Multivariate Statistical Inference

Yiqiao YIN
Statistics Department
Columbia University
Notes in LaTeX

April 19, 2018

Abstract

This document presents notes from STAT 5223 - Multivariate Statistical Inference. This course is concerned with statistical analysis of data sets containing multiple observations on each subject. We begin with a review of basic results from matrix algebra, and develop some probability theory for random vectors and matrices. The multi-variate normal, Wishart, and Hotelling's T^2 distributions provide the foundation for inference about mean vectors and covariance matrices. We then move on to exploratory and inferential multivariate methods including: principal component analysis, canonical correlation, factor analysis, discrimination and classification, clustering, and multidimensional scaling. The emphasis throughout will be on the application of statistical methods to real data.

This note is dedicated to Professor Ronald Neath.

Contents

1 MATRIX

1.1 Matrix Operation

Data matrix, as follows

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & & & \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

with x_{ij} the value for variable j on subject i. Here we denote rows of X to be units (subjects) and columns of X to be variables. We can also call this matrix X

$$X = (\tilde{x}_{(1)}, ..., \tilde{x}_{(p)})$$

where $\tilde{x}_{(j)}$ to be $n \times 1$ vector of realized values for variable j while $j = 1, ..., p$. Or as collection of row vectors:

$$\begin{pmatrix} \tilde{x}_1^T \\ \tilde{x}_2^T \\ \vdots \\ \tilde{x}_n^T \end{pmatrix}$$

where $\tilde{x}_i$ is the $p \times 1$ vector of values for unit i and $i = 1, 2, ..., n$.

1.2 Sample Mean Vector, Correlation, and Covariance

Sample mean for variable j is

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = (\frac{1}{n}, ..., \frac{1}{n}) \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \vdots \\ \tilde{x}_n \end{pmatrix} = \frac{1}{n}\mathbf{1}_n^T \tilde{x}_{(j)}$$

Sample mean vector is

$$\bar{X} = \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \vdots \\ \tilde{x}_n \end{pmatrix} = \frac{1}{n} \begin{pmatrix} \tilde{x}_1 \tilde{1}_n \\ \tilde{x}_2 \tilde{1}_n \\ \vdots \\ \tilde{x}_n \tilde{1}_n \end{pmatrix} = \frac{1}{n} \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \vdots \\ \tilde{x}_n \end{pmatrix} \tilde{\mathbf{1}}_n$$

Thus, we have

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n} X^T \tilde{1}_n = \bar{X}$$

Correlation matrix is

$$S = (s_{jk})_{p \times p} = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & & & \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{pmatrix}$$

where $s_{jj} = s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ which is sample variance for variable j. This gives us

$$S_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

which is the sample covariance between variables j and k. Next, we have

$$S = \frac{1}{n-1}\sum_{j=1}^{n}(\tilde{x}_i - \bar{x})(x_i - \bar{x})^T = \frac{1}{n-1}[\sum_{i=1}^{n} x_i x_i^T - n\bar{x}\bar{x}^T]$$

and this leads us to

$$\bar{x} = \frac{1}{n}X^T\tilde{1}_n$$

which implies

$$S = \frac{1}{n-1}[X^TX - \frac{1}{n}X^T\tilde{1}_n\tilde{1}_n^TX]$$

while $\tilde{1}\tilde{1}^T$ is the matrix with one-entry in every cell, and one can simplify S to the following

$$S = \frac{1}{n-1}X^T(I_n - \frac{1}{n}\tilde{1}_n\tilde{1}_n^T)X$$

We can also write the following way

$$S = \frac{1}{n-1}X^THX$$

and $H = I_n - \frac{1}{n}\tilde{1}_n\tilde{1}_n^T$. Note that H is the $n \times n$ centering matrix.

Note that from the center mean,

$$HX = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \dots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \dots & x_{2p} - \bar{x}_p \\ \vdots & & & \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \dots & x_{np} - \bar{x}_p \end{bmatrix}$$

we can center all data by its column. This gives us

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(\tilde{x}_i - \bar{x})(\tilde{x}_i - \bar{x})^T = \frac{1}{n-1}(HX)^T(HX) = \frac{1}{n-1}X^TH^THX$$

since H is symmetric ad idempotent. Note we have $X^TH^THX = X^THX$ is analogous to $\sum_{i=1}^{n}(\tilde{x}_i - \bar{x})(\tilde{x}_i - \bar{x})^T = \sum_{i=1}^{n}\tilde{x}_i(\tilde{x}_i - \bar{x})^T$. Thus, along with definition of H, we alsowrite

$$S = \frac{1}{n-1}X^THX$$

Proposition 1.2.1. *S is positive semi-definite.*

Proof. Note $a^TSa = \frac{1}{n-1}a^TX^TH^THXa$ can be simplify to $a^TSa = \frac{1}{n-1}(HXa)^T(HXa)$ which is the square of length. Thus, we have

$$a^TSa = \frac{1}{n-1}||HXa||^2 \geq 0$$

□

Correlation matrix is defined as

$$R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1p} \\ \vdots & & & \\ r_{p1} & r_{p2} & \dots & r_{pp} \end{pmatrix}$$

while $r_{jk} = \frac{s_{jk}}{s_j s_k}$ and $R = (r_{jk})$ with $r_{jk} = \frac{1}{s_j} s_{jk} \frac{1}{s_k}$ which becomes

$$\begin{pmatrix} 1/s_1 & \dots & 0 \\ 0 & 1/s_2 & \\ \vdots & & \\ 0 & & 1/s_p \end{pmatrix} \begin{pmatrix} s_{11} & \dots \\ \dots & s_{pp} \end{pmatrix} \begin{pmatrix} 1/s_1 & \dots \\ \dots & 1/s_p \end{pmatrix}$$

All the information needed to compute R is in S. $R = D^{-1/2} S D^{-1/2}$ where $D = \text{diag}(s_1^2, s_2^2, ..., s_p^2) = \text{diag}(S)$.

We have X sized $n \times p$ raw data. We also have HX sized $n \times p$ centered data (meaning averaged at 0) while

$$H = I_n \frac{1}{n} \tilde{1}_n \tilde{1}_n^T = \begin{pmatrix} 1 - 1/n & \dots \\ \dots & 1 - 1/n \end{pmatrix}$$

Conclude that we have $S = \frac{1}{n-1}(HX)^T HX = \frac{1}{n-1} X^T H X$ while $D = \text{diag}(S)$. $HXD^{-1/2}$ gives centered and standardized data matrix. Then its i, j entry is $(x_{ij} - \bar{x}_j)/s_j$. Hence, we have

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (\tilde{x}_i - \bar{x})(x\tilde{x}_i - \bar{x})^T = \frac{1}{n-1}(HX)^T (HX)$$

and

$$R = \frac{1}{R-1}(HXD^{-1/2})^T (HXD^{-1/2})$$

The correlation matrix is the cvariance for the (centered and) standardized data!

Suppose $y_i = a_1 x_{i1} + a_2 x_{i2} + \cdots + a_p x_{ip} = \tilde{a}^T \tilde{x}_i$ while $i = 1, 2, ..., n$. The sample mean and variance for the data set$\{y_1, y_2, ..., y_n\}$ are

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \sum_{i=1}^{n} \tilde{a}^T \tilde{x}_i = \tilde{a}^T (\frac{1}{n} \sum_{i=1}^{n} \tilde{x}_i) = \tilde{a}^T \tilde{x}$$

and variance is

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \frac{1}{n-1} (\tilde{a}^T x_i - \tilde{a}^T \bar{x})^2 = \frac{1}{n-1} \sum_{i=1}^{n} [\tilde{a}^T (\tilde{x}_i - \bar{x})(\tilde{x}_i - \bar{x})^T \tilde{a}]$$

which gives us

$$s_j^2 = \tilde{a}^T S \tilde{a}$$

since $(\tilde{a}^T \tilde{x}_i - \tilde{a}^T \bar{x})^2 = (\tilde{a}^T \tilde{x}_i - \tilde{a}^T \bar{x})(\tilde{x}_i^T \tilde{a} - \bar{x}^T \tilde{a}) = \tilde{a}^T (\tilde{x}_i - \bar{x})(\tilde{x}_i - \bar{x})^T \tilde{a}$.

Let us say we have a data set $y_{i1} = a_{11} x_{i1} + a_{12} x_{i2} + \cdots + a_{1p} x_{ip} = \tilde{a}_1^T \tilde{x}_i$ and $y_{i2} = a_{21} x_{i1} + \cdots + a_{2p} x_{ip} = \tilde{a}_2^T \tilde{x}_i$ and so on to y_{ip}. Let us put this in a vector. That is, let us build an $n \times q$ data matrix Y.

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iq} \end{pmatrix} = \begin{pmatrix} \tilde{a}_1^T \\ \vdots \\ \tilde{a}_p^T \end{pmatrix} \tilde{x}_i = A \tilde{x}_i$$

and $\bar{y} = A\tilde{x}$ with $s_y = AS_xA^T$? Let us check. Note that $\bar{y} = \frac{1}{n}\sum_{i=1}^n \tilde{y}_i = A\frac{1}{n}\sum_{i=1}^n \tilde{x}_i = A\bar{x}$ and $s_y = \frac{1}{n-1}\sum_{i=1}^n (y_i - \tilde{y})^T(y_i - \bar{y})^T = A\frac{1}{n-1}\sum_{i=1}^n (\tilde{x}_i - \bar{x})(\tilde{x}_i - \bar{x})^T A^T$.

The data matrix Y of $n \times q$ is

$$Y = \begin{pmatrix} y_1^T \\ \vdots \\ y_n^T \end{pmatrix} = \begin{pmatrix} (A\tilde{x}_1)^T \\ \vdots \\ (A\tilde{x}_n)^T \end{pmatrix} = \begin{pmatrix} \tilde{x}_1^T A^T \\ \vdots \\ \tilde{x}_n^T A^T \end{pmatrix}$$

If $y_i = A\tilde{x}_i$ with $i = 1, ..., n$ where A is $q \times p$ then

$$Y = \begin{pmatrix} y_i^T \\ \vdots \\ y_n^T \end{pmatrix} = XA^T$$

1.3 Standardization

If $X = (x_{ij})$ with $n \times p$ is the data matrix, then $HX = (x_{ij} - \bar{x}_j)$ gives the matrix of centered data. Note $H = I_n - \frac{1}{n}\tilde{1}_n\tilde{1}_n^T$ is called the centering matrix. H is symmetric and also idempotent. It is $n \times n$ matrix and its rank is $n-1$. If x is $n \times p$ while $(p < n)$ has rank $p$that is no variable redundant, then HX has rank p.

Recall $S = (s_{ij})$ is sized $p \times p$, the covariance matrix of X and s_{jk} gives the matrix of centered data. Let $D = \text{diag}(s_{11}, ..., s_{pp}) = \text{diag}(s_1^2, ..., s_p^2)$, then $HXD^{-1/2} = \frac{x_{ij} - \bar{x}_j}{s_j}$ is the centered and standardized data set and it's sized $n \times p$. Also recall covariance matrix S is found by

$$S = \frac{1}{n-1}\sum(\tilde{x}_i - \bar{x})(\tilde{x}_i - \bar{x}) = \frac{1}{n-1}X^T(I_n - \frac{1}{n}\tilde{1}\tilde{1}^T)X$$

that is, $S = \frac{1}{n-1}X^THX$. Then correlation matrix is

$$R = (r_{jk}) = \frac{s_{jk}}{\sqrt{s_{jj} - s_{kk}}} = D^{-1/2}SD^{-1/2}$$

Now we can consider the following questions. What are $\frac{1}{n-1}(HX)^TH(HX)$, i.e. covariance matrix of centered data? Next, what is $\frac{1}{n-1}(HXD^{-1/2})H(HXD^{-1/2})$, i.e. covariance matrix for centered and standardized? S is just the original data matrix. The process of centering and standardizing does nothing to the data matrix. Notice that $H^THH = H$. and standardization described above, $x \to H \times D^{-1/2}$. That is a univariate standardization. It is univariate because H standardizes each variable separately. A multivariate standardization is accomplished by so called the Mahalanobis transformation, related to ideal M-distance.

Let us review this idea. Let $X \in \mathbb{R}^P$. The norm or length of $\tilde{X}$ is

$$||\tilde{X}|| = \sqrt{x_1^2 + \cdots + x_p^2} = \sqrt{\tilde{X}^T X}$$

which is the length of the vector in a geometrical sense. The Euclidean distance between two vectors $\tilde{x}$ and $\tilde{y}$ is

$$d_E(\tilde{X}, \tilde{Y}) = \sqrt{(x_i - y_i)^2 + \cdots + (x_p - y_p)^2} = \sqrt{(\tilde{X} - \tilde{Y})^T(\tilde{X} - \tilde{Y})}$$

Thus, $||\tilde{X}|| = d_E(\tilde{O}, \tilde{X})$ and $d_E(\tilde{x}, \tilde{y}) = ||\tilde{x} - \tilde{y}||$. Now imagine $\tilde{x}$ and $\tilde{y}$ are two observations from a population with mean vector $\tilde{u}$ and covariance matrix Σ. A statistical distance between $\tilde{x}$ and $\tilde{y}$ should take into account not just the $|x_j - y_j|$. but also

(1) the variance of the p variables, and (2) the correlation. The Mahalanobis distance between $\tilde{x}$ and $\tilde{y}$ is

$$d_M^2(\tilde{x},\tilde{y}) = (\tilde{x}-\tilde{y})^T\Sigma^{-1}(\tilde{x}-\tilde{y}) = d_E^2(\Sigma^{-1/2}\tilde{x},\Sigma^{-1/2}\tilde{y}) = d_E^2(\Sigma^{-1/2}(\tilde{x}-\tilde{u})\Sigma^{-1/2}(\tilde{y}-\tilde{u}))$$

The practice Σ is replaced by sample covariance matrix S. The Mahalanobis transform of data matrix X is

$$\tilde{z}_i = S^{-1/2}(\tilde{x}_i - \bar{x})$$

with $i = 1, ..., n$. The M-transform of X with size $n \times p$ is $\tilde{z}_i = S^{-1/2}(\tilde{x}_i - \tilde{\bar{x}})$ with $i = 1, ..., n$ gives us

$$Z = (X - \tilde{1}_n\tilde{\bar{x}})S^{-1/2}$$

Accomplishes 3 things: (1) centers the data, (2) scales the data, and (3) removes correlation. An animation program would be code here.

Remark 1.3.1. Recall the spectral decomposition of S (symmetric positive definite) $S = \Gamma\Lambda\Gamma^T$ while λ is the diagonal matrix of evolves columns of Γ are normalized vector. Note that $\triangle = \text{diag}(x_1, ..., x_p)$, $\triangle^\alpha = \text{diag}(\lambda^\alpha, ..., \lambda_p^\alpha$, then $S^\alpha = \Gamma\Lambda^\alpha\Gamma^T$.

Note the first two things would not even be visible in successive plots. Only labels on axes would change. Third thing changes elliptical point cloud to a circular one.

