Clustering via Generalized Energy Statistics

Guilherme França **Joshua T. Vogelstein**
Johns Hopkins University

Abstract

Energy statistics introduces the notion of potential energy between probability distributions, in close analogy to Newton's gravitational potential in physics. We propose a principled approach to the clustering problem based on energy statistics theory. Our mathematical formulation establishes connection to kernel methods, leading to a quadratically constrained quadratic program (QCQP). To obtain local solutions of such NP-hard QCQP we introduce an iterative algorithm based on Hartigan's method, which has the same computational cost but offers several advantages compared to kernel k-means, based on Lloyd's heuristic. We provide carefully designed numerical experiments showing that our 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.

1 Introduction

Energy statistics (Székely and Rizzo, 2013) provides a hypothesis test for equality of distributions which is achieved under minimum statistical potential energy. When probability distributions are different the statistical potential energy diverges as sample size increases, while tends to a nondegenerate limit distribution when probability distributions are equal. The test statistic has 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, multisample tests of equality of distributions, analysis of variance (Rizzo and Székely, 2010), nonlinear dependence tests through distance covariance and distance correlation, which generalizes the Pearson correlation coefficient, and hierarchical clustering (Székely and Rizzo, 2005) by extending Ward's method of minimum variance. Moreover, an application of energy statistics to clustering, in Euclidean spaces, was proposed (Li, 2015), which in part motivated this paper. We refer to (Székely and Rizzo, 2013) for an overview.

Distance covariance was further generalized from Euclidean to metric spaces of strong negative type (Lyons, 2013). Furthermore, the missing link between energy distance based tests and kernel based tests has been recently resolved (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.

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 the-

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. JMLR: W&CP volume 7X.

ory 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 heuristic (Lloyd, 1982) of grouping points that are closer to a cluster center, now in 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 with kernel methods, showing that this QCQP is equivalent to kernel k-means optimization problem.

Our main algorithmic contribution is to use Hartigan's method (Hartigan and Wong, 1979) to find local solutions of the above mentioned QCQP, which is NP-hard in general. Hartigan's method was also used in (Li, 2015), without any connection to kernels. More importantly, its advantages over Lloyd's method was already demonstrated in some simple settings (Telgarsky and Vattani, 2010; Slonim et al., 2013), but apparently this method did not receive the deserved attention. To the best of our knowledge, Hartigan's method was not previously employed together with kernel methods. We provide a full 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, but Hartigan's method in kernel spaces offer several advantages. Furthermore, in the examples considered in this paper, it also provides superior performance compared to a spectral clustering. 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 different distributions, contrary to k-means and GMM for instance. The proposed algorithm 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.

2 Background

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

Consider random vectors X_i 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 $\sum_{i,j=1}^{n} c_i c_j \rho(X_i, X_j) \leq 0$, 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, $\mathcal{E}(P, Q) \geq 0$, 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* (Székely and Rizzo, 2013) in Euclidean spaces uses the semimetric

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

where $0 < \alpha \leq 2$ 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 data $\mathbb{X} = \{x_1, \ldots, x_n\}$ sampled from k unknown distributions $\{P_j\}_{j=1}^{k}$, with points lying on a space of negative type, $x_i \in \mathcal{X}$. Let $\mathbb{X} = \bigcup_{j=1}^{k} \mathcal{C}_j$ be a disjoint partition, $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$. 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 $n_i = |\mathcal{C}_i|$ is the number of elements 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

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

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 is a test statistic for equality of distributions (Székely and 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, i.e. $K(x_i, x_j) = K(x_j, x_i)$ and $\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) \geq 0$, or equivalently, if G is the Gram matrix with entries $G_{ij} = K(x_i, x_j)$ then $G = G^\top$ and $v^\top G v \geq 0$ for any $v \in \mathbb{R}^n$. 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 (Aronszajn, 1950), 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 $\langle \varphi(x), \varphi(y) \rangle = K(x, y)$.

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 $\langle h_P, h_Q \rangle_{\mathcal{H}_K} = \mathbb{E}\, K(X, Y)$, thus in practice we can estimate the inner product between embedded distributions by averaging the kernel function over sampled data.

The following important result shows that semimetrics of negative type and symmetric positive definite kernels are closely related (Berg et al., 1984). Let $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $x_0 \in \mathcal{X}$ 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 have a family of kernels, one for each choice of x_0. Conversely, if ρ is a semimetric of negative type and K is a kernel in this family, then

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

and the canonical feature map $\varphi : x \mapsto h_x$ is injective (Sejdinovic et al., 2013). When these conditions are satisfied 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

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

so we can compute the energy distance using the inner product of $\mathcal{H}_K$.

3 Clustering via Energy Statistics

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

Due to the energy test statistic for equality of distributions, the obvious criterion for clustering data is to maximize S 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 $\mathbb{X} = \{x_1, \ldots, x_n\}$ where each data point x_i lives in a space $\mathcal{X}$ of negative type. For a fixed integer k, the partition $\mathbb{X} = \bigcup_{j=1}^{k} \mathcal{C}_j$, where $\mathcal{C}_i \cap C_j = \emptyset$ for all $i \neq j$, maximizes the between-sample statistic S, defined in equation (5), if and only if*

$$\min_{\mathcal{C}_1, \ldots, C_k} W(\mathcal{C}_1, \ldots, \mathcal{C}_k), \tag{10}$$

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

In the Euclidean case, the optimization problem (10) 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. In this current form, the relation with kernels and other clustering methods is obscure. In the following, we show what is the explicit optimization problem behind (10) in the corresponding RKHS, establishing the connection with kernel methods.

Assume that the kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ generates ρ.

Define the Gram matrix $G \in \mathbb{R}^{n\times n}$ as

$$G_{ij} \equiv K(x_i, x_j). \tag{11}$$

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 $D = \mathrm{diag}(n_1, \ldots, n_k)$ contains the number of points in each cluster. We also introduce the rescaled matrix $Y \equiv ZD^{-1/2}$. In component form they are given by

$$Z_{ij} \equiv \begin{cases} 1 & \text{if } x_i \in \mathcal{C}_j \\ 0 & \text{otherwise} \end{cases} \quad Y_{ij} \equiv \begin{cases} \frac{1}{\sqrt{n_j}} & \text{if } x_i \in \mathcal{C}_j \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

Throughout the paper, we use the notation $M_{i\bullet}$ to denote the ith row of a matrix M, and $M_{\bullet j}$ denotes its jth column. Our next result shows that the optimization problem (10) is NP-hard since it is a quadratically constrained quadratic program (QCQP) in the RKHS.

Proposition 2. *The optimization problem* (10) *is equivalent to*

$$\begin{aligned} &\max_{Y} \mathrm{Tr}\left(Y^\top G Y\right) \\ &\text{s.t. } Y \geq 0,\ Y^\top Y = I,\ YY^\top e = e, \end{aligned} \tag{13}$$

where $e = (1, \ldots, 1)^\top \in \mathbb{R}^n$ *is the all-ones vector, and* G *is the Gram matrix* (11).

Therefore, to group data $\mathbb{X} = \{x_1, \ldots, x_n\}$ into k clusters we first compute the Gram matrix G and then solve the optimization problem (13) 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 $x_i \in \mathcal{C}_j$. This optimization problem is nonconvex, and also NP-hard, thus a direct approach is computational prohibitive even for small datasets. However, one can find approximate solutions by relaxing some of the constraints. For instance, the relaxed problem $\max_Y \mathrm{Tr}\left(Y^\top G Y\right)$ 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 following.

Note that the energy clustering problem (13) 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 would be 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 depend on the choice of a suitable ρ which measures the similarity between data points. If prior information is available to choose ρ it can be conveniently incorporated in the optimization problem (13).

Relation to Kernel k-Means. One may wonder how energy clustering relates to the well-known kernel k-means problem[1], extensively used in machine learning. For a positive semidefinite Gram matrix G, as defined in (11), 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, in feature space, is defined by

$$\min_{\mathcal{C}_1,\ldots,\mathcal{C}_k} \left\{ \sum_{j=1}^{k} \sum_{x\in\mathcal{C}_j} \|\varphi(x) - \varphi(\mu_j)\|^2 \right\} \tag{14}$$

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 is known (Dhillon et al., 2004, 2007) that kernel k-means problem can be cast into a trace maximization in the same form as (13). The next result makes this explicit.

Proposition 3. *For a fixed kernel, the energy clustering optimization problem* (10) *is equivalent to the kernel* k*-means optimization problem* (14)*, and both are equivalent to* (13).

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 (13). Thus one can view all these problems arising from energy statistics as well.

4 Hartigan's for Energy Clustering

We introduce an iterative algorithm based on Hartigan's method (Hartigan and Wong, 1979) to find a local maximizer of the optimization problem (13). Due to Proposition 3 we can use kernel k-means algorithm

[1]When we refer to kernel k-means problem we mean specifically the optimization problem (14), which should not be confused with kernel k-means algorithm that is just one possible recipe to solve (14).

(Dhillon et al., 2004, 2007), which will be compared with the proposed algorithm.

We can write the optimization problem (13) as

$$\max_{\{\mathcal{C}_1,\ldots,\mathcal{C}_k\}} \left\{ Q = \sum_{j=1}^{k} \frac{Q_j}{n_j} \right\}, \quad Q_j \equiv \sum_{x,y\in\mathcal{C}_j} K(x,y), \tag{15}$$

where Q_j represents an internal energy cost of cluster $\mathcal{C}_j$, and Q is the total energy 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 energy cost with the entire cluster $\mathcal{C}_\ell$ by

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

where we recall that $G_{i\bullet}$ ($G_{\bullet i}$) denotes the ith row (column) of matrix G.

For a given configuration, we consider the maximum 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 cost change $\Delta Q^{j\to\ell}(x_i) \equiv Q^{(\ell)} - Q^{(j)}$, for $\ell \neq j$. From (15), by explicitly writing the costs related to these two cluster we obtain

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

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

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

Therefore, if $\Delta Q^{j\to\ell}(x_i) > 0$ we get closer to a maximum of (15) by moving x_i to $\mathcal{C}_\ell$, otherwise we should keep x_i in $\mathcal{C}_j$.

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 $\ell = 1,\ldots,k$ with $\ell \neq j$. Then, choose

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

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. The entire procedure is explicitly described in Algorithm 1, which we call $\mathcal{E}^H$-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.

Algorithm 1 $\mathcal{E}^H$-clustering is Hartigan's method to find local solutions to the optimization problem (13).

input number of clusters k, Gram matrix G, initial label matrix $Z \leftarrow Z_0$
output label matrix Z

$q \leftarrow (Q_1,\ldots,Q_k)^\top$ have the energy costs of each cluster, defined in (15)
$n \leftarrow (n_1,\ldots,n_k)^\top$ have the number of points in each cluster
repeat
  for $i = 1,\ldots,n$ **do**
    let j be such that $x_i \in \mathcal{C}_j$
    $j^\star \leftarrow \arg\max_{\ell=1,\ldots,k \,|\, \ell\neq j} \Delta Q^{j\to\ell}(x_i)$
    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

The worst time complexity of $\mathcal{E}^H$-clustering is $\mathcal{O}(nk\max_\ell n_\ell) = \mathcal{O}(kn^2)$. This 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 and Vattani (2010)). *If $n > k$ the resulting partition obtained from Hartigan's method has 1. no empty clusters, and 2. distinct means.*

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

Theorem 5 (Telgarsky and 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 method can potentially escape local optima of Lloyd's method. 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 in 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 and Vattani, 2010, Theorems 2.4 and 3.1). Points in this excess volume force Hartigan's to iterate, contrary to Lloyd's. 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 and 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 Lloyd's local optima, while the number of Hartigan's local optima is small and correspond to true partitions of the data. Empirically, the number of Hartigan's local optima was considerably smaller.

