Journal of Multivariate Analysis 173 (2019) 1–17

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Good (K-means) clusterings are unique (up to small perturbations)

Marina Meilă *

Department of Statistics, University of Washington, Box 345322, Seattle, WA 98195-4322, USA

HIGHLIGHTS

- Cluster validation with no assumptions about the data generating process.
- Guarantees that given clustering is unique/optimal up to small perturbations.
- The guarantees can be computed in practice for any given clustering.
- The guarantees apply to K-means and spectral clustering cost functions, among others.
- For practitioners: tractable way to confirm that a clustering is uniquely supported by data.

ARTICLE INFO

Article history:
Received 15 February 2018
Received in revised form 23 December 2018
Accepted 23 December 2018
Available online 11 February 2019

AMS subject classificaition:
primary 62H30
secondary 91C20
15A18
15B57

Keywords:
K-means clustering
Spectral clustering
Cluster validation
Model free
Clusterability

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 $\mathcal{C}^{\text{opt}}$ of these data? Perhaps surprisingly, the answer to this question is sometimes yes. This paper gives spectral bounds on the distance $d(\mathcal{C}, \mathcal{C}^{\text{opt}})$ 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 only if the data admit a "good", low-cost clustering. The results in this paper are non-asymptotic and model-free, in the sense that no assumptions are made on the data generating process. The bounds do not depend on undefined constants, and can be computed tractably from the data.

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 [9,18,32]. 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.

* Corresponding author.
E-mail address: mmp@stat.washington.edu.

https://doi.org/10.1016/j.jmva.2018.12.008
0047-259X/

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 data set, 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 $d(\mathcal{C}, \mathcal{C}^{\text{opt}})$ between a given clustering $\mathcal{C}$ and the (unknown) optimal clustering $\mathcal{C}^{\text{opt}}$ of the given data. The bounds 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 $\mathcal{C}$.

Section 2 introduces the terminology and notation, defines the K-means and the NCut cost functions, and gives a precise meaning to the terms "good" and "close". 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. In Section 4 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 5. We discuss the related work in Section 6 and present experiments on synthetic and real data in Sections 7 and 8, respectively. The extended discussion in Section 9 compares our paradigm with other directions of research on the theoretical foundations of clustering. 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 data set of size n is a partition of the indices $\{1, \ldots, n\}$ into disjoint, nonempty subsets called clusters. If the partition has K clusters, we write $\mathcal{C} = \{C_1, \ldots, C_K\}$, set $n_1 = |C_1|, \ldots, n_K = |C_K|$ and $n_1 + \cdots + n_K = n$. If the data points have weights $w_1 > 0, \ldots, w_n > 0$, then the cluster sizes become cluster weights

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

and the total weight of the data is $W_{\text{all}} = w_1 + \cdots + w_n$. The weighted case reduces to the unweighted one when $w_1 = \cdots = w_n = 1$.

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

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

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, viz.

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

In the case of unweighted data, i.e., $w_1 = \cdots = w_n = 1$, the normalized representation becomes

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

In the future, we will refer to a clustering by any of its matrix representations. As we will typically work with two clusterings, one will be denoted by $\tilde{X}$ (respectively X) and 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 $\mathcal{C} = \{C_1, \ldots, C_K\}$ and $\mathcal{C}' = \{C'_1, \ldots, C'_{K'}\}$ 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 $M = \tilde{X}^\top \tilde{X}'$. 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

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

This distance represents the well-known cost of classification, minimized over the set Π_K of all permutations of the labels $1, \ldots, K$. Although the maximization is over a set of size $K!$, d can be computed in polynomial time by a maximum bipartite matching algorithm [28]. This distance is widely used, having very appealing properties as long as X and X' are close [21].

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

$$M^{\mathbf{w}} = \tilde{X}^\top \mathrm{diag}(\mathbf{w}) \tilde{X},$$

and the weighted misclassification error is written as

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

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 $A_{ij} = z_i^\top z_j$ or $A = ZZ^\top$. We will assume without loss of generality that the data are centered at the origin, i.e., $\sum_i z_i = 0$ or, in matrix notation $\mathbf{1}^\top Z = 0$. Therefore, Z and A will have rank at most d. The squared error distortion, often called "K-means" cost function, is defined as

$$\mathcal{D}(X) = \sum_{k=1}^{K} \sum_{i \in C_k} \|z_i - \mu_k\|^2. \tag{3}$$

In the above, $\mu_1, \ldots, \mu_K$ are the clusters' centers, whose coordinates in $\mathbb{R}^d$ are given, for all $k \in \{1, \ldots, K\}$, by

$$\mu_k = \frac{1}{n_k} \sum_{i \in C_k} z_i. \tag{4}$$

If one substitutes the expression of the centers (4) into (3), and if one represents a clustering by the orthonormal column matrix X defined above, one can show that the distortion is a quadratic function of X [12], viz.

$$\mathcal{D}(X) = \mathrm{tr}\, A - \mathrm{tr}\, X^\top A X. \tag{5}$$

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

$$c = \left(\sqrt{\frac{n_1}{n}} \cdots \sqrt{\frac{n_k}{n}} \cdots \sqrt{\frac{n_K}{n}} \right)^\top \tag{6}$$

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 1s. Denote

$$XV = \left(Y \;\; \mathbf{1}\frac{1}{\sqrt{n}} \right). \tag{7}$$

We can now rewrite the distortion (3) in terms of the $n \times (K-1)$ matrix Y, viz.

$$\mathrm{tr}\, A - \mathrm{tr} \left(Y \; \mathbf{1}\frac{1}{\sqrt{n}} \right)^\top A \left(Y \; \mathbf{1}\frac{1}{\sqrt{n}} \right) = \mathrm{tr}\, A - \mathrm{tr}\, Y^\top A Y - \frac{1}{n} \mathbf{1}^\top A \mathbf{1} = \mathrm{tr}\, A - \mathrm{tr}\, Y^\top A Y \;\equiv\; \mathcal{D}(Y). \tag{8}$$

In the above, with a slight abuse of notation, we identify Y with X and write $\mathcal{D}(Y)$ for $\mathcal{D}(YV^\top)$. The last equality holds because $A\mathbf{1} = ZZ^\top \mathbf{1} = 0$. It has been noted [12] that relaxing the integrality constraints in the above equation results in a trace maximization problem that is solved by an eigendecomposition, viz.

$$\mathcal{D}^* = \operatorname*{argmin}_{Y \in \mathbb{R}^{n \times (K-1)}, \text{orthogonal}} \mathcal{D}(Y) = \mathrm{tr}\, A - \sum_{k=1}^{K-1} \sigma_k = \mathcal{D}(U), \tag{9}$$

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. Hence, we have that for any clustering X represented by Y as above, $\mathcal{D}(X) \geq \mathcal{D}^*$.

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 $V = \{1, \ldots, n\}$. 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

$$w_i = \sum_{j \in V} S_{ij}.$$

Without loss of generality, 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 [20,40] is

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

where

$$\mathrm{Cut}(A, B) = \sum_{i\in A} \sum_{j\in B} S_{ij}.$$

It is known from [18,40] 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

$$\mathrm{NCut}(\mathcal{C}) = K - \mathrm{tr}\, X^\top L X, \tag{10}$$

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

$$L = \mathrm{diag}(\mathbf{w})^{-1/2} S \mathrm{diag}(\mathbf{w})^{-1/2}. \tag{11}$$

By a reasoning similar to the one leading to Eq. (9), one can show [18] that for any clustering X,

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

where U is the $n \times K$ matrix containing the principal eigenvectors of L and $\lambda_1 \geq \cdots \geq \lambda_n$ 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 (11) 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., $Z^\top \mathbf{1} = 0$. 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 $\mathrm{NCut}(X)$, 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 $d(X, X^{\mathrm{opt}})$ 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 3 *Let X be any clustering of a data set represented by the Gram matrix $A = (z_i^\top z_j)_{i,j=1}^n$. If $\delta = \{\mathcal{D}(X) - \mathcal{D}^*\}/(\sigma_{K-1} - \sigma_K)$ is sufficiently small, then $d(X, X^{\mathrm{opt}}) \leq \mathrm{bound}(\delta, X)$, 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.

