Model Selection and Stability in Spectral Clustering

Zeev Volkovich and Renata Avros

Ort Braude College of Engineering, Software Engineering Department, Karmiel, Israel

Keywords: Spectral Clustering, Model Selection.

Abstract: An open problem in spectral clustering concerning of finding automatically the number of clusters is studied. We generalize the method for the scale parameter selecting offered in the Ng-Jordan-Weiss (NJW) algorithm and reveal a connection with the distance learning methodology. Values of the scaling parameter estimated via clustering of samples drawn are considered as a cluster stability attitude such that the clusters quantity corresponding to the most concentrated distribution is accepted as true number of clusters. Numerical experiments provided demonstrate high potential ability of the offered method.

1 INTRODUCTION

The recent decades, have seen numerous applications of graph eigenvalues in many areas of combinatorial optimization (Chung, 1997), (Mohar, 1997), (Spielman, 2012). Spectral clustering methods became very popular in the 21st century following Shi and Malik (Shi and Malik, 2000) and Ng, Jordan and Weiss (Ng et al., 2001). Over the last decade, various spectral clustering algorithms have been developed and applied to computer vision (Ng et al., 2001), (Shi and Malik, 2000), (Yu and Shi, 2003), network science (Fortunato, 2010), (White and Smyth, 2005), biometrics (Wechsler, 2010), text mining (Liu et al., 2009), natural language processing (Dasgupta and Ng, 2009) and other areas. We note that spectral clustering methods have been found equivalent to kernel k-means (Dhillon et al., 2004), (Kulis et al., 2005) as well as to nonnegative matrix factorization (Ding et al., 2005). For surveys on spectral clustering, see (Nascimento and Carvalho, 2011), (Luxburg, 2007), (Filippone et al., 2008).

The main idea is to use eigenvectors of the Laplacian matrix, based on an affinity (similarity) function over the data. The Laplacian is a positive semidefinite matrix whose eigenvalues are nonnegative reals. It is well-known that the smallest eigenvalue of the Laplacian is 0, and it corresponds to an eigenvector with all entries equal. Moreover, viewing the data similarity function as an adjacency matrix of a graph, the multiplicity of the 0 eigenvalue is the number of connected components (Mohar, 1997). While in clustering problems the corresponding graph is typically connected, we partition the data into $k$ clusters using the $k$ eigenvectors corresponding to the $k$ smallest eigenvalues. These would either be the $k$ smallest eigenvectors or the $k$ largest eigenvectors, depending on the Laplacian version being used. For example, a simple way of partitioning the data into two clusters would be considering the second eigenvector as an indicator vector, assigning items with positive coordinate values into one cluster, and items with negative coordinate values to another cluster.

Spectral clustering algorithms have several significant advantages. First, they do not make any assumptions on the clusters, which allows flexibility in discovering various partitions (unlike the $k$-means algorithm, for example, which assumes that the clusters are spherical). Second, they rely on basic linear algebra operations. And finally, while spectral clustering methods can be costly for large and "dense" data sets, they are particularly efficient when the Laplacian matrix is sparse (i.e., when many pairs of points are of zero affinity). Spectral methods can also serve in dimensionality reduction for high-dimensional data sets (the new dimension being the number of clusters $k$).

Note, that the problem to determine the optimal ("true") number of groups for a given data set is very crucial in cluster analysis. This task arising in many applications. As usual, the clustering solutions, obtained for several numbers of clusters are compared according to the chosen criteria. The sought number yields the optimal quality in accordance with the chosen rule. The problem may have more than one solution and is known as an "ill posed" (Jain and Dubes, 1988) and (Gordon, 1999). For instance, an answer here can depend on the scale in which the data is measured. Many approaches were proposed to solve this...
problem, yet none has been accepted as superior so far.


In the papers Volkovich, Barzily and Morozensky (Volkovich et al., 2008), Barzily, Volkovich, Akteke-Ozturk and Weber (Barzily et al., 2009), Toledano-Kitai, Avros and Volkovich (Toledano-Kitai et al., 2011), methods using the goodness of fit concepts are suggested. Here, the source cluster distributions are constructed based on a model designed to represent well-mixed samples within the clusters.

