CS340 Machine learning
Final review

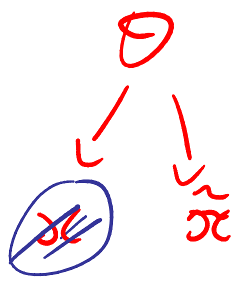
Covered in midterm review

- I – Basics:
  - Statistics (MLE, posteriors, Bayes factors, model selection, etc)
  - Info theory
  - Decision theory

Outline

- II- Models
  - Generative vs discriminative
  - Naïve Bayes
  - MVN
  - Markov chains
  - DGMs, including expert systems
  - UGMs, including Ising models
- III – Algorithms
  - Gibbs sampling

Unconditional density models



Eg x ~ bernoulli, theta ~ beta

x ~ multinomial, theta ~ dir

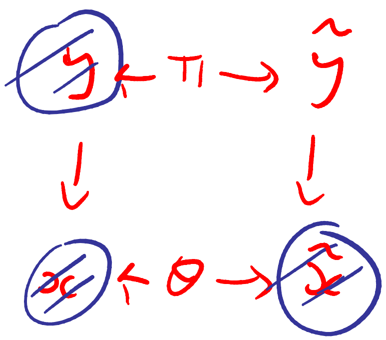
x ~ multinomial, theta ~ mixture of dir

x ~ gaussian, theta = (mu, lambda) ~ NormalGamma

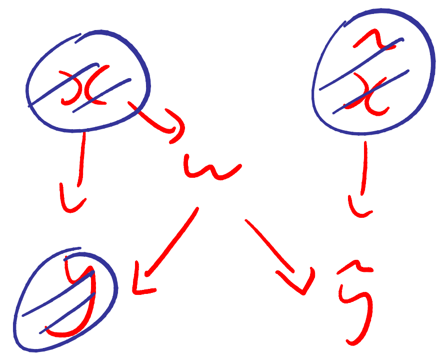
x ~ MVN, theta = (mu, Lambda) ~ NormalWishart

Generative vs discriminative models

Generative y->x



Discriminative x->y



$$p(\mathbf{x}, y|\boldsymbol{\pi}, \boldsymbol{\theta}) = p(y|\boldsymbol{\pi})p(\mathbf{x}|y, \boldsymbol{\theta})$$

$$p(y|\mathbf{x}, \mathbf{w})$$

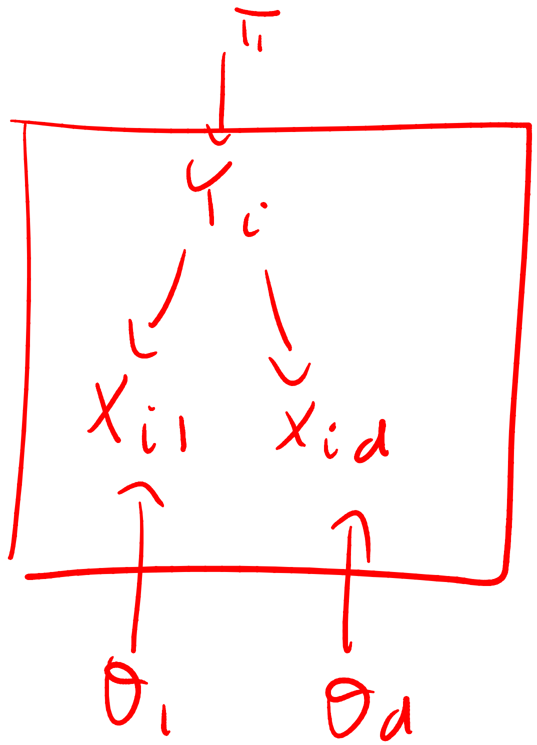
P(x|y) = class conditional density

Eg fully factored (naïve Bayes)

Markov chain

full covariance Gaussian

Naïve Bayes



P(xj|Y=c) = bernoulli, gaussian, …
Compute p(theta_j |D)
Handle missing data
Log sum exp trick

