Computational Cognitive Science

Lecture 8: Model Comparison

Frank Keller

15 October 2026

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

Slide credits: Chris Lucas, Benjamin Peters

Overview

Why Model Comparison?

Bias and Variance

Marginal Likelihood

Alternatives to Marginal Likelihood

- Held-out Data and Cross-validation
- Likelihood Ratio
- AIC and BIC

Reading:

Chapters 10 and 11.2.6 of Farrell and Lewandowsky (2018).
Jefferys and Berger (1992).

Why Model Comparison?

Model comparison

We have discussed estimating parameters conditional on a model.

- That may be all we need, if we can capture different theories as parameter choices in a single model
- In practice, we may want to compare qualitatively different models

How do we choose between models?

Criteria for choosing models

We prefer models that are

1. Predictively useful
2. Compatible with our data
3. Likely to be correct, or closer to a correct model
4. Provide good explanations
5. Simple and elegant (see geocentric vs. heliocentric models from lecture 1)

Two models of perceived intensity

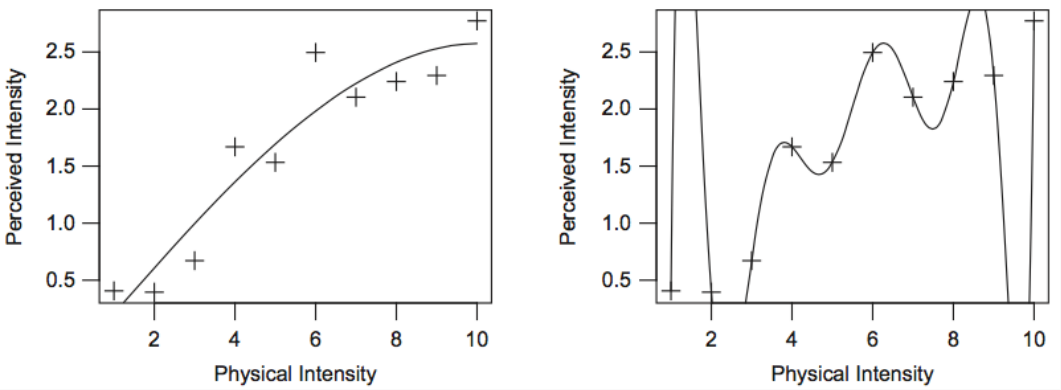
$\mathcal{M}_1$: Perceived intensity is a **2nd** order polynomial function of physical intensity

$\mathcal{M}_2$: Perceived intensity is a **9th** order polynomial function of physical intensity

(Ignore the fact that we could treat these as a single model with suitable priors over parameters.)

Two models of perceived intensity

Both models, with MLE fits:



Which one should we prefer?

Two models

Is the complex polynomial going to give good predictions?

- $p(y_{K+1}|\mathbf{y}, \mathcal{M}_2)$

Is the complex polynomial compatible with our data?

- $p(\mathbf{y}|\mathcal{M}_2)$

Is the complex polynomial the right generative model?

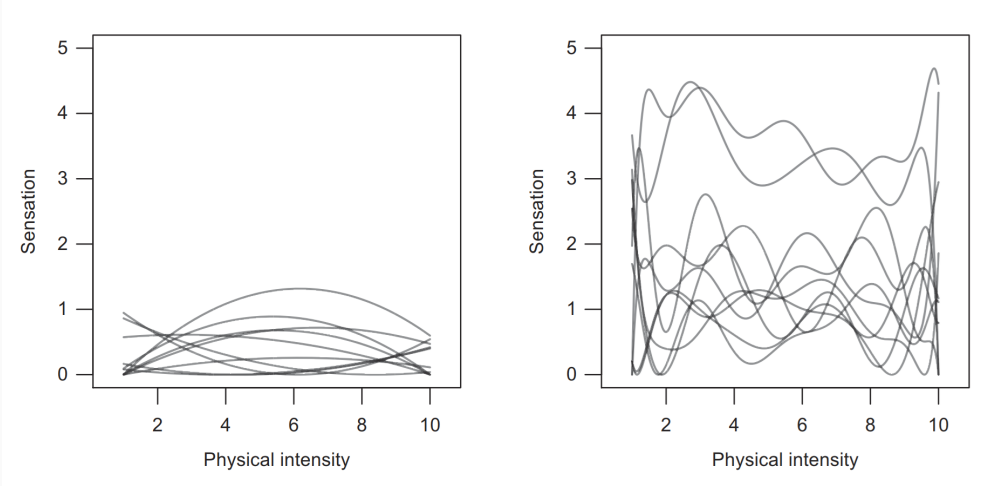
- $p(\mathcal{M}_2|\mathbf{y})$

Better fit to the observed data does not imply a better model of the generating process.

A worse fit can be preferable if it captures the signal rather than the noise!

