An Approach to Identify the Number of Clusters

Katelyn Gao* Heather Hardeman† Edward Lim‡
Cristian Potter§ Carl Meyer¶ Ralph Abbey‖

July 11, 2012

Abstract

In this technological age, vast amounts of data are generated. Various statistical methods are used to find patterns in data, including clustering. Many common methods for cluster analysis, such as k-means and Nonnegative Matrix Factorization, require input of the number of clusters in the data. However, usually that number is unknown. There exists a method that uses eigenvalues to compute the number of clusters, but sometimes it underestimates that number. In this paper, we propose a complementary method to identify the number of clusters. This method is used to analyze three data sets and gives fairly accurate estimates of the number of clusters.

1 Introduction

Cluster analysis is a type of data analysis that partitions observations into groups, or clusters, so that observations in the same group have similar characteristics. For numerical data, numbers that are close to one another would be clustered together. Nonnumerical data, such as text data, can be converted to numerical data and then clustered.

*Massachusetts Institute of Technology, Cambridge, MA 02139, katelyng@mit.edu
†University of Montevallo, Montevallo, AL 35115, hhardema@forum.montevallo.edu
‡Johns Hopkins University, Baltimore, MD 21218, ed1229@gmail.com
§East Carolina University, Greenville, NC 27858, cwpotter223@gmail.com
¶North Carolina State University, Raleigh, NC 27695, meyer@ncsu.edu
‖North Carolina State University, Raleigh, NC 27695, rwabbey@ncsu.edu

There are numerous algorithms that cluster data into k clusters when k is specified. Two common ones are k-means [1] and Nonnegative Matrix Factorization (NMF) [2]. However, in many situations k is unknown, so it is often necessary to estimate k before executing an algorithm. The Stochastic Clustering Algorithm (SCA) [3] does not require k to be input. Instead, the number of clusters is calculated in a step of the algorithm that we call the Stochastic Clustering (SC) method.

Our task was to explore an alternative method of deriving k, the number of clusters, from a data set. We devised a new, graphical method that, when applied to three well-known data sets, performs better than the SC method.

This paper is organized as follows. In Section 2, we introduce and discuss some important clustering terminology and describe the SC method. In Section 3, we describe and give intuition for our new approach to find the number of clusters. In Section 4, we test our method on three data sets and explore its limitations. Lastly, in Section 5, we conclude our discussion.

2 Background

A clustering algorithm takes as input data with n observations and groups them into k clusters. A **clustering assignment** is a surjective map f from the set of data points $D = \{d_1, d_2, ..., d_n\}$ to the set of clusters $C = \{1, 2, ..., k\}$; that is, $f : D \to C$.

Given a clustering assignment, we can construct a symmetric adjacency matrix A, which is of size $n \times n$ and is defined as follows:

$$a_{ij} = \begin{cases} 1 & \text{if } f(d_i) = f(d_j) \\ 0 & \text{otherwise} \end{cases}$$

With multiple runs of a clustering algorithm on a data set, we have multiple clustering assignments and adjacency matrices. Then, we may calculate a symmetric **consensus matrix**. Formally, with m runs of clustering algorithm and corresponding adjacency matrices A_1, A_2, . . . , A_m, the **consensus matrix** is defined as $S = \frac{1}{m} \sum_{i=1}^{m} A_i$.

To calculate the number of clusters in a data set, the SC method takes as input a $n \times n$ consensus matrix S, constructed after multiple runs of some

clustering method on a data set using a variety of numbers of clusters. First, it converts S into a $n \times n$ doubly stochastic matrix P using the Sinkhorn-Knopp Algorithm [4]. Then, it calculates the eigenvalues of P, which lie in $[0, 1]$: $1 \geq \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n \geq 0$. If the largest difference between consecutive eigenvalues is between λ_k and λ_{k+1}, the SC method concludes that the data set has k clusters. The theory behind the SC method can be found in Meyer and Wessell [3].

At the crux of our new method to calculate the number of clusters in a data set is the **uncoupling measure** of a consensus matrix, given in Meyer and Wessell [3]. It is defined as follows.

Let S be a $n \times n$ consensus matrix and n_1 and n_2 be positive integers such that $n_1 + n_2 = n$. S can be partitioned into the form

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$$

where S_{11} is $n_1 \times n_1$ and S_{22} is $n_2 \times n_2$ for $n_1 \in [1, n-1]$. The **uncoupling measure of S with respect to** n_1 is the function $\sigma(S, n_1) = \dfrac{e^T S_{12} e + e^T S_{21} e}{e^T S e} = \dfrac{2e^T S_{12} e}{e^T S e}$, where e is a column vector of ones. In other words, the **uncoupling measure** is the ratio of the sum of the entries in the off-diagonal blocks to the sum of all entries in the matrix.

3 Proposed Graphical Method: Local Maxima

We can make some observations about the uncoupling measures of a consensus matrix for a data set. Consider some such matrix S. The entry s_{ij} would be close to 1 if the data points d_i and d_j are often assigned to the same cluster by the algorithms used to construct S, and close to 0 otherwise. Suppose that we know the correct clustering assignment of the data. Using that, permute S so that the rows and columns corresponding to data points in the same cluster are adjacent. Then, assuming that the algorithms used to construct S are accurate, S would approximate a block diagonal matrix, with each block corresponding to a cluster. The entries in the blocks would be close to 1, and the other entries would be close to 0.

Now consider the uncoupling measure of S. If n_1 is such that none of the blocks are partitioned, then $\sigma(S, n_1)$ would be close to 0. If it is not, $\sigma(S, n_1)$ would be larger. Indeed, it would be largest when a block is partitioned in

half. Therefore, the plot of $\sigma(S, n_1)$ versus n_1 would have one local maximum for each cluster. That is, the number of clusters in the data set would be equal to the number of local maxima in the plot of the uncoupling measure of a consensus matrix.

Suppose now that we do not know the actual number of clusters and must run a clustering algorithm to derive the clustering assignment used to permute S. If the clustering assignment used to permute S were derived from an algorithm using l clusters, where $l < k$, then rows and columns corresponding to data in different clusters may be mixed together. Their distinct local maxima would be replaced by one local maximum, and we would see at least l and at most k local maxima in the plot of $\sigma(S, n_1)$. However, if $l = k$, we should see k local maxima, with similar reasoning as in the previous paragraph.

On the other hand, if $l > k$, some cluster must be divided into multiple clusters. Then, the rows and columns of S corresponding to the data in that cluster may not be adjacent. The numbers off the diagonal would increase. Thus, the local maxima in the plot of $\sigma(S, n_1)$ would become less pronounced and there may be less than l local maxima.

Based on these observations, we propose the following Local Maxima (LM) method to find the number of clusters in a data set:

1. Construct a consensus matrix S for the data set $D = \{d_1, d_2, ..., d_n\}$ by executing a clustering method multiple times with a variety of numbers of clusters.

2. Repeat the following steps for $l = 2, 3, 4, ...$:

    - Run the same clustering method on the data with l clusters to get a clustering assignment f.
    - Permute S such that if $f(d_i) = f(d_j)$, the rows of S corresponding to d_i and d_j are adjacent and the columns of S corresponding to d_i and d_j are adjacent.
    - Plot $\sigma(S, n_1)$ for $n_1 \in [1, n-1]$, and count the number of local maxima.
    - If this number is less than l, the algorithm stops, and we conclude there are $l - 1$ clusters.

4 Experimental Design and Results

We tested this method on three well-known data sets: the Ruspini data set [5], the Fisher Iris data set [6], and the Leukemia data set [7].

4.1 Ruspini

The Ruspini data set [5] consists of 75 observations with two variables each. It has four clusters. To illustrate the data, we carried out Principle Component Analysis. Figure 1 is a histogram of the coordinates of the first principal component for each data point. There are two clear clusters, although it may be argued that each contains two clusters since there seems to be two modes.

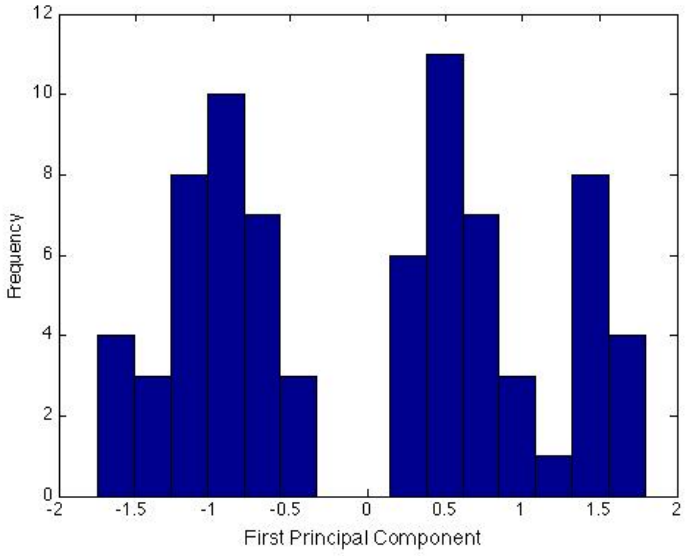


Figure 1: Ruspini: Data Projection on First Principal Component

To build the consensus matrix S, we used the results from executing the k-means algorithm on the data 10 times each for $k = 2$, $k = 3$, $k = 4$, $k = 5$, and $k = 6$.

For simplicity, we just show the last three steps of the method. For $l = 2$,

there were two local maxima in the plot of $\sigma(S, n_1)$ versus n_1. Figure 2(a) shows the next step of the method, the plot of $\sigma(S, n_1)$ versus n_1 for $l = 3$. There are either two or three local maxima. If there are two, the method ends and we say that there are two clusters. If there are three, we continue to the next step.

Figure 2(b) is the plot of $\sigma(S, n_1)$ for $l = 4$. There are two, three, or four local maxima. If there are two or three, the method ends and we say that there are three clusters. If there are four, we continue to the next step.

Figure 2(c) is the plot of $\sigma(S, n_1)$ for $l = 5$. Here, there are definitely less than five local maxima; there are at least two, and possibly two more. Therefore, the method ends.

The LM method predicts that the Ruspini data set to have at least two and up to four clusters.

4.2 Iris

The Fisher Iris data set, first introduced in [6], contains 150 observations with 4 dimensions corresponding to different features of the flower. It has three clusters, one for each species of iris: *Iris setosa*, *Iris versicolor*, and *Iris virginica*. However, a recent paper by Benson-Putnins et al. [8] asserts that there may be actually two clusters in the *Iris setosa* cluster.

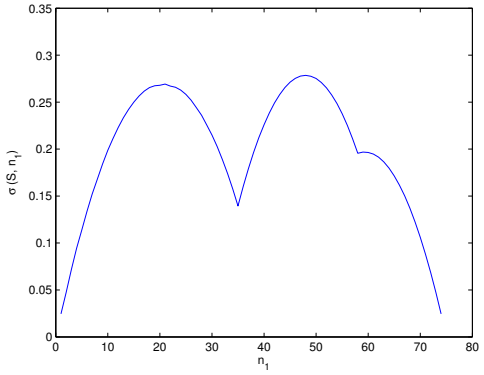
Again, we did Principle Component Analysis on the data. Figure 3 is a histogram of the coordinates of the first principal component for each data point. There are two clear clusters, although it is possible that the second consists of two or three clusters.

To build the consensus matrix, we executed the k-means algorithm 10 times each for $k = 2$, $k = 3$, $k = 4$, and $k = 5$.

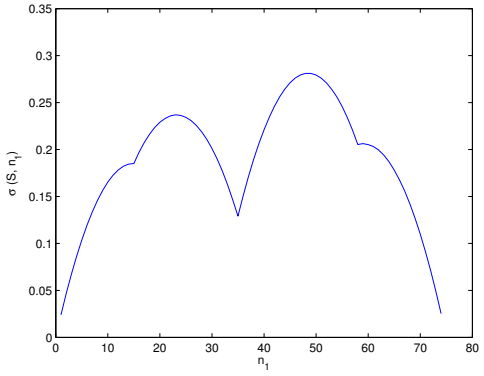
Figure 4(a) is the plot of $\sigma(S, n_1)$ versus n_1 for $l = 2$. There are clearly two local maxima, so we proceed to the next step of the method.

Figure 4(b) is the plot of $\sigma(S, n_1)$ for $l = 3$. There are either two or three local maxima. If there are two, the method ends and we say that there are two clusters. If there are three, we continue to the next step.

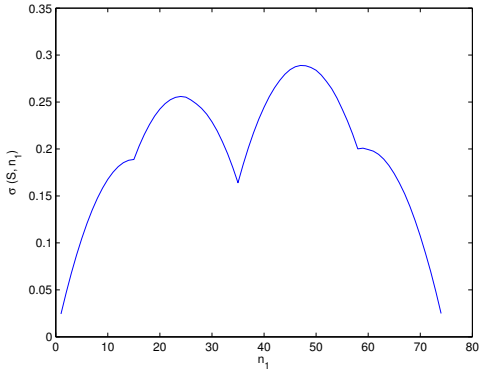
Figure 4(c) is the plot of $\sigma(S, n_1)$ for $l = 4$. There are two local maxima, even though there are several small bumps on the curve. Since this is less than 4, the method ends.



(a) $l = 3$



(b) $l = 4$



(c) $l = 5$

Figure 2: Ruspini: Uncoupling Measures

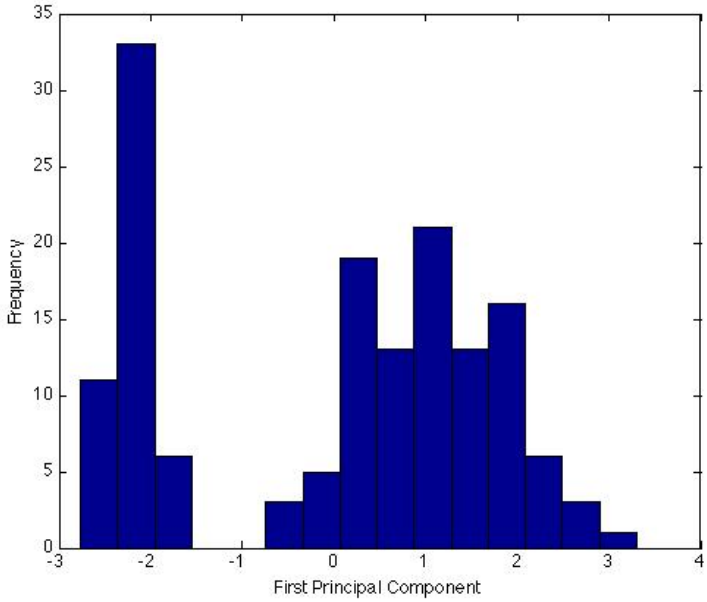


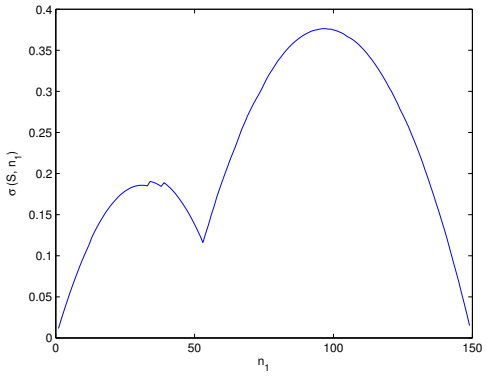
Figure 3: Iris: Data Projection on First Principal Component

We conclude that the Iris data set has two or three clusters.

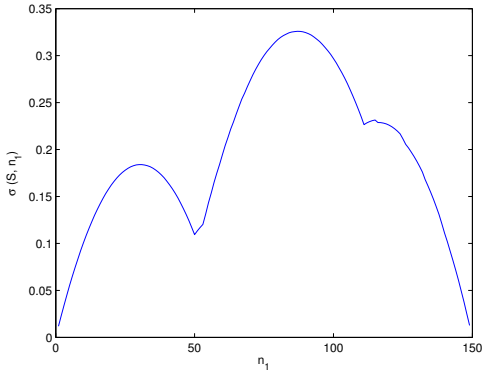
We may make an interesting observation here. As shown in Figure 4(b), there are two clusters that are close together but are further away from the third cluster. Similarly, in the plot corresponding to $l = 3$, there are two clear local maxima and one less apparent one. Thus, it is possible that the existence of an obvious local maximum for a cluster is determined by that cluster's dissimilarity from other clusters.

4.3 Leukemia

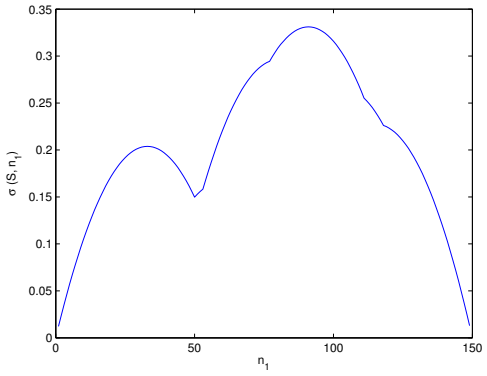
The Leukemia data set [7] consists of 38 observations with 5000 dimensions. It contains gene expression data for 38 cancer patients, provided by the Broad Institute of Harvard and MIT. There are three clusters, corresponding to three types of leukemia: acute lymphoblastic leukemia, B-cell subtype (ALL-B), acute lymphoblastic leukemia T-cell subtype (ALL-T), and acute myelogenous leukemia (AML).



(a) $l = 2$



(b) $l = 3$



(c) $l = 4$

Figure 4: Iris: Uncoupling Measures

To build the consensus matrix S, we executed the k-means algorithm on the data 10 times each for $k = 2$, $k = 3$, $k = 4$, and $k = 5$.

Figure 5a is the plot of $\sigma(S, n_1)$ versus n_1 for $l = 2$. There are clearly only two local maxima, so we proceed to the next step of the method.

Figure 5b is the plot of $\sigma(S, n_1)$ for $l = 3$. There are three local maxima, so we continue to the next step of the method.

Figure 5c is the plot of $\sigma(S, n_1)$ for $l = 4$. Again, there are three local maxima, less than $l = 4$. Therefore, the algorithm ends, and we conclude that there are three clusters in the Leukemia data set.

4.4 Comparison with Stochastic Clustering Method

We implemented the SC method on the Ruspini, Iris, and Leukemia data sets as well. Table 1 shows the number of clusters for each data set returned by each method.

Table 1: Comparison of the Number of Clusters Estimated

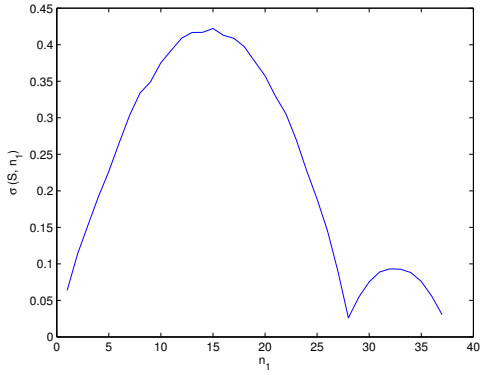
Number of clusters	Ruspini	Iris	Leukemia
SC	1-2	1-2	2
LM	2-4	2-3	3
Actual	4	3	3

We immediately see that the LM method performs just as well or better than the SC method on all three data sets. Both methods sometimes underestimate the number of clusters, but the SC does so more often and with greater error.

