BACHELOR

A Cramér-von-Mises Based Dip Test for Multimodality

Reinhoudt, Sjoerd

Award date:
2022

Department of Mathematics and Computer Science

A Cramér-von-Mises Based Dip Test for Multimodality

Bachelor End Project Report

Sjoerd Pieter Reinhoudt

Supervisors:

dr. Rui M. Castro

Ivo V. Stoepker, MSc

29-06-2022

Abstract

In this thesis, we explore two existing families of statistical tests that test for multimodality based on a finite i.i.d. data sample, and attempt to improve the calibration and power of one such test, called the dip test. This test is known to be very conservative. The dip test assesses modality by determining the distribution unimodal distribution closest to the ECDF to the data sample with respect to the Kolmogorov-Smirnov statistic. We propose to modify the test to minimise the Cramér-von-Mises statistic instead, since this is a global statistic, compared to the very local nature of the Kolmogorov-Smirnov statistic. It is expected that this results in an improved test, since the Cramér-von-Mises statistic takes into account several points of the respective distributions, and can therefore more accurately quantify similarity and discrepancy. This introduces a non-trivial optimisation problem, and we provide a method to numerically solve this problem. For this, we use an accelerated gradient descent procedure. Numerical evidence is then used to assess the power and calibration of the modified test. From this, we find that the new methodology is less well calibrated than the existing test that most closely resembles the new methodology. In particular, the new methodology appears to be rather conservative, especially for poorly behaved, heavy tailed distributions like the Cauchy distribution. Finally, from numerical experimentation on a Gaussian mixture distribution, the new methodology appears to be stronger in power than the original formulation of the dip test, though weaker than other existing tests.

Contents

1 Introduction **3**

2 Existing Methodology **4**
2.1 Silverman's test . . . 5
2.1.1 Testing procedure . . . 5
2.1.2 Advantages and disadvantages . . . 7
2.2 Dip test . . . 7
2.2.1 Testing procedure . . . 8
2.2.2 Advantages and disadvantages . . . 9

3 Improving the power of the dip test **10**
3.1 Optimising the CvM and AD statistics . . . 10
3.2 Simplification in a continuous setting . . . 10
3.2.1 Piecewise-linearity of optimal solution . . . 12
3.2.2 Simplifying the constraint set . . . 15
3.2.3 The structure of the solution . . . 15
3.3 Numerical optimisation . . . 17
3.3.1 Reparametrizing the problem . . . 17
3.3.2 Projected Gradient Descent . . . 18
3.3.3 Evaluating the prox operator . . . 21
3.3.4 Reducing the required number of steps at runtime . . . 24

4 Numerical experimentation on the new methodology **26**
4.1 The calibration of the test . . . 26
4.2 The power of the test . . . 27
4.3 Performance of the numerical optimisation . . . 31

5 Conclusion **35**

Appendices **37**

A Math appendix **37**
A.1 Discrete form of CvM statistic . . . 37
A.2 Extendability of piecewise-linear solution . . . 37
A.3 Redundancy of the various constraints . . . 38
A.4 Convexity of the objective functions . . . 39
A.5 The norm of the transition matrix M . . . 39
A.6 Projection in case of non-positive inner products . . . 40

B Additional data and measurements **41**
B.1 The calibration of the test . . . 41
B.2 The power of the test . . . 42

1 Introduction

In mathematics, science, and engineering, one often encounters data that can be modelled as a sample from some unknown underlying distribution. From this limited data sample, we wish to learn about the various characteristics of the underlying distribution, which then gives information about the dynamics of the system that is being examined. One such characteristic is whether the underlying distribution is *unimodal* or *multimodal*: whether it has one or several maxima[1], where the probability density is (locally) maximal. Such maxima are called modes, and can manifest themselves in samples from the distribution in the form of clusters: intervals where a large number of data points are concentrated. This can often be observed when a histogram is drawn from the data, where a large concentration of data points appears as a hill near the mode.

In particular, a multimodal underlying distribution can imply that there are two or more largely separate groups present in the data, which behave differently from one another. For example, suppose one observes two modes in the distribution of the incubation times of a disease. This could point towards there being two different strains of this disease that behave differently, where on average, one strain has a larger incubation time compared to the other strain. It could also point towards a certain subgroup of patients being more resistant to the disease than others, and several other hypotheses could be formulated. Either way, the observation of multimodality leads to several novel research questions, which contributes to a greater understanding of the dynamics at play.

We can now define the problem more formally. Suppose we have measured N univariate data points, in the form of real numbers. The collection of these data points is called the dataset. We assume these data points are a realisation of the random variables $\{X_i\}_{i=1}^N$. For notational simplicity, we define $\mathbf{X} = (X_i)_{i=1}^N$ to be the random vector collecting these random variables. We assume these random variables are independent and identically distributed (i.i.d) according to some (cumulative) distribution function $F : \mathbb{R} \to [0, 1]$ where $F(x) = \mathbb{P}(X_i \leq x)$ for all $x \in \mathbb{R}$. We wish to test if F corresponds to a unimodal or multimodal distribution. We thus arrive at two rival hypotheses:

$$H_0 : F \text{ is unimodal.} \quad \text{vs.} \quad H_1 : F \text{ is multimodal.}$$

The goal of this project is to design a practical, powerful test to differentiate between these hypotheses based on the measured data points. We restrict ourselves to a continuous setting, meaning that F is a continuous function. We further assume the system that is examined is not sufficiently well understood to reasonably assume the distribution function has a known form, and thus resort to non-parametric statistics.

In Section 2, existing methodologies to test for multimodality are introduced. The advantages and disadvantages of the various tests are also discussed. For one such test, the so-called dip test, numerical evidence suggests it might be possible to improve its calibration and increase its power by changing the distance measure the test minimises. This introduces a non-trivial optimisation problem, which is discussed in Section 3. The statistical performance of the modified test is then examined in Section 4, where numerical evidence is presented to assess its power and calibration. In addition, this is compared to the power and calibration of the existing tests that were introduced in Section 2. The usefulness of the method as a practical test is also assessed in this section, by determining how long the optimisation procedure takes to perform.

[1] Multimodality can be defined more generally and exactly. This is done in the subsequent sections.

2 Existing Methodology

Before introducing existing methodology, it is helpful to begin by explaining why the "visual inspection method" is not always sufficient to test for multimodality. In everyday life, when one suspects a distribution might be multimodal, one might simply draw a histogram (or kernel density plot), and visually inspect the maxima of this histogram. If several disjoint intervals contain a large concentration of data points, one might conclude the distribution is multimodal. Conversely, if only one maximum is shown (or the histogram does not show clear maxima at all), one might conclude the distribution is unimodal. For instance, we can consider the kernel density plots of the eruption durations of the Old Faithful Geyser shown in Figure 1. In all these kernel density plots, we see more than a single maximum. We may therefore conclude that the eruption duration of Old Faithful comes from a multimodal distribution.

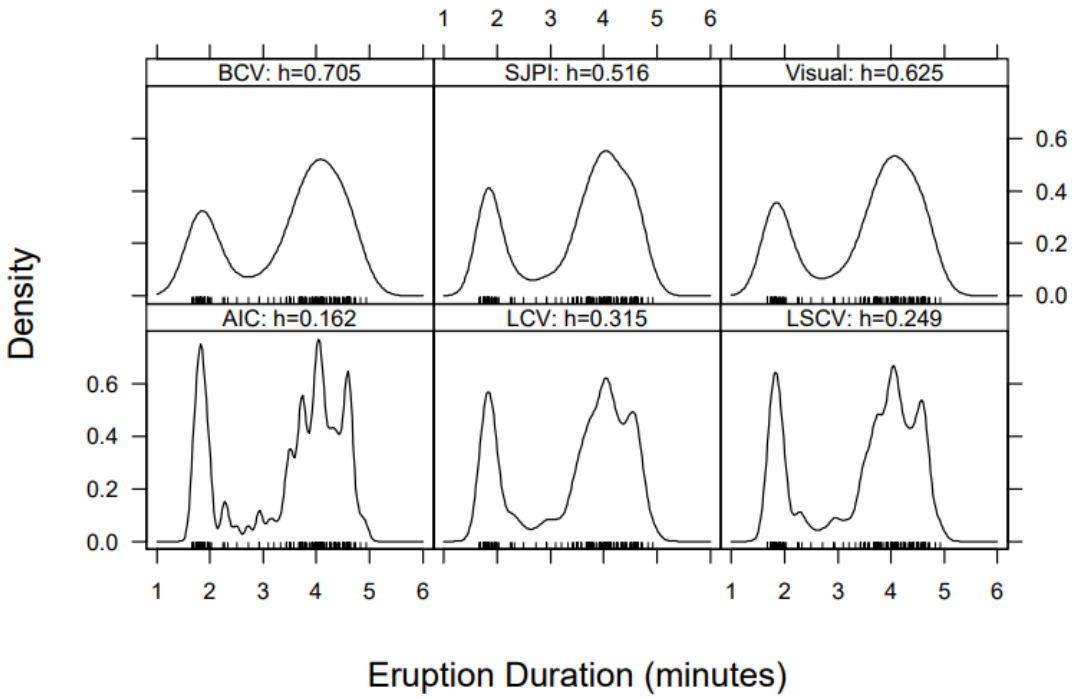


Figure 1: Various kernel density plots of the eruption duration of the Old Faithful Geyser in Yellowstone National Park, for various kernel widths h. In all cases a Gaussian kernel is used. Source: [1].

While this approach can certainly give correct results, it has several disadvantages. Firstly, one has to properly choose the bin size (or kernel width): If the bin size is chosen to be too large, several present modes can be combined into a seemingly single mode. Conversely, if the bin size is chosen to be too small, extraneous maxima that do not correspond to modes in the distribution can be introduced, that appear due to randomness in the finite sample. This is also visible in Figure 1. The smaller kernel widths in the bottom row show many maxima, whereas the larger kernel widths in the top row show only two maxima. From only looking at these plots, it is not obvious what kernel width should be preferred. While in the picture shown all kernel widths result in the conclusion that the distribution is multimodal, it is clear that the chosen kernel width has a large influence on the number of maxima in the density plot, and can therefore directly influence the conclusion following from this approach. Choosing a "correct" bin size is often a somewhat subjective matter. Secondly, even for a well-chosen bin size, it can still be subjective whether an observed maximum is an extraneous maximum or a mode that truly appears in the underlying distribution: a small hump in the histogram can represent a weak mode in the distribution, but can also be a point where several data points are concentrated purely by chance. Finally, conclusions drawn from this method are highly qualitative in nature, since no quantitative confidence level can be assigned. In many occasions, this makes it insufficiently reliable to use in science and industry.

For this reason, several more quantitative tests have been developed. In order to discuss these tests, we first have to make the definition of multimodality introduced in Section 1 more exact.

Definition 2.1 (Unimodality)**.** *A univariate distribution function $F : \mathbb{R} \to [0, 1]$ is defined to be unimodal if and only if there exists an $m \in \mathbb{R}$ such that F is convex on $(-\infty, m]$ and concave on $[m, \infty)$. The interval $(-\infty, m)$ is then called the convex part, and the interval (m, ∞) is called the concave part. A distribution function is defined to be multimodal if and only if it is not unimodal. A probability density function is defined to be unimodal if and only if its corresponding distribution function is unimodal.*

If F is a continuous unimodal distribution function, the value m manifests itself as the point where the slope of the distribution function is maximised, and is therefore the point of maximum probability density. This means that m is the mode of this unimodal distribution. This corresponds with the more intuitive definition given in Section 1.

Remark 2.2. *It should be noted that even for a unimodal distribution, the value of m might not be unique: there might be an interval on which F has its maximal slope. In this case, the corresponding density function does not have a single maximum, but rather a plateau, where the probability density is maximised. If a distribution has only one such plateau, it is still considered to be unimodal.*

If a unimodal F is not continuous, then it has exactly one point of discontinuity, namely at m. This can then be interpreted as a point where the probability density approaches ∞, and therefore also clearly corresponds to the mode of the distribution.

We now introduce two existing (families of) tests that test for multimodality. Note that alternative tests also exist, but these are by far the most commonly used tests for modality.

2.1 Silverman's test

Silverman's test, published in 1981 in [2], is an often-used test for multimodality. It can be interpreted as a more quantitative version of the visual inspection method. In full generality, it tests for slightly more general hypotheses than unimodality and multimodality: for a (fixed) value $k \in \mathbb{N}$, Silverman's test differentiates between the rival hypotheses

$$H_0 : F \text{ has at most } k \text{ modes.} \quad \text{vs.} \quad H_1 : F \text{ has more than } k \text{ modes.}$$

It is clear that for the value $k = 1$, Silverman's test reduces to a test for unimodality versus multimodality. The rest of this subsection assumes the value $k = 1$ is chosen. As mentioned before, if a kernel density plot has several maxima, increasing the kernel width can combine two of these maxima into a single maximum. As the kernel width is increased further, all maxima that were initially present will merge, and the kernel density plot will only have a single maximum. One can then consider the minimum kernel width that is required in order to reduce the number of maxima to 1. If the underlying distribution has several strong modes, we can expect this minimum kernel width to be very large. After all, a large amount of smoothing is necessary to combine two strong maxima into a single maximum. Conversely, if the underlying distribution only has a single mode, we may expect a small kernel density width to be sufficient to smooth out any extraneous maxima that may have appeared. This minimum kernel width can thus be used as a statistic to test for multimodality.

2.1.1 Testing procedure

Silverman's test begins by considering the kernel density function $\hat{f}_h(t)$. Here the parameter h denotes the kernel width, and t denotes the input variable. $\hat{f}_h$ is then defined by

$$\hat{f}_h(t) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{t - X_i}{h}\right), \tag{1}$$

where $K : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a kernel function. This means that K is symmetric and normalised. In other words, K satisfies the following conditions:

$$\forall_{x \in \mathbb{R}} : K(x) = K(-x), \text{ and}$$
$$\int_{-\infty}^{\infty} K(x)dx = 1.$$

In the original paper, K is taken to be the density function of the standard normal distribution. This has the advantage that for any dataset, the number of maxima of $\hat{f}_h$ is non-increasing as a function of h, as shown in [2]. For other kernels, this may not be the case. We can then define the *critical kernel width* h_{crit} to be

$$h_{crit} = \inf\{h \in \mathbb{R}^+ | \hat{f}_h(t) \text{ has a single maximum}\}.$$

Under the null hypothesis and some regularity assumptions, h_{crit} is of the order $N^{-\frac{1}{5}}$, as shown in [3]. This means that as $N \to \infty$, h_{crit} approaches 0 with probability 1. This is to be expected: after all, the empirical cumulative distribution function (ECDF) converges to the underlying (unimodal) distribution with probability 1 [4]. As $N \to \infty$, no smoothing will therefore be necessary.

The critical kernel density estimator $\hat{f}_{h_{crit}}$ can be interpreted as the unimodal density function that most closely resembles the measured data: After all, if h is chosen to be smaller than h_{crit}, $\hat{f}_h$ will not be unimodal, and if h is chosen to be smaller than h_{crit}, $\hat{f}_h$ will resemble the measured data less. This is shown in Figure 2.

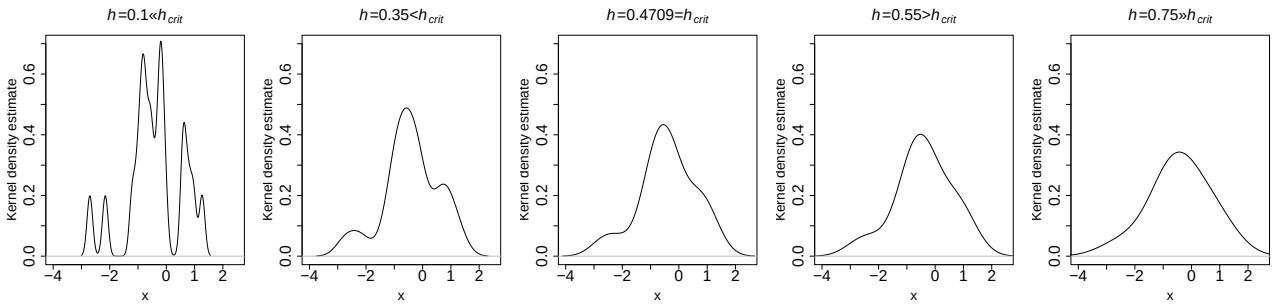


Figure 2: The kernel density plots of a sample of size 20 from a standard normal distribution for various kernel widths. Note the density estimates are multimodal for small kernel widths, and several small details that are visible in the plots for smaller kernel widths are lost when larger kernel widths are used.

The critical kernel width h_{crit} can be found efficiently using a binary search algorithm, where at each time-step the kernel width is lowered if the kernel density currently has a single maximum, and is increased if the kernel density currently has more than one maximum. By definition of the infimum, $\hat{f}_h$ has multiple maxima for all $h < h_{crit}$, and since the number of maxima is non-increasing as a function of h when a Gaussian kernel is used, $\hat{f}_h$ has only one maximum for $h > h_{crit}$. This means the binary search will always converge to h_{crit}.

To give a p-value, the test then uses a bootstrapping approach. The test begins by rescaling $\hat{f}_{h_{crit}}$ to have the same variance as the sample variance of the dataset. After this, n new samples of the same size N are drawn from this rescaled density function. For each of these samples, a kernel density function is then computed, with the same kernel function and the previously found window width h_{crit}. For each of these samples, the number of modes $\{j_i\}_{i=1}^n$ is counted. Finally, the test returns the p-value, given by

$$p = \frac{\sum_{i=1}^{N} \mathbb{1}\{j_i > 1\}}{n}.$$

This is based on the following idea: Let $h_{crit,i}$ be the h_{crit} value corresponding to bootstrap sample i. We wish to determine the fraction of bootstrap samples that require more smoothing than the original dataset, in other words $h_{crit,i} > h_{crit}$. However, calculating $h_{crit,i}$ for every bootstrap sample turns out to be somewhat computationally expensive. We can therefore test for an equivalent condition. As discussed before, the number of maxima of $\hat{f}_h$ is non-increasing as a function of h. Therefore, $h_{crit,i} > h_{crit}$ if and only if the kernel density of the bootstrap sample with kernel width h_{crit} has multiple maxima, in other words $j_i > 1$.

2.1.2 Advantages and disadvantages

Silverman's test has been found to be conservative, and this comes with a lack of power. As explained in Section 2.1.1, the test reports a p-value given by the fraction of bootstrap samples that require more smoothing than the original dataset. If the test is optimally calibrated, the reported p-value should be uniformly distributed between 0 and 1 under the null. However, even asymptotically, as the sample size N approaches ∞, the reported p-values are generally larger than would be expected from a uniform distribution. This is because under the null hypothesis, $h_{crit,i}$ is generally larger than h_{crit}, meaning that the bootstrap samples generally require more smoothing than the original data. Put more formally, the original formulation of Silverman's test rejects the null hypothesis at a significance level α if $\mathbb{P}(h_{crit,i} \leq h_{crit}|\mathbf{X}) \geq 1 - \alpha$, where this probability is estimated using the bootstrapping approach. Note that this outcome is still random, since it depends on the original data $\mathbf{X}$. Hall and York show in [5] that under the null hypothesis $\mathbb{P}(\mathbb{P}(h_{crit,i} \leq h_{crit}|\mathbf{X}) \leq 1-\alpha) < \alpha$. This means that the probability the test rejects the null hypothesis is less than α, when ideally they should be equal. Hall and York suggest to remedy this by instead choosing λ_α such that $\mathbb{P}(\mathbb{P}(h_{crit,i} \leq \lambda_\alpha h_{crit}|\mathbf{X}) \leq 1 - \alpha) = \alpha$. Generally, $\lambda_\alpha \neq 1$. Note that this value of λ_a depends on the distribution of $\mathbf{X}$, and therefore depends on both the sample size N and the underlying distribution F. Since F is unknown, the exact value of λ_α is unknown as well. Hall and York then show that under the null hypothesis, asymptotically, λ_α converges to a value only dependent on α, and independent of the underlying distribution, as $N \to \infty$. The finite-sample λ_α can then be approximated by this asymptotic result. They also note that this removes the need to rescale $\hat{f}_{h_{crit}}$ to have the same variance as the sample variance as the dataset.

