A Primer on Bayesian Model Selection

Subhayan De

Contents

Chapter 1

Introduction

The Bayesian method in model selection and inference is comparatively a recent field of research. Although Bayesian analysis has been used extensively in research work from Laplace's time, use of the Bayes' theorem in model selection is rather a new domain worth exploring. The application of it can also be far reached.

The first problem in Bayesian model class selection is to identify the parameters of a particular model class. A lot of articles has been dedicated to this purpose [31]. However, a number of articles addressing the second problem of order-ranking the models using Bayes' theorem with identified parameters is much less. At one point in time, in fact, Sir Ronald A. Fisher [1] believed that the model specification is out of the scope of mathematical statistics. The situation got changed when Akaike in 1973 derived his Akakaike's Information criterion (AIC) and with some modification AIC$_c$. Then came the Swartz criterion or Bayesian Information Criterion (BIC). These criteria are based on maximum likelihood theory and information theory. Even in the field of molecular biology software (e.g. Modelgenerator etc.) has been developed based on these criteria. However, the application was limited to mainly in the in the field of ecology and evolution [15]. The application of these criteria to other engineering application is however quite limited. Literature for model selection based on these criteria or information theory is limited to statistics and biology journals [9, 14, 15, 17, 29]. From 2004 Prof. Beck with his few students has been working on Bayesian model class selection [4, 5, 8, 20, 30, 31] with applications to structural engineering problems. A scope of extending the Bayesian model class selection problem is through model averaging. In statistics, Bayesian model averaging concept has been developed in late 90's and it is expanded in the paper 'Bayesian Model Averaging : A Tutorial' by Hoeting et al [13]. This primer discusses the theoretical foundation of Bayesian model selection and the challenges in applying this tool to practical problems.

[1] *Maximum Likelihood Theory* has been developed by Sir R. A. Fisher [1890-1962].

Chapter 2

Bayesian Inference

2.1 Introduction

Bayes' theorem has been developed by the British mathematician Thomas Bayes [1702-1761] in his well known paper 'An essay towards solving a problem in the doctrine of chances' [3]. In this chapter a short introduction of the application of this theorem in model class selection problem is presented.

2.2 Bayes' Theorem

2.2.1 Bayes' Theorem for Discrete Events

Let us assume A and B denote two events. Then using Bayes' theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad \text{if } P(B) > 0 \tag{2.1}$$

If the event A is partitioned into N mutually exclusive events, $A_1, A_2, ..., A_N$

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{N} P(B|A_i)P(A_i)} \qquad \text{for } i = 1, 2, ..., N \tag{2.2}$$

2.2.2 Bayes' Theorem for Discrete Events with Continuous-valued Parameters

If we take the event B as $B \in (b, b + db)$, we can use Equation (2.1) and write

$$\begin{aligned} P[A|B \in (b, b + db)] &= \frac{P[B \in (b, b + db)|A]P(A)}{P[B \in (b, b + db)]} \\ &= \frac{p(b|A)db.P(A)}{p(b)db} \end{aligned} \tag{2.3}$$

$$\therefore \quad P(A|b) = \frac{p(b|A)P(A)}{p(b)} \tag{2.4}$$

2.2.3 Bayes' Theorem for Continuous-valued Parameters

Similarly, as before for continuous valued parameters with $A \in (a, a + da)$ we can write

$$p(a|b) = \frac{p(b|a)p(a)}{p(b)} \tag{2.5}$$

2.3 Bayesian Inference

Bayes' Theorem offers the possibility for inferencing uncertain models/systems from their measurements. Let us assume $M_1, M_2, ..., M_{N_m}$ are the N_m numbers of different model classes for a physical system. Let the uncertain parameter vector for each model class be $\boldsymbol{\theta_j}$, $j = 1, 2, ..., N_m$ where the dimensions of each $\boldsymbol{\theta}_j$ vector are $s_1 \times 1, s_2 \times 1, ..., s_{N_m} \times 1$, respectively. Let D be the measurement data of the system.

