A Missing Technique for High Dimensional Data

Mian Arif Shams Adnan[1], Silvia Irin Sharna[2]
[1,2]Bowling Green State University, Bowling Green, OH 43403

Abstract
Since a high dimensional missing value resembles not only an unknown high dimensional data of an unknown high dimensional probability distribution but also their unknown characteristics, it is better to construct a basket of characteristics based on assumed high dimensional missing values. The missing technique, as demonstrated by Sharna *et al* (2017, 2016), is a check and balance method for estimating missing value(s). In this paper we offer an extended version of the iterative estimation method for high dimensional missing value. This paper also demonstrates a resampling method for generating 1 or 2 correlated observations from the same high dimensional distribution from where the original sample is drawn.

Key Words: Average Log Likelihood Function, Combination, Dummy Missing Value, Likelihood Rate, Simple Random Sample.

1. Introduction

Missing data pattern describes which values are observed in the data matrix and which values are missing, and missing data mechanism addresses the relationship between missing value and the available values in the data matrix. Missing value estimation is a common problem in several statistical studies. The problem synchronized a lot when the sample size is very small and sensitive. Missing data mechanisms addresses the dependencies among the missing data and the available data. Rubin (1976) developed a device of treating the missing data indicators as random variables along with a distribution.

The literature on analysis of partially missing data is inaugurated by Afifi and Elashoff (1966), Hartley and Hocking (1971), Orchard and Woodbury (1972), Dempster, Laird, and Rubin (1977), Litte and Rubin (1983 a), Little and Schenker (1994), and Little (1997) as addressed by the book written by Little. R. J. A and Rubin. D. B. (2002). Methods proposed by the aforesaid authors can be grouped into the following categories. The categories include Procedures Based on Completely Record Units, Weighting Procedures, Imputation-Based Procedures and Model-Based Procedures. Broadly there are two ways for estimating missing values. These are Missing Value Estimation in Experiment and Missing Value Estimation by Likelihood Based Method. Imputation Method, Weighted Methods by Complete Case and Available Case Analysis are from class one. And Inference based Likelihood method, Factored Likelihood Method, EM Algorithm, Large Sample Inference based Maximum Likelihood Method, Bayesian Iterative Simulation Method, Robust Method, Partially Classified Contingency Table Method (ML Estimation, Bayes Estimation, Log-linear Model, Logistic Regression Method) etc. are from class two.

Allan and Wishart's (1930), Wilkinson's (1958), Hartley's (1956), Westmacott (1956), Pearce (1965, p.111;1971), Bartlett (1937) demonstrated some methods/modifications for estimating missing values. A variety of techniques are available in the literature to estimate missing values. These will be reviewed briefly later. Sharna *et al* (2017, 2016) proposed Missing techniques to estimate one and more than one missing value(s). In this paper we extended the paper to develop a Missing Value Estimation Technique to estimate more than one missing values for High Dimensional Distribution.

2. New Method and Methodology

Let there be $(n-2)$ observations and 2 missing observations. We want to estimate the missing paired observations. We know nothing about missing value or the distribution of observation from where the observations are drawn. So, we know nothing about the missing value, or the distribution of the observations or the parameters of the distribution or other characteristics like mean, median, mode, variance, skewness, kurtosis, and higher order moments of the distribution. In this situation we will estimate all the aforesaid characteristics and their volatility due to the change of sample size. We will also measure the deviation of the estimated characteristics from those of the missing values. So, we adjust our estimates of various characteristics due to the exact sample size and bandwidth of each of the characteristics. Later all the estimated characteristics will be used to find out several relation among themselves to predict the probability distribution. The parameters will also be estimated under the predicted probability distribution. Later on the deviation of the theoretically estimated characteristics and practically observed characteristics can be found to check how better the predicted distribution was by checking the equivalence of the theoretical and observed characteristics. Average Maximum Likelihood function and the consistent rate of the mean sum of squares of error can be found to confirm that the performance of the estimated missing values and the error conducted due to the estimated missing values is the least.

2.1 Estimating First Missing Value from a Sample of Size *n*

Let the observations $X_{1,}X_{2,}\ldots, X_{n-2}$ be non-missing and two observations be missing. Let the missing observation be Y and Z. We want to estimate Y and Z. So out of $(n-2)$ non-missing observations $n-2_{C_{n-2-2}}$ samples each of size $(n-2-2)$ can be drawn assuming two observations for each sample are missing. Assuming two non-missing observation as tw missing ones we can generate $n-2_{C_{n-4}}$ samples each of which is consisting of $(n-4)$ non-missing observations pretending the rest non-missing observations as the missing observation. So the $n-2_{C_{n-4}}$ generated samples are as below:

$\text{n} - 2_{C_{\text{n}-4}}$ samples each of size (n − 4)	Assumed missing observation
$X_{1,}X_{2,}\ldots, X_{n-2}$	X_{n-1}, X_n
…	…
$X_{1,}X_{3,}\ldots, X_{n-1}$	X_2, X_n
$X_{3,}\ldots, X_n$	X_1, X_2

So we have calculated a class of characteristics (demonstrated in Table 1) to develop and observe several relationships among themselves (characteristics). For each of these characteristics, we will observe it's deviation from the same characteristic with the presence of two dummy missing observations. Let us at first explain the easiest characteristic say sample mean and its sample standard deviation from the assumed missing value as addressed in Table 2.

If we assume one of the aforesaid two missing observations as the estimate of the *n-1*th pretended missing observation, and (if we consider) the available original observations $X_{1,}X_{2,}\ldots,X_{n-2}$ as the $(n-2)$ other non-missing observations then the consecutive Maximum Likelihood Function or Likelihood Rate will be

$$L = f(X_1;\bar{X},S)f(X_2;\bar{X},S)\ldots f(X_{n-2};\bar{X},S)$$

$$\log\,(L) = log[f(X_1;\bar{X},S)f(X_2;\bar{X},S)\ldots f(X_{n-2};\bar{X},S)]$$

$$\log(L) = log\big(f(X_1;\bar{X},S)\big) + log\big(f(X_2;\bar{X},S)\big) + \cdots + log(f(X_{n-2};\bar{X},S))$$

$$\therefore \frac{1}{n-2}\log(L) = \frac{1}{n-2}\sum_{i=1}^{n-2}\log\big(f(X_i;\bar{X},\ S)\big)$$

which can be termed as the average expected log likelihood function or expected log-likelihood rate. Now, we should generate short incremented (various) values for X such that

$$\text{dis}(\bar{X},RS)\text{<kAD.}$$

Here *k* is a very small number.

Now, $$L' = f(X_1;\bar{X},S)f(X_2;\bar{X},S)\ldots\ f(X_{n-2};\bar{X},S)f(\boldsymbol{X_{n-1}};\bar{X},S)$$

$$\log\,(L') = log[f(X_1;\bar{X},S)f(X_2;\bar{X},S)\ldots\ f(X_{n-2};\bar{X},S)f(\boldsymbol{X_{n-1}};\bar{X},S)]$$

$$\log(L') = log\big(f(X_1;\bar{X},S)\big) + log\big(f(X_1;\bar{X},S)\big) + \cdots + log\big(f(\boldsymbol{X_{n-1}};\bar{X},S)\big)$$

$$\frac{1}{n-1}\log(L') = \frac{1}{n-1}\sum_{i=1}^{n-1}\log\big(f(X_i;\bar{X},S)\big)$$

We will search the incremented value of the *n-1*th observation for which the expected log likelihood rate and the observed log likelihood rate will be same i.e.

$$\frac{1}{n-1}\log(L') = \frac{1}{n-1}\sum_{i=1}^{n-1}\log\big(f(X_i;\bar{X},S)\big) \cong \frac{1}{n-2}\log(L) = \frac{1}{n-2}\sum_{i=1}^{n-2}\log\big(f(X_i;\bar{X},S)\big)$$

The incremented value of the *n-1*th observation for which the likelihood functions are same, will be an efficiently-estimated value of the n-1th missing observation.

However, if we get more than two estimates of the missing observation, we can check for which estimate of the missing value the first two moments are close to those of the original $(n-2)$ observations. Hence, we will find the closer estimate of the missing value. Therefore, if we get more than two or three or more estimates of a missing observation, we can use all the estimates to estimate that missing value. Hence, we will estimate the (n-1)th missing observation which is the estimate of one missing value out of two missing value.

So, we have described how $n-2_{C_{n-2-2}}$ samples have been generated assuming two non-missing observations as two missing ones in each case and calculated their sample averages to find out a bandwidth for the first missing value. Here the missing value has been determent adding the half of the bandwidth of the 1st missing value with the average of all of the available non-missing values. Similarly, several sample characteristics and their bandwidth can be calculated to find out different characteristics of the missing data as well as the distribution from which the sample (consisting of the 1st missing value and non-missing value) has been drawn. So, sample variance, sample higher order moments, sample median, mode, skewness, kurtosis, tail behaviors, etc. can be found using their respective bandwidth. Several relationships can be explored from the aforesaid estimated characteristics to recognize the pattern of the distribution and its relevant features. The relevant features, estimated parameters and the predicted distribution are used to fit the observed sample data. So least square fitting or least deviation fitting or any sort of other goodness of fit can be used to check the performance of the predicted probabilistic model along-with the bandwidth based estimated parameters and the characteristics. After checking the fitting performance of the predicted model for the observed data, we can observe whether the average log-likelihood function for both the non-missing and the first missing value is equivalent that of the average log-likelihood rate for the all non-missing values.

After estimating the first missing value, we will estimate the 2nd as well as the last missing value based on the non-missing values and the estimated 1st missing value. Hence, we will repeat the previously developed method of estimating one missing value by Sharna *et al* (2016) as follows.

2.2 Estimating Last Missing Value from a Sample of Size *n* using the Estimated Missing Value

Suppose there are n observations out of which $(n-1)$ non-missing observations and one missing observation. We also suppose that observations $X_{1,}X_{2,}\dots, X_{n-1}$ are non-missing and one observation x_n is missing. We want to estimate x_n. So out of $(n-1)$ non-missing observations, $(n-1)$ samples each of which is of size $(n-2)$ can be drawn assuming each sample has one missing observation. Assuming one non-missing observation as a missing one we can generate $(n-1)$ samples each of which is consisting of $(n-2)$ non-missing observations pretending the rest non-missing observations as the missing observation. So, the $(n-1)$ generated samples are as below:

$(n-1)$ samples each of size $(n-2)$	*Assumed missing observation*
$X_{1,}X_{2,}\dots,X_{n-2}$	X_{n-1}
$X_{1,}X_{2,}\dots,X_{n-1}$	X_{n-2}
$\dots$	$\dots$
$X_{1,}X_{3,}\dots,X_{n-2}$	X_2
$X_{2,}X_{3,}\dots,X_{n-1}$	X_1

So, we have calculated a class of characteristics (demonstrated in Table 1) to develop and observe several relationships among themselves (characteristics). For each of these characteristics, we will observe it's deviation from the same characteristic with the presence of dummy missing observation. Let us at first explain the easiest characteristic say sample mean and its deviation from the assumed missing value as addressed in Table 2.