Predictive accuracy

Predictions from a polynomial function of order 2 (left panel) and order 10 (right panel), with randomly sampled parameter values:



Predictive accuracy

More flexible models can fit a wider range of patterns.

But greater flexibility has a cost:

- simple models may underfit: their predictions are systematically wrong;
- very flexible models may overfit: their predictions depend too strongly on the particular sample of data.

How do we balance these two sources of prediction error?

This is the bias-variance dilemma, see also Geman et al. (1992).

Bias and Variance

Bias and variance

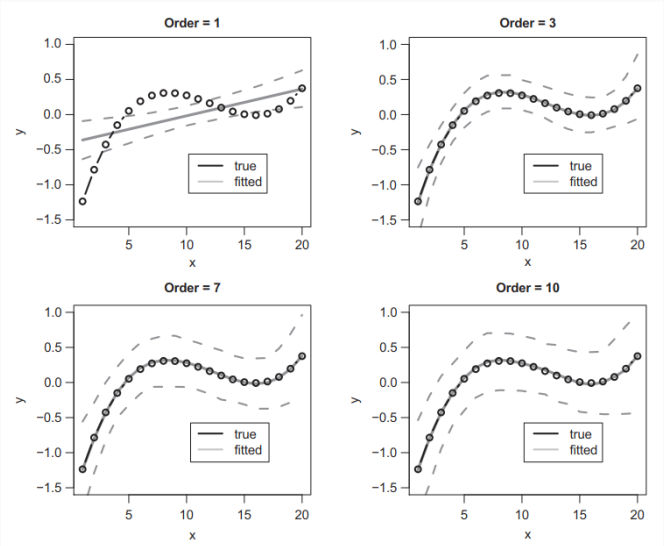
Two important contributions to expected prediction error are:

- Error due to bias: The difference between the expected predictions of the model (under all possible data) and the real mean
- Error due to variance: How much the model's predictions vary as a function of the specific data it has been given

$$\text{Expected prediction error} = \text{Bias}^2 + \text{Variance} + \text{Noise}$$

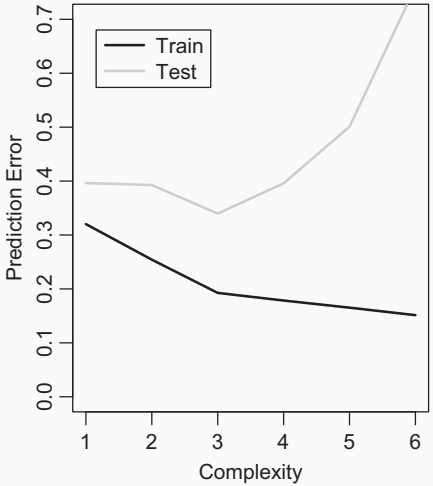
Bias and variance

Bias and variance are both related to model flexibility.



Bias and variance

More flexibility almost always helps training fit. What we really care about is performance on new data.



Marginal Likelihood

Two models

For probabilistic models, predictive accuracy relates to other desiderata:

Is the complex polynomial compatible with our data?

- $p(\mathbf{y}|\mathcal{M}_2)$

Is the complex polynomial the right generative model?

- $p(\mathcal{M}_2|\mathbf{y}) \propto p(\mathbf{y}|\mathcal{M}_2)P(\mathcal{M}_2)$

To answer these questions, we need the marginal likelihood of our data.

Two models

Marginal likelihood:

$$p(\mathbf{y}|\mathcal{M}) = \int_{\boldsymbol{\theta}} p(\mathbf{y}|\mathcal{M}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta}$$
$$\neq p(\mathbf{y}|\mathcal{M}, \hat{\boldsymbol{\theta}})$$

- Flexible models can accommodate a wide variety of patterns
- If those patterns are not present in our data, they're bad models

Marginal likelihood rewards models that predicted the observed data before seeing it.

Flexibility and overfitting: Likelihood

What if we specify $p(\boldsymbol{\theta})$ at the start, and compute $p(\mathbf{y}|\mathcal{M})$?

That's an excellent solution, when it's viable.

However:

- We must choose priors carefully
- Integrating over $\boldsymbol{\theta}$ is often expensive or impossible

Time for a short quiz on WooClap!



Alternatives to Marginal Likelihood

Held-out prediction

What if we can't compute the marginal likelihood, but can compute likelihoods and MLEs?

Use data the model has not seen.

1. Before model building, decide what to hold out: participants, conditions, items, or future observations.
2. Fit each model on the remaining training data.
3. Compare predictive log likelihood or prediction error on the held-out data.

If the evaluation data are genuinely unseen, we do not need an explicit parameter-count penalty: overfit models tend to predict new data poorly.

Cross-validation

When data are limited, repeat the held-out idea across the whole data set.

