Principal Component Analysis

Yingzhen Li

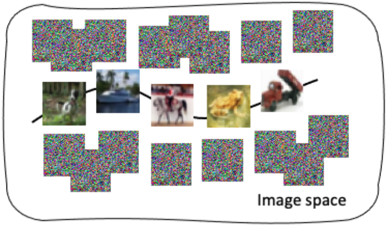
Department of Computing
Imperial College London

@liyzhen2
yingzhen.li@imperial.ac.uk

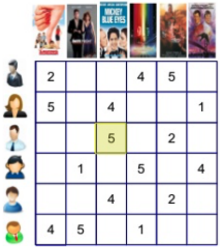
Nov 21, 2022

Dimensionality reduction

High-dimensional raw data are often sparse,
perhaps lying on a low-dimensional manifold:



natural images vs all RGB images

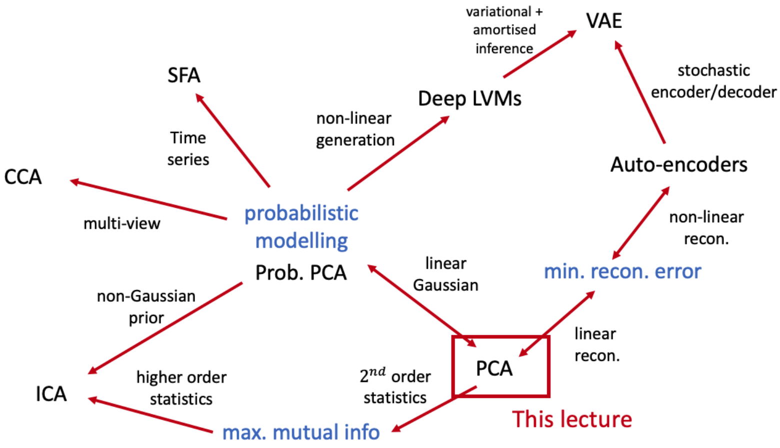


2			4	5	
5		4			1
		5		2	
	1		5		4
		4		2	
4	5		1		

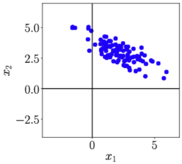
User ratings on items

Dimensionality reduction

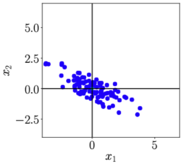
To name a few dimensionality reduction methods:



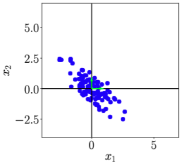
PCA in practise



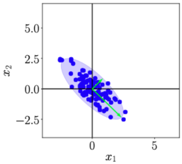
(a) Original dataset.



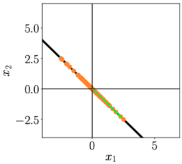
(b) Step 1: Centering by subtracting the mean from each data point.



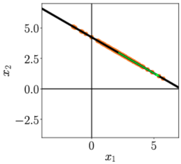
(c) Step 2: Dividing by the standard deviation to make the data unit free. Data has variance 1 along each axis.



(d) Step 3: Compute eigenvalues and eigenvectors (arrows) of the data covariance matrix (ellipse).



(e) Step 4: Project data onto the principal subspace.



(f) Undo the standardization and move projected data back into the original data space from (a).

Fig from the MML book.

PCA: set-up

Problem set-up:

- Data: $\mathcal{D} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}, \boldsymbol{x}_n \in \mathbb{R}^{D \times 1}$ s.t. $\text{mean}(\boldsymbol{x}_n) = \mathbf{0}$
- Find projections in a **lower-dimensional** space:

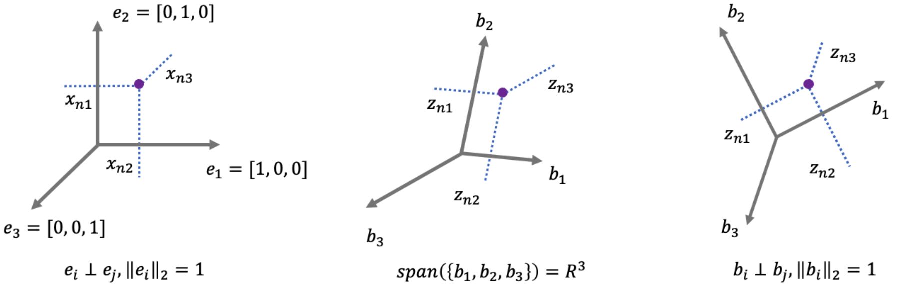
$$\boldsymbol{x}_n \approx \tilde{\boldsymbol{x}}_n := \sum_{j=1}^{M} z_{nj} \mathbf{b}_j, \quad z_{nj} := \mathbf{b}_j^\top \boldsymbol{x}_n$$

using an **orthonormal basis**

$$\mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_M], \quad \mathbf{b}_m \in \mathbb{R}^{D \times 1}, \quad M < D$$

Quick refresher: basis

For a given datapoint $\boldsymbol{x}_n = [x_{n1}, ..., x_{nD}]^\top \in \mathbb{R}^{D\times 1}$



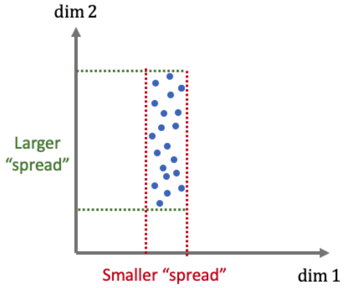
↑ orthonormal basis

Coordinates $\{z_{nj}\}$ are projections of the $\boldsymbol{x}_n$ vector onto a given basis:

$$\boldsymbol{x}_n = \sum_{j=1}^{D} z_{nj}\mathbf{b}_j, \quad z_{nj} := \mathbf{b}_j^\top \boldsymbol{x}_n$$

PCA: maximum variance perspective

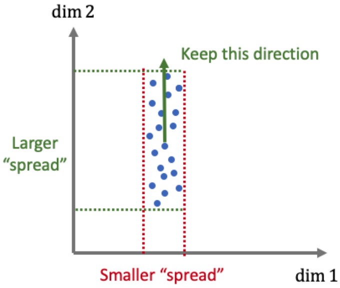
The "maximum variance" intuition of PCA:
project onto directions where the datapoints "vary the most"



"Spread" is defined as the variance along a given direction

PCA: maximum variance perspective

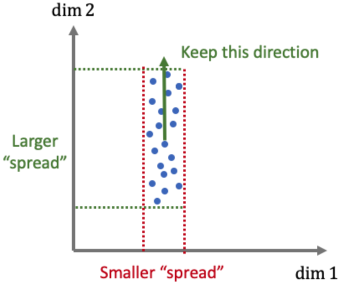
The "maximum variance" intuition of PCA:
project onto directions where the datapoints "vary the most"

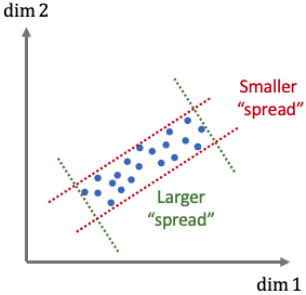


"Spread" is defined as the variance along a given direction

PCA: maximum variance perspective

The “maximum variance” intuition of PCA:
project onto directions where the datapoints “vary the most”

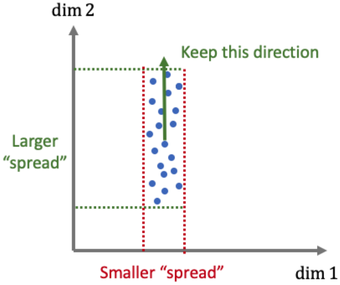


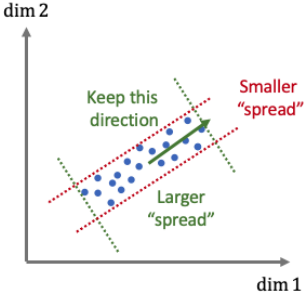


“Spread” is defined as the variance along a given direction

PCA: maximum variance perspective

The “maximum variance” intuition of PCA:
project onto directions where the datapoints “vary the most”





“Spread” is defined as the variance along a given direction

PCA: maximum variance perspective

Problem set-up:

- Data: $\mathcal{D} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$, $\boldsymbol{x}_n \in \mathbb{R}^{D \times 1}$ s.t. $\text{mean}(\boldsymbol{x}_n) = \mathbf{0}$
- Find projections in a **lower-dimensional** space:

$$z_n := \mathbf{B}^\top \boldsymbol{x}_n \quad \Leftrightarrow \quad z_{nj} := \mathbf{b}_j^\top \boldsymbol{x}_n$$

  using an **orthonormal basis**

$$\mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_M], \quad \mathbf{b}_m \in \mathbb{R}^{D \times 1}, \quad M < D$$

- Solve for $\mathbf{b}_1$ such that

$$\mathbb{V}[\mathbf{b}_1^\top \boldsymbol{x}_n] \quad \text{is maximised}$$

PCA: maximum variance perspective

Solve for $\mathbf{b}_1$ such that

$$\mathbb{V}[\mathbf{b}_1^\top \boldsymbol{x}_n] \quad \text{is maximised, subject to } ||\mathbf{b}_1||_2 = 1$$

- Variance after projection (recall that $\boldsymbol{x}_n$ has mean zero):

$$\mathbb{V}[\mathbf{b}_1^\top \boldsymbol{x}_n] := \frac{1}{N}\sum_{n=1}^{N}(\mathbf{b}_1^\top \boldsymbol{x}_n)^2 = \mathbf{b}_1^\top \underbrace{\left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n\boldsymbol{x}_n^\top\right)}_{:=\mathbf{S}=\mathbf{Q}\Lambda\mathbf{Q}^\top}\mathbf{b}_1$$

$$= \mathbf{b}_1^\top \mathbf{Q}\Lambda \underbrace{\mathbf{Q}^\top \mathbf{b}_1}_{:=\boldsymbol{\beta}_1} = \sum_{d=1}^{D}\lambda_d \beta_{1d}^2$$

- $||\mathbf{b}_1||_2^2 = 1 \quad \Rightarrow \quad ||\boldsymbol{\beta}_1||_2^2 = 1$

$$||\mathbf{b}_1||_2^2 := \mathbf{b}_1^\top \mathbf{b}_1 = \mathbf{b}_1^\top \underbrace{\mathbf{Q}\mathbf{Q}^\top}_{=\mathbf{I}} \mathbf{b}_1 = (\mathbf{Q}^\top \mathbf{b}_1)^\top (\underbrace{\mathbf{Q}^\top \mathbf{b}_1}_{:=\boldsymbol{\beta}_1}) = \boldsymbol{\beta}_j^\top \boldsymbol{\beta}_j = ||\boldsymbol{\beta}_1||_2^2$$

PCA: maximum variance perspective

Solve for $\mathbf{b}_1$ such that

$$\mathbb{V}[\mathbf{b}_1^\top \boldsymbol{x}_n] \quad \text{is maximised, subject to } ||\mathbf{b}_1||_2 = 1$$

- Equivalent to solving the following problem

$$\max_{\boldsymbol{\beta}_1} \sum_{d=1}^{D} \lambda_d \beta_{1d}^2 \quad \text{s.t.} ||\boldsymbol{\beta}_1||_2^2 = \sum_{d=1}^{D} \beta_{1d}^2 = 1.$$

- **Solution:** $\boldsymbol{\beta}_1 = \boldsymbol{e}_1 := (1, 0, ..., 0)^\top$

  $\Rightarrow \mathbf{b}_1 = \boldsymbol{q}_1$ (the eigenvector with the largest eigenvalue)

PCA: maximum variance perspective

Iteratively solve for the rest of the directions $\mathbf{b}_2, ...\mathbf{b}_M$:

For $m = 2, ..., M$:

- Compute the "remainder" of projection:

$$\hat{\boldsymbol{x}}_n = \boldsymbol{x}_n - \sum_{j=1}^{m-1} z_{nj}\mathbf{b}_j = \boldsymbol{x}_n - \sum_{j=1}^{m-1} (\mathbf{b}_j^\top \boldsymbol{x}_n)\mathbf{b}_j$$

PCA: maximum variance perspective

Iteratively solve for the rest of the directions $\mathbf{b}_2, ...\mathbf{b}_M$:

For $m = 2, ..., M$:

- Compute the "remainder" of projection:

$$\hat{\boldsymbol{x}}_n = \boldsymbol{x}_n - \sum_{j=1}^{m-1} z_{nj}\mathbf{b}_j = \boldsymbol{x}_n - \sum_{j=1}^{m-1} (\mathbf{b}_j^\top \boldsymbol{x}_n)\mathbf{b}_j$$

- maximise the following objective w.r.t. $\mathbf{b}_m$:

$$\max_{\mathbf{b}_m} \mathbb{V}[\mathbf{b}_m^\top \hat{\boldsymbol{x}}_n], \quad \text{s.t. } ||\mathbf{b}_m||_2 = 1, \mathbf{b}_m \perp \mathbf{b}_j, j = 1, ..., m-1$$

PCA: maximum variance perspective

Iteratively solve for the rest of the directions $\mathbf{b}_2, ...\mathbf{b}_M$:
For $m = 2, ..., M$:

- maximise $\mathbb{V}[\mathbf{b}_m^\top \hat{\boldsymbol{x}}_n]$, subject to $||\mathbf{b}_m||_2 = 1$, $\mathbf{b}_m \perp \mathbf{b}_j, j = 1, ..., m-1$
- Recall that $\boldsymbol{x}_n$ has mean zero:

$$\mathbb{V}[\mathbf{b}_m^\top \hat{\boldsymbol{x}}_n] := \frac{1}{N} \sum_{n=1}^{N} (\mathbf{b}_m^\top \boldsymbol{x}_n - \sum_{j=1}^{m-1} (\mathbf{b}_j^\top \boldsymbol{x}_n) \underbrace{\mathbf{b}_j^\top \mathbf{b}_m}_{=0})^2 = \mathbf{b}_m^\top \underbrace{(\frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top)}_{=\mathbf{S}=\mathbf{Q}\Lambda\mathbf{Q}^\top} \mathbf{b}_m$$

$$= \mathbf{b}_m^\top \mathbf{Q} \Lambda \underbrace{\mathbf{Q}^\top \mathbf{b}_m}_{:=\boldsymbol{\beta}_m} = \sum_{d=1}^{D} \lambda_d \beta_{md}^2$$

- Here $\Lambda = diag(\lambda_1, ..., \lambda_d)$ and $\lambda_1 \geqslant ... \geqslant \lambda_D \geqslant 0$.

PCA: maximum variance perspective

Iteratively solve for the rest of the directions $\mathbf{b}_2, ...\mathbf{b}_M$:

For $m = 2, ..., M$:

- maximise $\mathbb{V}[\mathbf{b}_m^\top \hat{\boldsymbol{x}}_n]$, subject to $||\mathbf{b}_m||_2 = 1$, $\mathbf{b}_m \perp \mathbf{b}_j, j = 1, ..., m-1$
- $||\mathbf{b}_m||_2^2 = 1 \quad \Rightarrow \quad ||\boldsymbol{\beta}_m||_2^2 = 1$

$$||\mathbf{b}_m||_2^2 := \mathbf{b}_m^\top \mathbf{b}_m = \mathbf{b}_m^\top \underbrace{\mathbf{Q}\mathbf{Q}^\top}_{=\mathbf{I}} \mathbf{b}_m = (\mathbf{Q}^\top \mathbf{b}_m)^\top (\underbrace{\mathbf{Q}^\top \mathbf{b}_m}_{:=\boldsymbol{\beta}_m}) = \boldsymbol{\beta}_m^\top \boldsymbol{\beta}_m = ||\boldsymbol{\beta}_m||_2^2$$

- $\mathbf{b}_m \perp \mathbf{b}_j \quad \Rightarrow \quad \mathbf{b}_m^\top \mathbf{b}_j = 0 \quad \Rightarrow \quad \boldsymbol{\beta}_m^\top \boldsymbol{\beta}_j = 0$

$$\mathbf{b}_m^\top \mathbf{b}_j = \mathbf{b}_m^\top \underbrace{\mathbf{Q}\mathbf{Q}^\top}_{=\mathbf{I}} \mathbf{b}_j = (\underbrace{\mathbf{Q}^\top \mathbf{b}_m}_{:=\boldsymbol{\beta}_m})^\top (\underbrace{\mathbf{Q}^\top \mathbf{b}_j}_{:=\boldsymbol{\beta}_j}) = \boldsymbol{\beta}_m^\top \boldsymbol{\beta}_j$$

