Principal Component Analysis

CSC311

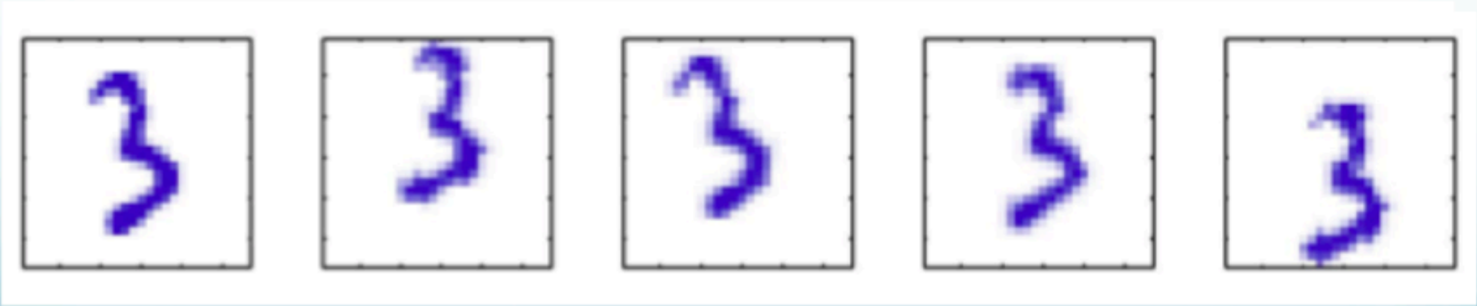
Based on Slides from Amir-massoud Farahmand & Emad A.M. Andrews

Dimensionality Reduction

- We have some data $X \in \mathbb{R}^{N \times D}$
- D may be huge, etc.
- We would like to find a new representation $Z \in \mathbb{R}^{N \times K}$ where K << D.
  - For computational reasons.
  - To better understand (e.g., visualize) the data.
  - For compression.
  - ...
- We will restrict ourselves to linear transformations for the time being.

Example

- In this dataset, there are only 3 degrees of freedom: horizontal and vertical translations, and rotations.
- Yet each image contains 784 pixels, so X will be 784 elements wide.



Setup: Multivariate Inputs

- Setup: Given an i.i.d. dataset $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\} \subset \mathbb{R}^D$.
- N instances/observations/examples

$$\mathbf{X} = \begin{bmatrix} [\mathbf{x}^{(1)}]^\top \\ [\mathbf{x}^{(2)}]^\top \\ \vdots \\ [\mathbf{x}^{(N)}]^\top \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_D^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_D^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(N)} & x_2^{(N)} & \cdots & x_D^{(N)} \end{bmatrix}$$

- Mean

$$\mathbb{E}[\mathbf{x}^{(i)}] = \boldsymbol{\mu} = [\mu_1, \cdots, \mu_D]^T \in \mathbb{R}^D$$

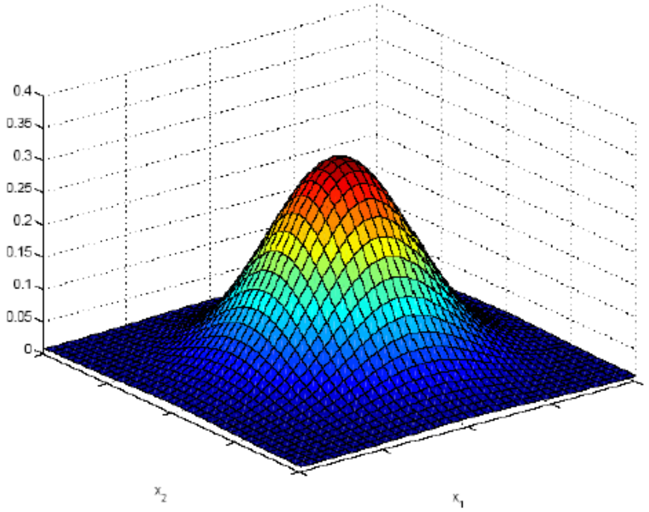
- Covariance

$$\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{x}^{(i)}) = \mathbb{E}[(\mathbf{x}^{(i)} - \boldsymbol{\mu})(\mathbf{x}^{(i)} - \boldsymbol{\mu})^\top] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1D} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{D1} & \sigma_{D2} & \cdots & \sigma_D^2 \end{bmatrix}$$

Multivariate Gaussian Model

- $\mathbf{x}^{(i)} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, a Gaussian (or normal) distribution defined as

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$



Mean and Covariance Estimators

- Observed data: $\mathcal{D} = \{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}\}$.
- Recall that the MLE estimators for the mean $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ under the multivariate Gaussian model is given by (previous lecture)

