Multivariate Normal Distribution

Based on module MATH 5772M: Multivariate and Cluster Analysis
Lectured by Arief Gusnanto

1 Introduction

When working with data, it is common to have multiple features of interest compared to just one, which we are introduced to when we first learn statistics. To work with this data, we need a new class of distributions that allows us to model the extra complexity arising from multiple variables. In particular, the multivariate normal distribution generalises the univariate normal distribution to the multidimensional case. Using the multivariate normal distribution, we can model individual variables' behaviour and the interaction between variables.

This report aims to introduce the multivariate normal distribution as a transformation of independent standard normal distributions. We begin by looking at an example to motivate the need for the multivariate normal distribution. Next, we look at the case where we have just two variables of interest, and starting from a vector of independent standard random variables, we derive the bivariate normal distribution through a series of transformations. Finally, we define the multivariate normal distribution for n-variables.

2 Motivation

We start by looking at an example of a situation we might encounter. A doctor wants to assess the effectiveness of an exercise scheme and has two groups of people. Group A participated in the scheme, and Group B, a control group, did not. For each group of patients, we record a resting heart rate and a weight value. We want to know if there is a statistical significance in the difference between the two groups. To solve this problem, we first need a model for the distribution of heart rate and weight.

3 Bivariate Normal Distribution

We will build the bivariate normal distribution in a series of steps. First, we begin by modelling the base structure of our data. We can assume the variables are normally distributed, so the first step is to start with two independent standard

normal distributions. We use two variables because people with the same weight might have different heart rates and vice versa. So, the space spanned by these values is two-dimensional.

Figure 1 below shows the density of 1000000 pseudo-random samples for $(Z_1, Z_2)^T$, where Z_1 and Z_2 are i.i.d $N(0, 1)$ random variables. We can see that we maintain the properties that make the normal distribution useful, such as symmetry in the form of roughly circular shaped contours, and bell-shaped surface in the form of concentric contours which show a decreased density of points as we get further away from the origin. We now have a way to model the base structure of our data, but we need to fit it to our specific situation.

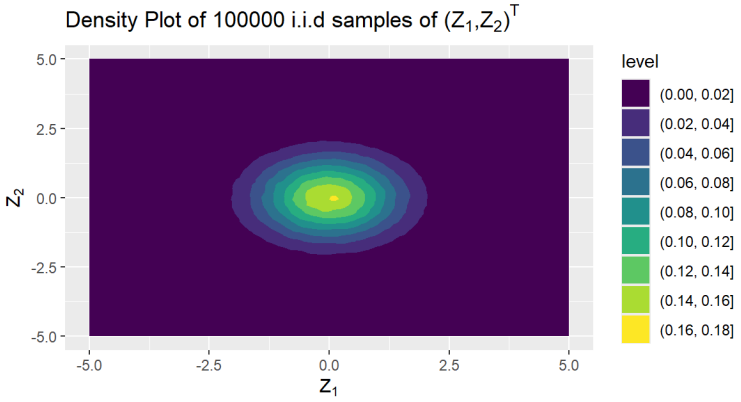

Figure 1: Density Plot for Standard Bivariate Normal Distribution

The first adjustment we make is that the variation can differ in different directions. For example, people's weights might vary greatly, but people's resting heart rates might be less varied. In the example given above, the variation is the same in all directions, so we must adjust it. We do this by scaling each of our variables.

Let $\begin{pmatrix} A_1 \\ A_2 \end{pmatrix} = \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$ where $\alpha_1, \alpha_2 \in \mathbb{R}_{>0}$. We can assume $\alpha_1, \alpha_2 \neq 0$ because if they did, we would remove all information from that dimension, which would be a degenerate case. Since the standard normal distribution is symmetric about 0, we can further assume that $\alpha_1, \alpha_2 > 0$.

Figure 2 shows the values sampled above scaled with $\alpha_1 = 4$ and $\alpha_2 = 2$. We can see that the contours become ellipses, but we still maintain symmetry about each axis and the bell-shaped surface.

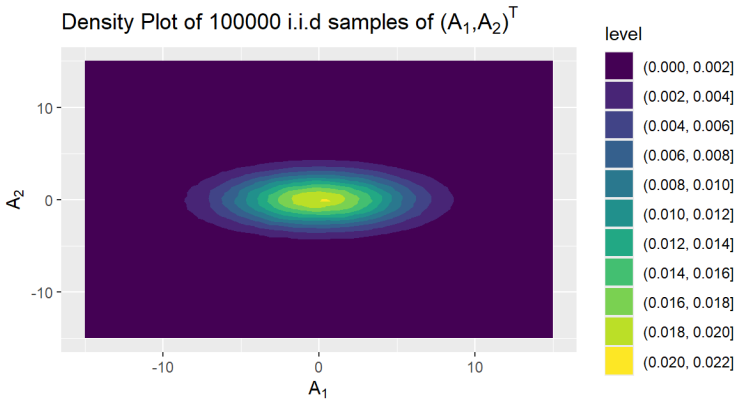

Figure 2: Density Plot for Scaled Bivariate Normal Distribution

The next addition we need to make is to describe relationships between variables. This is one of the key reasons we care about multivariate data; we want to know how variables interact. If we continue with our example, we might expect a heavier person to have a faster resting heart rate because that weight means potentially more fat and other factors, which means the heart needs to work harder. There is a correlation between weight and resting heart rate. To add this to our model, we need to perform a change of basis. Since we need to be careful to preserve the independence of our standard normal random variables, we need an orthonormal change of basis. We follow the details found in [1] for how to do this.

