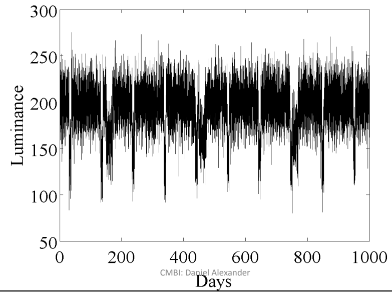
Model Selection

Daniel Alexander

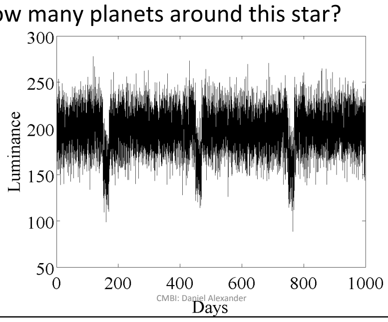
What is the problem?

- How many planets around this star?



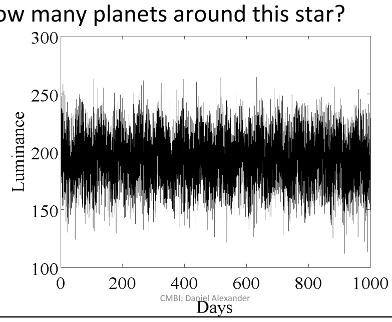
What is the problem?

- 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…
- Predict unseen data closely
- Provide accurate and precise parameter estimates
  - Balance goodness of fit and model complexity
- Reflect what's going on in the world
  - Scientific intuition

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 degrees of freedom N_2-N_1 and K-N_2-1 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.
  - No relative likelihood obtained

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 $p(\mathbf{A} \mid \mathbf{x})$.
- Loss of Kullback-Liebler information using model instead of truth
  - See Burnham and Anderson

Akaike IEEE Trans Automatic Control 1974

AIC – Gaussian noise

- Classic formula for Gaussian noise with unknown variance

$$AIC = 2N + K\log\left(K^{-1}\sum_{k=1}^{K}(A_k - S_k)^2\right)$$

- K must include the standard deviation, which is also estimated.

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.
- These provide *Akaike weights*, which represent model probabilities.
  - See Burnham and Anderson book

AIC

- Any noise model
- Models need not be nested

CMBI: Daniel Alexander

AIC Corrected (AICc)

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

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

- Favour over classic AIC for K/N<40.

CMBI: Daniel Alexander

Bayesian information criterion

- Works in a similar way.
- The criterion is

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

- Derived via a Bayesian argument
  - See Burnham and Anderson

Schwarz Annals of Statistics 1978

CMBI: Daniel Alexander

Application in Diffusion MRI

NeuroImage 59 (2012) 2241–2254

Contents lists available at SciVerse ScienceDirect

ELSEVIER

NeuroImage

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

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

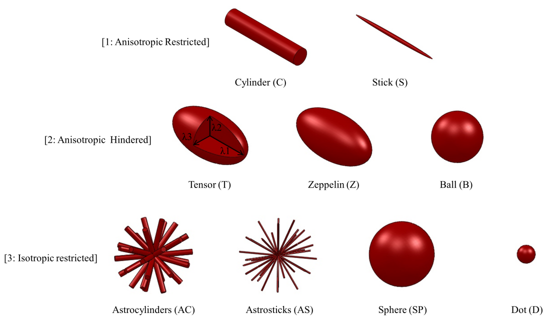
Eleftheria Panagiotaki [a,*], Torben Schneider [b], Bernard Siow [a,c], Matt G. Hall [a], Mark F. Lythgoe [c], Daniel C. Alexander [a]

[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

CMBI: Daniel Alexander

Panagiotaki Models

Panagiotaki et al Neuroimage 2012.



Picture from Richardson MRM 2013

Model ranking

Panagiotaki et al Neuroimage 2012.

Table 4
Table of the models ranked by the Bayesian information criterion (BIC) for dataset A, the ranking of the models according to BIC for dataset B and the number of parameters (counting S_0) for each model.