Another very common, in this area, methodology employs the stability concepts. Apparently, Jain and Moreau (Jain and Moreau, 1987) were the first to propose such a point of view in the cluster validation thematic and used the dispersions of empirical distributions of the cluster object function as a stability measure. Following this perception, differences between solutions obtained via rerunning a clustering algorithm on the same datum evaluate the partitions stability. Hence, the number of clusters minimizing partitions’ changeability is used to assess the “true” number of clusters. In papers of Levine and Domany (Levine and Domany, 2001), Ben-Hur, Elisseeff and Guyon (Ben-Hur et al., 2002), Ben-Hur and Guyon (Ben-Hur et al., 2003) and Dudoit and Fridlyand (Dudoit and Fridlyand, 2002) (the CLEST method), stability criteria are understood to be the fraction of times that pairs of elements maintain the same membership under reruns of the clustering algorithm. Mufti, Bertrand, and El Mouharki (Mufti et al., 2005) exploit Leowinger’s measure of isolation to determine a stability function.

In this paper we offer a new approach to an open problem in spectral clustering which concerns automatically finding the number of clusters. Our approach is based on the stability concept. Here we generalize the method for the scale parameter selecting offered in the Ng-Jordan-Weiss (NJW) algorithm and reveal a connection with the distance learning methodology. Values of the scaling parameter, estimated via clustering of the drawn samples for the number of clusters allocated in a given area, are considered as a cluster stability attitude such that the preferred number of clusters corresponds to the most concentrated empirical distribution of the parameter. Provided numerical experiments demonstrate high potential ability of the offered method. The rest of the paper is organized in the following way. Section 2 is devoted to statement of the base facts of cluster analysis used and to a discussion of the scale parameter selection approaches. In section 3 we propose an application of the offered methodology to the cluster validation problem. Section 4 is devoted to the numerical experiments provided.

2 CLUSTERING

We consider a finite subset \( X = \{x_1, \ldots, x_n\} \) of the Euclidean space \( \mathbb{R}^d \). A partition of the set \( X \) into \( k \) clusters is a collection of \( k \) non-empty of its subsets \( \Pi_k = \{\pi_1, \ldots, \pi_k\} \) satisfying the conditions:

\[
\bigcup_{i=1}^{k} \pi_i = X,
\]

\[\pi_i \cap \pi_j = \emptyset \text{ if } i \neq j.\]

The partition’s elements are named clusters. Two partitions are identical if and only if every cluster in the first partition is also presented in the second one and vice versa. In other words, both partitions have the same clusters up to a permutation. In cluster analysis a partition is chosen so that a given quality

\[
Q(\Pi_k) = \sum_{i=1}^{k} q(\pi_i)
\]

is optimized for some real valued function \( q \) whose domain is the set of subsets of \( X \). The function \( q \) is a distance-like function and, commonly, it is not required to be positive or to satisfy the triangle inequality. In case of the hard clustering the underlying distribution of \( X \) is assumed to be represented in the form

\[\mu_X = \sum_{i=1}^{k} p_i \eta_i,\]

where \( p_i, i = 1, \ldots, k \) are the clusters’ probabilities and \( \eta_i, i = 1, \ldots, k \) are the clusters’ distributions. Note, that this supposition is widespread in clustering, pattern recognition and multivariate density estimation (see, for example, McLachlan and Peel, 2000). Particularly, the most prevalent Gaussian model considers distributions \( \eta_i \) having densities

\[f_i(x) = \phi(x|m_i, \Gamma_i), i = 1, \ldots, k,\]

where \( \phi(x|m_i, \Gamma_i) \) denotes the Gaussian density with mean vector \( m_i \) and covariance matrix \( \Gamma_i \). Usually, the mixture parameters
\[ \theta = (p_i, m_i, \Gamma_i), \ i = 1, ..., k \]

are estimated in this case by maximizing the likelihood

\[ L(\theta|x_1, ..., x_n) = \sum_{j=1}^{n} \ln \left( \sum_{i=1}^{k} p_i \phi(x_j|m_i, \Gamma_i) \right). \quad (1) \]

The most common procedure for maximum likelihood clustering solution is the EM algorithm (see, for example (McLachlan and Peel, 2000)). The EM algorithm provides, in many cases, meaningful results. However, the algorithm often converges slowly and has a strong dependence on its starting position. One of the important EM related algorithms is a Classification EM algorithm (CEM) introduced by Celeux and Govaert in (Celeux and Govaert, 1992). CEM maximizes the Classification Likelihood criterion which is different from the Maximum Likelihood criterion (1). In fact, it does not yield maximum likelihood estimates and can lead to inconsistent values (see, for example (McLachlan and Peel, 2000), section 2.21).

