MH4501 Multivariate Analysis

- Final Revision -

Naoki Honda

May 2019

1 PCA: Principal Component Analysis

Population PCA (Result 6.1)

Assume the population covariance matrix Σ of random vector $X = (X_1, X_2, ..., X_p)^T$ is known, and Σ has eigenvalue-eigenvector pairs $(\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_p, e_p)$, subject to $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Then the kth **principal component** is given by

$$Y_k = e_k^T X = e_{k1}X_1 + e_{k2}X_2 + \cdots + e_{kp}X_p \qquad k = 1, 2, ..., p$$

In addition,

$$\begin{aligned} \text{Var}(Y_k) &= e_k^T \Sigma e_k = \lambda_k \qquad k = 1, 2, ..., p \\ \text{Cov}(Y_j, Y_k) &= e_j^T \Sigma e_k = 0 \qquad j \neq k \quad (j, k = 1, 2, ..., p) \end{aligned}$$

Proof of Result 6.1

Let $U = (e_1, e_2, ..., e_p)$ and $\Lambda = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_p)$. Thus $\Sigma = U\Lambda U^T$ is the eigenvalue decomposition of Σ. Denote $b = U^T a$. Then

$$a^T \Sigma a = \frac{a^T \Sigma a}{a^T a} = \frac{a^T U \Lambda U^T a}{a^T U U^T a} = \frac{b^T \Lambda b}{b^T b} = \frac{\sum_{j=1}^p \lambda_j b_j^2}{\sum_{j=1}^p b_j^2} \leq \frac{\lambda_1 \sum_{j=1}^p b_j^2}{\sum_{j=1}^p b_j^2} = \lambda_1$$

The maximum is attained when $a = e_1$, since it gives $b = U^T e_1 = (1, 0, 0, ..., 0)^T$, $b^T b = 1$ and $b^T \Lambda b = \lambda_1$.
This step shows that the 1st principal component is given by $Y_1 = e_1^T X$.

Secondly, we prove: when the first k $(k = 1, 2, ..., p-1)$ principal component(s) $Y_1 = e_1^T X, Y_2 = e_2^T X, ..., Y_k = e_k^T X$ are decided, to have the $(k+1)$th principal component $Y_{k+1} = a_{k+1}^T X$ orthogonal to every one of $e_1, e_2, ..., e_k$.
For any $j = 1, 2, ..., k$,

$$\text{Cov}(Y_{k+1}, Y_j) = \text{Cov}(a_{k+1}^T X, e_j X) = a_{k+1}^T \Sigma e_j = a_{k+1}^T (\lambda_j e_j) = \lambda_j a_{k+1}^T e_j$$

So to have $\text{Cov}(Y_{k+1}, Y_j)$ is equivalent to have $a_{k+1}^T e_j = 0$.
This step prepares for the next one.[1]

Variance (Result 6.2)

$$\sum_{k=1}^{p} \text{Var}(Y_k) = \sum_{j=1}^{p} \text{Var}(X_j)$$

Proof:

$$\sum_{k=1}^{p} \text{Var}(Y_k) = \sum_{k=1}^{p} \lambda_k = \text{tr}(\Lambda) = \text{tr}(\Lambda U^T U) = \text{tr}(U \Lambda U^T)$$
$$= \text{tr}(\Sigma) = \sum_{j=1}^{p} \sigma_{jj} = \sum_{j=1}^{p} \text{Var}(X_j)$$

In practice, some popular quantities of interest are:

% of variance explained by the kth principal component Y_k: $\frac{\lambda_k}{\sum_{j=1}^{p} \lambda_j}$ $\quad k = 1, 2, ..., p$

Cumulative % of variance explained by the first k principal components: $\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$ $\quad k = 1, 2, ..., p$

Correlation with the original variable (Result 6.3)

$$\text{Cor}(Y_k, X_j) = \frac{e_{kj}\sqrt{\lambda_k}}{\sqrt{\sigma_{jj}}} \qquad k, j = 1, 2, ..., p$$

Proof:
Denote $d_j = (0, ..., 0, 1_{j\text{th}}, 0, ..., 0)^T$ so that $X_j = d_j^T X$. Therefore

$$\text{Cov}(Y_k, X_j) = \text{Cov}(X_j, Y_k)$$
$$= \text{Cov}(d_j^T X, e_k^T X) = d_j^T \Sigma e_k = d_j^T (\lambda_k e_k) = \lambda_k (d_j^T e_k) = \lambda_k e_{kj}$$

In addition, $\text{Var}(Y_k) = \lambda_k$ and $\text{Var}(X_j) = \sigma_{jj}$. Thus

$$\text{Cor}(Y_k, X_j) = \frac{\text{Cov}(Y_k, X_j)}{\sqrt{\text{Var}(Y_k)}\sqrt{\text{Var}(X_j)}} = \frac{e_{kj}\sqrt{\lambda_k}}{\sqrt{\sigma_{jj}}} \qquad k, j = 1, 2, ..., p$$

[1] The last part of the proof is omitted due to the similarity to the first part. Detail in P.9 - Lec #6

1.1 Sample PCA

What we discussed previously still applies in sample PCA. While in addition, since we have a sample $x_1, x_2, ..., x_n$, in PCA, we would as well have a value of every principal component on every observation: y_{ik} ($i = 1, 2, ..., n$ and $k = 1, 2, ..., p$). These values are called PCA **scores**.

1.2 Sample PCA - Standardized Variables

Sample PCA on Standardized Variables (Result 6.6)

Assume we have realizations $x_1, x_2, ..., x_n$ of random vector $X = (X_1, X_2, ..., X_p)^T$, and the sample correlation matrix R has eigenvalue-eigenvector pairs $(\lambda_1^Z, u_1), (\lambda_2^Z, u_2), ..., (\lambda_p^Z, u_p)$, subject to $\lambda_1^Z \geq \lambda_2^Z \geq \cdots \geq \lambda_p^Z \geq 0$. Then the kth **sample principal component** obtained from standardized variables $Z = (Z_1, Z_2, ..., Z_p)^T$ is given by

$$Y_k^Z = u_k^T Z = u_{k1} Z_1 + u_{k2} Z_2 + \cdots + u_{kp} Z_p \qquad k = 1, 2, ..., p$$

and the value of Y_k^Z on the ith observation is given by

$$y_{ik}^Z = u_k^T z_i \qquad i = 1, 2, ..., n \quad k = 1, 2, ..., p$$

Again $\text{Var}(Y_k^Z) = \lambda_k^Z$ ($k = 1, 2, ..., p$) and $\text{Cov}(Y_j^Z, Y_k^Z) = 0$ ($j \neq k$). In addition,

$$\sum_{k=1}^{p} \text{Var}(Y_k^Z) = \sum_{j=1}^{p} \text{Var}(Z_j) = p$$

and

$$\text{Cor}(Y_k^Z, Z_j) = u_{kj} \sqrt{\lambda_k^Z} \qquad k, j = 1, 2, ..., p$$

2 CA: Cluster Analysis

2.1 Distance

Distance (No need to memorize)

Minkowski distance: $d(x,y) = \left[\sum_{j=1}^{p} |x_j - y_j|^m\right]^{\frac{1}{m}}$

When $m = 1$: $d(x,y) = \sum_{j=1}^{p} |x_j - y_j|$ (Manhattan distance)

When $m = 2$: $d(x,y) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}$ (Euclidiean distance)

When $m \to \infty$: $d(x,y) = \max_{j=1,2,\ldots,p} |x_j - y_j|$ (Chebyshev distance)

Mahalanobis distance: $d(x,y) = \sqrt{(x-y)^T S^{-1}(x-y)}$

(variance of Euclidean distance relative to S)

Canberra distance: $d(x,y) = \sum_{j=1}^{p} \frac{|x_j - y_j|}{(x_j + y_j)}$

(weighted version of Manhattan distance)

Czekanowski distance: $d(x,y) = 1 - \frac{2\sum_{j=1}^{p} \min(x_j, y_j)}{\sum_{j=1}^{p}(x_j + y_j)}$

2.2 K-means Clustering

K-means Clustering

Given a collection of observations $x_1, x_2, ..., x_n$ and a well defined measure of distance $d()$, the K-means clustering works in an iterative fashion as below, with K, the desired number of clusters, pre-specified:

1. Initialization: assign the n observations into K clusters arbitrarily.

2. Find the center C_k $(k = 1, 2, ..., K)$ of each cluster (based on current assignments), by taking average (mean) of the observations in each cluster respectively.

3. Assign each observation x_i $(i = 1, 2, ..., n)$ to Cluster $k(i)$, by the criterion that

$$k(i) = \arg\min_k d(C_k, x_i)$$

If any reassignment happens, go back to Step 2; Otherwise, the algorithm is completed.

2.3 Hierarchical Clustering

Motivation

1. One major drawback of K-means clustering is that the desired number of clusters must be pre-specified, which is not always feasible.

2. Hierarchical clustering instead provides a whole "path", from one extreme situation, namely that each individual observation constitutes a cluster, to the other extreme situation, namely that all observations are in one single cluster. Thus, one is able to review the whole path before specifying K.

3. There are different algorithms to implement hierarchical clustering, while in this course we only discuss the **agglomerative hierarchical clustering**, which involves merging small clusters which are similar (in some sence) into larger ones.

4. Therefore, besides a measure of distance between two observations, we also need a measure of distance between two clusters, which is conventionally called a **linkage** function.

Linkage functions (No need to memorize)

Single linkage: $\min_{x\in A, y\in B} d(x,y)$

Complete linkage: $\max_{x\in A, y\in B} d(x,y)$

Average linkage: $\frac{1}{|A||B|}\sum_{x\in A}\sum_{y\in B} d(x,y)$

Centroid linkage: $d(\bar{x}_A, \bar{x}_B)$

where $\bar{x}_A = \frac{1}{|A|}\sum_{x\in A} x$ and $\bar{x}_B = \frac{1}{|B|}\sum_{x\in B} x$

Given a collection of observations $x_1, x_2, ..., x_n$, a well defined measure of distance $d()$, and a pre-specified linkage function to measure distance between clusters, the agglomerative clustering works in an iterative fashion as below:

Hierarchical Clustering (agglomerative)

1. Initialization: every individual observation constitutes a cluster respectively.

2. Calculate the distance (linkage) between every pair of clusters, find the pair that are closest to each other. Merge these two clusters into one cluster.

3. If all observations are in a single cluster, the alrorithm is completed. Otherwise, go back to Step 2.

3 DA: Discriminant Analysis

3.1 Definition Preparation

- **Prior probabilities** measure the chance of an **unspecified** observation (x) belonging to the populations:

$$p_k = Pr[x \in \pi_k] \quad \text{where} \quad k = 1, 2, ..., K$$

  Clearly, it requires that $\sum_{k=1}^{K} p_k = 1$

  - When there is not enough prior knowledge, a naive choice of prior probability is simply $p_1 = p_2 = \cdots = p_K = \frac{1}{K}$
  - Alternatively, prior probabilities can be estimated from sample proportions of classes:

$$p_k = \frac{|\{i : y_i = k\}|}{n} \qquad k = 1, 2, ..., K$$

- **Conditional probability**

$$p(t|k) = Pr[x \in R_t | x \in \pi_k] = \int_{R_t} f_k(x)dx \qquad t, k = 1, 2, ..., K$$

  where $f_k(x)$ is the PDF of population π_k

  - $\sum_{t=1}^{K} p(t|k) = 1 \qquad (k = 1, 2, ..., K)$
  - $Pr[x \in R_t \ \& \ x \in \pi_k] = Pr[x \in \pi_k] \times Pr[x \in R_t | x \in \pi_k] = p_k p(t|k)$ $(t, k = 1, 2, ..., K)$
  - $\sum_{k=1}^{K} \sum_{t=1}^{K} p_k p(t|k) = 1$

- **Misclassification cost**

$$c(t|k) = \text{loss caused by misclassifying an observation into } \pi_t \text{ when it is actually from } \pi_k \quad (t, k = 1, 2, ..., K)$$

  - In this course, unless otherwise stated, $c(1|1) = c(2|2) = \cdots = c(K|K) = 0$
  - In general, $c(t|k) \neq c(k|t) \quad (t \neq k)$.[2]

[2]For example, misclassifying a person with a certain disease as healthy, v.s. misclassifying a healthy person as diseased

ECM: Expected Cost of Missclassification

$$ECM = \sum_{k=1}^{K} \sum_{t=1}^{K} p_k p(t|k) c(t|k)$$

[a]

[a]Note that ECM is a **function of classification rule**. If $p_k, f_k(x), p(t|k)$, and $c(t|k)$ are all known or can be estimated (from the sample $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$), the classification rule that minimizes ECM is called the **minimum ECM classification rule**.

