Kapitel 4 Dimensionsreduktion

§ 1 Hauptkomponentenanalyse (PCA)

PCA $\hat{=}$ principal component analysis

Gegeben: Datenpunkte $X = \begin{pmatrix} | & | & & | \\ x^1 & x^2 & \cdots & x^N \\ | & | & & | \end{pmatrix} \in \mathbb{R}^{D \times N}$.

Oft: D sehr groß.

Hoffnung: alle x^n liegen nahe zu einer niedrig-dimensionalen Mannigfaltigkeit, d.h. die Daten x^n haben Struktur.

einfaches Modell: alle x^n liegen nahe zu einem affinen Unterraum der Dimension M und $M \ll D$.

D.h.

$$x^n \approx c + \sum_{i=1}^{M} y_i^n b^i = \tilde{x}_n, \quad \text{wobei } b_1, \dots, b_M \in \mathbb{R}^D \text{ linear unabhängig.}$$

Ziel: Finde $c \in \mathbb{R}^D$, $Y = (y_i^n)_{i,n} \in \mathbb{R}^{M \times N}$, $B = \begin{bmatrix} | & & | \\ b^1 & \cdots & b^M \\ | & & | \end{bmatrix} \in \mathbb{R}^{D \times M}$,

so dass

$$E[B, Y, c] = \sum_{n=1}^{N} \| x_n - \tilde{x}_n \|^2.$$

minimal ist

Vereinfachung: Datenpunkte sind zentriert, d.h. es gilt
$$\frac{1}{N}\sum_{n=1}^{N} x^n = 0.$$

In diesem Fall kann man zeigen, dass $c = 0$ sein muss
(→ Blatt 9).

Satz Ang. die Datenpunkte $(x^n)_n$ sind zentriert.
Seien $\lambda_1, \dots, \lambda_D$ die Eigenwerte der Matrix $XX^T \in S^D_{\geq 0}$,
so sortiert, dass $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_D$ gilt. Sei
$b_1, \dots, b_D$ eine Orthonormalbasis von $\mathbb{R}^D$ bestehend aus
zugehörigen Eigenvektoren. Dann gilt
$$\min_{B,Y} E[B,Y] = \sum_{i=M+1}^{D} \lambda_i$$
und ein optimales B ist $B = [b_1 \cdots b_M]$ und $Y = B^T X$.

(Die Vektoren $b_1, \dots, b_M$ heißen Hauptkomponenten des Datensatz X.)

Bew.: Es gilt
$$E[B,Y] = \sum_{n=1}^{N} \| x^n - \tilde{x}^n \|^2 = \sum_{n=1}^{N} \sum_{i=1}^{D} (x_i^n - \tilde{x}_i^n)^2 = \dots$$

$$= \sum_{n=1}^{N} \sum_{i=1}^{D} \left(x_i^n - \sum_{j=1}^{M} y_j^n b_i^j \right)^2$$

$$= \mathrm{Tr}\left((X - BY)^T (X - BY) \right) \qquad (*)$$

Wir können oBdA annehmen, dass B aus orthonormalen Vektoren besteht, d.h. dass $B^T B = I_M$ gilt.

Für fixes B haben wir folgendes notwendiges Kriterium für die Optimalität von Y:

$$0 = \frac{\partial}{\partial y_k^n} E[B, Y] = \frac{\partial}{\partial y_k^n} \sum_{i=1}^{D} \left(x_i^n - \sum_{j=1}^{M} y_j^n b_i^j \right)^2$$

$$= \sum_{i=1}^{D} \frac{\partial}{\partial y_k^n} \left((x_i^n)^2 - 2 x_i^n \sum_{j=1}^{M} y_j^n b_i^j + \sum_{j,j'=1}^{M} y_j^n y_{j'}^n b_i^j b_i^{j'} \right)$$

$$= \sum_{i=1}^{D} \left(-2 x_i^n b_i^k \right) + 2 \sum_{j=1}^{M} y_j^n b_i^j b_i^k)$$

$$= -2 \sum_{i=1}^{D} x_i^n b_i^k + 2 \sum_{j=1}^{M} y_j^n \underbrace{\sum_{i=1}^{D} b_i^j b_i^k}_{\delta_{jk}}$$

