The Uniqueness of a Good Optimum for K-Means

Marina Meilă MMP@STAT.WASHINGTON.EDU

University of Washington, Department of Statistics, Box 354322, Seattle, WA 98195-4322 USA

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. When "goodness" is measured by the distortion of K-means clustering, this paper proves spectral bounds on the distance $d(\mathcal{C}, \mathcal{C}^{opt})$. The bounds exist in the case when the data admits a low distortion clustering.

1. Motivation

Optimizing clustering criteria like the minimum squared error of K-means clustering is theoretically NP-hard. 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. There is no reason to believe that the second 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 "cluster of near-optimal clusterings". To our knowledge, this is the first stability result for the K-means optimization problem. The terms "good" and "close" are defined in the next section, 2, which also introduces the rest of the terminology and notation. 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 and an extended discussion in section 5.

2. Definitions and Representations

A *clustering* $\mathcal{C}$ of a finite dataset, assumed w.l.o.g to be $\{1, 2, \ldots, n\} = [n]$, 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 $n_k = |C_k|$, $\sum_k n_k = n$. A clustering can be represented by a $n \times K$ matrix $\tilde{X}$, whose columns represent the indicator vectors of the the K clusters.

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

The columns of $\tilde{X}$ are mutually orthogonal vectors. If we normalize these to length 1, we obtained the *normalized* representation X.

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

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 $d(X, X') = d(\tilde{X}, \tilde{X}')$.

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.

Appearing in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006. Copyright 2006 by the author(s)/owner(s).

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

$$d(\tilde{X}, \tilde{X}') \;=\; 1 - \frac{1}{n} \max_{\pi \in \Pi_K} \sum_k m_{k,\pi(k)} \qquad (3)$$

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

2.2. The K-Means Clustering Objective

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 S be the Gram matrix given by $S_{ij} = z_i^T z_j$ or $S = ZZ^T$. 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 S will have rank at most d. The *squared error distortion*, often called "K-means" cost function, is defined as

$$\mathcal{D}(X) \;=\; \sum_{k=1}^{K} \sum_{i \in C_k} ||z_i - \mu_k||^2 \qquad (4)$$

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

$$\mu_k \;=\; \frac{1}{n_k} \sum_{i \in C_k} z_i, \quad \text{for } k = 1, \ldots K \qquad (5)$$

If one substitutes the expression of the centers (5) into (4) 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 & He, 2004)

$$\mathcal{D}(X) \;=\; \operatorname{tr} S - \operatorname{tr} X^T S X \qquad (6)$$

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

$$c \;=\; \left[\sqrt{\frac{n_1}{n}} \cdots \sqrt{\frac{n_k}{n}} \cdots \sqrt{\frac{n_K}{n}} \right]^T \qquad (7)$$

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

$$XV \;=\; \left[Y \;\; \mathbf{1} \frac{1}{\sqrt{n}} \right]. \qquad (8)$$

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

$$\begin{aligned} \mathcal{D}(Y) &= \mathit{constant} - \operatorname{tr} \left[Y \, \mathbf{1} \frac{1}{\sqrt{n}} \right]^T S \left[Y \, \mathbf{1} \frac{1}{\sqrt{n}} \right] \\ &= \mathit{constant} - \operatorname{tr} Y^T S Y - \frac{1}{n} \mathbf{1}^T S \mathbf{1} \\ &= \mathit{constant} - \operatorname{tr} Y^T S Y \qquad (9) \end{aligned}$$

The last equality holds because $S\mathbf{1} = ZZ^T\mathbf{1} = 0$. It has been noted (Ding & 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,

$$\mathcal{D}(Y) \;\geq\; \mathcal{D}^* \;=\; \operatorname{tr} S - \sum_{k=1}^{K-1} \sigma_k, \;\; \text{attained for } \; Y \;=\; U \qquad (10)$$

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

3. The Main Result

We call *good* a K-clustering whose distortion $\mathcal{D}(X)$ is not too large compared to the optimum $\mathcal{D}^*$, that is $\mathcal{D}(X) - \mathcal{D}^* \leq \epsilon$, for an ϵ to be determined. Let X^{opt} be the K-clustering of S 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 principal subspace U of S. 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.

3.1. A Bound on the "Error Subspace" Projection

Now we will show that any clustering which has a distortion close enough to $\mathcal{D}^*$ will be close to the $(K-1)$-th principal subspace of S. Let Y be a clustering with a corresponding c defined as in (7); Y can be written as

$$Y \;=\; [U\; U_e] \begin{bmatrix} R \\ E \end{bmatrix} \tag{11}$$

$U^{all} = [U \quad U_e] \in \mathbb{R}^{n\times n}$ is the orthogonal basis represented by the eigenvectors of S 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 now show that if $\mathcal{D}(Y)$ is small enough, then E is "almost 0".

