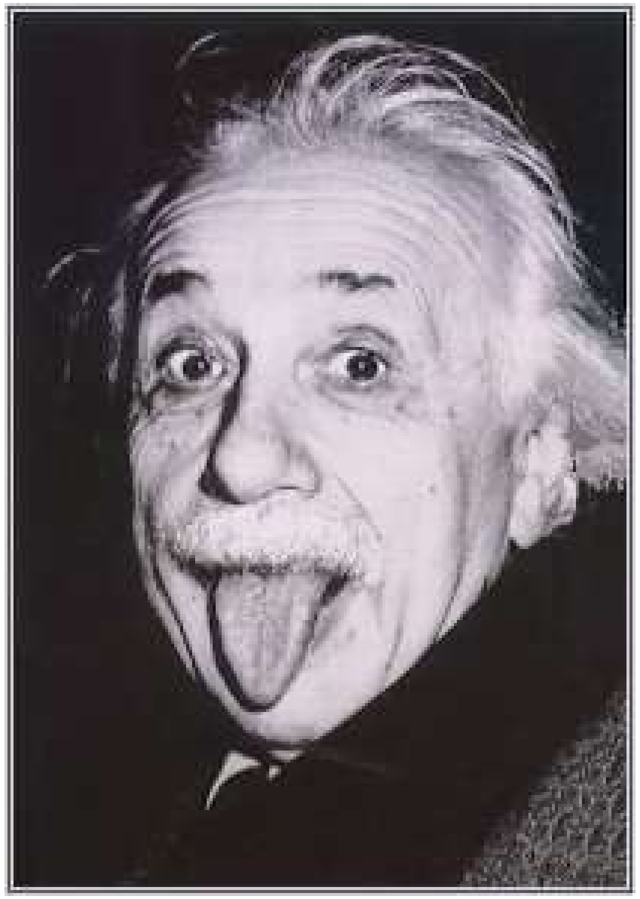
8. Bayesian Model Selection

University of Glasgow

SUPA



“Everything should be made as simple as possible, but not simpler”

Bayes' theorem:

$$p(Y \mid X) = \frac{p(X \mid Y) \times p(Y)}{p(X)}$$

Laplace rediscovered work of
Rev. Thomas Bayes (1763)

Thomas Bayes
(1702 – 1761 AD)

$$\overbrace{p(\theta \,|\, \text{data}, M)}^{\text{Posterior}} = \frac{\overbrace{p(\text{data} \,|\, \theta, M)}^{\text{Likelihood}} \times \overbrace{p(\theta \,|\, M)}^{\text{Prior}}}{\underbrace{p(\text{data} \,|\, M)}_{\text{Evidence}}}$$

Laplace rediscovered work of
Rev. Thomas Bayes (1763)

Thomas Bayes
(1702 – 1761 AD)

Posterior Likelihood Prior

$$p(\theta \mid \text{data}, M) = \frac{p(\text{data} \mid \theta, M) \times p(\theta \mid M)}{p(\text{data} \mid M)}$$

Evidence

$$\text{Evidence} = \underbrace{\int \text{p}(\text{data} \mid \theta, M)\, p(\theta \mid M)\, d\theta}_{\text{Average likelihood, weighted by prior}}$$

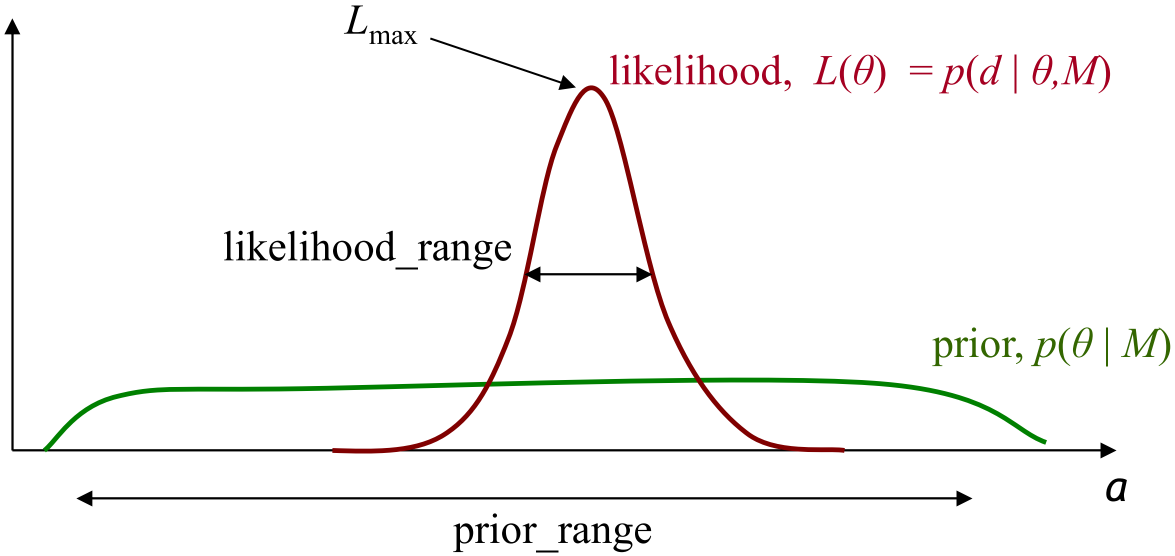
Selecting Between Competing Models

- We can compute the odds ratio of two competing models. This can be divided into the prior odds and the Bayes factor

$$O_{12} = \frac{\text{prob}(M_1 \mid d)}{\text{prob}(M_2 \mid d)} = \underbrace{\frac{\text{prob}(M_1)}{\text{prob}(M_2)}}_{\text{prior odds}} \times \underbrace{\frac{\text{prob}(d \mid M_1)}{\text{prob}(d \mid M_2)}}_{\text{Bayes factor}}$$

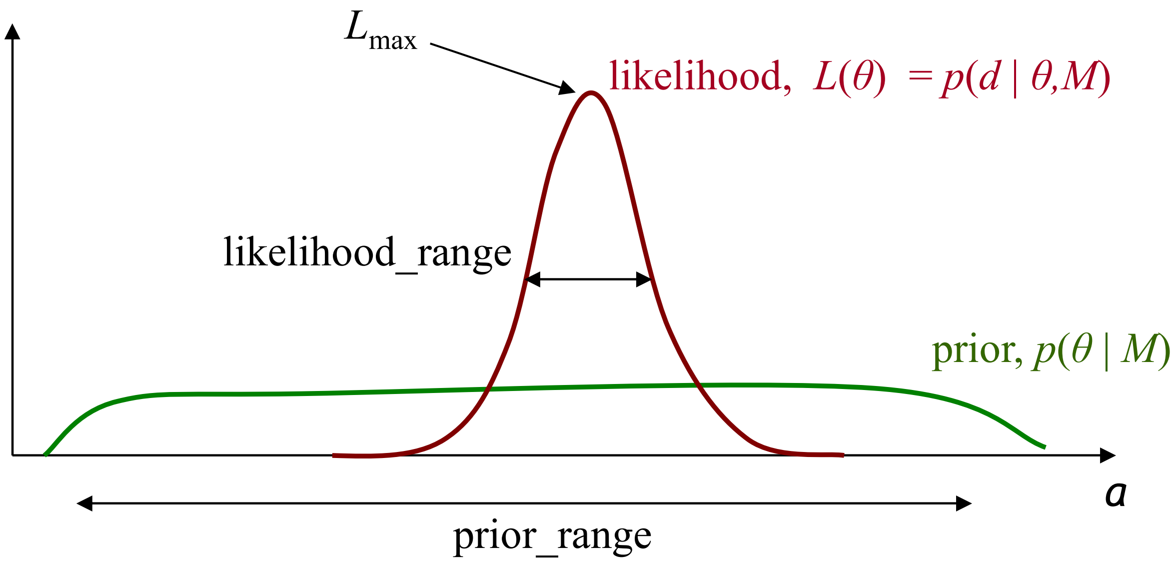
- The Bayes factor is just the ratio of the evidences.

We can split the evidence into two approximate parts:
the maximum of the likelihood and an "Occam factor":



$$p(d \mid M) = \int p(\theta \mid M)\, p(d \mid \theta, M)\, \mathrm{d}\theta \approx L_{\max} \underbrace{\frac{\text{likelihood\_range}}{\text{prior\_range}}}_{\text{the 'Occam factor'}}$$

We can split the evidence into two approximate parts:
the maximum of the likelihood and an “Occam factor”:



$$p(d \mid M) = \int p(\theta \mid M)\, p(d \mid \theta, M)\, \mathrm{d}\theta \approx L_{max} \underbrace{\frac{\text{likelihood\_range}}{\text{prior\_range}}}_{\text{the 'Occam factor'}}$$

