Structure Learning

EECS 395/495 Fall 2014

Road Map

- Basics of Probability and Statistical Estimation
- Bayesian Networks
- Markov Networks
- Inference
- Learning
  - Parameters, **Structure**, EM
- HMMs

Learning Structure

- Hard problem
  - Finding the BN structure with the highest "score" among those structures with at most *k* parents is NP hard for $k>1$ (Chickering, 1995)
- Inputs
  - Data (potentially incomplete)
- Outputs
  - Graphical model structure (we'll focus on Bayes Nets)
- Approaches
  - Constraint-based
  - Score-based approaches
    - Local search
  - Bayesian Model Averaging

Learning Structure

- Hard problem
  - Finding the BN structure with the highest "score" among those structures with at most *k* parents is NP hard for *k*>1 (Chickering, 1995)
- Inputs
  - Data (potentially incomplete)
- Outputs
  - Graphical model structure (we'll focus on Bayes Nets)
- Approaches
  - **Constraint-based**
  - Score-based approaches
    - Local search
  - Bayesian Model Averaging

Constraint-based Approaches

- Idea: we know how to construct a Bayes Net if we can perform **independence tests**
  - (**A** $\perp$ **B** | **C**) ?
- Naïve construction
  - depends on variable ordering
  - Issues potentially large number of independence queries
- A more sophisticated PDAG construction process works better (see book)

Constraint-based approach guarantees

- Can uncover a *perfect* map using a polynomial # of tests if:
  - Bounded in-degree *d* in *G** (the true graph)
  - Perfect independence queries up to size 2*d* + 2
    (Strong)
  - P* (true dist.) is *faithful* to G*
    (Also strong)
    - i.e., any independencies in P* reflected as d-separation in G*

Learning Structure

- Approaches
  - Constraint-based
  - **Score-based approaches**
    - **Local search**
  - Bayesian Model Averaging

Scoring Structures

- Maximum likelihood G
  - Choose $G = \arg\max_G \max_\theta P(\text{Data} \mid \theta)$
- Or MAP:
  - Choose $G = \arg\max_G \max_\theta P(\text{Data} \mid \theta)\, P(\theta)$

- …what's wrong with these?

