Strict Monotonicity of Sum of Squares Error and Normalized Cut in the Lattice of Clusterings

Nicola Rebagliati NICOLA.REBAGLIATI@GMAIL.COM
VTT Technical Research Centre of Finland
P.O. Box 1000, VTT 02044, Finland

Abstract

Sum of Squares Error and Normalized Cut are two widely used clustering functional. It is known their minimum values are monotone with respect to the input number of clusters and this monotonicity does not allow for a simple automatic selection of a correct number of clusters. Here we study monotonicity not just on the minimizers but on the entire clustering lattice. We show the value of Sum of Squares Error is strictly monotone under the strict refinement relation of clusterings and we obtain data-dependent bounds on the difference between the value of a clustering and one of its refinements. Using analogous techniques we show the value of Normalized Cut is strictly anti-monotone. These results imply that even if we restrict our solutions to form a chain of clustering, like the one we get from hierarchical algorithms, we cannot rely on the functional values in order to choose the number of clusters. By using these results we get some data-dependent bounds on the difference of the values of any two clusterings.

1. Introduction

Sum of Squares Error and Normalized Cut are two different clustering functionals widely used in Machine Learning applications (von Luxburg, 2007; Jain, 2010). Minimizing the Sum of Squares Error on a set of points allows the representation of each cluster with a single point called prototype, or centroid. On the other side, minimizing the Normalized Cut aims at dividing a graph into components with balanced volumes. These two minimization problems are NP-complete, but they both enjoy approximation algorithms which are quite successful in applications.

A principal area of research for these functionals is model validation, i.e. choosing a correct number of clusters. It is easy to empirically observe, on a given set of points, that the minimum Sum of Squares Error value decreases with the number of clusters and it is zero if we cluster each point within itself. Similarly in (Nagai, 2007) it was proved that the minimum Normalized Cut value is monotonically decreasing with respect to the number of clusters. This monotonicity property characterizes these functionals and does not allow for automatic selection of the number of clusters, which is a trivial solution separating each data, or collecting all of them together. This is in contrast with other model-based approaches, like Gaussian Mixture Models, where the number of clusters can be learned from the data (Figueiredo & Jain, 2002), or from Correlation Clustering (Bansal et al., 2004), where the functional is not monotone in the number of clusters.

Our main purpose here is to prove strict monotonicity, for both functionals, not just on the minimizers but on the entire clustering lattice, which is the algebraic structure of a set partitions (Meila, 2005). E.g., we prove that the split operation, which divides a cluster into smaller different ones, strictly decreases the Sum of Squares Error, or strictly increases the Normalized Cut. In general a *chain* of clusterings, the typical output of hierarchical clustering algorithms, leads to a strictly monotone function. As a further relevant result, we obtain data-dependent bounds on the change of the functional value for any pair of clusterings.

These results furnish further evidences that for these two functionals it is not appropriate to choose the number of clusters by using solely the functional value. This fact is the main result of (Nagai, 2007) for the Normalized Cut and it is largely recognized, see (Jain, 2010), for the minimum Sum of Squares Error. How-

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

ever, from a more general point of view, these results can be used as a base of reference for developing clustering techniques which allow the user to explore the clustering lattice.

The reason for dealing, at the same time, with these two different functionals is the unified treatment with Linear Algebra tools which allow for clean, direct proofs and provide data-dependent rates. By proving strict anti-monotonicity of Normalized Cut we improve a result of (Nagai, 2007) with a short proof.

1.1. Related Literature

In general, the problems of minimizing the Sum of Squares Error and the Normalized Cut are NP-complete (Aloise et al., 2009; Drineas et al., 2004; Shi & Malik, 2000). Minimization of Sum of Squares Error functional is usually approximated, in practice, by the K-means algorithm (Ding & He, 2004; Meila, 2006; Jain, 2010). On the other side, the Normalized Cut is approximated using spectral techniques (von Luxburg, 2007).

A main tool in our proofs is eigenvalue interlacing. This was used in (Bolla, 1993; Bollobás & Nikiforov, 2004; 2008) to prove lower and upper bounds on the Ratio Cut, a functional which has the same principle as Normalized Cut, and in (Zha et al., 2001; Steinley, 2011) for Sum of Squares Error. From this point of view our results can be considered as generalizations of theirs.

The rest of the paper is structured as follows. In section 2 we present the necessary preliminaries of Linear Algebra, in section 3 we prove Sum of Squares Error is, under certain conditions, strictly monotone in the clustering lattice and in in section 4 we prove Normalized Cut is strictly anti-monotone in the clustering lattice.

2. Definitions and Preliminaries of Linear Algebra 1

In a typical clustering problem we have n input objects that we want to divide into $K < n$ clusters, where K is a user-chosen parameter. To do so we usually minimize a functional which represent our notion of cluster. The resulting *clustering* $\mathcal{C} : \{1, \dots, n\} \to \{1, \dots, K\}$ is a surjective function which assigns a cluster to each of these data. Alternatively, we write $\mathcal{C} = \{C_1, \dots, C_K\}$, with $C_i = \mathcal{C}^{-1}(i)$. Let $\mathcal{C}$ be a clustering into K clusters and $\mathcal{C}'$ a clustering into $K' \geq K$ clusters. We say the $\mathcal{C}'$ is a *refinement* of $\mathcal{C}$ if for each pair $i, j \in \{1, \dots, n\}$ we have $\mathcal{C}'(i) = \mathcal{C}'(j) \Rightarrow \mathcal{C}(i) = \mathcal{C}(j)$. In that case we write $\mathcal{C}' \subseteq \mathcal{C}$. Equivalently, we say $\mathcal{C}$ is a *coarsening* of $\mathcal{C}'$. If $K' > K$ then the relation is *strict*. Transitivity, $\mathcal{C}'' \subset \mathcal{C}'$ and $\mathcal{C}' \subset \mathcal{C}$ implies $\mathcal{C}'' \subset \mathcal{C}$, clearly holds. The *join* $\mathcal{C} \vee \mathcal{C}'$ is the clustering with the maximum number of clusters such that $\mathcal{C} \subseteq \mathcal{C} \vee \mathcal{C}'$ and $\mathcal{C}' \subseteq \mathcal{C} \vee \mathcal{C}'$. The *meet* $\mathcal{C} \wedge \mathcal{C}'$ is the clustering with the minimum number of clusters such that $\mathcal{C} \wedge \mathcal{C}' \subseteq \mathcal{C}$ and $\mathcal{C} \wedge \mathcal{C}' \subseteq \mathcal{C}'$. For every pair $\mathcal{C}$ and $\mathcal{C}'$ both join and meet exist. Given these properties the pair $(C, \subseteq)$ is called *lattice*. This lattice has at least two elements, $\perp$ and $\top$, and it holds $\perp \subseteq \mathcal{C} \subseteq \top$. Thus $\perp$ is the clustering with $K = n$ and $\top$ is the clustering with $K = 1$. A chain of clusterings is an ordered refinement of clusterings, for example on the set $\{a, b, c\}$ we may have the following chain: $\perp = \{\{a\}, \{b\}, \{c\}\} \subset \{\{a, b\}, \{c\}\} \subset \{\{a, b, c\}\} = \top$. We refer to (Meila, 2005) for comparing elements of a clustering lattice.