Models	BIC				No. parameters
	Dataset A×10³	Rank A	Dataset B×10³	Rank B	
TensorCylinderSphere	−2.587	1	−2.636	1	11
TensorStickSphere	−2.558	2	−2.586	2	10
ZeppelinCylinderSphere	−2.495	3	−2.542	5	9
ZeppelinGDRCylindersSphere	−2.489	4	−2.507	7	10
ZeppelinStickSphere	−2.482	5	−2.488	9	8
TensorCylinderDot	−2.481	6	−2.561	3	10
TensorGDRCylindersDot	−2.446	7	−2.557	4	11
TensorGDRCylindersSphere	−2.431	8	−2.518	6	12
ZeppelinCylinderDot	−2.429	9	−2.499	8	8
ZeppelinGDRCylindersDot	−2.412	10	−2.484	10	9
BallGDRCylindersDot	−2.382	11	−2.466	11	8
BallCylinderSphere	−2.377	12	−2.448	13	8
BallCylinderDot	−2.375	13	−2.454	12	7
BallGDRCylindersSphere	−2.366	14	−2.420	14	9
ZeppelinStickDot	−2.329	15	−2.396	15	7
TensorCylinderAstrocylinders	−2.290	16	−2.331	17	10
TensorCylinderAstrosticks	−2.289	17	−2.331	18	10
TensorGDRCylindersAstrosticks	−2.286	18	−2.329	19	11
TensorStickAstrocylinders	−2.282	19	−2.324	20	10
TensorStickAstrosticks	−2.279	20	−2.324	21	9
ZeppelinGDRCylindersAstrocylinders	−2.278	21	−2.230	36	9
BallStickSphere	−2.272	22	−2.332	16	7

Other criteria

- Deviance information criterion
  - Designed to work with MCMC
  - Large sample size
  - Posterior distribution approximately normal.
- Minimum description length
  - Model selection as a coding problem
- Minimum message length
  - Bayesian version of MDL.

CMBI: Daniel Alexander

Bayesian Model Selection

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

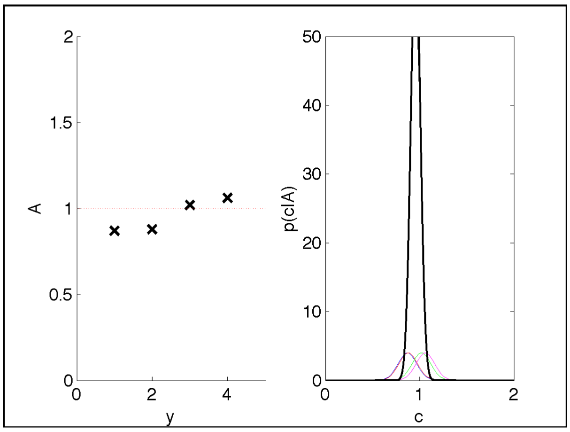
$$p(M_i \mid \mathbf{A}) = \frac{p(\mathbf{A} \mid M_i)p(M_i)}{p(\mathbf{A})} = \frac{p(\mathbf{A} \mid M_i)p(M_i)}{\sum_j p(\mathbf{A} \mid M_j)p(M_j)}$$

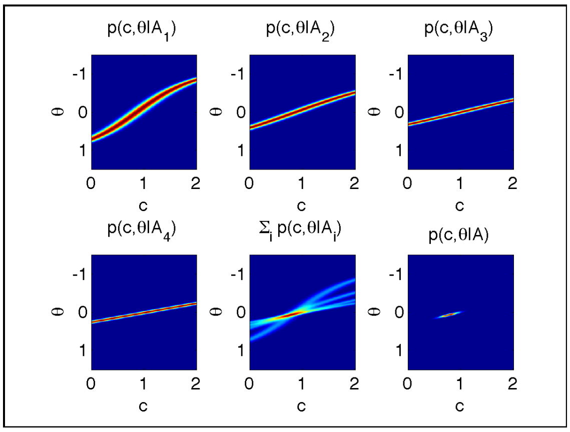
- $p(M_i)$ is the prior belief in M_i.
- $p(\mathbf{A} \mid M_i)$ is the likelihood of M_i.
- If M_i has parameters $\mathbf{x}$, then