The \( k \)-means approach has been introduced in (Forgy, 1965) and in (MacQueen, 1967). It provides the clusters which approximately minimize the sum of the items’ squared Euclidean distances from cluster centers, which are called centroids. The algorithm generates linear boundaries among clusters. Celeux and Govaert (Celeux and Govaert, 1992) showed that, in the case of the Gaussian mixture model, this procedure actually assumes that all mixture proportions are equal

\[ p_1 = p_2 = ... = p_k; \]

and the covariance matrix is of the form:

\[ \Gamma_i = \sigma^2 I, \ i = 1, ..., k, \]

where \( I \) is the identity matrix of order \( d \) and \( \sigma^2 \) is an unknown parameter. In other words, the \( k \)-means algorithm is, evidently, a particular case of the CEM algorithm.

Spectral clustering skills commonly leverage the spectrum of a given similarity matrix in order to perform dimensionality reduction for clustering in fewer dimensions. Note, that there is a large family of possible algorithms based on the spectral clustering methodology (see, for example (Nascimento and Carvalho, 2011), (Luxburg, 2007), (Filippone et al., 2008)).

Here, we concentrate on a relatively simple technique offered in (Ng et al., 2001) in order to demonstrate the ability of the proposed approach.

**Algorithm 2.1. Spectral Clustering (X, k, \( \sigma \))(NJW)**

**Input**
- \( X \) - the data to be clustered;
- \( k \) - number of clusters;
- \( \sigma \) - the scaling parameter.

**Output**
\( \Pi_{k, \sigma}(X) \) - a partition of \( X \) into \( k \) clusters depending on \( \sigma \).

**Algorithm 2.2. Self-Learning Spectral Clustering (X, k, F)**

**Input**
- \( X \) - the data to be clustered;
- \( k \) - number of clusters;
The goal is to learn a distance metric and pairs in maximizing the sum of dissimilar distances related to the distance metric in the form:

\[ d_{\text{dissimilar}} \text{ data points.} \]

To this end, we define a distance, (i.e., close to each other) while distinguishing the similarity between pairs of elements of data collection is what follows, it is presumed that the degree of similarity is characterized by a positive semi-definite matrix, \( C \geq 0 \) which is learned. We can formulate a constrained optimization problem where we aim to minimize the sum of similar distances concerning pairs in \( S \) while maximizing the sum of dissimilar distances related to pairs in \( D \) in the following way:

\[
\min_C \sum_{(x_i, x_j) \in S} ||x_i - x_j||_C^2
\]

s.t.

\[
\sum_{(x_i, x_j) \in D} ||x_i - x_j||_C^2 \geq 1, \quad C \succ 0
\]

If we suppose that the purported metric matrix is diagonal then minimizing the function is equivalent to solving the stated optimization problem (Xing et al., 2002) up to a multiplication of \( C \) by a positive constant. So, the second quality function can be offered as

\[
F_2(\Pi_k, \sigma(X)) = \sum_{(x_i, x_j) \in S, i \neq j} ||y_i - y_j||^2 - \log \left( \sum_{(x_i, x_j) \in D} ||y_i - y_j|| \right).
\]

Finally, in the spirit of the Fisher’s linear discriminant analysis we can consider the function:

\[
F_3(\Pi_k, \sigma(X)) = \sum_{(x_i, x_j) \in S, i \neq j} ||y_i - y_j||^2 \quad \text{and} \quad \sum_{(x_i, x_j) \in D} ||y_i - y_j||^2.
\]

3. AN APPLICATION TO THE CLUSTER VALIDATION PROBLEM

In this section we discuss an application of the offered methodology to the cluster validation problem. We suggest that these values should be learned from samples clustered for several clusters quantities such that the most stable behaviour of the parameter is exhibited when the cluster structure is the most stable. In our case, it means that the number of clusters is chosen by the best possible way. The drawbacks of the used algorithm together with the complexity of the dataset structure add to the uncertainty of the process outcome. To overcome this ambiguity, a sufficient amount of data has to be involved. This is achieved by drawing many samples and constructing an empirical distribution of the scaling parameter values. The most concentrated distribution corresponds to the appropriate number of clusters.

