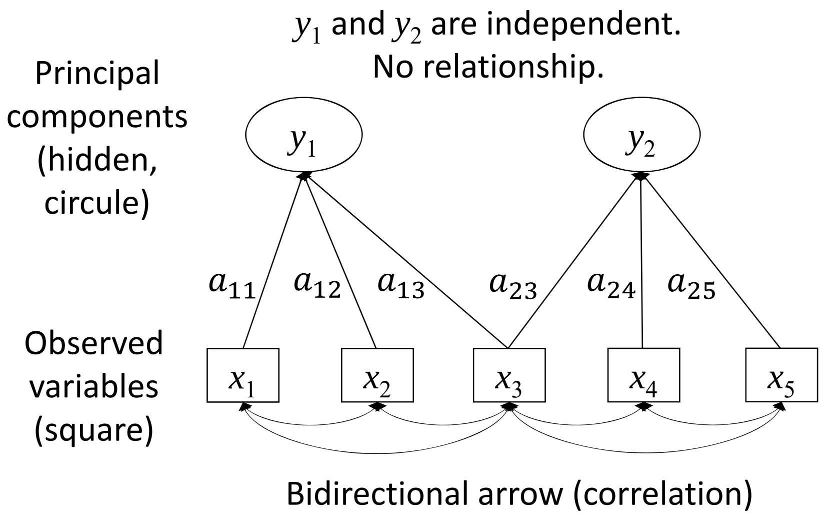
2 Principal component analysis

主成分分析

2.1 Objectives

- Often referred to as PCA.
- When multiple variables are recorded in a certain system, PCA is to find groups of variables that covary (共変する).
- Variables in a group are correlated with each other.
- Groups are independent from each other.
- Reduction of dimensions
  - A number of correlated variables can be summarized into a smaller number of independent groups. PCA can be used to reduce the number of variables.
- Variables in the same group produce a new imaginary variable called principal component. Principal components are hidden variables, which cannot be directly sampled.

2.1 Objectives



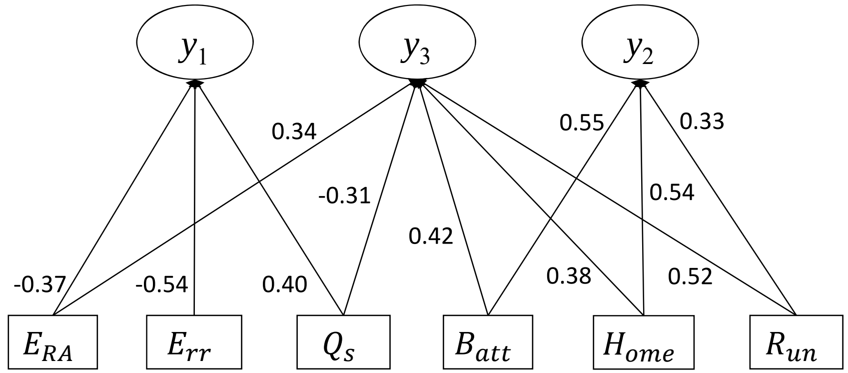
2.1 Objectives

- Reduction of variables or dimensions help interpret the data.
- PCA is used as for pre-processing before another analysis because it provides a smaller number of independent variables, which is convenient for some analytical methods.

2.2 Example of baseball data

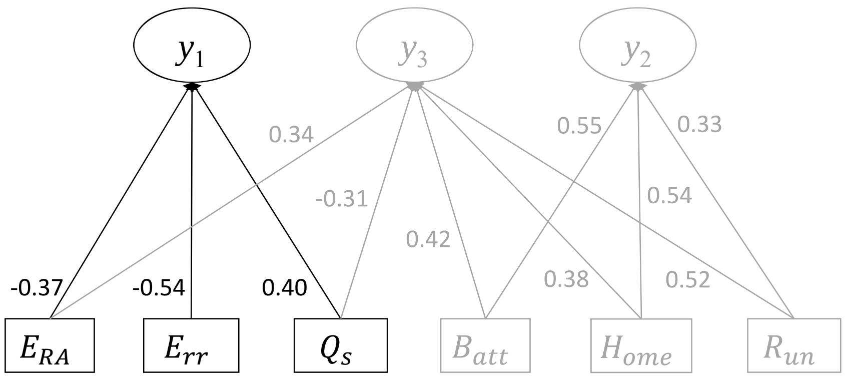
- Six variables relating to the annual performance of baseball clubs are analyzed by PCA.
  - ERA: Earned Run Average
  - Batting average
  - Homerun
  - Run
  - Errors
  - QS: Quality start … Number of games at which a pitcher permits less than 4 runs within the first six innings.

2.2 Example of baseball data



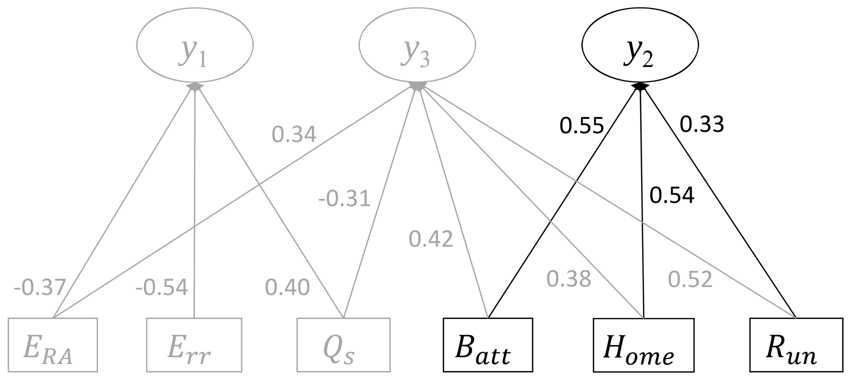
Three components are found from 6 variables.

