Model Selection

Daniel Alexander

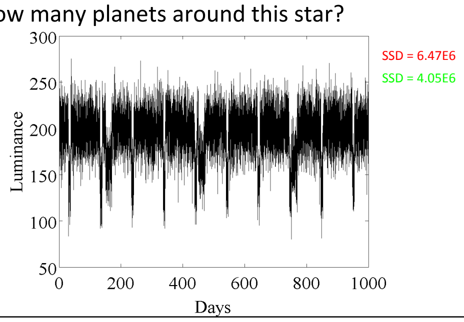
What can we do with a model?

- Learn about the world
- Estimate parameters
- Make predictions

2

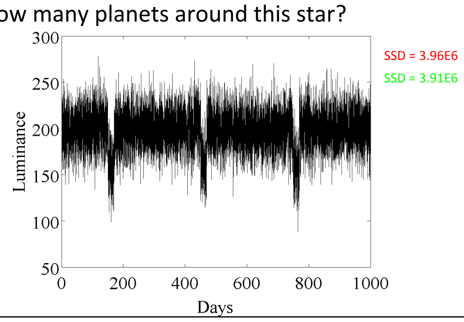
What is the problem?

- How many planets around this star?



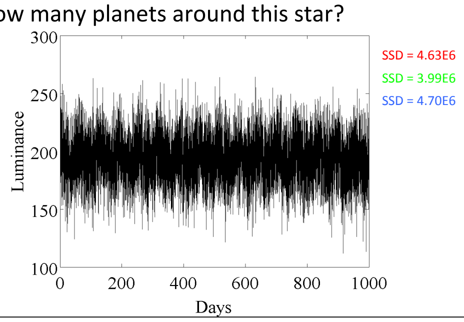
What is the problem?

- How many planets around this star?



What is the problem?

- How many planets around this star?



Occam's Razor

- **Lex parsimoniae** – the law of parsimony
- The best model is the simplest that explains the data.

Guiding Principles

- A good model should…
- Balance goodness of fit with simplicity
- Predict unseen data most closely
- Reflect what's going on in the world

Classical F-test

- Tests the null hypothesis that nested models are equivalent.
- Simple model M_1 with N_1 parameters $x_1, \ldots, x_{N1}$.
- Complex model M_2 with N_2 (> N_1) parameters $x_1, \ldots, x_{N1}, \ldots, x_{N2}$.

Classical F-test

- The statistic

$$F = \frac{(K - N_2 - 1)(\mathrm{Var}(M_2) - \mathrm{Var}(M_1))}{(N_2 - N_1)E(M_2)}$$

- has F-distribution with K-N_2-1 and N_2-N_1 degrees of freedom under the null hypothesis.

$$\mathrm{Var}(M) = \frac{1}{K-1}\sum_{i=1}^{K}\left(M(\tilde{\mathbf{x}};\mathbf{y}_i) - \bar{M}\right)^2; \text{and } \bar{M} = \frac{1}{K}\sum_{i=1}^{K} M(\tilde{\mathbf{x}};\mathbf{y}_i)$$

$$E(M) = \frac{1}{K}\sum_{i=1}^{K}\left(M(\tilde{\mathbf{x}};\mathbf{y}_i) - A_i\right)^2$$

Armitage P, Berry G. Statistical methods in medical research. Oxford, UK: Blackwell Scientific Publications; 1971.

Classical F-test

- Assumes nested models
- Assumes Gaussian noise model
- Rejects or does not reject the null hypothesis that the models are equivalent.

Akaike's information criterion

- The criterion is
  $AIC = 2N - 2\log L$
- where N is the number of parameters in the model and L is the likelihood $p(\mathbf{A} \mid \mathbf{x})$.
- "It is grounded in the concept of information entropy, in effect offering a relative measure of the information lost when a given model is used to describe reality."

Akaike IEEE Trans Automatic Control 1974

AIC

- With i models to choose from, compute each AIC_i.
- Smallest is best.
- Or, model i is $\exp((AIC_{\min} - AIC_i)/2)$ times as likely to be correct as the best model.

AIC

- Any noise model
- Models need not be nested
- More conservative than F-test.

AIC Corrected (AICc)

- Basic AIC valid only as the number of data points, K, tends to infinity.
- For finite K,

$$AICc = AIC + \frac{2N(N+1)}{K-N-1}$$

Bayesian information criterion

- Works in a similar way.
- The criterion is

$$BIC = N \log K - 2 \log L$$

- "The BIC was developed by Gideon E. Schwarz, who gave a Bayesian argument for adopting it."

Schwarz Annals of Statistics 1978

Application in Diffusion MRI

NeuroImage 59 (2012) 2241–2254

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

ELSEVIER

Compartment models of the diffusion MR signal in brain white matter: A taxonomy and comparison

Eleftheria Panagiotaki [a,*], Torben Schneider [b], Bernard Siow [a,c], Matt G. Hall [a], Mark F. Lythgoe [c], Daniel C. Alexander [a]

[a] Centre for Medical Image Computing, Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK
[b] NMR Research Unit, Department of Neuroinflammation, UCL Institute of Neurology, University College London, WC1N 3BG, UK
[c] Centre for Advanced Biomedical Imaging, University College London, Gower Street, London WC1E 6BT, UK

Other criteria

- Deviance information criterion
- Minimum description length
- Minimum message length

Bayesian Model Selection

- Suppose we have two models M_1 and M_2 and some measured data $\mathbf{A}$.

$$p(M_i \mid \mathbf{A}) = \frac{p(\mathbf{A} \mid M_i)p(M_i)}{p(\mathbf{A})} = \frac{p(\mathbf{A} \mid M_i)p(M_i)}{\sum_j p(\mathbf{A} \mid M_j)p(M_j)}$$

- $p(M_i)$ is the prior belief in M_i.
- $p(\mathbf{A} \mid M_i)$ is the likelihood of M_i.
- If M_i has parameters $\mathbf{x}$, then

$$p(\mathbf{A} \mid M_i) = \int p(\mathbf{A} \mid M_i, \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

- where $p(\mathbf{x})$ is the prior on $\mathbf{x}$.

The Bayes Factor

- The Bayes factor is the likelihood ratio

$$K = \frac{p(\mathbf{A} \mid M_1)}{p(\mathbf{A} \mid M_2)} = \frac{\int p(\mathbf{A} \mid M_1, \mathbf{x}_1) p(\mathbf{x}_1) d\mathbf{x}_1}{\int p(\mathbf{A} \mid M_2, \mathbf{x}_2) p(\mathbf{x}_2) d\mathbf{x}_2}$$

- Rule of thumb: if K>10, accept M_1 over M_2.

Cross validation

- Estimate parameters on training set
- Evaluate fit on unseen test set

- Cross validation divides the available data into multiple pairs of training and test sets:
  - ***k*-fold cross-validation**: randomly divide into k equal-sized subsets. Use k-1 sets to fit; compute error on remainder; average error over all k subsets.
  - **Repeated random subsampling**: as above, but draw a random sample each time.
  - **Leave-one-out validation**: $k = K$-1.

Summary

- Find the model that predicts unseen data the best
- Frequentist null hypothesis tests
- Information criteria
- Bayesian model selection
- Cross validation

Guiding Principles Revisited

- A good model should…

- Balance goodness of fit with simplicity

- Predict unseen data most closely

- Reflect what's going on in the world