Model comparison

Choose the best model!

Test if the model is correct!

(model checking)

Model comparison

Choose the <u>best</u> model!

Test if the model is <u>correct</u>!

(model checking)

Existing methods can say:

“If the model is true it produces such data extremely rarely.”

Or

“If the model is true it produces such data extremely rarely.”

Model comparison

Choose the <u>best</u> model!

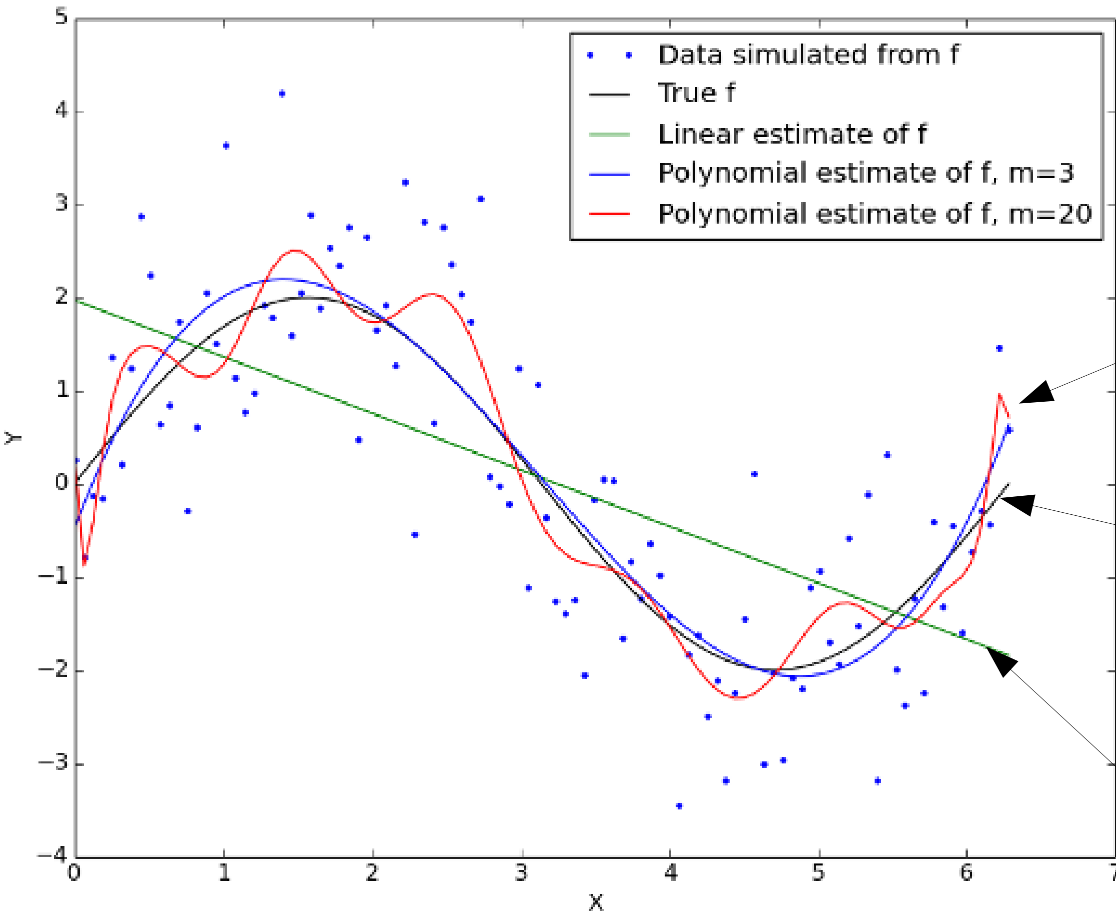
What is <u>best</u>?

- If it fits the data!

Create a model that equals the data
→ fits perfectly

different styles of model comparison

Bias-variance trade-off



Overfitting:

Prediction of left out data is bad == high variance

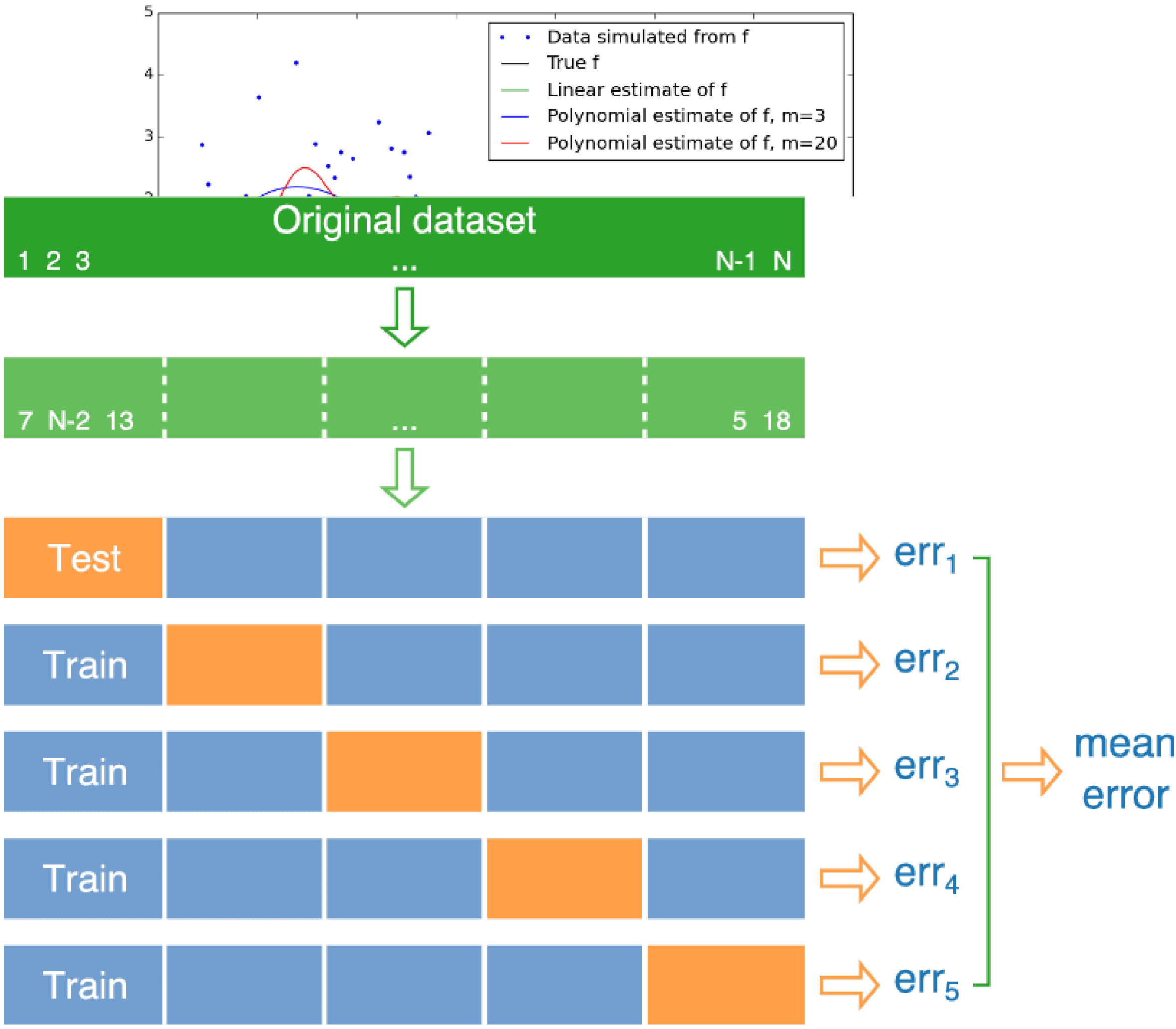
Balance between variance and bias

Underfitting:

Prediction of fitted data is bad == high bias

Empirical, ad-hoc models

- Polynomials, splines, …
- Capture effective behaviour
- Prediction quality of left-out data



K-folding
Boot-strapping

Leave-one-out (LOO), jackknife

Problem-dependent
how to split

Common in machine learning!

Best = predicts unseen data from the same pool of data

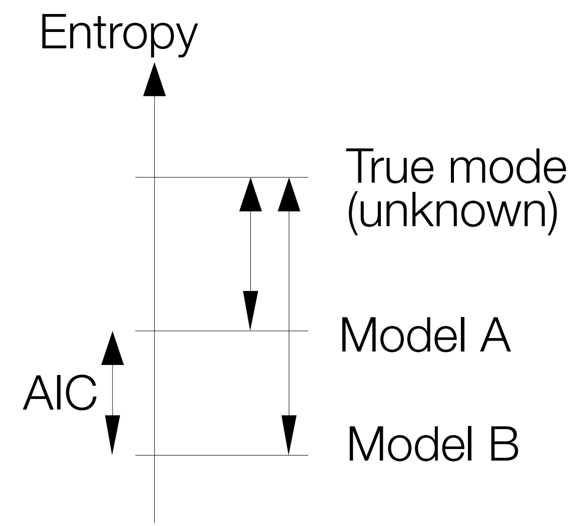
Information theory

- Store data into the fitted model
- Measure entropy, loss of information
- KL divergence between model result and true model
- Akaike information criterion:

$$AIC = -2 * \underbrace{\log(L)}_{\text{maximum likelihood}} + 2 * \underbrace{p}_{\text{number of parameters}}$$

