Gentle Introduction to Probabilistic Graphical Models

Seungjin Choi

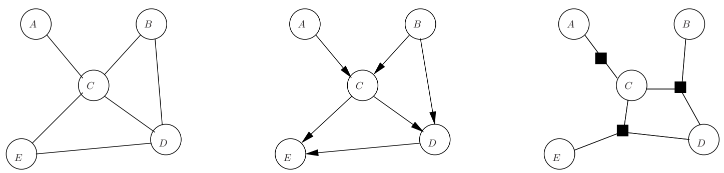
Department of Computer Science
Pohang University of Science and Technology, Korea
seungjin@postech.ac.kr

Textbook, References, and Web

- Primary textbook
  Draft chapters from M. Jordan's unpublished book, "An introduction to probabilistic graphical models".
- References
  - Z. Ghahramani's lecture in "Winter School: Mathematics for Data Modeling"
  - K. Murphy's note, "An introduction to graphical models"
  - C. Bishop's book, "Pattern Recognition and Machine Learning"
  - Selected articles in reading list
- Web
  http://www.postech.ac.kr/∼seungjin/courses/pgm/pgm10.html

Probabilistic Graphical Models

- Learning with probabilistic models: Model distributions over observed data using $p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N|\theta)$.
- Probabilistic graphical models:
  - A happy marriage between graph theory and probability theory.
  - Consider a (directed or undirected) graph where nodes are associated with random variables and edges represent statistical dependencies between variables.



undirected graph   directed graph   factor graph

Why Graphical Models?

- Graphs are an intuitive way of representing and visualizing the relationships between many variables.
- A graph allows us to abstract out the conditional independence relationships between the variables from the details of their parametric forms. Thus we can answer questions like: "Is A dependent on B given that we know the value of C ?" just by looking at the graph.
- Graphical models allow us to define general message passing algorithms that implement probabilistic inference directly. Thus we can answer queries like "What is $P(A|C = c)$ without enumerating all settings of all variables in the model.

"Taken from Z. Ghahramani's talk."

Topics in Graphical Models

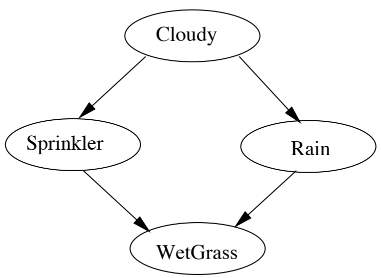
- Representation: How can a graphical model represent a joint probability distribution
  - Directed graphical models
  - Undirected graphical models
- Inference: How can we efficiently infer the hidden states of a system, given partial and possibly noisy observations?
  - Exact inference: belief propagation (sum product, max product), junction tree algorithm.
  - Approximate inference: Laplace approximation, variational Bayes, sampling methods, expectation propagation.
- Learning: structure learning, parameter estimation (maximum likelihood, MAP, Bayesian estimation), EM.

"Taken from K. Murphy's note."

Applications of Graphical Models

- Speech and language processing: HMM and dynamic Bayesian networks
- Computer vision: modeling image data (hierarchical Bayes)
- Computational biology
- Robotics: tracking and localization (Kalman, particle filter)
- Sensor networks: modeling sensor data (Markov networks)
- Planning under uncertainty: dynamic Bayesian networks, Markov decision processes
- Structured data (text, web pages ...): probabilistic relational models

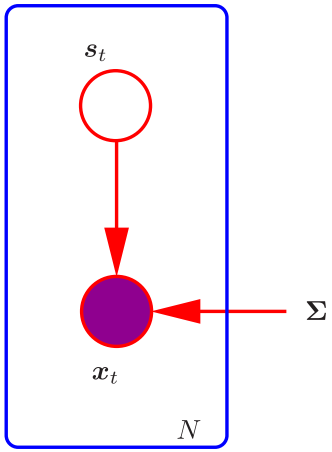
Bayesian Networks



- Directed acyclic graph.
- The representation of $P(C,S,R,W)$ needs $\mathcal{O}(2^4)$ parameters, whereas a BN may need exponentially fewer ($\mathcal{O}(4\cdot 2^2)$).

$$P(C,S,R,W) = P(C)P(S|C)P(R|C)P(W|S,R).$$

Factor Analysis (FA)



Given Gaussian latent variables $\mathbf{s}_t$, the linear generative model obeys

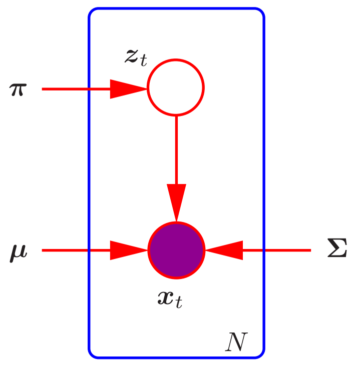
$$\mathbf{x}_t = \mathbf{A}\mathbf{s}_t + \epsilon_t,$$

leading to

$$p(\mathbf{x}_t) = \int p(\mathbf{x}_t|\mathbf{s}_t)p(\mathbf{s}_t)d\mathbf{s}_t = \mathcal{N}\left(\mathbf{x}_t|0, \mathbf{A}\mathbf{A}^\top + \mathbf{\Sigma}\right).$$

Finds a lower-dimensional projection of high-dimensional data that captures the correlation structure of the data.

