Regression Modeling and Performance Evaluation on Multimodal Gaussian Data: A Study on Linear, Polynomial, Ridge, and LASSO Methods

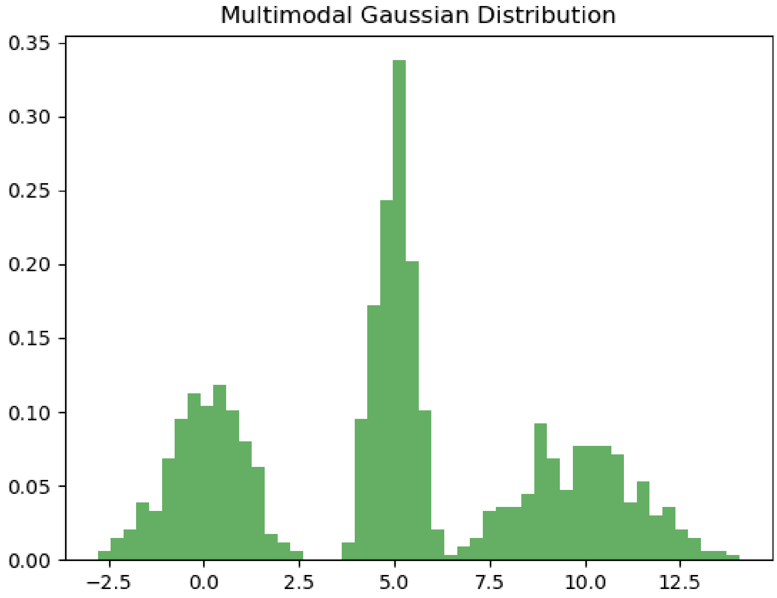
Introduction

This report outlines the analysis performed on a synthetic multimodal Gaussian distribution using various regression techniques. The tasks include constructing multimodal Gaussian data, applying piecewise linear and polynomial regressions, and evaluating performance using error metrics such as RMSE and R^2. Additionally, Ridge and LASSO regularization techniques are applied and compared with standard polynomial regression. The results are evaluated in terms of accuracy and complexity.

Question 1: Build a multimodal Gaussian distribution with synthetic data.

For this task, a multimodal Gaussian distribution was generated using synthetic data. The data was created with three distinct modes, each with different means, standard deviations, and weights. These parameters help in forming a distribution with multiple peaks, which visually reflects a multimodal Gaussian distribution.

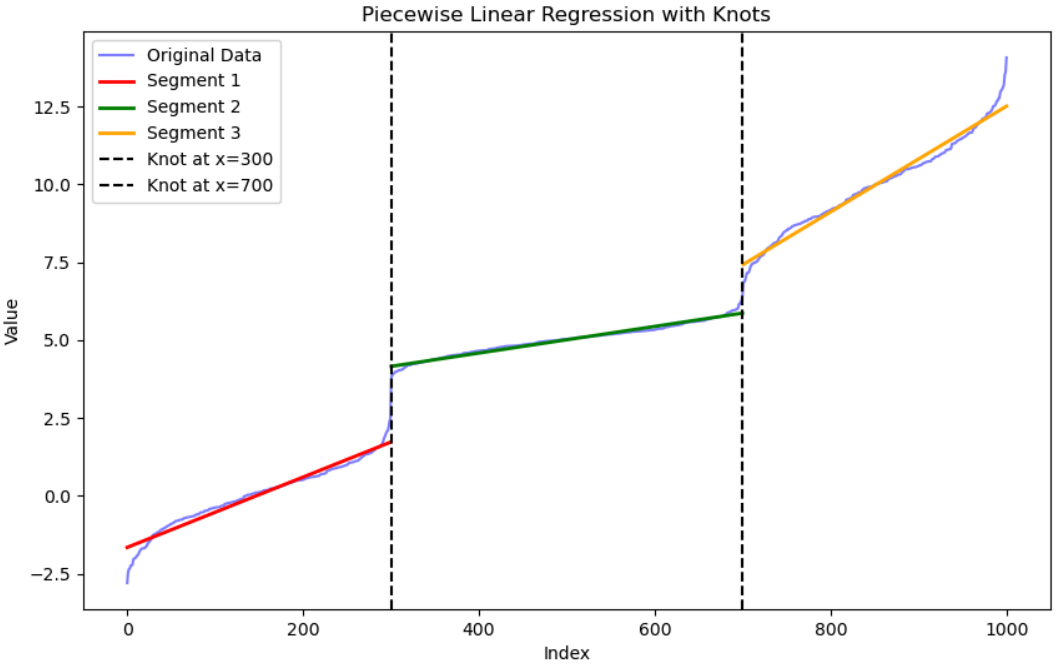
The image below shows the distribution, where the green bars represent the density of the generated data. The first peak occurs around a mean of 0, the second around a mean of 5, and the third around a mean of 10. The relative heights of these peaks are determined by the weights assigned to each mode.



