Daniel-Cristian-Marian Țăpuși

Knowledge representation and reasoning – HW2

Part I

Q1: Consider a dataset with two variables X and Y, where some values of Y are missing. Explain how the EM algorithm can be used to estimate the mean and variance of Y and the correlation between X and Y. Why is the log-likelihood expected to increase after each iteration?

The EM algorithm is commonly used for estimation of parameters in models where data is incomplete, that can be seen as latent variables (unobserved data that we want to infer). In our problem, the missing values of Y should be treated as latent variables, where the EM algorithm helps us estimate the parameters of the model that we need, such as the mean (μ_Y) and variance (σ_Y^2) of Y, plus the correlation (ρ_{XY}) between X and Y.

The EM algorithm treats our problem similarly to the one that appears in the context of mixture distributions, where data is assumed to come from a combination of multiple groups, each with its own mean and variance. The solution is to think of the missing values of Y as of latent variables, where the EM algorithm can run iteratively to estimate these values, while updating the parameters of the model.

The first step is to estimate the expected values of the missing data (E-step). This is done using the conditional expectation $E[Y|X]$ to fill the missing values of Y, thanks to the observed values of X and the current values of the parameters (μ_Y, σ_Y^2, ρ_{XY}).

$$E\,[Y \mid X] \;=\; \mu_Y + \rho_{XY} \cdot \frac{\sigma_Y}{\sigma_X}(X \,-\, \mu_X)$$

And then the parameters are updated based on the observed data and these estimates.

$$\mu_Y \;=\; \frac{1}{N}\left(\sum_{observed\,Y}^{Y} \quad + \sum_{mis\sin g\,Y} E[Y|X|]\right)\sigma_Y^2$$

$$= \frac{1}{N}\left(\sum_{observed\,Y} (Y \,-\, \mu_Y)^2 + \sum_{mis\sin g\,Y} (E[Y|X|] \,-\, \mu_Y)^2\right)$$

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_x \sigma_Y}$$

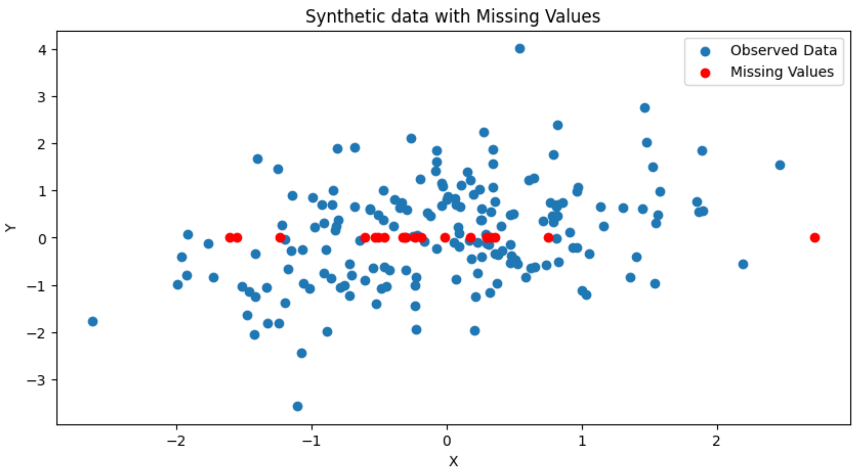
The algorithm iterates between these two steps until it converges to a maximum likelihood estimate. The E-step calculates the expectation of the log-likelihood, which serves as a lower bound for the true log-likelihood function. The M-step then maximizes this expectation. This process ensures that the log-likelihood either increases or remains constant at each iteration, thus guaranteeing convergence to a (local) maximum.

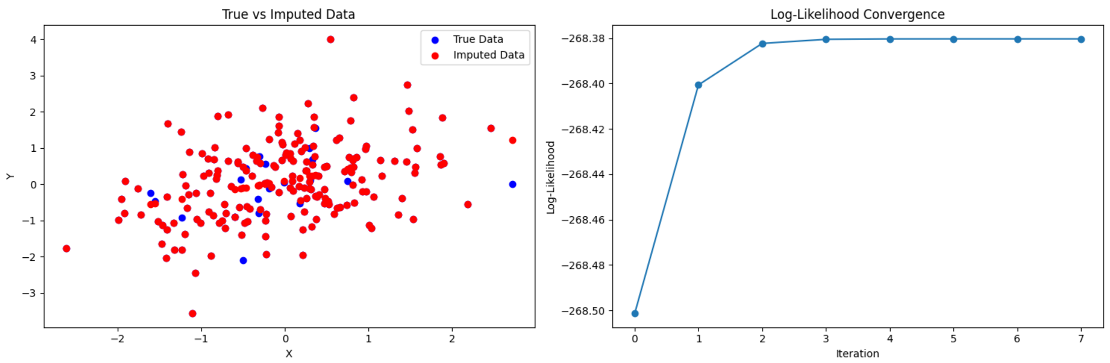
Q2: Implement the EM algorithm to handle missing data in a dataset. Your implementation should:

- Accept a dataset with missing values – generate a synthetic dataset then eliminate some values to obtain a set with missing values;
- Randomly initialize the missing values;
- Iteratively perform the E-step (estimating the missing values) and M-step (recomputing the parameters);
- Output the imputed dataset and the final parameter estimates (e.g., means, variances, correlations).

Test your implementation on the synthetic dataset:

- Visualize the original data with missing values and the imputed dataset;
- Compare your imputed values to the true values;
- Plot the log-likelihood values at each iteration to demonstrate convergence.





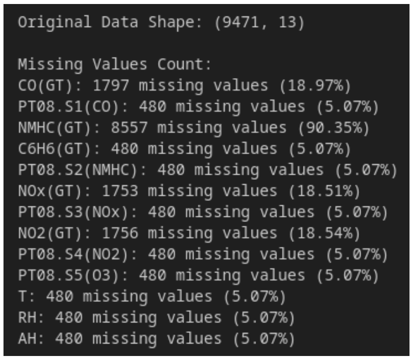
Q3: Choose a real-world dataset with missing values. e.g.,

- Air Quality Dataset (UCI Machine Learning Repository) - includes sensor data with missing entries;
- Housing Prices Dataset (Kaggle) - contains features like house prices, lot size, and number of rooms, with missing values in some columns.

You are free to choose another real-world dataset if you wish.

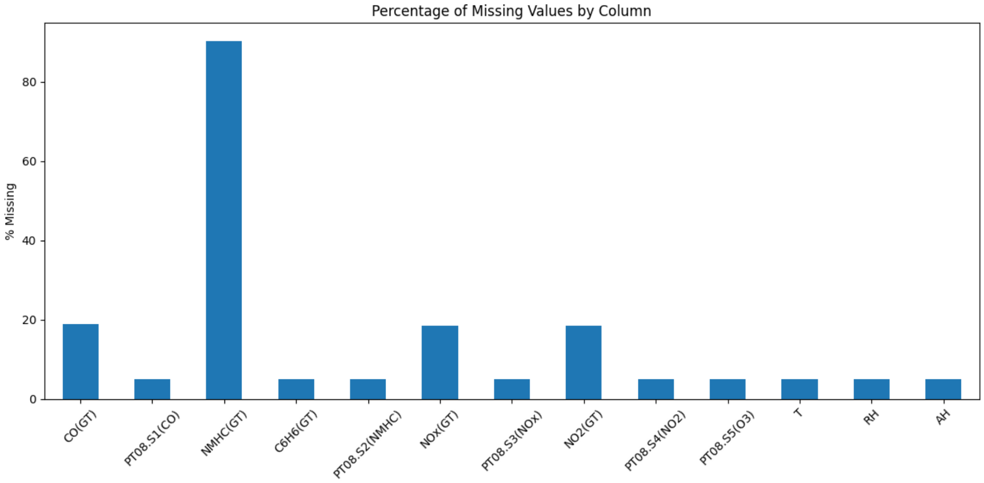
Preprocess the data as necessary (e.g., handle categorical variables, normalize numerical features).

