Problem Set 2: Linear Algebra for Data Sciences

1. Consider the yield curve for Treasury securities. The "yield" is the interest rate paid on the bonds, which depends on the time to maturity (called "tenor"). The yield is varied from day to day by the Federal Reserve. The table below shows the interest rate over 6 business days of 2001.

	Us Treasury Yield in 2001					
Tenor	**Jan 3**	**Jan 4**	**Jan 5**	**Jan 6**	**Jan 7**	**Jan 10**
3 MO	**5.87**	**5.69**	**5.37**	**5.12**	**5.19**	**5.24**
6 MO	**5.58**	**5.44**	**5.20**	**4.98**	**5.03**	**5.11**
1 Yr	**5.11**	**5.04**	**4.82**	**4.60**	**4.61**	**4.71**
2 Yr	**4.87**	**4.92**	**4.77**	**4.56**	**4.54**	**4.64**
3 Yr	**4.82**	**4.92**	**4.78**	**4.57**	**4.55**	**4.65**
5 Yr	**4.76**	**4.94**	**4.82**	**4.66**	**4.65**	**4.73**
7 Yr	**4.97**	**5.18**	**5.07**	**4.93**	**4.94**	**4.98**
10 Yr	**4.92**	**5.14**	**5.03**	**4.93**	**4.94**	**4.98**
20 Yr	**5.46**	**5.62**	**5.56**	**5.50**	**5.52**	**5.53**

Table 1: 6-days U.S. Treasury Yields

The PCA is used to analyse this data. First, we construct the differences of the interest rate between days with the mean difference subtracted from each row. Therefore, our data matrix $\mathbf{A}$ is 9-by-5 matrix with its rows adding to zero. The singular vectors $\mathbf{u}_i$ of $\mathbf{A}$ for $i = 1, \ldots, 5$ are given in the table below.

$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\mathbf{u}_4$	$\mathbf{u}_5$
0.3833	-0.5302	0.4795	-0.0682	0.1048
0.3366	-0.4376	0.0414	-0.2007	-0.1434
0.3584	-0.2643	-0.2330	0.5053	-0.1726
0.3492	0.0333	-0.4418	-0.1571	0.8106
0.3718	0.1302	-0.4405	-0.2266	-0.4496
0.3505	0.2925	-0.1224	0.2030	-0.1904
0.3242	0.3648	0.2276	-0.4517	-0.1182
0.2984	0.3771	0.3541	0.5696	0.1594
0.1848	0.2803	0.3642	-0.2322	0.0623

(a) Explain why we should expect at most 4 non-zero singular values of $\mathbf{A}$ instead of 5.

[sol]**:** Sine the five column vectors add to the zero vector, the rank of $\mathbf{A}$ is at most 4.

(b) The singular values of $\mathbf{A}$ (for the "compact" SVD) are $\sigma_1 = 36.39$, $\sigma_2 = 19.93$, $\sigma_3 = 5.85$, $\sigma_4 = 1.19$, $\sigma_5 = 0$. How many singular vector will you use to explore the important feature in this data? Explain your justification.

[sol]**:** The gap between the two consecutive singular values becomes "flat" at σ_3. So, 3 singular vectors should be adequate to explain all variation in the data.

(c) The singular vectors $\mathbf{u}_i$ of $\mathbf{A}$ for $i = 1, \ldots, 5$ are given in the above table. How would you interpret $\mathbf{u}_1$, $\mathbf{u}_2$ and $\mathbf{u}_3$ in the financial context?

[sol]**:** $\mathbf{u}_1$ measures a weight average of the daily change in the 9 yields. $\mathbf{u}_2$ measures the daily change in the yield spread between long and short bonds. $\mathbf{u}_3$ shows daily change in the curvature (i.e. short and long bonds vs medium).

2. Suppose that the SVD of a m-by-n (centred) data matrix $\mathbf{X}$ is given by $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$. When developing PCA, we solve the minimization problem $\min\|\mathbf{X} - \mathbf{C}\mathbf{T}\|$ s.t. $\mathbf{C}^T\mathbf{C} = \mathbf{I}$. The principle component scores (PCs) are defined by the columns of $\mathbf{C} := \mathbf{U}_d$ and the loadings are the columns of $\mathbf{T} := \Sigma_d\mathbf{V}_d^T$ (read again how the EYT is used to construct PCA).

   (a) The choice of $\mathbf{C}$ and $\mathbf{T}$ are non-unique. Provide another example of $\mathbf{C}$ and $\mathbf{T}$ under $\mathbf{C}^T\mathbf{C} = \mathbf{I}$. What is a possible implication of using other definitions of $\mathbf{C}$ and $\mathbf{T}$ in practice?

