Homework #15

Not collected: this is practice for the final exam

1. Consider the data matrix

$$A_0 = \begin{pmatrix} 20 & 11 & -6 & 13 & 24 & -17 & 18 \\ -6 & 7 & 1 & -32 & -9 & 3 & 8 \end{pmatrix}.$$

a) Subtract the means of the rows of A_0 to obtain the centered matrix A.

b) Compute the covariance matrix $S = \frac{1}{7-1}AA^T$ and the total variance $s^2 = \mathrm{Tr}(S)$.

c) Compute the singular value decomposition of $\frac{1}{\sqrt{7-1}}A$ in outer product form:

$$\frac{1}{\sqrt{6}}A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T.$$

Verify that $s^2 = \sigma_1^2 + \sigma_2^2$.

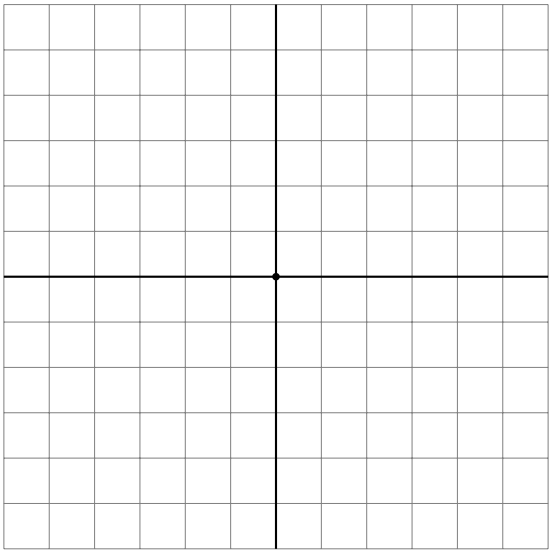
d) What is the direction of largest variance, and what is the variance in that direction?

e) What is the direction of smallest variance, and what is the variance in that direction?

f) What is the line of best fit to your *original* (non-centered) data points in the sense of orthogonal least squares? What is the error^2? Is this line a good fit for the data? Why or why not?

g) In the grid below, draw and label: **i)** the columns of A (as points), **ii)** the lines in the directions of largest and smallest variance, **iii)** the columns of $\sqrt{6}\,\sigma_1 u_1 v_1^T$ (as points), and **iv)** the columns of $\sqrt{6}\,\sigma_2 u_2 v_2^T$ (as points).

h) How is **ii)** related to **iii)** and **iv)**? How are **iii)** and **iv)** related to **i)**?



(Grid lines are 5 units apart.)

2. Consider the centered data points $d_1, \ldots, d_{10}$ in the following table:

d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9	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

Use SymPy (in the Sage cell on the webpage) or your favorite linear algebra calculator to put the data into a matrix A:

```
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.transpose() / (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()])
```

The following useful function computes orthogonal projections:

```
def project(B, v):
    """Compute the orthogonal projection of v onto Col(B)
       assuming B has full column rank"""
    return B*(B.transpose()*B).inv()*B.transpose()*v
```

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

a) What is the total variance s^2 of these data points?

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

i) $V_1 =$ the line through $\begin{pmatrix}1\\1\\1\end{pmatrix}$ **ii)** $V_2 =$ the plane $x_1 + x_2 + x_3 = 0$

iii) $V_3 = \text{Span}\left\{\begin{pmatrix}1\\-1\\3\end{pmatrix}, \begin{pmatrix}2\\1\\1\end{pmatrix}\right\}$ **iv)** $V_4 = \mathbf{R}^3$.

Explain why $s(V_1)^2 + s(V_2)^2 = s^2$ and $s(V_4)^2 = s^2$. You'll want to do something like this:

```
 # Compute the orthogonal projections of the columns of A
 # onto Span{(1,1,2),(1,-1,1)}:
l = [project(Matrix([[1,1],[1,-1],[2,1]]), A.col(i))
     for i in range(A.cols)]
 # Sum the squares of the lengths and normalize:
print(sum(v.dot(v) for v in l)/9)
```

c) Find the line L and plane P of best fit, compute the variances $s(L)^2$ and $s(P)^2$, and compute the errors[2] $s(L^\perp)^2$ and $s(P^\perp)^2$. Verify that $s(L)^2 \geq s(V_1)^2$, $s(P)^2 \geq s(V_2)^2$, and $s(P)^2 \geq s(V_3)^2$.

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

3. Find four centered *nonzero, distinct* data points d_1, d_2, d_3, 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 their mean is zero.)

4. 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 (columns 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 $\mathbf{1} = (1, 1, \ldots, 1)$.]

5. 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

Using SymPy (in the Sage cell on the webpage) or your favorite linear algebra calculator, put the data into a matrix:

```
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.transpose() / (10-1)
pprint(S)
```

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 entry of S tells you that people who liked *Prognosis Negative* generally did not like *Lenore's Promise*?

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()])
```

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

d) 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).

e) Use the fact that $\{u_1, u_2, u_3\}$ is orthonormal to find an implicit equation for $\text{Span}\{u_1, u_2\}$ of the form $x_3 = a_1 x_1 + a_2 x_2$.

f) 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.