An Introduction to
Bayesian Data Analysis

Lecture 3

D. S. Sivia
St. John's College
Oxford, England

April 29, 2012

Outline

- The basics
- Parameter estimation I
- Parameter estimation II
- Model selection

- Assigning probabilities
- Non-parametric estimation
- Experimental design
- Least-squares extensions*
- Nested sampling
- Quantification

D.S. Sivia (1996), Data analysis: a Bayesian tutorial, O.U.P.; 2nd edition with *J. Skilling* (2006).

Common simplifying approximations

Want to estimate M parameters $\mathbf{X}$ of a certain model, given N data $\mathbf{D}$.

$$\underbrace{\mathrm{prob}(\mathbf{X}|\mathbf{D},I)}_{Posterior} \propto \underbrace{\mathrm{prob}(\mathbf{D}|\mathbf{X},I)}_{Likelihood} \times \underbrace{\mathrm{prob}(\mathbf{X}|I)}_{Prior}$$

- **Prior**: $\mathrm{prob}(\mathbf{X}|I) = \text{constant}$

  $\Rightarrow \quad \mathrm{prob}(\mathbf{X}|\mathbf{D},I) \propto \mathrm{prob}(\mathbf{D}|\mathbf{X},I)$ — *maximum likelihood*

- Likelihood: $\mathrm{prob}(\mathbf{D}|\mathbf{X},I) \propto \exp\left(-\frac{\chi^2}{2}\right)$

  where $\chi^2 = \sum_{k=1}^{N} \left(\frac{F_k - D_k}{\sigma_k}\right)^2$ and $F_k = f(\mathbf{X},k)$

  — *least-squares*

Least-squares

$$\Rightarrow\ L = \log_e\left[\,\text{prob}(\mathbf{X}|\mathbf{D},I)\,\right] = \text{const} - \frac{\chi^2}{2}$$

- Want: $\boldsymbol{\nabla}L(\mathbf{X}_O) = -\frac{1}{2}\boldsymbol{\nabla}\chi^2(\mathbf{X}_O) = 0$

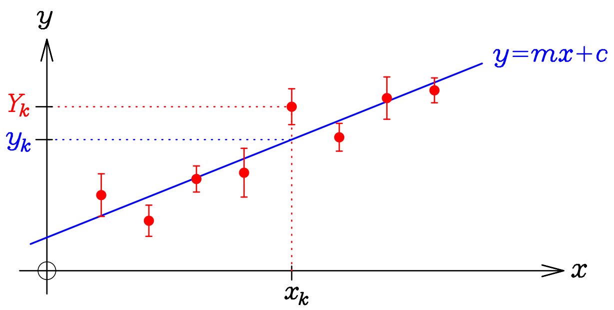
- Linear: $\boldsymbol{\nabla}L = \mathsf{H}\,\mathbf{X} + \boldsymbol{C}$ if $\mathbf{F} = \mathsf{T}\,\mathbf{X} + \boldsymbol{K}$

  $\Rightarrow$ "simple" optimisation problem

- Covariance: $\boldsymbol{\sigma^2} = 2\left[\boldsymbol{\nabla}\boldsymbol{\nabla}\chi^2(\mathbf{X}_O)\right]^{-1}$

- Goodness-of-fit: $\langle\chi^2\rangle \approx N \pm \sqrt{2N}$

Fitting a straight line (1)



$$\chi^2 = \sum_{k=1}^{N}\left(\frac{y_k - Y_k}{\sigma_k}\right)^2 = \sum_{k=1}^{N}\frac{(m\,x_k + c - Y_k)^2}{\sigma_k^2}$$

Fitting a straight line (2)

$$\boldsymbol{\nabla}\chi^2 = \begin{pmatrix} \partial\chi^2/\partial m \\ \partial\chi^2/\partial c \end{pmatrix} = \begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix} \begin{pmatrix} m \\ c \end{pmatrix} - \begin{pmatrix} p \\ q \end{pmatrix} \quad \text{and} \quad \boldsymbol{\nabla\nabla}\chi^2 = \begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix}$$

