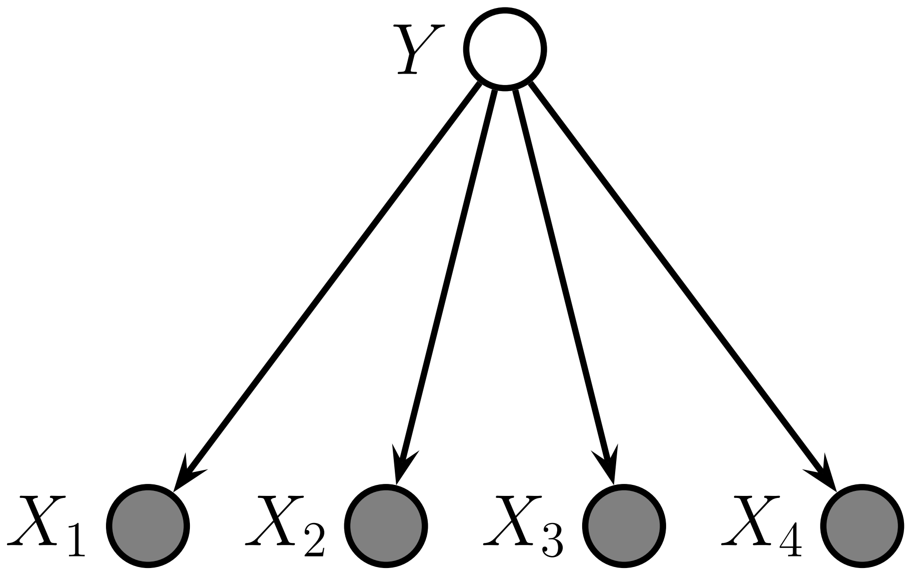
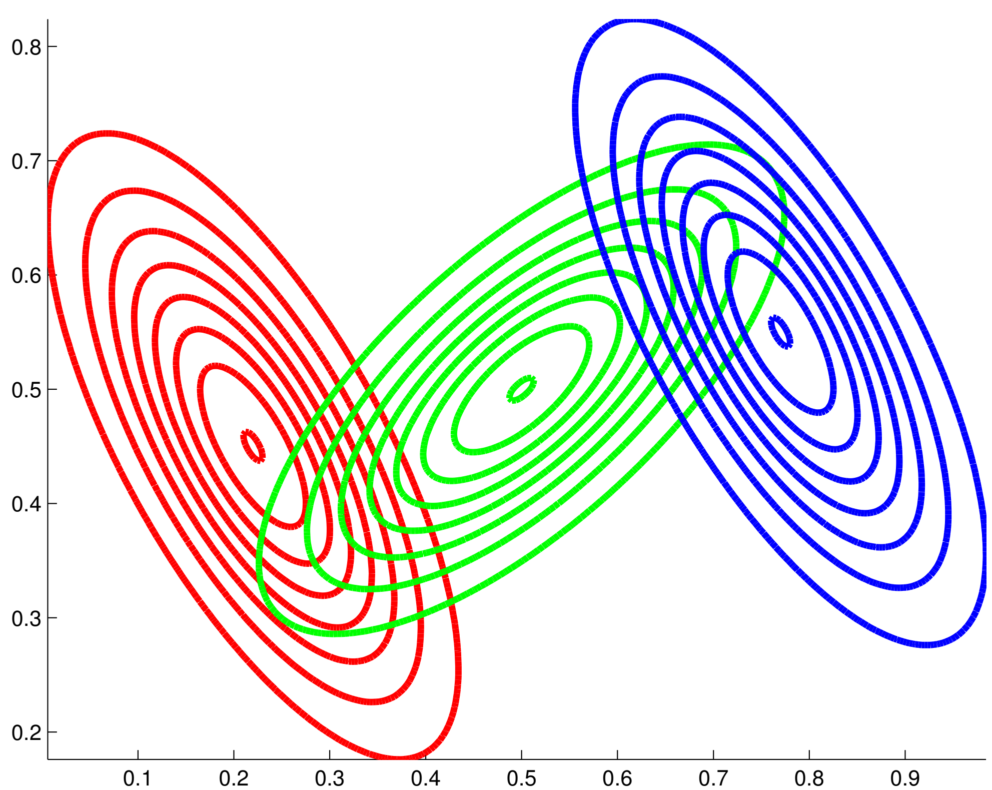
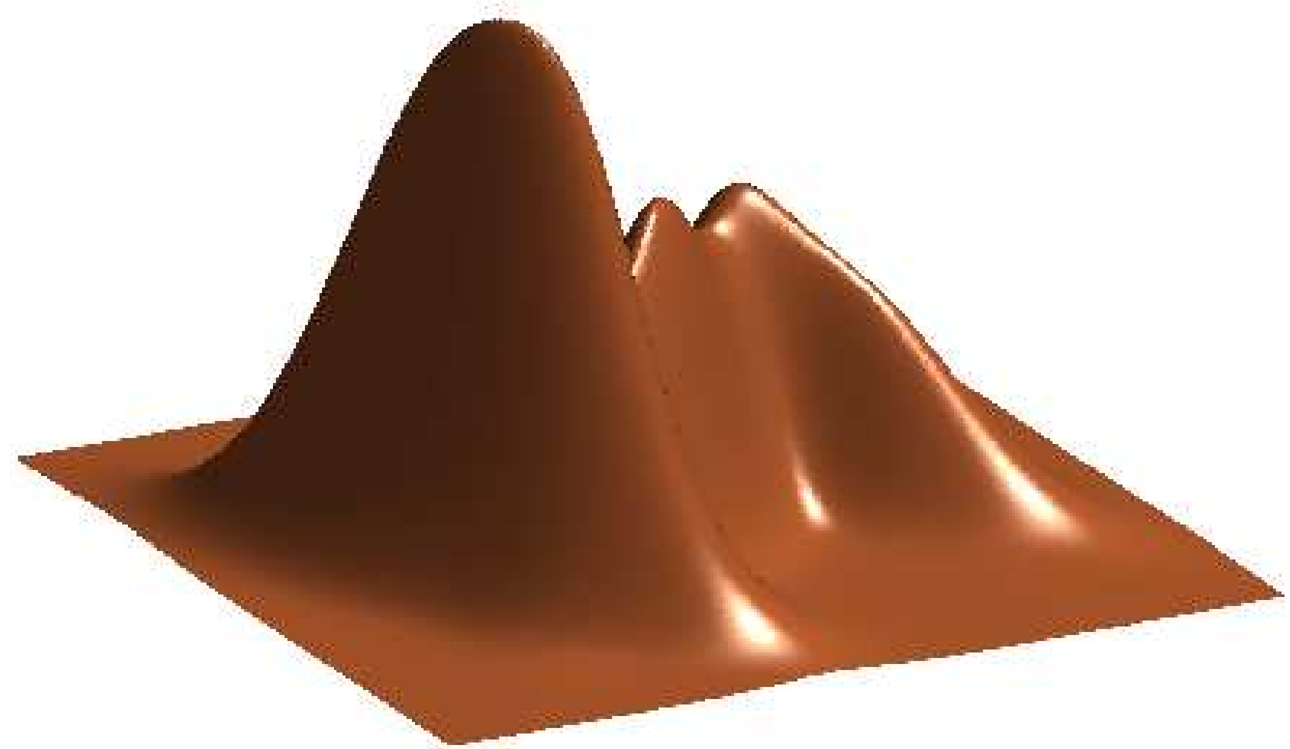
Chapter 11: Mixture models and EM


