The stability of a good clustering

Marina Meilă mmp@stat.washington.edu
University of Washington, Box 345322
Seattle, WA 98195-4322

September 18, 2011

Abstract

If we have found a "good" clustering $\mathcal{C}$ of a data set, can we prove that $\mathcal{C}$ is not far from the (unknown) best clustering $\mathcal{C}^{opt}$ of these data? Perhaps surprisingly, the answer to this question is sometimes yes. This paper proves spectral bounds on the distance $d(\mathcal{C}, \mathcal{C}^{opt})$ for the case when "goodness" is measured by a quadratic cost, such as the squared distortion of K-means clustering, or the Normalized Cut criterion of spectral clustering. The bounds exist if the data admits a "good", low-cost clustering.

1 Motivation

Optimizing clustering criteria like the minimum squared error of K-means clustering or the multiway Normalized Cut of spectral clustering is theoretically NP-hard [Brucker, 1978, Shi and Malik, 2000, Meila and Xu, 2003]. Abundant empirical evidence, however, shows that if the data are well clustered, then it is easy to find a near-optimal partition. This suggests the existence of at least two regimes for this optimality problem: the "difficult" regime, characterized by the worst-case situations, and the "easy" one, characterized by the existence of a "good" clustering. To be more precise, in the "easy" regime the global minimum of the clustering criterion, e.g of the Normalized Cut, is much lower relative to its average value. Hence, the cost function has a "deep" well at the global minimum. There is no reason to believe that the "easy" regime is typical. But, even if such a case is rare, this is the case of interest for the field of data clustering. If we define clustering as the task of finding a natural partition of the data – as opposed to data quantization, which is finding the best partition in data, no matter how "bad" this is – then it is in the easy regime that the interesting cases lie. This paper shows that, when a sufficiently "good" clustering $\mathcal{C}$ exists in a dataset, then $\mathcal{C}$ is also *stable*, in the sense that any other "good" clustering is "close" to it. Thus, our paper shows that, in such a case, there is a unique and compact group of near-optimal clusterings. To our knowledge, this is the first finite sample stability result for the K-means optimization problem.

Practically, this paper will produce computable bounds on the distance $d(\mathcal{C}, \mathcal{C}^{opt})$ between a given clustering $\mathcal{C}$ and the (unknown) optimal clustering $\mathcal{C}^{opt}$ of the given data, which will be valid whenever the distortion of $\mathcal{C}$ will be small. Both the bound on the distance and the threshold defining the existence of the bound are computable given the clustering $\mathcal{C}$.

Next section, 2, introduces the terminology and notation, defining terms as "good" and "close", the K-means and the $NCut$ cost functions. Section 3 is the core of the paper, describing how to arrive from a lower bound on the distortion to an upper bound on the distance to the optimum. We present validating experiments in section 4. In the next section, 5 we extend our results to weighted data. This lets us obtain an analog bound for the Normalized Cut criterion of spectral clustering. The case of general quadratic cost is treated in section 6. The discussion in section7 concludes the paper. To keep the paper readable, most of the proofs are relegated to the appendix.

2 Definitions and representations

A *clustering* $\mathcal{C}$ of a finite dataset, assumed w.l.o.g to be $\{1, 2, \ldots, n\}$, is a partition of the dataset into disjoint, nonempty subsets called *clusters*. If the partition has K clusters, we write $\mathcal{C} = \{C_1, C_2 \ldots, C_K\}$ and denote by $n_k = |C_k|$, $\sum_k n_k = n$. If the data points have *weights* $w_i > 0, i = 1, \ldots n$, then the cluster sizes become *cluster weights*

$$W_k = \sum_{i \in C_k} w_i, \tag{1}$$

the *total weight* of the data is $W_{all} = \sum_k W_k = \sum_{i=1}^{n} w_i$. The weighted case reduces to the unweighted one for $w_i \equiv 1$.

A clustering can be represented by a $n \times K$ matrix $\tilde{X}$, whose columns represent the indicator vectors of the the K clusters.

$$\tilde{X}_{ik} = \begin{cases} 1 & \text{if } i \in C_k \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The columns of $\tilde{X}$ are mutually orthogonal vectors. We normalize these to length 1 in a way that takes into account the point weights; we obtain thus the *normalized* representation $X \in \mathbb{R}^{n \times K}$ of a clustering.

$$X_{ik} = \begin{cases} \sqrt{w_i / W_k} & \text{if } i \in C_k \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

In the case of unweighted data, i.e $w_i = 1$ for all i, the normalized representation becomes

$$X_{ik} = \begin{cases} n_k^{-1/2} & \text{if } i \in C_k \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

In the future, we will refer to a clustering by any of its matrix representations. As we'll typically work with two clusterings, one will be denoted by $\tilde{X}$, (X)

while the other by $\tilde{X}'$ (respectively X'). For example, the distance between two clusterings can be denoted equivalently by $d(\mathcal{C}, \mathcal{C}')$ or $d(X, X')$ or $d(\tilde{X}, \tilde{X}')$.

2.1 The Misclassification Error (ME) distance between clusterings

The *confusion matrix* of two clusterings $\mathcal{C} = \{C_1, C_2 \ldots, C_K\}$ and $\mathcal{C}' = \{C'_1, C_2 \ldots, C'_{K'}\}$ is defined as the $K \times K'$ matrix $M = [m_{kk'}]$ with $m_{kk'} = |C_k \cap C'_{k'}|$. It can be easily shown that $M = \tilde{X}^T \tilde{X}'$. A distance between two clusterings is typically a permutation invariant function of the confusion matrix M. For the purpose of clustering stability, it is sufficient to handle the case $K = K'$. We will make this assumption implicitly in all that follows, including the definitions of the distances. The *Misclassification Error (ME)* distance is defined as

$$d(\tilde{X}, \tilde{X}') \;=\; 1 - \frac{1}{n} \max_{\pi \in \Pi_K} \sum_k m_{k,\pi(k)} \tag{5}$$

This distance represents the well known cost of classification, minimized over all permutations of the labels $\{1, 2, \ldots, K\}$. Although the maximization is over a set of size $K!$, d can be computed in polynomial time by a maximum bipartite matching algorithm [Papadimitriou and Steiglitz, 1998]. This distance is widely used, having very appealing properties as long as X, X' are close [Meilă, 2005].

For weighted data, the weighted confusion matrix is $M^{\mathbf{w}} = [m^{\mathbf{w}}_{kk'}]$ with $m^{\mathbf{w}}_{kk'} = \sum_{i \in C_k \cap C'_{k'}} w_i$. In matrix form we have

$$M^{\mathbf{w}} \;=\; \tilde{X}^T \mathrm{diag}(\mathbf{w}) \tilde{X} \tag{6}$$

and the weighted misclassification error is written as

$$d^{\mathbf{w}}(\tilde{X}, \tilde{X}') \;=\; 1 - \frac{1}{W_{all}} \max_{\pi \in \Pi_K} \sum_k m^{\mathbf{w}}_{k,\pi(k)} \tag{7}$$

2.2 The K-means clustering cost

In K-means clustering, the data points $\{z_1, \ldots, z_n\}$ are vectors in $\mathbb{R}^d$. Let Z be the $n \times d$ data matrix having z_i on row i, and A be the Gram matrix given by $A_{ij} = z_i^T z_j$ or $A = ZZ^T$. We will assume w.l.o.g. that the data are *centered* at the origin, i.e $\sum_i z_i = 0$ or, in matrix notation $\mathbf{1}^T Z = 0$. Therefore, Z and A will have rank at most d. The *squared error distortion*, often called "K-means" cost function, is defined as

$$\mathcal{D}(X) \;=\; \sum_{k=1}^{K} \sum_{i \in C_k} ||z_i - \mu_k||^2 \tag{8}$$

In the above, μ_k, $k = 1, \ldots K$ are the clusters' *centers*, whose coordinates in $\mathbb{R}^d$ are given by

$$\mu_k \;=\; \frac{1}{n_k} \sum_{i \in C_k} z_i, \quad \text{for } k = 1, \ldots K \tag{9}$$

If one substitutes the expression of the centers (9) into (8) and represents a clustering by the orthonormal column matrix X defined above, one can show that the distortion is a quadratic function of X [Ding and He, 2004]

$$\mathcal{D}(X) \;=\; \operatorname{tr} A - \operatorname{tr} X^T A X \tag{10}$$

