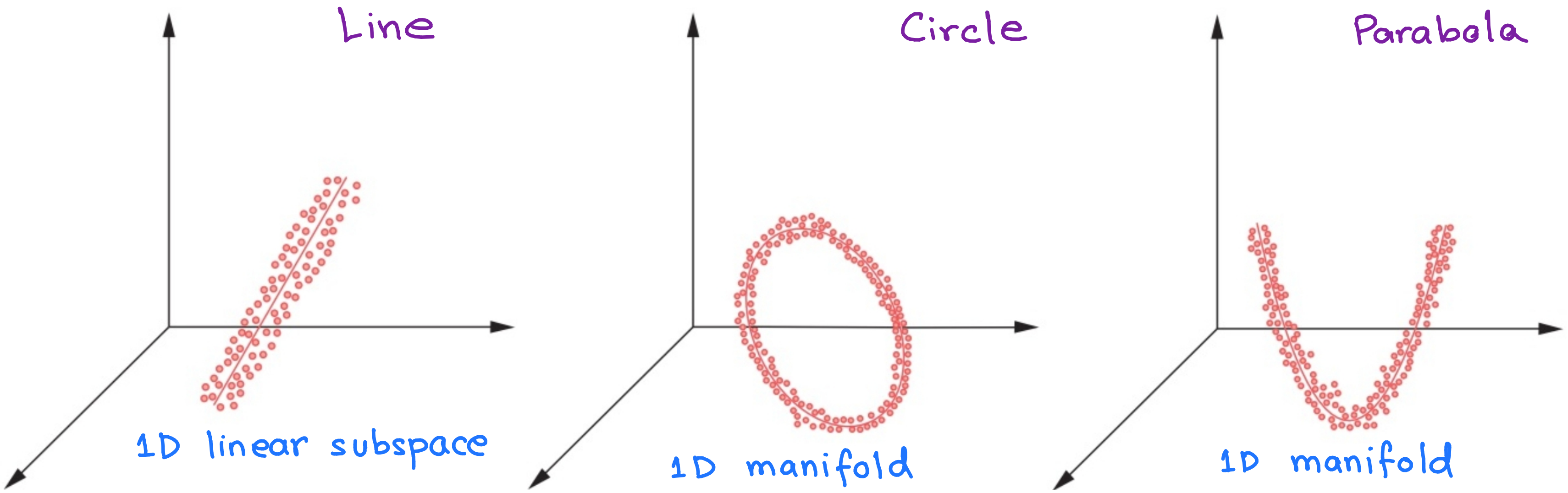
Dimensionality Reduction

- In unsupervised learning, we have seen clustering.

- In this lecture, we will look at dimensionality reduction

- In many practical applications, the input data $\underline{x} \in \mathbb{R}^p$ is a very high-dimensional, however, the intrinsic dimensionality may be quite small



In all three cases, the intrinsic dimensionality of data is 1

Intrinsic Dimensionality

- A data set $\{\underline{x}^{(i)}\}_{i=1}^{N}$, with $\underline{x} \in \mathbb{R}^P$, is said to have intrinsic dimensionality $M \leq P$, if the dataset can be described effectively in terms of 'M' free random variables

$$\underset{\mathbb{R}^P}{\underline{x}} = g(\underset{\mathbb{R}^M}{\underline{u}})$$

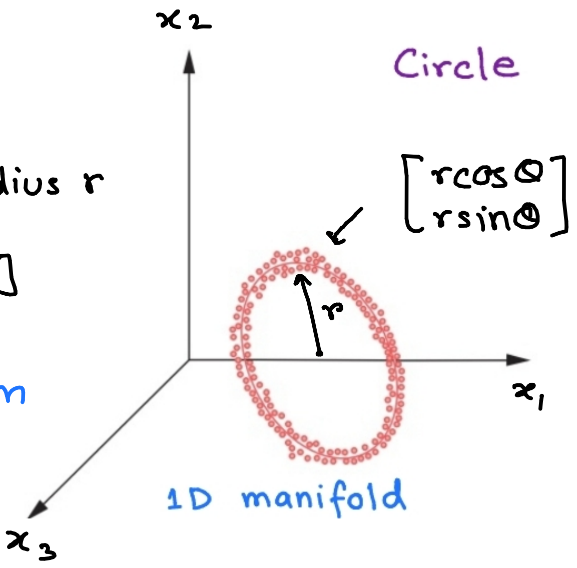
Example

Constant radius r

$\theta \sim [0, 2\pi]$

$\uparrow$

free random variable



The data lies along the circumference of a circle of radius r and a single free parameter θ suffices to describe the data

Intrinsic dimension = 1

Intrinsic Dimensionality

- An important concern in ML is learning from high-dimensional data $\underline{x}$
- Success of ML, in particular deep learning, is due to its capability of

learning a useful representation of high-dimensional data

- One of the goals of unsupervised learning:

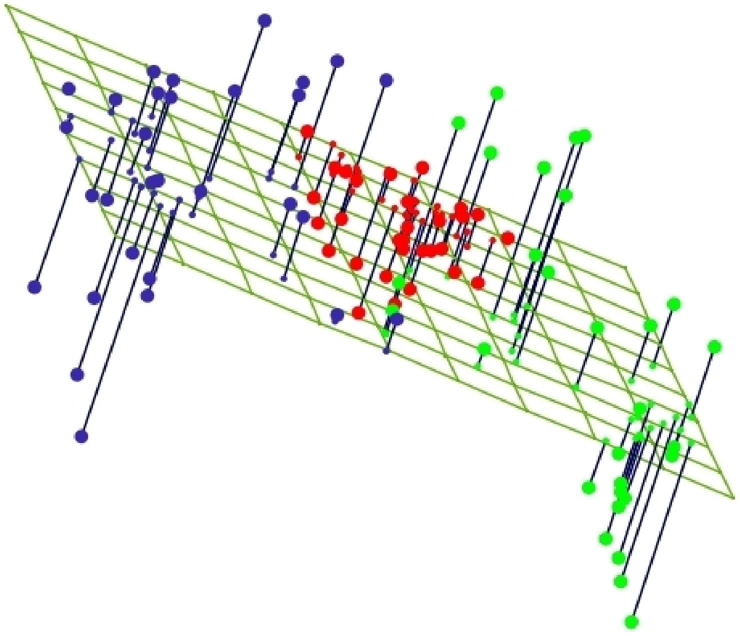
Learning a lower-dimensional subspace for encoding high-dimensional data set

- Idea of dimensionality reduction: Map data to a lower dimensional space
  - Save computational time in modelling high-dimensional data
  - Visualization in 2-dimensions can offer insights
  - Reduce overfitting and achieve better generalization

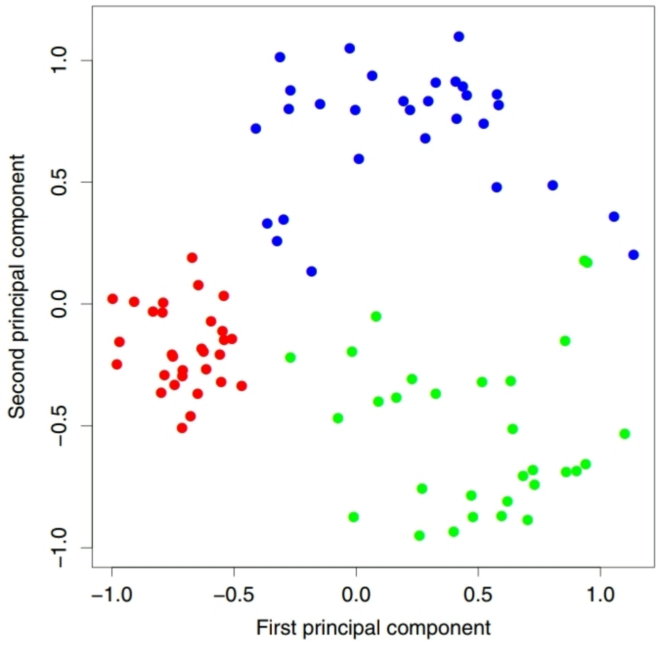
Linear Dimensionality Reduction

- We will introduce linear dimensionality reduction using Principal Component Analysis (PCA)

