MSR29

A Comparative Analysis of Missing Value Imputation Techniques: Parametric Imputation with eCDF (PIE) vs. Markov Chain Monte Carlo Imputation (MCMC)

Paul Choudhury S[1], Dutta Majumdar A[1], Sil A[1], Dutta SK[1]

[1]PharmaQuant Insights Pvt. Ltd., Kolkata, WB, India

INTRODUCTION

- Missing value imputation is critical to healthcare as it ensures accurate patient diagnosis and treatment planning, leading to comprehensive and trustworthy medical decisions.[1,2]
- It is essential to guarantee the completeness of data and the validity of statistics in healthcare research to achieve precise and reliable results.
- Various well-established methods, such as the Markov Chain Monte Carlo (MCMC) imputation exist to handle missing continuous data.[3,4]
- The MCMC method can be computationally intensive with large datasets, and convergence issues may arise due to poor mixing and high autocorrelation between samples. MCMC methods may struggle with high-dimensional data due to the curse of dimensionality. As the number of parameters in the model increases, it becomes harder to explore the posterior distribution effectively.[5,6]
- Thus, we propose a ***parametric imputation procedure*** that utilizes the available data as an alternative approach.

OBJECTIVES

- **Primary -** To compare the performance of the standard MCMC method with a proposed alternative parametric imputation method (PIE) that uses the empirical cumulative distribution function (eCDF) curve to assess goodness of fit.
- **Secondary -** To evaluate how parametric uncertainty impacts the effectiveness of the PIE imputation technique.

METHODS

Data preparation:

- Data on diastolic blood pressure (DBP) and 2-hour serum insulin levels (SIL) for 768 patients were selected for analysis from a publicly available dataset (Smith 1988[7]).
- We created a missing value dataset (MVD) by randomly ***removing 10% of the observations*** from the primary dataset.

PIE imputation procedure:

- Different parametric distributions (Normal, Cauchy, Log-normal, Gamma, Weibull, Exponential, Pareto, etc.) were fitted on MVD. Model fit was compared using Akaike's Information Criteria (AIC), Bayesian Information Criterion (BIC), eCDF curves, and probability density plots.
- We imputed each missing value by calculating the mean from one hundred samples drawn from the best-fitting distribution.

MCMC imputation procedure:

- The MCMC imputation method was used on MVD to create five different datasets by randomly selecting five different seeds.
- We replaced each missing value with the mean of one hundred samples drawn from the posterior distribution.

Comparison between two techniques:

- Root-mean-square error (RMSE) was estimated to compare the accuracy of the two techniques.

Sensitivity analysis for PIE imputation procedure:

- Sensitivity analysis was performed using bootstrap sampling[8] with 1,000 iterations to check for parametric uncertainty in the PIE imputation technique and the probable effect on the imputed data set was then summarized using RMSE values.

RESULTS

DBP data | **SIL data**

Table 1: *Summary of model fit*

DBP data

Distribution	AIC	BIC
Log-normal	4698.8	4707.9
Gamma	4702.6	4719.8
Log-logistic	4718.8	4727.9
Normal	4732.2	4741.3
Logistic	4735.1	4744.2
Weibull	4795.5	4804.6
Inverse weibull	4814.1	4823.2
Cauchy	4979.2	4988.3
Exponential	6332.1	6336.7
Pareto	6347.9	6357.0

SIL data

Distribution	AIC	BIC
Log-normal	4051.2	4058.9
Log-logistic	4062.7	4069.4
Gamma	4068.8	4076.4
Weibull	4094.3	4102.0
Inverse weibull	4128.0	4135.7
Exponential	4166.9	4170.8
Pareto	4187.1	4194.8
Logistic	4199.2	4206.9
Cauchy	4200.0	4207.7
Normal	4274.3	4282.0

