

Computational Cognitive Science

Lecture 6: Model Comparison; Generalization

Chris Lucas
(Slides adapted from Frank Keller's)

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

6 October, 2017

Model Comparison

Tests for comparing models

Likelihood Ratio Test

Akaike's Information Criterion

Bayesian Information Criterion

Generalization and Overfitting

Generalization and Overfitting

Crossvalidation

Reading: Chapters 5 and 6 of L&F.

Distinguishing between Models

Recall that in the likelihood function, the parameters θ determine the likelihood of the data \mathbf{y} , but only given a particular model M .

Last lecture, we assumed there was only one model to consider, so we omitted M from our equations.

Different models assign different likelihoods to the same data. This allows us to do *model comparison*, i.e., to distinguish models based on how well they fit the data.

Model comparison

If we suppose that:

- ▶ Theories are best expressed in precise terms (i.e., as models)
- ▶ Any given model is probably incomplete (or wrong)

then scientific progress is a matter of find better models than what came before.

What makes a model “better”?

- ▶ Better at explaining the data we have:
 - ▶ fit
 - ▶ being understandable
- ▶ Simplicity and plausibility
- ▶ Better at prediction and generalization
- ▶ Being likely to be true

What makes a model “better”?

- ▶ Better at explaining the data we have:
 - ▶ fit: $P(\mathbf{y}|M)$
 - ▶ being understandable (e.g., succinct description)
- ▶ Simplicity and plausibility: $P(M)$
- ▶ Better at prediction and generalization: $P(\mathbf{y}_{\text{new}}|M, \mathbf{y})$
- ▶ Being likely to be true: $P(M|\mathbf{y}) \propto P(M)P(\mathbf{y}|M)$

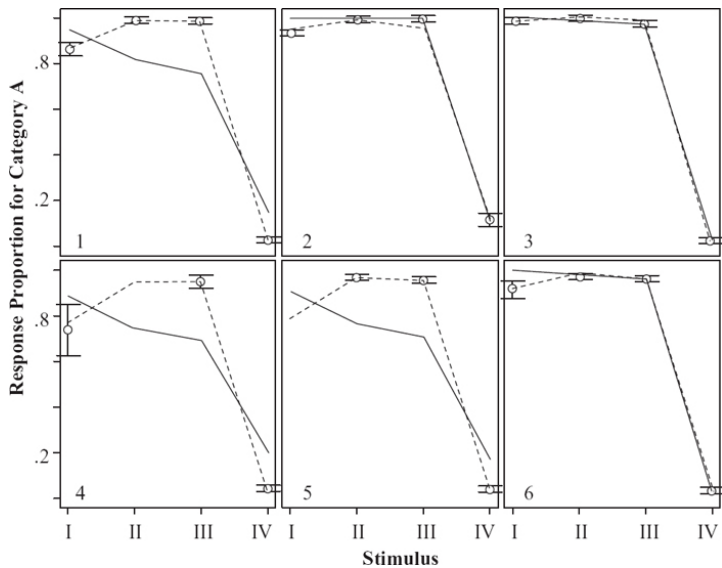
How can we compare models?

Sometimes model fit is visible with the naked eye. Compare:

- ▶ Generalized Context Model (Nosofsky, 1986), see lecture 1;
- ▶ General Recognition Theory (Ashby and Townsend, 1986)

on an experimental data set with four conditions I, II, III, and IV.

Distinguishing between Models



Distinguishing between Models

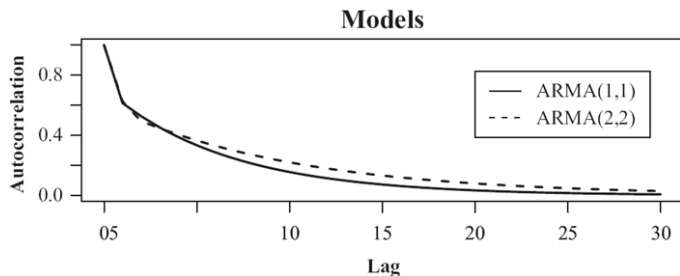
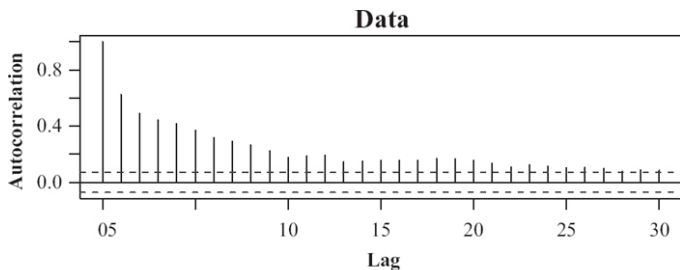
In other cases, it's difficult to distinguish between models visually.

Consider the Auto-regressive Moving Average (ARMA) model, fit on data from a time estimation task:

- ▶ participants have to repeatedly press a button after a given time interval has passed (here: 1 second);
- ▶ the different trials (button presses) are not independent, but correlated with each other to a certain extent;
- ▶ compute the auto-correlation function, a measure of how correlated a time series is with itself.

Use this to compare ARMA(1,1), which only considers the previous trial, with ARMA(2,2), which considers the previous two trials.

Distinguishing between Models



What do we do in this case? Even if one model looks better than another,

$$P(\mathbf{y}|M) \neq P(\mathbf{y}|\hat{\theta}, M)$$

Complicated models – ones with lots of flexibility – give an inflated estimate of fit when we optimize their parameters.

Ideally, we would compute the *marginal likelihood*

$$P(\mathbf{y}|M) = \int_{\theta} P(\mathbf{y}|\theta, M)P(\theta|M)d\theta$$

What do we do in this case? Even if one model looks better than another,

$$P(\mathbf{y}|M) \neq P(\mathbf{y}|\hat{\theta}, M)$$

Complicated models – ones with lots of flexibility – give an inflated estimate of fit when we optimize their parameters.

Ideally, we would compute the *marginal likelihood*

$$P(\mathbf{y}|M) = \int_{\theta} P(\mathbf{y}|\theta, M)P(\theta|M)d\theta$$

In practice, it can be difficult to choose $P(\theta|M)$, and the integral may be impossible to compute.

Instead, it's standard to use simple tests that use the MLE parameters, but apply a penalty based on complexity.

Tests for comparing models

These tests rely on two criteria:

- ▶ *model fit*: how well does the model capture the data we are trying to model?
- ▶ *simplicity*: how complicated is the model, i.e., how many parameters does it require to achieve the fit?

By balancing fit against simplicity, we can compensate for the fact that models with more parameters tend to fit the data better.

Sometimes we are dealing with *nested models*, i.e., one model is reduced version of the other. For instance ARMA(1,1) is a version of ARMA(2,2) that doesn't consider the second-to-last trial.

Likelihood Ratio Test

We can compare *nested models* using the likelihood ratio test. It determines whether we are justified in making the model more complicated by adding additional parameters.

The likelihood ratio test uses the χ^2 distribution to compare the likelihood of a *specific version* of the model with a *general version*:

$$\chi^2 \approx -2 \ln L_{\text{specific}} - (-2 \ln L_{\text{general}})$$

The general version of the model is identical to the specific version, but has K additional parameters.

If a χ^2 test with K degrees of freedom is significant, then we are justified in adding the K parameters to obtain a better model fit.

Akaike's Information Criterion

The likelihood ratio test can only be applied to nested models. For non-nested models, *information criteria* can be used.

Akaike's Information Criterion (AIC), where K is the number of parameters of model M :

$$AIC = -2 \ln L(\theta|y, M) + 2K$$

The AIC measures the distance between the probability distribution specified by model M and the (hypothetical) true distribution. (This can be formalized using the Kullback-Leibler divergence.)

Bayesian Information Criterion

An alternative to AIC is the *Bayesian Information Criterion* (BIC), with N the number of data points used to calculate the likelihood:

$$BIC = -2 \ln L(\theta|y, M) + K \ln N$$

Based on the BIC, we can compare two models M_1 and M_2 using the *Bayes factor*:

$$B = \exp\left(-\frac{1}{2}\Delta BIC\right)$$

where $\Delta BIC = BIC(M_1) - BIC(M_2)$.

Formally, the Bayes factor measures the ratio of the posterior probabilities of the two models, $P(M_1|y)/P(M_2|y)$.

