8. 차원축소 (PCA)

2019년 4월 14일 일요일 오후 2:02

데이타를 분석할때 피쳐가 많으면 데이타 분석이 어렵고, 특히 3개 이상 (3차원)의 피쳐가 존재할 경우 시각화가 어려워진다.
피쳐의 수를 줄인다는 말과 같고, 앞에서 언급한 바와 같이 데이타 분석에서는 차원을 줄여서 시각화를 가능하게 해서 데이타 분석을 용이하게 할 수 있다
차원을 줄여서 데이타의 특성을 파악할 필요가 있고, 또한 머신러닝에 있어서 학습 데이타의 수를 줄이고, 학습에 필요한 컴퓨팅 파워를 절약하기 위해서 차원 감소는 유용한 기법이 된다.

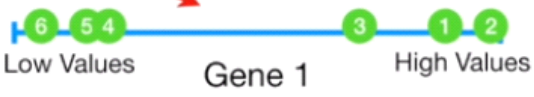
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

We've measured transcription of two genes, gene 1 and gene 2...

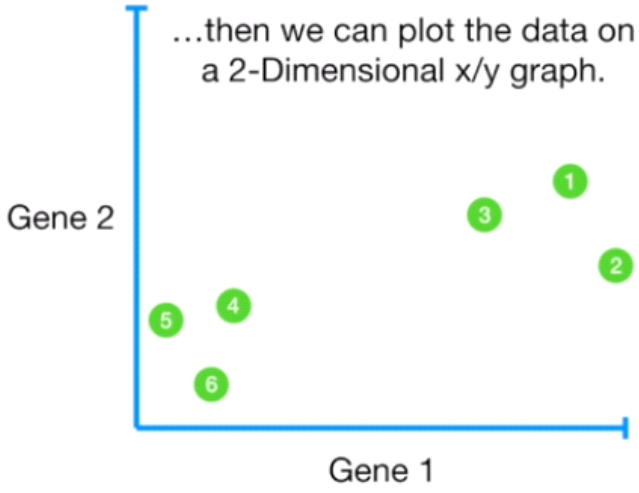
Mouse = student
Gene1 = math, english..

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1

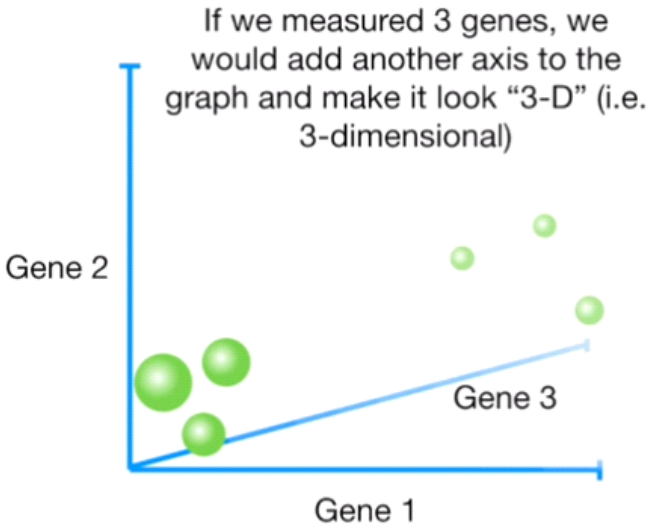
Even though it's a simple graph, it shows us that mice 1, 2 and 3 are more similar to each other than they are to mice 4, 5 6.



	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1



	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2



	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7

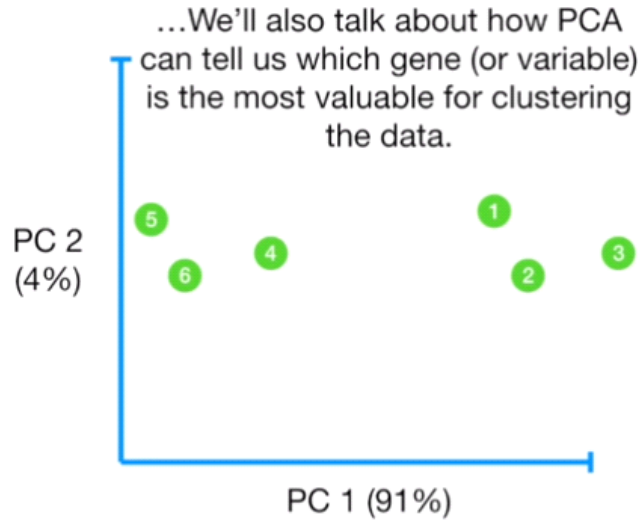
If we measured 4 genes, however, we can no longer plot the data - 4 genes require 4 dimensions.

>> 4차원 이상의 데이터를 어떻게 다루고, 2-d PCA 그래프로 만들 수 있을까?

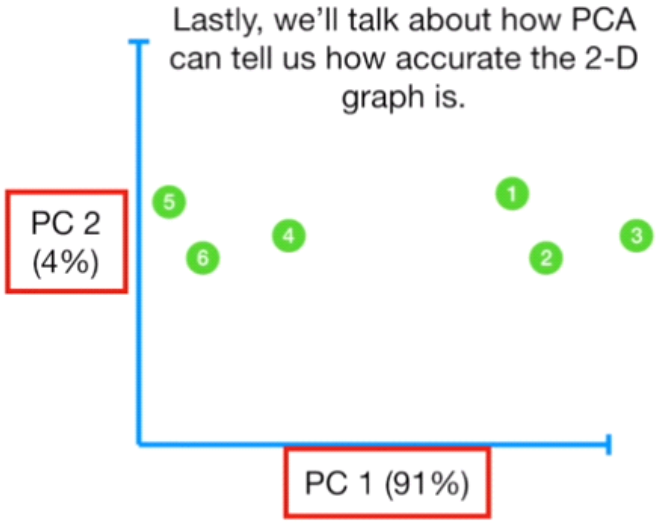
>> PCA는 어떤 변수가 데이터를 클러스터링 하는데 가장 valuable한지 어떻게 구분할까?

>> PCA가 얼마나 정확할까?

