Lecture 8 (March 5): PCA cont'd.

Recall:

→ population $X \sim (\mu, \Sigma)$

find u, $\|u\| = 1$, s.t. $\mathrm{Var}(u^T X)$ maxim.

→ $\Sigma = U \Lambda U^T$ $\lambda_1 \geq \dots \geq \lambda_m$

→ PCA tranformation $Z = U^T X$

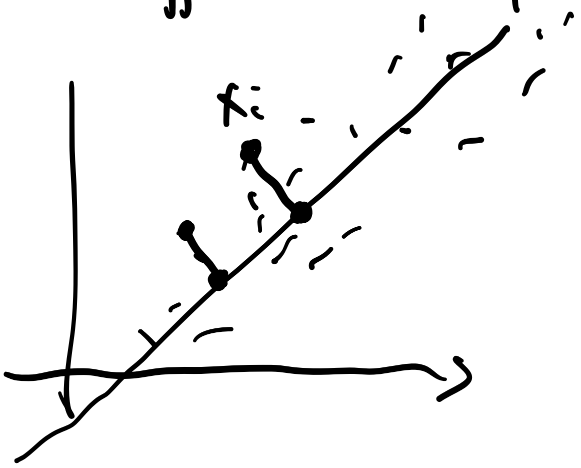
→ DATA $\underline{x}_1, \dots, \underline{x}_n \in \mathbb{R}^m$

$$S_n = U \Lambda U^T$$

<u>scores</u> : d principal directions

$$y_i = (u_1^T \underline{x}_i, \dots, u_d^T x_i)$$

Best d-dim'l affine subspace approx. the data



① affine subspace: $\mu + \mathcal{L}$ linear subspace $\mathcal{L} \subseteq \mathbb{R}^m$

note: w.l.o.g $\mu \in \mathcal{L}^\perp$

$$\left\{ \begin{array}{l} \mu = \mu' + \mu'' \quad \text{where } \mu' \in \mathcal{L},\ \mu'' \in \mathcal{L}^\perp \\ \mu + \mathcal{L} = \mu' + \mu'' + \mathcal{L} = \mu'' + \mathcal{L} \end{array} \right.$$

② $\mathcal{L} = \operatorname{span}\{\underbrace{w_1, \dots, w_d}_{\text{orthonormal}}\} \subseteq \mathbb{R}^m$

$$W = \begin{pmatrix} | & & | \\ w_1 & \cdots & w_d \\ | & & | \end{pmatrix} \in \mathbb{R}^{m \times d} \rightarrow W^T W = I_d$$

every elem. $y \in \mathcal{L}$ is of the form

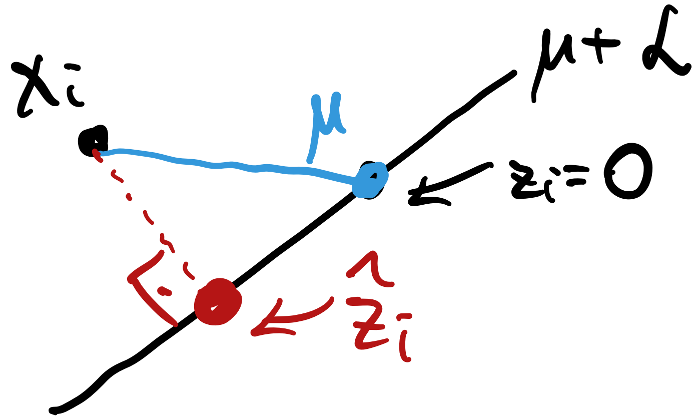
$$y = Wz \qquad z \in \mathbb{R}^d$$

$$y \in \mu + \mathcal{L} \implies y = \mu + Wz$$

note: $W^T \mu = 0$

③ find μ, W, z_i's st.

$$\sum_{i=1}^{n} \underbrace{\| x_i - (\mu + W z_i) \|^2} \longrightarrow \min.$$



$$\hat{z}_i = (W^T W)^{-1} W^T x_i = W^T x_i$$

④ minim. $\sum_{i=1}^{n} \| x_i - (\mu + W W^T x_i) \|^2$

wrt μ, W, $W^T W = I_d$

optimize wrt μ.

$\hat{\mu} = 0$ (data centered)

⑤ minim. $\sum_{i=1}^{n} \| x_i - W W^T x_i \|^2$

wrt W, $W^T W = I_d$

⑥ Recall

$$\|A\|_F^2 := \operatorname{tr}(A^T A) = \operatorname{tr}(A A^T) = \sum_{i,j} A_{ij}^2$$

minimize $\|\underset{n\times m}{X} - \underbrace{X \cdot W W^T}_{n\times d \; d\times m}\|_F$

ith row $\underline{x}_i - WW^T \cdot x_i$

XWW^T has rank $\le d$

7. Eckart - Young :

$$\min \|\underline{X} - M\|_F^2 \text{ s.t. } \operatorname{rank}(M) \le d$$

$X = VDU^T$ then

$\hat{M} = VD_dU^T$ $d<m$

$$\underset{n\times m}{D} = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_m \\ \hline & 0 & \end{pmatrix} \qquad D_d = \begin{pmatrix} \sigma_1 \cdots \sigma_d & \underbrace{0 \cdots 0}_{m-d} \\ \hline 0 \end{pmatrix}$$

8. show $\hat{M} = \underline{XWW^T}$

for some W

$W^T W = I_d$

take $W = \begin{pmatrix} | & & | \\ u_1 & \cdots & u_d \\ | & & | \end{pmatrix} =: \underset{m \times d}{U_d}$

$\hat{M} = V \boxed{D_d U^T}$ 1

$\text{RHS} = X \cdot U_d \cdot U_d^T$

$= V \cdot D \cdot \underbrace{U^T \cdot U_d}_{\underset{m \times d}{\begin{array}{c} \\ m-d \{ \end{array} \begin{pmatrix} I_d \\ 0 \end{pmatrix}^{d \times d}}} U_d^T$

$= V \cdot \boxed{D \cdot \begin{pmatrix} U_d^T \\ 0 \end{pmatrix}}$ 2

1

$=$

m

U^T

$$\begin{array}{c}\left(\begin{array}{cc}\begin{matrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_d\end{matrix} & 0 \\ 0 & \begin{matrix}0 & & \\ & \ddots & \\ & & 0\end{matrix} \\ 0 & 0\end{array}\right) \\ D_d\end{array}\begin{array}{l}\left.\vphantom{\begin{matrix}1\\1\\1\\1\\1\\1\end{matrix}}\right\} m \\ \left.\vphantom{1}\right\} n-m\end{array}\quad \begin{array}{c}\begin{pmatrix} U_d^T \\ \\ * \end{pmatrix} \\ U^T\end{array}$$

(2)

$$\begin{pmatrix}\begin{matrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_d\end{matrix} & 0 \\ 0 & \begin{matrix}\sigma_{d+1} & & \\ & \ddots & \\ & & \sigma_m\end{matrix} \\ 0 & 0\end{pmatrix}\begin{pmatrix} U_d^T \\ \\ 0 \end{pmatrix}$$

this show that

$W = U_d$ is an

optimum

optimal affine subspace is linear and it is spanned by $\underline{u}_1, \dots, u_d$

$$X = V D U^T \leftarrow \text{centered}$$

$$S_n = \frac{1}{n} X^T X$$

$$= \frac{1}{n} U D^T \underbrace{V^T V}_{} D U^T$$

$$= \frac{1}{n} U \underbrace{(D^T D)}_{} U^T$$

$$= U \left(\frac{1}{n} D^T D \right) U^T$$

so it is the same as in PCA.