   [sol]**:** When applying EYT, the solution is given by $\mathbf{C}\mathbf{T} = \mathbf{U}_d\Sigma_d\mathbf{V}_d^T$ but not how we will choose $\mathbf{C}$ and $\mathbf{T}$. Hence, $\mathbf{C}$ and $\mathbf{T}$ are non-unique. In particular, we can choose $\mathbf{C} = \mathbf{U}_d\mathbf{R}$ and $\mathbf{T} = \mathbf{R}^T\Sigma_d\mathbf{V}_d^T$ where $\mathbf{R}$ is a d-by-d matrix with orthonormal columns. It is easy to check that:(1) $\mathbf{C}\mathbf{T} = \mathbf{U}_d\mathbf{R}\mathbf{R}^T\Sigma_d\mathbf{V}_d^T = \mathbf{U}_d\Sigma_d\mathbf{V}_d^T$ since $\mathbf{R}\mathbf{R}^T = \mathbf{I}$. (2) $\mathbf{C}\mathbf{C}^T = \mathbf{U}_d\mathbf{R}\mathbf{R}^T\mathbf{U}_d^T = \mathbf{U}_d\mathbf{U}_d^T = \mathbf{I}$. However, the score vectors will not align with the direction of the largest variation of data.

   (b) If an arbitrary basis can be used to span S, i.e. no constraints of $\mathbf{C}^T\mathbf{C}$ at all, provide another example of $\mathbf{C}$ and $\mathbf{T}$.

   [sol]**:** We can choose $\mathbf{C} := \mathbf{U}_d\mathbf{H}$ and $\mathbf{T} = \mathbf{H}^{-1}\Sigma_d\mathbf{V}_d^T$. The columns of $\mathbf{C}$ will not be orthogonal in general but they can still

form the basis (i.e. all columns of $\mathbf{C}$ are still independent) since $\mathbf{H}$ is invertible.

3. Consider solving $\mathbf{Ax} = \mathbf{y}$ for $\mathbf{x}$. Suppose $\mathbf{A}$ is a $m-$by$-n$ matrix with a rank of r. If $r = n < m$, show that $\|\mathbf{Ax} - \mathbf{y}\|^2$ is minimized by

$$\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}.$$

[sol]**:** $L(\mathbf{x}) := (\mathbf{Ax} - \mathbf{y})^T(\mathbf{Ax} - \mathbf{y})$.

$L(\mathbf{x}) = \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - \mathbf{x}^T\mathbf{A}^T\mathbf{y} - \mathbf{y}^T\mathbf{Ax} - \mathbf{y}^T\mathbf{y} = \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - 2\mathbf{y}^T\mathbf{Ax} - \mathbf{y}^T\mathbf{y}$

The optimal condition is

$$\frac{\partial L}{\partial \mathbf{x}} = 2\mathbf{A}^T\mathbf{Ax} - 2\mathbf{A}^T\mathbf{y} = 2\mathbf{A}^T(\mathbf{Ax} - \mathbf{y})$$

$$\frac{\partial L}{\partial \mathbf{x}} = 0 \Rightarrow \mathbf{A}^T(\mathbf{Ax} - \mathbf{y}) = 0 \Rightarrow \mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}.$$

4. Consider solving $\mathbf{Ax} = \mathbf{y}$ for $\mathbf{x}$. Suppose $\mathbf{A}$ is a $m-$by$-n$ matrix with a rank of r. If $r = m < n$, show that the solution of the following constraint minimization

$$\begin{array}{ll} \text{minimize} & \|\mathbf{x}\|^2 \\ \text{subject to} & \mathbf{Ax} = \mathbf{y} \end{array}$$

is given by $\mathbf{x} = \mathbf{A}^T(\mathbf{AA}^T)^{-1}\mathbf{y}$.
[sol]**:**The Lagrange multiplier method can be used for this problem. To this end, we define the Lagrange multiplier:

$$L(\mathbf{x}, \lambda) := \mathbf{x}^T\mathbf{x} + \lambda^T(\mathbf{Ax} - \mathbf{y}).$$

The optimal conditions are $\frac{\partial L}{\partial \mathbf{x}} = 2\mathbf{x} + \mathbf{A}^T\lambda = 0$ and $\frac{\partial L}{\partial \lambda} = \mathbf{Ax} - \mathbf{y} = 0$. The first condition gives $\mathbf{x} = -\mathbf{A}^T\lambda/2$. Substituting this into the second condition gives $\lambda = -2(\mathbf{AA}^T)^{-1}\mathbf{y}$. Combining these two conditions give the desired solution.

