# FYS5429/9429, MARCH 22, 2023

## Basics of PCA

Data set $X \in \mathbb{R}^{n \times p}$

# samples   # features

$$X = \begin{bmatrix} | & | & & | \\ x_0 & x_1 & \cdots & x_{p-1} \\ | & | & & | \end{bmatrix}$$

$$\mathrm{cov}[x_i\, x_j] = \frac{1}{n}\sum_k (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

$$x_i = \begin{bmatrix} x_{0i} \\ x_{1i} \\ \vdots \\ x_{n-1\,i} \end{bmatrix}$$

$$\bar{x}_i = \frac{1}{n}\sum_k x_{ki}$$

$$\tilde{x}_{ki} = x_{ki} - \bar{x}_i \Rightarrow x_{ki}$$

$$X = \begin{bmatrix} x_{00} & x_{01} \\ x_{10} & x_{11} \\ x_{20} & x_{21} \end{bmatrix} = \begin{bmatrix} | & | \\ x_0 & x_1 \\ | & | \end{bmatrix}$$

$$\mathrm{cov}[x_0, x_1] = \frac{1}{n}\sum_k x_{k0}\, x_{k1}$$

$$= \frac{1}{n}\left( x_{00} x_{01} + x_{10} x_{11} + x_{20} x_{21} \right)$$

$$= \begin{bmatrix} x_{00} & x_{10} & x_{20} \end{bmatrix} \begin{bmatrix} x_{01} \\ x_{11} \\ x_{21} \end{bmatrix}$$

$$= x_0^T x_1 \frac{1}{n}$$

$$\mathrm{cov}[X] = \frac{1}{n} \cancel{X}^T X$$

Design matrix

$$X \in \mathbb{R}^{n \times p}$$

$$\mathrm{cov}[\cancel{X}] = \frac{1}{n} \begin{bmatrix} x_0^T x_0 & x_0^T x_1 \\ x_1^T x_0 & x_1^T x_1 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_0^2 & \mathrm{cov}[x_0 x_1] \\ \mathrm{cov}[x_1, x_0] & \sigma_1^2 \end{bmatrix}$$

$$\mathrm{corr}[\cancel{X}] \propto \frac{\mathrm{cov}[x_i, x_j]}{\sqrt{\mathrm{var}[x_i]\,\mathrm{var}[x_j]}}$$

$$\mathrm{corr}[X] = \begin{bmatrix} 1 & \mathrm{corr}[x_0, x_1] \\ \mathrm{corr}[x_1, x_0] & 1 \end{bmatrix}$$

SVD : singular value decomp

$$X = U \Sigma V^T$$

$$U \in \mathbb{R}^{n \times n} \qquad \Sigma \in \mathbb{R}^{n \times p}$$

$$V \in \mathbb{R}^{p \times p}$$

$$U^T U = U U^T = \mathbb{1}_{n \times n}$$

$$V V^T = V^T V = \mathbb{1}_{p \times p}$$

$$\Sigma = \begin{bmatrix} \lambda_0 & & & & \\ & \lambda_1 & & & \\ & & \ddots & & \\ & & & \lambda_{p-1} & \\ & & & & \ddots \\ & & & & 0 \end{bmatrix}$$

$$\lambda_0 > \lambda_1 > \lambda_2 > \dots > \lambda_{p-1} > 0$$

$$\Sigma = \begin{bmatrix} \tilde{\Sigma} \\ 0 \end{bmatrix}$$

$$\tilde{\Sigma} \in \mathbb{R}^{p \times p}$$

$$\mathrm{Cov}[X] = \frac{1}{n} X^T X$$

$$= \frac{1}{n} \left( V \Sigma^T \underbrace{U^T U}_{} \Sigma V^T \right)$$

$$= \frac{1}{n} V \Sigma^T \Sigma V^T \in \mathbb{R}^{p \times p}$$

$$X^T X = V \Sigma^T \Sigma V^T$$

multiply with $V$ from right

$$(X^T X) V = V \Sigma^T \Sigma \underbrace{V^T V}_{\mathbb{1}}$$

$$= V \Sigma^T \Sigma = V \Sigma^2$$

$$V = \begin{bmatrix} | & | & & | \\ v_0 & v_1 & \cdots & v_{p-1} \\ | & | & & | \end{bmatrix}$$

