# Nonparametric Clustering Based on Energy Distance

**Guilherme França**
Johns Hopkins University
guifranca@gmail.com

**Joshua Vogelstein**
Johns Hopkins University
jovo@jhu.edu

## Abstract

blabla

## 1 Introduction

Mention why energy is important, main results, where it was applied, etc. Motivate how this can be used for clustering. Mention most important papers on this . . . Explain main results of this paper and give a brief outline.

## 2 Energy Statistics and RKHS

In this section we briefly review the main concepts from energy statistics and reproducing kernel Hilbert spaces (RKHS) used through paper. For more details we refer the reader to [1] (and references therein) and also [2].

Consider random variables in $\mathbb{R}^D$ such that $X, X' \overset{iid}{\sim} P$ and $Y, Y' \overset{iid}{\sim} Q$, where $P$ and $Q$ are cumulative distribution functions with finite first moments. The quantity [1]

$$\mathcal{E}(P,Q) = 2\mathbb{E}\|X - Y\| - \|X - X'\| - \|Y - Y'\|, \tag{1}$$

called energy distance, is rotational invariant and nonnegative, $\mathcal{E}(P,Q) \geq 0$, where equality to zero holds if and only if $P = Q$. Above, $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^D$. Energy distance provides a characterization of equality of distributions, and $\mathcal{E}^{1/2}$ is a metric on the space of distributions.

The energy distance can be generalized as

$$\mathcal{E}_\alpha(P,Q) = 2\mathbb{E}\|X - Y\|^\alpha - \|X - X'\|^\alpha - \|Y - Y'\|^\alpha \tag{2}$$

where $0 < \alpha \leq 2$. This quantity is also nonnegative, $\mathcal{E}_\alpha(P,Q) \geq 0$. Furthermore, for $0 < \alpha < 2$ we have that $\mathcal{E}_\alpha(P,Q) = 0$ if and only if $P = Q$, while for $\alpha = 2$ we have $\mathcal{E}_2(P,Q) = 2\|\mathbb{E}(X) - \mathbb{E}(Y)\|^2$, showing that equality to zero only requires equality of the means and thus it does not imply equality of distributions. It is important to mention that (2) can be further generalized by replacing $\|X - Y\| \to \rho(X,Y)$, where $\rho$ is a so-called semimetric of negative type [2]. In this case there is a Hilbert space $\mathcal{H}$ and a map $\varphi : \mathbb{R}^D \to \mathcal{H}$ such that $\rho(X,Y) = \|\varphi(X) - \varphi(Y)\|^2_{\mathcal{H}}$. Even though the semimetric $\rho$ may not satisfy the triangle inequality, its square root $\rho^{1/2}$ does, since it is an actual metric.

There is an equivalence between energy distance, commonly used in statistics, and distances between embeddings of distributions in reproducing kernel Hilbert spaces (RKHS), commonly used in machine learning. This equivalence was established in [2]. We recall the definition of RKHS. Let $\mathcal{H}$ be a Hilbert space over real-valued functions over $\mathbb{R}^D$. A function $k : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ is a reproducing kernel of $\mathcal{H}$ if it satisfies the following two conditions:

1. $K_x \equiv k(\cdot, x) \in \mathcal{H}$ for any $x \in \mathbb{R}^D$.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2. $\langle K_x, f\rangle_{\mathcal{H}} = f(x)$ for any $x \in \mathbb{R}^D$ and $f \in \mathcal{H}$.

In other words, for any $x \in \mathbb{R}^D$ there is a unique function $K_x \in \mathcal{H}$ that reproduces $f(x)$ through the inner product of $\mathcal{H}$. If such a kernel function $k$ exists, then $\mathcal{H}$ is called a RKHS. From this we have $\langle K_x, K_y\rangle = K_y(x) = k(x,y)$. This implies that $k(x,y)$ is symmetric and positive definite, $\sum_{i,j=1}^{n} c_i c_j k(x_i,x_j) \geq 0$ for $c_i, c_j \in \mathbb{R}$.

