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 \( S = \frac{1}{\sqrt{n-1}}AA^T \) and the total variance \( s^2 = \text{Tr}(S) \).

c) Compute the singular value decomposition of \( \frac{1}{\sqrt{n-1}}A \) 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? 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)?
Solution.

a) \[ A = \begin{pmatrix} 11 & 2 & -15 & 4 & 15 & -26 & 9 \\ -2 & 11 & 5 & -28 & -5 & 7 & 12 \end{pmatrix}. \]

b) \[ S = \begin{pmatrix} 674/3 & -56 \\ -56 & 192 \end{pmatrix} \quad s^2 = \frac{1250}{3}. \]

c) We have
\[ \sigma_1 = \frac{40}{\sqrt{6}} \quad u_1 = \frac{1}{5} \begin{pmatrix} -4 \\ 3 \end{pmatrix} \quad v_1^T = \frac{1}{8} \begin{pmatrix} -2 & 1 & 3 & -4 & -3 & 5 & 0 \end{pmatrix} \]
\[ \sigma_2 = \frac{30}{\sqrt{6}} \quad u_2 = \frac{1}{5} \begin{pmatrix} 3 \\ 4 \end{pmatrix} \quad v_2^T = \frac{1}{6} \begin{pmatrix} 1 & 2 & -1 & -4 & 1 & -2 & 3 \end{pmatrix}. \]

Note that \[ \sigma_1^2 + \sigma_2^2 = \frac{800}{3} + 150 = \frac{1250}{3} = s^2. \]

d) The direction of largest variance is \( u_1 = \frac{1}{5} \begin{pmatrix} -4 \\ 3 \end{pmatrix} \), and the variance in that direction is \( \sigma_1^2 = \frac{800}{3}. \)

e) The direction of smallest variance is \( u_2 = \frac{1}{5} \begin{pmatrix} 3 \\ 4 \end{pmatrix} \), and the variance in that direction is \( \sigma_2^2 = 150. \)

f) We have to add the means to get the line of best fit to the columns of \( A_0. \) The line of orthogonal best fit is \( \text{Span}\{u_1\} + \begin{pmatrix} 4 \\ 0 \end{pmatrix}; \) the error \( ^2 \) is \( \sigma_2^2 = 150. \) This is not a good fit: the error has the same order of magnitude as the variance along that line \( (800/3 \approx 267). \)

2. Consider the centered data points \( d_1, \ldots, d_{10} \) in the following table:
Use SymPy (in the Sage cell on the webpage) or your favorite linear algebra calculator to put the data into a matrix $A$:

$$A = \begin{bmatrix} 1.6 & -0.1 & -3.8 & 3.6 & 3.8 & -3.2 & 0.1 & -2.6 & 0.5 & 0.1 \\ -1.1 & 0.76 & -0.74 & 0.96 & -0.04 & -0.74 & 0.16 & 0.76 & -0.94 & 0.96 \\ -1.7 & 0.66 & 0.66 & -0.74 & -1.3 & 1.2 & -0.24 & 1.7 & -0.74 & 0.66 \end{bmatrix}$$

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) $V_1 = \text{the line through } \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$
- ii) $V_2 = \text{the plane } x_1 + x_2 + x_3 = 0$
- iii) $V_3 = \text{Span } \left\{ \begin{pmatrix} 1 \\ -1 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} \right\}$
- iv) $V_4 = \mathbb{R}^3$.

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)}:
1 = [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 1)/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 $s(L)^2$ and $s(P)^2$. 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.

### Solution.

a) The total variance is $s^2 = \text{Tr}(S) \approx 8.81$.

b) The variances are:
\[
\begin{align*}
  s(V_1)^2 &\approx 1.88 \\
  s(V_2)^2 &\approx 6.93 \\
  s(V_3)^2 &\approx 5.08 \\
  s(V_4)^2 &\approx 8.81.
\end{align*}
\]
We have $s(V_1)^2 + s(V_2)^2 = s^2$ because $V_2 = V_1^\perp$, and $s(V_4)^2 = s^2$ because $d_i = (d_i)_{\mathbb{R}^3}$ and the total variance is $s^2 = \frac{1}{p} \sum_{i=1}^{10} ||d_i||^2$.

c) The line of best fit is $L = \text{Span}\{u_1\}$. The variance is $s(L)^2 = \sigma_1^2 \approx 7.75$ and the error$^2$ is $s(L)^2 = \sigma_1^2 + \sigma_2^2 \approx 1.05$. The plane of best fit is $P = \text{Span}\{u_1, u_2\}$. The variance is $s(P)^2 = \sigma_1^2 + \sigma_2^2 \approx 8.77$ and the error$^2$ is $s(P)^2 = \sigma_3^2 \approx 0.037$.

d) These data fit the plane $P$ very closely with small error$^2$ because $\sigma_3$ is much smaller than $\sigma_1$ and $\sigma_2$.

3. Find four centered nonzero, distinct data points $d_1, d_2, d_3, d_4$ in $\mathbb{R}^2$ 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.)

### Solution.

This will happen exactly when the data matrix $A$ containing the data points has two identical singular values, so it is easy to produce such a matrix by writing its SVD. For instance,
\[
\begin{pmatrix}
  1 & -1 \\
  1 & 1
\end{pmatrix}
= 2 \begin{pmatrix}
  1 \\
  0
\end{pmatrix} \frac{1}{2} \begin{pmatrix}
  1 & -1 & 1 & -1
\end{pmatrix} + 2 \begin{pmatrix}
  0 \\
  1
\end{pmatrix} \frac{1}{2} \begin{pmatrix}
  1 & 1 & -1 & -1
\end{pmatrix},
\]
so the data points $(1, 1), (-1, -1), (1, -1), (-1, 1)$ work.