> **Imprecise version of** Corollary 1 *Let X be any clustering of a data set represented by the symmetric similarity matrix $S = (S_{ij})$ with $S_{ij} \geq 0$. Let the vector of node weights be $\mathbf{w} = (w_i)$ and let $W_1, \ldots, W_K$ be defined as in (1). If $\delta = \{\mathrm{NCut}(X) - \mathcal{N}^*\}/(\lambda_K - \lambda_{K+1})$ is sufficiently small, then $d^{\mathbf{w}}(X, X^{\mathrm{opt}}) \leq \mathrm{bound}'(\delta, X)$, where X^{opt} represents the clustering with K clusters that minimizes the K-way Normalized Cut on the data S.*

In the above, $\mathcal{D}(X)$, $\mathcal{D}^*$, σ_{K-1}, σ_K are defined as in Section 2.2 and $\mathcal{N}(X)$, $\mathcal{N}^*$, λ_{K+1}, λ_K are defined as in Section 2.3. The exact expression of the functions $\mathrm{bound}(\delta, X)$, $\mathrm{bound}'(\delta, X)$ and the other technical conditions for which these inequalities hold are given in Theorem 3 and Corollary 1, respectively. Theorem 5 in Section 5 is a further generalization that includes as special cases both Theorem 3 and Corollary 1.

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 $\mathcal{D}^*$, i.e., $\mathcal{D}(X) - \mathcal{D}^* \leq \epsilon$, for an ϵ to be determined. Let X^{opt} be the K-clustering of A with the smallest distortion and note that $\mathcal{D}(X) \geq \mathcal{D}(X^{\mathrm{opt}}) \geq \mathcal{D}^*$. 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 $d(X, X^{\mathrm{opt}})$.

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 $(K-1)$st 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 (6); Y can be written as

$$Y = [\, U \; U_e \,] \begin{bmatrix} R \\ E \end{bmatrix}, \tag{12}$$

where $U^{\text{all}} = [U \;\; U_e] \in \mathbb{R}^{n\times n}$ 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, $[R^\top E^\top]^\top$ 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 (12), the following inequality holds:*

$$\|E\|_F^2 \le \delta = \frac{\mathcal{D}(Y) - \mathcal{D}^*}{\sigma_{K-1} - \sigma_K}. \tag{13}$$

By $\|\cdot\|_F$ we denote the Frobenius norm of a matrix, $\|M\|_F^2 = \operatorname{tr} M^\top M$. 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:

$$\phi(X, X') = \|X^\top X'\|_F^2. \tag{14}$$

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

$$0 \le \phi(X, X') = \|X^\top X'\|_F^2 \;\le\; \|X\|_F \|X'\|_F = K.$$

Lemma 1. *For any two clusterings X, X' denote by δ, respectively δ' the corresponding values of the right-hand side term of (13). For $\delta, \delta' \le (K-1)/2$, $\phi(X, X') \ge K - \epsilon(\delta, \delta')$ 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 2 (*After [22]*)**.** *For two weighted clusterings with K clusters each, if $\phi(X, X') \ge K - \epsilon, \epsilon \le p_{\min}$ then $d_{ME}^{\mathbf{w}}(X, X') \le \epsilon p_{\max}$, where $p_{\max} = \max_k W_k/W_{\text{all}}$, $p_{\min} = \min_k W_k/W_{\text{all}}$.*

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 3. *Let X be any clustering of a data set represented by the Gram matrix $A = (z_i^\top z_j)_{i,j=1}^n$, with $z_1 + \cdots + z_n = 0$. Let $p_{\max} = \max_k n_k/n$, $p_{\min} = \min_k n_k/n$, let δ be given by (13) and ϵ by (15). If $\delta \le (K-1)/2$ and $\epsilon(\delta, \delta) \le p_{\min}$, then $d(X, X^{\text{opt}}) \le \epsilon(\delta, \delta) p_{\max}$, where X^{opt} represents the clustering with K clusters that minimizes the distortion $\mathcal{D}$ on the data A.*

Proof. We know that $\mathcal{D}(Y^{\text{opt}}) \le \mathcal{D}(Y)$ and hence $\|E^{\text{opt}}\|_F^2 \le \delta$ from Theorem 1. By applying Lemma 1 and Theorem 2 we obtain the desired result. □

A few remarks are in order. 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 Sections 7 and 8. Second, Theorem 3 implies that $d(X, X^{\text{opt}}) \le p_{\min} p_{\max} \le p_{\min}$. Hence, for $p_{\max}$ not too large, the bound is informative, guaranteeing that all clusters in $\mathcal{C}^{\text{opt}}$ have been identified. It should be also noted that the condition $\epsilon \le p_{\min}$ in Theorem 2 is only sufficient, not necessary.

4. Extension to weighted data and the Normalized Cut cost

Extending Theorem 3 to weighted and kernel-based distortion functions is immediate. Assume that the data points are weighted with weights $\mathbf{w} = (w_1, \ldots, w_n)$. The weighted distortion is defined as

$$\mathcal{D}^{\mathbf{w}}(\mathcal{C}) = \min_{\mu_1,\ldots,\mu_K \in \mathbb{R}^d} \sum_k \sum_{i \in C_k} w_i \|z_i - \mu_k\|^2. \tag{16}$$

It can be easily checked that the centroids $\mu_1, \ldots, \mu_K$ that minimize the above expression for Z and $\mathcal{C}$ fixed are the weighted means of the data in each cluster. That is, for all $k \in \{1, \ldots, K\}$,

$$\mu_k = \sum_{i \in C_k} w_i z_i / W_k.$$

By replacing the above values in (16) we obtain after some calculations

$$\mathcal{D}^{\mathbf{w}}(\mathcal{C}) = \operatorname{tr} A - \operatorname{tr} X^\top A X, \tag{17}$$

with X defined as in and

$$A = \operatorname{diag}(\sqrt{\mathbf{w}}) Z Z^\top \operatorname{diag}(\sqrt{\mathbf{w}}). \tag{18}$$

An important application of Theorem 2 for weighted data is to the problem of graph partitioning with the Normalized Cut cost.

By comparing the quadratic representation of the NCut criterion (10)–(11) and of the weighted distortion (17)–(18), 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 same clustering for a mapping of the graph nodes $i \in \{1, \dots, n\}$ into d-dimensional vectors $z_1, \dots, z_n$; this fact was noted by [2] and used by [11] for the special case S positive definite. To find the mapping we set $A = L$ and obtain

$$Z = \operatorname{diag}(\mathbf{w})^{-1} \sqrt{S} \tag{19}$$

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 the matrix M. With the mapping Z as in (19), we have, for all X,

$$\operatorname{NCut}(X) = \mathcal{D}^{\mathbf{w}}(X) - \operatorname{tr} L + K.$$

Because for any clustering $\operatorname{NCut}(X)$ differs from $\mathcal{D}(X)$ by a constant independent of X, we can use Theorem 2 in order to obtain an analog for partitions 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 3 is assumed to be obtained from centered data. In the following lemma we show how to evaluate directly 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 (19). Let $Z_0 = Z - \mathbf{1} m^\top$ denote the embedded points shifted by the vector m, so that $Z_0^\top \mathbf{w} = 0$. That is, Z_0 represents the centered data. Let L_0 be the "centered L" matrix, i.e., the matrix obtained by applying the right-hand side of (18) to Z_0. Note that although Z, m, Z_0 may be complex, L, L_0 are always real and symmetric matrices.

Lemma 2. *Let n, $\mathbf{w}$, L, L_0 be defined as above. Let $\lambda_1 = 1 \geq \cdots \geq \lambda_n$ be the eigenvalues of L and $u_1, \dots, u_n$ be the corresponding eigenvectors. Then*

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

(ii) The eigenvalues and eigenvectors of L_0 are

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

(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; this decomposition is not possible in general, but it can be verified that it is always possible when X represents a clustering. Then

$$\operatorname{tr} L = \operatorname{tr} L_0 + 1, \tag{20}$$

$$\operatorname{tr} X^\top L X = \operatorname{tr} Y^\top L_0 Y + 1, \tag{21}$$

and

$$\mathcal{D}(X) = \operatorname{tr} L_0 - \operatorname{tr} Y^\top L_0 Y. \tag{22}$$

We can now apply Theorem 1 to the distortion expressed as in (22). If we take into account Lemma 2 and we assume in addition that

$$\lambda_{K+1} \geq 0, \tag{23}$$

we obtain

$$\delta = \frac{\lambda_2 + \cdots + \lambda_K - \operatorname{tr} Y^\top L_0 Y}{\lambda_K - \lambda_{K+1}} = \frac{1 + \lambda_2 + \cdots + \lambda_K - \operatorname{tr} X^\top L X}{\lambda_K - \lambda_{K+1}}. \tag{24}$$

Assumption (23) is often verified in practice. If it is true, then the $K-1$ largest eigenvalues of L_0 are $\lambda_2, \dots, \lambda_K$ and its $(K-1)$st eigengap is $\lambda_K - \lambda_{K+1}$. If (23) does not hold, then the modification of the bound in Eq. (24) 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.

