Model comparison

Overview

- We have already covered **nested model comparison** earlier in the semester – we'll review that briefly
- Then we'll look at model comparison for **causal models**
- This will lead us naturally to **non-nested model comparison**
- We will introduce **held-out evaluation techniques**
- And, time permitting, we will discuss **subtleties of held-out evaluation for multi-level data**

Nested model comparison – recap

- Model A is **nested** in model B if every parameter setting for A produces a predictive distribution that can also be produced by some parameter setting for model B
  - Often, but not always, this involves setting some subset of model B's parameters to 0
- If A is nested in B, and their # model parameters are k_A and k_B respectively, then under fairly general conditions:

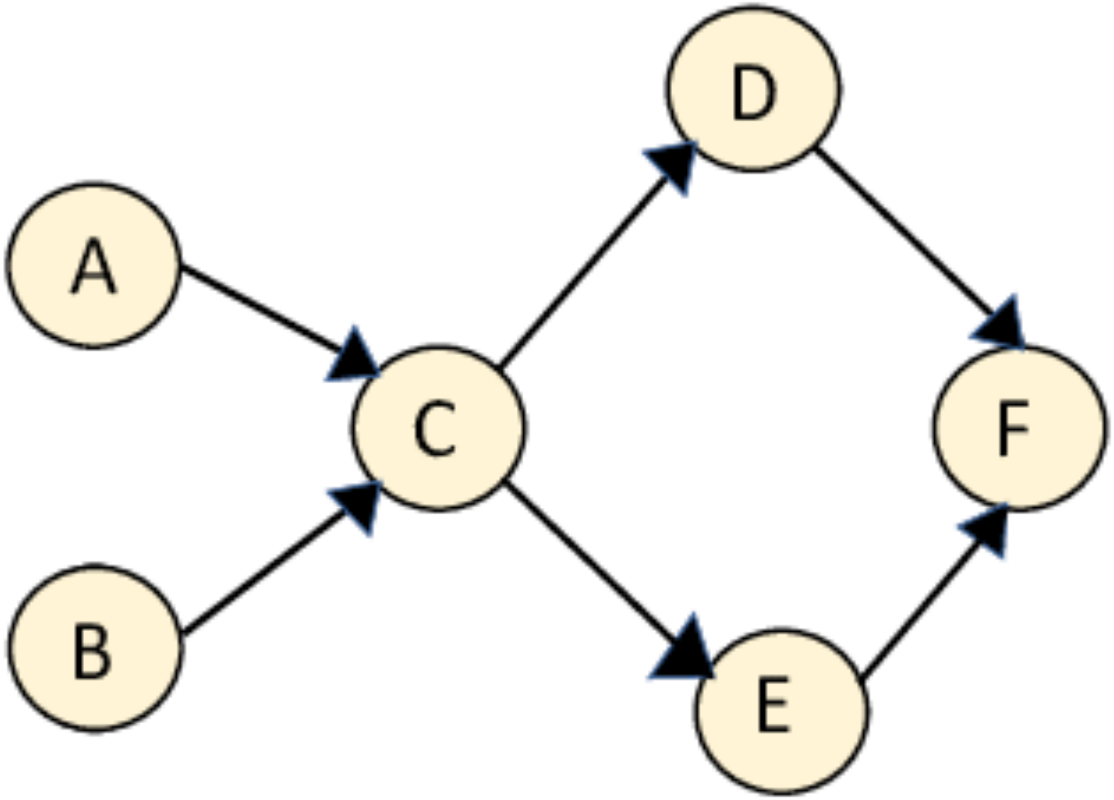
$$-2\log\frac{\max \mathrm{Lik}_A(y)}{\max \mathrm{Lik}_B(y)}$$

(where y are the data)

is asymptotically χ^2 distributed with $k_B - k_A$ deg. freedom
- (There are also more specialized tests for nested models, such as the F-test)