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 $1 = (1, 1, \ldots, 1)$.]

### Solution.

The sum of the columns of $A$ is $A1$. We have
\[
A1 = \sigma_1 (v_1 \cdot 1) u_1 + \cdots + \sigma_r (v_r \cdot 1) u_r.
\]
Since $\{u_1, \ldots, u_r\}$ is linearly independent, this is equal to zero if and only if each $v_i \cdot 1 = 0$. 

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:

<table>
<thead>
<tr>
<th>Prognosis Negative</th>
<th>Abe</th>
<th>Amy</th>
<th>Ann</th>
<th>Ben</th>
<th>Bob</th>
<th>Eve</th>
<th>Dan</th>
<th>Don</th>
<th>Ian</th>
<th>Meg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7.8</td>
<td>6.1</td>
<td>2.4</td>
<td>9.8</td>
<td>10</td>
<td>3.0</td>
<td>6.3</td>
<td>3.6</td>
<td>6.7</td>
<td>6.3</td>
</tr>
<tr>
<td>Ponce De Leon</td>
<td>6.0</td>
<td>7.9</td>
<td>6.4</td>
<td>8.1</td>
<td>7.1</td>
<td>6.4</td>
<td>7.3</td>
<td>7.9</td>
<td>6.2</td>
<td>8.1</td>
</tr>
<tr>
<td>Lenore’s Promise</td>
<td>5.8</td>
<td>8.2</td>
<td>8.2</td>
<td>6.8</td>
<td>6.2</td>
<td>8.7</td>
<td>7.3</td>
<td>9.2</td>
<td>6.8</td>
<td>8.2</td>
</tr>
</tbody>
</table>

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

\[
A_0 = \text{Matrix}([[7.8, 6.1, 2.4, 9.8, 10, 3.0, 6.3, 3.6, 6.7, 6.3],
[6.0, 7.9, 6.4, 8.1, 7.1, 6.4, 7.3, 7.9, 6.2, 8.1],
[5.8, 8.2, 8.2, 6.8, 6.2, 8.7, 7.3, 9.2, 6.8, 8.2]])
\]

Find the row averages and subtract them:

```python
# 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:

\[
S = A*A.transpose() / (10-1)
\]

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:

```python
[(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 \( \approx 68\% \) of the ratings are at a distance of \( \sigma_3 \approx 0.18 \) stars from the plane \( \text{Span}\{u_1, u_2\} \) (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 \emph{Prognosis Negative} a rating of 8.5 and \emph{Ponce De Leon} a rating of 6.2. How would you expect Joe to rate \emph{Lenore’s Promise}?

\textbf{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.

\textbf{Solution.}

You should have computed the following quantities:

\[
S \approx \begin{pmatrix}
6.85 & 0.47 & -2.39 \\
0.47 & 0.70 & 0.34 \\
-2.39 & 0.34 & 1.26
\end{pmatrix}, \quad \sigma_1^2 \approx 7.75
\]

\[
\sigma_2^2 \approx 1.04
\]

\[
\sigma_3^2 \approx 0.03
\]

\[
u_1 \approx \begin{pmatrix}
-0.94 \\
-0.05 \\
0.34
\end{pmatrix}, \quad u_2 \approx \begin{pmatrix}
0.17 \\
0.80 \\
0.57
\end{pmatrix}, \quad u_3 \approx \begin{pmatrix}
0.30 \\
-0.59 \\
0.75
\end{pmatrix}.
\]

\textbf{a) The variances of the movie ratings are the diagonal entries of } \( S \):

\[
s_1^2 \approx 6.85 \quad s_2^2 \approx 0.70 \quad s_3^2 \approx 1.26.
\]

The total variance is \( \text{Tr}(S) \approx 8.82 \).

\textbf{b) The } \((1,3)\)-entry of } \( S \) \text{ is the covariance of the first row with the third; in this case, it is } \approx -2.39 \text{, which means that } \emph{Prognosis Negative} \text{ and } \emph{Lenore’s Promise} \text{ are negatively correlated.}

\textbf{c) The direction with the most variance is } \( u_1 \), \text{ which has a variance of } \sigma_1^2.

\textbf{d) The variance in the } \( u_3 \)-direction is } \sigma_3^2 \text{. This means that the standard deviation of } \{(\text{columns of } A) \cdot u_3\} \text{ is } \sigma_3 \approx 0.18 \text{, and } 68\% \text{ of values lie within one standard deviation of the mean.}

\textbf{e) The plane spanned by } \{u_1, u_2\} \text{ is the orthogonal complement of } \{u_3\}. \text{ Therefore it is defined by } 0.30x_1 - 0.59x_2 + 0.75x_3 = 0 \text{, or }

\[x_3 = -0.40x_1 + 0.79x_2.\]

\textbf{f) After subtracting the average scores, Joe’s first two scores are } x_1 = 2.30 \text{ and } x_2 = -0.94 \text{. Substituting into the above equation gives } x_3 \approx -1.67 \text{; adding back the average third score translates into } \approx 5.9 \text{ stars.}