Beware: derived in the limit of high-data

Best = retains information in the data



Variations

DIC: Deviance information criterion "effective" number of parameters

AICc: correction for few data

WAIC: uses the likelihood contribution from each data point

Effect size

- Model B = Model A + some effect

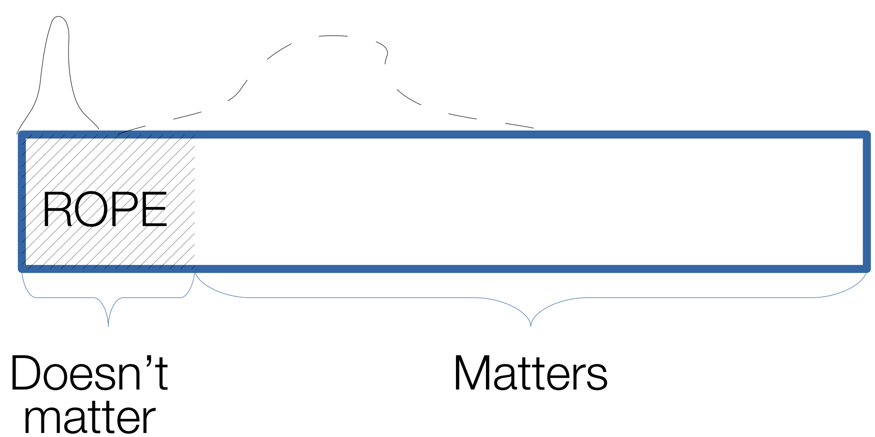
  (e.g., a line, process)

- Does the effect matter?