2.3.1 Problem-1: Parametric Identification

In this level of inference, a class of mathematical model M_j is given for a particular physical phenomenon or system and we are asked to identify the unknown parameters $\boldsymbol{\theta}_j$.
From the Bayes' Theorem (2.5),

$$\begin{aligned} p(\boldsymbol{\theta}_j|D, M_i) &= \frac{p(D|\boldsymbol{\theta}_j, M_j)p(\boldsymbol{\theta}_j|M_j)}{p(D|M_j)}; \qquad j = 1, 2, ..., N_m \\ &= k_0 p(D|\boldsymbol{\theta}_j, M_j)p(\boldsymbol{\theta}_j|M_j) \end{aligned} \tag{2.6}$$

where $k_0 = \frac{1}{p(D|M_j)}$.
Here, $p(\boldsymbol{\theta}_j|D, M_j)$ is the posterior distribution of the parameters, $p(D|\boldsymbol{\theta}_j, M_j)$ is known as 'likelihood', $p(\boldsymbol{\theta}_j|M_j)$ is known as 'prior' and $p(D|M_j)$ is known as 'evidence'. However, to get the posterior using MCMC (Markov Chain Monte Carlo) method we do not need to evaluate the evidence.

2.3.2 Markov Chain Monte Carlo[22, 27]

Markov chain Monte Carlo (MCMC), is a general computational approach that replaces analytic integration by summation over samples generated from iterative algorithms. Problems that are intractable using analytic approaches often become possible to solve using some form of MCMC, even with high-dimensional problems.

The goal of MCMC is to design a Markov chain such that the stationary distribution of the chain is exactly the distribution that we are interesting in sampling from. This is called the target distribution. There are a number of methods that achieve this goal using relatively simple procedures. I have used Metropolis-Hastings sampling in the current research work.

Metropolis-Hastings Algorithm[21, 27]

Let us assume our goal is to sample from the target density $p(\boldsymbol{\theta})$, then Metropolis sampler creates a Markov chain where $\boldsymbol{\theta}^{(t)}$ represents the state of a Markov chain at iteration t.

Algorithm 1 Metropolis-Hastings Sampling Algorithm

```
Set $t = 1$.
Generate a initial value u, and set $\boldsymbol{\theta}^{(t)} = u$.
while $t \leq T$ do
    $t = t + 1$
    Generate a proposal from $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)})$
    Evaluate the acceptance probability $\alpha = \min\left(1, \frac{p(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(t-1)})} \frac{q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})}\right)$.
    Generate a $u$ from a $Uniform(0,1)$ distribution.
    if $u < \alpha$ then
        accept the proposal and set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$
    else
        set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.
    end if
end while
```

When the proposal distribution $q(\boldsymbol{\theta})$ is symmetric then the above algorithm becomes Metropolis sampler.

2.3.3 Problem-2: Model Class Selection

The probability of a class of models conditioned on the data set D,

$$P(M_j|D) = \frac{p(D|M_j)P(M_j)}{p(D)}, \qquad j = 1, 2, ..., N_m. \tag{2.7}$$

where, the denominator $p(D) = \sum_{j=1}^{N_m} p(D|M_j)P(M_j)$ and $P(M_j)$ is a priori measure of plausibility assigned by user such that $\sum_{j=1}^{N_m} P(M_j) = 1$. However, specification of $P(M_i)$, the prior distribution over competing models is challenging. The algorithm for Bayesian model class selection problems is as follows

Algorithm 2 Bayesian Model Class Selection Algorithm

```
Collect measurement data $D$
for $i = 1, ...., N_m$ do
    Estimate $p(D|M_i)$.
end for
for $i = 1, ...., N_m$ do
    Calculate $P(M_i|D) = \frac{p(D|M_i)P(M_i)}{\sum_{i=1}^{N_m} p(D|M_i)P(M_i)}$.
end for
```

2.4 Occam's Razor

> "It is vain to do with more what can be done with fewer" - William of Occam (or Ockham in english)

It is obvious that a more complicated model will fit the data better than a less complicated one with fewer adjustable parameters. This approach is likely to over-fitted model classes. Over-fitted models perform poorly for future predictions. Therefore, it is necessary to penalize more complicated models.

2.4.1 Occams Razor in Bayesian inference [17]

In Bayesian Model Class Selection, the evidence for each model class provided by data automatically enforces a penalty for complicated models. The evidence

$$p(D|M_i) = \int_{\Theta} p(D|\theta_i, M_i)p(\theta_i|M_i)d\theta$$

For many problems the posterior distribution $p(\theta|D, M_i) \propto p(D|\theta_i, M_i)p(\theta_i|M_i)$ will have a strong peak at the most probable parameters. Then the evidence can be approximated by the height of the peak multiplied by the width $\Delta\theta$.

$$\begin{aligned} p(D|M_i) &\simeq p(D|\theta_{MP}, M_i)p(\theta_{MP}|M_i)\Delta\theta \\ \text{Evidence} &\simeq \text{Best fit likelihood} \times \text{Occam factor} \end{aligned} \tag{2.8}$$