2.2 Example of baseball data



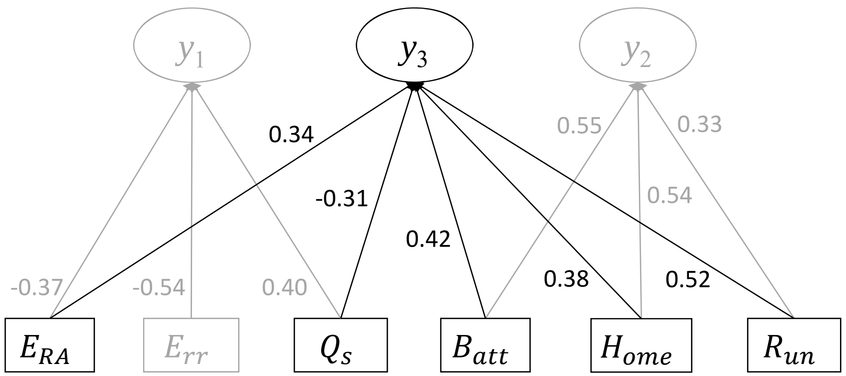
- 1st component (y_1): Defensive component
  - Greater y_1 value indicates better defensive performance of the club.

2.2 Example of baseball data



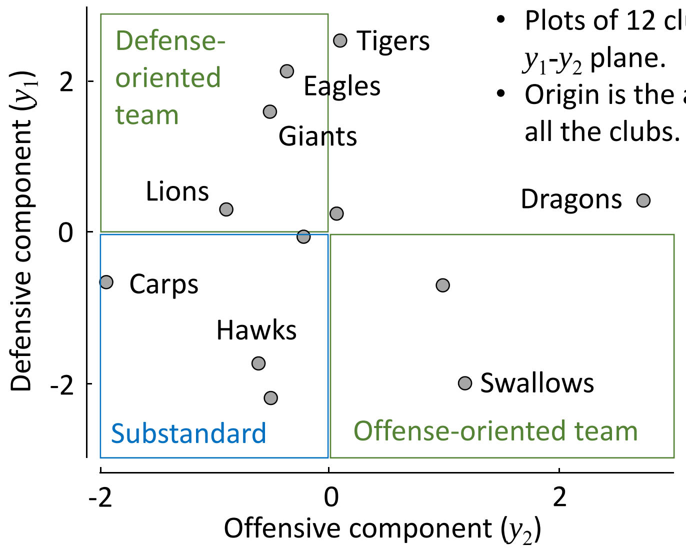
- 2nd component (y_2): Offensive component
  - Greater y_2 value indicates better offensive performance.

2.2 Example of baseball data



- 3rd component (y_3): Component of trade-off
  - Greater y_3 value indicates the club is more offensive but less defensive.

2.2 Example of baseball data



- Plots of 12 clubs on the y_1-y_2 plane.
- Origin is the average of all the clubs.

2.2 Example of baseball data

- 1. Defensive component: Defense ability
- 2. Offensive component: Offense ability
- 3. Trade-off component:

- 1st and 2nd components independently determine the defensive and offensive abilities of club teams, respectively.
- 3rd trade-off component is also independent from the others. It indicates that all the teams suffer from the imbalance of offensive and defensive abilities.

2.3 Covariance matrix

2.3.1 Covariance (共分散)

- Covariance indicates to what extent two variables covary and is defined by

$$\mathrm{cov}(x_1, x_2) = \frac{1}{n}\sum_{i=1}^{n}(x_{1i} - \overline{x_1})(x_{2i} - \overline{x_2}) \qquad (2.3.1)$$

- If the covariance is positive, one value (x_1) tends to increase when the other value (x_2) increases.
- If the covariance is negative, two values are likely to change in opposite directions.

2.3.2 Covariance matrix (共分散行列)

- Covariance values among multiple (p) variables are expressed in the form of matrix. It is $p \times p$ matrix of which cells are covariance values of two variables.

$$\boldsymbol{S} = \begin{bmatrix} \mathrm{cov}(x_1, x_1) & \dots & \mathrm{cov}(x_1, x_k) & \dots & \mathrm{cov}(x_1, x_p) \\ \vdots & & \vdots & & \vdots \\ \mathrm{cov}(x_k, x_1) & \ddots & \mathrm{cov}(x_k, x_k) & \ddots & \mathrm{cov}(x_k, x_p) \\ \vdots & & \vdots & & \vdots \\ \mathrm{cov}(x_p, x_1) & \dots & \mathrm{cov}(x_p, x_k) & \vdots & \mathrm{cov}(x_p, x_p) \end{bmatrix}$$

(2.3.2)

2.3.2 Covariance matrix

- A covariance matrix is symmetric, and usually only the upper triangular part is written.

$$\mathrm{cov}(x_1, x_2) = \mathrm{cov}(x_2, x_1) \qquad (2.3.3)$$

- Diagonal cells of a covariance matrix are variances of variables.

$$\mathrm{cov}(x_k, x_k) = \mathrm{var}(x_k) = \frac{1}{n}\sum_{i=1}^{n}(x_{ki} - \overline{x_k})^2 \qquad (2.3.4)$$

2.3.3 Normalization (Z-value)

- The multiple regression analyses often address multiple variables of different quantities or dimensions.
- We prefer dimension-less data. For example, we reach different results whether we use meter or centimeter as a unit of length.
- To compare variables by removing the effects of dimensions, normalization is used.
- The normalized value of variable x is defined as follows by using its variance and mean.

$$z_i = \frac{x_i - \bar{x}}{\sqrt{\mathrm{var}(x_i)}} \qquad (2.3.5)$$

2.3.3 Normalization (Z-value)

- The mean and variance of normalized variable are always 0 and 1, respectively.

$$\bar{z} = 0 \qquad (2.3.6)$$

$$\mathrm{var}(z) = 1 \qquad (2.3.7)$$

2.3.4 Correlation coefficient

- The correlation coefficient resembles covariance. It is a normalized value of covariance such that the value fits into [-1, 1].

$$\mathrm{corr}(x_1, x_2) = \frac{1}{n}\sum_{i=1}^{n} \frac{(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sqrt{\mathrm{var}(x_1)\mathrm{var}(x_2)}} \qquad (2.6)$$

- When $\mathrm{cov}(x_1, x_2) > 0$, the correlation coefficient is also positive: $\mathrm{corr}(x_1, x_2) > 0$.

