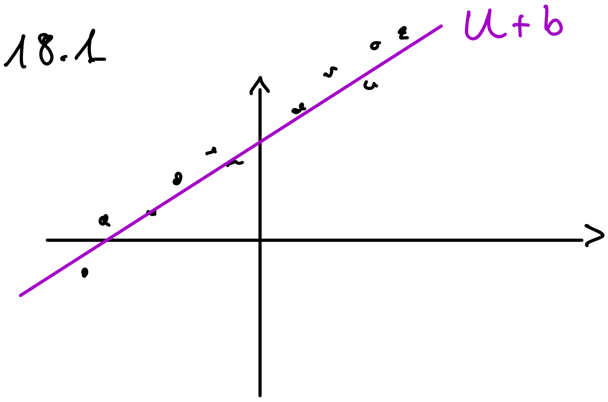
Vorlesung 18 PCA

In dieser Vorlesung behandeln wir Principal Component Analysis (PCA). Dies ist eine Methode zur Reduktion von Dimensionalität. Was heißt das? Wir haben Daten $x_1, \ldots, x_n \in \mathbb{R}^D$ und D könnte sehr groß sein, obwohl die "intrinsische" Dimension der Daten klein ist. In PCA heißt das, dass die x_i "nah" an einem (affin) linearen Unterraum $U+b$ von Dimension $d \ll D$ liegen.

Beispiel 18.1

$D=2$



Motivation für Reduktion der Dimensionalität ist z.B. Data-Compression oder die Motivation Informationen über die Geometrie der Daten zu lernen.

Im Folgenden werden wir PCA auf die Input-Daten $x_1, \ldots, x_n \in \mathbb{R}^D$ anwenden und die Response-Variablen ignorieren. Learning ohne Response-Variablen heißt auch Unsupervised Learning. Im Gegensatz dazu heißen die Learning-Probleme aus den letzten Vorlesung Supervised Learning.

Wir nehmen zunächst die Dimension d als gegeben an.

Oft sind Daten nicht-linear. Daher arbeiten wir wieder mit einer sogenannten Feature-Map

$$\phi : \mathbb{R}^D \to \mathbb{R}^M$$

und setzen

$$z_i := \phi(x_i). \qquad 1 \leq i \leq n$$

Wir wollen dann PCA auf die $z_1, \dots, z_n$ anwenden.

z.B. falls $\phi(x) = ($Monome vom Grad $\leq k$ an x ausgewertet$)$, dann bedeutet PCA zu verstehen, ob polynomielle Gleichungen vom Grad k auf den Daten verschwinden (= auf den Daten sollen die Gleichungen ≈ 0 sein).

Wir definieren wieder die Feature-Matrix

$$\begin{aligned} \Omega &= [\phi(x_1) \; \dots \; \phi(x_n)]^T \\ &= [z_1 \; \dots \; z_n]^T \quad \in \mathbb{R}^{n \times M}. \end{aligned}$$

Im Folgenden nehmen wir an, dass $x_1, \dots, x_n$ unabhängige Ziehungen einer Zufallsvariable $x \in \mathbb{R}^D$ sind. Wir setzen $z := \phi(x)$, so dass

$$z_1, \dots, z_n \overset{iid}{\sim} z \in \mathbb{R}^M$$

Wir setzen $\mu := \mathbb{E} z \in \mathbb{R}^M$ und die Covarianzmatrix.

$$\Sigma = \begin{pmatrix} \mathrm{Cov}(z^{(1)}, z^{(1)}) & \mathrm{Cov}(z^{(1)}, z^{(M)}) \\ \vdots & \\ \mathrm{Cov}(z^{(M)}, z^{(1)}) & \mathrm{Cov}(z^{(\ )}, z^{(M)}) \end{pmatrix} \in \mathbb{R}^{M \times M}.$$

Die Covarianzmatrix Σ ist positiv-semidefinit (siehe Übung) und hat daher Eigenwerte $\lambda_1 \geq \ldots \geq \lambda_M \geq 0$

Definition 18.2

Seien $z_1, \ldots, z_n \in \mathbb{R}^M$.

