The Projected Dip-means Clustering Algorithm

Theofilos Chamalis
Department of Computer Science & Engineering
University of Ioannina
GR 45110, Ioannina, Greece
thchama@cs.uoi.gr

Aristidis Likas
Department of Computer Science & Engineering
University of Ioannina
GR 45110, Ioannina, Greece
arly@cs.uoi.gr

ABSTRACT

One of the major research issues in data clustering concerns the estimation of number of clusters. In previous work, the dip-means clustering algorithm has been proposed as a successful attempt to tackle this problem. Dip-means is an incremental clustering algorithm that uses a statistical criterion called dip-dist to decide whether a set of data objects constitutes a homogeneous cluster or it contains subclusters. The novel aspect of dip-dist is the introduction of the notion of data unimodality when deciding for cluster compactness and homogeneity. More specifically, the use of dip-dist criterion for a set of data objects requires the application of a univariate statistic hypothesis test for unimodality (the so called dip-test) on each row of the distance (or similarity) matrix containing the pairwise distances between the data objects. In this work, we propose an alternative criterion for deciding on the homogeneity of a set of data vectors that is called projected dip. Instead of testing the unimodality of the the row vectors of the distance matrix, the proposed criterion is based on the application of unimodality tests on appropriate 1-d projections of the data vectors. Therefore it operates directly on the data vectors and not on the distance matrix. We also present the projected dip-means (pdip-means) algorithm that is an adaptation of dip-means using the proposed pdip criterion to decide on cluster splitting. We conducted experiments using the pdip-means and the dip-means algorithms on artificial and real datasets to compare their clustering performance and provide empirical conclusions from the obtained experimental results.

CCS CONCEPTS

• **Information systems → Clustering**; • **Theory of computation → Unsupervised learning and clustering**; • **Computing methodologies → Cluster analysis**;

KEYWORDS

Data clustering, number of clusters, k-means, dip-means, data projections, dip test for unimodality

ACM Reference Format:
Theofilos Chamalis and Aristidis Likas. 2018. The Projected Dip-means Clustering Algorithm. In *SETN '18: 10th Hellenic Conference on Artificial Intelligence, July 9–15, 2018, Rio Patras, Greece.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3200947.3201008

SETN '18, July 9–15, 2018, Rio Patras, Greece

ACM ISBN 978-1-4503-6433-1/18/07…$15.00
https://doi.org/10.1145/3200947.3201008

1 INTRODUCTION

Data clustering is an important machine learning and data mining task with wide applications in various scientific and practical fields. The goal of clustering is to automatically reveal the underlying structure of a given set of data. More specifically, given a dataset and a similarity (or distance) measure, clustering methods produce a partition of the dataset into clusters (ie. groups) of similar objects. Typical clustering methods (such a k-means or spectral clustering) require that the number of clusters is provided as input by the user. In the same spirit, other popular techniques (e.g. DBSCAN), require other types of parameters to be user defined such as cluster radius, minimum distance between clusters etc.

A very important issue in clustering is to develop methods that do not require as input critical user defined parameters. In this work the critical parameter is the number of clusters, thus we focus on the significant problem of automatically estimating the number of clusters k in a dataset.

A fundamental issue related to the number of clusters estimation problem is whether a given set of data objects constitutes a compact (ie. content homogeneous) cluster or it contains two or more subclusters. If a successful test is used to decide on this issue, then it is possible to develop incremental (top-down) clustering methods that proceed as follows. We start with a single cluster and proceed by splitting clusters. At each stage, assuming k is the number clusters i) we apply the decision test to every cluster of the current solution and determine the clusters that contain subclusters, ii) we split one of those clusters in two clusters (ie. the number of clusters is increased by one, k=k+1), iii) we refine the k+1 clusters (using for example the k-means algorithm). The above three steps are repeated until we reach a clustering solution where all clusters are considered to be compact. The basic components in this approach are: i) the decision test (the major one), ii) how to select the cluster to be splitted (if more than one candidate clusters exist) and iii) the cluster refinement method (usually the k-means or Expectation-Maximization algorithm are used).