Corollary 1. *Let X be any clustering of a data set represented by the symmetric similarity matrix $S = (S_{ij})$ with $S_{ij} \geq 0$. Let the vector of node degrees be $\mathbf{w} = (w_i)$ with $w_i > 0$, $W_1, \dots, W_K$ be defined as in (1), $p_{\max} = \max_k W_k / W_{\text{all}}$, $p_{\min} = \min_k W_k / W_{\text{all}}$; let δ be given by (24) and ϵ by (15). Assume $\lambda_{K+1} \geq 0$, where λ_{K+1} is the $(K+1)$st eigenvalue of $L = \operatorname{diag}(\mathbf{w})^{-1/2} S \operatorname{diag}(\mathbf{w})^{-1/2}$. Then, if $\delta \leq (K-1)/2$ and $\epsilon(\delta, \delta) \leq p_{\min}$, we have $d^{\mathbf{w}}(X, X^{\text{opt}}) \leq \epsilon(\delta, \delta) p_{\max}$, 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 [24], which we reproduce here.

Theorem 4 (*After [24], 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', $\lambda_1, \ldots, \lambda_{K+1}$ be defined as in Corollary 1 and let the function $\phi(X, X')$ be defined by (14). Then, whenever $\delta \leq 1$, $\phi(X, X') \geq K - \epsilon^{\text{old}}(\delta, \delta')$ with $\epsilon^{\text{old}}(\delta, \delta') = 2\sqrt{\delta\delta'(1-\delta)(K-\delta')} + K\delta + \delta' - 2\delta\delta'$.*

There are some slight differences in the requirements of the above theorem versus Corollary 1. 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. In contrast, assumption (23) is not necessary. We remind the reader that assumption (23) is a simplifying assumption 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 3. *$\epsilon^{\text{old}}(\delta, \delta) \geq K\epsilon(\delta, \delta)/2$ for all $\delta \leq 1$.*

Hence, the new bound improves the result of [24].

5. General quadratic cost function

Theorem 5. *Let $\mathcal{D}^{\mathbf{w}}$ be any clustering cost function that can be expressed in the form $\mathcal{D}^{\mathbf{w}}(X) = C_0 - \operatorname{tr} X^\top A_0 X$, where X is a (weighted) clustering defined as in , and $C_0 \in \mathbb{R}$, $A_0 \in \mathbb{R}^{n\times n}$ symmetric depend only on the data and on the data weights $\mathbf{w} = (w_i)$ with $w_i > 0$. Define $W_{\text{all}} = \sum_i w_i$, $W_1, \ldots, W_K$ as in (1), $p_{\max} = \max_k W_k/W_{\text{all}}$, $p_{\min} = \min_k W_k/W_{\text{all}}$.*

(i) Let $B = \sqrt{\mathbf{w}}\sqrt{\mathbf{w}}^\top/W_{\text{all}}$ and $A = (I-B)A_0(I-B)$. Then, for any clustering X, $A\sqrt{\mathbf{w}} = 0$ and $\mathcal{D}^{\mathbf{w}}(X) = C - \operatorname{tr} X^\top AX$ 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 (13) and ϵ by (15). Then, if $\delta \leq (K-1)/2$ and $\epsilon(\delta, \delta) \leq p_{\min}$, we have $d^{\mathbf{w}}(X, X^{\text{opt}}) \leq \epsilon(\delta, \delta)p_{\max}$, 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 Cut [32].

For the Average Cut, we have $\mathbf{w} = \mathbf{1}/n$ and $\mathcal{D}^{\mathbf{w}}(X) = -\operatorname{tr} X^\top SX$, where S is the graph similarity matrix defined in Section 2.3. In kernel K-means (see [31] 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 $\kappa(z_i, z_j) = h_i^\top h_j$. The Gram matrix A_0 is redefined to be

$$A_0 = [\kappa(z_i, z_j)]_{i,j=1}^n. \tag{25}$$

The kernel K-means clustering cost function is the distortion with respect to $\mathcal{H}$. It is easy to see that with A_0 defined as in (25) the distortion takes the same form as in Theorem 5.

In all cases presented in this paper, computing the bounds ϵ, δ requires computing the principal K or $K+1$ eigenvalues of a symmetric $n \times n$ matrix. This operation is of order n^2K, hence obtaining the bounds is tractable whenever spectral clustering is tractable for a data set.

6. Existing work on model-free guarantees for clustering

This paper expands and strengthens earlier work by the author in [19,24]. For K-means clustering, Ostrovsky et al. [27] give an algorithm with verifiable model-free guarantees. However, the conditions become too restrictive when $K > 2$, e.g., they do not hold on the data in Figs. 1–2. We are not aware of other model-free guarantees for clustering by K-means.

In the area of *graph partitioning* this problem has been studied more. Wan and Meilă [39] gave model-free guarantees for community detection in networks. They considered two classes of models, the Stochastic Block Model and the more general Preference Frame Model [37] and for each, they derived spectral bounds.

We also mention [6], which guarantees approximate and tractable recovery assuming A is close to a DC-SBM. The assumptions are testable given a clustering $\mathcal{C}$. This result is one of many in the clustering literature which present tractable algorithms to find a good $\mathcal{C}$, under the assumption that one with specified "robustness" exists. We will discuss this area of research below, as well as in Section 9.2.

Spectral graph partitioning is the area with most model-free guarantees. In the paper of Peng et al. [29], Theorem 1.2 states that if the K-way Cheeger constant of the graph is $\rho(K) \geq \lambda_{K+1}(L)/(cK^3)$, then the Spectral Clustering algorithm

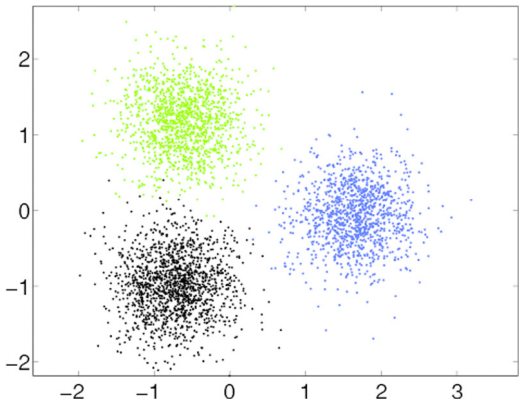


Fig. 1. A mixture of three normal distributions in $d = 35$ dimensions, with fixed centers and equal covariances $\sigma^2 I_d$, $\sigma = 0.4$, projected on its second principal subspace. The true mixture labels are shown in different colors.

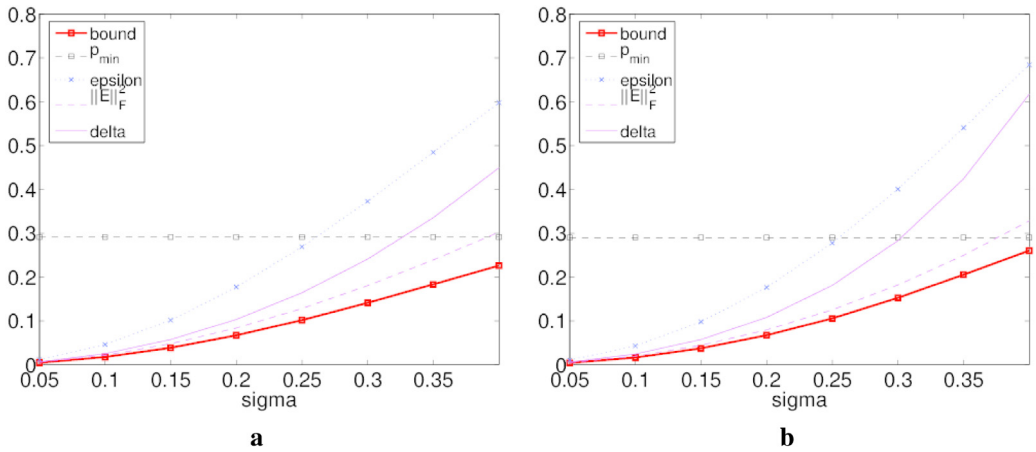


Fig. 2. The bound used as a certificate of correctness. The data represent a mixture of three normal distributions in $d = 35$ dimensions, with fixed centers and equal covariances $\sigma^2 I_d$; these data are depicted in Fig. 1 for $\sigma = 0.4$. The clustering X represents the K-means solution. In **(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 $n = 100$.

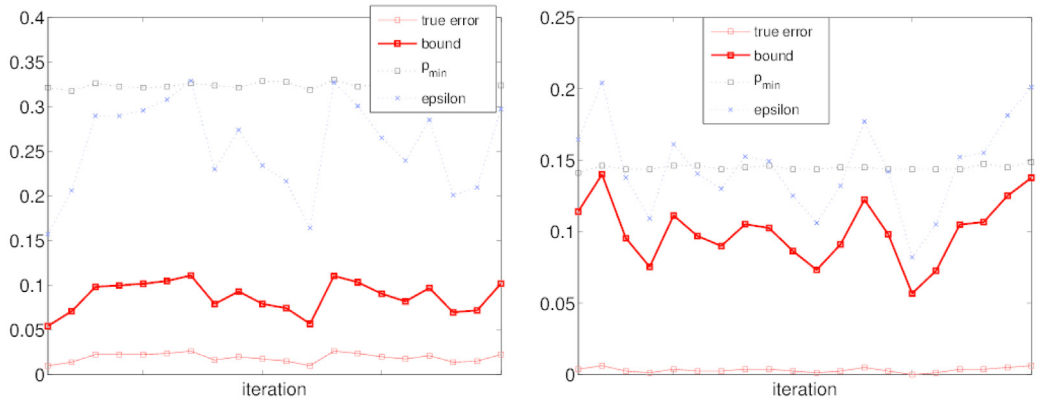
outputs $\mathcal{C}$ with $d^{\mathbf{w}}(\mathcal{C}, \mathcal{C}^{\text{opt}}) \leq C/c$ with $C = 2 \times 10^5, c \sim 1/K^3$. Since the distance $d^{\mathbf{w}}$ cannot exceed 1, while the right-hand side $C/c \ll 1$, this bound is too loose to be informative. In [4,26], guarantees are given when the matrix A is nearly block diagonal. The conditions in these papers are extremely restrictive in their applicability. This was studied by Wan and Meilă [38], who generated a set of weighted graphs on which spectral clustering could recover the original clustering. Then, it was verified whether the guarantees of [4,26] applied to these data; the experiments of [38] extended to several other recovery theorems for clustering in graphs. The outcome was uniformly negative: guarantees could not be obtained in even a single of the cases tested. In the experiments of Section 7.2, we show that with Corollary 1 we can obtain very tight bounds on the same examples.

Other significant results in graph partitioning by cuts. The work of Lee et al. [17] establishes a relationship between the kth Cheeger constant of a graph and the eigenvalues of the Laplacian of the graph. The aforementioned Cheeger constant is equal to the NCut whenever no cluster is larger than half the total weight W_V. Of interest to the present work is Theorem 4.10, which relates the existence of good r-way partitioning, where $r \geq K - 3K\delta$, to a large eigengap. More precisely, if $\lambda_{K+K\delta}/\lambda_K > c(\ln K)^2/\delta^9$ then the above partition is "better" than $\lambda_K/\delta^3 \times c'$ (for c, c' unspecified). While these results are remarkable for their generality, the previous theorem requires extremely large $\lambda_{K+K\delta}/\lambda_K$ to produce non-trivial bounds, no matter what c, c' are.

7. Experiments

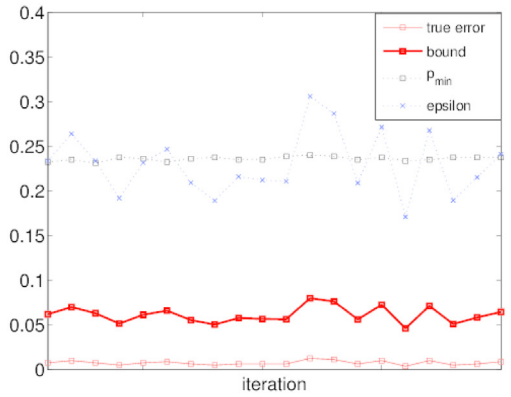
7.1. Experiments with the K-means distortion

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 illustrated by Fig. 2, we generated data from a mixture of spherical normal distributions,

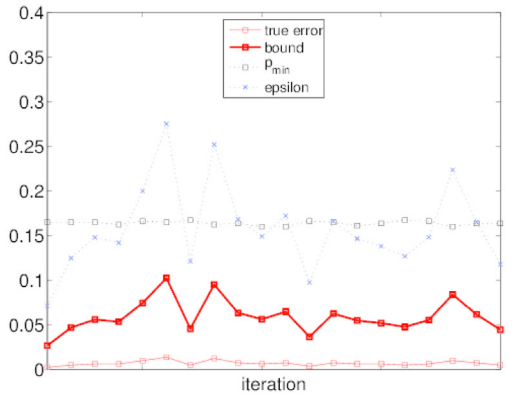


a $K = 3,\ \sigma = 0.1,\ p_{\text{err}} = 3\%,\ p_{\min} \approx 0.32$
$\|\mu_k - \mu_{k'}\| \in [2.3, 2.5]$

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



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



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

Fig. 3. The data represent 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 with probability 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.

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 $d(X, X^{\text{opt}})$ 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 3 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 (13). 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 Fig. 2a and 2b 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.

Fig. 3 shows a different experiment. Here the optimal clustering X is perturbed randomly into X'; we assume X to be represented by the true labels, which is extremely plausible as the clusters are well separated. We evaluate the true misclassification error $d(X, X')$ and its bound, together with other relevant quantities for $K \in \{3, 4\}$, each with a uniform and a non-uniform clustering. Note that the bound becomes looser when $d(X', X^{\text{opt}})$ and K increase, or when $p_{\min}$ decreases. For instance, in Fig. 3d, more than half of the clusterings have invalid bounds. Hence, the figures demonstrate both the informativeness of the bounds and the limitations of their applicability.

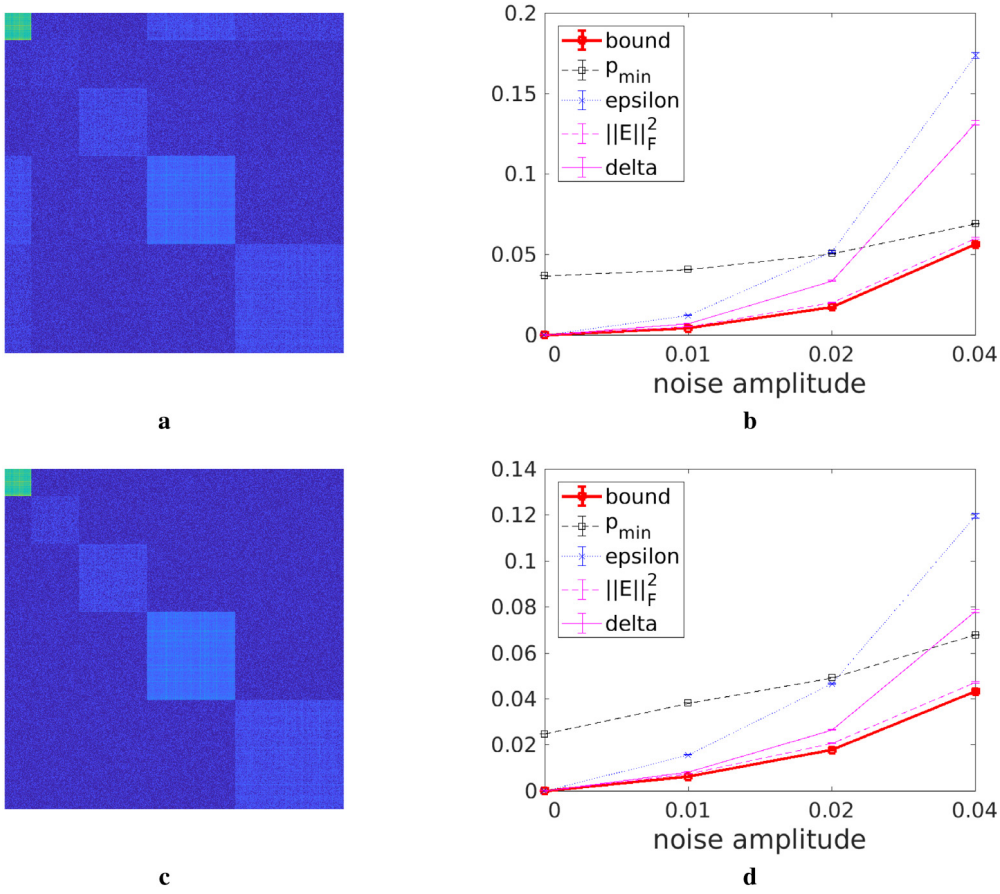


Fig. 4. Certificates of correctness for spectral clustering of a weighted graph. The data **(a)** represent a similarity matrix S with elements in $[0, 1]$, which were further perturbed by symmetric, zero-mean uniform iid noise with amplitude 0.04; for better contrast, the plot displays $\sqrt{S_{ij}}$ instead of S_{ij}. In **(b)**, the bound and the values of $p_{\min}$, ϵ, $\|E\|_F^2$, and δ are evaluated at different values of the noise amplitude; the plot shows mean and standard deviations over five replications. The data set has size $n = 3000$. In **(c)** and **(d)** the experiment is repeated with an easier, almost block diagonal, S matrix.