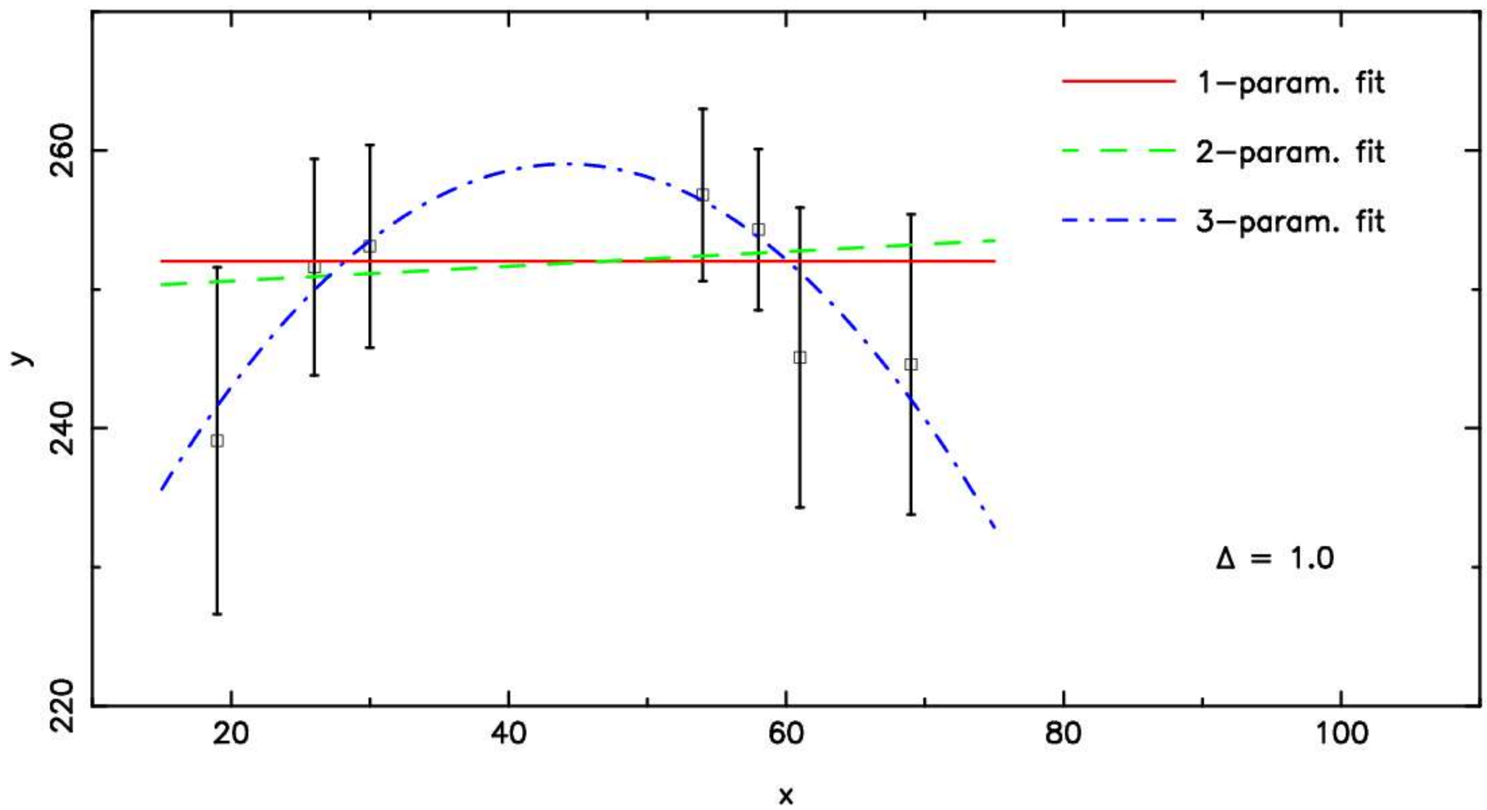
The Occam factor penalises models that include wasted parameter space, even if they show a good ML fit.

William of Ockham
(1288 – 1348 AD)

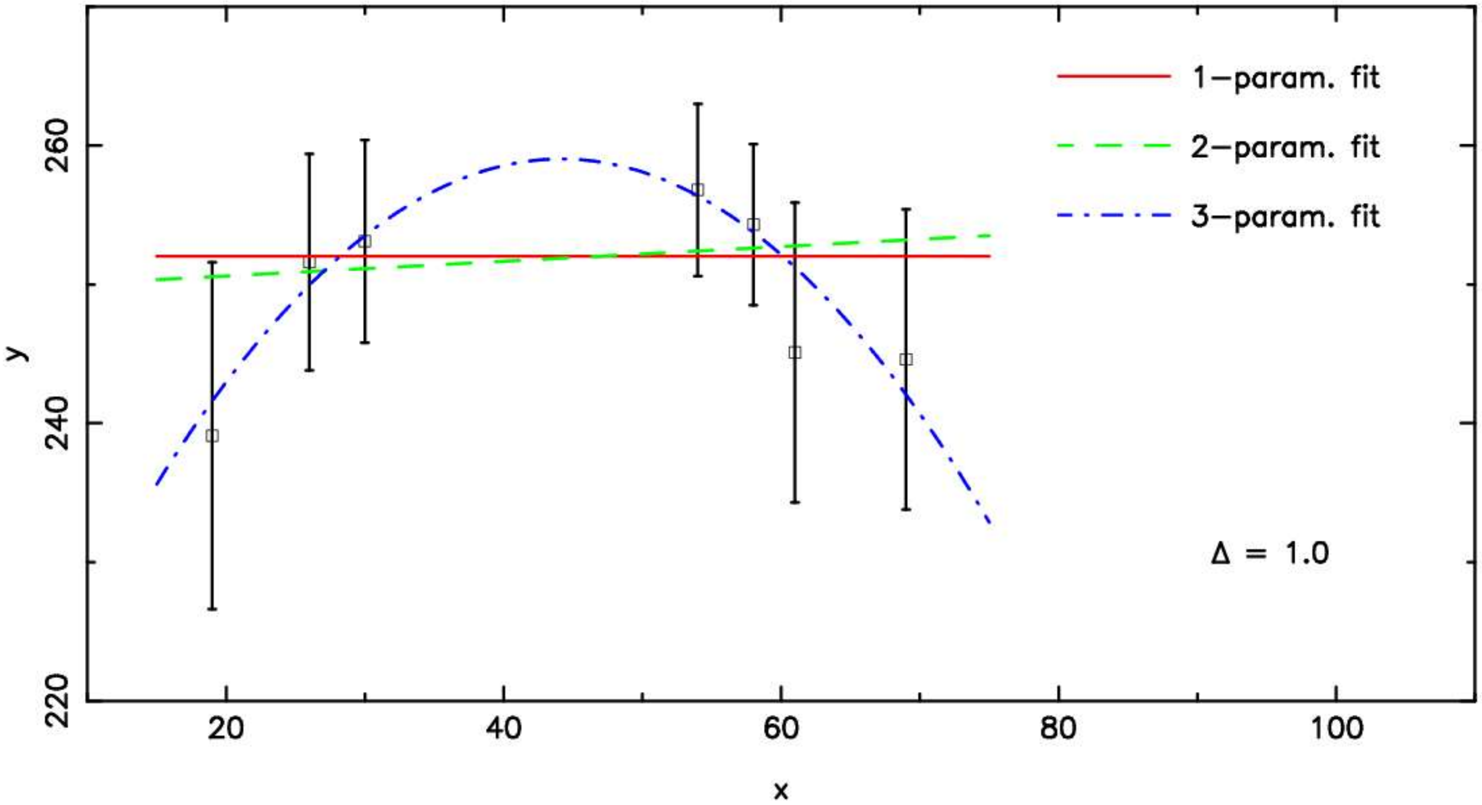
Occam's Razor

"It is vain to do with more what can be done with less."

Everything else being equal, we favour models which are ***simple***.

