Report HW3

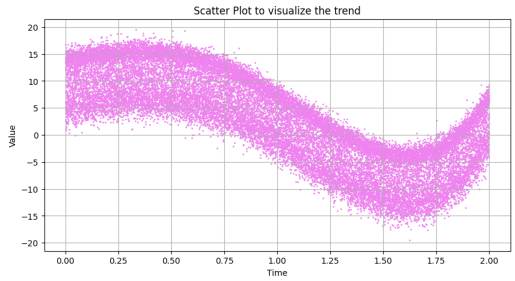
Anonymous Submission

Exercise 1

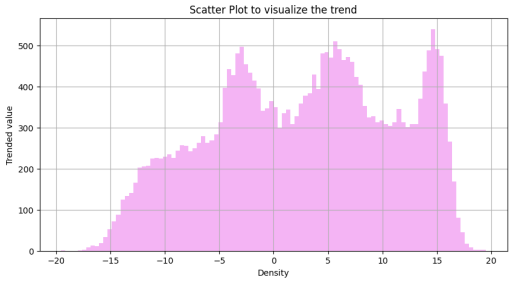
Given a dataset of approximately $3 \cdot 10^4$ entries of the form (x, y), where x is a time value and y is its corresponding value, we are interested in modeling the statistics of the dataset. There is a clear trend in the data.

Part 1: Visualize the trend

First, we need to visualize the trend in the data. To do this, we scatter plot the data points (1a) and we plot the distribution of the data points (1b) that we cannot trust.



(a) Scatter plot of the dataset.



(b) Histogram of the dataset.

Figure 1: Trended dataset.

Part 2: Fit a polynomial trend

We estimate a polynomial trend using the least squares method for degrees from 1 to 8 (2).

Part 3: Find the best degree and remove the trend

To remove the trend, we need to decide which polynomial degree is the best. Then we can de-trend the data.

Find the best degree

To decide which polynomial degree is the best, we use the root mean square error (RMSE) of the polynomial fit for each degree (3a). We also use cross-validation to estimate the RMSE of the polynomial fit for each degree (3b).

Both methods in Figure 3 show that the best degree is 4.

De-trend the data

To de-trend the data, we just subtract the best polynomial fit (degree 4) from the data. The result is shown in Figure 4.

Part 4: Fit a mixture of Gaussians

To fit a mixture of Gaussians to the de-trended data, we implement the EM algorithm.

We run the EM algorithm with a maximum of 10^3 iterations and a tolerance of 10^{-6} with two initilization settings:

- starting from a uniform prior, means randomly chosen from the data, and variances all equals to the dataset variance;
- starting from K-means clustering parameters, i.e. setting the priors as the fraction of points for each cluster, means as the cluster centroids and variances estimated as intra-cluster variances.

The results of the EM algorithm are shown in Table 1 and visualized in Figure 5.