Question 2: Construct a piecewise linear regression and plot the result with its splines and knots.

In this task, a piecewise linear regression was applied to the multimodal Gaussian distribution generated earlier. The data was segmented into three parts using two knots, located at indices 300 and 700. For each segment, a separate linear regression model was fitted to the data, allowing for distinct line segments in different parts of the data.

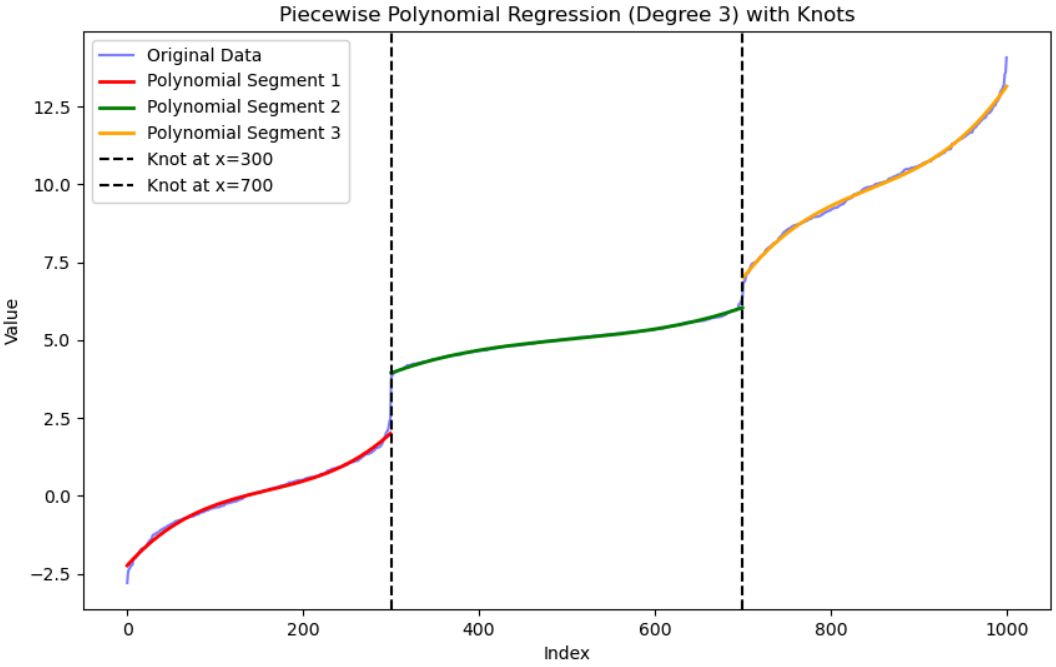
The image below shows the original data in blue, along with the three linear segments in red, green, and orange, representing the individual linear regressions for each segment. The dashed black lines indicate the positions of the knots at indices 300 and 700, where the behavior of the data changes.



Question 3: Construct a piecewise polynomial regression and plot the result, its splines, and knots.