The degradation with decreasing $p_{\min}$ is completely expected, based for example on the condition $\epsilon \leq p_{\min}$ in Theorem 2. 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.

7.2. Experiments with the Normalized Cut cost

In the first set of experiments, the data consist of symmetric, non-negative similarity matrices S of dimension $n = 3000$ with entries in $[0, 1]$. These matrices were used in [38], which compared experimentally the existing theoretical conditions in [4,26]. They were generated in a way that guaranteed that they would have an optimal clustering $\mathcal{C}^{\text{opt}}$ easy to find by the Spectral Clustering algorithm, e.g., the matrices in Fig. 4a, c. Then the conditions required in [4,26] were computed for these data. The aforementioned conditions are provably sufficient to guarantee that the Spectral Clustering algorithm will find an almost optimal $\mathcal{C}$. The question posed by the experiment was whether these conditions are also (close to being) necessary. In other words, if we generate cases where Spectral Clustering works well but the aforementioned theorems do not hold, it means that the state of knowledge in the area as a whole is far from understanding what is possible. This is what [38] showed: none of the theorems of [4,26] covered any of the test cases.

We applied Corollary 1 to the S matrices used in [38] for $n = 3000$. These were matrices with $K = 5$ clusters, where the eigengaps varied between 0.99 (almost block diagonal matrix) and 0.02, the cluster sizes were either equal or unequal (see Fig. 4) and the degree distribution in each cluster varied from uniform to highly non-uniform (as in Fig. 4). In all $3 \times 2 \times 3$ cases, the bound from Corollary 1 was below 10^{-3}, and much smaller than $0.1 \approx p_{\min}$. Hence, even though the conditions in this paper are sufficient but not necessary, they represent a major improvement over the state of the art.

Next, two of the matrices S from [38] were further perturbed by zero-mean uniform iid noise in each entry, where the noise amplitude s varied between 0.01 and 0.04, ensuring that no weight becomes negative. The case $s = 0$ corresponds to

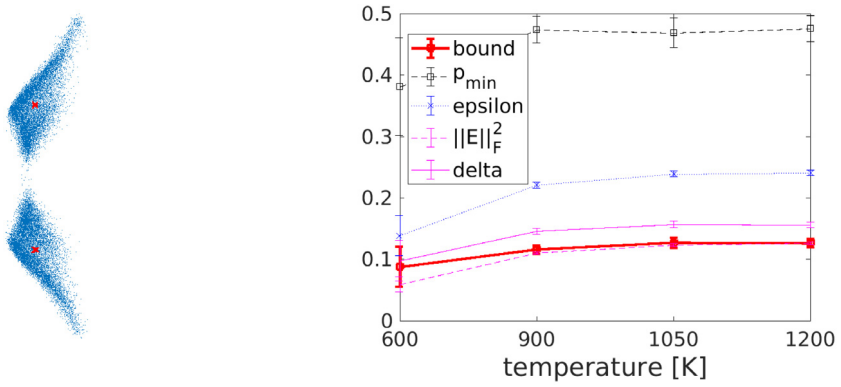


Fig. 5. Certificates of correctness for labeling the data from molecular dynamics simulations of the system described by Eq. (26). The data in **(a)** represent one simulation trajectory at 1200 Kelvin, containing $n = 23{,}566$ points. The preprocessed data are 12-dimensional, and they are displayed here in the plane of the first two principal components. The centroids obtained after K-means clustering with $K = 2$ are also displayed. In **(b)**, the bound and the values of $p_{\min}$, ϵ, $\|E\|_F^2$, δ for the obtained clustering are evaluated at different values of the noise amplitude; the plot shows mean and standard deviations over the 16 trajectories simulated at each temperature.

the original data from [38]. The results of this experiment are depicted in Fig. 4. Note that in the right panels, the values are computed on the weighted data points, as described in Section 4. Hence, although the smallest cluster contains 10% of the data points ($n_1 = 300$ points), when the points are weighted by their degree, the mass $p_{\min}$ amounts to only 5% (respectively 2%) of the total. The first of the matrices, depicted in Fig. 4a, is very far from being block diagonal. In fact, for the second largest cluster in this weighted graph, the intra-cluster connections have less weight than the connections with nodes outside the cluster. Yet, a good clustering exists in this S and Corollary 1 is able to guarantee its uniqueness. The second matrix, in Fig. 4c, is a nearly block diagonal matrix. It is included because the other existing guarantees for Spectral Clustering are taylored to this type of matrices.

8. An experiment on molecular simulation data

In Fig. 5, Theorem 3 is used to obtain guarantees for chemical simulation data. The data are obtained from a molecular dynamics simulation [13] of the reversible reaction

$$CH_3C\ell + C\ell^- \;\leftrightarrow\; CH_3C\ell + C\ell^-, \tag{26}$$

in which one of the chlorine ($C\ell$) atoms replaces the other in the methylchloride molecule. The simulation output represents a sequence of x, y, z spatial locations of the six atoms involved in the reaction at consecutive time steps during the simulation. When a chlorine atom binds to the methyl (CH_3) group, the energy of the system is lower; in the sequence of configurations in which the reaction takes place, and both chlorine atoms are dissociated, the energy of the system is higher, hence fewer configurations will be present in these regions. Therefore, molecular simulations of a chemical reaction will exhibit clusters, whose sizes will be roughly dependent on the potential energy of the respective energy wells. In this reaction, due to symmetry, the cluster sizes will be approximately equal if the simulation is long enough to allow several transitions from one cluster to the other, which is the case in our data. The density between the two clusters will depend on the absolute temperature T of the system, with lower density at lower temperatures. Our data, available at https://www.stat.washington.edu/spectral/data/MDsimulations2017/, consist of 16 simulated trajectories at each of the four temperatures $T \in \{600, 900, 1050, 1200\}$ degrees Kelvin. The length of the trajectories varies around $n = 6250$ in the $600K$ simulations, and around $n = 23{,}400$ in the other simulations.

Because the energy of the molecule depends only on the relative positions of the atoms, the original $6 \times 3 = 18$ dimensions are reduced to 12 degrees of freedom in the following way. First, all the plane angles within the molecule are computed, for a total of $\binom{6}{3} = 20$ angles. Then, the linear relationships between these are eliminated by Principal Component Analysis, keeping up to 12 components, or enough to reduce the residual variance to 10^{-4}. With the present data, the resulting dimension was always 12. These data are now clustered by K-means with $K = 2$. Fig. 5 shows that the data distribution in each cluster is non-Gaussian, non-symmetric around the centroids, and heavy-tailed.

The purpose of clustering is to label the data by energy well; this in turn allows chemists to filter out the states around the transitions. They form a very small fraction of the data; their study enables chemists to understand the conditions under which the reaction occurs. Currently, the labeling is done by ad hoc algorithms. Having guarantees of (almost) correctness for the grouping, such as those in Fig. 5 saves the time needed to validate the clustering by human inspection.

9. Discussion

This paper proves that if (i) the data are well clustered and (ii) 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, it provides a user with a certificate that the clustering X at hand 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 Z. The matrix $A = ZZ^\top$ 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. 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 $\mu_1, \ldots, \mu_K$ with $K > 2$ are along a line, no matter how well separated the clusters are, 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., [2,11,12,20,26,40]. This 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 are well clustered, $\mathcal{D}(X)$ has a unique "deep crater". When points are moved to other clusters with respect to 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.

