# Chapter 8 - Principal Components

Goal: Describe variance-covariance structure of set of variables through linear combinations of the variables. 1) Data Reduction, 2) Interpretation.

$n$ measurements on $p$ variables $\Rightarrow$ $n$ measurements on $k < p$ components.

## 8.2 Population Principal Components

$p$ random variables: $X_1, \ldots, X_p$

r.v: $\underset{\sim}{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$ w/ covariance matrix $\Sigma$

w/ eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$

Linear Combinations:

$$Y_1 = \underset{\sim}{a}_1' \underset{\sim}{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

$$Y_2 = \underset{\sim}{a}_2' \underset{\sim}{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$

$$\vdots$$

$$Y_p = \underset{\sim}{a}_p' \underset{\sim}{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p$$

$$V\{Y_i\} = \underset{\sim}{a}_i' V\{X\} \underset{\sim}{a}_i = \underset{\sim}{a}_i' \Sigma \underset{\sim}{a}_i \qquad i = 1, \ldots, p$$

$$Cov\{Y_i, Y_k\} = \underset{\sim}{a}_i' \Sigma \underset{\sim}{a}_k \qquad i, k = 1, \ldots, p$$

- First principal component is the linear combination $\underset{\sim}{a}_1' \underset{\sim}{X}$ that maximizes $V\{a_1'X\}$ s.t. $a_1'a_1 = 1$
- Second principal component $= a_2'X$ that maximizes $V\{\underset{\sim}{a}_2' \underset{\sim}{X}\}$ s.t. $\underset{\sim}{a}_2' \underset{\sim}{a}_2 = 1$ $Cov\{\underset{\sim}{a}_1' \underset{\sim}{X}, \underset{\sim}{a}_2' \underset{\sim}{X}\} = 0$ . . .

**Result 8.1** Random vector w/ $\underset{\sim}{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$ and $V\{\underset{\sim}{X}\} = \Sigma$

$\Sigma$ has eigenvalue-eigenvector pairs:

$$(\lambda_1, \underset{\sim}{e}_1), (\lambda_2, \underset{\sim}{e}_2), \ldots, (\lambda_p, \underset{\sim}{e}_p) \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

Then the $i^{th}$ Principal component is given as:

$$Y_i = \underset{\sim}{e}_i' \underset{\sim}{X} = e_{i1} X_1 + \cdots + e_{ip} X_p \quad i = 1, \ldots, p$$

with $V\{Y_i\} = \underset{\sim}{e}_i' \Sigma \underset{\sim}{e}_i = \lambda_i \quad i = 1, \ldots, p$

$$Cov\{Y_i, Y_k\} = \underset{\sim}{e}_i' \Sigma \underset{\sim}{e}_k = 0 \quad i \neq k$$

**Proof:** Equation (2-51): $B = \Sigma$ (Normalized $\Rightarrow$)

$$\max_{\underset{\sim}{a} \neq \underset{\sim}{0}} \frac{\underset{\sim}{a}' \Sigma \underset{\sim}{a}}{\underset{\sim}{a}' \underset{\sim}{a}} = \lambda_1 \quad \text{attained @ } \underset{\sim}{a} = \underset{\sim}{e}_1 \quad (\underset{\sim}{e}_1' \underset{\sim}{e}_1 = 1)$$

$$\Rightarrow \max_{\underset{\sim}{a} \neq \underset{\sim}{0}} \frac{\underset{\sim}{a}' \Sigma \underset{\sim}{a}}{\underset{\sim}{a}' \underset{\sim}{a}} = \lambda_1 = \frac{\underset{\sim}{e}_1' \Sigma \underset{\sim}{e}_1}{\underset{\sim}{e}_1' \underset{\sim}{e}_1} = \underset{\sim}{e}_1' \Sigma \underset{\sim}{e}_1 = V\{Y_1\}$$

Equation (2-52):

$$\max_{\underset{\sim}{a} \perp \underset{\sim}{e}_1, \ldots, \underset{\sim}{e}_k} \frac{\underset{\sim}{a}' \Sigma \underset{\sim}{a}}{\underset{\sim}{a}' \underset{\sim}{a}} = \lambda_{k+1} \quad @ \ a = \underset{\sim}{e}_{k+1} \quad k = 1, \ldots, p-1$$

$$\underset{\sim}{e}_{k+1}' \underset{\sim}{e}_i = 0 \quad i = 1, \ldots, k; \ k = 1, \ldots, p-1$$

$$\frac{\underset{\sim}{e}_{k+1}' \Sigma \underset{\sim}{e}_{k+1}}{\underset{\sim}{e}_{k+1}' \underset{\sim}{e}_{k+1}} = \underset{\sim}{e}_{k+1}' \Sigma \underset{\sim}{e}_{k+1} = V\{Y_{k+1}\}$$

where: $\underset{\sim}{e}_{k+1}' (\Sigma \underset{\sim}{e}_{k+1}) = \lambda_{k+1} \underset{\sim}{e}_{k+1}' \underset{\sim}{e}_{k+1} = \lambda_{k+1} (1) = V\{Y_{k+1}\}$

- If all eigenvalues are distinct $\Rightarrow$ all eigenvectors orthogonal
- If eigenvalues are not all distinct, eigenvectors corresponding to common eigenvalues can be chosen to be orthogonal.

$\Rightarrow$ $\underset{\sim}{e}_i' \underset{\sim}{e}_k = 0 \quad i \neq k$

$$Cov\{Y_i, Y_k\} = \underset{\sim}{e}_i' \underset{\sim}{\Sigma} \underset{\sim}{e}_k = \underset{\sim}{e}_i' \lambda_k \underset{\sim}{e}_k = \lambda_k \underset{\sim}{e}_i' \underset{\sim}{e}_k = 0 \quad i \neq k$$

**Result 8.2** R.V. $\underset{\sim}{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$ w/ $V\{X\} = \Sigma$ w/

eigenvalue/eigenvector pairs $(\lambda_1, \underset{\sim}{e}_1), (\lambda_2, \underset{\sim}{e}_2), \ldots, (\lambda_p, \underset{\sim}{e}_p)$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

$Y_1 = \underset{\sim}{e}_1' \underset{\sim}{X}, \ldots, Y_p = \underset{\sim}{e}_p' \underset{\sim}{X}$ $\equiv$ principal components

Then: $\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} V\{X_i\} = \lambda_1 + \cdots + \lambda_p = \sum_{i=1}^{p} V\{Y_i\}$

**Proof:** $\sigma_{11} + \cdots + \sigma_{pp} = tr(\Sigma)$ w/ $\Sigma = P \Delta P'$

$P = \left[ \underset{\sim}{e}_1 \cdots \underset{\sim}{e}_p \right]$ $\Delta = diag\{\lambda_i\}$ $PP' = P'P = I$

$$tr(\Sigma) = tr(P \Delta P') = tr(P'P\Delta) = tr(\Delta) = \sum_{i=1}^{p} V\{Y_i\}$$

Proportion of total population variance due to $k^{th}$ Prin. Comp.:

$$\frac{\lambda_k}{\lambda_1 + \cdots + \lambda_p} \qquad k = 1, \ldots, p$$

If most of variation for large $p$ can be explained by a few principal components, data can be reduced w/out major loss of information.

**Result 8.3** $Y_1 = \underset{\sim}{e}_1'\underset{\sim}{X}, \ldots, Y_p = \underset{\sim}{e}_p'\underset{\sim}{X}$ PC's from $\Sigma$

Then $\rho_{Y_i, X_k} = \dfrac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \quad i, k = 1, 2, \ldots, p$

**Proof:** Let $\underset{\sim}{a}_k' = [0 \ldots 0\ 1\ 0 \ldots 0] \Rightarrow X_k = \underset{\sim}{a}_k'\underset{\sim}{X}$

$\mathrm{Cov}\{\underset{\sim}{a}_k'\underset{\sim}{X}, \underset{\sim}{e}_i'\underset{\sim}{X}\} = \underset{\sim}{a}_k'\Sigma\underset{\sim}{e}_i \qquad \Sigma\underset{\sim}{e}_i = \lambda_i\underset{\sim}{e}_i$

$\Rightarrow \mathrm{Cov}\{X_k, Y_i\} = \underset{\sim}{a}_k'\lambda_i\underset{\sim}{e}_i = \lambda_i e_{ik}$

$V\{Y_i\} = \lambda_i \quad V\{X_k\} = \sigma_{kk} \Rightarrow \mathrm{Corr}\{X_k, Y_i\} = \dfrac{\lambda_i e_{ik}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = \dfrac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$

**EXAMPLE - LPGA DATA** ($N = 156$, $p = 4$ (Drive, Front, Greens, putted))

$$\Sigma = \begin{bmatrix} 86.23 & -25.41 & 15.60 & -1.72 \\ & 31.16 & 4.60 & 1.12 \\ & & 12.95 & 0.20 \\ & & & 1.06 \end{bmatrix}$$

$\mathrm{trace}(\Sigma) = 131.40$

$\lambda_1 = 98.18 \quad \lambda_3 = 5.02 \quad \sum_{i=1}^{4}\lambda_i = 131.41$