We say that a clustering functional $f : K^n \to \mathcal{R}^+$ is strictly monotone with respect to the lattice of clusterings if $\mathcal{C}' \subset \mathcal{C} \Rightarrow f(\mathcal{C}) < f(C^{'})$, or strictly anti-monotone if $\mathcal{C}' \subset \mathcal{C} \Rightarrow f(\mathcal{C}) > f(C^{'})$.

Given a matrix $\mathbf{M} \in \mathcal{R}^{n,m}$, with $m \leq n$, we denote its singular values with $\sigma(\mathbf{M}) = \sigma_1(\mathbf{M}) \geq \cdots \geq \sigma_m(\mathbf{M})$. If $\mathbf{A}$ is square and symmetric its eigenvalues are similarly represented as $\lambda(\mathbf{A}) = \lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$. It holds $\lambda(\mathbf{M}\mathbf{M}^t) = \sigma_i^2(\mathbf{M})$ for $i = 1, \dots, m$. The trace of a matrix product $\mathrm{Tr}(\mathbf{AB}) = \sum_{i,j} \mathbf{A}_{i,j}\mathbf{B}_{i,j}$.

Our main tool in proofs is the interlacing theorem. We say that a vector of numbers $\mu_1 \geq \cdots \geq \mu_m$ interlace $\lambda_1 \geq \cdots \geq \lambda_n$, with $n > m$, if

$$\lambda_i \geq \mu_i \geq \lambda_{n-m+i},\ i = 1, \dots, m.$$

Theorem 1. *(Haemers, 1995) Let $\mathbf{S} \in \mathcal{R}^{n,m}$ be a real matrix such that $\mathbf{S}^t\mathbf{S} = I$ and let $\mathbf{A} \in \mathcal{R}^{n,n}$ be a symmetric matrix with eigenvalues $\lambda_1 \geq \dots, \geq \lambda_n$. Define $\mathbf{B} = \mathbf{S}^t\mathbf{A}\mathbf{S}$ and let $\mathbf{B}$ have eigenvalues $\mu_1 \geq \cdots \geq \mu_m$ and respective eigenvectors $v_1, \dots, v_m$.*

i) *The eigenvalues of $\mathbf{B}$ interlace those of $\mathbf{A}$.*

ii) *If $\mu_i = \lambda_i$ or $\mu_i = \lambda_{n-m+i}$ for some $i \in [1, m]$, then $\mathbf{B}$ has a μ_i-eigenvector v such that $\mathbf{S}v$ is a μ_i-eigenvector of $\mathbf{A}$.*

iii) *If for some integer l, $\mu_i = \lambda_i$ for $i = 1, \dots, l$ (or $\mu_i = \lambda_{n-m+i}$ for $i = l, \dots, m$), then $\mathbf{S}v_i$ is a μ_i-eigenvector of $\mathbf{A}$ for $i = 1, \dots, l$ (respectively $i = l, \dots, m$).*

iv) *If the interlacing is tight then $\mathbf{SB} = \mathbf{AS}$.*

3. Sum of Squares Error

We suppose we are given as input a set of n distinct d-dimensional points and a target number of clusters $K \in 1, \ldots, n$. The input points are stacked as rows of a matrix $\mathbf{X} \in \mathcal{R}^{n,d}$. Given a subset C_r of points, $\mathbf{X_{C_r}}$ is the sub-matrix of $\mathbf{X}$ with the points belonging to C_r.

We define the *centroid* of a cluster C_r as the mean of its points, that is

$$\mu_r = \frac{1}{|C_r|} \sum_{j \in C_r} x_j$$

The Sum of Squares Error evaluates the sum of squared euclidean distances of input points from their own centroids:

$$\mathrm{SSE}(\mathbf{X}, \mathcal{C}) = \sum_{r=1}^{K} \sum_{j \in C_r} \|x_j - \mu_r\|^2 \tag{1}$$

When matrix $\mathbf{X}$ is clear from the context we simply write $\mathrm{SSE}(\mathcal{C})$. By minimizing the Sum of Squares Error we seek for the best set of K prototypes representing the data. It is possible that choosing a correct K is part of the problem but it is well known that we cannot rely only on the minimum value of $\mathrm{SSE}(.,.)$ itself because it is strictly monotone and an optimal clustering is the one putting each point by itself.

Observation 2 (Strict Monotonicity of Minimum Sum of Squares Error Clustering)**.** *Consider an input dataset $\mathbf{X} \in \mathcal{R}^{n,d}$ made of n distinct d-dimensional points. Then if $K < n$*

$$\min_{\{\mathcal{C} | |\mathcal{C}| > K\}} \mathrm{SSE}(\mathcal{C}) < \min_{\{\mathcal{C} | |\mathcal{C}| = K\}} \mathrm{SSE}(\mathcal{C}) \tag{2}$$

This observation is simple and provable in many different ways. Here we refer to our result of theorem 9.

Proof of Observation 2. Let $\mathcal{C}^*$ be the K-clusters minimizer. We can create a new cluster $\mathcal{C}'$ with $K+1$ clusters by choosing a cluster C_i from $\mathcal{C}$ with more than one point and putting in the new cluster just a point of C_i with higher distance from its centroid. By theorem 9 we obtain $\mathrm{SSE}(\mathcal{C}') < \mathrm{SSE}(\mathcal{C}^*)$. □

Our purpose is extending observation 2 with the additional hypothesis that the solutions are related by the refinement relation of the lattice. In order to appreciate the impact of this hypothesis we should consider that the minimizers of (1), with different values of K, can be substantially different within each other, i.e. the minimizer with $K+1$ clusters may not be a refinement of the minimizer with K clusters. A simple example of this fact is that of figure 1, a one-dimensional set of equally spaced points. The solutions here are equally-distributed partitions and clearly one solution with K' clusters refines one with K clusters only if K divides K'.

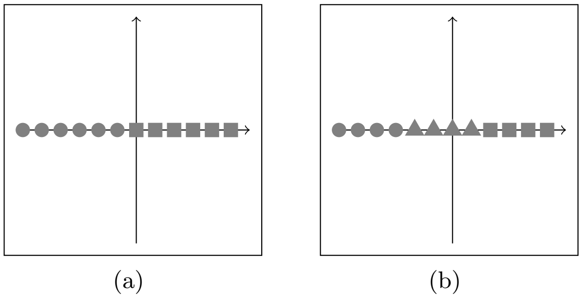
(a) (b)

Figure 1. Examples of datasets which admit non-proper clusterings.

It is central to our treatment to rewrite the Sum of Squares Error as the trace of a matrix product. To do so we use the following square indicator matrix of clusterings (Ding & He, 2004; Meila, 2006):

$$\mathbf{H}_{\mathcal{C}}(i,j) = \begin{cases} \dfrac{1}{|C_r|} & \text{if } i \in C_r \wedge j \in C_r \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Matrices $\mathbf{H}_\perp$ and $\mathbf{H}_\top$ represent, respectively, the clustering where every point is put by itself and the clustering where all points are put together. Their dimensions depend on the context. The following lemma summarizes the relation between the Sum of Squares Error and indicator matrices,

Lemma 3. *Let $\mathbf{X} \in R^{n,d}$ be an input set of n distinct d-dimensional points and let $\mathcal{C} = \{C_1, \ldots, C_K\}$ be a clustering.*

i) *The following equalities hold.*

$$\begin{aligned} \mathrm{SSE}(\mathcal{C}) &= \sum_{r=1}^{K} \mathrm{SSE}(\mathbf{X_{C_r}}, \mathbf{H}_\top) \\ &= \sum_{r=1}^{K} \mathrm{Tr}(\mathbf{X_{C_r}} \mathbf{X_{C_r}}{}^t) \\ &\quad - \mathrm{Tr}(\mathbf{H}_\top \mathbf{X_{C_r}} \mathbf{X_{C_r}}{}^t \mathbf{H}_\top) \\ &= \mathrm{Tr}(\mathbf{X}\mathbf{X}^t) - \mathrm{Tr}(\mathbf{H}_{\mathcal{C}} \mathbf{X}\mathbf{X}^t \mathbf{H}_{\mathcal{C}}) \\ &= \sum_{i=1}^{n} \lambda_i(\mathbf{X}\mathbf{X}^t) - \sum_{j=1}^{K} \lambda_j(\mathbf{H}_{\mathcal{C}} \mathbf{X}\mathbf{X}^t \mathbf{H}_{\mathcal{C}}). \end{aligned}$$

ii) *Matrix* $\mathbf{H}_\mathcal{C}\mathbf{X}\mathbf{X}^t\mathbf{H}_\mathcal{C}$ *has* K*, or less, strictly positive eigenvalues and these eigenvalues interlace those of* $\mathbf{X}\mathbf{X}^t$*.*

Proof of Lemma 3. **i)** The second equality is given by the well known Huygens theorem (Aloise & Hansen, 2009). The third and fourth equalities come from basic properties of matrix trace and eigenvalues.

ii) Define the indicator matrix

$$\mathbf{Q}_\mathcal{C}(i,r) = \begin{cases} \dfrac{1}{\sqrt{|C_r|}} & \text{if } i \in C_r \\ 0 & \text{otherwise.} \end{cases}$$

Since ${\mathbf{Q}_\mathcal{C}}^t\mathbf{Q}_\mathcal{C} = I$ by theorem 1 $\lambda({\mathbf{Q}_\mathcal{C}}^t\mathbf{X}\mathbf{X}^t\mathbf{Q}_\mathcal{C})$ interlace $\lambda(\mathbf{X}\mathbf{X}^t)$, but the greatest eigenvalues are the same as of $\lambda(\mathbf{Q}_\mathcal{C}{\mathbf{Q}_\mathcal{C}}^t\mathbf{X}\mathbf{X}^t\mathbf{Q}_\mathcal{C}{\mathbf{Q}_\mathcal{C}}^t)$ and $\mathbf{H}_\mathcal{C} = \mathbf{Q}_\mathcal{C}{\mathbf{Q}_\mathcal{C}}^t$.

□

Intuitively speaking, if a clustering $\mathcal{C}$ is coarser than $\mathcal{C}'$ then it contains less information than $\mathcal{C}'$, but the contained information is consistent with the one of $\mathcal{C}'$. Next lemma shows how this fact reflects into indicator matrices.

Lemma 4. *Let* $\mathcal{C}$ *be a clustering with* $K < n$ *clusters and* $\mathcal{C}'$ *a refinement of* $\mathcal{C}$*. Then* $\mathbf{H}_\mathcal{C}\mathbf{H}_{\mathcal{C}'} = \mathbf{H}_\mathcal{C} = \mathbf{H}_{\mathcal{C}'}\mathbf{H}_\mathcal{C}$*.*

Proof of Lemma 4. Suppose $i \in C_s$ and $j \in C'_r$, then

$$\begin{aligned} \mathbf{H}_\mathcal{C}\mathbf{H}_{\mathcal{C}'}(i,j) &= \sum_k \mathbf{H}_\mathcal{C}(i,k)\mathbf{H}_{\mathcal{C}'}(j,k) \\ &= \frac{1}{|C_s|}\sum_{k\in C_s}\mathbf{H}_{\mathcal{C}'}(j,k) \\ &= \frac{1}{|C_s||C'_r|}\sum_{k\in C_s}\mathbb{1}_{[k\in C'_r]}. \end{aligned}$$

So if $C'_r \subseteq C_s$ the last term is equal to $1/|C_s|$ and otherwise, since $C'_r \cap C_s = \emptyset$, it is zero. Since $\mathbf{H}_\mathcal{C}$ is symmetric we get $\mathbf{H}_\mathcal{C} = \mathbf{H}_{\mathcal{C}'}\mathbf{H}_\mathcal{C}$. □

The monotonicity result we obtain on the Sum of Squares Error does not hold in general for every dataset and every clustering, because there exist particular configurations of data which allow for not valid clusterings. The subset of clusterings we work on are characterized by the following property.