5 Numerical Experiments

The main goal of this section is threefold. First, to compare $\mathcal{E}$-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, able to perform accurately in different settings using the same kernel.

The following experimental setup holds unless specified otherwise. We consider $\mathcal{E}$-clustering, kernel k-means and spectral clustering with semimetrics

$$\rho_\alpha(x,y) = \|x-y\|^\alpha,$$
$$\widetilde{\rho}_\sigma(x,y) = 2 - 2e^{-\frac{\|x-y\|}{2\sigma}},$$
$$\widehat{\rho}_\sigma(x,y) = 2 - 2e^{-\frac{\|x-y\|^2}{2\sigma^2}}$$

These are generated by the kernel (8) where we fix $x_0 = 0$. The standard ρ_1, from energy statistics in Euclidean spaces, will always be present as a reference and it is implied unless explicitly specified. 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 and Vassilvitskii, 2007), and GMM with the output of k-means, making it more robust and not breaking in high dimensions. 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 the algorithms 5 times with different initializations, picking the result with best objective. We evaluate clustering quality by the *accuracy* based on the true labels. For each setting we show the average accuracy over 100 Monte Carlo trials, with error bars indicating the standard error.

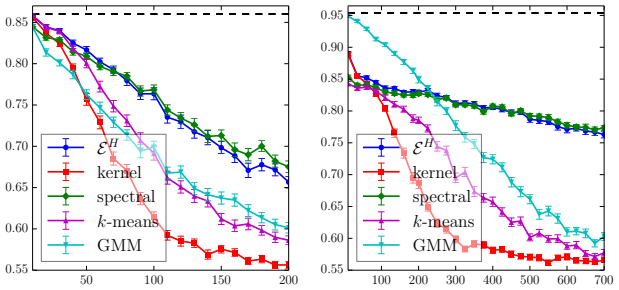

Figure 1: High dimensional Gaussian mixture. Comparison of $\mathcal{E}^H$-clustering, kernel k-means, spectral clustering, k-means and GMM (last three from scikit-learn). (a) Parameters as in (16). (b) Parameters as in (17). Dashed line is Bayes accuracy. We plot mean accuracy versus number of dimensions D over 100 Monte Carlo trials, error bars are standard error.

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

$$\mu_1 = (\underbrace{0,\dots,0}_{\times D})^\top,\ \mu_2 = 0.7(\underbrace{1,\dots,1}_{\times 10},\underbrace{0,\dots,0}_{\times(D-10)})^\top. \quad (16)$$

The Bayes error gives an optimal accuracy ≈ 0.86, which is fixed as D increases. We sample 200 points for each trial. The results are shown in Fig. 1a, where $\mathcal{E}^H$- and spectral-clustering have close performance, much better than kernel k-means, and also better than 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

$$\Sigma_2 = \left(\begin{array}{c|c} \widetilde{\Sigma}_{10} & 0 \\ \hline 0 & I_{D-10} \end{array}\right),$$
$$\widetilde{\Sigma}_{10} = \mathrm{diag}(1.367, 3.175, 3.247, 4.403, 1.249, 1.969, 4.035, 4.237, 2.813, 3.637). \quad (17)$$

We simply chose 10 number uniformly at random on the interval $[1,5]$, and any other choice would give

