Computational Problems.

1. Mixture model problem

This problem shows that projecting onto proper subspace helps greatly in learning mixture models compared to projecting onto random subspace.

(a) Create a mixture model with the following specifications.

\[ X \in \mathbb{R}^d \] is a random variable whose distribution is a mixture of 4 gaussian random variables with means at \( e_1, e_2, e_3 \) and \( e_4 \) (\( e_i \) are standard basis) of magnitude 1 and variance 0.1. All 4 random variables are chosen uniformly at random. Draw \( \left\lfloor 4 \times d \log(d) \right\rfloor \) samples of \( X \) for \( d = 10, 100 \) and 500.

(b) Cluster the samples based on nearest neighbour. One way to do it is to classify any two points as from the same cluster if the distance between them is less than a threshold \( r \) (Look for function ‘pdist’ in matlab). Count the fraction of pairs in error. The error will depend on the threshold chosen \( r \). One good value for threshold is the diameter of a cluster, which is \( \sqrt{2 \times \sigma^2 \times d} \).

(c) Now project all the points into a low dimensional subspace spanning the means of the four gaussian random variables. Take the space spanned by \( \{e_1, \ldots, e_{10}\} \) for this case. Compute the same error for different values of \( d \).

(d) This time project all points into a randomly generated subspace of dimension 10 (One way is to use the range space of a random matrix). Compute error in this case for different values of \( d \).

(e) (15 points) Give a plot comparing errors in the above three scenarios.

Solution: The plot is shown in Figure 1. The errors have been normalized by \( n(n-1)/2 \) (maximum number of possible pairs), where \( n \) is the number of samples in each case.

2. Planted Model

Recall the planted model from class. Let \( n = 200 \), and \( k = 5 \). Form the matrix \( P \) as in class, that has \( p \) on the five equal blocks on the diagonal, and \( q = 1 - p \) everywhere else. Let \( A \) be the random matrix, as in class, where each entry \( A_{ij} \) is a Bernoulli random variable with probability \( P_{ij} \). Note that you will have to construct \( A \) as a symmetric matrix, so generate elements above the diagonal, and then just replicate them below.

(a) (5 points) For \( p = 0.8, 0.7, 0.6 \) and 0.55, generate the eigenvalues of \( P \) and of (the random matrix) \( A \) and plot them. How many eigenvalues of \( P \) are non-zero?
(b) (5 points) Now run the spectral clustering algorithm on $A$ (rather than the Laplacian). The points that you get will be 5 dimensional. Pick a random projection onto two dimensions, and project all of the resulting five dimensional points onto these two randomly chosen dimensions. Plot the results for the different values of $p$.

(c) (5 points) Cluster according to $k$-Means.

**Solution:** The various plots are shown in Figure 2-9. In all the cases the top 5 eigenvalues of $P$ are non-zero.

**Linear Algebra.**

The following problems focus on important concepts and ideas from linear algebra that are critical in this class. Please read up about singular value decomposition, and other topics as necessary. Problems marked with a * are optional.

1. Range and Nullspace of Matrices: Recall from class the definition of the null space and the range of a linear transformation, $T : V \rightarrow W$:

   $\text{null}(T) = \{ \mathbf{v} \in V : T\mathbf{v} = 0 \}$
   $\text{range}(T) = \{ T\mathbf{v} \in W : \mathbf{v} \in V \}$

   • (5 points) Suppose $A$ is a 10-by-10 matrix of rank 5, and $B$ is also a 10-by-10 matrix of rank 5. What is the **smallest** and **largest** the rank the matrix $C = AB$ could be?
   • (5 points) Now suppose $A$ is a 10-by-15 matrix of rank 7, and $B$ is a 15-by-11 matrix of rank 8. What is the **largest** that the rank of matrix $C = AB$ can be?
Figure 2: Plot showing the eigenvalues of the $P$ and $A$ matrices for $p = .8$.

Figure 3: Plot showing the eigenvalues of the $P$ and $A$ matrices for $p = .7$. 
Figure 4: Plot showing the eigenvalues of the $P$ and $A$ matrices for $p = .6$.

