Demonstrating the Mechanics of Principal Component Analysis via Spreadsheets

Abstract

Principal component analysis (PCA) is a popular multivariate statistical method that is used for dimensionality reduction. When teaching PCA in a marketing research or business analytics course, the mechanics of the analysis are often not communicated to the students. Students observe computer output that contains information pertaining to eigenvalues, component loadings, and rotated loadings, yet an understanding of how these numbers were obtained is lacking. This paper presents an Excel workbook that demonstrates the mechanics of PCA, which include (1) the construction of the correlation matrix from the raw data, (2) the extraction of eigenvalues and eigenvectors from the correlation matrix and the computation of the component loadings and component scores, and (3) the rotation of the component loadings to improve interpretability.

Keywords

Spreadsheets, principal component analysis, eigenvalues and eigenvectors, rotation

1. Introduction

Spreadsheets have enormous practical value for communicating statistical methods and concepts to students in the fields of business and economics. For example, Barr and Scott [2, 3] described a variety of spreadsheet-based simulation tools and recently reported a specific application in the context of portfolio construction [4]. Kwan [11, 12] has demonstrated approaches for the shrinkage of covariance or correlation matrices to a prespecified target matrix, which also has important applications to portfolio theory. There are many other types of multivariate statistical procedures that center on the analysis of covariance or correlation matrices, such as factor analysis, canonical correlation, and structural equation modeling. Spreadsheet modeling can be particularly useful for communicating a basic understanding of the mechanics of such procedures to business students.

Principal component analysis (PCA) is a well-known multivariate statistical technique that is used for data reduction and an excellent treatment of the topic is provided by Jolliffe [9]. Given a data matrix, $\mathbf{X} = [x_{ij}]$, consisting of measurements for n observations on p variables, the goal of PCA is to select a few linear combinations (i.e., components) of the variables that explain most of the variation in the full data matrix. PCA is closely related to several other multivariate techniques, such a singular-value-decomposition [5, 8], exploratory factor analysis [16], multidimensional scaling [15], correspondence analysis [7], and biplots [6].

Although widely used in many scientific disciplines, the particular focus herein is on marketing applications. PCA has a rich history in the field of marketing research and its importance as a data reduction tool also extends to business analytics. When taught in a marketing research or business analytics course, PCA is generally implemented using a statistical software package such as SPSS, SAS, or R. Unfortunately, although a business student might leave the course with a reasonable understanding of the output produced by a statistical software package, they seldom have a sufficient grasp of how the results reported in that output were obtained. The goal of this paper is to rectify this problem using a spreadsheet demonstration that highlights the mechanics of PCA, thus providing an original, flexible and educational thrust that the major software packages do not afford.

The Excel workbook for PCA consists of three worksheets. The first worksheet, which is described in Section 2, is used to obtain the correlation matrix from the raw data. The second worksheet, explained in Section 3, is used to extract the eigenvectors from the correlation matrix. This is accomplished using the power method for finding the dominant eigenvector and corresponding eigenvalue of a matrix. Approaches for choosing the number of eigenvectors to retain are also described in this section. The selected eigenvectors and their corresponding eigenvalues are used to compute the principal component loadings (correlations between each variable and each component) and principal component scores for the respondents in the sample are also computed. The third worksheet, described in Section 4, is provided for the rotation of the component loadings to improve their interpretability. Practical experience with the Excel workbook and suggestions for adaptations and extensions are provided in Section 5.

2. Obtaining the Correlation Matrix from the Raw Data

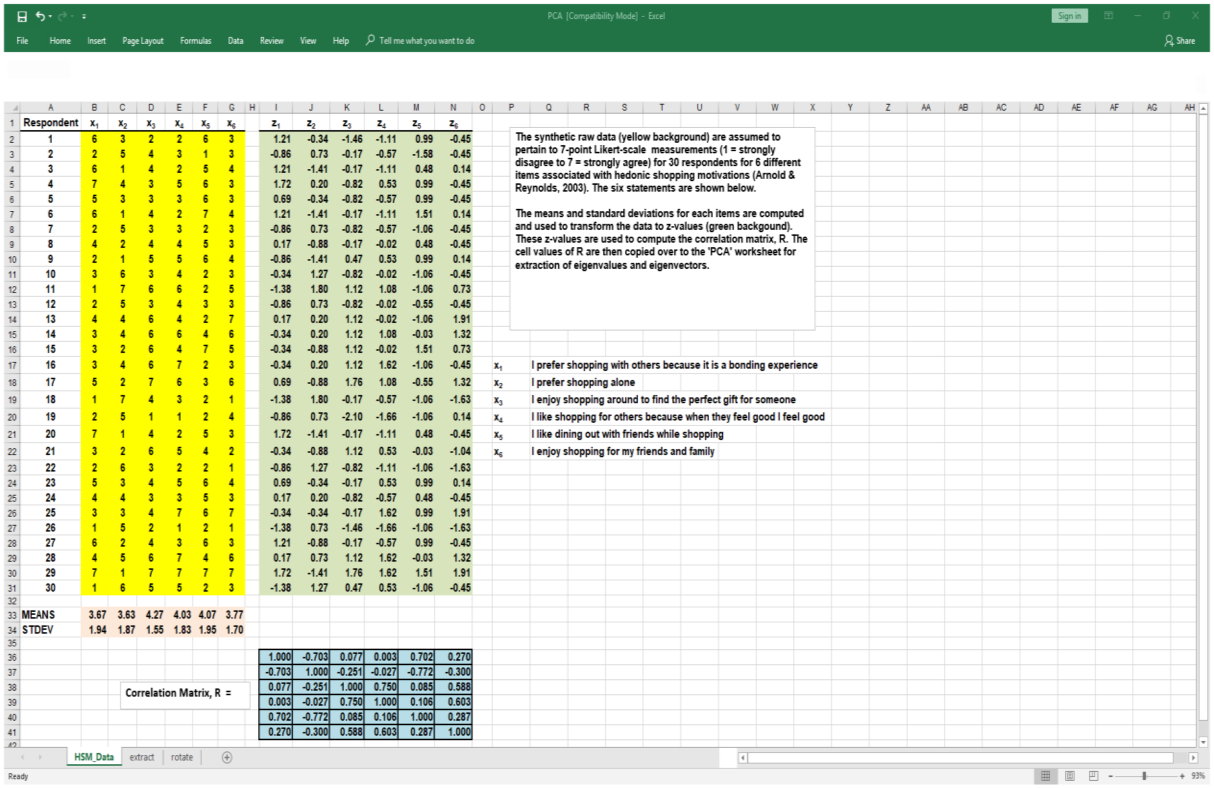
The first worksheet 'HSM_Data' of the Excel workbook PCA is shown in Figure 1. The worksheet contains the raw data in the form of an $n \times p$ data matrix, $\mathbf{X} = [x_{ij}]$. In the example, the data are 7-point Likert-scale measurements for $n = 30$ respondents for each of $p = 6$ variables (or scale item statements) pertaining to hedonic shopping motivations. More specifically, the measurements correspond to each respondent's level of agreement with each of the six statements in Figure 1, where the measurements range from 1= strongly disagree to 7 = strongly agree. The raw data occupy cells B2:C31 and are shaded in yellow in Figure 1. Although the data are synthetic, they are based on actual constructs and questionnaire items associated with an extensive study of hedonic shopping motivations conducted by Arnold and Reynolds [1]. The means ($\bar{x}_j$) and standard deviations (s_j) of each of the variables ($1 \leq j \leq p$) are computed (they are displayed in cells B33:G34 and shaded with a tan background in the worksheet). The means and standard deviations are used to transform the raw data to an $n \times p$ matrix of z-scores, $\mathbf{Z} = [z_{ij}]$, which is accomplished as follows:

$$z_{ij} = (x_{ij} - \bar{x}_j)/s_j, \forall\ 1 \leq i \leq n \text{ and } 1 \leq j \leq p \tag{1}$$

The z-scores, which occupy cells I2:N31 and are shaded in green in Figure 1, are then used to compute the correlation matrix, $\mathbf{R}$, as follows:

$$\mathbf{R} = (1/(n\text{-}1))\mathbf{Z}^{\mathrm{T}}\mathbf{Z}. \tag{2}$$

The correlation matrix (contained in cells I36:N41 and shaded in blue in Figure 1) is computed with the aid of the MMULT function based on the z-score columns. The correlation matrix could have been obtained directly using the Data Analysis Toolpack capabilities of Excel. However, the use of the z-scores and MMULT function to compute $\mathbf{R}$ is concordant with the goals of providing a thorough and detailed presentation of all of the computation aspects of PCA. The correlation matrix $\mathbf{R}$ is copied (cell values only, not the formulas) to the top left corner of the second worksheet ‘extract’.



Respondent	x_1	x_2	x_3	x_4	x_5	x_6		z_1	z_2	z_3	z_4	z_5	z_6
1	6	3	2	2	6	3		1.21	-0.34	-1.46	-1.11	0.99	-0.45
2	2	5	4	3	1	3		-0.86	0.73	-0.17	-0.57	-1.58	-0.45
3	6	1	4	2	5	4		1.21	-1.41	-0.17	-1.11	0.48	0.14
4	7	4	3	5	6	3		1.72	0.20	-0.82	0.53	0.99	-0.45
5	5	3	3	3	6	3		0.69	-0.34	-0.82	-0.57	0.99	-0.45
6	6	1	4	2	7	4		1.21	-1.41	-0.17	-1.11	1.51	0.14
7	2	5	3	3	2	3		-0.86	0.73	-0.82	-0.57	-1.06	-0.45
8	4	2	4	4	5	3		0.17	-0.88	-0.17	-0.02	0.48	-0.45
9	2	1	5	5	6	4		-0.86	-1.41	0.47	0.53	0.99	0.14
10	3	6	3	4	2	3		-0.34	1.27	-0.82	-0.02	-1.06	-0.45
11	1	7	6	6	2	5		-1.38	1.80	1.12	1.08	-1.06	0.73
12	2	5	3	4	3	3		-0.86	0.73	-0.82	-0.02	-0.55	-0.45
13	4	4	6	4	2	7		0.17	0.20	1.12	-0.02	-1.06	1.91
14	3	4	6	6	4	6		-0.34	0.20	1.12	1.08	-0.03	1.32
15	3	2	6	4	7	5		-0.34	-0.88	1.12	-0.02	1.51	0.73
16	3	4	6	7	2	3		-0.34	0.20	1.12	1.62	-1.06	-0.45
17	5	2	7	6	3	6		0.69	-0.88	1.76	1.08	-0.55	1.32
18	1	7	4	3	2	1		-1.38	1.80	-0.17	-0.57	-1.06	-1.63
19	2	5	1	1	2	4		-0.86	0.73	-2.10	-1.66	-1.06	0.14
20	7	1	4	2	5	3		1.72	-1.41	-0.17	-1.11	0.48	-0.45
21	3	2	6	5	4	2		-0.34	-0.88	1.12	0.53	-0.03	-1.04
22	2	6	3	2	2	1		-0.86	1.27	-0.82	-1.11	-1.06	-1.63
23	5	3	4	5	6	4		0.69	-0.34	-0.17	0.53	0.99	0.14
24	4	4	3	3	5	3		0.17	0.20	-0.82	-0.57	0.48	-0.45
25	3	3	4	7	6	7		-0.34	-0.34	-0.17	1.62	0.99	1.91
26	1	5	2	1	2	1		-1.38	0.73	-1.46	-1.66	-1.06	-1.63
27	6	2	4	3	6	3		1.21	-0.88	-0.17	-0.57	0.99	-0.45
28	4	5	6	7	4	6		0.17	0.73	1.12	1.62	-0.03	1.32
29	7	1	7	7	7	7		1.72	-1.41	1.76	1.62	1.51	1.91
30	1	6	5	5	2	3		-1.38	1.27	0.47	0.53	-1.06	-0.45
MEANS	3.67	3.63	4.27	4.03	4.07	3.77							
STDEV	1.94	1.87	1.55	1.83	1.95	1.70							

Correlation Matrix, R =

1.000	-0.703	0.077	0.003	0.702	0.270
-0.703	1.000	-0.251	-0.027	-0.772	-0.300
0.077	-0.251	1.000	0.750	0.085	0.588
0.003	-0.027	0.750	1.000	0.106	0.603
0.702	-0.772	0.085	0.106	1.000	0.287
0.270	-0.300	0.588	0.603	0.287	1.000

The synthetic raw data (yellow background) are assumed to pertain to 7-point Likert-scale measurements (1 = strongly disagree to 7 = strongly agree) for 30 respondents for 6 different items associated with hedonic shopping motivations (Arnold & Reynolds, 2003). The six statements are shown below.

The means and standard deviations for each items are computed and used to transform the data to z-values (green backgound). These z-values are used to compute the correlation matrix, R. The cell values of R are then copied over to the 'PCA' worksheet for extraction of eigenvalues and eigenvectors.

x_1	I prefer shopping with others because it is a bonding experience
x_2	I prefer shopping alone
x_3	I enjoy shopping around to find the perfect gift for someone
x_4	I like shopping for others because when they feel good I feel good
x_5	I like dining out with friends while shopping
x_6	I enjoy shopping for my friends and family

Figure 1: The hedonic shopping motivation data and corresponding correlation matrix.

3. Extraction of the Principal Components

In the 'extract' worksheet displayed in Figure 2, the correlation matrix **R** is in cells A3:F8. The $p \times p$ correlation matrix is a real symmetric positive semidefinite matrix and, therefore has real nonnegative eigenvalues that sum to the trace of **R**. Because the main diagonal elements of the correlation matrix are all one, the trace of **R** is equal to the number of variables, p. Moreover, based on the principles of eigen-decomposition, also known as spectral decomposition, the correlation matrix can be written as a function of its eigenvalues ($\lambda_1,\ldots,\lambda_p$) and eigenvectors ($\mathbf{u}_1,\ldots,\mathbf{u}_p$) as follows:

$$\mathbf{R} = \lambda_1 \mathbf{u}_1 \mathbf{u}_1^T + \lambda_2 \mathbf{u}_2 \mathbf{u}_2^T + \cdots + \lambda_p \mathbf{u}_p \mathbf{u}_p^T = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^T \qquad (3)$$

where $\boldsymbol{\Lambda}$ is a $p \times p$ diagonal matrix of containing the eigenvalues of **R** and $\mathbf{U} = [\mathbf{u}_1,\ldots,\mathbf{u}_p]$ is the $p \times p$ matrix of corresponding eigenvectors. Without loss of generality, we will assume that the eigenvalues are sequenced in nonincreasing order of magnitude (i.e., λ_1

$\geq \lambda_2, \ldots, \geq \lambda_{p-1} \geq \lambda_p$). The eigenvalues are measures of the explained variation in the correlation matrix and, accordingly, the eigenvalue-eigenvector pairs associated with the largest eigenvalues are those that make the greatest contribution to the decomposition of **R**.

Rather than a complete decomposition of **R** as in Equation (3), in PCA we would like a low-rank ($q << p$) approximation that explains the greatest amount of variation in **R**. The underlying optimization problem associated with the first principal component, **u**, is as follows:

$$\text{Maximize: } \mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u} \tag{4}$$

Subject to:

$$\text{Subject to: } \mathbf{u}^{\mathrm{T}}\mathbf{u} = 1 \tag{5}$$

The Lagrangian function associated with the optimization problem is:

$$\text{Maximize: } L = \mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u} - \lambda(\ \mathbf{u}^{\mathrm{T}}\mathbf{u} - 1), \tag{6}$$

The first order condition is:

$$\partial L/\partial \mathbf{u} = 2\mathbf{R}\mathbf{u} - 2\lambda\mathbf{u} = \mathbf{0} \text{ or } \mathbf{R}\mathbf{u} = \lambda\mathbf{u} \tag{7}$$

which is an eigen-structure where **0** is a $p \times 1$ vector of zeros, the Lagrange multiplier λ is the eigenvalue, and **u** is its corresponding eigenvector.

Although we know that there are typically p eigenvalues that solve the determinant polynomial corresponding to Equation (7), we also know from the eigen-decomposition in Equation (3) that it is the largest eigenvalue and its corresponding eigenvector that will explain the most variation in **R**. Therefore, we want to find the largest (or *dominant*) eigenvalue/eigenvector pair associated with Equation (7), that is, $\lambda = \lambda_1$ and $\mathbf{u} = \mathbf{u}_1$. Once this pair is identified, the correlation matrix can be deflated by removing the contribution from λ_1 and $\mathbf{u}_1$ using Equation (3). We use the notation $\mathbf{R}(q)$ to denote the deflated correlation matrix associated with the elimination of variation stemming from the first *q* principal components. After the first eigenvalue/eigenvector pair is extracted, the correlation matrix is deflated via:

$$\mathbf{R}(1) = \mathbf{R} - \lambda \mathbf{u}_1 \mathbf{u}_1^T \qquad (8)$$

The second eigenvalue/eigenvector pair (λ_2, $\mathbf{u}_2$) associated with **R** is the dominant eigenvalue associated with **R**(1) and would then be extracted in similar fashion. Thus, the process of extracting principal components from **R** is sequential [13, p. 98]. After the extraction of (λ_2, $\mathbf{u}_2$), deflation occurs by setting:

$$\mathbf{R}(2) = \mathbf{R}(1) - \lambda \mathbf{u}_2 \mathbf{u}_2^T, \qquad (9)$$

and the third eigenvalue/eigenvector pair (λ_3, $\mathbf{u}_3$) associated with **R** would be extracted as the dominant eigenvalue associated with **R**(2). We use this process in our Excel spreadsheet to sequentially extract all of the eigenvalue/eigenvector pairs for **R**. Each stage of the sequential extraction process is accomplished using successive approximation via the *power method* [14].

The power method is one of the most conceptually straightforward approaches for finding the dominant eigenvector and corresponding eigenvalue of a correlation matrix (as well as other symmetric matrices). Using the power method, the eigenvector at iteration k+1 ($\mathbf{u}^{k+1}$) is estimated from the multiplication of the eigenvector at iteration k ($\mathbf{u}^k$) by **R**. That is, $\mathbf{u}^{k+1} = \mathbf{R}\mathbf{u}^k$. To illustrate why this process will converge to the dominant eigenvector, let us assume that the initial estimate ($\mathbf{u}^0$) is expressed as a linear combination (with coefficients α_j, for $1 \leq j \leq p$) function of the p eigenvectors of **R** as follows: $\mathbf{u}^0 = \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \ldots + \alpha_p\mathbf{u}_p$. For the first iteration, we would have $\mathbf{u}^1 = \mathbf{R}\mathbf{u}^0 = \alpha_1\mathbf{R}\mathbf{u}_1 + \alpha_2\mathbf{R}\mathbf{u}_2 + \ldots + \alpha_p\mathbf{R}\mathbf{u}_p$ and, because of the eigen-structure relationship $\mathbf{R}\mathbf{u} = \lambda\mathbf{u}$, we can re-write this as $\mathbf{u}^1 = \mathbf{R}\mathbf{u}^0 = \alpha_1\lambda_1\mathbf{u}_1 + \alpha_2\lambda_2\mathbf{u}_2 + \ldots + \alpha_p\lambda_p\mathbf{u}_p$. Recalling that we have defined λ_1 as the largest eigenvalue, it is helpful to re-write this as: $\mathbf{u}^1 = \mathbf{R}\mathbf{u}^0 = \lambda_1[\alpha_1\mathbf{u}_1 + \alpha_2(\lambda_2/\lambda_1)\mathbf{u}_2 + \ldots + \alpha_p(\lambda_p/\lambda_1)\mathbf{u}_p]$. After m+1 iterations of the power method, we have: $\mathbf{u}^{m+1} = \mathbf{R}^m\mathbf{u}^0 = (\lambda_1)^m[\alpha_1\mathbf{u}_1 + \alpha_2(\lambda_2/\lambda_1)^m\mathbf{u}_2 + \ldots + \alpha_p(\lambda_p/\lambda_1)^m\mathbf{u}_p.]$. In the limit, as $m \rightarrow \infty$, the terms $(\lambda_2/\lambda_1)^m, \ldots, (\lambda_p/\lambda_1)^m$ go to zero as long as λ_1 is strictly larger than all of the other eigenvalues. Assuming $\alpha_1 \neq 0$, the result is convergence to an eigenvector that is a multiple of the dominant eigenvector: that is, $\mathbf{u}^{m+1} = \mathbf{R}^m\mathbf{u}^0 = \alpha_1(\lambda_1)^m\mathbf{u}_1$. Faster convergence is generally achievable if the updated eigenvector is normalized to unit length using the following equation:

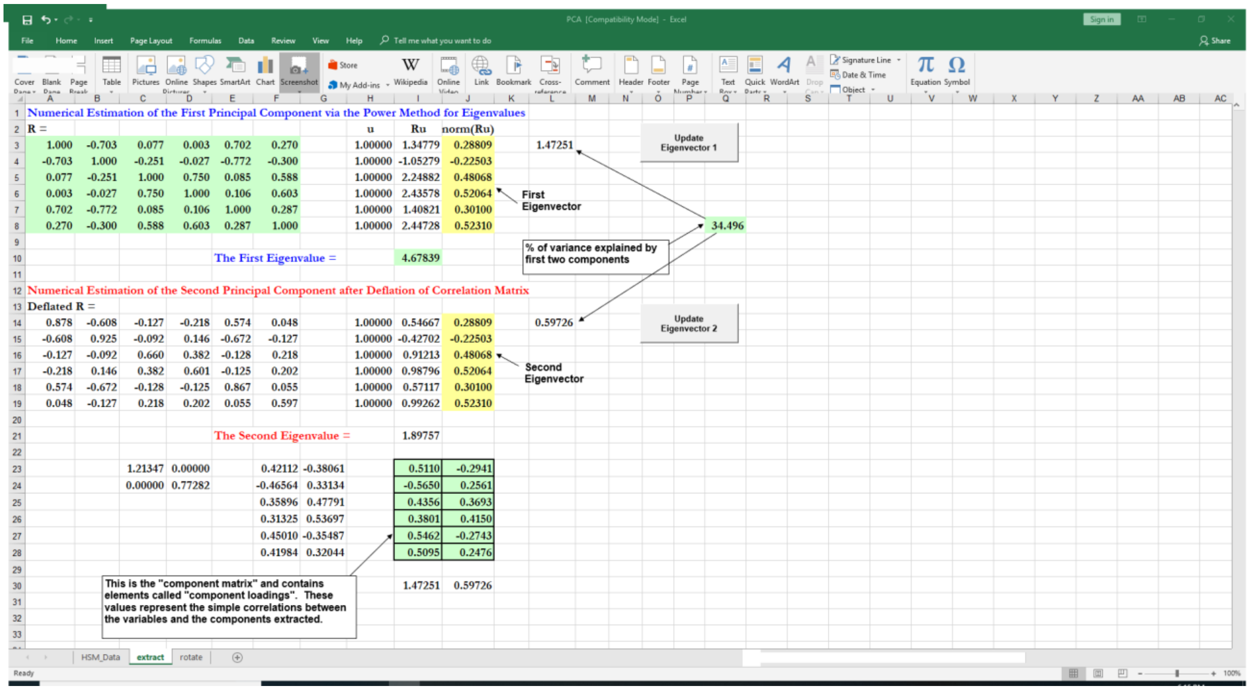
$$\mathbf{u}^{k+1} = \frac{\mathbf{Ru}^k}{\sqrt{\left(\mathbf{Ru}^k\right)^T\left(\mathbf{Ru}^k\right)}}. \tag{10}$$

Thus, assuming λ_1 is strictly greater than all other eigenvalues, the power method estimates the eigenvector via an iterative process of multiplying by **R** and normalizing the result. Upon convergence of the estimation process to the eigenvector **u**, the corresponding eigenvalue for an eigenvector can be obtained via the Rayleigh quotient as follows:

$$\lambda = \frac{\mathbf{u}^{\mathbf{T}}\mathbf{Ru}}{\mathbf{u}^{\mathbf{T}}\mathbf{u}}, \tag{11}$$

The Rayleigh quotient makes use of the facts that **u** is an eigenvector of **R** and $\mathbf{Ru} = \lambda\mathbf{u}$. Thus, the numerator of Equation (11) could be re-written as $\lambda\mathbf{u}^T\mathbf{u}$, which implies $\lambda = \lambda$. If the eigenvector is unit length, then the denominator is one and the eigenvalue is equal to the numerator, which is the objective function in Equation (4) and indicates that $\lambda = \mathbf{u}^T\mathbf{Ru}$ is the variance extracted.

For successive extraction of eigenvalues from correlation matrices, the convergence of the power method to the dominant eigenvectors is generally rapid; however, convergence can take more iterations as the ratio of the second largest eigenvalue to the largest eigenvalue approaches one. For example, when computing the dominant eigenvalue for **R**, the ratio of λ_2/λ_1 may have an effect on the rapidity of convergence. When computing the dominant eigenvalue for **R**(1), it is the ratio of λ_3/λ_2 that affects convergence.



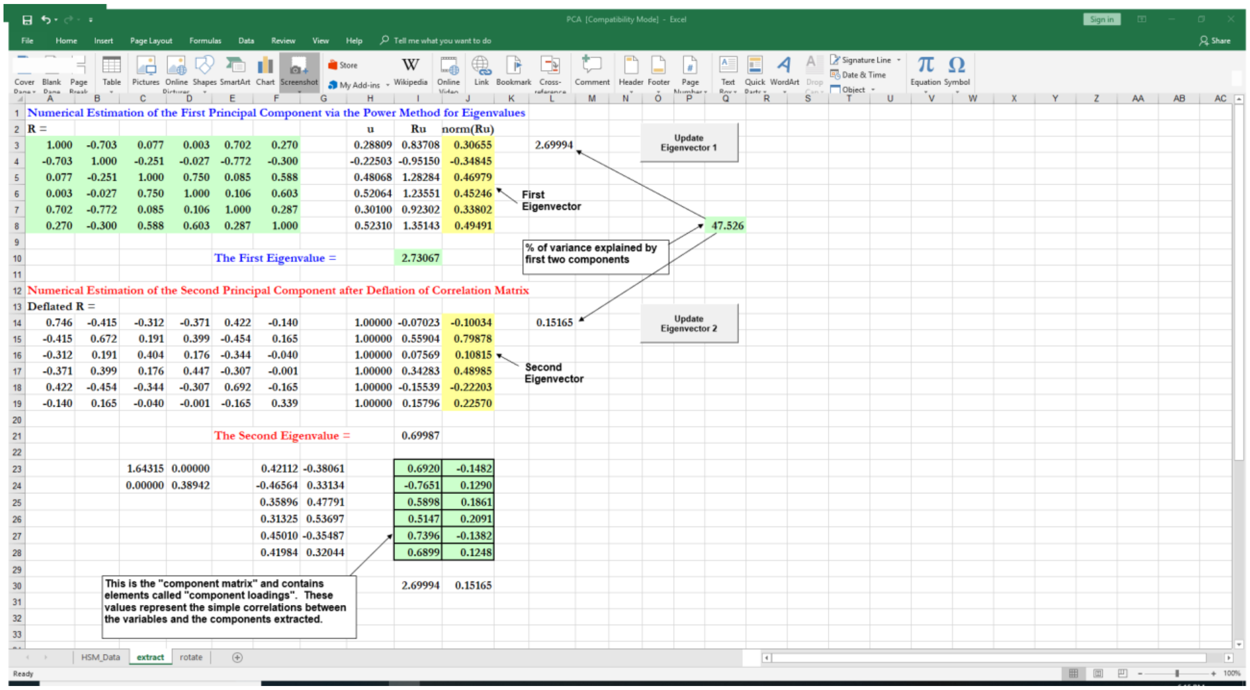
	A	B	C	D	E	F	G	H	I	J	K	L	Q
1	Numerical Estimation of the First Principal Component via the Power Method for Eigenvalues												
2	R =							u	Ru	norm(Ru)			
3	1.000	-0.703	0.077	0.003	0.702	0.270		1.00000	1.34779	0.28809		1.47251	
4	-0.703	1.000	-0.251	-0.027	-0.772	-0.300		1.00000	-1.05279	-0.22503			
5	0.077	-0.251	1.000	0.750	0.085	0.588		1.00000	2.24882	0.48068			
6	0.003	-0.027	0.750	1.000	0.106	0.603		1.00000	2.43578	0.52064			
7	0.702	-0.772	0.085	0.106	1.000	0.287		1.00000	1.40821	0.30100			
8	0.270	-0.300	0.588	0.603	0.287	1.000		1.00000	2.44728	0.52310			34.496
9													
10					The First Eigenvalue =				4.67839				
11													
12	Numerical Estimation of the Second Principal Component after Deflation of Correlation Matrix												
13	Deflated R =												
14	0.878	-0.608	-0.127	-0.218	0.574	0.048		1.00000	0.54667	0.28809		0.59726	
15	-0.608	0.925	-0.092	0.146	-0.672	-0.127		1.00000	-0.42702	-0.22503			
16	-0.127	-0.092	0.660	0.382	-0.128	0.218		1.00000	0.91213	0.48068			
17	-0.218	0.146	0.382	0.601	-0.125	0.202		1.00000	0.98796	0.52064			
18	0.574	-0.672	-0.128	-0.125	0.867	0.055		1.00000	0.57117	0.30100			
19	0.048	-0.127	0.218	0.202	0.055	0.597		1.00000	0.99262	0.52310			
20													
21					The Second Eigenvalue =				1.89757				
22													
23			1.21347	0.00000		0.42112	-0.38061		0.5110	-0.2941			
24			0.00000	0.77282		-0.46564	0.33134		-0.5650	0.2561			
25						0.35896	0.47791		0.4356	0.3693			
26						0.31325	0.53697		0.3801	0.4150			
27						0.45010	-0.35487		0.5462	-0.2743			
28						0.41984	0.32044		0.5095	0.2476			
29													
30									1.47251	0.59726			

% of variance explained by first two components

This is the "component matrix" and contains elements called "component loadings". These values represent the simple correlations between the variables and the components extracted.

Figure 2: The extraction worksheet with eigenvectors initialized to vectors of ones.

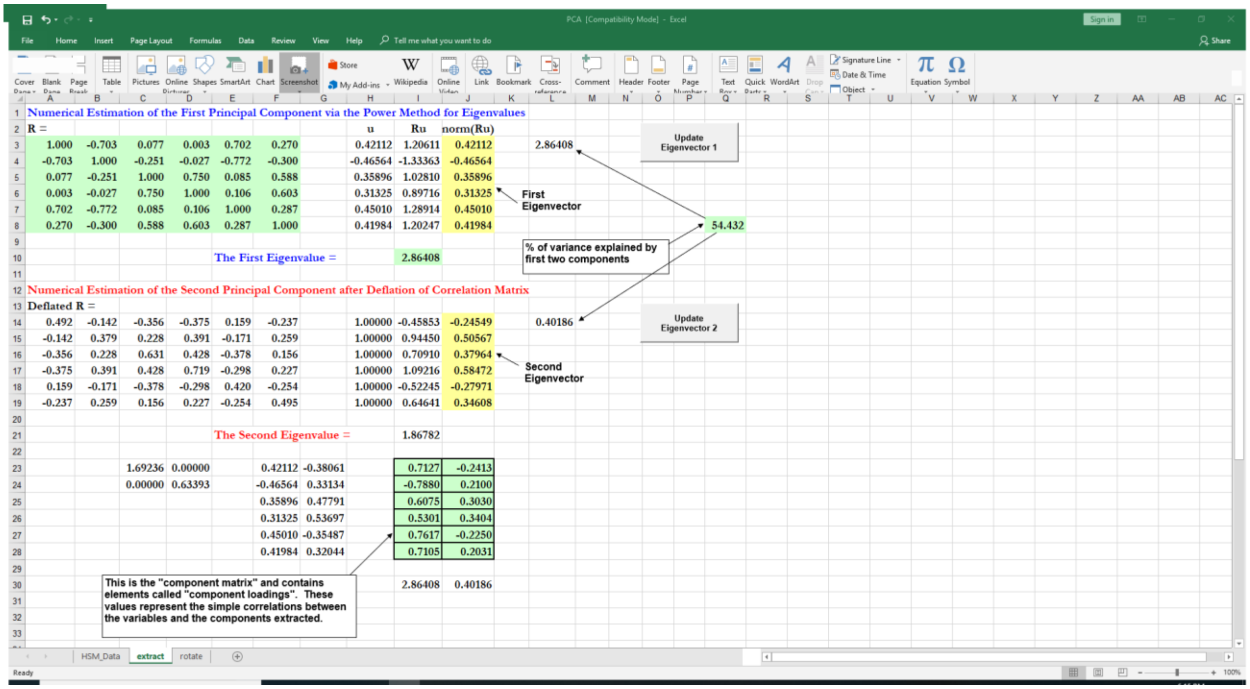
An initial estimate of the first eigenvector (**u**) is placed in cells H3:H8. I usually just start off with all ones in these cells. Cells I3:I8 contain the vector formed by the **Ru** product. The value in cell I10 is the constant $\mathbf{u}^T\mathbf{Ru}$. Again, if **u** is normalized, then this quantity is equal to the eigenvalue estimate. Cells J3:J8 contain **Ru** after it has been normalized to unit length. This is the updated eigenvector. Clicking on the button "Update Eigenvector 1" simply reads the cell values in J3:J8 and re-pastes them in cells H3:H8 to obtain the updated eigenvector for the next iteration. Tapping on this button once (assuming the initial eigenvector of all ones in Figure 2) yields the result in Figure 3. As we continue to tap this button, we can see the eigenvalue get larger. After 15-20 taps, the changes tend to be small. We can continue tapping until the values in H3:H8 and J3:J8 have stabilized (in other words the values in these two cell ranges do not change). The results after convergence are shown in Figure 4. The first eigenvalue is $\lambda = 2.86408$ and its corresponding eigenvector is shown in cells J3:J8.


PCA [Compatibility Mode] - Excel
File Home Insert Page Layout Formulas Data Review View Help Tell me what you want to do
Update Eigenvector 1
Update Eigenvector 2
First Eigenvector
Second Eigenvector
% of variance explained by first two components
HSM_Data extract rotate
Ready

	A	B	C	D	E	F	G	H	I	J	L	Q
1	Numerical Estimation of the First Principal Component via the Power Method for Eigenvalues											
2	R =							u	Ru	norm(Ru)		
3	1.000	-0.703	0.077	0.003	0.702	0.270		0.28809	0.83708	0.30655	2.69994	
4	-0.703	1.000	-0.251	-0.027	-0.772	-0.300		-0.22503	-0.95150	-0.34845		
5	0.077	-0.251	1.000	0.750	0.085	0.588		0.48068	1.28284	0.46979		
6	0.003	-0.027	0.750	1.000	0.106	0.603		0.52064	1.23551	0.45246		
7	0.702	-0.772	0.085	0.106	1.000	0.287		0.30100	0.92302	0.33802		
8	0.270	-0.300	0.588	0.603	0.287	1.000		0.52310	1.35143	0.49491		47.526
9												
10					The First Eigenvalue =				2.73067			
11												
12	Numerical Estimation of the Second Principal Component after Deflation of Correlation Matrix											
13	Deflated R =											
14	0.746	-0.415	-0.312	-0.371	0.422	-0.140		1.00000	-0.07023	-0.10034	0.15165	
15	-0.415	0.672	0.191	0.399	-0.454	0.165		1.00000	0.55904	0.79878		
16	-0.312	0.191	0.404	0.176	-0.344	-0.040		1.00000	0.07569	0.10815		
17	-0.371	0.399	0.176	0.447	-0.307	-0.001		1.00000	0.34283	0.48985		
18	0.422	-0.454	-0.344	-0.307	0.692	-0.165		1.00000	-0.15539	-0.22203		
19	-0.140	0.165	-0.040	-0.001	-0.165	0.339		1.00000	0.15796	0.22570		
20												
21					The Second Eigenvalue =				0.69987			
22												
23			1.64315	0.00000		0.42112	-0.38061		0.6920	-0.1482		
24			0.00000	0.38942		-0.46564	0.33134		-0.7651	0.1290		
25						0.35896	0.47791		0.5898	0.1861		
26						0.31325	0.53697		0.5147	0.2091		
27						0.45010	-0.35487		0.7396	-0.1382		
28						0.41984	0.32044		0.6899	0.1248		
29												
30									2.69994	0.15165		

This is the "component matrix" and contains elements called "component loadings". These values represent the simple correlations between the variables and the components extracted.

Figure 3: The extraction worksheet after tapping the "Update Eigenvector 1" button once.


PCA [Compatibility Mode] - Excel
File Home Insert Page Layout Formulas Data Review View Help Tell me what you want to do
Update Eigenvector 1
Update Eigenvector 2
First Eigenvector
Second Eigenvector
% of variance explained by first two components
HSM_Data extract rotate
Ready

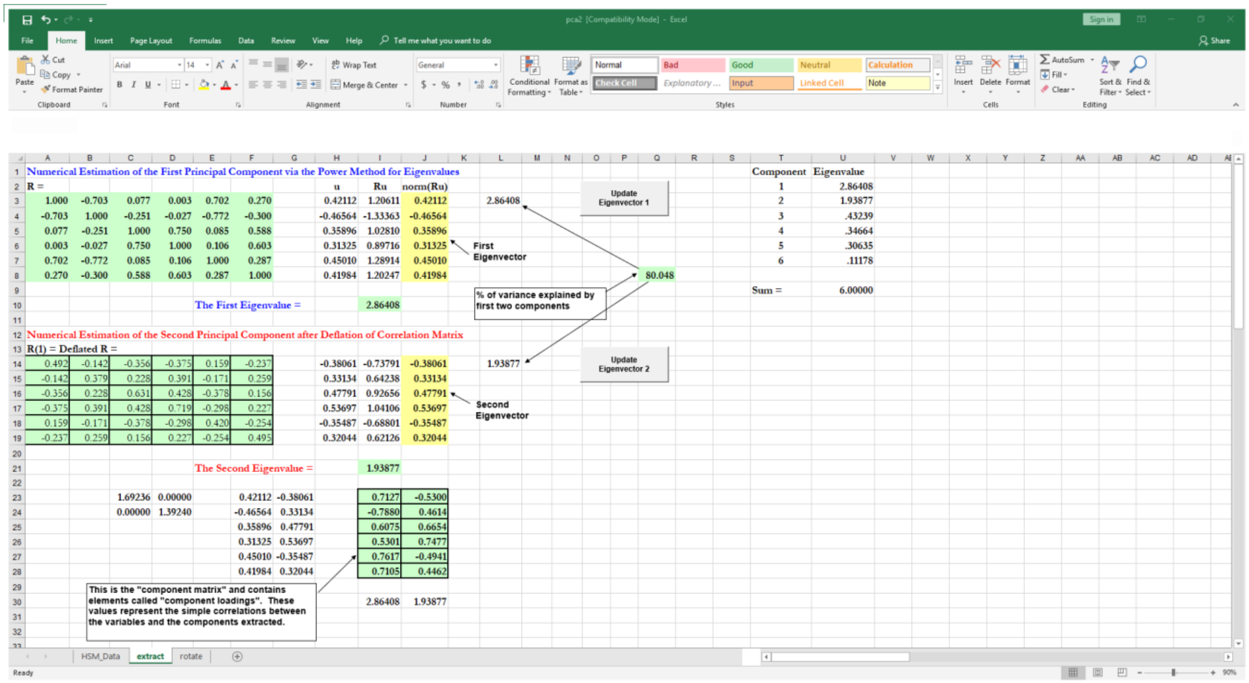
	A	B	C	D	E	F	G	H	I	J	L	Q
1	Numerical Estimation of the First Principal Component via the Power Method for Eigenvalues											
2	R =							u	Ru	norm(Ru)		
3	1.000	-0.703	0.077	0.003	0.702	0.270		0.42112	1.20611	0.42112	2.86408	
4	-0.703	1.000	-0.251	-0.027	-0.772	-0.300		-0.46564	-1.33363	-0.46564		
5	0.077	-0.251	1.000	0.750	0.085	0.588		0.35896	1.02810	0.35896		
6	0.003	-0.027	0.750	1.000	0.106	0.603		0.31325	0.89716	0.31325		
7	0.702	-0.772	0.085	0.106	1.000	0.287		0.45010	1.28914	0.45010		
8	0.270	-0.300	0.588	0.603	0.287	1.000		0.41984	1.20247	0.41984		54.432
9												
10					The First Eigenvalue =				2.86408			
11												
12	Numerical Estimation of the Second Principal Component after Deflation of Correlation Matrix											
13	Deflated R =											
14	0.492	-0.142	-0.356	-0.375	0.159	-0.237		1.00000	-0.45853	-0.24549	0.40186	
15	-0.142	0.379	0.228	0.391	-0.171	0.259		1.00000	0.94450	0.50567		
16	-0.356	0.228	0.631	0.428	-0.378	0.156		1.00000	0.70910	0.37964		
17	-0.375	0.391	0.428	0.719	-0.298	0.227		1.00000	1.09216	0.58472		
18	0.159	-0.171	-0.378	-0.298	0.420	-0.254		1.00000	-0.52245	-0.27971		
19	-0.237	0.259	0.156	0.227	-0.254	0.495		1.00000	0.64641	0.34608		
20												
21					The Second Eigenvalue =				1.86782			
22												
23			1.69236	0.00000		0.42112	-0.38061		0.7127	-0.2413		
24			0.00000	0.63393		-0.46564	0.33134		-0.7880	0.2100		
25						0.35896	0.47791		0.6075	0.3030		
26						0.31325	0.53697		0.5301	0.3404		
27						0.45010	-0.35487		0.7617	-0.2250		
28						0.41984	0.32044		0.7105	0.2031		
29												
30									2.86408	0.40186		

This is the "component matrix" and contains elements called "component loadings". These values represent the simple correlations between the variables and the components extracted.

Figure 4: The extraction worksheet after convergence of the first eigenvector/eigenvalue pair.

Now that the first eigenvalue and eigenvector have been extracted, the correlation matrix is deflated using Equation (8). The deflated correlation matrix is in cells A14:F19. Once again, the initial estimates for the eigenvector are all ones in cells H14:H19.

Tapping the "Update Eigenvector 2" button will lead to convergence of the estimation of the second eigenvalue and eigenvector in a manner similar to the first. The result is shown in Figure 5. The second eigenvalue is λ = 1.93877 and its corresponding eigenvector is shown in cells J14:J19.