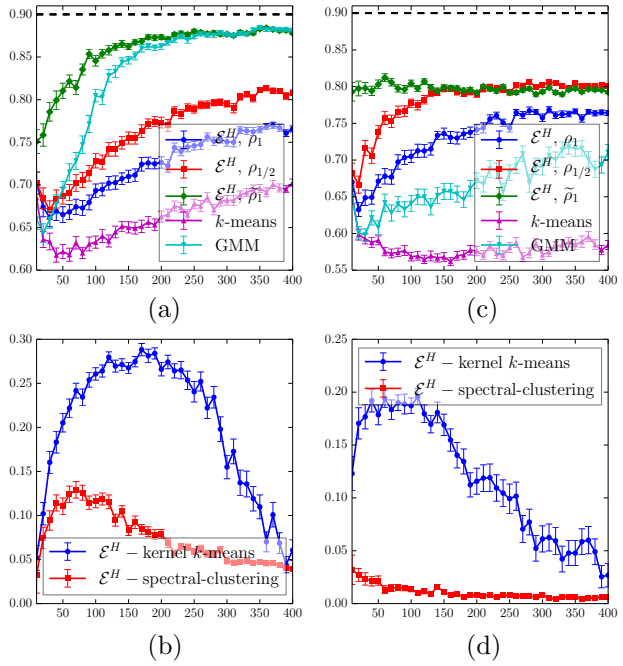

Figure 2: (a,b) Normal and (c,d) lognormal distributions in $\mathbb{R}^{20}$ with parameters (18). We use different kernels for $\mathcal{E}^H$-clustering, compared to k-means and GMM. Bayes accuracy is ≈ 0.9. We plot average accuracy (error bars are standard error) versus number of points for 100 Monte Carlo trials. The plots in (c) and (d) consider the difference in accuracy between $\mathcal{E}^H$ versus kernel k-means and spectral clustering, with $\widetilde{\rho}_1$.

analogous results. The Bayes accuracy is fixed at ≈ 0.95. Fig. 1b we show the results. GMM performs better in low dimensions, but it quickly degenerates as D increases, as (kernel) k-means, while $\mathcal{E}^H$- and spectral-clustering remains much more stable.

Now we sample n points from the Gaussian mixture $x \overset{iid}{\sim} \frac{1}{2}\left[\mathcal{N}(\mu_1, \Sigma_1) + \mathcal{N}(\mu_2, \Sigma_2)\right]$ in $\mathbb{R}^{20}$ with $\Sigma_1 = \frac{1}{2} I_{20}$, $\Sigma_2 = I_{20}$,

$$\mu_1 = (\underbrace{0, \dots, 0}_{\times 20})^\top, \quad \mu_2 = \tfrac{1}{2}(\underbrace{1, \dots, 1}_{5}, \underbrace{0, \dots, 0}_{15})^\top. \quad (18)$$

Bayes accuracy is ≈ 0.90. We increase the sample size in the range $n \in [10, 400]$. The results are shown in Fig. 2a, 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 number of points, however, it is superior for small number of points. Still for the same setting, in Fig. 2b we we show the difference in accuracy provided by $\mathcal{E}^H$ minus kernel k-means and $\mathcal{E}^H$ minus spectral clustering, when using the semimetric $\widetilde{\rho}_1$. Note that $\mathcal{E}^H$ was always superior (there would be points with negative values on the y-axis otherwise). Now consider the same parameters as in (18) but with lognormal mixtures in $\mathbb{R}^D$, $\log x \overset{iid}{\sim} \frac{1}{2}\left[\mathcal{N}(\mu_1, \Sigma_1) + \mathcal{N}(\mu_2, \Sigma_2)\right]$. The same experiments are shown in Fig. 2c and Fig. 2d. Note that $\mathcal{E}^H$-clustering still performs accurately, with any of those kernels, providing better results than k-means and GMM. These two experiments illustrate how energy clustering is flexible, since it is accurate on data from very different distributions with the same kernel.

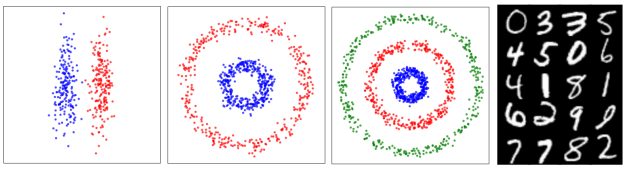

Figure 3: (a) Parallel cigars, 200 points each. (b,c) Two and three concentric circles with Gaussian noise, 400 points each. (d) MNIST handwritten digits. Clustering results are in Tables 1 and 2.