$\lambda_2 = 27.21 \quad \lambda_4 = 1.00$

$$\underset{\sim}{e}_1 = \begin{bmatrix} 0.927 \\ -0.342 \\ 0.151 \\ -0.020 \end{bmatrix} \quad \underset{\sim}{e}_2 = \begin{bmatrix} -0.223 \\ -0.829 \\ -0.512 \\ -0.025 \end{bmatrix} \quad \underset{\sim}{e}_3 = \begin{bmatrix} -0.299 \\ -0.442 \\ 0.844 \\ 0.047 \end{bmatrix} \quad \underset{\sim}{e}_4 = \begin{bmatrix} 0.027 \\ -0.007 \\ 0.049 \\ 0.998 \end{bmatrix}$$

- Component 1: Driving Distance (Note that $\sigma_{11}$ is large)
- Component 2: Fairway Pct / Greens in Reg
- Component 3: Greens in Reg.
- Component 4: Putting.

$$\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3+\lambda_4} = \frac{98.18}{131.41} = .747$$

$$\frac{\lambda_1+\lambda_2}{131.41} = \frac{98.18+27.21}{131.41} = .954$$

First 2 components could replace 4 vars w/ little loss of information.

$$\rho_{Y_1,X_1} = \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}} = \frac{0.927\sqrt{98.18}}{\sqrt{86.23}} = .989$$

$$\rho_{Y_1,X_2} = \frac{e_{12}\sqrt{\lambda_1}}{\sqrt{\sigma_{22}}} = \frac{-.342\sqrt{98.18}}{\sqrt{31.16}} = -.607$$

# Case where $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$

Density of $X$ constant on $\underset{\sim}{\mu}$ centred ellipsoids:

$$(\underset{\sim}{x}-\underset{\sim}{\mu})'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}) = c^2 \quad \text{w/ axes } \pm c\sqrt{\lambda_i}\,\underset{\sim}{e}_i \quad i=1,\ldots,p$$

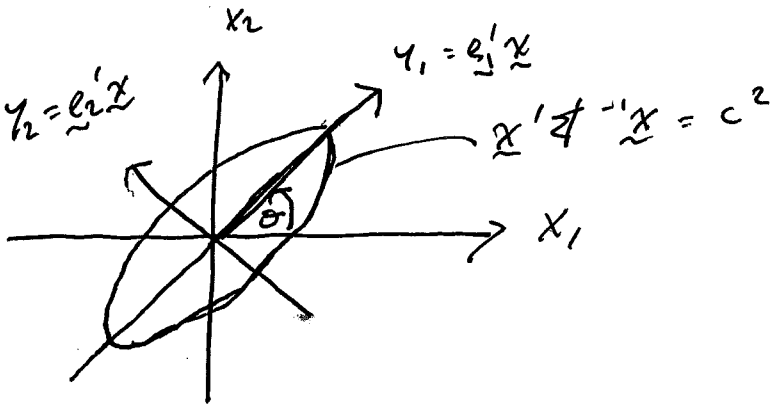
Setting $\underset{\sim}{\mu}=0$ $(\underset{\sim}{w}=\underset{\sim}{x}-\underset{\sim}{\mu}) \Rightarrow E\{w\}=0$, $Cov\{\underset{\sim}{w}\}=Cov\{\underset{\sim}{x}\}$

$$c^2 = \underset{\sim}{x}'\Sigma^{-1}\underset{\sim}{x} = \frac{1}{\lambda_1}(\underset{\sim}{e}_1'\underset{\sim}{x})^2 + \cdots + \frac{1}{\lambda_p}(\underset{\sim}{e}_p'\underset{\sim}{x})^2$$

Setting $y_i = \underset{\sim}{e}_i'\underset{\sim}{x} \Rightarrow c^2 = \frac{1}{\lambda_1}y_1^2 + \cdots + \frac{1}{\lambda_p}y_p^2$

$(\lambda_1,\ldots,\lambda_p > 0)$

$y_1 = \underset{\sim}{e}_1'\underset{\sim}{x}, \ldots, y_p = \underset{\sim}{e}_p'\underset{\sim}{x}$ lie in directions of axes of a constant density ellipsoid.