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7

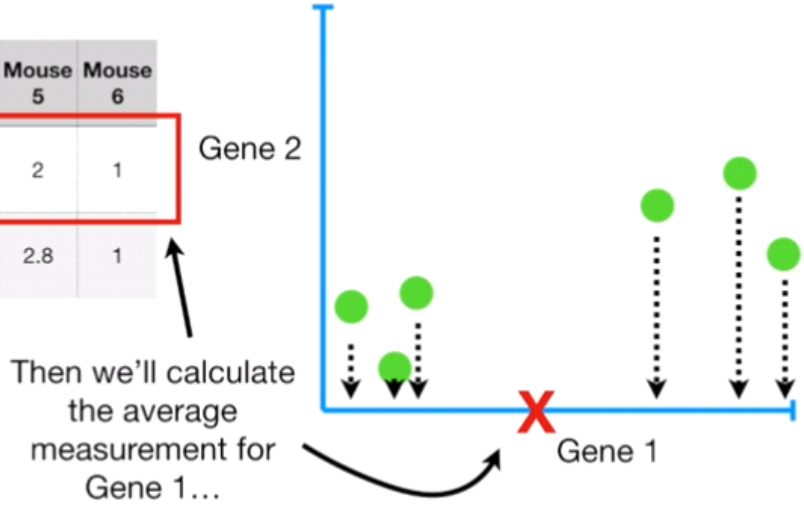


	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7



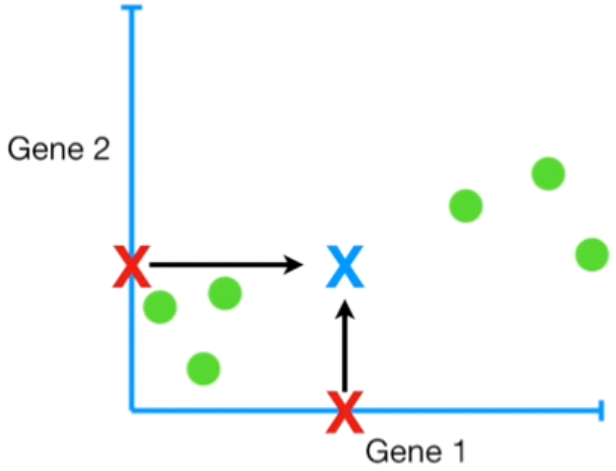
1. **Centering 하기**

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

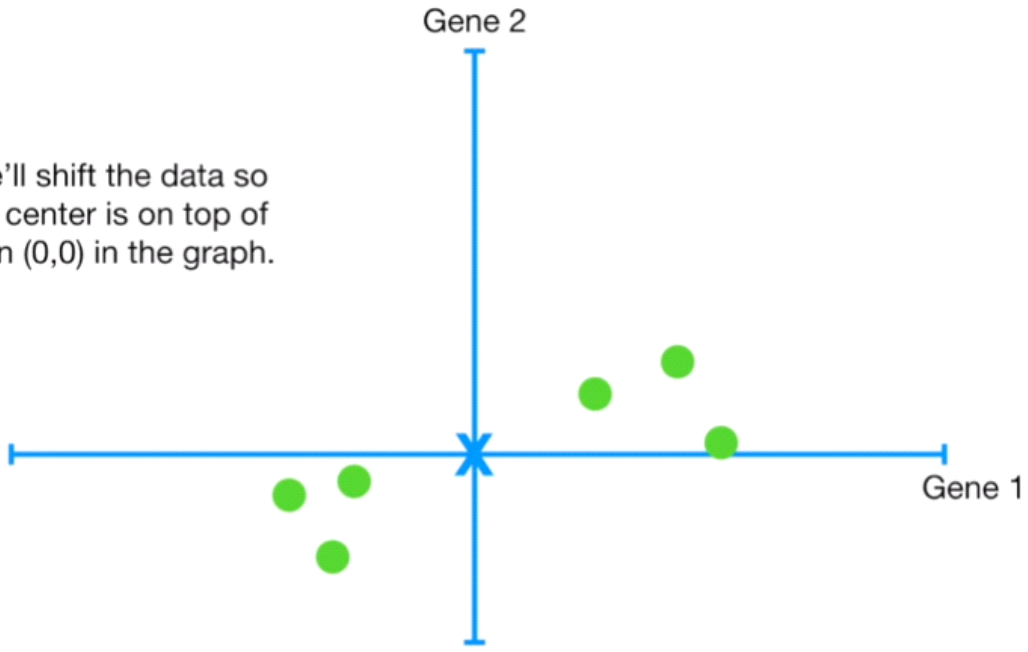


	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

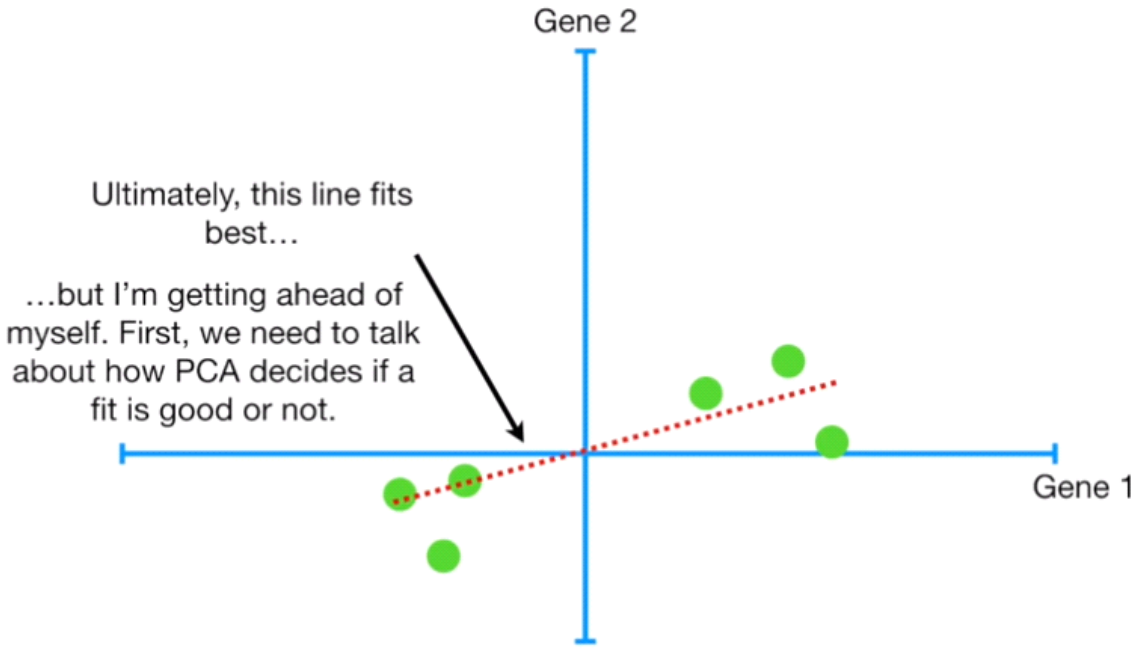
From this point on, we'll focus on what happens in the graph; we no longer need the original data…



Now we'll shift the data so that the center is on top of the origin (0,0) in the graph.