Furthermore, because the columns of $\tilde{X}$ sum to 1, the last column is determined by the other $K-1$ and therefore one can uniquely represent any clustering by a matrix with $K-1$ orthonormal columns Y as follows. Let $c \in \mathbb{R}^K$ be the vector

$$c \;=\; \left[\sqrt{\frac{n_1}{n}} \cdots \sqrt{\frac{n_k}{n}} \cdots \sqrt{\frac{n_K}{n}}\right]^T \tag{11}$$

with $||c|| = \sqrt{(\sum_k n_k)/n} = 1$. Let V be a $K \times K$ orthogonal matrix with c on its last column. It can be verified easily that $Xc = \mathbf{1}/\sqrt{n}$. Then, XV is a matrix with orthonormal columns, whose last column equals $\mathbf{1}/\sqrt{n}$, where $\mathbf{1}$ denotes the vector of all 1's. Denote

$$XV \;=\; \left[Y \;\; \mathbf{1}\frac{1}{\sqrt{n}}\right]. \tag{12}$$

We can now rewrite the distortion in terms of the $n \times (K-1)$ matrix Y, obtaining

$$\begin{aligned}
\mathcal{D}(Y) &= \mathit{constant} - \operatorname{tr}\left[Y\,\mathbf{1}\frac{1}{\sqrt{n}}\right]^T A \left[Y\,\mathbf{1}\frac{1}{\sqrt{n}}\right] \\
&= \mathit{constant} - \operatorname{tr} Y^T A Y - \frac{1}{n}\mathbf{1}^T A \mathbf{1} \\
&= \mathit{constant} - \operatorname{tr} Y^T A Y
\end{aligned} \tag{13}$$

The last equality holds because $A\mathbf{1} = ZZ^T\mathbf{1} = 0$. It has been noted [Ding and He, 2004] that relaxing the integrality constraints in the above equation results in a trace maximization problem that is solved by an eigendecomposition. Hence, we have that for any clustering X represented by Y as above,

$$\mathcal{D}(Y) \;\geq\; \mathcal{D}^* \;=\; \operatorname{tr} A - \sum_{k=1}^{K-1} \sigma_k \quad \text{attained for} \quad Y \;=\; U \tag{14}$$

where $\sigma_1, \ldots \sigma_{K-1}$ are the $K-1$ principal eigenvalues of A and U is the $n \times (K-1)$ matrix containing the principal eigenvectors.

2.3 The Multiway Normalized Cut clustering cost

In graph partitioning, the data is a set of *similarities* S_{ij} between pairs i, j of nodes in the set $V = \{1, 2, \ldots n\}$. The similarities satisfy $S_{ij} = S_{ji} \geq 0$. The matrix $S = [S_{ij}]_{i,j \in V}$ is called the *similarity matrix*. If we assimilate V with

the node set of a graph, in graph theory terminology S represents a *weighted adjacency matrix*. The *weight* of node i is defined as

$$w_i \;=\; \sum_{j \in V} S_{ij} \tag{15}$$

W.l.o.g we assume that no node has weight 0. The weight of a set $A \subseteq V$ is $W_A = \sum_{i \in A} w_i$.

The *multiway normalized cut (NCut)* clustering objective [Meilă, 2002, Yu and Shi, 2003] is

$$NCut(\mathcal{C}) \;=\; \sum_{k=1}^{K} \sum_{k' \neq k} \frac{Cut(C_k, C_{k'})}{W_{C_k}} \tag{16}$$

where

$$Cut(A, B) \;=\; \sum_{i \in A} \sum_{j \in B} S_{ij} \tag{17}$$

It is known from [Meila and Xu, 2003, Yu and Shi, 2003] that the multiway normalized cut of a clustering $\mathcal{C}$ with K clusters in the weighted graph represented by the similarity matrix S can be expressed as

$$NCut(\mathcal{C}) \;=\; K - \operatorname{tr} X^T L X \tag{18}$$

where X is the normalized matrix representation of clustering $\mathcal{C}$ and L is the normalized similarity matrix defined as

$$L \;=\; \operatorname{diag}(\mathbf{w})^{-1/2} S \operatorname{diag}(\mathbf{w})^{-1/2} \tag{19}$$

By a reasoning similar to the one leading to equation (14) one can show [Meila and Xu, 2003] that for any clustering X

$$NCut(X) \;\geq\; \mathcal{N}^* \;=\; K - \sum_{k=1}^{K} \lambda_k \quad \text{attained for } X \;=\; U \tag{20}$$

where U is the $n \times K$ matrix containing the principal eigenvectors of L and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ are the eigenvalues of L.

2.4 Summary of results

We now give a preview of the main results of this paper. Some of the technical conditions will be left vague here, to be explicated later.

We assume data given in the form of a matrix A defined as in section 2.2 for squared distortion clustering and as in (19) for spectral clustering. In addition, for spectral clustering, the node weights $\mathbf{w}$ are given and for K-means clustering the data are assumed centered, i.e $Z^T \mathbf{1} = 0$. We assume a fixed K and we denote by $\mathcal{D}^*$, respectively $\mathcal{N}^*$, the spectral lower bound for the cost functions $\mathcal{D}(X)$ and $NCut(X)$, and by X^{opt} the (unknown) optimal clustering according to the respective criterion.

We prove that for any K-clustering X whose cost is sufficiently low, the distance $d(X, X^{opt})$ can be bounded above by a value that depends only on known quantities and can be computed easily. For the squared distortion $\mathcal{D}(X)$, we have the following:

Imprecise version of Theorem 4 *Let X be any clustering of a data set represented by the Gram matrix $A = [z_i^T z_j]_{i,j=1}^n$. If $\delta = \frac{\mathcal{D}(X) - \mathcal{D}^*}{\sigma_{K-1} - \sigma_K}$ is sufficiently small, then*

$$d(X, X^{opt}) \;\leq\; \text{bound}(\delta, X) \tag{21}$$

where X^{opt} represents the clustering with K clusters that minimizes the distortion $\mathcal{D}$ on the data A.

An analog result holds in the case of the Normalized Cut cost objective.

Imprecise version of Corollary 6 *Let X be any clustering of a data set represented by the symmetric similarity matrix $S = [S_{ij}]$, $S_{ij} \geq 0$. Let the vector of node weights be $\mathbf{w} = [w_i]$, $w_i > 0$, and let W_k, $k = 1 \ldots K$ be defined as in (1). If $\delta = \frac{NCut(X) - \mathcal{N}^*}{\lambda_K - \lambda_{K+1}}$ is sufficiently small, Then*

$$d^{\mathbf{w}}(X, X^{opt}) \;\leq\; \text{bound}'(\delta, X) \tag{22}$$

where X^{opt} represents the clustering with K clusters that minimizes the K-way Normalized Cut on the data S.

In the above, $\mathcal{D}(X), \mathcal{D}^*, \sigma_{K-1}, \sigma_K$ are defined as in section 2.2 and $\mathcal{N}(X), \mathcal{N}^*, \lambda_{K+1}, \lambda_K$ are defined as in 2.3.

The exact expression of the functions bound(δ, X), bound'(δ, X) and the other technical conditions for which these inequalities hold are given in theorem 4 and corollary 6 respectively. Theorem 9 in section 6 is a further generalization that includes as special cases both theorem 4 and corollary 6.

3 A clustering with small K-means distortion is close to the optimal clustering

We call *good* a K-clustering whose distortion $\mathcal{D}(X)$ is not too large compared to the lower bound $\mathcal{D}^*$, that is $\mathcal{D}(X) - \mathcal{D}^* \leq \epsilon$, for an ϵ to be determined. Let X^{opt} be the K-clustering of A with the smallest distortion and note that $\mathcal{D}(X) \geq \mathcal{D}(X^{opt}) \geq \mathcal{D}^*$. We will show that under certain conditions which can be verified on the data, if a clustering X is good, then it is not too dissimilar from X^{opt}, as measured by the misclassification error distance $d(X, X^{opt})$.

This result will be proved in three steps. First, we will show that any good clustering represented by its Y matrix is close to the $(K-1)$-th principal subspace U of A. Second, we show that any two good clusterings must be close

to each other under the distance d. Based on this, in the third step we obtain the desired result.

Let Y be a clustering with a corresponding c defined as in (11); Y can be written as

$$Y \;=\; [\,U\; U_e\,] \left[\begin{array}{c} R \\ E \end{array}\right] \tag{23}$$

where $U^{all} = [U\; U_e] \in \mathbb{R}^{n\times n}$ is the orthogonal basis represented by the eigenvectors of A and $R \in \mathbb{R}^{(K-1)\times(K-1)}$, $E \in \mathbb{R}^{(n-K+1)\times(K-1)}$ are matrices of coefficients. Additionally, because Y, U^{all} are orthogonal, $[R^T\, E^T]^T$ is also orthogonal. We show that if $\mathcal{D}(Y)$ is small enough, then E is small.

Theorem 1 *For any clustering Y represented like in (23) the following inequality holds*

$$||E||_F^2 \;\leq\; \delta \;=\; \frac{\mathcal{D}(Y) - \mathcal{D}^*}{\sigma_{K-1} - \sigma_K} \tag{24}$$

By $||\;||_F$ we denote the Frobenius norm of a matrix, $||M||_F^2 = \operatorname{tr} M^T M$. The proof of the theorem is given in the appendix.

We now show that two clusterings Y, Y' for which δ is small must be close to each other. First we show that a certain function $\phi(X, X')$ taking values in $[0, K]$ is close to its maximum K when Y, Y' are both close to the subspace spanned by U. Then, we show that when $\phi(X, X')$ is large, the misclassification error $d(X, X')$ is small.

