An indicator for the number of clusters using a linear map to simplex structure

Marcus Weber[1], Wasinee Rungsarityotin[2], and Alexander Schliep[2]

[1] Zuse Institute Berlin ZIB
Takustraße 7, D-14195 Berlin, Germany
[2] Computational Molecular Biology, Max Planck Institute for Molecular Genetics
Ihnestraße 63–73, D-14195 Berlin, Germany

Abstract. The problem of clustering data can be formulated as a graph partitioning problem. In this setting, spectral methods for obtaining optimal solutions have received a lot of attention recently. We describe Perron Cluster Cluster Analysis (PCCA) and establish a connection to spectral graph partitioning. We show that in our approach a clustering can be efficiently computed by mapping the eigenvector data onto a simplex. To deal with the prevalent problem of noisy and possibly overlapping data we introduce the Min-chi indicator which helps in confirming the existence of a partition of the data and in selecting the number of clusters with quite favorable performance. Furthermore, if no hard partition exists in the data, the Min-chi can guide in selecting the number of modes in a mixture model. We close with showing results on simulated data generated by a mixture of Gaussians.

1 Introduction

In data analysis, it is a common first step to detect groups of data, or clusters, sharing important characteristics. The relevant body of literature with regard to methods as well as applications is vast (see Hastie et al. (2001) or Jain and Dubes (1988) for an introduction). There are a number of ways to obtain a mathematical model for the data and the concept of similarity between data points, so that one can define a measure of clustering quality and design algorithms for finding a clustering maximizing this measure. The simplest, classical approach is to model data points as vectors from $\mathbb{R}^n$. Euclidean distance between points measures their similarity and the average Euclidean distance between data points to the centroid of the groups they are assigned to is one natural measure for the quality of a clustering. The well-known k-means algorithm, Jain and Dubes (1988), will find a locally optimal solution in that setting.

One of the reasons why the development of clustering algorithms did not cease after k-means are the many intrinsic differences of data sets to be analyzed. Often the measure of similarity between data points might not fulfill all the properties of a mathematical distance function, or the measure of clustering quality has to be adapted, as for example the ball-shape assumption inherent in standard k-means does not often match the shape of clusters in real data.

An issue which is usually, and unfortunately, of little concern, is whether there is a partition of the data into a number of groups in the first place and how many possible groups the data support. Whenever we apply a clustering algorithm which computes a k-partition this is an assumption we imply to hold for the data set we analyze. The problem is more complicated when k is unknown. In the statistical literature, McLachlan et al. (1988) suggested mixture models as alternatives for problem instances where clusters overlap.

We address the problem of finding clusters in data sets for which we do not require the existence of a k-partition. The model we will use is a similarity graph. More specifically, we have $G = (V, E)$, where $V = \{1, \ldots, n\}$ is the set of vertices corresponding to the data points. We have an edge $\{i, j\}$ between two vertices iff we can quantify their similarity, which is denoted $w(i, j)$. The set of all edges is E and the similarities can be considered as a function $w : E \mapsto \mathbb{R}_0^+$. The problem of finding a k-partition of the data can now be formulated as the problem of partitioning V into k subsets, $V = \cup_{i=1}^k V_i$. Let us consider the problem of finding a 2-partition, say $V = A \cup B$. This can be achieved by removing edges $\{i, j\}$ from E for which $i \in A$ and $j \in B$. Such a set of edges which leaves the graph disconnected is called a *cut* and the weight function allows us to quantify cuts by defining their weight or *cut-value*,

$$cut(A, B) := \sum_{\{i,j\} \in E, i \in A, j \in B} w(i, j).$$

A natural objective is to find a cut of minimal value. A problem with this objective function is that sizes of partitions do not matter. As a consequence, using min-cut will often compute very unbalanced partitions, effectively splitting V into one single vertex, or a small number of vertices, and one very large set of vertices. We can alleviate this problem by evaluating cuts differently.

Instead of just considering partition sizes one can also consider the similarity within partitions, for which we introduce the so-called *association* value of a vertex set A denoted by $a(A) = a(A, V) := \sum\limits_{i \in A} \sum\limits_{j \in V} w_{ij}$. Defining the normalized cut by

$$\text{Normcut}(A, B) = \frac{cut(A, B)}{a(A, V)} + \frac{cut(A, B)}{a(B, V)},$$

we observe that the cut value is now measured in terms of the similarity of each partition to the whole graph. Vertices which are more similar to many data points are harder to separate. As we will see, the normalized cut is well suited as an objective function for minimizing because it keeps the relative size and connectivity of clusters balanced.

The min-cut problem can be solved in polynomial time for $k = 2$. Finding k-way cuts in arbitrary graphs for $k > 2$ is proven NP-hard by Dahlhaus et al. (1994). For the two other cut criteria, already the problem of finding a 2-way cut is NPC, for a proof, see appendix in Shi and Malik (2000).

However, we can find good approximate solutions to the 2-way normalized cut by considering a relaxation of the problem, see Kannan et al. (1999) and Shi and Malik (2000). Instead of discrete assignments to partitions consider a continuous indicator for membership. Let $D = diag(d(1), \ldots, d(n))$ and $d(i) = \sum_{j \in V, i \neq j} w(i,j)$. The relaxation of the 2-way normalized cut problem can be formulated as

$$(D - W)x = \lambda Dx. \qquad (1)$$

For solving the 2-partition problem, we are interested in the eigenvector x_2 for the second-smallest eigenvalue, compare Kannan et al. (1999) and Shi and Malik (2000). In particular, we will inspect its sign structure and use the sign of an entry $x_2(i)$ to assign vertex i to one or the other vertex set. Similarly, for direct computation of k-partitions one can use all k eigenvectors to obtain k-dimensional indicator vectors. Previous approaches in Shi and Malik (2000) and Ng et al. (2002) relied on k-means clustering of the indicator vectors to obtain a k-partition in this space.

In the next section, we will propose an indicator for the amount of overlapping in W which helps in deciding whether the recursive spectral method is applicable. Subsequently we will introduce an alternative approach to finding k-partitions even in absence of a perfect block structure. We first rephrase the problem equivalently in terms of transition matrices of Markov-chains and use perturbation analysis to arrive at the main result, a geometric interpretation of the eigenvector data as a simplex. This allows to devise an assignment of data into overlapping groups and a measure for the deviation from the simplex structure, the so-called *Min-chi value*. The advantages of our method are manifold: there are fewer requirements on the similarity measure, it is effective even for high-dimensional data and foremost, with our robust diagnostic we can assess whether a unique k-partition exists. The immediate application value is two-fold. On one hand, the Min-chi value indicates whether trying to partition the data into k groups is possible. On the other hand, if clusters arise from a mixture model, the indicator can be used as a guide for deciding on the number of modes in a mixture model. We close with showing results on simulated data generated by a mixture of Gaussians.

2 Clustering Method

2.1 Simplex Structure and Perturbation Analysis

One can transform equation (1) into an eigenvalue problem for a stochastic matrix:

$$\begin{aligned} (D - W)x &= \lambda Dx \\ \Leftrightarrow (I - D^{-1}W)x &= \lambda x \end{aligned}$$

$$\Leftrightarrow D^{-1}Wx = \underbrace{(1-\lambda)}_{=\bar{\lambda}} x.$$

In this equation $T = D^{-1}W$ is a stochastic matrix and the eigenvalues $1 \geq \bar{\lambda} \geq -1$ are real valued, because W is symmetric.

If W has a perfect block diagonal structure with k blocks, then clustering should lead to k perfectly separated index sets $C_1, \ldots, C_k$. With W the matrix T also has perfect block diagonal structure and due to the row sum of stochastic matrices the characteristic vectors[1] $\chi_1, \ldots, \chi_k$ of the sets $C_1, \ldots, C_k$ are eigenvectors of T for the k-fold maximal eigenvalue $\bar{\lambda}_1 = \ldots = \bar{\lambda}_k = 1$. The numerical eigenvector computation in this case provides an arbitrary basis $X = [x_1, \ldots, x_k]$ of the eigenspace corresponding to the eigenvalue $\bar{\lambda} = 1$, i.e. with $\chi = [\chi_1, \ldots, \chi_k]$ there is a transformation matrix $\mathcal{A} \in \mathbb{R}^{k \times k}$ with

$$\chi = X\mathcal{A}. \tag{2}$$

In other words: If one wants to find the clustering of a perfect block diagonal matrix T, one has to compute the transformation matrix $\mathcal{A}$ which transforms the eigenvector data into characteristic vectors. If $\widetilde{T}$ has almost block structure it can be seen as an ϵ-perturbed stochastic matrix of the ideal case T. For $\widetilde{T}$ the k-fold eigenvalue $\bar{\lambda} = 1$ degenerates into one Perron eigenvalue $\widetilde{\bar{\lambda}}_1 = 1$ with a constant eigenvector and a cluster of $k-1$ eigenvalues $\widetilde{\bar{\lambda}}_2, \ldots, \widetilde{\bar{\lambda}}_k$ near 1, the so-called Perron cluster. It has been shown, that there is a transformation matrix $\widetilde{\mathcal{A}}$ such that

$$\chi - \widetilde{\chi} = O(\epsilon^2)$$

for $\widetilde{\chi} = \widetilde{X}\widetilde{\mathcal{A}}$, see Deuflhard and Weber (2005). If the result $\widetilde{\chi}$ shall be interpretable, then the vectors $\widetilde{\chi}_1, \ldots, \widetilde{\chi}_k$ have to be "*close to*" characteristic: I.e., they have to be nonnegative and provide a partition of unity. In other words: The rows of $\widetilde{\chi}$ as points in $\mathbb{R}^k$ have to lie inside a simplex spanned by the k unit vectors. If clustering is possible, then additionally, for the reason of maximal separation of the clusters, for every almost characteristic vector $\widetilde{\chi}_i$ there should be an entry l with $\widetilde{\chi}_i(l) = 1$. It has been shown, that there is always a possibility to meet three of the four conditions (i) nonnegativity, (ii) partition of unity, (iii) $\widetilde{\chi} = \widetilde{X}\mathcal{A}$, and (iv) 1-entry in every vector. If all four conditions hold, the solution $\widetilde{\chi}$ is unique, see Deuflhard and Weber (2005). In this case the eigenvector data itself spans a simplex. This simplex can be found via the *inner simplex algorithm*, see Weber and Galliat (2002) and Deuflhard and Weber (2005). The result $\widetilde{\chi}$ of this algorithm always meets the conditions (ii)-(iv), but the solution may have negative components. The absolute value of the minimal entry of $\widetilde{\chi}$ is called the Min-chi indicator. As the uniqueness of the clustering increases, Min-chi goes to zero. Due to perturbation analysis it has been shown, that Min-chi$= O(\epsilon^2)$, see Weber (2004).

[1] A characteristic vector χ_i of an index subset C_i meets $\chi_i(l) = 1$ iff $l \in C_i$, and $\chi_i(l) = 0$ elsewhere.

2.2 Implementation: Min-chi in practice

Given an $n \times m$ data matrix, we compute pairwise-distances for all pairs and construct the $n \times n$ distance matrix A with a symmetric distance function $w : \mathbb{R}^m \times \mathbb{R}^m \mapsto \mathbb{R}_0^+$. We then convert the distance to a similarity matrix with $W = \exp(-\beta A)$ where β is a scaling parameter and the stochastic matrix is defined by $T = D^{-1}W$. We can use the error measure Min-chi to determine a locally optimal solution for the number of clusters. Given the matrix T, we can use our method to determine a number of clusters denoted by k as follows:

The Mode Selection Algorithm

1. Choose $k_{min}, \ldots, k_{max}$ such that the optimal k could be in the interval,
2. Iterate from $k_{min}, \ldots, k_{max}$ and for each k-th trial, calculate χ for cluster assignment via the *Inner Simplex* algorithm and Min-chi as an indicator for the number of clusters,
3. Choose the maximum k for which Min-chi $<$ Threshold as the number of clusters.