Let $\begin{pmatrix} B_1 \\ B_2 \end{pmatrix} = P \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix} P^T \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$ where $\alpha_1, \alpha_2 \in \mathbb{R}_{>0}$ and $P \in \mathbb{R}^{2,2}$ is an orthogonal matrix. Figure 3 below shows the case $P = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}$. We see a positive trend between B_1 and B_2. Essentially, we have just rotated our data to fit a trend.

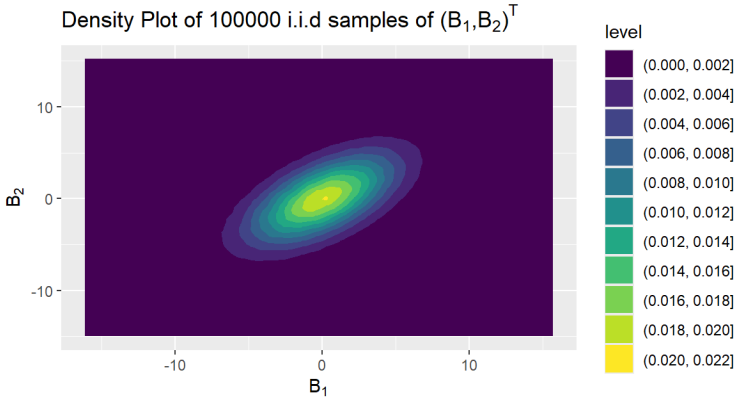

Figure 3: Density Plot for Correlated Bivariate Normal Distribution

We have now finished adding variations to our model. We now need to change the location of our distribution. If the average heart rate of a group of people is 0, then it is highly likely that they are dead. We do not want that, so we need to change the average heart rate to a more typical value, like 70. We do this by just shifting our values by a fixed amount. $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + S \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$ where $\mu_1, \mu_2 \in \mathbb{R}$ and $S = P \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix} P^T$, given above. The visualisation for this has been omitted due to space constraints. We have now fully modelled our situation.

4 Definition

The above steps should hopefully clarify the following definition of the multivariate normal distribution for the reader, which is a simplified version of the one found in [2]. The joint p.d.f, mean vector and covariance matrix calculations are left to the reader. From this definition, we can now begin looking into multivariate hypothesis tests, particularly the union-intersection test and likelihood ratio test.

Definition 1 (Multivariate Normal Distribution) *Let Z be a random n-vector of i.i.d $N(0,1)$ random variables. A random n-vector X is distributed according to a multivariate normal distribution with mean vector $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n,n}$, with Σ positive definite and symmetric, if*

$$X = \mu + \Sigma^{\frac{1}{2}} Z$$

.

References

[1] T. W. Körner. *Vectors, Pure and Applied: A General Introduction to Linear Algebra*, pages 192–210. Cambridge University Press, 2012.

[2] S. Ross. *A First Course in Probability*, pages 383–389. Pearson Education, Limited, 2019.

UNIVERSITY OF LEEDS

School of Mathematics

Declaration of Academic Integrity
for Individual Pieces of Work

I declare that I am aware that as a member of the University community at the University of Leeds I have committed to working with Academic Integrity and that this means that my work must be a true expression of my own understanding and ideas, giving credit to others where their work contributes to mine.

I declare that the attached submission is my own work.

Where the work of others has contributed to my work, I have given full acknowledgement using the appropriate referencing conventions for my programme of study.

I confirm that the attached submission has not been submitted for marks or credits in a different module or for a different qualification or completed prior to entry to the University.

I have read and understood the University's rules on Academic Misconduct. I know that if I commit an academic misconduct offence there can be serious disciplinary consequences.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties to verify that this is my own work, and for quality assurance purposes.

I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and I wish to have taken into account.

Student Signature: J Andy **Student Number:** 201396918

Student Name: James Andy **Date:** 13/3/25

Please note:

When you become a registered student of the University at first and any subsequent registration you sign the following authorisation and declaration:

"I confirm that the information I have given on this form is correct. I agree to observe the provisions of the University's Charter, Statutes, Ordinances, Regulations and Codes of Practice for the time being in force. I know that it is my responsibility to be aware of their contents and that I can read them on the University web site. I acknowledge my obligation under the Payment of Fees Section in the Handbook to pay all charges to the University on demand.

I agree to the University processing my personal data (including sensitive data) in accordance with its Code of Practice on Data Protection http://www.leeds.ac.uk/dpa . I consent to the University making available to third parties (who may be based outside the European Economic Area) any of my work in any form for standards and monitoring purposes including verifying the absence of plagiarised material. I agree that third parties may retain copies of my work for these purposes on the understanding that the third party will not disclose my identity.'"