

Computational Cognitive Science

Lecture 7: Simplicity and parameters

Chris Lucas

School of Informatics
University of Edinburgh
clucas2@inf.ed.ac.uk

10 October, 2017

Reading:

- ▶ Chapters 5 and 6 of L&F.;
- ▶ Ockham's razor and Bayesian analysis (Jefferys and Berger, 1992) ¹.

¹<http://www.jstor.org/stable/29774559>

Last time

We discussed model comparison, focusing on:

- ▶ Model fit, using data likelihood given parameters $\hat{\theta}$ obtained by MLE.
- ▶ Model complexity, using numbers of parameters.

Under this approach, we can compare models using straightforward tests and criteria: likelihood ratio test, BIC, AIC.

Today we'll take a longer look at model complexity.

Simplicity

“Simple” can mean many things, including:

- ▶ Easy to understand or compute
- ▶ Have few “moving parts” or degrees of freedom

LLR, BIC, and AIC approximate the latter by counting parameters.



WHEN PEOPLE ASK FOR STEP-BY-STEP DIRECTIONS, I WORRY THAT THERE WILL BE TOO MANY STEPS TO REMEMBER, SO I TRY TO PUT THEM IN MINIMAL FORM.

<https://xkcd.com/1155/>

Complexity isn't just $|\theta|$

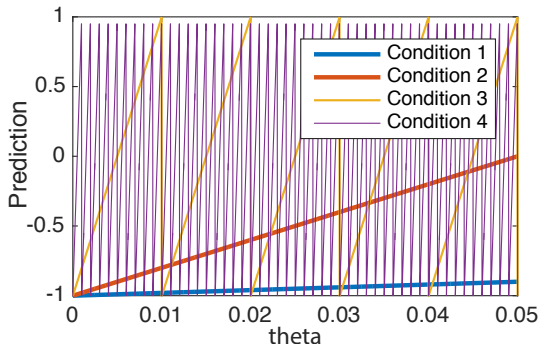
If we could compute $P(\mathbf{y}|M)$, we could estimate predictive accuracy and likelihood ratios directly.

Counting parameters is a way to compensate for using $P(\mathbf{y}|\hat{\theta}, M)$ rather than $P(\mathbf{y}|M)$.

Parameter counting can sometimes lead us astray. AIC and BIC make assumptions that are often violated. Sometimes this is acceptable; in other cases, less so.

Example 1: Hidden flexibility

Suppose we elicit judgments in $[-1, 1]$ from participants in four experimental conditions. Our model makes the predictions below.



Compare to a model that says conditions 1 and 2 should be identical (θ_1) and conditions 3 and 4 should be slower ($\theta_1 + \theta_2$)

Example 2: Adding parameters without increasing flexibility

Suppose we want to estimate decay-rate differences across words in our phonological loop model.

- ▶ Model 1: Every word has an idiosyncratic decay rate: $|\mathbf{w}|$ parameters.
- ▶ Model 2: There's a frequency effect, plus the above: $|\mathbf{w}| + 1$ parameters.

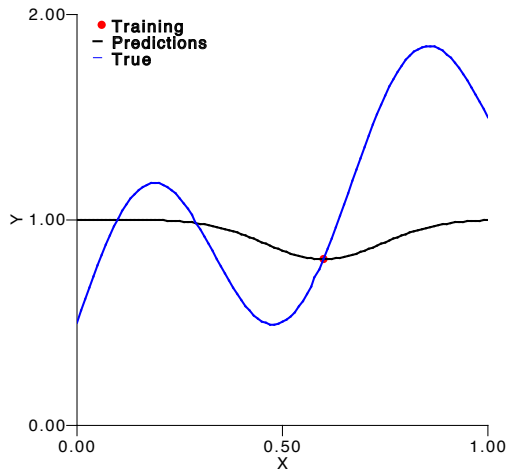
Model 2 has more parameters, but we can bundle frequency effects with the other $|\mathbf{w}|$ parameters.

Bonus questions:

- ▶ Why is MLE problematic here?
- ▶ Why might we prefer model 2?

