Chapter 7
Principal Component Analysis

In any enquiry into a problem, if we want to proceed scientifically, we often have a large number of variables. It is desired to know how many variables should be included or discarded. When planning an experiment, the experimenter usually tries to incorporate many variables to safeguard against potential gaps in knowledge before the experiment begins. Increasing the variables has its own consequences. For example, in any medical study, there can be many symptoms of the disease. Suppose each symptom is represented by a variable. Suppose there are $p = 10$ variables presented in a vector $\underset{\sim}{X}$ of order $p \times 1$. Then the

- mean vector has 10 elements,
- covariance matrix has 10^2 elements,
- number of unique variance and covariance terms are $\frac{p(p+1)}{2} = 55$,
- number of unique covariance terms are $\frac{p(p+1)}{2} - p = 45$ and
- number of unique variances is $p = 10$.

Suppose five more variables are increased, so now $p = 15$. Then the

- mean vector has 15 elements,
- covariance matrix has 15^2 elements,
- number of unique variance and covariance terms are $\frac{15\times 16}{2} = 120$,
- number of unique covariance terms are $\frac{p(p+1)}{2} - p = 120 - 15 = 105$ and
- number of unique variance terms is $p = 15$.

The situation becomes more complicated when more variables are added. Handling a large number of variables makes it more difficult to draw clear inferences and handle so many means, variances and covariances.

A possible way out to handle such a situation is to consider a normalized linear combination of the random variables such that the total variance (structure) is not lost, i.e., the total variance remains the same. These normalized linear combinations are known as principal components. Here, 'normalised' means that the sum of the squares of the coefficients is 1.

The principal components are the normalized linear combinations of random variables that have some special properties in terms of variances.

In effect, transforming the original vector variable into the vector of principal components amounts to a rotation of the coordinate axes into a new coordinate system with inherent statistical properties. In practice, the total variation is accounted for by only a few principal components.

Such a methodology helps reduce the number of variables by discarding linear combinations with lower variances and retaining those with higher variances. Such linear combinations are obtained by principal component analysis. So, principal component analysis is a method for reducing the dimensionality of the data without affecting the total variation. For example, if there are ten variables, then it may be possible that only, say, four variables are sufficient and contribute to a large part of the variance.

Principal components also address multicollinearity in multiple linear regression through principal component regression.

In practice, interpreting principal components can be difficult, but the operations on them are simple. For example, the interpretation of $\sum_{i=1}^{p} a_i x_i$ may be difficult for the variables $x_1, x_2, \ldots, x_p$ and scalars $a_1, a_2, \ldots, a_p$. Suppose $\underset{\sim}{X}$ is a $p \times 1$ random vector with covariance matrix as Σ, which is a positive definite symmetric matrix. Without loss of generality, assume that the mean of $\underset{\sim}{X}$ is $\underset{\sim}{0}$. We can assume that the variables are measured from their respective means. Since we are more concerned about the structure of Σ, so the actual distribution of $\underset{\sim}{X}$ is irrelevant unless otherwise stated. If $\underset{\sim}{X}$ is distributed as multivariate normal, then more meaning can be given to principal components, and calculations become simpler. The distribution of $\underset{\sim}{X}$ is needed for testing of hypotheses related to principal components.

Let $\lambda_1, \lambda_2, \ldots, \lambda_p$ be p distinct characteristic roots of $p \times p$ matrix Σ, obtained by solving

$$|\Sigma - \lambda I| = 0$$

where $\lambda_i \geq 0$ for all $i = 1, 2, \ldots, p$. Corresponding to each λ_i, there is a characteristic vector as

$$(\Sigma - \lambda_i I)\underset{\sim}{x}_i = 0.$$

Let $p = 2$, then λ_1 and λ_2 are distinct characteristic roots and associated characteristic vectors are $\underset{\sim}{x}_1$ and $\underset{\sim}{x}_2$, then $\underset{\sim}{x}_1$ and $\underset{\sim}{x}_2$ are orthogonal. This is seen as follows:

$$\begin{aligned}
& (A - \lambda_1 I)\underset{\sim}{x}_1 = 0 \Rightarrow A\underset{\sim}{x}_1 = \lambda_1 \underset{\sim}{x}_1 \\
& (A - \lambda_2 I)\underset{\sim}{x}_2 = 0 \Rightarrow A\underset{\sim}{x}_2 = \lambda_2 \underset{\sim}{x}_2 \\
\text{or} \quad & (\underset{\sim}{x}_2' A \underset{\sim}{x}_1)' = \lambda_2 (\underset{\sim}{x}_2' \underset{\sim}{x}_1)' \\
\text{or} \quad & \underset{\sim}{x}_2' A \underset{\sim}{x}_1 = \lambda_2 \underset{\sim}{x}_2' \underset{\sim}{x}_1
\end{aligned}$$

which using $A\underset{\sim}{x}_1 = \lambda_1 \underset{\sim}{x}_1$ gives

$$\begin{aligned}
& \lambda_1 \underset{\sim}{x}_2' \underset{\sim}{x}_1 = \lambda_2 \underset{\sim}{x}_2' \underset{\sim}{x}_1 \\
\text{or} \quad & (\lambda_1 - \lambda_2)\underset{\sim}{x}_2' \underset{\sim}{x}_1 = 0 \\
\text{or} \quad & \underset{\sim}{x}_2' \underset{\sim}{x}_1 = 0 \text{ as } \lambda_1 \neq \lambda_2.
\end{aligned}$$

This example will help in understanding the derivation of principal components.

Formulation of the Problem of Principal Components

Suppose a $p \times 1$ random vector $\underset{\sim}{X}$ is having a mean vector $\underset{\sim}{0}$ and covariance matrix Σ. Let $\underset{\sim}{\alpha}$ be a $p \times 1$ column vector such that $\underset{\sim}{\alpha}'\underset{\sim}{\alpha} = 1$ and we call it as a normalized vector. The variance of a linear function, say, $\underset{\sim}{\alpha}'\underset{\sim}{X}$ of $X_1, X_2, \ldots, X_p$ is

$$Var(\underset{\sim}{\alpha}'\underset{\sim}{X}) = \underset{\sim}{\alpha}'\Sigma\underset{\sim}{\alpha}.$$

We aim to find the linear function $\underset{\sim}{\alpha}'\underset{\sim}{X}$ such that the variance of $\underset{\sim}{\alpha}'\underset{\sim}{X}$ is maximum and we consider only normalized vector, so $\underset{\sim}{\alpha}'\underset{\sim}{\alpha} = 1$. So we maximize $\underset{\sim}{\alpha}'\Sigma\underset{\sim}{\alpha}$ such that $\underset{\sim}{\alpha}'\underset{\sim}{\alpha} = 1$. If we could find such $\underset{\sim}{\alpha}$, then $\underset{\sim}{\alpha}'\underset{\sim}{X}$ will be the principal component.

Analysis

Let $Y_1 = \underset{\sim}{\alpha}_1'\underset{\sim}{X}$ is to be obtained given $\underset{\sim}{\alpha}_1'\underset{\sim}{\alpha}_1 = 1$ such that $\underset{\sim}{\alpha}_1'\Sigma\underset{\sim}{\alpha}_1$ is maximum.

Consider the Lagrangian function

$$\phi_1 = \underset{\sim}{\alpha}_1'\Sigma\underset{\sim}{\alpha}_1 - \lambda(\underset{\sim}{\alpha}_1'\underset{\sim}{\alpha}_1 - 1)$$

where λ is a Lagrangian multiplier. The vector of partial derivatives is set to 0, i.e.,

$$\begin{aligned} \frac{\partial\phi_1}{\partial\underset{\sim}{\alpha}_1} &= 2\Sigma\underset{\sim}{\alpha}_1 - 2\lambda\underset{\sim}{\alpha}_1 = 0 \\ &\Rightarrow (\Sigma - \lambda I)\underset{\sim}{\alpha}_1 = 0, \qquad (1) \end{aligned}$$

i.e., $\underset{\sim}{\alpha}_1$ is the characteristic vector corresponding to the characteristic root λ. The respective characteristic equation is

$$|\Sigma - \lambda I| = 0$$

which is a polynomial in λ of degree p and therefore has p roots $\lambda_1, \lambda_2, \ldots, \lambda_p$.

From (1), we get

$$\begin{aligned} \Sigma\underset{\sim}{\alpha}_1 &= \lambda\underset{\sim}{\alpha}_1 \qquad (2) \\ \text{or} \qquad \underset{\sim}{\alpha}_1'\Sigma\underset{\sim}{\alpha}_1 &= \lambda\underset{\sim}{\alpha}_1'\underset{\sim}{\alpha}_1 = \lambda \end{aligned}$$

as $\underset{\sim}{\alpha}_1'\underset{\sim}{\alpha}_1 = 1$.

Since $Var(\underset{\sim}{\alpha}_1'\underset{\sim}{X}) = \underset{\sim}{\alpha}_1'\Sigma\underset{\sim}{\alpha}_1$, so

$$\begin{aligned} Var(\underset{\sim}{\alpha}_1'\underset{\sim}{X}) &= \lambda \\ \text{or} \qquad Var(Y_1) &= \lambda. \end{aligned}$$

First Principal Component

Our requirement for $Y_1 = \underset{\sim}{\alpha}'\underset{\sim}{X}$ to be a principal component is that the variance of Y_1, i.e., $Var(Y_1) = \underset{\sim}{\alpha}_1'\Sigma\underset{\sim}{\alpha}_1 = \lambda$ should be maximum. Since λ is one of the characteristic roots out of $\lambda_1, \lambda_2, \ldots, \lambda_p$, so we select the λ having the maximum value. Let λ is maximum at λ_1 and we denote it as

$$\lambda_{(1)} = \lambda_1.$$

Let $\underset{\sim}{\alpha}^{(1)}$ be the characteristic vector corresponding to λ_1. which is obtained as a normalized solution of

$$(\Sigma - \lambda_{(1)}I)\underset{\sim}{\alpha} = 0.$$

Let the solution is attained at $\underset{\sim}{\alpha}_1 = \underset{\sim}{\alpha}^{(1)}$. So $Y_1 = \underset{\sim}{\alpha}^{(1)'}\underset{\sim}{X}$ is the first principal component, which is a normalized function of $X_1, X_2, \ldots, X_p$ having the maximum variance.

Second Principal Component

Now we find the second principal component such that its linear function is normalized, has maximum variance and is uncorrelated with the first principal component.

Let $Y_2 = \underset{\sim}{\alpha}_2'\underset{\sim}{X}$ be the normalized linear function and we find Y_2 such that

(i) the variance of Y_2 is maximum under the normalized constraint $\underset{\sim}{\alpha}_2'\underset{\sim}{\alpha}_2 = 1$ and

(ii) Y_2 is uncorrelated with the first principal component Y_1, or equivalently Y_1 and Y_2 are orthogonal.

