# The UU-test: Deciding on Distribution Unimodality using Tests of Uniformity

A Thesis

submitted to the designated
by the General Assembly of Special Composition
of the Department of Computer Science and Engineering
Examination Committee

by

Paraskevi Chasani

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE
WITH SPECIALIZATION
IN SOFTWARE

University of Ioannina

February 2019

Examining Committee:

- Aristidis Likas, Professor, Department of Computer Science and Engineering, University of Ioannina (Supervisor)
- Konstantinos Blekas, Associate Professor, Department of Computer Science and Engineering, University of Ioannina
- Christophoros Nikou, Professor, Department of Computer Science and Engineering, University of Ioannina

# DEDICATION

To my family and Fotis.

# ACKNOWLEDGMENTS

I would like to thank my thesis advisor Prof. Aristidis Likas for the guidance and patience throughout my research. His positive outlook and his brilliant ideas not only made this thesis possible but also helped me develop strong research skills and my critical thinking.

Furthermore, I would like to thank my colleagues for their great help and the pleasure collaboration we had.

Finally, I would like to thank my parents for supporting me all these years and my sister and best friend Katerina who has always been there for me and always understands me whatever happens.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ALGORITHMS

# ABSTRACT

Paraskevi Chasani
MSc, Computer Science and Engineering, University of Ioannina, Greece
February 2019
Title: The UU-test: Deciding on Distribution Unimodality using Tests of Uniformity
Supervisor: Aristidis Likas

Recognizing unimodal data distributions is of great significance in statistics, machine learning and data science. Well-known distributions, such as Gaussian, Student's t, Gamma, Chi-square, Exponential and Cauchy are typical examples of unimodal distributions. Also the uniform distribution is considered as an extreme unimodal case. The characteristic property of a unimodal distribution is that data values are gathered around a single value (peak), which is the mode of the distribution. Due to this property, data can be characterized as homogeneous, forming a single and coherent group.

Unimodality tests have been proposed to decide on the unimodality of a set of data values, thus providing useful knowledge about the structure of the data. For example, if a dataset is unimodal, the data values are "gathered" thus applying a clustering method is unnecessary. Current unimodality tests decide exclusively about the existence (or not) of a single mode and do not focus on the statistical modeling of the data.

We propose a new unimodality test called Unimodal-Uniform test (UU-test) to decide if a set of data values has been a generated by a unimodal distribution or not. The method utilizes the empirical cumulative density function (ecdf) and attempts to obtain a unimodal piecewise linear approximation of the ecdf under the constraint that the data corresponding to each linear segment follow the uniform distribution.

An attractive feature of the proposed approach is that not only it decides on unimodality, but it also produces a generative model of the unimodal data in the form of a mixture of uniform distributions. Thus, it can be used for statistical data modeling. Modeling unimodal data is

typically performed by fitting a specific single unimodal distribution, usually a Gaussian distribution. This approach lacks flexibility since it cannot efficiently model data samples generated by asymmetric distributions. The uniform mixture model produced by the UU-test is able to model unimodal distributions with arbitrary shape.

In the experimental evaluation we conducted, it is shown that the UU-test is effective both in deciding unimodality/multimodality and also in providing accurate statistical models of unimodal data.

# ΕΚΤΕΤΑΜΕΝΗ ΠΕΡΙΛΗΨΗ ΣΤΑ ΕΛΛΗΝΙΚΑ

Παρασκευή Χασάνη
MSc, Τμήμα Μηχανικών Η/Υ και Πληροφορικής, Πανεπιστήμιο Ιωαννίνων
Φεβρουάριος 2019
Τίτλος: UU-τεστ: Ανίχνευση Μονοτροπικότητας Κατανομών χρησιμοποιώντας Τεστ Ομοιομορφίας
Επιβλέπων: Αριστείδης Λύκας

Η αναγνώριση μονοτροπικών κατανομών διαδραματίζει σημαντικό ρόλο στη στατιστική, τη μηχανική μάθηση και την ανάλυση δεδομένων. Γνωστές κατανομές, όπως οι: Κανονική, Student's t, Γάμμα, Χι-τετράγωνο, Εκθετική και Cauchy είναι παραδείγματα μονοτροπικών κατανομών. Επίσης, η Ομοιόμορφη κατανομή είναι μια ακραία περίπτωση μονοτροπικής κατανομής. Η χαρακτηριστική ιδιότητα των μονοτροπικών κατανομών είναι ότι τα δεδομένα βρίσκονται πολύ κοντά σε μία τιμή, η οποία είναι η κορυφή της κατανομής. Εξαιτίας αυτής της ιδιότητας, τα δεδομένα χαρακτηρίζονται ως ομοιογενή, σχηματίζοντας μία συνεκτική ομάδα.

Τα τελευταία χρόνια έχουν προταθεί τεστ μονοτροπικότητας που αποφασίζουν τη μονοτροπικότητα ενός συνόλου δεδομένων, παρέχοντας χρήσιμη γνώση για τη δομή των δεδομένων. Για παράδειγμα, εάν ένα σύνολο δεδομένων είναι μονοτροπικό, οι τιμές των δεδομένων είναι «συγκεντρωμένες», επομένως δεν χρειάζεται η εφαρμογή μεθόδων ομαδοποίησης. Τα τεστ μονοτροπικότητας που έχουν προταθεί αποφασίζουν αποκλειστικά για την ύπαρξη (ή όχι) μοναδικής κορυφής και δεν εστιάζουν στη στατιστική μοντελοποίηση δεδομένων.

Προτείνουμε ένα νέο τεστ μονοτροπικότητας που λέγεται Μονοτροπικό-Ομοιόμορφο τεστ (UU-τεστ) για να αποφασίζουμε εάν ένα σύνολο δεδομένων έχει παραχθεί από μονοτροπική κατανομή ή όχι. Η μέθοδος αυτή χρησιμοποιεί την εμπειρική συνάρτηση κατανομής και προσπαθεί να κατασκευάσει μια μονοτροπική κατά τμήματα γραμμική προσέγγιση αυτής υπό

τον περιορισμό ότι τα δεδομένα που αντιστοιχούν σε κάθε γραμμικό κομμάτι να ακολουθούν Ομοιόμορφη κατανομή.

Ένα ενδιαφέρον χαρακτηριστικό της προτεινόμενης προσέγγισης είναι ότι δεν αποφασίζει μόνο τη μονοτροπικότητα δεδομένων, αλλά συγχρόνως παράγει ένα μοντέλο για μονοτροπικά δεδομένα που έχει τη μορφή μεικτών Ομοιόμορφων κατανομών. Επομένως, μπορεί να χρησιμοποιηθεί για στατιστική μοντελοποίηση δεδομένων. Η μοντελοποίηση μονοτροπικών δεδομένων πραγματοποιείται «ταιριάζοντας» σε αυτά μια συγκεκριμένη μονοτροπική κατανομή, συνήθως την Κανονική κατανομή. Αυτή η προσέγγιση όμως, στερείται ευελιξίας καθώς δεν μπορεί να μοντελοποιήσει αποτελεσματικά δείγματα δεδομένων που έχουν παραχθεί από ασύμμετρες κατανομές. Το ομοιόμορφο μεικτό μοντέλο που παράγει το UU-τεστ μπορεί να μοντελοποιήσει μονοτροπικές κατανομές οποιουδήποτε μεγέθους.

# CHAPTER 1.

# INTRODUCTION



## 1.1 Introduction

Gaining knowledge of data distributions is a significant topic in data analysis. Many problems require assumptions related to the shape of the data in order to be solved. For example, it is important information whether the data are gathered or not. Distributions such as Gaussian, Student's t, Gamma generate data points forming one single and coherent cluster (unimodal distributions).

A probability density function is unimodal, if it has a single mode; a region where the density becomes maximum, while non-increasing density is observed when moving away from the mode. In other words, the data points make a single peak in the histogram. More details about unimodality are provided in section 1.2.

Unimodal distributions form a remarkable subset of probability distributions, which presents nice properties. Location, scale and skewness are important concepts for the description of a probability distribution. The study of tails for skewed distributions presents a particular aspect for unimodal distributions: when we consider the shape of the probability density function, the mode seems to be an appealing center. The mode, as location parameter, does not preserve the stochastic ordering. However, its interpretation makes it more rational in certain circumstances than mean or median.

To decide whether a distribution is unimodal or not (multimodal), we use unimodality tests. The last decades a few tests have been proposed using various properties of the distributions to prove unimodality/multimodality. The significance of these tests is great, since they provide fundamental information about the distributions themselves. For example, running a clustering algorithm in a unimodal dataset may be unnecessary. This dataset consists of homogeneous data values and being unimodal indicates that these values do not need to be clustered into subsets.

We propose a deterministic and non-parametric test for deciding unimodality, called Unimodal-Uniform test (UU-test), that is applied on one–dimensional data values. UU-test decides on the unimodality of data distribution (assuming this distribution is continuous), using the ecdf. Our method tries to build a unimodal piecewise linear approximation of the ecdf that successfully models the data in each interval using the uniform distribution. If an approximation like that cannot be found, UU-test decides that the distribution of the data is not unimodal (i.e. it is multimodal). Moreover, in case of unimodality, UU-test provides a generative model of the data values in the form of a uniform mixture model. Thus, except the decision of unimodality/multimodality, UU-test can also model data generated by unimodal distributions.

To summarize, detecting distribution's properties and trying to create a suitable model fit for a dataset are fundamental topics in the field of data analysis. We propose a new method (UU-test) which solves basic problems related with the property of unimodality. UU-test decides if a dataset is unimodal or not and in case of unimodality it returns a corresponding generative model. Therefore, it is expected to be useful in various problems in machine learning and data science.

## 1.2 Definitions

### 1.2.1 Basics

Some basic definitions are provided to make more clear the rest of the thesis. First, the definitions of a pdf, cdf and ecdf are provided. Subsequently, the terms of unimodality and multimodality are explained with some figures. Finally, the greatest convex minorant and the least concave majorant are explained, since they play essential role to construct our method.

The probability density function (pdf) of a continuous random variable 1-d $X$ with support $S$ is an integrable function $f(x)$ satisfying the following:

1) $f(x)$ is positive everywhere in the support $S$, that is, $f(x) > 0$, for all $x$ in $S$
2) The area under the curve $f(x)$ in the support $S$ is 1, that is: $\int_S f(x)dx = 1$
3) If $f(x)$ is the pdf of $x$, then the probability that $x$ belongs to $A$, where $A$ is some interval, is given by the integral of $f(x)$ over that interval, that is: $P(X \in A) = \int_A f(x)dx$

   More specifically, if $A = [a, b]$, the integral will be $P(a \le X \le b) = \int_a^b f(x)dx$

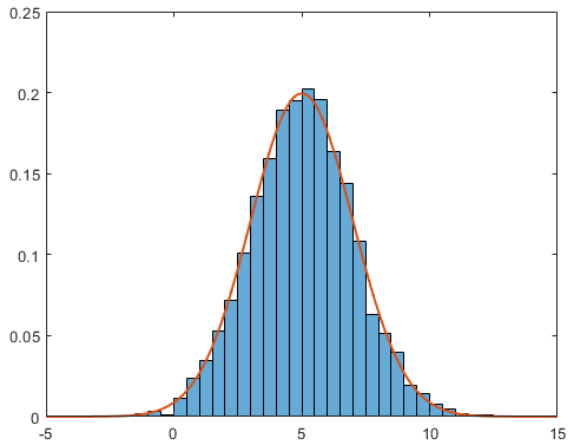



Figure 1.1: Histogram and pdf curve of a Gaussian (μ=5, $\sigma^2$=1) where μ is the mean value and $\sigma^2$ is the variation.

Figure 1.1 illustrates a density histogram where most of the values are close to 5, but some are a bit more and some a bit less. We can represent the probability distribution of *X*, not only as a density histogram, but rather as a curve (by connecting the "dots" at the tops of the rectangles).

The cumulative distribution function (cdf) of a real-valued random variable *X*, or just distribution function of *X*, evaluated at *x*, is the probability that *X* will take a value less than or equal to *x*. The cumulative distribution function of a real-valued random variable *X* is the function given by: $F_X(x) = P(X \leq x)$, where the right-hand side represents the probability that the random variable *X* takes on a value less than or equal to *x*. The probability that *X* lies in the semi-closed interval $(a, b]$, where $a < b$, is therefore $P(a < X \leq b) = F_X(b) - F_X(a)$.

An empirical distribution function (ecdf) is the distribution function associated with the empirical measure of a sample of N data points. It is a step function that jumps up by 1/N at each of the N data points. Its value at any specified value of the measured variable is the fraction of observations of the measured variable that are less than or equal to the specified value. In other words:

Given N ordered points $X_1, X_2, \ldots, X_N$, the ecdf is defined as:

$$F_N(x) = \frac{\text{number of elements in the sample} \leq x}{N} = \frac{1}{N}\sum_{i=1}^{N} I_{[-\infty,x]}(X_i),$$

where $I_{[-\infty,x]}(X_i)$ is the indicator function: $I_{[-\infty,x]}(X_i) = \begin{cases} 1, & \text{if } X_i \leq x \\ 0, & \text{otherwise} \end{cases}$

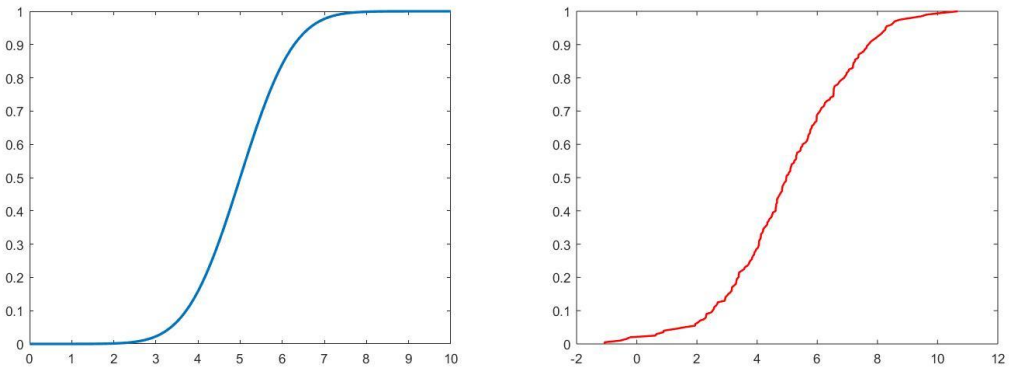



Figure 1.2: Cdf and ecdf of a Gaussian ($\mu$=5, $\sigma^2$=1).

### 1.2.2 Unimodality-Multimodality

Next, we provide the definitions of a unimodal and multimodal function. Note that there are two kinds of definitions depending on the type of distribution function.

A pdf is unimodal, if it has a single mode; a region where the density becomes maximum, while non-increasing density is observed when moving away from the mode. In other words, a function $f(x)$ is a unimodal function if for some value $m$, it is monotonically increasing for $x \leq m$ and monotonically decreasing for $x \geq m$. In that case, the maximum value of $f(x)$ is $f(m)$ and there are no other local maxima. The most widely used unimodal distribution function is the Gaussian. Other examples of unimodal functions are Student's t, Gamma, Chi-square, Cauchy and Exponential.

In the figures (histograms) below, the single mode is obvious in the case of Gaussian, Student's t and Gamma distribution.

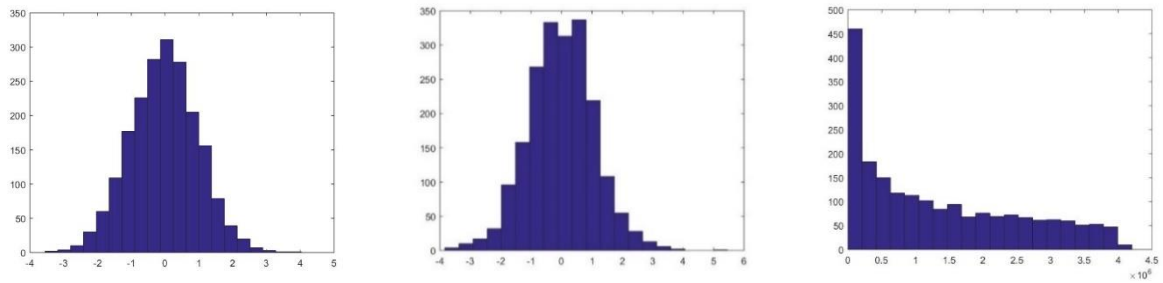


Figure 1.3: Histograms of a Gaussian, Student's t and Gamma distribution.

A cdf is unimodal if two points $x_l$ and $x_u$ exist such that the function can be divided into three parts: a) a convex part $(-\infty, x_l)$, b) a constant part $[x_l, x_u]$ and c) a concave part in $(x_u, \infty)$. It is worth mentioning that it is possible for either the first two parts or the last two parts of unimodality's definition to be missing.

Figure 1.4 illustrates the ecdfs of the previous unimodal functions. In the first two ecdfs, we clearly see first a convex part, after a linear, and last a concave part. In the third (Gamma function), the convex and linear parts are missing, but it still remains unimodal according to

unimodality's definition. Here, we need to mention that the red points in the figures of this thesis correspond to ecdf's values.

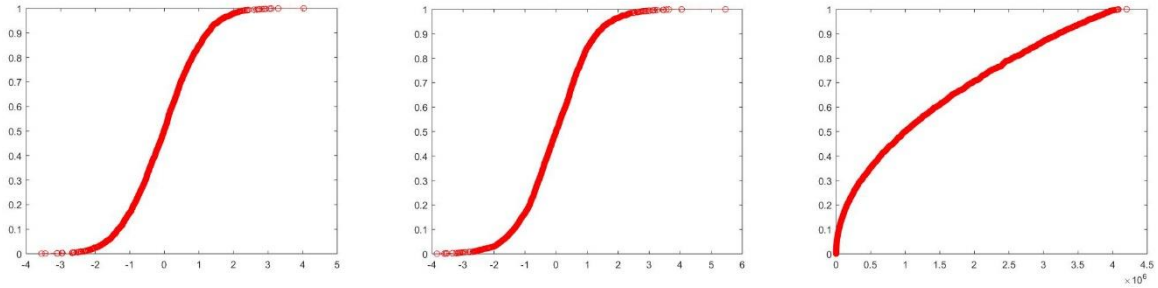


Figure 1.4: Ecdf figures of a Gaussian, Student's t and Gamma distribution.

Moreover, we should mention that the Uniform distribution is an extreme single mode case where the mode covers all the region with non-zero density. The cdf plot only contains the linear part of unimodality's definition (Figure 1.5).

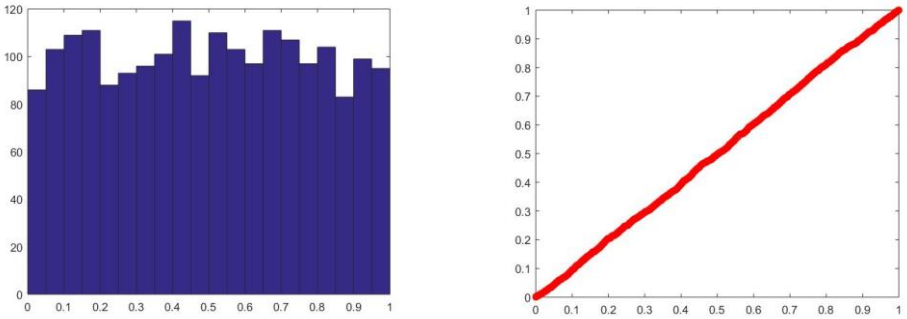


Figure 1.5: Histogram and ecdf figure of a Uniform distribution (a=0, b=1) where a is the minimum value and b is the maximum value.

On the other hand, a non-unimodal distribution is called multimodal with two or more modes. A common case is when a distribution has only two modes; these appear as distinct peaks (local maxima) in the pdf plot. The distribution with exactly two modes is called bimodal, while the distribution with exactly three modes is called trimodal. A bimodal distribution most commonly arises as a mixture of two different unimodal distributions (i.e. distributions having only one mode). For example, a mixture of two Gaussian distributions with the same variance, but different means is a bimodal distribution.

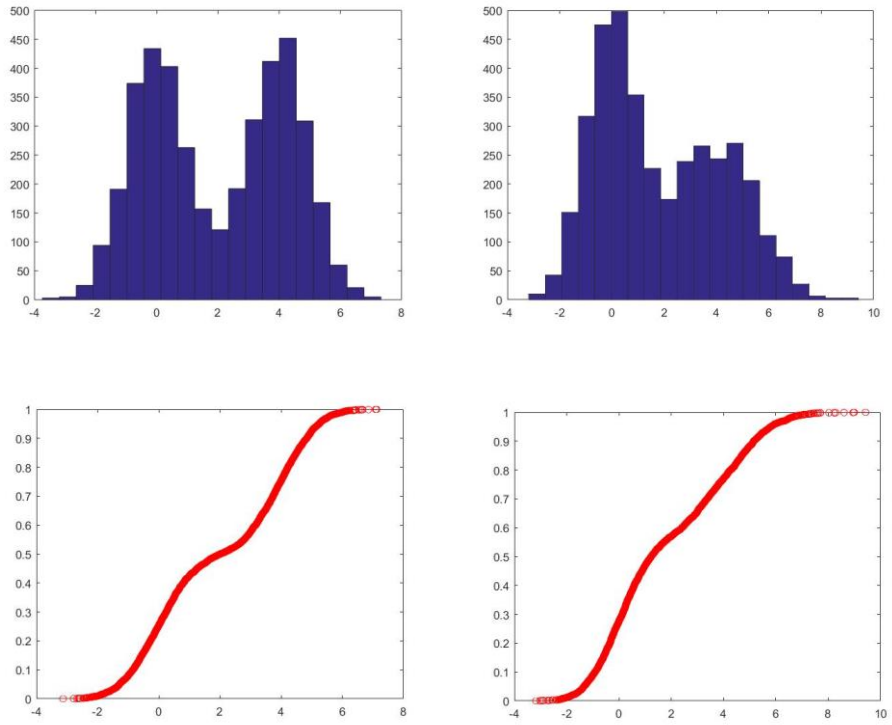


Figure 1.6: Histograms and corresponding ecdf figures of bimodal distributions.

In Figure 1.6, the existence of exactly two modes is clear. In the ecdf plots, we can observe the convex, linear and concave parts, and in the middle of the figure, a convex part is appearing again. This is against the unimodality's definition, so the two functions are multimodal (bimodal in this case).

The definition of unimodality in the case of pdf functions is sufficiently comprehensible and simple, while in the case of cdf functions may be a little tricky. However, the shape of a histogram - in which we visualize a pdf - varies depending on the number of bins (buckets), while a cdf plot independent of any parameters. So, the cdf has the clear advantage over the pdf, as a more stable and handier tool, and it is used in our method.

In Figure 1.7 we see the histograms of the same pdf for three different numbers of bins (10, 20, 50). This number affects significantly the plot and if we depend on the number of bins, we may end up in different conclusions.

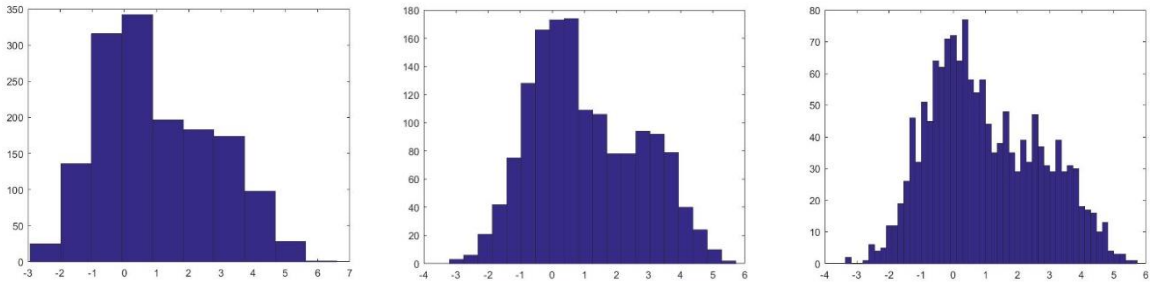


Figure 1.7: Histograms of a distribution function. The number of bins is 10, 20 and 50, respectively.

Moreover, since the underlying distribution function is not known, and we work with sample observations, we choose ecdf over cdf in our method. The ecdf is useful, since it approximates the true cdf well if the sample size (the number of data) is large, and knowing the distribution is helpful for statistical inference. Also, a plot of the ecdf can be visually compared to known cdfs of frequently used distributions to check if the data came from one of those common distributions.

### 1.2.3 Greatest Convex Minorant-Lowest Concave Majorant

A very useful and significant part of our method is the computation of the greatest convex minorant and the lowest concave majorant functions. We provide below the definitions of the two functions.

The greatest convex minorant (gcm) of a function F in $(-\infty, a]$ is sup G(x) for $x \le a$, where the sup is taken over all functions G that are convex in $(-\infty, a]$ and nowhere greater than F.

The least concave majorant (lcm) of a function F in $[a, \infty)$ is inf L(x) for $x \ge a$, where the inf is taken over all functions L that are concave in $[a, \infty)$ and nowhere less than F.

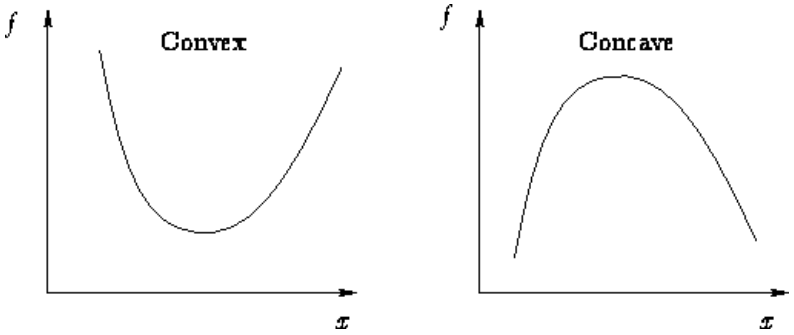



Figure 1.8: A Convex and a Concave distribution.

The gcm/lcm points of the ecdf of a unimodal and two bimodal distributions are illustrated in Figure 1.9. The blue circles on the lower line are the gcm's points of contact with F, while the black circles on the upper line are the lcm's points. The former are called gcm points and the latter lcm points. These points are a great approximation of function F, since they can be characterized as a lower and upper limit of F.

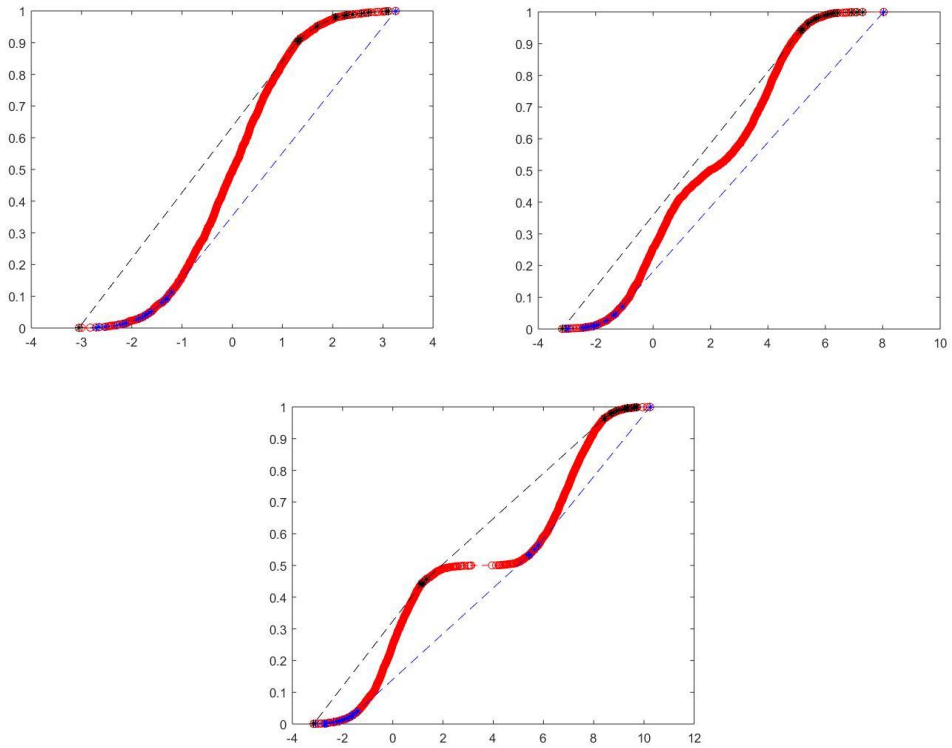


Figure 1.9: Visualization of gcm and lcm points of an ecdf. The blue circles on the lower line are the gcm points, while the black circles on the upper line are the lcm points.

## 1.3 Thesis Roadmap

The structure of this thesis is organized as follows. In Chapter 1, we provide basic information about the issue of unimodality and explain definitions related with the problem. Chapter 2 presents the basic background in statistical tests, describes the related work that focuses on unimodality tests and explains the usage of them. In Chapter 3, we present UU-test, a new unimodality test and analyze step by step the whole procedure. Then, we provide the experimental results on deciding distribution unimodality comparing to a popular alternative, which is the dip test. In Chapter 4, we begin with an introduction to statistical modeling and mixture distributions and continue with the statistical model provided by our test. Moreover, we compare this model with the Gaussian and Uniform model, on several datasets. Finally, Chapter 5 concludes this thesis by summarizing our findings and also presents potential future work.

# CHAPTER 2.

# RELATED WORK



## 2.1 Statistical Tests

### 2.1.1 Basics

Statistical tests are used to prove various properties in the field of statistics. They provide a mechanism for making quantitative decisions about a process of interest. In other words, we use statistical tests to decide whether a pattern we observe is due to chance or due to the program or intervention effects. Research often uses them to determine if there is a relationship between an intervention and an outcome as well as to quantify the strength of that relationship.

At first, a test statistic (a quantity derived from the sample) is computed and is considered as a numerical summary of a data-set that reduces the data to one value. The intent is to determine whether there is enough evidence to "reject" a conjecture or hypothesis about the process. The conjecture is called the null hypothesis. Not rejecting may be a good result if we want to continue to act as if we "believe" the null hypothesis is true. Or it may be a disappointing result, possibly indicating we may not yet have enough data to "prove" something by rejecting the null hypothesis.

The test statistic is used in testing the statistical hypothesis and is selected or defined in such a way as to quantify, using observed data, behaviors that would distinguish the null from the alternative hypothesis.

A common format for a hypothesis test is:

$H_0$: A statement of the null hypothesis, e.g., two population means are equal.

$H_a$: A statement of the alternative hypothesis, e.g., two population means are not equal.

Test Statistic: The test statistic is based on the specific hypothesis test.

Significance Level: The significance level, α, defines the sensitivity of the test. A value of $\alpha = 0.05$ means that we inadvertently reject the null hypothesis 5% of the time when it is in fact true. This is also called the type I error. The choice of α is somewhat arbitrary, although in practice values of 0.1, 0.05, and 0.01 are commonly used.

Common test statistics are one-sample, two-sample and paired tests.

One-sample tests are appropriate when a sample is being compared to the population from a hypothesis. The population characteristics are known from theory or are calculated from the population.

Two-sample tests are appropriate for comparing two samples, typically experimental and control samples from a scientifically controlled experiment.

Paired tests are appropriate for comparing two samples where it is impossible to control important variables. Rather than comparing two sets, members are paired between samples so the difference between the members becomes the sample. Typically, the mean of the differences is then compared to zero. The common example scenario for when a paired difference test is appropriate is when a single set of test subjects has something applied to them and the test is intended to check for an effect. For example, if we compare the weight of every person in a group of people before they went on a diet with their weight after they completed the diet program.

According to the null and alternative hypothesis, there are two kinds of tests: the two-sided and one-sided tests (or two-tailed and one-tailed tests). A two-tailed test is appropriate if the estimated value may be more than or less than the reference value, for example, whether a test

taker may score above or below the historical average. A one-tailed test is appropriate if the estimated value may depart from the reference value in only one direction, for example, whether a machine produces more than one-percent defective products.

A measure for evaluating the result of a test of hypothesis is Critical values. Critical values for a test of hypothesis depend upon a test statistic, which is specific to the type of test. They are essentially cut-off values that define regions where the test statistic is unlikely to lie; for example, a region where the critical value is exceeded with probability $\alpha$ if the null hypothesis is true. The null hypothesis is rejected if the test statistic lies within this region which is often referred to as the rejection region(s).

Another quantitative measure for reporting the result of a test of hypothesis is the p-value. The p-value is the probability of the test statistic being at least as extreme as the one observed given that the null hypothesis is true. A small p-value is an indication that the null hypothesis is false. The benefit of using p-value is that it calculates a probability estimate, we can test at any desired level of significance by comparing this probability directly with the significance level. It is good practice to decide in advance of the test how small a p-value is required to reject the test. This is exactly analogous to choosing a significance level, $\alpha$, for test. For example, we decide either to reject the null hypothesis if the test statistic exceeds the critical value (for $\alpha = 0.05$) or analogously to reject the null hypothesis if the p-value is smaller than 0.05.

Known statistical tests are: Z-test, T-test and Chi-square. In Z-test, the sample is assumed to be normally distributed and a z-score is calculated with population parameters, such as “population mean” and “population standard deviation” and is used to validate a hypothesis that the sample drawn belongs to the same population. A T-test is used to compare the mean of two given samples. Like a Z-test, a T-test also assumes a Normal distribution of the sample. A T-test is used when the population parameters (mean and standard deviation) are not known. Chi-square test is used to compare categorical variables. There are two types of Chi-square tests: (1) Goodness of fit test, which determines if a sample matches the population and (2) a Chi-square fit test for two independent variables is used to compare two variables in a contingency table to check if the data fits.

### 2.1.2 Gaussianity (Normality) Tests

There is a large number of tests for determining if a dataset is well-modeled by a Normal distribution and computing how likely it is for a random variable underlying the dataset to be normally distributed. In Statistics, these tests are called Gaussianity or Normality tests. An informal and simple approach to testing normality is to compare a histogram of the sample data to a Normal probability curve. The empirical distribution of the data (the histogram) should be bell-shaped and resemble the Normal distribution. This might be difficult to observe if the sample is small. Another test of normality is Shapiro-Wilk test, which tests the null hypothesis that a sample $x_1, ..., x_n$ comes from a normally distributed population.

There are more general tests, which check if a dataset is well-modeled not only by a Normal distribution but also by other distributions. For example, the Anderson–Darling test is a statistical test of whether a given sample of data is drawn from a given probability distribution. It works well for distributions such as: Normal, Exponential, Extreme-value, Weibull, Gamma, Logistic, Cauchy, etc. Kolmogorov-Smirnov test (KS test) is another statistical test which determines if a dataset comes from a given population. In our method we use KS test, so we are more specific in the next section.

### 2.1.3 Kolmogorov-Smirnov Test

The Kolmogorov–Smirnov (KS test) is a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample KS test), or to compare two samples (two-sample KS test).

The KS statistic quantifies a distance between the ecdf of the sample and the cdf of the reference distribution, or between the ecdfs of two samples. The null distribution of this statistic is calculated under the null hypothesis that the sample is drawn from the reference distribution (in the one-sample case) or that the samples are drawn from the same distribution (in the two-sample case). The KS test can be modified to serve as a goodness of fit test. The goodness of fit of a statistical model describes how well it fits a set of observations. In the

special case of testing for normality of the distribution, samples are standardized and compared with a standard normal distribution.

The KS test is defined as:

$H_0$: The data follow a specified distribution.

$H_a$: The data do not follow the specified distribution.

The KS test statistic is defined as: $D_N = \sup_x |F_N(x) - F(x)|$, where $\sup_x$ is the supremum of the set of distances, $F_N(x)$ is the ecdf and $F(x)$ is the cdf (specified distribution).

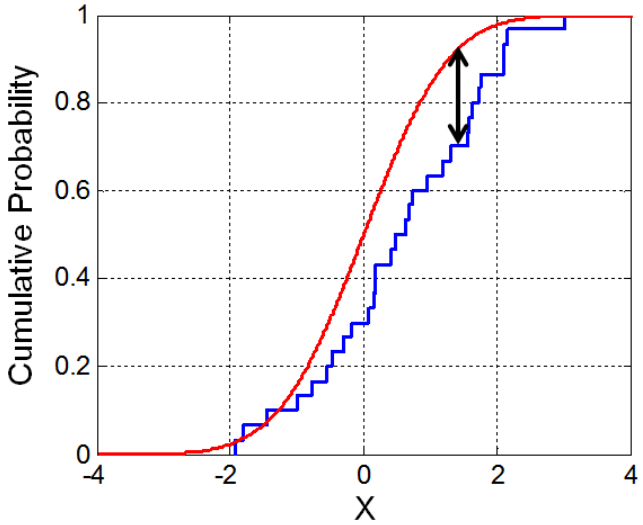



Figure 2.1: Illustration of the Kolmogorov–Smirnov statistic. Red line is cdf, blue line is an ecdf, and the length of the black arrow is the KS statistic.

The KS test decides to reject the null hypothesis by comparing the p-value with the significance level α, not by comparing the test statistic with the critical value. Since the critical value is approximate, comparing the statistic with the critical value occasionally leads to a different conclusion than comparing p-value with α.

As we mentioned, p-value is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. Small values of p cast doubt on the validity of the null hypothesis. Therefore, if p-value ≤ α, the KS test will reject the null hypothesis, otherwise, it will accept it. KS test computes the critical value using an

approximate formula or by interpolation in a table. The formula and table cover the range $0.005 \leq \alpha \leq 0.1$ for one-sided tests.

For testing uniformity, we use one-sample KS test with reference distribution being the Uniform distribution and significance level $\alpha$=0.01. KS test decides if the ecdf's segments come from a uniform population (are uniformly distributed).

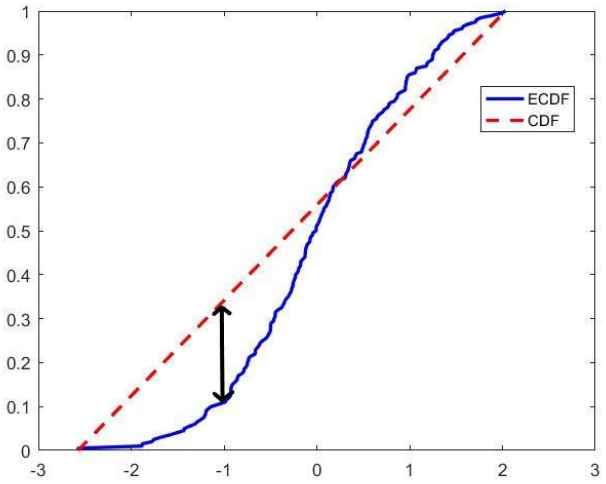



Figure 2.2: Plot of an ecdf with a Uniform cdf for 200 uniform random numbers. The KS test is based on the maximum distance between these two curves (black arrow).

An attractive feature of this test is that the distribution of the KS test statistic itself does not depend on the underlying cumulative distribution function being tested. Another advantage is that it is an exact test (e.g. the chi-square goodness-of-fit test depends on an adequate sample size for the approximations to be valid). Despite these advantages, the KS test has one important limitation which affects our method. It tends to be more sensitive near the center of the distribution than at the tails. This is a difficulty we encountered using KS test in unimodal distributions with a large dataset size. When we tested the middle linear part of a unimodal ecdf (e.g. Gaussian, Student's t), KS test sometimes failed to accept the null hypothesis that this part is drawn from the uniform distribution. More details are given in Chapter 3.

## 2.2 Unimodality Tests

### 2.2.1 Hartigans' Dip Test

There are few statistical tests for discovering the presence of more than one mode in a distribution, known as unimodality tests. One test, suggested in [1], uses the likelihood ratio for a two-component normal mixture against the normal null hypothesis. A related test [2] divides the sample into two subsets to maximize the likelihood ratio that the two subsets are sampled from normal with different means, against the null hypothesis that the means are equal. In [3] a test is suggested for multimodality, called k-critical windows. The idea is the following: The smallest window width is used, such that the resulting kernel density estimate is unimodal, as a test statistic for unimodality. A sample from a density with more than k modes will require more smoothing to exhibit k or less modes in the density estimate compared to a sample from a density with exactly k modes. The significance level of the test statistic is evaluated by empirically sampling from a rescaled version of the unimodal density estimate.

Another unimodality test is **Hartigans'** dip test [4], which does have the benefit of not requiring a kernel width. It is the most widely used test that computes the dip statistic as the maximum difference between the ecdf, and the unimodal distribution function that minimizes that maximum difference. The Uniform distribution is the asymptotically least favorable unimodal distribution, and the distribution of the test statistic is determined asymptotically and empirically when sampling from the Uniform.

The Hartigans' dip test [4] is a notable test statistic which decides on the unimodality of a real-valued dataset. It takes as input an 1-d dataset, examines the underlying ecdf of the set of numbers and decides whether it contains a single or more than one mode (peak). Given a set of real numbers $X = \{x_1, x_2, \ldots, x_n\}$ the dip-test computes the dip value, which is the departure from unimodality of the ecdf $F_n(x) = \frac{1}{n}\sum_n I(x_i \le x)$.

For bounded input functions $F, G$, let $\rho(F,G) = \max_x |F(x) - G(x)|$, and let $\boldsymbol{U}$ be the class of all unimodal distributions. Then the dip statistic of a distribution function *F* is given by:

$$dip(F) = \min_{G \in \boldsymbol{U}} \rho(F,G).$$

In other words, the dip statistic computes the minimum among the maximum deviations observed between the cdf $F$ and the cdfs from the class of unimodal distributions. A nice property of dip is that, if $X$ is a sample distribution of *n* observations from *F*, then $\lim_{n\to\infty} dip(F_n) = dip(F)$. It is argued that the class of Uniform distributions *U* is being used for the null hypothesis, since its dip values are stochastically larger than other unimodal distributions, such as those having exponentially decreasing tails.

Given a dataset $X = \{x_1, ..., x_n\}$, $x_i \in \Re$ the dip statistic is computed as follows:

i. Begin with $x_L = x_1, x_U = x_{n,} D = 0$.

ii. Compute the gcm G and lcm L for F in $[x_L, x_U]$; suppose the points of contact with F are respectively $g_1, g_2, \ldots, g_k$ and $l_1, l_2, \ldots, l_m$.

iii. Suppose $d = \sup|G(g_i) - L(g_i)| > \sup|G(l_i) - L(l_i)|$ and that the sup occurs at $l_j \le g_i \le l_{j+1}$. Define $x_L{}^0 = g_i, x_U{}^0 = l_{j+1}$.

iv. Suppose $d = \sup|G(l_i) - L(l_i)| > \sup|G(g_i) - L(g_i)|$ and that the sup occurs at $g_i \le l_j \le g_{i+1}$. Define $x_L{}^0 = g_i, x_U{}^0 = l_j$.

v. If $d \le D$, stop and set $D(F) = D$.

vi. If $d > D$, set $D = \sup\left\{D, \sup_{x_L \le x \le x_L{}^0} |G(x) - F(x)|, \sup_{x_U{}^0 \le x \le x_U} |L(x) - F(x)|\right\}$

vii. Set $x_U = x_U{}^0, x_L = x_L{}^0$ and return to ii.

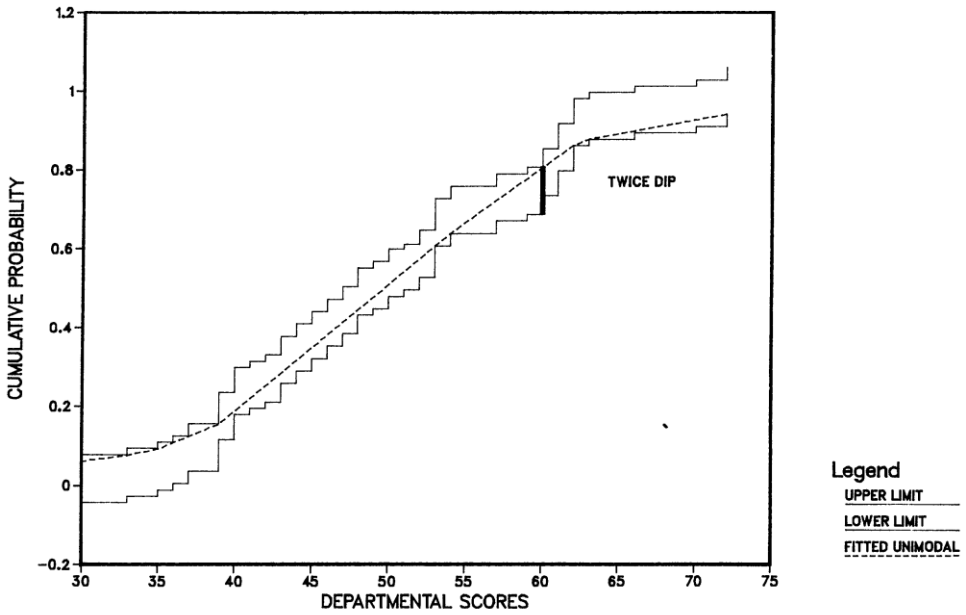



Figure 2.3: Visualization of dip quantity [4].

Dip test examines the $n(n-1)/2$ possible modal intervals $[x_L, x_U]$ between the sorted $n$ individual observations. For all these combinations it computes in $O(n)$ time the respective gcm and the lcm curves in $(\min_x X_n, x_L)$ and $(x_U, \max_x X_n)$, respectively. Fortunately, for a given $X_n$, the complexity of one dip computation is $O(n)$. The dip test returns not only the dip value, but also the statistical significance of the computed dip value, i.e. a p-value. The computation of the *p*-value for a unimodality test uses bootstrap samples and expresses the probability of $dip(X_n)$ being less than the dip value of a cdf $U_n^r$ of $n$ observations sampled from the $U[0,1]$ Uniform distribution:

$$P = \#[dip(X_n) \le dip(U_n^r)] / b, \; r = 1, \ldots, b$$

It should be stressed that for each value of $n$, the bootstrap samples $U_n^r$ do not depend on the dataset $X$, therefore they can be computed only once, along with the corresponding values $dip(U_n^r)$. The null and alternative hypothesis, $H_0$ and $H_a$, are given below:

$H_0$: $X_n$ is unimodal  $H_a$: $X_n$ is multimodal

$H_0$ is accepted at significance level α if p-value > α, otherwise $H_0$ is rejected in favor of the alternative hypothesis $H_a$, which suggests multimodality.

### 2.2.2 The Folding test

In [5] a multivariate unimodality test is proposed, which makes no distribution assumption and utilizes only a p-value. Given a multidimensional dataset of numerical attributes, the authors wonder about the "grouping behavior" of the data points. In Chapter 1, it was mentioned that a unimodality test can be easily utilized to decide if it needs to run a clustering algorithm or not. In [5] it is argued that in unimodal cases, the data points make a single peak in the histogram. It is obvious that if this fact is known, we will not proceed in a clustering method, since our data points make exactly one group (cluster).

The folding test of unimodality is proposed in [5]; a test which relies on a folding technique for univariate and multivariate cases. Their approach is the following: (1) fold up the distribution with respect to a pivot s*, (2) compute the variance of the folded distribution and (3) compare it with the initial variance. The main idea is that the resulting density of the folded distribution will have a far lower variance in multimodal distributions, while this phenomenon will not appear in unimodal cases (i.e. not with the same amplitude).

The folding step is performed with the transformation X ↦ |X - s*| and the folding ratio is computed as:

$$\varphi(X) = \frac{Var|X - s*|}{VarX}$$

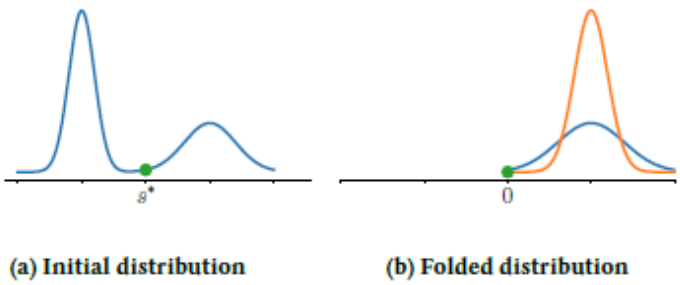
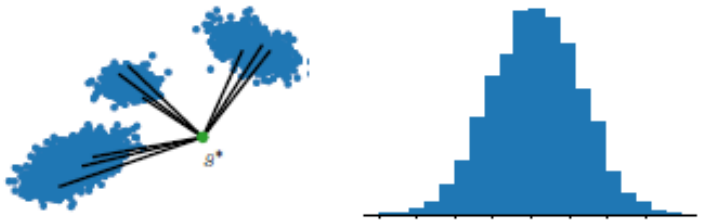



Figure 2.4: (a) Initial distribution (b) Folded distribution with respect to a pivot s*, in univariate case [5].

In higher dimensions the absolute value is replaced by the Euclidean norm $Var\|X - s*\|$ where X is a random vector of $\mathbb{R}^d$ and $s* \in \mathbb{R}^d$ is the pivot. The variance is replaced by $\mathrm{E}\left[\|X - \mathrm{E}[X]\|^2\right]$. In the unimodal case this expected value will be much lower than in multimodal case. Finally, the folding ratio is generalized through:

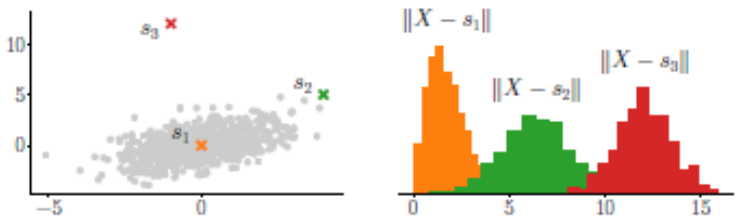
$$\varphi(X) = \frac{Var\|X - s*\|}{\mathrm{E}\left[\|X - \mathrm{E}[X]\|^2\right]}$$




(a) Initial distribution (b) Folded distribution

Figure 2.5: (a) Initial distribution (b) Folded distribution with respect to a pivot s*, in multivariate case [5].

To this end, more details should be given about pivot. According to [5], the right pivot should be found, so the variance can be significantly reduced through the folding process. Thus, the pivot should reduce the variance the most (if such a pivot exists). It is mentioned that the best pivot s* is likely to be close to the mode in the unimodal case, while is likely to stand "between" the modes in the multimodal case.




(a) Initial distribution (b) Folded distribution

Figure 2.6: Impact of the pivot location [5].

The best pivot s* is found as:

$$s^* = \arg\min_{s \in \mathbb{R}^d} Var \left\| X - s \right\|$$

The level of confidence of the test is computed with a p-value. The lower p-value is, the more significant the decision will be.

### 2.2.3 Nonparametric Testing of the Existence of Modes

In [6] a test is proposed for the weight of evidence of individual observed modes. The test statistic used is a measure of the size of the mode, the absolute integrated difference between the estimated density and the same density with the mode in question excised at the level of the higher of its two surrounding antimodes. More sprecifically,

$$M_i = \int_{u_{i-1}}^{u_{i+1}} [\, f(x) - \max(f(u_{i-1}),\, f(u_{i+1}))]_+ \, dx$$

$M_i$ is the minimal $L_1$ distance from the density estimate to the set of continuous functions without a local maximum between the observed antimodes in the density function. $M_i$ can be thought of as the area or probability mass of the mode above the higher of the two surrounding antimodes. The decision to use the smallest possible bandwidth makes sense; the smaller the bandwidth, the higher the modes and lower the antimodes, and the greater the probability mass in the region above the higher antimode.

Samples are simulated from a conservative member of the composite null hypothesis to estimate p-values within a Monte Carlo setting. Such a test can be used with the graphical "mode tree" [7] to examine, in a locally adaptive fashion, not only the existence of individual modes, but also (roughly) the overall number of modes of the density.

## 2.3 Exploitation of Unimodality Tests

The term of unimodality and various unimodality tests have been used in various scientific fields such as biology, ecology, etc.

In ecology, for example, the search for predictions of species diversity across environmental gradients has challenged ecologists for decades. [8] presents this topic and uses unimodality or HBM. The humped-back model (HBM) states that plant species richness peaks at intermediate productivity, taking above-ground biomass as a proxy for annual net primary productivity. This diversity peak is driven by two opposing processes. In unproductive ecosystems with low plant biomass, species richness is limited by abiotic stress, such as insufficient water and mineral nutrients, which few species are able to tolerate. In contrast, in the productive conditions that generate high plant biomass, competitive exclusion by a small number of highly competitive species is hypothesized to constrain species richness.

Over time the HBM has become increasingly controversial, and some studies refuse it. In [8] the authors provide evidence in support of the HBM pattern of the plant richness-productivity relationship at both global and regional extents. They used a global-extent regression model and found that plant richness formed a unimodal relationship with productivity that is characterized by a highly significant concave-down quadratic regression (negative binomial generalized linear model (GLM)). This relationship was not sensitive to the statistical model used; the hump-backed relationship was also evident when they used a negative binomial generalized linear mixed model (GLMM).

In the field of biology, in [9] Hartigans' dip and Silverman's test of unimodality are used in cytometry. Many automated gating algorithms for flow cytometry data are based on the concept of unimodal cell populations. A key problem is how to determine whether a cell subset represents one or multiple-possibly overlapping-cell populations. In manual gating this is left to the judgment of the analyst, supported by various visualizations. In automated gating the decision is often based on whether a group of events is well described by a density with a single peak-a unimodal density-or not.

