Multiple Theoretical Distributions Fitting to Empirical Distribution Using KDE and Modified KS-Tests

David Dorfman, Ella Kharakh

13 March 2023

1 Abstract

The aim of this project is to improve the goodness of fit by creating an upgraded version of KS-test. The traditional KS-test has limitations in detecting deviations from concatenation of theoretical distributions and can lead to inaccurate results. The proposed approach is to identify the KDE function's minimum and maximum points and find theoretical distributions between every pair of points. The approach is based on the idea that data can be thought of as a concatenation of some distributions. The identified theoretical distributions can help understand how the data behaves and predict data in the world beyond the sample. The proposed method was applied to different datasets, and the results showed that the upgraded KS-test was more accurate in detecting deviations from theoretical distributions. In conclusion, the proposed approach is a novel and effective way of analyzing data and improving the goodness of fit. It has the potential to contribute significantly to the field of statistical analysis and data science.

2 Problem Description

Our approach aims to enhance the goodness of fit by improving the agreement between empirical and theoretical distributions. This falls within the scope of Exploratory Data Analysis (EDA). The conventional approach involves assessing whether the values in each feature conform to a specific theoretical distribution.

However, this approach overlooks the possibility that different segments of the data may have unique distributions. By examining subsets of data for a given feature and identifying their distributions, we can obtain a better fit for our data. This will facilitate a deeper understanding of how the data behaves and enable us to make informed conclusions about the data, detect outliers, and identify imbalanced data.

3 Solution Overview

Our solution finds the best fitting theoretical distributions to the mixed distributed empirical distribution. First, we calculate the KDE of the empirical data, then we divide the empirical data using the minimum and maximum points (found by sampling the KDE) and we check the best fitting theoretical distribution using KS-tests for the following segments: from a minimum point to the next maximum point, from maximum point to the next minimum point, from minimum point to the next minimum point. For every segment we find the best fitting theoretical distribution out of the following continuous distributions we saw in class: Normal distribution, Exponential distribution, and Weibull distribution. For each of those distributions we find the best fitting parameters using the Maximum Likelihood Estimation method. Then, we choose the best matching distribution for each segment (using the highest P-value we got from the KS-test), and then we choose the segment with the highest P-value out of those 3 segments. After that, we check if expending or narrowing each segment by 10 percent improves the previous found P-value, and if the P-value indeed improved then we repeat that until the next P-value is lower then the previous. After we found all those segments, there may be some overlaps of segments and some gaps between segments. To get the most representative continuous concatenation of theoretical distributions we need to find the best balance points in those overlaps and gaps. In case of overlap, we find the point within the overlap which maximize the P-valve of those two new segments (which are similar to the previous segments but with the new point as end/start point). We find those balance point using a recursive method, similarly to binary search. In case of gaps, we act similarly. The P-values calculated are multiplied by the ratio between the segment length and the data length, to prefer a longer segment then short one. Finally, the total P-value of those KS-tests is the weighted average of the previously found P-values by the ratio of the segments length they belong

to and the total segment length.

4 Experimental Evaluation

We compared the weighted average of the P-values from the KS-tests we performed on each segment we found with the maximal P-value between the P-values of the KS-tests performed on the whole feature (using 3 different distributions which their parameters calculated using the Maximum Likelihood Estimation: Weibull distribution, Normal distribution, Exponential distribution). We discovered that our solution outperforms the second solution - the P-value we received is higher than the solution we compared with:

Dataset 1: our solution: 0.8457735233222444 taraditional solution: 2.626433987152613e-229

Dataset 2: our solution: 0.9292004785476264 taraditional solution: 0.0

Dataset 3: our solution: 0.8068496510416658 taraditional solution: 0.0

Dataset 4: our solution: 0.8172216621134196 taraditional solution: 1.5037528771650616e-235

5 Related Work

Unfortunately, there are no other innovative methods that relate to this specific problem of data having subsets with unique theoretical distributions in them. we hoped the paper " Mixture Models, Outliers, and the EM Algorithm" would be a good option to compare, in terms of kstest grade and most importantly the pvalue. It turned out this paper takes into consideration different aspects and relates to a different problem. That's why we ended up comparing with the traditional approach of kstest: comparing the whole data to one specific theoretical distribution.

Other papers didn't manage to solve our problem, but they did try to improve the kstest effectivity and accuracy, by finding a better implementation for ks test for a specific theoretical distribution. In our project, we got inspired by the papers and used several ideas and methods we saw in them.

