Bayesian Statistics

* We have seen that Monte Carlo and simulation methods can be used to evaluate the performance of statistical methods.

* Monte Carlo can also be used for inference from a single sample when integration is required.

* They can also be used for optimization problems but we shall not consider that use of Monte Carlo in this course.

* Integration most commonly appears in Bayesian inference problems.

Basics of Bayesian Inference

* In Bayesian statistics, all unknown quantities including the parameters of the distribution are considered to be random variables.

* The distribution of the data $X_1, \ldots, X_n$ is then a conditional distribution given a value of the random variable θ.

* We also need a distribution for the parameter θ.

* This represents our knowledge of the parameter before we see any data and is called the prior distribution.

Bayesian Statistics Process

1. Specify a conditional distribution of the data given the parameters. This is identical to the usual model specification in frequentist statistics.

2. Specify the prior distribution of the model parameter $\pi(\theta)$.

3. Collect the data, $x_1, \ldots, x_n$.

4. Update the prior distribution based on the data observed to give a Posterior Distribution of the parameter given the observed data $\boldsymbol{x}$, $\pi(\theta \mid \boldsymbol{x})$.

5. All inference is then based on this posterior distribution.

Finding the Posterior Distribution

* Prior distribution $\pi(\theta)$

* Data $x_1, \ldots, x_n$ and likelihood $L(\theta \mid x_1, \ldots, x_n)$.

* Posterior distribution

$$\begin{aligned} \pi(\theta \mid x_1, \ldots, x_n) &= \frac{\pi(\theta) L(\theta \mid x_1, \ldots, x_n)}{\int \pi(\theta) L(\theta \mid x_1, \ldots, x_n)\, d\theta} \\ &\propto \pi(\theta) L(\theta \mid x_1, \ldots, x_n) \end{aligned}$$

Bayesian Inference

* Typically we use the posterior mean and standard deviation of the parameters for inference.

$$\begin{aligned} \mathsf{E}(\theta \mid x_1, \ldots, x_n) &= \int_{-\infty}^{\infty} \theta \pi(\theta \mid x_1, \ldots, x_n)\, d\theta \\ \mathsf{Var}(\theta \mid x_1, \ldots, x_n) &= \int_{-\infty}^{\infty} \theta^2 \pi(\theta \mid x_1, \ldots, x_n)\, d\theta - \Big(\mathsf{E}(\theta \mid x_1, \ldots, x_n)\Big)^2 \end{aligned}$$

* Since these quantities involve integration we can use Monte Carlo.

Monte Carlo Bayesian Inference

* Find the posterior distribution

$$\pi(\theta \mid x_1, \ldots, x_n) \;\propto\; \pi(\theta) L(\theta \mid x_1, \ldots, x_n)$$

* Use one of the simulation methods described earlier to generate

$$\theta_1, \ldots, \theta_N \;\sim\; \pi(\theta \mid x_1, \ldots, x_n)$$

* Note that, in many instances, we only have the posterior distribution up to a constant which may depend on the data but not on the parameters. so we need to use methods (such as MCMC or accept-reject) that do not require this constant.

Monte Carlo Bayesian Inference

* Given a sample from the posterior distribution we can use simple Monte Carlo estimation of properties of the posterior distribution.

$$\widehat{\mathsf{E}}\big(\theta \mid x_1, \ldots, x_n\big) = \frac{1}{N}\sum_{i=1}^{N} \theta_i$$

$$\widehat{\mathsf{Var}}\big(\theta \mid x_1, \ldots, x_n\big) = \frac{1}{N}\sum_{i=1}^{N} \theta_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} \theta_i\right)^2$$

Bayesian Credible Intervals

Definition 8

A $100(1-\alpha)\%$ *Bayesian Credible Interval is an interval* $\left[\theta_l,\ \theta_u\right]$ *such that*

$$\mathsf{P}\Big(\theta_l < \theta < \theta_u \mid x_1, \ldots, x_n\Big) \ = \ 1-\alpha$$

* There are a number of ways to find these intervals.

* One method is to select θ_l and θ_u such that

$$\mathsf{P}\Big(\theta < \theta_l \mid x_1, \ldots, x_n\Big) \ = \ \frac{\alpha}{2} \qquad \mathsf{P}\Big(\theta > \theta_u \mid x_1, \ldots, x_n\Big) \ = \ \frac{\alpha}{2}$$

* These are called equi-tailed credible intervals.

* Note that the endpoints will depend on the data $x_1, \ldots, x_n$ since they are based on the posterior distribution.

Monte Carlo Credible Intervals

* Equi-tailed intervals require evaluation of the quantiles of the posterior distribution.

* Given a large sample from the posterior distribution we can estimate these.

