HBAT Factor Analysis
(July 2019)

Jason Huggy, Student, Lewis University

I. INTRODUCTION

The initiative of this study was to examine data from the HBAT dataset utilized in *Multivariate Data Analysis* [1]. The data set includes 99 lines of unlabeled data, and 23 variables. A variety of methods are utilized to perform factor analysis on the dataset. The primary purpose was to determine what variables could be grouped together into factors and if there was a need to reduce the data variables used. The work in this case can be used as a setup for following multivariate techniques such as discriminate analysis or multiple regression.

II. METHODOLOGY

To conduct this study, SAS Enterprise Guide (EG) is used to produce all figures and data tables. To begin, the HBAT data is imported into EG and verified to be free of missing values or errors. A quick examination of the data shows that the first five columns of data, as well as the last, are nonmetric features. The rest of the features are all metric. For this study only eleven features are utilized. Due to their metric nature, they are easily analyzed using factor analysis.

A. Correlation Matrix

To begin the first part of this study, a correlation matrix is created to observe the correlation between variables. This is done in EG by going to *Tasks, Multivariate,* and *Correlations.* No further edits were made, and all default settings were left the same. This displays a table of the simple statistics of the data and a correlation matrix between the variables.

B. Deriving Factors and Assessing Overall Fit

In the second part of the study, the variables are processed through factor analysis using SAS Enterprise Guide. This is done using *Tasks, Multivariate,* and then *Factor Analysis.* All default settings are left the same for the initial factor analysis test. However, "Kaiser's Measure of Sampling Adequacy" is checked. In the second part of the factor analysis study, an additional factor analysis is performed on the same data but using orthogonal Varimax rotation. The box to reorder the matrix rows by highest absolute loading was also selected. In the last part of this study, a random sample of the data is selected to perform factor analysis as a method to verify the results. This sample included 60 percent of the data. Orthogonal Varimax rotation is also used for this portion.

III. RESULTS AND DISCUSSION

A. Correlation Matrix

Table 1 in the appendix displays the correlation matrix for the HBAT data with the eleven variables chosen. The first number in each box displays the bivariate correlation between the two variables, and the second number displays the correlation significance. As you can see, every single variable shows significant correlation with at least one other variable. The closer the correlation is to 1 or -1, the stronger the relationship. In this case, any relationship that shows a correlation significance of less than 0.0001 represents a strong relationship. This is a good first step to understand the relationships in the data in an effort to begin to formulate factors and to reduce the number of variables needed.

B. Deriving the Factors

In the next part of the study, factor analysis is performed on the variables with the given data. Factor is analysis is performed with and without orthogonal Varimax rotation. The results of factor analysis without any rotation are shown in tables 2 and 3 in the appendix. In table 2, eleven factors are represented with their eigenvalues and proportion of the data. Each factor is sorted from largest to smallest based on their eigenvalue. We are only looking to use factors with an eigenvalue of one or greater, so only the top four factors will be utilized. As can be seen, this represents 79.5 percent of the data, which is not bad. However, the Kaiser's Measure of Sampling Adequacy for this data is only listed as at 0.65. This is considered mediocre, but the the study is continued. Table 3 shows the factor pattern for this analysis, displaying the weights each variable carries in each factor. The weights are not sorted, so it is slightly difficult to see any pattern. This will be improved upon in the next section.

In the next section, an orthogonal Varimax rotation is applied to the data. In table 4 you can see how this has affected the results. Not only does the Varimax rotation produce stronger results as far as the weights of each variable, but now the variables are sorted by their weights. This makes grouping the variables extremely easy. For example, for factor 1 in table 4, the first three entries have significant weights. This allows us to assume their relation. This is confirmed when you look back at the correlation matrix from table 1. Variables 9, 16, and 18 are all very strongly correlated. In turn, this allows us to label each factor based on the characteristics of the correlated variables. For example, factor one relates to Post Sale Customer Service.

In the last part of the experiment, 60 percent of the data was utilized in its own sample for factor analysis. Looking at table 5, there are some slight differences compared to table 4. The variables are in slightly different orders, and the group factors are different as well. Instead of the first factor only including 9, 16, and 18, it now includes variable 11. This change is due to variable 11. As seen in table 5, variable 11 is cross-loading, because it is shows a significant weight in both factor 1 and 4. For this reason, variable 11 should be deleted. After that, the results would look very similar to how they did in table 4. Therefore, even when only taking a portion of the data, the factors still check out.

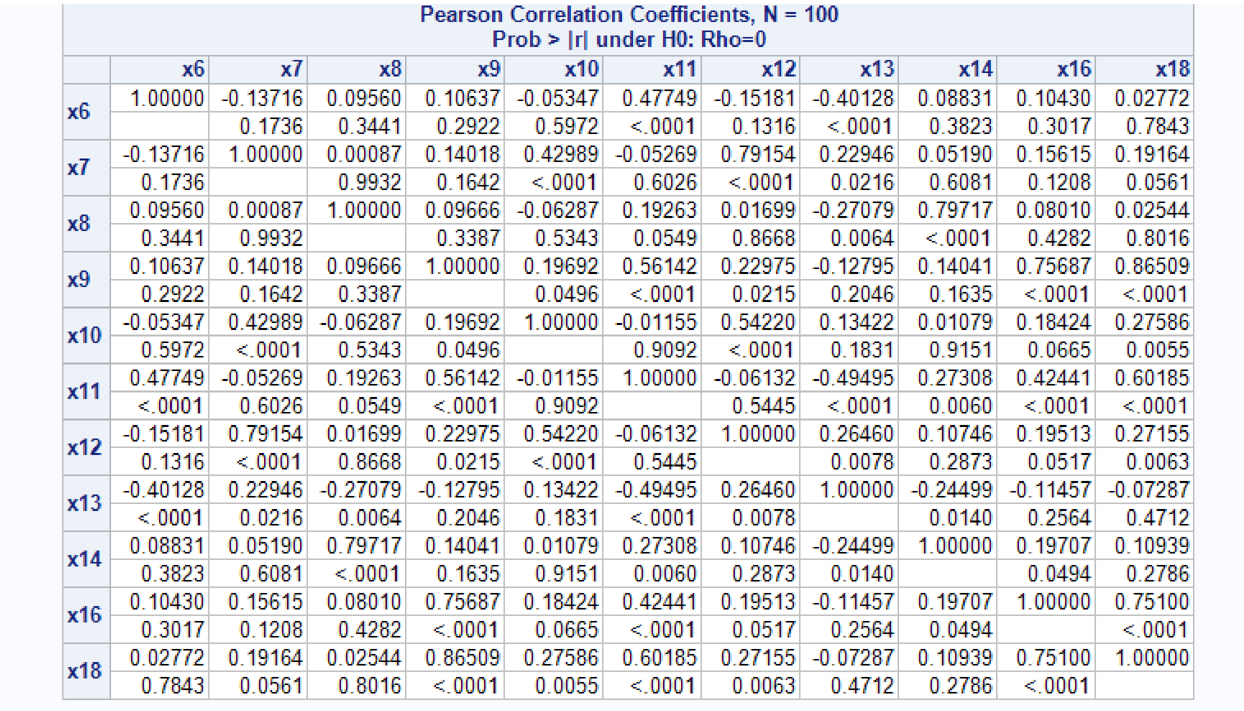
IV. CONCLUSION

This study showed the use of factor analysis for the purpose of deriving factors and for possible dimensionality reduction. The purpose of factor analysis is to try to determine the variables in the dataset that do the most to describe the data. However, there is not always a need to use variables that are highly correlated with another variable if that correlation already impacts the results. Instead, it can be useful to group variables together into factors. This also serves as a great way to learn about the relationships between features. Factor analysis also serves as a great way to identify variables that can cause problems in the use of other multivariate techniques. By reducing the clutter in this step, if needed, it can result in a better overall model.

REFERENCES

[1] J. Hair, W. Black, B. Babin, R. Anderson, *Multivariate Data Analysis: Pearson New International Edition*. 2013.

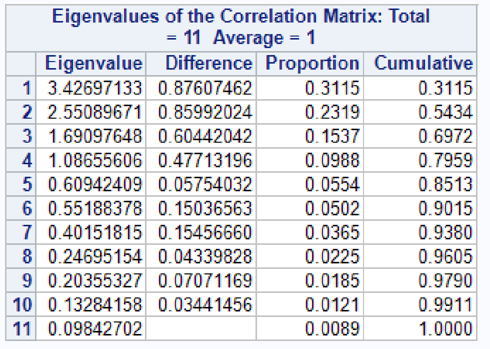
Appendix



Pearson Correlation Coefficients, N = 100
Prob > |r| under H0: Rho=0

	x6	x7	x8	x9	x10	x11	x12	x13	x14	x16	x18
x6	1.00000	-0.13716	0.09560	0.10637	-0.05347	0.47749	-0.15181	-0.40128	0.08831	0.10430	0.02772
		0.1736	0.3441	0.2922	0.5972	<.0001	0.1316	<.0001	0.3823	0.3017	0.7843
x7	-0.13716	1.00000	0.00087	0.14018	0.42989	-0.05269	0.79154	0.22946	0.05190	0.15615	0.19164
	0.1736		0.9932	0.1642	<.0001	0.6026	<.0001	0.0216	0.6081	0.1208	0.0561
x8	0.09560	0.00087	1.00000	0.09666	-0.06287	0.19263	0.01699	-0.27079	0.79717	0.08010	0.02544
	0.3441	0.9932		0.3387	0.5343	0.0549	0.8668	0.0064	<.0001	0.4282	0.8016
x9	0.10637	0.14018	0.09666	1.00000	0.19692	0.56142	0.22975	-0.12795	0.14041	0.75687	0.86509
	0.2922	0.1642	0.3387		0.0496	<.0001	0.0215	0.2046	0.1635	<.0001	<.0001
x10	-0.05347	0.42989	-0.06287	0.19692	1.00000	-0.01155	0.54220	0.13422	0.01079	0.18424	0.27586
	0.5972	<.0001	0.5343	0.0496		0.9092	<.0001	0.1831	0.9151	0.0665	0.0055
x11	0.47749	-0.05269	0.19263	0.56142	-0.01155	1.00000	-0.06132	-0.49495	0.27308	0.42441	0.60185
	<.0001	0.6026	0.0549	<.0001	0.9092		0.5445	<.0001	0.0060	<.0001	<.0001
x12	-0.15181	0.79154	0.01699	0.22975	0.54220	-0.06132	1.00000	0.26460	0.10746	0.19513	0.27155
	0.1316	<.0001	0.8668	0.0215	<.0001	0.5445		0.0078	0.2873	0.0517	0.0063
x13	-0.40128	0.22946	-0.27079	-0.12795	0.13422	-0.49495	0.26460	1.00000	-0.24499	-0.11457	-0.07287
	<.0001	0.0216	0.0064	0.2046	0.1831	<.0001	0.0078		0.0140	0.2564	0.4712
x14	0.08831	0.05190	0.79717	0.14041	0.01079	0.27308	0.10746	-0.24499	1.00000	0.19707	0.10939
	0.3823	0.6081	<.0001	0.1635	0.9151	0.0060	0.2873	0.0140		0.0494	0.2786
x16	0.10430	0.15615	0.08010	0.75687	0.18424	0.42441	0.19513	-0.11457	0.19707	1.00000	0.75100
	0.3017	0.1208	0.4282	<.0001	0.0665	<.0001	0.0517	0.2564	0.0494		<.0001
x18	0.02772	0.19164	0.02544	0.86509	0.27586	0.60185	0.27155	-0.07287	0.10939	0.75100	1.00000
	0.7843	0.0561	0.8016	<.0001	0.0055	<.0001	0.0063	0.4712	0.2786	<.0001	

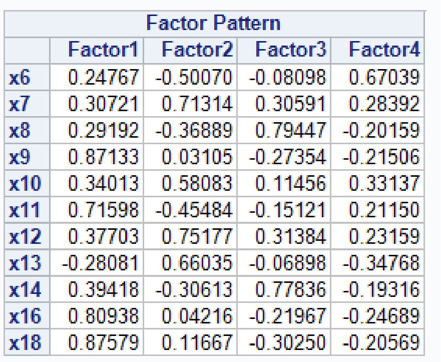
Table. 1. Correlation Matrix of HBAT data.



