TMA4267 Linear Statistical Models
Part 1: Multivariate random variables and the multivariate normal distribution
Recommended exercise 2 - V2017

January 12, 2017

Keywords: multivariate normal distribution, conditional distribution, linear combinations, independence and zero covariance, quadratic forms, ellipsoids, copulae, chisquare distributions, t-distribution, F-distribution, ratio of quadratic forms.

- The properties of the multivariate normal (mvN) distribution are studied in this exercise set: linear combinations of mvN variables are mvN, zero covariance implies independence, subsets of mvN random variables and conditional distributions of mvN are also mvN. (Problem 1, 2).
- It is also important to be able to sample data from the mvN (Problem 3) and to see from bivariate plots when data are (obviously) not multivariate normal (elliptic contours), even though the marginal distributions are univariate normal (Problem 4).
- The Student-t and Fisher distributions can be developed from the univariate normal distribution - and Problem 5 shows how this can be done with the aid of moment generating functions and the multivariate transformation formula. Problem 6 studies the same distributions using `R`.
- The last topic of Part 1 was "quadratic forms" and Problem 7a applies the formula for the mean of a quadratic form (no assumptions on distribution) and Problem 7b the distributional results based on a multivariate normal random vector.

Problem 1: Simple calculations with the multivariate normal distribution

Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{bmatrix}$

a) Find the distribution of $3X_1 - 2X_2 + X_3$.

b) Find a 2×1 vector $\boldsymbol{a}$ such that X_2 and $X_2 - \boldsymbol{a}^T \begin{bmatrix} X_1 \\ X_3 \end{bmatrix}$ are independent.

c) Find the conditional distribution of X_1 given that $X_2 = x_2$ and $X_3 = x_3$.

Problem 2: From correlated to independent variables

(Exam TMA4267, May 2013, Problem 1, slightly modified)

Assume that the random vector $\mathbf{X} = (X_1, X_2, X_3)^T$ has a trivariate normal distribution with mean vector $\boldsymbol{\mu} = (2, 6, 4)^T$ and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & -1 \\ 1 & -1 & 3 \end{bmatrix}.$$

a) Find out which of the random variables X_1 and X_2 is most correlated (in absolute value) with X_3. What is the distribution of the random vector $\mathbf{Z} = (X_2 - X_1, X_3 - X_1)^T$?

A company is measuring three quality characteristics in order to control the quality of a product. Their respective random variables can be arranged in a random vector $\mathbf{X} = (X_1, X_2, X_3)^T$. Based on previous experience, it is reasonable to assume that $\mathbf{X}$ is trivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, as given above.

The company would like to have a simplified quality control procedure where they only consider a bivariate random vector instead of a trivariate one. The eigenvalues λ_1, λ_2 and λ_3 of $\boldsymbol{\Sigma}$, as well as their respective eigenvectors $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$ are given below as `R`-output.

```
$values
[1] 3.8793852 1.6527036 0.4679111

$vectors
           [,1]       [,2]       [,3]
[1,] -0.2931284 -0.4490988  0.8440296
[2,]  0.4490988 -0.8440296 -0.2931284
[3,] -0.8440296 -0.2931284 -0.4490988
```

b) Define the bivariate vector $\mathbf{Y} = (\mathbf{e}_1^T\mathbf{X}, \mathbf{e}_2^T\mathbf{X})^T$. Why does $\mathbf{Y}$ have a bivariate normal distribution?

Show that $Y_1 = \mathbf{e}_1^T\mathbf{X}$ and $Y_2 = \mathbf{e}_2^T\mathbf{X}$ are independent. How much of the total variance in $\mathbf{X}$ is explained by $\mathbf{Y}$? Hint: the total variance is the trace of the covariance matrix, that is, the sum of the variances. Also, there is a relationship between the trace and eigenvalues of a matrix – which relationship?