The Moore-Aronszajn theorem establishes the converse. For every symmetric and positive definite function $k : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$, there is an associated RKHS, $\mathcal{H}_k$, with reproducing kernel $k$. The map $\varphi : x \mapsto K_x \in \mathcal{H}_k$ is called the canonical feature map. Given a kernel $k$, this theorem enables us to define an embedding of a probability measure $P$ into the RKHS: $P \mapsto K_P \in \mathcal{H}_k$ such that $\int f(x)dP(x) = \langle f, K_P\rangle$ for all $f \in \mathcal{H}_k$, or alternatively $K_P = \int k(\cdot, x)dP(x)$. We can now introduce the notion of distance between two probability measures using the inner product of $\mathcal{H}_k$. This is called the maximum mean discrepancy (MMD) and it is given by

$$\gamma_k(P,Q) = \|K_P - K_Q\|_{\mathcal{H}_k}, \tag{3}$$

which can also be written as [3]

$$\gamma_k^2(P,Q) = \mathbb{E}k(X,X') + \mathbb{E}k(Y,Y') - 2\mathbb{E}k(X,Y) \tag{4}$$

where $X, X' \overset{iid}{\sim} P$ and $Y, Y' \overset{iid}{\sim} Q$. From the equality between (3) and (4) we also have that

$$\langle K_P, K_Q\rangle_{\mathcal{H}_k} = \mathbb{E}\, k(X,Y). \tag{5}$$

Thus we can compute the inner product between the 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 [4]. Let $\rho : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ be a semimetric, and $x_0 \in \mathbb{R}^D$ an arbitrary but fixed point. Define

$$k(x,y) = \tfrac{1}{2}\left\{\rho(x,x_0) + \rho(y,x_0) - \rho(x,y)\right\}. \tag{6}$$

Then $k$ is positive definite if and only if $\rho$ is of negative type. Thus we have a family of kernels, one for each choice of $x_0$. Conversely, if $\rho$ is a semimetric of negative type and $k$ is a kernel in this family, then

$$\rho(x,y) = k(x,x) + k(y,y) - 2k(x,y) = \|K_x - K_y\|^2_{\mathcal{H}_k}, \tag{7}$$

and the canonical feature map $\varphi : x \mapsto K_x$ is injective [2]. We say that the kernel $k$ generates the semimetric $\rho$. If two different kernels generate the same $\rho$, they are equivalent kernels.

Now we can state the equivalence between energy distance $\mathcal{E}$ and inner products on RKHS, which is one of the main results of [2]. If $\rho$ is a semimetric of negative type and $k$ a kernel that generates $\rho$, then

$$\begin{aligned}
\mathcal{E}(P,Q) &\equiv 2\mathbb{E}\rho(X,Y) - \mathbb{E}\rho(X,X') - \mathbb{E}\rho(Y,Y') && (8)\\
&= 2\left[\mathbb{E}\, k(X,X') + \mathbb{E}\, k(Y,Y') - 2\mathbb{E}\, k(X,Y)\right] && (9)\\
&= 2\gamma_k^2(P,Q) && (10)
\end{aligned}$$

This result follows simply by substituting (7) into (8), and using (4). Since $\gamma_k^2(P,Q) = \|K_P - K_Q\|^2_{\mathcal{H}_k}$, we can compute the energy distance using the inner product in the RKHS. Moreover, this can be computed from the data according to (5).

## 3 Energy Distance based Clustering

Now we formulate an optimization problem for clustering based on energy statistics and RKHS. Assume we have data $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$, and a partition $\mathcal{X} = \cup_{k=1}^{K} \mathcal{A}_k$, where $\mathcal{A}_k \cap \mathcal{A}_{k'} = \emptyset$. Each expectation in (2) can be computed through the function

$$g_\alpha(\mathcal{A}_k, \mathcal{A}_{k'}) \equiv \frac{1}{N_k N_{k'}} \sum_{x\in\mathcal{A}_k} \sum_{y\in\mathcal{A}_{k'}} \|x-y\|^\alpha \tag{11}$$

where $N_k = |\mathcal{A}_k|$ is the number of elements in partition $\mathcal{A}_k$. In energy statistics [1] we have the within energy dispersion

$$\mathcal{W}_\alpha = \sum_{k=1}^{K} \frac{N_k}{2} g_\alpha(\mathcal{A}_k, \mathcal{A}_k), \tag{12}$$

and also the between-sample energy statistic

$$\mathcal{S}_\alpha = \sum_{1 \le k < k' \le K} \frac{N_k N_{k'}}{2N} \left[ 2 g_\alpha(\mathcal{A}_k, \mathcal{A}_{k'}) - g_\alpha(\mathcal{A}_k, \mathcal{A}_k) - g_\alpha(\mathcal{A}_{k'}, \mathcal{A}_{k'}) \right]. \tag{13}$$