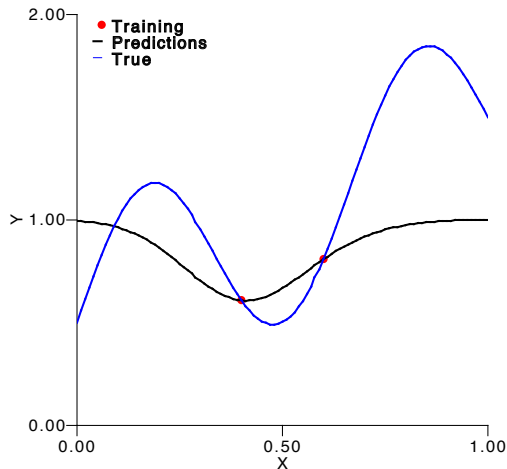
Example 3: Nonparametric models

Some models can fit complex patterns in data with a small number of parameters (3).



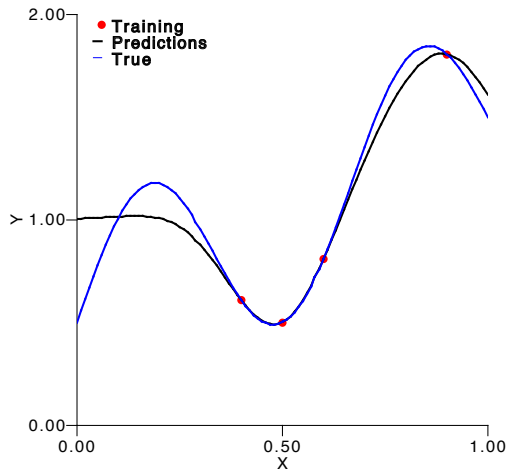
Example 3: Nonparametric models

Some models can fit complex patterns in data with a small number of parameters (3).



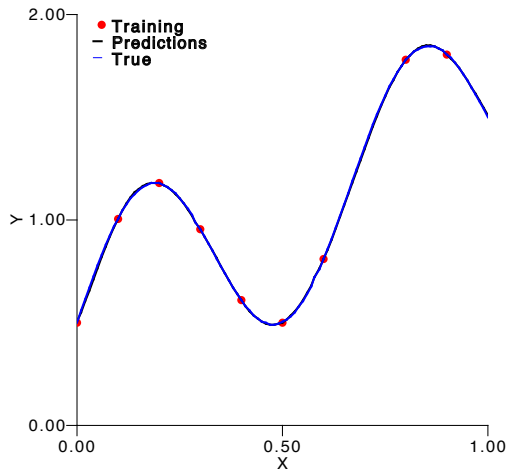
Example 3: Nonparametric models

Some models can fit complex patterns in data with a small number of parameters (3).



Example 3: Nonparametric models

Some models can fit complex patterns in data with a small number of parameters (3).



Flexibility and priors

Flexible models assigns non-negligible likelihood to a wide variety of outcomes, including many that never come to pass.

If they can fit the data we have seen, we should hesitate to conclude they generalize well.

This doesn't imply we should avoid flexible models. Having *priors* over parameters is a way to balance flexibility and generalization. See the Jefferys and Berger (1992) reading for a discussion of the *Bayesian Occam's razor*².

²Also spelled "Ockam's razor".

Model comparison – What to do?

- ▶ In a perfect world: Use held-out test data. However:
 - ▶ Data can be expensive to collect
 - ▶ How to make such a comparison fair?
- ▶ Parameter-counting works reasonably well for many models
- ▶ Cross-validation (leave-one-out, k-fold)
 - ▶ Pros: More general and robust than AIC, BIC
 - ▶ Cons: More computationally expensive
 - ▶ Take care what you leave out

Quiz 1

Which of the following models of planetary motion provides the simplest explanation of the observed trajectories of the planets?

- ▶ Keplerian model
- ▶ Copernican model
- ▶ Ptolomaic model

Quiz 1

Which of the following models of planetary motion provides the simplest explanation of the observed trajectories of the planets?

- ▶ Copernican model

Quiz 1

Which of the following models of planetary motion provides the best fit with the observed trajectories?

- ▶ Copernican model
- ▶ Keplerian model
- ▶ Ptolomaic model

Quiz 1

Which of the following models of planetary motion provides the best fit with the observed trajectories?

- ▶ Keplerian model

Quiz 1

Which of the following is not a way of quantifying model fit?

- ▶ χ^2 statistic
- ▶ Negative log likelihood
- ▶ Mean squared error
- ▶ Acceptance function

Quiz 1

Which of the following is not a way of quantifying model fit?