If we assume any of the aforesaid observations as the estimate of the *n*th pretended missing observation, and (if we consider) the available original observations $X_{1,}X_{2,}\dots,X_{n-1}$ as the $(n-1)$ other non-missing observations then the consecutive Maximum Likelihood Function or Likelihood Rate will be

$$L' = f(X_1;\bar{X}^*,S^*)f(X_2;\bar{X}^*,S^*)\dots f(X_{n-1};\bar{X}^*,S^*)$$

$$\log(L') = log[f(X_1;\bar{X}^*,S^*)f(X_2;\bar{X}^*,S^*)\dots f(X_{n-1};\bar{X}^*,S^*)]$$

$$\log(L') = log\big(f(X_1;\bar{X}^*,S^*)\big) + log\big(f(X_2;\bar{X}^*,S^*)\big) + \cdots + log\big(f(X_{n-1};\bar{X}^*,S^*)\big)$$

$$\therefore \tfrac{1}{n-1}\log(L') = \tfrac{1}{n-1}\textstyle\sum_{i=1}^{n-1}\log(f(X_i;\bar{X}^*\,S^*))$$

which can be termed as the average expected log likelihood function or expected log-likelihood rate. Now, we should generate short incremented (various) values for X such that

$$\text{dis}(\bar{X}^*,RS^*)<k^*AD^*.$$

Here k is a very small number.

Now,

$$L \;\; = f(X_1;\bar{X}^*,S^*)f(X_2;\bar{X}^*,S^*)\dots\; f(X_n;\bar{X}^*,S^*)$$

$$\log(L) = log[f(X_1;\bar{X}^*,S^*)f(X_2;\bar{X}^*,S^*)\dots\; f(X_n;\bar{X}^*,S^*)]$$

$$\log(L) = log\big(f(X_1;\bar{X}^*,S)\big) + log\big(f(X_2;\bar{X}^*,S)\big) + \cdots + log\big(f(X_n;\bar{X}^*,S)\big)$$

$$\tfrac{1}{n}\log(L) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\log(f(X_i;\bar{X}^*,S^*))$$

We will search the incremented value of the nth observation for which the expected log likelihood rate and the observed log likelihood rate will be same i.e.

$$\tfrac{1}{n}\log(L) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\log\big(f(X_i;\bar{X}^*,S^*)\big) \cong \tfrac{1}{n-1}\log(L') = \tfrac{1}{n-1}\textstyle\sum_{i=1}^{n-1}\log\big(f(X_i;\bar{X}^*,S^*)\big).$$

The incremented value of the nth observation for which the likelihood functions are same, will be an efficiently-estimated value of the missing observations.

However, if we get more than two estimates of the missing observation, we can check for which estimate of the missing value the first two moments are close to those of the original $(n-1)$ observations. Hence, we will find the closer estimate of the missing value. Therefore, if we get more than two or three or more estimates of a missing observation, we can use all the estimates to estimate that missing value.

So, we have described how $(n-1)$ samples have been generated assuming one non-missing observation as a missing one in each case and calculated their sample averages to find out a bandwidth for the missing value. Here the missing value has been determent adding the half of the bandwidth of the missing value with the average of all of the available non-missing values. Similarly, several sample characteristics and their bandwidth can be calculated to find out different characteristics of the missing data as well as the distribution from which the sample (consisting of missing value and non-missing value) has been drawn. So, sample variance, sample higher order moments, sample median, mode, skewness, kurtosis, tail behaviors, etc. can be found using their respective bandwidth. Several relationships can be explored from the aforesaid estimated characteristics to recognize the pattern of the distribution and its relevant features. The relevant features, estimated parameters and the predicted distribution are used to fit the observed sample data. So least square fitting or least deviation fitting or any sort of other goodness of fit can be used to check the performance of the predicted probabilistic model along-with the bandwidth based estimated parameters and the characteristics. After checking the fitting performance of the predicted model for the observed data, we can observe whether the average log-likelihood function for both the non-missing and missing values is equivalent that of the average log-likelihood rate for the all non-missing values.

2.3 Estimating First Missing Value from a Sample of Size 6

For more clarification let $n = 6$. So there are 4 non-missing observations and 2 missing observations. The non-missing observations are $X_{1,}X_{2,}X_3, X_4$ and the missing observations are X_6 and X_5. Assuming two non-missing observations as missing ones we can generate 6 samples each of which is consisting of 2 non-missing observations assuming the rest non-missing observations as the missing observations. So, the 6 samples are as below:

Samples of size 2	*Assumed missing observations*
$X_{1,}X_2$	X_3, X_4
$X_{1,}X_3$	X_2, X_4
$X_{1,}X_4$	X_2, X_3
$X_{2,}X_3$	X_1, X_4
$X_{2,}X_4$	X_1, X_3
$X_{3,}X_4$	X_1, X_2

Table 1: Sample means and sample variances for several samples.

