Networks

Network features and confidence assessment

Outline

- Identifying high-confidence features of learned Bayesian networks
- Markov blankets
- The Bootstrap method
- Permutation testing
- Yeast dataset case study

Bayes Net Structure Learning Case Study:
Friedman et al., *JCB* 2000

- expression levels in populations of yeast cells
- 800 genes
- 76 experimental conditions
- used two representations of the data
  - discrete representation (underexpressed, normal, overexpressed) with CPTs in the models
  - continuous representation with linear Gaussians

Bayes Net Structure Learning Case Study: Two Key Issues

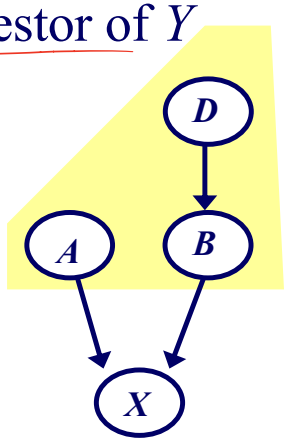
- Since there are many variables but data is sparse, there is not enough information to determine the "right" model. Instead, can we consider many of the high-scoring networks?

- How can we tell if the structure learning procedure is finding real relationships in the data? Is it doing better than chance?

Representing Partial Models

- How can we consider many high-scoring models? Use the *bootstrap* method to identify high-confidence *features* of interest.

- Friedman et al. focus on finding two types of "features" common to lots of models that could explain the data
  - Markov relations: is *Y* in the *Markov blanket* of *X*?
  - order relations: is *X* an ancestor of *Y*

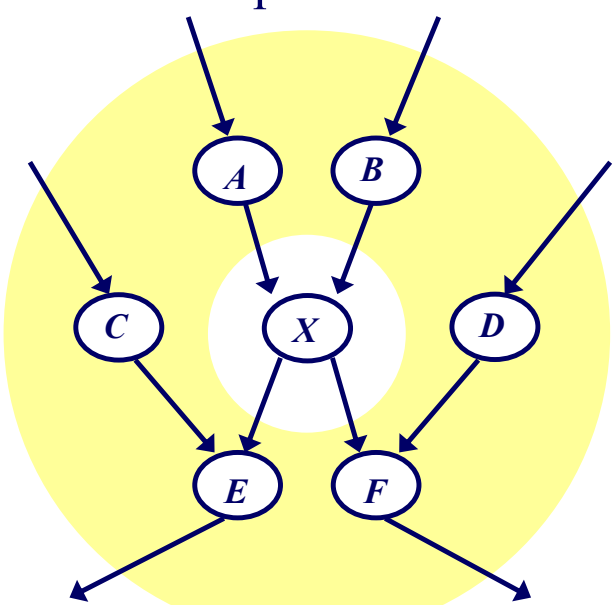


Markov Blankets

- every other node Y in the network is conditionally independent of X when conditioned on X's Markov blanket MB(X)

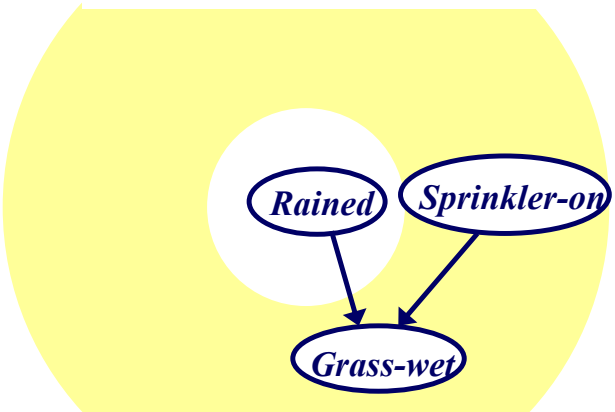
$$\Pr(X \mid \mathrm{MB}(X), Y) = \Pr(X \mid \mathrm{MB}(X))$$

- the Markov blanket for node X consists of its parents, its children, and its children's parents



Markov Blankets

- why are parents of X's children in its Markov blanket?
- suppose we're using the following network to infer the probability that it rained last night



we observe the grass is wet; is the *Sprinkler-on* variable now irrelevant?

no – if we observe that the sprinkler is on, this helps to "explain away" the grass being wet