This is a test statistic for equality of distributions, which is small if all datapoints comes from the same distribution, and diverges otherwise, in the limit $N \to \infty$. Therefore, our criteria for clustering data is to maximize $\mathcal{S}_\alpha$. It can be shown that the total dispersion of the data obeys [1]

$$\mathcal{T}_\alpha(\mathcal{X}) = \mathcal{W}_\alpha + \mathcal{S}_\alpha = \frac{N}{2} g_\alpha(\mathcal{X}, \mathcal{X}). \tag{14}$$

Note that $\mathcal{T}$ only depends on the pooled data, so it does not depend on how we partition $\mathcal{X}$. Therefore, maximizing $\mathcal{S}_\alpha$ is equivalent to minimizing $\mathcal{W}_\alpha$, which has a simpler form. Thus, our clustering problem corresponds to find the best partition of $\mathcal{X}$ such that

$$\min_{\{\mathcal{A}_k\}} \mathcal{W}_\alpha(\{\mathcal{A}_1, \dots, \mathcal{A}_K\}) \tag{15}$$

where each datapoint belongs to one and only one partition (hard assignments).

Based on the equivalence between semimetrics of negative type and kernel functions (7), and choosing $x_0 = 0$ for simplicity (one might choose any other point in $\mathbb{R}^D$), in the case of (11) we have the associated kernel

$$k_\alpha(x, y) = \tfrac{1}{2} \left( \|x\|^\alpha + \|y\|^\alpha - \|x - y\|^\alpha \right). \tag{16}$$

Now we can write

$$\mathcal{W}_\alpha(\{\mathcal{A}_1, \dots, \mathcal{A}_K\}) = -\sum_{k=1}^{K} \frac{N_k}{2} \, \mathbb{E} \, k_\alpha(\mathcal{A}_k, \mathcal{A}_k) = -\sum_{k=1}^{K} \frac{1}{2N_k} \sum_{x, x' \in \mathcal{A}_k} k_\alpha(x, x'). \tag{17}$$

Now let $Z \in \{0,1\}^{N \times K}$ be a binary matrix such that

$$Z_{nk} = \begin{cases} 1 & \text{if } x_n \in \mathcal{A}_k \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

Notice that $D = Z^T Z = \operatorname{diag}(N_1, N_2, \dots, N_K)$ contains the number of elements in each partition. Introducing the kernel matrix $K^\alpha \in \mathbb{R}^{N \times N}$ such that

$$K^\alpha_{ij} = k_\alpha(x_i, x_j), \tag{19}$$

we can write (17) as $-\frac{1}{2} \operatorname{Tr} D^{-1} Z^\top K^\alpha Z$. Thus our optimization problem (15) can be written as

$$\begin{aligned} &\max_Z \operatorname{Tr} \left\{ \left( Z D^{-1/2} \right)^\top K^\alpha \left( Z D^{-1/2} \right) \right\} \\ &\text{s.t. } Z_{ij} \in \{0,1\}, \textstyle\sum_{k=1}^{K} Z_{nk} = 1, \sum_{n=1}^{N} Z_{nk} = N_k, \text{ and } D = Z^\top Z. \end{aligned} \tag{20}$$

This is a quadratic problem with integer constraints, which is usually NP-hard. This problem has the same formulation as Kernel $K$-means, but with our non-parametric kernel function given by (16).

## 4 Numerical Experiments

## 5 Conclusion

### Acknowledgements

We thank ...

## References

[1] 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.

[2] 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.

[3] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.

[4] 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.