mgr Przemysław Spurek
mgr Krzysztof Misztal
Uniwersytet Jagielloński
Instytut Informatyki

k-means optimization of the Karhunen-Loève Transform

Our general problem concerns splitting a given data-set W into clusters with respect to their intrinsic dimensionality. The motivation to create such an algorithm is a desire to extract parts of data which can be easily described by a smaller number of parameters. More precisely, we want to find minimal number of affine spaces $S_1, \ldots, S_n$ (of possibly different dimensions) such that every element of W belongs (with certain maximal error) to one of the spaces $S_1, \ldots, S_n$.

We have constructed a simultaneous generalization of the k-means method and the Karhunen-Loève transform (called also PCA – Principle Component Analysis).

One of the natural applications of our method is in the image compression, since it is a generalization of the Karhunen-Loève transform. In the simplest form our algorithm needs the number of clusters K and the dimension D (for $D = 0$ we obtain the k-means while for $K = 1$ we obtain the PCA).

Our method can be described as follows:

1. choose a random distribution of the clusters,
2. compute barycenter of clusters,
3. follow the Karhunen-Loève method for each cluster (remember the corresponding bases and middle of cluster),
4. appoint new clusters,
5. repeat Steps 2, 3 and 4 until the barycenter of clusters no longer move.