Outline

• Review:
  – Graphical models (DGM, UGM)
  – Learning issues (approaches, observations etc.)

• Parameter learning:
  – Frequentist approach (Likelihood function, MLE)
  – Bayesian approach (Bayes rule, MAP)
  – Detailed example: Gaussian density estimation

• Structure learning:
  – Search-and-score approach

• Conclusion
Review: Graphical Models (GM)

GM = Probability theory + Graph theory

- Tool for dealing with **uncertainty** and **complexity**
- Notion of modularity
- Representation of a GM:
  - A graph is a pair $G = (V, E)$
    - Set of nodes $V = \{X_1, \ldots, X_N\}$
    - Set of edges $E = \{(X_i, X_j); i \neq j\}$
- Lack of edges: Conditional independence!
  - Factorisation of the joint probability distribution
  - Fewer parameters -> learning easier
Review: Directed Graphical Model

= Bayesian network, belief network

- uses Bayes rule for inference

- DAG: Directed acyclic graph (causal dependencies)

- Parent-child relationship: $p(x_i | x_{\pi_i})$

- Directed local Markov property

- Joint probability distribution:

$$p(x_1, \ldots, x_N) = \prod_{i=1}^{N} p(x_i | x_{\pi_i})$$

Factored representation
Review: Undirected Graphical Model

= Markov random field, Markov networks

- Global and local Markov property

- Joint probability distribution:

\[ p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_{X_C}(x_C) \]
**Parameter Vs. Structure Learning**

- **Parameter Learning:**
  - Parameter estimation
  - Discrete: CPD = table
    - For a binary variable
      \[ \theta_{ij} = P(X_i = 1|X_{\pi_i} = j) \]
  - Continuous: CPD = variable
    - For a Gaussian
      \[ \theta = (\mu, \sigma^2) \]

- **Structure Learning:**
  - Model selection
  - Inferring graph G
**Full Vs. Partial Observations**

- **Fully observed variables** (=complete data):
  - Data is obtainable on all variables in the network

- **Partially observed variables** (=incomplete data):
  - Missing data
  - Hidden variables
  - General assumption: *Missing at random*
  - Learning is harder (no close form solution for the likelihood)
Frequentists Vs. Bayesians 1/2

- **The Frequentists:**
  - Probability is an "objective" quantity
  - A parameter $\theta$ is an unknown but fixed quantity
    \( p(x|\theta) \) is a family of distributions indexed by $\theta$
  - Consider various estimators for $\theta$ and choose the "best" one (low bias, low variance)
  - *Likelihood:* Consider $p(x|\theta)$ as a function of $\theta$ for fixed $x$ (inverts relationship between them)
- Advantage:
  - Mathematically / computationally simple
Frequentists Vs. Bayesians 2/2

- **The Bayesians:**
  - Probability is a Person’s degree of belief and therefore “subjective”
  - A parameter $\theta$ is a random variable with a prior distribution (treat model $p(x|\theta)$ as CPD)
  - Update the degree of belief for $\theta$ using Bayes rule (inverts relationship between data and parameter)
  - Data is a quantity to be conditioned on
  - Advantage:
    - Works well when amount of data less than number of parameters
    - Can be used for model selection
Learning Issues

• What will we focus on?

Frequentist
Bayesian
Fully Observed
Partially Observed

Approach
Variables

Model
DGM
UGM

Task
Parameter
Structure
### Overview: Learning Approaches

<table>
<thead>
<tr>
<th>Complete Data</th>
<th>Known structure</th>
<th>Unknown structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter estimation:</td>
<td>- <em>ML, MAP</em></td>
<td>Optimization over structures</td>
</tr>
<tr>
<td><strong>Incomplete data</strong></td>
<td>Parametric optimization:</td>
<td>Optimization over structures and parameters:</td>
</tr>
<tr>
<td></td>
<td><em>EM, gradient descent, stochastic sampling methods</em></td>
<td><em>Structural EM</em></td>
</tr>
</tbody>
</table>
Where are we?