PCA: maximum variance perspective

Iteratively solve for the rest of the directions $\mathbf{b}_2, ...\mathbf{b}_M$:

For $m = 2, ..., M$:

- maximise $\mathbb{V}[\mathbf{b}_m^\top \boldsymbol{x}_n]$, subject to $||\mathbf{b}_m||_2 = 1$, $\mathbf{b}_m \perp \mathbf{b}_j, j = 1, ..., m-1$
- Equivalent to the following optimisation problem:

$$\max_{\boldsymbol{\beta}_m} \sum_{d=1}^{D} \lambda_d \beta_{md}^2 \quad \text{s.t.} ||\boldsymbol{\beta}_m||_2^2 = 1, \boldsymbol{\beta}_m^\top \boldsymbol{\beta}_j = 0, j = 1, ..., m-1.$$

Proof by induction: we show $\boldsymbol{\beta}_m = \boldsymbol{e}_m := (0, ..., 0, \underbrace{1}_{m\text{th element}}, 0, ..., 0)$

1. $\boldsymbol{\beta}_1 = \boldsymbol{e}_1 \quad \Rightarrow \quad \mathbf{b}_1 = \boldsymbol{q}_1$

PCA: maximum variance perspective

Iteratively solve for the rest of the directions $\mathbf{b}_2, ...\mathbf{b}_M$:

For $m = 2, ..., M$:

- maximise $\mathbb{V}[\mathbf{b}_m^\top \boldsymbol{x}_n]$, subject to $||\mathbf{b}_m||_2 = 1$, $\mathbf{b}_m \perp \mathbf{b}_j, j = 1, ..., m-1$
- Equivalent to the following optimisation problem:

$$\max_{\boldsymbol{\beta}_m} \sum_{d=1}^{D} \lambda_d \beta_{md}^2 \quad \text{s.t.} ||\boldsymbol{\beta}_m||_2^2 = 1, \boldsymbol{\beta}_m^\top \boldsymbol{\beta}_j = 0, j = 1, ..., m-1.$$

Proof by induction: we show $\boldsymbol{\beta}_m = \boldsymbol{e}_m := (0, ..., 0, \underbrace{1}_{m\text{th element}}, 0, ..., 0)$

1. $\boldsymbol{\beta}_1 = \boldsymbol{e}_1 \quad \Rightarrow \quad \mathbf{b}_1 = \boldsymbol{q}_1$
2. For $m = 2, ..., M$, assume $\boldsymbol{\beta}_j = \boldsymbol{e}_j, j = 1, ..., m-1$

    2a. $\boldsymbol{\beta}_m^\top \boldsymbol{\beta}_j = 0, j = 1, ..., m-1 \quad \Rightarrow \quad \beta_{mj} = 0, j = 1, ..., m-1$

PCA: maximum variance perspective

Iteratively solve for the rest of the directions $\mathbf{b}_2, ...\mathbf{b}_M$:

For $m = 2, ..., M$:

- maximise $\mathbb{V}[\mathbf{b}_m^\top \boldsymbol{x}_n]$, subject to $||\mathbf{b}_m||_2 = 1$, $\mathbf{b}_m \perp \mathbf{b}_j, j = 1, ..., m-1$
- Equivalent to the following optimisation problem:

$$\max_{\boldsymbol{\beta}_m} \sum_{d=1}^{D} \lambda_d \beta_{md}^2 \quad \text{s.t.} ||\boldsymbol{\beta}_m||_2^2 = 1, \boldsymbol{\beta}_m^\top \boldsymbol{\beta}_j = 0, j = 1, ..., m-1.$$

Proof by induction: we show $\boldsymbol{\beta}_m = \boldsymbol{e}_m := (0, ..., 0, \underbrace{1}_{m\text{th element}}, 0, ..., 0)$

1. $\boldsymbol{\beta}_1 = \boldsymbol{e}_1 \quad \Rightarrow \quad \mathbf{b}_1 = \boldsymbol{q}_1$
2. For $m = 2, ..., M$, assume $\boldsymbol{\beta}_j = \boldsymbol{e}_j, j = 1, ..., m-1$

   2a. $\boldsymbol{\beta}_m^\top \boldsymbol{\beta}_j = 0, j = 1, ..., m-1 \quad \Rightarrow \quad \boldsymbol{\beta}_{mj} = 0, j = 1, ..., m-1$

   2b. $||\boldsymbol{\beta}_m||_2 = 1 \quad \Rightarrow \quad \sum_{d=m}^{D} \beta_{md}^2 = 1$

PCA: maximum variance perspective

Iteratively solve for the rest of the directions $\mathbf{b}_2, ...\mathbf{b}_M$:

For $m = 2, ..., M$:

- maximise $\mathbb{V}[\mathbf{b}_m^\top \boldsymbol{x}_n]$, subject to $||\mathbf{b}_m||_2 = 1$, $\mathbf{b}_m \perp \mathbf{b}_j, j = 1, ..., m-1$
- Equivalent to the following optimisation problem:

$$\max_{\boldsymbol{\beta}_m} \sum_{d=1}^{D} \lambda_d \beta_{md}^2 \quad \text{s.t.} ||\boldsymbol{\beta}_m||_2^2 = 1, \boldsymbol{\beta}_m^\top \boldsymbol{\beta}_j = 0, j = 1, ..., m-1.$$

Proof by induction: we show $\boldsymbol{\beta}_m = \boldsymbol{e}_m := (0, ..., 0, \underbrace{1}_{m\text{th element}}, 0, ..., 0)$