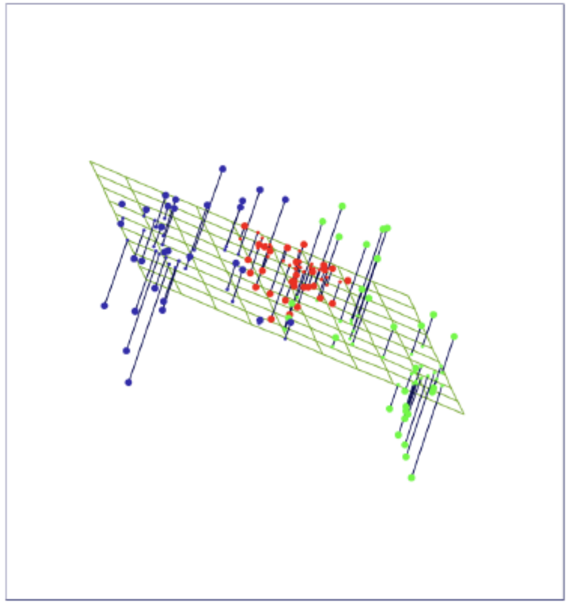
$$\text{Sample mean:} \quad \hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^{(i)}$$

$$\text{Sample covariance:} \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}})(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}})^\top$$

- $\hat{\boldsymbol{\mu}}$ quantifies (approximately) where your data is located in space.
- $\hat{\boldsymbol{\Sigma}}$ quantifies (approximately) how your data points are spread.

Low Dimensional Representation

- Sometimes in practice, even though data is very high dimensional, its important features can be accurately captured in a low dimensional subspace.



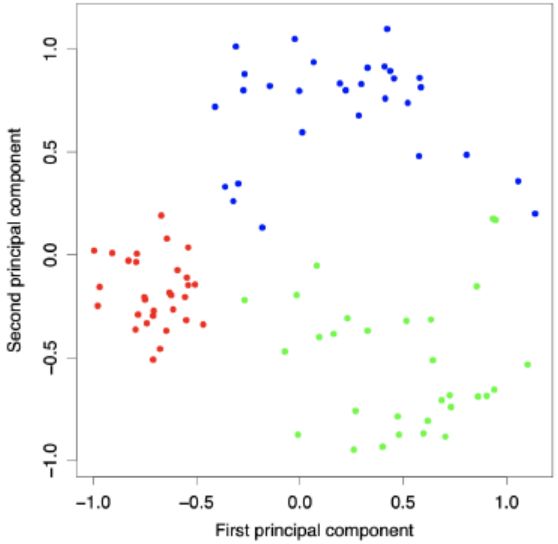
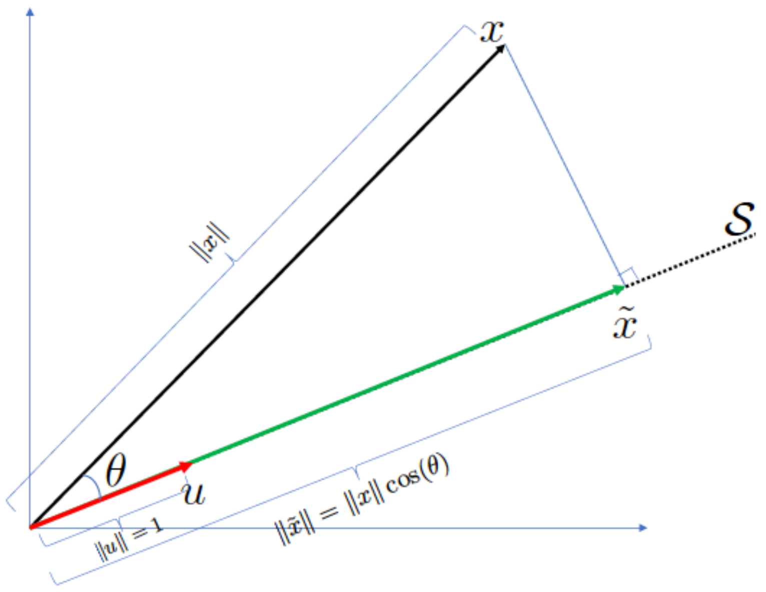


Image credit: Elements of Statistical Learning

Euclidean Projection



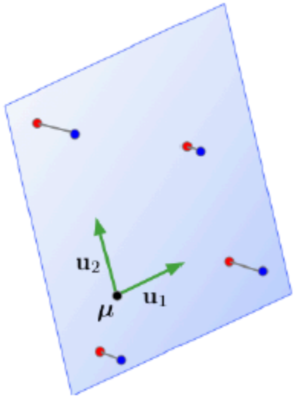
- Here, $\mathcal{S}$ is the line along the unit vector $\mathbf{u}$ (1-dimensional subspace)
  - $\mathbf{u}$ is a basis for $\mathcal{S}$: any point in $\mathcal{S}$ can be written as $z\mathbf{u}$ for some z.

