MONTE CARLO METHODS, BOOTSTRAP PROCEDURES, AND BAYESIAN STATISTICS

Wooyong Park
2025-04-23

This text is based on Hogg et al. [2013] - *Introduction to Mathematical Statistics(8 ed.)* - chapters 4 and 11[1]. I leave these notes as a personal studying material for midterm, and I am happy to learn from my mistakes and typos. Even grammar issues are welcome to be notified of. Feel free to contact me via email.

The Method of Monte Carlo

Limitations of Asymptotic Techniques

By now, we are familiar with asymptotic techniques developed from **central limit theorem(CLT)**. However, two issues can be brought up with regard to the asymptotic techniques:

1. It says nothing about finite properties of an estimator.
    - In reality, the sample size is finite. However, the asymptotic distribution tells nothing about how a statistical method would perform within a finite number of samples.
2. Some limit distributions are impossible to obtain.[2]

The method of Monte Carlo is a method of *generating observations*(i.e., sampling) to simulate a complicated statistical process and investigate the finite sample properties of statistical methodologies.

Keywords: `generate`, `simulate`, `investigate (the finite properties)`

How fast is the CLT?

One motivating example of Monte Carlo is to check whether CLT holds true and how fast the convergence is through sampling process. We can validate such process through the following steps:

1. Set $X_i \overset{iid}{\sim} f(x;\mu,\sigma)$
    - Note that bell-shaped dist^{n}s converge faster.
2. Sample n X_i's to obtain $\bar{X}^{(n)}$.
3. Generate fixed N amount of $\bar{X}^{(n)}$'s to obtain $\bar{X}_1^{(n)}, \bar{X}_2^{(n)}, \dots, \bar{X}_N^{(n)}$
    - N is the fixed number **unrelated to our validation of convergence**.
4. Calculate $\frac{\sqrt{n}\left(\bar{X}_k^{(n)}-\mu\right)}{\sigma}$ for each $\bar{X}_k^{(n)}$ and obtain the histogram.
5. Compare it to the PDF of standard normal.

[1] The chapter numbers might differ between editions. Here, I am referring to the 8th global edition. The name of the chapters are **Some Elementary Statistical Inferences** and **Bayesian Statistics**.
[2] See Athey et al. [2021], for example. NOTE: THIS IS NOT COVERED IN THE COURSE

But how do we sample from $f(x; \mu, \sigma)$? The following theorem will shed us light on how to do this. The trick is, due to the introduction of high computation power, it became possible to sample from a uniform distribution, $U(0, 1)$. The following technique delinates how we can sample from any distribution with a known CDF based on the sampling from $U(0, 1)$.

Theorem 0.1 (Inverse CDF)**.** *Let $U \sim unif(0, 1)$. Suppose F is a* **continuous** *CDF. Then, $X = F^{-1}(U)$ follows the distribution F.*

Monte Carlo Integration

Suppose we want to compute $\int_a^b g(u)du$. If the integration is difficult, we can use Monte Carlo sampling to obtain a consistent estimate of the integration result:

$$\int_a^b g(u)du = (b-a)\int_a^b g(u)\frac{1}{b-a}du = (b-a)\mathbb{E}_{U(a,b)}[g(U)]$$

where U follows a uniform distribution from a to b. By sampling $g(U_1), g(U_2), \dots, g(U_n)$ independently from the uniform distribution, we can obtain $(b-a)g(\bar{U})$, which is a consistent estimate of $\int_a^b g(u)du$.

Review Questions

1. What were the two limitations of asymptotic techniques mentioned in this text?
2. Prove Theorem 0.1. For an easier version of the proof, you can assume that F is strictly increasing.

Bootstrap Procedures

As mentioned earlier, some estimators have asymptotic distributions whose closed form is impossible to obtain. In such cases, we can use **bootstrapping** to make an approximate $100(1-\alpha)$% confidence interval for an estimator, denoted by $\hat{\theta}$.