Finally, Hall and York note that the test is sensitive to extraneous maxima, in particular if the support of the underlying distribution is the full real line and the tails are heavy. In this case, extraneous maxima will generally appear, at large distances from one another. If these extraneous maxima are much farther apart than the true modes, h_{crit} will mostly be determined by these extraneous maxima, rather than the true modes of the distribution. Therefore, the magnitude of h_{crit} will no longer be representative of the unimodality or multimodality of the underlying distribution, and the reliability of the test result decreases, though it is not discussed whether this makes the test more conservative or anti-conservative. Hall and York suggest to only consider maxima on a compact interval on which the true modes are presumed to lie, rather than the full real line. The influence of any extraneous modes outside this interval is then nullified. In [6], Izenman and Sommer suggest to reduce the influence of this effect instead, by using a kernel width h that depends on t, as defined in equation (1). In particular, the value h would be smaller when the concentration of data points near t is large, and larger when the concentration of data points near t is small. This means that extraneous maxima, which mainly appear in the tails where the concentration of data points is relatively small, are combined more quickly as (the average value of) h increases compared to true modes.

2.2 Dip test

Hartigan and Hartigan's dip test, published in 1985 in [7], makes use of a different characteristic of the data to assess multimodality: the overall distance of the ECDF to the closest unimodal distribution. From the data, we can estimate the distribution function F using the Empirical Cumulative Distribution Function (ECDF) $\hat{F}$, defined by $\hat{F}(x) = \sum_{i=1}^{N} \mathbb{1}\{X_i \leq x\}$. Even if the underlying distribution is unimodal, the finite sample size means the ECDF is discontinuous at every data point, meaning

the ECDF is never unimodal[2]. Nevertheless, we intuitively expect the ECDF to be somewhat close, with respect to some metric, to a unimodal distribution.

The dip test proceeds as follows. From the ECDF, the best fitting unimodal distribution is found, with respect to the Kolmogorov-Smirnov statistic (KS). This statistic is defined in Section 2.2.1. The value of this statistic is then a measure for how close the ECDF is to being unimodal. If this distance is small, this provides evidence for the underlying distribution indeed being unimodal. Conversely, if this distance is large, the underlying distribution is likely multimodal.

2.2.1 Testing procedure

As discussed in the previous subsection, the dip test begins by fitting the best unimodal distribution to the ECDF $\hat{F}$ with respect to the KS statistic. The KS statistic simply measures the maximal vertical distance between two distribution functions. Let G, H be two distribution functions. We then have that

$$\mathrm{KS}(G, H) = \sup_{x \in \mathbb{R}} |G(x) - H(x)|.$$

We now wish to find minimal value of $\mathrm{KS}(\hat{F}, \tilde{F})$ among all unimodal distributions $\tilde{F}$. More formally, let $\mathcal{U}$ be the set of unimodal distribution functions. We can then define

$$D = \inf_{\tilde{F} \in \mathcal{U}} \mathrm{KS}(\hat{F}, \tilde{F}).$$

Here D is the so-called dip statistic. While an infimum is used in the definition, a distribution $\tilde{F}$ that minimises this statistic always exists, as shown in [7]. This distribution can then be interpreted as a *hypothesised underlying distribution*, being an estimate for the true underlying distribution. One of the authors of [7] provides an algorithm in [8] to compute this hypothesised distribution in linear ($O(N)$) time. Since any uniform distribution is unimodal, and we can always find a uniform distribution function with a maximal vertical distance of $\frac{1}{2}$ to any given ECDF, we immediately see that the dip statistic is bounded above by $\frac{1}{2}$. Trivially, the dip statistic is bounded below by 0.

A low dip statistic is then evidence for the ECDF being close to unimodal, which means the underlying distribution is likely unimodal as well.

The original formulation of the dip test is calibrated against the uniform distribution. For the uniform distribution, the probability the dip exceeds certain values is tabulated in [7]. While the uniform distribution satisfies the definition of unimodality, its modal interval (its full support) is of the weakest possible kind, in the sense that the density at this mode is not strictly larger than the density at any other point on the support. Since the mode is very weak, this can be used to calculate a conservative p-value for the test.

In practice, the resulting p-value is very conservative. The very weak mode of the uniform distribution means that the ECDF of data generated from this distribution is often far from a unimodal distribution compared to the ECDFs of data generated from unimodal distributions with more pronounced modes. More precisely, it is shown in [7] that among a large class of light tailed unimodal distributions, the dip statistic is asymptotically the largest for the uniform distribution[3], as $N \to \infty$. However, this means that if the data comes from a unimodal distribution other than the uniform distribution, the

[2]The only case this does not hold is if N=1, in which case the ECDF is *always* unimodal. However, since we cannot draw any conclusions about the unimodality or multimodality of the underlying distribution based on a single data point, we neglect this case.

[3]In [7] it is shown that $\sqrt{N}D$ converges in distribution to a (nonzero) constant for the uniform distribution, and converges in distribution to 0 for distributions that decay exponentially away from the mode. Furthermore, it is conjectured that among all unimodal distributions, D is asymptotically stochastically largest for the unimodal distribution. This means that as $N \to \infty$, we have that $\mathbb{P}(D_{unif} > x) \geq \mathbb{P}(D_{other} > x)$ for all $x \in \mathbb{R}$.

test is very unlikely to reject the null hypothesis. The power of the test can be increased by calibrating the test against a distribution that is likely to be closer to the true distribution. Since the form of the true distribution is unknown, one has to rely on non-parametric methods to choose a calibration distribution. For instance, Stoepker uses bootstrap samples from the minimising distribution $\tilde{F}$ to calibrate the test at runtime [9]. This results in a modified test, called the string test. This calibration is performed by finding the dip statistic for every bootstrap sample, and calculating the fraction of these that do not exceed the dip statistic of the original data. The resulting fraction can then be used as an approximate p-value for the test, while ensuring the test is approximately exact.

2.2.2 Advantages and disadvantages

Like Silverman's test, the dip test is very conservative, and has somewhat low power. For the dip test, this is mainly caused by the choice of calibration distribution. As stated before, the original formulation uses the uniform distribution for calibration, which is a pessimistic choice. If the calibration distribution is chosen to be closer to (an estimate of) the true distribution, power increases. Since this calibration distribution depends on the dataset used, one cannot compute dip statistic thresholds beforehand, and has to resort to a bootstrapping approach. This adds additional computational load, since one does not only need to minimise the KS statistic for the gathered data, but also for every bootstrap sample.

In [9] Stoepker also found numerical evidence that power could be increased further by modifying the dip statistic to be based on the Cramér-von Mises (CvM) statistic or Anderson-Darling (AD) statistic respectively. These are defined below:

$$\mathrm{CvM}(\hat{F},\tilde{F}) = \int_{-\infty}^{\infty} \left(\hat{F}(x)-\tilde{F}(x)\right)^2 d\tilde{F}(x) = \mathbb{E}\left[(\hat{F}(X)-\tilde{F}(X))^2 \middle| \mathbf{X}\right], \text{ and} \tag{2}$$

$$\mathrm{AD}(\hat{F},\tilde{F}) = \int_{-\infty}^{\infty} \frac{(\hat{F}(x)-\tilde{F}(x))^2}{\tilde{F}(x)(1-\tilde{F}(x))} d\tilde{F}(x) \ = \mathbb{E}\left[\frac{(\hat{F}(X)-\tilde{F}(X))^2}{\tilde{F}(X)(1-\tilde{F}(X))} \middle| \mathbf{X}\right], \tag{3}$$

where $X \sim \tilde{F}$, independently of $\mathbf{X}$.

Since the minimising distribution $\tilde{F}$ of the KS statistic exists, the regular the dip statistic can be defined as $D = \mathrm{KS}(\hat{F},\tilde{F})$. Stoepker then redefined the dip statistic as $D = \mathrm{CvM}(\hat{F},\tilde{F})$ or $D = \mathrm{AD}(\hat{F},\tilde{F})$ respectively, while $\tilde{F}$ was kept to be the distribution that minimises the KS statistic. In effect, one measures the distance between $\hat{F}$ and $\tilde{F}$ using the CvM or AD statistics rather than the KS statistic. These statistics have the advantage of being global statistics, taking into account the entire distributions, rather than only the point where the vertical distance is maximised.

Furthermore, it is important to note that the differential variable used in both integrals is $d\tilde{F}(x)$. This means that points where $\tilde{F}(x)$ changes rapidly, i.e. points of high probability density, are given a larger weight. Finally, it should be noted that the CvM and AD statistics are very similar, only differing in the denominator. This denominator is especially close to zero if $\tilde{F}(x)$ is close to zero or one, near the tails of the distribution. This gives the tails additional weight compared to the centre in the AD statistic.

3 Improving the power of the dip test

In the previous section, the dip test and its advantages and disadvantages were introduced. The original dip test is calibrated for a uniform distribution, resulting in a lack of power when applied to other distributions. This was improved upon by using bootstrapping, resulting in a better calibration in [9]. Numerical evidence in this report also suggests that assessing distances between the distributions using the CvM and AD statistic results in a further increase in power, even when the hypothesised distribution $\tilde{F}$ is still determined by minimising the KS statistic. This suggests that it might be possible to improve the calibration of the test further by using the distribution that minimises the CvM or AD statistics for the calibration step instead.

3.1 Optimising the CvM and AD statistics

This suggests the following procedure:

1. From the data sample of size N, compute the ECDF $\hat{F}$.
2. Determine a unimodal distribution function $\tilde{F}$ that minimises $\mathrm{CvM}(\hat{F}, \tilde{F})$ or $\mathrm{AD}(\hat{F}, \tilde{F})$.
3. Generate n bootstrap samples of size N from $\tilde{F}$, and compute their respective ECDFs $\{\hat{F}_i\}_{i=1}^n$.
4. For each of these $\hat{F}_i$, again determine a unimodal distribution function $\tilde{F}_i$ that minimises the CvM or AD statistic.
5. Report a p-value, given by
   $\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\mathrm{CvM}(\hat{F}_i, \tilde{F}_i) \geq \mathrm{CvM}(\hat{F}, \tilde{F})\}$, or
   $\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\mathrm{AD}(\hat{F}_i, \tilde{F}_i) \geq \mathrm{AD}(\hat{F}, \tilde{F})\}$.

This procedure is very similar to the procedure used in bootstrap-based goodness-of-fit testing, described in [10]. However, there are also key differences. Firstly, [10] restricts itself to a parametric setting, whereas we consider a non-parametric setting. Moreover, the null-hypothesis in [10] is simple ($H_0 : \theta_1 = \bar{\theta}_1$). Our null-hypothesis, on the other hand, is compound (F is multimodal).

Minimising the CvM and AD statistics is a highly nontrivial step in this procedure, unlike minimising the KS statistic in [7], [8] and [9]. After all, the KS statistic is a local statistic, generally only determined by a single point where the distance is maximal. The CvM and AD statistics, on the other hand, are global statistics. This makes the optimisation step significantly more difficult. Solving this optimisation problem is the main focus of this thesis. In particular, this thesis focuses mainly on minimising the CvM statistic, although the techniques used can be modified to work for the AD statistic as well. Furthermore, *a-priori* it is not clear if there even exists a distribution that minimises the CvM and AD statistics. This is also be discussed in the following subsections.

Before continuing, it is helpful to explicitly state the constraints on $\tilde{F}$ to be a unimodal distribution function. Since $\tilde{F}$ is a distribution function, it has to satisfy:

1. $\tilde{F}$ is monotonically increasing.
2. $\tilde{F}$ is right-continuous. (Note: in Section 3.2 we explain why we only consider continuous $\tilde{F}$).
3. $\lim_{x\to-\infty} \tilde{F}(x) = 0$, $\lim_{x\to\infty} \tilde{F}(x) = 1$.

Finally, $\tilde{F}$ has to be unimodal, as defined by Definition 2.1.

3.2 Simplification in a continuous setting

The goal is to minimise the CvM and AD statistics, defined in equations (2) and (3). For convenience, their definitions are restated below.

$$CvM(\hat{F},\tilde{F}) = \int_{-\infty}^{\infty} \left(\hat{F}(x) - \tilde{F}(x)\right)^2 d\tilde{F}(x) = \mathbb{E}\left[(\hat{F}(X) - \tilde{F}(X))^2 \middle| \mathbf{X}\right], \text{ and}$$
$$AD(\hat{F},\tilde{F}) = \int_{-\infty}^{\infty} \frac{(\hat{F}(x) - \tilde{F}(x))^2}{\tilde{F}(x)(1-\tilde{F}(x))} d\tilde{F}(x) \ = \mathbb{E}\left[\frac{(\hat{F}(X) - \tilde{F}(X))^2}{\tilde{F}(X)(1-\tilde{F}(X))} \middle| \mathbf{X}\right],$$

where $X \sim \tilde{F}$, independently of $\mathbf{X}$.

In the following sections, a continuous unimodal distribution G "scoring better" than another continuous unimodal distribution H refers to $\mathrm{CvM}(\hat{F},G) < \mathrm{CvM}(\hat{F},H)$ and $\mathrm{AD}(\hat{F},G) < \mathrm{AD}(\hat{F},H)$.

In a continuous setting, F is continuous. It is therefore reasonable to restrict ourselves to continuous $\tilde{F}$. If this restriction is not made, we generally find rather meaningless solutions. This is because if $\tilde{F}$ increases continuously from some value a to another value $b > a$, the expression $(\hat{F}(x) - \tilde{F}(x))$ is evaluated by the integral at every value of $\tilde{F}(x)$ between a and b. If, on the other hand, $\tilde{F}(x)$ jumps discontinuously from a to b at some x, all the mass associated with the jump is applied at the value $\tilde{F}$ takes at x, which by right-continuity of $\tilde{F}$ is b. To illustrate further why this is problematic, we give an example. Suppose we consider the unimodal distribution $\tilde{F}(x) = \mathbb{1}\{x \geq \max_{1\leq i\leq N}\{X_i\} + 1\}$. This distribution clearly does not resemble the dataset, placing all probability mass at a single point outside the data range, while placing no mass at any of the data points. However, if we evaluate the CvM statistic, we find:

$$\begin{aligned}
\mathrm{CvM}(\hat{F},\tilde{F}) &= \mathbb{E}\left[(\hat{F}(X) - \tilde{F}(X))^2 \middle| \mathbf{X}\right] \\
&= \left(\hat{F}(\max_{1\leq i\leq N}\{X_i\} + 1) - \tilde{F}(\max_{1\leq i\leq N}\{X_i\} + 1)\right)^2 \\
&= (1-1)^2 = 0.
\end{aligned}$$

This clearly minimises the CvM statistic, while not resembling the dataset. To avoid such meaningless solutions, we therefore restrict ourselves to continuous $\tilde{F}$. Since F is continuous, the data points $\{X_i\}_{i=1}^N$ are almost surely all distinct. Let $y_i = \tilde{F}(X_{(i)})$, where $X_{(i)}$ is the i-th order statistic of the dataset, and let $\mathbf{y} = (y_i)_{i=1}^N$ be the vector whose components are these values y_i. We then find that:

$$\mathrm{CvM}(\mathbf{y}) = \frac{1}{12N^2} + \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \frac{2i-1}{2N}\right)^2, \text{ and} \tag{4}$$
$$\mathrm{AD}(\mathbf{y}) = -1 - \frac{1}{N^2}\sum_{i=1}^{N}((2i-1)\log(y_i) + (2N+1-2i)\log(1-y_i)). \tag{5}$$

The derivation of equation (4) can be found in Appendix A.1. Equation (5) can be obtained by a similar derivation. Note that, interestingly, the above statistics only depend on the values $y_i = \tilde{F}(X_{(i)})$. Altering the way $\tilde{F}$ interpolates or extrapolates the values it takes at the data points does not change the value of the respective statistics. In the absence of any restrictions, these expressions are minimised with the choice $y_i = \frac{2i-1}{2N}$, and decrease as any y_i is brought closer to this value. For the CvM statistic this can be seen immediately from the expression, and for the AD statistic the same can be shown to hold using elementary calculus. However, since $\mathbf{y}$ is constrained by the fact that it must correspond to a unimodal distribution $\tilde{F}$, which is in turn constrained by the criterion given by Definition 2.1, this unconstrained optimum (generally) cannot be achieved.

In the remainder of this section, we first show that among all possible interpolations and extrapolations of the data points, a piecewise-linear choice is the least restrictive option. More formally, for any

hypothesised continuous unimodal distribution, there exists a piecewise-linear continuous unimodal distribution yielding the same statistic values. This means we can consider only piecewise-linear distributions in our further analysis, without loss of generality. We then show that a piecewise-linear function whose segments connect is a unimodal distribution if and only if it satisfies three constraints. These three constraints are analogous to the requirements of *monotonicity* of $\tilde{F}$, the requirement that $\tilde{F}(x)$ approaches 0 and 1 as $x \to \pm\infty$, and the requirement of *unimodality* of $\tilde{F}$. Next, we show that if either of the first two of these constraints is not satisfied, there exists a piecewise-linear distribution that scores better, that satisfies all three constraints. This means that when searching for an optimum, only the third constraint needs to be enforced[4]. Next, we show that in this larger set an optimum indeed exists. Finally, we elaborate on the structure of the solution, which aids us in numerically finding the optimum.

3.2.1 Piecewise-linearity of optimal solution

In this section we prove that we can assume, without loss of generality, that the hypothesised distribution $\tilde{F}$ is piecewise-linear. We specifically exclude hypothesised distributions for which $y_1 = 1$ or $y_N = 0$. If this were the case, then monotonicity of the distribution implies that $y_i = 1$ for all i or $y_i = 0$ for all i. Such distributions are always be suboptimal with respect to the CvM statistic, and even lie outside the domain of the AD statistic.

We begin by defining the concepts of local convexity and local concavity.

Definition 3.1 (Local convexity and concavity)**.** *Let f be a function that is defined on an interval [a,b], with $a, b \in \mathbb{R} \cup \{-\infty, \infty\}$. f is called (strictly) locally convex in a point $x \in (a, b)$ if and only if there exists an $r > 0$ such that f is (strictly) convex on $[x - r, x + r]$. We define local concavity in an analogous way.*

It is easy to show that a continuous function f is convex on $[a, b]$ if and only if f is locally convex for all $x \in (a, b)$. For a continuous piecewise-linear function f, the following two properties trivially hold for every $x \in \mathbb{R}$.

1. f is either strictly locally convex or locally concave in x
2. f is either strictly locally concave or locally convex in x.

We now state a lemma that is helpful in showing that, without loss of generality, $\tilde{F}$ is piecewise-linear. Conceptually, this lemma has to do with the fact that we need to be able to extrapolate the piecewise-linear solution outside the data range $[X_{(1)}, X_{(N)}]$ to $y = 0$ and $y = 1$, without violating the unimodality constraint. The proof of this lemma can be found in Appendix A.2.

Lemma 3.2. *Let $N \geq 2$. Let $\tilde{F}$ be any continuous unimodal distribution, with $\tilde{F}(X_{(1)}) < 1$ and $\tilde{F}(X_{(N)}) > 0$. Define $y_i = \tilde{F}(X_{(i)})$. Then the following statements are true:*

- *$y_1 = 0$ or $y_2 > y_1$, and*
- *$y_N = 1$ or $y_N > y_{N-1}$*

We can now show that $\tilde{F}$ can be assumed to be piecewise-linear.

