Introduction to Machine Learning

Brown University CSCI 1950-F, Spring 2011

Prof. Erik Sudderth

Lecture 24: Gibbs Samplers, Directed Graphical Models

Many figures courtesy Kevin Murphy's textbook, *Machine Learning: A Probabilistic Perspective*

Announcements

- May 10 at noon: Homework 9 due
- May 10 at 1pm: Review session for final exam
- May 12 at 1pm: Graduate project presentations (rumor that free food will be served)
- May 19 at 2pm: Final exam

Monte Carlo Estimators

$$\mathbb{E}_p[f(x)] = \int_{\mathcal{X}} f(x)p(x)\,dx$$
$$\approx \frac{1}{L}\sum_{\ell=1}^{L} f(x^{(\ell)}) = \mathbb{E}_{\tilde{p}}[f(x)]$$

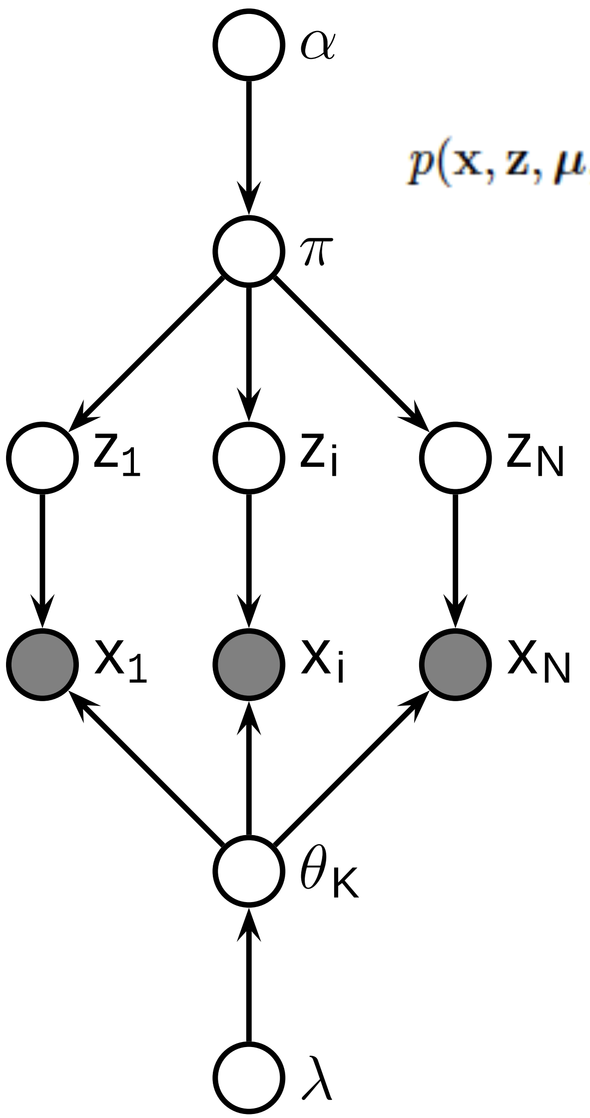
$\{x^{(\ell)}\}_{\ell=1}^{L}$ *independent samples*

$$\tilde{p}(x) = \frac{1}{L}\sum_{\ell=1}^{L} \delta(x, x^{(\ell)})$$

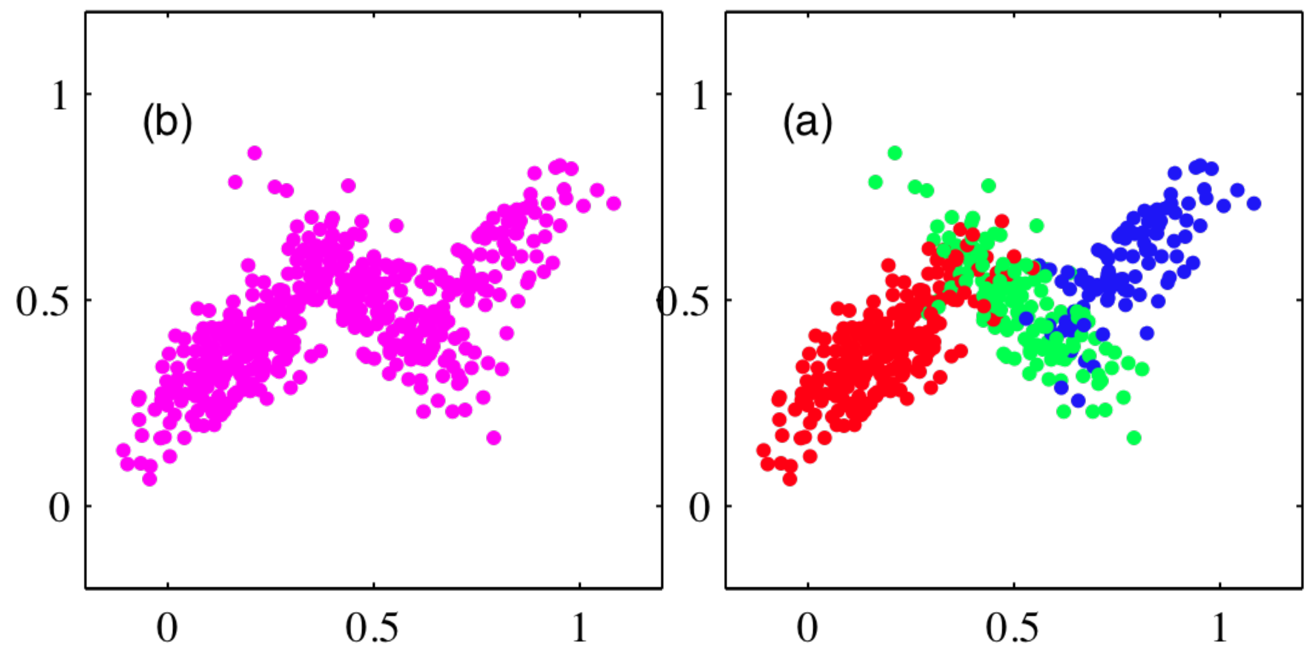
Good properties if *L* sufficiently large:

- Unbiased for any sample size
- Variance inversely proportional to sample size (and independent of dimension of space)
- Weak law of large numbers
- Strong law of large numbers
- **Problem:** Drawing samples from complex distributions…

Gibbs Sampler for Gaussian Mixtures



$$
\begin{aligned}
p(\mathbf{x}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) &= p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\sigma}) p(\mathbf{z}|\boldsymbol{\pi}) p(\boldsymbol{\pi}) \prod_{k=1}^{K} p(\boldsymbol{\mu}_k) p(\boldsymbol{\Sigma}_k) \\
&= \left(\prod_{i=1}^{N} \prod_{k=1}^{K} (\pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))^{\mathbb{I}(z_i=k)} \right) \\
&\quad \times \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \times \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_0, \mathbf{V}_0) \mathrm{IW}(\boldsymbol{\Sigma}_k|\mathbf{S}_0, \nu_0)
\end{aligned}
$$



Mixture Sampler Pseudocode

Given mixture weights $\pi^{(t-1)}$ and cluster parameters $\{\theta_k^{(t-1)}\}_{k=1}^K$ from the previous iteration, sample a new set of mixture parameters as follows:

1. Independently assign each of the N data points x_i to one of the K clusters by sampling the indicator variables $z = \{z_i\}_{i=1}^N$ from the following multinomial distributions:

$$z_i^{(t)} \sim \frac{1}{Z_i} \sum_{k=1}^{K} \pi_k^{(t-1)} f(x_i \mid \theta_k^{(t-1)})\, \delta(z_i, k) \qquad Z_i = \sum_{k=1}^{K} \pi_k^{(t-1)} f(x_i \mid \theta_k^{(t-1)})$$

2. Sample new mixture weights according to the following Dirichlet distribution:

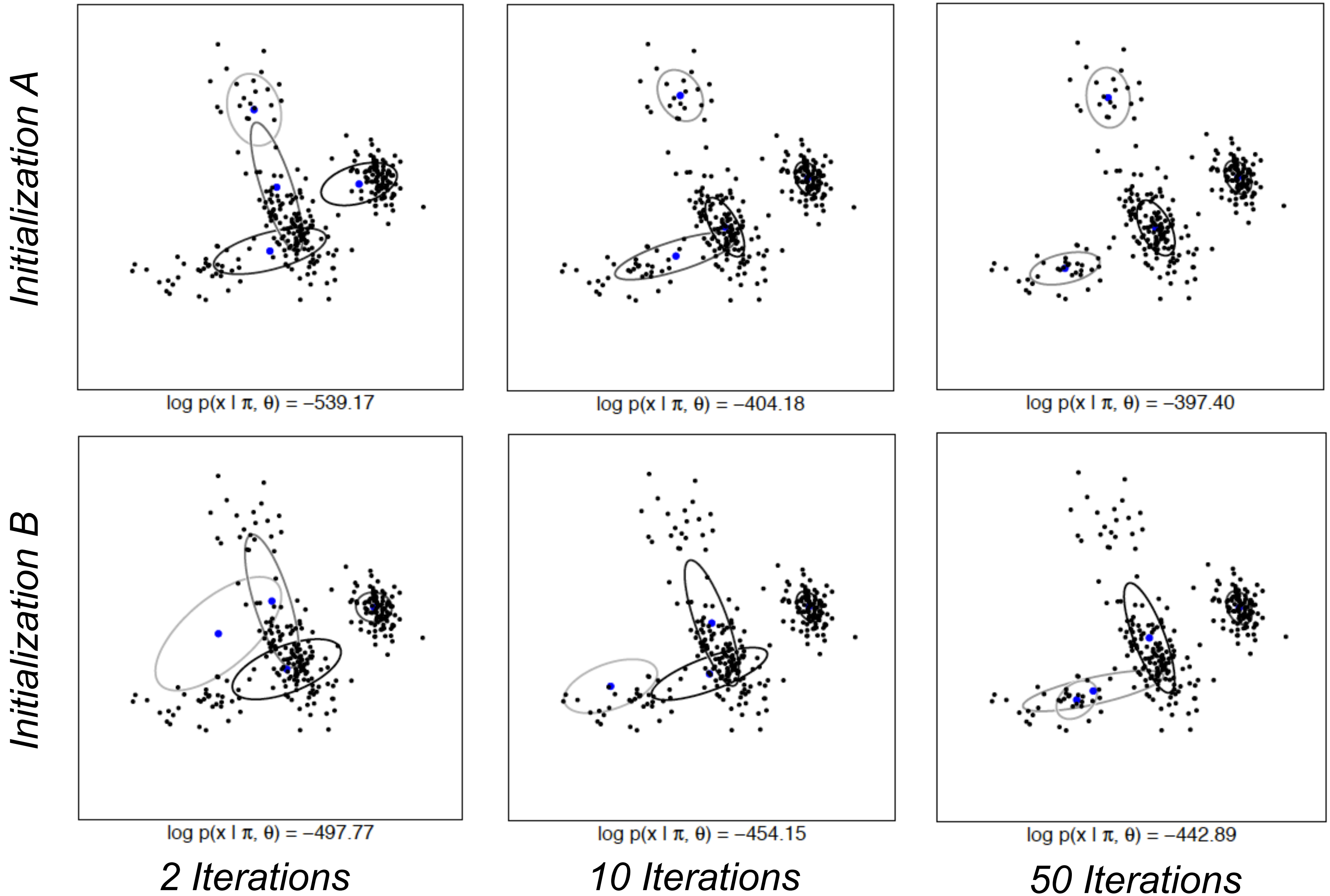
$$\pi^{(t)} \sim \mathrm{Dir}(N_1 + \alpha/K, \ldots, N_K + \alpha/K) \qquad N_k = \sum_{i=1}^{N} \delta(z_i^{(t)}, k)$$

