Principal component analysis (PCA)

Multivariate observations $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \cdots, \underset{\sim}{x}_n$

– n observations - p variables

Goals: Try to represent/approximate data by $r < p$ ($r << p$) dimensions, $r = 1, 2, 3$ optimal.

Earlier: Try to find "interesting" projections of the data → projection pursuit.

i.e. find $\underset{\sim}{a}$ s.t. $\underset{\sim}{a}^T \underset{\sim}{x}_1, \cdots, \underset{\sim}{a}^T \underset{\sim}{x}_n$ look ~~look~~ non-normal (e.g. maximize kurtosis)

Find structure (outliers, clusters etc)

Trying to find $\underset{\sim}{y}_1, \cdots, \underset{\sim}{y}_n$ (← r vectors) based on $\underset{\sim}{x}_1, \cdots, \underset{\sim}{x}_n$ having similar structure to $\underset{\sim}{x}_1, \cdots, \underset{\sim}{x}_n$

Example: Distance measure $d(\underset{\sim}{x}_i, \underset{\sim}{x}_j)$ { Euclidean / Manhattan } distances

Try to find $\underset{\sim}{y}_1, \cdots, \underset{\sim}{y}_n$ s.t. distance structure is approximately maintained

$$d(\underset{\sim}{y}_i, \underset{\sim}{y}_j) \doteq d(\underset{\sim}{x}_i, \underset{\sim}{x}_j)$$

PCA: Two approaches

① Low rank matrix approximation

② Projection pursuit → maximize variance

Low rank matrix approximation

– assume $p < n$ and define

$$X = \begin{pmatrix} \underset{\sim}{x}_1^T \\ \vdots \\ \underset{\sim}{x}_n^T \end{pmatrix} \qquad \text{rank}(X) = p$$

(X is $n \times p$)

– can we approximate X by a matrix with rank $r < p$ (and how)?

– find X^* with rank to minimize $\|X - M\|_F^2 = \sum_{i=1}^n \sum_{j=1}^p |X_{ij} - M_{ij}|^2$

($\|\cdot\|_F$: Frobenius norm)

Singular value decomposition (SVD)

Idea: Write $X = UDV^T$ (U: $n \times p$, D: diagonal $p \times p$, V: $p \times p$) Where U & V have orthonormal columns

$$U = (\underset{\sim}{u}_1, \cdots, \underset{\sim}{u}_p), \quad V = (\underset{\sim}{v}_1, \cdots, \underset{\sim}{v}_p) \quad \text{with } \underset{\sim}{u}_i^T \underset{\sim}{u}_j = \underset{\sim}{v}_i^T \underset{\sim}{v}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

$$D = \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_p \end{pmatrix} \qquad \underbrace{d_1 \geq d_2 \geq \cdots \geq d_p > 0}_{\text{singular values of } X}$$

Now Diagonal $D^* = \begin{pmatrix} d_1 & & & & \\ & \ddots & & & \\ & & d_r & & \\ & & & 0 & \\ & & & & \ddots \ 0 \end{pmatrix}$ and define $X^* = UD^*V^T$

$$\begin{aligned} \text{rank}(X^*) &= \min(\text{rank}(U), \text{rank}(D^*), \text{rank}(V)) \\ &= \text{rank}(D^*) \\ &= r \end{aligned}$$

X^* minimizes $\|X - M\|_F^2$ over M with $\text{rank}(M) = r$

– approximation is good if d_{r+1}, d_p close to 0.

– $X^*V = UD^*V^TV = UD^* = \begin{pmatrix} \underset{\sim}{u}_1 & \cdots & \underset{\sim}{u}_p \end{pmatrix} \begin{pmatrix} d_1 & & & 0 \\ & \ddots \ d_r & & \\ & & 0 & \\ 0 & & & \ddots \ 0 \end{pmatrix} = (d_1 \underset{\sim}{u}_1 \cdots d_r \underset{\sim}{u}_r \ 0 \cdots 0)$

- R function: svd

Example: Exam mark data

- look at centred & scaled data

$d_1 = 16.64$, $d_3 = 6.32$, $d_5 = 4.63$

$d_2 = 8.02$, $d_4 = 5.81$

$$V = \begin{pmatrix} -0.40 & -0.65 \\ -0.43 & -0.44 \\ -0.50 & 0.13 \\ -0.46 & 0.39 \\ -0.44 & 0.47 \end{pmatrix}$$

overall strength

comparison of (MECH, VECT) to (ALG, ANA, STAT)

From SVD to PCA

Define $\tilde{X}$ to be the $n \times p$ matrix of centred (and possibly scaled) data.

SVD: $\tilde{X} = UDV^T$

$$S = \frac{1}{n-1}\tilde{X}^T\tilde{X} = \frac{1}{n-1}VDD^TU^TDV^T$$

$$= V\left(\frac{D^2}{n-1}\right)V^T$$

sample covariance matrix (correlation matrix if vars scaled)

- Columns of V are the eigenvectors $\underset{\sim}{v}_1, \cdots, \underset{\sim}{v}_p$ of S with eigenvalues the diagonal elements of $\frac{D^2}{n-1} \Rightarrow \frac{d_1^2}{n-1}, \cdots \frac{d_r^2}{n-1}$

Principal components:

Define $n \times p$ matrix

$$y = \tilde{X}V = \tilde{X}(\underset{\sim}{v}_1, \cdots, \underset{\sim}{v}_p) = (\tilde{X}\underset{\sim}{v}_1, \cdots, \tilde{X}\underset{\sim}{v}_p)$$

p new variables → PC scores

Projection pursuit approach

Data $\underset{\sim}{x}_1, \cdots, \underset{\sim}{x}_n$ with sample covariance S or correlation matrix R

- find vector $\underset{\sim}{a}$ with $\|\underset{\sim}{a}\|^2 = \underset{\sim}{a}^T\underset{\sim}{a} = 1$ maximizing sample variance of $\{\underset{\sim}{a}^T\underset{\sim}{x}_i\}$

$$\frac{1}{n-1}\sum_{i=1}^{n}(\underset{\sim}{a}^T\underset{\sim}{x}_i - \underset{\sim}{a}^T\bar{\underset{\sim}{x}})^2 = \underset{\sim}{a}^TS\underset{\sim}{a}$$

$S = V\Lambda V^T \left(\Lambda = \frac{D^2}{n-1}\right)$

$\underset{\sim}{a}^TS\underset{\sim}{a} = \underset{\sim}{a}^TV\Lambda V^T\underset{\sim}{a} = \sum_{k=1}^{p}\lambda_k(\underset{\sim}{v}_k^T\underset{\sim}{a})^2$ $\qquad \Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}$

$\Rightarrow$ maximized at $\underset{\sim}{a} = \underset{\sim}{v}_1$

Now maximize $\underset{\sim}{a}^TS\underset{\sim}{a}$ s.t. $\|\underset{\sim}{a}\|^2 = 1$ and $\underset{\sim}{a}^T\underset{\sim}{v}_1 = 0 \Rightarrow$ maximized at $\underset{\sim}{a} = \underset{\sim}{v}_2$

Maximize $\underset{\sim}{a}^TS\underset{\sim}{a}$ s.t. $\|\underset{\sim}{a}\|^2 = 1$ and $\underset{\sim}{a}^T\underset{\sim}{v}_1 = \cdots = \underset{\sim}{a}^T\underset{\sim}{v}_{p-1} = 0$

$\Rightarrow$ maximized at $\underset{\sim}{a} = \underset{\sim}{v}_k$

Also note that $\frac{1}{n-1}\sum_{i=1}^{n}(\underset{\sim}{v}_k^T\underset{\sim}{x}_i - \underset{\sim}{v}_k^T\bar{\underset{\sim}{x}})^2 = \underset{\sim}{v}_k^TS\underset{\sim}{v}_k = \lambda_k$

Example: $p = 2$, $R = \begin{pmatrix} 1 & \hat{\rho} \\ \hat{\rho} & 1 \end{pmatrix}$:

eigenvalues

solve $\det(R-\lambda I)=0$

$\Rightarrow \lambda^2-2\lambda+1-\hat{\rho}^2=0$

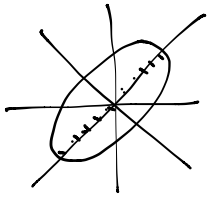
$\Rightarrow \lambda = 1 \pm |\hat{\rho}|$

$\lambda_1 = 1+|\hat{\rho}|$

$\lambda_2 = 1-|\hat{\rho}|$

eigenvectors (loadings) $\underset{\sim}{v_1} = \begin{pmatrix} 1/\sqrt{2} \\ \text{sgn}(\hat{\rho})\frac{1}{\sqrt{2}} \end{pmatrix}$ $\underset{\sim}{v_2} = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\text{sgn}(\hat{\rho})\frac{1}{\sqrt{2}} \end{pmatrix}$

PCs with $p=2$ using R = sums and differences between 2 variables



$\hat{\rho} > 0$

What happens if $\hat{\rho} = 0$? $R = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I$

– any two orthogonal vectors $\underset{\sim}{v_1}, \underset{\sim}{v_2}$ will define PC loadings.

How to use PCs?

① R vs S ← centred

↑ centred, scaled

Generally: R is the better choice.

– PCs are not invariant to scaling.
– If S is used in PCA, variables with high variance will tend to dominate.

– If all variables are measures of size (i.e. length, volume, mass) then taking logs and using S is often OK.

② What PCs to look at? ← corresponding to largest eigenvalues or or singular values
– Typically, first few PCs contain most information
– Often reveal interesting structure ⇒ projection pursuit

③ Plot of 1st 2 PC scores is particularly useful.
– biplot (R function: biplot): scatterplot of 1st 2 PC scores + information on original variables.

④ PCs with smallest variance can be useful also.
– reveals relationship between variables or a subset of variables.

Examples: Athlete records (Blackboard)
55 countries: USA, CAN, etc Cook Islands, Western Samoa
8 men's national records: 100m, 200m, 400m, 800m, 1500m, 5000m, 10000m, marathon
– from 1980s
– do PCA on correlation matrix
princomp(··· , cor = T)
– correlations highest for similar events

$\lambda_1, \cdots, \lambda_8$ = variances of PCs

$\lambda_1 + \cdots + \lambda_8 = 8$

$\frac{\lambda_1}{8} = 0.79$, $\frac{\lambda_1+\lambda_2}{8} = 0.90$, $\frac{\lambda_1+\lambda_2+\lambda_3}{8} = 0.94$,

Loadings:	PC#1	PC#2
100m	-0.33	-0.54
200m	-0.34	-0.47
400m	-0.36	-0.25
800m	-0.38	0.00
1500m	-0.39	0.13
5000m	-0.37	0.31
10000m	-0.33	0.35
marathon	-0.33	0.47

Interpretation?

- rank countries based on PC scores