MA676: Bayesian Methods

Sujit K. Sahu
Faculty of Mathematical Studies,
University of Southampton,
Highfield, Southampton, UK.

All these pages are available from the web.

http://www.maths.soton.ac.uk/staff/Sahu/teach/ma676/

May 9, 2000

MA676 – Bayesian Methods

Lecturer:	Dr S. K. Sahu	Email:	sks@maths.soton.ac.uk
Room:	9001	Telephone:	595123

This course consists of 12 lectures, including three example sheets.

Assessment is by closed book examination, although a formula sheet will be allowed.

Rough syllabus

1. Bayesian parametric statistical inference
   (a) Prior and Posterior distributions
   (b) Bayesian credible interval
   (c) Bayesian point estimator
   (d) Bayesian prediction
2. Model choice and hypotheses testing
3. Multiparameter problems and Markov Chain Monte Carlo methods

Books

G E P Box and G C Tiao – *Bayesian Inference in Statistical Analysis*

J M Bernardo and A F M Smith – *Bayesian Theory*

M H Degroot – *Probability and Statistics*

A Gelman, J B Carlin, H S Stern and D B Rubin – *Bayesian Data Analysis*

Contents

Chapter 1

Bayesian Statistics

1.1 Introduction

Bayesian theory (named after the Rev. Thomas Bayes, an amateur 18th century English mathematician), provides an approach to statistical inference which is different in spirit from the familiar classical approach. We do not think of Bayesian statistics as a separate area within Statistics. Any statistical problem (Survival analysis; Multivariate analysis; Generalised linear models *etc.*) can be approached in a Bayesian way.

The basic philosophy underlying Bayesian inference is that **the only sensible measure of uncertainty is probability**. Data are still assumed to come from one of a parameterized family of distributions. However, whereas classical statistics considers the parameters to be *fixed but unknown*, the Bayesian approach treats them as random variables in their own right. Prior beliefs about θ are represented by the **prior**, $\pi(\theta)$, a probability density (or mass) function. The **posterior** density (mass function), $\pi(\theta|x_1,\dots,x_n)$ represents our *modified* belief about θ in the light of the observed data. We will do this in quite detail. Let us start from the basics.

1.2 Prior and Posterior Distributions

Theorem 1 (Bayes Theorem)

Let $B_1, B_2, \dots, B_k$ be a set of mutually exclusive and exhaustive events. For any new event A,

$$P(B_i|A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}. \tag{1.1}$$

♡ **Example** 1.1. Suppose there are three production machines, I, II and III. Respectively 2%, 3% and 5% of the items produced by these machines are defective. Of total production machine I yields 30%, II yields 40% and III yields 30%. Suppose an item drawn randomly from the total production turns out to be defective. What is the chance that it was manufactured by machine I?

Notation We are used to the notation that X is the random variable and x is its value. Now we will relax that little bit for the random variable θ only. We will use θ to denote the random variable and Θ to denote the parameter space.

Theorem 2 (Bayes Theorem for Random variables)
Suppose that two random variables X and θ are given with pdf's $f(x|\theta)$ and $\pi(\theta)$.

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)d\theta}. \tag{1.2}$$

Bayesian Inference Framework:

- X = Data or the notation $x_1, x_2, \dots, x_n$.
- θ = Unknown parameters.
- $f(x_1, \dots, x_n|\theta)$ = Likelihood of data given unknown parameters θ.
- $\pi(\theta)$ = Prior distribution or prior belief. A priori what you know for the unknown parameters.

By the above Bayes Theorem,

$$\pi(\theta|x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n|\theta)\pi(\theta)}{\int_\Theta f(x_1, \dots, x_n \mid \theta)\pi(\theta)d\theta}.$$

This distribution is called the **posterior distribution**. Bayesian inference proceeds from this distribution. In practice, the denominator of the above equation needn't usually be calculated, and Bayes' rule is often just written,

$$\pi(\theta|x_1, \dots, x_n) \propto f(x_1, \dots, x_n \mid \theta)\pi(\theta).$$

Hence we always know the posterior distribution up-to a normalizing constant. Often we will be able to identify the posterior distribution of θ just by looking at the numerator. By Bayes Theorem we "update" $\pi(\theta)$ to $\pi(\theta|\mathbf{x})$.

Remark: Bayesian Learning:

$$\begin{aligned}\pi(\theta|x_1) &\propto f(x_1|\theta)\pi(\theta)\\ \pi(\theta|x_1,x_2) &\propto f(x_2|\theta)f(x_1|\theta)\pi(\theta)\\ &\propto f(x_2|\theta)\pi(\theta|x_1)\end{aligned}$$

Thus the Bayes theorem shows how the knowledge about the state of nature represented by θ is continually modified as new data becomes available.

♡ **Example** 1.2. Suppose $X \sim$ binomial(n,θ) where n is known and we assume Beta(α,β) prior for θ. Here the likelihood is

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

Prior is

$$\pi(\theta) = \frac{1}{Beta(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

Hence

$$\pi(\theta|x) \propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}.$$

Note that we have only written down the terms involving θ from the likelihood × the prior. We do not care care about the other terms which do not involve θ, like the $\binom{n}{x}$ or the constant $\frac{1}{Beta(\alpha,\beta)}$. Now the posterior is recognised to be a Beta distribution with parameters $x+\alpha$ and $n-x+\beta$. Hence:

$$\pi(\theta|x) = \frac{1}{Beta(x+\alpha,n-x+\beta)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}.$$

♡ **Example** 1.3. Suppose $X_1,\ldots,X_n$ is a random sample from the distribution with pdf $f(x|\theta) = \theta e^{-\theta x}$. Suppose the prior for θ is given by $\pi(\theta)$ and $\pi(\theta) = \mu e^{-\mu\theta}$ for some known $\mu > 0$.

Then the likelihood is:

$$f(x_1,x_2,\ldots,x_n|\theta) = \prod \theta e^{-\theta x_i} = \theta^n e^{-\theta\sum_{i=1}^n x_i}.$$

Hence the posterior $\propto$ Likelihood × Prior is:

$$\pi(\theta|x_1,\ldots,x_n) \propto \theta^n e^{-\theta\sum_{i=1}^n x_i}\ \mu e^{-\mu\theta}.$$

Now collecting the terms involving θ only we see that:

$$\pi(\theta|x_1,\dots,x_n) \propto \theta^n e^{-\theta(\mu+\sum_{i=1}^n x_i)}.$$

The above is the pdf of a Gamma random variable.

♡ **Example** 1.4. Suppose $X_1,\dots,X_n \sim N(\theta,\sigma^2)$, $\pi(\theta) \sim N(\mu,\tau^2)$ for known μ and τ^2. The likelihood is:

$$f(x_1,x_2,\dots,x_n|\theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x_i-\theta)^2}{\sigma^2}} = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2}\sum_{i=1}^n \frac{(x_i-\theta)^2}{\sigma^2}}.$$

The prior is:

$$\pi(\theta) = \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{1}{2}\frac{(\theta-\mu)^2}{\tau^2}}.$$

The posterior is proportional to the Likelihood × Prior. Hence we keep the terms involving θ only.

$$\pi(\theta|x_1,\dots,x_n) \propto e^{-\frac{1}{2}\left[\sum_{i=1}^n \frac{(x_i-\theta)^2}{\sigma^2}+\frac{(\theta-\mu)^2}{\tau^2}\right]} = e^{-\frac{1}{2}M}.\ \ M \text{ is what's inside [and].}$$

Now we look at the exponent carefully. Notice that it is a quadratic in θ. OK. We should try to complete the square in θ and ultimately we may have that the posterior distribution of θ is a normal distribution. Now:

$$\begin{aligned} M &= \textstyle\sum_{i=1}^n \frac{(x_i-\theta)^2}{\sigma^2} + \frac{(\theta-\mu)^2}{\tau^2} \\ &= \frac{\sum x_i^2 - 2\theta\sum x_i + n\theta^2}{\sigma^2} + \frac{\theta^2-2\theta\mu+\mu^2}{\tau^2} \\ &= \theta^2(n/\sigma^2+1/\tau^2) - 2\theta(\textstyle\sum x_i/\sigma^2+\mu/\tau^2) + \textstyle\sum x_i^2/\sigma^2 + \mu^2/\tau^2 \\ &= \theta^2\, a - 2\theta\, b + c \end{aligned}$$

where

$$a = n/\sigma^2 + 1/\tau^2, b = \sum x_i/\sigma^2 + \mu/\tau^2, c = \sum x_i^2/\sigma^2 + \mu^2/\tau^2.$$