In [9] it is shown that criteria previously used to make decisions on unimodality cannot adequately distinguish unimodal from bimodal densities. Moreover, it is shown that dip and bandwidth tests for unimodality can do this with consistent and low error rates. These tests

also have the possibility to adjust the significance level to handle the trade-off between failing to detect a second mode and observing a second mode when there is none.

In data analysis it is of great importance to discover information about the structure in data. There are significant topics, such as clustering, which are only appropriate when cluster structure is present. In [10] it is presented in a nice way why a unimodality test is so important in cluster analysis. First, it needs to mention the meaning of clusterability. Clusterability [10] depends on the presence of inherent structure and aims to quantify the degree of cluster structure. This analysis should precede the application of clustering algorithms, as the success of any clustering algorithm depends on the presence of underlying cluster structure. If such a structure exists, the next step of choosing a clustering algorithm will follow. In other cases, the results of any clustering technique become arbitrary and potentially misleading, so clustering possibly should not be applied.

For concreteness, consider a dataset randomly generated from a single Gaussian distribution. Because the data contains only one cluster, it makes no sense to divide the dataset into clusters. Most clustering algorithms (e.g. k-means with $k \geq 2$) would find multiple clusters in the data, even though no multi-cluster structure is present. Before the clustering algorithm, we could have checked with a unimodality test, if the dataset makes one single and coherent cluster or not. In other words, we would know that the data are homogeneous, so clustering would not be suitable in this case.

On the other hand, if a dataset contains multiple clusters, then there should be some separation between the clusters. For example, consider a dataset randomly generated from two Gaussian distributions with the two means being extremely different. There are two peaks in the Gaussians' histogram, so there are two clusters in the dataset and a unimodality test will decide multimodality. Thus, running a unimodality test before the clustering algorithm ensures us that the cluster structure is present in the dataset. When the data are multi-dimensional is a severe limitation which renders these unimodality methods unpredictable for real datasets, most of which have multiple, if not high dimensions, unless the user first reduces the data to one dimension.

In [11] the dip-means algorithm is presented, a combination of the dip statistic with the k−means algorithm. To decide on the content homogeneity of a new data region, the dip-dist

criterion is proposed for evaluating the cluster structure of a set of data objects. This criterion is based on Hartigans' dip test for unimodality.

The basic intuition behind dip-dist is that if the density distribution of a set of objects is unimodal, then the set is considered homogeneous. It is noteworthy that unimodality is tested using only the pairwise distances between data objects (i.e. the distance matrix). This implies that it is not checked in the original data space. Computing the dip-dist criterion, for a set of objects, is described as follows: each object of the set is treated as viewer that decides on the unimodality of the set by considering the set of the pairwise distances from the viewer to all other data objects. Then, the density of this set of distances is tested for unimodality using Hartigans' dip test and is characterized as either unimodal or multimodal. The viewer decides either unimodality or multimodality and is characterized as either unimodal or multimodal viewer, respectively. In the dip-dist criterion all data objects of the group are considered as viewers. If the percentage of viewers suggesting multimodality exceeds a given threshold, then the set of objects is characterized as multimodal, otherwise it is considered unimodal. In other words, if k-means provides a cluster which is not unimodal, dip-means re-run the algorithm assuming an additional cluster (with an initial state carefully chosen).

As unimodality test, Hartigans' dip test is also used in [12]. This work presents an almost parameter-free method - the DipTransformation - that is able to improve the structure of a dataset and thus allows k-means to cluster datasets better. Moreover, it is deterministic and requires no distance calculations. The algorithm does not assume a special distribution for the clusters or data. It simply enhances structure and thereby improves clustering. Thus, it is not only a preparatory technique for k-means, it can also be used to improve performance for various clustering techniques. The idea is the following: the dip test returns a value, the dip statistic, which provides a measure of the structure of a dimension and thus a measure of the "relevance" of the dimension. The more relevant a dimension is, the larger it will be scaled (in relation to the other dimensions) and the greater its influence on the clustering result from k-means.

In [13] it is demonstrated how the dip can be used in projection pursuit. Projection pursuit is the systematic search for interesting low-dimensional linear projections of high-dimensional data. Dimension reduction is an important problem in exploratory data analysis and visualization – the human ability for pattern recognition can only be exploited for very low dimensional data. Machine learning methods can also take advantage of this method since it

provides a viable way to overcome the “curse of dimensionality” problem. Projection pursuit optimizes projection indices, which increase with the interestingness of the projection image. Most classical approaches equate interestingness with non-gaussianity. However, in [13] the dip is not used to measure the distance of distributions from Gaussian distributions. Dip is a way of measurement of distance from the class of unimodal distributions in general. The authors establish properties and develop efficient algorithms to search for projections maximizing the dip (multimodal data) and extend them to find multiple interesting projections. They discuss two approaches to find higher dimensional projections. The first method is based on an iterative orthogonal search, where one fixes k-1 orthogonal directions already found and optimizes the selection of the k-th orthogonal direction in the remaining subspace. The second method removes the interesting structure among each interesting direction, resulting in a recursive procedure. In case of the dip, this means making the distribution unimodal along these directions. According to their experiments, they indicate that the dip is a highly robust projection index, successfully identifying interesting directions, even in very high dimensional spaces, with a minimum of data preprocessing.

# CHAPTER 3.

## THE UNIMODAL-UNIFORM TEST



## 3.1 UU-test description

In previous chapters, we explained why the property of unimodality is significant for a dataset and why unimodality tests are being useful in data analysis. We provided definitions related with unimodality and we described the methods used in current unimodality tests.

Here, we propose the UU-test (Unimodal-Uniform test), a deterministic non-parametric unimodality test that is applied on one-dimensional data values and decides on the unimodality of their ecdf (assuming this distribution is continuous). It is possible to identify unimodal distributions such as Gaussian, Student's t, Gamma, Exponential, Cauchy, etc. The main idea is the following: we try to build a unimodal piecewise linear approximation of the ecdf that models adequately the data points. This means that the data points in each line segment follow the Uniform distribution as indicated by the decisions of a uniformity test (KS test). If an approximation like that cannot be found, UU-test decides that the distribution of the data is multimodal.

First, the desirable approximation must be unimodal according to the definition of unimodality in a case of a cdf. It is optional for all three parts - the convex, the linear and the concave – to exist, if there are though, they should strictly be arranged in the above order. For example, in Chapter 1, Figure 1.4 (third ecdf) illustrates a Gamma distribution with the cdf having only the concave part. Furthermore, the convex and concave parts indicate the monotonically increasing and decreasing parts of a pdf, respectively, while the linear part is related with the maximum value of the function- the mode region (peak)- indicating there are no other local maxima. Therefore, the specific order mentioned above, should be explicitly followed to decide unimodality.

In Figure 3.1, the convex part follows the concave part in the ecdf. This ecdf is generated from two Truncated Gaussians ($\mu_1$=0, $\mu_2$=4, $\sigma_1^2$=$\sigma_2^2$=1). It is a case of a multimodal function and does not satisfy the conditions of the definition of unimodality.

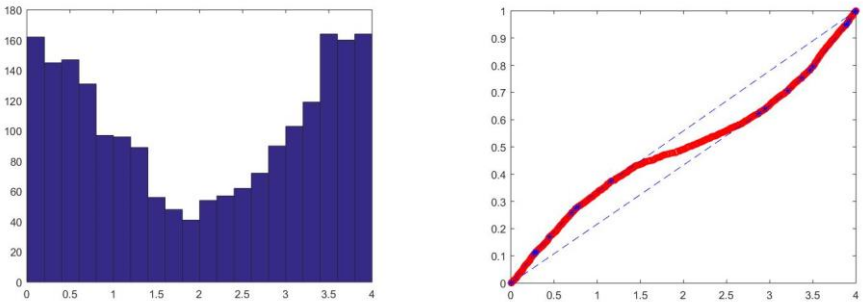


Figure 3.1: Example of a multimodal distribution. The ecdf has the concave part first and the convex part after.

It is equally important for the approximation to be piecewise linear. The UU-test uses the ecdf which is piecewise constant. So, its gcm and lcm functions are piecewise linear and therefore, piecewise linear is the closest unimodal approximation of the ecdf. The Uniform distribution has a linear cdf as shown in Figure 1.5 and it is a case of unimodal distribution, thus we want the successive data points to form linear segments, i.e. those sets of points to follow the Uniform distribution. To check uniformity, we use the KS test. In Figure 3.2, we provide an example of an ecdf which seems to satisfy the definition of unimodality. This happens if here is a convex part in [-4, 1], a linear part in [1,3] and a concave part in [3, 8]. Nevertheless, the histogram illustrates a bimodal distribution with two peaks. The difference between Figures 3.2 and 1.4 is that the intermediate part is not linear (uniform). More specifically, in Figure

3.2 there are two convex and concave parts in the ecdf, which is against the definition of unimodality.

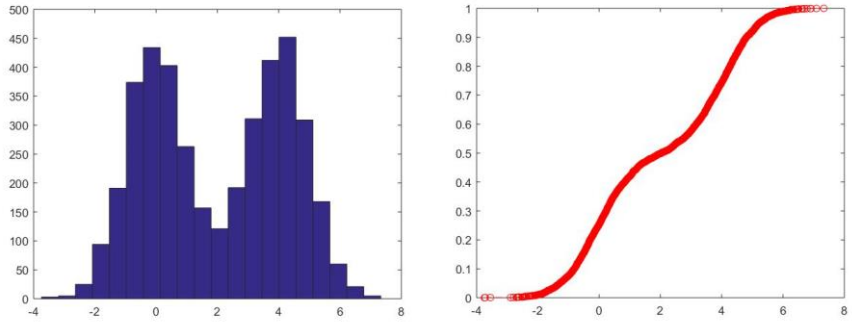



Figure 3.2: Example of a bimodal distribution. The ecdf has two convex parts and two concave parts.

## 3.2 Main methodology of UU-test

At first UU-test computes the ecdf F($x$) of the dataset $X = \{x_1, ..., x_n\},\ x_i \in \Re$. Then, the gcm and lcm points of set $X$ are computed excluding the minimum and maximum data values. These two values are the start and the end of the ecdf, respectively, and are considered to be both gcm and lcm points simultaneously. Note that since ecdf F(x) is piecewise constant, both gcm and lcm functions are piecewise linear and are actually defined as a sequence of line segments between successive (ordered) points $g_i$ and $l_j$ respectively. For the gcm function, we denote the sequence of points as ($g_i$, $G_i$) with $g_i \in X$ and $G_i = F(g_i)$, while for the lcm function, we denote the sequence of points as ($l_j$, $L_j$) with $l_j \in X$ and $L_j = F(l_j)$.

In UU-test we first check whether the whole ecdf is uniform. In this case we have a Uniform distribution which is an extreme single mode case, as we discussed in the definition of unimodality. Thus, UU-test ends and returns a positive unimodal decision without specific unimodal intervals, since the whole ecdf is uniform.

Computing the gcm/lcm function (gcm/lcm points) of the ecdf is the next step. Gcm and lcm points correspond to the convex and concave parts, respectively. As it was mentioned above, either the gcm or lcm points may be missing (e.g. Figure 3.3 illustrates a Gamma distribution; for specific parameters it may consist of lcm points exclusively).

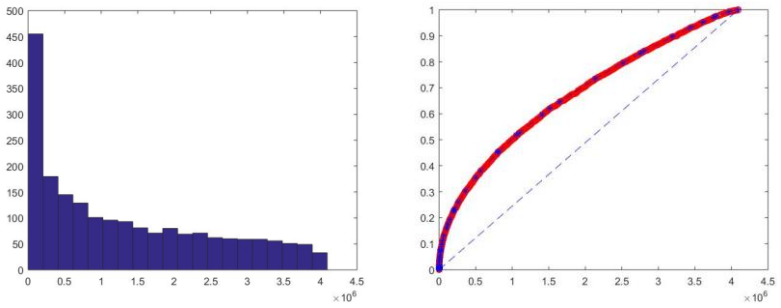


Figure 3.3: Histogram and ecdf of a Gamma distribution.

The existence of the gcm points that are followed by the lcm points is a powerful indication of unimodality, since the definition is respected; convex, linear and concave part, in this specific order. Unfortunately, there are cases where the gcm points precede the lcm points, but the function is multimodal, as illustrated in Figure 3.5 (a) and (b). The problem will be solved, if we utilize the second assumption of our method, the one of uniformity.

The main methodology of UU-test is summarized below:

Unimodality is respected, if we can draw line segments between:

i. successive gcm points
ii. a gcm point that is followed by an lcm point
iii. successive lcm points.

However, in order for the resulting piecewise linear cdf to be an acceptable approximation of the ecdf, the data points in each segment should follow the Uniform distribution.

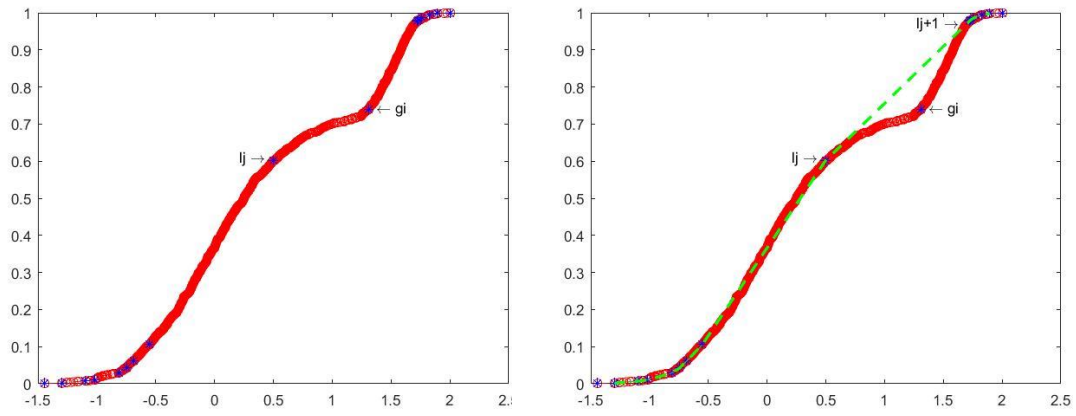
It should be mentioned that if the size of the original dataset is less than five data points, we will assume the function is unimodal without further testing. In addition, if any part of the dataset is that small, we assume this part of the function is unimodal, since gcm and lcm points of datasets with less than five data points cannot be computed. Furthermore, we define a maximum number of data points so that a segment is considered uniform with no use of KS test. For example, if our dataset's size is 2000 and we define the maximum number as the 1% of the dataset's size then we will not need to check uniformity with KS test for segments of less than 20 points. These segments are pretty small compared with the original dataset's size, so they will be directly considered as uniform with no use of KS test.

## 3.3 Case of a lcm point followed by a gcm point

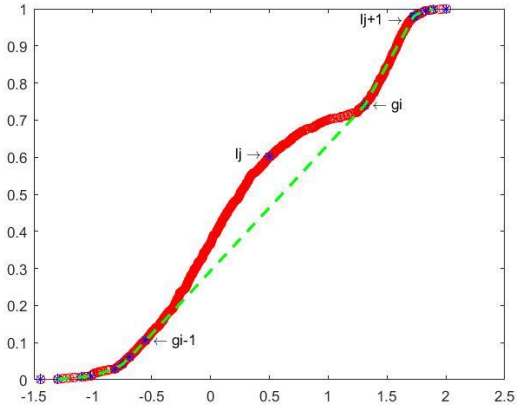
In the unfortunate case where we have an lcm point ($l_j$, $L_j$) followed by a gcm point ($g_i$, Gi), to ensure unimodality, we should try to avoid line segments from lcm to gcm points. This can be done considering two alternative approaches:

a) Case 1: Connect the lcm point ($l_j$, $L_j$) with the next lcm point ($l_{j+1}$, $L_{j+1}$) ignoring the in-between gcm points.
b) Case 2: Consider the line segment between the previous gcm point ($g_{i-1}$, $G_{i-1}$) and the gcm point ($g_i$, $G_i$) ignoring the in-between lcm points.

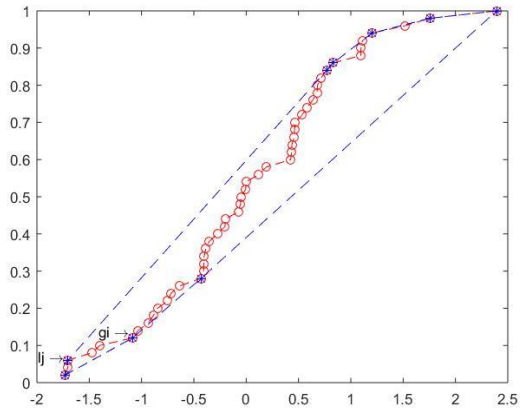
Then we check whether the data points in each line segment follow the Uniform distribution using KS test. In the case where a lcm point (or more than one) is followed by a gcm point (or more gcm points), this means that the concave part appears first, thus unimodality is not respected. However, there exist many cases where a lcm point precedes a gcm point and unimodality holds. In such cases, we do not draw line segments from $l_j$ to $g_i$, since we then accept our approximation will be first concave and after convex, i.e. multimodal. The idea is the following: if the points located before $g_i$ or after $l_j$ form uniform segments, our approximation will be possibly unimodal. We provide two examples of multimodal and unimodal cases where $l_j$ appears before $g_i$ (Figure 3.4).