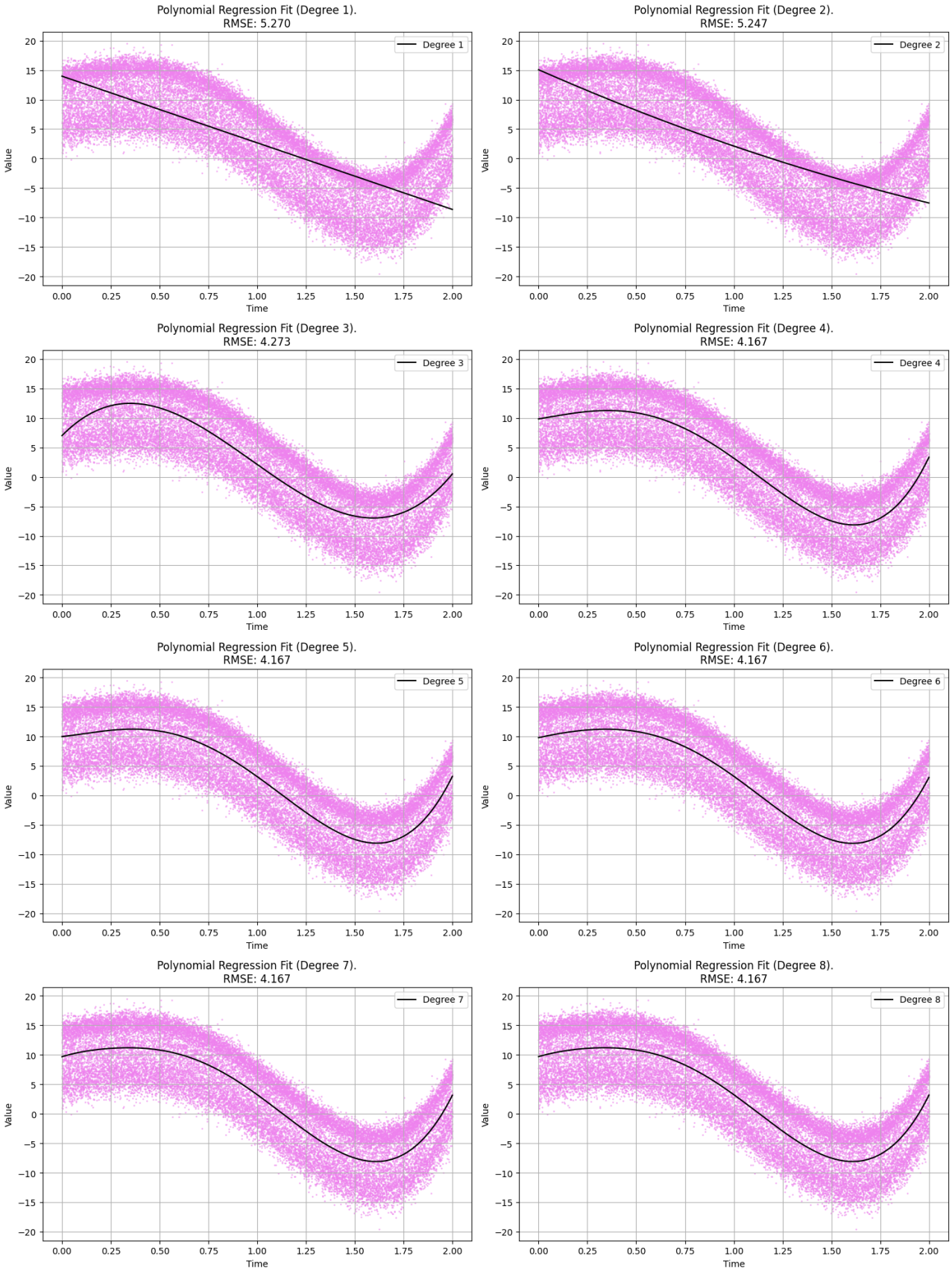
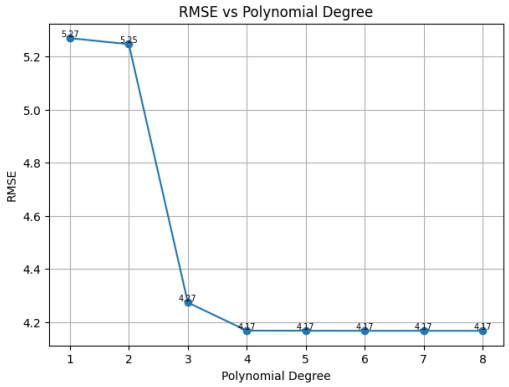
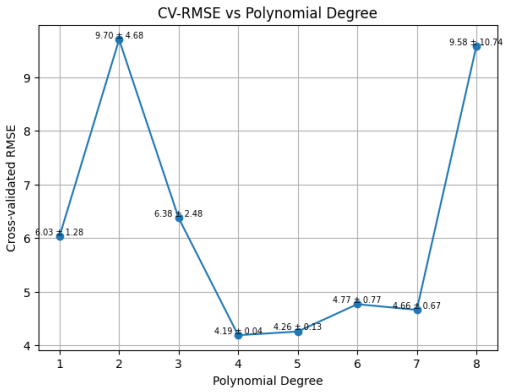


Figure 2: Polynomial trends fitted to the dataset.

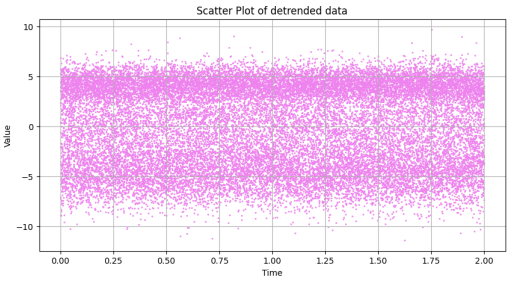


(a) RMSE of the polynomial for each degree.

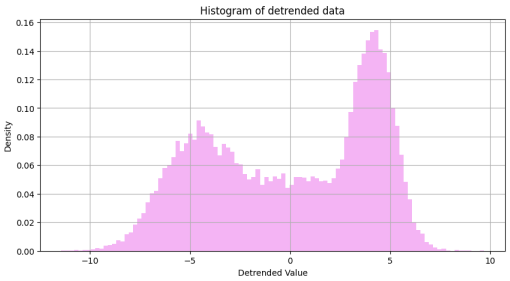


(b) Mean RMSE from the cross-validation for each degree.

Figure 3: Evaluation of the polynomial degree.



(a) Scatter plot of the de-trended dataset.



(b) Histogram of the de-trended dataset.

Figure 4: De-trended dataset.

	Original		KMeans init (iter. 569)			Random init (iter. 473)		
Component	μ	σ^2	Weight	μ	σ^2	Weight	μ	σ^2
1	-5	3	0.3443	-4.7274	3.0625	0.3443	-4.7274	3.0625
2	0	6	0.3024	0.4021	5.9683	0.3534	0.4021	5.9683
3	4	1	0.3534	4.2616	0.9841	0.3534	4.2616	0.9841

Table 1: Results from the EM algorithm for the two initialization settings compared to the original Gaussian parameters.

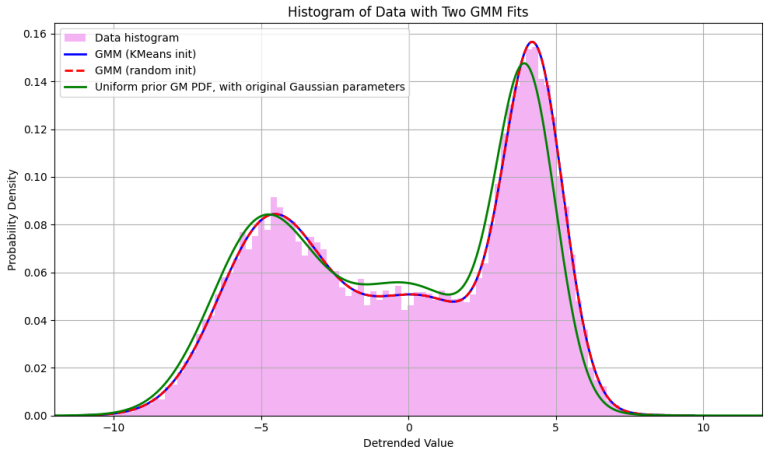


Figure 5: Gaussian Mixture Models fitted to the de-trended distribution in comparison with the original gaussians assuming uniform priors.

Part 5: Find the optimal number of gaussians

To find the optimal number of Gaussians, we use the Bayesian Information Criterion (BIC) as regularization term. The BIC is defined as:

$$\text{BIC} = -2 \log L + k \log n \tag{1}$$

where L is the likelihood of the model, k is the number of parameters in the model, and n is the number of data points. The BIC is a trade-off between the goodness of fit of the model and the complexity of the model that aims to avoid overfitting.

We fit GMM models with 1 to 10 Gaussians for a maximum of 10^3 iterations and a tolerance of 10^{-6} with random means initialization as before. The BIC scores are shown in Figure 6 and the optimal number of Gaussians is 3.

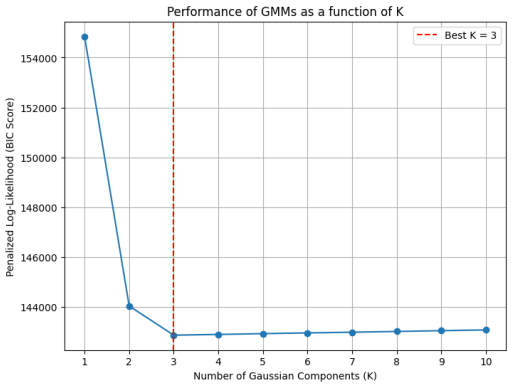


Figure 6: Bayesian Information Criterion (BIC) score for increasing number of gaussians.