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 9, 2023

Abstract

Datasets often have many missing values. There are many ways of imputing values in a data set 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

Bayesian inference is a powerful and flexible statistical approach that allows for the incorporation of prior knowledge and beliefs into the analysis of data James et al. (2013); Kruschke (2014); Gelman et al. (2013); Hoff (2009). One of the major advantages of Bayesian inference is its ability to handle complex and high-dimensional data. Bayesian inference refers to the application of Bayesian methods to data sets with a large number of features or variables. This type of analysis is becoming increasingly important as data sets in many fields, such as genomics, neuroscience, and machine learning, continue to grow in size and complexity.

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 (see Jones (1996)) 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 (see Van Buuren (2012)) 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. In this thesis, we consider the problem of data imputation using a Bayesian approach. As mentioned previously, listwise deletion and mean imputation both have drawbacks. The goal of a Bayesian approach proposed in this paper is to overcome these drawbacks. 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 these variables. Through numerical examples provided in this thesis, it is showed that the Bayesian imputation method is by far the best approach to dealing with missing data. The Bayesian imputed data set has almost the same predictive power as the original data set. The stronger the correlation between the variables, the better Bayesian imputation performs compared to the other methods.

The rest of the thesis is organized as follows: In Sections 2, 3 we recall some basic definitions of probability and statistics needed for later sections. Section 4 recalls the key concepts of Bayesian inference. Section 3 provides the preliminary work needed for Section 6. Section 6 presents the main results of the thesis where both theoretical foundations and numerical examples are provided. Section 7 concludes the thesis.

2 Random Variables

In this section, we recall a few basic concepts of probability and statistics including probability density function, expected value (mean), and variance of a random variable with particular attentions paying for normal random variables.

2.1 Continuous Random Variables

Continuous random variables can take on an infinite number of values in 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. $p(x) \geq 0, \quad \forall x \in \mathbb{R}$
2. $\displaystyle\int_{-\infty}^{\infty} p(x)dx = 1.$

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

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

For example, one of the most famous example is the normal random variable. In this case the PDF for a normal distribution is $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$. 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 **R**. This will be explored further in the next section. For a normal random variable with parameters μ and σ, it is often denoted by $N(\mu, \sigma^2)$. In the Figure 1, we plot the graph of $p(x)$ with $\mu = 0$ and $\sigma = 1$.

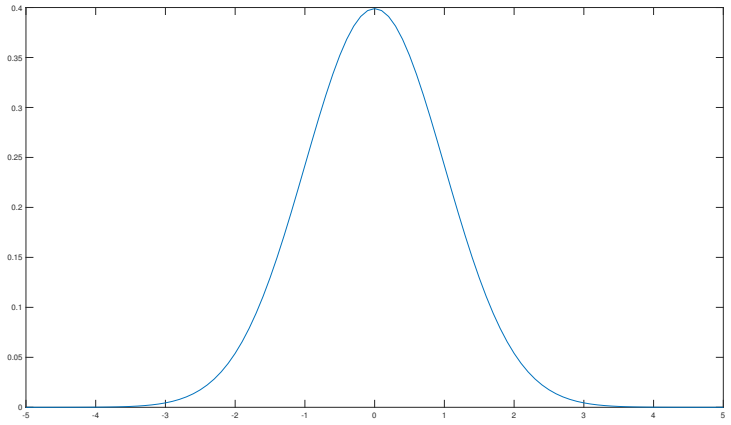

Figure 1: Graph of $N(\mu, \sigma^2)$ with $\mu = 0, \sigma = 1$

Normal distributions are considered to be the most important distribution in statistics and are often used in the natural and social sciences to represent real-valued random variables whose distributions are not known. Their importance is partly due to the central limit theorem, which will be considered in Theorem 4.

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:

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

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

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

The expected value of a continuous random variable is defined as

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

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 Durrett (2019)) *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:

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

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 $z = \dfrac{x-\mu}{\sqrt{2}\sigma}$, then it can be seen that

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

□

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 $p(x)$ is defined as

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

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 Durrett (2019)) *Let X be a random variable following a normal distribution: $X \sim N(\mu, \sigma^2)$. Then, the variance of X is* $\operatorname{var}(X) = \sigma^2$.

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

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

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

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

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

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

□

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

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

Theorem 3 (Durrett (2019)) *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,

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

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 Durrett (2019)) *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*

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

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

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

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*

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

3 Multivariate Normal Distribution

In probability theory and statistics, the multivariate normal distribution, multivariate Gaussian distribution is a generalization of the one-dimensional (univariate) normal distribution to higher dimensions. It is often used to model high dimensional data sets. We will start by recall some basic properties of multivariate normal distribution including its probability density function, maximum likelihood estimations and its conditional distribution. It will be used extensively in later sections.

3.1 Probability Density Function

Let $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ be a p-dimensional random vector. Also, let $\mu = (\mu_1, \mu_2, \ldots, \mu_p)^T \in \mathbb{R}^p$ and $\Sigma = (\sigma_{ij})_{p\times p} \in \mathbb{R}^{p\times p}$ be a non-singular matrix. It is said that $\mathbf{X}$ has a multivariate-normal distribution with the mean μ and covariance matrix Σ if the probability density function of X is given by

$$p(\mathbf{x}|\mu, \Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right).$$

Note that in the above equation, Σ^{-1} denotes the matrix inverse of Σ. That is $\Sigma\Sigma^{-1} = \mathbf{I}_{p\times p} = \Sigma^{-1}\Sigma$. Also, $|\Sigma|$ denotes the determinant of the matrix Σ. When $\mathbf{X}$ has a multivariate-normal distribution with the mean μ and covariance matrix Σ, we will denote it as

$$\mathbf{X} \sim \text{multivariate-normal}(\mu, \Sigma).$$

In Figure 2, we plot the probability density function of $\mathbf{X} = (X_1, X_2)$ with $\mu = (0, 0)^T$ and $\Sigma = \begin{pmatrix} 0.25 & 0.3 \\ 0.3 & 1 \end{pmatrix}$.

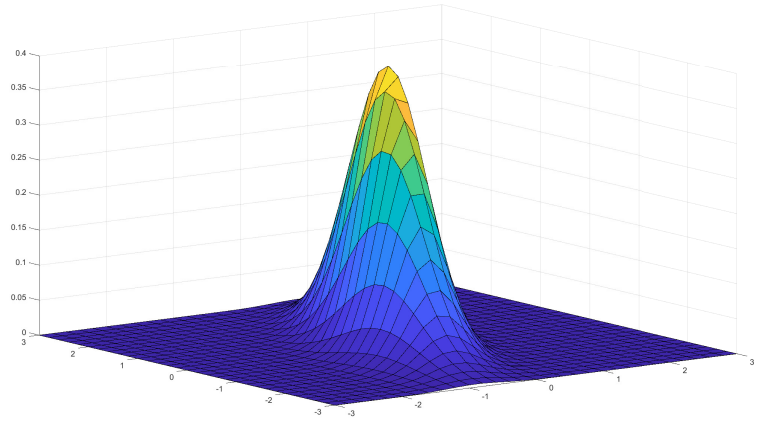

Figure 2: An image of a multivariate normal distribution

In general, it can be showed that $\mu = (E(X_1), E(X_2), \ldots, E(X_p))^T$ and

$$\Sigma = \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p,1} & \sigma_{p,2} & \cdots & \sigma_{p,p} \end{pmatrix},$$

where $\sigma_{i,j} = Cov(X_i, X_j) = E(X_iX_j) - E(X_i)E(X_j), i, j = 1, 2, \ldots, p$. From the covariance matrix Σ, we can define the correlation matrix $\rho = (\rho_{ij})_{p\times p} \in \mathbb{R}^{p\times p}$ where each entry ρ_{ij} is defined as $\rho_{ij} = \rho(X_i, X_j) = \frac{cov(X_i,X_j)}{\sqrt{Var(X_i)Var(X_j)}}$.

Figure 3 provides examples of a multivariate normal distribution $\mathbf{X} = (X_1, X_2)$ with different correlation values. 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.

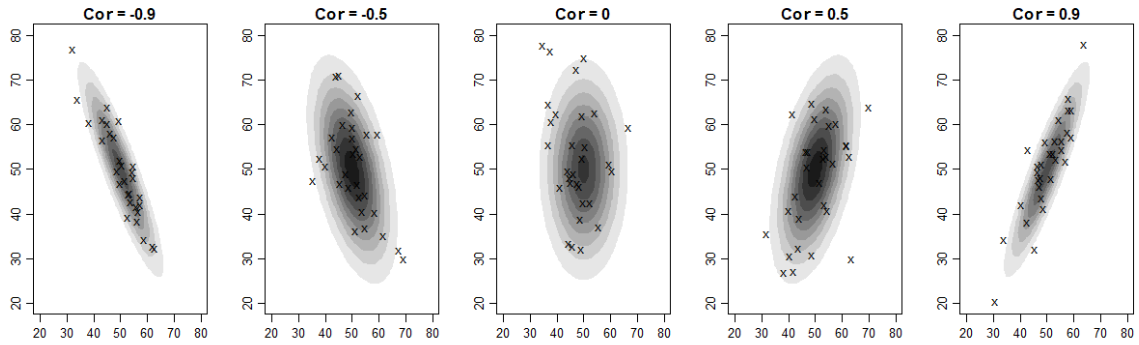

Figure 3: Multivariate normal distributions with 5 different correlation values

3.2 Maximum Likelihood Estimation

Assume that an iid sample $\{\mathbf{x}_i : i = 1, 2, \ldots, n\}$ is collected and $\mathbf{x}_i \sim \text{multivariate-normal}(\mu, \Sigma), i = 1, 2, \ldots, p$. We can use the maximum likelihood estimation (MLE) to estimate μ and Σ from the collected sample as follows.

Theorem 5 *If we have an iid sample* $\{\boldsymbol{x}_i \sim \text{multivariate-normal}(\mu, \Sigma) : i = 1, 2, \ldots, n\}$*, then the MLE for the parameters* μ *and* Σ *are, respectively, given by*

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \equiv \bar{\boldsymbol{x}},$$

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T.$$

Proof. First, the likelihood function can be found as

$$\begin{aligned} p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n | \mu, \Sigma) &= \prod_{i=1}^{n} (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left(-(\mathbf{x}_i - \mu)^T \Sigma^{-1} (\mathbf{x}_i - \mu)/2\right) \\ &= (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} (\mathbf{x}_i - \mu)^T \Sigma^{-1} (\mathbf{x}_i - \mu)\right). \end{aligned}$$

Let $\Lambda = \Sigma^{-1}$, then the log-likelihood function is given by

$$\log(\mu, \Sigma) = \log(p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n | \mu, \Sigma)) = \log((2\pi)^{-np/2}) + \frac{n}{2}\log|\Lambda| - \frac{1}{2}\sum_{i=1}^{n} (\mathbf{x}_i - \mu)^T \Lambda (\mathbf{x}_i - \mu).$$

Therefore,

$$\begin{aligned} \frac{\partial}{\partial \mu}\log(\mu, \Sigma) &= \frac{\partial}{\partial \mu}\left(-\frac{1}{2}\sum_{i=1}^{n} (\mathbf{x}_i - \mu)^T \Lambda (\mathbf{x}_i - \mu)\right) \\ &= -\frac{1}{2}\sum_{i=1}^{n} -2\Lambda(\mathbf{x}_i - \mu) \\ &= -\Lambda\sum_{i=1}^{n} (\mathbf{x}_i - \mu). \end{aligned}$$

Hence

$$\begin{aligned}
&\frac{\partial}{\partial \mu} \log(\mu, \Sigma) = 0 \\
&\Leftrightarrow \sum_{i=1}^{n} (\mathbf{x}_i - \mu) = 0 \\
&\Leftrightarrow \sum_{i=1}^{n} \mathbf{x}_i - \sum_{i=1}^{n} \mu = 0 \\
&\Leftrightarrow \sum_{i=1}^{n} \mathbf{x}_i - n\mu = 0 \\
&\Leftrightarrow \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i = \mu.
\end{aligned}$$

Next, we recall the following property of trace. For matrices of appropriate sizes A, B, C it is well-known that

$$tr(ABC) = tr(CAB) = tr(BCA).$$

Also, for any vector $\mathbf{x}$ and matrix A, we have

$$\mathbf{R} \ni \mathbf{x}^T A \mathbf{x} = tr(\mathbf{x}^T A \mathbf{x}) = tr(\mathbf{x}\mathbf{x}^T A) = tr(A\mathbf{x}\mathbf{x}^T).$$