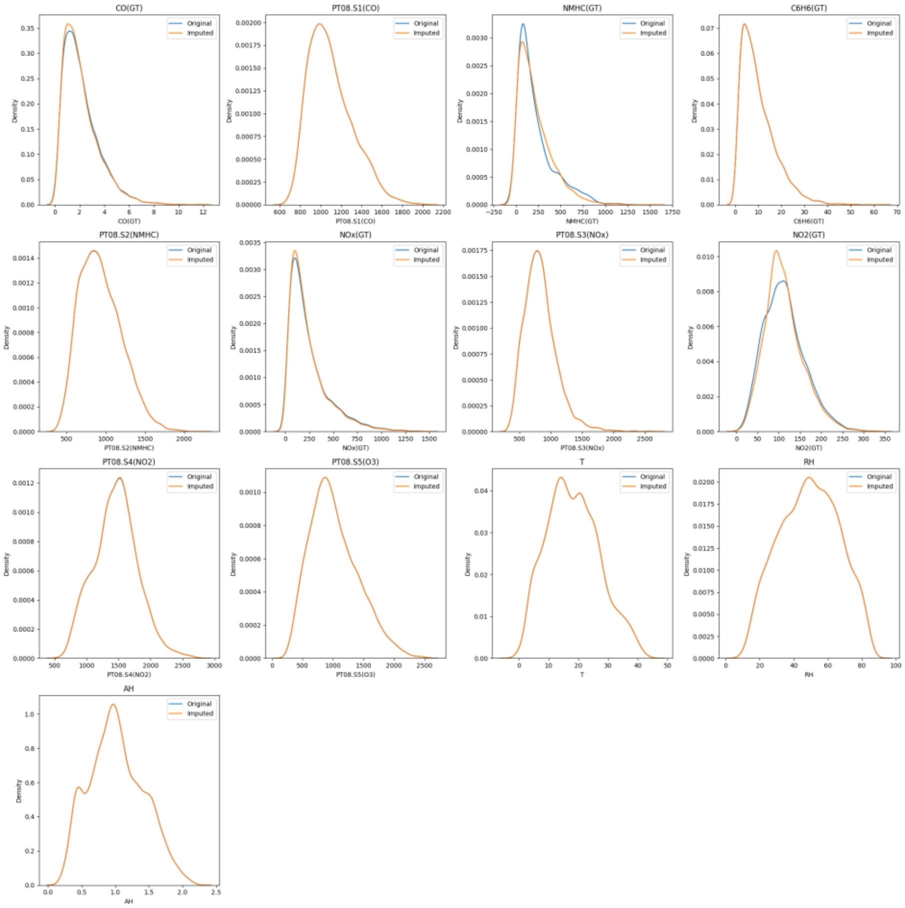
- Use your EM algorithm to impute the missing values;
- Visualize the dataset before and after imputation;
- Analyze how the imputed values compare with domain knowledge;
- Evaluate the overall impact of imputation on the dataset’s structure (e.g., correlation matrix).

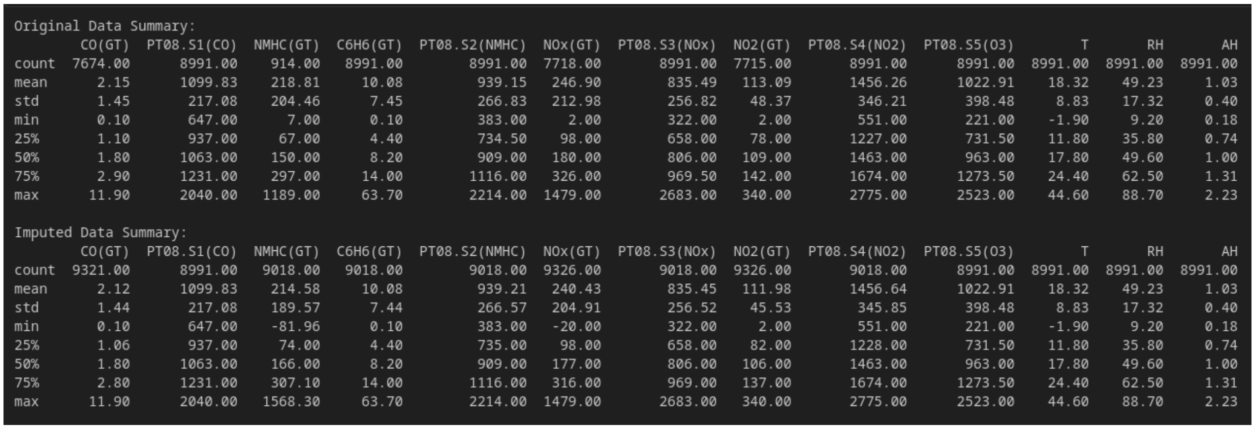


```
Original Data Shape: (9471, 13)

Missing Values Count:
CO(GT): 1797 missing values (18.97%)
PT08.S1(CO): 480 missing values (5.07%)
NMHC(GT): 8557 missing values (90.35%)
C6H6(GT): 480 missing values (5.07%)
PT08.S2(NMHC): 480 missing values (5.07%)
NOx(GT): 1753 missing values (18.51%)
PT08.S3(NOx): 480 missing values (5.07%)
NO2(GT): 1756 missing values (18.54%)
PT08.S4(NO2): 480 missing values (5.07%)
PT08.S5(O3): 480 missing values (5.07%)
T: 480 missing values (5.07%)
RH: 480 missing values (5.07%)
AH: 480 missing values (5.07%)
```

