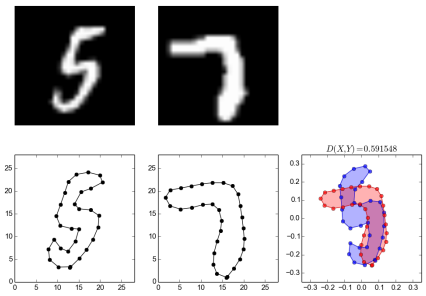
0.1 Non-Parametric Shape Clustering

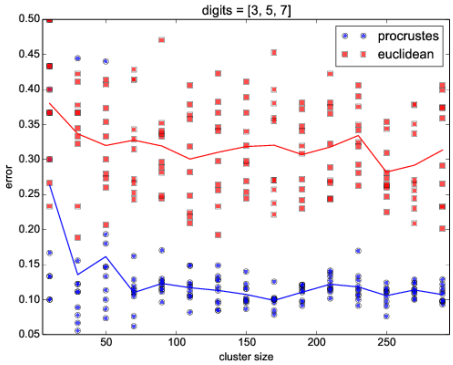
We developed a prototype algorithm for 2-dimensional shape clustering, which is invariant under affine transformations. This employs the procrustes distance between two objects, which requires feature extraction to obtain landmark points; see Fig. **??** for an example. We intend to apply a variant of this method to analyze neural synapses, since we have such datasets from our collaborators. However, to this particular dataset, the preprocessing techniques may be highly sophisticated, and a new metric to compare different objects may be necessary. Moreover, these shapes are 3-dimensional. We are currently working on this project with our collaborators, extending our existing techniques to this particular dataset.
We also started working on a related, and more general, project. We intend to develop non-parametric clustering algorithms with statistical guarantees. We will use an energy-statistics based approach. Given two datasets X and Y, there is an energy function $\mathcal{E}(X, Y)$ test statistic which allows us to infer if X and Y have the same distribution. Our results thus far suggest that this can be written as a quadratic optimization problem with quadratic constraints:

$$\max_{x,z \in \mathbb{R}^N} x^T \Delta z \qquad \text{s.t. } x_i^2 = 1, x + z = 0 \tag{1}$$

where Δ is a dissimilarity data matrix. There is not enough literature on this interesting problem, so this will very likely lead to new methods which can have interesting applications, in particular to neuroscience datasets.



(a) MNIST digits with extracted landmarks and alignment.



(b) Classification error against the size of each cluster (the three classes have the same number of points) is shown in blue. The red line is standard K-means with Euclidean distance for comparison.

Figure 1: MNIST handwritten digits and classification error results.

We have been mostly focused on developing non-parametric clustering methods. To this purpose, we are exploring ideas from energy statistics, which is non-parametric, robust, and rotational invariant, thus it incorporates the main ingredients that we are looking for. The main difficulty is to formulate an algorithm based on this, i.e. to identify the correct test statistic, or to formulate it as a feasible optimization problem. Consider K-Means clustering problem which is $\min_{\{\mathcal{C}_k\}} \sum_{k=1}^{K} \sum_{x \in \mathcal{C}_k} \|x - \mu_k\|^2$, where $\mathcal{C}_k$ is the kth cluster and μ_k the mean of its points. We showed that this problem is equivalent to

$$\max_{G} \operatorname{Tr}\left(G^T K G\right) \qquad \text{s.t.} \quad G \geq 0, G^T G = I, G G^T e_1 = e_1. \tag{2}$$

where $e_1 = (1, 1, \ldots, 1)^T$. This is a Quadratically Constrained Quadratic Problem (QCQP), which is usually NP-hard. Analogously, consider the energy function $\mathcal{E}(F, G) = 2\mathbb{E}\|X - Y\| - \mathbb{E}\|X - X'\| - \mathbb{E}\|Y - Y'\|$ between $X, X' \sim F$ and $Y, Y' \sim G$. We showed that this can be written as $\mathcal{E}(A, B) = e_1^T \Delta e_1$, where Δ is a dissimilarity matrix between the two sets of data points $A \overset{iid}{\sim} F$ and $B \overset{iid}{\sim} G$. Consequently, a simple two-class clustering problem would be

$$\max_{x, z \in \mathbb{R}^N} x^T \Delta z \qquad \text{s.t. } x_i^2 = 1, x + z = 0, \tag{3}$$

which is also a QCQP problem. We are currently investigating this problem and trying to generalize it correctly for more classes. A simple check of the energy function as a test statistic is shown in Fig. **??**. Under the null $F = G$, T converges to a quadratic form of normally distributed random variables. This seems to be the case in the first (blue) histogram, while it is definitely not the case in the other (red and green) histograms. For the blue histogram a single test gives $T \approx 0.32$ (small), for the red histogram $T \approx 4000$ (large), and for the green histogram $T \approx 105$ (large), with only a few points. Thus, energy statistics based approach is able to distinguish between different distributions, even when the clusters have the same mean, which is a property that K-Means cannot resolve.

