Introduction to Bayesian Analysis

Prof. Nagi Gebraeel
Industrial and Systems Engineering
Georgia Tech

1

Degradation-Based Prognostic Modeling

Introduction to Bayesian Statistics
- Bayesian vs. Frequentist perspectives
- Influence of Baye's Rule
- Prior Distribution and Likelihood Functions Posterior Distribution
- Conjugate Priors
- Detailed Example

2

Frequentist versus Bayesian Statistics

- In the classical statistical approach, the parameter θ is thought to be an unknown, but fixed quantity.
- A random sample $X = (X_1, \dots, X_n)$ is drawn from a population indexed by θ and based on the observed values in the sample $x = (x_1, \dots, x_n)$, where the knowledge about the value of θ is obtained.
- In contrast, the Bayesian statistical approach considers θ to be a quantity whose variation can be described by a probability distribution that is updated using new observations.

Frequentist versus Bayesian Statistics

- In the Bayesian statistical approach, θ is consider to be a quantity whose variation can be described by a probability distribution, which is called the ***prior distribution***.
  - This is a subjective distribution based on the experimenter's belief, and perhaps some empirical evidence. It is formulated before any experimental data is obtained.
  - A sample is then drawn from a population indexed by θ and the prior distribution is updated with the sample information. The updated distribution is called ***posterior distribution***.
  - The updating framework is done according to Bayes' Rule.

Overview of Bayesian Statistics

- If we denote the prior distribution of θ by $\pi(\theta)$ and the sampling distribution given θ by $f(\mathbf{x}|\theta)$, then the posterior distribution, i.e., the conditional distribution of θ given the sample x, is

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int f(\mathbf{x}|\theta)\pi(\theta)d\theta}$$

- In choosing a prior belonging to a specific distributional family, $\pi(\theta)$, some choices may be computationally more convenient than others.
- In particular, it might be possible to select a member of that family which is a ***conjugate*** to the likelihood function $f(\mathbf{x}|\theta)$, that is, one that leads to a posterior distribution $\pi(\theta|\mathbf{x})$ belonging to the same distribution family as the prior.

Explanatory Example

- The following table displays historical data for launches of new rockets conducted by “new” companies during the period 1980–2002.
- A total of 11 launches were performed; 3 were successes and 8 were failures.
- Our goal in presenting this data is to specify a statistical model that can be used for predicting the future success of new rocket systems.
- Because a launch outcome can be regarded as either a success or failure, we can model launch outcome as Bernoulli data

Vehicle	Outcome
Pegasus	Success
Percheron	Failure
AMROC	Failure
Conestoga	Failure
Ariane 1	Success
India SLV-3	Failure
India ASLV	Failure
India PSLV	Failure
Shavit	Success
Taepodong	Failure
Brazil VLS	Failure

Explanatory Example

- If we let π denote the probability that a new launch vehicle selected at random succeeds, then we can express the probability of observing the sequence of successes and failures reported in the previous table as follows:

$$\pi^3(1-\pi)^8$$

- The above expression can be generalized to the situation in which we observe y successes in n trials leading to the binomial probability density function, which we can write as:

$$f(y|n,\pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y}$$

  - $f(y|n,\pi)$ specifies the probability of observing an outcome of a future experiment conducted on a sample of items drawn from the population of interest and is referred to as *Sampling Distribution*.

Explanatory Example

- Using the classic statistical approach, a point estimate of the failure probability of a new launch system developed by an inexperienced manufacturer is provided by the MLE:

$$\widehat{\pi} = \frac{y}{n} = \frac{3}{11} = 0.272$$

- The standard error for this estimate is

$$se(\widehat{\pi}) = \sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n}} = \sqrt{\frac{0.272(1-0.272)}{11}} = 0.134.$$

