Clustering via Generalized Energy Statistics

Anonymous Authors[1]

Abstract

Energy statistics introduces the notion of potential energy between probability distributions, in close analogy to Newton's gravitational potential in physics. In this paper, we propose a principled approach to clustering based on energy statistics theory. Our mathematical formulation establishes connection to kernel methods, leading to a quadratically constrained quadratic program in the associated feature space. To obtain local solutions of such an NP-hard optimization problem, we introduce an iterative algorithm based on Hartigan's method. This algorithm has the same computational cost as kernel k-means but offers several advantages. We provide carefully designed numerical experiments illustrating that the proposed method is more flexible and outperforms kernel k-means, spectral clustering, standard k-means and Gaussian mixture models in a variety of settings, specially in high dimensions. We employ the method to an important real dataset describing protein expressions of neural synapses.

1. Introduction

Energy statistics (Székely & Rizzo, 2013) provides a hypothesis test for equality of distributions, which is achieved under the minimum of a "statistical energy". When probability distributions are different, this statistical energy diverges as sample size increases. On the other hand, it tends to a nondegenerate limit distribution when probability distributions are equal. The test statistic has a compact representation in terms of expectations of pairwise distances, providing straightforward empirical estimates. Energy statistics has been applied to several goodness-of-fit hypothesis tests, multi-sample tests of equality of distributions, analysis of variance (Rizzo & Székely, 2010), nonlinear dependence tests through distance covariance and distance correlation, which generalizes the Pearson correlation coefficient, and hierarchical clustering (Székely & Rizzo, 2005) by extending Ward's method of minimum variance. Moreover, an application of energy statistics to clustering in Euclidean spaces was considered (Li, 2015). We refer to (Székely & Rizzo, 2013) for an overview.

Lyons (2013) generalized distance covariance from Euclidean to arbitrary metric spaces of strong negative type. Furthermore, the missing link between energy distance based tests and kernel based tests has been recently resolved by Sejdinovic et al. (2013), where a unifying framework establishing an equivalence between generalized energy distances to maximum mean discrepancies (MMD), which are distances between embeddings of distributions in reproducing kernel Hilbert spaces (RKHS), was established. This equivalence immediately relates energy statistics to kernel methods often used in machine learning, and form the basis of our approach in this paper.

Clustering has such a long history in machine learning, making it impossible to mention all important contributions in a short space. Perhaps, the most used method is k-means (Lloyd, 1982; J. B. MacQueen, 1967; Forgy, 1965), which is based on Lloyd's heuristic (Lloyd, 1982) of assigning a data point to the cluster with closest center. The only statistical information about each cluster comes from its mean, making it sensitive to outliers. Nevertheless, k-means works very well when data is linearly separable in Euclidean space. Gaussian mixture models (GMM) is another very common approach, providing more flexibility than k-means, however, it still makes strong assumptions about the distribution of the data.

To account for nonlinearities, kernel methods were introduced (Schölkopf et al., 1998; Girolami, 2002). A mercer kernel (Mercer, 1909) is used to implicitly map data points to a RKHS, then clustering can be performed in the associated Hilbert space by exploiting its inner product. However, the kernel choice remains the biggest challenge since there is no principled theory to construct a kernel for a given dataset, and usually kernels introduce hyperparameters that need to be carefully chosen. The well-known kernel k-means optimization problem is nothing but k-means in the feature space (Girolami, 2002). Furthermore, kernel k-means algorithm (Dhillon et al., 2004; 2007) is still based on Loyd's

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

heuristic of grouping points that are closer to a cluster center, now in the feature space. We refer the reader to Filippone et al. (2008) for a survey of clustering methods.

Although clustering from energy statistics in Euclidean spaces was considered in Li (2015), the precise optimization problem behind this approach remains elusive, as well as the connection with kernel methods. The main theoretical contribution of this paper is to fill this gap. Since the statistical potential energy is minimum when distributions are equal, the principle behind clustering is to maximize the statistical energy, enforcing probability distributions associated to each cluster to be different from one another. We provide a precise mathematical formulation to this statement, leading to a quadratically constrained quadratic program (QCQP) in the associated RKHS. Our results immediately establish the connection between energy statistics based clustering and kernel methods. Moreover, we show that that this QCQP is equivalent to kernel k-means optimization problem.

Our main algorithmic contribution is to use Hartigan's method (Hartigan & Wong, 1979) to find local solutions of the above mentioned QCQP, which is NP-hard in general. Hartigan's method was also used by Li (2015), but without any connection to kernels. More importantly, the advantages of Hartigan's over Lloyd's method was already demonstrated in some simple settings by Telgarsky & Vattani (2010); Slonim et al. (2013), but apparently this method did not receive the deserved attention by the statistics and machine learning communities. To the best of our knowledge, Hartigan's method was not previously employed together with kernel methods. Here we provide a fully kernel based Hartigan's algorithm for clustering, where the kernel is fixed by energy statistics. We make clear the advantages of this proposal versus Lloyd's method, which kernel k-means is based upon and will also be used to solve our QCQP. We show that both algorithms have the same time complexity, however, Hartigan's method in kernel spaces is superior. Furthermore, in the examples considered in this paper, it also provides equal or better performance than spectral clustering, which in fact solves a relaxed version of our QCQP. Our numerical results provide compelling evidence that Hartigan's method applied to energy statistics based clustering is more accurate and robust than kernel k-means. Moreover, we illustrate the flexibility of energy clustering, showing that it is able to perform accurately on data coming from very different distributions, contrary to k-means and GMM, for instance. More specifically, the proposed method performs closely to k-means and GMM on normally distributed data, however, it is significantly better on other settings. Its superiority in high dimensions is striking, being more accurate than k-means and GMM even on Gaussian settings. We illustrate the applicability of the proposed method on a real dataset obtained by neuroscientists experts, which describes shapes of neural synapses.

2. Background

In this section we introduce the main concepts from (generalized) energy statistics (Székely & Rizzo, 2013; Lyons, 2013) and its relation to RKHS (Sejdinovic et al., 2013).

