Testing for multimodality

Stoepker, Ivo V.

Award date:
2016

Testing for multimodality

Author: Ivo Stoepker

Supervisors: dr. R.M. Pires da Silva Castro
prof.dr. E.R. van den Heuvel

TU/e, June 27, 2016

Abstract

In this thesis we explore the different tests for multimodality, devise new tests and touch upon the issue of explaining multimodality by using covariates of the data. Two new tests have been devised. The first test, the LOG test, uses logconcavity as a proxy for multimodality and is calibrated by bootstrapping. This test can be of great value since the MLE of logconcave densities is well defined. We show this test is reasonably well-calibrated and powerful but naturally has issues when used to test for a non-logconcave unimodal density. We propose a censoring approach and show it does not fix this issue. The test is nevertheless valuable when testing for logconcavity in its own right. The second test, the string test, is based on the dip test. It fits a unimodal CDF to the data which aims to minimize the Kolmogorov-Smirnov statistic. The test is also calibrated by bootstrapping and is a valuable substitute for the original dip test which has considerable less power. The string test is also appealing because of the simplicity of the methodology. We assess the calibration and power of both the string and LOG test with numerical results and compare the string test with an existing test. The results of the comparison must be approached with caution as the results of the competing test where not always completely specified in the original paper, but indicate the string test is more conservative and has less power. The difference in conservativeness is small and might be desirable in some cases as the competing test is anti-conservative under some distributions. Finally, we present a regression setting which can be used when testing for multimodality in the presence of covariates. We prove an attractive approach, where we seek parameters which will cause the error distribution to be unimodal, to be ill-defined and offer an alternative procedure where we test the residuals for multimodality. We also provide an iterative algorithm where the parameter values are updated after estimating a unimodal density of the error distribution.

Contents

List of abbreviations

AD statistic	Anderson-Darling statistic.
CDF	Cumulative distribution function.
CvM statistic	Cramér-Von Mises statistic.
ECDF	Empirical cumulative distribution function.
gcm	Greatest convex minorant.
i.i.d.	Independently and identically distributed.
KS statistic	Kolmogorov-Smirnov statistic.
lcm	Least concave majorant.
MLE	Maximum likelihood estimator.
PDF	Probability density function.

Chapter 1

Introduction

In many situations of practical interest, we would like to characterize the density function of a given set of data. When we have some idea of how the data came to be, we can assume a model and estimate the corresponding parameters. The characterization of the density function is then assumed by the statistician. However, in many situations, assumptions cannot be made and we have to resort to non-parametric statistics.

When examining the histogram or kernel density estimate of a dataset, we might observe multiple modes. A mode can informally be defined[1] as a local maximum in the density function. These modes are important features of the density and can give us a lot of information about the process which produced the data. However, we must first evaluate if the observed modes are true and not caused by, for instance, poor choice of binning or bandwidth.

We are therefore interested to test the density of the data for multimodality: the presence of multiple modes. To find out whether the underlying density f of the data is unimodal or multimodal, many tests have been devised. Parametric tests exists which assume mixture models of certain distributions. However, these tests can be inaccurate when the true density is not of this type of mixture. We are therefore most interested in the non-parametric tests which decide between the following two competing hypotheses:

$$H_0 : f \text{ is unimodal} \qquad \text{vs.} \qquad H_1 : f \text{ is multimodal.} \tag{1.1}$$

When we have evidence that multimodality is present, it would be desirable to explain the modes by using covariates of the data. This could be done by using a regression model. However, estimating both the density of the data and the regression parameters can be difficult. Furthermore, the non-parametric maximum-likelihood estimator (MLE) of a unimodal density is not well-defined. An approach which makes use of the non-parametric MLE must restrict the class of unimodal densities to, for instance, the class of logconcave densities.

In this thesis, our goal is to devise well-calibrated and powerful tests for the competing hypotheses in (1.1). We also aim to develop a test that can be of value when testing for multimodality in the presence of covariates.

[1] We provide formal definitions in the next chapter.

We accomplish this goal in three stages. First, we examine existing tests of multimodality and provide an overview of the tests and their general approach. We also examine the most significant tests in great detail and show the issues their approaches face. Secondly, we devise two test for multimodality. The first test will use logconcavity as a proxy for multimodality, which will be of value when testing for multimodality in the presence of covariates. The second test is inspired the dip test [11]. Both tests will use a bootstrap approach for calibration. Finally, we introduce the problem of explaining multimodality with covariates and propose an approach to the problem.

Chapter 2

Existing tests

We will first identify the existing tests for multimodality. A number of different tests and approaches have been developed. Both parametric and non-parametric tests can be found in literature.

The parametric tests aim to fit a mixture distribution to the data. The disadvantage of this approach is, of course, that the mixture distribution must be specified beforehand. If the assumption is incorrect, the results of the test will be unreliable.

Because of this disadvantage, non-parametric tests have been developed. These tests do not require a specification of a mixture model beforehand. However, these types of tests do have a problem concerning calibration. Since the null-hypothesis is composite, calibration can be challenging. To avoid anti-conservativeness of the test, the test could to be calibrated using an unfavourable unimodal null distribution, such as the uniform distribution. This results in conservativeness on other unimodal distributions and likely loss of power. However, the calibration issue can be somewhat resolved by using Monte-Carlo approaches.

2.1 Definitions

We have informally introduced the concept of multimodality. Before continuing, it is necessary to now state the formal definitions of unimodality, multimodality and modes which will be used throughout the text. We will present the definitions of the univariate case only.

Definition A probability density function (PDF) f is *unimodal* if for some value m, the function f is monotonically increasing for $x \leq m$ and monotonically decreasing for $x \geq m$. If no such value for m exists, f is *multimodal*.

Definition A cumulative distribution function (CDF) F is *unimodal* if for some value m, the function F is convex on the interval $(-\infty, m]$ and concave on the interval $[m, \infty)$. If no such value for m exists, F is *multimodal*.

Note that m is not necessarily unique. For example, m is not unique for the uniform distribution.

Definition The modes of a (multimodal) PDF f are the values of x such that $f(x)$ attains a (local) maximum. Such a maximum may be attained at a single point or inside a closed interval. If the maxima is attained at a single point, the corresponding value of x is referred to as a *mode*. If the local maximum is attained at a closed interval, the interval is referred to as a *modal interval*.

In some literature, the notion of strict unimodality or strong unimodality is mentioned. We provide a definition here for completeness:

Definition A PDF f is *strictly unimodal* if for some value m, the function f is strictly monotonically increasing for $x \leq m$ and strictly monotonically decreasing for $x \geq m$.

We will always use unimodality as opposed to strong unimodality.

Finally, we define a density with a shoulder, which we will use later in the text.

Definition A PDF f with one *shoulder* is defined as density which has one closed interval $[x_i, x_j]$ such that for all $x \in [x_i, x_j]$, the derivative $f'(x) = 0$, but the points in the interval $[x_i, x_j]$ do not attain a global or local maximum. Thus, this region, which we call the shoulder, is not a mode of the density f.

2.2 Overview of tests

Test for multimodality can generally be split into two groups. There exist tests for unimodality versus bimodality (two modes) and more general tests for unimodality versus multimodality (two or more modes).

We now present a list of some well-known tests for multimodality. The reader is referred to the original papers for a detailed description of the procedure.

The following test for unimodality versus bimodality:

1. Haldane's test [6], which uses tangents of the density to assess multimodality. The test is based on the distribution of second central differences.
2. Larkin's test [14], based the F test; it results in a small F ratio if the sample is unimodal and a large F ratio if it is bimodal. First, the variance of the sample when considered unimodal is computed. Then, the sample is divided into two parts repeatedly and the mean of the variance of both parts is computed for every division of the sample. The lowest mean of two variances is assumed to be the variance of the sample if it were bimodal. The ratio can now be computed between the variance of the sample regarded as unimodal and the variance when it is regarded bimodal. The author notes that the test can be modified to handle more than two modes, but does not do so in the original paper.
3. Tokeshi [24] developed a test in the context of distribution dynamics. He notes that his test is not statistically foolproof: "The method was used as

a convenient means of defining and recognizing modality patterns in the context of distribution dynamics, rather than as a statistically neutral, foolproof test procedure".

4. Holzmann & Vollmer propose a parametric test [10]. The likelihood ratio is employed when fitting two-component mixture densities.

The following test for unimodality versus multimodality:

1. Silverman's bandwidth test [22], which uses the bandwidth of the kernel density estimate to test for multimodality. If the bandwidth of the kernel density estimate is very large, then the estimate is smoothed and becomes unimodal. If we need a large amount of smoothing to find an unimodal estimate, this indicates multimodality. A thorough description of this test can be found in Section 2.3.1.

2. Dip test [11], which calculates the dip statistic. This statistic measures the maximum difference between the empirical cumulative distribution function (ECDF) and the CDF that minimizes this maximum difference. A thorough description of this test can be found in Section 2.3.2.

3. Excess mass test [17][19], which measures how much "excess" probability mass is present in the empirical density in comparison with multiples of the uniform density. They argue that a mode is present where an excess of mass is concentrated. The approach uses the excess mass functional $E(\lambda)$, which is the amount of excess mass compared to λ-multiple of the uniform density (which is generalized by using the λ-multiple of Lebesque measure). If there are multiple modes in the density, the excess mass will be split over different "clusters", which are used to test for multimodality.

4. Mode existence test [16], which examines potential modes on a local level. Usually, multimodality tests are formulated to assess whether the true density is unimodal or multimodal as a whole, and test for the null hypothesis "the true density is unimodal" versus the alternative "the true density is multimodal". However, this test examines each mode separately. The null hypothesis is therefore formulated in the paper as "the mode at x_0 of our density estimate is an artefact of the sample" versus the alternative "the mode at x_0 of our density estimate is a true feature of the sample".

5. RUNT test [9], a test which is based on single linkage clusters. The test first follows the regular single linkage clustering procedure. After the procedure we are left with a hierarchical tree of clusters, where each cluster divides into a number of subclusters. Now, for every cluster C, we can find the subcluster with the smallest amount of datapoints - this can be thought of as the "runt" of the cluster - and denote it by $n(C)$. The RUNT statistic is the maximum runt size over all clusters: $\max_C\{n(C)\}$. The author argues that, if the true density is multimodal, then (asymptotically) at one point during the single linkage procedure, the data about one mode is merged with most of the data about the other mode. Therefore, the runt of this merged cluster will be very large. If there is only one mode, the smallest subcluster should contain only a few points. Therefore, a large value of the RUNT statistic indicates multimodality.

6. MAP test [18], which tests for multimodality using minimum spanning

trees of the data. An additional constraint to the minimum spanning tree is added; the distances between vertices must be non-increasing on every path to the root node, starting from any vertex in the graph. The resulting graph is called the minimum ascending path spanning tree (MAPST). The definition is extended to accommodate multiple root nodes. The minimum ascending path statistic (MAP statistic) is then defined as the sum of the distances between the vertices. Multimodality is assessed by comparing the MAP statistic of a MAPST with a single root, corresponding to a unimodal density, to the statistic when the MAPST contains multiple roots, corresponding to a multimodal density.

2.3 Description of main tests

As we showed in Section 2.2, many tests for multimodality exists. However, there are three tests that seem to be the most popular and have been improved after their initial publication. These tests are the Silverman's bandwidth test, the excess mass test and the dip test. We will now describe these tests in detail and find the advantages and disadvantages of their approach.

Authors who have improved either the dip test or the excess mass test have noted that the two tests are equivalent when applied in a one-dimensional case [2]. Because of this, we will examine one of these tests. Since we will use some underlying theory of the dip test ourselves, we choose to examine the dip test in more detail here.

2.3.1 Description of Silverman's test

We now take a closer look at Silverman's test. Silverman initially published his test in 1981 [22].

Suppose we are given data $X = (x_i, \ldots, x_n)$ where we can assume all x_i's i.i.d. Silverman's test now tests the following two hypotheses:

$$H_0 : f \text{ has } m \text{ modes} \qquad \text{vs.} \qquad H_1 : f \text{ has more than } m \text{ modes.}$$

The test makes use of kernel density estimation. The kernel density estimator is given by:

$$\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right). \tag{2.1}$$

In (2.1), $K(\cdot)$ is the kernel. Silverman's test makes use of a Gaussian kernel. This choice has both theoretical and computational advantages. Importantly, the Gaussian kernel ensures that the number of modes of the kernel density estimate is non-increasing as h increases, which will be of importance in the test. Silverman's test is sometimes referred to as the "bandwidth test", since it uses the bandwidth of the Gaussian kernels to assess multimodality.

The test works as follows:

1. The kernel density estimator for the data X as a function of the bandwidth h is computed. Denote this function by $\hat{f}(h)$.
2. A critical bandwidth h_{crit} is computed as follows:

$$h_{crit} = \inf\{h \mid \hat{f}(h) \text{ has at most } m \text{ modes}\}.$$

   So the test finds h such that a decrease in h results in more than m modes. This h_{crit} is used as a test statistic; a large value of h_{crit} means that we need a large amount of smoothing such that the modes in the kernel density estimate disappear, which we can interpret as a evidence against unimodality.
3. B samples from the density $\hat{f}(h_{crit})$ are taken.
4. For every sample, the number of modes is computed. Denote the number of modes of sample j by m_j.
5. The p-value is computed as follows:

$$p = \frac{\#\{j : m_j \leq m\}}{B}.$$

The test has a few drawbacks. One of the main concerns is the sensitivity to tail behavior. This issue is described briefly in [8], but we illustrate the issue with an example here as well. Consider the following mixture of normal distributions:

$$X = \frac{1}{5}\mathcal{N}(-4, 2) + \frac{1}{5}\mathcal{N}(-\frac{3}{2}, \frac{6}{5}) + \frac{2}{5}\mathcal{N}(0, \frac{1}{2}) + \frac{1}{5}\mathcal{N}(1, \frac{1}{10}). \tag{2.2}$$

The probability density function f of X is plotted in Figure 2.1. Note that f has two modes; one for $x \approx -0.02$ and one for $x \approx 0.92$. Now, if we sample from f and construct the kernel density estimator, the resulting density estimator typically looks like the density in Figure 2.2. As can be seen, the two original modes have merged into one large mode, and an extra mode has been added to the left tail.

This issue is caused by the following. Suppose there is a uniform interval in the density of width $2d$. The kernel density estimator with too small a bandwidth might introduce several spurious modes in this interval. To annihilate these spurious modes, we need to use a bandwidth of width about d. However, if there are true modes in another interval of the density that are at most $2d$ apart, they will also be annihilated. When the true density contains uniform intervals which have a length greater than the distance between a pair of its true modes, Silverman's test will fail to identify them correctly. Since distributions with heavy tails contain these (near) uniform intervals, the test will be adversely affected in these cases. In essence, the bandwidth in the Silverman's test is determined globally, while this global bandwidth might adversely change the distribution's shape locally. Note that this drawback inspired the Mode existence test [16].

Another drawback of the test is that it does not account for the difference in probability mass associated with each mode. Small modes in the tails of the

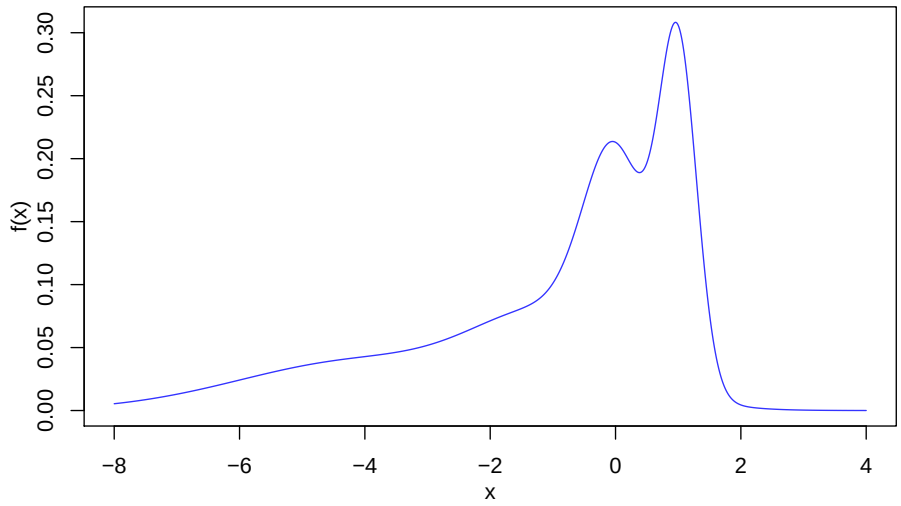

Figure 2.1: PDF of the mixture distribution X in (2.2).

kernel density might not be correct but are as significant when counting the number of modes as the larger modes present.

Furthermore, authors who aim to improve the original test note that the test is also conservative [2].

2.3.2 Description of the dip test

We now take a closer look at Hartigan's dip test. The test was originally proposed in 1985 [11].

