CS5100: Data Analysis

Resite

Chapter 9: Unsupervised Learning
Exploratory Data Analysis (EDA)

- Unsupervised learning
  - works with datasets that do not have labels.
  - Its main target is to understand the structure of the data.
- The main target of EDA
  - is to summarize the main characteristics of the data
  - It helps to identify the hidden pattern in the data.
  - More concretely, it is usefull for performing dimension reduction.
  - This makes visualization of high-dimensional data feasible.
    e.g. Principal Component Analysis (PCA)

Motivating Examples of EDA

- A dataset has observation $x_1, x_2, \ldots, x_N$, each is a vector in $\mathbb{R}^d$.
- In our example,
  - the dimension $d \leq 4$,
  - but keep in mind that
  - In many real-world dataset, d is very large (say $d \geq 100$, or even $d \geq 1000$).

- Quite often, only a few dimensions/attributes are
  - "relevant" to the data analysis problem you want to solve.
  - The other dimensions are either
    - "redundant", "irrelevant" or "boring".

- We are interested in a systematic way
  - to extract the relevant dimensions.

- Some concrete examples of redundant/irrelevant/boring dimensions:
  - contains both the ages & the birth years of people.
  - contains their father's weights.

- While insights/intuitions tell you these particular dimensions are redundant/irrelevant/boring,
  - but when $d \geq 1000$, you don't really want to look onto each attribute & judge yourself...

Redundancy

- When two or more attributes are strongly linearly correlated
- In mathematical language, the observations lie in an affine subspace of the whole space.
- In more general terminology, they lie on a lower dimension manifold.

Irrelevance

- In mathematical language, the irrelevant attributes can be viewed as random noises (or perturbations).

Boredom

- most of it values in different observations concentrate around the mean.

Example:

- We prefer attributes which show a wider range of value, since the observations can be distinguished easily
- (even under the influence of noises), or they form well-separated clusters.
- To quantify the range, as you should expect, a common measure is Variance.

Recaps of Linear Algebra

Dot Product, Norm and Unit Vector

- Given two vectors in $\mathbb{R}^d$

$$V = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_d \end{bmatrix} \qquad U = \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_d \end{bmatrix}$$

their dot product is

$$\langle V, U \rangle = V_1U_1 + V_2U_2 + \dots + V_dU_d$$

- The norm of a vector V is

$$\|V\| = \sqrt{\langle V, V \rangle}$$

Example:

(0,0) ——— (3,4)

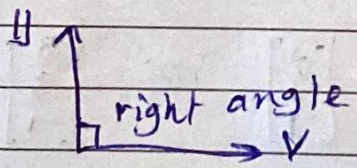
norm = length.

$$= \sqrt{(3-0)^2 + (4-0)^2} = \sqrt{3^2+4^2} = \sqrt{9+16} = \sqrt{25} = 5$$

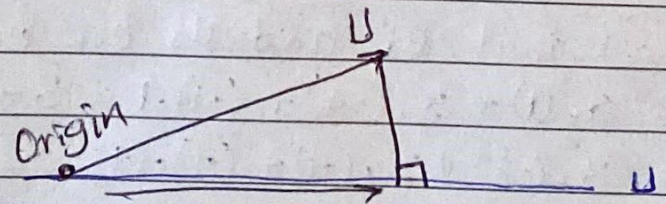
- We say V is a unit vector if $\|V\| = 1$

Orthogonality and Projection

- Two vectors $V, U \in \mathbb{R}^d$ are orthogonal if $\langle V, U \rangle = 0$
  - If their dot product is zero!
  - orthogonal = perpendicular
  - Math: orthogonal Data: uncorrelated.



Projection



projection of V on the subspace spaned by U.

Projection

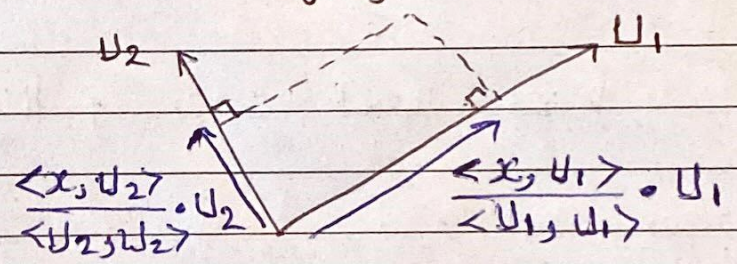
- Suppose $u_1, u_2, \dots, u_k \in \mathbb{R}^d$ are non-zero vectors
  - such that every pair of them are orthogonal.
  - their dot product is zero.