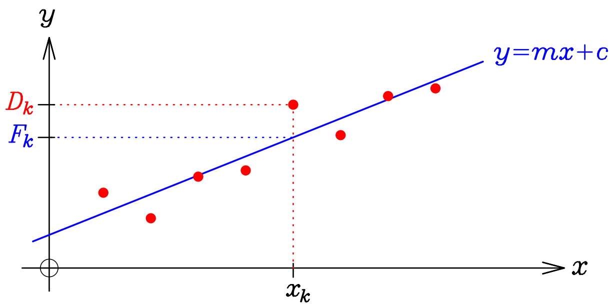
Where $\alpha = \sum w_k x_k^2\,,\ \ \beta = \sum w_k\,,\ \ \gamma = \sum w_k x_k\,,$

$p = \sum w_k x_k Y_k\,,\ \ q = \sum w_k Y_k$ and $w_k = 2/\sigma_k^2$

- $\boldsymbol{\nabla}\chi^2 = 0 \quad \Rightarrow \quad m_\mathrm{o} = \dfrac{\beta p - \gamma q}{\alpha\beta - \gamma^2}$ and $c_\mathrm{o} = \dfrac{\alpha q - \gamma p}{\alpha\beta - \gamma^2}$

- Covariance: $\begin{pmatrix} \sigma_m^2 & \sigma_{mc}^2 \\ \sigma_{mc}^2 & \sigma_c^2 \end{pmatrix} = 2 \begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix}^{-1} = \dfrac{2}{\alpha\beta - \gamma^2} \begin{pmatrix} \beta & -\gamma \\ -\gamma & \alpha \end{pmatrix}$

Data with unknown noise-level (1)



- Conditional likelihood: $\mathrm{prob}(\mathbf{D}|\mathbf{X},\sigma,I) \propto \exp\left(-\dfrac{\chi_\mathrm{o}^2}{2\sigma^2}\right)$

$$\text{where} \quad \chi_\mathrm{o}^2 = \sum_{k=1}^{N} (F_k - D_k)^2$$

Data with unknown noise-level (2)

- Marginal likelihood: $\mathrm{prob}(\mathbf{D}|\mathbf{X},I) = \int_0^\infty \mathrm{prob}(\mathbf{D},\sigma|\mathbf{X},I)\,\mathrm{d}\sigma$

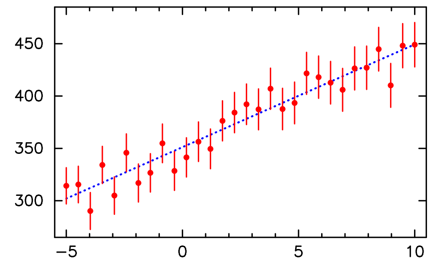
$$= \int_0^\infty \mathrm{prob}(\mathbf{D}|\mathbf{X},\sigma,I)\,\mathrm{prob}(\sigma|I)\,\mathrm{d}\sigma$$

$$\Rightarrow\ L = \log_e\left[\,\mathrm{prob}(\mathbf{X}|\mathbf{D},I)\,\right] = \mathrm{const} - \frac{(N-1)}{2}\log_e\left[\chi_o^2\right]$$

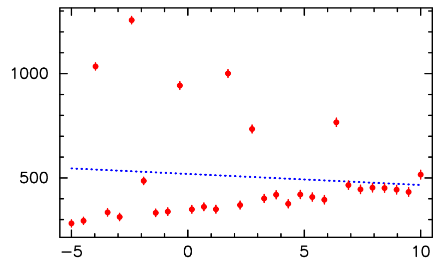
- $\nabla L(\mathbf{X}_O) = 0 \quad \Rightarrow\ \nabla\chi_o^2(\mathbf{X}_O) = 0$

- $\nabla\nabla L(\mathbf{X}_O) = -\dfrac{\nabla\nabla\chi_o^2(\mathbf{X}_O)}{2}\,\dfrac{(N-1)}{\chi_o^2(\mathbf{X}_O)}$