Sample Mean	*Sample Variance*
$\overline{X_1} = \frac{X_{1+}X_2}{2}$	$S_1 = \frac{(X_1 - \overline{X_1})\acute{}(X_1 - \overline{X_1})' + (X_2 - \overline{X_1})(X_2 - \overline{X_1})'}{2-1}$
$\overline{X_2} = \frac{X_{1+}X_3}{2}$	$S_2 = \frac{(X_1 - \overline{X_2})\acute{}(X_1 - \overline{X_2})' + (X_3 - \overline{X_2})(X_3 - \overline{X_2})'}{2-1}$
$\overline{X_3} = \frac{X_{1+}X_4}{2}$	$S_3 = \frac{(X_1 - \overline{X_3})\acute{}(X_1 - \overline{X_3})' + (X_4 - \overline{X_3})(X_4 - \overline{X_3})'}{2-1}$
$\overline{X_4} = \frac{X_{2+}X_3}{2}$	$S_4 = \frac{(X_2 - \overline{X_4})\acute{}(X_2 - \overline{X_4})' + (X_3 - \overline{X_4})(X_3 - \overline{X_4})'}{2-1}$
$\overline{X_5} = \frac{X_{2+}X_3}{2}$	$S_5 = \frac{(X_2 - \overline{X_5})\acute{}(X_2 - \overline{X_5})' + (X_3 - \overline{X_5})(X_3 - \overline{X_5})'}{2-1}$
$\overline{X_6} = \frac{X_{3+}X_4}{2}$	$S_6 = \frac{(X_3 - \overline{X_6})\acute{}(X_3 - \overline{X_6})' + (X_4 - \overline{X_6})(X_4 - \overline{X_6})'}{2-1}$
$\bar{X} = \frac{\overline{X_1}+\overline{X_2}+\overline{X_3}+\overline{X_4}++\overline{X_5}+\overline{X_6}}{6}$	$S = \frac{S_1 + S_2 + S_3 + S_4 + S_5 + S_6}{6}$

So, we have calculated a class of characteristics to develop and observe some relationships among them (characteristics). For each of these characteristics we will observe it's deviation from the same characteristic with the presence of assumed missing observation. Let us at first explain the easiest characteristics say sample mean and its deviation from the assumed missing value in the following table:

Table 2: Sample mean difference for several samples.

Sample Mean of size 2	*Assumed Missing Values*	*Difference*	$\lvert Difference \rvert$
$\overline{X_1} = \frac{X_1+X_2}{2}$	X_3, X_4	$\overline{X_1} - \frac{X_3+X_4}{2}$	$\mathrm{dis}(\overline{X_1}, \overline{X_1{}'})$
$\overline{X_2} = \frac{X_1+X_3}{2}$	X_2, X_4	$\overline{X_2} - \frac{X_2+X_4}{2}$	$\mathrm{dis}(\overline{X_2}, \overline{X_2{}'})$
$\overline{X_3} = \frac{X_1+X_4}{2}$	X_2, X_3	$\overline{X_3} - \frac{X_2+X_3}{2}$	$\mathrm{dis}(\overline{X_3}, \overline{X_3{}'})$
$\overline{X_4} = \frac{X_2+X_3}{2}$	X_1, X_4	$\overline{X_4} - \frac{X_1+X_4}{2}$	$\mathrm{dis}(\overline{X_4}, \overline{X_4{}'})$
$\overline{X_5} = \frac{X_2+X_3}{2}$	X_1, X_3	$\overline{X_5} - \frac{X_1+X_3}{2}$	$\mathrm{dis}(\overline{X_5}, \overline{X_5{}'})$
$\overline{X_6} = \frac{X_3+X_4}{2}$	X_1, X_2	$\overline{X_6} - \frac{X_1+X_2}{2}$	$\mathrm{dis}(\overline{X_6}, \overline{X_6{}'})$
		Total	$\mathrm{dis}(\overline{X_1}, \overline{X_1{}'}) + \mathrm{dis}(\overline{X_2}, \overline{X_2{}'}) + \mathrm{dis}(\overline{X_3}, \overline{X_3{}'}) + \mathrm{dis}(\overline{X_4}, \overline{X_4{}'}) + \mathrm{dis}(\overline{X_5}, \overline{X_5{}'}) + \mathrm{dis}(\overline{X_6}, \overline{X_6{}'})$
		Average Absolute Difference (AD)	$\frac{\mathrm{dis}(\overline{X_1},\overline{X_1{}'})+\mathrm{dis}(\overline{X_2},\overline{X_2{}'})+\mathrm{dis}(\overline{X_3},\overline{X_3{}'})+\mathrm{dis}(\overline{X_4},\overline{X_4{}'})+\mathrm{dis}(\overline{X_5},\overline{X_5{}'})+\mathrm{dis}(\overline{X_6},\overline{X_6{}'})}{6}$

Now,

$$L = f(X_1; \bar{X}, S) f(X_2; \bar{X}, S) f(X_3; \bar{X}, S) f(X_4; \bar{X}, S)$$

$$\log\,(L) = log[f(X_1; \bar{X}, S) f(X_2; \bar{X}, S) f(X_3; \bar{X}, S) f(X_4; \bar{X}, S)]$$

$$\log(L) = log(f(X_1; \bar{x}, S)) + logf(X_2; \bar{X}, S)) + log(f(X_3; \bar{X}, S)) + log(f(X_4; \bar{X}, S))$$

$$\frac{1}{4}\log(L) = \frac{1}{4}\sum_{i=1}^{4} \log(f(X_i; \bar{X}, S))$$

Which can termed as the average expected likelihood or expected likelihood rate.