First, we analyze the condition required for the lack of correlation as follows: We want

$$E(\underset{\sim}{\alpha}_2'\underset{\sim}{X} \cdot Y_1) = 0.$$

Then

$$\begin{aligned}
E(\underset{\sim}{\alpha}_2'\underset{\sim}{X}\cdot Y_1) &= E(\underset{\sim}{\alpha}_2'\underset{\sim}{X}\cdot \underset{\sim}{X}'\underset{\sim}{\alpha}^{(1)})\\
&= \underset{\sim}{\alpha}_2'\Sigma\underset{\sim}{\alpha}^{(1)}\\
&= \underset{\sim}{\alpha}_2'(\lambda_1\underset{\sim}{\alpha}^{(1)}) \quad (\text{using } (2), \Sigma\underset{\sim}{\alpha}_1 = \lambda\underset{\sim}{\alpha}_1)\\
&= \lambda_1(\underset{\sim}{\alpha}_2'\underset{\sim}{\alpha}^{(1)}). \qquad (3)
\end{aligned}$$

So

$$\begin{aligned}
E(\underset{\sim}{\alpha}_2'\underset{\sim}{X}\cdot Y_1) &= 0\\
\Rightarrow \quad \lambda_1\underset{\sim}{\alpha}_2'\underset{\sim}{\alpha}^{(1)} &= 0.
\end{aligned}$$

Since $\lambda_1 \neq 0$ as λ_1 is the variance of first principal component, so $\underset{\sim}{\alpha}_2'\underset{\sim}{\alpha}^{(1)} = 0$, i.e., $\underset{\sim}{\alpha}_2$ and $\underset{\sim}{\alpha}^{(1)}$ are orthogonal.

So we consider the following Lagrangian function for maximization:

$$\phi_2 = \underset{\sim}{\alpha}_2'\Sigma\underset{\sim}{\alpha}_2 - \lambda(\underset{\sim}{\alpha}_2'\underset{\sim}{\alpha}_2 - 1) - 2\mu_1(\underset{\sim}{\alpha}_2'\Sigma\underset{\sim}{\alpha}^{(1)})$$

where λ and μ_1 are the Lagrangian multipliers.

The vector of partial derivatives is set to 0 as follows:

$$\begin{aligned}
&\frac{\partial\phi_2}{\partial\underset{\sim}{\alpha}_2} = 2\Sigma\underset{\sim}{\alpha}_2 - 2\lambda\underset{\sim}{\alpha}_2 - 2\mu_1\Sigma\underset{\sim}{\alpha}^{(1)} = 0\\
\text{or}\quad &\Sigma\underset{\sim}{\alpha}_2 - \lambda\underset{\sim}{\alpha}_2 - \mu_1\Sigma\underset{\sim}{\alpha}^{(1)} = 0 \qquad (4)\\
\text{or}\quad &\underset{\sim}{\alpha}^{(1)'}\Sigma\underset{\sim}{\alpha}_2 - \lambda\underset{\sim}{\alpha}^{(1)'}\underset{\sim}{\alpha}_2 - \mu_1\underset{\sim}{\alpha}^{(1)'}\Sigma\underset{\sim}{\alpha}^{(1)} = 0\\
\text{or}\quad &0 - 0 - \mu_1\lambda_1 = 0.
\end{aligned}$$

Since $\lambda_1 \neq 0$, so $\mu_1 = 0$. Substituting $\mu_1 = 0$ in (4), we get

$$\begin{aligned}
\Sigma\underset{\sim}{\alpha}_2 - \lambda\underset{\sim}{\alpha}_2 &= 0\\
\text{or}\quad (\Sigma - \lambda I)\underset{\sim}{\alpha}_2 &= 0,
\end{aligned}$$

i.e., $\underset{\sim}{\alpha}_2$ should satisfy $(\Sigma - \lambda I)\underset{\sim}{\alpha}_2 = 0$ which means that $\underset{\sim}{\alpha}_2$ is the characteristic vector corresponding to the characteristic root of $|\Sigma - \lambda I| = 0$.

Let λ_2 be the second largest root out of $\lambda_1, \lambda_2, \ldots, \lambda_p$. So we denote $\lambda_{(2)} = \lambda_2$.

Let $\underset{\sim}{\alpha}_2$ be the characteristic root corresponding to $\lambda_{(2)}$ which satisfies

$$\begin{aligned}(\Sigma - \lambda_{(2)} I)\underset{\sim}{\alpha}_2 &= 0,\\ \underset{\sim}{\alpha}'_2\underset{\sim}{\alpha}_2 &= 1\end{aligned}$$

and satisfying $E(\underset{\sim}{\alpha}'_2\underset{\sim}{X} \cdot Y_1) = 0$. We denote such $\underset{\sim}{\alpha}_2$ as $\underset{\sim}{\alpha}^{(2)}$ and the corresponding function as a linear combination of $X_1, X_2, \ldots, X_p$ is

$$Y_2 = \underset{\sim}{\alpha}^{(2)'}\underset{\sim}{X}$$

with variance $\lambda_{(2)}$, i.e., $Var(Y_2) = \lambda_{(2)} = \lambda_2$. This is the second principal component and is uncorrelated with the first principal component Y_1, having the second largest variance λ_2.

(r+1)th Principal Component

The procedure is continued. At the $(r+1)^{th}$ step, we want to find $\underset{\sim}{\alpha}$ in $\underset{\sim}{\alpha}'\underset{\sim}{X}$ such that $\underset{\sim}{\alpha}'\underset{\sim}{X}$ has the maximum variance of all normalized linear combinations which are uncorrelated with first r principal components $Y_1, Y_2, \ldots, Y_r$.

Suppose the value of $\underset{\sim}{\alpha}$ at $(r+1)^{th}$ step is $\underset{\sim}{\alpha}_{r+1}$. The condition for $\underset{\sim}{\alpha}'_{r+1}$ to be uncorrelated with first Y_i's, $i = 1, 2, \ldots, r$ is

$$\begin{aligned}0 &= E(\underset{\sim}{\alpha}'_{r+1}\underset{\sim}{X} \cdot Y_i) = E(\underset{\sim}{\alpha}'_{r+1}\underset{\sim}{X} \cdot \underset{\sim}{X}'\underset{\sim}{\alpha}^{(i)})\\ &= \underset{\sim}{\alpha}'_{r+1}\Sigma\underset{\sim}{\alpha}^{(i)}\\ &= \lambda_{(i)}\underset{\sim}{\alpha}'_{r+1}\underset{\sim}{\alpha}^{(i)}, \quad i = 1, 2, \ldots, r.\end{aligned} \tag{5}$$

Since $\lambda_{(i)} \neq 0$, so $\underset{\sim}{\alpha}'_{r+1}\underset{\sim}{\alpha}^{(i)} = 0$, $i = 1, 2, \ldots, r$ is the condition for Y_{r+1} to be uncorrelated with $Y_1, Y_2, \ldots, Y_r$.

Now consider the following Lagrangian function for maximization

$$\phi_{r+1} = \underset{\sim}{\alpha}'_{r+1}\Sigma\underset{\sim}{\alpha}_{r+1} - \lambda(\underset{\sim}{\alpha}'_{r+1}\underset{\sim}{\alpha}_{r+1} - 1) - 2\sum_{i=1}^{r}\mu_i\underset{\sim}{\alpha}'_{r+1}\Sigma\underset{\sim}{\alpha}^{(i)}$$

where $\lambda, \mu_1, \mu_2, \dots, \mu_r$ are the Lagrangian multiplier.

The vector of partial derivatives is set to 0 as follows.

$$\frac{\partial \phi_{r+1}}{\partial \underset{\sim}{\alpha}_{r+1}} = 2\Sigma \underset{\sim}{\alpha}_{r+1} - 2\lambda \underset{\sim}{\alpha}_{r+1} - 2\sum_{i=1}^{r} \mu_i \Sigma \underset{\sim}{\alpha}^{(i)} = 0 \tag{6}$$

Let λ_{r+1} be the $(r+1)^{th}$ maximum of $\lambda_1, \lambda_2, \dots, \lambda_p$, denoted as $\lambda_{(r+1)} = \lambda_{r+1}$, such that $\underset{\sim}{\alpha}_{r+1}$ satisfies

$$(\Sigma - \lambda_{(r+1)})\underset{\sim}{\alpha}_{r+1} = 0$$

such that $\underset{\sim}{\alpha}'_{r+1}\underset{\sim}{\alpha}_{r+1} = 1$ and lack of correlation condition, like (3), are satisfied. Let this vector be $\underset{\sim}{\alpha}^{(r+1)}$ and the corresponding linear combination be

$$Y_{r+1} = \underset{\sim}{\alpha}^{(r+1)}\underset{\sim}{X}.$$

If $\lambda_{(r+1)} = 0$ and $\lambda_{(j)} = 0$ $(j \neq r+1)$, then $\underset{\sim}{\alpha}^{(j)'}\Sigma\underset{\sim}{\alpha}^{(r+1)} = 0$ does not imply that $\underset{\sim}{\alpha}^{(j)'}\underset{\sim}{\alpha}^{(r+1)} = 0$.

However $\lambda^{(r+1)}$ can be replaced by a linear combination of $\underset{\sim}{\alpha}^{(r+1)}$ and $\underset{\sim}{\alpha}^{(j)}$'s with $\lambda_{(j)} = 0$, so that the new $\underset{\sim}{\alpha}^{(r+1)}$ is orthogonal to all $\underset{\sim}{\alpha}^{(j)}$, $(j = 1, 2, \dots, r)$.

Continuing with (6), we have

$$\begin{aligned} & 2\Sigma\underset{\sim}{\alpha}_{r+1} - 2\lambda\underset{\sim}{\alpha}_{r+1} - 2\sum_{i=1}^{r}\mu_i\Sigma\underset{\sim}{\alpha}^{(i)} = 0 \\ \text{or} \quad & \underset{\sim}{\alpha}^{(j)'}\Sigma\underset{\sim}{\alpha}_{r+1} - \lambda\underset{\sim}{\alpha}^{(j)'}\underset{\sim}{\alpha}_{r+1} - \sum_{i=1}^{r}\mu_i\underset{\sim}{\alpha}^{(j)'}\Sigma\underset{\sim}{\alpha}^{(i)} = 0. \end{aligned}$$

Now using

$$\underset{\sim}{\alpha}^{(j)'}\Sigma\underset{\sim}{\alpha}^{(i)} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j = 1, 2, \dots, r, \end{cases}$$

we further simplify

$$\begin{aligned} & \underset{\sim}{\alpha}^{(j)'}\Sigma\underset{\sim}{\alpha}_{r+1} - \lambda\underset{\sim}{\alpha}^{(j)'}\underset{\sim}{\alpha}_{r+1} \\ & \quad - \{\mu_1\underset{\sim}{\alpha}^{(j)'}\Sigma\underset{\sim}{\alpha}^{(1)} + \mu_2\underset{\sim}{\alpha}^{(j)'}\Sigma\underset{\sim}{\alpha}^{(2)} + \ldots + \mu_j\underset{\sim}{\alpha}^{(j)'}\Sigma\underset{\sim}{\alpha}^{(j)} + \ldots + \mu_r\underset{\sim}{\alpha}^{(j)'}\Sigma\underset{\sim}{\alpha}^{(r)}\} = 0 \\ \text{or} \quad & 0 - 0 - \mu_j\lambda_{(j)} = 0. \end{aligned}$$

Since $\lambda_{(j)} \neq 0$, so $\mu_j = 0,\ j = 1, 2, \ldots, r$. Thus $\underset{\sim}{\alpha}$ must satisfy

$$(\Sigma - \lambda I)\underset{\sim}{\alpha}_{r+1} = 0$$

and therefore λ must satisfy $|\Sigma - \lambda I| = 0$.

