Examination Multivariate Statistical Methods

Linköpings Universitet, IDA, Statistik

Course code and name:	732A97 Multivariate Statistical Methods
Date:	2020/08/25, 8–12
Examinator:	Krzysztof Bartoszek phone 013–281 885
Provided aids:	Table with common formulæ and moment generating functions Table of integrals Table with distributions from Appendix in the course book
Grades:	A= $[18 - \infty)$ points B= $[16 - 18)$ points C= $[14 - 16)$ points D= $[12 - 14)$ points E= $[10 - 12)$ points F= $[0 - 10)$ points
Instructions:	Write clear and concise answers to the questions. Please number each page according to the pattern: Question number . page in question number i.e. Q1.1, Q1.2, Q1.3,. . ., Q2.1, Q2.2, . . ., Q3.1, ... Scan/take photos of your solutions preferably into a single pdf file but if this is not possible multiple pdf pages are fine and also multiple .bmp, .jpg, .png files. Please do not use other formats for scanned/photographed solutions. Please submit your solutions via LISAM or e–mail. Name your solution files as: **[your id]_[own file description].[format]** If emailing, please email them to **BOTH** krzysztof.bartoszek@liu.se and KB_LiU_exam@protonmail.ch . During the exam you may ask the examiner questions by emailing them to KB_LiU_exam@protonmail.ch **ONLY**. Other exam procedures in LISAM.

Problem 1 (4p)

Assume that $\mathbf{X} \in \mathbb{R}^{n \times d}$ is a matrix of n d–dimensional observations (each row is an observation), where $n > d$. Let $\mathbf{S}$ be the sample covariance matrix and define $\mathbf{Q} = (n-1)^{-1}\mathbf{X}^T\mathbf{X}$. What is the relationship between $\mathbf{S}$ and $\mathbf{Q}$?
TIP: Recall the matrix formula for the sample average

$$\overline{X} = n^{-1}\mathbf{M}\vec{1}_n$$

and sample covariance matrix,

$$\mathbf{S} = (n-1)^{-1}\left(\mathbf{M} - \overline{X}\vec{1}_n^T\right)\left(\mathbf{M} - \overline{X}\vec{1}_n^T\right)^T$$

where $\mathbf{M} \in \mathbb{R}^{d \times n}$ is the matrix of observations, where each column is an observation and $\vec{1}_n$ is a vector of n 1s.

Problem 2 (5p)

Suppose that the bivariate normal random vector $\vec{X} = [X_1, X_2]^T$ is normally distributed with mean vector $\vec{\mu} = [\mu_1, \mu_2]^T$ and covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

We know that the density function of a normal distribution is constant on the ellipsoids

$$\left(\vec{X} - \vec{\mu}\right)^T \mathbf{\Sigma}^{-1}\left(\vec{X} - \vec{\mu}\right) = c,$$

for every $c > 0$.
(a 1p) Under what conditions are X_1 and X_2 independent?
(a 4p) Consider a new random vector defined as $\vec{Y} := [(X_1 - \mu_1)/\sigma_1, (X_2 - \mu_2)/\sigma_2]^T \equiv [Y_1, Y_2]$. Show now that the constant density ellipses are given, for $k > 0$, by the equation

$$\frac{1}{1-\rho^2}\left(Y_1^2 - 2\rho Y_1 Y_2 + Y_2^2\right) = k.$$

Problem 3 (5p)

Consider a bivariate normal random vector $\vec{X} = [X_1, X_2]^T$. The covariance matrix of $\vec{X}$, $\mathbf{\Sigma}$, has one positive eigenvalue $\lambda > 0$ and the other eigenvalue equal 0. What can you say about the relationship between X_1 and X_2? What is the relationship between λ and the variances of X_1 and X_2?
TIP: Start working with the eigendecomposition of $\mathbf{\Sigma}$.

Problem 4 (6p)

You are provided with the following distributional results.

- Let $\mathbb{R}^p \ni \vec{X} \sim \mathcal{N}(\vec{\mu}, \mathbf{\Sigma})$, then
$$(\vec{X} - \vec{\mu})^T \mathbf{\Sigma}^{-1} (\vec{X} - \vec{\mu}) \sim \chi^2_p,$$
- Let $\mathbb{R}^p \ni \overline{x}$ be the sample mean of n normal observations and $\mathbf{S}$ the sample covariance. If the population expectation is $\vec{\mu}$, then
$$(\overline{x} - \vec{\mu})^T \left(\frac{1}{n}\mathbf{S}\right)^{-1} (\overline{x} - \vec{\mu}) \sim \frac{(n-1)p}{n-p} F_{p,n-p},$$
- If we have two independent samples, both of dimension p, first of size n_1 from $\mathcal{N}(\vec{\mu}, \mathbf{\Sigma}_1)$ and second of sizes n_2 from $\mathcal{N}(\vec{\mu}, \mathbf{\Sigma}_2)$, then denoting by $\overline{x}_1$, $\mathbf{S}_1$ and $\overline{x}_2$, $\mathbf{S}_2$ the respective sample averages and covariances
  - $$(\overline{x}_1 - \overline{x}_2)^T (\frac{1}{n_1}\mathbf{\Sigma}_1 + \frac{1}{n_2}\mathbf{\Sigma}_2)^{-1} (\overline{x}_1 - \overline{x}_2) \sim \chi^2_p,$$
  - if $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$
  $$(\overline{x}_1 - \overline{x}_2)^T \left(\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{S}_{\text{pooled}}\right)^{-1} (\overline{x}_1 - \overline{x}_2) \sim \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p,n_1+n_2-p-1},$$
  where
  $$\mathbf{S}_{\text{pooled}} = \frac{n_1 - 1}{n_1 + n_2 - 2}\mathbf{S}_1 + \frac{n_2 - 1}{n_1 + n_2 - 2}\mathbf{S}_2$$
  - if $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$ and n is large, then approximately
  $$(\overline{x}_1 - \overline{x}_2)^T (\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2)^{-1} (\overline{x}_1 - \overline{x}_2) \sim \chi^2_p.$$

A blood sample is meant to be tested for two indices. It is sent to two laboratories, so that they perform the tests independently. It is known that the resulting bivariate measurement from both laboratories is normally distributed and in expectation equal to the correct value (i.e. both laboratories are unbiased). However, the covariances of the two laboratories differ. For the first laboratory it is
$$\mathbf{\Sigma}_1 = \begin{bmatrix} 4 & 0.05 \\ 0.05 & 2 \end{bmatrix}$$
while for the second it is
$$\mathbf{\Sigma}_1 = \begin{bmatrix} 8 & 4 \\ 4 & 8 \end{bmatrix}.$$
The results of the sample from the first laboratory are $[10, 2]$, while from the second $[20, 12]$.
(a 3p) Perform a test at the 5% significance level if the two returned measurements, are actually measuring the same sample, or should you suspect some problem somewhere? Justify your choice of test. Does the test accept or reject the hypothesis?
(b 3p) Which of the two returned measurements would you trust more and why? Which laboratory seems to be better and why?