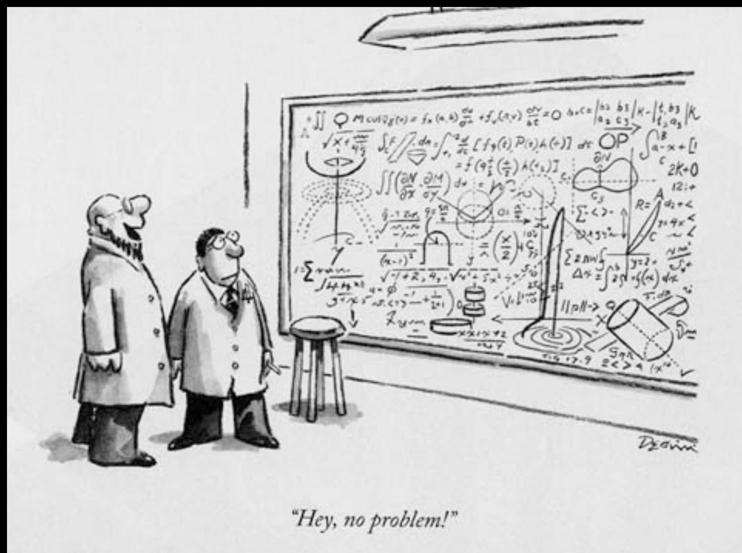
Model Fitting Workshop: Part 2

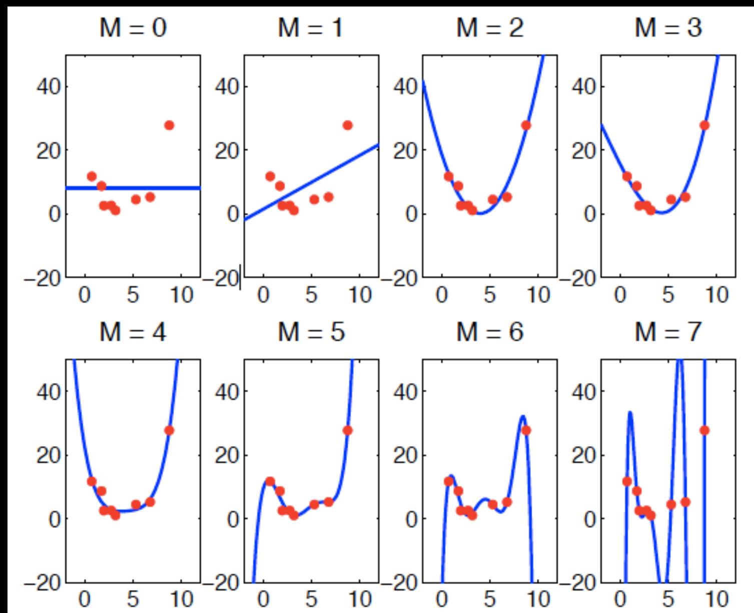


reminder: plan

- Snap introduction to Bayes rule
- Basic RL modeling & fitting the parameters of a model
- April: Model comparison

- background papers: Niv & Schoenbaum (2008) - dialogues on prediction errors
  Niv (2009) - RL in the brain
- primary paper: Daw (2011) - trial by trial data analysis using computational models
- newer: Wilson & Collins (2018) - Ten simple rules for the computational modeling of behavioral data

Act III: Which model is best?



Which model is best? Model comparison

P(Model | Data) = ?

wwBd?

Bayes rule

posterior → $P(model \,|\, data)$; likelihood → $P(data \,|\, model)$; prior → $P(model)$; marginalization (ignore!) → $P(data)$

$$P(model \,|\, data) = \frac{P(data \,|\, model) \cdot P(model)}{P(data)}$$

Which model is best? Model comparison

P(Model | Data) = ?

wwBd?