• **Review:**
  – Graphical models (DGM, UGM)
  – Learning issues (approaches, observations etc.)
• **Parameter learning:**
  – Frequentist approach (Likelihood function, MLE)
  – Bayesian approach (Bayes rule, MAP)
  – Detailed example: Gaussian density estimation
• **Structure learning:**
  – Search-and-score approach
• **Conclusion**
Learning Parameters From Data 1/2

- **Given:**  
  - Structure G known and fixed (DAG)  
  - Data set

- **Goal:**  
  - Learn the conditional probability distribution of each node

<table>
<thead>
<tr>
<th>Structure</th>
<th>Dataset</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A B C D E</td>
<td>$p(A)$</td>
</tr>
<tr>
<td></td>
<td>1 2 2 0 1</td>
<td>$p(B</td>
</tr>
<tr>
<td></td>
<td>1 1 0 2 1</td>
<td>$p(C)$</td>
</tr>
<tr>
<td></td>
<td>0 0 1 1 1</td>
<td>$p(D</td>
</tr>
<tr>
<td></td>
<td>1 1 1 1 2</td>
<td>$p(D</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$p(E</td>
</tr>
</tbody>
</table>
• **Maximum likelihood estimation:**
  – Parameter values are fixed but unknown
  – Estimate these values by maximizing the probability of obtaining the samples observed

• **Bayesian estimation:**
  – Parameters are random variables having some known prior distribution
  – Observing new samples converts the prior to a posterior density
Frequentist Approach 1/5

• Given:
  – Data set of M observations \( D = \{x^{(1)}, \ldots, x^{(M)}\} \)

• Assumptions:
  – Observations are \textit{independently} and \textit{identically} distributed according to the JPD (i.i.d. samples)

• Aim:
  – Use the data set \( D \) to estimate the unknown parameter vector \( \theta \)
Frequentist Approach 2/5

- Define the likelihood function:
  \[ L(\theta; D) = p(D|\theta) = p(x^{(1)}, \ldots, x^{(M)}|\theta) \]
- Due to i.i.d. assumption
  \[ L(\theta; D) = \prod_{j=1}^{M} p(x^{(j)}|\theta) \]

- **Maximum likelihood estimation:**
  - Choose the parameter vector \( \theta \) that maximizes the likelihood function
  \[ \hat{\theta}_{ML} = \arg \max_{\theta} L(\theta; D) \]
  - most likely to have generated the data \( D \)
- Trick: Maximize the log-likelihood instead
  \[ l(\theta; D) = \log L(\theta; D) = \sum_{j=1}^{M} \log p(x^{(j)}|\theta) \]
Frequentist Approach 3/5

Detailed example:

• Given: - Network structure
  - Choice of representation for the parameters
  - Data set \( D = \{ x^{(1)}, \ldots, x^{(M)} \} \)

• The log-likelihood function

\[
l(\theta; D) = \sum_{j=1}^{M} \log p(x^{(j)}|\theta)\]

• Factorization due to graph structure

\[
l(\theta; D) = \sum_{j=1}^{M} \log p(x_1^{(j)}|\theta)p(x_2^{(j)}|\theta)p(x_3^{(j)}|x_1^{(j)}, x_2^{(j)}, \theta)\]
Frequentist Approach 4/5

• Assume: Parameter independence

\[
l(\theta; D) = \sum_{j=1}^{M} \log p(x_1^{(j)}|\theta_1) + \sum_{j=1}^{M} \log p(x_2^{(j)}|\theta_2) \\
+ \sum_{j=1}^{M} \log p(x_3^{(j)}|x_1^{(j)}, x_2^{(j)}, \theta_3) \\
= \sum_{i=1}^{3} l(\theta_i; D)
\]

• \(\theta_i\) are the parameters associated with node \(i\)

• Reduced to learning three separate small DAGs
Frequentist Approach 5/5

• Generalizing for any Bayes net

\[
l(\theta; D) = \sum_{i=1}^{N} \sum_{j=1}^{M} \log p(x_i^{(j)} | x_{\pi_i}^{(j)}, \theta_i) \\
= \sum_{i=1}^{N} l(\theta_i; D)
\]

• The likelihood *decomposes* according to the structure of the graph

• **Independent estimation problems:**
  Maximize each likelihood function separately
Bayesian Approach 1/2