2.3.4 Correlation coefficient

- The correlation coefficient of two variables is identical to the covariance of two normalized variables.

$$\begin{aligned}\mathrm{corr}(x_1, x_2) &= \frac{1}{n}\sum_{i=1}^{n} \frac{(x_{1i} - \overline{x_1})(x_{2i} - \overline{x_2})}{\sqrt{\mathrm{var}(x_1)\mathrm{var}(x_2)}} \qquad (2.7)\\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{(x_{1i} - \overline{x_1})}{\sqrt{\mathrm{var}(x_1)}}\frac{(x_{2i} - \overline{x_2})}{\sqrt{\mathrm{var}(x_2)}}\\ &= \frac{1}{n}\sum_{i=1}^{n} z_{1i}z_{2i}\\ &= \mathrm{cov}(z_1, z_2)\end{aligned}$$

2.3.5 Correlation matrix

- Similar to the covariance matrix, the correlation matrix consists of correlation coefficients of all pairs of related variables.

$$\boldsymbol{C} = \begin{bmatrix} \mathrm{corr}(x_1, x_1) & \dots & \mathrm{corr}(x_1, x_k) & \dots & \mathrm{corr}(x_1, x_p) \\ \vdots & & \vdots & & \vdots \\ \mathrm{corr}(x_k, x_1) & \ddots & \mathrm{corr}(x_k, x_k) & \ddots & \mathrm{corr}(x_k, x_p) \\ \vdots & & \vdots & & \vdots \\ \mathrm{corr}(x_p, x_1) & \dots & \mathrm{corr}(x_p, x_k) & \vdots & \mathrm{corr}(x_p, x_p) \end{bmatrix}$$

(2.8)

2.3.5 Correlation matrix

$$= \begin{bmatrix} 1 & \dots & \mathrm{corr}(x_1, x_k) & \dots & \mathrm{corr}(x_1, x_p) \\ \vdots & & \vdots & & \vdots \\ \mathrm{corr}(x_k, x_1) & \ddots & 1 & \ddots & \mathrm{corr}(x_k, x_p) \\ \vdots & & \vdots & & \vdots \\ \mathrm{corr}(x_p, x_1) & \dots & \mathrm{corr}(x_p, x_k) & \vdots & 1 \end{bmatrix}$$

Mathematical review II:
Eigen value expansion (固有値展開)

- Suppose we have a square matrix $\boldsymbol{A} \in \mathbb{R}^{p \times p}$, it can be expressed by a product of a few matrices. The eigen value expansion is one of such methods.

$$\boldsymbol{A} = \left[\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_p\right]^T \qquad \mathrm{Rank}(\boldsymbol{A}) = p \qquad \boldsymbol{a}_i \in \mathbb{R}^{p \times 1}$$

- There exists a vector $y_i \in \mathbb{R}^{p \times 1}$ and scalor λ_i that satisfy

$$\boldsymbol{A}\boldsymbol{y}_i = \lambda_i \boldsymbol{y}_i \qquad (2.9)$$

- $\boldsymbol{y}_i$ and λ_i are called the eigen vector and eigen value. There exist p pairs of such vector and value.

Mathematical review II:
Eigen value expansion (固有値展開)

$$\begin{aligned} \boldsymbol{A}\boldsymbol{y}_1 &= \lambda_1 \boldsymbol{y}_1 \\ \boldsymbol{A}\boldsymbol{y}_2 &= \lambda_2 \boldsymbol{y}_2 \\ &\vdots \\ \boldsymbol{A}\boldsymbol{y}_p &= \lambda_p \boldsymbol{y}_p \end{aligned}$$

- These p simultaneous equations can be expressed by

$$\boldsymbol{A}[\boldsymbol{y}_1\ \boldsymbol{y}_2 \dots \boldsymbol{y}_p] = [\boldsymbol{y}_1\ \boldsymbol{y}_2 \dots \boldsymbol{y}_p] \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix} \qquad (2.10)$$

Mathematical review II:
Eigen value expansion (固有値展開)

$$\boldsymbol{A}[\boldsymbol{y}_1\ \boldsymbol{y}_2 \dots \boldsymbol{y}_p] = [\boldsymbol{y}_1\ \boldsymbol{y}_2 \dots \boldsymbol{y}_p] \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix}$$

$$\boldsymbol{A}\boldsymbol{Y} = \boldsymbol{Y}\boldsymbol{\Lambda} \qquad (2.11)$$

$$\boldsymbol{Y} = [\boldsymbol{y}_1\ \boldsymbol{y}_2 \dots \boldsymbol{y}_p] \qquad \Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix}$$

Mathematical review II:
Eigen value expansion (固有値展開)

- From (2.11), a square matrix $\boldsymbol{A}$ can be decomposed into the matrix of its eigen vectors and eigen values.

$$\boldsymbol{A} = \boldsymbol{Y}\boldsymbol{\Lambda}\boldsymbol{Y}^{-1} \qquad (2.12)$$

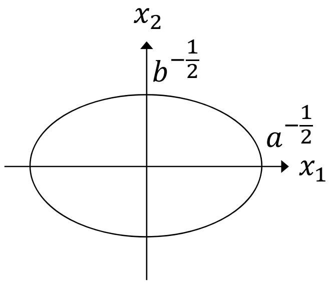
- When $\boldsymbol{A}$ is a symmetric matrix, $\boldsymbol{Y}$ can be an orthonormal matrix whose components are orthogonal (直交):

$$\boldsymbol{y}_i \perp \boldsymbol{y}_j$$

$$|\boldsymbol{y}_i| = 1$$

2.4 Ellipsoid and covariance matrix

- The equation of ellipsoid (2.4.1) is written using vectors and a matrix as (2.4.2), which is called the quadratic form (二次形式).



$$ax_1^2 + bx_2^2 = 1 \qquad (2.4.1)$$

$$[x_1\ x_2]\begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 1 \quad (2.4.2)$$

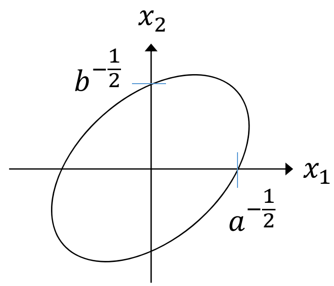
$$\boldsymbol{x}^T\boldsymbol{D}\boldsymbol{x} = 1 \qquad (2.4.3)$$

where

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \text{and} \quad D = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$$