Information Criteria

Example (see also tutorial 2):

- ▶ data: reaction times $y = [3 \ 4 \ 4 \ 4 \ 4 \ 5 \ 5 \ 6 \ 6 \ 7 \ 8 \ 9]$ ($N = 12$);
- ▶ Model 1: ex-Gaussian with free parameters μ , τ , and fixed σ ($K = 2$);
- ▶ Model 2: ex-Gaussian with free parameters μ , τ , and σ ($K = 3$).

	Model 1	Model 2	Comparison
$\ln L$	-23.21	-22.70	$\chi^2(1) = 1.02$ (not sign.)
AIC	50.42	51.40	$\Delta AIC = -0.98$
BIC	51.39	52.85	$B = 0.60$

Information Criteria

Choosing between AIC and BIC:

- ▶ AIC and BIC don't always select the same best model;
- ▶ the AIC is less suitable for nested models (which often only differ in only one parameter);
- ▶ the AIC puts more emphasis on model fit, and less emphasis on simplicity;
- ▶ the BIC puts more emphasis on simplicity (if $\ln N > K$);
- ▶ in general, the BIC makes a more complex trade-off: model fit L , number of parameters K , and size of the data set N .

Generalization and Overfitting

If we use MLE to estimate parameters for a model, it may fit the data we have, but generalize poorly to new data.

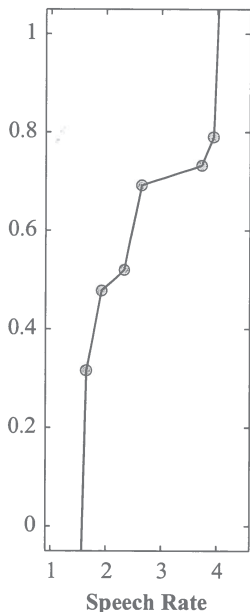
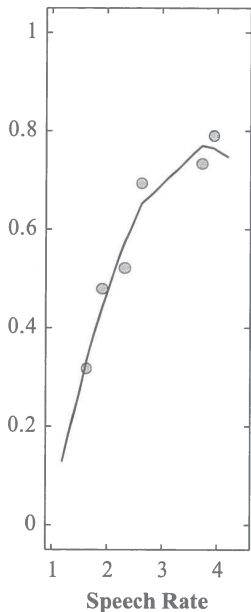
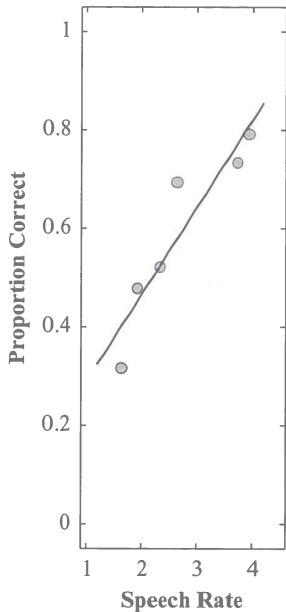
For example, a model that maximizes goodness of fit by treating noise, e.g., individual idiosyncrasies, as part of a generalizable pattern.

This is called *overfitting*.

Example: Fit the data on the effect of speech rate (word length) on recall using three different models:

- ▶ linear regression;
- ▶ third-order polynomial;
- ▶ fifth-order polynomial.

Generalization and Overfitting



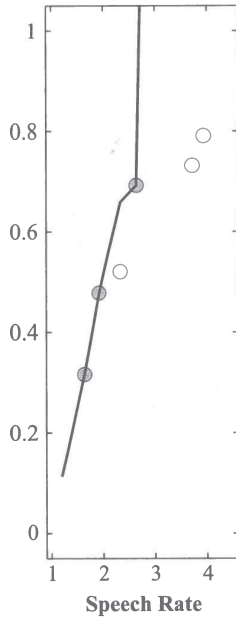
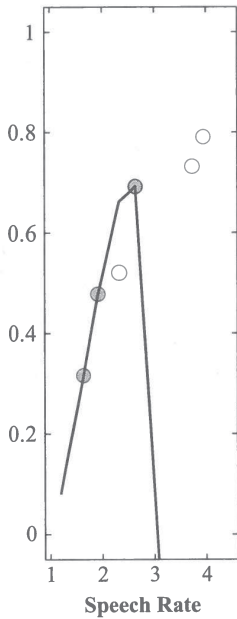
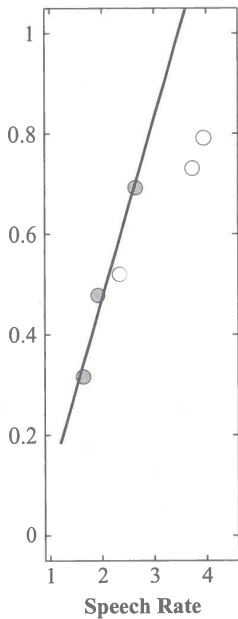
Generalization and Overfitting

A simple method for testing whether a model is able to generalize is to use *held-out data*.