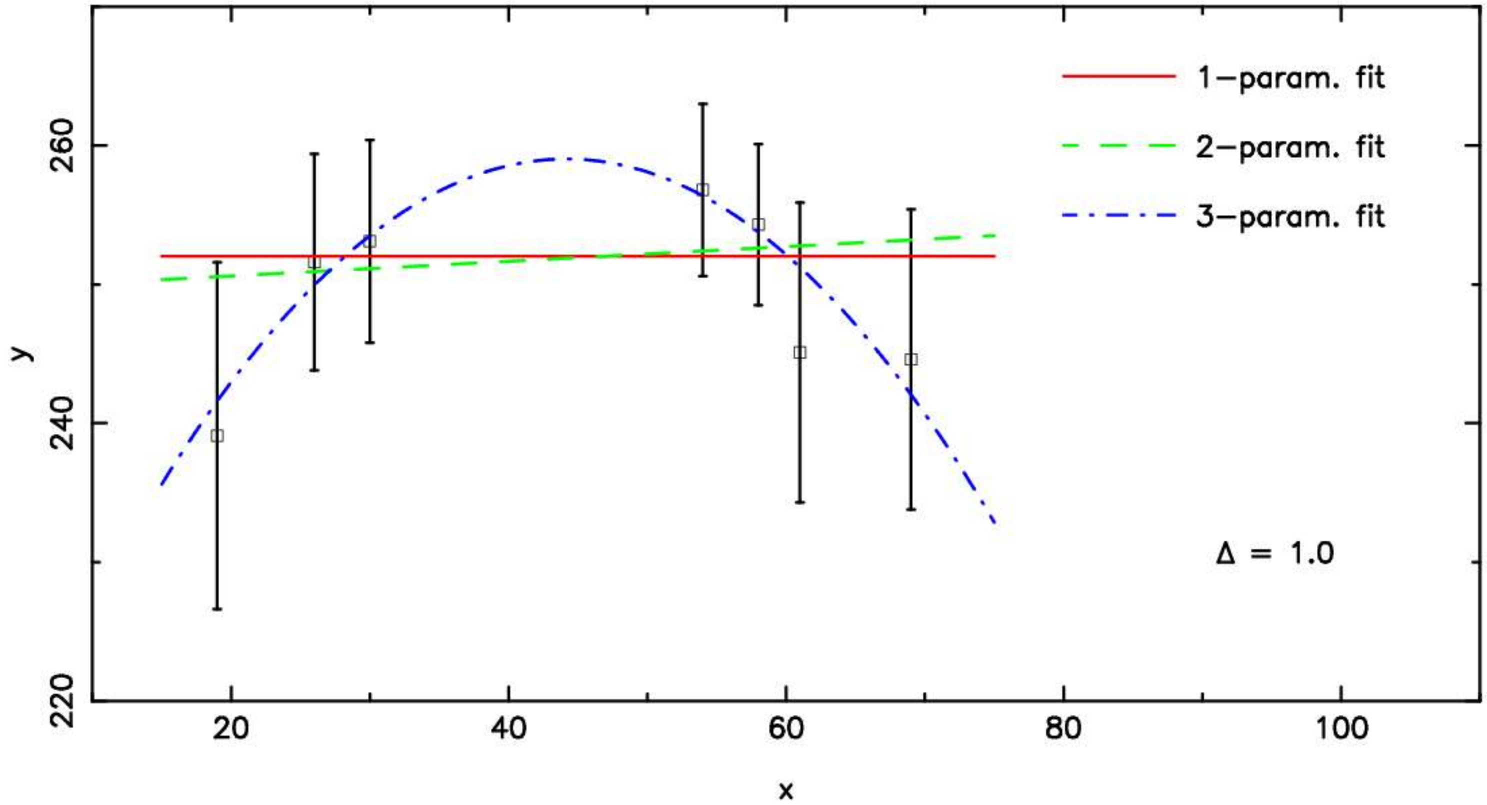


1−param. fit
2−param. fit
3−param. fit
Δ = 1.0
y
x
260
240
220
20
40
60
80
100



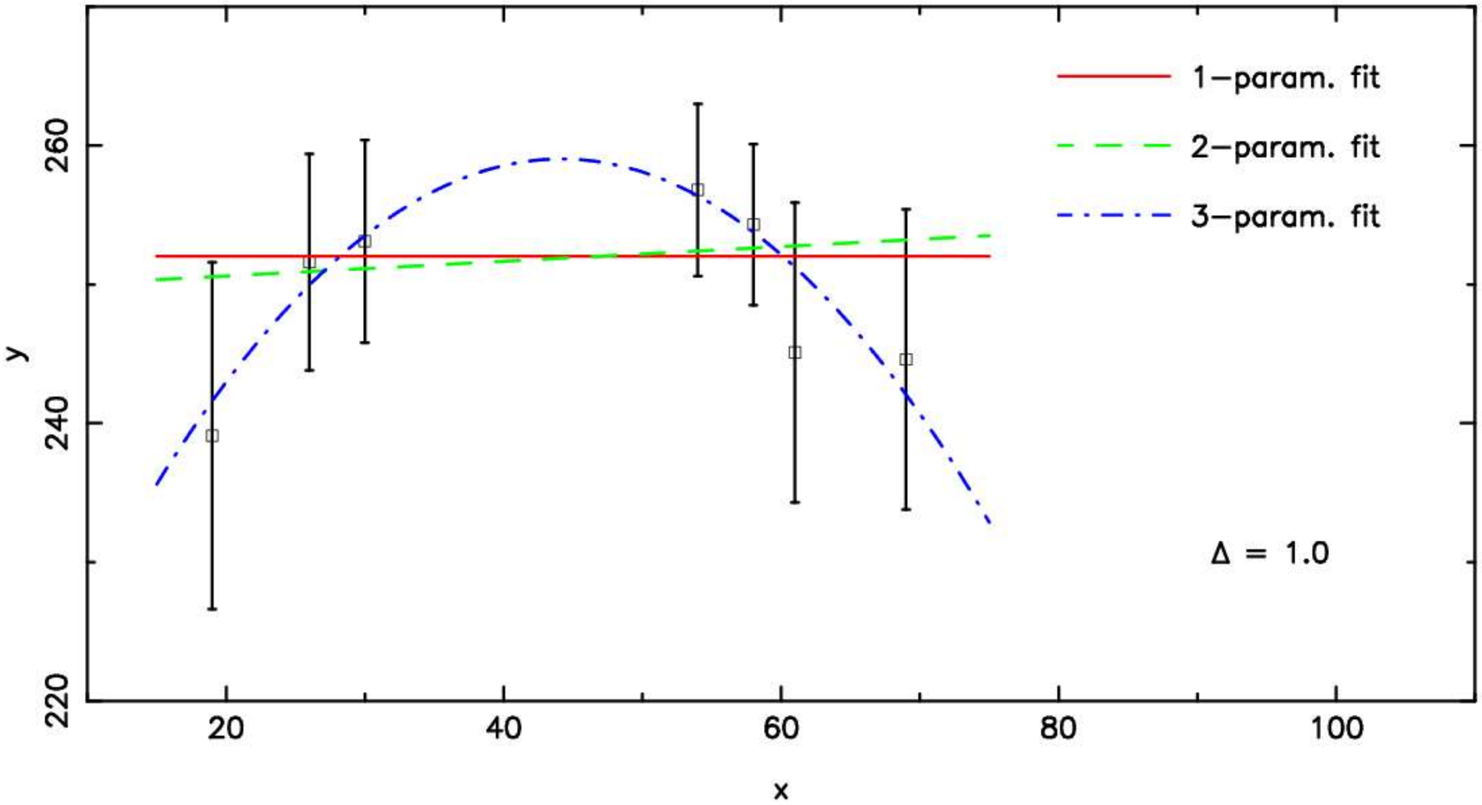
We can compute e.g. $$O_{12} = \frac{\text{prob}(m=1\,|\,d)}{\text{prob}(m=2\,|\,d)}$$

$$O_{13} = \frac{\text{prob}(m=1\,|\,d)}{\text{prob}(m=3\,|\,d)}$$



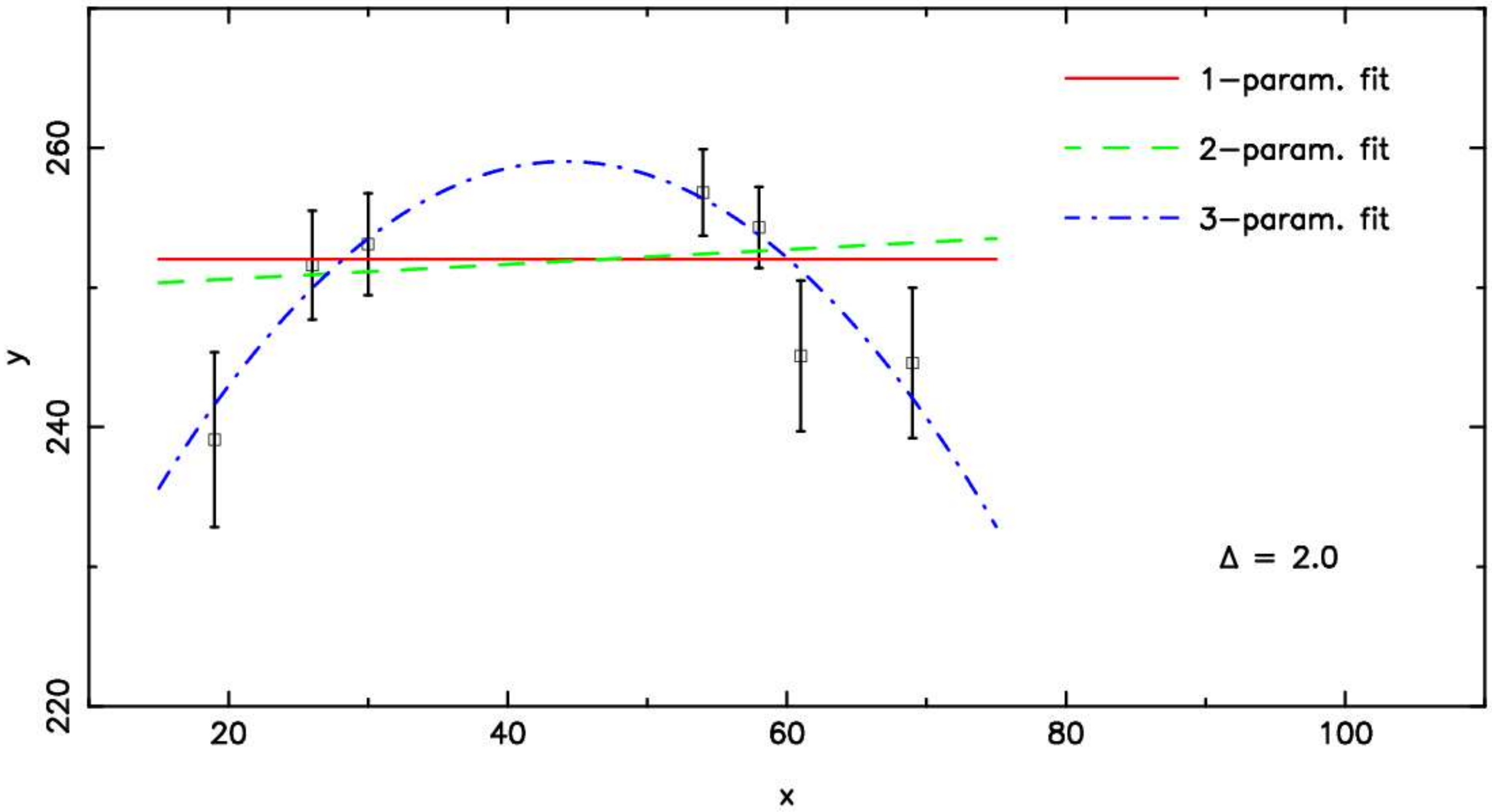
We can compute e.g. $$O_{12} = \frac{\text{prob}(m=1\,|\,d)}{\text{prob}(m=2\,|\,d)} = 7.2$$

$$O_{13} = \frac{\text{prob}(m=1\,|\,d)}{\text{prob}(m=3\,|\,d)} = 172.0$$



What if the error bars were over-estimated?