Suppose that we have $X = (x_1, \ldots, x_n)$, an i.i.d. sample. The dip test tests for multimodality of the sample by computing the dip statistic of the ECDF of the data.

The dip statistic of a CDF F is defined as follows. Let $\mathscr{U}$ be the class of unimodal CDFs. Now the dip statistic D is given by:

$$D(F) = \min_{G \in \mathscr{U}} \sup_x \mid F(x) - G(x) \mid .$$

To compute the dip statistic, the following theorem is used:

Theorem Let F be an arbitrary distribution function. Then $D(F) = d$ only if there exists a nondecreasing function G such that, for some $x_L \leq x_U$,

1. G is the greatest convex minorant (gcm) of $F + d$ in $(-\infty, x_L)$.
2. G has constant maximum slope in (x_L, x_U).
3. G is the least concave majorant (lcm) of $F - d$ in (x_U, ∞).
4. $d = \sup_{x \notin (x_L, x_U)} |F(x) - G(x)| \geq \sup_{x \in (x_L, x_U)} |F(x) - G(x)|$.

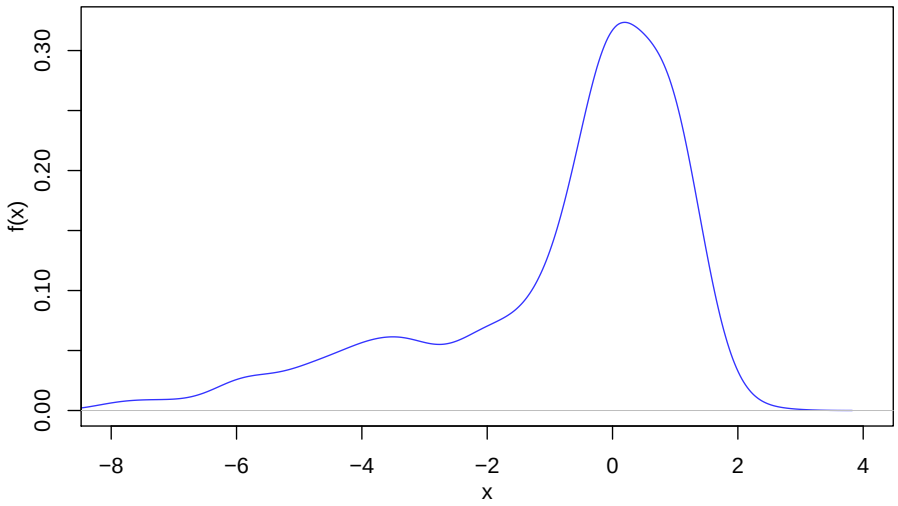

Figure 2.2: Kernel density function (with bandwidth chosen according to Gaussian approximation, or "Silverman's rule of thumb" [21]) of a sample of size 1000 from the mixture distribution X in (2.2).

The dip can now be calculated as follows. Suppose the data X is ordered and let x_i and x_j be two datapoints where $x_i \leq x_j$. Let F be the ECDF of the data. Now we first the gcm of F on the interval $(-\infty, x_i)$ and the lcm on (x_j, ∞). Let d_{ij} be the maximum absolute distance of F to these two curves.

Now we find the minimum value of d_{ij}, provided that the line segment $[x_i, F(x_i)+\frac{1}{2}d_{ij}]$ to $[x_j, F(x_j) - \frac{1}{2}d_{ij}]$ lies between $F(x) - \frac{1}{2}d_{ij}$ and $F(x) + \frac{1}{2}d_{ij}$.

We are now able to use the theorem. This is not directly evident, but we can show how we can use it in the following way: for the minimum value of d_{ij} just found, we now have a gcm of F on $(-\infty, x_i)$ and an lcm in (x_j, ∞). We can now translate the gcm curve by $\frac{1}{2}d_{ij}$ to the positive y-direction and the lcm $\frac{1}{2}d_{ij}$ to the negative y-direction. The result is a gcm of $F + \frac{1}{2}d_{ij}$ and lcm of $F - \frac{1}{2}d_{ij}$. The maximum distance between F and the shifted gcm/lcm is now $\frac{1}{2}d_{ij}$. Furthermore, the maximum distance of the line between $[x_i, F(x_i) + \frac{1}{2}d_{ij}]$ to $[x_j, F(x_j) - \frac{1}{2}d_{ij}]$ and F is also $\frac{1}{2}d_{ij}$ since the line was between the curves $F(x) - \frac{1}{2}d_{ij}$ and $F(x) + \frac{1}{2}d_{ij}$.

Using the theorem, with $d = \frac{1}{2}d_{ij}$, we find that $d_{ij} = 2D$.

As noted by [11], it seems as if all possible begin- and endpoints x_i, x_j must be considered. However, an n order algorithm exists. The algorithm can be interpreted by using a taut string metaphor.

Suppose that the curves $F + d$ and $F - d$ are solid. Now suppose we stretch a string between $[x_1, d]$ and $[x_n, F(x_n) - d]$. If d decreases, the string must bend between the two solid curves, and will form a convex minorant for $F+d$ between (x_1, x_L) and concave majorant for $F - d$ on (x_U, x_n).

As d gets smaller, the interval (x_L, x_U) gets smaller and eventually the string

will get bent between (x_L, x_U) in such a way that the total shape of the string is no longer unimodal. We now find the smallest value of d such that the string still has an unimodal shape. For that value, the maximum distance between the left part of the string (which is a convex minorant for $F + d$) and F or right part of string (which is a concave majorant of $F - d$) and F is then d. Also, the maximum distance between the middle part of the string (between x_L and x_U) and F is at most d, because it is bound between $F + d$ and $F - d$. So we can use the theorem, which results in $d = D$. A graphical example for given in Figure 2.3.

The greatest disadvantage, as noted by authors who have calibrated the test, is that the original implementation is very conservative [2]. The main cause is the fact that the original dip test is calibrated for the uniform distribution. In the dip test paper, the author argues the uniform distribution is asymptotically the least-favourable unimodal distribution. This causes the test to be conservative in more favourable situations.

However, the author in [2] also notes that one of the attractive features of the dip test is that it does not estimate a density explicitly, unlike it is done in the Silverman's test for instance.

2.4 Challenges for calibration

Both the original Silverman's test and dip test are known to be very conservative. This conservativeness also leads to reduced power, rendering the tests less useful in practice.

However, both tests have been improved after their initial publication by calibrating in different ways.

Silverman's test is calibrated using an asymptotic approach in [7], where the limiting distribution of the test statistic is identified. In the same paper, the distribution is also calibrated using Monte-Carlo methods.

The dip test (and thus also the excess mass) are calibrated in [2] using Monte-Carlo methods as well. The authors prove that the test statistic Δ (in their case, the excess mass statistic) can be refactored, where only one factor, denoted by c, depends on the density function and the location of the mode. The bootstrap samples come from a "calibration distribution" dependent a transformation of the value of c. The value of c should therefore be estimated from the data. When testing the bootstrap samples, the obtained values of Δ should be similar to the value of Δ of the original sample, should the null hypothesis be correct.

The dip test is also calibrated in [3]. The statistic is calibrated using a bootstrap approach. The bootstrap samples are drawn from $f(h_{\text{crit}})$, which is the kernel density estimate of the Silverman's test where the minimum bandwidth h is chosen such that the density has one mode, but a further decrease in h produces more modes.

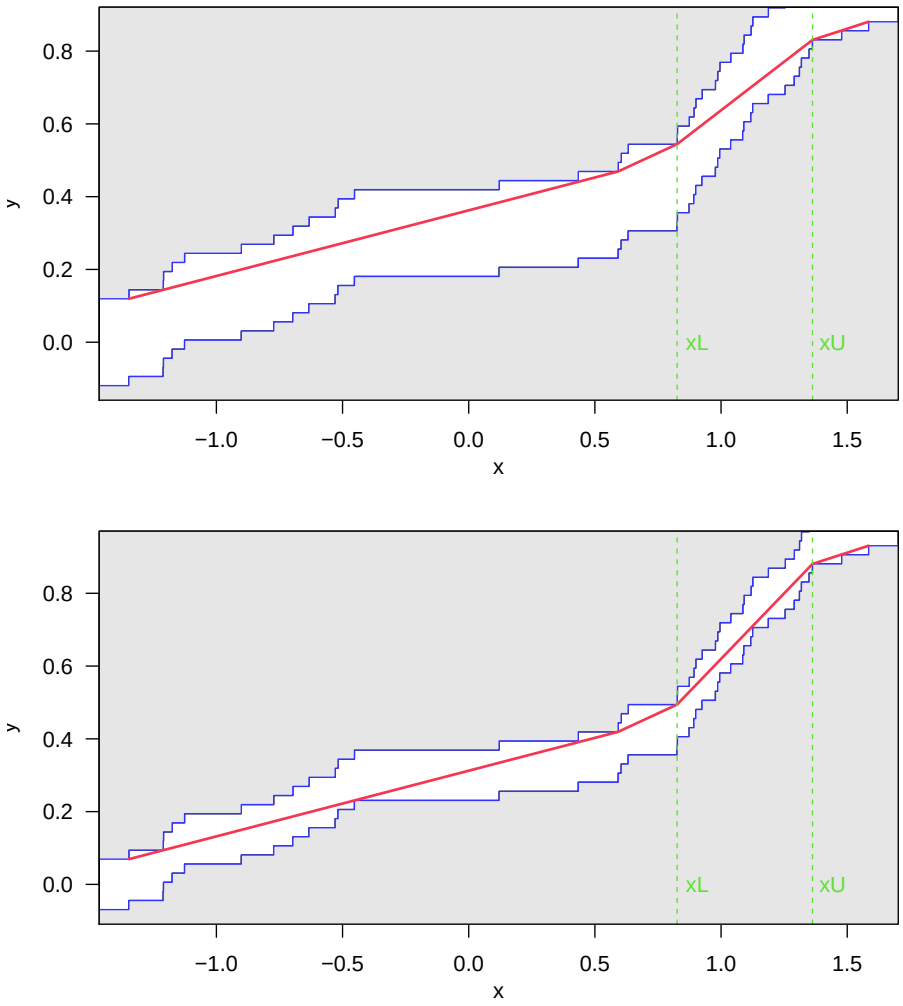

Figure 2.3: Graphical example of the taut string metaphor used in describing the algorithm for the dip test. The red line is the string, the bottom blue line is $F - d$ and the upper blue line is $F + d$. Note that the string forms the gcm on (x_1, x_L) for $F - d$, and the lcm on (x_U, x_n) for $F + d$. The bottom plot depicts the minimum value of d. If we would decrease d even further, the string would get bent out of its unimodal shape at around $x \approx -0.5$

Chapter 3

Our approach for devising new tests

In this chapter, we will outline our approach for devising new tests for multimodality.

Suppose we are given data $X = (x_1, \ldots, x_n)$ assumed to be an i.i.d. sample from unknown probability density function f. We are interested to test the following competing hypotheses:

$$H_0 : f \text{ is unimodal} \qquad \text{vs.} \qquad H_1 : f \text{ is multimodal.}$$

In Chapter 2 we have examined the existing tests for multimodality. We have shown that a test based on the global kernel density bandwidth faces some inherent issues. Furthermore, calibrating the tests for multimodality is not straightforward, since H_0 is composite. Because of these challenges, we focus on fitting a unimodal CDF to our data. We can then resample from the fitted CDF in a bootstrap procedure to obtain well-calibrated tests.

We will now outline the general bootstrap procedure which will be used for our tests. Next, we will introduce test statistics which can be used in conjunction with this bootstrap approach.

3.1 Outline of the bootstrap procedure

To devise new, well calibrated tests, our approach will use the bootstrap method studied in [23]. We first outline the procedure:

1. From the data X, compute the ECDF $\hat{F}$ and the CDF estimate $\hat{G}$.
2. Compute a discrepancy measure between $\hat{F}$ and the estimate $\hat{G}$, which will be used as a test statistic. Denote the statistic by D.
3. Generate $B \in \mathbb{N}$ bootstrap samples of size n from $\hat{G}$.

4. For each bootstrap sample $j \in B$, compute the respective ECDF $\hat{F}_j$ and the CDF estimate $\hat{G}_j$. With these, compute the test statistic used in step 2, denoted by D_j.
5. Compute the (approximate) p-value of the test as follows:

$$p \approx \frac{\#\{j : D_j \geq D\}}{B}.$$

Note that to use this procedure, we need to find a CDF estimate $\hat{G}$ that is sensible for the data we are given and must be in the set of null distributions. In the setting of multimodality, the hypothesized distribution will be a unimodal distribution function. We can then test how well a unimodal distribution fits the data - if it does, we have some confidence the data came from a unimodal distribution. If it does not, we have reason to suspect multimodality.

3.2 Test statistics

There are a number of well-known discrepancy measures that can be employed when using this bootstrap approach. We will use the Kolmogorov-Smirnov (KS), Cramér-Von Mises (CvM) and the Anderson-Darling (AD) statistic in our own devised tests. The statistics for two CDFs F and G, where F is the ECDF and G is the hypothesized distribution, are now defined.

The KS statistic measures the greatest absolute distance between the two CDFs:

$$D_{\text{KS}} = \sup_x \mid F(x) - G(x) \mid .$$

The CvM statistic measures the area between the two curves:

$$D_{\text{CvM}} = \int \big(F(x) - G(x)\big)^2 dG(x).$$

The AD statistic also measures the are between the two curves, but places more weight on the difference in the tails:

$$D_{\text{AD}} = \int \frac{\big(F(x) - G(x)\big)^2}{G(x)(1 - G(x))} dG(x).$$

The expressions can be simplified when the hypothesized distribution G is continuous. Note that the ECDF is piecewise constant. It is then easy to see that for the KS statistic, the maximum difference must occur on either the datapoints X_i or arbitrary close to the datapoints, say $X_i - \epsilon$ where epsilon is arbitrarily small but larger than 0. We can then compute the test statistic as follows:

$$D_{\text{KS}} = \max_{i \in \{1,\dots,n\}} \max \big\{ \mid F(X_i) - G(X_i) \mid, \mid F(X_i - \epsilon) - G(X_i - \epsilon) \mid \big\}.$$

The expression can be simplified even further when we note that $F(X_i) = \frac{i}{n}$ when X is sorted. Define $U_i = G(X_i)$ and let $U_{(i)}$ denote the correspond-

ing order statistics[1]. We can then simplify the expression of the KS statistic to:

$$D_{\mathrm{KS}} = \max_{i \in \{1,\ldots,n\}} \max \left\{ \mid \frac{i}{n} - U_{(i)} \mid, \mid \frac{(i-1)}{n} - U_{(i)} \mid \right\}.$$

In a similar way, we can also find simplified expressions for the CvM and AD statistics:

$$D_{\mathrm{CvM}} = \frac{1}{12n^2} + \frac{1}{n} \sum_{i=1}^{n} \left(U_{(i)} - \frac{2i-1}{2n} \right)^2,$$

and

$$D_{\mathrm{AD}} = -1 - \frac{1}{n^2} \sum_{i=1}^{n} (2i-1) \left(\log(U_{(i)}) + \log(1 - U_{(n-i+1)}) \right).$$

Note that, when $G(X_1) = 0$ and $G(X_n) = 1$, the expression for the AD statistic is not well-defined, as this will result in $\log(0) = -\infty$ in the sum.

[1]Note that since G is monotone, it holds that $U_{(i)} = G(X_{(i)})$ and the ordering of the data is not perturbed.

Chapter 4

LOG test

We now introduce a new test for multimodality; the LOG test. We will first explain why the class of logconcave densities is interesting in the context of multimodality. We will then describe procedure of the test. Finally, we offer some numerical results, both in regard to calibration and power of the test.

4.1 Introduction

In the context of multimodality, the class of logconcave densities is interesting because it can be used in conjunction with the non-parametric MLE. The non-parametric MLE of a logconcave density is well-defined. In contrast, the non-parametric MLE of a unimodal density does not exist. To see that this is true, note that we can fit a unimodal density to the data where we accumulate probability mass to one data point. This increases the likelihood by an arbitrary amount. We have formulated a rigorous proof of this interpretation, which is deferred to Appendix B for the sake of brevity in the introduction here.

Note that any logconcave density must be unimodal. Therefore, we can use logconcavity as a surrogate for unimodality.

However, not every unimodal density is logconcave. An example is a unimodal normal mixture[1] where the density has a shoulder, which we defined in Section 2.1. Examples of such a densities, and the logarithms of the density functions, are given in Figure 4.1. A less obvious example is the Student's t distribution, which is non-logconcave for any number of degrees of freedom.