Bayesian Score

- Bayesian Score for G =

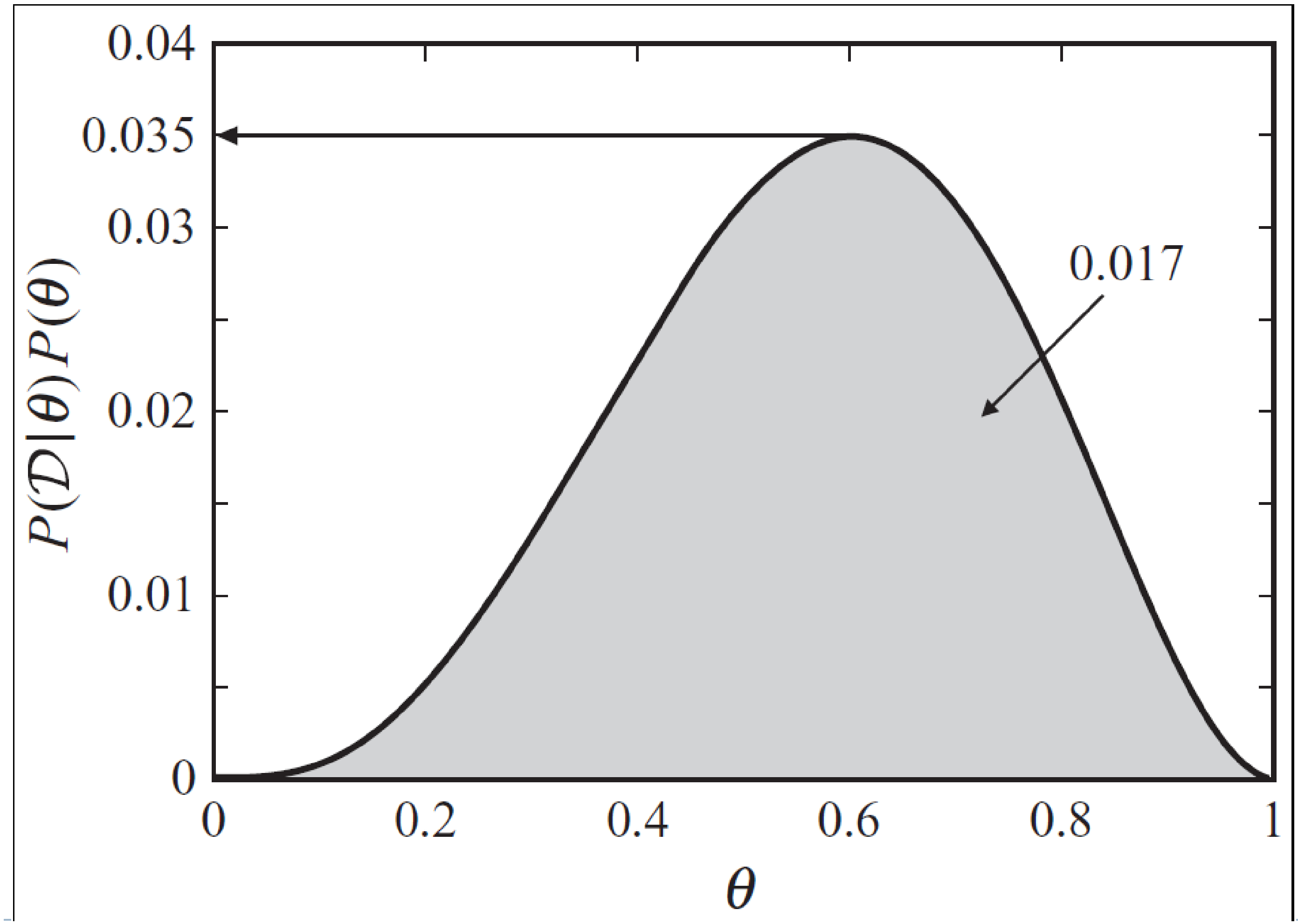
  prior for G

  +

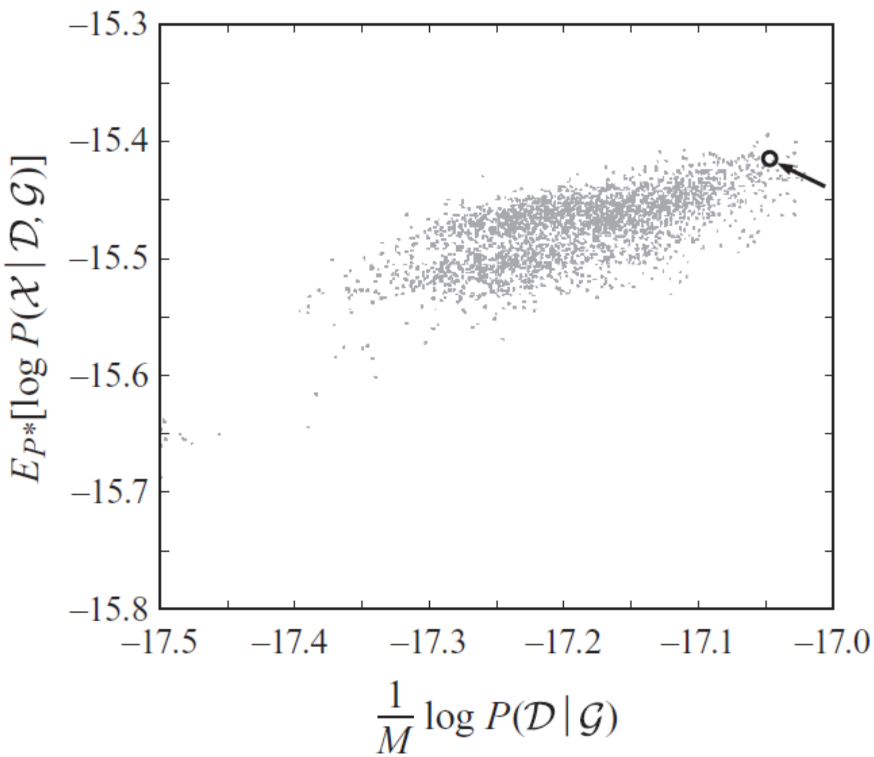
  likelihood integrated over all parameters for G

- BayesianScore(G : Data) = log P(Data | G) + log P(G)
- $P(\text{Data} \mid G) = \int_{\Theta_G} P(\text{Data} \mid \theta_G, G)\, P(\theta_G \mid G)\, d\theta_G$

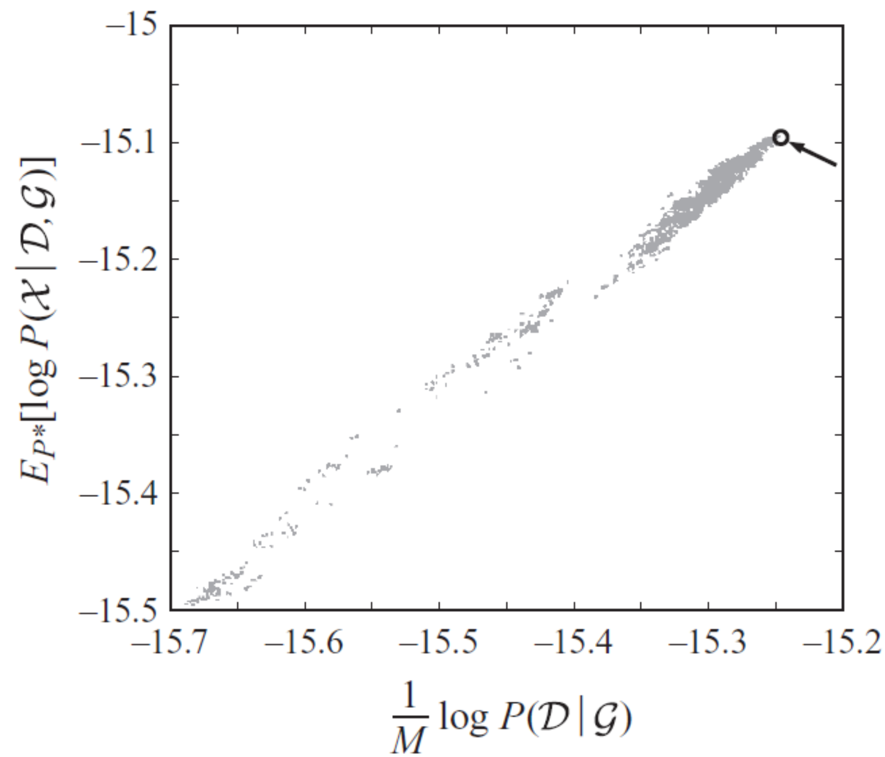
Integrating over parameters



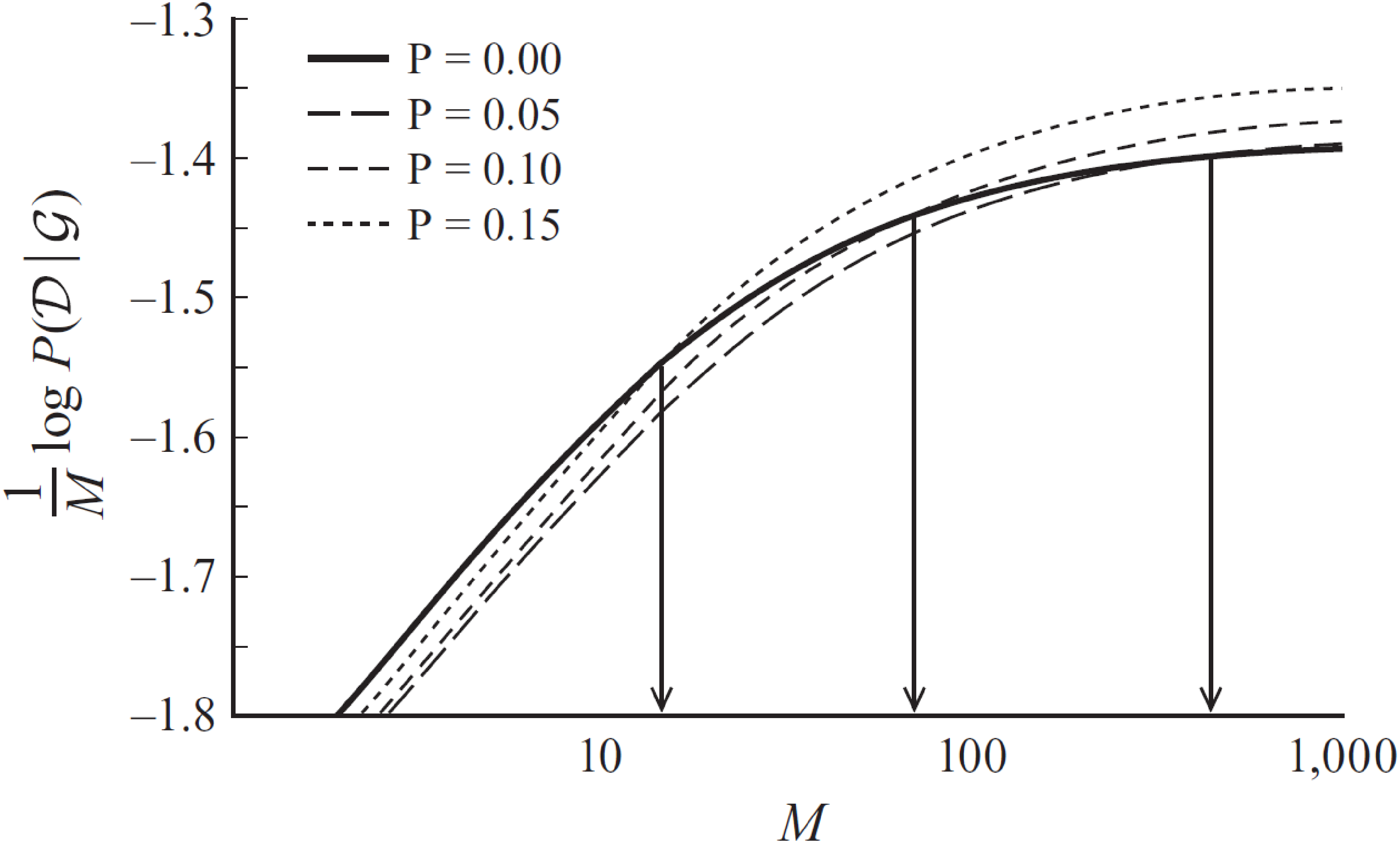
Training (x-axis) vs. Test (y-axis) Perf.