9.1. Related work: Probably correct algorithms for mixtures

While Section 6 discusses the existing model-free guarantees for clustering, here we describe work under a different paradigm, and mainly published in theoretical computer science. This work is concerned with guarantees for clustering under the assumptions that the data are sampled from a finite mixture model.

This area was pioneered by [10] 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 [35] shows that by projecting a mixture of spherical Gaussians on the $(K-1)$st 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 [1,15] results in algorithms for 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 classes of results. The papers in this section offer polynomial algorithms for clustering, plus guarantees that the output will be 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 with high probability which means that the aforementioned algorithms guarantee $d(X, X^{\text{opt}}) = 0$ with high probability. 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 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, namely 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 [12] or of the variant of EM with PC projection used by Srebro et al. [33] as associated algorithms. Alternatively, one can use our results after algorithms such as that in [35] as a certificate of correctness for the found clustering. This would allow such algorithms to be run with a lower confidence parameter, i.e., a larger δ.

Beyond the differences in posing the problem, there is also a fundamental similarity between our paper and the spectral embedding methods of [1,15] and especially of [35] that we discuss now. A crucial fact proved in [35] is that for a sufficiently large n and for well separated clusters, the Kth principal subspace of the data and the K dimensional subspace determined by the cluster centroids are close. Hence, the Kth principal subspace of the data contains information about the best clustering. Our result exploits the same fact, as Theorem 3 holding implicitly means that all good clusterings, when represented as subspaces, are close to the $(K-1)$st principal subspace of the data (the difference of 1 in the dimensions comes from centering the data). Hence, both groups of results rely on the informativity of the principal subspace with respect to a salient clustering in the data. In [1], the same fact is exploited, 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 in [1] is based on the distance preserving property of the projection on the principal subspace.

We also mention the results of Balakrishnan et al. [3], who analyzed the fixed points of the EM algorithm for models with hidden variables, and their basins of attraction, giving sufficient conditions for the basin to have radius R in parameter space.

One can also draw an analogy between our result and the VC type bounds for structural risk minimization (SRM); see [34]. 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 3 gives a bound on the error of an obtained clustering with respect to 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/eigengap(A) and increases when either component increases. It is known that the $(K-1)$st eigengap of a matrix measures the stability of its $(K-1)$st 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 does not always exist but is always informative when it does.

9.2. Related stability results

As mentioned before, Theorem 2 is a stability result. Here we discuss stability results of a different type, i.e., under a different paradigm, that have been obtained in the literature.

For spectral clustering with the Normalized Cut cost, Ben-David et al. [7] build on previous work by von Luxburg et al. [36] and Rakhlin and Caponnetto [30]. We briefly summarize [7]. 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 are 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 that (i) P has a unique minimizer, and (ii) 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, the present paper proves the necessary prerequisites for the main theorem in [7] to hold. The previous statement can be made more precise: our Theorem 5 proves that for a quadratic cost, under certain verifiable conditions, a distribution P with finite support has a unique minimizer, while the definition in [7] refers to general probability spaces. To bridge the gap, it remains to take the limit $n \to \infty$ in Theorem 5. We regard this as possibly subject to some distributional assumptions about the data, but as beyond the scope of the present paper.

One can think of [7] as proving that if in the limit of $n \to \infty$ all almost optimal clusterings are similar, and if one has an algorithm $\mathcal{A}$ which always produces an almost optimal clustering with respect to the cost, then the output of algorithm $\mathcal{A}$ 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.

In the literature, a data set that is well clustered is often called clusterable, and various definitions of clusterability have been proposed. The data properties we analyze here, namely admitting a good clustering (as in the conditions of Theorem 2 as well as 5), and the uniqueness (up to perturbations) of such a good clustering, can be considered as ways to define clusterability. The latter notion of clusterability was used under the name "uniqueness of optimum" in [5,25]. An alternative popular notion of clusterability is that of (weak) perturbation resilience introduced in [8]. In [5,25] it is shown that if a data set contains a clustering satisfying Theorem 2 then this clustering is weakly perturbation resilient. Hence, the theorems in this paper are the first tractable way to prove any form of perturbation resilience.

9.3. An alternative distance between clusterings

The function $\phi(X, X')$, defined by (14), and used as an intermediary vehicle for the proof of Theorem 3, can in fact represent a distance in its own right. Denote $d^2_\chi(X, X') = 1 - \phi(X, X')/\min(K, K')$. 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 [14] and is closely related to the χ^2 distance between two distributions [16]. 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 1 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.

9.4. 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. The experimental work by Srebro et al. [33] 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 does not seem to work well at minimizing the cost (the cost function in [33] is only slightly different from the quadratic distortion $\mathcal{D}$), but a good clustering may exist. The "easy" regime delimited empirically in [33] contains realistic, non-trivial data sets, and extends well beyond the current theoretical results for clustering with mixtures.

Our theoretical and experimental results 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 we can not only 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 still find the optimal clustering with high probability, but we cannot prove that they did. Fortunately, as the experiments in Sections 7–8 demonstrate, the easy and high-confidence regime extends to realistic data and covers practically relevant applications.

Acknowledgments

Thanks to Paul Tseng for substantially shortening the proof of Theorem 1. Christopher Fu generated the molecular dynamics simulation data studied in Section 8. The author was partially supported by National Science Foundation, USA awards IIS-0313339 and DMS-1810975.

Appendix

Proof of Theorem 1. Using Eq. (8), the notation of (12) and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_{K-1})$, $\Sigma_e = \mathrm{diag}(\sigma_K, \ldots, \sigma_n)$ we have that

$$\mathcal{D}(Y) - \mathcal{D}^* = \mathrm{tr}\,\Sigma - \mathrm{tr}\,(R^\top \Sigma R + E^\top \Sigma_e E).$$

We now construct the matrix

$$A^0 = U^{\mathrm{all}} \begin{bmatrix} \Sigma & \\ & \sigma I_{n-K+1} \end{bmatrix} U^{\mathrm{all}},$$

with $\sigma \in (\sigma_{K-1}, \sigma_K)$. If we replace A with A^0 in (9), the solution, which depends only on the first $K-1$ eigenvalues/vectors of A, remains unchanged. Hence, we have

$$U^\top A^0 U - Y^\top A^0 Y = \mathrm{tr}\,\Sigma - \mathrm{tr}\,(R^\top \Sigma R + \sigma E^\top E) \le 0.$$

Subtracting now from we obtain

$$\begin{aligned} \mathcal{D}(Y) - \mathcal{D}^* &\ge \mathrm{tr}\,(R^\top \Sigma R + \sigma E^\top E) - \mathrm{tr}\,(R^\top \Sigma R + E^\top \Sigma_e E) = \mathrm{tr}\, E^\top(\sigma I - \Sigma_e)E \\ &\ge \mathrm{tr}\, E^\top(\sigma I - \sigma_K I)E = (\sigma - \sigma_K)\|E\|_F^2. \end{aligned} \tag{A.1}$$

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 (A.1) we obtain

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

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

Proof of Lemma 1. 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 Eqs. (6)–(7). Denote by V_- the first $K-1$ columns of V. We can write X as

$$X = YV_-^\top + n^{-1/2}\mathbf{1}c^\top = URV_-^\top + U_e EV_-^\top + U_e qc^\top = URV_-^\top + U_e(EV_-^\top + qc^\top).$$

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

$$X' = UR'(V'_-)^\top + U_e\{E'(V'_-)^\top + q(c')^\top\}.$$

We now compute directly $X^\top X'$ and then $(X^\top X')(X^\top X')$, 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 $\phi(X, X') = \mathrm{tr}\,(X^\top X')(X^\top X') = \|X^\top X'\|^2$:

$$\begin{aligned} \mathrm{tr}\,(X^\top X')(X^\top X') &= K - 1 + 2\mathrm{tr}\, V'_- R'^\top R E^\top (q(c')^\top + E'V'_-) + \mathrm{tr}\,(EE^\top + qq^\top)(E'E'^\top + qq^\top) \\ &= K - 1 + 2\mathrm{tr}\, R'^\top R E^\top E' + \mathrm{tr}\,(EE^\top E'E'^\top) + q^\top E^\top E q + q^\top E'^\top E' q + qq^\top \\ &= K - 1 + 2\mathrm{tr}\,(RE^\top)(E'R'^\top) + \mathrm{tr}\,(EE^\top E'E'^\top) + 0 + 0 + 1 \end{aligned} \tag{A.2}$$

To see that $E^\top q = E'^\top q = 0$ recall that $[R^\top\ E^\top]^\top$ and $[0\ q]$ are the coefficients of Y and $\mathbf{1}$, respectively, in the basis U^{all}. As $\mathbf{1} \perp Y$ it must hold that $[0\ q] \perp [R^\top\ E^\top]^\top$, which implies $E^\top q = 0$.