- It follows that an approximate $(1-\alpha)\times 100\%$ confidence interval for π is given by (for $\alpha = 0.1$),

$$(\widehat{\pi} - z_{\alpha/2}\, se(\widehat{\pi}), \widehat{\pi} + z_{\alpha/2}\, se(\widehat{\pi})) = (0.052, 0.492)$$

Explanatory Example

- Alternatively, we can use experience from vehicles launched prior to 1980 to specify ***informative prior*** distribution for success probabilities of post-1980 launch vehicles
  - We can specify prior information regarding the value of this parameter by using a probability density function on the unit interval.
  - This probability density is called the prior density, since it reflects information about π prior to observing experimental data
- In practice, the distribution used to reflect prior information may be dispersed, reflecting the fact that little is known about the parameter, or it may be concentrated in a particular region of the parameter space, reflecting the fact that more specific information is available.
  - In the former case, the prior distribution is sometimes called diffuse, **noninformative**, or vague;
  - In the latter, it is called **informative**.

Explanatory Example

- For the noninformative case, we assume that all values of π between 0 and 1 are equally plausible, i.e., this can be summarized by assuming that the prior distribution for π is uniform on the unit interval.
  - $Unif(0,1)$ or $Beta(1,1,)$
- For the informative case we will assume that the prior distribution follows a Beta distribution with parameters $\alpha = 2.4$ and $\beta = 2$.
  - $Beta(2.4, 2)$

Informative vs. Non-informative Priors

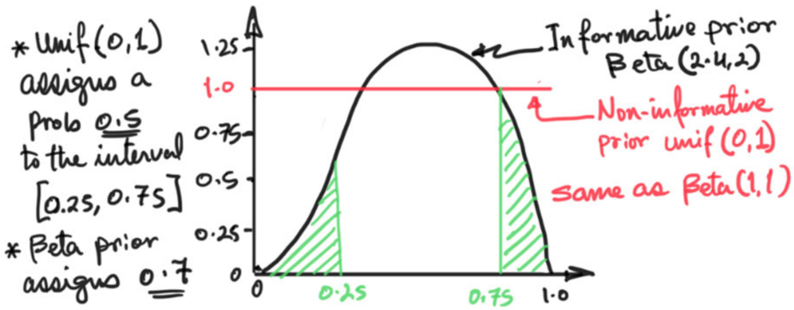
- Alternatively, we can use experience from vehicles launched prior to 1980 to specify ***informative prior*** distribution for success probabilities of post-1980 launch vehicles
- Once data are obtained, the prior distribution is updated using the new information.
- In this example, we will assume that the prior distribution follows a Beta distribution with parameters $\alpha = 2.4 \text{ and } \beta = 2$

Informative vs. Non-informative Priors

- Prior distributions can be:
  - Dispersed, reflecting the fact that little is known about the parameter, aka., ***Non-informative***
  - Concentrated in a particular region of the parameter space, reflecting the fact that more specific information is available, aka., ***informative.***
- **Example:** Suppose that little information is known π.
  - A priori, we might suppose that all values of π between 0 and 1 are equally plausible, i.e., this can be summarized by assuming that the prior distribution for π is uniform on the unit interval.
    - This prior distribution is an example of a diffuse prior since it reflects a lack of precise prior information about the true value of π.

Informative vs. Non-informative Priors

- **<u>Example</u>: Calculating posterior distributions for the launch vehicle failure data for two prior distributions**
  - *Beta*(1, 1) → equivalent to the non-informative uniform prior distribution
  - *Beta*(2.4, 2) → equivalent to the informative prior distribution



Combining Data with Prior Information

- Observations
  - The prior distribution is a beta distribution,

$$P(\pi|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma\alpha\,\Gamma\beta}\,\pi^{\alpha-1}(1-\pi)^{\beta-1}$$

  - The corresponding likelihood function is,

$$f(y|\pi,n) = \binom{n}{y}\pi^{y}(1-\pi)^{n-y}$$

  - Therefore, the posterior distribution is,

$$P(\pi|y) \propto f(y|\pi,n)\cdot P(\pi|\alpha,\beta)$$

Informative vs. Non-informative Priors

- Case 1:

- Case 2:

- It follows that an approximate $(1-\alpha) \times 100\%$ confidence interval for $\alpha = 0.1$ is given by $(0.13, 0.58)$.

Combining Data with Prior Information

- This resulting model is called the ***beta-binomial model***.
- Prior distributions that take the same functional form as the posterior distribution are called ***conjugate prior distributions*****.**
  - Conjugate prior distributions can make posterior analysis easy.
  - Prior distributions should not be specified simply for computational convenience.
  - If a conjugate prior that adequately represents the data prior to the experimentation cannot be found, then non-conjugate priors should be used
  - We explore numerical techniques handling non-conjugate and conjugate with do not admit simple analytical forms later.

Conjugate Pairs

Sampling Distribution (Parameter)	Conjugate Prior
Binomial (π)	Beta
Exponential (λ)	Gamma
Gamma (λ)	Gamma
Multinomial ($\boldsymbol{\pi}$)	Dirichlet
Multivariate Normal ($\boldsymbol{\mu}, \Sigma$)	Normal Inverse Wishart
Negative Binomial (π)	Beta
Normal (μ, σ^2 known)	Normal
Normal (σ^2, μ known)	Inverse Gamma
Normal (μ, σ^2)	Normal Inverse Gamma
Pareto (β)	Gamma
Poisson (λ)	Gamma
Uniform$(0, \beta)$	Pareto

Combining Data with Prior Information

- Posterior distributions represent all available information about π after both prior information and experimental data are combined.
- All inferences about the success probability π are based on these posterior distributions
  - Posterior probability intervals are the Bayesian analogues of classical confidence intervals and can be summarized using the $(1-\alpha) \times 100\%$ interval.
  - The posterior mean is given as

$$E(\pi|y) = \int_0^1 \pi \, p(\pi|y) d\pi$$

Combining Data with Prior Information

- To better understand the combination of prior information and data, consider the following explanation:
  - The mean of the Beta distribution is $\frac{\alpha}{\alpha+\beta}$
  - Based on y successes and $n-y$ failures, the posterior mean is:

Combining Data with Prior Information

- For the Binomial example with prior distribution $Beta(\alpha, \beta)$, the posterior is $Beta(y+\alpha, n-y+\beta)$
  1. When y and $n-y$ are large, the difference between $Beta(y+\alpha, n-y+\beta)$ and $Beta(y, n-y)$ becomes smaller.
  2. Thus, the influence of the prior distribution diminishes.
  3. For large values of y and $n-y$, a $Beta(y, n-y)$ looks very much like a normal distribution.

More on Bayesian Statistics

- **Example 2:** Let $X \sim N(\theta, \sigma^2)$, and suppose the prior distribution of is $\theta \sim N(\mu, \tau^2)$. Then, the posterior distribution of θ is also normal, with mean and variance given by:

$$E(\theta|x) = \left(\frac{\tau^2}{\tau^2+\sigma^2}\right)x + \left(\frac{\sigma^2}{\tau^2+\sigma^2}\right)\mu, \qquad Var(\theta|x) = \frac{\sigma^2\tau^2}{\tau^2+\sigma^2}$$

- The Bayes estimator of θ is the posterior mean, $E(\theta|x)$.
- Notice that the Bayes estimator is a linear combination of the prior and sample means.
- As τ^2 tends to infinity, the Bayes estimator tends toward the sample mean.