Definition 1 (Proper clustering)**.** *We say that a clustering* $\mathcal{C}$*, of a given dataset* $\mathbf{X}$*, with* $K \geq 2$ *clusters is* ***proper*** *if it has* K *different centroids* $\mu_1, \ldots, \mu_K$*.*

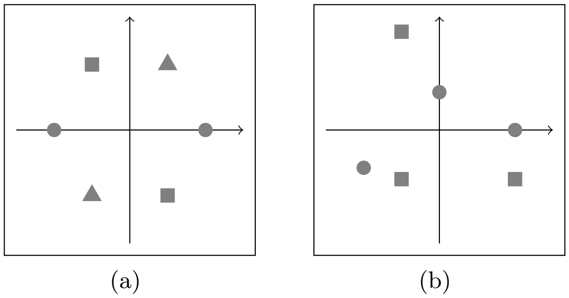

Figure 2. Examples of datasets admitting non-proper clusterings. Each class is marked with a different symbol. In both cases the origin of the axes is the centroid common to the different classes.

Strictly speaking, a clustering which is not proper has less than K clusters because at least two of them have the same centroid and can be considered the same cluster. Clustering which are not proper may arise in situations where data present symmetries, like in figure 2(a), where a clustering which pairs each point with its opposite w.r.t. the center, is non-proper. In general, also data without symmetries may admit non-proper clusterings, as in figure 2(b). On the other side, we have different arguments for considering a non-proper clustering unstable and with a marginal impact on practical applications. Firstly, a non-proper clustering can be made proper by slightly perturbing the data, whereas the opposite is unlikely. Secondly, since points are distinct, it just suffices to change the cluster of one of them to break the equality of two means and eventually obtaining, with few changes, a proper clustering out of a non-proper one.

In studying strict monotonicity for Sum of Squares Error we have to consider different cases for the dimensionality of the input set.

Definition 2 (Dimensionality)**.** *We define the dimensionality* $dim(\mathbf{X})$ *of a dataset, with* $\mathbf{X} \in \mathcal{R}^{n,d}$*, as the number of singular values, counted with multiplicities, which are different from zero. Clearly,* $dim(\mathbf{X}) \leq min(\{n,d\})$*.*

If $dim(\mathbf{X}) = n$ we can state strict monotonicity without any further conditions. If $dim(\mathbf{X}) < n$ we give a necessary and sufficient condition to have strict monotonicity. Since a non-proper dataset contains linear dependencies among points, it is easy to observe the following.

Observation 5. *Every clustering of a full dimensional dataset is proper.*

3.1. Full dimensional dataset

A necessary condition for having a full dimensional dataset $\mathbf{X}$ is $d \geq n$. Perhaps, the most frequent practical cases where this condition holds are high-dimensional datasets, characterised by very large d. In these cases it is likely that $\sigma_n(\mathbf{X}) > 0$.

Theorem 6 (Strict Monotonicity of Sum of Squares Error on $(\mathcal{C}, \subseteq)$, $dim(\mathbf{X}) = n$)**.** *Let $\mathbf{X} \in \mathcal{R}^{n,d}$ be a d-dimensional dataset of n points with $dim(\mathbf{X}) = n$. Let $\mathcal{C}$ be a clustering with $K < n$ clusters and $\mathcal{C}'$ a strict refinement with $K' > K$ clusters. Then*

$$\begin{aligned}\mathrm{SSE}(\mathcal{C}') &\leq \mathrm{SSE}(\mathcal{C}) - \sum_{i=1}^{K'-K} \sigma^2_{n+1-i}(\mathbf{X}) \\ &\vee\!| \\ \mathrm{SSE}(\mathcal{C}) &- \sum_{i=1}^{K'-K} \sigma^2_i(\mathbf{X}),\end{aligned}$$

in particular $\mathrm{SSE}(\mathcal{C}') < \mathrm{SSE}(\mathcal{C})$.

Proof of Theorem 6.

$$\begin{aligned}& \mathrm{SSE}(\mathcal{C}') - \mathrm{SSE}(\mathcal{C}) \\ = & \sum_{j=1}^{K} \lambda_j(\mathbf{H}_{\mathcal{C}}\mathbf{X}\mathbf{X}^t\mathbf{H}_{\mathcal{C}}) - \sum_{i=1}^{K'} \lambda_i(\mathbf{H}_{\mathcal{C}'}\mathbf{X}\mathbf{X}^t\mathbf{H}_{\mathcal{C}'}) \\ = & \sum_{j=1}^{K} \left(\lambda_j(\mathbf{H}_{\mathcal{C}}\mathbf{X}\mathbf{X}^t\mathbf{H}_{\mathcal{C}}) - \lambda_j(\mathbf{H}_{\mathcal{C}'}\mathbf{X}\mathbf{X}^t\mathbf{H}_{\mathcal{C}'})\right) \\ & - \sum_{j=K+1}^{K'} \lambda_i(\mathbf{H}_{\mathcal{C}'}\mathbf{X}\mathbf{X}^t\mathbf{H}_{\mathcal{C}'}) \\ \leq & - \sum_{i=1}^{K'-K} \sigma^2_{n+1-i}(\mathbf{X}) < 0.\end{aligned}$$

In the last step we used the fact, entailed by theorem 3, that the eigenvalues of $\mathbf{H}_{\mathcal{C}}\mathbf{H}_{\mathcal{C}'}\mathbf{X}\mathbf{X}^t\mathbf{H}_{\mathcal{C}'}\mathbf{H}_{\mathcal{C}}$ interlace those of $\mathbf{H}_{\mathcal{C}'}\mathbf{X}\mathbf{X}^t\mathbf{H}_{\mathcal{C}'}$, but since lemma 4 implies

$$\mathbf{H}_{\mathcal{C}}\mathbf{H}_{\mathcal{C}'}\mathbf{X}\mathbf{X}^t\mathbf{H}_{\mathcal{C}'}\mathbf{H}_{\mathcal{C}} = \mathbf{H}_{\mathcal{C}}\mathbf{X}\mathbf{X}^t\mathbf{H}_{\mathcal{C}}$$

the first sum is maximized by zero. The second sum, again because of interlacing, cannot be greater than the negative of the sum of the $K' - K$ smallest eigenvalues of $\lambda(\mathbf{X}\mathbf{X}^t)$. Similarly we get the other inequality. □

By using theorem 6 we obtain two different ways for lower bounding, and upper bounding, the quantity $\mathrm{SSE}(\mathcal{D}) - \mathrm{SSE}(\mathcal{C})$ between any two clusterings $\mathcal{C}, \mathcal{D}$.

