

STAT 3690 Lecture Note

Week Three (Jan 23, 25, & 27, 2023)

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca, zhiyanggeezhou.github.io)

2023/Jan/24 22:46:36

Statistical modelling (con'd)

Transformation of random vectors

- Derive the pdf of continuous $\mathbf{Y} = \mathbf{g}(\mathbf{X})$ from the pdf of continuous \mathbf{X}
- Prerequisite
 - $\mathbf{X} = [X_1, \dots, X_p]^\top$ and $\mathbf{Y} = [Y_1, \dots, Y_p]^\top$
 - $\mathbf{g} = (g_1, \dots, g_p): \mathbb{R}^p \rightarrow \mathbb{R}^p$ is a continuous one-to-one map with inverse $\mathbf{g}^{-1} = (h_1, \dots, h_p)$, i.e., $Y_i = g_i(\mathbf{X})$ and $X_i = h_i(\mathbf{Y})$
- Elaborate $\text{supp}(\mathbf{Y}) = \{[y_1, \dots, y_p]^\top : [h_1(y_1, \dots, y_p), \dots, h_p(y_1, \dots, y_p)]^\top \in \text{supp}(\mathbf{X})\}$
- Jacobian matrix of \mathbf{g}^{-1} is $\mathbf{J}_{\mathbf{g}^{-1}} = [\partial x_i / \partial y_j]_{p \times p} = [\partial h_i(y_1, \dots, y_p) / \partial y_j]_{p \times p}$
 - Also, $|\det(\mathbf{J}_{\mathbf{g}^{-1}})| = |\det([\partial y_i / \partial x_j]_{p \times p})|^{-1} = |\det([\partial g_i(x_1, \dots, x_p) / \partial x_j]_{p \times p})|^{-1}$
- Then

$$f_{\mathbf{Y}}(y_1, \dots, y_p) = f_{\mathbf{X}}(h_1(y_1, \dots, y_p), \dots, h_p(y_1, \dots, y_p)) |\det(\mathbf{J}_{\mathbf{g}^{-1}})| \mathbf{1}_{\text{supp}(\mathbf{Y})}(y_1, \dots, y_p)$$

-
- Exercise: Let $\mathbf{X} = [X_1, X_2]^\top$ follow the standard bivariate normal, i.e., its pdf is

$$f_{\mathbf{X}}(x_1, x_2) = (2\pi)^{-1} \exp\{-(x_1^2 + x_2^2)/2\} \mathbf{1}_{\mathbb{R}^2}(x_1, x_2).$$

Find out the joint pdf of $\mathbf{Y} = [Y_1, Y_2]^\top$, where $Y_1 = \sqrt{X_1^2 + X_2^2}$ and $0 \leq Y_2 < 2\pi$ is the angle from the positive x -axis to the ray from the origin to the point (X_1, X_2) , that is, Y is X in the polar coordinate.

-
- Exercise: Given positive α, β and θ , $\mathbf{X} = [X_1, X_2]^\top$ follow

$$f_{\mathbf{X}}(x_1, x_2) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)\theta^{\alpha+\beta}} x_1^{\alpha-1} x_2^{\beta-1} \exp\left(-\frac{x_1 + x_2}{\theta}\right) \mathbf{1}_{\mathbb{R}^+ \times \mathbb{R}^+}(x_1, x_2),$$

where $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$, e.g., $\Gamma(n) = (n-1)!$ for positive integer n . Find out the joint pdf of $\mathbf{Y} = [Y_1, Y_2]^\top$, where $Y_1 = X_1/(X_1 + X_2)$ and $Y_2 = X_1 + X_2$.