Denote by $\phi(X, X')$ the following function, defined for any two $n \times K$ matrices with orthonormal columns.

$$\phi(X, X') \;=\; ||X^T X'||_F^2 \tag{25}$$

Since the Frobenius norm $||\;||_F$ of an orthogonal matrix with K columns is $\sqrt{K}$ we have

$$0 \leq \phi(X, X') \;=\; ||X^T X'||_F^2 \;\leq\; ||X||_F ||X'||_F \;=\; K$$

Lemma 2 *For any two clusterings X, X' denote by δ, respectively δ' the corresponding values of the r.h.s term of (24). For $\delta, \delta' \leq (K-1)/2$*

$$\phi(X, X') \;\geq\; K - \epsilon(\delta, \delta') \tag{26}$$

with

$$\epsilon(\delta, \delta') \;=\; 2\sqrt{\delta\delta'(1 - \delta/(K-1))(1 - \delta'/(K-1))} \tag{27}$$

This lemma is proved in the appendix.

Theorem 3 *(after [Meilă, 2011]) For two weighted clusterings with K clusters each, if $\phi(X, X') \geq K - \epsilon$, $\epsilon \leq p_{min}$ then $d_{ME}^{\mathbf{w}}(X, X') \leq \epsilon p_{max}$, where $p_{max} = \max_k W_k/W_{all}$, $p_{min} = \min_k W_k/W_{all}$.*

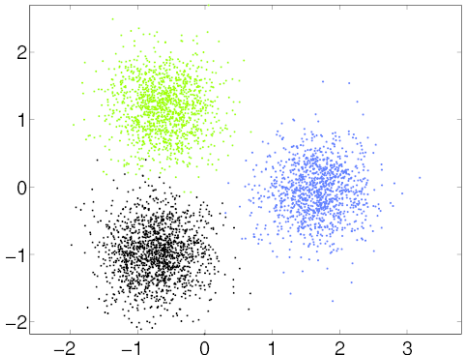


Figure 1: A mixture of 3 normal distributions in $d = 35$ dimensions, with fixed centers and equal covariances $\sigma^2 I_d$, $\sigma = 0.4$, projected on its second principal subspace. The true mixture labels are shown in different colors.

Note the asymmetry of this statement, which involves only the p_{max}, p_{min} values of one clustering. This is crucial in allowing us to prove the result we have been striving for.

Theorem 4 *Let X be any clustering of a data set represented by the Gram matrix $A = [z_i^T z_j]_{i,j=1}^n$, with $\sum_{i=1}^n z_i = 0$. Let $p_{max} = \max_k n_k/n$, $p_{min} = \min_k n_k/n$, let δ be given by (24) and ϵ by (27). Then, if $\delta \leq (K-1)/2$ and $\epsilon(\delta, \delta) \leq p_{min}$ then*

$$d(X, X^{opt}) \;\leq\; \epsilon(\delta, \delta) p_{max} \tag{28}$$

where X^{opt} represents the clustering with K clusters that minimizes the distortion $\mathcal{D}$ on the data A.

Proof: We know that $\mathcal{D}(Y^{opt}) \leq \mathcal{D}(Y)$ and hence $||E^{opt}||_F^2 \leq \delta$ from theorem 1. By applying lemma 2 and theorem 3 we obtain the desired result. QED

A few remarks are in place. First, the bound δ in theorem 1 is necessary only for the unknown clustering X^{opt}; for a known clustering, one can directly compute $||E||_F^2$ and therefore obtain a tighter bound. We have followed this route in the experiments of next section. Second, theorem 4 implies that $d(X, X^{opt}) \leq p_{min} p_{max} \leq p_{min}$. Hence, for p_{max} not too large, the bound is informative, it tells one that all clusters in X^{opt} have been identified.

It should be also noted that the condition $\epsilon \leq p_{min}$ in theorem 3 is only sufficient, not necessary. We are working currently toward a general condition that would extend the domain of theorem 3 to ϵ's larger than p_{min} (e.g. of the order $2p_{min}$).

4 Experiments

Worst case bounds are notoriously lax; therefore we conducted experiments in order to check that the bounds in this paper ever apply. In the experiments

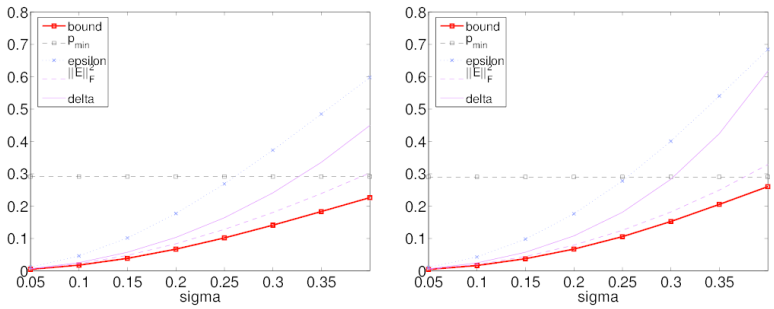


Figure 2: The bound used as a certificate of correctness. The data represents a mixture of 3 normal distributions in $d = 35$ dimensions, with fixed centers and equal covariances $\sigma^2 I_d$; this data is depicted in figure 1 for $\sigma = 0.4$. The clustering X represents the K-means solution. In **(a)**, the bound and the values of p_{min}, ϵ, $||E||_F^2$, δ for X are evaluated at different values of σ; the data set has size $n = 1000$. In **(b)** the same are plotted for $n = 100$.

illustrated by Figure 2 we generated data from a mixture of spherical normal distributions, clustered them with the K-means algorithm (with multiple initializations), then evaluated the bound and the other related quantities. The spread of the clusters, controlled by the standard deviation σ, varied from $\sigma = 0.05$ (very well separated) to $\sigma = 0.4$ (clusters touching). The centroids are fixed inside the $[0, 1]^d$ hypercube. In all cases we confirmed by visual inspection that K-means found a (nearly) optimal clustering. Therefore, the true $d(X, X^{opt})$ is practically identical 0. The bound worsens with the increase of σ, as expected, from 0.004 to 0.22. Up to values of $\sigma = 0.3$, however, the bound is lower than $p_{min}/2$. This confirms qualitatively that we have found a "correct" clustering, in the sense that the total number of misclustered points is a fraction of the smallest cluster size.

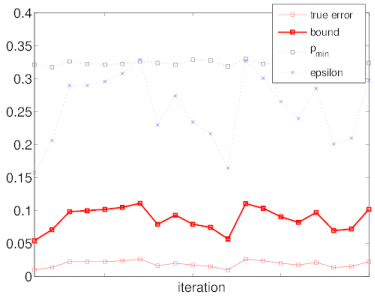
The values of ϵ are plotted to verify that corollary 4 applies. For the two largest values of σ, ϵ is outside the admissible domain, so the bound is not provably correct.

The lines with no markers display the quantity $||E||_F^2$ for the found clustering (with E defined in section 3) and its upper bound δ from (24). We see that the quality of this bound in absolute value also degrades with increasing σ; however, the ratio $\delta/||E||_F^2$ is approximatively constant around 1.4. This occurred uniformly over all our experiments with mixtures of Gaussians.

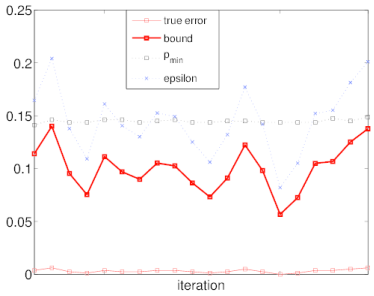
A comparison between Figure 2,a and 2,b shows that there is practically no variation due to the data set size, except for a slight improvement for larger n. This is consistent with the theory and with all our other experiments so far.