Outliers



$m = 9.8 \pm 0.8\ ,\ \ c = 351.2 \pm 3.8$



$m = -5.3 \pm 0.8\ ,\ \ c = 519.1 \pm 3.8$

$m = -5.3 \pm 10.4\ ,\ \ c = 519.1 \pm 50.7$
(noise scaling)

Gaussian datum with uncertainty

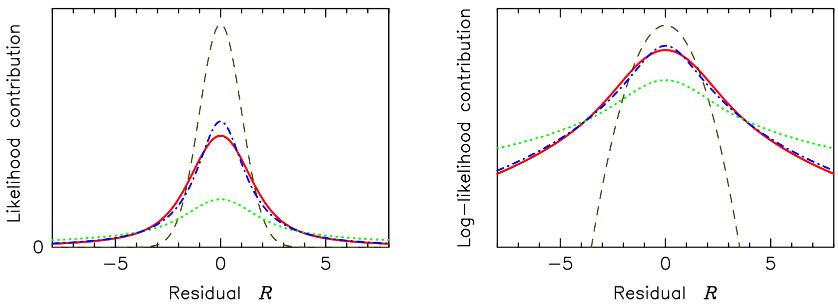
- Gaussian datum: $\text{prob}(D|F,\sigma,I) = \dfrac{e^{-R^2/2}}{\sigma\sqrt{2\pi}}$ where $R = \dfrac{F-D}{\sigma}$

- Lower-bound error-bar: $\text{prob}(\sigma|\sigma_o,I) = \begin{cases} \sigma_o/\sigma^2 & \text{for } \sigma \geqslant \sigma_o \\ 0 & \text{otherwise} \end{cases}$

- Lower-bound likelihood:

$$\begin{aligned}\text{prob}(D|F,\sigma_o,I) &= \int_0^\infty \text{prob}(D,\sigma|F,\sigma_o,I)\,d\sigma \\ &= \int_0^\infty \text{prob}(D|F,\sigma,I)\,\text{prob}(\sigma|\sigma_o,I)\,d\sigma \\ &= \frac{1-e^{-R^2/2}}{R^2\,\sigma_o\sqrt{2\pi}} \qquad \text{where } R = \frac{F-D}{\sigma_o}\end{aligned}$$

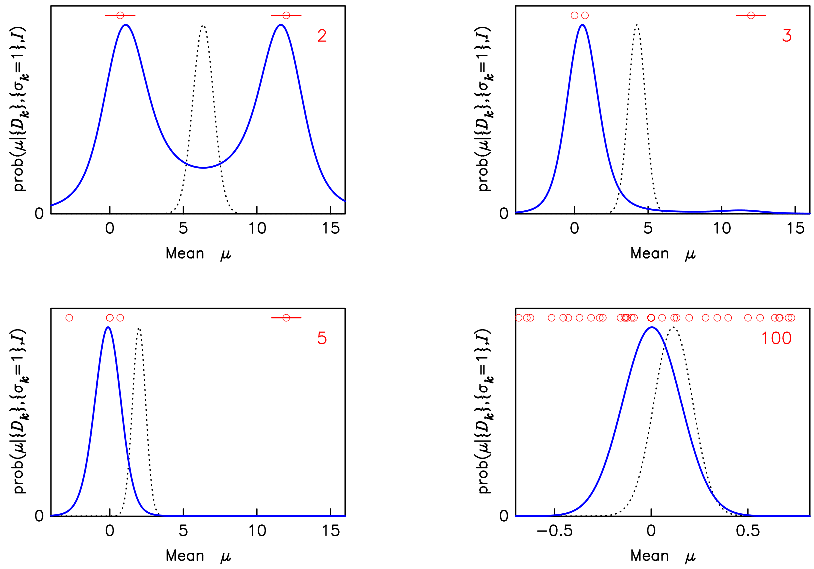
Lower-bound likelihood analysis



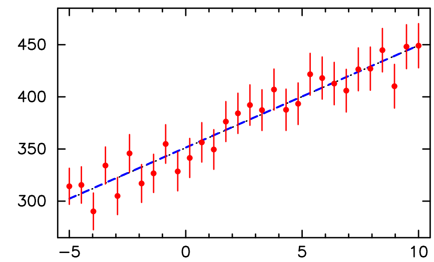
$$L = \log_e\left[\text{prob}(\mathbf{X}|\mathbf{D},I)\right] = \text{const} + \sum_{k=1}^{N} \log_e\left[\frac{1-e^{-R_k^2/2}}{R_k^2}\right], \quad R_k = \frac{F_k - D_k}{\sigma_k}$$