# Principal Components from Standardized Variables

$$Z_1 = \frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}}, \quad \dots, \quad Z_p = \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}}$$

$$\underset{\sim}{Z} = \left(V^{1/2}\right)^{-1}\left(\underset{\sim}{X} - \underset{\sim}{\mu}\right) \qquad V^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & & 0 \\ & \sqrt{\sigma_{22}} & \\ & & \ddots \\ 0 & & \sqrt{\sigma_{pp}} \end{bmatrix}$$

$$E\{\underset{\sim}{Z}\} = \left(V^{1/2}\right)^{-1}\left\{\underset{\sim}{\mu} - \underset{\sim}{\mu}\right\} = 0$$

$$V\{\underset{\sim}{Z}\} = \left(V^{1/2}\right)^{-1} \Sigma \left(V^{1/2}\right)^{-1} = \rho$$

**Result 8.4** $i^{th}$ principal component of standardized variables $Z' = [Z_1 \cdots Z_p]$ w/ $\mathrm{Cov}\{Z\} = \rho$ is:

$$Y_i = \underset{\sim}{e}_i' \underset{\sim}{Z} = \underset{\sim}{e}_i' \left(V^{1/2}\right)^{-1}\left(\underset{\sim}{X} - \underset{\sim}{\mu}\right) \qquad i = 1, \dots, p$$

$$\sum_{i=1}^{p} V\{Y_i\} = \sum_{i=1}^{p} V\{Z_i\} = p \qquad \rho_{Y_i, Z_k} = e_{ik}\sqrt{\lambda_i} \qquad i, k = 1, \dots, p$$

$(\lambda_i, \underset{\sim}{e}_i) \equiv$ eigenvalue eigenvector pairs of $\rho$

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$$

Proportion of (standardized) population variance due to the $k^{th}$ component $\equiv \dfrac{\lambda_k}{p} \qquad k = 1, \dots, p$

# EXAMPLE – LPGA DATA

$$\rho = \begin{bmatrix} 1.000 & -.490 & .467 & -.179 \\ & 1.000 & .229 & .194 \\ & & 1.000 & .053 \\ & & & 1.000 \end{bmatrix}$$

$\lambda_1 = 1.647 \quad \lambda_2 = 1.292 \quad \lambda_3 = 0.864 \quad \lambda_4 = 0.197$

$$\underset{\sim}{e}_1 = \begin{bmatrix} .715 \\ -.537 \\ .298 \\ -.334 \end{bmatrix} \quad \underset{\sim}{e}_2 = \begin{bmatrix} -.205 \\ -.491 \\ -.775 \\ -.342 \end{bmatrix} \quad \underset{\sim}{e}_3 = \begin{bmatrix} -.156 \\ .427 \\ .157 \\ -.877 \end{bmatrix} \quad \underset{\sim}{e}_4 = \begin{bmatrix} -.650 \\ -.537 \\ .535 \\ -.050 \end{bmatrix}$$

| # components | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Proportion of variation | $\frac{1.647}{4} = .412$ | $\frac{2.939}{4} = .735$ | $\frac{3.803}{4} = .951$ | 1 |

$\rho_{Y_1,z_1} = e_{11}\sqrt{\lambda_1} = .715\sqrt{1.647} = .918 \quad \rho_{Y_1,z_2} = e_{12}\sqrt{\lambda_1}$

$= -.537\sqrt{1.647} = -.689$

## Principal Components for Special Covariance Matrices

1) $\Sigma = \begin{bmatrix} \sigma_{11} & & 0 \\ & \sigma_{22} & \\ 0 & & \sigma_{pp} \end{bmatrix} \quad \underset{\sim}{e}_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

$\Sigma \underset{\sim}{e}_i = \sigma_{ii} \underset{\sim}{e}_i \Rightarrow (\sigma_{ii}, \underset{\sim}{e}_i) \equiv$ eigenvalue/eigenvector pair

$\rho = I \quad \rho \underset{\sim}{e}_i = \underset{\sim}{e}_i \quad (1, \underset{\sim}{e}_i) \equiv$

2) Compound Symmetry $\Sigma = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \ddots \\ & \sigma^2 \end{bmatrix}$ $R = \begin{bmatrix} 1 & \rho \\ \rho & \ddots \\ & 1 \end{bmatrix}$

When $\rho > 0 \Rightarrow$

$\lambda_1 = 1 + (p-1)\rho$ $\quad e_1' = \left[\frac{1}{\sqrt{p}}, \frac{1}{\sqrt{p}}, \ldots, \frac{1}{\sqrt{p}}\right]$

$\lambda_2 = \lambda_3 = \cdots = \lambda_p = 1 - \rho$

$\underset{\sim}{e}_2' = \left[\frac{1}{\sqrt{1\times 2}} \quad \frac{-1}{\sqrt{1\times 2}} \quad 0 \cdots 0\right]$