Note that none of a, b and c involves θ. These are defined just for writing convenience. Now

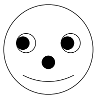
$$\begin{aligned} M &= a(\theta^2 - 2\theta b/a) + c \\ &= a(\theta^2 - 2\theta b/a + b^2/a^2 - b^2/a^2) + c \\ &= a(\theta - b/a)^2 + b^2/a + c \end{aligned}$$

Note again that none of a, b and c involves θ, hence the first term only involves θ and the last two are rubbish!

$$\pi(\theta|x_1,\dots,x_n) \propto e^{-\frac{1}{2}a(\theta-b/a)^2}$$

which is easily recognised to be the pdf of a normal distribution with mean b/a and variance $1/a$. More explicitly

$$\pi(\theta|\mathbf{x}) = N\left(b/a = \frac{\sum x_i/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2}, \frac{1}{n/\sigma^2 + 1/\tau^2}\right) = N\left(\frac{n\bar{x}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2}, \frac{1}{n/\sigma^2 + 1/\tau^2}\right)$$



1.3 Bayes Estimators

Given $\pi(\theta|x_1, \dots, x_n)$, we require a mechanism to choose a reasonable estimator $\hat{\theta}$. Suppose the true parameter is θ_0 which is unknown. Let a be our guess for it. In real life we may not have $a = \theta_0$. Then it is sensible to measure the penalty we have to pay for guessing incorrectly. The penalty may be measured by $(a - \theta_0)^2$ or $|a - \theta_0|$ or some other function. We should choose that value of a which minimizes the expected loss $E[L(a, \theta)]$, sometimes called the **risk**, where the expectation is taken with respect to the posterior distribution $\pi(\theta|x_1, \dots, x_n)$ of θ. Note that a should not be a function of θ, rather it should be a function of $x_1, \dots, x_n$, the random sample. The minimizer, $\hat{\theta}$ say, is called the Bayes estimator of θ.

1.3.1 Squared Error Loss Function

We consider the loss function:

$$L(a, \theta) = (a - \theta)^2.$$

See what happens to the expected loss = the risk, for squared error loss. Let

$$b = E_{\pi(\theta|x_1, x_2, \dots, x_n)}(\theta) = \int \theta \pi(\theta|x_1, x_2, \dots, x_n) d\theta.$$

$$\begin{array}{rcl} E[L(a,\theta)] & = & \int L(a,\theta)\pi(\theta|x_1, \dots, x_n)d\theta \\ & = & \int (a - b + b - \theta)^2 \pi(\theta|x_1, \dots, x_n)d\theta \\ & = & (a-b)^2 + \int (b - \theta)^2 \pi(\theta|x_1, \dots, x_n)d\theta \\ & \geq & \int (b-\theta)^2 \pi(\theta|x_1, \dots, x_n)d\theta, \end{array}$$

for any value of a. When will the above inequality be an equality? Ans: when $a = b$. Note that b is the posterior mean of θ. Hence we say that **the Bayes estimator under squared error loss is the posterior mean.**

1.3.2 Absolute Error Loss Function

What happens when we assume the absolute error loss, $L(a, \theta) = |a - \theta|$. Then by a theorem Degroot, page 210, $E[|a - \theta|]$ is minimized by taking a to be the median of the posterior distribution of θ. For this loss function the Bayes estimator is the posterior median. The median of a random variable Y with pdf $g(y)$ is defined as the value μ which solves:

$$\int_{-\infty}^{\mu} g(y)dy = \frac{1}{2}.$$

The is hard to find except for symmetric distributions. For example, for the normal example the Bayes estimator of θ under the absolute error loss is still the posterior mean because it is also the posterior median. For the other examples, we need a computer to find the posterior medians.

1.3.3 Step Function Loss

We consider the loss function:

$$\begin{aligned} L(a, \theta) &= 0 \text{ if } |a - \theta| \leq \delta \\ &= 1 \text{ if } |a - \theta| > \delta \end{aligned}$$

where δ is a given small positive number. Now let us find the expected loss, i.e. the risk. Note that expectation is to be taken under the posterior distribution.

$$\begin{aligned} E[L(a, \theta)] &= \int_{\Theta} I(|a - \theta| > \delta)\pi(\theta|\mathbf{x})d\theta \\ &= \int_{\Theta} (1 - I(|a - \theta| \leq \delta))\, \pi(\theta|\mathbf{x})d\theta \\ &= 1 - \int_{a-\delta}^{a+\delta} \pi(\theta|\mathbf{x})d\theta \\ &\approx 1 - 2\delta\pi(a|\mathbf{x}) \end{aligned}$$

where $I(\cdot)$ is the indicator function. In order to minimise the risk we need to maximise $\pi(a|\mathbf{x})$ with respect to a and the Bayes estimator is the maximiser.

Therefore, the Bayes estimator is that value of θ which maximises the posterior, i.e. the modal value. This estimator is called the maximum a-posteriori (MAP) estimator.

♡ **Example** 1.5. **Binomial** As in Example 1.2. The Bayes estimator under squared error loss is

$$\hat{\theta} = \frac{x+\alpha}{x+\alpha+n-x+\beta} = \frac{x+\alpha}{n+\alpha+\beta}.$$

♡ **Example** 1.6. **Exponential** As in Example 1.3. The Bayes estimator is

$$\hat{\theta} = \frac{n+1}{\mu + \sum_{1=1}^{n} x_i}.$$

Note that μ is a given constant.

♡ **Example** 1.7. **Normal** As in Example 1.4. The Bayes estimator under all three loss functions is

$$\hat{\theta} = \frac{n\bar{x}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2}$$

♡ **Example** 1.8. Let $X_1, \ldots, X_n \sim \mathit{Poisson}(\theta)$. Also suppose the prior is $\pi(\theta) = e^{-\theta}, \theta > 0$. The likelihood is:

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} \frac{1}{x_i!} e^{-\theta} \theta^{x_i} = e^{-n\theta} \theta^{\sum_{i=1}^{n} x_i} \frac{1}{x_1! x_2! \cdots x_n!}$$

Now try to get the posterior distribution of θ as the Likelihood times the prior. We only need to collect the terms involving θ only.

$$\begin{aligned} \pi(\theta|x_1, \ldots, x_n) &\propto e^{-n\theta} \theta^{\sum_{i=1}^{n} x_i} e^{-\theta} \\ &\propto e^{-(n+1)\theta} \theta^{\sum_{i=1}^{n} x_i} \end{aligned}$$

which is easily seen to be the pdf of Gamma$\left(1 + \sum_{i=1}^{n} x_i, \frac{1}{n+1}\right)$. Hence the Bayes estimator of θ under squared error loss is:

$$\hat{\theta} = \text{ Posterior mean } = \frac{1 + \sum_{i=1}^{n} x_i}{1+n}.$$

1.4 Credible Regions

Choose a set A such that $P(\theta \in A|\mathbf{x}) = 1-\alpha$. Such a set A is called $100(1-\alpha)\%$ credible region for θ.

The set A is called a *Highest Posterior Density* (HPD) credible region if $\pi(\theta|\mathbf{x}) \geq \pi(\psi|\mathbf{x})$ for all $\theta \in A$ and $\psi \notin A$.

♡ **Example** 1.9. **Normal–Normal** Also find a 95% HPD credible region for θ.

If we have a hypothesis concerning θ, e.g. $H : \theta \leq c$ for some known value of c we can calculate

$$P(H \text{ is true}|\mathbf{x}) = P(\theta \leq c|\mathbf{x}) = \int_{-\infty}^{c} \pi(\theta|\mathbf{x}) d\theta.$$

Simple hypothesis such as $H : \theta = c$ requires a different approach. We shall return to this later.

1.5 Non-Bayesian Inference Procedures

The essential difference between Bayesian and frequentist statistics is that: “Bayesian statistics directly produces statements about the uncertainty of unknown quantities, either parameter or future observations, conditional on known data; frequentist statistics produces probability statements about hypothetical repetitions of the data conditional on the unknown parameter”. The p-values or confidence intervals are **largely irrelevant** once the sample has been observed, since they are concerned with events which might have occurred but have not. Indeed, Bernardo and Smith quote Jeffreys:

> ...a hypothesis which may be true may be rejected because it has not predicted observable results which have **not** occurred. This seems a remarkable procedure.

We examine several cases demonstrating this phenomenon.

1.5.1 Confidence Intervals

Confidence intervals obtained from classical perspectives have lot of problems. These are not probability intervals. These do not exist for arbitrary confidence levels when the model is discrete.