1.4 Linear Models in Matrix Form

As an example, consider multiple linear regression

$$y_i = B_0 + B_1x_{i1} + \cdots + B_pX_{ip} + \epsilon_i$$

for $i = 1, ..., n$. This can be written as

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \ddots & \\ x_{n1} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} B_0 \\ \vdots \\ B_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

in this case we have $\tilde{y} = X\tilde{B} + \tilde{\epsilon}$ while X is constants $\tilde{B}$ when $\mathbb{E}\tilde{\epsilon} = 0$ and $\text{cov}(\tilde{\epsilon}) = \sigma^2 I$. Statistical model with $p + 2$ parameters, which are $B_0, B_1, ..., B_p, \sigma^2$. Least squares estimate (also ML if $\tilde{\epsilon} \sim N_n$) Then

$$\tilde{B} = \arg\min_B[(\tilde{y} - XB)^T(\tilde{y} - XB)] = (X^TX)^{-1}X^T\tilde{y}$$

An unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n(p+1)}(\tilde{y} - X\hat{B})^T(\tilde{y} - X\hat{B})$$

If $\tilde{\epsilon} \sim N_n$ then $\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2}$, and also $\hat{\sigma^2}$ is independent of $\hat{B}$ so

$$\frac{\hat{B}_j - B_j}{[\text{cov}(\hat{B})_{j+1,j+1}]^{1/2}} \sim \mathcal{N}(0,1)$$

and

$$\frac{\hat{B}_j - B_j}{[\sigma^2(\hat{B})_{j+1,j+1}]^{1/2}} \sim \mathcal{N}(0,1)$$

as well as

$$\frac{\hat{B}_j - B_j}{[\sigma^2(\hat{B})_{j+1,j+1}]^{1/2}} \sim t_{n-p-1}$$

2 MULTIVARIATE DISTRIBUTIONS

2.1 Joint Distribution

Given $\tilde{X} = (X_1, ..., X_p)^T$ which is a random vector. The joint distribution function (or joint cdf) of $\tilde{X}$ is F. If X is a discrete random variable, its probability mass function is concentrated on a finite or countable set of points $\{c_j : j \in J\}$. Then for any set $D \in \mathbb{R}$. Then $Pr(X \in D) = \sum Pr(X = S)$. If X is a continuous random variable, then its cdf $F(x) = Pr(x \in X)$.That is,

$$F(x) = \int_{-\infty}^{x} f(u)du = Pr(X \leq x)$$

Indeed for any $D \subset \mathbb{R}$ have $Pr(X \in D) = \int_D f(x)dx$. For a continuous random vector $\tilde{X} = (X_1, ..., X_p)^T$, the joint density (joint pdf) is $f(x) = \frac{\partial^p}{\partial x_1 ... \partial x_p} F(\tilde{x})$. Moreover, we have $F(x) = \int \cdots \int f(x_1, ..., x_p)dx_p \dots dx_1$, which is the joint/marginal/conditional pdf and cdfs.

Consider $f_x(x) = \int f(x_1, x_2)dx_2$. Then $F_{x_1}(x_1) = F(x_1, +\infty)$. One has $f_{x_2/x_1} = f(x_1, x_2)/f_{x_1}(x_1)$ and this gives us

$$F_{x_1/x_2}(x_2/x_1) = \int f_{x_2/x_1}(x_2/x_1)du_2$$

Example 2.1.1. Given $(X_1, X_2) \sim f(x_1, x_2) = x_1 + x_2$ if $0 < x_1, x_2 < 1$ and 0 else where. Then $f_{x_1}(x_1) = \int f(x_1, x_2)dx_2 = \int (x_1 + x_2)dx_2 = x_1 = 1/2$ while $0 < x_1 < 1$.

Definition 2.1.2. $\tilde{X}_1$ and $\tilde{X}_2$ are independent if, for any sets D_1 and D_2, we have $Pr(\tilde{X}_1 \in D_1, \tilde{X}_2 \in D_2) = Pr(\tilde{X}_1 \in D_1)Pr(\tilde{X}_2 \in D_2)$.

Proposition 2.1.3. *The following statements are equivalent:*
(1) $\tilde{X}_1$ and $\tilde{X}_2$ are independent
(2) $Pr(\tilde{X}_1 \in D_1), \tilde{X}_2 \in D_2) = Pr(\tilde{X}_1 \in D_1)Pr(\tilde{X}_2 \in D_2)$
(3) $f(\tilde{X}_1, \tilde{X}_2) = f_{\tilde{X}_1}(x_1)f_{\tilde{X}_2}(x_2)$ for any $\tilde{X}_1, \tilde{X}_2$.
(4) $F(\tilde{X}_1, \tilde{X}_2) = F_{\tilde{X}_1}(x_2|x_1) = f_{\tilde{X}_2}(x_2)$ for all $\tilde{X}_2$.

2.2 Multivariate Moment

The notion of solving moment is to compute $\mathbb{E}(x) = \int xf(x)dx = \mu$ and $\text{var}(x) = \int (x - \mu)^2 f(x)dx = \mathbb{E}[(x - \mu)^2] = \sigma^2$. The random vector $\tilde{X} = (X_1, ..., X_p)^T$ has expected value $\mathbb{E}(\tilde{X}) = (\mathbb{E}X_1, ..., \mathbb{E}X_p)^T = (\mu_1, ..., \mu_p)^T = \tilde{\mu}$. For covariance, we have $\text{var}(\tilde{X}) = \mathbb{E}[(\tilde{X} - \tilde{\mu})(\tilde{X} - \tilde{\mu})^T] = \Sigma$. Note that $\Sigma = \sigma_{ij}$ as a $p \times p$ matrix.

Next, $\sigma_i = \sigma_i^2 = \text{var}(x_i)$. More generally, given random p-vector $\tilde{X} = (X_1, ..., X_p)^T$ and random q-vector $\tilde{Y} = (Y_1, ..., Y_p)^T$ with expected values $\tilde{\mu}$ and $\tilde{\nu}$, respectively.

Definition 2.2.1. Define $\text{cov}(\tilde{X}, \tilde{Y}) = \mathbb{E}[(\tilde{X} - \tilde{\mu})(\tilde{Y} - \tilde{\nu})^T]$ which is a $p \times p$ matrix.

Let $\tilde{X} = (x_1, ..., x_p)^T$ be a p-vector. Let A be a $q \times p$ matrix and $\tilde{b}$ be a $q \times 1$ vector. The expected value and covariance matrix of $\tilde{Y} = A\tilde{X} + \tilde{b}$ are $\mathbb{E}\tilde{Y} = A\mathbb{E}\tilde{X} + \tilde{b}$ and $\text{var}\tilde{Y} = A\text{var}\tilde{X}A^T$.

Let $\tilde{X} = (x_1, ..., x_p)^T$ and $\tilde{Y} = (y_1, ..., y_p)^T$. Then the random vector $(\tilde{X}, \tilde{Y})$ has expected values and variance.

2.3 Conditional Expectation

Consider the following example.

Example 2.3.1. Consider the discrete random vector $\tilde{X} = (X_1, X_2)$. We are also given

x_1/x_2	3	6	8	
1	.05	.10	.05	.2
2	.20	.15	.05	.4
3	.05	.10	.25	.4

We can compute $\mathbb{E}(X_2) = 3(.3) + 6(.35) + 8(.35) = 5.8$, and also $\text{var}(X_2) = (3 - 5.8)^2.3 + (6-5.8)^2(.35) + (8-5.8)^2(.35) = 4.06$. Now suppose we want to find conditional pmf of $X_2|X_1 = x_1$ for $x_1 = 1, 2, 3$. Now we have

$p(x_1\|x_1)$	3	6	8	$\mathbb{E}(X_2\|X_1)$	$\text{var}(X_2\|X_1 = x_1)$	$\mathbb{E}(X_2\|X_1 = x_1)$	$\text{var}(X_2\|X_1 = x_1)$	$P_{X_1}(x_1)$
1	.25	.5	.25	5.75	18.75	5.75	3.18	.2
2	.5	.375	.125	4.75	3.43	4.75	3.44	.4
3	.125	.25	.625	6.875	2.85	6.8	2.85	.4

For any x_1 satisfies $p_{X_1}(x_1) > 0$ can compute $\mathbb{E}(X_2|X_1 = x_1)$ and $\text{var}(X_2|X_1 = x_1)$. X_1 is a random variable $m(\circ)$ and $v(\circ)$ are functions defined on the range of X_1. This $m(X_1)$ and $v(X_1)$ are random variables. Note that $\mathbb{E}(X_2|X_1)$ and $\text{var}(X_2|X_1)$ are random variables. They have expected values and variances also.

Proposition 2.3.2. *We have the following*

$$\mathbb{E}(X_2) = \mathbb{E}[\mathbb{E}(X_2|X_1)]$$

Proof. For any function $h(\cdot)$, we have

$$\begin{aligned}
\mathbb{E}[h(X_2)] &= \iint h(x_2) f(x_1, x_2) dx_2 dx_1 \\
&= \iint h(x_2) f(x_2|x_1) f(x_1) dx_2 dx_1 \\
&= \int\int\int h(x_2) f(x_2|x_1) dx_2 dx_1 \\
&= \int \mathbb{E}[h(X_2)|X_1 = x_1] f(x_1) dx_1 \\
&= \mathbb{E}(\mathbb{E}h(X_2)|X_1])
\end{aligned}$$

and then we have

$$\begin{aligned}
\text{var}(X_2) &= \iint (x_2 - \mu_2)^2 f(x_1, x_2) dx_2 dx_1 \\
&= \iint (x_2 - \mu_2)^2 f(x_2|x_1) dx_2 f(x_1) dx_1 \\
&= \iint [x_2 - m(x_1) + m(x_1) - \mu_2]^2 f(x_2|x_1) dx_2 f(x_1) dx_1 \\
&= \iint [x_2 - m(x)]^2 f(x_2|x_1) dx_2 + f(x_1) dx_2 + \iint [m(x_1) - \mu_2]^2 f(x_2|x_1) dx_2 f(x_1) dx_1 \\
&\quad + \text{cross-product terms that equal 0} \\
&= \mathbb{E}[\text{var}(X_2|X_1)] + \text{var}[\mathbb{E}(X_2|X_1)]
\end{aligned}$$

□

2.4 Transformation

Suppose we have random variable x and let $Y = g(X)$. The goal is to find the distribution of Y. Then

$$F_Y(y) = \Pr(Y \leq y) = \Pr(g(X) \leq y) = \begin{cases} \Pr(X \in g^{-1}(y)) & \text{if } g \text{ is monotonic increasing} \\ \Pr(X \in g^{-1}(y)) & \text{if } g \text{ is monotonic decreasing} \\ ? & \text{else} \end{cases}$$

Assume X is continuous with pdf f_X and g is monotone, then it has inverse function

$$F_Y(y) = \begin{cases} F_X(g^{-1}(y)) & \text{if } g \text{ is monotone increasing} \\ 1 - F_X(g^{-1}(y)) & \text{if } g \text{ is monotone decreasing} \end{cases}$$

so $f_Y(y) = f_x(g^{-1}(y))|\frac{d}{dy}g^{-1}(y)|$ and let us call it $\star$.

For a random vector $\tilde{X} = (X_1, ..., X_p)^T$, it has joint pdf $f_{\tilde{X}}(x_1, ..., x_p)$. Let $\tilde{Y} = g(\tilde{X})$ where $g : \mathbb{R}^p \to \mathbb{R}^p$ is one-to-one. How to generalize $\star$ to $p > 1$?

Let $u : \mathbb{R}^p \to \mathbb{R}^p$, and let $u(y_1, ..., y_p)^T = (u_1(y_1, ..., y_p), ..., u_p(y_1, ..., y_p))^T$. $p = 1$, $f_Y(y) = f_X(u(dy)))|\frac{du}{dy}|$; for $p \geq 1$, then $f_Y(y) = f_X(u(y))|\det(J)|$ where J denotes Jacobian matrix. That is,

$$J = \begin{pmatrix} \partial u_1/\partial y_1 & \dots & \partial u_1/\partial y_p \\ \vdots & & \\ \partial u_p/\partial y_1 & \dots & \partial u_p/\partial y_p \end{pmatrix}$$

which means $\tilde{X} = u(\tilde{Y}) = A^{-1}(\tilde{Y} - \tilde{b})$ and this gives us $J = (\partial u_i/\partial y_j) = A^{-1}$.

Example 2.4.1. $(x_1, x_2) \sim f(x_1, x_2) = x_1 + x_2$ while $0 < x_1, x_2 < 1$ and 0 elsewhere. Find the joint density of $(Y_1, Y_2) = (X_1^2, X - 1X_2)$. Then solve

$$y_1 = x_1 \text{ and } y_2 = x_1 x_2 \Rightarrow x_1 = \sqrt{y_1}, \text{ and } x_2 = y_2/\sqrt{y_1}$$

and we have Jacobian

$$J = \begin{pmatrix} \partial x_1/\partial y_1 & \partial x_1/\partial y_2 \\ \partial x_2/\partial y_1 & \partial x_2/\partial y_2 \end{pmatrix} = \begin{pmatrix} 1/2(y^{-1/2} & 0 \\ \star & y_1^{-1/2} \end{pmatrix}$$

and thus $\det(J) = 1/(2y_1)$. Range of (X_1, X_2) is unit square. That is, $0 < x_1 < 1 \Rightarrow 0 < \sqrt{y_1} < 1 \Rightarrow 0 < y_1 < 1$, and moreover $0 < x_2 < 1 \Rightarrow 0 < y_2/\sqrt{y_1} \Rightarrow 0 < y_2 < \sqrt{y_1} < 1$.

Example 2.4.2. Suppose X has pdf $f_X(x)$. What is the pdf of $Y = 3X$? Or if $X = (X_1, X_2, X_3)^T$, what is the pdf of

$$Y = \begin{pmatrix} 3X_1 \\ X_1 - 4X_2 \\ X_3 \end{pmatrix}$$

This is a special case of asking for the pdf of Y when

$$X = \mu(Y)$$

for a one-to-one transformation $u : \mathbb{R}^n \to \mathbb{R}^p$. Define the Jacobian of u is

$$\mathcal{J} = \left(\frac{\partial x_i}{\partial y_i}\right) = \left(\frac{\partial u_i(y)}{\partial y_j}\right)$$

and let abs($|\mathcal{J}|$) be the absolute value of the determinant of this Jacobian. The pdf of Y is given by

$$f_Y(y) = \text{abs}(|\mathcal{J}|) \cdot f_X(u(y)).$$

Using this to answer the the introductory question, namely

$$(x_1, ..., x_n)^T = u(y_1, ..., y_p) = \frac{1}{3}(y_1, ..., y_p)^T$$

with

$$\mathcal{J} = \begin{pmatrix} \frac{1}{3} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{3} \end{pmatrix}$$

and hence abs($|\mathcal{J}|$) $= (1/3)^p$. Thus, the pdf of Y is $\frac{1}{3^p} f_X(\frac{y}{3})$.

The introductory example above is a special case of

$$Y = \mathcal{A}X + b, \text{ where } \mathcal{A} \text{ is non singular}$$

The inverse transformation is

$$X = A^{-1}(Y - b)$$

Therefore, $\mathcal{J} = \mathcal{A}^{-1}$, and hence

$$f_Y(y) = \text{abs}(|\mathcal{J}|) f_X\{\mathcal{A}^{-1}(y - b)\}$$

This theorem forms a strong relationship and let us look at the following example

Example 2.4.3. Consider $X = (X_1, X_2) \in \mathbb{R}^2$ with density $f_X(x) = (\mathcal{A}^{-1}y)$,

$$\mathcal{A} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, b = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Then

$$Y = \mathcal{A}X + b = \begin{pmatrix} X_1 + X_2 \\ X - 1 - X_2 \end{pmatrix}$$

and

$$|\mathcal{A}| = -2, \text{abs}(|\mathcal{A}^{-1}|) = \frac{1}{2}, \mathcal{A}^{-1} = \frac{1}{2}\begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix}$$

Hence,

$$\begin{aligned} f_Y(y) &= \text{abs}(|\mathcal{J}|) \cdot f_X(\mathcal{A}^{-1}y) \\ &= \tfrac{1}{2} f_X\{\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\} \\ &= \tfrac{1}{2} f_X\{\tfrac{1}{2}(y_1 + y_2), \tfrac{1}{2}(y_1 - y_2)\} \end{aligned}$$

2.5 Multi-normal Distribution

Let us discuss multi-normal distribution.

Suppose there are two random variables. $X_2|X_1 = x_1 \sim \mathcal{N}(\alpha + \beta x_1, \sigma^2)$. Consider $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$. Find the marginal distribution of X_2. Find the joint distribution of (X_1, X_2).

Let us do mean and variance first (not necessary, simply practice). We have $\mathbb{E}(X_2) = \mathbb{E}[\mathbb{E}(X_2|X_1)] = \mathbb{E}(\alpha + \beta X_1)$ and this gives us $\alpha + \beta\mathbb{E}(X_1) = \alpha + \beta\mu_1$. Then variance $\text{var}(X_2) = \mathbb{E}[\text{var}(X_2|X_1)] + \text{var}[\mathbb{E}(X_2|X_1)] = \mathbb{E}(\sigma^2 + \text{var}(\alpha + \beta X_1) = \sigma^2 + \text{var}(\alpha + \beta X_1) = \sigma^2\beta^2\text{var}(X_1) = \sigma^2 + \beta^2\sigma_1^2$. In fact, $X_2 \sim \mathcal{N}(\alpha + \beta\mu_1, \sigma^2 + \beta^2\sigma_1^2)$.

