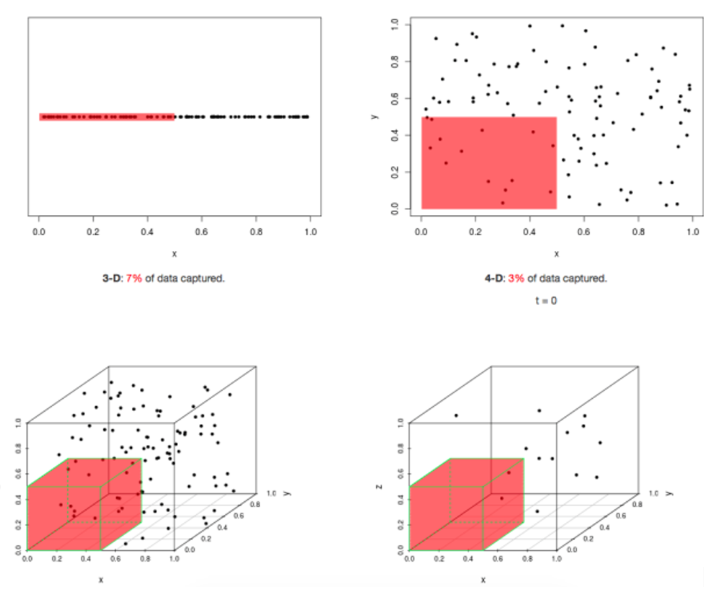
차원 축소

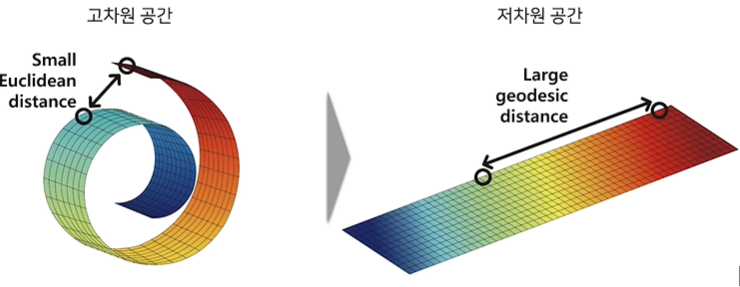
**차원의 저주** – 샘플 수가 늘어날수록 학습이 어려워지며 고차원 공간의 사상이 어려움.



고차원 데이터셋의 경우, 학습 난이도가 증가함. 이에 대비해 특성 수를 줄임으로써 차원 축소시켜 학습 가능하도록 함. 차원이 커졌지만 데이터가 부족할 경우 차원의 일부 공간에서 제한되어 소수의 데이터만 설명하기 때문에 과적합이 발생. – {데이터 유실, 데이터 시각화}

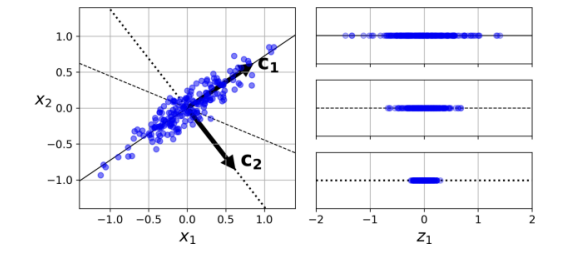
**투영** – n>d 일 때, n차원 공간의 존재하는 d차원 공간으로 투영해 표현

**매니폴드 학습** – n차원의 데이터를 n-1차원의 평면으로 축소하는 학습법으로 데이터를 잘 아우르는 sub-space(manifold)를 구하고 데이터를 투영시킴으로 차원 축소를 구현한다.



차원 축소 진행 시 분산 고려사항

**분산보존** – 투영 시에 가장 분산이 큰 축을 선택해야 데이터 손실이 가장 적음.



**PCA (주성분 분석)** – 고차원 데이터를 낮은 데이터로 바꿔주는 것으로 이 때 중요한 것은 어떻게 차원을 잘 낮추는지이다. 2차원의 데이터를 1차원의 데이터로 바꾼다고 하였을 때 최대한 특징을 살리며 차원을 낮추는 방법이 PCA 분석이다.

