

Nama : Fasya Hanifah

NIM : 1103200149

RANGKUMAN PRINCIPAL COMPONENT ANALYSIS (PCA)

Example graph : This graph was drawn from single cell RNA-seq. There were about 10.000 transcribed genes in each cell. Each dot represents a single-cell and its transcription profile. We see that in this graph the blood cells form one cluster that's different from pluripotent cells which is different from neuronal cells and dermal or epidermal cells. So that question is How does transcription 10.000 genes get compressed to a single dot on a graph?

Here the answer PCA (Principal Component Analysis) is a method for compressing a lot of data into something that captures the essence of the original data.

An Introduction to Dimensions

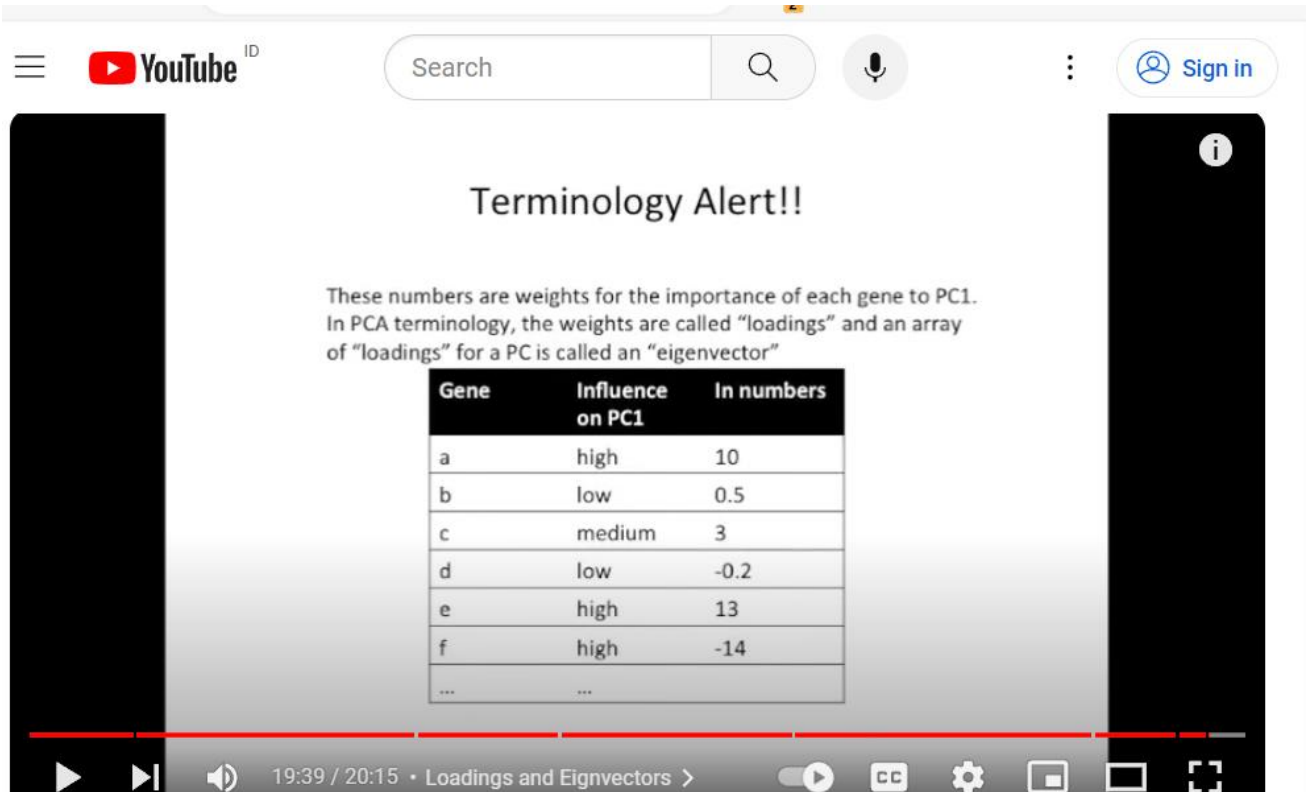
- 1-Dimension (1-D) = a number line, example (A pretend RNA-seq data set for a single cell). There are two output (A uniform distribution and A non uniform distribution) (if we have one data)
- 2-D = a normal graph, in two dimensional graphs we have two axes. We might see two expression in the two cells is correlated and not correlated. If the expression is correlated its mean genes that are highly transcribed in cell one are also highly transcribed in cell two and genes that are lowly transcribe in cell one are also lowly transcribed in self to and if the expression is not correlated its mean gene is highly transcribed in cell one that doesn't tell us anything about whether its highly or lowly transcribed in cell two. (if we have data from two cells)
- 3-D = a fancy graph that has depth = have three separate axes (if we have data from three cells)

PCA takes a dataset with a lot of dimensions and flattens it to 2 or 3 dimensions so we can look it. PC1 the direction of the most variation in gene expression, PC 2 the 2nd most variation in gene expression.

- If we had 2 cell = PC1 captures the direction where most of the variation is, PC2 captures the direction with the 2nd most variation.
- If we had 3 cells = PC1 span the direction of the most variation, PC2 span the direction 2nd most variation, PC3 span the direction of the 3rd most variation
- If we had 4 cells = PC1 span the direction of the most variation, PC2 span the direction 2nd most variation, PC3 span the direction of the 3rd most variation, PC4 = span the direction of the 4th most variation

HOW WE PLOT CELL? The length and direction of PC1 is mostly determined by the circled genes. Genes with little influence on PC1 get value close to zero, and genes with more influence get numbers further from zero.

For identify key genes we want to find out which genes had a big influence in putting dermal cells on the left neural cells on the right, we could look at the influence scores in PC1 and if we wanted to find put which genes help distinguish blood cells from neural and dermal cells, we could look at the influence score in PC2.



The image shows a YouTube video player interface. At the top, there is a search bar and a 'Sign in' button. The video content is a slide titled 'Terminology Alert!!'. The slide text explains that the numbers are weights for the importance of each gene to PC1, and in PCA terminology, these are called 'loadings'. An array of 'loadings' for a PC is called an 'eigenvector'. Below the text is a table with three columns: 'Gene', 'Influence on PC1', and 'In numbers'. The table lists genes a through f, along with their influence levels and numerical values. The video player controls at the bottom show the video is at 19:39 / 20:15, with a title 'Loadings and Eignvectors >'.

Terminology Alert!!

These numbers are weights for the importance of each gene to PC1. In PCA terminology, the weights are called "loadings" and an array of "loadings" for a PC is called an "eigenvector"

Gene	Influence on PC1	In numbers
a	high	10
b	low	0.5
c	medium	3
d	low	-0.2
e	high	13
f	high	-14
...

19:39 / 20:15 • Loadings and Eignvectors >