e.g. divide by factor Δ

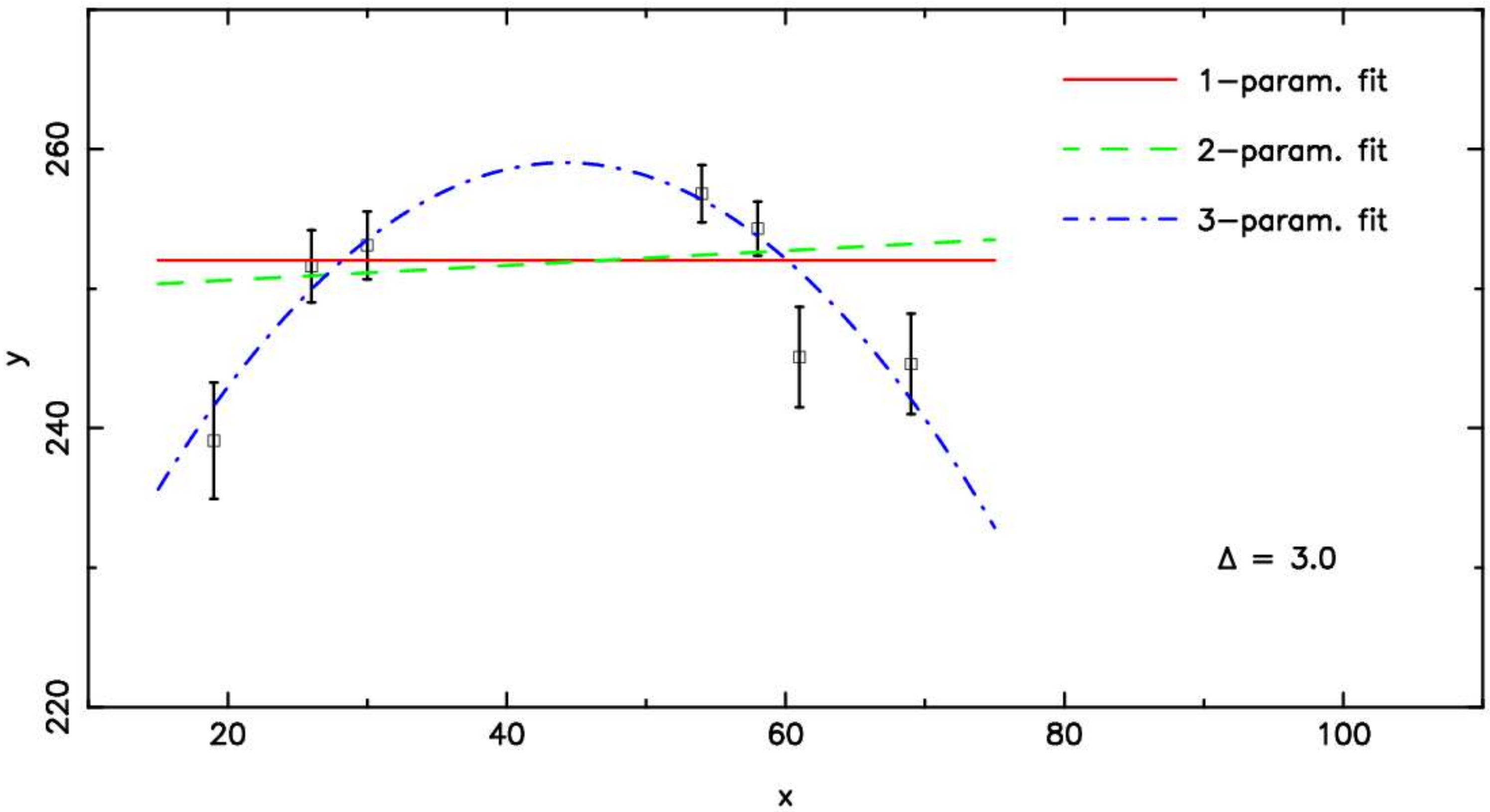


What if the error bars were over-estimated?

e.g. divide by factor $\Delta = 2.0$

$$O_{12} = 6.4$$

$$O_{13} = 5.9$$



What if the error bars were over-estimated?

e.g. divide by factor $\Delta = 3.0$

$$O_{12} = 5.2$$

$$O_{13} = 0.02$$

University of Glasgow

SUPA

Question 16 In this example, when the error bars are reduced by a factor of 3, then $O_{13} << 1$ can be interpreted as

A indicating a much better fit to the quadratic model than the constant model, sufficient that we can justify including an extra 2 parameters

B indicating that, with the smaller error bars, the constant model no longer gives an acceptable fit to the data

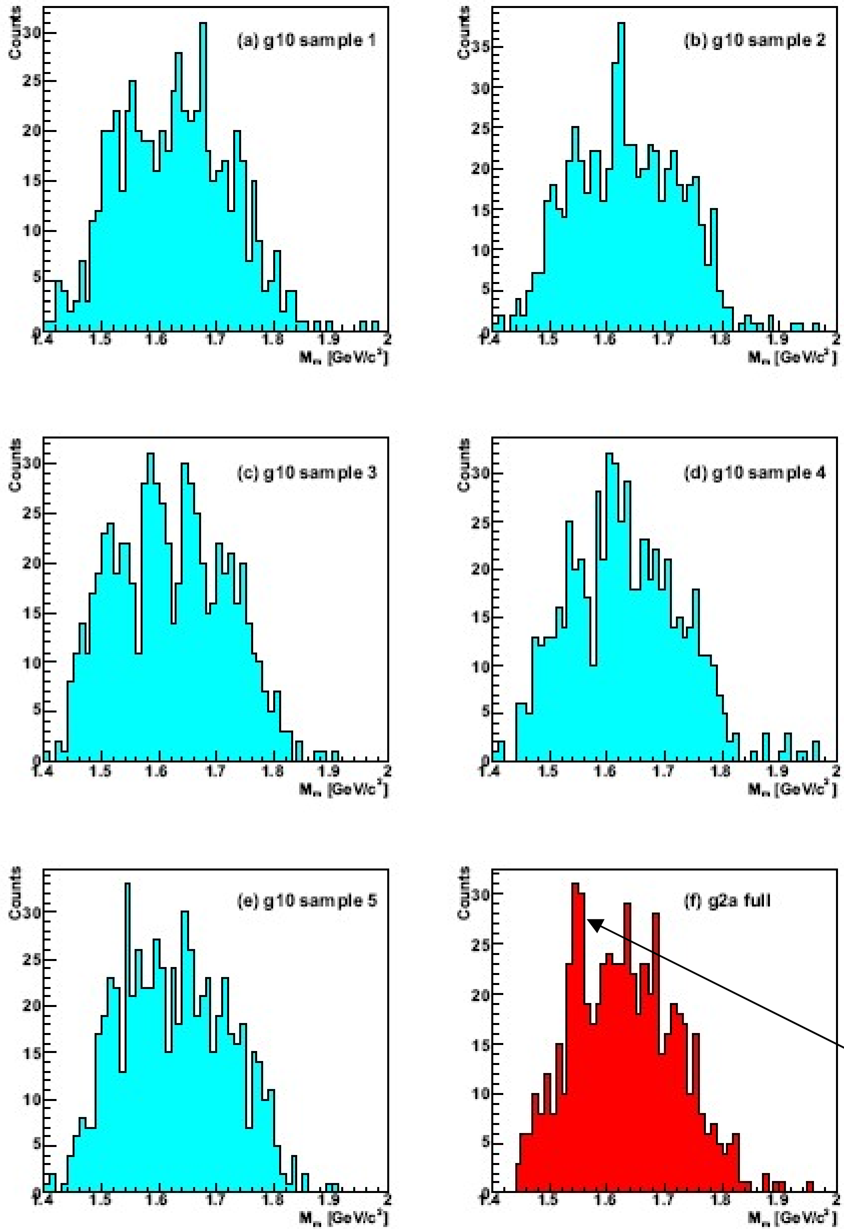
C indicating that the quadratic model is much more likely than the constant model

D all of the above

Ireland et al. (2008)