In this spirit, several algorithms have been proposed with the major difference being in the way that a cluster is decided as split candidate or not. The first approach was x-means [7] which uses Bayesian Information Criterion (BIC) that works well only in cases where there are plenty of data and well-separated spherical clusters. In [2] the G-means algorithm has been proposed, that uses a statistical test for the hypothesis that each cluster has been generated from a Gaussian distribution. The algorithm first projects the datapoints of a cluster on an axis of high variance and then applies Anderson-Darling statistic with a fixed significance level. Clusters

that are not accepted are split repeatedly until no further split is possible.

The dip-means algorithm has been proposed in [5] as a major extension of g-means that checks for unimodality instead of Gaussianity. Dip-means is an incremental clustering algorithm, that uses a statistical criterion for unimodality, called dip-dist, that can be applied into a data subset in order to determine if it contains a single or multiple cluster structures. The dip-dist criterion is based on the notion of viewer which is an arbitrary data object whose role is to suggest on the unimodality of the dataset by forming the set of its distances to all other data objects and applying a unimodality test, called dip-test [3], on this set of distances. In practice, the dip-dist works by applying the dip-test for unimodality on each row of the distance matrix containing the pairwise distances among the objects of the dataset. In case of a homogeneous cluster, the distribution of distances is expected to be unimodal. In the case where distinct subclusters exist, the distribution of distances should exhibit distinct modes. The dip-means algorithm has been empirically found [5] to be highly superior to previous approaches such as X-means and g-means in estimating the correct number of clusters.

In this work we propose an alternative way to decide on the homogeneity of a set of data vectors that we call projected dip (pdip) criterion. In this criterion the unimodality tests are applied on appropriate 1-d projections of the data vectors, while the dip-dist criterion used in the dip-means algorithm applies unimodality tests on the rows of the distance matrix. Then we present pdip-means (projected dip-means) which is an incremental clustering algorithm analogous to dip-means, but employing the projected dip criterion for deciding of whether to split a cluster or not.

In section 2 we review the dip-means algorithm along with the employed dip-dist criterion. In section 3, the projected dip criterion is presented followed by the proposed projected dip-means algorithm. Comparative experimental results on synthetic and real datasets are provided in section 4 along with the empirical conclusions that have been drawn. Finally section 5 provides conclusions and directions for future research.

2 DIP-MEANS AND THE DIP-DIST CRITERION

The dip-means algorithm [5] is an incremental clustering algorithm that relies on the dip-dist criterion [5] to decide on the homogeneity of a cluster and uses the k-means to algorithm to refine the clustering solutions after each cluster splitting. The dip-dist criterion is based on unimodality tests and more specifically on the dip-test [3] that is briefly described next.

2.1 Dip test for unimodality

The dip-test for unimodality [3] takes as input an 1-d dataset (e.g. a set of real numbers) and provides a statistical decision of whether this set is unimodal or not. It examines the underlying empirical probability distribution (e.g. the histogram) of the given set of numbers and decides whether it contains a single or more than one mode (peak). Given a set of real numbers $X = \{x_1, x_2, \ldots, x_n\}$ the dip-test computes the so-called dip value, which is the departure from unimodality of the empirical cumulative distribution (cdf) $F(x)$ of X, which is computed as:

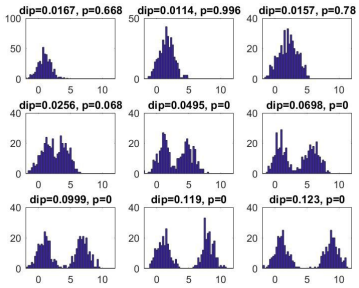


Figure 1: Unimodal and bimodal histograms and the corresponding dip and p-values.

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x)$$

where $I(z)$ is the indicator function. The cdf $F(x)$ is considered unimodal with mode the region $m = (tL, tU)$, if it is convex in (fk‼∞, tL], has constant slope in $[tL, tU]$, and concave in $[tU, \infty)$. This implies the non-increasing probability density behavior when moving away from the mode. The dip value provided by the dip test is the distance of the empirical cdf $F(x)$ from the closest unimodal distribution. Details on dip computations can be found in [3]. Low values of dip indicate unimodality of X, while high values indicate multimodality. Given a 1-d dataset X of size n, the complexity of computing dip(X) is $O(n)$ [3]. The dip-test returns not only the dip value, but also a p-value. The null hypothesis $H0$ that $F(x)$ is unimodal, is accepted at significance level α if $p - value > \alpha$, otherwise $H0$ is rejected in favor of the alternative hypothesis $H1$ which suggests multimodality. The computation of the p-value for a unimodality test uses bootstrap samples and expresses the probability of dip(X) being less than the dip value of a set U_{rn} of n observations (where n is the size of X) sampled from the $U[0, 1]$ uniform distribution:

$$p - value = \#[dip(X) < dip(U_{rn})]/b, r = 1, \ldots, b$$

where b is the number of bootstrap samples (b=1000 in our experiments).

It should be stressed that for each value of n, the bootstrap samples U_{rn} do not depend on the dataset X, therefore they can be computed only once, along with the corresponding values dip(U_{rn}). Those dip(U_{rn}) values are sorted, stored and subsequently used each time we need to compute the p-value corresponding to a given dip(X) value. In this way the computational cost of computing the p-value given the dip value is negligible.

To provide intuitive insight on the results provided by the dip-test, we present in Fig. 1 several unimodal and bimodal histograms along with the dip and p-values provided by the dip-test. It can be clearly observed that as we move from unimodality (1st row) to bimodality (2nd and 3rd rows) the dip value increases and the p-value decreases.

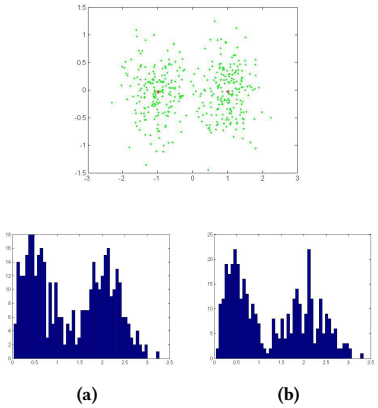
2.2 Dip-dist criterion

The dip-dist criterion has been proposed in [5] for deciding whether a set data objects X is homogeneous with respect to content or not. In other words it is used to decide whether dataset X contains subclusters or not. The novel aspect of dip-dist is that it relates homogeneity to unimodality. As stressed in [5], the empirical density of an acceptable cluster should have a single mode: a region where the density becomes maximum, while non-increasing density is observed when moving away from the mode. There are no other underlying assumptions about the shape of a cluster and the distribution that generated the empirically observed unimodal property. Unimodality is a very general assumption that allows to go far beyond Gaussianity (which was the typical case) and includes a very wide family of distributions. Note that even the uniform distribution is an extreme case of a unimodal distribution.

Although there exist powerful 1-d unimodality tests like the dip-test or the Silverman method [8], it is not straightforward to check unimodality for higher dimensions. As the dimensionality of the data increases, the tests require sufficient number of data points in order to be reliable. For this reason, the dip-dist criterion has been proposed [5] for determining unimodality in a set of datapoints using their pairwise distances (or similarities). The dip-dist criterion is based on the notion of viewer which is an arbitrary data object whose role is to suggest on the unimodality of the dataset by forming the set of its distances to all other data objects and applying the dip-test on this set of distances. The idea is that the distribution of the values in this distance vector could reveal information about the cluster structure. In presence of a homogneous cluster, the distribution of distances is expected to be unimodal. In the case where distinct subclusters exist, the distribution of distances should exhibit distinct modes, with each mode containing the distances to the data vectors of each cluster. Therefore, the result of a unimodality test could provide evidence on whether the dataset contains subclusters.