1. Let X_1, X_2 be a random sample of size 2 from $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$. Let $Y_1 = \min(X_1, X_2)$ and $Y_2 = \max(X_1, X_2)$. Then it is easy to see that

$$P(Y_1 < \theta < Y_2) = 0.5,$$

so that (Y_1, Y_2) is a 50% confidence interval for θ. However, if for the observed data it turns out that $y_2 - y_1 \geq 0.5$ then certainly $y_1 < \theta < y_2$, so that we know *for sure* that θ belongs to the interval (y_1, y_2), even though the confidence level is only 50%.

2. Suppose

$$Pr\{A(\mathbf{X}) < \theta < B(\mathbf{X})\} = 1 - \alpha.$$

If $A(\mathbf{x}) = a$ and $B(\mathbf{x}) = b$, then it is said that the interval (a, b) is a *confidence interval for θ with confidence coefficient* $1 - \alpha$. It is NOT correct to say that θ lies in the interval (a, b) with *probability* $1 - \alpha$. We can make the the statement *before* we have observed the data. After the specific values of $A(\mathbf{X})$ and $B(\mathbf{Y})$ are observed, the probability statement is no longer valid. However, if we take the Bayesian view, we interpret θ as a random variable and no problem arises. See Degroot (p400) for more details.

1.5.2 Unbiasedness

Recall that an estimator, $T(\mathbf{X})$, is unbiased for θ if $E[T(\mathbf{X}|\theta)] = \theta$. Most Bayes estimators are biased. See this in earlier examples. Does this matter?

Unbiasedness requires us to be able to specify S, the sample space. That is, not just what was observed, but also what wasn't.

♡ **Example** 1.10. **The taste test** In an experiment to determine whether an individual possesses discriminating powers, she has to identify correctly which of the two brands she is provided with, over a series of trials.

Let θ denote the probability of her choosing the correct brand in any trial and X_i be the Bernoulli r.v. taking the value 1 for correct guess in the ih trial. Suppose that in first 6 trials the results are 1, 1, 1, 1, 1, 0.

Case 1: Suppose that n, the number of trials, is fixed in advance. Then $\hat{\theta} = \frac{X}{n}$ where X is the total number of correct guesses. This is unbiased for θ.

Case 2: Suppose that the sampling design is to continue the trials until first zero (geometric sampling). Then $\hat{\theta} = \frac{X}{X+1}$. This is not unbiased for θ.

Also for many real problems, it is impossible to find unbiased estimators. Often the maximum likelihood estimators (mle) are used. But, in general, these do not provide unbiased estimators.

For example, the mle of σ^2 in a normal (with both parameters unknown) problem is biased.

♡ **Example** 1.11. Let X ~Poisson(θ). The **only** unbiased estimator of $g(\theta) = e^{-(k+1)\theta},\ k > 0$ is $T(X) = (-k)^X$. Hence, $T(x) > 0$ if x is even and < 0 if x is odd!

1.5.3 Likelihood Principle

Consider two experiments yielding, respectively data $\mathbf{y}$ and $\mathbf{z}$ with model representation involving the same parameter $\boldsymbol{\theta} \in \Theta$ and proportional likelihoods:

$$f(\mathbf{y}|\boldsymbol{\theta}) = g(\mathbf{y}, \mathbf{z}) f(\mathbf{z}|\boldsymbol{\theta}).$$

The *likelihood principle* says that the experiments produce same conclusion about θ. It is a trivial consequence of the Bayes theorem if we assume the same prior for $\boldsymbol{\theta}$. However, the frequentist procedure typically violates the principle, since long run behavior under hypothetical repetitions depends on the entire distribution $\{p(\mathbf{y}|\boldsymbol{\theta}), \mathbf{y} \in \mathcal{Y})\}$ where $\mathcal{Y}$ is the sample space and not only on the likelihood. The pure likelihood approach, i.e., the attempt to produce inferences solely based on the likelihood function breaks down immediately when there are nuisance parameters. The use of "marginal likelihoods" necessarily requires the elimination of nuisance parameters, but the suggested procedures for doing this, seem hard to justify in terms of the likelihood approach.

1.5.4 Sufficiency

Classical procedures are based heavily on sufficient statistics. In our context it is a simple consequence of Bayes theorem. Examine the posterior distribution with the help of the above factorization theorem. Two comments:

- Sufficiency is a global concept. Example: with univariate normal we say $(\bar{x}, s^2)$ is jointly sufficient for (μ, σ^2), but $\bar{x}$ is not sufficient for μ, nor is s^2 sufficient for σ^2.

- Sufficiency is a concept relative to a model; thus even a small perturbation to the assumed model may destroy sufficiency. For example, $(\bar{x}, s^2)$ is *not* sufficient if the true model is a t distribution with mean μ, scale parameter σ^2 and degrees of freedom 1000.

1.5.5 P-value

This is often incorrectly interpreted as the *probability* that H_0 is true is smaller than the p-value. We have seen how to find such probability under the Bayesian setup.

♡ **Example** 1.12. **Return to the taste test** Suppose the problem is to test $H_0 : \theta = \frac{1}{2}$ against $H : \theta > \frac{1}{2}$.

Case 1: Under binomial sampling,

$$\begin{aligned} \text{p-value} &= P(X = 5 \text{ or something more extreme } |\theta = \tfrac{1}{2}) \\ &= P(X = 5 \text{ or } X = 6|\theta = \tfrac{1}{2}) \\ &= 7 \times \left(\tfrac{1}{2}\right)^6 = 0.109. \end{aligned}$$

Case 2: Under geometric sampling,

$$\begin{aligned} \text{p-value} &= P(X = 5 \text{ or something more extreme } |\theta = \tfrac{1}{2}) \\ &= P(X = 5, 6, 7, \ldots |\theta = \tfrac{1}{2}) \\ &= \left(\tfrac{1}{2}\right)^6 + \left(\tfrac{1}{2}\right)^7 + \ldots \\ &= 0.031. \end{aligned}$$

Despite exactly the same sequence of events being observed, different inferences are made! Recall Jeffreys quotation again.

There are serious difficulties in incorporating any known restrictions in the parameter space. No general method is available. Multi-parameter situation poses problems. Nuisance parameters create even bigger problems. Plug-in things (*ad-hoc* approximations based on substituting estimates for parameters) should be treated with caution. Estimation of future observation is even more problematic.

1.6 Reading List

Read the Chapter 1 of Box and Tiao. It provides a gentle and easy to understand introduction. For arguments between the frequentist theory and Bayes read the Appendix B of Bernardo and Smith. You will already be familiar with some of the things they are saying there. Last but not least, read the referenced pages of Degroot to clear up the ideas.

Chapter 2

Priors, Predictions and Model Choice

2.1 Prior Distributions

If prior information from previous experimentation is available, use them. There are many methods of prior selection and elicitation based on expert opinions.

2.1.1 Conjugate Priors

Suppose that we have a hierarchical model $f(\mathbf{y}|\theta)$: the likelihood; $\pi(\theta|\eta)$ the prior. If $\pi(\theta|\mathbf{y},\eta)$ belongs to the same parametric family as $\pi(\theta|\eta)$, then we say that $\pi(\theta|\eta)$ is a conjugate prior for θ. In these cases if we assume that η is known, the analysis becomes much easier. Natural conjugacies:

Likelihood	Prior
Binomial	Beta
Poisson	Gamma
Normal	Normal
Exponential	Gamma

2.1.2 Locally Uniform Priors

What if one has no prior information with which to choose $\pi(\theta)$? Although this is rare in practice, this type situations can be overcome by the use of what are called non-informative (vague, diffuse, flat) priors.

A basic property of a pdf is that it integrates to 1, i.e. $\int \pi(\theta)d\theta = 1$. Sometimes we assume prior distributions which are constant over the whole real line. For example,

$$\pi(\theta) = k,\ k > 0,\ -\infty < \theta < \infty.$$

This pdf violates the above condition. This would be called an **improper** prior distribution. It is alright to assume improper prior distributions only if the resulting posterior distribution is proper, i.e. $\int_\Theta \pi(\theta|\mathbf{x})d\theta < \infty$. Further, suppose that $\pi(\theta) = k$ only for values of θ where the likelihood function has appreciable value, and $\pi(\theta) = 0$ otherwise. This $\pi(\theta)$ will then define a proper density and no theoretical problem arises. Prior distributions like the above are called locally uniform priors.

