Detecting Number of Clusters by Testing Block Diagonal Behavior of Similarity Matrix

Sharmodeep Bhattacharyya and Peter J. Bickel

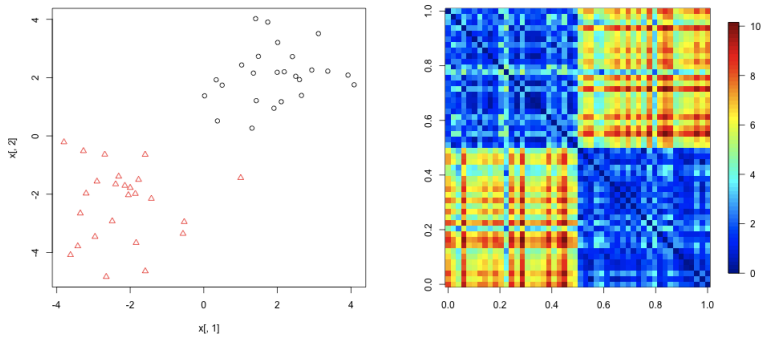
Abstract

In cluster analysis, one of the main challenges is the detection of number of clusters. Most clustering algorithms need the number of clusters to be specified beforehand. Previously, there has been some work related to choosing the number of clusters. We propose a new method of selecting number of clusters, based on hypothesis testing. One way to look at clustering is - getting hold of the most block-diagonal form of the similarity matrix. So, we test the hypothesis, whether the resulting similarity matrix after clustering is block-diagonal or not. The number of clusters for which we have the most block diagonal similarity matrix is considered to be the most suitable number of clusters for the data set. So, the method can be applied for any optimal partitioning algorithm (like k-means or spectral clustering). We show that this method works well compared to currently used methods for both simulated and real data sets.

1 Introduction

Cluster analysis is an important unsupervised classification technique. In clustering, a set of unlabeled patterns, usually vectors in a multidimensional space, are grouped into clusters in such a way that patterns in same cluster are similar in some sense and patterns in different clusters are dissimilar in the same sense. One of the main approaches of clustering is optimal partitioning algorithms. The first step of optimal partitioning algorithm is choosing the number of groups or clusters.

The method we have proposed for clustering depends upon exploiting the structure of the similarity (or distance) matrix after clustering. One way of viewing clustering is getting hold of the most block-diagonal form of the similarity matrix, by simultaneously permuting the rows and columns of the similarity matrix. In figure 1(a) we have a data set, whose distance matrix is given in figure 1(b) after permuting the row and columns according to the assignments obtained from output of k-means algorithm with 2 clusters on the data set. We see that the matrix in figure 1(b) is block-diagonal in nature. So, our method is based on the assumption that if the partitioning method is applied with correct number of clusters, then the resulting similarity (or distance) matrix will have a better block-diagonal structure.



(a) (b)

Figure 1: (a) Sample data set (b)Distance matrix after 2-means clustering

Now, we test the 'block-diagonal-ness' of a matrix by hypothesis test of location shift. We test if there is a location shift between the distances in a diagonal block with the distances in an off-diagonal block. If there is evidence of location shift, that means that cluster is well-separated from other clusters. So, it is also a cluster validation technique, which determines, whether the current cluster under consideration is actually a well-separated cluster from the remaining clusters. If there is evidence of location shift for all blocks/clusters, then, that means that number of blocks/clusters is one possible choice for number of clusters. So, we have several possible choices for the number of clusters and this is expected if we consider Hartigan's (1985) [6] definition of high-density clusters, where, depending on the level, the number of disjoint components of the level set of the density (that means, number of clusters) vary. However, if we have to specify one number as the number of clusters, we shall prefer the one with most deviation from the null distribution. Also, note that our method works for selecting number of clusters for any clustering/partitioning algorithms.

Several Methods have been proposed for choosing the number of clusters in the literature. Milligan and Cooper (1985) [8] performed a simulation study comparing different statistical heuristics for choosing number of clusters, among which the best were by Calinski and Harabasz (1974) [3]:

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)} \tag{1}$$

where, $B(k)$ and $W(k)$ are the between and within cluster sums of squares with k clusters. Rousseeuw (1987) citerousseeuw1987sihouettes proposed the cluster silhouette coefficient

$$SC(k) = \frac{1}{n}\sum_{i=1}^{n} s(i); \quad s(i) = \frac{b(i) - a(i)}{\max a(i), b(i)} \tag{2}$$

where, For observation i, let $a(i)$ be the average distance to other points in its cluster, and $b(i)$ the average distance to points in the nearest cluster besides its own and nearest is defined by the cluster minimizing this average distance. The $\hat{k}$ for which $SC(k)$ is maximized is considered the best number of clusters by this method. Tibshirani et. al. (2001) [12] proposed gap statistic for estimating the number of clusters. Another popular tool for selecting number of clusters is using cluster stability, as proposed by Ben-hur et. al. (2001) [1] and Lange et. al. (2004) [7]. There are also methods for selecting number of clusters through BIC in model-based clustering as proposed in Fraley and Raftery (1998) [5]. There are many other methods of finding number of clusters in the literature, but for brevity, we are not mentioning them here.

The paper is arranged as follows. In section 2, we have introduced our method. In section 3, we have carried out simulation study, showing the efficacy of our method compared to other methods. In section 4, we have applied our method to two real-life data sets - an astronomical data set and a microarray data on leukemia study. In section 5, we have provided the discussion of the results and the method.

2 Our Method

Let us consider, for data $\mathbf{X} = (X_1, \ldots, X_n)$, where, $X_i \in \mathbb{R}^p$, we start with a distance matrix $D = ((d_{ij}))_{i,j=1}^n$, where, d_{ij} = distance between the observations X_i and X_j. We also have a clustering/partitioning method, which partitions the data into clusters, after the number of clusters have been specified. Let us consider, that for number of clusters, k, the partitioning method partitions the data into clusters $(C_1, \ldots, C_k)$, where, $C_j \subset \mathbf{X}$ for $j = 1, \ldots, k$ and $\cup_j C_j = \mathbf{X}$, $C_i \cap C_j = \phi$, for all $i \neq j$.

Now, if we consider the distance matrix with the row-column entries of the matrix being ordered according to the clusters, that is, consider the permutation of the data entries according to the clusters, $\mathbf{X}_\pi = (X_{\pi(1)}, \ldots, X_{\pi(n)} = (C_1, \ldots, C_k)$. We form the distance matrix $D^\pi = ((d_{ij}^\pi))$ from $\mathbf{X}_\pi$ by d_{ij}^π = the distance between $X_{\pi(1)}$ and $X_{\pi(j)}$.

One of the necessary conditions for the matrix D^π is - it should be 'block-diagonal'. That means the entries in the diagonal blocks of the matrix D^π should have a lower values than the value of the entries in the off-diagonal blocks. We can denote $D^\pi = ((D_{ii'}^\pi))$ as the $k \times k$ block matrix, where, $D_{ii'}^\pi$ is $|C_i| \times |C_{i'}|$ matrix containing the distances between the observations in clusters C_i and $C_{i'}$, where, $i, i' = 1, \ldots, n$. So, in ideal case, the entries in D_{ii}^π should have lower values than entries in D_{ij}^π, where, $i \neq j$. We judge whether a clustering is valid by testing this statement. Also, this is a cluster-wise validation, as for each cluster (or block), we are testing whether the corresponding diagonal block in D^π has smaller values than the corresponding off-diagonal blocks of D^π.

2.1 Block Diagonal Hypothesis Testing

We want to test the hypothesis that D^{π} has block diagonal structure. We proceed as follows - for each block C_i $(i = 1, \ldots, k)$, consider the $\frac{|C_i|(|C_i-1)}{2}$ upper diagonal entries of D^{π}_{ii} in the vectorized form $(Y_1, \ldots, Y_m)$, where, $m = \frac{|C_i|(|C_i-1)}{2}$. Now, for each $i' = (i+1, \ldots, n)$, consider the $|C_i||C_{i'}|$ entries of $D^{\pi}_{ii'}$ in a vectorized form $(Z_1, \ldots, Z_{m'})$, where, $m' = |C_i||C_{i'}|$.