How Many Principal Components be Constructed?

The next question is: what value of r can we continue to use to create such principal components? The answer is that we can continue up to $r = p$, which is the number of variables in $\underset{\sim}{X}$, i.e., if there are p random variables in the random vector, then p principal components can be constructed. We prove this result as follows:

Suppose the procedure is carried on until at the $(m+1)^{st}$ stage one cannot find a vector $\underset{\sim}{\alpha}$ satisfying $\underset{\sim}{\alpha}'\underset{\sim}{\alpha} = 1$, $(\Sigma - \lambda I)\underset{\sim}{\alpha} = 0$ and condition for independence (as in (5)). Either $m = p$ or $m < p$ since $\underset{\sim}{\alpha}^{(1)}, \underset{\sim}{\alpha}^{(2)}, \ldots, \underset{\sim}{\alpha}^{(m)}$ must be linearly independent.

Since Σ is a positive definite matrix, it has p characteristic roots. Now we show that $m < p$ condition leads to a contradiction.

If $m < p$, then there exist a $(p - m)$ vectors, say $\underset{\sim}{e}_{m+1}, \underset{\sim}{e}_{m+2}, \ldots, \underset{\sim}{e}_p$ such that

$$\underset{\sim}{\alpha}^{(i)'}\underset{\sim}{e}_j = 0,\ \underset{\sim}{e}_i'\underset{\sim}{e}_j = \delta_{ij}$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

Using the following lemma

Lemma 1: Let A be $n \times m$ $(n > m)$ matrix such that $A'A = I_m$, then there exist an $n \times (n - m)$ matrix B such that $(A \quad B)$ is orthogonal,

we find that, $\underset{\sim}{\alpha}^{(2)}$ is orthogonal to $\underset{\sim}{\alpha}^{(1)}$; $\underset{\sim}{\alpha}^{(3)}$ is orthogonal to $\underset{\sim}{\alpha}^{(1)}$ and $\underset{\sim}{\alpha}^{(2)}$; and so on, i.e., all the characteristic vectors are orthogonal to every preceding characteristic vector.

Let $E = (\underset{\sim}{e}_{m+1}, \underset{\sim}{e}_{m+2}, \ldots, \underset{\sim}{e_p})$ with $\underset{\sim}{\alpha}^{(i)'}\underset{\sim}{e}_j = 0,\ i = 1, 2, \ldots, m;\ j = m+1, \ldots, p.$

Consider the roots $|E'\Sigma E - \theta I| = 0$ and a corresponding characteristic vector $\underset{\sim}{c}$, so

$$(E'\Sigma E - \theta I)\underset{\sim}{c} = 0. \tag{7}$$

We have

$$\begin{aligned} \underset{\sim}{\alpha}^{(i)}\Sigma &= \lambda_{(i)}\underset{\sim}{\alpha}^{(i)'}, \quad i = 1, 2, \ldots, m \\ \text{or} \quad \underset{\sim}{\alpha}^{(i)'}\Sigma E\underset{\sim}{c} &= \lambda_{(i)}\underset{\sim}{\alpha}^{(i)'}E\underset{\sim}{c} \\ &= \lambda_{(i)}\underset{\sim}{\alpha}^{(i)'}\sum_{j=m+1}^{p} c_j\underset{\sim}{e}_j \\ &= \lambda_{(i)}\sum_{j=m+1}^{p} c_j \\ &= 0 \qquad (\text{since } \underset{\sim}{\alpha}^{(i)'}\underset{\sim}{e}_j = 0) \end{aligned}$$

which implies that $\underset{\sim}{\alpha}^{(i)}$ is orthogonal to $\Sigma E\underset{\sim}{c}$ for all $i = 1, 2, \ldots, m$ and therefore $\Sigma E\underset{\sim}{c}$ is a vector in the space spanned by $\underset{\sim}{e}_{m+1}, \underset{\sim}{e}_{m+2}, \ldots, \underset{\sim}{e}_p$ and can be written as $E\underset{\sim}{g}$ where $\underset{\sim}{g}$ is $(p-m)$ component vector. Since $\underset{\sim}{\alpha}^{(i)'}\underset{\sim}{e}_j c_j$ and $\underset{\sim}{\alpha}^{(i)'}\underset{\sim}{e}_j$ are 0, so they belong to the same space. Thus

$$\begin{aligned} \Sigma E\underset{\sim}{c} &= E\underset{\sim}{g} \\ \text{or} \quad E'\Sigma E\underset{\sim}{c} &= E'E\underset{\sim}{g} \\ &= \underset{\sim}{g}. \end{aligned}$$

Thus

$$\begin{aligned} & \underset{\sim}{g} = \theta\underset{\sim}{c} \\ \text{or} \quad & \Sigma E\underset{\sim}{c} = \theta E\underset{\sim}{c} \qquad (\text{from } (7) \\ \text{or} \quad & (\Sigma - \theta I)E\underset{\sim}{c} = 0. \end{aligned}$$

Then $(E\underset{\sim}{c})'\underset{\sim}{X}$ is uncorrelated with $\underset{\sim}{\alpha}^{(j)'}\underset{\sim}{X}$, $j = 1, 2, \ldots, m.$ and thus leads to a new $\underset{\sim}{\alpha}^{(m+1)}$.

Since this implies $m < p$, we must have $m = p$.

So now we have characteristic roots $\lambda_{(1)}, \lambda_{(2)}, \ldots, \lambda_{(p)}$ with corresponding characteristic vectors $\underset{\sim}{\alpha}^{(1)}, \underset{\sim}{\alpha}^{(2)}, \ldots, \underset{\sim}{\alpha}^{(p)}$ respectively.

Computation of Characteristic Roots and Characteristic Vectors

Now we compile the characteristic roots and characteristic vectors as follows:

Let $A = (\underset{\sim}{\alpha}^{(1)}, \underset{\sim}{\alpha}^{(2)}, \ldots, \underset{\sim}{\alpha}^{(p)})$ and $\Lambda = diag(\lambda_{(1)}, \lambda_{(2)}, \ldots, \lambda_{(p)})$. The equation $\Sigma\underset{\sim}{\alpha}^{(r)} = \lambda_{(r)}\underset{\sim}{\alpha}^{(r)}$ can be written in matrix form as

$$\begin{aligned} \Sigma A &= A\Lambda \\ \text{or} \quad A'\Sigma A &= A'A\Lambda = \Lambda. \end{aligned}$$

We have $\underset{\sim}{\alpha}^{(r)'}\underset{\sim}{\alpha}^{(r)} = 1$ and $\underset{\sim}{\alpha}^{(r)'}\underset{\sim}{\alpha}^{(s)} = 0$ if $r \neq s$ and this can be summarized as $A'A = I$. Based on this, the following theorem describes the result.

Theorem: Let the p-component random vector $\underset{\sim}{X}$ have $E(\underset{\sim}{X}) = 0$ and $E(\underset{\sim}{X}\underset{\sim}{X}') = \Sigma$ is a positive definite symmetric matrix. Then there exists an orthogonal linear transformation $Y = A'\underset{\sim}{X}$ such that the covariance matrix of Y (which consists of vectors of principal components) is

$$E(YY') = \Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \ldots & \lambda_p \end{pmatrix}$$

where $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p \geqslant 0$ are the characteristic roots of $|\Sigma - \lambda I| = 0$. The r^{th} column of A, i.e., $\underset{\sim}{\alpha}^{(r)}$ satisfies $(\Sigma - \lambda_r I)\underset{\sim}{\alpha}^{(r)} = 0$. The r^{th} component of Y, i.e., $Y_r = \underset{\sim}{\alpha}^{(r)'}\underset{\sim}{X}$ has maximum variance of all normalized linear combination uncorrelated with $Y_1, Y_2, \ldots, Y_{r-1}$.

It follows that

$$\begin{aligned} tr\Sigma &= \sum_{i=1}^{p} Var(X_i) \\ &= tr\ A\Lambda A' \\ &= tr\ \Lambda AA' \\ &= tr\ \Lambda \\ &= \sum_{i=1}^{p} \lambda_i \end{aligned}$$

and $|\Sigma| = |\Lambda| = \prod_{i=1}^{p} \lambda_i$. Note that $tr\Sigma$ is the total variance in the population, which is nothing but the total variance of principal components. Thus, the total variance of the original observations equals the total variance of the new principal components.

For example, suppose there are five variables, and we want to reduce the dimensionality, i.e., reduce the dimension from five. So we determine the characteristic roots of its covariance matrix. Suppose

$$\begin{aligned}
\lambda_1 = 0.60 = Var(Y_1) &\text{ where } Y_1 = \underset{\sim}{\alpha}^{(1)'}\underset{\sim}{X} \\
\lambda_2 = 0.20 = Var(Y_2) &\text{ where } Y_2 = \underset{\sim}{\alpha}^{(2)'}\underset{\sim}{X} \\
\lambda_3 = 0.10 = Var(Y_3) &\text{ where } Y_3 = \underset{\sim}{\alpha}^{(3)'}\underset{\sim}{X} \\
\lambda_4 = 0.06 = Var(Y_4) &\text{ where } Y_4 = \underset{\sim}{\alpha}^{(4)'}\underset{\sim}{X} \\
\lambda_5 = 0.04 = Var(Y_5) &\text{ where } Y_5 = \underset{\sim}{\alpha}^{(5)'}\underset{\sim}{X}.
\end{aligned}$$

We observe that $\lambda_1 = 60\%, \lambda_2 = 20\%, \lambda_3 = 10\%, \lambda_4 = 6\%$ and $\lambda_5 = 4\%$ of the total variance $(\sum_{i=1}^{5} \lambda_i)$ as 100%.

The variation captured by Y_1 and Y_2 is $(60 + 20)\% = 80\%$ of the total variation. The variation captured by Y_1, Y_2 and Y_3 is $(60 + 20 + 10)\% = 90\%$ of the total variation. The variation captured by the first four principal components Y_1, Y_2, Y_3 and Y_4 is $(60 + 20 + 10 + 6)\% = 96\%$.

The first three principal components account for 90% of the total variation, reducing the dimension from five to three. Similarly, the first four principal components take care of 96% of the total variation, so the dimension is reduced to four from five. This is how the dimensionality reduction is achieved.