Example 2

- **Example 1:** Let $X \sim N(\theta, \sigma^2)$, and suppose the prior distribution of is $\theta \sim N(\mu, \tau^2)$. Then, the posterior distribution of θ is also normal, with mean and variance given by:

Example 2

Example 2

Example 2

More on Bayesian Statistics

$$E(\theta|x) = \left(\frac{\tau^2}{\tau^2+\sigma^2}\right)x + \left(\frac{\sigma^2}{\tau^2+\sigma^2}\right)\mu, \qquad Var(\theta|x) = \frac{\sigma^2\tau^2}{\tau^2+\sigma^2}$$

- Some observations:
  - As the prior information becomes more vague, the Bayes estimator tends to give more weight to the sample information.
  - On the other hand, if the prior information is good, i.e., $\sigma^2 > \tau^2$, then the prior mean is given more weight

More on Bayesian Statistics

- **Example 3:** Consider a random IID sample from a normal distribution, i.e., $X_i \sim N(\theta, \sigma^2)$ for $i = 1, \ldots, n$. Suppose the prior distribution of is $\theta \sim N(\mu, \tau^2)$. Then, the posterior distribution of θ is also normal, with mean and variance given by:

$$E(\theta|x_1, \ldots, x_n) = \frac{n\tau^2}{n\tau^2 + \sigma^2}\left(\frac{\sum_{i=1}^n x_i}{n}\right) + \frac{\sigma^2}{n\tau^2 + \sigma^2}\mu, \quad Var(\theta|x_1, \ldots, x_n) = \frac{\sigma^2\tau^2}{n\tau^2 + \sigma^2}$$

- Notice that as we get more and more sample data, i.e., as n increases, the posterior estimate places more weight on the sample information and less on the prior.
- Moreover, when $n \to \infty$, the Bayes estimator of θ, $E(\theta|x_1, \ldots, x_n)$, tends toward the sample mean $\frac{\sum_{i=1}^n x_i}{n}$.

Example 3

- Further, if we want to determine the distribution of a future draw from the population, X_{n+1}, which is IID with $X_1, \ldots, X_n$, we can jointly use the posterior distribution of θ based on the information from observations $X_1, \ldots, X_n$, and the distribution of X_{n+1}.
- In other words, we have the following:

$$X_{n+1} \sim N(\hat{\mu}, \sigma^2 + \hat{\tau}^2)$$

Where $\hat{\mu} = E(\theta|x_1, \ldots, x_n)$, and $\hat{\tau}^2 = Var(\theta|x_1, \ldots, x_n)$

Overview of "Empirical" Bayes Approach

- The basic empirical Bayes approach uses observed data to estimate the parameters of the prior distribution, which are called ***hyper parameters***.
- The name Empirical Bayes arises from the fact that data from experiments are used to estimate the parameters of the prior distribution.
- EB is sometimes classified into parametric EB and nonparametric EB.
  - The major difference is that the parametric approach specifies a parametric family of prior distributions, but the nonparametric approach leaves the prior completely unspecified, and thus the prior distribution is fitted using the observed data.

Overview of "Empirical" Bayes Approach

- We demonstrate how to get EB estimators for the Normal case in which the prior and likelihood functions are Normal.
- Suppose p random variables are observed, each from a normal population with different means but the same known variance, that is,

$$X_i \sim N(\theta_i, \sigma^2), \quad i = 1, \dots, p$$

- Then the Bayesian assumption is made as,

$$\theta_i \sim N(\mu, \tau^2), \quad i = 1, \dots, p$$

- According to Bayes' rule, the Bayes estimator for θ_i is given by

$$\mu^{EB}(X_i) = \left(\frac{\sigma^2}{\sigma^2+\tau^2}\right)\mu + \left(\frac{\tau^2}{\sigma^2+\tau^2}\right)X_i$$

