Advanced Signal Processing 2 SE

Parameter and Structure Learning in Graphical Models

02.05.2005

Stefan Tertinek
turtle@sbox.tugraz.at

TUG

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 $E = \left\{(X_i, X_j); i \neq j\right\}$
- Lack of edges: Conditional independence!
    - Factorisation of the joint probability distribution
    - Fewer parameters -> learning easier

Review: Directed Graphical Model

= Bayesian network, belief network

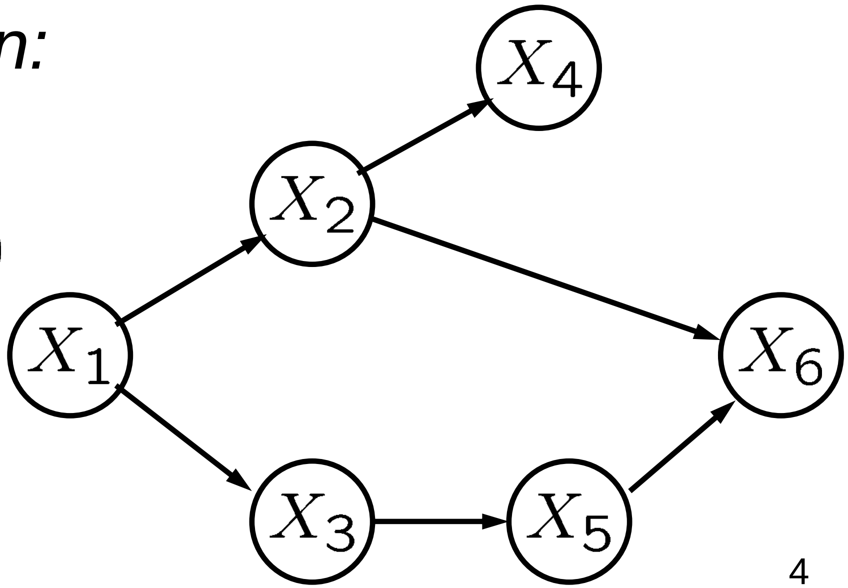
→ uses Bayes rule for inference

- DAG: Directed acyclic graph (causal dependencies)
- Parent-child relationsship: $p(x_i|\mathbf{x}_{\pi_i})$
- Directed local Markov property

- *Joint probability distribution:*

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

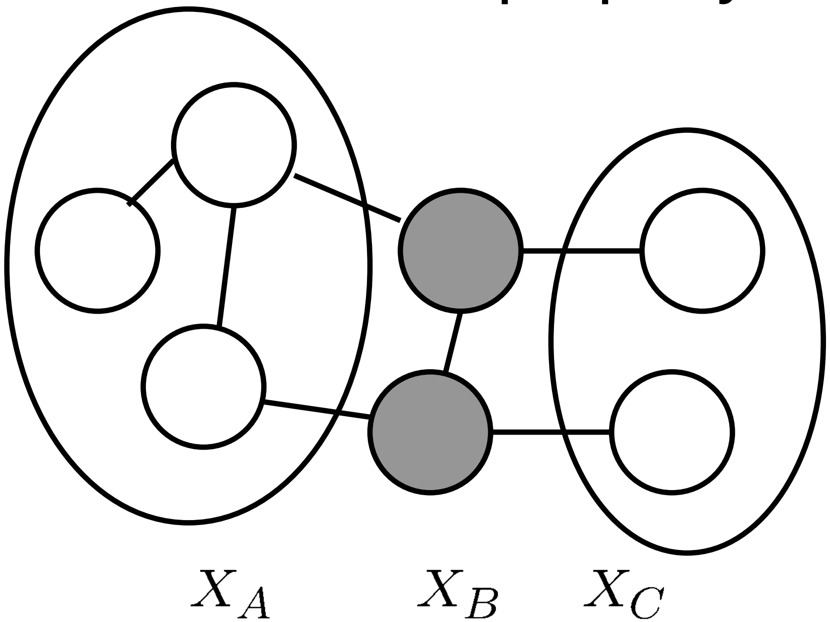
Factored representation



Review: Undirected Graphical Model

= Markov random field, Markov networks

- Global and local Markov property



- *Joint probability distribution:*

$$p(x) = \frac{1}{Z} \prod_{C \in \mathbf{C}} \psi_{X_C}(x_C)$$

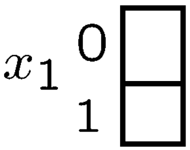
Parameter Vs. Structure Learning

- ***Parameter Learning*:**
  = parameter estimation
- Discrete: CPD = table
  - For a binary variable