	A	B	C	D	E	F	G	H	I	J	L	Q	T	U
1	Numerical Estimation of the First Principal Component via the Power Method for Eigenvalues												Component	Eigenvalue
2	R =							u	Ru	norm(Ru)			1	2.86408
3	1.000	-0.703	0.077	0.003	0.702	0.270		0.42112	1.20611	0.42112	2.86408		2	1.93877
4	-0.703	1.000	-0.251	-0.027	-0.772	-0.300		-0.46564	-1.33363	-0.46564			3	.43239
5	0.077	-0.251	1.000	0.750	0.085	0.588		0.35896	1.02810	0.35896			4	.34664
6	0.003	-0.027	0.750	1.000	0.106	0.603		0.31325	0.89716	0.31325			5	.30635
7	0.702	-0.772	0.085	0.106	1.000	0.287		0.45010	1.28914	0.45010			6	.11178
8	0.270	-0.300	0.588	0.603	0.287	1.000		0.41984	1.20247	0.41984		80.048		
9													Sum =	6.00000
10					The First Eigenvalue =				2.86408					
11														
12	Numerical Estimation of the Second Principal Component after Deflation of Correlation Matrix													
13	R(1) = Deflated R =													
14	0.492	-0.142	-0.356	-0.375	0.159	-0.237		-0.38061	-0.73791	-0.38061	1.93877			
15	-0.142	0.379	0.228	0.391	-0.171	0.259		0.33134	0.64238	0.33134				
16	-0.356	0.228	0.631	0.428	-0.378	0.156		0.47791	0.92656	0.47791				
17	-0.375	0.391	0.428	0.719	-0.298	0.227		0.53697	1.04106	0.53697				
18	0.159	-0.171	-0.378	-0.298	0.420	-0.254		-0.35487	-0.68801	-0.35487				
19	-0.237	0.259	0.156	0.227	-0.254	0.495		0.32044	0.62126	0.32044				
20														
21					The Second Eigenvalue =				1.93877					
22														
23			1.69236	0.00000		0.42112	-0.38061		0.7127	-0.5300				
24			0.00000	1.39240		-0.46564	0.33134		-0.7880	0.4614				
25						0.35896	0.47791		0.6075	0.6654				
26						0.31325	0.53697		0.5301	0.7477				
27						0.45010	-0.35487		0.7617	-0.4941				
28						0.41984	0.32044		0.7105	0.4462				
29														
30									2.86408	1.93877				

First Eigenvector

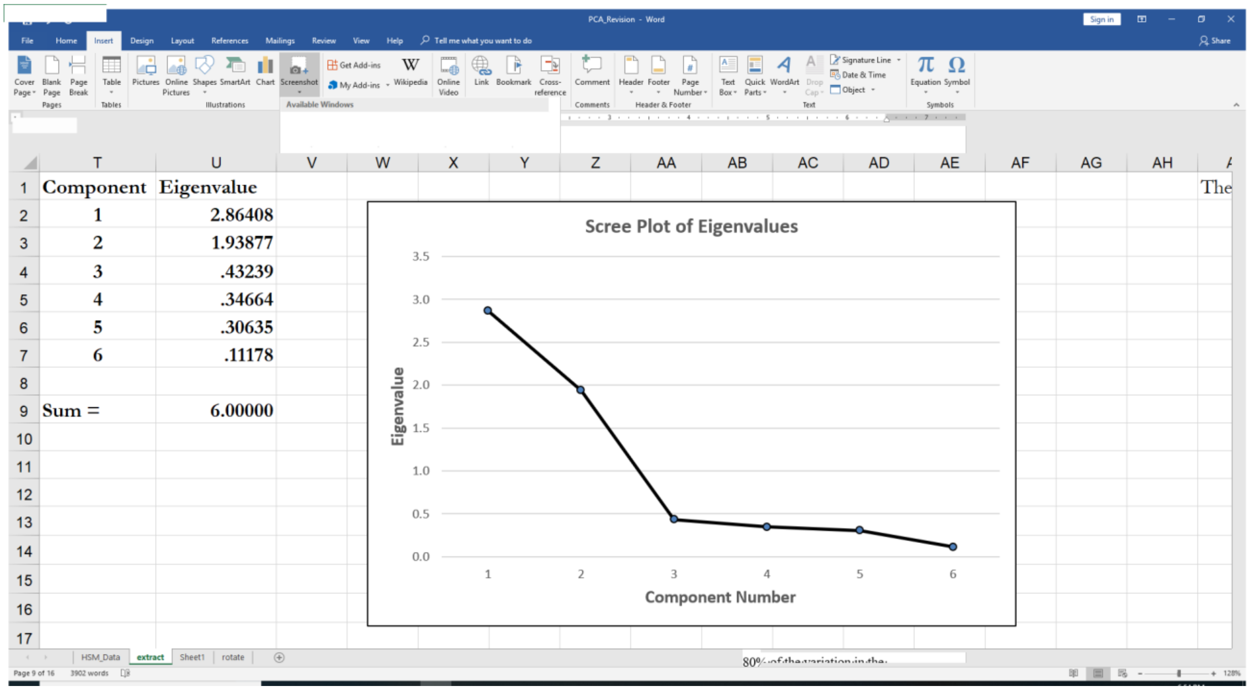
Second Eigenvector

% of variance explained by first two components

This is the "component matrix" and contains elements called "component loadings". These values represent the simple correlations between the variables and the components extracted.

Figure 5: The extraction worksheet after extraction of the first two components.

Continuing in this manner, the remaining four eigenvalue/eigenvector pairs are extracted in rows 34 to 76 of the worksheet. Cells A80:F85 show the fully delated correlation matrix after extraction of all six components. The six eigenvalues are copied into cells U2:U7 so as to facilitate the production of a scree plot, which is shown in Figure 6. The sharp elbow in the plot suggests that two components should be extracted because of the huge drop in the size of the third eigenvalue in comparison to the second. Two components would also be chosen based on the popular default rule (e.g. in SPSS) for correlation matrices of selecting all components with eigenvalues greater than one. The first two components explain just over 80% of the variation in the data set.



Component	Eigenvalue
1	2.86408
2	1.93877
3	.43239
4	.34664
5	.30635
6	.11178
Sum =	6.00000

Figure 6: A scree plot of the eigenvalues.

The 2 × 2 diagonal matrix (**D**) in cells C23:D24 contains the square roots of the eigenvalues on the main diagonal. The two columns of the 6 × 2 matrix ($\mathbf{U}_2$) in cells F23:F28 are the two eigenvectors. The matrix product $\mathbf{G} = \mathbf{U}_2\mathbf{D}$ yields the 6 × 2 matrix of *component loadings* in cells I23:I28. These (unrotated) loadings, which are interpreted as correlations between the six items and the two components, are visually displayed in the correlation circle plot in Figure 7. Five of the six variables (all but x_2) have fairly high positive loadings on component 1. Similarly, four of the six variables (all but x_1 and x_5) have fairly high positive loadings on component 2. Moreover, the component loadings all fall between 0.4 and 0.8 in absolute value. Ideally, we would like to have a *simple structure* whereby the values in each column are close (in absolute value) to either zero or one. This makes it easy to ascertain which variables correspond most heavily to each component. Therefore, in section 4, we will copy these component loadings into the next worksheet ('rotate') and rotate them to see if we can improve interpretability.

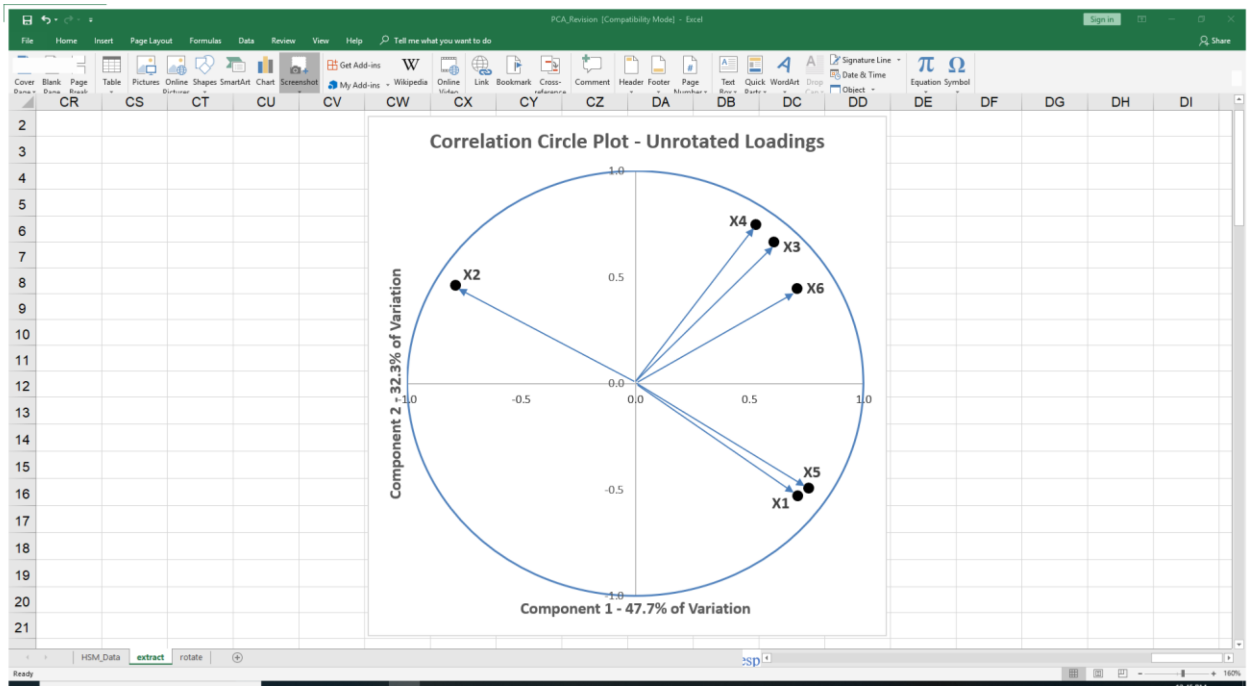

PCA_Revision [Compatibility Mode] - Excel
Correlation Circle Plot - Unrotated Loadings
Component 2 - 32.3% of Variation
Component 1 - 47.7% of Variation
1.0
0.5
0.0
-0.5
-1.0
X1
X2
X3
X4
X5
X6
HSM_Data
extract
rotate

Figure 7: Correlation circle plot for the unrotated component loadings.

Whereas the correlation circle plot in Figure 7 affords a visual display of the relationship between the variables and the two components, Figure 8 is a *component score* plot that provides a visual display of the $n = 30$ respondents in the component space. The component scores are computed by multiplying the 30 × 6 matrix of z-scores ($\mathbf{Z}$) in cells BA4:BF33 by the 6 × 2 matrix containing the first two eigenvectors ($\mathbf{U}_2 = [\mathbf{u}_1\ \mathbf{u}_2]$) in cells BH4:BI9. This matrix multiplication is accomplished using the MMULT function and the resulting 30 × 2 matrix of raw component scores are contained in cells BK4:BL33. Labels are provided in the plot of these component scores in Figure 8 to identify the respondent associated with each data point. For example, the label 'R29' indicates the position of the component score for respondent #29. This respondent provided extreme answers for all six variables (i.e., a response of 1 for x_2 and a response of 7 for the remaining variables and is, therefore, well separated from other respondents in the plot. Ideally, we would like to be able to identify groupings of the respondents (perhaps just by examining the respondents that fall in each quadrant) and characterize them based on the two dimensions of the plot. Unfortunately, this is difficult because the components are difficult to define based on the loadings.

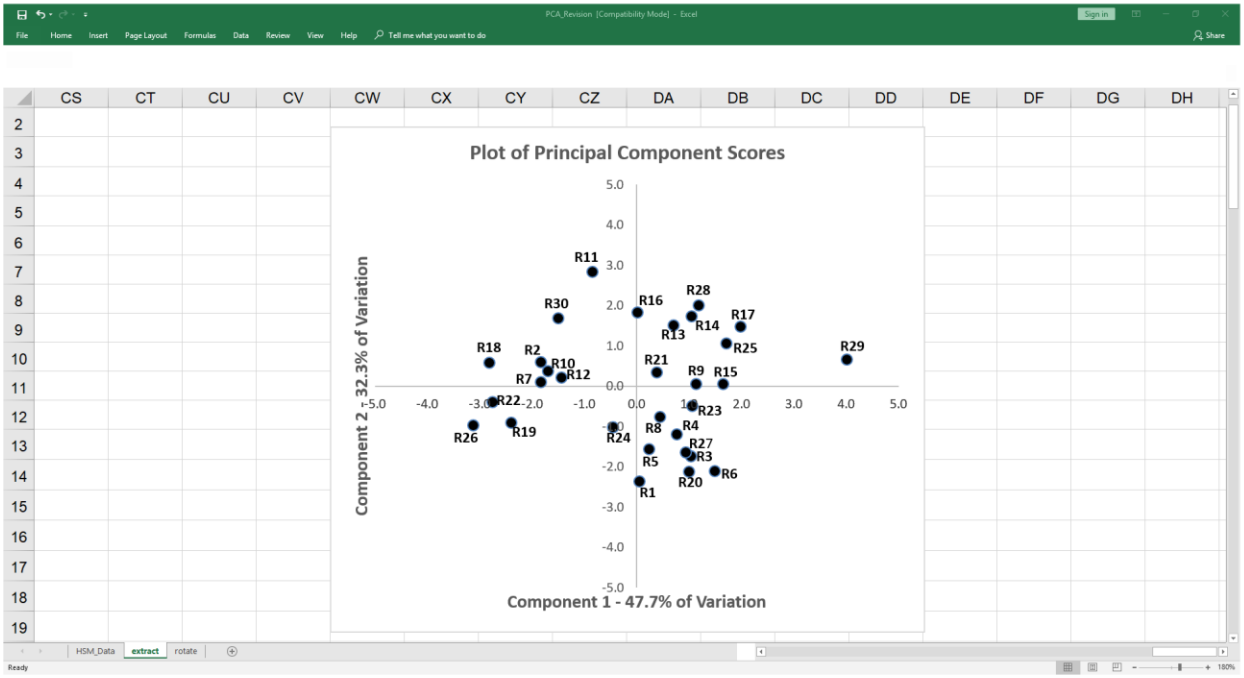


Figure 8. A plot of the raw principal component scores.

We know from the eigenvalues that the first two principal components explain (2.86408 + 1.93877)/6 = 80.048% of the variation in the data. However, we can also see this more concretely from the component scores. To begin the process, we compute the total variation in the data by computing the sum of squares for the *z*-scores in cells BA4:BF33 (note that the mean of the *z*-scores is zero and, therefore, these are squared deviations from the mean). The sum of these squared values, which represents the total variation in the data, is computed to be 174 in cell CD3 using the SUM function. The squared component scores are computed in cells BN4:BO33. The sum of squared values for components 1 and 2 are 83.058 and 56.224, respectively as shown in cells CD6:CD7. The sum for the two components is 139.283 as shown in cell CD9. The percentage of variation in the full data set that is explained by the two components is 139.283/174 = 80.048%. Moreover, the first and second components explain 83.058/174 = 47.7% and 56.224/174 = 32.3% of the variation, respectively.

4. Rotation of the Component Loadings

In the 'rotate' worksheet displayed in Figure 9, the matrix of unrotated component loadings (**G**) has been pasted into cells E6:F11. We seek to find a rotation of this coordinate system to a new set of coordinates, **H** = **GW**, that is more interpretable. The counterclockwise angle of rotation (θ, in degrees) is placed in cell A5. The 2 × 2 rotation matrix, **W**, is in cells A7:B8. To understand the rationale for the rotation matrix, **W**, it is helpful to think of a coordinate pair (g_1, g_2) for the unrotated loadings in terms of polar coordinates (d, φ), where $d = \sqrt{g_1^2 + g_2^2}$ and $\varphi = \arccos(d/g_1)$ if $g_2 \geq 0$ and $\varphi = -\arccos(d/g_1)$ if $g_2 < 0$. The unrotated component loading coordinates can then be expressed in terms of the polar coordinates as:

$$g_1 = d\cos\varphi \qquad (12)$$

$$g_2 = d\sin\varphi \qquad (13)$$

Likewise, recognizing that the counterclockwise angle of rotation will be θ, the coordinate pair (h_1, h_2) for the rotated loadings in terms of polar coordinates will be:

$$h_1 = d\cos(\varphi\text{-}\theta) \qquad (14)$$

$$h_2 = d\sin(\varphi\text{-}\theta) \qquad (15)$$

Based on the trigonometric rules for compound angles, (14) and (15) can be rewritten as:

$$h_1 = d\cos\varphi\cos\theta + d\sin\varphi\sin\theta \qquad (16)$$

$$h_2 = d\sin\varphi\cos\theta - d\cos\varphi\sin\theta \qquad (17)$$

Equations (12) and (13) can be used to simplify Equations (16) and (17) as follows:

$$h_1 = g_1\cos\theta + g_2\sin\theta \qquad (18)$$

$$h_2 = -g_1\sin\theta + g_2\cos\theta \qquad (19)$$

More generally, in matrix notation:

$$[h_1\ h_2] = [g_1\ g_2] \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \qquad (20)$$

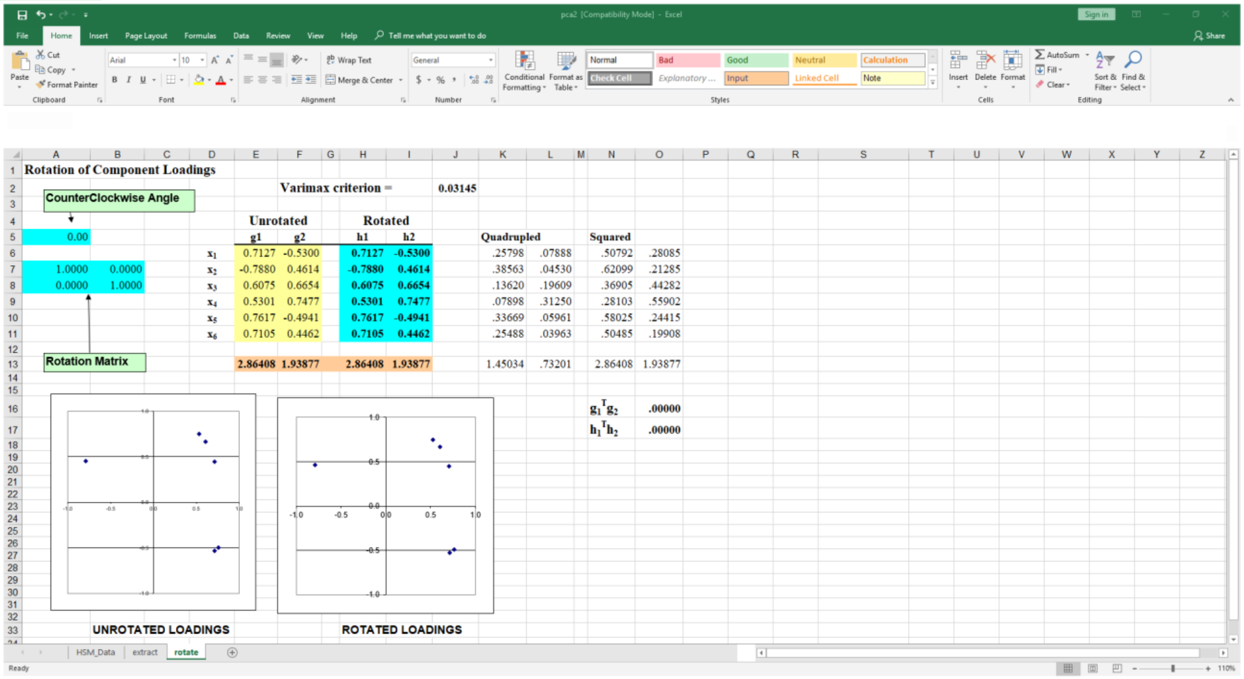
The 2 × 2 matrix on the right side of Equation (20) is the rotation matrix in cells A7:B8. The matrix of rotated component loadings, **H** =**GW** is computed using this matrix and the values are displayed in cells H6:I11. Because the angle of rotation in cell A5 is set to zero in Figure 9, **H** = **G**. Plots of the unrotated and rotated loadings are displayed at the bottom of Figure 9.

The student can manually change the angle of rotation in cell A5 and the rotated loadings in **H** will be updated automatically. Alternatively, an optimal selection of the angle of rotation can be made based on the well-known *varimax* criterion developed by Kaiser [10].

$$V = \frac{1}{p}\sum_{l=1}^{q}\left[\sum_{j=1}^{p} h_{jl}^{4} - \frac{1}{p}\left(\sum_{j=1}^{p} h_{jl}^{2}\right)^{2}\right], \tag{21}$$

where q is the number of selected components (in this example, q = 2). Kaiser [10] developed the *raw varimax criterion* in Equation (21) with the goal of producing an orthogonal rotation of the component axes so as to induce a *simple structure* in the rotated loadings. For this criterion, simplicity is operationalized *as the variance of the squared loadings*. Maximizing this quantity has a propensity to drive the loadings toward zero or one in absolute value, which makes it easier to ascertain the variables that correlate strongly with each component.

Kaiser also proposed a normalized varimax criterion that is particularly relevant in the context of factor analysis. The normalized version adjusts the formula in Equation (21) to account for the *communality* of each variable (i.e., the variance in the variable that is accounted for by the common factor). Because PCA makes no distinction between common and specific factors, communality and the normalized varimax criterion are not relevant in the context of PCA. Although other options for rotation of the loadings are available, varimax is one of the most common and easiest to implement. The values in cells K6:O13 are scratch computations used to facilitate the computation of the value of the varimax criterion, which is in cell J2.



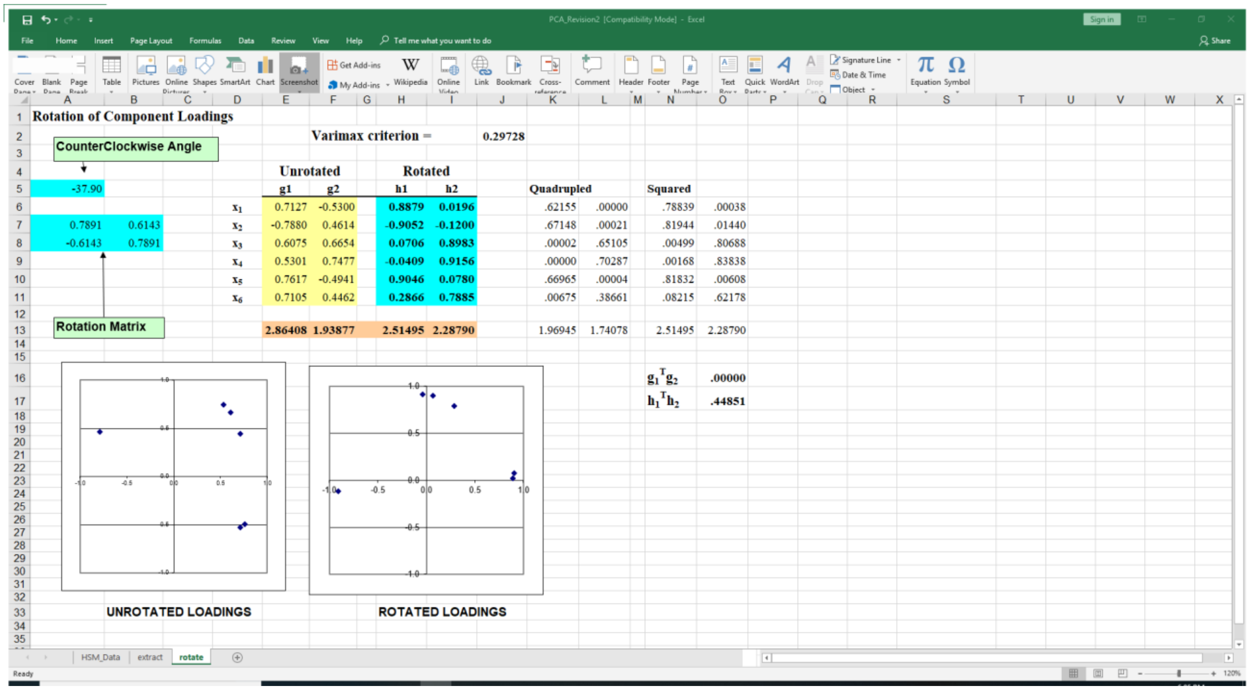
Rotation of Component Loadings

	A	B	D	E	F	H	I	J	K	L	N	O
2				Varimax criterion =				0.03145				
3	CounterClockwise Angle											
4				Unrotated		Rotated						
5	0.00			g1	g2	h1	h2		Quadrupled		Squared	
6			x_1	0.7127	-0.5300	0.7127	-0.5300		.25798	.07888	.50792	.28085
7	1.0000	0.0000	x_2	-0.7880	0.4614	-0.7880	0.4614		.38563	.04530	.62099	.21285
8	0.0000	1.0000	x_3	0.6075	0.6654	0.6075	0.6654		.13620	.19609	.36905	.44282
9			x_4	0.5301	0.7477	0.5301	0.7477		.07898	.31250	.28103	.55902
10			x_5	0.7617	-0.4941	0.7617	-0.4941		.33669	.05961	.58025	.24415
11			x_6	0.7105	0.4462	0.7105	0.4462		.25488	.03963	.50485	.19908
13	Rotation Matrix			2.86408	1.93877	2.86408	1.93877		1.45034	.73201	2.86408	1.93877
16											$g_1^T g_2$	.00000
17											$h_1^T h_2$	.00000

UNROTATED LOADINGS

ROTATED LOADINGS

Figure 9: The ‘rotate’ worksheet prior to varimax rotation (rotation angle θ = 0). The Excel Solver (using θ = 0° as the initial value in cell A5) was used to find the angle of rotation that maximizes the value of the varimax criterion in cell J2. The results are displayed in Figure 10. The optimal counterclockwise angle of rotation is θ = -37.9° (i.e., a clockwise rotation of θ = 37.9°). The rotated loadings have changed dramatically and the correlation circle plot of these loadings in Figure 11 now shows that the variables tend to be associated with one of the two components but not both. Cells O16:O17 show that unrotated loading vectors are orthogonal ($\mathbf{g}_1^T\mathbf{g}_2 = 0$) but the rotated loadings are not ($\mathbf{h}_1^T\mathbf{h}_2 \neq 0$). Moreover, it is noteworthy that, although the total variation explained by the two components is unchanged, the relative contribution of the explained variation for the two components is more equally distributed. That is, for the unrotated loadings it is noted that 2.86048 + 1.93877 = 4.80285 and for the rotated loadings 2.51495 + 2.28790 = 4.80285.


PCA_Revision2 [Compatibility Mode] - Excel
File Home Insert Page Layout Formulas Data Review View Help Tell me what you want to do
Share
HSM_Data extract rotate
Ready
UNROTATED LOADINGS
ROTATED LOADINGS
1.0
0.5
0.0
-0.5
-1.0

Rotation of Component Loadings

CounterClockwise Angle: -37.90

Varimax criterion = 0.29728

Rotation Matrix

0.7891	0.6143
-0.6143	0.7891

	Unrotated		Rotated		Quadrupled		Squared	
	g1	g2	h1	h2				
x_1	0.7127	-0.5300	0.8879	0.0196	.62155	.00000	.78839	.00038
x_2	-0.7880	0.4614	-0.9052	-0.1200	.67148	.00021	.81944	.01440
x_3	0.6075	0.6654	0.0706	0.8983	.00002	.65105	.00499	.80688
x_4	0.5301	0.7477	-0.0409	0.9156	.00000	.70287	.00168	.83838
x_5	0.7617	-0.4941	0.9046	0.0780	.66965	.00004	.81832	.00608
x_6	0.7105	0.4462	0.2866	0.7885	.00675	.38661	.08215	.62178
	2.86408	1.93877	2.51495	2.28790	1.96945	1.74078	2.51495	2.28790

$g_1^T g_2$ .00000

$h_1^T h_2$ .44851

Figure 10: The ‘rotate’ worksheet after varimax rotation (optimal angle -37.9°).

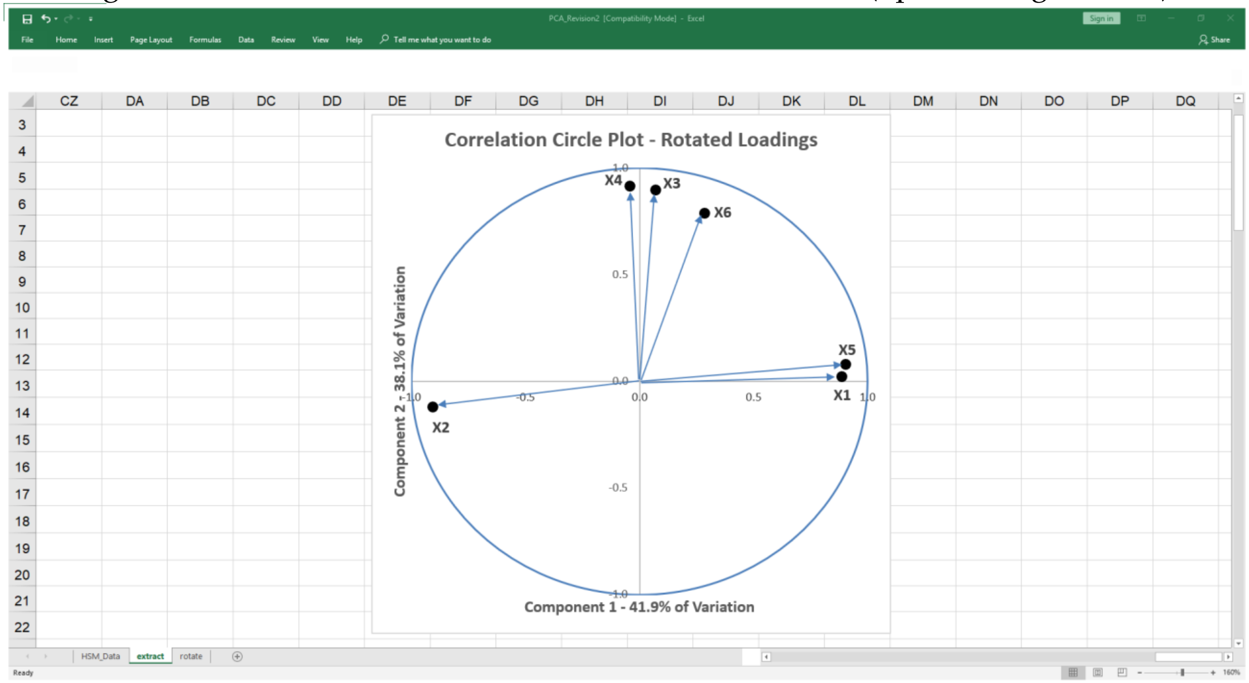


Figure 11. A correlation circle plot for the rotated loadings.

When examining the rotated loadings, it is evident that the variables x_1, x_2, and x_5 have large absolute values on component 1 and small absolute values on component 2. The reverse is true for the other three variables x_3, x_4, and x_6. Accordingly, the interpretation of the rotated components is much cleaner. As observed from Figure 1, Likert-scale items x_1, x_2, and x_5 all pertain to shopping *with* other people; whereas items x_3, x_4, and x_6 are related to shopping *for* other people. In the language used by Arnold and Reynolds [1], component 1 would be identified as a *social shopping* construct and component 2 as a *role shopping* construct.

The cell values of the rotation matrix from the 'rotate' worksheet were copied to cells BQ4:BR5 of the 'extract' worksheet to facilitate a rotation of the principal component scores. The rotated component scores in cells BT4:BU33 were obtained using the MMULT function. Cells CD12:CE15 verify that the rotated components still explain 80.048% of the variation in the data, but that the relative contribution of the two components has changed. A plot of the rotated principal scores is displayed in Figure 12. Once again, labels for each respondent are provided in the plot to facilitate interpretation. Respondents with large positive scores on component 1 are those who might generally perceived as social shoppers, whereas those respondents with large negative scores on component 1 have no proclivity for social shopping. Respondents with large positive scores on component 2 are role shoppers whereas those with large negative scores on component 2 are more averse to role shopping.

Respondent 29 (R29 in the plot in Figure 12) has the largest positive score on component 1 and largest positive score on component 2 and, therefore, this person is both a social and role shopper. This customer strongly agreed (Likert scale response of 7) with all of the role shopping statements (x_3, x_4, x_6), as well as the two social shopping statements (x_1, x_5) that pertained to shopping with others. Concordantly, the respondent strongly disagreed (with a Likert scale response of 1) with the social shopping statement pertaining to shopping alone (x_2). At the other corner of the plot, respondent 26 (R26) is someone who is neither a role shopper nor a social shopper. This individual strongly disagreed (Likert scale responses of 1 or 2) with all of the role shopping statements and the two social shopping statements associated with shopping with others. Respondent 26 generally agreed (Likert scale response of 5) with the statement on shopping alone. Respondent 1 (R1) is a fairly solid social shopper but not a role shopper. This individual strongly agreed (Likert scale response of 6) with the statements pertaining to shopping with others (x_1, x_5), but generally disagreed (Likert scale response of 3) with the statement pertaining to shopping alone (x_2). By contrast, R1 generally disagreed (Likert responses of 2 or 3) with all three of the role shopping statements (x_3, x_4, x_6). Respondent 11 (R11) is a fairly solid role shopper but not a social shopper. This individual agreed (Likert scale responses of 5 or 6) with the statements pertaining to shopping for others (x_3, x_4, x_6). However, R11 disagreed strongly (Likert responses of 1 or 2) with the statements pertaining to shopping with others (x_1, x_5), but strongly agreed (Likert scale response of 7) with the statement pertaining to shopping alone (x_2).

Extreme respondents such as R1, R11, R26, and R29 help to characterize the boundaries of the component score plot. It is also interesting to observe those respondents with very similar component scores. For example, {R2, R7, R10, R12} form a tight cluster of similar individuals who are generally not inclined to be either social or role shoppers. For all of these respondents, their Likert scale responses to statements x_1, x_3, x_4, x_5, and x_6 never exceed four. Moreover, with the exception of statement x_5, where R2 responded with a Likert scale measure of 1 and R12 with 3, the difference between the Likert scale responses for all pairs of these four respondents never differed by more than one-unit for any of the other statements.

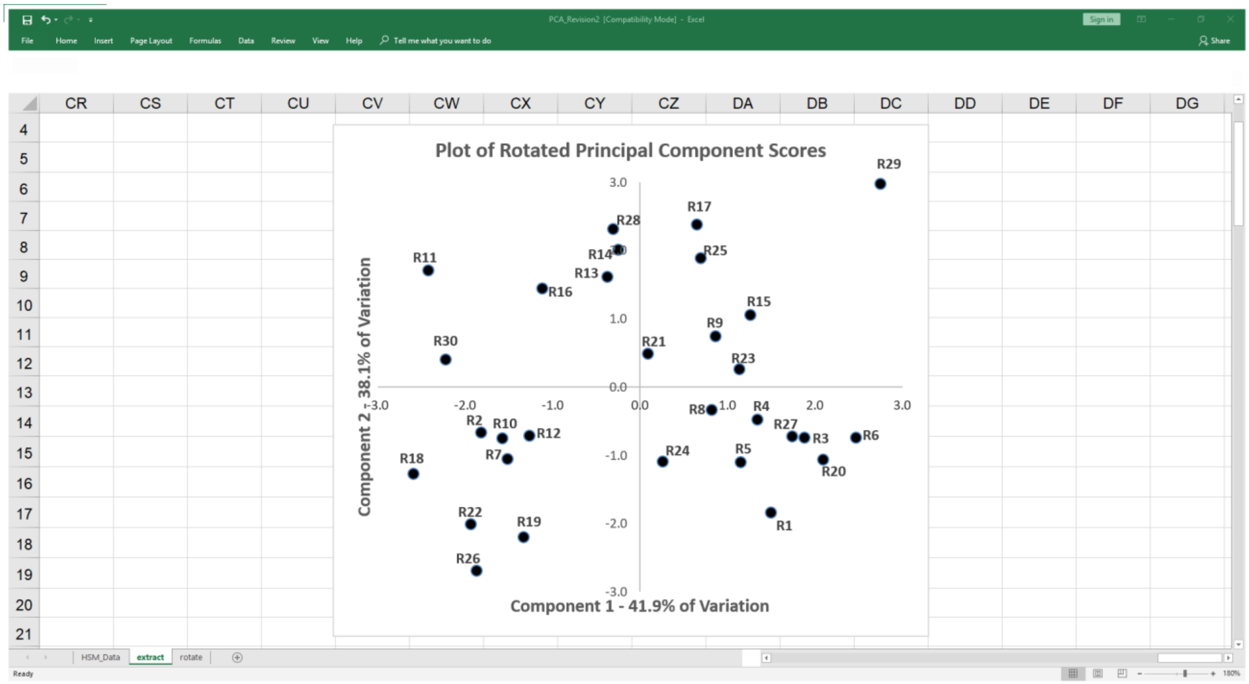


Figure 12: A plot of the rotated principal component scores

5. Practical Experience and Extensions

I have used the Excel workbook for PCA for three different courses: (1) an undergraduate marketing research course, (2) a masters level course in marketing analytics, and (3) a Ph.D. seminar in quantitative methods. Naturally, the coverage of the content in the workbook is apt to vary depending on the level of the course. The rigor of the presentation at the masters level typically falls somewhere between the undergraduate and Ph.D. level and is contingent on other factors such as class size. At the undergraduate level, the focus might be restricted to the generation of the correlation matrix, the scree plot of the eigenvalues for choosing the number of components, the correlation circle plots of the loadings, and the plot of the principal component scores. The coverage of the eigen-decomposition would likely be avoided in an undergraduate business course, but not necessarily a course in a more scientific discipline such as engineering. The details of the rotation of the loadings might also be skipped; however, over the years, I have found that many undergraduates are interested in having at least some rudimentary understanding of how the rotation process works.

At the PhD level, the entire content of the Excel workbook for PCA is presented. The PhD students are expected to have some knowledge of the fundamentals of linear algebra. I have typically required students to do some small matrix analysis examples by hand (i.e., matrix and vector products, inverse, determinants, eigenvalues and eigenvectors) to develop an understanding. The extraction of the eigenvalues and eigenvectors for a six-by-six correlation matrix, however, is virtually impossible by hand. However, the template shows how the power method can do this efficiently and, therefore, the Ph.D. students have some idea as to how eigen-decomposition is done by computer. The feedback that I received has been quite positive, as the students are generally appreciative of gaining a more rigorous understanding of the methodology.

Perhaps the most natural extension of the Excel workbook for PCA is to exploratory factor analysis. This would require the incorporation of an approach for computing the communality associated with each variable, which can be accomplished in different ways. Communality may also affect the nature of the varimax rotation. As noted previously, varimax rotation with Kaiser normalization may be preferable for factor analysis and is the default setting in some software packages (e.g., SPSS). Another important extension would be to adapt the workbook to generate biplots. Whereas PCA is based on the eigen-decomposition of the correlation matrix, biplots are grounded in the singular-value-decomposition [5] of the raw $n \times p$ data matrix. Singular-value-decomposition enables the extraction of eigenvector pairs for respondents and variables, thus enabling them to be plotted simultaneously in a two-dimensional space. There have been some important advancements in the design of biplots [6] that should also be taken into consideration when extending the PCA workbook for this interesting topic.

In addition to exploratory factor analysis and biplots, there are several other possible extensions of the Excel workbook for PCA. Many of these possibilities hinge on the fact that the key engine of the workbook is the use of the power method for eigenvalue and eigenvector estimation. Eigen-structure problems also arise in a variety of related multivariate statistical methods, such as multiple discriminant analysis, canonical correlation, and multidimensional scaling. The development of spreadsheet solutions for these and other applications is an interesting avenue for future research.

References

[1] Arnold, M. J., & Reynolds, K. E. (2003). Hedonic shopping motivations. *Journal of Retailing*, *79*, 77-95.

[2] Barr, G., & Scott, L. (2011). Teaching statistics in a spreadsheet environment using simulation. *eJournal of Spreadsheets in Education*, 4(3), Article 2. Available at: http://epublications.bond.edu.au/ejsie/vol4/iss3/2. [Accessed 17 August 2018].

[3] Barr, G., & Scott, L. (2015). Spreadsheets and simulation for teaching a range of statistical concepts. In: H. MacGillivray, B. Phillips, M. Martin, ed., *Topics from Australian conferences on teaching statistics: OZCOTS 2008-2012*, 1st edition, New York: Springer, pp 99-117.

[4] Barr, G. D., & Scott, L. (2018) An active learning exercise showing some fundamentals of financial portfolio construction*, eJournal of Spreadsheets in Education* Vol. 10(3), Article 5. Available at: https://epublications.bond.edu.au/ejsie/vol10/iss3/5 [Accessed 17 August 2018]

[5] Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*, 211-218.

[6] Gower, J. C., & Hand, D. J. (1996). *Biplots*. London: Chapman and Hall.

[7] Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

[8] Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th edition). Upper Saddle River, NJ: Pearson Prentice Hall.

[9] Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York: Springer.

[10] Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187-200.

[11] Kwan, C. C. Y. (2011). An introduction to shrinkage estimation of the covariance matrix: A pedagogic illustration. *eJournal of Spreadsheets in Education*, 4(3), Article 6. Available at https://epublications.bond.edu.au/ejsie/vol4/iss3/6/ [Accessed 17 August 2018].

[12] Kwan, C. C. Y. (2017). Shrinkage of the sample correlation matrix of returns towards a constant correlation target: A pedagogic illustration based on Dow Jones stock returns, *eJournal of Spreadsheets in Education*, 10(1), Article 3. Available at http://epublications.bond.edu.au/ejsie/vol10/iss1/3 [Accessed 17 August 2018]

[13] Lattin, J., Carroll, J. D., & Green, P. E. (2003). *Analyzing multivariate data*. Pacific Grove, CA: Thomson.

[14] Lay, D. C., Lay, S. R., & McDonald, J. J. (2015). *Linear algebra and its applications* (5th edition). New York: Pearson.

[15] Okada, A., & Tsurumi, H. (2012). Asymmetric multidimensional scaling of brand switching among margarine brands. *Behaviormetrika*, *30*, 111-126.

[16] Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, *15*, 201-292.