Statistical Modeling, Day 3

Model Comparison & Selection

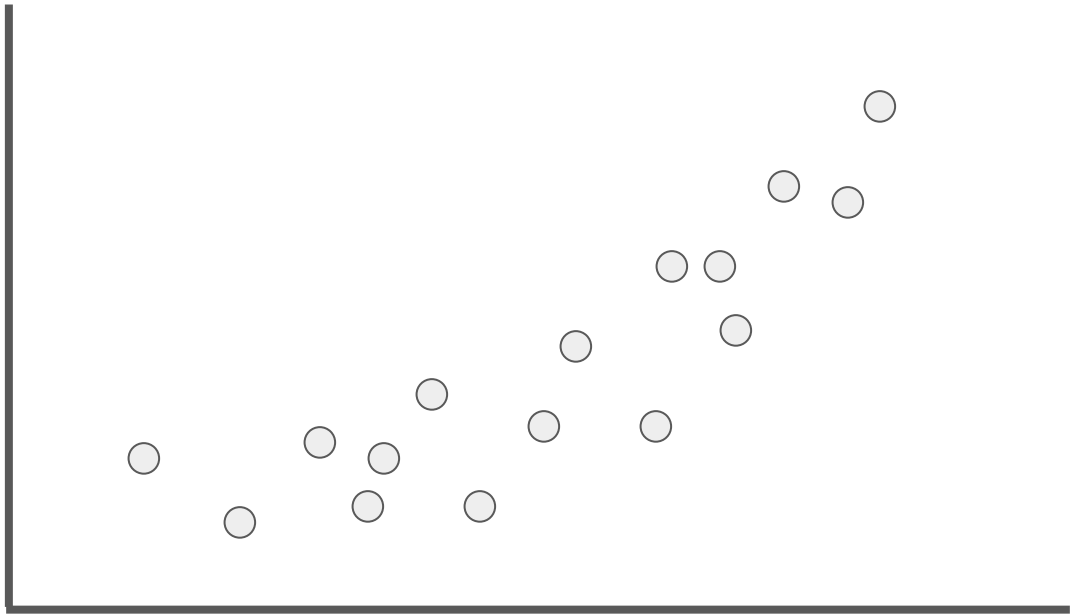
What’s the problem?

Scenario 1:
You’ve fit two or more models to some data. You want to decide which model is the “best” for explaining your data.