Comparing two models:

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1) \cdot P(M_1)}{P(D|M_2) \cdot P(M_2)}$$

Bayes factor

We prefer simple models

"Pluralitas non est ponenda sine necessitate"
Plurality should not be posited without necessity – William of Ockham (1349)

we should go for the simplest model that explains the data

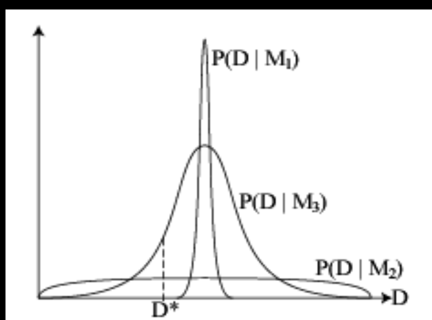
Comparing two models:

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1)}{P(D|M_2)} \cdot \frac{P(M_1)}{P(M_2)}$$

Bayes factor

automatic Occam's razor: simple models tend to make precise predictions

can put in preference for simple models here, but don't need to...



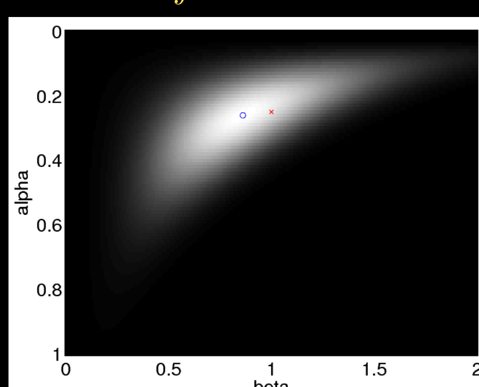
Which model is best? Model comparison

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1) \cdot P(M_1)}{P(D|M_2) \cdot P(M_2)}$$

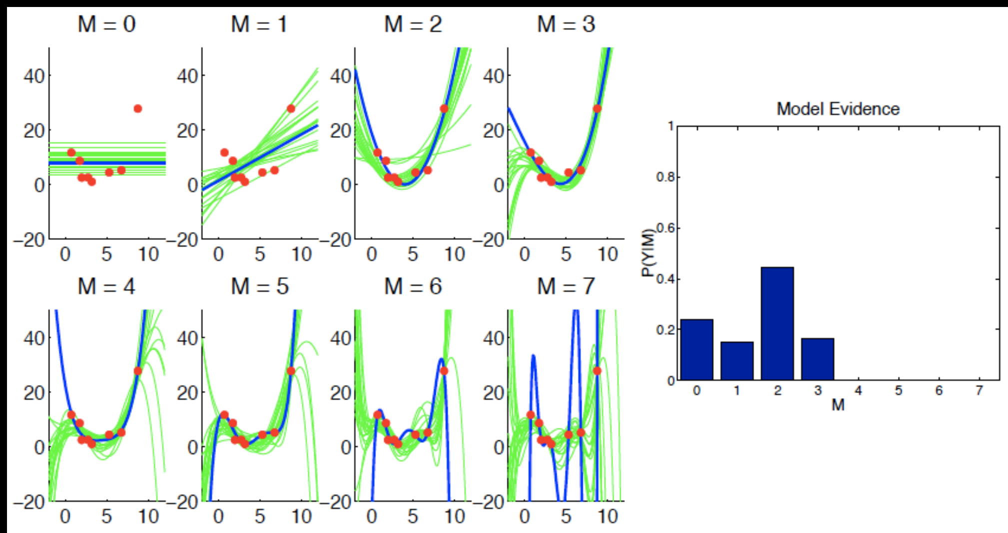
assuming uniform prior over models all we care about is P(*D*|*M*)

$$P(D|M) = \int d\theta P(D|M,\theta) \cdot P(\theta)$$

Bayesian evidence for model M (marginal likelihood)



Occam's razor at work: model evidence



$$P(D|M) = \int d\theta P(D|M,\theta) \cdot P(\theta)$$

Integrating over all settings of the parameters is, in most cases, too hard...

Approximate solutions:

1. sample posterior at many places to approximate integral and compute Bayes factor directly
2. Laplace approximation: make Gaussian approximation around MAP parameter estimate $\hat{\theta}$

$$P(D|M) = \int d\theta P(D|M,\theta) \cdot P(\theta)$$

Integrating over all settings of the parameters is, in most cases, too hard...