In the " A Modified Kolmogorov-Smirnov Test for Weibull Distributions with Unknown Location and Scale Parameters " we saw an interesting method of getting ks test result for an empirical distribution, suspected to be Weibull Distributions, even though there are unknown parameters such as Unknown

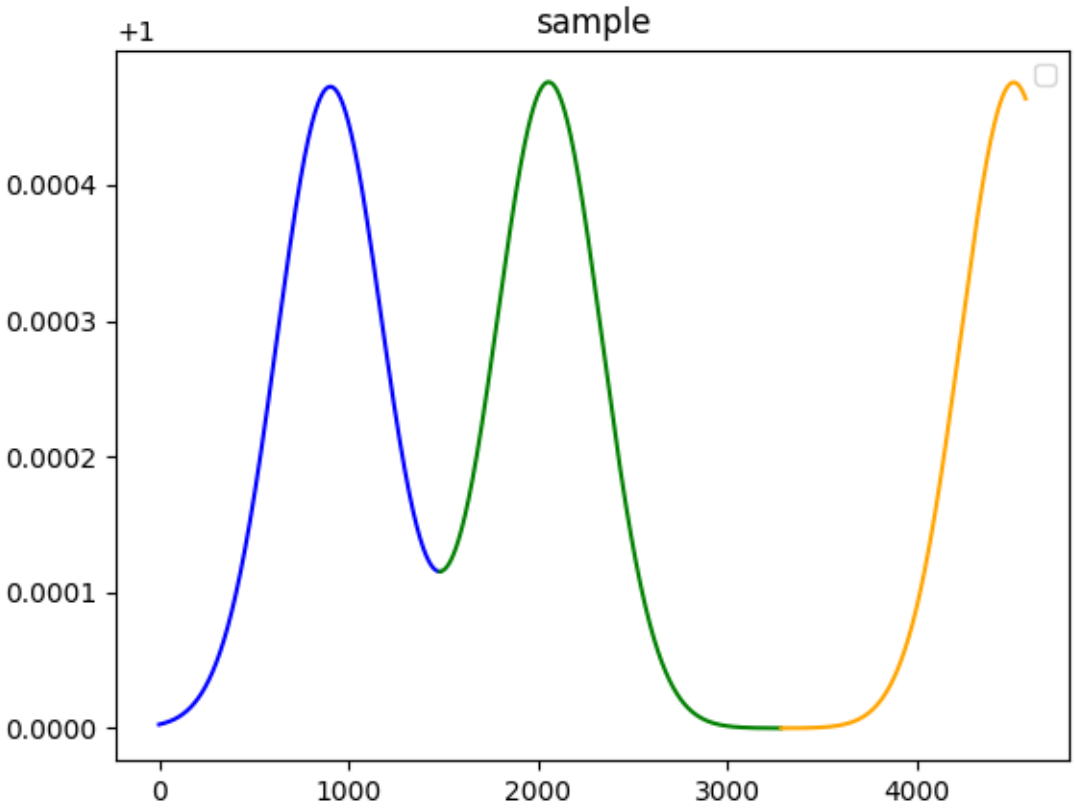

Figure 1: Dataset 1

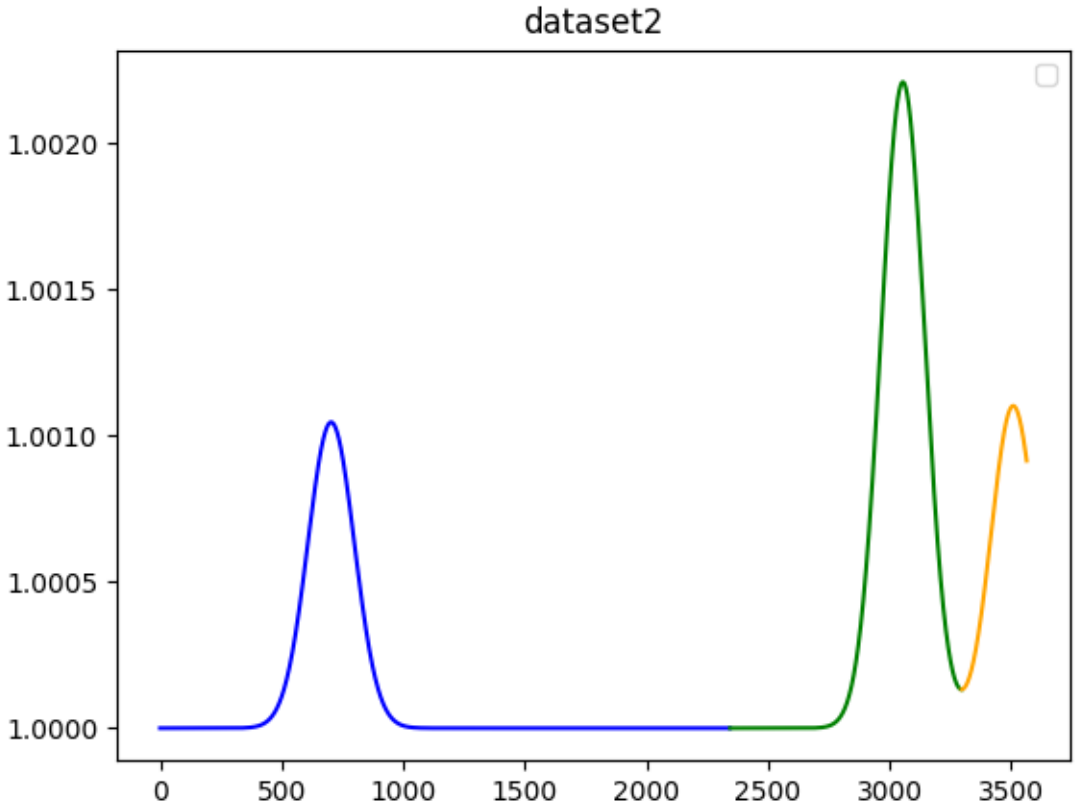

Figure 2: Dataset 2

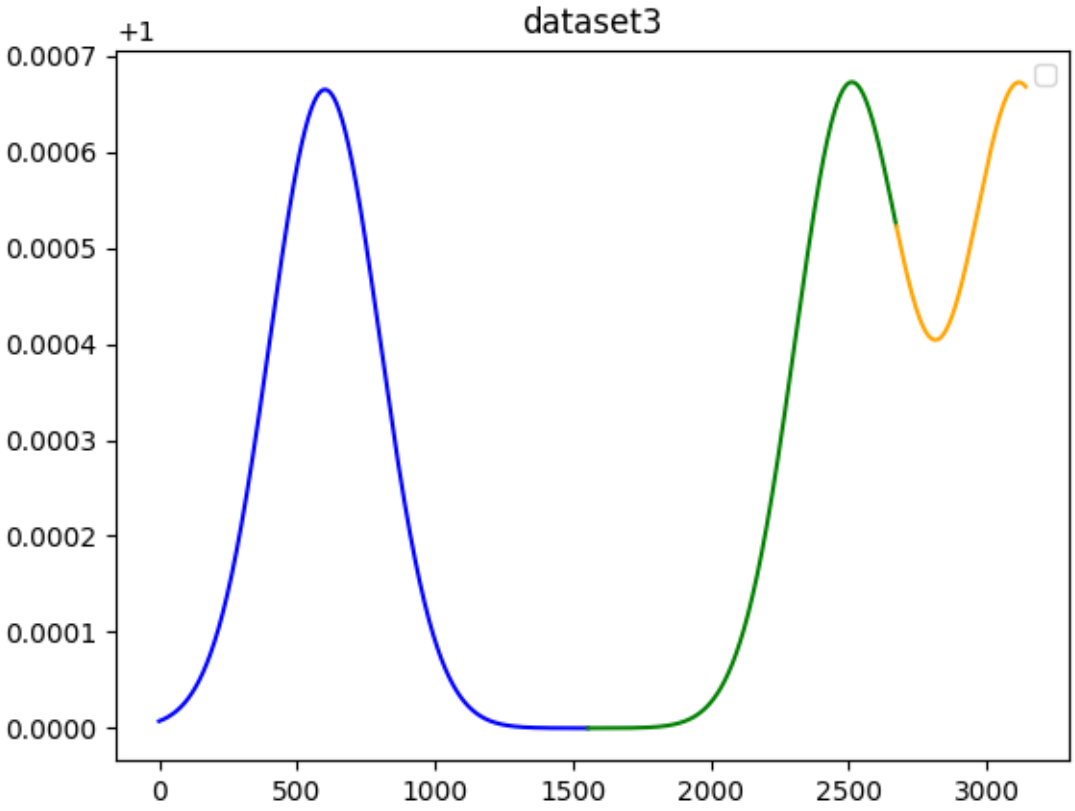

Figure 3: Dataset 3

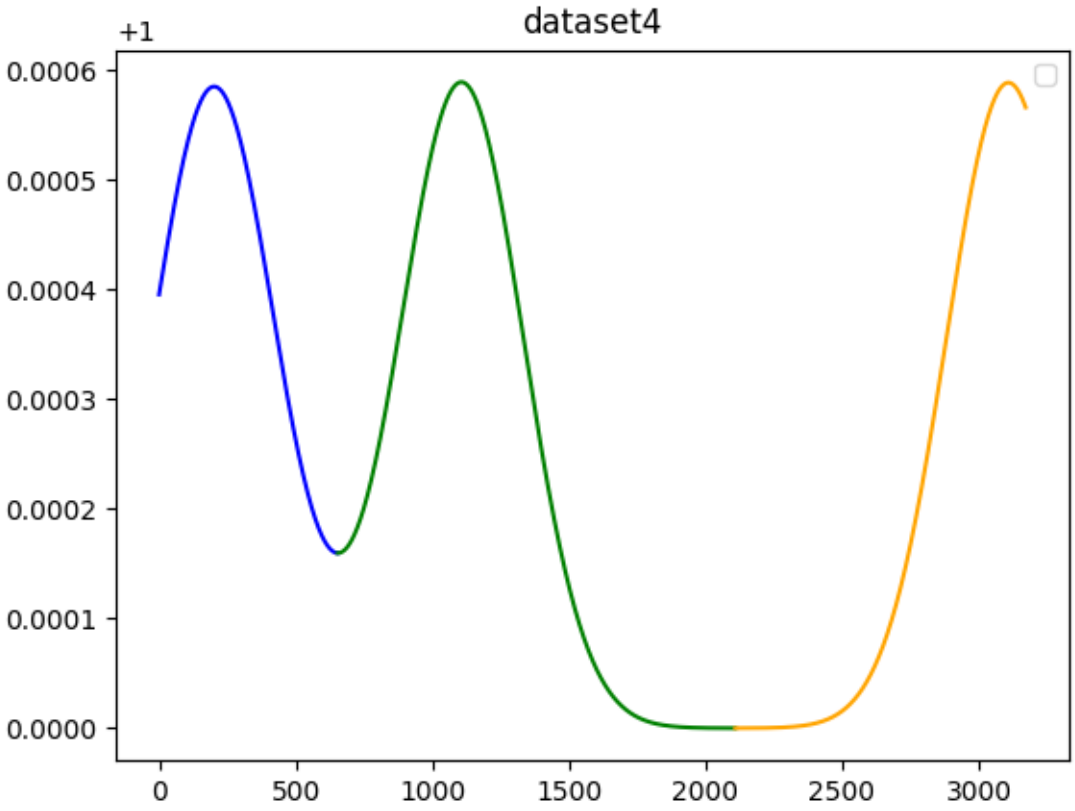

Figure 4: Dataset 4

Location and Scale Parameters. The paper takes into consideration the shape parameter. we couldn't apply this method in our project because we didn't find any information on how to estimate shape parameter when it's not given, but we saw that the maximum likelihood estimate is used in the paper on the scale and location parameters, and that inspired us to use It In our implementation. We used it in the code: params = stats.expon.fit(empirical), which uses MLE and by that finds the best parameters for the ks test func.