$$= -2 \sum_{i=1}^{D} x_i^n b_i^k + 2 y_k^n$$

$$\Rightarrow \boxed{Y = B^T X} \qquad (**)$$

Setze nun (**) in (*) ein:

$$E[B] = Tr\left((X - BB^T X)^T (X - BB^T X) \right)$$

$$= Tr\left((X^T - X^T B B^T)(X - BB^T X) \right)$$

$$= Tr\left(X^T X - X^T BB^T X - X^T BB^T X + X^T B \underbrace{B^T B}_{I_n} B^T X \right)$$

$$= Tr\left(X^T X - X^T BB^T X \right)$$

$$= Tr\left(X^T (I - BB^T) X \right)$$

$$= Tr\left((I - BB^T) X X^T \right)$$

$$= Tr\left(X X^T \right) - Tr\left(BB^T X X^T \right) \rightarrow \min.$$

D.h. $Tr(BB^T X X^T) \rightarrow \max.$, unter der Nebenbedingung $B^T B = I$.

Aus der VL Konvexe Optimierung (Lecture 10)

<u>Theorem</u> (Fan)

Sei $C \in S^n$ sym. Matrix mit Eigenwerten $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$.
Dann gilt

$$\lambda_1 + \ldots + \lambda_k = \max\{ Tr(CBB^T) : B \in \mathbb{R}^{n \times k}, B^T B = I_k \}$$

$$= \max\{ Tr(CY) : Y \in S^n, Tr(Y) = k, Y \succeq 0, I - Y \succeq 0 \}.$$

Wende dies auf $C = XX^T$ an. Dann

$$E[B] = \mathrm{Tr}(XX^T) - \mathrm{Tr}(BB^T XX^T)$$
$$= \sum_{i=1}^{D} \lambda_i - \sum_{i=1}^{M} \lambda_i = \sum_{i=M+1}^{D} \lambda_i .$$

Dass B aus den zu $\lambda_1, \dots, \lambda_D$ gehörigen Eigenvektoren von XX^T besteht, folgt ebenfalls aus dem Beweis des Theorems von Fan. $\boxtimes$

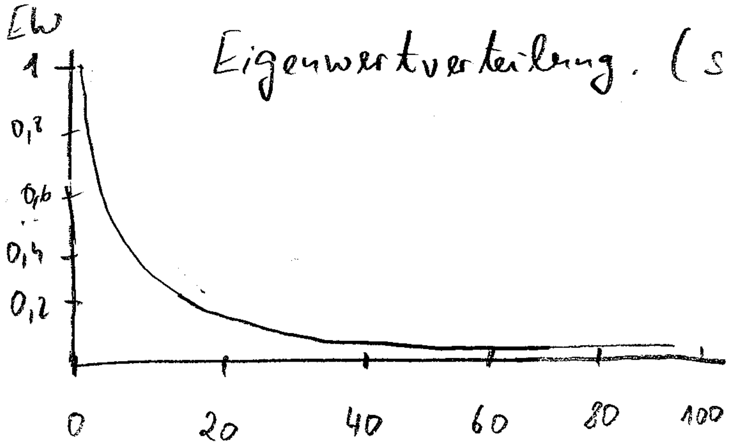
<u>Bsp.</u>: (für die Hauptkomponentenanalyse)

$X \in \mathbb{R}^{784 \times 892}$ Datenmatrix der handgeschriebenen 5 der MNIST-Datenbank

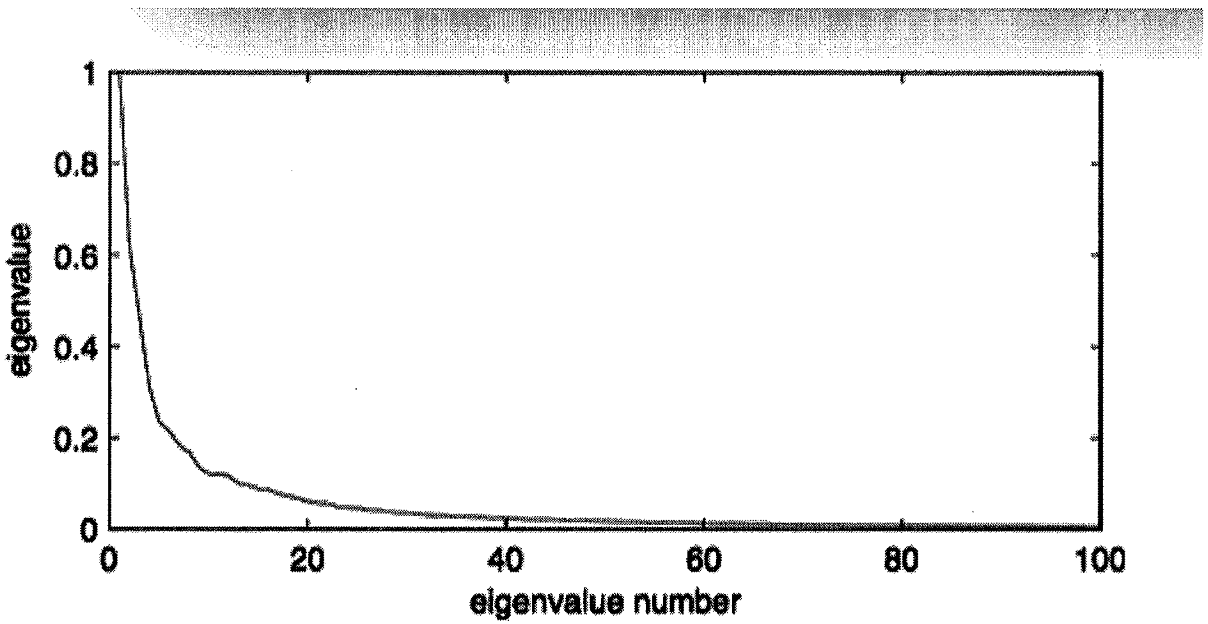
$N = 892$ Anzahl Beispiele

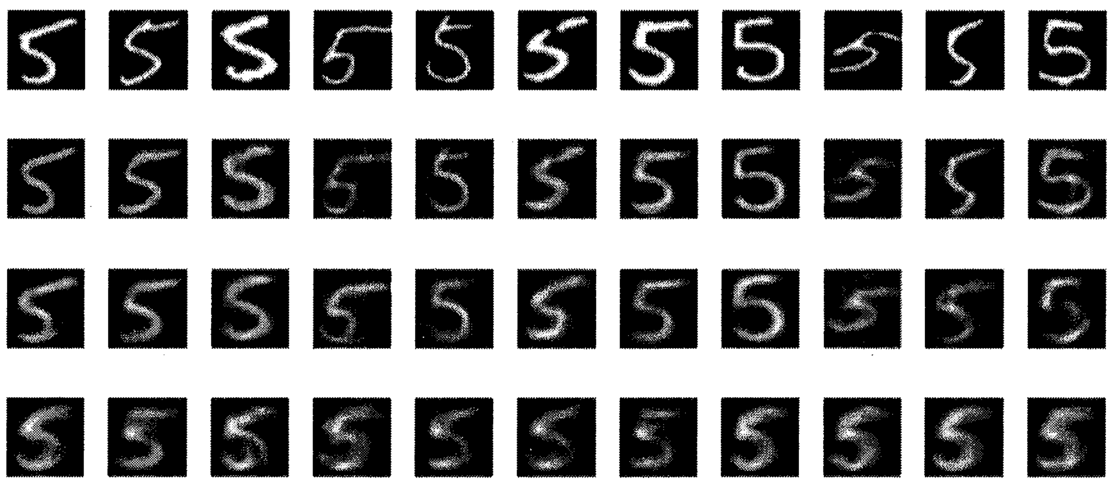
$D = 784 = 28 \times 28$ Pixel pro Buchstabe.

Eigenwertverteilung. (skaliert, so dass $\lambda_1 = 1$).



PCA: Beispiel aus MNIST Datenbank (aus Barber)





Original

100 Eigenvektoren

30 Eigenvektoren

5 Eigenvektoren