parametric assumptions
my data

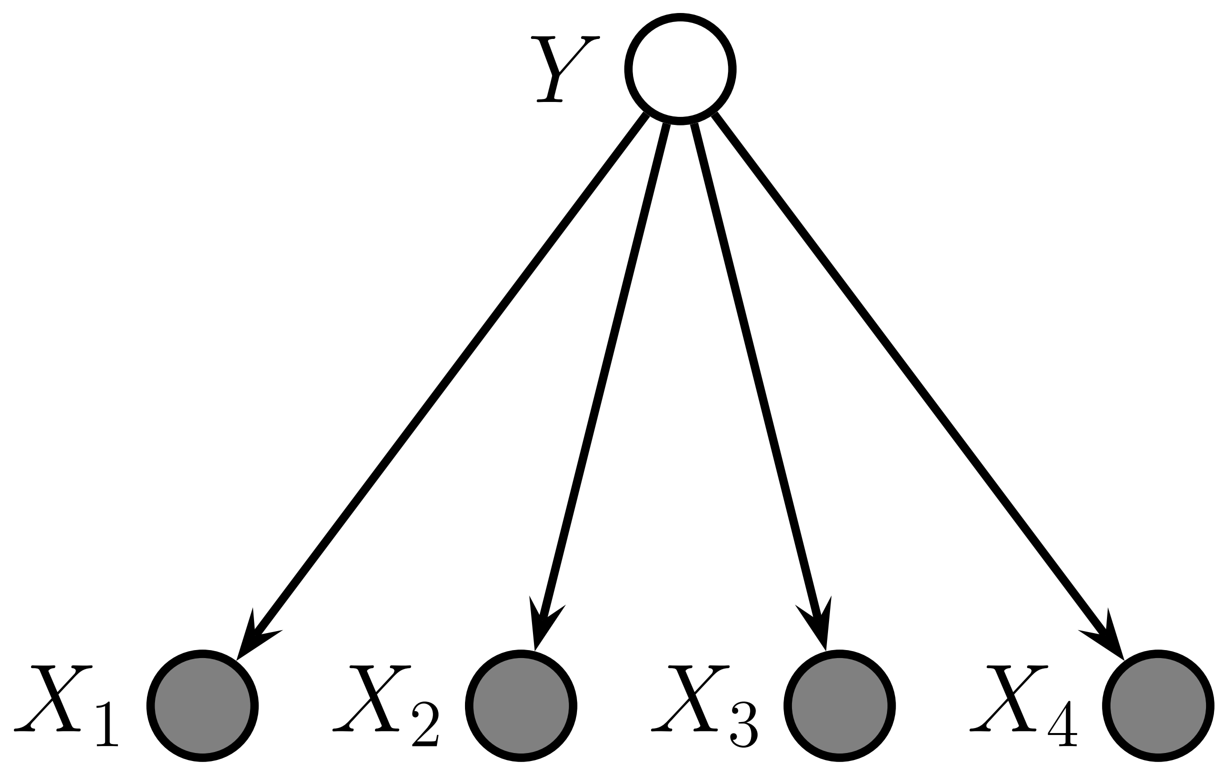
Example: Mixture of Gaussians







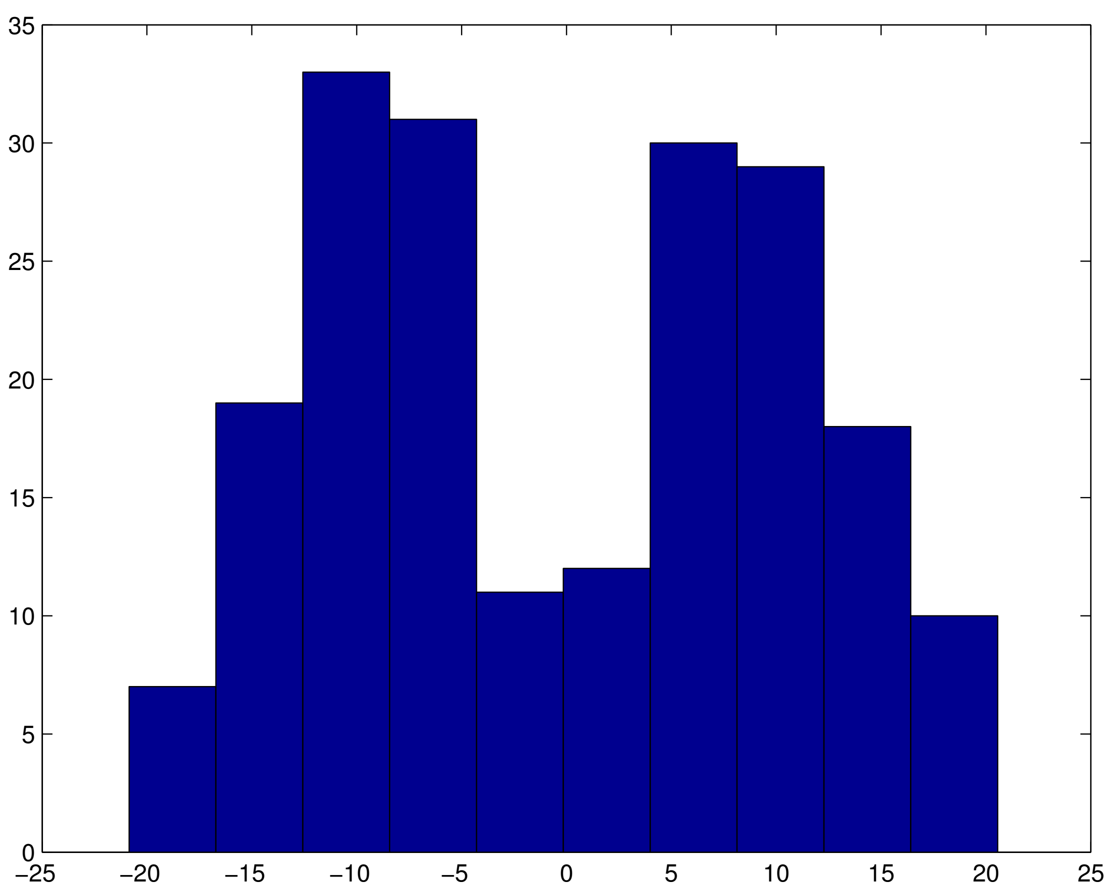
Example: Clustering with binary features



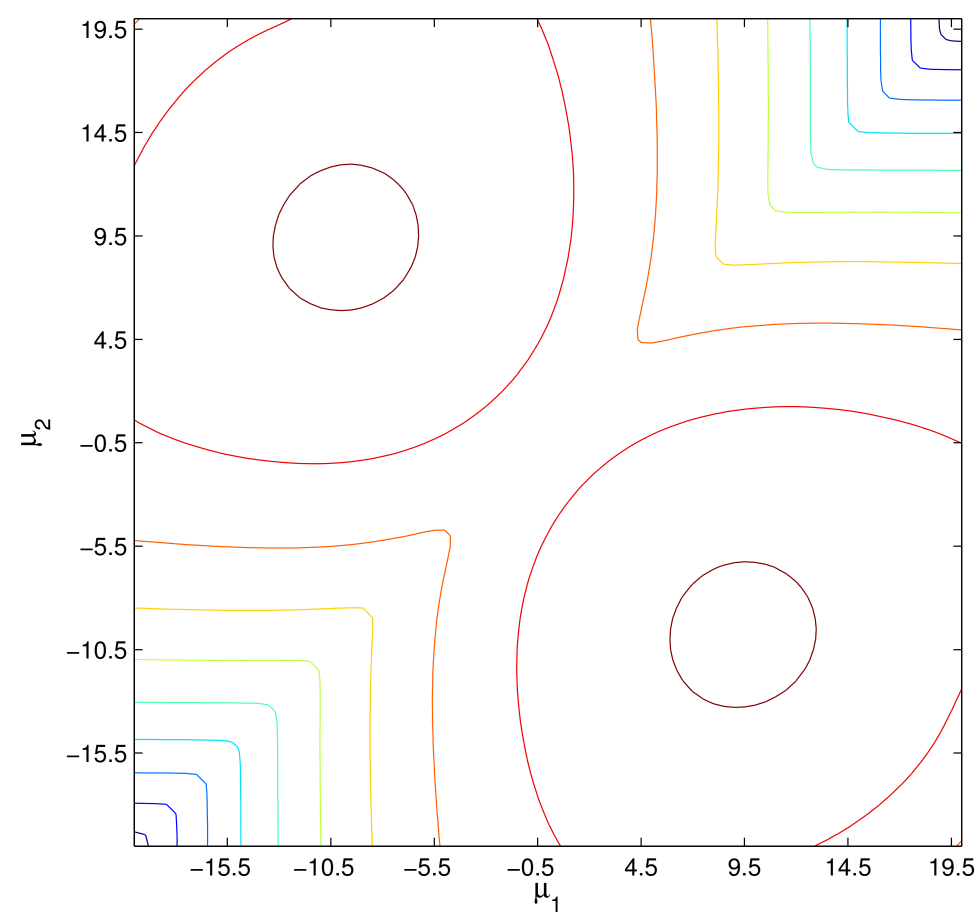
Parameter estimation and unidentifiability

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \left[\sum_{\mathbf{z}_i} p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta}) \right]$$

$P(\mu_1|D)$



μ_1



EM algorithm

$$Q(\theta, \theta^{t-1}) \triangleq \mathbb{E}_{\mathbf{z} \sim \theta^{t-1}} \left[\sum_{i=1}^{N} \log p(\mathbf{x}_i, \mathbf{z}_i | \theta^t) \right]$$

$$\theta^t \leftarrow \mathrm{argmax}_\theta Q(\theta, \theta^{t-1})$$

EM algorithm for GMMs

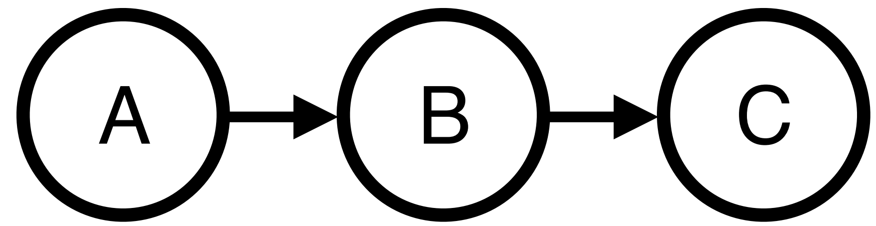
$$
\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) &\triangleq \mathbb{E}\left[\sum_i \log p(\mathbf{x}_i, z_i|\boldsymbol{\theta})\right] \\
&= \sum_i \mathbb{E}\left[\log\left[\prod_{k=1}^{K} (\pi_k p(\mathbf{x}_i|\boldsymbol{\theta}_k))^{\mathbb{I}(z_i=k)}\right]\right] \\
&= \sum_i \sum_k \mathbb{E}\left[\mathbb{I}(z_i = k)\right] \log[\pi_k p(\mathbf{x}_i|\boldsymbol{\theta}_k)] \\
&= \sum_i \sum_k p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}^{t-1}) \log[\pi_k p(\mathbf{x}_i|\boldsymbol{\theta}_k)] \\
&= \sum_i \sum_k r_{ik} \log \pi_k + \sum_i \sum_k r_{ik} \log p(\mathbf{x}_i|\boldsymbol{\theta}_k)
\end{aligned}
$$

$$
r_{ik} = \frac{\pi_k p(\mathbf{x}_i|\boldsymbol{\theta}_k^{(t-1)})}{\sum_{k'} \pi_{k'} p(\mathbf{x}_i|\boldsymbol{\theta}_{k'}^{(t-1)})}
$$

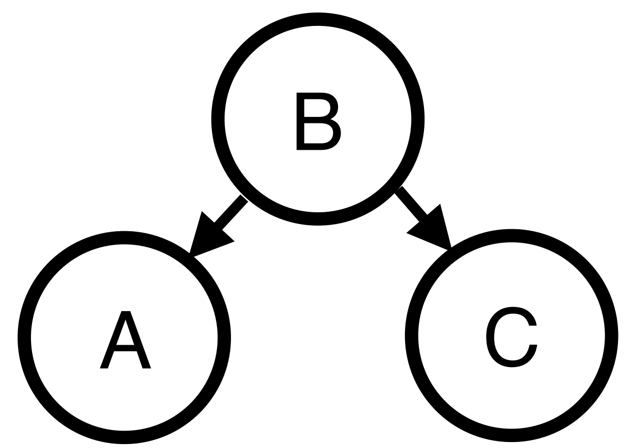
d-separation

d-separation

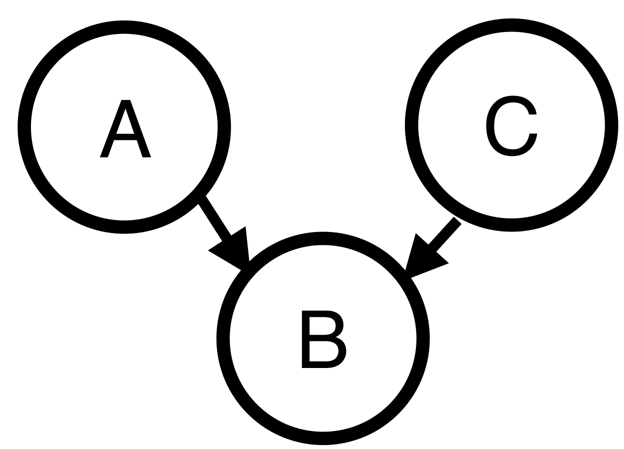
$$A \not\perp C$$



$$A \not\perp C$$

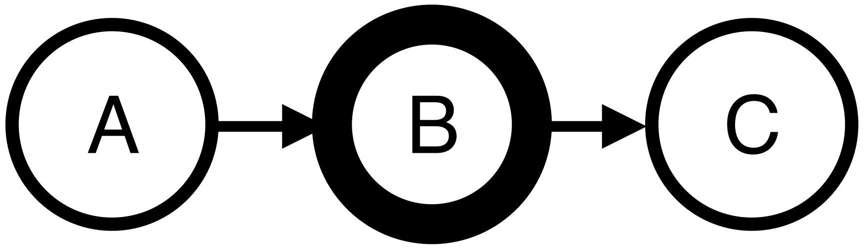


$$A \perp C$$

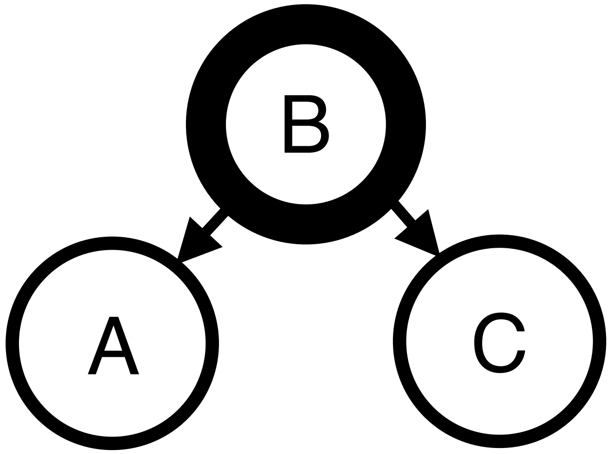


d-separation

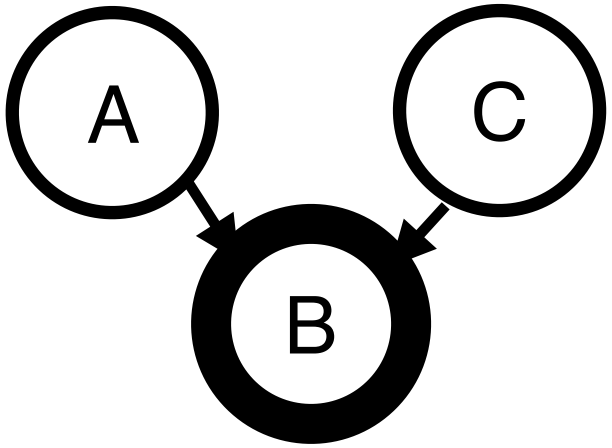