2.1.3 Non-informative priors

If a prior distribution $\pi(\theta)$ does not contain any information for θ, it is called a *non-informative prior*. Most widely used non-informative priors are Jeffreys (1961) priors:

$$\pi(\theta) = \{I(\theta)\}^{1/2}$$

where $I(\boldsymbol{\theta})$ is the Fisher information

$$I(\theta) = -E_\theta\left[\frac{\partial^2}{\partial\theta^2} f(\mathbf{x}|\theta)\right].$$

Note that we obtain improper priors in most situation. We have to guarantee that the resulting posterior is **proper**.

There is a huge literature on prior selection. Box and Tiao (Section 1.3) would be a good start. In our course we will assume (loosely) vague or flat priors, i.e., priors which are locally uniform.

♡ **Example** 2.1. **Binomial** For the binomial example show that

$$\pi(\theta) = \{\theta(1-\theta)\}^{-\frac{1}{2}}.$$

♡ **Example** 2.2. **Normal** For $N(0, \sigma^2)$ problem show that

$$\pi(\sigma^2) = \frac{1}{\sigma^2}.$$

2.2 Predictive Distributions

"What is the probability that the sun will rise tomorrow, given that it has risen without fail for the last n days?" In order to answer questions like these we need to learn what are called predictive distributions.

2.2.1 Posterior Predictive Distribution

Let $X_1, \ldots, X_n$ be an i.i.d. sample from the distribution $f(x|\theta)$. Let $\pi(\theta)$ be the prior distribution and $\pi(\theta|\mathbf{x})$ be the posterior distribution. We want the distribution (pdf or pmf) of $X_{n+1}|X_1, \ldots, X_n$. The given notation is to denote that $X_1, \ldots, X_n$ have already been observed, like the sun has risen for the last n days. We define the **posterior predictive distribution** to be:

$$f(x_{n+1}|x_1, \ldots, x_n) = \int_\Theta f(x_{n+1}|\theta)\pi(\theta|x_1, \ldots, x_n)d\theta. \tag{2.1}$$

It uses the conditional independence of x_{n+1} and $\mathbf{x}$ given θ. It is the density of a future observation given everything else, i.e., the 'model' and the observations. Intuitively, if θ is known then x_{n+1} will follow $f(x_{n+1}|\theta)$ since it is from the same population as $x_1, \ldots, x_n$ are. We do not know θ but the posterior $\pi(\theta|\mathbf{x})$ contains all that we know about θ. Therefore, the predictive distribution is obtained as an average over $\pi(\theta|\mathbf{x})$. Hence the definition. We now derive some predictive distributions.

♡ **Example** 2.3. We return to the sun example. Let

$$\begin{aligned} X_i &= 1 \quad \text{if its sunny on the } i\text{th day,} \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Note that X_{n+1} will be binary as well. We want $P[X_{n+1} = 1|\mathbf{x} = (1, 1, \ldots, 1)]$. Assume $f(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$, and X_i are independent. Therefore, the likelihood is

$$\begin{aligned} f(\mathbf{x}|\theta) &= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}, \\ &= \theta^n \text{ if } \mathbf{x} = (1, 1, \ldots, 1). \end{aligned}$$

Let us assume a uniform prior for θ, i.e. $\pi(\theta) = 1$ if $0 < \theta < 1$. Now the posterior is:

$$\begin{aligned} \pi(\theta|\mathbf{x}) &= \frac{\theta^n}{\int_0^1 \theta^n d\theta} \\ &= (n+1)\theta^n. \end{aligned}$$

Here $f(X_{n+1} = 1|\theta) = \theta$. Finally we can evaluate the posterior predictive distribution using (2.1).

$$\begin{aligned} P(X_{n+1} = 1|\mathbf{x}) &= \textstyle\int_0^1 \theta(n+1)\theta^n d\theta \\ &= (n+1)\textstyle\int_0^1 \theta^{n+1} d\theta \\ &= \tfrac{n+1}{n+2}. \end{aligned}$$

Intuitively, this probability goes to 1 as $n \to \infty$.

2.2.2 Prior Predictive Distribution

We sometimes need to define what is called the **prior predictive distribution** defined as

$$f(x) = \int_\Theta f(x|\theta)\pi(\theta)\,d\theta. \tag{2.2}$$

Note that it is simply the normalising constant in $\pi(\theta|x)$. It is also called the marginal distribution of the data. And it is of the same form as the posterior predictive distribution (2.1). The prior predictive distribution is obtained by replacing the posterior $\pi(\theta|x_1,\dots,x_n)$ by the prior $\pi(\theta)$ in (2.1).

With n samples, we define the (**joint**) prior predictive distribution of $x_1,\dots,x_n$ as

$$f(\mathbf{x}) = \int_\Theta f(\mathbf{x}|\theta)\pi(\theta)\,d\theta. \tag{2.3}$$

♡ **Example** 2.4. We return to the normal example. Suppose $X_1,\dots,X_n \sim N(\theta,\sigma^2)$, $\pi(\theta) \sim N(\mu,\tau^2)$ for known μ and τ^2. We had

$$\pi(\theta|\mathbf{x}) = N\left(\frac{n\bar{x}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2}, \frac{1}{n/\sigma^2 + 1/\tau^2}\right).$$

Satisfy yourself that X_{n+1} follows a normal distribution with

$$\text{mean } = \frac{n\bar{x}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2} \text{ and variance } \sigma^2 + \frac{1}{n/\sigma^2 + 1/\tau^2}.$$

For this example, the prior predictive distribution is,

$$f(\mathbf{x}) = \int \prod_{i=1}^n N(x_i|\theta,\sigma^2)N(\theta|\mu,\tau^2)d\theta.$$

For this distribution show that $E(X_i) = \mu$ and $V(X_i) = \tau^2 + \sigma^2$. Are X_i & X_j marginally independent? No, they have covariance τ^2. (Derive it!)

You may find the following useful:

Result For any two random variable with finite variances:

$$E(X) = EE(X|Y), \quad \text{Var}(X) = E\text{Var}(X|Y) + \text{Var}(E(X|Y)).$$

2.3 Model Choice

2.3.1 Bayes Factors

Suppose that we have to choose between two hypotheses H_0 and H_1 corresponding to assumptions of alternative models M_0 and M_1 for data $\mathbf{x}$. The likelihoods are denoted by $f_i(\mathbf{x}|\theta_i)$ and the priors by $\pi_i(\cdot), i = 0, 1$ in the following discussion. In many cases, the competing models have a common set of parameters, but this is not necessary; hence the notations f_i, π_i and θ_i. Recall that the prior predictive distribution (2.3) for model i is,

$$f(\mathbf{x}|M_i) = \int f_i(\mathbf{x}|\theta_i)\pi_i(\theta_i)d\theta_i.$$

Bayes factor is defined as:

$$B_{01}(\mathbf{x}) = \frac{f(\mathbf{x}|M_0)}{f(\mathbf{x}|M_1)}. \tag{2.4}$$

Note that the Bayes factor is the ratio of the marginal likelihood under two different models. Hence, intuitively $B_{01}(\mathbf{x}) > 1$ implies that M_0 is more relatively plausible in the light of $\mathbf{x}$. (Some authors use 3 as some sort of cut-off point.)

♡ **Example** 2.5. **Geometric versus Poisson** Suppose that:

$$M_0 : X_1, X_2, \ldots, X_n|\theta_0 \sim f_0(x|\theta_0) = \theta_0(1-\theta_0)^x, \quad x = 0, 1, \ldots .$$

$$M_1 : X_1, X_2, \ldots, X_n|\theta_1 \sim f_1(x|\theta_1) = e^{-\theta_1}\theta_1^x/x!, \quad x = 0, 1, \ldots .$$

Further, assume that θ_0 and θ_1 are known. How should we decide between the two models based on $x_1, x_2, \ldots, x_n$?

Since the parameters are known under the models, we do not need to assume any prior distributions for them. Consequently,

$$f(\mathbf{x}|M_0) = \theta_0^n(1-\theta_0)^{n\bar{x}}.$$

and

$$f(\mathbf{x}|M_1) = e^{-n\theta_1}\theta_1^{n\bar{x}}/\prod_{i=1}^{n} x_i!.$$

Now the Bayes factor is just the ratio of the above two. To illustrate, let $\theta_0 = 1/3$ and $\theta_1 = 2$ (then the two distributions have same mean). Now if $n = 2$ and $x_1 = x_2 = 0$ then $B_{01}(\mathbf{x}) = 6.1$, however if $n = 2$ and $x_1 = x_2 = 2$ then $B_{01}(\mathbf{x}) = 0.3$.

Why it is called a factor? Let $P(M_i)$ denote the prior probability for model i. Let us now calculate the posterior probability of M_i given the data using the Bayes theorem.

$$P(M_i|\mathbf{x}) = \frac{P(M_i)f(\mathbf{x}|M_i)}{\sum_{j=0}^{1} P(M_j)f(\mathbf{x}|M_j)}.$$

So the posterior odds ratio of the two models is given by

$$\frac{P(M_0|\mathbf{x})}{P(M_1|\mathbf{x})} = \frac{P(M_0)}{P(M_1)} \times \frac{f(\mathbf{x}|M_0)}{f(\mathbf{x}|M_1)}.$$

Now in words,

posterior odds ratio = prior odds ratio × the Bayes factor

That is why it is called a factor! Seen in this light we can define

$$\text{Bayes factor} = \frac{\text{posterior odds ratio}}{\text{prior odds ratio}}.$$

Intuitively, the Bayes factor provides a measure of whether the data $\mathbf{x}$ have increased or decreased the odds on M_0 relative to M_1. Thus $B_{01}(\mathbf{x}) > 1$ signifies that M_0 is more relatively plausible in the light of $\mathbf{x}$.

Remark: We do not need to know the prior probabilities $P(M_i), i = 0, 1$ to calculate the Bayes factor. Those are needed if we wished to calculate the posterior probability $P(M_i|\mathbf{x})$. If two models are equally likely a-priori, then Bayes factor is equal to the posterior odds ratio.

2.3.2 Hypothesis Testing

Suppose that we wish to test

$$H_0 : \theta \in \Theta_0 \text{ versus } H_A : \theta \in \Theta_1.$$

Let $f(\mathbf{x}|\theta)$ denote the likelihood of $\mathbf{x}$ given θ. Special forms:

$$\begin{array}{ll} B_{01}(\mathbf{x}) = \frac{f(\mathbf{x}|\theta_0)}{f(\mathbf{x}|\theta_1)} & \text{(simple versus simple test)} \\ B_{01}(\mathbf{x}) = \frac{f(\mathbf{x}|\theta_0)}{\int_{\Theta_1} f(\mathbf{x}|\theta)\pi_1(\theta)d\theta} & \text{(simple versus composite test)} \\ B_{01}(\mathbf{x}) = \frac{\int_{\Theta_0} f(\mathbf{x}|\theta)\pi_0(\theta)d\theta}{\int_{\Theta_1} f(\mathbf{x}|\theta)\pi_1(\theta)d\theta} & \text{(composite versus composite test)} \end{array}$$

♡ **Example** 2.6. **Taste-test** We wish to test that the tester does not have any discriminatory power against the alternative that she does. So our problem is:

$$H_0 : \theta = \frac{1}{2} \text{ versus } H_1 : \theta > \frac{1}{2}.$$

It is a simple versus composite case and we have $\Theta_0 = \frac{1}{2}$ and $\Theta_1 = (\frac{1}{2}, 1)$. Let us assume uniform prior on θ under the alternative. So the prior $\pi_1(\theta) = 2$ if $\frac{1}{2} < \theta < 1$. Recall that we had 6 Bernoulli trials with the results, 1, 1, 1, 1, 1, 0. Now the Bayes factor is

$$B_{01}(\mathbf{x}) = \frac{\frac{1}{2}^6}{\int_{\frac{1}{2}}^1 \theta^5(1-\theta)2\,d\theta} = \frac{1}{2.86}.$$

This suggests that she does appear to have some discriminatory power but not a lot.

2.4 Multi-parameter Situation

2.4.1 Basic Methods

In most realistic applications of statistical models, there are more than one unknown parameters. In principle, everything proceeds as before, except

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix},$$

where p is the number of parameters. We still start with

$$\pi(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}).$$

How do we summarise $\pi(\boldsymbol{\theta}|\mathbf{x})$? We use multi-variate statistical tools you are already familiar with. For example we can obtain the marginal posterior distribution of θ_1 as

$$\pi(\theta_1|\mathbf{x}) = \int \cdots \int \pi(\boldsymbol{\theta}|\mathbf{x})\, d\theta_2 d\theta_3 \ldots d\theta_p.$$

So we can calculate features of the above distribution, for example $E(\theta_1|\mathbf{x})$ and $\text{Var}(\theta_1|\mathbf{x})$. Also for example we can study correlations between θ_1 and θ_2.

♡ **Example** 2.7. Suppose that $X_1, X_2, \dots, X_n$ are i.i.d. $N(\theta, \sigma^2)$ and $\pi(\mu, \sigma^2) = \frac{1}{\sigma^2}$. Obtain the joint and the marginal posterior distributions of μ and σ^2.

♡ **Example** 2.8. **Pump Failure Data** The data set given below relates to 10 power plant pumps. The number of failures, y_i, follows a Poisson distribution with mean $\lambda_i = \theta_i t_i$ where θ_i is the failure rate for pump $i, i = 1, \dots, 10$ and t_i is the length of operation time of the pump (in 1000s of hours). A conjugate gamma prior distribution with density $\beta^\alpha \theta_i^{\alpha-1} e^{-\beta\theta_i}/\Gamma(\alpha)$ is adopted for each θ_i, where $\alpha = 1.802$ and $\beta = 0.1$.

Obtain the marginal posterior distribution of θ_1.

Pump	1	2	3	4	5	6	7	8	9	10
t_i	94.3	15.7	62.9	126	5.24	31.4	1.05	1.05	2.1	10.5
y_i	5	1	5	14	3	19	1	1	4	22

2.4.2 Nuisance Parameters

Suppose we partition $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\eta})$ and we are interested in $\boldsymbol{\gamma}$. How do we proceed? Review classical methods. We are given $L(\boldsymbol{\gamma}, \boldsymbol{\eta}; \mathbf{x})$. If we are lucky, there exists a sufficient statistics $\mathbf{T}$ such that

$$\mathbf{X}|\mathbf{T} \sim f(\mathbf{x}|\boldsymbol{\gamma}, \mathbf{T})$$

then base inference on the distribution of $\mathbf{T}$. This is called conditional inference. Otherwise, plug in maximum likelihood estimate $\hat{\boldsymbol{\eta}}(\boldsymbol{\gamma})$ of $\boldsymbol{\eta}$ in $L(\boldsymbol{\gamma}, \boldsymbol{\eta}; \mathbf{x})$. This is the 'profile likelihood' technique. In a third scenario, if we are able to work out the integral,

$$\int L(\boldsymbol{\gamma}, \boldsymbol{\eta}; \mathbf{x}) d\boldsymbol{\eta} = L(\boldsymbol{\gamma}; \mathbf{x}), \text{ say}$$

then use $L(\boldsymbol{\gamma}; \mathbf{x})$ to make inference. This is the 'marginal likelihood' technique.

From a Bayesian viewpoint, we have $\pi(\boldsymbol{\gamma}, \boldsymbol{\eta}|\mathbf{x})$. We use,

$$\pi(\boldsymbol{\gamma}|\mathbf{x}) = \int \pi(\boldsymbol{\gamma}, \boldsymbol{\eta}|\mathbf{x}) d\boldsymbol{\eta}.$$

This will be routinely done by numerical methods to be developed in the next chapter.

Chapter 3

Bayesian Computation

Notation and Setup:

We use the generic notation $\mathbf{x}$ to denote the parameters. We will use the notation $\pi(\mathbf{x})$ to denote the posterior distribution. We are suppressing the data part. This $\pi(\mathbf{x})$ is sometimes called the **target** distribution.

- $\mathbf{x}$: Parameters, earlier notation was $\boldsymbol{\theta}$.
- $f(\mathbf{x})$: Non-normalized posterior density. $f(\mathbf{x}) =$ Likelihood × Prior.
- $\pi(\mathbf{x}) = \frac{f(\mathbf{x})}{\int f(\mathbf{x})d\mathbf{x}}$: Normalized posterior density.
- $b(\mathbf{x})$: Any generic function of interest.

Problem: Evaluate features of $b(\mathbf{x})$, e.g.,

$$E_\pi(b) = \int b(\mathbf{x})\pi(\mathbf{x})d\mathbf{x} = \frac{\int b(\mathbf{x})f(\mathbf{x})d\mathbf{x}}{\int f(\mathbf{x})d\mathbf{x}}. \tag{3.1}$$

Evaluating the above under a multidimensional non-normalized situation is quite difficult. Thats why we want to develop machinery. Let us start with a very simple example.