Eigenvalues of the Correlation Matrix: Total = 11 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	3.42697133	0.87607462	0.3115	0.3115
2	2.55089671	0.85992024	0.2319	0.5434
3	1.69097648	0.60442042	0.1537	0.6972
4	1.08655606	0.47713196	0.0988	0.7959
5	0.60942409	0.05754032	0.0554	0.8513
6	0.55188378	0.15036563	0.0502	0.9015
7	0.40151815	0.15456660	0.0365	0.9380
8	0.24695154	0.04339828	0.0225	0.9605
9	0.20355327	0.07071169	0.0185	0.9790
10	0.13284158	0.03441456	0.0121	0.9911
11	0.09842702		0.0089	1.0000

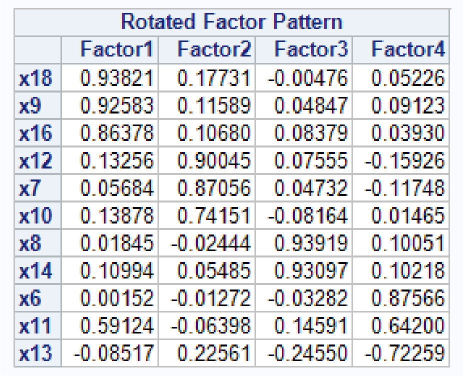
Table. 2. Eigenvalues of non-rotated data.



Factor Pattern

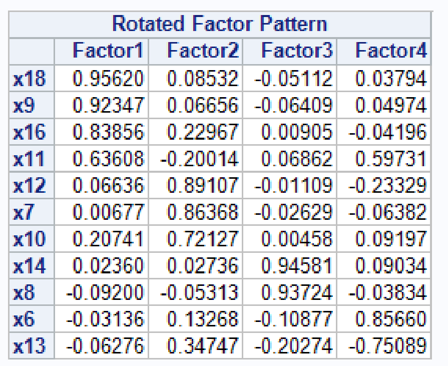
	Factor1	Factor2	Factor3	Factor4
x6	0.24767	-0.50070	-0.08098	0.67039
x7	0.30721	0.71314	0.30591	0.28392
x8	0.29192	-0.36889	0.79447	-0.20159
x9	0.87133	0.03105	-0.27354	-0.21506
x10	0.34013	0.58083	0.11456	0.33137
x11	0.71598	-0.45484	-0.15121	0.21150
x12	0.37703	0.75177	0.31384	0.23159
x13	-0.28081	0.66035	-0.06898	-0.34768
x14	0.39418	-0.30613	0.77836	-0.19316
x16	0.80938	0.04216	-0.21967	-0.24689
x18	0.87579	0.11667	-0.30250	-0.20569

Table. 3. Factor Pattern of non-rotated data.



Rotated Factor Pattern				
	Factor1	Factor2	Factor3	Factor4
x18	0.93821	0.17731	-0.00476	0.05226
x9	0.92583	0.11589	0.04847	0.09123
x16	0.86378	0.10680	0.08379	0.03930
x12	0.13256	0.90045	0.07555	-0.15926
x7	0.05684	0.87056	0.04732	-0.11748
x10	0.13878	0.74151	-0.08164	0.01465
x8	0.01845	-0.02444	0.93919	0.10051
x14	0.10994	0.05485	0.93097	0.10218
x6	0.00152	-0.01272	-0.03282	0.87566
x11	0.59124	-0.06398	0.14591	0.64200
x13	-0.08517	0.22561	-0.24550	-0.72259

Table. 4. Factor Pattern of Varimax rotated data.



Rotated Factor Pattern				
	Factor1	Factor2	Factor3	Factor4
x18	0.95620	0.08532	-0.05112	0.03794
x9	0.92347	0.06656	-0.06409	0.04974
x16	0.83856	0.22967	0.00905	-0.04196
x11	0.63608	-0.20014	0.06862	0.59731
x12	0.06636	0.89107	-0.01109	-0.23329
x7	0.00677	0.86368	-0.02629	-0.06382
x10	0.20741	0.72127	0.00458	0.09197
x14	0.02360	0.02736	0.94581	0.09034
x8	-0.09200	-0.05313	0.93724	-0.03834
x6	-0.03136	0.13268	-0.10877	0.85660
x13	-0.06276	0.34747	-0.20274	-0.75089

Table. 5. Factor pattern of random sample.