$$A \perp C \mid B$$



$$A \perp C \mid B$$



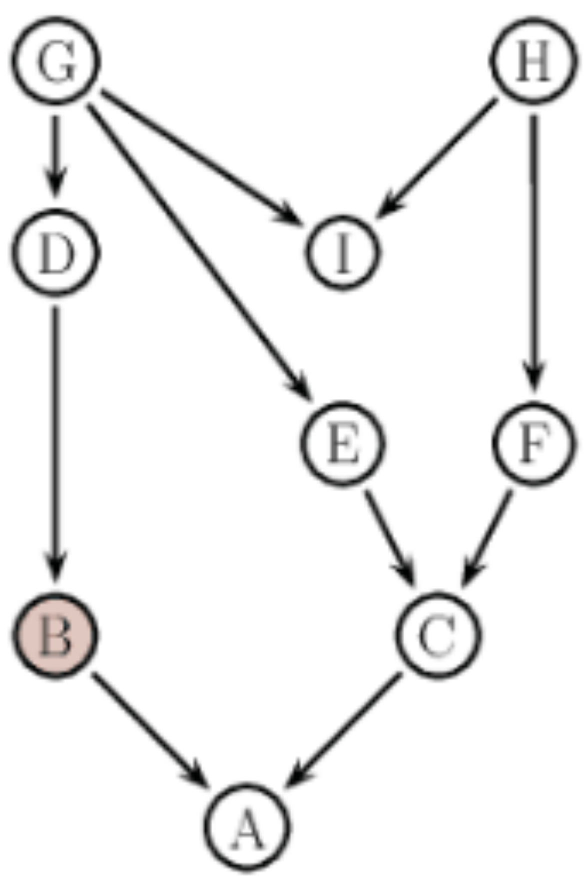
$$A \not\perp C \mid B$$



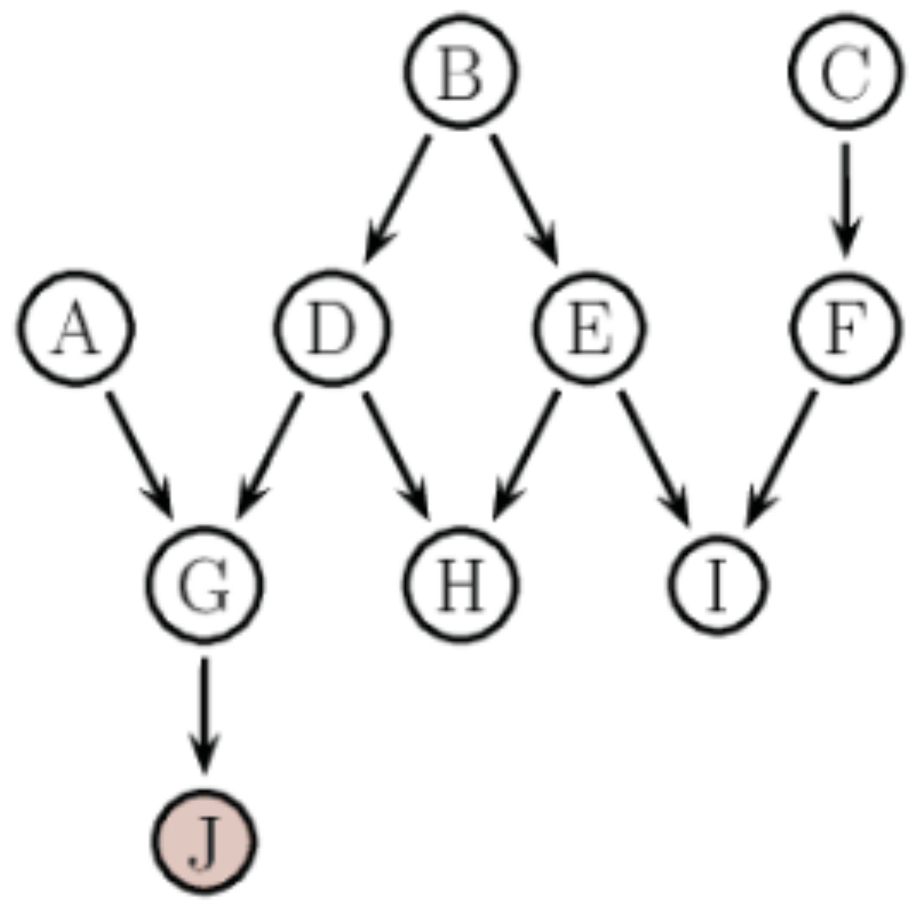
d-separation

Exercise: d-separation

In each BN, list all variables that are independent of *A*.



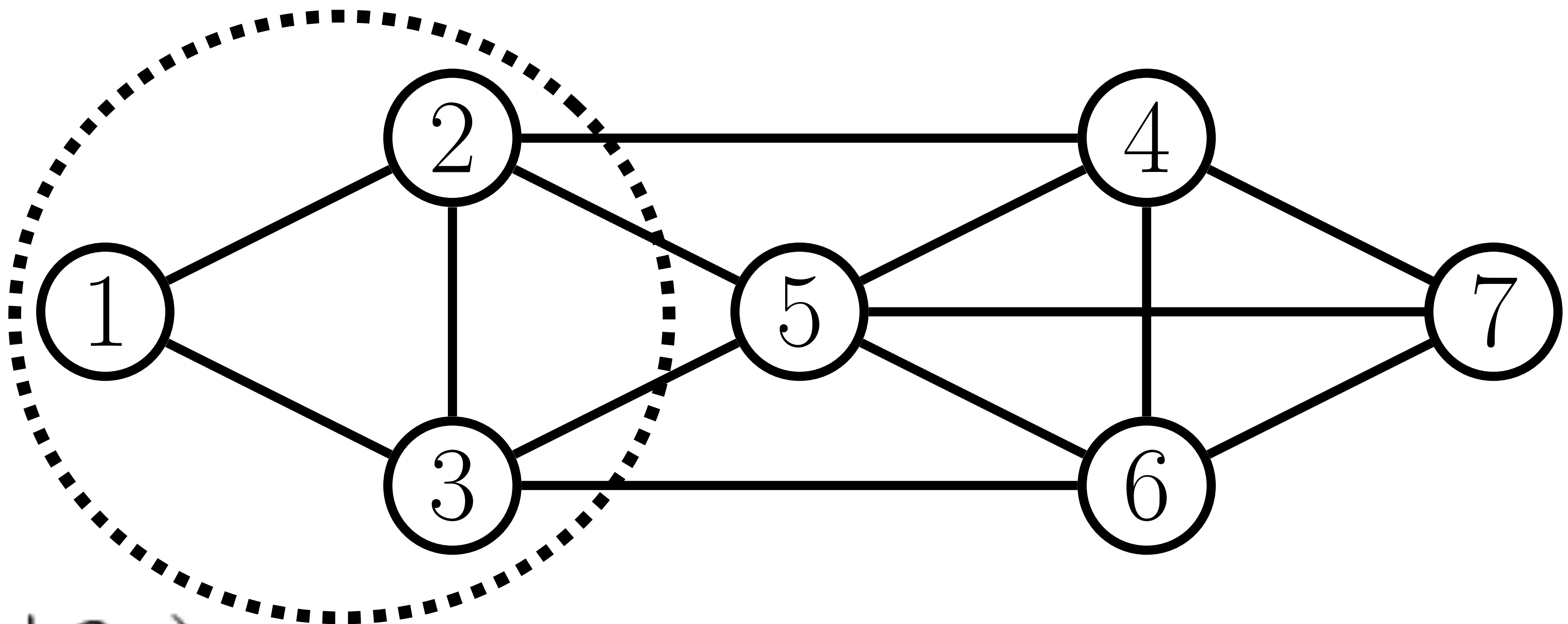
(a)