Randomly split the data set used to fit the model into a *training set* (calibration sample) and an unseen *test set* (validation sample):

- ▶ train the model (determine its parameters) on training set;
- ▶ apply model to both the training set and the test set; compute model fit on both sets;
- ▶ the difference between the fit on the training and test data indicates the model's ability to generalize.

Generalization and Overfitting



Crossvalidation

Potential problem: *data set too small* for a split into training and test set (often the case for experimental data).

Solution: *k-fold crossvalidation* maximizes the use of the data:

- ▶ divide data randomly into k folds (subsets) of equal size;
- ▶ train the model on $k - 1$ folds, use the other fold for testing;
- ▶ repeat this k times so that all folds are used for testing;
- ▶ compute the average model fit on the k test sets.

This effectively uses all the data for both training and testing. Often $k = 10$ is used. *Leave-one-out*: $k = N$ (size of data).

Crossvalidation

Example: Apply 5-fold crossvalidation to a data set with 20 items.

test set training set

i_1, i_2, i_3, i_4
i_5, i_6, i_7, i_8
$i_9, i_{10}, i_{11}, i_{12}$
$i_{13}, i_{14}, i_{15}, i_{16}$
$i_{17}, i_{18}, i_{19}, i_{20}$

$$\ln L = -12$$

Crossvalidation

Example: Apply 5-fold crossvalidation to a data set with 20 items.

test set training set

i_1, i_2, i_3, i_4
i_5, i_6, i_7, i_8
$i_9, i_{10}, i_{11}, i_{12}$
$i_{13}, i_{14}, i_{15}, i_{16}$
$i_{17}, i_{18}, i_{19}, i_{20}$

$$\ln L = -10$$

Crossvalidation

Example: Apply 5-fold crossvalidation to a data set with 20 items.

test set training set

i_1, i_2, i_3, i_4
i_5, i_6, i_7, i_8
$i_9, i_{10}, i_{11}, i_{12}$
$i_{13}, i_{14}, i_{15}, i_{16}$
$i_{17}, i_{18}, i_{19}, i_{20}$

$$\ln L = -8$$

Crossvalidation

Example: Apply 5-fold crossvalidation to a data set with 20 items.

test set training set

i_1, i_2, i_3, i_4
i_5, i_6, i_7, i_8
$i_9, i_{10}, i_{11}, i_{12}$
$i_{13}, i_{14}, i_{15}, i_{16}$
$i_{17}, i_{18}, i_{19}, i_{20}$

$$\ln L = -13$$

Crossvalidation

Example: Apply 5-fold crossvalidation to a data set with 20 items.

test set training set

i_1, i_2, i_3, i_4
i_5, i_6, i_7, i_8
$i_9, i_{10}, i_{11}, i_{12}$
$i_{13}, i_{14}, i_{15}, i_{16}$
$i_{17}, i_{18}, i_{19}, i_{20}$




$$\ln L = -15$$

$$\text{average } \ln L = -11.6$$

Summary

- ▶ When comparing models, we can use simple criteria:
 - ▶ model fit: how well does the model account for the data?
 - ▶ simplicity: how many parameters does the model use to achieve this fit?
- ▶ for nested models, we can use the likelihood ratio test (essentially a χ^2 test);
- ▶ for non-nested models, we can use information criteria (AIC or BIC), which trade off fit against number of parameters;
- ▶ sometimes a model overfits the data and doesn't generalize well to new data;
- ▶ we can test for overfitting by using held-out data or by performing crossvalidation.

References

-  Ashby, F. G. & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, 93, 154–179.
-  Lewandowsky, S. & Farrell, S. (2011). *Computational modeling in cognition: Principles and practice*. Thousand Oaks, CA: Sage.
-  Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1), 39–57.