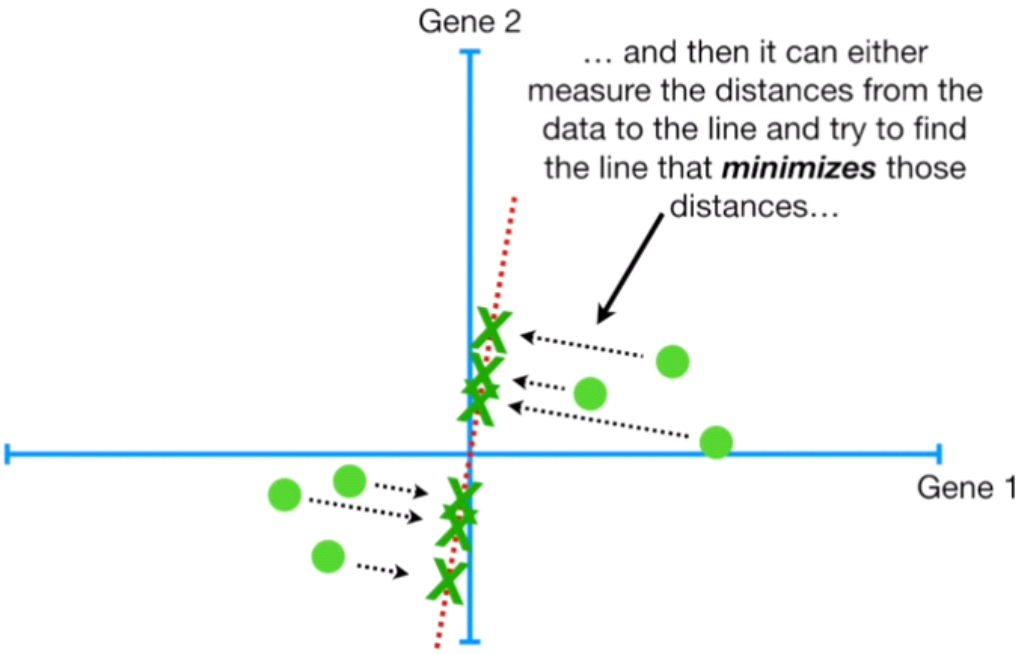


2. PC1 (principal component) 찾기.

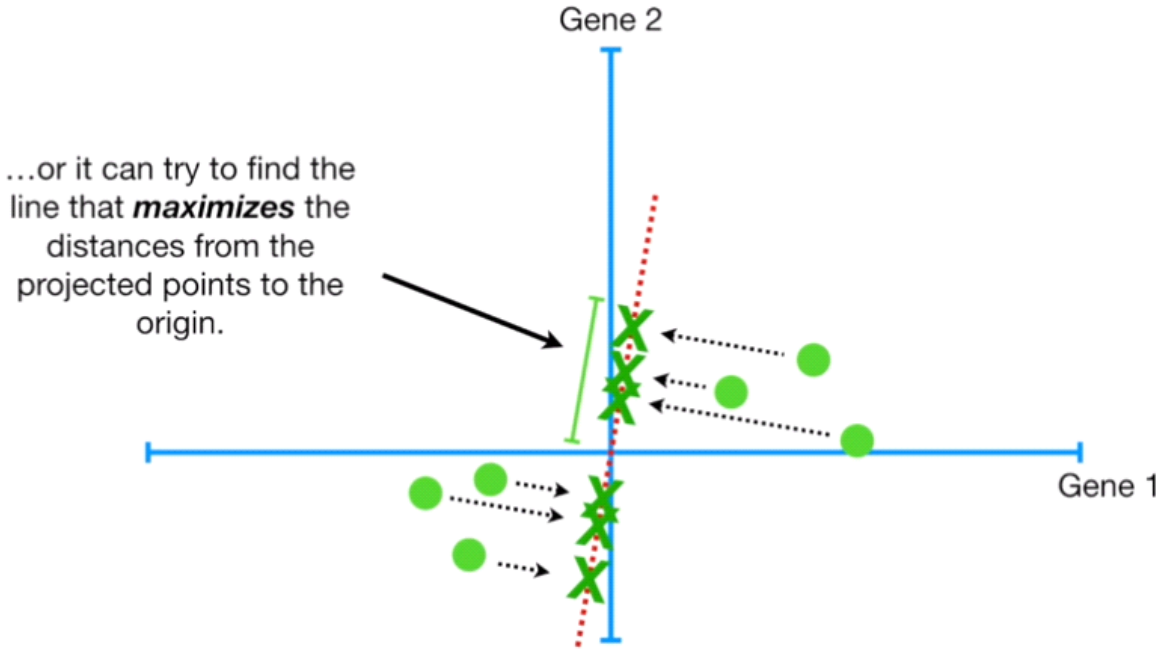

Gene 2
Ultimately, this line fits best…
…but I'm getting ahead of myself. First, we need to talk about how PCA decides if a fit is good or not.
Gene 1

- How?

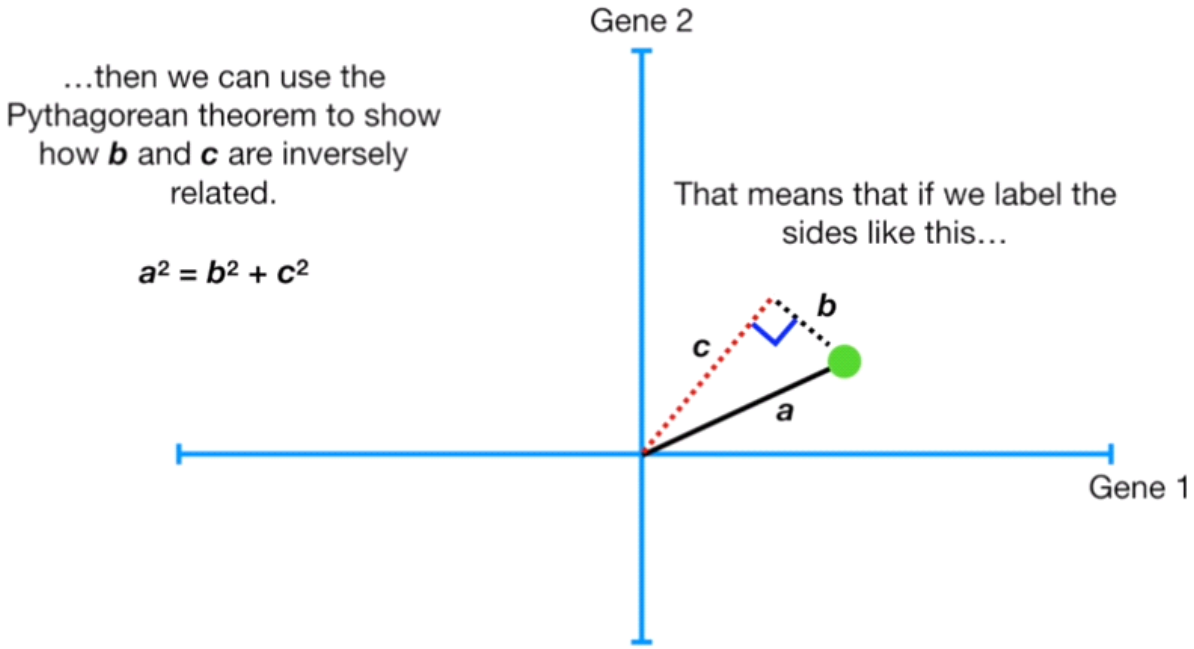
(1)


Gene 2
… and then it can either measure the distances from the data to the line and try to find the line that *minimizes* those distances…
Gene 1

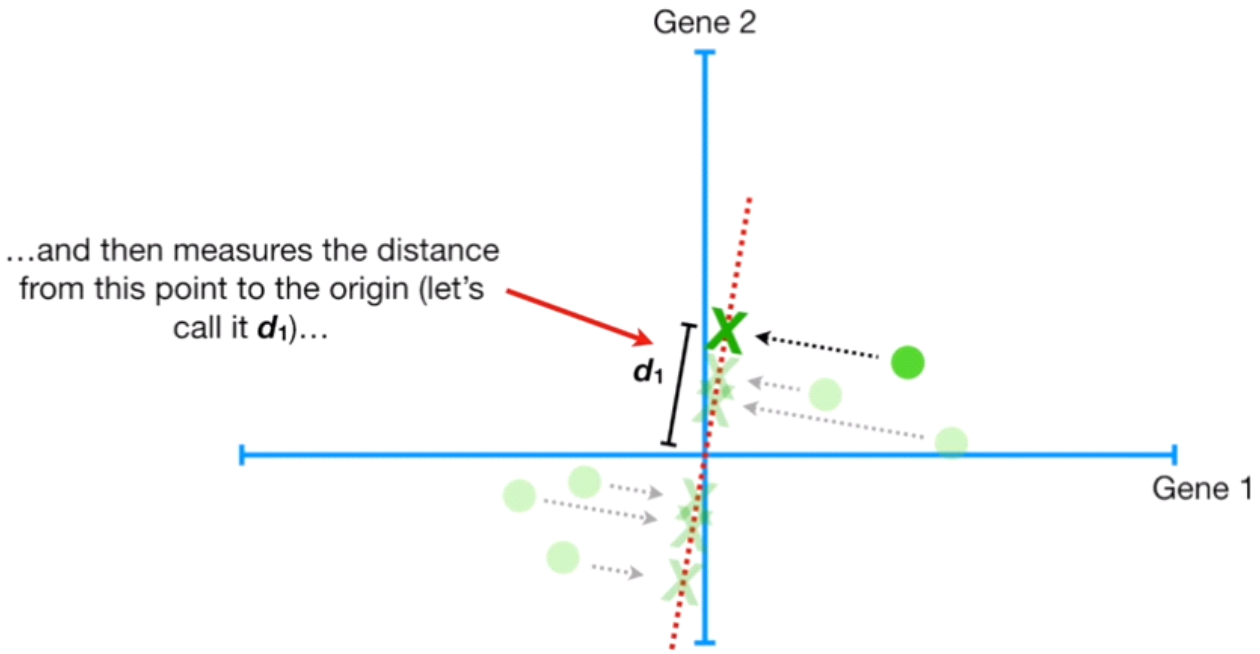
(2)



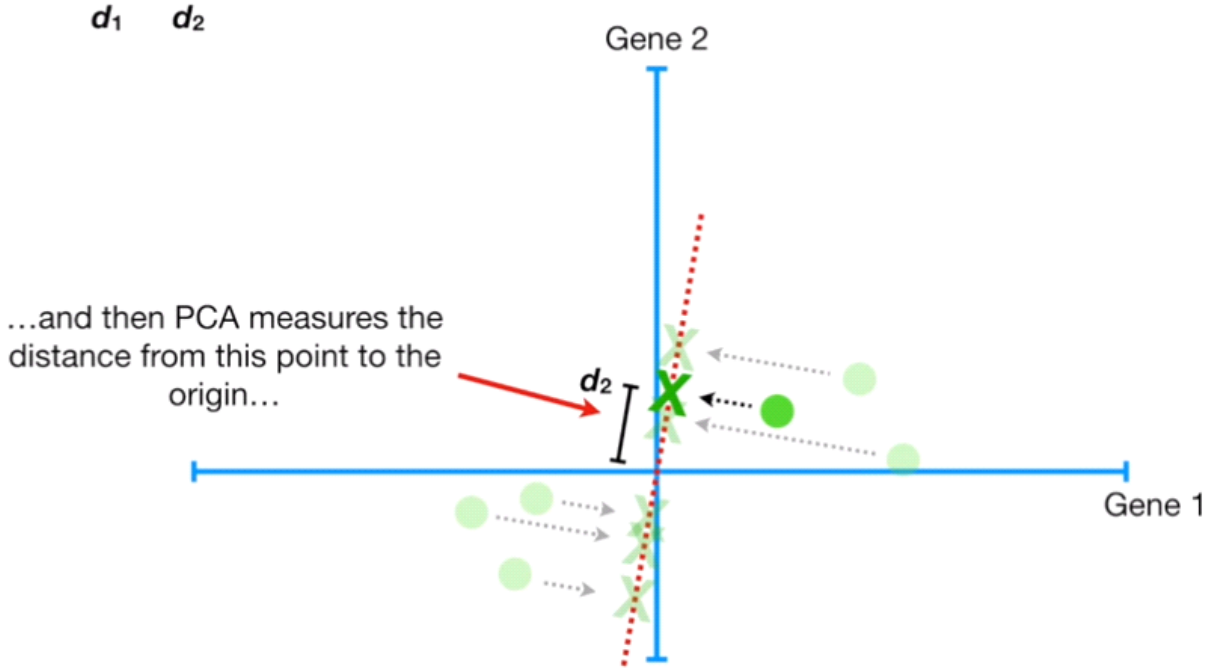
(1) = (2) 인 이유, a는 line과 상관없이 항상 일정. Line을 돌려보면서 생각해봅시당.