2.4 Ellipsoid and covariance matrix

- More generally, an inclined ellipsoid is (2.4.4)



$$ax_1^2 + 2cx_1x_2 + bx_2^2 = 1 \quad (2.4.4)$$

$$[x_1\ x_2]\begin{bmatrix} a & c \\ c & b \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 1 \quad (2.4.5)$$

$$\boldsymbol{x}^T\boldsymbol{D}\boldsymbol{x} = 1 \qquad (2.4.6)$$

where

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \text{and} \quad D = \begin{bmatrix} a & c \\ c & b \end{bmatrix}$$

2.4 Ellipsoid and covariance matrix

- $\boldsymbol{D}$ matrix of ellipsoid is always symmetric. It is decomposed into orthonormal matrix. Hence, (2.4.6) is identical to (2.4.7).

$$\boldsymbol{x}^T\boldsymbol{Y}\boldsymbol{\Lambda}\boldsymbol{Y}^{-\mathbf{1}}\boldsymbol{x} = 1 \qquad (2.4.7)$$

$$\Downarrow$$

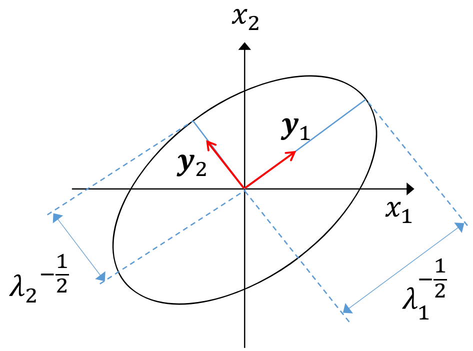
$$(\boldsymbol{Y}^T\boldsymbol{A})^T\boldsymbol{\Lambda}(\boldsymbol{Y}^T\boldsymbol{A})\boldsymbol{x} = 1 \quad (2.4.8)$$

$$\because \quad \boldsymbol{Y}^{-1} = \boldsymbol{Y}^T \qquad (2.4.9)$$

Note (2.4.9) is true only for orthonormal matrices.

2.4 Ellipsoid and covariance matrix

- $\boldsymbol{Y}$ is an orthonormal matrix and rotates the coordinate system of the ellipsoid. Its column vectors are the axes of the ellipsoid, and eigen values are the lengths of corresponding axes.



$$\boldsymbol{Y} = [\boldsymbol{y}_1\ \boldsymbol{y}_2]$$

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

2.4 Ellipsoid and covariance matrix

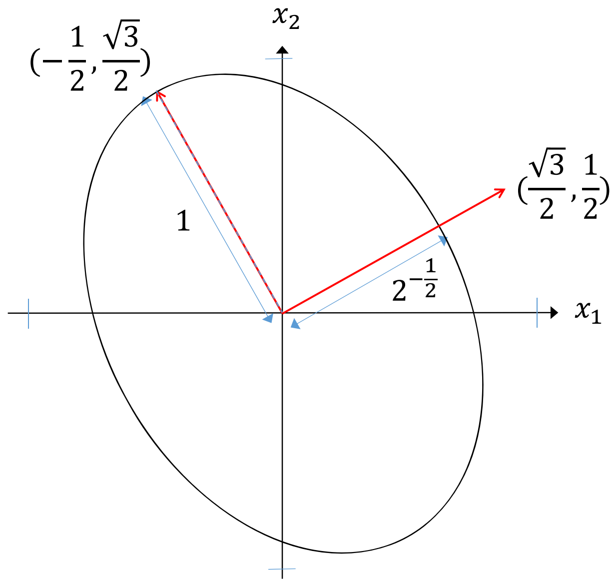
- Let's see a numerical example.

The ellipsoid
$$\frac{7}{8}x_1^2 + 2\frac{\sqrt{3}}{4}x_1x_2 + \frac{5}{4}x_2^2 = 1 \qquad (2.4.10)$$

Quadratic form
$$\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} \frac{7}{8} & \frac{\sqrt{3}}{4} \\ \frac{\sqrt{3}}{4} & \frac{5}{4} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 1 \qquad (2.4.11)$$

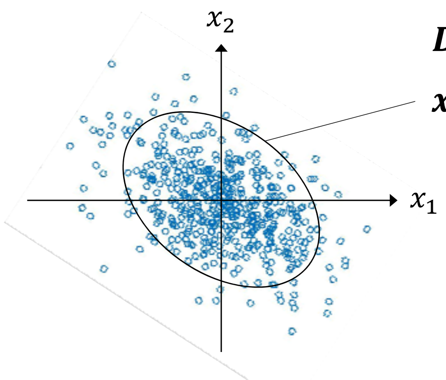
$$\boldsymbol{D} = \begin{bmatrix} \frac{7}{8} & \frac{\sqrt{3}}{4} \\ \frac{\sqrt{3}}{4} & \frac{5}{4} \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{3}}{2} & \frac{-1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix}\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \frac{\sqrt{3}}{2} & \frac{-1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix}^T \qquad (2.4.12)$$

2.4 Ellipsoid and covariance matrix



2.4 Ellipsoid and covariance matrix

- Samples from two Gaussian variables form an ellipsoid, of which $\boldsymbol{D}$ matrix matches the covariance matrix of the two variables. The covariance matrix expresses the distribution of the samples.



$$\boldsymbol{D} = \mathrm{cov}(x_1, x_2)$$
$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{x} = 1$$

2.5 Theory

- A vector $\boldsymbol{x}$ of p variables can be expressed by a linear combination of p independent vectors of the same length.

$$\boldsymbol{x} = b_1\boldsymbol{y}_1 + \cdots + b_k\boldsymbol{y}_k + \cdots + b_p\boldsymbol{y}_p \qquad (2.5.1)$$

$$\boldsymbol{x}, \boldsymbol{y}_k \in \mathbb{R}^{p\times 1}$$
$$\boldsymbol{y}_j \perp \boldsymbol{y}_k$$
$$|\boldsymbol{y}_j| = 1$$

2.5 Theory

- For any i-th zero-centered sample of $\boldsymbol{x}$, (2.5.1) also holds, but with different coefficients b.

$$\boldsymbol{x_i} - \overline{\boldsymbol{x}} = b_{i1}\boldsymbol{y}_1 + \cdots + b_{ik}\boldsymbol{y}_k + \cdots + b_{ip}\boldsymbol{y}_p \quad (2.5.2)$$

