


Principal Component Analysis

Yingzhen Li

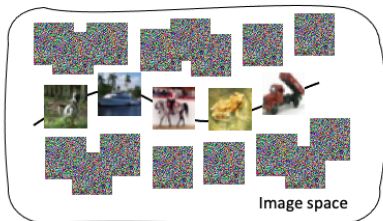
Department of Computing
Imperial College London

@liyzhen2
yingzhen.li@imperial.ac.uk

Nov 2, 2021

Dimensionality reduction

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



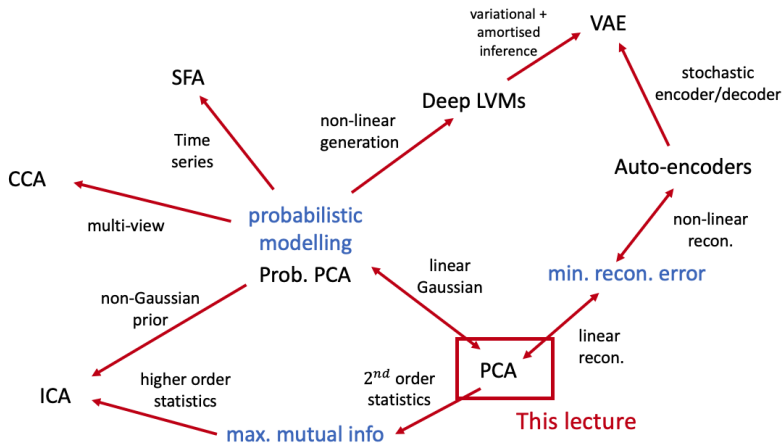
natural images vs all RGB images

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

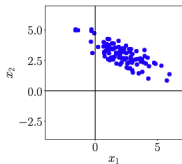
User ratings on items

Dimensionality reduction

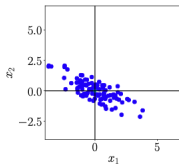
To name a few dimensionality reduction methods:



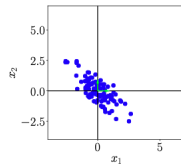
PCA in practise



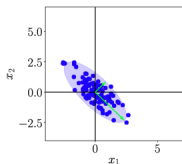
(a) Original dataset.



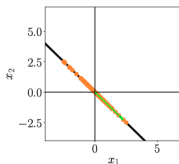
(b) Step 1: Centering by subtracting the mean from each data point.



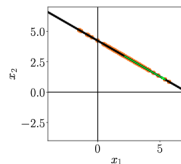
(c) Step 2: Dividing by the standard deviation to make the data unit free. Data has variance 1 along each axis.



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

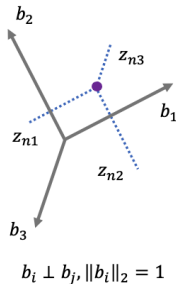
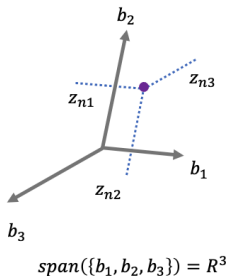
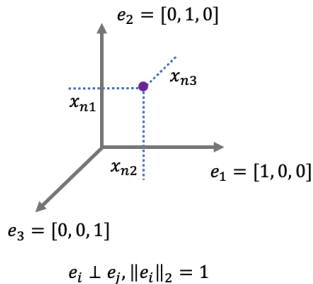
$$\mathbf{x}_n \approx \tilde{\mathbf{x}}_n := \sum_{j=1}^M z_{nj} \mathbf{b}_j, \quad z_{nj} := \mathbf{b}_j^\top \mathbf{x}_n$$

using an **orthonormal basis**

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

Quick refresher: basis

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



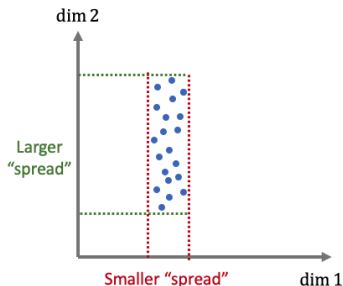
↑ orthonormal basis

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

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

PCA: maximum variance perspective

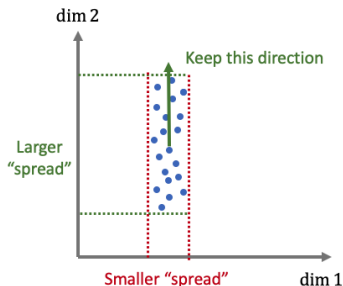
The “maximum variance” intuition of PCA:
project onto directions where the datapoints “vary the most”