So, we have two data sets $\mathbf{Y} = (Y_1, \ldots, Y_m)$ and $\mathbf{Z} = (Z_1, \ldots, Z_{m'})$ and we want to find whether $\mathbf{Y}$ has smaller values than $\mathbf{Z}$ in average. So, we perform a one-sided location shift test between $\mathbf{Y}$ and $\mathbf{Z}$, with the null hypothesis of no location shift. Now, note that, this is a non-standard location shift test, since, the data $\mathbf{Y}$ and $\mathbf{Z}$ are not independent. So, the null distribution of the standard location shift tests (t-test, Wilcoxon Rank-Sum test etc.) does not hold in this case. We have to go for a different route to get hold of a null distribution for the test statistic we use.

2.2 Using Permutation Test

Getting hold of a null or reference distribution for testing cluster structure is always a challenge, as indicated in Tibshirani et. al. [12]. In this case, let us consider $T_{ii'}$ is the test statistic we are using to test for a location shift between $\mathbf{Y}$ and $\mathbf{Z}$. Now, among among all $i' = i+1, \ldots, n$, consider T_i the test statistic $T_{ii'}$ that is least favorable towards the alternative, for example, for t-test statistic, $T_i = \max_{i'} T_{ii'}$. Now, the null distribution of this statistic is difficult to find theoretically. So, we perform permutation test instead. We permute the row-column entries of the distance matrix D^{π} to generate a new distance matrix $D^{\pi'}$ and generate the corresponding test-statistic T_i for the matrix $D^{\pi'}$. We, repeat this procedure, to get a null-distribution of the test-statistic T_i. Using the null distribution, we test for the location shift for i^{th} cluster.

Now, there is a problem with this approach. The objective of any clustering algorithm is to find the permutation of the row-column entries of the distance matrix D, such that, the permuted matrix D^{π} has most block-diagonal structure. So, through permutation, we are not actually finding the null distribution of T_i for each cluster i. So, we use the p-values generated from the permutation test to get a coefficient called permutation coefficient as follows - for fixed level α (usually 0.01) we find the α^{th} quantile, Q_i of the 'null' distribution of T_i generated through permutations. Now, we define, for each cluster, the *excess value* (ev_i) as

$$ev_i = Q_i - T_i \tag{3}$$

Then, we define the block-diagonal coefficient for the number of clusters k as

$$BDQ(k) = \min_i ev_i \tag{4}$$

The estimated number of clusters $\hat{k}$ is defined as

$$\hat{k} = \max_k BDQ(k) \tag{5}$$

Now, $\hat{k}$ is the most prominent number of clusters. But, there might be other possible number of clusters for which the partition of the data set makes sense. So, we also output a list of potential number of clusters. That is done as follows - let us consider i^* achieves the minimizer in (4). If the value of T_{i^*} is less than $Q_{i^*} - 1.5IQR$, where, IQR is the inter-quartile range of the distribution of permuted T_{i^*}, we call, the corresponding k as a potential cluster number.

Also, if the set of potential clusters is a null set. Then, it implies the lack of cluster structure, which can mean either that there is only one cluster in the data or the clustering algorithm is failing to properly cluster the data. So, by our method, we can also detect one cluster, which many of methods for detecting cluster number (like silhouette, C-H) fail to find.

2.3 An Example

We apply our method on a very well-known data set - Fisher's Iris data set [4]. The data set contains 4 measurements for a sample of 150 flowers. There are 3 types of flowers in the data set. The scatter plot based on first 2 types of measurements in given in figure 2.