The quantity $\Delta\theta$ is the posterior uncertainty in θ. For simplicity let us assume the prior $p(\theta|M_i)$ is uniform on some large interval $\Delta^0\theta$, representing the range of values of θ that M_i thought possible before the data arrived. Then,

$$p(\theta_{MP}|M_i) = \frac{1}{\Delta^0\theta} \tag{2.9}$$

$$\text{and Occam factor } = \frac{\Delta\theta}{\Delta^0\theta} \tag{2.10}$$

Hence, the occam factor is the ratio of the posterior accessible volume of M_i's parameter space to the prior accessible volume. Typically, a complicated model with many parameters each of which is free to vary over a large range $\Delta^0\theta$ will be penalized by a larger Occam factor than a simpler one. Also, the Occam factor provides a penalty for overfiiting of data by the term in the numerator, namely $\Delta\theta$. [1]

[1] Occam factor in Bayesian inference for several parameters has been thoroughly discussed in MacKay (1992)[17] and in Beck & Yuen (2004)[5]. From logarithm of evidence it can also be shown that Occam's razor principle is present inside the calculation of evidence[20].

2.5 Predictive Analysis using Model Classes (Model Averaging)

2.5.1 Bayesian Model Averaging (BMA)

Let $M = (M_1, ..., M_{N_m})$ be the set of models under consideration. If Δ is the quantity of interest, then the posterior distribution of Δ given data D is[13],

$$p(\Delta|D) = \sum_{i=1}^{N_m} p(\Delta|D, M_i) P(M_i|D) \tag{2.11}$$

Parameter estimates and other quantities of interest can be estimated by the principle described above. For example, the BMA estimate of a common parameter θ is,

$$\hat{\theta}_{BMA} = \sum_{i=1}^{N_m} \hat{\theta}_i P(M_i|D) \tag{2.12}$$

Many system performance measures can be expressed as the expectation of some function $g(X)$ with respect to the posterior pdf.

$$\langle g(X)|D, M_i\rangle = \int g(x) p(x|D, M_i) dx \tag{2.13}$$

Using the BMA concept,

$$\langle g(X|D)\rangle_{BMA} = \sum_{i=1}^{N_m} \langle g(X|D, M_i)\rangle P(M_i|D) \tag{2.14}$$

When $g(X) = 1_F(X)$ which is equal to 1 if $X \in F$ and 0 otherwise, where F is the region in the response space that corresponds to unsatisfactory system performance, then it gives posterior failure probability $P(F|D)$.

2.5.2 Implementing Bayesian Model Averaging

The first question to answer while doing model averaging is which model classes should we consider? Unfortunately there is no certain answer. But a guideline can be set, such as follows [13],

The model classes not belonging to

$$\mathscr{A} = \left\{ M_k : \frac{\max_l \{P(M_l|D)\}}{P(M_k|D)} \leq C \right\} \tag{2.15}$$

should be excluded from Equation (2.11).

Hence, we can replace Equation (2.11) by,

$$p(\Delta|D) = \sum_{M_i \in \mathscr{A}} p(\Delta|D, M_i) P(M_i|D) \tag{2.16}$$

2.5.3 Predictive Performance

The purpose of order ranking models is also to make forcasts. In the examples the data is split into two portions, D^I and D^{II}. D^I is used to get the order of merit of each model, i.e. $P(M_i|D^I)$ and performance is then measured on the second half of the data, D^{II}. The predictive ability is measured by the following formula (Logarithmic scoring rule) [13],

$$\text{performance}_i = \sum_{d \in D^{II}} \log p(d|M_i, D^I) \tag{2.17}$$

and the predictive performance of BMA is measured as,

$$\text{performance} = \sum_{d \in D^{II}} \log \left\{ \sum_{M_i \in \mathscr{A}} p(d|M_i, D^I) P(M|D^I) \right\} \tag{2.18}$$

The smaller the predictive log score for a given model or model average is, the better the predictive performance will be.

2.6 Comparison with Bayes Factor

Bayes factor is also used to compare competing models. Bayes factor is given by,

$$K = \frac{Z_i}{Z_j} = \frac{p(D|M_i)}{p(D|M_j)} = \frac{\int_\Theta p(D|\theta_i, M_i) p(\theta_i|M_i) d\theta}{\int_\Theta p(D|\theta_j, M_j) p(\theta_j|M_j) d\theta} \tag{2.19}$$

If $K > 1$ then the model M_i will be favoured. However, from the above relation it is clear that the using Bayes factor is same as using Bayesian model class selection approach with equal prior. Famous mathematician Sir Harold Jeffreys provided a table to interpret K and compare models [Table 2.1].

