**More on *k*-means clustering: Discussion of spherical *k*-means clustering and clustering with mixtures of von Mises-Fisher distributions**

**Spherical *k*-means clustering**

The text mining literature has also produced its own special variants of clustering methods. Hornik et al (2012) discuss **spherical** ***k*-means** clustering and develop efficient algorithms for *k*-means clustering when adopting the cosine distance in place of the commonly-used squared Euclidean distance. We had mentioned earlier that the cosine distance is not affected by the number of words in the document (that is, the length of the document) which makes it a useful distance measure and explains why it performs well for document clustering. While similar in spirit to Ward’s algorithm for the squared Euclidean distance, the spherical *k*-means clustering algorithm is tailor-made for the cosine distance. Hornik et al also generalize their algorithm to fuzzy clustering (also referred to as "soft" clustering) where a document can be assigned to more than one cluster. With fuzzy clustering, documents are assigned non-negative vectors of cluster assignments that sum to one. Calculations can be carried out with their R package **skmeans**.

**Clustering with mixtures of von Mises-Fisher distributions**

Hornik and Grün (2014) follow a mixture approach to clustering that preserves the spherical nature of vectors of word frequencies. Vectors of word frequencies are projected onto the sphere by normalizing the vectors to have length one. As example, consider the case when the corpus consists of just three words in which case the normalized vectors of document frequencies reduce to points on a three-dimensional ball of radius 1. Assume that one set of documents points to a certain area on that sphere (of course with some scatter around it) and that another set of different documents points to quite a different area. Then one would conclude that there are two clusters, and one would characterize the two clusters by their word distributions (that is, with the words that come up most often in each of the two clusters). This is the basic idea.

The von Mises-Fisher distribution, which is parametrized with a mean direction vector and a concentration coefficient, provides a parametric model for the variability of such unit length vectors. Hornik and Grün (2014) model the vectors of word frequencies of the *d* documents as mixtures of a smaller number of such von Mises-Fisher distributions. They use an iterative expectation-maximization (EM) algorithm to estimate the parameters of the component (von Mises-Fisher) distributions and the mixture proportions on each document. The fitted distributions reflect the most important words of the components, and the mixture proportions represent the soft-clustering weights. Another version of their estimation method uses a hard clustering algorithm which puts each document into a single cluster. Calculations can be carried with their R package **movMF**.

**References**

Hornik, K. and Grün, B. (2014): “movMF: An R Package for Fitting Mixtures of von Mises-Fisher Distributions.” Journal of Statistical Software, 58(10):1-31.

Hornik, K, Feinerer, I., Kober, M. and Buchta, C. (2012): “Spherical k-Means Clustering.” Journal of Statistical Software, 50(10):1-22.