Result 8.1

Explicit solution of the minimum ECM classification rule is given by

$$R_m = \{x : \arg\min_t \sum_{k=1}^{K} p_k f_k(x) c(t|k) = m\} \qquad m = 1, 2, ..., K$$

That is, to classify x into π_m, if $\sum_{k=1}^{K} p_k f_k(x) c(m|k)$ is the minimum among $\sum_{k=1}^{K} p_k f_k(x) c(t|k)$ for all $t = 1, 2, ..., K$

3.2 $K = 2$ multivariate normal populations

Now suppose $k = 2$, and the 2 populations are both multivariate normal:

$$\pi_1 : N_p(\mu_1, \Sigma_1) \qquad \& \qquad \pi_2 : N_p(\mu_2, \Sigma_2)$$

Recall the PDF of multivariate normal distributions:

$$f_k(x) = \frac{1}{(2\pi)^{p/2} \det(\Sigma_k)^{1/2}} \exp\left[-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right]$$

Minimum ECM when K is 2; MVN

The minimum ECM classification rule can be reformulated as:[a]

$$R_1 = \left\{ x : \frac{1}{2}\left(d_2(x) - d_1(x) + \ln\left[\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right] \right) \geq \ln\left[\frac{p_2 c(1|2)}{p_1 c(2|1)}\right] \right\}$$

$$R_2 = \Omega \setminus R_1 \qquad \text{where } \Omega \text{ is the sample space}$$

[a]LHS $= (1/2) \times (d_2(x) - d_1(x) + \ln[\det(\Sigma_2)/\det(\Sigma_1)]) = \ln[f_1(x)/f_2(x)]$

As for quantities used in the rule:

μ_k $(k = 1, 2)$: if not known, estimated by $\bar{x}_k$

Σ_k $(k = 1, 2)$: if not known, estimated by S_k

p_k $(k = 1, 2)$: if not known, estimated by $\frac{|\{i:y_i=k\}|}{n}$

$c(1|2)$ & $c(2|1)$: must be pre-specified

If we further have $\Sigma_1 = \Sigma_2 = \Sigma$,

$$\ln\left[\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\right] = \ln 1 = 0$$
$$d_k(x) = (x-\mu_k)^T\Sigma^{-1}(x-\mu_k) \qquad (k=1,2)$$

Thus

$$\begin{aligned} d_{12}(x) &= \frac{1}{2}[d_2(x) - d_1(x)] \\ &= \frac{1}{2}[(x-\mu_2)^T\Sigma^{-1}(x-\mu_2) - (x-\mu_1)^T\Sigma^{-1}(x-\mu_1)] \\ &= (\mu_1-\mu_2)^T\Sigma^{-1}x - \frac{1}{2}(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1+\mu_2) \\ &= (\mu_1-\mu_2)^T\Sigma^{-1}(x - \frac{1}{2}(\mu_1+\mu_2)) \end{aligned}$$

Minimum ECM when K is 2; MVN with same covariance matrix

The minimum EMC classification rule can be reformulated as

$$R_1 = \left\{ x : d_{12}(x) \geq \ln\left[\frac{p_2 c(1|2)}{p_1 c(2|1)}\right] \right\}$$
$$R_2 = \Omega \setminus R_1$$

In the rule, Σ, if unknown, can be estimated by

$$S_{pool} = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1+n_2-2}$$

3.3 $K \geq 3$ multivariate normal populations

- If $p_1 = p_2 = \cdots = p_K$ and $c(t|k)$ is constant for all pairs $t \neq k$, to find m that maximizes $f_m(x)$, is equivalent with to find m that minimizes

$$-2\ln[f_m(x)] - p\ln[2\pi] = d_m(x) + \ln[\det(\Sigma_m)]$$

  where $d_m(x) = (x-\mu_m)^T\Sigma_m^{-1}(x-\mu_m)$

- If $c(t|k)$ is constant for all pairs $t \neq k$ while $p_1, p_2, ..., p_K$ are general, to find m that maximizes $p_m f_m(x)$, is equivalent with to find m that minimizes

$$-2\ln[p_m f_m(x)] - p\ln[2\pi] = d_m(x) + \ln[\det(\Sigma_m)] - 2\ln p_m$$

4 FA: Factor Analysis

The essential purpose of FA is to describe, if possible, the covariabce relationships among many variables in terms of a fer underlying, but **unobservable**, random quantities (latent variables) called factors.