Theorem 1 *For any clustering Y represented like in (11) the following inequality holds*

$$||E||_F^2 \;\le\; \delta \;=\; \frac{\mathcal{D}(Y) - \mathcal{D}^*}{\sigma_{K-1} - \sigma_K} \tag{12}$$

The proof is given in the appendix.

3.2. Any Two Clusterings with Small Distortion are Close

We have proved so far that the projection of the matrix Y on the subspace U_e is bounded by the r.h.s of (12). We will now show that two clusterings Y, Y' for which this bound is small must be close to each other. First we show that a certain function $\phi(X, X')$ taking values in $[0, K]$ is large 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.

$$\phi(X, X') \;=\; ||X^T X'||_F^2 \tag{13}$$

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

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

Lemma 2 *For any two clusterings X, X' denote by δ, respectively δ' the corresponding values of the r.h.s term of (12). For $\delta, \delta' \le K/2$*

$$\phi(X, X') \;\ge\; K - \epsilon(\delta, \delta') \tag{14}$$

with

$$\epsilon(\delta, \delta') \;=\; 2\sqrt{\delta\delta'(1 - \delta/(K-1))(1 - \delta'/(K-1))} \tag{15}$$

This lemma is proved in the appendix.

Theorem 3 *(Meilă, 2006) For two clusterings with K clusters each, if $\phi(X, X') \ge K - \epsilon$, $\epsilon \le p_{min}$ then $d_{ME}(Y, Y') \le \epsilon p_{max}$, where $p_{max} = \max_k n_k/n$, $p_{min} = \min_k n_k/n$.*

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 following corollary, which is the result we have been striving for.

3.3. A Low Distortion Clustering is Close to the Optimal Clustering

Corrollary 4 *Let X be any clustering of a data set represented by the Gram matrix $S = [z_i^T z_j]_{i,j=1}^n$. Let the $p_{max} = \max_k n_k/n$ $p_{min} = \min_k n_k/n$, δ be given by (12) and ϵ by (15). Then, if $\delta \le (K-1)/2$ and $\epsilon(\delta, \delta) \le p_{min}$ then*

$$d(X, X^{opt}) \;\le\; \epsilon(\delta, \delta) p_{max} \tag{16}$$

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

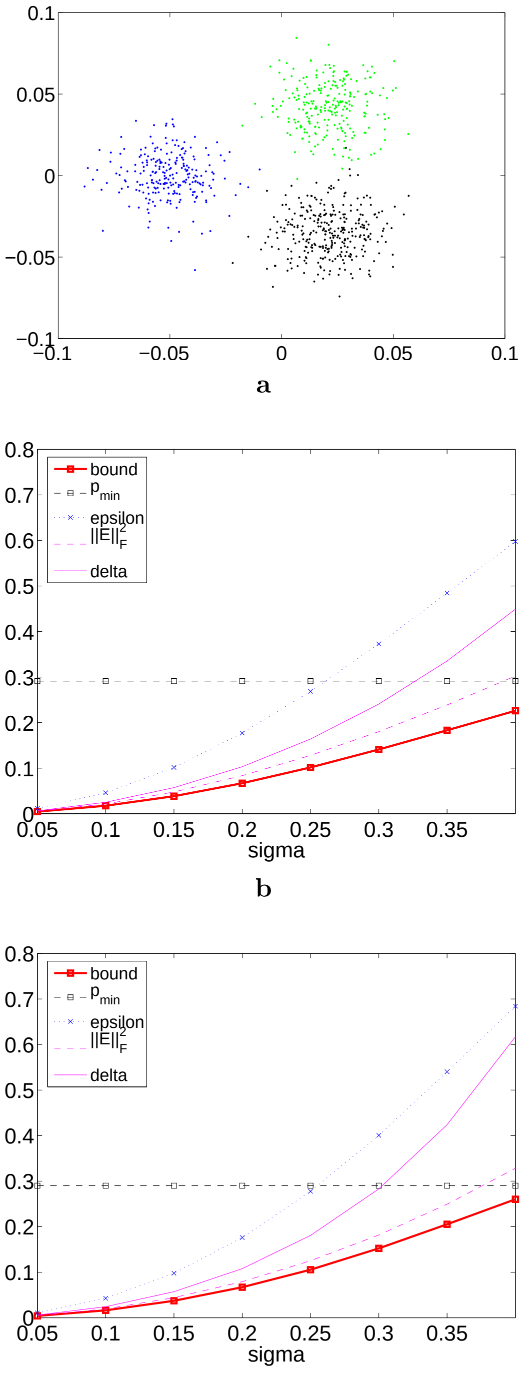
Proof: We know that $\mathcal{D}(Y^{opt}) \le \mathcal{D}(Y)$ and hence $||E^{opt}||_F^2 \le \delta$ from theorem 1. By applying lemma 2 and theorem 3 we obtain the desired result. QED

A few remarks are in place. First, the bound 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 below. Second, from the corollary it follows that $d(X, X^{opt}) \le p_{min} p_{max} \le p_{min}$. Hence, for p_{max} not too large, the bound is a good bound, it tells us that all clusters in X^{opt} have been identified.

It should be also noted that the condition $\epsilon \le p_{min}$ 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 are ever informative. In the experiments illustrated by Figure 1 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

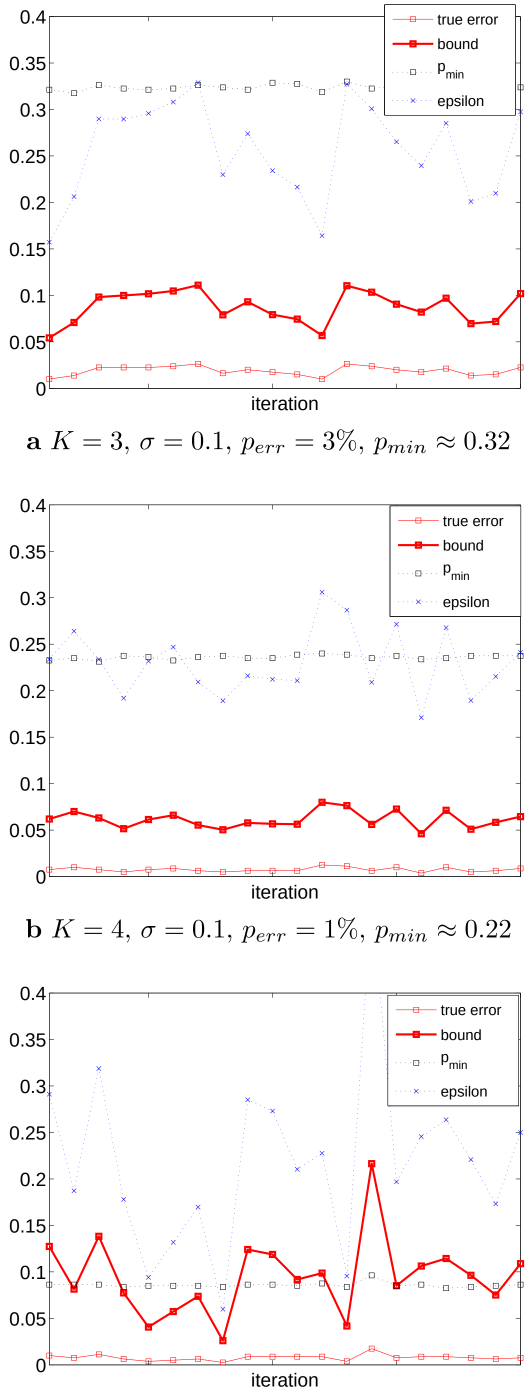


a

b

c

Figure 1. 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 $\sigma^2 I_d$; **(a)** shows the data for $\sigma = 0.4$, projected on its second principal subspace. The true mixture labels are shown in different colors. The clustering X represents the K-means solution. In **(b)**, the bound and the values of p_{min}, ϵ, $||E||_F^2, \delta$ for X are evaluated at different values of σ; the data set has size $n = 1000$. **(c)** Same as (b) for $n = 100$.



a $K = 3$, $\sigma = 0.1$, $p_{err} = 3\%$, $p_{min} \approx 0.32$

b $K = 4$, $\sigma = 0.1$, $p_{err} = 1\%$, $p_{min} \approx 0.22$

c $K = 4$, $\sigma = 0.02$, $p_{err} = 1\%$, $p_{min} \approx 0.08$

Figure 2. 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 clustering, 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.

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 3.1) and its upper bound δ from (12). 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 1,b and 1,c shows that there is practically no variation due to the data set size. This is consistent with the theory and with all our other experiments so far.