  Define

  <u>Region of practical equivalence (ROPE)</u>

  between the models

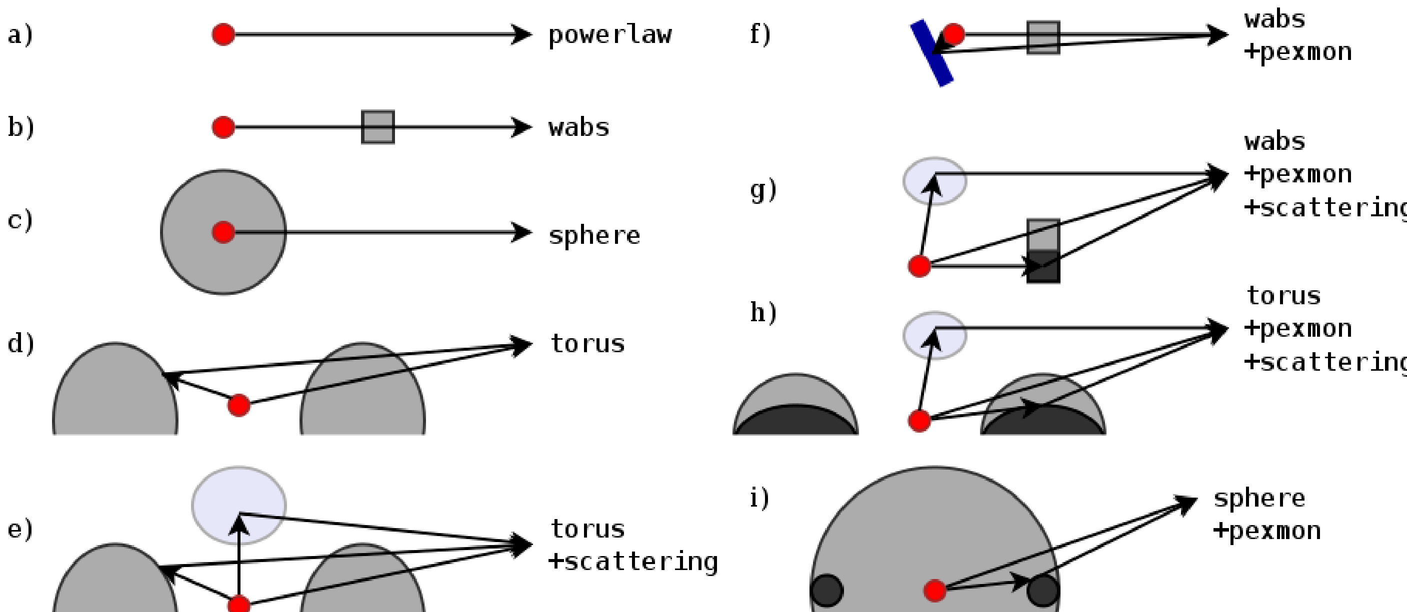


- Is parameter posterior contained in ROPE? → parameter estimation

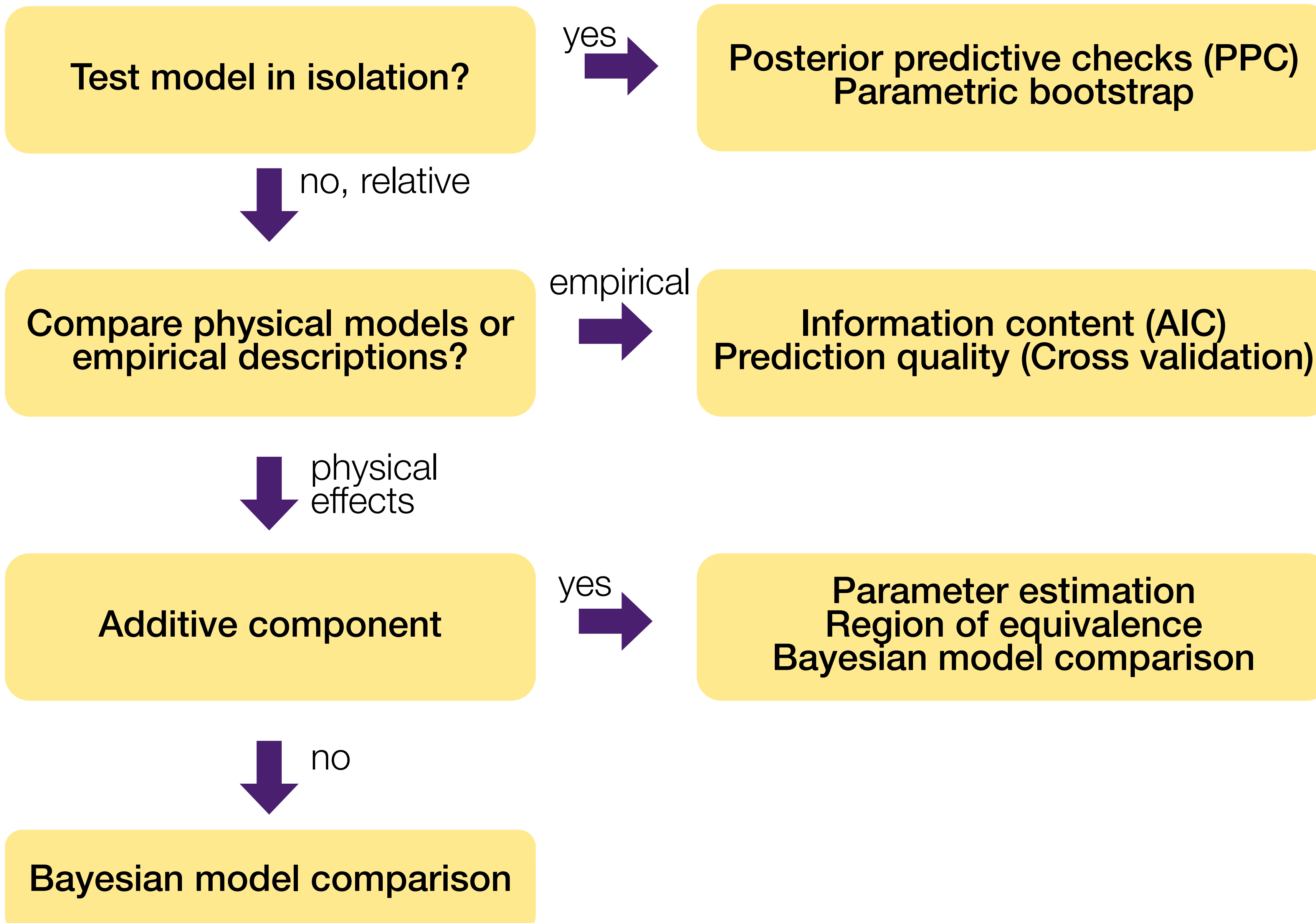
Model comparison

- Empirical models
  - Information content
  - Prediction quality
- Physical effects
  - Priors often well-justified
  - Bayesian model comparison

