Bayesian Model Comparison/Selection/Preference

QUESTION

More than one models are introduced to explain the data. Which model should be preferred or how should we rank the models given the data?

Occam's razor advises to buy the simplest model among all that adequately explain the data.

Bayesian methods can consistently and quantitatively solve the model selection problem.

Bayesian inference ranks the models and based on the rank can use all models for robust predictions, taking into account the relative preference/ranking of each model.

Bayesian Inference

Level 1: Parameter Estimation Given a Model

Posterior **Likelihood** **Prior**

$$p(\theta \mid D, \mathrm{M}) = \frac{p(D \mid \theta, \mathrm{M})\ \pi(\theta \mid \mathrm{M})}{p(D \mid \mathrm{M})}$$

Evidence

Asymptotic Approximation

Posterior PDF is Gaussian

$$p(\theta \mid D, \mathrm{M}) \approx N(\hat{\theta}, H^{-1}(\hat{\theta}))$$

Centered at the most Probable Model: $\hat{\theta} = \arg \min_{\theta}[L(\theta)]$

with covariance the inverse of the Hessian of: $L(\theta)$

$$L(\theta) = -\ln\ p(\theta \mid D, \mathrm{M})$$

$$= -\ln[\text{Likelihood x Prior}]$$

Bayesian Inference

Level 1: Parameter Estimation Given a Model

Posterior Likelihood Prior

$$p(\theta \mid D, \mathrm{M}) = \frac{p(D \mid \theta, \mathrm{M})\ \pi(\theta \mid \mathrm{M})}{p(D \mid \mathrm{M})}$$

Evidence

Level 2: Model Selection/Preference

$$P(\mathrm{M}_i \mid D) \propto p(D \mid \mathrm{M}_i)\ P(\mathrm{M}_i)$$

Preference Evidence Prior

Evidence of Model

$$p(D \mid \mathrm{M}_i) = \int p(D \mid \theta_i, \mathrm{M}_i)\ \pi(\theta_i \mid \mathrm{M}_i)\ d\theta_i$$

$$= \int \exp[-L(\theta_i)]\ d\theta_i$$

Laplace-type Integral

Laplace Asymptotic Approximation

$$I = \int h(\theta)\ \exp[-\lambda g(\theta)]\, d\theta$$

$$= h(\hat{\theta})\ \exp[-\lambda g(\hat{\theta})] \frac{\left(\sqrt{2\pi}\right)^n}{\sqrt{\lambda^n \det H(\hat{\theta})}} \left[1 + O\left(\frac{1}{\lambda}\right)\right]$$

Evidence of Model

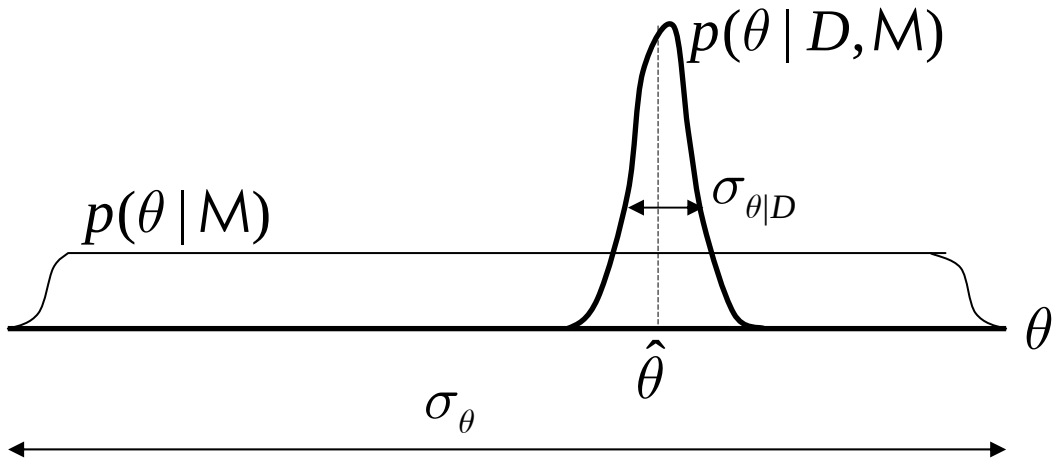
$$p(D \mid \mathrm{M}_i) = \int p(D \mid \theta_i, \mathrm{M}_i)\ \pi(\theta_i \mid \mathrm{M}_i)\ d\theta_i$$

$$= \int \exp[-L(\theta_i)]\, d\theta_i$$

$$p(D \mid \mathrm{M}_i)\ \approx p(D \mid \hat{\theta}, \mathrm{M}_i)]\ \pi_\theta(\hat{\theta} \mid \mathrm{M}_i) \frac{(2\pi)^{n/2}}{\sqrt{\det H(\hat{\theta})}}$$