그렇다면 언제 주성분 분석을 사용해야 할까? 주성분 분석의 본질은 탐색적 분석이다. 즉, 변인을 탐색해서 변환을 통해 주성분을 결정하는 방법이다. 주성분은 독립 변수들의 분산을 가장 잘 설명하는 성분이다. 하나의 변수는 하나의 차원을 뜻한다. 그렇다면 4개의 독립변수는 하나의 공간으로 표현되기 위해서 4차원으로 나타낼 수 있다. 즉 변수가 하나씩 추가될수록 축이 하나씩 더 생긴다. 차원이 증가하게 될수록 표현해야 하는 공간은 복잡해질 것이고, 그렇게 된다면 공간이 불안정해지고 신뢰구간이 넓어지게 된다. 따라서 주성분 분석은 전체 데이터의 분산을 가장 잘 설명하는 축의 개수를 선정해서 그 축에 따라 변환된 데이터를 배열한다.

**차원축소를 하는 이유는?**

고차원으로 구성되어 있는 데이터셋은 시각화에 큰 어려움을 겪기 때문에 차원축소를 진행함으로 데이터 패턴을 인지하고 3차원 이내의 데이터로 시각화를 진행하기 위함이다.

또한, 특성을 줄이는 과정에 있어 중요한 특성만 남게 되므로 분산을 유지하면서도 불필요한 노이즈를 크게 줄이며 모델의 퍼포먼스 향상에 크게 기여하게 된다.

전반적으로 모델의 성능을 향상시키는 방향으로 메모리 사용률 감소, 데이터 압축, 시각화로 인한 데이터 패턴 분석 용이를 일궈낸다.

다중공정성 문제는 변수들 간의 상관관계가 높을 때 문제가 발생한다. 변수들 간의 상관 관계가 높으면 분석의 부정적인 영향을 끼친다. 회귀 분석에서 어떤 변수의 영향력을 파악할 때 설명 변수들끼리 서로 독립이라고 가정을 하고 한다. 그런데 두 설명 변수가 서로에게 큰 영향을 끼치고 있다면 각각의 설명 변수들이 종속 변수에 대해서 얼마나 영향을 미치고 있는지 알 수가 없게 되고 그로 인해 설명 변수의 설명력이 낮아진다. PCA 알고리즘은 주성분 PC1과 PC2를 찾는 과정에서 두 변수가 직교해야 하므로, 두 변수 사이의 상관관계가 0으로 나타나 다중공정성 문제를 해결할 수 있다.

주성분 분석의 개념이 복잡해 보여서 어떻게 활용이 되는지 의구심을 가지곤 했는데 실제로 데이터 분석을 진행할 때 굉장히 많이 사용을 한다. 먼저 데이터에 대한 정보를 확인하고 데이터 탐색 분석을 진행을 하면서 상관분석을 많이 사용한다. 특히 회귀 분석에서 가장 영향력이 높은 독립변수를 몇 개만 선정할 때 주성분 분석을 통해서 변수들을 추출하고 선정된 변수들로 모델을 구현하게 된다. 평소에 데이터 분석을 할 경우 주성분 분석을 해야 된다는 것만 알고 개념이나 원리에 대해서 알지 못하였는데 이번 주성분 분석 수업을 통해서 왜 주성분 분석을 해야 하는지 주성분 분석이 무엇이고 차원의 저주는 또 무엇인지 코드는 어떻게 진행하는지에 대해 배울 수 있어서 매우 유익했다. 또한 PCA알고리즘 뿐만 아니라 커널 PCA, 점진적 PCA, LLE 등 새로운 개념들도 배울 수 있었다.