A Bayesian analysis of pentaquark signals from CLAS data

D.G. Ireland,[1] B. McKinnon,[1] D. Protopopescu,[1] P. Ambrozewicz,[13] M. Anghinolfi,[18] G. Asryan,[38] H. Avakian,[33] H. Bagdasaryan,[28] N. Baillie,[37] J.P. Ball,[3] N.A. Baltzell,[32] V. Batourine,[22] M. Battaglieri,[18] I. Bedlinskiy,[20] M. Bellis,[6] N. Benmouna,[15] B.L. Berman,[15] A.S. Biselli,[6,12] L. Blaszczyk,[14] S. Bouchigny,[19] S. Boiarinov,[33] R. Bradford,[6] D. Branford,[11] W.J. Briscoe,[15] W.K. Brooks,[33] V.D. Burkert,[33] C. Butuceanu,[37] J.R. Calarco,[25] S.L. Careccia,[28] D.S. Carman,[33] L. Casey,[7] S. Chen,[14] L. Cheng,[7] P.L. Cole,[16] P. Collins,[3] P. Coltharp,[14] D. Crabb,[36] V. Crede,[14] N. Dashyan,[38] R. De Masi,[8,19] R. De Vita,[18] E. De Sanctis,[17] P.V. Degtyarenko,[33] A. Deur,[33] R. Dickson,[6] C. Djalali,[32] G.E. Dodge,[28] J. Donnelly,[1] D. Doughty,[9,33] M. Dugger,[3] O.P. Dzyubak,[32] K.S. Egiyan,[38] L. El Fassi,[2] L. Elouadrhiri,[33] P. Eugenio,[14] G. Fedotov,[24] G. Feldman,[15] A. Fradi,[19] H. Funsten,[37] M. Garçon,[8] G. Gavalian,[28] N. Gevorgyan,[38] G.P. Gilfoyle,[31] K.L. Giovanetti,[21] F.X. Girod,[8,33] J.T. Goetz,[4] W. Gohn,[10] A. Gonenc,[13] R.W. Gothe,[32] K.A. Griffioen,[37] M. Guidal,[19] N. Guler,[28] L. Guo,[33] V. Gyurjyan,[33] K. Hafidi,[2] H. Hakobyan,[38] C. Hanretty,[14] N. Hassall,[1] F.W. Hersman,[25] I. Hleiqawi,[27] M. Holtrop,[25] C.E. Hyde-Wright,[28] Y. Ilieva,[15] B.S. Ishkhanov,[24] E.L. Isupov,[24] D. Jenkins,[35] H.S. Jo,[19] J.R. Johnstone,[1] K. Joo,[10] H.G. Juengst,[28] N. Kalantarians,[28] J.D. Kellie,[1] M. Khandaker,[26] W. Kim,[22] A. Klein,[28] F.J. Klein,[7] M. Kossov,[20] Z. Krahn,[6] L.H. Kramer,[13,33] V. Kubarovsky,[33,29] J. Kuhn,[6] S.V. Kuleshov,[20] V. Kuznetsov,[22] J. Lachniet,[28] J.M. Laget,[33] J. Langheinrich,[32] D. Lawrence,[23] K. Livingston,[1] H.Y. Lu,[32] M. MacCormick,[19] N. Markov,[10] P. Mattione,[30] B.A. Mecking,[33] M.D. Mestayer,[33] C.A. Meyer,[6] T. Mibe,[27] K. Mikhailov,[20] M. Mirazita,[17] R. Miskimen,[23] V. Mokeev,[24,33] B. Moreno,[19] K. Moriya,[6] S.A. Morrow,[8,19] M. Moteabbed,[13] E. Munevar,[15] G.S. Mutchler,[30] P. Nadel-Turonski,[15] R. Nasseripour,[32] S. Niccolai,[19] G. Niculescu,[21] I. Niculescu,[21] B.B. Niczyporuk,[33] M.R. Niroula,[28] R.A. Niyazov,[33] M. Nozar,[33] M. Osipenko,[18,24] A.I. Ostrovidov,[14] K. Park,[22] E. Pasyuk,[3] C. Paterson,[1] S. Anefalos Pereira,[17] J. Pierce,[36] N. Pivnyuk,[20] O. Pogorelko,[20] S. Pozdniakov,[20] J.W. Price,[5] S. Procureur,[8] Y. Prok,[36] B.A. Raue,[13,33] G. Ricco,[18] M. Ripani,[18] B.G. Ritchie,[3] F. Ronchetti,[17] G. Rosner,[1] P. Rossi,[17] F. Sabatié,[8] J. Salamanca,[16] C. Salgado,[26] J.P. Santoro,[7] V. Sapunenko,[33] R.A. Schumacher,[6] V.S. Serov,[20] Y.G. Sharabian,[33] D. Sharov,[24] N.V. Shvedunov,[24] L.C. Smith,[36] D.I. Sober,[7] D. Sokhan,[11] A. Stavinsky,[20] S.S. Stepanyan,[22] S. Stepanyan,[33] B.E. Stokes,[14] P. Stoler,[29] S. Strauch,[32] M. Taiuti,[18] D.J. Tedeschi,[32] A. Tkabladze,[15] S. Tkachenko,[28] C. Tur,[32] M. Ungaro,[10] M.F. Vineyard,[34] A.V. Vlassov,[20] D.P. Watts,[11] L.B. Weinstein,[28] D.P. Weygand,[33] M. Williams,[6] E. Wolin,[33] M.H. Wood,[32] A. Yegneswaran,[33] L. Zana,[25] J. Zhang,[28] B. Zhao,[10] and Z.W. Zhao[32]
(The CLAS Collaboration)



Significant peak?

- Model M_0: The spectrum can be described by a 3^{rd} order polynomial in the region of interest. This represents the assumption that there is no new particle. A 3^{rd} order polynomial was employed in the original analysis to model the background shape. This model depends on four parameters.
- Model M_P: The spectrum can be described by a "narrow" Gaussian peak sitting atop a 3^{rd} order polynomial background in the region of interest. "Narrow" in this case meaning that the width is significantly less than the region of interest in the mass spectrum. This model depends on seven parameters.

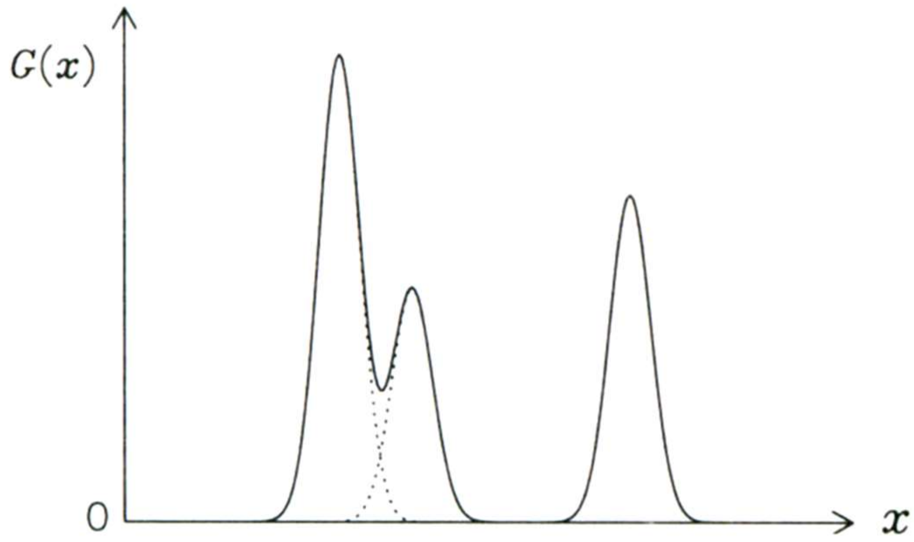
To compare the different models, a ratio of their probabilities in the light of data can be formed:

$$R_E = \frac{P(M_P \mid D)}{P(M_0 \mid D)} = \frac{P(D \mid M_P)}{P(D \mid M_0)} \times \frac{P(M_P)}{P(M_0)},$$

Example from Sivia, Section 4.2: How many spectral lines?

Model: Spectral lines
$$G(x) = \sum_{j=1}^{M} A_j \, f(x, x_j) \, ,$$

where
$$f(x, x_j) = \exp\left[- \frac{(x - x_j)^2}{2W^2} \right]$$

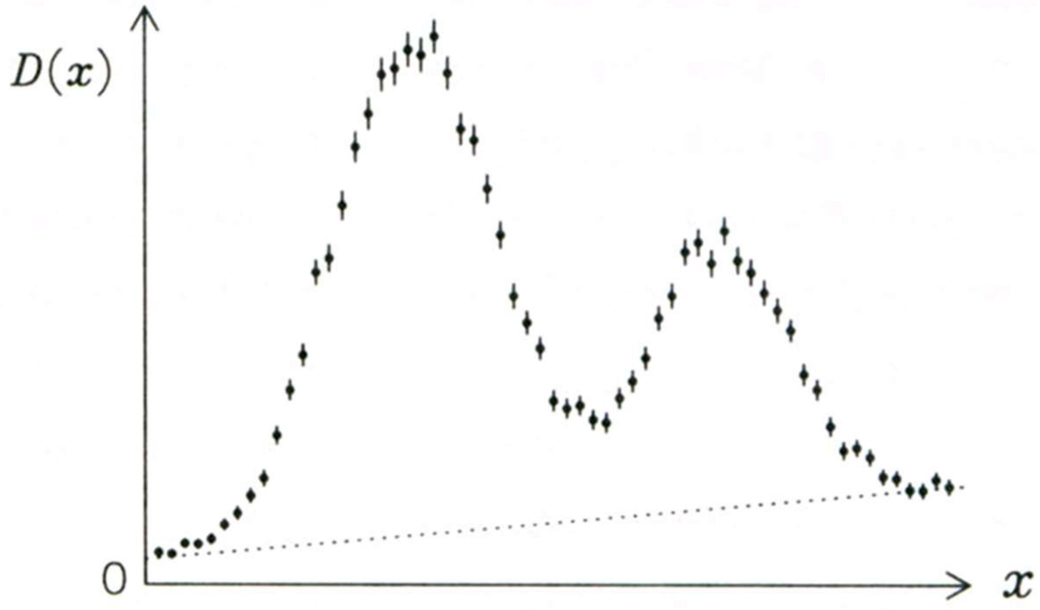


Observed data:

$$D(x_k) = \int G(x)\, R(x_k - x)\,\mathrm{d}x + B(x_k) \quad + \text{ noise}$$

Blurring function
(assumed known)

background



$$\text{prob}(M|\{D_k\},I) = \frac{\text{prob}(\{D_k\}|M,I) \times \text{prob}(M|I)}{\text{prob}(\{D_k\}|I)}$$

