MUTIVALIATE ANALYSIS

STAT 550

Principal Component Analysis and Factor Analysis

Weixiong Junyan Deng

April 11, 2016

Professor: Kim Sung

California State University of Long Beach

1 Abstract

Data reduction is the transformation of original data into a corrected, ordered and simplified form. This paper describes two methods of data reduction, principal components and Factor Analysis. Both methods attempt to approximate the covariance matrix or correlation matrix. We will use crime data recorded in the United States as an example to explain these two method in detail. Our goal is to reduce variables and interpret the reduced variables.

2 Data Description

The data consist of 50 observations recorded in 1985, and each observation contains 12 variables. Crimes are classified into 7 categories (X4-X10). The variables are recorded in different units described in Table 2.1.2. Larger stander deviation would have more weight on the covariance matrix. So we would use correlation matrix instead of covariance matrix to perform principal component analysis and factor analysis. This method is equivalent to standardized covariance matrix.

Table 2.1.1

X1: State
X2: land area (land)
X3: population 1985 (popu)
X4: murder (murd)
X5: rape
X6: robbery (robb)
X7: assault (assa)
X8: burglary (burg)
X9: larceny (larc)
X10: auto theft (auto)
X11: US State region number (reg)
X12: US State division number (div)
X1, X11 and X12 are categorical variables. The others are numeric variables.

Table 2.1.2

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
x2	50	72374	88408	3618701	1212	591004	land
x3	50	4762	5069	238113	509.00000	26365	popu
x4	50	6.85800	3.84798	342.90000	0.50000	15.30000	murd
x5	50	15.61600	7.34820	780.80000	3.60000	36.00000	rape
x6	50	101.51000	91.19338	5076	6.50000	443.30000	robb
x7	50	135.42000	68.16968	6771	21.00000	293.00000	assa
x8	50	930.80000	361.04977	46540	286.00000	1753	burg
x9	50	1944	709.82929	97182	694.00000	3550	larc
x10	50	367.86000	199.60952	18393	78.00000	878.00000	auto

3 Method

We use Principle Component Analysis and Factor Analysis to decompose the crime data.

Principal Component Analysis is to explain the variance-covariance with p original variables X through a few linear combinations of these variables to form a new set of variables Y, where X and Y are matrixes with p dimension. We choose k components (variables) from Y where k variables consist (almost) as much information as there is in the original p variables (X).

Factor Analysis is to describe variability among observed, correlated variables in terms of a few unobservable random variables called factors. The factors are viewed to describe an observed phenomenon. Factor analysis is an exploratory method which needs subjective determinant by the analyst.

Factor Analysis can be viewed as an extension of Principal Component Analysis. Both of them are to approximate the covariance matrix. In both PCA and FA, their dimension of the data are reduced.

4 Result

4.1 Pearson Correlation (preliminary analysis)

From Table 4.1, Pearson Correlation, we can find out that x2(land area) is not related to other variables (p-value greater than 0.05), only related to x5(rape). All the variables chosen for Principal analysis and factor analysis must be correlated to each other. Moreover, the only one correlation coefficient between x2(land area) and x5(rape) is 0.37, coefficients between x2 and others are all relatively small. Principal component analysis or factor analysis will not work well to reduce data if the coefficient is small. So we will exclude x2(land area) to perform PCA and FA.

X3(population) is not that much related to x4(murder) and x9(larceny). It seems that murder and larceny are two extreme sides in the measurement of violence. While murder is extreme severe violence, larceny is extreme soft.

X6(robbery), x7(assault), x8(burglary) are highly correlated to each other with similar coefficient (from 0.5 to 0.7). This seems that they form in one cluster.

X3(population), x5(rape) and x10(auto theft) seem to be group together because their coefficients of correlation are similar (from 0.35 to 0.45). This seems that rape and auto theft are more likely to happen in a state which has more population.

Table 4.1

Pearson Correlation Coefficients, N = 50
Prob > |r| under H0: Rho=0

	x2	x3	x4	x5	x6	x7	x8	x9	x10
x2 land	1.00000	0.07188 0.6198	0.24450 0.0870	0.37683 0.0070	-0.02054 0.8874	0.16203 0.2609	0.06765 0.6406	0.25319 0.0760	0.08236 0.5696
x3 popu	0.07188 0.6198	1.00000	0.27216 0.0559	0.41805 0.0025	0.62324 <.0001	0.42635 0.0020	0.42856 0.0019	0.23054 0.1072	0.37589 0.0071
x4 murd	0.24450 0.0870	0.27216 0.0559	1.00000	0.51987 0.0001	0.34106 0.0154	0.81256 <.0001	0.27672 0.0517	0.06478 0.6549	0.10983 0.4477
x5 rape	0.37683 0.0070	0.41805 0.0025	0.51987 0.0001	1.00000	0.55144 <.0001	0.69593 <.0001	0.68015 <.0001	0.60061 <.0001	0.44070 0.0014
x6 robb	-0.02054 0.8874	0.62324 <.0001	0.34106 0.0154	0.55144 <.0001	1.00000	0.56320 <.0001	0.62219 <.0001	0.43618 0.0015	0.61705 <.0001
x7 assa	0.16203 0.2609	0.42635 0.0020	0.81256 <.0001	0.69593 <.0001	0.56320 <.0001	1.00000	0.52072 0.0001	0.31670 0.0250	0.33038 0.0191
x8 burg	0.06765 0.6406	0.42856 0.0019	0.27672 0.0517	0.68015 <.0001	0.62219 <.0001	0.52072 0.0001	1.00000	0.80110 <.0001	0.70010 <.0001
x9 larc	0.25319 0.0760	0.23054 0.1072	0.06478 0.6549	0.60061 <.0001	0.43618 0.0015	0.31670 0.0250	0.80110 <.0001	1.00000	0.55478 <.0001
x10 auto	0.08236 0.5696	0.37589 0.0071	0.10983 0.4477	0.44070 0.0014	0.61705 <.0001	0.33038 0.0191	0.70010 <.0001	0.55478 <.0001	1.00000

