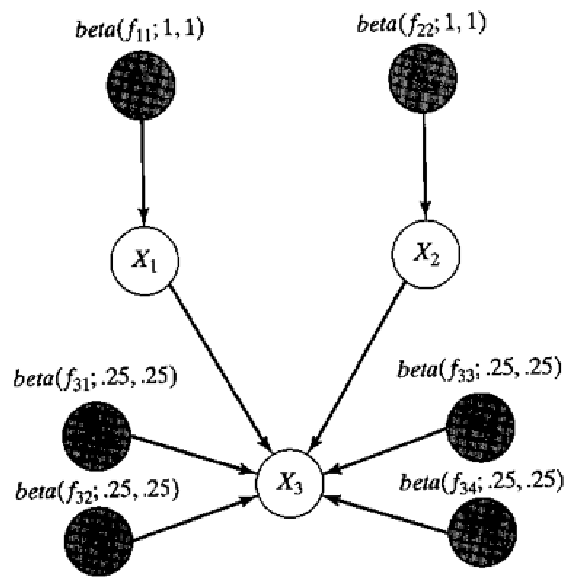


Probabilities in embedded network; $N = 15$

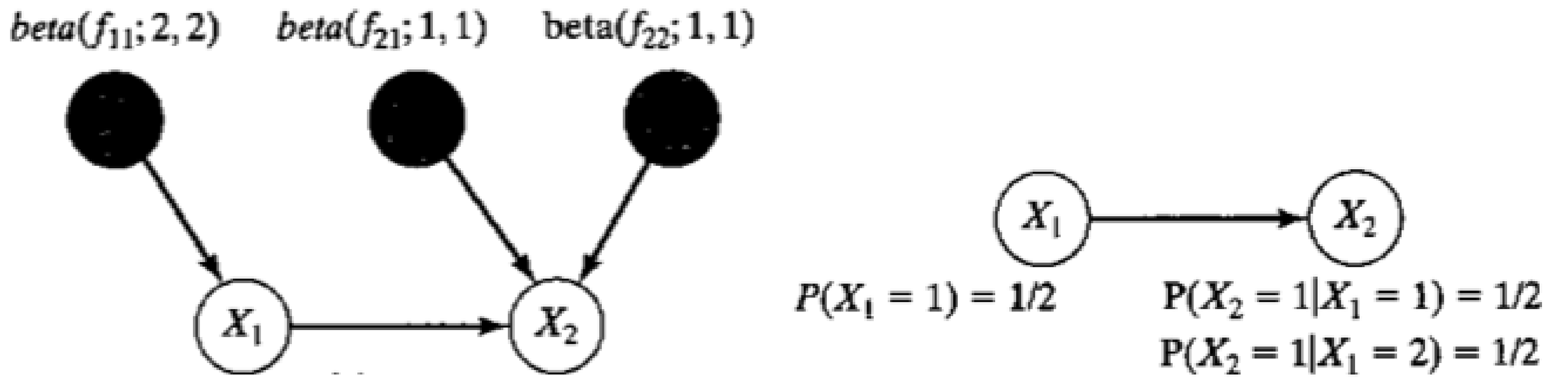
Prior Equivalent Sample Size

Choosing the Value of N

- We've established that $beta(f;1,1)$ is our ultimate uninformed prior
- But when establishing equivalent sample sizes, placing $beta(f;1,1)$ at nonroots resulted in stronger priors at the roots (e.g., $beta(f;2,2)$)
- To remain truly uninformed, recommended that we start with $beta(f;1,1)$ at the roots ($N=2$) and then use fractional parameters at the internal nodes (they still sum to 2)

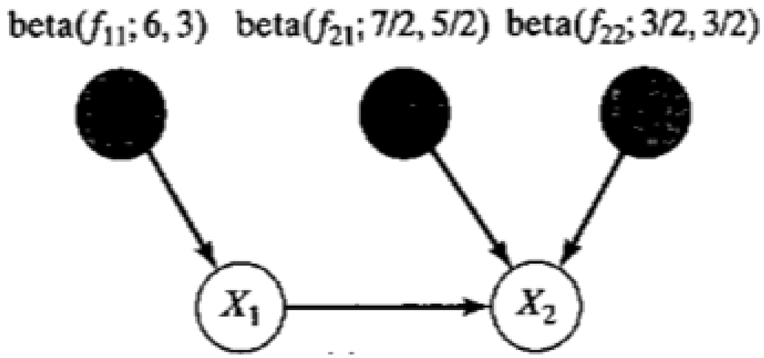


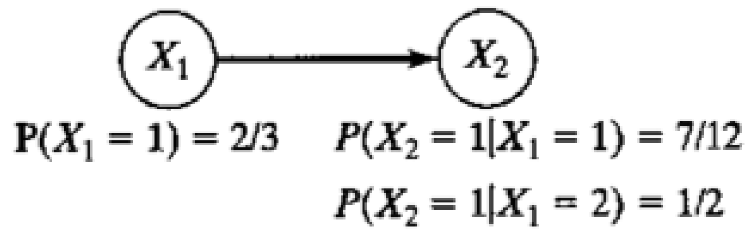
Handling Missing Attribute Values



- How do we update the Bayes net when we see partially-specified data $\mathbf{d} = \{(1, 1), (1, ?), (1, 1), (1, 2), (2, ?)\}$?
- Can handle specified values as before, e.g. number of times $X_1 = 1$ is $s_{11} = 4$, yielding $beta(f_{11}; 6, 3)$
- Since we already have a probability distribution over the values, we can fractionalize the examples with unspecified attributes, e.g. number of times $X_1 = 1$ and $X_2 = 1$ is $s_{21} = 2 + 1/2$, and number of times times $X_1 = 1$ and $X_2 = 2$ is $t_{21} = 1 + 1/2$, yielding $beta(f_{21}; 7/2, 5/2)$
  - The "1/2" fractions came from $P(X_2 = 1 \mid X_1 = 1)$, etc., from the embedded network

