A Multivariate Unimodality Test Harnessing the Dip Statistic of Mahalanobis Distances Over Random Projections

Prodromos Kolyvakis[1] **Aristidis Likas**[2]

[1] ORamaVR SA, Geneva, Switzerland & Talos Automata SMPC, Ioannina, Greece
[2] Department of Computer Science and Engineering, University of Ioannina, 45110 Ioannina, Greece

Abstract

Unimodality, pivotal in statistical analysis, offers insights into dataset structures and drives sophisticated analytical procedures. While unimodality's confirmation is straightforward for one-dimensional data using methods like Silverman's approach and Hartigans' dip statistic, its generalization to higher dimensions remains challenging. By extrapolating one-dimensional unimodality principles to multi-dimensional spaces through linear random projections and leveraging point-to-point distancing, our method, rooted in α-unimodality assumptions, presents a novel multivariate unimodality test named *mud-pod*. Both theoretical and empirical studies confirm the efficacy of our method in unimodality assessment of multidimensional datasets as well as in estimating the number of clusters. The implementation of *mud-pod* is publicly available at `https://github.com/prokolyvakis/mudpod`.

1 INTRODUCTION

Unimodality, a fundamental concept in statistical analysis, serves as a critical lens through which one can decipher the inherent structure and patterns within datasets. Understanding unimodality is paramount for multiple reasons. Firstly, it provides a rudimentary insight into the nature of the data, highlighting whether the data points converge towards a common central tendency or deviate significantly. Secondly, unimodality serves as a precursor to more complex analytical procedures, such as clustering algorithms, determining their necessity, and potentially influencing their outcomes [Kalogeratos and Likas, 2012, Daskalakis et al., 2013, 2014]. In essence, the importance of unimodality transcends mere statistical significance, extending its value to practical, real-world applications.

In one-dimensional data, unimodality can be fundamentally understood as the task of discerning whether a given distribution exhibits a single prominent peak or mode. A notable advantage is that one-dimensional unimodality can be confirmed using robust statistical hypothesis tests, particularly for one-dimensional data. Methods such as Silverman's approach, exploiting fixed-width kernel density estimates [Silverman, 1981], the widely recognized Hartigans' dip statistic [Hartigan and Hartigan, 1985] and more recently the UU-test [Chasani and Likas, 2022] are prime examples.

Nevertheless, when transitioning to higher dimensions, the process of defining unimodality becomes less straightforward, even when considering only symmetric distributions. The intricacies of multi-dimensional spaces impose challenges that are not present in one-dimensional settings, leading to diverse interpretations and approaches to gauge unimodality. Even worse, these intricacies make the generalization of unimodality tests notably challenging. Numerous efforts have been made to capture the geometric essence of unimodality in $\mathbb{R}^d$ ($d > 1$) and translate it into an analytical framework [Dai, 1989]. In a seminal work by Olshen and Savage [1970], a definition of generalized unimodality characterized by a positive parameter α was proposed, called α-unimodality, which is pertinent to distributions across $\mathbb{R}^d$ and offers a broader perspective that encompasses many aspects of 1-dimensional unimodality [Dharmadhikari and Joag-Dev, 1988, Chapter 3.2]. Despite the various definitions, however, few methods are available for assessing unimodality in multidimensional data vectors.

In parallel, another line of research has delved into the use of random projections as a strategy for capturing the essence of multi-dimensional distributions. Random projections, known for its efficacy in dimensionality reduction, have shown significant potential for learning mixtures of Gaussians [Dasgupta, 1999]. Additionally, the Diaconis-Freedman effect elucidates the behavior of random projections of probability distributions in the high-dimensional space [Diaconis and Freedman, 1984]. Specifically, for a given probability distribution P in a d-dimensional space, when we consider a dimension q much smaller than d, the majority of the q-dimensional projections of P resemble scaled mixtures of spherically symmetric Gaussian distributions [Dümbgen et al., 2013]. Consequently, random projections appear to be a potent tool for the analysis of unimodality as they facilitate the transformation of the problem into a seemingly simpler space. However, not every random projection is conducive to our analysis; many projections can obfuscate distinct modes by distorting distances. Consequently, we limit our approach to a family of random projections that, with a certain probability, maintain pairwise distances.

In this work, we tackle the complexity of extending unimodality testing to higher dimensions. Echoing the principles of α-unimodality, we introduce a novel algorithm for efficient multivariate unimodality testing. Our approach bridges the gap between the simplicity of one-dimensional unimodality confirmation and the intricacies of its higher-dimensional counterpart. Central to our investigation are the α-unimodality preserving properties of point-to-point distancing and linear random projections. We demonstrate that linear random projections preserve the α-unimodality property in Mahalanobis distances from a reference point. Leveraging these one-dimensional Mahalanobis distances, we apply the dip test for unimodality detection. Employing various random projections, akin to Monte Carlo simulations, we assess the α-unimodality of the original data distribution. In summary, the contribution of our work is two-fold. Firstly, to the best of our knowledge, we propose the first mathematically founded multivariate unimodality test, dubbed *mud-pod*[1]. Secondly, we present the *mp-means* incremental clustering method which is a wrapper around k-means exploiting mudpod for unimodality assessment. Both theoretical findings and empirical validations underpin our methods, showcasing efficacy in unimodality assessments and clustering scenarios.

2 RELATED WORK

A multivariate unimodality test that aligns closely with our work is the *dip-dist*, introduced by Kalogeratos and Likas [2012]. This criterion aims to ascertain the modality (unimodal vs. multimodal) of a dataset by applying the unidimensional dip test on each row of the pairwise distance matrix of the dataset. The rationale behind this approach is that by selecting an arbitrary data point and calculating its distances to all other points, we obtain a snapshot of the underlying cluster morphology. In presence of a single cluster, the distribution of distances is expected to be unimodal. To further the applicability of this criterion, it has been integrated into a clustering method named *dip-means*. This incremental algorithm employs cluster splitting based on the dip-dist to determine if a cluster should be divided. Consequently, dip-means can automatically estimate the cluster count. However, the dip-dist criterion is not without its limitations. A glaring drawback is its reliance on pairwise distances and its operation within the original data space, which can present challenges in certain scenarios. Moreover, the dip-dist method operates in an ad-hoc manner, lacking a rigorous mathematical foundation. In our work, we address these limitations by introducing random projections to assess unimodality on randomly projected distances. This approach permits Monte Carlo hypothesis testing by enabling sampling, previously not directly feasible in the original space. Additionally, we leverage the Mahalanobis distance and we empirically demonstrate its added benefits. Last but not least, we provide a mathematical formulation and prove the consistency of our test, thereby establishing the missing foundation for the dip-dist method.

The folding test, introduced in Siffer et al. [2018], offers a versatile evaluation method suitable for both univariate and multivariate scenarios. This test revolves around the concept of *folding*, involving three key steps: (a) folding the distribution with respect to a designated pivot s, (b) calculating the variance of the folded distribution, and lastly, (c) comparing this folded variance to the original variance. The central idea behind the folding test is that when applied to multimodal distributions, the folded distribution typically exhibits a significantly reduced variance compared to its unimodal counterparts. A limitation of the folding test is its reliance on the empirical assumption that folding a multimodal distribution leads to a reduction in variance. Consequently, the concept of unimodality is not explicitly integrated into the folding test computation. It is important to note that there are cases where this assumption does not hold true, resulting in incorrect outcomes for the folding test [Chasani and Likas, 2022]. Another research direction focuses on examining particular families of unimodal distributions. In the work of Dunn et al. [2021], a scalable test for log-concavity is elucidated building on maximum likelihood estimation (MLE), validated in finite samples across any dimension. A noteworthy empirical observation from their research is the pronounced efficacy achieved by adopting random projections. However, it is crucial to acknowledge that while log-concave functions capture a substantial subset of unimodal functions, they fall short of encompassing the entirety of the concept and may struggle to extend their applicability to more diverse and real-world scenarios.