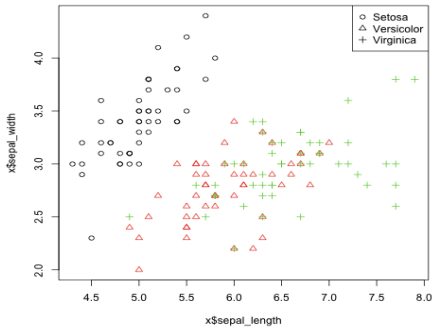
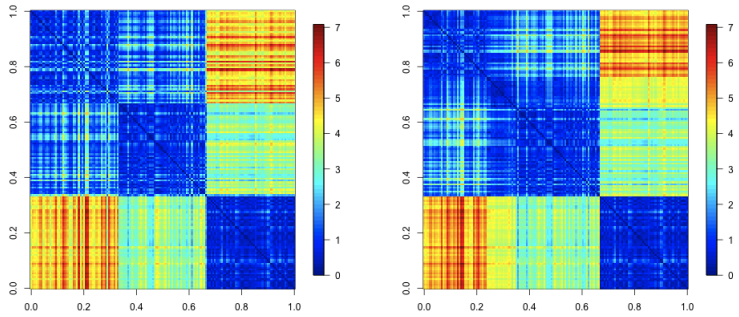


Figure 2: Iris Data with 2 dimensions sepal length and width.

We use partitioning around medoid (PAM) as the clustering method. We use t-test statistic as the statistic for hypothesis testing here. Then, if we use, our method to choose the number of clusters, then the value of $BDQ(k)$ is maximized for $k = 2$. Actually, the only positive values of BDQ come to be $BDQ(2) = 96.5$ and $BDQ(3) = 7.03$. So, we see that by our analysis, $\hat{k} = 2$. However, if we try to find the set of potential number of clusters, then, $k = 3$ also becomes a potential cluster number, as $Q_{i^*} = -6.27$, $T_{i^*} = -13.3$ and $IQR = 2.6$ for $k = 3$. So, we see that we can identify clusters for different hierarchies according to our method and we know, where to stop. The distance matrices for $k = 2$ and $k = 3$ also gives the proper intuitions.



(a) (b)

Figure 3: (a) Distance matrix after 2-means clustering (b) Distance matrix after 3-means clustering

3 Simulation Study

For simulation study, we generated data for four different scenarios -

(a) *No Cluster*: We have generated data uniformly over a unit square in 10 dimensions.

(b) *2, 3, 4, 5 Random Clusters in 7, 8, 9 and 10 dimensions respectively*: We generated clusters centers randomly from $N(0, 2I)$ distribution, such that, two clusters centers are at least one unit apart. Then, we generated clusters of size 25 or 50 (randomly chosen) from normal distributions with mean as cluster centers and identity covariance matrix.

(c) *2 elongated clusters in three dimensions* We generated each cluster as follows: For cluster 1, set $x_1 = x_2 = x_3 = t$ with t taking on 100 equally spaced values from -.5 to .5 and then Gaussian noise with standard deviation .1 is added to each feature. Cluster 2 is generated in the same way, except that the value 2 is then added to each feature. The result is two elongated clusters, stretching out along the main diagonal of a three-dimensional cube.

(d) *2 close and elongated clusters in three dimensions* As in the previous scenario, with cluster 2 being generated in the same way as cluster 1, except that the value 1 is then added to the first feature only.

The scenarios are motivated from Tibshirani and Walther (2005) [11].

We have repeated each experiment 50 times. For scenario (a), we compare our method with gap statistic. For scenario(b)-(d), we compare our method with CH, silhouette and stability criterion. The stability method has been adopted

from Brock et. al. (2008) [2]. We use PAM as the clustering method and t-test statistic as the statistic for location shift testing. We represent our methods by BDQ and BDQ.potential. The BDQ.potential lists the number of times a cluster number becomes a potential candidate for the data set. So, the sum of the elements in rows of BDQ.potential will not be 50, as each data set can have more than one potential clusters. The results are provided in table 1 - 4.

Table 1: Number of Clusters for Scenario (a)

	$k = 1$	$k = 2$	$k = 3$
gap	38	11	1
BDQ	50	0	0

We can see that for scenario (a) BDQ performs better. For scenario (b), for number of clusters 3 and 4, BDQ.potential performs best. For the comparatively hard scenarios (c) and (d), BDQ and BDQ.potential performs quite well. Especially, BDQ.potential almost always include the correct number of clusters within its potential choices.

4 Study on Two Real Data Sets

We apply our method to two real data sets. We compare our method in these cases with CH, silhouette and stability criterion.

4.1 Leukemia Data

The Leukemia data is obtained from Monti et. al. (2003) [9]. The data is composed by instances representing diagnosed samples of bone marrow from pediatric acute leukemia patients, corresponding to six prognostically important leukemia subtypes - 43 T-lineage ALL; 27 E2A-PBX1; 15 BCR-ABL; 79 TEL-AML1 and 20 MLL rearrangements; and 64 hyperdiploid > 50? chromosomes. There are 248 total patients and for each patient the number of attributes is 985.

We use hierarchical clustering method as the clustering algorithm for our method in this case. The performance of our method on this data set compared to other methods is given in table 5 -

We see here that BDQ identifies the correct number of clusters. Also, we see that when we consider the BDQ.potential method, it gives the most information about the clustering picture of the data set, since if we see the cluster membership, after the clustering, one of the classes is spuriously broken and two classes remain merged to form the 6 clusters for the hierarchical clustering method considered. So, when, we have seven clusters, we are actually having all the 6 classes plus a broken part of one class. So, when, BDQ.potential says that the data potentially also has 7 clusters, it gives insight into the data.

Table 2: Number of Clusters for Scenario (b)

$\mathbf{k} = 2$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$
Silhouette	0	48	0	0	0	0	0	2
CH	0	48	1	1	0	0	0	0
Stability	0	47	2	0	0	0	0	1
BDQ	11	39	0	0	0	0	0	0
BDQ.potential	11	39	0	0	0	0	0	0
$\mathbf{k} = 3$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$
Silhouette	0	25	25	0	0	0	0	0
CH	0	27	23	0	0	0	0	0
Stability	0	27	23	0	0	0	0	0
BDQ	4	30	16	0	0	0	0	0
BDQ.potential	4	40	33	0	0	0	0	0
$\mathbf{k} = 4$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$
Silhouette	0	11	16	23	0	0	0	0
CH	0	12	18	20	0	0	0	0
Stability	0	17	11	22	0	0	0	0
BDQ	4	24	7	15	0	0	0	0
BDQ.potential	4	35	30	28	0	0	0	0
$\mathbf{k} = 5$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$
Silhouette	0	5	7	13	25	0	0	0
CH	0	10	10	18	12	0	0	0
Stability	0	10	7	10	23	0	0	0
BDQ	5	26	9	3	7	0	0	0
BDQ.potential	5	31	18	12	17	0	0	0

Table 3: Number of Clusters for Scenario (c)

	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$
Silhouette	0	50	0	0	0	0	0	0
CH	0	0	0	7	0	32	3	8
Stability	0	50	0	0	0	0	0	0
BDQ	0	50	0	0	0	0	0	0

Table 4: Number of Clusters for Scenario (d)

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$
Silhouette	0	4	0	10	2	30	3	1	0	0
CH	0	0	0	0	0	10	2	26	7	5
Stability	0	44	1	5	0	0	0	0	0	0
BDQ	0	11	1	34	4	0	0	0	0	0
BDQ.potential	0	50	14	48	43	0	0	0	0	0

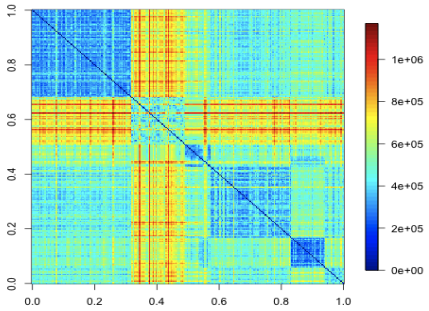


Figure 4: The distance matrix of Leukemia data with the classes arranged TEL-AML1, T-Lineage ALL, MLL, hyperdiploid, E2A-PBX1, BCR-ABL.

4.2 Astronomy Data

The astronomy data is obtained from Richards et. al. (2011) [10]. The data is composed by instances representing light sources from sky surveys. The light sources are composed of 5 types of stars - 191 Classical Capheid, 145 Beta Lyrae, 114 Delta Scuti, 144 Mira, 58 W Ursae Majoris. There are 652 total light sources and for each light source the number of features is 64.

