Singular Value Decomposition

Cheng-Ying Chou

Eigenvalues and Eigenvectors

- Eigenvalue problem (特徵值問題) (one of the most important problems in the linear algebra):

- Eigenvalue (特徵值) and Eigenvector (特徵向量):

A: an $n \times n$ matrix

λ: a scalar (could be **zero**)

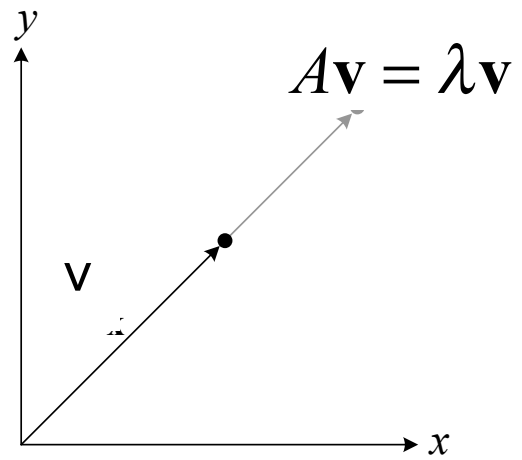
$\mathbf{v}$: a **nonzero** vector in R^n

Eigenvalue

$$A\mathbf{v} = \lambda\mathbf{v}$$

Eigenvector

※ Geometric Interpretation



We summarize the computational approach for determining eigenpairs (λ, **x**) (eigenvalues and eigen vector) as a two-step procedure:

Step I. To find the eigenvalues of **A** compute the roots of the characteristic equation $\mathbf{det(A - \lambda I_n) = 0}$.

Step II. To find an eigenvector corresponding to an eigenvalue μ, compute a nontrivial solution to the homogeneous linear system $\mathbf{(A - \mu I_n)x = 0}$.

Example: Find eigenpairs of $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ -2 & 4 \end{bmatrix}$.

Step I. Find the eigenvalues.

$$\mathbf{c(\lambda) = det(A - \lambda I_2)} = \mathbf{det}\left(\begin{bmatrix} 1-\lambda & 1 \\ -2 & 4-\lambda \end{bmatrix}\right)$$

$$= (1-\lambda)(4-\lambda) + 2 = \lambda^2 - 5\lambda + 6$$

Thus the characteristic polynomial is a quadratic and the eigenvalues are the solutions of $\lambda^2 - 5\lambda + 6 = 0$. We factor the quadratic to get $(\lambda - 3)(\lambda - 2) = 0$ so the eigenvalues are $\lambda_1 = 3$ and $\lambda_2 = 2$.

The eigenvalues are $\lambda_1 = 3$ and $\lambda_2 = 2$.

Step II. To find corresponding $\mathbf{(A - \lambda_i I_n)x = 0}$ for

Case $\lambda_1 = 3$: We have that $\mathbf{(A - 3I_2)x = 0}$ has augmented matrix $\left[\begin{array}{cc|c} -2 & 1 & 0 \\ -2 & 1 & 0 \end{array}\right]$ and its rref is $\left[\begin{array}{cc|c} 1 & \frac{-1}{2} & 0 \\ 0 & 0 & 0 \end{array}\right]$. (Verify.) Thus if $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ we have $x_1 = (1/2)x_2$ so $\mathbf{x} = \begin{bmatrix} (1/2)x_2 \\ x_2 \end{bmatrix} = x_2 \begin{bmatrix} 1/2 \\ 1 \end{bmatrix}$, $x_2 \neq 0$. Choosing $x_2 = 2$, to conveniently get integer entries, gives eigenvector $\mathbf{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

Case $\lambda_2 = 2$: We have that $\mathbf{(A - 2I_2)p = 0}$ has augmented matrix $\left[\begin{array}{cc|c} -1 & 1 & 0 \\ -2 & 2 & 0 \end{array}\right]$ and its rref is $\left[\begin{array}{cc|c} 1 & -1 & 0 \\ 0 & 0 & 0 \end{array}\right]$. (Verify.) Thus if $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ we have $x_1 = x_2$ so $\mathbf{x} = \begin{bmatrix} x_2 \\ x_2 \end{bmatrix} = x_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $x_2 \neq 0$. Choosing $x_2 = 1$ gives eigenvector $\mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

Note: rref means row reduced echelon form.

Example

- Let

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

Real, symmetric.

- Then

$$A - \lambda I = \begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix} \Rightarrow (2-\lambda)^2 - 1 = 0.$$

- The eigenvalues are 1 and 3 (nonnegative, real).
- The eigenvectors are orthogonal (and real):

$$\begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Plug in these values and solve for eigenvectors.

Eigenvalue example

- Consider,

$$\mathbf{A}=\begin{bmatrix}1 & 2\\ 2 & 4\end{bmatrix}\Rightarrow\begin{cases}\lambda^2-(a_{11}+a_{22})\lambda+(a_{11}a_{22}-a_{12}a_{21})=0\\ \lambda^2-(1+4)\lambda+(1\cdot 4-2\cdot 2)=0\\ \lambda^2=(1+4)\lambda\Rightarrow\lambda=0,\lambda=5\end{cases}$$

- The corresponding eigenvectors can be computed as

$$\lambda=0\Rightarrow\left[\begin{bmatrix}1 & 2\\ 2 & 4\end{bmatrix}-\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}\right]\cdot\begin{bmatrix}x\\ y\end{bmatrix}=0\Rightarrow\begin{bmatrix}1 & 2\\ 2 & 4\end{bmatrix}\cdot\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}1x+2y\\ 2x+4y\end{bmatrix}=\begin{bmatrix}0\\ 0\end{bmatrix}$$

$$\lambda=5\Rightarrow\left[\begin{bmatrix}1 & 2\\ 2 & 4\end{bmatrix}-\begin{bmatrix}5 & 0\\ 0 & 5\end{bmatrix}\right]\cdot\begin{bmatrix}x\\ y\end{bmatrix}=0\Rightarrow\begin{bmatrix}-4 & 2\\ 2 & -1\end{bmatrix}\cdot\begin{bmatrix}x\\ y\end{bmatrix}=\begin{bmatrix}-4x+2y\\ 2x-1y\end{bmatrix}=\begin{bmatrix}0\\ 0\end{bmatrix}$$

  - For $\lambda = 0$, one possible solution is $\mathbf{x} = (2, -1)$
  - For $\lambda = 5$, one possible solution is $\mathbf{x} = (1, 2)$

Eigen/diagonal Decomposition

- Let $\mathbf{S} \in \mathbb{R}^{m \times m}$ be a **square** matrix with m **linearly independent eigenvectors** (a "non-defective" matrix)
- **Theorem**: Exists an **eigen decomposition** $S=U\Lambda U^{-1}$
  - (cf. matrix diagonalization theorem)

Unique for distinct eigen-values

- Columns of ***U*** are **eigenvectors** of ***S***
- Diagonal elements of Λ are **eigenvalues** of $\Lambda = \text{diag}(\lambda_1 \ldots \lambda_m)$, $\lambda_i >= \lambda_{i+1}$

Diagonal decomposition: why/how

Let ***U*** have the eigenvectors as columns: $U = \begin{bmatrix} v_1 & \dots & v_n \end{bmatrix}$

Then, ***SU*** can be written

$$SU = S\begin{bmatrix} v_1 & \dots & v_n \end{bmatrix} = \begin{bmatrix} \lambda_1 v_1 & \dots & \lambda_n v_n \end{bmatrix} = \begin{bmatrix} v_1 & \dots & v_n \end{bmatrix}\begin{bmatrix} \lambda_1 & & \\ & \dots & \\ & & \lambda_n \end{bmatrix}$$

Thus $\boldsymbol{SU=U\Lambda}$, or $\boldsymbol{U^{-1}SU=\Lambda}$

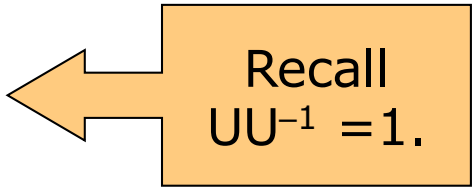
And $\boldsymbol{S=U\Lambda U^{-1}}$.

Diagonal decomposition - example

Recall $S = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}; \lambda_1 = 1, \lambda_2 = 3.$

The eigenvectors form $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ $\quad U = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$

Inverting, we have $U^{-1} = \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix}$



Then, $\boldsymbol{S = U\Lambda U^{-1} =}$ $\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix}$

Example continued

Let's divide $\boldsymbol{U}$ (and multiply $\boldsymbol{U^{-1}}$) by $\sqrt{2}$

Then, $\boldsymbol{S}=$

$$\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

$\boldsymbol{Q}$ $\qquad$ $\boldsymbol{\Lambda}$ $\qquad$ $\boldsymbol{(Q^{-1} = Q^{T})}$

Why? Stay tuned …

Symmetric Eigen Decomposition

- If $\mathbf{S} \in \mathbb{R}^{m \times m}$ is a **symmetric** matrix:
- **Theorem**: Exists a (unique) **eigen decomposition** $S = Q \Lambda Q^T$
- where ***Q*** is **orthogonal:**
  - $Q^{-1} = Q^T$
  - Columns of ***Q*** are normalized eigenvectors
  - Columns are orthogonal.
  - (everything is real)

- If matrix A can be eigen decomposed and if none of its eigenvalues are zero, then A is nonsingular and its inverse is given by

$$\mathbf{A}^{-1} = \mathbf{Q}\Lambda^{-1}\mathbf{Q}^{-1}$$

- Furthermore, because is a diagonal matrix, its inverse is easy to calculate:

$$[\Lambda^{-1}]_{ii} = \frac{1}{\lambda_i}$$

Exercise

- Examine the symmetric eigen decomposition, if any, for each of the following matrices:

$$\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 \\ -2 & 3 \end{bmatrix} \quad \begin{bmatrix} 2 & 2 \\ 2 & 4 \end{bmatrix}$$