As a result, we can rewrite the log-likelihood function as

$$\begin{aligned}
\log(\mu, \Sigma) &= \log((2\pi)^{-np/2}) + \frac{n}{2} \log|\Lambda| - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \mu)^T \Lambda (\mathbf{x}_i - \mu) \\
&= \log((2\pi)^{-np/2}) + \frac{n}{2} \log|\Lambda| - \frac{1}{2} \sum_{i=1}^{n} tr[(\mathbf{x}_i - \mu)^T \Lambda (\mathbf{x}_i - \mu)] \\
&= \log((2\pi)^{-np/2}) + \frac{n}{2} \log|\Lambda| - \frac{1}{2} \sum_{i=1}^{n} tr[(\mathbf{x}_i - \mu)^T (\mathbf{x}_i - \mu)\Lambda] \\
&= \log((2\pi)^{-np/2}) + \frac{n}{2} \log|\Lambda| - \frac{1}{2} tr[\sum_{i=1}^{n} (\mathbf{x}_i - \mu)^T (\mathbf{x}_i - \mu)\Lambda] \\
&= \log((2\pi)^{-np/2}) + \frac{n}{2} \log|\Lambda| - \frac{1}{2} tr[S_\mu \Lambda],
\end{aligned}$$

where $S_\mu = \sum_{i=1}^{n} (\mathbf{x}_i - \mu)^T (\mathbf{x}_i - \mu)$. Therefore,

$$\begin{aligned}
&\frac{\partial}{\partial \Lambda} \log(\mu, \Sigma) = \frac{n}{2} \Lambda^{-T} - \frac{1}{2} S_\mu^T = 0 \\
&\Leftrightarrow n\Lambda^{-1} - S_\mu = 0 \\
&\Leftrightarrow \Lambda^{-1} - \frac{1}{n} S_\mu = 0 \\
&\Leftrightarrow \Sigma = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \mu)^T (\mathbf{x}_i - \mu) = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})^T (\mathbf{x}_i - \bar{\mathbf{x}}).
\end{aligned}$$

This completes the proof of the theorem.

□

We note that the MLE estimates for μ and Σ will be useful later when one needs to initialize the Gibbs sampling algorithm. Please see Section 6 for more details.

3.3 Conditional Distribution

In this section, we are interested in deriving the closed functional form of a normal random variable conditioned on an another normal random variable. This will provide the key foundation for a Bayesian data imputation developed in Section 6. To this end, assume that $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{R}^{d_1 \times d_2}$ is jointly normally distributed $\text{multivariate-normal}(\mu, \Sigma)$ with parameters

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}.$$

We are interested in the conditional distribution $p(\mathbf{x}_1|\mathbf{x}_2)$. We have the following theorem regarding the closed form distribution of $p(\mathbf{x}_1|\mathbf{x}_2)$when $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ is normally distributed.

Theorem 6 *The conditional distribution $p(\boldsymbol{x}_1|\boldsymbol{x}_2)$ is given by*

$$p(\boldsymbol{x}_1|\boldsymbol{x}_2) \sim \text{multivariate-normal}(\mu_{1|2}, \Sigma_{1|2})$$

where

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \mu_2), \quad \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Proof. Recall that the joint probability density function of $(\mathbf{X}_1, \mathbf{X}_2)$ can be written as

$$p(\mathbf{x}_1, \mathbf{x}_2|\mu, \Sigma) \propto \exp\left(-\frac{1}{2}\begin{pmatrix} \mathbf{x}_1 - \mu_2 \\ \mathbf{x}_2 - \mu_2 \end{pmatrix}^T \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{x}_1 - \mu_2 \\ \mathbf{x}_2 - \mu_2 \end{pmatrix}\right).$$

The matrix inversion in the above equation can be factored as

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{pmatrix} \begin{pmatrix} (\Sigma/\Sigma_{22})^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix} \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix},$$

where $\Sigma/\Sigma_{22} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. As a result, we have

$$\begin{aligned} p(\mathbf{x}_1, \mathbf{x}_2|\mu, \Sigma) &\propto \exp\left(-\frac{1}{2}\begin{pmatrix} \mathbf{x}_1 - \mu_2 \\ \mathbf{x}_2 - \mu_2 \end{pmatrix}^T \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{x}_1 - \mu_2 \\ \mathbf{x}_2 - \mu_2 \end{pmatrix}\right) \\ &= \exp\left(-\frac{1}{2}(\mathbf{x}_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2))^T(\Sigma/\Sigma_{22})^{-1}(\mathbf{x}_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2))\right) \times \\ &\times \exp\left(-\frac{1}{2}(\mathbf{x}_2 - \mu_2)^T\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2)\right). \end{aligned}$$

Moreover, it can be seen that

$$\begin{aligned} (2\pi)^{(d_1+d_2)/2}|\Sigma|^{1/2} &= (2\pi)^{(d_1+d_2)/2}(|\Sigma/\Sigma_{22}||\Sigma_{22}|)^{1/2} \\ &= (2\pi)^{d_1/2}|\Sigma/\Sigma_{22}|^{1/2}(2\pi)^{d_2/2}|\Sigma_{22}|^{1/2}. \end{aligned}$$

Hence, we have

$$\begin{aligned} p(\mathbf{x}_1, \mathbf{x}_2|\mu, \Sigma) &= (2\pi)^{\frac{d_1}{2}}|\Sigma/\Sigma_{22}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2))^T(\Sigma/\Sigma_{22})^{-1}(\mathbf{x}_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2))\right) \\ &\times (2\pi)^{\frac{d_2}{2}}|\Sigma_{22}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}_2 - \mu_2)^T\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2)\right) \\ &= p(\mathbf{x}_1|\mathbf{x}_2)p(\mathbf{x}_2). \end{aligned}$$

Therefore, from the above expression, we can see that

$$\begin{aligned} p(\mathbf{x}_1, \mathbf{x}_2|\mu, \Sigma) &= (2\pi)^{\frac{d_1}{2}}|\Sigma/\Sigma_{22}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2))^T(\Sigma/\Sigma_{22})^{-1}(\mathbf{x}_1 - \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2))\right) \\ &\sim \text{multivariate-normal}(\mu_{1|2}, \Sigma_{1|2}), \end{aligned}$$

where

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2), \quad \Sigma_{1|2} = \Sigma/\Sigma_{22} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

This completes the proof of the theorem.

