BAYESIAN DATA IMPUTATION

by

WILLIAM HOLT
ADVISOR: DR. DUY NGUYEN

A thesis submitted in partial fulfillment
of the requirements for graduation with honors in
the major of data science and analytics

MARIST COLLEGE
Poughkeepsie, New York
May 2, 2023

Abstract

Datasets often have many missing values. There are many ways of imputing values in a dataset including listwise deletion and mean imputation. However, these cause significant penalties to the power of analysis. Bayesian imputation is a way of imputing missing values without incurring such steep penalties.

Contents

1 Introduction

Data is often messy and incomplete when collected. However, models cannot be trained on datasets with missing values. There are several methods to deal with the missing values. Listwise deletion is a rudimentary imputation method that involves deleting every row with a missing value. It is a poor imputation method because it loses a lot of information with the deleted rows. Mean imputation is slightly better than listwise deletion. It involves imputing the missing values with the mean or median of the column. This retains the missing information by not deleting rows like listwise deletion, but it doesn't take into account the relationship between the variables. A better approach is to use a Bayesian method to impute the data. The Bayesian approach uses Gibbs sampling to take the mean of the column and adjust it based on the other values in the row and the relationship between the variables.

2 Random Variables

2.1 Continuous Random Variables

Continuous random variables can take on an infinite number of values between a certain interval. This means that the probability that a continuous random variable will be any one value is zero. Instead of finding the probability of a continuous random variable taking on a specific value, the probability of the random variable being between a range of numbers is calculated. To do this a probability density function (PDF) must be used to calculate the random variable value at a specific point. A PDF must satisfy these two requirements:

1. $f(x) \geq 0, \quad \forall x \in \mathbb{R}$
2. $\displaystyle\int_{-\infty}^{\infty} f(x)dx = 1.$

These two requirements together guarantee that for any event A, the probability of A, denoted by $P(A)$, $0 \leq P(A) \leq 1$. Note that $P(A)$ is defined as follows

$$P(A) = \int_A f(x)dx.$$

For example, one of the most famous example is the normal random variable. In this case the PDF for a normal distribution is $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$. A finite integral of this equation does not admit a closed form. However, it can be computed up to any arbitrary accuracy level using statistical software such as $\mathbf{R}$. This will be explored further in the next section. For a normal random variable with mean μ and standard deviation σ, it is often denoted by $N(\mu, \sigma^2)$. In the Figure 1, we plot the graph of $f(x)$ with $\mu = 0$ and $\sigma = 1$.

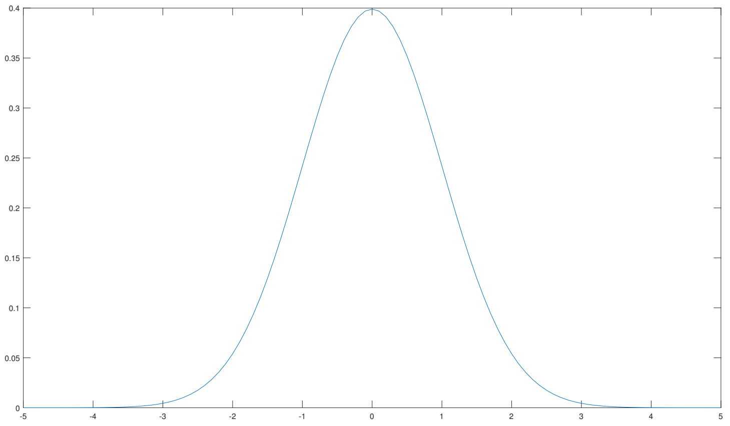


Figure 1: Graph of $f(x)$ with $\mu = 0, \sigma = 1$

2.2 Cumulative Density Function, Expected Value, and Variance

There are many properties of continuous random variables that will be needed further on in this thesis. To start, the probability density function can be integrated to become the cumulative probability function. Specifically, the cumulative density function (CDF) is defined as:

$$F(x) = P(X < x) = \int_{-\infty}^{x} f(x)dt.$$

An example of this is the CDF of a normal random variable $N(\mu, \sigma^2)$ is given by,

$$F(x) = P(X < x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

The expected value of a continuous random variable is defined as

$$\mu = E(x) = \int_{-\infty}^{\infty} xf(x)dt.$$

We have the following theorem regarding the expected value of a normal random variable $N(\mu, \sigma^2)$.

Theorem 1 (Expected Value of a Normal Distribution) *Let X be a random variable following a normal distribution: $X \sim N(\mu, \sigma^2)$. Then, the mean or expected value of X is $E(X) = \mu$.*

Proof. The expected value is the probability-weighted average over all possible values:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

With the probability density function of the normal distribution, this reads:

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \end{aligned}$$

Substituting $z = \dfrac{x-\mu}{\sqrt{2}\sigma}$

$$\begin{aligned} &= \frac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(\sqrt{2}\sigma z + \mu\right) \exp\left(-z^2\right) dz \\ &= \frac{1}{\sqrt{\pi}} \left(\sqrt{2}\sigma \int_{-\infty}^{\infty} z\exp(-z^2)dz + \mu \int_{-\infty}^{\infty} \exp(-z^2)dz\right) \\ &= \frac{1}{\sqrt{\pi}} \left(\sqrt{2}\sigma \left[-\frac{1}{2}\exp(-z^2)\right]_{-\infty}^{\infty} + \mu\sqrt{\pi}\right) \\ &= \frac{\mu\sqrt{\pi}}{\sqrt{\pi}} \\ &= \mu. \end{aligned}$$

The variance of a continuous random variable measures it's spread. The higher the variance the more spread out the distribution is. The variance of a random varible with CDF $f(x)$ is defined as

$$\sigma^2 = \text{var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx.$$

Let us consider an example. In the below, we will compute the variable of a normal random variable $N(\mu, \sigma^2)$.

Theorem 2 (Variance of a Normal Distribution) *Let X be a random variable following a normal distribution: $X \sim N(\mu, \sigma^2)$. Then, the variance of X is* $\mathrm{var}(X) = \sigma^2$.

Proof. The variance is the expectation of the squared deviation of a random variable from its mean

$$\mathrm{var}(X) = \int_{-\infty}^{\infty} x^2 f(x)dx - E(X)^2$$

With the probability density function of the normal distribution, this reads:

$$\begin{aligned}\mathrm{var}(X) &= \int_{-\infty}^{\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx - \mu^2 \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx - \mu^2\end{aligned}$$

Substituting $z = \dfrac{x-\mu}{\sqrt{2}\sigma}$

$$\begin{aligned}\mathrm{var}(X) &= \frac{\sqrt{2}\sigma}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(\sqrt{2}\sigma z + \mu\right)^2 \exp -t^2 dz - \mu^2 \\ &= \frac{1}{\sqrt{\pi}} \left(2\sigma^2 \int_{-\infty}^{\infty} z^2 \exp(-z^2)dz + 2\sqrt{2}\sigma\mu \int_{-\infty}^{\infty} z \exp(-t^2)dz + \mu^2 \int_{-\infty}^{\infty} \exp(-z^2)dz\right) - \mu^2 \\ &= \frac{1}{\sqrt{\pi}} \left(2\sigma^2 \int_{-\infty}^{\infty} z^2 \exp(-z^2)dz + 2\sqrt{2}\sigma\mu \left[-\frac{1}{2}\exp(-z^2)\right]_{-\infty}^{\infty} + \mu^2\sqrt{\pi}\right) - \mu^2 \\ &= \frac{1}{\sqrt{\pi}} \left(2\sigma^2 \int_{-\infty}^{\infty} z^2 \exp(-z^2)dz + 2\sqrt{2}\sigma\mu(0)\right) + \mu^2 - \mu^2 \\ &= \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} z^2 \exp(-z^2)dz \\ &= \frac{2\sigma^2}{\sqrt{\pi}} \left(\left[-\frac{z}{2}\exp(-z^2)\right]_{-\infty}^{\infty} + \frac{1}{2}\int_{-\infty}^{\infty} \exp(-z^2)dz\right) \\ &= \frac{2\sigma^2}{\sqrt{\pi}} \left(\frac{1}{2}\right) \int_{-\infty}^{\infty} \exp(-z^2)dz \\ &= \frac{2\sigma^2\sqrt{\pi}}{2\sqrt{\pi}} \\ &= \sigma^2\end{aligned}$$

The standard deviation of a continuous random variable distribution is the square root of the variance. That is, standard Deviation: $\sqrt{Var(x)}$.

In Table **??**, we provide a list of several random variables used in this thesis.

Name	PDF	Mean	Var
Normal	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	μ	σ^2
Normal			
Normal			

We have the following theorem regarding the average of a finite set of normal random variables.

Theorem 3 (Law of Large Number) *Assume that $X_1, X_2, \ldots, X_n$ is an independent and identically distributed sample and $X_i \sim N(\mu, \sigma^2)$. Then we have*

$$\bar{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \sim N(\mu, \frac{\sigma^2}{n}).$$

Proof. Let $X_1, X_2, X_3, ..., X_n$ be identically and independently distributed ($i.i.d$) variables with mean μ and standard deviation σ^2. First, to calculate $E[\bar{X}_n]$, we get,

$$\begin{aligned} E[\bar{X}_n] &= E[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}E[\sum_{i=1}^{n} X_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i E[x_i] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{n} = \mu \end{aligned}$$

Calculating $Var[Y]$, we get,

$$\begin{aligned} Var[\bar{X}_n] &= Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}Var(\sum_{i=1}^{n} X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\ &= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \end{aligned}$$

The following theorem give a reason why normal random variables are very common in the probability/statistics world.