$$\text{Instead of} \quad L = \text{const} - \frac{1}{2}\sum_{k=1}^{N} R_k^2 \quad [\text{least}-\text{squares}]$$

Dealing with outliers (1)

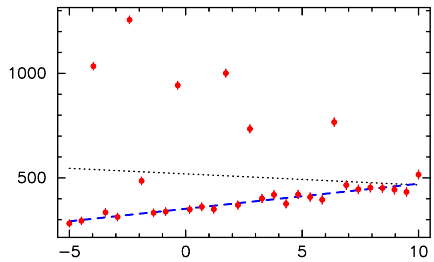


Dealing with outliers (2)



$m = 9.8 \pm 1.2$, $c = 351.4 \pm 5.9$

$m = 9.8 \pm 0.8$, $c = 351.2 \pm 3.8$
(least-squares)

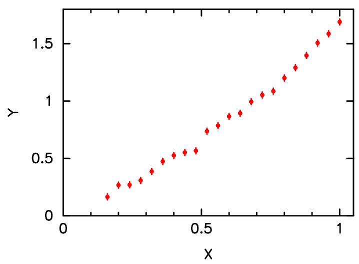


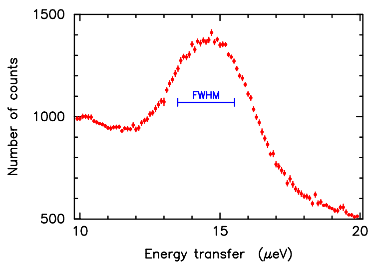
$m = 12.0 \pm 1.4$, $c = 352.1 \pm 7.0$

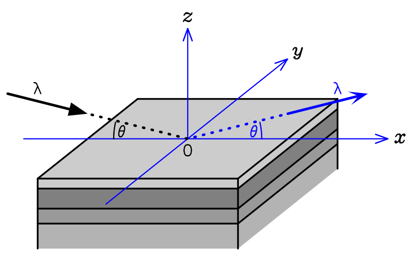
$m = -5.3 \pm 0.8$, $c = 519.1 \pm 3.8$
(least-squares)

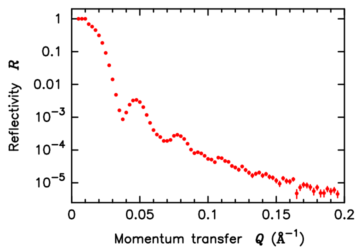
Propagation of errors

Model selection









The story of Mr. A and Mr. B

Mr. A has a theory; Mr. B also has a theory, but with an adjustable parameter λ. Whose theory should we prefer on the basis of data $\mathbf{D}$?

[Jeffreys, 1939, Gull 1988]

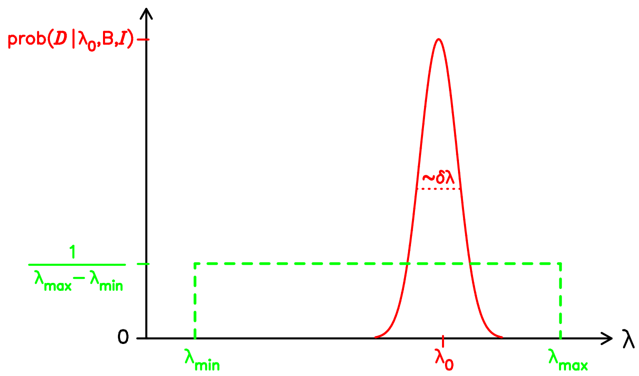
- Posterior ratio: $$\frac{\text{prob}(A|\mathbf{D},I)}{\text{prob}(B|\mathbf{D},I)} \begin{cases} \gg 1 & \text{prefer A} \\ \approx 1 & \text{undecided} \\ \ll 1 & \text{prefer B} \end{cases}$$

$$= \frac{\text{prob}(\mathbf{D}|A,I)}{\text{prob}(\mathbf{D}|B,I)} \times \frac{\text{prob}(A|I)}{\text{prob}(B|I)} \quad \text{(Bayes')}$$

  - Need predictions for data from both A and B.
    But, for B, need λ!

