Outliers

Four Bayesian approaches to deal with erratic data

1. A conservative model
2. Good-and-bad data model
3. The Cauchy formulation
4. Many nuisance parameters

Consider data $D = \{y_i\}_{i=1}^{N}$ with specified absolute error σ_0

our model: $y_M(x;\theta) = \theta_0 + \theta_1 x$

$$\Rightarrow y_i \sim N\left(y_M(x_i;\theta), \sigma_0^2\right)$$

The likelihood will be a function of residuals

$$R_i(\sigma,\theta) = \frac{y_i - y_M(x_i;\theta)}{\sigma} \; ; \quad R_i(\sigma_0,\theta) \equiv R_i(\theta)$$

Below, we will also assume uniform priors for θ

$$p(\theta|I) \propto 1$$

E.g the standard log-likelihood will give the log-posterior

$$L = \log[p(\theta|D,I)] = \text{constant} - \frac{1}{2}\sum_{i=1}^{N} R_i^2(\theta)$$

1/ We are skeptic to the error assignment; therefore we treat them as lower bounds using $\sigma \geq \sigma_0$. For this purpose we use a variant of Jeffrey's prior

$$p(\sigma|\sigma_0,I) = \begin{cases} \frac{\sigma_0}{\sigma^2} & \text{for } \sigma \geq \sigma_0 \\ 0 & \text{otherwise} \end{cases}$$

Now, we will have to marginalize over σ. Marginal likelihood for one datum

$$p(D_i|\theta,\sigma_0,I) = \int d\sigma\, p(D_i,\sigma|\theta,\sigma_0,I)$$

$$= \int d\sigma \underbrace{p(D_i|\theta,\sigma,I,\cancel{\sigma_0})}_{\text{assume Gaussian, width } \sigma} p(\sigma|\cancel{\sigma_0},I,\cancel{\theta}) \quad \text{(σ_0 does not depend)}$$

$$= \left\{\sigma = \frac{1}{t}\right\} = \frac{\sigma_0}{\sqrt{2\pi}} \int_0^{1/\sigma_0} t\, e^{-t^2\sigma_0^2 R_i^2(\theta)/2}\, dt$$

$$= \frac{1}{\sqrt{2\pi\sigma_0^2}} \left[\frac{1 - e^{-R_i^2(\theta)/2}}{R_i^2(\theta)}\right]$$

The log-posterior (using all data) becomes

$$L(\theta) = \text{constant} + \sum_{i=1}^{N} \log\left[\frac{1 - e^{-R_i^2(\theta)/2}}{R_i^2(\theta)}\right]$$

2/ The-good-and-bad data model

Let's be less pessimistic and allow two possibilities:

a/ the data and its error are reliable

b/ the data is bad and the error should be larger by a (large) factor γ

$$p(\sigma_i | \sigma_0, \gamma, \beta, I) = \beta \delta(\sigma_i - \gamma\sigma_0) + (1-\beta)\delta(\sigma - \sigma_0)$$

where $0 \leq \beta \leq 1$ and $\gamma \gg 1$

(β and γ are additional nuisance parameters)

This gives the log-posterior

$$L(\theta) = \text{constant} + \sum_{i=1}^{N} \log\left[\frac{\beta}{\gamma} e^{-R_i^2(\theta)/2\gamma^2} + (1-\beta) e^{-R_i^2(\theta)/2}\right]$$

(Note: reduces to the standard least-squares when $\beta \to 0$)

3/ The Cauchy formulation

Assume $\sigma \approx \sigma_0$ but could be either narrower or wider

$$p(\sigma | \sigma_0, I) = \frac{2\sigma_0}{\sqrt{\pi}\,\sigma^2} \exp\left(-\frac{\sigma_0^2}{\sigma^2}\right)$$

Marginalizing σ (using $\sigma = \frac{1}{t}$) gives the Cauchy form likelihood

$$p(D_i | \theta, \sigma_0, I) = \frac{1}{\sigma_0 \pi \sqrt{2}\left[1 + R_i^2(\theta)/2\right]}$$

The log posterior becomes

$$L(\theta) = \text{constant} - \sum_{i=1}^{N} \log\left[1 + \frac{R_i^2(\theta)}{2}\right]$$

4/ Let us take the good-and-bad to its extreme with a ~~likelihood~~ nuisance parameter

~~$p(D_i|$~~ β_i for each data

(let's assume that we fix γ, i.e. $p(\gamma | I) = \delta(\gamma - 50)$ at a large value)

$\Rightarrow$ N extra nuisance parameters