Obscurer geometry of AGN in the CDF-S
Buchner+14

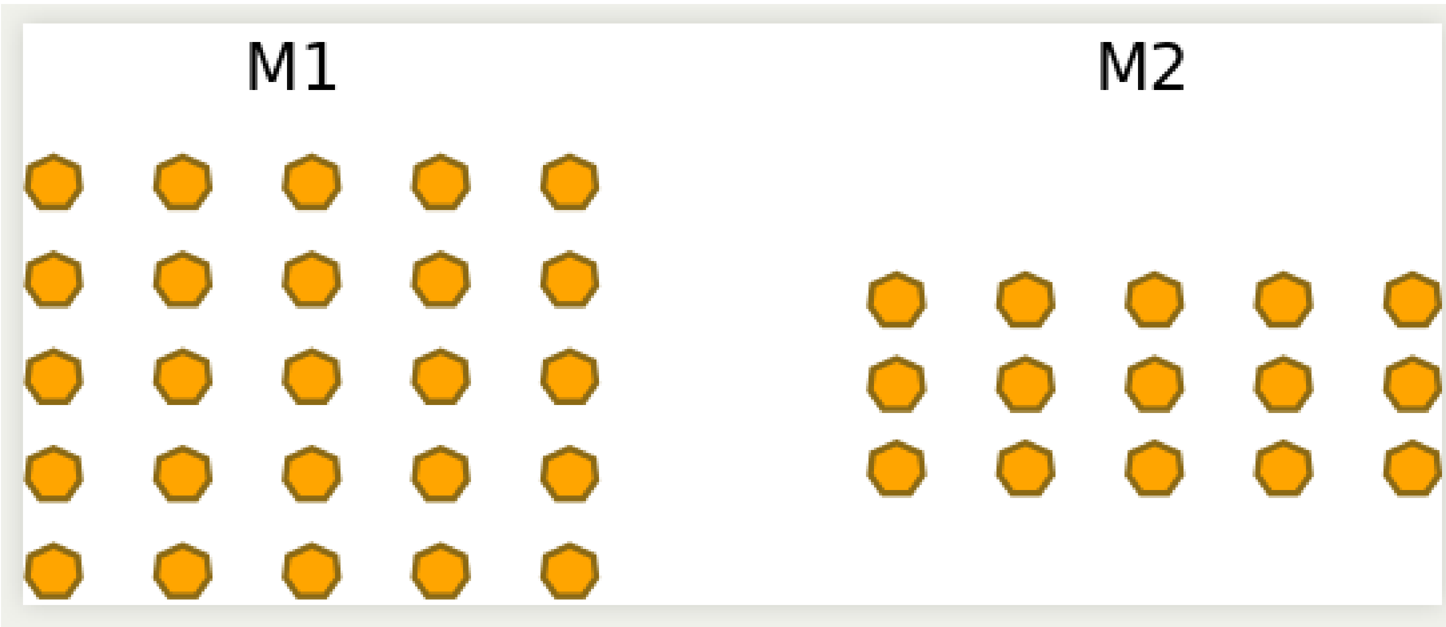


Model comparison



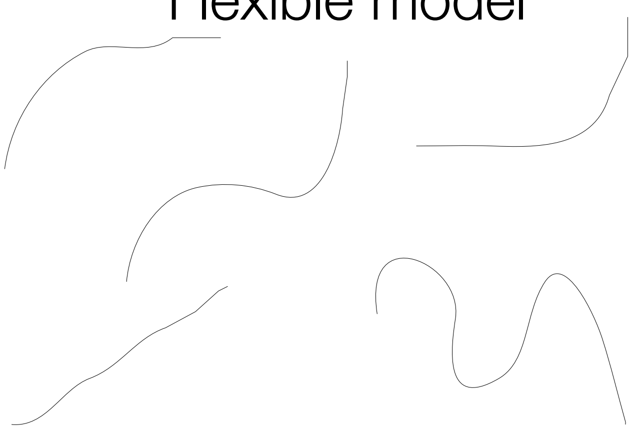
Punishing prediction diversity

(not number of parameters)