Since the class of unimodal densities is not equal to the class of logconcave densities, using logconcavity as surrogate for unimodality will be problematic in some cases. The devised test is fundamentally a test for logconcavity. It might be a valuable test when assessing multimodality, but conclusions from this test

[1] Note that a normal mixture is not generally multimodal. The weights, the standard deviations and the means of the components causes the mixture to be either uni- or multimodal. If the mixture model is made up of two components with equal variance, solving for the number of modes is not hard. However, solving for the number of modes in a general normal mixture density is not trivial. The paper [20] explains this issue.

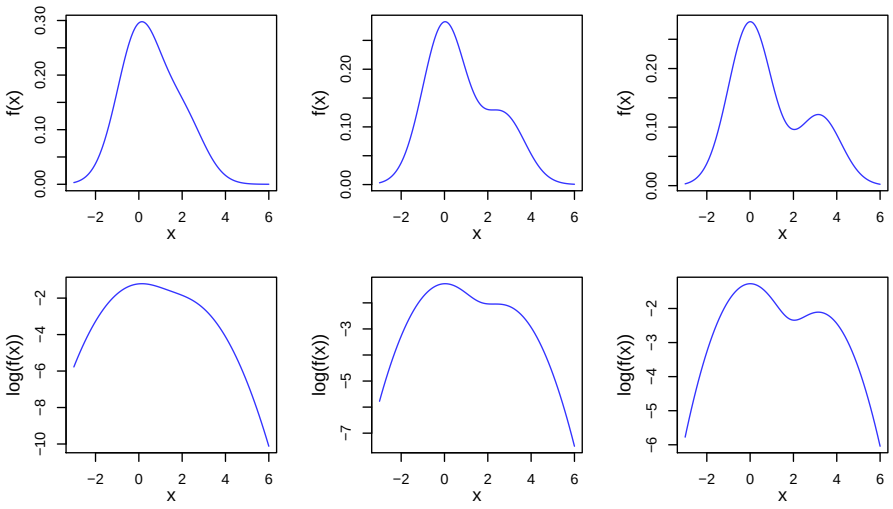

Figure 4.1: Examples of densities and the logarithm of these densities to illustrate that every logconcave density must be unimodal, but not every unimodal density is logconcave. Top row: densities of mixture models. Bottom row: corresponding logarithms of the density functions. As can be seen, the leftmost mixture model is unimodal and logconcave, the middle mixture model is unimodal but non-logconcave (and has a shoulder), and the rightmost mixture model is multimodal and therefore non-logconcave.

should be approached with great care. We will illustrate this point by example in Section 4.4. Nevertheless, this test will be useful in its own right to test for logconcavity - which might be an important modelling feature.

4.2 Description of the LOG test

The LOG test uses the bootstrap procedure outlined in Section 3.1. The CDF estimate used to generate bootstrap samples is a fitted logconcave density which we denote by F_{lc}.

The logconcave density is fitted by using the methods implemented in the *R* package called *logcondens* [5]. The underlying algorithm for fitting the logconcave density estimate is developed in [4]. We have included pseudocode of the full procedure of the LOG test in Appendix A.1.

4.3 Numerical results

We have conducted several simulations to assess both the calibration and the power of the test. The results of these simulations will be presented in this section. The results of both the calibration and power assessment are also given in tabulated form in Appendix C.1.

4.3.1 Assessing the calibration of the LOG test

We will now discuss the calibration of the LOG test. We first identify a few significant distributions to use for assessing the significance levels.

- Normal distribution. The normal distribution is naturally a very common unimodal distribution. The test should be calibrated well under this distribution.
- Student's t distribution. The Student's t distribution is not a logconcave distribution, so the test should reject the null under this distribution. However, the Student's t distribution is logconcave on a large part of it's domain. It is therefore interesting to see if the LOG test will be well-calibrated or not.
- Uniform distribution (continuous). The uniform distribution is asymptotically the least-favorable unimodal distribution according to [11] for a large family of distributions. The dip test is calibrated using the uniform distribution.
- Distribution with a shoulder. A distribution with a shoulder is asymptotically also an unfavorable distribution. Both [8] and [3] propose to use this distribution to calibrate tests for multimodality[2]. It is therefore useful to know if the LOG test is well-calibrated for this distribution. Note that a shoulder density is not logconcave, so we can expect to see some anti-conservativeness here.
- Laplace distribution. The Laplace distribution both unimodal and logconcave. However, the distribution is not strictly logconcave. If f is the density of a Laplace distribution with location parameter μ, then $\log(f)$ is linear and increasing on $(-\infty, \mu)$, linear and decreasing on (μ, ∞). This is therefore a borderline case for the LOG test. It is useful to know if the test is well-calibrated for this case.

We now identify the quantiles of the p-value distribution to assess the calibration of the LOG test. The actual significance level of the test at nominal level α is calculated by finding the fraction of times we reject H_0. That is, we find the amount of p-values observed to be smaller than α.

We ran the test 500 times for each of the five distributions considered above. Each individual sample had size $n = 100$. For each test, a total of $B = 500$ bootstrap samples where generated. Each test returned three p-values, based on the KS, CvM and AD statistic. We assess the calibration of the LOG test when based each of those three statistics.

The resulting quantiles are plotted in Figure 4.2.

[2] A simple argument for using a shoulder density to calibrate tests for multimodality from [3] can be summarized as follows: in a parametric setting, when dealing with a composite $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, we would first construct and calibrate a test for the simple null hypothesis $\theta = \theta_0$. This approach will then give uniformly most powerful tests, provided that the likelihood ratio is monotone. In our setting, the simple null hypothesis would be a density which is just on the boundary of being multimodal. A density with a shoulder corresponds to this description.

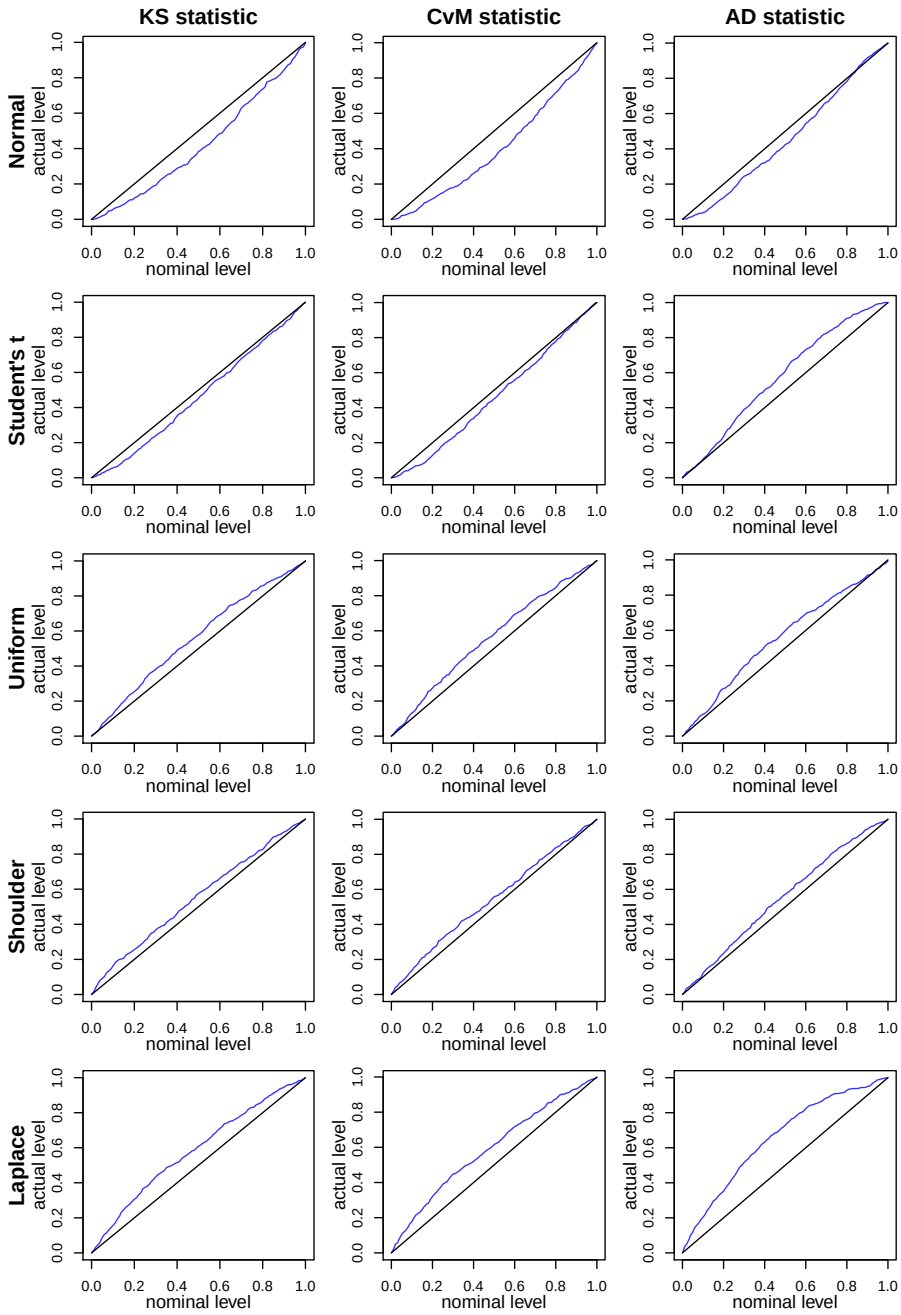

Figure 4.2: Overview of the calibration assessment. Each row corresponds to one distribution, which is denoted at the left side. Each column corresponds to the LOG test based on a different statistic, denoted at the top. Each sub-plot denotes the nominal or target significance level at the x-axis and the actual significance level at the y-axis. The black line represents perfect calibration: that is, the nominal level is equal to the actual level for every level.

We have not included the quantiles of the original dip test in Figure 4.2. The original dip test is very conservative under all distributions considered except for the uniform distribution, in which case it is well-calibrated. We see that the LOG test is reasonably well-calibrated, although it is usually somewhat anti-conservative. While this anti-conservativeness might be undesirable in some cases, the actual levels of the LOG test are much closer to the target level than those of the original dip test.

Note that for the Student's t distribution, the test is well-calibrated when using the KS or CvM statistic. However, the AD statistic shows anti-conservatism. This can easily be explained. The tails of the Student's t distribution are too heavy for a logconcave distribution and thus the logconcave estimate has lighter tails. This means that there is quite some difference in the tails of the ECDF of the data and the fitted logconcave distribution. The AD statistic places more weight on the differences in the tails of the fitted distribution, so the p-values under the AD statistic are lower and thus the test is anti-conservative. Strictly speaking however, the LOG test tests for logconcavity, so the AD statistic is indicating the non-logconcavity more strongly than the other two statistics.

The test is quite anti-conservative for the Laplace distribution, especially under the AD statistic. Since the Laplace distribution is a borderline logconcave distribution, we can expect the test to have difficulties deciding between logconcavity and non-logconcavity.

Note that the test is reasonably well-calibrated under the shoulder distribution, while the distribution is not logconcave.

4.3.2 Assessing the power of the LOG test

We will now discuss the power of the LOG test. To assess the power, we consider the following two mixture distributions:

$$\frac{7}{10}\mathcal{N}(0,1) + \frac{3}{10}\mathcal{N}(\mu,1), \tag{4.1}$$

$$\frac{1}{2}\mathcal{N}(0,1) + \frac{1}{2}\mathcal{N}(8,\sigma). \tag{4.2}$$

The densities, along with the logarithms of the density functions, of these mixture distributions are plotted for different values of μ and σ in Figure 4.3.

We now use the test for the two mixture distributions for different values of μ and σ. For each test, the input sample size is $n = 100$ and $B = 500$ bootstrap samples are generated.

For distribution (4.1), we consider $\mu \in \{1.5, 1.9, 2.3, \ldots, 4.7\}$. We run the test 100 times for every parameter value. Note that for $\mu \approx 2.7$, the distribution has a shoulder and is borderline unimodal. The distribution is borderline logconcave for $\mu \approx 2$.

For mixture distribution (4.2), we consider $\sigma \in \{4, 5, 6, \ldots, 17\}$. We run the test 50 times for every parameter value. Note that for $\sigma \approx 10$ the distribution

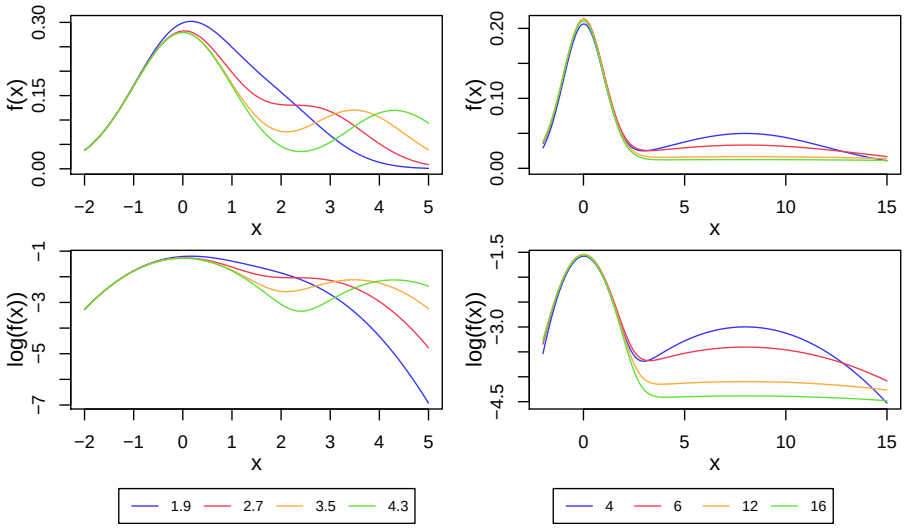

Figure 4.3: Examples of the density functions and the corresponding logarithms of the density functions, considered for different parameters. Left: density of the mixture distribution 4.1 for different values of μ (top) along with the logarithm of the density (bottom). Right: density of mixture distribution 4.2 for different values of σ (top) along with the logarithm of the density (bottom).

can be regarded unimodal[3]. The distribution is non-logconcave for any value of σ considered. The distribution was chosen specifically because of this. The distribution represents a very unfortunate case for the LOG test, since for large values of σ, the distribution is essentially unimodal with a very heavy right tail.

The results of the simulations can be found in Figures 4.4 and 4.5.

4.4 Drawbacks of the LOG test

As we saw in Section 4.3.2, the LOG test can draw incorrect conclusions when used to assess multimodality for certain distributions. We showed that for a unimodal distribution with very heavy tails, the test potentially rejects the null hypothesis even for low significance levels. If the test would be used purely to test for logconcavity, which can be valuable in some cases, the test works as expected.

It is important reiterate here that the test does not actually test for multimodality. The test fundamentally only tests for logconcavity. It the test rejects logconcavity, then this does not automatically mean we should reject unimodality. The possibility exists the test rejected the null based on, for instance, heavy tails, not on presence of multiple modes.

[3]Strictly speaking, the distribution still has a very slight mode, but this mode can be regarded insignificant.

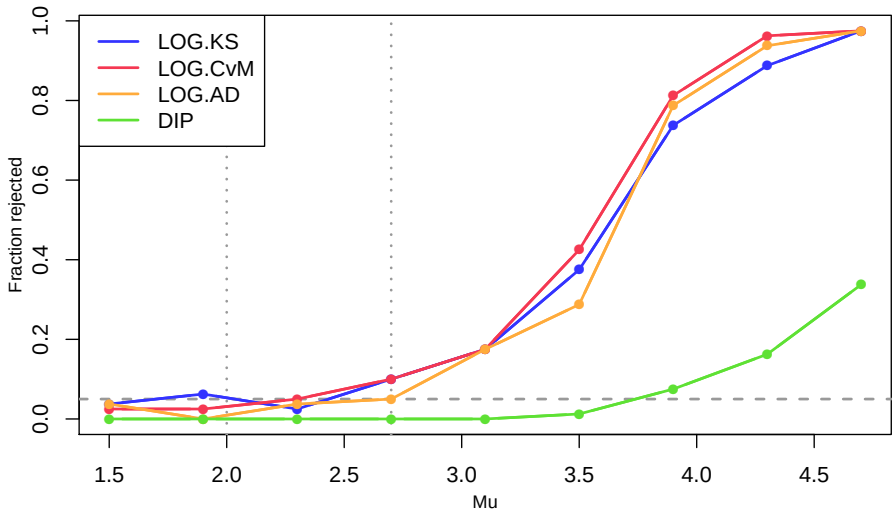

Figure 4.4: Power results of the LOG test for mixture distribution (4.1) for different values of μ. The leftmost dotted vertical line represents the value of μ such that the density is borderline concave, the rightmost line such that it is borderline unimodal. The y axis portrays the fraction of times the null hypothesis was rejected under significance level $\alpha = 0.05$. The dashed horizontal grey line represents $\alpha = 0.05$.

