Math 5364
Data Mining 2
Homework 31
Mary Barker

1. The file Hw31data.txt contains SAS code for generating two data sets. The first data set provides the correlation matrix of six measurement made on white leghorn fowls, including skull length(SL), skull breadth (SB), humerus length (HS), ulna length (UL), femur length (FL), and tibia length(TL);

```
%include '/folders/myshortcuts/sas_folder/Hw31Data.txt';
```

   (a) Perform a principal components analysis for this data set, and report the resulting eigenvalues and eigenvectors;

```
proc princomp data = leghorn;
run;
```

The PRINCOMP Procedure

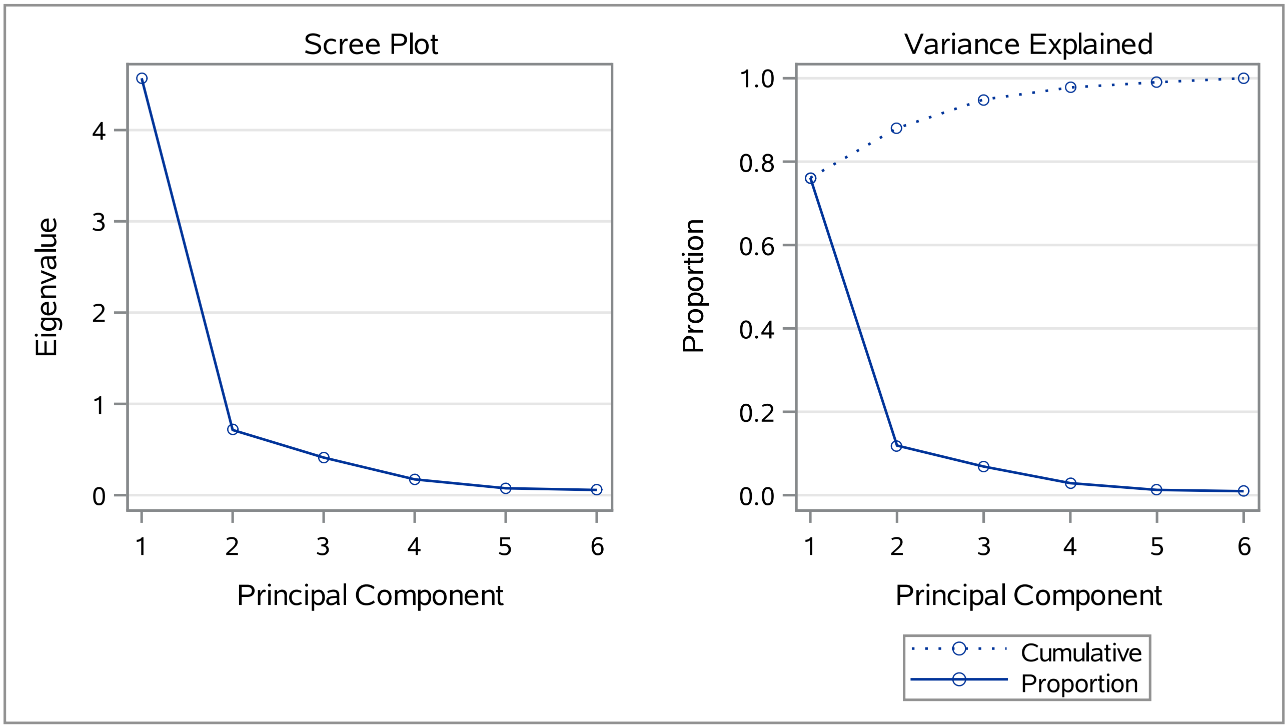
Observations	10000
Variables	6

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.56757080	3.85344753	0.7613	0.7613
2	0.71412326	0.30199429	0.1190	0.8803
3	0.41212898	0.23894007	0.0687	0.9490
4	0.17318890	0.09733018	0.0289	0.9778
5	0.07585872	0.01872938	0.0126	0.9905
6	0.05712934		0.0095	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
sl	0.347439	0.536974	-.766673	0.049099	0.027212	0.002372
sb	0.326373	0.696467	0.636305	0.002033	0.008044	0.058827
hl	0.443419	-.187301	0.040071	-.524077	0.168397	-.680939
ul	0.439983	-.251382	-.011196	-.488771	-.151153	0.693796
fl	0.434544	-.278168	0.059205	0.514259	0.669483	0.132738
tl	0.440150	-.225698	0.045735	0.468582	-.706953	-.184077



(b) How many principal components are required to explain at least 90% of the total variation in the data? ;

3

(c) Provide an intuitive interpretation for the principal components accounting for 90% of the total variation. (For example, the first principal component has large positive coefficients for all of the variables in the data set, so it roughly measure the overall size of a white leghorn fowl.);

The first principal component has relatively uniform values for coefficients, giving overall an indication of how big each animal is.