- ▶ Acceptance function

Quiz 1

What is the word length effect?

- ▶ Longer words are less likely to be recalled correctly
- ▶ Longer words are more likely to be recalled correctly
- ▶ Longer word lists lead to higher error rates
- ▶ Longer word lists take longer to recall in their entirety
- ▶ Longer words take longer to be recalled, individually
- ▶ Longer words take longer to be mentally rehearsed

Quiz 1

What is the word length effect?

- ▶ Longer words are less likely to be recalled correctly

Quiz 1

Which of the following best describes an error surface?

- ▶ a plot of the error dependent on the parameters of a model
- ▶ the quantity a parameter estimation algorithm tries to minimize
- ▶ a visualization of the local minima of model
- ▶ a visualization of the global minima of model

Quiz 1

Which of the following best describes an error surface?

- ▶ a plot of the error dependent on the parameters of a model

Quiz 1

How does simulated annealing deal with local minima?

- ▶ when it reaches a local minimum, it jumps to a random part of the error surface
- ▶ it uses the annealing schedule to control the search space
- ▶ it investigates multiple local minima at the same time and then returns the best one
- ▶ it allows updates to the parameters that increase the error

Quiz 1

How does simulated annealing deal with local minima?

- ▶ it allows updates to the parameters that increase the error

Quiz 1

In simulated annealing as described in lecture, which of the following must be tuned? (More than one may apply.)

- ▶ candidate function
- ▶ learning rate
- ▶ momentum
- ▶ annealing schedule

Quiz 1

In simulated annealing as described in lecture, which of the following must be tuned? (More than one may apply.)

- ▶ candidate function

- ▶ annealing schedule

Quiz 1

Which of the following conditions suggest that it would be unwise to use simulated annealing as described in lecture?

- ▶ there are many local optima
- ▶ there is an analytic solution for minimizing the error function
- ▶ the first derivative of the error function is easy to compute
- ▶ the second derivative of the error function is difficult to compute
- ▶ there are many (e.g., > 5000) parameters to optimize

Quiz 1

Which of the following conditions suggest that it would be unwise to use simulated annealing as described in lecture?

- ▶ there is an analytic solution for minimizing the error function
- ▶ the first derivative of the error function is easy to compute

- ▶ there are many (e.g., > 5000) parameters to optimize

Quiz 1

Which of the follow statements are true of maximum likelihood estimation?

- ▶ If we can compute the MLE for $g(\theta)$, we can always use that to find the value of θ that maximizes the likelihood function
- ▶ For a well-specified model, MLE can yield an arbitrarily precise estimate for a parameter given a data set with a bounded size
- ▶ MLE yields the lowest-possible variance in parameter estimates
- ▶ MLE finds the parameters that give the highest probability (or probability density) for the data
- ▶ MLE find the values of the parameters that have the highest probability (or probability density) given the data

Quiz 1

Which of the follow statements are true of maximum likelihood estimation?

- ▶ MLE finds the parameters that give the highest probability (or probability density) for the data

Quiz 1

Suppose you work at a factory making trick coins that are biased to flip heads or tails at different rates than 50 percent of the time. Your job is to understand the biases of coins that are coming out of the coin-making machine. For a given coin, you flip it twice and compute the MLE of its probability of coming up heads on a given flip.



- ▶ As you flip more coins, your distribution of MLEs will look more and more like a normal distribution.
- ▶ You can use a Binomial probability mass function to obtain the MLE for each coin's bias.
- ▶ If the coins are manufactured using a common process, a good way to estimate the typical coin's bias is to take the mode of the maximum likelihood estimates you've obtained.
- ▶ Testing a larger number of coins using the same experiment is a good way to learn about individual differences between the coins.

Quiz 1

Suppose you work at a factory making trick coins that are biased to flip heads or tails at different rates than 50 percent of the time. Your job is to understand the biases of coins that are coming out of the coin-making machine. For a given coin, you flip it twice and compute the MLE of its probability of coming up heads on a given flip.

- ▶ You can use a Binomial probability mass function to obtain the MLE for each coin's bias.

References

-  Jefferys, W. H. & Berger, J. O. (1992). Ockham's razor and bayesian analysis. *American Scientist*, 80(1), 64–72.
-  Lewandowsky, S. & Farrell, S. (2011). *Computational modeling in cognition: Principles and practice*. Thousand Oaks, CA: Sage.