Consider random vectors X_i $(i = 1, \dots, n)$ living in an arbitrary space $\mathcal{X}$ of *negative type*, which means that $\mathcal{X}$ is endowed with a *semimetric* $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfying $\sum_{i,j=1}^{n} c_i c_j \rho(X_i, X_j) \le 0$, where $c_i \in \mathbb{R}$ and $\sum_{i=1}^{n} c_i = 0$. Let $X, X' \overset{iid}{\sim} P$ and $Y, Y' \overset{iid}{\sim} Q$, where P and Q are cumulative distribution functions with finite first moments, and $X, X', Y, Y' \in \mathcal{X}$. The *generalized energy distance* between P and Q is given by

$$\mathcal{E}(P, Q) \equiv 2\mathbb{E}\rho(X, Y) - \mathbb{E}\rho(X, X') - \mathbb{E}\rho(Y, Y'). \quad (1)$$

This quantity is nonnegative, $\mathcal{E}(P, Q) \ge 0$, and $\mathcal{E}^{1/2}$ is a metric on the space of distributions. Energy distance provides a characterization of equality of distributions. For instance, the *standard energy distance* in Euclidean spaces introduced by Székely & Rizzo (2013) uses the semimetric

$$\rho_\alpha(X, Y) = \|X - Y\|^\alpha \quad (2)$$

where $0 < \alpha \le 2$ and $\| \cdot \|$ is the Euclidean norm in $\mathcal{X} = \mathbb{R}^D$. In this case, $\mathcal{E}(P, Q)$ is rotationally invariant. For $0 < \alpha < 2$ we have $\mathcal{E}(P, Q) = 0$ if and only if $P = Q$. However, for $\alpha = 2$ we get $\mathcal{E}(P, Q) = 2\|\mathbb{E}(X) - \mathbb{E}(Y)\|^2$, thus $\mathcal{E}(P, Q) = 0$ does not imply equality of distributions but only equality of the means.

Consider a sample $\mathbb{X} = \{x_1, \dots, x_n\}$ from k unknown distributions $\{P_j\}_{j=1}^{k}$, where $x_i \in \mathcal{X}$. Let $\mathbb{X} = \bigcup_{j=1}^{k} \mathcal{C}_j$ be a disjoint partition, i.e. $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$. Each expectation in the generalized energy distance can be empirically estimated with the aid of the function

$$g(\mathcal{C}_i, \mathcal{C}_j) \equiv \frac{1}{n_i n_j} \sum_{x \in \mathcal{C}_i} \sum_{y \in \mathcal{C}_j} \rho(x, y), \quad (3)$$

where $n_i = |\mathcal{C}_i|$ denotes the number of points in partition $\mathcal{C}_i$. Define the *within energy dispersion* as

$$W \equiv \sum_{j=1}^{k} \frac{n_j}{2} g(\mathcal{C}_j, \mathcal{C}_j), \quad (4)$$

and the *between-sample energy statistic* as

$$S \equiv \sum_{1 \le i < j \le k} \frac{n_i n_j}{2n} \left[2g(\mathcal{C}_i, \mathcal{C}_j) - g(\mathcal{C}_i, \mathcal{C}_i) - g(\mathcal{C}_j, \mathcal{C}_j) \right], \quad (5)$$

where $n = \sum_{j=1}^{k} n_j$. A given point x_i belongs to partition $\mathcal{C}_j$ if and only if $x_i \sim P_j$. The quantity S defined above is a test statistic for equality of distributions (Székely & Rizzo, 2013). When the sample size is large enough, $n \to \infty$, under the null hypothesis $H_0 : P_1 = P_2 = \cdots = P_k$ we have that $S \to 0$, and under the alternative hypothesis $H_1 : P_\ell \neq P_j$ for at least two $\ell \neq j$, we have that $S \to \infty$.

Let $\mathcal{H}_K$ be a Hilbert space of real-valued functions over $\mathcal{X}$ with an associated kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which is a symmetric and positive definite function. For every $x \in \mathcal{X}$ there exists $h_x \equiv K(\cdot, x) \in \mathcal{H}_K$ such that $\langle h_x, f \rangle = f(x)$ for any function $f \in \mathcal{H}_K$. Thus, $\langle h_x, h_y \rangle = K(x, y)$. Conversely, for every symmetric positive definite function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ there is a Hilbert space $\mathcal{H}_K$ with reproducing kernel K, with a *feature map* $\varphi : x \mapsto h_x \in \mathcal{H}_K$ such that $\langle \varphi(x), \varphi(y) \rangle = K(x, y)$ (Aronszajn, 1950).

Define the embedding of a probability measure $P \mapsto h_P \in \mathcal{H}_K$ through $h_P \equiv \int K(\,\cdot\,, x) dP(x)$. The distance between two probability measures, called maximum mean discrepancy (MMD), is thus given by

$$\gamma_K(P, Q) \equiv \| h_P - h_Q \|_{\mathcal{H}_K}, \tag{6}$$

which can also be written as (Gretton et al., 2012)

$$\gamma_K^2(P, Q) = \mathbb{E}K(X, X') + \mathbb{E}K(Y, Y') - 2\mathbb{E}K(X, Y) \tag{7}$$

where $X, X' \overset{iid}{\sim} P$ and $Y, Y' \overset{iid}{\sim} Q$. The equality between (6) and (7) gives $\langle h_P, h_Q \rangle_{\mathcal{H}_K} = \mathbb{E}\, K(X, Y)$.

The following important result of Berg et al. (1984) shows that semimetrics of negative type and symmetric positive definite kernels are closely related. Let $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and $x_0 \in \mathcal{X}$ be an arbitrary but fixed point. Define

$$K(x, y) \equiv \tfrac{1}{2} \left[\rho(x, x_0) + \rho(y, x_0) - \rho(x, y) \right]. \tag{8}$$

Then, it can be shown that K is positive definite if and only if ρ is a semimetric of negative type. We thus have a family of kernels, one for each choice of x_0. Conversely, as shown by Sejdinovic et al. (2013), if ρ is a semimetric of negative type and K is a kernel in this family, then

$$\begin{aligned} \rho(x, y) &= K(x, x) + K(y, y) - 2K(x, y) \\ &= \| h_x - h_y \|^2_{\mathcal{H}_K} \end{aligned} \tag{9}$$

and the canonical feature map $\varphi : x \mapsto h_x$ is injective. When the above conditions are met we say that the kernel K generates the semimetric ρ. If two different kernels generate the same ρ they are said to be equivalent kernels. Now we can state the equivalence between energy distance and inner products on RKHS, which is one of the main results of Sejdinovic et al. (2013). If ρ is a semimetric of negative type and K a kernel that generates ρ, then replacing (9) into (1), and using (7), yields

$$\begin{aligned} \tfrac{1}{2}\mathcal{E}(P, Q) &= \mathbb{E}\, K(X, X') + \mathbb{E}\, K(Y, Y') - 2\mathbb{E}\, K(X, Y) \\ &= \gamma_K^2(P, Q). \end{aligned} \tag{10}$$

Therefore, we can compute the energy distance using the inner product of the associated Hilbert space $\mathcal{H}_K$.

3. Clustering via Energy Statistics

This section contains our main theoretical contributions, where we formulate an optimization problem for clustering based on energy statistics in the associated RKHS. The proofs are contained in supplementary material.

Due to the energy test statistic for equality of distributions previously discussed, the obvious criterion for clustering data is to maximize S, defined in (5), which makes each cluster as different as possible from the other ones. In other words, given a set of points coming from different probability distributions, the test statistic S should attain a maximum when each point is correctly classified as belonging to the cluster associated to its probability distribution. The following straightforward result shows that maximizing S is, however, equivalent to minimizing W, which has a more convenient form.