To prove such result, we use moment generating function. Recall that $X \sim \mathcal{N}(\mu, \sigma^2)$ if and only if $\mathbb{E}(e^{tX_2}) = \mathbb{E}(\mathbb{E}(e^{tX_2}|X_1)]$. Then, by using conditional normality, we have $\mathbb{E}[\exp(t(\alpha + \beta X_1) + \frac{1}{2}t^2\sigma^2)] = \exp(t\alpha + \frac{1}{2}t^2\sigma^2)\mathbb{E}[e^{t\beta X_1}] = \exp(t\alpha + (1/2)t^2\sigma^2)\exp(t\beta\mu_1 + (1/2)t^2\beta^2\sigma_1) = \exp(t(\alpha + \beta\mu_1) + (1/2)t^2(\sigma^2 + \beta^2\sigma_1^2))$.

Proposition 2.5.1. *If $X_2|X| = X_1 \sim \mathcal{N}(\alpha + \beta X_1, \sigma^2)$, $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, then $X_2 \sim \mathcal{N}(\alpha + \beta\mu_1, \sigma^2 + \beta^2\sigma_1^2)$.*

Question: what about joint distribution of $(X_1, X_2)^T$?

Proposition 2.5.2. *If $(X_1, X_2)^T$ has expected values and covariance matrix $\tilde{\mu} = (\mu_1, \mu_2)^T$ and $\Sigma = (\sigma_1^2, \sigma_{12}; \sigma_{12}, \sigma_2^2)$.*

(a) the function $h(x_1)$ that minimizes the MSE $\mathbb{E}[[X_2 - h(X_1)]^2]$, i.e. the best function of X_1 for predicting X_2 (best in sense of minimize MSE) is

$$h(X_1) = \mathbb{E}(X_2|X_1 = x_1)$$

(b) the linear function $h(X_1) = a + bX_1$ which minimizes

$$\mathbb{E}[(X_2 - a - bX_1)^2]$$

is $b = \frac{\sigma_{12}}{\sigma_1^2}$ and $a = \mu_2 - b\mu_1$.

Lemma 2.5.3. *For any random variable X, we have expectation $\mathbb{E}[(X-a)^2]$. This is minimized by $a = \mathbb{E}(X)$.*

Proof. Consider $\frac{\partial}{\partial a}\mathbb{E}[(X-a)^2] = -2\mathbb{E}(X-a) \overset{\text{(set)}}{=} 0$. Done. □

Proof. This is a proof of (b) (directly above).

Consider $\mathbb{E}[(X_2 - a - bX_1)^2] = \mathbb{E}[[(X_2 - bX_1) - a]^2]$ is minimized at $a = \mathbb{E}(X_2 - bX_1) = \mu_2 - b\mu_1$. Thus $\mathbb{E}[(X_2 - \hat{a}(b) - bX_1)^2] = \mathbb{E}[(X_2 - \mu_2) - b(X_1 - \mu_1)]^2 = \mathbb{E}[(X_2 - \mu_2) + b(X_1 - \mu_1)]^2$.

Next, we have $\frac{\partial}{\partial b}() = -2\mathbb{E}[(X_2 - \mu_2) - b(X_1 - \mu_1)](X_1 - \mu) = -2\big(\mathbb{E}[(X_2 - \mu_2)(X_1 - \mu_1)] - b\mathbb{E}[(X_1 - \mu_1)^2]\big) = -2[\text{cov}(X_1, X_2) - b\text{var}(X_1)] \overset{\text{(set)}}{=} 0$ which gives us

$$b = \frac{\text{cov}(X_1, X_2)}{\text{var}(X_1)} = \frac{\sigma_{12}}{\sigma_1^2}$$

(a) Consider $\mathbb{E}[[X_2 - h(X_1)]^2] = \iint [X_2 - h(x_1)]^2 f(X_1, X_2) dx_2 dx_1 = \int \mathbb{E}[[X_2 - h(X_1)]^2 | X_1 = x_1]$. To minimize $\mathbb{E}[[X_2 - h(X_1)]^2 | X_1 = x_1]$ for each x, take $h(X_1) = \mathbb{E}[X_2 | X_1 = x_1]$.

This is just the lemma approached to the conditional distribution of $(X_2|X_1 = x_1)$. To minimize the integral of non-negative integrand, we minimize the integrand at every point. Done.

The best linear predictor of X_2 by $a + bX_1$ is $h(X_1) = \mu_2 + \frac{\sigma_{12}}{\sigma_1^2}(X_1 - \mu)$ is

$$h(X_1) = \mathbb{E}[X_2 | X_1 = x_1]$$

□

Proposition 2.5.4. *The predictor error $X_2 - \mathbb{E}(X_2|X_1)$ is uncorrelated with the predictor variable X_1.*

Proof. The proof is in Homework 2. □

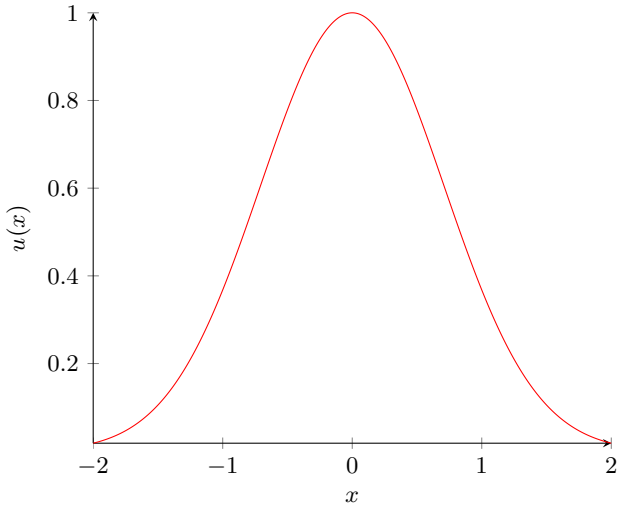
2.6 Normal Distribution

Let us review univariate normal distribution. For a random variable, we say $X \sim \mathcal{N}(0, \sigma^2)$ to be a normal distribution with

$$\mathbb{E}(e^{tX}) = \exp(t\mu + \frac{1}{2}t^2\sigma^2)$$

If $Y = aX + b$, then $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^{1/2})$. The density of $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$f(x) = \frac{1}{\sigma\sqrt{2\sigma}} \exp\{-\frac{1}{2}(X-\mu)^2\}$$



2.7 Bivariate Normal Distribution

Consider the following

$$\tilde{X} = (X_1, X_2)^T \sim \mathcal{B}_2\left[\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{pmatrix}\right]$$

$$\sim \text{BVN}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho) \sim \mathcal{N}_2\left[\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_{22}\end{pmatrix}\right]$$

Proposition 2.7.1. *It states that* $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ *and* $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ *and* $X_2|X_1 = x_1 \sim \mathcal{N}(\mu_x, \sigma_\star^2)$ *while* $\sigma_\star^2 = \sigma_t^2(1-\rho^2)$.

2.8 Multivariate Normal Distribution

Consider $\tilde{X} \sim \mathcal{N}_p(\tilde{\mu}, \Sigma)$ and $\tilde{\mu} = \mathbb{E}(\tilde{X}) = (\mu_1, ..., \mu_p)^T$ while $\mu_j = \mathbb{E}X_j$.

Moreover, $\Sigma = \text{var}(\tilde{X}) = (\sigma_{jk})$ and $\sigma_{jk} = \text{cov}(X_j, X_k)$. Then $Y = a_1X_1 + \cdots + a_pX_p + b = \tilde{a}^T\tilde{X} + b$ is also normal.

Let us discuss mean and variance. $\mathbb{E}(\tilde{a}^T\tilde{X} + b) = \mathbb{E}(\sum a_jX_j + b) = \sum a_j\mathbb{E}(X_j) + b$. $\text{var}(\tilde{a}^T\tilde{X}+b) = \text{var}(\tilde{a}^T\tilde{X}) = \text{var}(\sum a_jX_j) = \text{cov}(\sum a_jX_j, \sum a_kX_k) = \sum\sum a_ja_k\text{cov}X_j, X_k) = \sum\sum a_ja_k\sigma_{jk} = \tilde{a}^T\Sigma\tilde{a}$. Thus,

$$Y = \tilde{a}^T\tilde{X} + b \sim \mathcal{N}(\tilde{a}^T\tilde{X} + b, \tilde{a}^T\Sigma\tilde{a})$$

Now consider $\tilde{Y} = A\tilde{X} + \tilde{b}$ while A is $q \times p$ and $\tilde{b}$ is $q \times 1$. Given

$$\tilde{X} \sim \mathcal{N}_p(\tilde{\mu}, \Sigma)$$

if and only if

$$\mathbb{E}(e^{\tilde{t}^T \tilde{X}}) = \exp\{t^T\tilde{\mu} + \frac{1}{2}\tilde{t}^T\Sigma\tilde{t}\}$$

Let $t \in \mathbb{R}^q$, then $\mathbb{E}[e^{\tilde{t}^T(A\tilde{X}+b)}] = \exp\{\tilde{t}^T(A\tilde{\mu}+\tilde{b}) + \frac{1}{2}\tilde{t}^T A\Sigma A^T\}$. This is an exercise. This gives us

$$\tilde{Y} = A\tilde{X} + \tilde{b} \sim \mathcal{N}_q(A\tilde{\mu} + \tilde{b}, A\Sigma A^T)$$

Recall spectral decomposition of symmetric matrix

$$\Sigma = \Gamma\Lambda\Gamma^T$$

where $\Lambda = \text{diag}(\lambda_1, ..., \lambda_p)$ and Γ is orthogonal, meaning $\Gamma^T\Gamma = \Gamma\Gamma^T = \mathbb{I}$. Note that λ_j are the eigenvalues of Σ. If Σ is a covariance matrix, then $\lambda_j \geq 0$ and $\Sigma^{1/2} = \Gamma\Lambda^{1/2}\Gamma^T$ is square root matrix. The following conditions are equivalent for covariance matrix.

More specifically, we can have $Y = \tilde{a}^T\tilde{X} + b \sim \mathcal{N}_1(\tilde{a}^T\tilde{\mu} + b, \tilde{a}^T\Sigma\tilde{a}$. Let A be $q \times p$ with rank a. Then

$$\tilde{Y} = A\tilde{X} + \tilde{b} \sim \mathcal{N}_q(A\tilde{\mu} + \tilde{b}, A\Sigma A^T)$$

Proof. This proof uses uniqueness of moment generating function, which is different than the approach of the text [1]. □

Consider spectral decomposition a covariance matrix (i.e. symmetric, positive semi-definite). Then we write

$$\Sigma = \Gamma\Lambda\Gamma^T$$

where $\Lambda = \text{diag}(\lambda_1, ..., \lambda_p)$. The λ_j are the eigenvalues of Σ. We have $\Gamma = (\tilde{\gamma}_1, ..., \tilde{\gamma}_p)$ while $\tilde{\gamma}_j$ is associated eigenvector. Γ is an orthogonal matrix. $\Sigma^{-1/2} = \Gamma\Lambda^{1/2}\Gamma^T$ while $\Lambda^{1/2} = \text{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_p})$.

Proposition 2.8.1. *The following are equivalent for correlation matrix:*

(1) Σ has an inverse;

(2) All eigenvalues λ_j are positive;

(3) Σ is positive definite.

In this case, $\Sigma^{-1} = \Gamma\Lambda^{-1}\Gamma^T$ while $\Lambda = diag(1/\lambda_1, ..., 1/\lambda_p)$ and $\Sigma^{1/2} = \Gamma\Lambda^{-1/2}\Gamma^T = \Lambda^t = diag(1/\sqrt{\lambda_1}, ..., 1/\sqrt{\lambda_p})$.

Then

$$\tilde{z} = \Sigma^{-1/2}(\tilde{X} - \tilde{\mu}) \sim \mathcal{N}_p(\tilde{0}, I_p)$$

by Mahalanobis transformation.

Note that $\tilde{z} \sim \mathcal{N}_p(\tilde{0}, I_p) \Leftrightarrow z_1, ..., z_p \sim \text{iid}\mathcal{N}(-, 1)$. Then that means

$$\sum_{i=1}^{p} z_i^2 \sim \chi_p^2 \Leftrightarrow \tilde{z}^T\tilde{z} \sim \chi_p^2$$

The Mahalanobis transformation also provides the recipe for generating any multivariate normal distribution.

Let $z_1, ..., z_p \sim \text{iid}\mathcal{N}(0, 1)$. Consider $\tilde{z} = (z_1, ..., z_p)$. $\tilde{x} = \tilde{\mu} + \Sigma^{1/2}\tilde{z}$. Then we have $\tilde{X} \sim \mathcal{N}_p(\tilde{\mu}, \Sigma)$.

Multivariate normal density only exists if Σ is invertible, in which case,

$$f(\tilde{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(\tilde{x} - \tilde{\mu})^T\Sigma^{-1}(\tilde{x} - \tilde{\mu})\}$$

Sampling distribution. Recall if $X_1, ..., X_n \sim \text{iid}\mathcal{N}(\mu, \sigma^2)$ and with $\bar{x} = \frac{1}{n}\sum X_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$. In this case, we would have

$$\sum_{i=1}^{n}(X_i - \mu)^2 \sim \sigma^2 \chi_n^2$$

independent of $\bar{x}$. Call it $W_1(\sigma^2, n)$ distribution.

Then we have generalization from χ^2 to W distribution:

$$\sum_{i=1}^{n}(X_i \bar{X})^2 \sim \sigma^2 \chi_{p-1}^2 = W_1(\sigma^2, p-1_)$$

which is also independent of $\bar{X}$. If

$$\bar{\tilde{X}} = \frac{1}{n}\sum X_i \sim \mathcal{N}_p(\tilde{\mu}, \frac{1}{n}\Sigma)$$

$$\sum_{i=1}^{n}(\tilde{X} - \tilde{\mu})^T \sim W_p(\Sigma, n)$$

Lastly, we have

$$\sum(\tilde{X}_i - \bar{\tilde{X}})(\tilde{X}_i - \bar{\tilde{X}})^T \sim W_p(\Sigma, n-1)$$

which is Wishart Distribution from text [1].

Partitioning multivariate normal random vectors. Partition the random vector $\tilde{X} \in \mathbb{R}^p$ into $\tilde{X} = (\tilde{X}_1, \tilde{X}_2)^T$ while $\tilde{X}_i \in \mathbb{R}^{p_1}$, and $\tilde{X}_2 \in \mathbb{R}^{p_2}$ while $p_1 + p_2 = p$.

Suppose $\tilde{X} \sim \mathcal{N}_p(\tilde{\mu}, \Sigma)$. Then write

$$(\tilde{X}_1, \tilde{X}_2)^T \sim \mathcal{N}_{p_1+p_2}\left(\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \begin{pmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22}\end{pmatrix} \right)$$

In this case, we note that Σ_{11} is $p_1 \times p_1$, Σ_{22} is $p_2 \times p_2$, while $\Sigma_{12} = \Sigma_{21}^T$ and they are $p_1 \times p_2$.

Then $\tilde{X}_1 \sim \mathcal{N}_{p_1}(\tilde{\mu}_1, \Sigma_{11})$ and $\tilde{X}_2 \sim \mathcal{N}_{p_2}(\tilde{\mu}_2, \Sigma_{22})$. $\tilde{X}_1$ and $\tilde{X}_2$ are independent if and only if $\Sigma_{12} = 0$.

Also $\tilde{X}_2 | \tilde{X}_1 = \tilde{x}_1 \sim \mathcal{N}$. For proofs, see Section 3.1 of text [1].