Z = likelihood averaged over parameter space according to prior

Flexible model

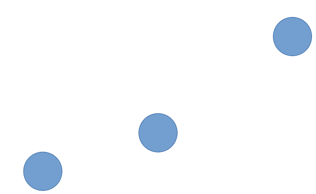


L high, V tiny

Inflexible model

L medium, V medium

Data



What to do with Z

- Z1, Z2

$$\underbrace{\frac{p(M1|D)}{p(M2|D)}}_{\text{Posterior odds ratio}} = \underbrace{\frac{Z1}{Z2}}_{\text{Bayes factor}} \cdot \underbrace{\frac{p(M1)}{p(M2)}}_{\text{Prior odds ratio}}$$

What to do with Z

- Z1, Z2

$$\frac{p(M1|D)}{p(M2|D)} = \frac{Z1 \cdot p(M1)}{Z2 \cdot p(M2)}$$

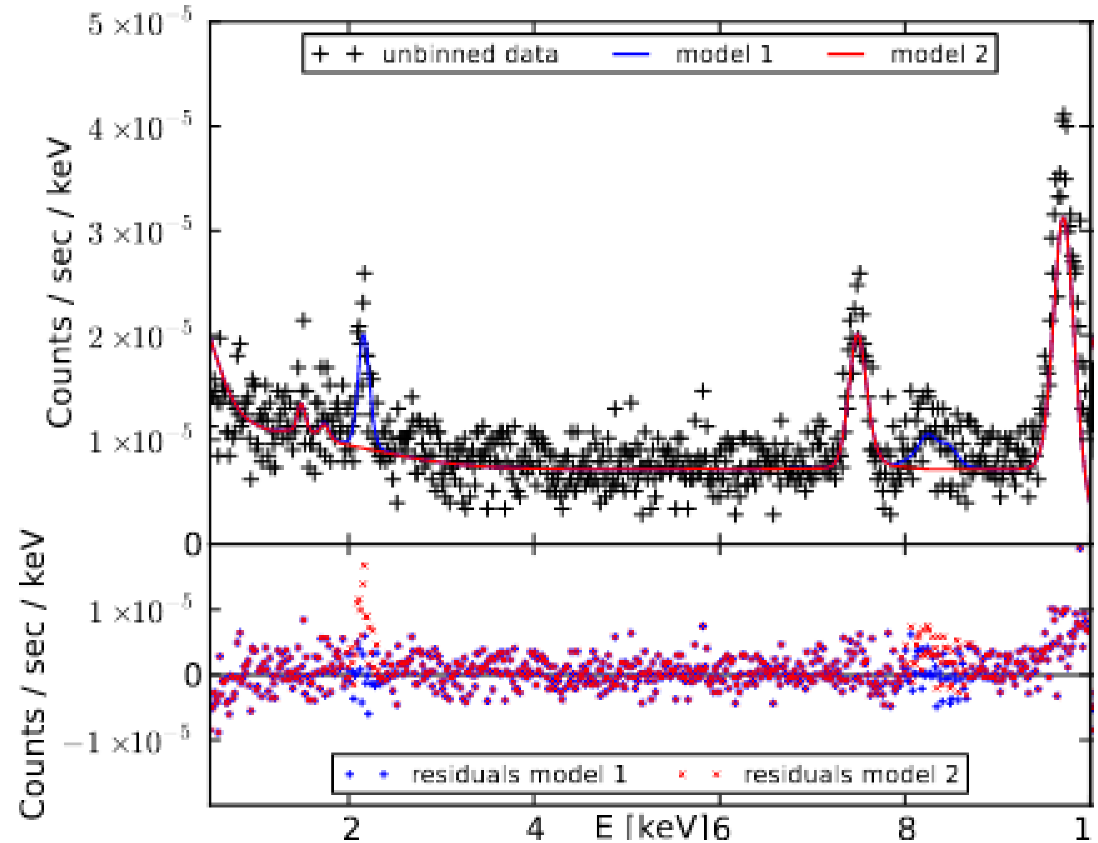
- model priors: leave to reader or motivated by theory
- Does $\frac{p(M1|D)}{p(M2|D)} = 3/1$ mean M2 is correct in a quarter of the cases?

Making decisions

- yes / no
- yes / no / unsure

Choosing between 2 models

- Test statistic
  - can be anything
  - Data counts
  - Likelihood ratio
  - Bayes factor
- Threshold



On a training sample where we know the truth (e.g., simulated data) how often are we right (diagonal)?

