차원의 저주

훈련 샘플의 특성 개수 ↑

- 훈련 속도 ↓
- 과대적합 위험성 ↑

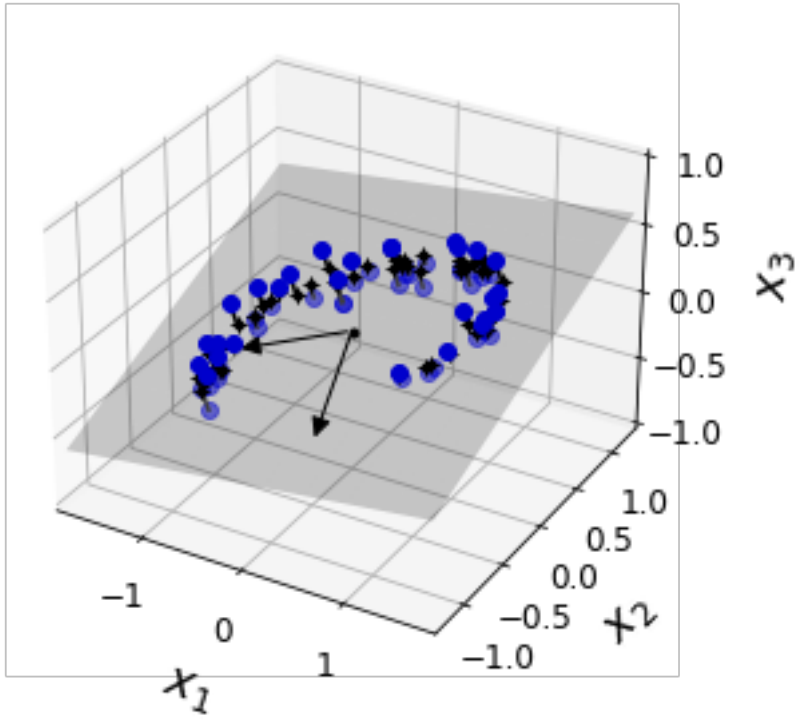
⊙ 높은 차원에서 과대적합이 일어나는 이유.

차원이 높을 수록 올바른 가설을 찾기 힘들어 짐.