Percentile Bootstrap Confidence Intervals

Suppose we have the data $x_1, x_2, \dots, x_n$ from $f(x;\theta)$. We construct the approximate $100(1-\alpha)$% confidence interval for $\hat{\theta}$ as follows:

1. Resample from the dataset **with replacement** and equal probabilities.

$$x_i^* = \begin{cases} x_1 & (1/n) \\ x_2 & (1/n) \\ \vdots\, x_n & (1/n) \end{cases} \text{ for } i = 1, \dots, n$$

   - Note: The resampled version of the sample should have equal size to the original sample.
2. Obtain $\hat{\theta}^* = \hat{\theta}(x_1^*, x_n^*)$.
3. Repeat steps 1 and 2 B times(usually 100 to 300) to obtain $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$
   - Theoretically, it is known that the histogram of these estimates is approximately the same as the true distribution of $\hat{\theta}$
4. Let $m = \left[\frac{\alpha}{2}B\right]$. Then, $(\hat{\theta}_{(m)}, \hat{\theta}_{(B+1-m)})$ is the approximate confidence interval for $\hat{\theta}$.

Discussions on Distributions

By now, you should have learned three types of techniques for obtaining the distribution of a random variable:

1. Traditional Techniques in MathStat(1) : CDF technique, Transformation of Variables(if invertible), MGF technique, and **Student's Theorem**
2. Asymptotic Distributions
3. Bootstrapping Procedure

Note that these techniques are listed in the order of how accurate they are. Such accurate distributions are more valuable, but they tend to require more assumptions, and is often more difficult to obtain. However, one should try to find the asymptotic distributions before heading straight into obtaining Bootstrap CIs.

Review Questions

1. Delineate how to obtain percentile bootstrap CIs.
2. List all the techniques you have learned with regard to obtaining the distribution of a random variable.

Bayesian Statistics

In this section, we will repeatedly use the following notation for model parameter:

- θ : the model parameter
  - constant to frequentists; random to Bayesians

Frequentist Statistics

Terminology

- θ: unknown constant
- probability: the limit of the relative frequency of an event under a **random experiment**
- random experiment
  1. We cannot perfectly forecast the outcome
  2. We can describe all possible outcomes prior the experiment
  3. It can be repeated independently under the same condition.

Likelihood-based approach

So far, we have focused on the likelihood function, and it was all that a frequentist would need to investigate the characteristics of a population of interest.

$$\underbrace{L(\theta; \mathbf{x})}_{\text{all that a frequentist needs}} = \prod_{i=1}^{n} \underbrace{f}_{\text{set by assumption}}(x_i; \underbrace{\theta}_{\text{constant}})$$

Two attacks can be made with the likelihood-based approach:

1. The researcher's setting of f can be arbitrary.
2. The inference is **fully driven by the data**; hence, there is no way to incorporate the researchers' subjective belief, past experience, and so on.

Indeed, some PDFs have justifiable reasons. For example, Bernoulli trials for binary data has no better distribution than the binomial for the probability parameter to be estimated. At this point, you should remember the three assumptions of Binomial distribution.

1. The outcome can be classified into either success or failure, not both.
2. Each trial is independent.
3. The probability of success is constant.

In such setting, we can obtain the asymptotic confidence interval of $\left(\hat{p} - 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$

(Ask yourself why.)

Frequentists' Confidence Intervals

This is where one major problem of Frequentist statistics begins.

What does the following mean in Frequentists' view?

$$PR\left[p \in (\hat{p} - 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}})\right] = 0.95$$

It is not the p that is moving; it is $\hat{p}$ that floats around, and the probability should be read as **"how often does our constructed interval contain the constant parameter in terms of the limit of the relative frequency?"**

However, it is often interpreted as the probability that the parameter will be contained in our constructed confidence interval. This is one of the weakest points in Frequentists' statistics. Since the parameter is constant to a Frequentist, the probability that θ is contained in the confidence interval is either zero or one.

Introduction to Bayesian Statistics

Bayesian critiques of Frequentist statistical methods

So far, we reviewed the basic concepts of Frequentists' statistical methods. Now, we will address their limitations, which are stronly demonstrated by Bayesians.

1. The problem with confidence intervals
    - As mentioned earlier, Frequentists' confidence intervals cannot be interpreted as the probability that the parameter will be contained in a suggested interval, but is often misinterpreted in such fashion.
2. Reproducibility of random experiments
    - To a frequentist, probability is the limit of the relative frequency under a random experiment, and such random experiment must be **reproducible**. Consider stock market prices. Such variables cannot be estimated in the same condition, which undermines the cogency of Frequentists' statistical methods.
3. Fully driven by the data
    - In some cases, researchers should speak louder than the data. Suppose the researcher wants to add his background knowledge to a statistical inference. In such case, since the MLE approach is fully determined by the data, it does not allow any introduction of prior beliefs.

