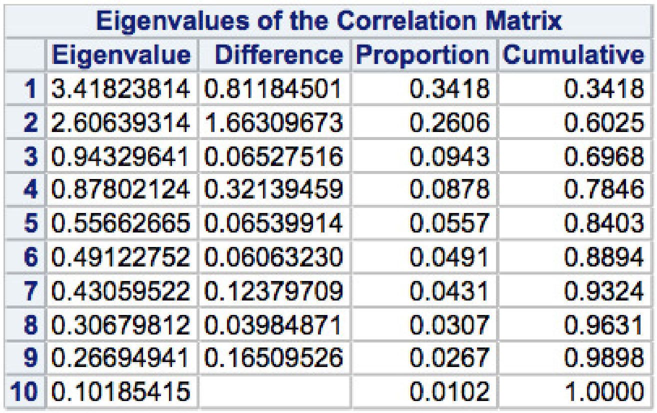
STAT 8320 Spring 2015 Assignment 6

Peng Shao 14221765

May 4, 2015

▶ **1. Solution.** (a). Because PCA is not invariant with respect to changes in scale, that is, the variable which is measured under larger scale will have larger variance, then produce a large eigenvalue when performing singular value decomposition. So standardization avoids the problems of having one variable with large variance unduly influencing the determination of factor loadings, and the correlation matrix is the covariance matrix of standardized variable.

(b). Before PCA, we should investigate the data roughly, and we find that there may be an outlier which have the total score less than 6000. We have no idea why cause this abnormality, but we should take it out of the following analysis because the PCA is very sensitive to outliers. Then we do a PCA based on the 10 single scores.



Eigenvalues of the Correlation Matrix

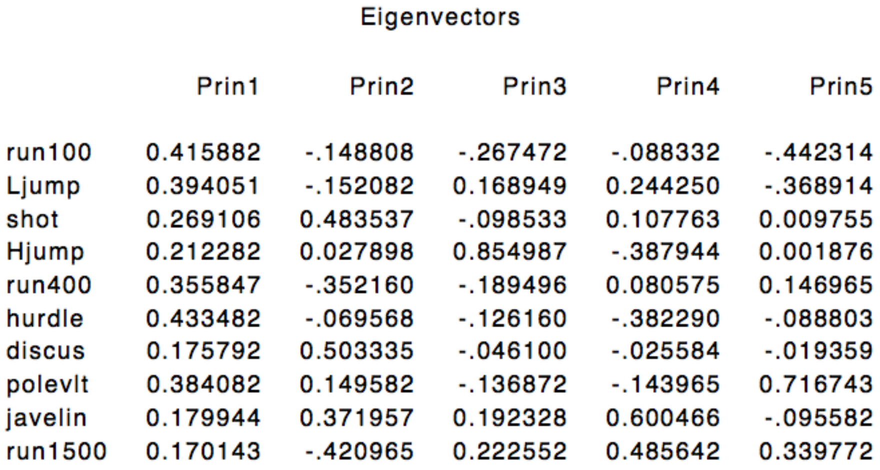
	Eigenvalue	Difference	Proportion	Cumulative
1	3.41823814	0.81184501	0.3418	0.3418
2	2.60639314	1.66309673	0.2606	0.6025
3	0.94329641	0.06527516	0.0943	0.6968
4	0.87802124	0.32139459	0.0878	0.7846
5	0.55662665	0.06539914	0.0557	0.8403
6	0.49122752	0.06063230	0.0491	0.8894
7	0.43059522	0.12379709	0.0431	0.9324
8	0.30679812	0.03984871	0.0307	0.9631
9	0.26694941	0.16509526	0.0267	0.9898
10	0.10185415		0.0102	1.0000

Figure 1: Principal Component Proportion

From Figure **??**, the first 2 PCs have accounted for 65.25% variability of the total.

(c). Because the first 2 PCs account for more than 50% variability of the data, and the only these 2 PCs have the eigenvalues more than 1.

(d). According to the part of PC table (Figure **??**), the first principal component may be a kind of overall means of all scores, not exactly equal to the average because of some slightly different weights; and the second principal component may be the comparison between the scores of running add long jump and those of throwing add pole vault.



Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
run100	0.415882	-.148808	-.267472	-.088332	-.442314
Ljump	0.394051	-.152082	0.168949	0.244250	-.368914
shot	0.269106	0.483537	-.098533	0.107763	0.009755
Hjump	0.212282	0.027898	0.854987	-.387944	0.001876
run400	0.355847	-.352160	-.189496	0.080575	0.146965
hurdle	0.433482	-.069568	-.126160	-.382290	-.088803
discus	0.175792	0.503335	-.046100	-.025584	-.019359
polevlt	0.384082	0.149582	-.136872	-.143965	0.716743
javelin	0.179944	0.371957	0.192328	0.600466	-.095582
run1500	0.170143	-.420965	0.222552	0.485642	0.339772

Figure 2: First 5 Principal Components

(e).

From the scatter plot Figure **??**, the data points have already been labeled as the their rank. So we can see that most high rank athletes (here higher rank means higher total score but smaller value of the rank) have comparably high 1st PC score. However, it is hard to see that whether there is any pattern of 2nd PC score relative to the ranks of athletes.

Hence, to make the relationship between PC scores and the ranks (or the total scores) more clearly, we can plot the total score versus two PC scores respectively(Figure **??** and Figure **??**). From Figure **??**, we can see that the 1st PC score has positive relationship with total score indeed, and we also believe that there should be no significant relationship between 2nd PC score and total score.

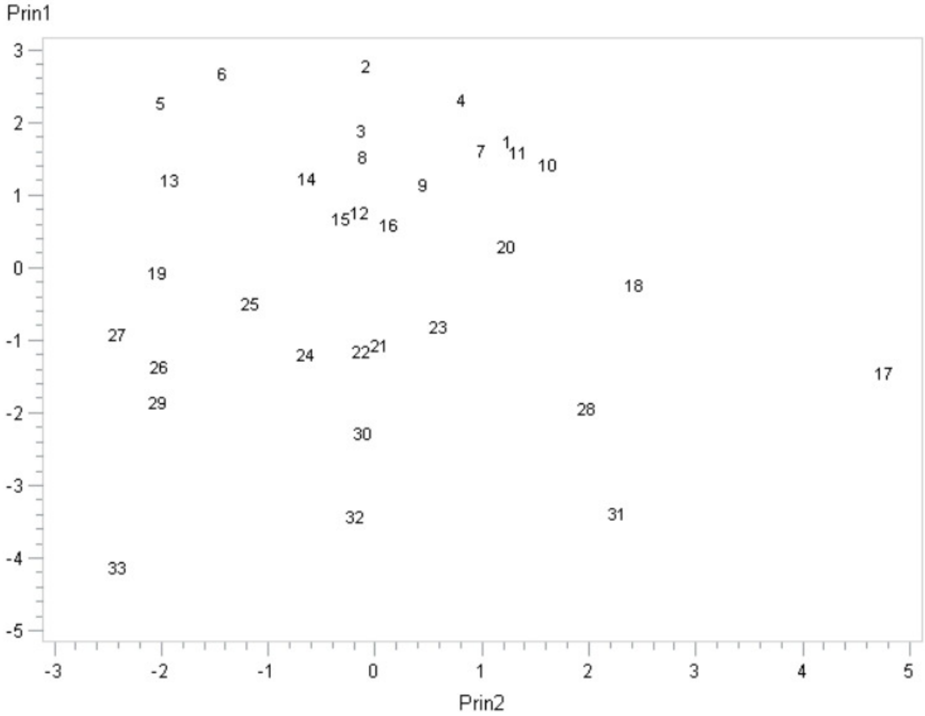
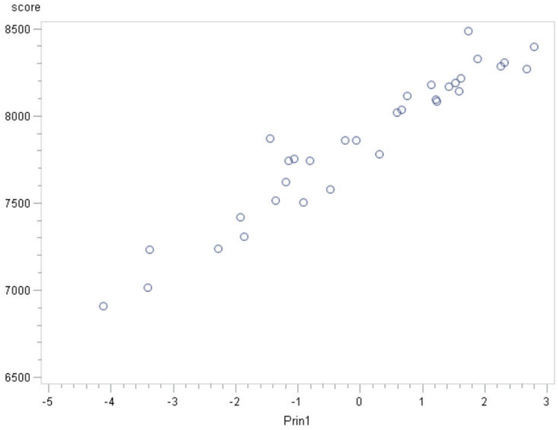


