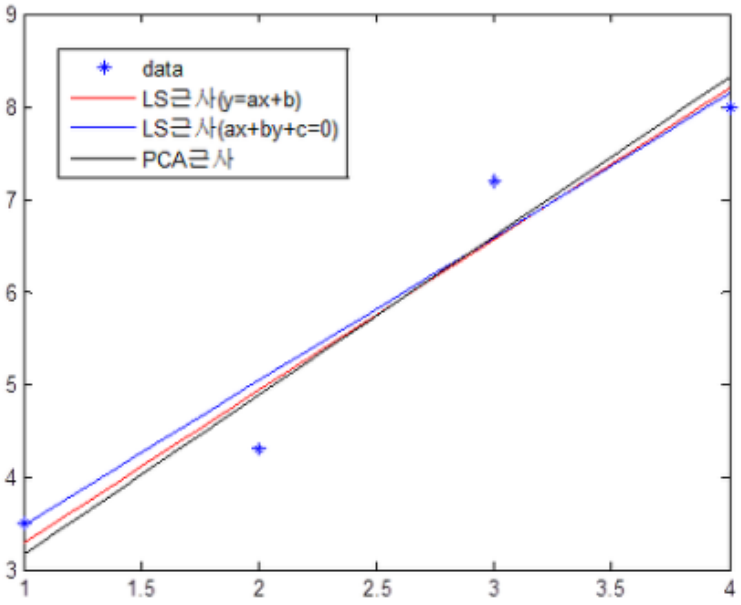

<그림 3> PCA 라인 근사

PCA로 직선을 근사하는 방법은 데이터들의 평균 위치를 지나면서 PCA로 나온 제 1 주성분 벡터와 평행인 직선을 구하면 된다.

실제로 위 네 점을 지나는 직선을 PCA로 구해보면 y = 1.7194x + 1.4515가 나오며 그 그래프는 <그림 3>과 같다. 그림에서 보듯이 PCA로 구한 직선과 최소자승법(LS 방법)으로 구한 직선이 모두 다름을 볼 수 있다. 그 이유는 최소자승법은 직선과 데이터와의 거리를 최소화하는 반면 PCA는 데이터의 분산이 가장 큰 방향을 구하기 때문이다.

~~계산적인 측면에서 보면 PCA는 최소자승법(LS)에 비해 훨씬 효율적이다. 왜냐하면 데이터의 차원이 n, 데이터의 개수가 m개일 때 LS는 n×m 행렬의 의사역행렬(pseudo inverse)를 계산해야하지만 PCA는 n×n 행렬의 고유값 분해만 계산하면 되기 때문이다 (2차원 평면에서 1,000개의 점을 근사하는 경우를 생각해 보면 LS는 2 x 1,000의 역행렬을 계산해야 하고 PCA는 2 x 2 행렬의 고유값 분해만 하면 된다).~~ => 잘못된 설명으로 삭제함 (2015.12.29 댓글 참조)

☞ 그림 3에서 LS근사(y=ax+b)는 직선과의 y축 거리를 최소화시키고, LS근사(ax+by+c=0)는 평면 z = ax+by+c와의 z축 거리를 최소화시킨다 (평면 z = ax+by+c과 xy평면과의 교선이 ax+by+c=0). PCA는 데이터들의 평균점을 지나는 직선들 중에서 데이터들을 직선에 투영(projection)시켰을 때 해당 직선을 따라서 데이터의 분산이 최대가 되는 방향의 직선을 구한다.