- When n samples are provided, the simultaneous equations are

$$\underset{(p \times n)}{[\boldsymbol{x}_1 - \overline{\boldsymbol{x}} \ \dots \boldsymbol{x}_k - \overline{\boldsymbol{x}} \dots \boldsymbol{x}_n - \overline{\boldsymbol{x}}]} = \underset{(p \times p)}{[\boldsymbol{y}_1 \ \dots \boldsymbol{y}_k \ \dots \boldsymbol{y}_p]} \underset{(p \times n)}{\begin{bmatrix} b_{11} & \cdots & b_{k1} & \cdots & b_{n1} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ b_{1k} & \cdots & b_{kk} & \cdots & b_{nk} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ b_{1p} & \cdots & b_{nk} & \cdots & b_{np} \end{bmatrix}} \quad (2.5.3)$$

2.5 Theory

- Using matrix variables, (2.5.3) is

$$\boldsymbol{X}_{\mathrm{c}} = \boldsymbol{YB} \quad (2.5.4)$$

- The product of $\boldsymbol{X}_{\mathrm{c}}$ and $\boldsymbol{X}_{\mathrm{c}}^T$ is

$$\boldsymbol{X}_{\mathrm{c}}\boldsymbol{X}_{\mathrm{c}}^T = \boldsymbol{YBB}^T\boldsymbol{Y}^T \quad (2.5.5)$$

- $\boldsymbol{X}_{\mathrm{c}}\boldsymbol{X}_{\mathrm{c}}^T$ is symmetric, and $\boldsymbol{Y}$ is orthonormal. From (2.12), (2.5.5) is

$$\begin{aligned} \boldsymbol{X}_{\mathrm{c}}\boldsymbol{X}_{\mathrm{c}}^T &= \boldsymbol{YBB}^T\boldsymbol{Y}^T \\ &= \boldsymbol{Y\Lambda Y}^{-1} \quad (2.5.6) \end{aligned}$$

$$\boldsymbol{BB}^T = \boldsymbol{\Lambda} = \mathrm{diag}(\lambda) \quad (2.5.7)$$

2.5 Theory

- From (2.5.7), we know that

$$\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{b}_j = 0 \quad (2.5.8)$$

$$\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{b}_i = \lambda_i \quad (2.5.9)$$

- Also, because $\overline{b_i} = 0$,

$$\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{b}_i = \sum_{j=1}^{n} b_{ji}^2 = n\,\mathrm{var}(b_i) = \lambda_i \quad (2.5.10)$$

2.5 Theory

- The principal component analysis is to find $\boldsymbol{Y'} \in \mathbb{R}^{p\times q}$ $(q < p)$ matrix, with which the matrix of sample vectors is best approximated:

$$\boldsymbol{X}_{\mathrm{c}} \cong \boldsymbol{Y'B'} \quad (2.5.11)$$

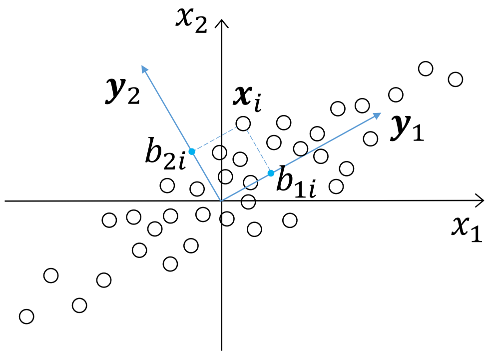
where

$\boldsymbol{Y'} \in \mathbb{R}^{p\times q}$ … matrix of principal component coefficient

$\boldsymbol{B'} \in \mathbb{R}^{q\times n}$ … matrix of principal component score

2.5 Theory

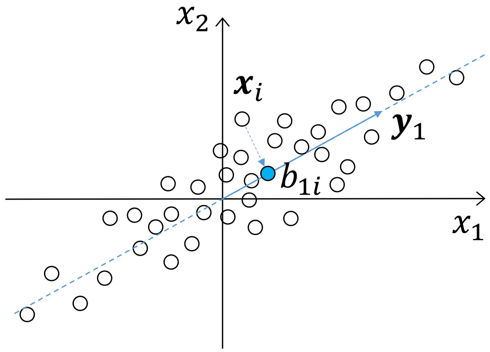
- Graphical understanding of PCA
  - $\boldsymbol{y}_i$ is the vector of the axis of the ellipsoid.
  - b_{ij} is the coordinate of $\boldsymbol{x}_i$ on the ellipsoidal axis $\boldsymbol{y}_i$.



$$\begin{bmatrix} x_{1i} \\ x_{2i} \end{bmatrix} = [\boldsymbol{y}_1 \ \boldsymbol{y}_2] \begin{bmatrix} b_{1i} \\ b_{2i} \end{bmatrix}$$

2.5 Theory

- Using only the major axis ($\boldsymbol{y}_1$), the samples are approximated.



$$\mathrm{var}(b_1) > \mathrm{var}(b_2)$$

$$\lambda_1 > \lambda_2$$

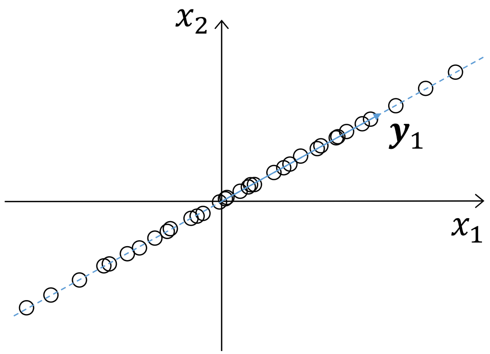
$$\begin{bmatrix} x_{1i} \\ x_{2i} \end{bmatrix} = [\boldsymbol{y}_1 \ \boldsymbol{y}_2] \begin{bmatrix} b_{1i} \\ b_{2i} \end{bmatrix}$$

Approximation by loosing the information along $\boldsymbol{y}_2$.

$$\begin{bmatrix} x_{1i} \\ x_{2i} \end{bmatrix} \cong [\boldsymbol{y}_1][b_{1i}]$$

2.5 Theory

- After approximation, the samples are solely expressed on the principal component y_1.