Figure 3: 1st PC v.s. 2nd PC



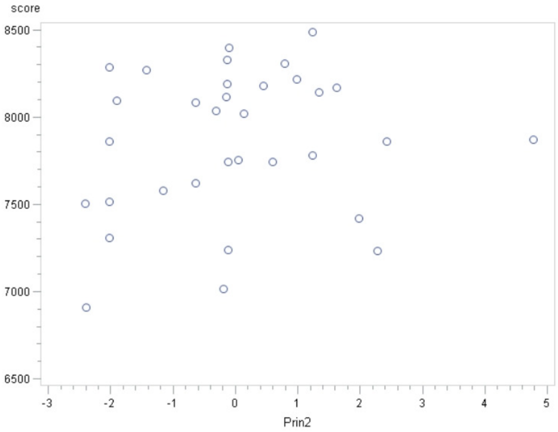
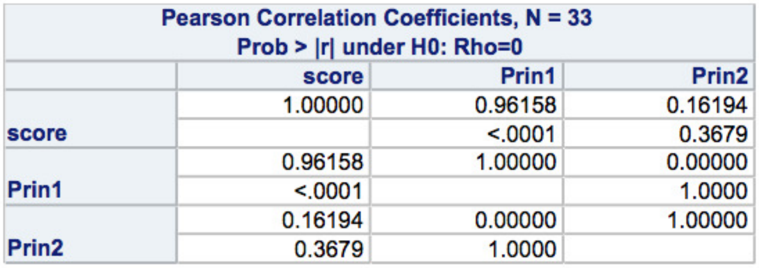


Figure 4: Total Score v.s. 1st PC Figure 5: Total Score v.s. 2nd PC

(f).



Pearson Correlation Coefficients, N = 33
Prob > |r| under H0: Rho=0

	score	Prin1	Prin2
score	1.00000	0.96158	0.16194
		<.0001	0.3679
Prin1	0.96158	1.00000	0.00000
	<.0001		1.0000
Prin2	0.16194	0.00000	1.00000
	0.3679	1.0000	

Figure 6: Correlation Matrix among PCs

To compute the Pearson correlation coefficients among the total score, 1st PC score and 2nd PC score. The correlation coefficient between total score and 1st PC score is 0.96158 while that between total score and 2nd PC score is only 0.16194. This is in accordance with the result of last part. Thus, we may guess that the total score should be measured in a way like 1st PC, i.e., a kind of overall mean, not some comparisons.

▶ **2. Solution.** (a). Because X_i are standardized random variables(zero mean and one standard deviation), then the covariance of X_i is exactly the correlation of X_i. So

$$\begin{aligned} corr(X_i, X_k) &= cov(X_i, X_k) = cov(a_iF + e_i, a_kF + e_K) \\ &= a_ia_kvar(F) + a_icov(F, e_k) + a_kcov(e_i, F) + cov(e_i, e_k) = a_ia_k \\ corr(X_j, X_k) &= cov(X_j, X_k) = cov(a_jF + e_j, a_kF + e_K) \\ &= a_ja_kvar(F) + a_jcov(F, e_k) + a_kcov(e_j, F) + cov(e_j, e_k) = a_ja_k \end{aligned}$$

where i, j, k are mutually different. Thus the ratio of pair of rows (i, j) is always

$$\frac{corr(X_i, X_k)}{corr(X_j, X_k)} = \frac{a_i}{a_j}$$

for all $k \neq i, j$.

(b).

$$corr(\boldsymbol{X}) = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{pmatrix} \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \end{pmatrix} + var(\boldsymbol{e})$$

(c).

$$\boldsymbol{X} = \boldsymbol{a}F + \boldsymbol{e}$$

where

- $\boldsymbol{X} = \begin{pmatrix} X_1 & X_2 & \cdots & X_6 \end{pmatrix}' \sim \mathrm{N}(\mathbf{0}, corr(\boldsymbol{X}))$
- $F \sim \mathrm{N}(0,1)$
- $\boldsymbol{a} = \begin{pmatrix} a_1 & a_2 & \cdots & a_6 \end{pmatrix}'$
- $\boldsymbol{e} = \begin{pmatrix} e_1 & e_2 & \cdots & e_6 \end{pmatrix}' \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Psi})$
- F and $\boldsymbol{e}$ are independent.

(d). For my perspective, only one factor should be included. Because

1. only the eigenvalue of the 1st factor is larger than 1, no matter how many factor we include in the analysis, and the proportion of variability accounted by the 1st factor is at least 100%.
2. to test the hypothesis

   $$H_0: \text{ 1 Factor is sufficient}$$

   the P-value is 0.9805, so we cannot reject the null hypothesis, that is, more factors are not necessary.
3. the AIC and BIC of one factor are lower than those of two factors.

(e). The result of "Factor Pattern" can give the information about factor loadings of different variable, i.e. the a_i in the model in part (c). For example, the variable C has a loading of 0.95611(a very high loading) in Factor1, and the variable F has a loading of 0.87081(a moderately to high loading) in Factor1, and so on. All variables have a at least moderately to high loading in Factor1, and the loadings are not different so much.

In addition, we can know how much proportion of variance of each variable is accounted by the communality from the "Factor Pattern" output or "Final Communality" output. For example,

$$\text{Communality of C} = a_1^2 = 0.95611^2 = 0.9141$$

which means 91.41% variability of C comes from common factor and only 8.69% variability come from specific factor.

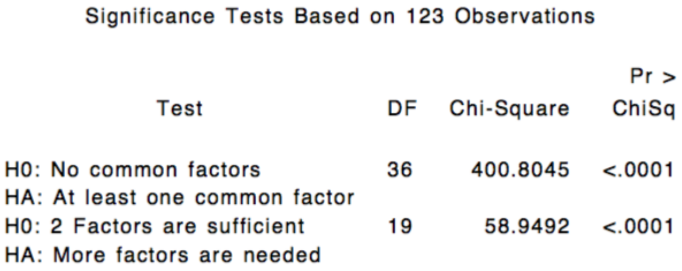
(f). No. If the rotation is help, let T become the rotation operator. Then we can have that

$$F^* = TF$$

where the F^* is the factor after rotating. So the F^* should also have the restriction of length equal to 1, i.e. $var(F^*) = 1$. This implies that $T^2 = 1$, and so $T = \pm 1$. So rotation cannot change the value of loadings except the sign. Thus, there is no need to perform a rotation.

(g). Because we choose to include only one factor in the model and the loadings and communality are high comparing to specific factor. That is to say, all variable perform like the only one common factor, so they should perform similarly to each other. Then the correlations between them should looks same.

▶ **3. Solution.** (a).



```
           Significance Tests Based on 123 Observations

                                                         Pr >
               Test                  DF   Chi-Square    ChiSq

H0: No common factors                36     400.8045   <.0001
HA: At least one common factor
H0: 2 Factors are sufficient         19      58.9492   <.0001
HA: More factors are needed
```

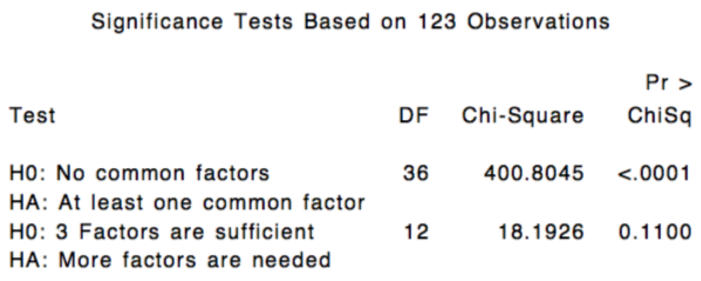
Figure 7: Likelihood Tests

No. From the second test in Figure **??**, the P-value less than 0.0001 means we should reject the null hypothesis, that is, two factors is not sufficient.

(b). Yes. From the second test in Figure **??**, the P-value=0.1100 means we do not reject the null hypothesis, that is, 3 factors is sufficient.

From the Figure **??**, we can see that $p1, p3, p4, p8, p9$ have high loadings in factor 1, and all there statements have mentioned doctor. So maybe the factor 1 is a "doctor factor". $p6, p7$ have high loadings in factor 2, and all these statements is about themselves without some specific reasons. So maybe the factor 2 is a "subjective personal factor". $p2, p5$ have high loadings in factor 3, and all these statements is also about themselves but more reasonable. So maybe the factor 3 is a "objective personal factor".

(c). From the Figure **??**, we can see that the factor pattern does not change

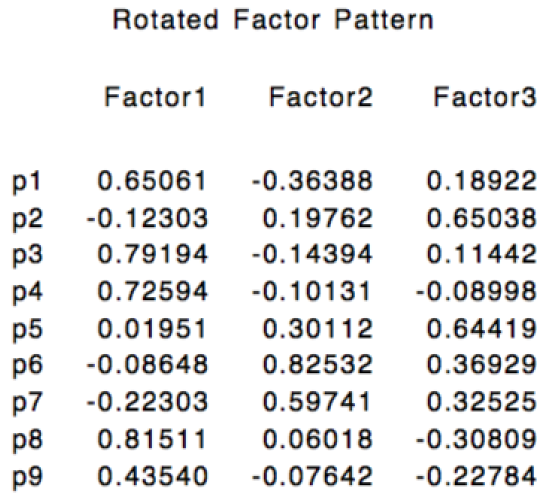


```
          Significance Tests Based on 123 Observations

                                                          Pr >
Test                                DF    Chi-Square     ChiSq

H0: No common factors               36      400.8045    <.0001
HA: At least one common factor
H0: 3 Factors are sufficient        12       18.1926    0.1100
HA: More factors are needed
```

Figure 8: Likelihood Tests

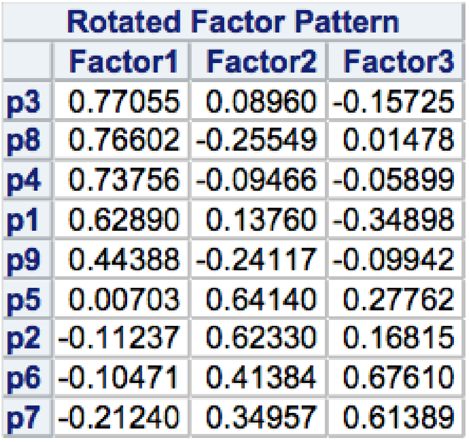


```
        Rotated Factor Pattern

         Factor1      Factor2      Factor3


p1       0.65061     -0.36388      0.18922
p2      -0.12303      0.19762      0.65038
p3       0.79194     -0.14394      0.11442
p4       0.72594     -0.10131     -0.08998
p5       0.01951      0.30112      0.64419
p6      -0.08648      0.82532      0.36929
p7      -0.22303      0.59741      0.32525
p8       0.81511      0.06018     -0.30809
p9       0.43540     -0.07642     -0.22784
```

Figure 9: Rotated Factors

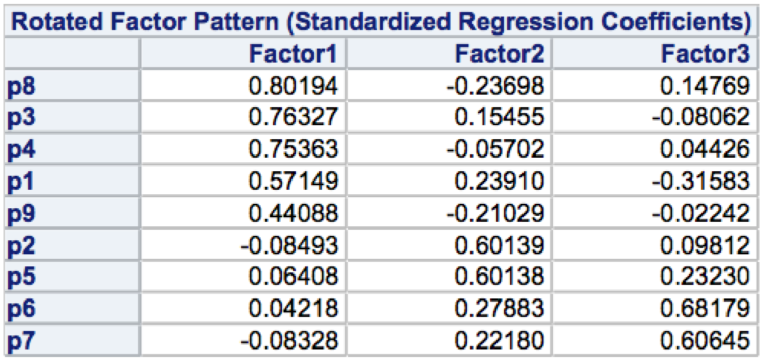


Rotated Factor Pattern			
	Factor1	Factor2	Factor3
p3	0.77055	0.08960	-0.15725
p8	0.76602	-0.25549	0.01478
p4	0.73756	-0.09466	-0.05899
p1	0.62890	0.13760	-0.34898
p9	0.44388	-0.24117	-0.09942
p5	0.00703	0.64140	0.27762
p2	-0.11237	0.62330	0.16815
p6	-0.10471	0.41384	0.67610
p7	-0.21240	0.34957	0.61389

Figure 10: Rotated Factors

too much between MLE and principle factor method; the clusters are still same: $\{p3, p8, p4, p1, p9\}$, $\{p5, p2\}$, $\{p6, p7\}$, only the loadings change slightly. But the interpretation becomes a little different. Because principle factor method uses the philosophy of PCA, so the factor interpretation may be like the interpretation of PCs. That is, factor 1 is the overall mean of $\{p3, p8, p4, p1, p9\}$, factor 2 is the overall mean of $\{p5, p2\}$ and factor 3 is the overall mean of $\{p6, p7\}$.

(d). The overall result of principle factor method with oblique rotation is not so different from the result of other methods. It also only has some differences about factor loadings. The clusters and the interpretations of factors are same as the principle factor method with orthogonal rotation. However, for oblique rotation, we



Rotated Factor Pattern (Standardized Regression Coefficients)			
	Factor1	Factor2	Factor3
p8	0.80194	-0.23698	0.14769
p3	0.76327	0.15455	-0.08062
p4	0.75363	-0.05702	0.04426
p1	0.57149	0.23910	-0.31583
p9	0.44088	-0.21029	-0.02242
p2	-0.08493	0.60139	0.09812
p5	0.06408	0.60138	0.23230
p6	0.04218	0.27883	0.68179
p7	-0.08328	0.22180	0.60645

Figure 11: Oblique Rotated Factors

should also see the correlation between latent factors. In the problem the maximum correlation coefficient is 0.33129, which is not significant. So we can stop here. If the correlation between factors is significant, we can go on and perform a factor analysis to get the second order factor, which may contain more general information.