Then for any vector x, its projection on the subspace spanned by $u_1, u_2, \dots, u_k$ is

$$\frac{\langle x, u_1\rangle}{\langle u_1, u_1\rangle}\cdot u_1 + \frac{\langle x, u_2\rangle}{\langle u_2, u_2\rangle}\cdot u_2 + \dots + \frac{\langle x, u_k\rangle}{\langle u_k, u_k\rangle}\cdot u_k$$

($\langle x, u_1\rangle$: dot product of x and u_1; $\langle u_1, u_1\rangle$: dot product of u_1 and u_1)

In this lecture, we call $\frac{\langle x, u_j\rangle}{\langle u_j, u_j\rangle}$ a coefficient.



Example 1 If $x = \begin{bmatrix} 8 \\ 3 \\ -1 \end{bmatrix}$, what is its projection on the subspace spanned by $\begin{bmatrix} 1 \\ -4 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} -3 \\ -1 \\ -1 \end{bmatrix}$?

Solution: $u_1 = \begin{bmatrix} 1 \\ -4 \\ 1 \end{bmatrix}$ $u_2 = \begin{bmatrix} -3 \\ -1 \\ -1 \end{bmatrix}$ $x = \begin{bmatrix} 8 \\ 3 \\ -1 \end{bmatrix}$

$\langle u_1, u_2\rangle = 1\cdot(-3) + (-4)\cdot(-1) + 1\cdot(-1) = -3 + 4 - 1 = 0$

- First, check that u_1 and u_2 are orthogonal, dot product is zero

$\langle x, u_1\rangle = 8\cdot(1) + 3\cdot(-4) + (-1)\cdot 1 = 8 - 12 - 1 = -5$

$\langle x, u_2\rangle = 8\cdot(-3) + 3\cdot(-1) + (-1)\cdot(-1) = -24 - 3 + 1 = -26$

$\langle u_1, u_1\rangle = 1\cdot 1 + (-4)\cdot(-4) + 1\cdot 1 = 1 + 16 + 1 = 18$

$\langle u_2, u_2\rangle = (-3)\cdot(-3) + (-1)\cdot(-1) + (-1)\cdot(-1) = 9 + 1 + 1 = 11$

$$\text{Projection} = \frac{\langle x, u_1\rangle}{\langle u_1, u_1\rangle}\cdot u_1 + \frac{\langle x, u_2\rangle}{\langle u_2, u_2\rangle}\cdot u_2$$

$$= \frac{-5}{18}\cdot\begin{bmatrix} 1 \\ -4 \\ 1 \end{bmatrix} + \frac{-26}{11}\cdot\begin{bmatrix} -3 \\ -1 \\ -1 \end{bmatrix} = \begin{bmatrix} -5/18 \\ 20/18 \\ -5/18 \end{bmatrix} + \begin{bmatrix} 78/11 \\ 26/11 \\ 26/11 \end{bmatrix} = \begin{bmatrix} 6.8131 \\ 3.4747 \\ 2.0859 \end{bmatrix}$$

Example ② If $x=\begin{bmatrix}8\\3\\-1\end{bmatrix}$, what is its projection on the subspace spanned by $\begin{bmatrix}1\\-4\\1\end{bmatrix}$ and $\begin{bmatrix}-3\\1\\-1\end{bmatrix}$?

Solution: let $u_1=\begin{bmatrix}1\\-4\\1\end{bmatrix}$ and $u_2=\begin{bmatrix}-3\\1\\-1\end{bmatrix}$

First, check that u_1 and u_2 are orthogonal.
- their dot product is zero

$\therefore \langle u_1, u_2\rangle = 1\cdot(-3)+(-4)\cdot 1+1\cdot(-1)=-3-4-1=-8$

so, no orthogonal.

Quiz ①: If $x=\begin{bmatrix}9\\-3\\-4\\7\end{bmatrix}$, what is its projection on the subspace spanned by $\begin{bmatrix}1\\-1\\2\\4\end{bmatrix}$ and $\begin{bmatrix}2\\4\\-1\\1\end{bmatrix}$?

Solution: let $u_1=\begin{bmatrix}1\\-1\\2\\4\end{bmatrix}$ and $u_2=\begin{bmatrix}2\\4\\-1\\1\end{bmatrix}$

First, check that u_1 and u_2 are orthogonal.

$\langle u_1, u_2\rangle = 1\cdot 2+(-1)\cdot 4+2\cdot(-1)+4\cdot 1=2-4-2+4=0$

∴ Since their dot product is zero, u_1 and u_2 are orthogonal.