4.1 The Orthogonal Factor Model

Notations

$$\begin{aligned} X_1 - \mu_1 &= l_{11}F_1 + l_{12}F_2 + \cdots + l_{1m}F_m + \epsilon_1 \\ X_2 - \mu_2 &= l_{21}F_1 + l_{22}F_2 + \cdots + l_{2m}F_m + \epsilon_2 \\ &\cdots \\ X_p - \mu_p &= l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pm}F_m + \epsilon_p \end{aligned}$$

- X_j: jth observable random variable ($j = 1, 2, ..., p$)
- μ_j: mean of X_j ($j = 1, 2, ..., p$)
- F_k: kth **common factor** ($k = 1, 2, ..., m$)
- l_{jk}: **loading** of the jth variable on the kth (common) factor ($j = 1, 2, ..., p$ and $k = 1, 2, ..., m$)
- ϵ_j: jth **specific factor** or **error** ($j = 1, 2, ..., p$)

Assumptions

- $\mathbb{E}[F] = \mathbf{0}_{m\times 1}$ and $\mathrm{Cov}(F) = l_m$

- $\mathbb{E}[\epsilon] = \mathbf{0}_{p\times 1}$ and

$$\mathrm{Cov}(\epsilon) = \Psi_{p\times p} = \begin{pmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{pmatrix}$$

- F and ϵ are independent, thus $\mathrm{Cov}(F, \epsilon) = \mathbf{0}_{m\times p}$

4.2 Covariance Structure; OFM

$$X = \mu + LF + \epsilon$$

implies

$$\begin{aligned} (X-\mu)(X-\mu)^T &= (LF+\epsilon)(LF+\epsilon)^T \\ &= (LF)(LF)^T + \epsilon(LF)^T + (LF)\epsilon^T + \epsilon\epsilon^T \\ &= LFF^TL^T + \epsilon F^TL^T + LF\epsilon^T + \epsilon\epsilon^T \end{aligned}$$

Thus

$$\begin{aligned}\Sigma &= E[(X-\mu)(X-\mu)^T] \\ &= \mathbb{E}[LFF^TL^T + \epsilon F^TL^T + LF\epsilon^T + \epsilon\epsilon^T] \\ &= L\mathbb{E}[FF^T]L^T + \mathbb{E}[\epsilon]\mathbb{E}[F^T]L^T + L\mathbb{E}[F]\mathbb{E}[\epsilon^T] + \mathbb{E}[\epsilon\epsilon^T] \\ &= L\text{Cov}(F)L^T + \mathbf{0}_{p\times p} + \mathbf{0}_{p\times p} + \text{Cov}(\epsilon) \\ &= LL^T + \Psi\end{aligned}$$

which implies

$$\sigma_{jj} = l_{j1}^2 + l_{j2}^2 + \cdots + l_{jm}^2 + \psi_j \qquad j = 1, 2, ..., p$$

σ_{jj}: variance of X_j

$h_j^2 = l_{j1}^2 + l_{j2}^2 + \cdots + l_{jm}^2$: jth **communality**
portion of σ_{jj} explained by the m common factors $\Sigma = LL^T +$

ψ_j: jth **specific variance** or **uniqueness**
protion of σ_{jj} explained by ϵ_j

Ψ also implies

$$\sigma_{ij} = l_{i1}l_{j1} + l_{i2}l_{j2} + \cdots + l_{im}l_{jm}$$

where $i, j = 1, 2, ..., p$ and $i \neq j$

Back to

$$X = \mu + LF + \epsilon$$

it also implies

$$(X-\mu)F^T = (LF+\epsilon)F^T = LFF^T + \epsilon F^T$$

Thus

$$\begin{aligned}\text{Cov}(X, F) &= \mathbb{E}[(X-\mu)F^T] = \mathbb{E}[LFF^T + \epsilon F^T] \\ &= L\mathbb{E}[FF^T] + \mathbb{E}[\epsilon]\mathbb{E}[F^T] = L\text{Cov}(F) + \mathbf{0}_{p\times m} \\ &= L\end{aligned}$$

i.e.

$$\text{Cov}(X_j, F_k) = l_{jk} \qquad \text{where} \qquad j = 1, 2, ..., p \quad \text{and} \quad k = 1, 2, ..., m$$

4.3 Principal Component Method

Principal Component Method

In PCA, we had

$$\begin{aligned}\Sigma &= U\Lambda U^T = U\Lambda^{1/2}\Lambda^{1/2}U^T \\ &= (U\Lambda^{1/2})(U\Lambda^{1/2})^T\end{aligned}$$

In consequence, the solution to FA is given by

$$\begin{aligned}L &= U\Lambda^{1/2} \\ \Psi &= \mathbf{0}_{p\times p}\end{aligned}$$

This solution is feasible in theory but definitely suboptimal, since $m = p$.
A strategy to address this is to use only the largest m ($m < p$ or even $m << p$) eigenvalues $\lambda_1, \lambda_2, ..., \lambda_m$, and their corresponding eigenvectors $e_1, e_2, ..., e_m$.

1. Let $U_m = (e_1, e_2, ..., e_m)_{p\times m}$ and $\Lambda_m = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_m)_{m\times m}$, then

$$\tilde{L}_{p\times m} = U_m\Lambda_m^{1/2} = (\sqrt{\lambda_1}e_1, \sqrt{\lambda_2}e_2, ..., \sqrt{\lambda_m}e_m)$$

2. After $\tilde{L}$ is determined, let

$$\tilde{\Psi} = \text{diag}(\Sigma - \tilde{L}\tilde{L}^T)$$

   which is a diagonal matrix with the same diagonal elements as $\Sigma - \tilde{L}\tilde{L}^T$

3. At this stage, we think, more or less,

$$\Sigma \approx \tilde{L}\tilde{L}^T + \tilde{\Psi}$$

Principal Component Method; Residual Matrix

$$\Sigma - (\tilde{L}\tilde{L}^T + \tilde{\Psi})$$

is thus called the **residual matrix**, which has the following properties:
1. Symmetric matrix.
2. All diagonal elements are zero.

