An Introduction to
Bayesian Data Analysis

Lecture 3

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

April 29, 2012

<table>
<thead>
<tr>
<th>Outline</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common simplifying approximations</td>
<td>3</td>
</tr>
<tr>
<td>Least-squares</td>
<td>4</td>
</tr>
<tr>
<td>Fitting a straight line (1)</td>
<td>5</td>
</tr>
<tr>
<td>Fitting a straight line (2)</td>
<td>6</td>
</tr>
<tr>
<td>Data with unknown noise-level (1)</td>
<td>7</td>
</tr>
<tr>
<td>Data with unknown noise-level (2)</td>
<td>8</td>
</tr>
<tr>
<td>Outliers</td>
<td>9</td>
</tr>
<tr>
<td>Gaussian datum with uncertainty</td>
<td>10</td>
</tr>
<tr>
<td>Lower-bound likelihood analysis.</td>
<td>11</td>
</tr>
<tr>
<td>Dealing with outliers (1)</td>
<td>12</td>
</tr>
<tr>
<td>Dealing with outliers (2)</td>
<td>13</td>
</tr>
<tr>
<td>Propagation of errors</td>
<td>14</td>
</tr>
<tr>
<td>Model selection</td>
<td>15</td>
</tr>
<tr>
<td>The story of Mr. A and Mr. B.</td>
<td>16</td>
</tr>
<tr>
<td>The story of Mr. A and Mr. B.</td>
<td>17</td>
</tr>
<tr>
<td>The story of Mr. A and Mr. B.</td>
<td>18</td>
</tr>
<tr>
<td>How many lines are there?</td>
<td>19</td>
</tr>
<tr>
<td>Test example (1)</td>
<td>20</td>
</tr>
<tr>
<td>Test example (2)</td>
<td>21</td>
</tr>
<tr>
<td>Test example (3)</td>
<td>22</td>
</tr>
<tr>
<td>Reflectivity: bi-polymer data</td>
<td>23</td>
</tr>
<tr>
<td>Reflectivity: model selection</td>
<td>24</td>
</tr>
<tr>
<td>Interlude: what not to compute.</td>
<td>25</td>
</tr>
<tr>
<td>Model selection: a summary</td>
<td>26</td>
</tr>
<tr>
<td>Model selection: the evidence</td>
<td>27</td>
</tr>
<tr>
<td>Testing for uniformity</td>
<td>28</td>
</tr>
<tr>
<td>Testing for uniformity</td>
<td>29</td>
</tr>
<tr>
<td>Conclusions</td>
<td>30</td>
</tr>
</tbody>
</table>
Common simplifying approximations

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

$$
\text{Posterior} \propto \text{Likelihood} \times \text{Prior}
$$

- **Prior**: $\text{prob}(X|I) = \text{constant}$
  \[\Rightarrow \text{prob}(X|D, I) \propto \text{prob}(D|X, I)\]  — maximum likelihood

- **Likelihood**: $\text{prob}(D|X, I) \propto \exp\left(-\frac{\chi^2}{2}\right)$
  
  where $\chi^2 = \sum_{k=1}^{N} \left(\frac{F_k - D_k}{\sigma_k}\right)^2$ and $F_k = f(X, k)$  — least-squares
Least-squares

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

- Want: \( \nabla L(X_0) = -\frac{1}{2} \nabla \chi^2(X_0) = 0 \)

- Linear: \( \nabla L = H X + C \) if \( F = T X + K \)
  \[ \Rightarrow \text{“simple” optimisation problem} \]

- Covariance: \( \sigma^2 = 2 \left[ \nabla \nabla \chi^2(X_0) \right]^{-1} \)

- 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)

\[ \nabla \chi^2 = \left( \frac{\partial \chi^2}{\partial m}, \frac{\partial \chi^2}{\partial c} \right) = \left( \frac{\alpha}{\beta}, \frac{\gamma}{\beta} \right) \]

and

\[ \nabla \nabla \chi^2 = \left( \frac{\alpha}{\beta}, \frac{\gamma}{\beta} \right) \]

