

Qualitative Model Comparisons

Q550: Models in Cognitive Science
Lecture 2



Comparing Models

Building a model to produce and explain a behavioral phenomenon is both challenging and rewarding

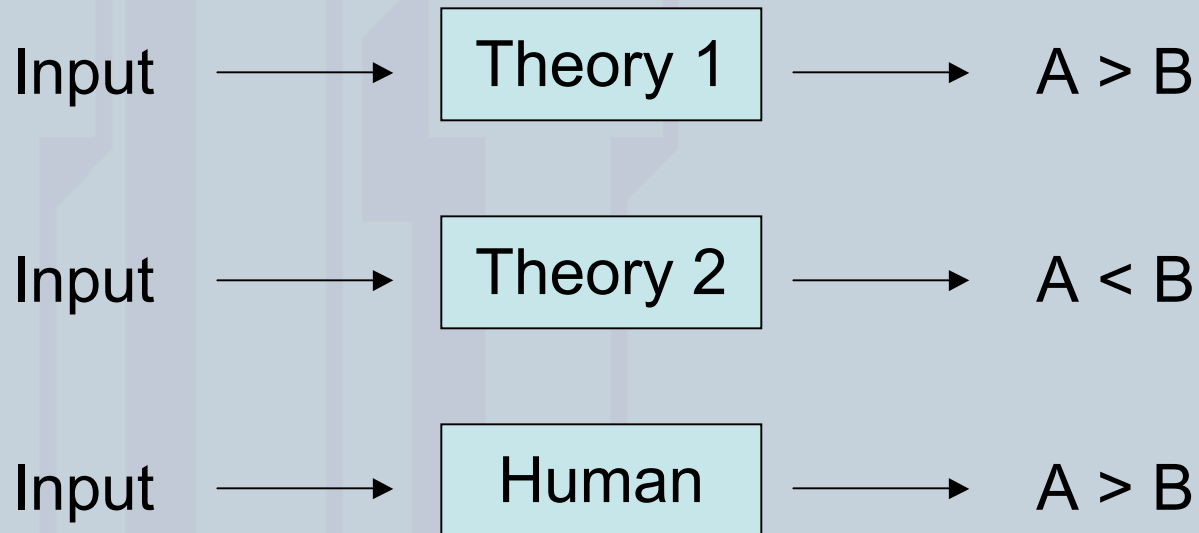
But, we further science by constraining between different models...formalize and reject models based on data

Ideally, we would love to do a **qualitative** comparison between models, that is, a comparison of formalized principles:

- free from ad hoc assumptions
- independent of specific model parameters

E.g., in Minerva, if we compare exemplar and prototype versions, the ordinal predictions should not differ at various levels of F

Our ideal situation:



Theory 1 is consistent w/ empirical data, whereas Theory 2 is not. Further, Theory 2 cannot produce the result for any set of parameters, whereas, Theory 1 produces the result for all parameter values and assumptions

...of course, this rarely occurs in practice

Comparing Models

...of course, this rarely occurs in practice

Often, we look at the proportion of total parameters that would produce the observed effect

OR: we find the best fitting set of parameters (estimated from the data) and compare the models qualitatively

Comparing Models

- Levels of model comparison:
 - Pure qualitative comparison independent of specific parameter values
 - Proportion of total parameter that produce the effect (signed-difference test)
 - Quantitative comparison to determine which model is most likely to have produced the data (assuming all models predict the ordinal differences)
 - Generalizing to other paradigms

Minimize task-specific parameters -- when these are changed, model should generalize to other tasks w/ the same general domain

A Qualitative Comparison in Minerva

We have an existence proof: It is unnecessary to posit both an abstract and episodic memory

But: do we have any reason to prefer one over the other?

Could we produce the data as well using a single prototype per category as well as multiple episodes?

➤ Let's keep everything the same in Minerva, but instead of storing multiple episodes, we'll create a composite for each category as they are learned, and will just store these three prototypes

What happens?