It can also be derived w/o the Lagrange multiplier. Notice that

$$\mathbf{x} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{y} \Rightarrow \mathbf{A}\mathbf{x} = (\mathbf{A}\mathbf{A}^T)(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{y} \Rightarrow \mathbf{A}\mathbf{x} = \mathbf{y}.$$

But the solution is not unique in this case (Why?). So, let assume $\mathbf{x}^*$ to be another solution (i.e. $\mathbf{A}\mathbf{x}^* = \mathbf{y}$). Now we need to show that $\|\mathbf{x}^*\| \geq \|\mathbf{x}\|$.

$$(\mathbf{x}^* - \mathbf{x})^T\mathbf{x} = (\mathbf{x}^* - \mathbf{x})^T\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{y} = (\mathbf{A}\mathbf{x}^* - \mathbf{A}\mathbf{x})^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{y} = 0$$

So, $(\mathbf{x}^* - \mathbf{x}) \perp \mathbf{x}$. And we have

$$\|\mathbf{x}^*\|^2 = \|\mathbf{x} + \mathbf{x}^* - \mathbf{x}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{x}^* - \mathbf{x}\|^2 \geq \|\mathbf{x}\|^2.$$

5. Let

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

Then

$$\mathbf{A}^\dagger = \begin{bmatrix} 1/4 & 1/4 \\ 1/4 & 1/4 \\ 0 & 0 \end{bmatrix}.$$

Decompose the vector $[2\ \ 3\ \ 4]^T$ uniquely into the sum of a vector in $N(\mathbf{A})^\perp$ and a vector in $N(\mathbf{A})$.
[sol]**:** $\mathbf{A}^\dagger\mathbf{A}$ is a projection onto $N(\mathbf{A})^\perp$ and $\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}$ is a projection onto $N(\mathbf{A})$.

$$\begin{aligned} \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} &= \mathbf{A}^\dagger\mathbf{A}\mathbf{x} + \left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right)\mathbf{x} \\ &= \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} + \begin{bmatrix} 1/2 & -1/2 & 0 \\ -1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} \\ &= \begin{bmatrix} 5/2 \\ 5/2 \\ 0 \end{bmatrix} + \begin{bmatrix} -1/2 \\ 1/2 \\ 4 \end{bmatrix} \end{aligned}$$

6. Consider the least square problem $\mathbf{y} = \mathbf{A}\mathbf{x}$ where $\mathbf{A}$ is a m-by-n matrix with rank r. Let $\mathbf{R}_n = \mathbf{A}^\dagger\mathbf{A}$ and $\mathbf{R}_m = \mathbf{A}\mathbf{A}^\dagger$.

   (a) Show that $\mathbf{R}_n = \mathbf{I}_n$ if $r = n$.
   [sol]**:** $\mathbf{A}^\dagger\mathbf{A} = \mathbf{V}_r\Sigma_r^{-1}\mathbf{U}_r^T\mathbf{U}_r\Sigma_r\mathbf{V}_r^T = \mathbf{V}_r\mathbf{V}_r^T$. If $r = n$, $\mathbf{V}_r$ has n orthonormal columns, so $\mathbf{V}_r$ is $n-$by$-n$ and $\mathbf{V}_r^T = \mathbf{V}_r -1$.

(b) Show that $\mathbf{R}_m = \mathbf{I}_m$ if $r = m$.
[sol]: $\mathbf{A}\mathbf{A}^\dagger\mathbf{A} = \mathbf{U}_r\Sigma_r\mathbf{V}_r^T\mathbf{V}_r\Sigma_r^{-1}\mathbf{U}_r^T = \mathbf{U}_r\mathbf{U}_r^T$. If $r = m$, $\mathbf{U}_r$ has m orthonormal columns, so $\mathbf{U}_r$ is $m-$by$-m$ and $\mathbf{U}_r^T = \mathbf{U}_r^{-1}$.

