Amalthea Technical report 1

Yixin Lin

May 18, 2016

Von Mises distribution

Intuitively, the Von Mises distribution[4] is a simple approximation for the normal distribution on a circle (known as the *wrapped normal distribution*). The Von Mises probability density function is defined by the following equation:

$$P(x) = \frac{e^{b\cos(x-a)}}{2\pi I_0(b)}$$

where $I_0(x)$ is the modified Bessel function of the first kind; the Von Mises cumulative density function has no closed form.

The mean $\mu = a$ (intuitively, the angle that the distribution clustered around), and the circular variance $\sigma^2 = 1 - \frac{I_1(b)}{I_0(b)}$ (intuitively, b is the "concentration" parameter). Therefore, as $b \to 0$, the distribution becomes uniform; as $b \to \infty$, the distribution becomes normal with $\sigma^2 = 1/b$.

Von Mises-Fisher distribution

THe Von-Mises Fisher distribution is the generalization of the Von Mises distribution to n-dimensional hyperspheres. It reduces to the Von-Mises distribution with $n = 2$. The probability density function is defined by the following equation:

$$f_p(\boldsymbol{x};\boldsymbol{\mu},\kappa) = C_p(\kappa)\exp(\kappa\boldsymbol{\mu}^T\boldsymbol{x})$$

where

$$C_p(\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^p I_{p/2-1}(\kappa)}$$

and intuitively is an approximation for the normal distribution on the hypersphere.

K-means clustering

k-means clustering is a simple and popular algorithm for the *clustering problem*, the task of grouping a set of observations so that a group is "similar" within itself and "dissimilar" to other groups. k-means partitions n observations into k clusters, with each observation belonging to the nearest mean of the cluster. This problem is NP-hard in general, but there are heuristics which guarantee convergence to a local optimum.

The standard heuristic (known as *Lloyd's algorithm*) is the following:

Algorithm 1 Lloyd's algorithm for k-means clustering

1: generate an initial set of k means
2: **while** not converged **do**
3: assign all data points to nearest Euclidean-distance mean
4: calculate new means to as the centroids of the observations in the cluster
5: **end while**

There is a choice of initialization method. The *Forgy method* randomly picks k observations as initial means, while the *Random Partition method* randomly picks a cluster for each observation.

The Lloyd's algorithm is a heuristic, so it does not guarantee a global optimum. Furthermore, there exists sets of points in which it converges in exponential time. However, it has been shown to have a smoothed polynomial running time, and in practice converges quickly.

Spherical k-means clustering

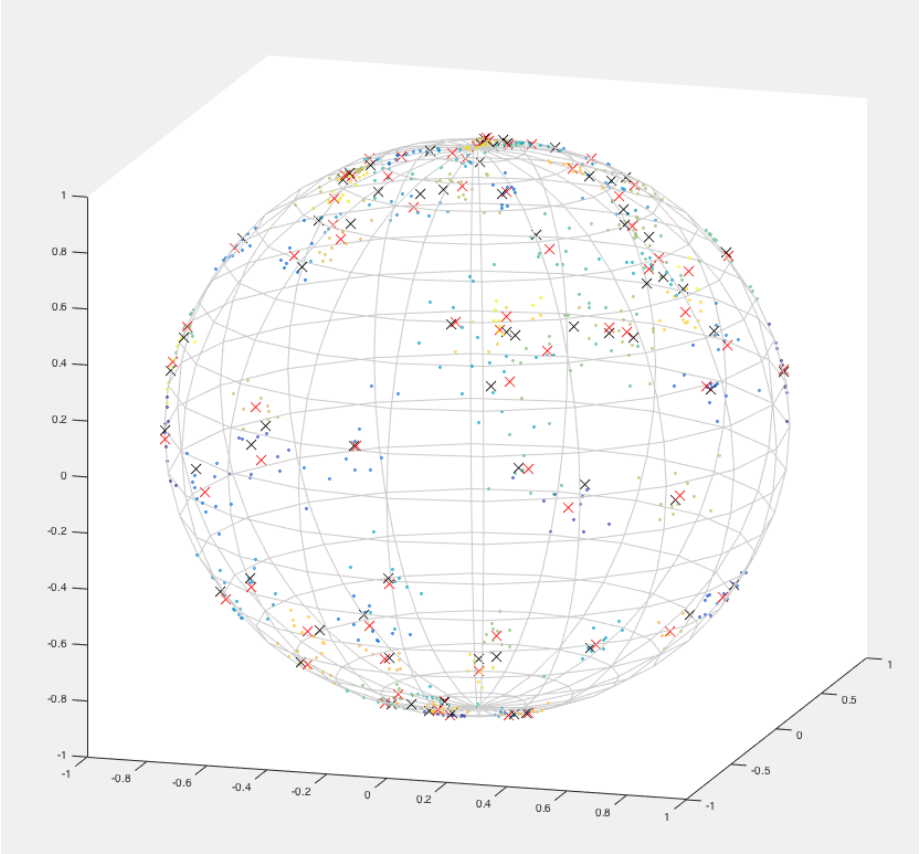
Spherical k-means clustering the same idea, but with points on a sphere. We investigated a MATLAB implementation by Nguyen[3, 2], which required a mean-and-norm-normalized dataset located on a hypersphere. Important aspects of this implementation include:

- When there exists an empty cluster, the largest cluster is split
- Use the dot product as "negative distance", which leverages the fact that observations are unit vectors on the hypersphere
- Use the normalized sum of observations as a centroid/mean, which leverages the fact that observations are unit vectors on the hypersphere. Note that this fails on pathological cases where the sum of observations is zero.

Our investigation

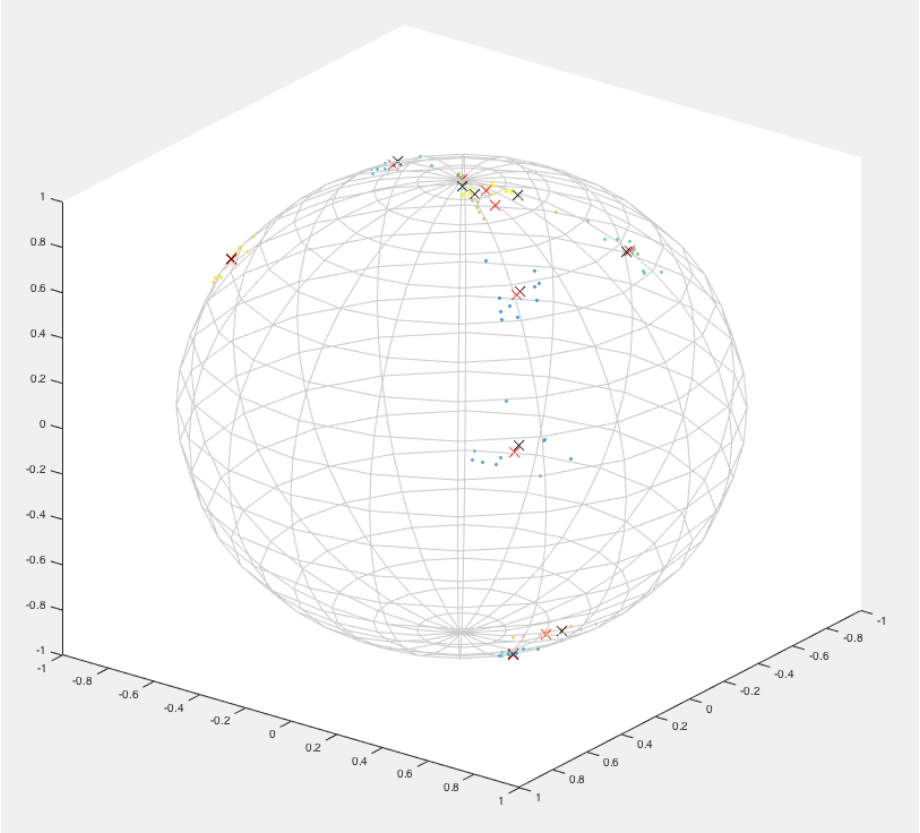
In order to test how well the spherical k-means clustering algorithm worked, we constructed a random data set using the Von Mises distribution random

Figure 1: Plot of 70 clusters and means on a 3-D sphere



sampling function from the MATLAB Circular Statistics Toolbox[1]. We constructed 70 clusters of 10 points each, using random points on the sphere as means and with $\kappa = 100$, and then applied the spherical k-means clustering and visualized the results on the unit three dimensional sphere, color coded for each cluster and with estimated means labeled as red X's and true means labeled as black X's.

Figure 2: Plot of 10 clusters and means on a 3-D sphere



References

[1] Phillip Berens. Matlab central file exchange: Circular statistics toolbox. *http://www.mathworks.com/matlabcentral/fileexchange/10676-circular-statistics-toolbox–directional-statistics-*. Accessed: 2016-05-18.

[2] Vinh Nguyen. Matlab central file exchange: The spherical k-means algorithm. *http://www.mathworks.com/matlabcentral/fileexchange/32987-the-spherical-k-means-algorithm*. Accessed: 2016-05-18.

[3] Vinh Nguyen. Gene clustering on the unit hypersphere with the spherical k-means algorithm: coping with extremely large number of local optima. In *World Congress in Computer Science, Computer Engineering, and Applied Computing (Hamid R. Arabnia and Youngsong Mun 14 July 2008 to 17 July 2008)*, pages 226–233. CSREA Press, 2008.

[4] Eric W. Weisstein. von mises distribution. *http://mathworld.wolfram.com/vonMisesDistribution.html*. Accessed: 2016-05-18.