□

4 Bayesian Statistics

4.1 Bayes' Theorem

In probability theory and statistics, Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

Theorem 7 (Bayes' Theorem Hoff (2009)) *Let A and B be events where $P(B) \neq 0$. Then*

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

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.

□

REMARK 4.1 *In case the sets $A_1, A_2, \ldots, A_n$ form a partition for the sample space Ω and B is an event. Then we have*

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}, \forall i \in \{1, 2, \ldots, n\}.$$

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.

Bayesian Inference Example: Let's us provide an example to demonstrate the Bayesian rule. Suppose that 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 $Beta(\alpha = 4.8, \beta = 19.2)$. Bart chooses to use a uniform distribution of $y = 1$ for his prior. Finally, Chris uses a piecewise 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.

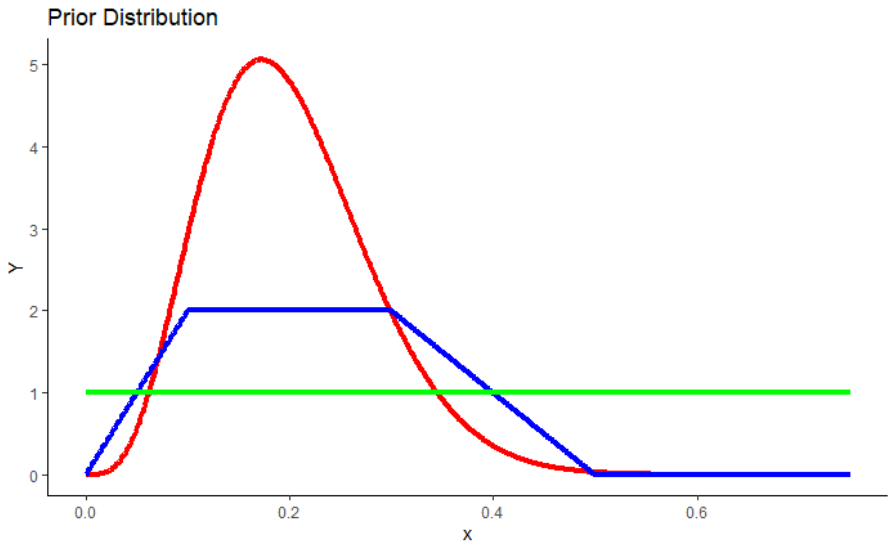

Figure 4: The prior distributions of the three students. The red line is Anna, The blue line is Chris, and the green line is Bart

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:

$$g(p) = \frac{1}{Beta(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1}$$
$$p(y|p) = \binom{n}{y} p^x (1-p)^{n-y}.$$

As a result, we can compute the posterior distribution as follows:

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

Here we can see that $p(p|x)$ 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.

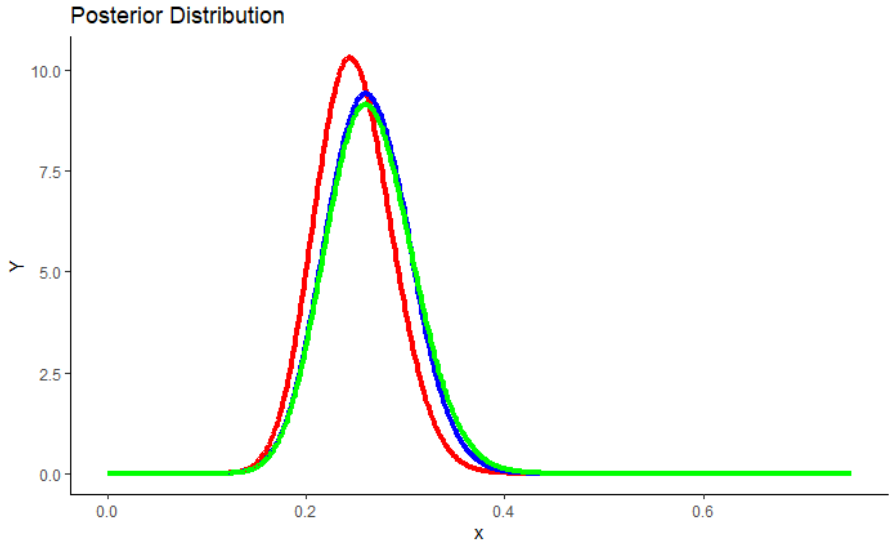

Figure 5: The posterior distributions of the three students. The red line is Anna, The blue line is Chris, and the green line is Bart

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 Multivariate Normal Bayesian Inference

In this section, it is assumed that the random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ has a multivariate normal distribution. That is, the probability density of X is given by

$$p(\mathbf{x}|\theta, \Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left(-(\mathbf{x}-\theta)^T\Sigma^{-1}(\mathbf{x}-\theta)/2\right), \tag{1}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T$, $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^T$, and the covariance matrix Σ

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}. \tag{2}$$

Given a realized sample $\{\mathbf{x}_i = (x_1, x_2, \ldots, x_p) : i = 1, 2, \ldots, n\}$ from the random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)$, our goal is to construct estimate for θ and Σ by taking the prior distribution knowledge of θ and Σ and the observed data into account. We will accomplish these goals through several steps which will be outlined in details in the following subsections.

5.1 The Posterior Conditional Distribution of θ

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:

$$p(\boldsymbol{\theta}) = \text{multivariate-normal}(\boldsymbol{\mu}_0, \Lambda_0).$$

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

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

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

where $\mathbf{A}_1 = n\Sigma^{-1}$, $\mathbf{b}_1 = n\Sigma^{-1}\bar{\boldsymbol{x}}$, and $\bar{\boldsymbol{x}} = \left(\frac{1}{n}\sum_{i=1}^{n}x_{i,1}, ..., \frac{1}{n}\sum_{i=1}^{n}x_{i,p}\right)^T$ denotes the sample average.