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

\[ \begin{align*}
p &= \sum w_k x_k y_k, \quad q = \sum w_k y_k \quad \text{and} \quad w_k = 2/\sigma_k^2
\end{align*} \]

\[ \nabla \chi^2 = 0 \Rightarrow m_o = \frac{\beta p - \gamma q}{\alpha \beta - \gamma^2} \quad \text{and} \quad c_o = \frac{\alpha q - \gamma p}{\alpha \beta - \gamma^2} \]

Covariance:

\[ \begin{bmatrix} \sigma_m^2 & \sigma_{mc}^2 \\ \sigma_{mc}^2 & \sigma_c^2 \end{bmatrix} = 2 \begin{bmatrix} \alpha & \gamma \\ \gamma & \beta \end{bmatrix}^{-1} = \frac{2}{\alpha \beta - \gamma^2} \begin{bmatrix} \beta & -\gamma \\ -\gamma & \alpha \end{bmatrix} \]

Data with unknown noise-level (1)

Conditional likelihood:

\[ \text{prob}(D | X, \sigma, I) \propto \exp \left( -\frac{\chi_o^2}{2\sigma^2} \right) \]

where \( \chi_o^2 = \sum_{k=1}^N (F_k - D_k)^2 \)
**Data with unknown noise-level (2)**

- **Marginal likelihood:** \( \text{prob}(D|X,I) = \int_0^\infty \text{prob}(D,\sigma|X,I) \, d\sigma \)
  \[ = \int_0^\infty \text{prob}(D,X,\sigma,I) \text{prob}(\sigma|I) \, d\sigma \]
  \[ \Rightarrow L = \log_e[\text{prob}(X|D,I)] = \text{const} - \frac{(N-1)}{2} \log_e[\chi^2_o] \]

- \( \nabla L(X_o) = 0 \quad \Rightarrow \quad \nabla \chi^2_o(X_o) = 0 \)

- \( \nabla \nabla L(X_o) = -\frac{\nabla \nabla \chi^2_o(X_o)}{2} \frac{(N-1)}{\chi^2_o(X_o)} \)

---

**Outliers**

\( m = 9.8 \pm 0.8, \quad c = 351.2 \pm 3.8 \)

\( m = -5.3 \pm 0.8, \quad c = 519.1 \pm 3.8 \)

\( m = -5.3 \pm 10.4, \quad c = 519.1 \pm 50.7 \)

(noise scaling)
**Gaussian datum with uncertainty**

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

- **Lower-bound error-bar:** \( \text{prob}(\sigma|\sigma_0, I) = \begin{cases} \frac{\sigma_0}{\sigma^2} & \text{for } \sigma \geq \sigma_0 \\ 0 & \text{otherwise} \end{cases} \)

- **Lower-bound likelihood:**
  \[
  \text{prob}(D|F, \sigma_0, I) = \int_0^\infty \text{prob}(D, \sigma|F, \sigma_0, I) \, d\sigma \\
  = \int_0^\infty \text{prob}(D|F, \sigma, I) \text{prob}(\sigma|\sigma_0, I) \, d\sigma \\
  = \frac{1 - e^{-R^2/2}}{R^2 \sigma_0 \sqrt{2\pi}} \quad \text{where} \quad R = \frac{F-D}{\sigma_0}
  \]

**Lower-bound likelihood analysis**

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

Instead of \( L = \text{const} - \frac{1}{2} \sum_{k=1}^{N} R_k^2 \) [least-squares]
Dealing with outliers (1)

\[ m = 9.8 \pm 1.2 , \ c = 351.4 \pm 5.9 \]

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

Dealing with outliers (2)

\[ m = 12.0 \pm 1.4 , \ c = 352.1 \pm 7.0 \]

\[ m = -5.3 \pm 0.8 , \ c = 519.1 \pm 3.8 \]
(least-squares)
The story of Mr. A and Mr. B

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

[Jeffreys, 1939, Gull 1988]

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

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

- Need predictions for data from both A and B.
  But, for B, need $\lambda$!