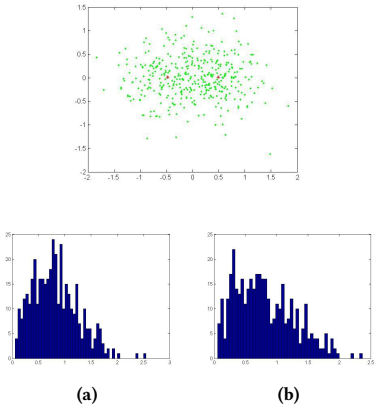
As mentioned in [5], there is a dependence of the results on the selected viewer. Intuitively, viewers at the boundaries of the set are expected to form distance vectors whose density modes are more distinct in case of more than one cluster. For this reason, in the dip-dist criterion all n data objects are considered as viewers, thus the dip-test is applied separately on each row of the distance matrix. If there exist viewers that reject unimodality (called split viewers), the dataset is characterized as multimodal.

In Fig. 2 and Fig. 3 we provide an illustrative example for datasets with two clusters and a single cluster respectively. We present the histograms of the set of distances observed by two viewers (marked with '+') which are clearly bimodal in Fig. 2 and unimodal in Fig. 3. Note that the dip values in Fig. 2 are higher than those in Fig. 3. Based on the results of the corresponding dip-tests we can correctly decide on the unimodality of each dataset.



(a) (b)

Figure 2: A 2-d dataset, two viewers (marked with '+') (top row) and the histograms of the distances of each of the two viewers from all data points (bottom row). Both histograms are found to be multimodal with dip values 0.054 and 0.043 and p-values equal to zero. Both viewers correctly characterize the dataset as multimodal.



(a) (b)

Figure 3: A 2-d dataset, two viewers (marked with '+') (top row) and the histograms of the distances of each of the two viewers from all data points (bottom row). Both histograms are found to be unimodal with dip values 0.012 and 0.014 and p-values 0.991 and 0.927. Both viewers correctly characterize the dataset as unimodal.

2.3 Dip-means algorithm

Dip-means is an incremental clustering algorithm that is based on cluster splitting and employs the dip-dist criterion to decide whether or not to split a cluster. Dip-means methodology takes as input the dataset X and, at each iteration, all current clusters are examined for unimodality using the dip-dist criterion. If a cluster is found multimodal under the dip-dist criterion, it is considered as a split candidate and obtains a non-zero score value.

At each dip-means iteration, among the clusters that are found multimodal according to the dip-dist criterion, only the candidate with maximum score is selected for splitting. This cluster is split into two clusters using a 2-means local search approach with multiple restarts from random initial centers, thus number of clusters increases by one. Then all clusters are refined using the k-means

algorithm. Starting with one cluster, the above procedure terminates when all clusters are found to be unimodal according to the dip-dist criterion.

It must be noted that in the dip-means implementation used in this work, the dip-dist criterion decides multimodality if at least one split viewer is found. Moreover, the score corresponding to a split candidate cluster is the maximum dip value among the split viewers, ie. the dip value of the 'most multimodal' viewer.

In the dip-means algorithm, the Euclidean distance among two data points is used to compute the distance matrix. Note, however, that in the dip-dist criterion only the pairwise distances (or similarities) between data points are used and not the vector representations themselves. This allows to apply dip-means even in kernel space, once a kernel matrix K with the pairwise similarities is provided. In this case the kernel k-means is used for cluster refinement instead of k-means.

3 PROJECTED DIP-MEANS

The projected dip-means algorithm (called pdip-means) constitutes an alternative to the dip-means algorithm that, instead of using the dip-dist criterion, it employs a different criterion (called projected dip) to decide on cluster homogeneity. The projected dip criterion also relies on the dip-test of unimodality, which is now applied on different 1-d datasets compared to dip-dist criterion.

3.1 The projected dip criterion

As previously mentioned the dip-dist criterion acts on the matrix with the pairwise distances among the data objects, applying the dip-test for unimodality on each row of this matrix. In this sense it is very general, and can be applied even in cases where the data objects are not available and only the distance matrix is given.

In the typical case where the objects of a dataset correspond to data vectors, it is possible to test the homogeneity of this set by employing dip-test for unimodality [3] in a different and more straightforward way. More specifically, we assume that a set of data vectors is homogeneous if its 1-d projections are unimodal (according to the dip-test). This is in analogy with the criterion employed in the g-means algorithm [2] where the set of projections on the first principal axis is tested for gaussianity. In this work, we extend this idea by considering 1-d projections on several axes and use the more general dip-test for unimodality, instead of a gaussianity test.