Terminology

For a Bayesian, following terms would have different definitions:

- θ: the realized value of a random variable, Θ
- probability: a reasonable expectation representing a state of knowledge(does not require random experiments)[3]

The name Bayesian statistics comes from Bayes Theorem indeed:

[3] NOTE: THIS WAS NOT COVERED IN THE COURSE

Theorem 0.2 (Bayes Theorem)**.**

$$P(B|A) = \frac{P(B)P(A|B)}{P(A)}$$

Suppose that event B is having lung cancer, and A is the event of having positive test results for a cancer. Indeed, $P(B|A)$ would be different from $P(B)$ unless $A \perp B$. This suggests the possibility of an **information update**, which was not discussed in detail in the Frequentist methods.

Likelihoods

Now **likelihood** is defined as the conditional probability of the data given a realized value of Θ:

$$X_1, X_2, \dots, X_n \overset{iid}{\sim} f(x|\theta)$$

$$\Theta \sim h(\theta)$$

Here, h is called the **prior** distribution of the model parameter. Since we are adding an assumption on how the model parameter is distributed, we can solve more practical issues related to the data.

It's all about the posterior

Posterior is the conditional probability of a parameter given the dataset:

$$g(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)h(\theta)}{\int f(\mathbf{x}, \theta)d\theta}$$

In Bayesian Statistics, the posterior distribution includes all the relevant information to our estimation of θ. One should note that the denominator of the posterior is a constant; say, C:

$$\int f(\mathbf{x}, \theta)d\theta = \int f(\mathbf{x}|\theta)h(\theta)d\theta = C$$

C can be easily computable or not. Sometimes, assumptions on the prior distribution, such as **conjugacy** (which will be discussed shortly), allow quick computation, but if not so, we rely on sampling methods.[4]

In our discussion, let us assume that C can be easily derived and write the relation between the **posterior**, **prior**, and the **likelihood** as follows:

$$g(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)h(\theta)$$

[4] See advanced texts such as Murphy [2012] or Murphy [2023].

Posterior and the Sufficient Statistic

As you might have learned earlier, **sufficient statistic**, $y = u(\mathbf{x})$, is a statistic such that for $X_1, \dots, X_n \overset{iid}{\sim} f(x|\theta)$,

$$P(\mathbf{x}|y) \equiv \frac{f_{\mathbf{x}}(\mathbf{x}|\theta)}{f_Y(y|\theta)} = g(\mathbf{x}, \cancel{\theta})$$

In other words,

$$f_{\mathbf{x}}(\mathbf{x}|\theta) = f_Y(y|\theta) \cdot g(\mathbf{x}, \cancel{\theta})$$

Note that $g(\mathbf{x})$ can be considered a constant in terms of the posterior, as **x** is what we observe and condition by.

Then, the proportionality relation can be rewritten using a sufficient statistic as below:

$$\begin{aligned} g(\theta|\mathbf{x}) &\propto f(\mathbf{x}|\theta)h(\theta) \\ &= f_Y(y|\theta)g(\mathbf{x})h(\theta) \\ &\propto f_Y(y|\theta)h(\theta) \\ &= g(\theta|y) \end{aligned}$$

Here, $g(\theta|y)$ is the posterior distribution wrt the sufficient statistic. That said, sufficient statistic is again all what we need to deduce the posterior distribution even in Bayesian statistics. ~~How amazing that there is something that both Frequentists and Bayesians can treasure.~~ Note that using the **factorization theorem** is very helpful to obtain a sufficient statistic.

