Bayesian Data Imputation

William Holt

Department of Computing Technology, Marist College, Poughkeepsie, NY

Abstract

Datasets often have missing values, as data collection is an imperfect process. There are many approaches to dealing with missing data such as listwise deletion and mean imputation. However, these methods cause significant penalties to the power of analysis. Bayesian imputation is a way of imputing missing values without incurring such steep penalties by using Bayesian inference.

Objectives

Let $\{\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ip}) : i = \overline{1,n}\}$ be iid sample from $MN(\theta, \Sigma)$ where $y_{ij} = NA$ for some $i \in \overline{1,n}, j \in \overline{1,p}$. Here $MN(\theta, \Sigma)$ denotes a multivariate normal distribution with mean $\theta = (\theta_1, \ldots, \theta_p)^T$ and covariance matrix $\Sigma = (\sigma_{ij})_{p \times p}$.

Our goal: To "fill in" those missing values using a Bayesian approach assuming the missing data are missing at random.

Background

The Bayesian approach to imputation uses gibbs sampling to take the mean of the column and adjust it based on the other values in the dataset.

For $i \in \overline{1,n}$, let $\boldsymbol{O}_i = \{O_1, \ldots, O_p)^T$ such that

$$O_{ij} = \begin{cases} 1 \text{ if } Y_{ij} \neq NA \\ 0 \text{ if } Y_{ij} = NA \end{cases}$$

Let $\boldsymbol{o}_i, y_{ij}$ be realizations of $\boldsymbol{O}_i, Y_{ij}$, respectively.

- $\boldsymbol{Y}_{obs} = \{y_{ij} : o_{ij} = 1\}$: the observed data
- $\boldsymbol{Y}_{miss} = \{y_{ij} : o_{ij} = 0\}$: the missing data

Note that

$$p(\boldsymbol{Y}_{miss} | \boldsymbol{Y}_{obs}, \theta, \Sigma) \propto \prod_{i=1}^{n} p(y_{i,miss} | y_{i,obs}, \theta, \Sigma). \qquad (1)$$

Gibbs Sampling

From our observed data, we want to obtain the posterior distribution $p(\theta, \Sigma, \boldsymbol{Y}_{miss} | \boldsymbol{Y}_{obs})$. We will use Gibbs sampling to generate a sequence of s random vectors $(\theta^{(s)}, \Sigma^{(s)}, \boldsymbol{Y}_{miss}^{(s)})$ that converge to the distribution $(\theta, \Sigma, \boldsymbol{Y}_{miss})$:

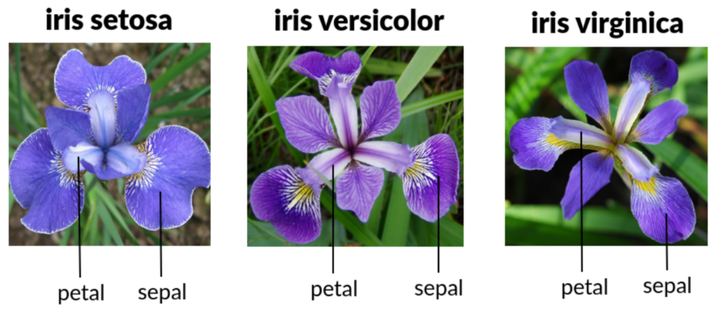
$$(\theta^{(0)}, \Sigma^{(0)}, \boldsymbol{Y}_{miss}^{(0)}) \to \ldots \to (\theta^{(s)}, \Sigma^{(s)}, \boldsymbol{Y}_{miss}^{(s)})$$

We will follow these steps:

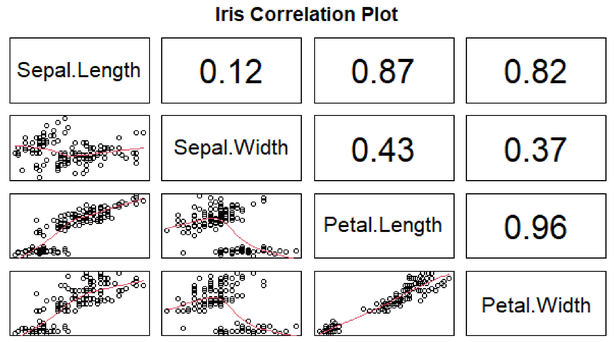
1. $s = 0$: Input the initial guesses $(\theta^{(s)}, \Sigma^{(s)}, \boldsymbol{Y}_{miss}^{(s)})$
2. Sampling $\theta^{(s+1)}$ from $p(\theta | \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{miss}, \Sigma^{(s)})$
3. Sampling $\Sigma^{(s+1)}$ from $p(\Sigma | \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{miss}, \theta^{(s+1)})$
4. Sampling $\boldsymbol{Y}_{miss}^{(s+1)}$ from $p(\boldsymbol{Y}_{miss} | \boldsymbol{Y}_{obs}, \theta^{(s+1)}, \Sigma^{(s+1)})$ using equation (1).

Note that in step 2 and step 3 above, the $p(\theta | \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{miss}, \Sigma)$ and $p(\Sigma | \boldsymbol{Y}_{obs}, \boldsymbol{Y}_{miss}, \theta)$ are known to be a multivariate normal distribution and inverse-Wishart distribution, respectively.

Example: Iris

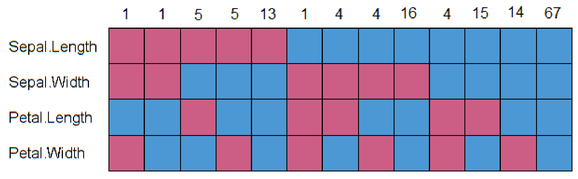


The dataset that will be imputed is the Iris flower dataset. The iris dataset is a 150 iris flower measurements done by Ronals Fisher in 1936. There are three flowers measured: Iris setosa, Iris virginica, and Iris versicolor. Each flower has four measurements: sepal length, sepal width, petal length, and petal width.



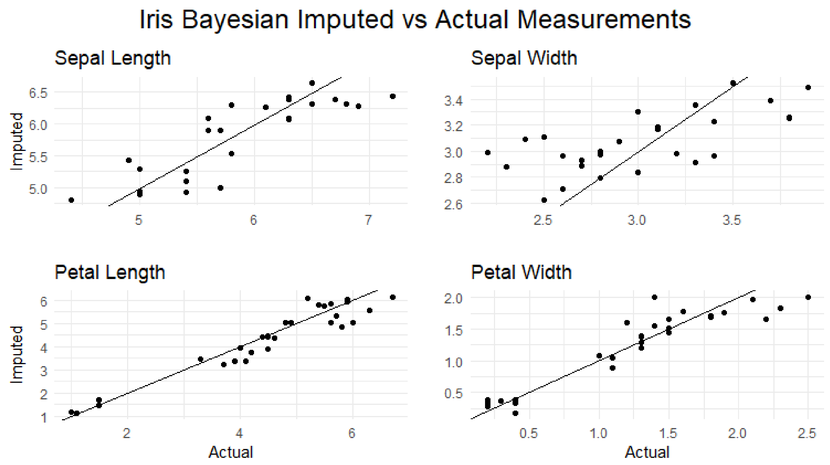
Value Deletion

20% of the values in each column were deleted giving the following missing data pattern with the number of rows that follow each pattern.



The three methods for dealing will missing data mentioned above (listwise deletion, mean imputation, and Bayesian imputation) were then compared against the original dataset.

Results



Mean Absolute Deviation of Imputation

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

Model Accuracy

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

Discussion

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.

Selected 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