Lecture 8 Principal Component Analysis

Shiwei Lan[1]

[1]School of Mathematical and Statistical Sciences
Arizona State University

STP533 Multivariate Analysis
Spring 2025

Table of Contents

Principal Component Analysis

- A principal component analysis (PCA) is concerned with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables.
- Its general objectives are (1) data reduction and (2) interpretation.
- Although p components are required to reproduce the total system variability, often much of the variability can be accounted by a small number of k of the principal components.
- PCA can reveal relationships that were not previously suspected or discovered, and is more of a means than an end.

Population Principal Components

- Algebraically, PCs are linear combinations of p random variables; geometrically, these linear combinations represents the selection of new coordinate system by rotating the original one with these variables.
- Let the random vector $\mathbf{X} = [X_1, \cdots, X_p]$ have a covariance matrix $\mathbf{\Sigma}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.
- Consider the linear combinations $Y_i = \mathbf{a}_i^T \mathbf{X} = \sum_{j=1}^p a_{ij} X_j$. We have

$$\mathrm{Var}(Y_i) = \mathbf{a}_i^T \mathbf{\Sigma} \mathbf{a}_i, \quad \mathrm{Cov}(Y_i, Y_k) = \mathbf{a}_i^T \mathbf{\Sigma} \mathbf{a}_k, \quad , i, k = 1, \cdots, p$$

- The *principal components* are those uncorrelated linear combinations Y_i's with variances as large as possible.

Principal Components

- The first principal component is Y_1 that maximizes $\mathrm{Var}(Y_1) = \mathbf{a}_1^T \boldsymbol{\Sigma} \mathbf{a}_1$.
- To make rigorous of the problem, we normalize the linear coefficients $\mathbf{a}_1$ such that $\|\mathbf{a}_1\|_2^2 = \mathbf{a}_1^T \mathbf{a}_1 = 1$.
- The i-th PC is $Y_i = \mathbf{a}_i^T \mathbf{X}$ such that
  - $\mathrm{Var}(Y_i) = \mathbf{a}_i^T \boldsymbol{\Sigma} \mathbf{a}_i$ is maximized subject to $\mathbf{a}_i^T \mathbf{a}_i = 1$,
  - $\mathrm{Cov}(Y_i, Y_k) = 0$ for $k < i$.
- Suppose $\boldsymbol{\Sigma} = \Gamma \Lambda \Gamma^{-1} = \Gamma \Lambda \Gamma^T$ with $\Lambda = \mathrm{diag}(\{\lambda_i\})$ and $\Gamma = [\mathbf{v}_1, \cdots, \mathbf{v}_p]$. Then we have

$$Y_i = \mathbf{v}_i^T \mathbf{X} = \sum_{j=1}^p v_{ij} X_j, \quad \mathrm{Var}(Y_i) = \lambda_i, \quad \mathrm{Cov}(Y_i, Y_k) = 0 \; for \; i \neq k.$$

Principal Components

- Let $Y_i = \mathbf{v}_i^T \mathbf{X}$'s be the PCs of $\mathbf{X}$ with covariance matrix $\boldsymbol{\Sigma}$.
- Then the total populations variance is

$$\sum_{i=1}^{p} \mathrm{Var}(Y_i) = \sum_{i=1}^{p} \lambda_i = \mathrm{tr}(\boldsymbol{\Sigma}) = \sum_{i=1}^{p} \mathrm{Var}(X_i)$$

- The correlation coefficients between Y_i and X_k becomes

$$\rho_{Y_i,X_k} = \frac{v_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$$

- The correlations ρ_{Y_i,X_k} measures the *univariate* contribution of individual X_k to the component Y_i.
- Some statisticians suggest the coefficients v_{ik} to interpret the contribution of x_k to the component Y_i.

