Math 218D-1: Homework #14

due Friday, December 5, at 11:59pm

1. (Practicing a Procedure) Consider the following data matrix holding 6 samples of 3 measurements each:

$$A_0 = \begin{pmatrix} 2 & 7 & 1 & 0 & 5 & 3 \\ 1 & 9 & 6 & 3 & 2 & 3 \\ 0 & 0 & 4 & 3 & 1 & 4 \end{pmatrix}.$$

a) Subtract the means of the rows to find the centered data matrix.

b) Compute the covariance matrix S.

c) What are the variances s_1^2, s_2^2, s_3^2 of each of the three measurements? What is the total variance s^2?

d) Compute the variance $s(u)^2$ in the directions of the following unit vectors u:

$$\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \qquad \frac{1}{3}\begin{pmatrix} 1 \\ -2 \\ 2 \end{pmatrix} \qquad \frac{1}{\sqrt{3}}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

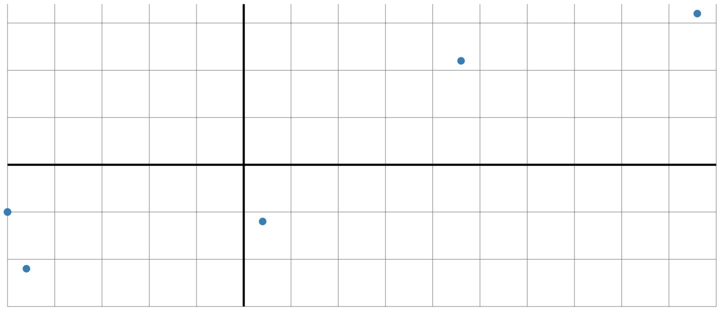
Check your answers using SymPy:

```
AO = Matrix([[2, 7, 1, 0, 5, 3],
             [1, 9, 6, 3, 2, 3],
             [0, 0, 4, 3, 1, 4]])
  # Compute the row averages.
  # Matrix.ones(m, n) is the m x n matrix with all 1's in it
averages = A * Matrix.ones(6, 1) / 6
  # This is the recentered data matrix.
A = AO - averages * Matrix.ones(1, 6)
  # This is the covariance matrix.
S = A * A.T / 5
  # Compute the variance in a direction.
u = Matrix([0, 1, 0])
su2 = u.T * S * u
```

2. **(Internalizing a Concept)** Consider the centered data matrix

$$A = \frac{1}{4}\begin{pmatrix} 23 & -25 & 2 & 48 & -25 & -23 \\ 11 & -5 & -6 & 16 & -5 & -11 \end{pmatrix}.$$

The columns are plotted as dots below.



- **a)** By looking at the plot, make a guess as to which direction maximizes variance, and draw that line.
- **b)** Compute the singular value decomposition of $\frac{1}{\sqrt{5}}A$ using SymPy:
$$\frac{1}{\sqrt{5}}A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T$$
Now plot Span$\{u_1\}$. How close did you come in part **a)**?
- **c)** What is the variance in the directions of u_1 and u_2? Compute the total variance without computing the covariance matrix.

3. Let A be a matrix with singular value decomposition

$$A = \sigma_1 u_1 v_1^T + \cdots + \sigma_r u_r v_r^T.$$

Show that A is a centered data matrix (the entries of each row sum to zero) if and only if the entries of each right singular vector v_i sum to zero.

[**Hint:** Multiply by the ones vector $\mathbf{1} = (1, 1, \ldots, 1)$.]

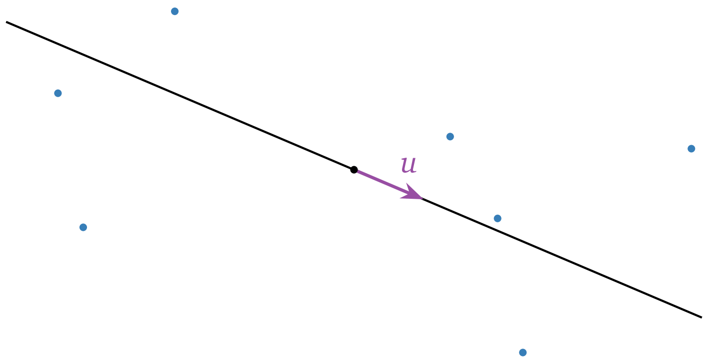
4. **(Internalizing a Concept)**

- **a)** Find an example of a nonzero 2×4 centered data matrix A and a unit vector u such that the variance in the u-direction of the data points is equal to zero.
- **b)** If A is an $m \times n$ centered data matrix, what has to be true about A in order for there to exist a unit vector u such that $s(u)^2 = 0$?