In the " A Modified Kolmogorov-Smirnov Test for Normality " described an innovative method for finding if an empirical distribution is actually the normal distribution and if so it finds the parameters of that normal distribution: its mean, its std, and the pvalue appropriate for the that (and other ks test grade parameters). Because aspired to find the pvalue of the kstest, so we could compare our method to the original one, we didn't use the method for finding whether an empirical distribution is actually normal, but we implement the algorithm given in the paper for finding the parameters of the normal distribution. Although the algorithm returns almost the same mean as of a tested normal distribution, its std value is far less accurate than what we would have wished for, that's why we ended up not using it. the implementation of the algorithm displayed in the code we provided. we did get inspiration from the algebraic methods and ways of implementation in the algorithm, such as using binary search, which helped us a lot in bettering our algorithm and making it more accurate.

6 Conclusion

In conclusion, our approach of examining subsets of data for a given feature and identifying their distributions has shown to be a promising technique for improving the agreement between empirical and theoretical distributions. Our results indicate that our method outperforms the conventional approach of assessing whether the values in each feature conform to a specific theoretical distribution. However, it is important to note that our method was only tested on specific datasets with theoretical distributions or empirical distributions close to a theoretical distribution. It is likely that on datasets with no theoretical distributions or empirical distributions close to a theoretical distribution, our approach may not be as successful. Thus, there is still much work to be done in the field of tackling this problem. Unfortunately, it seems that there are not

many situations where our method will make a big difference in a dataset with almost no theoretical distributions. Nonetheless, our approach has the potential to facilitate a deeper understanding of how data behaves, detect outliers, and identify imbalanced data, making it a valuable tool in exploratory data analysis.