A prototype version of Minerva cannot explain the immediate benefit for old exemplars

It also cannot explain the interaction (crossover) between old exemplars and the prototype as a function of forgetting

➤ A prototype version of Minerva would not explain the ordinal differences in conditions b/c we would lose differential activation of stored episodes

(actually, I'm guessing here --> test this intuition by actually manipulating the code)

Serial Learning and Serial/Free Recall

Let's consider a second example: we're interested in learning items in a sequence and recalling them (in the order presented, or in any order)

Random series of words: 1 word per second. Remember the words and recall them after the presentation

You may be asked to recall them in the order presented, or in any order you chose

Free Recall: What are we doing?

Rehearsing? We'll call this a strength account

Each goes in once? Decay account

Associative chaining? (S-R theory)

Combinations?

Serial position curves: How would these accounts produce them?

If a model doesn't predict $R > P > M$ then reject it

If our models all make the same ordinal predictions, then we move onto qualitative model comparisons

Remember Gallileo's balls

TODAM, SAM, Perturbation theory (Neath Website)

Example from Busemeyer

A well-designed 2-D category learning experiment

Describes sophisticated connectionist versions of exemplar and prototype models--bivariate receptive fields, delta learning error backprop, etc.

The models have the same parameters:

α : learning rate

b : sensitivity for category choice

σ : discriminability for width of generalization gradients

c : recognition sensitivity

β : recognition response bias

Example from Busemeyer

XOR learning:

	S2=1	S2=10
S1=10	.94	.03
S1=1	.05	.98

He shows w/ a closed form solution that a prototype model **cannot** account for the crossover due to the independence axiom, and w/ simulation that the exemplar model **can** for a large proportion of its parameters

...or, we could just simulate both

Nosofsky & Zaki (2002)

Using the exemplar-based model that Bussemeyer describes, they show that the dissociation between amnesiacs and normals can be accounted for by parameter shifts

Another existence proof: no need to postulate multiple memory systems

We'll talk about things like backprop, d-prime, etc. in detail later.

Quantitative Model Comparison

It is often impossible to differentiate between models based on a qualitative comparison, so we decide which model is most likely to have generated the data

A model should make quantitative predictions that are more accurate than its competitors

The qualitative test must be based on an optimal selection of parameters...otherwise we could reject a perfectly good model simply by choosing a poor set of parameters

For each model

1. Find the best-fitting model parameters
2. Compare the quantitative accuracy of predictions based on these optimal parameters

Quantitative Model Comparison

This is complicated b/c model complexity needs to be considered in the comparison

- # of free parameters, model assumptions, etc.
- Nested and non-nested models
- Selection of the best model must satisfy both accuracy and parsimony