The story of Mr. A and Mr. B

$$\text{prob}(\mathbf{D}|B,I) = \int \text{prob}(\mathbf{D},\lambda|B,I)\,d\lambda = \int \text{prob}(\mathbf{D}|\lambda,B,I)\ \text{prob}(\lambda|B,I)\,d\lambda$$



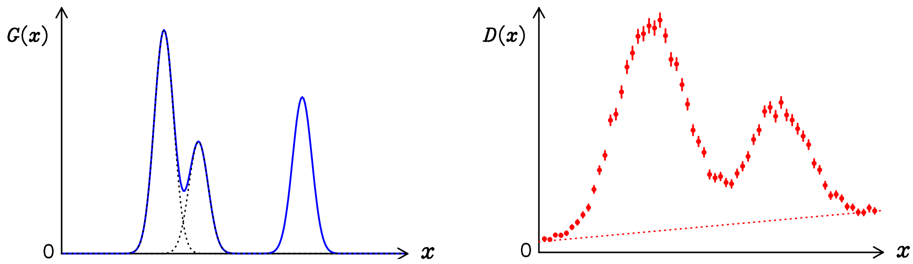
$$\text{prob}(\mathbf{D}|B,I) = \frac{1}{\lambda_{\max}-\lambda_{\min}} \int \text{prob}(\mathbf{D}|\lambda,B,I)\,d\lambda \approx \frac{\text{prob}(\mathbf{D}|\lambda_o,B,I)\,\delta\lambda}{\lambda_{\max}-\lambda_{\min}}$$

The story of Mr. A and Mr. B

$$\underbrace{\frac{\text{prob}(\text{A}\,|\,\mathbf{D},I)}{\text{prob}(\text{B}\,|\,\mathbf{D},I)}}_{\text{Posterior ratio}} = \underbrace{\frac{\text{prob}(\text{A}\,|\,I)}{\text{prob}(\text{B}\,|\,I)}}_{\text{Prior ratio}} \times \frac{\text{prob}(\mathbf{D}\,|\,\text{A},I)}{\text{prob}(\mathbf{D}\,|\,\text{B},I)}$$

$$\approx \frac{\text{prob}(\text{A}\,|\,I)}{\text{prob}(\text{B}\,|\,I)} \times \underbrace{\frac{\text{prob}(\mathbf{D}\,|\,\text{A},I)}{\text{prob}(\mathbf{D}\,|\,\lambda_o,\text{B},I)}}_{\text{Best}-\text{fit likelihood ratio}} \times \underbrace{\frac{\lambda_{\max}-\lambda_{\min}}{\delta\lambda}}_{\text{“Ockham factor”}}$$

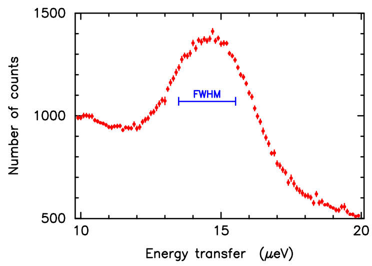
How many lines are there?