	Fold 1	Fold 2	Fold 3	Fold 4
Run 1	test	train	train	train
Run 2	train	test	train	train
Run 3	train	train	test	train
Run 4	train	train	train	test

- Define the partitioning and fitting policy *before* evaluating models.
- Train on $K - 1$ folds and evaluate on the remaining fold.
- Average held-out performance across folds to estimate out-of-sample predictive accuracy.

Cross-validation is not as clean as a fully untouched test set, but it is much better than training and testing on the same data.

Model comparison without test set

Sometimes we still compare models using the same data used for fitting:

- When alternative models were already fitted to the whole data set
- Very few data points, so that setting aside data is costly

Two common in-sample approaches:

1. Likelihood ratio tests for nested models
2. AIC and BIC: likelihood plus a complexity penalty

Nested models and likelihood ratios

Suppose $\mathcal{M}_1$ is a special case of $\mathcal{M}_2$; $\mathcal{M}_2$ has additional parameters and reduces to $\mathcal{M}_1$ for specific values of these parameters. We can say $\mathcal{M}_1$ is nested in $\mathcal{M}_2$.

Even if the additional parameters of $\mathcal{M}_2$ are useless – they just allow it to fit noise – the negative log likelihood will be slightly lower.

Nested models and likelihood ratios

However, under certain assumptions and as n goes to infinity, that improvement (times 2) will converge to a χ^2 distribution with df equal to the difference in dimensionality.[1]

Based on this, we can test whether the improvement in fit of the larger model is greater than expected under the restricted model:

$$\chi^2 \approx 2[\log(p(\mathbf{y}|\hat{\boldsymbol{\theta}}_2, \mathcal{M}_2) - \log(p(\mathbf{y}|\hat{\boldsymbol{\theta}}_1, \mathcal{M}_1)]$$

Caveat: Only applicable to nested models.

[1] To learn more, see Wilks' theorem

Akaike's Information Criterion (AIC)

Another approach: "How different is the distribution implied by my model from the real-world distribution of human behavior?"

How can we quantify this difference?

Kullback-Leibler divergence:[2]

$$\mathrm{KL} = \int_{\mathbf{y}} R(\mathbf{y}) \log \frac{R(\mathbf{y})}{p_M(\mathbf{y})} d\mathbf{y}$$

If these distributions are identical, divergence is zero. If the model assigns zero probability density to events that are possible, it's ∞.

[2] Wikipedia article. Don't call it a distance.

Akaike's Information Criterion (AIC)

AIC approximates relative KL divergences of models to target distribution (e.g., relative probabilities of behaviors):

$$\text{AIC} = -2\log(L(\hat{\boldsymbol{\theta}}|\mathbf{y}, M)) + 2K$$

- K is the number of parameters of the model.
- AIC penalizes more complex (i.e., flexible) models, because the same data is used to estimate $\hat{\boldsymbol{\theta}}$ and the relative KL divergence.
- Asymptotically agrees with leave-one-out cross-validation.
- There are many alternatives, but AIC is simple and popular.

Bayesian Information Criterion (BIC)

The BIC is a popular alternative to the AIC:

$$BIC = -2\log(L(\hat{\boldsymbol{\theta}}|\mathbf{y}, M)) + K \log N$$

where N is the number of data points and K is the number of parameters.

- BIC is motivated as an asymptotic approximation to Bayesian model comparison/marginal likelihood
- Can be understood as a "minimum description length" approach
- Like AIC, a model-comparison method that boils down to MLE likelihoods and counting parameters
- Like AIC, rests on assumptions; guarantees are asymptotic

See Farrell and Lewandowsky (2018) (Sec. 11.26) for details and comparison with AIC.

AIC and BIC: a comparison

	AIC	BIC
Main goal	Predictive accuracy	Model identification / evidence for a model
Penalty	$2K$	$K \log N$
As N grows	Penalty per parameter stays constant	Penalty per parameter increases
Connection	Asymptotically equivalent to leave-one-out cross-validation	Asymptotic approximation to Bayesian model comparison / log marginal likelihood
Typical use	When prediction is the main goal	When comparing models as candidate explanations

For both AIC and BIC: lower is better.

Summary

Key question in model comparison: How do we reward fit while preventing flexibility from winning automatically?

We discussed a set of methods, each of which give a different answer:

- marginal likelihood: average over parameter space;
- held-out data/cross-validation: generalization on unseen data;
- likelihood ratio: ask whether added flexibility significantly improves fit for nested models;
- AIC/BIC: in-sample likelihood penalized for complexity.

References

Farrell, Simon and Stephan Lewandowsky. 2018. *Computational Modeling of Cognition and Behavior*. Cambridge University Press, Cambridge.

Geman, Stuart, Elie Bienenstock, and René Doursat. 1992. Neural networks and the bias/variance dilemma. *Neural Computation* 4(1):1–58.

Jefferys, William H. and James O. Berger. 1992. Ockham's razor and Bayesian analysis. *American Scientist* 80(1):64–72.