3. For each of the K clusters, independently sample new parameters from the conditional distribution implied by those observations currently assigned to that cluster:

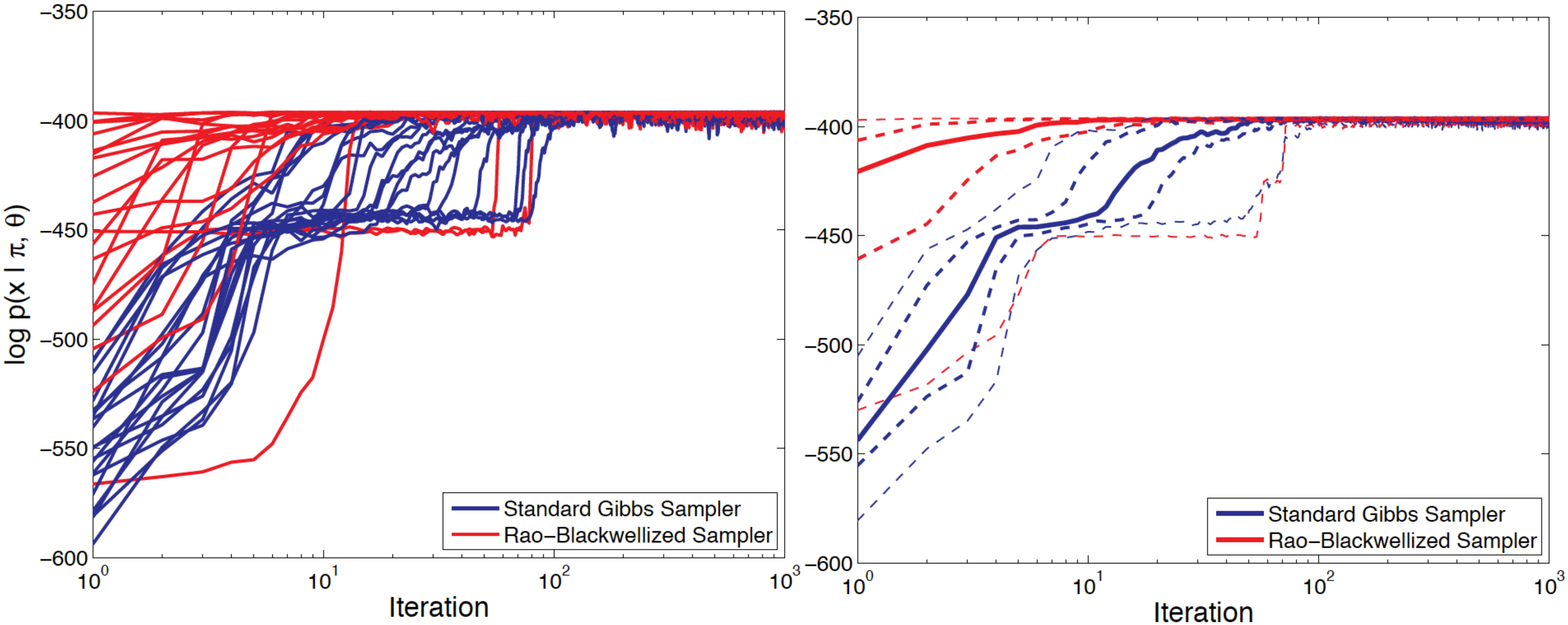
$$\theta_k^{(t)} \sim p(\theta_k \mid \{x_i \mid z_i^{(t)} = k\}, \lambda)$$

When λ defines a conjugate prior, this posterior distribution is given by Prop. 2.1.4.

Snapshots of Mixture Gibbs Sampler



Gibbs: Representation and Mixing

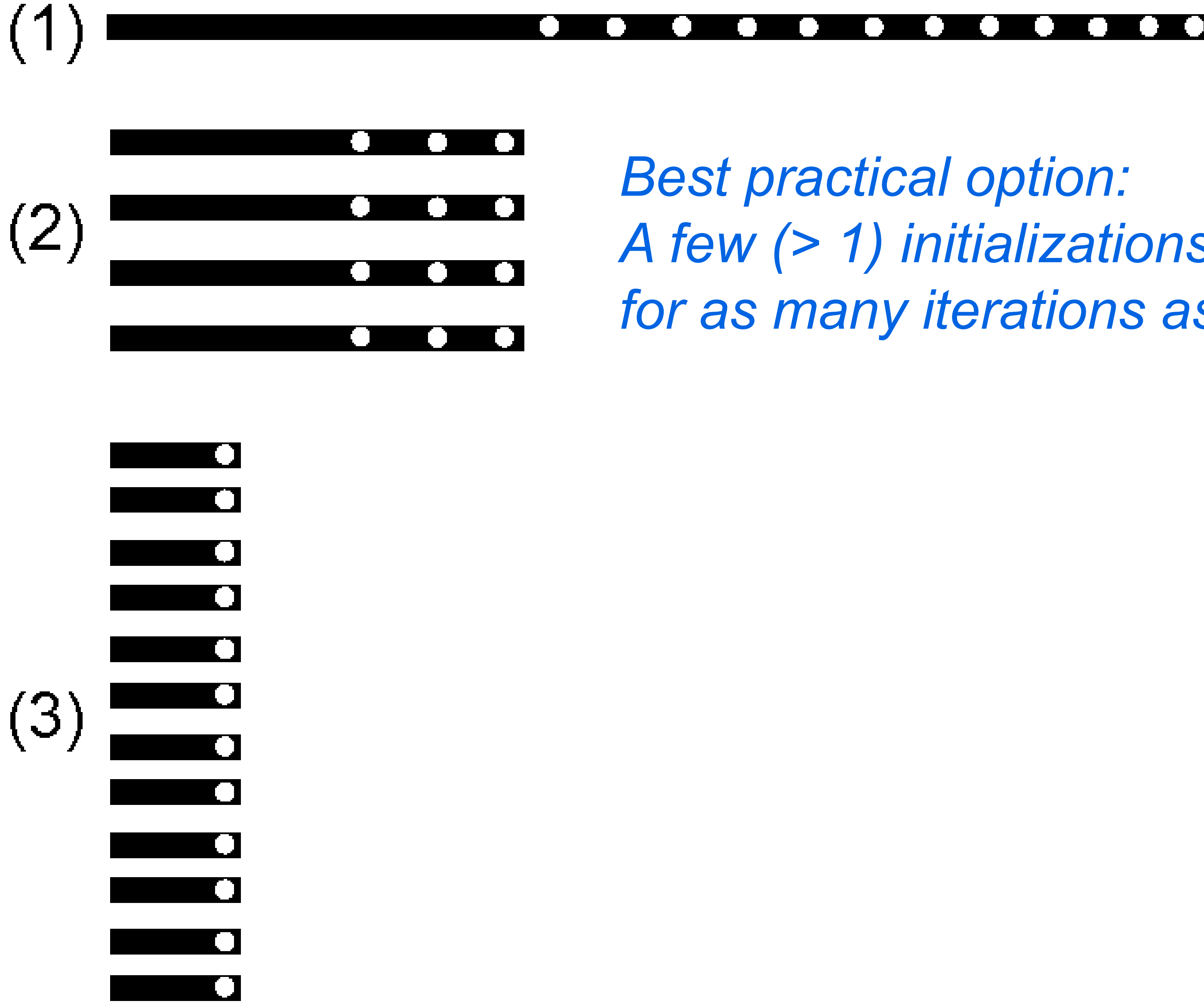


Multiple Initializations *Quantiles of 100 Chains*

Standard Gibbs: Alternatively sample assignments, parameters

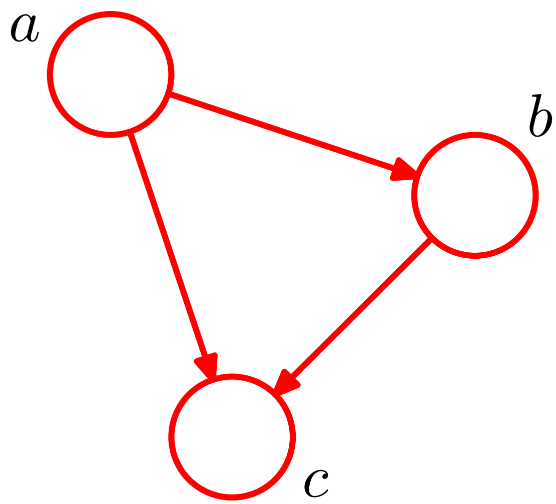
Collapsed Gibbs: Marginalize parameters, sample assignments

MCMC & Computational Resources

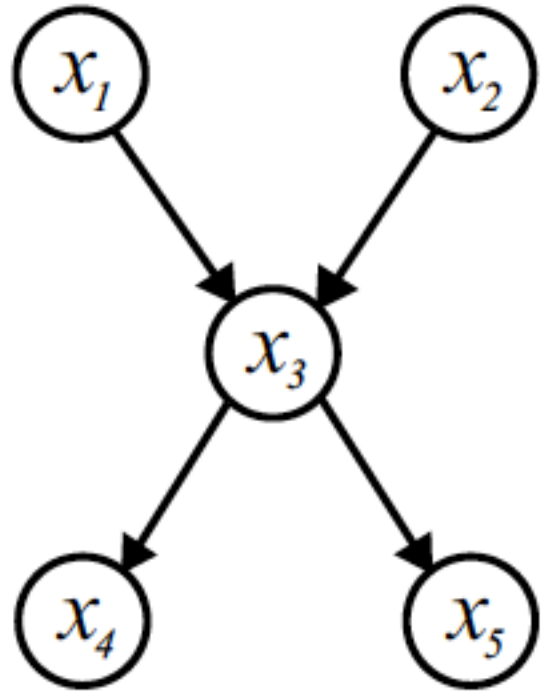


Best practical option:
A few (> 1) initializations
for as many iterations as possible