3 METHODS

3.1 PRELIMINARIES

In the following, we present key notation and foundational background applicable throughout this paper. We use capital letters to denote random variables or matrices and boldface

[1] Multivariate Unimodality Dip based on random Projections, Observers & Distances.

type to represent vectors.

3.1.1 α-Unimodal Distrubutions

Without loss of generality, we assume that the mode of the unimodal distribution is at $\mathbf{0}$. A random d-vector $\mathbf{X} \in \mathbb{R}^d$ is said to have an *α-unimodal distrubution* about $\mathbf{0}$ if, for every bounded, nonnegative, Borel measurable function g on $\mathbb{R}^d$ the quantity $t^\alpha \mathbb{E}[g(t\mathbf{X})]$ is nondecreasing in $t \in (0, \infty)$. In what follows, we use the notation $\mathbf{X} \sim \mathcal{P}_\alpha$ to represent a d-vector following an α-unimodal distribution. It follows from the defintion that if $\mathbf{X} \sim \mathcal{P}_\alpha$ and $\alpha < \beta$, then $\mathbf{X} \sim \mathcal{P}_\beta$. An important equivalent characterization for the set of α-unimodal distrubutions is the Decomposition theorem: $\mathbf{X} \sim \mathcal{P}_\alpha$ iff $\mathbf{X}$ is distributed as $U^{\frac{1}{\alpha}}\mathbf{Z}$, where U is uniform on $(0, 1)$ and $\mathbf{Z}$ is independent of U [Olshen and Savage, 1970].

This decomposition provides the intuition that an α-unimodal vector can be generated by first choosing a "direction" $\mathbf{Z}$ and then scaling it toward the mode at $\mathbf{0}$ by a random radial factor $U^{1/\alpha}$. In particular, larger values of α lead to stronger concentration of mass near the mode (and correspondingly lighter tails), neatly capturing the degree of unimodality through the power $1/\alpha$ applied to the uniform radial component. This theorem closely mirrors the intuition of one-dimensional unimodality, since Khintchine [1938] demonstrated that a real random variable X has a unimodal distribution iff $X \sim UZ$, where U is uniform on $[0, 1]$ and U and Z are independent. It follows that a scalar $X \sim \mathcal{P}_\alpha$ iff X^α is unimodal as per the standard definition in $\mathbb{R}$. Next, we present and prove three pivotal properties of α-unimodality, i.e., the translation, norm, and projection properties.

Lemma 3.1 (Translation Property)**.** *Let $\mathbf{X} \sim \mathcal{P}_\alpha$ and $\mathbf{c} \in \mathbb{R}^d$, then $\mathbf{X} + \mathbf{c} \sim \mathcal{P}_\alpha$.*

Proof. Let $t^\alpha \mathbb{E}[g(t\mathbf{X} + t\mathbf{c})] = t^\alpha \mathbb{E}[h(t\mathbf{X})]$, where $h(\mathbf{x}) = g(\mathbf{x} + \mathbf{c})$. Note that h is bounded, nonnegative, Borel measurable and that the first expression is nondecreasing iff the last expression is nondecreasing in t. □

Lemma 3.2 (Norm Property)**.** *If $\mathbf{X} \sim \mathcal{P}_\alpha$, then $||\mathbf{X}|| \sim \mathcal{P}_\alpha$.*

Proof. $||\mathbf{X}|| = \sqrt{(U^{\frac{1}{\alpha}}\mathbf{Z})^\top (U^{\frac{1}{\alpha}}\mathbf{Z})} = U^{\frac{1}{\alpha}} ||\mathbf{Z}||$ □

Lemma 3.3 (Projection Property)**.** *Let $\mathbf{X} \sim \mathcal{P}_\alpha$ and let A be a real matrix from $\mathbb{R}^d$ to $\mathbb{R}^q$, then $A\mathbf{X} \sim \mathcal{P}_\alpha$.*

Proof. A direct use of the Decomposition Theorem. □

Leveraging the foundational properties described above, we now present a salient result. This result not only fortifies our understanding of α-unimodal distributions but will also play an instrumental role in the rest of the paper.

Lemma 3.4 (Mahalanobis)**.** *Let $\mathbf{X} \sim \mathcal{P}_\alpha$ with a well-defined covariance matrix Σ and $\mathbf{o} \in \mathbb{R}^d$, then the distribution of the Mahalanobis distances with respect to $\mathbf{o}$, given by $\sqrt{(\mathbf{X} - \mathbf{o})^\top \Sigma^{-1} (\mathbf{X} - \mathbf{o})}$ is α-unimodal.*

Proof. Given a positive semidefinite covariance matrix Σ, and utilizing the matrix square root decomposition, the Mahalanobis distance can be expressed as:

$$\sqrt{(\mathbf{x} - \mathbf{o})^\top \Sigma^{-1} (\mathbf{x} - \mathbf{o})} = ||\Sigma^{-\frac{1}{2}} (\mathbf{x} - \mathbf{o})||$$

Proof stems from translation and projection lemmas. □

Mahalanobis distance possesses distinct properties: it is unitless, scale-invariant, and considers the covariance structure across all dimensions. Traditionally, it has been employed in multivariate hypothesis testing. Notably, the Hotelling's T^2 statistic [Hotelling, 1931], which generalizes the Student's t-statistic, exemplifies its usage. The Mahalanobis distance is pivotal to our multivariate unimodality test.

3.1.2 Dip Test

The dip test serves as a tool for discerning multimodality within a unidimensional dataset. It gauges this by examining the maximum deviation, i.e., the Kolmogorov-Smirnov statistic, between the empirical cumulative distribution function (e.c.d.f.), $F(t)$, and the nearest unimodal c.d.f., $G(t)$. The dip statistic for a distribution function F is defined as: $\text{dip}(F) = \min_{G \in U} \rho(F, G)$, where $\rho(F, G) = \max_t |F(t) - G(t)|$ and U represents the set of all possible unimodal distributions. The dip test's significance is highlighted by its ability to unveil the least among the most substantial deviations between the empirical cumulative distribution function F of the univariate dataset and the c.d.f.s of the class of unimodal distributions. A salient attribute of the dip statistic is its convergence as the sample size burgeons, such that $\lim_{n \to \infty} \text{dip}(F_n) = \text{dip}(F)$ [Hartigan and Hartigan, 1985]. Moreover, the class of uniform distributions U is acclaimed to be the most fitting for the null hypothesis, owing to its stochastically larger dip values compared to other unimodal distributions. To calculate the dip value, the e.c.d.f. of the data is considered, and the unimodal piecewise linear function with the smallest maximum distance to the e.c.d.f. is determined. The p-value for unimodality, derived via bootstrap samples, functions as a determinant for the dataset's modality. A dataset with a p-value greater than a indicates unimodality; otherwise, multimodality is suggested.

3.1.3 Random Projections

The Diaconis-Freedman effect can be a valuable tool for unimodality analysis, simplifying the problem by likely transforming it into a Gaussian mixture model. When considering a probability distribution P in a d-dimensional space, most q-dimensional projections of P with $q \ll d$ resemble scale mixtures of spherically symmetric Gaussian distributions. Additionally, linear random projections can preserve distances when projecting high-dimensional points into lower-dimensional spaces. This phenomenon is encapsulated in the celebrated Johnson and Lindenstraus lemma [Johnson and Lindenstraus, 1984, Fernandez-Granda, 2016] presented below:

JL Lemma: Let $S := \{x_i\}_{i=1}^{k}$ be a subset of $\mathbb{R}^d$ and $\epsilon > 0$. Then, let $\Pi \in \mathbb{R}^{d \times q}$, where $q \geq \mathit{8}\log(k)/\epsilon^{\mathit{2}}$, be a random matrix with i.i.d. entries $\Pi_{ij} \sim \mathcal{N}(0, \mathit{1}/d)$. With probability at least $\mathit{1}/k$, for any $x_i, x_j \in S$, we have:

$$(1-\epsilon)\|x_i - x_j\|^2 \leq \|\Pi x_i - \Pi x_j\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2.$$

We denote by $\mathcal{R}_{\mathbf{\Pi}}$ the set of matrices fulfilling the distance preservation criteria specified in the JL Lemma. According to the JL Lemma, if Π is sampled with i.i.d. $\mathcal{N}(0, \frac{1}{d})$ entries, then $P(\Pi \in \mathcal{R}_{\Pi}) \geq \frac{1}{k}$. By employing a square root decomposition, we can demonstrate the applicability of the JL lemma to the Mahalanobis distance [Bhattacharya et al., 2009], while mitigating the singularity issue inherent in inverting the covariance matrix in high dimensions [Lopes et al., 2011, Radhendushka Srivastava and Ruppert, 2016].

3.2 CONNECTING THE DOTS

Given a dataset of multidimensional data vectors, assessing its unimodality becomes intricate. Random projections offer a solution by maintaining key pairwise distances and performing unimodality assessment in a more Gaussian-like space. By picking an arbitrary *observer* data point and deriving its distances to all other points, we garner a snapshot of the underlying cluster morphology. In presence of a single cluster, the distribution of distances is proven to be unimodal. Notably, the narrative this observer presents is contingent upon its location. We integrate this idea of random projection and the observer's perspective into what we term a *view*. Our proposed algorithm focuses on analyzing these views, pinpointing those views that contradict unimodal narratives, and thus highlight multifaceted cluster formations. In the rest of the section, we will rigorously define the aforementioned concept.

Given a set $\mathcal{S}_{\mathbf{X}}$ of points from $\mathbf{X} \sim \mathcal{P}_\alpha$, let $\mathbf{o} \in \mathcal{S}_{\mathbf{X}}$ be a random point, dubbed observer. We define the set of Mahalanobis distances with regard to this observer as follows:

$$\mathcal{D}^{o}_{\mathbf{S}} = \left\{ ||\mathbf{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \mathbf{o})|| \mid \mathbf{x} \in \mathcal{S}_{\mathbf{X}} \setminus \{\mathbf{o}\} \right\}.$$

Let $\Pi \in \mathcal{R}_{\mathbf{\Pi}}$, we define $\Pi \circ \mathcal{D}^{o}_{\mathbf{S}}$ to be the set of the Mahalanobis distances of the randomly projected points with respect to an observer $\mathbf{o}$. Specifically, we have:

$$\Pi \circ \mathcal{D}^{o}_{\mathbf{S}} = \left\{ ||\mathbf{\Sigma}_{\Pi}^{-\frac{1}{2}} \Pi(\mathbf{x} - \mathbf{o})|| \mid \mathbf{x} \in \mathcal{S}_{\mathbf{X}} \setminus \{\mathbf{o}\} \wedge \Pi \in \mathcal{R}_{\mathbf{\Pi}} \right\},$$

where $\mathbf{\Sigma}_{\mathbf{\Pi}} = \Pi \mathbf{\Sigma} \Pi^T$. It is important to note that since $\mathbf{X} \sim \mathcal{P}_\alpha$, the elements of both $\mathcal{D}^{o}_{\mathbf{S}}$ and $\Pi \circ \mathcal{D}^{o}_{\mathbf{S}}$ exhibit α-unimodal distributions, as established earlier. This yields the subsequent observation.

Proposition 3.5 (Randomisation Hypothesis)**.** *Given* $\mathbf{X} \sim \mathcal{P}_\alpha$*, the distributions of elements within* $\mathcal{D}^{o}_{\mathbf{S}}$ *and* $\Pi \circ \mathcal{D}^{o}_{\mathbf{S}}$ *retain* α*-unimodality under any transformation* $\Pi \in \mathcal{R}_{\mathbf{\Pi}}$*.*

The Randomisation Hypothesis (RH) is central to our analysis. RH facilitates the execution of a series of one-dimensional unimodality tests, subsequently allowing the evaluation of the α-unimodality of the distribution that produces our data X. Under the RH, every randomly projected distances should exhibit unimodality. Any deviation from this expected behavior can signal a departure from unimodality in the original data distribution. Random projections confer several distinct merits. Firstly, they preserve the pairwise distances, as endorsed by the JL lemma. Secondly, given our observations, random projections serve as an invaluable tool for unimodality investigation. They transmute the challenge into a space resembling a mixture of Gaussians. Furthermore, they ameliorate the singularity problem associated with the inversion of the covariance matrix used by the Mahalanobis distance. Lastly, they pave the way for harnessing Monte Carlo simulation for hypothesis testing [Lehmann and Romano, 2005], i.e., they establish the bedrock for sampling from a distribution, specifically $\mathcal{R}_{\mathbf{\Pi}}$, that is ostensibly simpler than the original data distribution of $\mathbf{X}$.

3.3 MULTIVARIATE UNIMODALITY TESTING

Building on the aforementioned foundation, we now introduce our multivariate α-unimodality test called *mud-pod*. For a given α and a set of points $\mathcal{S}_{\mathbf{X}}$ from $\mathbf{X} \sim \mathcal{P}_\alpha$, we define our hypotheses:

$$H_0 : \mathbf{X} \sim \mathcal{P}_\alpha \quad \text{vs.} \quad H_1 : \mathbf{X} \not\sim \mathcal{P}_\alpha.$$

We define the pairing of a random projection Π with an observer $\mathbf{o}$ as a random view. We assume independence between the random projection Π and the observer $\mathbf{o}$. Given a set of N random vectors $\mathcal{S}$ and a random view, we can obtain the corresponding set of Mahalanobis distances $\Pi \circ \mathcal{D}^{o}_{\mathbf{S}}$. Under the null hypothesis, recall that $\Pi \circ \mathcal{D}^{o}_{\mathbf{S}} = \{d_i\}_{i=1}^{N-1}$ is α-unimodal, and the set $\{d_i^\alpha\}_{i=1}^{N-1}$ is unimodal, allowing the employment of the dip test. Let $T(\Pi \circ \mathcal{D}^{o}_{\mathbf{S}})$ denote the dip test p-value. If $a \in [0, 1]$ is the significance level, the null hypothesis is rejected iff $T(\Pi \circ \mathcal{D}^{o}_{\mathbf{S}}) \leq a$.

Utilizing the idea of random views, which, as previously discussed, preserve α-unimodality, we can employ them as a foundation for Monte Carlo simulations. The rationale is that as more views reject the null hypothesis, our confidence about data multimodality increases. Let $\{\Pi_i\}_{i=1}^{M} \subset \mathcal{R}_{\mathbf{\Pi}}$ be a set of M random projections. Leveraging Monte Carlo hypothesis testing theory [Lehmann and Romano, 2005, Chapter 11.2.2] and building on the fact that the dip statistic has a well defined c.d.f. [Hartigan and Hartigan, 1985], we explore the conditional c.d.f. of the dip test: $J_N(t) = P\left(T(\Pi \circ \mathcal{D}^o_{\mathbf{S}}) \leq t \mid \mathcal{S}_{\mathbf{X}}\right)$. The aforementioned probability is measured over the joint distribution of random projections $\mathcal{R}_{\mathbf{\Pi}}$ and the distribution of observers $\mathcal{O}$. Let $I\{.\}$ denote the indicator function, we define $\hat{J}_{n,M}(t)$ as the approximation of the true c.d.f $J_N(t)$ computed on the series of the random views:

$$\hat{J}_{N,M}(t) = M^{-1} \sum_{i=1}^{M} I\left\{T(\Pi \circ \mathcal{D}^o_{\mathbf{S}}) \leq t\right\}$$