Now, we should generate short incremented various values form the range

$$\mathrm{dis}(\bar{X}, RS) < \mathrm{kAD}$$

Here k may be $\frac{1}{1000}$ or $\frac{1}{100}$ or $\frac{1}{10}$ or so on. The increment h can take the value 0.01 or 0.05 or 0.10 and so on. The values could be

$$\mathrm{dis}\left(\frac{1}{4}\sum_{i=1}^{4} X_i\,, RS\right) < k\,\frac{\mathrm{dis}(\overline{X_1},\overline{X_1{}'}) + \mathrm{dis}(\overline{X_2},\overline{X_2{}'}) + \mathrm{dis}(\overline{X_3},\overline{X_3{}'}) + \mathrm{dis}(\overline{X_4},\overline{X_4{}'}) + \mathrm{dis}(\overline{X_5},\overline{X_5{}'}) + \mathrm{dis}(\overline{X_6},\overline{X_6{}'})}{6}$$

If we assume any one of the two afore said observations as the 5th observation and the four other observations are the given original observations $X_{1,}X_{2,}X_3, X_4$; then the consecutive average observed likelihood or observed likelihood rate will be

$$L' = f(X_1;\bar{X},S)f(X_2;\bar{X},S)f(X_3;\bar{X},S)f(X_4;\bar{X},S)\ f(\boldsymbol{RS} = \boldsymbol{X_5};\bar{X},S)$$

$$\log\,(L') = log[f(X_1;\bar{X},S)f(X_2;\bar{X},S)f(X_3;\bar{X},S)f(X_4;\bar{X},S)f(\boldsymbol{RS} = \boldsymbol{X_5};\bar{X},S)]$$

$$\log(L') = log(f(X_1;\bar{X},S)) + log(f(X_2;\bar{X},S)) + log(f(X_3;\bar{X},S)) + log(f(X_4;\bar{X},S)) + log(f(\boldsymbol{RS} = \boldsymbol{X_5};\bar{X},S))$$

$$\tfrac{1}{5}\log(L') = \tfrac{1}{5}\textstyle\sum_{i=1}^{5}\log(f(X_i;\bar{X},S))$$

If we get more than two estimates of the missing observation (since we get two values of the 5th observation for whom the likelihood rates are same), we can check for which estimate of the missing value the first two moments are close to those of the original 4 observations. Hence, we will find the estimate of the missing values.

We will search the incremented value of the 5th observation for which the expected likelihood rate and the observed likelihood rate will be same i.e.

$$\tfrac{1}{5}\log(L') = \tfrac{1}{5}\textstyle\sum_{i=1}^{5}\log(f(X_i;\bar{X},S)) \cong \tfrac{1}{4}\log(L) = \tfrac{1}{4}\sum_{i=1}^{4}\log(f(X_i;\bar{X},S)).$$

If we get more than two or three or more estimates of each of the missing observations, we can have the corresponding averages all the estimates of the missing values and can assume that as the estimate of that missing value. Hence, we have obtained the 5th observation. We will now estimate the 6th (last) observation in the next step.

2.4 Estimating Last Missing Value from a Sample of Size 6 using the Estimated Missing Value

Now let $n = 6$. So there are 4 non-missing observations and one missing observation. The non-missing observations are $X_{1,}X_{2,}X_3, X_4, X_5$ and the missing observation is X_6. So, assuming one non-missing observation as a missing one we can generate 5 samples each of which is consisting of 4 non-missing observations assuming the rest non-missing observations as the missing observation. So, the 5 samples are as below:

Samples of size 4	*Assumed missing observation*
$X_{1,}X_{2,}X_3, X_4$	X_5
$X_{1,}X_{2,}X_3, X_5$	X_4
$X_{1,}X_{2,}X_{4,}X_5$	X_3
$X_{1,}X_{3,}X_{4,}X_5$	X_2
$X_{2,}X_{3,}X_{4,}X_5$	X_1

Table 3: Sample means and sample variances for several samples.

Sample Mean	*Sample Variance*
$\overline{X_1} = \frac{X_1 + X_2 + X_3 + X_4}{4}$	$S_1 = \frac{(X_1 - \overline{X_1})(X_1 - \overline{X_1})' + (X_2 - \overline{X_1})(X_2 - \overline{X_1})' + (X_3 - \overline{X_1})(X_3 - \overline{X_1})' + (X_4 - \overline{X_1})(X_4 - \overline{X_1})'}{4-1}$
$\overline{X_2} = \frac{X_1 + X_2 + X_3 + X_5}{4}$	$S_2 = \frac{(X_1 - \overline{X_1})(X_1 - \overline{X_1})' + (X_2 - \overline{X_1})(X_2 - \overline{X_1})' + (X_3 - \overline{X_1})(X_3 - \overline{X_1})' + (X_5 - \overline{X_1})(X_5 - \overline{X_1})'}{4-1}$
$\overline{X_3} = \frac{X_1 + X_2 + X_4 + X_5}{4}$	$S_3 = \frac{(X_1 - \overline{X_1})(X_1 - \overline{X_1})' + (X_2 - \overline{X_1})(X_2 - \overline{X_1})' + (X_4 - \overline{X_1})(X_4 - \overline{X_1})' + (X_5 - \overline{X_1})(X_5 - \overline{X_1})'}{4-1}$
$\overline{X_4} = \frac{X_1 + X_3 + X_4 + X_5}{4}$	$S_4 = \frac{(X_1 - \overline{X_1})(X_1 - \overline{X_1})' + (X_3 - \overline{X_1})(X_3 - \overline{X_1})' + (X_4 - \overline{X_1})(X_4 - \overline{X_1})' + (X_5 - \overline{X_1})(X_5 - \overline{X_1})'}{4-1}$
$\overline{X_5} = \frac{X_2 + X_3 + X_4 + X_5}{4}$	$S_5 = \frac{(X_2 - \overline{X_1})(X_2 - \overline{X_1})' + (X_3 - \overline{X_1})(X_3 - \overline{X_1})' + (X_4 - \overline{X_1})(X_4 - \overline{X_1})' + (X_5 - \overline{X_1})(X_5 - \overline{X_1})'}{4-1}$
$\bar{X}^* = \frac{\overline{X_1} + \overline{X_2} + \overline{X_3} + \overline{X_4} + \overline{X_5}}{5}$	$S^* = \frac{S_1 + S_2 + S_3 + S_4 + S_5}{5}$

Table 4: Sample mean difference for several samples.