Z_j/Z_i	$log_2(Z_j/Z_i)$	$log_e(Z_j/Z_i)$	$log_{10}(Z_j/Z_i)$	Evidence against model M_i
$1 - 3.2$	$0 - 1.7$	$0 - 1.2$	$0 - 0.5$	Weak
$3.2 - 10$	$1.7 - 3.3$	$1.2 - 2.3$	$0.5 - 1$	Substantial
$10 - 100$	$3.3 - 6.6$	$2.3 - 4.6$	$1 - 2$	Strong
> 100	> 6.6	> 4.6	> 2	Decisive
> 1000	> 10	> 7	> 3	Beyond reasonable doubt

Table 2.1: Jeffrey's scale[19].

Chapter 3

Different Methods for Evaluating the ‘Evidence’

3.1 Introduction

The main challenge in Bayesian Model Class Selection problems lies in evaluating the evidence i.e. $p(\mathfrak{D}|\mathcal{M}_k)$. For a particular model class, $\mathcal{M}_k$, the evidence can be written as,

$$\mathcal{Z}^{(k)} = \int_{\Theta} p(\mathfrak{D}|\theta, \mathcal{M}_k)p(\theta|\mathcal{M}_k)d\theta = \int_{\Theta} \mathcal{L}(\theta, \mathcal{M}_k)p(\theta|\mathcal{M}_k)d\theta \tag{3.1}$$

where, $\mathcal{L}(\theta, \mathcal{M}_k)$ is the likelihood function and $p(\theta|\mathcal{M}_k)$ is the prior distribution.

For the past few decades many authors have proposed different methods. Among them few methods applicable to structural problems are discussed here.

3.2 Laplace’s Method

If $p(\theta|\mathfrak{D}, M_k)$ can be approximated closely by a Gaussian then we can use ‘Laplace’s method’ [**?**].

$$\mathcal{Z}^{(k)} \approx (2\pi)^{r/2}|\Sigma|^{-1/2}p(\mathfrak{D}|\tilde{\theta}, M_k)p(\tilde{\theta}|\mathcal{M}_k) \tag{3.2}$$

where $\Sigma = -(\nabla\nabla \log p(\mathfrak{D}|\hat{\theta}, \mathcal{M}_k))$, r is the dimension of θ vector for M_k model and $\hat{\theta}$ is the most probable value of θ i.e. which maximizes $p(\mathfrak{D}|\theta, \mathcal{M}_k)$.

3.3 Arithmetic Mean Estimator

We can sample θ_i from pdf $p(\theta)$ where, $i = 1$ to N and N can be taken suitably large. Then we can approximate the integral by Monte-Carlo approximation,

$$\mathcal{Z}^{(k)} = \int_{\Theta} \mathcal{L}(\theta, \mathcal{M}_k)p(\theta|\mathcal{M}_k)d\theta \approx \frac{1}{N_s}\sum_{i=1}^{N_s} \mathcal{L}(\theta_i, \mathcal{M}_k) \tag{3.3}$$

However, in these method if N is not large enough and likelihood is highly peaked then most of the samples we take are from low likelihood region and give a bad estimate.

3.4 Harmonic Mean Estimator [?]

Newton and Raftery in their 1994 paper proposed one method to calculate ‘Evidence’. In this method,

$$\mathcal{Z}^{(k)} = \int_{\Theta} \mathcal{L}(\theta, \mathcal{M}_k) p(\theta|\mathcal{M}_k) d\theta \approx \frac{1}{\left(\frac{1}{N_s}\sum_{i=1}^{N_s} \mathcal{L}(\theta_i, \mathcal{M}_k)\right)} \tag{3.4}$$

where, $\theta_i \sim p(\theta|\mathfrak{D}, \mathcal{M}_k)$. When N_s is large enough then from law of large numbers the above result can be shown.

However, as we take θ_i from posterior distribution in this method the dependence of $\mathcal{Z}_{(k)}$ on prior distribution is not present in here. For two totally different priors which encompass the posterior region and informative data we will get same result which will not show ‘Occam’s razor’ effect. Another problem with this method is that we have to take a very large number of θ_i to get the correct estimate of ‘evidence’.

