Assignment

Aryan Jain - EE22BTECH11011*

Question: Suppose that $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}, \mathbf{Y_1}, \mathbf{Y_2}, \ldots, \mathbf{Y_n}$ are independent and identically distributed random vectors each having $N_p\left(\boldsymbol{\mu}, \Sigma\right)$ distributions, where Σ is non-singular, $p > 1$ and $n > 1$. If $\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X_i}$ and $\mathbf{Y} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Y_i}$, then which one of the following statements is true?

(a) There exists $c > 0$ such that $c(\mathbf{X} - \boldsymbol{\mu})^T\Sigma^{-1}\left(\mathbf{X} - \boldsymbol{\mu}\right)$ has χ^2-distribution with p degrees of freedom.

(b) There exists $c > 0$ such that $c(\mathbf{X} - \mathbf{Y})^T\Sigma^{-1}\left(\mathbf{X} - \mathbf{Y}\right)$ has χ^2-distribution with $(p-1)$ degrees of freedom.

(c) There exists $c > 0$ such that $c\sum_{i=1}^{n}\left(\mathbf{X_i} - \mathbf{X}\right)^T\Sigma^{-1}\left(\mathbf{X_i} - \mathbf{X}\right)$ has χ^2-distribution with p degrees of freedom.

(d) There exists $c > 0$ such that $c\sum_{i=1}^{n}\left(\mathbf{X_i} - \mathbf{Y_i} - \mathbf{X} + \mathbf{Y}\right)^T\Sigma^{-1}\left(\mathbf{X_i} - \mathbf{Y_i} - \mathbf{X} + \mathbf{Y}\right)$ has χ^2-distribution with p degrees of freedom.

GATE ST Paper 2023

Solution:

We are given that,

$$\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}, \mathbf{Y_1}, \mathbf{Y_2}, \ldots, \mathbf{Y_n} \sim N_p\left(\boldsymbol{\mu}, \Sigma\right) \tag{1}$$

Also,

$$\mathbf{X} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X_i} \tag{2}$$

$$\mathbf{Y} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Y_i} \tag{3}$$

The mean of $\mathbf{X}$ is given by:

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{X}} &= E\left(\mathbf{X}\right) && (4)\\ &= \frac{1}{n}\sum_{i=1}^{n}E\left(\mathbf{X_i}\right) && (5)\\ &= \boldsymbol{\mu} && (6) \end{aligned}$$

Similarly,

$$\boldsymbol{\mu}_{\mathbf{Y}} = \boldsymbol{\mu} \tag{7}$$

The covariance of $\mathbf{X}$ is given by:

$$\begin{aligned}
\Sigma_{\mathbf{X}} &= E\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\right] \qquad (8)\\
&= E\left[\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X_i}-\boldsymbol{\mu}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X_i}-\boldsymbol{\mu}\right)^T\right] \qquad (9)\\
&= \frac{1}{n^2}E\left[\sum_{i=1}^{n}(\mathbf{X_i}-\boldsymbol{\mu})(\mathbf{X_i}-\boldsymbol{\mu})^T\right] \qquad (10)\\
&= \frac{1}{n^2}\left[\sum_{i=1}^{n}E\left(\mathbf{X_i}^2+\boldsymbol{\mu}^2-2\boldsymbol{\mu}\mathbf{X_i}\right)\right] \qquad (11)\\
&= \frac{1}{n^2}\left[\sum_{i=1}^{n}E\left(\mathbf{X_i}^2\right)+\sum_{i=1}^{n}E\left(\boldsymbol{\mu}^2\right)-2\boldsymbol{\mu}\sum_{i=1}^{n}E(\mathbf{X_i})\right] \qquad (12)\\
&= \frac{1}{n^2}\left[n\Sigma+n\boldsymbol{\mu}^2+n\boldsymbol{\mu}^2-2\boldsymbol{\mu}^2\right] \quad \left[\because E\left(\mathbf{X_i}^2\right)=\Sigma_{\mathbf{X_i}}+E(\mathbf{X_i})^2\right] \qquad (13)\\
&= \frac{\Sigma}{n} \qquad (14)
\end{aligned}$$

Similarly,

$$\Sigma_{\mathbf{Y}} = \frac{\Sigma}{n} \qquad (15)$$

(a) To check option (A):
let us say,

$$\mathbf{A} = c(\mathbf{X}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}) \qquad (16)$$

$$(17)$$

And,

$$\begin{aligned}
\Sigma^{-1} &= \mathbf{F}^T\mathbf{F} \qquad (18)\\
\mathbf{y} &= \mathbf{F}(\mathbf{X}-\boldsymbol{\mu}) \qquad (19)\\
\implies \mathbf{A} &= c\mathbf{y}^T\overline{\mathbf{y}} \qquad (20)\\
&= c||\mathbf{y}||^2 \qquad (21)
\end{aligned}$$