$\underset{\sim}{e}_3' = \left[\frac{1}{\sqrt{2\times 3}} \quad \frac{1}{\sqrt{2\times 3}} \quad \frac{-2}{\sqrt{2\times 3}} \quad 0 \cdots 0\right]$

$\underset{\sim}{e}_i' = \left[\frac{1}{\sqrt{(i-1)i}} \cdots \frac{1}{\sqrt{(i-1)i}} \quad \frac{-(i-1)}{\sqrt{(i-1)i}} \quad 0 \cdots 0\right]$

$\underset{\sim}{e}_p' = \left[\frac{1}{\sqrt{(p-1)p}} \cdots \frac{1}{\sqrt{(p-1)p}} \quad \frac{-(p-1)}{\sqrt{(p-1)p}}\right]$

$Y_1 = \underset{\sim}{e}_1' \underset{\sim}{z} = \frac{1}{\sqrt{p}} \sum_{i=1}^{n} z_i$ "equal weights" - index.

$\frac{\lambda_1}{p} = \frac{1 + (p-1)\rho}{p} = \rho + \frac{1-\rho}{p}$

$z_i \sim N_p(0, \rho) \Rightarrow$ ellipsoids of Constant density are cigar shaped w/ 1st principal Component $Y_1 = \frac{1}{\sqrt{p}} \underset{\sim}{1}' \underset{\sim}{z}$ (Projection of $\underset{\sim}{z}$ onto equiangular line $1' = [1 \ldots 1]$ Minor axes occur in spherically symmetric directions perpendicular to major axis.

## 8.3 Summarizing Sample Variation by Principal Components

Observe $X_1, \ldots, X_n \overset{iid}{\sim} (\mu, \Sigma) \Rightarrow \underset{\sim}{\bar{X}}, S, R$

$a_i' \underset{\sim}{X} = a_{i1} X_{j1} + \cdots + a_{ip} X_{jp} \quad j = 1, \ldots, n$

$\Rightarrow$ sample mean $a_i' \bar{X}$, sample variance $a_i' S a_i$

$Cov\{a_1' X_j, a_2' X_j\} = \underset{\sim}{a_i'} S \underset{\sim}{a_2}$

Sample Principal components and variances obtained in same manner as population, based on $S$, not $\Sigma$

$S = \{s_{ik}\} \equiv$ $p \times p$ Sample variance-covariance matrix w/ eigenvalue-eigenvector pairs $(\hat{\lambda}_1, \underset{\sim}{\hat{e}_1}), \ldots, (\hat{\lambda}_p, \underset{\sim}{\hat{e}_p})$

$\hat{Y}_i = \underset{\sim}{\hat{e}_i'} \underset{\sim}{x} = \hat{e}_{i1} x_1 + \hat{e}_{i2} x_2 + \cdots + \hat{e}_{ip} x_p \quad i = 1, \ldots, p$

$\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$ $\quad x$ is any observation on $X_1, \ldots, X_p$

Sample $var\{\hat{Y}_k\} = \hat{\lambda}_k \quad k = 1, \ldots, p$

Sample $cov\{\hat{Y}_i, \hat{Y}_k\} = 0 \quad i \neq k$

Total Sample variance $\sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p$

$$r_{\hat{Y}_i, X_k} = \frac{\hat{e}_{ik} \sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}} \qquad i, k = 1, \ldots, p$$

Sample Principal Components can be obtained from $\hat{\Sigma}, S, R$

Centered values: $\hat{y}_i = \hat{e}_i'(\underset{\sim}{x} - \underset{\sim}{\bar{x}})$ $i = 1, \ldots, p$ for any $\underset{\sim}{x}$

Observed values: $\hat{y}_{ji} = \hat{e}_i'(\underset{\sim}{x}_j - \underset{\sim}{\bar{x}})$ $j = 1, \ldots, n$

$$\Rightarrow \bar{\hat{y}}_i = \frac{1}{n}\sum_{j=1}^{n} \hat{e}_i'(\underset{\sim}{x}_j - \underset{\sim}{\bar{x}}) = \frac{1}{n}\hat{e}_i'\sum_{j=1}^{n}(\underset{\sim}{x}_j - \underset{\sim}{\bar{x}}) = 0$$

Example – Metals (Cd, Hg, Pb) in Mediterranean Gastropods.