Example 0.1 (Poisson and Gamma)**.** Find the posterior distribution $g(\theta|\mathbf{X})$

$$\begin{cases} X_i|\theta \overset{iid}{\sim} \text{Poisson}(\theta) \\ \Theta \sim \Gamma(\alpha, \beta) \\ (\alpha, \beta) \text{ are assumed to be known} \end{cases}$$

We will consider a simple example with a Poisson likelihood and a Gamma prior. This setup leads us naturally to **conjugate priors**, meaning that the prior and the posterior follow the same known distribution with different parameters.

$$L(\mathbf{x}|\theta) = e^{-n\theta}\theta^{\sum x_i} \cdot \prod_{i=1}^{n} \frac{1}{x_i!}$$

$$h(\theta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}}\theta^{\alpha-1}e^{-\frac{1}{\beta}\theta}$$

Thus, the posterior $g(\theta|\mathbf{x})$ is proportional to $\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\theta^{\sum x_i+\alpha-1}e^{-(n+1/\beta)\theta}$, which, in turn, is proportional to the PDF of $\Gamma(\sum x_i + \alpha, \frac{\beta}{\beta n+1})$.

Bayesian Point Estimation

Terminology

- **Decision Function**
  - Within $X_1, X_2, \dots, X_n \overset{iid}{\sim} f(x|\theta)$, let $Y = u(\mathbf{X})$ be a statistic we wish to base a point estimate of Θ. Then, $\delta(Y)$ is called a **decision function** on θ.
- **Loss**
  - Since in Bayesian estimation, our estimation is a prediction problem[5], we have a concept of loss. For a given pair of $\theta, \delta(y)$, **loss** is a non-negative number from $L(\theta, \delta(y))$ that reflects the failure of estimation with $\delta(y)$

$$L(\theta, \delta(y)) = \begin{cases} (\theta - \delta(y))^2 & (L^2\text{-norm}) \\ |\theta - \delta(y)| & (L^1\text{-norm}) \end{cases}$$

- **Risk Function**
  - **Risk is the average of loss over X**. In other words,

$$R(\theta, \delta(X)) = \int_{\mathbf{x}} L(\theta, \delta(\mathbf{x}))L(\mathbf{x}|\theta)d\mathbf{x}$$

Bayes Point Estimate

The **Bayesian point Estimate** on θ is the minimizer of $\mathbb{E}(L(\Theta, \delta)|\mathbf{X} = \mathbf{x})$. That is,

$$\delta(\mathbf{x}) = \arg\min_{\theta} \int L(\theta, \delta)g(\theta|\mathbf{x})d\theta$$

Note that if the loss function uses L1 norm, the median of the posterior should be the point estimate, and if the loss uses L2 norm, the mean of the posterior becomes the point estimate.

Bayes Point estimate minimizes the expected risk as well.

Bayes estimate satisfies

$$\delta(\mathbf{x}) = \arg\min_{\theta} \mathbb{E}_{g(\theta|\mathbf{x})}\Big[L(\theta, \delta)|\mathbf{x}\Big]$$

for any **x**. The expectation of $\mathbb{E}_{g(\theta|\mathbf{x})}\Big[L(\theta, \delta)\Big]$ can be rewritten with respect to the risk function as follows:

[5] Note that in Frequentist statistics, we used to rigorously distinguish estimation from prediction. The target of estimation was a constant, whereas the target of prediction was a random variable. Bayesians have a much more ambiguous position in such terminology. In the eyes of a Frequentist, however, Bayesian estimation should be called *prediction*.

$$\begin{aligned}\mathbb{E}_{p(\mathbf{x})}\left[\mathbb{E}_{g(\theta|\mathbf{x})}\left[L[\Theta,\delta(\mathbf{x})]|\mathbf{x}\right]\right] &= \int\int L[\Theta,\delta(\mathbf{x})]g(\theta|\mathbf{x})d\theta p(\mathbf{x})d\mathbf{x} \\ &= \int\int L[\Theta,\delta(\mathbf{x})]L(\mathbf{x}|\theta)d\mathbf{x}h(\theta)d\theta \\ &= \int R(\theta,\delta)h(\delta)d\delta\end{aligned}$$

Example 0.2 (Poisson and Gamma Revisited)**.** Note that the posterior from Example 0.1 was $\Gamma(\sum x_i + \alpha, \frac{\beta}{\beta n+1})$. That said, the Bayes point estimate for θ using L2 norm would be the product of shape and scale parameters, $(\sum x_i + \alpha)\cdot\frac{\beta}{\beta n+1}$

One must note that using L2 norm, the Bayes estimate is the **weighted mean of MLE and the prior mean**:

$$(\sum x_i + \alpha)\cdot\frac{\beta}{\beta n+1} = \frac{\beta n}{\beta n+1}\cdot\frac{\sum x_i}{n} + \frac{1}{\beta n+1}\cdot\alpha\beta$$

That said, one can say that $1/\beta$ is **the worth of the prior opinion** relative to the sample size of n.