Figure 3 shows a different experiment. Here the optimal clustering[1] X is per-

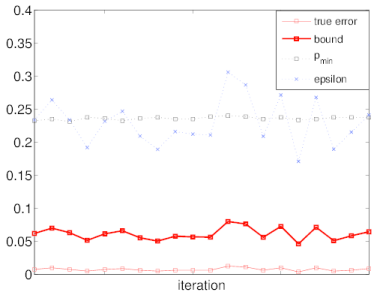
[1]We assume X to be represented by the true labels, which is extremely plausible as the



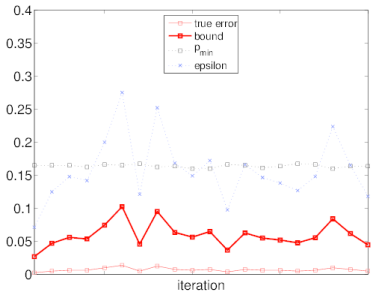
a $K = 3$, $\sigma = 0.1$, $p_{err} = 3\%$, $p_{min} \approx 0.32$
$||\mu_k - \mu_{k'}|| \in [2.3, 2.5]$



b $K = 3$, $\sigma = 0.1$, $p_{err} = 0.5\%$, $p_{min} \approx 0.15$
$||\mu_k - \mu_{k'}|| \in [2.2, 2.3]$



c $K = 4$, $\sigma = 0.1$, $p_{err} = 1\%$, $p_{min} \approx 0.22$
$||\mu_k - \mu_{k'}|| \in [2.0, 2.2]$



d $K = 4$, $\sigma = 0.03$, $p_{err} = 1\%$, $p_{min} \approx 0.15$
$||\mu_k - \mu_{k'}|| \approx 1.4$

Figure 3: The data represents a mixture of K normal distributions in $d = 25$ dimensions, with fixed centers and equal covariances $\sigma^2 I_d$; X represents the true mixture labels, which can be assumed to be the optimal clustering for these data. We construct X' by perturbing the labels of X randomly w.p. p_{err}. The figure displays the value of $d(X, X')$ and the values for the bound, ϵ and p_{min} for 20 randomly sampled X's; $n = 800$ in all cases.

turbed randomly into X'. We evaluate the true misclassification error $d(X, X')$ and its bound, together with other relevant quantities for $K = 3, 4$, each with a uniform and a non-uniform clustering. Note that the bound becomes looser when $d(X', X^{opt})$ and K increase, or when p_{min} decreases. For instance, in Figure 3, d, almost than half of the clusterings have invalid bounds. Hence, the figures demonstrate both the informativeness of the bounds and the limitations of their applicability.

The degradation with decreasing p_{min} is completely expected, based for example on the condition $\epsilon \leq p_{min}$ in theorem 3. This behavior also agrees with the common wisdom that small clusters in the data make clustering more difficult practically (higher chance of missing a cluster) and harder to analyze theoretically. In our framework, we can say that small clusters in the data reduce the confidence that a clustering X is optimal, even when it is so.

5 Extension to weighted data and the Normalized Cut cost

Extending theorem 4 to weighted and kernel-based distortion functions is immediate. Assume that the data points are weighted with weights $\mathbf{w} = [w_i]_{i=1}^n$. The *weighted distortion* is defined as

$$\mathcal{D}^{\mathbf{w}}(\mathcal{C}) \;=\; \min_{\mu_1,\ldots\mu_K \in \mathbb{R}^d} \sum_k \sum_{i \in C_k} w_i ||z_i - \mu_k||^2 \tag{29}$$

It can be easily checked that the centroids μ_k that minimize the above expression for Z and $\mathcal{C}$ fixed are the weighted means of the data in each cluster

$$\mu_k \;=\; \frac{\sum_{i \in C_k} w_i z_i}{W_k} \;\text{ for } k = 1, \ldots K \tag{30}$$

By replacing the above values in (29) we obtain after some calculations

$$\mathcal{D}^{\mathbf{w}}(\mathcal{C}) \;=\; \operatorname{tr} A - \operatorname{tr} X^T A X \tag{31}$$

with X defined as in (3) and

$$A \;=\; \operatorname{diag}(\sqrt{\mathbf{w}}) Z Z^T \operatorname{diag}(\sqrt{\mathbf{w}}) \tag{32}$$

An important application of theorem 3 for weighted data it to the problem of graph partitioning with the Normalized Cut cost.

By comparing the quadratic representation of the *NCut* criterion (18,19) and of the weighted distortion (31,32) one can see that the normalized cut of any clustering X in S equals (up to a constant) the weighted distortion $\mathcal{D}(X)$ of the

clusters are well-separated.

same clustering for a mapping of the graph nodes $i = 1, \ldots n$ into d-dimensional vectors[2] $z_1, \ldots z_n$. To find the mapping we set $A = L$, and obtain

$$Z \;=\; \mathrm{diag}(\mathbf{w})^{-1}\sqrt{S} \tag{33}$$

In the above, the $\sqrt{S}$ is the matrix square root of S, which is real if S is non-negative definite and complex otherwise. The matrix square root $\sqrt{S}$ satisfies $\sqrt{S}\sqrt{S}^* = S$ where M^* denotes the transpose complex conjugate of matrix M.

With the mapping Z as in (33), we have

$$NCut(X) \;=\; \mathcal{D}(X) - \mathrm{tr}\, L + K \quad \text{for all } X \tag{34}$$

Because for any clustering $NCut(X)$ differs from $\mathcal{D}(X)$ by a constant independent of X, we can use Theorem 3 in order to obtain an analog for partititions in a graph that are "good" under the *NCut* criterion.

A necessary preparation for this is "centering" the data Z, as the matrix A in theorem 4 is assumed to be obtained from centered data. In the following lemma we show how to directly evaluate the effect of centering on the eigenvalues and eigenvectors of L. We start with some notation. Let Z be the embedding of the graph nodes according to (33). Let $Z_0 = Z - \mathbf{1}m^T$ denote the embedded points shifted by the vector m, so that $Z_0^T\mathbf{w} = 0$. That is, Z_0 represents the centered data. Let L_0 be the "centered L" matrix, i.e. the matrix obtained by applying the r.h.s of (32) to Z_0. Note that although Z, m, Z_0 may be complex, L, L_0 are always real and symmetric matrices.

Lemma 5 *Let $n, \mathbf{w}, L, L_0$ be defined as above. Let $\lambda_1 = 1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ be the eigenvalues of L and $u_1, \ldots u_n$ be the corresponding eigenvectors. Then*

i) $L_0 \;=\; (I - B)L(I - B)$ where the matrix $B = \sqrt{\mathbf{w}}\sqrt{\mathbf{w}}^T/W_{all}$ represents the projection onto the direction $\sqrt{\mathbf{w}}$.

ii) The eigenvalues and eigenvectors of L_0 are

$$\lambda_j^0 \;=\; \begin{cases} 0, & \text{if } j = 1 \\ \lambda_j & \text{if } j > 1 \end{cases}, \qquad u_j^0 \;=\; u_j \quad \text{for all } j \tag{35}$$

iii) Let X be a clustering and Y be an orthogonal $n \times (K-1)$ matrix satisfying $XV \;=\; [\,u_1\; Y]$ for V an orthogonal matrix[3]. Then

$$\mathrm{tr}\, L \;=\; \mathrm{tr}\, L_0 + 1 \tag{36}$$

$$\mathrm{tr}\, X^T L X \;=\; \mathrm{tr}\, Y^T L_0 Y + 1 \tag{37}$$

$$\mathcal{D}(X) \;=\; \mathrm{tr}\, L_0 - \mathrm{tr}\, Y^T L_0 Y \tag{38}$$

[2]This fact was noted by [Bach and Jordan, 2004] and used by [Dhillon et al., 2004] for the special case S positive definite.

[3]This decomposition is not possible in general, but it can be verified that it's always possible when X represents a clustering.

We can now apply theorem 1 to the distortion expressed as in (38). If we take into account lemma 5 and we assume in addition that

$$\lambda_{K+1} \geq 0 \tag{39}$$

we obtain

$$\delta \;=\; \frac{\lambda_2 + \ldots + \lambda_K - \operatorname{tr} Y^T L_0 Y}{\lambda_K - \lambda_{K+1}} \;=\; \frac{1 + \lambda_2 + \ldots + \lambda_K - \operatorname{tr} X^T L X}{\lambda_K - \lambda_{K+1}} \tag{40}$$

Assumption (39) is often verified in practice. If it is true, then the $K-1$ largest eigenvalues of L_0 are $\lambda_2, \ldots \lambda_K$ and its $(K-1)$-th eigengap is $\lambda_K - \lambda_{K+1}$. If (39) does not hold, then the modification of the bound in equation (40) is immediate.

With this, we have succeeded in bounding the distance of a clustering with small *NCut* to the optimal clustering possible for data S.

Corrollary 6 *Let X be any clustering of a data set represented by the symmetric similarity matrix $S = [S_{ij}]$, $S_{ij} \geq 0$. Let the vector of node degrees be $\mathbf{w} = [w_i]$, $w_i > 0$, W_k, $k = 1 \ldots K$ be defined as in (1), $p_{max} = \max_k W_k/W_{all}$, $p_{min} = \min_k W_k/W_{all}$; let δ be given by (40) and ϵ by (27). Assume $\lambda_{K+1} \geq 0$, where λ_{K+1} is the $(K+1)$-th eigenvalue of $L = \operatorname{diag}(\mathbf{w})^{-1/2} S \operatorname{diag}(\mathbf{w})^{-1/2}$.*

Then, if $\delta \leq (K-1)/2$ and $\epsilon(\delta, \delta) \leq p_{min}$

$$d^{\mathbf{w}}(X, X^{opt}) \;\leq\; \epsilon(\delta, \delta) p_{max}$$

where X^{opt} represents the clustering with K clusters that minimizes the K-way Normalized Cut on the data S.

We now compare this bound with the previously obtained bound of [Meilă et al., 2005], which we reproduce here.