Theorem 4 (Central Limit Theorem) *Assume that $X_1, X_2, \ldots, X_n$ is an independent and identically distributed sample, $E(X_i) = \mu = 0$, $Var(X_i) = \sigma^2$, and let $S_n = \sum_{i=1}^{n} X_i$, then*

$$\lim_{n\to\infty} P\left(\frac{S_n}{\sigma\sqrt{n}} \le x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

Proof. Let $Z_n = \frac{S_n}{\sigma\sqrt{n}}$. Recall that the moment generating function of $N(0,1)$ is given by $e^{t^2/2}$. To proof the theorem, we will show that the moment generating function of Z_n is approaching to that of the standard normal distribution. Note that by independent, we have the moment generating function of Z_n is given by

$$\begin{aligned} M_{S_n}(t) &= E(e^{tS_n}) = E(e^{t\sum_{i=1}^{n} X_i}) \\ &= E(e^{tX_1})E(e^{tX_2}) \cdot \ldots \cdot E(e^{tX_n}) = E(e^{tX_1})E(e^{tX_1}) \cdot \ldots \cdot E(e^{tX_1}) \\ &= \prod_{i=1}^{n} \left(E(e^{tX_1})\right) = \left(E(e^{tX_1})\right)^n = \left(M_{X_1}(t)\right)^n. \end{aligned}$$

Hence we have by the above calculation

$$M_{Z_n}(t) = E(e^{tZ_n}) = E(e^{t\frac{S_n}{\sigma\sqrt{n}}}) = E(e^{\frac{t}{\sigma\sqrt{n}}S_n}) = \left(M_{X_1}(\frac{t}{\sigma\sqrt{n}})\right)^n.$$

Consider the moment generating function of X_1, $M_{X_1}(t)$. By the Taylor's expansion, we have

$$M_{X_1}(t) = M_{X_1}(0) + tM'_{X_1}(0) + \frac{t^2}{2}M''_{X_1}(0) + \epsilon(t),$$

where $\epsilon(t)/t^2 \to 0$ as $t \to 0$. Recall that

$$M_{X_1}(0) = 1, M'_{X_1}(0) = E(X_1) = \mu = 0, M''_{X_1}(0) = E(X_1^2) = \mu + \sigma^2 = 0 + \sigma^2.$$

Hence

$$M_{X_1}(t) = 1 + t0 + \frac{t^2}{2}\sigma^2 + \epsilon(t).$$

By the above calculation, we have

$$\begin{aligned} M_{X_1}(\frac{t}{\sigma\sqrt{n}}) &= 1 + \frac{1}{2}\sigma^2\left(\frac{t}{\sigma\sqrt{n}})\right)^2 + \epsilon(\frac{t}{\sigma\sqrt{n}})) \\ &= 1 + \frac{t^2}{2n} + \epsilon = \left(1 + \frac{t^2}{2n} + \epsilon(\frac{t}{\sigma\sqrt{n}}))\right). \end{aligned}$$

Recall that $\lim_{n\to\infty}(1+\frac{z}{n}) = e^z$, we have,

$$\begin{aligned} \lim_{n\to\infty} M_{Z_n}(t) &= \lim_{n\to\infty}\left(M_{X_1}(\frac{t}{\sigma\sqrt{n}})\right)^n \\ &= \lim_{n\to\infty} = \left(1 + \frac{t^2}{2n} + \epsilon(\frac{t}{\sigma\sqrt{n}}))\right)^n = e^{t^2/2}. \end{aligned}$$

Hence, the moment generating function of Z_n is converging to the moment generating of the standard normal variable. As a result, we have

$$P(Z_n \le x) = \lim_{n\to\infty} P\left(\frac{S_n}{\sigma\sqrt{n}} \le x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt.$$

This completes the proof of the central limit theorem.

Remark 2.1 *In case, $X_1, X_2, \ldots, X_n$ is an independent and identically distributed sample, $E(X_i) = \mu \neq 0$, $Var(X_i) = \sigma^2$, and let $S_n = \sum_{i=1}^{n} X_i$, then*

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

3 Multivariate Normal Distribution

3.1 Multivariate Statistics

Multivariate distributions involve using multiple independent variables for one dependent variable. A typical univariate distribution is 2-D, so it has one X variable and one Y variable. These distributions can be viewed on a 2d graph such as a univariate normal distribution in figure 1. Multivariate distributions are 3-D and higher. 3-D distributions can be viewed in 3-D space, but 4-D and higher cannot be viewed in their entirety, but marginal distributions of the data can be. The marginal distributions are the univariate subsets of the multivariate distribution.

3.2 Multivariate Normal Equations

The multivariate normal equation is notated by $N(\mu, \Sigma)$ containing k variables or in other words having k dimensions. μ is a vector of size k containing the expected value of each dimension of the normal distribution. $\mu = E(X) = (E(X_1), E(X_2), ..., E(X_k))$. Σ is a variance-covariance matrix of size kxk containing the covariance values.

The Multivariate Normal PDF is as follows: $P(X|\mu, \Sigma) = (2\pi)^{-k/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$.
Where X is k dimensional.
Covariance is the measure of spread for the multivariate normal distribution. Along the main diagonal of the variance-covariance matrix are the variance values and all the other values are the covariance values between the different variables. Covariance is equated by the following equation:

Covariance: $cov(X, Y) = E((X - \mu(X))(Y - \mu(Y)))$

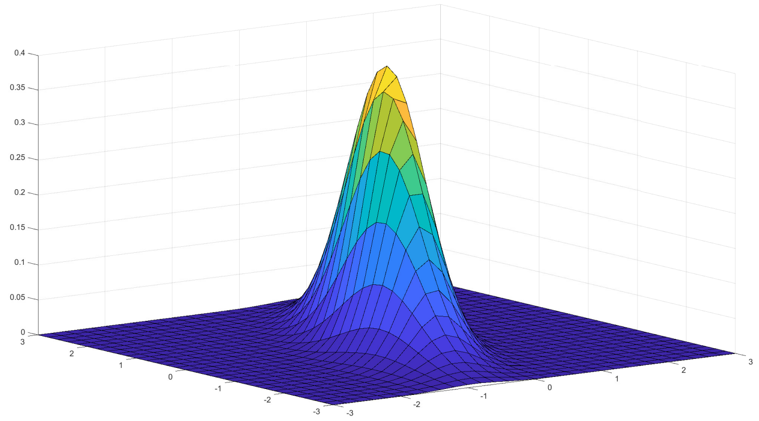


Figure 2: An image of a multivariate normal distribution and its marginal distributions

$$\sum = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2}^2 & \cdots & \sigma_{1,p}^2 \\ \sigma_{2,1}^2 & \sigma_2^2 & \cdots & \sigma_{2,p}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p,1}^2 & \sigma_{p,2}^2 & \cdots & \sigma_p^2 \end{pmatrix}$$

Figure 3: This is an example of an uncorrelated variance-covariance matrix. Notice how only the variance values are filled in, while the covariance values are zero.

Correlation is a measure of how related two variables are. The closer to positive or negative one the correlation is the more connected two variables are. This can be either a positive or a negative relationship. It is found by the equation below:

Correlation: $\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$

A covariance matrix is a square matrix giving the covariance between every pair of variables in a multivariate model. Every covariance matrix is symmetrical along the diagonal with the main diagonal denoting the variance of each variable.

Covariance Matrix: $\Sigma = cov(X) = \begin{pmatrix} E(X_1^2) - E(X_1)^2 & E(X_1X_2) - E(X_1)E(X_2) \\ E(X_1X_2) - E(X_1)E(X_2) & E(X_2^2) - E(X_2)^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix}$

3.3 Multivariate Normal Correlation

Here is an example of a multivariate normal distribution with two variables y_1 and y_2. Using different values for the covariance matrix creates very different distributions. The closer the correlation values are to zero the more circular the plot appears. This is showing how much of an effect correlation values have for a multivariate normal distribution.

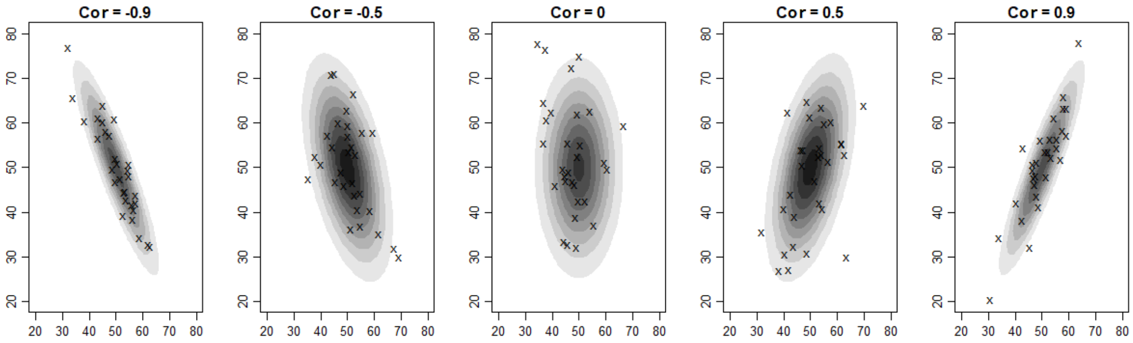


Figure 4: An image of multivariate normal distributions with 5 different correlation values

4 Bayesian Statistics

4.1 Bayes' Theorem

Bayes' Theorem measures the probability of an event given a prior set of beliefs related to the event. For example, this can be used to make more accurate predictions on the risk of developing heart disease based on the prior knowledge of the patients' age or weight.

Theorem 5 (Bayes' Theorem) *Let A and B be events where $P(B) \neq 0$. Then Bayes' Theorem is*

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

Proof. Bayes' theorem can be derived from conditional probability. The conditional probability formula is:

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

Solving for $P(A \cap B)$ from the above equation, we have

$$P(A \cap B) = P(A)P(B|A).$$

Similarly, we have

$$P(A \cap B) = P(B)P(A|B).$$

From the above two equation, we obtain

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

This completes the proof of the theorem.

4.2 Bayesian Inference

Bayesian inference is one of the many applications of Bayes' theorem. Bayesian inference involves taking a prior belief and updating this belief based on new evidence that is applied. The Bayes' Rule for Bayesian Inference is:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

- $P(\theta)$ is our prior belief. It is our knowledge known going into the Bayesian inference. It is what we believe the probability of θ is going into the Bayesian inference.