We refer to the proposed criterion as projected dip (pdip) criterion that can be applied to decide on the homogeneity of a set of data vectors. To use this criterion, first a set of M 1-d projections is specified. Then for each projection j ($j = 1, \ldots, M$), the corresponding projected value for each data vector is computed, thus the 1-d set P_j is formed, and the dip test is applied on set P_j to obtain the values dip_j and $p - value_j$. In analogy with the dip-dist criterion, each projection can be considered as a 'point-of-view'. Therefore if multimodality is observed for several 'point-of-views' (at least one in our experiments) then the dataset is considered multimodal, otherwise it is considered unimodal.

Three types of 1-d data projections could have been considered:

- Projections on each of the d the original axes, ie. we apply the dip test on each column of the dataset.
- PCA projections, where Principal Component Analysis is applied to the dataset to extract projections on each principal axis. In our experiments, the dip-test is applied to each of the d principal projections.
- Projections on randomly selected axes (random projections). A sufficient number of random projections could reveal the existence of Gaussian clusters with error probability 1% for 12 projections and 0.1% for 18 projections. In the current work we did not use random projections to avoid the randomness that would have been introduced in the computation of the pdip criterion.

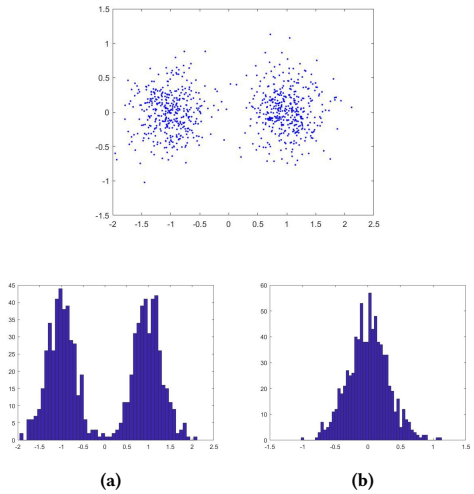
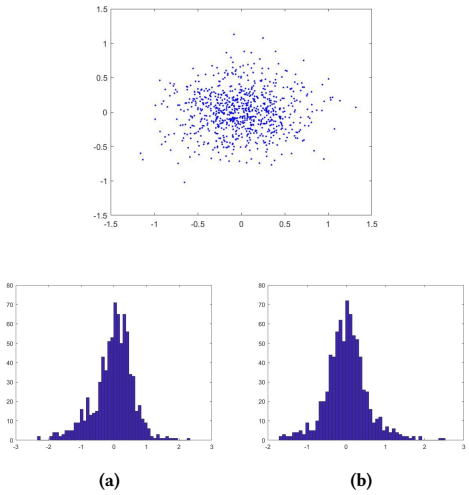


Figure 4: A 2-d dataset (top row) and the histograms of data projections on each of the two original axes (bottom row). The histogram on the left is found to be multimodal, with dip=0.1 and with p-value=0. The histogram on the right is found to be unimodal with dip=0.01 and p-value=0.94 Since at least one multimodal projection is found, the pdip criterion characterizes this dataset as multimodal.

In summary, to decide on content homogeneity of a set of real valued data vectors using the pdip criterion, we apply the dip-test $2d$ times: for each of the d columns of the data matrix and for each the d PCA projections. If at least one dip-test indicates multimodality, then the dataset is considered multimodal, otherwise it is considered unimodal. In the case of multimodality, the largest among the dip values of the projections is considered as the multimodality score of the dataset.

In Fig. 4 and Fig. 5 we provide an illustrative example using artificial 2-d datasets with two clusters and a single cluster respectively. For each dataset we present the two histograms corresponding to the data projections on the two main axes. In Fig. 4 it is clear that one histogram is bimodal and one unimodal and this also in accordance with the results of the dip-test. In Fig. 5 both histograms are multimodal and the dip-test also correctly decides on this issue. Based on the results of the dip-tests, the pdip criterion can correctly decide on the unimodality of each dataset.