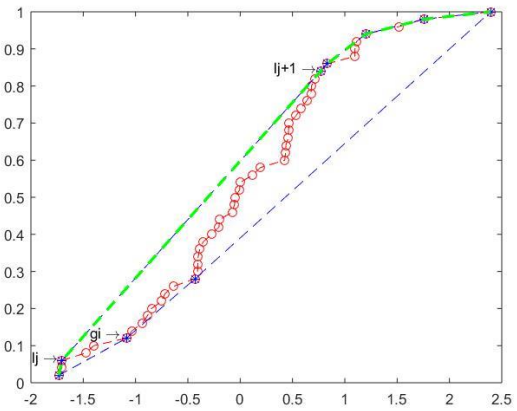

(a) (b)



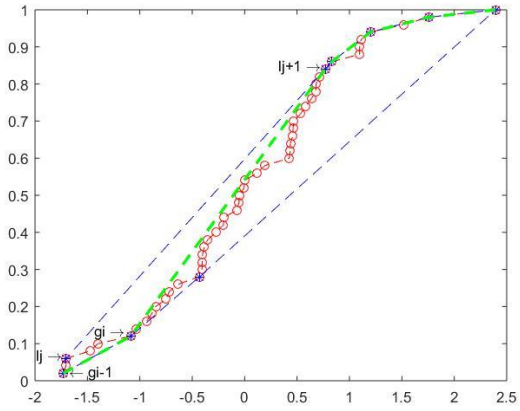

(c)




(d)




(e)




(f)

Figure 3.4: (a) and (d) Ecdfs of a multimodal and a unimodal distribution, respectively. A lcm point precedes a gcm point. (b) and (e) Trying to fix the above problem ignoring $g_i$ and connecting $l_j$ with the next lcm point $l_{j+1}$ (green line). (c) and (f) Trying to fix the above problem ignoring $l_j$ and connecting $g_i$ with the previous gcm point $g_{i-1}$ (green line).

In Figure 3.4, (a) and (d) illustrate the case of a lcm point $l_j$ followed by a gcm point $g_i$. In general, this implies that a concave part precedes a convex part, which is against the definition of unimodality. For this reason, in (b) and (e), we ignore the $g_i$ point and draw line segments according to case 1. Alternatively, we can ignore $l_j$ point and draw line segments according to case 2, as shown in (c) and (f). UU-test suggests uniformity in line segments, so the red and green line are pretty close. This implies that the points' distribution corresponding in each segment of the green line should be uniform i.e. it succeeds in KS test. Here we should mention that a function should satisfy at least one of cases 1 or 2 in order to be unimodal. Otherwise, UU-test decides multimodality.

In general, we need at least of one of the segments in cases 1 or 2 being uniform to proceed to next steps. When cases 1 or/and 2 is/are satisfied, we keep the uniform intervals. If there is a single lcm to gcm segment and both checks fail in KS test, UU-test will end and decide multimodality. There are cases where lcm to gcm segments appear more than once. UU-test requires at least one of these segments to satisfy cases 1 or 2 in order to build the desirable unimodal piecewise approximation (if exists). Though, lcm to gcm segments may appear more than once. We desire one of them to satisfy cases 1 or/and 2 and succeed in next checks (sections 3.4 and 3.5). Only in that case, UU-test will decide unimodality.

An important consideration is that in the first case, if ($l_j$, $L_j$) is the last lcm point and there is not a next one, we consider as ($l_{j+1}$, $L_{j+1}$) as the second maximum point of ecdf. Similarly, in the second case, if ($g_i$, $G_i$) is the first gcm point and there is not a previous one, we consider as ($g_{i-1}$, $G_{i-1}$) as the second minimum point of ecdf. Note that the minimum and maximum values of the data values have been excluded.

Algorithm 3.3.1 describes the steps of UU-test in the case where a lcm to gcm segment exists.

---

**Algorithm 3.3.1 Lcmtogcm ($g_i$, $l_j$, gcm_list, lcm_list)**

---

Input: $l_j$ (lcm point) precedes $g_i$ (gcm point), gcm_list= $\{g_1, ..., g_p\}$ (sorted list of gcm points), lcm_list= $\{l_1, ..., l_k\}$ (sorted list of lcm points)

Output: unimodality (1/0 ← unimodal/ multimodal), intervals: uniform intervals, S=possible uniform interval from gcm to lcm point

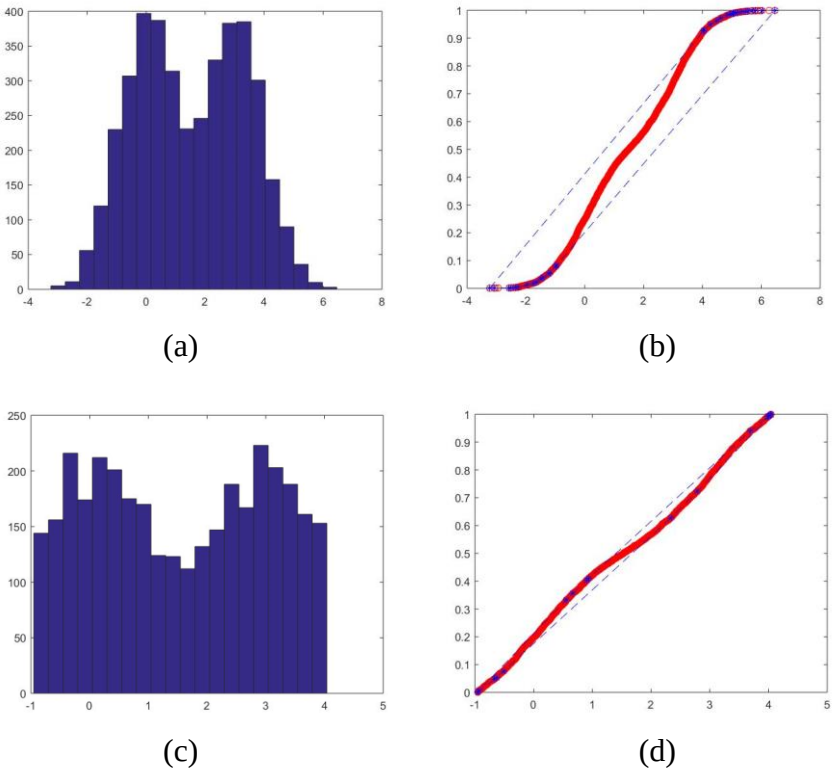
1. unimodality ← 1
2. {S, intervals} ← ∅
3. if KS ($[g_{i-1}, g_i]$) = true /* check for uniformity with KS test
4. intervals ← intervals ∪ $[g_{i-1}, g_i]$
5. S ← $[g_p, l_{j+1}]$ /* $g_p$ = the last gcm point after $g_i$ and before $l_{j+1}$
6. end if
7. if KS ($[l_j, l_{j+1}]$) = true
8. intervals ← intervals ∪ $[l_j, l_{j+1}]$
9. S ← S ∪ $[g_{i-1}, l_j]$
10. end if
11. if KS ($[g_{i-1}, g_i]$) = false and KS ($[l_j, l_{j+1}]$) = false
12. unimodality ← 0
13. end if
14. return {unimodality, intervals, S}

Of course, it does not necessarily exist a lcm to gcm segment in all the functions. For example, there are distribution functions consisting of exclusively the gcm part first and the lcm part second, such as the Gaussian, Student's t etc. The next steps are related with the checks on the gcm to lcm segment and the segments between successive gcm or lcm points.

## 3.4 The gcm to lcm segment

The next check is related with the gcm to lcm segment. Obviously, when both gcm and lcm points appear in an ecdf we expect to exist a gcm to lcm segment. One condition of unimodality is this segment being linear; in our method this is equivalent to uniformity tested by KS test. Whether lcm to gcm segments exist or not, the gcm to lcm segment should be uniform in order for the whole distribution to be unimodal. In unimodal distributions, this segment consists of the greatest part of the dataset. This is reasonable, since the gcm to lcm segment corresponds to the peak of the histogram; the region with the highest density. In case the gcm to lcm segment is not uniform, we call UU-test for this segment recursively. In other words, UU-test runs once more, with data points in the gcm to lcm segment as input, and tries to build the desirable approximation. If it deals with the above task, the input will be

unimodal and piecewise linear. UU-test will return the unimodal-uniform intervals and will continue with the rest checks. However, if UU-test does not decide unimodality for the gcm to lcm segment, it will definitely decide multimodality for the whole dataset.



(a) (b)

(c) (d)

Figure 3.5: (a) and (b) Histogram and ecdf of a bimodal distribution. (c) and (d) Histogram and ecdf of the middle segment from the gcm to lcm point.

In Figure 3.5, we provide an example of an ecdf consisting of a convex part, a gcm to lcm segment and a concave part. We need to check the middle segment's uniformity, so we call recursively the UU-test for this segment. In (c) and (d), we provide the results of UU-test. As we see, there are clearly two peaks at the histogram and a lcm to gcm segment which fails to meet cases 1 or 2, as mentioned in section 3.3. Finally, UU-test decides multimodality, since it cannot build a piecewise linear approximation for the gcm to lcm segment.

Figure 3.6 illustrates an example of a unimodal distribution. Although, the middle part in (b) is not uniform; this may be caused by the sensitivity of KS test in the center of distributions. The UU-test runs again with input being the data points of this segment, and finally decides

unimodality. Note that, UU-test does not end here, since there is still one last check (section 3.5).

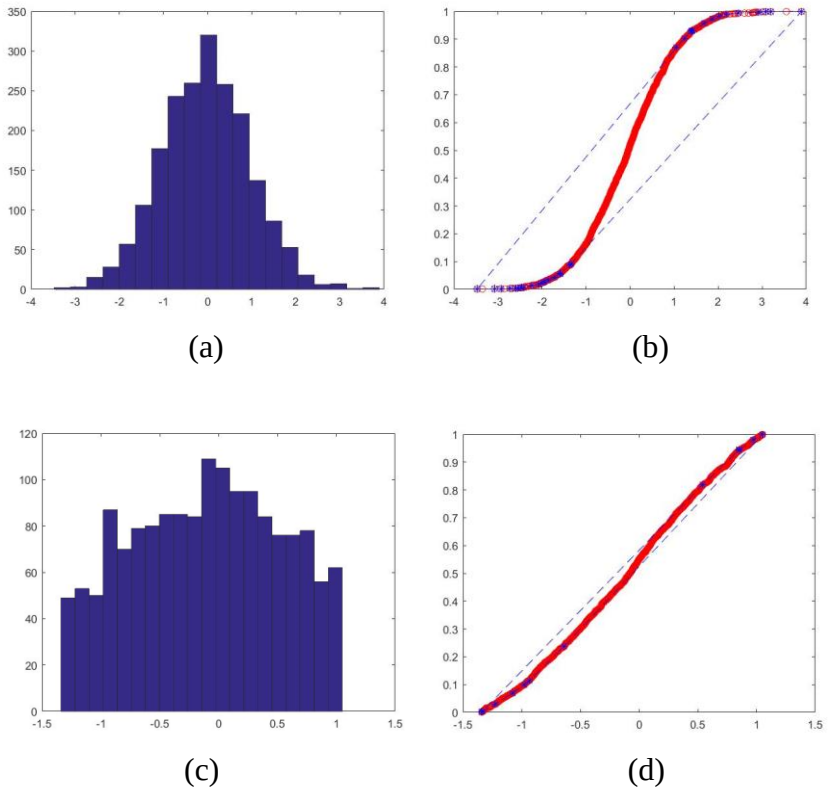



Figure 3.6: (a) and (b) Histogram and ecdf of a unimodal distribution. (c) and (d) Histogram and ecdf of the middle segment from the gcm to lcm point.

## 3.5 Checking uniformity between successive gcm or lcm points

For ensuring unimodality, UU-test should check uniformity in the segments defined by successive gcm or lcm points. This is the last step of UU-test and we will proceed only if the previous steps are successful. This implies that at least one of lcm to gcm segments (if exists) satisfy cases 1 or 2 and the gcm to lcm segment is uniform (or unimodal and piecewise uniform). Suppose, we are interested in the convex part (gcm points). First, we check the whole convex segment from the second minimum point to the last gcm point. If it is not uniform according to KS test, the check is carried out in successive segments of gcm points. If a segment is not uniform, we will test backwards and forwards the previous and next segment of gcm points, respectively, including the original non-uniform segment. The pseudocodes of the functions Backcheck and Frontcheck are given below (Algorithms 3.5.1, 3.5.2).

We try to be more clear using an example. Let $\{x_1, ..., x_n\}$ be the dataset in an ascending order and $\{g_1, ..., g_k\}$ be the list of successive gcm points, with $k << n$. Unimodality is respected, since our check is made in a convex part (the first condition of unimodality's definition). Thus, we need to ensure uniformity in each segment of gcm points. First, we check the segment between $x_1$ and $g_k$ points (the whole convex part). If it is uniform, we proceed to the lcm points, in a similar way. However, the above segment is usually too large and complex to be directly uniform. Cutting in smaller segments is needed to achieve the piecewise uniformity. So, we check the successive segments $(x_1, g_1), (g_1, g_2), (g_2, g_3), ..., (g_{k-1}, g_k)$. Let's assume that the $(g_i, g_{i+1})$ segment, with $i < k$, is not uniform.

Backward checks: The function Backcheck is utilized to check uniformity of successive segments in a backward direction. Starting with greater gcm values, we end up in lower values. Note that the gcm points are sorted in an ascending order. Having the non-uniform $(g_i, g_{i+1})$ we check the successive segments $(g_{i-1}, g_{i+1}), (g_{i-2}, g_{i+1}), (g_{i-3}, g_{i+1}), ..., (x_1, g_{i+1})$, one by one, and if one of them is uniform we continue with the segment after $g_{i+1}$. For example, if $(g_{i-2}, g_{i+1})$ is found uniform in Backcheck, we will save this interval and proceed to $(g_{i+1}, g_{i+2})$ segment. Segments before $g_{i-2}$ have already been checked and proved uniform. In case where all backward checks fail, we will make forward checks.

Forward checks: Similarly to the backward checks, the function Frontcheck is utilized to check uniformity of successive segments in a forward direction. Gcm values come out from the lower to the greater. Having the non-uniform $(g_i, g_{i+1})$ segment, we check the successive segments $(g_i, g_{i+2}), (g_i, g_{i+3}), (g_i, g_{i+4}), ..., (g_i, g_k)$, one by one. If one of them is uniform, the checks will be made for the segments of gcm points that follow the interval's right endpoint. For example, if $(g_i, g_{i+3})$ is found uniform in Frontcheck, we will keep this interval and proceed to checking $(g_{i+3}, g_{i+4})$ segment. In case where all forward checks fail, since backward checks have also failed, then UU-test decides multimodality.

The procedure is analogous in the case of the segments between successive lcm points. Initially, we check the segment from the first lcm point to the second maximum point of the ecdf function and continue with the rest checks as described for the convex part.

**Algorithm 3.5.1 Backcheck (gcm_list, not_unif)**

Input: gcm_list = $\{g_1, ..., g_p\}$ , not_unif = [$g_i$, $g_{i+1}$] a non-uniform segment.

Output: unimodality (1/0 ← unimodal/ multimodal), interval: the not_unif segment inside a uniform segment, in case of unimodality.

1. unimodality ← 0
2. interval ← ∅
3. maxval ← $g_{i+1}$
4. for j=i-1 to 1
5. minval ← $g_j$
6. if KS ([minval, maxval]) = true /* check for uniformity with KS test
7. unimodality ← 1
8. interval ← [minval, maxval]
9. break
10. end if
11. end for
12. return {unimodality, interval}

**Algorithm 3.5.2 Frontcheck (gcm_list, not_unif)**

Input: gcm_list = $\{g_1, ..., g_p\}$ , not_unif = [$g_i$, $g_{i+1}$] a non-uniform segment.

Output: unimodality (1/0 ← unimodal/ multimodal), interval: the not_unif segment inside a uniform segment, in case of unimodality.

1. unimodality ← 0
2. interval ← ∅
3. minval ← $g_i$
4. for j=i+2 to p
5. maxval ← $g_j$
6. if KS ([minval, maxval]) = true /* check for uniformity with KS test
7. unimodality ← 1
8. interval ← [minval, maxval]

9. break
10. end if
11. end for
12. return {unimodality, interval}

To sum up, UU-test fails i.e. a function is multimodal in either of three main cases: (i) lcm to gcm segments appear and none of them satisfies the cases 1 or 2, (ii) the gcm to lcm is not uniform (even if UU-test calls itself recursively for this segment) and (iii) the uniformity is not respected in successive gcm or lcm points.

In Figure 3.7, we present a unimodal ecdf as well as a multimodal ecdf with a non-uniform segment between successive lcm points.