$n = 34$

$$R = \begin{bmatrix} 1.0000 & 0.3112 & -0.0395 \\ & 1.0000 & -0.1065 \\ & & 1.0000 \end{bmatrix}$$

$\lambda_1 = 1.3426$ $\lambda_2 = 0.9760$ $\lambda_3 = 0.6814$

$$\underset{\sim}{e}_1 = \begin{bmatrix} .6614 \\ .6912 \\ -.2911 \end{bmatrix} \quad \underset{\sim}{e}_2 = \begin{bmatrix} .3157 \\ .0955 \\ .9441 \end{bmatrix} \quad \underset{\sim}{e}_3 = \begin{bmatrix} .6804 \\ -.7163 \\ -.1550 \end{bmatrix}$$

Cumulative Proportion: $\frac{1.3426}{3} = .4475$ $\frac{2.3186}{3} = .7729$ $1$

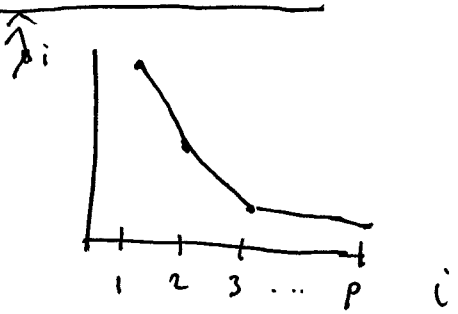
| $z_k$ | $r_{\hat{y}_1, z_k}$ | $r_{\hat{y}_2, z_k}$ | $r_{\hat{y}_3, z_k}$ |
|---|---|---|---|
| $z_1$ | $.6614\sqrt{1.3426} = .7664$ | $.3157\sqrt{.9760} = .3119$ | $.6804\sqrt{.6814} = .5616$ |
| $z_2$ | $.6912\sqrt{1.3426} = .8009$ | $.0955\sqrt{.9760} = .0943$ | $-.7163\sqrt{.6814} = -.5913$ |
| $z_3$ | $-.2911\sqrt{1.3426} = -.3373$ | $.9441\sqrt{.9760} = .9327$ | $-.1550\sqrt{.6814} = -.1279$ |

Interpretation: Mean of Cd, Hg | Pb | Difference between Cd, Hg

## Number of Principal Components

Scree plot

Look for elbow/bend in plot.




EXAMPLE 8.4 - PROGRAM posted on webpage

## Interpretation of Sample Principal Components

$\hat{y}_i = \hat{e}_i'(\underset{\sim}{x} - \bar{\underset{\sim}{x}})$ realizations of population $Y_i = \underset{\sim}{e}_i'(\underset{\sim}{X} - \underset{\sim}{\mu})$

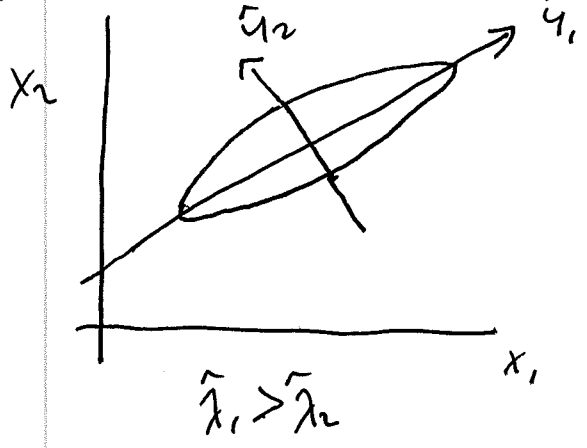
From sample values $x_j$, approximate $\underset{\sim}{\mu}$ by $\bar{\underset{\sim}{x}}$, $\Sigma$ by $S$

$S \equiv$ Pos. def. $\Rightarrow (\underset{\sim}{x} - \bar{\underset{\sim}{x}})' S^{-1} (\underset{\sim}{x} - \bar{\underset{\sim}{x}}) = c^2$ estimates

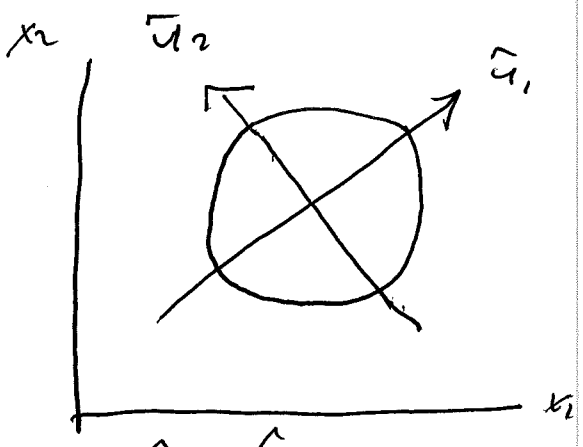
Contours of $(\underset{\sim}{x} - \underset{\sim}{\mu})' \Sigma^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}) = c^2$

Lengths of hyperellipsoid axes are $\sqrt{\hat{\lambda}_i}$ $i = 1, \ldots, p$

$p = 2$




$\hat{\lambda}_1 > \hat{\lambda}_2$




$\hat{\lambda}_1 \doteq \hat{\lambda}_2$

# Standardizing the Sample Principal Components

$$\underset{\sim}{Z}_j = D^{-1/2}(\underset{\sim}{X}_j - \bar{\underset{\sim}{X}}) = \begin{bmatrix} \frac{X_{j1} - \bar{X}_1}{\sqrt{s_{11}}} \\ \vdots \\ \frac{X_{jp} - \bar{X}_p}{\sqrt{s_{pp}}} \end{bmatrix} \quad j = 1, \ldots, n$$

$$\underset{n \times p}{Z} = \begin{bmatrix} \underset{\sim}{Z}_1' \\ \underset{\sim}{Z}_2' \\ \vdots \\ \underset{\sim}{Z}_n' \end{bmatrix} = \begin{bmatrix} \frac{X_{11} - \bar{X}_1}{\sqrt{s_{11}}} & \cdots & \frac{X_{1p} - \bar{X}_p}{\sqrt{s_{pp}}} \\ \vdots & & \\ \frac{X_{n1} - \bar{X}_1}{\sqrt{s_{11}}} & \cdots & \frac{X_{np} - \bar{X}_p}{\sqrt{s_{pp}}} \end{bmatrix}$$

$$\bar{\underset{\sim}{Z}} = \frac{1}{n} Z' \underset{\sim}{1}_n$$

$$S_Z = \frac{1}{n-1}\left(Z - \frac{1}{n}\underset{\sim}{1}\underset{\sim}{1}'Z\right)'\left(Z - \frac{1}{n}\underset{\sim}{1}\underset{\sim}{1}'Z\right)$$

$$= \frac{1}{n-1}\left(Z - \underset{\sim}{1}\bar{\underset{\sim}{Z}}'\right)'\left(Z - \underset{\sim}{1}\bar{\underset{\sim}{Z}}'\right) = \frac{1}{n-1} Z'Z$$

Note: $$\sum_{j=1}^{n}\left[\frac{(X_{ji} - \bar{X}_i)}{\sqrt{s_{ii}}}\,\frac{(X_{jk} - \bar{X}_k)}{\sqrt{s_{kk}}}\right] = \frac{(n-1)s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}$$

$$= (n-1) r_{ik}$$

$$\Rightarrow S_Z = \frac{1}{n-1} Z'Z = R$$

$z_1, \ldots, z_n$ ≡ Standardized observations w/ covariance matrix $R$.

$$\hat{y}_i = \hat{e}_i' \underset{\sim}{z} = \hat{e}_{i1} z_1 + \cdots + \hat{e}_{ip} z_p \quad i = 1, \ldots, p$$

where $(\hat{\lambda}_i, \hat{e}_i)$ is $i^{th}$ eigenvalue-eigenvector pair of $R$

w/ $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$

Sample variance $\{\hat{y}_i\} = \hat{\lambda}_i \quad i = 1, \ldots, p$

Sample covariance $\{\hat{y}_i, \hat{y}_k\} = 0 \quad i \neq k$

Total (standardized) Sample variance $= tr(R) = p = \hat{\lambda}_1 + \cdots + \hat{\lambda}_p$

$$r_{\hat{y}_i, z_k} = \hat{e}_{ik} \sqrt{\hat{\lambda}_i} \quad i, k = 1, \ldots, p$$

Proportion of Standardized sample variance due to $p^{th}$ PC:

$$\frac{\hat{\lambda}_i}{p} \quad i = 1, \ldots, p$$

EXAMPLE 8.5 - ON WEBPAGE

## 8.4 Graphing Principal Components

$$\underset{\sim}{x}_j = (\underset{\sim}{x}_j'\hat{\underset{\sim}{e}}_1)\hat{\underset{\sim}{e}}_1 + \cdots + (\underset{\sim}{x}_j'\hat{\underset{\sim}{e}}_p)\hat{\underset{\sim}{e}}_p = \hat{y}_{j1}\underset{\sim}{e}_1 + \cdots + \hat{y}_{jp}\hat{\underset{\sim}{e}}_p$$

1. Check for normality: (a) Scatterplots of first few PC's
   (b) Q-Q plots of sample values for each pc.
2. Scatter diagrams and Q-Q plots for last few PCs to identify suspect data cases.