Sample Mean of size 5	*Assumed Missing Values*	*Difference*	$|Difference|$
$\overline{X_1}$	X_5	$\overline{X_1} - X_5$	$\mathrm{dis}(\overline{X_1}, X_5)$
$\overline{X_2}$	X_4	$\overline{X_2} - X_4$	$\mathrm{dis}(\overline{X_2}, X_4)$
$\overline{X_3}$	X_3	$\overline{X_3} - X_3$	$\mathrm{dis}(\overline{X_3}, X_3)$
$\overline{X_4}$	X_2	$\overline{X_4} - X_2$	$\mathrm{dis}(\overline{X_4}, X_2)$
$\overline{X_5}$	X_1	$\overline{X_5} - X_1$	$\mathrm{dis}(\overline{X_5}, X_1)$
Total			$\mathrm{dis}(\overline{X_1}, X_5) + \mathrm{dis}(\overline{X_2}, X_4) + \mathrm{dis}(\overline{X_3}, X_3) + \mathrm{dis}(\overline{X_4}, X_2) + \mathrm{dis}(\overline{X_5}, X_1)$
Average Absolute Difference	AD^*		$\frac{\mathrm{dis}(\overline{X_1}, X_5) + \mathrm{dis}(\overline{X_2}, X_4) + \mathrm{dis}(\overline{X_3}, X_3) + \mathrm{dis}(\overline{X_4}, X_2) + \mathrm{dis}(\overline{X_5}, X_1)}{5}$

So, we have calculated a class of characteristics (Table 3) to develop and observe some relationships among them (characteristics). For each of these characteristics we will observe it's deviation from the same characteristic with the presence of assumed missing observation. Let us at first explain the easiest characteristics say sample mean and its deviation from the assumed missing value in the Table 4.

Now,

$$L = \ f(X_1;\bar{X}^*,S^*)f(X_2;\bar{X}^*,S^*)f(X_3;\bar{X}^*,S^*)f(X_4;\bar{X}^*,S^*)f(X_5;\bar{X}^*,S^*)$$

$$\begin{aligned}&\log{(L)}\\&= log[f(X_1;\bar{X}^*,S^*)f(X_2;\bar{X}^*,S^*)f(X_3;\bar{X}^*,S^*)f(X_4;\bar{X}^*,S^*)f(X_5;\bar{X}^*,S^*)]\end{aligned}$$

$$\begin{aligned}\log(L) = log(f(X_1;\bar{X}^*,S^*)) + log(f(X_2;\bar{X}^*,S^*)) + log(f(X_3;\bar{X}^*,S^*))\\ + log(f(X_4;\bar{X}^*,S^*)) + log(f(X_5;\bar{X}^*,S^*))\end{aligned}$$

$$\tfrac{1}{5}\log(L) = \tfrac{1}{5}\textstyle\sum_{i=1}^{5}\log(f(X_i;\bar{X}^*,S^*))$$

which can termed as the average expected log likelihood or expected log likelihood rate.

Now, we should generate short incremented various values form the range

$$\text{dis}(\bar{X}^*,RS^*)<k^*AD^*$$

$$\text{or, dis}\left(\tfrac{1}{5}\textstyle\sum_{i=1}^{5}X_i\,,RS^*\right) < k^*\,\frac{\overline{|X_1}-X_5|+\overline{|X_2}-X_{4|}+\overline{|X_3}-X_{3|}+\overline{|X_4}-X_2|+\overline{|X_5}-X_1|}{5}$$

Here k may be $\frac{1}{1000}$ or $\frac{1}{100}$ or $\frac{1}{10}$ or so on.

If we assume any of the afore said observations as the 6th observation and the four other observations are the given original observations $X_{1,}X_{2,}X_3,X_4,X_5$; then the consecutive maximum likelihood function or observed likelihood rate will be

$$L' = f(X_1;\bar{X}^*,S^*)f(X_2;\bar{X}^*,S^2)f(X_3;\bar{X}^*,S^2)f(X_4;\bar{X}^*,S^2)\ f(X_5;\bar{X}^*,S^2)f(RS^* = \boldsymbol{X_6};\bar{X}^*,S^2)$$

$$\log{(L')} = log[f(X_1;\bar{X}^*,S^*)f(X_2;\bar{X}^*,S^2)f(X_3;\bar{X}^*,S^2)f(X_4;\bar{X}^*,S^2)\ f(X_5;\bar{X}^*,S^2)f(RS^* = \boldsymbol{X_6};\bar{X}^*,S^2)]$$

$$\begin{aligned}\log(L') = log(f(X_1;\bar{X}^*,S^*)) + log(f(X_2;\bar{X}^*,S^*)) + log(f(X_3;\bar{X}^*,S^*))\\ + log(f(X_4;\bar{X}^*,S^*)) + log(f(X_5;\bar{X}^*,S^*))\\ + log(f(RS^* = X_6;\bar{X}^*,S^*)))\end{aligned}$$

$$\tfrac{1}{6}\log(L') = \tfrac{1}{6}\textstyle\sum_{i=1}^{6}\log(f(X_i;\bar{X}^*,S^*))$$

We will search the incremented value of the 6th observation for which the expected log likelihood rate and the observed log likelihood rate will be same i.e.

$$\tfrac{1}{6}\log(L') = \tfrac{1}{6}\textstyle\sum_{i=1}^{6}\log(f(X_i;\bar{X}^*,S^*)) \cong \tfrac{1}{5}\log(L) = \tfrac{1}{5}\textstyle\sum_{i=1}^{5}\log(f(X_i;\bar{X}^*,S^*)).$$

The incremented value of the 5th observation for which the likelihood functions are same, will be the estimated value of the missing observations. If we get more than two estimates of the missing observation (since we get two values of the 5th observation for whom the likelihood rates are same), we can check for which estimate of the missing value the first two moments are close to those of the original 4 observations. Hence, we will find the estimate of the missing value.

3. Real Life Examples

We like to simulate a couple of samples each of which is of size n from a probability distribution with specified parameters. Later we will keep one observations a complete missing observation and pull it out from the original sample. Hence the original sample turns to a sample of size $n-2$. Out of $n-2$ available observations of the sample, we will draw samples each of which is of size $n-2$. For each of the $n_{C_{n-2}}$ samples of size $n-2$, we will assume the two absent observations as two dummy missing values of the sample. So, for each of the $n_{C_{n-2}}$ samples, there are $n-2$ available observations and two dummy missing values. From each of the $n_{C_{n-2}}$ samples, we will have one absolute dispersion between the average of $n-2$ available observations and the average of the two dummy missing observations. So, we will have $n_{C_{n-2}}$ absolute between differences for $n_{C_{n-2}}$ pairs of averages and dummy missing values. Averaging the $n_{C_{n-2}}$ absolute differences, we will calculate average absolute difference. Based on the average absolute difference, we will generate a possible range of the original missing value. We will generate several values of that range starting from the lower limit and will get several valued for fixed increment upto to upper limit of that range. We will check whether the average likelihood of the $n-2$ original observations is similar for which n-1th nth observed missing values from the generating range and the $n-2$ observations.

Let $n = 10$. So there are 8 non-missing observations and two missing observations. The non-missing observations from Normal with mean vector and variance covariance matrix

$$(7.788 \quad 170.760 \quad 65.540 \quad 21.232),$$

$$\begin{pmatrix} 18.970465 & 291.0624 & 4.386204 & 22.991412 \\ 291.0624 & 6945.1657 & 312.2751 & 519.2691 \\ 4.386204 & 312.2751 & 209.518776 & 55.768082 \\ 22.991412 & 519.2691 & 55.768082 & 87.72916 \end{pmatrix}$$

are

$$\begin{aligned}
X_1 &= (12.620704 \quad 214.43815 \quad 80.47159 \quad 31.211845)\\
X_2 &= (7.551823 \quad 188.28706 \quad 59.65852 \quad 19.726793)\\
X_3 &= (1.345236 \quad 12.45771 \quad 58.83213 \quad -1.312026)\\
X_4 &= (13.707797 \quad 303.77786 \quad 83.23562 \quad 34.149916)\\
X_5 &= (9.777255 \quad 131.52619 \quad 62.66918 \quad 16.799206)\\
X_6 &= (7.984191 \quad 263.11509 \quad 83.55288 \quad 41.809556)\\
X_7 &= (5.948454 \quad 168.88802 \quad 84.60518 \quad 27.419772)\\
X_8 &= (\ 7.807614 \quad 162.97018 \quad 55.55015 \quad 29.249890)
\end{aligned}$$

and the missing observations are

$$\begin{aligned}
X_9 &= (\ 6.593975 \quad 172.79172 \quad 55.72927 \quad 19.105670)\\
X_{10} &= (10.267980 \quad 160.04332 \quad 59.76657 \quad 13.548488)
\end{aligned}$$

Now, assuming two non-missing observations as two missing ones we can generate 28 samples each of which is consisting of 6 non-missing observations assuming the rest two non-missing observations as two missing observations. So, the 28 samples (as addressed in table 3) each consisting of 6 non-missing values are as below (the bold numbers in the last row are representing here the assumed missing value for each sample):

Table 3: The 28 samples each consisting of 6 non-missing values.

Sample |	*Non-Missing Part*						*Missing Part*	
1	X_3	X_4	X_5	X_6	X_7	X_8	X_1	X_2
2	X_2	X_4	X_5	X_6	X_7	X_8	X_1	X_3
3	X_2	X_3	X_5	X_6	X_7	X_8	X_1	X_4
4	X_2	X_3	X_4	X_6	X_7	X_8	X_1	X_5
5	X_2	X_3	X_4	X_5	X_7	X_8	X_1	X_6
6	X_2	X_3	X_4	X_5	X_6	X_8	X_1	X_7
7	X_2	X_3	X_4	X_5	X_6	X_7	X_1	X_8
8	X_1	X_4	X_5	X_6	X_7	X_8	X_2	X_3
9	X_1	X_3	X_5	X_6	X_7	X_8	X_2	X_4
10	X_1	X_3	X_4	X_6	X_7	X_8	X_2	X_5
11	X_1	X_3	X_4	X_5	X_7	X_8	X_2	X_6
12	X_1	X_3	X_4	X_5	X_6	X_8	X_2	X_7
13	X_1	X_3	X_4	X_5	X_6	X_7	X_2	X_8
14	X_1	X_2	X_5	X_6	X_7	X_8	X_3	X_4
15	X_1	X_2	X_4	X_6	X_7	X_8	X_3	X_5
16	X_1	X_2	X_4	X_5	X_7	X_8	X_3	X_6
17	X_1	X_2	X_4	X_5	X_6	X_8	X_3	X_7
18	X_1	X_2	X_4	X_5	X_6	X_7	X_3	X_8
19	X_1	X_2	X_3	X_6	X_7	X_8	X_4	X_5
20	X_1	X_2	X_3	X_5	X_7	X_8	X_4	X_6
21	X_1	X_2	X_5	X_5	X_6	X_8	X_4	X_7
22	X_1	X_2	X_3	X_5	X_6	X_7	X_4	X_8
23	X_1	X_2	X_3	X_4	X_7	X_8	X_5	X_6
24	X_1	X_2	X_3	X_4	X_6	X_8	X_5	X_7
25	X_1	X_2	X_3	X_4	X_6	X_7	X_5	X_8
26	X_1	X_2	X_3	X_4	X_5	X_8	X_6	X_7
27	X_1	X_2	X_3	X_4	X_5	X_7	X_6	X_8
28	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8

Table 4: The Bandwidth for each of the 28 samples.