- Projection of $\mathbf{x}$ on $\mathcal{S}$ is denoted by $\text{Proj}_{\mathcal{S}}(\mathbf{x})$
- Recall: $\mathbf{x}^\top \mathbf{u} = \|\mathbf{x}\|\|\mathbf{u}\| \cos(\theta) = \|\mathbf{x}\| \cos(\theta)$
- $\text{Proj}_{\mathcal{S}}(\mathbf{x}) = \underbrace{\mathbf{x}^\top \mathbf{u}}_{\text{length of proj}} \cdot \underbrace{\mathbf{u}}_{\text{direction of proj}} = \|\tilde{\mathbf{x}}\|\mathbf{u}$

General Subspaces

- In general, $\mathcal{S}$ is not one dimensional (i.e., line), but a (linear) subspace with a dimension K.

- In this case, we have K basis vectors $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_K \in \mathbb{R}^D$: any vector $\mathbf{y}$ in $\mathcal{S}$ can be written as $\mathbf{y} = \sum_{i=1}^{K} z_i \mathbf{u}_i$ for some $z_1, \ldots, z_K$.



- Projection of $\mathbf{x} \in \mathbb{R}^D$ on this subspace is given by

$$\text{Proj}_{\mathcal{S}}(\mathbf{x}) = \sum_{i=1}^{K} z_i \mathbf{u}_i \quad \text{where} \quad z_i = \mathbf{x}^\top \mathbf{u}_i.$$

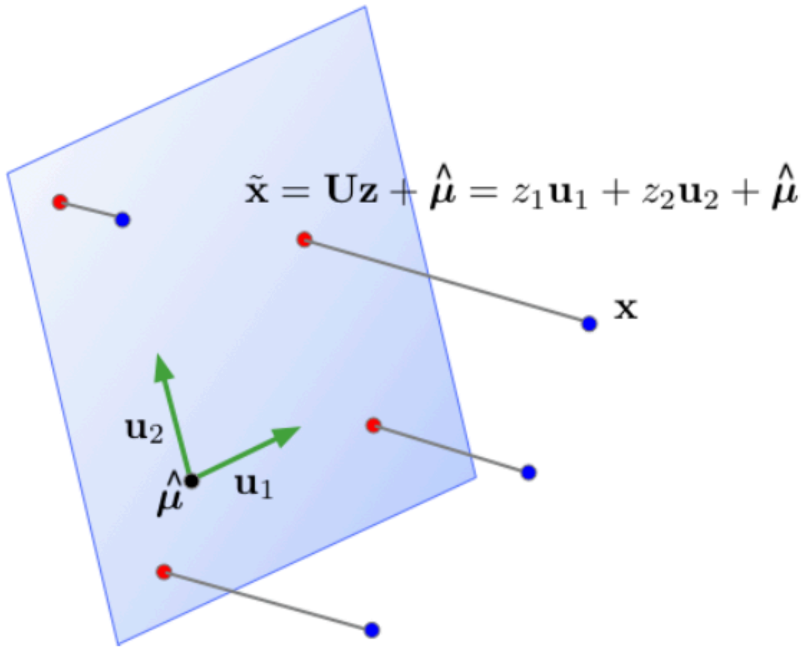
Projection onto a Subspace

- Let $\{\mathbf{u}_k\}_{k=1}^K$ be an orthonormal basis of the subspace $\mathcal{S}$ (a K-dimensional linear subspace of $\mathbb{R}^D$).
- Approximate each data point $\mathbf{x} \in \mathbb{R}^D$ as:
  1. Center (subtract the mean)
  2. Project onto $\mathcal{S}$
  3. Add the mean back

$$\tilde{\mathbf{x}} = \hat{\boldsymbol{\mu}} + \text{Proj}_{\mathcal{S}}(\mathbf{x} - \hat{\boldsymbol{\mu}})$$
$$= \hat{\boldsymbol{\mu}} + \sum_{k=1}^{K} z_k \mathbf{u}_k$$

- We also know: $z_k = \mathbf{u}_k^T(\mathbf{x} - \hat{\boldsymbol{\mu}})$
- Let $\mathbf{U} \in \mathbb{R}^{D \times K}$ be a matrix with columns $\{\mathbf{u}_k\}_{k=1}^K$.
- Then $\mathbf{z} = \mathbf{U}^T(\mathbf{x} - \hat{\boldsymbol{\mu}})$ (Note that $\mathbf{z} \in \mathbb{R}^K$).
- Also: $\tilde{\mathbf{x}} = \hat{\boldsymbol{\mu}} + \mathbf{U}\mathbf{z} = \hat{\boldsymbol{\mu}} + \mathbf{U}\mathbf{U}^T(\mathbf{x} - \hat{\boldsymbol{\mu}})$ (Note that $\tilde{\mathbf{x}} \in \mathbb{R}^D$).
- Here, $\mathbf{U}\mathbf{U}^T$ is the projector onto $\mathcal{S}$, and $\mathbf{U}^T\mathbf{U} = I$.