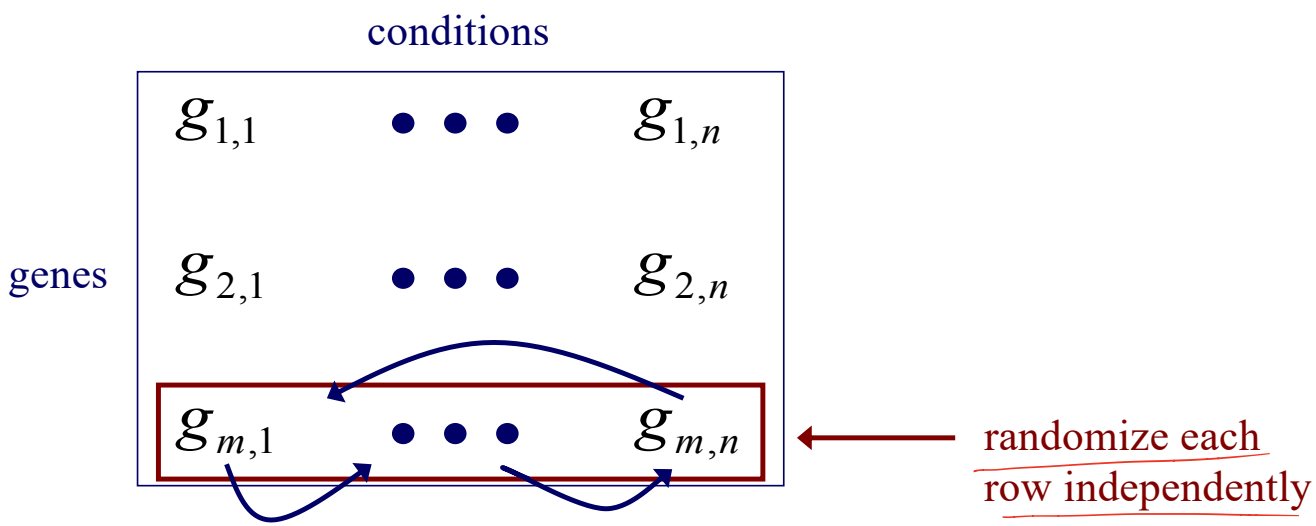
Estimating Confidence in Features: The *Bootstrap* Method

- for $i = 1$ to m data sets
  - randomly draw sample S_i of N expression experiments from the original N expression experiments with replacement
  - learn a Bayesian network B_i from S_i

- some expression experiments will be included multiple times in a given sample, some will be left out.
- the confidence in a feature is the fraction of the m models in which it was represented

Permutation Testing: Do the Networks Represent Real Relationships

- how can we tell if the high-confidence features are meaningful?
- compare against confidence values for *randomized* data – genes should then be independent and we shouldn't find "real" features



Confidence Levels of Features: Real vs. Randomized Data

Markov features

order features

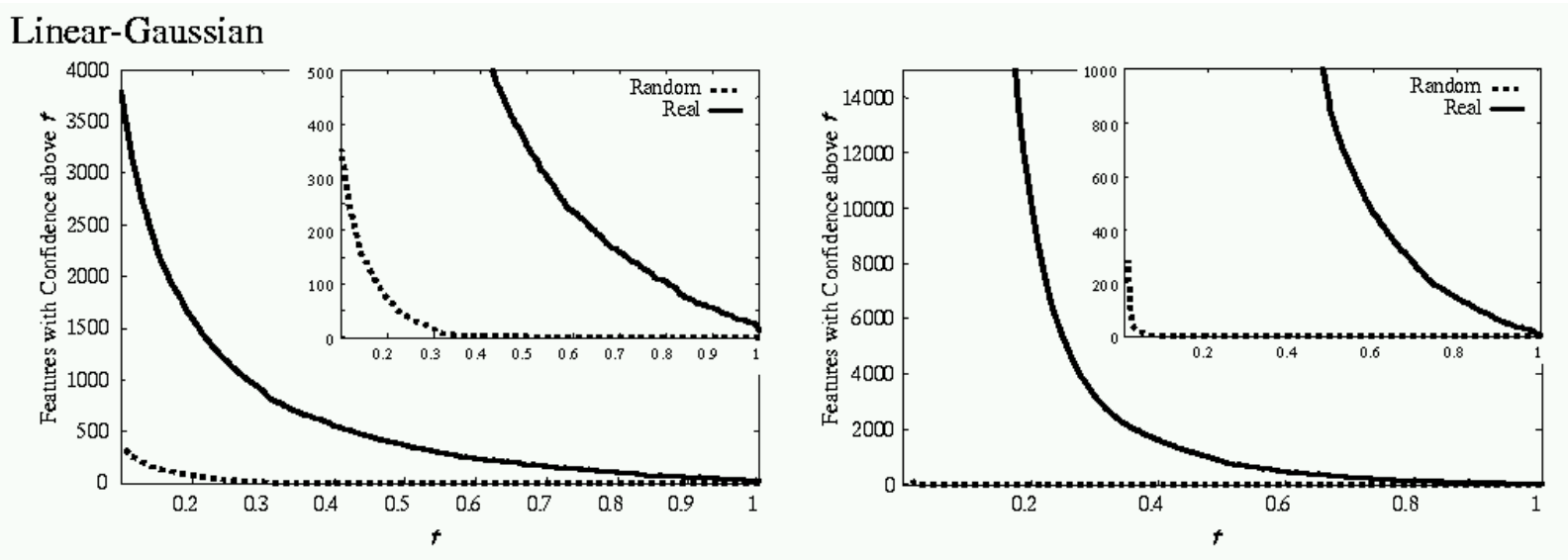


figure from Friedman et al., *Journal of Computational Biology*, 2000

Summary

- We can use the bootstrap to identify high-confidence features in our learned Bayesian networks
- Possible features included ancestor relationships between random variables and Markov blankets
- Permutation testing can be used to establish a confidence threshold in order to minimize prediction of false positive features