Directed Graphical Models



$$p(\mathbf{x}_{1:D}|G) = \prod_{s=1}^{D} p(x_s|\mathbf{x}_{\mathrm{pa}(s)})$$

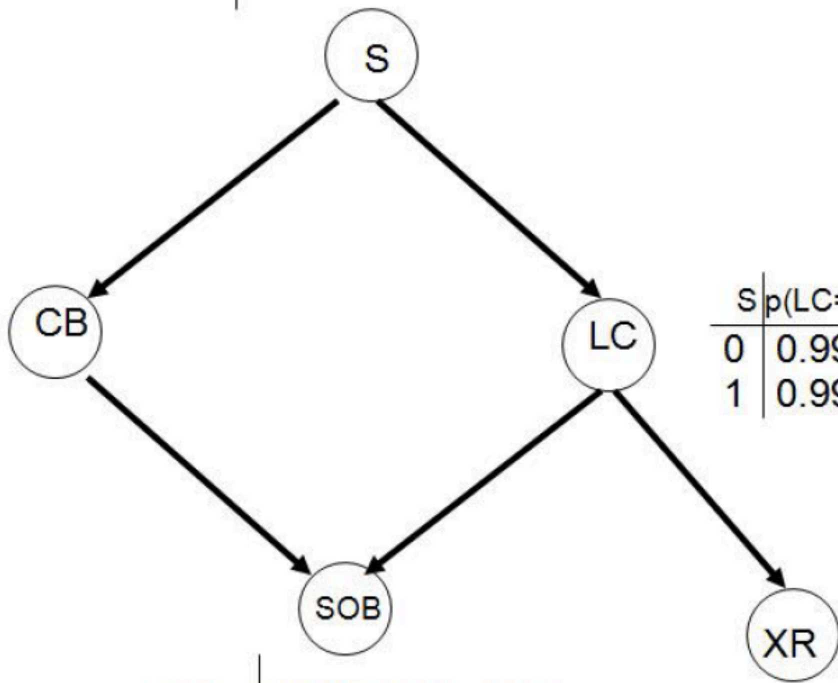


$$p(x) = p(x_1)\, p(x_2)\, p(x_3 \mid x_1, x_2)\, p(x_4 \mid x_3)\, p(x_5 \mid x_3)$$

- Allows easy visualization of model similarities and differences
- Suggests ways of generalizing and combining models
- General-purpose learning and inference algorithms

Smoking, Bronchitis, & Lung Cancer

p(S=1)	p(S=0)
0.8	0.2



S	p(CB=1|.)	p(CB=0|.)
0	0.95	0.05
1	0.75	0.25

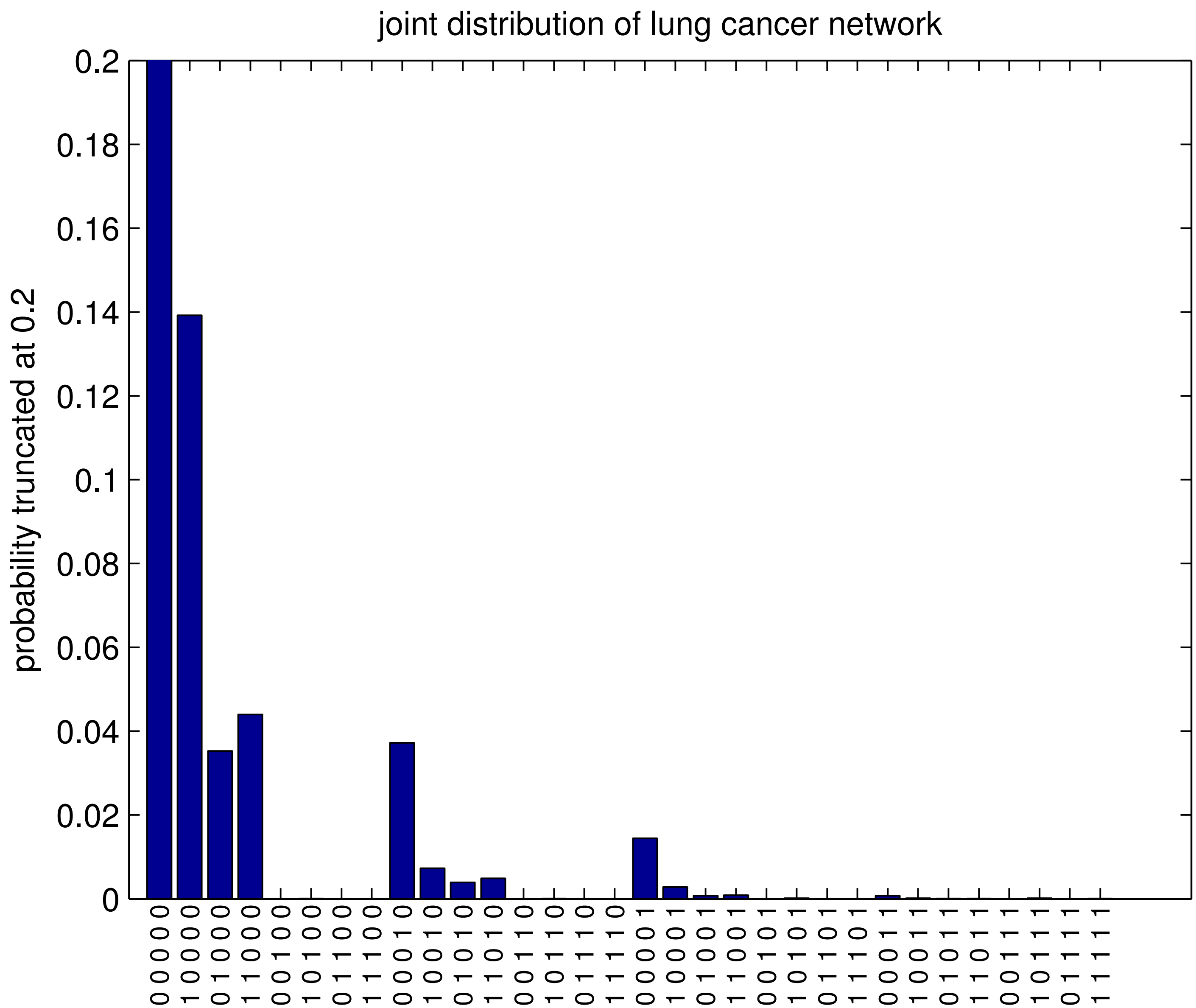
S	p(LC=1|.)	p(LC=0|.)
0	0.99995	0.00005
1	0.997	0.003

CB	LC	p(SOB=1|.)	p(SOB|.)
0	0	0.95	0.05
1	0	0.5	0.5
0	1	0.5	0.5
1	1	0.25	0.75

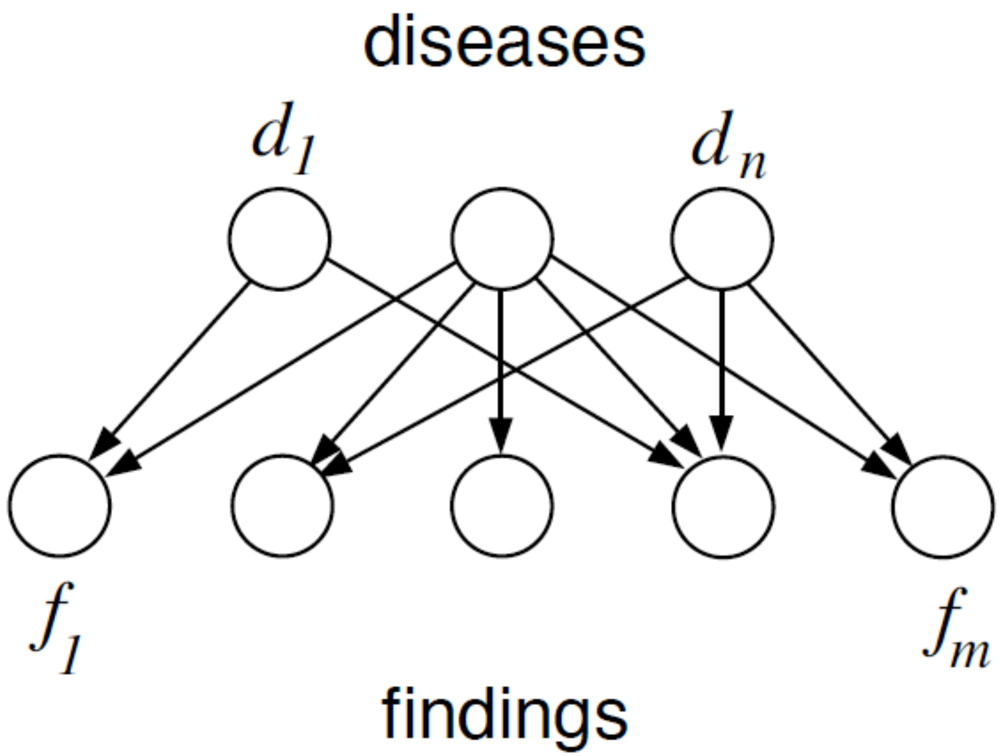
LC	p(XR=1|.)	p(XR=0|.)
0	0.98	0.02
1	0.40	0.60

$$
\begin{aligned}
p(X_S, X_{CB}, X_{LC}, X_{SOB}, X_{XR}) &= p(X_S)p(X_{CB}|X_S)p(X_{LC}|\cancel{X_{CB}}, X_S)p(X_{SOB}|X_{LC}, X_{CB}, \cancel{X_s}) \\
&\quad \times p(X_{XR}|\cancel{X_{SOB}}, X_{LC}, \cancel{X_{CB}}, \cancel{X_S}) \\
&= p(X_S)p(X_{CB}|X_S)p(X_{LC}|X_S)p(X_{SOB}|X_{CB}, X_{LC})p(X_{XR}|X_{LC})
\end{aligned}
$$