♡ **Example** 3.9. Suppose $\mathbf{x}$ is one dimensional and x follows a normal distribution with mean 0 and variance 1. We are interested in the mean of x. We know the mean here is 0. But we are going to learn how to obtain this by using a computer. Then:

- $f(\mathbf{x}) = \exp\left(-\frac{x^2}{2}\right)$

- $\pi(\mathbf{x}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$
- $b(\mathbf{x}) = x.$

3.1 Numerical Integration

Suppose the problem is to evaluate

$$I = \int_a^b h(x)dx.$$

(In our context $h(x) = b(x)\pi(x)$.) Then the *trapezoidal rule* is the following. Let $x_1 = a$ and $x_n = b$. We take a grid $a = x_1 < x_2 < \cdots < x_n = b$. Then use

$$\hat{I} = \frac{1}{2}\sum_{i=1}^{n-1}(x_{i+1} - x_i)\{h(x_i) + h(x_{i+1})\},$$

to approximate I. There are other variations using what are called Simpson's rule. However, as our models become more complex it becomes increasingly difficult to perform the required integrations by this method. If the dimension of the posterior distribution is quite low e.g., 1, 2 or 3 this method may work well.

3.2 Laplace Approximation

Here the posterior is first approximated by a normal distribution. Then the integral (3.1) can be evaluated sometimes analytically by replacing $\pi(\mathbf{x})$ by its normal approximation. Let $L_n(\mathbf{x}) = \log \pi(\mathbf{x})$. Let $\mathbf{x}_0$ denote the mode of $\pi(x)$, i.e.

$$L_n'(\mathbf{x}_0) = \left.\frac{\partial L_n'(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{x}_0} = 0.$$

Assume the existence and positive-definiteness of

$$\Sigma = \left(-L_n''(\mathbf{x}_0)\right)^{-1}$$

where $L_n''(\mathbf{x}_0)$ is the Hessian matrix with elements

$$\left[L_n''(\mathbf{x}_0)\right]_{ij} = \left.\frac{\partial^2 L_n(\mathbf{x})}{\partial x_i \partial x_j}\right|_{\mathbf{x}=\mathbf{x}_0}.$$

Once we are able to find $\mathbf{x}_0$ and Σ we can replace $\pi(\mathbf{x})$ by the normal distribution with mean $\mathbf{x}_0$ and variance-covariance matrix Σ.

3.3 Monte Carlo Integration

Suppose that we can draw samples $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}$ from $\pi(\mathbf{x})$. Then we can estimate

$$E_\pi[b(\mathbf{x})] \approx \bar{b}_N = \frac{1}{N}\sum_{i=1}^{N} b\left(\mathbf{x}^{(i)}\right). \tag{3.2}$$

We can use the laws of large numbers to show that $\bar{b}_N$ approaches $E_\pi[b(\mathbf{x})]$. This is *Monte Carlo integration.* Think of plenty of examples here like the 95th percentile etc.

♡ **Example** 3.10. If we are interested in the ratio: $b(\mathbf{x}) = \frac{x_1}{x_2}$, then

$$E_\pi\left[\frac{x_1}{x_2}\right] \approx \frac{1}{N}\sum_{i=1}^{N}\frac{x_1^{(i)}}{x_2^{(i)}}$$

♡ **Example** 3.11. Kernel Density Estimates: Select a 'kernel' $k(\cdot)$ such that $\int_{-\infty}^{\infty} k(z)dz = 1$, e.g., the standard normal density. Given the samples $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$, form

$$\hat{f}_N(z) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h}k\left(\frac{z - x^{(i)}}{h}\right).$$

Evaluate this at a number of grid points (z) on the support and plot it. Thats how kernel density estimates are formed. h is called the window width or the smoothing parameter. There are many different optimal choices for h. Recommended one is $h = 1.06sN^{-1/5}$ where s is the sample standard deviation. A software package like S-Plus can do this with a command or two.

3.4 Non-iterative Monte Carlo

The methods based on importance sampling, rejection method and weighted bootstrap method fall in this category.

Importance Sampling: Suppose that $g(\mathbf{x})$ is an importance sampling density (ISD) for $\pi(\mathbf{x})$. The density g should have two properties.

1. g resembles π, i.e., g looks like π

2. g is easy to sample from.

Implicitly we also assume that g and π have the same support. See Figure 3.1. Then we can evaluate $E_\pi(b)$ as follows.

$$E_\pi(b) = \frac{\int b(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}}{\int f(\mathbf{x}) d\mathbf{x}} = \frac{\int b(\mathbf{x}) \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x}) d\mathbf{x}}{\int \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x}) d\mathbf{x}} = \frac{\int b(\mathbf{x}) w(\mathbf{x}) g(\mathbf{x}) d\mathbf{x}}{\int w(\mathbf{x}) g(\mathbf{x}) d\mathbf{x}}.$$

Suppose that we have N samples, $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}$ from the distribution $g(\mathbf{x})$. Clearly,

$$\hat{b} \overset{def}{=} \frac{\sum b(\mathbf{x}^{(i)}) w_i}{\sum w_i}, \text{ where } w_i = \frac{f(\mathbf{x}^{(i)})}{g(\mathbf{x}^{(i)})},$$

provide a Monte Carlo integration for $E_\pi(b)$. Notice that in effect we have done two Monte Carlo integrations: one for the numerator and the other for the denominator. For $\hat{b}$ to be a good estimate of $E_\pi(b)$ we need w's to be well behaved, i.e., should be roughly equal. This is very hard to do when dimensionality is high. We can refine the estimator of b in two ways.

Rejection Sampling: Suppose that,

$$M = \sup_{\mathbf{x}} \frac{f(\mathbf{x})}{g(\mathbf{x})}$$

is available. Draw $\mathbf{x}^* \sim g$ and $u \sim U(0,1)$. Retain $\mathbf{x}^*$ if $u \leq \frac{f(\mathbf{x}^*)}{Mg(\mathbf{x}^*)}$. This is the **rejection** method. Satisfy yourself that a sample drawn this way has the exact distribution π.

Proof First see:

$$P\left(\mathbf{x}^* \text{ retained } | \mathbf{x}^* \sim g\right) = \frac{f(\mathbf{x}^*)}{Mg(\mathbf{x}^*)}.$$

Hence

$$P\left(\mathbf{x}^* \text{ retained }\right) = \int \frac{f(\mathbf{x}^*)}{Mg(\mathbf{x}^*)} g(\mathbf{x}^*) d\mathbf{x}^* = \frac{1}{M} \int f(\mathbf{x}^*) d\mathbf{x}^*.$$

Now see the effect of M on retaining a sample. Our aim is to prove that $\mathbf{x}^*$ retained through the rejection method has the distribution π.

$$\begin{aligned} P\left(\mathbf{x}^* < c | \mathbf{x}^* \text{ retained }\right) &= \frac{P(\mathbf{x}^* < c \text{ and } \mathbf{x}^* \text{ retained })}{P(\mathbf{x}^* \text{ retained })} \\ &= \frac{\int_{\mathbf{x}^* < c} \int_0^{\frac{f(\mathbf{x}^*)}{Mg(\mathbf{x}^*)}} g(\mathbf{x}^*) du d\mathbf{x}^*}{\frac{1}{M} \int f(\mathbf{x}^*) d\mathbf{x}^*} \\ &= \frac{\int_{\mathbf{x}^* < c} f(\mathbf{x}^*) d\mathbf{x}^*}{\int f(\mathbf{x}^*) d\mathbf{x}^*}, \end{aligned}$$

which shows that the retained $\mathbf{x}^*$ has the exact distribution π as was claimed. □

Draw N samples and form the estimator $\bar{b}_N$ in (3.2).

♡ **Example** 3.12. Suppose that we are working with proper priors. Let g be the prior distribution. Then

$$M = \sup_{\mathbf{x}} \frac{f(\mathbf{x})}{g(\mathbf{x})} = \sup \frac{\text{Likelihood } \times \text{ Prior}}{\text{Prior}} = \sup \text{ Likelihood} \;.$$

Hence M will be attained at the maximum likelihood estimate (mle). However, it is not a good idea in high dimension. For simple problems this is ideal.

Weighted Bootstrap: Draw m samples, (m is usually huge) $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}$ from the distribution $g(\mathbf{x})$. Calculate the weights w_i and then calculate the normalized weights $q_i = \frac{w_i}{\sum w_i}$. Note that $\mathbf{x}^{(i)}, q_i$ define a discrete probability distribution. A random sample $\mathbf{x}^*$ drawn from this discrete distribution is an approximate sample from π. Draw N random samples this way and form $\bar{b}_N$ given by (3.2).

♡ **Example** 3.13. Can experiment with the normal distribution in one dimension. Take π to be the standard normal distribution and g to be a normal distribution with zero mean and variance σ^2. We need to take $\sigma^2 > 1$ for good results. Why?