4.4 Covariance Structure; PCM

As for the (estimated) factor loading matrix

$$\tilde{L} = \begin{pmatrix} \tilde{l}_{11} & \tilde{l}_{12} & \cdots & \tilde{l}_{1m} \\ \tilde{l}_{21} & \tilde{l}_{22} & \cdots & \tilde{l}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{l}_{p1} & \tilde{l}_{p2} & \cdots & \tilde{l}_{pm} \end{pmatrix}$$

- By row, we have interpreted it as
1. Portion of σ_{jj} explained by the m common factors: $\tilde{h}_j^2 = \tilde{l}_{j1}^2 + \tilde{l}_{j2}^2 + \cdots + \tilde{l}_{jm}^2$
2. Portion of σ_{jj} explained by ϵ_j: $\tilde{\psi}_j = \sigma_{jj} - \tilde{h}_j^2$

- In addition, by column, we interpreted it as

$$\text{Portion of total variance explained by } F_k = \sum_{j=1}^{p} \tilde{l}_{jk}^2$$

1. In consequence,

$$\text{Portion of total variance explained by } F_k = \frac{\sum_{j=1}^{p} \tilde{l}_{jk}^2}{\sum_{j=1}^{p} \sigma_{jj}}$$

$$\text{Cumulative portion of total variance explained by the first } k \text{ common factors } = \frac{\sum_{i=1}^{k}\sum_{j=1}^{p} \tilde{l}_{ji}^2}{\sum_{j=1}^{p} \sigma_{jj}}$$

2. In principle component method of estimation for FA,

$$\sum_{j=1}^{p} \tilde{l}_{jk}^2 = (\sqrt{\lambda_k} e_k)^T (\sqrt{\lambda_k} e_k) = \lambda_k e_k^T e_k = \lambda_k$$

4.5 The Maximum Likelihood Method

If F and ϵ both can be assumed to be normally distributed, so is $X = \mu + LF + \epsilon$.
If we have a collection of observations $x_1, x_2, ..., x_n$, the likelihood function would be

$$L(\mu, \Sigma | \{x_i\}_{i=1}^n) = \frac{1}{(2\pi)^{np/2} \det(\Sigma)^{n/2}} e^{-\sum_{i=1}^n (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)/2}$$

Substitute $\Sigma = LL^T + \Psi$ in we have

$$L(\mu, L, \Psi | \{x_i\}_{i=1}^n) = \frac{1}{(2\pi)^{np/2} \det(LL^T + \Psi)^{n/2}} e^{-\sum_{i=1}^n (x_i - \mu)^T (LL^T + \Psi)^{-1} (x_i - \mu)/2}$$

Maximum likelihood solution, $\hat{L}$ and $\hat{\Psi}$, can be obtained by maximizing the likelihood function, with assistance of computers.

4.6 Test of Sufficiency; MLM (No need to memorize)

Test of Sufficiency

Whenever a FA model is build using sample covariance matrix S, a common question of interest is whether the model constructed fits the observed data well enough. That is, we want to test the (null) hypothesis

$$H_0 : \Sigma = \hat{L}\hat{L}^T + \hat{\Psi}$$

If the maximum likelihood method of estimation is employed, a testing statistic is proposed:

$$TS = \left(n - 1 - \frac{2p + 4m + 5}{6}\right) \ln \frac{\det(\hat{L}\hat{L}^T + \hat{\Psi})}{\det(\hat{\Sigma})}$$

Remarks (No need to memorize):

1. $\hat{L}$ and $\hat{\Psi}$ are the maximum likelihood estimate of L and Ψ respectively
2. $\hat{\Sigma}$ is the maximum likelihood estimate of Σ, that is,

$$\hat{\Sigma} = \frac{1}{n} SSCP = \frac{n-1}{n} S$$

3. If n is large, under H_0, TS follows a χ^2 distribution with degree of freedom

$$d.f. = \frac{(p-m)^2}{2} - \frac{p+m}{2}$$

To have a positive $d.f.$, it must be the case that

$$m < p - \frac{\sqrt{8p+1} - 1}{2}$$

4.7 Factor Rotation

In a FA model

$$\Sigma = LL^T + \Psi$$

it is noticed that, for any $m \times m$ orthogonal matrix G,

$$(LG)(LG)^T = LGG^TL^T = LI_mL^T = LL^T$$

That is,

$$L^* = LG$$
$$\Psi^* = \Psi$$

is also a feasible solution $\Sigma = L^*L^{*T} + \Psi^*$.
In linear algebra, to right-multiply G to L can be interpreted as to **rotate** L.

Since any orthogonal matrix will keep L^* a feasible solution, an idea is proposed to keep rotating L until a simple structure is achieved. Here by "simple", it means every factor loading l_{jk} is either large or close to 0. The so-called **varimax criterion** is thus proposed:

Varimax Criterion (No need to memorize)

$$V = \sum_{k=1}^{m} \left[\sum_{j=1}^{p} l_{jk}^4 - \frac{1}{p} \left(\sum_{j=1}^{p} l_{jk}^2 \right)^2 \right]$$

Although it looks a little bit forbidding, the criterion in fact has a quite direct interpretation, since

$$\sum_{j=1}^{p} l_{jk}^4 - \frac{1}{p} \left(\sum_{j=1}^{p} l_{jk}^2 \right)^2 = \text{variance of } \{l_{1k}^2, l_{2k}^2, ..., l_{pk}^2\}$$

Therefore, to maximize the varimax criterion, is effectively "spreading out" l_{jk}^2 as much as possible, so that some of them are large, while the others are close to 0.

A weighted version of varimax criterion is defined to be

$$V_W = \sum_{k=1}^{m} \left[\sum_{j=1}^{p} \left(\frac{l_{jk}^2}{h_j^2} \right) - \frac{1}{p} \left(\sum_{j=1}^{p} \frac{l_{jk}^2}{h_j^2} \right)^2 \right]$$

What are different before and after factor rotation:

1. The factor loadings: l_{jk} $(j = 1, 2, ..., p$ and $k = 1, 2, ..., m)$
2. The variance explained by each factor: $\sum_{j=1}^{p} l_{jk}^2$ $(k = 1, 2, ..., m)$

While what are the same:

1. The communalities: $h_j^{*2} = \sum_{k=1}^{m} l_{jk}^{*2} = (L^* L^{*T})_{jj} = (LL^T)_{jj} = h_j^2$ $(j = 1, 2, ..., p)$
2. The specific variances: $\psi_j^* = 1 - h_j^{*2} = 1 - h_j^2 = \psi_j$ $(j = 1, 2, ..., p)$
3. The cumulative variance explained by **all** the m common factors:

$$\sum_{k=1}^{m}\sum_{j=1}^{p} l_{jk}^{*2} = \sum_{j=1}^{p}\sum_{k=1}^{m} l_{jk}^{*2} = \sum_{j=1}^{p} h_j^{*2} = \sum_{j=1}^{p} h_j^2 = \sum_{k=1}^{m}\sum_{j=1}^{p} l_{jk}^2$$