Figure 5: Plot showing the eigenvalues of the $P$ and $A$ matrices for $p = .55$. 
Figure 6: Projected points in random 2 direction, and clustered points by K-means for $p = .8$

Figure 7: Projected points in random 2 direction, and clustered points by K-means for $p = .7$
Figure 8: Projected points in random 2 direction, and clustered points by K-means for $p = .6$

Figure 9: Projected points in random 2 direction, and clustered points by K-means for $p = .55$
Solution: We know for any two matrices $A \ (m \times n)$ and $B \ (n \times k),$

$$\text{rank}(A) + \text{rank}(B) - n \leq \text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$$

Part I: Using the above bounds,

$$\text{rank}(C) \geq 5 + 5 - 10 = 0$$
$$\text{rank}(C) \leq \min(5, 5) = 5$$

The bounds are tight. For example take,

$$A = B = \begin{bmatrix} I_5 & 0 \\ 0 & 0 \end{bmatrix}$$

Then $C = AB = A$, therefore $\text{rank}(C) = 5$. Now if we take,

$$A = \begin{bmatrix} I_5 & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 0 & I_5 \end{bmatrix}$$

then $C = AB = 0$ the all zero matrix, hence $\text{rank}(C) = 0$.

Part II: $\text{rank}(C) \leq \min(7, 8) = 7$. The bound is also tight. Let

$$A = \begin{bmatrix} I_7 & 0_{7\times8} \\ 0_{3\times7} & 0_{3\times8} \end{bmatrix}, \quad B = \begin{bmatrix} I_8 & 0_{8\times3} \\ 0_{7\times8} & 0_{7\times3} \end{bmatrix}$$

Then,

$$C = AB = \begin{bmatrix} I_7 & 0_{7\times4} \\ 0_{3\times7} & 0_{3\times4} \end{bmatrix}$$

Therefore $\text{rank}(C) = 7$.

2. Riesz Representation Theorem: (5 points) Consider the standard basis for $\mathbb{R}^n$: $e_1 = (1, 0, \ldots, 0), e_2 = (0, 1, 0, \ldots, 0)$, etc. Recall that the inner-product of two vectors $w_1 = (\alpha_1, \ldots, \alpha_n), w_2 = (\beta_1, \ldots, \beta_n) \in \mathbb{R}^n$, is given by:

$$\langle w_1, w_2 \rangle = \sum_{i=1}^{n} \alpha_i \beta_i.$$ 

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a linear map. Show that there exists a vector $x \in \mathbb{R}^n$, such that

$$f(w) = \langle x, w \rangle,$$

for any $w \in \mathbb{R}^n$.

Remark: It turns out that this result is true in much more generality. For example, consider the vector space of square-integrable functions (something we will see much more later in the course). Let $F$ denote a linear map from square integrable functions to $\mathbb{R}$. Then, as a
consequence similar to the finite dimensional exercise here, there exists a square integrable function, \( g \), such that:

\[
F(f) = \int f g.
\]

**Solution:** Let \( x = \sum_{i=1}^{n} f(e_i)e_i \) and \( w = \sum_{i=1}^{n} w_ie_i \). Then using linearity of \( f() \) we have,

\[
f(w) = f(\sum_{i=1}^{n} w_ie_i) = \sum_{i=1}^{n} w_i f(e_i) = \langle x, w \rangle
\]

3. (10 points) Recall from class that the spectral theorem for symmetric \( n \times n \) real matrices, says, among other things, that if \( A \) is a symmetric (real) \( n \times n \) matrix, then it has an basis of orthonormal eigenvectors, \( \{v_1, \ldots, v_n\} \). Use \( A \) and \( \{v_i\} \) to construct a matrix \( T \), such that the matrix \( T^T A T \) is diagonal.