Algorithm 3.1. Spectral Clustering Validation

\( (X, K, F, J, m, Ind) \)

Input

- \( X \) - the data to be clustered;
- \( K \) - maximal number of clusters to be tested;
- \( F \) - cluster quality function to be minimized;
- \( J \) - number of samples to be drawn;
- \( m \) - size of samples to be drawn;
- \( Ind \) - concentration index.
• \( k^* \) – an estimated number of clusters in the dataset.

---

For \( k = 2 \) to \( K \) do
- For \( j = 1 \) to \( J \) do
  - \( S = \text{sample}(X, m) \);
  - \( \sigma_j = \text{Self-Learning Spectral Clustering}(X, k, F) \);
- end For \( j \)
- Compute \( C_k = \text{Ind}\{\sigma_1, ..., \sigma_J\} \)
- end For \( k \)

The “true” number of clusters \( k^* \) is chosen according to the most concentrated distribution indicated by an appropriate value of \( C_k \), \( k = 2, ..., K \).

3.1 Remarks Concerning the Algorithm

Here, \( \text{sample}(X, m) \) denotes a procedure of drawing a sample of size \( m \) from the population \( X \) without repetitions. Concentration indexes can be provided in several ways. The most widespread instrument used for the evaluation of a distribution’s concentration is the standard deviation. However, it is sensitive to outliers and can be principally dependent, in our situation, on the number of clusters examined. To counterbalance this reliance, the values have been normalized. Unfortunately, it has been specified in the clustering literature that the standard “correct” strategy, for normalization and scaling, does not exist (see, for example (Roth et al., 2004) and (Tibshirani et al., 2001)). We use the coefficient of variation (CV) which is defined as the ratio of the sample standard deviation to the sample mean. For comparison between arrays with different units this value is preferred to the standard deviation because it is a dimensionless number.

4 NUMERICAL EXPERIMENTS

We exemplify the described approach by means of various numerical experiments on synthetic and real datasets provided for the three functions mentioned in 2-4. We choose \( K = 7 \) in all tests and perform 10 trials for each experiment. The results are presented via the error-bar plots of the coefficient of variation within the trials.

4.1 Synthetic Data

The first example consists of a mixture of 5 two-dimensional Gaussian distributions with independent coordinates with the same standard deviation \( \sigma = 0.25 \). The components means are placed on the unit circle with the angular neighboring distance \( 2\pi/5 \). The dataset contains (denoted as \( G5 \)) 4000 items. The scatterplot of this data is presented in the next figure.
The CV index demonstrates approximately the same performance for all object functions hinting to a 5 or 7 clusters structure. However, the bars do not overlap only in the first case where a 5 cluster partition is properly indicated.

4.2 Real-world Data

4.2.1 Three Texts Collection

The first real dataset is chosen from the text collection http://ftp.cs.cornell.edu/pub/smart/. This set (denoted as T3) includes the following three text collections:

- DC0–Medlars Collection (1033 medical abstracts);
- DC1–CISI Collection (1460 information science abstracts);
- DC2–Cranfield Collection (1400 aerodynamics abstracts).

This dataset was considered in many works (Dhillon and Modha, 2001), (Kogan et al., 2003a), (Kogan et al., 2003b), (Kogan et al., 2003c) and (Volkovich et al., 2004). Usually, following the well-known “bag of words” approach, 600 “best” terms were selected (see, (Dhillon et al., 2003) for term selection details). So, the dataset was mapped into Euclidean spaces with dimensions 600. A dimension reduction is provided by the Principal Component Analysis (PCA). The considered dataset is recognized to be well-separated by means of the two leading principal components. We use this data representation in our experiments. The results presented in Fig. 5-7 for \( m = J = 100 \) show that the number of clusters was properly determined for all functions \( F \).