Lemma 3.3 (Piecewise-linearity of hypothesised distribution)**.** *Let $N \geq 2$. Let $\tilde{F}$ be any continuous unimodal distribution, with $\tilde{F}(X_{(1)}) < 1$ and $\tilde{F}(X_{(N)}) > 0$. Define $y_i = \tilde{F}(X_{(i)})$. Then there exists a continuous piecewise-linear unimodal distribution, that is equal to $\tilde{F}$ at every data point and therefore yields the same statistic values as $\tilde{F}$.*

Proof. Let $\bar{F}$ be the piecewise-linear function connecting the various points $(X_{(i)}, y_i)$, linearly

[4]Effectively, we are extending the search space to a larger set. This is allowed, since we prove that we will certainly not find the optimum in a point in this larger set that is not in the original set

extrapolated outwards, and capped at 0 and 1. In other words:

$$\bar{F}(x) = \left\{ \begin{array}{lll} 0, & & \text{for } x < x_l \\ y_1 + \frac{y_2 - y_1}{X_{(2)} - X_{(1)}} & (x - X_{(1)}), & \text{for } x_l \leq x < X_{(1)} \\ y_i + \frac{y_{i+1} - y_i}{X_{(i+1)} - X_{(i)}} & (x - X_{(i)}), & \text{for } X_{(i)} \leq x < X_{(i+1)}, 1 \leq i \leq N-1 \\ y_N + \frac{y_N - y_{N-1}}{X_{(N)} - X_{(N-1)}} & (x - X_{(N)}), & \text{for } X_{(N)} \leq x < x_r \\ 1, & & \text{for } x \geq x_r \end{array} \right\},$$

where

$$x_l = X_{(1)} - \frac{X_{(2)} - X_{(1)}}{y_2 - y_1} \cdot y_1, \text{ and}$$
$$x_r = X_{(N)} + \frac{X_{(N)} - X_{(N-1)}}{y_N - y_{N-1}} \cdot (1 - y_N).$$

are the x-coordinates where the left linear extrapolation intersects $y = 0$ and the right linear extrapolation intersects $y = 1$ respectively. If $y_1 = y_2 = 0$, we omit the segment between x_l and $X_{(1)}$. Similarly, if $y_{N-1} = y_N = 1$, we omit the segment between $X_{(N)}$ and x_r. The construction of $\bar{F}$ is shown in Figure 3.

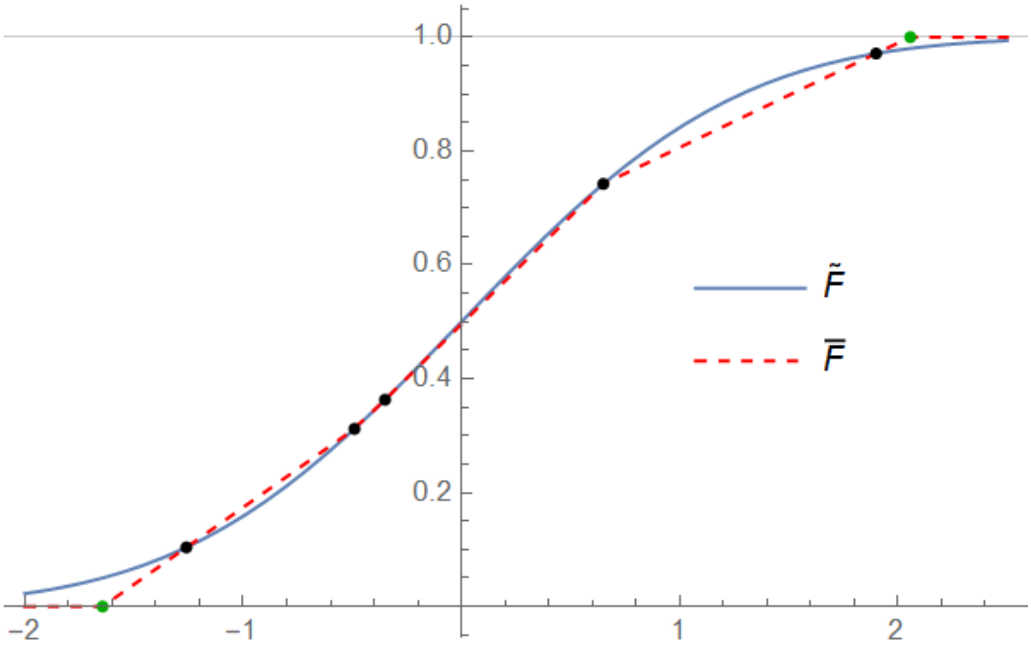


Figure 3: The construction of $\bar{F}$ from a hypothesised distribution $\tilde{F}$. The x-coordinates of the black points indicate the data points, and the green points indicate x_l and x_r.

Then we can show that $\bar{F}$ is a continuous unimodal distribution. It is easy to verify that the linear segments in $\bar{F}$ connect, meaning $\bar{F}$ is continuous. Monotonicity of $\bar{F}$ is guaranteed by the monotonicity of $\tilde{F}$. The limits of $\bar{F}$ as $x \to \pm\infty$ are also trivial. We thus only need to show that $\bar{F}$ is unimodal.

Since $\bar{F}$ is piecewise-linear, it is both locally convex and locally concave in all points other than the points where the linear segments meet. These are the data points $X_{(2)}$ through $X_{(N-1)}$, and x_l and x_r. At x_l, the function is locally convex. At x_r, the function is locally concave. We thus only need to consider local convexity and concavity in $X_{(2)}$ through $X_{(N-1)}$. Since $\bar{F}$ consists of secant lines of $\tilde{F}$, $\tilde{F}$ being convex or concave on $[X_{(i-1)}, X_{(i+1)}]$ for some i implies that $\bar{F}$ is respectively locally convex or locally concave in $X_{(i)}$. This argument is used repeatedly in the remainder of the proof.

Let $m \in \mathbb{R}$ be a value such that $\tilde{F}$ is convex on $(-\infty, m]$ and concave on $[m, \infty)$. Then either m lies outside the data range, m lies on a data point, or m lies in between two data points. In other words, either $m < X_{(1)}$ (or $m > X_N$), $m = X_{(i)}$ for some i, or $m \in (X_{(i)}, X_{(i+1)})$ for some i.

If $m < X_{(1)}$ (or $m > X_N$), then $\tilde{F}$ is convex (or concave) on $[X_{(1)}, X_{(N)}]$. This implies that $\bar{F}$ is locally convex (or locally concave) in every $X_{(i)}$ with $2 \leq i \leq N-1$. This means that $\bar{F}$ is unimodal. Similarly, if $m = X_{(i)}$ for some i, $\tilde{F}$ is convex on $(-\infty, X_{(i)}]$ and concave on $[X_{(i)}, \infty)$. We then find that $\bar{F}$ is locally convex in the data points $X_{(2)}$ through $X_{(i-1)}$, and locally concave in the data points $X_{(i+1)}$ through $X_{(N-1)}$. It is then irrelevant whether $\bar{F}$ is locally concave or locally convex in $X_{(i)}$, in both cases $\bar{F}$ is unimodal.

Finally, we consider the case where $m \in (X_{(i)}, X_{(i+1)})$ for some i. Since $\tilde{F}$ is convex on $(-\infty, X_{(i)}] \subset (-\infty, m]$ and concave on $[X_{(i+1)}, \infty)$, it holds that $\bar{F}$ is locally convex in $X_{(2)}$ through $X_{(i-1)}$ and locally concave in $X_{(i+2)}$ through $X_{(N-1)}$. We therefore only need to prove that $\bar{F}$ is either locally convex in $X_{(i)}$ or locally concave in $X_{(i+1)}$. After all, if $\bar{F}$ is locally convex in $X_{(i)}$ it is locally convex for all $x < X_{(i+1)}$, and locally concave for all $x > X_{(i+1)}$, and therefore unimodal. If $\bar{F}$ is locally concave in $X_{(i+1)}$, it is locally concave for all $x > X_{(i)}$ and locally convex for all $x < X_{(i)}$, and therefore also unimodal.

By convexity of $\tilde{F}$ on $(-\infty, m]$, we have that the slope of $\tilde{F}$ is at least as large as $\frac{y_i - y_{i-1}}{X_{(i)} - X_{(i-1)}}$ on $[X_{(i)}, m]$. Similarly, by concavity of $\tilde{F}$ on $[m, \infty)$, we have that the slope of $\tilde{F}$ is at least as large as $\frac{y_{i+2} - y_{i+1}}{X_{(i+2)} - X_{(i+1)}}$ on $[m, X_{(i+1)}]$. On $[X_{(i)}, X_{(i+1)}]$, being the union of these intervals, the slope of $\tilde{F}$ is then at least as large as the minimum of these quantities. We thus find
$\frac{y_{i+1} - y_i}{X_{(i+1)} - X_{(i)}} \geq \min(\frac{y_i - y_{i-1}}{X_{(i)} - X_{(i-1)}}, \frac{y_{i+2} - y_{i+1}}{X_{(i+2)} - X_{(i+1)}})$.

If $\frac{y_{i+1} - y_i}{X_{(i+1)} - X_{(i)}} \geq \frac{y_i - y_{i-1}}{X_{(i)} - X_{(i-1)}}$, then $\bar{F}$ is locally convex in $X_{(i)}$.
If $\frac{y_{i+1} - y_i}{X_{(i+1)} - X_{(i)}} \geq \frac{y_{i+2} - y_{i+1}}{X_{(i+2)} - X_{(i+1)}}$, then $\bar{F}$ is locally concave in $X_{(i+1)}$.

In all cases, we therefore find that $\bar{F}$ is a continuous unimodal distribution.

Since by definition, $\bar{F}(X_{(i)}) = \tilde{F}(X_{(i)})$ for all i, we find that $\mathrm{CvM}(\hat{F}, \bar{F}) = \mathrm{CvM}(\hat{F}, \tilde{F})$ and $\mathrm{AD}(\hat{F}, \bar{F}) = \mathrm{AD}(\hat{F}, \tilde{F})$. □

Without loss of generality, we can therefore assume that our hypothesised distribution $\tilde{F}$ is piecewise-linear.

This piecewise-linear distribution is fully determined by the values y_i. This transforms the continuous optimisation problem for a distribution function $\tilde{F}$ into a finite-dimensional optimisation problem for a vector $\mathbf{y}$. This vector $\mathbf{y}$ still needs to correspond to a unimodal distribution. We therefore translate the constraints on $\tilde{F}$ to constraints on $\mathbf{y}$.

1. **monotonicity**: $\tilde{F}$ is a monotonically increasing function $\iff$ $\{y_i\}_{i=1}^N$ is a monotonically increasing sequence .
2. **extendability**: $(\lim_{x \to -\infty} \tilde{F}(x) = 0, \lim_{x \to \infty} \tilde{F}(x) = 1) \iff$
   - $y_1 = 0$ or $y_2 > y_1 \geq 0$
   - $y_N = 1$ or $y_{N-1} < y_N \leq 1$
3. **unimodality**: F is unimodal $\iff \exists_{1 \leq m \leq N-1}$ such that:
   - $\frac{y_{i+1} - y_i}{X_{(i+1)} - X_{(i)}} \geq \frac{y_i - y_{i-i}}{X_{(i)} - X_{(i-1)}}$ for $1 < i \leq m$
   - $\frac{y_{i+1} - y_i}{X_{(i+1)} - X_{(i)}} \leq \frac{y_i - y_{i-i}}{X_{(i)} - X_{(i-1)}}$ for $m < i < N$

3.2.2 Simplifying the constraint set

The goal then becomes to minimise the functions given by equations (4) and (5), subject to the constraints given above. This is a finite-dimensional optimisation problem. However, the constraint set can be further simplified. In this section, we show show that monotonicity and extendability do not have to be enforced externally, in the sense that they follow automatically from unimodality and optimality of the solution.

The argument proceeds as follows: Firstly, we show that if $\mathbf{y}$ satisfies unimodality but violates monotonicity, there exists a vector that scores better that satisfies both unimodality and monotonicity. Secondly, we show that if $\mathbf{y}$ satisfies unimodality and monotonicity but not extendability, there exists a vector that scores better than $\mathbf{y}$ and satisfies all three constraints. This completes the argument. Since the proofs for these statements are rather mundane and give little further insight into the problem, this section only contains the statements themselves. For their proofs, the reader is referred to Appendix A.3. In short, these statements are proven by making small modifications to the vectors $\mathbf{y}$, making them score better and simultaneously satisfy the respective constraints.

Lemma 3.4 (Redundancy of the monotonicity constraint)**.** *Let* $\mathbf{y}$ *satisfy unimodality, but not monotonicity. Then there exists a* $\hat{\mathbf{y}}$ *that satisfies both unimodality and monotonicity, and scores better than* $\mathbf{y}$*.*

Lemma 3.5 (Redundancy of the extendability constraint)**.** *Let* $\mathbf{y}$ *satisfy unimodality and monotonicity, but not extendability. Then there exists a* $\hat{\mathbf{y}}$ *that satisfies unimodality, monotonicity and extendability, and scores better than* $\mathbf{y}$*.*

Therefore, only unimodality needs to be enforced. We can now prove an optimum indeed exists. Let S be the set of all vectors $\mathbf{y}$ that satisfy the unimodality constraint. Then S is a closed subset of $\mathbb{R}^n$. Take some $\mathbf{y} \in S$, and let $A = \{\mathbf{z} \in S | \mathrm{CvM}(\mathbf{z}) \leq \mathrm{CvM}(\mathbf{y})\}$. We can do the same for the AD statistic, except that we need to take care to only include vectors $\mathbf{y}$ for which $y_i \in (0, 1)$ for all i, in order for the AD statistic to be defined. Since the objective functions are continuous, and go to ∞ at the boundaries of the domain[5], A is a closed subset of S, and therefore a closed subset of $\mathbb{R}^n$. Moreover, since the domain is bounded ($y_i \in [0, 1]$), A is a bounded set. Therefore, Weierstrass' extremum theorem guarantees an optimum indeed exists.

3.2.3 The structure of the solution

In order to find the optimal solution, we could determine the optimal hypothesised solution for every possible value of m, and then determine the minimum of these values. This would require a total of N fits, of N data points each. However, this problem can be reduced to a pair of simpler optimisation problems.

Recall the goal is to find the vector $\mathbf{y}$ that satisfies the unimodality constraint and minimises the AD or CvM statistic. Removing constants that are irrelevant for the optimisation process from equations (4) and (5), we arrive at the following objective functions:

$$g_{CvM}(\mathbf{y}) = \sum_{i=1}^{N} \left(y_i - \frac{2i-1}{2N} \right)^2, \text{ and}$$

$$g_{AD}(\mathbf{y}) = -\sum_{i=1}^{N} \left((2i-1)\log(y_i) + (2N+1-2i)\log(1-y_i) \right).$$

When considering properties that hold for both g_{CvM} and g_{AD}, we simply write g.

[5]From equations (4) and (5) it is apparent that the CvM and AD statistics approach ∞ as any $y_i \to \pm\infty$ or any $y_i \to 0$ or 1 respectively.

We can show that both these objective functions are strictly convex. This is done in Appendix A.4. We now define the functions $g_m^-(\mathbf{y})$ and $g_m^+(\mathbf{y})$ for every m with $0 \leq m \leq N+1$. These functions are defined as

$$g_m^-(\mathbf{y}) = \sum_{i=1}^{m} \left(y_i - \frac{2i-1}{2N} \right)^2, \text{ and} \tag{6}$$
$$g_m^+(\mathbf{y}) = \sum_{i=m}^{N} \left(y_i - \frac{2i-1}{2N} \right)^2.$$

We define g_m^- and g_m^+ analogously for the AD statistic, by replacing the bounds of the sum in the objective function by 1 to m and m to N respectively. Note that $g(\mathbf{y}) = g_m^+(\mathbf{y}) + g_{m+1}^-(\mathbf{y})$ for all m. These newly defined functions can be thought of as the objective function, where only the first m or last $N+1-m$ points are taken into consideration respectively. Note that if $m = 0$ or $m = N+1$, one of these becomes an empty sum. We show that the optimal solution to the problem is the concatenation of the solution that minimises g_m^- under a convexity constraint and the solution that minimises g_{m+1}^+ under a concavity constraint for some $0 \leq m \leq N$. Importantly, this means we can treat the convex part and concave part as entirely separate, and do not need to take into account the constraints at $X_{(m)}$ and $X_{(m+1)}$ where the convex and concave part meet. In practice, this means that optimising the objective function can be viewed as the collection of two separate optimisation problems: optimising g_m^- and g_{m+1}^+ for every value of m.

Theorem 3.6 (Structure of the solution)**.** *Let $\mathbf{y}^* = (y_i^*)_{i=1}^N$ be the vector that optimises the objective function under the unimodality constraint. Then there exists a $0 \leq k \leq N$ such that $(y_i^*)_{i=1}^k$ is the convex solution that minimises g_k^- and $(y_i^*)_{i=k+1}^N$ is the concave solution that minimises g_{k+1}^+.*

Proof. Let $\mathbf{y}^*$ be given as above. There are two possibilities:

Case 1: the solution is neither locally convex in every $X_{(i)}$ with $2 \leq i \leq N-1$ (fully convex) nor locally concave in every $X_{(i)}$ with $2 \leq i \leq N-1$ (fully concave).

If the solution is neither fully convex nor fully concave, then there exists a largest k with $1 < k < N-1$ for which the corresponding $\tilde{F}$ is strictly locally convex in $X_{(k)}$, in other words $\frac{y_{k+1}^* - y_k^*}{X_{(k+1)} - X_{(k)}} > \frac{y_k^* - y_{k-1}^*}{X_{(k)} - X_{(k-1)}}$. This k exists since the solution is not fully concave, and $k < N-1$ since if $k = N-1$ the solution would be fully convex.

Now let $\mathbf{y}^- = (y_i^-)_{i=1}^k$ be the vector that optimises g_k^- under the constraint that $\frac{y_{i+1}^- - y_i^-}{X_{(i+1)} - X_{(i)}} \geq \frac{y_i^- - y_{i-1}^-}{X_{(i)} - X_{(i-1)}}$ for any $1 < i < k$. Furthermore, let $\mathbf{y}^+ = (y_i^+)_{i=k+1}^N$ be the vector that optimises g_{k+1}^+ under the constraint that $\frac{y_{i+1}^+ - y_i^+}{X_{(i+1)} - X_{(i)}} \leq \frac{y_i^+ - y_{i-1}^+}{X_{(i)} - X_{(i-1)}}$ for any i with $k+1 < i < N$. Finally, let $\hat{\mathbf{y}}$ be the concatenated vector, defined by $\hat{y}_i = \left\{ \begin{array}{ll} y_i^-, & \text{for } 1 \leq i \leq k \\ y_i^+, & \text{for } k < i \leq N \end{array} \right\}$. We show that $\hat{\mathbf{y}} = \mathbf{y}^*$.

Since $\hat{\mathbf{y}}$ is less constrained than $\mathbf{y}^*$[6], we must have that $g(\hat{\mathbf{y}}) \leq g(\mathbf{y}^*)$. However, $\hat{\mathbf{y}}$ might not satisfy the unimodality constraint: it might be the case that $\mathbf{y}$ is both strictly locally concave in $X_{(k)}$ and strictly locally convex in $X_{(k+1)}$. However, since $\frac{y_{k+1}^* - y_k^*}{X_{(k+1)} - X_{(k)}} > \frac{y_k^* - y_{k-1}^*}{X_{(k)} - X_{(k-1)}}$, one can choose $\lambda > 0$ small enough such that, $\mathbf{z} = \lambda \mathbf{y}^* + (1-\lambda)\hat{\mathbf{y}}$ is still locally convex in $X_{(k)}$. It then no longer matters if $\mathbf{z}$ is locally convex or locally concave in $X_{(k+1)}$, it will satisfy the unimodality constraint either way. This vector $\mathbf{z}$ cannot score better than $\mathbf{y}^*$, since by assumption, $\mathbf{y}^*$ is optimal. Since the objective function is strictly convex, the only way this can be the case is if $\mathbf{y}^* = \hat{\mathbf{y}}$.

[6]Note $\hat{\mathbf{y}}$ satisfies the same constraints that $\mathbf{y}^*$ does, except (possibly) the constraints at $X_{(k)}$ and $X_{(k+1)}$.

The optimal solution must therefore be a concatenation of the best convex fit for points 1 through k, and the best concave fit for points $k+1$ through N, for some $1 < k < N-1$.

Case 2: The solution is fully convex or fully concave.