Problem 3: The bivariate normal distribution

Let X and Y be random variables with pdf $f(x, y)$ parameterized by $(\mu_X, \mu_Y, \rho, \sigma_X^2, \sigma_Y^2)$.

$$\begin{aligned} f(x, y) &= ce^{-\frac{1}{2}Q(x,y)} \\ c &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \\ Q(x, y) &= \frac{1}{(1-\rho^2)}[(\frac{x-\mu_X}{\sigma_X})^2 + (\frac{y-\mu_Y}{\sigma_Y})^2 - 2\rho(\frac{x-\mu_X}{\sigma_X})(\frac{y-\mu_Y}{\sigma_Y})] \end{aligned}$$

This is the bivariate normal distribution.

a) We will study the quadratic form $Q(x, y)$. Show that $Q(x, y)$ can be written as

$$(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$$

where

$$\boldsymbol{x} = \begin{bmatrix} x \\ y \end{bmatrix}$$
$$\boldsymbol{\mu} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}$$
$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathrm{Var}(Y) \end{bmatrix} = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

Remark: $\boldsymbol{\Sigma}$ is called the variance-covariance matrix of $\boldsymbol{X}$.

b) Now we focus on rewriting the pdf $f(x, y)$ using vectors and matrices. From a) we saw that $\det(\boldsymbol{\Sigma}) = \sigma_X^2 \sigma_Y^2 (1 - \rho^2)$. Use this to rewrite the pdf $f(x, y)$ in a vector-matrix form, that is as a function of $\boldsymbol{x}$, with parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Call the new pdf $f(\boldsymbol{x})$.

c) Now we turn to look at contours of $f(\boldsymbol{x})$, that is, the graphical form (in $\boldsymbol{x}$) given for constant values of $f(\boldsymbol{x})$.

Why can the contours be seen as solutions to the equation

$$(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = d^2$$

for a given constant d^2?

Use the method of diagonalization (spectral decomposition) to explain that the contours are ellipses with center in $\boldsymbol{\mu}$, axes in the direction of the eigenvectors of $\boldsymbol{\Sigma}$, with halflengths $\sqrt{\lambda_1} d$ and $\sqrt{\lambda_2} d$, where λ_1 and λ_2 are the eigenvalues of $\boldsymbol{\Sigma}$.

Remark: remember that there is a simple connection between the eigenvectors and eigenvalues of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$.

d) For the special case that $\sigma_X = \sigma_Y$ find the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$. Make a drawing of the contours by hand.

e) Now we turn to R to draw ellipses. This can be done using

```
library(ellipse)
```

First look at $\mu_X = \mu_Y = 1$, and $\sigma_X = 1$, $\sigma_y = 1$ and $\rho = 0.5$.

```
mu1=mu2=1
sigma1=1
sigma2=1
rho=0.5
plot(ellipse(rho, scale=c(sigma1, sigma2),centre=c(mu1,mu2)),type="l")
```

Try varying the parameters in $\boldsymbol{\Sigma}$ and observe.

Problem 4: Normal marginals, but not multivariate normal?

When assessing whether a data set is multivariate normal, it is important to not only look at univariate marginal plots, but also at multivariate plots.

In finance - in general - one may find data sets with marginal normals, but with a different correlation structure than for the mulitivariate normals. For instance (as we will see below in a very extreme version) dependence structures that are asymmetric may occur.

a) Reading data The data has been generated and dumped to a file using `dput`. You read the data using the command `dget`. Let us call the data set `ccdata`.

```
ccdata <- dget("http://www.math.ntnu.no/emner/TMA4267/2017v/ccdata.dd")
```

Check the dimensions of the data set, i.e. number of observations and number of variables.

We now would like to examine the data to see if it may come from a multivariate normal population.

b) Marginal plots Look first at the data marinally. Make boxplots, histograms and normal-qq-plots for for each variable. What do the normal-qq-plots show?

c) Joint plot with ellipse Then turn to the data set in two dimensions. First make a scatterplot. Does this look multivariate normal?

Let us assume that the data came from a mulitvariate normal distribution with mean vector `c(0,0)` and covariance matrix `matrix(c(1,0.8,0,8,1),2,2)`. Add an ellipse with probability 95% to the plot (assuming that data come from a mulitvariate normal distribution with the given mean and covariance matrix), use library `ellipse`. Are around 95% of the data inside the ellipse? Do you see a trend?

Comments: I hope it is clear to you that the data are not taken from a mulitivariate normal distribution. If you want to take a peek at the "correct" contours of this density (that are not ellipse) a imageplot is found from the www-page of the course, in the table for this exercise.

The data set was generated with a use of copulas, and a depencence structure called a Clayton Copula was used.

Problem 5: The chi-square, t and F-distribution

Why this problem? We believe that it is important that you see how these distributions are derived, since they play a major role in the course. The calculations performed here are rather technical, and this makes this problem not particularly suited to be an exam question. However, the concepts and way of thinking is useful in general.

This problem is building on Recommended Exercise 1, Problem 5, consult that before you start.

a) Use the multivariate transformation formula to find the pdf of the F-distribution.
Hint: Let $V \sim \chi^2_p$ and $W \sim \chi^2_q$, where V and W are independent. Let then $F = \frac{V/p}{W/q}$, and $G = W$, and use the multivariate transformation formula to find the joint pdf of F and G. Find the marginal distribution of F from this joint distribution. For the last part it will help you to recognize the integral of a χ^2 pdf.

b) Let $U \sim N(0,1)$ and $V \sim \chi^2_p$, and U and V are independent. Find the pdf of the random variable $T = \frac{U}{\sqrt{V/p}}$.
Hint: First find the joint pdf of U and V, then use the multivariate transformation formula for T and $W = V$ to find the joint pdf of T and W, and then find the marginal pdf of T. For the last part it will help you to recognize the integral of gamma-pdf with parameters $\alpha = (p+1)/2$ and $\beta = 2/(1+t^2/p)$. That is if X is gamma(α,β), then

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$

c) Let $T \sim t_q$ (t-distribution with q degrees of freedom). Then show that $T^2 \sim F_{1,q}$ (Fisher-distribution with 1 and q degrees of freedom).

d) Make a map visualizing the relationships between the normal, chisquare, t and F-distribution,

Problem 6: t and F by simulation - in R

This is the sequel to Recommended Exercise 1, Problem 6.

Let $B = 10000$ and $n = 10$.

a) Now to the t. First simulate B data points from the standard normal distribution, and then simulate B data independently from the chi-square distribution with $n-1$ degrees of freedom. Make the t-ratio (see the previous problem) and plot a histogram of the data. Add the pdf of the t_{n-1} pdf to the histogram. Then add vertical lines for the the critical values for the 0.15 and 0.85 quantiles. Repeat this for other values of n.

b) Then, the F-distribution. Take the B t-ratios in d) and square them. Plot a histogram of the squared t-ratios. Add the pdf of the $F_{1,n-1}$ to the histogram. Then add vertical lines for the the critical values for the 0.05 and 0.95 quantiles.

c) The F-distribution can also be constructed as a ratio of chi-square variables. Let $n_1 = 5$ and $n_2 = 40$. Simulate B data independently from the chi-square distributions with n_1 and n_2 degrees of freedom. Make the F-ratio as defined in the previous problem. Plot a histogram of the data. Add the pdf of the F_{n_1,n_2} to the histogram. Then add vertical lines for the the critical values for the 0.05 and 0.95 quantiles.

Problem 7: Linear combinations and quadratic forms

(Exam TMA4267, spring 2014, Problem 1 - see also Recommended Exercise 1, Problem 2, for first part of the exam question.)

Let $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$ be a trivariate random vector with mean $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ and covariance matrix $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. Further, let $\boldsymbol{A} = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix}$ be a matrix of constants.

Define $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \boldsymbol{AX}$.

a) Find the mean of $\boldsymbol{X}^T\boldsymbol{AX}$.

Hint: formula involving the trace.

b) Now assume that $\boldsymbol{X}$ is trivariate normal.

Show that $\boldsymbol{A}$ is a symmetric and idempotent matrix. Find the rank of $\boldsymbol{A}$.

Derive the distribution of $\boldsymbol{X}^T\boldsymbol{AX}$.

Find the probability that $\boldsymbol{X}^T\boldsymbol{AX}$ is smaller than 6.