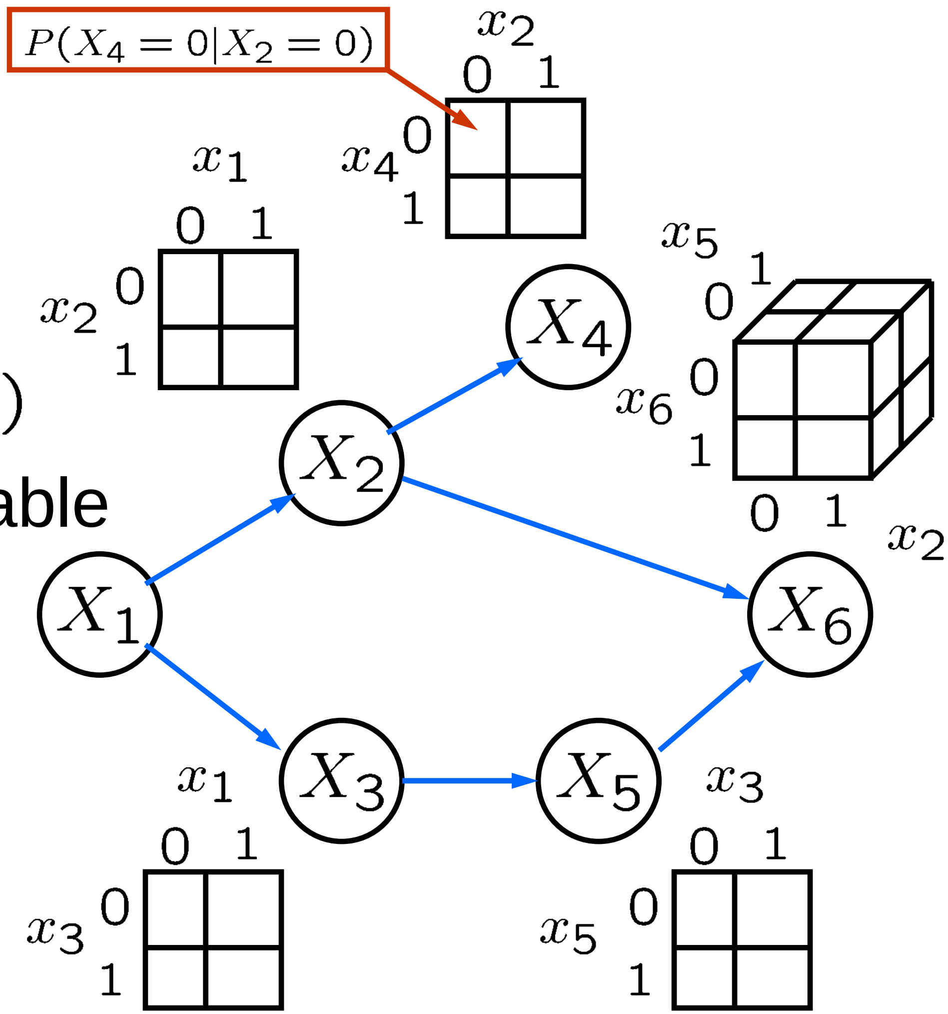
$$\theta_{ij} = P(X_i = 1 | X_{\pi_i} = j)$$

