CS376 Machine Learning

Dimensionality Reduction & PCA

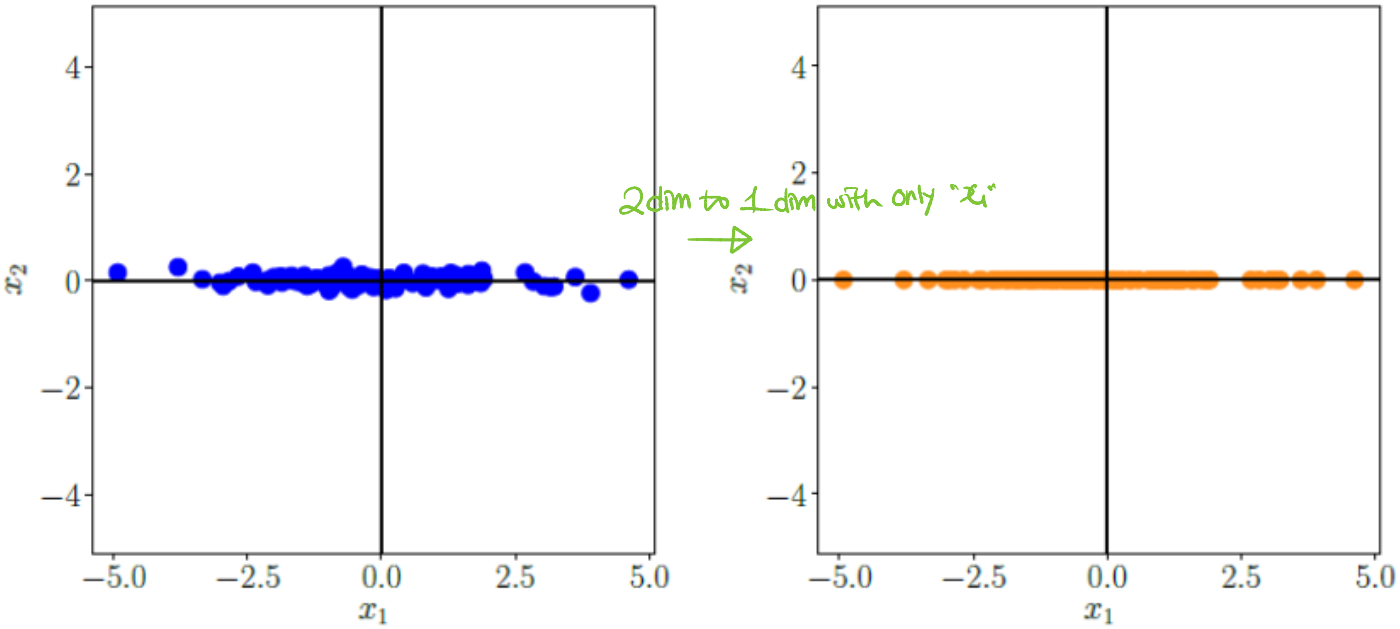
김기응

Kee-Eung Kim
KAIST
kekim@kaist.ac.kr

KAIST
Korea Advanced Institute of Science and Technology
한국과학기술원

Motivation

Figure 10.1 Illustration: Dimensionality reduction. (a) The original dataset does not vary much along the x_2 direction. (b) The data from (a) can be represented using the x_1-coordinate alone with nearly no loss.



(a) Dataset with x_1 and x_2 coordinates.

(b) Compressed dataset where only the x_1 coordinate is relevant.

KAIST
Korea Advanced Institute of Science and Technology
한국과학기술원

Problem Setting

- Find projections $\tilde{x}_n$ of data points x_n that are as similar as possible, but with significantly lower dimensionality (intrinsic)
- $X = \{x_1 \dots x_N\}$, $x_n \in \mathbb{R}^D$ with mean 0. (important)
- Data covariance matrix

$$S = \frac{1}{N} \sum_{n=1}^{N} x_n x_n^T$$

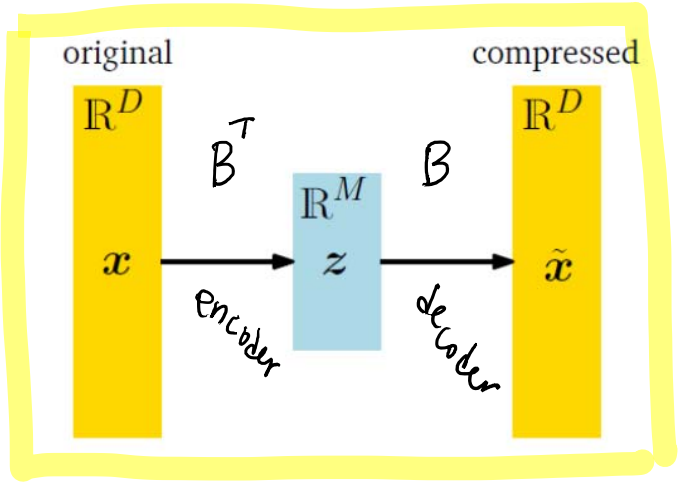
the columns $\boldsymbol{b}_1, \dots, \boldsymbol{b}_M$ of $\boldsymbol{B}$. Retaining most information after data compression is equivalent to capturing the largest amount of variance in the low-dimensional code (Hotelling, 1933).

- Find $z_n = B^T x_n \in \mathbb{R}^M$ $\quad B \triangleq [b_1 \dots b_M] \in \mathbb{R}^{D \times M}$

  $\{b_1 \dots b_M\}$ are orthonormal.

- $B^{-1} = B^T$, so $\tilde{x}_n = B z_n \in \mathbb{R}^D$

Figure 10.2 Graphical illustration of PCA. In PCA, we find a compressed version $\tilde{\boldsymbol{x}}$ of original data $\boldsymbol{x}$ that has an intrinsic lower-dimensional representation $\boldsymbol{z}$.



KAIST
Korea Advanced Institute of Science and Technology
한국과학기술원

$x \xrightarrow{B^T} z \xrightarrow{B} \tilde{x}$

$x \in \mathbb{R}^D$ $z \in \mathbb{R}^M$ $\tilde{x} \in \mathbb{R}^D$

10.2 Maximum Variance Perspective.

- Retain most informations after data compression 을 위해서는 capture the largest amount of variance in the low-dimensional code.

10.2.1 Direction with Maximal Variance

* single vector $b_1 \in \mathbb{R}^D$ that maximize the variance of projected data 구하기.

* maximize the variance of first coordinate z_1 of $z \in \mathbb{R}^M$

$\Rightarrow V_1 := \mathbb{V}[z_1] = \frac{1}{N}\sum_{n=1}^{N} z_{1n}^2$ $(z_n \in \mathbb{R}^M,\ x_n \in \mathbb{R}^D)$

z_1의 variance 구하는 식.

$z_{1n} = b_1^T x_n \rightarrow Z = B^T X$

$\therefore V_1 = \frac{1}{N}\sum_{n=1}^{N}(b_1^T x_n)^2 = \frac{1}{N}\sum_{n=1}^{N} b_1^T x_n x_n^T b_1 = b_1^T\left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n^T\right) b_1 = b_1^T S b_1$ ∵ 이때 $\|b_1\|^2 = 1$

(data covariance matrix)

$\Rightarrow$ Optimization problem

$\max_{b_1} b_1^T S b_1 \quad s.t.\ \|b_1\|^2 = 1$

$\mathcal{L}(b_1, \lambda_1) = b_1^T S b_1 + \lambda_1(1 - b_1^T b_1)$

$\frac{d\mathcal{L}}{db_1} = 2b_1^T S - 2\lambda_1 b_1^T \qquad \frac{d\mathcal{L}}{d\lambda_1} = 1 - b_1^T b_1$

$\therefore Sb_1 = \lambda_1 b_1$ & $b_1^T b_1 = 1 \Leftrightarrow \lambda_1$: eigenvalue, b_1: eigenvector

$\Rightarrow V_1 = b_1^T S b_1 = \lambda_1 b_1^T b_1 = \lambda_1 \Leftrightarrow$ Variance 값 = eigenvalue 값

$\therefore$ variance 값이 클려면 eigenvalue 값 커야해.

이런식으로 $z \in \mathbb{R}^M$ 이내의 M-1 개의 큰 eigenvalue 값 뽑고, 해당 eigenvector로 $\underset{\sim}{B}$ 구성하기

$x \xrightarrow{B^T} z \xrightarrow{B} \tilde{x}$

$x \in \mathbb{R}^D$ $z \in \mathbb{R}^M$ $\tilde{x} \in \mathbb{R}^D$

10.2.2 M-dimensional subspace with Maximal Variance

※ Symmetric한 matrix ⟷ eigenvalue는 REAL

eigenvector는 orthonormal.

* M-1번째까지는 큰 eigenvalue를 구하고, M번째부터는 남는 애들 아무거나 값 쓰고 넘어

→ subtract the effect of the first m-1 principal components from the data, thereby trying to find principal components that compress the remaining information.

$\hat{X} := X - \sum_{i=1}^{m-1} b_i b_i^T X = X - B_{m-1} X$

⇒ $V_m = V[z_m] = \frac{1}{N}\sum_{n=1}^{N} z_{mn}^2 = \frac{1}{N}\sum_{n=1}^{N} (b_m^T \hat{x}_n)^2 = b_m^T \hat{S} b_m$.

⇒ 똑같이 optimize 하면. "b_m은 eigenvector of $\hat{S}$ with largest eigenvalue for $\hat{S}$"

그런데 사실 $S, \hat{S}$은 같은 sets of eigenvectors 가지고 있음.

↳ D개의 distinct한 eigenvector를 지님.

증명 만약 S의 m-1개의 eigenvector $b_1, b_2 \cdots b_{m-1}$ 찾고. b_i는 S의 eigenvector라면 $Sb_i = \lambda b_i$

$\hat{S} b_i = \frac{1}{N}\hat{X}\hat{X}^T b_i = \frac{1}{N}(X - B_{m-1}X)(X - B_{m-1}X)^T b_i = (S - SB_{m-1} - B_{m-1}S + B_{m-1}SB_{m-1}) b_i$

① $i \geq m$ $B_{m-1} b_i = 0$

$\hat{S} b_i = S b_i = \lambda b_i$

② $i < m$. $B_{m-1} b_i = b_i$

$\hat{S} b_i = 0 b_i = 0$ ⇒ eigenvalue = 0.

$b_1, b_2 \cdots b_{m-1}$: span NULL SPACE of $\hat{S}$

∴ every eigenvector of $S = \hat{S}$

S의 eigenvector이 (m-1)dim principal subspace라면 eigenvalue of $\hat{S} = 0$.

* M dimensional subspace에 project 되면, variance of data equals sum of the m개의 eigenvalues with high values : $V_M = \sum_{m=1}^{M} \lambda_m$

Maximum Variance Perspective

- First, we aim to maximize variance of the first coord z_1

$$V_1 \triangleq V[z_1] = \frac{1}{N}\sum_{n=1}^{N} z_{1n}^2$$

where $z_{1n} = b_1^T x_n$ (From $z = B^T x$)

$x \xrightarrow{B^T} z \xrightarrow{B} \tilde{x}$

- $V_1 = \frac{1}{N}\sum_{n=1}^{N} (b_1^T x_n)^2 = \frac{1}{N}\sum_{n=1}^{N} b_1^T x_n x_n^T b_1 = b_1^T \left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n^T\right) b_1 = b_1^T S b_1$

S = "data covariance matrix"

- Thus, we need solve constrained optimization problem

$$\max_{b_1} b_1^T S b_1 \quad \text{subject to } \|b_1\|^2 = 1$$

$$\Leftrightarrow \mathcal{L}(b_1, \lambda) = b_1^T S b_1 + \lambda_1 (1 - b_1^T b_1)$$

$$\frac{\partial \mathcal{L}}{\partial b_1} = 2 b_1^T S - 2\lambda_1 b_1^T = 0 \qquad \frac{\partial \mathcal{L}}{\partial \lambda} = 1 - b_1^T b_1 = 0$$

$$\Leftrightarrow \begin{cases} S b_1 = \lambda_1 b_1 \\ b_1^T b_1 = 1 \end{cases} \Rightarrow$$

b_1 is an eigenvector of S.

λ_1 is the corresponding eigval.

$V_1 = b_1^T S b_1 = \lambda_1 b_1^T b_1 = \lambda_1 \Rightarrow$ Choose eigenvector with largest eigenvalue

"First principal component"

$\Rightarrow$ largest variance dir.

KAIST
Korea Advanced Institute of Science and Technology
한국과학기술원

Maximum Variance Perspective

- Assume we have found $m-1$ principal component. 지금까지는 이렇게 찾은 방식으로 eigenvalue 큰 거 골라.

  That is, we have found $b_1 \cdots b_{m-1}$, and find remaining PCs that compress remaining information → 마지막 m개는 남아있는 정보 가장 잘 압축할 수 있는 애로 고르기

  $$\hat{X} \triangleq X - \sum_{i=1}^{m-1} b_i b_i^T X \quad \text{where } X = [x_1 \cdots x_N] \in \mathbb{R}^{D \times N}$$

  원래 − 집어넣은 애 (!)

- $$\max V_m = V[z_m] = \frac{1}{N}\sum_{n=1}^{N} z_{mn}^2 = \frac{1}{N}\sum_{n=1}^{N} (b_m^T \hat{x}_n)^2 = b_m^T \hat{S} b_m$$

  Subject to $\|b_m\| = 1$ and $b_i^T b_m = 0$, $i = 1 \cdots m-1$

  $\Rightarrow$ b_m is the eigenvec with largest eigenval of $\hat{S}$

- This is also eigenvector of S

$$\hat{S} = \frac{1}{N}\sum_{n=1}^{N} \hat{x}_n \hat{x}_n^T = \frac{1}{N}\sum_{n=1}^{N} \left(x_n - \sum_{i=1}^{m-1} b_i b_i^T x_n\right)\left(x_n - \sum_{i=1}^{m-1} b_i b_i^T x_n\right)^T$$

$$= \frac{1}{N}\sum_{n=1}^{N}\Bigg[x_n x_n^T + \left(\sum_{i=1}^{m-1} b_i b_i^T\right) x_n x_n^T \left(\sum_{i=1}^{m-1} b_i b_i^T\right) - x_n x_n^T \left(\sum_{i=1}^{m-1} b_i b_i^T\right) - \left(\sum_{i=1}^{m-1} b_i b_i^T\right) x_n x_n^T \Bigg]$$

KAIST
Korea Advanced Institute of Science and Technology
한국과학기술원

Maximum Variance Perspective

- This is also eigenvector of S

$$\hat{S} = \frac{1}{N}\sum_{n=1}^{N} \hat{x}_n \hat{x}_n^T = \frac{1}{N}\sum_{n=1}^{N} \left(x_n - \sum_{i=1}^{m-1} b_i b_i^T x_n\right)\left(x_n - \sum_{i=1}^{m-1} b_i b_i^T x_n\right)^T$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left[x_n x_n^T + \left(\sum_{i=1}^{m-1} b_i b_i^T\right) x_n x_n^T \left(\sum_{i=1}^{m-1} b_i b_i^T\right) - x_n x_n^T \left(\sum_{i=1}^{m-1} b_i b_i^T\right) - \left(\sum_{i=1}^{m-1} b_i b_i^T\right) x_n x_n^T\right]$$

$$\hat{S} b_m = \frac{1}{N}\sum_{n=1}^{N}\left[x_n x_n^T b_m + 0\right] - 0 = S b_m$$

last 0: proof

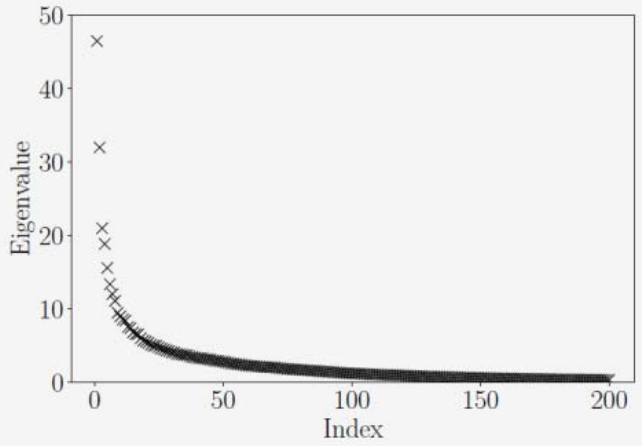
$$\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{m-1} b_i b_i^T x_n x_n^T b_m = \sum_{i=1}^{m-1} b_i b_i^T \underbrace{\left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n^T\right)}_{S} b_m$$

$$= \sum_{i=1}^{m-1} b_i \lambda_i b_i^T b_m = 0$$

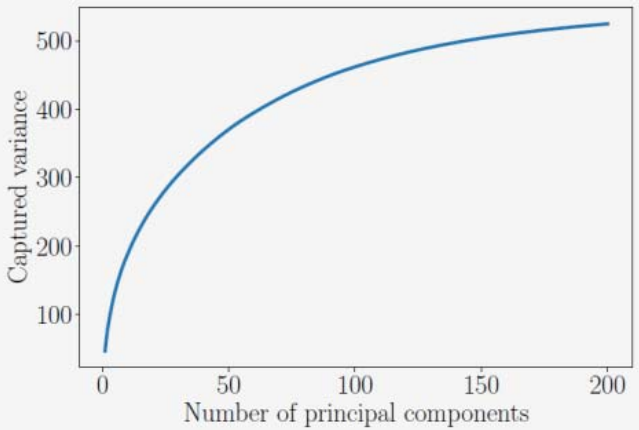
$$\therefore V_m = b_m^T \hat{S} b_m = b_m^T S b_m$$

KAIST
Korea Advanced Institute of Science and Technology
한국과학기술원

Maximum Variance Perspective



(a) Eigenvalues (sorted in descending order) of the data covariance matrix of all digits '8' in the MNIST training set.



(b) Variance captured by the principal components.

- The amount of variance captured by PCA using M PCs

$$V_M = \sum_{m=1}^{M} \lambda_m$$

The amount of variance lost by PCA

$$J_M = \sum_{j=M+1}^{D} \lambda_j = V_D - V_M$$

relative measure: $\frac{V_M}{V_D}$, $1 - \frac{V_M}{V_D}$

KAIST
Korea Advanced Institute of Science and Technology
한국과학기술원

Projection Perspective

- PCA as an algorithm that directly minimizes avg reconstruction error ($\|x_n - \tilde{x}_n\|$)
- Assume ordered orthonormal basis $B = (b_1 \cdots b_D)$, ie $b_i^T b_j = 1$ iff $i = j$ and 0 otherwise
- Any $x \in \mathbb{R}^D$, $x = \sum_{d=1}^{D} \zeta_d b_d = \sum_{m=1}^{M} \zeta_m b_m + \sum_{j=M+1}^{D} \zeta_j b_j$

$$\tilde{x} = \sum_{m=1}^{M} z_m b_m \in U \subseteq \mathbb{R}^D \text{ with } \dim(U) = M$$

RM z_m and ζ_m are not identical? identical? $\in \mathbb{R}$!

Projection Perspective

a $\min \quad J_M \triangleq \frac{1}{N} \| x_n - \tilde{x}_n \|^2$ ($\tilde{x}_n = B^T z$)

$z_1 \cdots z_n$

$b_1 \cdots b_m$

Finding $z_1 \cdots z_n$

e $\frac{\partial J_M}{\partial z_{in}} = \frac{\partial J_M}{\partial \tilde{x}_n} \frac{\partial \tilde{x}_n}{\partial z_{in}}$,

$$\begin{cases} \frac{\partial J_M}{\partial \tilde{x}_n} = -\frac{2}{N}(x_n - \tilde{x}_n)^T \in \mathbb{R}^{1 \times D} \\ \frac{\partial \tilde{x}_n}{\partial z_{in}} = \frac{\partial}{\partial z_{in}} \left(\sum_{m=1}^{M} z_{mn} b_m \right) = b_i \end{cases}$$

$$= -\frac{2}{N}(x_n - \tilde{x}_n)^T b_i = -\frac{2}{N}\left(x_n - \sum_{m=1}^{M} z_{mn} b_m\right)^T b_i$$

나머지 같은 orthonormal이니 없애도

$$= -\frac{2}{N}(x_n^T b_i - z_{in} b_i^T b_i) = -\frac{2}{N}(x_n^T b_i - z_{in}) = 0$$

$\therefore z_{in} = x_n^T b_i = b_i^T x_n$ (z_{in} is the coordinate of ortho. projection of x_n onto b_i)

Note ① ξ_m and z_m are identical for $m = 1 \cdots M$

KAIST
Korea Advanced Institute of Science and Technology
한국과학기술원

Low-Rank Approximation Perspective

finding $b_1 \cdots b_M$

② $z_n = B^T x_n$, $\tilde{x}_n = B z_n$ (D, D×M, M)

$B = [b_1 \cdots b_M] \in \mathbb{R}^{D \times M}$

D → M → D

$$\tilde{x}_n = \sum_{m=1}^{M} z_{mn} b_m = \sum_{m=1}^{M} (x_n^T b_m) b_m = \left(\sum_{m=1}^{M} b_m b_m^T \right) x_n$$

$$x_n = \sum_{d=1}^{D} z_{dn} b_d = \sum_{d=1}^{D} (x_n^T b_d) b_d = \left(\sum_{d=1}^{D} b_d b_d^T \right) x_n$$

$$= \left(\sum_{m=1}^{M} b_m b_m^T \right) x_n + \left(\sum_{j=M+1}^{D} b_j b_j^T \right) x_n$$

$$x_n - \tilde{x}_n = \left(\sum_{j=M+1}^{D} b_j b_j^T \right) x_n = \sum_{j=M+1}^{D} (x_n^T b_j) b_j$$

Low rank Approx. View

$$\tilde{x}_n = \left(\sum_{m=1}^{M} b_m b_m^T \right) x_n = BB^T x_n$$ (BB^T is of rank $\leq M$)

$$J_M = \frac{1}{N} \sum_{n=1}^{N} \| x_n - \tilde{x}_n \|^2 = \frac{1}{N} \sum_{n=1}^{N} \| x_n - BB^T x_n \|^2 = \frac{1}{N} \sum_{n=1}^{N} \| (I - BB^T) x_n \|^2$$

(Try to find the best rank-M approximation of BB^T to I)

KAIST
Korea Advanced Institute of Science and Technology
한국과학기술원

Low-Rank Approximation Perspective

$$J_M = \frac{1}{N}\sum_{n=1}^{N} \|x_n - \tilde{x}_n\|^2 = \frac{1}{N}\sum_{n=1}^{N} \left\| \sum_{j=M+1}^{D} (b_j^T x_n) b_j \right\|^2 = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=M+1}^{D} (b_j^T x_n)^2$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{j=M+1}^{D} b_j^T x_n x_n^T b_j = \sum_{j=M+1}^{D} b_j^T \left(\frac{1}{N}\sum_{n=1}^{N} x_n x_n^T\right) b_j = \sum_{j=M+1}^{D} b_j^T S b_j$$

$$= \sum_{j=M+1}^{D} \mathrm{tr}(b_j^T S b_j) = \sum_{j=M+1}^{D} \mathrm{tr}(b_j b_j^T S) = \mathrm{tr}\left(\left(\sum_{j=M+1}^{D} b_j b_j^T\right) S\right)$$

$$= \sum_{j=M+1}^{D} \lambda_j$$

(minimize the variance of data when projected on to the subspace we ignore)

= (select $(D-M)$ eigenvectors with smallest eigvalues)

= (select D eigenvectors with largest eigvalues)

KAIST
Korea Advanced Institute of Science and Technology
한국과학기술원

PCA Illustration

Example 10.3 (MNIST Digits Embedding)

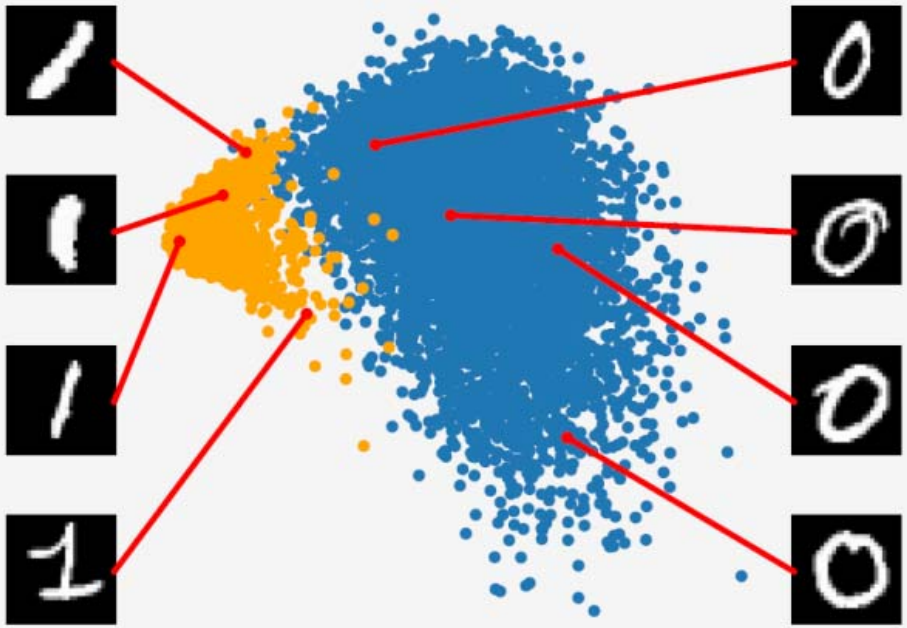


Figure 10.10 Embedding of MNIST digits 0 (blue) and 1 (orange) in a two-dimensional principal subspace using PCA. Four embeddings of the digits '0' and '1' in the principal subspace are highlighted in red with their corresponding original digit.