Theorem 7. *Let $\mathbf{X} \in \mathcal{R}^{n,d}$ be a d-dimensional dataset of n points with $dim(\mathbf{X}) = n$. Let $\mathcal{C}$ be a clustering with K_1 clusters and $\mathcal{D}$ a clustering with K_2 clusters. Let $n_1 = K_1 - |\mathcal{C} \wedge \mathcal{D}|$ and $n_2 = K_2 - |\mathcal{C} \wedge \mathcal{D}|$. Then*

$$\begin{aligned}\mathrm{SSE}(\mathcal{D}) - \mathrm{SSE}(\mathcal{C}) &\leq \sum_{i=1}^{n_1} \sigma^2_i(\mathbf{X}) - \sum_{i=1}^{n_2} \sigma^2_{n+1-i}(\mathbf{X}) \\ &\vee\!| \\ \sum_{i=1}^{n_1} \sigma^2_{n+1-i}(\mathbf{X}) &- \sum_{i=1}^{n_2} \sigma^2_i(\mathbf{X}).\end{aligned} \tag{4}$$

Proof. Using theorem 6 we get

$$\begin{aligned}\mathrm{SSE}(\mathcal{D}) &\leq \mathrm{SSE}(\mathcal{C} \wedge \mathcal{D}) - \sum_{i=1}^{n_2} \sigma^2_{n+1-i}(\mathbf{X}) \\ &\vee\!| \\ \mathrm{SSE}(\mathcal{C} \wedge \mathcal{D}) &- \sum_{i=1}^{n_2} \sigma^2_i(\mathbf{X}),\end{aligned} \tag{5}$$

and

$$\begin{aligned}-\mathrm{SSE}(\mathcal{C}) &\leq -\mathrm{SSE}(\mathcal{C} \wedge \mathcal{D}) + \sum_{i=1}^{n_1} \sigma^2_i(\mathbf{X}) \\ &\vee\!| \\ -\mathrm{SSE}(\mathcal{C} \wedge \mathcal{D}) &+ \sum_{i=1}^{n_1} \sigma^2_{n+1-i}(\mathbf{X}),\end{aligned} \tag{6}$$

by adding (5) and (6) we get (4). □

Theorem 8. *Let $\mathbf{X} \in \mathcal{R}^{n,d}$ be a d-dimensional dataset of n points with $dim(\mathbf{X}) = n$. Let $\mathcal{C}$ be a clustering with K_1 clusters and $\mathcal{D}$ a clustering with K_2 clusters. Let $m_1 = |\mathcal{C} \vee \mathcal{D}| - K_1$ and $m_2 = |\mathcal{C} \vee \mathcal{D}| - K_2$. Then*

$$\begin{aligned}\mathrm{SSE}(\mathcal{D}) - \mathrm{SSE}(\mathcal{C}) &\leq \sum_{i=1}^{m_2} \sigma^2_i(\mathbf{X}) - \sum_{i=1}^{m_1} \sigma^2_{n+1-i}(\mathbf{X}) \\ &\vee\!| \\ \sum_{i=1}^{m_2} \sigma^2_{n+1-i}(\mathbf{X}) &- \sum_{i=1}^{m_1} \sigma^2_i(\mathbf{X}).\end{aligned} \tag{7}$$

Proof. From theorem 6 we have

$$\begin{aligned} \mathrm{SSE}(\mathcal{C} \vee \mathcal{D}) \leq \mathrm{SSE}(\mathcal{C}) - \sum_{i=1}^{m_1} \sigma^2_{n+1-i}(\mathbf{X}) \\ \vee\!\mid \\ \mathrm{SSE}(\mathcal{C}) - \sum_{i=1}^{m_1} \sigma^2_i(\mathbf{X}) \end{aligned}$$

and the same with $\mathcal{D}$ instead of $\mathcal{C}$. Thus we get

$$\begin{aligned} \mathrm{SSE}(\mathcal{C}) - \sum_{i=1}^{m_1} \sigma^2_{n+1-i}(\mathbf{X}) \\ \vee\!\mid \\ \mathrm{SSE}(\mathcal{D}) - \sum_{i=1}^{m_2} \sigma^2_i(\mathbf{X}), \end{aligned}$$

and

$$\begin{aligned} \mathrm{SSE}(\mathcal{D}) - \sum_{i=1}^{m_2} \sigma^2_{n+1-i}(\mathbf{X}) \\ \vee\!\mid \\ \mathrm{SSE}(\mathcal{C}) - \sum_{i=1}^{m_1} \sigma^2_i(\mathbf{X}), \end{aligned}$$

from which we obtain the two sides of inequality (7). □

3.2. Non full dimensional dataset

We treat separately the low dimensional case because if $\sigma_{dim(\mathbf{X})+1} = \cdots = \sigma_n = 0$ then theorem 6 does not necessarily imply strict monotonicity.

Theorem 9 (Strict Monotonicity of Sum of Squares Error, $d < n$)**.** *Let $\mathcal{C}$ be a clustering into $K < n$ clusters and $\mathcal{C}'$ one of its refinements with $K+1$ clusters. If $\mathcal{C}'$ is proper, as in definition 1, then $\mathrm{SSE}(\mathcal{C}') < \mathrm{SSE}(\mathcal{C})$.*

Proof of Theorem 9. Since $|\mathcal{C}'| = K+1$ and $\mathcal{C}' \subset \mathcal{C}$ there exists a cluster C_r such that $|C_r| \geq 2$ and a pair $C'_j \in \mathcal{C}', C'_k \in \mathcal{C}'$ with $C'_j \cup C'_k = C_r$.

Let $\mathbf{X_r} = \mathbf{X_{C_r}}$. By using lemma 3 we have $\mathrm{SSE}(\mathcal{C}) - \mathrm{SSE}(\mathcal{C}') = \mathrm{T{\scriptsize R}}(\mathbf{H_{C'_j \uplus C'_k} X_r X_r}^t \mathbf{H_{C'_j \uplus C'_k}}) - \mathrm{T{\scriptsize R}}(\mathbf{H_r X_r X_r}^t \mathbf{H_r})$, $\lambda(\mathbf{H_r X_r X_r}^t \mathbf{H_r}) = \tau_1 \geq \tau_2 = 0$ and $\lambda(\mathbf{H_{C'_j \uplus C'_k} X_r X_r}^t \mathbf{H_{C'_j \uplus C'_k}}) = \lambda_1 \geq \lambda_2 \geq \lambda_3 = 0$.

Since

$$\mathbf{H_r H_{C'_j \uplus C'_k} X_r X_r}^t \mathbf{H_{C'_j \uplus C'_k} H_r} = \mathbf{H_r X_r X_r}^t \mathbf{H_r}$$

The eigenvalues interlace and we have $\lambda_1 \geq \tau_1 \geq \lambda_2 \cdots \geq 0$. If $\tau_1 = 0$ then the set $\mathbf{X_r}$ has mean zero, but since $\mathcal{C}'$ is proper at least one mean of $\mathbf{X_j}$ and $\mathbf{X_k}$ is different from zero, giving $\lambda_1 > 0$, and the conclusion follows.

Suppose $\tau_1 > 0$. Then, unless $\lambda_1 = \tau_1$ and $\lambda_2 = 0$, $\lambda_1 + \lambda_2 > \tau_1$. Suppose, by contradiction, that $\lambda_1 = \tau_1$ and $\lambda_2 = 0$, we have by lemma 3 and theorem 1

$$\frac{1}{|C_r|^2} \mathbf{E X_r X_r}^t \mathbf{E} = \mathbf{H_{C'_j \uplus C'_k} X_r X_r}^t \mathbf{H_{C'_j \uplus C'_k}}$$

so that $\mu_{C'_j} = \mu_r = \mu_{C'_k}$ which cannot happen since $\mathcal{C}'$ is proper. By lemma 3 we have $\mathrm{SSE}(\mathcal{C}') < \mathrm{SSE}(\mathcal{C})$. □

Finally we have the following corollary.

Corollary 10. *Any optimal K clusters solution to the minimization of Sum of Squares Error is proper.*

4. Normalized Cut and Ratio Cut

Let $G = (V, E)$ be an undirected, weighted and connected graph. Given a clustering $\mathcal{C}_K = \{C_1, C_2, \ldots, C_K\}$ let

$$\mathrm{cut}(C_i, C_j) = \sum_{r \in C_i} \sum_{s \in C_j} w_{r,s}.$$

The degree d_r of a vertex r is $\mathrm{cut}(r, V)$ and the volume $\mathrm{vol}(C_i) = \mathrm{cut}(C_i, V)$. The Normalized Cut (Chung, 1997; Shi & Malik, 2000) is defined as

$$\mathrm{NCUT}(\mathcal{C}) = \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \frac{\mathrm{cut}(C_i, C_j)}{\mathrm{vol}(C_j)}$$

As for the Sum of Squares Error choosing the right K is a difficult problem. Indeed, in (Nagai, 2007) the following lemma was proved.

Lemma (Nagai, 2007)**.** *Let $G = (V, E)$ be an undirected, weighted and connected graph. Let $\mathcal{C}$ be a clustering with $K < n$ clusters and $\mathcal{C}'$ a strict refinement with $K' > K$ clusters. Then*

$$\mathrm{NCUT}(\mathcal{C}') \geq \mathrm{NCUT}(\mathcal{C}).$$

Here we first prove strict anti-monotonicity on the clustering lattice, so if $\mathcal{C}' \subset \mathcal{C}$ we obtain $\mathrm{NCUT}(\mathcal{C}') > \mathrm{NCUT}(\mathcal{C})$. This result turns out to imply strict anti-monotonicity also for the lemma of (Nagai, 2007), improving it. Then we obtain data-dependent bounds for the difference of the Normalized Cut of any two clusterings.

Let $\mathbf{d}_i = d_i$, $\mathbf{W}$ be the adjacency matrix of the graph and $\mathbf{D} = \operatorname{diag}(\mathbf{d})$. The normalized laplacian is the matrix $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$. If the graph is connected then $\lambda_n(\mathbf{L}) = 0$ and $\lambda_{n-1}(\mathbf{L}) > 0$ (Chung, 1997). The vector $\mathbf{d}/\operatorname{vol}(G)$ is the eigenvector of $\lambda_n(\mathbf{L})$.

The following matrix is an indicator matrix of partitions:

$$\mathbf{G}_{\mathcal{C}}(i,j) = \begin{cases} \dfrac{\sqrt{d_i d_j}}{\operatorname{vol}(C_r)} & \text{if } i \in C_r \wedge j \in C_r \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

The following lemma summarizes the relation between the Normalized Cut and indicator matrices.

Lemma 11. *Let $G = (V, E)$ be an undirected, weighted and connected graph and $\mathbf{L}$ its normalized laplacian. Let $\mathcal{C} = \{C_1, \dots, C_K\}$ be a clustering. Then*

i) $\operatorname{NCUT}(\mathcal{C}) = \operatorname{Tr}(\mathbf{G}_{\mathcal{C}}\mathbf{L}\mathbf{G}_{\mathcal{C}})$

ii) *Matrix $\mathbf{G}_{\mathcal{C}}\mathbf{L}\mathbf{G}_{\mathcal{C}}$ has $K-1$, or less, strictly positive eigenvalues and these eigenvalues interlace those of $\mathbf{L}$.*

Proof of Lemma 11. **i)** See (von Luxburg, 2007).

ii) As in lemma 3 **ii)**, but with the following indicator matrix

$$\mathbf{P}_{\mathcal{C}}(i,r) = \begin{cases} \sqrt{\dfrac{d_i}{\operatorname{vol}(C_r)}} & \text{if } i \in C_r \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

and by observing that for every $\mathcal{C}$, the null eigenvector $\mathbf{d}/\operatorname{vol}(G)$ is in the span of the columns of $\mathbf{P}_{\mathcal{C}}$. □

Lemma 12 (Lemma)**.** *Let $\mathcal{C}$ be a clustering with $K \leq n-1$ clusters and $\mathcal{C}'$ a refinement of $\mathcal{C}$. Then $\mathbf{G}_{\mathcal{C}}\mathbf{G}_{\mathcal{C}'} = \mathbf{G}_{\mathcal{C}} = \mathbf{G}_{\mathcal{C}'}\mathbf{G}_{\mathcal{C}}$.*

Proof of Lemma 12. Suppose object $i \in C_s$ and $j \in C'_r$.

$$\begin{aligned} \mathbf{G}_{\mathcal{C}}\mathbf{G}_{\mathcal{C}'}(i,j) &= \sum_k G_{\mathcal{C}}(i,k)\mathbf{G}_{\mathcal{C}'}(j,k) \\ &= \frac{1}{\operatorname{vol}(C_s)} \sum_{k \in C_s} \sqrt{d_k d_i}\,\mathbf{G}_{\mathcal{C}}(j,k) \\ &= \frac{\sqrt{d_i d_j}}{\operatorname{vol}(C_s)\operatorname{vol}(C'_r)} \sum_{k \in C_s} d_k \mathbb{1}_{[k \in C'_r]} \end{aligned}$$