**Solution:** Let \( T = \begin{bmatrix} v_1 & v_2 & \ldots & v_n \end{bmatrix} \), i.e. the matrix with orthonormal eigenvectors of \( A \) as columns. Let \( \lambda_i \) be the eigenvalue corresponding to the eigenvector \( v_i \). Then,

\[
T^T A T = \begin{bmatrix} v_1^T \\ \vdots \\ v_n^T \end{bmatrix} \begin{bmatrix} Av_1 & Av_2 & \ldots & Av_n \end{bmatrix} = \begin{bmatrix} v_1^T \\ \vdots \\ v_n^T \end{bmatrix} \begin{bmatrix} \lambda_1 v_1 & \lambda_2 v_2 & \ldots & \lambda_n v_n \end{bmatrix}
\]

\[
= \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix}
\]

Since \( ||v_i||^2 = 1 \) and \( v_i^T v_j = 0 \) for all \( i \neq j \).

4. (10 points) Let \( A \) be a rank \( n \) matrix (of any dimensions, not necessarily square). Let its singular value decomposition be given by

\[
A = U \Sigma V^*,
\]

where \( \Sigma \) is the matrix of singular values given in *descending order*: \( \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \). Let \( \Sigma_k \) denote the matrix with \( \sigma_{k+1}, \ldots, \sigma_n \) set to zero, and let \( \hat{A} = U \Sigma_k V^* \). Show that \( \hat{A} \) solves the optimization problem:

\[
\min_{\hat{A} : \text{rk}(\hat{A}) \leq k} ||A - \hat{A}||_F.
\]

Hint: you can use the fact that the optimal solution should satisfy \( (A - \hat{A}) \perp \hat{A} \), where orthogonality is defined with respect to the natural matrix inner product compatible with the Frobenius norm:

\[
\langle M, N \rangle = \sum_{ij} M_{ij}N_{ij} = \text{Trace}(M^*N).
\]

If you choose to use this hint, please do show that \( \hat{A} \) and \( A \) satisfy the orthogonality, as claimed.

**Solution:** There can be several proofs. The following has been adapted from [1]. Let \( x^* \) be the optimal value of the optimization problem. Now, \( \hat{A} = U \Sigma_k V^* \) is feasible and

\[
||A - U \Sigma_k V^*||_F^2 = \sum_{i=k+1}^{n} \sigma_i^2.
\]

Hence,
\[ x^* \leq \sqrt{\sum_{i=k+1}^{n} \sigma_i^2} \]  

(1)

Now we find a lower bound for \( x^* \). First we prove the following lemmas.

**Lemma 1:** For any two matrices \( A, B \in \mathbb{C}^{m \times n} \) with \( \text{rank}(B) \leq k \),

\[ \sigma_1(A - B) \geq \sigma_{k+1}(A). \]

**Proof:** For any unit vector \( v \) we have \( \sigma_1(A - B)^2 = \|A - B\|^2 \geq v^*(A - B)(A - B)v \). Now \( \text{rank}(B) \leq k \) implies \( \text{dim}(\text{Null}(B)) > n - k \). Let \( U = \text{span}\{v_1, \ldots, v_{k+1}\} \) be the span of top \( k+1 \) right singular vectors of \( A \). From dimensionality argument it follows \( \text{Null}(B) \cap U \neq \phi \). So let \( v = \sum_{i=1}^{k+1} a_iv_i \in \text{Null}(B) \cap U \). Then, \( v^*(A - B)^*(A - B)v = v^*Av = \sum_{i=1}^{k+1} \sigma_i(A)^2a_i^2 \geq \sigma_{k+1}(A)^2 \). Hence \( \sigma_1(A - B) \geq \sigma_{k+1}(A) \).

**Lemma 2:** Let \( A = B + C \), then \( \sigma_{i+j-1}(A) \leq \sigma_i(B) + \sigma_j(C) \).

**Proof:** For \( i = j = 1 \) we have \( \sigma_1(A) = u_1^*Av_1 = u_1^*(B + C)v_1 = u_1^*Bv_1 + u_1^*Cv_1 \leq \sigma_1(B) + \sigma_1(C) \). Now for any matrix \( X \), let \( X_k \) be the matrix with last \( n - k \) singular values of \( X \) set to 0. Hence \( \sigma_1(A - A_k) = \sigma_k(A) \). Now using Lemma 1 and the fact that \( \text{rank}(B_{i-1} + C_{j-1}) \leq i + j - 2 \) we get,

\[ \sigma_i(B) + \sigma_j(C) = \sigma_1(B - B_{i-1}) + \sigma_1(C - C_{j-1}) \geq \sigma_1(B + C - (B_{i-1} + C_{j-1})) \geq \sigma_{i+j-1}(A). \]

**Lemma 3:** For any two matrices \( A, B \in \mathbb{C}^{m \times n} \) with \( \text{rank}(B) \leq k \), \( \sigma_i(A - B) \geq \sigma_{i+k}(A) \).