- PCA is also known as Karhunen-Loève (KL) transform
  - It falls under linear dimensionality reduction techniques



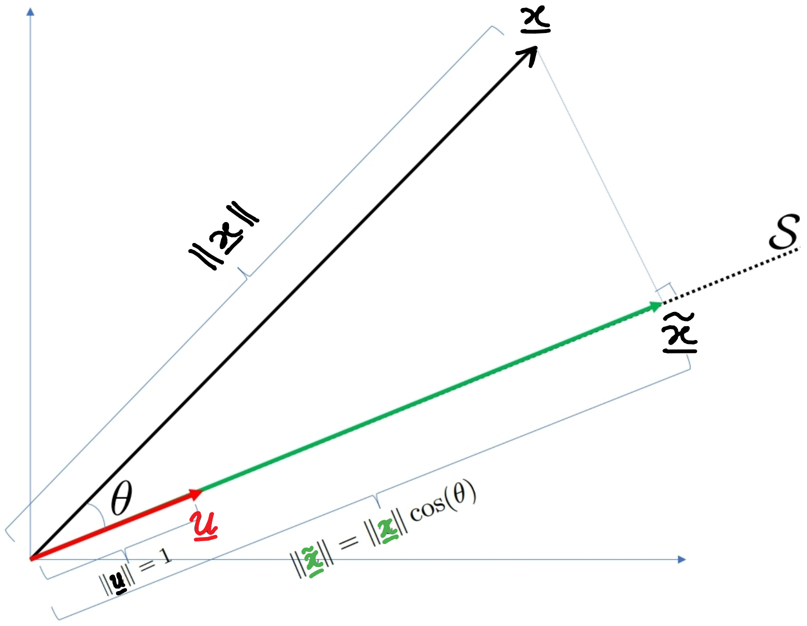
Projection on a linear subspace



3D space $\xrightarrow{\text{PCA}}$ 2D - space

Idea of projection

- Consider projection onto 1-D subspace (a line)



- Subspace S is the line along the unit vector $\underline{u}$
  - $\underline{u}$ is the basis of S: Any point in S can be written as $z\underline{u}$ for some scalar z

- Projection of vector $\underline{x}$ on S is denoted by $\underline{\tilde{x}} = \mathrm{Proj}_S(\underline{x})$

- Recall that: $\underline{x}^T\underline{u} = \|\underline{x}\| \, \overset{1}{\|\underline{u}\|} \cos(\theta) = \|\underline{x}\| \cos\theta = \|\underline{\tilde{x}}\|$

- $$\underline{\tilde{x}} = \mathrm{Proj}_S(\underline{x}) = \underbrace{\underline{x}^T\underline{u}}_{\text{length of projection}} \cdot \underbrace{\underline{u}}_{\text{direction of projection}} = \|\underline{\tilde{x}}\| \, \underline{u}$$

Idea of projection

- How to project onto a M-dimensional subspace?
  - Idea: Choose an orthonormal bases $\{\underline{u}_1, \underline{u}_2, \dots, \underline{u}_M\}$ for S
  - Project onto each unit vector individually (as in previous slide) and sum together the projections

- Mathematically, the projection is given as:

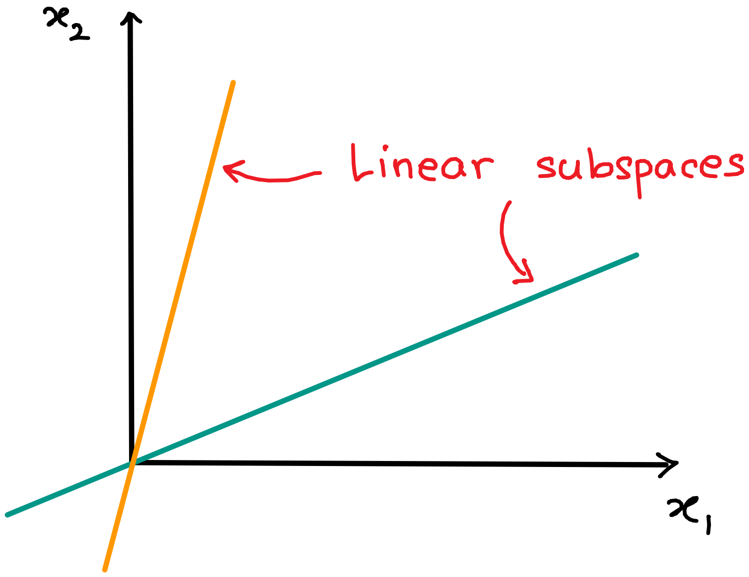
$$\tilde{\underline{x}} = \mathrm{Proj}_S(\underline{x}) = \sum_{i=1}^{M} z_i \, \underline{u}_i \qquad \text{where} \qquad z_i = \underline{x}^T \underline{u}_i$$

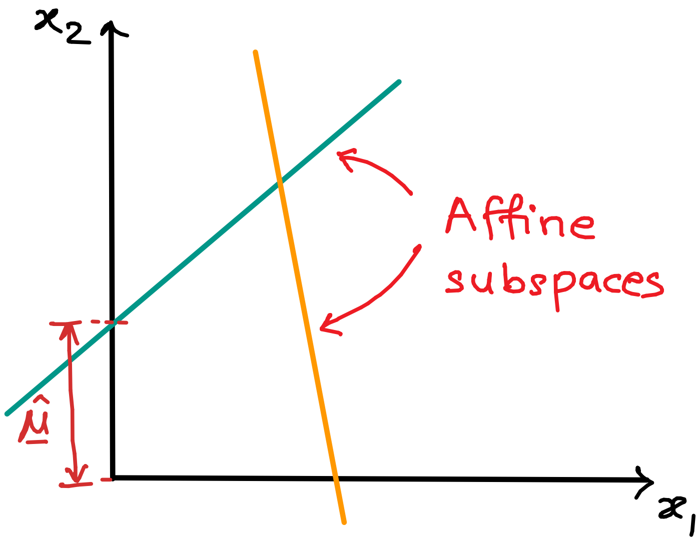
- In vector form:

$$\tilde{\underline{x}} = \mathrm{Proj}_S(\underline{x}) = \underline{\underline{U}} \, \underline{z} = \begin{bmatrix} | & | & & | \\ \underline{u}_1 & \underline{u}_2 & \cdots & u_M \\ | & | & & | \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_M \end{bmatrix}, \qquad \text{where} \qquad \underline{z} = \underline{\underline{U}}^T \underline{x}$$