Corresponding Joint Distribution

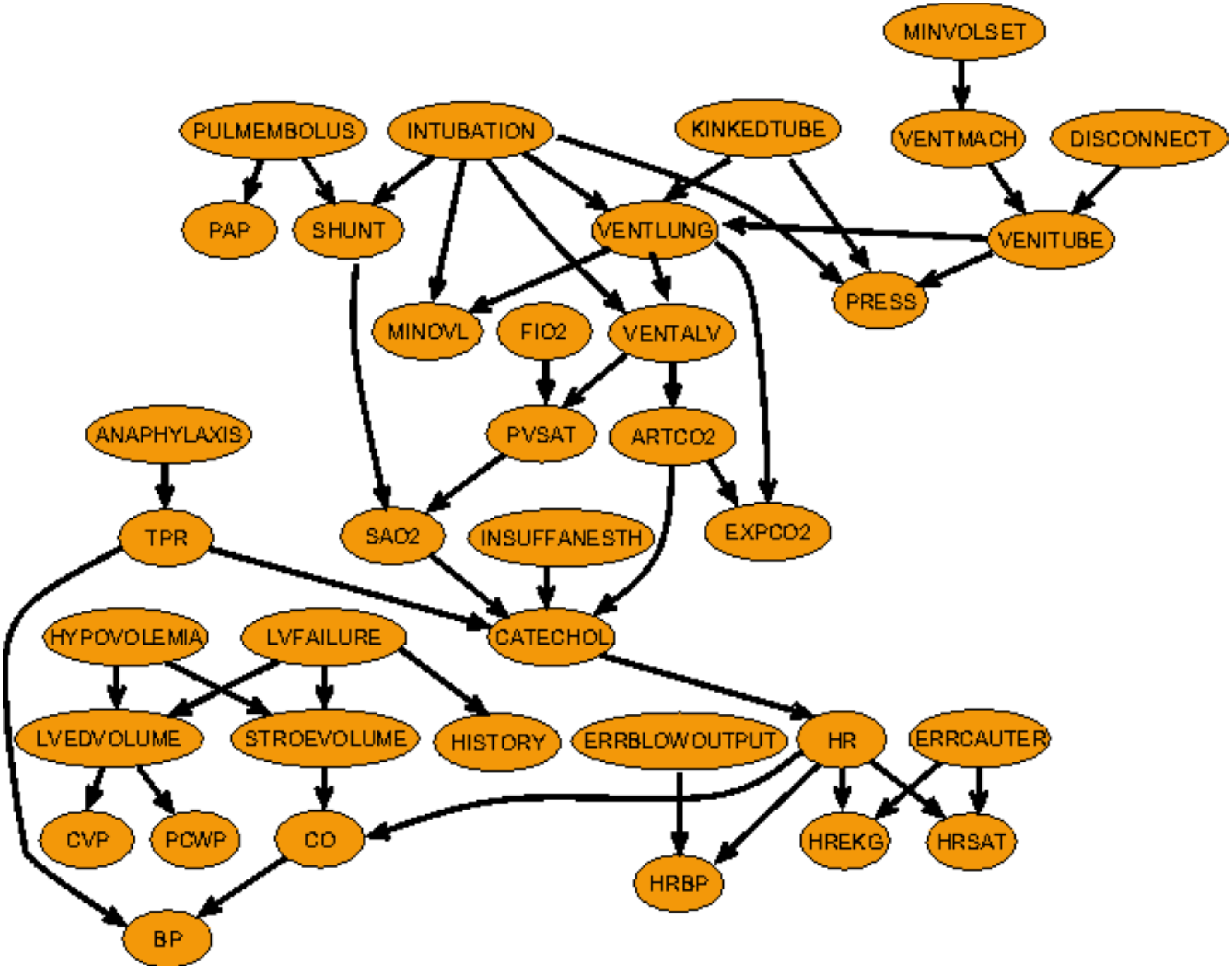


Medical Diagnosis



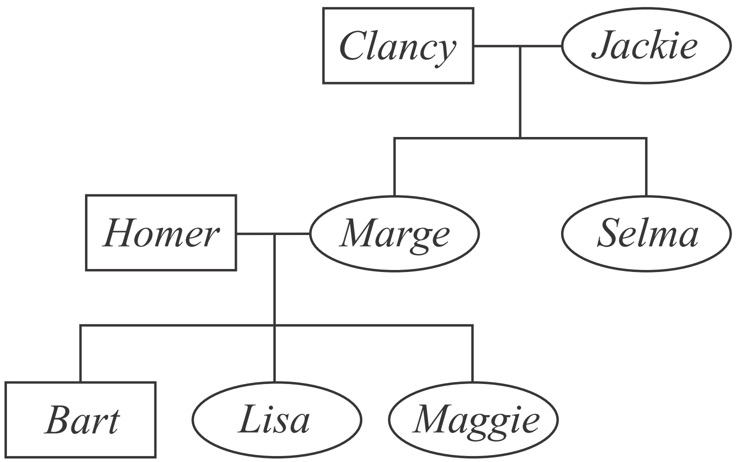
Parameterization: Noisy-OR, logistic regression, generalized linear models…

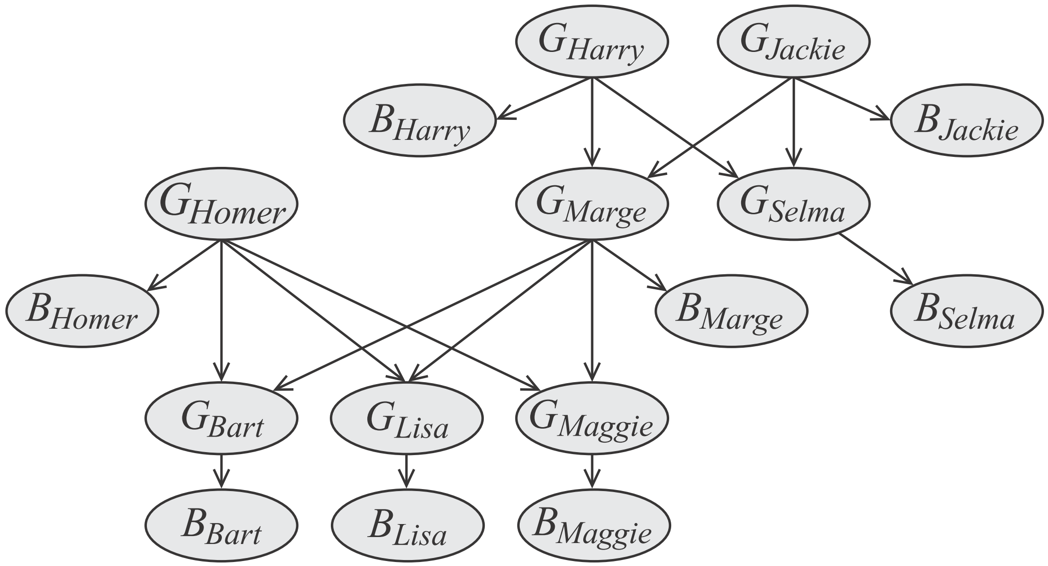
Expert Systems



Alarm Network for ICU Monitoring

Bloodtypes and Genotypes





Learning for Graphical Models

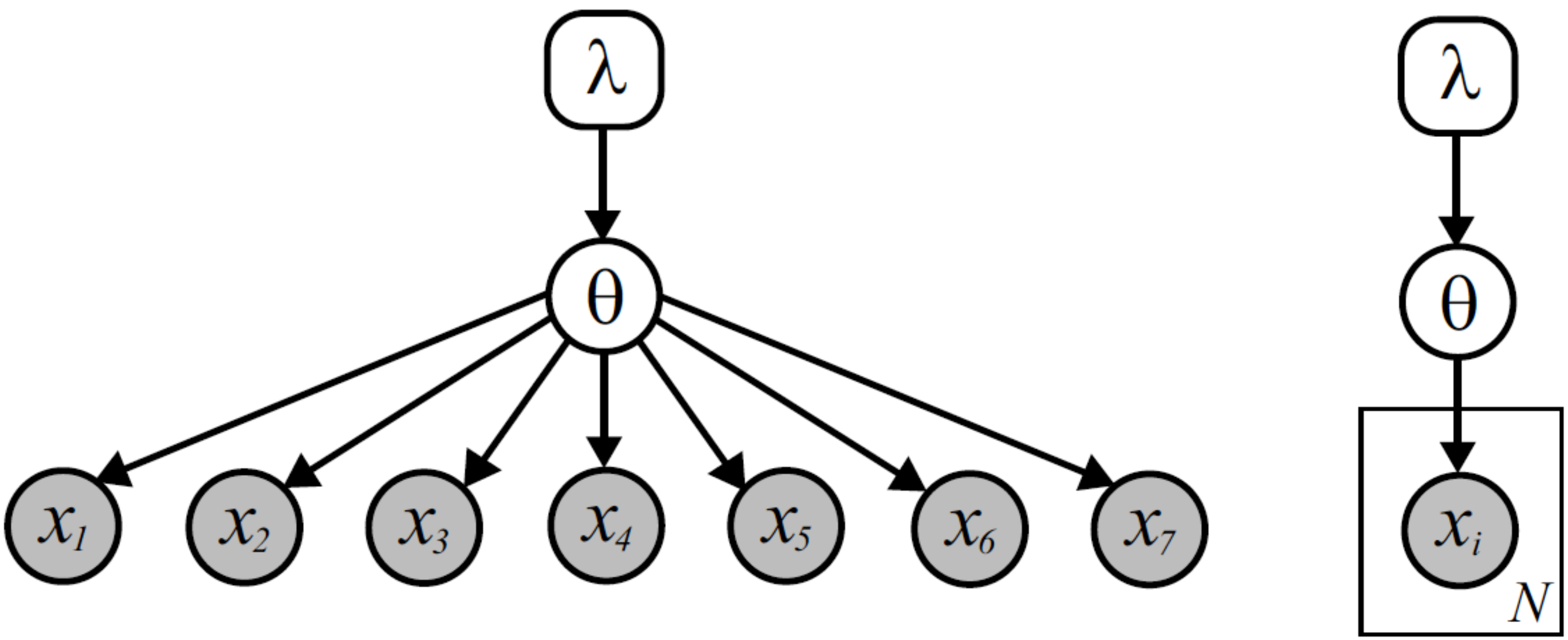
$$p(\mathbf{x}_{1:D}|G) = \prod_{s=1}^{D} p(x_s|\mathbf{x}_{\mathrm{pa}(s)})$$

- Given a sequence of complete observations of all variables in graph
  - Log-likelihood decomposes: Predict each child node given its parent nodes
  - Discrete child node: classification problem
  - Continuous child node: regression problem
- What if some variables are unobserved?
  - Expectation-Maximization (EM) algorithm
  - MCMC algorithms
  - Alternate between "filling in" and re-estimating parameters

