Final Exam

จัดทำโดย

48850226 นายวฐา มินเสน

49850126 นายชวการ ธรแพทย์

49850159 นายสุเทพ วรรณะศิลปิน

Note: เอกสารต้นฉบับ , ไฟล์ Project จากโปรแกรม Minitab และ การไฟล์โปรแกรมช่วยคำนวณ MATLAB สามารถ Download ได้จาก http://beam.to/statistics

Exercise 8.13. In the radiotherapy data listed in Table 1.7 (see also the radiotherapy data on the CD-ROM), the n = 98 observations on p = 6 variables represent patients' reactions to radiotherapy.

(a) Obtain the covariance and correlation matrices S and R for these data.

$$S = \begin{bmatrix} 4.6548 & 0.9313 & 0.5897 & 0.2769 & 1.0749 & 0.1582 \\ 0.9313 & 0.6128 & 0.1109 & 0.1185 & 0.3889 & -0.0249 \\ 0.5897 & 0.1109 & 0.5714 & 0.0870 & 0.3480 & 0.1101 \\ 0.2769 & 0.1185 & 0.0870 & 0.1104 & 0.2174 & 0.0218 \\ 1.0749 & 0.3889 & 0.3480 & 0.2174 & 0.8622 & -0.0088 \\ 0.1582 & -0.0249 & 0.1101 & 0.0218 & -0.0088 & 0.8615 \end{bmatrix}$$

$$R = \begin{bmatrix} 1 & 0.5514 & 0.3616 & 0.3863 & 0.5366 & 0.079 \\ 0.5514 & 1 & 0.1875 & 0.4554 & 0.5350 & -0.0342 \\ 0.3616 & 0.1875 & 1 & 0.3464 & 0.4958 & 0.1570 \\ 0.3863 & 0.4554 & 0.3464 & 1 & 0.7046 & 0.0707 \\ 0.5366 & 0.5350 & 0.4958 & 0.7046 & 1 & -0.0102 \\ 0.079 & -0.0342 & 0.1570 & 0.0707 & -0.0102 & 1 \end{bmatrix}$$

(b) Pick one of the matrices S or R (justify your choice), and determine the eigenvalues and eigenvectors. Prepare a table showing, in decreasing order of size, the percent that each eigenvalue contributes to the total sample variance.

Justify your choice

การตัดสินใจเลือกใช้แบบใดนั้นขึ้นอยู่กับเหตุผลดังนี้ (ประกอบการให้เหตุผลจาก Richard A. Johnson and Dean W. Wichern, Applied multivariate statistics analysis, fifth edition, page 435.)

"Variables should probably be standardized if they are measured on scales with widely differing ranges or if the units of measurement are not commensurate. For example, if x_1 represents annual sales in the the \$10,000 to \$350,000 range and x_2 is the ratio (net annual income)/(total assets) that falls in the .01 to .60 range, then the total variation will be due almost exclusively to dollar sales. In this case, we would expect a single (important) principal component with a heavy weighting of x_1. Alternatively, if both variables are standardized, their subsequent magnitudes will be of the same order, and x_2 (or z_2) will play a larger role in the construction of the principal components."

ดังนั้นเมื่อพิจารณาข้อมูลใน Table 1.7 พบว่า ตัวแปรทั้ง 6 ตัวมี Scale ดังนี้

The data consist of average ratings over the course of treatment for patients undergoing radiotherapy. Variables measured include x_1 (number of symptoms, such as sore throat or nausea, on a ≥ 0 **scale**); x_2 (amount of activity, on a **1-5 scale**); x_3 (amount of sleep, on a **1-5 scale**); x_4 (amount of food consumed, on a **1-3 scale**); x_5 (appetite consumed, on a **1-5 scale**); x_6 (skin reaction, on a **0-3 scale**).

จาก Scale ของตัวแปรทั้ง 6 ข้างต้น ตัวแปรมี Scale แตกต่างกันโดยตัวแปร x_1 มีขนาดกว้างกว่าตัวแปรอื่น ดังนั้นการตัดสินใจใช้การพิจารณาหาค่า The sample principal components จาก **Sample correlation matrix (R)** จึงมีความเหมาะสมกว่าการหา The sample principal components จาก Sample variance-covariance matrix (**S**)

Eigenvalues and eigenvectors

Eigenvalues analysis of the Correlation Matrix

$\lambda_1 = 2.8643, \lambda_2 = 1.0764, \lambda_3 = 0.7776, \lambda_4 = 0.6503, \lambda_5 = 0.3880$ and $\lambda_6 = 0.2433$

$$e_1 = \begin{bmatrix} 0.445 \\ 0.429 \\ 0.359 \\ 0.463 \\ 0.521 \\ 0.056 \end{bmatrix}, e_2 = \begin{bmatrix} -0.027 \\ -0.292 \\ 0.380 \\ -0.021 \\ -0.074 \\ 0.874 \end{bmatrix}, e_3 = \begin{bmatrix} 0.339 \\ 0.499 \\ -0.628 \\ -0.125 \\ -0.203 \\ 0.430 \end{bmatrix}, e_4 = \begin{bmatrix} 0.551 \\ 0.061 \\ 0.421 \\ -0.666 \\ -0.201 \\ -0.179 \end{bmatrix}, e_5 = \begin{bmatrix} 0.601 \\ -0.687 \\ -0.332 \\ 0.207 \\ 0.103 \\ -0.053 \end{bmatrix}, e_6 = \begin{bmatrix} 0.146 \\ 0.076 \\ 0.212 \\ 0.533 \\ -0.794 \\ -0.116 \end{bmatrix}$$

The percent that each eigenvalue contributes to the total sample variance

(The proportion of the total sample variance)

$$\text{สูตร} \begin{pmatrix} \text{the proportion of} \\ \text{the total sample variance} \\ \text{explained by } \hat{y}_1 \end{pmatrix} = \frac{\hat{\lambda}_1}{p} = \frac{\hat{\lambda}_1}{tr(R)} = \frac{\hat{\lambda}_1}{\hat{\lambda}_1 + \hat{\lambda}_2} = 0.477 \text{ or } 47.7\%$$

ตารางที่ 1 The proportion of the total sample variance

The proportion of the total sample variance by	Percent (%)	Cumulative percent (%)
$\hat{y}_1 = 0.445z_1 + 0.429z_2 + 0.359z_3 + 0.463z_4 + 0.521z_5 + 0.056z_6$	**47.7**	**47.7**
$\hat{y}_2 = -0.027z_1 - 0.292z_2 + 0.380z_3 - 0.021z_4 - 0.074z_5 + 0.874z_6$	**17.9**	**65.7**
$\hat{y}_3 = 0.339z_1 + 0.499z_2 - 0.628z_3 - 0.125z_4 - 0.203z_5 + 0.430z_6$	**13.0**	**78.6**
$\hat{y}_4 = 0.551z_1 + 0.061z_2 + 0.421z_3 - 0.666z_4 - 0.201z_5 - 0.179z_6$	**10.8**	**89.5**
$\hat{y}_5 = 0.601z_1 - 0.687z_2 - 0.332z_3 + 0.207z_4 + 0.103z_5 - 0.053z_6$	**6.5**	**95.9**
$\hat{y}_6 = 0.146z_1 + 0.076z_2 + 0.212z_3 + 0.533z_4 - 0.794z_5 - 0.116z_6$	**4.1**	**100**

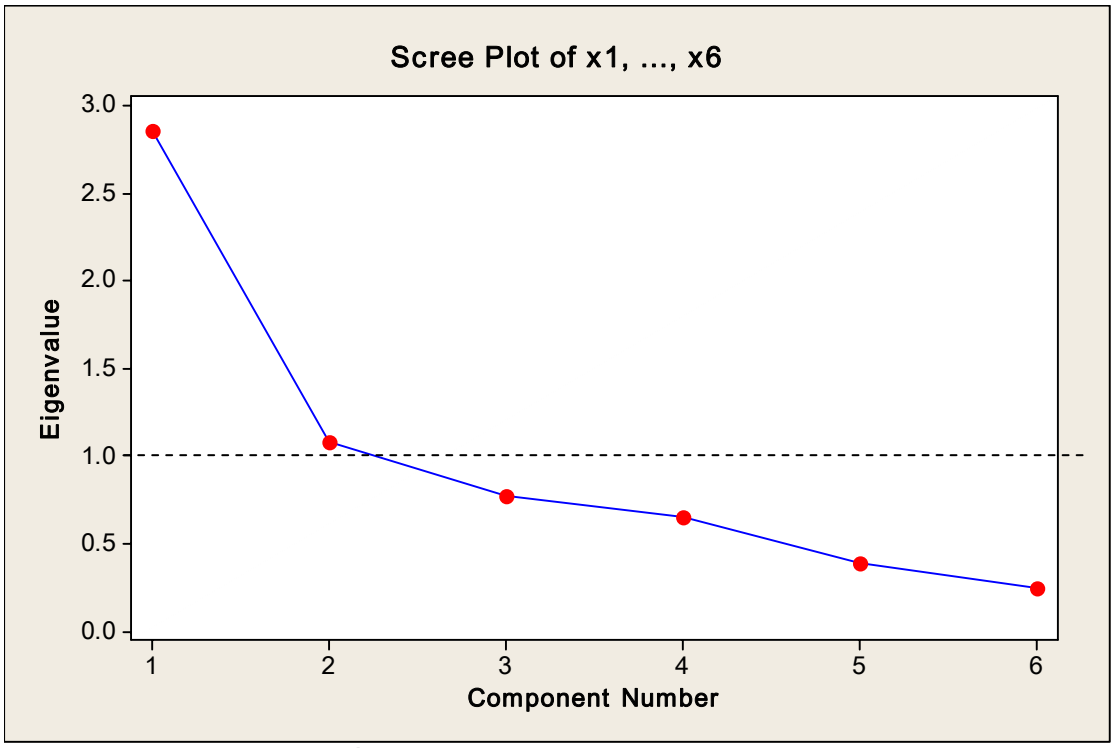
(c) Given the results in Part b, decide on the number of important sample principal components. Is it possible to summarize the radiotherapy data with a single reaction-index component? Explain.

ในการพิจารณาเลือกจำนวน The sample principal components นั้นขึ้นอยู่กับผู้วิจัยที่จะนำไปใช้เป็นสำคัญ อย่างไรก็ตามการตัดสินใจสามารถกระทำได้ด้วยเหตุผลดังนี้ (ประกอบการให้เหตุผลจาก Richard A. Johnson and Dean W. Wichern, Applied multivariate statistics analysis, fifth edition, page 429.)

"It most **(for instance, 80% to 90%)** of the total population variance, for large *p*, can be attributed to the first one, two, or three components, then these components can "replace" the original *p* variables **without much loss of information**."

และเหตุผลดังนี้(ประกอบการให้เหตุผลจาก Richard A. Johnson and Dean W. Wichern, Applied multivariate statistics analysis, fifth edition, page 440.)

"Things to consider include the amount of total sample variance explained, the relative sizes of the eigenvalues(the variances of the sample components), and the subject-matter interpretations of the components. In addition, as we discuss later, a component associated with an eigenvalue **near zero** and, hence, **deemed unimportant**, may indicate an unsuspected linear dependency in the data."



รูปที่ 1 แสดง Scree Plot of each Component

ดังนั้นการตัดสินใจจึงขอยึดตามหลักเหตุผลข้างต้นเป็นเกณฑ์ จากตารางที่ 1 Part b จึงสรุปได้ว่าจำนวนของ The sample principal components นั้น**<u>ควรจะเป็น 3 หรือ 4 components (3 components อธิบายได้ 78.6%, 4 components อธิบายได้ 89.5%)</u>** เพื่อให้ข้อมูลศูนย์เสียไปน้อยและความผิดพลาดที่จะได้รับเมื่อนำไปใช้ต่อไปอยู่ในเกณฑ์ต่ำ