Taking a uniform prior on M implies

$$\text{prob}(M|\{D_k\},I) \propto \text{prob}(\{D_k\}|M,I)$$

where $\text{prob}(\{D_k\}|M,I) = \iint\cdots\int \text{prob}\big(\{D_k\},\{A_j,x_j\}|M,I\big)\,\mathrm{d}^M A_j\,\mathrm{d}^M x_j$

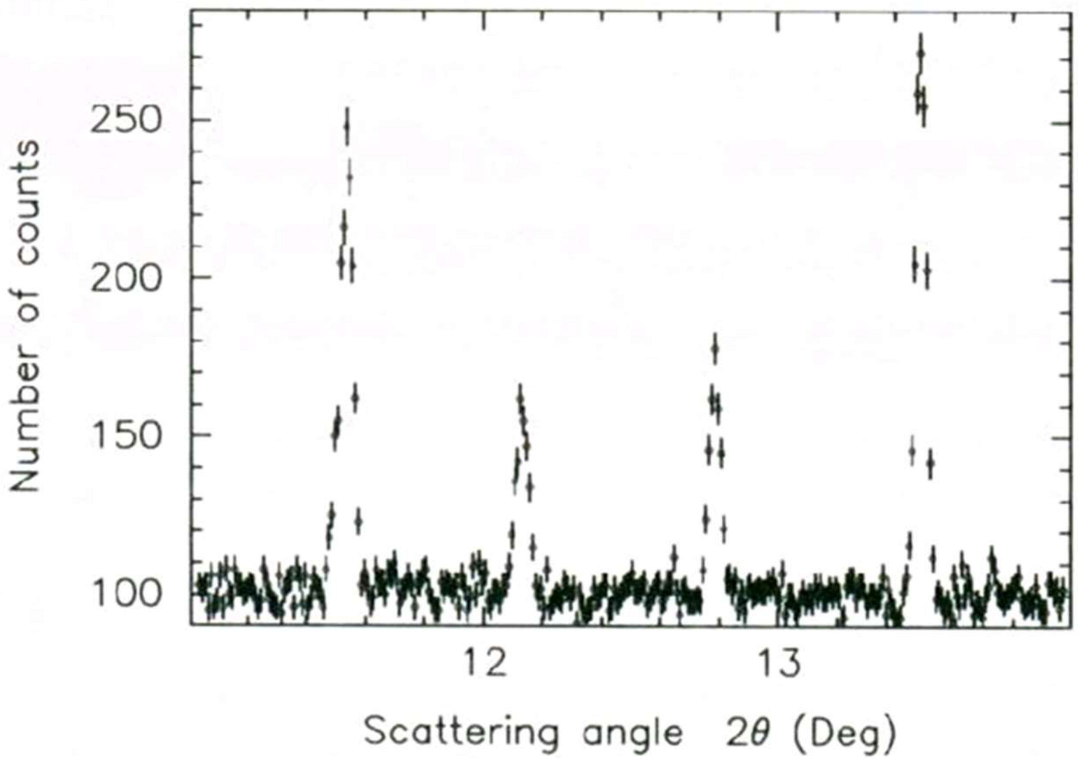
and

$$\text{prob}\big(\{D_k\},\{A_j,x_j\}|M,I\big) = \underbrace{\text{prob}\big(\{D_k\}|\{A_j,x_j\},M,I\big)}_{\text{likelihood}}\ \underbrace{\text{prob}\big(\{A_j,x_j\}|M,I\big)}_{\text{prior}}$$

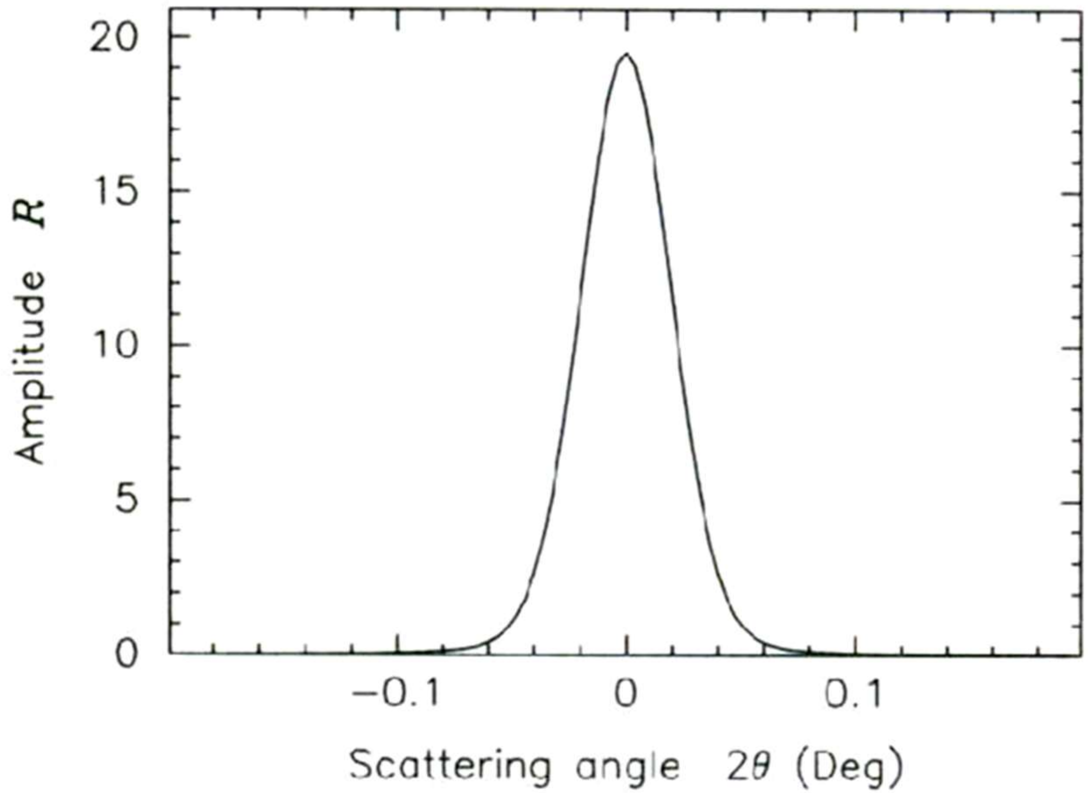
Taking uniform priors on the $\{A_j, x_j\}$ implies

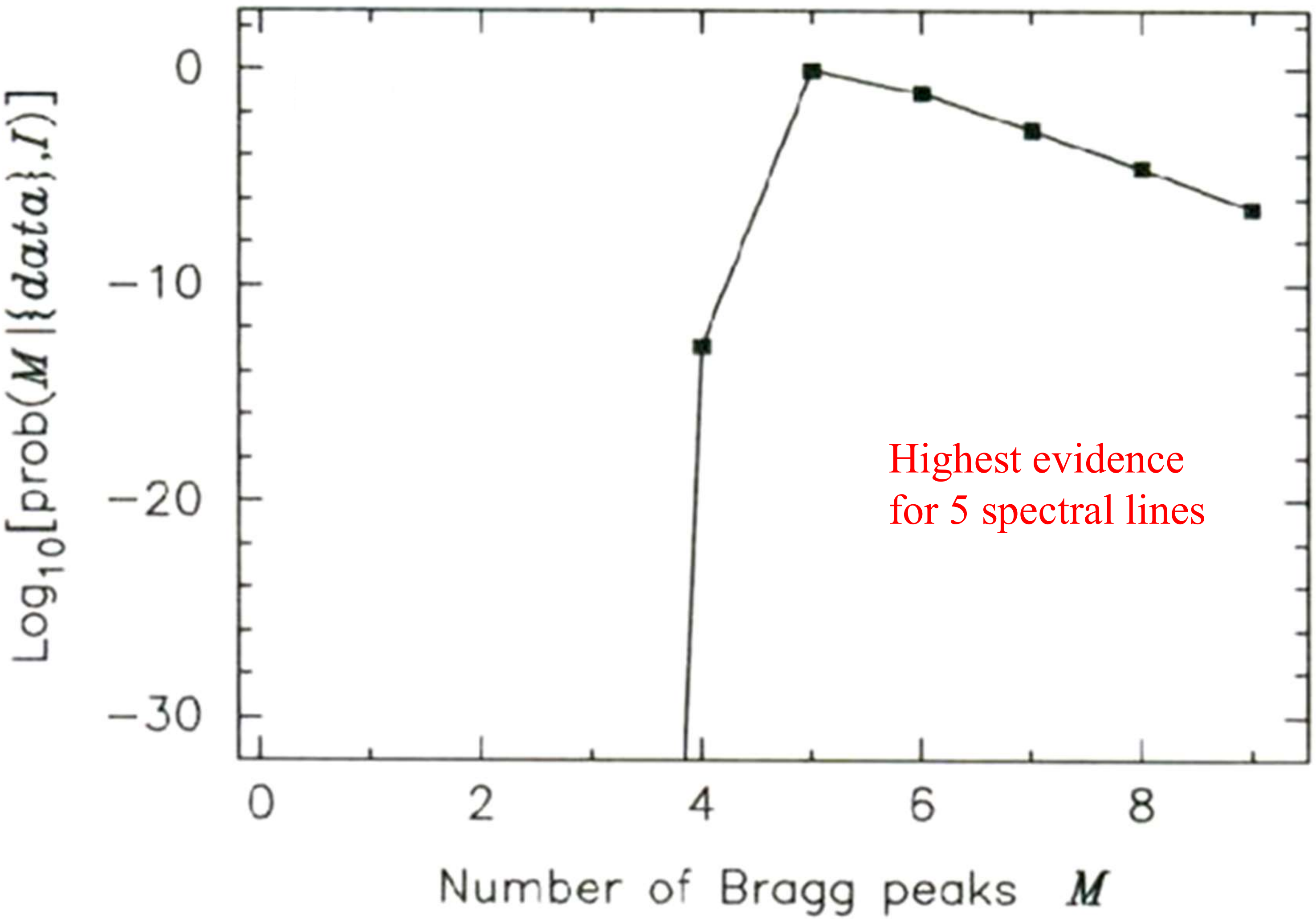
$$\mathrm{prob}(M\,|\,\{D_k\}, I) \propto \left[(x_{\max} - x_{\min})\,A_{\max}\right]^{-M} \iint \cdots \int \exp\left(-\frac{\chi^2}{2}\right) \mathrm{d}^M A_j \, \mathrm{d}^M x_j$$

Simulated example



Assume blurring function known….