• Assumptions:
  1) \( \theta \) is a quantity whose variation can be described by a prior probability distribution \( p(\theta) \)
  2) Samples in the data set \( D = \{x^{(1)}, \ldots, x^{(M)}\} \) are drawn independently from the density \( p(x|\theta) \) whose form is assumed to be known but \( \theta \) is not known exactly
Bayesian Approach 2/2

- Given $D$, the prior distribution can be updated to form the posterior distribution using **Bayes rule**

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

- Link between Frequentist and Bayesian view

$$\text{Posterior} \propto \text{Likelihood} \times \text{prior}$$

- **Maximum a-posterior** (MAP) estimate:

$$\hat{\theta}_{MAP} = \arg \max_{\theta} p(\theta|D)$$

$$= \arg \max_{\theta} p(x|\theta)p(\theta)$$

- MAP = MLE if the prior is *uniform*
Gaussian Density Estimation 1/7

- Univariate Gaussian distribution

\[ p(x|\theta) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\} \]

- Parameter vector: \( \theta = (\mu, \sigma^2) \)
- Given:
  - Multiple observations \( x = \{x_1, \ldots, x_N\} \) which are IID (assumption no necessary)
- Aim:
  - Estimate \( \theta \) based on the observations of \( X \) using a Frequentist and Bayesian approach
FREQUENTIST APPROACH:

• Graphical model:

\[
\begin{array}{cccc}
X_1 & X_2 & \cdots & X_N \\
\end{array}
\]

• „The Frequentists“:
  – No conditioning on the data
  – Use maximum likelihood estimation

• JP written as the product of local probabilities

\[
p(x|\theta) = \prod_{i=1}^{N} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x_i - \mu)^2 \right\} \\
= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2 \right\}
\]
Gaussian Density Estimation 3/7

- The log-likelihood function
  \[ l(\theta; x) = \log p(x|\theta) \]
- Maximization with respect to the parameters \( \mu \) and \( \sigma^2 \)
  \[ \frac{\partial l(\theta;x)}{\partial \mu} = 0 \quad \text{and} \quad \frac{\partial l(\theta;x)}{\partial \sigma^2} = 0 \]
- For a Gaussian distribution:
  - The MLE of the mean = sample mean
    \[ \hat{\mu}_{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i \]
  - The MLE of the variance = sample variance
    \[ \hat{\sigma}^2_{ML} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu}_{ML})^2 \]
BAYESIAN APPROACH:

• „The Bayesians“:
  – Data is conditionally independent given the parameters
  – Choose a prior distribution
• Assume:
  – Variance $\sigma^2$ is a known constant
• Goal:
  – Find the mean $\mu$ to form the posterior $p(\mu|\mathbf{x})$
• Modeling decision:
  – What prior should we take for $\mu$?
Gaussian Density Estimation 5/7

- Take the prior distribution to be Gaussian
  \[ p(\mu) = \frac{1}{(2\pi\sigma_0^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2 \right\} \]

- Hierarchical Bayesian Modeling
  - **Hyperparameter**: Fixed mean \( \mu_0 \) and variance \( \sigma_0^2 \) for \( p(\mu) \)
  - Graphical model:

![Graphical model diagram](image)

- Data is assumed to be *conditionally independent given the parameters*
Gaussian Density Estimation 6/7

• Multiply the prior with the likelihood to obtain the posterior

\[ p(\mu|x) = \frac{1}{(2\pi\tilde{\sigma}^2)^{1/2}} \exp \left\{ -\frac{1}{2\tilde{\sigma}^2}(\mu - \tilde{\mu})^2 \right\} \]

where

\[ \tilde{\mu} = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \bar{x} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 \]

and

\[ \tilde{\sigma}^2 = \left( \frac{1}{N} + \frac{1}{\sigma_0^2} \right)^{-1} \]

• The posterior PD is Gaussian with \((\tilde{\mu}, \tilde{\sigma}^2)\)
  – Linear combination of sample mean and prior mean
  – Inverse of data variance and prior variance add
Gaussian Density Estimation 7/7

• Interpretation of the result:
  – $\tilde{\mu}$ is our best guess after observing $x$
  – $\tilde{\sigma}^2$ is the uncertainty about this guess
  – $\tilde{\mu}$ always lies between $\bar{x}$ and $\mu_0$
    • If $\sigma_0^2 = 0$, then $\tilde{\mu} = \mu_0$ and $\tilde{\sigma}^2 = \sigma^2/N$
      (no prior knowledge can change our opinion)
    • If $\sigma_0^2 >> \sigma^2$, then $\tilde{\mu} \approx \bar{x}$
      (we are very uncertain about our prior guess)
    • With $N \to \infty$ we get $\tilde{\mu} = \bar{x} = \hat{\mu}_{ML}$
      (For set large data the two approaches provide the same result)
Where are we?