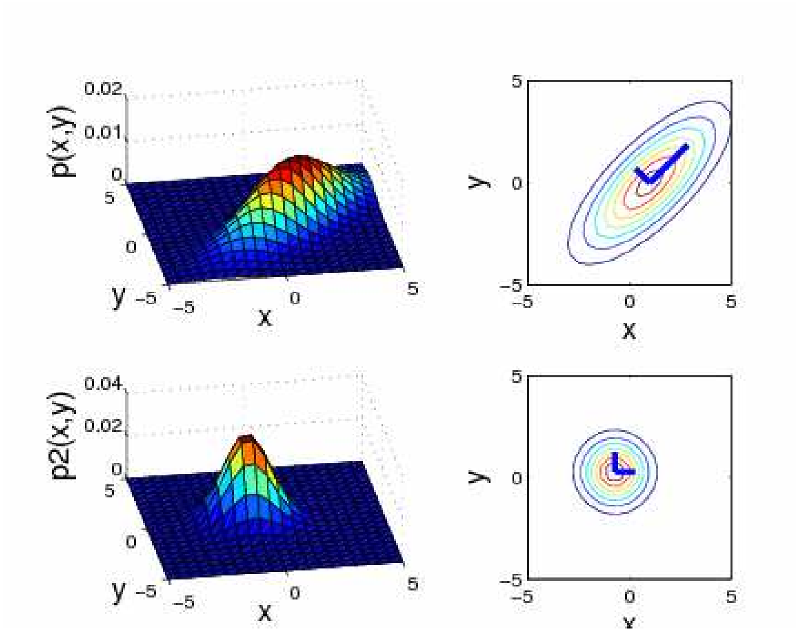
$$p(Y=c|x,\theta,\pi) \quad \propto \quad \exp\left[\log \pi_c + \sum_i I(x_i=1)\log\theta_{ic} + I(x_i=0)\log(1-\theta_{ic})\right]$$

$$\begin{aligned} x' &= [1, I(x_1=1), I(x_1=0), \ldots, I(x_d=1), I(x_d=0)] \\ \beta_c &= [\log\pi_c, \log\theta_{1c}, \log(1-\theta_{1c}), \ldots, \log\theta_{dc}, \log(1-\theta_{dc})] \end{aligned}$$

$$p(Y=c|x,\beta) = \frac{\exp[\beta_c^T x']}{\sum_{c'}\exp[\beta_{c'}^T x']}$$

Becomes sigmoid in 2-class case

Multivariate normal



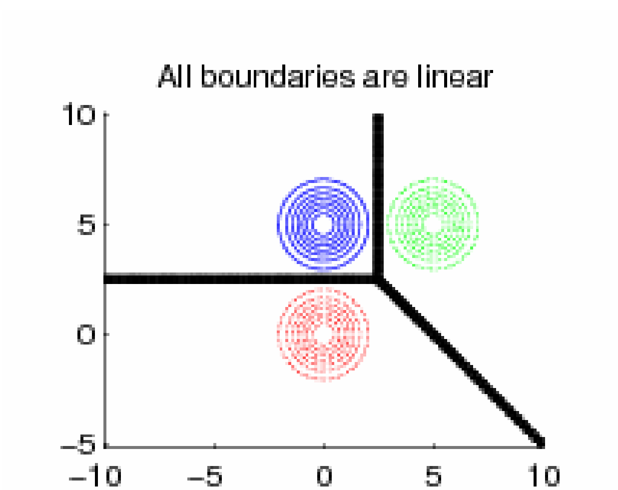
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\Sigma) \stackrel{\text{def}}{=} \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})]$$

MLE: $\hat{\mu} = \frac{1}{n}\sum_i x_i \qquad \hat{\Sigma} = \frac{1}{n}\sum_i (x_i - \bar{x})(x_i - \bar{x})^T$