Nonlinear parameter estimation

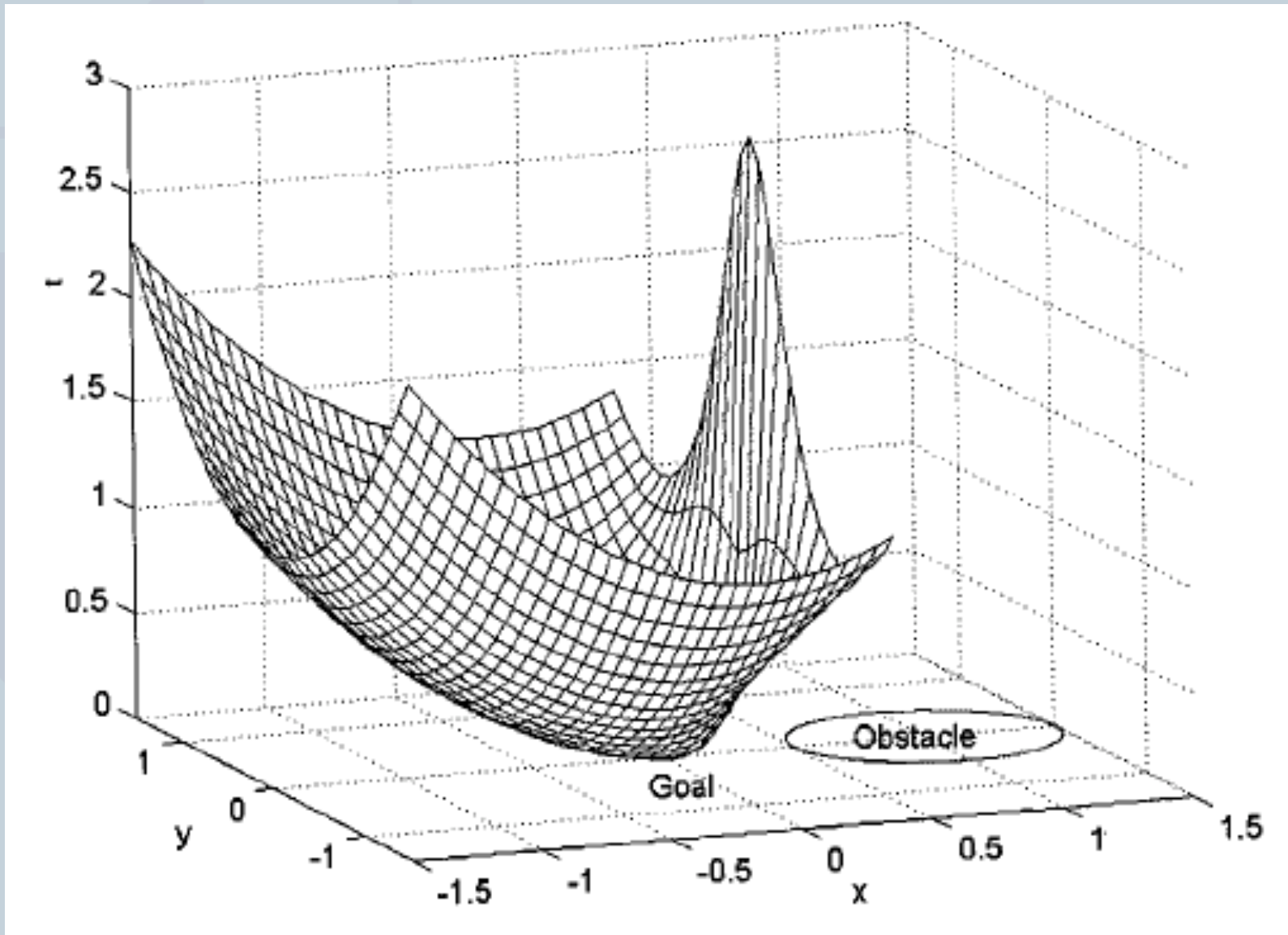
Objective functions to minimize/maximize (e.g., chi-square, least squares, log-likelihood)