Projection onto an affine subspace

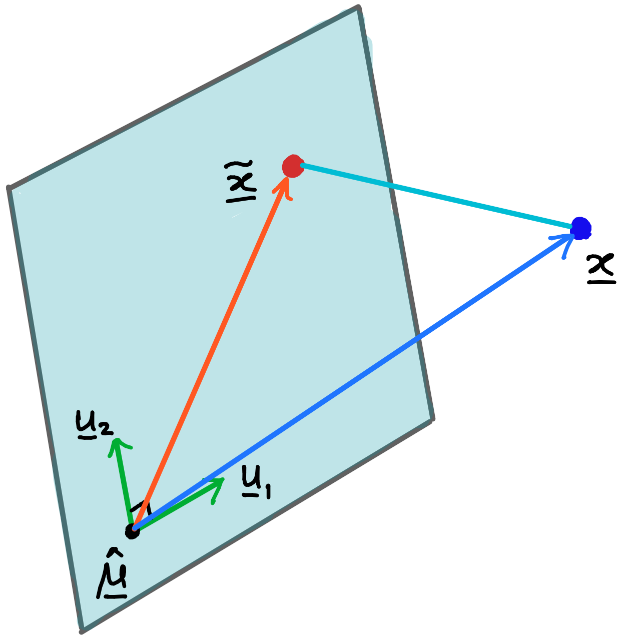
- So far, we have assumed a subspace that passes through zero
- However, the subspaces that we want to project onto can also be affine subspaces, which need not pass through zero





The affine subspaces can have an arbitrary origin $\hat{\underline{\mu}}$

Projection onto an affine subspace



data point projected on the affine subspace

$$\underline{\tilde{x}} = \text{Proj}_S(\underline{x}) = \underline{\underline{U}}\,\underline{z} + \underline{\hat{\mu}} = z_1\underline{u}_1 + z_2\underline{u}_2 + \underline{\hat{\mu}}$$

$$\underline{z} = \underline{\underline{U}}^T(\underline{x} - \underline{\hat{\mu}})$$

The affine subspace has an origin $\underline{\hat{\mu}}$

- $\underline{\tilde{x}}$ is called the reconstruction of $\underline{x}$
- $\underline{z}$ is its feature / code
- If all the data points $\underline{x}$ lie close to the subspace, we could approximate $\underline{x}$ with its reconstructions $\underline{\tilde{x}}$

$$\underline{x} \approx \underline{\underline{U}}\,\underline{z} + \underline{\hat{\mu}}$$

How to choose a good subspace?

- We want to choose a subspace S which is low-dimensional compared to the dimension of the input space

- How to choose such a subspace S?
  - We need to find appropriate $\hat{\underline{\mu}}$ and the orthogonal bases $\underline{\underline{U}}$
  - origin $\hat{\underline{\mu}}$ can be set equal to the mean of the dataset

- To find $\underline{\underline{U}}$, one of the two equivalent criteria could be followed:

  - Minimize the reconstruction error:

$$\boxed{\underset{\underline{\underline{U}}}{\arg\min} \; \frac{1}{N} \sum_{i=1}^{N} \left\| \underline{x}^{(i)} - \underline{\tilde{x}}^{(i)} \right\|_2^2}$$

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

$$\boxed{\underset{\underline{\underline{U}}}{\arg\max} \; \frac{1}{N} \sum_{i=1}^{N} \left\| \underline{\tilde{x}}^{(i)} - \hat{\underline{\mu}} \right\|_2^2}$$

(You can show that $\underline{x}$ and $\underline{\tilde{x}}$ have same mean)

Principal Component Analysis

- Choosing a subspace to maximize the projected variance, or minimize the reconstruction error, is called PCA

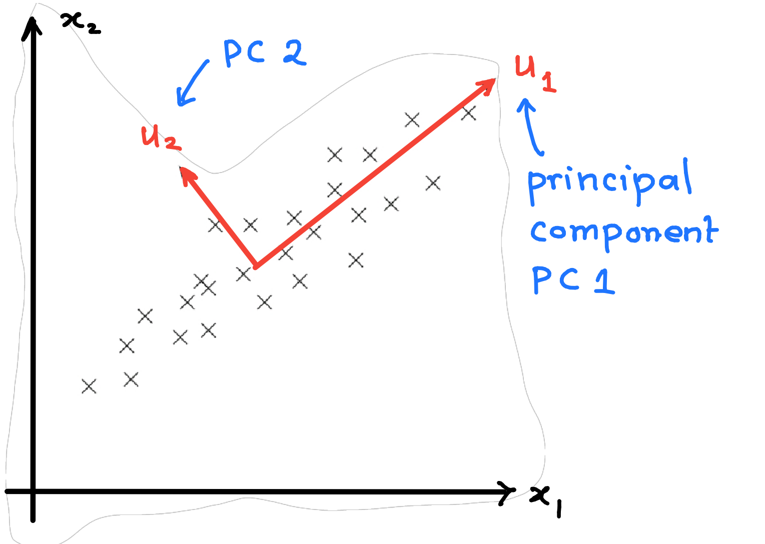
- Consider the sample covariance matrix :

$$\underline{\underline{\hat{\Sigma}}} = \frac{1}{N} \sum_{i=1}^{N} (\underline{x}^{(i)} - \underline{\hat{\mu}})(\underline{x}^{(i)} - \underline{\hat{\mu}})^T$$

- $\underline{\underline{\hat{\Sigma}}}$ is symmetric and Positive semi-definite (PSD)

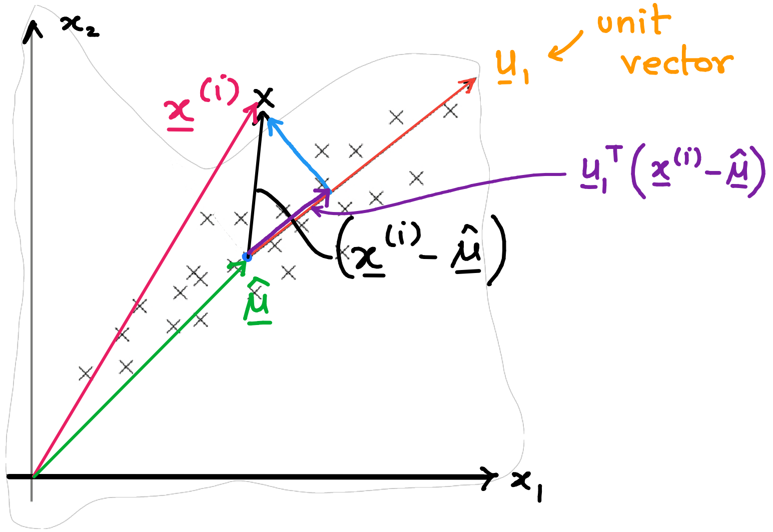
- The optimal PCA subspace is spanned by the top 'M' eigenvectors of $\underline{\underline{\hat{\Sigma}}}$

- These eigenvectors are called principal components or principal directions, much like the principal axes of an ellipse



Derivation of PCA

- Let us consider the simplest case of finding a 1-D subspace
  - The goal then is to find a single direction represented by unit vector $\underline{u}_1$



x_2
x_1
$\underline{x}^{(i)}$
$\underline{u}_1$
unit vector
$\underline{u}_1^T(\underline{x}^{(i)} - \underline{\hat{\mu}})$
$(\underline{x}^{(i)} - \underline{\hat{\mu}})$
$\underline{\hat{\mu}}$

- Lets maximize the projected variance

$$J(\underline{u}_1) = \frac{1}{N}\sum_{i=1}^{N}\left(\underline{u}_1^T\,(\underline{x}^{(i)} - \underline{\hat{\mu}})\right)^2$$

$$= \frac{1}{N}\sum_{i=1}^{N} \underline{u}_1^T (\underline{x}^{(i)} - \underline{\hat{\mu}})(\underline{x}^{(i)} - \underline{\hat{\mu}})^T \underline{u}_1$$

$$= \underline{u}_1^T\, \underline{\underline{\hat{\Sigma}}}\, \underline{u}_1$$

- So the optimization task is :

$$\boxed{\begin{aligned} \underline{u}_1 &= \underset{\underline{u}}{\operatorname{argmax}}\ \underline{u}^T \underline{\underline{\hat{\Sigma}}}\, \underline{u} \\ \text{s.t.}\quad & \underline{u}^T\underline{u} = 1 \end{aligned}}$$