- *Continuous:* CPD = variable
  - For a Gaussian

$$\boldsymbol{\theta} = (\mu, \sigma^2)$$



- ***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

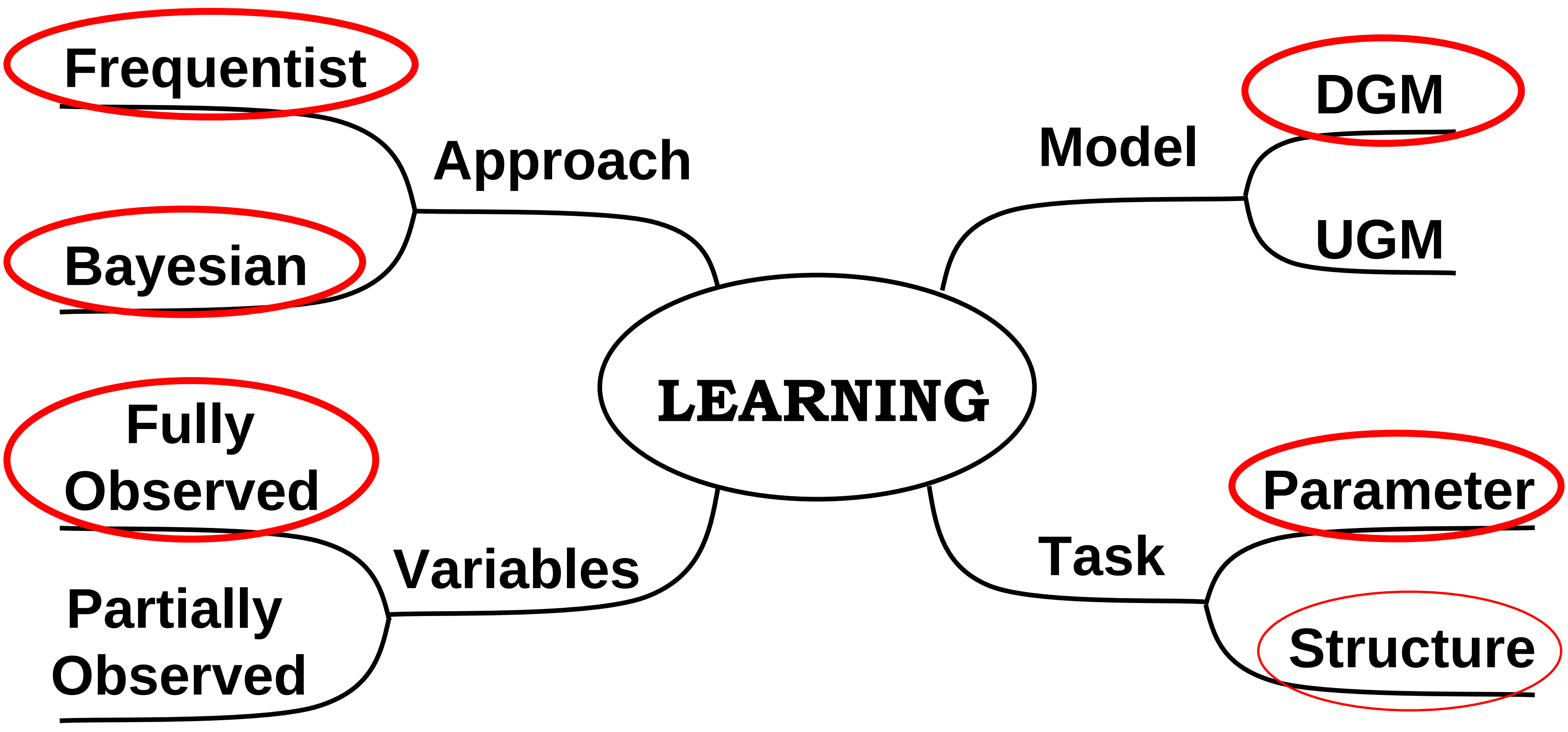
- **The Frequentists:**
  - Probability is an „objective“ quantity
  - A parameter θ is an unknown but fixed quantity ($p(\mathbf{x}|\theta)$ is a family of distributions indexed by θ)
  - Consider various estimators for θ and choose the „best“ one (low bias, low variance)
  - *Likelihood:* Consider $p(\mathbf{x}|\theta)$ as a function of θ 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 θ is a random variable with a prior distribution (treat model $p(\mathbf{x}|\theta)$ as CPD)
  - Update the degree of belief for θ 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?