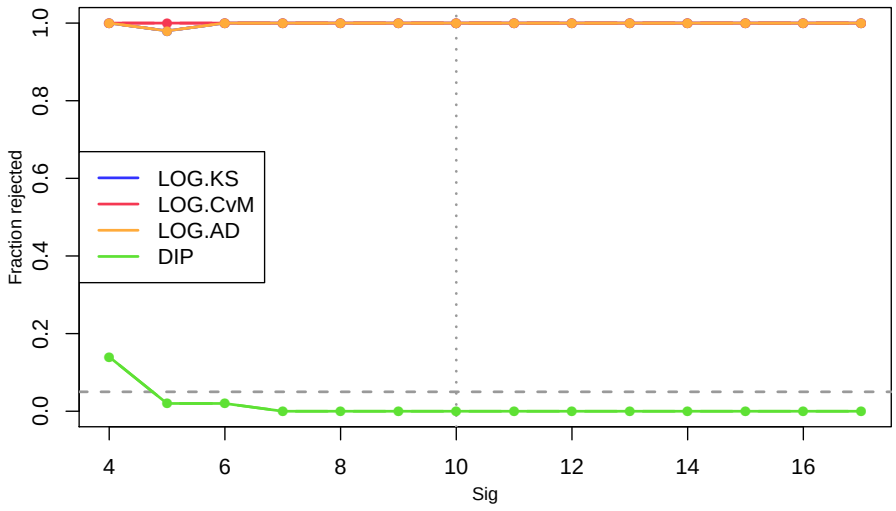

Figure 4.5: Power results of the LOG test for mixture distribution (4.2) for different values of σ. The dotted vertical line represents (approximate) borderline unimodality. The y axis portrays the fraction of times the null hypothesis was rejected under significance level $\alpha = 0.05$. The dashed horizontal grey line represents $\alpha = 0.05$.

Ideally, we would find a transformation of the data that maps a unimodal density to a logconcave density, and a multimodal density to a non-logconcave density. We have not found such a transformation.

However, we can solve the problem partially by using a censoring approach which aims to remove the influence of heavy tails[4]. Our censoring approach is discussed in the next section.

4.4.1 Censoring

The LOG test rejects unimodal densities with heavy tails, as these densities are non-logconcave. Ideally it would fail to reject the null hypothesis in these cases, if we aim to use the test for multimodality. One way to mitigate this problem is to censor the data in the tails, so only the center of the density is evaluated.

We now introduce a censoring procedure. The procedure is iterative and based on the following idea which is first outlined informally.

Suppose we have data $X = (x_1, \ldots, x_n)$ for which we would like to test for multimodality. Using the logconcave test we obtain a p-value. Now we remove a small number of outliers from the data, and use the logconcave test again.

We expect the resulting p-value to be close to the value obtained when using the whole dataset. After all, the small number of outliers removed should not have a significant effect on the conclusion of the test.

If, however, the change in p-value is large, then the small number of outliers heavily influenced the tests decision. However, this is not what we expect; the removed datapoints are outliers and only a small amount is removed, after all. So we censor these datapoints, and base our decision on the censored dataset.

Now, this approach can be used repeatedly, removing "chunks" of data in every iteration until the p-values stabilize. Of course, we need to formalize this notion of stabilization. We introduce a parameter ρ, chosen by the user of the test, which dictates whether two p-values are, in the sense of this approach, close enough. We name this parameter ρ the tolerance.

The censoring procedure is now outlined step-by-step. Let $c \in \mathbb{N}_+$ be the *chunk size*. Let $\rho \geq 0$ be the *tolerance*. Let $X = (x_1, \ldots, x_n)$ be the sample for which we want to test multimodality.

1. Denote the original sample by $X^{(1)}$. Use the LOG test for this sample and denote the corresponding p-value by p_1.
2. Identify the c largest outliers in the sample and remove them from $X^{(1)}$. Denote the resulting dataset by $X^{(2)}$.
3. Use the LOG test for $X^{(2)}$ and denote the corresponding p-value by p_2.

[4]Note that problematic regions, which cause the density to be non-logconcave but retain unimodality, are not necessarily present at the tails of the density. An example is the density with a shoulder. The non-logconcavity of the density is caused by the flat piece around the shoulder, which is not necessarily at the tails.

4. Set $i = 2$. While $|p_{i-1} - p_i| > \rho$:
    - Identify the c largest outliers in $X^{(i)}$ and remove them from $X^{(i)}$. Denote the resulting dataset by $X^{(i+1)}$.
    - Use the LOG test for $X^{(i+1)}$ and denote the corresponding p-value by p_{i+1}. Increment i.
5. Denote the last obtained p-value p_i by p. This is the resulting p-value of the test.

It is unclear if this procedure always yields accurate results. The results can heavily rely on the parameters c and ρ, both of which are not chosen in a natural way but instead need to be chosen on the users discretion.

We have tested the procedure both in terms of calibration and power. We chose $c = 2$ and $\rho = 0.1$. The sample size was $n = 100$, the number of bootstrap samples was $B = 500$ and we simulated 500 results per distribution. We evaluated the censored LOG test when using the AD statistic only as it is the most anti-conservative under some distributions. First, we use the procedure on the Student's t and Laplace distribution to see if the calibration of the LOG test can be improved. As shown in Section 4.3.1, the LOG test is somewhat anti-conservative for the Student's t and Laplace distribution. We include the results when evaluating the normal distribution to verify the test is significantly affected in logconcave cases. The results are given in Figure 4.6.

We also use the procedure on the problematic mixture distribution 4.2, for which the LOG test rejects nearly always for any value of the parameter σ. The results are given in Figure 4.7.

Note that the calibration has been improved for the test under the Student's t and the Laplace distribution, but the conservativeness under the normal distribution increased. Furthermore, the power results show the issues are not resolved.

While no proof of guarantee of the censoring procedure is given, the numerical results show the procedure can be valuable when adjusting the calibration of the LOG test. However, as the LOG test is not improved in the difficult case of mixture distribution (4.2) the procedure does not fix all issues, and the LOG test should be employed with care even when using this censoring approach.

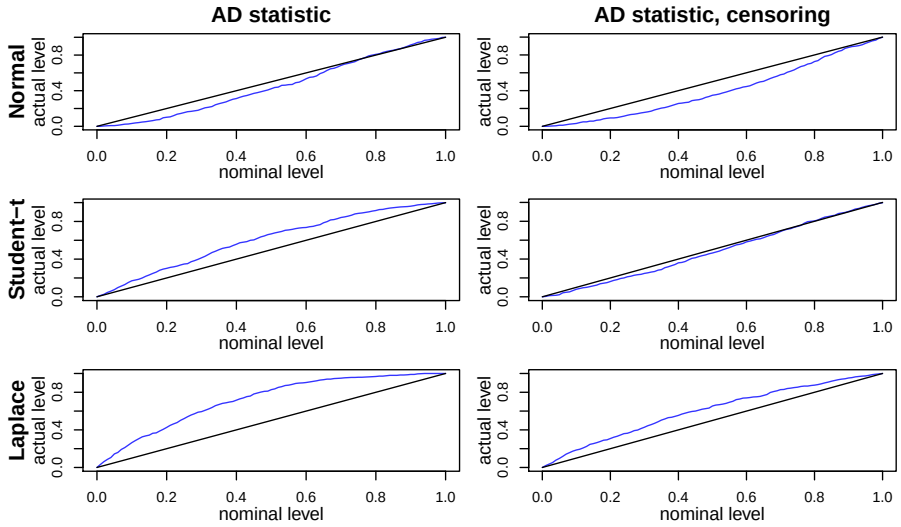

Figure 4.6: Calibration assessment for the LOG test when using the censoring procedure and the AD statistic. Each row corresponds to one distribution, which is denoted at the left side. The left column corresponds to the calibration of the LOG test without censoring, the right column with the censoring procedure. Each subplot denotes the nominal or target significance level at the x-axis and the actual significance level at the y-axis. The black line represents perfect calibration: that is, the nominal level is equal to the actual level for every level.

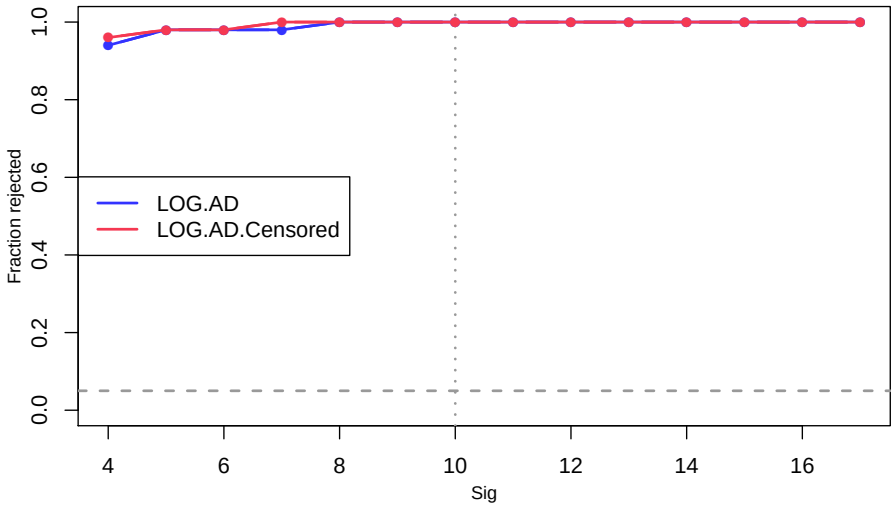

Figure 4.7: Power results of the LOG test when using the censoring approach for mixture distribution (4.2) for different values of σ. The dotted vertical line represents (approximate) borderline unimodality. The y axis portrays the fraction of times the null hypothesis was rejected under significance level $\alpha = 0.05$. The dashed horizontal grey line represents $\alpha = 0.05$.

Chapter 5

String test

In this chapter, we will introduce a new test for multimodality: the string test. The test is inspired by the dip test, which is studied in Section 2.3.2. The test is named after the taut string metaphor (illustrated in Figure 2.3) that explains the workings of the dip test. The string will, with a slight modification, serve as our unimodal CDF estimate. We can use this estimate in our bootstrap approach outlined in Section 3.1. The interpretation of the test is therefore quite intuitive.

We will first introduce the test and describe the procedure. We also offer numerical results, both in regard to calibration and power of the test. We will analyse the test critically and show that the test does not always work as expected.

5.1 Introduction

The string test is developed to be used in the bootstrap procedure outlined in Section 3.1. To use this procedure, we need a unimodal CDF estimate for the sample. We first show the issues when trying to find a best-fitting unimodal CDF from scratch. We then demonstrate how the theory from the dip test paper can be used to find a best-fitting unimodal CDF.

5.1.1 Challenges when fitting a unimodal CDF

Since a unimodal CDF is convex in $(-\infty, m]$ and concave in $[m, \infty)$, we could do the following. Define:

$$U_m(x) = \begin{cases} f_1(x) & x \in (-\infty, m) \\ f_2(x) & x \in [m, \infty) \end{cases},$$

where f_1 is convex, f_2 is concave and both are chosen such that U_m is a distribution function. Suppose F is the ECDF of the data. Now we must find

f_1, f_2, m such that U_m minimizes a discrepancy measure. For the KS statistic, we would thus find f_1, f_2, m such that:

$$\min_{f_1, f_2, m} \sup_x |F(x) - U_m(x)|.$$

However, finding this minimum is not straightforward. We can reduce the problem to an easier problem by specifying the curves f_1 and f_2. Define:

$$F_m(x) = \begin{cases} g(x) & x \in (-\infty, m] \\ l(x) & x \in [m, \infty) \end{cases},$$

where g is the gcm of F and l is the lcm of F, thus resulting in a unimodal CDF. Now, we can obtain a distribution estimate by finding the value of m that minimizes a discrepancy measure. For the KS statistic, we would thus find m in the following way:

$$\min_m \sup_x |F(x) - F_m(x)|.$$

However, finding this minimum can be very inefficient. If we restrict m to be any of the datapoints, then n different values for m need to be considered. In a naive implementation[1] we need to to order n calculations for every value of m, resulting in order n^2 calculations for computing the estimate.

More importantly, since we restricted the shape of F_m more than strictly needed and thus made assumptions about the data, F_m is generally not the minimizing unimodal distribution over all unimodal distributions:

- F_m is continuous, since $g(m) = l(m)$. However, the best fitting unimodal distribution might be discontinuous.
- The left part of the curve, the gcm of F, is by definition always less or equal to F and could be a better fit by shifting the curve to the positive y direction. Analogously, the lcm might fit better when shifted in the negative y direction. However, when the curves are shifted in these directions, F_m is no longer nondecreasing and thus no longer a CDF.

We can solve the second problem by introducing a connecting line in between the gcm and lcm curves. This introduces some new problems:

- We then need to estimate two parameters m_1 and m_2, introducing some complexity.
- We also need to take care the estimated CDF is unimodal by only allowing solutions where the connecting line is of maximum slope, which makes finding the minimum harder.
- Finding the amount of shifting required for the best fit is not straightforward.

We can solve these problems effectively by using theory developed in the dip test paper [11]. This will be explained in the next section.

[1] A more efficient implementation exists. In the dip test paper [11], it is noted that in a similar situation it is possible to do all necessary gcm and lcm calculations beforehand once in order n.

5.1.2 Finding a best-fitting unimodal CDF

The problems when finding a best-fitting unimodal CDF found in Section 5.1.1 can be solved in the following way. Consider the theorem from [11] which was covered in Section 2.3.2, restated here for convenience:

Theorem Let F be an arbitrary distribution function. Then $D(F) = d$ only if there exists a nondecreasing function G such that, for some $x_L \leq x_U$,

1. G is the gcm of $F + d$ in $(-\infty, x_L)$.
2. G has constant maximum slope in (x_L, x_U).
3. G is the lcm of $F - d$ in (x_U, ∞).
4. $d = \sup_{x \notin (x_L, x_U)} |F(x) - G(x)| \geq \sup_{x \in (x_L, x_U)} |F(x) - G(x)|$.

Note that the function G is the "string" in the taut string metaphor, if G is continuous in $[x_L, x_U]$ or $x_L = x_U$. We will now prove the following corollary:

Corollary Suppose F is the ECDF of data $X = (x_1, \ldots, x_n)$. If G satisfies the conditions in the previous theorem, then G^* is a unimodal CDF estimate which minimizes the KS statistic for the distributions F and G^*, where G^* is defined as:

$$G^*(x) = \begin{cases} 0 & x \in (-\infty, x_1 - \frac{G(x_1)}{a_1}) \\ a_1(x - x_1) + G(x_1) & x \in [x_1 - \frac{G(x_1)}{a_1}, x_1) \\ G(x) & x \in [x_1, x_n] \\ a_2(x - x_n) + G(x_n) & x \in (x_n, x_n - \frac{G(x_n)}{a_2}] \\ 1 & x \in (x_n - \frac{G(x_n)}{a_2}, \infty) \end{cases},$$

where $a_1 = \frac{G(x_2) - G(x_1)}{x_2 - x_1}$ and $a_2 = \frac{G(x_{n-1}) - G(x_n)}{x_{n-1} - x_n}$.

Note that G^* is the function G extended outside the interval $[x_1, x_n]$ to have range $[0, 1]$. This was done so G^* is a well-defined CDF.

Proof. Let G be a function satisfying the requirements in the theorem and F the ECDF of $X = (x_1, \ldots, x_n)$.

Since G meets the four requirements in the theorem, it holds that for some $x_L \leq x_U$:

$$d = \sup_{x \notin (x_L, x_U)} |F(x) - G(x)| \geq \sup_{x \in (x_L, x_U)} |F(x) - G(x)|,$$

and

$$D(F) = d.$$

Therefore:

$$D(F) = \sup_x |F(x) - G(x)|.$$

Since F is the ECDF of X, $F(x) = 0$ for $x \leq x_1$. Since G^* is a non-decreasing function:

$$\forall_{x < x_1} : |F(x) - G^*(x)| \leq G(x_1) = d.$$

Similarly, $F(x) = 1$ for $x \geq x_n$:

$$\forall_{x>x_n} : |F(x) - G^*(x)| \leq |1 - G(x_n)| = d.$$

Therefore:

$$\sup_x |F(x) - G^*(x)| = \sup_{x \in [x_1, x_n]} |F(x) - G^*(x)| = \sup_x |F(x) - G(x)| = D(F).$$

Now, note the definition of the dip statistic:

$$D(F) = \min_{H \in \mathscr{U}} \sup_x | F(x) - H(x) | .$$

Therefore:

$$\min_{H \in \mathscr{U}} \sup_x | F(x) - H(x) | = \sup_x |F(x) - G^*(x)|.$$

Note that G^* is unimodal, since G^* is non-decreasing, convex on $(-\infty, x_L)$, concave on (x_U, ∞) and has maximum slope on (x_L, x_U).