Lemma 1. *Let $\mathbb{X} = \{x_1, \dots, x_n\}$, where each data point x_i lives in a space $\mathcal{X}$ of negative type. For a fixed integer k, the partition $\mathbb{X} = \bigcup_{j=1}^k \mathcal{C}_j^\star$, where $\mathcal{C}_i^\star \cap \mathcal{C}_j^\star = \emptyset$ for all $i \neq j$, maximizes the between-sample statistic S, defined in* (5)*, if and only if*

$$\{\mathcal{C}_1^\star, \dots, \mathcal{C}_k^\star\} = \operatorname*{arg\,min}_{\{\mathcal{C}_1, \dots, \mathcal{C}_k\}} W(\mathcal{C}_1, \dots, \mathcal{C}_k), \tag{11}$$

where the within energy dispersion W is defined by (4)*.*

In the particular Euclidean case, the optimization problem (11) based on energy statistics was already proposed in Li (2015). However, it is important to note that this is equivalent to maximizing S, which is the test statistic for equality of distributions.

The clustering problem as stated in the form (11) makes the relation with kernels and other clustering methods obscure. In the following, we demonstrate what is the explicit optimization problem behind (11) in the associated RKHS, directly establishing the connection with kernel methods commonly used in machine learning.

Assume that the kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ generates ρ. Define the Gram matrix $G \in \mathbb{R}^{n \times n}$ with components

$$G_{ij} \equiv K(x_i, x_j). \tag{12}$$

Let $Z \in \{0, 1\}^{n \times k}$ be the label matrix, with only one nonvanishing entry per row, indicating to which cluster (column) each point (row) belongs to. This matrix satisfies $Z^\top Z = D$, where $D = \operatorname{diag}(n_1, \dots, n_k)$ contains the number of points in each cluster. We also introduce the rescaled matrix $Y \equiv Z D^{-1/2}$. In component form they are given by

$$Z_{ij} \equiv \begin{cases} 1 & \text{if } x_i \in \mathcal{C}_j \\ 0 & \text{otherwise} \end{cases}, \qquad Y_{ij} \equiv \begin{cases} \frac{1}{\sqrt{n_j}} & \text{if } x_i \in \mathcal{C}_j \\ 0 & \text{otherwise} \end{cases}. \tag{13}$$

Our next result shows that the optimization problem (11) is NP-hard since it is a quadratically constrained quadratic program (QCQP) in the associated RKHS.

Proposition 2. *Problem (11) is equivalent to*

$$\begin{aligned} &\max_{Y} \operatorname{Tr}\left(Y^\top G Y\right) \\ &\text{s.t. } Y \geq 0,\, Y^\top Y = I,\, YY^\top e = e, \end{aligned} \tag{14}$$

where $e = (1, \dots, 1)^\top \in \mathbb{R}^n$ is the all-ones vector, and G is the Gram matrix (12).

Thus, to group data $\mathbb{X} = \{x_1, \dots, x_n\}$ into k clusters we first compute the Gram matrix G and then solve the optimization problem (14) for $Y \in \mathbb{R}^{n \times k}$. The ith row of Y will contain a single nonzero element in some jth column, indicating that $x_i \in \mathcal{C}_j$.

Besides NP-hard, the optimization problem (14) is nonconvex, and a direct computational approach is prohibitive even for small datasets. However, one can find approximate solutions by relaxing some of the constraints. For instance, the relaxed problem $\max_Y \operatorname{Tr}\left(Y^\top G Y\right)$ s.t. $Y^\top Y = I$, has a well-known closed form solution $Y^\star = UR$, where the columns of $U \in \mathbb{R}^{n \times k}$ contain the top k eigenvectors of G corresponding to the k largest eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$, and $R \in \mathbb{R}^{k \times k}$ is an arbitrary orthogonal matrix. *Spectral clustering* is based on (variants of) this approach and will be compared to the iterative method that will be proposed in the next section.

Note also that the optimization problem (14) is valid for data living in an *arbitrary* space of negative type, where a semimetric ρ, and thus the kernel K, are assumed to be known. Standard energy statistics in Euclidean spaces fixes a family of choices through (2). The same is valid for data living in more general spaces $(\mathcal{X}, \rho)$. In any case, energy clustering is model-free, contrary to k-means and GMM, for example. In practice, however, the clustering quality strongly depends on the choice of a suitable ρ which measures the similarity between data points. If prior information is available to make this choice, then it can be immediately incorporated into (14) through the kernel matrix G.

3.1. Relation to Kernel k-Means Optimization Problem

One may wonder how energy clustering relates to the well-known kernel k-means optimization problem[1], extensively used in machine learning. For a positive semidefinite Gram matrix G, as defined in (12), there exists a feature map $\varphi : \mathcal{X} \to \mathcal{H}_K$ such that $G_{ij} = \langle \varphi(x_i), \varphi(x_j) \rangle$. The kernel k-means optimization problem is defined by

$$\min_{\mathcal{C}_1, \dots, \mathcal{C}_k} \left\{ \sum_{j=1}^{k} \sum_{x \in \mathcal{C}_j} \| \varphi(x) - \varphi(\mu_j) \|^2 \right\}, \tag{15}$$

where $\mu_j = \frac{1}{n_j} \sum_{x \in \mathcal{C}_j} x$ is the mean of cluster $\mathcal{C}_j$ in the ambient space $\mathcal{X}$. Notice that the above objective function is strongly tied to the idea of minimizing distances between points and cluster centers, which arises from k-means objective function based on Lloyd's heuristic (Lloyd, 1982). It was shown by Dhillon et al. (2004; 2007) that kernel k-means problem can be cast into a trace maximization in the same form as (14). The next result makes this explicit.

Proposition 3. *For a fixed kernel, the energy clustering optimization problem (11) is equivalent to kernel k-means optimization problem (15), and both are equivalent to (14).*

The above result shows that kernel k-means problem is equivalent to the clustering problem formulated in the energy statistics framework, when operating on the same kernel. Notice, however, that energy statistics theory is valid for arbitrary semimetric spaces of negative type, fixing the kernel function in the associated RKHS, which is guaranteed to be positive definite. As shown by Dhillon et al. (2004; 2007), kernel k-means, spectral clustering, and graph partitioning problems such as ratio association, ratio cut, and normalized cut are all equivalent to a QCQP of the form (14). Therefore, one can view all of these problems as arising from energy statistics as well.

4. Hartigan's Method for Energy Clustering

We now introduce an iterative algorithm based on Hartigan & Wong (1979) to find a local maximizer of the optimization problem (14). Due to Proposition 3 we can also use kernel k-means algorithm (Dhillon et al., 2004; 2007) to solve this optimization problem, and it will be compared with the proposed algorithm in our numerical experiments.