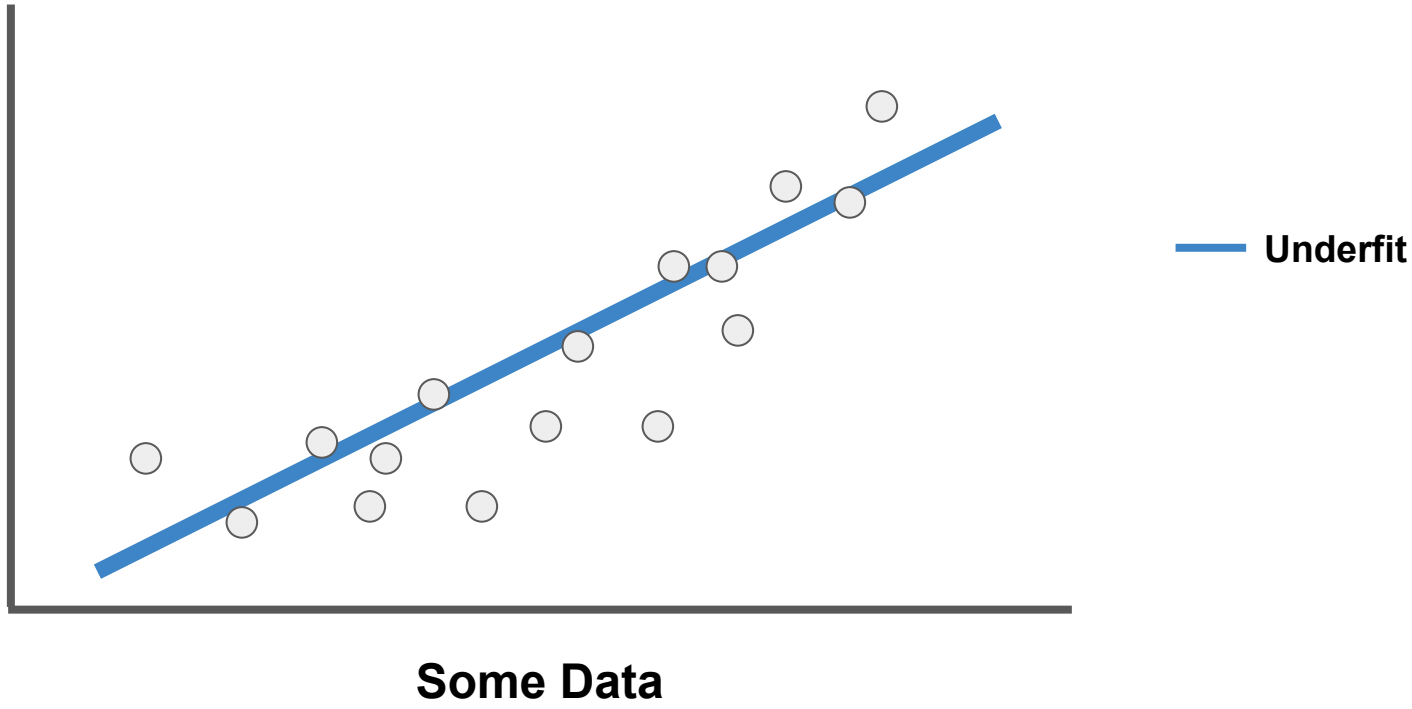
Scenario 2:
You’re reading a paper where a group cites the “best fitting” model of some data as evidence of a particular inference. You want to assess the validity of that claim.

