# 1 Principal Component Analysis

## Computing PCA

Here is the full procedure for computing the principal components of a data-set represented by the matrix $A$:

1. **Compute the covariance matrix**:
   - Center the matrix along the attributes (columns in this case), so that $\tilde{A} = A - \frac{1}{m}\vec{1}\vec{1}^T A$.
   - Find the covariance matrix $S = \frac{1}{m}\tilde{A}^T\tilde{A}$ or $S = \frac{1}{m-1}\tilde{A}^T\tilde{A}$ depending on whether your data is a sample from a population or the full population.
2. **Diagonalize the covariance matrix** $S$:
   - Diagonalize the covariance matrix: $S = P\Lambda P^T$.
   - Since $S$ is a symmetric matrix, we know $P^T = P^{-1}$.
   - The columns of $P$ are the principal components.

**Result**: The result of this process is the matrix $P$ consisting of the principal components. Changing the data into this basis can reveal many useful insights about the data.

# 2 Understanding PCA

What is this PCA basis, and why is it useful? There are two equivalent ways of understanding PCA:

1. The principal component vectors $\vec{p}_i$ maximize the variance of data when it is projected on the line corresponding to $\vec{p}_i$.
2. The subspace spanned by $k$ PC vectors (i.e. span$\{\vec{p}_1, \dots, \vec{p}_k\}$ ) is the best $k$-dimensional approximation to the given data.

**Variance Maximization perspective**: Given a unit vector $\vec{u}$, we can compute the projection of the data along the line defined by $\vec{u}$ by computing $\tilde{A}\vec{u}$. The variance of the projection is formally given by $\frac{1}{m}\vec{u}^T\tilde{A}^T\tilde{A}\vec{u}$. The main idea behind PCA is to find trends in the data by choosing the $\vec{w}$ that maximizes this variance.

- Therefore, $\vec{p}_1$ is simply the unit vector that maximizes $\vec{w}^T\tilde{A}^T\tilde{A}\vec{w}$ .
- $\vec{p}_2$ is the unit vector that maximizes $\vec{w}^T\tilde{A}^T\tilde{A}\vec{w}$ while **also** being orthogonal to $\vec{p}_1$. Note that the possible choices of $\vec{p}_2$ are more restricted due to orthogonality constraint. This constraint ensures that we get an orthonormal basis.
- $\vec{p}_k$ is the unit vector that maximizes $\vec{w}^T\tilde{A}^T\tilde{A}\vec{w}$ while **also** being orthogonal to $\vec{p}_1, \dots, \vec{p}_{k-1}$.

Formally, we can write this as:

$$\vec{p}_k = \max_{w} \vec{w}^T\tilde{A}^T\tilde{A}\vec{w}$$

$$\text{subject to } \vec{w}^T\vec{w} = 1, \vec{w}^T\vec{p}_1 = 0, \vec{w}^T\vec{p}_2 = 0, \dots, \vec{w}^T\vec{p_{k-1}} = 0$$

It turns out that the solutions to this optimization problem are the eigenvectors of $\tilde{A}^T\tilde{A}$, i.e. the eigenvectors of the covariance matrix!

**Reconstruction Error Minimization perspective**: Let's say we want to compress our high-dimensional data by approximating each sample point $\vec{x}_i$ as a linear combination of a small number of vectors $\vec{v}_1, \dots, \vec{v}_k$,

That is, $\vec{x}_i \approx \hat{\alpha}_1\vec{v}_1 + \cdots + \alpha_k\vec{v}_k$ . Then $\vec{x}_i$ can be (approximately) described by the coefficients $\alpha_1, \dots, \alpha_k$, and we can just store the $\alpha$'s instead of storing every attribute of $\vec{x}_i$. These $\alpha$'s can be found by solving a least sqaures problem $V\vec{\alpha} = \vec{x}_i$. The question then is: which vectors $\vec{v}_1, \dots, \vec{v}_k$ will allow us to make the best reconstructions? These vectors have to minimize the total error in reconstructing the data points $\vec{x}_i$. It can be proven that the first $k$ vectors of the PCA basis $\{\vec{p}_1, \dots, \vec{p}_k\}$ are the best choice of vectors to use.