If $\mathbf{y}^*$ is fully convex or fully concave, we again find that $\mathbf{y}^*$ is the best convex or concave fit respectively for points 1 through N. This corresponds to the cases $k = N$ or $k = 1$ respectively. □

The problem thus reduces to two simpler problems: finding the best convex fit of the leftmost m points, and finding the best concave fit for points $m+1$ through N, for every value of m with $1 \leq m \leq N$. We can then concatenate them as done before with $\hat{\mathbf{y}}$, and choose the minimum value among all possibilities for m.

Note that, generally, not all of these concatenated vectors will be unimodal[7]. Nevertheless, we have proven that the optimal vector $\mathbf{y}^*$ within the class of vectors corresponding to a unimodal distribution *will* be one of these concatenated vectors. This means multimodal concatenated vectors can be discarded immediately.

3.3 Numerical optimisation

In the remainder of this section, we only address how we numerically determine the best convex fit of the leftmost m points. After all, determining the best concave fit of the rightmost $N+1-m$ points is analogous.

3.3.1 Reparametrizing the problem

In order to numerically treat the problem, it is helpful to parameterise the problem in a way that better captures the convexity requirement. We therefore observe that $\mathbf{y} = \sum_{i=1}^{m} y_i \mathbf{e}_i$, with $\{\mathbf{e}_i\}_{i=1}^{m}$ the standard basis of $\mathbb{R}^m$. While this is a highly trivial result, it allows us to switch to a different set of basis vectors. These basis vectors in turn represent piecewise-linear basis functions, a linear combination of which forms the final solution. There are several such bases that better capture the convexity requirement. In general, this comes at a cost of complicating the objective functions, so a compromise has to be made. In the end, a basis that works particularly well is the step-function basis $\{\mathbf{v}_i\}_{i=1}^{m}$ given by $(\mathbf{v}_i)_j = \mathbb{1}\{j \geq i\}$. This basis makes the convexity requirement decently simple, while only complicating the objective functions by a small amount. Furthermore, if one views these vectors purely as vectors in $\mathbb{R}^m$, and not as objects representing a piecewise-linear function, the vectors do not depend on the data. This makes reasoning about numerical optimisation in the absence of concrete data much simpler. The transition matrices are given below:

$$\mathbf{y} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 1 & 0 & \dots & 0 & 0 \\ 1 & 1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & 1 & \dots & 1 & 0 \\ 1 & 1 & 1 & \dots & 1 & 1 \end{pmatrix} \cdot \boldsymbol{\delta} =: M^{-1} \cdot \boldsymbol{\delta},$$

$$\boldsymbol{\delta} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 1 \end{pmatrix} \cdot \mathbf{y} =: M \cdot \mathbf{y}.$$

[7] In subsequent sections, we use the terms "unimodal" and "multimodal" vectors to refer to vectors that respectively correspond to unimodal and multimodal distributions.

We thus see that for $i \geq 2$, $\delta_i = y_i - y_{i-1}$. The coefficients δ_i can thus be viewed as the slope of the solution between $X_{(i-1)}$ and $X_{(i)}$, scaled with the distance between these two adjacent data points. The convexity requirement, which can be rephrased as a monotonically increasing slope, then reduces to $\frac{\delta_{i+1}}{X_{(i+1)}-X_{(i)}} \geq \frac{\delta_i}{X_{(i)}-X_{(i-1)}}$ for $2 \leq i \leq m-1$.

We define $\boldsymbol{\delta}^* = M \cdot \mathbf{y}^*$ to be the optimum, written in the new parameters.

3.3.2 Projected Gradient Descent

From now on, we restrict ourselves to the CvM statistic. However, similar methods can be used to optimise the AD statistic as well.

We can use the new parameters in a projected gradient descent procedure. Such methods consist of two parts: Firstly, one calculates the gradient of the objective function in the current point. In addition to information from previous iterations, this is used to make a step to a different point. This point may lie outside the feasible set. If this is the case, the point is projected into the feasible set, by finding the point in the feasible set which is closest with respect to the Euclidean norm. In this section, $\|\cdot\|$ always refers to the Euclidean norm of a vector. Due to its quadratic convergence, compared to a linear convergence for regular gradient descent, we opt for Nesterov accelerated projected gradient descent [11]. The algorithm is given below[8].

`Nesterov accelerated projected gradient descent.`
Input: `Objective function` g, `initial guess` $\boldsymbol{\delta}^{(0)} = \boldsymbol{\delta}^{(-1)}$, `time-step` t
for *k=1,2,...* **do**
  $\mathbf{q} \leftarrow \boldsymbol{\delta}^{(k-1)} + \frac{k-2}{k+1}(\boldsymbol{\delta}^{(k-1)} - \boldsymbol{\delta}^{(k-2)})$
  $\boldsymbol{\delta}^{(k)} \leftarrow \texttt{prox}(\mathbf{q} - t\nabla g(\mathbf{q}))$
end

The for loop is terminated after an appropriate number of iterations. For numerical stability, it is required that the gradient of g is Lipschitz continuous, and that $t \leq \frac{1}{L}$, where L is the Lipschitz constant of ∇g. In other words, we must have:

$$\forall_{\boldsymbol{\delta},\hat{\boldsymbol{\delta}}} : \left\|\nabla g(\boldsymbol{\delta}) - \nabla g(\hat{\boldsymbol{\delta}})\right\| \leq L \left\|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}\right\|.$$

The algorithm consists of three steps. Firstly, from the current point, a step is taken to a point called $\mathbf{q}$ using the momentum carried from the previous iteration. The magnitude of this momentum step is modulated by the fraction $\frac{k-2}{k+1}$. For $k = 1$, the momentum term is zero, since $\boldsymbol{\delta}^{(0)} = \boldsymbol{\delta}^{(-1)}$. For $k = 2$, the momentum step is again zero, since $\frac{k-2}{k+1} = 0$. As k grows, the momentum step approaches 1, and therefore becomes more and more prevalent. It is important to note that $\mathbf{q}$ might already lie outside the feasible set. Therefore, the point $\mathbf{q}$ should not be viewed as an improved guess for the point $\boldsymbol{\delta}^*$, but rather purely as an intermediate point that is used to compute such a improved guess.

Next, from $\mathbf{q}$ a step is taken opposing the gradient of g. Since the gradient points in the direction of steepest ascent, this results in a reduction in the function value. To make this more rigorous, for any point $\boldsymbol{\sigma}$ connecting $\mathbf{q}$ and $\mathbf{q} - t\nabla g(\mathbf{q})$ we have $\|\boldsymbol{\sigma} - \mathbf{q}\| \leq t\|\nabla g(\mathbf{q})\|$. By the Lipschitz condition on ∇g, it holds that $\|\nabla g(\boldsymbol{\sigma}) - \nabla g(\mathbf{q})\| \leq Lt\|\nabla g(\mathbf{q})\| \leq \|\nabla g(\mathbf{q})\|$. This implies for any such point $\boldsymbol{\sigma}$ the movement still opposes the gradient, hence a reduction in function value is guaranteed. Finally, the result is projected into the feasible set using the prox operator, which simply returns the closest (with respect to the Euclidean norm) point in the feasible set.

[8] In literature, such as in [12] and [13], a different momentum coefficient is sometimes used, given by $\frac{t_k - 1}{t_{k+1}}$, where $(t_k)_{k\in\mathbb{N}}$ is sequence recursively defined by $t_{k+1} = \frac{1}{2} + \frac{1}{2}\sqrt{1 + 4t_k^2}$ and $t_0 = 1$. Though this choice generally results in faster convergence in practice, the same theoretical convergence rate can also be used for the simpler coefficient.

This method guarantees a quadratic convergence rate for convex functions with a Lipschitz continuous gradient. In particular, for any k, we have that [11]:

$$g(\boldsymbol{\delta}^{(k)}) - g(\boldsymbol{\delta}^*) \leq \frac{2\left\|\boldsymbol{\delta}^{(0)} - \boldsymbol{\delta}^*\right\|^2}{t(k+1)^2} \approx \frac{2L\left\|\boldsymbol{\delta}^{(0)} - \boldsymbol{\delta}^*\right\|^2}{(k+1)^2}. \tag{7}$$

Where the last approximate equality holds if $t \approx \frac{1}{L}$. It is immediately apparent from the inequality that a larger time-step t results in a faster convergence. However, this is limited by the numerical stability requirement on t. We use this procedure to minimise the previously defined function g_m^-, up to some predetermined accuracy level. In order to estimate the error after k time-steps, we need to estimate L and $\left\|\boldsymbol{\delta}^{(0)} - \boldsymbol{\delta}^*\right\|^2$. This then allows us to calculate the minimum number of time-steps k before we are guaranteed to find a solution with $f(\boldsymbol{\delta}^{(k)}) - f(\boldsymbol{\delta}^*) < \epsilon$, where $\epsilon > 0$ is some predetermined accuracy level.

We begin by estimating L. Plugging the reparametrization into the objective function given by equation (6), we find:

$$g_m^-(\boldsymbol{\delta}) = \sum_{i=1}^{m}\left(\sum_{j=1}^{i}\delta_j - \frac{2i-1}{2N}\right)^2.$$

We can now calculate its partial derivatives with respect to δ_k, $1 \leq k \leq m$. Doing this, we obtain

$$\frac{\partial g_m^-}{\partial \delta_k}(\boldsymbol{\delta}) = 2\sum_{i=k}^{m}\left(\sum_{j=1}^{i}\delta_j - \frac{2i-1}{2N}\right).$$

Note the bounds of the outer sum changed, since only the terms with $i \geq k$ contain δ_k. We can now derive an upper bound for the Lipschitz constant of ∇g.

$$\left\|\nabla g_m^-(\boldsymbol{\delta}) - \nabla g_m^-(\hat{\boldsymbol{\delta}})\right\|^2 = \sum_{k=1}^{m}\left(2\sum_{i=k}^{m}\sum_{j=1}^{i}(\delta_j - \hat{\delta}_j)\right)^2$$

Switching the order of summation, we get

$$\begin{aligned}\left\|\nabla g_m^-(\boldsymbol{\delta}) - \nabla g_m^-(\hat{\boldsymbol{\delta}})\right\|^2 &= \sum_{k=1}^{m}\left(2\sum_{j=1}^{m}(\delta_j - \hat{\delta}_j)\sum_{i=\max(k,j)}^{m}1\right)^2 \\ &= \sum_{k=1}^{m}\left(2\sum_{j=1}^{m}(\delta_j - \hat{\delta}_j)(m+1-\max(k,j))\right)^2.\end{aligned}$$

Using the arithmetic mean-quadratic mean inequality, we can write the above as

$$\begin{aligned}\left\|\nabla g_m^-(\boldsymbol{\delta}) - \nabla g_m^-(\hat{\boldsymbol{\delta}})\right\|^2 &\leq 4m\sum_{k=1}^{m}\sum_{j=1}^{m}(\delta_j - \hat{\delta}_j)^2(m+1-\max(k,j))^2 \\ &\leq 4m\sum_{k=1}^{m}(m+1-k)^2\sum_{j=1}^{m}(\delta_j - \hat{\delta}_j)^2 \\ &= 4m\sum_{k=1}^{m}(m+1-k)^2\left\|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}\right\|^2\end{aligned}$$

Finally, reversing the summation:

$$\begin{aligned}\left\|\nabla g_m^-(\boldsymbol{\delta}) - \nabla g_m^-(\hat{\boldsymbol{\delta}})\right\|^2 &\leq 4m \left\|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}\right\|^2 \sum_{k=1}^{m} k^2 \\ &= \left(\frac{4}{3}m^4 + 2m^3 + \frac{2}{3}m^2\right) \left\|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}\right\|^2\end{aligned}$$

Hence $L \leq \sqrt{\frac{4}{3}m^4 + 2m^3 + \frac{2}{3}m^2}$.

Next, we estimate $\left\|\boldsymbol{\delta}^{(0)} - \boldsymbol{\delta}^*\right\|^2$. For this, we give an argument based on the convexity of the objective function. Intuitively, the argument goes as follows: If $\boldsymbol{\delta}^*$ and $\boldsymbol{\delta}^{(0)}$ are very far apart, then their function values must also be very far apart. Otherwise, by convexity of the objective function there exists a point in between which has a lower function value than the optimum, which cannot happen. We can therefore bound the distance between $\boldsymbol{\delta}^*$ and $\boldsymbol{\delta}^{(0)}$ by this difference in function value. While we do not know the function value of the optimum, we can estimate it by realising that it must be larger than or equal to the optimal function value for $m-1$. After all, the CvM statistic is a sum of non-negative terms, so adding another term will never decrease the value. We now inductively assume we have already found the solution for $m-1$ up to an accuracy of ϵ, called $\boldsymbol{\delta}_{prev}$. We can then estimate $\left\|\boldsymbol{\delta}^{(0)} - \boldsymbol{\delta}^*\right\|^2$.

In Appendix A.4 we show the objective function g is convex. In particular, for any $\lambda \in (0,1)$, $\mathbf{y}$ and $\hat{\mathbf{y}}$, we have that $g(\lambda\mathbf{y} + (1-\lambda)\hat{\mathbf{y}}) = \lambda g(\mathbf{y}) + (1-\lambda)g(\hat{\mathbf{y}}) - \lambda(1-\lambda)\left\|\mathbf{y} - \hat{\mathbf{y}}\right\|^2$. Since the argument does not depend on the summation bounds, the same holds for g_m^- as well. If we now plug in $\mathbf{y} = \mathbf{y}^{(0)}$ and $\hat{\mathbf{y}} = \mathbf{y}^*$, we find that:

$$\begin{aligned}&\lambda g_m^-(\mathbf{y}^{(0)}) + (1-\lambda)g_m^-(\mathbf{y}^*) - \lambda(1-\lambda)\left\|\mathbf{y}^{(0)} - \mathbf{y}^*\right\|^2 \overset{\circ}{\geq} g_m^-(\mathbf{y}^*) \\ &\lambda g_m^-(\mathbf{y}^{(0)}) - \lambda g_m^-(\mathbf{y}^*) \geq \lambda(1-\lambda)\left\|\mathbf{y}^{(0)} - \mathbf{y}^*\right\|^2 \\ &g_m^-(\mathbf{y}^{(0)}) - g_m^-(\mathbf{y}^*) \geq (1-\lambda)\left\|\mathbf{y}^{(0)} - \mathbf{y}^*\right\|^2,\end{aligned}$$

Where $(\circ)$ holds by optimality of $\mathbf{y}^*$. But if this holds for any $\lambda \in (0,1)$, then in particular for $\lambda \to 0^+$, we have $\left\|\mathbf{y}^{(0)} - \mathbf{y}^*\right\|^2 \leq g_m^-(\mathbf{y}^{(0)}) - g_m^-(\mathbf{y}^*) \leq g_m^-(\mathbf{y}^{(0)}) - (g_m^-(\mathbf{y}_{prev}) - \epsilon)$.

We now have an upper bound on the distance between the initial and optimal $\mathbf{y}$, but we need the distance between the initial and optimal $\boldsymbol{\delta}$. However, since $\boldsymbol{\delta} = M\mathbf{y}$ for the transition matrix M given in Section 3.3.1, we can estimate $\left\|\boldsymbol{\delta}^{(0)} - \boldsymbol{\delta}^*\right\|^2 \leq |M|^2 \left\|\mathbf{y}^{(0)} - \mathbf{y}^*\right\|^2 \leq 4\left\|\mathbf{y}^{(0)} - \mathbf{y}^*\right\|$. The proof that this matrix norm is bounded above by 2 can be found in Appendix A.5.

Substituting the estimate for L and the estimate for $\left\|\boldsymbol{\delta}^{(0)} - \boldsymbol{\delta}^*\right\|^2$ into equation (7), we find

$$f(\boldsymbol{\delta}^{(k)}) - f(\boldsymbol{\delta}^*) \leq \frac{8\sqrt{\frac{4}{3}m^4 + 2m^3 + \frac{2}{3}m^2} \cdot (f(\boldsymbol{\delta}^{(0)}) - f(\boldsymbol{\delta}_{prev}) + \epsilon)}{(k+1)^2}, \forall_{k\in\mathbb{N}}.$$

We want the right-hand side of this inequality to be below ϵ. We then find

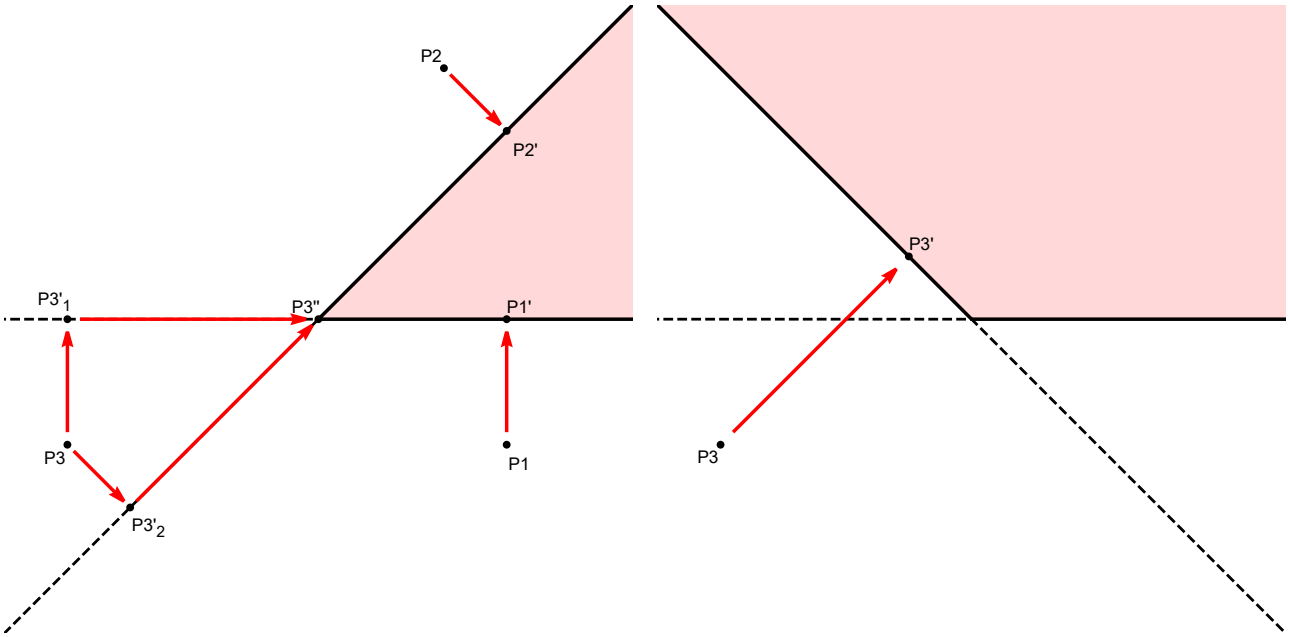
$$k+1 \geq \frac{\sqrt[4]{\frac{4}{3}m^4 + 2m^3 + \frac{2}{3}m^2} \cdot \sqrt{8(f(\boldsymbol{\delta}^{(0)}) - f(\boldsymbol{\delta}_{prev}) + \epsilon)}}{\sqrt{\epsilon}} =: k_{min} + 1.$$

This gives us a way to estimate the minimum number of time-steps $\lceil k_{min} \rceil$, depending on m, ϵ, the previous solution and the current guess, to reach this accuracy of ϵ. Note this error will eventually

appear in both the convex and concave fits. The final solution is therefore determined up to an accuracy of 2ϵ.

3.3.3 Evaluating the prox operator

The final challenge in applying projected gradient descent is evaluating the projection operator. In general, this is a difficult step, even for linearly constrained sets. However, projection greatly simplifies if the (outward pointing) normal vectors of the various constraint planes point in opposite directions, in the sense that their inner product is always non-positive. In this case, any constraint that is *violated* in the initial configuration will always be a *binding* constraint after projection. If any two such normal vectors have a positive inner product, this is not necessarily the case. This is shown visually in $\mathbb{R}^2$ in Figure 4. By Riesz representation theorem, any linear functional on $\mathbb{R}^m$ can be rewritten as an inner product with some constraint vector $\mathbf{n}_i$ [14, p.188], meaning the constraints can be rewritten as $\mathbf{n}_i \cdot \boldsymbol{\delta} \leq 0$. These constraint vectors correspond to the outward pointing normal vectors of the constraint planes, on which the inequalities in the constraints are replaced with equality.



