Student Number 999292509

Name Qiu Rui
SURNAME GIVEN NAMES
(Please print in block letters)

Faculty or School of Registration

BOOK NO:
TOTAL NUMBER OF BOOKS USED

Course STA437
(e.g. PHL 202S, ANTAOIY)

Instructor

Date of Examination

Place of Examination

TERM

INSTRUCTIONS

Write the information sought in the spaces above.

Write the answers on the RULED SIDE ONLY; all calculations or rough drafts of answers should be shown, preferably on the unruled side.

Clearly identify the question to which each answer applies; whenever the answer to a question is divided, note at the end of the first section "see also work on page ."

If a page is left blank write on it "see work on page ."

If more than one book is used, indicate the total number on the cover of each. At the conclusion of the examination, place all other books inside Book No. 1.

Do not tear any paper out of this book.

EXAMINER'S REPORT

1	15
2	0
3	13+4
4	8
5	
6	
7	
8	
9	
10	
11	
12	
Total	40

THIS BOOK MUST NOT BE TAKEN FROM THE EXAMINATION ROOM

University of Toronto

Q1.

(a). The joint ~~distribu~~tion of $(X_1, X_2, X_3)^T$ is a normal distribution with $\underset{\sim}{\mu}' = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$

and $C' = \begin{pmatrix} 55 & 7 & -5 \\ 7 & 59 & -13 \\ -5 & -13 & 19 \end{pmatrix}$

This is because the one-dimension projections of a multi-dimensional Normal distribution ~~is~~ are always normal, and joint of one-dimensional Normal is also Normal.

(b). Then the sum of Normal distributions is Normal.

$\mu'' = \mu_4 + \mu_5 = 0$

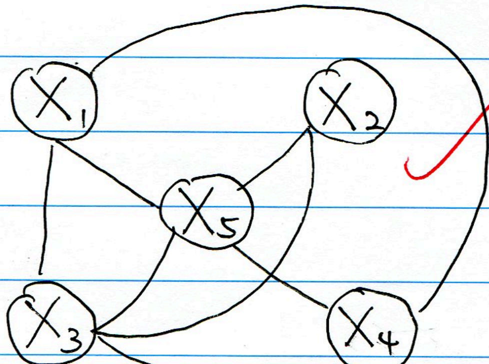
$\sigma^2 = (1, 1) \begin{pmatrix} 55 & -6 \\ -6 & 60 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = (55-6 \quad -6+60) \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

$= 55-6-6+60$

$= 103$

so $\underline{X_4 + X_5 \sim N(0, 103)}$

(c). By the information given in concentration matrix K, $\sigma_{12} = \sigma_{24} = 0$, so there is no edge between X_1 and X_2, X_2 and X_4.



Q2

(a). It's true because Covariance matrix is non-negative definite.

(b). It's true. We can prove for 1-dim normal case:

$x_1 \sim N(0, \sigma_1^2)$, $x_2 \sim N(0, \sigma_2^2)$, $x_1 \perp x_2$.

$$\mu_{1-2} = \mu_1 - \mu_2 = 0, \quad Var(x_1 - x_2) = E[(x_1-x_2)^2] - [E(x_1-x_2)]^2$$
$$= E[x_1^2 + x_2^2 - 2x_1x_2] - 0^2$$
$$= E(x_1^2) + E(x_2^2) - \underbrace{2E(x_1x_2)}_{=0 \text{ b/c } x_1 \perp x_2}$$
$$= Var(x_1) + E(x_1)^2$$
$$+ Var(x_2) + E(x_2)^2$$
$$= \sigma_1^2 + \sigma_2^2.$$

Similarly, we can show for $\underset{\sim}{X_1} \sim N_p(\underset{\sim}{0}, C_1)$, $\underset{\sim}{X_2} \sim N_p(\underset{\sim}{0}, C_2)$.

(c). It's probably false. (but I'm not sure).