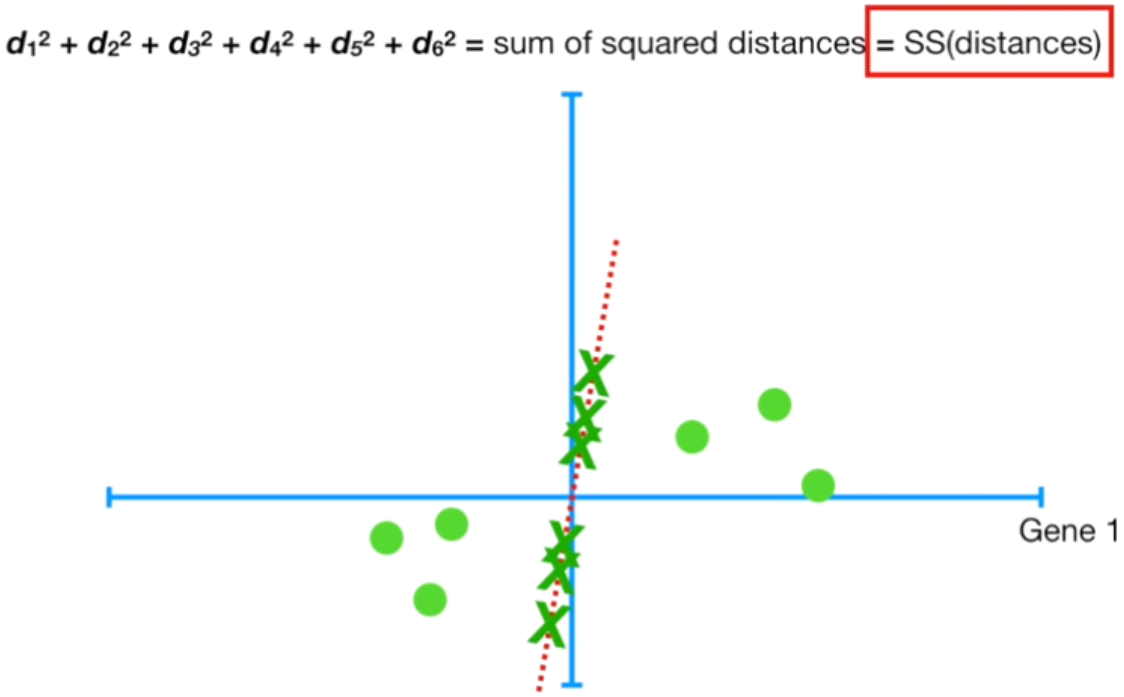


- Largest SS distance 를 갖는 선분을 찾자. => PC1 (

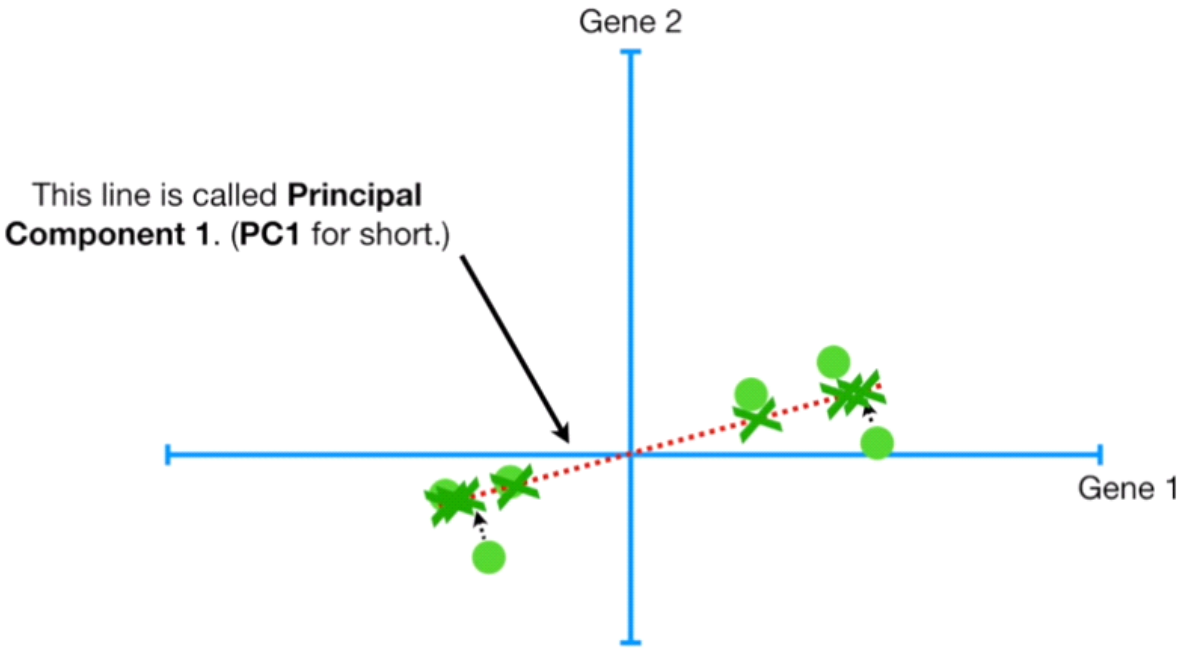

Gene 2
…and then measures the distance from this point to the origin (let's call it d_1)…
d_1
Gene 1

d_1 d_2


Gene 2
…and then PCA measures the distance from this point to the origin…
d_2
Gene 1