Equation (21) shows that $\mathbf{A}$ can have χ^2-distribution.
To confirm that we will find the mean and covariance-matrix of $\overline{\mathbf{y}}$.

$$\begin{aligned}
\boldsymbol{\mu}_{\mathbf{y}} &= E(\mathbf{y}) \qquad (22)\\
&= E(\mathbf{F}(\mathbf{X}-\boldsymbol{\mu})) \qquad (23)\\
&= \mathbf{F}[E(\mathbf{X})-E(\boldsymbol{\mu})] \qquad (24)\\
&= \mathbf{F}[\boldsymbol{\mu}-\boldsymbol{\mu}] \quad from(6) \qquad (25)\\
&= 0 \qquad (26)
\end{aligned}$$

And,

$$\begin{aligned}
\Sigma_{\mathbf{y}} &= E\left[\left(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}}\right)\left(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}}\right)^T\right] \qquad (27)\\
&= E\left[(\mathbf{F}(\mathbf{X}-\boldsymbol{\mu}))(\mathbf{F}(\mathbf{X}-\boldsymbol{\mu}))^T\right] \qquad (28)\\
&= E\left[\mathbf{F}(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\mathbf{F}^T\right] \qquad (29)\\
&= \mathbf{F}\left[E\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\right]\right]\mathbf{F}^T \qquad (30)\\
&= \mathbf{F}\Sigma\mathbf{F}^T \qquad (31)
\end{aligned}$$

since,

$$\Sigma^{-1} = \mathbf{F}^T\mathbf{F} \tag{32}$$

$$\Sigma\Sigma^{-1} = \Sigma\mathbf{F}^T\mathbf{F} \tag{33}$$

$$\mathbf{I} = \Sigma\mathbf{F}^T\mathbf{F} \tag{34}$$

$$\mathbf{I}\mathbf{F}^{-1} = \Sigma\mathbf{F}^T \tag{35}$$

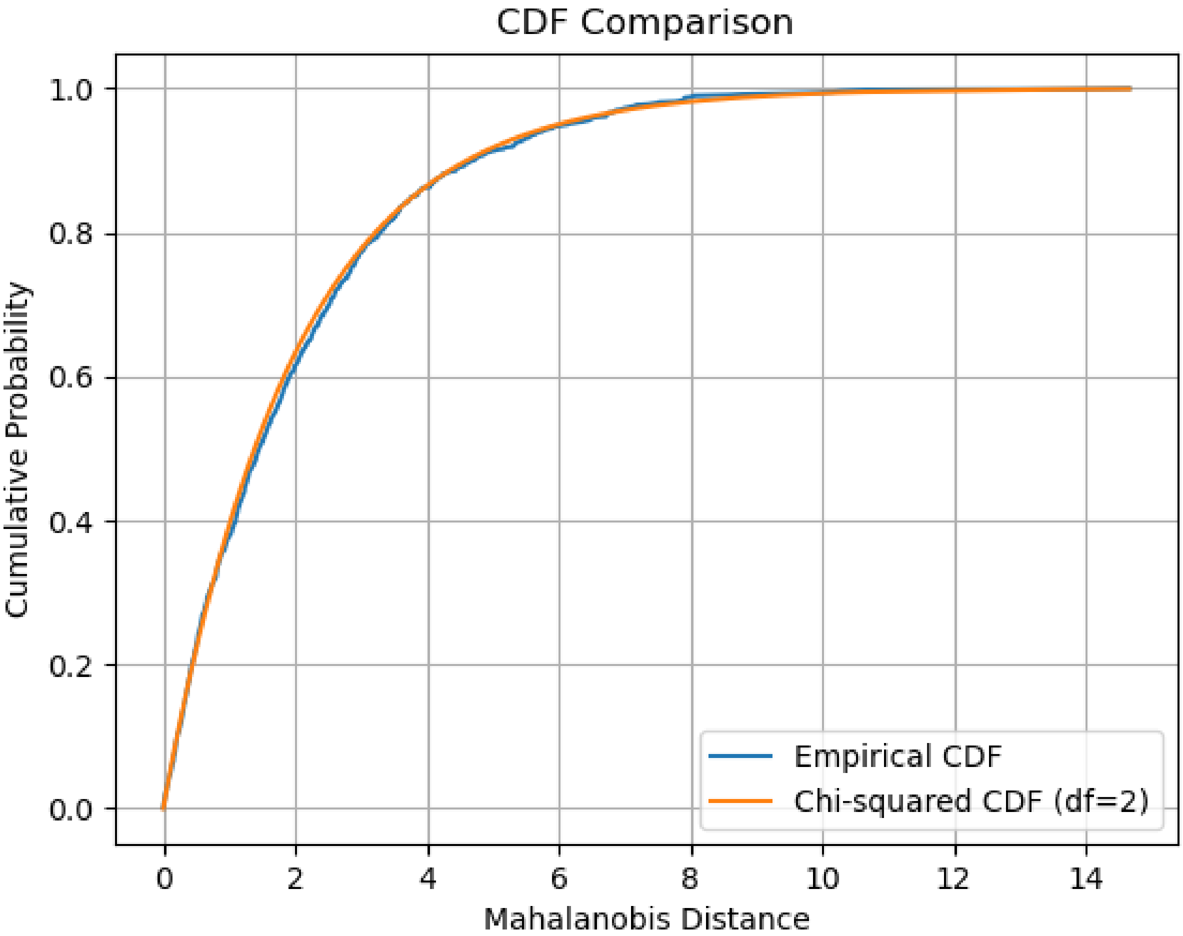
$$\mathbf{F}\mathbf{F}^{-1} = \mathbf{F}\Sigma\mathbf{F}^T \tag{36}$$

$$\mathbf{I} = \mathbf{F}\Sigma\mathbf{F}^T \tag{37}$$

So using (37),

$$\Sigma_{\mathbf{y}} = \mathbf{I} \tag{38}$$

Hence, For $c = 1$ **A** has χ^2-distribution with p degrees of freedom.
So option (A) is correct.



(b) To check option (B):
Let us say,

$$\mathbf{B} = c(\mathbf{X} - \mathbf{Y})^T\Sigma^{-1}(\mathbf{X} - \mathbf{Y}) \tag{39}$$

And,

$$\begin{aligned} \Sigma^{-1} &= \mathbf{F}^T\mathbf{F} && (40)\\ \mathbf{y} &= \mathbf{F}\left(\mathbf{X}-\mathbf{Y}\right) && (41)\\ \implies \mathbf{B} &= c\mathbf{y}^T\overline{\mathbf{y}} && (42)\\ &= c\|\mathbf{y}\|^2 && (43) \end{aligned}$$

Equation (43) shows that **B** can have χ^2-distribution.
To confirm that we will find the mean and covariance-matrix of $\overline{\mathbf{y}}$.

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{y}} &= E\left(\mathbf{y}\right) && (44)\\ &= E\left[F\left(\mathbf{X}-\mathbf{Y}\right)\right] && (45)\\ &= F\left[E\left(\mathbf{X}\right)-E\left(\mathbf{Y}\right)\right] && (46)\\ &= F\left[\boldsymbol{\mu}-\boldsymbol{\mu}\right] && (47)\\ &= 0 && (48) \end{aligned}$$

And,

$$\begin{aligned} \Sigma_{\mathbf{y}} &= E\left[\left(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}}\right)\left(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}}\right)^T\right] && (49)\\ &= E\left[\left(\mathbf{F}\left(\mathbf{X}-\mathbf{Y}\right)\right)\left(\mathbf{F}\left(\mathbf{X}-\mathbf{Y}\right)\right)^T\right] && (50)\\ &= E\left[\mathbf{F}\left(\mathbf{X}-\mathbf{Y}\right)\left(\mathbf{X}-\mathbf{Y}\right)^T\mathbf{F}^T\right] && (51)\\ &= \mathbf{F}\left[E\left[\left(\mathbf{X}-\mathbf{Y}\right)\left(\mathbf{X}-\mathbf{Y}\right)^T\right]\right]\mathbf{F}^T && (52)\\ &= \mathbf{F}\left[E\left[\|\mathbf{X}-\mathbf{Y}\|^2\right]\right]\mathbf{F}^T && (53)\\ &= \mathbf{F}\left[E\left(\mathbf{X}^2\right)+E\left(\mathbf{Y}^2\right)-E\left(2\mathbf{X}\mathbf{Y}\right)\right]\mathbf{F}^T && (54)\\ &= \mathbf{F}\left[\frac{\Sigma}{n}+\boldsymbol{\mu}^2+\frac{\Sigma}{n}+\boldsymbol{\mu}^2-2\boldsymbol{\mu}^2\right]\mathbf{F}^T \quad \left[\because E\left(\mathbf{X}^2\right)=\Sigma_{\mathbf{X}}+E\left(\mathbf{X}\right)^2\right] && (55)\\ &= \frac{2}{n}\mathbf{F}\Sigma\mathbf{F}^T && (56)\\ &= \frac{2}{n}\mathbf{I} && (57) \end{aligned}$$

Hence, for $c = \frac{n}{2}$, **B** has χ^2-distribution with p degrees of freedom.
So option (B) is incorrect.