Mean matrix

- $E(\mathbf{X}) = [E(X_{ij})]_{n \times p}$, where
 - Random $n \times p$ matrix $\mathbf{X} = [X_{ij}]_{n \times p}$
 - (Linearity) $E(\mathbf{A}\mathbf{X} + \mathbf{B}\mathbf{Y}) = \mathbf{A}E(\mathbf{X}) + \mathbf{B}E(\mathbf{Y})$, where
 - Fixed $\mathbf{A} \in \mathbb{R}^{\ell \times n}$ and $\mathbf{B} \in \mathbb{R}^{\ell \times m}$
 - Random matrices $\mathbf{X} = [X_{ij}]_{n \times p}$ and $\mathbf{Y} = [Y_{ij}]_{m \times p}$
-

Covariance matrix

- Random p -vector $\mathbf{X} = [X_1, \dots, X_p]^\top$ and random q -vector $\mathbf{Y} = [Y_1, \dots, Y_q]^\top$
 - Covariance matrix (defined via expectation) $\Sigma_{\mathbf{XY}} = \text{cov}(\mathbf{X}, \mathbf{Y}) = E[\{\mathbf{X} - E(\mathbf{X})\}\{\mathbf{Y} - E(\mathbf{Y})\}^\top]$
 - Also, $\Sigma_{\mathbf{XY}} = E(\mathbf{XY}^\top) - E(\mathbf{X})E(\mathbf{Y}^\top)$
 - The (i, j) -entry of $\Sigma_{\mathbf{XY}}$ is $\text{cov}(X_i, Y_j)$
 - $\Sigma_{\mathbf{AX} + \mathbf{a}, \mathbf{BY} + \mathbf{b}} = \mathbf{A}\Sigma_{\mathbf{XY}}\mathbf{B}^\top$ for fixed $\mathbf{A} \in \mathbb{R}^{m \times p}$, $\mathbf{a} \in \mathbb{R}^m$, $\mathbf{B} \in \mathbb{R}^{\ell \times q}$ and $\mathbf{b} \in \mathbb{R}^\ell$
 - $\Sigma_{\mathbf{X}} \geq 0$, where $\Sigma_{\mathbf{X}} = \text{cov}(\mathbf{X})$ is short for $\Sigma_{\mathbf{XX}} = \text{cov}(\mathbf{X}, \mathbf{X})$
-

- Exercise: Verify the following properties of covariance matrix
 1. $\Sigma_{\mathbf{AX} + \mathbf{a}, \mathbf{BY} + \mathbf{b}} = \mathbf{A}\Sigma_{\mathbf{XY}}\mathbf{B}^\top$
 2. $\Sigma_{\mathbf{X}} \geq 0$

Sample covariance matrix

- Samples $\mathbf{X}_k = [X_{k1}, \dots, X_{kp}]^\top$ and $\mathbf{Y}_k = [Y_{k1}, \dots, Y_{kq}]^\top$, $k = 1, \dots, n$
- $(\mathbf{X}_k, \mathbf{Y}_k) \stackrel{\text{iid}}{\sim} (\mathbf{X}, \mathbf{Y})$, where $\mathbf{X} = [X_1, \dots, X_p]^\top$ and $\mathbf{Y} = [Y_1, \dots, Y_q]^\top$
- Sample mean vectors
 - $\bar{\mathbf{X}} = n^{-1} \sum_{k=1}^n \mathbf{X}_k = [\bar{X}_{\cdot 1}, \dots, \bar{X}_{\cdot p}]^\top$
 - $\bar{\mathbf{Y}} = n^{-1} \sum_{k=1}^n \mathbf{Y}_k = [\bar{Y}_{\cdot 1}, \dots, \bar{Y}_{\cdot q}]^\top$
- Sample covariance matrix:

$$\mathbf{S}_{\mathbf{XY}} = \frac{1}{n-1} \sum_{k=1}^n \{(\mathbf{X}_k - \bar{\mathbf{X}})(\mathbf{Y}_k - \bar{\mathbf{Y}})^\top\}$$