7. Consider solving a problem $\mathbf{Ax} = \mathbf{y}$. Let $\mathbf{x}^+$ be the solution associated with the data $\mathbf{y}$ and $\mathbf{x}_*^+$ be the solution associated with $\mathbf{y}_*$. Suppose that both $\mathbf{y}$ and $\mathbf{y}_*$ are in the column space of $\mathbf{A}$. Show that

$$\frac{\|\mathbf{x}^+ - \mathbf{x}_*^+\|_2}{\|\mathbf{x}^+\|_2} \leq \frac{\sigma_1}{\sigma_r}\frac{\|\mathbf{y} - \mathbf{y}_*\|_2}{\|\mathbf{y}\|_2}.$$

(Hint: Use $\|\mathbf{Bz}\| \leq \|\mathbf{B}\|\|\mathbf{z}\|$).Explain how this inequality can be used to relate a numerical stability of the generalized inverse solution with the singular values of $\mathbf{A}$.
[sol]: $\mathbf{x}^+ - \mathbf{x}_*^+ = \mathbf{A}^\dagger(\mathbf{y} - \mathbf{y}_*)$ Then, we have $\|\mathbf{x}^+ - \mathbf{x}_*^+\| \leq \|\mathbf{A}^\dagger\|\|\mathbf{y} - \mathbf{y}_*\|$. Since $\|\mathbf{A}^\dagger\|$ is the largest singular values of $\mathbf{A}^\dagger$, $\|\mathbf{A}^\dagger\| = 1/\sigma_r$. Thus, $\|\mathbf{x}^+ - \mathbf{x}_*^+\| \leq \sigma_r^{-1}\|\mathbf{y} - \mathbf{y}_*\|$. Also, $\|\mathbf{y}\| = \|\mathbf{Ax}^+\| \leq \|\mathbf{A}\|\|\mathbf{x}^+\| = \sigma_1\|\mathbf{x}^+\|$. Combining these inequalities gives the desired result.
This inequality suggests that if σ_1/σ_r is very large, a small change in the vector $\mathbf{y}$ can lead to a large difference in the generalized inverse solution. This has an implication in practice. If a true vector $\mathbf{x}$ has to be reconstructed from a true data $\mathbf{y}$ and a given model $\mathbf{A}$, then even a small perturbation of $\mathbf{y}$ would make it difficult to recover the true vector $\mathbf{x}$ through $\mathbf{A}^\dagger$.

8. Consider 8 samples of (width,length)-measurements of leaves (both in cm): $(2,10), (2,5), (8,4), (5,8), (7,5), (6,4), (1,2), (4,9)$.

(a) Use the k-mean algorithm to group the above data into 3 clusters. Note you can choose any initialization of the centroid.
[sol]: With the initialization of $\mu_1 = [2,10]^T$, $\mu_2 = [5,8]^T$ and $\mu_3 = [1,2]^T$. The K-mean algorithm takes 3 iterations to converge to the following clusters: $\{(2,10),(5,8),(4,9)\}$, $\{(2,5),(1,2)\}$ and $\{(8,4),(7,5),(6,4)\}$. The number of iteration depends on your initialization. Your solution might also be different since the K-mean algorithm only guarantees a local minimum.

(b) If you lack of a good prior knowledge of the number of clusters, discuss how will use the k-mean algorithm to help you decide on the "optimal" number of clusters?
[sol]: Running the K-mean algorithm to obtain the partition into $2, 3, \ldots, 7$ clusters. The cases of 1-cluster and 8-cluster are trivial.

Then, we can evaluate these clusters based on the distortion function. Then, choose the number of clusters that gives the minimum distortion function.

(c) It was later reported that the above data came from 3 types of leaves: $\{(2,10),(5,8),(4,9)\},\{(2,5),(1,2)\},\{(8,4),(7,5),(6,4)\}$. Suppose that we construct a linear classification model for this data in a form of $\mathbf{T} = \mathbf{XW}$. Determine $\mathbf{T}$ and $\mathbf{X}$. Explain how you will solve for $\mathbf{W}$ (without carrying out the computation). Can the classification data be fitted exactly, explain your reason?
[sol]**:** The arrangement of $\mathbf{T}$ and $\mathbf{X}$ depends on how you order the data. Here is just one example:

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 2 & 10 \\ 1 & 5 & 8 \\ 1 & 4 & 9 \\ 1 & 2 & 5 \\ 1 & 1 & 2 \\ 1 & 8 & 4 \\ 1 & 7 & 5 \\ 1 & 6 & 4 \end{bmatrix}.$$

This is the "tall and thin" case. If $\mathbf{X}$ is full-rank, we have $\mathbf{W} = \mathbf{X}^{\dagger}\mathbf{T}, \quad \mathbf{X}^{\dagger} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. The solution $\mathbf{W}$ is unique but the data would not be fitted exactly since the $N(\mathbf{A})$ is non-trivial and only the part of data in the $C(\mathbf{A})$ can be fitted.