(a) Projection in the case of constraint vectors with negative inner products.

(b) Projection in the case of constraint vectors with positive inner products.

Figure 4: The projection of points P1, P2, and P3 into the feasible set (the pink regions).

In Figure 4a, P1 only violates the horizontal constraint, and after projection onto the horizontal plane to P1′, it satisfies all constraints. P2 is similar. P3 violates both constraints, and can firstly be projected onto either the diagonal plane or the horizontal plane, to $\mathrm{P3}'_1$ or $\mathrm{P3}'_2$. After this projection, $\mathrm{P3}'_1$ still violates the diagonal constraint and $\mathrm{P3}'_2$ still violates the horizontal constraint. If we restrict these points to their respective planes while projecting it onto the other plane, we arrive at P3″, which satisfies all constraints. Indeed, P1, P2 and P3 are all projected to the closest point in the feasible set. A similar strategy generally does not work if at least one pair of constraint vectors have a positive inner product, as in Figure 4b. After all, if one starts with projecting P3 onto the horizontal plane, and then onto the diagonal plane, one will again arrive in the origin, which is not the point in the feasible set closest to P3. In particular, if at least one pair of constraint vectors has a positive inner product, the order the projections are performed in plays a role. This is not the case if all inner products are non-positive. As the number of constraints increases, the number of possible orders increases quickly, and the procedure becomes unpractical to use in the former case.

This leads to a simple projection algorithm.

Projection in case of non-positive inner products.

Input: $l \in \mathbb{Z}_{\geq 0}$ constraint vectors $\{\mathbf{n}_i\}_{i=1}^{l} \in \mathbb{R}^m$, with $\mathbf{n}_i \cdot \mathbf{n}_j \leq 0$ for $i \neq j$, target vector $\mathbf{v}^0 \in \mathbb{R}^m$

Output: The vector $\mathbf{v}$ that minimises the Euclidean distance to $\mathbf{v}^0$ among all vectors for which $\mathbf{n}_i \cdot \mathbf{v} \leq 0$ for all i

$\mathbf{v} \leftarrow \mathbf{v}^0$

while $\exists_{\mathbf{n}_i}$ *with* $\mathbf{n}_i \cdot \mathbf{v} > 0$ **do**

  Subtract $\mathbf{n}_i$ from $\mathbf{v}$ until $\mathbf{n}_i \cdot \mathbf{v} = 0$ (i.e. $\mathbf{v} \leftarrow \mathbf{v} - \lambda \mathbf{n}_i$, where $\lambda > 0$ is chosen such that (before assignment) $\mathbf{n}_i \cdot (\mathbf{v} - \lambda \mathbf{n}_i) = 0$).

  Orthogonally project all $\mathbf{n}_j$ with $j \neq i$ onto $\langle \mathbf{n}_i \rangle^{\perp}$ $\left(\text{i.e. } \mathbf{n}_j \leftarrow \mathbf{n}_j - \frac{\mathbf{n}_i \cdot \mathbf{n}_j}{\|\mathbf{n}_i\|^2} \mathbf{n}_i\right)$

end

return $\mathbf{v}$

The details of this algorithm, and the proof that it gives the correct result can be found in Appendix A.6. Intuitively, this algorithm performs the procedure shown in Figure 4*a*, except that we do not limit ourselves to $\mathbb{R}^2$ or only two constraints.

After passing through the while loop for $\mathbf{n}_i$, we say constraint i has been enforced.

This algorithm allows us to perform projections very efficiently. As explained in Section 3.3.1, the constraints are given by $\frac{\delta_{i+1}}{X_{(i+1)} - X_{(i)}} \geq \frac{\delta_i}{X_{(i)} - X_{(i-1)}}$ for $2 \leq i \leq m-1$. Bringing the right-hand side of this equation to the left and multiplying by -1, we find the constraints to be given by

$$(..., 0, \frac{1}{X_{(i)} - X_{(i-1)}}, \frac{-1}{X_{(i+1)} - X_{(i)}}, 0, ...) \cdot \boldsymbol{\delta} \leq 0.$$

In other words, let $(\mathbf{n}_i)_j$ denote the j-th entry of the i-th constraint vector. These entries are then given by

$$(\mathbf{n}_i)_j = \left\{ \begin{array}{ll} \frac{1}{X_{(i)} - X_{(i-1)}}, & \text{if } j = i \\ \frac{-1}{X_{(i+1)} - X_{(i)}}, & \text{if } j = i+1 \\ 0, & \text{otherwise} \end{array} \right\}. \tag{8}$$

We thus see that a constraint vector $\mathbf{n}_i$ has an inner product of zero with every other constraint vector, except with $\mathbf{n}_{i-1}$ and $\mathbf{n}_{i+1}$, with which it has a negative inner product. This means we can apply the projection algorithm[9]. For the sake of clearer writing, suppose the constraints are given by $w_i \delta_i \leq w_{i+1} \delta_{i+1}$, with $w_i = \frac{1}{X_{(i)} - X_{(i-1)}}$ for $1 \leq i < m$. This makes the constraint vectors of the form $\mathbf{n}_i = (..., 0, w_i, -w_{i+1}, 0, ...)$. We now want to know how these constraint vectors change after going through the algorithm several times, and the constraint vectors have been projected onto the orthogonal complements of several other constraint vectors. This is given by the following theorem.

Theorem 3.7. *Let i be any index for which constraint i has not been enforced. Let L_i (possibly 0) be the largest integer such that constraints $i - L_i$ through $i - 1$ have all been enforced. Similarly, let R_i (possibly 0) be the largest number such that constraints $i + 1$ through $i + R_i$ have all been enforced. By construction, $L_{i+R_i+1} = R_i$. We show $\mathbf{n}_i$ then takes the following form:*

$$(\mathbf{n}_i)_j = \left\{ \begin{array}{ll} \frac{1}{w_j \sum_{k=i-L_i}^{i} \frac{1}{w_k^2}}, & \textit{if } i - L_i \leq j \leq i \\ \frac{-1}{w_j \sum_{k=i+1}^{i+1+R_i} \frac{1}{w_k^2}}, & \textit{if } i+1 \leq j \leq i+1+R_i \\ 0, & \textit{otherwise} \end{array} \right\}. \tag{9}$$

Proof. We show this holds using an inductive proof. Initially, $L_i = R_i = 0$ for all i, since none of the constraints have been enforced yet. Substituting $L_i = R_i = 0$ into equation (9), we find the form given by equation (8).

[9] In the original parameters, the constraint vectors $\mathbf{n}_i$ and $\mathbf{n}_{i+2}$ always have a positive inner product. This is an important reason why it is helpful to reparametrize the problem.

Now suppose all constraint vectors are of the form given by equation (9). Suppose constraint i has not been enforced yet. When we enforce constraint i, L_{i+R_i+1} will change from R_i to $R_i + L_i + 1$. Similarly, R_{i-L_i-1} will change from L_i to $R_i + L_i + 1$. We thus need to show that after orthogonal projection onto $\langle \mathbf{n}_i \rangle^\perp$, $\mathbf{n}_{i-L_i-1}$ and $\mathbf{n}_{i+R_i+1}$ are updated to the form predicted by equation (9). We now show $\mathbf{n}_{i-L_i-1}$ is updated to the correct form, since $\mathbf{n}_{i+R_i+1}$ is analogous.

We begin by calculating the norm of $\mathbf{n}_i$. We then find:

$$\begin{aligned}\|\mathbf{n}_i\|^2 &= \sum_{j=i-L_i}^{i} \frac{1}{w_j^2 (\sum_{k=i-L_i}^{i} \frac{1}{w_k^2})^2} + \sum_{j=i+1}^{i+1+R_i} \frac{1}{w_j^2 (\sum_{k=i+1}^{i+1+R_i} \frac{1}{w_k^2})^2} \\ &= \frac{1}{\sum_{k=i-L_i}^{i} \frac{1}{w_k^2}} + \frac{1}{\sum_{k=i+1}^{i+1+R_i} \frac{1}{w_k^2}}.\end{aligned}$$

By a similar computation, we find that $\mathbf{n}_i \cdot \mathbf{n}_{i-L_i-1} = \frac{-1}{\sum_{k=i-L_i}^{i} \frac{1}{w_k^2}}$.

If we now project $\mathbf{n}_{i-L_i-1}$ onto $\langle \mathbf{n}_i \rangle^\perp$, we find that for $i - L_i \leq j \leq i$, $(\mathbf{n}_{i-L_i-1})_j$ becomes:

$$\begin{aligned}(\mathbf{n}_{i-L_i-1})_j &= \frac{-1}{w_j \sum_{k=i-L_i}^{i} \frac{1}{w_k^2}} + \frac{\frac{1}{\sum_{k=i-L_i}^{i} \frac{1}{w_k^2}}}{\frac{1}{\sum_{k=i-L_i}^{i} \frac{1}{w_k^2}} + \frac{1}{\sum_{k=i+1}^{i+1+R_i} \frac{1}{w_k^2}}} \cdot \frac{1}{w_j \sum_{k=i-L_i}^{i} \frac{1}{w_k^2}} \\ &= \frac{-\frac{1}{\sum_{k=i+1}^{i+1+R_i} \frac{1}{w_k^2}}}{\frac{1}{\sum_{k=i-L_i}^{i} \frac{1}{w_k^2}} + \frac{1}{\sum_{k=i+1}^{i+1+R_i} \frac{1}{w_k^2}}} \cdot \frac{1}{w_j \sum_{k=i-L_i}^{i} \frac{1}{w_k^2}}.\end{aligned}$$

Multiplying the numerator and denominator by $\left(\sum_{k=i-L_i}^{i} \frac{1}{w_k^2}\right) \cdot \left(\sum_{k=i+1}^{i+1+R_i} \frac{1}{w_k^2}\right)$

$$\begin{aligned}(\mathbf{n}_{i-L_i-1})_j &= \frac{-1}{w_j \left(\sum_{k=i-L_i}^{i} \frac{1}{w_k^2} + \sum_{k=i+1}^{i+1+R_i} \frac{1}{w_k^2}\right)} \\ &= \frac{-1}{w_j \sum_{k=i-L_i}^{i+1+R_i} \frac{1}{w_k^2}}.\end{aligned}$$

For $i + 1 \leq j \leq 1 + R_i$, $(\mathbf{n}_{i-L_i-1})_j$ instead becomes:

$$\begin{aligned}(\mathbf{n}_{i-L_i-1})_j &= 0 + \frac{\frac{1}{\sum_{k=i-L_i}^{i} \frac{1}{w_k^2}}}{\frac{1}{\sum_{k=i-L_i}^{i} \frac{1}{w_k^2}} + \frac{1}{\sum_{k=i+1}^{i+1+R_i} \frac{1}{w_k^2}}} \cdot \frac{-1}{w_j \sum_{k=i+1}^{i+1+R_i} \frac{1}{w_k^2}} \\ &= \frac{-1}{w_j \sum_{k=i-L_i}^{i+1+R_i} \frac{1}{w_k^2}}.\end{aligned}$$

The rest of $\mathbf{n}_{i-L_i-1}$ remains unchanged. After projecting orthogonally onto $\langle \mathbf{n}_i \rangle^\perp$, we therefore indeed find the form given by equation (9). □

While the constraint vectors initially contain only two nonzero entries, allowing us to calculate inner products very quickly, the number of nonzero entries increases after several orthogonal projections, as shown by equation (9). This increases the complexity of calculating the inner product, slowing down the algorithm. However, by transforming $\boldsymbol{\delta}$ in a smart way, we can remedy this.

The algorithm tells us that in the first iteration, we need to subtract the constraint vector $\mathbf{n}_i$, given by equation (8) from $\boldsymbol{\delta}$ until $w_i\delta_i = w_{i+1}\delta_{i+1}$. Equivalently, we can define $\boldsymbol{\delta'}$ by $(\boldsymbol{\delta'})_i = w_i\delta_i$ for all $1 \leq i \leq m$, and $\mathbf{n}'_i$ by

$$(\mathbf{n}'_i)_j = w_j(\mathbf{n}_i)_j = \left\{ \begin{array}{ll} \frac{1}{\sum_{k=i-L_i}^{i} \frac{1}{w_k^2}}, & \text{if } i - L_i \leq j \leq i \\ \frac{-1}{\sum_{k=i+1}^{i+1+R_i} \frac{1}{w_k^2}}, & \text{if } i+1 \leq j \leq i+1+R_i \\ 0, & \text{otherwise} \end{array} \right\}$$

In the first iteration, we then need to subtract $\mathbf{n}'_i$, with $L_i = R_i = 0$, from $\boldsymbol{\delta'}$ until $\delta'_i = \delta'_{i+1}$. This is equivalent to replacing both δ'_i and δ'_{i+1} by their weighted average, where the weights are given by $\frac{1}{w_i^2}$ and $\frac{1}{w_{i+1}^2}$ respectively. After doing this, δ'_i and δ'_{i+1} are equal, and remain equal since the constraint vectors now have equal entries for $j = i$ and $j = i+1$. Therefore, it is more efficient to only "track" and perform computations on one of these (say δ'_i), and neglect the other. Changing δ'_i can then be interpreted to represent changing both δ'_i and δ'_{i+1}. At the end of the process, we can then reconstruct the vector (i.e. set δ'_{i+1} and any others that are represented by δ'_i equal to δ'_i). If we neglect such negligible entries, $\mathbf{n}_i$ will always have exactly two non-negligible entries. Furthermore, note that these entries, given by $\frac{1}{\sum_{k=i-L_i}^{i} \frac{1}{w_k^2}}$ and $\frac{-1}{\sum_{k=i+1}^{i+1+R_i} \frac{1}{w_k^2}}$ do not need to be recomputed at every step: enforcing constraint i corresponds to taking the weighted average of two non-negligible entries, with the weights given by $\sum_{k=i-L_i}^{i} \frac{1}{w_k^2}$ and $\sum_{k=i+1}^{i+1+R_i} \frac{1}{w_k^2}$. These are simply the sums of the weights of all the entries the two non-negligible entries represent.

To perform the projection, we therefore first calculate $\boldsymbol{\delta'}$ from $\boldsymbol{\delta}$. This requires m multiplications. We then repeatedly take weighted averages between any pair of two adjacent non-negligible entries for which the former is smaller than the latter. After taking this average, we neglect one of these entries, and the other entry will represent it from that point on. Its weight will be updated to the combined weight of the pair. We repeat this process until the non-negligible entries of $\boldsymbol{\delta'}$ form a monotonically increasing sequence. After reconstructing $\boldsymbol{\delta'}$, we can again calculate $\boldsymbol{\delta}$ from $\boldsymbol{\delta'}$. This allows us to very efficiently perform the projection step.

3.3.4 Reducing the required number of steps at runtime

In Section 3.3.2, we explained how the factor $\left\|\boldsymbol{\delta}^{(0)} - \boldsymbol{\delta}^*\right\|^2$ is bounded above by $4(g_m^-(\boldsymbol{\delta}^{(0)}) - g_m^-(\boldsymbol{\delta}^*))$, where $g_m^-(\boldsymbol{\delta}^*)$ can in turn be estimated by the function value of $\boldsymbol{\delta}_{prev}$, the optimum for the previous value of m. In practice, this is a fairly crude lower bound: especially for large values of m, $g_m^-(\boldsymbol{\delta}^*)$ is often much larger than the estimated value, making us overestimate the required number of iterations to reach the desired accuracy level. However, this estimate can be improved at runtime. Suppose we run k iterations, which ensures an accuracy, based on the crude bound, of $g_m^-(\boldsymbol{\delta}^{(k)}) - g_m^-(\boldsymbol{\delta}^*) < q$ for some $q > 0$. Nevertheless, we see that $g_m^-(\boldsymbol{\delta}^{(k)}) - g_m^-(\boldsymbol{\delta}_{prev}) \gg q$. This means that $g_m^-(\boldsymbol{\delta}^*) \gg g_m^-(\boldsymbol{\delta}_{prev})$, and gives us a new, much less conservative estimate for $g_m^-(\boldsymbol{\delta}^{(0)}) - g_m^-(\boldsymbol{\delta}^*)$. This allows us to lower the number of iterations required, while maintaining the guaranteed level of accuracy ϵ. This idea is quantified in the following theorem.

Theorem 3.8 (Re-estimating the error)**.** *Let $\boldsymbol{\delta}^{(0)}$ be the initial vector, and $\boldsymbol{\delta}^{(k)}$ the vector after k iterations of the accelerated gradient descent procedure, with $\frac{8}{t(k+1)^2} < 1$. Then*
$g_m^-(\boldsymbol{\delta}^{(0)}) - g_m^-(\boldsymbol{\delta}^*) \leq \frac{g_m^-(\boldsymbol{\delta}^{(0)}) - g_m^-(\boldsymbol{\delta}^{(k)})}{1 - \frac{8}{t(k+1)^2}}$.

Proof.

$$\begin{aligned} g_m^-(\boldsymbol{\delta}^{(0)}) - g_m^-(\boldsymbol{\delta}^*) &= g_m^-(\boldsymbol{\delta}^{(0)}) - g_m^-(\boldsymbol{\delta}^{(k)}) + g_m^-(\boldsymbol{\delta}^{(k)}) - g_m^-(\boldsymbol{\delta}^*) \\ &\leq g_m^-(\boldsymbol{\delta}^{(0)}) - g_m^-(\boldsymbol{\delta}^{(k)}) + \frac{8(g_m^-(\boldsymbol{\delta}^{(0)}) - g_m^-(\boldsymbol{\delta}^*))}{t(k+1)^2}. \end{aligned}$$

Moving the rightmost term to the left, we then find that

$$\left(1 - \frac{8}{t(k+1)^2}\right)(g_m^-(\boldsymbol{\delta}^{(0)}) - g_m^-(\boldsymbol{\delta}^*)) \leq g_m^-(\boldsymbol{\delta}^{(0)}) - g_m^-(\boldsymbol{\delta}^{(k)}),$$

and the conclusion follows directly. □

The right-hand side of this inequality is entirely known at runtime, after performing k iterations. We can then substitute this expression into equation (7) to improve the estimate of the number of iterations we still need to perform in order to reach the desired accuracy level.

Note that this theorem only helps us if $(k+1)^2 \geq \frac{8}{t}$. In general, $t \approx \frac{1}{L}$ will be rather small. This means that the theorem will only help us after we have already performed a certain number of iterations. The theorem mainly prevents the procedure from performing a very large number of iterations with only marginal improvement, while maintaining a strong accuracy estimate and without having to rely on heuristics to decide when to stop iterating.

4 Numerical experimentation on the new methodology

In the previous section, a numerical method to optimise the modified dip statistic, based on the CvM statistic, was described. This was implemented in the programming language R[10]. In this section, we numerically assess the properties of this test. Firstly, we consider the calibration of the test. To this end, data is generated from unimodal distributions. We then visually compare the actual significance to the nominal significance level of the test. Secondly, we assess the power of the test if the null hypothesis is violated. This is done by generating data from a family of distributions, containing both a unimodal and several multimodal distributions, and determining the fraction of times the null-hypothesis is rejected by the test for these distributions. This is also compared to the existing tests introduced in Section 2. Finally, we assess the numerical performance of the test, quantifying how long it takes to perform the numerical optimisation for various distributions, and determining how this scales with the accuracy level ϵ and sample size N. While this section presents the results in the form of graphs, some of the collected data is also tabulated in Appendix B.

4.1 The calibration of the test

To test the calibration of the new methodology, three distributions are considered: the standard normal distribution, the standard uniform distribution and the Cauchy distribution. The normal distribution was chosen because of its ubiquity in daily life. The uniform distribution was chosen because it is on the boundary between the class of unimodal distributions and the class of multimodal distributions. Finally, the Cauchy distribution was chosen because it is rather poorly behaved. It does not have any well-defined moments (including a mean), and has very heavy tails, decaying as $\frac{1}{x^2}$. This can therefore be used to stress-test the method, and see how it behaves in unfavourable circumstances. For each of these distributions, the procedure was repeated 100 times on different data of size $N = 200$, sampled from the respective distribution. Each repetition, $n = 100$ bootstrap samples were generated, and a fit was performed at an accuracy level of $\epsilon = 10^{-6} \cdot N = 10^{-6} \cdot 200$. Since this accuracy level applies to both the convex part and the concave part, this ensures a final CvM error of at most $2 \cdot 10^{-6}$.