Figure 1: *Empirical and theoretical CDFs*

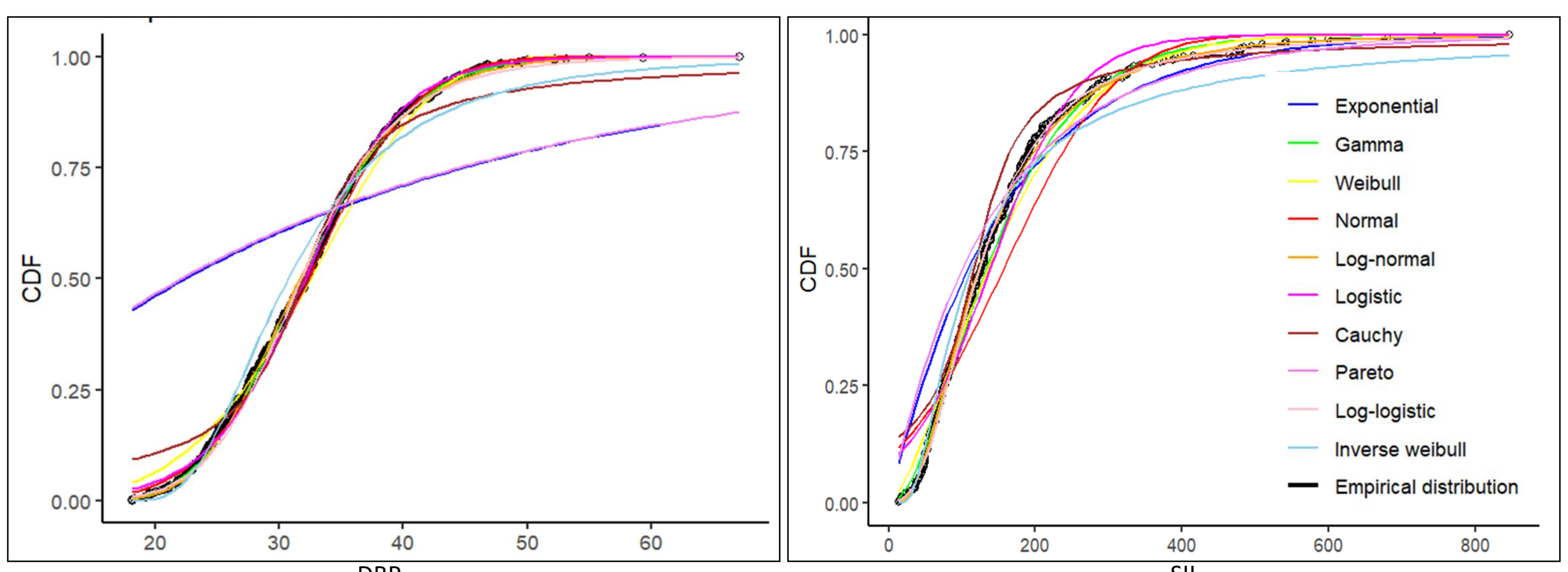
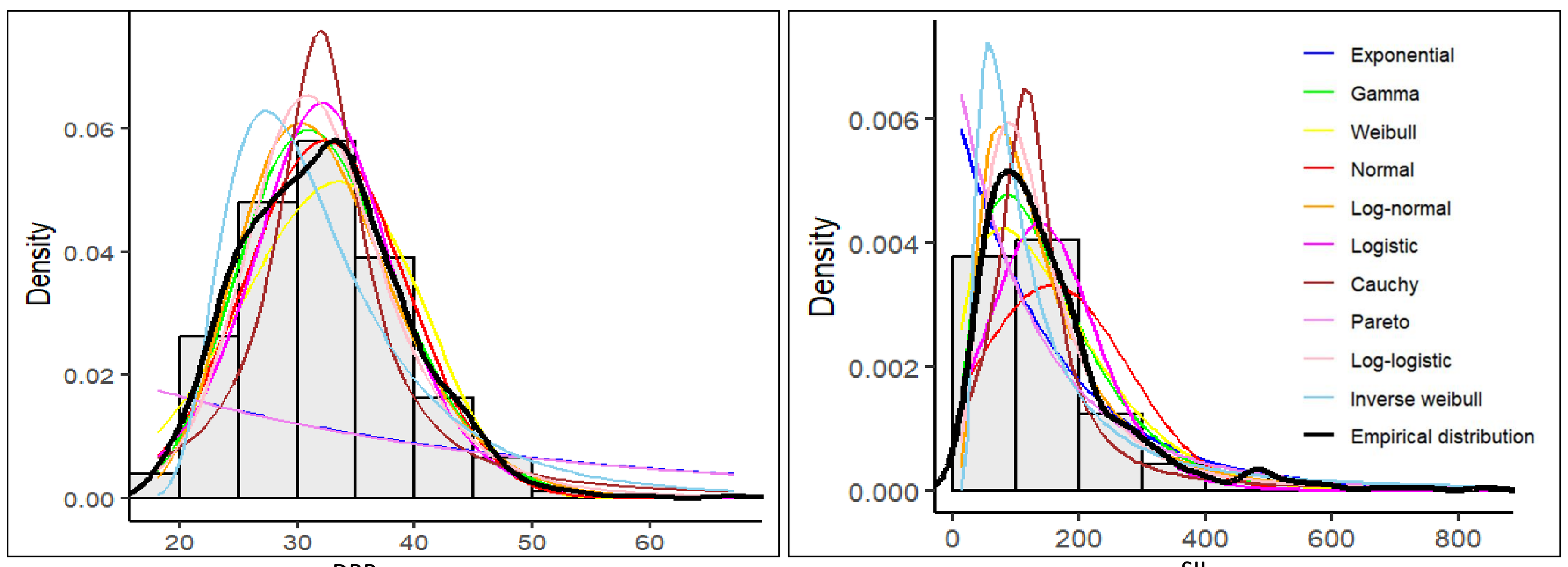


Figure 2: *Histogram and theoretical densities*



- **Log-normal distribution** was the best-fitting distribution for both scenarios with the lowest AIC & BIC **(Table 1)**.
- For visual examination, eCDF plots **(Figure 1)** and density plots **(Figure 2)** were used.
- Both plots demonstrated that the log-normal distribution closely approximated the empirical data.
- Hence log-normal distribution was chosen for imputation purposes for both scenarios.