9. Consider a linear regression problem in which the annual salary Y of a data scientist in the US is modeled by a fifth-degree polynomial of the scaled age X (in years).

$$Y = a + b_1X + b_2X^2 + b_3X^3 + b_4X^4 + b_5X^5.$$

Note that the age is separately mean-subtracted and scaled for each variable $X, X^2, \ldots, X^5$ using the standard deviation.

(a) The table below shows the estimate of the model parameters using the Tikhonov regularisation under 3 different values of the regularisation parameters $\alpha^2 = 0$, $\alpha^2 = 0.02$ and $\alpha^2 = 0.1$ Match the values of α^2 to the parameter estimates in each row of the table.

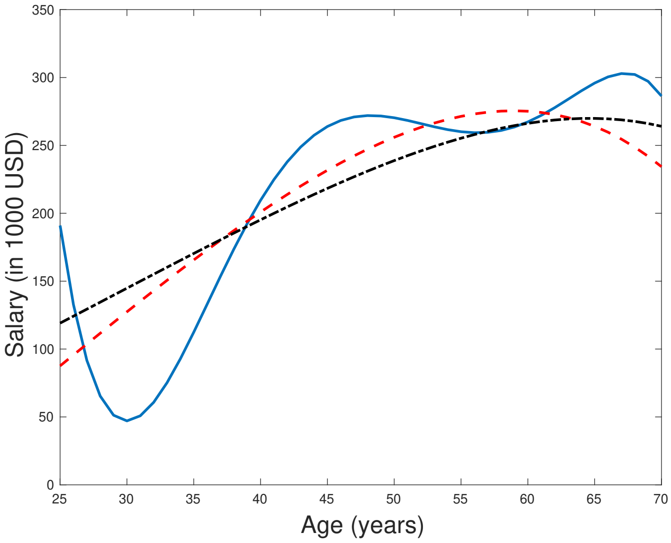
Justify your answers without using any computation.

α^2	a	b_1	b_2	b_3	b_4	b_5
?	216.5	97.8	36.6	-8.5	-35.0	-44.6
?	216.5	56.5	28.1	3.7	-15.1	-28.4
?	216.5	$32,622.6$	$13,5402.7$	$-215,493.1$	$-155,314.6$	$-42,558.8$

[sol]: From top to bottom, $\alpha^2 = 0.02, 0.1, 0$, respectively. The Tikhonov regularisation penalises the large parameter values via α^2; the greater the value of α^2, the more penalty added to the cost function for a minimisation problem. Therefore, it is expected that the greater the values of model parameters, the smaller the values of α^2.

(b) The predictions of salary (in a unit of 1000 USD) for the above values of α^2 are shown. Match the predicted curve in the plot with the value of α^2. Justify your answers.

[sol]: The greater the value of α^2 the smoother the predicted curve. In this case, $\alpha^2 = 0.02$ looks more reasonable. The curve is too irregular with no regularisation ($\alpha^2 = 0$) but become oversmooth for $\alpha^2 = 0.1$.



10. Consider a quotient $\frac{3x_1^2+2x_1x_2+3x_2^2}{x_1^2+x_2^2}$. For $\mathbf{x} = [x_1, x_2]^T$, write this quotient in the form $\frac{\mathbf{x}^T\mathbf{S}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$, i.e., finding the matrix $\mathbf{S}$.

[sol]: $\mathbf{S} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$ The eigenvalues of $\mathbf{S}$ are 2 and 4. Hence, the maximum of this quotient is 4.

11. Consider a maximization problem

$$\max_{\mathbf{x}} \ \mathbf{x}^T\mathbf{A}\mathbf{x} \quad \text{subject to } \mathbf{x}^T\mathbf{S}\mathbf{x} = C,$$

where both $\mathbf{A}$ and $\mathbf{S}$ are symmetric and positive-semidefinite. Show that the optimal solution satisfies $\mathbf{A}\mathbf{x} = \lambda_{\max}\mathbf{B}\mathbf{x}$ where $\lambda_{\max}$ is the largest number to satisfy this equation (i.e. the largest generalized eigenvalue). Hint: Use the Lagrange multiplier method.
[sol]**:** Define the Lagrangian $L(\mathbf{x}, \lambda) := \mathbf{x}^T\mathbf{A}\mathbf{x} - \lambda(\mathbf{x}^T\mathbf{S}\mathbf{x} - C)$. The derivative wrt $\mathbf{x}$ gives $2(\mathbf{A} - \lambda\mathbf{S})\mathbf{x} = 0$; hence $\mathbf{A}\mathbf{x} = \lambda\mathbf{S}\mathbf{x}$. This implies $\mathbf{x}^T\mathbf{A}\mathbf{x} = \lambda\mathbf{x}^T\mathbf{S}\mathbf{x} = \lambda C$. Thus, to maximize $\mathbf{x}^T\mathbf{A}\mathbf{x}$, we need to choose the vector $\mathbf{x}$ corresponding to the largest value of λ that satisfies $\mathbf{A}\mathbf{x} = \lambda\mathbf{S}\mathbf{x}$.