The performance of our method on this data set compared to other methods is given in table 6 -

So, BDQ.potential also performs good in this case and provides a nice insight to the data. Though number of clusters selected by BDQ is 4 in this case, we see that, $k = 5$ is one one of the potential cluster numbers.

5 Discussion

So, we can see that methods of selecting cluster number by testing for block-diagonality of a matrix works nicely in practice. This method is highly general

Table 5: Number of Clusters for Leukemia Data

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
Silhouette	0	0	0	0	0	1	0	0
CH	0	1	0	0	0	0	0	0
BDQ	0	0	0	0	0	1	0	0
BDQ.potential	0	0	0	0	0	1	1	0

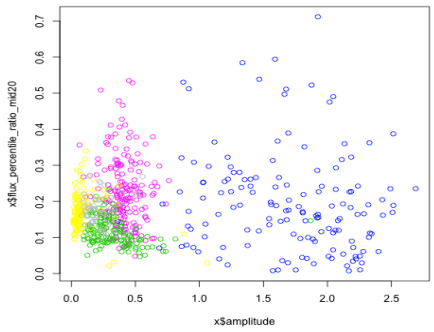


Figure 5: The astronomy data in two of its features.

and can be applied in conjunction with any clustering method and any similarity (or distance) matrix. Also, the method can also provide a list of potential number of clusters, which is quite suggestive, since, a data set usually can be considered to have different number of clusters depending on the level of inspection, we are going to perform on the data set. Also, note that this a completely non-parametric approach, so it can be applied to quite general class of models

However, note that this method of selecting number of clusters is dependent on the performance of the clustering method itself. If the clustering method does not perform well, then, this method might produce unstable results.

Another issue is selecting the number of permutations. We have generally considered 1000 permutations to construct the 'null' distribution. However, it might be better to first sequentially test for the p-value 0.01, to see how many permutations are needed for the sequential rule to stop. Then, we can use a number of permutations slightly greater than the stopping number, to form the null distribution.

Lastly, we have not derived any theoretical results for this method. However, assuming some underlying model space, we can try to prove the consistency and variance bounds of this method. Considering, gaussian model, we can

Table 6: Number of Clusters for Astronomy Data

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
Silhouette	0	0	0	1	0	0	0	0
CH	0	0	1	0	0	0	0	0
BDQ	0	0	0	1	0	0	0	0
BDQ.potential	0	0	0	1	1	0	0	0

easily see that, our procedure of finding location shift is a correct one. However, theoretically deriving the 'null' distribution is a challenge and we wish to address this issue later on.

Acknowledgments

We would like to thank Center for Time-domain Informatics and Prof. Joshua Bloom for providing the astronomy data set.

References

[1] A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing 2002: Kauai, Hawaii, 3-7 January 2002*, page 6. World Scientific Pub Co Inc, 2001.

[2] G. Brock, V. Pihur, S. Datta, and S. Datta. clvalid: An r package for cluster validation. *Journal of Statistical Software*, 25(4):1–22, 2008.

[3] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-Theory and Methods*, 3(1):1–27, 1974.

[4] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, 7(2):179–188, 1936.

[5] C. Fraley and A.E. Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal*, 41(8):578, 1998.

[6] J.A. Hartigan. Statistical theory in clustering. *Journal of classification*, 2(1):63–76, 1985.

[7] T. Lange, V. Roth, M.L. Braun, and J.M. Buhmann. Stability-based validation of clustering solutions. *Neural computation*, 16(6):1299–1323, 2004.

[8] G.W. Milligan and M.C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.

[9] S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1):91–118, 2003.

[10] J.W. Richards, D.L. Starr, N.R. Butler, J.S. Bloom, J.M. Brewer, A. Crellin-Quick, J. Higgins, R. Kennedy, and M. Rischard. On machine-learned classification of variable stars with sparse and noisy time-series data. *The Astrophysical Journal*, 733:10, 2011.

[11] R. Tibshirani and G. Walther. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528, 2005.

[12] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.