1. $\boldsymbol{\beta}_1 = \boldsymbol{e}_1 \quad \Rightarrow \quad \mathbf{b}_1 = \boldsymbol{q}_1$
2. For $m = 2, ..., M$, assume $\boldsymbol{\beta}_j = \boldsymbol{e}_j, j = 1, ..., m-1$
   - 2a. $\boldsymbol{\beta}_m^\top \boldsymbol{\beta}_j = 0, j = 1, ..., m-1 \quad \Rightarrow \quad \beta_{mj} = 0, j = 1, ..., m-1$
   - 2b. $||\boldsymbol{\beta}_m||_2 = 1 \quad \Rightarrow \quad \sum_{d=m}^{D} \beta_{md}^2 = 1$
   - 2c. Solve for maximum of $\sum_{d=m}^{D} \lambda_d \beta_{md}^2$ w.r.t. β_{md}:

     Solution: $\boldsymbol{\beta}_m = \boldsymbol{e}_m \quad \Rightarrow \quad \mathbf{b}_m = \boldsymbol{q}_m$

PCA: maximum variance perspective

For $m = 1, ..., M$:

- maximise $\mathbb{V}[\mathbf{b}_m^\top \boldsymbol{x}_n]$, subject to $||\mathbf{b}_m||_2 = 1$, $\mathbf{b}_m \perp \mathbf{b}_j, j = 1, ..., m-1$

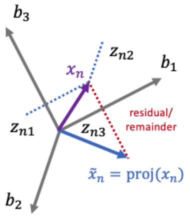
Solutions: $\mathbf{b}_m = \boldsymbol{q}_m$ for $m = 1, ..., M$

$\Rightarrow$ Projecting $\boldsymbol{x}_n$ to a subspace

$$span(\{\boldsymbol{q}_m\}_{m=1}^{M}) = span(\{\boldsymbol{q}_j\}_{j=M+1}^{D})^{\perp}$$

$$\boldsymbol{x}_n = \underbrace{\sum_{j=1}^{M} z_{nj}\boldsymbol{q}_j}_{:=\tilde{\boldsymbol{x}}_n} + \underbrace{\sum_{j=M+1}^{D} z_{nj}\mathbf{b}_j}_{\text{dropped}}, \quad \mathbf{b}_i \perp \boldsymbol{q}_j$$

$$\tilde{\boldsymbol{x}}_n \in span(\{\boldsymbol{q}_m\}_{m=1}^{M})$$



$b_i \perp b_j, \|b_i\|_2 = 1$

$span(\{b_1, b_2, b_3\}) = R^3$

PCA: minimum reconstruction error perspective

Goal: find orthonormal basis $\{\mathbf{b}_1, ..., \mathbf{b}_M\}$ to minimise ℓ_2 reconstruction error:

$$L = \frac{1}{N}\sum_{n=1}^{N} ||\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n||_2^2, \quad \tilde{\boldsymbol{x}}_n := \sum_{j=1}^{M} z_{nj}\mathbf{b}_j, \; z_{nj} = \mathbf{b}_j^\top \boldsymbol{x}_n.$$

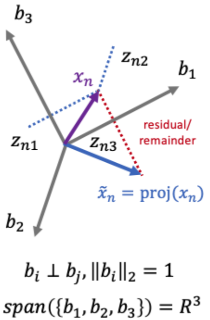
Rewriting the loss:

- Consider the full orthonormal basis:

$$\mathbf{B}_{full} = [\underbrace{\mathbf{b}_1, ..., \mathbf{b}_M}_{\text{will be used in new basis}}, \underbrace{\mathbf{b}_{M+1}, ..., \mathbf{b}_D}_{\text{will be dropped}}]$$

- Representing $\boldsymbol{x}_n$ using basis $\mathbf{B}_{full}$:

$$\boldsymbol{x}_n = \underbrace{\sum_{j=1}^{M} z_{nj}\mathbf{b}_j}_{=\tilde{\boldsymbol{x}}_n} + \sum_{j=M+1}^{D} z_{nj}\mathbf{b}_j, \quad z_{nj} := \mathbf{b}_j^\top \boldsymbol{x}_n$$



PCA: minimum reconstruction error perspective

Goal: find orthonormal basis $\{\mathbf{b}_1, ..., \mathbf{b}_M\}$ to minimise ℓ_2 reconstruction error:

$$L = \frac{1}{N}\sum_{n=1}^{N} ||\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n||_2^2, \quad \tilde{\boldsymbol{x}}_n := \sum_{j=1}^{M} z_{nj}\mathbf{b}_j, \; z_{nj} = \mathbf{b}_j^\top \boldsymbol{x}_n.$$

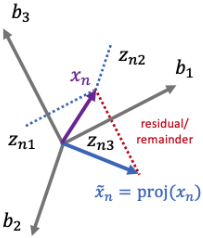
Rewriting the loss:

- Consider the full orthonormal basis:

$$\mathbf{B}_{full} = [\underbrace{\mathbf{b}_1, ..., \mathbf{b}_M}_{\text{will be used in new basis}}, \underbrace{\mathbf{b}_{M+1}, ..., \mathbf{b}_D}_{\text{will be dropped}}]$$

- Representing $\boldsymbol{x}_n$ using basis $\mathbf{B}_{full}$:

$$\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n = \sum_{j=M+1}^{D} z_{nj}\mathbf{b}_j, \quad z_{nj} := \mathbf{b}_j^\top \boldsymbol{x}_n$$



