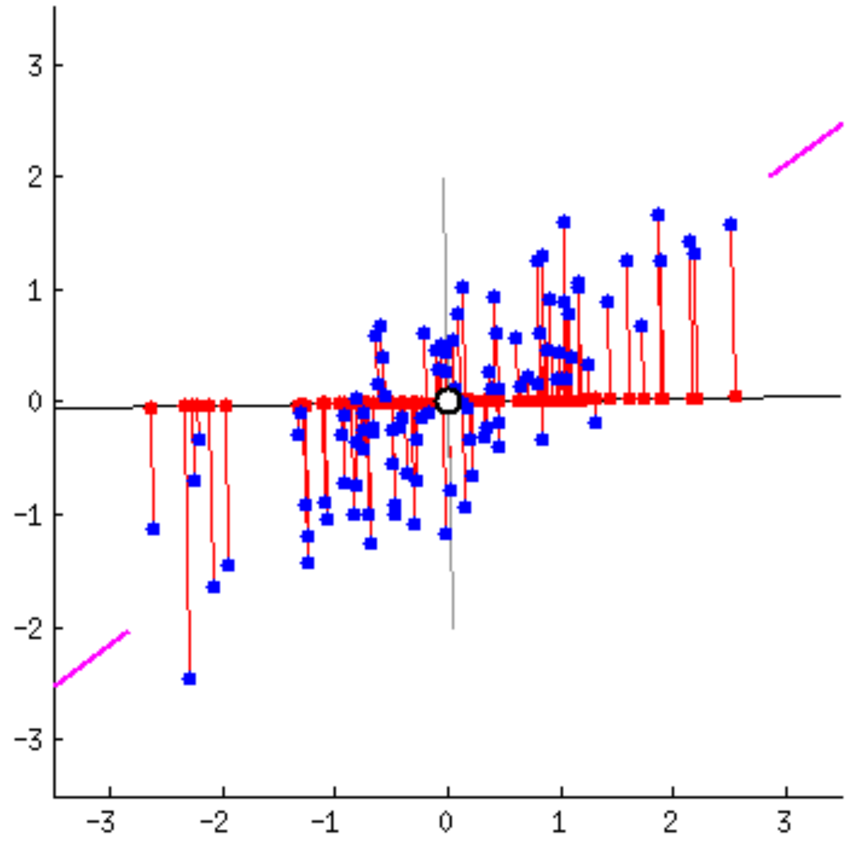
Principle Component Analysis

Jing Qin

03/05/2023

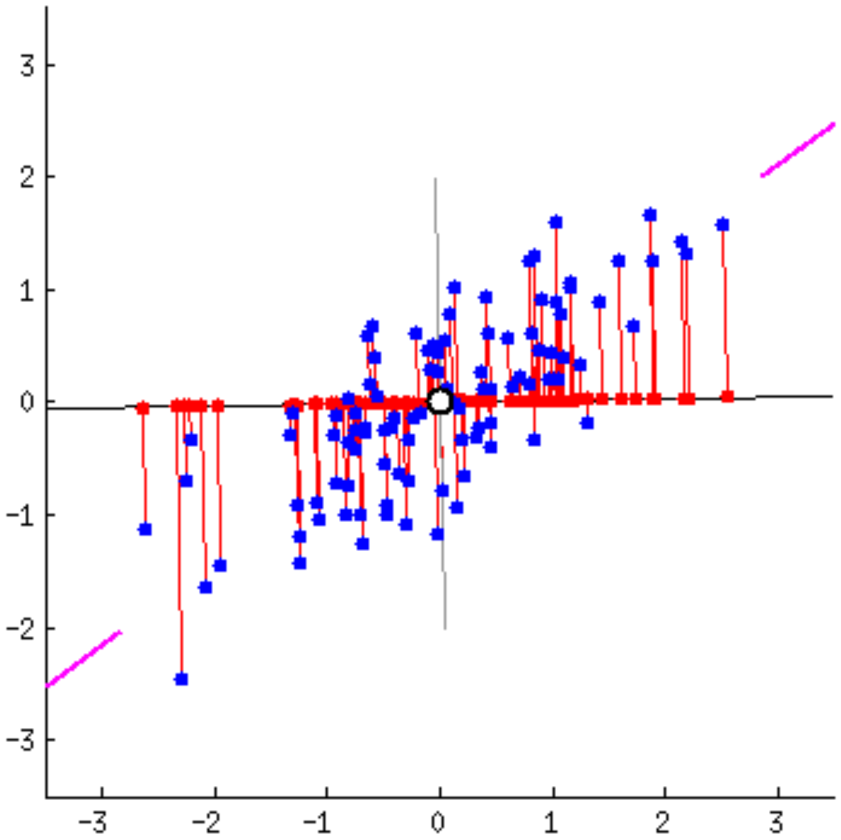
What is PCA? Start with a movie…



The **first** principal component accounts for the **largest possible variance** in the data set.

The second principal component is calculated as the one that is **uncorrelated with** (perpendicular to) the first principal component and that it accounts for the **next highest variance**...

What is PCA? Start with a movie…



How to do this?

The **first** principal component accounts for the **largest possible variance** in the data set.

The second principal component is calculated as the one that is **uncorrelated with** (perpendicular to) the first principal component and that it accounts for the **next highest variance**...

Main result:

This is the answer to the 'How' question.
Step-by-step gets into it ~

Principle components are ***linear combinations*** of the random variables, and the coefficients are determined by the ***eigenvectors*** of covariance matrix.

Result 8.1. Let Σ be the covariance matrix associated with the random vector $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$. Let Σ have the eigenvalue-eigenvector pairs $(\lambda_1, \mathbf{e}_1)$, $(\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ where $\boxed{\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0}$. Then the *ith principal component* is given by

$$Y_i = \mathbf{e}_i'\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p, \qquad i = 1, 2, \ldots, p \tag{8-4}$$

With these choices,

$$\boxed{\begin{aligned} \operatorname{Var}(Y_i) &= \mathbf{e}_i'\Sigma\mathbf{e}_i = \lambda_i \\ \operatorname{Cov}(Y_i, Y_k) &= \mathbf{e}_i'\Sigma\mathbf{e}_k = 0 \end{aligned}} \qquad \begin{aligned} &i = 1, 2, \ldots, p \\ &i \neq k \end{aligned} \tag{8-5}$$

If some λ_i are equal, the choices of the corresponding coefficient vectors, $\mathbf{e}_i$, and hence Y_i, are not unique.

All the eigenvalues of Σ are non-negative (Σ is positive semi-definite)

- Σ has eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$
- Alternative definition: For all $\boldsymbol{y}$, the quadratic from $\boldsymbol{y}' \cdot \Sigma \cdot \boldsymbol{y} \geq 0$.
- Proof of Σ is positive semi-definite:

Assume that $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ is a random sample and $\overline{\boldsymbol{X}}$ is the sample mean vector

$$\begin{aligned} \Sigma &= \frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{X}_j - \overline{\boldsymbol{X}})(\boldsymbol{X}_j - \overline{\boldsymbol{X}})' \\ \boldsymbol{y}' \cdot \Sigma \cdot \boldsymbol{y} &= \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{y}'(\boldsymbol{X}_j - \overline{\boldsymbol{X}})(\boldsymbol{X}_j - \overline{\boldsymbol{X}})'\boldsymbol{y} \\ &= \frac{1}{n}\sum_{j=1}^{n}\left((\boldsymbol{X}_j - \overline{\boldsymbol{X}})'\boldsymbol{y}\right)^2 \geq 0 \end{aligned}$$

Inner product (2-4)

In sum...

Original components

Principle components are p linear combinations of $\boxed{X_1, X_2, \ldots, X_p}$

Principle components

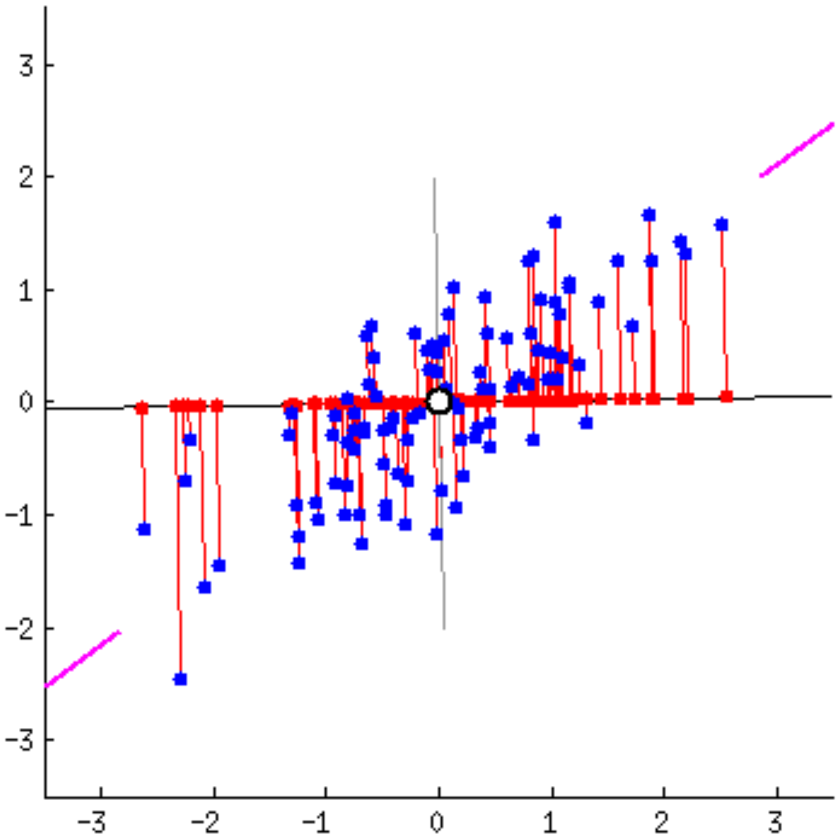
$$
\begin{aligned}
Y_1 &= \boldsymbol{a}_1'\boldsymbol{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
Y_2 &= \boldsymbol{a}_2'\boldsymbol{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\
&\ldots \\
Y_p &= \boldsymbol{a}_p'\boldsymbol{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p
\end{aligned}
$$

satisfying

- variance as large as possible $Var(Y_1) \geq Var(Y_2) \geq \cdots \geq Var(Y_p)$
- uncorrelated $Cov(Y_j, Y_k) = 0$ for any pair (j, k)
- "standardized" $\boldsymbol{a}_i' \cdot \boldsymbol{a} = 1$, where $\boldsymbol{a}_i' = (a_{i1}, a_{i2}, \ldots, a_{ip})$

(Result 8.1) $\boldsymbol{a}_i = \boldsymbol{e}_i$, where $(\lambda_i, \boldsymbol{e}_i)$ is eigenvalue-eigenvector pair of Σ.

The movie, again



The **first** principal component accounts for the **largest possible variance** in the data set.

The second principal component is calculated as the one that is **uncorrelated with** (perpendicular to) the first principal component and that it accounts for the **next highest variance**...

How many PCs we need (not all of them)

Result 8.2. Let $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ have covariance matrix $\boldsymbol{\Sigma}$, with eigenvalue-eigenvector pairs $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Let $Y_1 = \mathbf{e}_1'\mathbf{X}, Y_2 = \mathbf{e}_2'\mathbf{X}, \ldots, Y_p = \mathbf{e}_p'\mathbf{X}$ be the principal components. Then

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} \operatorname{Var}(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \operatorname{Var}(Y_i)$$

Result 8.2 says that

$$\begin{aligned} \text{Total population variance} &= \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} \\ &= \lambda_1 + \lambda_2 + \cdots + \lambda_p \end{aligned} \tag{8-6}$$

and consequently, the proportion of total variance due to (explained by) the kth principal component is

$$\begin{pmatrix} \text{Proportion of total} \\ \text{population variance} \\ \text{due to } k\text{th principal} \\ \text{component} \end{pmatrix} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \qquad k = 1, 2, \ldots, p \tag{8-7}$$

$\geq 95\%$

How many PCs we need (not all of them)

Result 8.2 says that

$$\begin{aligned}\text{Total population variance} &= \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} \\ &= \lambda_1 + \lambda_2 + \cdots + \lambda_p \end{aligned} \qquad (8\text{-}6)$$

and consequently, the proportion of total variance due to (explained by) the kth principal component is

$$\begin{pmatrix}\text{Proportion of total} \\ \text{population variance} \\ \text{due to } k\text{th principal} \\ \text{component}\end{pmatrix} = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \qquad k = 1, 2, \ldots, p \qquad (8\text{-}7)$$

- Proportion of total variance explained by the first k component

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

How to do PCA in R with prcomp()

```
> pca <- prcomp(df)
> pca
Standard deviations (1, .., p=3):
[1] 0.15265434 0.02446027 0.01896934

Rotation (n x k) = (3 x 3):
         PC1        PC2        PC3
V1 0.6831023 -0.1594791  0.7126974
V2 0.5102195 -0.5940118 -0.6219534
V3 0.5225392  0.7884900 -0.3244015
> summary(pca)
Importance of components:
                          PC1     PC2     PC3
Standard deviation     0.1527 0.02446 0.01897
Proportion of Variance 0.9605 0.02466 0.01483
Cumulative Proportion  0.9605 0.98517 1.00000
```

How many PCs we need (scree plot) with screeplot()

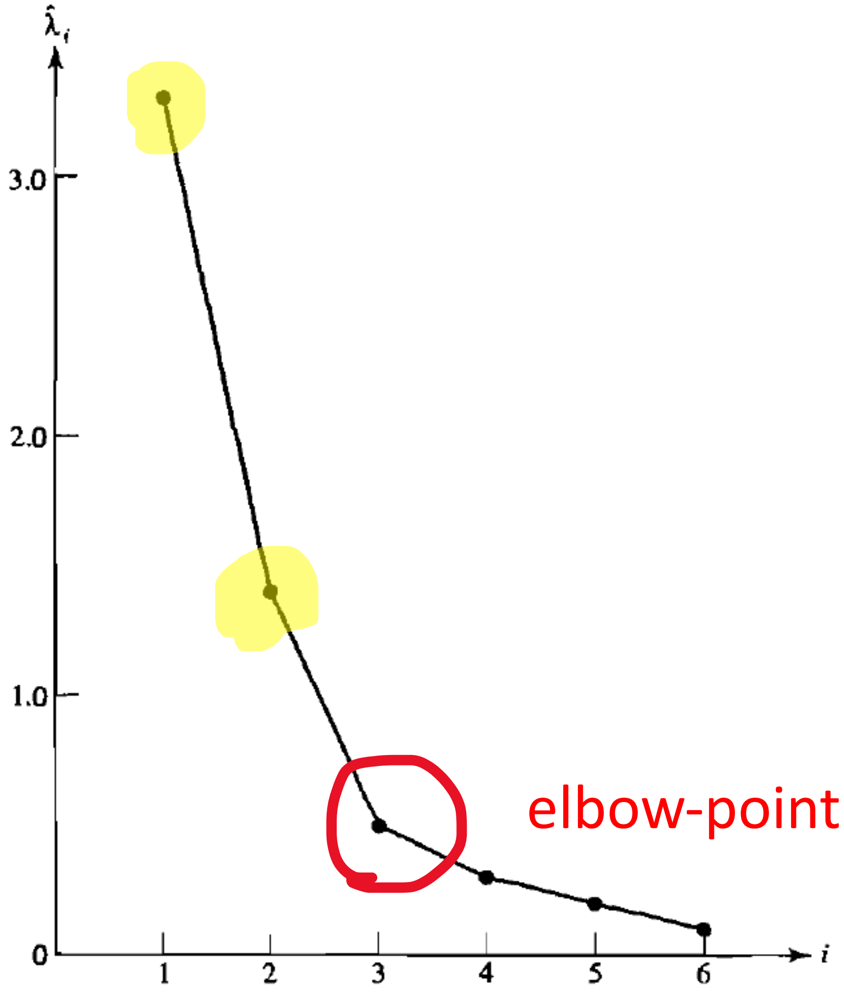


Figure 8.2 A scree plot.

Standardization (PCA on correlation matrix $\boldsymbol{\rho}$ not Σ)

```r
pca2 <- prcomp(df, scale=T)
```

by (2-37). The principal components of $\mathbf{Z}$ may be obtained from the eigenvectors of the *correlation* matrix $\boldsymbol{\rho}$ of $\mathbf{X}$. All our previous results apply, with some simplifications, since the variance of each Z_i is unity. We shall continue to use the notation Y_i to refer to the ith principal component and $(\lambda_i, \mathbf{e}_i)$ for the eigenvalue–eigenvector pair from either $\boldsymbol{\rho}$ or Σ. *However, the* $(\lambda_i, \mathbf{e}_i)$ *derived from* Σ *are, in general, not the same as the ones derived from* $\boldsymbol{\rho}$.

Result 8.4. The ith principal component of the standardized variables $\mathbf{Z}' = [Z_1, Z_2, \ldots, Z_p]$ with $\text{Cov}(\mathbf{Z}) = \boldsymbol{\rho}$, is given by

$$Y_i = \mathbf{e}_i'\mathbf{Z} = \mathbf{e}_i'(\mathbf{V}^{1/2})^{-1}(\mathbf{X} - \boldsymbol{\mu}), \qquad i = 1, 2, \ldots, p$$

Moreover,

$$\sum_{i=1}^{p} \text{Var}(Y_i) = \sum_{i=1}^{p} \text{Var}(Z_i) = p \tag{8-11}$$

that the total (standardized variables) population variance he diagonal elements of the matrix $\boldsymbol{\rho}$. Using (8-7) with $\mathbf{Z}$ in place of $\mathbf{X}$, we find that the proportion of total variance explained by the kth principal component of $\mathbf{Z}$ is

$$\begin{pmatrix} \text{Proportion of (standardized)} \\ \text{population variance due} \\ \text{to } k\text{th principal component} \end{pmatrix} = \frac{\lambda_k}{p}, \qquad k = 1, 2, \ldots, p \tag{8-12}$$

where the λ_k's are the eigenvalues of $\boldsymbol{\rho}$.


MANGE TUSIND
AF HJERTET
TIL DIG
FRA MIG
TAK tak
FOR OPMÆRKSOMHEDEN
ER KUN ET FATTIGT ORD
FOR HJÆLPEN
FOR SIDST
FORDI DU ER DIG
FOR ALT
FOR INDBYDELSEN
TAK


LARS LILHOLT BAND
HELD OG LYKKE