12. A contact-lens factory measures the qualities of its product in term of curvature and diameter. The quality control result is "Passed" or "Not Passed". A sample of 7 contact lens is randomly chosen and the results are listed below: Find the direction of most separation between "Passed" and "Not Passed" using Linear Discriminant Analysis (LDA). Then classify if a new item with 7 mm curvature and 11 mm diameter would pass the quality control.

Curvature (mm)	Diameter (mm)	Result
8	12	Passed
9	14	Passed
8	13	Passed
8	12	Passed
11	14	Not Passed
9	15	Not Passed
10	10	Not Passed

Table 2:

[sol]**:** The means for the two classes (1 for "Passed" and 2 for "Not Passed") are

$$\mu_1 = \begin{bmatrix} 8.25 \\ 12.75 \end{bmatrix} \quad \mu_2 = \begin{bmatrix} 10 \\ 13 \end{bmatrix} \quad \mu_1 - \mu_2 = \begin{bmatrix} -1.75 \\ -0.25 \end{bmatrix}$$

$$\mathbf{B} = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T = \begin{bmatrix} 3.0625 & 0.4375 \\ 0.4375 & 0.0625 \end{bmatrix}$$

The mean-subtracted data within each class is given by

$$\mathbf{X}_1 = \begin{bmatrix} -0.25 & -0.75 \\ 0.75 & 1.25 \\ -0.25 & 0.25 \\ -0.25 & -0.75 \end{bmatrix} \quad \mathbf{X}_2 = \begin{bmatrix} 1 & 1 \\ -1 & 2 \\ 0 & -3 \end{bmatrix}$$

$$\mathbf{S}_1 = \mathbf{X}_1^T\mathbf{X}_1 = \begin{bmatrix} 0.75 & 1.25 \\ 1.25 & 2.75 \end{bmatrix} \quad \mathbf{S}_2 = \mathbf{X}_2^T\mathbf{X}_2 = \begin{bmatrix} 2 & -1 \\ -1 & 14 \end{bmatrix}$$

$$\mathbf{S}^{-1} = (\mathbf{S}_1 + \mathbf{S}_2)^{-1} = \begin{bmatrix} 2.75 & 0.25 \\ 0.25 & 16.75 \end{bmatrix}^{-1} = \begin{bmatrix} 0.3641 & -0.0054 \\ -0.0054 & 0.0598 \end{bmatrix}$$

Therefore, we have

$$\mathbf{S}^{-1}\mathbf{B} = \begin{bmatrix} 1.1128 & 0.1590 \\ 0.0095 & 0.0014 \end{bmatrix} \to \lambda_1 = 1.1141 \quad \mathbf{w} = \begin{bmatrix} 1 \\ 0.0085 \end{bmatrix}$$

Note that $\mathbf{w}$ can also be computed from

$$\mathbf{w}^* = \mathbf{S}^{-1}(\mu_1 - \mu_2) = [-0.6359 \;\; -0.0054]^T$$

and then normalized to $\mathbf{w}^*/\|\mathbf{w}^*\| = \mathbf{w}$. To classify a new point $\mathbf{x} = [7 \;\; 11]^T$, we project it onto a unit vector $\mathbf{w}$, which gives $\mathbf{p} = (\mathbf{w}^T\mathbf{x})\mathbf{w} = [7.0935 \;\; 0.0606]^T$. Similarly, we project μ_1 and μ_2 onto $\mathbf{w}$, called μ_1' and μ_2' respectively. Check that $\|\mu_1' - \mathbf{p}\| = 1.27$ and $\|\mu_2' - \mathbf{p}\| = 3.02$; hence this item is "Passed".