Recall the following

Theorem 3: (Spectral Decomposition Theorem) Any symmetric matrix A of order $p \times p$ can be written as $A = \Gamma\Lambda\Gamma' = \sum_{i=1}^{p} \lambda_i \underset{\sim}{\gamma}_{(i)} \underset{\sim}{\gamma}'_{(i)}$ where $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$ is a diagonal matrix of characteristic roots of A and $\Gamma = (\underset{\sim}{\gamma}_{(1)}, \underset{\sim}{\gamma}_{(2)}, \ldots, \underset{\sim}{\gamma}_{(p)})$ is an orthogonal matrix of standardized characteristic vectors $\underset{\sim}{\gamma}_{(i)}$'s.

Using the spectral decomposition theorem, we can express the symmetric matrix Σ

as

$$\begin{aligned}\Sigma &= \Gamma\Lambda\Gamma' \\ &= \lambda_1 \underset{\sim}{\alpha}^{(1)}\underset{\sim}{\alpha}^{(1)'} + \lambda_2 \underset{\sim}{\alpha}^{(2)}\underset{\sim}{\alpha}^{(2)'} \ldots + \lambda_p \underset{\sim}{\alpha}^{(p)}\underset{\sim}{\alpha}^{(p)'}\end{aligned}$$

where Γ is an orthogonal matrix. Note that $\lambda_1 \underset{\sim}{\alpha}^{(1)}\underset{\sim}{\alpha}^{(1)'}$ is the contribution to Σ from the first principal component, $\lambda_2 \underset{\sim}{\alpha}^{(2)}\underset{\sim}{\alpha}^{(2)'}$ is the contribution to Σ from the second principal component, lastly $\lambda_p \underset{\sim}{\alpha}^{(p)}\underset{\sim}{\alpha}^{(p)'}$ is the contribution to Σ from the last p^{th} principal component and λ_i is the maximum variance contributed by the i^{th} principal component.

Usually, the values of λ_i become generally small after a few λ_i's and Σ can be approximated by the first few λ_i's. For example, suppose the values of λ_i's become very small after three λ_i's $(i = 1, 2, 3)$ other λ_i's $(i = 4, 5, \ldots, p)$ become negligible. So

$$\begin{aligned}\Sigma &\approx \lambda_1 \underset{\sim}{\alpha}^{(1)}\underset{\sim}{\alpha}^{(1)'} + \lambda_2 \underset{\sim}{\alpha}^{(2)}\underset{\sim}{\alpha}^{(2)'} + \lambda_3 \underset{\sim}{\alpha}^{(3)}\underset{\sim}{\alpha}^{(3)'} \\ \text{and} \quad tr\Sigma &= \sum_{i=1}^{p} \lambda_i.\end{aligned}$$

Case When More than One λ_i's Are the Same

If $\lambda_{r+1} = \lambda_{r+2} = \ldots = \lambda_{r+m} = \lambda$ then $\text{Rank}(\Sigma - \lambda I) = p - m$. The corresponding characteristic vector $(\underset{\sim}{\alpha}_{r+1}, \underset{\sim}{\alpha}_{r+2}, \ldots, \underset{\sim}{\alpha}_{r+m})$ is uniquely determined except for multiplication from the right by an orthogonal matrix using the following result:

Result 4: Given any symmetric matrix B, there exists an orthogonal matrix C such that $C'BC = D = diag(d_1, d_2, \ldots, d_p)$. If B is positive definite, then $d_i > 0$; If B is positive semi definite, then $d_i \geqslant 0$ for all $i = 1, 2, \ldots, p$,
we find an orthogonal matrix P by writing $P'\Sigma P = \Lambda = diag(\lambda_{(1)}, \lambda_{(2)}, \ldots, \lambda_{(p)})$. Now find P. The columns of P are independent and not identical for the corresponding λ's, but their variances are the same. So we get the corresponding values of $\underset{\sim}{\alpha}$'s from the columns of P. Also

$$|P'\Sigma P| = |P'||\Sigma||P| = |\Sigma||PP'| = |\Sigma|.$$

Maximum Likelihood Estimates of the Principal Components and their Variances

Now we estimate the characteristic vectors and characteristic roots using maximum likelihood estimation.

Theorem 5: Let $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots, \underset{\sim}{x}_N$ be $N(> p)$ observations from $N_p(\underset{\sim}{\mu}, \Sigma)$ where Σ is a matrix with p different characteristic roots. Then a set of maximum likelihood estimates of $\lambda_1, \lambda_2, \ldots, \lambda_p$ and $\underset{\sim}{\alpha}^{(1)}, \underset{\sim}{\alpha}^{(2)}, \ldots, \underset{\sim}{\alpha}^{(p)}$ consists of the roots $\hat{\lambda}_1 > \hat{\lambda}_2 > \ldots > \hat{\lambda}_p$ of

$$|\hat{\Sigma} - \hat{\lambda} I| = 0$$

and a set of corresponding vectors $\underset{\sim}{a}^{(1)}, \underset{\sim}{a}^{(2)}, \ldots, \underset{\sim}{a}^{(p)}$ satisfying

$$(\hat{\Sigma} - \hat{\lambda}_i I)\underset{\sim}{a}^{(i)} = 0$$

with $\underset{\sim}{a}^{(i)\prime}\underset{\sim}{a}^{(i)} = 1$ where $\hat{\Sigma}$ is the maximum likelihood estimate of Σ.

Proof: When the roots of $|\Sigma - \lambda I| = 0$ are different, each $\underset{\sim}{\alpha}^{(i)}$ is uniquely defined except that $\underset{\sim}{\alpha}^{(i)}$ can be replaced by $-\underset{\sim}{\alpha}^{(i)}$. If we require that the first nonzero component of $\underset{\sim}{\alpha}^{(i)} > 0$, then $\underset{\sim}{\alpha}^{(i)}$ is uniquely defined and $\underset{\sim}{\mu}$, Λ and A is a single valued function of $\underset{\sim}{\mu}$ and Σ.

By the invariance property of maximum likelihood estimation, the set of maximum likelihood estimates of $\underset{\sim}{\mu}$, Λ and A is the same function of $\hat{\underset{\sim}{\mu}}$ and $\hat{\Sigma}$. This function is defined by

$$\begin{aligned} |\hat{\Sigma} - \hat{\lambda} I| &= 0 \\ (\hat{\Sigma} - \hat{\lambda}_i I)\underset{\sim}{a}^{(i)} &= 0 \\ \underset{\sim}{a}^{(i)\prime}\underset{\sim}{a}^{(i)} &= 1 \end{aligned}$$

with the corresponding restriction that the first nonzero component of $\underset{\sim}{a}^{(i)}$ must be positive.

It can be shown that if $|\Sigma| \neq 0$, the probability is 1 that the roots of

$$|\hat{\Sigma} - kI| = 0$$

are different, because the condition on $\hat{\Sigma}$ for the roots to have multiplicities higher than 1 determines a region in the space of $\hat{\Sigma}$ of dimensionality less than $\frac{p(p+1)}{2}$.

Since

$$\begin{aligned} A'\Sigma A &= \Lambda \\ \text{So} \quad \Sigma &= A\Lambda A' \\ &= \sum_{i=1}^{p} \lambda_i \underset{\sim}{\alpha}^{(i)} \underset{\sim}{\alpha}^{(i)'}. \end{aligned}$$

By the same algebra, we can show

$$\hat{\Sigma} = \sum_{i=1}^{p} \hat{\lambda}_i \underset{\sim}{a}^{(i)} \underset{\sim}{a}^{(i)'}.$$

Replacing $\underset{\sim}{a}^{(i)}$ by $-\underset{\sim}{a}^{(i)}$ clearly does not change $\sum_{i=1}^{p} \hat{\lambda}_i \underset{\sim}{a}^{(i)} \underset{\sim}{a}^{(i)'}$. Since the likelihood function depends only on $\hat{\Sigma}$, the maximum of the likelihood function is attained by taking any set of solutions of $(\hat{\Sigma} - \hat{\lambda}_i I)\underset{\sim}{a}^{(i)} = 0$ and $\underset{\sim}{a}^{(i)'}\underset{\sim}{a}^{(i)} = 1$.

Overcoming the Problem of Units of Variables

When different variables $X_1, X_2, \ldots, X_p$ in the random vector $\underset{\sim}{X}$ have different units of measurement, then it is sometimes difficult to interpret the principal components. To overcome this situation, one can use the correlation matrix instead of the covariance matrix. This is explained in the following.

Suppose Y_i is the i^{th} principal component and X_k is the k^{th} variable from the original data set. The correlation coefficient between Y_i and X_k is

$$\rho(Y_i, X_k) = \frac{Cov(Y_i, X_k)}{\sqrt{Var(Y_i)}\sqrt{Var(X_k)}}.$$

Since $Y_i = \alpha_{i1}X_1 + \alpha_{i2}X_2 + \ldots + \alpha_{ip}X_p$, so

$$\begin{aligned} \rho(Y_i, X_k) &= \frac{\alpha_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \\ \text{or} \quad \rho &\propto \alpha_{ik} \end{aligned}$$

where and 1 occurs at the k^{th} place.

Since the correlation coefficient is independent of units of measurement, the problem is overcome when the variables have different units.

Thus when each X_i in $\underset{\sim}{X}$ has the same units of measurement, we use Σ matrix and when X_i's have different units of measurements, we use the matrix of correlation coefficients to perform the principal component analysis.

Here α_{ik} is the weight assigned to the k^{th} variable X_k in $\underset{\sim}{X}$ in the principal component $\underset{\sim}{\alpha}^{(i)'}\underset{\sim}{X}$.

Another way to deal with variables having different units of measurement is to consider the standardized variables as

$$z = \frac{x - \text{mean of } x}{\text{standard deviation of } x}$$

and the observations are standardized. Now work with standardized variables instead of $X_1, X_2, \ldots, X_p$. In such a case, the covariance matrix of X becomes the correlation matrix as

$$\rho = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1p} \\ \rho_{21} & \rho_{22} & \rho_{23} & \cdots & \rho_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \rho_{p3} & \cdots & 1 \end{pmatrix},$$

whereas

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \sigma_{p3} & \cdots & \sigma_{pp} \end{pmatrix}$$

$$\rho = \Delta\Sigma\Delta$$

where $\Delta = diag\left(\frac{1}{\sqrt{\sigma_{11}}}, \frac{1}{\sqrt{\sigma_{22}}}, \ldots, \frac{1}{\sqrt{\sigma_{pp}}}\right)$.

Find the characteristic roots, say δ, and characteristic vector $\underset{\sim}{q}$ from

$$|\rho - \delta I| = 0$$
$$\text{and} \quad (\rho - \delta I)\underset{\sim}{q}_i = 0.$$

Here, we proceed exactly as with λ and $\underset{\sim}{\alpha}$ to obtain the principal components. In general, there is no relationship between the two sets of vectors and roots of Σ and ρ. So it is not necessary to get the same inferences from Σ and ρ matrices.