(a) (b)

Figure 5: A 2-d dataset (top row) and the histograms of data projections on each of the two original axes (bottom row). Both histograms are found to be unimodal with dip values 0.098, 0.089 and p-values 0.92 and 0.91. Both projections correctly characterize the dataset as unimodal.

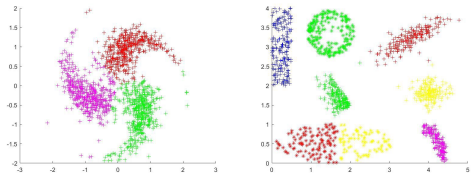
3.2 The projected dip-means algorithm

Given a set of real valued data vectors, the projected dip-means (pdip-means) can be obtained from the original dip-means algorithm by replacing the dip-dist criterion with the projected dip (pdip) criterion. Therefore, pdip-means is an incremental clustering algorithm that starts with a single cluster and iteratively adds clusters to the solution through cluster splitting based on the pdip criterion.

More specifically, the pdip criterion is applied on every cluster of the current solution with say k clusters, and each cluster is characterized as either multimodal or unimodal. In the case where a cluster is found multimodal, the maximum dip value (maxdip) computed in the dip-tests for this cluster is also retained. If one or more multimodal clusters exist in the currest solution, then the cluster with the highest multimodality score is selected and splitted into two clusters. Thus, the number of clusters increases to k+1 and the k-means with k+1 clusters is applied to further refine the k+1 clusters and provide the solution with k+1 clusters. The above steps are repeated until all clusters in the current solution are found to be unimodal, thus the algorithm terminates since no further cluster splitting is suggested.

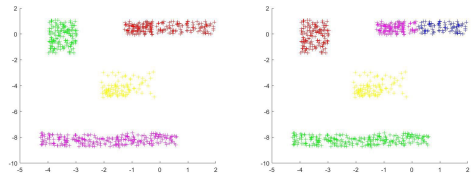
4 EXPERIMENTAL RESULTS

In our experimental evaluation we compare the proposed pdip-means method with the original dip-means [5]. The two compared methods were executed starting with a single cluster and, at each iteration, the most multimodal cluster is selected and splitted in two clusters. To implement splitting of a cluster in two subclusters, 10 trials are performed on the cluster data using a randomly initialized 2-means algorithm and the split with lower clustering error is kept.



(a) (b)

Figure 6: The same clustering solution provided by both pdip-means and dip-means on two artificial datasets.



(a) (b)

Figure 7: The clustering solution provided by (a) pdip-means and (b) dip-means on an artificial dataset four with uniform clusters of rectangular shape.

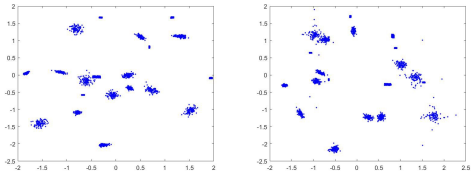
Therefore, the major difference between the two methods concerns the criterion that decides on cluster homogeneity. The parameters of the dip-dist criterion are set as a=0 for significance level of dip test and b=1000 for the number of bootstraps, the same as in the case of pdip. In all examined datasets ground truth cluster labels are used. In artificial datasets they are specified by the data generation mechanism, while in real datasets we assume that cluster labels coincide with the available class labels. In order to compare the ground truth labeling and the grouping produced by the clustering methods, we utilize the Variation of Information (VI) metric [6] and Rand Index (RI) [4]. Note that lower values of VI and higher for RI indicate better solutions.

It should be emphasized that in all clustering experiments the number of clusters was not given as input, instead it was automatically determined by the clustering algorithms.

4.1 Synthetic datasets

We first provide clustering results for synthetic 2-d datasets. In Fig. 6 we provide the solution obtained (a) on the three wings dataset and (b) a difficult dataset with seven clusters of various shape and density. Both dip-means and pdip-means provided the same solution (shown in the figures). In the first dataset (Fig. 6a) the number of clusters is correctly estimated, while in the second dataset (Fig. 6b) eight clusters were found, while the correct number is seven (the bottom left cluster has been split in two).