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

PCA: maximum variance perspective

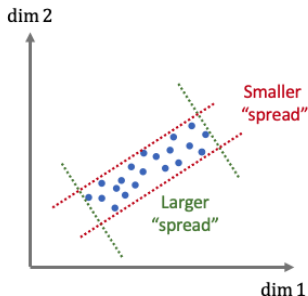
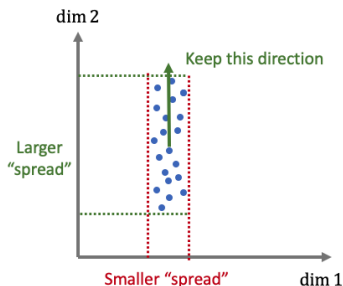
The “maximum variance” intuition of PCA:
project onto directions where the datapoints “vary the most”



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

PCA: maximum variance perspective

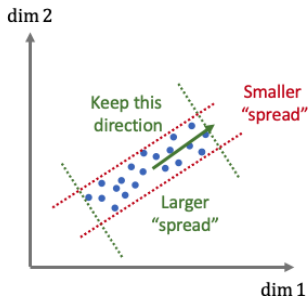
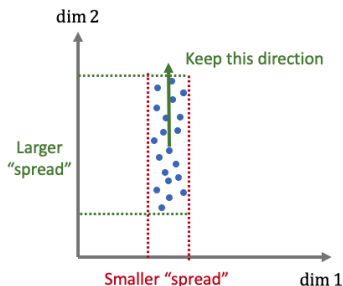
The “maximum variance” intuition of PCA:
project onto directions where the datapoints “vary the most”



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

$$\mathbf{z}_n := \mathbf{B}^\top \mathbf{x}_n \quad \Leftrightarrow \quad \mathbf{z}_{nj} := \mathbf{b}_j^\top \mathbf{x}_n$$

using an **orthonormal basis**

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

- Solve for \mathbf{b}_1 such that

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

PCA: maximum variance perspective

Solve for \mathbf{b}_1 such that

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

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

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

$$= \mathbf{b}_1^\top \mathbf{Q} \mathbf{\Lambda} \underbrace{\mathbf{Q}^\top \mathbf{b}_1}_{:= \beta_1} = \sum_{d=1}^D \lambda_d \beta_{1d}^2$$

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

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

PCA: maximum variance perspective

Solve for \mathbf{b}_1 such that

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

- 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:** $\boldsymbol{\beta}_1 = \mathbf{e}_1 := (1, 0, \dots, 0)^\top$
 $\Rightarrow \mathbf{b}_1 = \mathbf{q}_1$ (the eigenvector with the largest eigenvalue)

PCA: maximum variance perspective

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

For $m = 2, \dots, M$:

- Compute the “remainder” of projection:

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

PCA: maximum variance perspective

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

For $m = 2, \dots, M$:

- ▶ Compute the “remainder” of projection:

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

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

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

PCA: maximum variance perspective

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

For $m = 2, \dots, M$:

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

$$\begin{aligned}\mathbb{V}[\mathbf{b}_m^\top \hat{\mathbf{x}}_n] &:= \frac{1}{N} \sum_{n=1}^N (\mathbf{b}_m^\top \mathbf{x}_n - \sum_{j=1}^{m-1} (\mathbf{b}_j^\top \mathbf{x}_n) \underbrace{\mathbf{b}_j^\top \mathbf{b}_m}_{=0})^2 = \mathbf{b}_m^\top \underbrace{\left(\frac{1}{N} \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^\top \right)}_{=\mathbf{S}=\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top} \mathbf{b}_m \\ &= \mathbf{b}_m^\top \mathbf{Q} \mathbf{\Lambda} \underbrace{\mathbf{Q}^\top \mathbf{b}_m}_{:=\beta_m} = \sum_{d=1}^D \lambda_d \beta_{md}^2\end{aligned}$$

- ▶ Here $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_D)$ and $\lambda_1 \geq \dots \geq \lambda_D \geq 0$.

PCA: maximum variance perspective

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

For $m = 2, \dots, M$:

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

▸ $\|\mathbf{b}_m\|_2^2 = 1 \Rightarrow \|\boldsymbol{\beta}_m\|_2^2 = 1$

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

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

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

PCA: maximum variance perspective

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

For $m = 2, \dots, M$:

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

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

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

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

PCA: maximum variance perspective

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

For $m = 2, \dots, M$:

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

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

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

1. $\boldsymbol{\beta}_1 = \mathbf{e}_1 \Rightarrow \mathbf{b}_1 = \mathbf{q}_1$
2. For $m = 2, \dots, M$, assume $\boldsymbol{\beta}_j = \mathbf{e}_j, j = 1, \dots, m-1$
 - 2a. $\boldsymbol{\beta}_m^\top \boldsymbol{\beta}_j = 0, j = 1, \dots, m-1 \Rightarrow \beta_{mj} = 0, j = 1, \dots, m-1$

PCA: maximum variance perspective

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

For $m = 2, \dots, M$:

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

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

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

1. $\boldsymbol{\beta}_1 = \mathbf{e}_1 \Rightarrow \mathbf{b}_1 = \mathbf{q}_1$
2. For $m = 2, \dots, M$, assume $\boldsymbol{\beta}_j = \mathbf{e}_j, j = 1, \dots, m-1$
 - 2a. $\boldsymbol{\beta}_m^\top \boldsymbol{\beta}_j = 0, j = 1, \dots, m-1 \Rightarrow \beta_{mj} = 0, j = 1, \dots, m-1$
 - 2b. $\|\boldsymbol{\beta}_m\|_2 = 1 \Rightarrow \sum_{d=m}^D \beta_{md}^2 = 1$

PCA: maximum variance perspective

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

For $m = 2, \dots, M$:

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

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

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

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

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

PCA: maximum variance perspective

For $m = 1, \dots, M$:

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

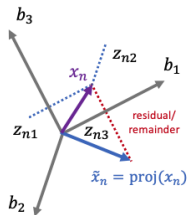
Solutions: $\mathbf{b}_m = \mathbf{q}_m$ for $m = 1, \dots, M$

\Rightarrow Projecting \mathbf{x}_n to a subspace

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

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

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



$$\mathbf{b}_i \perp \mathbf{b}_j, \|\mathbf{b}_i\|_2 = 1$$

$$\text{span}(\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}) = \mathbb{R}^3$$

PCA: minimum reconstruction error perspective

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

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

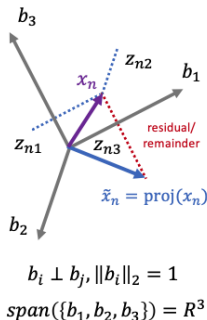
Rewriting the loss:

- Consider the full orthonormal basis:

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

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

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



PCA: minimum reconstruction error perspective

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

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

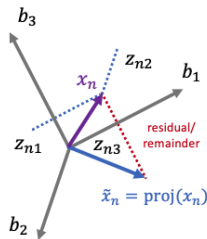
Rewriting the loss:

- Consider the full orthonormal basis:

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

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

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



$$b_i \perp b_j, \|b_i\|_2 = 1$$
$$\text{span}(\{b_1, b_2, b_3\}) = \mathbb{R}^3$$

PCA: minimum reconstruction error perspective

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

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

Rewriting the loss:

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

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

PCA: minimum reconstruction error perspective

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

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

Rewriting the loss:

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

$$\begin{aligned} L &= \frac{1}{N} \sum_{n=1}^N \sum_{j=M+1}^D (\mathbf{b}_j^\top \mathbf{x}_n)^2 = \frac{1}{N} \sum_{n=1}^N \sum_{j=M+1}^D \mathbf{b}_j^\top (\mathbf{x}_n \mathbf{x}_n^\top) \mathbf{b}_j \\ &= \sum_{j=M+1}^D \mathbf{b}_j^\top \underbrace{\left(\frac{1}{N} \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^\top \right)}_{:= \mathbf{S} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^\top} \mathbf{b}_j = \sum_{j=M+1}^D \mathbf{b}_j^\top \mathbf{Q} \mathbf{\Lambda} \underbrace{\mathbf{Q}^\top \mathbf{b}_j}_{:= \beta_j} = \sum_{j=M+1}^D \sum_{d=1}^D \lambda_d \beta_{jd}^2 \end{aligned}$$

PCA: minimum reconstruction error perspective

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

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

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

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

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

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

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

PCA: minimum reconstruction error perspective

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

An iterative approach for solutions:

Solve β_D first and then solve for β_j for $j = D - 1, \dots, M + 1$.

- Optimisation objective for β_D :

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

- Notice: $\lambda_1 \geq \dots \geq \lambda_D$
- **Solution:** $\beta_D = e_D := (0, \dots, 0, 1)^\top$
 $\Rightarrow \mathbf{b}_D = \mathbf{q}_D$ (the eigenvector with the smallest eigenvalue)

PCA: minimum reconstruction error perspective

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

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

1. For $j = D$: $\beta_D = e_D$, i.e., $\mathbf{b}_D = \mathbf{q}_D$
2. For $j = D-1, \dots, M+1$, assume for $i > j$, $\beta_i = e_i$, i.e., $\mathbf{b}_i = \mathbf{q}_i$
 - 2a. $\beta_i^\top \beta_j = 0, i > j \Rightarrow \beta_j = (\beta_{j1}, \dots, \beta_{jj}, 0, \dots, 0)^\top$

PCA: minimum reconstruction error perspective

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

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

1. For $j = D$: $\beta_D = e_D$, i.e., $\mathbf{b}_D = \mathbf{q}_D$
2. For $j = D-1, \dots, M+1$, assume for $i > j$, $\beta_i = e_i$, i.e., $\mathbf{b}_i = \mathbf{q}_i$
 - 2a. $\beta_i^\top \beta_j = 0, i > j \Rightarrow \beta_j = (\beta_{j1}, \dots, \beta_{jj}, 0, \dots, 0)^\top$
 - 2b. $\|\beta_j\|_2^2 = 1 \Rightarrow \sum_{d=1}^j \beta_{jd}^2 = 1$

PCA: minimum reconstruction error perspective

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

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

1. For $j = D$: $\beta_D = e_D$, i.e., $\mathbf{b}_D = \mathbf{q}_D$
2. For $j = D-1, \dots, M+1$, assume for $i > j$, $\beta_i = e_i$, i.e., $\mathbf{b}_i = \mathbf{q}_i$
 - 2a. $\beta_i^\top \beta_j = 0, i > j \Rightarrow \beta_j = (\beta_{j1}, \dots, \beta_{jj}, 0, \dots, 0)^\top$
 - 2b. $\|\beta_j\|_2^2 = 1 \Rightarrow \sum_{d=1}^j \beta_{jd}^2 = 1$
 - 2c. Solve for the following minimisation problem w.r.t. β_{jd} :

PCA: minimum reconstruction error perspective

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

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

1. For $j = D$: $\beta_D = e_D$, i.e., $\mathbf{b}_D = \mathbf{q}_D$
2. For $j = D-1, \dots, M+1$, assume for $i > j$, $\beta_i = e_i$, i.e., $\mathbf{b}_i = \mathbf{q}_i$
 - 2a. $\beta_i^\top \beta_j = 0, i > j \Rightarrow \beta_j = (\beta_{j1}, \dots, \beta_{jj}, 0, \dots, 0)^\top$
 - 2b. $\|\beta_j\|_2^2 = 1 \Rightarrow \sum_{d=1}^j \beta_{jd}^2 = 1$
 - 2c. Solve for the following minimisation problem w.r.t. β_{jd} :

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

PCA: minimum reconstruction error perspective

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

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

1. For $j = D$: $\beta_D = e_D$, i.e., $\mathbf{b}_D = \mathbf{q}_D$
2. For $j = D-1, \dots, M+1$, assume for $i > j$, $\beta_i = e_i$, i.e., $\mathbf{b}_i = \mathbf{q}_i$
 - 2a. $\beta_i^\top \beta_j = 0, i > j \Rightarrow \beta_j = (\beta_{j1}, \dots, \beta_{jj}, 0, \dots, 0)^\top$
 - 2b. $\|\beta_j\|_2^2 = 1 \Rightarrow \sum_{d=1}^j \beta_{jd}^2 = 1$
 - 2c. Solve for the following minimisation problem w.r.t. β_{jd} :

Solution: $\beta_j = e_j$, i.e., $\mathbf{b}_j = \mathbf{q}_j$

PCA: minimum reconstruction error perspective

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

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

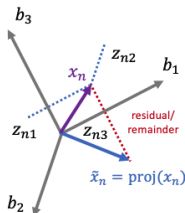
Solutions: $\mathbf{b}_j = \mathbf{q}_j$ for $j = M+1, \dots, D$

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

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

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

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



$$\mathbf{b}_i \perp \mathbf{b}_j, \|\mathbf{b}_i\|_2 = 1$$

$$\text{span}(\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}) = \mathbb{R}^3$$

PCA: comparing both views

- ▶ Maximum variance view:

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

- ▶ Minimum reconstruction error view:

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

- ▶ 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)