Assignment #8: Multivariate Analysis (0 points)

Data Directory: Data can be accessed on the SAS OnDemand server using this libname statement.

```
libname mydata        '/courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/' access=readonly;
```

Data Set: mydata.european_employment

Data Description: Employment in various industry segments reported as a percent for thirty European nations. See the data dictionary for full details. Note that EU stands for European Union, EFTA stands for European Free Trade Association, and Eastern stand for Eastern European nations or the former Eastern Block.

Assignment Instructions: Note that this assignment will not use our assignment template, nor will it follow the guidelines for report writing that we have used all quarter. Instead, you will be able to paste your output and type your answers directly into the Word version of this assignment, convert your solution document to a pdf, and submit your pdf document into Blackboard. **Please color code your answers in green.**

In this assignment we will take a guided tour of the multivariate analysis capabilities in SAS. These capabilities will include PROC PRINCOMP, PROC FACTOR, and PROC CLUSTER. Since none of these methods are covered in our SAS books, our only reference will be the SAS User's Guide.

PROC FACTOR	Chapter 34	SAS 9.3 User's Guide
PROC PRINCOMP	Chapter 72	SAS 9.3 User's Guide
PROC CLUSTER	Chapter 30	SAS 9.3 User's Guide

http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#titlepage.htm

The assignment is broken down into parts for your convenience. Each part will instruct you to generate a particular set of SAS output and interpret this output. In addition a section may have some particular questions that you should address. These questions will be written in **bold black type**.

Note that the SAS code provided in this assignment will produce an extensive amount of output. You will probably want to run the code piece by piece and answer each Part of the assignment completely before moving to the next Part.

For convenience here are the definitions of the abbreviated industries.

AGR: agriculture
MIN: mining
MAN: manufacturing
PS: power and water supply
CON: construction
SER: services
FIN: finance
SPS: social and personal services
TC: transport and communications

Part 1: An Initial Correlation Analysis

We will conclude this tutorial by applying cluster analysis to this data. When we perform a cluster analysis, we will always want to perform the cluster analysis in a low dimensional setting. Only in low dimensions can points be "close together". As we move towards this cluster analysis we want to perform some basic examinations of the data and consider using factor analysis and principal components as means to reduce the dimensionality of our data.

Of course, before we conclude this tutorial we must begin this tutorial. We will begin this tutorial by examining the two dimensional scatterplots of the variables. Use PROC CORR to produce the Pearson correlation coefficients and the scatterplot matrix. Looking at the scatterplots, is there any scatterplot that looks like it would yield interesting cluster results? For the two variables of your choice make this scatterplot (replace Yvar and Xvar with your two variables).

```
data temp;
set mydata.european_employment;
run;

ods graphics on;
proc sgplot data=temp;
title 'Scatterplot of Raw Data';
scatter y=Yvar x=Xvar / datalabel=country group=group;
run; quit;
ods graphics off;
```

In this data set there are four counties that do not belong to any of the three primary groups. If you had to assign each of these countries to a group to which group would you assign each country.

Note: In this assignment our observations are assigned to *classes* or are said to have *labels* (EU, EFTA, Eastern, or Other). Typically we use cluster analysis as an *unsupervised learner* (a situation with no response variable or label) and not as a *supervised learner* (a situation with a response variable or label). If we wanted to be able to correctly assign each country to its group affiliation, then we would define a *classification problem* (see Chapter 11 in *Applied Multivariate Data Analysis*). Throughout this assignment we will be interested in grouping countries together (creating a *segmentation*), but we can also observe their group affiliation to see if these groups have similarities.

