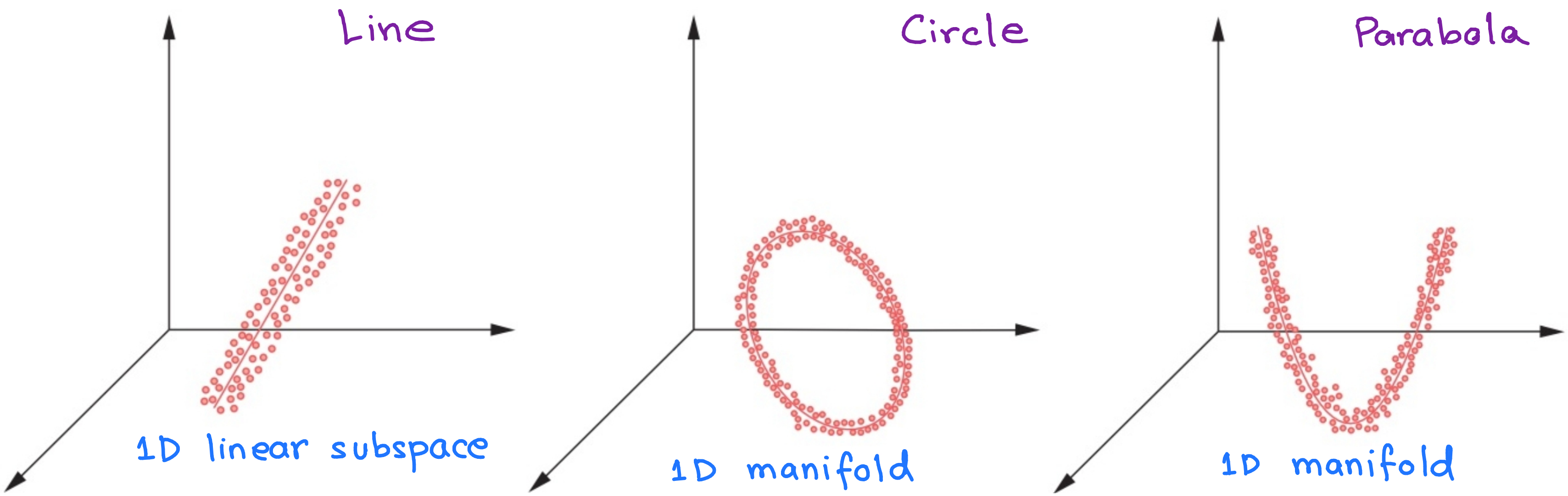
Dimensionality Reduction

- 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

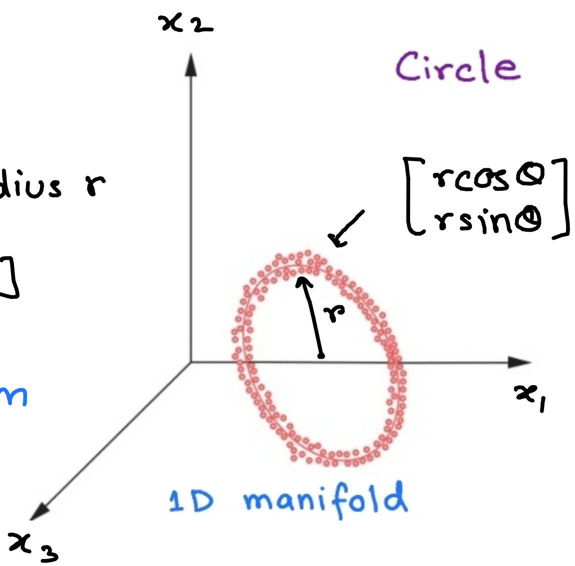
$$\underbrace{\underline{x}}_{\mathbb{R}^P} = g(\underbrace{\underline{u}}_{\mathbb{R}^M})$$

Example

Constant radius r

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

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

$$\boxed{\text{learning a useful representation of high-dimensional data}}$$

- One of the goals of unsupervised learning:

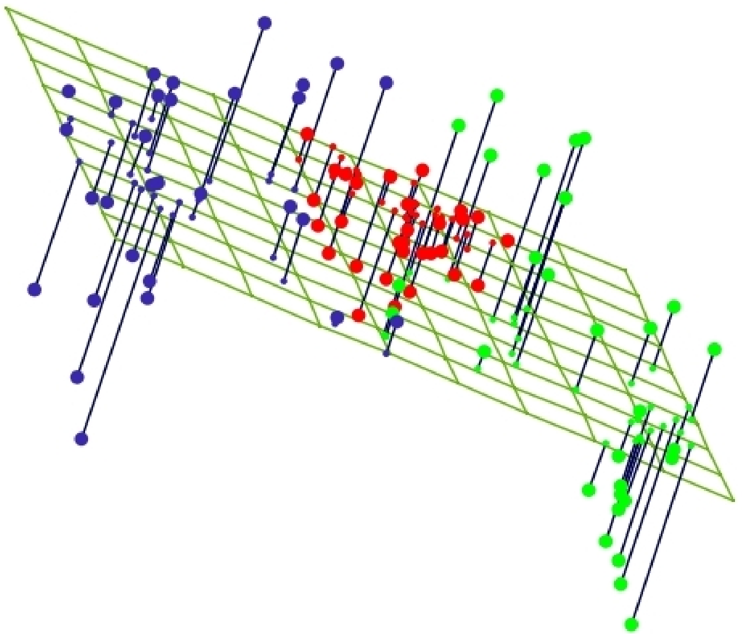
  Learning a lower-dimensional subspace for encoding high-dimensional data set

- 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

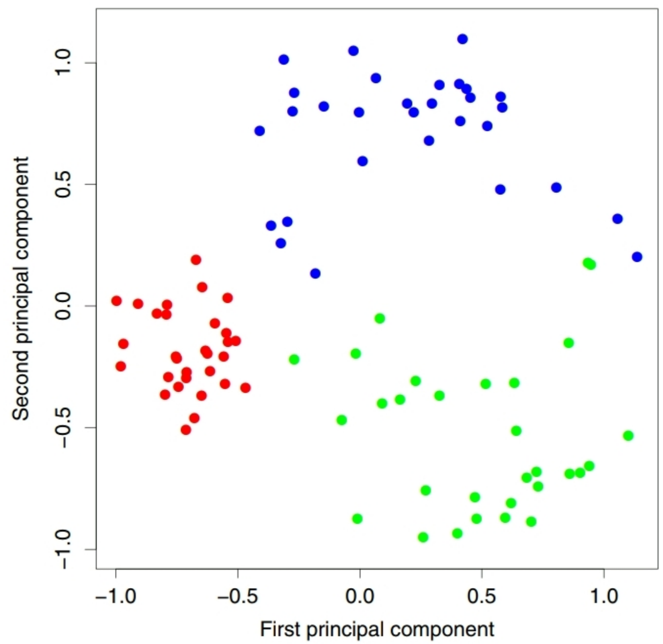
- We will introduce linear dimensionality reduction using Principal Component Analysis (PCA)

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



Projection on a linear subspace

$\xrightarrow{\text{PCA}}$

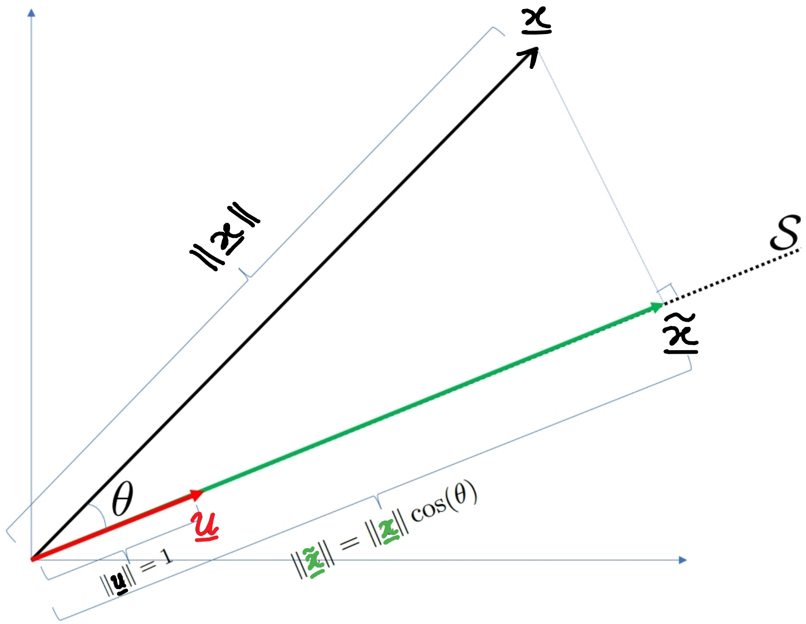


3D space

2D - space

Idea of projection

- Consider projection onto 1-D subspace of 2D



low dimensional

- 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 $\tilde{\underline{x}} = \text{Proj}_S(\underline{x})$

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

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

Idea of projection

from a D > M space

- How to project onto an 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:

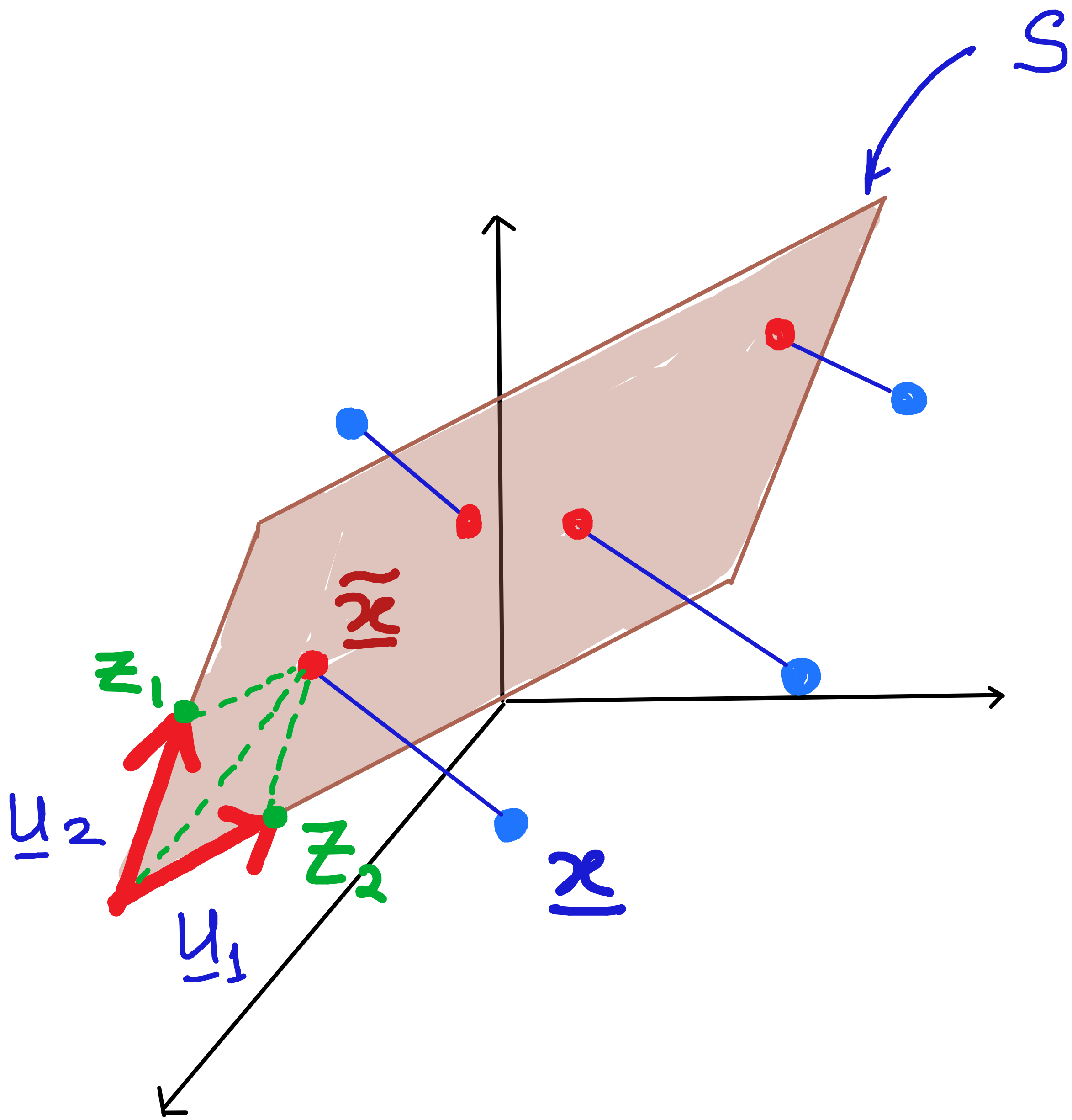
$$\underline{\tilde{x}} = \mathrm{Proj}_S(\underline{x}) = \sum_{i=1}^{M} z_i \underline{u}_i \qquad \text{where} \qquad z_i = \underline{x}^T \underline{u}_i$$

  - each u_i is the basis vector
  - z_i is the magnitude along that projection

- In vector form:

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

2D subspace in 3D space

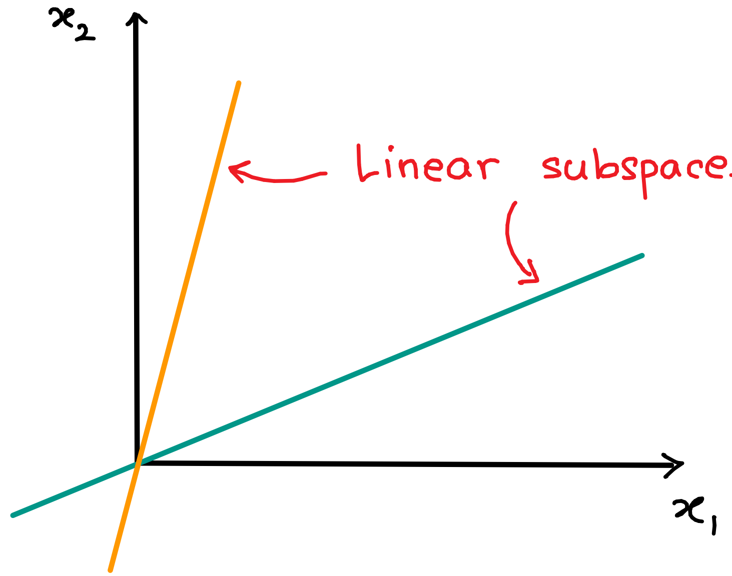


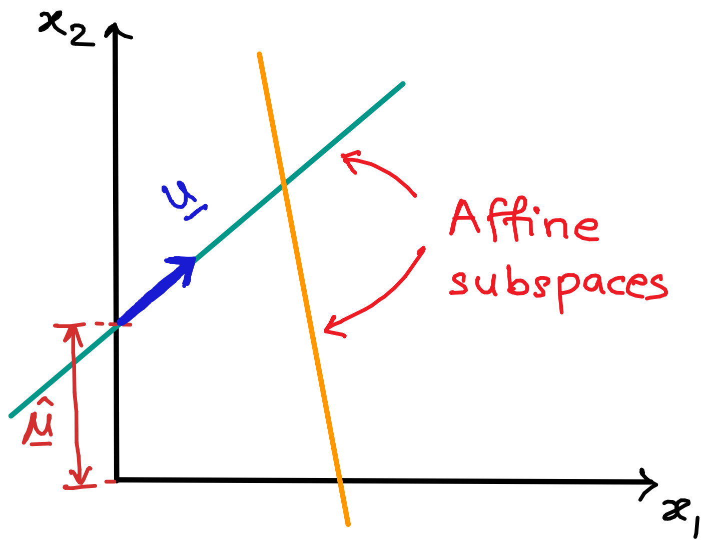
- Recall that every point in subspace S can be represented as $\{\underline{u}_1, \underline{u}_2\}$ & its linear combination

$$\tilde{\underline{x}} = z_1 \underline{u}_1 + z_2 \underline{u}_2$$

Projection onto an affine subspace

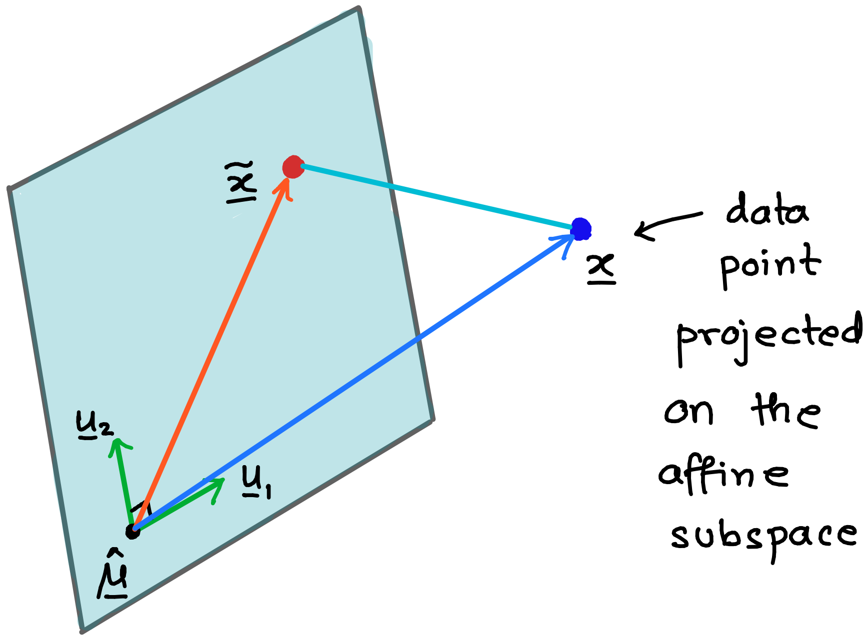
- So far, we have assumed a subspace that passes through zero
- However, the subspaces that we want to project onto can also be affine subspaces, which need not pass through zero





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

Projection onto an affine subspace



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

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

($\underline{z} = \underline{\underline{U}}^T \underline{x}$ earlier)

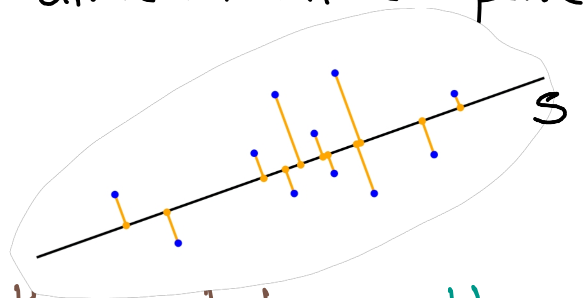
The affine subspace has an origin $\hat{\underline{\mu}}$

- $\tilde{\underline{x}}$ 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 $\tilde{\underline{x}}$

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

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

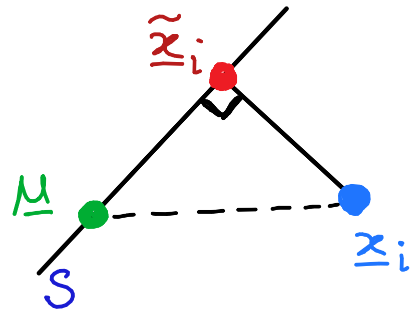


- How to choose such a subspace S?
  - We need to find appropriate $\underline{\hat{\mu}}$ and the orthogonal bases $\underline{\underline{U}}$
  - origin $\underline{\hat{\mu}}$ can be set equal to the mean of the dataset