(b)

Chapter 19: Undirected graphical models

Undirected graphical models

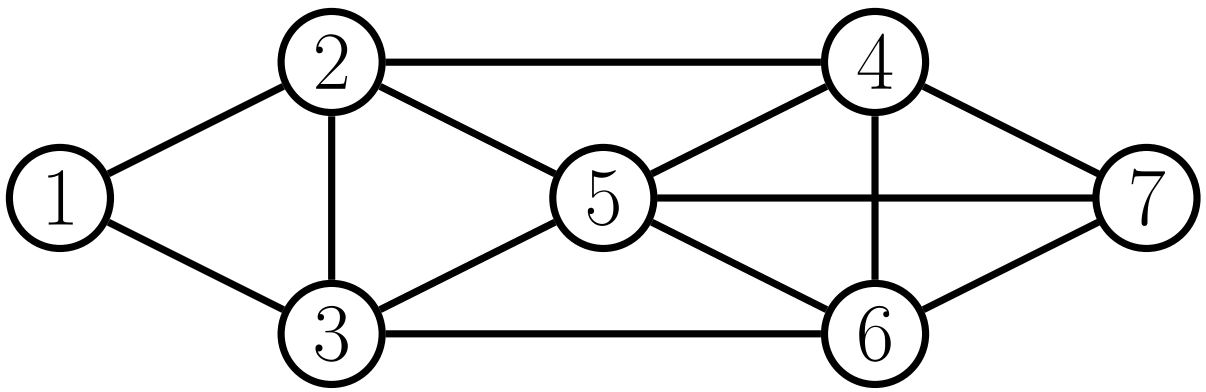


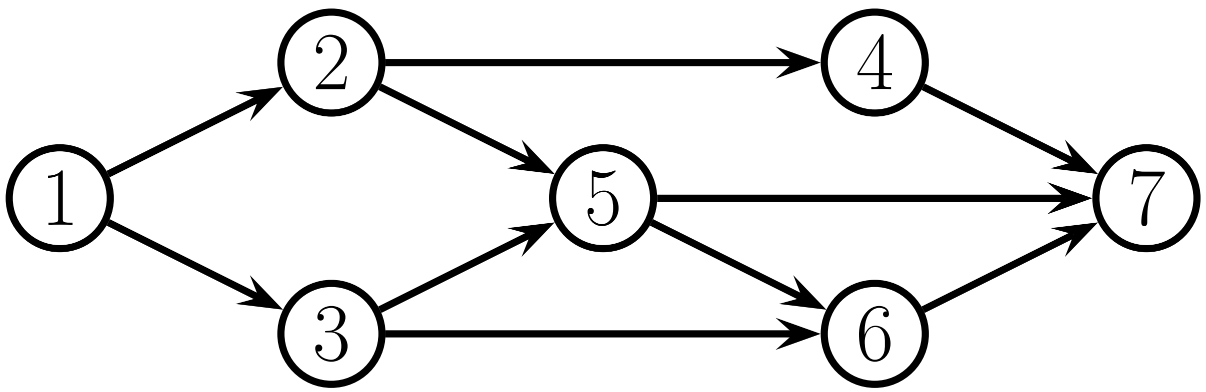
$$\psi_c(\mathbf{y}_c|\boldsymbol{\theta}_c)$$

$$p(\mathbf{y}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{y}_c|\boldsymbol{\theta}_c)$$

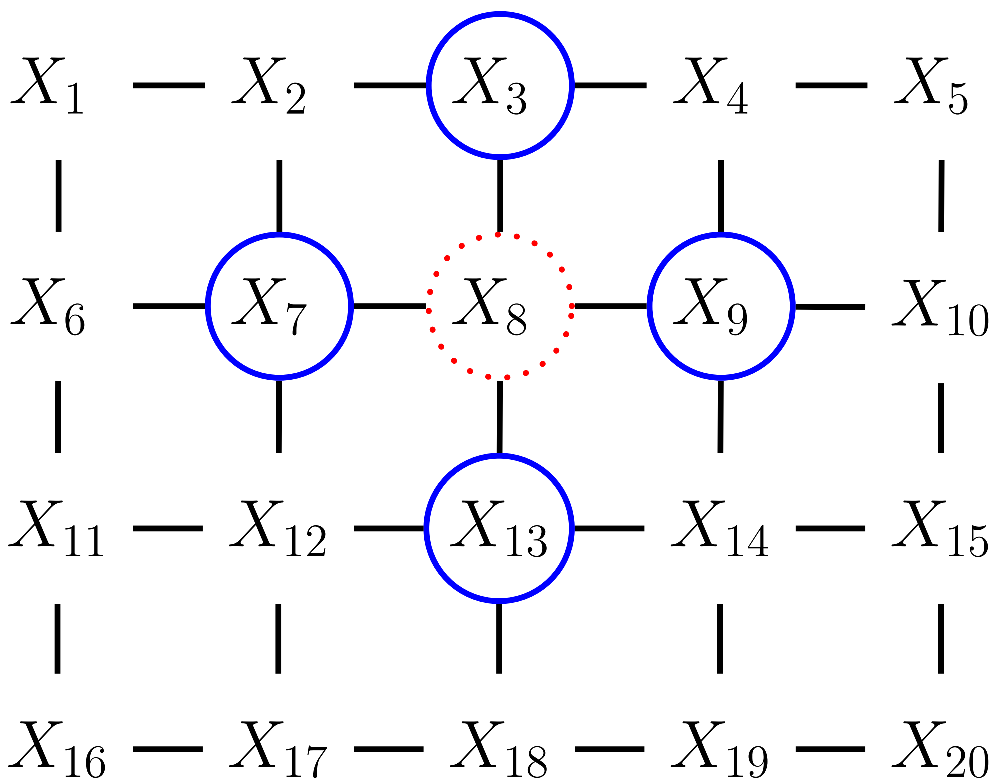
$$Z(\boldsymbol{\theta}) \triangleq \sum_{\mathbf{y}} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{y}_c|\boldsymbol{\theta}_c)$$