- The (i, j) -entry of $\mathbf{S}_{\mathbf{XY}}$ is $(n-1)^{-1} \sum_{k=1}^n (X_{ki} - \bar{X}_{\cdot i})(Y_{kj} - \bar{Y}_{\cdot j})$, i.e., the sample covariance between X_i and Y_j
 - Unbiasedness: $E(\mathbf{S}_{\mathbf{XY}}) = \Sigma_{\mathbf{XY}}$
 - $\mathbf{S}_{\mathbf{AX} + \mathbf{a}, \mathbf{BY} + \mathbf{b}} = \mathbf{A}\mathbf{S}_{\mathbf{XY}}\mathbf{B}^\top$ for $\mathbf{A} \in \mathbb{R}^{m \times p}$, $\mathbf{a} \in \mathbb{R}^m$, $\mathbf{B} \in \mathbb{R}^{\ell \times q}$ and $\mathbf{b} \in \mathbb{R}^\ell$
 - $\mathbf{S}_{\mathbf{X}} \geq 0$
 - Implementation in R: `cov()` (or `var()` if $\mathbf{X} = \mathbf{Y}$)
-

- Exercise: Verify the following properties of sample covariance matrix
 1. $E(\mathbf{S}_{\mathbf{XY}}) = \Sigma_{\mathbf{XY}}$
 2. $\mathbf{S}_{\mathbf{AX} + \mathbf{a}, \mathbf{BY} + \mathbf{b}} = \mathbf{A}\mathbf{S}_{\mathbf{XY}}\mathbf{B}^\top$
 3. $\mathbf{S}_{\mathbf{X}} \geq 0$

Computing sample mean vectors and sample covariance matrices via R

Multivariate normal (MVN) distribution (J&W Sec 4.2)

Definition

- Standard MVN
 - $\mathbf{Z} = [Z_1, \dots, Z_p]^\top \sim \text{MVN}_p(\mathbf{0}, \mathbf{I}) \Leftrightarrow Z_1, \dots, Z_p \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$
 - pdf

$$f_{\mathbf{Z}}(\mathbf{z}) = (2\pi)^{-p/2} \exp(-\mathbf{z}^\top \mathbf{z} / 2) \cdot \mathbf{1}_{\mathbb{R}^p}(\mathbf{z})$$
- General MVN
 - $\mathbf{X} = [X_1, \dots, X_p]^\top \sim \text{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Leftrightarrow$ there exists $\boldsymbol{\mu} \in \mathbb{R}^p$, $\mathbf{A} \in \mathbb{R}^{p \times p}$ and $\mathbf{Z} \sim \text{MVN}_p(\mathbf{0}, \mathbf{I})$ such that $\mathbf{X} = \mathbf{A}\mathbf{Z} + \boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\top$
 - * Limited to non-degenerate cases, i.e., invertible \mathbf{A} ($\Leftrightarrow \boldsymbol{\Sigma} > 0$)
 - pdf

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-p/2} (\det \boldsymbol{\Sigma})^{-1/2} \exp\{-(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) / 2\} \cdot \mathbf{1}_{\mathbb{R}^p}(\mathbf{x})$$

-
- Exercise: Density of $\text{MVN}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ evaluated at $(4, 7)$, where

$$\boldsymbol{\mu} = [3, 6]^\top, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 10 & 2 \\ 2 & 5 \end{bmatrix}.$$

Properties of MVN

- \mathbf{X} is of MVN $\Leftrightarrow a^\top \mathbf{X}$ is normally distributed for ALL non-zero $a \in \mathbb{R}^p$.
 - Warning: marginal normals do not imply the joint normal.
- If $\mathbf{X} \sim \text{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{A}\mathbf{X} + \mathbf{b} \sim \text{MVN}_q(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$ for $\mathbf{A} \in \mathbb{R}^{q \times p}$ of full-row-rank. Hence,
 - (Stochastic representation of MVN) if $\mathbf{X} \sim \text{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then there is $\mathbf{Z} \sim \text{MVN}_p(\mathbf{0}, \mathbf{I})$ such that $\mathbf{X} = \boldsymbol{\Sigma}^{1/2} \mathbf{Z} + \boldsymbol{\mu}$. Actually, $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2} (\mathbf{X} - \boldsymbol{\mu})$.
- $(\mathbf{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(p)$ if $\mathbf{X} \sim \text{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

-
- Exercise: Generate six iid samples following bivariate normal $\text{MVN}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = [3, 6]^\top, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 10 & 2 \\ 2 & 5 \end{bmatrix}.$$

