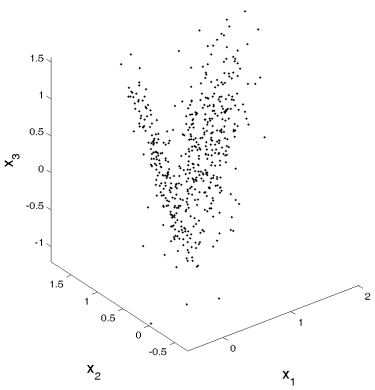
EE613 - Subspace Clustering - Exercises - Oct. 21, 2015

The main folder contains four examples `demo_GMM01.m`, `demo_MFA01.m`, `demo_MPPCA01.m` and `demo_HDDC01.m`. These codes can be run either from Matlab or from GNU Octave. Each example loads a dataset and fits a GMM, MFA, MPPCA or HDDC 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 sparse data

- Generate a random dataset containing N datapoints of D dimension that lie within K clusters. The number of datapoints in each cluster should follow a given proportion $\{\pi_1, \pi_2, \ldots, \pi_K\}$. The center of each cluster should be randomly selected by following a uniform distribution $\mathcal{U}(\mathbf{0}, \boldsymbol{I})$ (i.e., within a hypercube of unit size). The datapoints in each cluster should be normally distributed within a subspace of d dimensions, characterized by two axes of length $1/K$ with random directions. All datapoints are observed with noise $\mathcal{N}(\mathbf{0}, .001\,\boldsymbol{I})$. The functions `rand(D,T)` and `randn(D,T)` can for example be used to generate T random datapoints of D dimensions with uniform and normal distributions.



- 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 of 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 $d=2$ (same parameters as for data sampling).

- With the help of the `plotGMM3D` function, find a way to 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 the results in graphs.

Exercise 4: Subspace clustering *Vs* global dimensionality reduction

- With the dataset generated in *Exercise 1* and the models learned in *Exercise 2*, show that MPPCA is not the same as a global reduction of dimensionality with PCA, followed by GMM clustering of the data in the latent space.

- Show that MPPCA can produce different results compared to a GMM clustering of the original dataset followed by a projection of the datapoints of each cluster in a latent space with PCA.