Principal Component Analysis

Yingzhen Li

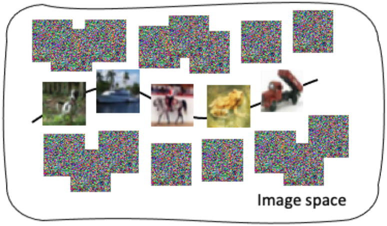
Department of Computing
Imperial College London

@liyzhen2
yingzhen.li@imperial.ac.uk

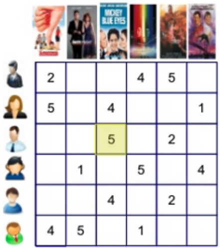
Nov 2, 2021

Dimensionality reduction

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



natural images vs all RGB images

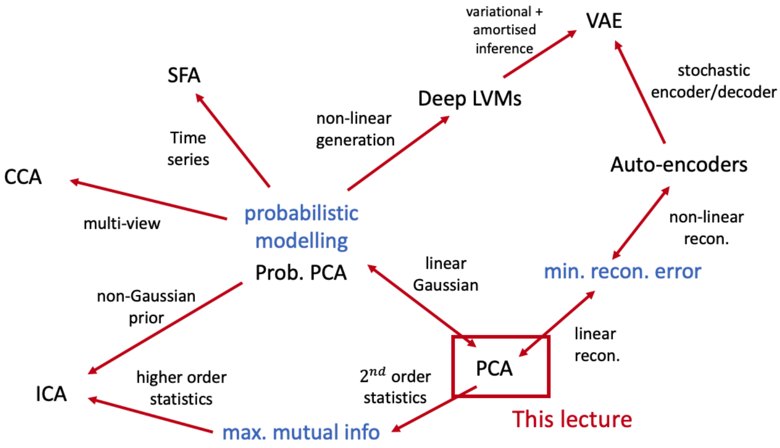


2			4	5	
5		4			1
		5		2	
	1		5		4
		4		2	
4	5		1		

User ratings on items

Dimensionality reduction

To name a few dimensionality reduction methods:



Dimensionality reduction

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:

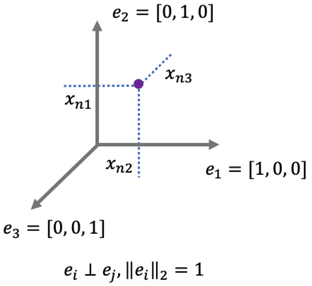
$$\boldsymbol{x}_n \approx \tilde{\boldsymbol{x}}_n := \sum_{j=1}^{M} \boldsymbol{z}_{nj}\mathbf{b}_j, \quad \boldsymbol{z}_{nj} := \mathbf{b}_j^\top \boldsymbol{x}_n$$

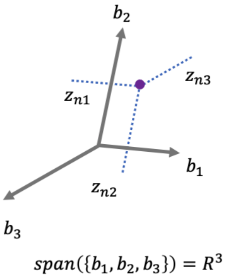
using an **orthonormal basis**

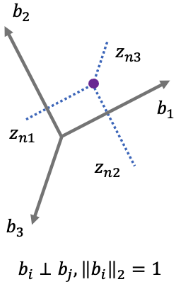
$$\mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_M], \quad \mathbf{b}_m \in \mathbb{R}^{D\times 1}, \quad M < D$$

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: minimum reconstruction error perspective

Goal: minimising ℓ_2 reconstruction error:

$$L = \frac{1}{N}\sum_{n=1}^{N} ||\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n||_2^2$$

Rewriting the loss:

- Consider the full orthonormal basis:

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

PCA: minimum reconstruction error perspective

Goal: minimising ℓ_2 reconstruction error:

$$L = \frac{1}{N} \sum_{n=1}^{N} ||\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n||_2^2$$

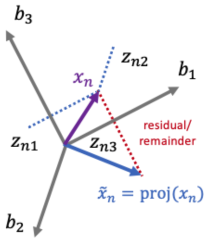
Rewriting the loss:

- 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 $\mathbf{B}_{full}$:

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



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

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

PCA: minimum reconstruction error perspective

Goal: minimising ℓ_2 reconstruction error:

$$L = \frac{1}{N}\sum_{n=1}^{N} ||\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n||_2^2$$