Example 0.3 (Binomial and Beta Distribution)**.** Suppose that

$$\begin{cases} X_1, X_2, \dots, X_N \overset{iid}{\sim} B(n_i, p) \\ P \sim \mathrm{B}(\alpha,\beta)\end{cases}$$

Then,

$$\begin{aligned} g(p|\mathbf{x}) &\propto L(\mathbf{x}|p)h(p) \\ &= \prod_{i=1}^{N}\binom{n_i}{x_i}\cdot p^{\sum x_i+\alpha-1}(1-p)^{\sum n_i+\beta-\sum x_i-1} \\ &\propto \mathrm{B}(\sum x_i+\alpha, \sum n_i+\beta-\sum x_i) \text{ (Conjugate Prior)}\end{aligned}$$

Based on L2 loss function,

$$\begin{aligned}\delta(\mathbf{x})_{L2} &= \frac{\sum x_i+\alpha}{\alpha+\beta+\sum n_i} \\ &= \left(\frac{\sum n_i}{\alpha+\beta+\sum n_i}\right)\underbrace{\frac{\sum x_i}{\sum n_i}}_{\text{MLE}} + \left(\frac{\alpha+\beta}{\alpha+\beta+\sum n_i}\right)\underbrace{\frac{\alpha}{\alpha+\beta}}_{\text{Prior Mean}}\end{aligned}$$

One can say that $\alpha+\beta$ is the worth of the prior opinion relative to $\sum n_i$.

Remark 0.1 (How to choose the right prior). As opposed to Frequentist statistics, we can apply our prior knowledge to the inference process. Then, in Example 0.3, how can we construct the prior so that it matches our knowledge?

Set (α,β) in a way that:

1. $\frac{\alpha}{\alpha+\beta}$ is equal to the desired mean.
2. $\alpha+\beta$ matches our desired worth of the prior opinion.

~~indeed, doing so is attacked by a lot of Frequentists.~~

Bayesian Interval Estimation

Interval estimation is about determining the boundaries, rather than a single point estimate of θ. Just as the point estimates depended on our data **x**, the boundaries, denoted by u and v will be $u = u(\mathbf{x}), v = v(\mathbf{x})$.

To distinguish from Frequentists' confidence intervals, Bayesian intervals are called **credible intervals**. A credible interval of size γ is defined as follows:

$$P(u(\mathbf{x}) < \Theta < v(\mathbf{x})|\mathbf{X} = \mathbf{x}) = \int_{u(\mathbf{x})}^{v(\mathbf{x})} g(\theta|\mathbf{x})d\theta = \gamma$$

Note that credible intervals need not be unique, just as confidence intervals are not always unique as well. One way of choosing the best credible interval is to select the **highest density region(HDR)** or the **highest density interval(HDI)**. It is defined as follows:

If $\mathcal{R}$ is the **HDR** s.t.

$$\int_{\mathcal{R}} g(\theta|\mathbf{x})d\theta = \gamma$$

there is $k > 0$ such that $\mathcal{R} = \{\theta : g(\theta|\mathbf{x}) > k\}$. i.e., $\mathcal{R}$ is the shortest possible interval that covers the size of γ.

Bayesian Testing

Bayesian testing solely relies on the posterior: which hypothesis has higher posterior probability.

A test, given a dataset $X_1, X_2, \dots, X_n \overset{iid}{\sim} f(x|\theta)$ for $\theta \in \Omega$, is choosing rather to take the null hypothesis(H_0) of the alternative hypothesis(H_1) such that

$$\begin{cases} H_0 : \theta \in \omega_0 \\ H_1 : \theta \in \omega_1 \end{cases}$$

where $\omega_0 \cap \omega_1 = \emptyset$ and $\omega_0 \cup \omega_1 = \Omega$.

THe Bayesian testing procedure can be simply put as:

Accept H_0 if $P(\Theta \in w_0|\mathbf{x}) \geq P(\Theta \in w_1|\mathbf{x})$. Otherwise, accept H_1.

Noninformative and Improper Priors

Shrinkage Estimate

Suppose $X_1, X_2, \dots, X_n \overset{iid}{\sim} b(1, p)$ and p is the realized outcome from $P \sim \mathrm{B}(\alpha, \beta)$. We want to consider a situation **where we do not have any prior information on** P. In other words, we want to treat all possible P's equally, so we set $\alpha = \beta = 1$ to obtain a uniform prior.

However, the Bayes estimate with a squared loss function would be

$$\left(\frac{n}{n+2}\right)\frac{y}{n} + \left(\frac{2}{n+2}\right)\frac{1}{2}$$

which means that the prior mean $1/2$ still has the worth of 2 relative to the sample size. Such estimate is called a **shrinkage estimate** in that the MLE is pulled a little toward the prior mean, although we tried to avoid the influence of the prior distribution.

Noninformative Priors

A **noninformative prior** is a prior that does not inject prior opinion into our Bayesian estimate. More strictly speaking, it is a prior that minimizes the influence of the prior distribution. A uniform distribution often satisfies the role of noninformative prior; however, if the parameter is a scale parameter, $1/\theta \cdot \mathbb{1}(\theta > 0)$ is used instead.

One commonly used noninformative prior is **Jeffrey's prior**:[6]

$$h(\theta) \propto \sqrt{I(\theta)}$$

where $I(\theta)$ is the **Fisher information**.

So in the case of Bernoulli random variables,

$$\begin{aligned} g(p|\mathbf{x}) &\propto L(\mathbf{x}|p) \cdot \sqrt{I(p)} \\ &= \binom{n}{\sum x_i} p^{\sum x_i}(1-p)^{n-\sum x_i} \cdot \Big(p(1-p)\Big)^{-1/2} \\ &= \binom{n}{\sum x_i} p^{\sum x_i - \frac{1}{2}}(1-p)^{n-\sum x_i - \frac{1}{2}} \end{aligned}$$

whose mean would be equal to $\frac{\sum x_i + 0.5}{n+1}$. However, note that the estimate is still a shrinkage estimate.

Improper prior

An **improper prior** is a prior distribution whose integration does not sum to a constant. In other words, a proper prior would integrated to a positive constant. Often, a noninformative prior would be an improper one as it has to treat all values across $\mathbb{R}$ equally. However, this does not always have to be true. Also, an improper prior is not necessarily a noninformative prior as well.

[6] NOTE: JEFFREY'S PRIOR IS NOT DISCUSSED IN THE CLASS

Gibbs Sampling

Let $k(\theta|\mathbf{x}) = L(\mathbf{x}|\theta)h(\theta)$. Then, the posterior distribution is

$$g(\theta|\mathbf{x}) = \frac{k(\theta|\mathbf{x})}{\int k(\theta|\mathbf{x})d\theta}$$

In a lot of cases, it is difficult to compute the denominator, which is why Bayesians often resort to Monte Carlo techniques. One solution is the **rejection sampling**:[7]

Rejection Sampling

1. Sample θ from a proposal distribution $q(\theta)$ such that $Mq(\theta) \geq k(\theta|\mathbf{x})$ for some $M > 0$.
2. Sample u from Unif$(0, Mq(\theta))$.
3. If $u > k(\theta|\mathbf{x})$, reject that θ; otherwise, accept it.
4. Repeat the above steps until you have enough samples.
5. The samples that are accepted follow the distribution $g(\theta|\mathbf{x})$.

Theorem 11.4.1. (Another technqiue of Sampling)

Suppose now, that it is difficult to sample from a distribution $f_X(x)$. However, it is easy to sample from the following distributions

1. $Y \sim f_Y(y)$
2. $X \sim f_{X|y}(x)$

Then, if we generate Y from the first distribution and sequentiall generate X from the second distribution, the obtained X follows the distribution $f_X(x)$.

Exploitation of Theorem 11.4.1.

Suppose we want to compute the mean of $\mathbb{E}[W(x)]$ where $\mathbb{E}[W^2(x)] < \infty$.