Selections of the threshold depends on the value β or variance which controls the perturbation from the perfect block structure of T. As a rule, when β is large, the threshold can be small because T is almost block-diagonal.

3 Result and Discussion

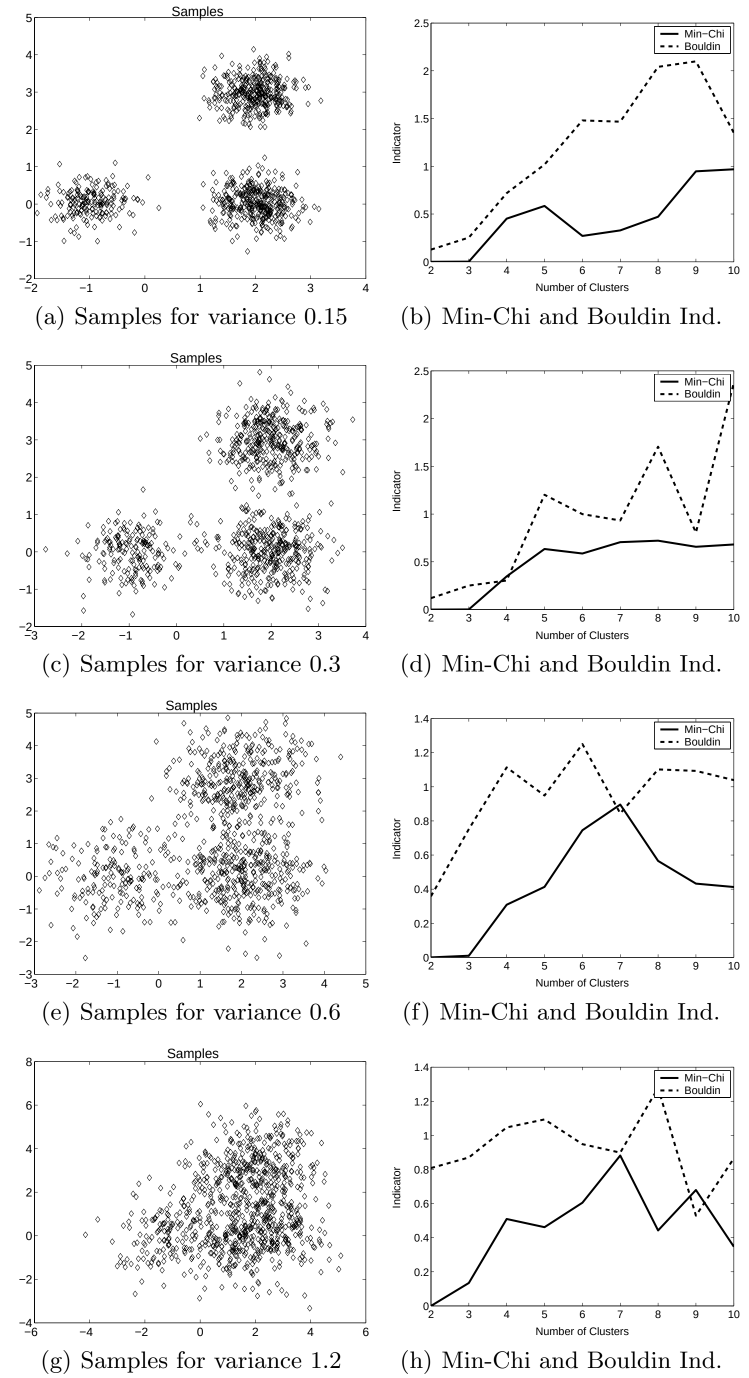
We compare the Min-chi indicator with the Bouldin index defined in Jain and Dubes (1988) applied to the result from the Inner Simplex algorithm described in details by Weber and Galliat (2002) and Deuflhard and Weber (2005). Given a partition into k clusters by a clustering algorithm, one first defines the measure of within-to-between cluster spread for the ith cluster with the notation $R_i = \max\limits_{j \neq i} \frac{e_j + e_i}{m_{ji}}$, where e_i is the average distance within the ith cluster and m_{ij} is the Euclidean distance between the means. The Bouldin index for k is