Overview: Learning Approaches

	Known structure	**Unknown structure**
Complete Data	**Parameter estimation**: *ML, MAP*	**Optimization over structures**
Incomplete data	**Parametric optimization:** *EM*, *gradient descent, stochastic sampling methods*	**Optimization over structures and parameters:** *Structural EM*

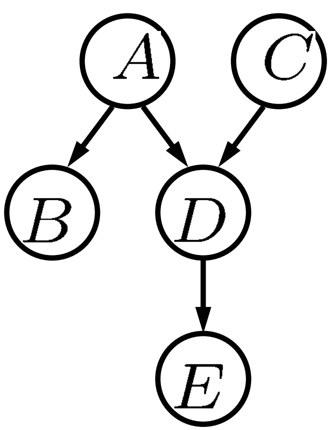
Where are we?

- **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

Structure



Dataset

A	B	C	D	E
1	2	2	0	1
1	1	0	2	1
0	0	1	1	1
1	1	1	1	2

Parameters

$p(A)$

$p(B|A)$

$p(C)$

$p(D|A,B)$

$p(D|A)$

$p(E|D)$

Learning Parameters From Data 2/2

- **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 $\mathbf{D} = \left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}\right\}$
- Assumptions:
  - Observations are *independently* and *identically* distributed according to the JPD (i.i.d. samples)
- Aim:
  - Use the data set $\mathbf{D}$ to estimate the unknown parameter vector $\boldsymbol{\theta}$

Frequentist Approach 2/5

- Define the likelihood function:

$$L(\boldsymbol{\theta}; \mathbf{D}) = p(\mathbf{D}|\boldsymbol{\theta}) = p(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}|\boldsymbol{\theta})$$

- Due to i.i.d. assumption

$$L(\boldsymbol{\theta}; \mathbf{D}) = \prod_{j=1}^{M} p(\mathbf{x}^{(j)}|\boldsymbol{\theta})$$

- **Maximum likelihood estimation:**
  - Choose the parameter vector $\boldsymbol{\theta}$ that *maximizes* the likelihood function

$$\hat{\boldsymbol{\theta}}_{ML} = \text{arg max}_{\boldsymbol{\theta}} \, L(\boldsymbol{\theta}; \mathbf{D})$$

  - most likely to have generated the data $\mathbf{D}$
- Trick: Maximize the log-likelihood instead

$$l(\boldsymbol{\theta}; \mathbf{D}) = \log L(\boldsymbol{\theta}; \mathbf{D}) = \sum_{j=1}^{M} \log p(\mathbf{x}^{(j)}|\boldsymbol{\theta})$$

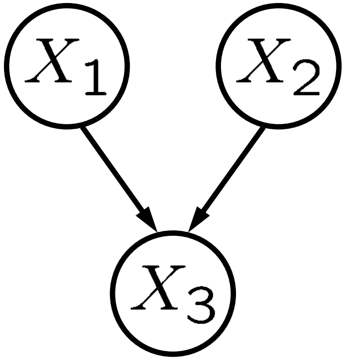
Frequentist Approach 3/5

Detailed example:

- Given: - Network structure

  - Choice of representation for the parameters

  - Data set $\mathbf{D} = \left\{ \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)} \right\}$



- The log-likelihood function

$$l(\boldsymbol{\theta}; \mathbf{D}) = \sum_{j=1}^{M} \log p(\mathbf{x}^{(j)} | \boldsymbol{\theta})$$

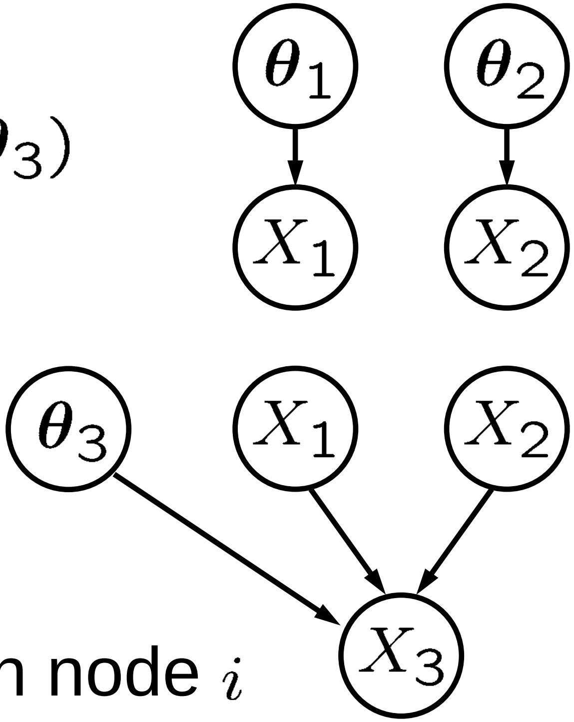
- Factorization due to graph structure

$$l(\boldsymbol{\theta}; \mathbf{D}) = \sum_{j=1}^{M} \log p(x_1^{(j)} | \boldsymbol{\theta}) p(x_2^{(j)} | \boldsymbol{\theta}) p(x_3^{(j)} | x_1^{(j)}, x_2^{(j)}, \boldsymbol{\theta})$$

Frequentist Approach 4/5

- Assume: Parameter independence

$$
\begin{aligned}
l(\boldsymbol{\theta}; \mathrm{D}) &= \sum_{j=1}^{M} \log p(x_1^{(j)} | \boldsymbol{\theta}_1) + \sum_{j=1}^{M} \log p(x_2^{(j)} | \boldsymbol{\theta}_2) \\
&+ \sum_{j=1}^{M} \log p(x_3^{(j)} | x_1^{(j)}, x_2^{(j)}, \boldsymbol{\theta}_3) \\
&= \sum_{i=1}^{3} l(\boldsymbol{\theta}_i; \mathrm{D})
\end{aligned}
$$



- $\boldsymbol{\theta}_i$ are the parameters associated with node i
- Reduced to learning *three* sparate small DAGs

Frequentist Approach 5/5

- Generalizing for any Bayes net

$$\begin{aligned} l(\boldsymbol{\theta}; \mathbf{D}) &= \sum_{i=1}^{N} \sum_{j=1}^{M} \log p(x_i^{(j)} | \mathbf{x}_{\pi_i}^{(j)}, \boldsymbol{\theta}_i) \\ &= \sum_{i=1}^{N} l(\boldsymbol{\theta}_i; \mathbf{D}) \end{aligned}$$

- The likelihood *decomposes* according to the structure of the graph
- **Independent estimation problems:**

  Maximize each likelihood function separately

Bayesian Approach 1/2

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

Bayesian Approach 2/2

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

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

- Link between Frequentist and Bayesian view

Posterior $\propto$ Likelihood x prior

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

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{MAP} &= \text{arg max}_{\boldsymbol{\theta}}\ p(\boldsymbol{\theta}|\mathrm{D}) \\ &= \text{arg max}_{\boldsymbol{\theta}}\ p(x|\boldsymbol{\theta})p(\boldsymbol{\theta}) \end{aligned}$$

- MAP = MLE if the prior is *uniform*

Gaussian Density Estimation 1/7

- Univariate Gaussian distribution

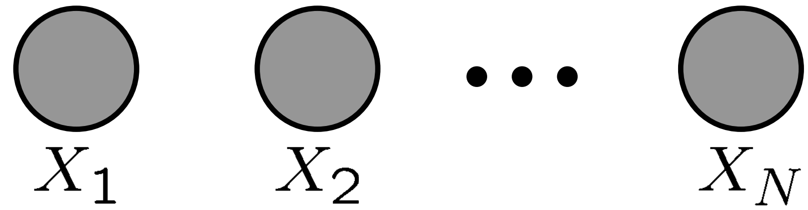
$$p(x|\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

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

Gaussian Density Estimation 2/7

FREQUENTIST APPROACH:

- Graphical model:



- *„The Frequentists“:*
  - No conditioning on the data
  - Use maximum likelihood estimation
- JP written as the product of local probabilites

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\}$$

$$= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right\}$$

Gaussian Density Estimation 3/7

- The log-likelihood function

$$l(\boldsymbol{\theta}; \mathrm{x}) = \log p(\mathrm{x}|\boldsymbol{\theta})$$

- Maximization with respect to the parameters μ and σ^2

$$\frac{\partial l(\boldsymbol{\theta};\mathrm{x})}{\partial \mu} = 0 \quad \text{and} \quad \frac{\partial l(\boldsymbol{\theta};\mathrm{x})}{\partial \sigma^2} = 0$$

