LECTURE 16: PCA AND SVD

Resource:
- PCA Slide by Iyad Batal
- Chapter 12 of PRML
- Principal Component Analysis (PCA)
- Singular Value Decomposition (SVD)
PRINCIPLE COMPONENT ANALYSIS

- PCA finds a **linear** projection of high dimensional data into a lower dimensional subspace such as:
  - The variance retained is maximized.
  - The least square reconstruction error is minimized.
PCA STEPS

Linearily transform an $N \times d$ matrix $X$ into an $N \times m$ matrix

- Centralize the data (subtract the mean).
- Calculate the $d \times d$ covariance matrix: $C = \frac{1}{N-1} X^T X$
  
  $C_{i,j} = \frac{1}{N-1} \sum_{q=1}^{N} X_{q,i}X_{q,i}$
  
  + $C_{i,i}$ (diagonal) is the variance of variable $i$.
  + $C_{i,j}$ (off-diagonal) is the covariance between variables $i$ and $j$.
- Calculate the eigenvectors of the covariance matrix (orthonormal).
- Select $m$ eigenvectors that correspond to the largest $m$ eigenvalues to be the new basis.
If \( A \) is a square matrix, a non-zero vector \( \mathbf{v} \) is an eigenvector of \( A \) if there is a scalar \( \lambda \) (eigenvalue) such that

\[
A \mathbf{v} = \lambda \mathbf{v}
\]

Example:

\[
\begin{pmatrix}
2 & 3 \\
2 & 1
\end{pmatrix}
\begin{pmatrix}
3 \\
2
\end{pmatrix}
= 
\begin{pmatrix}
12 \\
8
\end{pmatrix}
= 4 \begin{pmatrix}
3 \\
2
\end{pmatrix}
\]

If we think of the squared matrix \( A \) as a transformation matrix, then multiply it with the eigenvector do not change its direction.
PCA EXAMPLE

$X$ : the data matrix with $N=11$ objects and $d=2$ dimensions
Step 1: subtract the mean and calculate the covariance matrix $C$.

\[
C = \begin{pmatrix}
0.716 & 0.615 \\
0.615 & 0.616
\end{pmatrix}
\]
Step 2: Calculate the eigenvectors and eigenvalues of the covariance matrix:

\[ \lambda_1 \approx 1.28, \ v_1 \approx [-0.677 \ -0.735]^T, \ \lambda_2 \approx 0.49, \ v_2 \approx [-0.735 \ 0.677]^T \]

Notice that \( v_1 \) and \( v_2 \) are orthonormal:

\[
|v_1| = 1 \\
|v_2| = 1 \\
v_1 \cdot v_2 = 0
\]
Step 3: project the data

Let \( V = [v_1, \ldots, v_m] \) is \( d \times m \) matrix where the columns \( v_i \) are the eigenvectors corresponding to the largest \( m \) eigenvalues.

The projected data: \( Y = X V \) is \( N \times m \) matrix.

If \( m = d \) (more precisely \( \text{rank}(X) \)), then there is no loss of information!
Step 3: project the data

\[ \lambda_1 \approx 1.28, \mathbf{v}_1 \approx [-0.677, -0.735]^T, \lambda_2 \approx 0.49, \mathbf{v}_2 \approx [-0.735, 0.677]^T \]

- The eigenvector with the highest eigenvalue is the principle component of the data.
- If we are allowed to pick only one dimension, the principle component is the best direction (retain the maximum variance).
- Our PC is \( \mathbf{v}_1 \approx [-0.677, -0.735]^T \)
USEFUL PROPERTIES

- The covariance matrix is always symmetric

\[ C = \left( \frac{1}{N - 1} X^T X \right)^T = \frac{1}{N - 1} X^T X^T = C \]

- The principal components of \( X \) are orthonormal

\[ v_i^T v_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \]

- \( V = [v_1, \ldots, v_m] \), then \( V^T = V^{-1} \), i.e. \( V^T V = I \)
**Theorem 1:** If square \( d \times d \) matrix \( S \) is a real and symmetric matrix \((S = S^T)\) then

\[
S = V \Lambda V^T
\]

Where \( V = [v_1, ..., v_d] \) are the eigenvectors of \( S \) and \( \Lambda = diag(\lambda_1, ..., \lambda_d) \) are the eigenvalues.