Log$_{10}$[prob(M|{$data$},I)]
0
−10
−20
−30
0
2
4
6
8
Number of Bragg peaks M
Highest evidence
for 5 spectral lines

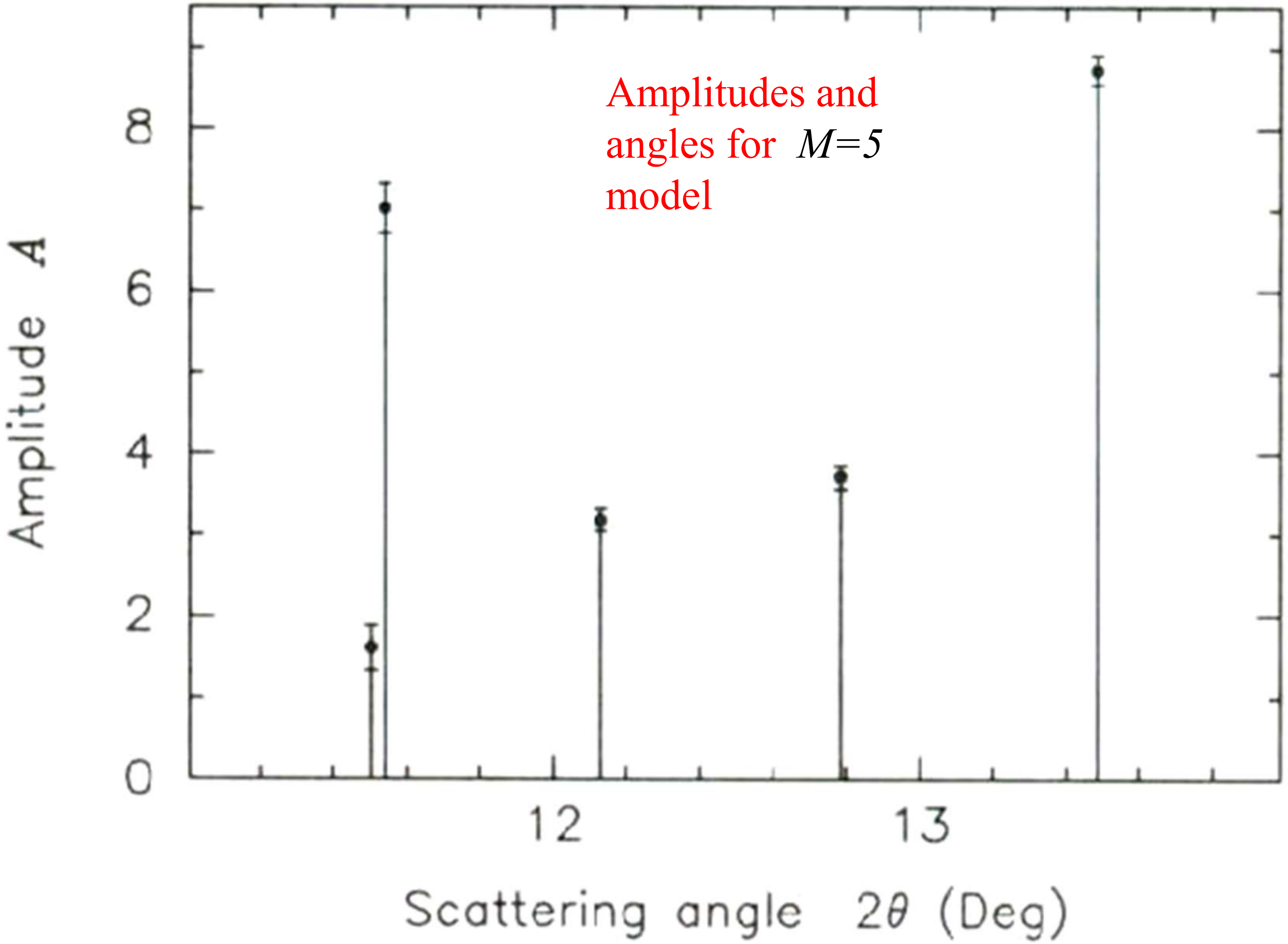
Question 17 The evidence is smaller for $M > 5$ most likely because

A the ML fit is poorer for $M > 5$

B the prior on M is smaller for $M > 5$

C the improvement in the ML fit for $M > 5$ is more than offset by the reduced Occam factor

D none of the above


Amplitudes and
angles for M=5
model
Amplitude A
0
2
4
6
8
12
13
Scattering angle 2θ (Deg)

Taking uniform priors on the $\{A_j, x_j\}$ implies

$$\mathrm{prob}(M \,|\, \{D_k\}, I) \propto \left[(x_{\max} - x_{\min})\, A_{\max}\right]^{-M} \int\!\!\int \cdots \int \exp\left(-\frac{\chi^2}{2}\right) \mathrm{d}^M A_j \, \mathrm{d}^M x_j$$

Evaluating this integral can be a major computational challenge

Approximating the Evidence

$$\text{Evidence} = \underbrace{\int \mathrm{p}(\mathrm{data} \mid \theta, M)\, p(\theta \mid M)\, d\theta}_{\text{Average likelihood, weighted by prior}}$$

- Calculating the evidence can be computationally very costly (e.g. CMBR C_ℓ spectrum in cosmology)

- How to proceed?...

1. Information criteria (see e.g. Liddle 2004, 2007)

1. Laplace and Savage-Dickey approximations (see e.g. Trotta 2005)

3. Nested sampling (Skilling 2004, 2006; http://www.inference.phy.cam.ac.uk/bayesys/)

University of Glasgow

SUPA

Akaike Information Criterion (Akaike 1974)

$$\boxed{\mathrm{AIC} = -2\ln L_{\max} + 2k}$$

Number of parameters in model

- Models with too few parameters give poor fit → first term large
- Models with too many parameters penalised by second term
- MC testing (e.g. Kass & Rafferty 1995): can favour models with too many parameters
- ‘dimensionally inconsistent’
- Can give useful upper limit on number of parameters

Bayesian Information Criterion (Schwarz 1978)

$$\mathrm{BIC} = -2\ln L_{\max} + k\ln N$$

Number of datapoints used in fit

- Approximation to the Bayes factor
- Dimensionally consistent
- If BIC(1) – BIC(2) > 2 $\Rightarrow$ positive evidence favouring Model 2
- If BIC(1) – BIC(2) > 6 $\Rightarrow$ strong evidence favouring Model 2

(Jeffreys 1961; Mukherjee et al. 1998)

Can we do better than the BIC?

- Laplace approximation to the Bayes factor:
  assume posterior well described by a multivariate Gaussian around best-fit parameters

Following Trotta (2005)

$$\ln \frac{\bar{\mathcal{P}}(\boldsymbol{\theta}|\mathbf{D},\mathcal{M})}{\bar{\mathcal{P}}(\boldsymbol{\theta}_*|\mathbf{D},\mathcal{M})} \approx -\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_*)^T\mathbf{C}^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_*)$$

Unnormalised posterior

Best-fit (i.e. ML) parameters

Covariance matrix

Comparing models $\mathcal{M}_0$ and $\mathcal{M}_1$, the Bayes factor B_{01} satisfies

$$\ln B_{01} \approx \mathcal{L}_{01} + \mathcal{C}_{01} + \mathcal{F}_{01}\,,$$

where

$$\mathcal{L}_{01} \quad \equiv \quad \ln \frac{L(\mathbf{D}|\boldsymbol{\theta}_*^{(0)}, \mathcal{M}_0)}{L(\mathbf{D}|\boldsymbol{\theta}_*^{(1)}, \mathcal{M}_1)},$$