Approximate solutions:

1. sample posterior at many places to approximate integral and compute Bayes factor directly
2. Laplace approximation: make Gaussian approximation around MAP parameter estimate $\hat{\theta}$
3. BIC approximation: if N is large, can drop whatever doesn't grow with the amount of data

$$lnP(D|M) \approx lnP(D|\hat{\theta}, M) - \frac{d}{2} ln(N)$$

BIC approximation

$$lnP(D|M) \approx lnP(D|\hat{\theta}, M) - \frac{d}{2} ln(N)$$

- For each model, compute the maximum log likelihood and add to that a penalty that depends on d (# of parameters).
- Compare the results between the models

- Advantages: easy to compute; can use ML rather than MAP estimate (i.e., don't care about prior on parameters)
- Disadvantage: can overpenalize as it is hard to determine d (only identifiable parameters) and N (only samples used to fit parameters; note: not necessarily same for different parameters)

$$P(D|M) = \int d\theta P(D|M,\theta) \cdot P(\theta)$$

Integrating over all settings of the parameters is, in most cases, too hard...

Approximate solutions:

1. sample posterior at many places to approximate integral and compute Bayes factor directly
2. Laplace approximation: make Gaussian approximation around MAP parameter estimate $\hat{\theta}$
3. BIC approximation: if N is large, can drop whatever doesn't grow with the amount of data
4. (W)AIC approximation: ranks models by penalizing by # parms
5. Likelihood ratio test (for nested models)

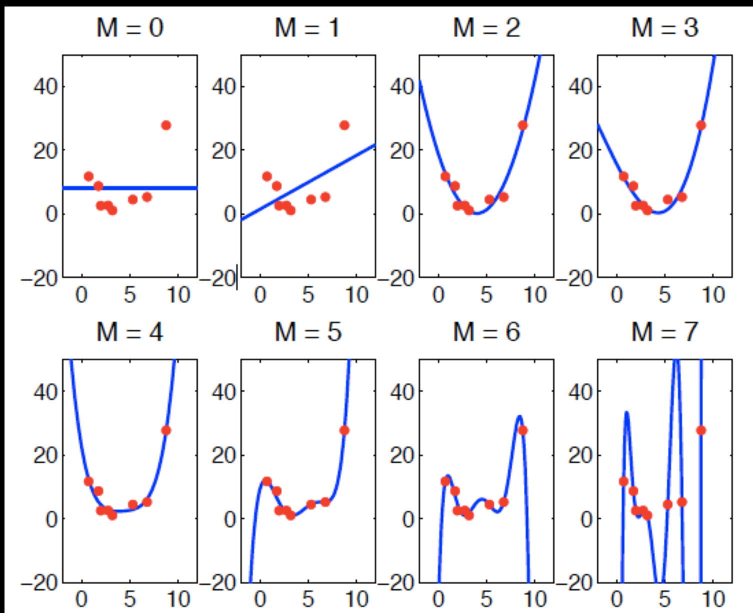
Likelihood ratio test

For nested models (one is a special case of the other)

Compares hypothesis H_1 to one where some parameters are fixed, H_0

Statistical test on the likelihood differences: compare 2*difference in log likelihood (ML) to χ^2 statistic with df=#additional parameters

Holy grail: cross validation



Fit models on training set and validate fit on hold-out set.

Problem: often hard to find i.i.d. data sets in a learning setting

Summary

Learning models are detailed hypotheses about trial-by-trial overt and covert variables

trial-by-trial model fitting lets us test these hypotheses
...and compare alternatives

special premium on detailed model fitting when considering learning data:
non-stationary, can't use traditional averaging techniques

Modeling trial-by-trial data: last thoughts

We've discussed modeling each subject separately (each have different parameters). Alternatives?

- Population level fit: assumes subjects all have the same parameter
- Hierarchical fit: learn distribution over parameters from population (use population as constraints on individual fits)

And: does each subject have to have one parameter value throughout the experiment?

Examples to the contrary?

What can you do about this?

Some more practical tips:

BEFORE running an experiment:

- Simulate data from your favorite model
- Fit the model to the data: do you get the parameters you put in the simulation? If not, what can you do?
- Compare to other models: do the data *identify* your model correctly? If not, what can you do?

- background papers: Niv & Schoenbaum (2008) - dialogues on prediction errors
  Niv (2009) - RL in the brain
- primary paper: Daw (2011) - trial by trial data analysis using computational models
- newer: Wilson & Collins (2018) - Ten simple rules for the computational modeling of behavioral data

The only stupid question is the one you did not ask

-Rich Sutton