3.5 Annealed Sampling Method [?]

R. M. Neal in 2001 proposed Annealed Sampling method which can also be used to calculate normalizing constants. The calculation of evidence by this method is nothing but an implementation of Thermodynamic Integration[?]. In this method we have to sample $\{\theta\}_{i=1}^{N_s}$ for model M_k from the following density,

$$p(\theta|\mathfrak{D}, \mathcal{M}_k, \gamma) = \hat{p}(\theta|\mathfrak{D}, \mathcal{M}_k)^{\gamma} p(\theta|\mathcal{M}_k)^{(1-\gamma)} \tag{3.5}$$

where, $\gamma \in [0, 1]$.

To evaluate evidence let us denote $p(\theta|\mathfrak{D}, \mathcal{M}_k) = \hat{p}(\theta|\mathfrak{D}, \mathcal{M}_k)/\mathcal{Z}^{(k)}$ i.e. $\hat{p}(\theta|\mathfrak{D}, \mathcal{M}_k) = p(\mathfrak{D}|\theta, \mathcal{M}_k)p(\theta|\mathcal{M}_k) = \mathcal{L}(\theta, \mathcal{M}_k)p(\theta|\mathcal{M}_k)$ (i.e. $\hat{p}$ is not normalized). Now let us vary γ from 0 to 1 e.g. $0 = \gamma_0 < \gamma_1 < .... < \gamma_{N_t-1} < \gamma_{N_t} = 1$. When $j = 0$, p_j becomes prior distribution and when $j = N_t$, p_j becomes posterior distribution. At the first step we generate θ_0 from prior distribution p_0. Then using the idea of importance sampling we can generate sample from p_{j+1} by taking θ_j generated from p_j at the previous step via Metropolis-Hastings or Gibbs sampling algorithm. At the end of N_tth step we get posterior samples of θ (as θ_i, $i = 1$ to N_s) from $p(\theta|\mathfrak{D}, \mathcal{M}_k)$. From this calculation as a byproduct we can get the importance weights($w_i^{(j)}$) at each step and using those weights,

$$W_i = \prod_{j=1}^{N_t} w_i^{(j)} = \prod_{j=1}^{N_t} \frac{p(\theta_j|\mathfrak{D}, \mathcal{M}_k, \gamma_j)}{p(\theta_{j-1}|\mathfrak{D}, \mathcal{M}_k, \gamma_{j-1})} \tag{3.6}$$

Clearly,

$$\begin{aligned}\frac{\sum_{i=1}^{N_s} W_i}{N_s} &\approx \frac{\mathcal{Z}_{N_s}^{(k)}}{\mathcal{Z}_0^{(k)}} \\ &= \frac{\int_\Theta p_{N_s}(\theta|\mathfrak{D}, \mathcal{M}_k)d\theta}{\int_\Theta p_0(\theta|\mathfrak{D}, \mathcal{M}_k)d\theta} \\ &= \int_\Theta \hat{p}(\theta|\mathfrak{D})d\theta \\ &= \int_\Theta \mathcal{L}(\theta, \mathcal{M}_k)p(\theta|\mathcal{M}_k)d\theta\end{aligned} \tag{3.7}$$

[$\because p_{N_s}(\theta|\mathfrak{D}, \mathcal{M}_k) = \hat{p}(\theta|\mathfrak{D}, \mathcal{M}_k)$ and $\int_\Theta p_0(\theta|\mathcal{M}_k)d\theta = 1$].
Therefore, in this method evidence is calculated as,

$$\mathcal{Z}^{(k)} = \frac{\sum_{i=1}^{N_s} W_i}{N_s} \tag{3.8}$$

3.6 Power Posterior Method [?]

Nial Friel and Pettitt defined 'power posterior' as,

$$p(\theta|\mathfrak{D}, \mathcal{M}_k, \Gamma(\gamma)) \propto p(\mathfrak{D}|\theta, \mathcal{M}_k)^{\Gamma(\gamma)} p(\theta|\mathcal{M}_k) \tag{3.9}$$

where, $\Gamma(\gamma) : [0,1] \to [0,1]$. For simplicity we can take simply $\Gamma(\gamma) = \gamma$.
This method is also related to Thermodynamic Integration principles or path sampling. In this method we take samples from the power posterior for different values of γ (e.g. $0 = \gamma_0 < \gamma_1 < .... < \gamma_{N_t - 1} < \gamma_{N_t} = 1$).
The normalizing constants for power posterior will be,

$$p(\mathfrak{D}|\mathcal{M}_k, \gamma) = \int_\Theta p(\mathfrak{D}|\theta, \mathcal{M}_k)^{\Gamma(\gamma)} p(\theta|\mathcal{M}_k)d\theta \tag{3.10}$$