แต่การตัดสินใจนี้ เมื่อพิจารณาใช้จำนวน The sample principal components ของตัวแปร $\hat{y}_i =$ 3 หรือ 4 ตัว มีจำนวนมากจนเกือบเท่ากับจำนวนตัวแปรตั้งต้นเดิม $x_i =$ 6 ตัว ผู้วิจัยอาจจะพิจารณายอมศูนย์เสียข้อมูลบางส่วนไปเพื่อลดจำนวนของตัวแปร $\hat{y}_i$ ได้ จากการพิจารณารูปที่ 1 ก็มีเหตุผลเพียงพอที่จะใช้จำนวน 2 components ทั้งนี้เนื่องจาก $\lambda_3, \lambda_4, \lambda_5$ and λ_6 มีค่าน้อยกว่า 1 หรือกล่าวได้ว่าเข้าใกล้ 0 การใช้ 2 components นี้จะอธิบายได้เพียง 65.7% ที่ระดับนี้ผู้วิจัยจะต้องยอมรับข้อผิดพลาดจำนวนมากที่อาจจะเกิดขึ้นได้

Is it possible to summarize the radiotherapy data with a single reaction-index component? Explain.

No.

การใช้เพียง a single reaction-index component นั้นไม่สมควรเป็นอย่างยิ่งทั้งนี้เนื่องจากการใช้ The first sample principal component

$$\hat{y}_1 = 0.445z_1 + 0.429z_2 + 0.359z_3 + 0.463z_4 + 0.521z_5 + 0.056z_6$$

จะอธิบายได้เพียง 47.7% (จากตารางที่ 1) ยังไม่ถึง 50% ซึ่งถือว่ามีค่าน้อยเกินกว่าจะนำไปใช้ในการอธิบายตัวแปรตั้งต้นเดิม x_i จำนวน 6 ตัว การนำไปใช้จะมีความน่าเชื่อต่ำ และมีความผิดพลาดสูง

ถ้าในกรณีที่ผู้วิจัยมีจำเป็นที่จะใช้จำนวนตัวแปร (The sample principal components) น้อยๆ การเลือกตัว 2 Components คือ The sample principal components ลำดับที่ 1 และ 2 จะมีความเหมาะสมกว่า โดยที่การใช้ 2 components นี้จะอธิบายได้ 65.7% ซึ่งมากกว่า 47.7% อยู่มากพอสมควร

Note: ที่ระดับ 65.7% นี้ ผู้วิจัยจะต้องยอมรับข้อผิดพลาดจำนวนมากที่อาจจะเกิดขึ้นได้

(d) Prepare a table of the correlation coefficients between each principal component you decide to retain and the original variables. If possible, interpret the components.

สูตรคำนวณ $r_{\hat{y}_i, z_k} = \hat{e}_{ik}\sqrt{\hat{\lambda}_i}$; i, k = 1, 2, …, 6

ตารางที่ 2 the correlation coefficients between each principal component and the original variables

$r_{\hat{y}_i, z_k}$	z_1	z_2	z_3	z_4	z_5	z_6	
$\hat{y}_1$	0.75289	0.72656	0.6072	0.78335	0.88222	0.094568	ตัดสินใจ
$\hat{y}_2$	-0.02766	-0.30268	0.3944	-0.021745	-0.076455	0.90675	
$\hat{y}_3$	0.29923	0.43969	-0.55393	-0.10986	-0.17931	0.37909	
$\hat{y}_4$	0.44446	0.049488	0.33955	-0.53676	-0.16171	-0.14412	
$\hat{y}_5$	0.37428	-0.42813	-0.20671	0.1292	0.06427	-0.033071	
$\hat{y}_6$	0.072251	0.037685	0.10438	0.26273	-0.39167	-0.057341	

เมื่อตัดสินใจที่ระดับการอธิบายได้ 89.5% ซึ่งใช้ $\hat{y}_i$ จำนวน 4 components ดังนี้

$$\hat{y}_1 = 0.445z_1 + 0.429z_2 + 0.359z_3 + 0.463z_4 + 0.521z_5 + 0.056z_6$$
$$\hat{y}_2 = -0.027z_1 - 0.292z_2 + 0.380z_3 - 0.021z_4 - 0.074z_5 + 0.874z_6$$
$$\hat{y}_3 = 0.339z_1 + 0.499z_2 - 0.628z_3 - 0.125z_4 - 0.203z_5 + 0.430z_6$$
$$\hat{y}_4 = 0.551z_1 + 0.061z_2 + 0.421z_3 - 0.666z_4 - 0.201z_5 - 0.179z_6$$