Figure 2 shows a different experiment. Here the optimal clustering[1] X is perturbed randomly into X'. We evaluate the true misclassification error $d(X, X')$ and its bound, together with other relevant quantities. Here $K = 3, 4$, and for $K = 4$ we have a uniform and a non-uniform clustering. Note that the bound becomes looser with increasing K, and for the same K with the decreasing size of the smallest cluster. For instance, in Figure 2, c, less than half of the clusterings have valid bounds[2]. The degradation with decreasing p_{min} is completely expected, based for example on the condition $\epsilon \leq p_{min}$ in theorem 3. It 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 we have that the clustering we have found is optimal, even when it is.

5. 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 $\mu_1, \dots \mu_K$ is parallel to the K principal components of the data[3] 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 $\mu_1, \dots \mu_K,\ K > 2$ 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 only 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 X representation lies close to the principal subspace U. Our result says 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.

To our knowledge, this result is the first of this kind for the K-means distortion. There is however a large body of work, pioneered by (Dasgupta, 1999), dealing with estimating mixtures of normal distributions with high probability, by using projections on a subspace of lower dimension, in particular the $K-1$-th principal subspace (Vempala & Wang, 2004). These papers offer algorithms for fitting a mixture of (sufficiently round) normals with known K to data, plus guarantees that the estimates will be close to the truth with high probability. From their results, bounds on the "distortion" (i.e log-likelihood) could be derived. Our paper does not offer an algorithm[4] except for the practical ob-

[1] We assume X to be represented by the true labels, which is extremely plausible as the clusters are well-separated.

[2] In a, b, all the bounds have been checked individually and are valid, even if ϵ is occasionally larger than p_{min}.

[3] The matrix $S = ZZ^T$ and the (scaled) covariance matrix Z^TZ have the same non-zero eigenvalues; U is the projection of the data on the principal subspace.

[4] The associated algorithm could be however the spec-

servation that K-means works well when the clusters are well separated. We have not conducted quantitative comparisons, but we expect that the bounds offered by the mixture papers hold for less well separated data than ours. On the other hand, one should notice the fundamental difference between this line of work and ours: the former relies heavily on the gaussianity of the clusters (or on slightly weaker knowledge); our work is *model free* in the sense that it makes no assumption about the data distribution. By analogy with empirical risk minimization, it connects the observed distortion to the best possible distortion. In ERM, the bound is sometimes greater than 1, here the bound doesn't always exist (but is usually informative when it does, as ϵp_{max} is usually small in comparison to p_{min}). It is also a *worst case* bound.

Spectral methods for K-means and graph clustering have been popular in recent years, and this work builds on previous results, especially on (Ding & He, 2004) and (Meilă et al., 2005). But, for the purpose of this paper, beyond theorem 1 the mathematical tools had to be developed anew. Unlike previous work in spectral methods, e.g. (Meilă et al., 2005), our bounds depend, besides the distortions and the eigengap, on information about the clustering itself: p_{min}, p_{max}, the projection of Y on U.

Extensions. Our result relies on the quadratic form of $\mathcal{D}(X)$ from (6). As such, it applies to any clustering criterion that is quadratic in X. Such settings include weighted K-means clustering, kernel K-means, and their combinations.

An alternative distance between clusterings. The function $\phi(X, X')$ used as an intermediary vehicle for the proof of corollary 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 & 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.

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. 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.

Finally, theorem 3 is a *stability* result. Recent work, e.g (Lange et al., 2004; Ben-Hur et al., 2002), uses the stability of a clustering as a criterion for model selection. Therefore we are currently investigating the use of our result in the selection of K.

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 (9), the notation of (11) and $\Sigma = \mathrm{diag}\{\sigma_1, \ldots, \sigma_{K-1}\}$, $\Sigma_e = \mathrm{diag}\{\sigma_K, \ldots, \sigma_n\}$ we have that

$$\mathcal{D}(Y) - \mathcal{D}^* \;=\; \mathrm{tr}\,\Sigma - \mathrm{tr}\,[R^T\Sigma R + E^T\Sigma_e E] \qquad (17)$$

We now construct the matrix S^0

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

If we replace S with S^0 in (10) the solution which depends only on the first $K-1$ eigenvalues/vectors of S, remains unchanged. Hence, we have

$$\begin{aligned} & US^0U - Y^TS^0Y \\ &\quad = \;\; \mathrm{tr}\,\Sigma - \mathrm{tr}\,[R^T\Sigma R + \sigma E^TE] \;\leq\; 0 \qquad (18)\end{aligned}$$

Subtracting now (18) from (17) we obtain

$$\begin{aligned} & \mathcal{D}(Y) - \mathcal{D}^* \\ &\quad \geq \;\; \mathrm{tr}\,[R^T\Sigma R + \sigma E^TE] - \mathrm{tr}\,[R^T\Sigma R + E^T\Sigma_e E]\end{aligned}$$

tral algorithm of (Ding & He, 2004).