Overview of "Empirical" Bayes Approach

- The posterior distribution of θ_i given X_i, denoted by $\pi(\theta_i|X_i)$, is given by,

$$\pi(\theta_i|X_i) \sim N[\mu^{EB}(X_i), \sigma^2\tau^2/(\sigma^2+\tau^2)\,]$$

- The EB model agrees with the Bayes model, but refuses to specify values for μ and τ^2.
- Instead, the EB model uses the observed data to estimate the parameters in statistical way.
- All of the information about μ and τ^2 is contained in the marginal distribution of X_i (unconditional on θ_i) and some standard calculation shows that this marginal distribution of X_i, $f(X_i)$, is given by:

$$f(X_i) \sim N(\mu, \sigma^2 + \tau^2), \quad i = 1 \dots, p$$

Overview of "Empirical" Bayes Approach

- Using this fact, the unknown parameters in the expression of $\mu^{EB}(X_i)$, namely, μ, $\left(\frac{\sigma^2}{\sigma^2+\tau^2}\right)$, and $\left(\frac{\tau^2}{\sigma^2+\tau^2}\right)$, can be estimated.
- From Casella*, the following two equalities hold true:

$$E(\bar{X}) = \mu, \quad E\left(\frac{(p-3)\sigma^2}{\sum_{i=1}^{p}(X_i-\bar{\mathrm{X}})^2}\right) = \frac{\sigma^2}{\sigma^2+\tau^2}$$

- Then the EB estimators of those three parameters mentioned above are,

$$\bar{X}, \quad \frac{(p-3)\sigma^2}{\sum_{i=1}^{p}(X_i-\bar{\mathrm{X}})^2}, \quad 1 - \frac{(p-3)\sigma^2}{\sum_{i=1}^{p}(X_i-\bar{\mathrm{X}})^2}$$

* Casella, G. "An Introduction to Empirical Bayes Data Analysis," The American Statistician, May 1985, vol. 39, no.2, pp. 83-87.

Overview of "Empirical" Bayes Approach

- Thus the EB estimator of θ_i, $\mu^{EB}(X_i)$, is

$$\mu^{EB}(X_i) = \left(\frac{(p-3)\sigma^2}{\sum_{i=1}^{p}(X_i - \overline{\mathrm{X}})^2}\right)\bar{X} + \left(1 - \frac{(p-3)\sigma^2}{\sum_{i=1}^{p}(X_i - \overline{\mathrm{X}})^2}\right)X_i$$

- Casella demonstrates that $\mu^{EB}(X_i)$ is a good estimator of θ_i through several examples.
- In addition, EB estimation, on the average, is closer to θ_i than X_i, which is the usual/classical estimator of θ_i. Also if measured by the mean squared error (MSE), $\mu^{EB}(X_i)$ has the minimal MSE.
- The variance of EB estimator of θ_i, $V^{EB}(X_i)$, is

$$V^{EB}(X_i) = \sigma^2\left(1 - \frac{(p-1)(p-3)\sigma^2}{p\,\sum_{i=1}^{p}(X_i - \overline{\mathrm{X}})^2}\right) + \frac{2}{(p-3)}\left(\frac{(p-3)\sigma^2}{\sum_{i=1}^{p}(X_i - \overline{\mathrm{X}})^2}\right)^2 (X_i - \overline{\mathrm{X}})^2$$

Informative vs. Non-informative Priors

- In many cases, the goal of an analysis is to predict values of a future sample.
  - For example, estimate the number of new launch vehicles that will succeed in, say, *m* future launches scheduled.
  - If we knew the success probability for the launch of a new vehicle, π, the problem would be simple. However, we only know its posterior distribution.
  - In this case, the predictive probability of z (for a future sample of size m), given a posterior distribution on π based on past data y, is given by the integral

$$p(z|y) = \int_0^1 f(z|\pi)\, p(\pi|y) d\pi \qquad z = 0,1,\dots,m$$

Informative vs. Non-informative Priors

- In essence, by integrating the sampling distribution $f(z|\pi)$ over the posterior distribution on the parameter π. We average over the uncertainty in this parameter.
- The predictive distribution $p(z|y)$ provides a full account for the uncertainty in the unknown parameter, in this case π.

Section Summary

Georgia Tech

Covered the Basics of Bayesian Statistics.

Bayesian vs. Frequentist perspectives
Influence of Baye's Rule
Prior Distribution and Likelihood Functions Posterior
Conjugate Priors
Detailed Examples