$b_i \perp b_j, \|b_i\|_2 = 1$

$span(\{b_1, b_2, b_3\}) = R^3$

PCA: minimum reconstruction error perspective

Goal: find orthonormal basis $\{\mathbf{b}_1, ..., \mathbf{b}_M\}$ to minimise ℓ_2 reconstruction error:

$$L = \frac{1}{N} \sum_{n=1}^{N} ||\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n||_2^2, \quad \tilde{\boldsymbol{x}}_n := \sum_{j=1}^{M} z_{nj} \mathbf{b}_j, \; z_{nj} = \mathbf{b}_j^\top \boldsymbol{x}_n.$$

Rewriting the loss:
First notice that $\mathbf{B}_{full}$ is an **orthonormal** basis:

$$L = \frac{1}{N} \sum_{n=1}^{N} || \sum_{j=M+1}^{D} z_{nj} \mathbf{b}_j ||_2^2$$

$$= \frac{1}{N} \sum_{n=1}^{N} \sum_{j=M+1}^{D} ||z_{nj} \mathbf{b}_j||_2^2$$

$$= \frac{1}{N} \sum_{n=1}^{N} \sum_{j=M+1}^{D} z_{nj}^2$$

PCA: minimum reconstruction error perspective

Goal: find orthonormal basis $\{\mathbf{b}_1, ..., \mathbf{b}_M\}$ to minimise ℓ_2 reconstruction error:

$$L = \frac{1}{N}\sum_{n=1}^{N} ||\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n||_2^2, \quad \tilde{\boldsymbol{x}}_n := \sum_{j=1}^{M} z_{nj}\mathbf{b}_j, \ z_{nj} = \mathbf{b}_j^\top \boldsymbol{x}_n.$$

Rewriting the loss:
Plugging-in that $z_{nj} = \mathbf{b}_j^\top \boldsymbol{x}_n$:

$$L = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=M+1}^{D} (\mathbf{b}_j^\top \boldsymbol{x}_n)^2 = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=M+1}^{D} \mathbf{b}_j^\top (\boldsymbol{x}_n \boldsymbol{x}_n^\top)\mathbf{b}_j$$

$$= \sum_{j=M+1}^{D} \mathbf{b}_j^\top \underbrace{(\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top)}_{:=\mathbf{S}=\mathbf{Q}\Lambda\mathbf{Q}}\mathbf{b}_j = \sum_{j=M+1}^{D} \mathbf{b}_j^\top \mathbf{Q}\Lambda \underbrace{\mathbf{Q}^\top \mathbf{b}_j}_{:=\boldsymbol{\beta}_j} = \sum_{j=M+1}^{D}\sum_{d=1}^{D} \lambda_d \beta_{jd}^2$$

PCA: minimum reconstruction error perspective

Assume the eigenvalue decomposition as $\mathbf{S} = \mathbf{Q}\Lambda\mathbf{Q}^\top$,
with $\Lambda = diag([\lambda_1, ..., \lambda_D]), \lambda_1 \geqslant \cdots \geqslant \lambda_D$

$$\mathbf{b}_j^\top \mathbf{S}\mathbf{b}_j = \mathbf{b}_j^\top \mathbf{Q}\Lambda\mathbf{Q}^\top \mathbf{b}_j := \boldsymbol{\beta}_j^\top \Lambda \boldsymbol{\beta}_j = \sum_{d=1}^{D} \lambda_d \beta_{jd}^2$$

$$\boldsymbol{\beta}_j := \mathbf{Q}^\top \mathbf{b}_j = [\beta_{j1}, ..., \beta_{jD}] = [\boldsymbol{q}_1^\top \mathbf{b}_j, ..., \boldsymbol{q}_D^\top \mathbf{b}_j]$$

- $||\mathbf{b}_j||_2^2 = 1 \quad \Rightarrow \quad ||\boldsymbol{\beta}_j||_2^2 = 1$

$$||\mathbf{b}_j||_2^2 := \mathbf{b}_j^\top \mathbf{b}_j = \mathbf{b}_j^\top \underbrace{\mathbf{Q}\mathbf{Q}^\top}_{=\mathbf{I}} \mathbf{b}_j = (\mathbf{Q}^\top \mathbf{b}_j)^\top (\underbrace{\mathbf{Q}^\top \mathbf{b}_j}_{:=\boldsymbol{\beta}_j}) = \boldsymbol{\beta}_j^\top \boldsymbol{\beta}_j = ||\boldsymbol{\beta}_j||_2^2$$

- $\mathbf{b}_i \perp \mathbf{b}_j \quad \Rightarrow \quad \mathbf{b}_i^\top \mathbf{b}_j = 0 \quad \Rightarrow \quad \boldsymbol{\beta}_i^\top \boldsymbol{\beta}_j = 0$

$$\mathbf{b}_i^\top \mathbf{b}_j = \mathbf{b}_i^\top \underbrace{\mathbf{Q}\mathbf{Q}^\top}_{=\mathbf{I}} \mathbf{b}_j = (\underbrace{\mathbf{Q}^\top \mathbf{b}_i}_{:=\boldsymbol{\beta}_i})^\top (\underbrace{\mathbf{Q}^\top \mathbf{b}_j}_{:=\boldsymbol{\beta}_j}) = \boldsymbol{\beta}_i^\top \boldsymbol{\beta}_j$$

PCA: minimum reconstruction error perspective

$$\min_{\boldsymbol{\beta}_{M+1:D}} L = \sum_{j=M+1}^{D} \sum_{d=1}^{D} \lambda_d \beta_{jd}^2, \quad \text{s.t. } ||\boldsymbol{\beta}_j||_2^2 = 1, \ \boldsymbol{\beta}_i^\top \boldsymbol{\beta}_j = 0.$$

An iterative approach for solutions:
Solve $\boldsymbol{\beta}_D$ first and then solve for $\boldsymbol{\beta}_j$ for $j = D-1, ..., M+1$.