Mixture of Gaussians (MoG)



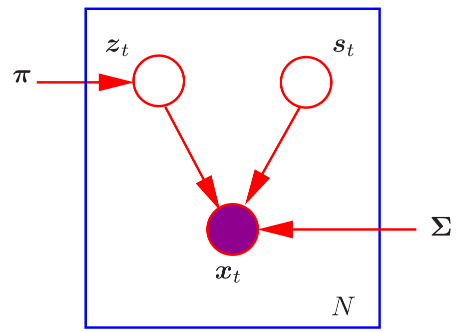
Given multinomial latent variables $\mathbf{z}_t$, the likelihood is written as

$$\prod_{t=1}^{N} p(\mathbf{x}_t) = \prod_{t=1}^{N} \sum_{\mathbf{z}_t} p(\mathbf{x}_t|\mathbf{z}_t)p(\mathbf{z}_t),$$

where

$$\begin{aligned} p(\mathbf{z}_t) &= \prod_{j=1}^{K} \pi_j^{z_{jt}}, \\ p(\mathbf{x}_t|\mathbf{z}_t) &= \mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)^{z_{jt}}. \end{aligned}$$

Mixture of Factor Analyzers (MFA)



Given Gaussian latent variables $\mathbf{s}_t$ and multinomial latent variables $\mathbf{z}_t$, the linear generative model obeys

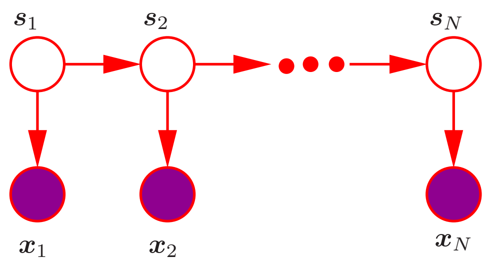
$$\mathbf{x}_t = \mathbf{A}_j \mathbf{s}_t + \boldsymbol{\mu}_j + \boldsymbol{\epsilon}_t,$$

leading to

$$p(\mathbf{x}_t) = \sum_{\mathbf{z}_t} \int p(\mathbf{x}_t|\mathbf{s}_t, \mathbf{z}_t)p(\mathbf{s}_t|\mathbf{z}_t)p(\mathbf{z}_t)d\mathbf{s}_t,$$

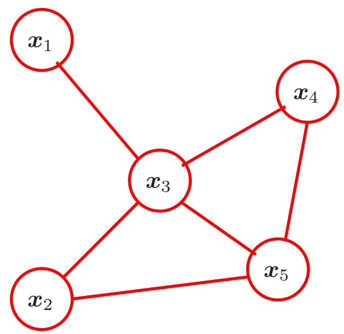
$$\begin{aligned} p(\mathbf{z}_t) &= \prod_{j=1}^{K} \pi_j^{z_{jt}}, \\ p(\mathbf{s}_t|\mathbf{z}_t) &= p(\mathbf{s}_t) = \mathcal{N}(\mathbf{s}_t|0, \mathbf{I}), \\ p(\mathbf{x}_t|\mathbf{s}_t, \mathbf{z}_t) &= \mathcal{N}(\mathbf{x}_t|\mathbf{A}_j\mathbf{s}_t + \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)^{z_{jt}}. \end{aligned}$$

Hidden Markov Model (HMM)



- HMMs are a ubiquitous tool for modeling time series data.
- Assume that the observation x_t was generated by some process whose state s_t is hidden from the observer.
- Discrete hidden state satisfies the Markov property.
- HMMs can be viewed as a particular instance of Bayesian network.

Markov Networks (Undirected Graphical Models)



- Examples of Markov networks
  - Boltzmann machines
  - Markov random fields
- The joint distribution of a Markov network is defined by

$$p(\mathbf{x}) = \frac{1}{Z}\psi_C(\mathbf{x}_C),$$

where $\psi_C(\mathbf{x}_C)$ is a potential function on the clique $\mathbf{x}_C$.