Lagrangian: $L(\underline{u}, \lambda) = \underline{u}^T \underline{\underline{\hat{\Sigma}}}\, \underline{u} - \lambda(\underline{u}^T\underline{u} - 1)$

Take gradient and set to zero:

eigenvalue

$$\underline{\underline{\hat{\Sigma}}}\, \underline{u} = \lambda\, \underline{u}$$

eigenvector

∴ Principal direction $\underline{u}_1$ is an eigenvector

- Since $\hat{\underline{\underline{\Sigma}}}$ is symmetric and PSD, all eigenvalues are real and non-negative: $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geqslant 0$

- The 2nd principal component $\underline{u}_2$ is selected such that:

  (a) $\underline{u}_2$ is orthogonal to $\underline{u}_1$

  (b) $\underline{u}_2$ maximizes the variance after projecting the data onto the direction of $\underline{u}_2$

  (c) The 2nd principal component (or direction) is the eigenvector corresponding to the 2nd largest eigenvalue of $\hat{\underline{\underline{\Sigma}}}$, λ_2

- Similar arguments can be used to show that the 'm' th principal component is the 'm' th eigenvector of $\hat{\underline{\underline{\Sigma}}}$

- The process continues until M principal components (corresponding to the M largest eigenvalues)

PCA decorrelates features

- The features (or code) are decorrelated by PCA

$$\begin{aligned} Cov(\underline{z}) &= Cov\left(\underline{\underline{U}}^T(\underline{x}-\hat{\underline{\mu}})\right) \\ &= \underline{\underline{U}}^T \, Cov(\underline{x}) \, \underline{\underline{U}} \\ &= \underline{\underline{U}}^T \, \hat{\underline{\underline{\Sigma}}} \, \underline{\underline{U}} \\ &= \underline{\underline{U}}^T \, \underline{\underline{Q}} \, \underline{\underline{\Lambda}} \, \underline{\underline{Q}}^T \, \underline{\underline{U}} \\ &= \begin{bmatrix} \underline{\underline{I}} & \underline{\underline{0}} \end{bmatrix} \underline{\underline{\Lambda}} \begin{bmatrix} \underline{\underline{I}} \\ \underline{\underline{0}} \end{bmatrix} \\ &= \text{top left } M\times M \text{ block of } \underline{\underline{\Lambda}} \end{aligned}$$

Spectral decomposition

$$\underset{p\times p}{\hat{\underline{\underline{\Sigma}}}} = \underline{\underline{Q}} \, \underline{\underline{\Lambda}} \, \underline{\underline{Q}}^T$$

eigenvector matrix (orthonormal)

eigenvalues matrix

$$\underset{p\times p}{\underline{\underline{Q}}} = \Big[\underset{p\times M}{\underline{\underline{U}}} \;\vdots\; \underset{p\times(p-M)}{\underline{\underline{U}}_{\perp}} \Big]$$

