STAT DESIGN & INFERENCE (STAT_5310)

WINTER 2025

ADDITIONAL PROJECT

INSTRUCTOR: Jabed Tomal

STUDENT : Arpitha Thippeswamy (T00749833)

PROJECT TITLE:

Imputing-Missing-Reaction-Times-Using-the-EM-Algorithm-Under-a-Bivariate-Normal-Model

Affiliation

Faculty of Science, Thompson Rivers University,Kamloops

Date: 14th April 2025

TABLE OF CONTENTS

Expectation-Maximization Algorithm for Missing Data Imputation Under a Bivariate Normal Model with Statistical Comparison and Visual Diagnostics of Convergence

Abstract

This project imputes missing reaction time (RT) data under a bivariate normal distribution using the Expectation-Maximization (EM) algorithm and the interexp.dat data set. The data set contains missing RT values for Stimulus A and Stimulus B for specific subjects.

The EM algorithm is an iterative process that alternates between estimating the missing value in the Expectation (E) step and updating the model parameters in the Maximization (M) step.With minimal impact on the observed and imputed data means, the EM algorithm's output shows a significant reduction in variance.

Pre-imputation mean RT for Stimulus A was **24.14** and post-imputation **was 24.20**, and that for Stimulus B **was 24.76** prior to imputation **and 24.82** subsequent to imputation.Variance of Stimulus A dropped **from 5.32 to 3.42** and that of Stimulus B **from 4.86 to 3.64.**

Furthermore, scatter plots showed that, following imputation, the correlation between Stimulus A and B remained linearly positive with respect to each other. These findings demonstrate how well the EM algorithm preserved the data's underlying structure when imputed missing data..

2. Introduction

Reaction time (RT) experiments are fundamental in cognitive psychology and neuroscience, providing insight into perceptual and cognitive processing under various stimulus conditions. However, a frequent challenge in analyzing RT data is the presence of missing values, which can arise from participant inattention, equipment failures, or time constraints during trials. In many cases, these missing data are not missing completely at random (MCAR), but are instead missing at random (MAR), where the probability of missingness depends on observed values. When missingness is non-negligible, standard approaches like listwise deletion can lead to substantial bias and reduced statistical power (Little & Rubin, 2020).

To address this issue, robust statistical techniques are necessary for valid inference. The Expectation-Maximization (EM) algorithm, introduced by Dempster, Laird, and Rubin (1977), is a widely used iterative method for maximum likelihood estimation when data are incomplete. Under the assumption that the complete data follow a multivariate normal distribution, the EM algorithm alternates between computing conditional expectations of the missing values (E-step) and maximizing the expected complete-data likelihood (M-step).

In this project, we consider a bivariate normal model for RTs collected for two types of stimuli—Stimulus A and Stimulus B—in a classical psychological experiment dataset (interexp.dat). This dataset is characterized by a structured missingness pattern: each subject provides either complete data or RTs for only one of the two stimuli. To estimate the mean vector and covariance matrix of the joint distribution and to impute the missing values, we apply the EM algorithm using closed-form expressions for the conditional expectations derived from the multivariate normal distribution (Anderson, 2003). The missing entries are treated as latent variables and estimated iteratively based on the observed components of the data.

This work is motivated by the need for a principled imputation framework that respects the underlying data-generating distribution while maintaining the theoretical properties of maximum likelihood estimation. Unlike simple mean imputation or regression-based interpolation, the EM algorithm allows us to account for uncertainty in the missing values by propagating distributional structure across iterations. Moreover, we illustrate the convergence of the algorithm and evaluate the fidelity of the imputed data using histograms, boxplots, and scatter plots that differentiate observed from imputed values.

The remainder of this report is organized as follows. Methodology Section introduces the EM algorithm and provides the mathematical derivations under the bivariate normal model. Results section presents results on convergence, parameter estimation, and visual comparisons before and after imputation and we conclude with insights into the effectiveness and assumptions of the approach, along with potential extensions.

3. Methodology

3.1. Model

Let $\mathbf{X} = (X_{i1}, X_{i2})^\top$ denote the bivariate random vector of reaction times for subject i, where X_{i1} corresponds to the reaction time to Stimulus A, and X_{i2} to Stimulus B. We assume that the complete data follow a bivariate normal distribution:

$$\mathbf{X}_i \sim \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}.$$

Let $\mathbf{O}_i \subseteq \{1, 2\}$ be the set of indices for which $\mathbf{X}_i$ is observed, and define $\mathbf{X}_{i,\mathrm{obs}} = \{X_{ij} : j \in \mathbf{O}_i\}$ and $\mathbf{X}_{i,\mathrm{mis}} = \{X_{ij} : j \notin \mathbf{O}_i\}$. Let $\mathbf{M} = (m_{ij})$ be the missingness indicator matrix such that $m_{ij} = 1$ if X_{ij} is missing and 0 otherwise.