We can rewrite the optimization problem (14) as

$$\max_{\{\mathcal{C}_1, \dots, \mathcal{C}_k\}} \left\{ Q = \sum_{j=1}^{k} \frac{Q_j}{n_j} \right\}, \quad Q_j \equiv \sum_{x, y \in \mathcal{C}_j} K(x, y), \tag{16}$$

where Q_j represents an internal cost of cluster $\mathcal{C}_j$, and Q is the total cost where each Q_j is weighted by the inverse of the number of points in $\mathcal{C}_j$. For a data point x_i we denote its own cost with the entire cluster $\mathcal{C}_\ell$ by

$$Q_\ell(x_i) \equiv \sum_{y \in \mathcal{C}_\ell} K(x_i, y) = G_{i\bullet} \cdot Z_{\bullet \ell}. \tag{17}$$

Above, $G_{i\bullet}$ ($G_{\bullet i}$) denotes the ith row (column) of the matrix G. For a given configuration, we consider the maximum

[1] By this we mean specifically the optimization problem (15), which should not be confused with kernel k-means *algorithm* that is just one possible recipe to solve (15). The distinction between kernel k-means problem and kernel k-means algorithm should be clear from the context.

change in the total cost function Q when moving each data point to another cluster. More specifically, suppose x_i is currently assigned to cluster $\mathcal{C}_j$, yielding a total cost function denoted by $Q^{(j)}$. Moving x_i to cluster $\mathcal{C}_\ell$ yields another total cost function denoted by $Q^{(\ell)}$. We are interested in computing the maximum $\Delta Q^{j \to \ell}(x_i) \equiv Q^{(\ell)} - Q^{(j)}$, for $\ell \neq j$. From (16), by explicitly writing the costs related to these two cluster we obtain

$$\Delta Q^{j \to \ell}(x_i) = \frac{Q_\ell^+}{n_\ell + 1} + \frac{Q_j^-}{n_j - 1} - \frac{Q_j}{n_j} - \frac{Q_\ell}{n_\ell}, \qquad (18)$$

where Q_ℓ^+ denote the cost of the new ℓth cluster with the point x_i added to it, and Q_j^- is the cost of new jth cluster with x_i removed from it. Observe that $Q_\ell^+ = Q_\ell + 2Q_\ell(x_i) + G_{ii}$ and $Q_j^- = Q_j - 2Q_j(x_i) + G_{ii}$, hence

$$\begin{aligned} \Delta Q^{j \to \ell}(x_i) i = {} & \frac{1}{n_j - 1}\left[\frac{Q_j}{n_j} - 2Q_j(x_i) + G_{ii}\right] \\ & - \frac{1}{n_\ell + 1}\left[\frac{Q_\ell}{n_\ell} - 2Q_\ell(x_i) - G_{ii}\right]. \qquad (19) \end{aligned}$$

Therefore, if $\Delta Q^{j \to \ell}(x_i) > 0$ we get closer to a maximum of (16) by moving x_i to $\mathcal{C}_\ell$, otherwise we should keep x_i in $\mathcal{C}_j$. A similar analysis leading to (19) in the particular case of Euclidean distances was considered in Li (2015).

We thus propose the following algorithm. Start with an initial label matrix $Z = Z_0$, then for each point x_i, assuming it currently belongs to cluster $\mathcal{C}_j$, compute the cost of moving it to $\mathcal{C}_\ell$, i.e. $\Delta Q^{j \to \ell}(x_i)$ for $\ell = 1, \dots, k$ with $\ell \neq j$. Choose

$$j^\star = \operatorname*{arg\,max}_{\ell = 1, \dots, k \,|\, \ell \neq j} \Delta^{j \to \ell}(x_i). \qquad (20)$$

If $\Delta Q^{j \to j^\star}(x_i) > 0$ move x_i to $\mathcal{C}_{j^\star}$, otherwise keep x_i in its original cluster $\mathcal{C}_j$. Repeat this process until no new assignments are made. This procedure is explicitly described in Algorithm 1, which we name $\mathcal{E}^H$-clustering to emphasize that it is based on Hartigan's method. Note that the objective function is monotonically increasing at each iteration, consequently, the algorithm converges in a finite number of steps.

The worst time complexity of the above algorithm is $\mathcal{O}(nk \max_\ell n_\ell) = \mathcal{O}(kn^2)$, which is the same cost as kernel k-means algorithm (Dhillon et al., 2004; 2007). If G is sparse this can be further reduced to $\mathcal{O}(kn')$ where n' is the number of nonzero entries.

There are known results about Hartigan's method indicating its advantages over Lloyd's method.

Theorem 4 (Telgarsky & Vattani (2010))**.** *If $n > k$ the resulting partition obtained from Hartigan's method has no empty clusters, and distinct means.*

Algorithm 1 $\mathcal{E}^H$-clustering is Hartigan's method to find local solutions to the optimization problem (14).

```
input  number of clusters $k$, Gram matrix $G$, initial label matrix $Z \leftarrow Z_0$
output  label matrix $Z$
$q \leftarrow (Q_1, \dots, Q_k)^\top$ have the energy costs of each cluster, defined in (16)
$n \leftarrow (n_1, \dots, n_k)^\top$ have the number of points in each cluster
repeat
    for $i = 1, \dots, n$ do
        let $j$ be such that $x_i \in \mathcal{C}_j$
        $j^\star \leftarrow \arg\max_{\ell=1,\dots,k \,|\, \ell \neq j} \Delta Q^{j \to \ell}(x_i)$
        if $\Delta Q^{j \to j^\star}(x_i) > 0$ then
            move $x_i$ to $\mathcal{C}_{j^\star}$: $Z_{ij} \leftarrow 0$ and $Z_{ij^\star} \leftarrow 1$
            update $n$: $n_j \leftarrow n_j - 1$ and $n_{j^\star} \leftarrow n_{j^\star} + 1$
            update $q$: $q_j \leftarrow q_j - 2Q_j(x_i) + G_{ii}$ and
                        $q_{j^\star} \leftarrow q_{j^\star} + 2Q_{j^\star}(x_i) + G_{ii}$
        end if
    end for
until convergence
```

Neither of these two conditions are guaranteed to hold for Lloyd's method, and consequently for (kernel) k-means algorithm.

Theorem 5 (Telgarsky & Vattani (2010))**.** *The set of local optima of Hartigan's method is a (possibly strict) subset of local optima of Lloyd's method.*

This means that Hartigan's can potentially escape local optima of Lloyd's method. Therefore, kernel k-means cannot improve on a local optima of $\mathcal{E}^H$-clustering, but on the other hand, $\mathcal{E}^H$-clustering might improve on a local optima of kernel k-means. Lloyd's method forms Voronoi partitions, while Hartigan's method groups data into regions called circlonoi cells. The circlonoi cells are within a smaller volume of a Voronoi cell, and this excess volume grows exponentially with the dimension of $\mathcal{X}$ (Telgarsky & Vattani, 2010, Theorems 2.4 and 3.1). Points in this excess volume force Hartigan's method to iterate, contrary to Lloyd's method. Moreover, this improvement should be more prominent as dimension increases. Also, the improvement grows as the number of clusters k increases. The empirical results of Telgarsky & Vattani (2010) show that an implementation of Hartigan's method has comparable execution time to an implementation of Lloyd's method, but no explicit complexity was provided. In our case, both $\mathcal{E}^H$-clustering and kernel k-means have the same time complexity. To the best of our knowledge, Hartigan's method was not previously considered together with kernels, as we are proposing here.