- Covariance of feature $\underline{z}$ is diagonal → uncorrelated

Summary of PCA

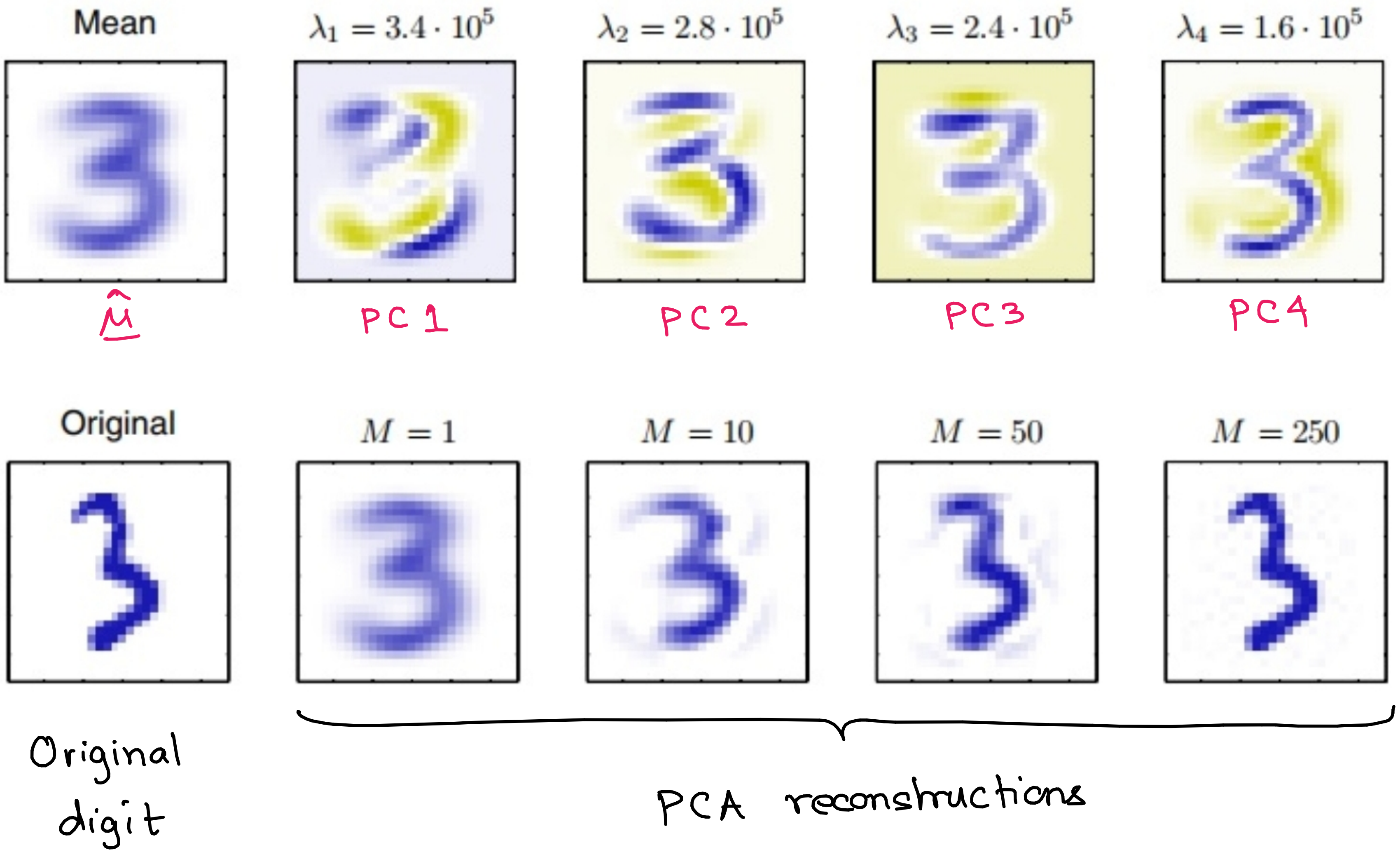
- Dimensionality reduction aims to find a low-dimensional representation of the data

- PCA projects the data onto an affine subspace that maximizes projected variance or minimizes the reconstruction error

- The optimal subspace is given by the top M eigenvectors of the sample covariance matrix, corresponding to the M largest eigenvalues

- PCA gives a set of decorrelated features

Example of data compression


Mean
$\lambda_1 = 3.4 \cdot 10^5$
$\lambda_2 = 2.8 \cdot 10^5$
$\lambda_3 = 2.4 \cdot 10^5$
$\lambda_4 = 1.6 \cdot 10^5$
$\hat{\mu}$
PC 1
PC 2
PC 3
PC 4
Original
$M = 1$
$M = 10$
$M = 50$
$M = 250$
Original digit
PCA reconstructions