Time Series HW2

Matthew Leong

March 2021

Problem 1

The SAS code used to impor the data and run the default PCA is as follows:

```
* Read data into SAS;
libname PCA "/home/u56680950/HW2";
data work.ratings;
  input JOB KNOWHOW PROBLEM_SOLVING ACCOUNTABILITY SALARY;
cards;
0  800  608  1056  102000
2  528  304  460  75740
3  460  264  460  75740
5  528  304  304  79172
4  460  264  400  70000
0  460  264  400  66536
0  528  304  264  70000
7  460  230  264  68000
10  400  200  350  73140
7  400  175  230  66016
7  400  200  200  66016
5  400  175  200  71840
5  304  115  175  71580
2  264  100  175  65860
3  264  100  175  66432
10  230  100  132  64040
10  230  100  132  62610
7  230  87  132  65002
7  230  76  115  64001
5  230  76  115  66900
5  230  87  100  63000
5  230  87  100  63780
7  200  87  100  62000
7  200  76  100  61960
7  200  76  100  62012
7  200  76  87  62300
```

```
5  200  76  87  61960
7  200  66  87  61700
7  175  66  100  61440
2  175  57  100  62220
3  175  57  100  63260
7  175  57  100  59880
2  175  57  100  62480
3  175  57  100  63000
2  175  57  100  63260
3  175  57  100  62480
4  175  57  87  62480
7  175  57  87  61440
2  175  57  87  62064
3  175  57  87  61180
2  175  57  87  59100
3  175  57  87  59620
5  175  66  76  59880
5  175  66  76  60200
7  175  57  76  60140
7  175  57  76  61700
5  175  66  66  60000
7  152  50  87  60920
7  152  50  76  59100
3  152  50  76  61700
2  152  50  76  59880
3  152  50  76  61700
5  152  50  66  59360
5  152  43  66  60660
2  152  43  66  59984
2  152  43  66  60660
3  152  43  66  60920
3  152  43  66  60920
2  152  43  66  60920
3  152  43  66  60660
3  152  43  66  60660
7  152  43  66  58320
5  152  43  66  59360
2  152  43  66  60920
3  152  43  66  60920
4  152  43  66  60660
7  152  43  57  59880
RUN;

* Extract principal components;
proc princomp data=work.ratings out=ratings_PC;
  var  KNOWHOW PROBLEM_SOLVING ACCOUNTABILITY;
```

```
RUN;
```

Problem 2

For the PDF, I include both the formula and nonformula solutions.

	A	B	C	D	E	F	G	H	I	J
1	Eigenvectors (from SAS)					Vector length being 1 signifies normal transformation				
2		Prin1	Prin2	Prin3		Prin1^2	Prin2^2	Prin3^2		
3	KNOWHOW	0.576251	-0.618121	0.53466		0.332065	0.382074	0.285861		
4	PROBLEM_SOLVING	0.584343	-0.145758	-0.79831		0.341457	0.021245	0.637299		
5	ACCOUNTABILITY	0.571383	0.772451	0.277201		0.326479	0.596681	0.07684		
6					SUM	1.00	1.00	1.00		
7	Dot Product = 0 signifies orthogonal transformation									
8	Prin1,Prin2	0.00								
9	Prin2,Prin3	0.00								
10	Prin1,Prin3	0.00								

	A	B	C	D	E	F	G	H
1	Eigenvectors (from SAS)					Vector length being 1		
2		Prin1	Prin2	Prin3		Prin1^2	Prin2^2	Prin3^2
3	KNOWHOW	0.576251	-0.618121	0.53466		=B3^2	=C3^2	=D3^2
4	PROBLEM_SOLVING	0.584343	-0.145758	-0.79831		=B4^2	=C4^2	=D4^2
5	ACCOUNTABILITY	0.571383	0.772451	0.277201		=B5^2	=C5^2	=D5^2
6					SUM	=SUM(F3:F5)	=SUM(G3:G5)	=SUM(H3:H5)
7	Dot Product = 0 signifies orthogonal transfo							
8	Prin1,Prin2	=SUMPRODUCT(B3:B5,C3:C5)						
9	Prin2,Prin3	=SUMPRODUCT(C3:C5,D3:D5)						
10	Prin1,Prin3	=SUMPRODUCT(B3:B5,D3:D5)						

As seen from the excel calculations, both conditions are fulfilled which verifies that the principal component transformation is orthonormal.

Problem 3

To get the standardized original vectors with the principal components, I use the following SAS code:

```
* Standardize the data first;
Proc STDIZE Data = ratings_PC out = ratings_PC_STD;
Var KNOWHOW PROBLEM_SOLVING ACCOUNTABILITY;
RUN;

* Export the data from SAS;
Proc Export Data = ratings_PC_STD outfile = '/home/u56680950/HW2/ratingsSTDPC.xlsx'
DMBS = XLSX REPLACE;
Run;
```

For this problem, I am basing my angle solution off of the following rearrangement of the definition of the dot product:

$$a \bullet b = |a| * |b| cos\theta$$

$$cos\theta = \frac{a \bullet b}{|a| * |b|}$$

$$\theta = cos^{-1}(\frac{a \bullet b}{|a| * |b|})$$

The answer here is in radians.

	A	B	C	D	E	F	G	H	I
1	KNOWHOW	PROBLEM_SOLVING	ACCOUNTABILITY	Vector Lengths	Dot Product	Dot Product/Vector Length	Arcosine to get angle in radians		
2	4.35032492	5.283939792	6.116424944	9.17910675					
3	2.25124045	2.122082877	2.124774933	3.753130402	34.002612	0.987002391	0.16140547		
4									
5									
6									
7	Prin1	Prin2	Prin3	Vector Lengths	Dot Product	Dot Product/Vector Length			
8	9.089332156	1.265430074	-0.19679536	9.17910675					
9	3.75136318	-0.059567149	0.098558914	3.753130402	34.002612	0.987002391	0.16140547		
10									

	A	B	C	D	E	F	G
1	KNOWHOW	PROBLEM_SOLVING	ACCOUNTABILITY	Vector Lengths	Dot Product	Dot Product/Vector Length t	Arcosine to get angle
2	4.35032491978601	5.28393979241028	6.11642494361894	=SQRT(SUMSQ(A2:C2))			
3	2.25124045003092	2.12208287685083	2.12477493301896	=SQRT(SUMSQ(A3:C3))	=A2*A3+B2*B3+C2*C3	=E3/(D2*D3)	=ACOS(F3)
4							
5							
6							
7	Prin1	Prin2	Prin3	Vector Lengths	Dot Product	Dot Product/Vector Length t	
8	9.08933215604007	1.26543007390455	-0.196795359919366	=SQRT(SUMSQ(A8:C8))			
9	3.75136318039807	-0.059567148942867	0.0985589142854691	=SQRT(SUMSQ(A9:C9))	=A8*A9+B8*B9+C8*C9	=E9/(D8*D9)	=ACOS(F9)

As seen in the calculations, the principal component rotation preserves the vector lengths and the angle between the two vectors.

Problem 4

K	L	M	N	O	P	Q	R	S
Variance				From SAS				
Prin1	Prin2	Prin3		the Correlation				
2.908081137	0.08369737	0.008221492			Eigenvalue	Difference	Proportion	Cumulative
				1	2.90808114	2.82438377	0.9694	0.9694
				2	0.08369737	0.07547588	0.0279	0.9973
				3	0.00822149		0.0027	1
*Slight differences due to rounding on the SAS output. But they are the same.								

K	L	M
Variance		
Prin1	Prin2	Prin3
=VAR.S(G2:G68)	=VAR.S(H2:H68)	=VAR.S(I2:I68)
*Slight differences due to roundir		

For the formulas and no formulas, I elect to just show the relevant portion. The variances are the same. The rest of the values are simple the principal components of the 67 jobs obtained from the following SAS code:

```
* Export the data from SAS;
Proc Export Data = ratings_PC outfile = '/home/u56680950/HW2/ratingsPC.xlsx'
DMBS = XLSX REPLACE;
Run;
```

Problem 5

The following SAS code was run:

```
* Problem 5: Regress Prin1 on three ratings
* Standardize the data first;
Proc STDIZE Data = ratings_PC out = ratings_PC_STD;
Var KNOWHOW PROBLEM_SOLVING ACCOUNTABILITY;
RUN;

Proc Reg Data = Ratings_PC_STD;
model Prin1 = KNOWHOW PROBLEM_SOLVING ACCOUNTABILITY / noint;
RUN;
```

It resulted in the following table:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
KNOWHOW	1	0.57625	0	Infty	<.0001
PROBLEM_SOLVING	1	0.58434	0	Infty	<.0001
ACCOUNTABILITY	1	0.57138	0	Infty	<.0001

The table results in the following linear equation:

$$Prin_1 = 0.57625 * KNOWHOW + 0.58434 * PROBLEM\_SOLVING + 0.57138 * ACCOUNTABILITY$$

This regression equation is not surprising. As explained in class, the linear regression aims to model the dependent variable linearly and it just so happens that a principal component's linear equation results from the eigenvectors of the principal component which these coefficients match with.

Problem 6

```
*Problem 6: Regress standardized knowhow on the 3 prin components;
Proc Reg Data = Ratings_PC_STD;
model KNOWHOW = Prin1 Prin2 Prin3 / noint;
RUN;
```

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Prin1	1	0.57625	0	Infty	<.0001
Prin2	1	-0.61812	0	-Infty	<.0001
Prin3	1	0.53466	0	Infty	<.0001

$$KNOWHOW = 0.57625 * Prin1 + 0.58434 * Prin2 + 0.57138 * Prin3$$

Again, since PCs are an orthonormal transformation which by definition is a linear transformation. It is unsurprising to see the eigenvalues here in KNOWHOW when it is modelled by a linear regression.

Problem 7

For the loadings matrix, I used the SAS Pearson correlation matrix whose coefficients represent the loading matrix for the PCA transformation.

```
*Problem 7: Write the loadings matrix;
Proc corr data = ratings_PC_STD;
Var Prin1 Prin2 Prin3;
with KNOWHOW PROBLEM_SOLVING ACCOUNTABILITY;
RUN;
```

Pearson Correlation Coefficients, N = 67 Prob > |r| under H0: Rho=0			
	Prin1	Prin2	Prin3
KNOWHOW	0.98269 <.0001	-0.17883 0.1476	0.04848 0.6968
PROBLEM_SOLVING	0.99648 <.0001	-0.04217 0.7347	-0.07238 0.5605
ACCOUNTABILITY	0.97439 <.0001	0.22347 0.0691	0.02513 0.8400

For interpreting these coefficients, I note that principal component 1 has high correlations with all three variables. This means this principal component should encompass jobs with all three of those attributes. This would most likely be leading positions or heads of a team.
Principal component 2 on the other hand has negative correlations with knowhow and problem solving but a positive correlation with accountability. This might be essential work but not so essential that it demands high skill levels. This might be something signifying jobs like a janitor for instance.
Lastl, principal component 3 does not have high correlations positive or negative. This means that it might signify roles that aren't important in the grand scheme of things.

Problem 8

For this problem, I refer to the eigenvalue and proportion table that was shown in problem 4.
The Kaiser rule disregards principal components with eigenvalues less than 1.

This would just leave principal component 1.
The Joliffe rule disregards principal components with eigenvalues less than 1. Again, this would just leave PC1.
The 80% rule signifies that we keep principal components that explain up to 80% of the total variance. Since PC1 explains about 96.94% of it, it is again the only one retained.

Problem 9

```
*Problem 9: Regress salary on three prin components;
Proc Reg Data = Ratings_PC_STD;
model salary = Prin1 Prin2 Prin3;
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: SALARY

Number of Observations Read	67
Number of Observations Used	67

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2465105931	821701977	189.55	<.0001
Error	63	273111655	4335106		
Corrected Total	66	2738217587			

Root MSE	2082.09165	R-Square	0.9003
Dependent Mean	63929	Adj R-Sq	0.8955
Coeff Var	3.25686		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	63929	254.36798	251.33	<.0001
Prin1	1	3557.20641	150.28811	23.67	<.0001
Prin2	1	2316.12408	885.87403	2.61	0.0112
Prin3	1	3540.61136	2826.52316	1.25	0.2150

From the R^2 value, this regression on salary explains about 90.03% of the variation in salary.

Problem 10

Judging from significance level, the order of importance is $PC1 > PC2 > PC3$ as the first one is significant even at really low α levels while the third is not significant even at an α level of 0.1 and 2 is somewhere in between.

part c

```
*Problem 10: Regress salary Prin1 only;
Proc Reg Data = Ratings_PC_STD;
model salary = Prin1;
RUN;
```

The REG Procedure
Model: MODEL1
Dependent Variable: SALARY

Number of Observations Read	67
Number of Observations Used	67

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2428670447	2428670447	509.98	<.0001
Error	65	309547140	4762264		
Corrected Total	66	2738217587			

Root MSE	2182.26114	R-Square	0.8870
Dependent Mean	63929	Adj R-Sq	0.8852
Coeff Var	3.41355		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	63929	266.60563	239.79	<.0001
Prin1	1	3557.20641	157.51847	22.58	<.0001

If we choose to only use PC1 to explain salary, the R^2 falls to 0.8870 meaning that about $0.9003 - 0.8870 = 0.0133$ of the R^2 is lost. Subsequently this translates to a loss of about 1.33% explanatory power in the model.