How do we get a g?

- CLT approximation.
- Quadratic Regression.
- West (1992) adaptive mixtures.

Remark: It is important to understand the points about all the sampling ideas described above. These together are called Sampling-Importance-Re-sampling *SIR* methodology.

3.5 Markov Chain Monte Carlo

As models become more complex in high dimension the posterior distributions become analytically intractable. All of the previous methods fail. The solution is to consider iterative Monte Carlo or Markov chain Monte Carlo. Simulate a Markov chain with stationary distribution given by the posterior distribution, $\pi(\mathbf{x})$. Features of π are discovered (accurately) by forming *ergodic averages* as in (3.2). It turns out that $\bar{b}_N$ still accurately estimates $E_\pi(b)$ if we generate samples using a Markov chain. Theorems like the CLT and laws of large numbers can be proven.

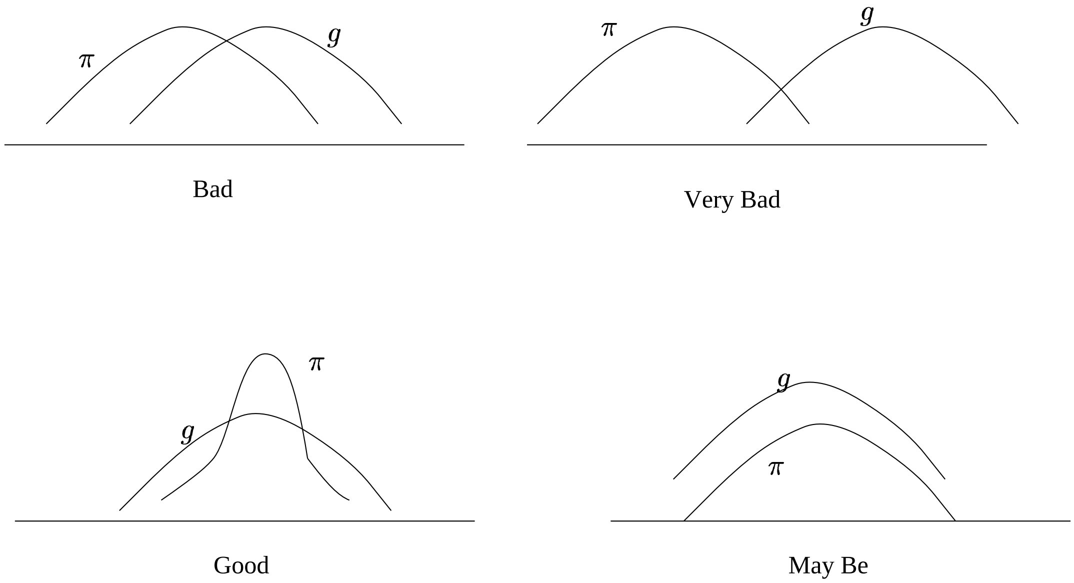


Figure 3.1: How different choices of g influences the computation.

3.5.1 Notions of Markov Chains

A Markov chain is generated by sampling

$$x^{(t+1)} \sim p(x|x^{(t)}).$$

This p is called the *transition kernel* of the Markov chain. So $x^{(t+1)}$ depends only on $x^{(t)}$, not on $x^{(0)}, x^{(1)}, \ldots, x^{(t-1)}$.

♡ **Example** 3.14.

$$x^{(t+1)} \sim N(x^{(t)}, 1)$$

Stationarity As $t \to \infty$, the Markov chain converges in distribution to its *stationary* distribution. This is also called its invariant distribution.

Irreducibility *Irreducible* means any set of states can be reached from any other state in a finite number of moves (or transitions).

Ergodicity Suppose that we have an irreducible Markov chain with stationary distribution $\pi(x)$. Then we have an *ergodic* theorem:

$$\begin{aligned} \bar{b}_N &= \tfrac{1}{N} \textstyle\sum_{i=1}^{N} b\left(\mathbf{x}^{(i)}\right) \\ &\to E_\pi[b(\mathbf{x})] \text{ as } N \to \infty. \end{aligned}$$

Can we estimate the standard error of $\bar{b}_N$? Yes! the numerical standard error is given by:

$$\text{nse}\left(\bar{b}_N\right) = \sqrt{\frac{1}{N}\text{var}_\pi(b)\left\{1 + 2\sum_{t=1}^{N-1}\rho_t(b)\right\}}$$

where $\rho_t(b)$ is the lag$-t$ correlation in $\left\{b(x^{(i)})\right\}$. Hence we can make nse as small as we like by increasing N. For independent sampling, the correlation term disappears. The correlation term can decrease the nse but usually increases it. The nse may not be finite for general Markov chains, but is finite when the Markov chain is 'good': geometrically ergodic.

3.5.2 Metropolis-Hastings

The most general algorithm known as the Metropolis-Hastings algorithm is given as follows.

1. Start anywhere and say we are at $\mathbf{x}^{(t)} = \mathbf{x}$.
2. Generate $\mathbf{y}$ from $q(\mathbf{y}|\mathbf{x})$. $\mathbf{y}$ is called a *candidate point* and q is called a *proposal distribution.*
3. Calculate $\alpha(\mathbf{x},\mathbf{y}) = \min\left\{1, \frac{\pi(\mathbf{y})q(\mathbf{x}|\mathbf{y})}{\pi(\mathbf{x})q(\mathbf{y}|\mathbf{x})}\right\}$
4. Accept $\mathbf{x}^{(t+1)} = \mathbf{y}$ with prob. $\alpha(\mathbf{x},\mathbf{y})$.
5. Else set $\mathbf{x}^{(t+1)} = \mathbf{x}$.

Note the involvement of π : the target density. It only enters through the ratio $\frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}$. Hence we do not need to know the normalizing constant to implement the algorithm.

Special cases:

1. $q(\mathbf{y}|\mathbf{x}) = q(\mathbf{x}|\mathbf{y})$: Metropolis Algorithm.
2. $q(\mathbf{y}|\mathbf{x}) = g(\mathbf{y})$: Independence Sampler.
3. $q(\mathbf{y}|\mathbf{x}) = \prod \pi(y_i|\mathbf{y}_{<i}, \mathbf{x}_{>i}) \Rightarrow \alpha(\mathbf{x},\mathbf{y}) = 1$: Gibbs Sampler.

Generally it is easier to analyse the first two versions theoretically. The last version, the Gibbs sampler however, is much easier to implement. We will try to understand each of the above.

Metropolis Algorithm: Here the proposal distribution is symmetric. i.e.,

$$q(\mathbf{y}|\mathbf{x}) = q(\mathbf{x}|\mathbf{y}).$$

A special case of this is

$$q(\mathbf{y}|\mathbf{x}) = q(|y-x|).$$

Note that for this choice we have

$$\alpha(\mathbf{x},\mathbf{y}) = \min\left\{1, \frac{\pi(\mathbf{y})}{\pi(\mathbf{x})}\right\}$$

This is called the *Random-Walk* Metropolis algorithm. Proposals depend on the current value.

♡ **Example** 3.15. Suppose that the target distribution is $\pi(x)$ is the standard normal distribution. Take $q(\mathbf{y}|\mathbf{x})$ to be the normal with mean x and variance σ^2. Experiment different values for σ^2. Calculate the percentage of acceptance. Form the estimator for the mean of the distribution.

♡ **Example** 3.16. Modify the above program to generate from $\Gamma(\alpha,\beta)$ distribution for given α and β using the same proposal distribution.

♡ **Example** 3.17. Write programs to generate from $\pi(\mathbf{x}) = N_p(\mathbf{0}, I)$, the p dimensional normal distribution.

Independence Sampler Here we choose

$$q(\mathbf{y}|\mathbf{x}) = g(\mathbf{y}).$$

Proposals are drawn from a fixed density g, just as we had done in the importance sampling case. Independence samplers are either very good or very bad. We can get independent samples from here by choosing $g = \pi$. Then $\alpha(\mathbf{x},\mathbf{y}) = 1$ and the chain will never reject a candidate. However, the above is perhaps a dream and hard to achieve in practice. Tails of $g(\cdot)$ must be heavier than tails of $\pi(\cdot)$ for geometric convergence.

♡ **Example** 3.18. $\pi(x) = N(0,1)$ is the standard normal distribution. $q(y|x) = g(\mathbf{y}) = N(0,\sigma^2)$.