Combining the two previous equation we have as follows:

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

where

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

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

5.2 The Posterior Conditional Distribution of Σ

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:

$$\begin{aligned} p(\Sigma) = &\left[2^{v_0p/2}\pi^{\binom{p}{2}/2}|\boldsymbol{S}_0|^{-v_0/2}\dot{\prod}\,p_{j=1}\Gamma([v_0+1-j]/2)\right]^{-1}\times \\ &|\Sigma|^{-(v_0+p+1)/2}\times\exp[-\operatorname{tr}(\boldsymbol{S}_0\Sigma^{-1})/2] \end{aligned}$$

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

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

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:

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

where

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

The matrix $\boldsymbol{S}_\theta$ is the residual sum of squares for the vectors $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ conditional on $\boldsymbol{\theta}$. Therefore, from Theorem 5, $\frac{1}{n}\boldsymbol{S}_\theta$ 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:

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

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

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

5.3 Gibbs Sampling

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 $\boldsymbol{S}_n = \boldsymbol{S}_0 + \boldsymbol{S}_\theta$. Let

$$\begin{aligned}
\boldsymbol{\theta}^{(1)} &= (\theta_1, \theta_2, ..., \theta_p) \\
\Sigma^{(1)} &= \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p,1} & \sigma_{p,2} & \cdots & \sigma_{p,p} \end{pmatrix},
\end{aligned}$$

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

1. compute $\boldsymbol{\mu}_n$ and Λ_n from $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ 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 $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ 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.

Example: Marist College is testing the effectiveness of it's statistics curriculum by administering a pretest at the beginning of an intro to statistics course and a post-test 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 students 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:

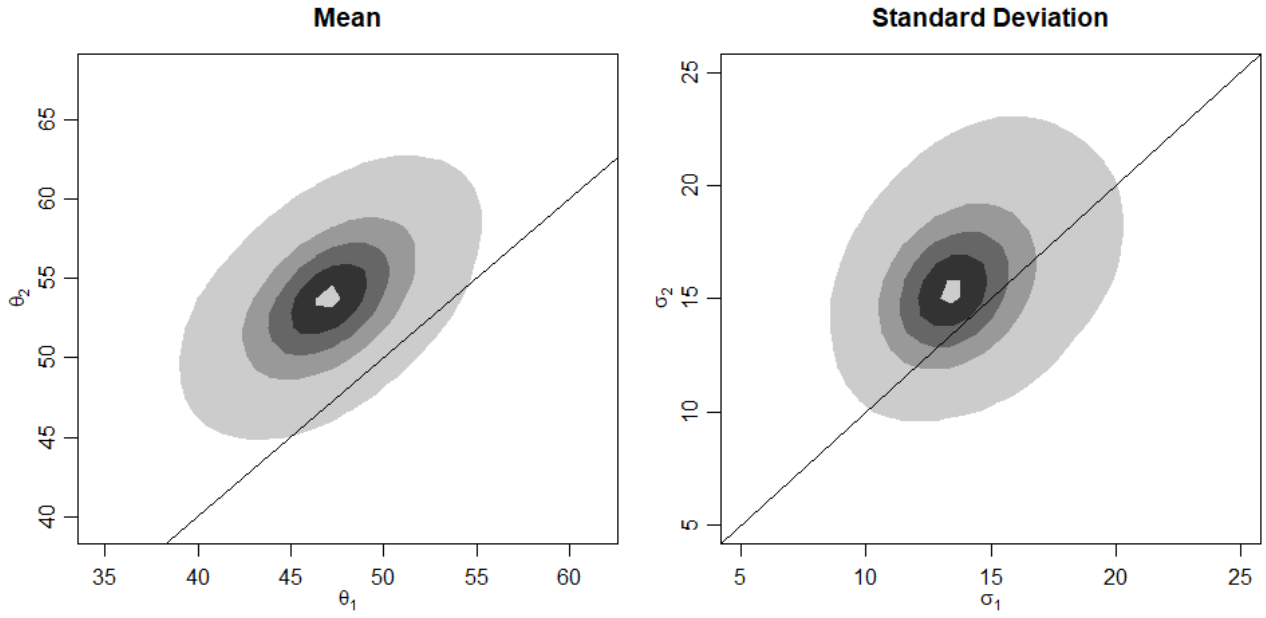

Figure 6: Theta, Sigma, and Correlation plots for the posterior distribution approximation

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.

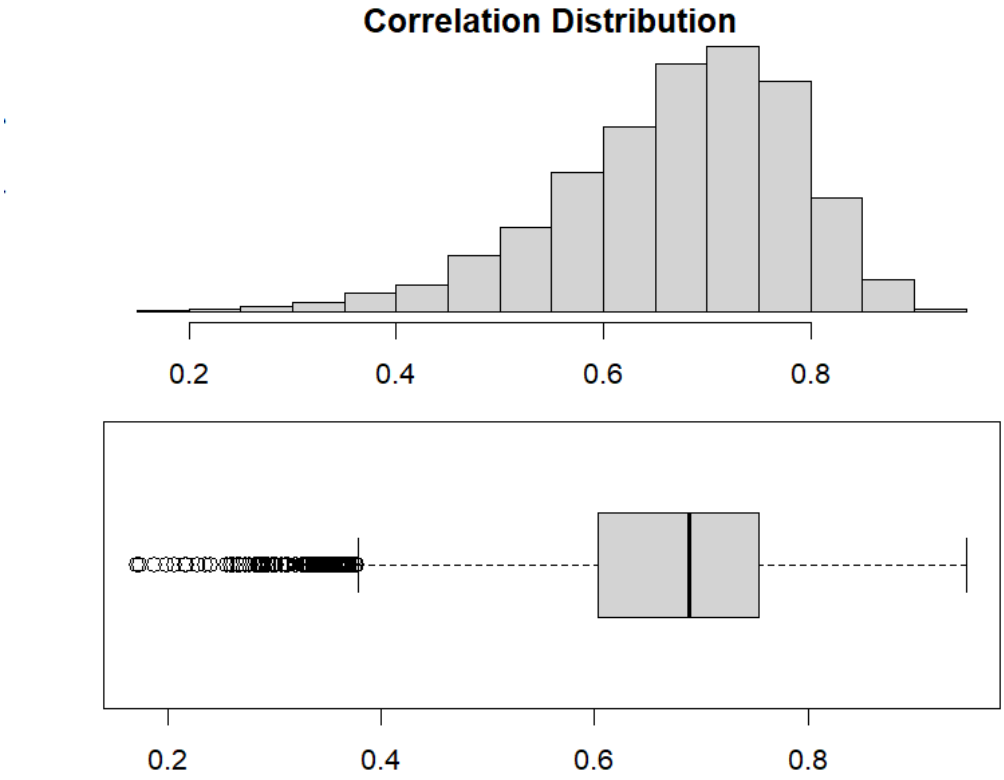

Figure 7: Covariance Plots

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.

