Lab Assignment No. 7: Answer Key

1) How many principal components account for at least 80% of the data? How many principal components would you retain based on the usual eigenvalue criterion?

The first (and largest) eigenavalue produced by the *proc princomp* procedure accounts for 82.78% of the variance in the data. The matrix that is being decomposed by *proc princomp* to get the eigenvalues (and corresponding eigenvectors) is the correlation matrix of the submitted raw data.

Table 1. Eigenvalues of the correlation matrix

No.	Eigenvalue	Proportion
1	6.62	.8278
2	.88	.1097
3	.16	.0199
4	.12	.0155
5	.08	.0100
6	.07	.0085
7	.05	.0058
8	.02	.0028

It follows from Table 1 that using the standard Kaiser criterion of keeping components (and factors) for all eigenvalues greater than one, that we would keep a single component accounting for the dataset at hand.

2) Provide a correlation matrix of all principal components and items analyzed. What information can you gather from the correlation matrix?

As it can be seen from Table 2 the correlations of the variables with the first component scores are the highest. This is unsurprising seeing as the first component accounts for over 80% of the data's variance. Five additional variables (m100, m200, m1500, m2000 and m2500) also correlated significantly with the second component. It is probable that a good portion of the remainder of the unaccounted variance for those variables is explained by the second component. Interestingly, the correlations for the last four variables are all negative. Given that these variables are time measurements for distances ran it does seem peculiar that the second component should have negative correlations.

Table 2. Correlations between variables and principal components

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
m100	.82**	.53**	.13	.04	.07	-.15	.03	.02
m200	.87**	.43**	.14	-.09	-.04	.17	-.02	-.01
m300	.92**	.23	-.22	.23	-.06	.04	-.00	-.00
m500	.95**	.01	-.21	-.17	.15	-.00	-.05	-.00
m1000	.96**	-.13	-.06	-.14	-.14	-.04	.13	.02
m1500	.94**	-.29*	.08	.01	-.07	-.04	-.13	.08
m2000	.94**	-.29*	.07	.03	-.04	-.06	-.04	-.12
m2500	.88**	-.41**	.10	.10	.14	.08	.08	.02

* $p < .05$, ** $p < .01$

3) Conduct a factor analysis (using the principal components method) extracting the maximum number of factors. Explain the utility and meaning (in your own words) of the eigenvalues, eigenvectors and factor loadings.

Using the principal components analysis method, the *proc factor* will produce output almost identical to that produced by *proc princomp*. Again, the procedure will decompose the correlation matrix and the results are comparable to those we have seen earlier.

Without rotation all items appear to load on the same factor. This is because of the first factor being equivalent to the first component in this method and therefore it account for most of the communality between the items.

4) Run a factor analysis (using the maximum likelihood method) using the eigenvalue criterion as an initial guide for the number of factors to be extracted. What are your initial conclusions? What corrective steps would you have to take?

As can be seen from the pre-rotated eigenvalues the maximum likelihood procedure does not decompose the observed data the same way the principal components method did.

Also, we can observe that because of this discrepancy there appears to be a difference in the total number of factors that would be retained. Here the eigenvalues have two that are over one. Therefore the SAS maximum likelihood (ML) procedure would retain two factors (rather than one which we would have

retained if the previous eigenvalues would have been observed). This peculiarity in the ML procedure is further substantiated when looking at the un-rotated factor patterns. No noticeable differentiation between the two factors can be made across the two factors. All items load highly on the first factor and not cohesively on the second one. This provides further evidence that perhaps a one factor solution would be adequate.

The corrective steps to be taken here would be to investigate the data further. It is noteworthy at this point that there seems to be a discrepancy in the units of measurement in times (after all it is not possible that one should run the 1000 meter distance in less time than the 500 meter one). But given the limited information on the dataset it may not be feasible for us to try and rescale the latter few variables.

Judging from the SAS provided output the two factor solution is an adequate fit in explaining the model's underlying latent structure, $\chi^2\ (13) = 16.36$, $p > .05$.

5) Report and discuss whether the correlation matrix is adequate to be used in this factor analysis. Report the measures of sampling adequacy and the appropriate Chi-square statistic.

One of the first measures consulted when considering the adequacy of a correlation matrix to be decomposed is Bartlett's test of sphericity. This test assesses whether the correlation matrix's off diagonal elements are significantly different from zero. Bartlett's test is thus defined as:

$$\chi^2 = -\left[(N-1) - \left(\frac{2k+5}{6} \right) \right] \times \ln(|R|) \ .$$

In our particular example the corresponding test statistic was significant indicating that the correlation matrix has appropriate off-diagonal elements, $\chi^2\ (28) = 719.11$, $p < .001$.

Further the measures of sampling adequacy are employed in order to determine whether the items in the observed dataset contribute to the overall underlying communality. The *Kaiser-Meyer-Olkin* (KMO) test measures the sampling adequacy (MSA) by comparing the magnitudes of correlation coefficients. Ideally, KMO ought to be above .70 for the overall dataset and similar for each individual item considered.

The ML procedure provides this information, summarized in Table 3, with an overall KMO = .91. Each of the items in the original dataset appear to be adequate for the purpose of decomposing the derived correlation matrix. Furthermore, of the eight variables it is noteworthy that not a single variable had a MSA value less than .84 (see Table 3).

Table 3. Measures of sampling adequacy for variables

m100	m200	m300	M500
.84	.87	.96	.95
m1000	m1500	m2000	m2500
.93	.91	.87	.93

6) Fit a four factor (maximum likelihood) model. Is this model a better fit than the one in question 4?

Much like in the two factor solution without a rotation little discernable pattering is possible. Again the model appears to be of an adequate fit, $\chi^2\ (2) = .76$, $p > .05$, however, this may be deceptive. Due to the over factoring there is little utility in knowing the factor loadings of the items since again only the first factor dominates the loadings of the individual variables.

If a chi-square difference test was to be computed one could see that the difference between the two model fits was not significant, $\chi^2\ (11) = 15.60$, $p > .05$. By the parsimonious rule we would therefore prefer the smaller and simpler model rather than the more complex one.

7) Which distances load on which factors in the four factor model?

This is somewhat of a trick question. As can be seen from the chi-square difference test and the previous (2-factor) factor loadings the variables do not load well on the four factor solution.

Table 4. Four factor solution factor loadings

	Factor1	Factor2	Factor3	Factor4
m100	0.84	-0.53	0.05	0.00
m200	0.87	-0.37	-0.06	-0.04
m300	0.90	-0.17	-0.11	0.06
m500	0.93	0.03	-0.26	0.12
m1000	0.95	0.17	-0.16	-0.11
m1500	0.93	0.31	0.05	0.00
m2000	0.94	0.31	0.08	0.00
m2500	0.86	0.40	0.09	0.08

Alternatively, if one were to consider the rotated (using a Varimax or a Promax rotation) solutions some items would load more noticeably on the other factors.

Table 5. Varimax rotated factor loadings

	Factor1	Factor2	Factor3	Factor4
m100	0.28	0.95	0.02	-0.01
m200	0.39	0.85	0.12	0.08
m300	0.53	0.72	0.22	0.01
m500	0.67	0.59	0.40	0.03
m1000	0.79	0.50	0.23	0.22
m1500	0.90	0.39	0.09	0.04
m2000	0.91	0.40	0.07	0.03
m2500	0.91	0.28	0.09	-0.05

Table 5 shows the rotated factor loadings where items one, tow and three appear to be loading more highly on the second factors. However, item four already shows problematic irreconcilable cross loadings between the first three factors. This is yet another piece of evidence demonstrating that the four factor solution simply is not a manageable/workable one.

8) Discuss the two rotations (Varimax and Promax) in a 2 factor model. Which one is a better fit? Why?

Orthogonal rotations, of which Varimax is one of, are rotations which leave the factors orthogonal (un-correlated) to one another. Oblique rotations, such as Promax, allow for the factors to be correlated, which is the case in most behavioral and social science latent constructs.

Table 6 summarizes the factor loadings following a Varimax rotation. With the exception of the fourth item, m500, the remaining seven items loaded clearer on one of the two factors. Items one, tow and three had all loadings above .70 on the second factor, while items five through eight had all loadings above .80 on the first factors. The Promax solution, on the other hand, was not as simple as the previous orthogonal rotation. Here the factor patterns showed relatively high loadings on both factors, mainly because we allowed for the factors to be correlated. Apparently the correlation, r = -.72, was high enough to create the rather non-ignorable cross loadings. As can be seen in Table 7 the general patterns earlier observed in Table 6 are still present, however, the cross loadings have no significantly increased.

Table 6. Varimax rotated two-factor solution

	Factor1	Factor2
m100	0.29	0.91
m200	0.38	0.88
m300	0.54	0.75
m500	0.69	0.62
m1000	0.80	0.53
m1500	0.90	0.40
m2000	0.91	0.40
m2500	0.91	0.28

Table 7. Promax rotated two-factor solution

	Factor1	Factor2
m100	0.63	0.95
m200	0.70	0.96
m300	0.79	0.90
m500	0.88	0.84
m1000	0.94	0.80
m1500	0.98	0.72
m2000	0.99	0.72
m2500	0.95	0.62

9) Briefly describe the differences and similarities between principal components and factor analysis. How do these two approaches provide different information in the analysis you have conducted?

Principal components as was discussed earlier is predominantly a data reduction method that uses a linear combination of the observed variables, using all of the variance available, to create components. These components are unaffected by

the inclusion or reduction in total number of components (not to exceed the number of variables of course) since the linear combination property still holds no matter how many components are modeled.

Factor analysis on the other hand is a way of investigating an underlying latent structure and interrelationship between variables. This approach allows for a much more insightful understanding of the communality of variables and uniqueness between them.

In light of the current data both approaches provide similar information in that we have seen that one or two factors/components account for most of the data's variance. Whereas the factor approach would suggest possible two factors, the principal components would have been happy with a single component rather than two. More so a side product of the nature of the data, both solutions had a difficult time with the fourth item (m500). Further research and review of this data would be needed in order to address some of the potential confounds in this dataset.

10) What are some of your final remarks/observations regarding your analysis?

As can be seen from the two rotated (Varimax and Promax) solutions of the two-factor model there seems to be a qualitative difference between the distances run. Perhaps there is an issue of endurance or extended training that sets apart distances below 500 meters and above 500 meters, which would account for the two factors with the observed respective item factor loadings. This, however, still leaves the questions of the 500 meter item. Unclear whether it belongs to the first or second factor it can be said that perhaps it lies at such a distance midpoint were it poses an increased level of difficulty for only some of the runners, but not for others. Perhaps those who had participated in long distance matches did not experience this distance as trying as those who specialize in shorter ones. This would certainly account for some of the peculiarity that is going on with this item given the rest of this data.