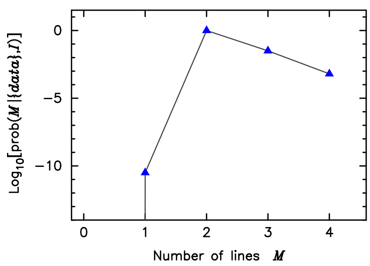


- With a least-squares likelihood, and the earlier approximations,

$$\text{prob}(M\,|\,\mathbf{D},I) \propto \frac{\text{prob}(M\,|\,I)\;\; M!\,(4\pi)^M}{\left[(x_{\max}-x_{\min})A_{\max}\right]^M\sqrt{\det(\nabla\nabla\chi^2)}}\;\exp\left(-\frac{\chi^2_{\min}}{2}\right)$$

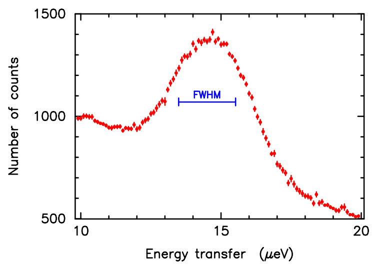
Test example (1)

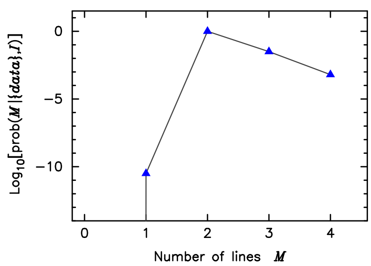


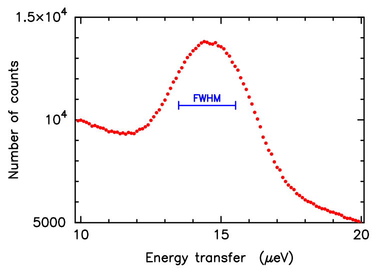


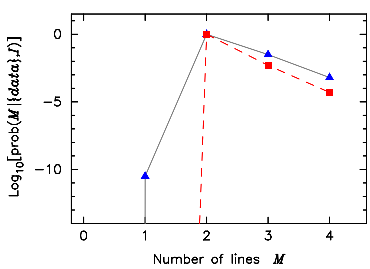
- $E_1 = 13.98 \pm 0.03\,\mu\text{eV}$ and $A_1 = 953 \pm 20$
- $E_2 = 15.47 \pm 0.02\,\mu\text{eV}$ and $A_2 = 1035 \pm 20$
- Intrinsic FWHM: $1.03 \pm 0.08\,\mu\text{eV}$

Test example (2)

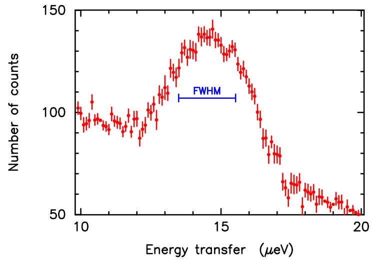


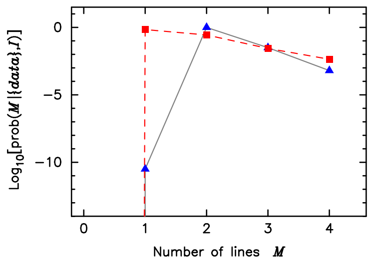




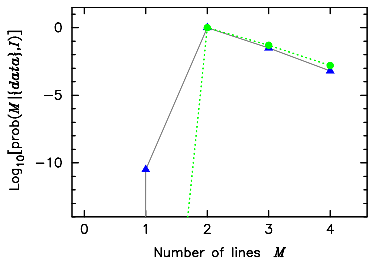


Test example (3)

