1 Latent semantic indexing (Matlab)

The file lsiDocuments.pdf contains 9 documents on various topics. A list of all the 460 unique words/terms that occur in these documents is in lsiWords.txt. A document by term matrix is in lsiMatrix.txt. Load this matrix and convert it to a standard term by document matrix as follows (note the transpose):

\[ X = \text{load('lsiMatrix.txt')}'; \]

Also, load the words as follows

\[
\begin{align*}
\text{fid} &= \text{fopen('lsiWords.txt')} ; \\
\text{tmp} &= \text{textscan(fid,'s')} ; \\
\text{fclose(fid)} ; \\
\text{words} &= \text{tmp(1)} ;
\end{align*}
\]

1. Compute the SVD of \( X \) and make an approximation to it \( \hat{X} \) using the first 2 singular values/vectors. Plot the low dimensional representation of the 9 documents in 2D. You should get something like Figure 1.

2. Consider finding documents that are about alien abductions. If you look at lsiWords.txt, there are 3 versions of this word, term 23 ("abducted"), term 24 ("abduction") and term 25 ("abductions"). Suppose we want to find documents containing the word "abducted". Documents 2 and 3 contain it, but document 1 does not. However, document 1 is clearly related to this topic. Thus LSI should also find document 1. Create a test document \( q \) containing the one word "abducted", and project it into the 2D subspace to make \( \hat{q} \). Now compute the cosine similarity between \( \hat{q} \) and the low dimensional representation of all the documents. What are the top 3 closest matches?

![Figure 1: Projection of 9 documents into 2 dimensions.](image-url)
2 Derivation of M step for GMM

Prove that the stationary points of

\[ J(\mu, \Sigma) = -\frac{1}{2} \sum_n \sum_k r_{nk} \left[ \log |\Sigma_k| + (x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k) \right] \]  

are given by

\[ r_k = \sum_n r_{nk} \]  
\[ \mu_k = \frac{\sum_n r_{nk} x_n}{r_k} \]  
\[ \Sigma_k = \frac{\sum_n r_{nk} (x_n - \mu_k^{\text{new}})(x_n - \mu_k^{\text{new}})^T}{r_k} \]  

Hint: you may find the following identities helpful

\[ \frac{\partial x^T A x}{\partial x} = (A + A^T)x \]  
\[ \frac{\partial \log |X|}{\partial X} = (X^{-1})^T = (X^T)^{-1} \]  
\[ \log |X| = -\log |X^{-1}| \]  
\[ \frac{\partial a^T X b}{\partial X} = ab^T \]

3 EM for a scale mixture of Gaussians

Consider the graphical model in Figure 2 which defines the following:

\[ p(x; \theta) = \sum_{j=1}^m \sum_{k=1}^l p_j q_k N(x; \mu_j, \sigma_k^2) \]

where

\[ N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2\sigma^2}(x - \mu)^2\right] \]

and \( \theta = \{p_1, \ldots, p_m, \mu_1, \ldots, \mu_m, q_1, \ldots, q_l, \sigma_1^2, \ldots, \sigma_l^2\} \) are all the parameters. (Here \( p_j \overset{\text{def}}{=} P(J = j) \) and \( q_k \overset{\text{def}}{=} P(K = k) \) are the equivalent of mixture weights.)

[We could view this as a simple mixture model with \( m \times l \) Gaussian components indexed by \((j, k)\). However, unlike before, the parameters of the \( ml \) components cannot be set independently. For example, there are only \( m \) possible means, not \( ml \). Alternatively, we could view this as a mixture of \( m \) non-Gaussian components, where each component distribution is a scale mixture, \( p(x|j; \theta) = \sum_{k=1}^l q_k N(x; \mu_j, \sigma_k^2) \), combining Gaussians with different variances (scales). These \( m \) components are again not parameterized independently of each other.]

We will now derive a generalized EM algorithm for this model. (Recall that in generalized EM, we do a partial update in the M step, rather than finding the exact maximum.)

1. Derive an expression for the responsibilities, \( P(J_n = j, K_n = k|x_n, \theta) \), needed for the E step.

2. Write out a full expression for the expected complete log-likelihood

\[ Q(\theta^{\text{new}}, \theta^{\text{old}}) = E_{\theta^{\text{old}}} \sum_{n=1}^N \log P(J_n, K_n, x_n|\theta^{\text{new}}) \]
3. Solving the M-step would require us to jointly optimize the means $\mu_1, \ldots, \mu_m$ and the variances $\sigma_1^2, \ldots, \sigma_l^2$. It will turn out to be simpler to first solve for the $\mu_j$’s given fixed $\sigma_j^2$’s, and subsequently solve for $\sigma_j^2$’s given the new values of $\mu_j$’s. For brevity, we will just do the first part. Derive an expression for the maximizing $\mu_j$’s given fixed $\sigma_{1,j}^2$, i.e., solve $\frac{\partial Q}{\partial \mu_j} = 0$. 