

Bayesian UQ in Engineering Science

Bayesian Model Selection

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

$$p(\theta | D, \mathcal{M}) = \frac{\overset{\text{Likelihood}}{p(D | \theta, \mathcal{M})} \overset{\text{Prior}}{\pi(\theta | \mathcal{M})}}{\underset{\text{Evidence}}{p(D | \mathcal{M})}}$$

Asymptotic Approximation

Posterior PDF is Gaussian

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

Centered at the most Probable Model:

$$\hat{\theta} = \arg \min_{\theta} [L(\theta)]$$

with covariance the inverse of the Hessian of:

$$L(\theta)$$

$$L(\theta) = -\ln p(\theta | D, \mathcal{M})$$

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

Bayesian Inference

Level 1: Parameter Estimation Given a Model

$$p(\theta | D, \mathcal{M}) = \frac{\overset{\text{Likelihood}}{p(D | \theta, \mathcal{M})} \overset{\text{Prior}}{\pi(\theta | \mathcal{M})}}{\underset{\text{Evidence}}{p(D | \mathcal{M})}}$$

Level 2: Model Selection/Preference

$$P(M_i | D) \propto \underset{\text{Preference}}{p(M_i)} \underset{\text{Evidence}}{p(D | M_i)} \underset{\text{Prior}}{P(M_i)}$$

Evidence of Model

$$\begin{aligned} p(D | M_i) &= \int p(D | \theta_i, M_i) \pi(\theta_i | M_i) d\theta_i \\ &= \int \exp[-L(\theta_i)] d\theta_i \\ &\text{Laplace-type Integral} \end{aligned}$$

Laplace Asymptotic Approximation

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

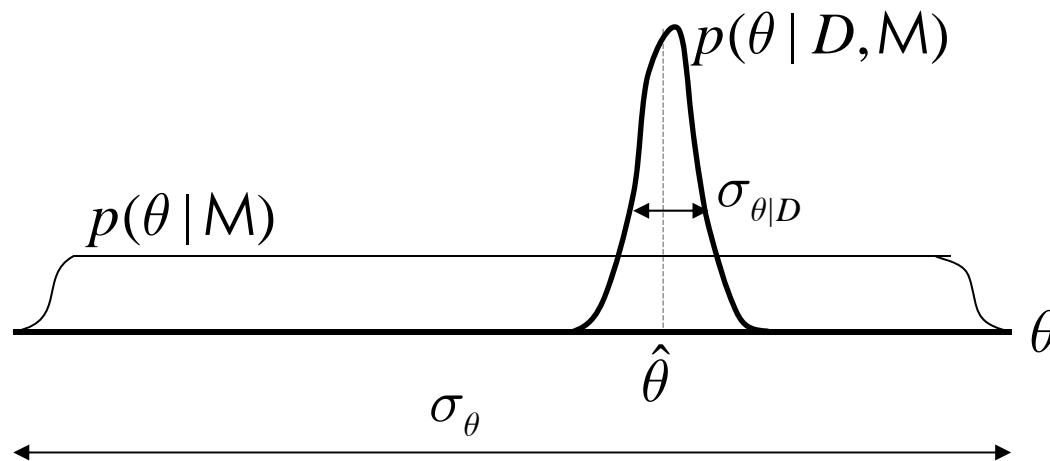
Evidence of Model

$$\begin{aligned} p(D | M_i) &= \int p(D | \theta_i, M_i) \pi(\theta_i | M_i) d\theta_i \\ &= \int \exp[-L(\theta_i)] d\theta_i \end{aligned}$$

$$p(D | M_i) \approx p(D | \hat{\theta}, M_i) \pi_{\theta}(\hat{\theta} | M_i) \frac{(2\pi)^{n/2}}{\sqrt{\det H(\hat{\theta})}}$$