Using the algorithm, we can generate $(Y_1, X_1), (Y_2, X_2), \dots (Y_m, X_m)$ where Y's are drawn independently from $f_Y(y)$ and X is drawn independently from $f_{X|y}(x)$. Then, we obtain $X_1, X_2, \dots, X_m$ that follows $f_X(x)$, so that we can obtain our estimator:

$$\bar{W} = \frac{1}{m}\sum_{i=1}^{m} W(x_i)$$

which, by WLLN,

$$\bar{W} \xrightarrow{p} \mathbb{E}[W(x)]$$

[7] NOTE: REJECTION SAMPLING IS NOT A PART OF THE EXAM.

Gibbs Sampling

Now, we discuss the **Gibbs sampler**, which is one of the groundbreaking sampling methods widely used in Bayesian statistics, motivated by Theorem 11.4.1.

Let m be a positive integer, and X_0 be a given initial value of X. The algorithm is as follows:

For $i = 1, 2, \dots, m$,

1. Generate $Y_i|X_{i-1} \sim f(y|x)$
2. Generate $X_i|Y_i \sim f(x|y)$

Then, as $i \to \infty$,

$$Y_i \xrightarrow{D} Y \sim f_Y(y)$$
$$X_i \xrightarrow{D} X \sim f_X(x)$$

and as $m \to \infty$,

$$\frac{1}{m}\sum_{i=1}^{m} W(X_i) \xrightarrow{P} \mathbb{E}[W(X)]$$

Markov Chain Monte Carlo

It must be noted that Gibbs sampling is one type of **Markov Chain Monte Carlo(MCMC)** methods. Markov chain means that a sampled variable's distribution is affected only from the most recent state of that variable. That is, a random variable's future is not affected by its past after conditioning on its present state. Since in Gibbs sampling the random variables are independent of its past state given the current state, it can be said that the method is a Monte Carlo method based on Markov Chain sampling process. In the next section, we will see how **Gibbs sampling** augments Bayesian inference.

Heirarchical Bayes Model

In Bayesian estimation, the prior distribution has an important influence over our inference. Until now, however, we had a fixed prior distribution that was too arbitrary without sufficient justification. **Heirarchical Bayes model** introduces a **hyperprior** $\phi(\gamma)$ to resolve such issues:

$$\begin{cases} X|\theta \sim f(x|\theta) \\ \Theta|\gamma \sim h(\theta|\gamma) \\ \Gamma \sim \phi(\gamma) \end{cases}$$

In this model the hyperparameter γ can be regarded as a **nuisance parameter**, whereas θ is the parameter of interest.

The posterior distribution can be written as:

$$
\begin{aligned}
g(\theta|\mathbf{x}) &= \int g(\theta, \gamma|\mathbf{x}) d\gamma \\
&= \frac{\int f(\mathbf{x}|\theta) h(\theta|\gamma) \phi(\gamma) d\gamma}{\int\int f(\mathbf{x}|\theta) h(\theta|\gamma) \phi(\gamma) d\gamma d\theta}
\end{aligned}
$$

If we use a squared-error loss function, the Bayes estimate of $W(\theta)$ would be the mean of the posterior:

$$
\begin{aligned}
\delta_W(\mathbf{x}) &= \int\int W(\theta) g(\theta|\mathbf{x}) d\theta \\
&= \frac{\int W(\theta) f(\mathbf{x}|\theta) h(\theta|\gamma) \phi(\gamma) d\gamma d\theta}{\int\int f(\mathbf{x}|\theta) h(\theta|\gamma) \phi(\gamma) d\gamma d\theta}
\end{aligned}
$$

~~I hated reading this integration from the textbook; now I hate more to write it down.~~

This is where the Gibbs sampler can give us a lot of help. Consider:

$$
\begin{cases}
\Theta_i|\mathbf{x}, \gamma_{i-1} \sim g(\theta|\mathbf{x}, \gamma_{i-1}) \\
\Gamma_i|\mathbf{x}, \theta_i \sim g(\gamma|\mathbf{x}, \theta_i)
\end{cases}
$$

By repeatedly sampling $W(\theta_1), W(\theta_2), ...$, one can show that with large enough n, m with $n \gg m$,

$$
\frac{1}{n-m} \sum_{i=m}^{n} W(\Theta_i) \xrightarrow{p} \mathbb{E}[W(\Theta)|\mathbf{x}] = \delta_W(\mathbf{x})
$$

as $n - m \to \infty$ by WLLN.