By a direct application of the Glivenko-Cantelli theorem, we have that $\hat{J}_{N,M}(t)$ converges w.p. 1 to $J_N(t)$ [Sharipov, 2011]. Interestingly, the Dvoretsky, Kiefer, Wolfowitz inequality [Massart, 1990] provides bounds on the closeness between $\hat{J}_{N,M}(t)$ and $J_N(t)$ for a given M. Specifically, we have:

$$P\left(\sup_{t \in \mathbb{R}} |\hat{J}_{N,M}(t) - J_N(t)| > \tau\right) \leq 2e^{-2M\tau^2}$$

Hitherto, we have not delineated the methodology for selecting observers from the set of points obtained from a random projection. Several sampling strategies can exist, with the most intuitive being uniform random sampling. However, empirical results suggest that uniformly selecting observers based on a specific percentile of the distance distribution from the samples' mean yields superior performance. The underlying rationale is that points situated farther away from the means possess a better capability to discern the topographical elevations formed by distinct clusters [Kalogeratos and Likas, 2012]. It is important to note that despite the dependency introduced between the observer $\mathbf{o}$ and the random projection Π by the *percentile strategy*, our analysis remains valid thanks to the extension of the Glivenko-Cantelli theorem to strictly stationary sequences [Sharipov, 2011]. Ultimately, the projection dimension is the minimal integer that satisfies the JL lemma for a specified ϵ. Algorithm 1 details the complete mud-pod test. It is important to note that dip-dist [Kalogeratos and Likas, 2012] is a special case of mud-pod, omitting α exponent, operating in the original space using the Euclidean distance and considering all data points as observers. It is also worth noting that the algorithm computes an empirical Monte Carlo rejection rate, denoted as $\hat{\rho}^{\text{MC}}_{\text{rej}}$. If a rigorous global Monte Carlo p-value is desired, we can define it by ranking the observed statistic relative to those obtained from the Monte Carlo projections, or by aggregating individual p-values through a well-established method such as Fisher's combined test.

4 EXPERIMENTS

In this section, we present a comprehensive suite of experiments conducted for both multivariate unimodality testing and estimating cluster counts in clustering tasks. Our decision to assess our algorithm for cluster estimation is driven by its complexity and wide practical relevance [Schubert, 2023]. A pertinent query pertains to which α-unimodality family we aim to detect. Despite its rigor, we opted to assess 1-unimodality regardless of the underlying data space. Empirical results indicated its efficacy even on challenging real-world datasets. Our algorithm is characterized by three parameters: M, ϵ, p. Following an initial exploration, we identified parameter values that consistently produced favorable outcomes. Specifically, we set $M = 100$, $\epsilon = 0.99$, $p = 0.99$, and chose a significance level $a = 0.01$.

Algorithm 1 mud-pod (α, X, a, M, p, ϵ)

```
Input: $\alpha$ (the positive unimodality index), $X$ (a set of real vectors), $a$ (a significance threshold), $M$ (number of simulations), $p$ (p-th percentile), $\epsilon$ (distance distortion)
Output: $\hat{\rho}^{\text{MC}}_{\text{rej}}$: an empirical Monte Carlo rejection rate
for $i = 1$ to $M$ do
    Project the points via a $\left\lceil \frac{8\log(|X|)}{\epsilon^2} \right\rceil$ random projection, resulting in $X\Pi_i$.
    Select an observer $\mathbf{o}$ from the p-th percentile of the projected Mahalanobis distances from the mean.
    Compute the set of distances from $o$, i.e., $\Pi \circ \mathcal{D}^o_{\mathbf{S}}$.
    Conduct a dip test on exponentiated $\Pi \circ \mathcal{D}^o_{\mathbf{S}}$ distances.
end for
return $\hat{\rho}^{\text{MC}}_{\text{rej}} = \frac{1}{M} \sum_{i=1}^{M} I\{T(\Pi_i \circ \mathcal{D}^o_{\mathbf{S}}) \leq a\}$
```

4.1 UNIMODALITY EXPERIMENTS

Table 1 presents an intricate assessment of the capability of three distinct tests — dip-dist (DD), mudpod (MP), and folding (F) — in discerning unimodal and multimodal datasets. DD and F tests are non-parametric and we also set $a = 0.01$. It is important to note that DD is a special case of mudpod, omitting α exponent, operating in the original space using the Euclidean distance and considering all data points as observers. The evaluation was carried out over ten distinct runs for each test on a combination of both synthetic and real-world data drawn from the MNIST dataset [LeCun et al., 1998]. Starting with synthetic unimodal datasets, namely the single 2D and 3D Gaussian distributions, we find a unanimous agreement across the three tests, with none indicating any instances of multimodality. Examining the synthetic bimodal distributions, 2D Moons and Circles show clear multimodality, confirmed by DD and MP tests with 100% detection. DD and MP also report 100% detection for bimodal Gaussians in 2D and 3D. However, DD's performance declines with three closely aligned Gaussians in both 2D and 3D, unlike MP's consistent 100% detection. The F test only identifies multimodality in the 2D Circles dataset.

Transitioning to real-world datasets, such as MNIST, offers a more intricate perspective. For our tests, we utilized the

Table 1: Performance comparison of dip-dist (DD), mudpod (MP), and folding (F) tests in determining unimodality or multimodality. The table displays the *percentage of multimodality* cases identified over 10 runs. 1000 points randomly sampled from synthetic sets and MNIST training set per experiment. For space constraints, Single MNIST experiment results are compressed, with digits divided by semi-colons summarizing the outcomes across all tests. G_n denotes a nD Gaussian distribution. $C(r)$ symbolizes the 2D equation of a circle with radius r. Utilizing parametric equations $U(\theta) = (\cos(\theta), \sin(\theta))$ and $L(\theta) = (1 - \cos(\theta), 1 - \sin(\theta) - 0.5)$ with $\theta \in [0, \pi]$.

Experiment	Distribution Details	DT	**MP**	F
Single 2D Gaussian	$G([0,0],\mathrm{I})$	0	0	0
Single 3D Gaussian	$G([0,0,0],\mathrm{I})$	0	0	0
Two 2D Circles	$\frac{1}{2}\left(C(0.5) + \mathcal{N}(0, 0.05^2 I)\right) + \frac{1}{2}\left(C(1) + \mathcal{N}(0, 0.05^2 I)\right)$	100	100	100
Two 2D Moons	$\frac{1}{2}\left(U(\theta) + \mathcal{N}(0, 0.05^2 I)\right) + \frac{1}{2}\left(L(\theta) + \mathcal{N}(0, 0.05^2 I)\right)$	100	100	0
Two 2D Gaussians	$0.5 \cdot G_1([1,4],\mathrm{I}) + 0.5 \cdot G_2([2,1],\mathrm{I})$	100	100	0
Three 2D Gaussians	$\frac{1}{3} \cdot G_1([t,t],\mathrm{I}) + \frac{1}{3} \cdot G_2([0,0],\mathrm{I}) + \frac{1}{3} \cdot G_3(-[t,t],\mathrm{I}) \mid t = 2.5$	50	100	0
Two 3D Gaussians	$0.5 \cdot G_1([1,4,2],\mathrm{I}) + 0.5 \cdot G_2([1,-2,3],\mathrm{I})$	100	100	0
Three 3D Gaussians	$\frac{1}{3} \cdot G_1([t,t,t],\mathrm{I}) + \frac{1}{3} \cdot G_2([0,0,0],\mathrm{I}) + \frac{1}{3} \cdot G_3(-[t,t,t],\mathrm{I}) \mid t = 2.9$	10	100	0
Single Digit MNIST	0; 2; 3; 4; 7; 8	0	0	100
Single Digit MNIST	1	100	100	100
Single Digit MNIST	5	0	10	100
Single Digit MNIST	6	10	10	100
Single Digit MNIST	9	10	20	100
Even Digits MMNIST	$\{0, 2, 4, 6, 8\}$	10	90	100
Odd Digits MMNIST	$\{1, 3, 5, 7, 9\}$	80	100	100
All Digits MMNIST	$\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$	0	100	100

flattened MNIST representations without any transformations. It is noteworthy that, while MNIST has labels, their alignment with clustering in the original space is not guaranteed. For instance, consider the digit “1”, representable as a single stroke or a combination of two distinct ones. It can be observed that for digits 0, 2, 3, 4, 7, and 8, the DD and MP tests consistently reject multimodality, while the F test falsely decides multimodality in all cases. Notably, all tests unanimously flag digit 1 as multimodal. The digits 5, 6, and 9 reveal varied outcomes, with the F test persistently recognizing multimodality. The conduction of three additional tests to assess multimodality of even, odd, and all MNIST subsets in multi-digit scenarios reveals a decline in DD’s efficacy. In summary, we observe a pronounced tendency of the F test to detect multimodality across various datasets, with DD and MP performance being comparable in simpler scenarios, but DD falling short in multi-digit scenarios.

Table 2: Impact of space, distance, and observer selection strategy on mudpod’s unimodality detection performance. Mudpod’s result agreement is shown for a mixture of two 2D Gaussians with confirmed ground truth unimodality. *Notation:* ‘O’ for Original, ‘RP’ for Randomly Projected, ‘E’ for Euclidean, ‘M’ for Mahalanobis, ‘R’ for Random and and ‘P’ for Percentile.

Space	Distance	Observer	Agreement (%)
O	E	R	0.80
O	E	P	0.87
O	M	R	0.82
O	M	P	0.87
RP	E	R	0.85
RP	E	P	0.90
RP	M	R	0.92
RP	M	P	0.95

Ablation Study: Our test comprises several components, including random projections, Mahalanobis distance, and uniform sampling of distances from origin percentiles. In Table 2, we present an ablation study to evaluate the significance of each component. We have generated datasets from a mixture of two 2D Gaussians as the target distribution, for which the bimodality or unimodality can be determined analytically [Konstantellos, 1980]. The reported performance is aggregated from a series of experiments conducted with four different significance levels $0.001, 0.005, 0.01, 0.05$, across 1000 distinct data sets, with each experiment executed 10 times. Our ablation study reveals key insights into the performance of different observer picking strategies, spaces, and distances for unimodality detection. Primarily, the percentile strategy (P) for observer picking demonstrated superiority over the random strategy (R) across all tested combinations of space and distance. Furthermore, the Ma-

Table 3: The table presents the number of clusters (K) and the associated NMI values obtained by various methods on different datasets. Values marked with † could not be computed due to memory constraints or were terminated after 8 hours. All results are represented as mean ± standard deviation over 10 executions. For the k-means algorithm, the correct number of clusters was always predefined.

Method	USPS		MNIST		F-MNIST		HAR	
	k	NMI	k	NMI	k	NMI	k	NMI
Ground truth	10	1.0	10	1.0	10	1.0	5	1.0
k-means	-	0.61±0.00	-	0.49±0.00	-	0.51±0.00	-	0.61±0.01
x-means	35±0	0.61±0.01	35±0	0.55±0.00	35±0	0.51±0.00	41±2	0.56±0.01
g-means	35±0	0.61±0.00	35±0	0.55±0.00	35±0	0.51±0.00	931±29	0.42±0.00
pg-means	2±1	0.14±0.07	2±1	0.18±0.09	4±2	0.31±0.11	2±1	0.14±0.05
dip-means	4±0	0.44±0.00	1±0	0.01±0.05	9±2	0.50±0.01	3±0	0.73±0.00
hdbscan	13±0	0.38±0.00	36±0	0.33±0.00	3±0	0.05±0.00	3±0	0.52±0.00
SpecialK	1±0	0.00±0.00	5±2	0.04±0.02	1±0	0.00±0.00	1±0	0.00±0.00
fold-means	31±0	0.59±0.01	31±0	0.55±0.01	31±0	0.54±0.01	11±0	0.61±0.00
mp-means	8±2	0.62±0.03	9±2	0.55±0.04	9±1	0.54±0.01	3±0	0.73±0.00

Method	Optdigits		Pendigits		Isolet		TCGA	
	k	NMI	k	NMI	k	NMI	k	NMI
Ground truth	10	1.0	10	1.0	26	1.0	5	1.0
k-means	-	0.69±0.01	-	0.69±0.01	-	0.73±0.01	-	0.80±0.01
x-means	35±0	0.71±0.01	35±0	0.70±0.01	233±4	0.66±0.01	20±1	0.68±0.01
g-means	35±0	0.72±0.01	35±0	0.70±0.01	101±6	0.69±0.01	283±48	0.49±0.04
pg-means	1±0	0.02±0.07	3±1	0.34±0.18	†	†	1±0	0.02±0.04
dip-means	1±0	0.00±0.00	16±1	0.71±0.02	4±0	0.44±0.01	2±0	0.50±0.01
hdbscan	21±0	0.71±0.00	38±0	0.72±0.00	4±0	0.04±0.00	7±0	0.75±0.00
SpecialK	3±0	0.05±0.00	3±1	0.34±0.21	1±0	0.00±0.00	1±0	0.00±0.00
fold-means	4±1	0.45±0.11	1±0	0.00±0.00	1±0	0.00±0.00	†	†
mp-means	8±1	0.67±0.06	14±1	0.70±0.01	20±7	0.63±0.14	6±1	0.95±0.03

halanobis distance consistently emerged as a more effective metric compared to the Euclidean distance. This performance difference was especially evident in the Randomly Projected space, suggesting that the intrinsic characteristics of the Mahalanobis distance, e.g., accounting for data covariance, plays a pivotal role in enhancing detection reliability. Notably, the randomly projected (RP) space exhibited a pronounced advantage over the original space in our assessments. This superiority held true irrespective of the distance metric or observer strategy employed. Such a trend strongly indicates that the RP space aligns more coherently with the ground truth unimodality, offering better detection capabilities compared to the original space. In summary, our findings recommend a strategic combination of the P strategy, Mahalanobis distance, and the RP space.

4.2 CLUSTERING EXPERIMENTS

This section analyzes the performance of *mp-means*, an approach that incorporates mud-pod into the dip-means wrapper method, replacing dip-dist. Specifically, both mp-means and dip-means employ incremental k-means clustering based on testing clusters for unimodality. Starting with one cluster (the entire dataset), they incrementally increase k by splitting multimodal clusters, terminating when all clusters are deemed unimodal. They differ in the cluster unimodality assessment. Dip-means uses the dip-test criterion, while mp-means uses our proposed unimodality test. Upon detecting a multimodal cluster, they split the one with the highest dip value. It's split into two using the 2-means algorithm or assigning clusters at mean ± standard deviation, using the cluster's mean and standard deviation. In our work, we opt for the latter for computational efficiency. In this way, the number of clusters is increased to $k+1$ and the $k+1$ centers are updated via k-means. A similar integration of the folding test yields the *fold-means* algorithm.