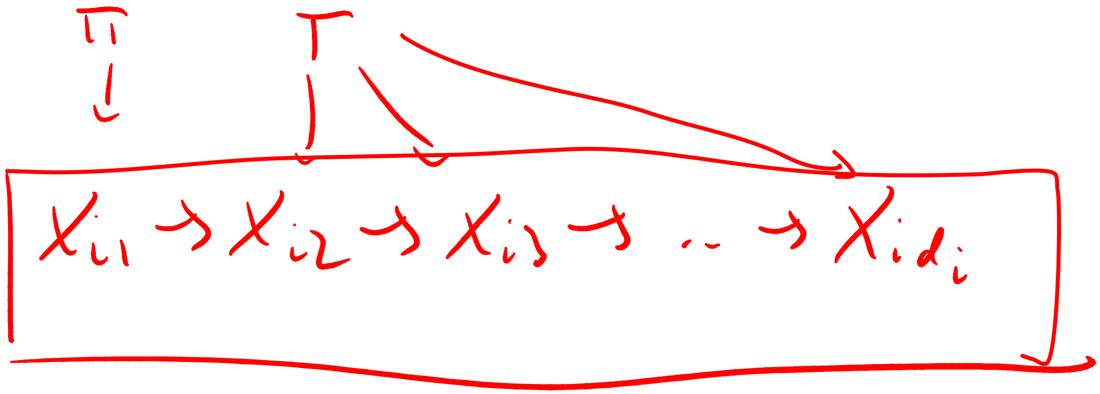
Gaussian classifiers

Tied Sigma, many classes

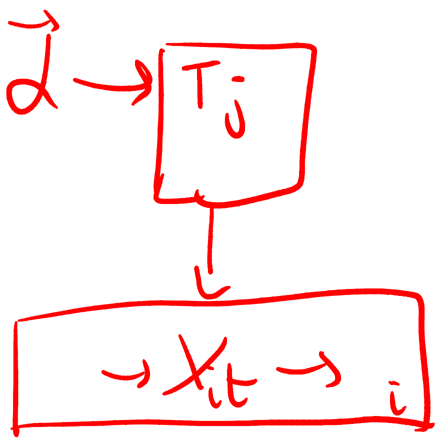
$$
\begin{aligned}
p(Y=c|\mathbf{x}) &= \frac{\pi_c \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_c)^T \Sigma_c^{-1}(\mathbf{x}-\mu_c)\right]}{\sum_{c'} \pi_{c'} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_{c'})^T \Sigma_{c'}^{-1}(\mathbf{x}-\mu_{c'})\right]} \\
&= \frac{\exp\left[\mu_{,}^T \Sigma^{-1} x - \frac{1}{2}\mu_c^T \Sigma^{-1}\mu_c + \log \pi_c\right]}{\sum_{c'} \exp\left[\mu_{c'}^T \Sigma^{-1}\mathbf{x} - \frac{1}{2}\mu_{c'}^T \Sigma^{-1}\mu_{c'} + \log \pi_{c'}\right]} \\
\theta_c &\stackrel{\text{def}}{=} \begin{pmatrix} -\mu_c^T \Sigma^{-1}\mu_c + \log \pi_c \\ \Sigma^{-1}\mu_c \end{pmatrix} = \begin{pmatrix} \gamma_c \\ \beta_c \end{pmatrix} \\
p(Y=c|\mathbf{x}) &= \frac{e^{\theta_c^T \mathbf{x}}}{\sum_{c'} e^{\theta_{c'}^T \mathbf{x}}} = \frac{e^{\beta_c^T \mathbf{x}+\gamma_c}}{\sum_{c'} e^{\beta_{c'}^T \mathbf{x}+\gamma_{c'}}}
\end{aligned}
$$