Probabilistic PCA

$X = (X_1, \dots, X_m)$

$$\boxed{X = \mu + \underset{m \times r}{W} \cdot Z + \varepsilon}$$

$\mu \in \mathbb{R}^m$, $Z \sim N_r(0, I)$

$\varepsilon \sim N_m(0, \sigma^2 I_m)$, $\varepsilon \perp\!\!\!\perp Z$

X is then also Gaussian

$$\mathbb{E}X = \mu$$

$$Var(X) = \Sigma = WW^T + \sigma^2 I_m$$

Z, ε are latent

Data: $\underline{X}_1, \dots, \underline{X}_n \in \mathbb{R}^m$

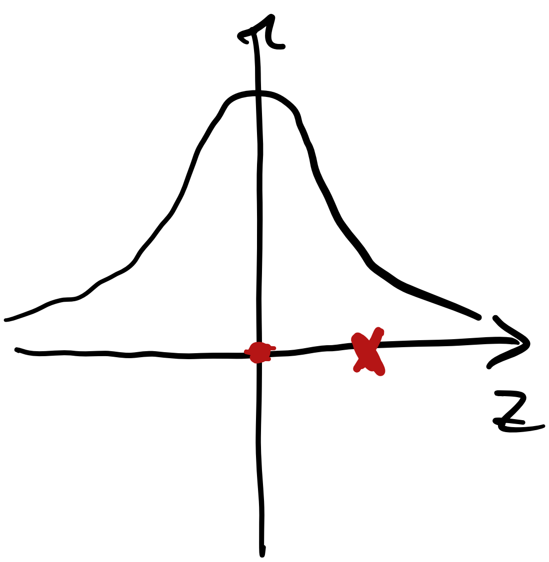
observations of X

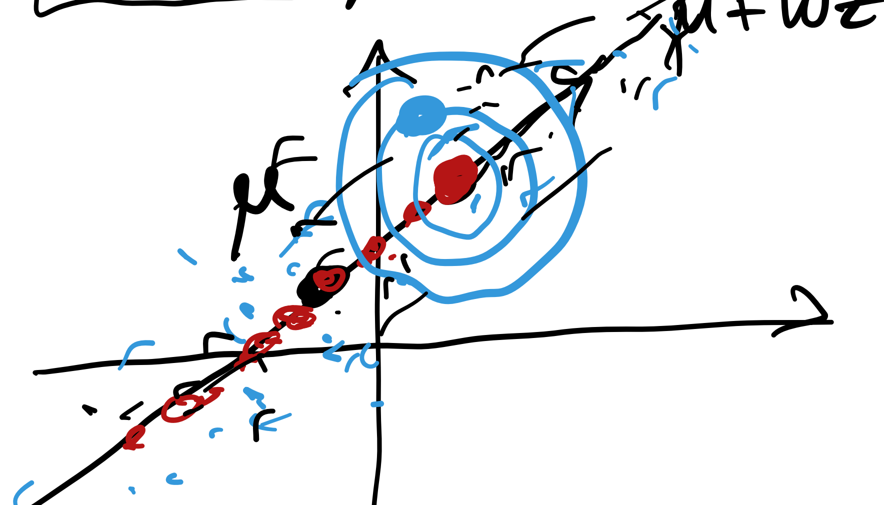
e.g., $m = 2$, $r = 1$

$W \in \mathbb{R}^{2\times 1}$ $\quad w = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$

$Z \sim N(0, 1)$

$$X = \underbrace{\mu + \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \cdot Z}_{\mu + Wz} + \underline{\varepsilon}$$





⚠ MLE given in a closed form

$X \sim N(\mu, WW^T + \sigma^2 I_m)$

est. μ, W, σ^2

$$(WR)(WR)^T = WW^T$$

if $R \in O(m)$

⚠ the model is not identifiable.

MLE:

$\hat{\mu} = \overline{X}_n$

$\hat{\sigma}^2 = \frac{1}{m-r} \sum^{m} \lambda_i$

$S_n = U \Lambda U^T$

$\lambda_1 \geq \dots \geq \lambda_m$

$i=r+1$

$$\hat{W} = U_r \underbrace{\left(\Lambda_r - \hat{\sigma}^2 I_r\right)^{1/2} R}_{\tilde{R}} \quad \text{any } O(m)$$

Observation: The column span of $\hat{W}$ is the same as the column span of U_r

$x \in \mathrm{colspan}(\hat{W})$

$$x = \hat{W} y \quad \text{for some } y$$

$$= U_r \cdot \underbrace{\tilde{R} \cdot y}_{\tilde{y}} = U_r \cdot \tilde{y}$$

$\Rightarrow x \in \text{colspan}(U_r)$

---

take $x = U_r y$

$$= U_r \tilde{R} \underbrace{\tilde{R}^{-1} y}_{\tilde{y}} = \underbrace{U_r \tilde{R}}_{\hat{W}} \tilde{y}$$