13. Consider an undirected, non-weighted network with n (non-isolated) nodes and m links.

   (a) Let $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$ be the eigenvalues of the weighted matrix $\mathbf{W}$ of the network. Let $k_{\max}$ be the maximum degree of the nodes in the network and $\bar{k}$ be the average degree of the nodes. Show that

$$\bar{k} \leq \mu_1 \leq k_{\max}$$

[sol]**:** By using the optimality of the Rayleigh quotient, we have

$$\mu_1 = \max_{\mathbf{x}} \frac{\mathbf{x}^T\mathbf{W}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \geq \frac{\mathbf{1}^T\mathbf{W}\mathbf{1}}{\mathbf{1}^T\mathbf{1}} = \frac{\sum_i \sum_j w_{ij}}{n} = \frac{\sum_i d_i}{n} = \bar{k}$$

Let ϕ_1 be the eigenvector associated with μ_1. Let k be the node on which the value of $\phi_1(k)$ is maximum, i.e., $\phi_1(k) \geq \phi_1(i)$ for all other i. Then evaluate the k-th element of $\mathbf{W}\mathbf{x} = \lambda_{\max}\mathbf{x}$,

$$\mu_1 = \frac{(\mathbf{W}\phi_1)(k)}{\phi_1(k)} = \frac{\sum_{i\sim k}\phi_1(i)}{\phi_1(k)} = \sum_{i\sim k}\frac{\phi_1(i)}{\phi_1(k)} \leq \sum_{i\sim k} 1 \leq k_{\max}.$$

(b) (Optional) Let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ be the eigenvalues of the Laplacian matrix of this network. Show that

$$\lambda_n \leq \max\{d_i + d_j - c(i,j) : 1 \leq i < j \leq n, i \sim j\},$$

where $i \sim j$ means the existence of the link between i and j and $c(i,j)$ is the number of nodes that are connected to both i and j. [sol]**:** Let $\lambda_n \mathbf{x} = \mathbf{L}\mathbf{x}$. Choose the node i such that $x_i \geq x_j$ for all other $j \neq i$. For i chosen as above, choose node k such that $x_k := \min\{x_s : s \sim i\}$. Note that $x_k \leq x_i$. Then we have (just try to evaluate the i-th element of $\mathbf{L}\mathbf{x}$)

$$\lambda_n x_i = d_i x_i - \sum_{j:j\sim i} x_j = d_i x_i - \sum_{j:j\sim i \& j\sim k} x_j - \sum_{j:j\sim i \& j\nsim k} x_j.$$

We also have that

$$\lambda_n x_k = d_k x_k - \sum_{j:j\sim k} x_j = d_k x_k - \sum_{j:j\sim k \& j\sim i} x_j - \sum_{j:j\sim k \& j\nsim i} x_j.$$

It follows that

$$\begin{aligned}\lambda_n(x_i - x_k) &= d_i x_i - d_k x_k - \sum_{j:j\sim i \& j\nsim k} x_j + \sum_{j:j\sim k \& j\nsim i} x_j \\ &\leq d_i x_i - d_k x_k - (d_i - c(i,k))x_k + (d_k - c(i,k))x_i \\ &= (d_i + d_k - c(i,k))(x_i - x_k)\end{aligned}$$

The above inequality implies the desired result.

14. Consider a network consisting of K identical clusters (i.e. all clusters have n nodes and m links).

(a) If there is no links between these K clusters, find the ratio cut of these K clusters
[sol]**:** $J(\mathcal{C}) = \sum_{i=1}^{K} \frac{W(C_i, C_i')}{|C_i|} = 0$ since $W(C_i, C_i') = 0$ (i.e. no links between clusters).

(b) Suppose that exactly one link exists between each pair of clusters, find the ratio cut of these K clusters
[sol]**:** $J(\mathcal{C}) = \sum_{i=1}^{K} \frac{W(C_i, C_i')}{|C_i|} = \sum_{i=1}^{k} \frac{1}{|C_i|} = K/m$.

15. Let $\mathbf{D}$ be the degree matrix and $\mathbf{W}$ be the weighted adjacency matrix of a network with n nodes. The Laplacian matrix is $\mathbf{L} = \mathbf{D} - \mathbf{W}$. Defined the normalized Laplacian matrix by

$$\mathbf{L}_s = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}.$$

The above definition assumes that all node degrees are greater than zero.

(a) Show that $\mathbf{L}_s$ is positive semi-definite
[sol]**:** For any $\mathbf{x}$, let $\mathbf{z} = \mathbf{D}^{-1/2}\mathbf{x}$, so $\mathbf{z}^T = \mathbf{x}^T\mathbf{D}^{-1/2}$. Thus, $\mathbf{x}^T\mathbf{L}_s\mathbf{x} = \mathbf{z}^T\mathbf{L}\mathbf{z} > 0$ since $\mathbf{L}$ is positive semi-definite.

(b) Find the smallest eigenvalue and its corresponding eigenvector of $\mathbf{L}_s$
[sol]**:** Check that $\mathbf{L}_s$ is given as

$$\begin{aligned}\mathbf{L}^s &= \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}\\ &= \begin{pmatrix} \frac{\sum_{j\neq 1} w_{1j}}{\sqrt{d_1 d_1}} & -\frac{w_{12}}{\sqrt{d_1 d_2}} & \cdots - & \frac{w_{1n}}{\sqrt{d_1 d_n}} \\ -\frac{w_{21}}{\sqrt{d_2 d_1}} & \frac{\sum_{j\neq 2}{}^{w_{2j}}}{\sqrt{d_{2 d_2}}} & \cdots - & \frac{w_{2n}}{\sqrt{d_2 d_n}} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{w_{n1}}{\sqrt{d_n d_1}} & -\frac{w_{n2}}{\sqrt{d_n d_2}} & \cdots & \frac{\sum_{j\neq n} w_{nj}}{\sqrt{d_n d_n}} \end{pmatrix}.\end{aligned}$$

Let $L_s(i)$ be the i-th column of $\mathbf{L}_s$. Notice that $\sqrt{d_1}L_s(1) + \cdots + \sqrt{d_n}L_s(n) = \mathbf{0}$. So, the columns of $\mathbf{L}_s$ are not linearly independent. Thus, $\mathbf{L}_s$ has the rank at most $n-1$ and the smallest eigenvalue $\lambda_n = 0$. The above equation also suggests that $\mathbf{x}_n = \left[\sqrt{d_1}, \ldots, \sqrt{d_n}\right]^T$ is an eigenvector for $\lambda_n = 0$. Normalizing this vector to a unit vector gives the eigenvector $\frac{1}{\sqrt{\sum d_i}}\mathbf{D}^{-1/2}\mathbf{1}$.

(c) Let $\mathbf{c} \in \mathbb{R}^n$. Show that the minimum value of the following function

$$J(\mathbf{c}_1, \ldots, \mathbf{c}_k) := \sum_{i=1}^{k} \frac{\mathbf{c}_i^T\mathbf{L}\mathbf{c}_i}{\mathbf{c}_i^T\mathbf{D}\mathbf{c}_i}$$

is the sum of the k smallest eigenvalues of $\mathbf{L}_s$.
[sol]**:** Note that the problem $\sum_{i=1}^{k} \frac{\mathbf{c}_i^T\mathbf{L}\mathbf{c}_i}{\mathbf{c}_i^T\mathbf{D}\mathbf{c}_i}$ was solved in Week6 Lec1 and the solution is the sum of the k smallest eigenvalues of $\mathbf{L}$. We need to show that we can also use the eigenvalues of $\mathbf{L}_s$ instead.

$$\begin{aligned}\sum_{i=1}^{k} \frac{\mathbf{c}_i^T\mathbf{L}\mathbf{c}_i}{\mathbf{c}_i^T\mathbf{D}\mathbf{c}_i} &= \sum_{i=1}^{k} \frac{\mathbf{c}_i^T\left(\mathbf{D}^{1/2}\mathbf{D}^{-1/2}\right)\mathbf{L}\left(\mathbf{D}^{-1/2}\mathbf{D}^{1/2}\right)\mathbf{c}_i}{\mathbf{c}_i^T\left(\mathbf{D}^{1/2}\mathbf{D}^{1/2}\right)\mathbf{c}_i}\\ &= \sum_{i=1}^{k} \frac{\left(\mathbf{D}^{1/2}\mathbf{c}_i\right)^T\left(\mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}\right)\left(\mathbf{D}^{1/2}\mathbf{c}_i\right)}{\left(\mathbf{D}^{1/2}\mathbf{c}_i\right)^T\left(\mathbf{D}^{1/2}\mathbf{c}_i\right)}\\ &= \sum_{i=1}^{k} \left(\frac{\mathbf{\Delta}^{1/2}\mathbf{c}_i}{\|\mathbf{D}^{1/2}\mathbf{c}_i\|}\right)^T \mathbf{L}^s \left(\frac{\mathbf{D}^{1/2}\mathbf{c}_i}{\|\mathbf{D}^{1/2}\mathbf{c}_i\|}\right) = \sum_{i=1}^{k} \mathbf{u}_i^T\mathbf{L}_s\mathbf{u}_i\end{aligned}$$

The term $\sum_{i=1}^{k} \mathbf{u}_i^T \mathbf{L}_s \mathbf{u}_i$ is then minimized by the sum of the k smallest eigenvalues of $\mathbf{L}_s$.