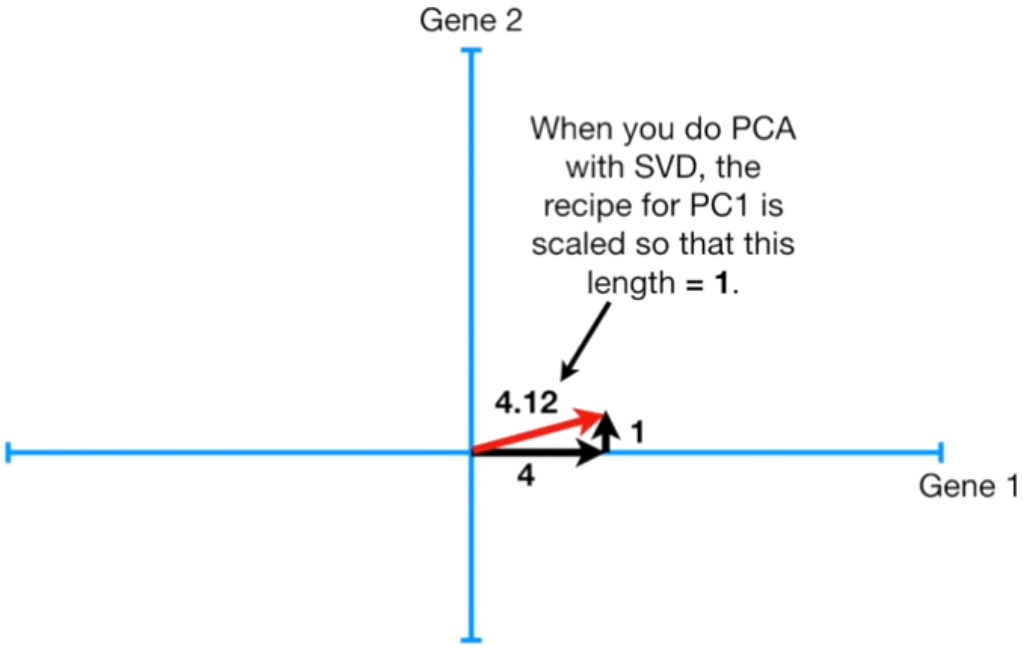


➢ 기울기 : 0.25



- (4,1) 좌표를 갖는 선분, 즉 x축을 따라서 넓게 분포되어 있다는 것을 알 수 있다. (gene1)
- 이것은 PC1을 구성하는데, Gene1이 4만큼, Gene2가 1만큼 기여했다고 볼 수 있으며, 즉 기여도를 측정하는 지표가 된다. (recipe를 생각해)
  - **Linear combination of** Genes1 and 2

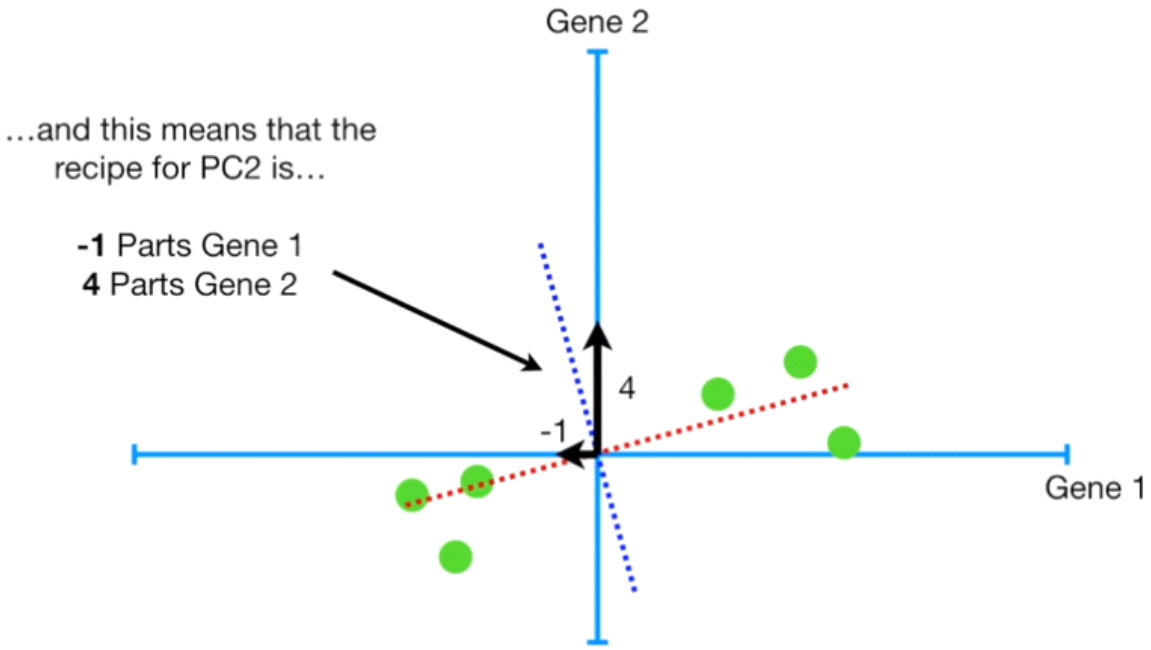
3. Scaling



- 피타고라스를 이용하여 a 크기를 찾고, 해당 길이를 1로 맞춘다.

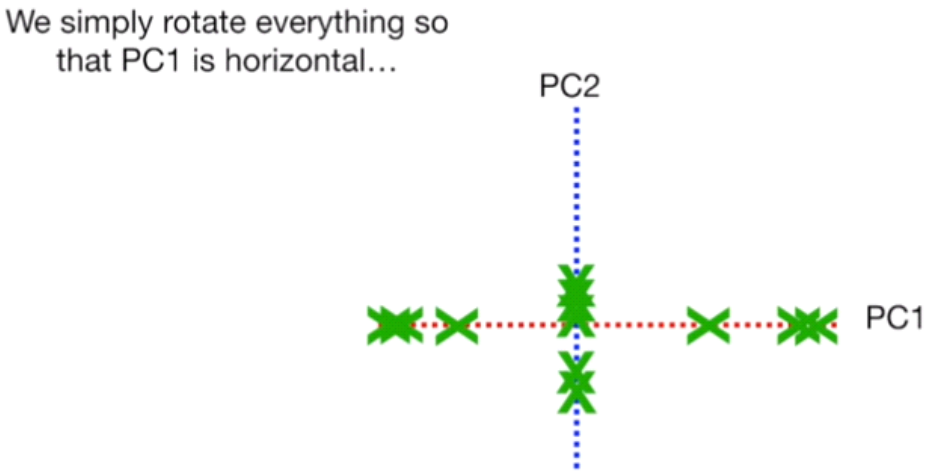
빨간선은 **Eigenvector** (= 루트 씌우면 **Singular** vector)

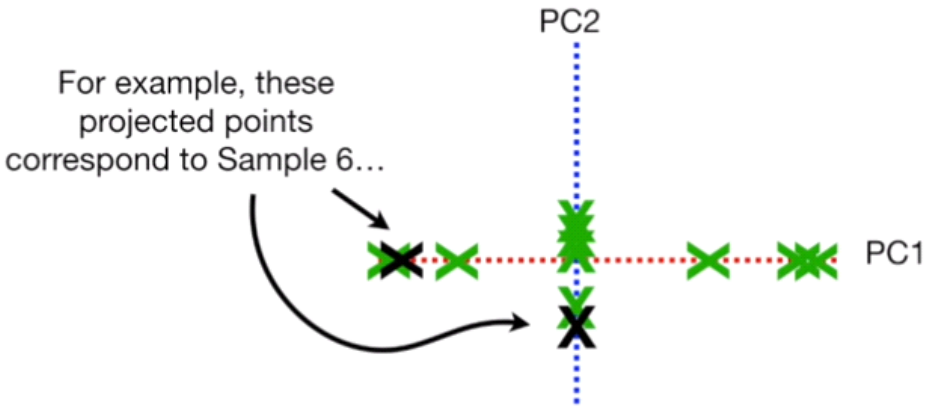
4. PC2 (수직벡터)

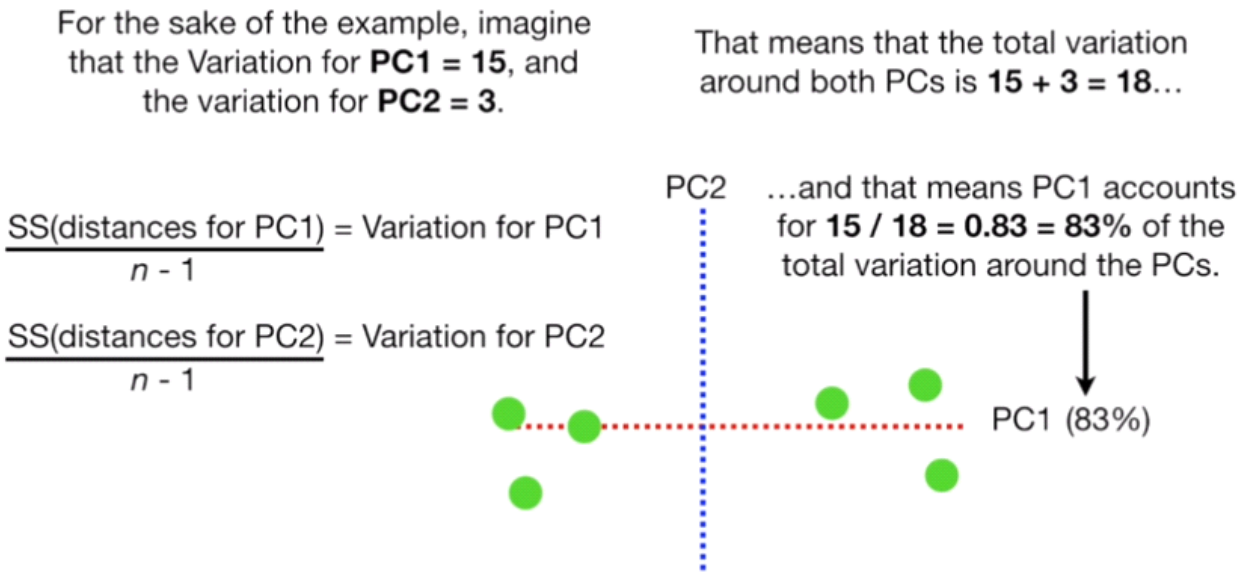


5. 회전

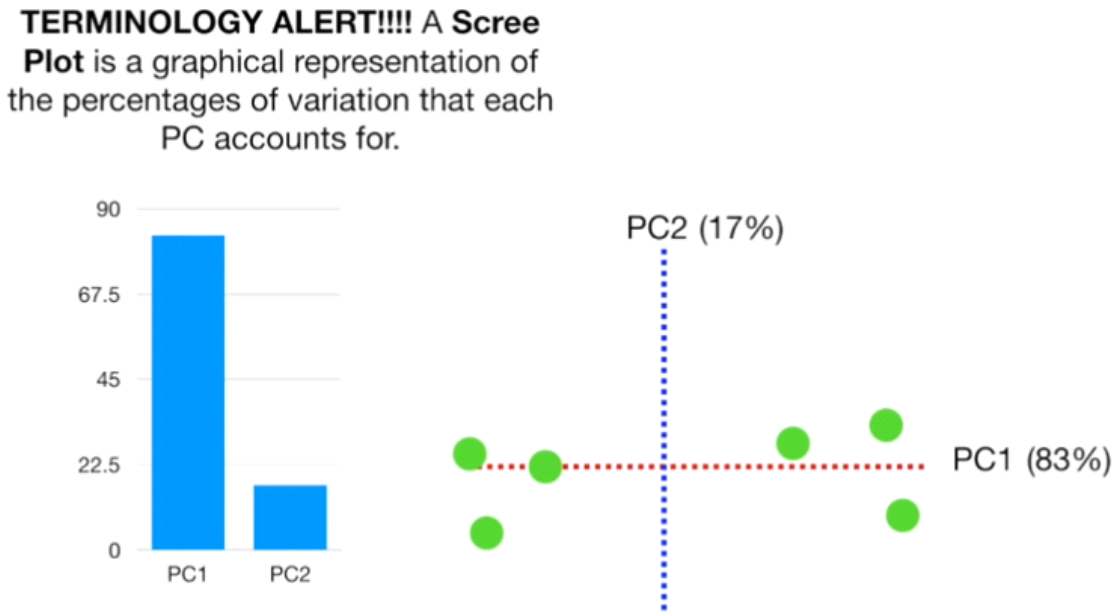
- Project to each line and locate everything so that pc1 is horizontal