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

$$\boxed{\arg\min_{\underline{\underline{U}}} \frac{1}{N} \sum_{i=1}^{N} \| \underline{x}_i - \underline{\tilde{x}}_i \|_2^2}$$

equivalent (show)



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

$$\boxed{\arg\max_{\underline{\underline{U}}} \frac{1}{N} \sum_{i=1}^{N} \| \underline{\tilde{x}}_i - \underline{\hat{\mu}} \|_2^2}$$

(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

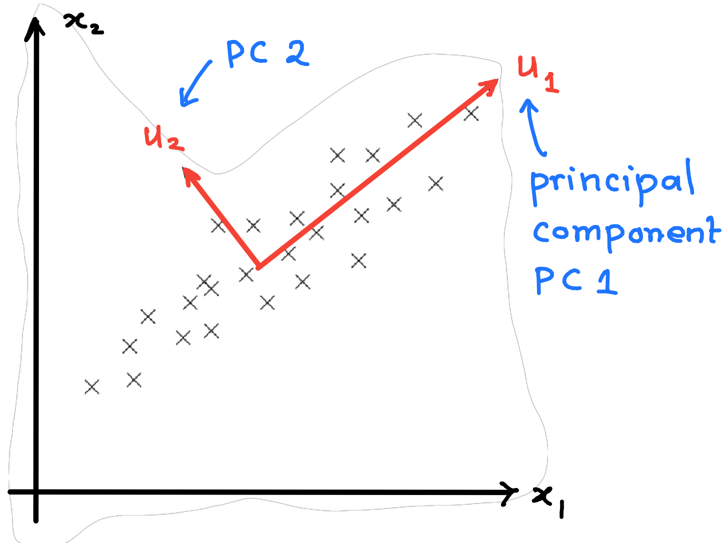
- Consider the sample covariance matrix:

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

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

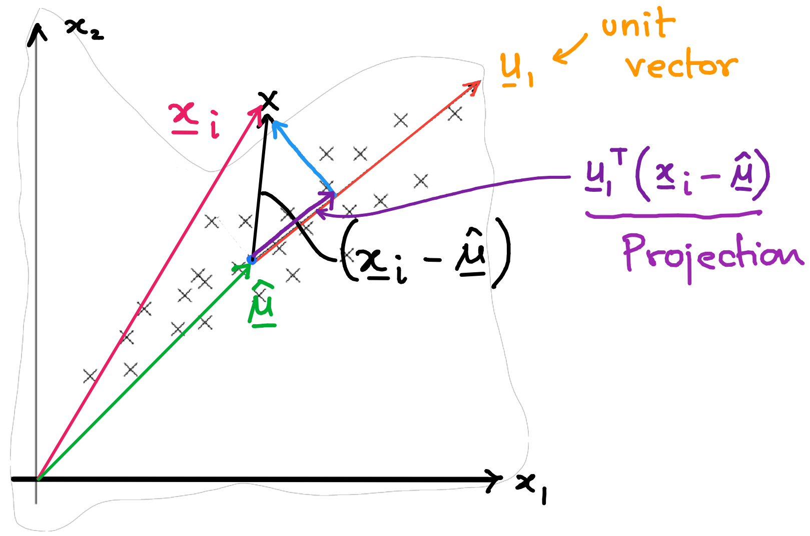
- The optimal PCA subspace is spanned by the top 'M' eigenvectors of $\hat{\underline{\underline{\Sigma}}}$

- 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 $\underline{u}_1$



- Lets maximize the projected variance

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

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

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

- So the optimization task is:

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

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

Take gradient and set to zero:

eigenvalue

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

eigenvector

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

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

- 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

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

Spectral decomposition

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

eigenvector matrix (orthonormal)

eigenvalues matrix

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

- 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