Interpret correlation coefficients

การพิจารณา $r_{\hat{y}_i, z_k}$ ในสอดคล้องกับ coefficient แต่ละตัวแปร z_k ของแต่ละ the sample principal components นั้นคือถ้าระดับความสัมพันธ์มาก($r_{\hat{y}_i, z_k}$ มาก) coefficient ก็จะมากด้วยเมื่อเทียบกับตัวอื่นๆ ใน the sample principal components เดียวกัน

เช่น ที่ $\hat{y}_1 = 0.445z_1 + 0.429z_2 + 0.359z_3 + 0.463z_4 + 0.521z_5 + 0.056z_6$

จากสมการดังกล่าวนี้จะเห็นว่า coefficient $z_6 = 0.056$ มีค่าน้อยเมื่อพิจารณา $r_{\hat{y}_1, z_6} = 0.094568$ ก็มีค่าความสัมพันธ์กับ $\hat{y}_1$ น้อยมากๆ ส่วนตัวแปร z_1 ถึง z_5 มีค่า coefficient สูง และค่า correlation จึงสูงตามไปด้วยซึ่งมีค่ามากกว่า .6

ที่ $\hat{y}_2 = -0.027z_1 - 0.292z_2 + 0.380z_3 - 0.021z_4 - 0.074z_5 + 0.874z_6$

z_1, z_4 และ z_5 มีค่า coefficient และ ค่า correlation น้อยเมื่อเทียบกับ z_2, z_3 และ z_6

การพิจารณา components อื่นๆ ก็ทำเช่นเดียวกัน

Interpret the components

ไม่ใช่เรื่องง่ายในการอธิบายการเกิดขึ้นของ components เนื่องจากผู้อธิบายจำเป็นต้องทราบถึงงานที่ต้องการวิเคราะห์เป็นอย่างดี ถึงจะให้เหตุผลของการเกิดขึ้นของ components ที่พิจารณาได้ แต่ทางกลุ่มจะอธิบายตามสมควร จากเรื่องที่โจทย์ศึกษาไว้ดังนี้ (อ้างอิงการอธิบายจาก: Richard A. Johnson and Dean W. Wichern, Applied multivariate statistics analysis, fifth edition, page 448.)

The first component มีน้ำหนัก coefficient ใกล้เคียงกัน(ยกเว้น z_6) และเป็นเครื่องหมายเดียวกัน $\hat{y}_1 = 0.445z_1 + 0.429z_2 + 0.359z_3 + 0.463z_4 + 0.521z_5 + 0.056z_6$ จึงเป็นไปได้ที่จะเรียก component นี้ว่า General the course of treatment for patients undergoing radiotherapy component, or simply a patients undergoing radiotherapy data.

The second component มีกลุ่มที่เครื่องหมายแตกต่างกันไป 2 กลุ่ม

$\hat{y}_2 = -0.027z_1 - 0.292z_2 + 0.380z_3 - 0.021z_4 - 0.074z_5 + 0.874z_6$

คือ กลุ่ม x_1 (number of symptoms, such as sore throat or nausea, on a ≥ 0 scale); x_2 (amount of activity, on a 1-5 scale); x_4 (amount of food consumed, on a 1-3 scale)และ x_5 (appetite consumed, on a 1-5 scale) กับกลุ่ม x_3 (amount of sleep, on a 1-5 scale) และ x_6 (skin reaction, on a 0-3 scale). จะเห็นว่าเป็นไปได้ยากในการสรุปผล component นี้

The third component มีกลุ่มที่เครื่องหมายแตกต่างกันไป 2 กลุ่ม

$\hat{y}_3 = 0.339z_1 + 0.499z_2 - 0.628z_3 - 0.125z_4 - 0.203z_5 + 0.430z_6$

x_1, x_2 และ x_6 กับกลุ่ม x_3, x_4 และ x_5 จะเห็นว่ากลุ่มแรกเป็นกลุ่มแสดงถึงอาการของโรคเป็นสำคัญ ส่วนอีกกลุ่มที่เป็น negative เป็นกลุ่มแสดงอาการของพฤติกรรมทั่วไป คือการกระหาย การกิน และการนอน อาจเป็นไปได้ว่าจะเรียก component นี้ว่า Activities of patients undergoing radiotherapy component.