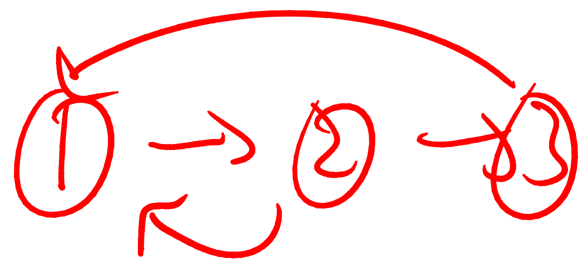
Markov chains



Language Models:
Empirical Bayes on rows of T
leads to backoff smoothing



Theory

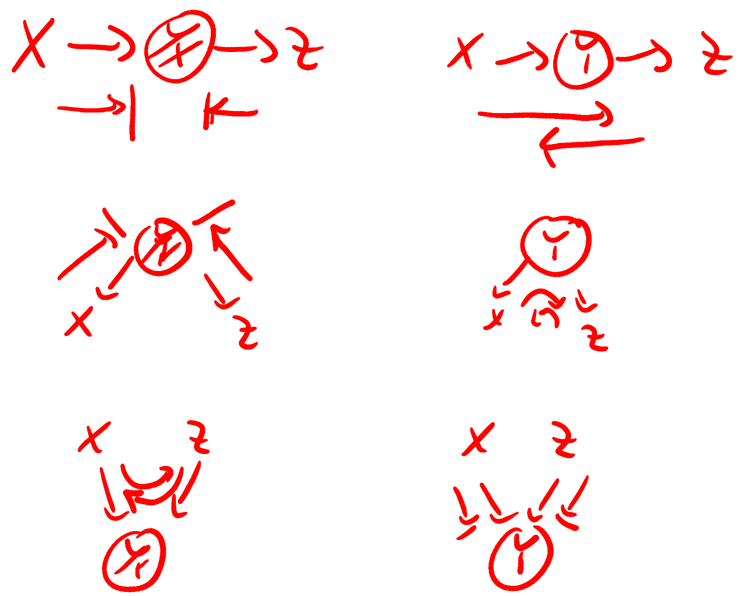


PageRank

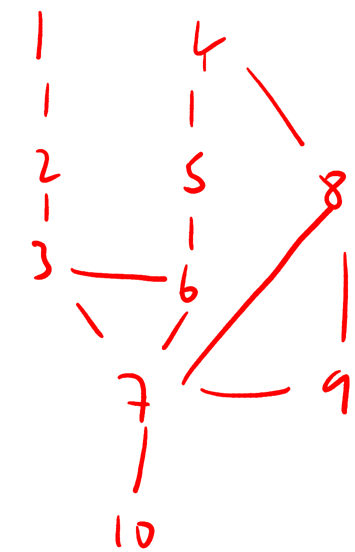
$$T_{ij} = \begin{cases} pG_{ij}/c_j + \delta & \text{if } c_j \neq 0 \\ 1/n & \text{if } c_j = 0 \end{cases}$$