In Table 3, we compare the performance of various clustering methods estimating the number of clusters across several datasets. It is important to note that in all our experiments, we use the raw flattened data encoding and apply only a feature-wise z-transformation. Although our method employs the Mahalanobis distance and is scale-agnostic, we ob-

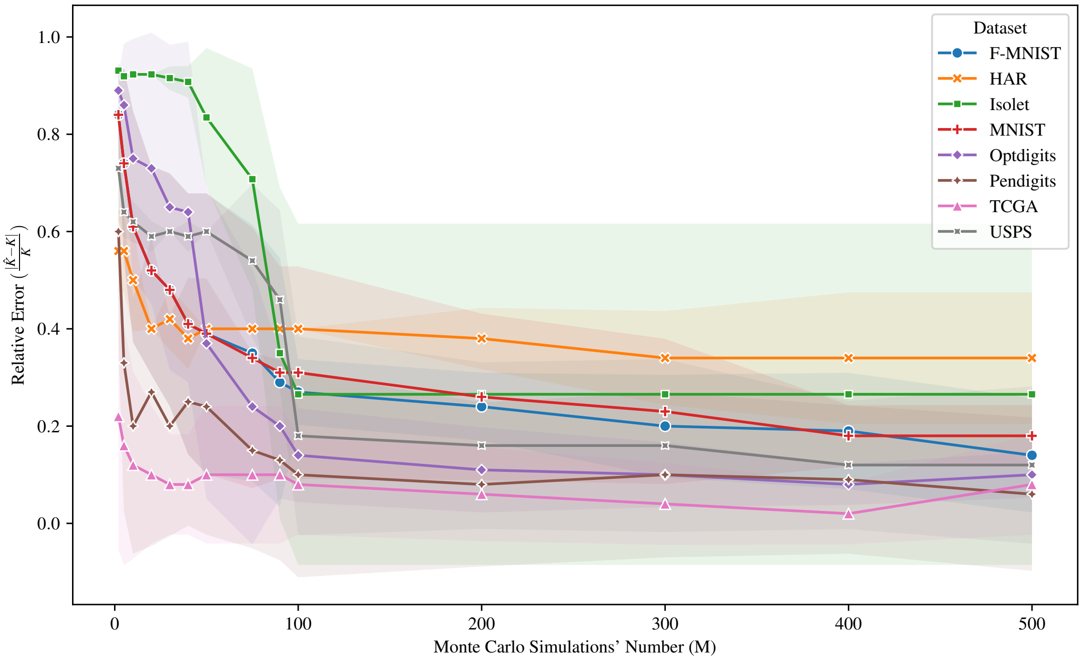

Figure 1: The plot shows the relative error between the estimated and actual number of clusters generated by mp-means, against increasing Monte Carlo simulations. Ten executions per experiment, variance depicted. The plot is more discernible in color.

served that scaling profoundly affects other algorithms. Our experiments include datasets like USPS, MNIST, Fashion-MNIST (F-MNIST), Human Activity Recognition (HAR), Optdigits, Pendigits, Isolet, and TCGA. Detailed dataset descriptions are provided in Appendix 6. As ground truth number of clusters, the number of classes was considered in all datasets. We benchmark against classic algorithms, including x-means [Pelleg and Moore, 2000], g-means [Hamerly and Elkan, 2003], pg-means [Feng and Hamerly, 2006], dip-means [Kalogeratos and Likas, 2012], SpecialK [Hess and Duivesteijn, 2020] and hdbscan [Campello et al., 2013]. A detailed overview of the baseline algorithms and their hyper-parameter setups are provided in Appendix 7. Our evaluation metrics consist of the estimated number of clusters (k) and the Normalized Mutual Information (NMI). NMI values lie between [0, 1], where 1 signifies a perfect match and 0 represents an arbitrary result.

This experiment set, encompassing both the complexity of many modes and inherent clustering errors, is notably more challenging than unimodality testing. We observe that most of the classical approaches, notably x-means and g-means, tend to predict a high value for k, frequently estimating around 35 clusters for datasets such as USPS, MNIST, and F-MNIST. Similarly, x-means, consistently overestimates the k value on these datasets, accompanied by NMI values not consistently reaching the top tier. In contrast, pg-means often significantly underestimates the number of clusters, suggesting a notable diversion from the ground truth. This assertion is further supported by its suboptimal NMI values across multiple datasets, with values as low as 0.14 for the USPS dataset. Hdbscan shows a varying range in k values across datasets, highlighting its adaptability. However, this variability does not always correlate with high NMI values. Its performance on the USPS dataset, where it estimates 13 clusters with an NMI of 0.38, is a case in point.

SpecialK, which leverages nonparametric concentration bounds on a spectral embedding, consistently underestimates the number of clusters across all datasets. As shown in Table 3, it typically returns $k = 1$ (or very low values) and yields near-zero NMI scores on datasets like MNIST and OptDigits, even when it identifies a larger number of clusters. This behavior suggests that the Bernstein-inequality criterion is overly conservative on these high-dimensional and heterogeneous real-world data, collapsing all points into a single cluster. Turning our attention to the HAR dataset, the dip-means and mp-means methods are particularly notable. They estimate k values closely aligned with the ground truth of 5 and simultaneously achieve the highest NMI scores, specifically 0.73. This is indicative of their capability in deciphering the cluster structure inherent to the data. In the context of the Optdigits and Pendigits datasets, dip-means struggles to align its k estimations with the ground truth, predicting values of 1 and 16, respectively. Notably, mp-means appears more consistent, with k values of 8 and 14 for Optdigits and Pendigits, respectively, and corresponding NMI scores that are commendable. In an experiment with k-means using accurate cluster count, the NMI of mp-means closely aligns with, or even exceeds, that of k-means. Intriguingly, the projection dimension of mp-means varies between subclusters. The experiments also underscore its robustness across multiple projection spaces. In conclusion, the consistent performance of mp-means underscores its utility in solving clustering problems with unknown number of clusters.

Ablation Study on Simulations' Number: In Figure 1, the relative error between the estimated number of clusters K and the true value, as deduced by the mp-means algorithm, is depicted against the number of Monte Carlo simulations (M). The shaded regions around each line represent the variance over 10 executions, highlighting result consistency. As observed, for all datasets, an increase in the number of Monte Carlo simulations tends to correspond with a decline in the relative error, signifying an enhancement in the accuracy of k estimation. While some datasets exhibit minimal spread, others display significant variance, suggesting the need for more Monte Carlo simulations. This suggests a need for more projections on certain datasets. However, this variance does not significantly impede k estimation. Summarizing, Figure 1 substantiates the notion that increasing the number of Monte Carlo simulations augments the precision of the mp-means algorithm's k estimation, albeit with varying degrees of improvement across different datasets. Interestingly, we note that convergence occurs at around 100 simulations, which is significant for practical use.

5 CONCLUSIONS

In this paper, we addressed the challenges associated with generalizing unimodality testing to higher dimensions. Building upon the notion of α-unimodality, we presented a novel methodology that provides a robust approach to multivariate unimodality testing. Utilizing point-to-point distancing and linear random projections, our approach bridges the gap between the simplicity of one-dimensional unimodality confirmation and the intricacies of its higher-dimensional counterpart. By integrating our unimodality test with k-means, we introduced a new incremental clustering approach that automatically estimates the cluster count while performing clustering. Empirical evaluations affirm our test's efficacy in unimodality assessments and clustering scenarios, highlighting its vital applicability across diverse data-driven domains. Future exploration will focus on identifying additional α-unimodality preserving operations to enhance our test, e.g, exploiting JL lemma's relaxations through matrix sketching, experimenting with α-unimodality for $\alpha \neq 1$, and applying the test to various data analysis tasks, such as covariate shift and time series change detection.

Acknowledgements

We thank the anonymous reviewers for their valuable comments and suggestions, which have helped improve the quality of this work. We also thank Ralph Urlus for his implementation of the diptest library[2], which we used for computing dip values and statistics.

References

Arnab Bhattacharya, Purushottam Kar, and Manjish Pal. On low distortion embeddings of statistical distance measures into low dimensional spaces. In Sourav S. Bhowmick, Josef Küng, and Roland Wagner, editors, *Database and Expert Systems Applications*, pages 164–172, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-03573-9.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-37456-2.

Paraskevi Chasani and Aristidis Likas. The uu-test for statistical modeling of unimodal data. *Pattern Recognition*, 122:108272, 2022. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2021.108272. URL https://www.sciencedirect.com/science/article/pii/S0031320321004520.

Tao Dai. On multivariate unimodal distributions, 1989. URL https://open.library.ubc.ca/collections/ubctheses/831/items/1.0097413.