**Proof:**

- \( S V = V \Lambda \)
- \([S v_1 ... S v_d] = [\lambda_1. v_1 ... \lambda_d. v_d] \): the definition of eigenvectors.
- \( S = V \Lambda V^{-1} \)
- \( S = V \Lambda V^T \) because \( V \) is orthonormal \( V^{-1} = V^T \)
USEFUL PROPERTIES

- The projected data: $Y = X V$
- The covariance matrix of $Y$ is

$$C_Y = \frac{1}{N-1} Y^T Y = \frac{1}{N-1} V^T X^T X V = V^T C_X V$$

$$= V^T V \Lambda V^T V$$ because the covariance matrix $C_X$ is symmetric

$$= V^{-1} V \Lambda V^{-1} V$$ because $V$ is orthonormal

$$= \Lambda$$

*After the transformation, the covariance matrix becomes diagonal.*
DERIVATION OF PCA: 1. MAXIMIZING VARIANCE

- Assume the best transformation is one that maximize the variance of project data.

- Find the equation for variance of projected data.

- Introduce constraint

- Maximize the un-constraint equation. (find derivative w.r.t projection axis and set to zero)
DERIVATION OF PCA:

2. MINIMIZING TRANSFORMATION ERROR

- Define error
- Identify variables that need to be optimized in the error
- Minimize and solve for the variables
- Interpret the information
Any $N \times d$ matrix $X$ can be uniquely expressed as:

$$X = U \times \Sigma \times V^T$$

- $r$ is the **rank** of the matrix $X$ (# of linearly independent columns/rows).
  - $U$ is a **column-orthonormal** $N \times r$ matrix.
  - $\Sigma$ is a **diagonal** $r \times r$ matrix where the singular values $\sigma_i$ are sorted in descending order.
  - $V$ is a **column-orthonormal** $d \times r$ matrix.
PCA AND SVD RELATION

Theorem:
Let \( X = U \Sigma V^T \) be the SVD of an \( N \times d \) matrix \( X \) and
\[
C = \frac{1}{N-1} X^T X
\]
be the \( d \times d \) covariance matrix.

The eigenvectors of \( C \) are the same as the right singular vectors of \( X \).

Proof:
\[
X^T X = V \Sigma U^T U \Sigma V^T = V \Sigma \Sigma V^T = V \Sigma^2 V^T
\]
\[
C = V \frac{\Sigma^2}{N-1} V^T
\]
But \( C \) is symmetric, hence \( C = V \Lambda V^T \)
Therefore, the eigenvectors of the covariance matrix \( C \) are the same as matrix \( V \) (right singular vectors) and

the eigenvalues of \( C \) can be computed from the singular values \( \lambda_i = \frac{\sigma_i^2}{N-1} \)
The singular value decomposition and the eigendecomposition are closely related. Namely:

- The **left-singular vectors** of $X$ are eigenvectors of $XX^T$.
- The **right-singular vectors** of $X$ are eigenvectors of $X^TX$.
- The **non-zero singular values** of $X$ (found on the diagonal entries of $\Sigma$) are the square roots of the non-zero eigenvalues of both $X^TX$ and $XX^T$. 

$$X = U \times \Sigma \times V^T$$
ASSUMPTIONS OF PCA

- I. Linearity
- II. Mean and variance are sufficient statistics.
  - Gaussian distribution assumed
- III. Large variances have important dynamics.
- IV. The principal components are orthogonal
function [signals,PC,V] = pca1(data)

% PCA1: Perform PCA using covariance.
% data - MxN matrix of input data
% (M dimensions, N trials)
% signals - MxN matrix of projected data
% PC - each column is a PC
% V - Mx1 matrix of variances

[M,N] = size(data);

% subtract off the mean for each dimension
mn = mean(data,2);
data = data - repmat(mn,1,N);

% calculate the covariance matrix
covariance = 1 / (N-1) * data * data';

% find the eigenvectors and eigenvalues
[PC, V] = eig(covariance);

% extract diagonal of matrix as vector
V = diag(V);

% sort the variances in decreasing order
[junk, rindices] = sort(-1*V);
V = V(rindices);
PC = PC(:,rindices);

% project the original data set
signals = PC' * data;

function [signals,PC,V] = pca2(data)

% PCA2: Perform PCA using SVD.
% data - MxN matrix of input data
% (M dimensions, N trials)
% signals - MxN matrix of projected data
% PC - each column is a PC
% V - Mx1 matrix of variances

[M,N] = size(data);

% subtract off the mean for each dimension
mn = mean(data,2);
data = data - repmat(mn,1,N);

% construct the matrix Y
Y = data' / sqrt(N-1);

% SVD does it all
[u,S,PC] = svd(Y);

% calculate the variances
S = diag(S);
V = S .* S;

% project the original data
signals = PC' * data;