In Slonim et al. (2013), Hartigan's method was applied to k-means problem with any Bregman divergence. It was shown that the number of Hartigan's local optima is upper bounded by $\mathcal{O}(1/k)$. In addition, examples were provided where *any* initial partition correspond to a local optima of Lloyd's method, while the number of local optima of Hartigan's method is small and correspond to true partitions of the data. Empirically, the number of local optima of Hartigan's method was considerably smaller than that of Lloyd's method.

5. Numerical Experiments

5.1. Synthetic Experiments

The goal of this section is threefold. First, to compare $\mathcal{E}^H$-clustering in Euclidean space to k-means and GMM. Second, to compare $\mathcal{E}^H$-clustering to kernel k-means and also to spectral clustering, when they all operate on the same kernel. Third, to show the flexibility of energy clustering since it is able to perform accurately in different settings using the same kernel.

Experimental Setup. The following setup holds unless specified otherwise. We consider $\mathcal{E}^H$-clustering, kernel k-means and spectral clustering with generating kernels to the following metrics:

$$\rho_\alpha(x,y) = \|x-y\|^\alpha, \tag{21a}$$

$$\widetilde{\rho}_\sigma(x,y) = 2 - 2e^{-\frac{\|x-y\|}{2\sigma}}, \tag{21b}$$

$$\widehat{\rho}_\sigma(x,y) = 2 - 2e^{-\frac{\|x-y\|^2}{2\sigma^2}}. \tag{21c}$$

These are generated by formula (8) and we fix $x_0 = 0$. The standard ρ_1 will always be present in the experiments as a reference, being the implied choice unless explicitly mentioned. For k-means, GMM and spectral clustering we use the robust implementations of *scikit-learn* library (Pedregosa et al., 2011), where k-means is initialized with k-means++ (Arthur & Vassilvitskii, 2007), and GMM with the output of k-means, making it more robust and preventing it from breaking in high dimensions. Spectral clustering is based on Shi & Malik (2000). We implemented kernel k-means as described in Dhillon et al. (2004; 2007), and $\mathcal{E}^H$-clustering as described in Algorithm 1. Both will also be initialized with k-means++. We run each algorithm 5 times with different initializations, picking the result with best objective function value. We evaluate clustering quality by the *accuracy*, which is the fraction of correctly clustered points based on the ground truth. More precisely, let Z be the true label matrix and $\hat{Z}$ the estimated label matrix produced by an algorithm. The accuracy is then defined as

$$\text{accuracy}(\hat{Z}) \equiv \max_{\pi} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \hat{Z}_{i\pi(j)} Z_{ij}, \tag{22}$$

where π is a permutation of the k cluster groups. The accuracy is always between $[0,1]$, where 1 correspond to all points correctly clustered, and 0 to all points wrongly clustered. For each setting we show the average accuracy over 100 Monte Carlo trials, with error bars being standard error.

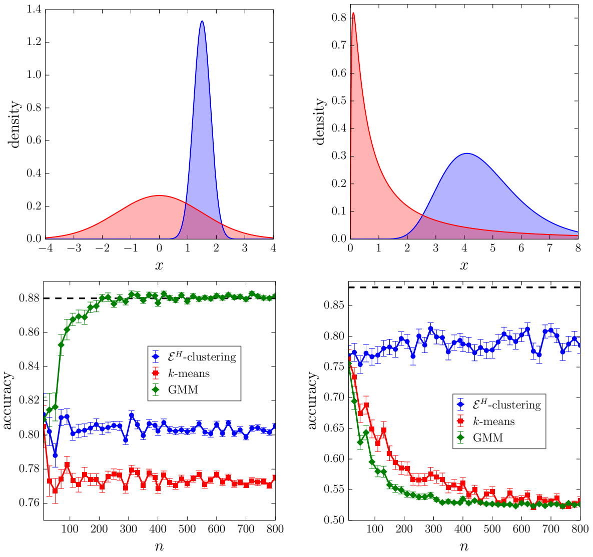

Figure 1. Clustering results for one dimensional data (23)–(25). *Top Left:* Probability density of (23). *Top Right:* Probability density of (24). *Bottom Left:* Clustering data from (23). *Bottom Right:* Clustering data from (24). The dashed lines show the optimal Bayes accuracy, which in both cases are ≈ 0.88.

The first two experiments involve univariate normal and lognormal mixtures:

$$x \overset{iid}{\sim} \tfrac{1}{2}\mathcal{N}(\mu_1, \sigma_1^2) + \tfrac{1}{2}\mathcal{N}(\mu_2, \sigma_2^2) \tag{23}$$

$$x \overset{iid}{\sim} \tfrac{1}{2}\exp\left\{\mathcal{N}(\mu_1, \sigma_1^2)\right\} + \tfrac{1}{2}\exp\left\{\mathcal{N}(\mu_2, \sigma_2^2)\right\} \tag{24}$$

$$\mu_1 = 1.5,\ \sigma_1 = 0.3,\ \mu_2 = 0,\ \sigma_2 = 1.5. \tag{25}$$

We increase the sample size n and show clustering accuracy (22) versus n in Fig. 1. $\mathcal{E}^H$-clustering performs better than k-means for normally distributed data, and worse than GMM, as expected. However, for lognormal distributions k-means and GMM are only slightly better than chance while $\mathcal{E}^H$-clustering is still accurate. The reason is that k-means and GMM models are inconsistent with this lognormal data. Energy clustering on the other hand is model-free.

Next we analyze how the algorithms degrade as the number of dimensions increase. Consider data from a Gaussian mixture $x \overset{iid}{\sim} \frac{1}{2}\mathcal{N}(\mu_1, \Sigma_1) + \frac{1}{2}\mathcal{N}(\mu_2, \Sigma_2)$ in $\mathbb{R}^D$, with $\Sigma_1 = \Sigma_2 = I_D$, and

$$\mu_1 = (\underbrace{0, \dots, 0}_{\times D})^\top,\ \mu_2 = 0.7(\underbrace{1, \dots, 1}_{\times 10}, \underbrace{0, \dots, 0}_{\times (D-10)})^\top. \tag{26}$$

Bayes error is fixed as D increases giving an optimal accuracy of ≈ 0.86. We sample 200 points on each trial. As

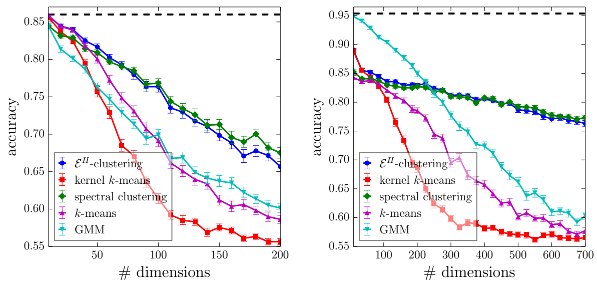

Figure 2. High dimensional Gaussian mixtures. *Left:* Parameters as in (26). *Right:* Parameters as in (27). The dashed lines are Bayes accuracy.