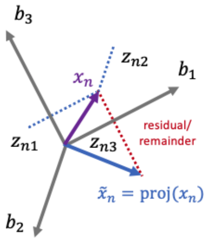
Rewriting the loss:

- 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: minimising ℓ_2 reconstruction error:

$$L = \frac{1}{N}\sum_{n=1}^{N} ||\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n||_2^2$$

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

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

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

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

PCA: minimum reconstruction error perspective

Goal: minimising ℓ_2 reconstruction error:

$$L = \frac{1}{N} \sum_{n=1}^{N} ||\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n||_2^2$$

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

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

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

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

PCA: minimum reconstruction error perspective

Define $\mathbf{S} := \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top$:
$\Rightarrow$ Goal: find orthonormal basis $\mathbf{B}_{full}$ to minimise

$$L = \sum_{j=M+1}^{D} \mathbf{b}_j^\top \mathbf{S} \mathbf{b}_j$$

PCA: minimum reconstruction error perspective

Define $\mathbf{S} := \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top$:
$\Rightarrow$ Goal: find orthonormal basis $\mathbf{B}_{full}$ to minimise

$$L = \sum_{j=M+1}^{D} \mathbf{b}_j^\top \mathbf{S} \mathbf{b}_j$$

Properties of **S**:

- $\mathbf{S} \in \mathbb{R}^{D \times D}$ is symmetric
- **S** is positive semi-definite: for any $\boldsymbol{x} \in \mathbb{R}^{D \times 1}$

$$\boldsymbol{x}^\top \mathbf{S} \boldsymbol{x} = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{x}^\top \boldsymbol{x}_n)^2 \geqslant 0$$

PCA: minimum reconstruction error perspective

Define $\mathbf{S} := \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top$:
$\Rightarrow$ Goal: find orthonormal basis $\mathbf{B}_{full}$ to minimise

$$L = \sum_{j=M+1}^{D} \mathbf{b}_j^\top \mathbf{S} \mathbf{b}_j$$

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

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

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

PCA: minimum reconstruction error perspective

Define $\mathbf{S} := \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top$:
$\Rightarrow$ Goal: find orthonormal basis $\mathbf{B}_{full}$ to minimise

$$L = \sum_{j=M+1}^{D} \mathbf{b}_j^\top \mathbf{S} \mathbf{b}_j$$

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

$$\min_{\mathbf{B}_{full}} L = \sum_{j=M+1}^{D} \sum_{d=1}^{D} \lambda_d \beta_{jd}^2, \quad \text{s.t. } ||\mathbf{b}_j||_2^2 = 1, \mathbf{b}_i \perp \mathbf{b}_j$$

PCA: minimum reconstruction error perspective

$$\min_{\mathbf{B}_{full}} L = \sum_{j=M+1}^{D} \sum_{d=1}^{D} \lambda_d \beta_{jd}^2, \quad \text{s.t. } ||\mathbf{b}_j||_2^2 = 1, \mathbf{b}_i \perp \mathbf{b}_j$$

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

PCA: minimum reconstruction error perspective

$$\min_{\mathbf{B}_{full}} L = \sum_{j=M+1}^{D} \sum_{d=1}^{D} \lambda_d \beta_{jd}^2, \quad \text{s.t. } ||\mathbf{b}_j||_2^2 = 1, \mathbf{b}_i \perp \mathbf{b}_j$$

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

- Optimisation objective for $\mathbf{b}_D$:

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

PCA: minimum reconstruction error perspective

$$\min_{\mathbf{B}_{full}} L = \sum_{j=M+1}^{D} \sum_{d=1}^{D} \lambda_d \beta_{jd}^2, \quad \text{s.t. } ||\mathbf{b}_j||_2^2 = 1, \mathbf{b}_i \perp \mathbf{b}_j$$

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

- Optimisation objective for $\mathbf{b}_D$:

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

- Notice: $\beta_{Dd} = \mathbf{b}_D^\top \boldsymbol{q}_D, \sum_{d=1}^{D} \beta_{Dd}^2 = 1,$
  $\lambda_1 \geqslant \cdots \geqslant \lambda_D$

PCA: minimum reconstruction error perspective

$$\min_{\mathbf{B}_{full}} L = \sum_{j=M+1}^{D} \sum_{d=1}^{D} \lambda_d \beta_{jd}^2, \quad \text{s.t. } ||\mathbf{b}_j||_2^2 = 1, \mathbf{b}_i \perp \mathbf{b}_j$$

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

- Optimisation objective for $\mathbf{b}_D$:

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

- Notice: $\beta_{Dd} = \mathbf{b}_D^\top \boldsymbol{q}_D, \sum_{d=1}^{D} \beta_{Dd}^2 = 1,$
  $\lambda_1 \geqslant \cdots \geqslant \lambda_D$
- **Solution**: $\mathbf{b}_D = \boldsymbol{q}_D$ (the eigenvector with the smallest eigenvalue)

PCA: minimum reconstruction error perspective

$$\min_{\mathbf{B}_{full}} L = \sum_{j=M+1}^{D} \sum_{d=1}^{D} \lambda_d \beta_{jd}^2, \quad \text{s.t. } ||\mathbf{b}_j||_2^2 = 1, \mathbf{b}_i \perp \mathbf{b}_j$$

Iteratively solve for the rest of $\mathbf{b}_j$ for $j = D-1, ..., M+1$:

1. $\mathbf{b}_D = \boldsymbol{q}_D$

PCA: minimum reconstruction error perspective

$$\min_{\mathbf{B}_{full}} L = \sum_{j=M+1}^{D} \sum_{d=1}^{D} \lambda_d \beta_{jd}^2, \quad \text{s.t. } ||\mathbf{b}_j||_2^2 = 1, \mathbf{b}_i \perp \mathbf{b}_j$$

Iteratively solve for the rest of $\mathbf{b}_j$ for $j = D - 1, ..., M + 1$:

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

PCA: minimum reconstruction error perspective

$$\min_{\mathbf{B}_{full}} L = \sum_{j=M+1}^{D} \sum_{d=1}^{D} \lambda_d \beta_{jd}^2, \quad \text{s.t. } ||\mathbf{b}_j||_2^2 = 1, \mathbf{b}_i \perp \mathbf{b}_j$$

Iteratively solve for the rest of $\mathbf{b}_j$ for $j = D-1, ..., M+1$:

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

   2a. $\mathbf{b}_j \perp \mathbf{b}_i, i > j \quad \Rightarrow \quad \mathbf{b}_j = \sum_{d=1}^{j} \beta_{jd} \boldsymbol{q}_d$

PCA: minimum reconstruction error perspective

$$\min_{\mathbf{B}_{full}} L = \sum_{j=M+1}^{D} \sum_{d=1}^{D} \lambda_d \beta_{jd}^2, \quad \text{s.t. } ||\mathbf{b}_j||_2^2 = 1, \mathbf{b}_i \perp \mathbf{b}_j$$

Iteratively solve for the rest of $\mathbf{b}_j$ for $j = D-1, ..., M+1$:

1. $\mathbf{b}_D = \boldsymbol{q}_D$
2. For $j = D-1, ..., M+1$, assume $\mathbf{b}_i = \boldsymbol{q}_i, i > j$
   2a. $\mathbf{b}_j \perp \mathbf{b}_i, i > j \quad \Rightarrow \quad \mathbf{b}_j = \sum_{d=1}^{j} \beta_{jd} \boldsymbol{q}_d$
   2b. $||\mathbf{b}_j||_2^2 = 1 \quad \Rightarrow \quad \sum_{d=1}^{j} \beta_{jd}^2 = 1$

PCA: minimum reconstruction error perspective

$$\min_{\mathbf{B}_{full}} L = \sum_{j=M+1}^{D} \sum_{d=1}^{D} \lambda_d \beta_{jd}^2, \quad \text{s.t. } ||\mathbf{b}_j||_2^2 = 1, \mathbf{b}_i \perp \mathbf{b}_j$$

Iteratively solve for the rest of $\mathbf{b}_j$ for $j = D-1, ..., M+1$:

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

   2a. $\mathbf{b}_j \perp \mathbf{b}_i, i > j \quad \Rightarrow \quad \mathbf{b}_j = \sum_{d=1}^{j} \beta_{jd} \boldsymbol{q}_d$

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

   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

$$\min_{\mathbf{B}_{full}} L = \sum_{j=M+1}^{D} \sum_{d=1}^{D} \lambda_d \beta_{jd}^2, \quad \text{s.t. } ||\mathbf{b}_j||_2^2 = 1, \mathbf{b}_i \perp \mathbf{b}_j$$