Then, we compute: $\frac{\langle x, u_1\rangle}{\langle u_1, u_1\rangle}=$ and $\frac{\langle x, u_2\rangle}{\langle u_2, u_2\rangle}$

$\langle x, u_1\rangle = 9\cdot 1+(-3)\cdot(-1)+(-4)\cdot 2+7\cdot 4=9+3-8+28=32$

$\langle x, u_2\rangle = 9\cdot 2+(-3)\cdot 4+(-4)\cdot(-1)+7\cdot 1=18-12+4+7=17$

$\langle u_1, u_1\rangle = 1\cdot 1+(-1)\cdot(-1)+2\cdot 2+4\cdot 4=1+1+4+16=22$

$\langle u_2, u_2\rangle = 2\cdot 2+4\cdot 4+(-1)\cdot(-1)+1\cdot 1=4+16+1+1=22$

$$\text{Projection}=\frac{\langle x, u_1\rangle}{\langle u_1, u_1\rangle}\cdot u_1+\frac{\langle x, u_2\rangle}{\langle u_2, u_2\rangle}\cdot u_2=\frac{32}{22}\cdot\begin{bmatrix}1\\-1\\2\\4\end{bmatrix}+\frac{17}{22}\cdot\begin{bmatrix}2\\4\\-1\\1\end{bmatrix}$$

$$=\begin{bmatrix}32/22\\-32/22\\32/11\\32/5.5\end{bmatrix}+\begin{bmatrix}17/11\\17/5.5\\-17/22\\17/22\end{bmatrix}=\begin{bmatrix}3\\36/22\\47/22\\145/22\end{bmatrix}$$

Projection and PCA

- As we have seen from the previous examples,
  - The projection of x can be far from x.
  - But when $u_1, u_2, \ldots, u_K$ are chosen appropriately,
  - the projection can be pretty close to the value of x.

- In PCA, the target is to compute $u_1, u_2, \ldots, u_K$ (for some $K \ll d$)
  - such that the projection of every observation x_i
  - is close to the value of x_i
  - The target is not always achievable.
  - But if the data has hidden patterns, this might be possible.

- When $K \ll d$,
  - x_i can be represented approximately
  - by the following vector in $\mathbb{R}^K$
  - after specifying $u_1, u_2, \ldots, u_K$,
  - thus achieving dimension reduction:

$$\left(\frac{\langle x_i, u_1 \rangle}{\langle u_1, u_1 \rangle}, \frac{\langle x_i, u_2 \rangle}{\langle u_2, u_2 \rangle}, \ldots, \frac{\langle x_i, u_K \rangle}{\langle u_K, u_K \rangle} \right)'$$

Transpose and Symmetric Matrix

- The transpose of a $d_1 \times d_2$ matrix A
  - is a $d_2 \times d_1$ matrix denoted by A'
  - such that for any $1 \le j \le d_1$ and $1 \le k \le d_2$, $A_{jk} = A'_{kj}$

Example: $[1\ 2\ 3]' = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}' = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$

- A $d \times d$ square matrix A is symmetric
  - If and only if $A' = A$

Example:

$$\begin{bmatrix} 3 & -2 & -6 & 7 \\ -2 & 11 & 0 & -1 \\ -6 & 0 & -4 & 10 \\ 7 & -1 & 10 & 5 \end{bmatrix}$$

Eigenvectors and Eigenvalues

- Given a $d \times d$ square matrix A
  - a non-zero vector $u \in \mathbb{R}^d$ is an eigenvector of A
  - if there exists real number λ such that

$$A \cdot u = \lambda \cdot u$$

  - λ is called the eigenvalue of the eigenvector.

Example: $\begin{bmatrix} 1 & -3 & 1 \\ 0 & 1 & -2 \\ 2 & 2 & -4 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} -2 \\ -4 \\ -6 \end{bmatrix} = (-2) \cdot \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$

Thus $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ is an eigenvector of the 3×3 square matrix, with eigenvalue -2.

Quiz ② Which of the following vector is an eigenvector of the matrix $\begin{bmatrix} 1 & -3 & 1 \\ -3 & -1 & 2 \\ 1 & 2 & 3 \end{bmatrix}$?

$\det(A - \lambda I) = 0$?

$$\det \begin{pmatrix} 1-\lambda & -3 & 1 \\ -3 & -1-\lambda & 2 \\ 1 & 2 & 3-\lambda \end{pmatrix} = 0$$

$$\lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix}$$

$$\begin{bmatrix} 1 & -3 & 1 \\ -3 & -1 & 2 \\ 1 & 2 & 3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix}$$

$$= \begin{bmatrix} 1-\lambda & -3 & 1 \\ -3 & -1-\lambda & 2 \\ 1 & 2 & 3-\lambda \end{bmatrix}$$

$$(1-\lambda)[(-1-\lambda)(3-\lambda) - (2)(2)] - (-3)[(-3)\cdot(3-\lambda) - 2 \times 1] + 1[(-3)\cdot 2 - (-1-\lambda)1] = 0$$

$$(1-\lambda)[$$

$$\begin{bmatrix} 1 & -3 & 1 \\ -3 & -1 & 2 \\ 1 & 2 & 3 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2+3+1 \\ 6+1+2 \\ 2-2+3 \end{bmatrix} = \begin{bmatrix} 6 \\ -3 \\ 3 \end{bmatrix} = 3 \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}$$

eigenvector

eigenvalue.

Special Theorem.

1. Any $d \times d$ symmetric matrix has d eigenvectors
   - which are pairwise orthogonal.

2. Each eigenvector corresponds to an eigenvalue
   - which is a real number

- In PCA,
  - we will use the observations to construct a
  - symmetric matrix.
  - Then we select K of the eigenvectors as $u_1, u_2, \ldots, u_K$ for projection

When $d \geq 4$, it is difficult to compute eigenvectors by hand;

Theory of Principal Component Analysis

Pre-processing: Centering

- Before carrying out PCA,
  - we need to do a centering step
  - to ensure that centroid of $\{x_1, x_2, \ldots, x_N\}$
  - is the zero vector:

1. Compute the centroid of $\{x_1, x_2, \ldots, x_N\}$.
   - Denote the centroid by $\bar{x}$.
2. Replace each x_i by $x_i - \bar{x}$.

Example: Suppose there are 5 observations:

$$x_1 = \begin{bmatrix} -3 \\ 1 \end{bmatrix}, \quad x_2 = \begin{bmatrix} -6 \\ -2 \end{bmatrix}, \quad x_3 = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \quad x_4 = \begin{bmatrix} 0 \\ -7 \end{bmatrix}, \quad x_5 = \begin{bmatrix} 1 \\ -6 \end{bmatrix}$$

Their centroid is

$$\bar{x} = \frac{1}{5}(x_1 + x_2 + x_3 + x_4 + x_5)$$

$$\bar{x} = \frac{1}{5}\begin{bmatrix} -3-6+3+0+1 \\ 1-2+4-7-6 \end{bmatrix} = \frac{1}{5}\begin{bmatrix} -5 \\ -10 \end{bmatrix} = \begin{bmatrix} -1 \\ -2 \end{bmatrix}$$

Then we replace each observation x_i by $x_i - \bar{x}$

$$x_1 = \begin{bmatrix} -3 \\ 1 \end{bmatrix} - \begin{bmatrix} -1 \\ -2 \end{bmatrix} = \begin{bmatrix} -2 \\ 3 \end{bmatrix} \quad x_2 = \begin{bmatrix} -6 \\ -2 \end{bmatrix} - \begin{bmatrix} -1 \\ -2 \end{bmatrix} = \begin{bmatrix} -5 \\ 0 \end{bmatrix}$$

$$x_3 = \begin{bmatrix} 3 \\ 4 \end{bmatrix} - \begin{bmatrix} -1 \\ -2 \end{bmatrix} = \begin{bmatrix} 4 \\ 6 \end{bmatrix} \quad x_4 = \begin{bmatrix} 0 \\ -7 \end{bmatrix} - \begin{bmatrix} -1 \\ -2 \end{bmatrix} = \begin{bmatrix} 1 \\ -5 \end{bmatrix} \quad x_5 = \begin{bmatrix} 1 \\ -6 \end{bmatrix} - \begin{bmatrix} -1 \\ -2 \end{bmatrix} = \begin{bmatrix} 2 \\ -4 \end{bmatrix}$$

Pre-processing: Normalization

- When different dimensions exhibit significantly different ranges,
  - it is advisable to do normalization.

- In PCA,
  - a common way to do normalization is
  - after centering, for each dimension,
  - divide the values by the standard deviation (i.e. $\sqrt{\text{variance}}$).

Example: In the past page, after centering, the observations are

$$x_1=\begin{bmatrix}-2\\3\end{bmatrix},\ x_2=\begin{bmatrix}-5\\0\end{bmatrix},\ x_3=\begin{bmatrix}4\\6\end{bmatrix},\ x_4=\begin{bmatrix}1\\-5\end{bmatrix},\ x_5=\begin{bmatrix}2\\-4\end{bmatrix}$$

- The variance along the first dimension is

$$\frac{1}{5}\cdot\left[(-2)^2+(-5)^2+4^2+1^2+2^2\right]=10$$

- The variance along the second dimension is

$$\frac{1}{5}\cdot\left[3^2+0^2+6^2+(-5)^2+(-4)^2\right]=17.2$$

$\therefore$ s.d $=\sqrt{\text{variance}}$, s.d $=\sqrt{10}$ and s.d $=\sqrt{17.2}$

Hence, after normalization, the observations become

$$x_1=\begin{bmatrix}-2/\sqrt{10}\\3/\sqrt{17.2}\end{bmatrix}\quad x_2=\begin{bmatrix}-5/\sqrt{10}\\0\end{bmatrix}\quad x_3=\begin{bmatrix}4/\sqrt{10}\\6/\sqrt{17.2}\end{bmatrix}\quad x_4=\begin{bmatrix}1/\sqrt{10}\\-5/\sqrt{17.2}\end{bmatrix}$$

$$x_5=\begin{bmatrix}2/\sqrt{10}\\-4/\sqrt{17.2}\end{bmatrix}$$

$$x_1=\begin{bmatrix}-0.6325\\0.7234\end{bmatrix}$$

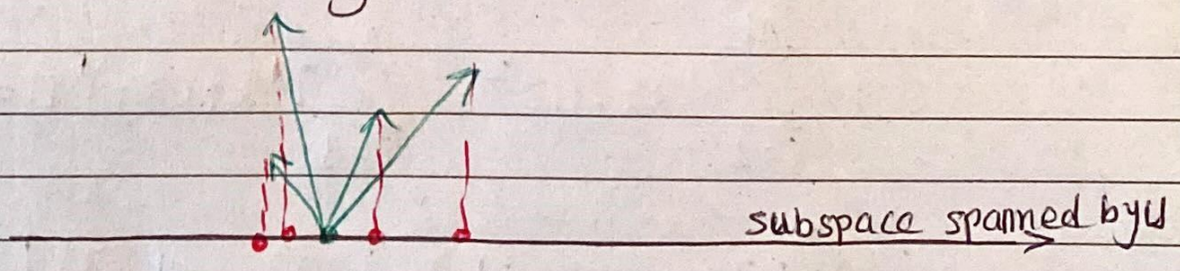
Key Motivation: Variance of Projections

- Earlier in this lecture, we have discussed the ff idea:
- If the variance of a attribute is large
  - it is more interesting (less boring).

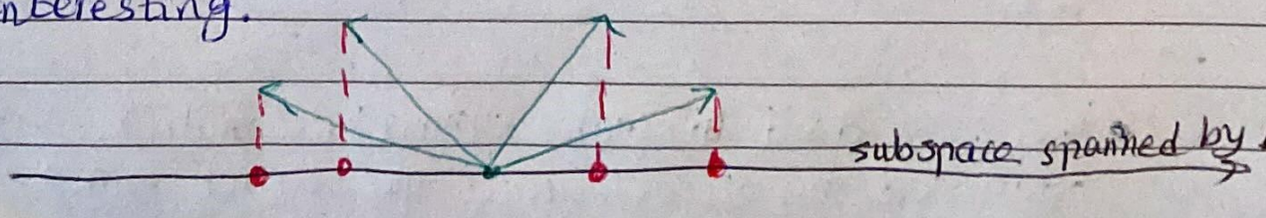
- By the same mentality,
  - for high-dimensional data,
  - the projection along vector u
  - is more interesting if the variance along u is large.

- In other words, we seek u
  - such that the variance is maximized.

- When the variance along u is small...
  - The red dots are the projections of the 4 observations along u.
  - when the variance along u is small,
  - the projections are close to each other,
  - and this is an indication that the projections along u are not interesting.



- When the variance along u is large...
  - when the variance along u is large, the projections
  - are dispare so they can be distinguished easily,
  - or they form some clusters which are far away from each other.
  - This is an indication that the projections along u are interesting.



- Up to now, you should be clear about the reason on
  - Why we are interested in maximizing variance.
- The next question:
  - How can we can do algebraically?

- Recall: The projection formula:

$$\frac{\langle x, u_1\rangle}{\langle u_1, u_1\rangle}\cdot u_1 + \frac{\langle x, u_2\rangle}{\langle u_2, u_2\rangle}\cdot u_2 + \cdots + \frac{\langle x, u_K\rangle}{\langle u_K, u_K\rangle}\cdot u_K$$

- Let u be a unit vector, i.e. $\langle u, u\rangle = 1$
  - The projection of x along vector u is
    - * coefficient $\Rightarrow \langle x, u\rangle \cdot u$

- When focusing on the coefficient, the variance is

$$\frac{1}{N}\sum_{i=1}^{N}\langle x_i, u\rangle^2$$

- When focusing on the coefficient, N times the variance is

$$\sum_{i=1}^{N}\langle x_i, u\rangle^2 = \sum_{i=1}^{N} u' x_i (x_i)' u = u'\left(\sum_{i=1}^{N} x_i (x_i)'\right) u$$

$\langle x_i, u\rangle = u' \cdot x_i$ $\quad \begin{bmatrix} - & u' & - \end{bmatrix} \begin{bmatrix} | \\ x_i \\ | \end{bmatrix}$

(Recall that: u' is the transpose of vector u.)

$$\langle x_i, u\rangle^2 = \begin{bmatrix} - & u' & - \end{bmatrix} \begin{bmatrix} | \\ x_i \\ | \end{bmatrix} \begin{bmatrix} - & x_i' & - \end{bmatrix} \begin{bmatrix} | \\ u \\ | \end{bmatrix}$$

$$\sum_{i=1}^{N}\langle x_i, u\rangle^2 = \sum_{i=1}^{N} \begin{bmatrix} - & u' & - \end{bmatrix} \left(\begin{bmatrix} | \\ x_i \\ | \end{bmatrix} \cdot \begin{bmatrix} - & x_i' & - \end{bmatrix} \right) \begin{bmatrix} | \\ u \\ | \end{bmatrix}$$

- This is a $d \times d$ square symmetric matrix, denoted by S, is called the scatter matrix of the dataset.

Example: If $x_i = \begin{bmatrix} 2 \\ -1 \\ -3 \end{bmatrix}$, then $x_i (x_i)'$?

$$x_i (x_i)' = \begin{bmatrix} 2 \\ -1 \\ -3 \end{bmatrix} \cdot [2 \ -1 \ -3] = \begin{bmatrix} 2\times2 & 2\times-1 & 2\times-3 \\ -1\times2 & -1\times-1 & -1\times-3 \\ -3\times2 & -3\times-1 & -3\times-3 \end{bmatrix} = \begin{bmatrix} 4 & -2 & -6 \\ -2 & 1 & 3 \\ -6 & 3 & 9 \end{bmatrix}$$

(3×1) $\quad$ $1\times(3)$

Quiz (3): Suppose that after pre-processing, the three observations on a dataset are $\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} -2 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -3 \end{bmatrix}$

What is the scatter matrix of this dataset?

Solution:

Let $x_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $x_2 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$, and $x_3 = \begin{bmatrix} 1 \\ -3 \end{bmatrix}$

$$x_1 (x_1)' = \begin{bmatrix} 1 \\ 2 \end{bmatrix} [1 \ 2] = \begin{bmatrix} 1\times1 & 1\times2 \\ 2\times1 & 2\times2 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

$$x_2 (x_2)' = \begin{bmatrix} -2 \\ 1 \end{bmatrix} [-2 \ 1] = \begin{bmatrix} -2\times-2 & -2\times1 \\ 1\times-2 & 1\times1 \end{bmatrix} = \begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix}$$

$$x_3 (x_3)' = \begin{bmatrix} 1 \\ -3 \end{bmatrix} [1 \ -3] = \begin{bmatrix} 1\times1 & 1\times-3 \\ -3\times1 & -3\times-3 \end{bmatrix} = \begin{bmatrix} 1 & -3 \\ -3 & 9 \end{bmatrix}$$

$\therefore$

The scatter matrix $= x_1 (x_1)' + x_2 (x_2)' + x_3 (x_3)'$

$$= \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} + \begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix} + \begin{bmatrix} 1 & -3 \\ -3 & 9 \end{bmatrix} = \begin{bmatrix} 1+4+1 & 2-2-3 \\ 2-2-3 & 4+1+9 \end{bmatrix}$$

$$= \begin{bmatrix} 6 & -3 \\ -3 & 14 \end{bmatrix}$$

- We focusing on the coefficient, N times the variance is

$$\sum_{i=1}^{N} \langle x_i, u \rangle^2 = \sum_{i=1}^{N} u' x_i (x_i)' u = u' \underbrace{\left(\sum_{i=1}^{N} x_i (x_i)' \right)}_{S: \text{ the scatter matrix}} u$$

- The unit vector u that maximizes $u'Su$
  - is same as the eigenvector with largest eigenvalue.
- Indeed, the eigenvalue is N times the variance.

- The eigenvector is called the first principal component (PC).
- In words, the projections along the 1st PC is the most interesting.

Principal Components

- The desired unit vector u
  - is the eigenvector of S with the largest eigenvalue,
  - which is called the first principal component (PC).

- By analogous reasoning,
  - the next most interesting projection
  - is along the eigenvector with 2nd largest eigenvalue.