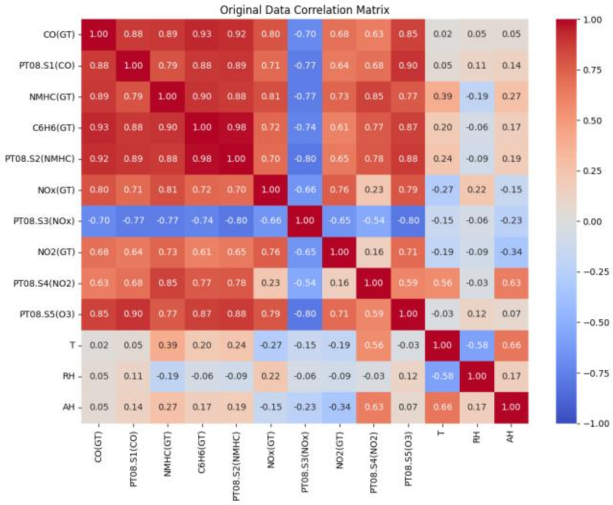


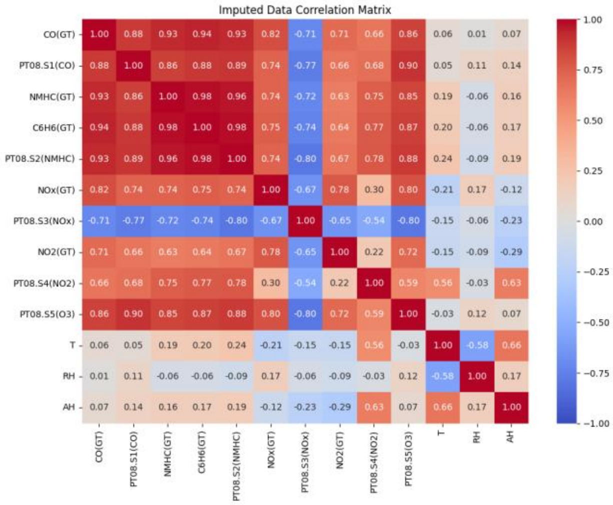


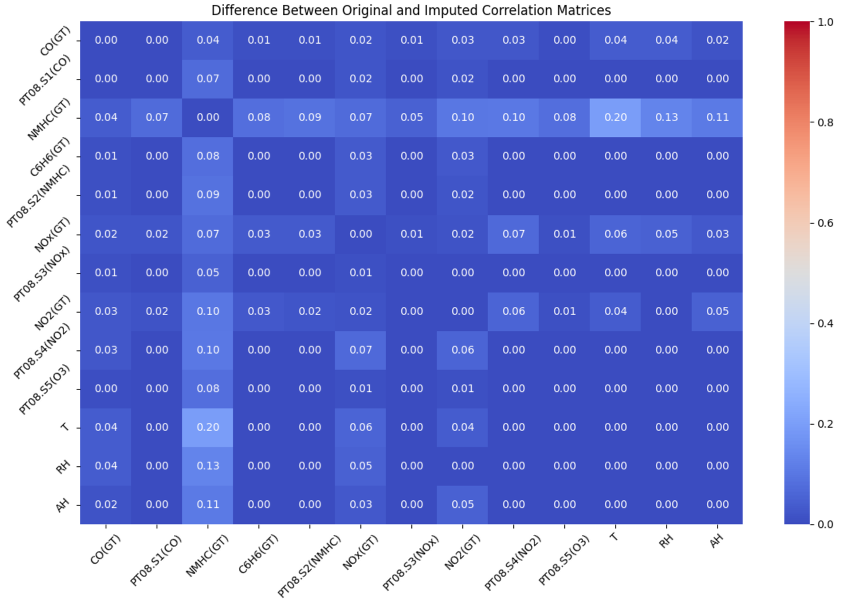


```
Original Data Summary:
       CO(GT)  PT08.S1(CO)  NMHC(GT)  C6H6(GT)  PT08.S2(NMHC)  NOx(GT)  PT08.S3(NOx)  NO2(GT)  PT08.S4(NO2)  PT08.S5(O3)        T       RH       AH
count  7674.00      8991.00    914.00   8991.00        8991.00  7718.00       8991.00  7715.00       8991.00      8991.00  8991.00  8991.00  8991.00
mean      2.15      1099.83    218.81     10.08         939.15   246.90        835.49   113.09       1456.26      1022.91    18.32    49.23     1.03
std       1.45       217.08    204.46      7.45         266.83   212.98        256.82    48.37        346.21       398.48     8.83    17.32     0.40
min       0.10       647.00      7.00      0.10         383.00     2.00        322.00     2.00        551.00       221.00    -1.90     9.20     0.18
25%       1.10       937.00     67.00      4.40         734.50    98.00        658.00    78.00       1227.00       731.50    11.80    35.80     0.74
50%       1.80      1063.00    150.00      8.20         909.00   180.00        806.00   109.00       1463.00       963.00    17.80    49.60     1.00
75%       2.90      1231.00    297.00     14.00        1116.00   326.00        969.50   142.00       1674.00      1273.50    24.40    62.50     1.31
max      11.90      2040.00   1189.00     63.70        2214.00  1479.00       2683.00   340.00       2775.00      2523.00    44.60    88.70     2.23

Imputed Data Summary:
       CO(GT)  PT08.S1(CO)  NMHC(GT)  C6H6(GT)  PT08.S2(NMHC)  NOx(GT)  PT08.S3(NOx)  NO2(GT)  PT08.S4(NO2)  PT08.S5(O3)        T       RH       AH
count  9321.00      8991.00   9018.00   9018.00        9018.00  9326.00       9018.00  9326.00       9018.00      8991.00  8991.00  8991.00  8991.00
mean      2.12      1099.83    214.58     10.08         939.21   240.43        835.45   111.98       1456.64      1022.91    18.32    49.23     1.03
std       1.44       217.08    189.57      7.44         266.57   204.91        256.52    45.53        345.85       398.48     8.83    17.32     0.40
min       0.10       647.00    -81.96      0.10         383.00   -20.00        322.00     2.00        551.00       221.00    -1.90     9.20     0.18
25%       1.06       937.00     74.00      4.40         735.00    98.00        658.00    82.00       1228.00       731.50    11.80    35.80     0.74
50%       1.80      1063.00    166.00      8.20         909.00   177.00        806.00   106.00       1463.00       963.00    17.80    49.60     1.00
75%       2.80      1231.00    307.10     14.00        1116.00   316.00        969.00   137.00       1674.00      1273.50    24.40    62.50     1.31
max      11.90      2040.00   1568.30     63.70        2214.00  1479.00       2683.00   340.00       2775.00      2523.00    44.60    88.70     2.23
```


Original Data Correlation Matrix


Imputed Data Correlation Matrix


Difference Between Original and Imputed Correlation Matrices

Nota Bene:

- For completing this homework, I have used AI tools for generating plots and finding the right formulas to answer question 1.
- My personal contribution was to preprocess the datasets, generate it for the second question, write the EM algorithm and adapt it for question 3 and analyze the results.