In Fig. 3a–c we have complex two dimensional datasets. We apply $\mathcal{E}^H$-clustering with different kernel choices. We also consider the best kernel choice for spectral clustering, besides k-means and GMM. Here we perform only 10 Monte Carlo runs. In (a) we initialize all algorithms with k-means++, and in (b) and (c) we initialize at random. The results are in Table 1. $\mathcal{E}^H$-clustering has superior performance in every example, and in particular better than the spectral clustering. For the data in Fig. 3a the semimetrics ρ_1 and $\rho_{1/2}$ still provide accurate results, however, for the examples in (b) and (c) the kernel choice is more sensitive.

Next, we consider the infamous MNIST handwritten digits as illustrated in Fig. 3d. Each data point is an 8-bit gray scale image forming a 784-dimensional vector corresponding to the digits $\{0, 1, \dots, 9\}$. We compute the parameter

$$\sigma^2 = \frac{1}{n^2} \sum_{i,j=1}^{n} \|x_i - x_j\|^2$$

from a separate training set, to be used in the kernels. We consider subsets of $\{0, 1, \dots, 9\}$, sampling 100 points for each class over 10 Monte Carlo trials. The results are shown in Table 2. Unsupervised clustering on MNIST without any feature extraction is not trivial. This same experiment was performed in (Qui and Sapiro, 2015) where a low-rank transformation is learned then subsequently used in subspace clustering. It would be interesting to explore methods for learning a better representation of the data and subsequently apply $\mathcal{E}^H$-clustering.

Table 1: Clustering data from Fig. 3a–c.

		Fig. 3a		*Fig. 3b*		*Fig. 3c*
$\mathcal{E}^H$*-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

Table 2: Clustering MNIST data from Fig. 3d.

Class Subset		$\{0,1,2\}$	$\{0,1,\ldots,4\}$	$\{0,1,\ldots,6\}$	$\{0,1,\ldots,8\}$
parameter	σ	10.34	10.41	10.41	10.37
$\mathcal{E}^H$*-clustering*	ρ_1	0.937 ± 0.006	0.873 ± 0.025	0.731 ± 0.016	0.687 ± 0.016
	$\rho_{1/2}$	0.939 ± 0.006	0.874 ± 0.027	0.722 ± 0.017	0.647 ± 0.017
	$\widetilde{\rho}_\sigma$	0.939 ± 0.006	0.847 ± 0.031	0.695 ± 0.023	0.657 ± 0.014
	$\widehat{\rho}_\sigma$	0.933 ± 0.005	$\mathbf{0.891 \pm 0.009}$	$\mathbf{0.759 \pm 0.011}$	$\mathbf{0.704 \pm 0.011}$
spectral-clustering	$\widehat{\rho}_\sigma$	0.823 ± 0.015	0.769 ± 0.012	0.678 ± 0.014	0.649 ± 0.018
k-means	✗	0.927 ± 0.004	0.878 ± 0.010	0.744 ± 0.008	0.695 ± 0.012
GMM	✗	$\mathbf{0.952 \pm 0.005}$	0.839 ± 0.015	0.694 ± 0.010	0.621 ± 0.009

6 Discussion

We considered clustering from the perspective of generalized energy statistics, valid for arbitrary spaces of negative type. Our mathematical formulation of energy clustering reduces to a QCQP in the associated RKHS, as demonstrated in Proposition 2. We showed that the optimization problem is equivalent to kernel k-means, once the kernel is fixed; see Proposition 3. Energy statistics, however, fixes a family of standard kernels in Euclidean space, and more general kernels on spaces of negative type can also be obtained. We proposed the iterative $\mathcal{E}^H$-clustering algorithm based on Hartigan's method, which was compared to kernel k-means algorithm based on Lloyd's heuristic. Both have the same time complexity, however, numerical and theoretical results provide compelling evidence that $\mathcal{E}^H$-clustering is more robust with a superior performance, specially in high dimensions. Furthermore, energy clustering, with standard kernels from energy statistics, outperformed k-means and GMM on several settings, even on Gaussian ones, illustrating the flexibility of the proposed method. In some examples the iterative $\mathcal{E}^H$-clustering also surpassed spectral clustering, and in others performed similarly but never worse.