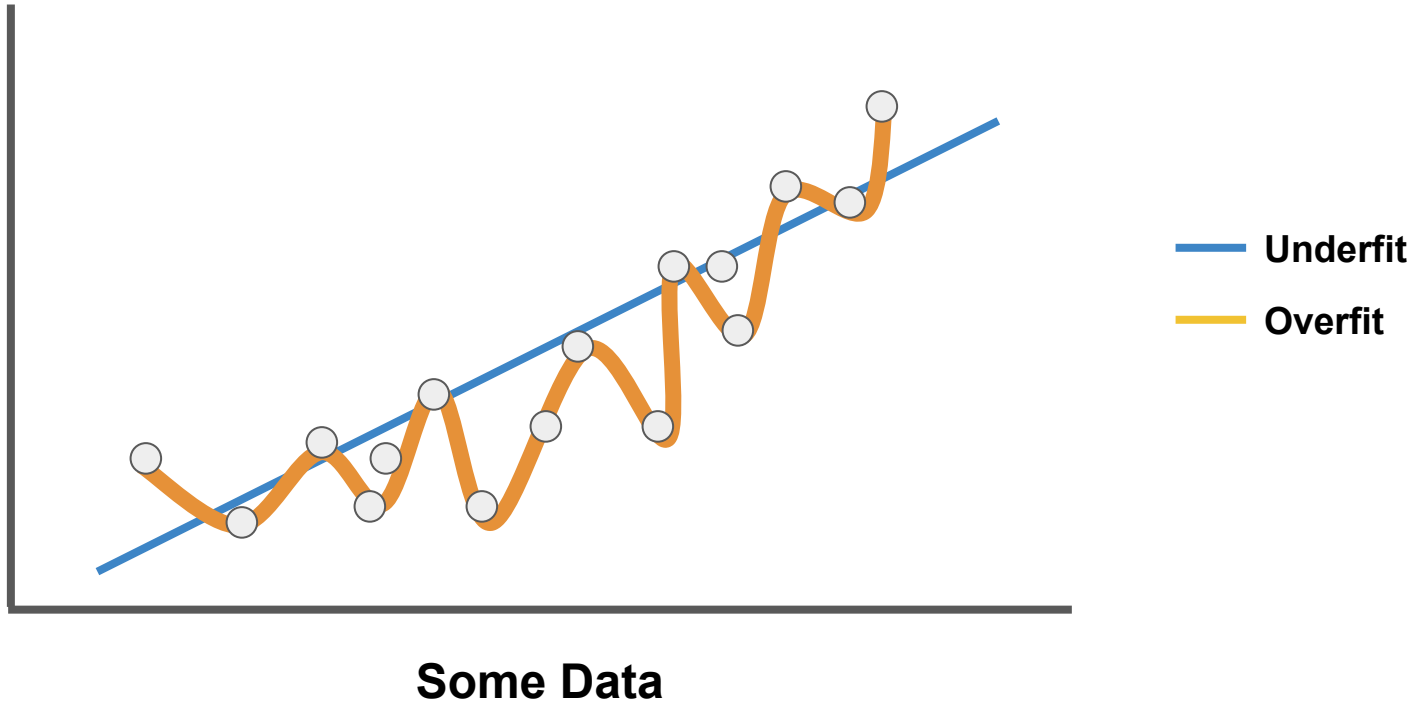
Roadmap

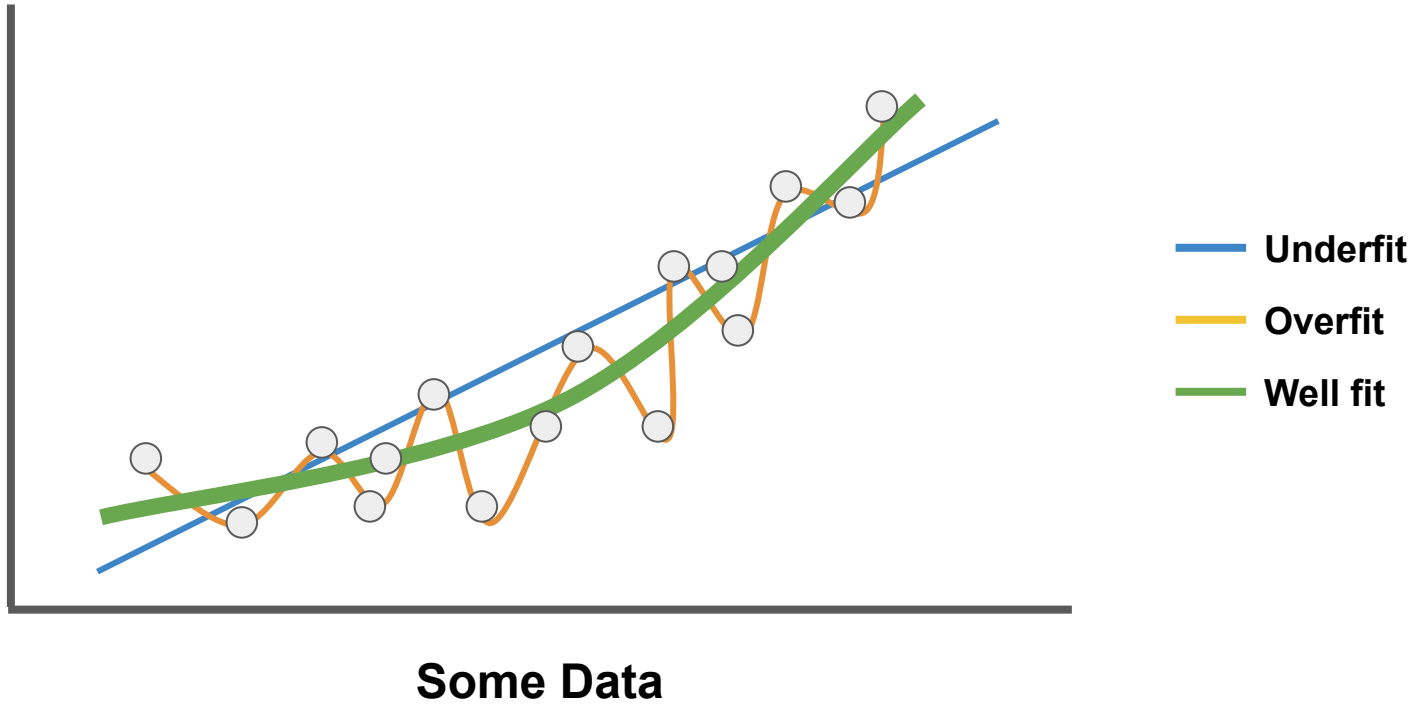
- On what criteria should we compare models?
- How do we normally compare models?
- How should you evaluate model comparison?