* Suppose that we order the sample

$$\theta_{(1)} < \theta_{(2)} < \cdots < \theta_{(N)}$$

and let $N_\alpha = \lfloor N\alpha/2 \rfloor$.

* Then we can take

$$\widehat{\theta}_l = \theta_{(N_\alpha)} \qquad \widehat{\theta}_u = \theta_{(N-N_\alpha)}$$

* We require a large N for this to be a good approximation since the variability will be much too high if $N\alpha/2$ is small.

Multiparameter Bayesian Inference

* In many settings we have more than one unknown parameter.

* Bayesian inference works in the same way with multiple parameters except that the prior and posterior distributions are for multivariate random vectors.

* This can really complicate the calculations in Bayesian inference.

* MCMC methods, however, are very useful in this setting.

* We can then simulate from the multivariate posterior distribution and use Monte Carlo methods to approximate whatever functions of the parameter vector we are interested in.

Example

Consider Bayesian inference for the Normal(μ, σ^2) based on a sample $X_1, \ldots, X_n$

$$\begin{aligned} X_i \mid \mu, \sigma^2 &\overset{iid}{\sim} \text{Normal}(\mu, \sigma^2) \\ \mu &\sim \text{Normal}(m, v) \\ \frac{1}{\sigma^2} &\sim \text{Gamma}(a, b) \end{aligned}$$

where m, v, a, b are all specified constants and μ and σ^2 are independent a priori.

We wish to use Monte Carlo methods to examine the posterior distribution of $\theta = (\mu, \sigma^2)$ for the dataset

17.03	18.45	18.59	18.58	18.23	21.78	13.71	17.66
21.93	24.40	14.81	26.19	16.36	22.97	26.67	24.68

Hierarchical Bayesian Inference

* In the standard Bayesian model we have the likelihood (interpreted as the conditional distribution of the data given the parameters) and a prior for the parameter vector, $\boldsymbol{\theta}$.

* The prior often depends on a vector of *hyperparameters* $\boldsymbol{\gamma}$ and these are usually assumed known.

* An extension, however, could consider the $\boldsymbol{\gamma}$ as random variables also and put a prior $g(\gamma_k)$ on the components of $\boldsymbol{\gamma}$

* Often these priors on the hyperparameters are defined to be *non-informative* priors.

Non-informative Priors

* An attempt to define a prior distribution that will have no or minimal effect on the posterior inference.

* For location parameters we will often use $\pi(\mu) \propto 1$ for $\mu \in \mathbb{R}$.

* For scale parameters we often use the same prior but on the log scale so we get $\pi(v) \propto \frac{1}{v}$ for $v > 0$.

* Note that neither of these are proper densities since

$$\int_{-\infty}^{\infty} \pi(\gamma)\, d\gamma \;=\; \infty$$

* Nonetheless using these *improper priors* usually still results in proper posterior distributions.

Hierarchical Bayesian Inference

* Consider the following hierarchical structure

$$\begin{aligned} X_i &\overset{iid}{\sim} f(x_i \mid \boldsymbol{\theta}) \qquad i = 1, \ldots, n; \boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \\ \theta_j &\sim \pi_j(\theta_j \mid \boldsymbol{\gamma}), \qquad j = 1, \ldots, p; \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d) \\ \gamma_k &\sim g_k(\gamma_k) \qquad k = 1, \ldots, d \end{aligned}$$

  and at each stage we assume conditional independence of the components.

* We can then write the joint posterior as

$$\pi(\boldsymbol{\theta}, \boldsymbol{\gamma} \mid \boldsymbol{X}) \propto \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta}) \prod_{j=1}^{p} \pi_j(\theta_j \mid \boldsymbol{\gamma}) \prod_{k=1}^{d} g(\gamma_k)$$

Hierarchical Bayesian Inference

* Hence we can write down the full conditional posteriors

$$\pi(\theta_j \mid \boldsymbol{\theta}_{-j}, \boldsymbol{\gamma}, \boldsymbol{x}) \propto \pi_j(\theta_j \mid \boldsymbol{\gamma}) \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta}) \qquad j = 1, \ldots, p$$

$$\pi(\gamma_k \mid \boldsymbol{\theta}, \boldsymbol{\gamma}_{-k}, \boldsymbol{x}) \propto g(\gamma_k) \prod_{j=1}^{p} \pi_j(\theta_j \mid \boldsymbol{\gamma})$$

* We can then use the Gibbs Sampler to simulate from these full conditional distributions.

* In some cases the full conditionals are not easy to simulate from. In such situations it is common to use a **single step** of a Metropolis-Hastings algorithm to generate from the full conditional. This is often called Metropolis-Within-Gibbs Sampling.