4.8 Factor Scores

When sample FA is performed, similar to the component scores in sample PCA, every factor is supposed to have a value on every observation.
However, due to different approaches of PCA and FA, the factor scores are not able to be directly calculated. Here we mention two methods of estimation:

Methods of Factor Scoring

1. **The Bartlett method** (the weighted least squares method):

$$\hat{f}_i = (\hat{L}^T \hat{\Psi}^{-1} \hat{L})^{-1} \hat{L}^T \hat{\Psi}^{-1} (x_i - \bar{x})$$

2. **The regression method**:

$$\hat{f}_i = \hat{L}^T S^{-1} (x_i - \bar{x})$$

where $\hat{f}_i = (\hat{f}_{i1}, \hat{f}_{i2}, ..., \hat{f}_{im})^T$ are estimated values of $F_1, F_2, ..., F_m$ on x_i.

5 CCA: Canonical Correlation Analysis (For the final, only expected to conduct CCA as we did in Example 10.1)

CCA seeks to identify and quantify the associations between two sets of variables.

5.1 Population CCA

With the partitioned r.v., let

$$U = a^T X^{(1)}$$
$$V = b^T X^{(2)}$$

where $a_{p\times 1}$ and $b_{q\times 1}$ are a pair of coefficient vectors. In consequence,

$$\text{Var}(U) = a^T \Sigma_{11} a$$
$$\text{Var}(V) = b^T \Sigma_{22} b$$
$$\text{Cov}(U, V) = a^T \Sigma_{12} b = b^T \Sigma_{21} a$$

and

$$\text{Cor}(U, V) = \frac{a^T \Sigma_{12} b}{\sqrt{a^T \Sigma_{11} a}\sqrt{b^T \Sigma_{22} b}}$$

Result 10.1

- $e_1, e_2, ..., e_p$ (each with unit length) to be the eigenvectors of $p \times p$ matrix $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$, with corresponding eigenvalues $\rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_p^2$
- $f_1, f_2, ..., f_p$ (each with unit length) to be the **first p** $(p \leq q)$ eigenvectors of $q \times q$ matrix $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$, with corresponding eigenvalues $\rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_p^2$

Then the kth pair of canonical variables $(k = 1, 2, ..., p)$ is explicitly given by

$$U_k = e_k^T \Sigma_{11}^{-1/2} X^{(1)}, \quad \text{that is} \quad a_k = \Sigma_{11}^{-1/2} e_k$$
$$V_k = f_k^T \Sigma_{22}^{-1/2} X^{(2)}, \quad \text{that is} \quad b_k = \Sigma_{22}^{-1/2} f_k$$

For $k = 1, 2, ..., p$

$$f_k = \frac{1}{\rho_k}\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2} e_k$$
$$e_k = \frac{1}{\rho_k}\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2} f_k$$

For $k, l = 1, 2, ..., p$ and $k \neq l$

$$\begin{aligned} \text{Var}(U_k) &= 1, & \text{Cor}(U_k, U_l) &= 0 \\ \text{Var}(V_k) &= 1, & \text{Cor}(V_k, V_l) &= 0 \\ \text{Cor}(U_k, V_k) &= \rho_k, & \text{Cor}(U_k, V_l) &= 0 \end{aligned}$$

5.2 Variation explained by canonical variables

Now denote

$$A = \begin{pmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_p^T \end{pmatrix} \qquad \text{and} \qquad U = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_p \end{pmatrix}$$

Then $U = AX^{(1)}$. Therefore,

$$\text{Cov}(U, X^{(1)}) = \text{Cov}(AX^{(1)}, X^{(1)}) = A \cdot \text{Cov}(X^{(1)}) = A\Sigma_{11}$$

At the meantime, $X^{(1)} = A^{-1}U$, thus

$$\text{Cov}(U, X^{(1)}) = \text{Cov}(U, A^{-1}U) = \text{Cov}(U)(A^{-1})^T = (A^{-1})^T$$

That is,

$$A\Sigma_{11} = (A^{-1})^T, \qquad \text{or equivalently,} \qquad A\Sigma_{11}A^T = I_p$$

Variation explained by canonical variables

We have [a]

$$\text{Total variation in } X^{(1)} = \sum_{j=1}^{p} \sigma_{jj}$$

$$\text{Portion of variation in } X^{(1)} \text{ explained by } U_k = \sum_{j=1}^{p} \text{Cov}(U_k, X_j^{(1)})$$

where $k = 1, 2, ..., p$.
$\text{Cov}(U_k, X_j^{(1)})$ is given by the kj-th element of matrix $A\Sigma_{11} = (A^{-1})^T$.

[a] The similar also applies to $X^{(2)}, V, B$ and Σ_{22}.
But notice that $V_{p+1}, V_{p+2}, ..., V_q$ are not canonical variables.

5.3 Sample CCA

Can be obtain by replacing $\Sigma \to S$.

In addition, every sample canonical variable has an observed value on each of the observations:

$$U_{ik} = a_k^T x_i^{(1)} = e_k^T S_{11}^{-1/2} x_i^{(1)}$$
$$V_{ik} = b_k^T x_i^{(2)} = f_k^T S_{22}^{-1/2} x_i^{(2)}$$

where $i = 1, 2, ..., n$ and $k = 1, 2, ..., p$.

Test for Σ_{12} (Not required)

$$H_0 : \Sigma_{12} = \mathbf{0}_{p \times q}$$

If we can assume that X follows the multivariate normal distribution $N_{p+q}(\mu, \Sigma)$, and the sample size n is large, a testing statistic is proposed

$$TS = -\left(n - 1 - \frac{1}{2}(p + q + 1)\right) \ln \prod_{k=1}^{p} (1 - \rho_k^2)$$
$$\sim \chi^2(pq)$$

where ρ_k $(k = 1, 2, ..., p)$ are the canonical correlations obtained from sample CCA.