In this task, a piecewise polynomial regression of degree 3 was applied to the multimodal Gaussian distribution generated earlier. The data was segmented into three parts using two knots located at indices 300 and 700. A separate polynomial regression model was fitted to each segment, allowing for a more flexible fit compared to linear regression.

The image below shows the original data in blue, along with the polynomial regression segments in red, green, and orange, representing the individual polynomial fits for each segment. The dashed black lines indicate the positions of the knots at indices 300 and 700, where the behavior of the data changes.



Question 4: In addition to F-statistics that you use to decide about the number of knots, report RMSE and R^2 for all models.

For this task, the performance of both piecewise polynomial regression (degree 3) and piecewise linear regression was evaluated on three segments of the multimodal Gaussian data. The evaluation was done using two key metrics:

- **Root Mean Squared Error (RMSE)**: Measures the standard deviation of the differences between predicted and observed values. Lower RMSE indicates a better fit.
- **R^2 Score**: Also known as the coefficient of determination, It measures how well the regression predictions approximate the real data points. A value closer to 1 suggests a better fit.

Piecewise Polynomial Regression Results:

Segment 1:

RMSE = 0.1137: The prediction error is quite low, indicating that the polynomial regression fits the data well.

R^2 = 0.9872: The model explains 98.72% of the variance in the data for segment 1, indicating an excellent fit.

Segment 2:

RMSE = 0.0428: The prediction error is even lower in this segment, suggesting a very tight fit between the predicted and actual values.

R^2 = 0.9921: The model explains 99.21% of the variance in the data, which means it's performing exceptionally well for this segment.

Segment 3:

RMSE = 0.1395: The prediction error is slightly higher compared to the other segments but still relatively low.

R^2 = 0.9910: The model explains 99.10% of the variance in the data, which is still a very strong fit for segment 3.

Piecewise Linear Regression Results:

Segment 1:

RMSE = 0.2051: The prediction error is higher than that of the polynomial regression in this segment, which is expected because linear regression is less flexible compared to polynomial regression.

R^2 = 0.9583: The model explains 95.83% of the variance, which is still a good fit but not as strong as the polynomial model.

Segment 2:

RMSE = 0.0993: The prediction error is relatively low, but higher than the polynomial regression for this segment.

R^2 = 0.9576: The model explains 95.76% of the variance, indicating a solid fit, though not as tight as the polynomial regression.

Segment 3:

RMSE = 0.2321: This is the highest RMSE among all the segments and models, indicating that the linear model struggles more in this segment.

R^2 = 0.9751: Despite the higher error, the model still explains 97.51% of the variance, which shows it performs reasonably well in this segment.

Conclusion:

The results demonstrate that **piecewise polynomial regression** is a superior model for this dataset compared to piecewise linear regression. The polynomial regression captures the underlying trends in the data more effectively, as evidenced by lower RMSE values and higher R^2 scores. On the other hand, the linear regression model, although simple and computationally efficient, struggles to model the complexity of the multimodal distribution.

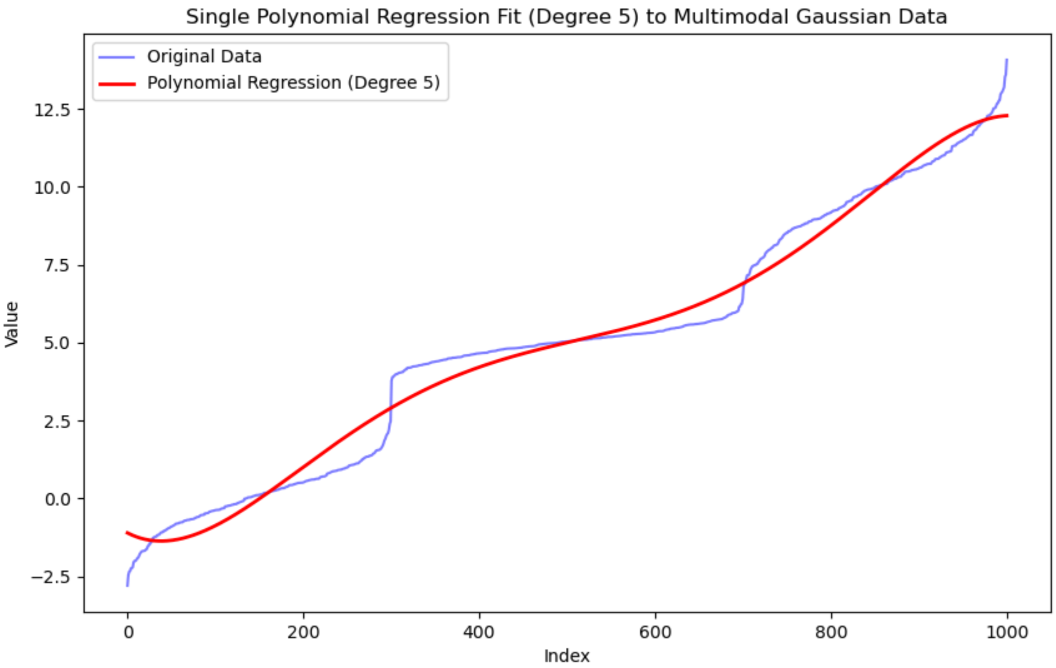
Question 5: Try to model the multimodal Gaussian distribution built with single polynomial regression.

In this task, a single polynomial regression model of degree 5 was applied to the entire multimodal Gaussian distribution. Unlike the previous tasks, where the data was split into segments, here the entire dataset was modeled with one polynomial regression curve.

The degree of the polynomial was set to 5 to allow the model to capture the complexity of the data, including the multiple peaks and nonlinear trends present in the multimodal distribution.

Results:

The graph below shows the original data in blue and the fitted polynomial curve in red. The polynomial regression model fits reasonably well but struggles to capture some of the finer details in the data, particularly the transitions between modes.



Performance Evaluation:

- **RMSE (Root Mean Squared Error)**: 0.5476
- **R^2**: 0.9814

Interpretation:

- The RMSE value of 0.5432 indicates that, on average, the model's predictions are about 0.543 units away from the actual values in the dataset.
- The R^2 score of 0.9814 suggests that the model explains about 98.14% of the variance in the data, indicating a good fit overall. However, the fit could be improved further by breaking the data into segments (as done in previous tasks), which would allow the model to more accurately capture the complexity of the multimodal distribution.

Question 6: Measure and report the execution time of tasks (2), task (3), and task (5). Then report the differences in execution time (no plot required, but you need to report them in a table). Hint: use excel for table drawing and add it in the word file.

In this task, the execution times for three different regression models were measured:

1. **Question 2**: Piecewise Linear Regression
2. **Question 3**: Piecewise Polynomial Regression (degree 3)
3. **Question 5**: Single Polynomial Regression (degree 5)

The objective was to compare the computational time required for each of these models, given that they handle the complexity of the data in different ways.

Task 2: Piecewise Linear Regression	0.0136
Task 3: Piecewise Polynomial Regression	0.0155
Task 5: Single Polynomial Regression	0.0131
Difference (Task 2 vs Task 3)	0.0019
Difference (Task 3 vs Task 5)	0.0024
Difference (Task 2 vs Task 5)	0.0005

Interpretation:

- **Piecewise Linear Regression** (Task 2) was slightly faster than **Piecewise Polynomial Regression** (Task 3). This is expected, as polynomial regression involves more computations (especially with the transformation of features).
- **Single Polynomial Regression** (Task 5) was the fastest among the three tasks, with an execution time of 0.0131 seconds. This is likely because the model fits the entire dataset in one go, rather than fitting multiple segments as in the piecewise approach. However, this doesn't necessarily mean it's a better model in terms of accuracy; the computation time is just lower due to the model complexity and method used.
- The differences in execution times are small, given the size of the dataset, but they become more noticeable as the complexity of the data or model increases.

Conclusion:

Although the differences in execution time between the tasks are minimal for this specific dataset, **piecewise models** tend to take slightly longer to compute due to the multiple regressions being fit on different segments of the data. In contrast, **single polynomial regression** can be computationally faster but might not capture the nuances in the data as effectively as piecewise models.

Question 7: Use the Multimodal Gaussian distribution of Tasks 5, apply Ridge, LASSO and compare their accuracy and number of parameters, parameters coefficient with Polynomial regression. Here you should also report and discuss the differences (if there are any differences).

In this task, **Ridge** and **LASSO** regression models were applied to the multimodal Gaussian distribution generated in Task 5. These models were compared to the polynomial regression model (degree 5) to analyze their performance in terms of:

- **Accuracy**: Measured using Root Mean Squared Error (RMSE) and R^2 score.
- **Model Coefficients**: Comparing the number and magnitude of the coefficients produced by each model.

Results:

Model Performance:

- **Polynomial Regression**:
  - **RMSE** = 0.5476
  - **R^2** = 0.9814
- **Ridge Regression** (Regularization applied):
  - **RMSE** = 0.7211
  - **R^2** = 0.9678
- **LASSO Regression** (L1 regularization applied):
  - **RMSE** = 0.7391
  - **R^2** = 0.9662

Model Coefficients:

- **Polynomial Regression Coefficients**:
  ```
  [ 0.00000000e+00 -2.04783936e-02 2.63837183e-04 -7.21005730e-07
  8.01374236e-10 -3.10652585e-13 ]
  ```
- **Ridge Regression Coefficients**:
  ```
  [ 0. 3.62216923 0.12076224 -0.50591088 0.01609331 0.75539543 ]
  ```
- **LASSO Regression Coefficients**:
  ```
  [0. 3.557389 0. 0. 0. 0.35080369 ]
  ```

Interpretation:

Model Performance:

- **Polynomial Regression** performed the best, achieving the lowest RMSE (0.5476) and the highest R^2 score (0.9814). This suggests that the polynomial regression model fits the data more accurately than both Ridge and LASSO.
- **Ridge Regression**, which applies L2 regularization, slightly increased the RMSE to 0.7211 and reduced the R^2 score to 0.9678. This is because Ridge attempts to shrink coefficients and reduce model complexity, which slightly decreases its ability to fit the data perfectly.
- **LASSO Regression** performed similarly to Ridge but had the highest RMSE (0.7391) and the lowest R^2 score (0.9662). LASSO tends to drive coefficients toward zero, resulting in a sparser model.

Model Coefficients:

- **Polynomial Regression** produced small but non-zero coefficients across all terms. These coefficients allow the model to fit the data well but can lead to overfitting in more complex datasets.
- **Ridge Regression** retained non-zero coefficients across all terms but shrank the magnitude of most coefficients. This suggests that Ridge controls overfitting by applying a penalty on the size of the coefficients while keeping them small and non-zero.
- **LASSO Regression** forced several coefficients to zero (notably for higher-degree polynomial terms), creating a more parsimonious model with fewer parameters. LASSO's feature selection property can be useful for reducing complexity, but in this case, it led to a slightly less accurate model.

Conclusion:

- **Polynomial Regression** provided the most accurate fit, with the lowest error and highest R^2, but at the cost of using more parameters.
- **Ridge Regression** achieved a balance between model complexity and performance, with slightly reduced accuracy but better control over overfitting.
- **LASSO Regression** applied stronger regularization, removing some polynomial terms entirely. This resulted in a simpler, more interpretable model, but the fit was not as accurate as Polynomial Regression or Ridge.

The choice between these models depends on the trade-off between accuracy and model complexity. If interpretability and avoiding overfitting are priorities, Ridge or LASSO might be better choices. However, if the best possible fit is desired, Polynomial Regression should be preferred.