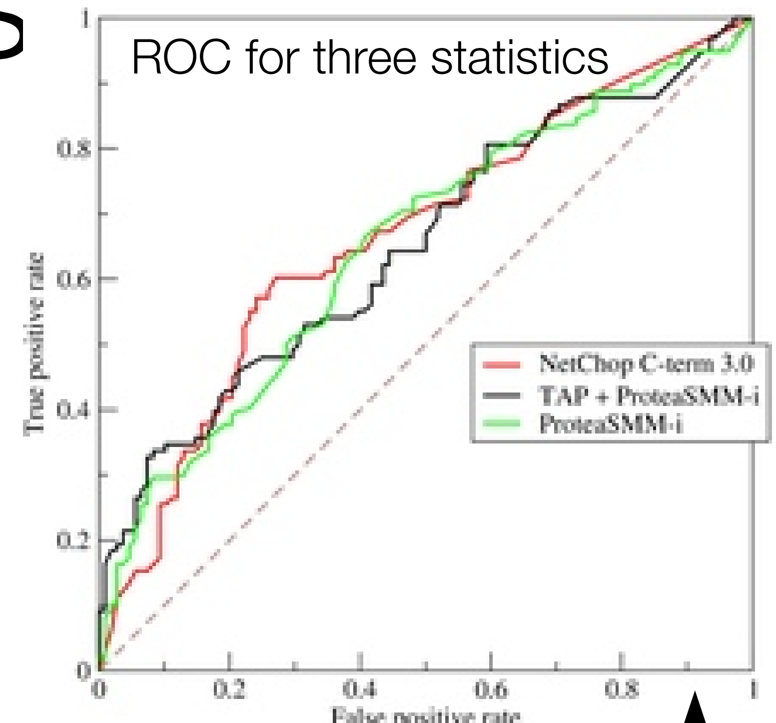
Predicted class / Actual class	Cat	Dog
Cat	6	2
Dog	1	3

Confusion Matrix

Classical statistical tests (KS, F, …) give error rate for H_0 analytically

Comparing selection methods

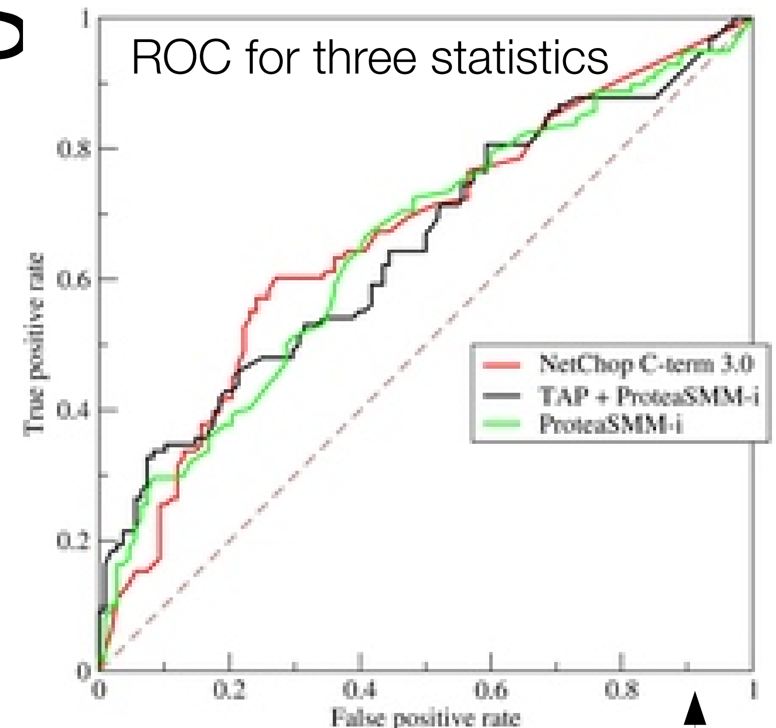
- You have a <u>statistic</u> (a indicator) – (signal strength, likelihood ratio, Bayes factor, classifier probability)
- If you apply <u>threshold:</u>
  - if above → select model A
  - if below → select model B
- There will be
  - cases where you will be right
  - cases where you will be wrong
- determine the rates with simulations from model A and model B.
- Plot the rates as a function of threshold: ROC curve



