Machine Learning (CSE 446):
Principal Component Analysis and
The Singular Value Decomposition

Sham M Kakade
© 2019

University of Washington
cse446-staff@cs.washington.edu
PCA formulation 1: Dimension of Greatest Variance

Let \( \mathbf{u} \) be the dimension of greatest variance, and (without loss of generality) let \( \| \mathbf{u} \|_2^2 = 1 \).

- \( \mathbf{x}_i \cdot \mathbf{u} \) is the projection of the \( n \)th example onto \( \mathbf{u} \).
- Since the mean of the data is \( 0 \), the mean of \( \langle p_1, \ldots, p_N \rangle \) is also \( 0 \).
- This implies that the variance of \( \langle p_1, \ldots, p_N \rangle \) is \( \frac{1}{N} \sum_{i=1}^{N} p_i^2 \).
- The \( \mathbf{u} \) that gives the greatest variance, then, is:

\[
\arg\max_{\mathbf{u}} \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i \cdot \mathbf{u})^2 \\
\text{s.t. } \| \mathbf{u} \|_2^2 = 1
\]

(This is PCA in one dimension!)
The optimization problem, in terms of matrices

- $X$ is $N \times d$ data matrix.

$$\arg\max_u \|Xu\|_2^2$$

s.t. $\|u\|_2^2 = 1$

- The covariance matrix (assuming mean is subtracted):

$$\Sigma = \frac{1}{N} X^\top X = \frac{1}{N} \sum_{i=1}^{N} x_i x_i^\top$$

and, equivalently,

$$\arg\max_u u^\top \Sigma u$$

s.t. $\|u\|_2^2 = 1$
Deriving the Solution

(“relax”. not as difficult as it looks

\[
\max_u u^\top \Sigma u, \text{ s.t. } \|u\|_2^2 = 1
\]

- The Lagrangian encoding: “relax” the constraint into the objective:

\[
\max_u u^\top \Sigma u - \lambda \|u\|_2^2
\]

- \(\lambda\) provides a ‘soft’ penalty
- (no we just do calculus). first derivatives with respect to \(u\): \(2\Sigma u - 2\lambda u\)
- Setting equal to 0 leads to: \(\lambda u = \Sigma u\)
- (maybe your recognize it?) this is the def. of an eigenvector (\(u\)) and eigenvalue (\(\lambda\)) for the matrix \(\Sigma\).
- We take the first (largest) eigenvalue.
The solution.

- This is the 'eigenvalue' problem. We know the solution must satisfy:

\[ \Sigma u = \lambda u \]

for some \( \lambda \).

- This means \( u \) is an eigenvector and \( \lambda \) is an eigenvalue of \( \Sigma \).

- So to solve the optimization problem the eigenvector \( u \) corresponding to the largest eigenvalue \( \lambda \).

- How do we find this?
  The Singular Value Decomposition.
Alternate View of PCA: Minimizing Reconstruction Error

Assume that the data are centered.
Find a line which minimizes the squared reconstruction error.
Alternate View of PCA: Minimizing Reconstruction Error

Assume that the data are centered.
Find a line which minimizes the squared reconstruction error.
Alternate View: Minimizing Reconstruction Error with $K$-dim subspace.

Equivalent ("dual") formulation of PCA: find an "orthonormal basis" $u_1, u_2, \ldots, u_K$ which minimizes the total reconstruction error on the data:

$$\arg\min_{\text{orthonormal basis: } u_1, u_2, \ldots, u_K} \frac{1}{N} \sum_i \|x_i - \text{Proj}_{u_1, \ldots, u_K}(x_i)\|^2$$

Recall the projection of $x$ onto $K$-orthonormal basis is:

$$\text{Proj}_{u_1, \ldots, u_K}(x) = \sum_{j=1}^{K} (u_i \cdot x) u_i$$

The SVD "simultaneously" finds all $u_1, u_2, \ldots, u_K$
Principal Components Analysis: the algorithm

- Input: unlabeled data $X = [x_1 | x_2 | \cdots | x_N]^\top$; dimensionality $K < d$
- Output: $K$-dimensional “subspace”.
- Algorithm:
  1. Compute the mean $\mu$
  2. Compute the covariance matrix:
     \[
     \Sigma = \frac{1}{N} \sum_i (x_i - \mu)(x_i - \mu)^\top
     \]
  3. Let $\langle \lambda_1, \ldots, \lambda_K \rangle$ be the top $K$ eigenvalues of $\Sigma$ and $\langle u_1, \ldots, u_K \rangle$ be the corresponding eigenvectors
- Let $\tilde{U} = [u_1 | u | \cdots | u_K]$
- Return $\tilde{U}$

You can read about many algorithms for finding eigendecompositions of a matrix.
Projection and Reconstruction: the one dimensional case

- Take out mean $\mu$:
- Find the “top” eigenvector $u$ of the covariance matrix.
- What are your projections?

- What are your reconstructions, $\hat{X} = [\hat{x}_1 | \hat{x}_2 | \cdots | \hat{x}_N]^{\top}$?

- What is is your reconstruction error?

$$\frac{1}{N} \sum_i (x_i - \hat{x}_i)^2 = ?$$
The singular value decomposition

- Let $M$ be a symmetric matrix. SVDs also work for asymmetric matrices, with a slightly modified thm.
- SVD theorem: there exists a decomposition of the following form:

$$M = U D U^\top$$

where $D$ is a diagonal matrix and $U$ is an orthogonal matrix (i.e. the columns of $U$ are unit length and orthogonal to each other).
- The columns of $U$ are eigenvectors of $M$.
- For PCA, you will take $\Sigma$ to be $M$. 