This makes PCA an excellent choice of basis when compressing data. Intuitively, the first few vectors in the basis explain most of the variation in the data, and so we can get away with just storing the first $k$ coefficients.

## 3 PCA

Suppose we had the following data points $(x_i, y_i) \in \mathbb{R}^2$ aggregated in the following matrix.

$$A = \begin{bmatrix} 5 & -6 \\ 7 & 0 \\ 11 & -4 \\ 5 & -6 \end{bmatrix}$$

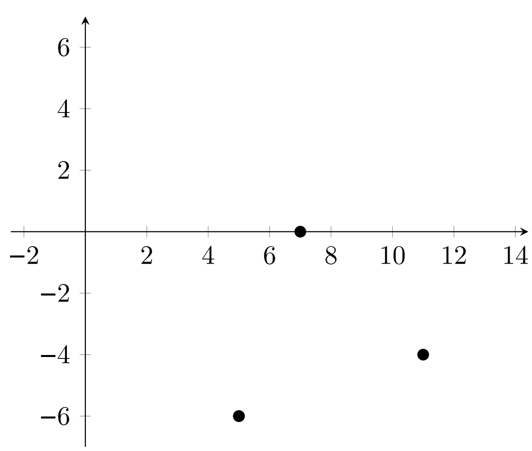
a) Find the covariance matrix $S$ of $A$ if each column is a type of data and each row is a measurement.

b) Since $S$ is a symmetric matrix, we can eigendecompose it into the form $S = P\Lambda P^T$, where $P$ contains the orthonormal principal components of $S$ and $\Lambda$ is a diagonal matrix with the squared weights of the corresponding principal components. Find the eigenvalues of $S$ and order them from largest to smallest, $\lambda_1 > \lambda_2$.

c) Find the orthonormal eigenvectors $\vec{p}_i$ of $S$ (all eigenvectors are mutually orthogonal and have unit length).

d) What are the principal components of the system? What are the weights of each principal component?

e) Plot the two principal components scaled by their weights on the following graph. Remember that we subtracted the column means from each column.



## 4 Using the SVD for PCA

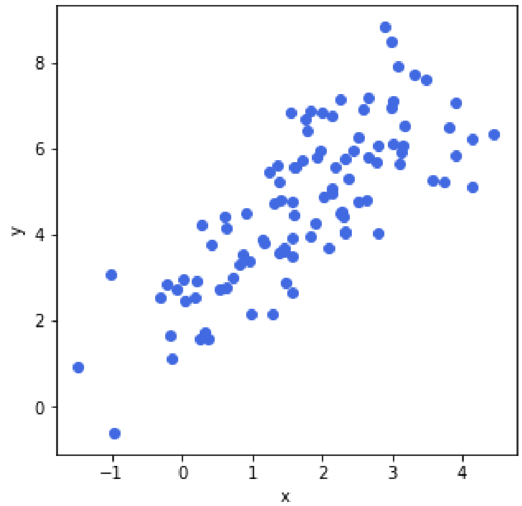
In the previous question, we viewed the principal components as the eigenvectors of the covariance matrix $S = \frac{1}{m}\tilde{A}^T\tilde{A}$. In this question, we see how Principal Component Analysis relates to the Singular Value Decomposition.

a) Given $m$ data points $(x_i, y_i)_{i=1}^m$ in $\mathbb{R}^2$, what is our data matrix $A$?

b) If the SVD of $M$ is $M = U\Sigma V^T$, what are the eigenvectors and eigenvalues of $M^T M$?

c) How can we use the SVD to compute our principal components?

d) If the given data looked like the following figure, what would you expect $\sigma_1\vec{v}_1$ and $\sigma_2\vec{v}_2$ to be?




e) Sketch the projection of the demeaned data onto the principal components $\vec{v}_1$ and $\vec{v}_2$.