Predicted class / Actual class	Cat	Dog
Cat	6	2
Dog	1	3

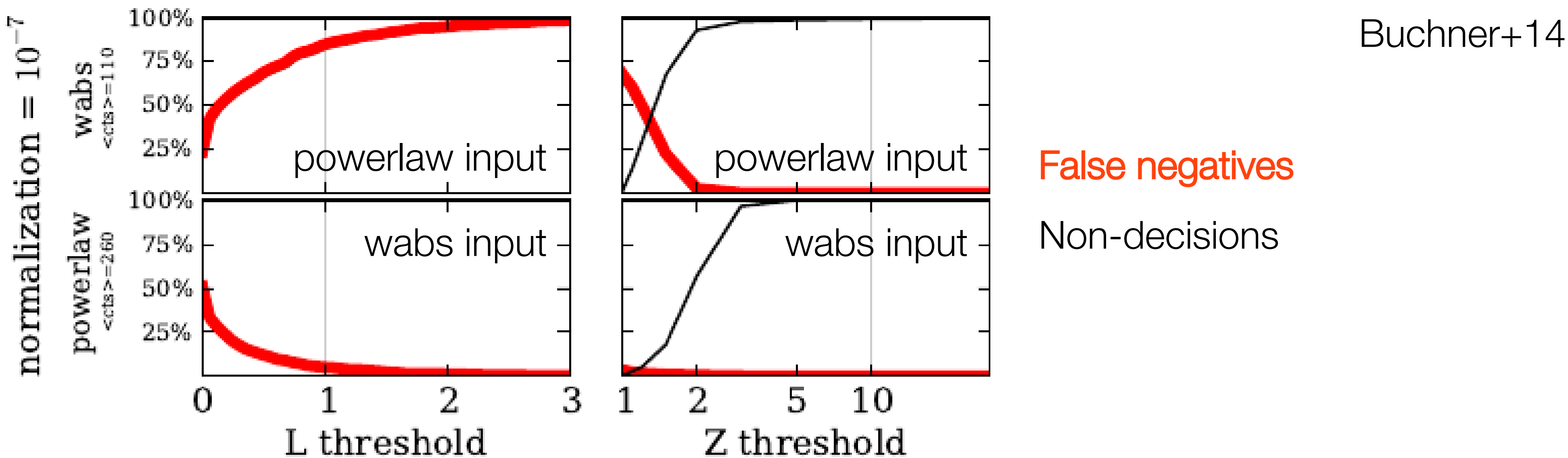
Comparing selection methods

- You have a <u>statistic</u> (a indicator) – (signal strength, likelihood ratio, Bayes factor, classifier probability)
- If you apply <u>threshold:</u>
  - if above → select model A
  - if below → select model B
- There will be
  - cases where you will be right
  - cases where you will be wrong
- determine the rates with simulations from model A and model B.
- Plot the rates as a function of threshold: ROC curve



Predicted class / Actual class	Cat	Dog
Cat	6	2
Dog	1	3

Calibrating model decisions



Advantages:

- Get rid of parameter prior dependences
- Have frequentist properties of Bayesian method
- Completely Bayesian treatment + decisions

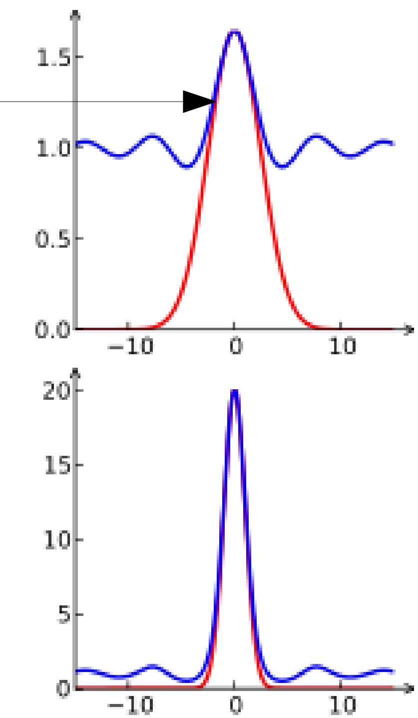
Disadvantages:

- Can be computationally expensive

Computation of Z

Easier to compute

- Monte carlo integration methods
  - Nested sampling, Importance sampling
  - Multi-modality, asymmetries, bananas, etc.
- Laplace approximation
  - Local Gaussian fit to posterior
- Deviance information criterion (DIC)
  - Use only posterior samples
- Bayesian information criterion (BIC)
  - Ignore width
- Maximum of the likelihood



Book:
Information Criteria and Statistical Modeling (Konishi & Kitagawa)

Exercise

- Create a Bayes factor distribution

  P(model A|D) / P(model B|D)

- from simulations from the simpler model

→ example-sine-modelcomparison.ipynb