# Exercise 10

***Learning objectives:*** *The aim of this exercise is to understand how NMF, AA, sparse coding and ICA can be used to decompose the data into prominent patterns and how in real data these approaches can be useful for the identification of the underlying structure of the data.*

***Resources:*** *Download the Matlab tools for today and run the script setup.m in order to install the tools used for the current exercise. Notice all methods are based on the optimization approaches described in today's lecture while the ICA method is based on the FAST_ICA toolbox.*

## *A qualitative comparison of unsupervised learning approaches*

1. Inspect and run the script *ex1.m* . The script analyses various synthetically generated data using the following unsupervised learning methods:
   **SVD, NMF, Archetypical Analysis (AA), Sparse Coding (SC), Non-negative Sparse Coding (NSC), ICA, and K-means**
   By changing the variable *method* you can switch between each of the approaches. The red lines generated in each plot indicate how the features extracted by each method accounts for the data while the variation explained (VE) gives the proportion of the variance in the data accounted for by the model. Explain for each of the six synthetically generated datasets which method you consider the most suited and discuss also how many components should be used to model the data.

## *Feature extraction and classification of hand-written digits*

2. Classification is an important problem in data mining. A simple and widely used classification approach is based on K-nearest neighbor (KNN). We will initially investigate if dimensionality reduction improves on classification of handwritten digits. Inspect the script *ex2.m*. Try analyzing the data using SVD with 25, 49, 100 and 256 components. What number of components results in the best classification performance? How will you describe the features extracted by SVD?
3. Try analyzing the data with 49 components using at least NMF and AA. What is the qualitative difference between the features extracted by each of the unsupervised learning methods? Discuss what the benefits/drawbacks of the approaches may be.

## *Finding the spectral profiles and concentrations in an NMR data experiment*

4. Load the data set *NMR_mix_DoEcompressed.mat.* The variable *xData* contains nuclear magnetic resonance measurements of mixtures of the three alcohols; propanol, butanol and ethanol. I.e. each of the 231 samples contains a given concentration fraction of propanol, butanol and ethanol (data taken from http://www.models.life.ku.dk/NMR_Mix_DoE). We are interested in identifying the spectral profiles of each compound as well as their concentration in each of the 231 measured samples using unsupervised learning on the measurements *xData* alone. Which of the methods:
   **SVD, NMF, Archetypical Analysis (AA), Sparse Coding (SC), Non-negative Sparse Coding (NSC), ICA, and K-means**

do you think will be the most suited for recovering the underlying concentrations of the three alcohols in the samples? Verify the method is indeed performing well by comparing the estimated concentrations by the model to the actual concentrations given in the variable *yData*.

## *Solving the cocktail party problem*

5. Load the data set *MixedSound.mat* . The data contains recording obtained from three microphones during a cocktail party where three bands unfortunately played at the same time. Can you separate the sounds coming from each of the bands playing by the use of any of the methods described today? (Notice: to listen to each recorded sound track use the matlab function *soundsc.m*).

## *Sparse Coding revisited*

6. Inspect the script *ex6.m* . The script loads the natural image data used in the Nature paper of Olshausen and Field. The method used for sparse coding is given in the matlab script *SparseCoding.m.* Run the script and inspect the features that are generated. (Note: If you are patient change *Npatch* to analyze the full dataset). Compare the features to the features generated by SVD. How do the features differ?

## *Implement the multiplicative updates of NMF*

7. Implement the multiplicative updates for NMF based on the least squares and KL-divergence objectives by inserting the following updates of W and H in the scripts *NMFLS.m* and *NMFKL.m*.

$$
\begin{aligned}
X_{i,j} \geq 0 \quad &, \quad w_{i,d} \geq 0 \quad and \quad h_{d,j} \geq 0 \\
C_{LS} &= \frac{1}{2}\|\mathbf{X} - \mathbf{WH}\|_F^2 = \frac{1}{2}\sum_{i,j}(\mathbf{X}_{i,j} - (\mathbf{WH})_{i,j})^2 \\
w_{i,d} &\leftarrow w_{i,d}\frac{(\mathbf{XH}^\mathrm{T})_{i,d}}{(\mathbf{WHH}^\mathrm{T})_{i,d}} \\
h_{d,j} &\leftarrow h_{d,j}\frac{(\mathbf{W}^\mathrm{T}\mathbf{X})_{d,j}}{(\mathbf{W}^\mathrm{T}\mathbf{WH})_{d,j}} \\
C_{KL} &= \sum_{i,j} x_{i,j}\log\frac{x_{i,j}}{(\mathbf{WH})_{i,j}} - x_{i,j} + (\mathbf{WH})_{i,j} \\
w_{i,d} &\leftarrow w_{i,d}\frac{\sum_j \frac{x_{i,j}}{(\mathbf{WH})_{i,j}} h_{d,j}}{\sum_j h_{d,j}} \\
h_{d,j} &\leftarrow h_{d,j}\frac{\sum_i w_{i,d}\frac{x_{i,j}}{(\mathbf{WH})_{i,j}}}{\sum_i w_{i,d}}
\end{aligned}
$$

Note that all elements in the W-update are updated simultaneously and all elements in the H-updated are updated simultaneously.

8. Try evaluating your implementation by running the script ex8.m. The script analyses the handwritten digits data by your NMF methods.