The obvious results in the case of ρ matrix are as follows:

- Characteristic roots: $\delta_{(1)} \geqslant \delta_{(2)} \geqslant \ldots \geqslant \delta_{(p)} \geqslant 0$.
- Characteristic vectors: $\underset{\sim}{q}^{(1)}, \underset{\sim}{q}^{(2)}, \ldots, \underset{\sim}{q}^{(p)}$ corresponding to $\delta_1, \delta_2, \ldots, \delta_p$, respectively.
- i^{th} Principal component: $P_i = \underset{\sim}{q}^{(i)'}\underset{\sim}{z}$ where $\underset{\sim}{z} = (z_1, z_2, \ldots, z_p)$ is the $p \times 1$ vector of standardized variables.
- $Var(P_i) = \delta_{(i)}, i = 1, 2, \ldots, p.$
- $\sum_{i=1}^{p} Var(P_i) = \sum_{i=1}^{p} \delta_{(i)} = p = \sum_{i=1}^{p} \delta_i.$

The decision on the extent of dimensionality reduction is taken on the same lines as in the case of Σ.

Some Large Sample Properties

Suppose $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$ and for Σ, the characteristic roots are $\lambda_1 > \lambda_2 > \ldots > \lambda_p \geqslant 0$ and the corresponding characteristic vectors are $\underset{\sim}{\alpha}^{(1)}, \underset{\sim}{\alpha}^{(2)}, \ldots, \underset{\sim}{\alpha}^{(p)}$ respectively. The sample covariance matrix is S with characteristic roots $\hat{\lambda}_1 \geqslant \hat{\lambda}_2 \geqslant \ldots \geqslant \hat{\lambda}_p$ with corresponding characteristic vectors $\underset{\sim}{a}^{(1)}, \underset{\sim}{a}^{(2)}, \ldots, \underset{\sim}{a}^{(p)}$, respectively.

Let

$$d_i = \sqrt{n}(\hat{\lambda}_i - \lambda_i)$$
$$g_i = \sqrt{n}(\underset{\sim}{a}^{(i)} - \underset{\sim}{\alpha}^{(i)}), i = 1, 2, \ldots, p.$$

We have the following results:

1. Then in the limiting normal distribution, the sets $d_1, d_2, \ldots, d_p$ and $g_1, g_2, \ldots, g_p$ are independent. Further, $d_1, d_2, \ldots, d_p$ are mutually independent.

2. The limiting distribution of d_i is $N(0, 2\lambda_i^2)$.

3. The asymptotic covariance matrix of $g_1, g_2, \ldots, g_p$ in the limiting distribution are

$$Var(g_i) = \sum_{\substack{k=1\\k\neq i}}^{p} \frac{\lambda_i\lambda_k}{(\lambda_i-\lambda_k)^2}\underset{\sim}{\alpha}^{(k)}\underset{\sim}{\alpha}^{(k)'}$$

$$align{*}Cov(g_i, g_j) = -\frac{\lambda_i\lambda_j}{(\lambda_i-\lambda_j)^2}\underset{\sim}{\alpha}^{(j)}\underset{\sim}{\alpha}^{(i)'},\ i \neq j.$$

4. For a single ordered root $\hat{\lambda}_i$, the distribution of $\hat{\lambda}_i$ is approximately normal with mean λ_i and variance $\frac{2\lambda_i^2}{n}$. Since $\hat{\lambda}_i$ is a consistent estimate of λ_i, the limiting distribution of $\frac{\sqrt{n}(\hat{\lambda}_i-\lambda_i)}{\sqrt{2}\lambda_i} \sim N(0,1)$.

   A two tailed test of $H_0 : \lambda = \lambda_i^0$ has (asymptotic) acceptance region

$$-Z_{\epsilon/2} \leq \sqrt{\frac{n}{2}}\frac{\hat{\lambda}_i - \lambda_i^0}{\lambda_i^0} \leq Z_{\epsilon/2}$$

   when value of $N(0,1)$ beyond $Z(\epsilon)$ is $\frac{\epsilon}{2}$. This interval can be inverted to give a confidence interval for λ_i with confidence $1-\epsilon$ as follows:

$$\frac{\hat{\lambda}_i}{1+\sqrt{\frac{2}{n}}Z_{\frac{\epsilon}{2}}} \leq \lambda_i \leq \frac{\hat{\lambda}_i}{1-\sqrt{\frac{2}{n}}Z_{\frac{\epsilon}{2}}}.$$

   The confidence coefficient should be taken large enough so $\sqrt{\frac{2}{n}}Z_{\frac{\epsilon}{2}} < 1$. Alternatively, one can use the limiting distribution of $\sqrt{n}(\log\hat{\lambda}_i - \log\lambda_i)$ is $N(0,2)$.

5. $\underset{\sim}{a}^{(i)}$ is approximately normally distributed with mean $\underset{\sim}{\alpha}^{(i)}$ and (singular) covariance matrix with

$$Var(\underset{\sim}{a}^{(i)}) = \frac{1}{n}\sum_{\substack{k=1\\k\neq i}}^{p} \frac{\lambda_i\lambda_k}{(\lambda_i-\lambda_k)^2}\underset{\sim}{\alpha}^{(k)}\underset{\sim}{\alpha}^{(k)'}$$

$$Cov(\underset{\sim}{a}^{(i)}, \underset{\sim}{a}^{(j)}) = -\frac{1}{n}\frac{\lambda_i\lambda_j}{(\lambda_i-\lambda_j)^2}\underset{\sim}{\alpha}^{(j)}\underset{\sim}{\alpha}^{(i)'},\ (i \neq j).$$

Principal Component Regression

The principal component regression in multiple linear regression model $\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$ is useful when the independent variables or regressors $X_1, X_2, \ldots, X_p$ are not independent and consequently multicollinearity problem is present in the data.

Let T be a $p \times p$ orthogonal matrix whose columns are the characteristic vectors associated with $\lambda_1, \lambda_2, \ldots, \lambda_p$ of $X'X$. So the model

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$$

is written as

$$\begin{aligned} \underset{\sim}{y} &= XTT'\underset{\sim}{\beta} + \underset{\sim}{\epsilon} \\ &= Z\underset{\sim}{\alpha} + \underset{\sim}{\epsilon} \end{aligned}$$

where $Z = XT$ and $\underset{\sim}{\alpha} = T'\underset{\sim}{\beta}$. Note that $T'X'XT = Z'Z = \Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$ where Λ is a diagonal matrix of characteristic roots of $X'X$.

Columns of Z now define a new set of orthogonal regressors $Z = (z_1, z_2, \ldots, z_p)$ which are based on principal components. The ordinary least squares estimator of $\underset{\sim}{\alpha}$ is

$$\hat{\underset{\sim}{\alpha}} = (Z'Z)^{-1}Z'\underset{\sim}{y} = \Lambda^{-1}Z'\underset{\sim}{y}$$

and covariance matrix of $\hat{\underset{\sim}{\alpha}}$ is

$$Var(\hat{\underset{\sim}{\alpha}}) = \sigma^2(Z'Z)^{-1} = \sigma^2\Lambda^{-1}.$$

If $\lambda_j = 0$, it indicates a perfect linear relationship between the original regressors. So if one or more λ_j's are close to zero, it indicates the presence of multicollinearity in the data. Thus

$$Var(\hat{\underset{\sim}{\beta}}) = Var(T\hat{\underset{\sim}{\alpha}}) = T\Lambda^{-1}T'\sigma^2$$

and so

$$Var(\hat{\beta}_j) = \sigma^2 \sum_{j=1}^{p} \frac{t_{ji}^2}{\lambda_j}.$$

To obtain the principal component regression estimator of $\underset{\sim}{\beta}$, obtain $\lambda_{(1)} > \lambda_{(2)} > \lambda_{(3)} > \ldots > \lambda_{(p)} > 0$. Suppose the last k λ's are extremely small. Then remove the principal components corresponding to these k λ's and apply the ordinary least squares estimation to the remaining components as follows:

$$\underset{\sim}{\hat{\alpha}}_{pc} = B\underset{\sim}{\hat{\alpha}}$$

where $b_1 = b_2 = \ldots = b_{p-k} = 1$ $b_{p-k+1} = b_{p-k+2} = \ldots = b_p = 0$. Then

$$\underset{\sim}{\hat{\alpha}}_{pc} = (\hat{\alpha}^{(1)}, \hat{\alpha}^{(2)}, \ldots, \hat{\alpha}^{(p-k)}, 0, \ldots, 0).$$

Since

$$\begin{aligned} & \underset{\sim}{\alpha} = T'\underset{\sim}{\beta} \\ \text{or} \quad & T\underset{\sim}{\alpha} = TT'\underset{\sim}{\beta} \\ \text{or} \quad & \underset{\sim}{\beta} = T\underset{\sim}{\alpha}. \end{aligned}$$

So the principal component regression based estimator of $\underset{\sim}{\beta}$ is obtained as

$$\begin{aligned} \underset{\sim}{\hat{\beta}}_{pc} &= T\underset{\sim}{\hat{\alpha}}_{pc} \\ &= \sum_{j=1}^{p-s} \frac{\underset{\sim}{t}'_j X' \underset{\sim}{y} \underset{\sim}{t}_j}{\lambda_j} \end{aligned}$$

Test of Hypothesis for the Last Few roots to be Zero

So far, we have decided the number of principal components to retain by observing the smaller values of the λ_i's. Now we develop a test of hypothesis to test whether the last k characteristic roots, out of p characteristic roots, are zero or not, i.e., the last k values of λ_i's are zero or not.

The null hypothesis is

H_0: All characteristic roots are equal (not necessarily zero),
i.e., all the principal components have the same variance. This null hypothesis is equiv-

alent to say that

$$
\begin{aligned}
|\Sigma - \lambda I| &= 0 \\
\text{or} \quad \Sigma &= \sigma^2 I.
\end{aligned}
$$

Thus the null hypothesis becomes $H_0 : \Sigma = \sigma^2 I.$

We use the likelihood ratio test to test $H_0 : \Sigma = \sigma^2 I.$ We can also test the hypothesis about the characteristic vector $\underset{\sim}{\alpha}^{(i)}$, say is equal to $\underset{\sim}{\alpha}$ or not, given $\underset{\sim}{\alpha}$, by the likelihood ratio test.

Now we further develop likelihood-ratio tests for testing covariance matrices. This will also serve as a test of the number of principal components to retain. So, essentially, we will cover two topics together:
(i) Testing of hypothesis for the equality of covariance matrices and
(ii) testing the hypothesis for the number of principal components to be retained.