- Optimisation objective for $\boldsymbol{\beta}_D$:

$$\min_{\boldsymbol{\beta}_D} \sum_{d=1}^{D} \lambda_d \beta_{Dd}^2, \quad \text{s.t. } ||\boldsymbol{\beta}_D||_2^2 = \sum_{d=1}^{D} \beta_{Dd}^2 = 1$$

- Notice: $\lambda_1 \geqslant \cdots \geqslant \lambda_D$
- **Solution**: $\boldsymbol{\beta}_D = \boldsymbol{e}_D := (0, ..., 0, 1)^\top$
  $\Rightarrow \ \mathbf{b}_D = \boldsymbol{q}_D$ (the eigenvector with the smallest eigenvalue)

PCA: minimum reconstruction error perspective

$$\min_{\boldsymbol{\beta}_{M+1:D}} L = \sum_{j=M+1}^{D} \sum_{d=1}^{D} \lambda_d \beta_{jd}^2, \quad \text{s.t. } ||\boldsymbol{\beta}_j||_2^2 = 1, \ \boldsymbol{\beta}_i^\top \boldsymbol{\beta}_j = 0.$$

Proof by induction: for $j = D, D-1, ..., M+1$, $\boldsymbol{\beta}_j = \boldsymbol{e}_j$, i.e., $\mathbf{b}_j = \boldsymbol{q}_j$

1. For $j = D$: $\boldsymbol{\beta}_D = \boldsymbol{e}_D$, i.e., $\mathbf{b}_D = \boldsymbol{q}_D$
2. For $j = D-1, ..., M+1$, assume for $i > j$, $\boldsymbol{\beta}_i = \boldsymbol{e}_i$, i.e., $\mathbf{b}_i = \boldsymbol{q}_i$

   2a. $\boldsymbol{\beta}_i^\top \boldsymbol{\beta}_j = 0, i > j \quad \Rightarrow \quad \boldsymbol{\beta}_j = (\beta_{j1}, ..., \beta_{jj}, 0, ..., 0)^\top$

PCA: minimum reconstruction error perspective

$$\min_{\boldsymbol{\beta}_{M+1:D}} L = \sum_{j=M+1}^{D} \sum_{d=1}^{D} \lambda_d \beta_{jd}^2, \quad \text{s.t. } ||\boldsymbol{\beta}_j||_2^2 = 1,\ \boldsymbol{\beta}_i^\top \boldsymbol{\beta}_j = 0.$$

Proof by induction: for $j = D, D-1, ..., M+1$, $\boldsymbol{\beta}_j = \boldsymbol{e}_j$, i.e., $\mathbf{b}_j = \boldsymbol{q}_j$

1. For $j = D$: $\boldsymbol{\beta}_D = \boldsymbol{e}_D$, i.e., $\mathbf{b}_D = \boldsymbol{q}_D$
2. For $j = D-1, ..., M+1$, assume for $i > j$, $\boldsymbol{\beta}_i = \boldsymbol{e}_i$, i.e., $\mathbf{b}_i = \boldsymbol{q}_i$

   2a. $\boldsymbol{\beta}_i^\top \boldsymbol{\beta}_j = 0, i > j \quad \Rightarrow \quad \boldsymbol{\beta}_j = (\beta_{j1}, ..., \beta_{jj}, 0, ..., 0)^\top$

   2b. $||\boldsymbol{\beta}_j||_2^2 = 1 \quad \Rightarrow \quad \sum_{d=1}^{j} \beta_{jd}^2 = 1$

PCA: minimum reconstruction error perspective

$$\min_{\boldsymbol{\beta}_{M+1:D}} L = \sum_{j=M+1}^{D} \sum_{d=1}^{D} \lambda_d \beta_{jd}^2, \quad \text{s.t. } ||\boldsymbol{\beta}_j||_2^2 = 1,\ \boldsymbol{\beta}_i^\top \boldsymbol{\beta}_j = 0.$$

Proof by induction: for $j = D, D-1, ..., M+1$, $\boldsymbol{\beta}_j = \boldsymbol{e}_j$, i.e., $\mathbf{b}_j = \boldsymbol{q}_j$

1. For $j = D$: $\boldsymbol{\beta}_D = \boldsymbol{e}_D$, i.e., $\mathbf{b}_D = \boldsymbol{q}_D$
2. For $j = D-1, ..., M+1$, assume for $i > j$, $\boldsymbol{\beta}_i = \boldsymbol{e}_i$, i.e., $\mathbf{b}_i = \boldsymbol{q}_i$

   2a. $\boldsymbol{\beta}_i^\top \boldsymbol{\beta}_j = 0, i > j \quad \Rightarrow \quad \boldsymbol{\beta}_j = (\beta_{j1}, ..., \beta_{jj}, 0, ..., 0)^\top$

   2b. $||\boldsymbol{\beta}_j||_2^2 = 1 \quad \Rightarrow \quad \sum_{d=1}^{j} \beta_{jd}^2 = 1$

   2c. Solve for the following minimisation problem w.r.t. β_{jd}:

PCA: minimum reconstruction error perspective

$$\min_{\boldsymbol{\beta}_{M+1:D}} L = \sum_{j=M+1}^{D} \sum_{d=1}^{D} \lambda_d \beta_{jd}^2, \quad \text{s.t. } ||\boldsymbol{\beta}_j||_2^2 = 1,\ \boldsymbol{\beta}_i^\top \boldsymbol{\beta}_j = 0.$$

Proof by induction: for $j = D, D-1, ..., M+1$, $\boldsymbol{\beta}_j = \boldsymbol{e}_j$, i.e., $\mathbf{b}_j = \boldsymbol{q}_j$