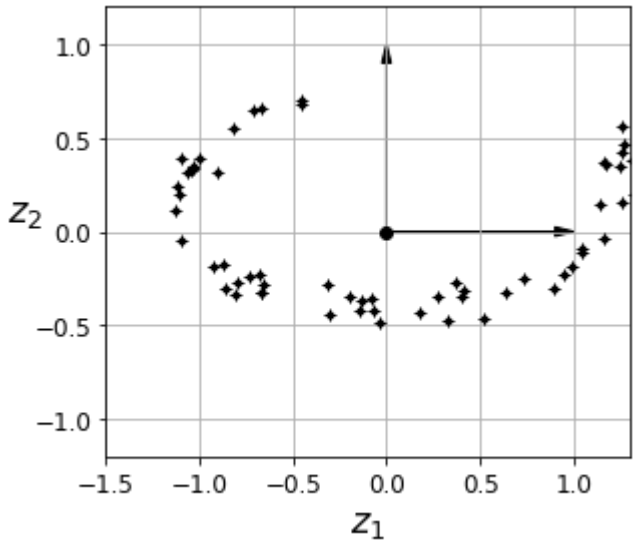
⊙ 모델 만들기 : 데이터를 설명할 하나의 가설이 존재함을 가정

데이터를 통해 그 가설과 같은 작용을 할 모델을 만드는 것

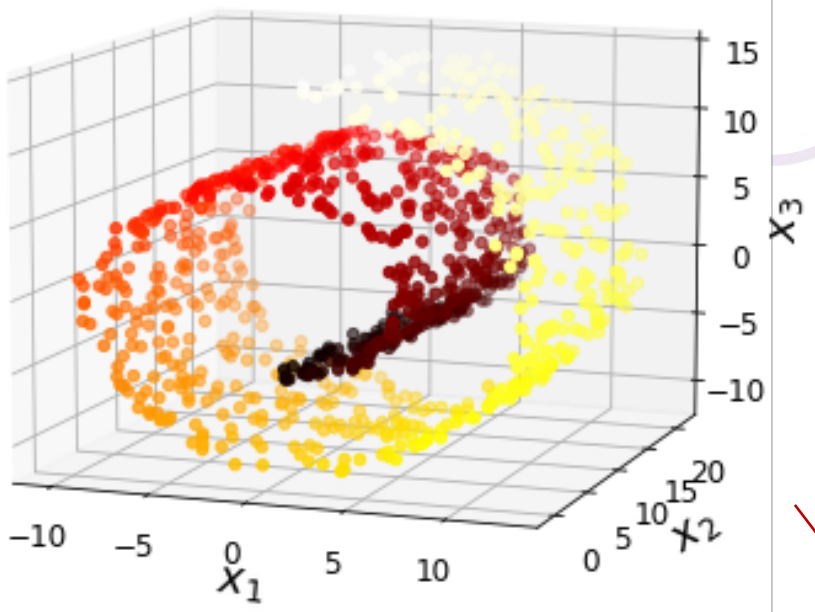
투영



훈련 샘플을 고차원 공간 안의 저차원 부분 공간에 투영시키는 방법.