• Review:
  – Graphical models (DGM, UGM)
  – Learning issues (approaches, observations etc.)

• Parameter learning:
  – Frequentist approach (Likelihood function, MLE)
  – Bayesian approach (Bayes rule, MAP)
  – Detailed example: Gaussian density estimation

• Structure learning:
  – Search-and-score approach

• Conclusion
Learning Structure From Data

• **Given:**  - Possible prior knowledge about the network structure $G$
  - Data set $D$

• **Goal:**  - Learn the full network structure $G$
  (parameter learning often as sub-problem)
First Approach

• How could we learn a structure?

*Naive approach:*
  – Enumerate all possible network structures
  – Choose the one which maximizes some criteria

Problem:
  – Enumeration becomes feasible for an increasing number of nodes
    E.g. 10 nodes leads to $O(10^{18})$ structures

• Unless we have prior (expert) knowledge to eliminate some possible structures, use statistically efficient search strategies
Equivalent Probability Models

• Given: GM with 3 nodes (binary random variables)
• Number of possible structure: 25

\( S_A \)
\[
\begin{array}{c}
X_1 \\
\downarrow \\
X_3 \\
\end{array}
\quad X_2
\]

\( S_B \)
\[
\begin{array}{c}
X_1 \\
\downarrow \\
X_3 \\
\end{array}
\quad X_2
\]

\( S_C \)
\[
\begin{array}{c}
X_1 \\
\downarrow \\
X_3 \\
\end{array}
\quad X_2
\]

\( S_D \)
\[
\begin{array}{c}
X_1 \\
\downarrow \\
X_3 \\
\end{array}
\quad X_2
\]

• **Structure** \( S_C \): \[ p_C(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2) \]

• **Structure** \( S_D \): \[ p_D(x_1, x_2, x_3) = p(x_1|x_2)p(x_2|x_3)p(x_3) \]

Using Bayes rule: \[ p_C(x_1, x_2, x_3) = p_D(x_1, x_2, x_3) \]

\( \iff \) Equivalent probability models
Search-And-Score Approach 1/2

• Idea:
  – Define a score function for measuring model quality (e.g. penalized likelihood)
  – Use search algorithm to find a (local) maximum of the score

• Scoring function:
  – Statistically motivated
  – Assigns a score $S(G)$ to the graph $G$

• Goal:
  – Find the structure with the best score $S(G|D)$ given the data set $D$
Search-And-Score Approach 2/2

• **Frequentist way:**
  – Maximize the likelihood of the data

\[
S(G) = p(D|G, \hat{\theta}_{ML}) = \prod_{i=1}^{N} p(x_i|x_{\pi_i}, G, \hat{\theta}_{ML})
\]

• **Bayesian score:**
  – \(S(G)\) is proportional to the posterior probability of a network structure given the data \(D\)

\[
S(G) = p(G|D) = \frac{p(D|G)p(G)}{p(D)}
\]

where

\[
p(D|G) = \int p(D|G, \theta)p(\theta|G)d\theta
\]

• Use **search methods** to find the optimal structure
Where are we?

• Review:
  – Graphical models (DGM, UGM)
  – Learning issues (approaches, observations etc.)

• Parameter learning:
  – Frequentist approach (Likelihood function, MLE)
  – Bayesian approach (Bayes rule, MAP)
  – Detailed example: Gaussian density estimation

• Structure learning:
  – Search-and-score approach

• Conclusion
Conclusion

• **Parameter learning:**
  – *Frequentist approach:*
    • Use Maximum likelihood estimate
  – *Bayesian approach:*
    • Use Maximum a-posteriori estimate
  – Approaches are equivalent for large data sizes

• **Structure learning:**
  – *Search-and-score approach:*
    • Optimize according to some scoring function
    • Use search methods to find the optimal structure