IN2P3/CNRS School of Statistics (Autrans 2012): Bayesian Lecture 3 16 / 30

The story of Mr. A and Mr. B

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

\[ \text{prob}(D|B, I) = \frac{1}{\lambda_{\text{max}} - \lambda_{\text{min}}} \int \text{prob}(D|\lambda, B, I) \, d\lambda \approx \frac{\text{prob}(D|\lambda_0, B, I) \, \delta\lambda}{\lambda_{\text{max}} - \lambda_{\text{min}}} \]

IN2P3/CNRS School of Statistics (Autrans 2012): Bayesian Lecture 3 17 / 30
The story of Mr. A and Mr. B

\[
\frac{\text{prob}(A | D, I)}{\text{prob}(B | D, I)} = \frac{\text{prob}(A | I)}{\text{prob}(B | I)} \times \frac{\text{prob}(D | A, I)}{\text{prob}(D | B, I)}
\]

Posterior ratio

Prior ratio

\[
\approx \frac{\text{prob}(A | I)}{\text{prob}(B | I)} \times \frac{\text{prob}(D | A, I)}{\text{prob}(D | \lambda_0, B, I)} \times \frac{\lambda_{\text{max}} - \lambda_{\text{min}}}{\delta \lambda}
\]

Best-fit likelihood ratio

“Ockham factor”

---

How many lines are there?

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

\[
\text{prob}(M | D, I) \propto \frac{\text{prob}(M | I)}{\left[ (x_{\text{max}} - x_{\text{min}}) A_{\text{max}} \right]^M} \sqrt{\text{det}(\nabla \nabla \chi^2)} \exp \left( -\frac{\chi^2_{\text{min}}}{2} \right)
\]
- $E_1 = 13.98 \pm 0.03 \, \mu eV$ and $A_1 = 953 \pm 20$
- $E_2 = 15.47 \pm 0.02 \, \mu eV$ and $A_2 = 1035 \pm 20$
- Intrinsic FWHM: $1.03 \pm 0.08 \, \mu eV$
Test example (3)

Reflectivity: bi-polymer data

Intrinsic FWHM = 1.0 µeV
Reflectivity: model selection

Interlude: what not to compute
Model selection: a summary

- Previously, $X \equiv M$ parameters of a given model (or hypothesis) $H$.
  
i.e. $\text{prob}(X|D, I) \equiv \text{prob}(X|D, H, I)$

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

\[
\begin{array}{c|c|c}
\text{Posterior ratio} & \approx 1 & \ll 1, \text{ prefer } H_2 \\
\text{Evidence ratio} & \gg 1, \text{ prefer } H_1 & \ll 1, \text{ not sure}
\end{array}
\]

\[
\underbrace{\frac{\text{prob}(D|H_1, I)}{\text{prob}(D|H_2, I)}} \times \frac{\text{prob}(H_1|I)}{\text{prob}(H_2|I)} = \text{(Bayes') Evidence ratio} \times \text{Prior ratio}
\]

Model selection: the evidence

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

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

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

- Prior-weighted ‘average likelihood’.

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

IN2P3/CNRS School of Statistics (Autrans 2012): Bayesian Lecture 3
Testing for uniformity
Given a set of data comprising $x$-values between 0 and 1, $\{x_k\}$,

\[
\text{do they support the hypothesis of a uniform underlying distribution?}
\]

\[
\begin{align*}
\text{prob}(\text{flat}|\{x_k\}, I) &= \begin{cases} 
1 & \text{uniform} \\
1/2 & \text{not sure} \\
0 & \text{not uniform}
\end{cases}
\end{align*}
\]

\[
\text{prob}(\{x_k\}|\text{flat}, I) \times \text{prob}(\text{flat}|I) = \frac{1 \times 1/2}{\text{prob}(\{x_k\}, \text{flat} | I) + \text{prob}(\{x_k\}, \overline{\text{flat}} | I)} = \left[ 1 + \text{prob}(\{x_k\}|\overline{\text{flat}}, I) \right]^{-1}
\]
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