Part 2: Principal Components Analysis

Our data set has nine variables. One method of reducing the dimensionality of our data set is to use principal components analysis. If we perform a principal components analysis, what would the resulting dimensionality be, i.e. how many components should we keep? What decision rule are you using to determine how many of the principal components to keep? Are there any other competing decision rules that you could use? Include the table of the eigenvalues of the correlation matrix, the scree plot, and the "Component Pattern Profiles" plot. Interpret these plots and make the appropriate comments. See Chapter 3 of *Applied Multivariate Data Analysis* for a statistical reference to principal components analysis.

```
ods graphics on;
title Principal Components Analysis using PROC PRINCOMP;
proc princomp data=temp out=pca_9components outstat=eigenvectors plots=all;
run;
ods graphics off;
```

Part 3: Factor Analysis

A second approach to reducing the dimensionality of our data set is to use factor analysis. Before we begin applying a factor analysis, you will need to answer a question? Provide your answer in green.

Are principal components analysis and factor analysis the same statistical method? How are they different?

The SAS procedure for performing a variety of implementations of factor analysis is PROC FACTOR. Let's perform a factor analysis on our data using different methods of factor analysis. See Chapter 12 of *Applied Multivariate Data Analysis* for a statistical reference to factor analysis (Exploratory Factor Analysis).

Principal Components Using PROC FACTOR:

In addition to using PROC PRINCOMP to perform a principal components analysis SAS will allow you to perform a principal components analysis using PROC FACTOR. Run this code and compare the output from PROC FACTOR to the output from PROC PRINCOMP.

```
ods graphics on;
title Principal Components Analysis using PROC FACTOR;
proc factor data=temp method=principal out=pca_factors
      nfactors=9 score plots=scree;
run;
ods graphics off;
```

Iterated Principal Factor Analysis:

Now let's perform a legitimate factor analysis using PROC FACTOR. We will run an Iterated Principal Factor Analysis using the following SAS code.

```
ods graphics on;
title Iterated Principal Factor Analysis using PROC FACTOR;
proc factor data=temp method=prinit out=pfa_factors
      nfactors=9 score plots=scree;
run;
ods graphics off;
```

Is this a valid factor analysis? (Hint: the answer is no.) Why is this not a valid factor analysis? Keep reducing the number for *nfactors* until you get a valid factor analysis. Report the "Eigenvalues of the Reduced Correlation Matrix", the "Factor Pattern", the "Variance Explained by Each Factor", and the "Final Communality Estimates" tables and make the appropriate comments on the results in these tables. As part of your comments do you have an interpretation of the factor loadings.

```
ods graphics on;
title Iterated Principal Factor Analysis using PROC FACTOR;
proc factor data=temp method=prinit out=pfa_factors
      nfactors=2 score plots=scree;
run;
ods graphics off;
```

Maximum Likelihood Factor Analysis:

An alternative to iterated principal factor analysis is maximum likelihood factor analysis.

```
ods graphics on;
title Maximum Likelihood Factor Analysis using PROC FACTOR;
proc factor data=temp method=ml out=fa_ml
      outstat=fa_ml_stats mineigen=0 priors=smc nfactors=2 score ;
run;
ods graphics off;
```

Is this a valid factor analysis? If this is not a valid factor analysis, then why is this not a valid factor analysis? If this is a valid factor analysis then report the "Eigenvalues of the Reduced Correlation Matrix", the "Factor Pattern", the "Variance Explained by Each Factor", and the "Final Communality Estimates" tables and make the appropriate comments on the results in these tables.

Unweighted Least Squares Factor Analysis:

Another type of factor analysis, which is an alternative to both iterated principal factor analysis and maximum likelihood factor analysis, is unweighted least squares factor analysis.

```
ods graphics on;
title Unweighted Least Squares Factor Analysis using PROC FACTOR;
proc factor data=temp method=uls out=fa_uls
      outstat=uls_stats mineigen=0 priors=smc nfactors=2 score ;
run;
ods graphics off;
```

Is this a valid factor analysis? If this is not a valid factor analysis, then why is this not a valid factor analysis? If this is a valid factor analysis then report the "Eigenvalues of the Reduced Correlation Matrix", the "Factor Pattern", the "Variance Explained by Each Factor", and the "Final Communality Estimates" tables and make the appropriate comments on the results in these tables. Are the estimated factor loadings from the unweighted least squares factor analysis significantly different from the factor loadings from iterated principal factor analysis?