5.4 Sample CCA - Standardized variables

Sample CCA performed on original variables and sample CCA performed on standardized variables are relates in a simple manner:

Result 10.3

- The canonical correlations are the same up to a change of sign:

$$\rho_k^2 = (\rho_k^Z)^2 \qquad (k = 1, 2, ..., p)$$

- If $\bar{x}_j^{(1)} = 0$ (for all $j = 1, 2, ..., p$) and $\bar{x}_j^{(2)} = 0$ (for all $j = 1, 2, ..., q$)[a], then the canonical variables are the same:

$$U_k = U_k^Z \qquad \text{and} \qquad V_k = V_k^Z \qquad (k = 1, 2, ..., p)$$

- Denote

$$\text{diag}(S) = D = \begin{pmatrix} D_1 & \mathbf{0} \\ \mathbf{0} & D_2 \end{pmatrix}$$

Then

$$a_k^Z = D_1^{1/2} a_k \qquad \text{and} \qquad b_k^Z = D_2^{1/2} b_k \qquad (k = 1, 2, ..., p)$$

[a] Or equivalently "both set of variables have mean 0".

Example 10.1

Observations are collected on $p = 2$ population variables and $q = 3$ economic variables for each of $n = 50$ countries:

$$\begin{aligned} X_1^{(1)} &= \text{Percentage of population over 75 years old} \\ X_2^{(1)} &= \text{Percentage of population under 15 years old} \\ X_1^{(2)} &= \text{Aggregate personal saving} \\ X_2^{(2)} &= \text{Per capita disposable income} \\ X_3^{(2)} &= \text{Percentage growth rate of disposable income} \end{aligned}$$

The sample correlation is found to be

$$R = \left(\begin{array}{cc|ccc} 1 & -0.91 & -0.46 & -0.76 & -0.05 \\ -0.91 & 1 & 0.32 & 0.79 & 0.03 \\ \hline -0.46 & 0.32 & 1 & 0.22 & 0.30 \\ -0.76 & 0.79 & 0.22 & 1 & -0.13 \\ -0.05 & 0.03 & 0.30 & -0.13 & 1 \end{array}\right) = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}$$

Example 10.1 (Cont.)

Eigenvalues and eigenvectors of $R_{11}^{-1/2}R_{12}R_{22}^{-1}R_{21}R_{11}^{-1/2}$ are

$$\begin{pmatrix} \rho_1^2 \\ \rho_2^2 \end{pmatrix} = \begin{pmatrix} 0.6877 \\ 0.1374 \end{pmatrix}, \qquad e_1^Z = \begin{pmatrix} 0.7314 \\ -0.6819 \end{pmatrix}, \qquad e_2^Z = \begin{pmatrix} 0.6819 \\ 0.7314 \end{pmatrix}$$

Thus

$$U_1 = (e_1^Z)^T R_{11}^{-1/2} Z^{(1)} = 0.5939 Z_1^{(1)} - 0.4288 Z_2^{(1)}$$
$$U_2 = (e_2^Z)^T R_{11}^{-1/2} Z^{(1)} = 2.3377 Z_1^{(1)} + 2.3735 Z_2^{(1)}$$

Eigenvalues and eigenvectors of $R_{22}^{-1/2}R_{21}R_{11}^{-1}R_{12}R_{22}^{-1/2}$ are

$$\begin{pmatrix} \rho_1^2 \\ \rho_2^2 \\ \rho_3^2 \end{pmatrix} = \begin{pmatrix} 0.6877 \\ 0.1374 \\ 0 \end{pmatrix},$$
$$f_1^Z = \begin{pmatrix} 0.3844 \\ 0.9211 \\ 0.0611 \end{pmatrix}, \qquad f_2^Z = \begin{pmatrix} 0.9299 \\ -0.3817 \\ -0.0513 \end{pmatrix}, \qquad f_3^Z = \begin{pmatrix} 0.0239 \\ -0.0761 \\ 0.9968 \end{pmatrix}$$

Thus

$$V_1 = (f_1^Z)^T R_{22}^{-1/2} Z^{(2)} = 0.2694 Z_1^{(2)} + 0.9049 Z_2^{(2)} + 0.0882 Z_3^{(2)}$$
$$V_2 = (f_2^Z)^T R_{22}^{-1/2} Z^{(2)} = 1.0463 Z_1^{(2)} - 0.5295 Z_2^{(2)} - 0.2595 Z_3^{(2)}$$

Conclusion

1. $U_1 = 0.5939 Z_1^{(1)} - 0.4288 Z_2^{(1)}$, a greater value of U_1 corresponds to a more aging population.
2. $V_1 = 0.2694 Z_1^{(2)} + 0.9049 Z_2^{(2)} + 0.0882 Z_3^{(2)}$, a greater valur of V_1 mostly corresponds to greater per capita disposable income.
3. $\rho_1^Z = -0.8293$ is indicating a strong negative correlation between U_1 and V_1

6 Supplementary Topics

6.1 Matrix Partitioning (might be used, but not solely tested about this)

No content to review, it's too basic.

6.2 Simultaneous CI Revisited (Required)

Generalization of SCIs

Let θ denote any one of the parameters for which the SCIs are to be constructed. Then the common structure of SCIs could be expressed as:

$$\hat{\theta} \pm t_{\alpha^*/2}[\text{d.f.}]\sqrt{\hat{\text{Var}}(\hat{\theta})}$$

- $\hat{\theta}$ is the point estimate of θ, and $\hat{\text{Var}}(\hat{\theta})$ is the estimate of $\text{Var}(\hat{\theta})$
- $\alpha^* = \alpha/(\text{number of SCIs})$, where α is the esired significance level
- d.f. is the corresponding degree of freedom, which depends on the specific situation, and is usually highly related to the sample size(s) (when the sample size is large enough, $t_{\alpha^*/2}[\text{d.f.}]$ is replaced by $z_{\alpha^*/2}$)