Example 8.1 (Calculating the population principal components) Suppose the random variables X_1, X_2 and X_3 have the covariance matrix

$$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

It may be verified that the eigenvalue–eigenvector pairs are

$$\begin{aligned} \lambda_1 &= 5.83, & \mathbf{e}_1' &= [.383, -.924, 0] \\ \lambda_2 &= 2.00, & \mathbf{e}_2' &= [0, 0, 1] \\ \lambda_3 &= 0.17, & \mathbf{e}_3' &= [.924, .383, 0] \end{aligned}$$

Principal Components for Normal Random Variables

- Now suppose $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- It is known that $(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2$ has axes $\pm c\sqrt{\lambda_i}\mathbf{v}_i$, $i = 1, \cdots, p$.
- Assume $\boldsymbol{\Sigma} = \Gamma\Lambda\Gamma^T$. Then $\boldsymbol{\Sigma}^{-1} = \Gamma\Lambda^{-1}\Gamma^T$.
- Assume $\boldsymbol{\mu} = \mathbf{0}$. Let $y_i = \mathbf{v}_i^T\mathbf{x}$ be the PCs. Then

$$c^2 = \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} = \sum_{i=1}^p \lambda_i^{-1}(\mathbf{v}_i^T\mathbf{x})^2 = \sum_{i=1}^p \lambda_i^{-1} y_i^2$$

- This equation defines an ellipsoid in a new coordinate systems $\{y_i\}$ with axes lying in the directions of $\{\mathbf{v}_i\}$.

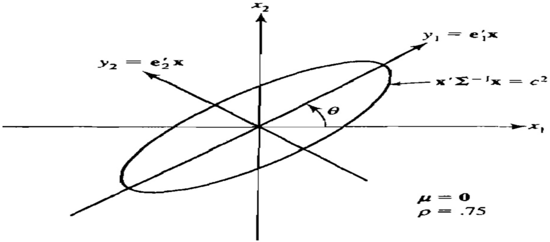

Figure 8.1 The constant density ellipse $\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} = c^2$ and the principal components y_1, y_2 for a bivariate normal random vector $\mathbf{X}$ having mean $\mathbf{0}$.

Principal Components for Normal Random Variables

- PCs may also be obtained for the standardized variables $Z_i = \frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}$.
- Denote $\mathbf{D} = \mathrm{diag}(\{\sqrt{\sigma_{ii}}\})$. Then $\mathbf{Z} = \mathbf{D}^{-1}(\mathbf{X} - \boldsymbol{\mu})$. Therefore the covariance of $\mathbf{Z}$ is actually the correlation matrix

$$\mathrm{Cov}(\mathbf{Z}) = \mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1} = \boldsymbol{P}$$

- The PCs of standardized normal random variables $\mathbf{Z}$ is given by

$$Y_i^* = \mathbf{v}_i^T\mathbf{Z} = \mathbf{v}_i^T\mathbf{D}^{-1}(\mathbf{X} - \boldsymbol{\mu})$$

- Moreover, we have

$$\sum_{i=1}^{p} \mathrm{Var}(Y_i^*) = \sum_{i=1}^{p} \mathrm{Var}(Z_i) = p, \quad \rho_{Y_i^*, Z_k} = v_{ik}\sqrt{\lambda_i}, \; i, k = 1, \cdots, p.$$

- We say the proportion of total variance explained by the k-th PC of $\mathbf{Z}$ is $\frac{\lambda_k^*}{p}$ with λ_k^* being the k-th eigenvalue of $\boldsymbol{P}$.

- Suppose the covariance matrix is diagonal $\boldsymbol{\Sigma} = \mathrm{diag}(\{\sigma_{ii}\})$. What are the PCs?

- How about the following covariance matrix coming from biology:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \cdots & \rho\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \cdots & \sigma^2 \end{bmatrix}$$

Hint: Check the correlation matrix.

Summarizing Sample Variation

- Now we consider the problem of summarizing the variation of n measurements on p variables with a few linear combinations.
- Let $\mathbf{X}_{n\times p}$ be a sample of size n from population with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
- We have the sample mean vector $\bar{\mathbf{X}}$, the sample covariance matrix $\mathbf{S}$, and the sample correlation matrix $\mathbf{R}$.
- The *sample principal components* are uncorrelated combinations that maximize the sample covariance matrix.
- The first sample PC is $\hat{\mathbf{y}}_1 = \mathbf{X}\mathbf{a}_1$ that maximizes the sample variance $\widehat{\mathrm{Var}}(\mathbf{y}_1) = \mathbf{a}_1^T\mathbf{S}\mathbf{a}_1$ subject to $\mathbf{a}_1^T\mathbf{a}_1 = 1$.
- The i-th sample PC is $\hat{\mathbf{y}}_i = \mathbf{X}\mathbf{a}_i$ such that $\widehat{\mathrm{Var}}(\hat{\mathbf{y}}_i) = \mathbf{a}_i^T\mathbf{S}\mathbf{a}_i$ is maximized subject to $\mathbf{a}_i^T\mathbf{a}_i = 1$, and $\widehat{\mathrm{Cov}}(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_k) = 0$ for $k < i$.

Sample Principal Components

- Suppose $\mathbf{S} = \Gamma\Lambda\Gamma^T$ with $\Lambda = \mathrm{diag}(\{\hat\lambda_i\})$ and $\Gamma = [\hat{\mathbf{v}}_1, \cdots, \hat{\mathbf{v}}_p]$. Then we have sample PCs

$$\hat{\mathbf{y}}_i = \mathbf{X}\hat{\mathbf{v}}_i = \sum_{j=1}^p \hat v_{ij}\mathbf{x}_j, \quad \widehat{\mathrm{Var}}(\hat{\mathbf{y}}_i) = \hat\lambda_i, \quad \widehat{\mathrm{Cov}}(\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_k) = 0 \textit{ for } i \neq k.$$

- In addition, we have the total sample variance and sample correlation

$$\sum_{i=1}^p s_{ii} = \sum_{i=1}^p \hat\lambda_i, \quad r_{\hat{\mathbf{y}}_i, \mathbf{x}_k} = \frac{\hat v_{ik}\sqrt{\hat\lambda_i}}{\sqrt{s_{kk}}}, \; i, k = 1, \cdots, p.$$

- The observations $\mathbf{x}_j$ are often "centered" to have centered sample PCs

$$\hat{\mathbf{y}}_i = (\mathbf{X} - \bar{\mathbf{X}})\hat{\mathbf{v}}_i, \quad i = 1, \cdots, p.$$

Calculating Two Sample PCs

Example 8.3 (Summarizing sample variability with two sample principal components) A census provided information, by tract, on five socioeconomic variables for the Madison, Wisconsin, area. The data from 61 tracts are listed in Table 8.5 in the exercises at the end of this chapter. These data produced the following summary statistics:

$$\bar{\mathbf{x}}' = \underset{\substack{\text{total}\\\text{population}\\\text{(thousands)}}}{[4.47,} \quad \underset{\substack{\text{professional}\\\text{degree}\\\text{(percent)}}}{3.96,} \quad \underset{\substack{\text{employed}\\\text{age over 16}\\\text{(percent)}}}{71.42,} \quad \underset{\substack{\text{government}\\\text{employment}\\\text{(percent)}}}{26.91,} \quad \underset{\substack{\text{median}\\\text{home value}\\\text{(\$100,000)}}}{1.64]}$$

and

$$\mathbf{S} = \begin{bmatrix} 3.397 & -1.102 & 4.306 & -2.078 & 0.027 \\ -1.102 & 9.673 & -1.513 & 10.953 & 1.203 \\ 4.306 & -1.513 & 55.626 & -28.937 & -0.044 \\ -2.078 & 10.953 & -28.937 & 89.067 & 0.957 \\ 0.027 & 1.203 & -0.044 & 0.957 & 0.319 \end{bmatrix}$$

Can the sample variation be summarized by one or two principal components?