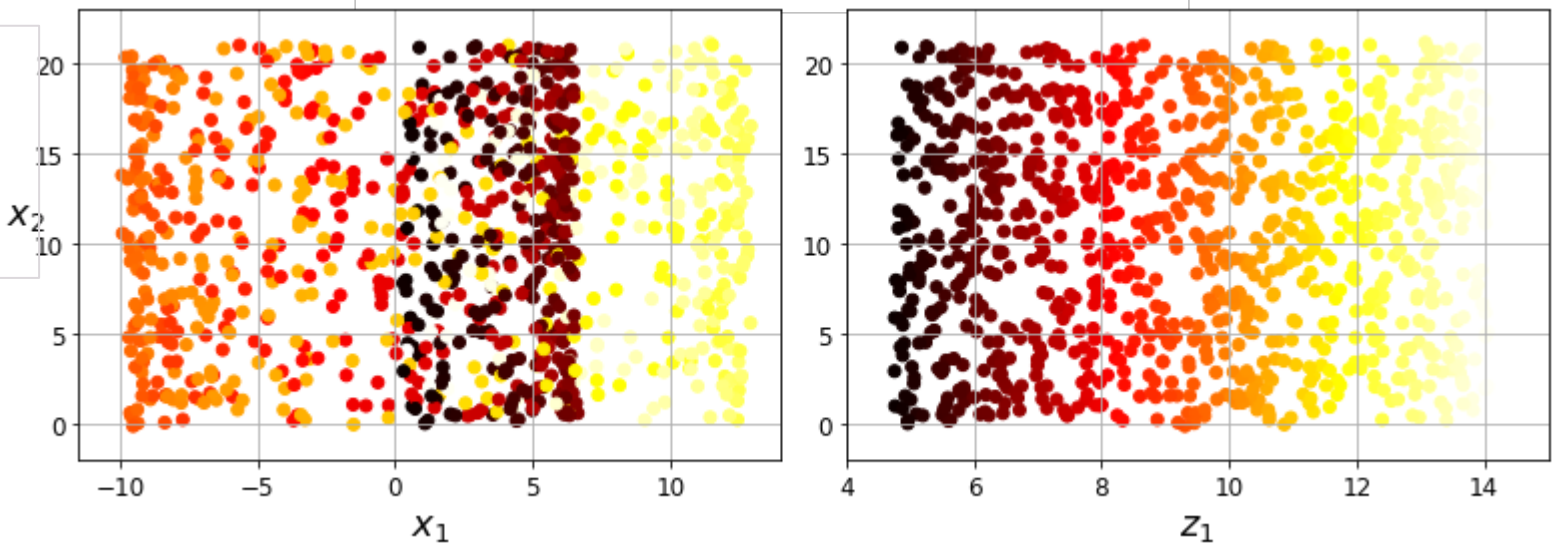


스위스 롤



투영

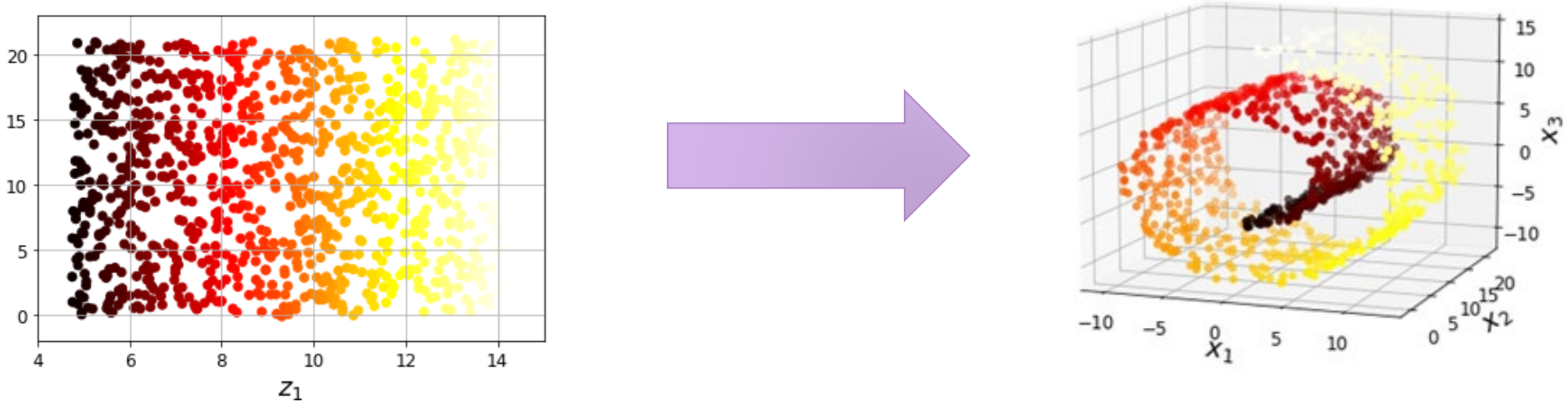
예 : 평면 X3 = 5에 투영

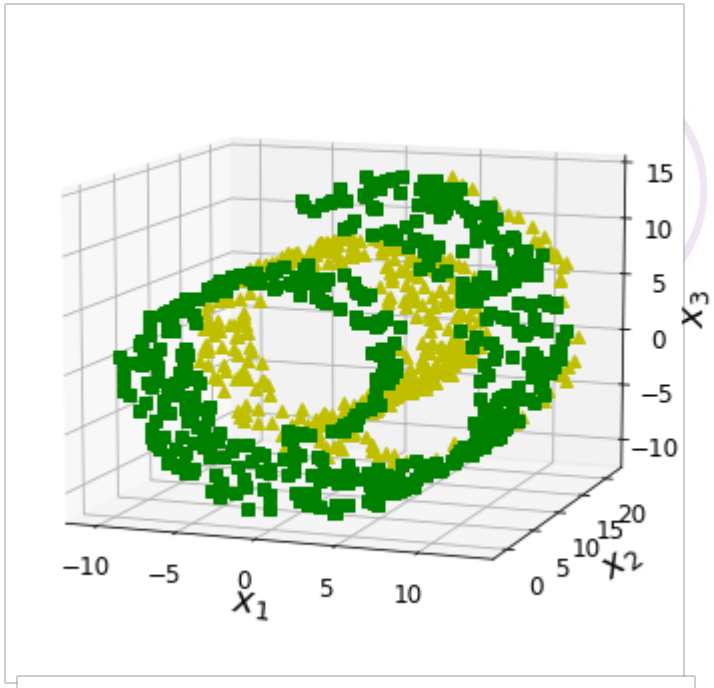


매니폴드

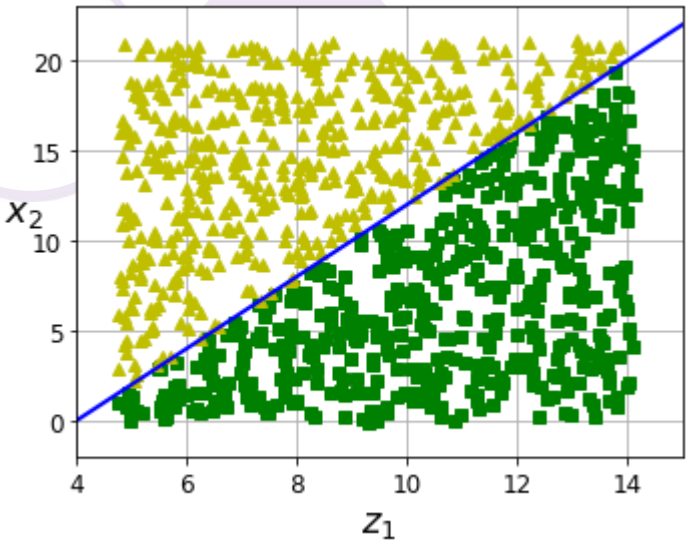