shown in Fig. 2 (left), $\mathcal{E}^H$- and spectral clustering have close performance, much better than kernel k-means, k-means and GMM. The improvement is noticeable in higher dimensions. Still for a two-class Gaussian mixture we now choose different numbers for the diagonal covariance Σ_2. We have $\Sigma_1 = I_D$, $\mu_1 = (0,\dots,0)^\top \in \mathbb{R}^D$, $\mu_2 = (1,\dots,1,0,\dots,0)^T \in \mathbb{R}^D$ with signal in the first 10 dimensions, and

$$\begin{aligned} \Sigma_2 &= \left(\begin{array}{c|c} \widetilde{\Sigma}_{10} & 0 \\ \hline 0 & I_{D-10} \end{array}\right), \\ \widetilde{\Sigma}_{10} &= \operatorname{diag}(1.367, 3.175, 3.247, 4.403, 1.249, \\ &\qquad 1.969, 4.035, 4.237, 2.813, 3.637). \end{aligned} \tag{27}$$

We simply chose 10 numbers uniformly at random between $[1,5]$ and other choice would give analogous results. Bayes accuracy is fixed at ≈ 0.95. From Fig. 2 (right) we see that GMM is better in low dimensions, but it quickly degenerates as D increases, as kernel k-means and k-means, while $\mathcal{E}^H$- and spectral clustering remains much more stable.

Consider $x \overset{iid}{\sim} \frac{1}{2}\mathcal{N}(\mu_1, \Sigma_1) + \frac{1}{2}\mathcal{N}(\mu_2, \Sigma_2)$ in $\mathbb{R}^{20}$ with $\Sigma_1 = \frac{1}{2} I_{20}$, $\Sigma_2 = I_{20}$, and

$$\mu_1 = (\underbrace{0,\dots,0}_{\times 20})^\top, \qquad \mu_2 = \tfrac{1}{2}(\underbrace{1,\dots,1}_{5}, \underbrace{0,\dots,0}_{15})^\top. \tag{28}$$

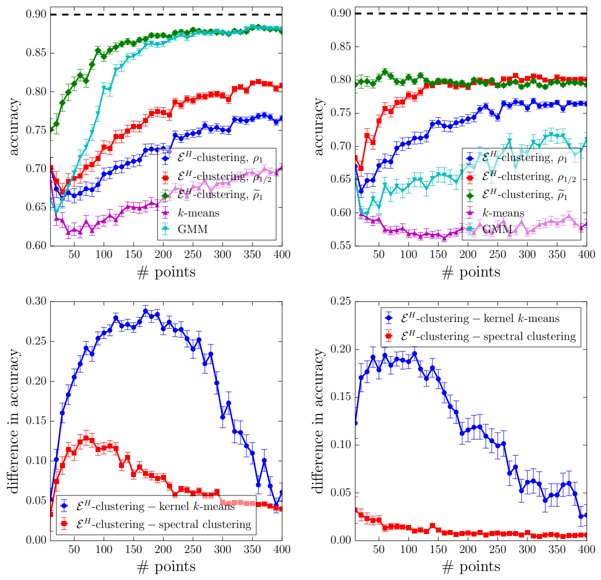

Figure 3. Normal *(Left)* and lognormal *(Right)* mixtures with parameters (28). We use different kernels for $\mathcal{E}^H$-clustering. Bayes accuracy is ≈ 0.9. The two plots in the *Bottom* show the difference in accuracy between $\mathcal{E}^H$-clustering versus kernel k-means and spectral clustering, with $\widetilde{\rho}_1$.

Bayes accuracy is ≈ 0.90. We increase the sample size $n \in [10, 400]$ and show the accuracy versus n in Fig. 3 (top-left), where we compare $\mathcal{E}^H$-clustering with different kernels, indicated in the legend, to k-means and GMM. Note that $\mathcal{E}^H$-clustering with $\widetilde{\rho}_1$ is as accurate as GMM for large n but superior for small n. For the same setting, in Fig. 3 (bottom-left) we we show the difference in accuracy provided by $\mathcal{E}^H$-clustering minus kernel k-means and spectral clustering when using $\widetilde{\rho}_1$. $\mathcal{E}^H$-clustering was always superior, otherwise there would be points with negative y-values. Consider the same parameters as in (28) but now with lognormal mixtures in $\mathbb{R}^D$. The same experiments are shown in Fig. 3 (top-right and bottom-right), where $\mathcal{E}^H$-clustering still performs accurately with any of those kernels, contrary to k-means and GMM.

In Fig. 4 we have examples of complex two dimensional datasets. We apply $\mathcal{E}^H$-clustering with different kernel choices, and also spectral clustering with the best kernel choice, besides k-means and GMM. Here we perform only 10 Monte Carlo runs. For parallel cigars we initialize all algorithms with k-means++, and the concentric circles example we initialize at random. The results are shown in Table 1. $\mathcal{E}^H$-clustering has superior performance in every example, in particular better than the spectral clustering. For parallel cigars the metrics ρ_1 and $\rho_{1/2}$ still provide accurate results, however, for the concentric circles the kernel choice is more sensitive.

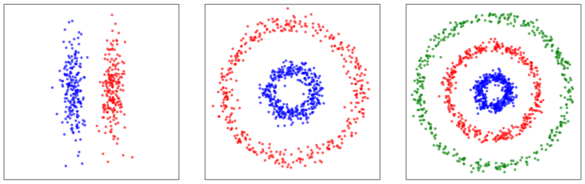

Figure 4. Left: Parallel cigars, 200 points each. *Center and Right:* Two and three concentric circles, respectively, with Gaussian noise. 400 points for each class.

Table 1. Clustering data from Figure 4.

		Parallel Cigars		*Two Circles*		*Three Circles*
$\mathcal{E}^H$*-clustering*	ρ_1	0.705 ± 0.065	ρ_1	0.521 ± 0.005	ρ_1	0.393 ± 0.020
	$\rho_{1/2}$	0.952 ± 0.048	$\rho_{1/2}$	0.522 ± 0.004	$\rho_{1/2}$	0.486 ± 0.040
	$\widetilde{\rho}_2$	$\mathbf{0.9987 \pm 0.0008}$	$\widetilde{\rho}_1$	0.778 ± 0.075	$\widetilde{\rho}_2$	0.666 ± 0.007
	$\widehat{\rho}_2$	0.956 ± 0.020	$\widehat{\rho}_1$	$\mathbf{1.0 \pm 0.0}$	$\widehat{\rho}_2$	$\mathbf{0.676 \pm 0.002}$
spectral-clustering	$\widetilde{\rho}_2$	0.557 ± 0.014	$\widehat{\rho}_1$	0.732 ± 0.002	$\widehat{\rho}_2$	0.364 ± 0.004
k*-means*	✗	0.550 ± 0.011	✗	0.522 ± 0.004	✗	0.368 ± 0.005
GMM	✗	0.903 ± 0.064	✗	0.595 ± 0.011	✗	0.465 ± 0.030