$$DB(k) = \frac{1}{k} \sum_{i>1} R_i.$$

According to the Bouldin indicator, the number of clusters is k^* such that

$$k^* = \underset{k_{min} \ldots k_{max}}{\operatorname{argmin}} \; DB(k).$$

In the examples of Fig. 3 we compute a sampling of 900 points from three spherical Gaussians with different variances and means. 180 points with mean $(-1, 0)$ and 360 points with mean $(2, 0)$ and $(2, 3)$ respectively. For three different variances $0.15, 0.3, 0.6$ and 1.2 we compute the Bouldin index and



(a) Samples for variance 0.15 (b) Min-Chi and Bouldin Ind.

(c) Samples for variance 0.3 (d) Min-Chi and Bouldin Ind.

(e) Samples for variance 0.6 (f) Min-Chi and Bouldin Ind.

(g) Samples for variance 1.2 (h) Min-Chi and Bouldin Ind.

Fig. 1. Simulated data: Mixture of three spherical gaussians with different variances. Comparison of Min-chi with the Bouldin index.

the Min-chi indicator for $k_{min} = 2$ and $k_{max} = 10$. For a low variance in Fig. 1(a) both indicators give the same result $k = 3$, but for increasing variance in Fig. 1(c) and Fig. 1(e) the Bouldin indicator fails, whereas the Min-chi indicator still finds three clusters. For very high variance in Fig. 1(g), the Bouldin index finds 9 clusters. In this experiment, the Min-chi indicator is not unique. Depending on the threshold, two or three clusteres are indicated. This behaviour becomes worse for increasing variance.

4 Conclusion

In this paper we have shown the relation between Perron Cluster Cluster Analysis and spectral clustering methods. Some changes of PCCA with regard to geometrical clustering have been proposed, e.g. the Min-chi indicator for the number k of clusters. We have shown that this indicator is valuable also for noisy data. It evaluates the deviation of some eigenvector data from simplex structure and, therefore, it indicates the possibility of a "fuzzy" clustering, i.e. a clustering with a certain number of almost characteristic functions. A simple linear mapping of the eigenvector data has to be performed in order to compute these almost characteristic functions. Therefore, the cluster algorithm is easy to implement and fast in practice. We have also shown, that PCCA does not need strong assumptions like other spectral graph partitioning methods, because it uses the full eigenvector information and not only signs or less than k eigenvectors.

References

DAHLHAUS, E., JOHNSON, D. S., PAPADIMITRIOU, C. H., SEYMOUR, P. D. and M. YANNAKAKIS (1994): The complexity of multiterminal cuts. *SIAM J. Comput.*, 23(4):864–894.

DEUFLHARD, P. and WEBER, M. (2005): Robust Perron Cluster Analysis in Conformation Dynamics. *Lin. Alg. App., Special Issue on Matrices and Mathematical Biology*, 398c:161–184.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001): *The Elements of Statistical Learning.* Springer, Berlin.

JAIN, A.K. and DUBES, R.C. (1988): *Algorithms for clustering data.* Prentice Hall, Engelwood Cliffs.

KANNAN, R., VEMPALA, S. and VETTA, A. (1999): On Clusterings: Good, Bad and Spectral. *Proceedings of IEEE Foundations of Computer Science.*

MCLACHLAN, G.J. and BASFORD, K.E. (1988): *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, Inc., New York, Basel.

NG, A. Y., JORDAN, M. and WEISS, J (2002): On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems 14.*

SHI, J. and MALIK, J. (2000): Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

WEBER, M. (2004): Clustering by using a simplex structure. Technical report, ZR-04-03, Zuse Institute Berlin.

WEBER, M. and GALLIAT, T (2002): Characterization of transition states in conformational dynamics using Fuzzy sets. Technical Report 02–12, Zuse Institute Berlin (ZIB).