DBP data | **SIL data**

Table 4: *Comparison between PIE and MCMC imputation procedures*

DBP data

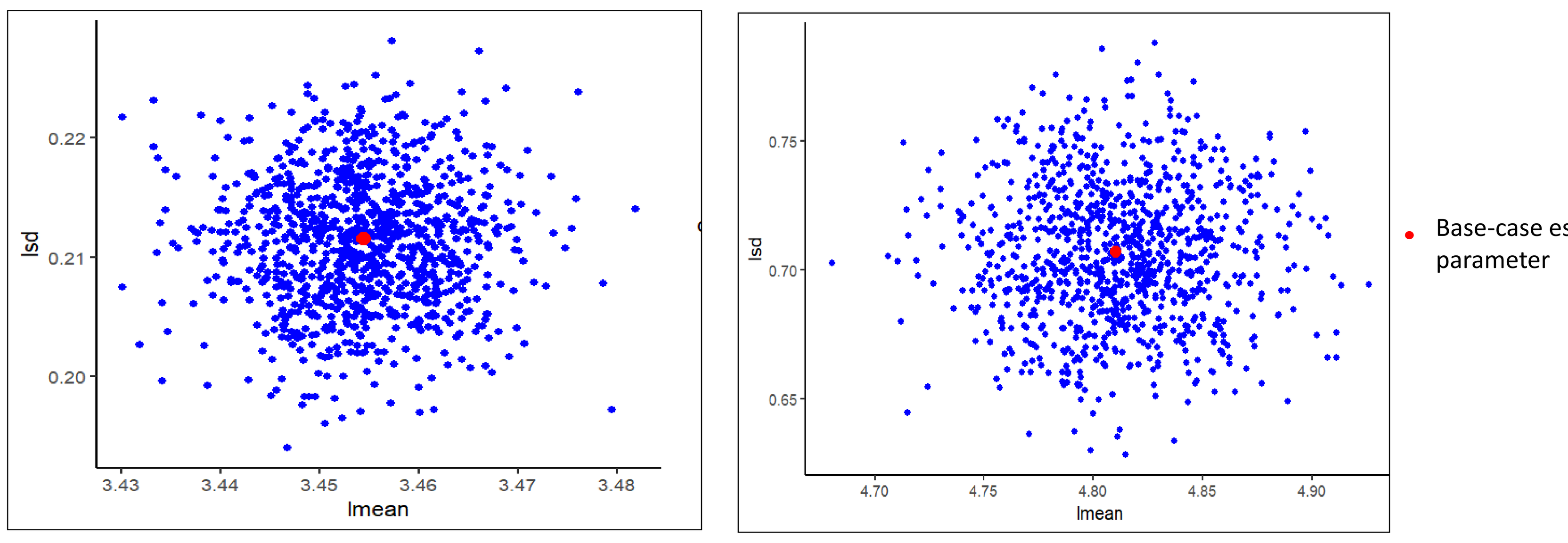
Analysis type		RMSE	Mean
Original data		-	32.5
Parametric Imputation with eCDF (PIE)		7.9	32.4
MCMC (Seeds)	2379392	9.8	33.6
	873287284	12.4	32.9
	144252	11.9	33.1
	123456	9.7	34.0
	88834883	9.6	32.9

SIL data

Analysis type		RMSE	Mean
Original data		-	155.6
Parametric Imputation with eCDF (PIE)		108.1	156.7
MCMC	2379392	196.6	166.8
	873287284	199.7	132.0
	144252	201.1	165.0
	123456	209.5	166.8
	88834883	207.8	158.1

- Root Mean Square Errors (RMSEs) were compared between the PIE and MCMC methods for both DBP and SIL datasets. RMSE values were lower when using the PIE method than the MCMC method for both datasets.
- For both the variables, PIE outperformed MCMC imputation when compared with the original data based on means difference **(Table 4).**