A limitation of $\mathcal{E}^H$-clustering is that it cannot handle accurately highly unbalanced clusters. An interesting problem would be to extend $\mathcal{E}^H$-clustering to these cases. Moreover, it would also be interesting to formally demonstrate cases where energy clustering is a consistent. A soft version of energy clustering is also an interesting extension. Finally, 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 algorithm more scalable.

Acknowledgements

We would like to thank Carey Priebe for discussions. This work was supported by NIH TRA grant.

References

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

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

C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups: Theory of Positive*

Definite and Related Functions. Graduate Text in Mathematics 100. Springer, New York, 1984.

I. S. Dhillon, Y. Guan, and B. Kulis. 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, pages 551–556, New York, NY, USA, 2004. ACM.

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

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

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

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

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

J. A. Hartigan and M. A. Wong. 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, pages 281–297. University of California Press, 1967.

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

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

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

J. Mercer. 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.

A. Y. Ng, M. I. Jordan, and Y. Weiss. On Spectral Clustering: Analysis and an Algorithm. In *Advances in Neural Information Processing Systems*, volume 14, pages 849–856, Cambridge, MA, 2001. MIT Press.

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

Q. Qui and G. Sapiro. Learning Transformations for Clustering and Classification. *Journal of Machine Learning Research*, 16:187–225, 2015.

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

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

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

N. Slonim, E. Aharoni, and K. Crammer. 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*, pages 1677–1684. AAI Press, 2013.

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

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

M. Telgarsky and A. Vattani. Hartigan's Method: k-Means Clustering without Voronoi. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9, pages 313–319. JMLR, 2010.

A Supplementary Material

Here we colect the proofs of our main results.

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

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

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

$$\begin{aligned} W &= \frac{1}{2} \sum_{j=1}^{k} \frac{1}{n_j} \sum_{x,y \in \mathcal{C}_j} \rho(x, y) \\ &= \sum_{j=1}^{k} \sum_{x \in \mathcal{C}_j} \left(K(x, x) - \frac{1}{n_j} \sum_{y \in \mathcal{C}_j} K(x, y) \right). \end{aligned} \tag{19}$$

Note that the first term is global so it does not contribute to the optimization problem. Therefore, minimizing (19) is equivalent to

$$\max_{\mathcal{C}_1, \ldots, \mathcal{C}_k} \sum_{j=1}^{k} \frac{1}{n_j} \sum_{x,y \in C_j} K(x, y). \tag{20}$$

But

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

where we used the definitions (11) and (12). Notice that $n_j^{-1} = D_{jj}^{-1}$, where the diagonal matrix $D = \operatorname{diag}(n_1, \ldots, n_k)$ contains the number of points in each cluster, thus the objective function in (20) is equal to $\sum_{j=1}^{k} D_{jj}^{-1} \left(Z^\top G Z \right)_{jj} = \operatorname{Tr}\left(D^{-1} Z^\top G Z \right)$. Now we can use the cyclic property of the trace, and by the definition of the matrix Z in (12), we obtain the following integer programing problem:

$$\begin{aligned} &\max_{Z} \operatorname{Tr}\left(\left(Z D^{-1/2} \right)^\top G \left(Z D^{-1/2} \right) \right) \\ &\text{s.t. } Z_{ij} \in \{0, 1\},\ \textstyle\sum_{j=1}^{k} Z_{ij} = 1,\ \sum_{i=1}^{n} Z_{ij} = n_j. \end{aligned} \tag{21}$$

Now we write this in terms of the matrix $Y = ZD^{-1/2}$. The objective function immediately becomes $\operatorname{Tr}\left(Y^\top G Y \right)$. 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 (13), which comes from the last constraint in (21). 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 (21) is equivalent to (13) . □

Proof of Proposition 3. Notice that

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

therefore, kernel k-means objective function becomes

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

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 (14) is equivalent to

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

which is exactly the same as (20) from the energy statistics formulation. Therefore, once the kernel K is fixed, the function W given by (4) is the same as J in (14). The remaining of the proof proceeds as already shown in the proof of Proposition 2, leading to the optimization problem (13). □