4.2.2 Iris Flower Dataset

Another real dataset chosen is the well-known Iris flower dataset or Fisher’s Iris dataset available, for example, at http://archive.ics.uci.edu/ml/datasets/Iris. The collection includes 50 samples from each of three species of Iris flowers:

- I. setosa;
- I. virginica;
- I. versicolor.

These species compose three clusters situated in a manner that one cluster is linearly separable from the others, but the other two are not. This dataset was analyzed in many papers. A two cluster structure was detected in (Roth et al., 2002). Here, we selected 100 samples of size 140 for each tested number of clusters. As it can be seen, the “true” number of clusters has been successfully found for the \( F_2 \) and \( F_3 \) objective functions. The experiments with \( F_1 \) offer a two clusters configuration.

4.2.3 The Wine Recognition Dataset

The last real dataset contains 178 results of a chemical analysis of three different types (cultivates) of wine given by their 13 ingredients. This collection is available at http://archive.ics.uci.edu/ml/machine-learning-databases/wine. This data collection is relatively small however it exhibits a high dimension. The parameters in use were \( J = 100 \) and \( m = 100 \). Fig. 4 demonstrates undoubtedly that for \( F_2 \) and \( F_3 \) the true number of clusters is revealed, however \( F_1 \) detects a wrong structure.
4.2.4 The Glass Dataset

This dataset is taken from the UC Irvine Machine Learning Repository collection. (http://archive.ics.uci.edu/ml/index.html). The study of classification of glass types was motivated by criminology investigation. The glass found at the place of a crime, can be used as evidence. Number of Instances: 214. Number of Attributes: 9. Type of glass: (class attribute)

- building_windows_float_processed;
- building_windows_non_float_processed;
- vehicle_windows_float_processed;
- vehicle_windows_non_float_processed (not presented);
- containers;
- tableware;
- headlamps.

Fig. 14-16 demonstrate outcomes obtained for $J = 100$. Note, that this relatively small dataset possess a comparatively large dimension and a significantly larger, in comparison with previous collection, suggested number of clusters. To eliminate the influence of the sample size on the clustering solutions we draw samples with growing sizes $m = \max((k - 1) \times 40, 214)$. The minimal value depicted in the graph corresponding to the $F3$ function is 6, however the bars of "2" and "6" overlap. Since the index behavior is more stable once the number of clusters is 6, this value is accepted as the true number of clusters. Other function do not success in determining the true number of clusters.

4.2.5 Comparison of the Partition Quality Function used

Table 1 summarizes the results of the numerical experiments provided. As can be seen, the functions $F2$
and Lai index (KL) (Krzanowski and Lai, 1985), the Sugar and James index (SJ) (Sugar and James, 2003), the GAP-index (Tibshirani et al., 2001) and the Clest-index (Dudoit and Fridlyand, 2002). Our method succeeds quite well in the comparison in case ones an appropriate quality function was chosen.

Table 2: Comparison with other methods.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CH</th>
<th>KL</th>
<th>SJ</th>
<th>Gap</th>
<th>Clest</th>
</tr>
</thead>
<tbody>
<tr>
<td>G5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>T3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Iris</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Wine</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Glass</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>3</td>
</tr>
</tbody>
</table>

5 CONCLUSIONS AND FUTURE WORK

In this paper a new approach to determine the number of the groups in spectral clustering was presented. An empirical distribution of the scaling parameter, found resting upon samples clusterization, is considered as a new cluster stability feature. We analyze three cost functions which can be used in a self-tuning version of a spectral clustering algorithm. In the future research we plan to generalize our method to the Local Scaling methodology (Zelnik-manor and Perona, 2004) and compare the obtained outcomes. Another research direction can consist of a study of the model behavior when the number of clusters is suggested to be relatively big. An essential ingredient of each resampling cluster validation approach is the selection of the parameters values in an implementation. It is difficult to treat this task from a theoretical point of view (see, e.g. (Dudoit and Fridlyand, 2002), (Roth et al., 2004) and (Levine and Domany, 2001)). We are going to investigate this matter in our future papers.

REFERENCES