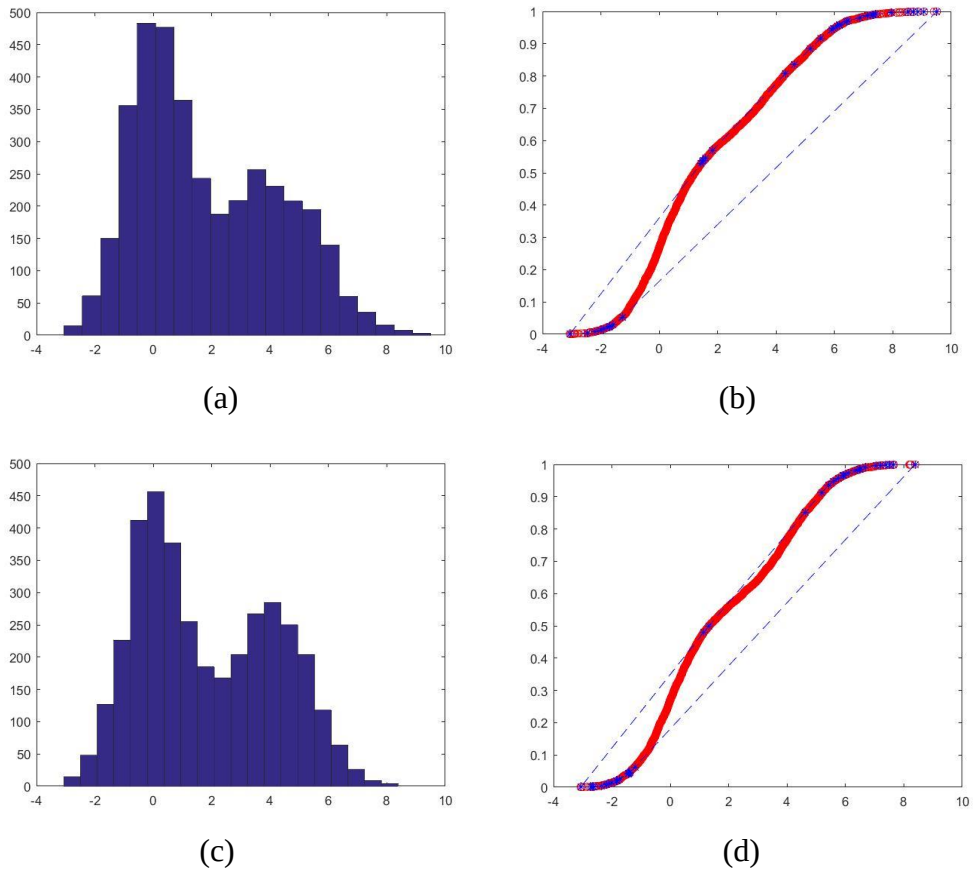
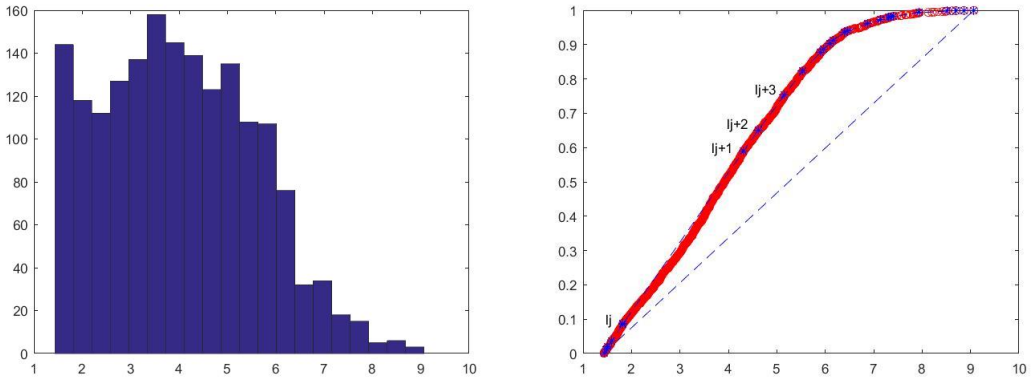


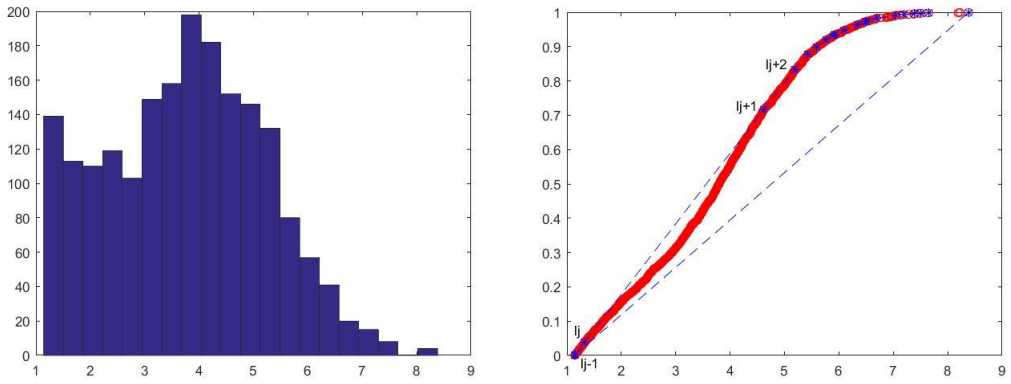

Figure 3.7: (a) and (b) Histogram and ecdf of a unimodal distribution. (c) and (d) Histogram and ecdf of a multimodal distribution.

In Figure 3.8, we present the lcm segment of the ecdfs of Figure 3.7. In (c) and (d) of Figure 3.8, the interval $(l_j, l_{j+1})$ is not uniform, according to KS test. In that case, we utilize the functions Backcheck and Frontcheck to check uniformity of previous and next segments. First, UU-test calls Backcheck which it uses backward checks, but the KS test fails in all of the checks. After, since lcm points after $l_{j+1}$ do exist, UU-test calls Frontcheck and KS test checks the $(l_j, l_{j+2})$ segment and decides non-uniformity. It continues checking the next segments, but it fails, too. Thus, UU-test decides multimodality, since neither Backcheck nor Frontcheck can fix the problem of non-uniformity.

Figure 3.8 ((a) and (b)) presents the histogram and ecdf of the lcm segments of a unimodal distribution (Figure 3.7 (a) and (b)). Similarly to (c) and (d), the interval $(l_j, l_{j+1})$ is not uniform. At first, Backcheck is called by UU-test and fails, and then Frontcheck is called to fix the non-uniformity of the segment. The KS test decides that the $(l_j, l_{j+2})$ segment is not uniform and continues with the $(l_j, l_{j+3})$ segment, where it finally decides uniformity and Frontcheck ends and returns the above uniform segment. The total intervals will consist of the successive segments created by successive gcm points $(x_1, g_1), (g_1, g_2), ..., (g_{i-1}, g_i)$, the gcm to lcm segment $(g_i, l_{j-3})$, the uniform segments before $l_j$: $(l_{j-3}, l_{j-2}), (l_{j-2}, l_{j-1}), (l_{j-1}, l_j)$ and the rest segments $(l_j, l_{j+3}), (l_{j+3}, l_{j+4}), ..., (l_m, x_n)$ including the interval returned by Frontcheck. In the specific example, we supposed that: i is the number of gcm points, m is the number of lcm points and there are three lcm points preceding $l_j$.




(a) (b)




(c) (d)

Figure 3.8: Histogram and ecdf of the lcm part of a unimodal ((a) and (b)) and multimodal ((c) and (d)) distribution. The segment is not uniform, so we add previous and next segments and check uniformity with KS test using the functions Backcheck and Frontcheck, respectively (Algorithm 3.5.1 and Algorithm 3.5.2).

Algorithm 3.5.3 provides the pseudocode of UU-test.

**Algorithm 3.5.3 UU_test (X)**

Input: X = $\{x_1, ..., x_n\}$ the dataset.

Output: unimodality (1/0 ← unimodality/ multimodality), intervals: the intervals in which the ecdf is unimodal and uniform (in case of unimodality).

1. xs ← sorted X in ascending order
2. [F, f] ← ecdf(xs)
3. unimodality ← 0  intervals ← ∅
4. ($g_i$, $G_i$) ← gcm points of (F, f)  /* $G_i$ = F($g_i$), gcm_list= set of $g_i$ points
5. ($l_j$, $L_j$) ← lcm points of (F, x)  /* $L_j$ = F($l_j$), lcm_list= set of $l_j$ points
6. if KS (xs) = true  /* the whole dataset is uniform
7. unimodality ← 1
8. else
9. find possible lcm_to_gcm_segments
10. {unimodality, intervals, m, S1, S2} ← *Lcmtogcm* ($g_i$, $l_j$, gcm_list, lcm_list)
11. if unimodality = 0 (for all possible lcm_to_gcm_segments)

```
12.             return {unimodality, intervals}      /* End
13.       end if
14.       if gcm_to_lcm_segment not uniform
15.             {unimodality, intervals} ← UU_test (gcm_to_lcm_segment)
16.             if unimodality = 0
17.                   return {unimodality, intervals}    /* End
18.             end if
19.       end if
20.       for all in-between segments in gcm and lcm list
21.             if KS (segment) = false
22.                   call Backcheck (gcm_list, segment)         /* similarly for lcm_list
23.             end if
24.             if unimodality = 0       /* if backward checks fail, UU-test will call the Frontcheck
25.                   call Frontcheck (gcm_list, segment)        /* similarly for lcm_list
26.             end if
27.       end for
28. end if
29. return {unimodality, intervals}
```

## 3.6 Examples of UU-test

As we previously mentioned, in case of unimodality, UU-test provides the intervals where the function is unimodal and uniform. We can connect all the existing points inside the intervals and that line (green line in Figure 3.9) will be the piecewise approximation.

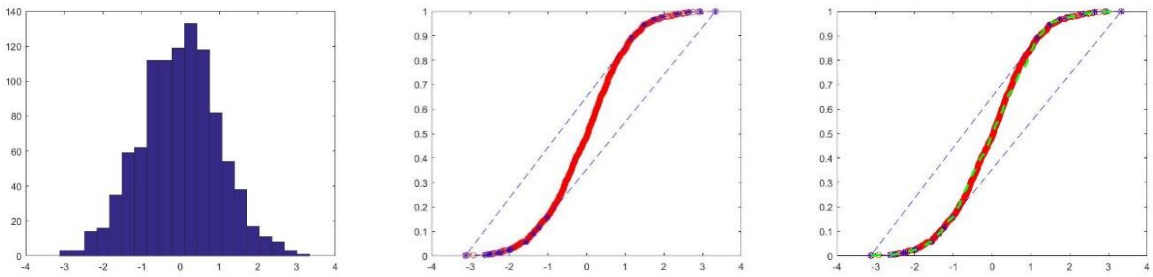
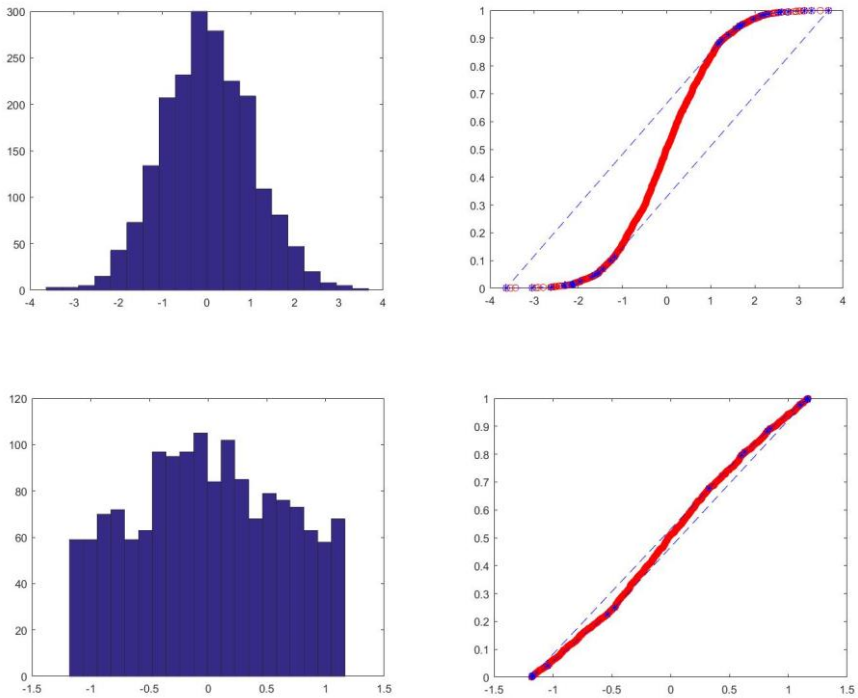


Figure 3.9: Construction of unimodal piecewise linear approximation (green line).

The lower and upper bounds of the intervals correspond to gcm and lcm points of the function. The simplest case is when there are exclusively the gcm part, the gcm to lcm segment and the lcm part. Note that we construct intervals only if our function is unimodal and the uniform tests between the segments are successful. Thus, the intervals consist of segments which are uniformly tested by the KS test. The gcm to lcm segment may consist of a single interval (if the whole segment is uniform) or divided in smaller uniform segments (if UU-test has been called recursively).

For example: a case where the whole gcm to lcm segment is not uniform - but is uniform in segments - is provided in Figure 3.10. UU-test is called again, decides unimodality and the intervals will consist of the segments returned by the recursive calls of UU-test. The rest plots show the successful approximation of the gcm to lcm part and of the original ecdf (green lines).



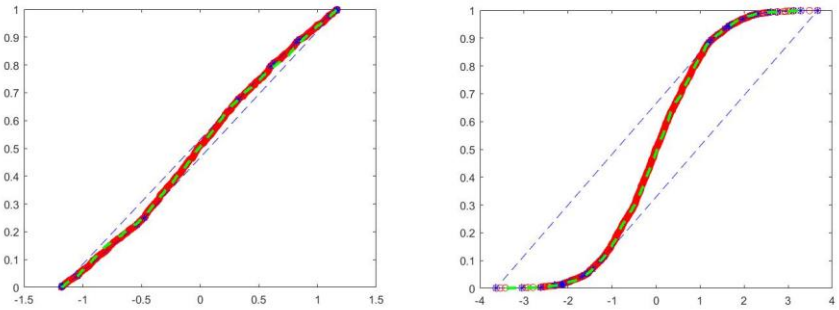


Figure 3.10: Unimodal function with the whole gcm to lcm segment being uniform.

As regards the gcm/lcm part, in section 3.5, we extensively explained the process of creating the uniform intervals (if they exist). In summary, we check the successive gcm/lcm segments with KS test, and if it decides uniformity, these segments will be our final intervals. If KS test fails in a ″tricky″ segment, we will connect it with the next or previous segment, and utilize the functions Frontcheck or Backcheck, respectively.

## 3.7 Experiments – Decision on Distribution Unimodality

The first part of the experimental evaluation we made is related to the decision on distribution unimodality and the results are shown in Table 3.1. We tested a variety of distributions using data samples from of different sizes and the names of distributions, the parameters and the sizes are shown in the first two columns of Table 3.1. The number of datasets for each distribution tested was 50, while both unimodal and multimodal distributions were examined. More specifically, the unimodal distributions tested were: Gaussian, Student's, Gamma, Exponential, Cauchy, Triangular, Asymmetric Triangular, two close Gaussians, two Truncated Gaussians, mixed Student's t & Uniform and mixed Uniform & Gaussian. The multimodal distributions tested were a mixture of two and three Gaussians.

We compared the results of UU-test with Hartigans' dip test using the same significance level ($\alpha$=0.01). The results of UU-test seem very encouraging, since it achieves high rates of agreement with the dip test. We can see that the total success rates of UU-test are greater than 94%, while in symmetric distributions (e.g. Gaussian, Student's t, Triangular, etc.) UU-test presents no failure. Furthermore, in multimodal distributions with clearly distinct modes (e.g 2 or 3 Gaussians), UU-test does not fail, too. In some asymmetric unimodal distributions, UU-

test fails to accept unimodality; this may be caused by the sensitivity of KS test in the center of distributions, as it was mentioned in Chapter 2. In general, the results of the two tests are pretty close and UU-test, in most cases, provides correct decisions.

| Distributions | Parameters | Dip test | UU-test | Agreement of two tests |
|---|---|---|---|---|
| Gaussian ($\mu$, $\sigma^2$)<br>$\mu$: mean value<br>$\sigma^2$: variance<br>size=2000 | $\mu=0$, $\sigma=1$ | 100% | 100% | 100% |
| Student's t ($\nu$)<br>$\nu$: degrees of freedom<br>size=2000 | $\nu=4$ | 100% | 100% | 100% |
| Gamma(k, $\theta$)<br>k: shape, $\theta$: scale<br>size=2000 | k=1, $\theta=2$ | 100% | 100% | 100% |
| Exponential ($\lambda$)<br>$\lambda$: rate<br>size=2000 | $\lambda=3$ | 100% | 100% | 100% |
| Cauchy($\nu$)<br>$\nu$: degrees of freedom<br>size=2000 | $\nu=1$ | 100% | 100% | 100% |
| Triangular (L, U, m)<br>L: Lower limit<br>U: Upper limit<br>m: mode<br>size=3700 | L=-1, U=1, m=0 | 100% | 100% | 100% |
| Asymmetric Triangular<br>size=6500 | L=-2, U=3, m=0 | 100% | 96% | 96% |
| Two Gaussians<br>size1=2000<br>size2=2000 | $\mu_1=0$, $\sigma_1=1$<br>$\mu_2=4$, $\sigma_2=1$ | 100% | 100% | 100% |

| | | | | |
|---|---|---|---|---|
| Two Gaussians<br>size1=2000<br>size2=1000 | $\mu_1=0$, $\sigma_1=1$<br>$\mu_2=4$, $\sigma_2=1$ | 100% | 100% | 100% |
| Two Gaussians<br>size1=1000<br>size2=1000 | $\mu_1=0$, $\sigma_1=1$<br>$\mu_2=4$, $\sigma_2=2$ | 100% | 100% | 100% |
| Two Truncated Gaussians<br>size1=1000<br>size2=1000 | $\mu_1=0$, $\sigma_1=1$ (right tail)<br>$\mu_2=0$, $\sigma_2=3$ (left tail) | 100% | 94% | 94% |
| Three Gaussians<br>size1=1000<br>size2=1000<br>size3=1000 | $\mu_1=0$, $\mu_2=4$, $\mu_3=8$<br>$\sigma_1 = \sigma_2 = \sigma_3 = 1$ | 100% | 100% | 100% |
| Three Gaussians<br>size1=1000<br>size2=1000<br>size3=2000 | $\mu_1=0$, $\mu_2=4$, $\mu_3=7$<br>$\sigma_1 = \sigma_2 = \sigma_3 = 1$ | 100% | 100% | 100% |
| Student's t ($\nu$) & Uniform (a, b)<br>a: minimum value<br>b: maximum value<br>size=15000 | $\nu=10$ &<br>a=0 b=10 | 100% | 96% | 96% |
| Uniform (a, b) & Gaussian($\mu$, $\sigma^2$)<br>size=16000 | a=-10 b=5 &<br>$\mu=3$, $\sigma=1$ | 100% | 96% | 96% |

Table 3.1: Comparative results of the two tests (dip and UU) on distribution unimodality.