Sample #	*Sample Mean*	*Absolute Difference (Euclidean Distance) or Bandwidth*
1	$a_{12} = \frac{X_{3+}X_4 + X_{5+}X_6 + X_7 + X_8}{6}$	$d(a_{12}, m_{12}) = d\left(\frac{X_{3+}X_4 + X_{5+}X_6 + X_7 + X_8}{6}, \frac{X_1 + X_2}{2}\right)$
2	a_{13}	$d(a_{13}, m_{13})$
3	a_{14}	$d(a_{14}, m_{14})$
4	a_{15}	$d(a_{15}, m_{15})$
5	a_{16}	$d(a_{16}, m_{16})$
6	a_{17}	$d(a_{17}, m_{17})$
7	a_{18}	$d(a_{18}, m_{18})$
8	a_{23}	$d(a_{23}, m_{23})$
9	a_{24}	$d(a_{24}, m_{24})$
10	a_{25}	$d(a_{25}, m_{25})$
11	a_{26}	$d(a_{26}, m_{26})$
12	a_{27}	$d(a_{27}, m_{27})$
13	a_{28}	$d(a_{28}, m_{28})$
14	a_{34}	$d(a_{34}, m_{34})$
15	a_{35}	$d(a_{35}, m_{35})$
16	a_{36}	$d(a_{36}, m_{36})$
17	a_{37}	$d(a_{37}, m_{37})$
18	a_{38}	$d(a_{38}, m_{38})$
19	a_{45}	$d(a_{45}, m_{45})$
20	a_{46}	$d(a_{46}, m_{45})$
21	a_{47}	$d(a_{47}, m_{47})$
22	a_{48}	$d(a_{48}, m_{48})$
23	a_{56}	$d(a_{56}, m_{56})$
24	a_{57}	$d(a_{57}, m_{57})$
25	a_{58}	$d(a_{58}, m_{58})$
26	a_{67}	$d(a_{67}, m_{67})$
27	a_{68}	$d(a_{68}, m_{68})$
28	a_{78}	$d(a_{78}, m_{78})$
Mean		AD = $\frac{\sum_{i<j=1}^{8} d(a_{ij}, m_{ij})}{28}$ = 86.68563

The Expected Average Log Likelihood Rate for 9 observations (8 non-missing and one from the generating interval) is -15.11826. By using the formula shown above, we get average distance [AD = $\frac{\sum_{i<j=1}^{8} d(a_{ij}, m_{ij})}{28}$ = 86.68563] from the known vector to the missing vector is 86.68563; where k= $\frac{1}{10000}$. We will get average likelihood rate for 9 observations. And for k= $\frac{1}{10000}$, we get the same value for the Average Likelihood and Observed Average Likelihood. So, our estimated value of the 1st missing observation is $(8.743209 \quad 210.316754 \quad 77.413469 \quad 20.149505)$.

Now depending on the 1st missing value and the missing value based, or 9 observations based mean and variance, the likelihood function and likelihood rate for 10 observations have been found. The Expected Log Likelihood Rate is -15.24454. By using the formula shown above, we get the distance as 68.85106; where k = $\frac{1}{1000000}$. For each increment we will get average likelihood rate for 10 observations (8 non-missing, one estimate of the 1st missing and one from the generating interval for the 2nd missing value). We get the same value for the Expected Average Likelihood and Observed Average Likelihood. So, our estimated value of the 2nd missing observation is (16.33900 218.84061 87.58071 22.81895).

So, the estimates of the two missing values

$$(6.593975 \quad 172.79172 \quad 55.72927 \quad 19.105670),$$
$$(10.267980 \quad 160.04332 \quad 59.76657 \quad 13.548488)$$

are (8.743209 210.316754 77.413469 20.149505)

and (16.33900 218.84061 87.58071 22.81895)

along with the distances 43 and 66 respectively.

Conclusion

The missing technique is a kind of check and balance method in estimating the missing value. In each step it checks the fluctuation due to sample size and balance it by capturing the dispersion of the estimate of the known data from the assumed unknown data which is really known. So, this method is trying to find the original rate of change of the deviation from the missing value for the exact size of the realized sample. So, from two directions, one direction from sample size and other direction for the deviation from the missing values, the missing technique has been aided to estimate the missing value efficiently maintaining a good performance through several goodness of fit tests. This paper also demonstrates a resampling method for generating 1 or 2 correlated observations from the same distribution from where the original sample is drawn. This paper can also be extended to get a resampling method for (n > 2) three or more correlated observations.

Reference

Little. R. J. A, Rubin. D. B. (2002). Statistical Analysis with Missing Data. 2nd edition. Wiley Publishers.

Sharna, S. I., Adnan, M. A. S., and Imon, R. (2016). A Missing Technique for Estimating a missing value. In JSM *Proceedings*, Statistical Computing Section. Alexandria, VA: American Statistical Association. 398 – 409.

Sharna, S. I., Adnan, M. A. S., and Imon, R. (2017). A Missing Technique for Estimating Univariate Multiple Missing Values: An Advanced Resampling Method for Correlated Observations. In JSM *Proceedings*, Statistical Computing Section. Alexandria, VA: American Statistical Association. 2522-2535.