Null and saturated models

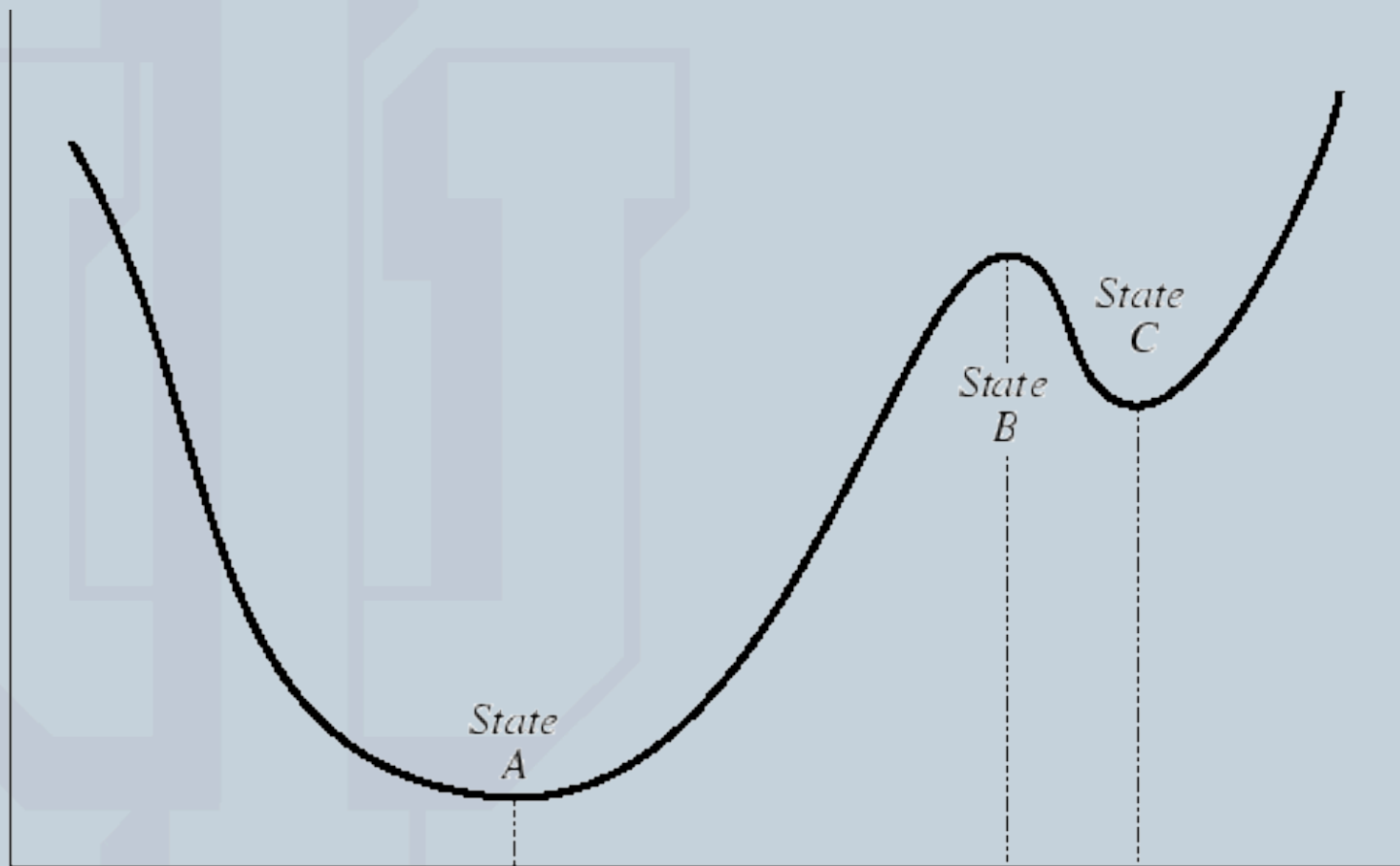
Linear vs. nonlinear models (special condition: mean of predictions from two different sets of parameters = the prediction produced by the average of the two sets; nonlinear makes parameter estimation more complicated)

Nonlinear parameter space estimation:

- grid search, hill climbing, steepest descent, simplex
- constraints on parameters
- flat/local minima problem



G (Gibbs Free Energy)



Any One Extensive Variable of the System

Model comparison techniques

When we have the optimal set of parameters for each model, we can compare their fit to data using a variety of techniques:

1. If nested: use classic chi-square
2. Bayesian model comparison (BIC: which of 2 models is most likely to be correct given the sample data); provides a measure of evidence for model A compared to model B, relative to # of free parameters
3. AIC method: Selects the model that generates a distribution closest to the true distribution --> select the most likely generating model
4. Minimum description length (equivalent to BIC for big N)

Fit is not everything

These techniques allow us to compare non-nested models that differ in complexity, but as Roberts and Pashler (2000) note, this may not be a good test of a model

Beware of overfitting

N-fold cross-validation: estimate parms from part of the data, and make predictions for the other part (bootstrap)

Generalization methods: estimate parms from one experimental condition, and make predictions for the other