6 Bayesian Data Imputation

6.1 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.

6.2 Data Imputation

In this section, assume that we can collect an iid data set $\{\mathbf{x}_i : i = 1, 2, \ldots, n\}$ such that $\mathbf{x}_i \sim \text{multivariate-normal}(\theta, \Sigma)$. However, it is also assumed that $x_{ij} = NA$ for some $i \in \{1, 2, \ldots, n\}, j \in \{1, 2, \ldots, p\}$. This is often the case for most of data set including medial data, economic data, etc. The major goal of this section is to "fill in" those missing values using a Bayesian approach. 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 $\boldsymbol{O}_i = (O_1, ..., O_p)^T$ be a binary list populated with zeros and ones such that $O_{i,j} = 1$ means that the value $X_{i,j}$ is present and $O_{i,j} = 0$ means that the value $X_{i,j}$ is missing. In the following, let $\mathbf{o}_i, x_{ij}$ be realizations of $\mathbf{O}_i, X_{ij}$, respectively. The sampling probability for the data for subject i is then as follows:

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

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, $\int p(x_{i,1}, ..., x_{i,p}|\boldsymbol{\theta}, \Sigma) \prod_{x_{i,j}:o_{i,j}=0} dx_{i,j}$ has no closed form solution. Therefore, we will use Gibbs sampling to find estimates for the missing values. Let $\boldsymbol{X}$ 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 $X_{i,j}$ is observed and $o_{i,j} = 0$ if $X_{i,j}$ is not observed. We will then split the matrix $\boldsymbol{X}$ into two parts:

$$\begin{aligned} \boldsymbol{X}_{obs} &= \{x_{i,j} : o_{i,j} = 1\} \\ \boldsymbol{X}_{miss} &= \{x_{i,j} : o_{i,j} = 0\}. \end{aligned}$$

From our observed data we want to obtain $p(\boldsymbol{\theta}, \Sigma, \boldsymbol{X}_{miss}|\boldsymbol{X}_{obs})$ which is the posterior distribution of the data set 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 $p(\boldsymbol{\theta}|\boldsymbol{X}_{obs}, \boldsymbol{X}_{miss}^{(s)}, \Sigma^{(s)})$
2. sample $\Sigma^{(s+1)}$ from $p(\Sigma|\boldsymbol{X}_{obs}, \boldsymbol{X}_{miss}^{(s)}, \boldsymbol{\theta}^{(s+1)})$
3. sample $\boldsymbol{X}_{miss}^{(s+1)}$ from $p(\boldsymbol{Y}_{miss}|\boldsymbol{X}_{obs}, \boldsymbol{\theta}^{(s+1)}, \Sigma^{(s+1)})$.

For steps one and two, $\boldsymbol{X}_{obs}$ and $\boldsymbol{X}_{miss}$ combine to form $\boldsymbol{X}$. 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:

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

For each i we need to sample the missing values conditionally to the observed elements in the row. Next, from Theorem 6, we have

$$\{\boldsymbol{x}_{i,miss}, \boldsymbol{x}_{i,obs}, \boldsymbol{\theta}, \Sigma\} \sim \text{multivariate-normal}(\boldsymbol{\theta}_{miss|obs}, \Sigma_{miss|obs}),$$

where

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

Note that here we have used the notation $\Sigma_{[A,B]}$ to denote the matrix Σ with rows in A and columns in B. In the formulae 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. In the next sections, we will provide several examples to demonstrate the power of the Bayesian data imputation.

6.3 Example: Iris Flowers

The data set that will be imputed is the Iris flower data set Fisher (1936). The iris data set 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. In Figure 8, we plot the correlations among the variables in the data set. We note that for later comparison, the Iris data set is a complete data set 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. All code can be found at `https://github.com/Will-Holt60/Bayesian_Imputation`.

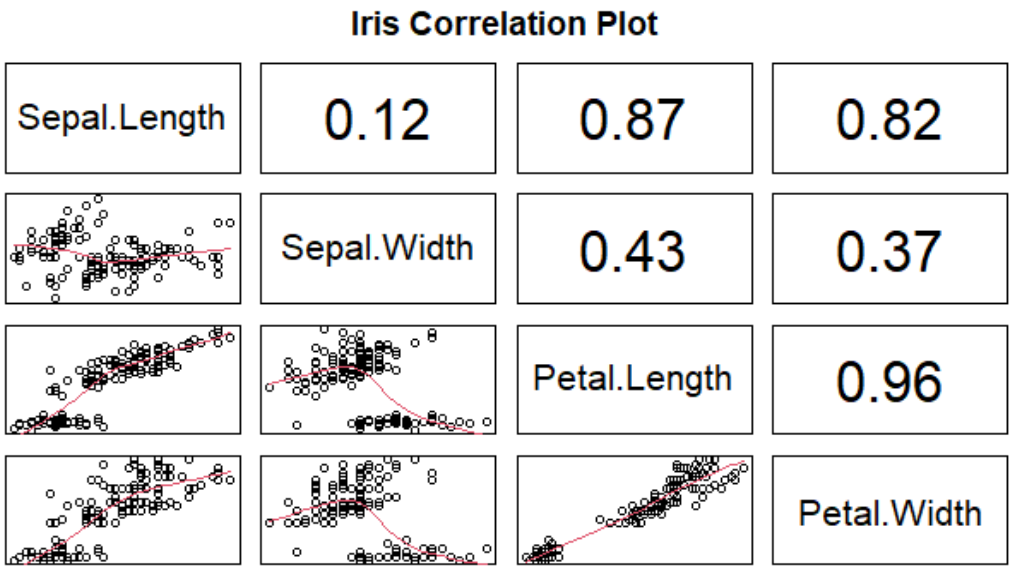

Figure 8: Correlations among Iris's variables

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 data set. 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. The this distribution of missing values are reported in Figure 9. In Figure 9, 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.