The fourth component มีกลุ่มที่เครื่องหมายแตกต่างกันไป 2 กลุ่ม

$\hat{y}_4 = 0.551z_1 + 0.061z_2 + 0.421z_3 - 0.666z_4 - 0.201z_5 - 0.179z_6$

x_1, x_2 และ x_3 กับกลุ่ม x_4, x_5 และ x_6 จะเห็นว่าเป็นไปได้ยากในการสรุปผล component นี้

Exercises 9.20. Using the air-pollution variables X_1, X_2, X_5, and X_6 given in Table 1.5, generate the sample covariance matrix.

$$S = \begin{bmatrix} 2.5 & -2.78049 & -0.58537 & -2.23171 \\ -2.78049 & 300.51568 & 6.76307 & 30.79094 \\ -0.58537 & 6.76307 & 11.36353 & 3.12660 \\ -2.23171 & 30.79094 & 3.12660 & 30.97851 \end{bmatrix}$$

(a) Obtain the principal component solution to a factor model with m = 1 and m = 2.

ตารางที่ 3 เมื่อใช้ **the sample covariance matrix** คำนวณ

Variable	Estimated factor loadings $\tilde{l}_{ij} = \sqrt{\hat{\lambda}_i}\hat{e}_{ij}$		Communalities $\tilde{h}_i^2$	Specific variances $\tilde{\Psi}_i = 1 - \tilde{h}_i^2$
	F_1	F_2		
Wind (x_1)	0.175	0.405	0.194	0.8060
Solar radiation (x_2)	**-17.325**	0.609	300.515	**-299.5150**
$NO_2(x_5)$	-0.421	-0.742	0.728	0.2720
$O_3(x_6)$	**-1.959**	**-5.187**	30.739	**-29.7390**
Eigenvalues	304.19	27.99		
Cumulative proportion of total sample variance	0.881	0.962		

Comment:

การตัดสินใจ Common factors ในที่นี้มี Common factor 1 ตัว(m=1) Cumulative proportion of total sample variance = 88.1% และถ้า Common factors 2 ตัว(m=2) Cumulative proportion of total sample variance = 96.2%

แต่ จากหนังสือ Richard A. Johnson and Dean W. Wichern, Applied multivariate statistics analysis, fifth edition, page 482. ในตัวอย่างที่ 9.2 **Nonexisence of proper solution** กล่าวโดยสรุปไว้ว่าในกรณีนี้ที่ Ψ_i เป็น negative values เราจะไม่สนใจการแก้ปัญหาด้วย Variance-covariance Matrix และมันเป็นผลทำให้การนำคำตอบเหล่านี้ไปใช้นั้น ไม่เหมาะสม (หนังสืออธิบายอีกครั้งใน Suppement 9A.)

$$S = \begin{bmatrix} 2.5 & -2.78049 & -0.58537 & -2.23171 \\ -2.78049 & 300.51568 & 6.76307 & 30.79094 \\ -0.58537 & 6.76307 & 11.36353 & 3.12660 \\ -2.23171 & 30.79094 & 3.12660 & 30.97851 \end{bmatrix}$$

จาก S Matrix เมื่อพิจารณาความแปรปรวน x_2 และ x_6 เห็นว่ามีความแปรปรวนมากกว่าตัวแปร x_1 และ x_5 มากเนื่องจากหน่วยวัด(Scale)(หรือความกว้าง) ของตัวแปรแตกต่างกัน ดังนั้นการจัดทำ Factor Analysis ด้วย R Matrix จึงน่าจะเหมาะสมมากกว่า

ตารางที่ 4 เมื่อใช้ **the sample correlation matrix** คำนวณ

Variable	Estimated factor loadings $\tilde{l}_{ij} = \sqrt{\hat{\lambda}_i}\hat{e}_{ij}$		Communalities $\tilde{h}_i^2$	Specific variances $\tilde{\Psi}_i = 1 - \tilde{h}_i^2$
	F_1	F_2		
Wind (x_1)	0.564	-0.243	0.377	0.6230
Solar radiation (x_2)	-0.645	-0.521	0.688	0.3120
$NO_2(x_5)$	-0.477	0.735	0.768	0.2320
$O_3(x_6)$	-0.771	-0.196	0.633	0.3670
Eigenvalues	1.5556	0.9097		
Cumulative proportion of total (Standardized) sample variance	0.389	0.6163		