- Exercise: Suppose $X_1 \sim \mathcal{N}(0, 1)$. In the following two cases, verify that $X_2 \sim \mathcal{N}(0, 1)$ as well. Does $\mathbf{X} = [X_1, X_2]^\top$ follow an MVN in both cases?
 - $X_2 = -X_1$;
 - $X_2 = (2Y - 1)X_1$, where $Y \sim \text{Ber}(p)$ and $Y \perp\!\!\!\perp \mathbf{X}$. (Hint: $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \Leftrightarrow f_{\mathbf{Z}}(\mathbf{z}) = f_{\mathbf{X}}(\mathbf{x})f_{\mathbf{Y}}(\mathbf{y})$, where $\mathbf{Z} = [\mathbf{X}^\top, \mathbf{Y}^\top]^\top$.)
-

Marginal and conditional MVN

- If $\mathbf{X} \sim \text{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

with $\Sigma_{11} > 0$ and $\Sigma_{22} > 0$, then

- (Marginals of MVN are still MVN) $\mathbf{X}_i \sim \text{MVN}_{p_i}(\boldsymbol{\mu}_i, \Sigma_{ii})$
- (Conditionals of MVN are MVN) $\mathbf{X}_i | \mathbf{X}_j = \mathbf{x}_j \sim \text{MVN}_{p_i}(\boldsymbol{\mu}_{i|j}, \Sigma_{i|j})$
 - * $\boldsymbol{\mu}_{i|j} = \boldsymbol{\mu}_i + \Sigma_{ij}\Sigma_{jj}^{-1}(\mathbf{x}_j - \boldsymbol{\mu}_j)$
 - * $\Sigma_{i|j} = \Sigma_{ii} - \Sigma_{ij}\Sigma_{jj}^{-1}\Sigma_{ji}$
- $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2 \Leftrightarrow \Sigma_{12} = \mathbf{0}$
 - * Warning: the prerequisite for this equivalence is the joint normal of \mathbf{X}_1 and \mathbf{X}_2 .

- Exercise: The argument $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2 \Leftrightarrow \Sigma_{12} = 0$ is based on $\mathbf{X} = [\mathbf{X}_1^\top, \mathbf{X}_2^\top]^\top \sim \text{MVN}$. That is, if \mathbf{X}_1 and \mathbf{X}_2 are both MVN BUT they are not jointly normal, the zero Σ_{12} doesn't suffice for the independence between \mathbf{X}_1 and \mathbf{X}_2 . Recall the instance in the previous exercise: $X_1 \sim \mathcal{N}(0, 1)$ and $X_2 = (2Y - 1)X_1$. Verify that X_1 and X_2 are not independent of each other.

Checking normality (J&W Sec 4.6)

- Checking the univariate marginal distributions
 - Normal Q-Q plot
 - * `qqnorm()`; `car::qqPlot()`
 - Univariate normality test
 - * `shapiro.test()`; `nortest::ad.test()`
- Testing the multivariate normality
 - `MVN::mvn()`
- Checking the quadratic form
 - χ^2 Q-Q plot
 - * $D_i^2 = (\mathbf{X}_i - \bar{\mathbf{X}})^\top \mathbf{S}^{-1}(\mathbf{X}_i - \bar{\mathbf{X}}) \approx \chi^2(p)$ if $\mathbf{X}_i \stackrel{\text{iid}}{\sim} \text{MVN}_p(\boldsymbol{\mu}, \Sigma)$
 - * `qqplot()`; `car::qqPlot()`

Detecting outliers (J&W Sec 4.7)

- Scatter plot of standardized values
- Check the points farthest from the origin in χ^2 Q-Q plot

Improving normality (J&W Sec 4.8)

- (Original) Box-Cox (power) transformation: transform positive x into

$$x^* = \begin{cases} (x^\lambda - 1)/\lambda & \lambda \neq 0 \\ \ln(x) & \lambda = 0 \end{cases}$$

with λ selected with certain criterion

- If $x \leq 0$, change it to be positive first.
- See J. Tukey (1977). *Exploratory Data Analysis*. Boston: Addison-Wesley.
- Multivariate Box-Cox transformation