This accuracy level adds a certain amount of uncertainty when evaluating whether the CvM statistic of a bootstrap sample exceeds the CvM statistic of the original data, since both of these values are only approximated up to an accuracy of 2ϵ. We will call the approximations of the found dip[11] statistics $\hat{D}$ and $\hat{D}_i$. The most conservative way to evaluate the p-value, is to say that we consider the CvM statistic to exceed the CvM statistic of the original data if it is *possible* that it truly exceeds it based on the approximated fits and the accuracy level. This results in a reported p-value given by

$$p_{cons} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{\hat{D}_i \geq \hat{D} - 2\epsilon\}.$$

This generally results in slightly inflated reported p-values. Similarly, the most anti-conservative way to evaluate this is to say that we consider the CvM statistic to exceed the CvM statistic of the original data if it is *guaranteed* that it truly exceeds it. This results in a reported p-value given by

$$p_{anti-cons} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{\hat{D}_i - 2\epsilon \geq \hat{D}\}.$$

This generally results in slightly deflated reported p-values. Finally, the most natural way to evaluate this is to simply compare the approximated values. After all, there is no reason to assume the fitting procedure systematically favours either the original data or bootstrap samples. This results in a p-value given by

$$p_{natural} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{\hat{D}_i \geq \hat{D}\}.$$

[10] For efficiency, the projection algorithm was implemented in C++, using the Rcpp library.

[11] The modified dip statistic, based on the CvM statistic.

From this, the correspondence of the reported p-values to the actual significance level of the test can be determined. We estimate this actual significance level by the fraction of repetitions where the null-hypothesis is rejected at the nominal significance level. As described in Section 2.1.2, the reported p-value should ideally be uniformly distributed between 0 and 1 under the null. This corresponds to the actual significance level matching the nominal significance level. If the actual significance level is (mostly) smaller than the nominal significance level, the test is conservative. Conversely, if the actual significance level is (mostly) larger than the nominal significance level, the test is anti-conservative. The results are shown in Figure 5.

For the normal distribution, the test is somewhat conservative. This is to be expected: the normal distribution has a very strong mode. Since the distribution $\tilde{F}$ is found by minimising the distance to the (by definition multimodal) ECDF, it will generally have a weaker mode than the original normal distribution. It is therefore not surprising that generally, ECDFs of the bootstrap samples will be at a larger distance to their respective fitted distributions than the ECDF of the original data is to $\tilde{F}$ with respect to the CvM statistic. For the uniform distribution, the test is very well calibrated, being slightly anti-conservative. Applying similar reasoning, this is again to be expected. Since the uniform distribution has no true modes, $\tilde{F}$ cannot have a weaker mode than the original uniform distribution, and may, by chance, have a stronger mode. We therefore observe the opposite effect we saw when considering the normal distribution. Finally, the test is extremely conservative for the Cauchy distribution. For instance, the median of the observed p-values is 0.76. While performing the test, it was noticed that the heavy tails often caused a small number of points to be chosen far away from the mode. This often caused even heavier tails in the fitted distribution $\tilde{F}$, resulting in very large distances between the ECDFs of the bootstrap samples and their respective fitted distributions. This shows that the outcome of this test should be interpreted with care, especially when it is plausible the data comes from a somewhat poorly behaved distribution.

Comparing the calibration of the new methodology to that of the CvM string test, it is clear that the new methodology is less well calibrated. For the Cauchy distribution in particular, the difference in quality of calibration is especially large. For the other distributions, the calibration figures are more similar, though a significant difference is still clearly visible.

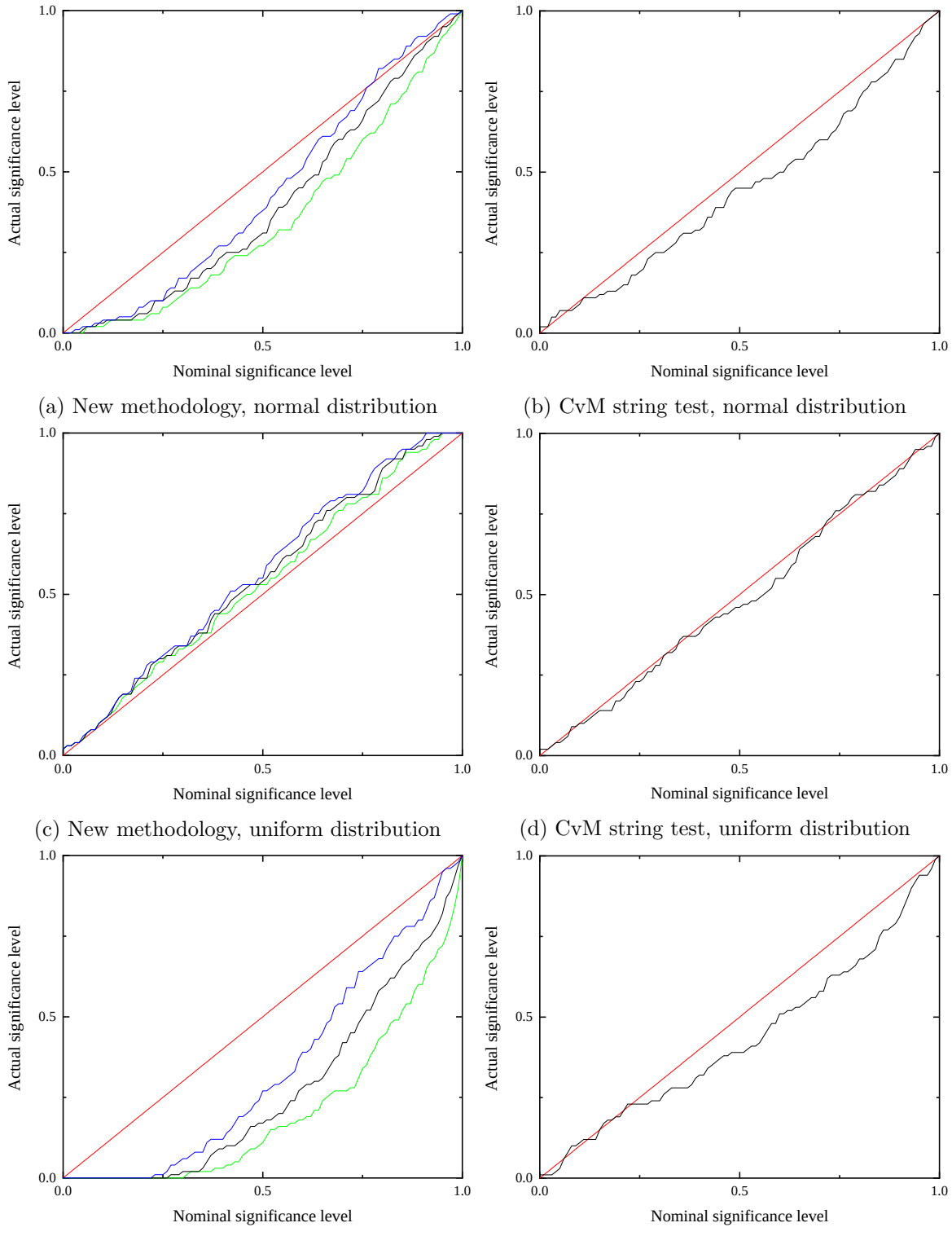
4.2 The power of the test

Next, we wish to assess the power of the test. For this, we consider the family of Gaussian mixtures with the density functions given by:

$$f_\mu(x) = \frac{7}{10}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right) + \frac{3}{10}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2}\right) \tag{10}$$

This is a mixture of two normal distributions, of weights $\frac{7}{10}$ and $\frac{3}{10}$ respectively, where both have a standard deviation of 1, and the means are at 0 and μ respectively. This distribution is chosen since it is also used[12] in [9] and [15]. This allows the reader to compare results from this simulation study to the results in these sources. For μ, the values 2, 3, 3.25, 3.5, 3.75, 4, 4.25, 4.5, and 4.75 were used. The density functions of these distributions are shown in Figure 6.

[12] [9] estimates the distribution from figures in [15], since [15] does not explicitly describe the distributions that are used.



(a) New methodology, normal distribution

(b) CvM string test, normal distribution

(c) New methodology, uniform distribution

(d) CvM string test, uniform distribution

(e) New methodology, Cauchy distribution

(f) CvM string test, Cauchy distribution.

Figure 5: The actual significance level compared to the nominal significance level of the test for the new methodology and the CvM string test, when a sample of size $N = 200$ is drawn from the respective distributions. For the new methodology, the three curves indicate the *convervative* approach (green), *natural* approach (black) and the *anti-conservative* approach (blue). Note that these figures are equivalent to the ECDFs of the reported p-values. The red line is a reference line, indicating perfect calibration.

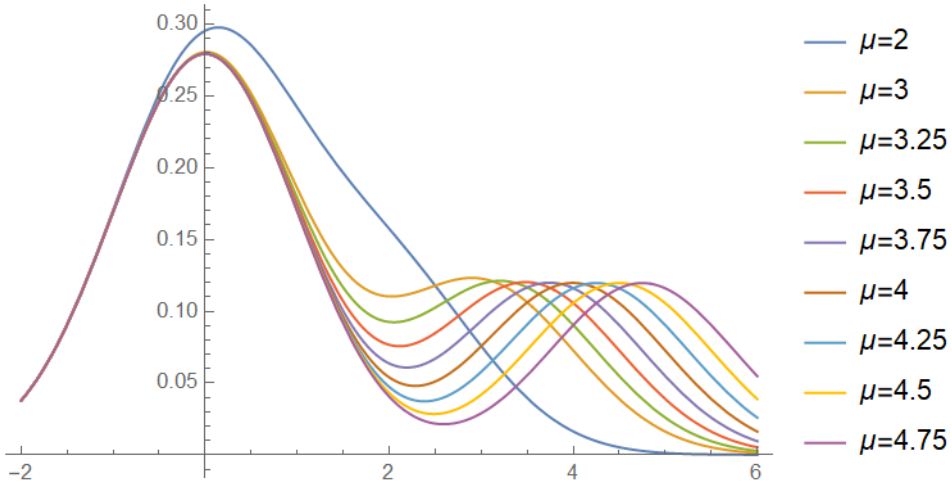
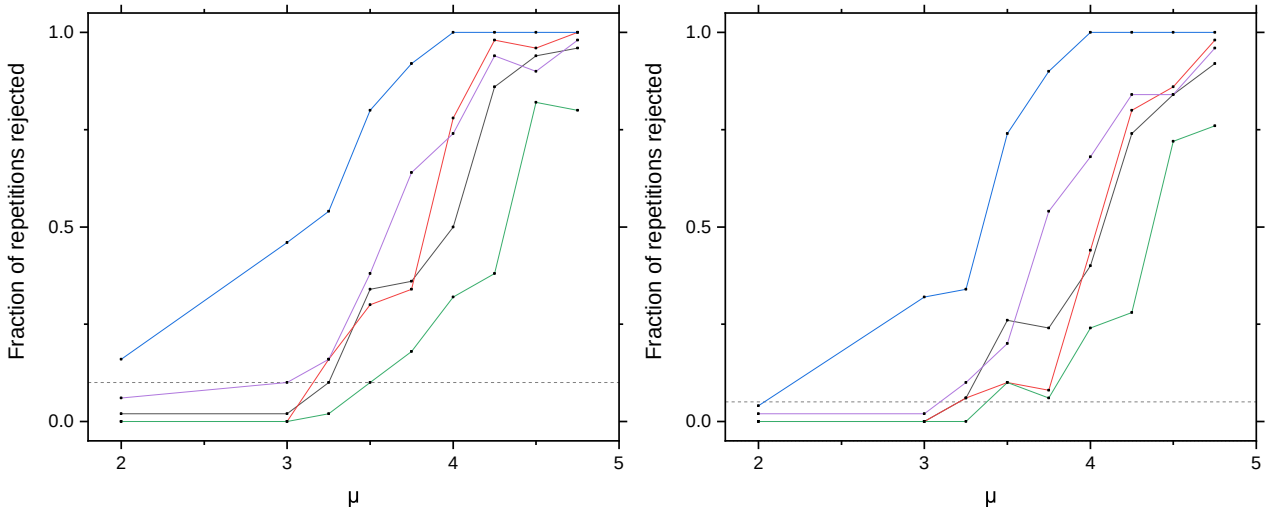


Figure 6: The density functions of the used Gaussian mixtures.

Only the distribution with $\mu = 2$ is unimodal. As μ increases, the second mode becomes increasingly pronounced, and we therefore expect the power to increase as μ increases.

For each of these distributions, the test was repeated 50 times. As in Section 4.1, for each repetition a sample of size $N = 200$ was drawn, and $n = 100$ bootstrap samples were generated. The fits were performed at an accuracy level of $\epsilon = 10^{-6} \cdot N$, and the natural evaluation of the p-value was used. We also perform the dip test from [7], the CvM string test from [9], Silverman's test from [2], and the improved Silverman's test from [5], using the "multimode" R library. Note that each test was performed on freshly generated data, and it is therefore possible that certain tests received more favourable data. However, since each test was repeated 50 times, we expect this effect to be limited in influence.

In Figure 7, the fraction of times the various tests rejected the null hypothesis at significance levels $\alpha = 0.1$ and $\alpha = 0.05$ is plotted.



(a) $\alpha = 0.1$ (b) $\alpha = 0.05$

Figure 7: A plot of the power of the original Silverman's test from [2] (red), the improved Silverman's test from [5] (blue), the original dip test from [7] (green), the CvM string test from [9] (purple), and the new methodology (black) at significance levels $\alpha = 0.1$ and $\alpha = 0.05$, when a sample of size $N = 200$ is drawn from density (10). The dotted lines indicate the significance levels $y = 0.1$ and $y = 0.05$ respectively.

From these figures, it is clear that the new methodology is significantly less powerful than the improved Silverman's test introduced in [5] when applied to data from (10). When $\mu = 2$ and $\alpha = 0.1$, the improved Silverman's test does appear to have a higher power than the significance level though, meaning it might be anti-conservative under these circumstances. Interestingly, under these circumstances, the power of the new methodology is very similar to the power of Silverman's original test. However, since Silverman's test and the new methodology use inherently different characteristics of the data to assess multimodality, it is unclear if this also extends to other families of distributions. Furthermore, since Silverman's test uses Gaussian kernels to estimate the density, it is plausible that Silverman's test is uniquely well-suited for the Gaussian mixture given by equation (10).

It is therefore more interesting to compare the power of the new methodology to the dip test and string test. For both confidence levels, the new methodology is consistently more powerful than the dip tests, by a significant margin. However, the new methodology is still fairly limited in power. For the (multimodal) distribution with $\mu = 3$, not a single repetition resulted in a p-value below 0.1, the lowest reported p-value being 0.14. This is expected from the conservativeness observed for the normal distribution in Section 4.1. It is also clear that the new methodology has low power compared to the CvM string test. From the figures it appears that this is mostly the case in the region $\mu \in [3.5, 4.25]$, though this should be taken with a grain of salt: Since each experiment was only repeated 50 times, random variations could play a large role when comparing single measurements.

Figure 7 quantifies the power based on a hard cut-off value: it shows the fraction of repetitions for which the p-value is less than α. However, it is also interesting to consider the distribution of the reported p-values of the various tests, in order to gain insight into *how much* the reported p-values generally differ from α. To this end, we can again plot the ECDFs of the reported p-values, as was done in Figure 5. This is shown for the distribution with $\mu = 3.5$ in Figure 8.

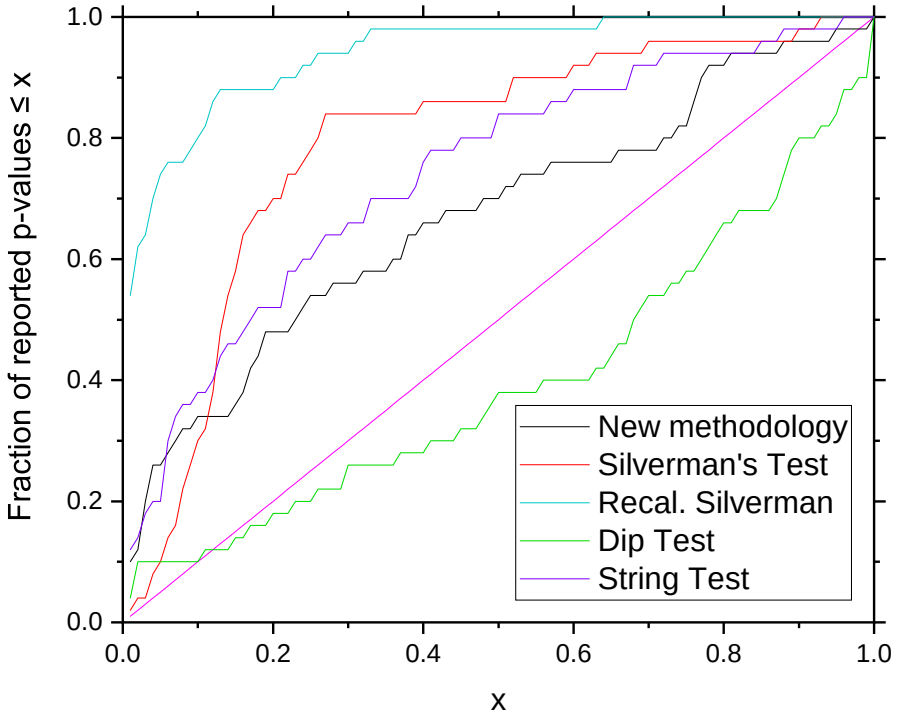


Figure 8: The ECDFs of the reported p-values of the five different tests, for data from density (10) with $\mu = 3.5$. The magenta line is a reference line.

We observe that for each test except the dip test, the ECDF-curve is far above the reference line. This means that for any significance level α, $\mathbb{P}(p \leq \alpha) \gg \alpha$. The test is therefore likely to reject the null-hypothesis for this multi-modal distribution, and thus give the correct result. For the dip test, this is not the case. This again shows that the dip test is very conservative. Comparing the new

methodology and the CvM string test, we again see that the new methodology is less powerful than the string test at almost all significance levels.

4.3 Performance of the numerical optimisation

In the previous sections, the calibration and power of the test was assessed. Finally, we need to determine if the test is practical to use. For this, it is important to quantify how long the fitting procedure takes. In this section, we assess the performance of this optimisation procedure. In particular, we wish to determine how long the optimisation procedure takes for a single dataset, which can either be the measured data or a bootstrap sample, and how this varies with the number of data points N, the accuracy level ϵ and the distribution the data points are drawn from.