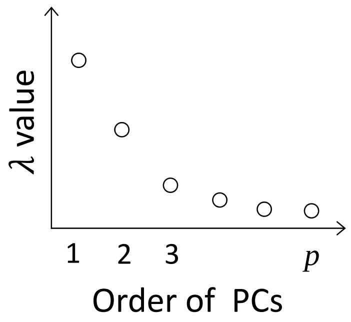
- The PCA is to find the principal components along witch the sample variances are large.
- Even after ignoring minor components, of which λ values are small, the large part of properties of samples are remained.

2.6 Criteria of principal components

- There are some methods to determine how many principal components should be considered. Here, popular two methods are introduced:
  - Scree diagram
  - Contribution ratio.
- Both methods are based on the λ values of PCs.

2.6.1 Scree diagram

- When the λ values are plot in descending order, their profiles typically appear to be a scree, which is a rocky slope of mountain.
- Based on its shape, PCs on a rapid slope are selected, and other small components on the moderate slope are truncated.



- If the scree diagram looks like this, then typically the first two components are used.

2.6.2 Contribution ratio

- The contribution ratio (寄与率) of a principal component is defined as the proportion its λ value in the sum of all λ values.
- Contribution ratio of i-th component

$$\frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} \in [0,1] \qquad (2.6.1)$$

- The numerator corresponds to the variance of i-th component
- The denominator corresponds to the sum of the variances along all the components.

2.6.2 Contribution ratio

- The number of PCs is determined such that the cumulative ratio of the PCs exceeds for example 0.9.
- We then may say that more than 90% of the variance of all samples are expressed by PCs.
- The greater cumulative ratio leads to more adequate approximation in exchange of variable reduction.
- In the baseball example, the three PCs account for 82% of the data variance.
  - 1^{st} comp: 0.33
  - 2^{nd} comp: 0.30
  - 3^{rd} comp: 0.17
  - 4^{th} comp: 0.09 (not employed)

Mathematical review III:

Trace and determinant of matrices

- Determinant of a square matrix $A = \{a_{ij}: i, j = 1, \dots, n\}$ is the product of its eigen values.

$$\det(\boldsymbol{A}) = \prod \lambda_i$$

- Trace of a square matrix $A = \{a_{ij}: i, j = 1, \dots, n\}$ is the sum of its eigen values, which is equal to the sum of diagonal elements.

$$\begin{aligned} \mathrm{tr}(\boldsymbol{A}) &= \Sigma a_{ii} \\ &= \Sigma \lambda_i \end{aligned}$$

Mathematical review III:
Trace and determinant of matrices

- Proof of determinant. In terms of $\boldsymbol{A}$, there exists λ and $\boldsymbol{x}$ that satisfy

$$(\lambda\boldsymbol{I}-\boldsymbol{A})\boldsymbol{x}=\boldsymbol{0} \qquad \boldsymbol{I}\text{: unit matrix}$$

- The determinant of the first component should be zero.

$$\det(\lambda\boldsymbol{I}-\boldsymbol{A})=0$$

- Using the eigen values of $\boldsymbol{A}$ (λ_1, …, λ_i,…, λ_n),

$$\det(\lambda\boldsymbol{I}-\boldsymbol{A})=(\lambda_1-\lambda)\ldots(\lambda_i-\lambda)\ldots(\lambda_n-\lambda)=0$$

- This should be also true for $\lambda=0$, and then,

$$\det(\boldsymbol{A})=\lambda_1\ldots\lambda_i\ldots\lambda_n$$

Mathematical review III:
Trace and determinant of matrices

- Proof of trace. $\boldsymbol{A}$ is expanded into matrices of eigen vectors and values as follow:

$$\boldsymbol{A}=\boldsymbol{X\Lambda X}^{-1}$$

$\boldsymbol{X}$: Matrix of eigen vectors

$\boldsymbol{\Lambda}$: Diagonal matrix of eigen values

- Using $\mathrm{tr}(\boldsymbol{CD})=\mathrm{tr}(\boldsymbol{DC})$, the trace of this equation is

$$\mathrm{tr}(\boldsymbol{A})=\mathrm{tr}(\boldsymbol{X\Lambda X}^{-1})$$

Mathematical review III:
Trace and determinant of matrices

- Proof of trace. $\boldsymbol{A}$ is expanded into matrices of eigen vectors and values as follow:

$$\boldsymbol{A}=\boldsymbol{X\Lambda X}^{-1}$$

$\boldsymbol{X}$: Matrix of eigen vectors

$\boldsymbol{\Lambda}$: Diagonal matrix of eigen values

- Using $\mathrm{tr}(\boldsymbol{CD})=\mathrm{tr}(\boldsymbol{DC})$, the trace of this equation is

$$\begin{aligned}\mathrm{tr}(\boldsymbol{A})&=\mathrm{tr}(\boldsymbol{X\Lambda X}^{-1})\\&=\mathrm{tr}(\boldsymbol{\Lambda X}^{-1}\boldsymbol{X})\\&=\mathrm{tr}(\boldsymbol{\Lambda})\\&=\Sigma\lambda_i\end{aligned}$$

2.7 Example of PCA

- Scores of five subjects (math, physics, chemistry, language, and history) in Junior high school.

Student	Math I (x_1)	Math II (x_2)	Chemistry (x_3)	Language (x_4)	History (x_5)
1	58	62	40	20	17
2	27	24	20	18	17
3	79	78	67	61	60
4	53	53	69	81	82
…	…	…	…	…	…
n	72	72	58	50	45

2.7 Example of PCA

- Compute the eigen value expansion of $\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$ and acquire eigen vectors and values as follows.

Eigen vectors
$$\boldsymbol{Y} = \begin{bmatrix} .70 & .38 & .24 & .13 & -.05 \\ .83 & .27 & .25 & -.20 & .32 \\ .62 & .53 & .03 & .19 & .19 \\ .47 & -.18 & .38 & .22 & .08 \\ .70 & -.59 & .12 & -.05 & -.05 \end{bmatrix} \sim [\boldsymbol{y}_1\ \boldsymbol{y}_2\ ]$$

Eigen values
$$\boldsymbol{\Lambda} = \begin{bmatrix} 2.9 & 0 & 0 & 0 & 0 \\ & 1.2 & 0 & 0 & 0 \\ & & .68 & 0 & 0 \\ & & & .54 & 0 \\ & & & & .32 \end{bmatrix}$$

→ Two largest components are employed.