- For a Gaussian distribution:
  - The MLE of the mean = sample mean

$$\widehat{\mu}_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

  - The MLE of the variance = sample variance

$$\widehat{\sigma}^2_{ML} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \widehat{\mu}_{ML})^2$$

Gaussian Density Estimation 4/7

BAYESIAN APPROACH:

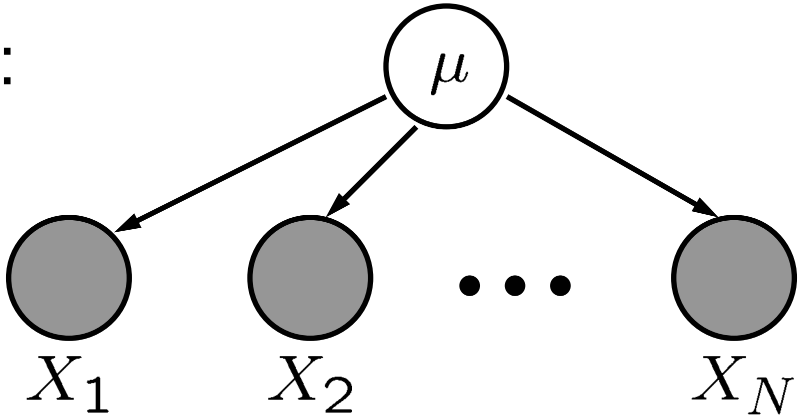
- *„The Bayesians“:*
  - Data is conditionally independent given the parameters
  - Choose a prior distribution
- Assume:
  - Variance σ^2 is a known constant
- Goal:
  - Find the mean μ to form the posterior $p(\mu|\mathbf{x})$
- Modeling decision:
  - What prior should we take for μ?

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 μ_0 and variance σ_0^2 for $p(\mu)$
- Graphical model:



- 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

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

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

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

- The posterior PD is Gaussian with $\left(\tilde{\mu}, \tilde{\sigma}^2\right)$
  - 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 $\mathbf{x}$
  - $\tilde{\sigma}^2$ is the uncertainty about this guess
  - $\tilde{\mu}$ always lies between $\bar{x}$ and μ_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 $N \rightarrow \infty$ 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**

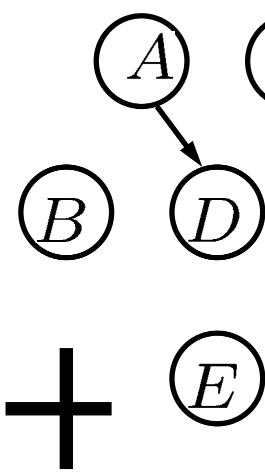
Learning Structure From Data

- <u>Given:</u> - Possible prior knowledge about the network structure G

  - Data set D
- <u>Goal:</u> - Learn the full network structure G

  (parameter learning often as sub-problem)

A	B	C	D	E
1	2	2	0	1
1	1	0	2	1
0	0	1	1	1
1	1	1	1	2



$p(A)$

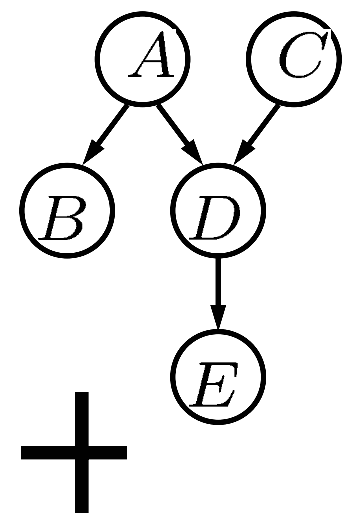
$p(B|A)$

$p(C)$

$p(D|A,B)$

$p(D|A)$

$p(E|D)$



First Approach

- How could we learn a structure?