Directed Graphical models

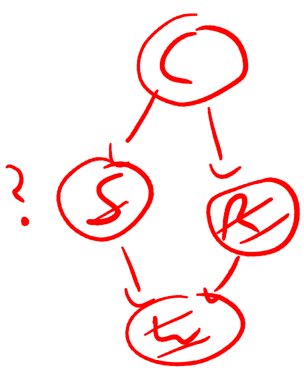
Bayes Ball



Moralization, ancestral graphs



State estimation



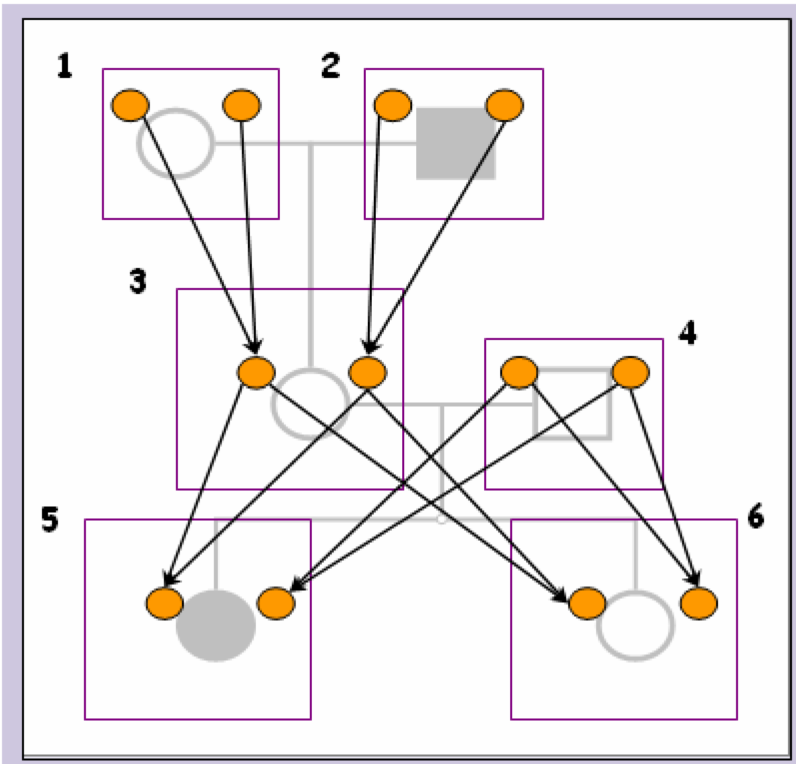
$$p(S=1|W=1,R=1) = \frac{p(S=1,W=1,R=1)}{p(W=1,R=1)} = 0.19$$