- $P(D|\theta)$ is the likelihood. It is a measure of how well the new evidence matches with the prior belief.
- $P(\theta|D)$ is the posterior belief. It is what we believe the new probability pf θ to be.
- $P(D)$ is the marginal likelihood or the model evidence. This ia how strongly we believe that the posterior distribution is correct. It is evenly applied to all possible values of D, so it doesn't affect the relative probabilities of the different values of D.

4.3 Bayesian Inference Example

Three students are trying to determine what proportion of Marist students support building a water park on campus. They have tree different idea about what the prior distribution should be. Anna chooses to use the prior distribution of a beta distribution with an alpha of 4.8 and a beta of 19.2. Bart chooses to use a uniform distribution of $y = 1$ for his prior. Finally, Chris uses a peicewise function for his prior given by

$$g(x) = \begin{cases} 20x & \text{if } x \in [0, 0.1] \\ 0.2 & \text{if } x \in [0.1, 0.3] \\ 5 - 10x & \text{if } x \in [0.3, 0.5] \end{cases}$$

The prior distribution of the students is shown below in figure 4.

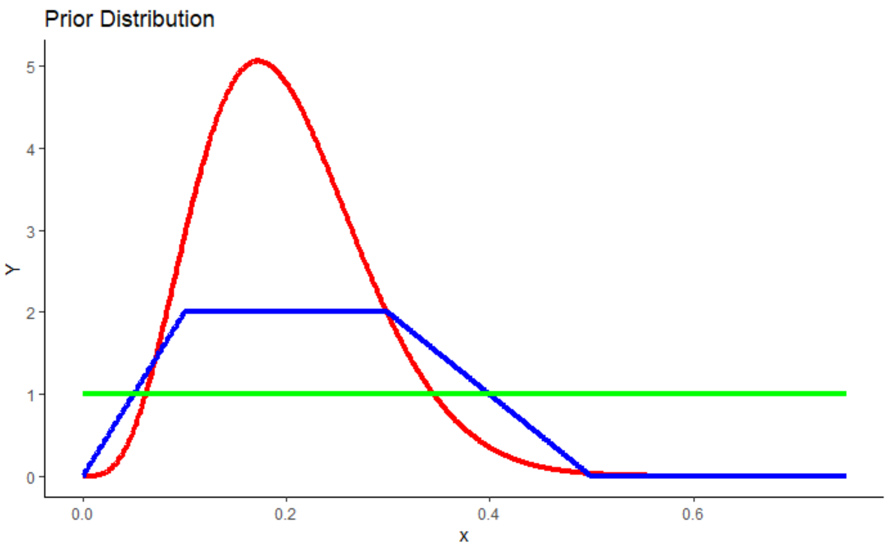


Figure 5: The prior distributions of the three students. The red line is Anna, The blue line is Chris, and the green line is Bart

The three students take a random sample of 100 Marist students and find their views on the waterpark. Out of the random sample, 26 say thet support the waterpark. The posterior distributions are found by taking this new information and applying it to the prior beliefs of the three students. The first two students can use a beta distribution prior. The posterior distribution for a beta distribution prior where the prior is g(x) $\sim$ Beta(α, β) and the likelihood function is $f(x|p)$ is as follows:

$$
\begin{aligned}
g(p) &= \frac{1}{Beta(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1} \\
f(y|p) &= \binom{n}{y} p^{x}(1-p)^{n-y} \\
f(y|x) &= \frac{f(y|p)}{f_y(y)} g(p) \\
&= \frac{\binom{n}{y} p^{y}(1-p)^{n-y}}{f_y(y)} \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1} \\
&\propto p^{y}(1-p)^{n-y} p^{\alpha-1}(1-p)^{\beta-1} \\
&\propto p^{\alpha+y-1}(1-p)^{\beta+n-y-1}
\end{aligned}
$$

Here we can see that $f(p|x)$ has a Beta distribution of $B(\alpha+y, \beta+n-y)$. It isn't necessary to find $f_y(y)$ and we can ignore the constants of both the prior and the likelihood. Using this equation, Anna has a prior of $B(4.8+y, 19.2+n-y) = B(30.8, 93.2)$ and Bart has a posterior of $B(1+y, 1+n-y) = B(27, 75)$. Chris's prior is found by using R to approximate the integral of Chris's prior distribution and then multiplying it by the beta likelihood equation from above. The posterior distributions are graphed below in figure 5.

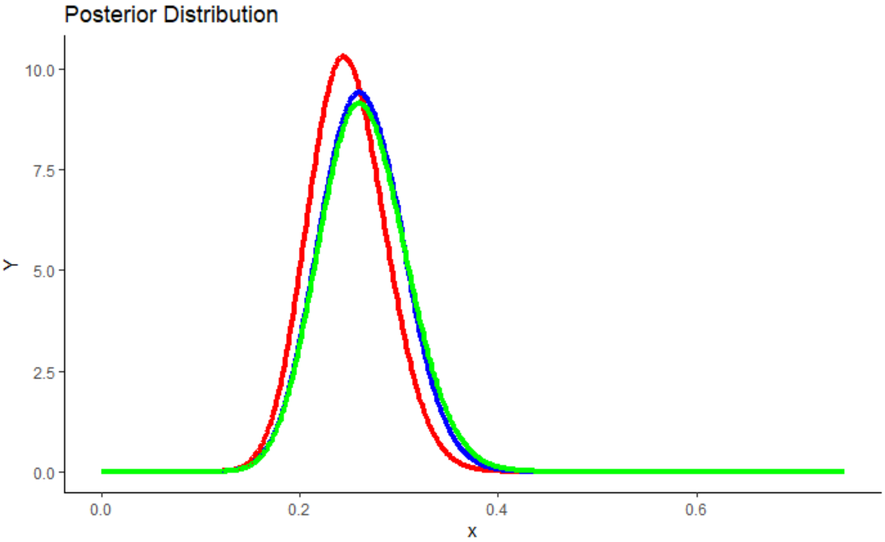


Figure 6: The posterior distributions of the three students. The red line is Anna, The blue line is Chris, and the green line is Bart

As can be seen above, the posterior distributions are all very similar. This means that in this instance, the likelihood from the new information from the polled students was much stronger than the prior beliefs of the students. This made three of the distributions' means fall within .015 of each other. This shows how in some instances, one's prior doesn't have much of an effect on the final posterior distribution.

5 Normal Bayesian Inference

5.1 Estimating the Mean

Estimating the mean and variance of a normal distribution using Bayesian Inference is quite a bit more difficult than the previous example using a beta distribution. That is because it is impossible to integrate the normal distribution probability density function. To get around this a sampling method called Gibbs sampling is used to approximate the mean and variance. This will be explored later. However, when one of the parameters is already known normal Bayesian inference does have a closed form solution. A common way to find a closed form solution is to use a known variance and use Bayesian inference to find the mean.

The probability density function for a normal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Because all values of x are independent, the likelihood function is

$$p(x|\mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

$$p(x|\mu) \propto \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

The prior is a normal distribution with a mean of μ_0 and a variance of τ_0^2. This assumes that μ has a normal distribution. The equation is as follows:

$$p(\mu) = \frac{1}{\tau_0\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\mu-\mu_0}{\tau}\right)^2\right)$$

$$p(\mu) \propto \exp\left(-\frac{1}{2}\left(\frac{\mu-\mu_0}{\tau}\right)^2\right)$$

Given the above prior and likelihood, the posterior distribution is

$$p(\mu) \times p(x|\mu) \propto \exp\left(-\frac{1}{2}\left[\left(\frac{\mu-\mu_0}{\tau}\right)^2 + \left(\frac{x-\mu}{\sigma}\right)^2\right]\right)$$

This simplifies as follows:

$$\begin{aligned}
p(\mu) \times p(x|\mu) &\propto \exp\left(-\frac{1}{2}\left[\frac{\sigma^2(\mu^2 - 2\mu\mu_0 + \mu_0^2) + \tau^2(x^2 - 2x\mu + \mu^2)}{\sigma^2\tau^2}\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left[\frac{(\sigma^2+\tau^2)\mu^2 - 2(\sigma^2\mu_0 + \tau^2 x)\mu + \mu_0^2\sigma^2 + x^2\tau^2}{\sigma^2\tau^2}\right]\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2\tau^2/(\sigma^2+\tau^2)}\left[\mu^2 - 2\mu\frac{(\sigma^2\mu_0 + \tau^2 x)}{\sigma^2+\tau^2} + (\frac{(\sigma^2\mu_0 + \tau^2 x}{\sigma^2+\tau^2})^2\right]\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2\tau^2/(\sigma^2+\tau^2)}\left[\mu - \frac{(\sigma^2\mu_0 + \tau^2 x)}{\sigma^2+\tau^2}\right]^2\right)
\end{aligned}$$

From this equation we can see that it is a normal distribution with the following parameters:

$$\begin{aligned}\mu_n &= \frac{(\sigma^2\mu_0 + \tau^2 x)}{\sigma^2 + \tau^2}\\ \sigma_n^2 &= \frac{\sigma^2\tau^2}{(\sigma^2 + \tau^2}\end{aligned}$$

Now that the posterior distribution is derived, all that is left is to find the updating rule for updating the parameters. First we will look at the inverse of variance which is also known as precision. The posterior precision is as follows:

$$\frac{1}{\sigma_n^2} = \frac{\sigma^2 + \tau^2}{\sigma^2\tau^2} = \frac{1}{\tau^2} + \frac{1}{\sigma^2}$$

Therefore, the posterior mean can be written as the following:

$$\begin{aligned}\mu_n &= \frac{(\sigma^2\mu_0 + \tau^2 x)}{\sigma^2 + \tau^2}\\ &= \mu_0\frac{\sigma^2}{\sigma^2 + \tau^2} + x\frac{\tau^2}{\sigma^2 + \tau^2}\\ &= \mu_0\frac{1/\tau^2}{1/\sigma^2 + 1/\tau^2} + x\frac{1/\sigma^2}{1/\sigma^2 + 1/\tau^2}\end{aligned}$$

When a random sample is taken from a normally distributed with mean μ and variance σ^2. Therefore, the sample is normally distributed by the central limit theorem and law of large numbers. The likelihood function will use the sample mean of $\bar{x}$, a sample variance of $\frac{\sigma^2}{n}$, and a precision of $\frac{n}{\sigma^2}$. Thus the posterior precision is:

$$\begin{aligned}\frac{1}{\sigma_n^2} &= \frac{1}{\tau^2} + \frac{n}{\sigma^2}\\ &= \frac{\sigma^2 + n\tau^2}{\sigma^2\tau^2}\end{aligned}$$

The posterior distribution mean standard deviation is then equal to:

$$\frac{1}{1/\sigma_n^2} = \sigma_n^2 = \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}$$

The posterior mean is equal to the weighted average of the prior mean and $\bar{x}$ where the weight are the proportions of the posterior precision. It is as follows:

$$\mu_n = \mu_0\frac{1/\tau^2}{n/\sigma^2 + 1/\tau^2} + \bar{x}\frac{n/\sigma^2}{n/\sigma^2 + 1/\tau^2}$$

When the population variance is unknown, the sample variance can be used as a substitute. This is much less accurate and as a result the posterior variance will be much greater than if the population variance was known. The credible interval will also be much wider.

