Multivariate Analysis (MATH5855)

Dr. Atefeh Zamani

Based on notes by **Prof. Spiridon Penev** and **Dr. Pavel Krivitsky**
University of New South Wales
School of Mathematics & Statistics
Department of Statistics

Term 3, 2022

Section 2:
The Multivariate Normal Distribution

Definition, Properties of multivariate normal, Tests for Multivariate Normality, Software, Examples, Additional resources, Exercises

Definition of the Multivariate Normal Distribution

Generalising the Normal Distribution

- Generalisation of the univariate normal for $p \geq 2$ dimensions
- Consider replacing $\left(\frac{x-\mu}{\sigma}\right)^2 = (x-\mu)\left(\sigma^2\right)^{-1}(x-\mu)$ in

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(x-\mu)/\sigma]^2/2}, \quad x \in \mathbb{R} \tag{1}$$

  by $(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})$.
  - $\boldsymbol{\mu} = E(\boldsymbol{x}) \in \mathbb{R}^p$ the expected value of $\boldsymbol{x} \in \mathbb{R}^p$
  - covariance matrix

$$\Sigma = E(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix} \in \mathcal{M}_{p,p} \tag{2}$$

  - diagonal elements of Σ : variances of the p random variables
  - $\sigma_{ij} = E[(X_i - E(X_i))(X_j - E(X_j))]$, $i \neq j$ stands for the covariances between the i-th and j-th random variables
  - $\sigma_{ii} \equiv \sigma_i^2$
- **Only makes sense if Σ is pos. def**

Multivariate Normal Distribution density

If Σ is pos. def then density of $\boldsymbol{X}$ is

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2}, x_i \in \mathbb{R},\ i = 1, \cdots, p. \tag{3}$$

- $E\boldsymbol{X} = \boldsymbol{\mu}$
- $E\left[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\right] = \Sigma$
- Notation: $N_p(\boldsymbol{\mu}, \Sigma)$.

Cramer–Wold argument

- We also want MVN for singular Σ (nonneg. def.)
- Use Cramer–Wold argument:

  The distribution of a p-dimensional random vector $\boldsymbol{X}$ is completely characterised by the one-dimensional distributions of **all** linear transformations $\boldsymbol{t}^T\boldsymbol{X}$, $\boldsymbol{t} \in \mathbb{R}^p$
  i.e., consider $E\left[\exp\{it\left(\boldsymbol{t}^T\boldsymbol{X}\right)\}\right]$, (assumed known for every $t \in \mathbb{R}$ and $\boldsymbol{t} \in \mathbb{R}^p$)

  - Substitute $t = 1$ to get $E\left[\exp\{i\left(\boldsymbol{t}^T\boldsymbol{X}\right)\}\right]$, the cf of the vector $\boldsymbol{X}$.

Definition

The random vector $\boldsymbol{X} \in \mathbb{R}^p$ has a multivariate normal distribution if and only if (iff) any linear transformation $\boldsymbol{t}^T\boldsymbol{X}$, $\boldsymbol{t} \in \mathbb{R}^p$ has a univariate normal distribution.

Lemma

The characteristic function of the (univariate) standard normal random variable $X \sim N(0,1)$ is

$$\varphi_x(t) = e^{-t^2/2}$$

Sketch of the Proof:

- $\varphi_x(t) = E\left(e^{itX}\right) = \int_{-\infty}^{\infty} e^{itX} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$
- Completing the square and factoring,

$$\varphi_x(t) = e^{-t^2/2} \lim_{h\to\infty} \int_{-h+it}^{h+it} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

- Use Cauchy's Theorem and contour integration to show that the complex integral above equals 1.

Aside: We could use the moment generating function (mgf) instead:

$$M_X(t) = E(e^{tX}) = e^{-t^2/2}.$$

Theorem

Suppose that for a random vector $\boldsymbol{X} \in \mathbb{R}^p$ with a multivariate normal distribution, we have $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $D(\boldsymbol{X}) = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T = \Sigma,$ Then:

- *For any fixed $\boldsymbol{t} \in \mathbb{R}^p$, $\boldsymbol{t}^T\boldsymbol{X} \sim N(\boldsymbol{t}^T\boldsymbol{\mu}, \boldsymbol{t}^T\Sigma\boldsymbol{t})$.*
- *The cf of $\boldsymbol{X} \in \mathbb{R}^p$ is*

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = e^{i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\Sigma\boldsymbol{t}} \tag{4}$$

Proof

Part (i) is obvious. For Part (ii),

- cf of the standard univariate normal random variable Z is $e^{-t^2/2}$.
- Any $U \sim N_1(\mu_1, \sigma_1^2)$ has a distribution that coincides with the distribution of $_1 + \sigma_1 Z$.
- Then,

$$\begin{aligned}\varphi_U(t) &= e^{it\mu_1}\varphi_{\sigma_1 Z}(t) = e^{it\mu_1}E(e^{it\sigma_1 Z}) \\ &= e^{it\mu_1} varphi_Z(\sigma_1 t) = e^{it\mu_1 - \frac{1}{2}t^2\sigma_1^2}\end{aligned}$$

- So, for $\boldsymbol{t}^T\boldsymbol{X} \sim N_1(\boldsymbol{t}^T\boldsymbol{\mu}, \boldsymbol{t}^T\Sigma\boldsymbol{t})$ (univariate), cf is $\varphi_{\boldsymbol{t}^T\boldsymbol{X}}(t) = e^{i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}t^2\boldsymbol{t}^T\Sigma\boldsymbol{t}}$

Proof

$\Rightarrow$ Given $\boldsymbol{\mu}$ and Σ use cf formula (4) rather than the density formula (3).

- cf formula defined for singular Σ.
- Still need to show density (3) for invertible Σ.

Theorem

Assume the matrix Σ in (4) is nonsingular. Then the density of the random vector $\boldsymbol{X} \in \mathbb{R}^p$ with cf as in (4) is given by (3).

Proof

- Consider the vector $\boldsymbol{Y} = \Sigma^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}) \in \mathbb{R}^p$.
- $E(\boldsymbol{Y}) = 0$
- $D(\boldsymbol{Y}) = E(\boldsymbol{Y}\boldsymbol{Y}^T) = \Sigma^{-1/2}E((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T)\Sigma^{-1/2} = \boldsymbol{I}$
  $\Rightarrow$ substitute to get the cf of $\boldsymbol{Y}$: $\varphi_{\boldsymbol{Y}}(\boldsymbol{t}) = e^{-\frac{1}{2}\sum_{i=1}^{p} t_i^2}$
  - This is cf of p independent $N(0, 1)$
- $\boldsymbol{Y} = \Sigma^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}) \Rightarrow \boldsymbol{X} = \boldsymbol{\mu} + \Sigma^{1/2}\boldsymbol{Y}$
- Density of $\boldsymbol{Y}$: $f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}\sum_{i=1}^{p} y_i^2}$
- Use density transformation:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{Y}}(\Sigma^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})^T)|J(x_1, \cdots, x_p)|$$

  - By linearity: $|J(x_1, \cdots, x_p)| = |\Sigma^{-1/2}| = |\Sigma^{1/2}|^{-1}$
  - $\sum_{i=1}^{p} y_i^2 = \boldsymbol{y}^T\boldsymbol{y} = (\boldsymbol{X} - \boldsymbol{\mu})^T\Sigma^{-1/2}\Sigma^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}) = (\boldsymbol{X} - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$

$\Rightarrow$ density formula (3) for $f_{\boldsymbol{X}}(\boldsymbol{x})$

Properties of multivariate normal

Property 1

If $\Sigma = D(\boldsymbol{X}) = \Lambda$ is a diagonal matrix then the p components of $\boldsymbol{X}$ are independent.

- i.e., then $\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = e^{i\sum_{j=1}^{p} t_j\mu_j - \frac{1}{2}t_j^2\sigma_j^2}$, decomposes into cf's of p independent components each distributed according to $N(\mu_j, \sigma_j^2)$, $j = 1, \cdots, p$.
- "for a multivariate normal, if its components are uncorrelated they are also independent"
- converse (if independent, then uncorrelated) true for **any distribution**
- For the multivariate normal distribution, we can conclude that its components are independent **if and only if** they are uncorrelated!

Example (Random variables that are marginally normal and uncorrelated but not independent).

Consider two variables $Z_1 = (2W - 1)Y$ and $Z_2 = Y$, where $Y \sim N_1(0, 1)$ and, independently, $W \sim \mathit{Binomial}(1, 1/2)$ (so $2W - 1$ takes -1 and +1 with equal probability).

Property 2

If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ and $C \in \mathcal{M}_{q,p}$ is an arbitrary matrix of real numbers then

$$\boldsymbol{Y} = C\boldsymbol{X} \sim N_q(C\boldsymbol{\mu}, C\Sigma C)$$

- for any $\boldsymbol{s} \in \mathbb{R}^q$,

  $$\varphi_{\boldsymbol{Y}}(\boldsymbol{s}) = \varphi_{\boldsymbol{X}}(C^T\boldsymbol{s}) = e^{i\boldsymbol{s}^T C\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{s}^T C\Sigma C\boldsymbol{s}}$$

  $$\Rightarrow \boldsymbol{Y} =\sim N_q(C\boldsymbol{\mu}, C\Sigma C)$$
- C is full rank and if $rk(\Sigma) = p$ then the rank of $C\Sigma C$ is also full, which means that the distribution of $\boldsymbol{Y}$ would **not** be **degenerate** in this case.

Property 3

Assume the vector $\boldsymbol{X} \in \mathbb{R}^p$ is divided into subvectors $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_{(1)} \\ \boldsymbol{X}_{(2)} \end{pmatrix}$ and according to this subdivision the vector means are $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{(1)} \\ \boldsymbol{\mu}_{(2)} \end{pmatrix}$ and the covariance matrix Σ has been subdivided into $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. Then the vectors $\boldsymbol{X}_{(1)}$ and $\boldsymbol{X}_{(1)}$ are independent iff $\Sigma_{12} = 0$.
Proof. (Exercise)

Property 4

Let the vector $\boldsymbol{X} \in \mathbb{R}^p$ be divided into subvectors $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_{(1)} \\ \boldsymbol{X}_{(2)} \end{pmatrix}$, $\boldsymbol{X}_{(1)} \in \mathbb{R}^r$, $r < p$, $\boldsymbol{X}_{(2)} \in \mathbb{R}^{p-r}$, and according to this subdivision the vector means are $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{(1)} \\ \boldsymbol{\mu}_{(2)} \end{pmatrix}$ and the covariance matrix Σ has been subdivided into $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. Assume for simplicity that the rank of Σ_{22} is full. Then the conditional density of $\boldsymbol{X}_{(1)}$ given that $\boldsymbol{X}_{(2)} = \boldsymbol{x}_{(2)}$ is

$$N_r(\boldsymbol{\mu}_{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_{(2)} - \boldsymbol{\mu}_{(2)}), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \tag{5}$$

Proof

- Expression $\boldsymbol{\mu}_{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_{(2)} - \boldsymbol{\mu}_{(2)})$ is a function of $\boldsymbol{x}_{(2)}$; denote is as $g(\boldsymbol{x}_{(2)})$. Construct r.v. $\boldsymbol{Z} = \boldsymbol{X}_{(1)} - g(\boldsymbol{X}_{(2)})$ and $\boldsymbol{Y} = \boldsymbol{X}_{(2)} - \boldsymbol{\mu}_{(2)}$. Observe $E\boldsymbol{Z} = 0$ and $E\boldsymbol{Y} = 0$.
- $\begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{Y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_r & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & \boldsymbol{I}_{p-r} \end{pmatrix} (\boldsymbol{X} - \boldsymbol{\mu}) \Rightarrow$ Normal (Property 2).
- $Var \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{Y} \end{pmatrix} = A\Sigma A^T = \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & -0 \\ 0 & \Sigma_{22} \end{pmatrix}$ block multiplication
  $\Rightarrow$ $\boldsymbol{Z}$ and $\boldsymbol{Y}$ uncorr. Normal $\Rightarrow$ independent (Property 3).
  - $\boldsymbol{Y}$ is a linear transformation of $\boldsymbol{X}_{(2)} \Rightarrow \boldsymbol{Z}$ and $\boldsymbol{X}_{(2)}$ indep.
    $\Rightarrow$ Conditional density of $\boldsymbol{Z}$ given $\boldsymbol{X}_{(2)} = \boldsymbol{x}_{(2)}$ will **not** depend on $\boldsymbol{x}_{(2)}$ and coincides with the unconditional density of $\boldsymbol{Z}$.

  $\Rightarrow$ $\boldsymbol{Z}$ normal with $Cov(\boldsymbol{Z}) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} := \Sigma_{1|2}$
  $\Rightarrow \boldsymbol{X}_{(1)} - g(\boldsymbol{x}_{(2)}) \sim N(0, \Sigma_{1|2})$

Example

As an immediate consequence of Property 4 we see that if $p = 2$ and $r = 1$, then for a two-dimensional normal vector

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}\right),$$

its conditional density $f(x_1|x_2)$ is

$$N(\mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(x_2 - \mu_2), \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2}.$$

As an exercise, try to derive the above result by direct calculations starting from the joint density $f(x_1, x_2)$, going over to the marginal $f(x_2)$ by integration and finally getting $f(x_1|x_2) = \frac{f(x_1,x_2)}{f(x_2)}$.

Property 5

If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ and Σ is nonsingular then

$$(\boldsymbol{X} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi_p^2$$

where χ_p^2 denotes the chi-square distribution with p degrees of freedom.

Proof.

It suffices to use the fact that the vector $\boldsymbol{Y} \in \mathbb{R}^p$ defined as $\boldsymbol{Y} = \Sigma^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}) \sim N(0, \boldsymbol{I}_p)$, i.e. it has p independent standard normal components. Then

$$(\boldsymbol{X} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{Y}^T \boldsymbol{Y} = \sum_{i=1}^{p} Y_i^2 \sim \chi_p^2$$

according to the definition of χ_p^2 as a distribution of the sum of squares of p independent standard normals.

Prediction: "Best Predictor"

- A corollary of Property 4
- Predict Y from p predictors $\boldsymbol{X} = (X_1\ X_2\ \cdots X_p)$ by choosing $g(\cdot)$ to minimise $E_Y([Y - g(\boldsymbol{X})]^2 | g(\boldsymbol{X}))$ s.t. $Eg(\boldsymbol{X})^2 < \infty$
- Optimal $g^*(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x})$: regression function

Prediction: Best Predictor for MVN

- In general, $g^*(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x})$ can be complicated.
- For MVN, much simpler.
- If $\begin{pmatrix} Y \\ \boldsymbol{X} \end{pmatrix} \in \mathbb{R}^{1+p}$ is normal, apply Property 4
  $\Rightarrow g^*(\boldsymbol{x}) = b + \boldsymbol{\sigma}_0^T C^{-1} \boldsymbol{x}$ for $b = E(Y) - \boldsymbol{\sigma}_0^T C^{-1} E(\boldsymbol{X})$,
  $C = Cov(\boldsymbol{X})$, and $\boldsymbol{\sigma}_0 = Cov(\boldsymbol{X}, Y)$
- i.e.,
$$g^*(\boldsymbol{x}) = E(Y) + \boldsymbol{\sigma}_0^T C^{-1}(\boldsymbol{x} - E(\boldsymbol{X}))$$
- In case of joint normality, prediction turns out linear.
- $C^{-1}\boldsymbol{\sigma}_0 \in \mathbb{R}^p$ is the vector of the regression coefficients.
  - results in variance $Var(Y) - \boldsymbol{\sigma}_0^T C^{-1} \boldsymbol{\sigma}_0$

Tests for Multivariate Normality

Graphical diagnostics

- Normality makes things easier.
- Is also sometimes an important assumption.
- Since margins and linear combinations of MVN are normal,
  1. check marginal distributions (e.g., Q-Q plots, the Shapiro-Wilk test);
  2. check scatterplots of pairs of observations;
  3. note outliers.
- Only good for bivariate normality.

Mardia's Multivarate Skewness and Kurtosis

Multivariate skewness: For $\boldsymbol{Y}$ independent of $\boldsymbol{X}$ but with the same distribution,

$$\beta_{1,p} = E[(\boldsymbol{Y} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu})]^3 \tag{6}$$

Multivariate kurtosis:

$$\beta_{2,p} = E[(\boldsymbol{Y} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu})]^2 \tag{7}$$

- Assuming these expectations exist.
- For $N_p(\boldsymbol{\mu}, \Sigma)$, $\beta_{1,p} = 0$ and $\beta_{2,p} = p(p+2)$.
- $p = 1 \Rightarrow \beta_{1,1} = \left(\frac{E(X-\mu)^3}{\sigma^3}\right)^2, \beta_{2,1} = \frac{E(X-\mu)^4}{\sigma^4}$
- Estimated as

$$\hat{\beta}_{1,p} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij}^3, \quad \hat{\beta}_{2,p} = \frac{1}{n} \sum_{i=1}^{n} g_i^2,$$

where $g_{ij} = (\boldsymbol{x} - \bar{\boldsymbol{x}})^T \boldsymbol{S}_n^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}})$

Mardia's test statistics

- $\hat{\beta}_{1,p} \geq 0$, and $\hat{\beta}_{2,p} \geq 0$
- For MVN, $\hat{\beta}_{1,p} \approx 0$, and $\hat{\beta}_{2,p} \approx p(p+2)$, respectively.
- Asymptotically, $k_1 = n\hat{\beta}_{1,p}/6 \sim \chi_{p(p+1))p+2)/6}$ and

$$k_2 = [\hat{\beta}_{2,p} - p(p+2)]/[8p(p+2)/n]^{1/2} \sim N(0,1)$$

$\Rightarrow$ Use k_1 and k_2 to test the null hypothesis of multivariate normality.

  - If neither hypothesis is rejected MVN assumption is in reasonable agreement with the data.
  - Mardia's multivariate kurtosis can also be used to detect outliers.

Caveat: Overreliance on tests

- Shapiro–Wilk, Mardia, etc.
  H_0 : population is (multivariate) normal
- Any deviation from normality then $Pr(reject\ H_0) \overset{n\to\infty}{\to} 1$
- CLT: $\bar{\boldsymbol{X}} \overset{n\to\infty}{\to}$ MNV regardless of population distribution
  - $\boldsymbol{S}$ too, but much more slowly
    $\Rightarrow$ As n increases,
  - more likely for test to conclude population non-normality.
  - need population normality less in the first place. $\Rightarrow$
    Particularly for large datasets, don't overrely on tests.

Software

SAS Use CALIS procedure. The quantity k2 is called Normalised Multivariate Kurtosis there, whereas $\hat{\beta}_{2,p} - p(p+2)$ bears the name Mardia's Multivariate Kurtosis.
R MVN::mvn, psych::mardia

Examples

Example

Testing multivariate normality of microwave oven radioactivity measurements (JW).

Additional resources

JW Sec. 4.1–4.2, 4.6.