Principal Component Analysis with SVD.

Recall from last week

$x_1, \dots, x_n \in \mathbb{R}^D$

$z_1, \dots, z_n \in \mathbb{R}^M$

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

$\phi(x_i) = z_i$ for $i = 1, \dots, n$

- empirical average

$$\bar{z} = \frac{1}{n}(z_1 + \dots + z_n)$$

- empirical covariance matrix

$$S = \frac{1}{n}\left(\Omega^T - \bar{z}e^T\right)\left(\Omega^T - \bar{z}e^T\right)^T = \frac{1}{n} W W^T.$$

$\Omega = [z_1, \dots, z_n]^T \in \mathbb{R}^{n \times M}$ [uncetered feature matrix] [W = centered feature matrix]

$\boxed{\text{Theorem}}$ If $\lambda_1 \geq \dots \geq \lambda_M \geq 0$ are the eigenvalues of S,

$u_1, \dots, u_M$ orthonormal basis of eigenvectors, then

$$U = \{u_1, \dots, u_d\}$$

minimizes squared distance $\displaystyle\sum_{i=1}^{n} \|(z_i - \bar{z}) - P_U(z_i - \bar{z})\|^2.$

$\boxed{\text{Def}}$ $u_1, \dots, u_d$ are called the principle components.

$\boxed{\text{The problem}}$ if M is very large, not practical to compute the eigenvalues.

$\boxed{\text{example}}$ in "Geometry of Images" section of NB6, $M > 3 \cdot 10^{12}$, but at most n nonzero eigenvalues

Assuming noisy data, $\text{rank}(W) = n$.

SVD of W:
$$W = U \, D \, V^T \in \mathbb{R}^{M \times n}$$

where $U \in \mathbb{R}^{M \times n}$ $\quad V \in \mathbb{R}^{n \times n}$ $\quad D = \text{diag}(\sigma_1, \dots, \sigma_n)$

where $\sigma_1 \geq \dots \geq \sigma_n \geq 0$ are the singular values.

Recall ($\sigma_i = \sqrt{\lambda_i}$ for the positive λ_i)

Then $S = \frac{1}{n}(UDV^T)(VDU^T) = \frac{1}{n} U D^2 U^T$

and the eigendecomposition is $\quad W^T W = \underline{V D^2 V^T} \in \mathbb{R}^{n \times n}$

eigendecomposition of $n \times n$ matrix is much smaller than $M \times M$ matrix!

Moreover if $U = [u_1 \cdots u_n]$
$V = [v_1, \dots v_n]$ then $u_i = \dfrac{W v_i}{\|W v_i\|}$

So computing eigendecomposition is sufficient to recover principal components.

$\boxed{\text{Lemma}}$ $R = W^T W \in \mathbb{R}^{n \times n}$ $\left\{ \begin{array}{l} \text{Symmetric + positive semidefinite} \\ \Rightarrow \text{eigenvalue} = \text{singular values.} \end{array} \right\}$

Then $$R = \left(\mathbb{1}_n - \frac{1}{n} e e^T\right) G \left(\mathbb{1}_n - \frac{1}{n} e e^T\right)$$

where $G = \Omega \Omega^T = \left(K(x_i, x_j)\right)_{1 \le i,j \le n}$

$\boxed{\text{Note}}$ We assume G full rank, and $\underbrace{\left(\mathbb{1}_n - \frac{1}{n} e e^T\right)\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} = \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} - \frac{1}{n}\begin{pmatrix}n\\ \vdots \\ n\end{pmatrix} = 0}_{\text{Rank} \le n-1}$

$\Rightarrow \text{rank}(R) \le n-1.$

Given $\lambda_1 \ge \cdots \ge \lambda_n \ge 0$ we need to pick d to find "size" of set for projection.

$\boxed{\text{ex}}$
- $\lambda_d > 0$, $\lambda_{d+1} \approx 0$
- maximize $\dfrac{\lambda_d}{\lambda_{d+1}}$
- maximize $\lambda_d - \lambda_{d+1}$
- maximize relative distance $\dfrac{\lambda_d - \lambda_{d+1}}{\lambda_d}$

$\boxed{\text{To do}}$ Notebook 6

- Points on a circle :
  - How is equation found?
  - what if you add a higher order polynomial (x_1^3)?

- Geometry of images :
  - What happens as we change k? what do we detect in eigenvalues?
  - What do you detect with other measurements than $\frac{\lambda_d - \lambda_{d+1}}{\lambda_d}$.

- Generative models
  - Follow code, what do we gain by adding the probabilistic component?

Probabilistic method:

$x_1, \dots, x_n$ chosen independently from random variable $x \in \mathbb{R}^D$.

Let $d \leq D$. Take Gaussian latent variable

$$\zeta \sim N(\nu, B) \quad \text{with} \quad \begin{matrix} \nu \in \mathbb{R}^d \\ B \in \mathbb{R}^{d \times d} \text{ positive definite} \end{matrix}$$

So

$$x = \underset{A \in \mathbb{R}^{D \times d}}{A}\, \zeta + \underset{b \in \mathbb{R}^D}{b} + \underset{\varepsilon \sim N(0,\, \sigma^2 \mathbb{I}_D)}{\varepsilon}$$

That is $(x \mid \zeta) \sim N(A\zeta + b,\ \sigma^2 \mathbb{I}_D)$

<u>Lemma</u> $x \sim N(A\nu + b,\ ABA^T + \sigma^2 \mathbb{I}_D)$

$\boxed{\text{Theorem}}$ Suppose we have a prior $\xi \sim N(\nu, B)$

Then the posterior distribution of ξ given x is

$$(\xi \mid x) \sim N(m, C)$$

with covariance matrix $C = \left(\sigma^{-2} A^T A + B^{-1}\right)^{-1}$

$$m = C\left(\sigma^{-2} A^T (x-b) + B^{-1}\nu\right)$$

$\boxed{\text{Pf}}$ By Bayes theorem

$$P(\xi \mid x) = P(x \mid \xi) \cdot \frac{P(\xi)}{P(x)}$$

$$\log(P(\xi \mid x)) = \log P(x \mid \xi) + \log(P(\xi)) - \underset{\substack{\uparrow \\ \text{independent of } \xi.}}{c}$$

We assumed

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

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

We've seen this before with MAP

See Theorem 3.17 of notes.

$\blacksquare$