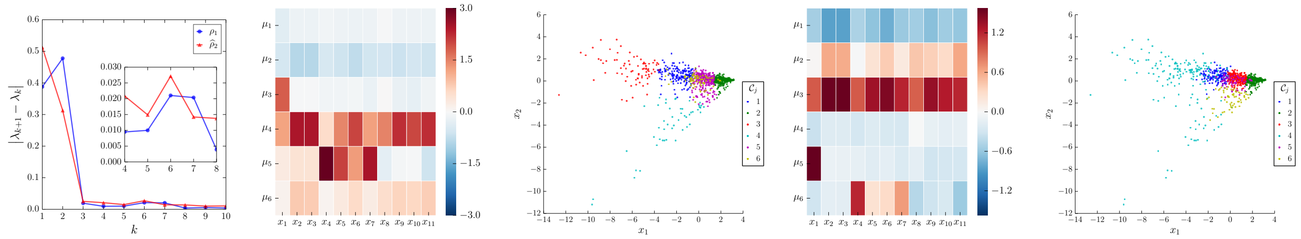

Figure 5. Clustering synapse dataset. Plots in order from *left* to *right*. *First:* Eigenvalues of random walk Laplacian. We choose $k = 6$ since its the first time we see a peak. We use energy clustering with the two metrics ρ_1 and $\widehat{\rho}_2$; see (21). *Second:* Heat map of the cluster means versus features using energy clustering with standard metric ρ_1 from energy statistics. *Third:* Clustered points projected into the two principal components, using ρ_1. *Fourth:* Heat map of the cluster means versus features using energy clustering with Gaussian metric $\widehat{\rho}_2$. *Fifth:* Clustered points projected into the two principal components, using $\widehat{\rho}_2$.

5.2. Real Data Experiment

The considered dataset describes protein expression of neural synapses (chemical connections between neurons), being crucial for understanding the diversity of synapses. This is part of a large NIH funded consortium within the BRAIN initiative; we refer to Collman et al. (2015) for details. We have 1025 data points with 11 features. Each point corresponds to a (x, y, z) location in the brain. We normalize this dataset before applying $\mathcal{E}^H$-clustering.

We use the two metrics ρ_1 and $\widehat{\rho}_2$; see (21). We compute the difference between eigenvalues of the random walk Laplacian obtained from the kernel matrix G (12). The choice of k corresponds to the first time we see a meaningful peak in the plot, which occurs at $k = 6$; first plot in Fig. 5. This procedure to find k is common in the literature (von Luxburg, 2007). Having found k we now cluster the data. First we use the metric ρ_1, which is the standard Euclidean choice in energy statistics. In the second and third plots of Fig. 5 we show a heat map of the cluster means versus features and the clustered data points projected into the 2 principal components, respectively. The fourth and fifth plots in Fig. 5 show the same experiment but using the metric $\widehat{\rho}_2$ corresponding to a Gaussian kernel.

We remark that we also used the gap statistic (Tibshirani et al., 2001) based on k-means to determine the number of clusters as a comparison. The results slightly suggest $k = 6$, but it does not pass the gap statistic test.

6. Final Remarks

We proposed clustering from the perspective of generalized energy statistics, valid for arbitrary spaces of negative type. Our mathematical formulation yields a QCQP in the associated RKHS; Proposition 2. We showed that such QCQP is equivalent to kernel k-means optimization problem, once the kernel is fixed; Proposition 3. However, energy statistics fixes a family of standard kernels in Euclidean space, and more general kernels on spaces of negative type as well.

We proposed the iterative $\mathcal{E}^H$-clustering (Algorithm 1) which is a kernelized version of Hartigan's method. It was compared to kernel k-means algorithm, based on Lloyd's heuristic. Both have the same complexity, however, numerical and theoretical results provide compelling evidence that $\mathcal{E}^H$-clustering is more robust with superior performance, specially in high dimensions. Moreover, $\mathcal{E}^H$-clustering with standard kernels from energy statistics outperformed k-means and GMM on several settings, even on Gaussian ones. In some cases $\mathcal{E}^H$-clustering also surpassed spectral clustering, and in others performed similarly but never worse. Note that computing eigenvectors of large matrices can be unfeasible and iterative methods are preferred. We provided an application of $\mathcal{E}^H$-clustering to an important real dataset describing protein expression of neural synapses. Finally, we remark that kernel methods can benefit from sparsity and fixed-rank approximations of the Gram matrix and there is plenty of room to make $\mathcal{E}^H$-clustering more scalable.

References

Aronszajn, N. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

Arthur, D. and Vassilvitskii, S. k-means++: The Advantage of Careful Seeding. In *Proceedings of the Eighteenth annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

Berg, C., Christensen, J. P. R., and Ressel, P. *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Graduate Text in Mathematics 100. Springer, New York, 1984.

Collman, F., Buchanan, J., Phend, K. D., Micheva, K. D., Weinberg, R. J., and Smith, S. J. ”mapping synapses by conjugate light-electron array tomography”. *The Journal of Neuroscience*, 34(14):5792–5807, 2015.

Dhillon, I. S., Guan, Y., and Kulis, B. Kernel K-means: Spectral Clustering and Normalized Cuts. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’04, pp. 551–556, New York, NY, USA, 2004. ACM.

Dhillon, I. S., Guan, Y., and Kulis, B. Weighted Graph Cuts without Eigenvectors: A Multilevel Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11):1944–1957, 2007.

Filippone, M., Camastra, F., Masulli, F., and Rovetta, S. A Survey of Kernel and Spectral Methods for Clustering. *Pattern Recognition*, 41:176–190, 2008.

Forgy, E. Cluster Analysis of Multivariate Data: Efficiency versus Interpretabiliby of Classification. *Biometrics*, 21 (3):768–769, 1965.

Girolami, M. Kernel Based Clustering in Feature Space. *Neural Networks*, 13(3):780–784, 2002.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Hartigan, J. A. and Wong, M. A. Algorithm AS 136: A k-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1): 100–108, 1979.

J. B. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 281–297. University of California Press, 1967.

Li, S. k-Groups: A Generalization of k-Means by Energy Distance. PhD Thesis, Bowling Green State University, 2015.

Lloyd, S. P. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

Lyons, R. Distance Covariance in Metric Spaces. *The Annals of Probability*, 41(5):3284–3305, 2013.

Mercer, J. Functions of Positive and Negative Type and their Connection with the Theory of Integral Equations. *Proceedings of the Royal Society of London*, 209:415–446, 1909.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Rizzo, M. L. and Székely, G. J. DISCO Analysis: A Nonparametric Extension of Analysis of Variance. *The Annals of Applied Statistics*, 4(2):1034–1055, 2010.