2.7 Example of PCA

- Coefficients of two principal components
- The 1st component equally includes all subjects, which indicates average academic performance of students.
- The 2nd component includes positive math and physics scores and negative language and history scores, which indicates science- or arts-students.

	1st PC	2nd PC
Math I	.70	.38
Math II	.83	.27
Chemistry	.62	.53
Language	.47	-.18
History	.70	-.59

2.7 Example of PCA

- Remember that normalized data matrix $\boldsymbol{Z}$ is factorized into

$$\boldsymbol{Z} = \boldsymbol{Y}\boldsymbol{B} \qquad (2.7.1)$$

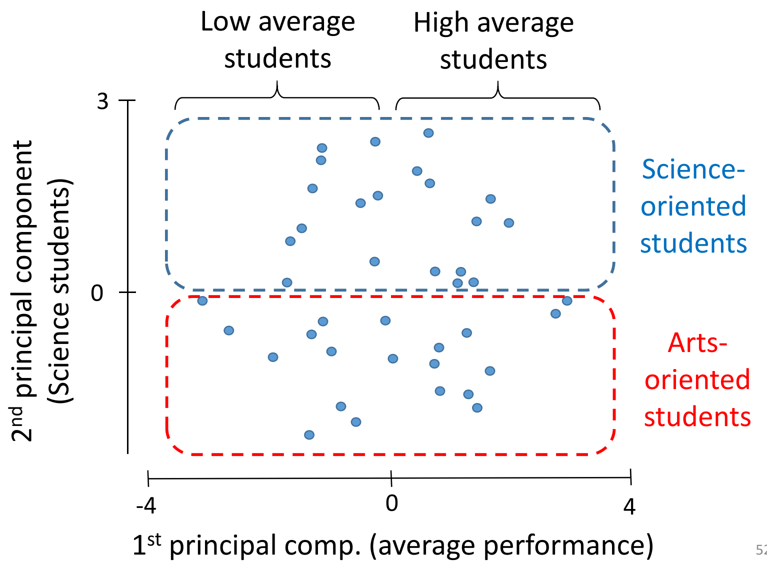
- The score matrix is

$$\boldsymbol{B} = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Z} \qquad (2.7.2)$$

- Using two principal components, the scores of each students (coordinates of 1st and 2nd PCs) are computed by

$$\widehat{\boldsymbol{B}} \sim \begin{bmatrix} \boldsymbol{y}_1^{\mathrm{T}} \\ \boldsymbol{y}_2^{\mathrm{T}} \end{bmatrix} \boldsymbol{Z} \qquad (2.7.3)$$

2.7 Example of PCA



2.8 Rotation of PC coefficients
2.8.1 Simple structure (単純構造)

- The matrix of PC coefficients is given by

$$\widehat{\boldsymbol{Y}} = [\boldsymbol{y}_1\ \boldsymbol{y}_2\ \ldots\ \boldsymbol{y}_q]$$

- If the magnitudes of some elements in $\boldsymbol{y}_i$ are large, and those of the others are nearly zero, then the coefficient matrix is of simple structure and its interpretation is easy.

2.8.1 Simple structure

	1st PC	2nd PC
Math I	.70	.38
Math II	.83	.27
Chemistry	.62	.53
Language	.47	-.18
History	.70	-.59

1st PC: Average ability; 2nd PC: Scientific ability

Varimax rotation ⇒

1st PC	2nd PC
.76	.23
.78	.08
.81	.41
.16	.43
.06	.91

1st PC: Scientific ability; 2nd PC: Arts ability

- After rotation (right), the structure of the coefficient matrix is simple.

2.8.2 Rotation

- Using PCA, the data matrix is approximated by

$$\boldsymbol{Z} \sim \widehat{\boldsymbol{Y}}\widehat{\boldsymbol{B}} \qquad (2.8.1)$$

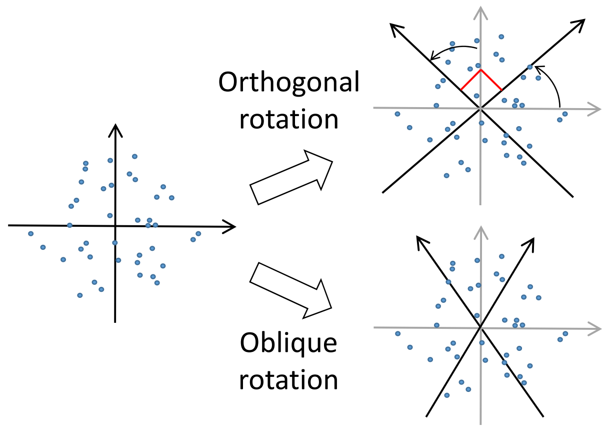
- Rotation matrices ($\boldsymbol{R}$) are orthonormal. Then, (2.8.1) can be transformed into

$$\begin{aligned} \boldsymbol{Z} &\sim \widehat{\boldsymbol{Y}}\widehat{\boldsymbol{B}} \qquad (2.8.2) \\ &= \widehat{\boldsymbol{Y}}\boldsymbol{R}\boldsymbol{R}^{-1}\widehat{\boldsymbol{B}} \\ &= (\widehat{\boldsymbol{Y}}\boldsymbol{R})(\boldsymbol{R}^{-1}\widehat{\boldsymbol{B}}) \\ &= \widehat{\boldsymbol{Y}}_r\widehat{\boldsymbol{B}}_r \end{aligned}$$

- PC coefficients can be rotated freely such that their meanings can be easily interpreted.

2.8.2 Rotation

- Two types of rotation algorithms are introduces:
  - Orthogonal rotation: Varimax rotation
  - Oblique rotation: Promax rotation



2.8.3 Orthogonal rotation (直交回転)

- Orthogonal rotation uses an orthonormal rotation matrix. Hence, even after rotation, the axes are still orthogonal.
- Varimax rotation. Determine the rotation matrix to maximize the variance of the squared elements in $\widehat{\boldsymbol{Y}}$.