Bayesian Inference: Single Parameter Case

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



Evidence of Model

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

Likelihood at the

best fit that the

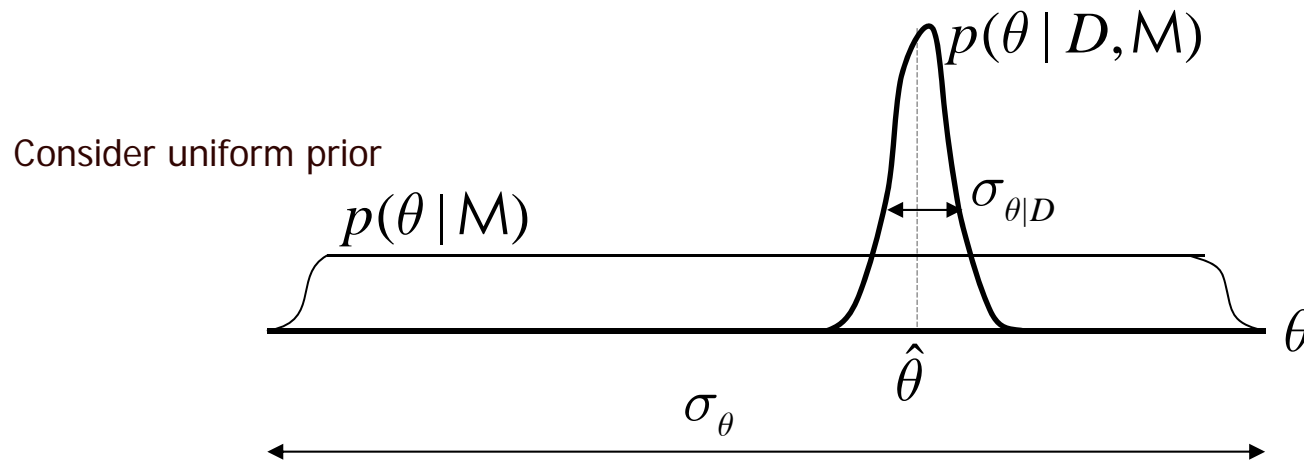
model can achieve

Occam Factor

Penalizing model for having

the parameter

Interpretation of Occam Factor: 1-d Case

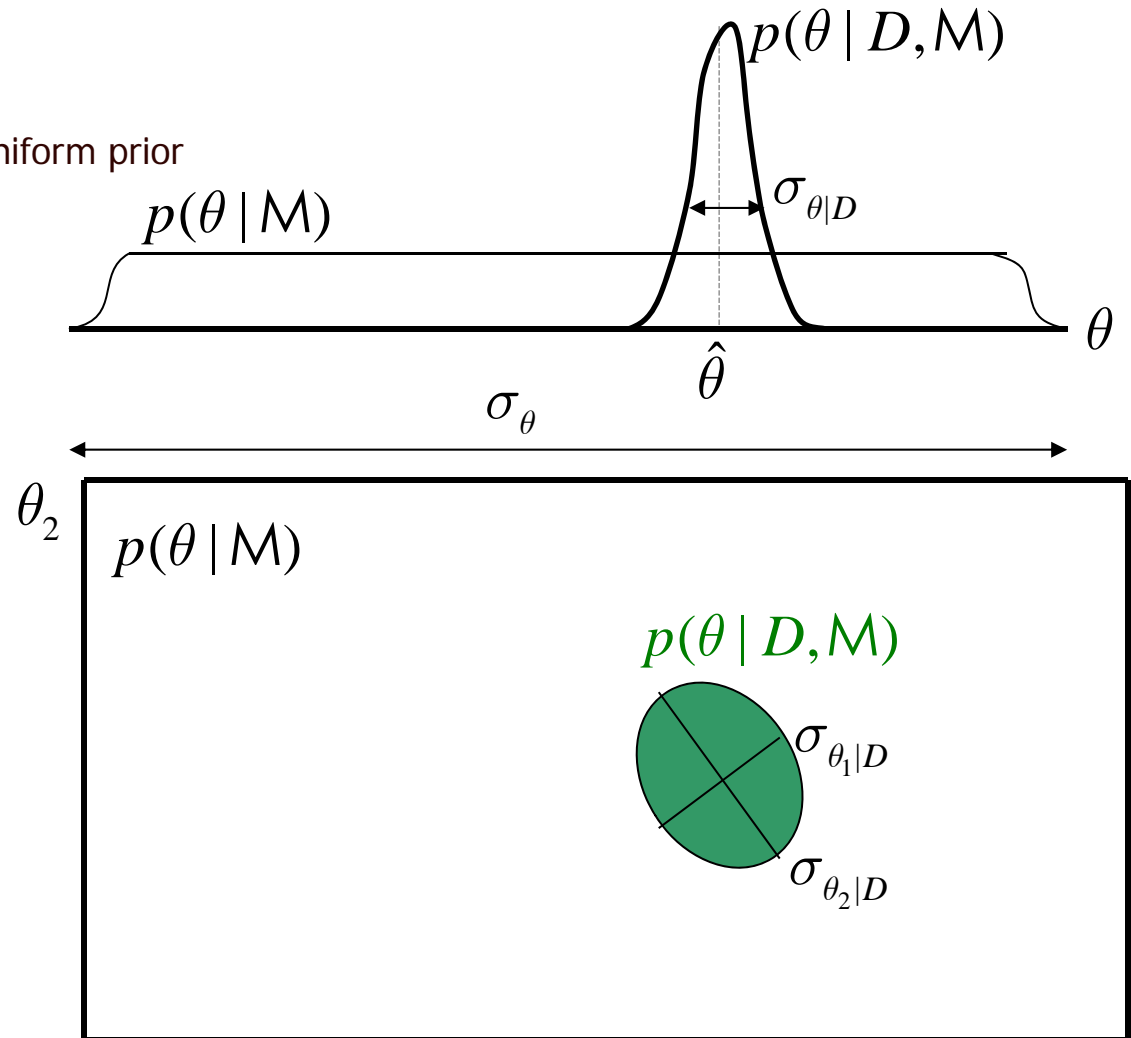


Occam Factor:
$$p(\hat{\theta} | M) \sigma_{\theta|D} = \frac{\sigma_{\theta|D}}{\sigma_{\theta}}$$

- 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 the amount of the information we gain about the model 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: $\pi_\theta(\hat{\theta} | M) \sqrt{\det \Sigma_{\theta|D}} = \frac{\sigma_{\theta_1|D} \sigma_{\theta_2|D}}{\sigma_{\theta_1} \sigma_{\theta_2}}$

Occam Factor for Large Number of Data

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

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

Occam Factor:
$$\beta = \ln \left[p(\hat{\theta} | M) \frac{(2\pi)^{n/2}}{\sqrt{\det H(\hat{\theta})}} \right]$$
$$= -\sum_{i=1}^n \ln \frac{\sigma_{\theta_i}}{\sigma_{\theta_i|D}} - \frac{1}{2} \sum_{i=1}^n \frac{(\hat{\theta}_i - \theta_{0i})^2}{\sigma_{\theta_i}^2}$$

- Occam factor is negative
- Is expected to decrease if the number of model parameters is increased
- Using the fact that the posterior variances is inversely proportional to the number of data one can derive for large number of data

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

Beck, J.L. and Yuen, K.Y. (2004), Model selection using response measurements: Bayesian probabilistic approach, *Journal of Engineering Mechanics*, 130(2), 192-203.

Model Evidence

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

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

- **First asymptotic approximation requires Hessian evaluation and is more accurate for small number of data**
- **Second asymptotic approximation does not require Hessian evaluation**