Following the identity proved in [**?**, **?**] we can write,

$$\begin{aligned}\log(\mathcal{Z}^{(k)}) &= \log\{p(\mathfrak{D}|\mathcal{M}_k)\} \\ &= \log\left\{\frac{p(\mathfrak{D}|\mathcal{M}_k, \gamma = 1)}{p(\mathfrak{D}|\mathcal{M}_k, \gamma = 0)}\right\} \\ &= \int_0^1 E_{\theta|\gamma}[\log\{p(\mathfrak{D}|\theta, \mathcal{M}_k)\}]d\gamma \\ &= \int_0^1 z(\gamma)d\gamma\end{aligned} \tag{3.11}$$

where, $z(\gamma) = E_{\theta|\gamma}[\log\{p(\mathfrak{D}|\theta, \mathcal{M}_k)\}]$.
The expectation is taken with respect to $p(\theta|\mathfrak{D}, \mathcal{M}_k, \Gamma(\gamma))$. To evaluate this inte-

gration we can use standard quadrature rules e.g. using 'Trapezoidal rule' we get,

$$\begin{aligned}\log(\mathcal{Z}^{(k)}) &= \int_0^1 E_{\theta|\gamma}[\log\{p(\mathfrak{D}|\theta, \mathcal{M}_k)\}]d\gamma \\ &\approx \frac{1}{2}\sum_{j=1}^{N_t-1}(\gamma_{j+1}-\gamma_j)[z(\gamma_{j+1})+z(\gamma_j)]\end{aligned} \tag{3.12}$$

3.7 Nested Sampling Method

John Skilling in 2004-2006 proposed a means of estimating the integral that tries to sample from high likelihood region. This technique is "nested sampling". In nested sampling method we use the following result,

Problem 1. *Let X be a non-negative random variable, that is, $P[X < 0] = 0$. Then the shaded area in the figure is equal to the expected value of X .*

Proof. [18]

$$\begin{aligned}E[X] &= \int_0^\infty x p_X(x)dx \\ &= \int_0^\infty x P'{}_X(x)dx \\ &= -\int_0^\infty x[1-P_X(x)]'dx \\ &= -\{x[1-P_X(x)]\}_0^\infty + \int_0^\infty [1-P_X(x)]dx \\ &\text{Now,} \quad \lim_{x\to 0} x[1-P_X(x)] \to 0 \qquad [\because P_X(0)=0] \\ &\int_0^\infty x p_X(x)dx < \infty \Rightarrow \lim_{k\to\infty}\int_k^\infty x p_X(x)dx \to 0, \\ &\text{and, since} \quad k\int_k^\infty p_X(x)dx \le \int_k^\infty x p_X(x)dx, \\ &\lim_{k\to\infty} k[1-P_X(k)] \to 0 \\ &\Rightarrow E[X] = \int_0^\infty [1-P_X(x)]dx\end{aligned} \tag{3.13}$$

□

3.7.1 Distribution of Likelihood Function

We can define the distribution of likelihood function as,

$$P_{\mathcal{L}}(\lambda) = P[\mathcal{L}(\theta) < \lambda] = \int_{\mathcal{L}(\theta)<\lambda} p(\theta)d\theta \tag{3.14}$$

$$\begin{aligned}\int_{\Theta} \mathcal{L}(\theta)p(\theta)d\theta = E_{\theta}[\mathcal{L}(\theta)] &= \int_0^{\infty} [1 - P_{\mathcal{L}}(\lambda)]d\lambda \\ &= \int_0^{\infty} X(\lambda)d\lambda\end{aligned} \tag{3.15}$$

Now, taking

$$X(\lambda) = s; \lambda = X^{-1}(s) \tag{3.16}$$

.
Therefore,

$$\int_{\Theta} \mathcal{L}(\theta)p(\theta)d\theta = \int_0^{\infty} X(\lambda)d\lambda = \int_0^1 X^{-1}(s)ds \tag{3.17}$$

3.7.2 Calculation of Evidence

The integral we need to evaluate can be approximated as a weighted sum using some quadrature rules.

$$\mathcal{Z} = \int_0^1 X^{-1}(s)ds \cong \sum_i \mathcal{L}_i \omega_i \tag{3.18}$$