	1st PC	2nd PC
Math I	.70	.38
Math II	.83	.27
Chemistry	.62	.53
Language	.47	-.18
History	.70	-.59

Varimax rotation ⇒

1st PC	2nd PC
.76	.23
.78	.08
.81	.41
.16	.43
.06	.91

2.8.3 Orthogonal rotation (直交回転)

	1st PC	2nd PC
Math I	.70	.38
Math II	.83	.27
Chemistry	.62	.53
Language	.47	-.18
History	.70	-.59
	Average ability	Scientific ability

Varimax rotation ⇒

1st PC	2nd PC
.76	.23
.78	.08
.81	.41
.16	.43
.06	.91
Scientific ability	Arts ability

- Another interpretation of principal components is enabled by rotation.
- Varimax rotation is provided by *rotatefactors* function.

2.8.4 Oblique rotation (斜交回転)
2.8.4.1 Procrustes transformation

$$\underset{(p\times q)}{\boldsymbol{T}_{\text{gt}}} = \underset{(p\times q)}{\widehat{\boldsymbol{Y}}}\ \underset{(q\times q)}{\boldsymbol{R}} \tag{2.8.1}$$

- Original matrix $\widehat{\boldsymbol{Y}}$ is rotated such that it resembles target matrix $\boldsymbol{T}_{\text{gt}}$, which is simpler than $\widehat{\boldsymbol{Y}}$.
- The rotation matrix $\boldsymbol{R}$ is computed by

$$\boldsymbol{R} = \widehat{\boldsymbol{Y}}^{+}\boldsymbol{T}_{\text{gt}} \tag{2.8.2}$$

2.8.4.2 Target matrix

- It is preferred that the target matrix is of simple structure. Here, a typical method to prepare for a simple-structured target matrix.

1. Normalize each row of $\widehat{\boldsymbol{Y}}$ s.t. the L2-norm of each row is 1.
2. Multiply each column by a fixed unique number s.t. the maximum element in the column is 1/-1.
3. Raise each element of the matrix to the k-th power ($k =$ 3 or 5, typically).
4. Normalize each column (L2-norm is 1) not to extend or shrink the axes.

2.8.4.2 Target matrix

- Example when the original matrix is given by $\widehat{\boldsymbol{Y}}$.

$$\widehat{\boldsymbol{Y}} = \begin{bmatrix} -.33 & .38 \\ .09 & -.90 \\ .94 & .22 \end{bmatrix}$$

1. Normalize each row of $\widehat{\boldsymbol{Y}}$ s.t. the L2-norm of each row is 1.

$$\boldsymbol{T}_{gt1} = \begin{bmatrix} -.66 & .75 \\ .10 & -.99 \\ .97 & .23 \end{bmatrix} \begin{matrix} .66^2 + .75^2 = 1 \\ .10^2 + .99^2 = 1 \\ .97^2 + .23^2 = 1 \end{matrix}$$

2.8.4.2 Target matrix

2. Multiply each column by a fixed unique number s.t. the maximum element in the column is 1/-1.

$$\boldsymbol{T}_{\mathrm{gt1}} = \begin{bmatrix} -.66 & .75 \\ .10 & -.99 \\ .97 & .23 \end{bmatrix}$$

$$\times 1.04 \downarrow \qquad \downarrow \times 1.01$$

$$\boldsymbol{T}_{\mathrm{gt2}} = \begin{bmatrix} -.68 & .75 \\ .10 & -1.00 \\ 1.00 & .23 \end{bmatrix}$$

2.8.4.2 Target matrix

3. Raise each element of the matrix to the k-th power ($k =$ 3 or 5, typically).

$$\boldsymbol{T}_{\mathrm{gt3}} = \begin{bmatrix} -.68^3 & .75^3 \\ .10^3 & -1.00^3 \\ 1.00^3 & .23^3 \end{bmatrix} = \begin{bmatrix} -.31 & .42 \\ .00 & -1.00 \\ 1.00 & .01 \end{bmatrix}$$

2.8.4.2 Target matrix

4. Normalize each column (L2-norm is 1) not to extend or shrink the axes.

$$\boldsymbol{T}_{\mathrm{gt3}} = \begin{bmatrix} -.31 & .42 \\ .00 & -1.00 \\ 1.00 & .01 \end{bmatrix}$$

$$\Downarrow$$

$$\boldsymbol{T}_{\mathrm{gt4}} = \begin{bmatrix} -.29 & .39 \\ .00 & -.92 \\ .96 & .00 \end{bmatrix}$$

2.8.4.2 Target matrix

- Compute the rotation matrix by using T_{gt4} as follows.

$$\boldsymbol{R} = \widehat{\boldsymbol{Y}}^{+}\boldsymbol{T}_{\mathrm{gt}} \tag{2.8.3}$$

$$= \begin{bmatrix} -.33 & .38 \\ .09 & -.90 \\ .94 & .22 \end{bmatrix}^{+} \begin{bmatrix} -.29 & .39 \\ .00 & -.92 \\ .96 & .00 \end{bmatrix}$$

$$= \begin{bmatrix} .99 & -.21 \\ .10 & .97 \end{bmatrix}$$

- $\boldsymbol{R}$ is not orthonormal.

2.8.4.2 Target matrix

- Promax rotation produces non-orthogonal coefficient vectors.

$$\overline{\boldsymbol{Y}} = \widehat{\boldsymbol{Y}}\boldsymbol{R} \tag{2.8.4}$$

$$\begin{bmatrix} -.29 & .04 \\ .00 & -.89 \\ .95 & .02 \end{bmatrix} = \begin{bmatrix} -.33 & .38 \\ .09 & -.90 \\ .94 & .22 \end{bmatrix} \begin{bmatrix} .99 & -.21 \\ .10 & .97 \end{bmatrix}$$

$\widehat{\boldsymbol{Y}}$ became more sparse (simple-structured) by promax rotation, resulting in easy interpretation of PCs.

Second report

1. Title (your name, student ID)
2. Background of data (what kind of data? why do you want to analyze personally?)
3. Scree diagram and contribution ratios of components (how many components should you use?)
4. Table or matrix of PC coefficients before/after rotation
5. Interpretation of the results (is the result reasonable? what did you find behind the data?)

You may use the data used for the previous report (multiple regression analysis).