Since $G^* \in \mathscr{U}$, G^* is a unimodal CDF which minimizes $\sup_x |F(x) - G^*(x)| = D_{\mathrm{KS}}$. □

Thus, the function G^* is a sensible unimodal CDF estimate which can be used to generate bootstrap samples. Note that, while G^* minimizes the KS statistic, it is not necessarily unique.

The dip statistic, including the values of x_L and x_U, can be computed in order n. The theorem and corollary can therefore be used to compute a best-fitting unimodal CDF in terms of the KS statistic efficiently.

5.2 Description of the string test

The string test uses the bootstrap procedure outlined in Section 3.1. The CDF estimate used to generate bootstrap samples is the function G^* which was defined in Section 5.1.2 dependent on G. Note that the theorem does not fully specify the function G; the function G may have a discontinuity at a point in the interval $[x_L, x_U]$. We choose to connect the gcm and lcm curves by a straight line:

$$G(x) = \begin{cases} g(x) & x \in [x_1, x_L) \\ h(x) = \frac{g(x_L) - l(x_U) + 2D_n}{x_U - x_L} x + \frac{x_L(l(x_U) - D_n) - x_U(g(x_L) + D_n)}{x_L - x_U} & x \in [x_L, x_U) \\ l(x) & x \in [x_U, x_n] \end{cases},$$

where g is the gcm of $F + d$, l is the lcm of $F - d$ and D_n is the dip statistic. The gcm and lcm are connected by the straight line h. The function G is thus defined as the "string" in the taut string metaphor of the dip test - hence the name of the string test.

Resulting from our definition, G^* can only be discontinuous if $x_L = x_U$. We provide an example in Figure 5.1. We will prove in Section 5.4 that G^* is continuous when the data came from a continuous distribution. However, the

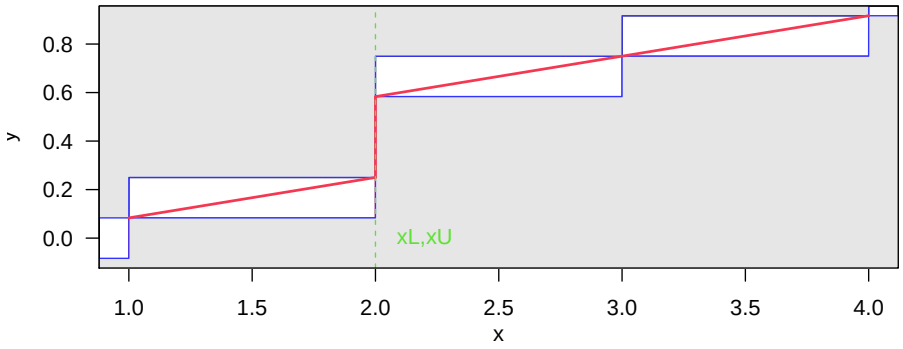

Figure 5.1: Graphical example of the taut string metaphor used in describing the algorithm for the dip test, when the shape of the string is discontinuous. The red line is the string, the bottom blue line is $F - d$ and the upper blue line is $F + d$. Note that the string forms the gcm on (x_1, x_L) for $F - d$, and the lcm on (x_U, x_n) for $F + d$.

theorem and our corollary which guarantee optimality of G^* do permit one discontinuity for the function G^*. Since we force our G^* to be continuous, we can potentially get suboptimal estimates for the unimodal CDF in regard to the KS statistic. We will explore this issue further in Section 5.4.1.

Note that we can use the string test with the CvM or AD statistic in the bootstrapping approach as well. However, the unimodal CDF is still fitted to minimize the KS statistic as before. So, the unimodal CDF is always fitted to minimize the KS statistic, but the discrepancy between the ECDF and the fitted unimodal CDF of the sample is evaluated using either the KS, CvM or AD statistic.

In Figure 5.2, the different steps taken to fit the unimodal CDF are depicted graphically for clarity. We have also included pseudocode of the procedure of the string test in Appendix A.2.

5.3 Numerical results

To assess the calibration and the power of the string test, we have conducted several simulations. The results of these simulations will be presented in this section. We conducted simulations with distributions also used in the assessment of the LOG test as well as simulation with distributions other authors have used. This allows us to compare the test with its alternatives. The results of both the calibration and power assessment (except for the comparison results) are also given in tabulated form in Appendix C.2.

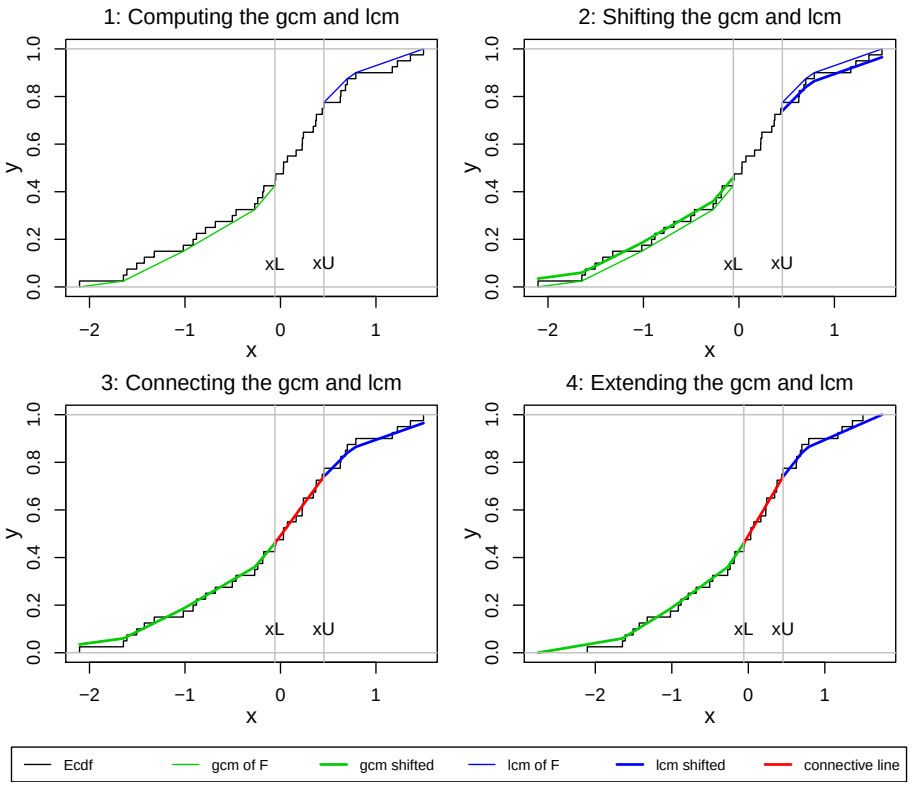

Figure 5.2: Example of the different steps in the string test to fit the unimodal CDF. The procedure is depicted here in 4 steps. Topleft: the gcm and lcm are computed on the intervals (x_1, x_L) and (x_L, x_U) respectively. Topright: the gcm is shifted upwards and the lcm is shifted downwards by the dip statistic, obtaining g and l. Bottomleft: The connective line h between the shifted gcm and lcm curves is computed. This plot depicts the function G. Bottomright: the shifted gcm and lcm curves are extended to 0 and 1 respectively to obtain a well-defined CDF function. This plot depicts the function G^*.

5.3.1 Assessing the calibration of the string test

We assess the calibration of the string test. We first conduct a study using the five distributions also considered for the LOG test in Section 4.3.1 which will allow us to compare the two tests. We identify the quantiles of the p-values similarly as we did for the LOG test.

We ran the test 500 times for each of the five distributions considered. Each individual sample had size $n = 100$. For each test, a total of $B = 500$ bootstrap samples where generated. Each test returned three p-values, based on the KS, CvM and AD statistic. We assess the calibration of the string test when based each of those three statistics. As noted before, the unimodal CDF is fitted to minimize the KS statistic regardless of the choice of test statistic, but the discrepancy between the ECDF and the fitted unimodal CDF is computed using the choice of test statistic.

The resulting quantiles are plotted in Figure 5.3.

We also compare the calibration of the string test with the calibrated dip test in [2]. In this paper, the results from a calibration study are given numerically for a select number of significance levels and sample sizes. The study is replicated for the string test exactly as in the paper: The sample size used in the paper[2] and in our study is $n = 200$, the amount of bootstrap samples generated is $B = 500$ and the number of simulations per distribution is 500. The results from the paper, along with the calibration results we obtained when using the string test with the AD statistic, are given in Figure 5.4. We show the results obtained with the AD statistic as it performed better than the other statistics. The distributions considered are specified in the paper except for the skewed normal and skewed Student's t distribution. We have estimated the parameters of these distributions from figures given in the paper, so results from these distributions must be taken with a grain of salt. The specification of these distributions as we have used them is given in Table 5.1.

The calibration results in Figure 5.3 show the string test is well-calibrated under all distributions considered, especially when using the AD statistic. The string test is calibrated better than the LOG test when examining these results. The original dip test is very conservative under all distributions considered, except for the uniform distribution, in which case it is well-calibrated. The string test is thus calibrated much better than the original dip test according to these results.

The calibration results in Figure 5.4 shows us the calibration from [2] comparable to the calibration of the string test but performs a little better. However, the actual level of the string test is never higher than the nominal level, while the actual level of the test from [2] is higher than the nominal level for the beta distributions for three of the four nominal levels considered. From these results it seems the calibration of the test in [2] is better than the calibration of the string test. However, since not all distributions were specified, the results must be taken with a grain of salt. Comparisons where both the calibrated dip test in [2] and the string test are implemented and used on precisely the same distribution must be done before conclusions should be drawn.

[2] In [2], $n = 50$ is also considered, but we only consider $n = 200$.

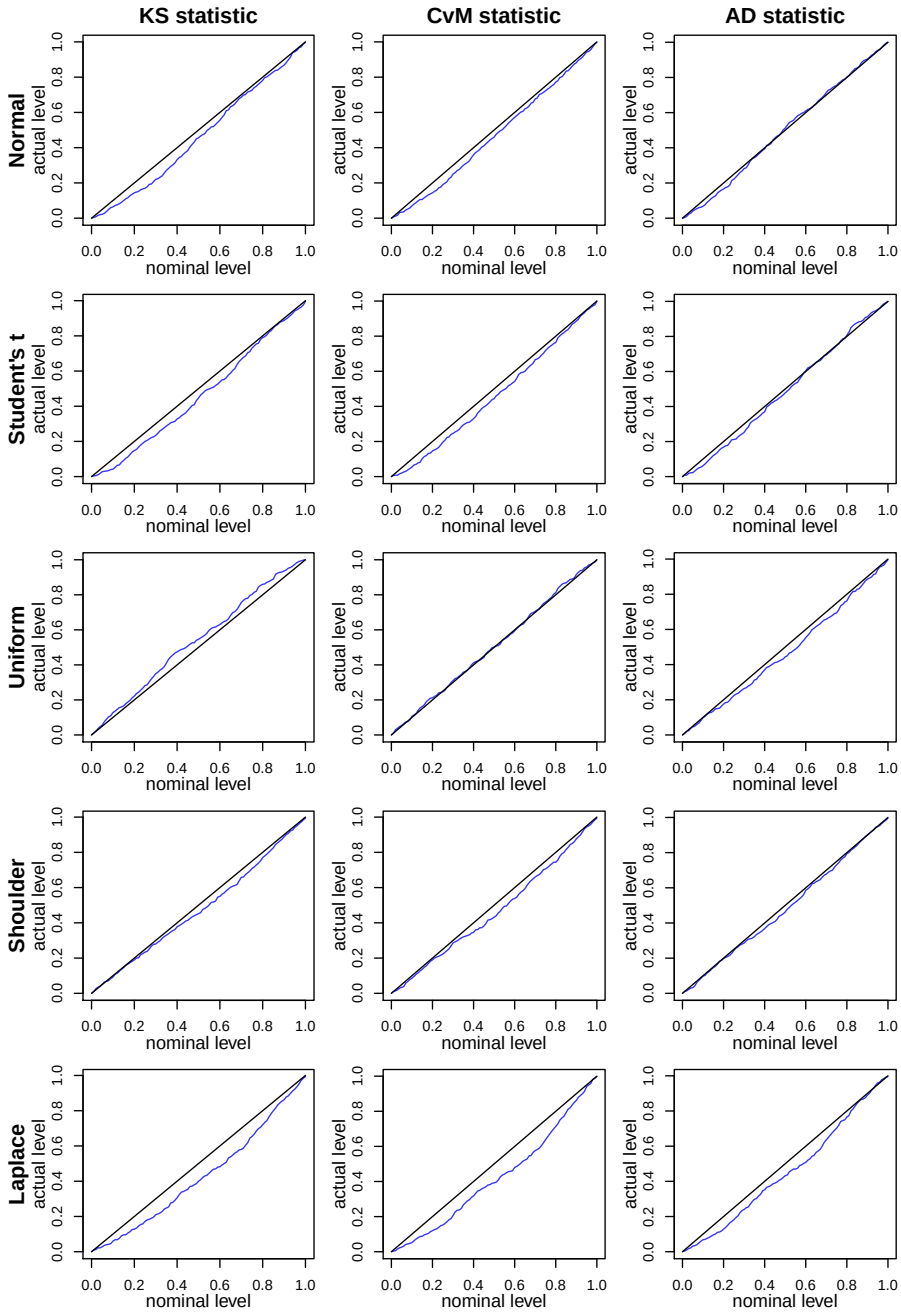

Figure 5.3: Overview of the calibration assessment of the string test. Each row corresponds to one distribution, which is denoted at the left side. Each column corresponds to the string test based on a different statistic, denoted at the top. Note that the unimodal CDF is still fitted using the KS statistic regardless of choice of statistic; only the discrepancy between the ECDF and the unimodal CDF estimate is calculated using the choice of statistic. Each sub-plot denotes the nominal or target significance level at the x-axis and the actual significance level at the y-axis. The black line represents perfect calibration: that is, the nominal level is equal to the actual level for every level.

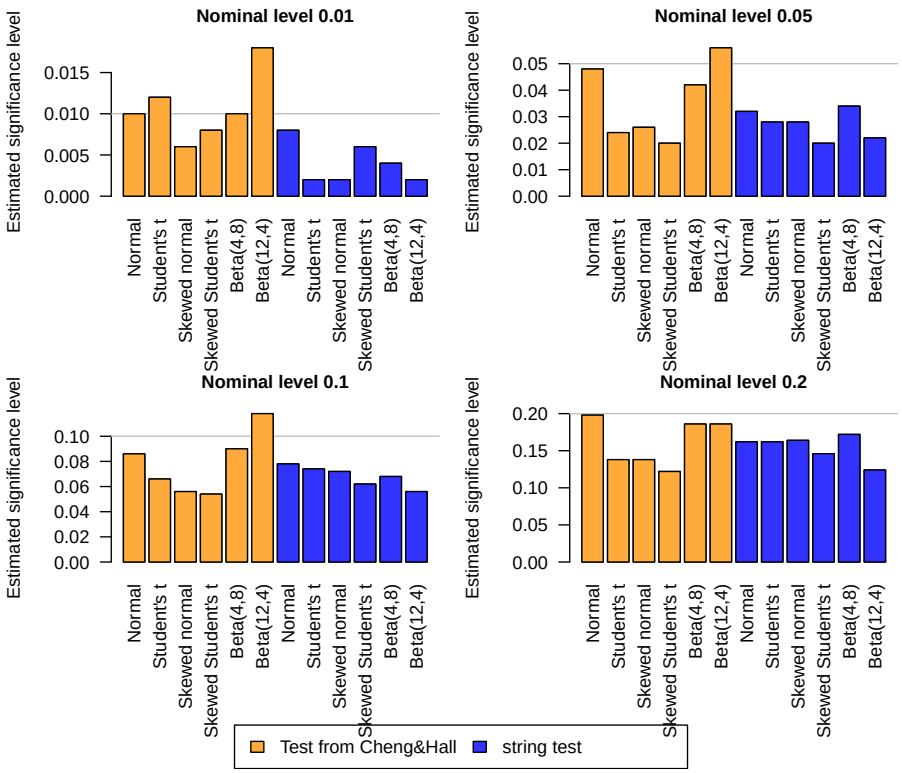

Figure 5.4: Calibration of the string test based on the AD statistic compared to the calibration in [2]. In this paper, the skewed normal and skewed Student's t distribution are not specified and were estimated from figures, so the results should be taken with a grain of salt. The specifications of these distributions as used by us are given in Table 5.1. In [2], the actual significance level is given for a select number of nominal levels. Each subplot corresponds to a nominal level. Each bar corresponds to the actual level under a different distribution denoted below. The green bars represent the results in [2], the blue bars represent the results of the string test when using the AD statistic. The gray horizontal line denotes the nominal level.

Name distribution	Specification
Skewed normal distribution	Skewed normal distribution with location parameter $\xi = 1.5$, scale parameter $\omega = 1.3$, slant parameter $\alpha = -3$.
Skewed Student's t distribution	Skewed Student's t distribution with location parameter $\xi = 1.5$, scale parameter $\omega = 1.3$, slant parameter $\alpha = -3$ and degrees of freedom $\nu = 6$.
Light tailed distribution 1	$0.5\mathcal{N}(0, 0.33) + 0.5\mathcal{N}(0.95, 0.4)$
Light tailed distribution 2	$0.5\mathcal{N}(-1.7, 1) + 0.5\mathcal{N}(1.7, 1)$
Light tailed distribution 3	$0.5\mathcal{N}(-1, 1) + 0.5\mathcal{N}(1, 0.4)$
Heavy tailed distribution 1	$0.5\mathcal{S}(0, 0.33) + 0.5\mathcal{S}(0.9, 0.4)$
Heavy tailed distribution 2	$0.5\mathcal{S}(-1.6, 0.95) + 0.5\mathcal{S}(1.6, 0.95)$
Heavy tailed distribution 3	$0.5\mathcal{S}(-1, 1) + 0.5\mathcal{S}(1, 0.4)$

Table 5.1: The specifications of the distributions used in the comparison of the string test and the calibrated dip test in [2], both in terms of calibration and in terms of power. Note that we used $\mathcal{S}(\mu, \sigma)$ here as a shorthand to indicate a Student's t distribution with location parameter μ, scale parameter σ and degrees of freedom $\nu = 6$.

5.3.2 Assessing the power of the string test

We now assess the power of the string test in two ways. First, we use the same mixture distributions used in Section 4.3.2. This allows us to compare the two tests in terms of power. The sample size, number of bootstrap samples generated, parameter values considered and runs per parameter value are the same as in Section 4.3.2. The results are plotted in Figures 5.5 and 5.6.

We also compare the power of the string test with the power of the calibrated dip test from [2]. In this paper, results from a power power study are not given numerically. We therefore resorted to estimating coordinates from the plots, so the comparison must be taken with a grain of salt. We have replicated the study from the paper: the sample size in the paper[3] and our study is $n = 200$, the amount of bootstrap samples generated is $B = 500$ and the number of simulations per distribution is 500. The results from the paper, along with the power results we obtained when using the string test with the AD statistic, are given in Figure 5.7. We show the results obtained with the AD statistic as it performed better than the other statistics. As with the comparison in terms of calibration in the previous section, the distributions used were not specified in the paper, but figures were given. We have estimated the distributions from these figures, furthermore increasing errors in this comparison - and thus to be taken with a grain of salt. The specifications of these distributions as we have used them are given in Table 5.1.

Compared to the LOG test, the string test has less power when considering distribution (4.1). This can be explained by the fact that the LOG test rejects the null when the density is non-logconcave which is true for lower values of

[3]In [2], $n = 50$ is also considered, but we only consider $n = 200$.

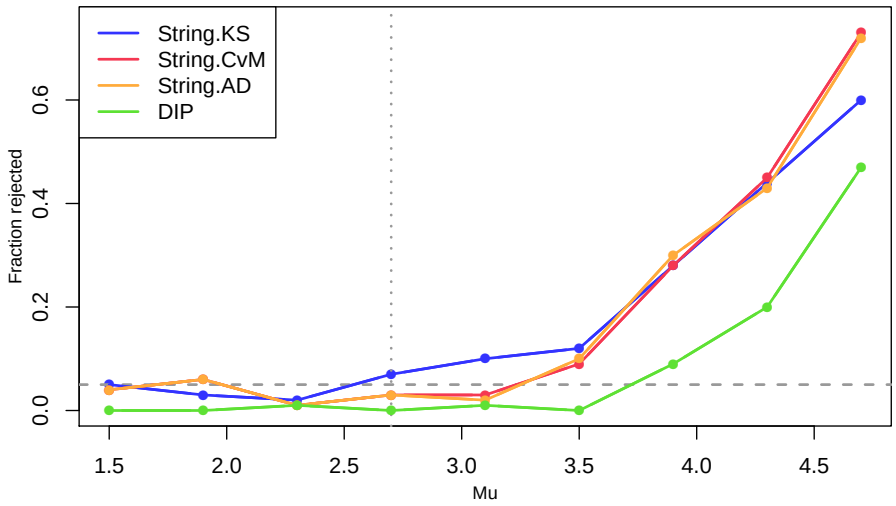

Figure 5.5: Power results of the string test for mixture distribution (4.1) for different values of μ. The dotted vertical line represents the value of μ such that the density is borderline unimodal. The y axis portrays the fraction of times the null hypothesis was rejected under significance level $\alpha = 0.05$. The dashed horizontal grey line represents $\alpha = 0.05$.

μ. However, the string test performs markedly better when considering distribution (4.2). While the LOG test had great difficulty as the distribution is non-logconcave, the string test does not face this issue and rejects for small values of σ while accepting the null for larger values of σ. It also performs much better than the original dip test.

Tt seems the power of the test from [2] is superior in all but one of the cases considered. The string test performs similar under "heavy tailed distribution 3". The test from [2] suffer in terms of power when applied to heavy-tailed distributions. However, the string test does not; note there is no loss of power between the light tailed and respective heavy tailed distribution when using the string test. However, as said before, the comparison is prone to many errors and should therefore be taken with a grain of salt. A firm conclusion cannot be drawn based on these results. Comparisons where both the calibrated dip test in [2] and the string test are implemented and used on precisely the same distribution must be done before conclusions should be drawn.

5.4 Analysis of the string test

In this section we conduct a critical analysis of the string test. We show the string test does not always do what is expected and is not guaranteed to fit the best-fitting unimodal CDF in terms of the KS statistic.

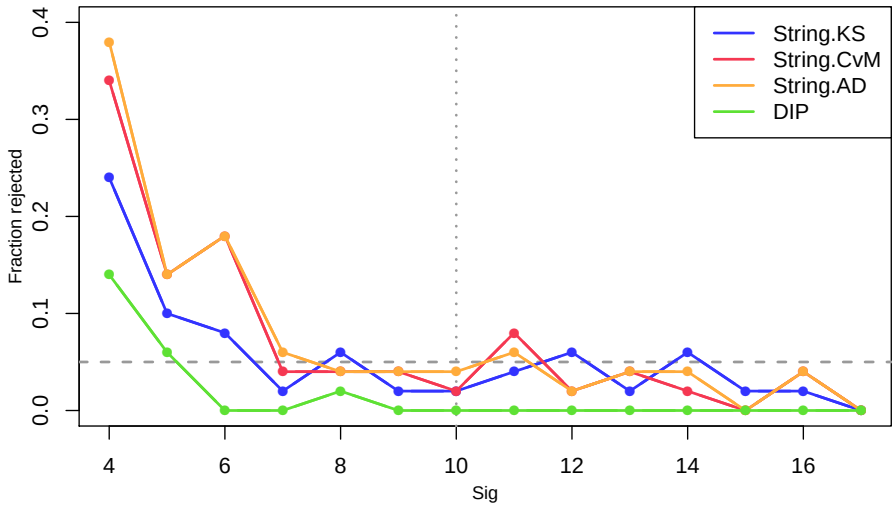

Figure 5.6: Power results of the string test for mixture distribution (4.2) for different values of σ. The dotted vertical line represents (approximate) borderline unimodality. The y axis portrays the fraction of times the null hypothesis was rejected under significance level $\alpha = 0.05$. The dashed horizontal grey line represents $\alpha = 0.05$.

5.4.1 Continuity of the unimodal CDF estimate

We first note that the string test always fits a continuous CDF if all data is distinct. We prove the following theorem:

Theorem When all data is distinct, G^* is continuous.

Proof. Note that G^* can only be discontinuous, if $x_L = x_U$.

Consider the algorithm described in the original dip test paper [11]. The algorithm of the dip test stops when the maximum absolute difference between the gcm and lcm on the interval (x_L, x_U) is less than or equal to the maximum absolute difference between the gcm and lcm outside this interval.

Note that the ECDF function F is piecewise constant with discontinuities of size $\frac{1}{n}$, since all data is distinct. Therefore, the minimum value of the absolute maximum difference between the gcm and lcm curves of F must be $\frac{1}{n}$.

Note that when $x_L = x_i$ and $x_U = x_{i+1}$ for some i, the maximum absolute difference between the gcm and lcm on the interval (x_L, x_U) is $\frac{1}{n}$, since all data is distinct. The maximum absolute difference thus attains the minimum value possible on this interval (x_i, x_{i+1}). Since the maximum absolute difference between the gcm and lcm outside the interval must necessarily be larger or equal than $\frac{1}{n}$, the algorithm must therefore have stopped before or when setting $x_L = x_i$ and $x_U = x_{i+1}$.

So the algorithm always stops before setting $x_L = x_U$. So for distinct data it always holds that $x_L \neq x_U$. Therefore G^* is continuous if all data is distinct. □

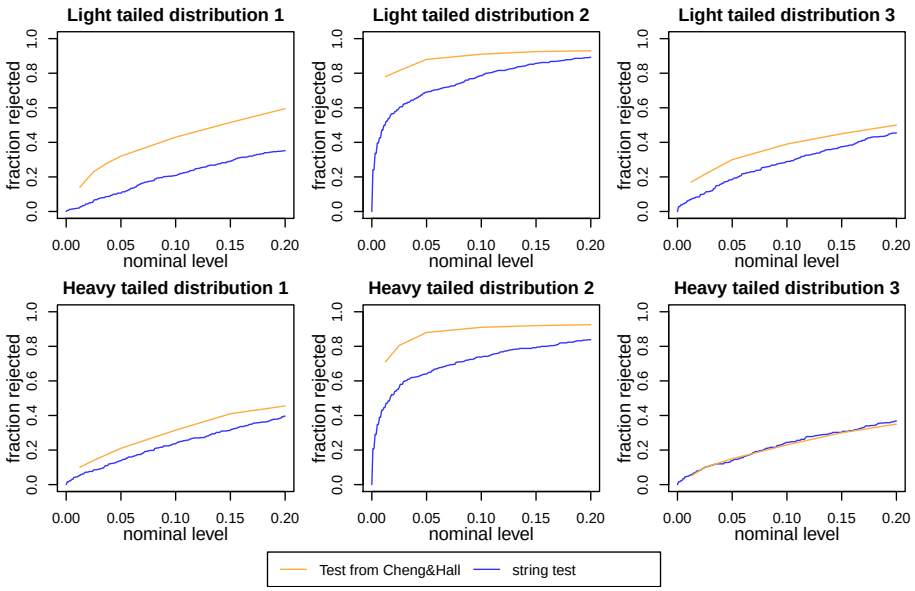

Figure 5.7: Power results of the string test when using the AD statistic compared to the power of the calibrated dip test in [2]. Each subplot denotes the nominal significance level at the x-axis and the fraction of times the null hypothesis is rejected at the y-axis. The distribution specifications were not given in [2] and were estimated by examining figures in the paper. The specifications of the distributions as we have used them are given in Table 5.1. The results were also not given numerically and were also estimated by examining figures. The results must therefore be taken with a grain of salt.

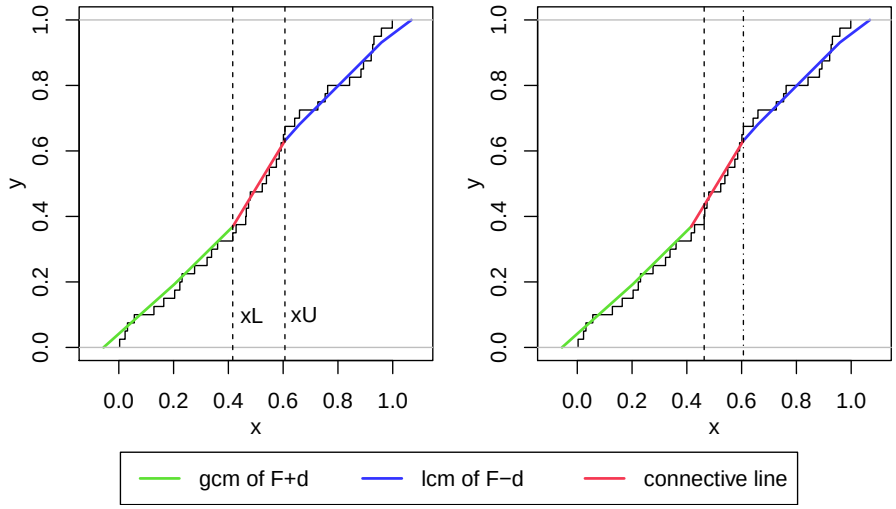

Figure 5.8: The fitted unimodal CDF of a particular dataset. Left: the dashed lines separate the different sections of the unimodal CDF estimate and thus enclose the interval (x_L, x_U). Right: the largest maximum difference is obtained at the dashed line. The largest maximum difference outside the interval (x_L, x_U) is obtained at the dash-dotted line.

5.4.2 Case study of a suboptimal fit

Since the original theorem permits a discontinuity, G^* is not necessarily the best fitting unimodal distribution with regard of the KS statistic.

However, the string test does not always fit the best-fitting continuous CDF as well, even though this is suggested by the theorem. We now show an example where this is the case.

Consider the ECDF of a particular dataset along with the fitted unimodal CDF as described by the string test. The result is plotted in Figure 5.8.

Let F be the ECDF of the data and let F_u denote the fitted unimodal distribution. We expect the maximum difference between F and F_u inside the interval (x_L, x_U) to be smaller than the maximum difference outside of the interval. Furthermore, we expect the maximum difference to be equal to the dip statistic. However, in this case, this is not true. The maximum difference between F and F_u is larger than the dip statistic and lies within (x_L, x_U).

One can expect this is caused by the test forcing a continuous unimodal CDF. When we permit a discontinuity at either x_L or x_U, the best-fitting unimodal distribution then has a discontinuity at x_U. However, the maximum difference between F and F_u is still larger than then dip statistic.

When we permit a discontinuity m everywhere in the interval (x_L, x_U), but require the slope on (x_L, m) and (m, x_U) to be the same (as to get a constant slope between (x_L, x_U) as required by the theorem) the fitted unimodal distribution is still not optimal; the maximum difference between F and F_u is larger than

the dip statistic.

Finally, when we permit a discontinuity m everywhere in the interval (x_L, x_U) and no requirements on the slope on (x_L, m) and (m, x_U), the fitted unimodal density is optimal; the maximum difference between F and F_u is equal to the dip statistic.

A striking result here is that we can in fact find a continuous unimodal CDF which is also optimal in terms of the KS statistic. Note that it is not striking because we found two optimal unimodal CDF's. The optimality as stated by the dip test does not include uniqueness. It is striking because there exists an optimal continuous unimodal CDF which is obtained by adjusting the interval (x_L, x_U) and then defining F_u as before, without discontinuities.

Note that the dip test was explained using a taut-string metaphor in Section 2.3.2. Examine the "string" fitted in the dip test in Figure 5.9. The string passes through the "solid" curves. When we adjust the interval a little, the string no longer passes through them.

When we examine G^* when using the new value for x_L, the maximum difference inside the interval (x_L, x_U) is smaller than outside the interval. Furthermore, the maximum difference is equal to the dip statistic. This result could suggest an algorithm that selects a valid x_L and x_U to be used in the theorem and should be investigated.

It is important to understand why an invalid x_L and x_U to be used in the theorem is selected by the dip test. The implementation could also calculate the KS statistic incorrectly. We now list a few reasons which could be causing the issue along with our findings.

- One might suspect that the implemented dip test did not select a value for x_L according to the algorithm in the paper. However, we implemented the dip test ourselves as well and both the public implementation and our own gave the same values for the modal interval.
- The function which calculates the KS statistic of the curves is not calculating the values correctly. However, we have checked this function using public implementation as well. Furthermore, the values are easily verifiable by hand when the function values are known.
- The fitted unimodal CDF plotted in the figures does not reflect the true function called when we calculate the KS statistic. This is also verified by plotting the CDF using true numerical results of the function.

We can conclude the following. The algorithm of the dip test selects the modal interval (x_L, x_U) such that it can find the minimum value for the KS statistic which can be attained when fitting a unimodal CDF. However, the modal interval does not necessarily represent the correct interval for the theorem employed to guarantee an optimal unimodal CDF.

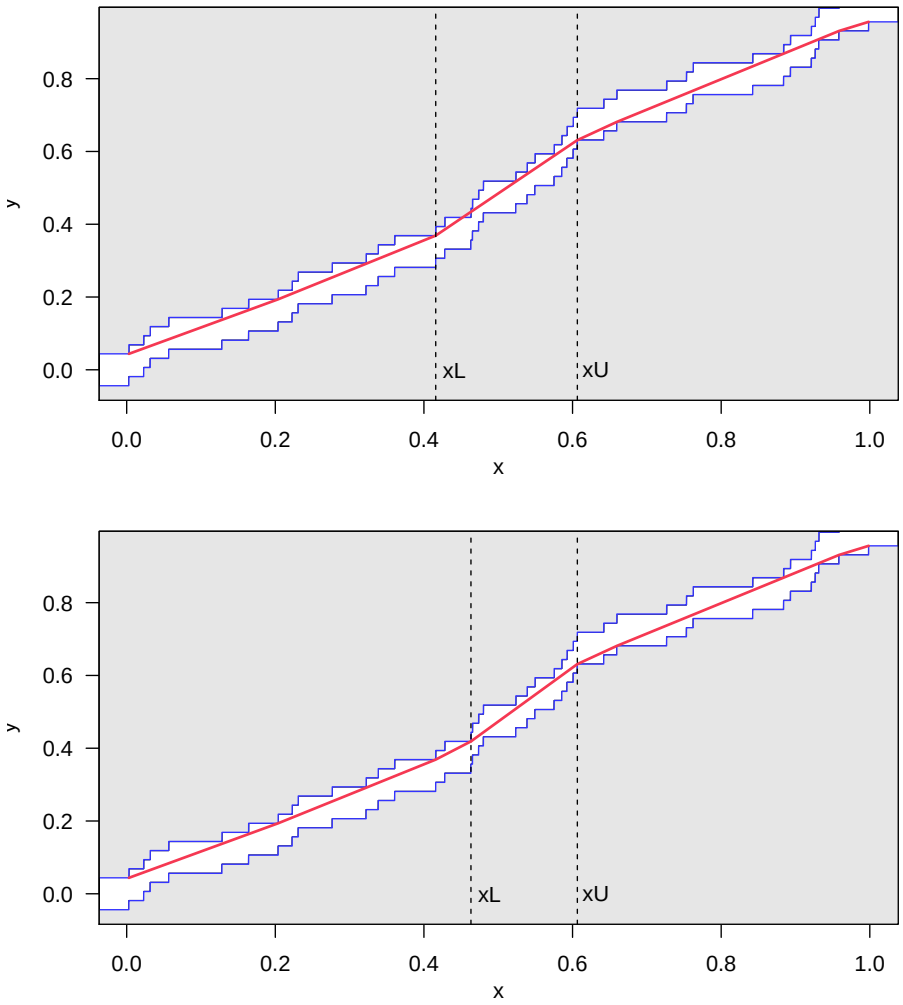

Figure 5.9: Graphical examples of the taut string metaphor used in describing the algorithm for the dip test applied to the data considered. The red line is the string, the bottom blue line is $F - d$ and the upper blue line is $F + d$. Note that the string forms the gcm on (x_1, x_L) for $F - d$, and the lcm on (x_U, x_n). In the plot, we set d equal to the dip statistic. Note that in the top figure, the string "passes through" the "solid" curve $F + d$ at around $x = 0.45$. If we shift x_L by 2 datapoints to the right, the figure looks like the one on the bottom. The string is not passing through any "solid" curves there.

Chapter 6

Testing multimodality in the presence of covariates

This chapter touches upon the issue of explaining multimodality with covariates. Especially the LOG test introduced in Chapter 4 might have some useful future applications in this regard, since the non-parametric MLE for logconcave densities exists. We now propose a way in which this test can be used in a regression setting. First, we introduce the regression setting and mathematically formulate the problem we aim to solve. We will also show a different formulation, which seems attractive but does not result in a well-defined problem and propose an alternative. Finally, we devise possible algorithms that can be used to solve the problem.

6.1 Introduction of the regression problem

We first consider the following simple regression setting. Suppose given data is a realization of $(Y_1, x_1), \ldots, (Y_n, x_n)$, where the Y_i's are response variables and x_i's are covariates. We assume the response variables Y_i to be independent samples from a distribution with density:

$$f\big(y - (\beta_0 + \beta_1 x_i)\big).$$

We are trying to find out if there are sources of multimodality that cannot be explained by the covariates using this simple linear regression model. Note that the marginal distribution of the Y_i's must not necessarily be multimodal for there to be sources of multimodality.

The problem can then be interpreted in the following way. In this simple regression model, it holds that:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon.$$

where (for identifiability reasons) we assume the errors have zero mean. If the errors ϵ are distributed multimodally, then the covariates alone cannot explain the multimodality of $f\big(y - (\beta_0 + \beta_1 x_i)\big)$.

6.2 Approach by finding parameters to obtain unimodal errors

One way to approach the problem is as follows. We will test the following two competing hypotheses:

$$H_0 : \exists_{\beta_0,\beta_1,f} : f\big(y - (\beta_0 + \beta_1 x_i)\big) \text{ unimodal} \\ \text{vs. } H_1 : \forall_{\beta_0,\beta_1,f} : f\big(y - (\beta_0 + \beta_1 x_i)\big) \text{ multimodal.} \quad (6.1)$$

The reasoning behind these hypotheses is as follows. We are trying to find out if there are sources of multimodality that cannot be explained by the covariates. To that end, we want to find if there is a set of parameters β_0, β_1 and density function f such that the covariates are the only source of multimodality. According to our interpretation, this means we try to find parameters such that the errors are distributed unimodally. So, if there is a set of parameters such that the error distribution $f\big(y - (\beta_0 + \beta_1 x_i)\big)$ is unimodal, we could argue that we have found a set of parameters such that the covariates explain the multimodality.

However, this formulation is not very sensible. The problem is illustrated below.

Suppose that $Y_i = \beta_0 + \beta_1 x_i + \epsilon$, where ϵ is distributed multimodally. Then $f\big(y - (\beta_0 + \beta_1 x_i)\big)$ is multimodal. Furthermore, suppose that the marginal distribution of the Y_i's, which is $f(y)$, is unimodal. We will now test the data for the two competing hypotheses.

Now, we can pick $\beta_0 = \beta_1 = 0$. This results in $f\big(y - (\beta_0 + \beta_1 x_i)\big) = f(y)$. Since the true density $f(y)$ is unimodal, if we would estimate f from the data with perfect accuracy, we have found a β_0, β_1, f such that $f\big(y - (\beta_0 + \beta_1 x_i)\big)$ is unimodal.

However, the choice of picking $\beta_0 = \beta_1 = 0$ is not based on fitting the data, but only based on the goal of finding parameters such that $f\big(y - (\beta_0 + \beta_1 x_i)\big)$ is unimodal. This means that the choice of the parameters is in general nonsensical. Moreover, according to this formulation of the problem, since we found a set of parameters such that $f\big(y - (\beta_0 + \beta_1 x_i)\big)$ is unimodal, we should now conclude that the covariates are the only source of multimodality. However, ϵ is distributed multimodally, so the covariates alone cannot explain all sources of multimodality.

Such an example, where the errors are distributed multimodally and the marginal distribution of the response variables is unimodal, is now given. Let $Y_i = \beta_0 + \beta_1 x_i + \epsilon$, where $\beta_0 = 0$, $\beta_1 = \frac{7}{10}$. Let $x = (x_1, \ldots, x_n)$ be a random sample from a discrete distribution, where $\mathbb{P}(x_i = 1) = \mathbb{P}(x_i = -1) = \frac{1}{2}$. Let ϵ be a random sample from the mixture normal distribution $\frac{1}{2}\mathcal{N}(-1, \frac{7}{10}) + \frac{1}{2}\mathcal{N}(1, \frac{7}{10})$. The density functions are plotted in Figure 6.1. As can be seen, the errors are distributed multimodally, while Y is distributed unimodally.

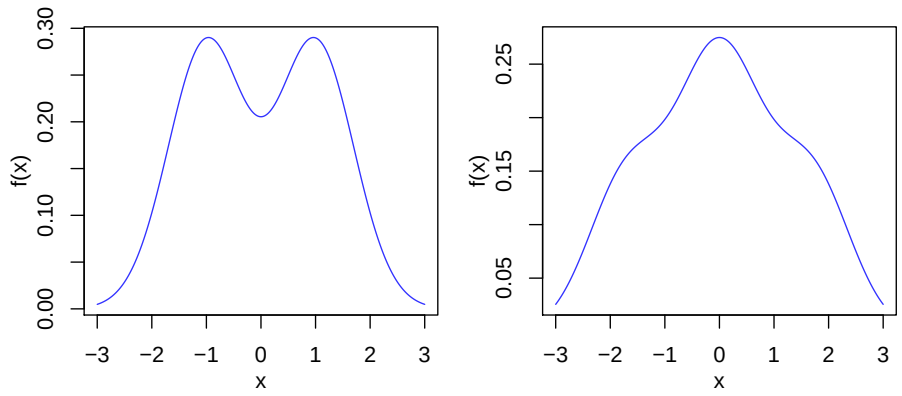

Figure 6.1: Plots of the PDF of the example where the errors are distributed multimodally and the marginal distribution of the response variables is unimodal. Left: density function of the errors ϵ. Right: Density function of the response variables Y.

6.3 Approach by using plugged-in parameters in multimodality tests

We now propose a different approach that will use plugged-in parameters obtained from the least-squares estimates.

First, using least-squares estimates, we will estimate the parameters β_0, β_1 in the simple linear regression setting. We then consider the residuals $e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$. The residuals are then tested for multimodality using an appropriate test.

What exactly is being tested is not clear. However, this approach could be calibrated well using the bootstrap approach outlined in Section 3.1.

6.4 Description of possible algorithms

We will now describe two possible algorithms which can be used for the regression problem.

First, we outline an algorithm that can be used in the approach formulated in Section 6.3 by using the bootstrap approach outlined in Section 3.1. However, instead of a regular bootstrap sample from the fitted distribution G, we must construct a regression sample. We obtain this sample in the following way.

Let $(Y_1, x_1), \ldots, (Y_n, x_n)$ be the data for which we aim to find sources of multimodality, where Y_i is a response variable and x_i is a covariate. Assume a simple linear regression model.

Let $\hat{\beta}_0, \hat{\beta}_1$ be the least squares-estimates obtained from the original dataset. Let e be the residuals obtained with the least-squares parameter estimates:

$e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. Now fit a distribution $\hat{G}$ to the residuals e. Let e^* be a bootstrap sample from the fitted distribution $\hat{G}$. Now a regression bootstrap sample of the data Y is given by:

$$Y^* = \hat{\beta}_0 + \hat{\beta}_1 x + e^*.$$

We can use this regression sampling in the bootstrap approach outlined in Section 3.1.

Another possibility would be to take an iterative approach. We will first find the least-squares estimate for the regression parameters and compute the residuals. We then fit a distribution to the residuals. We then use the fitted distribution to find MLE estimates for the regression parameters. Note that we use logconcave estimates for the density, so the MLE in step 5 is well-defined.

The steps can be done repeatedly if desired. We can thus outline the algorithm with the following steps:

1. Let $(Y_1, x_1), \ldots, (Y_n, x_n)$ be the data for which we aim to find sources of multimodality, where Y_i is a response variable and x_i is a covariate. Assume a simple linear regression model.
2. Use the least-squares method to compute estimates for β_0, β_1.
3. Given estimates $\hat{\beta}_0$, $\hat{\beta}_1$, compute the residuals $e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_0 x_i)$.
4. Compute the logconcave density estimate $\hat{f}_{lc}$ for the residuals e_i.
5. Update $(\hat{\beta}_0, \hat{\beta}_1) = \text{argmax}_{\beta_0, \beta_1} \prod_{i=1}^n \hat{f}_{lc}(y_i - (\beta_0 + \beta_0 x_i))$.
6. If desired, the procedure can now be restarted at step 3.

We have not studied this procedure in great detail, so firm conclusions on the validity of the approach cannot be drawn yet. Note that since the procedure is iterative, we should be careful the procedure converges. From preliminary testing, the procedure seems stable. However, the objective function in step 5 can be very irregular, in which case the global maximum can be hard to find. When suboptimal, local maxima are used instead, this could cause the procedure to diverge.

Chapter 7

Conclusion

Our goal was to devise tests for multimodality which were both well-calibrated and powerful. Furthermore, we aimed to develop a test that could be of value when testing for multimodality in the presence of covariates. We now conclude our findings of both tests. Our first devised test, the LOG test for multimodality, can be a very powerful tool for detecting multimodality. We have found the test to be somewhat anti-conservative, but reasonably well-calibrated. It has an inherent issue however; when the true density is non-logconcave and unimodal, the test can reject the null based on the non-logconcavity. The censoring approach which we proposed to mitigate this problem has several issues.

We conclude that the LOG test shows great promise as a test for multimodality, but must be used with great care. Censoring might mitigate some problems, but the censoring approach we have proposed does not seem adequate in solving the issues when the test is faced with a heavy-tailed distribution. The LOG test does not face any of these issues when used as a test for logconcavity in its own right, and can thus be valuable in these settings where a logconcave density is an important feature of the density. For instance, the test can be of great value when testing for multimodality in the presence of covariates, since the nonparametric MLE of a logconcave density exists. We have proposed an approach in this setting along with two algorithms. Further research needs to be done before conclusions on the validity of this approach can be drawn.

Our second devised test, the string test, did not face the same issues the LOG test did. The test can be of great value and has some appealing properties. Firstly, the interpretation of the test is quite intuitive. Secondly, the mechanics of the test are easily understood and easy to implement. Furthermore, the computations associated with the test are also efficient. Finally, the assessments of power and calibration show the test performs well and is superior to the original dip test.

However, the original dip test for multimodality has been improved significantly after its first proposal. The comparisons of the calibrated dip test by [2] and our string test seem to indicate our test is more conservative and less powerful than the test by [2]. However, our comparisons must be taken with a heavy grain of salt, as the distributions used by [2] were not always specified and we thus

estimated them purely by considering figures given in that paper. Furthermore, power results were not given numerically, so these were estimated from figures as well. It is therefore necessary to conduct an additional study where both tests are implemented before a fair comparison can be given.

The string test can be improved in two regards. First, the string test does not always fit a best-fitting CDF in terms of the KS statistic. The test might be improved when this issue is fixed. In the case study we conducted we have indicated a direction which could be taken to solve the issue. The issue is related to the insight we describe below. Secondly, our numerical results show the string test performs best when the AD statistic is used. We suspect the string test to perform better when we do not only measure the discrepancy between the ECDF and the fitted unimodal CDF with the AD statistic, but also fit the unimodal CDF to minimize the AD statistic.

Finally, we have found the following important insight. The original dip test finds a modal interval (x_L, x_U). This interval is not necessarily an interval which is usable in theorem 6 in the dip test paper [11] which guarantees a unimodal CDF which minimizes the KS statistic. Because of this, the string test, which is based on this theorem, does not always fit the best-fitting unimodal CDF in terms of the KS statistic.

Bibliography

[1] Lucien Birgé. Estimation of unimodal densities without smoothness assumptions. *Annals of Statistics*, 25(3):970–981, jun 1997.

[2] M Y Cheng and P Hall. Calibrating the excess mass and dip tests of modality. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 60(3):579–589, 1998.

[3] Ming-Yen Cheng and Peter Hall. Mode testing in difficult cases. *The Annals of Statistics*, 27(4):1294–1315, aug 1999.

[4] Lutz Dümbgen, Andre Huesler, and Kaspar Rufibach. Active set and EM algorithms for log-concave densities based on complete and censored data. *arXiv preprint arXiv:0707.4643*, (2006):1–19, 2007.

[5] Lutz Dümbgen and Kaspar RufiBach. logcondens: Computations Related to Univariate Log-Concave Density Estimation. *Journal of Statistical Software*, 39(6), 2011.

[6] J. B. S. Haldane. Simple Tests for Bimodality and Bitangentiality. *Annals of Eugenics*, 16(1):359–364, jan 1951.

[7] Peter Hall and Matthew York. On the calibration of silverman's test for multimodality. *Statistica Sinica*, 11(2):515–536, 2001.

[8] J A Hartigan. *Testing for antimodes*, pages 169–181. Springer Berlin Heidelberg, Berlin, Heidelberg, 2000.

[9] J. A. Hartigan and Surya Mohanty. The runt test for multimodality. *Journal of Classification: Classification Literature Automatic Search Service / plus CLASS*, 9(1):63–70, 1992.

[10] Hajo Holzmann and Sebastian Vollmer. A likelihood ratio test for bimodality in two-component mixtures with application to regional income distribution in the EU. *AStA Advances in Statistical Analysis*, 92(1):57–69, 2008.

[11] P M Hartigan J. A. Hartigan. The Dip Test of Unimodality. *The Annals of Statistics*, 13(1):70–84, 1985.

[12] Bernd Klaus and Korbinian Strimmer. Estimation of (Local) False Discovery Rates and Higher Criticism, 2015.

[13] Andreas Krause and V Liebscher. Multimodal projection pursuit using the dip statistic. *Preprint-Reihe Mathematik*, (13):1–25, 2005.

[14] Ronald P. Larkin. An algorithm for assessing bimodality vs. unimodality in a univariate distribution. *Behavior Research Methods & Instrumentation*, 11(4):467–468, jul 1979.

[15] Martin Maechler. Hartigan's dip test statistic for unimodality - corrected code, 2011.

[16] Michael C. Minnotte. Nonparametric testing of the existence of modes. *Annals of Statistics*, 25(4):1646–1660, 1997.

[17] D W Muller and G Sawitzki. Excess Mass Estimates and Tests for Multimodality. *Journal of the American Statistical Association*, 86(415):738–746, 1991.

[18] Gregory Paul M Rozál and J. A. Hartigan. The MAP test for multimodality. *Journal of Classification*, 11(1):5–36, 1994.

[19] G Sawitzki. The excess mass approach and the analysis of multi-modality. In *From Data to Knowledge*, pages 1–9. Springer, 1996.

[20] M F Schilling, A E Watkins, and W Watkins. Is human height bimodal? *American Statistician*, 56(3):223–229, 2002.

[21] B W Silverman. Density estimation for univariate and bivariate data. *Interpreting multivariate data*, pages 37–53, 1981.

[22] B W Silverman. Using Kernel Density Estimates to Investigate Multimodality. *Journal of the Royal Statistical Society. Series B (Methodological)*, 43(1):97–99, 1981.

[23] Winfried Stute, Wenceslao Gonzeles Manteiga, and Manuel Presedo Quindimil. Bootstrap based goodness-of-fit-tests. *Metrika: International Journal for Theoretical and Applied Statistics*, 40(1):243–256, 1993.

[24] Mutsunori Tokeshi. Dynamics of distribution in animal communities: Theory and analysis. *Researches on Population Ecology*, 34(2):249–273, 1992.

Appendix A

Pseudocode

This appendix contains pseudocode for the tests developed in the main thesis.

A.1 Pseudocode LOG test

This section contains the pseudocode for the LOG test. In our R implementation, the computation of the estimated logconcave CDF F_{lc} is done through the *logcondens* package [5].

The test statistics implemented are the KS, CvM and AD statistic. For each of these statistics, the simplified expression was used, since the estimated CDF is continuous on the domain $[x_{(1)}, x_{(n)}]$. The simplified expressions are given in Section 3.2.

Algorithm 1: Pseudocode LOG test

Input Data X of length n, choice of number of bootstrap samples B, choice of test statistic D.
Sort the data X.
Compute the ECDF F of the data X.
Compute the estimated logconcave CDF F_{lc} of the data X.
if $D \neq D_{AD}$ **then**
    Compute D using F and F_{lc}.
else
    Compute D using F and F_{lc} without the values $\min\{X\}$ and $\max\{X\}$[a].
Generate $B \cdot n$ samples from F_{lc} in vector S.
Divide S in B vectors of length n. Denote the new vectors by S_i
for $i = 1$ **to** B **do**
    Compute the ECDF F_i^* of S_i.
    Compute the estimated logconcave CDF $F_{lc,i}^*$ of S_i.
    if $D \neq D_{AD}$ **then**
        Compute D_i using F_i^* and $F_{lc,i}^*$.
    else
        Compute D_i using F_i^* and $F_{lc,i}^*$ without the values $\min\{S_i\}$ and $\max\{S_i\}$.
Compute $p = \frac{\#\{i:D_i \geq D\}}{B}$.
Return: p

[a]Note that AD statistic is not well-defined if $F_{lc}(\min\{X\}) = 0$ and $F_{lc}(\max\{X\}) = 1$. Refer to Section 3.2 for an explanation.

A.2 Pseudocode string test

This section contains the pseudocode for the string test. In our R implementation, the computation of the interval (x_L, x_U) is done through the *diptest* package [15]. The gcm and lcm are computed using the package *fdrtool* [12].

The test statistics implemented are the KS, CvM and AD statistic. For each of these statistics, the simplified expression was used, since the estimated CDF is continuous as proven in Section 5.4.1. The simplified expressions are given in Section 3.2.

Algorithm 2: Pseudocode string test

Input Data X of length n, choice of number of bootstrap samples B, choice of test statistic D.
Sort the data X.
Compute the ECDF F of the data X.
Compute x_L, x_U and the dip statistic d.
Compute the gcm of $F + d$ on (x_1, x_L) and the lcm of $F - d$ on (x_U, x_n).
Compute the slope between x_1 and x_2 and find the intersection with 0.
Compute the slope between x_{n-1} and x_n and find the intersection with 1.
Compute G^*.
Compute D using F and G^*.
Generate $B \cdot n$ samples from G^* in vector S.
Divide S in B vectors of length n. Denote the new vectors by S_i
for $i = 1$ **to** B **do**
    Compute the ECDF F_i of S_i.
    Compute the estimated unimodal CDF G_i^* of S_i, in the same fashion as the original sample
    Compute D_i using F_i and G_i^*.
Compute $p = \frac{\#\{i : D_i \geq D\}}{B}$.
Return: p

Appendix B

Proof non-existence of non-parametric MLE unimodal density

This appendix provides a rigorous proof of the non-existence of the non-parametric MLE of a unimodal density. We will prove the MLE is not well-defined by proposing an example unimodal density where we accumulate all probability mass to a single datapoint. Note that our example unimodal density is discontinuous, but the MLE is also not well-defined when we restrict to only continuous densities, noted for example in [1].

Suppose that we are given data $X = (x_1, \ldots, x_n)$ where all x_i's are i.i.d. Let $\mathscr{U}$ be the class of all unimodal density functions. We will now try to find the underlying density function using the non-parametric MLE where we only restrict the density function to be in the class of unimodal densities. That is, we try to find:

$$\arg\max_{f \in \mathscr{U}} L(f).$$

Since we would like to use this estimator later, we must make sure that this is well-defined. We will show that the class $\mathscr{U}$ is too large to use this estimator to find a sensible density function.

Examine the following function $\hat{f}_a^*(x)$:

$$\hat{f}_a^*(x) = \begin{cases} \frac{1}{2}(2x_{max} - \frac{1}{2^a})^{-1} & -x_{max} \leq x \leq \frac{-1}{2^{a+1}} \\ 2^{a-1} & \frac{-1}{2^{a+1}} < x < \frac{1}{2^{a+1}} \\ \frac{1}{2}(2x_{max} - \frac{1}{2^a})^{-1} & \frac{1}{2^{a+1}} \leq x \leq x_{max} \\ 0 & \text{otherwise} \end{cases},$$

where $x_{max} = \max_{i \in 1,\ldots,n} \{|x_i|\}$ (assume $n \geq 2$). We will show that this is a valid

density function. First we show that $\int_{-\infty}^{\infty} \hat{f}_a^*(x)dx = 1$. Note that:

$$\begin{aligned}
\int_{-\infty}^{\infty} \hat{f}_a^*(x)dx &= \int_{-x_{max}}^{\frac{-1}{2^{a+1}}} \hat{f}_a^*(x)dx + \int_{\frac{-1}{2^{a+1}}}^{\frac{1}{2^{a+1}}} \hat{f}_a^*(x)dx + \int_{\frac{1}{2^{a+1}}}^{x_{max}} \hat{f}_a^*(x)dx \\
&= \frac{1}{2}(2x_{max} - \frac{1}{2^a})^{-1}((\frac{-1}{2^{a+1}}) - -x_{max}) + \frac{2}{2^{a+1}} \cdot 2^{a-1} + \\
&\quad \frac{1}{2}(2x_{max} - \frac{1}{2^a})^{-1}(x_{max} - (\frac{1}{2^{a+1}})) \\
&= \frac{1}{2}(2x_{max} - \frac{1}{2^a})^{-1}(2x_{max} - 2 \cdot \frac{1}{2^{a+1}}) + \frac{1}{2} \\
&= \frac{1}{2} + \frac{1}{2} \\
&= 1.
\end{aligned}$$

We now show that $\hat{f}_a^*(x) \geq 0$. Note that $2^{a-1} \geq 0$ for every value of a. Furthermore, note that:

$$\begin{aligned}
&\frac{1}{2}(2x_{max} - \frac{1}{2^a})^{-1} \geq 0 \\
\Rightarrow\ & 2x_{max} - \frac{1}{2^a} \geq 0 \\
\Rightarrow\ & a \geq -\frac{\log(2x_{max})}{\log(2)}.
\end{aligned}$$

So $\hat{f}_a^*(x) \geq 0$ for a sufficiently large. We will now show that the density function is unimodal. Note that:

$$\begin{aligned}
&\frac{1}{2}(2x_{max} - \frac{1}{2^a})^{-1} \leq 2^{a-1} \\
\Rightarrow\ & (2x_{max} - \frac{1}{2^a})^{-1} \leq 2^a \\
\Rightarrow\ & 1 \leq 2^{a+1}x_{max} - 1 \\
\Rightarrow\ & \frac{\log(\frac{1}{x_{max}})}{\log(2)} \leq a.
\end{aligned}$$

So for a sufficiently large $\hat{f}_a^*(x) \in \mathscr{U}$.

Now examine the likelihood function for $\hat{f}_a^*(x)$. Suppose that, without loss of generality $0 \in \{x_1, \ldots, x_n\}$. If this would not be the case, then we can translate $\hat{f}^*(x)$ to $\hat{f}^*(x - x_{i^*})$ where $x_{i^*} \in \{x_1, \ldots, x_n\}$. Suppose that a is sufficiently

large such that there is exactly one datapoint $x_{i^*} \in (\frac{-1}{2^{a+1}}, \frac{1}{2^{a+1}})$:

$$\begin{aligned}
L(\hat{f}_a^*(x)) &= \prod_{i=1}^{n} \hat{f}_a^*(x_i) \\
&= 2^{a-1} \prod_{i=1, i\neq i^*}^{n} \frac{1}{2}(2x_{max} - \frac{1}{2^a})^{-1} \\
&= 2^{a-1} \left(\frac{1}{2}(2x_{max} - \frac{1}{2^a})^{-1} \right)^{n-1} \\
&= 2^{a-1} 2^{-(n-1)} (2x_{max} - \frac{1}{2^a})^{-(n-1)} \\
&= 2^{a-n} (2x_{max} - \frac{1}{2^a})^{1-n} \xrightarrow{a\to\infty} \infty.
\end{aligned}$$

So the likelihood for $\hat{f}_a^*(x)$ can be arbitrarily large by increasing a. This means that the class of unimodal distributions $\mathscr{U}$ is too large to compute a sensible density function with the non-parametric MLE.

Appendix C

Full results

In this chapter, we provide the numerical results obtained in the assessment of calibration and power of both the LOG and the string test in tabulated form.

C.1 Tabulated results of the LOG test

We now provide tables with results from the calibration and power assessment of the LOG test.

α	LOG.KS	LOG.CvM	LOG.AD
0.01	0.00	0.00	0.00
0.05	0.02	0.02	0.02
0.10	0.06	0.04	0.04
0.20	0.12	0.12	0.12

Table C.1: Target significance levels and the actual significance levels of the LOG test when applied to a standard normal distribution.

α	LOG.KS	LOG.CvM	LOG.AD
0.01	0.00	0.00	0.01
0.05	0.03	0.02	0.05
0.10	0.06	0.05	0.11
0.20	0.14	0.13	0.24

Table C.2: Target significance levels and the actual significance levels of the LOG test when applied to a Student's t distribution.

α	LOG.KS	LOG.CvM	LOG.AD
0.01	0.01	0.01	0.01
0.05	0.07	0.06	0.06
0.10	0.13	0.13	0.12
0.20	0.25	0.27	0.27

Table C.3: Target significance levels and the actual significance levels of the LOG test when applied to a uniform distribution.

α	LOG.KS	LOG.CvM	LOG.AD
0.01	0.01	0.02	0.03
0.05	0.09	0.10	0.12
0.10	0.16	0.18	0.20
0.20	0.31	0.32	0.35

Table C.4: Target significance levels and the actual significance levels of the LOG test when applied to a Laplace distribution.

α	LOG.KS	LOG.CvM	LOG.AD
0.01	0.01	0.02	0.01
0.05	0.09	0.07	0.06
0.10	0.16	0.14	0.12
0.20	0.26	0.26	0.23

Table C.5: Target significance levels and the actual significance levels of the LOG test when applied to a shoulder distribution.

μ	DIP	LOG.KS	LOG.CvM	LOG.AD
1.50	0.00	0.04	0.02	0.04
1.90	0.00	0.06	0.02	0.00
2.30	0.00	0.02	0.05	0.04
2.70	0.00	0.10	0.10	0.05
3.10	0.00	0.17	0.17	0.17
3.50	0.01	0.38	0.42	0.29
3.90	0.07	0.74	0.81	0.79
4.30	0.16	0.89	0.96	0.94
4.70	0.34	0.97	0.97	0.97

Table C.6: Power results of the LOG test for mixture distribution (4.1) for different values of μ. The values shown are the fraction of times the null hypothesis is rejected under significance level $\alpha = 0.05$.

σ	DIP	LOG.KS	LOG.CvM	LOG.AD
4.00	0.14	1.00	1.00	1.00
5.00	0.02	0.98	1.00	0.98
6.00	0.02	1.00	1.00	1.00
7.00	0.00	1.00	1.00	1.00
8.00	0.00	1.00	1.00	1.00
9.00	0.00	1.00	1.00	1.00
10.00	0.00	1.00	1.00	1.00
11.00	0.00	1.00	1.00	1.00
12.00	0.00	1.00	1.00	1.00
13.00	0.00	1.00	1.00	1.00
14.00	0.00	1.00	1.00	1.00
15.00	0.00	1.00	1.00	1.00
16.00	0.00	1.00	1.00	1.00
17.00	0.00	1.00	1.00	1.00

Table C.7: Power results of the LOG test for mixture distribution (4.2) for different values of σ. The values shown are the fraction of times the null hypothesis is rejected under significance level $\alpha = 0.05$.

C.2 Tabulated results of the string test

We now provide tables with results from the calibration and power assessment of the string test.

α	GCM.KS	GCM.CvM	GCM.AD
0.01	0.00	0.01	0.00
0.05	0.02	0.03	0.04
0.10	0.06	0.07	0.07
0.20	0.14	0.14	0.17

Table C.8: Target significance levels and the actual significance levels of the string test when applied to a standard normal distribution

α	GCM.KS	GCM.CvM	GCM.AD
0.01	0.00	0.01	0.01
0.05	0.03	0.02	0.02
0.10	0.04	0.05	0.06
0.20	0.15	0.15	0.17

Table C.9: Target significance levels and the actual significance levels of the string test when applied to a Student's t distribution

α	GCM.KS	GCM.CvM	GCM.AD
0.01	0.01	0.01	0.01
0.05	0.06	0.05	0.05
0.10	0.13	0.11	0.10
0.20	0.22	0.21	0.18

Table C.10: Target significance levels and the actual significance levels of the string test when applied to a uniform distribution

α	GCM.KS	GCM.CvM	GCM.AD
0.01	0.01	0.00	0.00
0.05	0.03	0.03	0.03
0.10	0.06	0.05	0.07
0.20	0.13	0.12	0.13

Table C.11: Target significance levels and the actual significance levels of the string test when applied to a Laplace distribution

α	GCM.KS	GCM.CvM	GCM.AD
0.01	0.01	0.01	0.01
0.05	0.05	0.03	0.03
0.10	0.09	0.09	0.10
0.20	0.19	0.19	0.20

Table C.12: Target significance levels and the actual significance levels of the string test when applied to a shoulder distribution

μ	DIP	String.KS	String.CvM	String.AD
1.50	0.00	0.05	0.04	0.04
1.90	0.00	0.03	0.06	0.06
2.30	0.01	0.02	0.01	0.01
2.70	0.00	0.07	0.03	0.03
3.10	0.01	0.10	0.03	0.02
3.50	0.00	0.12	0.09	0.10
3.90	0.09	0.28	0.28	0.30
4.30	0.20	0.44	0.45	0.43
4.70	0.47	0.60	0.73	0.72

Table C.13: Power results of the string test for mixture distribution (4.1) for different values of μ. The values shown are the fraction of times the null hypothesis is rejected under significance level $\alpha = 0.05$.

σ	DIP	String.KS	String.CvM	String.AD
4.00	0.14	0.24	0.34	0.38
5.00	0.06	0.10	0.14	0.14
6.00	0.00	0.08	0.18	0.18
7.00	0.00	0.02	0.04	0.06
8.00	0.02	0.06	0.04	0.04
9.00	0.00	0.02	0.04	0.04
10.00	0.00	0.02	0.02	0.04
11.00	0.00	0.04	0.08	0.06
12.00	0.00	0.06	0.02	0.02
13.00	0.00	0.02	0.04	0.04
14.00	0.00	0.06	0.02	0.04
15.00	0.00	0.02	0.00	0.00
16.00	0.00	0.02	0.04	0.04
17.00	0.00	0.00	0.00	0.00

Table C.14: Power results of the string test for mixture distribution (4.2) for different values of σ. The values shown are the fraction of times the null hypothesis is rejected under significance level $\alpha = 0.05$.