The Number of PCs

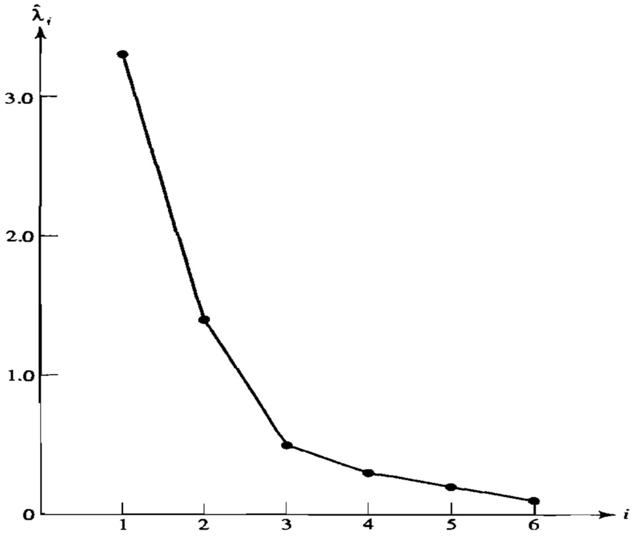

Figure 8.2 A scree plot.

Interpretation of Sample PCs

- Suppose $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- Then sample PCs $\hat{\mathbf{y}}_i = (\mathbf{X} - \bar{\mathbf{X}})\hat{\mathbf{v}}_i \sim N_p(\mathbf{0}, \Lambda)$ with $\Lambda = \mathrm{diag}(\{\lambda_i\})$.
- Then $(\mathbf{X} - \bar{\mathbf{X}})^T \mathbf{S}^{-1}(\mathbf{X} - \bar{\mathbf{X}}) = c^2$ approximates the constant density contour $(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c^2$.
- This hyperellipsoid is centered at $\bar{\mathbf{X}}$ with axes given by $\hat{\mathbf{v}}_i$ and length proportional to $\sqrt{\hat{\lambda}_i}$.

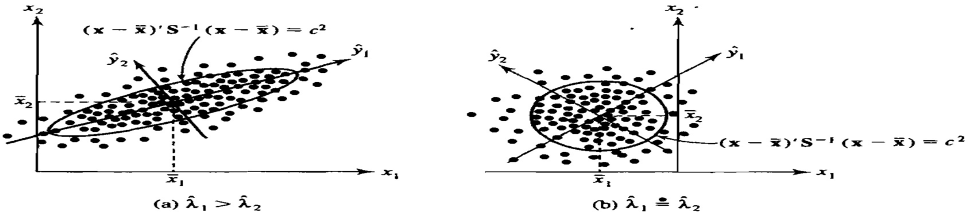

Figure 8.4 Sample principal components and ellipses of constant distance.

Standardizing the Sample PCs

- Sample PCs may also be obtained for the standardized variables $\mathbf{Z} = (\mathbf{X} - \bar{\mathbf{X}})\mathbf{D}^{-1}$ where $\mathbf{D} = \mathrm{diag}(\{\sqrt{s_{ii}}\})$.
- The sample covariance of $\mathbf{Z}$ is

$$\widehat{\mathrm{Cov}}(\mathbf{Z}) = \mathbf{S}_{\mathbf{Z}} = \frac{1}{n-1}(\mathbf{Z} - \bar{\mathbf{Z}})^T(\mathbf{Z} - \bar{\mathbf{Z}}) = \frac{1}{n-1}\mathbf{Z}^T\mathbf{Z} =: \mathbf{R}$$

- Let $\{\hat{\lambda}_i, \hat{\mathbf{v}}_i\}$ be eigen-pairs of $\mathbf{R}$. The sample PCs of standardized observations are given by

$$\mathbf{y}_i^* = \mathbf{Z}\hat{\mathbf{v}}_i = (\mathbf{X} - \bar{\mathbf{X}})\mathbf{D}^{-1}\hat{\mathbf{v}}_i$$