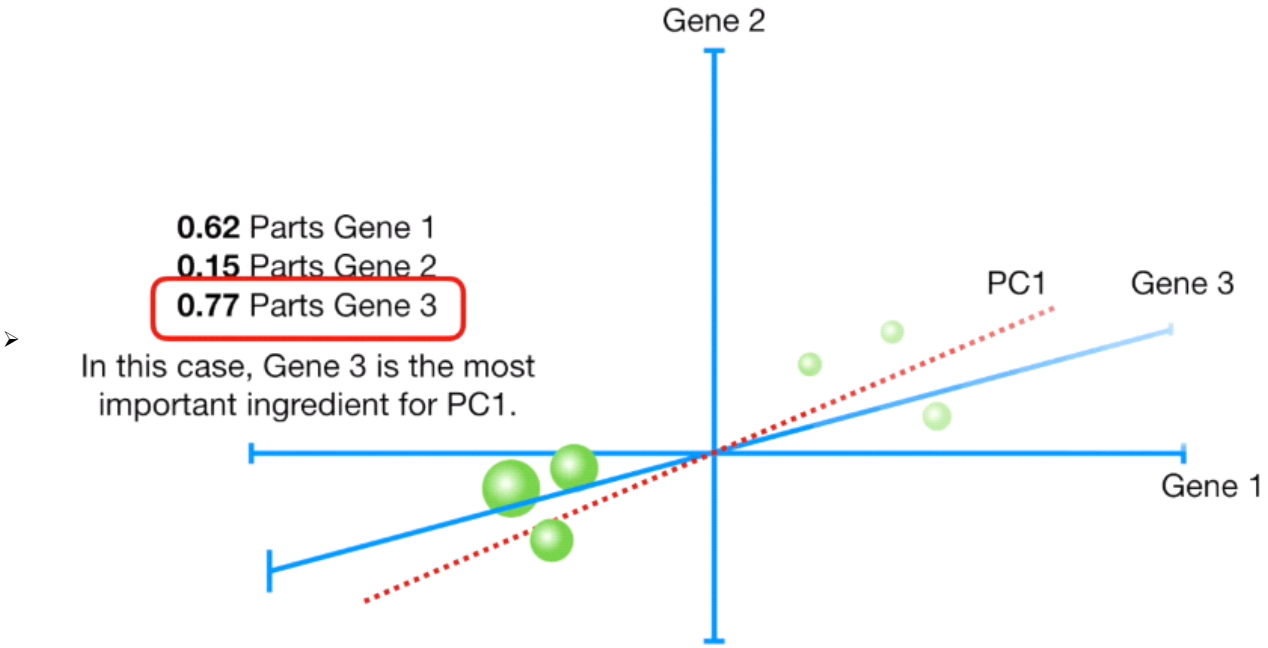


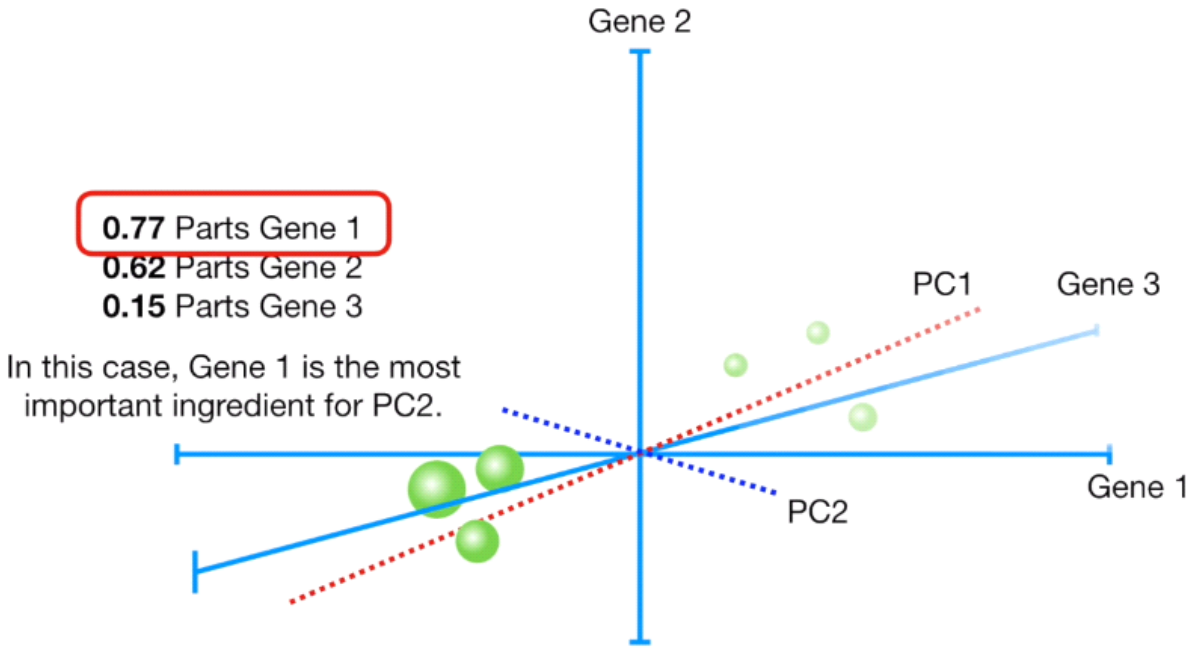


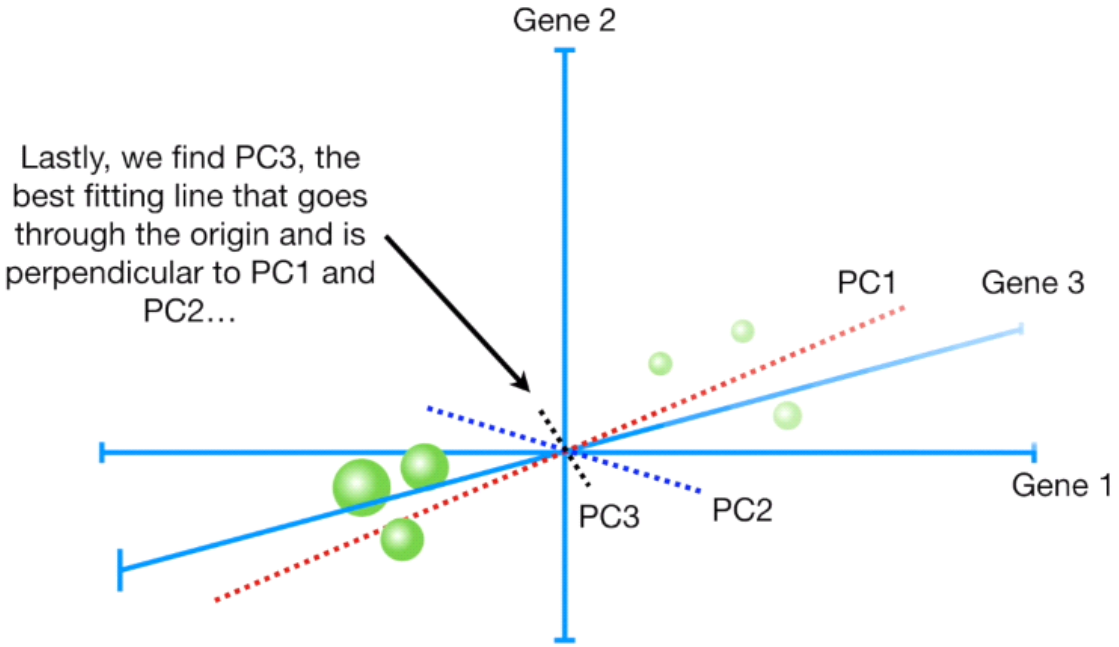
6. **Scree Plot**

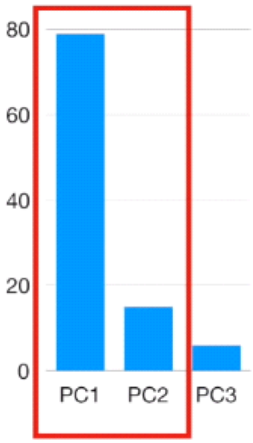


- Other


Gene 2
0.62 Parts Gene 1
0.15 Parts Gene 2
0.77 Parts Gene 3
In this case, Gene 3 is the most important ingredient for PC1.
PC1
Gene 3
Gene 1


Gene 2
0.77 Parts Gene 1
0.62 Parts Gene 2
0.15 Parts Gene 3
In this case, Gene 1 is the most important ingredient for PC2.
PC1
Gene 3
PC2
Gene 1





That means that a 2-D graph, using just PC1 and PC2, would be a good approximation of this 3-D graph since it would account for 94% of the variation in the data.