Schölkopf, B., Smola, A. J., and Müller, K. R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10:1299–1319, 1998.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. Equivalence of Distance-Based and RKHS-Based Statistic in Hypothesis Testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

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

Slonim, N., Aharoni, E., and Crammer, K. Hartigan’s k-Means versus Lloyd’s k-Means — Is it Time for a Change? In *Proceedings of the 20th International Conference on Artificial Intelligence*, pp. 1677–1684. AAI Press, 2013.

Székely, G. J. and Rizzo, M. L. Hierarchical Clustering via Joint Between-Within Distances: Extending Ward’s Minimum Variance Method. *Journal of Classification*, 22(2):151–183, 2005.

Székely, G. J. and Rizzo, M. L. Energy Statistics: A Class of Statistics Based on Distances. *Journal of Statistical Planning and Inference*, 143:1249–1272, 2013.

Telgarsky, M. and Vattani, A. Hartigan’s Method: k-Means Clustering without Voronoi. In *Proceedings of the 13th International Conference on Artificial Intelligence and*

Statistics (AISTATS), volume 9, pp. 313–319. JMLR, 2010.

Tibshirani, R., Walther, G., and Hastie, T. ”estimating the number of clusters in a data set via the gap statistic”. *J. Royal Statistical Society B*, 63(2):411–423, 2001.

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

A. Supplementary Material

Here we collect the proofs of our main results.

Proof of Lemma 1. From (4) and (5) we have

$$
\begin{aligned}
S + W &= \frac{1}{2n} \sum_{\substack{i,j=1 \\ i\neq j}}^{k} n_i n_j g(\mathcal{C}_i, \mathcal{C}_j) \\
&\quad + \frac{1}{2n} \sum_{i=1}^{k} \Big[n - \sum_{j\neq i=1}^{k} n_j \Big] n_i g(\mathcal{C}_i, \mathcal{C}_i) \\
&= \frac{1}{2n} \sum_{i,j=1}^{k} n_i n_j g(\mathcal{C}_i, \mathcal{C}_j) \\
&= \frac{1}{2n} \sum_{x\in\mathbb{X}} \sum_{y\in\mathbb{X}} \rho(x,y) \\
&= \frac{n}{2} g(\mathbb{X},\mathbb{X}).
\end{aligned}
\tag{29}
$$

Note that the right hand side of this equation only depends on the pooled data, so it is a constant independent of the choice of partition. Therefore, maximizing S over the choice of partition is equivalent to minimizing W. □

Proof of Proposition 2. From (9), (3), and (4) we have

$$
\begin{aligned}
W &= \frac{1}{2} \sum_{j=1}^{k} \frac{1}{n_j} \sum_{x,y\in\mathcal{C}_j} \rho(x,y) \\
&= \sum_{j=1}^{k} \sum_{x\in\mathcal{C}_j} \bigg(K(x,x) - \frac{1}{n_j} \sum_{y\in\mathcal{C}_j} K(x,y) \bigg).
\end{aligned}
\tag{30}
$$

Note that the first term is global so it does not contribute to the optimization problem. Therefore, minimizing (30) is equivalent to

$$
\max_{\mathcal{C}_1,\dots,\mathcal{C}_k} \sum_{j=1}^{k} \frac{1}{n_j} \sum_{x,y\in C_j} K(x,y). \tag{31}
$$

But

$$
\sum_{x,y\in\mathcal{C}_j} K(x,y) = \sum_{p=1}^{n} \sum_{q=1}^{n} Z_{pj} Z_{qj} G_{pq} = (Z^\top G\, Z)_{jj}, \tag{32}
$$

where we used the definitions (12) and (13). Notice that $n_j^{-1} = D_{jj}^{-1}$, where the diagonal matrix $D = \operatorname{diag}(n_1,\dots,n_k)$ contains the number of points in each cluster, thus the objective function in (31) is equal to $\sum_{j=1}^{k} D_{jj}^{-1} \big(Z^\top G Z\big)_{jj} = \operatorname{Tr}\big(D^{-1} Z^\top G Z\big)$. Now we can use the cyclic property of the trace, and by the definition of the matrix Z in (13), we obtain the following integer programing problem:

$$
\begin{aligned}
&\max_{Z} \operatorname{Tr}\Big(\big(ZD^{-1/2}\big)^\top G \big(ZD^{-1/2}\big) \Big) \\
&\text{s.t. } Z_{ij}\in\{0,1\},\ \textstyle\sum_{j=1}^{k} Z_{ij} = 1,\ \sum_{i=1}^{n} Z_{ij} = n_j.
\end{aligned}
\tag{33}
$$

Now we write this in terms of the matrix $Y = ZD^{-1/2}$. The objective function immediately becomes $\operatorname{Tr}\big(Y^\top G\,Y\big)$. Notice that the above constraints imply that $Z^T Z = D$, which in turn gives $D^{-1/2} Y^T Y D^{-1/2} = D$, or $Y^\top Y = I$. Also, every entry of Y is positive by definition, $Y \geq 0$. Now it only remains to show the last constraint in (14), which comes from the last constraint in (33). In matrix form this reads $Z^T e = De$. Replacing $Z = YD^{1/2}$ we have $Y^\top e = D^{1/2} e$. Multiplying this last equation on the left by Y, and noticing that $YD^{1/2}e = Ze = e$, we finally obtain $YY^\top e = e$. Therefore, the optimization problem (33) is equivalent to (14) . □

Proof of Proposition 3. Notice that

$$
\begin{aligned}
\|\varphi(x) - \varphi(\mu_j)\|^2 = \langle \varphi(x), \varphi(x)\rangle &- 2\langle \varphi(x), \varphi(\mu_j)\rangle \\
&+ \langle \varphi(\mu_j), \varphi(\mu_j)\rangle,
\end{aligned}
\tag{34}
$$

therefore, kernel k-means objective function becomes

$$
\begin{aligned}
\sum_{j=1}^{k} \sum_{x\in\mathcal{C}_j} \bigg(K(x,x) &- \frac{2}{n_j} \sum_{y\in\mathcal{C}_j} K(x,y) \\
&+ \frac{1}{n_j^2} \sum_{y,z\in\mathcal{C}_j} K(y,z) \bigg).
\end{aligned}
\tag{35}
$$

The first term is global so it does not contribute to the optimization problem. Notice that the third term gives $\sum_{x\in\mathcal{C}_j} \frac{1}{n_j^2} \sum_{y,z\in\mathcal{C}_j} K(y,z) = \frac{1}{n_j} \sum_{y,z\in\mathcal{C}_j} K(y,z)$, which is the same as the second term. Thus, problem (15) is equivalent to

$$
\max_{\mathcal{C}_1,\dots,\mathcal{C}_k} \sum_{j=1}^{k} \frac{1}{n_j} \sum_{x,y\in C_j} K(x,y) \tag{36}
$$

which is exactly the same as (31) from the energy statistics formulation. Therefore, once the kernel K is fixed, the function W given by (4) is the same as J in (15). The remaining of the proof proceeds as already shown in the proof of Proposition 2, leading to the optimization problem (14). □