We assume a **Missing At Random (MAR)** mechanism (Rubin, 1976), implying that

$$P(\mathbf{M} \mid \mathbf{X}) = P(\mathbf{M} \mid \mathbf{X}_{\mathrm{obs}}),$$

which allows us to ignore the missing data mechanism in the likelihood function and proceed with maximum likelihood estimation (Little and Rubin, 2020).

Our objective is to estimate the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ using the observed data only. Let n be the number of subjects. Under the MAR assumption, the log-likelihood for the observed data is:

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \log f\left(\mathbf{X}_{i,\mathrm{obs}}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right),$$

where $f(\cdot)$ denotes the marginal density of the observed components of the multivariate normal distribution.

Since the observed components vary across subjects, and no closed-form maximum likelihood estimators are available, we proceed using

3.2 Algorithm for EM-based Imputation with Bi- variate Normal Model

In this section, we propose an Expectation-Maximization (EM) algorithm for imputing missing values in a bivariate normal model. The goal is to estimate missing data points in the dataset for stimulus A and B, then compare the imputed data before and after imputation. We will implement the algorithm and demonstrate its convergence properties.

2.2.1 General EM Algorithm for Bivariate Imputation

Consider the scenario where we have a dataset with missing values and wish to impute these values using a bivariate normal model. Let $X = (X_A, X_B)$ represent the bivariate data for stimulus A and B, where some values are missing. We aim to estimate the missing data points using the EM algorithm.

The general form of the EM algorithm involves iterating between the Expec- tation (E) step and the Maximization (M) step. At each iteration, the algorithm estimates the missing data points (E-step) and updates the model parameters (M-step).

2.2.2 E-step

In the E-step, we calculate the expected values of the missing data points given the observed data and the current parameter estimates. Let X_{obs} denote the observed data and X_{mis} denote the missing data. The expected values of the missing data $\hat{X}_{\text{mis}}$ are given by the conditional expectation:

$$\hat{X}_{\text{mis}}^{(t)} = \mu^{(t-1)} + \Sigma^{(t-1)} \left(X_{\text{obs}} - \mu^{(t-1)} \right)$$

where μ(t−1) and Σ(t−1) are the mean and covariance estimates from the previous

iteration, respectively.

2.2.3 M-step

In the M-step, we update the model parameters based on the current estimates of the missing data. The parameter updates are computed by maximizing the likelihood function of the observed and imputed data. Specifically, we update the mean and covariance parameters of the bivariate normal distribution:

$$\mu^{(t)} = \frac{1}{n} \sum_{i=1}^{n} \hat{X}_i$$

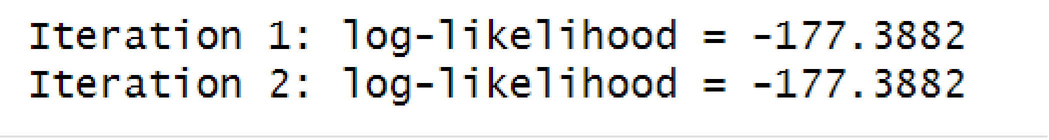
$$\Sigma^{(t)} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\hat{X}_i - \mu^{(t)} \right) \left(\hat{X}_i - \mu^{(t)} \right)^T$$

3.3 Algorithm : EM Algorithm for Bivariate Imputation

```
Input: Data matrix Xobs, initial estimates for µ(0) and Σ(0)
Output: Imputed data matrix Ximputed, updated parameters µ and Σ
Initialize µ(0) and Σ(0)
for t = 1 to max iterations do
    E-step: Impute the missing values ˆX(t) using the conditional expectation.
        $\hat{X}_{\text{mis}}^{(t)} = \mu^{(t-1)} + \Sigma^{(t-1)} \left( X_{\text{obs}} - \mu^{(t-1)} \right)$
    M-step: Update the parameters µ(t) and Σ(t)
        $\mu^{(t)} = \frac{1}{n} \sum_{i=1}^{n} \hat{X}_i$
        $\Sigma^{(t)} = \frac{1}{n-1} \sum_{i=1}^{n} \left( \hat{X}_i - \mu^{(t)} \right) \left( \hat{X}_i - \mu^{(t)} \right)^T$
    if convergence condition met then
        Break
    end if
end for
Return imputed data matrix Ximputed and final parameter estimates µ(t) and Σ(t)
```