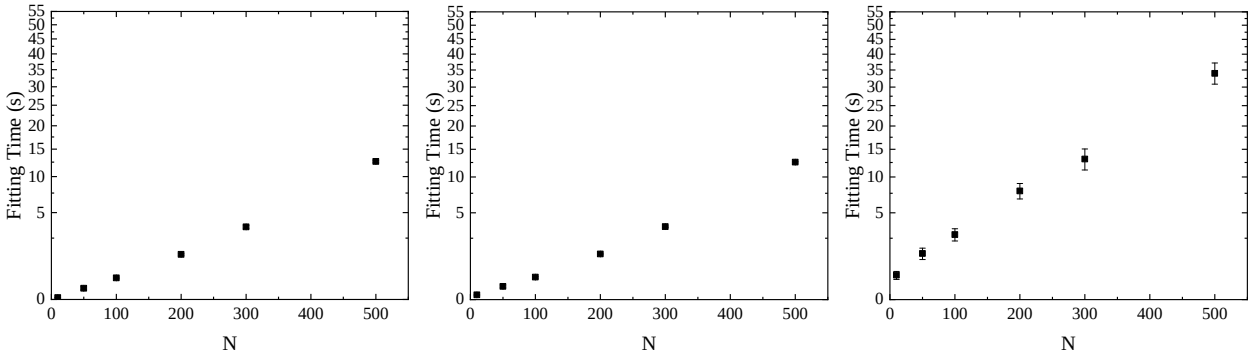
To this end, samples were drawn from four distributions:

1. The standard normal distribution
2. A Gaussian mixture, consisting of two normal distributions with equal weight, with means $\mu_1 = 0, \mu_2 = 4$, and standard deviations of $\sigma_1 = \sigma_2 = 1$
3. The standard uniform distribution
4. The standard exponential distribution

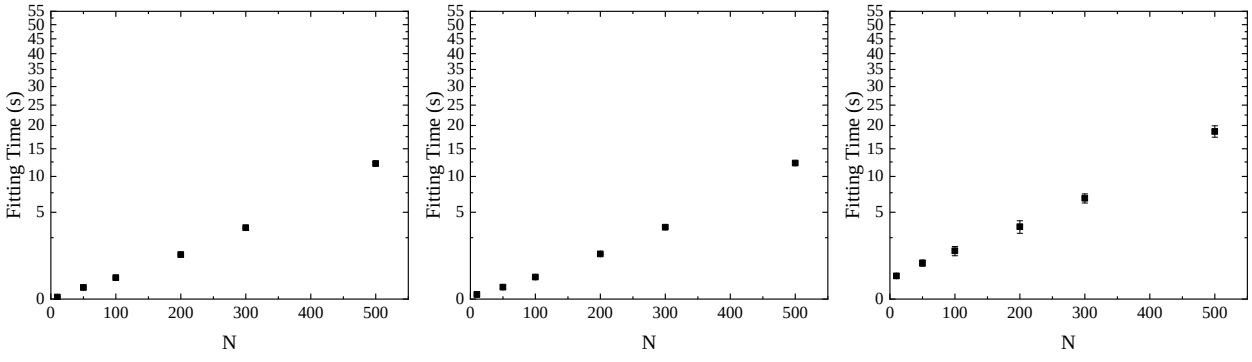
The standard normal distribution was chosen since it has a strong, central mode. This gives insight into how the procedure performs on (near) unimodal samples. Conversely, the Gaussian mixture was chosen because it has two strong modes. This gives insight into how the procedure performs on (near) multimodal samples. The uniform distribution was chosen because while it is technically unimodal, its mode is very weak. This gives insight into how the procedure performs with samples that are neither obviously unimodal nor obviously multimodal. Finally, the exponential distribution was chosen because of its asymmetry: the other three distributions are all symmetric, meaning that fitting the convex part and fitting the concave part are essentially equivalent. The exponential distribution, on the other hand, is very asymmetric, with its mode lying on the leftmost point of its support. This gives insight into how the procedure works with asymmetric samples. Since the distribution itself is concave everywhere, we may expect that fitting the concave part is easier than fitting the convex part for this distribution, and therefore takes less time than fitting the convex part.

For each of these four distributions, the procedure was performed for the sample sizes $N = 10$, $N = 50$, $N = 100$, $N = 200$, $N = 300$, and $N = 500$, at three accuracy levels $\epsilon = 10^{-4} \cdot N$, $\epsilon = 10^{-6} \cdot N$ and $\epsilon = 10^{-9} \cdot N$. These accuracy levels ensure a maximum error in the CvM statistic, given by equation (4), of $2 \cdot 10^{-4}$, $2 \cdot 10^{-6}$ and $2 \cdot 10^{-9}$ respectively. For each of these combinations, the procedure was performed a total of fifteen times. This was done to be able to assess the spread in the time taken. The time these fits took was measured using the built-in function *Sys.time()*[13]. The results are shown in Figure 9.

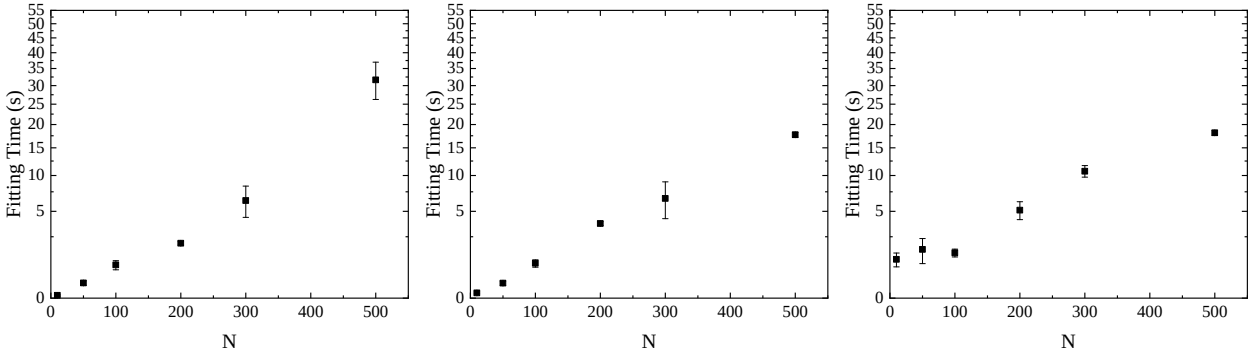
[13] All tests were performed on a Lenovo Thinkpad P1. Performance may vary depending on the hardware used.



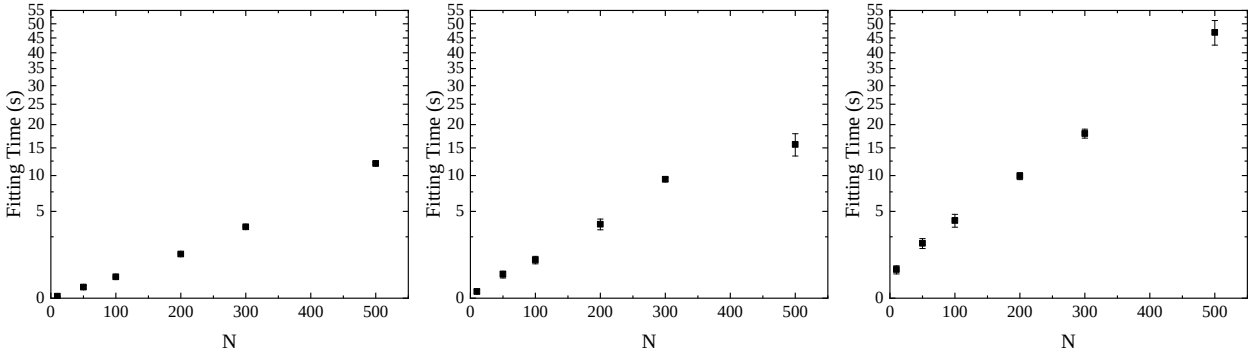
(a) Normal, $\epsilon = 10^{-4} \cdot N$ (b) Normal, $\epsilon = 10^{-6} \cdot N$ (c) Normal, $\epsilon = 10^{-9} \cdot N$



(d) Mixture, $\epsilon = 10^{-4} \cdot N$ (e) Mixture, $\epsilon = 10^{-6} \cdot N$ (f) Mixture, $\epsilon = 10^{-9} \cdot N$



(g) Uniform, $\epsilon = 10^{-4} \cdot N$ (h) Uniform, $\epsilon = 10^{-6} \cdot N$ (i) Uniform, $\epsilon = 10^{-9} \cdot N$



(j) Exponential, $\epsilon = 10^{-4} \cdot N$ (k) Exponential, $\epsilon = 10^{-6} \cdot N$ (l) Exponential, $\epsilon = 10^{-9} \cdot N$

Figure 9: Plots of the mean time taken to perform the optimisation procedure for a single dataset against the size of the dataset, for various distributions and accuracy levels. Error bars indicate the sample standard deviation. Note the quadratic scaling on the y-axis.

In most plots, we observe that the points approximately lie on a straight line. Due to the quadratic scaling of the y-axis, this means the fitting time scales approximately as N^2.

For the normal distribution, the time taken to perform the fits is extremely similar for the two weaker accuracy levels. At first, this may seem paradoxical, since the maximum error differs by a factor of 100. A possible explanation for this is that the procedure uses the solution for the previous value of m as a starting point for the next value of m. If ϵ is taken to be smaller, the fit for the previous value of m will be closer to the optimum. This effect seems to largely cancel out the effect of decreasing ϵ to $10^{-6} \cdot N$. For these accuracy levels, the spread is also very small, causing the error bars to be barely visible. However, when ϵ decreases further to $10^{-9} \cdot N$, the time taken increases considerably, approximately by a factor of 3. The spread also becomes more noticeable.

Interestingly, for the Gaussian mixture, the results are very similar to the results for the standard normal distribution for the accuracy levels $\epsilon = 10^{-4} \cdot N$ and $\epsilon = 10^{-6} \cdot N$. This may be surprising: the normal distribution is unimodal, whereas the Gaussian mixture has two strong modes. For the accuracy level $\epsilon = 10^{-9} \cdot N$, the fits are performed about 50% faster than for the normal distribution. We hypothesise this is caused by the following: for the normal distribution, the first half of the convex fits will most likely be easy to perform, since this is the part where the distribution itself is convex. After $m \approx \frac{N}{2}$, however, these convex fits take increasingly more time to perform, since the distribution can no longer be approximated well by a convex function. Especially for $\epsilon = 10^{-9}$, this takes a large amount of additional computation time. A similar argument holds for the concave fits. The first m points from the Gaussian mixture, however, will be closer to being fully convex for large values of m. This saves time, compared to the normal distribution, causing the fits to be performed faster.

For the uniform distribution, the spreads are generally larger compared to the previous distributions and seem to differ erratically with the sample size, especially for the weaker accuracy levels. This can possibly be explained by the following: Both the normal and mixed Gaussian distributions have strong modes. This means that clusters in the data are largely determined by these modes, and the number of clusters and the number of points per cluster is therefore somewhat predictable. For the uniform distribution, this is not the case. Clusters in the data are completely determined by randomness in the specific sample, and the number of clusters and their height is more unpredictable. This likely causes a larger spread in the measured fitting times. Another, somewhat paradoxical observation is that for the weakest accuracy level of $\epsilon = 10^{-4} \cdot N$ and the sample size $N = 500$, the fits take considerably longer than for the other, stricter accuracy levels. This can possibly again be explained by the fact that the procedure uses the solution for the previous value of m as a starting point for the next value of m. Since the distribution function of the uniform distribution is linear, the ECDF will generally be close to linear as well. This means the optimal fit for $m - 1$ will generally be very close to the optimum for m. If this previous fit is somewhat inaccurate, this starting point is likely worse as well, and more time will be taken.

Finally, for the exponential distribution at the strictest accuracy level, the fits take much longer to perform than for any of the other three distributions. An explanation for this has not been found. For the exponential distribution, it is also interesting to consider the difference in time the convex fits and the concave fits take. While the other distributions are symmetric, and therefore the convex and concave fits are (approximately) equivalent, this is not the case for the exponential distribution, whose distribution function is fully concave. It is therefore expected that the convex fits take longer than the concave fits for this distribution. This is indeed the case, as shown in Figure 10. Additionally, there is a greater spread in the time taken for the convex fits than for the concave fits.

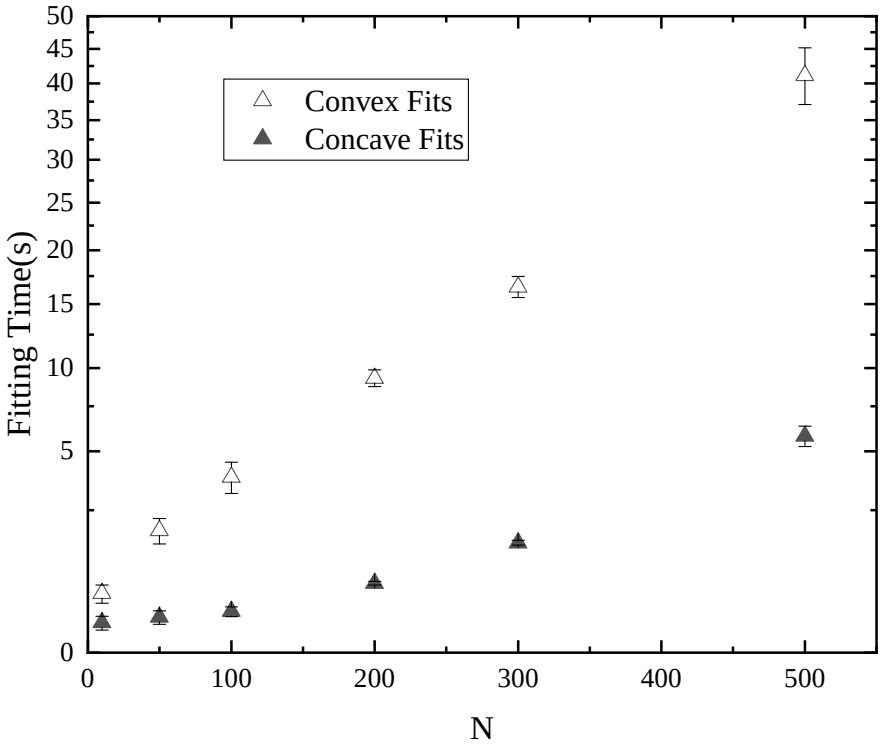


Figure 10: The time the convex fits and concave fits take for a single dataset of size $N = 200$ from the exponential distribution, at an accuracy level of $\epsilon = 10^{-6} \cdot N$. Error bars indicate the sample standard deviation. Note the quadratic scaling on the y-axis.

5 Conclusion

The goal of this project was to design a practical, powerful test to test for multimodality. This was done by modifying an existing test, called the dip test, which is known to be very conservative. In particular, the test was modified to minimise the CvM statistic instead of the KS statistic. A similar idea was already employed in the string test from [9]. However, this test still used the distribution that minimised the KS statistic to calibrate the test. It was expected that if the distribution that minimised the CvM statistic would be used for this calibration instead, the power and calibration would improve.

From the numerical experiments performed in Section 4, this does not appear to be the case. Although some caution should be exercised when drawing strong conclusions from a limited number of experiments, the results uniformly indicate that the test is less well calibrated and is weaker in power than the string test. For data from a normal distribution, the new methodology is more conservative than the string test, and for data generated from a Cauchy distribution, the new methodology is extremely conservative, with half of the reported p-values being larger than 0.76. The string test, while also being somewhat conservative for this distribution, is substantially better calibrated in this case. For the uniform distribution, both tests are reasonably well calibrated, although the new methodology is slightly anti-conservative.

We also compared the power of the new methodology to two formulations of Silverman's test, the dip test and the string test, where data was drawn from a Gaussian mixture. Additional caution should be exercised here, since only a single distribution and sample size is examined. From this, it appears that the new methodology is generally more powerful than the dip test, and similar in power to the original formulation of Silverman's test, though it lacks in power compared to the string test and the improved formulation of Silverman's test.

In order to perform the test, we needed an efficient method to minimise the CvM statistic. Whereas the local nature of the KS statistic enables one to minimise this statistic in a very short amount of time, even for very large sample sizes, minimising the CvM statistic turned out to be more difficult. In this project, a projected gradient descent method was used. In Section 3 we discussed how such a method can be used to minimise the CvM statistic, and how the method can be optimised for this specific case. In particular, the projection step turned out to simplify to a very simple algorithm for this specific case.

We conclude this optimisation procedure to be adequate for the use-case of this test, allowing us to perform the test for moderate sample sizes ($N \approx 200$) in a reasonable amount of time (≈ 3 minutes) on the hardware used. From experimentation, it appears the time taken scales approximately quadratically with the sample size, though the scaling constants differ between distributions and accuracy levels. It is also likely the implementation can be further optimised. For instance, the projection algorithm consists of taking a large number of weighted averages between numbers, several of which can be performed parallel to one another. Introducing parallel programming could therefore significantly reduce the time taken. However, this is beyond the scope of this project.

Finally, it is interesting to consider optimising the AD statistic instead of the CvM statistic. This comes with a unique set of challenges, that slightly complicate the used procedure. For instance, the discrete form of this statistic given by equation (5) contains logarithms, meaning its gradient is not Lipschitz continuous. Since the gradient most rapidly changes when the AD statistic becomes large, and therefore not in the region where the minimising distribution lies, the method can be adapted to work with the AD statistic as well. Of the three versions of the string test described in [9], the AD-variant had the most power. Due to the inherent similarities between the new methodology and the string test, it is expected that using the AD statistic for calibration instead could lead to an increase in power.

References

[1] C. Loader, *Local Regression and Likelihood*. Statistics and Computing, New York, NY: Springer, 1999.

[2] B. W. Silverman, "Using Kernel Density Estimates to Investigate Multimodality," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 43, pp. 97–99, 5 1981.

[3] E. Mammen, J. S. Marron, and N. I. Fisher, "Some asymptotics for multimodality tests based on kernel density estimates," *Probability Theory and Related Fields*, vol. 91, no. 1, pp. 115–132, 1992.

[4] H. G. Tucker, "A Generalization of the Glivenko-Cantelli Theorem," *The Annals of Mathematical Statistics*, vol. 30, pp. 828–830, 9 1959.

[5] P. Hall and M. York, "On the calibration of Silverman's test for multimodality," *Statistica Sinica*, vol. 11, pp. 515–536, 5 2001.

[6] A. J. Izenman and C. J. Sommer, "Philatelic Mixtures and Multimodal Densities," *Journal of the American Statistical Association*, vol. 83, pp. 941–953, 6 1988.

[7] J. A. Hartigan and P. M. Hartigan, "The Dip Test of Unimodality," *The Annals of Statistics*, vol. 13, pp. 70–84, 3 2007.

[8] P. M. Hartigan, "Algorithm AS 217: Computation of the Dip Statistic to Test for Unimodality," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 34, pp. 320–325, 5 1985.

[9] I. Stoepker, *Testing for multimodality*. B.S. thesis, Eindhoven University of Technology, Eindhoven, 2016.

[10] W. Stute, W. G. Manteiga, and M. P. Quindimil, "Bootstrap based goodness-of-fit-tests," *Metrika*, vol. 40, no. 1, pp. 243–256, 1993.

[11] R. Tibshirani, *Lecture Slides Convex Optimization*. Pittsburgh, PA: Carmegie Mellon University, 2019.

[12] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[13] Y. Nesterov, "A method of solving a convex programming problem with convergence rate O(1/kˆ2)," *Soviet Mathematics Doklady*, vol. 27, no. 3, pp. 372–376, 1983.

[14] E. Kreyszig, *Introductory Functional Analysis with Applications*. New York, NY: John Wiley & Sons, 1 ed., 1978.

[15] 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)*, vol. 60, pp. 579–589, 6 1998.

Appendices

A Math appendix

A.1 Discrete form of CvM statistic

Let $\{X_{(i)}\}_{i=1}^{N}$ be distinct order statistics, and $\hat{F}$ be their ECDF. Let $\tilde{F}$ be a continuous distribution function, and define $y_i = \tilde{F}(X_{(i)})$. We then have that:

$$\begin{aligned}
CvM(\hat{F}, \tilde{F}) &= \int_{-\infty}^{\infty} \left(\hat{F}(x) - \tilde{F}(x)\right)^2 d\tilde{F}(x) \\
&= \int_{-\infty}^{X_{(1)}} \left(\hat{F}(x) - \tilde{F}(x)\right)^2 d\tilde{F}(x) \\
&+ \sum_{i=1}^{n-1} \int_{X_{(i)}}^{X_{(i+1)}} \left(\hat{F}(x) - \tilde{F}(x)\right)^2 d\tilde{F}(x) \\
&+ \int_{X_{(n)}}^{\infty} \left(\hat{F}(x) - \tilde{F}(x)\right)^2 d\tilde{F}(x) \\
&= \frac{1}{3}\left[\tilde{F}(x)^3\right]_{-\infty}^{X_{(0)}} + \frac{1}{3}\sum_{i=1}^{N-1}\left[(\tilde{F}(x) - \frac{i}{N})^3\right]_{X_{(i)}}^{X_{(i+1)}} + \frac{1}{3}\left[(\tilde{F}(x) - 1)^3\right]_{X_{(N)}}^{\infty} \\
&= \frac{1}{3}\sum_{i=1}^{N}\left(\left(y_i - \frac{i-1}{N}\right)^3 - \left(y_i - \frac{i}{N}\right)^3\right).
\end{aligned}$$

We can then use the difference of cubes formula to expand the expression above.

$$\begin{aligned}
CvM(\hat{F}, \tilde{F}) &= \frac{1}{3N}\sum_{i=1}^{N}\left((y_i - \frac{i-1}{N})^2 + (y_i - \frac{i-1}{N})(y_i - \frac{i}{N}) + (y_i - \frac{i}{N})^2\right) \\
&= \frac{1}{3N}\sum_{i=1}^{N}(3y_i^2 - 3\cdot\frac{2i-1}{N}y_i + \frac{3i^2-3i+1}{N^2}) \\
&= \frac{1}{N}\sum_{i=1}^{N}(y_i^2 - 2\frac{2i-1}{2N}y_i + \frac{4i^2-4i+1}{(2N)^2} + \frac{1}{12N^2}) \\
&= \frac{1}{12N^2} + \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \frac{2i-1}{2N}\right)^2.
\end{aligned}$$

A.2 Extendability of piecewise-linear solution

Lemma 3.2. *Let $N \geq 2$. Let $\tilde{F}$ be any continuous unimodal distribution, with $\tilde{F}(X_{(1)}) < 1$ and $\tilde{F}(X_{(N)}) > 0$. Define $y_i = \tilde{F}(X_{(i)})$. Then the following statements are true:*

- *$y_1 = 0$ or $y_2 > y_1$, and*
- *$y_N = 1$ or $y_N > y_{N-1}$*

Proof. We show that either $y_1 = 0$ or $y_2 > y_1$. The second condition is analogous.