To start, i is set as zero and N number of samples are taken from prior (i.e. $s_0 = 1$) and $\mathcal{L}$ are evaluated for all N samples. The samples are sorted according to their values of likelihood$\mathcal{L}$. Then the sample with lowest likelihood $\mathcal{L}_0$ is replaced with a new sample which gives $\mathcal{L} > \mathcal{L}_0$. The corresponding prior-volume subjected to the constrained $\mathcal{L} > \mathcal{L}_0$. This is equivalent to taking sample from prior volume which can be given by the random variable, $s_1 = \tau_1 s_0$, where τ_1 follows the distribution, $P(\tau) = N\tau^{(N-1)}$ [1]. For the next iterations the same procedure is repeated i.e. replacing by drawing from prior with $\mathcal{L} > \mathcal{L}_i$ ($\mathcal{L}_i$ is the lowest likelihood at i^{th} iteration). Corresponding reduced prior volume will be $s_i = \tau_i s_{i-1}$. By this procedure the algorithm moves to go to high likelihood region with decreasing prior volumes.

The mean and standard deviation of $\ln t$ are, respectively[10],

$$\langle \ln t \rangle = -\frac{1}{N}; \quad \sigma[\ln t] = \frac{1}{N}. \tag{3.19}$$

after i iterations the prior volume will be approximately $\ln s_i \approx -(i \pm \sqrt{i})/N$. Thus, we can take $s_i = \exp(-i/N)$. Alternatively, the mean and standard deviation of t are[10],

$$\langle t \rangle = \frac{N}{N+1}; \quad \sigma[t] = \frac{N}{N+2} - \left(\frac{N}{N+1}\right)^2. \tag{3.20}$$

[1] This is the distribution for largest of N samples drawn from $\mathcal{U}(0,1)$

Algorithm 3 Nested Sampling Algorithm [6]

```
Sample $N$ points $\theta_1, ., \theta_N$ from prior.
Initialize $\mathcal{Z} = 0, X_0 = 1$.
for $i = 1, 2, ., j$; do
    Store the lowest of the current likelihood values as $\mathcal{L}_i$,
    Set $X_i = \exp(-i/N)$ or $[N/(N+1)]^i$,
    Set $\omega_i = X_{i-1} - X_i$,
    Increment $\mathcal{Z}$ by $\mathcal{L}_i \omega_i$,
    Replace the point corresponding to lowest likelihood value $\mathcal{L}_i$ by a new sample drawn from the prior with $\mathcal{L}(\theta) > \mathcal{L}_i$.
end for
```

3.7.3 Discussion on Nested Sampling

Few of the importance features of nested sampling is discussed below:

Posterior samples $\{\theta_i\}_{i=1}^{N_s}$ can also be calculated as a byproduct of the ‘Nested Sampling’ algorithm. According to Skilling [24], we can get samples from posterior by sampling from the area we already calculated while evaluating evidence or we can take the $\{\theta\}_{i=1}^{N_s}$ values multiplied by the importance weight $\mathcal{L}_i \omega_i$.

Therefore, if we use this method model class selection problem and system identification problem for each model can be done without any extra cost.

Chopin [**?**] proposed a ‘nested importance sampling’ technique where sampling from prior distribution is difficult. So instead we can sample from $\hat{p}(\theta|\mathcal{M}_k)$, calculate $\hat{\mathcal{L}}(\theta_i|\mathcal{M}_k)$ (or $\hat{\mathcal{L}}_i$)and finally multiply $\hat{\mathcal{L}}_i \omega_i$ by importance weight, $W_i = p(\theta_i|\mathcal{M}_k)/\hat{p}(\theta_i|\mathcal{M}_k)$.

Nested sampling technique samples more from the prior in regions where likelihood is high and less from low-likelihood region leading to higher efficiency. Also, it reduces multidimensional integral to one-dimensional integration. On the other hand this technique is not useful for likelihood with multiple peaks. In that case some modifications of this algorithn has been proposed.

3.8 Closure

In the last two chapters the steps of Bayesian model class selection has been discussed. In the next few chapters a formulation to incorporate results from experimental validation in Bayesian model class selection will be presented.

References

[1] Adhikari, S. 2003, *Damping Models for Structural Vibration*, PhD Thesis, University of Cambridge.

[2] Barney, B. 2012,*Introduction to Parallel Computing*, Lawrence Livermore National Laboratory.

[3] Bayes, T. 1763, *An Essay towards solving a Problem in the Doctrine of Chances*, Philosophical Transactions of the Royal Society of London 53 , 370-418.