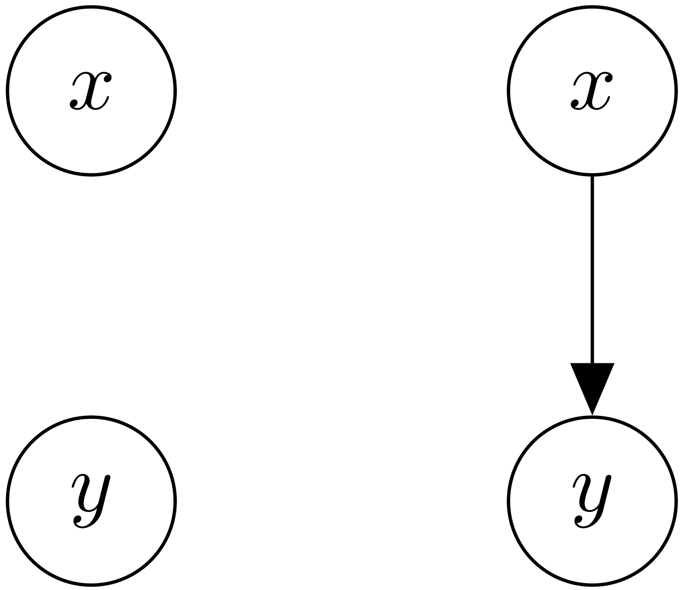
Applications to directed graphical models

- How can we apply these to directed graphical models?



Applications to directed graphical models

- Here are two models that are nested with respect to observational data:



- The number of parameters by which they differ depends on the representations & distributional assumptions regarding x and y

Applications to directed graphical models

- What is an example of **non-nested** directed graphical models?

-

Applications to causal graphical models

- Two DAGs can be **observationally equivalent** but have different **interventional distributions** (exercise: come up with an example)

Applications to causal graphical models

- Can two DAGs be in nested with respect to observational data, but non-nested with respect to the set of possible interventional distributions?

Non-nested model comparison

- There are some "classical" type statistical tests for comparing non-nested models
- One of the best known is **Vuong's test**
- This essentially involves comparing the distributions of point-by-point data likelihoods for maximum likelihood-fitted models A and B
- You can look it up, but it is a finicky test for technical reasons and I'm not going to go into it!
- Rather, I want to introduce model comparison using **held-out data**

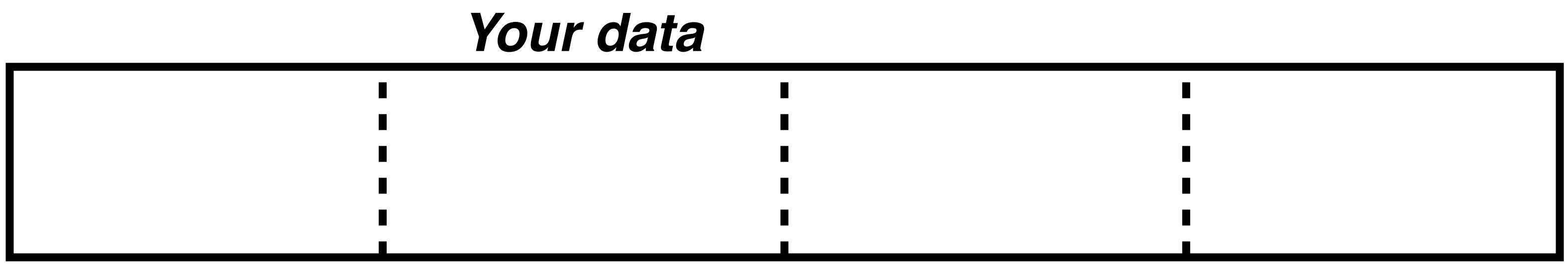
Model comparison on held-out data

- Train model A on the training set, calculate its performance on the test set
- Train model B on the training set, calculate its performance on the test set
- Which one does better?
- Advantages: **extremely general** – models don't need to be ML fit, "performance" can be defined in many ways
- But: how do we know when one model is "better

Model comparison on held-out data

- One option: you can take the paired performances of each model on each data point in your test set $\langle A(y_i), B(y_i) \rangle$ and do frequentist hypothesis testing on them as paired data
  - paired t-test
  - McNemar's test

Cross-validation



A caution regarding held-out evaluation

- Suppose you have **multi-level** (i.e. hierarchical) structure in your data–such that you might, for example, use a mixed-effects model to analyze it
- How do you do the train/test split for your data? What could go wrong?
- **Exercise:** create a simple multi-level dataset, show what could go wrong with a naive train/test split, and see if you can figure out how to avoid the problem