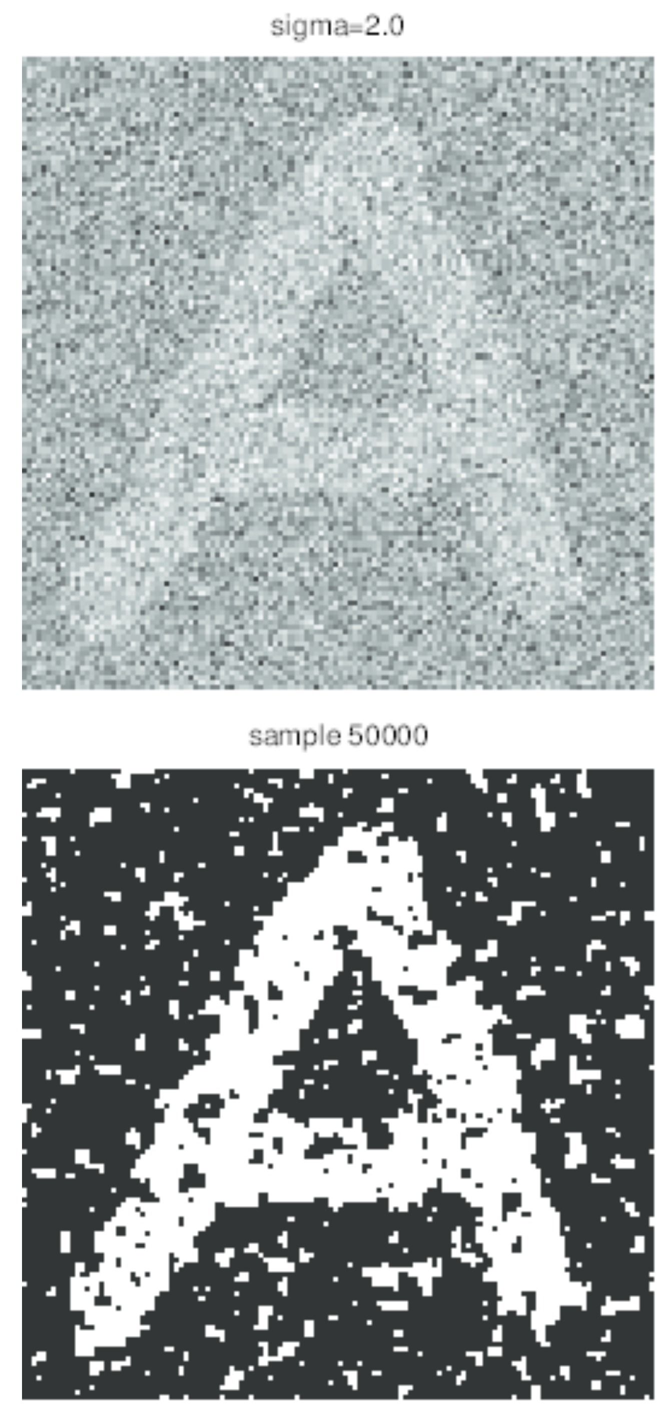
Undirected graphical models





Ising model

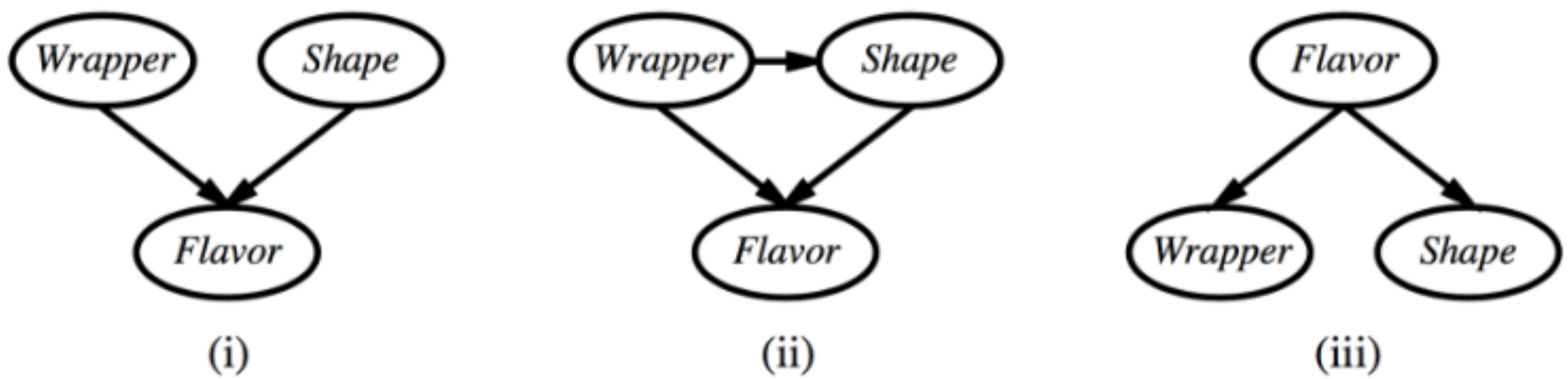

X_1 X_2 X_3 X_4 X_5
X_6 X_7 X_8 X_9 X_{10}
X_{11} X_{12} X_{13} X_{14} X_{15}
X_{16} X_{17} X_{18} X_{19} X_{20}


sigma=2.0
sample 50000

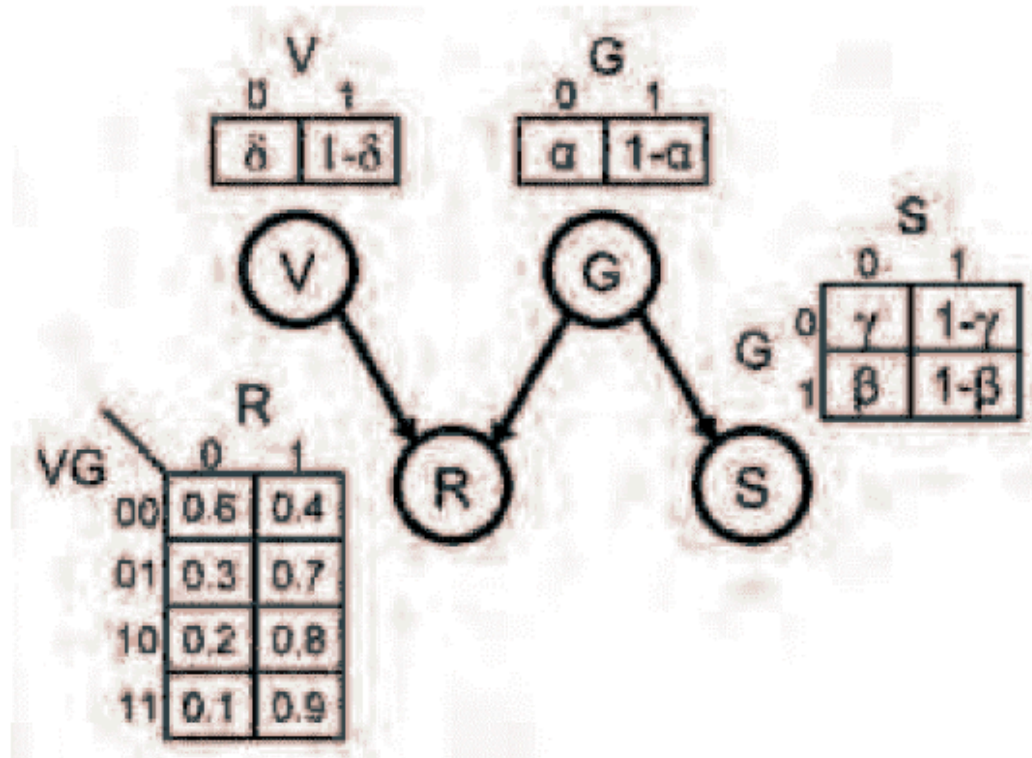
The Surprise Candy company makes candy in two flavors: 70% are strawberry flavor and 30% are anchovy flavor. Each new piece of candy starts out with a round shape; as it moves down the production line, a machine randomly selects a certain percentage to be trimmed into a square; then, each piece is wrapped in a wrapper whose color is chosen randomly to be red or brown. 80% of strawberry candies are round and 80% have a red wrapper, while 90% of the anchovy candies are square and 90% have a brown wrapper. All candies are sold individually in sealed, identical black boxes.

Now you, the customer, have just bought a Surprise candy at the store but have not yet opened the box.

Consider these three Bayes nets.



a) Which network(s) can correctly represent p(Flavor, Wrapper, Shape)?
b) Which network is the best representation?
c) True/False: Network (i) asserts that p(Wrapper | Shape) = p(Wrapper).
d) What is the probability that your candy has a red wrapper?
e) In the box is a round candy with a red wrapper. What is the probability that the flavor is strawberry?
f) An unwrapped strawberry candy is worth *x* on the open market and anchovy is worth *a*. Write an expression for the value of an unopened candy box.