Handling Missing Attribute Values





- After updating, get the above network
- Hmmmmm. Now $P(X_2 = 1 \mid X_1 = 1) \neq 1/2$, which is what we used in our fractional update
- What if we used the new probabilities to fractionalize the data?
- Then we still get $s_{11} = 4$ and $s_{22} = 1/2$ (why?), but now have $s_{21} = 2 + 7/12$ and $t_{21} = 1 + 5/12$

  $\Rightarrow \; beta(f_{11}; 6, 3), beta(f_{21}; 43/12, 29/12), beta(f_{22}; 3/2, 3/2)$

  $\Rightarrow \; P(X_2 = 1 \mid X_1 = 1) = 43/72$

- Can repeat again, and again, ...
- What does this look like?

Handling Missing Attribute Values

The Algorithm

- Yes, it's our old friend, the EM Algorithm!
- First, initialize f'_{ij} either to $a_{ij}/(a_{ij}+b_{ij})$ (deterministic) or to arbitrary values (to avoid local optima)
- Then compute ($M =$ number of examples)

$$\begin{aligned} s'_{ij} &= E(s_{ij} \mid \mathbf{d}, \mathbf{f}') = \sum_{h=1}^{M} P(X_i^{(h)} = 1, \mathbf{pa}_{ij} \mid \mathbf{x}^{(h)}, \mathbf{f}') \\ t'_{ij} &= E(t_{ij} \mid \mathbf{d}, \mathbf{f}') = \sum_{h=1}^{M} P(X_i^{(h)} = 2, \mathbf{pa}_{ij} \mid \mathbf{x}^{(h)}, \mathbf{f}') \end{aligned}$$

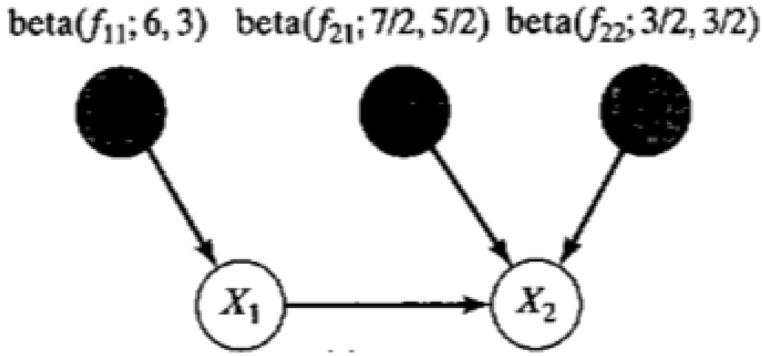
- Then compute

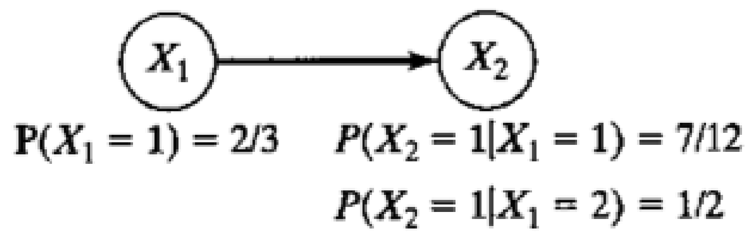
$$\overbrace{f'_{ij} = \frac{a_{ij} + s'_{ij}}{a_{ij} + s'_{ij} + b_{ij} + t'_{ij}}}^{\text{MAP: } \rho(\mathbf{f} \mid \mathbf{d})} \qquad \text{or} \qquad \overbrace{f'_{ij} = \frac{s'_{ij}}{s'_{ij} + t'_{ij}}}^{\text{ML: } P(\mathbf{d} \mid \mathbf{f})}$$

Handling Missing Attribute Values

The Algorithm

Example





$$\mathbf{d} = \{(1,1),(1,?),(1,1),(1,2),(2,?)\}, \mathbf{f}' = \{2/3, 7/12, 1/2\}$$

$$\begin{aligned}
s'_{21} &= E(s_{21} \mid \mathbf{d}, \mathbf{f}') = \sum_{h=1}^{5} P(X_1^{(h)} = 1, X_2^{(h)} = 1 \mid \mathbf{x}^{(h)}, \mathbf{f}') \\
&= P(X_1^{(1)} = 1, X_2^{(1)} = 1 \mid (1,1), \mathbf{f}') + P(X_1^{(2)} = 1, X_2^{(2)} = 1 \mid (1,?), \mathbf{f}') \\
&\quad + P(X_1^{(3)} = 1, X_2^{(3)} = 1 \mid (1,1), \mathbf{f}') + P(X_1^{(4)} = 1, X_2^{(4)} = 1 \mid (1,2), \mathbf{f}') \\
&\quad + P(X_1^{(5)} = 1, X_2^{(5)} = 1 \mid (2,?), \mathbf{f}') \\
&= 1 + 7/12 + 1 + 0 + 0 = 31/12
\end{aligned}$$