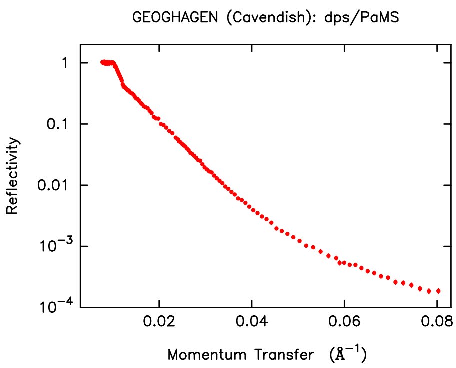


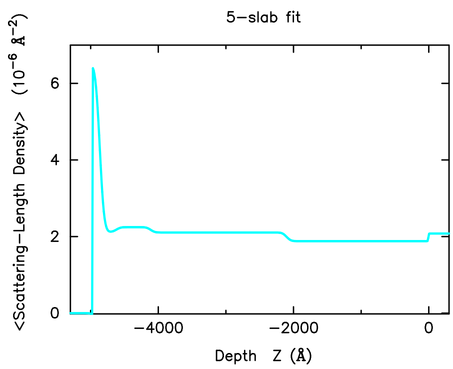


Intrinsic FWHM $= 1.0\,\mu\mathrm{eV}$

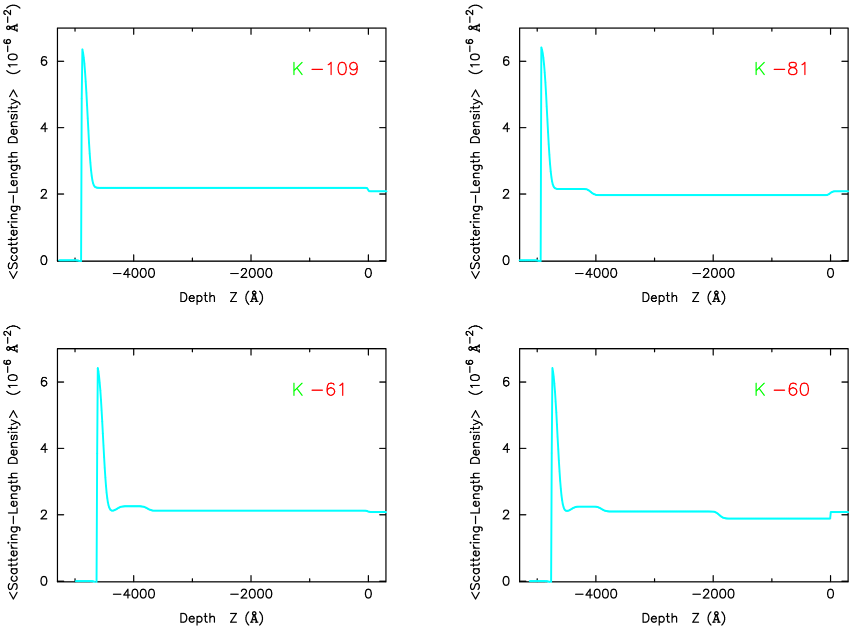


Reflectivity: bi-polymer data

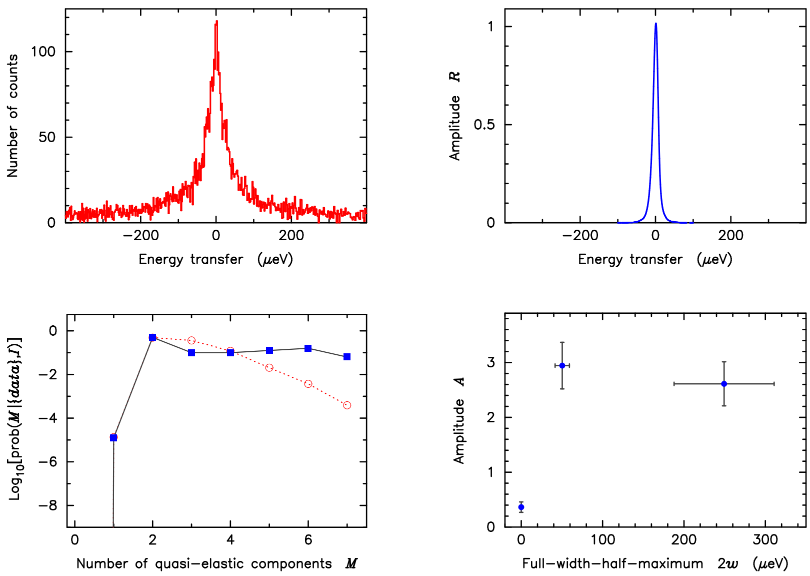




Reflectivity: model selection



Interlude: what not to compute



Model selection: a summary

- Previously, $\mathbf{X} \equiv M$ parameters of a given model (or hypothesis) H.