5. **(Internalizing a Concept)** Let A be an $m \times n$ centered data matrix, and let u_1 and v_1 be the first left and right singular vectors of A, respectively. Explain why the following quantities are equal:

$$||A^T u_1||^2 = ||A^T||^2 = ||A||^2 = ||Av_1||^2 = ||A^T A v_1|| = (n-1)s(u_1)^2.$$

6. **(Picture Problem)** The columns of a certain 2×7 centered data matrix A are plotted below, along with a unit vector u.



a) Draw the orthogonal projections of the columns of A onto $\mathrm{Span}\{u\}$.

b) Using a ruler to measure distances (keeping in mind that u is one unit long), compute $s(u)^2$.

7. **(Practicing a Procedure)** Consider the centered data matrix

$$A = \begin{pmatrix} 2 & 6 & -2 & -4 & -2 \\ 2 & 10 & -8 & -2 & -2 \\ 0 & 2 & 4 & -2 & -4 \\ 6 & 4 & -2 & -4 & -4 \end{pmatrix}$$

and the subspaces

$$V = \mathrm{Span}\left\{ \begin{pmatrix} 1 \\ 2 \\ -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 7 \\ 0 \\ 1 \\ 4 \end{pmatrix} \right\} \qquad W = \{(x_1, x_2, x_3, x_4) \colon x_1 + 2x_2 - x_3 - x_4 = 0\}.$$

a) Compute the measurement variances $s_1^2, s_2^2, s_3^2, s_4^2$.

b) Compute $s(\mathbf{R}^4)^2$. How is this value related to your answer to **a)**?

c) Compute $s(V)^2$ and $s(V^\perp)^2$. How are these values related to your answer to **b)**?

d) Compute $s(W)^2$ without doing elimination or Gram–Schmidt.

8. Find four centered *nonzero, distinct* data points $\overline{d}_1, \overline{d}_2, \overline{d}_3, \overline{d}_4$ in $\mathbf{R}^2$ that admit *infinitely many* best-fit lines in the sense of orthogonal least squares. (Recall that data points are *centered* if they sum to zero, i.e. if the mean of each coordinate is zero.)

9. (Practicing a Procedure) Consider the centered data points $\overline{d}_1, \ldots, \overline{d}_{10}$ in the following table:

$\overline{d}_1$	$\overline{d}_2$	$\overline{d}_3$	$\overline{d}_4$	$\overline{d}_5$	$\overline{d}_6$	$\overline{d}_7$	$\overline{d}_8$	$\overline{d}_9$	$\overline{d}_{10}$
1.6	−0.1	−3.8	3.6	3.8	−3.2	0.1	−2.6	0.5	0.1
−1.1	0.76	−0.74	0.96	−0.04	−0.74	0.16	0.76	−0.94	0.96
−1.7	0.66	0.66	−0.74	−1.3	1.2	−0.24	1.7	−0.74	0.66

Put the data into a matrix A in Sympy:

```
A = Matrix([[ 1.6,  -0.1,  -3.8,   3.6,   3.8,  -3.2,   0.1,  -2.6,   0.5,   0.1],
            [-1.1,  0.76, -0.74,  0.96, -0.04, -0.74,  0.16,  0.76, -0.94,  0.96],
            [-1.7,  0.66,  0.66, -0.74,  -1.3,   1.2, -0.24,   1.7, -0.74,  0.66]])
```

Compute the singular values and left singular vectors of A

```
  # The covariance matrix:
S = A*A.T / (10-1)
  # The eigenvalues, in order, and unit eigenvectors:
[(sigma3sq, u3), (sigma2sq, u2), (sigma1sq, u1)] = \
    sorted([(x[0], x[2][0].normalized())
            for x in S.eigenvects()])
```

In this problem, please write your answers to two decimal places.