1) Der empirische Mittelwert der z_i ist $\bar{z} := \frac{1}{n} \sum_{i=1}^{n} z_i$

2) Die empirische Covarianzmatrix

$$S = (s_{ij}) \in \mathbb{R}^{M \times M} \quad \text{mit}$$

$$s_{ij} = \frac{1}{n} \sum_{k=1}^{n} ((z_k)_i - \bar{z}_i)((z_k)_j - \bar{z}_j)$$

Bemerkung Oft werden die Daten durch $(z_k)_i \to \frac{(z_k)_i}{\sqrt{s_{ii}}}$

Es gilt dann für $e = (1, \ldots, 1)^T \in \mathbb{R}^n$:

$$\bar{z} = \frac{1}{n} \Omega^T e \quad \text{und} \quad S = \frac{1}{n} (\Omega^T - \bar{z} e^T)(\Omega^T - \bar{z} e^T)^T$$

Insbesondere ist auch S positiv semidefinit.

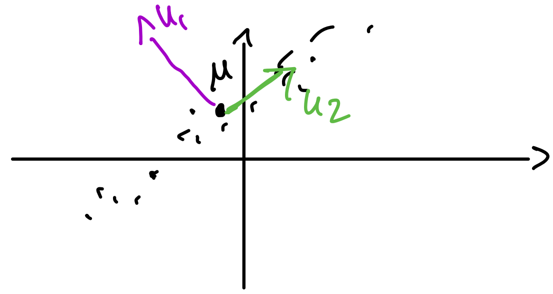
Definition 18.3

Sei $z \in \mathbb{R}^M$ eine Zufallsvariable mit $\mu := \mathbb{E} z \in \mathbb{R}^M$ und sei $d < M$. Ein Raum von maximaler Varianz $U \subset \mathbb{R}^M$ ist ein linearer Unterraum von Dimension $d = \dim(U)$, so dass

$$U \in \operatorname*{argmax}_{\dim U = d} \mathbb{E} \, \| P_U (z - \mu) \|^2,$$

wobei P_U = orthogonale Projektion auf U.

<u>Geometrische Idee:</u> Ein Raum von max. Varianz ist ein lin. Unterraum in dessen Richtung sich die Daten am ehesten ausstrecken.



Die Daten strecken sich eher in Richtung u_2 aus.

Unsere erste Interpretation eines Raumes, der "nah" an den Daten liegt, ist ein Raum maximaler Varianz.

<u>Theorem 18.4</u> Sei $d < D$.
Seien $\lambda_1 \geq \ldots \geq \lambda_M \geq 0$ die Eigenwerte der Covarianzmatrix Σ und sei u_i ein Eigenvektor von λ_i, s.d. $\langle u_i, u_j \rangle = \delta_{ij} = \begin{cases} 1, & i=j \\ 0, & i \neq j \end{cases}$
Dann ist

$$U = \operatorname{span}\{u_1, \ldots, u_d\}$$

ein Raum maximaler Varianz. Außerdem ist U eindeutig bestimmt, falls $\lambda_d > \lambda_{d+1}$.

- In der Praxis haben wir Σ und μ nicht zur Hand und approximieren sie durch S und $\bar{z}$.
- Das Theorem zeigt, dass d wie folgt wählen können: $\lambda_d > 0$, $\lambda_{d+1} \approx 0$.
  Oder wir wählen d, so dass $\lambda_d / \lambda_{d+1}$ maximal.

<u>Beweis von Theorem 18.4</u>
Sei $U \subset \mathbb{R}^M$ ein lin. Unterraum von Dimension d.
Sei $\{u_1, \ldots, u_d\}$ eine ONB von U und setze

$$A = [u_1, \dots, u_d] \in \mathbb{R}^{n \times d}.$$

Es gilt: $A^T A = \mathbb{1}_d$. Es gilt dann

$$P_U = A A^+.$$

und $$A^+ = (A^T A)^{-1} A^T = A^T$$

Daher
$$P_U = A A^T.$$

Somit:
$$\begin{aligned}
\| P_U (z - \mu) \|^2 &= \| A A^T (z - \mu) \|^2 \\
&= (A A^T (z-\mu))^T (A A^T (z - \mu)) \\
&= (z-\mu)^T A A^T A A^T (z-\mu) \\
&= (z-\mu)^T A A^T (z - \mu).
\end{aligned}$$

Außerdem gilt: $$A A^T = \sum_{i=1}^{d} u_i u_i^T$$

Daher:
$$\begin{aligned}
\| P_U (z-\mu) \|^2 &= (z-\mu)^T \sum_{i=1}^{d} u_i u_i^T (z-\mu) \\
&= \sum_{i=1}^{d} (z-\mu)^T u_i u_i^T (z-\mu) \\
&= \sum_{i=1}^{d} u_i^T (z-\mu)(z-\mu)^T u_i
\end{aligned}$$

Wir erhalten: $$\mathbb{E} \| P_U (z-\mu) \|^2 = \sum_{i=1}^{d} u_i^T \left(\mathbb{E} (z-\mu)(z-\mu)^T \right) u_i = \sum_{i=1}^{d} u_i^T \Sigma u_i$$

Wir maximieren diesen Ausdruck per Lagrange-Multiplier.

$$\mathcal{L}(u_1,\dots,u_d,l_{ij}) = \sum_{i=1}^{d} u_i^T \Sigma u_i - \sum_{1\le i\le j\le d} (u_i^T u_j - \delta_{ij}) l_{ij}$$

Es gilt:

$$(*) \quad \frac{\partial \mathcal{L}}{\partial u_j} = 2\Sigma u_j - 2 u_j l_{jj} - \sum_{i<j} u_i l_{ij} = 0 \quad 1\le j\le d$$

$$(**) \quad \frac{\partial \mathcal{L}}{\partial l_{ij}} = u_i^T u_j - \delta_{ij} = 0.$$

(*) für $j=1$ gibt: $2\Sigma u_1 - 2u_1 l_{11} = 0$

$\Rightarrow \Sigma u_1 = u_1 l_{11} \quad \Rightarrow u_1$ ist Eigenvektor von Σ.

(*) für $j=2$ gibt: $2\Sigma u_2 - 2u_2 l_{22} - u_1 l_{12}$

$\Rightarrow 2u_1^T\Sigma u_2 - 0 \quad - l_{12} = 0 \Rightarrow l_{12} = 0$

$\Rightarrow u_2$ ist Eigenvektor von Σ

.... $u_3,\dots,u_d$ sind auch Eigenvektoren von Σ.

$$\Rightarrow \sum_{i=1}^{d} u_i^T\Sigma u_i = \sum_{i=1}^{d} \lambda_i. \quad \text{wobei } \Sigma u_i = \lambda_i u_i.$$

Dieser Ausdruck wird maximiert, indem wir $\lambda_1 \geq \dots \geq \lambda_d \geq \lambda_{d+1} \dots \geq 0$ wählen.

Außerdem, wenn $\lambda_d > \lambda_{d+1}$, ist U eindeutig bestimmt als Summe der Eigenräume von $\lambda_1,\dots,\lambda_d$. □

Nachdem wir einen Raum maximaler Varianz mit Hilfe von S und $\bar{z}$ bestimmt haben, würden wir die Daten wie folgt repräsentieren

$$z_i \longmapsto P_U(z_i - \bar{z}) + \bar{z}.$$

Im zweiten Ansatz würden wir den quadratischen Abstand

$$\sum_{i=1}^{n} \| (z_i - \overline{z}) - P_U (z_i - \overline{z}) \|^2$$

über U minimieren.