1. For $j = D$: $\boldsymbol{\beta}_D = \boldsymbol{e}_D$, i.e., $\mathbf{b}_D = \boldsymbol{q}_D$
2. For $j = D-1, ..., M+1$, assume for $i > j$, $\boldsymbol{\beta}_i = \boldsymbol{e}_i$, i.e., $\mathbf{b}_i = \boldsymbol{q}_i$

   2a. $\boldsymbol{\beta}_i^\top \boldsymbol{\beta}_j = 0, i > j \quad \Rightarrow \quad \boldsymbol{\beta}_j = (\beta_{j1}, ..., \beta_{jj}, 0, ..., 0)^\top$

   2b. $||\boldsymbol{\beta}_j||_2^2 = 1 \quad \Rightarrow \quad \sum_{d=1}^{j} \beta_{jd}^2 = 1$

   2c. Solve for the following minimisation problem w.r.t. β_{jd}:

$$\min_{\boldsymbol{\beta}_j} \sum_{d=1}^{j} \lambda_d \beta_{jd}^2, \quad \text{s.t. } \sum_{d=1}^{j} \beta_{jd}^2 = 1$$

PCA: minimum reconstruction error perspective

$$\min_{\boldsymbol{\beta}_{M+1:D}} L = \sum_{j=M+1}^{D} \sum_{d=1}^{D} \lambda_d \beta_{jd}^2, \quad \text{s.t. } ||\boldsymbol{\beta}_j||_2^2 = 1, \ \boldsymbol{\beta}_i^\top \boldsymbol{\beta}_j = 0.$$

Proof by induction: for $j = D, D-1, ..., M+1$, $\boldsymbol{\beta}_j = \boldsymbol{e}_j$, i.e., $\mathbf{b}_j = \boldsymbol{q}_j$

1. For $j = D$: $\boldsymbol{\beta}_D = \boldsymbol{e}_D$, i.e., $\mathbf{b}_D = \boldsymbol{q}_D$
2. For $j = D-1, ..., M+1$, assume for $i > j$, $\boldsymbol{\beta}_i = \boldsymbol{e}_i$, i.e., $\mathbf{b}_i = \boldsymbol{q}_i$

    2a. $\boldsymbol{\beta}_i^\top \boldsymbol{\beta}_j = 0, i > j \quad \Rightarrow \quad \boldsymbol{\beta}_j = (\beta_{j1}, ..., \beta_{jj}, 0, ..., 0)^\top$

    2b. $||\boldsymbol{\beta}_j||_2^2 = 1 \quad \Rightarrow \quad \sum_{d=1}^{j} \beta_{jd}^2 = 1$

    2c. Solve for the following minimisation problem w.r.t. β_{jd}:

    Solution: $\boldsymbol{\beta}_j = \boldsymbol{e}_j$, i.e., $\mathbf{b}_j = \boldsymbol{q}_j$

PCA: minimum reconstruction error perspective

$$\min_{\mathbf{B}_{full}} L = \frac{1}{N}\sum_{n=1}^{N} ||\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n||_2^2, \quad \tilde{\boldsymbol{x}}_n := \sum_{j=1}^{M} z_{nj}\mathbf{b}_j$$

$$\text{s.t. } ||\mathbf{b}_j||_2^2 = 1, \mathbf{b}_i \perp \mathbf{b}_j$$

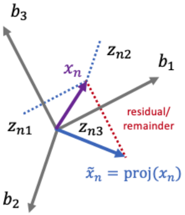
Solutions: $\mathbf{b}_j = \boldsymbol{q}_j$ for $j = M+1, ..., D$

$\Rightarrow$ Projecting $\boldsymbol{x}_n$ to an orthogonal complement space

$$span(\{\boldsymbol{q}_j\}_{j=M+1}^{D})^{\perp} = \{\boldsymbol{x} \in \mathbb{R}^{D\times 1} : \boldsymbol{x}^{\top}\boldsymbol{q}_j = 0, j = M+1, ..., D\}$$

$$\boldsymbol{x}_n = \underbrace{\sum_{j=1}^{M} z_{nj}\mathbf{b}_j}_{:=\tilde{\boldsymbol{x}}_n} + \underbrace{\sum_{j=M+1}^{D} z_{nj}\boldsymbol{q}_j}_{\text{dropped}}, \quad \mathbf{b}_i \perp \boldsymbol{q}_j$$

$$\tilde{\boldsymbol{x}}_n \in span(\{\boldsymbol{q}_j\}_{j=M+1}^{D})^{\perp}$$



$b_i \perp b_j, \|b_i\|_2 = 1$

$span(\{b_1, b_2, b_3\}) = R^3$

PCA: comparing both views

- Maximum variance view:

$$\mathbf{B}^*_{full} = \{\boldsymbol{q}_1, ..., \boldsymbol{q}_M, \mathbf{b}_{M+1}, ..., \mathbf{b}_D\}, \quad \mathbf{b}_i \perp \mathbf{b}_j, \mathbf{b}_i \perp \boldsymbol{q}_j$$

- Minimum reconstruction error view:

$$\mathbf{B}^*_{full} = \{\mathbf{b}_1, ..., \mathbf{b}_M, \boldsymbol{q}_{M+1}, ..., \boldsymbol{q}_D\}, \quad \mathbf{b}_i \perp \mathbf{b}_j, \mathbf{b}_i \perp \boldsymbol{q}_j$$

- No unique solution! By convention we often use $\mathbf{B}^*_{full} = \mathbf{Q}$
- Relates to the equivalence between PCA and **linear auto-encoder** (exercise for you)