Bayesian Inference: Single Parameter Case

$$p(\theta \mid D, M_i)] = \frac{1}{\sqrt{2\pi}\sigma_{\theta|D}} \exp\left[-\frac{\left(\theta - \hat{\theta}\right)^2}{2\sigma_{\theta|D}^2} \right]$$



Evidence of Model

$$p(D \mid M) \approx p(D \mid \hat{\theta}, M) \; \pi_\theta(\hat{\theta} \mid M) \; \sigma_{\theta|D}$$

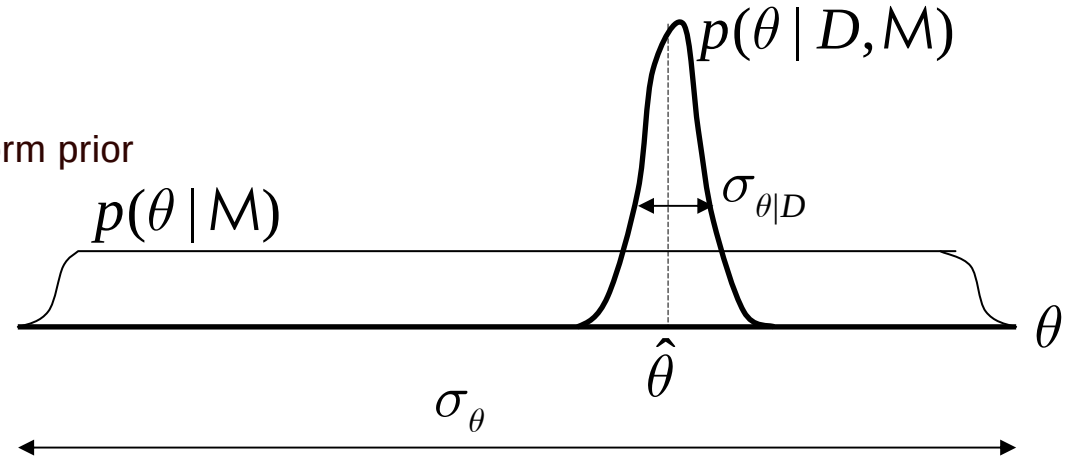
Likelihood at the best fit that the model can achieve

Occam Factor

Penalizing model for having the parameter

Interpretation of Occam Factor: 1-d Case

Consider uniform prior

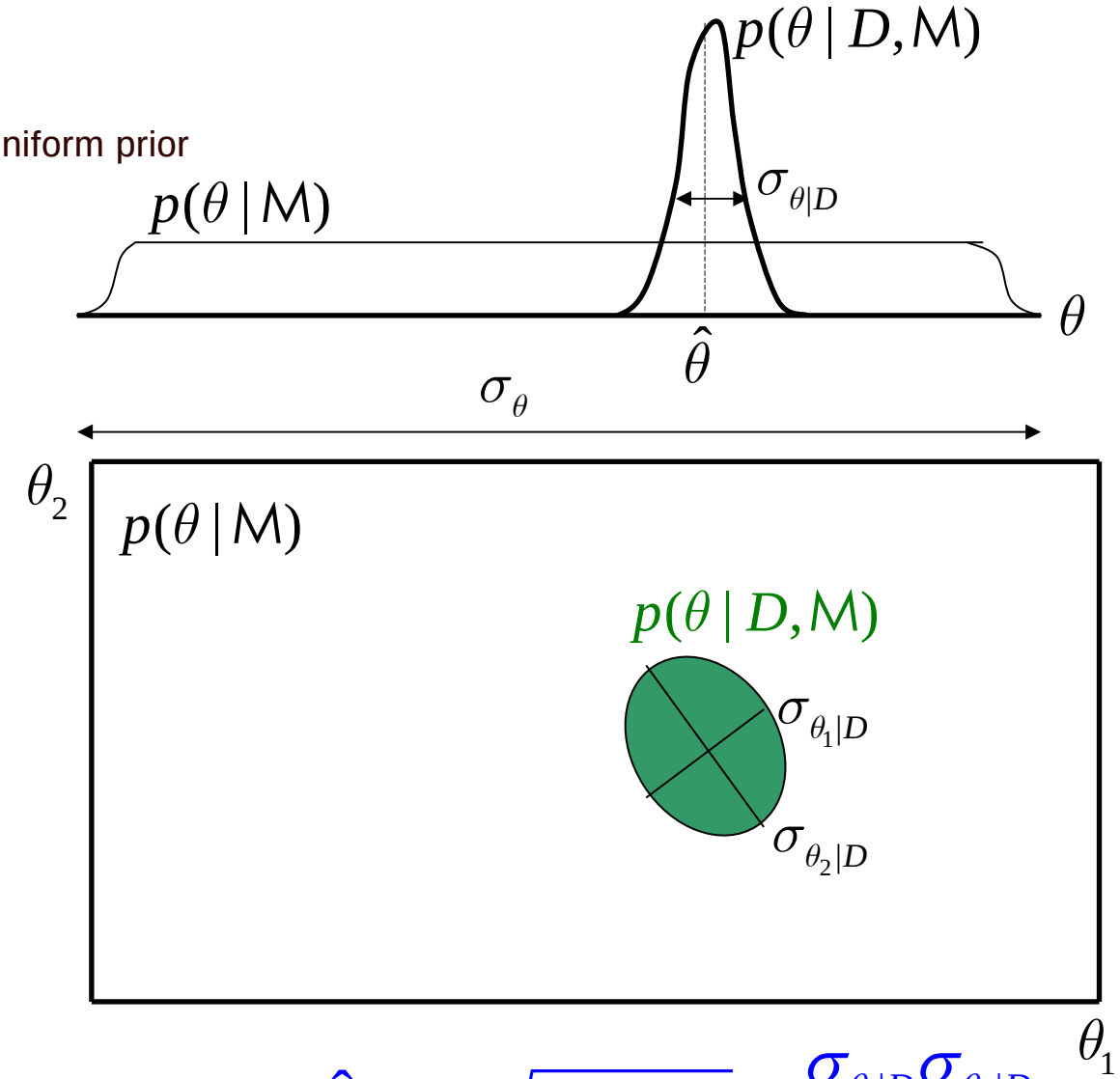


Occam Factor: $p(\hat{\theta} \mid M)\ \sigma_{\theta|D} = \dfrac{\sigma_{\theta|D}}{\sigma_\theta}$

- **Ratio of the posterior accessible volume of the parameter space to the prior accessible volume**
- **Factor by which the model hypothesis space collapses when the data arrives**
- **Measure of theamount of the information we gain about themodel parameters given the data**
- **Represents a penalty against parameterization. Depends on the number of model parameters (complexity of the model) and their assigned prior probability**
- **The greatest evidence between models is determined by a trade-off minimizing the model complexity measure and minimizing the data misfit**

Interpretation of Occam Factor: 2-d case

Consider uniform prior



Occam Factor: $\pi_\theta(\hat{\theta} \mid M)\sqrt{\det \Sigma_{\theta|D}} = \dfrac{\sigma_{\theta_1|D}\sigma_{\theta_2|D}}{\sigma_{\theta_1}\sigma_{\theta_2}}$

Occam Factor for Large Number of Data

Occam Factor: $$\exp[\beta] = \pi_\theta(\hat{\theta}\,|\,\mathrm{M}_i)\frac{(2\pi)^{n/2}}{\sqrt{\det H(\hat{\theta})}}$$

Consider Gaussian prior, then the log of the Occam factor is

Occam Factor: $$\beta = \ln\left[\, p(\hat{\theta}\,|\,\mathrm{M})\frac{(2\pi)^{n/2}}{\sqrt{\det H(\hat{\theta})}}\,\right]$$

$$= -\sum_{i=1}^{n}\ln\frac{\sigma_{\theta_i}}{\sigma_{\theta_i|D}} - \frac{1}{2}\sum_{i=1}^{n}\frac{(\hat{\theta}_i - \theta_{0i})^2}{\sigma_{\theta_i}^2}$$

- **Occam factor is negative**

- **Is expected to decrrease if the number of model parameters is inclreased**

- **Using the fact that the posterior variances is inversely proportional to the number of data one can derive for large number of data**

$$\beta = -\frac{1}{2}n\log N + R$$

Model Evidence

$$\ln p(D \mid \mathcal{M}) \approx \ln p(D \mid \hat{\theta}, \mathcal{M})] + \ln \pi_{\theta}(\hat{\theta} \mid \mathcal{M}) + \frac{1}{2}\left[n \ln(2\pi) - \det H(\hat{\theta}) \right]$$

$$\ln p(D \mid \mathcal{M}) = \ln p(D \mid \hat{\theta}, \mathcal{M}) - \frac{1}{2} n \log N$$

- **First asymptotic approximation requires Hessian evaluation and is more accurate for small number of data**

- **Second asymptotic approximation does not require Hessian evaluation**