- Moreover, we have

$$\widehat{\mathrm{Var}}(\hat{\mathbf{y}}_i^*) = \hat{\lambda}_i, \quad \widehat{\mathrm{Cov}}(\hat{\mathbf{y}}_i^*, \hat{\mathbf{y}}_k^*) = 0 \; for \; i \neq k.$$

$$\sum_{i=1}^p \widehat{\mathrm{Var}}(\hat{\mathbf{y}}_i^*) = \mathrm{tr}(\mathbf{R}) = \sum_{i=1}^p \hat{\lambda}_i = p, \quad r_{\hat{\mathbf{y}}_i^*, \mathbf{z}_k} = \hat{v}_{ik}\sqrt{\hat{\lambda}_i}, \; i, k = 1, \cdots, p.$$

- The proportion of total sample variance explained by the i-th sample PC: $\frac{\lambda_i}{p}$.

Calculating Sample Principal Components

Example 8.6 (Components from a correlation matrix with a special structure) Geneticists are often concerned with the inheritance of characteristics that can be measured several times during an animal's lifetime. Body weight (in grams) for $n = 150$ female mice were obtained immediately after the birth of their first four litters.[4] The sample mean vector and sample correlation matrix were, respectively,

$$\bar{\mathbf{x}}' = [39.88, 45.08, 48.11, 49.95]$$

and

$$\mathbf{R} = \begin{bmatrix} 1.000 & .7501 & .6329 & .6363 \\ .7501 & 1.000 & .6925 & .7386 \\ .6329 & .6925 & 1.000 & .6625 \\ .6363 & .7386 & .6625 & 1.000 \end{bmatrix}$$

The eigenvalues of this matrix are

$$\hat{\lambda}_1 = 3.085, \quad \hat{\lambda}_2 = .382, \quad \hat{\lambda}_3 = .342, \quad \text{and} \quad \hat{\lambda}_4 = .217$$

Graphing the PCs

- Plots of PCs can reveal suspect observations, as well as provide checks on the normality assumption.
- The last PCs can help pinpoint suspect observations. Each observation can be expressed

$$\mathbf{x}_i = \sum_{j=1}^{p} (\mathbf{x}_i^T \hat{\mathbf{v}}_j)\hat{\mathbf{v}}_j = \sum_{j=1}^{p} \hat{y}_{ij}\hat{\mathbf{v}}_j$$

- The suspect observations often have large values in one of the coordinates $\hat{y}_{ij}$.

1. To help check the normal assumption, construct scatter diagrams for pairs of the first few principal components. Also, make *Q–Q* plots from the sample values generated by *each* principal component.
2. Construct scatter diagrams and *Q–Q* plots for the last few principal components. These help identify suspect observations.

Example 8.7 (Plotting the principal components for the turtle data) We illustrate the plotting of principal components for the data on male turtles discussed in Example 8.4. The three sample principal components are

$$\begin{aligned}
\hat{y}_1 &= .683(x_1 - 4.725) + .510(x_2 - 4.478) + .523(x_3 - 3.703) \\
\hat{y}_2 &= -.159(x_1 - 4.725) - .594(x_2 - 4.478) + .788(x_3 - 3.703) \\
\hat{y}_3 &= -.713(x_1 - 4.725) + .622(x_2 - 4.478) + .324(x_3 - 3.703)
\end{aligned}$$

where $x_1 = \ln(\text{length})$, $x_2 = \ln(\text{width})$, and $x_3 = \ln(\text{height})$, respectively.