Sanjoy Dasgupta. Learning mixtures of gaussians. *40th Annual Symposium on Foundations of Computer Science (Cat. No.99CB37039)*, pages 634–644, 1999. URL https://api.semanticscholar.org/CorpusID:8338511.

Constantinos Daskalakis, Ilias Diakonikolas, Rocco A. Servedio, Gregory Valiant, and Paul Valiant. *Testing k-Modal Distributions: Optimal Algorithms via Reductions*, pages 1833–1852. 2013. doi: 10.1137/1.9781611973105.131. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611973105.131.

[2] https://pypi.org/project/diptest/

Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A. Servedio. Learning k-modal distributions via testing. *Theory of Computing*, 10(20): 535–570, 2014. doi: 10.4086/toc.2014.v010a020. URL https://theoryofcomputing.org/articles/v010a020.

Sudhakar Dharmadhikari and Kumar Joag-Dev. *Unimodality, convexity, and applications*. Elsevier, 1988.

Persi Diaconis and David Freedman. Asymptotics of Graphical Projection Pursuit. *The Annals of Statistics*, 12(3):793 – 815, 1984. doi: 10.1214/aos/1176346703. URL https://doi.org/10.1214/aos/1176346703.

Lutz Dümbgen, Del Conte-Zerial, et al. On low-dimensional projections of high-dimensional distributions. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, volume 9, pages 91–105. Institute of Mathematical Statistics, 2013.

Robin Dunn, Aditya Gangrade, Larry Wasserman, and Aaditya Ramdas. Universal inference meets random projections: a scalable test for log-concavity. *arXiv preprint arXiv:2111.09254*, 2021.

Yu Feng and Greg Hamerly. Pg-means: learning the number of clusters in data. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL https://proceedings.neurips.cc/paper_files/paper/2006/file/a9986cb066812f440bc2bb6e3c13696c-Paper.pdf.

Carlos Fernandez-Granda. Random projections and compressed sensing. Lecture Notes in Optimization-based Data Analysis, 2016. URL https://cims.nyu.edu/~cfgranda/pages/OBDA_spring16/material/random_projections.pdf. New York University.

Greg Hamerly and Charles Elkan. Learning the k in k-means. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL https://proceedings.neurips.cc/paper_files/paper/2003/file/234833147b97bb6aed53a8f4f1c7a7d8-Paper.pdf.

J. A. Hartigan and P. M. Hartigan. The Dip Test of Unimodality. *The Annals of Statistics*, 13(1):70 – 84, 1985. doi: 10.1214/aos/1176346577. URL https://doi.org/10.1214/aos/1176346577.

Sibylle Hess and Wouter Duivesteijn. k is the magic number—inferring the number of clusters through nonparametric concentration inequalities. In Ulf Brefeld, Elisa Fromont, Andreas Hotho, Arno Knobbe, Marloes Maathuis, and Céline Robardet, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 257–273, Cham, 2020. Springer International Publishing. ISBN 978-3-030-46150-8.

Harold Hotelling. The Generalization of Student's Ratio. *The Annals of Mathematical Statistics*, 2(3):360 – 378, 1931. doi: 10.1214/aoms/1177732979. URL https://doi.org/10.1214/aoms/1177732979.

Jonathan J Hull. Database for handwritten text recognition research. In *IEEE Transactions on pattern analysis and machine intelligence*, volume 16, pages 550–554. IEEE, 1994.

William B. Johnson and Joram Lindenstraus. Extensions of lipschitz mappings into hilbert space. *Contemporary mathematics*, 26:189–206, 1984.

Argyris Kalogeratos and Aristidis Likas. Dip-means: an incremental clustering method for estimating the number of clusters. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/a8240cb8235e9c493a0c30607586166c-Paper.pdf.

Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The uci machine learning repository, 2023. URL https://archive.ics.uci.edu/ml. Last accessed: October 31, 2023.

A. Ya. Khintchine. On unimodal distributions. *Izv. Nsuchno-Issled. Inst. Mat. Meh. Tomsk. Goa. Univ.*, 2:1–7, 1938. in Russian.

A. Konstantellos. Unimodality conditions for gaussian sums. *IEEE Transactions on Automatic Control*, 25(4):838–839, 1980. doi: 10.1109/TAC.1980.1102410.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

E. L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005. ISBN 0-387-98864-5.

Miles Lopes, Laurent Jacob, and Martin J Wainwright. A more powerful two-sample test in high dimensions using random projection. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/5487315b1286f907165907aa8fc96619-Paper.pdf.

Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, 18, 07 1990. doi: 10.1214/aop/1176990746.

Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205, 2017.

Andrei Novikov. PyClustering: Data mining library. *Journal of Open Source Software*, 4(36):1230, apr 2019. doi: 10.21105/joss.01230. URL https://doi.org/10.21105/joss.01230.

Richard A. Olshen and Leonard J. Savage. A generalized unimodality. *Journal of Applied Probability*, 7(1):21–34, 1970. ISSN 00219002. URL http://www.jstor.org/stable/3212145.

Dan Pelleg and Andrew W. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, page 727–734, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

Ping Li Radhendushka Srivastava and David Ruppert. Raptt: An exact two-sample test in high dimensions using random projections. *Journal of Computational and Graphical Statistics*, 25(3):954–970, 2016. doi: 10.1080/10618600.2015.1062771. URL https://doi.org/10.1080/10618600.2015.1062771.

Erich Schubert. Stop using the elbow criterion for k-means and how to choose the number of clusters instead. *SIGKDD Explor. Newsl.*, 25(1):36–42, jul 2023. ISSN 1931-0145. doi: 10.1145/3606274.3606278. URL https://doi.org/10.1145/3606274.3606278.

Olimjon Shukurovich Sharipov. *Glivenko-Cantelli Theorems*, pages 612–614. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-04898-2. doi: 10.1007/978-3-642-04898-2_280. URL https://doi.org/10.1007/978-3-642-04898-2_280.

Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouët. Are your data gathered? In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2210–2218, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3219994. URL https://doi.org/10.1145/3219819.3219994.

B. W. Silverman. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. Series B (Methodological)*, 43(1):97–99, 1981. ISSN 00359246. URL http://www.jstor.org/stable/2985156.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

A Multivariate Unimodality Test Harnessing the Dip Statistic of Mahalanobis Distances Over Random Projections: Supplementary Materials

Prodromos Kolyvakis[1] **Aristidis Likas**[2]

[1] ORamaVR SA, Geneva, Switzerland & Talos Automata SMPC, Ioannina, Greece
[2] Department of Computer Science and Engineering, University of Ioannina, 45110 Ioannina, Greece

6 CLUSTERING DATASETS

Table 4 summarizes the benchmark datasets used in our experiments, varying in size, dimensions, number of classes k (considered as ground-truth number of clusters), complexity, and domain. The datasets USPS, MNIST, and F-MNIST feature images of handwritten digits and fashion items, respectively. Specifically, USPS and MNIST contain grayscale images of handwritten digits from 0 to 9, with MNIST having a resolution of 28 x 28 pixels. In contrast, F-MNIST, or Fashion MNIST, encompasses grayscale images of ten clothing types, also at a resolution of 28 x 28 pixels. The Human Activity Recognition (HAR) dataset captures data from sensors to classify human activities, such as walking and sitting. OptDigits and Pendigits both pertain to handwritten digits: OptDigits consists of 8 x 8 resolution images, while Pendigits uses 16-dimensional vectors containing pixel coordinates. The Isolet dataset is an assemblage of speech recordings, representing the sounds of spoken letters, characterized by vectors with 617 spectral coefficients derived from the speech. Finally, TCGA is a compendium of gene expression profiles garnered from RNA sequencing of diverse cancer specimens, inclusive of clinical data, normalized counts, gene annotations, and pathways for five cancer types. For USPS, MNIST, and F-MNIST, we use the test datasets with 10000 points. For other datasets, we utilize the full set, typically containing fewer than 10000 points, except for Pendigits. Despite better preliminary results, we we constrained the sizes of USPS, MNIST, and F-MNIST to their test sets to avoid sample number bias.

7 CLUSTERING ALGORITHMS

In this section, we provide an overview of the clustering algorithms employed in our experiments. We focus on methods that automatically estimate the number of clusters. We benchmark our approach against well-established algorithms, namely x-means [Pelleg and Moore, 2000], g-means [Hamerly and Elkan, 2003], pg-means [Feng and Hamerly, 2006], dip-means

Table 4: Datasets for our experiments: Size indicates data instances, Dimension shows original encoding's flattened size, and k represents ground-truth class labels.

Dataset	**Type**	**Description**	**Size**	**Dimension**	**k**	**Source**
USPS	Image	Handwritten digits	10000	256	10	Hull [1994]
MNIST	Image	Handwritten digits	10000	784	10	LeCun et al. [1998]
F-MNIST	Image	Zalando's article image	10000	784	10	Xiao et al. [2017]
HAR	Time-series	Smartphone-based activity	2947	561	5	Kelly et al. [2023]
Optdigits	Image	Handwritten digits	1797	64	10	Kelly et al. [2023]
Pendigits	Time-series	Handwritten digits	10992	16	10	Kelly et al. [2023]
Isolet	Spectral	Speech recordings pronouncing letters	6238	617	26	Kelly et al. [2023]
TCGA	Tabular	Cancer gene expression profiles	801	20531	5	Kelly et al. [2023]

[Kalogeratos and Likas, 2012], SpecialK [Hess and Duivesteijn, 2020] and hdbscan [Campello et al., 2013]. Given our design's adherence to the original dataspace, we exclude methods combining learning embeddings and clustering, like deep clustering approaches. To select the best number of clusters, x-means incorporates a regularization penalty guided by the Bayesian Information Criterion (BIC), which accounts for model complexity. However, this approach excels mainly with abundant data and distinct spherical clusters. Another extension, g-means, tests the assumption that each cluster originates from a Gaussian distribution. Given the challenges of statistical tests in high dimensions, g-means first projects cluster datapoints onto a high variance axis and then employs the Anderson-Darling test for normality. Clusters failing this test are iteratively split to identify the Gaussian mixture. Conversely, projected g-means (pg-means) assumes a Gaussian mixture for the entire dataset, evaluating the model as a whole. It relies on the EM algorithm, constructing one-dimensional projections of both the dataset and the learned model, and subsequently assessing model fit in the projected space using the Kolmogorov-Smirnov (KS) test. This approach's strength lies in identifying overlapping Gaussian clusters of varying scales and covariances.

Moreover, hdbscan enhances dbscan by transforming it into a hierarchical clustering method and subsequently employs a technique to derive a flat clustering based on cluster stability. Hdbscan aims to obtain an optimal cluster solution by maximizing the aggregate stability of chosen clusters. Both mp-means and fold-means build upon the wrapper method (dip-means) introduced in the work of Kalogeratos and Likas [2012]. Dip-means draws upon this approach and uses the dip-dist criterion internally. These methods incrementally increase cluster count and apply unimodality tests to clusters shaped by k-means. The process concludes when all clusters are characterized as unimodal. SpecialK introduces a nonparametric statistical approach to determine the number of clusters without assuming specific data distributions. It assesses whether two clusters likely originate from the same distribution by computing a bound on the probability that they do, using only sample means and variances. This method serves as a wrapper applicable to any clustering algorithm minimizing the k-means objective, including Gaussian mixtures and Spectral Clustering.

For baseline model hyperparameters, we adopt a significance level $a = 0.01$ for all relevant algorithms. The default setups are used for x-means and g-means from Novikov [2019], pg-means as per Feng and Hamerly [2006], SpecialK as per Hess and Duivesteijn [2020] and hdbscan following McInnes et al. [2017]. The default hyperparameters for dip-means are those provided by the authors[1]. For the folding test, we utilize the publicly available Python version[2]. Given the folding test's propensity to indicate multimodality, we set a cap on k for fold-means. Specifically, for all algorithms necessitating a maximum k value, we set $k_{\text{max}} = 300$.

8 RUNTIME COMPLEXITY

The computational efficiency of the *mud-pod* algorithm is crucial for its practical applicability, especially when dealing with large datasets. Here, we provide a detailed analysis of its runtime complexity. The *mud-pod* algorithm (Algorithm 1) operates within a main loop that runs for M iterations. In each iteration, the following key computational steps are performed:

1. **Random Projection**: Each iteration begins by projecting the dataset X, containing n points in $\mathbb{R}^d$, into a lower-dimensional space $\mathbb{R}^q$ using a random projection matrix Π_i. The dimension q is determined by the Johnson-Lindenstrauss lemma to be $q = \left\lceil \frac{8 \log n}{\epsilon^2} \right\rceil$, where ϵ is the allowable distortion.

   Generating the random projection matrix $\Pi_i \in \mathbb{R}^{d \times q}$ involves sampling dq independent Gaussian random variables. Projecting the dataset requires computing the matrix product $X\Pi_i$, which has a computational complexity of $\mathcal{O}(ndq)$.

2. **Covariance Computation & Inversion**: Computing the covariance matrix Σ_Π of the projected data involves $\mathcal{O}(nq^2)$ operations, given that each element of the $q \times q$ covariance matrix requires summing over n points. Inverting Σ_Π requires $\mathcal{O}(q^3)$ time using standard matrix inversion algorithms.

3. **Mahalanobis Distance Computation**: Calculating the Mahalanobis distances from the observer $\mathbf{o}$ to each of the $n - 1$ other points requires $\mathcal{O}(nq^2)$ operations. Each distance computation involves a quadratic form in q dimensions, which takes $\mathcal{O}(q^2)$ time, and this is done for $n - 1$ points.

4. **Observer Selection**: Selecting an observer from the p-th percentile of the distances from the mean involves computing all distances from the mean and then selecting based on the percentile. Computing distances from the mean takes $\mathcal{O}(nq)$ time. Finding the p-th percentile can be done in $\mathcal{O}(n)$ time using selection algorithms.

[1] https://kalogeratos.com/psite/material/dip-means/

[2] https://github.com/asiffer/python3-libfolding

5. **Dip Test Execution**: The dip test is performed on the set of Mahalanobis distances, which contains $k = n - 1$ elements. The computational complexity of the dip test is $\mathcal{O}(k \log k)$ [Hartigan and Hartigan, 1985], primarily due to the sorting step required for computing the empirical cumulative distribution function.

Considering these steps, the per-iteration computational complexity is dominated by the random projection step, which is $\mathcal{O}(ndq)$. The other steps involve operations that are either of lower order or involve dimensions (q) significantly smaller than d and n.

Since the main loop runs for M iterations, the total computational complexity of the *mud-pod* algorithm is:

$$\mathcal{O}\left(M\left(ndq + nq^2 + q^3 + nq + n \log n\right)\right).$$

Given that $q = \left\lceil \frac{8 \log n}{\epsilon^2} \right\rceil$, we observe that $q = \mathcal{O}(\log n)$. Therefore, q, q^2, and q^3 grow logarithmically with n, and for large n, they are significantly smaller than n and d. Simplifying the complexity expression by considering the dominant terms, we have:

$$\mathcal{O}\left(M\left(nd \log n + n \log^2 n + \log^3 n + n \log n\right)\right).$$

Since d is typically much larger than $\log n$, and $\log^3 n$ is negligible compared to $nd \log n$ for large n, the total runtime complexity simplifies to:

$$\mathcal{O}\left(Mnd \log n\right).$$

Therefore, *mud-pod* is well-suited for high-dimensional data analysis, providing a computationally feasible method for multivariate unimodality testing.