**Proof:** As \( \text{rank}(B) \leq k \), \( \sigma_{k+1}(B) = 0 \). Therefore by applying Lemma 2 with \( A - B \) and \( B \) we get,

\[ \sigma_i(A - B) = \sigma_i(A - B) + \sigma_{k+1}(B) \geq \sigma_{i+k+1-1}(A - B + B) = \sigma_{i+k}(A) \]

Now we show the lower bound. For any matrix \( A \) and \( \hat{A} \) with \( \text{rank}(\hat{A}) \leq k \) using Lemma 3 we have,

\[ \|A - \hat{A}\|_F^2 = \sum_{i=1}^{n+k} \sigma_i(A - B)^2 \geq \sum_{i=1}^{n-k} \sigma_i(A - B)^2 \geq \sum_{i=1}^{n-k} \sigma_{i+k}(A)^2 = \sum_{i=k+1}^{n} \sigma_i(A)^2 \]  

(2)

Therefor \( x^* \geq \sqrt{\sum_{i=k+1}^{n} \sigma_i(A)^2} \). Combining with the upper bound in equation (1) we have \( x^* = \sqrt{\sum_{i=k+1}^{n} \sigma_i(A)^2} \). Hence \( \hat{A} = U\Sigma_k V^* \) is an optimal solution.

5. * Prove the above hint, namely, that the optimal solution must satisfy the orthogonality condition (in the previous hint, you are assuming that this is true).

**Solution:** Let \( A \) be \( m \times n \) matrix. We have \( \|A - \hat{A}\|_F^2 = \sum_{i=1}^{m} |A_i - \hat{A_i}|^2 \), the sum of the squared distances between the rows of \( A \) and \( \hat{A} \). Rows of \( \hat{A} \) spans at most a dimension \( k \) subspace \( V_k \). Let \( \text{proj}_{V_k}(A) \) be the matrix of rows being projection of the corresponding rows of \( A \) onto \( V_k \). We argue that the optimal solution will always have the rows of \( A \) projected onto the optimal subspace \( V_k \). This is because from Pythagoras’ theorem for any \( B \in V_k \), \( \|A - B\|_F^2 \geq \|A - \text{proj}_{V_k}(B)\|_F^2 \) and \( \text{proj}_{V_k}(B) \) is also a rank \( \leq k \) matrix, hence it
is optimum. Now let $\hat{A}$ be the optimal solution and $V_k$ the optimal subspace. Then rows of $A - \hat{A}$ are perpendicular to subspace $V_k$ hence to the rows of $\hat{A}$. Therefore $\langle A - \hat{A}, \hat{A} \rangle = \sum_{i=1}^{m} \langle A_i - \hat{A}_i, \hat{A}_i \rangle = 0$. Hence proved.

6. (10 points) Let $A \in \mathbb{R}^{n \times n}$ be symmetric. Recall that by the spectral theorem, $A$ will have real eigenvalues. Therefore we can order the eigenvalues of $A$: $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A)$. Show that:

$$
\lambda_1(A) = \max_{y \neq 0} \frac{\langle y, Ay \rangle}{\langle y, y \rangle},
$$

$$
\lambda_k(A) = \max_{V: \dim V = k} \min_{0 \neq y \in V} \frac{\langle y, Ay \rangle}{\langle y, y \rangle}.
$$

Solution: Let $\{v_1, \ldots, v_n\}$ be the set of orthonormal eigenvectors of $A$, $\lambda_i(A)$ is the eigenvalue corresponding to $v_i$. Let $V$ be any subspace of dimension $k$, and let $U = \text{span}\{v_k, v_{k+1}, \ldots, v_n\}$. Then $V \cap U \neq \emptyset$ else dim$(V) + \dim(U) = k + (n - k + 1) = n + 1 > n$ a contradiction. Let $w = \sum_{i=k}^{n} w_i v_i \in V \cap U$. Now,

$$
\frac{\langle w, Aw \rangle}{\langle w, w \rangle} = \frac{\sum_{i=k}^{n} w_i^2 \lambda_i}{\sum_{i=k}^{n} w_i^2} \leq \lambda_k \quad \text{therefore}
$$

$$
\min_{0 \neq y \in V} \frac{\langle y, Ay \rangle}{\langle y, y \rangle} \leq \lambda_k
$$

$$
\max_{V: \dim V = k} \min_{0 \neq y \in V} \frac{\langle y, Ay \rangle}{\langle y, y \rangle} \leq \lambda_k \quad (3)
$$

Now take $V = \text{span}\{v_1, \ldots, v_k\}$. Then for any $y = \sum_{i=1}^{k} y_i v_i \in V$,

$$
\frac{\langle y, Ay \rangle}{\langle y, y \rangle} = \frac{\sum_{i=1}^{k} y_i^2 \lambda_i}{\sum_{i=1}^{k} y_i^2} \geq \lambda_k
$$

The equality is achieved when $y = v_k$. Hence the inequality in equaltion $[3]$ is achieved with equality. Therefore,

$$
\max_{V: \dim V = k} \min_{0 \neq y \in V} \frac{\langle y, Ay \rangle}{\langle y, y \rangle} = \lambda_k
$$

7. * The spectral radius of a matrix $A$ is defined as:

$$
\rho(A) = \max\{|\lambda| : \lambda \text{ an e-value of } A\}.
$$

Note that the spectral radius is invariant under similarity transformations, and thus we can speak of the spectral radius of a linear operator.

(a) Show that $\rho(A) \leq \sigma_1(A)$. 


(b) Show that

$$\rho(A) = \inf_{\{S: \det S \neq 0\}} \sigma_1(S^{-1}AS).$$

**Solution:** Let $v$ be the normalized eigenvector corresponding to the maximum magnitude eigenvalue $\lambda$.

(a) $\sigma_1(A) = \max_{u_1, u_2: ||u_1||=||u_2||=1} |u_1^*Au_2| \geq |v^*(\lambda v)| = |\lambda| = \rho(A)$

(b) Note that matrix eigenvalues remain the same under similarity transforms. Hence for any non-degenerate matrix $S$, eigenvalues of $A$ and $S^{-1}AS$ are the same. Therefore $\rho(A) = \rho(S^{-1}AS)$. Now,

$$\rho(S^{-1}AS) \leq \sigma_1(S^{-1}AS)$$

$$\rho(A) = \inf_{\{S: \det S \neq 0\}} \rho(S^{-1}AS) \leq \inf_{\{S: \det S \neq 0\}} \sigma_1(S^{-1}AS) \quad (4)$$

Now let $A = U^*\Delta U$ be the Schur decomposition of $A$, where $\Delta$ is an upper triangular matrix and $U$ an unitary matrix. The diagonal entries of $\Delta$ are the eigenvalues of $A$. Now let $D_t = \text{diag}(t, t^2, \ldots, t^n)$. Then,

$$A_t = D_t\Delta D_t^{-1} = \begin{bmatrix}
\lambda_1 t^{-1}\Delta_{12} & t^{-2}\Delta_{13} & \cdots & t^{-n+1}\Delta_{1n} \\
0 & \lambda_2 t^{-1}\Delta_{23} & \cdots & t^{-n+1}\Delta_{2n} \\
0 & 0 & \cdots & t^{-1}\Delta_{n-1,n} \\
0 & 0 & 0 & \lambda_n
\end{bmatrix}$$

Now for any $\epsilon > 0$, by making $t$ large enough we can ensure that the sum of all off-diagonal components of $A_t$ is less than $\epsilon$. Hence choose $t$ such that, $||A_t||_1 \leq \rho(A) + \epsilon$ and $||A_t||_\infty \leq \rho(A) + \epsilon$. Define $S_t = U^*D_t^{-1}$. Then,

$$\inf_{\{S: \det S \neq 0\}} \sigma_1(S^{-1}AS) \leq \sigma_1(S_t^{-1}AS_t) = \sigma_1(D_t^{-1}\Delta D_t) = ||A_t||_2 \leq \sqrt{||A_t||_1||A_t||_\infty} \leq \rho(A) + \epsilon$$

Combining with equation (4) and taking $\epsilon \rightarrow 0$ we get the required result.