Parameter estimation

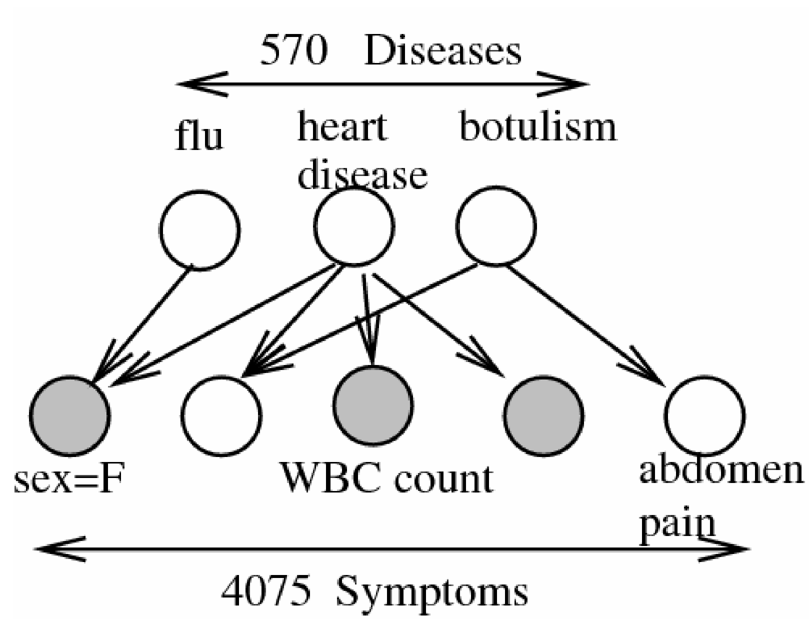
i	C	S	R	W
1	0	0	0	0
2	0	0	1	1
3	1	1	1	1

Expert systems

Pedigree trees

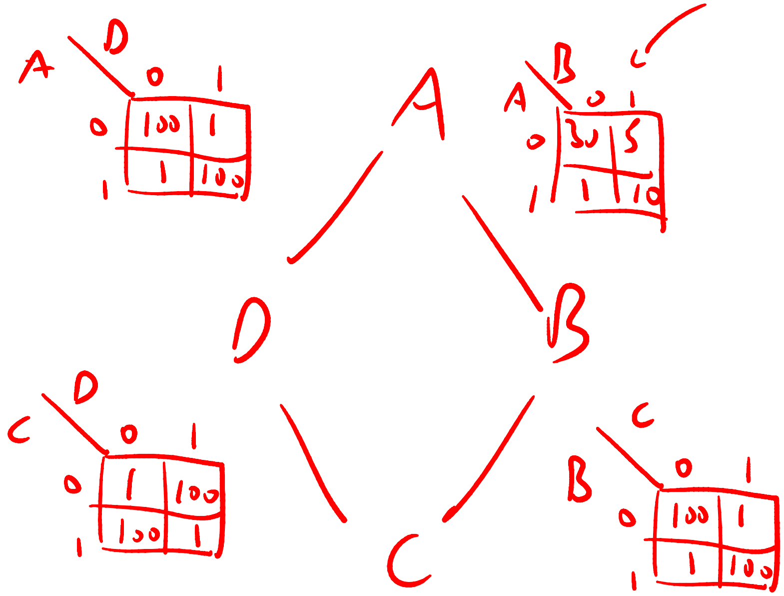


QMR



Noisy-or

Undirected graphical models



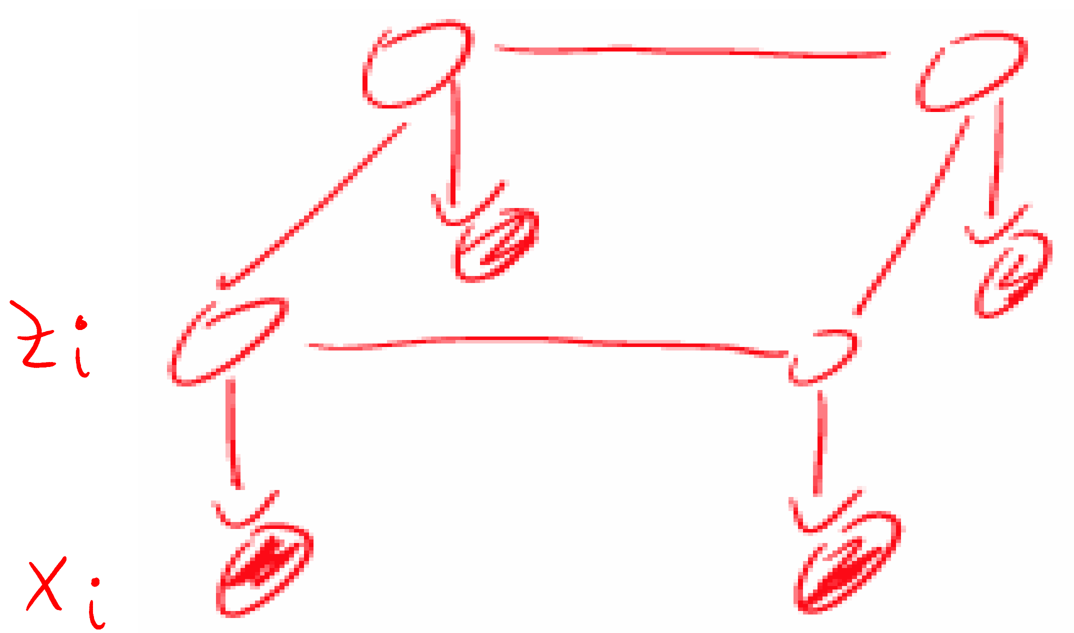
$$p(\mathbf{x}) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c)$$

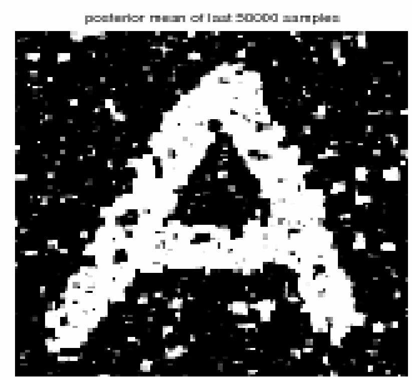
State estimation

Assignment				Unnormalized	Normalized
a^0	b^0	c^0	d^0	300000	0.04
a^0	b^0	c^0	d^1	300000	0.04
a^0	b^0	c^1	d^0	300000	0.04
a^0	b^0	c^1	d^1	30	$4.1 \cdot 10^{-6}$
a^0	b^1	c^0	d^0	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^0	d^1	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^1	d^0	5000000	0.69
a^0	b^1	c^1	d^1	500	$6.9 \cdot 10^{-5}$
a^1	b^0	c^0	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^0	d^1	1000000	0.14
a^1	b^0	c^1	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^1	d^1	100	$1.4 \cdot 10^{-5}$
a^1	b^1	c^0	d^0	10	$1.4 \cdot 10^{-6}$
a^1	b^1	c^0	d^1	100000	0.014
a^1	b^1	c^1	d^0	100000	0.014
a^1	b^1	c^1	d^1	100000	0.014

Ising models

$$p(\mathbf{x},\mathbf{z}) \quad = \quad p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \left[\frac{1}{Z}\prod_{<ij>}\psi_{ij}(z_i,z_j)\right]\left[\prod_i p(x_i|z_i)\right]$$

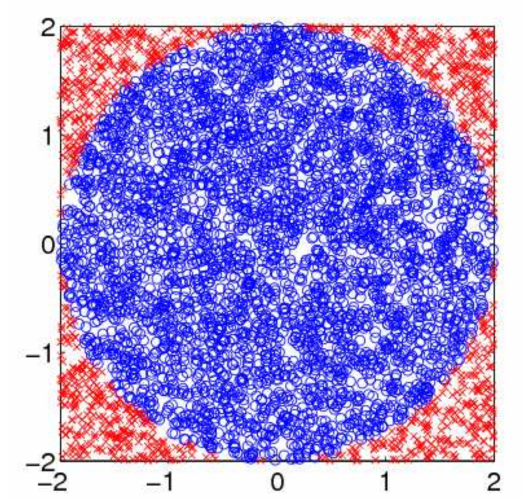




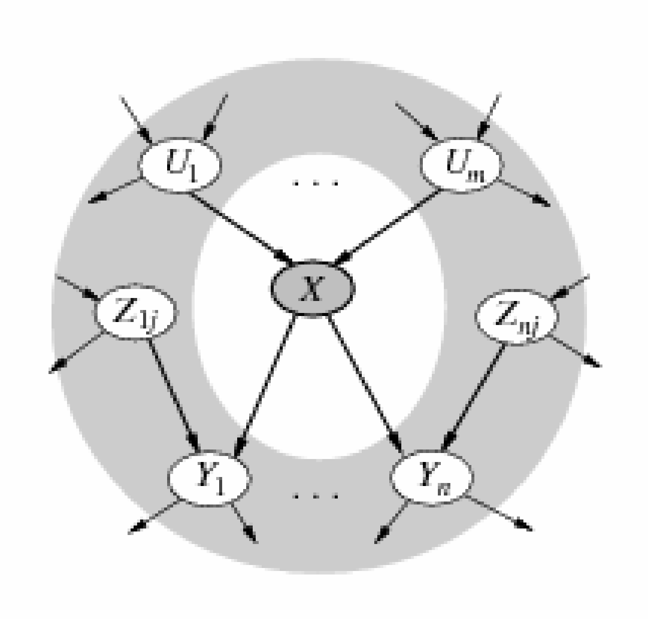
Gibbs sampling

1. $x_1^{s+1} \sim p(x_1 | x_2^s, \ldots, x_D^s)$
2. $x_2^{s+1} \sim p(x_2 | x_1^{s+1}, x_3^s, \ldots, x_D^s)$
3. $x_i^{s+1} \sim p(x_i | x_{1:i-1}^{s+1}, x_{i+1:D}^s)$
4. $x_D^{s+1} \sim p(x_D | x_1^{s+1}, \ldots, x_{D-1}^{s+1})$

Monte Carlo integration



Markov blanket



Full conditional

$$p(X_i | X_{-i}) \propto p(X_i | Pa(X_i)) \prod_{Y_j \in ch(X_i)} p(Y_j | Pa(Y_j))$$