Vorlesung 19 PCA cont'd

Erinnerung: Wir haben Daten $z_1, \dots, z_n \in \mathbb{R}^M$ und $d < M$.
In PCA suchen wir einen linearen Raum $U \subset \mathbb{R}^M$ von Dim. d und $b \in \mathbb{R}^M$, s.d.

$$U + b \subset \mathbb{R}^M$$

"nah" an den Daten.

Unser erstes Konzept von "nah" war ein Raum von maximaler Varianz. Dazu haben wir angenommen, dass $z_1, \dots, z_n$ unabh. Samples einer ZV $z \in \mathbb{R}^M$. Dann war

$$U = \operatorname{span}\{u_1, \dots, u_d\}, \quad b = \mu,$$

wobei $\mu = \mathbb{E}z$, $u_1, \dots, u_d$ Eigenvektoren der Covarianzmatrix Σ von z zu den größten Eigenwerten.

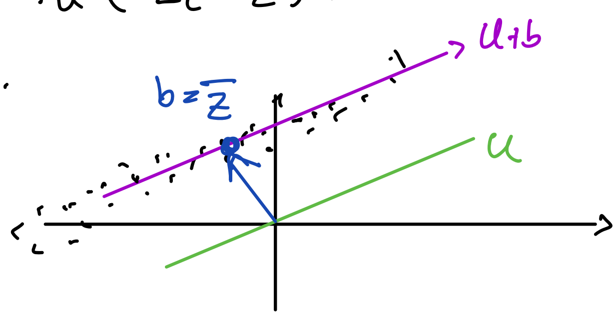
In der Praxis approximieren wir Σ durch die empirische Covarianzmatrix S und μ durch den empirischen Mittelwert $\bar{z}$.

Bemerkung: $u_1, \dots, u_d$ heißen Hauptkomponenten ("Principal Components")

Nun betrachten wir ein zweites Konzept von "nah".
Wir wählen U, s.d.

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

minimiert wird und $b = \bar{z}$.



Theorem 18.1

Seien $\lambda_1 \geq \dots \geq \lambda_M \geq 0$ die Eigenwerte der emp. Covarianzmatrix S. Sei u_i Eigenvektor zu λ_i, s.d. $\langle u_i, u_j \rangle = \delta_{ij}$. Dann minimiert

$$\mathcal{U} = \operatorname{span}\{u_1, \dots, u_d\}$$

den Ausdruck in (*). Falls $\lambda_d > \lambda_{d+1}$, ist $\mathcal{U}$ eindeutig bestimmt.

Beweis Sei $u_1, \dots, u_M$ eine Orthonormal-Basis von $\mathbb{R}^M$, d.h. $\langle u_i, u_j \rangle = \delta_{ij}$.
Sei $w_i = z_i - \bar{z}$. Sei außerdem $A = [u_1, \dots, u_d] \in \mathbb{R}^{M \times d}$
und $W = [w_1 \dots w_n] \in \mathbb{R}^{M \times n}$
Wir haben außerdem die Feature-Matrix $\Omega = [z_1 \dots z_n]^T \in \mathbb{R}^{n \times M}$.
Dann gilt:

$$W = \Omega^T - \bar{z} e^T, \quad e = (1, \dots, 1) \in \mathbb{R}^n.$$

Dann gilt:

$$\begin{aligned} WW^T &= (\Omega^T - \bar{z}e^T)(\Omega^T - \bar{z}e^T)^T \\ &= nS. \end{aligned}$$

Außerdem: $AA^T = \sum_{i=1}^{d} u_i u_i^T$ und $\mathbb{1}_M = \sum_{i=1}^{M} u_i u_i^T$
Dann gilt:

$$\begin{aligned} W - P_{\mathcal{U}} W &= W - AA^T W \\ &= (\mathbb{1}_M - AA^T) W \\ &= \left(\sum_{i=d+1}^{M} u_i u_i^T\right) W \end{aligned}$$

Außerdem, $W - P_{\mathcal{U}} W = [w_1 - P_{\mathcal{U}}(w_1), \dots, w_n - P_{\mathcal{U}}(w_n)]$.
Es gilt:

$$(*) = \operatorname{Trace}\left((W - P_U W)^T (W - P_U W) \right).$$

$$= \operatorname{Trace}\left(W^T \left(\sum_{i=d+1}^{M} u_i u_i^T\right)\left(\sum_{i=d+1}^{M} u_i u_i^T\right) W \right)$$

$$= \operatorname{Trace}\left(W^T \left(\sum_{i=d+1}^{M} u_i u_i^T\right) W \right) \quad , \text{ weil } \langle u_i, u_j \rangle = \delta_{ij}$$

$$= \sum_{i=d+1}^{M} \operatorname{Trace}\left(W^T u_i u_i^T W \right)$$

$$= \sum_{i=d+1}^{M} \operatorname{Trace}\left(u_i^T W W^T u_i \right)$$

$$= n \cdot \sum_{i=d+1}^{M} u_i^T S u_i$$

Wie im Beweis von Thm. 18.4 erhalten wir:

$u_1, \ldots, u_d$ sind Eigenvektoren zu $\lambda_1, \ldots, \lambda_d$.

Eindeutigkeit folgt ebenfalls wie in Thm. 18.4. $\square$.

In beiden Ansätzen müssen wir eine Eigenzerlegung von $S = \frac{1}{n} W W^T$ berechnen. Es ist $S \in \mathbb{R}^{M \times M}$. $W \in \mathbb{R}^{M \times n}$

Falls $n << M$, können wir wie folgt vorgehen. Wir nehmen an, dass $r(W) = n$. Dann sei

$$W = U D V^T$$

ein SVD von W, d.h. $U \in \mathbb{R}^{M \times n}$, $D = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$, $V \in \mathbb{R}^{n \times n}$.

Dann gilt:

$$S = \frac{1}{n} W W^T = U D^2 U^T$$

und die Diagonaleinträge von D^2 = nicht-null Eigenwerte von S.

Außerdem: $\mathbb{R}^{n \times n} \ni W^T W = V D^2 V^T$

$\rightsquigarrow$ eine Eigenzerlegung von $\frac{1}{n} W^T W$ gibt uns die gleichen Eigenwerte wie für S.

Außerdem $u_i = \frac{W v_i}{\|W v_i\|}$, so dass wir auch die Hauptkomponenten von S erhalten

Insbesondere zeigt dies, dass PCA durch SVD verstanden werden kann.

Außerdem kann $W^T W$ durch die Kernel-Map $K(x_i, x_j) = \langle z_i, z_j \rangle = \langle \phi(x_i), \phi(x_j) \rangle$ berechnet werden.

<u>Lemma 18.2</u>

Sei $G = (K(x_i, x_j))_{i,j=1}^n$ für Daten $x_1, \dots, x_n \in \mathbb{R}^D$, $z_i = \phi(x_i)$.
Dann gilt für $W = [z_1 - \bar{z}, \dots, z_n - \bar{z}]$ wie oben, dass

$$W^T W = (\mathbb{1}_n - \tfrac{1}{n} e e^T) G (\mathbb{1}_n - \tfrac{1}{n} e e^T).$$

<u>Beweis</u>

Es ist $W = \Omega^T - \bar{z} e^T$ und $\bar{z} = \frac{1}{n}(z_1 + \dots + z_n) = \frac{1}{n} \Omega^T e$

$$\begin{aligned}
\rightsquigarrow W^T W &= (\Omega^T - \tfrac{1}{n} \Omega^T e e^T)^T (\Omega^T - \tfrac{1}{n} \Omega^T e e^T) \\
&= (\Omega^T (\mathbb{1}_n - \tfrac{1}{n} e e^T))^T (\Omega^T (\mathbb{1}_n - \tfrac{1}{n} e e^T)) \\
&= (\mathbb{1}_n - \tfrac{1}{n} e e^T) \Omega \Omega^T (\mathbb{1}_n - \tfrac{1}{n} e e^T)
\end{aligned}$$

Es ist: $\Omega = \begin{bmatrix} - z_1^T - \\ \vdots \\ - z_n^T - \end{bmatrix}$. D.h. $\Omega \Omega^T = (\langle z_i, z_j \rangle)_{i,j=1}^n = G$

$\rightsquigarrow W^T W = (\mathbb{1}_n - \frac{1}{n} e e^T) G (\mathbb{1}_n - \frac{1}{n} e e^T)$. $\square$

Als nächstes betrachten wir folgendes statistisches Setting für PCA. Wir nehmen an, dass die Daten $x_1, \dots, x_n \in \mathbb{R}^D$ von folgender Zufallsvariable gesampled sind:

$$X = A\xi + b + \varepsilon,$$

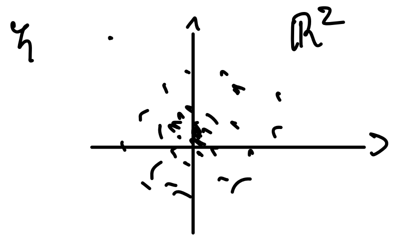
mit $\underset{r(A)=d.}{A \in \mathbb{R}^{D\times d}}$, $\xi \sim N(0, \mathbb{1}_d)$, $b \in \mathbb{R}^D$, $\varepsilon \sim N(0, \sigma^2 \mathbb{1}_D)$.

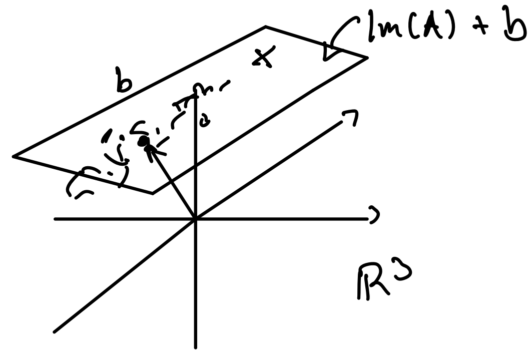
Im Folgenden arbeiten wir ohne Feature map.

D.h.

$$(x \mid \xi) \sim N(A\xi + b, \sigma^2 \mathbb{1}_D).$$

Cartoon: $d=2$, $D=3$





<u>Proposition</u> 18.3

Im obigen Setting haben wir

$$x \sim N(b, AA^T + \sigma^2 \mathbb{1}_D).$$

<u>Beweis</u>

Es gilt: $P(X) = \int_{\mathbb{R}^d} P(x \mid \xi) P(\xi) \, d\xi$

Es ist:

$$\log P(x \mid \xi) + \log P(\xi)$$

$$= -\frac{1}{2\sigma^2} \| x - (A\xi + b) \|^2 - \frac{1}{2} \|\xi\|^2 + c$$

$$= -\frac{1}{2} (x-\mu) \Sigma^{-1} (x-\mu) - \frac{1}{2} (\xi - \nu) B^{-1} (\xi - \nu) + c'$$

mit c, c' unabh. von x und η und $\mu \in \mathbb{R}^D$, $\gamma \in \mathbb{R}^d$
$\Sigma \in \mathbb{R}^{D \times D}$, $B \in \mathbb{R}^{d \times d}$ und μ und Σ unabh. von η.
D.h., nachdem wir η ausintegriert haben, bleibt ein Ausdruck der Form $e^{-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)} \frac{1}{\sqrt{(2\pi)^D \det(\Sigma)}}$.

$$\leadsto \quad x \sim N(\mu, \Sigma).$$

Es ist
$$\mu = \mathbb{E}_x x = \mathbb{E}_{\varepsilon,\eta}(A\eta + b + \varepsilon) = b$$
$$\Sigma = \mathbb{E}_x (x-b)(x-b)^T$$
$$= \mathbb{E}_{\eta,\varepsilon}(A\eta + b + \varepsilon - b)(A\eta + b + \varepsilon - b)^T$$
$$= \mathbb{E}_{\eta,\varepsilon}(A\eta + \varepsilon)(A\eta + \varepsilon)^T$$
$$= \mathbb{E}_{\eta,\varepsilon}(A\eta\eta^T A^T + A\eta\varepsilon^T + \varepsilon\eta^T A^T + \varepsilon\varepsilon^T)$$
$$= \mathbb{E}_{\eta,\varepsilon}(A\eta\eta^T A^T + \varepsilon\varepsilon^T).$$
$$= A\, \mathbb{E}\eta\eta^T A^T + \mathbb{E}\varepsilon\varepsilon^T$$
$$= AA^T + \sigma^2 \mathbb{1}_D. \qquad \square.$$

<u>Korollar 18.4</u>

Angenommen wir haben den Prior $\eta \sim N(\gamma, B)$. Dann erhalten wir
$$x \sim N(A\gamma + b, ABA^T + \sigma^2 \mathbb{1}_D).$$

<u>Beweis</u>

Es ist $\eta = R\eta' + \gamma$ mit $\eta' \sim N(0, \mathbb{1}_D)$, und $RR^T = B$.
Dahr:
$$x = A\eta + b = AR\eta' + A\gamma + b.$$
Wir wenden Thm. 18.3 mit AR und $A\gamma + b$ an. $\square$.

Mit Hilfe von Kor. 18.4 können wir jetzt MLE für die Parameter A, b bestimmen. Wir können auch die Verteilung von $(\xi | x)$ ausrechnen, um den Prior von ξ nach Sicht von x upzudaten.

<u>Theorem 18.5</u>

Sei $\xi \sim N(v, B)$ und $(x | \xi) \sim N(A\xi + b, \sigma^2 1_D)$.

Dann gilt:

$$(\xi | x) \sim N(m, C).$$

mit $C = (\sigma^{-2} A^T A + B^{-1})^{-1}$ und $m = C(\sigma^{-2} A^T (x-b) + B^{-1} v)$

<u>Beweis</u>

Es gilt, nach Bayes' Theorem:

$$\log P(\xi | x) = \log P(x | \xi) + \log P(\xi) + c$$

mit c unabh. von ξ und x.

Es ist $(x|\xi) \sim N(A\xi + b, \sigma^2 1_D)$ und $\xi \sim N(v, B)$, s.d.

$$\log P(\xi | x) = -\frac{1}{2\sigma^2} \| x - (A\xi + b) \|^2 - \frac{1}{2}(\xi - v)^T B^{-1} (\xi - v) + c''$$

mit c'' unabh. von x und ξ.

$$= -\frac{1}{2\sigma^2} \| A\xi - (x-b) \|^2 - \frac{1}{2}(\xi - v)^T B^{-1} (\xi - v).$$

Das wollen wir nun nach ξ auflösen. Wir gehen wie in Thm 14.2 vor und erhalten die Formeln für m und C. $\square$