Nanyang Technological University, SPMS, MAS

MH4501 Multivariate Analysis, Semester 2 AY 2022-23

Assignment 3 (Due date: 4:30pm, 23 March 2023)

Remark: NTU places very high importance on honesty in academic work submitted by students, and adopts a policy of **ZERO** tolerance on cheating and plagiarism.

1. Suppose that $X = (X_1, X_2, X_3, X_4)^\top$ is a random vector with the covariance matrix
$$\Sigma = \begin{pmatrix} 5 & 3 & 2 & 1 \\ 3 & 5 & 1 & 2 \\ 2 & 1 & 5 & 3 \\ 1 & 2 & 3 & 5 \end{pmatrix}.$$

   (a) Without the computer's aid, verify that the principal components for this covariance matrix are
$$\begin{aligned} Y_1 &= \frac{1}{2}(X_1 + X_2 + X_3 + X_4), \quad Y_2 = \frac{1}{2}(X_1 + X_2 - X_3 - X_4), \\ Y_3 &= \frac{1}{2}(X_1 - X_2 + X_3 - X_4), \quad Y_4 = \frac{1}{2}(X_1 - X_2 - X_3 + X_4). \end{aligned}$$

   (b) Determine the least number of principal components that can account for at least 85% of the total variance.

   (c) Compute the correlation coefficients between Y_3 and $X_k, k = 1, 2, 3, 4$.

2. The eigenvalues and eigenvectors of the covariance matrix for three standardized random variables Z_1, Z_2, and Z_3 are
$$\begin{aligned} \lambda_1 &= 1.96 \qquad e_1' = (0.625, 0.593, 0.507), \\ \lambda_2 &= 0.68 \qquad e_2' = (-0.219, -0.491, 0.843), \\ \lambda_3 &= 0.36 \qquad e_3' = (0.749, -0.638, -0.177). \end{aligned}$$

   (a) Assuming an $m = 1$ factor model for (Z_1, Z_2, Z_3), calculate the loading matrix L and matrix of specific variances Ψ using the principal component solution method.

   (b) Calculate communalities $h_i^2,\ i = 1, 2, 3$.

   (c) What proportion of the total population variance is explained by the first common factor?

   (d) Calculate $\mathrm{Corr}(Z_i, F_1)$ for $i = 1, 2, 3$. Which variable might carry the greatest weight in "naming" the common factor?

3. The following table shows the estimates of factor loadings and specific variances before and after rotation, based on the correlation matrix of the stock return data. Assume that l_{32}, l_{52}, and l^*_{ij} are positive and two-factor model is assumed. Calculate the unknown values in the table.

Variable	Factor loading $\tilde{l}_{ij}$ F_1	F_2	Communalities $\tilde{h}_i^2$	Specific variances $\tilde{\psi}_i$
1. J.P. Morgan	0.115	0.755	$\tilde{h}_1^2$	$\tilde{\psi}_1$
2. Citibank	0.322	0.788	$\tilde{h}_2^2$	$\tilde{\psi}_2$
3. Wells Fargo	0.182	l_{32}	$\tilde{h}_3^2$	$\tilde{\psi}_3$
4. Toyal Dutch Shell	1.000	0	$\tilde{h}_4^2$	$\tilde{\psi}_4$
5. ExonMobil	0.683	l_{52}	$\tilde{h}_5^2$	$\tilde{\psi}_5$
Cumulative Proportion of total sample variance explained	p_1	p_2		
Variable	Rotated factor loading $\tilde{l}^*_{ij}$ F_1^*	F_2^*	Communalities $\tilde{h}_i^{*2}$	Specific variances $\tilde{\psi}_i^*$
1. J.P. Morgan	l^*_{11}	0.024	$\tilde{h}_1^{*2}$	$\tilde{\psi}_1^*$
2. Citibank	l^*_{21}	0.227	$\tilde{h}_2^{*2}$	$\tilde{\psi}_2^*$
3. Wells Fargo	l^*_{31}	0.104	$\tilde{h}_3^{*2}$	$\tilde{\psi}_3^*$
4. Toyal Dutch Shell	0.118	l^*_{42}	$\tilde{h}_4^{*2}$	$\tilde{\psi}_4^*$
5. ExonMobil	0.113	l^*_{52}	$\tilde{h}_5^{*2}$	$\tilde{\psi}_5^*$
Cumulative Proportion of total sample variance explained	0.352	0.653		

4. Use R and the dataset "iris" (attached with this assignment) to do the following principal component analysis step by step. For part (a) and (b), just need to write down the R codes. For part (c) and (d), answer based on the R results.

   (a) Delete the column containing the species type and use "**newiris**" for this new dataset.

   (b) Perform a principal component analysis on the dataset **newiris** created above with "scale=TRUE". Save your PCA results in "**irpc**". Print "**irpc**" to see the output.

   (c) Compare output in "**irpc**" with "**eigen(cor(newiris))**". What do you notice?

   (d) Write down the first two PCs and calculate the proportion of variance explained by PC1 and PC2, respectively according to the output in "**irpc**".

[Optional Survey] How do you rate the difficulty of this assignment (1=easy, 5=hard)? Any feedback about lectures, assignments, tutorials, and labs so far?