การตัดสินใจ Common factors ในที่นี้มี Common factor 1 ตัว(m=1) Cumulative proportion of total sample variance = 38.9% และถ้า Common factors 2 ตัว(m=2) Cumulative proportion of total sample variance = 61.63% จะเห็นว่าการใช้ Common factors 2 ตัว(m=2) จะทำให้ดีขึ้นจากเดิมในระดับที่น่าพอใจ

(b) Find the maximum likelihood estimates of L and Ψ for m = 1 and m = 2.

ตารางที่ 5 เมื่อใช้ **the sample correlation matrix** คำนวณตามที่โปรแกรม Minitab กำหนด โดย m = 1

Variable	Estimated factor loadings $\hat{l}_{ij}$ F_1	Communalities $\hat{h}_i^2$	Specific variances $\hat{\Psi}_i = 1 - \hat{h}_i^2$
Wind (x_1)	**-0.324**	**0.105**	**0.8950**
Solar radiation (x_2)	**0.410**	**0.168**	**0.8320**
$NO_2(x_5)$	**0.232**	**0.054**	**0.9460**
$O_3(x_6)$	**0.771**	**0.595**	**0.4050**
Cumulative proportion of total (Standardized) sample variance	**0.23**		

การตัดสินใจ Common factors ในที่นี้มี Common factor 1 ตัว(m=1) Cumulative proportion of total sample variance = 23.0%

ตารางที่ 6 เมื่อใช้ **the sample correlation matrix** คำนวณตามที่โปรแกรม Minitab กำหนด โดย m = 2

Variable	**Estimated factor loadings** $\hat{l}_{ij}$		**Communalities** $\hat{h}_i^2$	**Specific variances** $\hat{\Psi}_i = 1 - \hat{h}_i^2$
	F_1	F_2		
Wind (x_1)	**-0.101**	**-0.412**	**0.180**	**0.8200**
Solar radiation (x_2)	**1.000**	**-0.000**	**1.000**	**0**
NO_2 (x_5)	**0.116**	**0.241**	**0.071**	**0.9290**
O_3 (x_6)	**0.319**	**0.536**	**0.389**	**0.6110**
Cumulative proportion of total (Standardized) sample variance	**0.281**	**0.410**		

การตัดสินใจ Common factors ในที่นี้มี Common factor 1 ตัว(m=1) Cumulative proportion of total (Standardized) sample variance = 28.1% และถ้า Common factors 2 ตัว(m=2) Cumulative proportion of total (Standardized) sample variance = 41.0% **แต่**…

```
* WARNING * Too many factors, solution is not unique
Factor Analysis: x1, x2, x5, x6
Maximum Likelihood Factor Analysis of the Correlation Matrix
* NOTE * Heywood case
```

เกิดการแจ้งเตือนในการ Run โปรแกรมผ่านโปรแกรม Minitab คือการเตือน * WARNING * Too many factors, solution is not unique และ *** NOTE * Heywood case** จาก Supplement 9A page 532. กล่าวไว้ดังนี้

"It often happens that the objective function in (9A-3) has a relative minimum corresponding to negative values for some $\hat{\Psi}_i$. This solution is clearly inadmissible and is said to be improper, or a **Heywood case."**

จากตารางที่ 6 จะพบว่า **Solar radiation** (x_2) เป็นตัวแปรที่มีโอกาสเกิด **Heywood case** เพราะเป็นไปได้ ที่ค่า $\hat{\Psi}_i$ นั้นติดลบอยู่แม้จะมีค่าติดลบน้อยๆ ก็ตาม ดังนั้นการใช้ Common Factor m=2 จึงเห็นสมควรว่าไม่เหมาะสม

(c) Compare the factorization obtained by the principal component and maximum likelihood methods.

Compare

ใน Part a เราตัดสินใจใช้ Common Factor m=2 ซึ่งได้จากการคำนวณผ่าน the sample correlation matrix ทั้งนี้เนื่องจาก the sample variance-covariance matrix เกิด $\hat{\Psi}_i$ เป็น negative values ดังนั้นนำผลที่ได้จากตารางที่ 4 มาคำนวณ The residual matrix ได้ผลดังนี้

$$R-\tilde{L}\tilde{L}'-\tilde{\Psi}=\begin{bmatrix}1 & -0.101 & -0.110 & -0.254\\ -0.101 & 1 & 0.116 & 0.319\\ -0.110 & 0.116 & 1 & 0.167\\ -0.254 & 0.319 & 0.167 & 1\end{bmatrix}\begin{bmatrix}0.564 & -0.243\\ -0.645 & -0.521\\ -0.477 & 0.735\\ -0.771 & -0.196\end{bmatrix}\begin{bmatrix}0.564 & -0.645 & -0.477 & -0.771\\ -0.243 & -0.521 & 0.735 & -0.196\end{bmatrix}$$

$$-\begin{bmatrix}0.6230 & 0 & 0 & 0\\ 0 & 0.3120 & 0 & 0\\ 0 & 0 & 0.2320 & 0\\ 0 & 0 & 0 & 0.3670\end{bmatrix}$$

The residual matrix of Principal component:

$$=\begin{bmatrix}0 & 0.136 & 0.338 & 0.133\\ 0.136 & 0 & 0.191 & -0.280\\ 0.338 & 0.191 & 0 & -0.057\\ 0.133 & -0.280 & -0.057 & 0\end{bmatrix}$$

ใน Part b เราตัดสินใจใช้ Common Factor m=1 ทั้งนี้เนื่องจากในกรณี m=2 เกิด Heywood case ซึ่งทางกลุ่มยังไม่ทราบวิธีการแก้ปัญหานี้ จึงตัดสินใจเพื่อให้ได้ผลลัพธ์ที่เหมาะสมตามเงื่อนไขของเอกสารวิชาเรียน จึงไม่ใช้ Common Factor m=2 ถึงแม้จะให้ค่า Cumulative proportion of total sample variance = 41.0%

$$R-\hat{L}\hat{L}'-\hat{\Psi}=\begin{bmatrix}1 & -0.101 & -0.110 & -0.254\\ -0.101 & 1 & 0.116 & 0.319\\ -0.110 & 0.116 & 1 & 0.167\\ -0.254 & 0.319 & 0.167 & 1\end{bmatrix}\begin{bmatrix}-0.324\\ 0.410\\ 0.232\\ 0.771\end{bmatrix}\begin{bmatrix}-0.324 & 0.410 & 0.232 & 0.771\end{bmatrix}$$

$$-\begin{bmatrix}0.8950 & 0 & 0 & 0\\ 0 & 0.8320 & 0 & 0\\ 0 & 0 & 0.9460 & 0\\ 0 & 0 & 0 & 0.4050\end{bmatrix}$$

The residual matrix of Maximum likelihood:

$$= \begin{bmatrix} 0 & 0.0314 & -0.0347 & -0.0037 \\ 0.0314 & 0 & 0.0206 & 0.0030 \\ -0.0347 & 0.0206 & 0 & -0.0122 \\ -0.0037 & 0.0030 & -0.0122 & 0 \end{bmatrix}$$

ถึงแม้ว่าค่า Cumulative proportion of total sample variance ของวิธี Principal component มีค่ามาก คือ 61.63% ซึ่ง Cumulative proportion of total sample variance ของวิธี Maximum likelihood มีค่าเพียง 23% ซึ่งผลดังกล่าวนี้เป็นปกติที่วิธี Principal component จะให้ผล Cumulative proportion of total sample variance ที่สูง แต่จากผล The residual matrix Maximum likelihood estimates $\hat{L}$ and $\hat{\Psi}$ ได้ผลลัพธ์**ดีกว่าอย่างชัดเจน** เมื่อเทียบกับ The residual matrix principal component estimates $\tilde{L}$ and $\tilde{\Psi}$ ดังนั้นการใช้ Maximum likelihood m=1 ในโจทย์ปัญหานี้ถือได้ว่าเหมาะสม (การพิจารณาอ้างอิง Example 9.5 และ Example 9.6 หนังสือ Richard A. Johnson and Dean W. Wichern, Applied multivariate statistics analysis, fifth edition, page 493.)