Recall $X_1, X_2, ..., X_n \sim \text{iid}\mathcal{N}(\mu, \sigma^2)$ with $\bar{X} = \frac{1}{n}\sum X_i$ while $s_\mu^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ then we have

(1) $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ while

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1), \text{ i.e., } \frac{n(\bar{x} - \mu)^2}{\sigma^2} \sim \chi_1^2$$

(2) also $(n-1)s_\mu^2 \sim \sigma^2 \chi_{n-1}^2$, independent of $\bar{X}$. Then for $\tilde{X}_1, ..., \tilde{X}_n \sim \text{iid}\mathcal{N}_p(\tilde{\mu}, \Sigma)$ while $\bar{\tilde{X}} = \frac{1}{n}\sum \tilde{X}_i$, $S_\mu = \frac{1}{n-1}\sum(\tilde{X}_i - \bar{\tilde{X}})(\tilde{X}_i - \bar{\tilde{X}})^T$.

(3) moreover, $\bar{\tilde{X}} \sim \mathcal{N}_p(\tilde{\mu}, \frac{1}{n}\Sigma)$ implies $n(\bar{\tilde{X}} - \tilde{\mu})^T \Sigma^{-1}(\bar{\tilde{X}} - \tilde{\mu}) \sim \chi_p^2$, while notice that $\sqrt{n}\Sigma^{-1/2}(\bar{\tilde{X}} - \tilde{\mu}) \sim \mathcal{N}_p(\tilde{0}, I_p)$

(4) $(n-1)S_\mu \sim W_p(\Sigma, n-1)$ independent of $\bar{\tilde{X}}$ which led to

$$n(\bar{\tilde{X}} - \tilde{\mu})^T S_\mu^{-1}(\bar{\tilde{X}} - \tilde{\mu}) \sim T^2(p, n-1)$$

3 THEORY OF MULTIVNORMAL

Before we proceed, let us recall the following. The pdf of $X \sim \mathcal{N}_p(\mu, \Sigma)$ is

$$f(x) = |2\pi\Sigma|^{-1/2} \exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}$$

The expectation is $\mathbb{E}X = \mu$, the covariance can be calculated as $\text{var}X = \mathbb{E}(X-\mu)(X-\mu)^T = \Sigma$.

Linear transformations turn normal random variables into normal random variables. If $X \sim \mathcal{N}_p(\mu, \Sigma)$ and $\mathcal{A}(p \times p)$, $c \in \mathbb{R}^p$, then $Y = \mathcal{A}X + c$ is a p-variate Normal, i.e.

$$Y \sim \mathcal{N}_p(\mathcal{A}\mu + c, \mathcal{A}\Sigma\mathcal{A}^T)$$

If $X \sim \mathcal{N}_p(\mu, \Sigma)$, then the Mahalanobis transformation is

$$Y = \Sigma^{-1/2}(X-\mu) \sim \mathcal{N}_p(0, \mathcal{I}_p)$$

and it holds that

$$Y^TY = (X-\mu)^T\Sigma^{-1}(X-\mu) \sim \chi^2_p$$

Often it is interesting to partition X into sub-vectors X_1 and X_2. The following theorem tells us how to correct X_2 to obtain a vector which is independent of X_1.

Theorem 3.0.1. *Let* $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, $X_1 \in \mathbb{R}^r$, $X_2 \in \mathbb{R}^{p-r}$. *Define* $X_{21} = X_2 - \Sigma_{21}\Sigma_{11}^{-1}X_1$ *from the partitioned covariance matrix*

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Then

$$X_1 \sim \mathcal{N}_r(\mu_1, \Sigma_{11})$$
$$X_{21} \sim \mathcal{N}_{p-4}(\mu_{21}, \Sigma_{22,1})$$

are independent with

$$\mu_{2,1} = \mu_2 - \Sigma_{21}\Sigma_{11}^{-1}\mu_1, \Sigma_{22,1} = \Sigma_{22} - \Sigma_{22}\Sigma_{11}^{-1}\Sigma_{12}$$

Theorem 3.0.2. *If* $X \sim \mathcal{N}_p(\mu, \Sigma)$, $\mathcal{A}(q \times p)$, $c \in \mathbb{R}^q$ *and* $q \leq p$, *then* $Y = \mathcal{A}X + c$ *is a* q*-variate Normal, i.e.*

$$Y \sim \mathcal{N}_q(\mathcal{A}\mu + c, \mathcal{A}\Sigma\mathcal{A}^T)$$

The conditional distribution of X_2 *given* X_1 *is given the next.*

Theorem 3.0.3. *The conditional distribution of* X_2 *given* $X_1 = x_1$ *is normal with mean* $\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1)$ *and covariance* $\Sigma_{22,1}$, *i.e.*

$$(X_2|X_1 = x_1) \sim \mathcal{N}_{p-r}(\mu_2 + \Sigma_{22}\Sigma_{11}^{-1}(x_1-\mu_1), \Sigma_{22,1})$$

3.1 Hypothesis Testing

A 1-α confidence for μ is $\{\mu : \frac{n(\bar{x}-\mu)^2}{s^2} < F_{1-\alpha,1,n-1}\}$ which implies that $\mu \in [\bar{x} \pm t_{1-\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}]$. For arbitrary p, we want test null hypothesis $H_0 : \tilde{\mu} = \tilde{\mu}_0$. We want to compute $T^2 = n(\tilde{\bar{X}} - \tilde{\mu}_0)^T S_u^{-1}(\tilde{\bar{X}} - \tilde{\mu})$. For $p > 2$, cannot draw it, it is an ellipsoid.

4 MAXIMUM LIKELIHOOD ESTIMATION

Let $\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_n \sim \text{iid} f(\tilde{X}, \tilde{\theta})$. Consider f is a density on $\mathbb{R}^p$, $\tilde{\theta} \in \Theta \subset \mathbb{R}^k$. Given an observation ed data matrix

$$\mathbf{X} = \begin{pmatrix} \tilde{X}_1^T \\ \vdots \\ \tilde{X}_n^T \end{pmatrix}, \text{ for } n \times p$$

the likelihood function

$$L : \Theta \to \mathbb{R}^+, \text{ i.e. } L(\tilde{\Theta}, \mathbf{X}) = \prod_{i=1}^{n} f(\tilde{X}_i, \tilde{\Theta})$$

and

$$\hat{\tilde{\Theta}} = \hat{\tilde{\Theta}} = \arg\max_{\Theta} L(\tilde{\Theta}, X) = \arg\max_{\Theta} (\tilde{\Theta}, X)$$

Example 4.0.1. Consider $\tilde{X}_1, ..., \tilde{X}_n$ i.i.d. $\mathcal{N}_p(\tilde{\mu}, \Sigma)$ with Σ is known, then $l(\tilde{\Theta}_i, X) = \sum_{i=1}^{n} \log f(\tilde{x}_i, \tilde{\Theta})$. Then

$$\begin{aligned} l(\tilde{\theta}, X) &= \sum_{i=1}^{n} \log f(\tilde{X}_i, \tilde{\Theta}) \\ &= -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log(|\Sigma|) - \frac{1}{2} \sum_{i=1}^{n} (\tilde{X}_i - \tilde{\mu})^T \Sigma^{-1} (\tilde{X}_i - \tilde{\mu}) \\ \tilde{\Theta} &= \tilde{\mu} \end{aligned}$$

and hence

$$\hat{\tilde{\mu}} = \arg\min \big[\sum_{i=1}^{n} (\tilde{X}_i - \tilde{\mu})^T \Sigma^{-1} (\tilde{X}_i - \tilde{\mu}) \big]$$

Now we have

$$(\tilde{X}_i - \tilde{\mu})^T \Sigma^{-1} (\tilde{X}_i - \tilde{\mu}) = [(\tilde{X}_i - \bar{\tilde{X}}) + (\tilde{X} - \tilde{\mu})]^T \Sigma^{-1} [(\tilde{X}_i - \tilde{\tilde{\mu}}) + (\bar{\tilde{X}} - \tilde{\mu})]$$

note that $\hat{\tilde{\Theta}} = \hat{\tilde{\Theta}} = \arg\max L(\tilde{\Theta}_i, X) = \arg\max l(\tilde{\Theta}, X)$. If Σ is unknown, still get $\tilde{\mu} = \bar{X}$. The first term has no $\tilde{\mu}$ so we ignore it. Cross product term sums to zero. The second implies $\hat{\tilde{\mu}} = \arg\min\{(\bar{\tilde{X}} - \tilde{\mu})^T \Sigma^{-1} (\bar{\tilde{X}} - \tilde{\mu})\}$. Hence, Σ^{-1} is positive definite, and $\tilde{\mu} = \bar{\tilde{X}}$.

Example 4.0.2. Consider $\tilde{Y} \sim \mathcal{N}_n(X\tilde{\beta}, \sigma^2 I)$ as a linear regression model. X is known with $n \times p$ full rank of p. Then

$$L(\tilde{\beta}, \sigma^2, y, X) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\big\{ -\frac{1}{2\sigma^2} (y_i - \tilde{X}^T \tilde{\beta})^2 \big\}$$

Then you solve for MLE for parameter $\hat{\tilde{\beta}}$ to be

$$\hat{\tilde{\beta}} = (X^T X)^{-1} X^T \tilde{y} = \hat{\tilde{\beta}}_{OLS}$$

and

$$\frac{\partial}{\partial \sigma^2}() = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} (\tilde{y} - X\tilde{\beta})^T (\tilde{y} - X\tilde{\beta}) \stackrel{\text{set}}{=} 0$$

The Fisher information in a sample $\tilde{X}_1, ..., \tilde{X}_n \sim \text{iid} f(\tilde{X}, \tilde{\theta})$ is

$$\mathcal{F}_n = -\mathbb{E}\Big[\frac{\partial^2}{\partial}\tilde{\theta}\partial\tilde{\theta}^T l(\theta, (\frac{x_i}{x_n}))\Big]$$

The Fisher info matrix for $\tilde{\Theta}$ in a single observation of $\tilde{X} \sim f(\tilde{X}, \tilde{\theta})$ is

$$\mathcal{F}_1 = -\mathbb{E}\Big[\frac{\partial^2}{\partial\tilde{\theta}...\partial\tilde{\theta}^T}\log f(\tilde{X}, \tilde{\theta})\Big]$$

and we can prove this.

Moreover, we have

$$\mathcal{F}_1 = \Sigma^{-1}$$
$$\mathcal{F}_n = n\Sigma^{-1} = \frac{1}{n}\Sigma^{-1} = \mathcal{F}_1 \text{ for } \tilde{\tilde{X}} \sim \mathcal{N}_p(\tilde{\mu}, \frac{1}{n}\Sigma)$$

Remark 4.0.3. HW3 is posted. It is due Tue. Feb. 27. First exam is Tuesday March 8th (through Section 7.1).

For HW3, note

(1) problem 1 will post a handout soon

(2) problem 2-3, end of chapter 6 problems (math)

(3) problem 4-7, try them now, note helpful handout on the way. Problem 6 and 7 are not graded.

Next Tuesday start Ch 7. What's important in 7.1?

Test Problem 1? No.

Test Problem 2? Yes but T^2/F-dist version of this test, the asymptotic chi-square approximation is unnecessary. Confidence region for μ? Yes.

Test Problem 3? No.

Test Problem 4? No.

Suppose that $x_i{}_{i=1}^n$ is an i.i.d. sample from a population with pdf $f(x;\theta)$. The aim is to estimate $\theta \in \mathbb{R}^k$ which is a vector of unknown parameters. The likelihood function is defined as the joint density $L(\chi, \theta)$ of the observations x_i considered as a function of θ:

$$L(\chi;\theta) = \prod_{i=1}^n f(x_i;\theta),$$

where χ denotes the sample of the data matrix with the observations $x_1^T, ..., x_n^T$ in each row. The MLE of θ is defined as

$$\hat{\theta} = \arg\max_\theta L(\chi;\theta),$$

which is equivalent since the logarithm is a monotone one-to-one function. Hence

$$\hat{\theta} = \arg\max_\theta L(\chi;\theta) = \arg\max_\theta l(\chi;\theta).$$

Example 4.0.4. Consider a sample $\{x_i\}_{i=1}^n$ from $\mathcal{N}_p(\mu, \mathcal{I})$, i.e., from the pdf

$$f(x;\theta) = (\pi)^{-p/2}\exp\Big\{-\frac{1}{2}(x-\theta)^T(x-\theta^T)\Big\}$$

where $\theta = \mu \in \mathbb{R}^p$ is the mean vector parameter. The log-likelihood is in this case

$$l(\chi;\theta) = \sum_{i=1}^n \log\big\{f(x_i;\theta)\big\} = \log(2\pi)^{-np/2} - \frac{1}{2}\sum_{i=1}^n (x_i-\theta)^T(x_i-\theta).$$

The term $(x_i - \theta)^T(x_i - \theta)$ equals

$$(x_i - \bar{x})^T(x_i - \bar{x}) + (\bar{x} - \theta)^T(\bar{x} - \theta) + 2(\bar{x} - \theta)^T(\bar{x} - \theta).$$

Summing this term over $i = 1, ..., n$ we see that

$$\sum_{i=1}^{n}(x_i - \theta)^T(x_i - \theta) = \sum_{i=1}^{n}(x_i - \bar{x})^T(x_i - \bar{x}) - \frac{n}{2}(\bar{x} - \theta)^T(\bar{x} - \theta).$$

Only the last term depends on θ for this family of pdfs $f(x, \theta)$.

Consider $X_1, X_2, ..., X_n \sim$ iid$\mathcal{N}(\mu, \Sigma)$. For MVN assumption reasonable, look at QQ-plot. Pairwise scatter plots for close enough to SLR model. There is no good way to test for MVN. Instead, we look for reasonable to respect it. Absent those, proceed with normal methods. Transformation can be useful as well. For

$$\bar{\tilde{X}} = \frac{1}{n}\sum X_i, S = \frac{1}{n-1}\sum(\tilde{X}_i - \bar{\tilde{x}})(\tilde{x}_i - \bar{\tilde{x}})^T$$

while $S = \frac{1}{n-1}X^THX$ while $H = I - \frac{1}{n}11^T$. Then we have

$$n(\bar{\tilde{X}} - \tilde{\mu})^T S^{-1}(\bar{\tilde{X}} - \tilde{\mu}) \sim T^2(p, n-1) = \frac{(n-1)p}{n-p}F_{p,n-p}$$

To test

$$H_0 : \tilde{\mu} = \tilde{\mu}_0$$

We want to compute

$$T^2 = n(\bar{\tilde{X}} - \tilde{\mu}_0)^T S^{-1}(\bar{\tilde{X}} - \tilde{\mu})$$

Then compare

$$n - p(n-1)pT^2$$

with

$$F_{p,n-p}$$

and we have

$$p - val = \mathbb{P}(F_{p,n-p} > \frac{n-p}{(n-1)p}T^2$$

Reject H_0 at level of $p - val < \alpha$. This is to say we reject if $T^2 > \frac{(n-1)p}{n-p}F_{1-\alpha;p,n-p}$. The $1 - \alpha$ confidence allowed for $\tilde{\mu}$ is

$$\{\tilde{\mu}_0 : H_0 : \tilde{\mu} = \tilde{\mu}_0 \text{ is accepted.}\}$$

4.1 Types of Confidence Intervals

There are three types of confidence intervals.

(1) An individual $1 - \alpha$ C.I. for $\tilde{a}^T\tilde{\mu}$ with

$$\bar{\tilde{X}} \sim \mathcal{N}_p(\tilde{\mu}, \frac{1}{n}\Sigma)$$

$$a^T\bar{\tilde{X}} \sim \mathcal{N}_1(\tilde{a}^T\tilde{\mu}, \frac{1}{n}\tilde{a}^T\Sigma\tilde{a})$$

Then for $1 - \alpha$ C.I. is $\tilde{a}^T\bar{\tilde{X}} \pm z_{1-\alpha/2}\sqrt{\frac{1}{n}\tilde{a}^T\Sigma\tilde{a}}$ or $\tilde{a}^T\bar{\tilde{X}} \pm t_{1-\alpha/2;n-1}\sqrt{\frac{1}{n}\tilde{a}^TS\tilde{a}}$.