Testing Equality of Several Covariance Matrices

Suppose we have q samples from a multivariate normal distribution with different mean vectors and different covariance matrices as

$$\underset{\sim}{x}_{\alpha}^{(g)} \sim N_p(\underset{\sim}{\mu}^{(g)}, \Sigma_g), \;\; \alpha = 1, 2, \ldots, N_g; \;\; g = 1, 2, \ldots, q.$$

The null hypothesis to be tested is

$$H_0 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_q.$$

Let

$$
\begin{aligned}
\sum_{g=1}^{q} N_g &= N, \\
A_g &= \sum_{\alpha=1}^{N_g} (\underset{\sim}{x}_{\alpha}^{(g)} - \bar{\underset{\sim}{x}}^{(g)})(\underset{\sim}{x}_{\alpha}^{(g)} - \bar{\underset{\sim}{x}}^{(g)})', \\
A &= \sum_{g=1}^{q} A_g.
\end{aligned}
$$

Now we develop the likelihood ratio test. The likelihood function is

$$L = \prod_{g=1}^{q} \frac{1}{(2\pi)^{\frac{N_g p}{2}} |\Sigma_g|^{\frac{N_g}{2}}} \exp\left[-\frac{1}{2}\sum_{\alpha=1}^{N_g} (\underset{\sim}{x}_\alpha^{(g)} - \underset{\sim}{\mu}^{(g)})'\Sigma_g^{-1}(\underset{\sim}{x}_\alpha^{(g)} - \underset{\sim}{\mu}^{(g)})\right].$$

The sample spaces are

$\Omega : \{\text{Each } \Sigma_g \text{ is positive definite and } \underset{\sim}{\mu}^{(g)} \text{ is any vector}\}$

$\omega : \{\Sigma_1 = \Sigma_2 = \ldots = \Sigma_q \text{ and } \underset{\sim}{\mu}^{(g)} \text{ is any vector.}\}$

The maximum likelihood estimates of $\underset{\sim}{\mu}^{(g)}$ and Σ_g under Ω and ω are as follows:

Under $\Omega : \hat{\underset{\sim}{\mu}}{}_{\Omega}^{(g)} = \bar{\underset{\sim}{x}}^{(g)},\ \hat{\Sigma}_{g\Omega} = \frac{1}{N_g}A_g.$

Under $\omega : \hat{\underset{\sim}{\mu}}{}_{\omega}^{(g)} = \bar{\underset{\sim}{x}}^{(g)},\ \hat{\Sigma}_{\omega} = \frac{A}{N}.$

Note that when $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_q = \Sigma$, then

$$L = \frac{1}{(2\pi)^{\frac{Np}{2}} |\Sigma|^{\frac{N}{2}}} \exp\left[-\frac{1}{2}\sum_{g=1}^{q}\sum_{\alpha=1}^{N_g} (\underset{\sim}{x}_\alpha^{(g)} - \bar{\underset{\sim}{x}}^{(g)})'\Sigma^{-1}(\underset{\sim}{x}_\alpha^{(g)} - \bar{\underset{\sim}{x}}^{(g)})\right]$$

has to be maximized with respect to Σ which gives $\hat{\Sigma} = \frac{A}{N}$.

Next we find the L under ω and Ω. as follows:

Under ω,

$$\begin{aligned}
L_\omega &= \prod_{g=1}^{q} \frac{1}{(2\pi)^{\frac{N_g p}{2}} |\hat{\Sigma}_\omega|^{\frac{N_g}{2}}} \exp\left[-\frac{1}{2}\sum_{\alpha=1}^{N_g} (\underset{\sim}{x}_\alpha^{(g)} - \bar{\underset{\sim}{x}}^{(g)})'\hat{\Sigma}_\omega^{-1}(\underset{\sim}{x}_\alpha^{(g)} - \bar{\underset{\sim}{x}}^{(g)})\right] \\
&= \frac{1}{(2\pi)^{\frac{Np}{2}} |\hat{\Sigma}_\omega|^{\frac{N}{2}}} \exp\left[-\frac{1}{2}\sum_{g=1}^{q}\sum_{\alpha=1}^{N_g} (\underset{\sim}{x}_\alpha^{(g)} - \bar{\underset{\sim}{x}}^{(g)})'\hat{\Sigma}_\omega^{-1}(\underset{\sim}{x}_\alpha^{(g)} - \bar{\underset{\sim}{x}}^{(g)})\right] \\
&= \frac{1}{(2\pi)^{\frac{Np}{2}} |\hat{\Sigma}_\omega|^{\frac{N}{2}}} \exp\left[-\frac{Np}{2}\right].
\end{aligned}$$

Under Ω,

$$\begin{aligned}
L_\Omega &= \prod_{g=1}^{q} \frac{1}{(2\pi)^{\frac{N_g p}{2}} |\hat{\Sigma}_{g\Omega}|^{\frac{N_g}{2}}} \exp\left[-\frac{1}{2}\sum_{\alpha=1}^{N_g}(\underset{\sim}{x}_\alpha^{(g)} - \bar{\underset{\sim}{x}}^{(g)})'\hat{\Sigma}_{g\Omega}^{-1}(\underset{\sim}{x}_\alpha^{(g)} - \bar{\underset{\sim}{x}}^{(g)})\right] \\
&= \frac{1}{(2\pi)^{\frac{Np}{2}}\left[\prod_{g=1}^{q} |\hat{\Sigma}_{g\Omega}|^{\frac{N_g}{2}}\right]} \exp\left[-\frac{1}{2}\sum_{g=1}^{q}\sum_{\alpha=1}^{N_g}(\underset{\sim}{x}_\alpha^{(g)} - \bar{\underset{\sim}{x}}^{(g)})'\hat{\Sigma}_{g\Omega}^{-1}(\underset{\sim}{x}_\alpha^{(g)} - \bar{\underset{\sim}{x}}^{(g)})\right] \\
&= \frac{1}{(2\pi)^{\frac{Np}{2}}\left[\prod_{g=1}^{q} |\hat{\Sigma}_{g\Omega}|^{\frac{N_g}{2}}\right]} \exp\left(-\frac{Np}{2}\right).
\end{aligned}$$

The likelihood ratio criterion is

$$\begin{aligned}
\lambda_1 &= \frac{\prod_{g=1}^{q} |\hat{\Sigma}_{g\Omega}|^{\frac{N_g}{2}}}{|\hat{\Sigma}_\omega|^{\frac{N}{2}}} \\
&= \frac{\prod_{g=1}^{q} |A_g|^{\frac{N_g}{2}}}{|A|^{\frac{N}{2}}} \cdot \frac{N^{\frac{Np}{2}}}{\prod_{g=1}^{q} N_g^{\frac{N_g p}{2}}}.
\end{aligned}$$

[Recall that a similar algebra was used in deriving Hotelling's T^2 statistic.]

The critical region is

$$\lambda_1 \leq \lambda_1(\epsilon)$$

where $\lambda(\epsilon)$ is defined so that

$$P[\lambda_1 \leq \lambda_1(\epsilon)] = \epsilon.$$

Testing the Independence of Sets of Variables

Now we develop a likelihood ratio test for testing the independence of sets of variates.

Let $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$. Partition the $p \times 1$ vector $\underset{\sim}{X}$ into q subvectors $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_q$ having $p_1, p_2, \ldots, p_q$ components, respectively. Denote $\underset{\sim}{X}$ and partition $\underset{\sim}{\mu}$, and Σ as follows:

$$\underset{\sim}{X} = \begin{pmatrix} \underset{\sim}{X}_1 \\ \underset{\sim}{X}_2 \\ \vdots \\ \underset{\sim}{X}_q \end{pmatrix}, \quad \underset{\sim}{\mu} = \begin{pmatrix} \underset{\sim}{\mu}_1 \\ \underset{\sim}{\mu}_2 \\ \vdots \\ \underset{\sim}{\mu}_q \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \ldots & \Sigma_{1q} \\ \Sigma_{21} & \Sigma_{22} & \ldots & \Sigma_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{q1} & \Sigma_{q2} & \ldots & \Sigma_{qq} \end{pmatrix}$$

with $p_1 + p_2 + \ldots + p_q = p$. The null hypothesis is

$H_0 : \underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_q$ are independent,

i.e., $f(\underset{\sim}{X})$ can be partitioned into the densities of $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_q$.

So H_0 can be expressed as

$H_0 : N_p(\underset{\sim}{X}|\underset{\sim}{\mu}, \Sigma) = \prod_{i=1}^{q} N_{p_i}(\underset{\sim}{X}_i|\underset{\sim}{\mu}_i, \Sigma_{ii})$.

If $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \ldots, \underset{\sim}{X}_q$ are independent subvectors then $E[(\underset{\sim}{X}_i - \underset{\sim}{\mu}_i)(\underset{\sim}{X}_j - \underset{\sim}{\mu}_j)'] = \Sigma_{ij} = 0$.

Conversely, if $E[(\underset{\sim}{X}_i - \underset{\sim}{\mu}_i)(\underset{\sim}{X}_j - \underset{\sim}{\mu}_j)'] = \Sigma_{ij} = 0$, then

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 & \ldots & 0 \\ 0 & \Sigma_{22} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \Sigma_{qq} \end{pmatrix}.$$

The likelihood ratio test statistic is

$$\lambda = \frac{\underset{\underset{\sim}{\mu}, \Sigma_0}{\text{Max}} L(\underset{\sim}{\mu}, \Sigma_0)}{\underset{\underset{\sim}{\mu}, \Sigma}{\text{Max}} L(\underset{\sim}{\mu}, \Sigma)}.$$

Under Ω,

$$\underset{\underset{\sim}{\mu}, \Sigma}{\text{Max}} L(\underset{\sim}{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{Np}{2}} |\hat{\Sigma}_\Omega|^{N/2}} \exp\left(-\frac{Np}{2}\right)$$

where $\hat{\Sigma}_\Omega = \frac{1}{N} A = \frac{1}{N} \sum_{\alpha=1}^{N} (\underset{\sim}{x}_\alpha - \underset{\sim}{\bar{x}})(\underset{\sim}{x}_\alpha - \underset{\sim}{\bar{x}})'$.

Under H_0,

$$L(\underset{\sim}{\mu}, \Sigma) = \prod_{i=1}^{q} L_i(\underset{\sim}{\mu}_i, \Sigma_{ii}),$$

so

$$\begin{aligned} \underset{\underset{\sim}{\mu}, \Sigma_0}{\text{Max}} L(\underset{\sim}{\mu}, \Sigma) &= \prod_{i=1}^{q} \underset{\underset{\sim}{\mu}_i, \Sigma_{ii}}{\max} L_i(\underset{\sim}{\mu}_i, \Sigma_{ii}) \\ &= \prod_{i=1}^{q} \frac{1}{(2\pi)^{\frac{Np_i}{2}} |\hat{\Sigma}_{ii\omega}|^{\frac{N}{2}}} \cdot \exp\left(-\frac{Np_i}{2}\right) \\ &= \frac{1}{(2\pi)^{\frac{Np}{2}} \prod_{i=1}^{q} |\hat{\Sigma}_{ii\omega}|^{N/2}} \exp\left(-\frac{Np}{2}\right) \end{aligned}$$

where $\hat{\Sigma}_{ii\omega} = \frac{1}{N}\sum_{\alpha=1}^{N}(\underset{\sim}{x}_{\alpha i} - \bar{\underset{\sim}{x}}_i)(\underset{\sim}{x}_{\alpha i} - \bar{\underset{\sim}{x}}_i)'$.