EXAMPLE 8.7 R PROGRAM ON WEBPAGE

## 8.5 Large-Sample Inferences

Assumptions: $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ ≡ Random sample from Normal Population

Eigenvalues of $\Sigma$ are distinct positive $\lambda_1 > \cdots > \lambda_p > 0$

(Case where the number of equal eigenvalues is known is an exception (e.g. Testing Compound Symmetry)).

Large-sample Theory of $\hat{\lambda}' = [\hat{\lambda}_1, \ldots, \hat{\lambda}_p]$, $\hat{\underset{\sim}{e}}_1, \ldots, \hat{\underset{\sim}{e}}_p$ for $S$

(1) $\Lambda = \text{diag}\{\lambda_i\}$ for $\Sigma \Rightarrow \sqrt{n}(\hat{\underset{\sim}{\lambda}} - \underset{\sim}{\lambda}) \dot{\sim} N_p(\underset{\sim}{0}, 2\Lambda^2)$

(2) Let $E_i \equiv \lambda_i \sum_{\substack{k=1 \\ k \neq i}}^{p} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} \underset{\sim}{e}_k \underset{\sim}{e}_k' \Rightarrow \sqrt{n}(\hat{\underset{\sim}{e}}_i - \underset{\sim}{e}_i) \dot{\sim} N_p(\underset{\sim}{0}, E_i)$

(3) Each $\hat{\lambda}_i$ is distributed independently of elements of associated $\hat{\underset{\sim}{e}}_i$.

$i = 1, \ldots, p$:

① $\Rightarrow Pr\left\{ -z(\frac{\alpha}{2})\sqrt{2\lambda_i^2} \le \sqrt{n}(\hat{\lambda}_i - \lambda_i) \le z(\alpha/2)\sqrt{2\lambda_i^2} \right\} = 1-\alpha$

$= Pr\left\{ -z(\frac{\alpha}{2})\sqrt{2} \le \sqrt{n}\left(\frac{\hat{\lambda}_i}{\lambda_i} - 1\right) \le z_{\alpha/2}\sqrt{2} \right\}$

$= Pr\left\{ -z(\frac{\alpha}{2})\sqrt{\frac{2}{n}} \le \frac{\hat{\lambda}_i}{\lambda_i} - 1 \le z(\frac{\alpha}{2})\sqrt{\frac{2}{n}} \right\}$

$= Pr\left\{ 1 - z(\frac{\alpha}{2})\sqrt{\frac{2}{n}} \le \frac{\hat{\lambda}_i}{\lambda_i} \le 1 + z(\frac{\alpha}{2})\sqrt{\frac{2}{n}} \right\}$

$= Pr\left\{ \frac{1}{\hat{\lambda}_i}\left(1 - z(\frac{\alpha}{2})\sqrt{\frac{2}{n}}\right) \le \frac{1}{\lambda_i} \le \frac{1}{\hat{\lambda}_i}\left(1 + z(\frac{\alpha}{2})\sqrt{\frac{2}{n}}\right) \right\}$

$= Pr\left\{ \frac{\hat{\lambda}_i}{1 + z(\frac{\alpha}{2})\sqrt{\frac{2}{n}}} \le \lambda_i \le \frac{\hat{\lambda}_i}{1 - z(\frac{\alpha}{2})\sqrt{\frac{2}{n}}} \right)$

$\equiv$ Large-sample $100(1-\alpha)\%$ CI for $\lambda_i$

Bonferroni simultaneous CI's can be computed for $m$ $\lambda_i$'s by $z(\frac{\alpha}{2}) \to z(\frac{\alpha}{2m})$

② for $\epsilon_i$, must use estimates $\{\hat{\lambda}_i\}, \{\hat{e}_i\}$ to obtain estimated standard errors.

## Testing for equal Correlation Structure

$$H_0: \rho = \rho_0 = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \qquad H_1: \rho \neq \rho_0$$

## Lawley's Procedure

$$\bar{r}_k = \frac{1}{p-1} \sum_{\substack{i=1 \\ i \neq k}}^{p} r_{ik} \qquad k = 1, \ldots, p \qquad \bar{r} = \frac{2}{p(p-1)} \sum\sum_{i<k} r_{ik}$$

$$\hat{\gamma} = \frac{(p-1)^2 \left[ 1 - (1-\bar{r})^2 \right]}{p - (p-2)(1-\bar{r})^2}$$

Reject $H_0$ in favor of $H_A$ if:

$$T = \frac{n-1}{(1-\bar{r})^2} \left[ \sum\sum_{i<k} (r_{ik} - \bar{r})^2 - \hat{\gamma} \sum_{k=1}^{p} (\bar{r}_k - \bar{r})^2 \right] > \chi^2_{(p+1)(p-2)/2}(\alpha)$$