We now try to bound (A.2) from below. We bound from below the last term $\operatorname{tr}(EE^\top E'E'^\top)$ by 0. The middle term $\operatorname{tr}(RE^\top)(E'R'^\top)$ requires more work:

$$|\operatorname{tr}(RE^\top)(E'R'^\top)| = |\langle ER^\top, E'R'^\top\rangle_F| \le \|ER^\top\|_F \|E'R'^\top\|_F.$$

Furthermore,

$$\|ER^\top\|_F^2 = \operatorname{tr} RE^\top ER^\top = \operatorname{tr} E^\top ER^\top R = \operatorname{tr} E^\top E(I - E^\top E) = \operatorname{tr} E^\top E - \operatorname{tr} E^\top EE^\top E \le \|E\|_F^2 - \frac{1}{K-1}\|E\|_F^4. \tag{A.3}$$

The last inequality follows from Lemma 4 stated below.

Now, because the function $x\{1 - x/(K-1)\}$ increases on $[0, (K-1)/2]$, we can combine (A.3) with $\|E\|_F^2 \le \delta$, $\|E'\|_F^2 \le \delta'$ and with (A.2) to obtain that $\|X^\top X'\|^2 \ge K - 2\sqrt{\delta\{1 - \delta/(K-1)\}\delta'\{1 - \delta'/(K-1)\}}$. □

Lemma 4. *For any matrix $A \in R^{m\times m}$, $\|A^\top A\|_F \ge \|A\|^2/m$.*

The proof is left to the reader.

Proof of Lemma 2. Part (i): Let m be the centroid of the data, $m = \sum_i w_i z_i / W_{\text{all}} \in \mathbb{C}^d$. Then, the centered data points z_i^0 can be expressed as $z_i^0 = z_i - m$, or, in matrix notation $Z_0 = Z - \mathbf{1}m^\top = (I - \mathbf{1}\mathbf{w}^\top/W_{\text{all}})Z$. It can easily be verified that $Z_0\mathbf{w} = 0$. Hence,

$$\begin{aligned}
L_0 &= \operatorname{diag}(\sqrt{\mathbf{w}})Z_0Z_0^*\operatorname{diag}(\sqrt{\mathbf{w}}) \qquad \text{(from (18))}\\
&= \operatorname{diag}(\sqrt{\mathbf{w}})(I - \mathbf{1}\mathbf{w}^\top/W_{\text{all}})ZZ^*(I - \mathbf{1}\mathbf{w}^\top/W_{\text{all}})^\top\operatorname{diag}(\sqrt{\mathbf{w}})\\
&= \operatorname{diag}(\sqrt{\mathbf{w}})(I - \mathbf{1}\mathbf{w}^\top/W_{\text{all}})\operatorname{diag}(\mathbf{w})^{-1}\sqrt{S}\sqrt{S}^*\operatorname{diag}(\mathbf{w})^{-1}\\
&\qquad \times (I - \mathbf{1}\mathbf{w}^\top/W_{\text{all}})^\top\operatorname{diag}(\sqrt{\mathbf{w}}) \qquad \text{(from (19))}\\
&= \operatorname{diag}(\sqrt{\mathbf{w}})(I - \mathbf{1}\mathbf{w}^\top/W_{\text{all}})\operatorname{diag}(\sqrt{\mathbf{w}})^{-1}\underbrace{\operatorname{diag}(\sqrt{\mathbf{w}})^{-1}S\operatorname{diag}(\sqrt{\mathbf{w}})^{-1}}_{L}\\
&\qquad \times \operatorname{diag}(\sqrt{\mathbf{w}})^{-1}(I - \mathbf{1}\mathbf{w}^\top/W_{\text{all}})^\top\operatorname{diag}(\sqrt{\mathbf{w}})\\
&= (I - \sqrt{\mathbf{w}}\sqrt{\mathbf{w}}^\top/W_{\text{all}})L(I - \sqrt{\mathbf{w}}\sqrt{\mathbf{w}}^\top/W_{\text{all}})^\top\\
&= (I - B)L(I - B)
\end{aligned} \tag{A.4}$$

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

$$B\sqrt{\mathbf{w}} = \sqrt{\mathbf{w}}, \tag{A.5}$$

$$Bu = 0 \qquad \text{for all } u \perp \sqrt{\mathbf{w}}, \tag{A.6}$$

L is a symmetric real matrix, hence it has real eigenvalues $\lambda_1 \ge \cdots \ge \lambda_n$ and real, orthogonal eigenvectors $u_1, \ldots, u_n$. The largest eigenvalue of L has value $\lambda_1 = 1$ and its corresponding eigenvector is $u_1 = \sqrt{\mathbf{w}}/\sqrt{W_{\text{all}}}$; see [23].

Applying (A.4)–(A.6) we obtain after some simple calculations

$$L_0u_j = \begin{cases} 0 & \text{if } j = 1,\\ \lambda_j u_j & \text{if } j > 1. \end{cases}$$

Part (iii): Eq. (20) follows from (ii).

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

$$\mathcal{D}(X) = \operatorname{tr} L_0 - \operatorname{tr} X^\top L_0 X = \operatorname{tr} L_0 - \operatorname{tr} V[Y\ u_1]^\top L_0[Y\ u_1]V^\top = \operatorname{tr} L_0 - u_1^\top L_0 u_1 - \operatorname{tr} Y^\top L_0 Y = \operatorname{tr} L_0 - \operatorname{tr} Y^\top L_0 Y$$

To obtain Eq. (21), it is sufficient to equate the right-hand sides of (5) and (22). □

Proof of Lemma 3. We have $\epsilon(\delta,\delta) = 2\delta\{1 - \delta/(K-1)\}$ and

$$\begin{aligned}
\epsilon^{\text{old}}(\delta,\delta) &= 2\delta\sqrt{(1-\delta)(K-\delta)} + (K+1)\delta - 2\delta^2 = 2\delta(K-\delta)\left\{\sqrt{\frac{1-\delta}{K-\delta}} + \frac{(K+1)/2 - \delta}{K-\delta}\right\}\\
&\ge 2\delta(1 - \frac{\delta}{K-1})K\left\{\sqrt{\frac{1-\delta}{K-\delta}} + \frac{(K+1)/2 - \delta}{K-\delta}\right\} = \epsilon(\delta,\delta)KF(\delta).
\end{aligned}$$

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

$$F(x) = \sqrt{\frac{1-x}{K-x}} + \frac{(K+1)/2 - x}{K-x}$$

and hence, for $x < 1$,

$$F'(x) = \frac{1}{2\sqrt{\frac{1-x}{K-x}}} \frac{x-K-(x-1)}{(x-K)^2} + \frac{x-K-\{x-(K+1)/2\}}{(x-K)^2} = -\frac{K-1}{2(x-K)^2}\left(\sqrt{\frac{K-x}{1-x}} - 1\right) \leq 0$$

Hence, for all $x \in [0, 1]$,

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

Proof of Theorem 5. Part (i): It is easy to check that $B\sqrt{\mathbf{w}} = \sqrt{\mathbf{w}}$ and, therefore, $A\sqrt{\mathbf{w}} = (I-B)A_0(I-B)\sqrt{\mathbf{w}} = 0$. It suffices to prove that $X^\top A = X^\top A_0 X$ + constant. To simplify notation, assume without loss of generality that $W_{\text{all}} = 1$. Then $B = \sqrt{\mathbf{w}}\sqrt{\mathbf{w}}^\top$ and the Kth column of X is $X_{:k} = \text{diag}(\sqrt{\mathbf{w}})\tilde{X}_{:k}\text{diag}(W_1^{-1/2}, \ldots, W_K^{-1/2})$. Hence $BX_{:k} = \sqrt{\mathbf{w}}\sqrt{W_k}$ and

$$\begin{aligned} BX &= (\sqrt{\mathbf{w}} \cdots \sqrt{\mathbf{w}})\text{diag}(W_k^{1/2}) = \text{diag}(\sqrt{\mathbf{w}})(\mathbf{1}_n \cdots \mathbf{1}_n)\text{diag}(W_k^{1/2}) = \text{diag}(\sqrt{\mathbf{w}})\tilde{X}(\mathbf{1}_K \cdots \mathbf{1}_K)\text{diag}(W_k^{1/2}) \\ &= \text{diag}(\sqrt{\mathbf{w}})\tilde{X}\text{diag}(W_k^{-1/2})\text{diag}(W_k^{1/2})(\mathbf{1}_K \cdots \mathbf{1}_K)\text{diag}(W_k^{1/2}) = X\underbrace{[\sqrt{W_k W_{k'}}]_{k,k'=1:K}}_{B'} = XB'. \end{aligned}$$