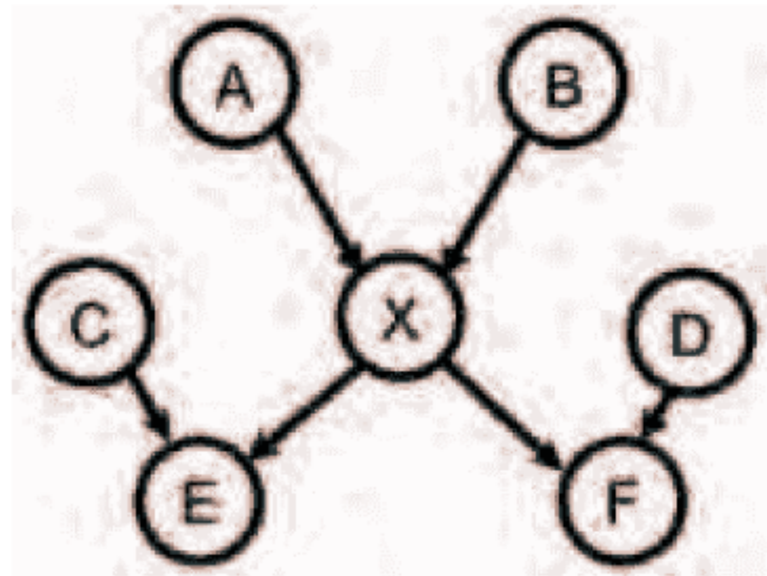
Exercise 10.5 Bayes nets for a rainy day

(Source: Nando de Freitas.). In this question you must model a problem with 4 binary variables: $G =$"gray", $V =$"Vancouver", $R =$"rain" and $S =$"sad". Consider the directed graphical model describing the relationship between these variables shown in Figure 10.15(a).

a. Write down an expression for $P(S = 1|V = 1)$ in terms of $\alpha, \beta, \gamma, \delta$.
b. Write down an expression for $P(S = 1|V = 0)$. Is this the same or different to $P(S = 1|V = 1)$? Explain why.
c. Find maximum likelihood estimates of α, β, γ using the following data set, where each row is a training case. (You may state your answers without proof.)

V	G	R	S
1	1	1	1
1	1	0	1
1	0	0	0

(10.61)



Exercise 10.1 Marginalizing a node in a DGM

(Source: Koller.)

Consider the DAG G in Figure 10.14(a). Assume it is a minimal I-map for $p(A, B, C, D, E, F, X)$. Now consider marginalizing out X. Construct a new DAG G' which is a minimal I-map for $p(A, B, C, D, E, F)$. Specify (and justify) which extra edges need to be added.

Exercise 11.3 EM for mixtures of Bernoullis

- Show that the M step for ML estimation of a mixture of Bernoullis is given by

$$\mu_{kj} = \frac{\sum_i r_{ik} x_{ij}}{\sum_i r_{ik}} \tag{11.116}$$

- Show that the M step for MAP estimation of a mixture of Bernoullis with a $\beta(\alpha, \beta)$ prior is given by

$$\mu_{kj} = \frac{(\sum_i r_{ik} x_{ij}) + \alpha - 1}{(\sum_i r_{ik}) + \alpha + \beta - 2} \tag{11.117}$$