Topic 2: The multivariate normal distribution

The multivariate normal distribution

This course will make heavy use of multivariate normal distribution. This distribution generalises the familiar normal distribution to multiple dimensions and provides a convenient way to represent many variables and their (pairwise) dependence on each other.

The *multivariate normal* (*MVN*) density is a generalisation of the univariate normal for $p \geq 2$ dimensions. Looking at the term $(\frac{x-\mu}{\sigma})^2 = (x-\mu)(\sigma^2)^{-1}(x-\mu)$ in the exponent of the well known formula

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-[(x-\mu)/\sigma]^2/2}, -\infty < x < \infty \tag{1.4}$$

for the univariate density function, a natural way to generalise this term in higher dimensions is to *replace* it by $(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$. Here $\boldsymbol{\mu} = \mathrm{E}\,\boldsymbol{X} \in \mathbb{R}^p$ is the expected value of the random vector $\boldsymbol{X} \in \mathbb{R}^p$ and the matrix

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

is the *covariance matrix*. Note that on the diagonals of Σ we get the *variances* of each of the p random variables whereas $\sigma_{ij} = \mathrm{E}[(X_i - \mathrm{E}(X_i))(X_j - \mathrm{E}(X_j))], i \neq j$ are the *covariances* between the ith and jth random variables. Sometimes, we will also denote σ_{ii} by σ_i^2.

Of course, the above replacement would only make sense if Σ were positive definite. In general, however, we can only claim that Σ is (as any covariance matrix) non-negative definite.

If Σ were positive definite, then the density of the random vector $\boldsymbol{X}$ can be written as

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{\frac{1}{2}}} \mathrm{e}^{-(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2}, \ -\infty < x_i < \infty, \ i = 1, 2 \ldots, p. \tag{1.5}$$

It can be directly checked that the random vector $\boldsymbol{X} \in \mathbb{R}^p$ has $\mathrm{E}\,\boldsymbol{X} = \boldsymbol{\mu}$ and

$$\mathrm{E}[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\top] = \Sigma.$$

Since the density is uniquely defined by the *mean vector and the covariance matrix* we will denote it by $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Notably, while the multivariate normal distribution does not have a density if Σ is singular (i.e., the variables are linearly dependent), it still has a valid expectation, variance, etc..

Properties of multivariate normal

The following *properties* of multivariate normal are useful in the rest of the course.

Property 1. If $\Sigma = D(\boldsymbol{X}) = \Lambda$ is a diagonal matrix then the p components of $\boldsymbol{X}$ are independent.

The above property can be paraphrased as "for a multivariate normal, if its components are uncorrelated they are also independent". On the other hand, it is well known that *always, i.e. not only for normal* from the fact that certain components are independent we can conclude that they are also uncorrelated. Therefore, for the **multivariate normal distribution** we can conclude that its components are **independent if and only if they are uncorrelated!**

Property 2. If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $C \in \mathcal{M}_{q,p}$ is an arbitrary matrix of real numbers then

$$Y = CX \sim N_q(C\boldsymbol{\mu}, C\Sigma C^\top).$$

Note also that if it happens that the rank of C is full and if $\mathrm{rk}(\boldsymbol{\Sigma}) = p$ then the rank of $C\Sigma C^\top$ is also full, i.e. the distribution of $\boldsymbol{Y}$ would not be degenerate in this case.

Property 3. (This is a finer version of Property 1). 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 $\boldsymbol{X}_{(2)}$ are independent iff $\Sigma_{12} = 0$.

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

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

Example 1.1. As an immediate consequence of Property 4 we see that if $p = 2, r = 1$ 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 $N(\mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(x_2 - \mu_2), \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2})$.

Optional 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 $(\boldsymbol{X}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \sim \chi^2_p$ where χ^2_p denotes the chi-square distribution with p degrees of freedom.

Remark 1.1. This stems from the fact that the vector $\boldsymbol{Y} \in \mathbb{R}^p : \boldsymbol{Y} = \Sigma^{-\frac{1}{2}}(\boldsymbol{X}-\boldsymbol{\mu}) \sim N(0, I_p)$ i.e. it has p independent standard normal components, where $\Sigma^{-\frac{1}{2}}$ is defined either via the spectral decomposition (0.8) or the Cholesky Decomposition on slide "Numerical stability and Cholesky decomposition" Then

$$(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = \boldsymbol{Y}^\top \boldsymbol{Y} = \sum_{i=1}^{p} Y_i^2 \sim \chi^2_p$$

according to the definition of χ^2_p as a distribution of the sum of squares of p independent standard normals. However, this multivariate version of standardising the normal distribution is useful in other situations, such as multivariate ANOVA and regression diagnostics.