Some Data


Underfit
Some Data


Underfit
Overfit
Some Data


Underfit
Overfit
Well fit
Some Data

On what criteria should we compare models?

Jacobs & Grainger (1994)

On what criteria should we compare models?

Descriptive accuracy (i.e., goodness of fit)

How well does the model account for the data?

Jacobs & Grainger (1994)

On what criteria should we compare models?

Descriptive accuracy (i.e., goodness of fit)

How well does the model account for the data?

Simplicity

How parsimonious is the explanation posited by the model? Linked to falsifiability by Popper (1935)

Jacobs & Grainger (1994)

On what criteria should we compare models?

Descriptive accuracy (i.e., goodness of fit)

How well does the model account for the data?

Simplicity

How parsimonious is the explanation posited by the model? Linked to falsifiability by Popper (1935)

Generality

What is the model's ability to generalize to new units of measurement (different tasks, new participants, new stimulus sets?)

Jacobs & Grainger (1994)

On what criteria should we compare models?

Descriptive accuracy (i.e., goodness of fit)

How well does the model account for the data?

Simplicity

How parsimonious is the explanation posited by the model? Linked to falsifiability by Popper (1935)

Generality

What is the model's ability to generalize to new units of measurement (different tasks, new participants, new stimulus sets?)

Explanatory adequacy

Does the model invoke plausible assumptions to explain the data?

Jacobs & Grainger (1994)

With what metrics do we compare models?

Camp 1: Identify “true” data-generating model

- Marginal likelihood
- Bayes factor
- Bayesian information criterion (BIC)
- Integrated BIC

Camp 2: Maximize out-of-sample predictive accuracy

- Cross validation
- Akaike information criterion (AIC)
- Deviance information criterion (DIC)
- Widely applicable information criterion (WAIC)

With what metrics do we compare models?

Camp 1: Identify "true" data-generating model

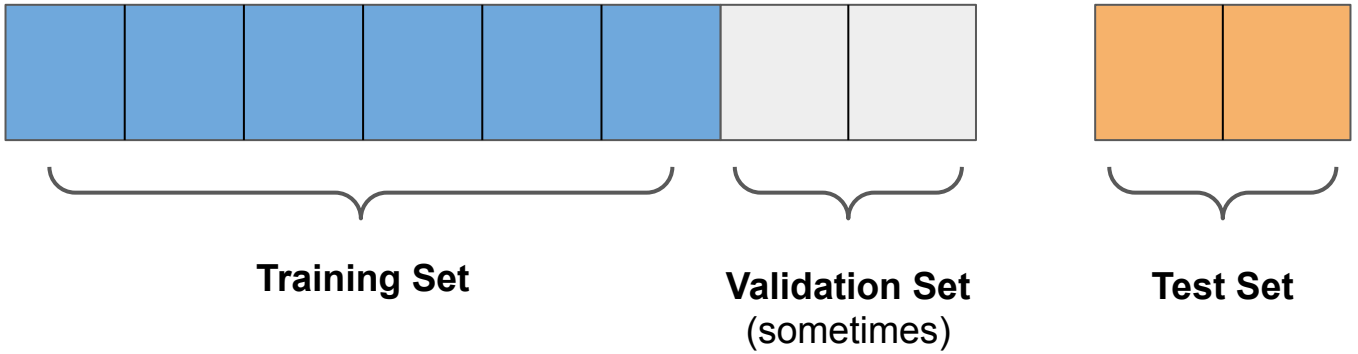
- Marginal likelihood
- Bayes factor
- Bayesian information criterion (BIC)
- Integrated BIC

Camp 2: Maximize out-of-sample predictive accuracy

- Cross validation
- Akaike information criterion (AIC)
- Deviance information criterion (DIC)
- Widely applicable information criterion (WAIC)

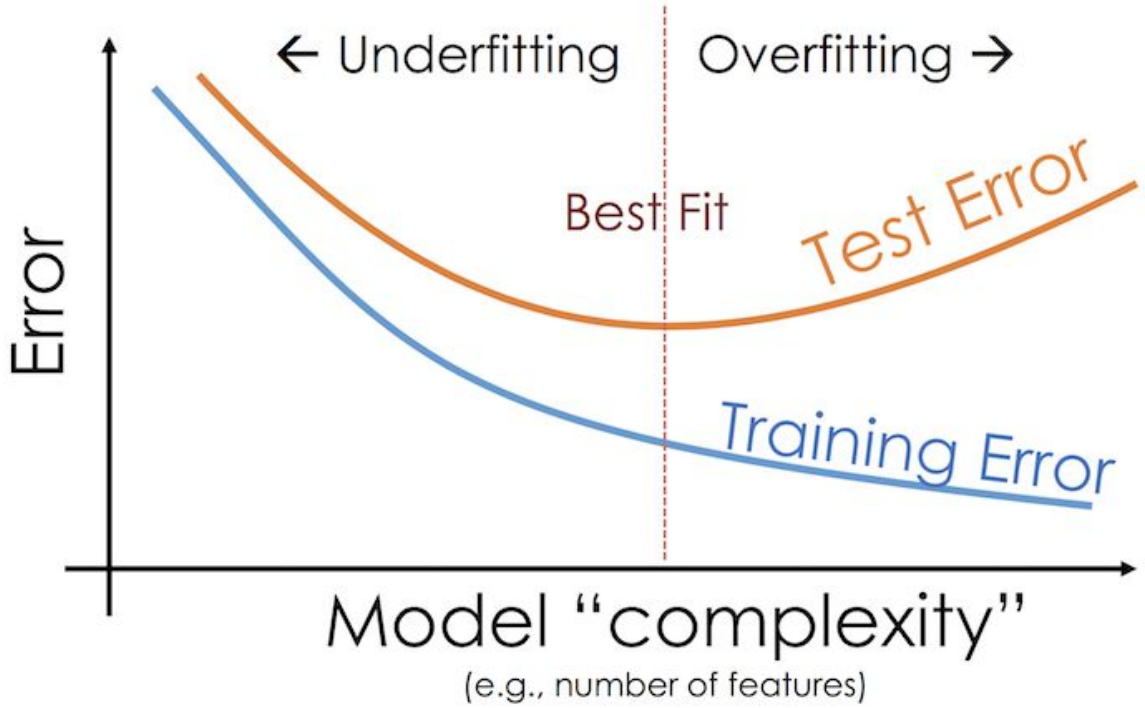
1. In principle, these camps are diametrically opposed (even if not in practice)
2. All these metrics compromise between *goodness-of-fit* and *simplicity*

Train/Test Split and Cross Validation



1. Partition data in training and test sets (validation if possible)
2. Fit model to training set (and optimize fitting routine using validation)
3. Test predictive accuracy on held-out test set

Train/Test Split and Cross Validation



Information criteria: approximating CV**

$$\log p(D \mid \theta_M) - \text{penalty}$$

Likelihood of data under model

Complexity term

Gelman et al. (2013)

Information criteria: approximating CV**

$$\mathrm{AIC} = 2k - 2\log p(D \mid \theta_M)$$

Number of parameters

Gelman et al. (2013)

Information criteria: approximating CV**

$$\mathrm{BIC} = k \log(n) - 2 \log p(D \mid \theta_M)$$

Number of parameters

Number of datapoints

Schwarz (1978)

What does this look like in practice?

Table S1. Performance comparison in terms of the trade-off between model fit and model complexity, Related to **Figure 2**.

Model	*MF alone*	*MB alone*	*dualBayesArb -mean*	*dualBayesArb -reliability*	*dualBayesArb -dynamic*	*mixedArb -mean*	*mixedArb -reliability*	*mixedArb -dynamic*
# param	2	2	4	4	6	4	4	6
AIC	1115.9	546.5	523.1	525.7	519.1	528.2	527.7	517.1*
AICc	1115.9	546.5	523.1	525.8	519.3	528.6	527.9	517.2*
BIC	1125.2	555.8	541.4	544.0	545.5	549.3	546.1	535.6*

Lee et al. (2014)

time for Sam to be annoying

What can we infer from model comparison?

Q: Can we be confident that the best-fitting model as determined by model comparison (e.g. lowest AIC) is a good-fitting model?

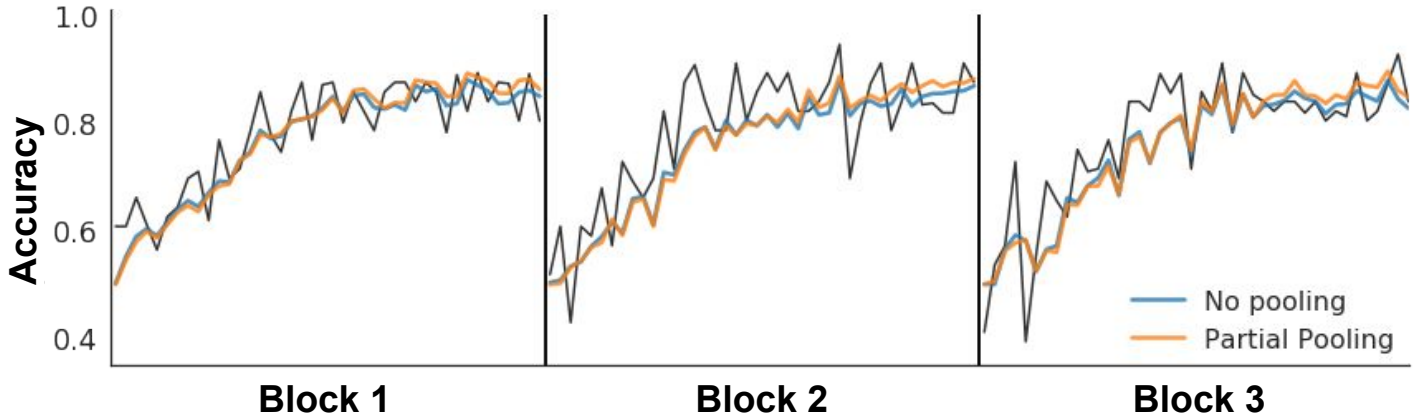
What can we infer from model comparison?

Q: Can we be confident that the best-fitting model as determined by model comparison (e.g. lowest AIC) is a good-fitting model?

A: No. The best-fitting model of 3 shit models is still shit.

Important Takeaway #1a: Predictive Checks

- Simulate data from the estimated parameters of a model
- Graphically compare simulated data to observed data
- Are they similar? Where do they diverge?



Gelman & Shelizi (2013)

Important Takeaway #1b: Baseline Checks

- Fit the simplest *plausible* model to your data
- Does your model outperform the baselines?

Wilson & Collins (2019)

What can we infer from model comparison?

Q: We perform an experiment where a subset of trials are of interest (e.g. optogenetic inhibition). We fit two models – a baseline and an alternative – and find that model comparison based on all trials prefers the alternative.

Can we be confident is the alternative model is favored because it better predicts the experimental trials?

What can we infer from model comparison?

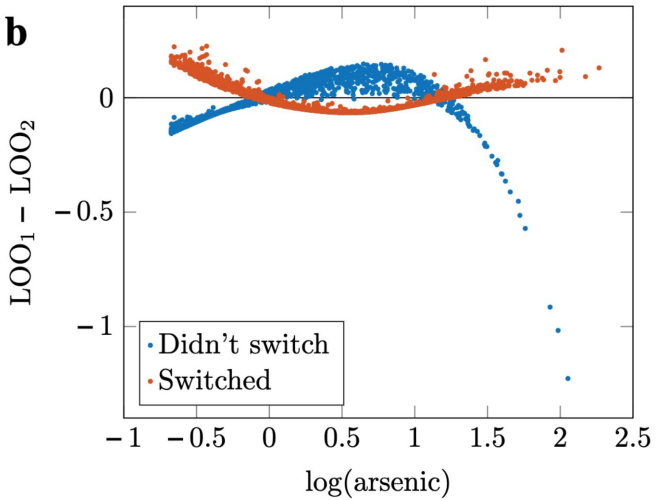
Q: We perform an experiment where a subset of trials are of interest (e.g. optogenetic inhibition). We fit two models – a baseline and an alternative – and find that model comparison based on all trials prefers the alternative.

Can we be confident is the alternative model is favored because it better predicts the experimental trials?

A: No. Differences in model metrics computed from all trials may not be due to differences in the subset of trials of interest.

Important Takeaway #2: Goal-Driven Analysis

- Evaluate your model w.r.t. your hypotheses / goals
- Check that your model outperforms others specifically at the points of interest



Vehtari et al. (2017)

What can we infer from model comparison?

Q: You fit two models to a dataset using cross-validation. The baseline model returns a cross-validated predictive accuracy of 72%. The alternate model returns a cross-validated predictive accuracy of 73%.

Can we be confident this is a meaningful improvement in model performance?

What can we infer from model comparison?

Q: You fit two models to a dataset using cross-validation. The baseline model returns a cross-validated predictive accuracy of 72%. The alternate model returns a cross-validated predictive accuracy of 73%.

Can we be confident this is a meaningful improvement in model performance?

A: It depends! What constitutes “meaningful” in your subfield?

Model comparison is (sometimes) necessary,
but never sufficient.

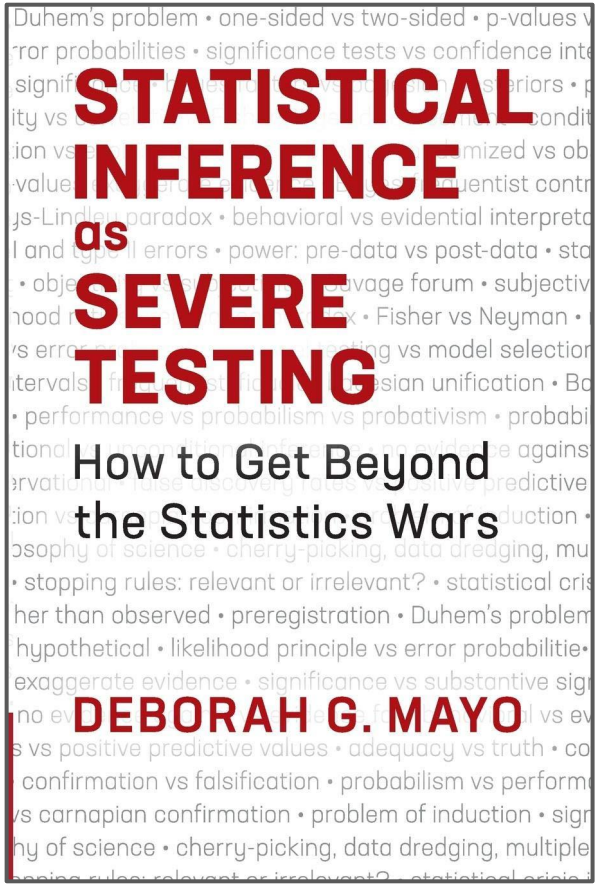
Trends in Cognitive Sciences

Opinion

The Importance of Falsification in Computational Cognitive Modeling

Stefano Palminteri,[1,2,*,‡] Valentin Wyart,[1,2,*,‡] and Etienne Koechlin[1,2,*]

In the past decade the field of cognitive sciences has seen an exponential growth in the number of computational modeling studies. Previous work has indicated why and how candidate models of cognition should be compared by trading off their ability to predict the observed data as a function of their complexity. However, the importance of falsifying candidate models in light of the observed data has been largely underestimated, leading to important drawbacks and unjustified conclusions. We argue here that the simulation of candidate models is necessary to falsify models and therefore support the specific claims about cognitive function made by the vast majority of model-based studies. We propose practical guidelines for future research that combine model comparison and falsification.



Design experiments to falsify models.