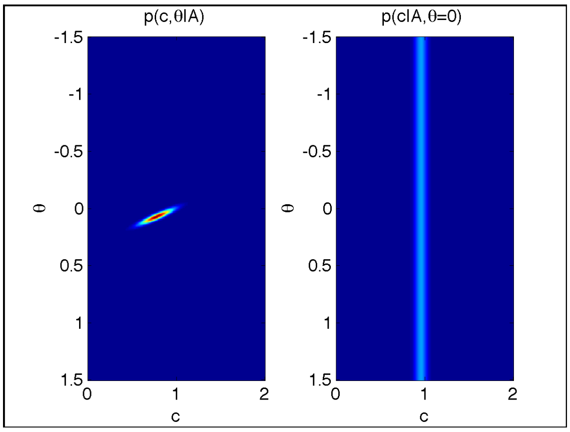
$$p(\mathbf{A} \mid M_i) = \int p(\mathbf{A} \mid M_i, \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

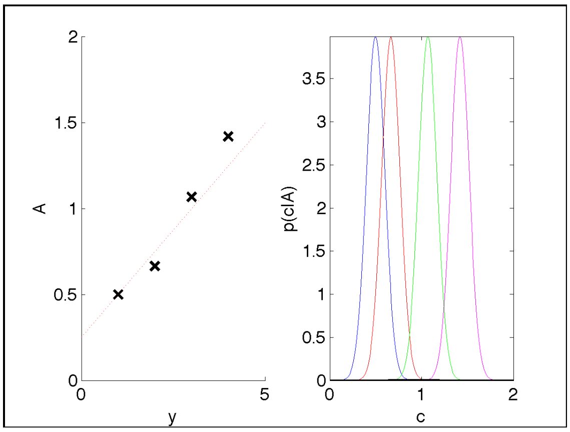
- where $p(\mathbf{x})$ is the term for the prior on $\mathbf{x}$.

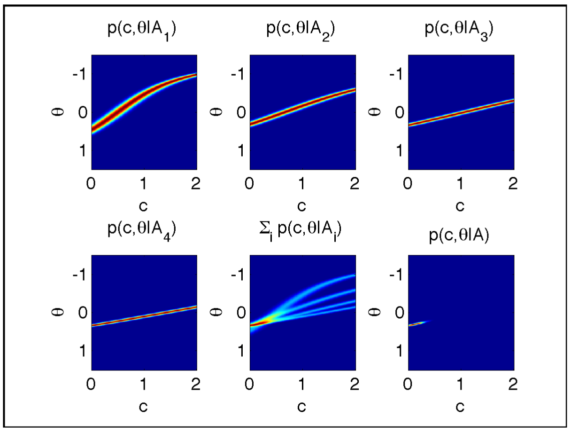
CMBI: Daniel Alexander

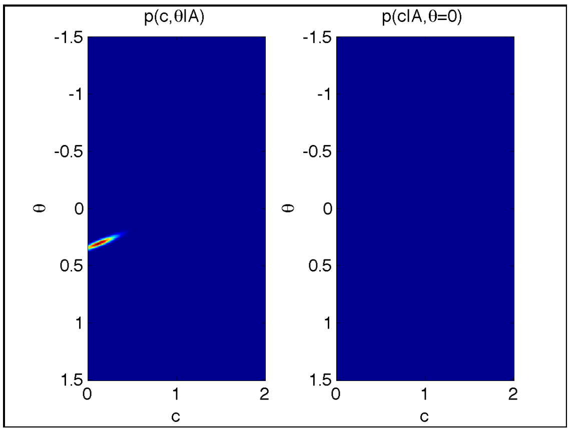












The Bayes Factor

- 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.

CMBI: Daniel Alexander

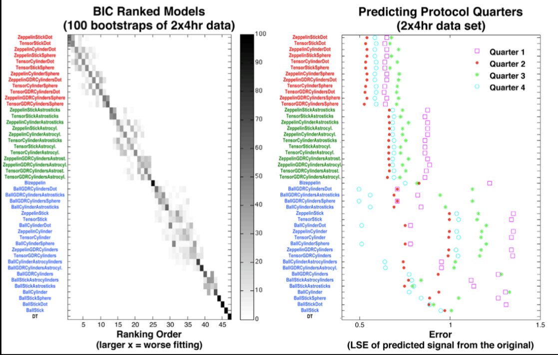
Cross validation

- Ultimate empirical model selection:
  - 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.

CMBI: Daniel Alexander

Ferizi Ranking

Ferizi et al MRM 2014.



Guiding Principles Revisited

- A good model should…
- Predict unseen data closely
- Provide accurate and precise parameter estimates
- Reflect what's going on in the world

CMBI: Daniel Alexander