Proposal - Energy Stats vs. Clustering.

k-means objective function:

$$E = \frac{1}{2} \sum_{k=1}^{K} \sum_{x_i \in C_k} \| x_i - m_k \|^2$$

$$= \frac{1}{2} \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ki} \| x_i - m_k \|^2$$

$$z_{ki} = \begin{cases} 1 & \text{if } x_i \in C_k \\ 0 & \text{otherwise} \end{cases}$$

Notice that $\sum_{k=1}^{K} z_{ki} = 1$, $\sum_{i=1}^{n} z_{ki} = n_k$. Introducing

$$X = \left(x_1 \,|\, x_2 \,|\, \cdots \,|\, x_n \right)_{D \times N} \quad M = \left(m_1 \,|\, m_2 \,|\, \cdots \,|\, m_K \right)_{D \times K}$$

$Z = z_{ij} \in \mathbb{R}^{K \times N}$.

Then we have

$$E = \| X - MZ \|_F^2$$

we can minimize this over M yielding:

$$\min_{Z} \| X - X Z^T (Z Z^T)^{-1} Z \|_F^2$$
$$\text{s.t. } z_{ij} \in \{0, 1\}, \ \sum_k z_{ki} = 1$$

Define $\tilde{z} = z^T(zz^T)^{-1}z$ then

$$\|X - X\tilde{z}\|_F^2 = Tr\left((X - X\tilde{z})^T(X - X\tilde{z})\right)$$
$$= Tr\left(X^TX - X^TX\tilde{z} - \tilde{z}^TX^TX + \tilde{z}^TX^TX\tilde{z}\right)$$
$$= Tr\left(X^TX - 2X^TX\tilde{z} + \tilde{z}^TX^TX\tilde{z}\right)$$

Gram matrix: $K \equiv X^TX$.

$$Tr\left(K - 2K\tilde{z} + \tilde{z}^TK\tilde{z}\right)$$

$\|X - MZ\|^2 = Tr\left((I-M)^TK(I-M)\right)$

So k-means can be written as an Integer Quadratic Optimization Problem:

$$\min Tr\left(K - 2K\tilde{z} + \tilde{z}^TK\tilde{z}\right)$$
$$\text{s.t. } z_{ij} \in \{0,1\},\ \sum_n z_{ni} = 1,\ \tilde{z} = z^T(zz^T)^{-1}z.$$

Another formulation is ($K = X^TX$)

$$\max Tr(G^TKG)$$
$$\text{s.t. } G \geq 0,\ G^TG = I,\ GG^T\vec{1} = \vec{1}$$

where $G_{ik} = \begin{cases} \frac{1}{\sqrt{n_k}} & \text{if } x_i \in C_k \\ 0 & \text{otherwise} \end{cases}$

We can still manipulate and write this in different forms. In any case, the problem should be expressed in terms of K, the Gram matrix.

We can also write $Y = Z^T(ZZ^T)^{-1}$. Then $M = XY$. Y is a doubly stochastic matrix, thus $Y \geqslant 0$.

Thus

$$\begin{cases} \min_{Y,Z} \|X - XYZ\|_F^2 \\ \text{s.t.} \quad Y \geqslant 0 \quad \vec{1}^T y_i = 1 \quad H(y_i) \gg 0 \\ \qquad Z \geqslant 0 \quad \vec{1}^T z_i = 1 \quad H(z_i) = 0 \end{cases}$$

$\rightarrow \mathrm{Tr}(I - YZ)^T K (I - YZ)$ ($\rightarrow X^T X$ gram matrix

$\leftarrow$ entropy constraints

In archetypal analysis we have

$$\begin{cases} \min_{Y,Z} \|X - XYZ\|_F^2 \\ \text{s.t.} \quad Y \geqslant 0 \quad \vec{1}^T y_i = 1 \\ \qquad Z \geqslant 0 \quad \vec{1}^T z_i = 1 \end{cases}$$

Non-negative matrix factorization:

$$\begin{cases} \min_{Y,Z} \|X - XYZ\|_F^2 \\ \text{s.t.} \quad Y \geqslant 0 \\ \qquad Z \geqslant 0 \end{cases}$$

There are few options to solve these optimization problems.

1. Frank-Wolfe Algorithm.

(i) Initialize Y, Z
(ii) Fix Z, solve for Y
(iii) Fix Y, solve for Z
(iv) Repeat until convergence

This will require the gradient of the objective function, which will depend on the Gram matrix K:

$$\nabla(\dots) \sim KY(ZZ^T) - KZ^T - \dots$$

2. SDP Relaxation

Standard Prime Dual SDP

$$\begin{aligned} &\min QX \\ &\text{s.t} \quad X_{ii} = 1 \\ &\qquad X \succcurlyeq 0 \end{aligned} \quad \longrightarrow \quad \begin{aligned} &\max \Lambda \\ &\text{s.t.} \quad Q \succcurlyeq \Lambda \\ &\qquad \Lambda \text{ diagonal} \end{aligned}$$

This involves an augmented Lagrangian. It is also possible to impose a <u>rank constraint</u> in the problem, and end up with an SDP with rank constraint. One can relax the constraint, solve it, then enforce the rank on a "rounding scheme" in the final solution.

3. Low Rank SDP.

We can employ the same technique used in that paper which provides the low-rank approx already in the optimization prob. itself. They did that for weighted k-means.

4. Alternating Direction Methods.

We can check how to solve these kind of optimization problems using alternating methods with augmented Lagrangians. A very nice option would be ADMM. This would be appealing and very novel. ADMM does not need gradients, so it would be very robust also.

How Energy Statistics Enters the Story?

In the above proposed formulation, all we need to do is exploit the kernel trick on k!

- From Sejdinovic et al, Ann. Stats (2013): Let $\rho: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a semi-metric on $\mathcal{X}$. Then $k(x, x') = \frac{1}{2}(\rho(x, x_0) + \rho(x', x_0) - \rho(x, x'))$ is positive definite <u>iff</u> ρ is of <u>negative type</u>.

Now the energy distance between two CDF's,

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

where $X, X' \overset{iid}{\sim} F$, $Y, Y' \overset{iid}{\sim} G$ is of negative type.

We can embed F, G into a RKHS with kernel function given by the energy distance above.

They proved an equivalence between energy distance and MMD in RKHS. Thus we can replace the gram matrix K in the optimization problem to a kernel defined in terms of the energy distance!

If $x_i \sim \mu_i$, $x_j \sim \mu_j$

$$\|x_i - x_j\| = \mathcal{D}(\mu_i, \mu_j)$$

↖ compute this from data "empirical measure".

The replace $K_{ij} \to k(x_i, x_j)$.

- Understand this better, on how to incorporate energy distance as a kernel.

Sketch of the problem: