Homework 5

1) Imagine you have a dataset, X. You wish to send the data to a friend. You decide to discover the regularities in the data by fitting a model. You send the model-specification (say, the cluster means), a code-vector for each data-case (say, the assignment of each data-case to a cluster) and the errors in predicting the data-case from the code and the model (say, the vector $x_i - \mu_{z_i}$). Your objective is to send as few bits as possible without losing any information (assuming you are only interested in knowing the data up to finite precision or a fixed quantization level).

Argue why for small datasets you expect a simple models to be optimal while for large datasets you would expect more complex models to be optimal for this purpose.

2) Derive K-means from the cost function C (see slides) by showing that the two steps correspond to coordinate descend on C with respect to the variables $\{z_i, \mu_c\}$.

3) Is it possible for a probability density such as a normal density to have $p(x) > 1$?

4) Show that the K-means update equations are a special case of the EM update equations (hint: take some limit).

5) Derive a linear transformation $y=Ax$ based on principal components which has the property that the covariance $\text{Cov}(y)=\text{Identity}$.

6) Imagine you have a collection of gray-level images of faces. Provide pseudo-code to compute the “eigenfaces” of this collection. (see below for 4 examples of eigenfaces).