Using the EM Algorithm to Impute Missing Values and Compare Stimuli A and B

Haoran He ID: T00749480

April 2025

1 Abstract

This project explored an unfinished experiment that compared human reaction times under two stimuli, A and B. The data set consisted of 26 complete responses, 15 to stimulus A only and 17 to stimulus B only. To recover missing values, we applied the expectation maximization (EM) algorithm under the assumption of a bivariate normal distribution. After filling in the missing values, we analyzed the complete data set to assess whether the difference in reaction times between stimulus A and stimulus B was statistically significant. A paired t-test showed that responses under Stimulus A were significantly faster than under Stimulus B. These results suggest that Stimulus A may lead to faster reaction times.

2 Introduction

This project uses data from an experiment that was stopped before completion. The goal of the experiment was to compare reaction times under two different conditions: Stimulus A and Stimulus B. Each subject was supposed to be tested under both conditions, but due

to interruption, only 26 subjects have complete data. An additional 15 subjects were tested with stimulus A only and 17 subjects were tested with stimulus B only.

The objective of this project is to handle the missing data and compare the two stimuli fairly. We use the Expectation-Maximization (EM) algorithm to estimate the missing values. This method is based on the assumption that the two variables (reaction times under A and B) follow a bivariate normal distribution.

After filling in the missing values, we compare the reaction times under both stimuli and test whether the difference is statistically significant.

3 Method

We assume that the reaction times under stimuli A anda bivariate normal distribution:

$$\begin{pmatrix} y_A \\ y_B \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \begin{pmatrix} \sigma_A^2 & \rho\sigma_A\sigma_B \\ \rho\sigma_A\sigma_B & \sigma_B^2 \end{pmatrix} \right)$$

Let $\mu = (\mu_A, \mu_B)$ and Σ be the 2×2 covariance matrix. For subjects with a missing value (either y_A or y_B), we estimate the missing entry using the conditional expectation of the current parameters.

Applying the Expectation-Maximization (EM) algorithm to estimate the missing values:

- **E-step:** For each missing value, compute the conditional expectation using the current estimates of μ and Σ.

  For example, if y_A is missing:

$$\mathbb{E}[y_A \mid y_B] = \mu_A + \frac{\sigma_{AB}}{\sigma_B^2}(y_B - \mu_B)$$

If y_B is missing:

$$\mathbb{E}[y_B \mid y_A] = \mu_B + \frac{\sigma_{AB}}{\sigma_A^2}(y_A - \mu_A)$$

- **M-step:** Update μ and Σ using the filled-in data:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} y_{Ai} \\ y_{Bi} \end{pmatrix}, \quad \hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mu})(\mathbf{y}_i - \hat{\mu})^T$$

- Repeat until convergence (change in μ and Σ is small).

4 Results

First, data using only the 26 subjects with complete responses were examined. A scatterplot of Stimulus A versus Stimulus B (Figure 1) shows a positive correlation between the two conditions.

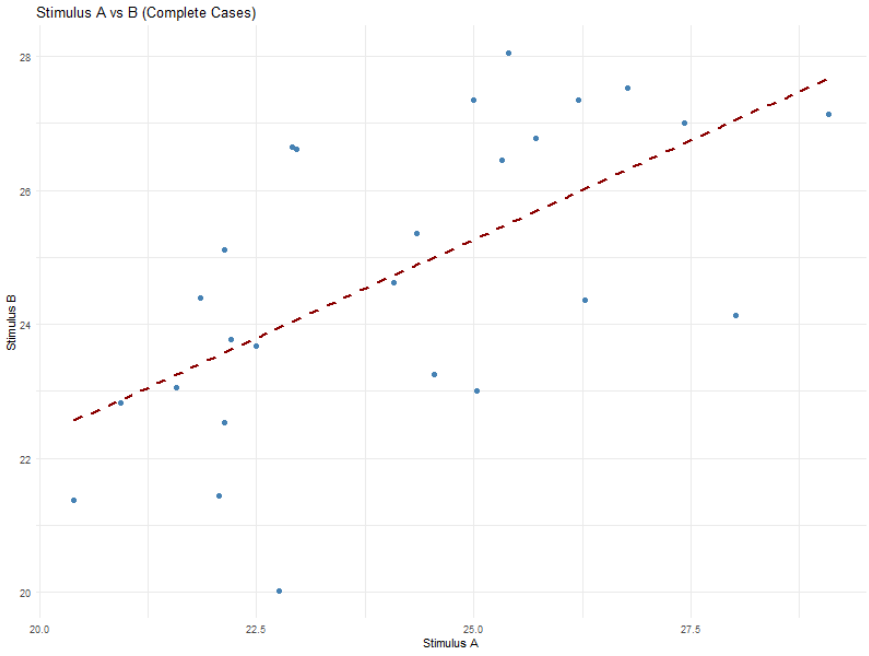

Figure 1: Reaction times under Stimulus A and B (complete cases only)

To recover missing data, we applied the EM algorithm under the assumption of a bivariate normal distribution. The algorithm converged after a few iterations. The estimated mean vector and covariance matrix after convergence were:

$$\hat{\mu} = \begin{pmatrix} 24.21 \\ 24.83 \end{pmatrix}, \quad \hat{\Sigma} = \begin{pmatrix} 3.576 & 2.756 \\ 2.756 & 3.776 \end{pmatrix}$$

After imputation, the completed dataset contains 58 subjects. A new scatterplot (Figure 2) shows the complete data including the imputed values.

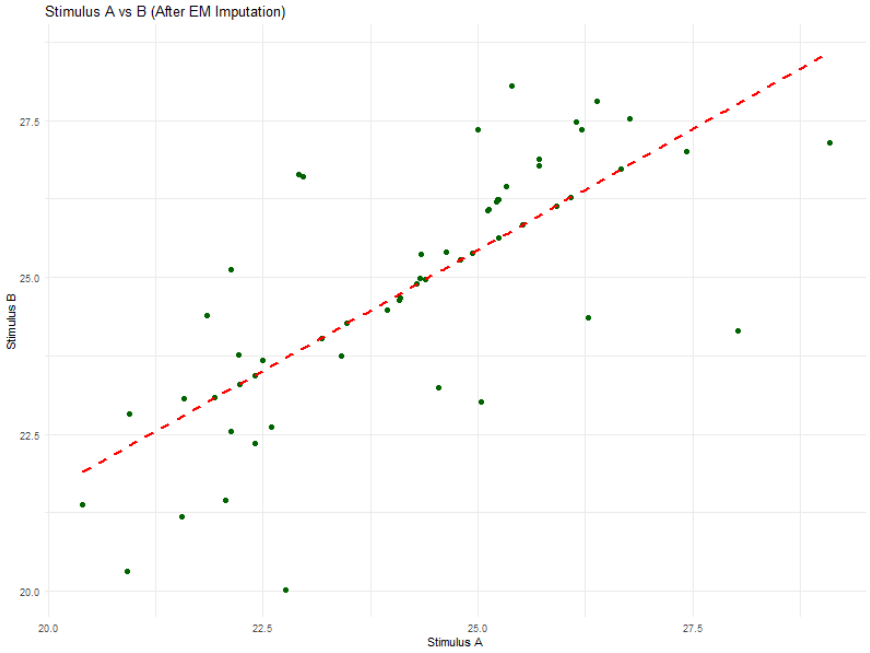

Figure 2: Reaction times under Stimulus A and B (after EM imputation)

A paired t-test was performed on the completed data to test for a difference between the two stimuli. The results were as follows:

- $t = -3.47$
- $p = 0.001$

- 95% CI for $\mu_A - \mu_B$: [-0.97, -0.26]
- Mean difference: -0.62

This suggests that on average, subjects responded faster to Stimulus A than to Stimulus B. The difference is statistically significant.

5 Conclusion

This project used the EM algorithm to handle missing data in a reaction time experiment. By assuming a bivariate normal model for the responses under Stimulus A and B, we were able to estimate missing values and recover a full dataset.

After imputation, a paired t-test showed that reaction times under Stimulus A were significantly faster than those under Stimulus B. The average difference (A minus B) was -0.62, and the 95% confidence interval for the difference was $[-0.97, -0.26]$. Since the entire interval is below zero, conclude that Stimulus A led to faster responses than Stimulus B.

The EM method allowed us to make use of all available data, not just the complete cases. This approach is useful when data are missing at random and the joint distribution can be reasonably modeled.

References

[1] C. Bee, *The EM Algorithm Explained*, Medium, 2018. Available at: `https://medium.com/@chloebee/the-em-algorithm-explained-52182dbb19d9`

[2] S. Lauritzen, *Fast and Stable EM Algorithm for Gaussian Mixture Models*, Department of Statistics, University of Oxford. Available at: `https://www.stats.ox.ac.uk/~steffen/teaching/fsmHT07/fsm407c.pdf`

Appendix

Note: The EM algorithm maximizes the expected complete-data log-likelihood. In our case, since the bivariate normal has closed-form conditional expectations, we can directly use them to impute missing values at each E-step.

The full R code used in this project is available at the following GitHub repository:

`https://github.com/Andyhhr7/STAT5310-Additional-Project.git`