Projection onto a Subspace



$$\mathbf{z} = \mathbf{U}^\top(\mathbf{x} - \hat{\boldsymbol{\mu}})$$

- In machine learning, $\tilde{\mathbf{x}}$ is also called the reconstruction of $\mathbf{x}$.
- $\mathbf{z}$ is its representation or code.

Learning a Subspace

- How to choose a good subspace $\mathcal{S}$?
  - Need to choose $D \times K$ matrix $\mathbf{U}$ with orthonormal columns.
- Two criteria:
  - Minimize the reconstruction error: Find vectors in a subspace that are closest to data points.

$$\min_{\mathbf{U}} \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{x}^{(i)} - \tilde{\mathbf{x}}^{(i)} \right\|^2$$

  - Maximize the variance of reconstructions: Find a subspace where data has the most variability.

$$\max_{\mathbf{U}} \frac{1}{N} \sum_{i} \left\| \tilde{\mathbf{x}}^{(i)} - \hat{\boldsymbol{\mu}} \right\|^2$$

  - The data and its reconstruction has the same means (exercise)!

PCA in General

- We can compute the entire PCA solution by just computing the eigenvectors with the top-k eigenvalues.
- These can be found using the singular value decomposition of Σ

Example: PCA

- Let our data matrix $\boldsymbol{X}$ be the score of three subjects :

Student	Math	English	Art
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

Example: PCA

- We can write then $\boldsymbol{X}$ as:

$$\mathrm{X} = \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix}$$

- Let’s then Compute the mean of every dimension:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^{(i)}$$

$$\mu = [66 \quad 60 \quad 60]$$

Example: PCA

- Compute the *covariance matrix* of the whole dataset:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}})(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}})^{\top}$$

$$\Sigma = \begin{bmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{bmatrix}$$

Example: PCA

- Compute Eigenvectors and corresponding Eigenvalues:
  - The eigenvalues of Σ are the roots of the characteristic equation:

$$\det(\Sigma - \lambda I) = 0$$

$$\det\begin{bmatrix} 504-\lambda & 360 & 180 \\ 360 & 360-\lambda & 0 \\ 180 & 0 & 720-\lambda \end{bmatrix} = 0$$

$$-\lambda^3 + 1584\lambda^2 - 641520\lambda + 25660800 = 0$$

Example: PCA

- After solving the previous equation for λ, we get:

$$\lambda_1 = 44.82, \lambda_2 = 629.11, \lambda_3 = 910.07$$

- And the corresponding orthonormal basis corresponding to the above values:

$$u_1 = \begin{bmatrix} -0.649 \\ 0.742 \\ 0.173 \end{bmatrix}, u_2 = \begin{bmatrix} -0.386 \\ -0.516 \\ 0.765 \end{bmatrix}, u_3 = \begin{bmatrix} 0.656 \\ 0.429 \\ 0.621 \end{bmatrix}$$

Example: PCA

- Let's reduce the dimension of $X = \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix}$ from 3 to 2.
- We have to chooses two basis that corresponds to the highest eigenvalues.

$$U = \begin{bmatrix} 0.656 & -0.386 \\ 0.429 & -0.516 \\ 0.621 & 0.765 \end{bmatrix}$$

Example: PCA

- We know z is the representation or the projection onto the new subspace.

$$\mathbf{z} = \mathbf{U}^\top(\mathbf{x} - \hat{\boldsymbol{\mu}})$$

$$Z = \begin{bmatrix} 34.374 & 13.686 \\ 9.984 & -47.694 \\ -3.936 & 2.316 \\ 14.694 & 25.266 \\ -55.116 & 6.426 \end{bmatrix}$$

Example: PCA

- Let's reconstruct $\tilde{X}$ from Z and U:

$$\tilde{\mathbf{x}} = \hat{\boldsymbol{\mu}} + \mathbf{U}\mathbf{z}$$

$$\tilde{X} = \begin{bmatrix} 83.266 & 67.684 & 91.816 \\ 90.9594 & 88.893 & 29.714 \\ 62.524 & 57.116 & 59.327 \\ 65.886 & 53.266 & 88.454 \\ 27.364 & 33.039 & 30.6889 \end{bmatrix}, X = \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix}$$

Applying PCA to digits