500 instances



10,000 instances



P = 0.00
P = 0.05
P = 0.10
P = 0.15
–1.3
–1.4
–1.5
–1.6
–1.7
–1.8
$\frac{1}{M} \log P(\mathcal{D} \mid \mathcal{G})$
10
100
1,000
M

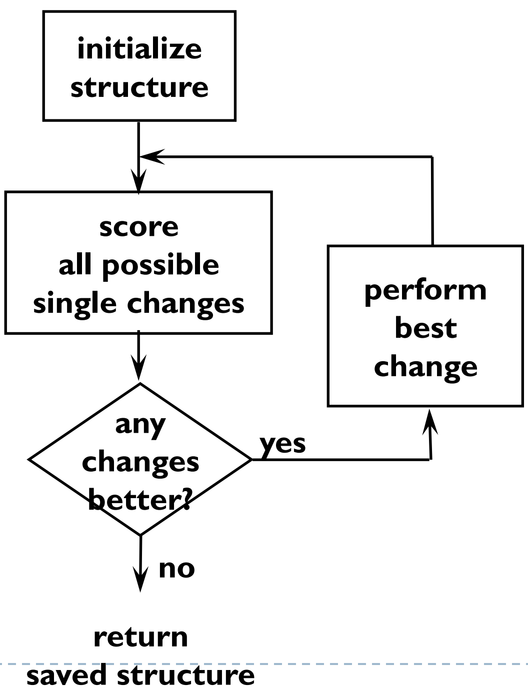
Bayesian Information Criterion

- Bayes Score includes:
  - $P(\text{Data} \mid G) = \int_{\Theta_G} P(\text{Data} \mid \theta_G, G)\, P(\theta_G \mid G)\, d\theta_G$
- Integral sometimes difficult
- Approximation:

$$\text{score}_{\text{BIC}}(G) = -(\text{Dim}[G]/2) \log M + \log \max_{\theta_G} P(\text{Data} \mid \theta_G)$$

Structure search

- Finding the BN structure with the highest score among those structures with at most *k* parents is NP hard for *k*>1 (Chickering, 1995)
- Heuristic methods
  - Greedy
  - Greedy with restarts



Structure priors

- Lots of options
  - All possible structures equally likely
  - Partial ordering, required / prohibited arcs
  - Prior(*G*) α Similarity(*G*, *G*prior)

Learning Structure

- Approaches
  - Constraint-based
  - Score-based approaches
    - Local search
  - **Bayesian Model Averaging**

Bayesian Model Averaging

- Previous methods all find a single graph G
- *Bayesian model averaging* instead makes predictions by averaging over structures:

  $$P(\text{test example} \mid \text{Data}) = \sum_G P(\text{test example} \mid \text{Data}, G)\, P(G \mid \text{Data})$$