Iteratively solve for the rest of $\mathbf{b}_j$ for $j = D-1, ..., M+1$:

1. $\mathbf{b}_D = \boldsymbol{q}_D$
2. For $j = D-1, ..., M+1$, assume $\mathbf{b}_i = \boldsymbol{q}_i, i > j$
   - 2a. $\mathbf{b}_j \perp \mathbf{b}_i, i > j \quad \Rightarrow \quad \mathbf{b}_j = \sum_{d=1}^{j} \beta_{jd} \boldsymbol{q}_d$
   - 2b. $||\mathbf{b}_j||_2^2 = 1 \quad \Rightarrow \quad \sum_{d=1}^{j} \beta_{jd}^2 = 1$
   - 2c. Solve for the following minimisation problem w.r.t. β_{jd}:
     Solution: $\mathbf{b}_j = \boldsymbol{q}_j$ **(i.e.** $\beta_{jj} = 1, \beta_{jd} = 0, d \neq j$**)**

PCA: minimum reconstruction error perspective

$$\min_{\mathbf{B}_{full}} L = \sum_{j=M+1}^{D} \sum_{d=1}^{D} \lambda_d \beta_{jd}^2, \quad \text{s.t. } ||\mathbf{b}_j||_2^2 = 1, \mathbf{b}_i \perp \mathbf{b}_j$$

Iteratively solve for the rest of $\mathbf{b}_j$ for $j = D-1, ..., M+1$:

1. $\mathbf{b}_D = \boldsymbol{q}_D$
2. For $j = D-1, ..., M+1$, assume $\mathbf{b}_i = \boldsymbol{q}_i, i > j$
   - 2a. $\mathbf{b}_j \perp \mathbf{b}_i, i > j \quad \Rightarrow \quad \mathbf{b}_j = \sum_{d=1}^{j} \beta_{jd} \boldsymbol{q}_d$
   - 2b. $||\mathbf{b}_j||_2^2 = 1 \quad \Rightarrow \quad \sum_{d=1}^{j} \beta_{jd}^2 = 1$
   - 2c. Solve for the following minimisation problem w.r.t. β_{jd}:
     Solution: $\mathbf{b}_j = \boldsymbol{q}_j$ (i.e. $\beta_{jj} = 1, \beta_{jd} = 0, d \neq j$)
3. **"Proof by induction": $\mathbf{b}_j = \boldsymbol{q}_j$ for $j = M+1, ..., D$**

PCA: minimum reconstruction error perspective

$$\min_{\mathbf{B}_{full}} L = \sum_{j=M+1}^{D} \sum_{d=1}^{D} \lambda_d \beta_{jd}^2, \quad \text{s.t. } ||\mathbf{b}_j||_2^2 = 1, \mathbf{b}_i \perp \mathbf{b}_j$$

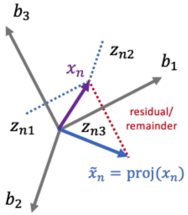
Solutions: $\mathbf{b}_j = \boldsymbol{q}_j$ for $j = M+1, ..., D$

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

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

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

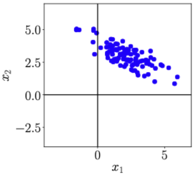
$$\tilde{\boldsymbol{x}}_n \in span(\{\boldsymbol{q}_j\}_{j=M+1}^{D})^{\perp}$$



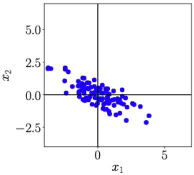
$b_i \perp b_j, \|b_i\|_2 = 1$

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

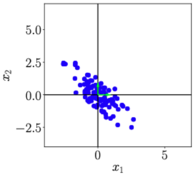
PCA in practise



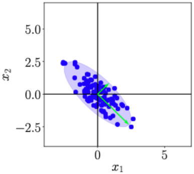
(a) Original dataset.



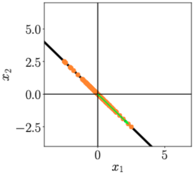
(b) Step 1: Centering by subtracting the mean from each data point.



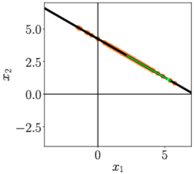
(c) Step 2: Dividing by the standard deviation to make the data unit free. Data has variance 1 along each axis.



(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).