(2) Introduce simultaneous $1 - \alpha$ C.I.s for $\tilde{a_1}^T\tilde{\mu}$, ..., $\tilde{a_q}^T\tilde{\mu}$. Bonferroni method! If I have q indivudla $1 - \alpha$ C.I.s, my simultaneous confidence interval that all q of them

cover is at least $1 - q\alpha$. If one needs $1 - \alpha$ simultaneous confidence level, one needs to construct each individual C.I. to have confidence level $1 - \frac{\alpha}{q}$.

Aside, equal allocation of error rate is not required.

More, 96% confidence interval for intervals. We could od 99% for each or even 99.5%. If $1 - \alpha$ is designed level, need $1 - \alpha = 1 - \sum \alpha_i$ Using $1 - \alpha/q$ confidence for each,

$$\tilde{a}_1^T \bar{\tilde{X}} \pm t_{1-\frac{\alpha}{2q};n-1} \sqrt{\frac{1}{n} \tilde{a}_1^T S \tilde{a}_1}$$

$$\tilde{a}_q^T \bar{\tilde{X}} \pm t_{1-\frac{\alpha}{2q};n-1} \sqrt{\frac{1}{n} \tilde{a}_q^T S \tilde{a}_q}$$

(3) Simultaneous $1 - \alpha$ C.I. for all possible linear combinations $\tilde{a}^T \tilde{\mu}$ Then

$$\tilde{a}^T \bar{\tilde{X}} \pm \sqrt{\frac{p(n-1)}{n-p} F_{1-\alpha;p,n-p} \frac{1}{n} \tilde{a}^T S \tilde{a}}$$

Discuss. When to use (2) and when to use (3)? If you want to simultaneous $1 - \alpha$ C.I.s for the compound means, use method (2) Bonferrroni the intervals are shorter. When to use (3)? We are $1-\alpha$ confident that $\tilde{a}^T \tilde{\mu} \in \tilde{a}^T \bar{\tilde{X}} \pm \sqrt{\frac{(p(n-1)}{n-p} F_{1-\alpha;p,n-p} \frac{1}{n} \tilde{a}^T S \tilde{a}}$ for all $\tilde{a} \in \mathbb{R}^P$. If you are studying a linear combination $\tilde{a}^T \tilde{\mu}$ that is only of interest because of the data observed, use method (3). Method (3) gives simultaneous $1 - \alpha$ confidence in intervals for all $\tilde{a}^T \tilde{\mu}$, including those suggested by the data, i.e. method (3) protects against data snooping. What is "protect against" exactly? Protect against reaching conclusions that event genuinely supported by the data, i.e., type I error.

5 PRINCIPLE COMPONENT ANALYSIS

Principle Component Analysis finds linear combinations of variables that best explain covariation structure of the variables. The common purposes:

(1) Dimension reduction: explain covariance structure of p variables using q variables (QCP);

(2) Interpretation: Find features of data that are important for explaining covariance.

Consider population PCA, we have $X = (X_1, ..., X_p)^T$ has mean vector $\tilde{\mu}$ covariance matrix Σ, we are not assuming MVN here. Consider linear transformation

$$\delta_1^T \tilde{X} = \delta_{11} X_1 + \delta_{21} X_2 + \cdots + \delta_{p1} X_p$$

...

$$\delta_p^T \tilde{X} = \delta_{1p} X_1 + \delta_{2p} X_2 + \cdots + \delta_{pp} X_p$$

Let $\triangle_{p \times p} = (\delta_1, ..., \delta_p)$ and let $\tilde{Y} = \triangle^T (\tilde{X} - \tilde{\mu})$, then

$$\text{var}(Y) = \triangle \Sigma \triangle^T, \mathbb{E}(\tilde{Y}) = \tilde{0}$$

Definition 5.0.1. PCs of $\tilde{X}$ are the uncorrelated standardized linear combinations (SLCs) $\delta_1^T (X - \mu), ..., \delta_P^T (X - \mu)$, whose variance is as large as possible.

$\star$ standardized have means $||\delta_j||^2 = \delta_j^T \delta_j = 1$:

(1) The first PC is $\delta_1^T (X - \mu)$ where δ_1 is solution to $\max_{\{||\delta||=1\}} \text{var}(\delta^T X) = \max_{||\delta||=1} \delta^T \Sigma \tilde{\delta}$

(k) The kth PC is $\delta_k^T (X - \mu)$ where δ_k is solution to

$$\max \text{var}(\delta^T X) = \max \delta^T \Sigma \delta$$

subject to $\text{cov}(\delta^T X, \delta_j^T X) = \delta^T \Sigma \delta_j = 0$ for each $j = 1, ..., k-1$. Ends at step p. Result: let $\text{var}(X) = \Sigma = \Gamma\Lambda\Gamma^{-1}$ where $\Lambda = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_p)$ satisfies $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$; then $\tilde{Y} = \Gamma^T(X - \mu)$ where $\Sigma = \Gamma\Lambda\Gamma^T$ is spectral decomposition.

Proposition 5.0.2. *Theorem 11.1 on page 323 [1]. Given $X \sim (\mu, \Sigma)$ and $\Sigma = \Gamma\Lambda\Gamma^T$ and the PC transform $\tilde{Y} = \Gamma^T(X - \mu)$ which satisfies*

(1) $\mathbb{E}(Y_j) = \mathbb{E}[\gamma_j^T(\tilde{X} - \mu)] = 0$

(2) Then $var(Y_j) = var(\gamma_j^T X) = \gamma_j^T \Sigma \gamma_j = \gamma_j^T \Gamma\Lambda\Gamma^T \gamma_j = (\gamma_j^T\gamma_1 \dots \gamma_j^T\gamma_p)\Lambda(\gamma_1^T\gamma_j, ..., \gamma_p^T\gamma_j)^T = \tilde{e}_j^T \Lambda \tilde{e}_j$ where $e_j = (0, ..., 0, 1, 0, ...,)^T$ where one in jth position and 0 elsewhere. This is the jth diagonal entry of Λ which is λ_j.

(3) $cov(Y_j, Y_k) = \gamma_j^T \Gamma\Lambda\Gamma^T \gamma_k$

(4) $tr(\Sigma) = \lambda_1 + \cdots + \lambda_p$. This gives us $\sigma_1^2 + \dots \sigma_p^2 = \lambda_1 + \cdots + \lambda_p$ and also $var(x_1) + \dots var(x_p) = var(Y_1) + \cdots + var(Y_p)$. Total variance of P(s) = Total variance of original variables.

(5) The correlation between jth variable X_j and the kth PC Y_k, $corr(X_j, Y_k) = p239 = \gamma_{jk}\sqrt{\lambda_k/\sigma_j^2}$.

Principal Components for Data:

Pretty much replace μ by $\bar{X}$ and Σ by S. Let $X_1, ..., X_n$ be random sample from MV population $X \sim (\mu, \Sigma)$ while μ and Σ unknown. Let $X = (X_1^T, ..., X_n^T)'$ with $n \times p$ data matrix. Let $\bar{X}^T \frac{1}{n}\sum X_i$ and $S = \frac{1}{n-1}\sum(X - \bar{X})(X_i - \bar{X})^T$. Y here is the matrix of sample PCs. Recall $S_x = \frac{1}{n-1}X^T H X$ while $H = I - \frac{1}{n}11^T$ thus, $S_Y = \frac{1}{n-1}Y^T H Y$. Estimate $Y = \Gamma^T(X - \mu)$ by $Y_i = G^T(X_i - \bar{X})$ for $i = 1, ..., n$ if and only if $(y_1^T, ..., y_n^T)' = Y = (X - 1_n\bar{X}^T)G = HXG$. Then this solves

$$\frac{1}{n-1}G^T X^T H X G = G^T S_X G = G^T G L G^T G$$

6 CANONICAL CORRELATION ANALYSIS

6.1 Canonical Correlation

Consider a company is training to find exam that will review potential for good performance in sales. Let response

$$\tilde{Y} = (Y_1, Y_2, Y_3)^T$$

which are growth in sales, profitability, and new accounting sales, respectively. More, we have

$$\tilde{X} = (X_1, X_2, X_3, X_4)^T$$

which are creativity, mechanical reasoning, abstract reasoning, and mathematics, respectively. What aspects of $\tilde{X}$ and $\tilde{Y}$ are most correlated? Mathematical formulation is: find the vectors $\tilde{a} \in \mathbb{R}^q$ and $\tilde{b} \in \mathbb{R}^p$ such that

$$\rho = \text{corr}(\tilde{a}^T\tilde{X}, \tilde{b}^T\tilde{Y})$$

is maximized. Note that

$$\rho = \text{corr}(\tilde{a}\tilde{X}, \tilde{b}^T\tilde{Y})$$

is called canonical correlation effect $\tilde{a}$ and $\tilde{b}$ vectors. The random variables $\eta = \tilde{a}^T\tilde{X}$ and $\varphi = \tilde{b}^T\tilde{Y}$ are canonical correlation variables. The method is called canonical correlation analysis.

Let $k = \min\{p, q\}$. There are k canonical correlations. Note that $\tilde{a}_1, ..., \tilde{a}_k$, and $\tilde{b}_1, ..., \tilde{b}_k$ are found by

Step 1. Choose $\tilde{a}_1$ and $\tilde{b}_1$ are maximized $\text{corr}(\tilde{a}_1^T X, \tilde{b}_1^T Y) = \rho_1$.
For $j = 2, ...k$
step j, choose $\tilde{a}_j$ and $\tilde{b}_j$ to
maximize $\text{corr}(\tilde{a}_j^T X, \tilde{b}_j^T Y) = \rho_j$
subject to $\text{corr}(\tilde{a}_j^T X, \tilde{b}_j^T Y) = 0$ and $\text{corr}(\tilde{b}_j^T X, \tilde{b}_j^T Y) = 0$ while $i = 1, ..., j-1$.

Here ρ_i is the ith canonical correlation while $\eta_i = \tilde{a}_i^T \tilde{X}$ and $\varphi_i = \tilde{b}_i^T \tilde{Y}$ are ith canonical correlation variables $\tilde{a}_i$ and $\tilde{b}_i$ are the ith canonical correlation vectors. That is the "definition" of canonical correlation analysis, specified as an optimization problem. The solution is, assume

$$(\tilde{X}, \tilde{Y})^T \sim q + p\left[(\tilde{\mu}, \tilde{\nu})^T, \begin{pmatrix} \Sigma_{XX(q\times q)} & \Sigma_{XY(q\times p)} \\ \Sigma_{YX(p\times q)} & \Sigma_{YY(p\times p)} \end{pmatrix}\right]$$

then we have

$$\text{corr}(\tilde{a}^T \tilde{X}, \tilde{b}^T \tilde{Y}) = \frac{\tilde{a}^T \Sigma_{XY} \tilde{b}}{\sqrt{\tilde{a}^T \Sigma_{XX} \tilde{a} + \tilde{b}^T \Sigma_{YY} \tilde{b}}}$$

To maximize this correlation and guarantee a unique solution we maximize $\tilde{a}^T \Sigma_{XY} \tilde{b}$ subject to $\tilde{a}^T \Sigma_{XX} \tilde{a} = 1 = \text{var}(\tilde{a}^T \tilde{X})$ and $\tilde{b}^T \Sigma_{YY} \tilde{b} = 1 = \text{var}(\tilde{b}^T \tilde{Y})$.

One can pause here to read Single Value Decomposition below before moving on.

Non apply the Single Value Decomposition to

$$K = \Sigma_{XX}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2}$$

with size $q \times p$. Then

$$\text{SVD} = (\gamma_1, ..., \gamma_k) \begin{pmatrix} \lambda_1^{1/2} & & \\ & \ddots & \\ & & \lambda_k^{1/2} \end{pmatrix} \begin{pmatrix} \delta_1^T \\ \vdots \\ \delta_k^T \end{pmatrix}$$

Letting $\eta_i = \tilde{a}^T \tilde{X}$ and $\varphi_i = \tilde{b}^T \tilde{Y}$ for $i = 1, ..., k$. Thus

$$\text{corr}(\varphi_i, \eta_i) = \frac{\tilde{a}_i^T \Sigma_{XY} \tilde{b}_i}{\sqrt{\tilde{a}_i^T \Sigma_{XX} \tilde{a}_i \tilde{b}_i^T \Sigma_{YY} \tilde{b}}}$$

and then $\tilde{a}_i^T \Sigma_{XX} \tilde{a}_i = \gamma_i^T \Sigma_{XX}^{-1/2} \Sigma_{XX} \Sigma_{XX}^{-1/2} \gamma_i = \gamma_i^T \gamma_i = 1$ and $\tilde{b}_i \Sigma_{YY} \tilde{b}_i - \tilde{\delta}_i^T \Sigma_{YY}^{-1/2} \Sigma_{YY} \Sigma - YY^{-1/2} \tilde{\delta}_i = \tilde{\delta}_i \tilde{\delta}_i = 1$, then

$$\text{corr}(\varphi_i, \eta_i) = \tilde{a}_i \Sigma_{XY} \tilde{b}_i$$

Note that $\Gamma^T \gamma_i = (\gamma_1^T \gamma_i, ..., \gamma_k^T \gamma_i)^T$. Then $\Delta^T \delta_i = $ sth with 1 in ith position and 0 elsewhere. Thus, $(\Gamma^T \gamma_i)^T \Lambda (\Delta^T \delta_i) = $ ith diagonal entry of $\Lambda = \lambda_i^{-1/2}$. An exercise can be to verify that

$$\text{corr}(\tilde{a}_i \tilde{X}, \tilde{b}_j^T \tilde{Y}) = 0$$

for $j \neq i$. Also

$$\begin{aligned} \text{corr}(\eta_i, \eta_j) &= \text{corr}(\tilde{a}_i^T \tilde{X}, \tilde{a}_j^T \tilde{X}) \\ &= \tilde{a}_i^T \Sigma_{XX} \tilde{a}_j \\ &= \gamma^T \Sigma_{XX}^{-1/2} \Sigma_{XX}^{-1/2} \gamma_j \\ &= \gamma_i^T \gamma_j = 0 \end{aligned}$$

Moreover, $\text{corr}(\varphi_i, \varphi_j) = 0$ if $j \neq i$. The canonical correlation variables are maximally correlated to each other within their pairs and uncorrelated between pairs. That is, $\text{corr}(\eta_i, \eta_j)$ is maximized, but $\text{corr}(\eta_i, \eta_j) = \text{corr}(\eta_i, \varphi_j) = \text{corr}(\varphi_i, \varphi_j) = 0$ for $j \neq i$.

The $k = \min(q, p)$ for canonical correlation are found by computing SVD

$$K = \Sigma_{XX}^{-1/2}\Sigma_{XX}, \Sigma_{YY}^{-1/2} = \Gamma\Lambda\Lambda^T = (\gamma_1, ..., \gamma_k) \begin{pmatrix} \lambda_1^{1/2} & & \\ & \ddots & \\ & & \lambda_k^{1/2} \end{pmatrix} \begin{pmatrix} \delta_1^T \\ \vdots \\ \delta_k^T \end{pmatrix}$$

The$i$th canonical correlation vectors are

$$\tilde{a}_i = \Sigma_{XX}^{-1/2}\tilde{\gamma}_i = \eta_i$$

and

$$\tilde{b}_i = \Sigma_{YY}^{-1/2}\tilde{\gamma}_i = \varphi_i$$

and the ith canonical correlation coefficient is

$$\eta_i = \text{corr}(\eta_i, \varphi_i) = \lambda_i^{1/2}$$

6.1.1 Single Value Decomposition

Let K be a $q \times p$ matrix with rank $= k = \min(q, p)$. Then

$$K = \Gamma\Lambda\Delta^T$$

where $\Lambda = \text{diag}(\lambda_1^{1/2}, ..., \lambda_k^{1/k})$ with $k \times k$ while $\lambda_1, ..., \lambda_k$ are the nonzero eigen values of KK^T which is a $q \times q$ matrix and of K^TK. Then we have $\Gamma_{q\times k} = (\gamma_1, ..., \gamma_k)$ with γ_i is the ith eigen vector of KK^T, i.e., $KK^T\gamma_i = \lambda_i\gamma_i$ while $\Gamma^T\Gamma = I_k$.