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

5.2 Estimating the Variance

Given a random sample from a normal distribution $N(\mu, \sigma^2)$ where the mean is assumed known, but the variance is unknown, how can we estimate the variance of the population? Using a Bayesian approach it can be summarized as follows:

$$g(\sigma^2|x_1, ..., x_n) \propto g(\sigma^2) \times f(x_1, ..., x_n|\sigma^2)$$

Because the prior is continuous and the standard deviation has to be above 0, the variance can be described as follows:

$$g(\sigma^2|x_1,...,x_n) = \frac{g(\sigma^2) \times f(x_1,...,x_n|\sigma^2)}{\int g(\sigma^2) \times f(x_1,...,y_n|\sigma^2) d\sigma^2}$$

To estimate the variance, we will use an inverse gamme prior as the prior for the variance is often difficult to properly guess and is quite vague. Inverse gamma distribution has a pdf of:

$$g(\sigma^2;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} \exp\left(-\frac{\beta}{\sigma^2}\right)$$

Let K equal degrees of freedom of the sample and S be the multiplyer to the chi-squared distribution. We can see that if $\alpha = \frac{K}{2}, \beta = 1/2$, then the inverse gamma distribution is the same as an inverse chi-squared distribution with K degrees of freedom. If instead $\beta = S/2$, then it is equal to S times an inverse chi-squared distribution with K degrees of freedom. The posterior distribution is then as follows:

$$g(\sigma^2|x_1,...,x_n) \propto \frac{1}{(\sigma^2)^{\alpha+1}} \exp\left(-\frac{\beta}{\sigma^2}\right) \times \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{SS_T}{2\sigma^2}\right)$$

Where SS_T equals:

$$SS_T = \sum_{i=1}^{n} (x_i - \mu)^2$$

5.3 Estimating Mean and Variance

Both of the above inferences assumed that either the mean or the variance was known. This is necessary to have a closed form solution to the posterior distribution. If neither the variance nor the mean are known, then more advanced inferencing techniques need to be used to approximate the posterior distribution which is itself an approximation of the population distribution. The technique that will be discussed here is Gibbs sampling.

Gibbs sampling is a form Markov chain Monte Carlo meaning that is chaining together samples using the previous sample to draw the next one. There is also the Metropolis-Hastings algorithm which is another form of Markov chain Monte Carlo methods, but it will not be discussed further. Gibbs sampling cycles through parameters using the previous sample parameters one at a time to calculate the next one. When allowed to run for many iterations, the Gibbs sampler algorithm will converge to the posterior distribution.

To start pick an arbitrary starting value for each parameter. In this case, we will use the sample mean and variance as the starting values. ϕ will hold all of the parameters where $\phi_1 = \mu_n$ and $\phi_2 = 1/\sigma^2$

$$\phi^{(1)} = (\phi_1^{(1)} = \mu_0, \phi_2^{(1)} = 1/\tau^2)$$

Next, draw $\phi_1^{(2)}$ to get the next μ_n

$$\begin{aligned}
\mu_n &= \mu_0 \frac{1/\tau^2}{n\phi_2^{(1)} + 1/\tau^2} + \bar{x} \frac{n\phi_2^{(1)}}{n\phi_2^{(1)} + 1/\tau^2} \\
\tau_n^2 &= \frac{1}{n\phi_2^{(1)} + 1/\tau^2} \\
\phi_1^{(2)} &= \text{rnorm}(\mu_n, \tau_n^2)
\end{aligned}$$

Then, draw $\phi_2^{(2)}$ to get the next $1/\sigma^2$

$$\begin{aligned}
\nu_n &= \nu_0 + n \\
s_n^2 &= \frac{(\nu_0 s_0^2 + (n-1)\sigma_0^2 + n(\bar{x} - \phi_1^{(2)})^2}{\nu_n} \\
\phi_2^{(2)} &= \text{rgamma}\left(\frac{\nu_n}{2}, \frac{\nu_n s_n^2}{2}\right)
\end{aligned}$$

Repeat these steps until the Gibbs sampler converges which typically takes a few thousand loops. Then, the average of the ϕ values will be the average for each parameter. Quantiles can be used to estimate the variance for each of the posterior distributions for the two variables.

5.4 Example

In this example, Marist students are measuring the wing length of finches. The accepted mean wing length is 1.9 inches with a standard deviation of 0.1 inches. The students catch 9 finches and measure them. There results are [1.64,1.70,1.72,1.74,1.82,1.82,1.82,1.90,2.08]. As this might be a new species, the mean and variance are unknown. Is there significant evidence that they caught a new species of finch?

For this calculation, a Gibbs Sampler is needed. Using θ to represent the mean and $\tilde{\sigma}$ to calculate standard deviation, the Gibbs sampler results look like the following.

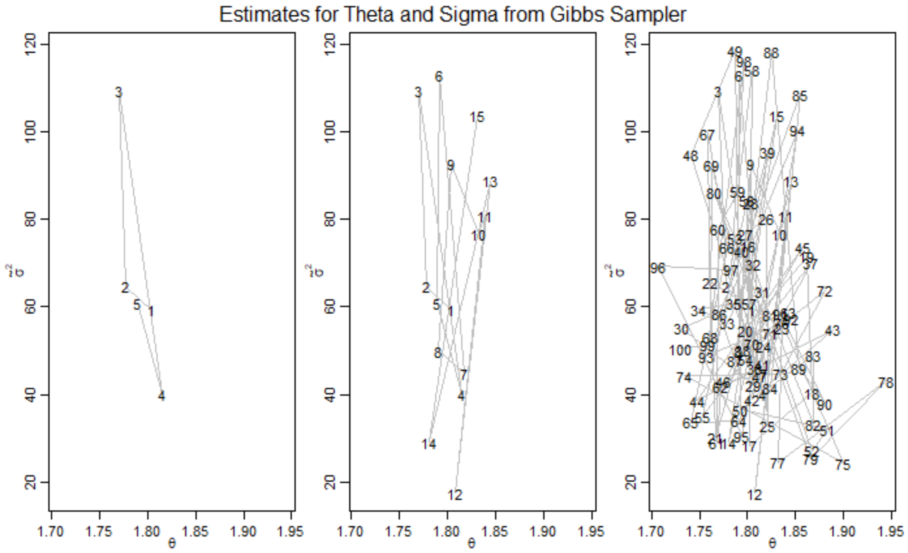


Figure 7: This shows the progression of the Gibbs sampler as it makes samples from 5 to 100.

As we can see from the first few samples from the Gibbs sampler, the samples appear to be withing 1.7 and 1.95. After calculating, the standard deviation appears to be within 0.1 and 0.25. Now the Gibbs sampler will be allowed to iterate 1000 times. The final distribution of samples is below.

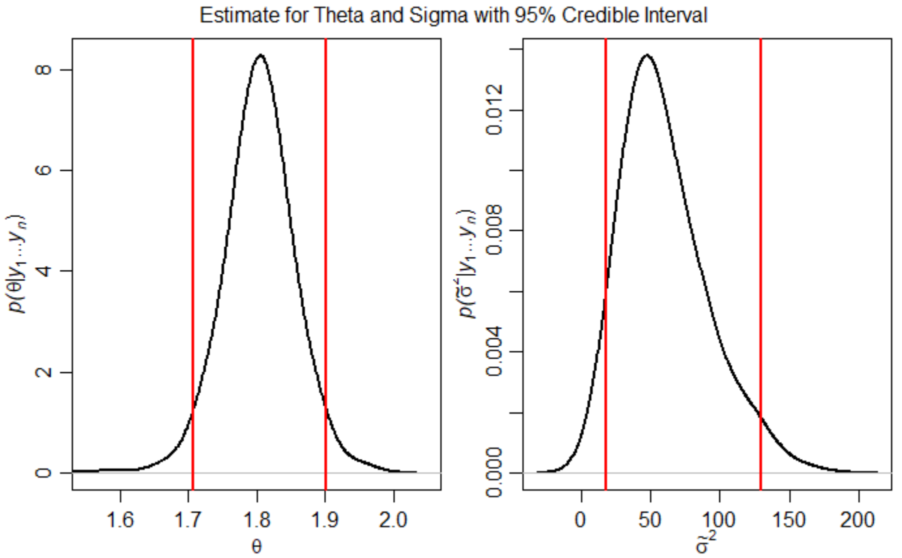


Figure 8: The distributions of θ and $\tilde{\sigma}$.

After the Gibbs sampler is allowed to run for 1000 iterations, θ has a mean of 1.804 with a 95% Credible Interval of $[1.707, 1.901]$. The σ has a mean of 0.137 with a 95% Credible Interval of $[0.088, 0.239]$. This means that there is insufficient evidence that this is a new species of finch as 1.9 inches falls within the 95% credible interval for the mean wing length.