  Naive approach:
  - Enumterate all possible network structures
  - Choose the one which maximizes some criteria

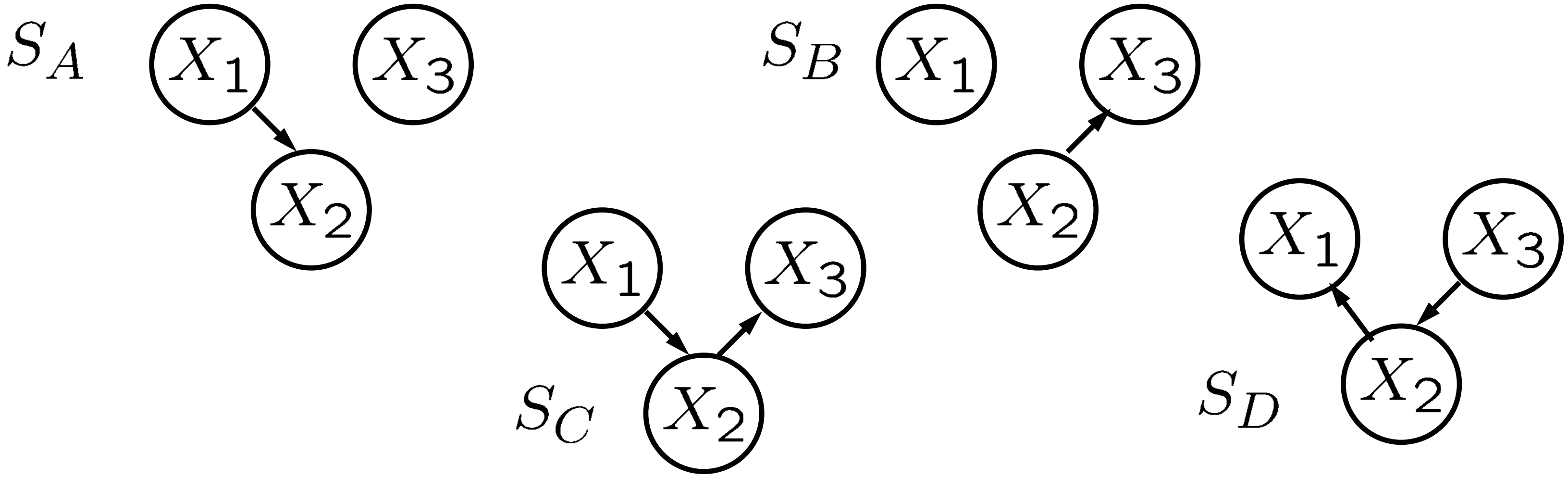
  Problem:
  - Enumeration becomes feasible for an increasing number of nodes

    E.g. 10 nodes leads to $O(10^{18})$ structures
- Unless we have prior (expert) knowledge to eliminiate some possible structures, use

  statistically efficient search strageties

Equivalent Probability Models

- Given: GM with 3 nodes (binary random variables)
- Number of possible structures: 25



- 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: $p_C(x_1, x_2, x_3) = p_D(x_1, x_2, x_3)$

  ⇨ 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* $S(G|\mathbf{D})$ given the data set $\mathbf{D}$

Search-And-Score Approach 2/2

- Frequentist way:
  - Maximize the likelihood of the data

$$S(G) = p(\mathbf{D}|G, \hat{\boldsymbol{\theta}}_{ML}) = \prod_{i=1}^{N} p(x_i|\mathbf{x}_{\pi_i}, G, \hat{\boldsymbol{\theta}}_{ML})$$

- Bayesian score:
  - $S(G)$ is proportional to the posterior probability of a network structure given the data $\mathbf{D}$

$$S(G) = p(G|\mathbf{D}) = \frac{p(\mathbf{D}|G)p(G)}{p(\mathbf{D})}$$

where

$$p(\mathbf{D}|G) = \int p(\mathbf{D}|G, \boldsymbol{\theta})p(\boldsymbol{\theta}|G)d\boldsymbol{\theta}$$

- 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

References

- Heckerman, D. (1995). A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research.
- Buntine, W. (1996) A Guide to the Literature on Learning Probabilistic Networks From Data. IEEE transactions On Knowledge and Data Engineering
- P.J. Krause (1998), Learning Probabilistic Networks, Knowledge Engineering Review 13, 321-351.
- Selim Aksoy, Lecture slides, CS 551Pattern Recognition

  http://www.cs.bilkent.edu.tr/~saksoy/courses/cs551/index.html