- We sort the eigenvectors of S by descending order of
  - their eigenvalues.
- The eigenvalues are called 1st PC, 2nd PC, etc...

Eigenvectors of S: u_1 (1st PC) u_2 (2nd PC) ... u_d (dth PC)

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$$

- When we perform dimension reduction
  - you only retain a few of the PCs,
  - and of course we retain the ones with largest eigenvalues.

- The eigenvalues guide you how to cut off.
  - One common way is to set a threshold $\theta \in (0,1)$
  - and choose the minimum k such that

$$\sum_{j=1}^{k} \lambda_j \Big/ \sum_{j=1}^{d} \lambda_j > \theta \qquad \text{e.g. } \frac{\geq 90}{100} = 0.9$$

Summary of the Theory of PCA

1. Given a dataset, before performing PCA
   - we need to do centering.
2. After centering, it is advisable to do normalization.
3. We are interested in finding a unit vector u,
   - such that the variance of
   - the projections along direction u is maximized.
4. By the calculations we just did,
   - The variance is $\frac{1}{N} \cdot u'Su$

   where,

   $S := \left(\sum_{i=1}^{N} x_i (x_i)' \right)$ is a $d \times d$ symmetric matrix

   called the scatter matrix.
5. We are therefore interested in computing the unit vector u
   - which maximizes $u'Su$
   - It is the eigenvector of S with the largest eigenvalue.
6. The desired unit vector $u_1, u_2, \ldots, u_d$
   - are the eigenvectors of S
   - sorted by decreasing eigenvalues.

   * For dimension reduction, select the top k eigenvectors for a reasonable choice of k.

Algorithm of Principal Component Analysis (PCA)

- Assuming you treat computing eigenvectors as a blackbox:
  1. centering
  2. normalization (
  3. compute the scatter matrix
     - using the centered & normalized dataset
  4. compute the eigenvectors & eigenvalues of the scatter matrix.

- The scatter matrix $S := \left(\sum_{i=1}^{N} x_i (x_i)'\right)$

- There is a more compact formula for S.
  - let X denote the $N \times d$ matrix
  - which contains all the observations after centering and normalization,
  - one row per observation.
  - Then

$$S = X'X$$

Example: suppose the matrix X is already given in R,
- then you can compute S simply by

```
S <- t(X) %*% X
```

- After computing S in R
  - you can compute the eigenvector & eigenvalues
  - using eigen() function.

```
eigen(S)
```

Example: If you want to compute the eigenvector and eigenvalues of the matrix $M := \begin{bmatrix} 1 & -3 & 1 \\ -3 & -1 & 2 \\ 1 & 2 & 3 \end{bmatrix}$

In R, you should code

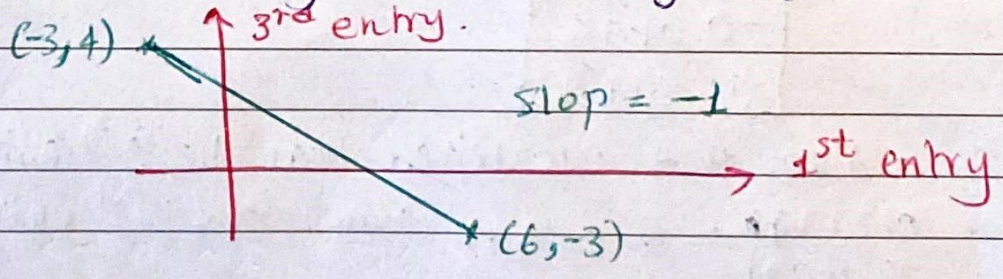
```
M <- matrix(c(1,-3,1,-3,-1,2,1,2,3), ncol=3, byrow=TRUE)
eigen_of_M <- eigen(M)
```

A Concrete Example

- Now, let's use the built-in function to do PCA on a small dataset,
  - to see how PCA copes with
    - redundant,
    - irrelevant and
    - boring dimensions.
- Suppose the original dataset (before centering + normalization) is

$$\begin{bmatrix} -3 \\ 2.2 \\ 4 \\ 1.1 \\ -5.3 \end{bmatrix} \begin{bmatrix} -2 \\ 1.9 \\ 3 \\ 1.2 \\ -4.1 \end{bmatrix} \begin{bmatrix} 0 \\ 2.1 \\ 1 \\ 1.2 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1.7 \\ 0 \\ 1.2 \\ 2.1 \end{bmatrix} \begin{bmatrix} 2 \\ 1.8 \\ -1 \\ 1.2 \\ 3.7 \end{bmatrix} \begin{bmatrix} 4 \\ 2.3 \\ -3 \\ 0 \\ 7.9 \end{bmatrix} \begin{bmatrix} 6 \\ 2.2 \\ -5 \\ 1.2 \\ 12.3 \end{bmatrix}$$

Pattern 1: the 1st entry + the 3rd entry = 1 (redundance)



Pattern 2: the 5th entry ≈ 2 times the 1st entry (redundance)

Observation 3: the 2nd entries look like perturbation from 2 (irrelevance).

Observation 4: most of the 4th entries are 1.2 (boring)

- Now, we enter the dataset into R and we prcomp() to carry out PCA.
  - we first do centering but not normalization.

```
pca.noscale <- prcomp(X, center = TRUE)
```

$s.d = \sqrt{variance}$

$variance = (s.d)^2$.

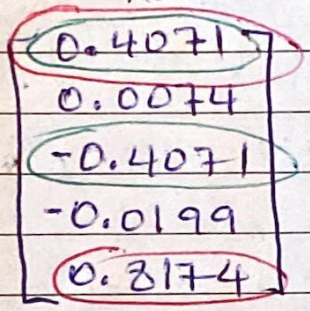
s.d

7.82 0.44

- You can see that the first PC has a very large eigenvalue $(7.82)^2$
  - compared with the remaining eigenvalues.

Note that: Variance = eigenvalue.

- Focus on the 1st PC.
  - PCA successfully identified that the 1st entry and the 3rd entry are related by a coefficient of -1. (recall that their sum is constant).



  - Also, PCA identified that the 5th entry is roughly $\frac{0.81745}{0.4071} = 2.00767$ times the 1st entry.
    - It is very close to 2; the small difference is due to the noises.

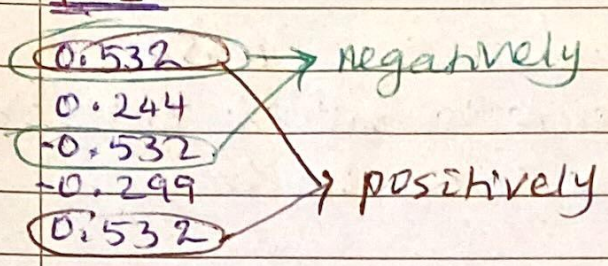
- Next, we do both centering and normalization.
  - scale = TRUE do centering & do normalization

s.d

1.82 1.12 0.66 0.019 8.06×10^{-17}

- You can see that the 1st PC has a large eigenvalue $(1.82)^2$, the 2nd PC has a smaller eigenvalue $(1.12)^2$, and the other eigenvalues are fairly small.

PC1



- Now, the 1st PC does not reveal to you the linear relationship between the 1st, 3rd & 5th entries.

- But it still demonstrates that the 1st & 3rd entries are negatively correlated,
  - while the 1st & the 5th entries are positively correlated.

- In this example, it appears that
  - PCA finds out the hidden pattern better without normalization than with normalization.
- This is not unexpected,
  - as the numbers in the dataset are of roughly the same magnitudes.

- However, in some datasets,
  - it may happen that the numbers in one of the dimensions are of small magnitude (say $\sim 10^{-3}$),
  - while the numbers in another dimension are of much higher magnitude (say $\sim 10^{3}$).

- In this case, normalization is essential to avoid PCA being "dominated"
  - by the dimension with high magnitudes.

Key Lemma

- The unit vector u that maximizes $u'Su$
  - is same as the eigenvector with the largest eigenvalue
  - Indeed, the eigenvalue is the Variance

Proof Sketch: (This is will not be in Exam).

- This proof relies on the Spectral Theorem
  - Let $u_1, u_2, \ldots, u_d$ denote the eigenvectors of S
  - and let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$ denoted the corresponding eigenvectors.
  - By normalization, we may assume that they are unit vectors.

- Since there are d such vectors which are pairwise orthogonal,
  - they form a basis of $\mathbb{R}^d$.

- Thus, for any unit vector $u \in \mathbb{R}^d$, we can write u as

$$u = \sum_{j=1}^{d} \langle u, u_j \rangle \cdot u_j, \quad \text{whereas} \sum_{j=1}^{d} \langle u, u_j \rangle^2 = 1$$

- Then by a simple matrix calculation

$$u'Su = \sum_{j=1}^{d} \lambda_j \langle u, u_j \rangle^2$$

- This summation on the RHS is clearly maximized
  - when $\langle u, u_1 \rangle = 1$
  - and $\langle u, u_2 \rangle = \langle u, u_3 \rangle = \ldots \langle u, u_d \rangle = 0$

In other words, $u = u_1$.