So if $i \in C_r$ then $C_s \subseteq C'_r$ and the last term is equal to $\sqrt{d_i d_j}/\operatorname{vol}(C_s)$ and it is zero otherwise. Since $\mathbf{G}_{\mathcal{C}}$ is symmetric we get $\mathbf{G}_{\mathcal{C}} = \mathbf{G}_{\mathcal{C}'}\mathbf{G}_{\mathcal{C}}$. □

In section 3 we proved strict monotonicity for the Sum of Squares Error in two different conditions, whether the input dataset $\mathbf{X}$ was full dimensional or not. Here we obtain strict anti-monotonicity using basically the same techniques. However the graph is connected and we know $\lambda_{n-1}(\mathbf{L}) > 0$. So we work with dimensionality $n-1$. We have the following.

Theorem 13 (Strict Anti-Monotonicity of Normalized Cut on $(\mathcal{C}, \subseteq)$)**.** *Let $G = (V, E)$ be an undirected, weighted and connected graph and $\mathbf{L}$ its normalized laplacian. Let $\mathcal{C}$ be a clustering with $2 \leq K < n$ clusters and $\mathcal{C}'$ a strict refinement with $K' > K$ clusters. Then*

$$\begin{aligned} &\operatorname{NCUT}(\mathcal{C}') \geq \operatorname{NCUT}(\mathcal{C}) + \sum_{i=2}^{K'-K+1} \lambda_{n+1-i}(\mathbf{L}) \\ &\wedge\!| \\ &\operatorname{NCUT}(\mathcal{C}) + \sum_{i=1}^{K'-K} \lambda_i(\mathbf{L}), \end{aligned} \tag{9}$$

in particular $\operatorname{NCUT}(\mathcal{C}') > \operatorname{NCUT}(\mathcal{C})$.

Proof of Theorem 13. The proof is similar to that of theorem 6. The only difference being that matrix $\mathbf{G}_{\mathcal{C}}\mathbf{L}\mathbf{G}_{\mathcal{C}}$ has at maximum $K-1$ eigenvalues greater than zero, instead of K. Similarly as before we have, by lemma 11, that the eigenvalues of $\mathbf{G}_{\mathcal{C}}\mathbf{G}_{\mathcal{C}'}\mathbf{L}\mathbf{G}_{\mathcal{C}'}\mathbf{G}_{\mathcal{C}}$ interlace those of $\mathbf{G}_{\mathcal{C}'}\mathbf{L}\mathbf{G}_{\mathcal{C}'}$, and lemma 12 implies $\mathbf{G}_{\mathcal{C}}\mathbf{G}_{\mathcal{C}'}\mathbf{L}\mathbf{G}_{\mathcal{C}'}\mathbf{G}_{\mathcal{C}} = \mathbf{G}_{\mathcal{C}}\mathbf{L}\mathbf{G}_{\mathcal{C}}$.

Now

$$\begin{aligned} &\operatorname{NCUT}(\mathcal{C}') - \operatorname{NCUT}(\mathcal{C}) \\ = &\sum_{i=1}^{K'-1} \lambda_i(\mathbf{G}_{\mathcal{C}'}\mathbf{L}\mathbf{G}_{\mathcal{C}'}) - \sum_{j=1}^{K-1} \lambda_j(\mathbf{G}_{\mathcal{C}}\mathbf{L}\mathbf{G}_{\mathcal{C}}) \\ = &\sum_{j=1}^{K-1} (\lambda_j(\mathbf{G}_{\mathcal{C}'}\mathbf{L}\mathbf{G}_{\mathcal{C}'}) - \lambda_j(\mathbf{G}_{\mathcal{C}}\mathbf{L}\mathbf{G}_{\mathcal{C}})) \\ &+ \sum_{j=K}^{K'-1} \lambda_i(\mathbf{H}_{\mathcal{C}'}\mathbf{X}\mathbf{X}^t\mathbf{H}_{\mathcal{C}'}) \\ \geq &\sum_{i=2}^{K'-K+1} \lambda_{n+1-i}(\mathbf{L}) > 0. \end{aligned}$$

In the last step we used the fact that the first sum is greater or equal than zero. The second sum, again because of interlacing, cannot be greater than the sum of the $K'-K+1$ smallest eigenvalues of $\lambda(\mathbf{L})$, excluding the last one which is zero. Similarly we get the other inequality. □

Using theorem 13 we get the following corollary which improves the result of (Nagai, 2007) with *strict* monotonicity.

Corollary 14 (Strict Anti-Monotonicity of Normalized Cut)**.** *Let $G=(V,E)$ be an undirected, weighted and connected graph and* $\mathbf{L}$ *its normalized laplacian. We have*

$$\min_{\{\mathcal{C}||\mathcal{C}|<K\}} \mathrm{NCUT}(\mathcal{C}) < \min_{\{\mathcal{C}||\mathcal{C}|=K\}} \mathrm{NCUT}(\mathcal{C}). \tag{10}$$

Proof of theorem 14. Let $\mathcal{C}^*$ be the minimizer with K clusters. By theorem 13 a coarsening $\mathcal{C}'$ will have $\mathrm{NCUT}(\mathcal{C}') < \mathrm{NCUT}(\mathcal{C}^*)$ and the conclusion follows. □

Similarly as in the Sum of Squares Error, theorems 7 and 8, we can bound the difference $\mathrm{NCUT}(\mathcal{D}) - \mathrm{NCUT}(\mathcal{C})$ between two clusterings $\mathcal{C}$ and $\mathcal{D}$ in two different ways. The proofs are omitted here, but can be easily derived by the proofs of theorems 7 and 8 using the indicator matrix $\mathbf{G}_{\mathcal{C}}$ instead of $\mathbf{H}_{\mathcal{C}}$, $\mathbf{L}$ instead of $\mathbf{XX}^t$ and invoking theorem 13 instead of theorem 6.

Theorem 15. *Let $G=(V,E)$ be an undirected, weighted and connected graph and* $\mathbf{L}$ *its normalized laplacian. Without loss of generality, let $\mathcal{C}$ be a clustering with $K_1<n$ clusters and $\mathcal{D}$ a clustering with $K_2 \leq K_1$ clusters. Let $n_1 = K_1 - |\mathcal{C}\wedge\mathcal{D}|$ and $n_2 = K_2 - |\mathcal{C}\wedge\mathcal{D}|$. Then*

$$\begin{gathered}\mathrm{NCUT}(\mathcal{D})\text{-}\mathrm{NCUT}(\mathcal{C}) \geq \sum_{i=1}^{n_1}\lambda_i(\mathbf{L})\text{-}\sum_{i=2}^{n_2+1}\lambda_{n+1-i}(\mathbf{L}) \\ \wedge| \\ \sum_{i=2}^{n_1+1}\lambda_{n+1-i}(\mathbf{L})\text{-}\sum_{i=1}^{n_2}\lambda_i(\mathbf{L}).\end{gathered} \tag{11}$$

Theorem 16. *Let $G=(V,E)$ be an undirected, weighted and connected graph and* $\mathbf{L}$ *its normalized laplacian.Without loss of generality, let $\mathcal{C}$ be a clustering with $K_1<n$ clusters and $\mathcal{D}$ a clustering with $K_2 \leq K_1$ clusters. Let $m_1 = |\mathcal{C}\vee\mathcal{D}| - K_1$ and $m_2 = |\mathcal{C}\vee\mathcal{D}| - K_2$. Then*

$$\begin{gathered}\mathrm{NCUT}(\mathcal{D})\text{-}\mathrm{NCUT}(\mathcal{C}) \geq \sum_{i=1}^{m_2}\lambda_i(\mathbf{L})\text{-}\sum_{i=2}^{m_1+1}\lambda_{n+1-i}(\mathbf{L}) \\ \wedge| \\ \sum_{i=2}^{m_2+1}\lambda_{n+1-i}(\mathbf{L})\text{-}\sum_{i=1}^{m_1}\lambda_i(\mathbf{L}).\end{gathered} \tag{12}$$

As a final observation, the results provided for the Normalized Cut holds also for the Ratio Cut

$$\mathrm{RCUT}(\mathcal{C}) = \sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\frac{\mathrm{cut}(C_i,C_j)}{|C_j|}$$

but using the unnormalized laplacian $\mathbf{L_u} = \mathbf{D}-\mathbf{W}$ and using the representation matrix defined for the Sum of Squares Error in section 3.

5. Discussion

In this paper we proved strict monotonicity for two clustering functionals, Sum of Squares Error and Normalized Cut, with respect to the refinement relation of the lattice of clusterings. As a consequence of these results we could get data-dependent bounds on the difference between any two clusterings and, for the minimizers of the Normalized Cut, we could improve the result of (Nagai, 2007) with strict monotonicity.

These results are interesting for model validation, i.e. choosing the right number of classes. From one side they confirm that we cannot rely only on the functional value even if we constraint our solutions to form a chain of clusterings. From the other side they give quantitative ways to estimate how much a dataset is clusterable. For example, for the Normalized Cut we see that if a graph G has a small gap between $\lambda_1(\mathbf{L}) - \lambda_{n-1}(\mathbf{L})$, like in expander graphs, all clusterings of G, with the same number of classes K, will have similar values implying that G consists of just one cluster. This direction of work deserves attention for future developments.

Acknowledgments

This work was carried out during the tenure of an ERCIM ”Alain Bensoussan” Fellowship Programme. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 246016.

References

Aloise, Daniel and Hansen, Pierre. A branch-and-cut sdp-based algorithm for minimum sum-of-squares clustering. *Pesquisa Operacional*, 29:503–516, 2009.

Aloise, Daniel, Deshpande, Amit, Hansen, Pierre, and Popat, Preyas. NP-hardness of euclidean sum-of-squares clustering. *Machine Learning*, 75(2):245–248, 2009.

Bansal, Nikhil, Blum, Avrim, and Chawla, Shuchi. Correlation clustering. *Machine Learning*, 56(1-3): 89–113, 2004.

Bolla, Marianna. Spectra, euclidean representations and clusterings of hypergraphs. *Discrete Mathematics*, 117:19–39, 1993.

Bollobás, Béla and Nikiforov, Vladimir. Graphs and hermitian matrices: eigenvalue interlacing. *Discrete Mathematics*, 289(1-3):119–127, 2004.

Bollobás, Béla and Nikiforov, Vladimir. Graphs and hermitian matrices: Exact interlacing. *Discrete Mathematics*, 308(20):4816–4821, 2008.

Chung, F. R. K. *Spectral Graph Theory.* American Mathematical Society, 1997.

Ding, Chris H. Q. and He, Xiaofeng. *K*-means clustering via principal component analysis. In Brodley, Carla E. (ed.), *ICML*, volume 69 of *ACM International Conference Proceeding Series.* ACM, 2004.

Drineas, Petros, Frieze, Alan M., Kannan, Ravi, Vempala, Santosh, and Vinay, V. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56(1-3):9–33, 2004.

Figueiredo, Mário A. T. and Jain, Anil K. Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(3):381–396, 2002.

Haemers, W.H. Interlacing eigenvalues and graphs. *Linear Algebra Applications*, 226/228:593–616, 1995.

Jain, Anil K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

Meila, Marina. Comparing clusterings: an axiomatic view. In Raedt, Luc De and Wrobel, Stefan (eds.), *ICML*, volume 119 of *ACM International Conference Proceeding Series*, pp. 577–584. ACM, 2005. ISBN 1-59593-180-5.

Meila, Marina. The uniqueness of a good optimum for k-means. In Cohen, William W. and Moore, Andrew (eds.), *ICML*, volume 148 of *ACM International Conference Proceeding Series*, pp. 625–632. ACM, 2006. ISBN 1-59593-383-2.

Nagai, Ayumu. Inappropriateness of the criterion of k-way normalized cuts for deciding the number of clusters. *Pattern Recognition Letters*, 28(15):1981–1986, 2007.

Shi, Jianbo and Malik, Jitendra. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

Steinley, D. & Brusco, M. J. Testing for validity and choosing the number of clusters in k-means clustering. *Psychological Methods*, 16:285–297, 2011.

von Luxburg, Ulrike. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

Zha, Hongyuan, He, Xiaofeng, Ding, Chris H. Q., Gu, Ming, and Simon, Horst D. Spectral relaxation for k-means clustering. In Dietterich, Thomas G., Becker, Suzanna, and Ghahramani, Zoubin (eds.), *NIPS*, pp. 1057–1064. MIT Press, 2001.