6 Multivariate Normal Bayesian Inference

6.1 When Means are Unknown and Covariance is Known

A conveinient semiconjugate prior distribution to find the mean for a multivariate normal distribution is a multivaariate normal distribution. This is the same as using a univariate normal prior for a univariate normal population as shown in the section above. We will parameterize this as:

$$p(\boldsymbol{\theta}) = \text{multivariatenormal}(\boldsymbol{\mu}_0, \Lambda_0)$$

The full prior distribution is then as follows:

$$\begin{aligned}
p(\boldsymbol{\theta}) &= (2\pi)^{-p/2}|\Lambda_0|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu}_0)^T\Lambda_0^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu}_0)\right)\\
&= (2\pi)^{-p/2}|\Lambda_0|^{-1/2}\exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\Lambda_0^{-1}\boldsymbol{\theta}+\boldsymbol{\theta}^T\Lambda_0^{-1}\boldsymbol{\mu}_0-\frac{1}{2}\boldsymbol{\mu}_0^T\Lambda_0^{-1}\boldsymbol{\mu}_0\right)\\
&\propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\Lambda_0^{-1}\boldsymbol{\theta}+\boldsymbol{\theta}^T\Lambda_0^{-1}\boldsymbol{\mu}_0\right)\\
&= \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\,\mathbf{A}_0\,\boldsymbol{\theta}+\boldsymbol{\theta}^T\,\mathbf{b}_0\right)
\end{aligned}$$

Where $\mathbf{A}_0 = \Lambda^-1_0$ and $\mathbf{b}_0 = \Lambda_0^{-1}\boldsymbol{\mu}_0$.

The joing sampling density or the likelihood can also be observed as a sampling of a normal population $\{\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n|\boldsymbol{\theta}, \Sigma\}$, so it can be shown as follows:

$$\begin{aligned}
p(\boldsymbol{y}_1, ..., \boldsymbol{y}_n|\boldsymbol{\theta}, \Sigma) &= \prod_{i=1}^{n}(2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{y}_i-\boldsymbol{\theta})^T\Sigma^{-1}(\boldsymbol{y}_i-\boldsymbol{\theta})\right)\\
&= (2\pi)^{-np/2}|\Sigma|^{-n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left[(\boldsymbol{y}_i-\boldsymbol{\theta})^T\Sigma^{-1}(\boldsymbol{y}_i-\boldsymbol{\theta})\right]\right)\\
&\propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\,\mathbf{A}_1\,\boldsymbol{\theta}+\boldsymbol{\theta}^T\,\mathbf{b}_1\right)
\end{aligned}$$

Where $\mathbf{A}_1 = n\Sigma^-1$, $\mathbf{b}_1 = n\Sigma^-1\bar{\boldsymbol{y}}$, and $\bar{\boldsymbol{y}} = \left(\frac{1}{n}\sum_{i=1}^{n} y_{i,1}, ..., \frac{1}{n}\sum_{i=1}^{n} y_{i,p}\right)^T$.

Combining the two previous equation we have as follows:

$$\begin{aligned}
p(\boldsymbol{\theta}|\boldsymbol{y}_1, ..., \boldsymbol{y}_n, \Sigma) &\propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\,\mathbf{A}_0\,\boldsymbol{\theta}+\boldsymbol{\theta}^T\,\mathbf{b}_0\right)\times\exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\,\mathbf{A}_1\,\boldsymbol{\theta}+\boldsymbol{\theta}^T\,\mathbf{b}_1\right)\\
&= \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\,\mathbf{A}_n\,\boldsymbol{\theta}+\boldsymbol{\theta}^T\,\mathbf{b}_n\right)
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{A}_n &= \mathbf{A}_0+\mathbf{A}_1 = \Lambda_0^{-1}+n\Sigma^{-1}\\
\mathbf{b}_n &= \mathbf{b}_0+\mathbf{b}_1 = \Lambda_0^{-1}\boldsymbol{\mu}_0+n\Sigma^{-1}\bar{\boldsymbol{y}}.
\end{aligned}$$

This is the posterior distribution for estimating the mean when the covariance matrix is known for multivariate normal Bayesian inference.

6.2 When Means are Known and Covariance is Unknown

Recall that for a univariate normal distribution an inverse-Gamma distribution was used to as the prior to estimate the posterior variance. For multivariate normal models an inverse-Wishart distribution is used. A Wishart distribution is a multivariate form of a Gamma distribution, so it makes sense to use as a prior for estimating posterior variance. The inverse-Wishart density function is given as follows:

$$p(\Sigma) = \left[2^{v_0 p/2} \pi^{\binom{p}{2}/2} |\boldsymbol{S}_0|^{-v_0/2} \prod_{j=1}^{p} \Gamma([v_0 + 1 - j]/2) \right]^{-1} \times \\ |\Sigma|^{-(v_0+p+1)/2} \times \exp[-\operatorname{tr}(\boldsymbol{S}_0 \Sigma^{-1})/2]$$

The sampling distribution is again a multivariate normal sample distribution which is as follows:

$$p(\boldsymbol{y}_1, ..., \boldsymbol{y}_n | \boldsymbol{\theta}, \Sigma) = (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \left[(\boldsymbol{y}_i - \boldsymbol{\theta})^T \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\theta}) \right] \right)$$

Using matrix algebra the sum $\Sigma_{k=1}^{K} \mathbf{b}_k^T \, \mathbf{A} \, \mathbf{b}_k = \operatorname{tr}(\mathbf{B}^T \, \mathbf{B} \, \mathbf{A})$ where $\mathbf{B}$ is the matrix whose kth row is $\mathbf{b}_k^T$. Using this knowledge, the sampling distribution from above can be re written as the following:

$$p(\boldsymbol{y}_1, ..., \boldsymbol{y}_n | \boldsymbol{\theta}, \Sigma) = (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\left(-\frac{1}{2} \operatorname{tr}(\boldsymbol{S}_\theta \Sigma^{-}1) \right)$$

where

$$\boldsymbol{S}_\theta = \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\theta})^T \Sigma^{-1} (\boldsymbol{y}_i - \boldsymbol{\theta})$$

The matrix $\boldsymbol{S}_\theta$ is the residual sum of squares for the vectors $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ conditional on $\boldsymbol{\theta}$. Therefore, $\frac{1}{n}\boldsymbol{S}_\theta$ is an unbiased estimator of covariance matrix for the population.

Combining the two previous equations to get the full posterior distribution gives us the following:

$$\begin{aligned} p(\boldsymbol{\theta} | \boldsymbol{y}_1, ..., \boldsymbol{y}_n, \Sigma) &\propto p(\Sigma) \times p(\boldsymbol{y}_1, ..., \boldsymbol{y}_n | \boldsymbol{\theta}, \Sigma) \\ &\propto \left(|\Sigma|^{-(v_0+p+1)/2} \times \exp[-\operatorname{tr}(\boldsymbol{S}_0 \Sigma^{-1})/2] \right) \times \left(|\Sigma|^{-n/2} \exp\left(-\frac{1}{2} \operatorname{tr}(\boldsymbol{S}_\theta \Sigma^{-}1) \right) \right) \\ &= |\Sigma|^{-(v_0+p+1)/2} \exp(-\operatorname{tr}([\boldsymbol{S}_0 + \boldsymbol{S}_\theta] \Sigma^{-1})/2) \end{aligned}$$

Therefore, the posterior distribution for variance when the mean is known for a multivariate normal model is an inverse-Wishart distribution.

$$p(\boldsymbol{\theta} | \boldsymbol{y}_1, ..., \boldsymbol{y}_n, \Sigma) \;\; \text{inverse-Wishart}(v_0 + n, [\boldsymbol{S}_0 + \boldsymbol{S}_\theta]^{-1})$$

6.3 Gibbs Sampling When Means and Covariance are Unknown

As done above with univariate normal Bayesian inference, we will use Gibbs sampling when both the means and covariance are unknown as there is no closed form solution to this problem. We will use a multivariate normal distribution for estimating the means and an inverse-Wishart distribution for estimating the covariance matrix.

To start pick an arbitrary starting value for each parameter. In this case, we will use the sample means $\boldsymbol{\mu}_0$ and sample covariance matrix $\boldsymbol{S}_0$ as the starting values. $\boldsymbol{\theta}$ will hold the means, and Σ will hold the covariance matrices. Note that $v_n = n_0 + n$ and $\boldsymbol{S}_n = \boldsymbol{S}_0 + \boldsymbol{S}_\theta$.

$$\boldsymbol{\theta}^{(1)} = (\theta_1, \theta_2, ..., \theta_p)$$

$$\Sigma^{(1)} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2}^2 & \cdots & \sigma_{1,n}^2 \\ \sigma_{2,1}^2 & \sigma_2^2 & \cdots & \sigma_{2,n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n,1}^2 & \sigma_{n,2}^2 & \cdots & \sigma_n^2 \end{pmatrix}$$

Next, sample $\boldsymbol{\theta}^{(2)}$:

1. compute $\boldsymbol{\mu}_n$ and Λ_n from $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ and $\Sigma^{(1)}$
2. sample $\boldsymbol{\theta}^{(2)}$ from rmvnorm$(\boldsymbol{\mu}_n, \Lambda_n)$

Then, sample $\Sigma^{(2)}$:

1. compute $\boldsymbol{S}_n$ from $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ and $\boldsymbol{\theta}^{(2)}$
2. sample $\Sigma^{(2)}$ from rwish$(v_n, \boldsymbol{S}_n^{-1})$

Repeat these steps until the gibbs sampler converges which typically takes a few thousand loops. Then, the average of the $\boldsymbol{\theta}$ and Σ values will be the average for each parameter. Quantiles can be used to estimate the variance for each of the posterior distributions for each variables.

6.4 Example

Marist College is testing the effectiveness of it's statistics cariculum by administering a pretest at the beginning of an intro to statistics course and a posttest at the end of the course. Is there a difference between the two sets of test scores and does the student's post-test score correlate with their pre-test score? Here is the data from the 22 studemts in the class.