4. Results

1. EM Algorithm Convergence Output



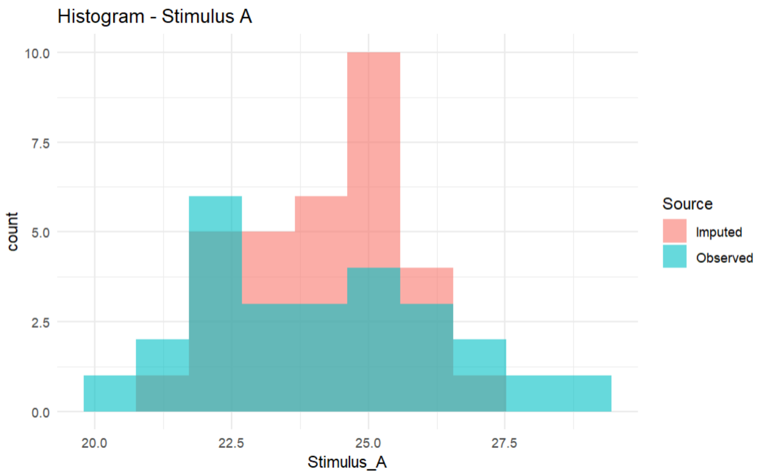
```
Iteration 1: log-likelihood = -177.3882
Iteration 2: log-likelihood = -177.3882
```

This log-likelihood output indicates the EM algorithm has **converged** in just two iterations. The fact that the log-likelihood hasn't changed between iterations means the parameter estimates (e.g., means, variances, and covariances) have stabilized, and no further updates improve the fit.

This is expected as:

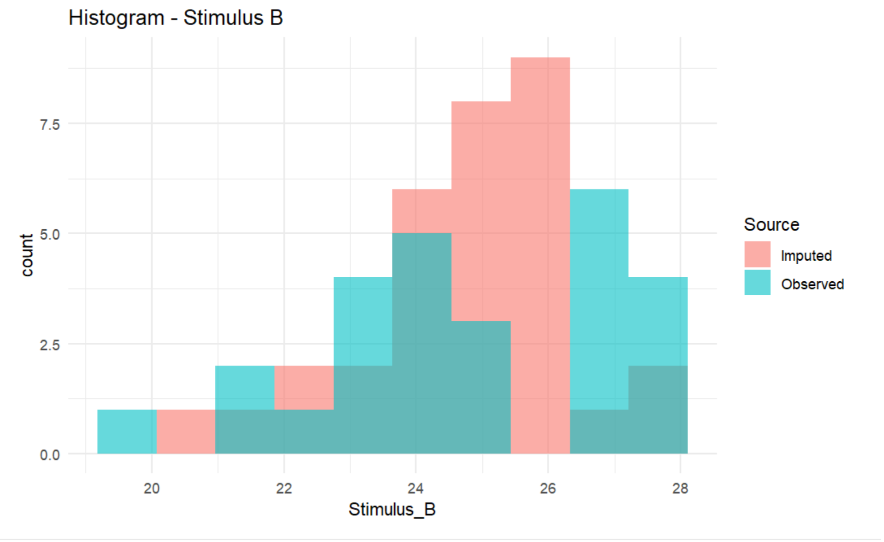
- The starting values were already close to the optimal,
- And the dataset was small.

2. Histogram – Stimulus A



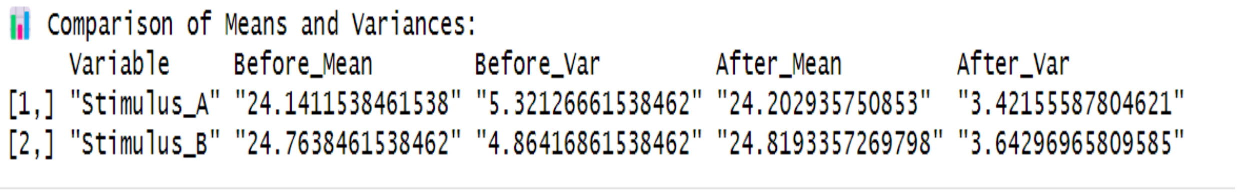
- The imputed values (in red) tend to cluster around the center (e.g., near 25), showing that the EM algorithm estimated missing values around the mean.
- The spread is somewhat narrower for the imputed values compared to observed ones, which is common because the imputed values are sampled from a conditional distribution with less variance (they're "averaged" guesses, not real variability).