3.5.3 Gibbs Sampler

Let $\mathbf{x} = (x_1, \ldots, x_k)^T$. Suppose that we start the algorithm at $\mathbf{x}^{(0)}$. This could be any point in the support of π. Given $\mathbf{x}^{(t)}$ we iterate one step as follows. To sample from the k dimensional distribution $\pi(\mathbf{x})$, the Gibbs sampler makes a Markov transition from $\mathbf{x}^{(t)}$ to $\mathbf{x}^{(t+1)}$ as follows.

$$\begin{array}{lll}
x_1^{(t+1)} & \sim & \pi(x_1|x_2^{(t)}, x_3^{(t)}, \cdots, x_k^{(t)}) \\
x_2^{(t+1)} & \sim & \pi(x_2|x_1^{(t+1)}, x_3^{(t)}, \cdots, x_k^{(t)}) \\
\vdots & \vdots & \vdots \\
x_k^{(t+1)} & \sim & \pi(x_k|x_1^{(t+1)}, x_2^{(t+1)}, \cdots, x_{k-1}^{(t+1)}).
\end{array}$$

The densities on the right hand sides of the above are called *complete conditional distributions* or the *full conditional distributions*. Note that always the most up-to-date version $\mathbf{x}$ is used.

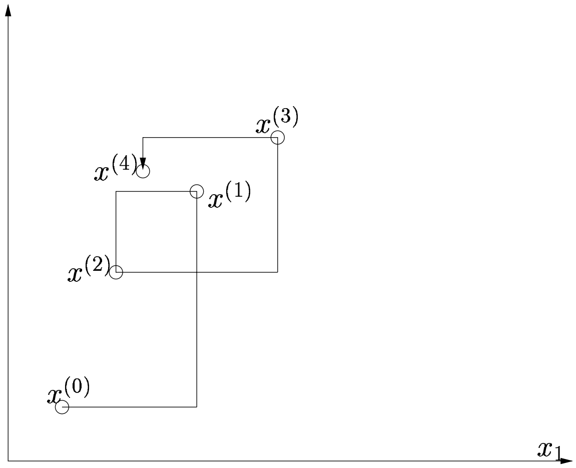


Figure 3.2: Diagram showing how the Gibbs sampler works.

♡ **Example** 3.19. Suppose $y_i \overset{iid}{\sim} N(\mu, \sigma^2), i = 1, 2, \ldots, n$. Assume flat prior for μ and σ^2 as $\pi(\mu, \sigma^2) = \sigma^{-2} I(\sigma^2 > 0)$. The posterior distribution of μ and σ^2 is given by

$$\pi(\mu, \sigma^2|\mathbf{y}) \propto \frac{1}{(\sigma^2)^{n/2+1}} \exp\left\{-\frac{1}{2}\sum\left(\frac{y_i-\mu}{\sigma^2}\right)^2\right\}.$$

The Gibbs sampler for this problem samples μ and σ^2 alternatively. It samples μ from the distribution, $N(\bar{y}, \sigma^2/n)$ and $\frac{1}{\sigma^2}$ from the distribution $\Gamma(\frac{n}{2}, \frac{1}{2}\sum\left(y_i-\mu)^2\right)$. Write a computer program to implement this Gibbs sampler. Check your answers from the exact results.

♡ **Example** 3.20. Try writing the complete conditional distributions for the pump failure data. Then write a Gibbs sampler program for sampling from the posterior distribution.

In the above examples the complete conditional distributions are easy to sample from because they are standard. Alternative strategies are used when some of the complete conditionals are non-standard. When they are *log-concave*, i.e., second derivative of the log density is strictly decreasing, the *adaptive rejection sampling* can be used. C programs are available to do this. Otherwise a Metropolis step can be used to update any complete conditional which is non-standard.

3.5.4 Implementation Issues

Theory suggests that if we run the chain long enough, it will converge to the stationary distribution. However, sometimes there can be slow convergence. Hence, some efforts are made to 'diagnose' convergence. Note that even if your Markov chain gives good convergence diagnostics you can not make the claim that you have *proved* that it has converged. Rather these diagnostics are treated as a 'feel-good-factor'.

Diagnostics 1: Visual plots Do a time series plot of the different components of the Markov chain. Calculate autocorrelations and see how quickly they die down as you increase the lag. For a fast mixing Markov chain they should die quite rapidly. Rerun the Markov chain with different starting points and overlay the time series of the same component. The replicates should criss-cross many times. If they do not, you should investigate whether there is multi-modality or something.

How many Chains? Advice in the literature is conflicting.

- One very long run. Geyer (1992), Raftery and Lewis (1992). Reaches part other schemes can not. If you are a proabilist, you would prefer this.

- Several long runs. Gelman and Rubin (1992). Gives indication of convergence. Argue that a single chain can be misleading.

Single Chain: Raftery and Lewis (1992) consider the problem of calculating the number of iterations necessary to estimate a posterior quantile from a single run of a Markov chain. They propose a 2-state Markov chain model fitting procedure based upon pilot analysis of output from the original chain.

Suppose that we wish to estimate a particular posterior quantile for some function b of a parameter (or set of parameters) $\mathbf{X}$, i.e., we wish to estimate u such that

$$P(b(\mathbf{X}) \leq u) = q$$

for some pre-specified q and so that, given our estimate $\hat{u}$, $P(b(\mathbf{X}) \leq \hat{u})$ lies within $\pm r$ of the true value, say, with probability p.

Raftery and Lewis propose a method to calculate n_0, the initial number of iterations to discard (which we call the 'burn-in'), k, the thinning size, i.e., every k^{th} iterate of the Markov chain and n

the number of further iterations required to estimate the above probability to within the required accuracy.

The details of the methodology can be found on the cited reference. Basically the methodology construct a discrete two state Markov chain $Z^t = I_{b^t = b(\mathbf{X}^t) \leq u}$ from $\mathbf{X}^t$. Then find the eigenvalues of this Markov chain to find the answers.

They also suggest looking at a quantity called the dependence factor:

$$I = \frac{n_0 + n}{n_{min}}$$

where n_{min} is the number of initial iterations performed to produce the estimates. Ideally, this factor should be close to 1.

Multi Chain: Gelman and Rubin (1992) propose a method which assesses convergence by monitoring the ratio of two variance estimates. In particular their method uses multiple replications and is based upon a comparison, for scalar functions of $\mathbf{X}$, of the within sample variance for each of m parallel chains, and the between sample variance of different chains. This is essentially a classical analysis of variance.

The method consists of analysing the independent sequences to form a distributional estimate for what is known about the target random variable, given the observations simulated so far. This distributional estimate, based upon the *Student's t* distribution, is somewhere between the starting and target distributions, and provides a basis for an estimate of how close the process is to convergence and, in particular, how much we might expect the estimate to improve with more simulations. The method proceeds as follows. We begin by independently simulating $m \geq 2$ sequences of length $2n$, each beginning at different starting points which are over dispersed with respect to the stationary distribution. We discard the first n iterations and retain only the last n. Then for each scalar functional of interest, b, we calculate B/n, the variance between the m sequence means, which we denote by $\bar{b}_{i.}$. Thus we define

$$\frac{B}{n} = \frac{1}{m-1} \sum_{i=1}^{m} (\bar{b}_{i.} - \bar{b}_{..})^2,$$

where

$$\bar{b}_{i.} = \frac{1}{n} \sum_{t=n+1}^{2n} b_i^t, \;\; \bar{b}_{..} = \frac{1}{m} \sum_{i=1}^{m} b_i^t.$$

and $b_i^t = b(\mathbf{x}_i^t)$ is the t^{th} observation of b from chain i. We then calculate W, the mean of the m within-sequence variances, s_i^2. Thus, W is given by

$$W = \frac{1}{m}\sum_{i=1}^{m} s_i^2,$$

where

$$s_i^2 = \frac{1}{n-1}\sum_{t=n+1}^{2n} (\bar{b}_{i.} - \bar{b}_{..})^2.$$

We can then estimate the target variance by a weighted average of W and B, given by

$$\hat{\sigma}^2 = \frac{n-1}{n}W + \frac{1}{n}B,$$

which overestimates the true value. We can improve upon the $N(\hat{\mu}, \hat{\sigma}^2)$ estimate of the stationary distribution by allowing for the variability of both $\hat{\mu}$ and $\hat{\sigma}^2$ and thus we obtain a t-distribution with with mean $\hat{\mu}$ and variance

$$\hat{V} = \hat{\sigma}^2 + \frac{B}{mn}$$

with a certain df to be estimated. Gelman and Rubin diagnostic is then calculates as

$$R = \frac{\hat{V}}{\hat{\sigma}^2}.$$

As $n \to \infty$ this should shrink to 1. For more details see the cited references.