Pre-test	59	43	34	32	42	38	55	67	64	45	49	72	34	70	34	50	41	52	60
Post-test	77	39	46	26	38	43	68	86	77	60	50	59	38	48	55	58	54	60	75

Just from eyeballing the data it appears that the post-test is higher than the pre-test and there does appear to be some correlation between the two variables. To find out let's use a normal prior for the mean and an inverse-Wishart for the covariance matrix with a normal sampling distribution for our Gibbs sampler. After running for 5000 iterations here are the results:

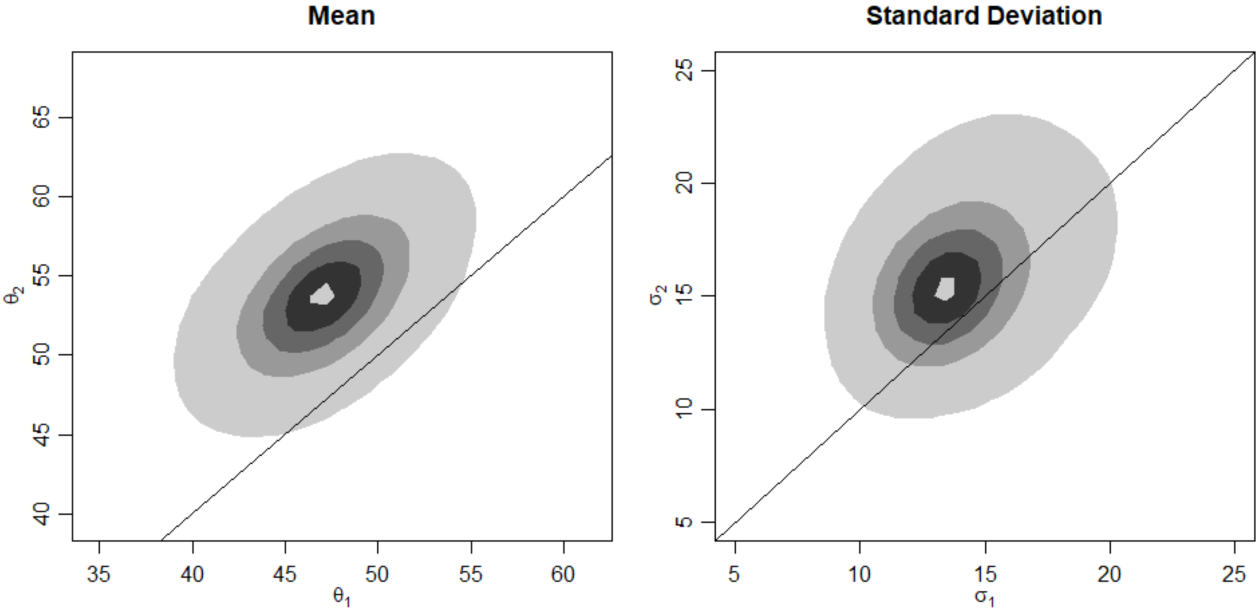


Figure 9: Theta, Sigma, and Correlation plots for the posterior distribution approximation

The mean for $\theta_2 - \theta_1$ is 6.689 with a 95% credible interval of [1.141, 11.772]. This shows that there is strong evidence that students performed better on the post-test than the pre-test. This means that the intro to statistics course curriculum is effective at teaching students.

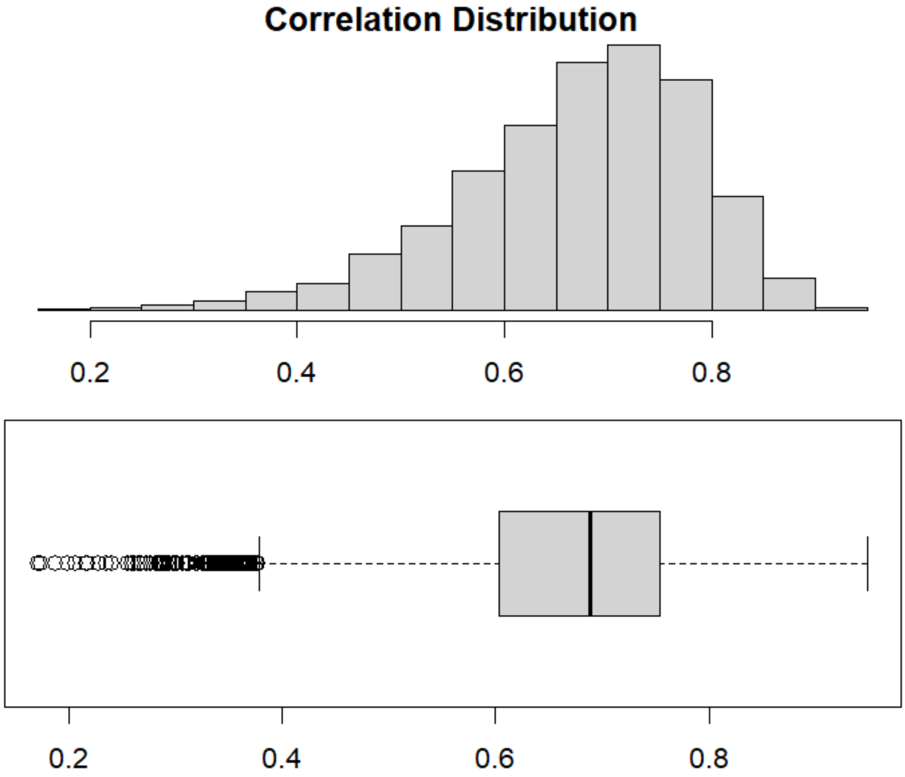


Figure 10: Covariance Plots

The correlation mean is .689 with a 95% credible interval of [0.398, 0.848]. This means that it is likely that there is a correlation with very strong evidence to support that as 0 isn't in the credible interval. This means that the higher a student's base knowledge of statistics, the better they tend to do on the post-test. This in conjunction means that while the course is effective, a student's background knowledge is also an important factor in how well they will succeed in staistics courses.

7 Bayesian Data Imputation

7.1 Types of Missing Values

Collecting data is often a messy process. Incomplete datasets are very common when working with real world data. The data could be missing due to wide variety of reasons such as a page on a survey being easy to skip, combing datasets with different variables, or data loss dut to storage corruption. There are four types of missing data: missing completely at random, missing at random, missing not at random, and structurally missing.

Missing completely at random means that the missing data points don't follow a pattern and are completely independent from the variables. For example, this could occur naturally if each subject is given a random set of survey questions to complete. This would cause data to be missing completely at random independent of how the other questions were answered. However, it is usually not the case that data are missing completely at random, as there is usually an underlying pattern that would make certain variables more or less likely to be missing.

The next type of missing data is missing at random. This is when there is a pattern to the missing data, but the missing values can be predicted from the existing data. This means that the missing data occurs with a wide range of existing values in the same row. For example, a sensor may have malfuncitioned for several minutes causing there to be a gap in the data. By using the readings of the other sensor and previous data the missed readings can be accurately predicted.

Missing not at random is similar to missing at random except that the missing data cannot be accurately predicted. This means that there is a lack of data from key subgroups in the data. For example, a survey might ask the income of the subject. Those with low income may be less likely to answer causing the average income to appear far higher than it actually is.

The final type of missing data is structurally missing data. This is data that is missing for a reason done on purpose by the researcher. For example, the income from employment for people without jobs would be null. Data engineering can also cause missing data. For example, when working with time series data there may be dates that don't have any data causing the appearance of 'holes' in the data.

7.2 Data Imputation

There are several methods to deal with the missing values when they are missing completely at random or missing at random. Listwise deletion is a rudimentary imputation method that involves deleting every row with a missing value. It is a poor imputation method because the deleted rows contain lots of information that is now lost. Becuase of this, listwise deletion is almost never the best solution for dealing with missing data. Mean/-median imputation is slightly better than listwise deletion, but it still has significant drawbacks. It involves imputing the missing values with the mean or median of the column. This retains the missing information by not deleting rows like listwise deletion, but it doesn't take into account the relationship between the variables. This causes the confidence interval for predictions to be much wider than they would be with a better imputation method. Mean/median imputation is very simple way to analyze data, but there are more sophisticated options better suited for more in depth analysis.

A better approach is to use a Bayesian method to impute the data. The Bayesian approach uses gibbs sampling to take the mean of the column and adjust it based on the other values in the row and the relationship between the variables. To start, let $O_i = (O_1, ..., O_p)^T$ be a binary list populated with zeros and ones such that

$O_{i,j} = 1$ means that the value $Y_{i,j}$ is present and $O_{i,j} = 0$ means that the value $Y_{i,j}$ is missing. The sampling probability for the data for subject i is then as follows:

$$\begin{aligned} p(\boldsymbol{o}_i, y_{i,j} : o_{i,j} = 1|\boldsymbol{\theta}, \Sigma) &= p(\boldsymbol{o}_i) \times p(y_{i,j} : o_{i,j} = 1|\boldsymbol{\theta}, \Sigma) \\ &= p(\boldsymbol{o}_i) \times \int \{p(y_{i,1}, ..., y_{i,p}|\boldsymbol{\theta}, \Sigma) \prod_{y_{i,j}:o_{i,j}=0} dy_{i,j}\} \end{aligned}$$

This means that the sampling probability for data for subject i is $p(\boldsymbol{o}_i)$ times the marginal probability of the existing variables after integrating out the unobserved values. When using a multivariate normal model for bayesian inference, $\int \{p(y_{i,1}, ..., y_{i,p}|\boldsymbol{\theta}, \Sigma) \prod_{y_{i,j}:o_{i,j}=0} dy_{i,j}$ has no closed form solution. Therefore, we will use gibbs sampling to find estimates for the missing values. Let $\boldsymbol{Y}$ be an $n \times p$ matrix that holds all of the data obtained and $\boldsymbol{O}$ be an $n \times p$ matrix in which $o_{i,j} = 1$ if $Y_{i,j}$ is observed and $o_{i,j} = 0$ if $Y_{i,j}$ is not observed. We will then split the matrix $\boldsymbol{Y}$ into two parts:

$$\begin{aligned} Y_{obs} &= \{y_{i,j} : o_{i,j} = 1\} \\ Y_{miss} &= \{y_{i,j} : o_{i,j} = 0\} \end{aligned}$$

From our observed data we want to obtain $p(\boldsymbol{\theta}, \Sigma, \boldsymbol{Y}_{miss}|\boldsymbol{Y}_{obs})$ which is the posterior distribution of the dataset and the missing values. A gibbs sampler can be used to approximate this distribution and the missing values by adding a step to the previous section's gibbs sampler.