Symmetric Eigen Decomposition

- If $S \in \mathbb{R}^{m \times m}$ is a **symmetric** matrix:
- **Theorem**: Exists a (unique) **eigen decomposition** $S = Q \Lambda Q^T$
- where ***Q*** is **orthogonal:**
  - $Q^{-1} = Q^T$
  - Columns of ***Q*** are normalized eigenvectors
  - Columns are orthogonal.
  - (everything is real)

Spectral Decomposition theorem

- If **A** is a symmetric and positive definite k **x** k matrix ($\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$) with λ_i ($\lambda_i > 0$) and $\mathbf{e}_i$, i = 1 ⋯ k being the k eigenvector and eigenvalue pairs, then

$$\underset{(k\times k)}{\mathbf{A}} = \lambda_1 \underset{(k\times 1)}{\mathbf{e}_1}\underset{(1\times k)}{\mathbf{e}_1^T} + \lambda_2 \underset{(k\times 1)}{\mathbf{e}_2}\underset{(1\times k)}{\mathbf{e}_2^T} \cdots + \lambda_k \underset{(k\times 1)}{\mathbf{e}_k}\underset{(1\times k)}{\mathbf{e}_k^T} \Rightarrow \underset{(k\times k)}{\mathbf{A}} = \sum_{i=1}^{k} \lambda_i \underset{(k\times 1)}{\mathbf{e}_i}\underset{(1\times k)}{\mathbf{e}_i^T} = \mathbf{P}\Lambda\mathbf{P}^T$$

$$\underset{(k\times k)}{\mathbf{P}} = [\mathbf{e}_1, \mathbf{e}_2 \cdots \mathbf{e}_k], \underset{(k\times k)}{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix}$$

  - This is also called the eigen decomposition theorem
- Any symmetric matrix can be reconstructed using its eigenvalues and eigenvectors.

Example for spectral decomposition

- Let A be a symmetric, positive definite matrix

$$A = \begin{bmatrix} 2.2 & 0.4 \\ 0.4 & 2.8 \end{bmatrix} \Rightarrow \det(A - \lambda I) = 0$$

$$\Rightarrow \lambda^2 - 5\lambda + (6.16 - 0.16) = (\lambda - 3)(\lambda - 2) = 0$$

- The eigenvectors for the corresponding eigenvalues are

$$\mathbf{e}_1^T = \left[1/\sqrt{5}, 2/\sqrt{5} \right], \mathbf{e}_2^T = \left[2/\sqrt{5}, -1/\sqrt{5} \right]$$

- Consequently,

$$A = \begin{bmatrix} 2.2 & 0.4 \\ 0.4 & 2.8 \end{bmatrix} = 3 \begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix} \begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \end{bmatrix} + 2 \begin{bmatrix} 2/\sqrt{5} \\ -1/\sqrt{5} \end{bmatrix} \begin{bmatrix} 2/\sqrt{5} & -1/\sqrt{5} \end{bmatrix}$$

$$= \begin{bmatrix} 0.6 & 1.2 \\ 1.2 & 2.4 \end{bmatrix} + \begin{bmatrix} 1.6 & -0.8 \\ -0.8 & 0.4 \end{bmatrix}$$

Spectral Decomposition

- The eigenvectors of an n x n symmetric matrix S are mutually orthogonal $Q = (x_1, x_2, \ldots, x_n)$, $QQ^T = Q^TQ = I$.

$$\mathbf{S} = \mathbf{S}QQ^T = \mathbf{S}(\mathbf{x_1}, \mathbf{x}_2, \cdots, \mathbf{x_n})Q^T = (\mathbf{S}\mathbf{x_1}, \mathbf{S}\mathbf{x}_2, \cdots, \mathbf{S}\mathbf{x_n})Q^T$$
$$= (\lambda_1\mathbf{x_1}, \lambda_2\mathbf{x}_2, \cdots, \lambda_n\mathbf{x_n})Q^T = Q\Lambda Q^T$$

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} = Q^T\mathbf{S}Q$$

Singular Value Decomposition

- If **A** is a rectangular m x k matrix of real numbers, then there exists an m x m orthogonal matrix **U** and a k x k orthogonal matrix **V** such that

$$\underset{(m\times k)}{\mathbf{A}} = \underset{(m\times m)}{\mathbf{U}}\,\underset{(m\times k)}{\Sigma}\,\underset{(k\times k)}{\mathbf{V}^T} \qquad \mathbf{U}\mathbf{U}^T = \mathbf{V}\mathbf{V}^T = \mathbf{I}$$

  - Σ is an m x k matrix where the $(i, j)^{th}$ entry λ_i ¸ 0, i = 1 ⋯ min(m, k) and the other entries are zero.
    - The positive constants λ_i are the singular values of **A**
- If A has rank r, then there exists r positive constants $\lambda_1, \lambda_2, \cdots \lambda_r$, r orthogonal m x 1 unit vectors $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_r$ and r orthogonal k x 1 unit vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_r$ such that

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$$

  - Similar to the spectral decomposition theorem

Singular Value Decomposition (SVD)

- Handy mathematical technique that has application to many problems
- Given any $m \times n$ matrix **A**, algorithm to find matrices **U**, **V**, and **Σ** such that

$$\mathbf{A} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^T$$

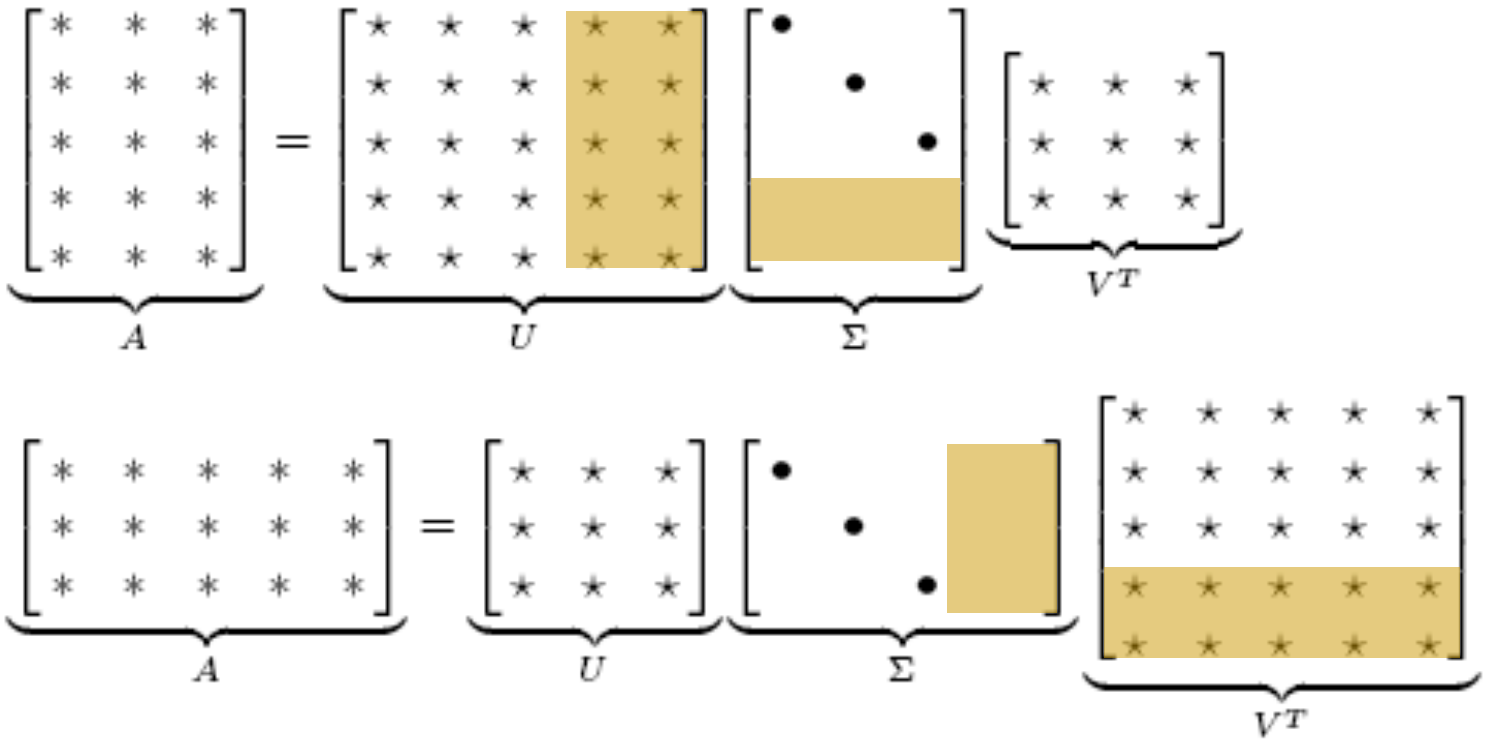
U is $m \times m$ and orthonormal

Σ is $m \times n$ and diagonal

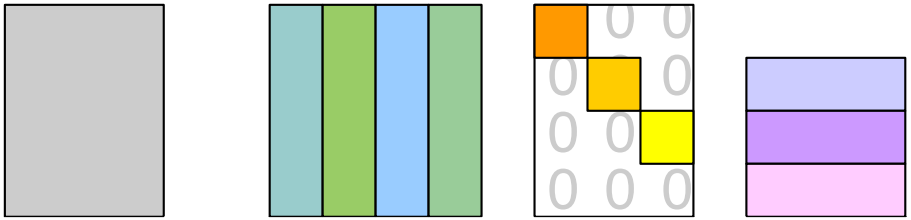
V is $n \times n$ and orthonormal

Singular Value Decomposition

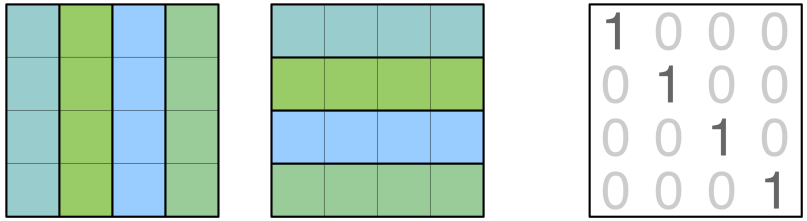
- Illustration of SVD dimensions and sparseness