  $i.e.$ $\quad \text{prob}(\mathbf{X}|\mathbf{D}, I) \equiv \text{prob}(\mathbf{X}|\mathbf{D}, \text{H}, I)$

- Competing hypotheses H_1 and H_2; which one is better?

$$\underbrace{\frac{\text{prob}(\text{H}_1|\mathbf{D}, I)}{\text{prob}(\text{H}_2|\mathbf{D}, I)}}_{\textit{Posterior ratio}} \begin{cases} \gg 1 & \text{prefer } \text{H}_1 \\ \approx 1 & \text{not sure} \\ \ll 1 & \text{prefer } \text{H}_2 \end{cases}$$

$$= \underbrace{\frac{\text{prob}(\mathbf{D}|\text{H}_1, I)}{\text{prob}(\mathbf{D}|\text{H}_2, I)}}_{\textit{Evidence ratio}} \times \underbrace{\frac{\text{prob}(\text{H}_1|I)}{\text{prob}(\text{H}_2|I)}}_{\textit{Prior ratio}} \qquad \text{(Bayes')}$$

Model selection: the evidence

- $\text{prob}(\mathbf{D}|\text{H}, I)$ is just the normalisation constant in Bayes' theorem for the posterior pdf of $\mathbf{X}$!

  - $\text{prob}(\mathbf{X}|\mathbf{D}, \text{H}, I) = \dfrac{\text{prob}(\mathbf{D}|\mathbf{X}, \text{H}, I) \times \text{prob}(\mathbf{X}|\text{H}, I)}{\text{prob}(\mathbf{D}|\text{H}, I)}$

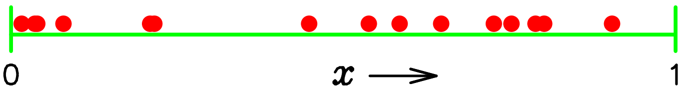
  - $\text{prob}(\mathbf{D}|\text{H}, I) = \displaystyle\int\!\!\int \cdots \int \text{prob}(\mathbf{D}|\mathbf{X}, \text{H}, I)\ \text{prob}(\mathbf{X}|\text{H}, I)\ \text{d}^M\mathbf{X}$

  - Prior-weighted 'average likelihood'.

- Since $\text{prob}(\mathbf{X}|\text{H}, I)$ must now be normalised properly, need to think about a suitable prior-range for $\mathbf{X}$.

Testing for uniformity

Given a set of data comprising x-values between 0 and 1, $\{x_k\}$,

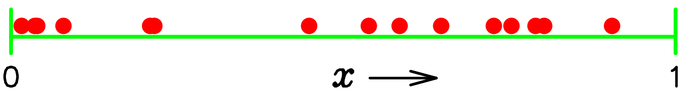


do they support the hypothesis of a uniform underlying distribution?

$$\text{prob}(\text{flat}|\{x_k\}, I) \begin{cases} \to 1 & \text{uniform} \\ \approx 1/2 & \text{not sure} \\ \to 0 & \text{not uniform} \end{cases}$$

Testing for uniformity

Given a set of data comprising x-values between 0 and 1, $\{x_k\}$,



do they support the hypothesis of an underlying uniform distribution?

$$\text{prob}(\text{flat}|\{x_k\}, I) = \frac{\text{prob}(\{x_k\}|\text{flat}, I) \times \text{prob}(\text{flat}|I)}{\text{prob}(\{x_k\}|I)}$$

$$= \frac{1 \times \frac{1}{2}}{\text{prob}(\{x_k\}, \text{flat}|I) + \text{prob}\big(\{x_k\}, \overline{\text{flat}}|I\big)}$$

$$= \Big[1 + \text{prob}\big(\{x_k\}|\overline{\text{flat}}, I\big)\Big]^{-1}$$

Conclusions

- The Bayesian approach to probability theory gives a logical and unified view of data analysis.
  - It provides the justification for many conventional procedures, and gives improved prescriptions when they fail.

- *"La théorie des probabilités n'est que le bon sens reduit au calcul."*

  – Laplace

- *"Data analysis is simply a dialogue with the data."*

  – Gull