Figure 8.5 shows the Q–Q plot for $\hat{y}_2$ and Figure 8.6 shows the scatter plot of $(\hat{y}_1, \hat{y}_2)$. The observation for the first turtle is circled and lies in the lower right corner of the scatter plot and in the upper right corner of the Q–Q plot; it may be suspect. This point should have been checked for recording errors, or the turtle should have been examined for structural anomalies. Apart from the first turtle, the scatter plot appears to be reasonably elliptical. The plots for the other sets of principal components do not indicate any substantial departures from normality. ■

Large Sample Properties

- We have seen that the eigenvectors of (empirical) covariance determine the directions of the maximum variability; while the eigenvalues specify the variances.
- So far, all the investigate has been based on the normality assumption. If this fails, we could still have large sample properties.
- Suppose $\mathbf{X}_i \overset{iid}{\sim} (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\{\lambda_i, \mathbf{v}_i\}$ be the eigen-pairs of $\boldsymbol{\Sigma}$, and $\{\hat{\lambda}_i, \hat{\mathbf{v}}_i\}$ be the eigen-pairs of sample covariance matrix $\mathbf{S}$. Then the following results are due to Anderson and Girshick:
  1. $\sqrt{n}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \xrightarrow{L} N_p(\mathbf{0}, 2\Lambda^2)$, with $\Lambda = \mathrm{diag}(\{\lambda_i\})$.
  2. Let $\mathbf{E}_i = \lambda_i \sum_{k\neq i} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} \mathbf{v}_k \mathbf{v}_k^T$. Then $\sqrt{n}(\hat{\mathbf{v}}_i - \mathbf{v}_i) \xrightarrow{L} N_p(\mathbf{0}, \mathbf{E}_i)$.
  3. Each $\hat{\lambda}_i$ is distributed independently of the elements of $\hat{\mathbf{v}}_i$.

Large Sample Properties

- Based on result 1, we construct $100(1-\alpha)\%$ confidence interval for each λ_i:

$$\frac{\hat{\lambda}_i}{1 + z_{1-\alpha/2}\sqrt{2/n}} \le \lambda_i \le \frac{\hat{\lambda}_i}{1 - z_{1-\alpha/2}\sqrt{2/n}}$$

- $100(1-\alpha)\%$ Bonferroni-type SCI for λ_i's can be obtained by replacing $z_{1-\alpha/2}$ with $z_{1-\alpha/(2m)}$.
- Result 2 can also be used to derive the approximate CI for v_{ij} substituting λ_i with $\hat{\lambda}_i$ and $\mathbf{v}_i$ with $\hat{\mathbf{v}}_i$ in $\mathbf{E}_i$.

Confidence Interval for λ_1

Example 8.8 (Constructing a confidence interval for λ_1) We shall obtain a 95% confidence interval for λ_1, the variance of the first population principal component, using the stock price data listed in Table 8.4 in the Exercises.

Assume that the stock rates of return represent independent drawings from an $N_5(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ population, where $\boldsymbol{\Sigma}$ is positive definite with distinct eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_5 > 0$. Since $n = 103$ is large, we can use (8-33) with $i = 1$ to construct a 95% confidence interval for λ_1. From Exercise 8.10, $\hat{\lambda}_1 = .0014$ and in addition, $z(.025) = 1.96$. Therefore, with 95% confidence,

$$\frac{.0014}{\left(1 + 1.96\sqrt{\frac{2}{103}}\right)} \le \lambda_1 \le \frac{.0014}{\left(1 - 1.96\sqrt{\frac{2}{103}}\right)} \quad \text{or} \quad .0011 \le \lambda_1 \le .0019 \qquad \blacksquare$$

Testing Equal Correlation Structure

- The special correlation structure $\mathrm{Corr}(X_i, X_k) = \rho$ for $i \neq k$ is important. We can test

$$H_0 : \boldsymbol{P} = \boldsymbol{P}_0, \quad \boldsymbol{P}_0 = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

