1 Introduction

- Principal component analysis (PCA) reduces the dimensionality of a data set by finding a new set of variables, smaller than the original set of variables. It retains most of the sample’s variance.
- It is useful for the compression and classification of data.
- The new variables, called principal components (PCs), are uncorrelated, and are ordered by the fraction of the total information each retains.

2 Computation of PCs

- Let $A$ be a $n \times m$ data matrix in which each row is a data vector, each column represents a variable. $A$ is centered: the estimated mean is subtracted from each row, so $A$ is of zero row mean.
- Let $w$ be the column vector in $\mathbb{R}^m$ of projection weights that result in the largest variance when the data $A$ is projected along $w$. We require $w^T w = 1$.
- Projection of a vector $u$ onto $w$ is $w^T u = \sum_{j=1}^{m} u_j w_j$. Thus, projection of data along $w$ is $Aw$.
- The variance of the data after the projection by $w$ is $(Aw)^T (Aw) = w^T A^T A w = w^T C w$ where $C = A^T A$ is the covariance matrix of the data (note that $A$ is centered).
- Compute the optimal projection $w$, which maximizes the variance subject to constraint $w^T w = 1$.
- Define the following Lagrange function:
  \[ f = w^T C w - \lambda (w^T w - 1) \]
  where $\lambda$ is the Lagrange multiplier.
- Take the directive of $f$ in terms of $w$ and set it to zero, we get
  \[ \frac{\partial f}{\partial w} = 2Cw - 2\lambda w = 0. \]
  This leads to the eigenvalue problem: $Cw = \lambda w$, where $C = A^T A$.
- Show that the solution is given by the singular values and singular vectors of $A$. 
• We aim to maximize the variance, which is given by

\[ w^T C w = w^T \lambda w = \lambda w^T w = \lambda. \]

Choose \( \lambda = \sigma_1^2 \). More precisely: the first principal component of \( A \) is exactly the first right singular vector \( v_1 \) of \( A \).

• Once the first principal component is found, we continue in the same fashion to look for the next one, which is orthogonal to (all) the principal component(s) already found.
  - The solutions are the right singular vectors \( v_k \) of \( A \), and the variance in each direction is given by the corresponding singular values \( \sigma_k \).

• It is natural to first form the covariance matrix \( C = A^T A \) of the centered data matrix \( A \) for computing the PCs. However, this is not a good idea. Why?
  - The condition number of \( A^T A \) is much larger than that of \( A \).
  - For a sparse \( A \), \( C = A^T A \) may not be sparse any more.

• How to compute the PCA:
  - Center the data by subtracting the mean of the rows.
  - Compute the SVD of the centered data matrix \( A \) as \( A = U \Sigma V^T \).
  - The principal components are the columns of \( V \), the coordinates of the data in the basis defined by the principal components are \( U \Sigma \).

3 How to choose the number of PCs

• The variance in the direction of the \( k \)-th principal component is given by the corresponding singular value: \( \sigma_k^2 \).

• Singular values can be used to estimate how many principal components to keep.

• Rule of thumb: keep enough to explain 85% of the variation:

\[
\frac{\sum_{j=1}^k \sigma_j^2}{\sum_{j=1}^n \sigma_j^2} \approx 0.85.
\]

• Typically, the squared singular values drop rapidly. Thus, the first few principal components are enough to capture most of the variation in the data. This leads to data compression.

4 Summary

• PCA is SVD done on centered data.

• PCA looks for such a direction that the data projected onto it has maximal variance.

• When found, PCA continues by seeking the next direction, which is orthogonal to all the previously found directions, and which explains as much of the remaining variance in the data as possible.

• PCA is useful for data exploration, visualizing data, compressing data, outlier detection, and etc.