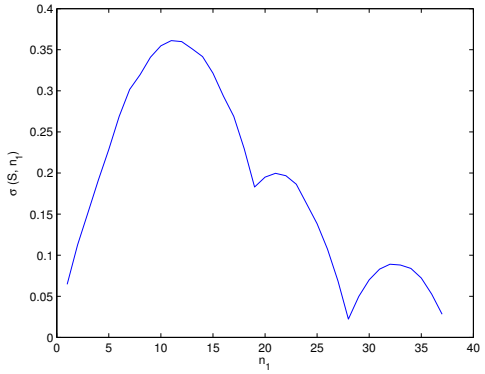
The range of values given for the SC method is due to randomness in the algorithms used to construct the consensus matrices. The range of values given for the LM method is due to this randomness as well as the ambiguity of interpretation of the graphs of $\sigma(S, n_1)$; for example, in the $l = 3$ plot for the Ruspini data set, the number of local maxima is debatable.

4.5 Discussion

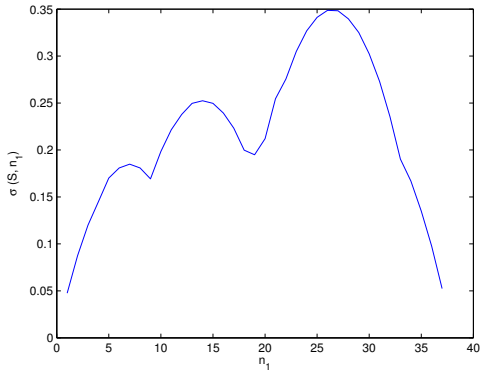
As shown in the experiments, the local maximum for a cluster may not be apparent if the data in that cluster is similar to the data in another cluster.



(a) $l = 2$



(b) $l = 3$



(c) $l = 4$

Figure 5: Leukemia: Uncoupling Measures

Therefore, the LM method may fail if clusters are not sufficiently distinct from one another.

However, we believe that the biggest weakness of the current LM method is its lack of rigor. Now the number of local maxima is counted by eye, but ideally there should be a mathematically rigorous method to analyze the graphs.

5 Conclusion and Future Work

In this paper, we introduced, discussed, and applied our Local Maxima method to identify the number of clusters in three data sets with fairly accurate results. This method is complementary to the Stochastic Clustering method given in Meyer and Wessell [3]. Combining these two methods could allow us to identify the number of clusters in data sets with acceptable accuracy.

We tested the Local Maxima method on three data sets. In future, we plan to apply it to many others to verify its capabilities. Moreover, we would like to examine the theoretical basis for this method and develop a rigorous technique to count the number of local maxima.

6 Acknowledgements

The work contained in this paper was done at the Research Experience for Undergraduates in Mathematics: Modeling and Industrial Applied Mathematics at North Carolina State University in Summer 2011. We would like to thank NC State University for hosting this program and the National Science Foundation and National Security Agency for providing funding that made it possible. Lastly, we would like to thank Dr. Carl Meyer, our faculty mentor, and Ralph Abbey, our graduate student mentor, for guiding us through the research process.

References

[1] T. Hastie, R. Tibshrani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd Ed. Springer,

2009.

[2] D. D. Lee and H. S. Seung, *Learning the parts of objects by non-negative matrix factorization*, **Nature**, 401(1999), pp. 788-791.

[3] C. D. Meyer and C. D. Wessell, *Stochastic Data Clustering*, **SIMAX**, under revision, 2010.

[4] R. Sinkhorn, *A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices*, **The Annals of Mathematical Statistics**, 35(1964), pp. 876-879.

[5] E. H. Ruspini, *Numerical methods for fuzzy clustering.* **Inform. Sci.**, 2(1970), pp. 319-350.

[6] R. A. Fisher, *The use of multiple measurements in taxonomical problems*, **Annals of Eugenics**, 7(1936), pp. 179-188.

[7] The Broad Institute of Harvard and MIT, http://www.broadinstitute.org/mpr/publications/projects/NMF, accessed July 2011.

[8] D. Benson-Putnins, M. Bonfardin, M. E. Magnoni, and D. Martin, *Spectral Clustering and Visualization: A Novel Clustering of Fisher's Iris Data Set*, **SIAM Undergraduate Research Online**, Vol. 4, 2011.