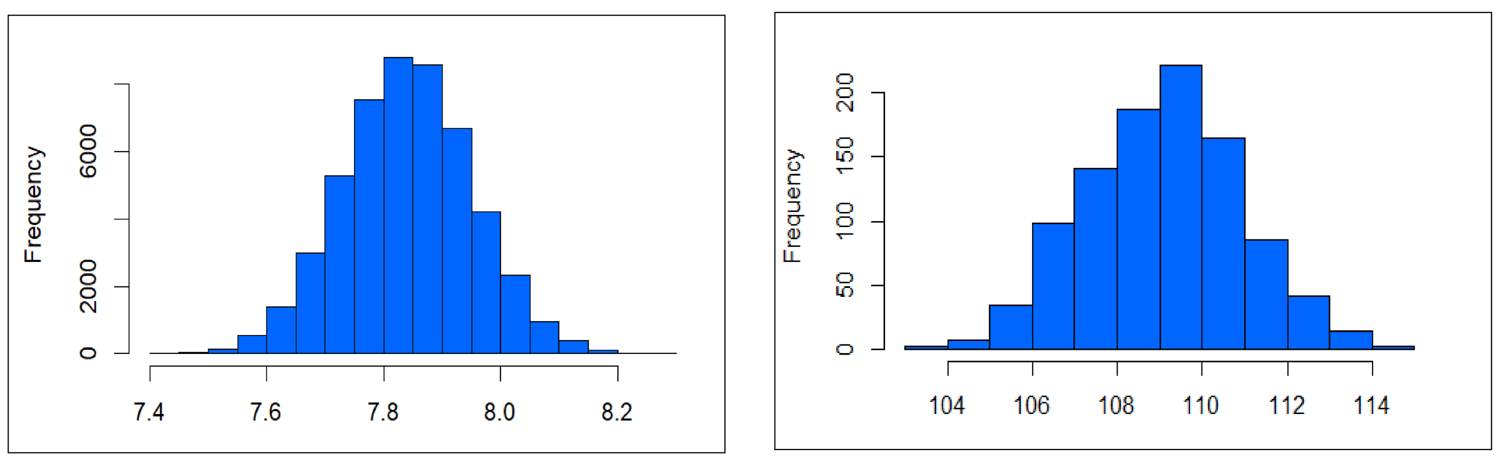
Figure 3: *Plot showing parametric uncertainty around model parameters*



lmean: Log(mean); lsd: Log(sd)

- **After conducting bootstrap sampling:**
  - Minimal variation was observed in the estimated parameters of the log-normal distribution.
- **For the DBP dataset:**
  - The mean of the log-normal distribution falls between 3.43 and 3.48.
  - The standard deviation ranges from **0.19 to 0.23. (Figure 3)**.
- **For the SIL dataset:**
  - The mean of the log-normal distribution falls between 4.68 and 4.92.
  - The standard deviation ranges from **0.63 to 0.79 (Figure 3)**.
- Missing values were imputed for each of the 1,000 pairs of parameters estimated from the iterations.
- Finally, RMSEs were calculated for all the imputed datasets.

Figure 4: *Plot showing uncertainty around RMSE values*



- The RMSE ranges obtained from bootstrapped sampling were **7.5 to 8.2** for the *DBP dataset* and **103.1 to 120.8** for the *SIL dataset* **(Figure 4).** These ranges were found to be lower than the RMSE values calculated using the MCMC method. This suggests that PIE performed better in terms of accuracy or error measurement when compared to MCMC for these two variables, DBP and SIL.

CONCLUSION

The results of this comparative analysis suggest that PIE imputation outperforms MCMC imputation in terms of RMSE, indicating its potential as a viable alternative.

REFERENCES

1. Li J, Yan XS, Chaudhary D, Avula V, Mudiganti S, Husby H, et al. Imputation of missing values for electronic health record laboratory data.npj Digital Medicine. 2021;4(1):147
2. Liu M, Li S, Yuan H, Ong MEH, Ning Y, Xie F, et al. Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. Artificial Intelligence in Medicine. 2023;142:102587.
3. Schunk D. A Markov chain Monte Carlo algorithm for multiple imputation in large surveys. AStA Advances in Statistical Analysis. 2008;92(1):101-14.
4. Harel O, Zhou XH. Multiple imputation: review of theory, implementation and software. Statistics in medicine. 2007 Jul 20;26(16):3057-77.
5. Brooks, S., & Gelman, A. (1998). "General methods for monitoring convergence of iterative simulations." Journal of Computational and Graphical Statistics, 7(4), 434-455.
6. Robert CP. The Bayesian choice: from decision-theoretic foundations to computational implementation. New York: Springer; 2007 Jun.
7. Smith JW, Everhart JE, Dickson W, Knowler WC, Johannes RS, editors. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proceedings of the annual symposium on computer application in medical care; 1988: American Medical Informatics Association.
8. Davison AC, Hinkley DV. Bootstrap methods and their application. Cambridge university press; 1997 Oct 28.

PharmaQuant

This study was funded by PharmaQuant Insights Pvt. Ltd., India

linkedin.com/company/pharmaquant

www.pharmaquant.org