다양체多樣體 (manifold)

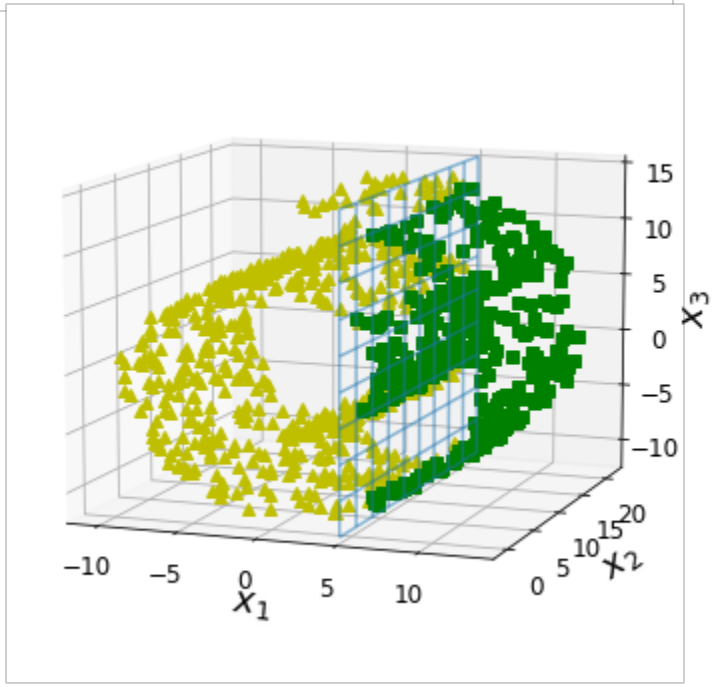
- 어디든 국소좌표계를 그릴 수 있는 공간
- 국소좌표계 : 공간의 제한된 범위에 그려진 좌표계
- 매니폴드 가정 : 대부분 실제 고차원 데이터셋이 더 낮은 차원의 다양체에 가깝게 놓여있다.





