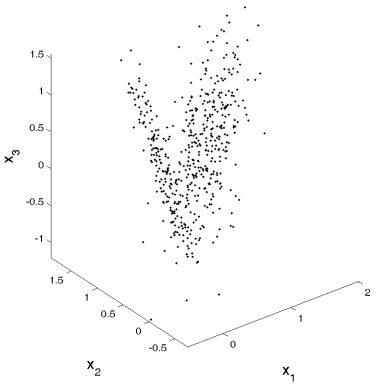
EE613 - Subspace Clustering - Exercises - Oct. 25, 2017

The main folder contains five examples `demo_GMM01.m`, `demo_MFA01.m`, `demo_MPPCA01.m`, `demo_HDDC01.m` and `demo_semitiedGMM01.m`. These codes can be run either from Matlab or from GNU Octave. Each example loads a dataset and fits a GMM, MFA, MPPCA, HDDC or GMM with semi-tied covariances model to the data with a dedicated EM algorithm. First run these codes and try to change the model parameters and visualize the results.

Exercise 1: Stochastic generation of data

- Generate a random dataset containing N datapoints of D dimensions that belong to K clusters. The number of datapoints in each cluster should follow a proportion $\{\pi_1, \pi_2, \ldots, \pi_K\}$. The centers of each cluster are sampled from a uniform distribution $\mathcal{U}(\mathbf{0}, \boldsymbol{I})$. The datapoints in each cluster are normally distributed within a subspace of d dimensions, characterized by a covariance $\boldsymbol{\Sigma}_i = \sum_{j=1}^{d} \boldsymbol{v}_j \boldsymbol{v}_j^\top$ built from d random vectors $\boldsymbol{v}_j$ of length $\|\boldsymbol{v}_j\| = 1/K$. An additional noise $\mathcal{N}(\mathbf{0}, 0.001\, \boldsymbol{I}_D)$ is finally added to all generated datapoints.

  In Matlab, the functions `rand(D,T)` and `randn(D,T)` can be used to generate T random datapoints of D dimensions with uniform distribution $\mathcal{U}(\mathbf{0}, \boldsymbol{I})$ and normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_D)$, respectively.



- Plot the data in a 3D graph for the special case $N = 500$, $D = 3$, $d = 2$, $K = 2$ and $\boldsymbol{\pi} = [0.3, 0.7]$. An example is given in the figure above.

Exercise 2: Fitting an MFA, MPPCA or HDDC model to the generated data

- With the help of the example codes, fit an MFA, MPPCA or HDDC model to the dataset generated in *Exercise 1*, by setting the model parameters $K = 2$ and $d = 2$.
- With the help of the `plotGMM3D` function, visualize the learned MFA parameters $\boldsymbol{\Theta}^{\text{MFA}} = \{\pi_i, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i, \boldsymbol{\Psi}_i\}_{i=1}^{K}$ in the 3D graph. You can do the same for MPPCA and HDDC.

Exercise 3: Analysis of the estimated parameters

- With the dataset generated in *Exercise 1* and the models learned in *Exercise 2*, analyse the effect of initialization by running EM from different initial estimates (initialization with k-means clustering and random initialization).
- With the help of the `gaussPDF` function, analyse the effect of N, D, K and d on the likelihood, and plot some of the results of this analysis in 2D graphs.

Exercise 4: Subspace clustering *Vs* global dimensionality reduction and clustering

- With the dataset generated in *Exercise 1* and the models learned in *Exercise 2*, show that MPPCA does not provide the same result compared to the approach of first reducing the dimension of the original data with PCA, and then clustering the projected data with GMM.