(c) To check option (C):
let us say,

$$\mathbf{C} = c\sum_{i=1}^{n}\left(\mathbf{X_i}-\mathbf{X}\right)^T\Sigma^{-1}\left(\mathbf{X_i}-\mathbf{X}\right) \tag{58}$$

And,

$$\begin{aligned} \Sigma^{-1} &= \mathbf{F}^T\mathbf{F} && (59)\\ \mathbf{y} &= \mathbf{F}\left(\sum_{i=1}^{n}\left(\mathbf{X_i}-\mathbf{X}\right)\right) && (60)\\ \implies \mathbf{C} &= c\mathbf{y}^T\overline{\mathbf{y}} && (61)\\ &= c\|\mathbf{y}\|^2 && (62) \end{aligned}$$

Equation (62) shows that **C** can have χ^2-distribution.
To confirm that we will find the mean and covariance-matrix of $\overline{\mathbf{y}}$.

$$\boldsymbol{\mu}_{\mathbf{y}} = E\left(\mathbf{y}\right) \tag{63}$$
$$= E\left[\mathbf{F}\left(\sum_{i=1}^{n}\left(\mathbf{X_i} - \mathbf{X}\right)\right)\right] \tag{64}$$
$$= \mathbf{F}\left[\sum_{i=1}^{n}\left(E\left(\mathbf{X_i}\right) - E\left(\mathbf{X}\right)\right)\right] \tag{65}$$
$$= \mathbf{F}\left[E\left(X_1\right) - E\left(X\right) + E\left(X_2\right) - E\left(X\right) + \ldots + E\left(X_n\right) - E\left(X\right)\right] \tag{66}$$
$$= 0 \tag{67}$$

And,

$$\Sigma_{\mathbf{y}} = E\left[\left(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}\right)\left(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}\right)^T\right] \tag{68}$$
$$= \mathbf{F}E\left[\left(\sum_{i=1}^{n}\left(\mathbf{X_i} - \mathbf{X}\right)\right)\left(\sum_{i=1}^{n}\left(\mathbf{X_i} - \mathbf{X}\right)\right)^T\right]\mathbf{F}^T \tag{69}$$
$$= \mathbf{F}E\left[\left(\sum_{i=1}^{n}\mathbf{X_i} - n\mathbf{X}\right)\left(\sum_{i=1}^{n}\mathbf{X_i} - n\mathbf{X}\right)^T\right]\mathbf{F}^T \tag{70}$$
$$= \mathbf{F}E\left[\left(n\mathbf{X} - n\mathbf{X}\right)\left(n\mathbf{X} - n\mathbf{X}\right)^T\right]\mathbf{F}^T \tag{71}$$
$$= \mathbf{F}\mathbf{0}\mathbf{F}^T \tag{72}$$
$$= \mathbf{0} \tag{73}$$

Hence, There is no value of $c > 0$ for which **C** have χ^2-distribution.
So option (C) is incorrect.

(d) To check option (D):
let us say,

$$\mathbf{D} = c\sum_{i=1}^{n}\left(\mathbf{X_i} - \mathbf{Y_i} - \mathbf{X} + \mathbf{Y}\right)^T\Sigma^{-1}\left(\mathbf{X_i} - \mathbf{Y_i} - \mathbf{X} + \mathbf{Y}\right) \tag{74}$$

And,

$$\Sigma^{-1} = \mathbf{F}^T\mathbf{F} \tag{75}$$
$$\mathbf{y} = \mathbf{F}\left(\sum_{i=1}^{n}\left(\mathbf{X_i} - \mathbf{Y_i} - \mathbf{X} + \mathbf{Y}\right)\right) \tag{76}$$
$$\implies \mathbf{C} = c\mathbf{y}^T\overline{\mathbf{y}} \tag{77}$$
$$= c||\mathbf{y}||^2 \tag{78}$$

Equation (78) shows that **D** can have χ^2-distribution.
To confirm that we will find the mean and covariance-matrix of $\overline{\mathbf{y}}$.

$$\begin{aligned}
\boldsymbol{\mu}_{\mathbf{y}} &= E\left(\mathbf{y}\right) && (79)\\
&= E\left(\mathbf{F}\left(\sum_{i=1}^{n}\left(\mathbf{X_i}-\mathbf{Y_i}-\mathbf{X}+\mathbf{Y}\right)\right)\right) && (80)\\
&= \mathbf{F}E\left[\sum_{i=1}^{n}\mathbf{X_i}-\sum_{i=1}^{n}\mathbf{Y_i}-n\mathbf{X}+n\mathbf{Y}\right] && (81)\\
&= \mathbf{F}\left[\sum_{i=1}^{n}E\left(\mathbf{X_i}\right)-\sum_{i=1}^{n}E\left(\mathbf{Y_i}\right)-nE\left(\mathbf{X}\right)+nE\left(\mathbf{Y}\right)\right] && (82)\\
&= \mathbf{F}\left[n\boldsymbol{\mu}-n\boldsymbol{\mu}-n\boldsymbol{\mu}+n\boldsymbol{\mu}\right] && (83)\\
&= 0 && (84)
\end{aligned}$$

And,

$$\begin{aligned}
\Sigma_{\mathbf{y}} &= E\left[\left(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}}\right)\left(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}}\right)^{T}\right] && (85)\\
&= \mathbf{F}E\left[\left(\sum_{i=1}^{n}\left(\mathbf{X_i}-\mathbf{Y_i}-\mathbf{X}+\mathbf{Y}\right)\right)\left(\sum_{i=1}^{n}\left(\mathbf{X_i}-\mathbf{Y_i}-\mathbf{X}+\mathbf{Y}\right)\right)^{T}\right]\mathbf{F}^{T} && (86)\\
&= \mathbf{F}E\left[\left(\sum_{i=1}^{n}\mathbf{X_i}-\sum_{i=1}^{n}\mathbf{Y_i}-n\mathbf{X}+n\mathbf{Y}\right)\left(\sum_{i=1}^{n}\mathbf{X_i}-\sum_{i=1}^{n}\mathbf{Y_i}-n\mathbf{X}+n\mathbf{Y}\right)^{T}\right]\mathbf{F}^{T} && (87)\\
&= \mathbf{F}E\left[\left(n\mathbf{X}-n\mathbf{Y}-n\mathbf{X}+n\mathbf{Y}\right)\left(n\mathbf{X}-n\mathbf{Y}-n\mathbf{X}+n\mathbf{Y}\right)^{T}\right]\mathbf{F}^{T} && (88)\\
&= \mathbf{F0F}^{T} && (89)\\
&= \mathbf{0} && (90)
\end{aligned}$$

Hence, There is no value of $c > 0$ for which **D** have χ^2-distribution.
So option (D) is incorrect.

Mahalanobis Distance:

It is the measure of the distance between a point **X** and a distribution Q. It is a multi-dimensional generalization of the idea of measuring how many standard deviations away **X** is from the mean of Q. So, Given a probability distribution Q on R^N with,

$$\boldsymbol{\mu} = \left(\mu_1, \mu_2, \cdots, \mu_N\right)^T \tag{91}$$

and positive covariance matrix Σ, the mahalanobis distance of a point,

$$\mathbf{X} = \left(X_1, X_2, \cdots, X_N\right)^T \tag{92}$$

is given by,

$$d_M\left(\mathbf{X}, Q\right) = \sqrt{\left(\mathbf{X}-\boldsymbol{\mu}\right)^T\Sigma^{-1}\left(\mathbf{X}-\boldsymbol{\mu}\right)} \tag{93}$$

Steps for simulation:

1) Firstly in the the file ”gauss.c”, I have generated 1000 random vectors with dimension 2 using Box-Muller method and listed the data in the file ”randomvectors.dat”.
2) Then in the file ”distance.c”, using the random vectors generated in the first step, I found the value of $c(\mathbf{X}-\boldsymbol{\mu})^T\Sigma^{-1}\left(\mathbf{X}-\boldsymbol{\mu}\right)$ distribution which will give us 1×1 matrix.
3) So as we have generated 1000 random vectors in first step, we will have 1000 values of the distribution.

4) Then I have listed the values of the distribution $c(\mathbf{X}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu})$ that I got in the file ”mahalanobisdistances.dat”.
5) Now in the file ”cdf.py”, I have plotted the cdf of the distribution $c(\mathbf{X}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu})$ and also plotted the theoretical cdf plot of a χ^2 distribution.

The variables that are used in the simulation are:

Variable	Definition
$\mathbf{X}$	random vector
p	dimension of vector
n	number of vectors
$\boldsymbol{\mu}$	mean
$\boldsymbol{\Sigma}$	Covariance