1. sample $\boldsymbol{\theta}^{(s+1)}$ from $p(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}^{(s)}_{miss}, \Sigma^{(s)})$
2. sample $\Sigma^{(s+1)}$ from $p(\Sigma|\boldsymbol{Y}_{obs}, \boldsymbol{Y}^{(s)}_{miss}, \boldsymbol{\theta}^{(s+1)})$
3. sample $\boldsymbol{Y}^{(s+1)}_{miss}$ from $p(\boldsymbol{Y}_{miss}|\boldsymbol{Y}_{obs}, \boldsymbol{\theta}^{(s+1)}, \Sigma^{(s+1)})$

For steps one and two, $\boldsymbol{Y}_obs$ and $\boldsymbol{Y}_miss$ combine to form $\boldsymbol{Y}$. This means that these steps can be completed in the same way as done in the previous section where $\boldsymbol{\theta}$ is sampled from a multivariate norma distribution and Σ is sampled from an inverse-Wishart distribution. Step 3 is sampled as follows:

$$\begin{aligned} p(\boldsymbol{Y}_{miss}|\boldsymbol{Y}_{obs}, \boldsymbol{\theta}, \Sigma) &\propto p(\boldsymbol{Y}_{miss}, \boldsymbol{Y}_{obs}|\boldsymbol{\theta}, \Sigma) \\ &= \prod_{i=1}^{n} p(\boldsymbol{y}_{i,miss}, \boldsymbol{y}_{i,obs}|\boldsymbol{\theta}, \Sigma) \\ &\propto \prod_{i=1}^{n} p(\boldsymbol{y}_{i,miss}, \boldsymbol{y}_{i,obs}, \boldsymbol{\theta}, \Sigma) \end{aligned}$$

For each i we need to sample the missing values conditionally to the observed elements in the row.

$$\{\boldsymbol{y}_{i,miss}, \boldsymbol{y}_{i,obs}, \boldsymbol{\theta}, \Sigma\} \ \text{multivariatenormal}(\boldsymbol{\theta}_{miss|obs}, \Sigma_{miss|obs})$$

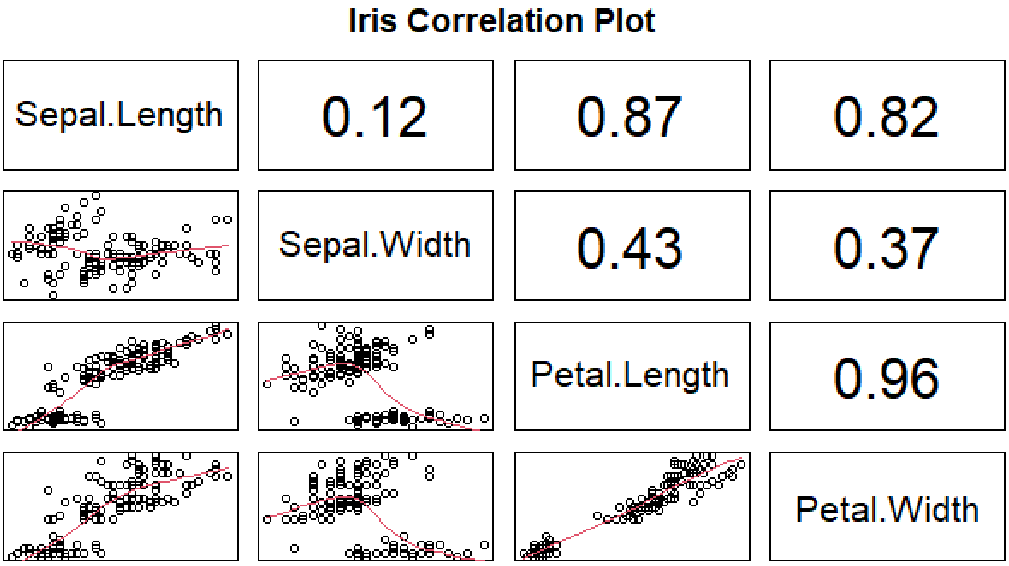
where

$$\begin{aligned} \boldsymbol{\theta}_{miss|obs} &= \boldsymbol{\theta}_{[miss]} + \Sigma_{[miss,obs]}(\Sigma_{[obs,obs]})^{-1}(\boldsymbol{y}_{[obs]} - \boldsymbol{\theta}_{[obs]}) \\ \Sigma_{miss|obs} &= \Sigma_{[miss,miss]} - \Sigma_{[miss,obs]}(\Sigma_{[obs,obs]})^{-1}\Sigma_{[obs,miss]} \end{aligned}$$

In the formulas above, we can see that the missing values are calculated by taking the unconditional mean and then modifying it by looking at the observed values. Similarly, the missing data covariance matrix is equal to the unconditional covariance matrix with a little bit subtracted from it. This makes sense, as having more information about the variables should decrease our uncertainty of their actual values.

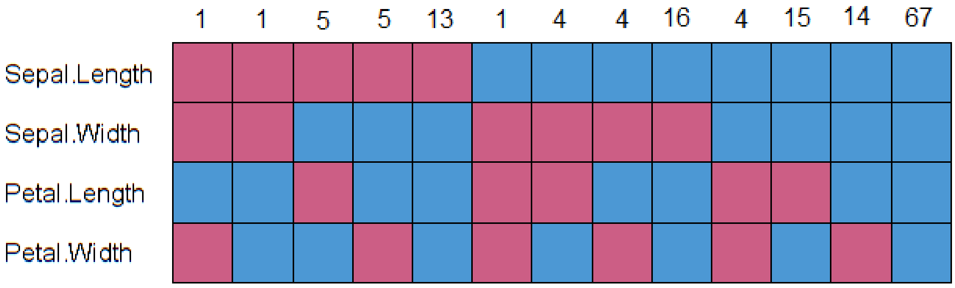
7.3 Example: Iris Flowers

The dataset that will be imputed is the Iris flower dataset. The iris dataset is a 150 iris flower measurements done by Ronald Fisher in his 1936 paper ”The use of multiple measurements in taxonomic problems”. There are three flowers measured: Iris setosa, Iris virginica, and Iris versicolor. Each flower has four measurements: sepal length, sepal width, petal length, and petal width. From these four measurements it is possible to create a highly accurate model for predicting the type of iris flower it is. This is a complete dataset with no missing values. This allows us to delete values at random and then compare the imputed values to the original to gauge the effectiveness of the imputation.



As can be seen in the above figure, the Sepal Length, Petal Length, and Petal Width values are all highly correlated. This should make it easy to impute these values. The Sepal Width value is less correlated with the others, so it will likely be more difficult to impute.

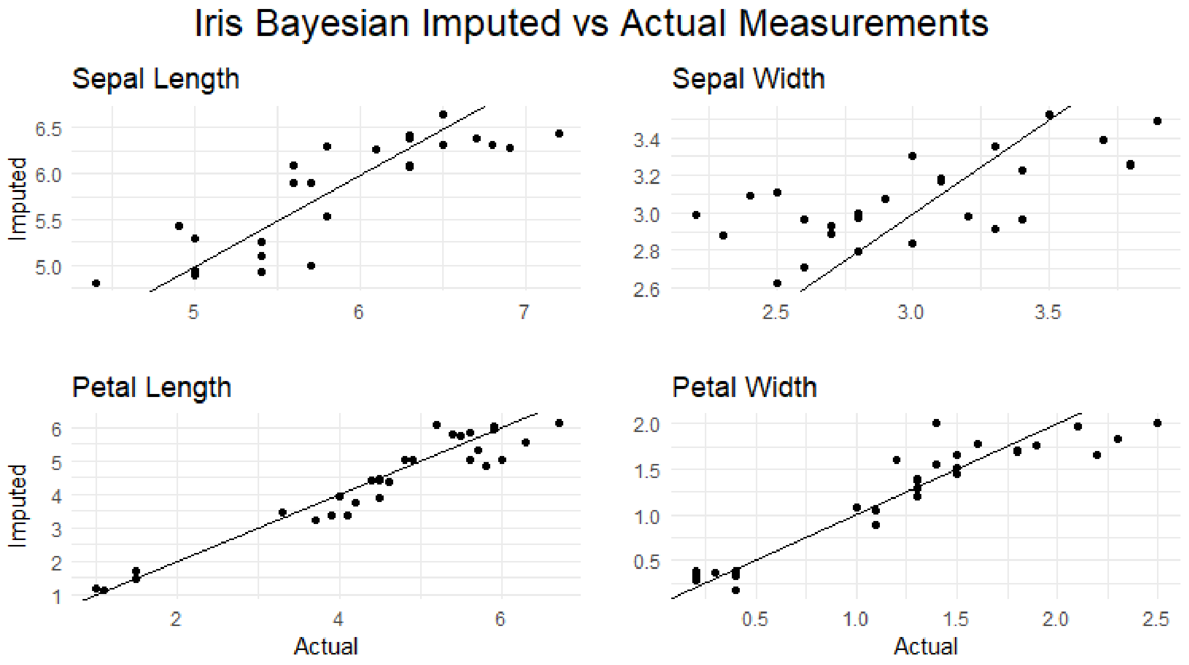
Now to test the imputation values must be deleted from the dataset. This was be done in a way to create data that is missing completely at random. 20% of the values in each column were deleted randomly giving the following pattern of missing data.



	1	1	5	5	13	1	4	4	16	4	15	14	67
Sepal.Length	🟪	🟪	🟪	🟪	🟪	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
Sepal.Width	🟪	🟪	🟦	🟦	🟦	🟪	🟪	🟪	🟪	🟦	🟦	🟦	🟦
Petal.Length	🟦	🟦	🟪	🟦	🟦	🟪	🟪	🟦	🟦	🟪	🟪	🟦	🟦
Petal.Width	🟪	🟦	🟦	🟪	🟦	🟪	🟦	🟪	🟦	🟪	🟦	🟪	🟦

Figure 11: Here the purple represents values that are missing and blue represents values that are not missing. The number on top is the number of rows that follow that pattern for missing data.

Gibbs sampling was then used to implement Bayesian imputation on the missing data. The results are as follows:



The three methods for dealing will missing data mentioned above (listwise deletion, mean imputation, and Bayesian imputation) were then compared against the original dataset. The methods for comparison were mean absolute deviation of the imputed values from the original values and the accuracy of the logistic regression and random forest models trained on the datasets.

Mean Absolute Deviation of the Imputed Data

Imputation Method	Sepal Length	Sepal Width	Petal Length	Petal Width
Mean	0.59	0.40	1.56	0.55
Bayesian	0.32	0.30	0.36	0.17

Accuracy of Logistic Regression and Random Forest Models

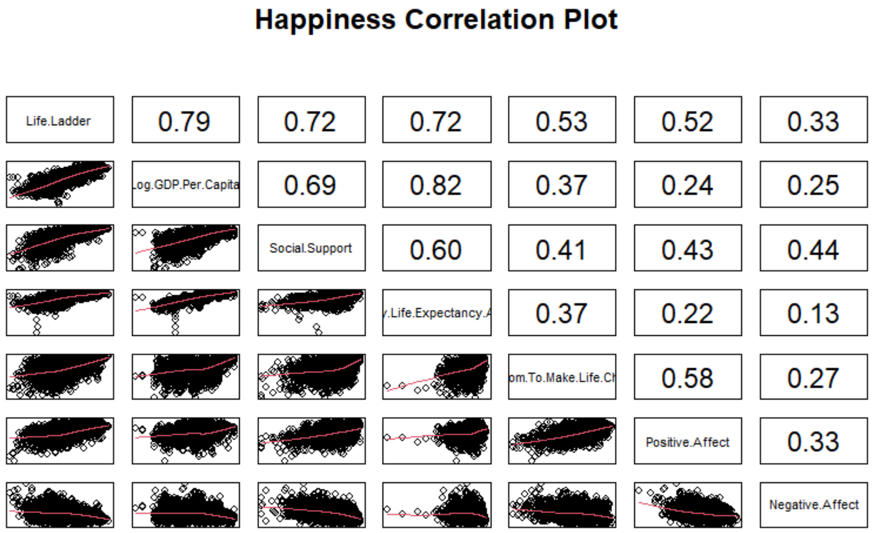
Model	Original Data	Listwise Deletion	Mean Imputation	Bayesian Imputation
Logistic Regression	0.96	0.89	0.87	0.94
Random Forest	0.95	0.83	0.91	0.95

The results above show that the Bayesian imputation method is by far the best approach to dealing with missing data. The Bayesian dataset has almost the same predictive power as the original dataset. The stronger the correlation between the variables, the better Bayesian imputation performs compared to the other methods. This can be seen in the variable with the lowest correlation, sepal width, for which Bayesian imputation is only marginally better than mean imputation. The dataset that was imputed using a Bayesian method has almost the same predictive power of the original dataset. The other methods fall far behind in predictive accuracy.

7.4 Example: World Happiness Report

The next dataset that will be imputed is the World Happiness Report dataset. The world happiness report is a study on the state of world happiness done by the Gallup World Poll from 2005 to 2020. The dataset has 12 features, but the subset we will be focusing on is Life Ladder, Log GDP per Capita, Social Support, Healthy

Life Expectancy at Birth, Positive Affect, and Negative Effect. The dataset has a few missing values, so in order to test the effectiveness of the data imputation, the rows with missing values were removed.



As can be seen in the above figure, there is a wide spread of correlation between the values with most correlations being .3 to .6. This may negatively affect the Bayesian imputation because it is dependent on correlation between the values in order to impute values.

Now to test the imputation values must be deleted from the dataset. This was be done in a way to create data that is missing completely at random. 30% of the values in each column were deleted randomly giving the following pattern of missing data. Afterwards, only about 25% or about 500 of the rows remained complete.

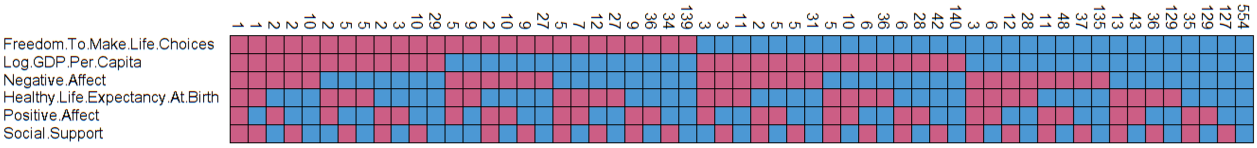
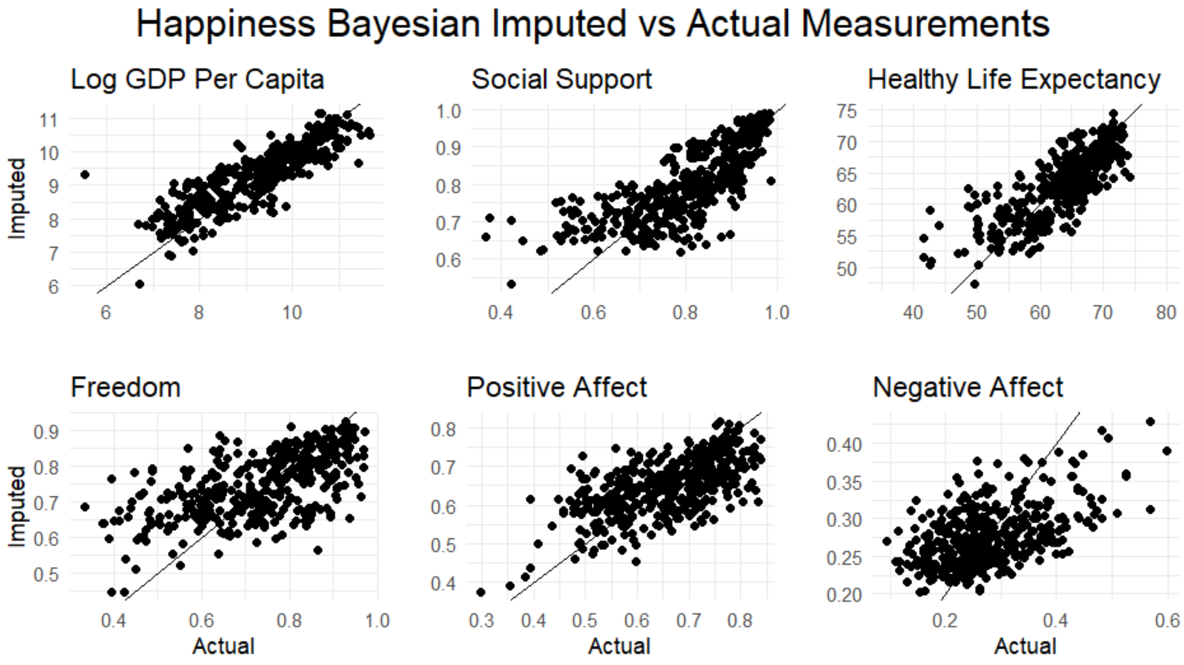


Figure 12: Here the purple represents values that are missing and blue represents values that are not missing. The number on top is the number of rows that follow that pattern for missing data.

Gibbs sampling was then used to implement Bayesian imputation on the missing data. The results are as follows:



Log GDP per Capita and Health Life Expectancy appear to be the best imputations. Social support and Positive Affect appear to be decent predictions as well. Freedom and Negative affect, however, seem to be lacking compared to the others. They all appear to be much better predictions than just using a mean imputation, though.

The three methods for dealing will missing data mentioned above (listwise deletion, mean imputation, and Bayesian imputation) were then compared against the original dataset. The methods for comparison were mean absolute deviation of the imputed values from the original values and the R^2 of the multiple linear regression and random forest models trained on the datasets.

Mean Absolute Deviation of the Imputed Data

Imputation Method	Log GDP	Social Support	Healthy Life Expectancy	Freedom	Positive Affect	Negative Affect
Mean	1.01	0.10	5.33	0.12	0.08	0.07
Bayesian	0.42	0.06	2.89	0.09	0.06	0.06

R^2 of Logistic Regression and Random Forest Models

Model	Original Data	Listwise Deletion	Mean Imputation	Bayesian Imputation
Linear Regression	0.76	0.76	0.70	0.81
Random Forest	0.86	0.82	0.79	0.86

The results above show that the Bayesian imputation method is by far the best approach to dealing with missing data. The Bayesian dataset has almost the same predictive power as the original dataset. The stronger the correlation between the variables, the better Bayesian imputation performs compared to the other methods. This can be seen in the variables with the lowest correlation, Negative Affect and Positive effect, for which Bayesian imputation is only marginally better than mean imputation. The dataset that was imputed using a Bayesian method has almost the same predictive power of the original dataset. The other methods fall far behind in predictive power.

8 Conclusion

Datasets are often incomplete due to a variety of reasons. In order to utilize all of the existing data, imputation methods are needed. As seen above, Bayesian imputation offers far better results than the traditional methods of listwise deletion and mean/median imputation. Bayesian imputation is an essential tool to working with incomplete datasets.

There are a few downsides to Bayesian such as a slow run time and the need for variable correlation, but for most cases the benefits far outweigh the costs. The two previous examples offer strong evidence that Bayesian imputation is very powerful and useful for increasing the predictive power of an incomplete dataset.

All code can be found at `https://github.com/Will-Holt60/Bayesian_Imputation`

9 Bibliography

References

Hoff, P. D.(2009). A first course in Bayesian statistical methods (Vol. 580). New York: Springer.

Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). Bayesian data analysis. CRC press.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, vol. 7, no. 2, pp. 179–188.

James, G. (2021). An Introduction to Statistical Learning: With Applications in R. Springer.