Next, $\Delta_{p\times k} = (\delta_1, ..., \delta_k)$ where δ_i is the ith eigen vector of K^TK, i.e., $K^TK\delta_i = \lambda_i\delta_i$. Then $\Delta^T\Delta = I_k$.

Remark 6.1.1. The following about Homework 4.

Handout 14 is for Problem 1.

Handout 13 is for Problem 2a.

Handout 12 is for Problem 2b.

From homework 4, we can discuss Problem 2.

$$\bar{X}_1 - \bar{X}_2 \sim \mathcal{N}_p[\tilde{\mu}_1 - \tilde{\mu}_2 = \tilde{\delta}, (\frac{1}{n_1} + \frac{1}{n_2})\Sigma]$$

which gives us

$$(\frac{1}{n_1}+\frac{1}{n_2})^{-1}(\bar{X}_1-\bar{X}_2-\tilde{\delta})^T S^{-1}(\bar{X}_1-\bar{X}_2-\tilde{\delta}) \sim T^2(p, n_1+n_2-2) = \frac{(n_1+n_2-2)p}{n_1+n_2-p-1}F_{p,n_1+n_2-p-1}$$

which implies the confidence intervals $1-\alpha$, a confidence region, for $\tilde{\delta} = \tilde{\mu_1} - \tilde{\mu_2}$. Then simultaneous $1-\alpha$ confidence intervals for all $\tilde{a}^T(\tilde{\mu}_1 - \tilde{\mu}_2)$ are given by

$$\tilde{a}^T(\bar{X}_1 - \bar{X}_2) \pm \sqrt{\frac{p(n_1+n_2-2)}{n_1+n_2-p-1}F_{1-\alpha;p,n_1+n_2-p-1}} \cdot \sqrt{(\frac{1}{n_1}+\frac{1}{n_2})\tilde{a}^T S\tilde{a}}$$

Then one could take

$$\tilde{a}_1 = (1, 0, 0)^T, \tilde{a}_2 = (0, 1, 0)^T, \tilde{a}_3 = (0, 0, 1)^T$$

But since component mean differences are a pre-planned influence, there is no need for this overly conservative approach. Instead, we apply Bonferroni method.

We show the following.

For $j = 1$, $p = 3$, $\bar{X}_{1j} - \bar{X}_{2j} \pm t_{1-\frac{\alpha}{2p},n_1+n_2-2}\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})S_j^2}$. Note that divided by 2 means two sided and divided by p means there are p degrees of freedom.

Handout 15 is for Problem 3. Don't want to wait? Please try

(1) Try prcomp() or prmcomp();

(2) If you are cool, you'll solve PCA by first principles using only eigen().

Now we are back to canonical correlation. Find the linear combinations $\tilde{a}^T\tilde{X}$ and $\tilde{b}^T\tilde{X}$ (satisfying $\tilde{a}^T\Sigma_{XX}\tilde{a} = \tilde{b}^T\Sigma_{YY}\tilde{b} = 1$) that are most highly correlated, i.e., maximize $\tilde{a}^T\Sigma_{XY}\tilde{b}$. Then $\tilde{X} \in \mathbb{R}^q$, $\tilde{Y} \in \mathbb{R}^P$, $k = \min(p, q) = \text{rank}(\Sigma_{XY})$. The k canonical correlation are found by

$$\max \text{cor}(\tilde{a}_1^T\tilde{X}_1, \tilde{b}_1^T\tilde{Y})$$

For $j = 2, ..., k$, we want

$$\max \text{corr}(\tilde{a}_j^T\tilde{X}, \tilde{b}_j^T\tilde{Y})$$

subject to

$$\text{corr}(\tilde{a}_i^T\tilde{X}, \tilde{a}_j^T\tilde{X}_j) = 0, \text{corr}(\tilde{b}_i^T\tilde{Y}, \tilde{b}_j^T\tilde{Y}_j) = 0$$

The solution depends on doing a singular value decomposition on $K = \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}$, which sized to $q \times p$. After doing the above, one would have

$$K_{q\times p} = \Gamma_{q\times k}\Lambda_{k\times k}\Delta_{k\times p}^T$$

where $\Lambda = \text{diag}(\lambda_1^{1/2}, ..., \lambda_k^{1/2})$ while $\lambda_1 \geq \cdots \geq \lambda_k \geq 0$, $\Gamma = (\gamma_1, ..., \gamma_k)$ sized $q \times k$, and $\Delta = (\delta_1, ..., \delta_k)$ sized $p \times k$. Let (λ_i, γ_i) are ith eigenpair of KK^T while (λ_i, δ_i) are ith eigenpair of K^TK. The canonical correlation vectors are

$$\tilde{a}_i = \Sigma_{XX}^{-1/2}\gamma_i \text{ and } \tilde{b}_i = \Sigma_{YY}^{-1/2}\tilde{\delta}_i$$

for $i = 1, ..., k$. The canonical correlation variables are

$$\eta_i = \tilde{a}_i^T\tilde{X} \text{ and } \varphi_i = \tilde{b}_i^T\tilde{Y}$$

for $i = 1, ..., k$. The cc coefficients are

$$\rho_i = \text{corr}(\eta_i, \varphi_i) = \lambda_i^{1/2}$$

for $i = 1, ..., k$.

6.2 Practical Canonical Correlation

In practice we write

$$\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}$$

is unknown. Hence, we estimate it like sample covariance matrix,

$$S = \begin{pmatrix} S_{XX} & S_{XY} \\ S_{YX} & S_{YY} \end{pmatrix}$$

Then we do singular value decomposition on $\hat{K} = S_{XX}^{-1/2}S_{XY}S_{YY}^{-1/2}$ and we get

$$\hat{K} = GLD^T = (g_1, ..., g_k)\begin{pmatrix} l_1^{1/2} & & \\ & \ddots & \\ & & l_k^{1/2} \end{pmatrix}\begin{pmatrix} \delta_1^T \\ \delta_2^2 \\ \vdots \\ \delta_k^T \end{pmatrix}$$

and $\tilde{a}_i = S_{XX}^{-1/2}\tilde{g}_i$ and $\tilde{b}_i = S_{YY}^{-1/2}\tilde{d}_i$ for $i = 1, ..., k$ and $r_i = l_i^{1/2} = i$th sample cc coefficient. If S is the sample covariance matrix for $(\tilde{X}, \tilde{Y})^T$, then r_i is the sample correlation between $\tilde{a}^T\tilde{X}$ and $\tilde{b}^T\tilde{Y}$.

6.3 Inference for Canonical Correlation

Suppose we have data $((x_1, y_1)^T, ..., (x_n, y_n)^T) \sim \text{iid}\mathcal{N}_{q+p}[(\mu, \nu)^T, \Sigma]$ which can be written as

$$((x_1, y_1)^T, ..., (x_n, y_n)^T) \sim \text{iid}\mathcal{N}_{q+p}[\begin{pmatrix}\mu \\ \nu\end{pmatrix}, \begin{pmatrix}\Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY}\end{pmatrix}]$$

Suppose we wish to test $H_0 : \Sigma_{XY} = 0_{q\times p}$. Let $\rho_1 \geq \cdots \geq \rho_k \geq 0$ for $k = \min(q, p)$ denote the CC coefficients. Then $H_0 : \Sigma_{XY} = 0$ is equivalently expressed as $H_0 : \rho_1 = \rho_2 = \cdots = \rho_k = 0$. If S is the sample covariance matrix, and R is the sample correlation matrix, namely,

$$S = \begin{pmatrix}S_{XX} & S_{XY} \\ S_{YX} & S_{YY}\end{pmatrix}, R = \begin{pmatrix}R_{XX} & R_{XY} \\ R_{YX} & R_{YY}\end{pmatrix}$$

then the Wilk's likelihood ratio test stat is

$$\Lambda_0 = \frac{|S|}{|S_{XX}| \cdot |S_{YY}|} = \frac{|R|}{|R_{XX}| \cdot |R_{YY}|} = \prod_{i=1}^{k}(1 - l_i)$$

in which $l_i^{1/2}$ is the ith sample cc coefficient. Notice if the null hypothesis is true, we expect $l_i \approx 0$ thus $\Lambda_0 \approx 1$. Thus we should reject H_0 for small values of Λ_0. The distribution theory for Λ_0 is complicated. But under the multivariate normal assumption for large n,

$$-[n - \frac{1}{2}(p + q + 3)] \log \Lambda \sim \chi^2_{p\times q}$$

Suppose we reject $H_0 : \Sigma_{XY} = 0_{q\times p}$. We conclude not all CC coefficients are zero. Likewise if we reject

$$\rho_2 = \ldots \rho_k = 0$$

which suggest a strategy for for inference about CC's. Test $H_0 : \rho_1 = \rho_2 = \cdots = \rho_k = 0$. If reject test, $H_0 : \rho_2 = \cdots = \rho_k = 0$. If reject test, $H_0 := \cdots = \rho_k = 0$.

We can answer the question. How many CC coefficients are significantly nonzero? All k of them? Or less than k? To test that only S of the k cc coefficient are nonzero, $H_0 : \rho_{s+1} = \cdots = \rho_k = 0$ test stat is

$$\Lambda_S = \prod_{i=s+1}^{k}(1 - l_i)$$

Find the p-value by comparing

$$-[n - \frac{1}{2}(p + q + S)] \log \Lambda_S$$

to $\chi^2_{(p-s)(q-s)}$ distribution.

7 FACTORING ANALYSIS

Factor analysis is a model-based technique. for "explaining" the covariance (or correlation) between components of a random vector. Factor analysis resembles PCA, but it is not the same thing. Charles Spearman (in the early 20th century) invented this method. He developed a theory of general intelligence.

Suppose

$$\tilde{X} = (X_1, X_2, X_3)^T = (\text{classics}, \text{french}, \text{english})^T$$

One way to model the test score is

$$X_1 = \mu_1 + q_1 f + u_1, X_2 = \mu_2 + q_2 f + u_2, X_3 = \mu_3 + q_3 f + u_3$$

and μ_j is the overall mean score on test j and f is underlying intelligence of subject while "f" stands for factor where $\mathbb{E}(f) = 0$ and $\text{var}(f) = 1$. Note that q_j is the factor loading for test j and μ_j is the unique individual response and measurement error. "f" is built in the model but it is not observable and hence it is a latent variable. Then

$$\mathbb{E}(u) = 0, \text{var}(u) \text{ is diagonal } \text{cov}(f, u) = 0$$

General $X = (X_1, X_2, X_3)^T$ has mean μ and cov matrix Σ where

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

where $\Psi = \text{diag}(\Psi_{11}, \Psi_{22}, ..., \Psi_{pp})$ and Q is $p \times k$ of rank $k < p$. Why this structure? We have p variables.

$$\begin{aligned} X_1 &= \mu_1 + q_{11}F_1 + q_{12}F_2 + \cdots + q_{1k}F_k + v_1 \\ X_2 &= \mu_2 + q_{21}F_1 + q_{22}F_2 + \cdots + q_{2k}F_k + v_2 \\ &\vdots \\ X_p &= \mu_p + q_{p1}F_1 + q_{p2}F_2 + \cdots + q_{pk}F_k + v_p \end{aligned}$$

and thus we have

$$\underbrace{\tilde{X}}_{p\times 1 \text{r.o.}} = \underbrace{\tilde{\mu}}_{p\times 1} + \underbrace{Q}_{p\times k,\text{ fixed}} \underbrace{\tilde{F}}_{k\times 1} + \underbrace{\tilde{v}}_{p\times 1}$$

and thus

$$\tilde{F} = (F_1 \dots F_k)^T$$

while F_l is the lth common factor $l = 1, ..., k$ satisfies $\mathbb{E}(\tilde{F}) = \tilde{0}$ and $\text{var}(\tilde{F}) = I_k$. $\text{cov}(\tilde{F}, \tilde{u}) = 0_{k\times p}$ Then

$$Q_{p\times k} = q_{jl}, \text{ loading of variable } j \text{ on factor } l$$

$$\begin{aligned} \text{var}(X) = \mathbb{E}[(X-\mu)(X-\mu)^T] &= \mathbb{E}[(QE+U)(QF=U)^T] \\ &= Q\mathbb{E}(FF^T)Q^T + \mathbb{E}(UU^T) \end{aligned}$$

while cross product term is zero since $\text{cov}(F, u) = 0$. Now this becomes

$$\text{var}(X) = QI_kQ^T + \Psi$$

Thus, $\Sigma = \text{var}(X) = QQ^T + \Psi$. The jth diagonal entry of $\Sigma = QQ^T + \Psi$ is $\sigma_{jj} = \text{var}(X_j) = q_{j1}^2 + q_{j2} + \cdots + q_{jk}^T + \Psi_{jj}$. Let us call this $h_j^2 + \Psi_{jj}$, which is communality of variables j. Also,

$$q_{jl} = \text{cov}(X_j, F_l)$$

Factor loadings are not unique.

Example 7.0.1. Spearman's one-factor general intelligence model. Replace intelligence factor f, with a stupidity factor $f^\star = -f$. Then $X_j = \mu_j + q_j f + u_j$ where $q_j^\star = -q_j$. If $k > 1$, we have factors

$$\tilde{F} = (F_1 \dots F_k)^T$$

the factor model

$$\tilde{X} = \tilde{\mu} + Q\tilde{F} + \tilde{U}$$

can be equivalently expressed as

$$\begin{aligned} X &= \mu + QI_kF + U \\ &= \mu + QGG^TF + U \\ &= \mu + Q^\star F^\star + U \end{aligned}$$

where $Q^\star = QG$ and $F^\star = G^T\tilde{F}$ where G is any $k \times k$ orthogonal matrix.

Factors and factor loadings are not unique expressing a factor model $F^\star = G^T\tilde{F}$ and $Q^\star = QG$ is called factor rotation. In practices μ, Q and Ψ are unknown we have data $X_1, X_2, ..., X_n$ from which we estimate them.

Factor analysis has two steps

(1) Estimation Q and Ψ of QQ^T and Ψ with rank k and Ψ diagonal matrix, respectively. Then Q is not unique but QQ^T is.

(2) Choice of $\hat{Q}$ given $\hat{Q}\hat{Q}^T$ (not unique solution, have the choice).

8 DISCRIMINATION AND CLASSIFICATION

Classification refers to procedures that allocate objects into one of two or more well defined groups. Observation $\tilde{X} \in \mathbb{R}^p$ drawn from one of the population Pop 1, Pop 2, ..., Pop J. Which one is it? We want to allocate role:

$$R_j = \{\tilde{X} : \tilde{X} = \tilde{x} \text{ is allocated to Pop } j\}$$

with $j = 1, ..., J$. So $R_1, R-2, ..., R_J$ form a partition of $\mathbb{R}^p$. Note that authors do not seem to care about this distinction. It is all just Discriminant Analysis and that is okay.

8.1 Fisher Discriminant

Referring to page 418 of text [2]. Consider

$$X_{11}, X_{21}, ..., X_{n_1 1} \sim \text{iid}p(\mu_1, \Sigma)$$

$$X_{12}, X_{22}, ..., X_{n_2 2} \sim \text{iid}p(\mu_2, \Sigma)$$

Then we have

$$X_1 = \begin{pmatrix} X_{11}^T \\ \vdots \\ X_{n1}^T \end{pmatrix} X_2 = \begin{pmatrix} X_{12}^T \\ \vdots \\ X_{n2}^T \end{pmatrix}$$

Calculate sample mean vectors and covariance matrices in

$$\tilde{X}_1 = \frac{1}{n_1}X_1\tilde{1}, S_1 = \frac{1}{n_1 - 1}X_1^T H_{n_1} X_1$$

$$\tilde{X}_2 = \frac{1}{n_2}X_2\tilde{1}, s_2 = \frac{1}{n_2 - 2}X_2^T H_{n_2} X_2$$

Recall $H_n = I_n - \frac{1}{n}\tilde{1}\tilde{1}_n^T$ and pooled covariance matrix

$$S = \frac{1}{n_1 + n_2 - 2}[(n-1)S_1 + (n-1)S_2]$$

Choose a that maximizes the separation between the two groups separation between the two groups. Specifically, choose a to maximize

$$\frac{1}{S_y^2}(\overline{y_1} - \overline{y_2})^2 = \frac{[a^T(\overline{x_1} - \overline{x_2})]^2}{a^T S a}$$

Consider a one-dimensional linear transformation

$$y_{ij} = a^T X_{ij} \text{ for } i = 1, ..., n, j = 1, 2$$

Then

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{[a^T(\bar{x}_1 + \bar{x}_2)^2]}{a^T S a}$$

This is great. The max possible scaled square distance between means of one-dimensional linear transformations is exactly the Mahalonobis distance between the mean vector. The max possible separation between two p-variate data samples is gotten from the one dimensional linear transformation $y = a^T X$ for $a = S^{-1}(\bar{X}_1 - \bar{X}_2)$.

Now consider $J = 2$. groups $X_{1j}, ..., X_{nj} \sim \text{iid}p(\mu, \Sigma)$. Then

$$S_j = \frac{1}{n_j - 1} X_j^T H X_j$$

The overall mean is

$$X = \frac{1}{n_1 + \ldots n_j} \sum_{j=1}^{J} \sum_{i=1}^{n} X_{ij} = \frac{1}{n_1 + \cdots + n_j} (X_1^T, ..., X_J^2) \begin{pmatrix} 1_n \\ \vdots \\ 1_{n_j} \end{pmatrix}$$

Pooled covariance matrix

$$S = \frac{1}{n_1 + \cdots + n_j - J} [(n_1 - 1)S_1 + \cdots + (n_j - 1)S_J]$$

Recall the $J = 2$ situation where we sought $a \in \mathbb{R}^P$. For generalized $J > 2$, we can simplify to

$$\text{const} \times \frac{a^T B a}{a^T W a}$$

while B is between groups sum of squares matrix and W is within group sum of squares matrix. By the Theorem from linear algebra Theorem 2.5 on page 63, it is exactly that a which maximizes λ is

$$W^{-1} B a = \lambda a$$

Thus, λ is the largest value of $W^{-1}B$.

For $J = 2$, this has reduced form of

$$a = S^{-1}(\bar{X}_1 - \bar{X}_2)$$

and thus we conclude that one-dimensional linear transformation $y = a^T X$.

A recurring theme in this course has been "find the linear transformation of $\tilde{X}$ that maximizes...."

(1) Given random vector $\tilde{X}$ find $\gamma \in \mathbb{R}^P$ with $||\gamma|| = 1$ to maximize $\text{var}(\gamma^T X)$. PCA! Take γ to be first eigenvector of $\Sigma = \text{var}(X)$.

(2) Given $\tilde{X} \in \mathbb{R}^q$ and $\tilde{Y} \in \mathbb{R}^P$ find $\tilde{a}$ and $\tilde{b}$ find $\tilde{a}$ and $\tilde{b}$ to maximize $\text{corr}(a^T X, b^T Y)$. CCA! Answer is $\tilde{a} = \Sigma_{xx}^{-1/2} \gamma$ and $\tilde{b} = \Sigma_{YY}^{-1/2} \tilde{\delta}$ where $\tilde{\gamma}$ is first left eigenvector while $\tilde{\gamma}$ is right eigen vector of $K = \Sigma_{xx}^{-1/2} \Sigma_{XY} \Sigma_{YY}^{-1/2} = \Gamma \Lambda \Delta^T$ by singular value decompositiopn.

(3) Given J data matrices,

$$X_j = \begin{pmatrix} X_{1j} \\ \vdots \\ X_{n_j j} \end{pmatrix}, n_j \times p, j = 1, ..., J$$

Find $\tilde{a} \in \mathbb{R}^P$ to maximize the "separation". Fisher discrimination! Define the matrices B and W sized $p \times p$ while

$$B = \sum_{j=1}^{J} n_j(\tilde{x}_j - \tilde{x})(\tilde{x}_j - \bar{x})^T$$

then it goes

$$W = \sum_{j=1}^{J}\sum_{i=1}^{n}(\tilde{x}_{ij} - \tilde{x})(\tilde{X}_{ij} - \tilde{X})^T = \sum_{j=1}^{J}(n_j - 1)S_j = \sum_{j=1}^{J} X_j^T H X_j$$

The maximal separation by one-dimensional linear transformation. is achieved by $\tilde{Y}_j = X_j a$ for $a \in \mathbb{R}^P$ that maximizes

$$\frac{a^T B a}{a^T W a}$$

By a theorem from linear algebra the a that maximizes.

8.2 Discriminant Analysis

Consider J distinct population and we have Pop 1, Pop 2, ..., Pop J with density $f_1, ..., f_J$, respectively. The f_j are densities on $\mathbb{R}^P$. Observe $X = x$ drawn from one of the population's. But which one? The posterior probability of population j is

$$\mathbb{P}(\text{Pop } j | X = x) = \frac{}{\pi_j} f_j(x)\pi_1 f_1(x) + \cdots + \pi_j f_j(x)$$

A classification rule partition $\mathbb{R}^P$ into $R_1, R_2, ..., R_j$ so $R_j = \{s : X = x \text{ is allocated to population } j\}$. All we have for allocation rule is the observed values. The rule defines what we would allocate to which population.

Theorem 8.2.1. *The Bayes' Discriminant Rule Consider* $R_j = \{x : \pi_j f_j(x) > \pi_i f_i(x)$ *for* $j \neq j\} = \{x :$ *Pop jhashighestposteriorprobability*$\}$. *Let* $C(i|j)$ *to be cost of allocating object to population* i *when in fact it is from population* j. *Assume* $C(j|j) = 0$ *and* $C(i|j) > 0$ *for* $i \neq j$.

$$EC_i(x) = \sum_{k \neq i} \pi_k f_k(x)$$

Allocate $X = x$ to the population with the lowest expected cost. For $j = z$, this simplifies.

$$R_1 = \{x : \pi_2 f_2(x) C(1/2) \leq \pi_1 f_1(x) C(2/1)\}$$
$$R_2 = \{x : \pi_1 f_1(x) C(2/1) < \pi_2 f_2(\tilde{x}) C(1/2)\}$$
$$R_1 = \{x : \frac{f_1(x)}{f_2(x)} \geq \frac{\pi_2 C(1/2)}{\pi_1 C(2/1)}\} = \{x : \log[\frac{f_1}{f_2} \geq \log[\frac{\pi_2 C(1/2)}{\pi_1 c(2/1)}\}$$

Suppose population $1 \sim \mathcal{N}_p(\mu_1, \Sigma)$ and population $2 \sim \mathcal{N}_p(\mu_2, \Sigma)$. The expected cost minimizing rule. $R_1 = \{x : (\mu_1 - \mu_2)^T \Sigma^{-1} X\} \frac{1}{2}(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2)$. If population $1 \sim \mathcal{N}_p(\mu, \Sigma)$ and population $2 \sim \mathcal{N}_p(\mu_2, \Sigma_2)$, then $R_1 = \{x : \frac{1}{2} x^T(\Sigma_2^{-1} - \Sigma_1^{-1})x + (\Sigma_1^{-1}\mu + \Sigma_2^{-1}\mu_2)X\} +$

Remark 8.2.2. Midterm 2:
Ch 2-6 are background
Ch 7, 11, 16, 12, 14 (on Discriminant), and 13 (on Clustering)

Remark 8.2.3. Take home exam: soon
Last day of class: April 26th; the last day of the semester is the following Monday
Due: by the end of Tuesday, May 8th

9 CLUSTER ANALYSIS

In Discriminant Analysis (Ch 14) we have sample data for j distinct populations. Construct methods to allocate future observations to those populations. We know how populations there are we know there prior probabilities we have data from each population, have can estimate within group distributions. In clustering, it is difficult. We just have data $X_1, ..., X_n \in \mathbb{R}^n$. This is

$$X_i = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{ip} \end{pmatrix}$$

for $i = 1, ..., n$. We do not know anything about the within group distribution. We do not know within group proportion of cases belonging to each group. The inputs required for a clustering algorithm are: the dissimilarities between every pair of observations. The inputs required for a clustering algorithm are X_i and X_j.

Definition 9.0.1.

$$d_{ij} = d(X_i, X_j)$$

for $i, j = 1, ..., n$.

Generally d is a distance which means

$$d(x, x) = 0, d(x, y) \geq 0, d(x, y) = d(y, x)$$

and we do not even know the number of groups but we still want to group the data. Then

$$d(x, y) \leq d(x, z) + d(y, z)$$

Different clustering algorithm use different distances. Euclidean distance is highly scale dependent

$$d(x, y) = \sqrt{\sum_{k=1}^{\infty} \left(\frac{x_k y_k}{s_k} \right)^2}$$

where s_k a scale estimate for kth component.

Scaled Euclidean distance

$$d(x, y) = \sum_{k=1}^{p} \sum_{k=1}^{\infty} |x_k - y_k|$$

gives Manhattan distance city blocks. Then

$$\sum_{k=1}^{p} \frac{|x_k - y_k|}{s_k}$$

gives scaled Manhattan distance and

$$d(x, y) = \max\{|x_k - y_k| : k = 1, ..., p\}$$

gives maximal distance.

Standardized maximal distance:

$$d(x, y) = \max\{\frac{|x_k y_k|}{s_k}\}, k = 1, ..., p$$

of course these are all special cases of Minkowski matrix. Then

$$d(x,y) = \left(\sum_{k=1}^{p} |x_k - y_k| \right)^2$$

which can also be scaled $r = 1$ Manhattan and $r = 2$ Euclidean while $r \to \infty$ maximal. The generalized distance or statistical distance is

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

In most clustering problems, there does not exist a candidate for the S (scale) matrix. The sample correlation matrix will not match within-group correlations.

Take

$$S = \begin{pmatrix} S_1^2 & & & \\ & S_2^2 & & \\ & & \ddots & \\ & & & S_p^2 \end{pmatrix}$$

to adjust for scale only.

9.1 Hierarchical Clustering

Hierarchical Clustering and more specifically we call agglomerative linkage methods.

(1) Start with n clusters of one point each and $n \times n$ of distances;

(2) Merge the two closest two distances (most similar);

(3) Update the distance matrix; (a) Remove rows and columns corresponding to the clusters that merged; (b) Add a new row or column for that new cluster from the matrix.

(4) Repeat step (2) and (3) $n - 1$ times until there is only one cluster.

Methods will differ in how they accomplish Step 3b. How to calculate distance between new (merged) cluster and all the other clusters? There are four methods accordingly: (1) single linkage, (2) complete linkage, (3) average linkage, and (4) wards method.

Two (not obviously) equivalent ways this can be expressed.

(1) Book. Let P, Q, and R be clusters. Suppose P and A are merged. Then

$$D_{SL}(P+Q, R) = \min\{D(P,R), D(Q,R)\}$$

and

$$D_{CL}(P+Q, R) = \min\{D(P,R), D(Q,R)\}$$

AL and wards method can also be expressed this way, not illuminating

(2) If A and B are two clusters, define $D(A,B)$ in terms of the d_{ij}, $i \in A$, $j \in B$. Then

$$D_{SL}(A,B) = \min\{d_{ij} : i \in A, j \in B\}$$
$$D_{CL}(A,B) = \max\{d_{ij} : i \in A, j \in B\}$$
$$D_{AL}(A,B) = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{i \in B} d_{ij}$$

Note each of these linkage methods works with any of the distances matrix before. Wards method uses a dissimilarity that is not technically a distance (violates triangle inequality). Wards dissimilarity between two clusters is the increase in within-cluster sum of squared distances from cluster means.

Example 9.1.1. Let us solve the example in page 394 in text [1]. This gives us a dendrogram with vertical position represents the distance between the merged clusters.

Step 1. $3 \to 5$
Step 2. $4 \to \{3, 5\}$
Step 3. $7 \to 8$.
Step 4. $6 \to \{7, 8\}$
Step 5. $1 \to 2$
Step 6. $\{3, 4, 5\} \Rightarrow \{6, 7, 8\}$
Step 7. $\{1, 2\} \to \{3, 4, 5, 6, 7, 8\}$
That is, for two clusters, they are

$$\{1, 2\} \text{ and } \{3, 4, 5, 6, 7, 8\}$$

For three clusters, they are

$$\{1, 2\}, \{3, 4, 5\}, \text{ and }, \{6, 7, 8\}$$

For 4 clusters, we have

$$\{1\}, \{2\}, \{3, 4, 5\}, \{6, 7, 8\}$$

This is not discussed in text.

A popular nonhierarchical (partitioning), clustering algorithm is the k-means algorithm (not in the book). Specify how many clusters you want in advance. And an initial clustering (or more common rules for determining one). Inputs. The data $X_1, ..., X_n \in \mathbb{R}^p$ and an initial clustering of k clusters $\{A_1, A_2, ..., A_k\}$ with partition of $\{1, 2, ..., n\}$. Let $n_i = \#$ of objects in cluster A_i and $n_1 + n_2 + \ldots n_k = n$ and then

$$\bar{X}_i = \frac{1}{n_i} \sum_{j \in A_i} X_j$$

the centroid of cluster A_i. The k-means algorithm iteratively adjusts the clustering to minimize an objective function, commonly,

$$\sum_{i=1} k \sum_{j \in A_i} (X_j - X_i)^T (X_j - X_i), \text{ write it as } \star$$

Given an initial clustering $\{A_1, A_2, ..., A_k\}$, for $i = 1, 2, ..., n$: move object x_i to the cluster for which $\star$ is minimized. (may result in staying put). Then repeat the above, until you can cycle through with no objects being moved.

Remark 9.1.2. Discuss:

(1) The 2nd in-class exam is Thur, April 26. Room + seat assignments will be communicated to you by coursework/email.

(2) Chapters 2-6: background only

(3) Chapters 7, 11, 16, 12, 13, 14: just read textbook

(4) Couple more handouts coming supporting Chapters 14 + 13. Handouts exist to demo the computing, may be useful study tool.

(5) As last time, one sheet of original hand-written notes (both sides) and calculator.

(6) Take-home final will be assigned on the last day of class April 30, and due by May 8th.

(7) Take-home final: it will essentially be a 7th assignment. You are expected to work entirely on your own. Students who turn in similar answers will get a zero!

(8) Allowed materials: textbook, course handouts/lecture notes/solutions, R Help. Sort of Allowed: any published work other than the H+S text. Any course material from a course other than this one. You must cite these in your answer. Disallowed: any person other than instructor. Any on-line searches.

(9) No class on Tuesday April 24. Study for the exam. I will be here. Have papers to return and I'll take questions. OH.

10 TEST PROBLEMS

10.1 Problem 1 - Hypothesis Testing

Example 10.1.1. Test Problem 1. Hypothesis testing. We want to test $H_0 : \tilde{\mu} = \tilde{\mu}_0$ with Σ known. This is only of academic interest.

Example 10.1.2. Test Problem 2. We want to test $H_0 : \tilde{\mu} = \tilde{\mu}_0$. We know this one. We want

$$T^2 = n(\bar{\tilde{X}} - \tilde{\mu}_0)^T S^{-1}(\bar{\tilde{X}} - \tilde{\mu}_0) \sim T^2(p, n-1) = \frac{(n-1)p}{n-p} F_{p,n-p}$$

and we have

$$p - val = \mathbb{P}(F_{p,n-p} > \frac{(n-p)}{(n-1)p} T^2$$

Thus, we reject at level α if and only if $p-val < \alpha$ which means that $T^2 > \frac{(n-1)p}{n-p} F_{1-\alpha;p,n-p}$.

Notice that $1-\alpha$ confidence region for $\tilde{\mu}$ is

$$\tilde{\mu}_0 : \{H_0 : \tilde{\mu} = \tilde{\mu}_0 \text{ is accepted at level } \alpha\}$$

$$= \{\tilde{\mu} : n(\bar{\tilde{\mu}} - \tilde{\mu})^T S^{-1}(\bar{\tilde{X}} - \tilde{\mu} \leq \frac{(n-1)p}{n-p} F_{1-\alpha;p,n-p}\}$$

Example 10.1.3. Test Problem 3. Test $H_0 : \Sigma = \Sigma_0$.

10.2 Problem 2 - Regression

Example 10.2.1. Test Problem 4. Let $Y_i \sim \text{indep}\mathcal{N}(\tilde{X}_i^T, \tilde{\beta}, \sigma^2)$ for $i = 1, ..., n$ while $\tilde{x}_i \in \mathbb{R}^p$ are known and $\tilde{\beta} \in \mathbb{R}^p$ is unknown, and σ^2 is unknown. Often times we have $\tilde{x}_{i1} = 1$ for each i but not necessarily. Let $\tilde{Y} = (Y_1, ..., Y_n)^T$ and $X = (\tilde{x}_1^T, ..., \tilde{x}_p^T)^T$ with model to be

$$\tilde{Y} \sim \mathcal{N}_n(X\tilde{\beta}, \sigma^2 I_n)$$

We want to test $H_0 : \tilde{\beta} = \tilde{\beta}_0$. Assume X to be $n \times p$ with rank p. Then

$$\hat{\tilde{\beta}} = (X^T X)^{-1} X^T y, \hat{\sigma}^2 = \frac{1}{n-p} ||\tilde{y} - X\hat{\tilde{\beta}}||^2 = \frac{1}{n-p} (\tilde{y} - X\hat{\tilde{\beta}})^T (\tilde{y} - X\hat{\tilde{\beta}})$$

while

$$\hat{\tilde{\beta}} \sim \mathcal{N}_n(\tilde{\beta}, \sigma^2 (X^T X)^{-1})$$

$$\hat{\sigma}^2 \frac{1}{n-p} ||y - X\hat{\tilde{\beta}}||^2 = \frac{1}{n-p} (\tilde{y} - X\hat{\tilde{\beta}})^T (\tilde{y} - X\hat{\tilde{\beta}})$$

$$\Rightarrow (X^T X)^{1/2} (\hat{\tilde{\beta}} - \tilde{\beta}) \sim \mathcal{N}_p(\tilde{0}, \sigma^2 I)$$

and this gives me

$$(\hat{\tilde{\beta}} - \tilde{\beta})^T X^T X (\hat{\tilde{\beta}} - \tilde{\beta}) \sim \sigma^2 \chi_p^2$$

which implies

$$\Rightarrow (n-p)\hat{\sigma}^2 \sim \sigma^2 \chi_{n-p}^2$$

are independent. Thus, this gives us our F-stat, which goes the following

$$(X^T X)^{1/2} (\hat{\tilde{\beta}} - \tilde{\beta}) \sim \mathcal{N}_p(\tilde{0}, \sigma^2 I_p)$$

$$\Rightarrow \frac{(\hat{\tilde{\beta}} - \tilde{\beta})^T X^T X (\hat{\tilde{\beta}} - \tilde{\beta})}{p\hat{\sigma}^2} \sim F_{p,n-p}$$

To test $H_0 : \tilde{\beta} = \tilde{\beta}_0$, we have

$$\text{F-stat } = \frac{(\hat{\tilde{\beta}} - \tilde{\beta}_0)^T X^T X (\hat{\tilde{\beta}} - \tilde{\beta}_0)}{p\hat{\sigma}^2}, \text{ [1]}$$

so that we have $p{-}val = \mathbb{P}(F_{p,n-p} > F{-}stat)$. That is, $\{\tilde{\beta}_0 : H_0 : \tilde{\beta}{-}\tilde{\beta}_0 \text{ is accepted at level } \alpha\} = \{\tilde{\beta} : (\hat{\tilde{\beta}} - \tilde{\beta})^T X^T X (\hat{\tilde{\beta}} - \tilde{\beta}) \leq p\hat{\sigma}^2 F_{1-\alpha;p,n-p}\}$.

More we can discuss: H_0 : Reduced model or alternatively H_a : full model by using

$$\text{F-stat } = \frac{(SS_R - SS_F)/(df_R - df_F)}{SS_f/df_F}$$

under H_0 we have $F_{df_R - df_F, df_F}$.

Alternatively, we can say reject H_0 if p-val $< \alpha$ if and only if F-stat $> F_{1-\alpha;p,n-p}$. Reduced model is $\tilde{\beta} = \beta_0$, full model is $\tilde{\beta} \in \mathbb{R}^p$. Then we have

$$SS_F = ||y - X\hat{\tilde{\beta}}||^2, df_F = n - p$$

$$SS_R = ||y - X\tilde{X}_0||^2, df_R = n$$

A $1 - \alpha$ confidence region for $\tilde{\beta}$ is

$$\{\tilde{\beta} : (\hat{\tilde{\beta}} - \tilde{\beta})^T X^T X (\hat{\tilde{\beta}} - \tilde{\beta}) \leq p\hat{\sigma}^2 F\}$$

Recall general form for testing, then we have

$$\text{F-stat} = \frac{n-p}{n}\left(\frac{||y - X\beta||^2}{||y - X\hat{\tilde{\beta}}||^2} - 1\right)$$

Remark 10.2.2. Prove this is equivalent to

$$\text{F-stat} = \frac{(\hat{\tilde{\beta}} - \tilde{\beta}_0)^2 X^T X (\hat{\tilde{\beta}} - \tilde{\beta}_0)}{p\hat{\sigma}^2}$$

Example 10.2.3. Test Problem 5. $H_0 : A\tilde{\mu} = \tilde{a}$ with Σ known.

Example 10.2.4. Test Problem 6. $H_0 : A\tilde{\mu} = \tilde{a}$ where A is $q \times p$ rank q. Consider

$$\bar{\tilde{X}} \sim \mathcal{N}_p(\tilde{\mu}, \frac{1}{n}\Sigma)$$

$$(n-1)S_n \sim W_p(\Sigma, n-1)$$

and then we have

$$A\bar{\tilde{X}} \sim \mathcal{N}_q(A\tilde{\mu}, \frac{1}{n}A\Sigma A^T)$$

$$(n-1)ASA^T \sim W_q(A\Sigma A^T, n-1)$$

Consider the general rule to construct independent r.v.

$$[\mathcal{N}_p(\tilde{0}, \Sigma)]^T \left[\frac{W_p(\Sigma, df)}{df}\right]^{-1} \mathcal{N}_p(\tilde{0}, \Sigma) \sim T^2(p, df)$$

Then

$$n(A\bar{\tilde{X}} - \tilde{a})^T (ASA^T)^{-1} (A\bar{\tilde{X}} - \tilde{a}) \sim T^2(q, n-1) = \frac{(n-1)q}{n-q} F_{q,n-q}$$

use this result to test $H_0 : A\tilde{\mu} = \tilde{a}$. The exercise is to derive the confidence region for $A\tilde{\mu}$.

This is p-dimension $X_1, ..., X_n \sim \text{iid}\mathcal{N}_p(\mu, \Sigma)$. We want to test

$$H_0 : Au = a$$

while A is $q \times p$ rank q and we have X and S are the sample mean and covariance matrix. We have $AX \sim \mathcal{N}_p(A\mu, \frac{1}{n}A\Sigma A^T)$. Then we have $(n-1)ASA^T \sim W_q(A\Sigma A^T, n-1)$, which is independent of the above. Under null, $H_0 : A\mu = a$. Then $n(AX - a)^T(ASA)^{-1}(AX - a) = T^2$. Then $T^2(q, n-1) = \frac{(n-1)q}{n-q}F_{q,n-q}$. Thus we have

$$p - val = \mathbb{P}\bigg(F_{a,n-q} > \frac{n-q}{(n-1)q}T^2\bigg)$$

and we conclude that we reject H_0 at level α if $p - val < \alpha$ if and only reject if $T^2 > \frac{(n-1)q}{n-q}F_{1-\alpha;q,n-q}$.

Connected to above example, we have

Example 10.2.5. Let X_{ij} be vocab score of student i in grade 7 + j (same students) for $j = 1, 2, 3, 4$. That is,

$$X_i^T = (X_{i1}, X_{i2}, X_{i3}, X_{i4}), \mu = (\mu_1, \mu_2, \mu_3, \mu_4)$$

Then we have null

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

which is a special case of linear transformation. We can write

$$H_0 : c\mu = 0$$

where

$$C = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

also we can attempt

$$C = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix}$$

or

$$C = \begin{pmatrix} -1 & 0 & 0 & 1 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

or

$$C = \begin{pmatrix} -1 & 1/3 & 1/3 & 1/3 \\ -1/3 & -1/3 & -1/3 & 1 \\ 0 & -1 & 0 & 1 \end{pmatrix}$$

To test $H_0 := \mu_1 = \mu_2 = \cdots = \mu_p$, it is as if we write $H_0 : c\mu = 0$ where C is $(p-1) \times p$ and $C\mathbf{1} = \tilde{0}$. Comment on notation letter C chosen for contrast. Then

$$CX \sim \mathcal{N}_{p-1}(C\mu, \frac{1}{n}C\Sigma C^T)$$

$$(n-1)CSC^T \sim W_{p-1}(C\Sigma C^T, n-1)$$

of which are independent. Under null $H_0 : c\mu = 0$ we have

$$T^2 = nX^TC^T(CSC^T)^{-1}CX$$

$$\sim T^2(p-1, n-1) = \frac{(n-1)(p-1)}{n-p+1}F_{p-1,n-p+1}$$

We conclude that we reject null H_0 if

$$\frac{n-p+1}{(n-1)(p-1)}T^2 > T_{1-\alpha,p-1,n-p+1}$$

Confidence interval: an individual $1-\alpha$ C.I. for $a^T\mu$ is

$$a^T X \pm t_{1-\alpha/2;n-1}\sqrt{\frac{1}{n}a^T S a}$$

or

$$a^T X \pm \sqrt{\frac{1}{n}a^T S a}$$

while simultaneous $1-\alpha$ C.I. for all $a^T u$ and

$$a^T X \pm \sqrt{\frac{(n-1)p}{n-p}F_{1-\alpha;p,n-p}}\sqrt{\frac{1}{n}a^T S a}\ \star$$

Essentially simultaneous confidence is gotten by making intervals wider by factor of $\sqrt{p}$.

Back to repeated measures.Simultaneous $1-\alpha$ C.I.s for all $a^T C\mu$ where $C1_p = \tilde{0}_{p-1}$. are

$$a^T Cx \pm \sqrt{\frac{(n-1)(p-1)}{n-p+1}F_{1-\alpha;p-1,n-p+1}}\sqrt{\frac{1}{n}a^T CSC^T a}\ \star\star$$

We have simultaneous CIs for all $a^T C\mu$ if and only if simultaneous confidence intervals for all $b^T\mu$ satisfying $c^T b = b$ if and only if simultaneous confidence intervals for all $b^T\mu$ satisfying $b^T 1 = 0$.

Compare $\star$ with $\star\star$. $\star$ is simultaneous for all linear combinations. $\star\star$ is simultaneous for all contrasts.

Example 10.2.6. Test Problem 7. Regression. Set up $Y_i \sim \text{ind}\mathcal{N}(\tilde{X}^T\tilde{\beta}, \sigma^2)$ for $i = 1, ..., n$. Test hypothesis $H_0 : A\tilde{\beta} = \tilde{a}$ where A is $q \times p$ rank q.Then

$$\hat{\tilde{\beta}} \sim \mathcal{N}_p(\tilde{\beta}, \sigma^2(X^T X)^{-1})$$

$$A\hat{\tilde{\beta}} \sim \mathcal{N}_p(A\tilde{\beta}, \sigma^2 A(X^T X)^{-1}, A^T)$$

independent of $(n-p)\hat{\sigma}^2 \sim \sigma^2\chi^2_{n-p}$. Thus, we just check F-stat

$$\text{F-stat} = \frac{(A\hat{\tilde{\beta}} - \tilde{a})^T[A(X^T X)^{-1}A^T]^{-1}(A\hat{\tilde{\beta}} - \tilde{a})}{q\hat{\sigma}^2} \sim F_{q,n-p}$$

A similar form can be multiple linear regression,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq} + \epsilon_q$$

and we want to test $H_0 : \beta_1 = \cdots = \beta_q = 0$ with (β_0 arbitrary). Compute

$$A_{q\times p} = A_{(p-1)\times p} = [\tilde{0}, I_q]$$

Example 10.2.7. Test Problem 8. Two sample problem. That is, comparing two mean vectors. Consider

$$X_{11}, ..., X_{n1} \sim \text{iid}\mathcal{N}_p(\tilde{\mu}_1, \Sigma_1)$$
$$X_{12}, ..., X_{n2} \sim \text{iid}\mathcal{N}_p(\tilde{\mu}_2, \Sigma_2)$$

Then we have

$$\bar{\tilde{X}}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} \tilde{X_i} j$$

$$S_j = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(\tilde{X}_{ij} - \tilde{X}_j)^T$$

for $j = 1, 2$.

Let us assume that $\Sigma_1 = \Sigma_2 = \Sigma$ and estimate Σ by

$$S_p = \frac{1}{n_1 + n_2 - 2}[(n_1 - 1)S_1 + (n_2 - 1)S_2]$$

and independent of that we have $(n_1 + n_2 - 2)S_p \sim W_p(\Sigma, n_1 + n_2 - 2)$, independent. The test statistics for $H_0 : \tilde{\mu}_1 - \tilde{\mu}_2 = \tilde{0}$ is

$$T^2 = (\frac{1}{n_1} + \frac{1}{n_2})^{-1}(\bar{\tilde{X}}_1 - \bar{\tilde{X}}_2)^T S_p^{-1}(\bar{\tilde{X}}_1 - \bar{\tilde{X}}_2) \sim T^2(p, n_1 + n_2 - 2) = \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1}F$$

So the p-value for $H_0 : \tilde{\mu}_1 = \tilde{\mu}_2$ is $\mathbb{P}(F >$ F-stat). Reject at level α if and only if p-value $< \alpha$ if and only if $T^2 > \frac{(n_1+n_2-2)p}{n_1+n_2-p-1}$F-stat. A $1 - \alpha$ confidence region $\tilde{\mu}_1 - \tilde{\mu}_2$ is

$$\{\tilde{\delta} : \text{ accepted } H_0 : \tilde{\mu}_1 - \tilde{\mu} = \delta \text{ at level } \alpha\}$$

$$\leq \{\tilde{\delta} : (\bar{\tilde{x}}_1 - \bar{\tilde{X}}_2 - \tilde{\delta})^T[(\frac{1}{n_1} + \frac{1}{n_2}S_p]^{-1}(\tilde{X}_1 - \tilde{X}_2 - \tilde{\delta}\}$$

$$\leq \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1}\text{F-stat}$$

Let us use T-stat

$$T^2 = (X_1 - X_2)^T[(\frac{1}{n_1} + \frac{1}{n_2}S]^{-1}(X_1 - X_2)$$

$$\sim T^2(p, n_1 + n_2 - 2) = \frac{n_1 + n_2 - 2)p}{n_1 + n_2 - p + 1}F_{p,n_1+n_2-p+1}$$

This example comes along with Test Problem 8. Now consider the samples of repeated measures.

Example 10.2.8. The different population of students. A test of $H_0 : \mu_1 - \mu_2 = 0$. Then we want to test "are profiles identical?"

Another example: a test of $H_0 : c(\mu_1 - \mu_2) = 0$ where $C1 = \tilde{0}_{p-1}$ is testing "are the two profiles parallel?" Parallel in a plot of μ_j as j. Then we have

$$T^2 = (\frac{1}{n_1} + \frac{1}{n_2})^{-1}(X_1 - X_2)^T C^T[CSC^2]^{-1}C(X_1 - X_2)$$

and we reject H_0 if

$$T^2 > \frac{(n_1 + n_2 - 2)(p - 1)}{n_1 + n_2 - p}F_{1-\alpha;p-1,n_1+n_2-p}$$

Exam is next Thursday. We will skip Chapters 8-10. Chapter 11 next class.

Remark 10.2.9. First exam is Thur., March., 8th. Closed book, one sheet (both sides) of notes. We have extra room. Some of you will be assigned seats in 428 Pupin. Exam will cover:

Section 7.1,

Section 7.2, its application of material in 7.1

Inference about mean vectors, yes! Covariance matrices, no. Regression problem, no.

Index

References

[1] Hardle, Wolfgang, Leopold Simar (2015), "Applied Multivariate Statistical Analysis", 4th Edition. *Springer*.

[2] Izenman, Alan Julian (2008), "Modern multivariate statistical techniques : regression, classification, and manifold learning", *Springer*.