We argue by contradiction: suppose $y_1 \neq 0$, and $y_2 \leq y_1$. Since $y_i \in [0, 1]$, we have that $y_1 > 0$. Since $\tilde{F}$ is monotonically increasing, $y_2 = y_1$.

By monotonicity, we have that $\tilde{F}(x) \geq y_1$ for all $x > X_{(1)}$.

Since $y_1 > 0$ and $\lim_{x\to\infty} \tilde{F}(x) = 0$, there exists an $x < X_{(1)}$ for which $\tilde{F}(x) < y_1$. But $y_1 = y_2$, so we conclude $X_{(1)}$ must lie in the concave part. $X_{(2)} > X_{(1)}$, so then $X_{(2)}$ must lie in the concave part as well.

However, since y_1=y_2, concavity of $\tilde{F}$ at $X_{(2)}$ implies that $\tilde{F}(x) \leq y_2 = y_1$ for all $x > X_{(2)}$. But we previously concluded that $\tilde{F}(x) \geq y_1$ for all $x > X_{(1)}$. Combining these facts, we conclude that $\tilde{F}(x) = y_1$ for all $x > X_{(2)}$. Since $\lim_{x\to\infty} \tilde{F}(x) = 1$, we find that $y_1 = 1$. This contradicts the assumption that $y_1 \neq 1$. □

A.3 Redundancy of the various constraints

Lemma 3.4 (Redundancy of the monotonicity constraint)**.** *Let* $\mathbf{y}$ *satisfy unimodality, but not monotonicity. Then there exists a* $\hat{\mathbf{y}}$ *that satisfies both unimodality and monotonicity, and scores better than* $\mathbf{y}$*.*

Proof. Since $\mathbf{y}$ is not monotonic, it either has a local minimum other than y_1 in the convex part, or a local maximum other than y_N in the concave part. We explain how we can modify $\mathbf{y}$ so that it no longer has a local minimum, since removing a local maximum is analogous. Suppose this minimum runs from $X_{(k)}$ to $X_{(l)}$. Put more formally, $\exists_{1<k\leq l\leq N}$ such that $y_k = y_{k+1} = ... = y_l$, $y_k < y_{k-1}$ and $y_l < y_{l+1}$ (if $l \neq N$).

Case 1: If $y_k \geq \frac{2k-1}{2N}$, then by convexity, $y_i > y_k = \frac{2k-1}{2N} > \frac{2i-1}{2N}$ for all $i < k$. The vector can thus trivially be improved by replacing y_1 through y_{k-1} by $\frac{2k-1}{2N}$, which removes the local minimum.

Case 2: If $y_k < \frac{2k-1}{2N}$, then $y_i < \frac{2i-1}{2N}$ for all i with $k \leq i \leq l$. We can thus improve $\mathbf{y}$ by increasing y_k through y_l to $\min\{\frac{2k-1}{2N}, y_{k-1}, y_{l+1}\}$, maintaining unimodality. If $\frac{2k-1}{2N} < \min(y_{k-1}, y_{l+1})$, we have $y_k = \frac{2k-1}{2N}$ after this modification, and we arrive at the first case, and the local minimum can be removed. If one of the two adjacent points is the binding, then we have increased the number of points constituting the local minimum by one, and we can repeat the procedure.

Since at every step of this procedure either the local minimum is removed or the number of points constituting the local minimum is increased by one, this procedure is guaranteed to terminate at some point, and the local minimum will be removed. □

Lemma 3.5 (Redundancy of the extendability constraint)**.** *Let* $\mathbf{y}$ *satisfy unimodality and monotonicity, but not extendability. Then there exists a* $\hat{\mathbf{y}}$ *that satisfies unimodality, monotonicity and extendability, and scores better than* $\mathbf{y}$*.*

Proof. Since $\mathbf{y}$ does not satisfy extendability, we have that either $y_1 < 0$, $y_1 = y_2$ and $y_1 \neq 0$, $y_N > 1$, or $y_{N-1} = y_N$ and $y_N \neq 1$. We will assume one of the former two conditions holds, since the latter two are analogous.

If $y_1 < 0$, let $l = \arg\max_i \mathbb{1}\{y_i < \frac{1}{2N}\}$. Replacing y_1 through y_l by $\frac{1}{2N}$ will maintain monotonicity and unimodality, while trivially improving the solution.

Suppose $y_1 = y_2$. Let $l = \arg\max_i \mathbb{1}\{y_i = y_1\}$. Since by definition $y_{l-1} = y_l$, and $y_{l+1} > y_l$, $\mathbf{y}$ is strictly locally convex in $X_{(l)}$.

Case 1: If $y_l \geq \frac{2l-1}{2N}$, then $y_i > \frac{2i-1}{2N}$ for all $i < l$. We can thus trivially improve the solution by replacing y_i by $y_i - \epsilon(X_{(l)} - X_{(i)})$ for $i < l$ where $\epsilon > 0$ is chosen small enough that the new values y_i are still larger than $\frac{2i-1}{2N}$ and convexity at $X_{(l)}$ is not violated. This improves the solution, and satisfies extendability.

Case 2: If $y_l < \frac{2l-1}{2N}$, then y_l can be increased by some $\epsilon > 0$, such that convexity at $X_{(l)}$ is not violated, y_l remains below y_{l+1} and y_l remains below $\frac{2l-1}{2N}$. We can then repeat the procedure, but now the value of l has decreased by one. This guarantees eventual termination.

A.4 Convexity of the objective functions

Theorem A.1 (Convexity of the objective functions)**.** *Let* $\mathbf{y}, \hat{\mathbf{y}} \in \mathbb{R}^n$*, and* $\lambda \in (0,1)$*. We show that*

$$g(\lambda\mathbf{y} + (1-\lambda)\hat{\mathbf{y}}) \leq \lambda g(\mathbf{y}) + (1-\lambda)g(\mathbf{y}) \leq \max(g(\mathbf{y}), g(\hat{\mathbf{y}}))$$

for both objective functions given in Section 3.2.3, with equality if and only if $\mathbf{y} = \hat{\mathbf{y}}$*. This means the objective functions are strictly convex.*

Proof. We can immediately conclude that the CvM objective function is convex since it is a quadratic function. Nevertheless, working out the expression results in an inequality that is useful in the error estimation of the projected gradient descent procedure.

$$\begin{aligned}
&g_{CvM}\left(\lambda\mathbf{y} + (1-\lambda)\hat{\mathbf{y}}\right) - \lambda g_{CvM}(\mathbf{y}) - (1-\lambda)g_{CvM}(\hat{\mathbf{y}}) \\
&= \sum_{i=1}^{N}\left(\lambda y_i + (1-\lambda)\hat{y}_i - \frac{2i-1}{2N}\right)^2 - \lambda\sum_{i=1}^{N}\left(y_i - \frac{2i-1}{2N}\right)^2 - (1-\lambda)\sum_{i=1}^{N}\left(\hat{y}_i - \frac{2i-1}{2N}\right)^2 \\
&= -\lambda(1-\lambda)\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \\
&= -\lambda(1-\lambda)\left\|\mathbf{y} - \hat{\mathbf{y}}\right\|^2 \leq 0, \text{ with equality only when } \mathbf{y} = \hat{\mathbf{y}}.
\end{aligned}$$

$$\begin{aligned}
&g_{AD}\left(\lambda\mathbf{y} + (1-\lambda)\hat{\mathbf{y}}\right) - \lambda g_{AD}(\mathbf{y}) - (1-\lambda)g_{AD}(\hat{\mathbf{y}}) \\
&= -\sum_{i=1}^{N}((2i-1)\log\left(\frac{\lambda y_i + (1-\lambda)\hat{y}_i}{y_i^{\lambda}\hat{y}_i^{1-\lambda}}\right) + (2N+1-2i)\log\left(\frac{\lambda(1-y_i) + (1-\lambda)(1-\hat{y}_i)}{(1-y_i)^{\lambda}(1-\hat{y}_i)^{1-\lambda}}\right) \\
&\leq 0, \text{ with equality only when } \mathbf{y} = \hat{\mathbf{y}}.
\end{aligned}$$

□

The last line holds because by the weighted AM-GM inequality, the arguments of both logarithms are ≥ 1, with equality only if $y_i = \hat{y}_i$.

A.5 The norm of the transition matrix M

Let $\mathbf{y}$ be given, and let $\boldsymbol{\delta} = M \cdot \mathbf{y}$. We then have that:

$$\begin{aligned}
\|\boldsymbol{\delta}\|^2 &= \sum_{i=1}^{m}\delta_i^2 \\
&= \delta_1^2 + \sum_{i=2}^{m}\delta_i^2 \\
&= y_1^2 + \sum_{i=2}^{m}(y_i - y_{i-1})^2 \\
&\leq y_1^2 + 2\sum_{i=2}^{m}(y_i^2 + y_{i-1}^2) \\
&\leq 4\sum_{i=1}^{m}y_i^2 \\
&= 4\|\mathbf{y}\|^2
\end{aligned}$$

This means that $\|\boldsymbol{\delta}\| \leq 2\|\mathbf{y}\|$. We therefore find that the norm of M is bounded above by 2. □

A.6 Projection in case of non-positive inner products

Let L be linear subspace of $\mathbb{R}^m$, and let $\mathbf{v}^0$ be a vector in L. Furthermore, let $\{\mathbf{n}_i\}_{i=1}^l \in L$ for some $l \in \mathbb{Z}_{\geq 0}$ be a number of independent vectors, with $\mathbf{n}_i \cdot \mathbf{n}_j \leq 0$ for all $i \neq j$, and let the orthogonal complement of any of these vectors $\mathbf{n}_i$ not be contained in L. The goal is to find the vector $\mathbf{v}^* \in L$ that minimises $\left\|\mathbf{v}^* - \mathbf{v}^0\right\|$, subject to the constraints $\mathbf{n}_i \cdot \mathbf{v}^* \leq 0$ for all i.

If $\mathbf{n}_i \cdot \mathbf{v}^0 \leq 0$ for all i, then the solution is $\mathbf{v}^* = \mathbf{v}^0$. If not, let i be an index such that $\mathbf{n}_i \cdot \mathbf{v}^0 > 0$.

Using orthogonal projection, we can write $\mathbf{v}^* = \mathbf{v}^0 + \alpha\mathbf{n}_i + \mathbf{v}^\perp$ for some scalar $\alpha \in \mathbb{R}$ where $\mathbf{v}^\perp \perp \mathbf{n}_i$. Since $\mathbf{v}^*, \mathbf{v}^0, \mathbf{n}_i \in L$, we see that $\mathbf{v}^\perp \in L$. Furthermore, we have that $0 \geq \mathbf{n}_i \cdot \mathbf{v}^* = \mathbf{n}_i \cdot \mathbf{v}^0 + \alpha\mathbf{n}_i \cdot \mathbf{n}_i$. Since $\mathbf{n}_i \cdot \mathbf{v}^0 > 0$, we see $\alpha < 0$. By Pythagoras' theorem, $\left\|\mathbf{v}^* - \mathbf{v}^0\right\|^2 = \alpha^2 \left\|\mathbf{n}_i\right\|^2 + \left\|\mathbf{v}^\perp\right\|^2$.

Now suppose $\mathbf{n}_i \cdot \mathbf{v}^* < 0$. We consider $\mathbf{v}^* + \epsilon\mathbf{n}_i$, where $0 < \epsilon < -\alpha$ is chosen small enough that $\mathbf{n}_i \cdot (\mathbf{v}^* + \epsilon\mathbf{n}_i) = \mathbf{n}_i \cdot \mathbf{v}^* + \epsilon\mathbf{n}_i \cdot \mathbf{n}_i \leq 0$. For any $j \neq i$ we have $\mathbf{n}_j \cdot (\mathbf{v}^* + \epsilon\mathbf{n}_i) = \mathbf{n}_j \cdot \mathbf{v}^* + \epsilon\mathbf{n}_j \cdot \mathbf{n}_i \leq \mathbf{n}_j \cdot \mathbf{v}^* \leq 0$. We thus see that $\mathbf{v}^* + \epsilon\mathbf{n}_i$ satisfies all constraints, and is closer to $\mathbf{v}^0$ than $\mathbf{v}^*$ is. This gives rise to a contradiction.

We must therefore have that $\mathbf{n}_i \cdot \mathbf{v}^* = 0$. This fixes the value of α, such that $\mathbf{v}^0 + \alpha\mathbf{n}_i \in \langle\mathbf{n}_i\rangle^\perp$. Let $\mathbf{v}^1 = \mathbf{v}^0 + \alpha\mathbf{n}_i$. We then wish to find the $\mathbf{v}^* \in \langle\mathbf{n}_i\rangle^\perp \cap L =: L'$ that minimises $\left\|\mathbf{v}^\perp\right\| = \left\|\mathbf{v}^1 - \mathbf{v}^*\right\|^2$. Furthermore, note that since $\mathbf{n}_i \cdot \mathbf{v}^* = 0$, we have that $\mathbf{n}_j \cdot \mathbf{v}^* \leq 0 \iff (\mathbf{n}_j - \frac{\mathbf{n}_i \cdot \mathbf{n}_j}{\|\mathbf{n}_i\|^2}\mathbf{n}_i) \cdot \mathbf{v}^* \leq 0$, and that $(\mathbf{n}_j - \frac{\mathbf{n}_i \cdot \mathbf{n}_j}{\|\mathbf{n}_i\|^2}\mathbf{n}_i) \in L'$. We thus arrive at exactly the same optimisation problem. Since we restrict ourselves to L' from now on, we can drop the constraint pertaining to $\mathbf{n}_i$. Since the number of constraints drops by one at every time-step, this process is guaranteed to terminate. This gives rise to a very simple algorithm to find the vector $\mathbf{v}^*$:

`Projection in case of non-positive inner products.`

Input: $l \in \mathbb{Z}_{\geq 0}$ `constraint vectors` $\{\mathbf{n}_i\}_{i=1}^l \in \mathbb{R}^m$`, with` $\mathbf{n}_i \cdot \mathbf{n}_j \leq 0$ `for` $i \neq j$`, target vector` $\mathbf{v}^0 \in \mathbb{R}^m$

$\mathbf{v} \leftarrow \mathbf{v}^0$

while $\exists_{\mathbf{n}_i}$ *with* $\mathbf{n}_i \cdot \mathbf{v} > 0$ **do**

`Subtract` $\mathbf{n}_i$ `from` $\mathbf{v}$ `until` $\mathbf{n}_i \cdot \mathbf{v} = 0$ `(i.e.` $\mathbf{v} \leftarrow \mathbf{v} - \lambda\mathbf{n}_i$`, where` $\lambda > 0$ `is chosen such that (before assignment)` $\mathbf{n}_i \cdot (\mathbf{v} - \lambda\mathbf{n}_i) = 0$`).`

`Orthogonally project all` $\mathbf{n}_j$ `with` $j \neq i$ `onto` $\langle\mathbf{n}_i\rangle^\perp$ $\left($`i.e.` $\mathbf{n}_j \leftarrow \mathbf{n}_j - \frac{\mathbf{n}_i \cdot \mathbf{n}_j}{\|\mathbf{n}_i\|^2}\mathbf{n}_i\right)$

end

return $\mathbf{v}$

B Additional data and measurements

B.1 The calibration of the test

In this appendix, the actual significance levels at several nominal significance levels from the calibration experiment are tabulated.

α	con	nat	a-c	c.s
0.05	0.00	0.01	0.02	0.07
0.10	0.02	0.03	0.04	0.08
0.20	0.04	0.06	0.08	0.14
0.50	0.27	0.30	0.37	0.045

Table 1: The actual significance level versus the nominal significance level α from the calibration experiment, when data comes from the standard normal distribution. *con*, *nat* and *a-c* indicate the conservative approach, natural approach and anti-conservative approach of the new methodology respectively. *c.s* indicates the CvM string test.

α	con	nat	a-c	c.s
0.05	0.05	0.05	0.06	0.04
0.10	0.11	0.11	0.11	0.10
0.20	0.23	0.24	0.25	0.17
0.50	0.53	0.54	0.56	0.46

Table 2: The actual significance level versus the nominal significance level α from the calibration experiment, when data comes from the standard uniform distribution. *con*, *nat* and *a-c* indicate the conservative approach, natural approach and anti-conservative approach of the new methodology respectively. *c.s* indicates the CvM string test.

α	con	nat	a-c	c.s
0.05	0.00	0.00	0.00	0.03
0.10	0.00	0.00	0.00	0.11
0.20	0.00	0.00	0.00	0.19
0.50	0.11	0.17	0.27	0.39

Table 3: The actual significance level versus the nominal significance level α from the calibration experiment, when data comes from the standard Cauchy distribution. *con*, *nat* and *a-c* indicate the conservative approach, natural approach and anti-conservative approach of the new methodology respectively. *c.s* indicates the CvM string test.

B.2 The power of the test

μ	n.m	sil	ims	dip	c.s
2.00	0.02	0.00	0.16	0.00	0.06
3.00	0.02	0.00	0.46	0.00	0.10
3.25	0.10	0.16	0.54	0.02	0.16
3.50	0.34	0.30	0.80	0.10	0.38
3.75	0.36	0.34	0.92	0.18	0.64
4.00	0.50	0.78	1.00	0.32	0.74
4.25	0.86	0.98	1.00	0.38	0.94
4.50	0.94	0.96	1.00	0.82	0.90
4.75	0.96	1.00	1.00	0.80	0.98

Table 4: The power of the various tests at the significance level $\alpha = 0.1$ from the power experiment, versus the value μ as defined in equation (10). *n.m* indicates the new methodology, *sil* indicates the original Silverman's test, *ims* indicates the improved Silverman's test, *dip* indicates the original dip test and c.s indicates the CvM string test.

μ	n.m	sil	ims	dip	c.s
2.00	0.00	0.00	0.04	0.00	0.02
3.00	0.00	0.00	0.32	0.00	0.02
3.25	0.06	0.06	0.34	0.00	0.10
3.50	0.26	0.10	0.74	0.10	0.20
3.75	0.24	0.08	0.90	0.06	0.54
4.00	0.40	0.44	1.00	0.24	0.68
4.25	0.74	0.80	1.00	0.28	0.84
4.50	0.84	0.86	1.00	0.72	0.84
4.75	0.92	0.98	1.00	0.76	0.96

Table 5: The power of the various tests at the significance level $\alpha = 0.05$ from the power experiment, versus the value μ as defined in equation (10). *n.m* indicates the new methodology, *sil* indicates the original Silverman's test, *ims* indicates the improved Silverman's test, *dip* indicates the original dip test and c.s indicates the CvM string test.