# CHAPTER 4.

# MODELING UNIMODAL DATA



## 4.1 Statistical Data Modeling

A statistical model is a mathematical model that embodies a set of statistical assumptions concerning the generation of some sample data and similar data from a larger population. A statistical model represents, often in considerably idealized form, the data-generating process. The assumptions embodied by a statistical model describe a set of probability distributions, some of which are assumed to adequately approximate the distribution from which a dataset is sampled. In simple terms, statistical modeling is a simplified, mathematically-formalized way to approximate reality (i.e. what generates our data) and optionally to make predictions from this approximation. The statistical model is the mathematical equation that is used. It should summarize the data as closely as possible (be 'a good fit') but also be as simple as possible. We cannot measure a population, so the best we can do is generalize from a sample to a population using a representative summary, i.e. a statistical model. Fitting a model to data means choosing the statistical model that predicts values as close as possible to the ones observed in our population. We need to find the values for the parameters in the model that are most appropriate to predicting the data.

### 4.1.1 Gaussian Model

The Gaussian Model is a widely used statistical model which works well as a good fit of many sets of data. It is often the case that we don't know the parameters of the Gaussian distribution, but instead want to estimate them. That is, having a sample ($x_1$, …, $x_n$) from a Gaussian N($\mu$, $\sigma^2$) population we would like to learn the approximate values of parameters $\mu$ and $\sigma^2$. The standard approach to this problem is the maximum likelihood method, which requires maximization of the log-likelihood function:

$$\ln L(\mu,\sigma^2) = \sum_{i=1}^{n} \ln f(x_i \mid \mu,\sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Taking derivatives with respect to $\mu$ and $\sigma^2$ and solving the resulting system of first order conditions yields the maximum likelihood estimates:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Modeling our data with a Gaussian Model needs estimation of the parameters of the Gaussian distribution. Estimator $\hat{\mu}$ is called the sample mean, since it is the arithmetic mean of all observations. The estimator $\hat{\sigma}^2$ is called the sample variance, since it is the variance of the sample ($x_1$, …, $x_n$). In practice, another estimator is often used instead of the $\hat{\sigma}^2$. This other estimator is denoted $s^2$, and is also called the sample variance, which represents a certain ambiguity in terminology; its square root s is called the sample standard deviation. The estimator $s^2$ differs from $\hat{\sigma}^2$ by having (n − 1) instead of n in the denominator.

$$s^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

The Gaussian Model fits very well in datasets coming from Gaussian distributions or distributions pretty close to the Gaussian, such as Student's t. A significant disadvantage, though, is its failure in asymmetric distributions.

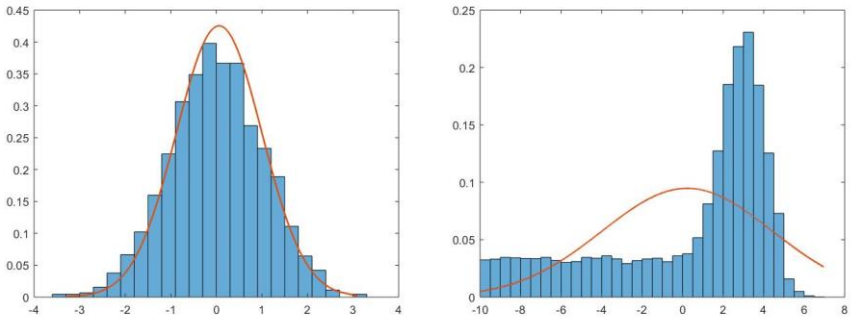


Figure 4.1: A Gaussian model fits in a Gaussian dataset and a mixture of a Uniform & a Gaussian.

Figure 4.1 illustrates two examples of a Gaussian Model fitting in samples of two distributions. The first is a sample from a standard Gaussian distribution ($\mu=0$, $\sigma^2=1$). The second is a sample from a Uniform ($a=-10$, $b=5$) and a Gaussian ($\mu=3$, $\sigma^2=1$). In the first case the parameters of the model were estimated and resulted in: $\bar{x}=0.06$ and $s^2=0.9378$. These values are really close to the original and undoubtfully the figure shows a very accurate fit. The similar process was made for the second case and the estimated parameters are: $\bar{x}=0.279$ and $s^2=17.457$. This is a case of asymmetric distribution and the Gaussian Model cannot efficiently fit.

Another model which can model our data sample is the Uniform Model. Similar to the Gaussian Model, the parameters of the Uniform Model (a and b) are estimated through the maximum likelihood method. In Figure 4.2, we work with similar distributions as in Figure 4.1. In both cases, the Uniform Model does not fit well.

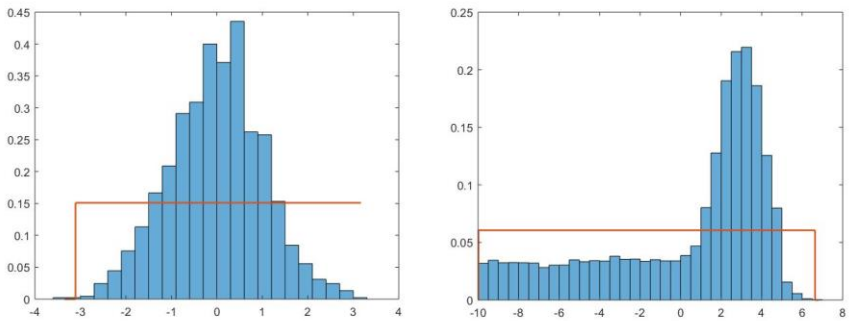


Figure 4.2: A Uniform model fits in a Gaussian dataset and a mixture of a Uniform & a Gaussian.

## 4.2 Mixture distribution

In Probability and Statistics, a mixture distribution is the probability distribution of a random variable that is derived from a collection of other hidden random variables as follows: first, a random variable is selected by chance from the collection according to given probabilities of selection, and then the value of the selected random variable is realized. The underlying random variables may be random real numbers, or they may be random vectors (each having the same dimension), in which case the mixture distribution is a multivariate distribution.

In cases where each of the underlying random variables is continuous, the outcome variable will also be continuous, and its probability density function is sometimes referred to as a mixture density. The cdf (and the pdf) can be expressed as a convex combination (i.e. a weighted sum, with non-negative weights that sum to 1) of other distribution functions and density functions. The individual distributions that are combined to form the mixture distribution are called the mixture components, and the probabilities (or weights) associated with each component are called the mixture weights.

Given a finite set of pdfs $p_1(x), ..., p_K(x)$, or corresponding cdfs $P_1(x), ..., P_K(x)$ and weights $w_1, ..., w_K$ such that $w_i \geq 0$ and $\sum_{i=1}^{K} w_i = 1$, the mixture distribution can be represented by writing either the density ( $f$ ), or the distribution function ( $F$ ), as a sum (which in both cases is a convex combination):

$$F(x) = \sum_{i=1}^{K} w_i P_i(x), \qquad f(x) = \sum_{i=1}^{K} w_i p_i(x)$$

Mixture distributions arise in many contexts in the literature and arise naturally where a statistical population contains two or more subpopulations. It is frequently the case that data is not explained by a single underlying distribution. Typically, this is because there are multiple phenomena occurring in the data set, each with their own underlying distribution. If we want to try to recover the underlying distributions, we need to have a model which has multiple components. An example could be sensor readings where the majority of the time a sensor shows no signal, but sometimes it detects some phenomena. Modeling both phenomena as a single distribution would be inaccurate because the readings would come from two distinct phenomena.

A common type of mixture model, called Gaussian Mixture Model, is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.

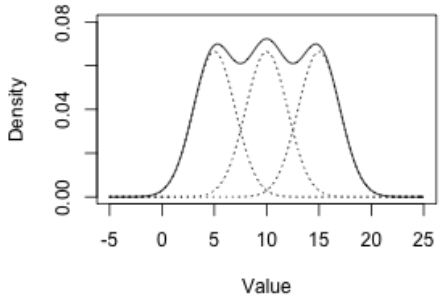



Figure 4.3: A Gaussian Mixture of three Gaussian distributions.

## 4.3 Uniform Mixture Model

In our approach, we use the Uniform Mixture Model in which each component is the Uniform distribution. In unimodal cases, it successfully models unimodal data and it is noteworthy that such a model is directly provided by the UU-test. Since each segment provided by UU-test comes from a uniform population, the uniform mixture is being a sensible choice.

Thus, in case the UU-test succeeds, it also provides a generative model of the 1-d dataset. More specifically, if K segments $[a_i, b_i]$ $(i = 1, ..., K)$ are provided with $N_i$ data points in each segment, then the following uniform mixture can be used as a pdf model of the data distribution:

$$p(x) = \sum_{i=1}^{K} \pi_i U(x; a_i, b_i) I(x \in [a_i, b_i]),\ \pi_i = N_i / N$$

where U is the uniform distribution pdf.

The cdf F(x) corresponding to the above pdf p(x) is:

$$F(x) = \sum_{j=1}^{i-1} \pi_j + \pi_i \frac{x - a_i}{b_i - a_i}, \quad a_i \le x \le b_i$$

and it is expected to be 'close' to the ecdf.

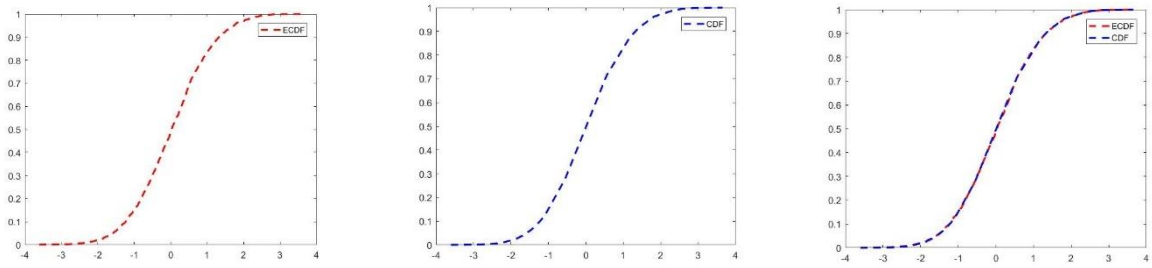



Figure 4.4: Gaussian distribution ($\mu$=0, $\sigma^2$=1), size=2000. The red line is the ecdf and the blue line is the cdf F(x) corresponding to p(x). The two curves are almost identical, as we expected.

## 4.4 Sampling from our Uniform Mixture Model

When UU-test succeeds, we can generate data points with the following steps:

(1) we compute the vector of probability weights $\pi = (\pi_1, \pi_2, ..., \pi_k)$, $k$: the number of segments, $\pi_i = N_i / N$, $N_i$: the number of data points in each segment and $N$: the total number of data points.

(2) we use a Multinomial distribution with parameters $n$ and $\pi$, $n$: sample size (number of trials). If $n$=1, it returns a random value in $[1, k]$ corresponding to which segment we should choose to pick a value.

(3) According to the value returned, we generate a uniformly distributed random number in the corresponding segment.

This procedure can be repeated for $n$ > 1.

Figure 4.5 illustrates a Gaussian distribution ((a) and (b)) of size 2000 and a dataset sampled from the Gaussian of the same size ((c) and (d)). The procedure followed in generating the data points is described above.

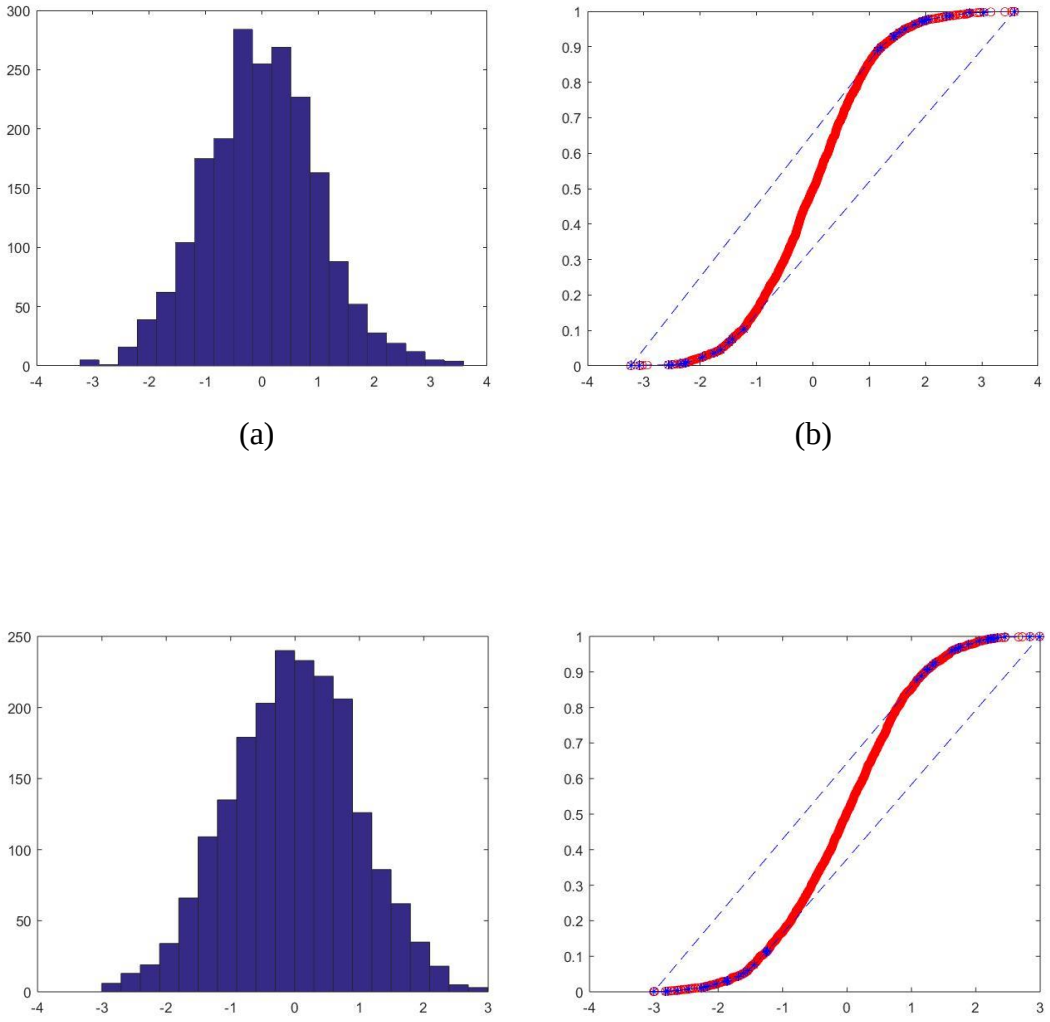



Figure 4.5: (a) and (b) Gaussian distribution ($\mu$=0, $\sigma^2$=1), size=2000. (c) and (d) A sample consisting of 2000 points, sampled from the original Gaussian, according to the above three steps. It is notable that the two histograms and ecdf figures are almost identical.

## 4.5 Experiments – Modeling Unimodal Data

We conducted a series of experiments using artificial data in order to test and compare the efficiency of the three models: Gaussian, Uniform and UU-model, in modeling unimodal data. In order to measure the quality of the fitted models, we used two criteria. First, we used the max log-likelihood on test sets.

We used a sample without replacement of the 75% of the dataset size as a test set. The rest 25% was used as a training set. Then, we computed the pdf of each model, evaluated at the values in test set and also computed the log-likelihood of the pdf. The maximum value among the three models provides the best fit.

Moreover, we used the two-sample Kolmogorov-Smirnov test as a suitable criterion to evaluate the three models. The two-sample KS test is a nonparametric hypothesis test that evaluates the difference between the cdfs of the distributions of the two data samples, through the maximum absolute difference between the cdfs. This test decides if two datasets come from the same continuous distribution. We used a data sample (test set) from the ground truth distribution and compared it (using the two-sample KS test) with a dataset sampled from each of the three fitted models. The smaller the distance provided by the KS test, the better the fitted model.

Our experiments consist of a great variety of unimodal distributions functions, such as: a Gaussian, Student's t, Gamma, Triangular, Asymmetric Triangular and two Gaussians closed to each other (unimodal though). Moreover, we tested mixed distributions i.e. mixtures of observations generated from the above distributions. Mixtures of Uniform and Student's t, Uniform and Gaussian are unimodal examples of them (Figure 4.6). Table 4.1 gives the ground-truth distributions, their parameters and the sizes of the training and test sets.

| Distributions | Parameters | Size of Training set | Size of Test set |
|---|---|---|---|
| Gaussian ($\mu$, $\sigma^2$)<br>$\mu$: mean value $\sigma^2$: variance | $\mu$=0, $\sigma$ =1 | 650 | 2000 |
| Student's t ($\nu$)<br>$\nu$: degrees of freedom | $\nu$=4 | 650 | 2000 |
| Gamma(k, $\theta$)<br>k: shape, $\theta$: scale | k=1, $\theta$=2 | 650 | 2000 |
| Triangular (L, U, m)<br>L: Lower limit<br>U: Upper limit<br>m: mode | L = -1, U = 1, m = 0 | 1250 | 3700 |
| Triangular (L, U, m) | L = -1, U = 1, m = 0 | 12500 | 37000 |
| Asymmetric Triangular | L = -2, U = 3, m = 0 | 2150 | 6500 |
| Two Gaussians | $\mu_1$=0, $\sigma_1$=1<br>$\mu_2$=2.4, $\sigma_2$=1 | 5850 | 17500 |
| Student's t ($\nu$) &<br>Uniform (a, b)<br>a: minimum value<br>b: maximum value | $\nu$=10 &<br>a=0 b=10 | 5000 | 15000 |
| Uniform (a, b) &<br>Gaussian($\mu$, $\sigma^2$) | a=-10 b=5 &<br>$\mu$=3, $\sigma$=1 | 5300 | 16000 |

Table 4.1: Names and parameters of distributions functions used, and sizes of training and test sets.

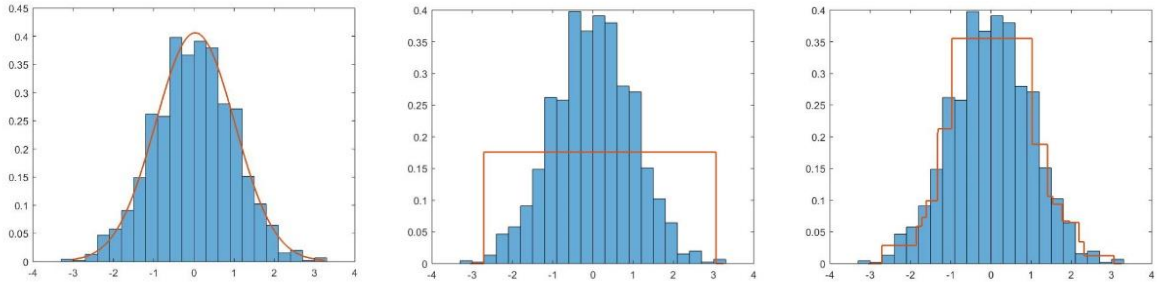
In Table 4.2, we provide the values of the test set log-likelihood for the three models (Gaussian, Uniform and UU). The maximum value (bold number) gives the model with the best fit. In Table 4.3, we compare the three models using the criterion of two-sample KS test. The minimum value (shortest distance) indicates the model with the best fit.

| Distribution | Gaussian model | Uniform model | UU model |
|---|---|---|---|
| Gaussian | **-13338** | -17694 | -14027 |
| Student's t | -16331 | -26681 | **-16149** |
| Gamma | -38283 | -37044 | **-34591** |
| Triangular | **-1873.6** | -2574.2 | -1986.7 |
| Triangular | -19451 | -25697 | **-18899** |
| Asymmetric Triangular | -93852 | -104910 | **-89273** |
| Two Gaussians | -32877 | -42027 | **-32483** |
| Student's t & Uniform | -40245 | -43284 | **-36288** |
| Uniform & Gaussian | -45959 | -45828 | **-39153** |

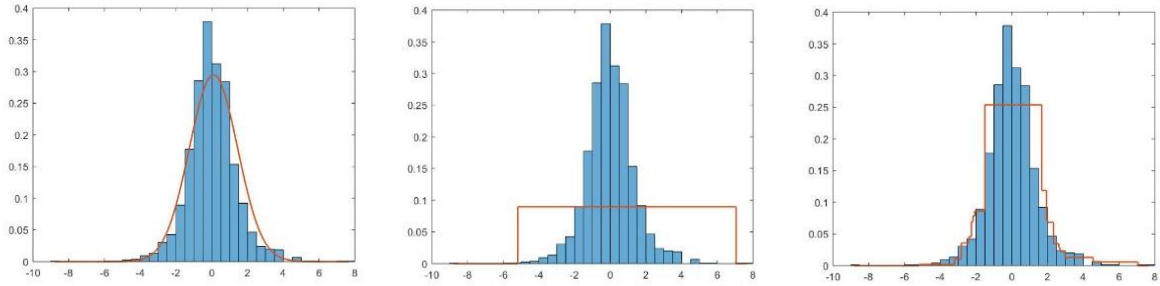
Table 4.2: Experiments: comparing the three models with criterion the max log-likelihood in the test set.

| Distribution | Gaussian model | Uniform model | UU model |
|---|---|---|---|
| Gaussian | **0.0133** | 0.1945 | 0.0232 |
| Student's t | 0.0366 | 0.3451 | **0.0186** |
| Gamma | 0.1064 | 0.2350 | **0.0164** |
| Triangular | 0.0179 | 0.1234 | **0.0097** |
| Triangular | 0.0180 | 0.1260 | **0.0062** |
| Asymmetric Triangular | 0.0929 | 0.2072 | **0.0055** |
| Two Gaussians | 0.0365 | 0.2336 | **0.0055** |
| Student's t & Uniform | 0.1488 | 0.2720 | **0.0065** |
| Uniform & Gaussian | 0.2041 | 0.2703 | **0.0048** |

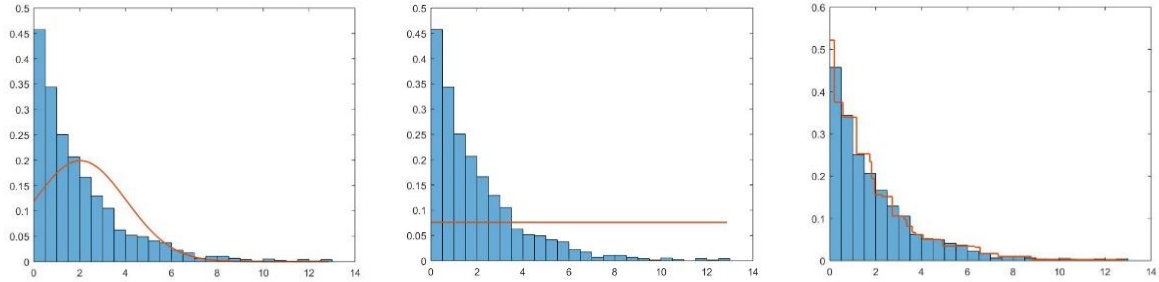
Table 4.3: Experiments: comparing the three models with criterion the two-sample KS test.



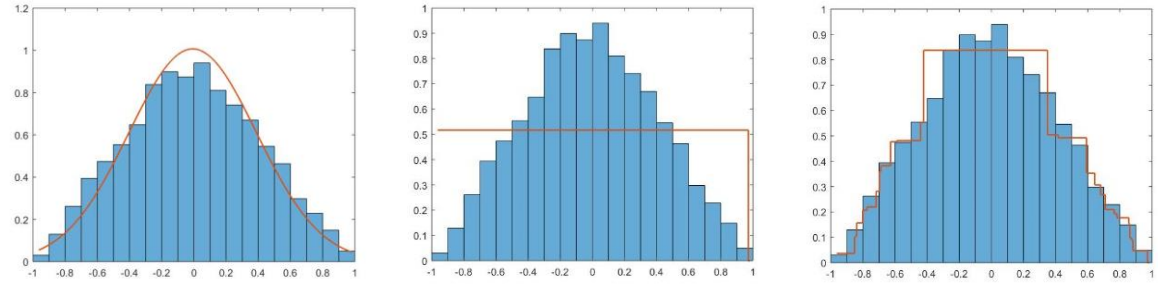
(a)



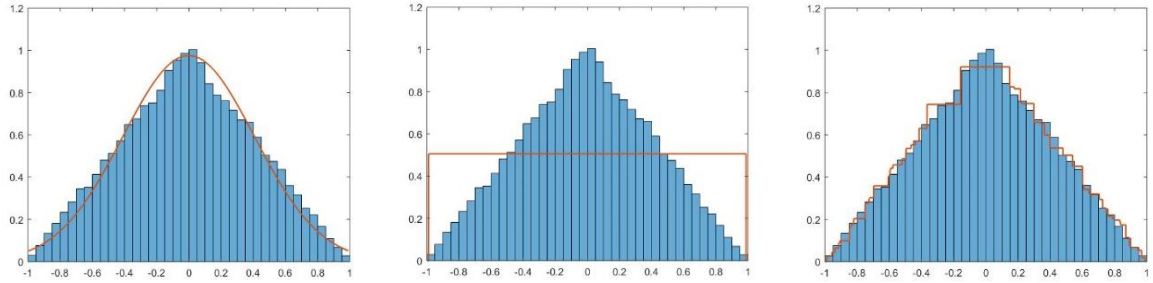
(b)



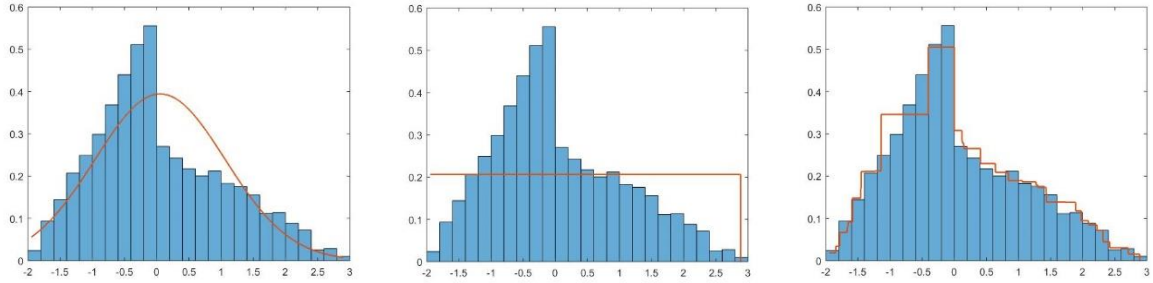
(c)



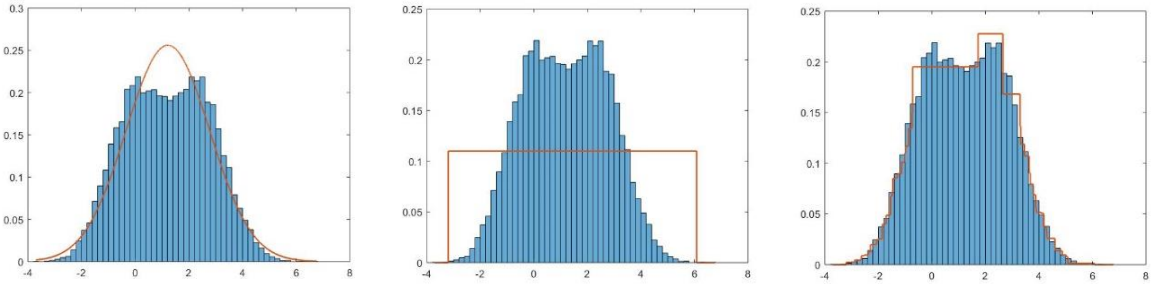
(d)



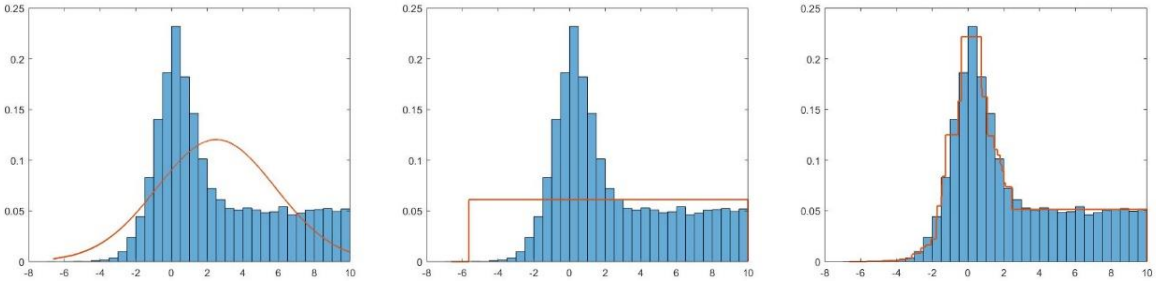
(e)



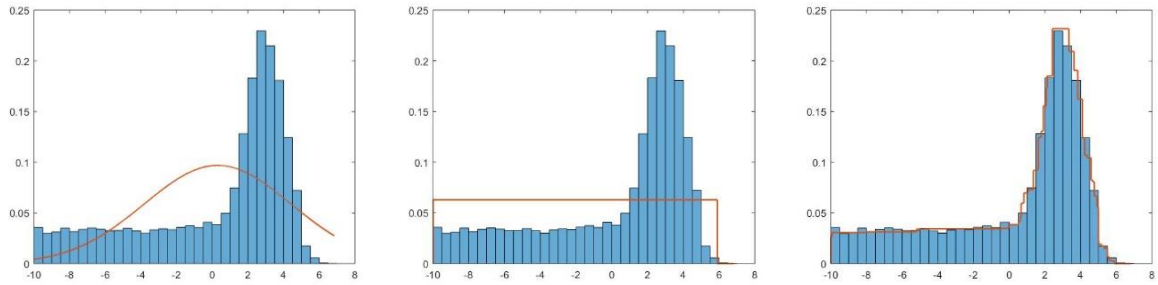
(f)



(g)



(h)



(i)

Figure 4.6: Examples of Gaussian, Uniform and UU-model fit in a variety of unimodal distributions. (a) Gaussian, (b) Student's t, (c) Gamma, (d) Triangular, (e) Triangular, (f) Asymmetric Triangular, (g) Two Gaussians, (h) Student's t & Uniform, (i) Uniform & Gaussian.

The experiments we conducted show that UU-model fits better in almost every case of function. As it is expected, the Gaussian model fits more efficiently in a Gaussian and a Triangular than our model, according to max log-likelihood criterion. It is noteworthy that UU-model gives better results in the Student's t case than the Gaussian model. Though, Gaussian's results are pretty close. The difference between Gaussian and UU-model is clear in asymmetric distributions, such as the mixed or the asymmetric Triangular. An important observation should be mentioned; the Uniform fits better than Gaussian in Gamma and mixed Uniform & Gaussian cases, according to max log-likelihood criterion. Thus, the Gaussian model extremely fails in asymmetric data while UU-model provides the best fit.

# CHAPTER 5.

## CONCLUSION AND FUTURE WORK



## 5.1 Conclusion

In this thesis, we studied the property of unimodality for continuous distributions. We suggested a new non-parametric method (UU-test), to decide whether a continuous 1-d dataset is unimodal or not. UU-test includes the computation of gcm/lcm points and checks for uniformity with Kolmogorov-Smirnov test, to provide the appropriate decision. Furthermore, our method directly provides a generative mixture model of the data points, in unimodal cases. We used the Uniform Mixture Model, which is constructed using the unimodal and uniform intervals returned by UU-test.

In the experimental evaluation we conducted, the first aim was to compare UU-test's results with the widely known Hartigans' dip test. In the most of the cases we examined, the decisions of unimodality/multimodality of the two tests were equivalent. Then, we tested how well our UU-model fits unimodal data. We compared UU-model with the Gaussian and Uniform model. In order to measure the quality of the fitted models, we used two criteria. First, we used the log-likelihood on test sets; the maximum value among the three models provided the best fit model. Secondly, the two-sample Kolmogorov-Smirnov test was used as a criterion to evaluate the three models. This test decides if two datasets come from the same continuous distribution. We used a data sample from the ground truth distribution and tested it with a data sample from each of the three compared models. We used various examples of ground truth distributions, such as the Gaussian, Student's t, Gamma, Triangular, Asymmetric

Triangular, two Gaussians closed to each other (unimodal though), Uniform-Student's t, Uniform-Gaussian. In Gaussian distributions, our UU-model does not perform as well as the Gaussian model, while in the rest distributions it fits better than the other two models.

## 5.2 Future Work

Given the encouraging results obtained from UU-test, there are several research directions to be followed. At first it would be interesting to study how the UU-test could be used in multidimensional datasets. One possible solution would be to apply UU-test in each dimension or in appropriate projections (e.g. PCA projections).

Another research direction is to consider a method to break a multimodal dataset in unimodal segments. This is a significant topic, since it is a kind of clustering. For example, a multimodal dataset may consist of two peaks (bimodal), so the result would be two discrete unimodal datasets (two clusters). Breaking the multimodal segments determined by the UU-test is a promising solution to this problem.

It would be also noteworthy to know in advance whether the result of UU-test will change, if we add new data values in the original dataset. What number of data values will turn a unimodal dataset into multimodal or what kind of data values should be added to fix multimodality?

Finally, another research direction concerns the implementation of UU-test. The employed KS test seems to be less accurate in the center of distributions. If another analogous test is used, the results might be more accurate.

# REFERENCES


[1] J. H. Wolfe, Pattern clustering by multivariate mixture analysis, Multivariate Behavioural Res. 5, pp. 329-350, 1970.

[2] L. Engelman and J. A. Hartigan, Percentage points of a test for clusters, J. Amer. Statist. Assoc, 64, pp. 1647-1648, 1969.

[3] B. W. Silverman, Using kernel density estimates to investigate multimodality, J. Roy. Statist. Soc. B 43, pp. 97-99, 1981.

[4] J.A. Hartigan and P. M. Hartigan, The dip test of unimodality, The Annals of Statistics, 13(1), pp. 70-84, 1985.

[5] A. Siffer, P.- A. Fouque, A. Termier, C. Largouët, Are your data gathered? The Folding Test of Unimodality, KDD 2018 - 24th ACM SIGKDD International Conference on Knowledge Discovery, pp.2210-2218, 2018.

[6] M. C. Minnotte, Nonparametric Testing of the Existence of Modes, The Annals of Statistics, 25(4), pp. 1646-1660, 1997.

[7] M. C. Minnotte and D. W. Scott, The Mode Tree: A Tool for Visualization of Nonparametric Density Features, Journal of Computational and Graphical Statistics, 2, pp. 51-68, 1993.

[8] L. H. Fraser, J. Pither, A. Jentsch, et al., Plant ecology. Worldwide evidence of a unimodal relationship between productivity and plant species richness, Science, 349(6245), pp. 302-305, 2015.

[9] K. Johnsson, M. Linderoth, M. Fontes, What Is a “Unimodal” Cell Population? Using Statistical Tests as Criteria for Unimodality in Automated Gating and Quality Control, Cytometry Part A Journal of Quantitative Cell Science, 91A, pp. 908-916, 2017.

[10] A. Adolfsson, M. Ackerman, N. C. Brownstein, To Cluster, or Not to Cluster: How to Answer the Question, KDD‘17, Canada, 2017

[11] A. Kalogeratos and A. Likas, Dip-means: an incremental clustering method for estimating the number of clusters, In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 25, Curran Associates, Inc., pp. 2393–2401, 2012.

[12] B. Schelling and C. Plant, DipTransformation: Enhancing the Structure of a Dataset and thereby improving Clustering, IEEE International Conference on Data Mining 2018 (ICDM), 2018.

[13] A. Krause and V. Liebscher, Multimodal projection pursuit using the dip statistic, Preprint-Reihe Mathematik, 13, 2005.

# SHORT CV

Paraskevi Chasani was born in Ioannina, Greece in 1993. In 2011, she enrolled in the department of Mathematics of University of Ioannina and received the BSc degree in 2015. In continuation to her studies, she enrolled as a MSc Student in the Computer Science & Engineering Department of University of Ioannina. After fulfilling her responsibilities as a graduate student, she presented her thesis in February 2019 in order to complete the Master's Degree. Her main interests are in the area of data analysis, statistics and machine learning.