Then,

$$\begin{aligned} \text{tr}\, X^\top A X &= \text{tr}\,(I-B')X^\top A_0 X(I-B') = \text{tr}\,(I-B')^2 X^\top A_0 X' = \text{tr}\,(I-B')X^\top A_0 X \;\; \text{because } (B')^2 = B' \\ &= \text{tr}\, X^\top A_0 X + \text{tr}\, B' X^\top A_0 X \end{aligned} \tag{A.7}$$

and

$$\text{tr}\, B' X^\top A_0 X = \text{tr}\sqrt{[W_k]_{k=1:K}}\sqrt{[W_k]_{k=1:K}}^\top X^\top A_0 X = \sqrt{[W_k]_{k=1:K}}^\top X^\top A_0 \underbrace{X\sqrt{[W_k]_{k=1:K}}}_{\sqrt{\mathbf{w}}} = \sqrt{\mathbf{w}}^\top A_0 \sqrt{\mathbf{w}}.$$

In the above, $[W_k]_{k=1:K} = (W_1, \ldots, W_K)^\top$ represents the column vector of cluster weights. Replacing the last equation into (A.7) we obtain the desired result.

Part (ii): Since $A_0\sqrt{\mathbf{w}} = 0$, this part is proved in the same way as (9) in Section 2.2.

Part (iii): The proof follows closely the proof of Theorem 3 and is therefore omitted. □

References

[1] D. Achlioptas, F. McSherry, On spectral learning of mixtures of distributions, in: P. Auer, R. Meir (Eds.), 18th Annual Conference on Learning Theory, COLT 2005, Springer, Berlin/Heidelberg, 2005, pp. 458–471.

[2] F. Bach, M.I. Jordan, Learning spectral clustering, in: S. Thrun, L. Saul (Eds.), Advances in Neural Information Processing Systems, vol. 16, MIT Press, Cambridge, MA, 2004.

[3] S. Balakrishnan, M.J. Wainwright, B. Yu, Statistical Guarantees for the EM Algorithm: From Population to Sample-Based Analysis, Technical Report, 2018, arXiv:1408.2156.

[4] S. Balakrishnan, M. Xu, A. Krishnamurthy, A. Singh, Noise thresholds for spectral clustering, in: J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F.C.N. Pereira, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12–14 December 2011, Granada, Spain, 2011, pp. 954–962.

[5] M. Balcan, A. Blum, A. Gupta, Clustering under approximation stability, J. ACM 60 (2013) 8.

[6] M.-F. Balcan, C. Borgs, M. Braverman, J. Chayes, S.-H. Teng, Finding endogenously formed communities, 2012, arxiv preprint arXiv:1201.4899v2.

[7] S. Ben-David, U. von Luxburg, D. Pal, A sober look at clustering stability, in: 19th Annual Conference on Learning Theory, COLT 2006, Springer, 2006.

[8] Y. Bilu, N. Linial, Are stable instances easy?, Combin. Probab. Comput. 21 (2012) 643–660.

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

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

[11] I.S. Dhillon, Y. Guan, B. Kulis, Kernel K-means, spectral clustering and normalized cuts, in: R. Kohavi, J. Gehrke, J. Ghosh (Eds.), Proceedings of The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD), ACM Press, 2004, pp. 551–556.

[12] C. Ding, X. He, K-means clustering via principal component analysis, in: C.E. Brodley (Ed.), Proceedings of the International Machine Learning Conference (ICML), Morgan Kauffman, 2004.

[13] K.L. Fleming, P. Tiwary, J. Pfaendtner, New approach for investigating reaction dynamics and rates with ab initio calculations, J. Phys. Chem. A 120 (2016) 299–305.

[14] L. Hubert, P. Arabie, Comparing partitions, J. Classification 2 (1985) 193–218.

[15] R. Kannan, H. Salmasian, S. Vempala, The spectral method for general mixture models, in: P. Auer, R. Meir (Eds.), 18th Annual Conference on Learning Theory, COLT 2005, Springer, Berlin/Heidelberg, 2005, pp. 444–457.

[16] H. Lancaster, The Chi-Squared Distribution, Wiley, 1969.

[17] J.R. Lee, S.O. Gharan, L. Trevisan, Multi-way spectral partitioning and higher-order cheeger inequalities, 2014, arXiv:1111.1055.

[18] M. Meila, L. Xu, Multiway Cuts and Spectral Clustering, Technical Report 442, University of Washington, 2003.

[19] M. Meilă, The uniqueness of a good optimum for K-means, in: A. Moore, W. Cohen (Eds.), Proceedings of the International Machine Learning Conference (ICML), International Machine Learning Society, pp. 625–632.

[20] M. Meilă, The Multicut Lemma, Technical Report 417, University of Washington, 2002.

[21] M. Meilă, Comparing clusterings – an axiomatic view, in: S. Wrobel, L. De Raedt (Eds.), Proceedings of the International Machine Learning Conference (ICML), ACM Press, 2005.
[22] M. Meilă, Local equivalence of distances between clusterings – a geometric perspective, Mach. Learn. 86 (2012) 369–389.
[23] M. Meilă, J. Shi, A random walks view of spectral segmentation, in: T. Jaakkola and T. Richardson (Eds.), Artificial Intelligence and Statistics AISTATS.
[24] M. Meilă, S. Shortreed, L. Xu, Regularized spectral learning, in: R. Cowell and Z. Ghahramani (Eds), Proceedings of the Artificial Intelligence and Statistics Workshop(AISTATS 05).
[25] J. Moore, M. Ackerman, Foundations of perturbation robust clustering, in: 16th International Conference on Data Mining (ICDM), IEEE, pp. 1089–1094.
[26] A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, in: T.G. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14, MIT Press, Cambridge, MA, 2002.
[27] R. Ostrovsky, Y. Rabani, L.J. Schulman, C. Swamy, The effectiveness of lloyd-type methods for the k-means problem, J. ACM 59 (2012) 28.
[28] C. Papadimitriou, K. Steiglitz, Combinatorial Optimization. Algorithms and Complexity, Dover Publication, Inc., Minneola, NY, 1998.
[29] R. Peng, H. Sun, L. Zanetti, Partitioning well-clustered graphs: Spectral clustering works!, in: P. Grünwald, E. Hazan (Eds.), Proceedings of The 28th Conference on Learning Theory (COLT), vol. 40, pp. 1–33.
[30] A. Rakhlin, A. Caponnetto, Stability of k-means clustering, in: Advances in Neural Information Processing Systems (NIPS), MIT Press, Cambridge,MA, 2006, pp. 1121–1128.
[31] B. Schölkopf, A. Smola, K.-R. Müller, Non-linear analysis as a kernel eigenvalue problem, Neural Comput. 10 (1998) 1299–1319, Kernel k-means.
[32] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. (2000).
[33] N. Srebro, G. Shakhnarovich, S. Roweis, An investigation of computational and informational limits in gaussian mixture clustering, in: Proceedings of the 23rd International Conference on Machine Learning (ICML).
[34] V. Vapnik, Statistical Learning Theory, Wiley, 1998.
[35] S. Vempala, G. Wang, A spectral algorithm for learning mixture models, J. Comput. System Sci. 68 (2004) 841–860.
[36] U. von Luxburg, O. Bousquet, M. Belkin, Limits of spectral clustering, in: L.K. Saul, Y. Weiss, L. Bottou (Eds.), Advances in Neural Information Processing Systems 17, MIT Press, 2005.
[37] Y. Wan, M. Meila, A class of network models recoverable by spectral clustering, in: D. Lee, M. Sugiyama (Eds.), Advances in Neural Information Processing Systems (NIPS).
[38] Y. Wan, M. Meila, Benchmarking recovery theorems for the DC-SBM, in: International Symposium on Artificial Intelligence and Mathematics (ISAIM).
[39] Y. Wan, M. Meilă, Graph clustering: block-models and model-free results, in: Advances in Neural Information Processing Systems (NIPS).
[40] S.X. Yu, J. Shi, Multiclass spectral clustering, in: International Conference on Computer Vision.