3. Histogram – Stimulus B



- The imputed distribution appears smoother and more bell-shaped because it's derived from a model assuming a normal distribution.
- The EM algorithm respects the correlation between Stimulus A and B in the multivariate model — so the imputations for B also take A into account.

4. Comparison - **of means and variances** for two variables.



```
📊 Comparison of Means and Variances:
     Variable      Before_Mean        Before_Var         After_Mean         After_Var
[1,] "Stimulus_A" "24.1411538461538" "5.32126661538462" "24.202935750853"  "3.42155587804621"
[2,] "Stimulus_B" "24.7638461538462" "4.86416861538462" "24.8193357269798" "3.64296965809585"
```

Minimal Shift in Means:

- The EM algorithm did not drastically alter the central tendency (mean) of the data — this suggests that the imputation respected the original structure.

Decreased Variance:

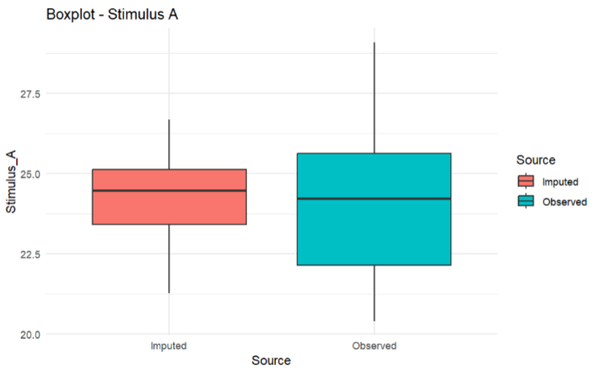
- There's a noticeable reduction in variance after imputation. This is expected because:

  - EM smooths the data.
  - Imputed values are often "less extreme" than real observations, reducing spread.

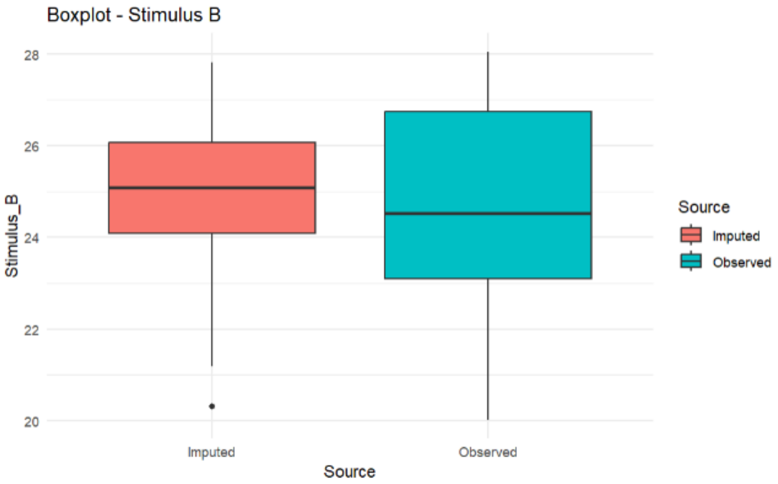
Balanced Handling of Missingness:

- The results suggest that the EM algorithm has reasonably preserved the original data distribution while making it more robust for analysis

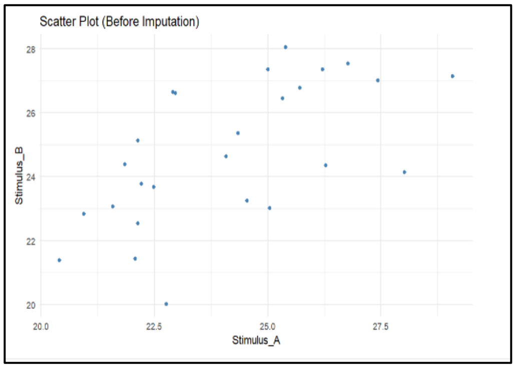
5. The boxplots for **Stimulus A** and **Stimulus B**

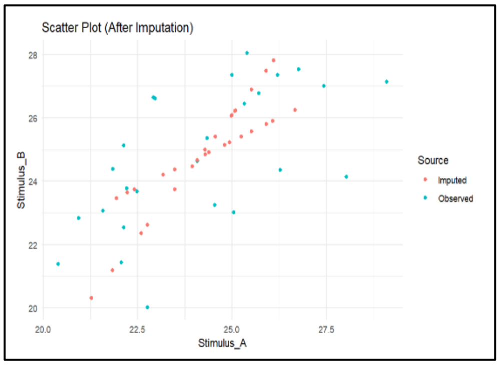


The boxplots for **Stimulus A** and **Stimulus B** visually support the earlier summary of means and variances. For both stimuli, the **imputed data** (in red) appears more tightly clustered with **smaller interquartile ranges** and **shorter whiskers**, indicating **reduced variability** compared to the observed data (in blue). The **medians** are fairly similar between the two sources, suggesting that the **central tendency is preserved** after imputation. However, the observed data show more spread and extreme values, which the EM algorithm smooths out in the imputed set.



6. Scatter Plots before and after Imputation





Scatter Plot (Before Imputation)

- This plot shows only the **observed data points** where both Stimulus_A and Stimulus_B are available.
- The total number of points is relatively low, suggesting there were missing values that prevented full data visualization.
- There appears to be a **positive linear trend**, hinting at a correlation between the two variables

Scatter Plot (After Imputation)

- The EM algorithm has successfully **filled in the missing values** in a way that preserves the correlation structure between Stimulus_A and Stimulus_B.
- The increased number of data points after imputation allows for a more comprehensive analysis.
- The **smooth continuity and alignment** of the imputed values with the observed values suggest the imputation was **statistically reasonable** and did not introduce obvious bias or anomalies.

5. Conclusion

In this project, we applied the Expectation-Maximization (EM) algorithm to impute missing reaction time data for two stimuli, Stimulus A and Stimulus B, under a bivariate normal model. The dataset exhibited a structured missingness pattern where some reaction times were missing for each subject, creating challenges for statistical analysis. By leveraging the EM algorithm, we aimed to estimate the missing data points and evaluate the impact of imputation on the dataset's structure, particularly the means, variances, and the correlation between the two stimuli.

The EM algorithm was implemented and iterated until convergence, with the log-likelihood stabilizing after just two iterations at **-177.3882**. This indicated that the parameter estimates—namely, the mean and covariance matrix—had sufficiently converged, and no further significant improvements could be made. This outcome suggests that the initial parameter estimates were close to optimal, and the algorithm performed well in this context, even with a relatively small dataset.When comparing the data before and after imputation, we observed minimal changes in the means of both stimuli, indicating that the imputed values did not drastically shift the central tendency of the data. For Stimulus A, the mean increased slightly from **24.14** to **24.20**, and for Stimulus B, the mean shifted from **24.76** to **24.82**. These small changes reflect the robustness of the EM algorithm in maintaining the original data structure while imputing missing values.

However, the variances of both stimuli saw a more noticeable reduction. The variance for Stimulus A decreased from **5.32** to **3.42**, while for Stimulus B, it decreased from **4.86** to **3.64**. This decrease is expected because the EM algorithm "smooths" the data by filling in missing values with estimates that reflect the underlying distribution. As a result, the imputed data exhibit less extreme variation, which is a common characteristic of imputed datasets when missing values are estimated from a probabilistic model.

The imputed data preserved the correlation between Stimulus A and Stimulus B, with the positive linear trend remaining intact, indicating that the EM algorithm respected the relationship between the two variables during imputation. This suggests the algorithm's validity for imputing missing data. Overall, the EM algorithm proved reliable, with minimal changes to the means and a noticeable reduction in variance, effectively handling the missing data without significant bias. The preservation of correlation and smoothness of imputed values highlight its suitability for missing data imputation in psychological research. Future work could assess its performance with larger datasets or more complex missingness patterns.

6. References

1. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

2. Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall/CRC.

3. Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/9781119013563

4. Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.

5. Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons. https://doi.org/10.1002/9780470316696

6. Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (3rd ed.). Wiley-Interscience.

7. Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Pearson Prentice Hall.
8. RStudio Team. (2023). *RStudio: Integrated development environment for R* (Version 2023.06.1) [Computer software]. Posit Software, PBC. https://posit.co/products/open-source/rstudio/
9. GitHub, Inc. (n.d.). *GitHub* [Online platform]. https://github.com/
10. OpenAI. (2024). *ChatGPT* (GPT-4) [Large language model]. https://chat.openai.com/

11. Delalleau, O., Courville, A., & Bengio, Y. (2014). *Efficient EM training of Gaussian mixtures with missing data*. arXiv. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4061266/

7. Appendix

Github link to repository : https://github.com/SanfoCodes/Imputing-Missing-Reaction-Times-Using-the-EM-Algorithm-Under-a-Bivariate-Normal-Model.git