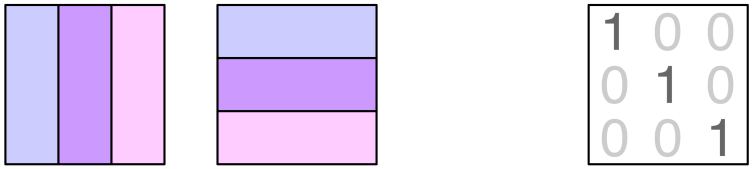
SVD



$$\underset{m\times n}{\mathbf{M}} = \underset{m\times m}{\mathbf{U}} \quad \underset{m\times n}{\mathbf{\Sigma}} \quad \underset{n\times n}{\mathbf{V}^*}$$

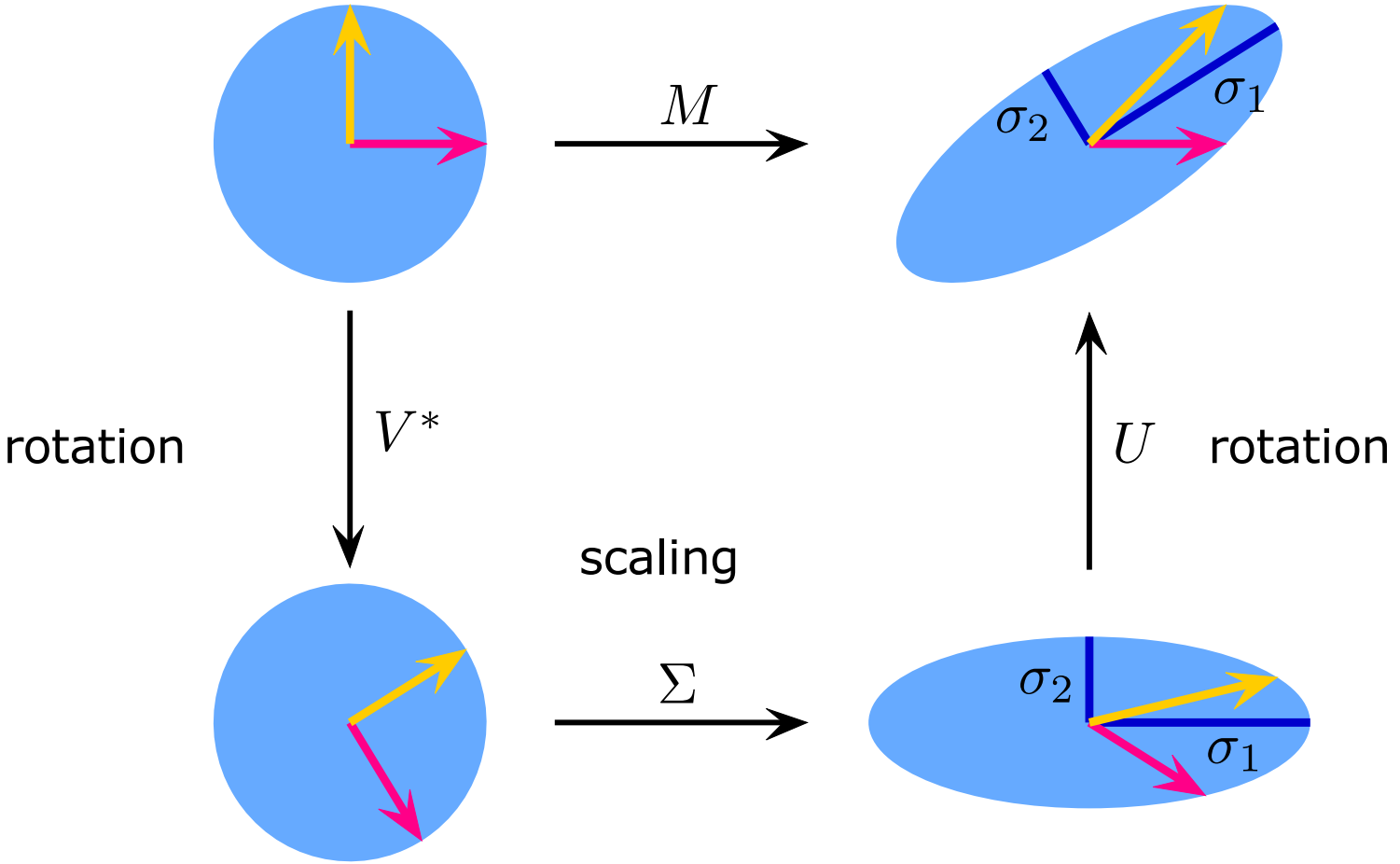


$$\mathbf{U} \quad \mathbf{U}^* = \mathbf{I}_m$$



$$\mathbf{V} \quad \mathbf{V}^* = \mathbf{I}_n$$

SVD: rotation and scaling



$$M = U \cdot \Sigma \cdot V^*$$

- Consider the fitted response

$$\hat{y} = \mathbf{X}\beta_{\text{ridge}} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

- SVD of $\mathbf{X}$ $\quad \mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$

$$\hat{y} = \mathbf{U}\Sigma\mathbf{V}^T\left((\mathbf{U}\Sigma\mathbf{V}^T)^T(\mathbf{U}\Sigma\mathbf{V}^T) + \lambda\mathbf{I}\right)^{-1}(\mathbf{U}\Sigma\mathbf{V}^T)^T\mathbf{y}$$

$$= \mathbf{U}\Sigma\mathbf{V}^T\left(\mathbf{V}\Sigma^T\mathbf{U}^T\mathbf{U}\Sigma\mathbf{V}^T + \lambda\mathbf{I}\right)^{-1}\mathbf{V}\Sigma^T\mathbf{U}^T\mathbf{y}$$

$$= \mathbf{U}\Sigma\mathbf{V}^T\left(\mathbf{V}\Sigma^2\mathbf{V}^T + \mathbf{V}\lambda\mathbf{I}\mathbf{V}^T\right)^{-1}\mathbf{V}\Sigma^T\mathbf{U}^T\mathbf{y}$$

$$= \mathbf{U}\Sigma\mathbf{V}^T\left(\mathbf{V}(\Sigma^2 + \lambda\mathbf{I})\mathbf{V}^T\right)^{-1}\mathbf{V}\Sigma^T\mathbf{U}^T\mathbf{y}$$

$$= \mathbf{U}\Sigma\mathbf{V}^T(\mathbf{V}^T)^{-1}(\Sigma^2 + \lambda\mathbf{I})^{-1}(\mathbf{V})^{-1}\mathbf{V}\Sigma^T\mathbf{U}^T\mathbf{y}$$

$$= \mathbf{U}\Sigma(\Sigma^2 + \lambda\mathbf{I})^{-1}\Sigma^T\mathbf{U}^T\mathbf{y}$$

$$= \sum_{i=1}^{M} u_i \frac{\sigma_i^2}{\sigma_i^2 + \lambda} u_i^T\mathbf{y}, \quad \text{where} \sum_{i=1}^{M} \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \triangleq df(\lambda) \in \mathcal{R}$$

Ridge regression shrinks the coefficients with respect to the orthonormal basis formed by the principal components

SVD example

Let $A = \begin{bmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$

Thus m=3, n=2. Its SVD is

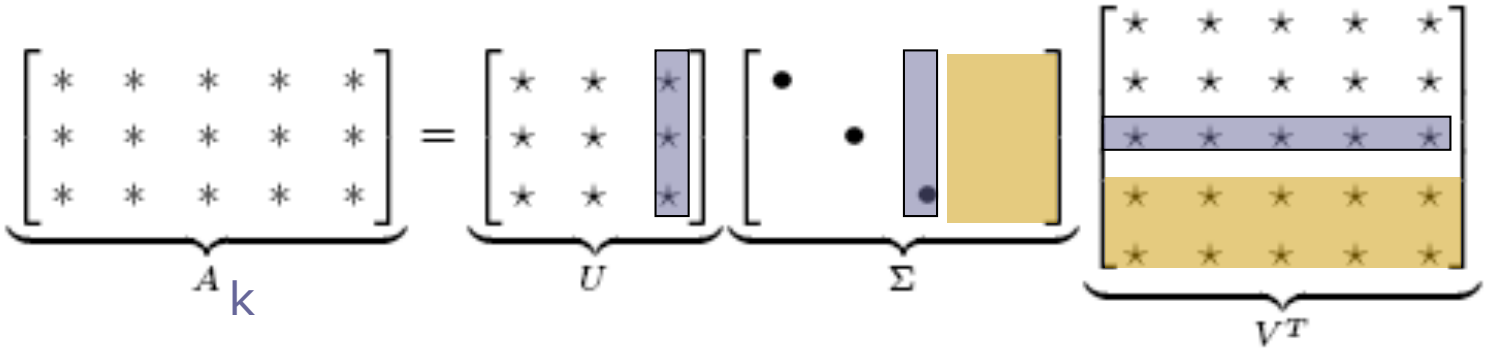
$$\begin{bmatrix} 0 & 2/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{2} & 1/\sqrt{6} & -1/\sqrt{3} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \sqrt{3} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

Typically, the singular values arranged in decreasing order.

Low-rank Approximation

- Solution via SVD

$$A_k = U\ \mathrm{diag}(\sigma_1,\ldots,\sigma_k,\underbrace{0,\ldots,0}_{\text{set smallest r-k singular values to zero}})V^T$$



$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

column notation: sum of rank 1 matrices

Approximation error

- How good (bad) is this approximation?
- It's the best possible, measured by the Frobenius norm of the error:

$$\min_{X:rank(X)=k} \|A - X\|_F = \|A - A_k\|_F = \sigma_{k+1}$$

where the σ_i are ordered such that $\sigma_i \geq \sigma_{i+1}$.

Suggests why Frobenius error drops as *k* increased.

Frobenius norm

$$\| \mathbf{A} \|_F \equiv \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} | a_{ij}^2 |}$$

- Equal to the square root of the matrix trace of AA^H,

$$\| \mathbf{A} \|_F \equiv \sqrt{\mathrm{Tr}(\mathbf{A}\mathbf{A}^T)}$$

- where A^T is the conjugate transpose.

SVD

$$\left(\mathbf{A} \right) = \left(\mathbf{U} \right) \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_n \end{pmatrix} \left(\mathbf{V} \right)^{\mathrm{T}}$$

```
library(MASS)
a <- matrix(c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0,
0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0,
0, 1, 1, 1), 9, 4)

a.svd <- svd(a)
a.svd$d
```

SVD

- The w_i are called the singular values of **A**
- If **A** is singular, some of the s_i will be 0
- In general *rank*(**A**) = number of nonzero s_i
- SVD is mostly unique (up to permutation of singular values, or if some s_i are equal)

SVD and Inverses

- Why is SVD so useful?
- Application #1: inverses
- $\mathbf{A}^{-1} = (\mathbf{V}^T)^{-1}\,\mathbf{\Sigma}^{-1}\,\mathbf{U}^{-1} = \mathbf{V}\,\mathbf{\Sigma}^{-1}\,\mathbf{U}^T$
  - Using fact that inverse = transpose for orthogonal matrices
  - Since $\mathbf{\Sigma}$ is diagonal, $\mathbf{\Sigma}^{-1}$ also diagonal with reciprocals of entries of $\mathbf{\Sigma}$.

SVD and Inverses

- $\mathbf{A}^{-1}=(\mathbf{V}^T)^{-1}\,\mathbf{\Sigma}^{-1}\,\mathbf{U}^{-1} = \mathbf{V}\,\mathbf{\Sigma}^{-1}\,\mathbf{U}^T$
- This fails when some σ_i are 0
  - It's *supposed* to fail – singular matrix
- Pseudoinverse: if $\sigma_i=0$, set $1/\sigma_i$ to 0 (!)
  - "Closest" matrix to inverse
  - Defined for all (even non-square, singular, etc.) matrices
  - Equal to $(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ if $\mathbf{A}^T\mathbf{A}$ invertible

SVD and Least Squares

- Solving **Ax**=**b** by least squares
- **x**=pseudoinverse(**A**) times **b**
- Compute pseudoinverse using SVD
  - Lets you see if data is singular
  - Even if not singular, ratio of max to min singular values (condition number) tells you how stable the solution will be
  - Set $1/\sigma_i$ to 0 if σ_i is small (even if not exactly 0)

SVD and Eigenvectors

- Let $\mathbf{A}=\mathbf{U\Sigma V}^T$, and let x_i be i^{th} column of $\mathbf{V}$
- Consider $\mathbf{A}^T\mathbf{A}\ x_i$:

$$\mathbf{A}^T\mathbf{A}x_i = \mathbf{V\Sigma}^T\mathbf{U}^T\mathbf{U\Sigma V}^T x_i = \mathbf{V\Sigma}^2\mathbf{V}^T x_i = \mathbf{V\Sigma}^2 \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{V} \begin{pmatrix} 0 \\ \vdots \\ \sigma_i^2 \\ \vdots \\ 0 \end{pmatrix} = \sigma_i^2 x_i$$

- So elements of $\mathbf{\Sigma}$ are sqrt(eigenvalues) and columns of $\mathbf{V}$ are eigenvectors of $\mathbf{A}^T\mathbf{A}$
  - What we wanted for robust least squares fitting!

SVD and Matrix Similarity

- One common definition for the norm of a matrix is the Frobenius norm:

- Frobenius norm can be computed from SVD

$$\|\mathbf{A}\|_F = \sum_i \sum_j a_{ij}^2$$

- So changes to a matrix can be evaluated by looking at changes to singular values

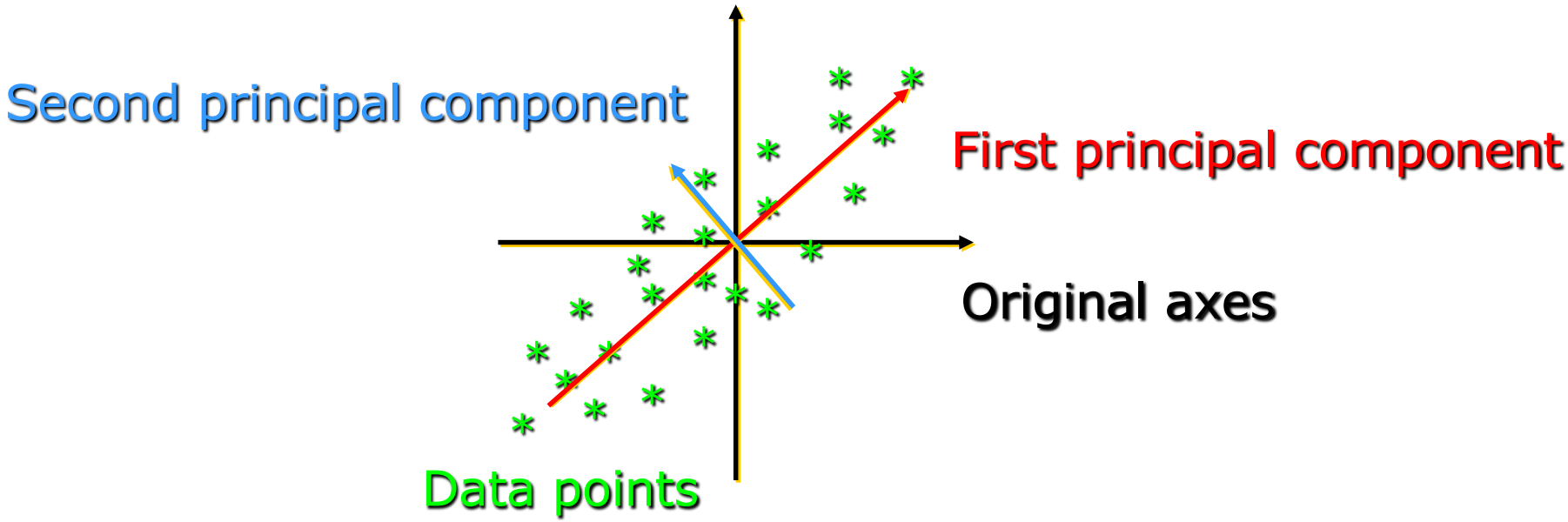
$$\|\mathbf{A}\|_F = \sum_i \sigma_i^2$$

SVD and Matrix Similarity

- Suppose you want to find best rank-*k* approximation to **A**
- Answer: set all but the largest *k* singular values to zero.
- Can form compact representation by eliminating columns of **U** and **V** corresponding to zeroed σ_i.

SVD and PCA

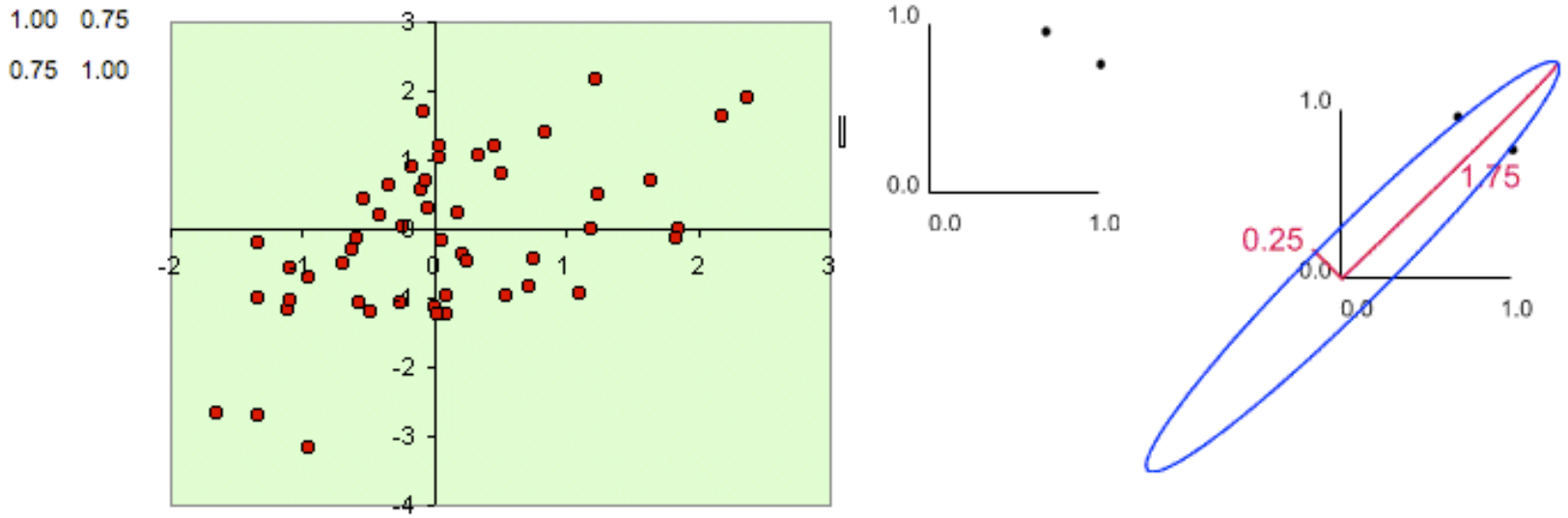
- Principal Components Analysis (PCA): approximating a high-dimensional data set with a lower-dimensional subspace



Physical interpretation

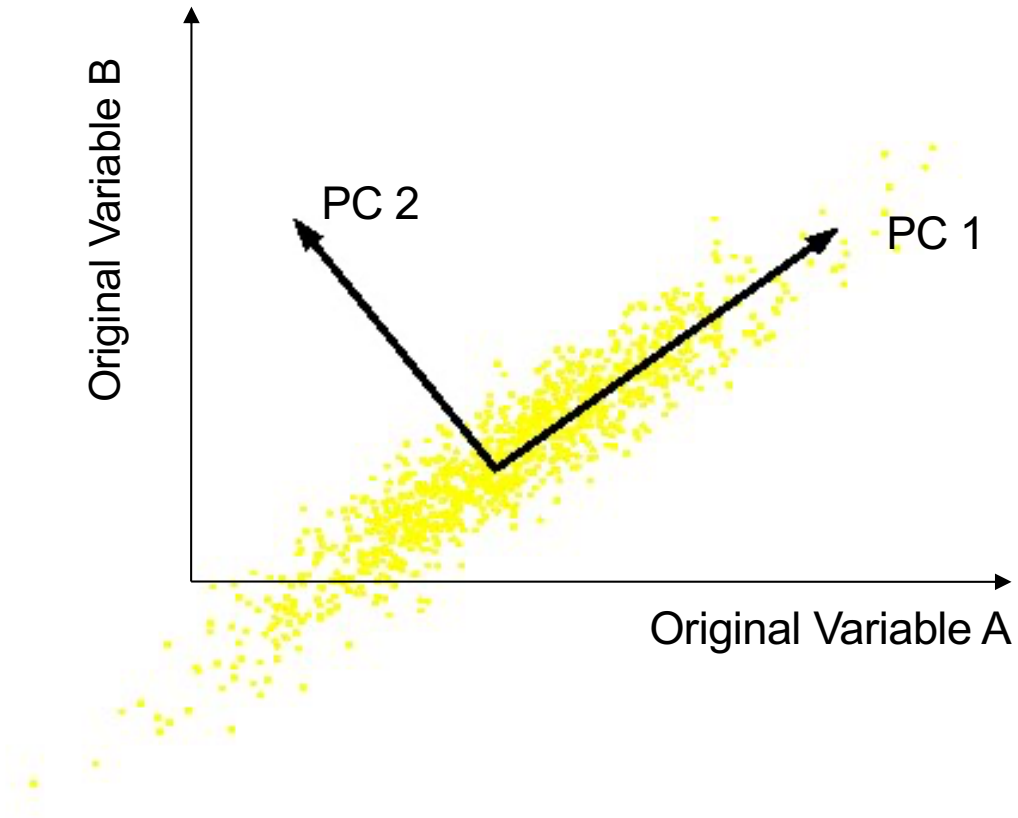
- Consider a covariance matrix, **A**, i.e., $A = 1/n\ S\ S^T$ for some S

$$\mathbf{A} = \begin{bmatrix} 1 & .75 \\ .75 & 1 \end{bmatrix} \Rightarrow \lambda_1 = 1.75, \lambda_2 = 0.25$$



- Error ellipse with the major axis as the larger eigenvalue and the minor axis as the smaller eigenvalue

Physical interpretation



- Orthogonal directions of greatest variance in data
- Projections along PC1 (Principal Component) discriminate the data most along any one axis.

Introduction

- Described by K. Pearson (1901)
- Computing methods by Hotelling (1933)
- Objective
  - To transform the original variables $X_1,\ldots,X_p$ into index variables $Z_1,\ldots,Z_p$
    - $Z_1,\ldots,Z_p$ are linear combinations of $X_1,\ldots,X_p$
    - $Z_1,\ldots,Z_p$ are independent and are in order of important
    - To describe the variation in the data

PCA

- Multiplying each observation vector by an orthogonal matrix A

$$\mathbf{z} = \mathbf{A}\mathbf{x}$$

- The new variables z_1, z_2, … z_p are uncorrelated principal components.
- The sample covariance matrix of z

$$\begin{aligned}\mathrm{cov}(\mathbf{z}) &= \mathrm{E}(\mathbf{z}\mathbf{z}^T) - \mathrm{E}(\mathbf{z})\mathrm{E}(\mathbf{z}^T) \\ &= \mathrm{E}(\mathbf{A}\mathbf{x}\mathbf{x}^T\mathbf{A}^T) - \mathrm{E}(\mathbf{A}\mathbf{x})\mathrm{E}(\mathbf{x}^T\mathbf{A}^T) \\ &= \mathbf{A}\left[\mathrm{E}(\mathbf{x}\mathbf{x}^T) - \mathrm{E}(\mathbf{x})\mathrm{E}(\mathbf{x}^T)\right]\mathbf{A}^T \\ &= \mathbf{A}\mathbf{S}\mathbf{A}^T\end{aligned}$$

$$s_{z_i z_j} = \mathbf{a}_i' \mathbf{S} \mathbf{a}_j = 0 \quad \text{for} \quad i \neq j$$

PCA

$$M = Q\Lambda Q^T$$

- Multiplying each observation vector by an orthogonal matrix A

$$\mathbf{z} = \mathbf{Ax} \qquad \text{cov}(\mathbf{z}) = \mathbf{ASA}^T \qquad \Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} = Q^T \mathbf{M} Q$$

- The new variables z_1, z_2, … z_p are uncorrelated principal components.
- The sample covariance matrix of z

$$\mathbf{S}_\mathbf{z} = \mathbf{ASA}^T = \begin{pmatrix} s_{z_1}^2 & 0 & \cdots & 0 \\ 0 & s_{z_2}^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & s_{z_p}^2 \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} \qquad s_{z_i z_j} = \mathbf{a}_i' \mathbf{S} \mathbf{a}_j = 0 \quad \text{for} \quad i \neq j$$

SVD and PCA

- Data matrix with points as rows, take SVD
  - Subtract out mean ("whitening")
- Columns of $\mathbf{V}_k$ are principal components
- Value of σ_i gives importance of each component

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$$

Symmetric matrix

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{V}\Sigma^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}} = \mathbf{V}\Sigma^2\mathbf{V}^{\mathrm{T}} = \mathbf{V}\Lambda\mathbf{V}^{\mathrm{T}}$$

$\mathbf{S}$

Principal Components Analysis

- Introduction
- Procedures
- Properties
- Examples
- Summary

Principal components

$\mathbf{z} = \mathbf{A}\mathbf{x}$ $z_i = a_i^T \mathbf{x}$

- The eigenvalues of covariance of observations are variances of the principal components.

- "The proportion of variance explained" by the first k components:

$$\text{Proportion of variance} = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\sum_{i=1}^{p} \lambda_i = \sum_{j=1}^{p} s_{jj} = \text{tr}(\mathbf{S})}$$

PC extracted from covariance or correlation matrix?

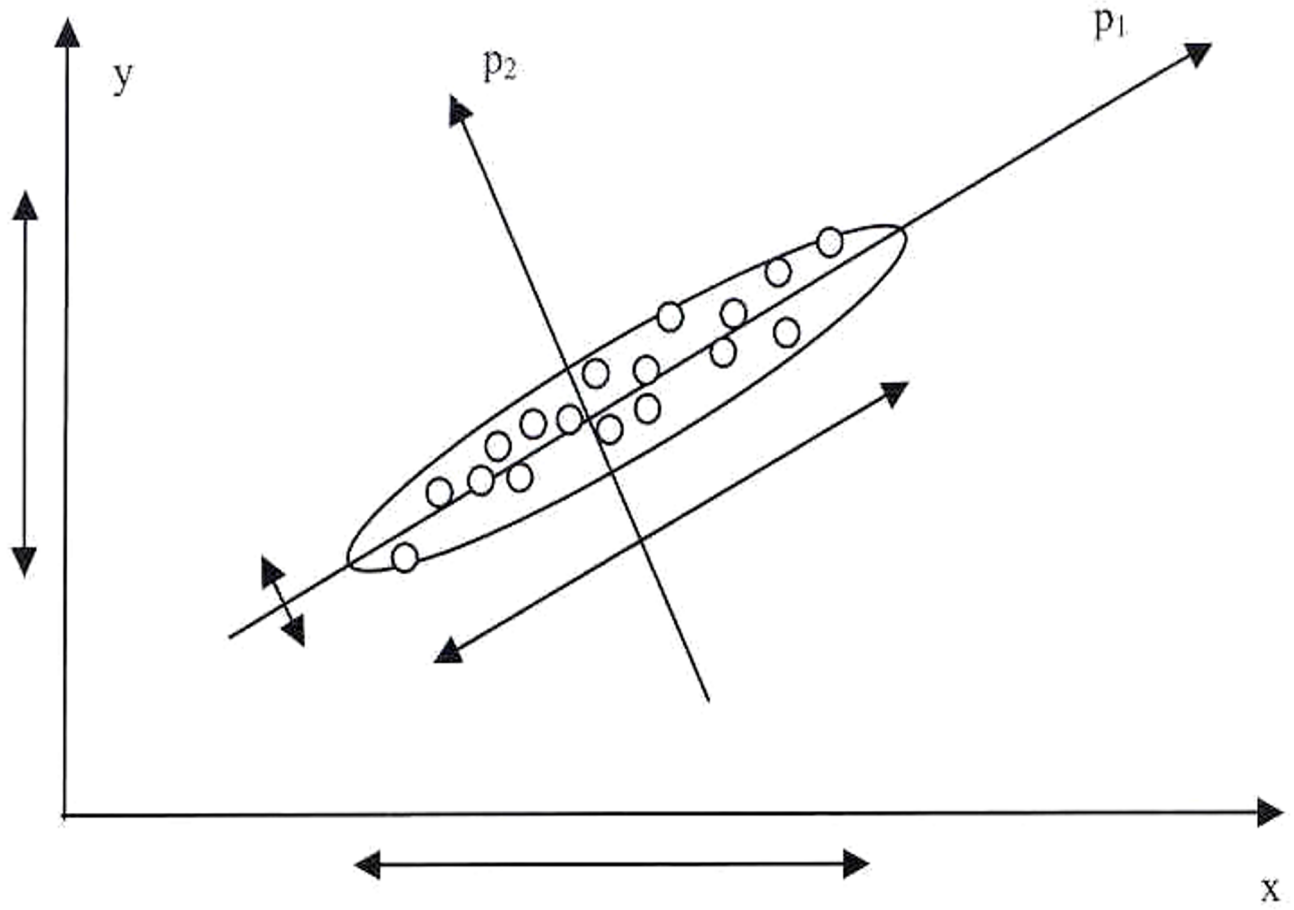
- Principal components analysis is not scale-invariant.
  - Weight (pound), height (ft), age (yr)
  - Weight x16 (ounce), height x12 (inch), age /10 (decade) → PCA
  - Weight (pound), height (ft), age (yr) → PCA

≠

x16 x12 /10

- If there are large differences between the variances of the original variables, then those whose variances are largest will tend to dominate the early components.
- PC should only be extracted from the sample covariance matrix when all the original variables have roughly the same scale.
- Thus, in practice, principal components are extracted from the correlation matrix, R.
  - Extracting the components as the eigenvectors of R is equivalent to calculating the principal components from the original variables after each has been standardized to have unit variance.

Introduction



Introduction

Correlations of Female Sparrows

	X1	X2	X3	X4	X5
Total length (X1)	1.000				
Alar length (X2)	0.399	1.000			
Length of beak and Head (X3)	0.015	0.059	1.000		
Length of humerus (X4)	0.247	0.590	-0.038	1.000	
Length of keel of sternum (X5)	0.137	0.323	-0.130	0.320	1.000

```
## Load Bumpus sparrow Data
bird<-read.table("bumpus.dat", header=TRUE); head(bird)
# length, alar, lbh, lhum, lkeel;  # col 2, 3, 5, 6, 10
X1_5<-c(2:3,5:6,10);
cor(bird[,X1_5])
```

Covariance of bumpus sparrows

```
cov(bird[,X1_5])
```

	length	alar	lbh	lhum	lkeel
length	8.581				
alar	4.454	14.527			
lbh	0.029	0.150	0.440		
lhum	0.018	0.056	-0.001	0.001	
lkeel	0.014	0.043	-0.003	0.000	0.001

```
cor(bird[,X1_5])
```

	1.000				
alar	0.399	1.000			
lbh	0.015	0.059	1.000		
lhum	0.247	0.590	-0.038	1.000	
lkeel	0.137	0.323	-0.130	0.320	1.000

Principal components

```
>result<- eigen(cor(bird[,X1_5]))
>round(result$values, digits=3)
[1] 2.050 1.077 0.822 0.677 0.376
>round(result$vectors,digits=3)
```

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	-0.410	-0.253	0.756	0.380	0.229
[2,]	-0.594	-0.164	-0.036	-0.223	-0.755
[3,]	0.032	-0.866	-0.426	0.235	0.115
[4,]	-0.557	0.018	-0.211	-0.533	0.601
[5,]	-0.410	0.399	-0.450	0.683	0.056

```
> prcomp(bird[,X1_5],scale=TRUE)
```

	PC1	PC2	PC3	PC4	PC5
length	0.410	0.253	-0.755	0.380	0.229
alar	0.594	0.164	0.036	-0.223	-0.755
lbh	-0.032	0.866	0.426	0.235	0.115
lhum	0.557	-0.018	0.211	-0.533	0.601
lkeel	0.410	-0.399	0.450	0.683	0.056

```
> bird_pca<-princomp(bird[,X1_5],cor=TRUE)
> summary(bird_pca, loadings=TRUE)
```

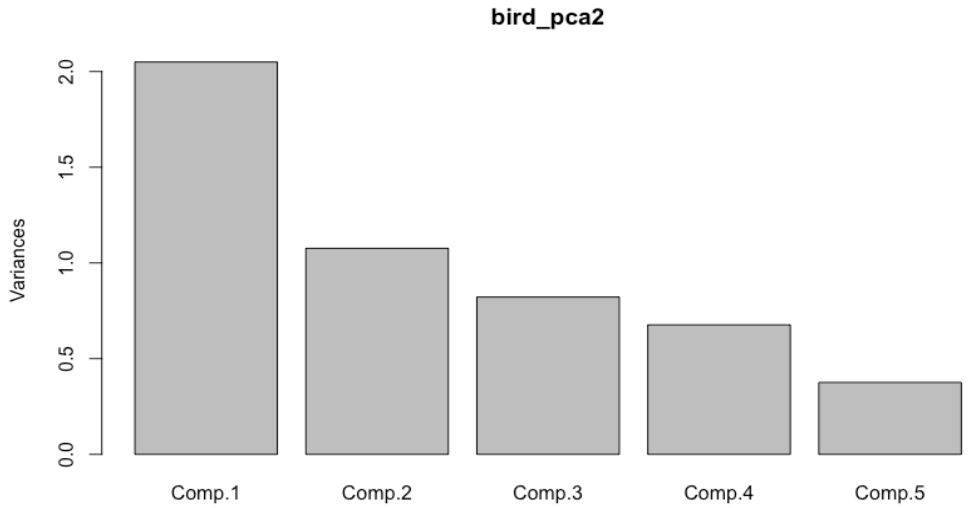
Importance of components:

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.4316	1.0376	0.9065	0.8225	0.61284
Proportion of Variance	0.4099	0.2153	0.1643	0.1353	0.07511
Cumulative Proportion	0.4099	0.6252	0.7896	0.9249	1.00000

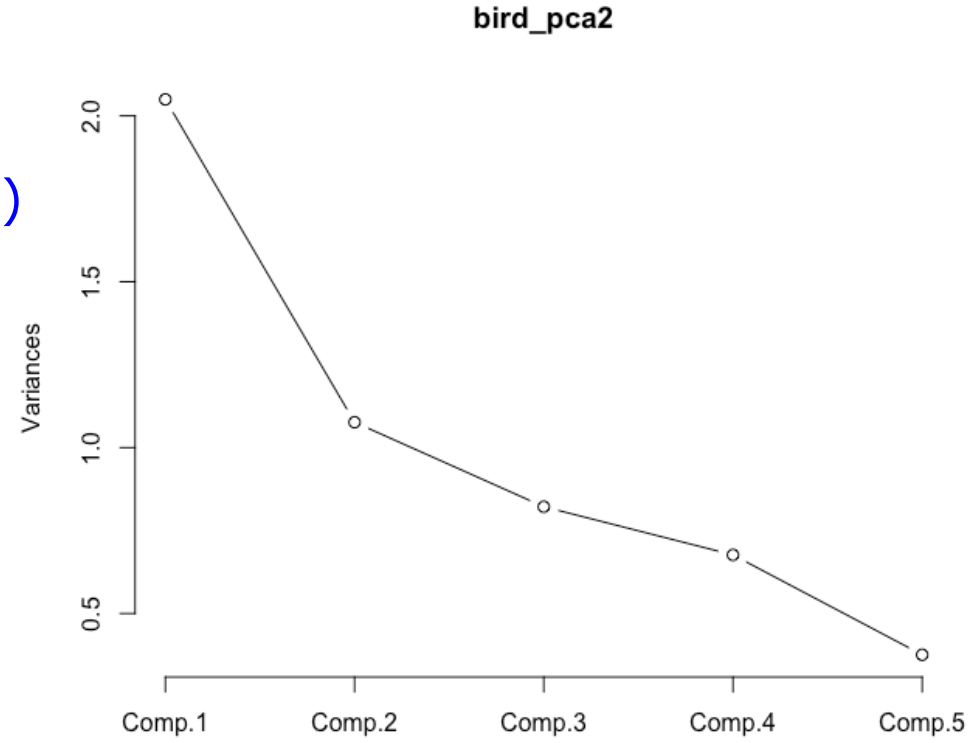
Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
length	0.410	0.253	0.756	0.380	0.229
alar	0.594	0.164		-0.223	-0.755
lbh	0.866	-0.426	0.235	0.115	
lhum	0.557		-0.211	-0.533	0.601
lkeel	0.410	-0.399	-0.450	0.683	

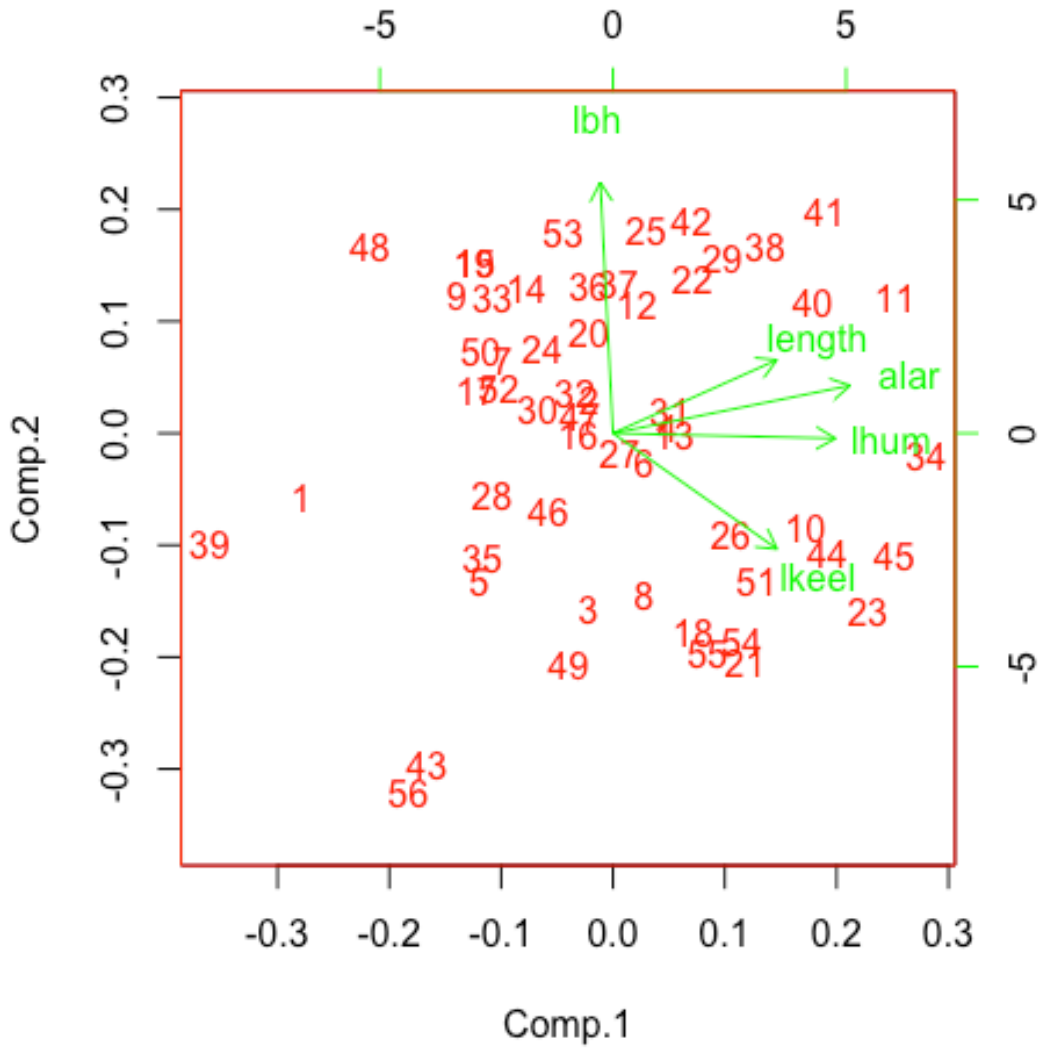
>plot(bird_pca2)



> plot(bird_pca2, type="lines")
> screeplot(bird_pca2, type="lines")



>biplot(bird_pca2,col=c("red","green"))



Introduction

$$Z_1=0.410X_1+0.594X_2-0.032X_3+0.557X_4+0.410X_5$$

- Variance of Z_1 is 2.05
- Variance of Z_1 accounts for 40.99% (2.05/5.00) of the total variation
- All coefficients of Z_1 are smaller than 1 and sum of squares of these coefficients is equal to 1
- Z_1 is in fact as the average (or sum) of X_1, X_2, X_3, X_4, and X_5
- Z_1 can be interpreted as the index for the size of the sparrow

Procedures

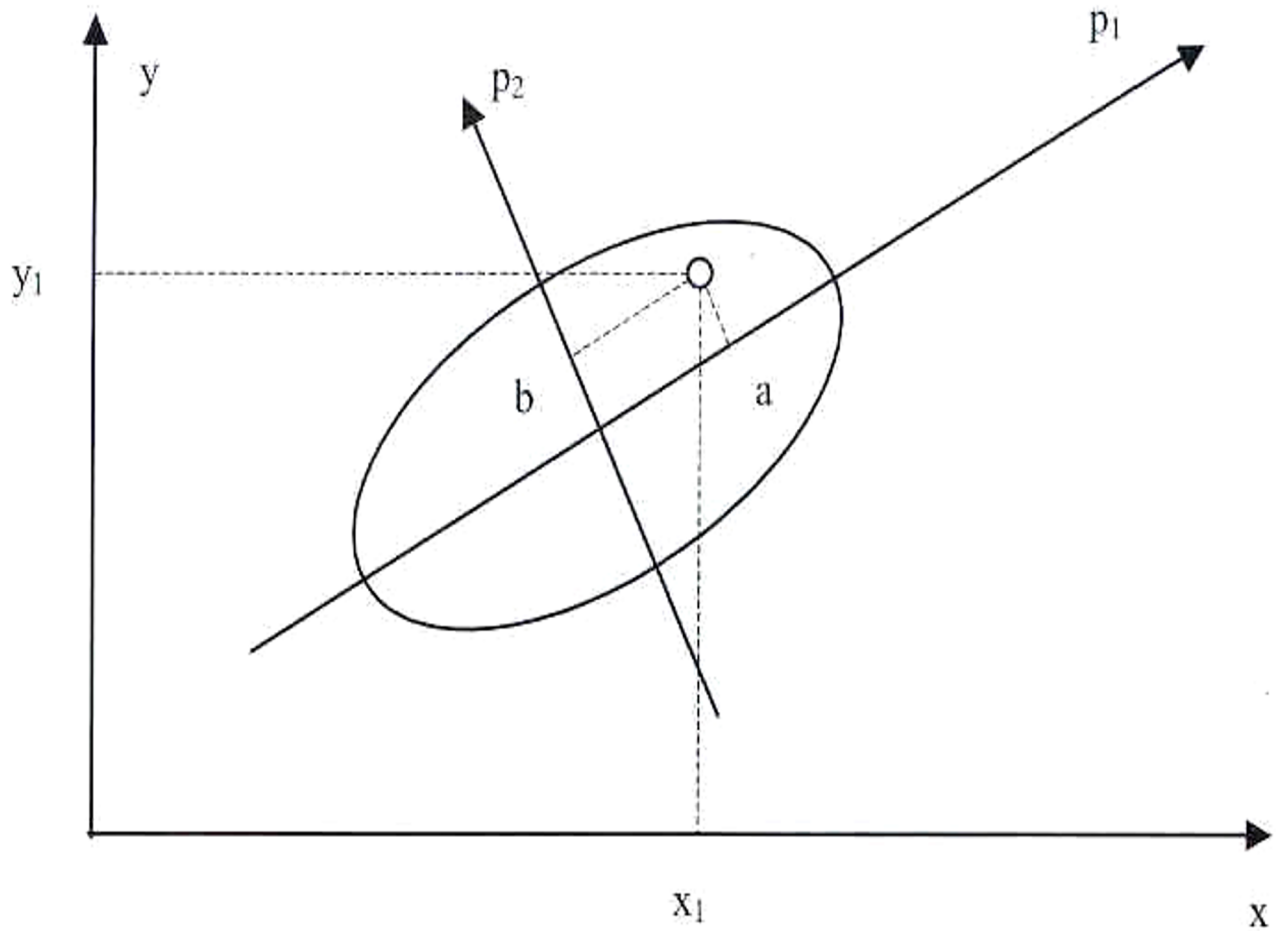
	Data Structure			
Case	X_1	X_2	...	X_p
1	x_{11}	x_{12}	...	x_{1p}
2	x_{21}	x_{22}	...	x_{2p}
.				
.				
N	x_{n1}	x_{n2}	...	x_{np}

Procedures

- The First Component
  - The first component is a linear combination of $X_1, X_2, \ldots, X_p$
  - $Z_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$
  - $\mathrm{Var}(Z_1)$ is as large as possible subject to condition that

$$a_{11}^2 + a_{12}^2 + \ldots + a_{1p}^2 = 1$$

Procedures



Procedures

- The second Component
  - The second component is also a linear combination of $X_1, X_2, \ldots,$ and X_p
  - $Z_1 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$
  - $\mathrm{Var}(Z_2)$ is as large as possible subject to condition that

$$a_{21}^2 + a_{22}^2 + \ldots + a_{2p}^2 = 1,$$

$\mathrm{Var}(Z_2)$ is the second largest,
Z_1 and Z_2 are not correlated

Procedures

- The third Component
  - The third component is also a linear combination of X_1, X_2, ..., and X_p
  - $Z_1 = a_{31}X_1 + a_{32}X_2 + \ldots + a_{3p}X_p$
  - $\mathrm{Var}(Z_2)$ is as large as possible subject to condition that

$$a_{31}^2 + a_{32}^2 + \ldots + a_{3p}^2 = 1,$$

$\mathrm{Var}(Z_3)$ is the second largest,

Z_1, Z_2 and Z_3 are not correlated

Procedures

- Continue until all p principal components are computed
- Covariance matrix of p variables

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{12} & c_{22} & \cdots & c_{2p} \\ . & . & \cdots & . \\ c_{1p} & c_{2p} & \cdots & c_{pp} \end{pmatrix}$$

Procedures

$\mathbf{C}\mathbf{a}_i = \lambda_i \mathbf{a}_i$

or

$(\mathbf{C} - \lambda_i \mathbf{I}) = \mathbf{a}_i$

λ_i is the eigenvalues of $\mathbf{C}$ and $\mathbf{a}_i$ is corresponding eigenvector.

Since $\mathbf{C}$ is a symmetric covariance matric, all eignvalues are nonnegative.

Procedures

$\mathbf{a}_i$ is the vector of the normalized coefficients for the ith principal components.

λ_i is the variance of Z_i.

The sum of eigenvalues = the sum of variances of X_i

$$\lambda_i + \lambda_2 + \ldots + \lambda_p = c_{11} + c_{22} + \ldots + c_{pp}$$

Procedures

- Different variables might have different units and magnitudes.
- PCA might be influenced by these magnitudes and units.
- Standardization to have zero mean and unit variance.
- Covariance on standardized variables is the correlation matrix.

Procedures

- Steps of (PCA)
  - Standardizing variables X_1, X_2,...,X_p to have zero means and unit variances unless that the importance of variables is reflected in their variances.
  - Calculate the covariance matrix (correlation matrix).

Procedures

- Steps of (PCA)
  - Find the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ and their corresponding eigenvectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$.
  - The coefficients of the ith principal component Z_i is the element of $\mathbf{a}_i$ and λ_i the variance of Z_i.
  - Discard any components that accounts for only a small proportion of the variation in the data.

Properties

PCA is a linear combination of $X_1,...,X_p$

that can be represented as

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ . \\ Z_p \end{pmatrix} = \begin{pmatrix} a_{11}\ a_{12}\ ...\ a_{1p} \\ a_{21}\ a_{22}\ ...\ a_{2p} \\ ..\quad ..\quad ...\ .. \\ a_{p1}\ a_{p2}\ ...\ a_{pp} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ . \\ X_p \end{pmatrix}$$

1st eigenvector of S
2nd eigenvector of S

$$= \begin{pmatrix} \mathbf{a_1'} \\ . \\ \mathbf{a_p'} \end{pmatrix} \mathbf{X}$$

$$= \mathbf{AX}$$

$$z_i = a_i' \mathbf{X}$$

Examples

- Determination of the number of principal components
  - Depends upon the needs of practitioners
  - The proportion of the total variation explained by the selected principal components is high, e.g., at least 80%
  - If correlation matrix is used, select the principal component with the variance greater than 1 because they accounts for more variation than the original variables (=1)
  - Use scree plot

Examples

- Evaluation of Statistics Course
  - 16 students for 11 items (variables)
  - Evaluation scales: 1(poor or not at all) to 5(excellent, strongly, or difficult)
  - The first two principal components explain 76.0% of total variation and the last four principal components explain only 2.2%

Examples

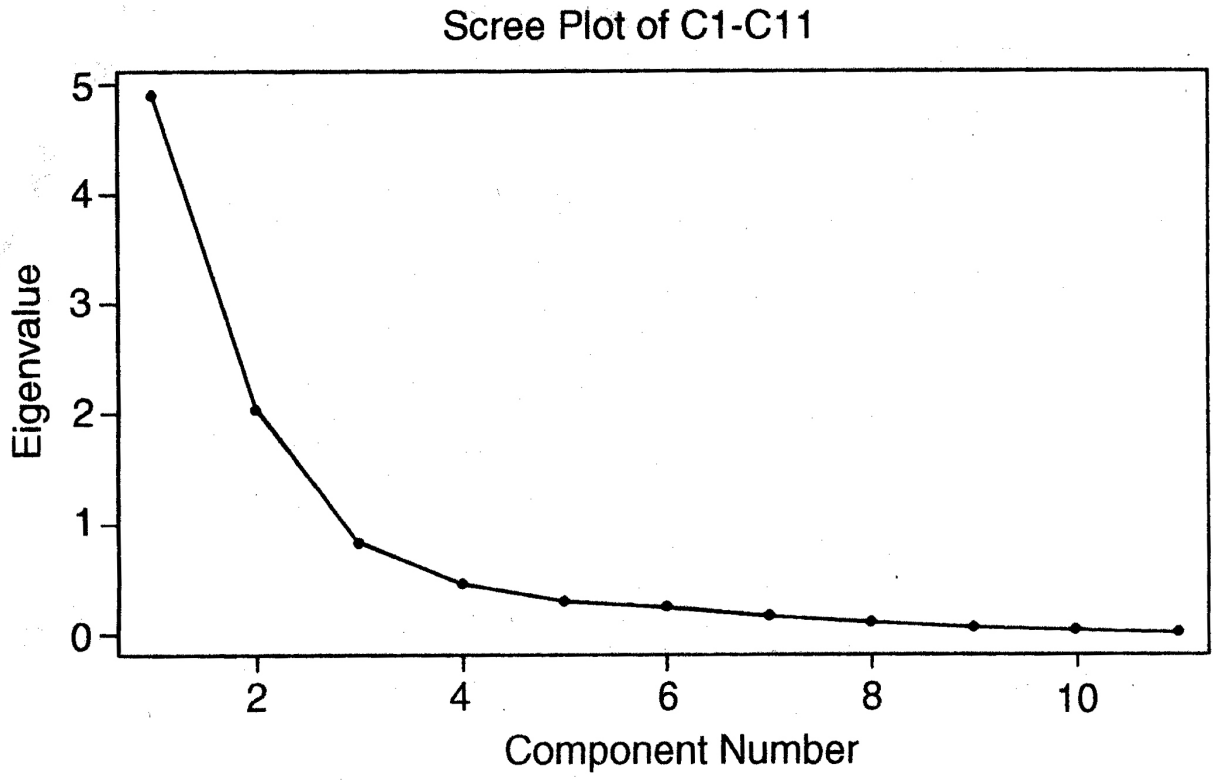


FIGURE 6.1 Scree plot of the course evaluation component variances

Examples

- Test scores of 10 students in 4 subjects

	Student									
Subject	1	2	3	4	5	6	7	8	9	10
Chinese (X1)	85	90	60	70	68	77	50	80	85	55
English (X2)	76	95	45	65	56	80	30	70	75	60
Math (X3)	60	80	38	60	70	65	40	60	65	40
Social (X4)	85	72	80	76	70	68	80	66	84	50

Source: Shen (1998)

```
#### Test scores of 10 students in 4 subjects
Ch<-c(85,90,60,70,68,77,50,80,85,55);
En<-c(76,95,45,65,56,80,30,70,75,60);
Math<-c(60,80,38,60,70,65,40,60,65,40);
Social<-c(85,72,80,76,70,68,80,66,84,50);
student<-data.frame(Ch,En,Math,Social)
cor(student)
eigen(cor(student))
stu_pca2<-princomp(student,cor=TRUE);
summary(stu_pca2,loadings=TRUE);
plot(stu_pca2, type="lines")

zdat<-scale(as.matrix(student));
pca.score<- zdat %*% (stu_pca2$loadings);
cor(as.matrix(student)[,1:4],pca.score[,1])
```

Examples

Correlation Matrix

	X1	X2	X3	X4
X1	1	0.8999	0.8375	0.2784
X2		1	0.7986	-0.0808
X3			1	0.1118
X4				1

Examples

- Eigenvalues and Eigenvectors

Eigenvalue	Prop.	Cum. Prop.	Eigenvector X1	X2	X3	X4
2.71055	0.6776	0.6776	0.5915	0.5684	0.5619	0.1061
1.04821	0.2661	0.9397	0.1100	-0.2522	-0.0420	0.9605
0.21014	0.0525	0.9922	-0.3282	-0.4624	0.8223	-0.0479
0.03110	0.0078	1.0000	-0.7283	0.6320	0.0794	0.2528

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4
Ch	0.591	0.110	0.328	0.728
En	0.568	-0.252	0.462	-0.632
Math	0.562		-0.822	
Social	0.106	0.960		-0.253

Examples

- Because the first two principal components account for 94.0%, we can just use these two principal components.
- The first principal component can be interpreted as the index for the sum of Chinese, English and math.
- The second principal component can be thought as social science.

Examples

- The above results can be also obtained by inspecting the correlation matrix.
- Correlations among Chinese, English, and math exceed 0.8.
- Correlations between Chinese, English, and math with social science are below 0.3.

Examples

Student	Comp 1	Comp 2	Comp3	Comp 4
[1,]	1.10066	1.04367	0.50666	0.02364
[2,]	2.56440	-0.42753	-0.12743	-0.15592
[3,]	-1.85671	0.87063	0.39681	-0.00554
[4,]	0.02469	0.24637	-0.16806	-0.18204
[5,]	0.00078	-0.22807	-1.05145	0.10661
[6,]	0.90464	-0.65066	0.04476	-0.15531
[7,]	-2.66813	0.98821	-0.33322	-0.03794
[8,]	0.50822	-0.65943	0.15031	0.42119
[9,]	1.25936	0.95055	0.18527	0.05365
[10,]	-1.83789	-2.13374	0.39635	-0.06835

Examples

Correlations of Female Sparrows

	X1	X2	X3	X4	X5
length	1.000				
alar	0.399	1.000			
lbh	0.015	0.059	1.000		
lhum	0.247	0.590	-0.038	1.000	
lkeel	0.137	0.323	-0.130	0.320	1.000

Examples

	PC1	PC2	PC3	PC4	PC5
length	0.410	0.253	-0.755	0.380	0.229
alar	0.594	0.164	0.036	-0.223	-0.755
lbh	-0.032	0.866	0.426	0.235	0.115
lhum	0.557	-0.018	0.211	-0.533	0.601
lkeel	0.410	-0.399	0.450	0.683	0.056
variance	2.050	1.077	0.822	0.677	0.376

Summary

- A linear combination of the original variables
- Try to reduce a large number of variables to a few index variables
- Index variables are not correlated and ordered in the magnitude of variation
- Illustration with real examples

PCA on Faces: "Eigenfaces"