1 Isotropic PCA

Principal Component Analysis fails to identify the principal axis when the data points has equal variance (Covariance matrix is Identify matrix) in all directions. Isotropic PCA [1] is an extension of the classical PCA, which weighs the data points according to a spherical Gaussian centered at the mean of original set of points and with appropriately chosen variance. Now, the eigen vector corresponding to the maximum eigen value on this weighted data set is returned as the principal direction. This technique has been applied to unravel the individual components of a convex combination of k Gaussian distributions. Assuming that each of these Gaussians are separable by a hyperplane, the algorithm correctly identifies the component distribution from which a sample point \( x \) is chosen.

2 Project Scope

The scope of this project is to explore the use of Isotropic PCA in document clustering, where each document is viewed as a sample from a convex combination of multinomial distributions rather than Gaussian distribution.

- The first step is to enforce the isotropy, data being centered at the origin with covariance being identify matrix, on the data points (i.e. documents). We conjecture that in Multinomial distributions, Mutual Information substitutes the variance in Gaussian distributions. So we need to transform the documents such that Mutual Information is same in all dimensions.

- Now reweight the points and apply PCA to find the principal directions.

- Find the separating hyperplane (normal to the principal direction) and recurse to subdivide the data.

References