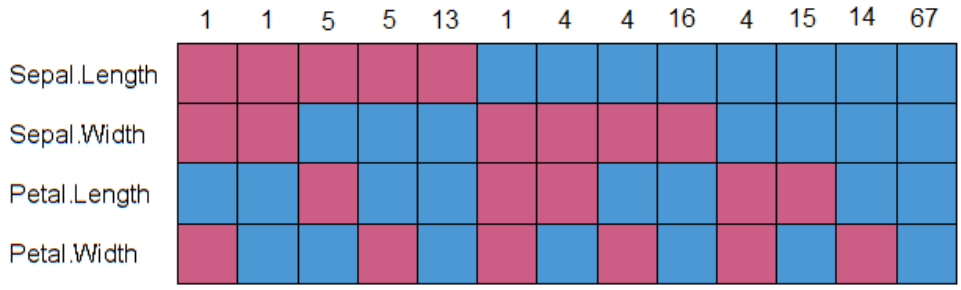

Figure 9: Distribution of missing values in Iris data set

Gibbs sampling was then used to implement Bayesian imputation on the missing data. We use the posterior mean of the $\mathbf{X}_{miss}$ to fill in the missing values in the data set. Since the true data is known so we can compare the imputed data with the true data. In Figure 10, we report the true values of missing values versus their Bayesian posterior mean for all variables under considerations. As expected, the Bayesian approach performs very well for this data set.

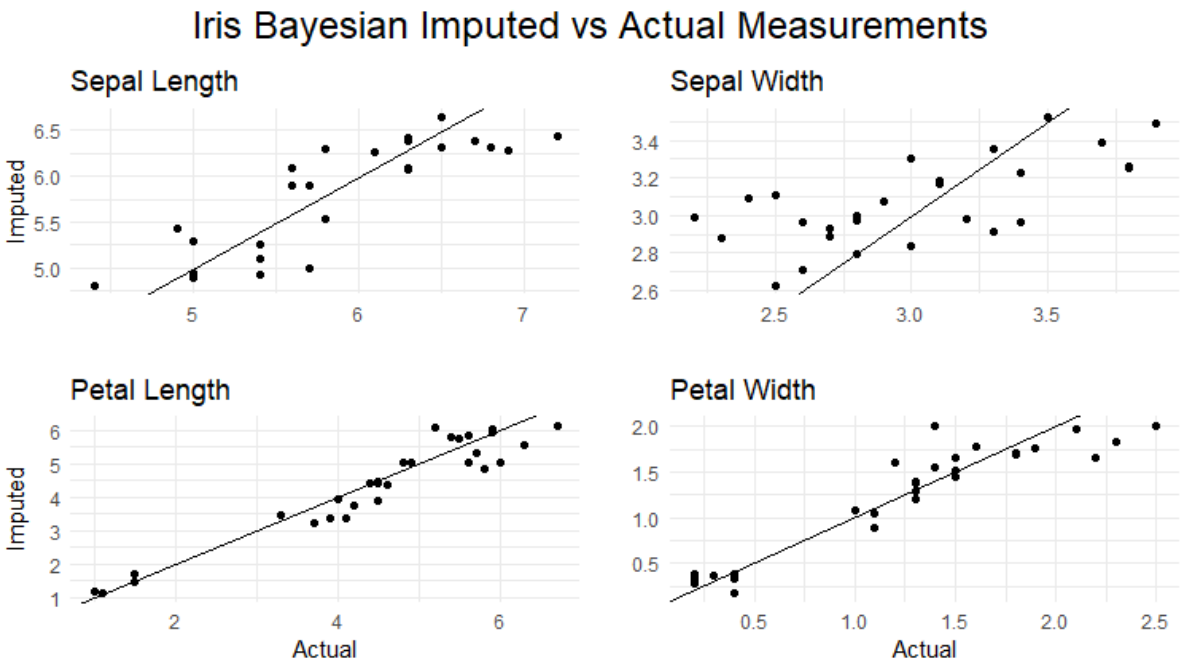

Figure 10: True values of missing values versus their posterior means

To further our comparison, we use three methods for dealing will missing data mentioned above (listwise deletion, mean imputation, and Bayesian imputation) were then compared against the original data set. The metrics for comparison are mean absolute deviation of the imputed values from the original values and the success rate of the classification algorithms such as random forest models trained on the data set. In Table 1, we report the mean absolute deviation of two approaches: Mean imputation and Bayesian imputation. It can be see that the Bayesian approach has much smaller mean absolute deviation, which means that it performs better than the mean imputation approach.

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

Table 1: Mean absolute deviation: Mean imputation versus Bayesian imputation

Next, we use original data set and the imputed data set obtained from listwise deletion, mean imputation and Bayesian imputation algorithms to perform classification using all variables. The model accuracy for the eight models created is in Table 2.

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

Table 2: Classification Accuracy

The results from Table 2 show that the Bayesian imputation method is by far the best approach to dealing with missing data. The Bayesian data set has almost the same predictive power as the original data set. 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 data set that was imputed using a Bayesian method has almost the same predictive power of the original data set. The other methods fall far behind in predictive accuracy.

6.4 Example: World Happiness Report

The next data set that will be imputed is the World Happiness Report data set. The world happiness report is a study on the state of world happiness done by the Gallup World Poll from 2005 to 2022 World Gallup Pole (2022). The data set has 12 features and 2200 rows, 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 data set has a few missing values, so in order to test the effectiveness of the data imputation, the rows with missing values were removed. Similar to the Iris data, the correlations among variables is displayed in Figure 11.

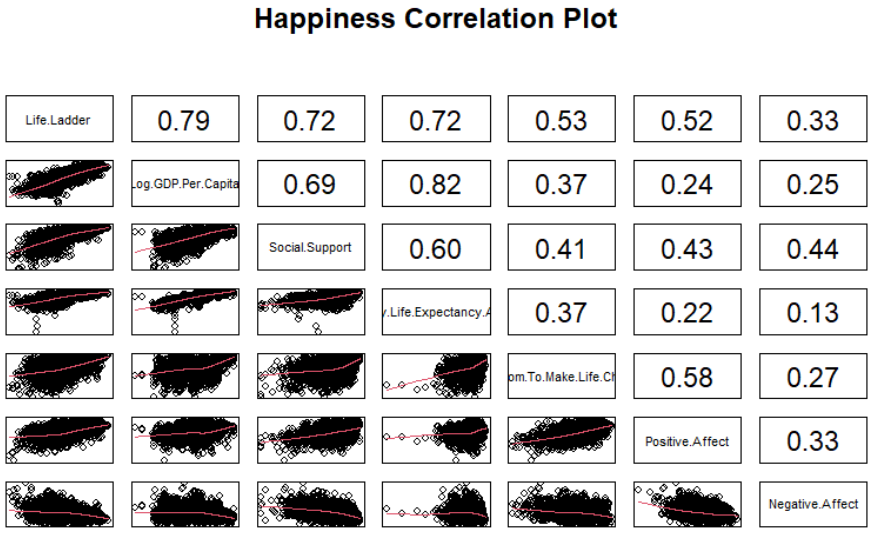

Figure 11: Correlations among variables

As can be seen in Figure 11, 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 data set. 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. The distribution of missing values are reported in Figure 12 In Figure 12, 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.

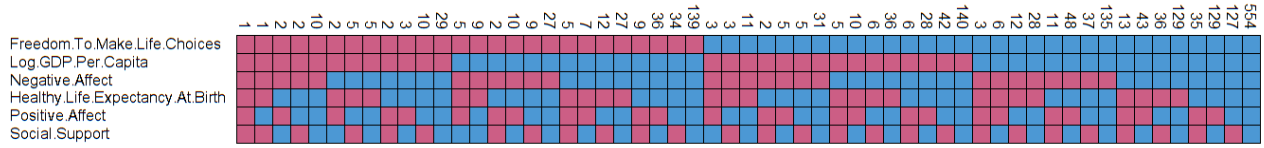

Figure 12: Distribution of missing values

Gibbs sampling was then used to implement Bayesian imputation on the missing data. We then plot the true values of missing values versus its Bayesian posterior means. The results are are reported in Figure 13.

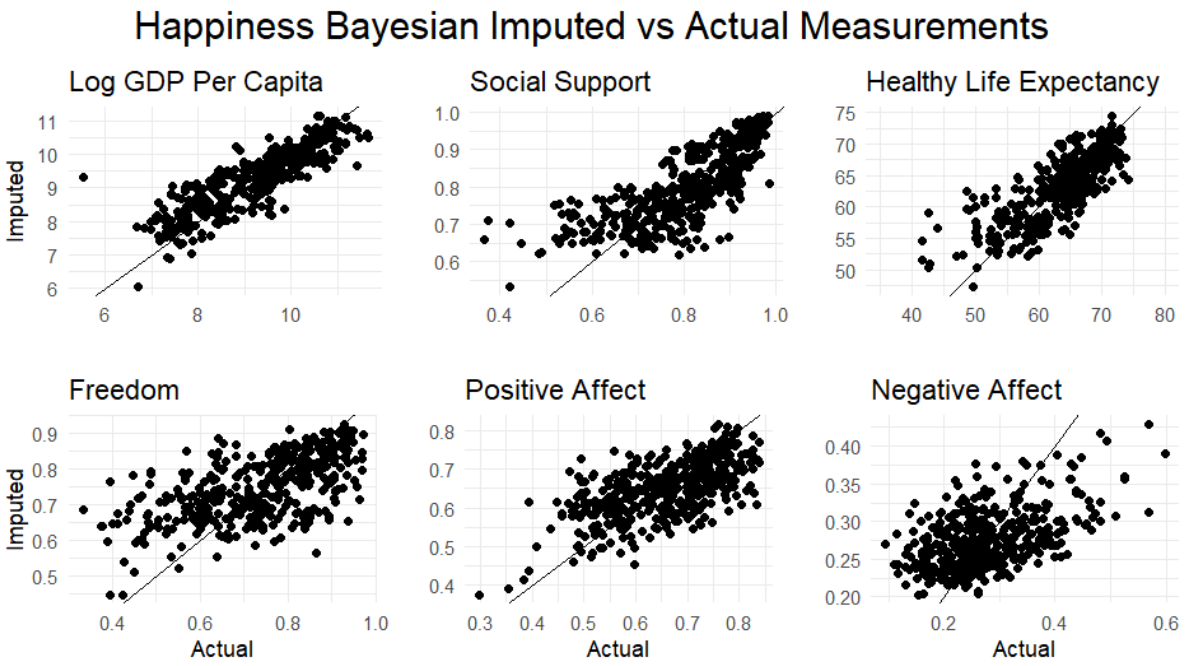

Figure 13: True values vs Bayesian imputed values

From Figure 13, 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 data set. 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.

First, the mean absolute deviation of the imputed data from the original data was calculated. To do this the imputed values were compared against the original values. The results are reported in Table 3.

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

Table 3: Mean Absolute Deviation of the Imputed Data

As can be seen in Table 3, Bayesian imputation has a lower mean absolute deviation value for every variable. This shows that Bayesian imputation tends to give more accurate values than mean imputation especially when there is strong correlation between the variables. 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 next comparison we will be looking at is the R^2 value of models trained on the four datasets created, the original dataset and the three imputed datasets. A random forest model and a linear regression model was then trained and tested on each of the four datasets giving a total of eight models. The R^2 of the models is in Table 4.

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

Table 4: Linear regression/Random Forest R^2 values

The results in Table 4 show that the Bayesian imputation method is by far the best approach to dealing with missing data. The Bayesian data set has almost the same predictive power as the original data set. The other methods fall far behind in predictive power.

7 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 data set.

References

Durrett, R. (2019). *Probability: theory and examples*, volume 49. Cambridge university press.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics* 7(2), 179–188.

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

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*, volume 580. Springer.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An introduction to statistical learning*, volume 112. Springer.

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American statistical association* 91(433), 222–230.

Kruschke, J. (2014). Doing bayesian data analysis: A tutorial with r, jags, and stan .

Van Buuren, S. (2012). Chapman & hall/crc interdisciplinary statistics series.

World Gallup Pole (2022). *World Happiness Report*. Sustainable Development Solutions Network.
URL `https://worldhappiness.report`