Principle Components Analysis

(1) PCs of Population:

① Definition:

$X = (X_1 \cdots X_p)^T \quad E(X) = \mu . \quad Var(X) = \Sigma_{p \times p}$.

Consider use $Z_i = a_i^T X = \sum_{j=1}^{p} a_{ji} X_j$ linear combination of $\{X_i\}$ to replace $\{Z_i\}$.

$Var(Z_i) = a_i^T \Sigma a_i . \quad Cov(Z_i, Z_j) = a_i^T \Sigma a_j$

Note that: if $Var \uparrow$. then the data include more information.

Besides. we don't want to the information of $Z_1 \cdots Z_i$. overlaps. i.e. $Cov(Z_i, Z_j) = 0 . \ i \neq j$.

Def: $Z_i = a_i^T X$ is i^{th} component of X if.

i) $a_i^T a_i = 1 . \quad \forall \ 1 \le i \le p$.

ii) $a_i^T \Sigma a_j = 0 . \quad \forall \ i \neq j$.

iii) $Var(Z_i) = \max \{ Var(\alpha^T X) \mid \alpha^T \alpha = 1 . \ \alpha^T \Sigma a_j = 0$ for $\forall \ 1 \le j \le i-1 \}$

② Find Principle components:

i) For Z_1:

Note that $\frac{\alpha^T \Sigma \alpha}{\alpha^T \alpha} \le \lambda_1$. $\lambda_1 \ge \lambda_2 \cdots \lambda_p$ is eigenvalues of Σ. Therefore, the 1^{st} component is e_1 eigenfunc. correspond λ_1. s.t. $\|e_1\|_2^2 = 1$.

ii) <u>For Z_k:</u>

$Var(\alpha^T X) \le \lambda_k$. if $\alpha^T\alpha = 1$. $\alpha^T \Sigma a_i = 0$. $1 \le i \le k-1$.

Then $Z_k = e_k$. correspond eigenvalue λ_k. $\|e_k\|_2^2 = 1$.

<u>Thm.</u> $Z = (Z_1 \cdots Z_p)^T$ is principle components of X.

if i) $Z = A^T X$. A is orthonormal. $A = (a_1 \cdots a_p)$

ii) $Var(Z) = diag\{\lambda_1 \cdots \lambda_p\}$. $\lambda_1 \ge \lambda_2 \cdots \ge \lambda_p$.

<u>Rmk.</u> To find $\{Z_i\}_1^p$. Apply Σ by a orthonormal diagonalization.

③ <u>Properties:</u>

i) $\sum_1^p \sigma_{ii} = tr(\Sigma) = \sum_1^p \lambda_i$

<u>Rmk:</u> If $\exists m \in \mathbb{Z}^+$. s.t. $\sum_1^p \sigma_{ii} \approx \sum_1^m \lambda_i$. Then replace $\{X_i\}_1^p$ by $\{Z_i\}_1^m$ to reduce data

ii) $\rho(Z_k, X_i) =: Cor(Z_k, X_i) = \sqrt{\lambda_k}\, a_{ik} / \sqrt{\sigma_{ii}}$.

<u>Pf:</u> $\rho(Z_k, X_i) = \frac{Cov(a_k^T X, e_i^T X)}{\sqrt{Var(Z_k)\, Var(X_i)}} = \frac{\lambda_k a_{ik}}{\sqrt{\lambda_k \sigma_{ii}}}$

<u>Rmk:</u> We call $\rho(Z_k, X_i)$ is factor loading.

iii) $\sum_{k=1}^{p} \rho^2(Z_k, X_i) = \sum_{k}^{p} \frac{\lambda_k a_{ik}^2}{\sigma_{ii}} = 1$

<u>Pf</u>: $\Sigma = A \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{pmatrix} A^T \quad \sigma_{ii} = a_i \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{pmatrix} a_i^T$

<u>Rmk</u>: Sum of squre of factor loading on X_i is 1. (Full correlated)

iv) $\sum_{i=1}^{p} \sigma_{ii} \rho^2(Z_k, X_i) = \lambda_k$

<u>Pf</u>: $A^T \Sigma A = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{pmatrix}$. $\lambda_k = a_k^T \Sigma a_k$.

<u>Def</u>: i) $\lambda_k / \sum_{1}^{p} \lambda_i$ is rate of contribution of Z_k.

$\sum_{1}^{m} \lambda_k / \sum_{1}^{p} \lambda_i$ is accumulation of rate

ii) Contribution of $\{Z_k\}_1^m$ to X_i : V_i^m def by sum of squre of factor loading of Z_i.

i.e. $V_i^m = \sum_{k=1}^{m} \lambda_k a_{ik}^2 / \sigma_{ii}$

$\Rightarrow$ Consider two factors when select PCs

④ PCA of Standardilization:

To eliminate the influence of units, we can standardilization X. i.e. $X_i^* = \frac{X_i - E(X_i)}{\sqrt{\sigma_{ii}}}$

Then $Var(X^*) = R$. correlation matrix of X.

$\Rightarrow \sum_{1}^{p} Var(Z_i^*) = p = \sum_{1}^{p} \lambda_i^*$. $Z_i^* = a_i^T X^*$.

Rmk: If the Variance of X_i differs a lot

Then the direction of PCA will differ

a lot from PCA on Standardilization data.

(2) PCs of Samples:

When μ. Σ are unknown. We should infer from

the data matrix $X = (X_{ij})_{n\times p} = \begin{pmatrix} X_{(1)}^T \\ \vdots \\ X_{(n)}^T \end{pmatrix} = (X_1 \cdots X_p)$

① Common case:

To find principle components:

i) replace population dist. by empirical dist.

ii) replace Σ by $S = \frac{1}{n-1} \Sigma (X_{(i)} - \bar{X})^T (X_{(i)} - \bar{X})$

$\Rightarrow$ Find $(\hat{\lambda}_1, \hat{e}_1) \cdots (\hat{\lambda}_p, \hat{e}_p)$ eigenvalue-func pair

of S. $\hat{\lambda}_1 \geq \hat{\lambda}_2 \cdots \geq \hat{\lambda}_p$. Then i^{th} PC is:

$\hat{\eta}_i = \hat{e}_i^T X = \sum_{j=1}^{p} \hat{e}_{ij} X_j$. for $1 \leq i \leq p$.

We obtain:

i) $\sum_1^p S_{ii} = tr(S) = \sum_1^p \hat{\lambda}_i$

ii) $\ell(\hat{\eta}_i, X_k) = \hat{e}_{ik} \sqrt{\hat{\lambda}_i} / \sqrt{s_{kk}}$.

② Standardilization:

If the data matrix is observed after standardilization

Then $R = \frac{1}{n-1} X^T X$ is Correlation Sample matrix.

Find p pair eigenvalue-func (λ_i, a_i). $1 \le i \le p$. of R.

$\lambda_1 \ge \lambda_2 \cdots \ge \lambda_p$. $A = (a_1 \cdots a_p)$ is orthonormal.

<u>Def</u>: PC Score of i^{th} PC $a_i^T X$ at t^{th} Sample

is $a_i^T X_{(t)}$. Denote by z_{ti} (of $Z_i = a_i^T X$)

<u>properties</u>:

i) $\bar{Z} = \frac{1}{n} \sum_1^n Z_{(t)} = 0$. $Z_i^T Z_j = (n-1)\lambda_i \delta_{ij}$

where $Z = XA = \begin{pmatrix} Z_{(1)}^T \\ \vdots \\ Z_{(n)}^T \end{pmatrix} = (z_1 \cdots z_p)$.

<u>Pf</u>: $\begin{pmatrix} Z_{(1)}^T \\ \vdots \\ Z_{(n)}^T \end{pmatrix} = \begin{pmatrix} X_{(1)}^T \\ \vdots \\ X_{(n)}^T \end{pmatrix} (a_1 \cdots a_p)$

$= (X_{(i)}^T a_j)_{n \times p} = (a_j^T X_{(i)})_{n \times p}$

ii) Principle components can minimize SSE.

Consider linear model:
(Extract m PCs. Others as residue ε_i)

$$\begin{cases} X_1 = b_{11} z_1 + \cdots + b_{1m} z_m + \vec{\varepsilon}_1 \\ \vdots \\ X_p = b_{p1} z_1 + \cdots + b_{pm} z_m + \vec{\varepsilon}_p \end{cases}$$

$B = (b_{ij})_{p \times m}$ $\quad Z^* = (z_1 \cdots z_m)$. $X = Z^* B^T + E$.

LSE is $\hat{B}^T = ((Z^*)^T Z^*)^{-1} (Z^*)^T X$. $\quad A^* = (a_1 \cdots a_m)$

$= (A^{*T} X^T X A^*)^{-1} A^{*T} X^T X$

$= \mathrm{diag}[\hat{\lambda}_1^{-1} \cdots \hat{\lambda}_m^{-1}] (RA^*)^T$

$= (a_1 \cdots a_m)^T = A^{*T}$

$\Rightarrow Q(A^*) = \min Q(B)$

<u>Rmk</u>: See geometry of PC line in (3) ①.

③ Large Sample:

For $(\hat{\lambda}_k, \hat{e}_k, \hat{\eta}_k)_{k=1,2\cdots p}$ in ①.

Assume: i) $X_{(k)}$ is random sample from normal dist.

ii) eigenvalues of Σ satisfies: $\lambda_1 > \lambda_2 \cdots > \lambda_p > 0$

Then: $\hat{\lambda} = (\hat{\lambda}_1 \cdots \hat{\lambda}_k)$. $\sqrt{n}(\hat{\lambda} - \vec{\lambda}) \sim AN_p(0, 2\Lambda^2)$.

where $\Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{pmatrix}$. $\vec{\lambda} = (\lambda_1 \cdots \lambda_p)$.

Denote: $E_i = \lambda_i \sum_{k \neq i}^{p} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} e_k e_k^T$. e_k is eigenfunc. correspond to λ_k w.r.t. Σ.

Then $\sqrt{n}(\hat{e}_i - e_i) \sim AN_p(0, E_i)$

Rmk: i) $\hat{\lambda}_i$ is indept with $\hat{e}_i$.

ii) Obtain confidence interval with α level:

$$\lambda_i \in [\hat{\lambda}_i / (1 \pm z(\tfrac{\alpha}{2})\sqrt{2/n})].$$

(3) Application:

① Connection with SVD:

· Apply SVD on $X(I - P_n) = X - J\bar{X}^T_{n\times p} = U_{n\times p} L_{p\times p} V^T_{p\times p}$

where $U^T U = I_p$. $V^T V = I_p$ $L = \text{diag}\{l_1 \cdots l_p\}$

$\Rightarrow S = \frac{1}{n-1}(X - J\bar{X}^T)^T (X - J\bar{X}^T) = V(\frac{L^2}{n-1})V^T$

It means:

i) $C(V)$ is eigenfunction of S.

ii) $l_i^2/(n-1)$: eigenvalues of S.

iii) $Y = (X - J_n\bar{X}^T)V = UL$: PC score.

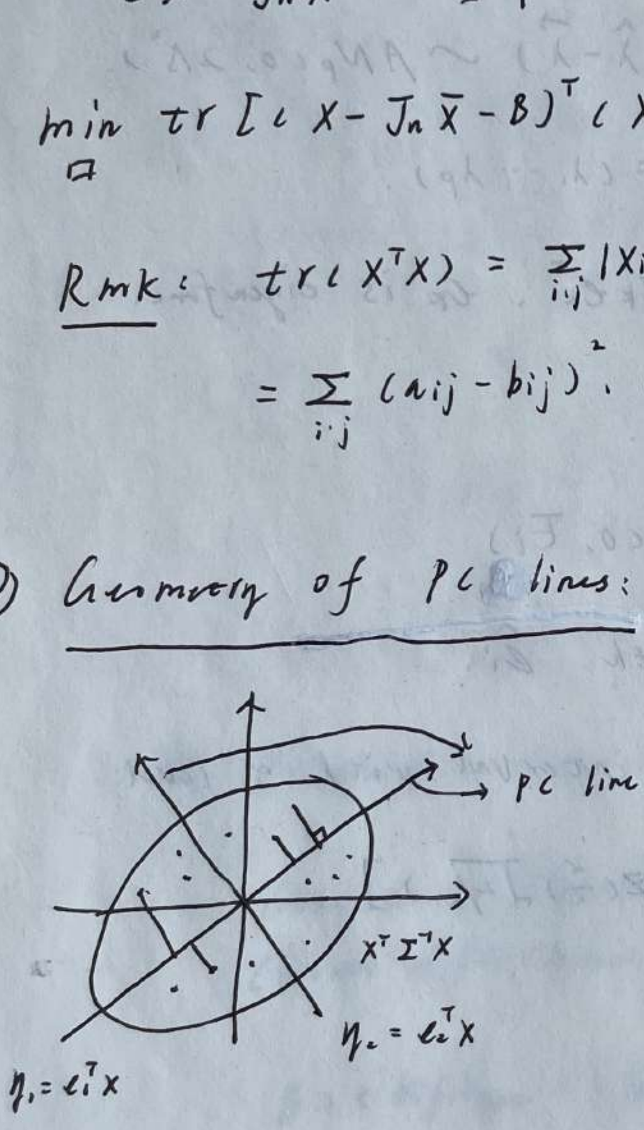
$$\tilde{B} = U\begin{pmatrix} l_1 & & & \\ & \ddots & & \\ & & l_k & \\ & & & O_{p-k}\end{pmatrix}V^T = \arg\min\{tr[(X - J_n\bar{X} - B)^T(X - J_n\bar{X} - B)] \mid r(B) \le k\}.$$

i.e.

$$\min_{B} tr[(X - J_n\bar{X} - B)^T(X - J_n\bar{X} - B)] = \sum_{k+1}^{p}\hat{\lambda}_i^2$$

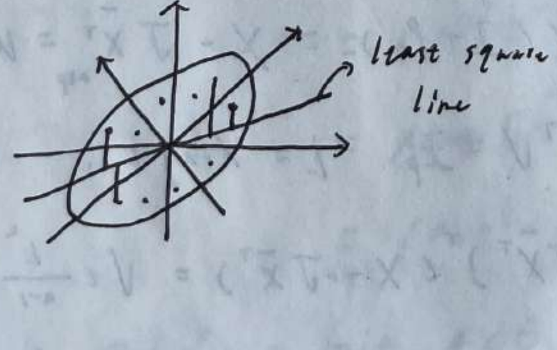
<u>Rmk</u>: $tr(X^TX) = \sum_{i,j}|X_{ij}|^2$. $tr((A-B)^T(A-B)) = \sum_{i,j}(a_{ij}-b_{ij})^2$. means SS of difference.

② <u>Geometry of PC lines:</u>



PC lines minimize the sum of squared orthogonal distances from each data to PC plane

<u>Rmk</u>: Compare to regression line: (least square line)



It minimizes the sum of vertical distance from data points to this line

③ Classification:

If X is standardilized. $R = \frac{1}{n-1} X^T X$. has eigenvalues $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p \geq 0$. PCs: $Z = (z_1 \cdots z_p)$

If we take the first m components $Z^* = (z_1 \cdots z_m)$

$$X^* = (X_1^* \cdots X_p^*) = (Z^* \; 0_{n \times (p-m)}) A = Z^* A^{*T}. \quad A = \begin{pmatrix} A^* \\ \tilde{A} \end{pmatrix}$$

prop. $tr((X-X^*)^T (X-X^*)) = (n-1) \sum_{m+1}^{p} \lambda_k$.

Pf: $X - X^* = X(I - A^* A^{*T})$. $A^* = \begin{pmatrix} a_1^T \\ \vdots \\ a_m^T \end{pmatrix}$

$$\Rightarrow (X-X^*)^T (X-X^*) = (n-1)(I - A^* A^{*T}) R (I - A^* A^{*T})$$

$$= (n-1) \left(\sum_{m+1}^{p} \lambda_i a_i a_i^T \right)$$

i) If $r_{ij} \simeq 1$. Then: X_i, X_j can be classified as one class.

$\| X_i - X_j \|_2^2 = 2(n-1)(1 - r_{ij}) \simeq \| X_i^* - X_j^* \|_2^2$. For $1 - r_{ij}$:

$$1 - r_{ij} = \frac{1}{2(n-1)} \left\| \sum_{1}^{m} (a_{ik} - a_{jk}) Z_k \right\|_2^2 = \frac{1}{2(n-1)} \sum_{k=1}^{m} \lambda_k (a_{jk} - a_{ik})^2$$

$$= \frac{1}{2(n-1)} \sum_{1}^{m} (\ell_{ik} - \ell_{jk})^2. \quad \ell_{jk} = \ell(Z_k, X_i) = \sqrt{\lambda_k}\, a_{ik}$$

prop. If $\sum_{1}^{m} (\ell_{ik} - \ell_{jk})^2 \simeq 0$. Then classify X_i, X_j as the same class

ii) Analogously. $\| X_{(i)} - X_{(j)} \|_2^2 \simeq 0 \Rightarrow X_{(i)}, X_{(j)}$ can be recognized as samples from same class

$$\| X_{(i)} - X_{(j)} \|_2^2 \simeq \| X_{(i)}^* - X_{(j)}^* \|_2^2 \quad (X_{(i)}^* = A^* Z_i^{*T})$$

$$= \sum_{k=1}^{m} (z_{ik} - z_{jk})^2$$

④ Drawback:

PCA doesn't use information of k^{th} moments $(k \geq 2)$

It may miss nonlinear structure or distort by outlier.