In Fig. 7 the clustering solutions for a dataset with four clusters of orthogonal shape and uniform density are presented. In this case pdip means (Fig. 7a) provides the correct solution and performs better than dip-means (Fig. 7b) which splits the top right cluster in two subclusters.



(a) (b)

Figure 8: Representative 2-d datasets: (a) case 1 dataset (with Gaussian and uniform clusters), (b) case 2 dataset (with Gaussian, uniform and Student-t clusters). Note that in case 2, the Student-t distribution introduces noise and outliers.

Table 1: Results with synthetic datasets (20 clusters) (case 1)

Dataset Case 2	k pdip means	RI pdip means	VI pdip means	k dip means	RI dip means	VI dip means
d=2	20	0.999	0.02	19.8	0.998	0.03
d=4	20	1	0	20	1	0
d=8	20	1	0	20	1	0

Table 2: Results with synthetic datasets (20 clusters) (case 2)

Dataset Case 2	k pdip means	RI pdip means	VI pdip means	k dip means	RI dip means	VI dip means
d=2	19.4	0.982	0.095	19.1	0.965	0.152
d=4	19.6	0.987	0.088	19.4	0.976	0.111
d=8	19.8	0.991	0.082	19.7	0.990	0.081

Based on the experimental protocol suggested in [5], we created synthetic datasets with true number k=20 clusters, with 100 datapoints in each cluster (thus n=2000), in d=2, 4, 8 dimensions and with low separation degree. Two cases were considered: 1) datasets with 50% Gaussian clusters and 50% Uniform clusters and 2) datasets with 40% Gaussian clusters, 40% Uniform clusters and 20% Student-t clusters. Note that the Student-t clusters introduce outliers to the dataset, thus the datasets of case 2 are more difficult to cluster correctly. Fig. 8 provides a representative 2-d dataset for case 1 (Fig. 8a) and for case 2 (Fig. 8b) respectively. Note the existence of noise and outliers in the case 2 dataset.

For each case and value of d, we generated 20 datasets that were clustered with pdip-means and dip-means respectively. Average performance results concerning the estimated k, RI and VI are presented in Table 1 and Table 2. As the results indicate, both p-dip and dip-means provide excellent clustering performance in all cases estimating almost perfectly the true number of clusters, even in the case of noisy datasets. Based on results, pdip-means seems to be slightly superior to dip-means, however, this difference cannot be considered as significant.

Table 3: Results on real datasets

Dataset	k pdip means	RI pdip means	VI pdip means	k dip means	RI dip means	VI dip means
Iris (k=3)	3	0.87	0.56	2	0.71	0.6
PD:047 (k=3)	5	0.82	0.97	5	0.84	0.95
PD:02468 (k=5)	5	0.57	0.98	6	0.88	0.78
PD:13579 (k=5)	7	0.73	1.66	6	0.82	1.41
HD:047 (k=3)	3	0.85	0.49	1	0.5	0.69
HD:02468 (k=5)	6	0.79	1.51	2	0.56	1.82
HD:13579 (k=5)	7	0.85	1.56	1	0.49	1.61

We also conducted clustering experiments using several real world datasets [1], where the provided class labels were considered as ground truth. The clustering performance results are presented in Table 3.

The first dataset is the well-known Iris dataset containing 150 4-dimensional examples belonging to 3 classes. It can be observed that the proposed pdip-means outperforms dip-means and correctly estimates the number of clusters. We also conducted experiments with several subsets of i) the Pendigits dataset (PD) that contains 16-dimensional vectors, each one representing a digit from 0-9 as written by a human subject and ii) the USPS Handwritten digits dataset (HD) that contains 64-dimensional vectors corresponding to 8x8 images of the digits 0-9. The datasets are provided in the form of training and test sets. We clustered three subsets of the test sets containing the data vectors of the digits {0,4,7} (k=3), {0,2,4,6,8} (k=5) and {1,3,5,7,9} (k=5). We do not apply any preprocessing on the data vectors.