8. (10 points) A linear operator $N$ is called *nilpotent* if for some integer $k$, $N^k = 0$. Show that if $N$ is nilpotent, then $(I + N)$ has a square root. (Hint: consider the Taylor expansion of $(\sqrt{1+x}$, and use that as a starting point).

**Solution:** Let the Taylor series expansion of $\sqrt{1+x}$ about $x=0$ is given by,

$$\sqrt{1+x} = a_0 + \sum_{i=1}^{\infty} a_i x^i$$

Where $a_0 = 1$. Since, $1 + x = (\sqrt{1+x})^2 = (1 + \sum_{i=1}^{\infty} a_i x^i)^2$. We have,

$$2a_1 = 1$$

$$\sum_{j=0}^{i} a_j a_{i-j} = 0 \quad \forall i \geq 2$$
Define $A_N = a_0 I + \sum_{i=1}^{\infty} a_i N^i = a_0 I + \sum_{i=1}^{k-1} a_i N^i$ (since $N^i = 0$ for all $i \geq k$). The sum converges since there are only finite number of terms, hence $A_N$ is well defined. Identically $A_N * A_N = (a_0 I + \sum_{i=1}^{\infty} a_i N^i)^2 = I + N$. Hence $I + N$ has a square root given by $A_N$.

9. (5 points) Find an example of a matrix that is diagonalizable, but not unitarily so. That is, produce an example of an $n \times n$ matrix $A$ (up to you) for which there is some invertible matrix $T$ that satisfies $T^{-1}AT = D$ for some diagonal matrix, but the columns of $T$ cannot be taken to be orthonormal. Hint: related to one of the problems above.

**Solution:** We have $T^{-1}AT = D$. Take any non-orthogonal but invertible matrix $T$ and a diagonal matrix $D$. Then the $A$ matrix which is diagonalizable by $T$ is given by $A = TDT^{-1}$.

For example,

$$
T = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}
$$

$$
A = \begin{bmatrix} 2 & 0 \\ 1 & 1 \end{bmatrix}
$$

10. (10 points) Suppose $A \in \mathbb{C}^{m \times n}$ has full column rank $n$. Show that:

$$
\min_{\Delta \in \mathbb{C}^{m \times n}} \{ \|\Delta\|_2 \mid \text{rank}(A + \Delta) < n \} = \sigma_n(A),
$$

where $\sigma_n(A)$ denotes the smallest singular value of $A$.

**Solution:** Let the SVD of $A$ be given by: $A = U\Sigma V^*$. Note that we can equivalently write this using a sum of rank-one matrices that are each outer products of the left and right singular vectors:

$$
A = \sum_{i=1}^{n} \sigma_i u_i v_i^*.
$$

Then letting $\Delta = -\sigma_n u_n v_n^*$, we conclude that the value of the minimization above is no more than $\sigma_n(A)$ (since we have exhibited a perturbation that causes $A$ to drop rank, and has spectral norm $\sigma_n(A)$). We have left to show that it is not possible to do any better.

Since $A$ has rank $n$, by rank-nullity, its null-space contains only the zero vector. Suppose $\Delta$ is any perturbation that causes $A$ to drop rank. Let $v$ be any unit vector in the null space of $(A + \Delta)$ (note that this vector is not in the nullspace of $A$). Then we have $(A + \Delta)v = 0$, and hence $\|Av\|_2 = \|\Delta v\|_2$, whence

$$
\|\Delta\| \geq \|\Delta v\|_2 = \|Av\|_2 \geq \min_{\|z\|_2 = 1} \|Az\|_2 = \sigma_n(A).
$$

References

[1] M. Raghupati, "Low Rank Approximation in the Frobenius Norm and Perturbation of Singular Values".

12