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:

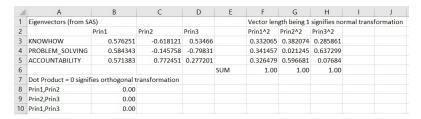
```
* Read data into SAS;
libname PCA "/home/u56680950/HW2";
data work.ratings;
  input JOB KNOWHOW PROBLEM_SOLVING ACCOUNTABILITY SALARY;
cards;
   800
             1056 102000
0
        608
   528
        304
             460
                  75740
   460
                  75740
        264
     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

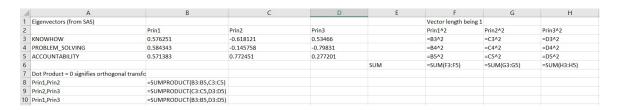
```
200
       76
           87 61960
7
   200
            87
                61700
        66
7
                61440
   175
        66
            100
2
   175
        57
            100
                 62220
3
   175
        57
            100
                 63260
7
   175
        57
         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

Problem 2

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





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:

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

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

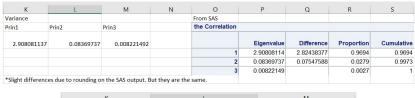
The answer here is in radians.

4	Α	В	С	D	E	F	G	Н	1
1	KNOWHOW	PROBLEM_SOLVING	ACCOUNTABILITY	Vector Lengths	Dot Product	Dot Product/Vector Length	Arcosine to g	et 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									

	Α	В	С	D	E	F	G
1	KNOWHOW	PROBLEM_SOLVING	ACCOUNTABILITY	Vector Lengths	Dot Product	Dot Product/Vector Length	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	
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
Variance		
Prin1	Prin2	Prin3
=VAR.S(G2:G68)	=VAR.S(H2:H68)	=VAR.S(12:168)
*Slight differences due to rour	ndir	

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:

	Pai	rameter Estim	nates		
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;
```

		Paramete	r 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 Correla Prob > r	ation Coeff under H0:		: 67
	Prin1	Prin2	Prin3
KNOWHOW	0.98269	-0.17883	0.04848
	<.0001	0.1476	0.6968
PROBLEM_SOLVING	0.99648	-0.04217	-0.07238
	<.0001	0.7347	0.5605
ACCOUNTABILITY	0.97439	0.22347	0.02513
	<.0001	0.0691	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 V	ariance		
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	-	

		Paramete	r 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 \mathbb{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;
```

		Marre	horo	f Observ	otion	Dood	67	7		
				f Observ						
			An	alysis o	f Varia	ance				
Sourc	е	DF		Sum o Squares	1	Me Squa		F Va	lue	Pr > F
Model		1 242		28670447 24286704		147	509	9.98	<.0001	
Error	or		30	30954714		4762264				
Correc	cted Total	66	273	2738217587						
		~=		2402.2	C44.4	D.C.	uaro	0.8	3870	
	Root M	SE		2182.2	6114	R-Sq	uare	0.0		
	Root M Depend	55	lean		3929	Adj R			8852	
		lent N	lean	6						
	Depend	lent N	lean	6	3929					
	Depend	lent N		6	3929 1355	Adj R				
V	Depend	lent N	Pa Para	6 3.4	3929 1355 Estim	Adj R	!-Sq			> t
-	Depend Coeff V	lent N	Pa Para	6 3.4 rameter	3929 1355 Estim Star	Adj R	t Va	3.0	3852	100

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.