From the results in Table 3, it can be observed that for the PD dataset, both methods provide reasonable estimates of the number of clusters, with dip-means being more accurate than the proposed pdip-means. For the HD dataset, the dip-means encounters difficulty in deciding to split clusters thus, it underestimates the cluster number, while pdip means does not seem to have such problem providing solutions of much better quality. A general empirical remark is that for some real datasets, the dip-means algorithm sometimes does not manage to split the initial clusters. This problem could be alleviated by running dip-means with larger initial k (e.g. k=3). In the pdip-means case such a problematic behavior has not been observed.

4.2 Computational Cost

In what concerns the comparison of the computational complexity of the two methods, assume a cluster with n data vectors of dimensionality d. Both the pdist and dip-dist use the dip-test on 1-d dimensional sets of size n, thus the dip-test complexity is O(n). The computation of pdist requires $2d$ applications of dip-test, while the computation of dip-dist requires n applications of dip-test. Usually $d \ll n$, thus in typical cases the pdist can be considered much faster. However, it must be stressed that application of pdip has the additional cost of computing the PCA projections that is in the general case O(n^3). In practice, and for the datasets examined in this paper where $d \ll n$, pdip-means was always faster than dip-means

if both provided solutions with similar number of clusters. More specifically, we conducted experiments using the case 1 (section 4.1) synthetic datasets with 20 clusters, having N=2000, 4000 and 8000 data vectors and with d=2, 4 and 8 dimensions. All the combinations of N and d were examined. We empirically found that execution of pdip-means was 1.5 to 4 times faster than dip-means.

5 CONCLUSIONS

Estimating the number of clusters is a difficult and important issue in data clustering. We have presented the pdip-means algorithm, a clustering approach that automatically determines the number of clusters. The method can be applied in the typical case where the dataset is given as a set of data vectors. It constitutes an adaptation of the dip-means algorithm [5] employing a different criterion (pdip criterion) for deciding on the homogeneity of a given cluster. More specifically, in the pdip criterion the unimodality tests are applied on 1-d projections of the data vectors, while the dip-dist criterion used in the dip-means algorithm applies unimodality tests on the rows of the distance matrix.

Experiments on difficult synthetic datasets with many non Gaussian clusters indicate that both methods provide solutions of very good quality, and estimate accurately the correct number of clusters. In the case where noise and outliers are added to the datasets, pdip-means seems to be slightly superior to dip-means. In the case of real datasets, pdip-means seems to be more effective than dip-means, in the cases where the latter has difficulty in deciding to split the initial clusters.

One direction of future work aims at assessing the performance of the pdip-means algorithm in real world clustering applications (e.g. face clustering). We also plan to experimentally test whether the applications of dip-test on random projections (that is not considered in this work) would improve the performance of pdip. Finally, it is worthwhile to examine the combination of the pdip and dip-dist criteria into a composite decision criterion for cluster unimodality. Such a criterion would take into account the unimodality observed both in various data projections (pdip) and on the rows of the distance matrix (dip-dist).

REFERENCES

[1] A. Asuncion and D. Newman. [n. d.]. UCI Machine Learning Repository. University of California at Irvine, Irvine, CA. ([n. d.]). http://www.ics.uci.edu/mlearn/MLRepository.html

[2] G. Hamerly and C. Elkan. 2003. Learning the k in k-means. In *Advances in Neural Information Processing Systems (NIPS '03)*. 281–288.

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

[4] L. Hubert and P. Arabie. 1985. Comparing partitions. *Journal of Classification* 2, 1 (1985), 193–218.

[5] A. Kalogeratos and A. Likas. 2012. Dip-means: an incremental clustering method for estimating the number of clusters. In *Advances in Neural Information Processing Systems (NIPS '12)*. 2393–2401.

[6] M. Meila. 2007. Comparing clusterings Ű an information based distance. *Multivariate Analysis* 98, 5 (2007), 873–895.

[7] D. Pelleg and A. Moore. 2000. X-means: extending k-means with efficient estimation of the number of clusters. In *Interantional Conference on Machine Learning (ICML '00)*. 727–734.

[8] B. Silverman. 1981. Using Kernel density estimates to investigate multimodality. *Journal of Royal Statistic Society B* 43, 1 (1981), 97–99.