- Lawley's procedure requires the quantities

$$\bar{r}_k = \frac{1}{p-1} \sum_{i \neq k} r_{ik}, \quad \bar{r} = \frac{2}{p(p-1)} \sum_{i<k} r_{ik}, \quad \hat{\gamma} = \frac{(p-1)^2[1-(1-\bar{r})^2]}{p-(p-2)(1-\bar{r})^2}$$

- The large sample approximate α-level test is to reject H_0 if

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

Testing for Equicorrelation Structure

Example 8.9 (Testing for equicorrelation structure) From Example 8.6, the sample correlation matrix constructed from the $n = 150$ post-birth weights of female mice is

$$\mathbf{R} = \begin{bmatrix} 1.0 & .7501 & .6329 & .6363 \\ .7501 & 1.0 & .6925 & .7386 \\ .6329 & .6925 & 1.0 & .6625 \\ .6363 & .7386 & .6625 & 1.0 \end{bmatrix}$$

We shall use this correlation matrix to illustrate the large sample test in (8-35). Here $p = 4$, and we set

$$H_0: \boldsymbol{\rho} = \boldsymbol{\rho}_0 = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

$$H_1: \boldsymbol{\rho} \neq \boldsymbol{\rho}_0$$

Table of Contents

Probabilistic PCA*

- Probabilistic PCA (Tipping and Bishop, 1999) generalizes PCA into a probabilistic model whose maximum likelihood estimates corresponds to the traditional version.
- The probabilistic PCA assumes that high-dimensional data $\mathbf{Y}_n \in \mathbb{R}^D$ lie on a lower-dimensional space spanned by latent variables $\mathbf{X}_n \in \mathbb{R}^Q$ with $Q \ll D$.
- Assuming a linear (factor) function $\mathbf{W} \in \mathbb{R}^{D\times Q}$, we have the following model

$$\mathbf{Y}_n | \mathbf{X}_n \sim N_D(\mathbf{W}\mathbf{X}_n, \sigma^2 \mathbf{I}_D)$$
$$\mathbf{X}_n \sim N_Q(\mathbf{0}, \mathbf{I}_Q)$$

- Marginalizing the latent variables $\mathbf{X}_i$ yields the following Gaussian

$$\mathbf{Y}_n \sim N_D(\mathbf{0}, \mathbf{C}), \quad \mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2 \mathbf{I}_D$$

with log-likelihood as

$$L = -\frac{N}{2}[\log|\mathbf{C}| + \mathrm{tr}(\mathbf{C}^{-1}\mathbf{S_Y})], \quad \mathbf{S_Y} = \frac{1}{N}\sum_{n=1}^N \mathbf{Y}_n \mathbf{Y}_n^T.$$

Probabilistic PCA*

- The maximum likelihood estimator (MLE) of $\mathbf{W}$ can be derived as

$$\hat{\mathbf{W}}_{ML} = \mathbf{U}(\Lambda_Q - \sigma^2 \mathbf{I})^{1/2}\mathbf{R}$$

  where $\mathbf{U}_{D\times Q}$ is formed the Q principal eigenvectors of $\mathbf{S_Y}$, $\Lambda_Q = \mathrm{diag}(\{\lambda_i\}_{i=1}^Q)$ is formed by the Q principal eigenvalues, and $\mathbf{R}_{Q\times Q}$ is an arbitrary orthogonal matrix.
- One can also show that $\hat{\sigma}^2_{ML} = \frac{1}{D-Q}\sum_{j=Q+1}^D \lambda_j$.
- Lawrence (2003) consider the dual problem of probabilistic PCA by marginalizing the weight parameter $\mathbf{W}$ and generalized probabilistic PCA to Gaussian process latent variable model (GP-LVM) by replacing the linear kernel $\mathbf{C_X} = \mathbf{X}\mathbf{X}^T + \sigma^2\mathbf{I}$ with more general kernel $\mathbf{C}$.
- Obite et al (2025) further generalized GP-LVM with more flexible Q-Exponential process `https://openreview.net/pdf?id=VOoJEQlLW5`.