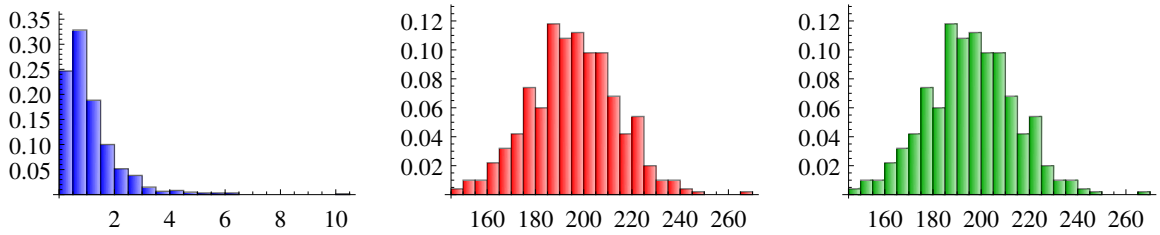


Figure 2: Distribution of test statistic $T \equiv \frac{nm}{n+m}\mathcal{E}(A, B)$ for an ensamble obtained from two distributions: $A \overset{iid}{\sim} \mathcal{N}(\mu_A, \sigma_A^2)$ and $B \overset{iid}{\sim} \mathcal{N}(\mu_B, \sigma_B^2)$, where $|A| = n$ and $|B| = m$. Blue histogram: $\mu_A = \mu_B = 0$ and $\sigma_A = \sigma_B = 1$; Red histogram: $\mu_A = -\mu_B = 1$ and $\sigma_A = \sigma_B = 1$; Green histogram: $\mu_A = \mu_B = 0$, $\sigma_A = 1$ and $\sigma_B = 1.5$.

Energy statistics provides a nonparametric test for equality of distributions. It is rotational invariant which is a highly desirable quality for clustering. For a two-class problem, $X, X' \sim \mu$ and $Y, Y' \sim \nu$, where μ, ν are CDFs, it reads

$$\mathcal{E}(X,Y) = 2\mathbb{E}\|X-Y\| - \mathbb{E}\|X-X'\| - \mathbb{E}\|Y-Y'\|. \tag{4}$$

We are developing a clustering framework based on **??**. Our criteria is that $\mathcal{E}$ should be a maximum when data points are correctly classified. It is possible to show that there is a map from the data space of X, Y to the probability space of μ, ν which is a Hilbert space whose inner product can be obtained from a kernel function related to **??**, $\langle \mu, \nu \rangle = k(x,y)$. This enables us to formulate our clustering problem as follows:

$$\max \left\{ \mathrm{Tr} L^{1/2} Z^T K Z L^{1/2} \right\} \quad \text{s.t.} \quad Z_{ij} \in \{0,1\}, \sum_i Z_{ij} = N_j, \sum_j Z_{ij} = 1, Z^T Z = L^{-1} \tag{5}$$

where N_j is the number of elements in the jth cluster, $L^{-1} = \mathrm{diag}(N_1, \ldots, N_k)$, and K is the Gram matrix obtained from the kernel. Let $\mathcal{X}$ be the pooled data matrix. If we replace $K \to \mathcal{X}^T \mathcal{X}$ we recover the well-known k-means problem, which in this formulation is related to spectral clustering and normalized cuts. Problem **??** is NP-hard and a numerical implementation is prohibitive even for small data sets. We are investigating how to solve **??** in a feasible way. As an evidence that **??** is the correct optimization problem, and more importantly, it illustrates the power behind our proposal, in Fig. **??** we generate data and plot the objective in **??** versus n, where n is the number of points randomly shuffled from one class to the other. Therefore, for $n = 0$ the function must be a maximum. We do this for the kernel related to **??** (blue dots) and compare with the kernel related to k-means (red dots). Clearly, **??** based on **??** is able to distinguish between different cluster even for complex data sets that are not linearly separable. Moreover, in our formulation there are no free-parameters in the kernel.

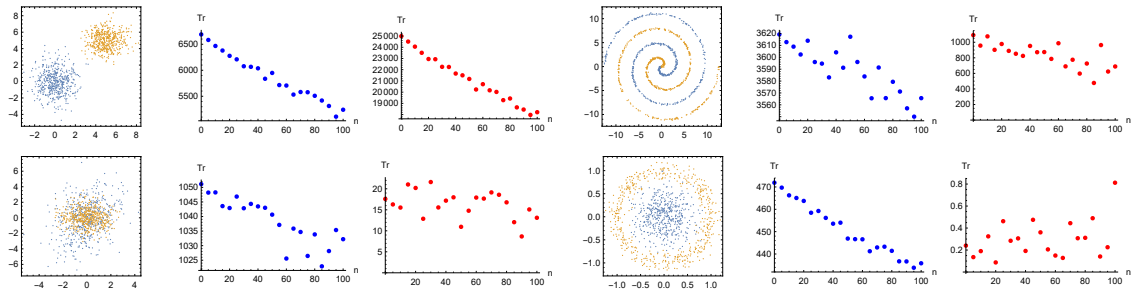


Figure 3: Two dimensional datasets and the objective function in **??** as a function of n, where n is the number of shuffled points from its correct class to the wrong class. Blue dots are for **??** and red dots for k-means. A good function must be monotonically decreasing. We can clearly see that **??** is way more powerful than k-means.