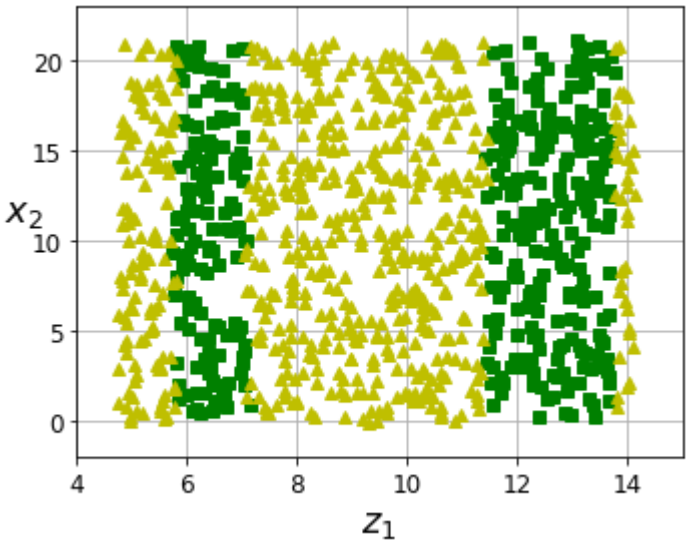
매니폴드 적절





매니폴드 부적절

3D 경계면, 투영

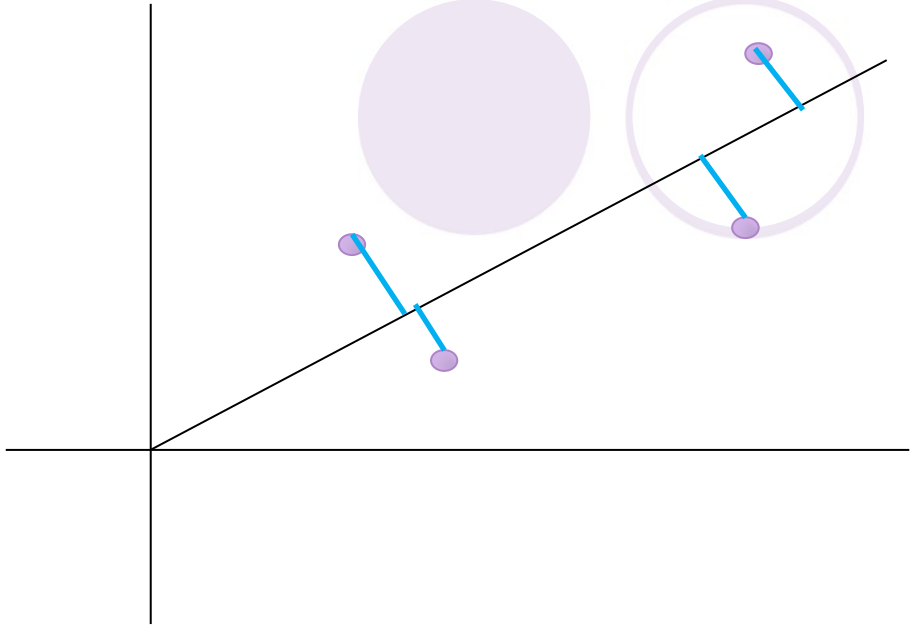


주성분 분석

- 주성분 분석은 분산을 최대로 보존하는 축을 찾는다.