a) What is the total variance s^2 of these data points? (Use `S.trace()`.)

b) Compute the variance of these data points along the following subspaces:

$$W_1 = \text{the line through } \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \qquad W_2 = \text{Span}\left\{ \begin{pmatrix} 1 \\ -1 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} \right\}$$

c) Find an orthonormal basis for the line of best fit V_1 and the plane of best fit V_2.

d) Compute the variances $s(V_1)^2$, $s(V_2)^2$ and errors² $s(V_1^\perp)^2$, $s(V_2^\perp)^2$. Verify that

$$s(V_1^\perp)^2 \geq s(V_2^\perp)^2 \qquad s(V_1)^2 \geq s(W_1)^2 \qquad s(V_2)^2 \geq s(W_2)^2,$$

and explain why these had to be true.

e) Do these data points best fit a line or a plane? Justify your answer.

f) Compute the rank-2 matrix A_2 that best approximates A. Now compute $\frac{1}{9}$ times the sum of the squares of the entries of $A - A_2$, and explain why this number has appeared earlier in the problem.

(An easy way to compute the sum of the squares of the entries of a matrix B is by taking the trace of $B^T B$. See HW13#10.)

10. An online movie-streaming service collects star ratings from its viewers and uses these to predict what movies you will like based on your previous ratings. The following are the ratings that ten (fictitious) people gave to three (fictitious) movies, on a scale of 0–10:

	Abe	Amy	Ann	Ben	Bob	Eve	Dan	Don	Ian	Meg
Prognosis Negative	7.8	6.1	2.4	9.8	10	3.0	6.3	3.6	6.7	6.3
Ponce De Leon	6.0	7.9	6.4	8.1	7.1	6.4	7.3	7.9	6.2	8.1
Lenore's Promise	5.8	8.2	8.2	6.8	6.2	8.7	7.3	9.2	6.8	8.2

Put the data into a matrix in SymPy:

```
A0 = Matrix([[7.8, 6.1, 2.4, 9.8,  10, 3.0, 6.3, 3.6, 6.7, 6.3],
             [6.0, 7.9, 6.4, 8.1, 7.1, 6.4, 7.3, 7.9, 6.2, 8.1],
             [5.8, 8.2, 8.2, 6.8, 6.2, 8.7, 7.3, 9.2, 6.8, 8.2]])
```

Find the row averages and subtract them:

```
  # Multiplying by (1,1,...,1) sums the rows
averages = A0 * Matrix.ones(10,1) / 10
A = A0 - averages * Matrix.ones(1, 10)
```

Now compute the covariance matrix:

```
S = A*A.T / (10-1)
pprint(S)
```

Let us compute the eigenvalues of S in order, and the corresponding unit eigenvectors:

```
[(sigma3sq, u3), (sigma2sq, u2), (sigma1sq, u1)] \
    = sorted([(x[0], x[2][0]) for x in S.eigenvects()])
  # Verify the sum is equal to the total variance
print(sigma1sq + sigma2sq + sigma3sq, S.trace())
  # Print the eigenvalues
print(sigma1sq, sigma2sq, sigma3sq)
  # Compute unit eigenvectors
pprint([u1.normalized(), u2.normalized(), u3.normalized()])
```

In this problem, please write your answers to two decimal places.

a) What is the variance in the number of stars given each of the three movies? What is the total variance? (Use `S.trace()`)

b) Which is the direction with the most variance? What is the variance in that direction?

c) Explain how these calculations tell you that ≈68% of the ratings are at a distance of $\sigma_3 \approx 0.18$ stars from the plane $\mathrm{Span}\{u_1, u_2\}$ (assuming the scores fit a normal distribution).

d) Use the fact that $\{u_1, u_2, u_3\}$ is orthonormal to find an implicit equation for $\mathrm{Span}\{u_1, u_2\}$ of the form $x_3 = a_1x_1 + a_2x_2$.

e) Suppose that Joe gave *Prognosis Negative* a rating of 8.5 and *Ponce De Leon* a rating of 6.2. How would you expect Joe to rate *Lenore's Promise*?

Remark: According to a New York Times Magazine article, this really is the idea behind Netflix's algorithm—which earned its creator a \$1 000 000 prize.