$$(X^T X) v_i = v_i \lambda_i^2$$

multiply with $\frac{1}{n}$

$$\left(\frac{1}{n} X^T X\right) v_i = \frac{1}{n} v_i \lambda_i^2 =$$

$$\mathrm{Cov}[x] \quad = v_i \frac{\lambda_i^2}{n}$$

The covariance matrix

$$\mathrm{cov}[X] = \frac{1}{n} X^T X$$

$$= \mathbb{E}[X^T X]$$

$$S^T \mathrm{cov}[X]\, S = D$$

$$= \begin{bmatrix} \sigma_0^2 & & \\ & \ddots & \\ & & \sigma_{p-1}^2 \end{bmatrix}$$

$$S^T S = S S^T = \mathbb{1}$$

$$\mathbb{E}[S^T X^T X S] =$$

$$S^T \mathbb{E}[X^T X] S =$$

$\mathrm{cov}[y]$ is uncorrelated matrix (is diagonal), along the diagonal we have the variance

$\mathrm{var}[y]$, $\sigma_0^2/n > \sigma_1^2/n$

$-- \quad \sigma_{p-1}^2/n$

$$S = \begin{bmatrix} | & | & & | \\ s_0 & s_1 & \cdots & s_{p-1} \\ | & | & & | \end{bmatrix}$$

$S = V$ from the SVD

$$\lambda_0^2/n = \sigma_0^2/n \Rightarrow$$

$$\lambda_i^2/n = \sigma_i^2/n$$

$$\mathrm{var}[y_0] > \mathrm{var}[y_1] > \dots$$

$$> \mathrm{var}[y_{p-1}]$$

The PCA theorem states that we can define an optimal dimension $m < p$ $(m \ll p)$, which defines a cutoff

$$\sum_{i=0}^{m-1} \sigma_i^2 \leq \text{cutoff}$$

$\Rightarrow$ Reduced dimensionality

PCA theorem

We assume there is an orthogonal transformation

$$S S^T = S^T S = \mathbb{1}$$

$$S = \begin{bmatrix} | & | & & | \\ s_0 & s_1 & \dots & s_{p-1} \\ | & | & & | \end{bmatrix}$$

$s_0^T s_0 = 1$

we define a Lagrangian

$$s_0^T \mathrm{cov}[x] s_0 + \lambda_0 (1 - s_0^T s_0)$$

optimize wrt $s_0$ and $\lambda_0$

wrt $s_0^T$:

$$\mathrm{cov}[x] s_0 = \lambda_0 s_0$$

multiply by $s_0^T$ =>

$$s_0^T \mathrm{cov}[x] s_0 = \lambda_0$$

= eigenvalue = variance of $\mathrm{cov}[x]$ for $\sigma_0^2/n$

$s_0$ is the first principal component.

Next $s_1$, $s_1^T s_0 = 0$

$$\mathcal{L} = s_1^T \mathrm{cov}[x] s_1 + \lambda_1 (1 - s_1^T s_1)$$
$$+ \gamma s_1^T s_0$$

Take derivatives wrt $\lambda_1, \gamma, s_1$

wrt $s_1$

$$\mathrm{cov}[x]\, s_1 + \gamma/2\, s_0 = \lambda s_1$$

multiply from the left with $s_0^T$

$$\underbrace{\left(s_0^T \mathrm{cov}[x]\right)}_{\lambda_0 s_0^T} s_1 + \gamma/2\, \underbrace{s_0^T s_0}_{1}$$

$$= \lambda s_0^T s_1 \quad \Rightarrow$$

$$\gamma = 2[\lambda_1 - \lambda_0] \underbrace{s_0^T s_1}_{=0} \quad \Rightarrow \gamma = 0$$