Part 4: Factor Rotations

We will now consider rotating a set of factors. Before we begin you will need to answer a question? Provide your answer in green.

What is the difference between an oblique and an orthogonal factor rotation? Is there any reason to choose an oblique rotation over an orthogonal rotation, or vice-versa?

VARIMAX Factor Rotation

First we will perform an orthogonal factor rotation using a VARIMAX rotation.

```
ods graphics on;
title A VARIMAX Rotation of a Unweighted Least Squares Factor Analysis using
PROC FACTOR;
proc factor data=temp method=uls rotate=varimax out=uls_varimax
      outstat=varimax_stats mineigen=0 priors=max nfactors=2 score
      plots=(initloadings preloadings loadings scree) ;
run;
ods graphics off;
```

Report the "Eigenvalues of the Reduced Correlation Matrix", the "Factor Pattern", the "Variance Explained by Each Factor", and the "Final Communality Estimates" tables and the same output for the rotated factors. Make the appropriate comments on the results in these tables. Did the factor rotation change the variance explained by each factor? Did the factor rotation change the interpretation of the factor loadings? Did the factor rotation change the final communality estimates?

PROMAX Factor Rotation

Now we will perform an oblique factor rotation using a PROMAX rotation.

```
ods graphics on;
title A PROMAX Rotation of a Unweighted Least Squares Factor Analysis using
PROC FACTOR;
proc factor data=temp method=uls rotate=promax out=uls_promax
      outstat=promax_stats mineigen=0 priors=max nfactors=2 score
      plots=(initloadings preloadings loadings scree) ;
run;
ods graphics off;
```

Report the "Eigenvalues of the Reduced Correlation Matrix", the "Factor Pattern", the "Variance Explained by Each Factor", and the "Final Communality Estimates" tables and the same output for the rotated factors. Make the appropriate comments on the results in these tables. Did the factor rotation change the variance explained by each factor? Did the factor rotation change the interpretation of the factor loadings? Did the factor rotation change the final communality estimates?

Part 5: Cluster Analysis

We will begin our discussion of cluster analysis by making a pair of scatterplots.

```
ods graphics on;
proc sgplot data=temp;
title 'Scatterplot of Raw Data: FIN*SER';
scatter y=fin x=ser / datalabel=country group=group;
run; quit;
ods graphics off;

ods graphics on;
proc sgplot data=temp;
title 'Scatterplot of Raw Data: MAN*SER';
scatter y=man x=ser / datalabel=country group=group;
run; quit;
ods graphics off;
```

How many clusters do you see in the scatterplot of FIN*SER? How many clusters do you see in the scatterplot of MAN*SER?

Clearly different projections of the data will produce different clustering results. We need to be cognizant of this fact.

Now we will use PROC CLUSTER to create a set of clusters algorithmically. Note that PROC CLUSTER performs *hierarchical clustering* (see Chapter 6 in *Applied Multivariate Data Analysis*) so we do not need to specify the number of clusters in advance. We will use the SAS procedure PROC TREE to assign observations to a specified number of clusters after we have performed the hierarchical clustering.

```
ods graphics on;
proc cluster data=temp method=average outtree=tree1 pseudo ccc plots=all;
var fin ser;
id country;
run; quit;
ods graphics off;
```

How do we interpret the measures of CCC, Pseudo F, and Pseudo T-Squared? How do we interpret the plots for these three measures?

We can use PROC TREE to assign our data to a set number of clusters. Let's compare the output when we assign the observations to four clusters and then to three clusters.

```
ods graphics on;
proc tree data=tree1 ncl=4 out=_4_clusters;
copy fin ser;
run; quit;
ods graphics off;

ods graphics on;
proc tree data=tree1 ncl=3 out=_3_clusters;
copy fin ser;
run; quit;
ods graphics off;
```