The second principal component has large coefficients for SB and SL, and negative, not to say small coefficients for the other variables, indicating a description of just how big the skull is.

The third component has larger absolute values for SB and SL, as with the second, but the value for SL is negative, so this seems to evaluate the difference between skull width and height.

2. Perform a factor analysis on the leghorn data;

```
proc factor data=leghorn res;
run;
```

(a) How many factors are retained using the MINEIGEN criterion?

```
proc factor data=leghorn mineigen=0.05;
run;
```

Only one factor was retained.

The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 6 Average = 1				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	4.56757080	3.85344753	0.7613	0.7613
2	0.71412326	0.30199429	0.1190	0.8803
3	0.41212898	0.23894007	0.0687	0.9490
4	0.17318890	0.09733018	0.0289	0.9778
5	0.07585872	0.01872938	0.0126	0.9905
6	0.05712934		0.0095	1.0000

1 factor will be retained by the MINEIGEN criterion.

Factor Pattern	
	Factor1
sl	0.74254
sb	0.69752
hl	0.94767
ul	0.94033
fl	0.92870
tl	0.94068

Variance Explained by Each Factor
Factor1
4.5675708

Final Communality Estimates: Total = 4.567571					
sl	**sb**	**hl**	**ul**	**fl**	**tl**
0.55137044	0.48653453	0.89807737	0.88421400	0.86248933	0.88488513

Residual Correlations With Uniqueness on the Diagonal						
	sl	**sb**	**hl**	**ul**	**fl**	**tl**
sl	0.44863	0.06606	-0.08869	-0.09723	-0.11960	-0.09850
sb	0.06606	0.51347	-0.08502	-0.12590	-0.12179	-0.10115
hl	-0.08869	-0.08502	0.10192	0.04888	-0.00510	-0.01346
ul	-0.09723	-0.12590	0.04888	0.11579	0.00372	0.00145

The FACTOR Procedure
Initial Factor Method: Principal Components

Residual Correlations With Uniqueness on the Diagonal						
	sl	sb	hl	ul	fl	tl
fl	-0.11960	-0.12179	-0.00510	0.00372	0.13751	0.05038
tl	-0.09850	-0.10115	-0.01346	0.00145	0.05038	0.11511

Root Mean Square Off-Diagonal Residuals: Overall = 0.08123310					
sl	sb	hl	ul	fl	tl
0.09559288	0.10247468	0.05948076	0.07444404	0.07964388	0.06731137

Partial Correlations Controlling Factors						
	sl	sb	hl	ul	fl	tl
sl	1.00000	0.13764	-0.41474	-0.42662	-0.48153	-0.43343
sb	0.13764	1.00000	-0.37164	-0.51633	-0.45834	-0.41603
hl	-0.41474	-0.37164	1.00000	0.44997	-0.04311	-0.12423
ul	-0.42662	-0.51633	0.44997	1.00000	0.02945	0.01256
fl	-0.48153	-0.45834	-0.04311	0.02945	1.00000	0.40046
tl	-0.43343	-0.41603	-0.12423	0.01256	0.40046	1.00000

Root Mean Square Off-Diagonal Partials: Overall = 0.36163739					
sl	sb	hl	ul	fl	tl
0.39816909	0.40170085	0.32554128	0.36113729	0.34786336	0.32769073

(b) What is the overall RMS off-diagonal residuals in this case?
0.08123310

3. Continuing with the leghorn data set, increase the number of factors until the overall residual RMS is less than 0.05;

```
proc factor data=leghorn nfact=3 res;
run;
```

The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 6 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.56757080	3.85344753	0.7613	0.7613
2	0.71412326	0.30199429	0.1190	0.8803
3	0.41212898	0.23894007	0.0687	0.9490
4	0.17318890	0.09733018	0.0289	0.9778
5	0.07585872	0.01872938	0.0126	0.9905
6	0.05712934		0.0095	1.0000

3 factors will be retained by the NFACTOR criterion.

Factor Pattern			
	Factor1	Factor2	Factor3
sl	0.74254	0.45377	-0.49218
sb	0.69752	0.58856	0.40849
hl	0.94767	-0.15828	0.02572
ul	0.94033	-0.21243	-0.00719
fl	0.92870	-0.23507	0.03801
tl	0.94068	-0.19073	0.02936

Variance Explained by Each Factor		
Factor1	Factor2	Factor3
4.5675708	0.7141233	0.4121290

Final Communality Estimates: Total = 5.693823					
sl	sb	hl	ul	fl	tl
0.99952600	0.99979667	0.92379166	0.92939318	0.91919114	0.92212437

Residual Correlations With Uniqueness on the Diagonal						
	sl	sb	hl	ul	fl	tl
sl	0.00047	0.00004	-0.00420	-0.00437	0.00577	0.00250
sb	0.00004	0.00020	-0.00237	0.00207	0.00104	-0.00089
hl	-0.00420	-0.00237	0.07621	0.01544	-0.04329	-0.04440
ul	-0.00437	0.00207	0.01544	0.07061	-0.04595	-0.03886
fl	0.00577	0.00104	-0.04329	-0.04595	0.08081	0.00443
tl	0.00250	-0.00089	-0.04440	-0.03886	0.00443	0.07788

The FACTOR Procedure
Initial Factor Method: Principal Components

Root Mean Square Off-Diagonal Residuals: Overall = 0.02282157					
sl	**sb**	**hl**	**ul**	**fl**	**tl**
0.00390799	0.00153294	0.02866001	0.02786663	0.02842195	0.02648714

Partial Correlations Controlling Factors

	sl	sb	hl	ul	fl	tl
sl	1.00000	0.13484	-0.69899	-0.75611	0.93278	0.41152
sb	0.13484	1.00000	-0.60212	0.54563	0.25549	-0.22242
hl	-0.69899	-0.60212	1.00000	0.21052	-0.55161	-0.57635
ul	-0.75611	0.54563	0.21052	1.00000	-0.60828	-0.52399
fl	0.93278	0.25549	-0.55161	-0.60828	1.00000	0.05590
tl	0.41152	-0.22242	-0.57635	-0.52399	0.05590	1.00000

Root Mean Square Off-Diagonal Partials: Overall = 0.53049547					
sl	**sb**	**hl**	**ul**	**fl**	**tl**
0.65082823	0.39829632	0.55351877	0.55826743	0.56793631	0.40710947

(a) How many factors are required to achieve this?
3 factors

(b) Report the estimated matrices hatL and hat Phi

(c) What is the communality for skull length?
≈ 0.9995.

(d) Find the unique variance of ulna length
0.07061

(e) What is the correlation between femur length and the 2nd factor?

4. The second data set in Hw31data.txt contains responses of 122 diabetes patients to 25 survey questions, on a Likert scale (a scale typically used on surveys, where 1 = Strongly Disagree, 2 = Somewhat Disagree, 3 = Neither DIsagree Nor Agree, 4 = Somewhat Agree, and 5 = Strongly Agree).

   (a) Perform a factor analysis with nfact=17 on this data and store the factor scores in a data set.

```
proc factor data=diabetes score nfact=17 res out = fact_scores;
run;
```

The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 25 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.82744353	1.85080952	0.1931	0.1931
2	2.97663401	1.37717999	0.1191	0.3122
3	1.59945402	0.06823351	0.0640	0.3761
4	1.53122051	0.10057945	0.0612	0.4374
5	1.43064106	0.12163229	0.0572	0.4946
6	1.30900876	0.08035333	0.0524	0.5470
7	1.22865544	0.05601538	0.0491	0.5961
8	1.17264006	0.13721370	0.0469	0.6430
9	1.03542637	0.17273454	0.0414	0.6844
10	0.86269182	0.03464371	0.0345	0.7190
11	0.82804811	0.02776081	0.0331	0.7521
12	0.80028730	0.08836069	0.0320	0.7841
13	0.71192661	0.08944024	0.0285	0.8126
14	0.62248637	0.02826278	0.0249	0.8375
15	0.59422360	0.03504807	0.0238	0.8612
16	0.55917552	0.08099790	0.0224	0.8836
17	0.47817762	0.03572211	0.0191	0.9027
18	0.44245551	0.07581700	0.0177	0.9204
19	0.36663851	0.04246366	0.0147	0.9351
20	0.32417485	0.00818368	0.0130	0.9481
21	0.31599117	0.01818378	0.0126	0.9607
22	0.29780739	0.04030049	0.0119	0.9726
23	0.25750690	0.02065067	0.0103	0.9829
24	0.23685623	0.04642752	0.0095	0.9924
25	0.19042871		0.0076	1.0000

17 factors will be retained by the NFACTOR criterion.

The FACTOR Procedure
Initial Factor Method: Principal Components

Factor Pattern										
	Factor1	**Factor2**	**Factor3**	**Factor4**	**Factor5**	**Factor6**	**Factor7**	**Factor8**	**Factor9**	**Factor10**
x1	0.24141	-0.14170	-0.40100	0.28573	-0.15424	-0.03261	0.24639	0.23705	0.62435	0.02075
x2	0.06313	0.41673	0.26577	-0.22142	0.30725	0.30155	0.38113	0.08680	-0.05168	0.08866
x3	0.44312	0.51789	-0.17461	0.08515	0.19031	-0.04759	-0.42145	-0.21045	-0.01099	-0.07860
x4	0.02892	0.43713	-0.09896	-0.37412	0.04161	-0.27051	0.35450	-0.19115	-0.01274	0.39987
x5	-0.19535	-0.40344	0.06617	0.36516	0.01143	0.26708	0.49647	-0.35243	-0.05685	-0.06061
x6	0.52891	-0.37879	0.22067	0.12386	0.44229	0.02575	-0.01895	-0.18655	0.00008	-0.03887
x7	-0.20342	0.52388	0.05456	0.29786	0.22401	0.22527	-0.35219	0.11804	-0.08260	0.16773
x8	0.67178	-0.16134	-0.25770	0.16308	-0.19492	-0.16009	-0.06561	-0.06067	0.10400	0.08213
x9	0.25373	0.40675	0.25681	0.04541	-0.55449	-0.00696	0.23151	0.10938	-0.23557	0.27685
x10	0.51183	0.43687	-0.19568	-0.09793	0.07796	-0.04696	0.22053	-0.13227	-0.08851	-0.29176
x11	0.63204	0.44020	-0.07242	0.20035	-0.04693	0.04978	-0.07277	0.00535	-0.10828	0.00758
x12	0.12452	0.54656	-0.24744	0.09200	0.44948	-0.04530	0.16709	0.07176	0.25116	0.16262
x13	0.30655	0.33490	0.27594	0.20215	-0.09814	0.45263	-0.00396	-0.40197	0.10976	-0.13355
x14	0.23177	-0.50354	0.32044	0.17911	0.20034	0.02258	-0.16987	0.34523	-0.02547	0.27708
x15	0.45777	-0.04731	-0.13110	0.10419	-0.09228	0.53327	0.13448	0.29244	-0.12009	0.00911
x16	0.49019	-0.01773	0.30913	-0.25095	-0.04784	0.29673	-0.04305	0.01936	0.46502	-0.05206
x17	0.50346	-0.14367	0.04376	-0.38366	0.23609	-0.15975	0.13877	0.16821	-0.06432	-0.40475
x18	0.55138	0.38030	0.22162	0.15371	-0.25004	-0.16453	-0.15231	0.17318	0.09629	-0.08205
x19	0.46047	-0.11452	0.15290	0.08549	0.43617	-0.18927	0.13299	0.11337	0.04968	0.20193
x20	0.33369	-0.06316	0.57939	-0.26707	-0.21600	-0.24412	-0.09488	-0.28105	0.27774	0.08926
x21	-0.26833	0.38205	0.28036	0.25510	-0.10678	-0.24600	0.23983	0.42428	0.00292	-0.34376
x22	0.75743	-0.17707	-0.01699	-0.08445	0.01927	-0.12931	0.01132	0.06034	-0.31536	-0.06316
x23	-0.54127	0.21634	0.48738	0.16249	0.19311	-0.09392	0.02285	0.13744	0.07329	-0.08530
x24	0.29823	-0.10764	0.12783	0.65193	0.03632	-0.38702	0.17697	-0.24754	-0.11968	0.01632
x25	0.71418	-0.28588	0.01077	-0.09605	-0.08171	0.15735	0.03641	0.19032	-0.16771	0.11656

The FACTOR Procedure
Initial Factor Method: Principal Components

Factor Pattern

	Factor11	Factor12	Factor13	Factor14	Factor15	Factor16	Factor17
x1	-0.11615	0.17173	-0.09705	-0.03525	0.03700	0.17475	-0.01953
x2	-0.08845	0.16256	-0.11666	-0.51712	-0.09703	0.04990	-0.01280
x3	0.17845	0.19635	-0.13776	-0.00942	-0.06264	0.00511	-0.00860
x4	0.15782	0.22200	-0.03710	0.30962	0.08611	0.09129	0.12155
x5	-0.02599	-0.08182	0.26523	0.02582	0.24487	0.07034	-0.01329
x6	-0.07452	0.08983	-0.14349	-0.01267	0.06752	-0.23353	0.23179
x7	-0.31184	-0.03139	0.25655	0.13188	0.09994	0.18621	0.04671
x8	0.28706	-0.04745	0.18750	-0.14364	-0.14643	-0.06706	0.12776
x9	-0.10579	-0.17845	0.07976	0.02283	-0.12189	-0.10395	0.11053
x10	-0.03981	0.17012	0.29311	0.07641	-0.14524	0.05678	-0.35803
x11	0.07392	-0.11541	0.12290	-0.17495	-0.06052	0.21390	0.16277
x12	0.08324	-0.12664	0.14229	-0.02613	0.25759	-0.27188	0.04084
x13	0.25410	-0.09836	-0.25343	0.14790	0.09280	0.17840	0.00191
x14	0.19638	0.26637	0.20426	-0.00534	0.05174	0.21645	-0.07062
x15	-0.14349	0.34661	-0.16422	0.28276	-0.13655	-0.17061	0.02878
x16	0.03240	-0.19967	0.26440	0.08324	-0.18716	-0.17724	-0.00670
x17	-0.15465	-0.03440	0.13261	0.14204	-0.00672	0.22125	0.33775
x18	-0.28086	-0.00003	-0.07675	-0.00190	0.28647	-0.11202	-0.03789
x19	-0.09920	-0.45381	-0.26547	0.17142	-0.20104	0.13659	-0.21644
x20	-0.15514	0.22572	0.01125	-0.05226	0.10756	0.09111	-0.07220
x21	0.27663	-0.02953	-0.14862	0.00983	0.10501	0.04269	0.02549
x22	-0.00766	-0.01132	0.02478	-0.02229	0.26093	-0.15261	-0.20797
x23	0.26141	0.15418	0.17027	0.13951	-0.14031	-0.16652	-0.02106
x24	-0.14333	0.14865	-0.00317	-0.01080	-0.22703	-0.05030	0.04833
x25	0.33465	-0.05302	-0.02015	-0.01075	0.14414	0.04906	-0.04119

Variance Explained by Each Factor

Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8	Factor9	Factor10
4.8274435	2.9766340	1.5994540	1.5312205	1.4306411	1.3090088	1.2286554	1.1726401	1.0354264	0.8626918

Factor11	Factor12	Factor13	Factor14	Factor15	Factor16	Factor17
0.8280481	0.8002873	0.7119266	0.6224864	0.5942236	0.5591755	0.4781776

(b) One of the assumptions of the factor model is that cov(f) = identity. Verify that the sample covariance matrix of the factor scores is equal to I (This occurs exactly, because we are using the principal component method for this factor analysis. There are other methods where this does not occur.)

```
proc corr data=fact_scores cov;
        var Factor1-Factor17;
run;
```

The CORR Procedure

Covariance Matrix, DF = 108								
	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8
Factor1	1.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Factor2	0.000000000	1.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Factor3	0.000000000	0.000000000	1.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Factor4	0.000000000	0.000000000	0.000000000	1.000000000	-0.000000000	0.000000000	0.000000000	0.000000000
Factor5	0.000000000	0.000000000	0.000000000	-0.000000000	1.000000000	0.000000000	0.000000000	0.000000000
Factor6	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	1.000000000	0.000000000	0.000000000
Factor7	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	1.000000000	0.000000000
Factor8	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	1.000000000
Factor9	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Factor10	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Factor11	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Factor12	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Factor13	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Factor14	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Factor15	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Factor16	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000
Factor17	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000	0.000000000

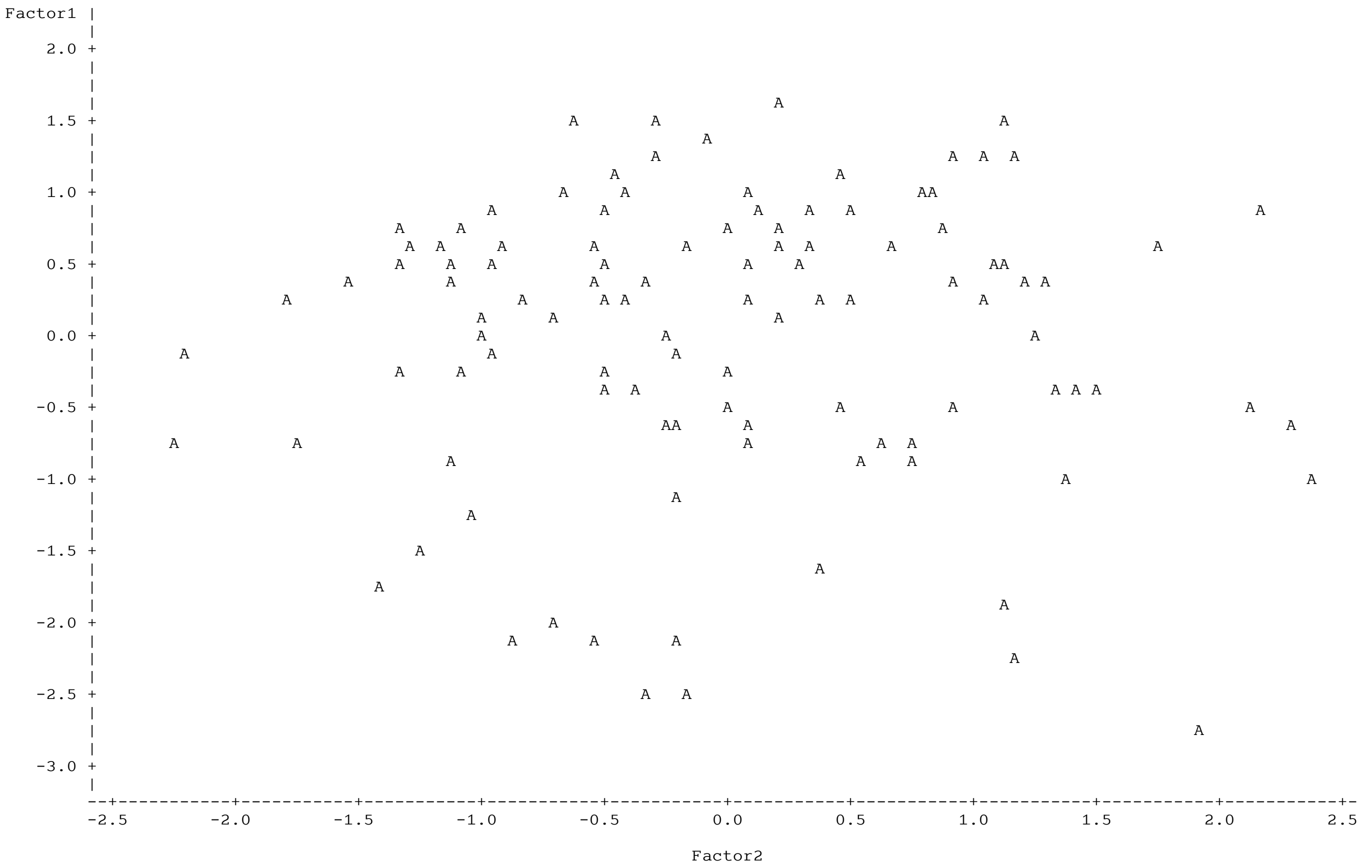
(c) What does $\text{cov}(f) = I$ say about the correlation between the two different factors? What would you expect the scatter plot of the two different factors to look like?

That the covariance matrix is I measn that the factors are uncorrelated, or independent. I would expect the plot to look random, with no underlying pattern between the two variables.

(d) Create a scatter plot of f1 vs f2. Does this plot agree with your expectations? ;

```
proc plot data=fact_scores;
        plot Factor1*Factor2;
run;
```

Plot of Factor1*Factor2. Legend: A = 1 obs, B = 2 obs, etc.



NOTE: 13 obs had missing values.

My initial guess does seem to be supported by the graph, where there is no trend discernible.

(e) It would be interesting to interpret the factors in this problem, but in order to do that, we will need to consider rotations of the factors, so this will have to wait until a future homework assignment.