Finally, one more interpretation of the result in Property 4 will be given. Assume we want, as is a typical situation in statistics, to predict a random variable Y that is correlated with some p random variables (predictors) $\boldsymbol{X} = \begin{pmatrix} X_1 & X_2 & \cdots & X_p \end{pmatrix}$. Trying to find the *best predictor* of Y we would like to minimise the Mean Squared Error (MSE), the expected value $\mathrm{E}_Y[\{Y-g(\boldsymbol{X})\}^2|\boldsymbol{X}=\boldsymbol{x}]$ over all possible choices of the function g such that $\mathrm{E}\, g(\boldsymbol{X})^2 < \infty$. A little careful work and use of basic properties of conditional expectations, i.e.,

$$\begin{aligned}\frac{\partial}{\partial g^*(\boldsymbol{X})}\mathrm{E}_Y[\{Y-g^*(\boldsymbol{X})\}^2|\boldsymbol{X}=\boldsymbol{x}] &= \mathrm{E}_Y[\frac{\partial}{\partial g^*(\boldsymbol{X})}\{Y-g^*(\boldsymbol{X})\}^2|\boldsymbol{X}=\boldsymbol{x}] \\ &= -2\,\mathrm{E}_Y[Y-g^*(\boldsymbol{X})|\boldsymbol{X}=\boldsymbol{x}] \stackrel{\text{set}}{=} 0,\end{aligned}$$

leads us to the conclusion that the optimal solution to the above minimisation problem is $g^*(\boldsymbol{x}) = \mathrm{E}(Y|\boldsymbol{X}=\boldsymbol{x})$.

This optimal solution is also called the *regression function*. Thus given a particular realisation $\boldsymbol{x}$ of the random vector $\boldsymbol{X}$ the regression function is just the conditional expected value of Y given $\boldsymbol{X}=\boldsymbol{x}$.

In general, the conditional expected value may be a complicated nonlinear function of the predictors. However, if we assume *in addition* that the joint $(p+1)$-dimensional distribution of Y and $\boldsymbol{X}$ is **normal** then by applying Property 4 we see that given the realisation $\boldsymbol{x}$ of $\boldsymbol{X}$, the best prediction of the Y value is given by $b+\sigma_0^\top C^{-1}\boldsymbol{x}$ where $b=\mathrm{E}(Y)-\sigma_0^\top C^{-1}\,\mathrm{E}(\boldsymbol{X})$, C is the covariance matrix of the vector $\boldsymbol{X}$, σ_0 is the vector of Covariances of Y with $X_i, i=1,\ldots,p$.

Indeed, we know that when the joint $(p+1)$-dimensional distribution of Y and $\boldsymbol{X}$ is **normal** the regression function is given by

$$\mathrm{E}(Y)+\sigma_0^\top C^{-1}(\boldsymbol{x}-\mathrm{E}(\boldsymbol{X})).$$

By introducing the notation $b=\mathrm{E}(Y)-\sigma_0^\top C^{-1}\,\mathrm{E}(\boldsymbol{X})$ we can write this as $b+\sigma_0^\top C^{-1}\boldsymbol{x}$.

That is, **in case of normality, the optimal predictor of Y in the least squares sense turns out to be a very simple linear function of the predictors**. The vector $C^{-1}\sigma_0 \in \mathbb{R}^p$ is the *vector of the regression coefficients*. Substituting the optimal values we get the minimal value of the sum of squares which is equal to $V(Y) - \sigma_0^\top C^{-1}\sigma_0$.

Tests for multivariate normality

We have seen that the assumption of multivariate normality may bring essential simplifications in analysing data. But applying inference methods based on the multivariate normality assumption in cases where it is grossly violated may introduce serious defects in the quality of the analysis. It is therefore important to be able to check the multivariate normality assumption. Based on the properties of normal distributions discussed in this topic, we know that all linear combinations of normal variables are normal and the contours of the multivariate normal density are ellipsoids.

Therefore we can (to some extent) check the multivariate normality hypothesis by:

1. checking if the marginal distributions of each component appear to be normal (by using Q–Q plots, for example);
2. checking if the scatterplots of pairs of observations give the elliptical appearance expected from normal populations;
3. are there any outlying observations that should be checked for accuracy.

All this can be done by applying univariate techniques and by drawing scatterplots, which are well developed in R. To some extent, however, there is a price to be paid for concentrating on univariate and bivariate examinations of normality.

There is a need to construct a "good" overall test of multivariate normality. One of the simple and tractable ways to verify the multivariate normality assumption is by using tests based on **Mardia's multivariate skewness and kurtosis measures**. For any general multivariate distribution we define these respectively as

$$\beta_{1,p} = \mathrm{E}[(\boldsymbol{Y}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu})]^3 \tag{1.6}$$

provided that $\boldsymbol{X}$ is independent of $\boldsymbol{Y}$ but has the same distribution and

$$\beta_{2,p} = \mathrm{E}[(\boldsymbol{X}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu})]^2 \tag{1.7}$$

(if the expectations in (1.6) and (1.7) exist). For the $N_p(\boldsymbol{\mu}, \Sigma)$ distribution: $\beta_{1,p} = 0$ and $\beta_{2,p} = p(p+2)$.

(Note that when $p = 1$, the quantity $\beta_{1,1}$ is the square of the skewness coefficient $\frac{\mathrm{E}(X-\mu)^3}{\sigma^3}$ whereas $\beta_{2,1}$ coincides with the kurtosis coefficient $\frac{\mathrm{E}(X-\mu)^4}{\sigma^4}$.)

For a sample of size n *consistent estimates* of $\beta_{1,p}$ and $\beta_{2,p}$ can be obtained as

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

$$\hat{\beta}_{2,p} = \frac{1}{n}\sum_{i=1}^{n} g_{ii}^2$$

where $g_{ij} = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top \boldsymbol{S}_n{}^{-1}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})$. Notice that for $\hat{\beta}_{1,p}$, we take advantage of our sample being independent and use observations $\boldsymbol{x}_j$ for $j \neq i$ as the "$\boldsymbol{Y}$" values for $\boldsymbol{x}_i$.

Both quantities $\hat{\beta}_{1,p}$ and $\hat{\beta}_{2,p}$ are nonnegative and for multivariate data, one would expect them to be around zero and $p(p+2)$, respectively. Both quantities can be utilised to detect departures from multivariate normality.

Mardia has shown that asymptotically, $k_1 = n\hat{\beta}_{1,p}/6 \sim \chi^2_{p(p+1)(p+2)/6}$, and $k_2 = [\hat{\beta}_{2,p} - p(p+2)]/[8p(p+2)/n]^{\frac{1}{2}}$ is standard normal. Thus we can use k_1 and k_2 to test the null hypothesis of multivariate normality. If neither hypothesis is rejected, the multivariate normality assumption is in reasonable agreement with the data. It also has been observed that Mardia's multivariate kurtosis can be used as a measure to detect outliers from the data that are supposedly distributed as multivariate normal. Mardia's multivariate kurtosis can also be used to detect outliers.

***Remark* 1.2** (Overreliance on hypothesis tests). Shapiro–Wilk, Mardia, and other distribution tests have, as their null hypothesis, that the true population distribution is (multivariate) normal. This means that if the population distribution deviates from normality even a little, then as the sample size increases, the power of the test (the probability of rejecting the null hypothesis of normality) approaches 1.

At the same time, as the sample size increases, the Central Limit Theorem tells us that many statistics, including sample means and (much more slowly) sample variances and covariances, approach normality—and multivariate statistics generally approach multivariate normality. This means that regardless of the underlying distribution, the statistical procedures depending on the normality assumption become valid—even as the chances that a statistical hypothesis test will detect what non-normality there is approaches 1.

This means that we must not rely on hypothesis testing blindly but consider the situation on a case-by-case basis, particularly when dealing with large datasets. For a decent sample size, the “symmetric, bell-shaped” heuristic may indicate an adequate distribution, even if a hypothesis test reports a small p-value.

Check your understanding

Complete the below exercises to check your understanding of concepts presented so far.

1. Let X_1 and X_2 denote i.i.d. $N(0,1)$ random variables.

a) Show that the random variables $Y_1 = X_1 - X_2$ and $Y_2 = X_1 + X_2$ are **independent**, and find their marginal densities.

b) Find $P\left(X_1^2 + X_2^2 < 2.41\right)$.

2. Let $\boldsymbol{X} \sim N_3(\boldsymbol{\mu}, \Sigma)$ where

$$\boldsymbol{\mu} = \begin{pmatrix} 3 \\ -1 \\ 2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

a) For $A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \end{pmatrix}$ find the distribution of $\boldsymbol{Z} = A\boldsymbol{X}$ and find the correlation between the two components of $\boldsymbol{Z}$.

b) Find the conditional distribution of $[X_1, X_3]^\top$ given $X_2 = 0$.

Solutions to these exercises will be provided in the 'Exercises' discussion forum.

Demonstration: Multivariate normal distribution

This demonstration can be completed using the provided RStudio or your own RStudio.

Transcript

To complete this task select the 'MVN_Examples.demo.Rmd' in the 'Files' section of RStudio. Follow the demonstration contained within the RMD file.

If you choose to complete the example in your own RStudio, upload the following file:

MVN_Examples.demo.Rmd

After working through the demonstration, complete the associated 'Challenge' on the next slide.

The output of the RMD file is also displayed below:

Challenge: Multivariate normal distribution

If you choose to complete this task in your own RStudio, upload the following file:

Click on the 'MVN_Examples.challenge.Rmd' in the 'Files' section to begin. Enter your response to the tasks in the 'Enter your code here' section.

MVN_Examples.challenge.Rmd

This activity and the solution will be discussed at the Collaborate session this week. In the meantime, share and discuss your results in the 'Tutorials' discussion forum.

The solution will also be available here on Friday of this week by clicking on the 'Solution' tab in the top right corner.

The output of the RMD file is also displayed below: