CS5489
Lecture 8.1: Principal Component Analysis

Kede Ma

City University of Hong Kong (Dongguan)

Slide template by courtesy of Benjamin M. Marlin

Outline

Eigenvectors

- Assume $\mathbf{A} \in \mathbb{R}^{N \times N}$, $\mathbf{v} \in \mathbb{C}^{N \times 1}$, and $\lambda \in \mathbb{C}$
- If $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ then $\mathbf{v}$ is a right eigenvector of $\mathbf{A}$ with eigenvalue λ
- If $\mathbf{A}^T\mathbf{v} = \lambda\mathbf{v}$ then $\mathbf{v}$ is a left eigenvector of $\mathbf{A}$ with eigenvalue λ (equivalently $\mathbf{v}^T\mathbf{A} = \lambda\mathbf{v}^T$)
- If $\mathbf{A}$ is symmetric so that $\mathbf{A} = \mathbf{A}^T$, then the left and right eigenvectors of $\mathbf{A}$ are the same with the same eigenvalues

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix} = 3 \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} = 3 \begin{bmatrix} 1 & 1 \end{bmatrix}$$

Linear Independence

- Linear independence is arguably the most important concept in linear algebra
- A set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N\}$ is linear independent if the vector equation

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_N\mathbf{v}_N = \mathbf{0}$$

  has only the trivial solution $a_1 = a_2 = \cdots = a_N = 0$
- $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N\}$ is linearly dependent if there exist numbers $a_1, a_2, \ldots, a_N$ not all equal to zero, such that

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_N\mathbf{v}_N = \mathbf{0}$$

  - Assuming $a_1 \neq 0$, we have $\mathbf{v}_1 = -\frac{a_2}{a_1}\mathbf{v}_2 - \cdots - \frac{a_N}{a_1}\mathbf{v}_N$

Some Matrices

- An $N \times N$ square matrix $\mathbf{A}$ is **invertible** if there exists an $N \times N$ square matrix $\mathbf{B}$ such that $\mathbf{AB} = \mathbf{BA} = \mathbf{I}_N$
  - Equivalently, the columns/rows of $\mathbf{A}$ are linearly independent
- A square matrix $\mathbf{Q}$ is **orthogonal** if its columns and rows are orthogonal unit vectors (orthonormal vectors)
  - I.e., $\mathbf{Q}\mathbf{Q}^T = \mathbf{Q}^T\mathbf{Q} = \mathbf{I}$
- A square matrix $\mathbf{A}$ is **diagonalizable** if there exists an invertible matrix $\mathbf{P}$ and a diagonal matrix $\mathbf{D}$ such that $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D}$, or equivalently $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$
- **Real symmetric** matrices are diagonalizable by orthogonal matrices
  - Can be proved using the Spectral Theorem

Eigendecomposition

- Let $\mathbf{V} \in \mathbb{R}^{N \times N}$ be a matrix whose columns $\mathbf{v}_i$ are N linearly independent eigenvectors of $\mathbf{A}$ with $\mathbf{\Lambda}$ the corresponding diagonal matrix of eigenvalues such that $\Lambda_{ii} = \lambda_i$.Then:

$$\mathbf{AV} = \mathbf{V\Lambda}$$

$$\mathbf{A} = \mathbf{V\Lambda V}^{-1}$$

$$\mathbf{V}^{-1}\mathbf{AV} = \mathbf{\Lambda}$$

- Only *diagonalizable* matrices have eigendecomposition

Eigendecomposition of a Symmetric Matrix

- If $\mathbf{A}$ is real symmetric, we can choose N orthonormal eigenvectors so that $||\mathbf{v}_i||_2^2 = 1$, $\mathbf{v}_i^T\mathbf{v}_j = 0$ and N real eigenvalues $\lambda_i \in \mathbb{R}$. As a result, we have

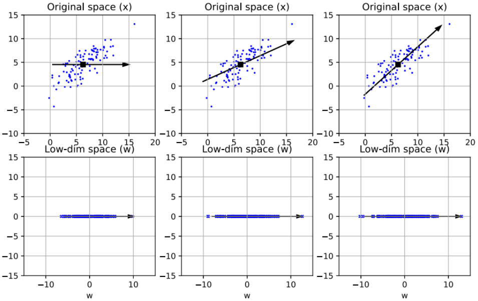
$$\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^T = \sum_{i=1}^{N} \lambda_i \mathbf{v}_i \mathbf{v}_i^T,$$

$$\mathbf{V}^T\mathbf{A}\mathbf{V} = \Lambda$$

Outline

Principal Component Analysis (PCA)

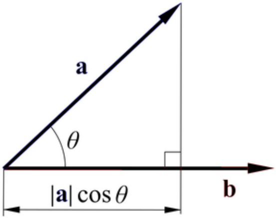
- Unsupervised method
- Given a data matrix $\mathbf{X} \in \mathbb{R}^{M \times N}$, the goal of PCA is to identify the directions of maximum variance contained in the data
  - Choose basis vectors along the maximum variance (longest extent) of the data
  - The basis vectors are called principal components (PC)



Sample Variance in a Given Direction

- Let $\mathbf{v} \in \mathbb{R}^N$ such that $||\mathbf{v}||_2^2 = \mathbf{v}^T\mathbf{v} = 1$
- The variance in the direction $\mathbf{v}$ is given by the expression:

$$\frac{1}{M}\sum_{i=1}^{M}(\mathbf{v}^T\mathbf{x}^{(i)} - \boldsymbol{\mu})^2, \quad \text{where} \quad \boldsymbol{\mu} = \frac{1}{M}\sum_{i=1}^{M}\mathbf{v}^T\mathbf{x}^{(i)}$$



https://www.mit.edu/~hlb/StantonGrant/18.02/details/tex/lec1snip2-dotprod.pdf

Pre-Centering

- Under the assumption that the data are pre-centered so that $\frac{1}{M}\sum_{i=1}^{M}\mathbf{x}^{(i)} = 0$, this expression simplifies to:

$$
\begin{aligned}
\frac{1}{M}\sum_{i=1}^{M}(\mathbf{v}^T\mathbf{x}^{(i)})^2 &= \frac{1}{M}\sum_{i=1}^{M}\left(\mathbf{v}^T\mathbf{x}^{(i)}\right)\cdot\left(\mathbf{v}^T\mathbf{x}^{(i)}\right) \\
&= \frac{1}{M}\sum_{i=1}^{M}\left(\mathbf{v}^T\mathbf{x}^{(i)}\right)\cdot\left((\mathbf{x}^{(i)})^T\mathbf{v}\right) \\
&= \frac{1}{M}\mathbf{v}^T\left(\sum_{i=1}^{M}\mathbf{x}^{(i)}(\mathbf{x}^{(i)})^T\right)\mathbf{v} \\
&= \frac{1}{M}\mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v}
\end{aligned}
$$

The Direction of Maximum Variance

- Suppose we want to identify the direction $\mathbf{v}_1$ of maximum variance given the data matrix $\mathbf{X}$. We can formulate this optimization problem as follows:

$$\begin{aligned} \max_{\mathbf{v}} \quad & \frac{1}{M}\mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v} \\ \text{subject to} \quad & \|\mathbf{v}\|_2^2 = 1 \end{aligned}$$

- Letting $\mathbf{\Sigma} = \frac{1}{M}\mathbf{X}^T\mathbf{X}$, we form the Lagrangian

$$L(\mathbf{v}, \lambda) = \mathbf{v}^T\mathbf{\Sigma}\mathbf{v} + \lambda(1 - \|\mathbf{v}\|_2^2)$$

The Direction of Maximum Variance

- Take the derivative of $L(\mathbf{v}, \lambda)$ w.r.t. $\mathbf{v}$ and set it to zero:

$$\frac{\partial L(\mathbf{v}, \lambda)}{\partial \mathbf{v}} = 2\boldsymbol{\Sigma}\mathbf{v} - 2\lambda\mathbf{v} = 0$$
$$\boldsymbol{\Sigma}\mathbf{v} = \lambda\mathbf{v}$$

- As $\mathbf{v} \neq 0$, $\mathbf{v}$ must be an eigenvector of $\boldsymbol{\Sigma}$ with eigenvalue λ. Assuming $\{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ are the eigenvectors of $\boldsymbol{\Sigma}$, corresponding to eigenvalues $\sigma_1 \geq \ldots \geq \sigma_N$, respectively, we have

$$\mathbf{v}^\star = \mathbf{v}_1,$$
$$p^\star = \mathbf{v}_1^T \boldsymbol{\Sigma} \mathbf{v}_1 = \mathbf{v}_1^T \lambda \mathbf{v}_1 = \lambda \mathbf{v}_1^T \mathbf{v}_1 = \lambda = \sigma_1$$

K Largest Directions of Variance

- Suppose instead of just the direction of maximum variance, we want the K largest directions of variance that are all mutually *orthogonal*
- Finding the second-largest direction of variance corresponds to solving the problem:

$$\begin{aligned} \max_{\mathbf{v}} \quad & \mathbf{v}^T \mathbf{\Sigma} \mathbf{v} \\ \text{subject to} \quad & \|\mathbf{v}\|_2^2 = 1 \\ & \mathbf{v}^T \mathbf{v}_1 = 0 \end{aligned}$$

- We form the Lagrangian

$$L(\mathbf{v}, \lambda, \nu) = \mathbf{v}^T \mathbf{\Sigma} \mathbf{v} + \lambda(1 - \|\mathbf{v}\|_2^2) + \nu \mathbf{v}^T \mathbf{v}_1$$

K Largest Directions of Variance

- Taking the derivative of $L(\mathbf{v}, \lambda, \nu)$ w.r.t. $\mathbf{v}$ and setting it to zero, we have

$$\frac{\partial L(\mathbf{v}, \lambda, \nu)}{\partial \mathbf{v}} = 2\mathbf{\Sigma}\mathbf{v} - 2\lambda\mathbf{v} + \nu\mathbf{v}_1 = 0$$

- If we left multiply $\mathbf{v}_1^T$ on both sides

$$\begin{aligned} 2\mathbf{v}_1^T\mathbf{\Sigma}\mathbf{v} - 2\lambda\mathbf{v}_1^T\mathbf{v} + \nu\mathbf{v}_1^T\mathbf{v}_1 &= 0 \\ 2(\mathbf{\Sigma}\mathbf{v}_1)^T\mathbf{v} - 0 + \nu &= 0 \\ 2\sigma_1\mathbf{v}_1^T\mathbf{v} - 0 + \nu &= 0 \\ \nu &= 0 \end{aligned}$$

- Therefore, we arrive at the eigenvalue equation again

$$\mathbf{\Sigma}\mathbf{v} = \lambda\mathbf{v}$$

K Largest Directions of Variance

- It is easy to see that $\mathbf{v}^\star$ is the eigenvector corresponding to the second largest eigenvalue
- In general, the top K directions of variance $\mathbf{v}_1, \ldots, \mathbf{v}_K$ are given by the K eigenvectors corresponding to the K largest eigenvalues of $\frac{1}{M}\mathbf{X}^T\mathbf{X}$
- PCA can also be derived by picking the principal vectors that minimize the approximation error arising from projecting the data onto the K-dimensional subspace spanned by these vectors

$$\min_{\mathbf{v}} \frac{1}{M}\sum_{i=1}^{M}\|\mathbf{x}^{(i)} - (\mathbf{v}^T\mathbf{x}^{(i)})\mathbf{v}\|_2^2$$

Dimensionality Reduction with PCA (Informal)

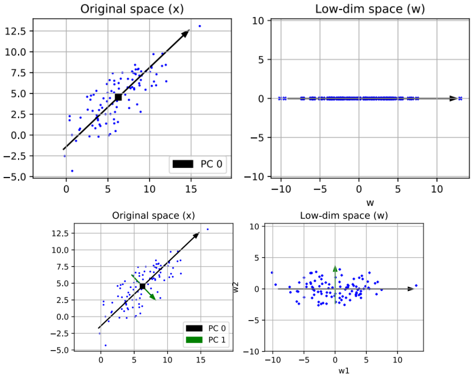
1. Subtract the mean of the data
2. The first PC $\mathbf{v}_1$ is the direction that explains the most variance of the data
3. The second PC $\mathbf{v}_2$ is the direction perpendicular to $\mathbf{v}_1$ that explains the most variance
4. The third PC $\mathbf{v}_3$ is the direction perpendicular to $\{\mathbf{v}_1, \mathbf{v}_2\}$ that explains the most variance
5. ...

Dimensionality Reduction with PCA (Formal)

1. Data preprocessing: Compute $\boldsymbol{\mu} = \frac{1}{M}\sum_i \mathbf{x}^{(i)}$ and replace each $\mathbf{x}^{(i)}$ with $\mathbf{x}^{(i)} - \boldsymbol{\mu}$
2. Given pre-processed data matrix $\mathbf{X} \in \mathbb{R}^{M\times N}$, compute the sample covariance matrix $\boldsymbol{\Sigma} = \frac{1}{M}\mathbf{X}^T\mathbf{X}$
3. Compute the K leading eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_K$ of $\boldsymbol{\Sigma}$ where $\mathbf{v}_i \in \mathbb{R}^N$
4. Stack the eigenvectors together into an $N \times K$ matrix $\mathbf{V}$ where each column i of $\mathbf{V}$ corresponds to $\mathbf{v}_i$
5. Project the matrix $\mathbf{X}$ into the rank-K subspace of maximum variance by computing the matrix product $\mathbf{Z} = \mathbf{X}\mathbf{V}$
6. To reconstruct $\mathbf{X}$ given $\mathbf{Z}$ and $\mathbf{V}$, we use $\hat{\mathbf{X}} = \mathbf{Z}\mathbf{V}^T$

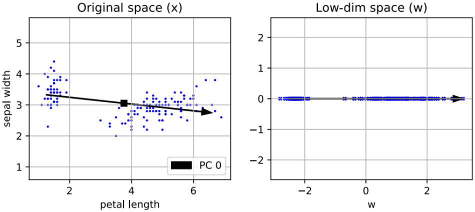
Example on Blob Data

- First and Second PC

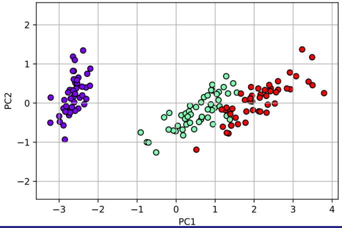


Example on Iris Data

- 2D (petal length, sepal width) to 1D



- 4D (sepal length, sepal width, petal length, petal width) to 2D

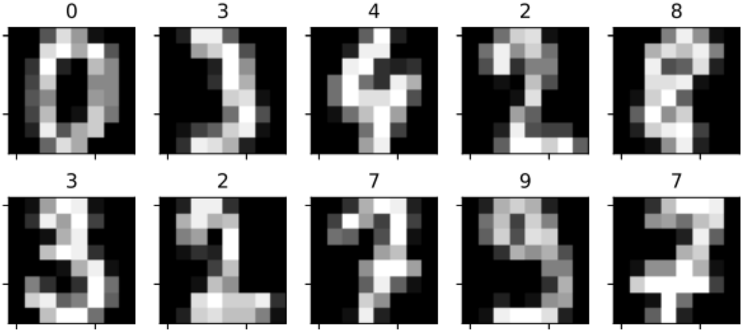


How to Choose the Number of PCs?

- Two methods to set the number of components K:
  - Preserve some percentage of the variance (e.g., 95%)
  - Whatever works well for our final task (e.g., classification, regression)

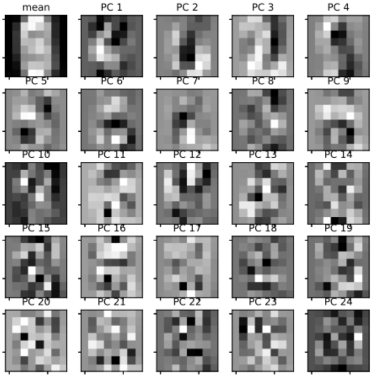
Handwritten Digits Data

- $1,797$ images of handwritten digits 0-9
  - Each image is 8×8
  - Flattened into a 64 dimensional vector



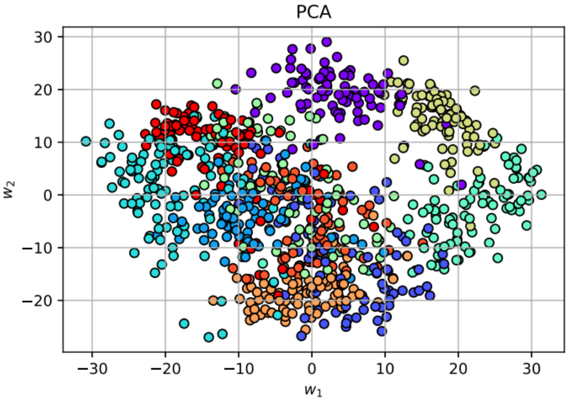
Run PCA on the Data

- Split data into training and testing sets
- Run PCA on training set, apply to test set
- The top 25 PCs are shown



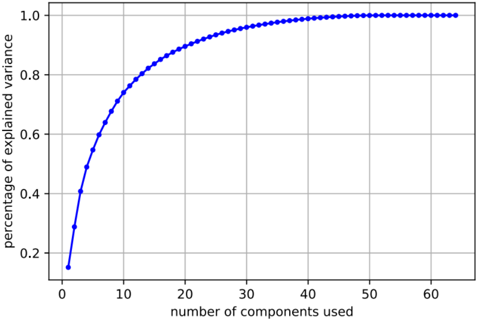
Run PCA on the Data

- Visualize the coefficients for the first two PCs
  - Grouping of different digits is sometimes preserved



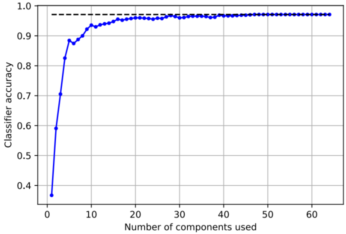
Explained Variance

- Each PC explains a percentage of the original data
  - This is called the explained variance
  - PCs are already sorted by explained variance from highest to lowest
- Pick the number of PCs to get a certain percentage of explained variance, typically 95%



Task-Dependent Selection

- Use results on the final task (in this case classification) to select the best number of components
- Note: no need to rerun PCA for each number of components
  - Just select the desired subset of PCs



- Classification accuracy is stable after using 20 PCs
  - Not much loss in performance if using only 20 PCs

Outline

Connection to SVD

- We have saw that the minimum Frobenius norm linear dimensionality reduction problem could be solved using the rank-K SVD of $\mathbf{X}$:

$$\underset{\mathbf{U},\mathbf{S},\mathbf{V}}{\arg\min} \, ||\mathbf{X} - \mathbf{U}\mathbf{S}\mathbf{V}^T||_F^2$$

where $\mathbf{U} \in \mathbb{R}^{M\times K}$, $\mathbf{S} \in \mathbb{R}^{K\times K}$, and $\mathbf{V} \in \mathbb{R}^{N\times K}$. The matrix product $\mathbf{Z} = \mathbf{U}\mathbf{S}$ gives the optimal rank-K representation of $\mathbf{X}$ with respect to Frobenius norm minimization

Connection to SVD

- If we let $K = N$ then $\mathbf{X} = \mathbf{USV}^T$ and $\mathbf{X}^T\mathbf{X} = \mathbf{VSU}^T\mathbf{USV}^T$
- Due to orthogonality of $\mathbf{U}$ this gives: $\mathbf{X}^T\mathbf{X} = \mathbf{VS}^2\mathbf{V}^T$
- This means that the right singular vectors of $\mathbf{X}$ are exactly the eigenvectors of $\mathbf{X}^T\mathbf{X}$
- We can also see that the eigenvalues of $\mathbf{X}^T\mathbf{X}$ are the squares of the diagonal elements of $\mathbf{S}$
- This means that the K largest singular values and K largest eigenvalues correspond to the same K basis vectors

Connection to SVD

- According to PCA, the projection operation is $\mathbf{Z} = \mathbf{X}\mathbf{V}$, therefore

$$\mathbf{Z} = \mathbf{X}\mathbf{V} = (\mathbf{U}\mathbf{S}\mathbf{V}^T)(\mathbf{V}) = \mathbf{U}\mathbf{S}$$

- Finally, note that if the decompositions are based only on the K leading principal vectors, the projections $\mathbf{Z} = \mathbf{X}\mathbf{V}$ and $\mathbf{Z} = \mathbf{U}\mathbf{S}$ will still be identical

Connection to SVD

- These manipulations show that PCA on $\mathbf{X}^T\mathbf{X}$ and SVD on $\mathbf{X}$ identify exactly the same subspace and result in exactly the same projection of the data into that subspace
- As a result, generic linear dimensionality reduction simultaneously minimizes the Frobenius norm of the reconstruction error of $\mathbf{X}$ and maximizes the retained variance in the learned subspace
- Both SVD and PCA provide the same description of generic linear dimensionality reduction: an orthogonal basis for exactly the same optimal linear subspace

When Does PCA Fail?

- The primary motivation behind PCA is to decorrelate the dataset, *i.e.*, remove second-order dependencies. If higher-order dependencies exist between the features in the data, PCA may be insufficient at revealing all structure in the data
- PCA requires that each component must be perpendicular to the previous ones, but clearly this requirement is overly stringent and the data might be arranged along non-orthogonal axes