[4] Beck, J. L. and Cheung, S. H. 2009, *Probability Logic, Model Uncertainty and Robust Predictive System Analysis*, Proceedings of the 10th International Conference on Structural Safety and Reliability.

[5] Beck, J. L. and Yuen, K. 2004, *Model Selection Using Response Measurements: Bayesian Probabilistic Approach*, J. Engg. Mech.,130:192-203.

[6] Bullard, F. 2006, *On Nested Sampling*, ISDS, Duke University.

[7] Burnham, K. P. and Anderson, D. R. 2002, *Model Selection and Multimodel Inference*, 2nd Ed., Springer.

[8] Cheung, S. H. and Beck, J. L. 2009, *Comparison of Different Model Classes for Bayesian Updating and Robust Predictions using Stochastic State-Space System Models*, Proceedings of the 10th International Conference on Structural Safety and Reliability.

[9] Eklund, M., Spjuth, O. and Wikberg, J. ES 2008, *The C^1C^2: A framework for simultaneous model selection and assessment*, BMC Biometric 9:360.

[10] Feroz, F. and Hobson, M. P. 2008, *Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses*, Mon. Not. R. Astron. Soc. 384, 449-463.

[11] Goulet, J. A., Kripakaran, P. and Smith, I. 2010, *Multimodel Structural Performance Monitoring*, J. Struct. Engg., Oct. 1309-1318.

[12] Grama, A., Gupta, A., Karypis, G. and Kumar, V. 2003, *Introduction to Parallel Computing*, 2nd Ed., Addison Wesley.

[13] Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. 1999, *Bayesian Model Averaging: A Tutorial*, J. Statistical Science, Vol. 14, No. 4, 382-401.

[14] Horra, J. de la, Rodriguez-Bernal, M. T. 2005,*Bayesian model selection: a predictive approach with losses based on distances L^1 and L^2*, Statistics & Probability Letters 71 257-265.

[15] Johnson, J. B. and Omland,K. S. 2004, *Model selection in ecology and evolution*, TRENDS in Ecology and Evolution Vol.19 No.2, 101-108.

[16] Kass, R. E. 1992, *Bayes Factors in Practice*, The Statistician, 42, No. 5, Special Issue, 551-560.

[17] MacKay, D. J. C. 1992, *Bayesian Interpolation*, Neural Computation 4, 415-447.

[18] Manohar, C. S. 2011. *Stochastic Structural Dynamics Lecture Notes*, NPTEL.

[19] Mthembu, L., Marwala, T., Friswell, M. I., Adhikari, S. 2011, *Model Selection in finite element model updating using the Bayesian evidence statistic*, Mechanical Systems and Signal Processing 25, 2399-2412.

[20] Muto, M. and Beck, J. L. 2008, *Bayesian Updating of Hysteretic Structural Models Using Stochastic Simulation*, J. Vibration & Control 14: 7-34.

[21] Navarro, D. and Perfors, A. ,*The Metropolis-Hastings Algorithm*, Course Material for COMPSCI 3016: Computational Cognitive Science, University of Adelaide.

[22] Rubinstein, R. Y. and Kroese, D. P 2008, *Simulation and the Monte Carlo Method*, Wiley-Interscience.

[23] Skiling J., 2004. *Nested Sampling*, American Institute of Physics Conf. Proc. 735, 395-405.

[24] Skiling J., 2006. *Nested Sampling for General Bayesian Computation*, International Society for Bayesian Analysis, 1, Number 4, pp. 833-860.

[25] Skiling J., 2005, *Nested Sampling for General Bayesian Computation*, Maximum Entropy Data Consultants Ltd.

[26] Smith, I. and Saitta, S. 2008, *Improving Knowledge of Structural System Behavior through Multiple Models*, J. Struct. Engg., Apr. 553-561.

[27] Steyvers, M., 2011, *Computational Statistics with Matlab*, UC Irvine.

[28] Stoica, P., Selen, Y. and Li, J. 2004, *Multi-model approach to model selection*, J. Digital Signal Processing 14 399-412.

[29] Ward, E. J. 2008, *A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools*, J. Ecological Modeling 211 1-10.

[30] Yuen, K. 2010, *Recent Developments of Bayesian model class selection and applications in civil engineering*, J. Structural Safety 32, 338-346.

[31] Yuen, K. 2010, *Bayesian Methods for Structural Dynamics and Civil Engineering*, John Wiley & Sons Pte Ltd.