Possible modifications:

We need some constraints for $\underset{\sim}{a}$, e.g. $\|\underset{\sim}{a}\|^2 = 1$

or e.g. $\underset{\sim}{a} \neq \underset{\sim}{0}$ (otherwise it's meaningless)

(d). It's true because on each dimension there is 1 degree of freedom, so in total there are p degrees of freedom.

Q3.

(a). Since $\rho_{24} = \rho_{42} = 0.78$ is the largest among all values in correlation matrix $\hat{R}$, so variable x_1 and x_4 are most highly correlated.

0

(b). We know that the standard deviations are square roots of eigenvalues, and sum of eigenvalues equals to the number of variables in PCA.

In this case, $1.6336942^2 + A^2 + 0.59706982^2 + 0.41051087^2 = 4$

so $A \doteq 0.898$

For loadings, we know sum of squares of each loading is 1.

so $(-0.360)^2 + (-0.547)^2 + B^2 + (-0.529)^2 = 1$

$B \doteq \pm 0.540$

And we can determine the sign of B by the fact that PC loadings are orthogonal vectors:

$\cancel{(-0.36 \times 0.183) + (-0.547 \times 0.753) + B \times (-0.421) + (-0.529 \times -0.472) = 1}$

~~So B =~~

$-0.36 \times 0.239 + B \times (-0.726) + (-0.529 \times 0.642) = 1$

$B \doteq -1.964$ (the value differs a lot from previous calculation, due to rounding errors. Besides, the variance of 4th PC is quite large as well).

so the sign of B is negative

Therefore, $B \doteq -0.540$.

4+4

(c).

Since $y_1, \cdots, y_{100}$ are already standardized, we use the transpose of loadings vectors to matrix multiple the observation so that we get a PC score as the result.

For example, for observation $y_1 = (y_{11}, y_{12}\ y_{13}\ y_{14})^T$

the 1st PC score

$$= (-0.360\quad -0.547\quad -0.540\quad -0.529)\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \end{pmatrix} = \begin{matrix} -0.36y_{11} - 0.547y_{12} \\ -0.540y_{13} \\ -0.529y_{14} \end{matrix}$$

the 2nd PC score

$$= (0.883\quad -0.362\quad 0\quad -0.292)\begin{pmatrix} y_{11} \\ y_{2} \\ y_{3} \\ y_{14} \end{pmatrix}$$

$$= 0.883y_{11} - 0.362y_{12} - 0.292y_{14}$$

the 3rd PC score

$$= 0.239y_{11} - 0.726y_{13} + 0.642y_{14}$$

the 4th PC score

$$= 0.183y_{11} + 0.753y_{12} - 0.421y_{13} - 0.472y_{14}$$

Similarly, we can substitute any y_i as previous y_1 and get the four PC scores.

5

(d). First of all, it's a 4x4 matrix with $\text{diag}(R) = \begin{pmatrix} 1 & & 0 \\ & 1 & \\ & & 1 \\ 0 & & & 1 \end{pmatrix}$

Then since the PCs are orthogonal,

$\text{cov}(PC_i, PC_j) = 0,\ \forall i \neq j$

so that $\text{corr}(PC_i, PC_j) = 0,\ \forall i \neq j$.

Therefore the correlation matrix of PC scores is identity matrix $I = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$

4

4.

(b). Since the goal of PCA is to lower the dimensionality of data, (to ~~avoid~~ avoid "curse of dimentionality"), we transform our old coordinate system into a new coordinate system with ~~so~~ lower "rank". Each new coordinate is a principal component. (PC).

Generally, the first 2 PC's contain the most information about our data, with highest and second highest variance.

So usually, instead of drawing scatterplots of lots of pairs of variables or 3-d scatterplots, we pick those 1st and 2nd PCs ~~[illegible]~~ to draw the scatterplot, and this is called a biplot which tells a lot about our data.

Hence, we usually do PCA on data ("princomp()" in R)

Find the first 2 PCs and use ~~"[illegible]~~ R to draw their pairwise scatterplot, whose xlab is usually PC score 1 and ylab is PC score 2.