Theorem 7 (after [Meilă et al., 2005], Theorem 1) *Let $\mathcal{C}, \mathcal{C}'$ be two K-way clusterings of the weighted graph represented by the similarity matrix S, let δ, δ', $\lambda_1, \lambda_2, \ldots \lambda_{K+1}$ be defined as in Corollary 6 and let the function $\phi(X, X')$ be defined by (25). Then, whenever $\delta \leq 1$,*

$$\phi(X, X') \;\geq\; K - \epsilon^{old}(\delta, \delta')$$

with

$$\epsilon^{old}(\delta, \delta') \;=\; 2\sqrt{\delta\delta'(1-\delta)(K-\delta')} + K\delta + \delta' - 2\delta\delta' \tag{41}$$

There are some slight differences in the requirements of the above theorem versus corollary 6. The expression for ϵ^{old} is defined only for $\delta \leq 1$, a more restrictive requirement than $\delta \leq (K-1)/2$ in the definition of ϵ. On the other hand, assumption (39) is not necessary. We remind the reader that assumption (39) is a simplifying assumptions, which allows one to compute δ according to the same formula in all cases. If this assumption is not satisfied, our main results will not be invalidated. Merely, the equation of δ will be changed in a way in which the comparison between the old and new criterion will be less straightforward. More interesting is the comparison between the bounds given by the two criteria. This is the object of the next lemma.

Lemma 8 *$\epsilon^{old}(\delta, \delta) \geq \frac{K}{2}\epsilon(\delta, \delta)$ for all $\delta \leq 1$.*

Hence, the new bound improves the result of [Meilă et al., 2005].

6 General quadratic cost function

Theorem 9 *Let $\mathcal{D}^{\mathbf{w}}$ be any clustering cost function that can be expressed in the form*

$$\mathcal{D}^{\mathbf{w}}(X) \;=\; C_o - \operatorname{tr} X^T A_o X$$

where X is a (weighted) clustering defined as in (3), and $C_o \in \mathbb{R}$, $A_o \in \mathbb{R}^{n\times n}$ symmetric depend only on the data and on the data weights $\mathbf{w} = [w_i]$, $w_i > 0$. Define $W_{all} = \sum_i w_i$, W_k, $k = 1 \dots K$ as in (1), $p_{max} = \max_k W_k/W_{all}$, $p_{min} = \min_k W_k/W_{all}$.

i) Let $B = \sqrt{\mathbf{w}}\sqrt{\mathbf{w}}^T/W_{all}$ and $A = (I - B)A_0(I - B)$ Then, for any clustering X, $A\sqrt{\mathbf{w}} = 0$ and

$$\mathcal{D}^{\mathbf{w}}(X) \;=\; C - \operatorname{tr} X^T A X \tag{42}$$

with C a constant independent of X.

ii) $\mathcal{D}^* = C - \sum_{k=1}^{K-1} \sigma_k(A)$ is a lower bound for $\mathcal{D}^{\mathbf{w}}(X)$.

iii) Let δ be given by (24) and ϵ by (27). Then, if $\delta \leq (K-1)/2$ and $\epsilon(\delta, \delta) \leq p_{min}$

$$d^{\mathbf{w}}(X, X^{opt}) \;\leq\; \epsilon(\delta, \delta)p_{max}$$

where X^{opt} represents the clustering with K clusters that minimizes the cost $\mathcal{D}^{\mathbf{w}}$ for the given data and weights.

In this form, our result encompasses the K-means distortion and the *NCut* as well as several other clustering cost functions. The most notable are the *kernel K-means distortion* and various graph partitioning criteria like for example the *Average Association* [].

For the Average Association, we have $\mathbf{w} = \mathbf{1}/n$ and $\mathcal{D}^{\mathbf{w}}(X) = -\operatorname{tr} X^T S X$ where S is the graph similarity matrix defined in section 2.3.

In kernel K-means (see [Schölkopf et al., 1998] for details) the data points z_i are mapped in a high (possibly infinite) dimensional Hilbert space $\mathcal{H}$ called the *feature space* by $z_i \xrightarrow{h} h_i = h(z_i)$. The dot product in $\mathcal{H}$ between two feature vectors h_i, h_j can be pulled back in the original z_i, z_j by the Mercer kernel $\kappa(z_i, z_j) = h_i^T h_j$. The Gram matrix A_o is redefined to be

$$A_o \;=\; [\kappa(z_i, z_j)]_{i,j=1}^n \tag{43}$$

The kernel K-means clustering cost function is the distortion w.r.t $\mathcal{H}$. It is easy to see that with A_o defined as in (43) the distortion takes the same form as in theorem 9.

7 Discussion

Intuitively, we have proved that, if (1) the data is well clustered, and (2) by some algorithm a good clustering X is found, then we can bound the distance between X and the unknown optimal clustering X^{opt} of this data set. Hence, we will have a "certificate" that our clustering X is almost optimal.

In the present context, "well clustered" means that the affine subspace determined by the centroids $\mu_1, \ldots \mu_K$ is parallel to the $K-1$ principal components of the data[4] Z. In other words, the first $K-1$ principal components of the variance are mainly due to the inter-cluster variability. This in turn implies that the bound will not exist (or will not be useful) when the centroids span an affine subspace of lower dimension than $K-1$. For example, if $\mu_1, \ldots \mu_K$, $K > 2$ are along a line, no matter how well separated the clusters, then the vectors U will give only partial information on the optimal clustering. Practically, this means that "well separated" refers not just to the distances between the clusters, but to the volume (of the polyhedron) spanned by them, which should be as large as possible.

By the same geometric view, a "good clustering" is one whose Y representation lies close to the principal subspace U. This is implied in much of the prior work, e.g [Ding and He, 2004, Meilă, 2002, Yu and Shi, 2003, Ng et al., 2002, Dhillon et al., 2004, Bach and Jordan, 2004]. Paper adds that all the clusterings that are near U must be very similar.

From the perspective of the function $\mathcal{D}(X)$, we have shown *quantitatively*, that if the data is well clustered, $\mathcal{D}(X)$ has a unique "deep crater". When points are moved to other clusters w.r.t X^{opt} the distortion grows fast because the clusters are far apart. Conversely, if the distortion is small, it means that we cannot be elsewhere than near X^{opt}. "Small" is measured as deviation from the lower bound $\mathcal{D}^*$ in $\sigma_{K-1} - \sigma_K$ units.

7.1 Related work – Probably correct algorithms for mixtures

To our knowledge, this result is the first of this kind for the K-means distortion. There is however a large body of work in theoretical computer science dealing with estimating mixtures of normal distributions with high probability (w.h.p.). This was pioneered by [Dasgupta, 1999] who presented an algorithm that estimates Gaussian mixtures with sufficiently "rounded" and separated clusters by projecting the data on a random subspace of dimension $\mathcal{O}(k)$. The paper of [Vempala and Wang, 2004] shows that by projecting a mixture of spherical gaussians on the $K-1$-th principal subspace of the data instead of a random subspace, the mixture components (clusters) can be identified at lower separations. More sophisticated use of the spectral projection by [Achlioptas and McSherry, 2005, Ravindran et al., 2005] result in algorithms for

[4]The matrix $A = ZZ^T$ and the data covariance matrix have the same non-zero eigenvalues up to a factor n; U is the projection of the data on the principal subspace.

mixtures of general log-concave distributions with arbitrary covariance matrices working at lower separations.

While technically our results do not rely on the above mentioned papers, it is instructive to look at both the similarities and the differences between the two approaches to clustering. The computer science papers offer polynomial algorithms for finding the clustering, plus guarantees that the clustering will correct with high probability. More precisely, these papers contain theorems saying that under certain separation conditions, with sufficiently large sample sizes, and with probability at least δ, the clustering returned by the proposed algorithm will correspond to the true mixture labels. There is a subtle difference here, as the "true clustering" (let us denote it by X^{true}) is not always the same as the Maximum Likelihood clustering X^{opt}. However, the two clusterings are the same w.h.p which means that the aforementioned algorithms guarantee $d(X, X^{opt}) = 0$ w.h.p. Hence, the "computer science" techniques give stronger theoretical guarantees, and provide algorithms. These guarantees rely on strong assumptions about the data distribution, in particular knowledge of the shape of the clusters (Gaussian or log-normal, sometimes spherical symmetry) and of the cluster separation. The proofs use concentration results (e.g. Chernoff bounds) for these distribution classes.

Our paper's results rely on much weaker assumptions. We do make no (explicit) assumptions about the sample size, nor about the distribution of the data inside each cluster. Hence our results are worst case results and necessarily the weakest possible. We make only one explicit assumption, that a clustering with low cost exists, where "low" is measured in δ units. Finding distributions and sample size when this condition is met is one of the areas that this research is opening. We do not explicitly offer an algorithm, but one can think of the spectral algorithm of [Ding and He, 2004] or of the variant of EM with PC projection used by [Srebro et al., 2006] as associated algorithms. Or, one can use our results after algorithms like that of [Vempala and Wang, 2004] as a certificate of correctness for the found clustering. This would allow such algorithms to be run with a lower confidence parameter (larger δ).

Beyond the differences in posing the problem, there is also a fundamental similarity between our paper and the spectral embedding methods of [Achlioptas and McSherry, 2005, Ravindran et al., 2005] and especially of [Vempala and Wang, 2004] that we discuss now. A crucial fact proved in [Vempala and Wang, 2004] is that for a sufficiently large n and for well separated clusters the K-th principal subspace of the data and the K dimensional subspace determined by the cluster centroids are close. Hence, the K-th principal subspace of the data contains information about the best clustering. Our result exploits the same fact, as theorem 4 holding implicitly means that all good clusterings, when represented as subspaces, are close to the $K-1$-th principal subspace of the data[5]. Hence, both groups of results rely on the informativity of the principal subspace w.r.t to a salient clustering in the data. In [Achlioptas and McSherry, 2005], the same fact is ex-

[5]The difference of 1 in the dimensions comes from the fact that we assume the data is centered.

ploited albeit in a slightly different way: even if the clusters are not isotropic (e.g ellipsoidal instead of spherical) there is a separation at which the principal subspace will coincide with the subspace spanned by the cluster centers. The algorithm presented by [Achlioptas and McSherry, 2005] is based on the distance preserving property of the projection on the principal subspace.

One can also draw an analogy between our result and the VC type bounds for *structural risk minimization (SRM)* (see e.g [Vapnik, 1998]). These are distribution free worst case bounds for the expected risk of a learned model on a data set. They depend on the empirical risk (i.e. the observed error rate) and on the complexity of the learned model. The bound is looser if either of the two components is higher. Similarly, theorem 4 gives a bound on the error of an obtained clustering w.r.t the best possible clustering of the same data set. The bound is worst case and distribution free and depends on the observed distortion; it also depends on $1/\text{eigengap}(A)$ and increases when either component increases. It is known that the $K-1$-th eigengap of a matrix measures the stability of its $K-1$-th principal subspace to perturbations. Thus, its inverse can be regarded as the analog of a "complexity" measure. There are also obvious differences: while in SRM the bound is for the expected error over unseen samples from the same distribution, in our case the sample is fixed. Our bound is not a generalization bound. The SRM bounds are sometimes greater than 1; here the bound doesn't always exist but is always informative when it does.

7.2 Related stability results

As mentioned before, theorem 3 is a *stability* result. Recently, another stability result for spectral clustering with the Normalized Cut cost appeared [Ben-David et al., 2006] (BLP) which builds on previous work by [von Luxburg et al., 2005] and [Rakhlin and Caponnetto, 2006]. We briefly summarize BLP. This paper is concerned with *algorithmic stability*; i.e can we bound the clustering algorithm's output variability that is due to sampling noise? If the variability tends to 0 when the sample size tends to infinity, then the algorithm is *stable*. The paper establishes general conditions for the stability of an algorithm $\mathcal{A}$ when data is sampled according to P, the clustering cost function is $\mathcal{D}^P(\mathcal{C})$ and the clusterings are compared with some distance d. These are (1) that P has a *unique minimizer*, and (2) that the algorithm $\mathcal{A}$ is *R-minimizing*.

The second condition essentially means that algorithm $\mathcal{A}$ is an ϵ-optimizer of the cost R for any finite sample. The first condition states that any low $\mathcal{D}^P$ clustering is similar to the optimal clustering on the given probability space. In other words, results like the one proved in this paper, are necessary prerequisites for the main theorem in BLP[6]

[6]The previous statement can be made more precise: Our theorem 9 proves that for a quadratic cost, under certain verifiable conditions, a distribution P with finite support has a unique minimizer, while the definition in BLP refers to general probability spaces. To bridge the gap it remains to take the limit $n \to \infty$ in theorem 9. We regard this as possible subject to some distributional assumptions about the data, but as beyond the scope of the present paper.

One can think of BLP as proving that: if in the limit of $n \to \infty$ all almost optimal clusterings are similar, and if I have an algorithm which always produces an almost optimal clustering w.r.t the cost, then the output of algorithm A on a large finite sample will not vary much with the sample.

Our paper makes no general assumption about the clustering algorithm. It only assumes that it was capable once to find a low cost clustering, where "low cost" means low enough for $\epsilon \leq p_{min}$. From this it follows that, *on the current data set* all low cost clusterings are similar.

7.3 An alternative distance between clusterings.

The function $\phi(X, X')$ used as an intermediary vehicle for the proof of theorem 4 can in fact represent a distance in its own right. Denote $d^2_\chi(X, X') = 1 - \frac{1}{min(K,K')}\phi(X, X')$. This function is 0 when the clusterings are identical and 1 when they are independent as random variables. It has been introduced by [Hubert and Arabie, 1985] and is closely related to the χ^2 distance between two distributions [Lancaster, 1969]. Another possible advantage of this distance, at least for theoretical analysis, is that it is a quadratic function in each of its arguments. From lemma 2 we have that $d^2_\chi(X, X') \leq \epsilon(\delta, \delta')/K$ whenever $\delta, \delta' \leq (K-1)/2$. This bound is tighter than the one in the subsequent theorem by virtue of making fewer approximations. Moreover, because the condition on ϵ is no longer necessary, it also holds for a much broader set of conditions (e.g larger perturbations away from the optimum) than the bound for d. Remembering also that the misclassification error has been criticized for becoming coarser as the clusterings become more dissimilar, we suggest that paying attention to the χ^2 distance will prove fruitful in theoretical and practical applications alike.

7.4 Regimes in clustering data

Let us return to the idea expressed in the introduction, of the existence of two regimes, "hard" and "easy" for the K-means optimization problem. Recent work by [Srebro et al., 2006] show experimental evidence for the existence of at least three regimes in clustering: the "hard" one where no clustering is known to be significantly better than the others, the "easy" one where clustering algorithms successfully find what is believed to be the best clustering, and an "interesting" regime where clustering algorithm do not seem to work well at minimizing the cost[7], but a good clustering may exist.

Our theoretical results together with the experiments suggest that the "easy" regime, the one where a good clustering can be found, may in turn contain two zones: the "high-confidence" one, where not only can we find a good clustering (in polynomial time), but we can also prove that we did so; outside this zone lies the "low-confidence" zone, where algorithms are still likely to find the optimal clustering with high probability, but one is not able to also prove that the obtained clustering is good.

[7]The cost function in [Srebro et al., 2006] is only slightly different from the quadratic distortion $\mathcal{D}$.

Acknowledgments

Thanks to Paul Tseng for substantially shortening the proof of theorem 1. The author was partially supported by NSF grant IIS-0313339.

Proofs

Proof of theorem 1

Using equation (13), the notation of (23) and $\Sigma = \mathrm{diag}\{\sigma_1, \ldots, \sigma_{K-1}\}$, $\Sigma_e = \mathrm{diag}\{\sigma_K, \ldots, \sigma_n\}$ we have that

$$\mathcal{D}(Y) - \mathcal{D}^* \;=\; \mathrm{tr}\,\Sigma - \mathrm{tr}\,[R^T \Sigma R + E^T \Sigma_e E] \tag{44}$$

We now construct the matrix A^0

$$A^0 \;=\; U^{all} \begin{bmatrix} \Sigma & \\ & \sigma I_{n-K+1} \end{bmatrix} U^{all} \;\text{ with } \sigma \in (\sigma_{K-1},\, \sigma_K)$$

If we replace A with A^0 in (14) the solution which depends only on the first $K-1$ eigenvalues/vectors of A, remains unchanged. Hence, we have

$$\begin{aligned} & U^T A^0 U - Y^T A^0 Y \\ &= \;\; \mathrm{tr}\,\Sigma - \mathrm{tr}\,[R^T \Sigma R + \sigma E^T E] \;\leq\; 0 \end{aligned} \tag{45}$$

Subtracting now (45) from (44) we obtain

$$\begin{aligned} & \mathcal{D}(Y) - \mathcal{D}^* \\ &\geq \;\; \mathrm{tr}\,[R^T \Sigma R + \sigma E^T E] - \mathrm{tr}\,[R^T \Sigma R + E^T \Sigma_e E] \\ &= \;\; \mathrm{tr}\, E^T (\sigma I - \Sigma_e) E \\ &\geq \;\; \mathrm{tr}\, E^T (\sigma I - \sigma_K I) E \\ &= \;\; (\sigma - \sigma_K) ||E||_F^2 \end{aligned} \tag{46}$$

The last inequality holds because $\sigma I - \Sigma_e \succeq (\sigma - \sigma_K) I \succeq 0$ for all σ in the chosen interval. Now, by taking the limit $\sigma \to \sigma_{K-1}$ in (46) we obtain

$$\mathcal{D}(Y) - \mathcal{D}^* \;\geq\; (\sigma_{K-1} - \sigma_K) ||E||_F^2 \tag{47}$$

From the above, whenever $\sigma_K - \sigma_{K-1}$ is nonzero, we obtain the desired result. QED.

Proof of lemma 2 Note first that since $A\mathbf{1} = 0$ we have $\mathbf{1} \perp U$ and therefore its normalized version $n^{-1/2}\mathbf{1} = U_e q$ where $q \in \mathbb{R}^{n-K+1}$ is a length 1 vector of coefficients.