Likelihood ratio

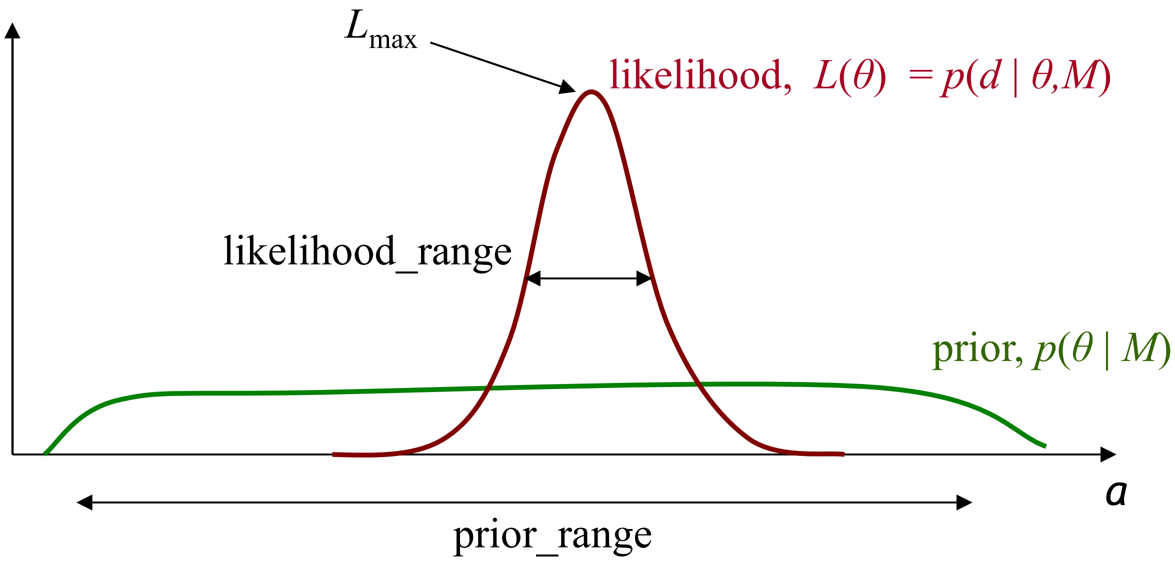
Occam factor

$$\mathcal{C}_{01} \quad \equiv \quad \frac{1}{2}\left(\ln\left[(2\pi)^{d^{(0)}-d^{(1)}}\right] + \ln \frac{\det \mathbf{C}^{(0)}}{\det \mathbf{C}^{(1)}}\right),$$

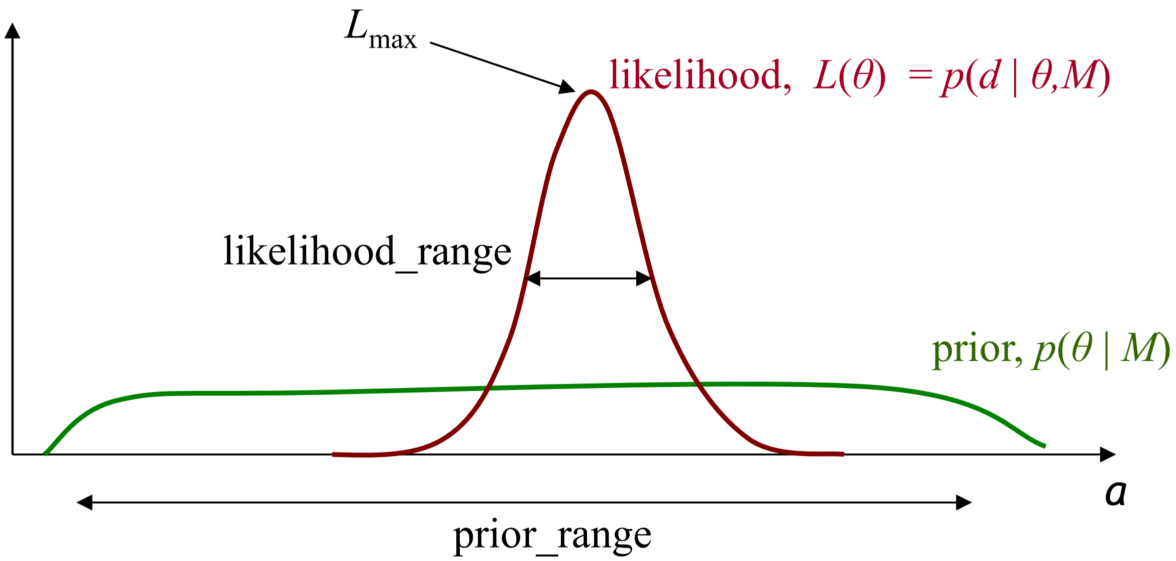
$$\mathcal{F}_{01} \quad \equiv \quad \ln \frac{\boldsymbol{\Delta\theta}^{(1)}}{\boldsymbol{\Delta\theta}^{(0)}}$$

Number of parameters

'Width' of prior



$$p(d \mid M) = \int p(\theta \mid M)\, p(d \mid \theta, M)\, \mathrm{d}\theta \approx L_{\max} \underbrace{\frac{\text{likelihood\_range}}{\text{prior\_range}}}_{\text{the 'Occam factor'}}$$



$$p(d \mid M) = \int p(\theta \mid M)\, p(d \mid \theta, M)\, \mathrm{d}\theta \approx L_{max} \underbrace{\frac{\boxed{\text{likelihood\_range}}}{\boxed{\text{prior\_range}}}}_{\text{the 'Occam factor'}}$$

$\mathcal{L}_{01}$

$\mathcal{F}_{01}$

$\mathcal{C}_{01}$

Comparing models $\mathcal{M}_0$ and $\mathcal{M}_1$, the Bayes factor B_{01} satisfies

$$\ln B_{01} \approx \mathcal{L}_{01} + \mathcal{C}_{01} + \mathcal{F}_{01},$$

where

$$\mathcal{L}_{01} \equiv \ln \frac{L(\mathbf{D}|\boldsymbol{\theta}_*^{(0)}, \mathcal{M}_0)}{L(\mathbf{D}|\boldsymbol{\theta}_*^{(1)}, \mathcal{M}_1)},$$

Likelihood ratio

Occam
factor

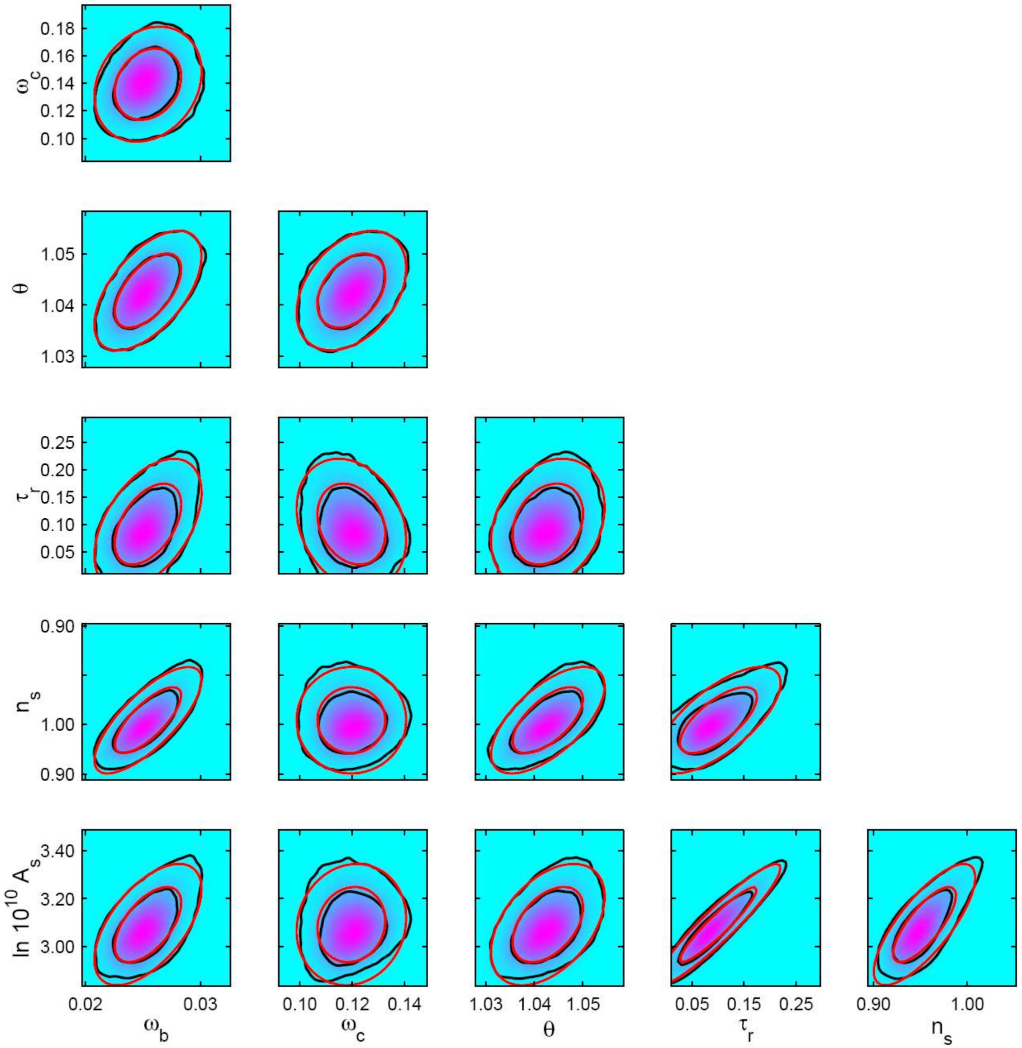
$$\mathcal{C}_{01} \equiv \frac{1}{2}\left(\ln\left[(2\pi)^{d^{(0)}-d^{(1)}}\right] + \ln\frac{\det \mathbf{C}^{(0)}}{\det \mathbf{C}^{(1)}}\right),$$

$$\mathcal{F}_{01} \equiv \ln \frac{\boldsymbol{\Delta\theta}^{(1)}}{\boldsymbol{\Delta\theta}^{(0)}}$$

Number of parameters

'Width' of prior

Testing the Laplace approximation



From Trotta (2005)

Good agreement between (MCMC sampled) posteriors and Laplace approximation.