$$\begin{aligned} &= \operatorname{tr} E^T(\sigma I - \Sigma_e)E \\ &\geq \operatorname{tr} E^T(\sigma I - \sigma_K I)E \\ &= (\sigma - \sigma_K)||E||_F^2 \qquad (19) \end{aligned}$$

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 (19) we obtain

$$\mathcal{D}(Y) - \mathcal{D}^* \;\geq\; (\sigma_{K-1} - \sigma_K)||E||_F^2 \qquad (20)$$

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 $S\mathbf{1} = 0$ 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 (7,8). Denote by V_- the first $K-1$ columns of V. We can write X as

$$\begin{aligned} X &= YV_-^T + n^{-1/2}\mathbf{1}c^T \\ &= URV_-^T + U_eEV_-^T + U_eqc^T \\ &= URV_-^T + U_e(EV_-^T + qc^T) \qquad (21) \end{aligned}$$

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 $\phi(X, X') = \operatorname{tr}(X^TX')(X^TX') = ||X^TX'||^2$

$$\begin{aligned} &\operatorname{tr}(X^TX')(X^TX') \\ &\quad= K - 1 + 2\operatorname{tr} V'_-R'^TRE^T(q(c')^T + E'V'_-) \\ &\qquad+\operatorname{tr}(EE^T + qq^T)(E'E'^T + qq^T) \qquad (22) \\ &\quad= K - 1 + 2\operatorname{tr} R'^TRE^TE' + \operatorname{tr}(EE^TE'E'^T) \\ &\qquad+q^TE^TEq + q^TE'^TE'q + qq^T \qquad (23) \\ &\quad= K - 1 + 2\operatorname{tr}(RE^T)(E'R'^T) + \operatorname{tr}(EE^TE'E'^T) \\ &\qquad+0 + 0 + 1 \qquad (24) \end{aligned}$$

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 (24). We lower bound the last term $\operatorname{tr}(EE^TE'E'^T)$ by 0. The middle term $\operatorname{tr}(RE^T)(E'R'^T)$ requires more work.

$$\begin{aligned} |\operatorname{tr}(RE^T)(E'R'^T)| &= |< ER^T, E'R'^T >_F| \\ &\leq ||ER^T||_F||E'R'^T||_F \qquad (25) \end{aligned}$$

Furthermore,

$$\begin{aligned} ||ER^T||_F^2 &= \operatorname{tr} RE^TER^T \;=\; \operatorname{tr} E^TER^TR \\ &= \operatorname{tr} E^TE(I - E^TE) \\ &= \operatorname{tr} E^TE - \operatorname{tr} E^TEE^TE \\ &\leq ||E||_F^2 - \frac{1}{K-1}||E||_F^4 \qquad (26) \end{aligned}$$

The last inequality follows from lemma 5 stated below.

Now, because the function $x[1 - x/(K-1)]$ increases on $[0, (K-1)/2]$, we can combine (26) with $||E||_F^2 \leq \delta$, $||E'||_F^2 \leq \delta'$ and with (24) to obtain that $||X^TX'||^2 \geq K - 2\sqrt{\delta(1-\delta/(K-1))\delta'(1-\delta'/(K-1))}$. QED

Lemma 5 *For any matrix* $A \in R^{n\times n}$, $||A^TA||_F \geq ||A||^2/n$.

The proof is left to the reader.

References

Ben-Hur, A., Elisseeff, A., & Guyon, I. (2002). A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing* (pp. 6–17).

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

Ding, C., & He, X. (2004). K-means clustering via principal component analysis. *Proceedings of the International Machine Learning Conference (ICML)*. Morgan Kauffman.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

Lancaster, H. (1969). *The Chi-squared distribution*. Wiley.

Lange, T., Roth, V., Braun, M. L., & Buhmann, J. M. (2004). Stability-based validation of clustering solutions. *Neural Comput.*, *16*, 1299–1323.

Meilă, M. (2005). Comparing clusterings – an axiomatic view. *Proceedings of the International Machine Learning Conference (ICML)*. ACM Press.

Meilă, M. (2006). The local equivalence of two distances between finite random variables: the misclassification error metric and the χ^2 distance. *(submitted)*.

Meilă, M., Shortreed, S., & Xu, L. (2005). Regularized spectral learning. *Proceedings of the Artificial Intelligence and Statistics Workshop(AISTATS 05).*

Papadimitriou, C., & Steiglitz, K. (1998). *Combinatorial optimization. algorithms and complexity.* Minneola, NY: Dover Publication, Inc.

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