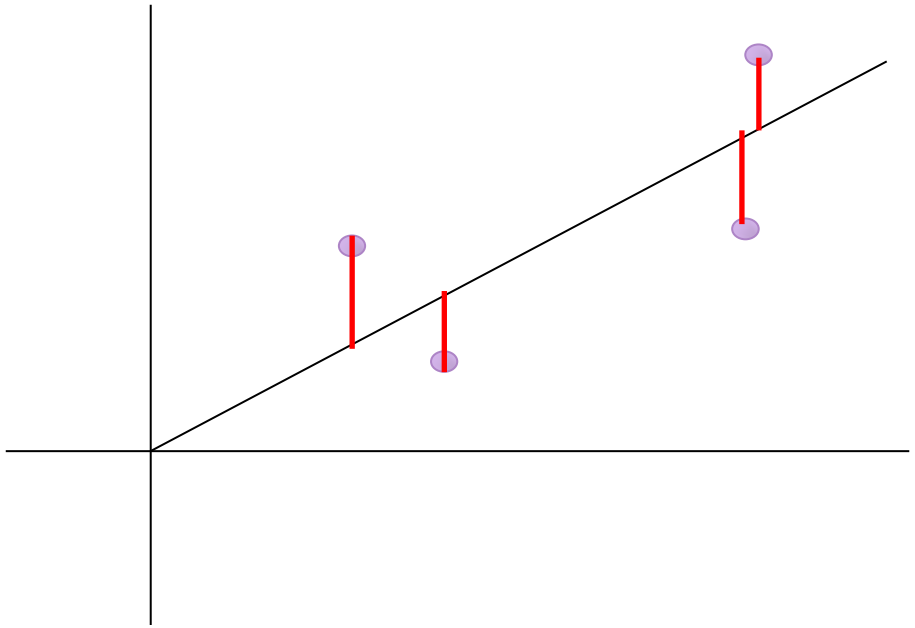
- 분산을 최대로 보존

⇔ 원본 데이터셋과 투영된 것 사이의 평균 제곱 거리를 최소화



분산 보존

: 파란색 선분 길이의 제곱의 평균을 최소화



회귀

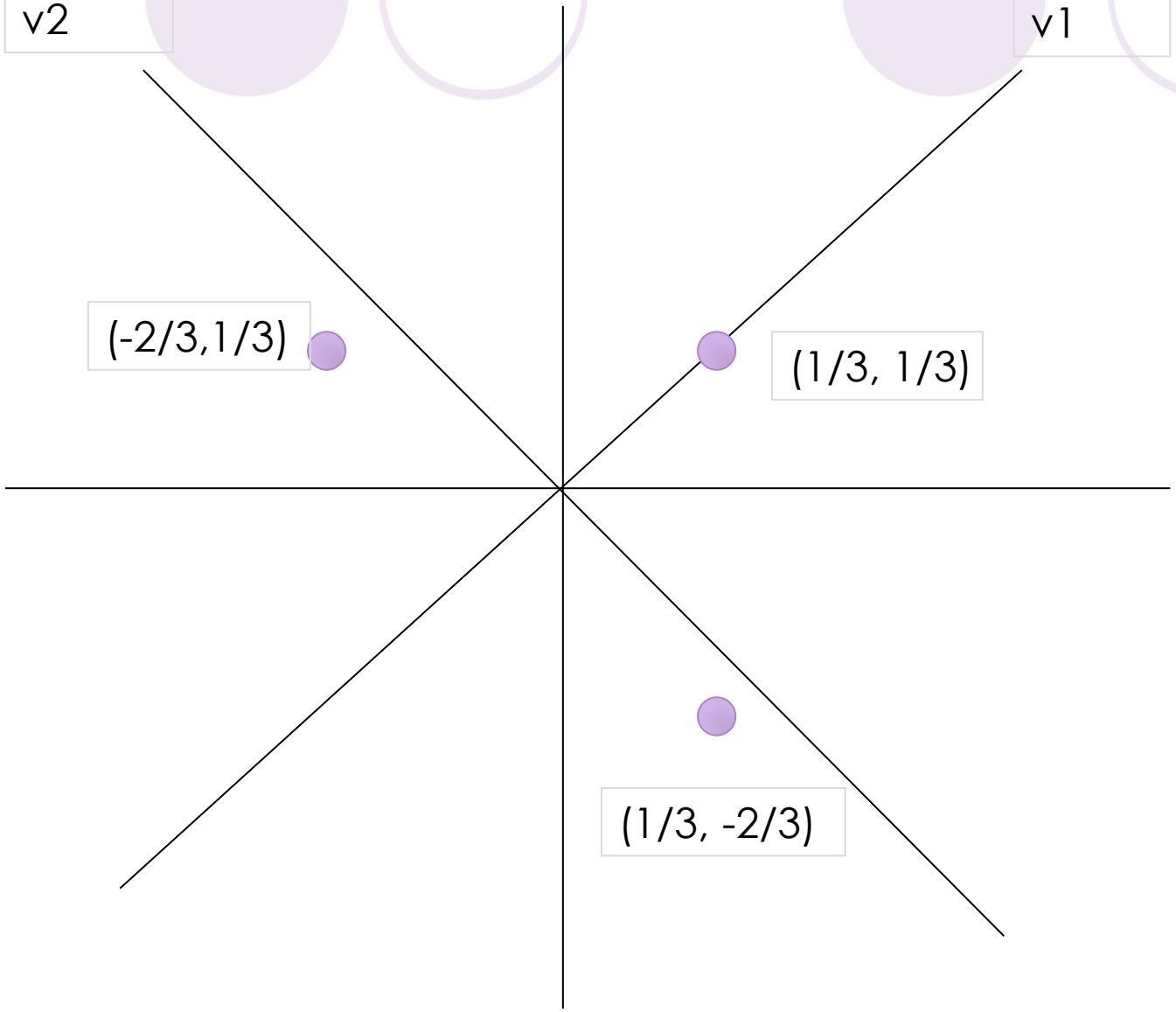
: 빨간색 선분 길이의 제곱의 평균을 최소화

주성분 찾기 (SVD)

	특성 1	특성 2
데이터 1	1	1
데이터 2	0	1
데이터 3	1	0

$$U\Sigma V^T$$

$$\begin{pmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}$$



V1 : 1

V2 : 11/9