Let X be a clustering, and c, V, Y be the same as in (11,12). Denote by V_- the first $K-1$ columns of V. We can write X as

$$\begin{aligned} X &= Y V_-^T + n^{-1/2} \mathbf{1} c^T \\ &= U R V_-^T + U_e E V_-^T + U_e q c^T \\ &= U R V_-^T + U_e (E V_-^T + q c^T) \end{aligned} \tag{48}$$

For a second clustering X' we define V', V'_-, c', R', E' similarly and have

$$X' \;=\; UR'V_-'^{\,T} + U_e(E'V_-'^{\,T} + q(c')^T).$$

We now calculate directly X^TX' and then $(X^TX')(X^TX')$, remembering that U, U_e and $[V_-\ c]$ $[V'_-\ c']$ represent pairs of orthogonal subspaces. After all the cancellations, we obtain the following formula for $\phi(X, X') = \mathrm{tr}\,(X^TX')(X^TX') = ||X^TX'||^2$

$$\begin{aligned}
&\mathrm{tr}\,(X^TX')(X^TX') \\
&\quad= K - 1 + 2\mathrm{tr}\,V'_-R'^TRE^T(q(c')^T + E'V'_-) \\
&\qquad +\mathrm{tr}\,(EE^T + qq^T)(E'E'^T + qq^T) \qquad (49)\\
&\quad= K - 1 + 2\mathrm{tr}\,R'^TRE^TE' + \mathrm{tr}\,(EE^TE'E'^T) \\
&\qquad +q^TE^TEq + q^TE'^TE'q + qq^T \qquad (50)\\
&\quad= K - 1 + 2\mathrm{tr}\,(RE^T)(E'R'^T) + \mathrm{tr}\,(EE^TE'E'^T) \\
&\qquad +0 + 0 + 1 \qquad (51)
\end{aligned}$$

To see that $E^Tq = E'^Tq = 0$ recall that $[R^T\ E^T]^T$ and $[0\ q]$ are respectively the coefficients of Y and $\mathbf{1}$ in the basis U^{all}. As $\mathbf{1} \perp Y$ it must hold that $[0\ q] \perp [R^T\ E^T]^T$ which implies $E^Tq = 0$.

We now try to lower bound (51). We lower bound the last term $\mathrm{tr}\,(EE^TE'E'^T)$ by 0. The middle term $\mathrm{tr}\,(RE^T)(E'R'^T)$ requires more work.

$$\begin{aligned}
|\mathrm{tr}\,(RE^T)(E'R'^T)| &= |< ER^T, E'R'^T >_F| \\
&\le ||ER^T||_F||E'R'^T||_F \qquad (52)
\end{aligned}$$

Furthermore,

$$\begin{aligned}
||ER^T||_F^2 &= \mathrm{tr}\,RE^TER^T \;=\; \mathrm{tr}\,E^TER^TR \\
&= \mathrm{tr}\,E^TE(I - E^TE) \\
&= \mathrm{tr}\,E^TE - \mathrm{tr}\,E^TEE^TE \\
&\le ||E||_F^2 - \frac{1}{K-1}||E||_F^4 \qquad (53)
\end{aligned}$$

The last inequality follows from lemma 10 stated below.

Now, because the function $x[1 - x/(K-1)]$ increases on $[0, (K-1)/2]$, we can combine (53) with $||E||_F^2 \le \delta$, $||E'||_F^2 \le \delta'$ and with (51) to obtain that $||X^TX'||^2 \ge K - 2\sqrt{\delta(1-\delta/(K-1))\delta'(1-\delta'/(K-1))}$. QED

Lemma 10 *For any matrix $A \in R^{m\times m}$, $||A^TA||_F \ge ||A||^2/m$.*

The proof is left to the reader.

Proof of Lemma 5 i) Let m be the centroid of the data, $m = \sum_i w_iz_i/W_{all} \in \mathbb{C}^d$. Then, the centered data points z_i^0 can be expressed as $z_i^0 = z_i - m$, or, in

matrix notation $Z_0 = Z - \mathbf{1}m^T = (I - \mathbf{1}\mathbf{w}^T/W_{all})Z$. It can easily be verified that $Z_0\mathbf{w} = 0$. Hence,

$$\begin{aligned}
L_0 &= \mathrm{diag}(\sqrt{\mathbf{w}})Z_0Z_0^*\mathrm{diag}(\sqrt{\mathbf{w}}) \qquad \text{(from (32))} \\
&= \mathrm{diag}(\sqrt{\mathbf{w}})(I - \mathbf{1}\mathbf{w}^T/W_{all})ZZ^*(I - \mathbf{1}\mathbf{w}^T/W_{all})^T\mathrm{diag}(\sqrt{\mathbf{w}}) \\
&= \mathrm{diag}(\sqrt{\mathbf{w}})(I - \mathbf{1}\mathbf{w}^T/W_{all})\mathrm{diag}(\mathbf{w})^{-1}\sqrt{S}\sqrt{S}^*\mathrm{diag}(\mathbf{w})^{-1} \\
&\quad \times(I - \mathbf{1}\mathbf{w}^T/W_{all})^T\mathrm{diag}(\sqrt{\mathbf{w}}) \qquad \text{(from (33))} \\
&= \mathrm{diag}(\sqrt{\mathbf{w}})(I - \mathbf{1}\mathbf{w}^T/W_{all})\mathrm{diag}(\sqrt{\mathbf{w}})^{-1}\underbrace{\mathrm{diag}(\sqrt{\mathbf{w}})^{-1}S\mathrm{diag}(\sqrt{\mathbf{w}})^{-1}}_{L} \\
&\quad \times\mathrm{diag}(\sqrt{\mathbf{w}})^{-1}(I - \mathbf{1}\mathbf{w}^T/W_{all})^T\mathrm{diag}(\sqrt{\mathbf{w}}) \\
&= (I - \sqrt{\mathbf{w}}\sqrt{\mathbf{w}}^T/W_{all})L(I - \sqrt{\mathbf{w}}\sqrt{\mathbf{w}}^T/W_{all})^T \qquad (54) \\
&= (I - B)L(I - B) \qquad (55)
\end{aligned}$$

ii) The matrix B above is symmetric, idempotent (i.e. $B^2 = BB^T = B$) and satisfies

$$\begin{aligned}
B\sqrt{\mathbf{w}} &= \sqrt{\mathbf{w}} \qquad (56) \\
Bu &= 0 \qquad \text{for all } u \perp \sqrt{\mathbf{w}} \qquad (57)
\end{aligned}$$

L is a symmetric real matrix, hence it has real eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n$ and real, orthogonal eigenvectors $u_1, \ldots u_n$. The largest eigenvalue of L has value $\lambda_1 = 1$ and its corresponding eigenvector is $u_1 = \sqrt{\mathbf{w}}/\sqrt{W_{all}}$ [Meilă and Shi, 2001].

Applying (55, 56) and (57) we obtain after some simple calculations

$$L_0u_j = \begin{cases} 0 & j = 1 \\ \lambda_j u_j & j > 1 \end{cases} \qquad (58)$$

□

iii) Equation (36) follows from (ii).

The distortion $\mathcal{D}$ is invariant to translations in the data Z and therefore to centering.

$$\begin{aligned}
\mathcal{D}(X) &= \mathrm{tr}\, L_0 - \mathrm{tr}\, X^T L_0 X \qquad (59) \\
&= \mathrm{tr}\, L_0 - \mathrm{tr}\, V[Y\ u_1]^T L_0 [Y\ u_1]V^T \qquad (60) \\
&= \mathrm{tr}\, L_0 - u_1^T L_0 u_1 - \mathrm{tr}\, Y^T L_0 Y \qquad (61) \\
&= \mathrm{tr}\, L_0 - \mathrm{tr}\, Y^T L_0 Y \qquad (62)
\end{aligned}$$

To obtain equation (37) it is sufficient to equate the right hand sides of (10) and (38).

□

Proof of Lemma 8

$$\begin{aligned}
\epsilon(\delta,\delta) &= 2\delta\left(1-\frac{\delta}{K-1}\right) && (63)\\
\epsilon^{old}(\delta,\delta) &= 2\delta\sqrt{(1-\delta)(K-\delta)}+(K+1)\delta-2\delta^2 && (64)\\
&= 2\delta(K-\delta)\left[\sqrt{\frac{1-\delta}{K-\delta}}+\frac{\frac{K+1}{2}-\delta}{K-\delta}\right] && (65)\\
&\geq 2\delta(1-\frac{\delta}{K-1})K\left[\sqrt{\frac{1-\delta}{K-\delta}}+\frac{\frac{K+1}{2}-\delta}{K-\delta}\right] && (66)\\
&= \epsilon(\delta,\delta)KF(\delta) && (67)
\end{aligned}$$

We show now that $F(x)\geq 1/2$ for all $x\in[0,1]$.

$$\begin{aligned}
F(x) &= \sqrt{\frac{1-x}{K-x}}+\frac{(K+1)/2-x}{K-x} && (68)\\
F'(x) &= \frac{1}{2\sqrt{\frac{1-x}{K-x}}}\frac{x-K-(x-1)}{(x-K)^2}+\frac{x-K-[x-(K+1)/2]}{(x-K)^2} && (69)\\
&= -\frac{K-1}{2(x-K)^2}\left(\sqrt{\frac{K-x}{1-x}}-1\right) \;\leq\; 0 \quad \text{for } x<1 && (70)
\end{aligned}$$

Hence, for $0\leq x\leq 1$

$$F(x) \;\geq\; F(1) \;=\; 0+\frac{(K+1)/2-1}{K-1} \;=\; \frac{1}{2} \qquad (71)$$

Proof of Theorem 9 i) It is easy to check that $B\sqrt{\mathbf{w}}=\sqrt{\mathbf{w}}$ and therefore $A\sqrt{\mathbf{w}}=(I-B)A_o(I-B)\sqrt{\mathbf{w}}=0$. To prove (42) it suffices to prove that $X^TA=X^TA_oX+constant$. To simplify notation, assume w.l.o.g. that $W_{all}=1$. Then $B=\sqrt{\mathbf{w}}\sqrt{\mathbf{w}}^T$ and the k-th column of X is $X_{:k}=\mathrm{diag}(\sqrt{\mathbf{w}})\tilde{X}_{:k}\mathrm{diag}(W_k^{-1/2},k=1:K)$. Hence $BX_{:k}=\sqrt{\mathbf{w}}\sqrt{W_k}$ and

$$\begin{aligned}
BX &= [\sqrt{\mathbf{w}}\;\sqrt{\mathbf{w}}\;\ldots\sqrt{\mathbf{w}}]\mathrm{diag}(W_k^{1/2}) && (72)\\
&= \mathrm{diag}(\sqrt{\mathbf{w}})[\,\mathbf{1}_n\;\mathbf{1}_n\ldots\mathbf{1}_n]\mathrm{diag}(W_k^{1/2}) && (73)\\
&= \mathrm{diag}(\sqrt{\mathbf{w}})\tilde{X}[\,\mathbf{1}_K\;\mathbf{1}_K\ldots\mathbf{1}_K]\mathrm{diag}(W_k^{1/2}) && (74)\\
&= \mathrm{diag}(\sqrt{\mathbf{w}})\tilde{X}\mathrm{diag}(W_k^{-1/2})\mathrm{diag}(W_k^{1/2})[\,\mathbf{1}_K\;\mathbf{1}_K\ldots\mathbf{1}_K]\mathrm{diag}(W_k^{1/2}) && (75)\\
&= X\underbrace{[\sqrt{W_kW_{k'}}]_{k,k'=1:K}}_{B'} \;=\; XB' && (76)
\end{aligned}$$

Then,

$$\begin{aligned}
\operatorname{tr} X^T A X &= \operatorname{tr}(I - B')X^T A_o X (I - B') && (77)\\
&= \operatorname{tr}(I - B')^2 X^T A_o X && (78)\\
&= \operatorname{tr}(I - B')X^T A_o X \ \text{ (because) } (B')^2 = B' && (79)\\
&= \operatorname{tr} X^T A_o X + \operatorname{tr} B' X^T A_o X && (80)\\
\operatorname{tr} B' X^T A_o X &= \operatorname{tr} \sqrt{[W_k]_{k=1:K}} \sqrt{[W_k]_{k=1:K}}^T X^T A_o X && (81)\\
&= \sqrt{[W_k]_{k=1:K}}^T X^T A_o \underbrace{X \sqrt{[W_k]_{k=1:K}}}_{\sqrt{\mathbf{w}}} && (82)\\
&= \sqrt{\mathbf{w}}^T A_o \sqrt{\mathbf{w}} && (83)
\end{aligned}$$

In the above, $[W_k]_{k=1:K}$ represents the column vector of cluster weights. Replacing the last equation into (80) we obtain the desired result.

ii) Since $A_0\sqrt{\mathbf{w}} = 0$, this part is proved in the same way as (14) in section 2.2.

iii) The proof follows closely the proof of Theorem 4 and is therefore omitted. □

References

[Achlioptas and McSherry, 2005] Achlioptas, D. and McSherry, F. (2005). On spectral learning of mixtures of distributions. In Auer, P. and Meir, R., editors, *18th Annual Conference on Learning Theory, COLT 2005*, pages 458–471, Berlin/Heidelberg. Springer.

[Bach and Jordan, 2004] Bach, F. and Jordan, M. I. (2004). Learning spectral clustering. In Thrun, S. and Saul, L., editors, *Advances in Neural Information Processing Systems 16,*, Cambridge, MA. MIT Press.

[Ben-David et al., 2006] Ben-David, S., von Luxburg, U., and Pal, D. (2006). A sober look at clustering stability. In *19th Annual Conference on Learning Theory, COLT 2006*. Springer.

[Brucker, 1978] Brucker, P. (1978). On the complexity of clustering algorithms. In Henn, R., Corte, B., and Oletti, W., editors, *Optimierung und Operations Research*, Lecture Notes in Economics and Mathematical Systems, pages 44–55. Springer Verlag, New York, NY.

[Dasgupta, 1999] Dasgupta, S. (1999). Learning mixtures of gaussians. In *FOCS '99: Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, page 634, Washington, DC, USA. IEEE Computer Society.

[Dhillon et al., 2004] Dhillon, I. S., Guan, Y., and Kulis, B. (2004). Kernel K-means, spectral clustering and normalized cuts. In Kohavi, R., Gehrke, J., and Ghosh, J., editors, *Proceedings of The Tenth ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining(KDD), pages 551–556. ACM Press.

[Ding and He, 2004] Ding, C. and He, X. (2004). K-means clustering via principal component analysis. In Brodley, C. E., editor, *Proceedings of the International Machine Learning Conference (ICML)*. Morgan Kauffman.

[Hubert and Arabie, 1985] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218.

[Lancaster, 1969] Lancaster, H. (1969). *The Chi-Squared Distribution*. Wiley.

[Meila and Xu, 2003] Meila, M. and Xu, L. (2003). Multiway cuts and spectral clustering. Technical Report 442, University of Washington.

[Meilă, 2002] Meilă, M. (2002). The multicut lemma. Technical Report 417, University of Washington.

[Meilă, 2005] Meilă, M. (2005). Comparing clusterings – an axiomatic view. In Wrobel, S. and De Raedt, L., editors, *Proceedings of the International Machine Learning Conference (ICML)*. ACM Press.

[Meilă, 2011] Meilă, M. (2011). Local equivalence of distances between clusterings – a geometric perspective. *Machine Learning*, 86(3):369–389.

[Meilă and Shi, 2001] Meilă, M. and Shi, J. (2001). A random walks view of spectral segmentation. In Jaakkola, T. and Richardson, T., editors, *Artificial Intelligence and Statistics AISTATS*.

[Meilă et al., 2005] Meilă, M., Shortreed, S., and Xu, L. (2005). Regularized spectral learning. In Cowell, R. and Ghahramani, Z., editors, *Proceedings of the Artificial Intelligence and Statistics Workshop(AISTATS 05)*.

[Ng et al., 2002] Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA. MIT Press.

[Papadimitriou and Steiglitz, 1998] Papadimitriou, C. and Steiglitz, K. (1998). *Combinatorial optimization. Algorithms and complexity*. Dover Publication, Inc., Minneola, NY.

[Rakhlin and Caponnetto, 2006] Rakhlin, A. and Caponnetto, A. (2006). Stability of k-means clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1121–1128, Cambridge,MA. MIT Press.

[Ravindran et al., 2005] Ravindran, K., Salmasian, H., and Vempala, S. (2005). The spectral method for general mixture models. In Auer, P. and Meir, R., editors, *18th Annual Conference on Learning Theory, COLT 2005*, pages 444–457, Berlin/Heidelberg. Springer.

[Schölkopf et al., 1998] Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Non-linear analysis as a kernel eigenvalue problem. *Neural computation*, 10:1299–1319. kernel k-means.

[Shi and Malik, 2000] Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *PAMI*.

[Srebro et al., 2006] Srebro, N., Shakhnarovich, G., and Roweis, S. (2006). An investigation of computational and informational limits in gaussian mixture clustering. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*.

[Vapnik, 1998] Vapnik, V. (1998). *Statistical Learning Theory*. Wiley.

[Vempala and Wang, 2004] Vempala, S. and Wang, G. (2004). A spectral algorithm for learning mixture models. *J. Comput. Syst. Sci.*, 68(4):841–860.

[von Luxburg et al., 2005] von Luxburg, U., Bousquet, O., and Belkin, M. (2005). Limits of spectral clustering. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, number 17. MIT Press.

[Yu and Shi, 2003] Yu, S. X. and Shi, J. (2003). Multiclass spectral clustering. In *International Conference on Computer Vision*.