Machine Learning Problems

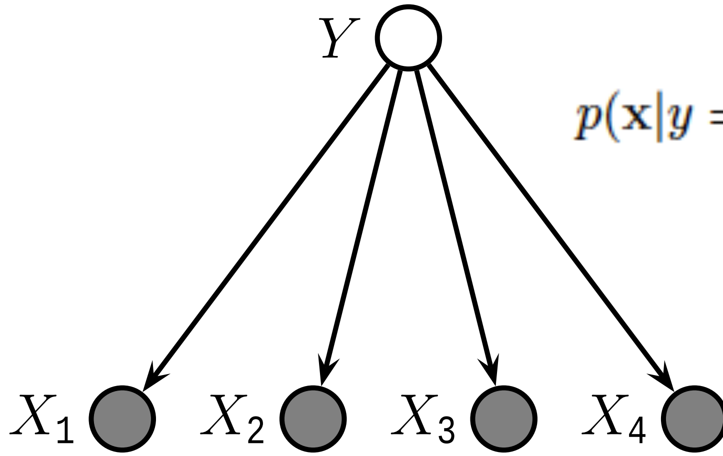
	Supervised Learning	Unsupervised Learning
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

Learning with a Prior and Independent Observations



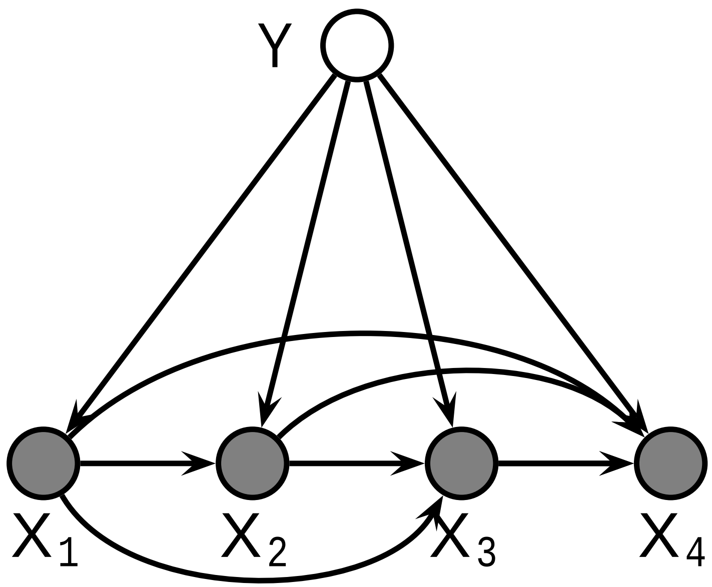
Sometimes-used convention: Shaded nodes are observed

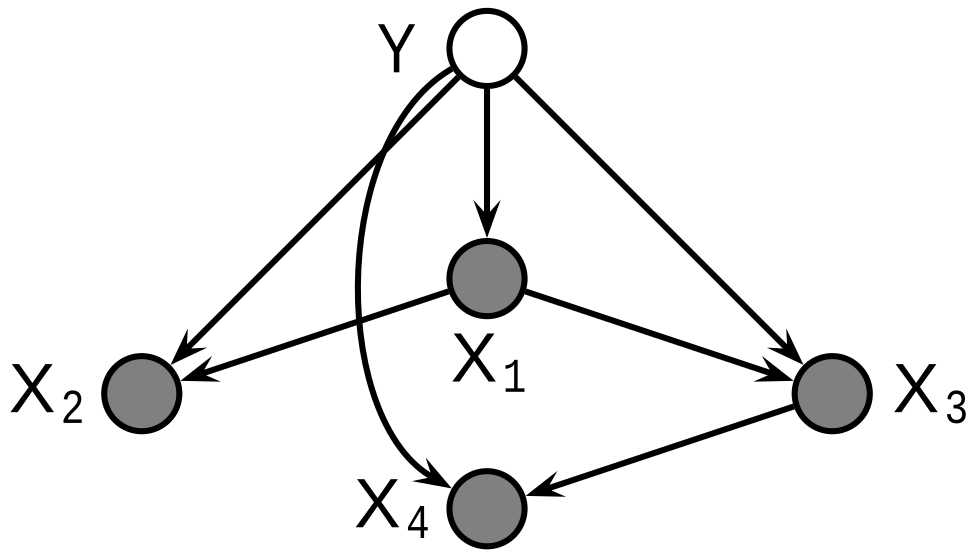
Naïve Bayes Classifier



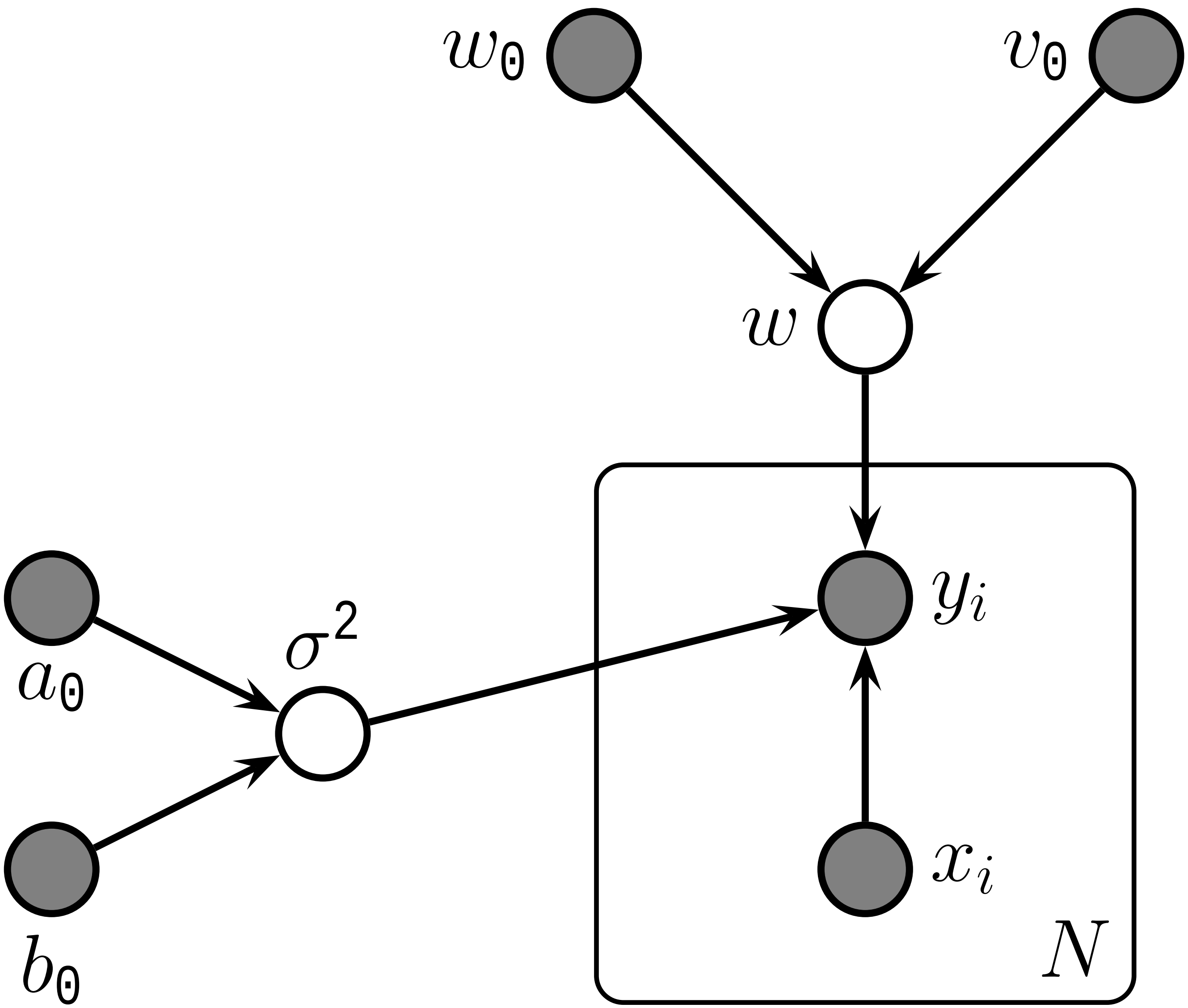
$$p(\mathbf{x}|y = c, \boldsymbol{\theta}) = \prod_{j=1}^{D} p(x_j|y = c, \boldsymbol{\theta}_j)$$

Allowing more dependence among features:

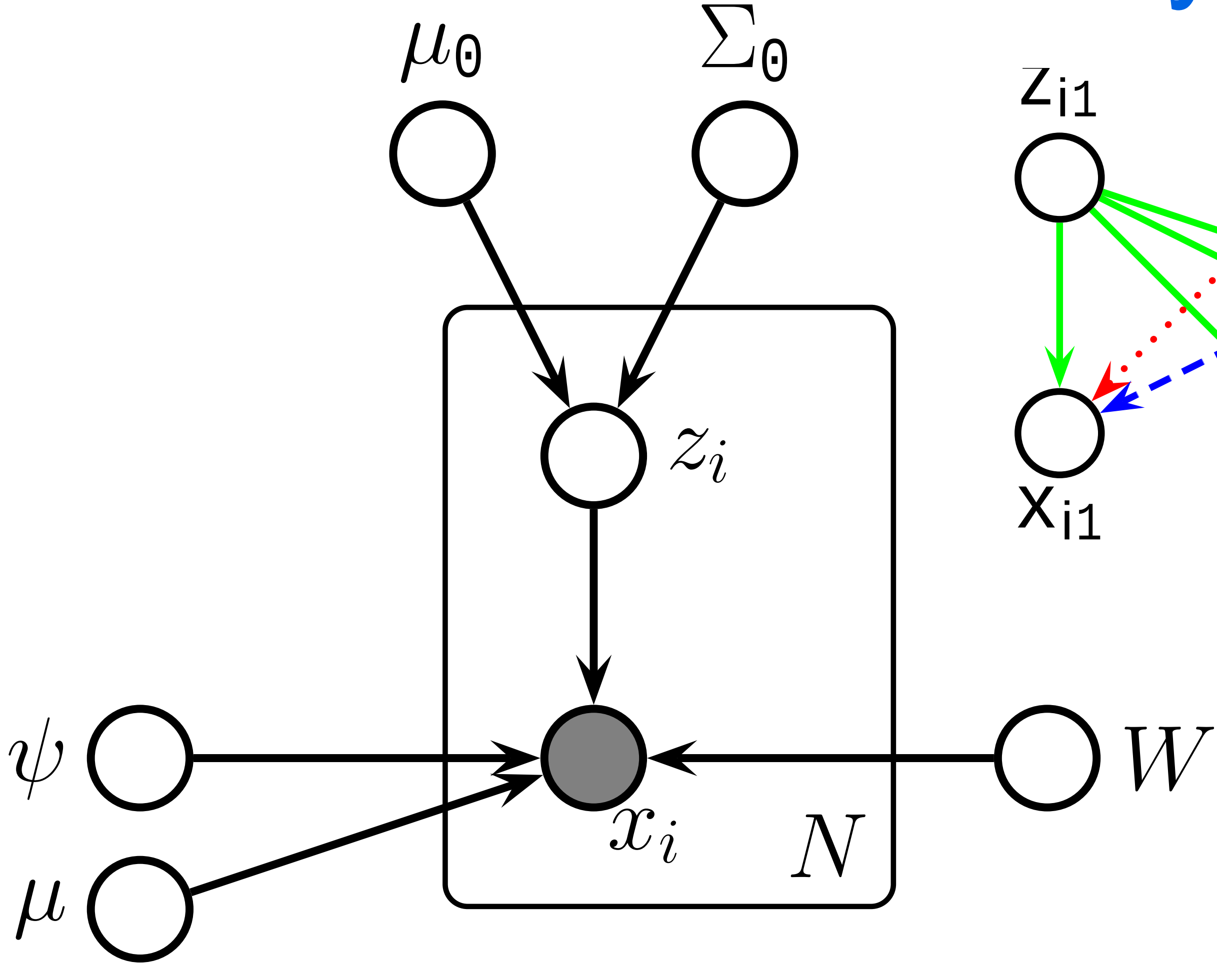


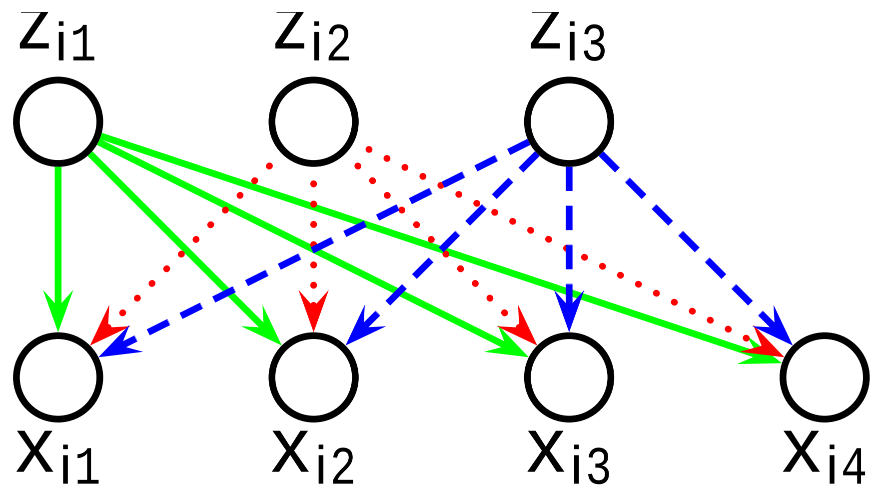


Bayesian Linear Regression

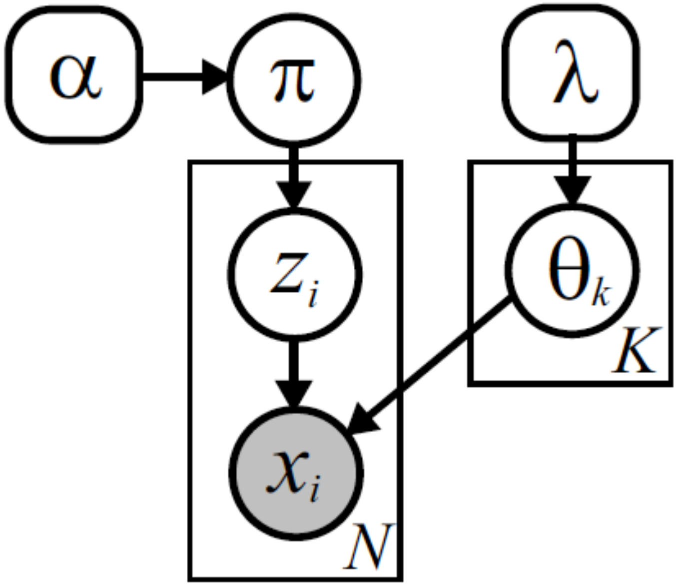


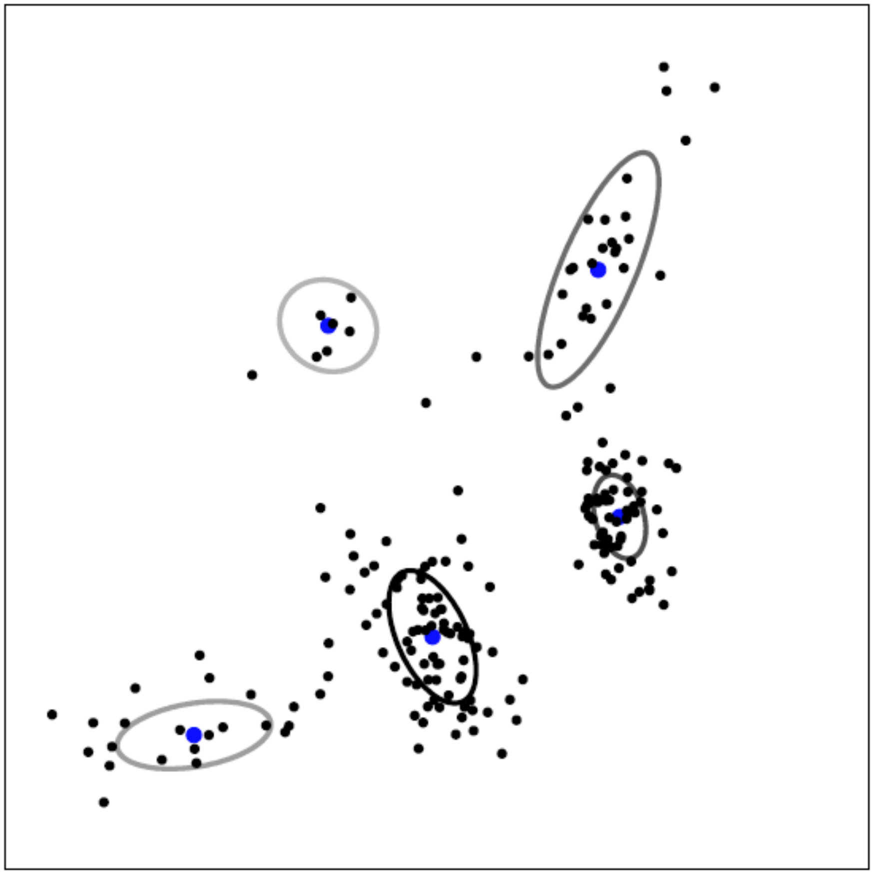
Factor Analysis



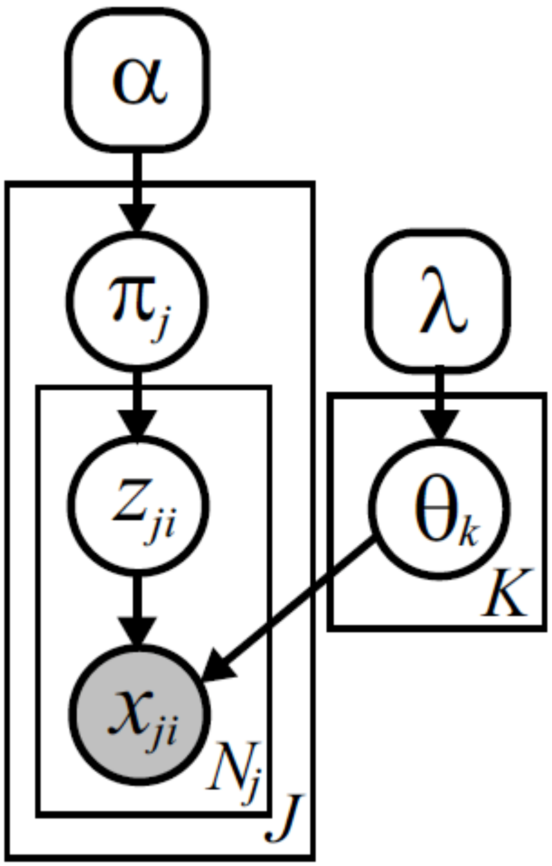


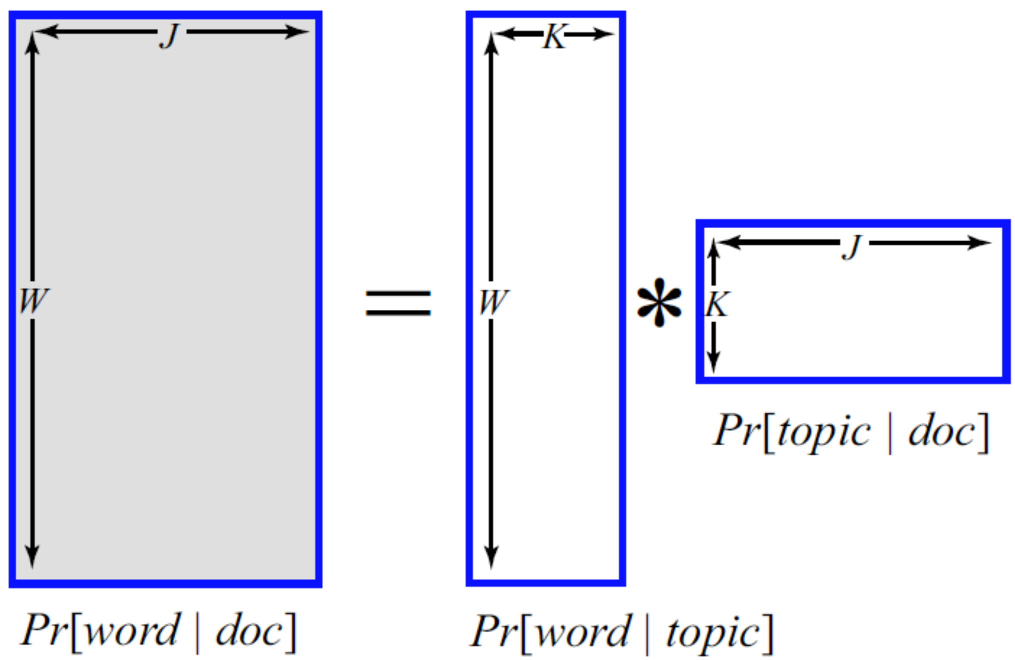
Mixture Models



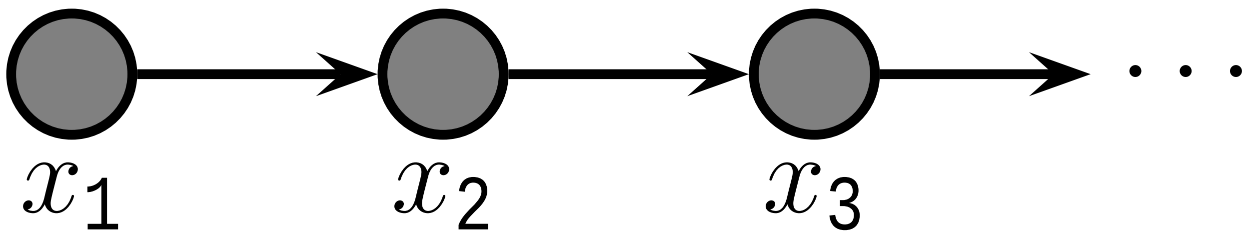


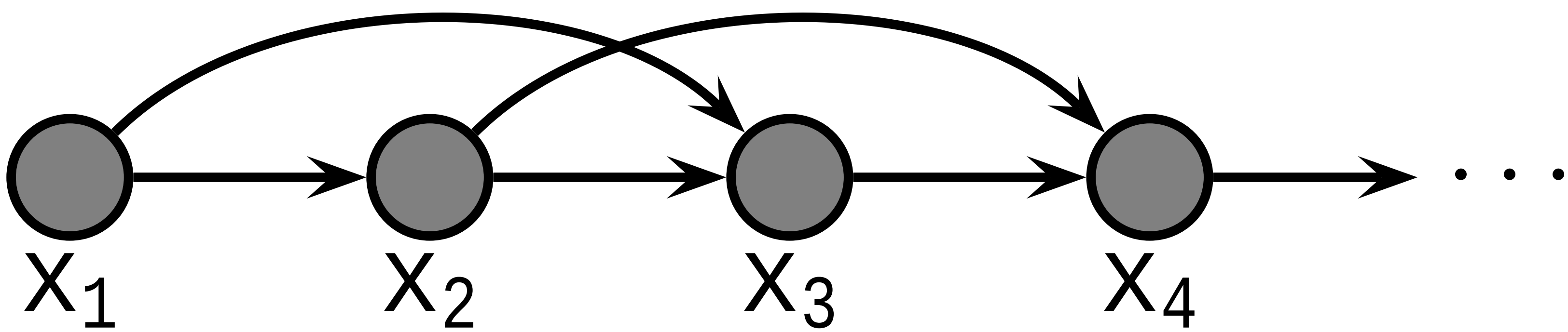
Admixture or Topic Models (Latent Dirichlet Allocation)



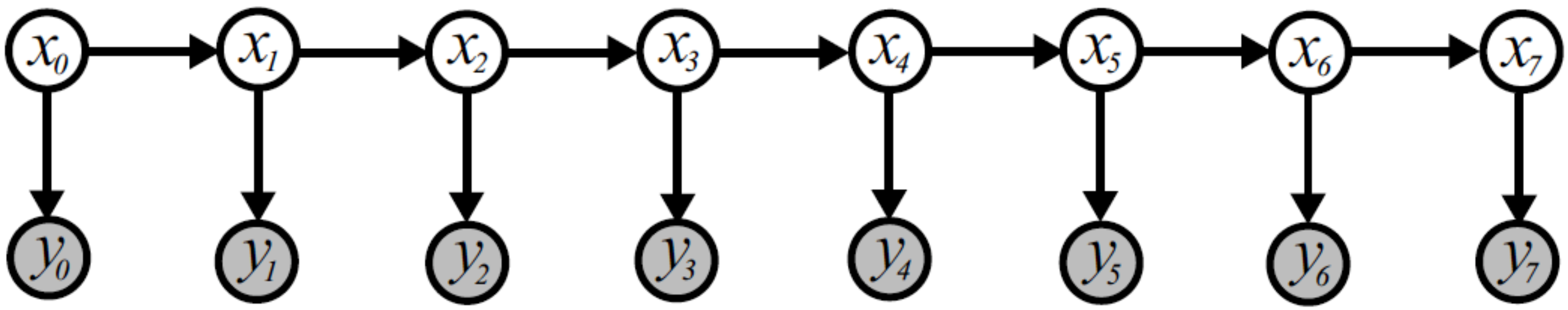


Markov Chains of Varying Orders



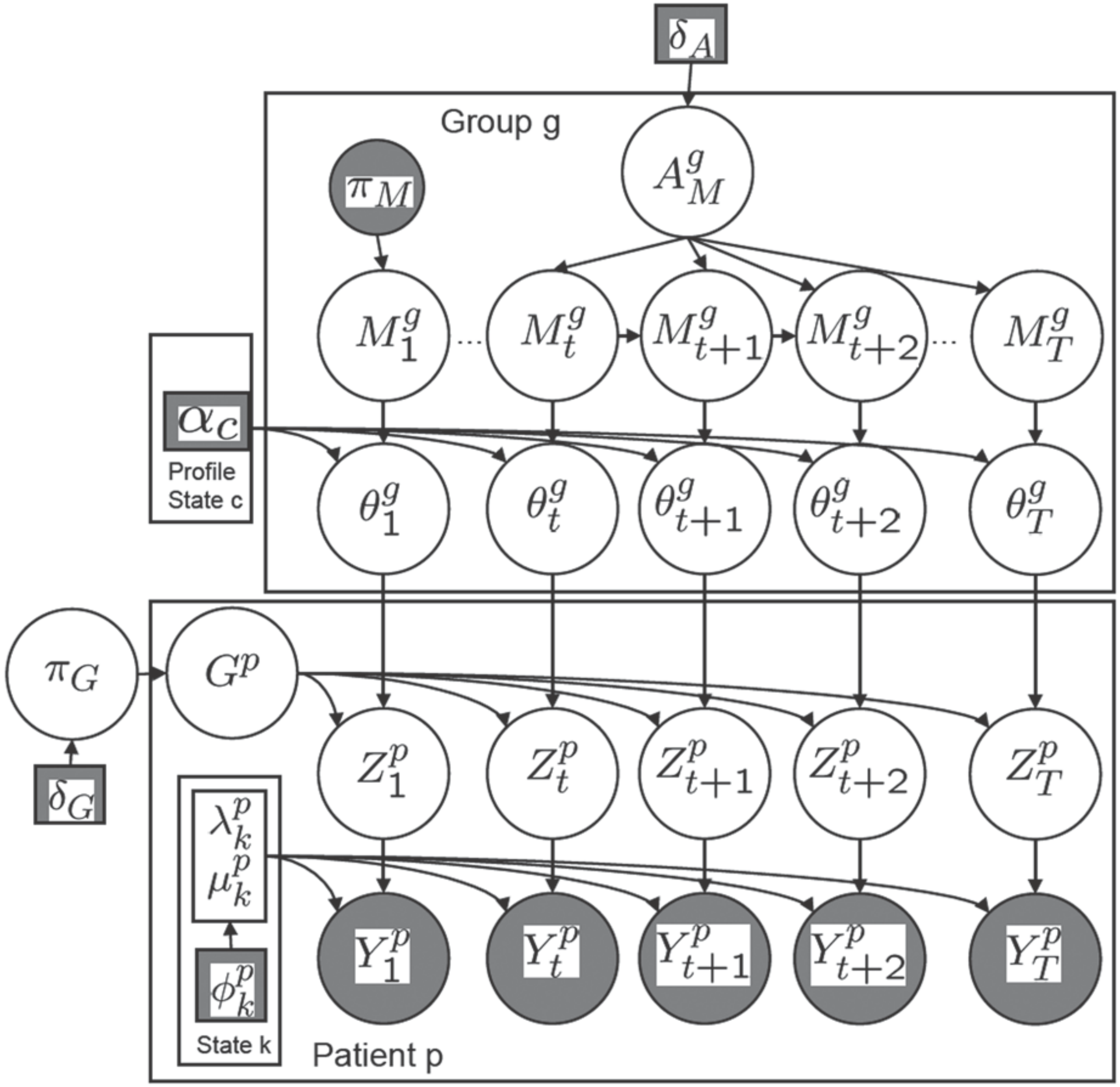


Hidden Markov Model
(& linear dynamical system)

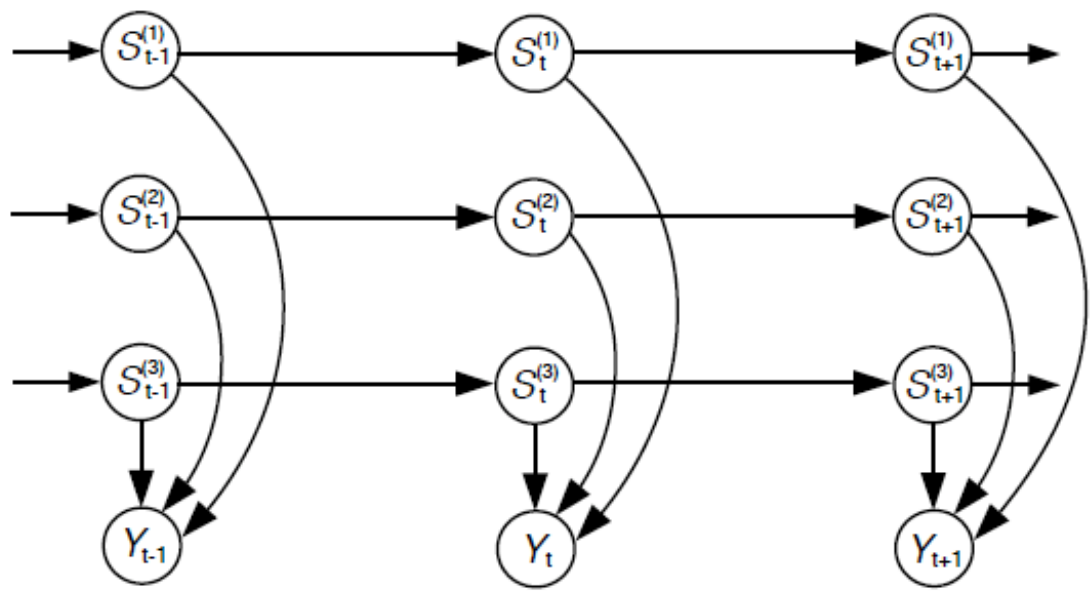


Advanced Models
(presented for your amusement)