Examples

1. For one-sample inference, $\theta = \mu_j$ $(j = 1, 2, ..., p)$:

$$\bar{x}_j \pm t_{\alpha^*/2}[n-1]\sqrt{\frac{s_{jj}}{n}}$$

2. For paired two-sample comparison, $\theta = \mu_{d,j}$ $(j = 1, 2, ..., p)$:

$$\bar{d}_j \pm t_{\alpha^*/2}[n-1]\sqrt{\frac{s_{d,jj}}{n}}$$

3. For unpaired two-sample comparison with assumption of equal population covariance matrix, $\theta = \delta_j = \mu_{1,j} - \mu_{2,j}$ $(j = 1, 2, ..., p)$:

$$(\bar{x}_1 - \bar{x}_2)_j \pm t_{\alpha^*/2}[n_1 + n_2 - 2]\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s_{pool,jj}}$$

4. For unpaired two-sample comparison with assumption of large sample size, $\theta = \delta_j = \mu_{1,j} - \mu_{2,j}$ $(j = 1, 2, ..., p)$:

$$(\bar{x}_1 - \bar{x}_2)_j \pm z_{\alpha^*/2}\sqrt{\left(\frac{s_{1,jj}}{n_1} + \frac{s_{2,jj}}{n_2}\right)}$$

SCIs for treatment effects in MANOVA

In MANOVA, if the null hypothesis of equal treatment effect $H_0 : \tau_1 = \tau_2 = \cdots = \tau_G$ is rejected, it is natural to ask about how $\tau_1, \tau_2, ..., \tau_G$ are different in details. In fact, we are able to construct SCIs for

$$\delta_j^{(gh)} = (\tau_g - \tau_h)_j$$

where $g, h = 1, 2, ..., G$ with $g > h$ (no need to consider $\tau_g - \tau_h$ and $\tau_h - \tau_g$ simultaneously) and $j = 1, 2, ..., p$

- $\hat{\delta}_j^{(gh)} = (\hat{x}_g - \hat{x}_h)_j$
- $\text{Var}(\hat{\delta}_j^{(gh)}) = \text{Var}((\hat{x}_g - \hat{x}_h)_j) = \left(\frac{1}{n_g} + \frac{1}{n_h}\right)\sigma_{jj}$, thus
  $\hat{\text{Var}}(\hat{\delta}_j^{(gh)}) = \left(\frac{1}{n_g} + \frac{1}{n_h}\right) s_{pool,jj} = \left(\frac{1}{n_g} + \frac{1}{n_h}\right) \frac{w_{jj}}{n-G}$
  where w_{jj} is the jth diagonal element of treatment SSCP matrix W,
  and $n = \sum_{g=1}^{G} n_g$
- Number of SCIs $= pG(G-1)/2$, thus $\alpha^* = \frac{2\alpha}{pG(G-1)}$
- d.f. $= \sum_{g=1}^{G}(n_g - 1) = n - G$

To sum up, the SCIs are

$$(\bar{x}_g - \bar{x}_h)_j \pm t_{\alpha^*/2}[n-G]\sqrt{\left(\frac{1}{n_g} + \frac{1}{n_h}\right)\frac{w_{jj}}{n-G}}$$

6.3 Fisher's Discrimination Method (two population)

Consider a classification problem with two populations. Suppose $x_{11}, x_{12}, ..., x_{1n}$ are sampled from π_1, and $x_{21}, x_{22}, ..., x_{2n}$ are sampled from π_2.
The key idea of Fisher's DA is to find a proper linear combination of X, $Z = a^T X$, so that a classification rule can be easily constructed by employing a cut-off to the linear combination. That is,

$$R_1 = \{x : a^T x \geq c\}$$
$$R_2 = \{x : a^T x < c\}$$

Therefore, Fisher's DA mostly focused on the determination of vector a and cut-off c
To determine c, a reasonable and convenient solution is that

$$c = a^T \left(\frac{\bar{x}_1 + \bar{x}_2}{2} \right)$$

because $(\bar{x}_1 + \bar{x}_2)/2$ is seen to be in the middle between the two sets of observations, $\{x_{11}, x_{12}, ..., x_{1n_1}\}$ and $\{x_{21}, x_{22}, ..., x_{2n_2}\}$. This is equivalent to

$$c = \frac{\bar{z}_1 + \bar{z}_2}{2}$$

since $\bar{z}_1 = a^T \bar{x}_1$ and $\bar{z}_2 = a^T \bar{x}_2$

Now it remains to solve a. The criterion Fisher proposed is to maximize the separation between the two sets of transformed observations:

$$z_{11} = a^T x_{11}, \quad z_{12} = a^T x_{12}, \quad \cdots \quad z_{1n_1} = a^T x_{1n_1}$$

and

$$z_{21} = a^T x_{21}, \quad z_{22} = a^T x_{22}, \quad \cdots \quad z_{2n_2} = a^T x_{2n_2}$$

When it is assumed that π_1 and π_2 share a same population covariance matrix, the separation is quantified by:

$$\frac{(\bar{z}_1 - \bar{z}_2)^2}{s_{pool,z}}$$

- The above criterion is interpreted as squared distance between sample means of z relative to (pooled) sample variance of z
- $s_{pool,z} = \frac{(n_1-1)s_{z,1}+(n_2-1)s_{z,2}}{n_1+n_2-2}$, where

  $$s_{z,1} = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (z_{1i} - \bar{z}_1)^2 \quad \text{and} \quad s_{z,2} = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (z_{2i} - \bar{z}_2)^2$$

  It can be easily seen that $s_{pool,z} = a^T S_{pool,z} a$

The solution of a that maximizes the separation can be proved to be:

$$a = S_{pool,x}^{-1}(\bar{x}_1 - \bar{x}_2)$$

Fisher's DA method

In consequence,

$$R_1 = \left\{ x : (\bar{x}_1 - \bar{x}_2)^T S_{pool,x}^{-1} x \geq (\bar{x}_1 - \bar{x}_2)^T S_{pool,x}^{-1} \left(\frac{\bar{x}_1 + \bar{x}_2}{2} \right) \right\}$$
$$R_2 = \Omega \setminus R_1$$

This coincides with the minimum ECM classification rule with assumptions on multivariate normality, as well as equal prior probabilities ($p_1 = p_2$) and equal misclassification costs ($c(1|2) = c(2|1)$).