We partition A and $\hat{\Sigma}_\Omega$ on the lines of Σ,

$$A = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1q} \\ A_{21} & A_{22} & \dots & A_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ A_{q1} & A_{q2} & \dots & A_{qq} \end{pmatrix}, \quad \hat{\Sigma}_\Omega = \begin{pmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} & \dots & \hat{\Sigma}_{1q} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} & \dots & \hat{\Sigma}_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\Sigma}_{q1} & \hat{\Sigma}_{q2} & \dots & \hat{\Sigma}_{qq} \end{pmatrix}.$$

Then we see that

$$\hat{\Sigma}_{ii\omega} \equiv \hat{\Sigma}_{ii} = \frac{A_i}{N}.$$

Thus

$$\lambda = \frac{\max\limits_{\underset{\sim}{\mu},\Sigma_0} L(\underset{\sim}{\mu},\Sigma_0)}{\max\limits_{\underset{\sim}{\mu},\Sigma} L(\underset{\sim}{\mu},\Sigma)} = \frac{|\hat{\Sigma}_\Omega|^{N/2}}{\prod_{i=1}^{q}|\hat{\Sigma}_{ii}|^{\frac{N}{2}}} = \frac{|A|^{N/2}}{\prod_{i=1}^{q} A_{ii}^{N/2}}.$$

The critical region for λ is

$$\lambda \leq \lambda(\epsilon)$$

where $\lambda(\epsilon)$ is a constant such that

$$P[\lambda \leq \lambda(\epsilon)] = \epsilon$$

with $\Sigma = \Sigma_0$.

Please note that it remains to show that such a number can be found. Let

$$V = \frac{|A|}{\prod_{i=1}^{q}|A_{ii}|}.$$

Then $\lambda = V^{\frac{N}{2}}$ is a monotonic increasing function of V. The critical region $\lambda \leq \lambda(\epsilon)$ can be written equivalently as $V \leq V(\epsilon)$. The distribution of V under H_0 is the distribution of $V_2, V_3, \dots, V_q$ where $V_2, V_3, \dots, V_q$ are independently distributed with V_i distributed as Wilks Lambda distribution with parameters p_i (dimension), $\bar{p}_i$ ($\bar{p}_i$ components of conditioning vector) and $n - \bar{p}_i$ (degree of freedom), and this is denoted as $\Lambda(p_i, \bar{p}_i, n - \bar{p}_i)$.

The Sphericity Test

Now we test the hypothesis that a covariance matrix is proportional to a given matrix. This is also the sphericity test.

The null hypothesis is

$$H_0 : \Sigma = \sigma^2 I$$

where σ^2 is not specified. This test can be given an algebraic interpretation in terms of characteristic roots of Σ, i.e., the roots of

$$|\Sigma - \phi I| = 0.$$

The null hypothesis is true if and only if all the characteristic roots of $|\Sigma - \phi I| = 0$ are equal. Suppose the characteristic roots are $\phi_1, \phi_2, \ldots, \phi_p$.

Another way to say the same statement is that the arithmetic mean of characteristic roots $\phi_1, \phi_2, \ldots, \phi_p$ is the same as the geometric mean of $\phi_1, \phi_2, \ldots, \phi_p$, i.e.,

$$\frac{\prod_{i=1}^{p} \phi_i^{1/p}}{\sum_{i=1}^{p}(\frac{\phi_i}{p})} = \frac{|\Sigma|^{1/p}}{(\frac{tr\Sigma}{p})} = 1.$$

The $H_0 : \Sigma = \sigma^2 I$ can be expressed in a more general framework as

$$H : \Psi = \sigma^2 \Psi_0$$

where Ψ_0 is specified and observations $\underset{\sim}{y}_1, \underset{\sim}{y}_2, \ldots, \underset{\sim}{y}_n \sim N_p(\underset{\sim}{\gamma}, \Psi)$. Such a hypothesis can be transformed to $H_0 : \Psi = \sigma^2 I$ as follows:

Let C be a matrix such that

$$C\Psi_0 C' = I$$

and let

$$\underset{\sim}{\mu}^* = C\underset{\sim}{\gamma} \ ,$$

$$\Sigma^* = C\Psi C' \ ,$$

$$\underset{\sim}{x}_\alpha^* = C\underset{\sim}{y}_\alpha .$$

Furthrr, let $\underset{\sim}{x}_\alpha^* \sim N_p(\underset{\sim}{\mu}^*, \Sigma^*), \alpha = 1, 2, \ldots, N$. Then H can be written as

$$H : \Sigma^* = \sigma^2 I$$

where $H$is a combination of the hypotheses:

$H_1 : \Sigma^*$ is a diagonal matrix or the components of $\underset{\sim}{X}$ are independent.
and

H_2: The diagonal elements of Σ^* are equal given that the components of $\underset{\sim}{X}$ are independent.

Now we use the following lemma:

Lemma 6: Let

$$\begin{aligned} & \underset{\sim}{y} \sim f(\underset{\sim}{\theta}), \underset{\sim}{\theta} \in \Omega \ , \\ & H_a : \underset{\sim}{\theta} \in \Omega_a \subset \Omega \ , \\ & H_b : \underset{\sim}{\theta} \in \Omega_b \subset \Omega_a \quad \text{given} \ \ \theta \in \Omega_a \ , \\ \text{and} \quad & H_{ab} : \underset{\sim}{\theta} \in \Omega_b \quad \text{given} \ \ \underset{\sim}{\theta} \in \Omega. \end{aligned}$$

Further, let λ_a, λ_b and λ_{ab} are the likelihood ratio test criterion for testing H_a, H_b and H_{ab} respectively are uniquely defined for $\underset{\sim}{y}$, then

$$\lambda_{ab} = \lambda_a \cdot \lambda_b \ .$$

Using this lemma, the likelihood ratio test statistic ($LRTS$) for H can be written as

$LRTS$ λ for $H = (LRTS \ \lambda_1$ for $H_1) \times (LRTS \ \lambda_2$ for $H_2)$.

The $LRTS$ for $H_1 : \Sigma = \sigma^2 I$ is

$$\lambda_1 = \frac{|A|^{N/2}}{\prod_{i=1}^{p} a_{ii}^{N/2}} = |r_{ij}|^{N/2}$$

where $A = \sum_{\alpha=1}^{N} (\underset{\sim}{x}_\alpha - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_\alpha - \bar{\underset{\sim}{x}})' = ((a_{ij}))$ and $r_{ij} = \frac{a_{ij}}{\sqrt{a_{ii} a_{jj}}}$.

To get the $LRTS$ for H_2, we use the results from the testing $H_0 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_q$ as follows:

Consider i^{th} component of $\underset{\sim}{x}_\alpha$ and α^{th} observations from i^{th} population. Here p, N and pN are q, N_g and N, respectively from earlier notations.

Thus

$$\begin{aligned}\lambda_2 &= \frac{\prod_{i=1}^{p}\left[\sum_{\alpha=1}^{N}(x_{i\alpha}-\bar{x}_i)^2\right]^{N/2}}{\left[\sum_i\sum_\alpha(x_{i\alpha}-\bar{x}_i)^2/p\right]^{\frac{Np}{2}}} \\ &= \frac{\prod_{i=1}^{p} a_{ii}^{N/2}}{(trA/p)^{\frac{Np}{2}}}.\end{aligned}$$

The $LRTS$ for H is

$$\lambda = \lambda_1 \cdot \lambda_2 = \frac{|A|^{N/2}}{(trA/p)^{\frac{Np}{2}}} = \left(\frac{|A|^{\frac{1}{p}}}{trA/p}\right)^{\frac{Np}{2}}.$$

[Note that λ resembles with $\left[\frac{\prod_{i=1}^{p}\phi_i^{1/p}}{\sum_{i=1}^{p}(\frac{\phi_i}{p})}\right]^{\frac{Np}{2}} = \left[\frac{|\Sigma|^{1/p}}{tr\Sigma/p}\right]^{\frac{Np}{2}} = 1$]

If $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_p$ are the characteristic roots of $|S - \hat{\lambda}I| = 0$, where $S = \frac{A}{N-1}$, then the likelihood ratio test criterion is

$$\begin{aligned}\lambda^* &= \left[\frac{\prod_{i=1}^{p}\hat{\lambda}_i^{1/p}}{\sum_{i=1}^{p}(\frac{\hat{\lambda}_i}{p})}\right]^{\frac{Np}{2}} \\ &= \left(\frac{\text{Geometric mean of } \hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_p}{\text{Arithmetic mean of } \hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_p}\right)^{\frac{Np}{2}}.\end{aligned}$$

Now consider

$H : \Psi = \sigma^2\Psi_0$ given $\underset{\sim}{y} \sim N(\underset{\sim}{\gamma}, \Psi)$.

In the transformed variables $\{\underset{\sim}{x}_{\alpha^*}\}$, the criterion is

$$|A^*|^{\frac{N}{2}}\left(\frac{trA^*}{p}\right)^{-\frac{Np}{2}}$$

where

$$\begin{aligned}A^* &= \sum_{\alpha=1}^{N}(\underset{\sim}{x}_\alpha^* - \bar{\underset{\sim}{x}}^*)(\underset{\sim}{x}_\alpha^* - \bar{\underset{\sim}{x}}^*)' \\ &= C\sum_{\alpha=1}^{N}\underset{\sim}{y}_\alpha - \bar{\underset{\sim}{y}})\underset{\sim}{y}_\alpha - \bar{\underset{\sim}{y}})'C' \\ &= CBC'\end{aligned}$$

and $B = \sum_{\alpha=1}^{N}(\underset{\sim}{y}_\alpha - \bar{\underset{\sim}{y}})(\underset{\sim}{y}_\alpha - \bar{\underset{\sim}{y}})'$.

Since

$$\Psi_0 = C^{-1}(C')^{-1} = (C'C)^{-1},$$

thus

$$|A^*| = \frac{|B|}{|\Psi_0|} = |B\Psi_0^{-1}|$$

and

$$trA = trCBC' = trBC'C = trB\Psi_0^{-1}.$$

Finally, we have the following theorem:

Theorem 7: Given $\underset{\sim}{y}_1, \underset{\sim}{y}_2, \ldots, \underset{\sim}{y}_n \sim N_p(\underset{\sim}{\gamma}, \Psi)$, the likelihood ratio test criterion for $H : \Psi = \sigma^2\Psi_0$ where Ψ_0 is specified and σ^2 is not specified is

$$\frac{|B\Psi_0^{-1}|^{\frac{N}{2}}}{\left(\frac{trB\Psi_0^{-1}}{p}\right)^{\frac{Np}{2}}}.$$

It is to be noted that the maximum likelihood estimator of σ^2 under H_0 is

$$\frac{trB\Psi_0^{-1}}{p(N-1)} = \frac{trA}{p(N-1)},$$

then

$$\frac{B\Psi_0^{-1}}{\sigma^2} \sim \chi^2_{p(N-1)}.$$

Testing the Equality of Characteristic Roots

Let $\underset{\sim}{x}_\alpha \sim N_p(\underset{\sim}{\mu}, \Sigma), \alpha = 1, 2, \ldots, N(N > p)$ Suppose that we want to test that the last $(p-k)$ principal components have the same variances. So the null hypothesis is

$$H_0 : \lambda_{k+1} = \lambda_{k+2} = \ldots = \lambda_p.$$

We use the likelihood ratio test. The likelihood function is given by

$$L(\underset{\sim}{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{Np}{2}} |\Sigma|^{N/2}} \exp\left[-\frac{1}{2} \{ tr\Sigma^{-1}S + N(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu})'\Sigma^{-1}(\underset{\sim}{\bar{x}} - \underset{\sim}{\mu}) \} \right].$$

Under Ω,

$$\begin{aligned} \max_{\Omega} L(\underset{\sim}{\mu}, \Sigma) &= \max_{\Sigma} L(\Sigma) \\ &= \max_{\Sigma} \frac{1}{(2\pi)^{\frac{Np}{2}} |\hat{\Sigma}_{\Omega}|^{N/2}} \exp\left(-\frac{1}{2} tr\hat{\Sigma}_{\Omega}^{-1} S \right) \\ &= \frac{1}{(2\pi)^{\frac{Np}{2}} |\frac{S}{N}|^{N/2}} \exp\left(-\frac{Np}{2} \right) \\ &= \frac{1}{(2\pi)^{\frac{Np}{2}} (\prod_{i=1}^{p} \hat{\lambda}_{(i)})^{N/2}} \exp\left(-\frac{Np}{2} \right). \end{aligned}$$

Note that the covariance matrix of the principal components is a diagonal matrix, as the principal components are independent.

Next, to maximize $L(\underset{\sim}{\mu}, \Sigma)$ under $H_0 : \lambda_{k+1} = \lambda_{k+2} = \ldots = \lambda_p = \lambda$, first we find the maximum likelihood estimators of λ which turn out to be the arithmetic mean of the sample characteristic roots $\hat{\lambda}_{k+1}, \hat{\lambda}_{k+2}, \ldots, \hat{\lambda}_p$, as $\hat{\lambda} = \frac{\sum_{i=k+1}^{p} \hat{\lambda}_i}{p-k} \equiv \bar{\lambda}$. Also

$$\begin{aligned} |\hat{\Sigma}_0| &= |diag(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_k, \bar{\lambda}, \bar{\lambda}, \ldots, \bar{\lambda})| \\ &= \hat{\lambda}_1 \cdot \hat{\lambda}_2 \ldots \hat{\lambda}_k (\bar{\lambda})^{p-k} \\ &= \bar{\lambda}^{p-k} \prod_{i=1}^{k} \hat{\lambda}_i. \end{aligned}$$

Thus

$$\max_{H_0} L(\underset{\sim}{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{Np}{2}} [\bar{\lambda}^{p-k} \prod_{i=1}^{k} \hat{\lambda}_i]^{N/2}} \exp\left(-\frac{Np}{2} \right).$$

So the likelihood ratio test statistic λ^* is obtained as

$$\begin{aligned}\lambda^* &= \frac{\max_{H_0} L(\underset{\sim}{\mu}, \Sigma)}{\max_{\Omega} L(\underset{\sim}{\mu}, \Sigma)}\\ &= \left(\frac{\prod_{i=1}^{p} \hat{\lambda}_{(i)}}{\bar{\lambda}^{p-k} \prod_{i=1}^{k} \hat{\lambda}_i}\right)^{\frac{N}{2}}\\ &= \left(\frac{\prod_{i=k+1}^{p} \hat{\lambda}_i}{\left(\frac{\sum_{i=k+1}^{p} \hat{\lambda}_i}{p-k}\right)^{p-k}}\right)^{\frac{N}{2}}.\end{aligned}$$

Now, the test criterion is to reject H_0 if $\lambda^* \leq c$.

Also, λ^* is equivalent to

$$q = (p-k)(N-1)\log\left(\frac{\sum_{i=k+1}^{p} \hat{\lambda}_i}{p-k}\right) - (N-1)\sum_{i=k+1}^{p} \hat{\lambda}_i > c$$

and asymptotically

$$q \overset{asy.}{\sim} \chi^2_{\left(\frac{(k)(k+1)}{2}-1\right)}.$$

Testing the Hypothesis About the Sum of the Smallest Characteristic Roots

Suppose we want to know whether the last $(p-m)$ principal components may be ignored or not, or whether the first m principal components furnish a good approximation to $\underset{\sim}{X}$. This could be done if the sum of the variances of the last principal components is less than some specified amount, say γ.

Consider the hypothesis

$$\begin{aligned}H_0 :& \lambda_{m+1} + \lambda_{m+2} + \ldots + \lambda_p \geq \gamma\\ H_1 :& \lambda_{m+1} + \lambda_{m+2} + \ldots + \lambda_p < \gamma\end{aligned}$$

where γ is specified.

If the characteristic roots of Σ are different, it follows from large sample results that $\sqrt{n}(\sum_{i=m+1}^{p} \hat{\lambda}_i - \sum_{i=m+1}^{p} \lambda_i)$ has a limiting distribution $N(0, 2\sum_{i=m+1}^{p} \lambda_i^2)$.

Now estimate the variance consistently by $2\sum_{i=m+1}^{p}\hat{\lambda}_i^2$. The large sample rejection region at significance level ϵ is

$$\sum_{i=m+1}^{p}\hat{\lambda}_i < \gamma - \frac{\sqrt{2\sum_{i=m+1}^{p}\hat{\lambda}_i^2}}{\sqrt{n}} Z_{2\epsilon},$$

where $Z_{2\epsilon}$ is the upper significance point on normal distribution for ϵ level of significance.

The investigator may alternatively want an upper confidence interval for $\sum_{i=m+1}^{p}\lambda_i$ with at least approximate confidence level $1-\epsilon$. It is given by

$$\sum_{i=m+1}^{p}\lambda_i \leq \sum_{i=m+1}^{p}\hat{\lambda}_i + \frac{\sqrt{2\sum_{i=m+1}^{p}\hat{\lambda}_i^2}}{\sqrt{n}} Z_{2\epsilon}.$$

If the right-hand side is sufficiently small (in particular, less than γ), the investigator has confidence that the sum of the variances of the smallest $(p-m)$ principal components is so small that they can be neglected.

Choice of Number of Principal Components

How to address and decide that the first k principal components account for a certain percentage of the total variance, which is parsimonious yet fairly accurate.

Using the spectral decomposition theorem, the symmetric matrix Σ is expressed as

$$\Sigma = \lambda_1 \underset{\sim}{\alpha}^{(1)}\underset{\sim}{\alpha}^{(1)'} + \lambda_2 \underset{\sim}{\alpha}^{(2)}\underset{\sim}{\alpha}^{(2)'} + \ldots + \lambda_p \underset{\sim}{\alpha}^{(p)}\underset{\sim}{\alpha}^{(p)'},$$

we write as

$$\Sigma_p = \sum_{i=1}^{p}\lambda_i \underset{\sim}{\alpha}^{(i)}\underset{\sim}{\alpha}^{(i)'}.$$

As successive components are extracted from Σ_p, the matrices can be formed by retaining first r components as

$$\Sigma_r = \sum_{i=1}^{r}\lambda_i \underset{\sim}{\alpha}^{(i)}\underset{\sim}{\alpha}^{(i)'}.$$

Now Σ_r is compared with Σ_p to determine how well the matrix Σ is being generated by a small number of variates, i.e., principal components. Unless $rank(\Sigma) < p$, some variance will always remain unexplained if fewer than p principal components are taken to describe the system.

How to decide the first k principal components that account for a certain percentage of the total variance, so as to avoid serious loss of information. So first we need to know or test whether all p principal components should be retained, i.e., the original X's are just as good as Y's (principal component) for the purpose of representations, i.e., – Are X's independent (or uncorrelated) can be tested using

$$-N\left(1-\frac{2p+11}{6N}\right)\log|R| \sim \chi^2_{\frac{p(p-1)}{2}}$$

where R is the correlation matrix.

If

H_0 : X's are independent, and all characteristic roots are equal,

then

$$-\left[N-\frac{2p^2+p+2}{6p(N-1)}\right]\left[\log\frac{|S|}{(\frac{trS}{p})^p}\right] \sim \chi^2_{\frac{p(p+1)}{2}}.$$

Suppose that the first k characteristic roots are large and account for most of the variance. Do the remaining roots differ significantly among themselves? Or are the remaining roots and their associated characteristic vector distinguishable?

This can be formulated as

$H_0 : \lambda_{k+1} = \lambda_{k+2} = \ldots = \lambda_p$

and

$$\begin{aligned} M = & \left(\underset{\sim}{\alpha}^{(k+1)}\cdot\underset{\sim}{\alpha}^{(k+2)}\ldots\underset{\sim}{\alpha}^{(p)}\right)^{-1}\left(\frac{\underset{\sim}{\alpha}^{(k+1)}+\underset{\sim}{\alpha}^{(k+2)}+\ldots+\underset{\sim}{\alpha}^{(p)}}{p-k}\right)^{p-k} \\ = & (\underset{\sim}{\alpha}^{(k+1)}\cdot\underset{\sim}{\alpha}^{(k+2)}\ldots\underset{\sim}{\alpha}^{(p)})^{-1}(\underset{\sim}{\bar{\alpha}})^{p-k} \end{aligned}$$

which is the $(p-k)^{th}$ power of ratio of arithmetic mean and geometric mean of $\underset{\sim}{\alpha}^{(k+1)}, \underset{\sim}{\alpha}^{(k+2)}, \ldots, \underset{\sim}{\alpha}^{(p)}$.

Then

$$-\ln M \sim \chi^2_{\frac{(p-k-1)(p-k+2)}{2}}$$

is a crude approximation and

$$\chi^2 = \left[N - k - 1 - \frac{1}{6}\left\{ 2(p-k) + 1 + \frac{2}{p-k} \right\} + \underset{\sim}{\bar{\alpha}}^2 \sum_{i=1}^{k} \frac{1}{(\alpha^{(j)} - \bar{\alpha})^2} \right] \ln M.$$

The term $\bar{\alpha}^2 \sum_{i=1}^{k} \frac{1}{(\alpha^{(j)} - \bar{\alpha})^2}$ can be dropped if $\underset{\sim}{\alpha}^{(1)}, \underset{\sim}{\alpha}^{(2)}, \ldots, \underset{\sim}{\alpha}^{(k)}$ are large compared to $\bar{\alpha}$.

In practice, we should proceed sequentially.

Test $H_0 : \lambda_1 = \lambda_2 = \ldots = \lambda_p$ that all λ's are equal.

If rejected, then test $H_0 : \lambda_2 = \lambda_3 = \ldots = \lambda_p$ (i.e., last $(p-1)$ λ's) otherwise stop.

If rejected, then test $H_0 : \lambda_3 = \lambda_4 = \ldots = \lambda_p$ and so on.

If $H_0 : \lambda_{k+1} = \lambda_{k+2} = \ldots = \lambda_p$ is accepted then stop.

Thus, $(p-k)$ is the number by which the dimensionality can be reduced.