Mixture of HMMs

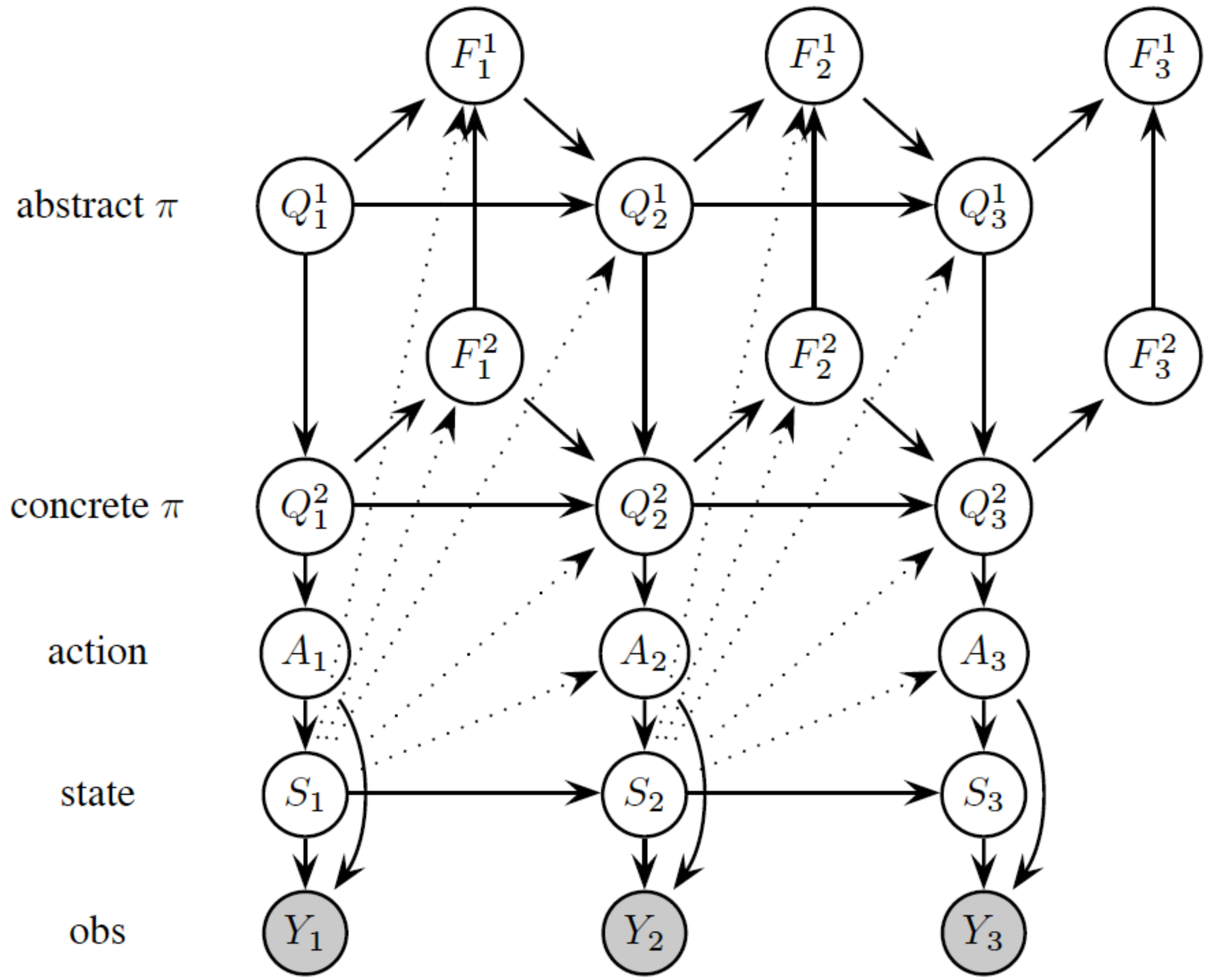


Factorial HMM



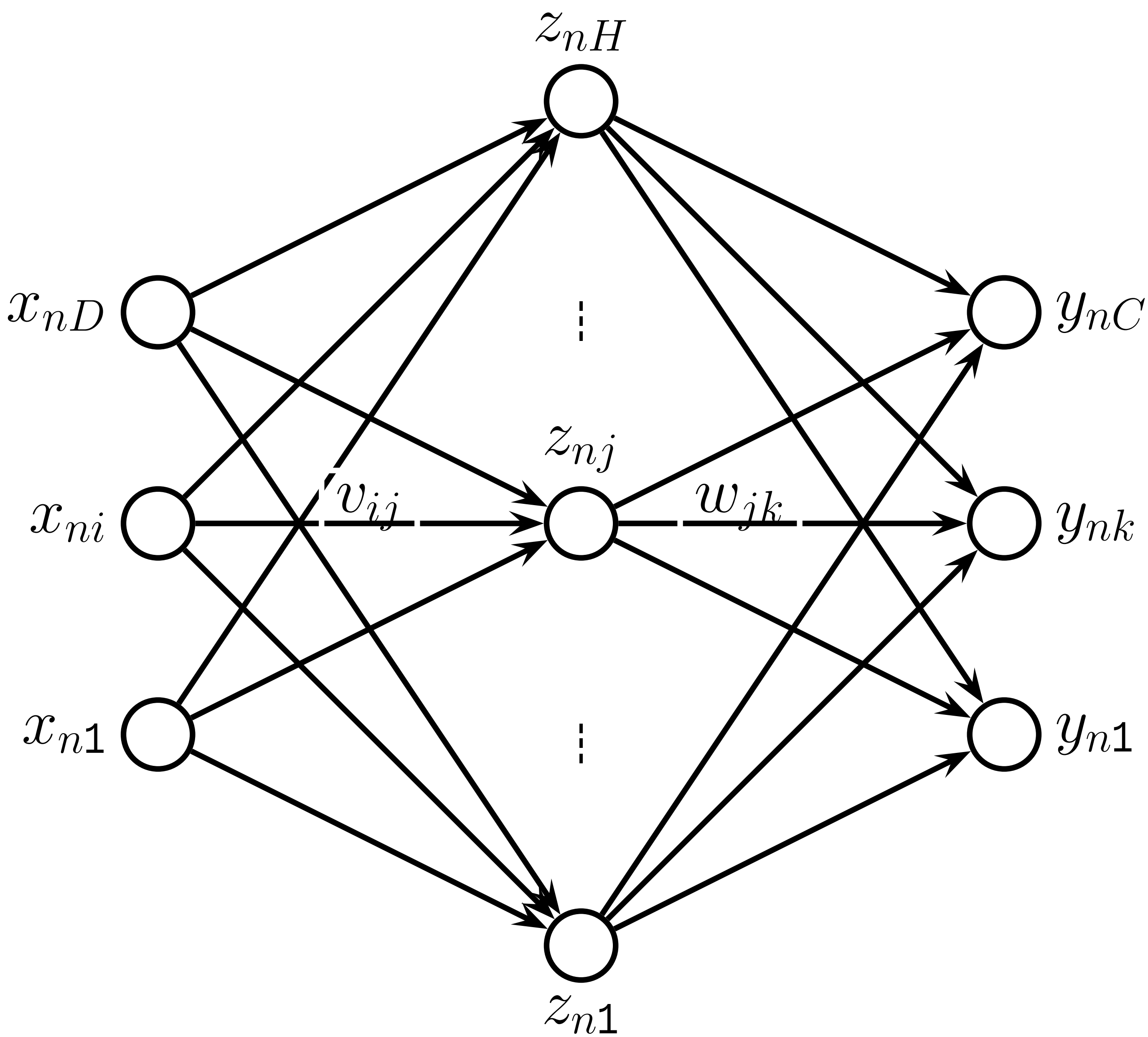
Ghahramani & Jordan, 1997

Dynamic Bayesian Networks

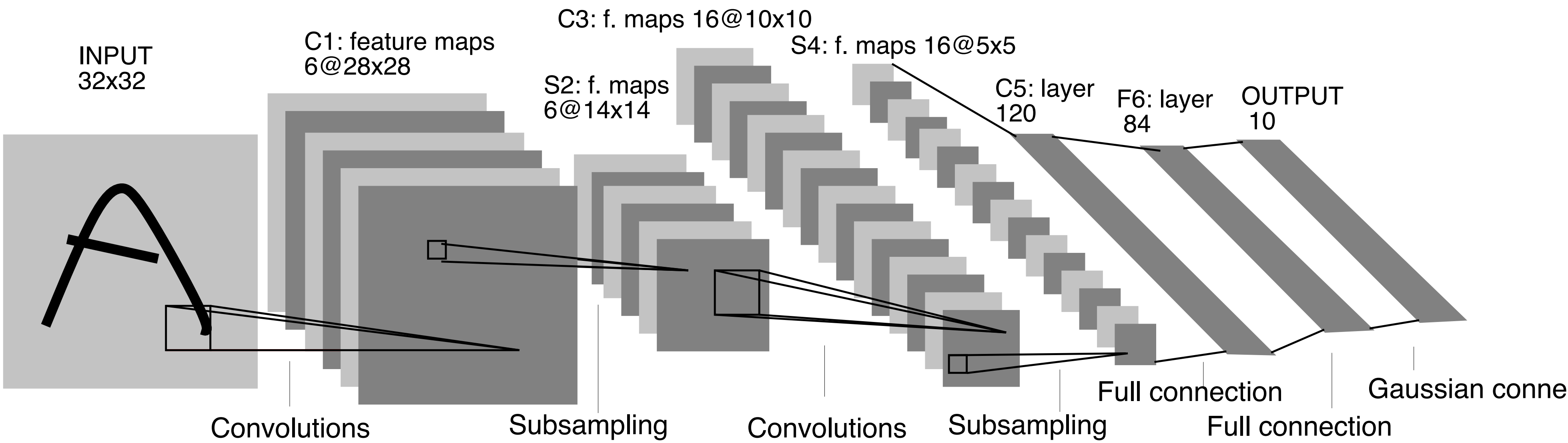


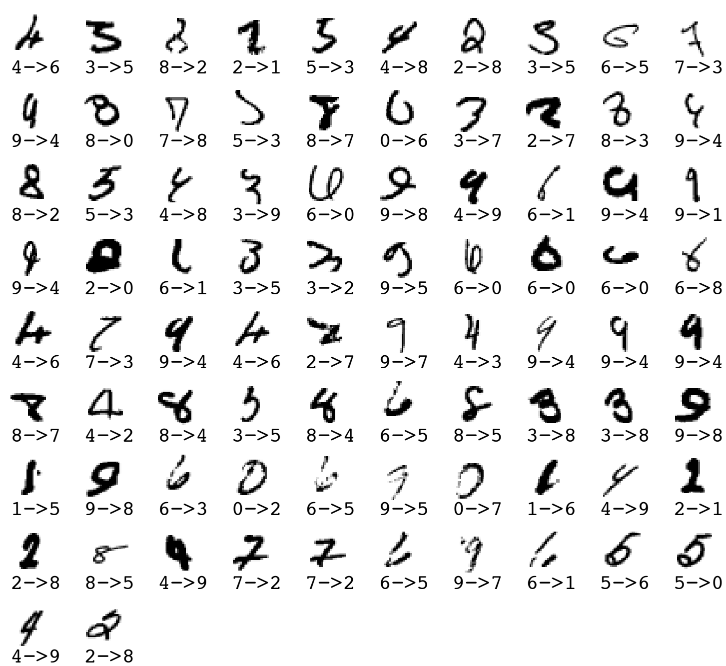
Murphy, 2002

Neural Network



Convolutional Neural Network





LeNet 5