We will use this macro to make tables displaying the assignment of the observations to the determined clusters.

```
%macro makeTable(treeout,group,outdata);
data tree_data;
      set &treeout.(rename=(_name_=country));
run;

proc sort data=tree_data; by country; run; quit;

data group_affiliation;
      set &group.(keep=group country);
run;

proc sort data=group_affiliation; by country; run; quit;

data &outdata.;
      merge tree_data group_affiliation;
      by country;
run;

proc freq data=&outdata.;
table group*clusname / nopercent norow nocol;
run;
%mend makeTable;
```

```
* Call macro function;
%makeTable(treeout=_3_clusters,group=temp,outdata=_3_clusters_with_labels);

* Plot the clusters for a visual display;
ods graphics on;
proc sgplot data=_3_clusters_with_labels;
title 'Scatterplot of Raw Data';
scatter y=fin x=ser / datalabel=country group=clusname;
run; quit;
ods graphics off;


%makeTable(treeout=_4_clusters,group=temp,outdata=_4_clusters_with_labels);

* Plot the clusters for a visual display;
ods graphics on;
proc sgplot data=_4_clusters_with_labels;
title 'Scatterplot of Raw Data';
scatter y=fin x=ser / datalabel=country group=clusname;
run; quit;
ods graphics off;
```

Display the tables and comment on these results. Did the members of each membership group get clustered into the same cluster? Which number of clusters do you prefer?

Now perform a similar cluster analysis using the following cluster commands. Which of these four cluster analyses do you prefer?

```
*****************************************************************;
* Using the first 2 principal components;
*****************************************************************;
ods graphics on;
proc cluster data=pca_9components method=average outtree=tree3 pseudo ccc
plots=all;
var prin1 prin2;
id country;
run; quit;
ods graphics off;


ods graphics on;
proc tree data=tree3 ncl=4 out=_4_clusters;
copy prin1 prin2;
run; quit;

proc tree data=tree3 ncl=3 out=_3_clusters;
copy prin1 prin2;
run; quit;
ods graphics off;

%makeTable(treeout=_3_clusters,group=temp,outdata=_3_clusters_with_labels);
%makeTable(treeout=_4_clusters,group=temp,outdata=_4_clusters_with_labels);
```

```
* Plot the clusters for a visual display;
ods graphics on;
proc sgplot data=_3_clusters_with_labels;
title 'Scatterplot of Raw Data';
scatter y=prin2 x=prin1 / datalabel=country group=clusname;
run; quit;
ods graphics off;


* Plot the clusters for a visual display;
ods graphics on;
proc sgplot data=_4_clusters_with_labels;
title 'Scatterplot of Raw Data';
scatter y=prin2 x=prin1 / datalabel=country group=clusname;
run; quit;
ods graphics off;


*************************************************************************;
* Using the first 2 factor components from ULS with VARIMAX rotation;
*************************************************************************;
ods graphics on;
proc cluster data=uls_varimax method=average outtree=tree4 pseudo ccc
plots=all;
var factor1 factor2;
id country;
run; quit;
ods graphics off;

ods graphics on;
proc tree data=tree4 ncl=4 out=_4_clusters;
copy factor1 factor2;
run; quit;

proc tree data=tree4 ncl=3 out=_3_clusters;
copy factor1 factor2;
run; quit;
ods graphics off;

%makeTable(treeout=_3_clusters,group=temp,outdata=_3_clusters_with_labels);
%makeTable(treeout=_4_clusters,group=temp,outdata=_4_clusters_with_labels);

* Plot the clusters for a visual display;
ods graphics on;
proc sgplot data=_3_clusters_with_labels;
title 'Scatterplot of Raw Data';
scatter y=factor2 x=factor1 / datalabel=country group=clusname;
run; quit;
ods graphics off;

* Plot the clusters for a visual display;
ods graphics on;
proc sgplot data=_4_clusters_with_labels;
title 'Scatterplot of Raw Data';
scatter y=factor2 x=factor1 / datalabel=country group=clusname;
run; quit;
ods graphics off;
```

Assignment Document:

As mentioned in the beginning we will not be using our typical assignment format. You will be given a Word document of the assignment, and you will write your answers directly into the document near the questions in green. As always the document should be submitted in pdf format.