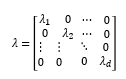
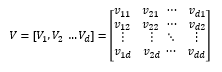
**숙제: PCA에서 압축을 위한 최적 eigenvector 개수 구하기**

1. Eigenvector의 개수는 원래 데이터의 차원(dimension)과 같으므로 모든 eigenvector를 이용하여 원래 데이터를 투영하면 데이터 압축의 효과를 볼 수 없다. 또한 데이터의 분산 특성을 고려하지 않고 특정 몇 개의 eigenvector를 이용하여 투영시키면 데이터의 유용한 정보가 손실될 수 있다. 유용한 정보를 손실하지 않으면서 최대로 압축 효과를 얻을 수 있도록 eigenvector의 개수를 설정할 수 있는 방법이 있는 지 설명해 보자.  
     
   **=>** N dimension에는 N개의 eigenvector가 존재하는데, 여기서 어떤 eigenvector를 선택하는가에 따라 데이터의 압축 결과가 달라진다. 데이터의 유용한 정보를 손실하지 않으려면 분산이 큰 eigenvector를 선택해야 한다. 즉, 분산과 비례하는 eigenvalue가 큰 eigenvector를 선택해야한다.  
     
   실습에서, eigenvalue (lambda)는 다음과 같다.  
     
     
   또한 eigenvector는 (V) 다음과 같다.  
     
     
   위에서 전체 분산은 lambda matrix의 trace 값 (lambda(1) + lambda(2) + … + lambda(d))이 된다. 여기서 선택된 eigenvector에 상응하는 eigenvalue들을 모두 합했을 때 전체 분산에서 높은 비율(90~95% 정도) 이상을 차지해야 한다. 이렇게 lambda 값들을 선택하면 데이터를 압축했을 때 데이터 손실을 최소화할 수 있다. 만약 선택된 lambda가 lambda(k), lambda(k+1), … 이라면 (lambda(k)+lambda(k+1)+….)/tr(lambda) >= 0.9 가 되게 해야 한다.