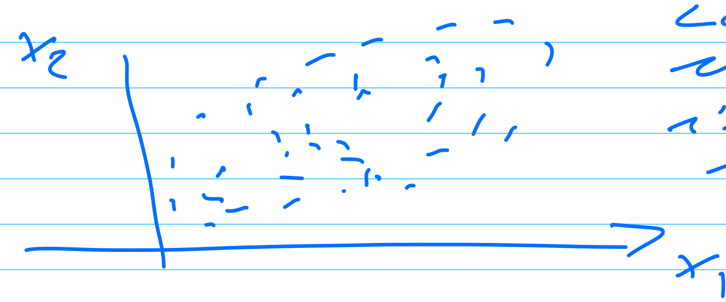
$$s_1^T \mathrm{cov}[x]\, s_1 = \lambda_1 = \mathrm{var}[y_1]$$

$$\mathrm{cov}[y] = S^T \mathrm{cov}[x]\, S$$

$s_i^T s_j = \delta_{ij}$, can construct by induction

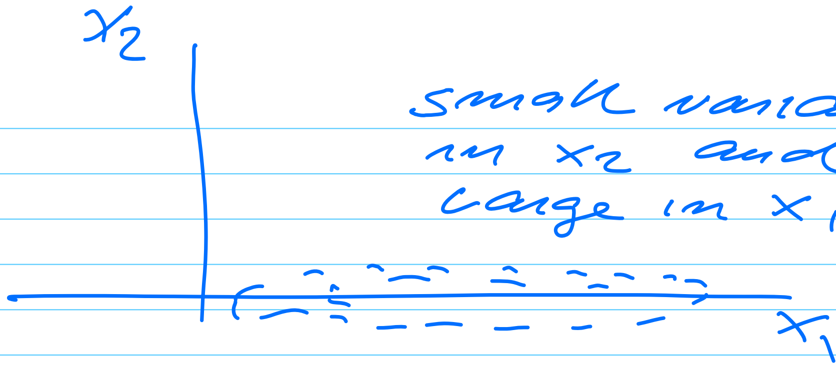
$$\mathcal{L} = s_i^T \mathrm{cov}[x]\, s_i + \lambda_i (1 - s_i^T s_i) + \sum_{j=0}^{i-1} \gamma_j\, s_i^T s_j$$

Two-dim set




Large variance in $x_1$ and in $x_2$

small variance in $x_2$ and large in $x_1$




could drop $x_2$ as a dof.

Reconstruction error

$$J(x) = \frac{1}{m} \sum_{i=0}^{m-1} \left( x_i - \tilde{x}_i \right)^2$$

$\tilde{x}_i$ is our approximation

For autoencoders, with linear dependence

$$\tilde{X} = VWx,$$

$$\tilde{J}(x) = \frac{1}{m} \sum_{i=0}^{m-1} \left( x_i - (VW)x_i \right)^2$$

we have an approximation

$$\underbrace{s_0 s_0^T}_{\text{outer product (matrix)}} x = \tilde{x}$$

$$s_0 = \begin{bmatrix} s_{00} \\ s_{10} \end{bmatrix}$$

$$s_0^T s_0 = s_{00}^2 + s_{10}^2$$

$$S_0 S_0^T = \begin{bmatrix} S_{00}^2 & S_{00} S_{10} \\ S_{10} S_{00} & S_{10}^2 \end{bmatrix}$$

The PCA theorem from the reconstruction error gives the smallest reconstruction error if $S_0$ are the eigen vectors of the $\mathrm{cov}[X]$ with largest variance $\sigma_0^2$

$$\underset{S_0}{\arg\min} \frac{1}{n} \| X - S_0 S_0^T X \|_2^2$$

Think of the autoencoder with linear dependence

$$\underset{V\,W}{\arg\min} \frac{1}{n} \| X - V W X \|_2^2$$

$$VW \propto S_0 S_0^T$$