Empirical Bayes

The **empirical Bayes(EB)** model is another modern Bayesian method to relax Bayesians' assumptions on the prior distribution.

The original Bayes model has the following setup:

$$
\begin{cases}
X|\theta \sim f(x|\theta) \\
\Theta \sim h_\gamma(\theta)
\end{cases}
$$

In such formation, we have implicitly made two assumptions:

1. Θ has the distribution form characterized by function h.
2. The function h can be summarized by a fixed parameter, γ, which is usually determined by the researcher.

Just as in **heirarchical Bayes model**, EB also assumes that γ is a realized value of a random variable Γ. However, it does not assume a hyperprior over that rv. Thus, EB can be presented as:

$$\begin{cases} X|\theta \sim f(x|\theta) \\ \Theta \sim h(\theta|\gamma) \end{cases}$$

EB can either relax the assumption on γ or relax the assumption on the entire form of h itself. In this document, I only delineate how the former works, but you can find information on the latter by searching **Robbin's Method**.

Likelihood-based approach

To begin with, we will discuss how EB relaxes assumptions on γ. It Consider the likelihood function

$$m(\mathbf{x}|\gamma) = \int_{-\infty}^{\infty} f(\mathbf{x}|\theta)h(\theta|\gamma)d\theta$$

In EB, we obtain the MLE of γ from the likelihood above and plug the estimate into our model:

$$\hat{\gamma} = \arg\max_{\gamma} m(\mathbf{x}|\gamma)$$

$$\begin{cases} X|\theta \sim f(x|\theta) \\ \Theta \sim h(\theta|\hat{\gamma}) \end{cases}$$

Example 0.4 (A Medical Example of Surgery)**.** Consider a cancer surgery where the surgeon tries to remove the cancer cells. It is not easy to remove the malignant nodes only, and some healty nodes are removed as well in practice. Let X denote the nodes found malignant. Then for each surgery $k = 1, 2, \dots, n$, the frequency of malignant cells can be denoted by $p_k = X_k/n_k$ where n_k is the total number of nodes removed.

If the probability of each node being malignant varies by individuals but is constant and independent within an individual, we can model that probability as the probability of a binomial rv X_k:

$$X_k \sim B(n_k, \theta_k)$$

By the approximation of binomial to a normal,

$$p_k \overset{a}{\sim} \mathcal{N}(\theta_k, \frac{\theta_k(1-\theta_k)}{n_k})$$

Indeed, frequentists would arrive at the conclusion that with larger n_k, one can estimate p_k with more precision, but cannot give any further information.

Bayesians would model θ_k with the prior density $g(\theta)$:

$$\theta_k \sim g(\theta) (\text{for } k = 1, 2, \dots, N)$$

In EB, we can model this prior by a fourth-order polynomial:

$$\log(g(\theta)) = \alpha_0 + \sum_{j=1}^{4} \alpha_j \theta^j$$

The hyperparameters would be $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$, because once α is given α_0 is automatically determined by the condition $\int_0^1 \exp[\alpha_0 + \sum_{j=1}^4 \alpha_j \theta^j] d\theta = 1$

Let $f_\alpha(x_k)$ be the individual likelihood given α:

$$f_\alpha(x_k) = \int_0^1 \binom{n_k}{x_k} \theta_k^{n_k} (1-\theta_k)^{n_k - x_k} g_\alpha(\theta) d\theta$$

Then, EB is completed by finding $\hat{\alpha}$ s.t.

$$\hat{\alpha} = \arg\max_\alpha \prod_{k=1}^{N} f_\alpha(x_k)$$

which would be our empirical Bayes estimate.

Review Questions

1. Discuss how the definitions of `probability` differ between Frequentist statistics and Bayesian statistics.
2. List the three key problems with Frequentist methods metioned above.
3. Suppose

$$X_1, X_2 \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma_0)$$
$$\mathrm{M} \sim N(\mu_0, \sigma_0)$$

   Find the Bayes point estimate of μ using an L2 loss function.
4. Illustrate the algorithm of Gibbs sampling.

References

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730, 2021.

Robert V Hogg, Joseph W McKean, and Allen T Craig. *Introduction to mathematical statistics*. Pearson Education India, 2013.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Kevin P Murphy. *Probabilistic machine learning: Advanced topics*. MIT press, 2023.