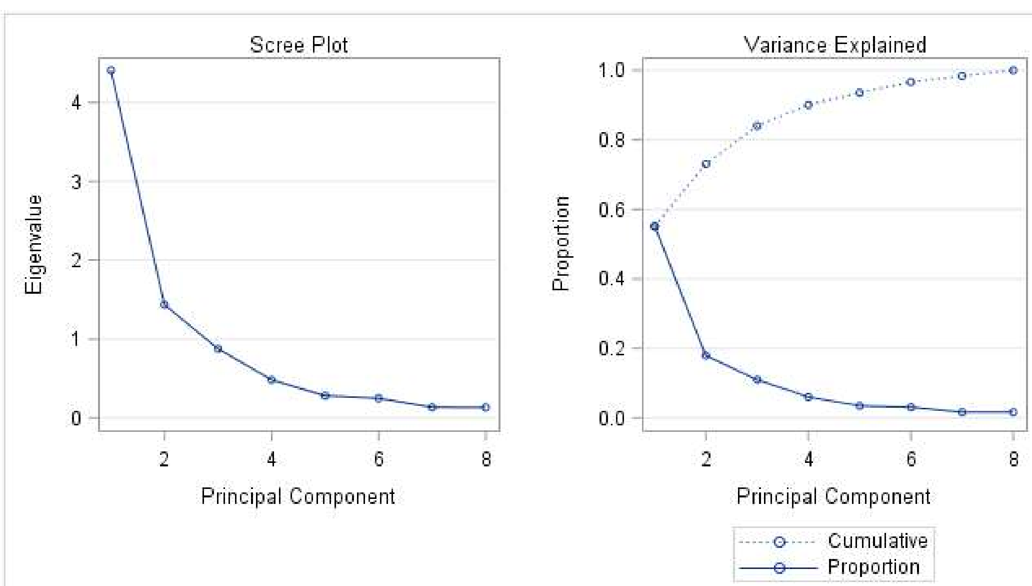
4.2.1 Principal Component Analysis

Perform a complete principal component analysis. From Table 4.2.1 and Figure 4.2.1 we can find out that the cutting point is 3, where the remaining eigenvalue is relative small. The first 3 principal components explained 84.02% of the correlation which is almost as much as there is in the original variables. This means the first 3 principal components are adequate. X3 - X10 can be replaced by Y1 – Y3 without significant loss of information.

Table4.2.1

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.40900906	2.97421466	0.5511	0.5511
2	1.43479441	0.55724268	0.1793	0.7305
3	0.87755173	0.39557198	0.1097	0.8402
4	0.48197976	0.20086410	0.0602	0.9004
5	0.28111566	0.03286929	0.0351	0.9356
6	0.24824637	0.11304429	0.0310	0.9666
7	0.13520207	0.00310114	0.0169	0.9835
8	0.13210094		0.0165	1.0000

Figure4.2.1



4.2.1 Interpretation of PCA

First Principal Component Analysis---PCA1

The PCA1 describes the corelations of all crimes execpt x4(mudering) with lest correlation. Every crime has similar weight on PC1 except x4. The PCA1 score increase with increasing scroes of x3(population), x5(rape), x6(robbery), x7(assualt), x8(burglary), x9(larceny), and x10(auto theft). This suggest that these six variables change together. This component can be viewed as a measurement of overall cimre correlations.

Second Principal Component Analysis---PCA2

The second component increases with increasing of x4(murdering) and x7(assault) and with decreasing of x9(larceny). Here we use cut off scroes 0.4 and -0.4. This component can be viewed as a measurement of violance.The more of x4 (murdering) and x7(assault), the less of x9(larceny). Larceny maybe just only need money and is not that eager to perform in a violent way. But Murder and assalt are more violent and brutal.

Third Principal Component Analysis----PCA3

The third component increases with increasing of x3(population) and x6(robbery) with decreasing x9(larceny). Here we use a cut off criteria 0.4. and -0.4. The coefficient between prin3 and x9 is very closed to -0.4, so it can be treated to contribute to prin3.

This component can be viewed as a measurement of degree of criminal eager for money. The more people live in a state, there will be more robbery and less larceny.

Table 4.2.2

		Prin1	Prin2	Prin3
x3	popu	0.298672	0.063303	0.708735
x4	murd	0.262779	0.632799	-.176338
x5	rape	0.399587	0.101583	-.278409
x6	robb	0.385958	-.040462	0.404520
x7	assa	0.371976	0.444816	-.126847
x8	burg	0.413155	-.269873	-.180174
x9	larc	0.331591	-.418809	-.399904
x10	auto	0.337778	-.370821	0.130287

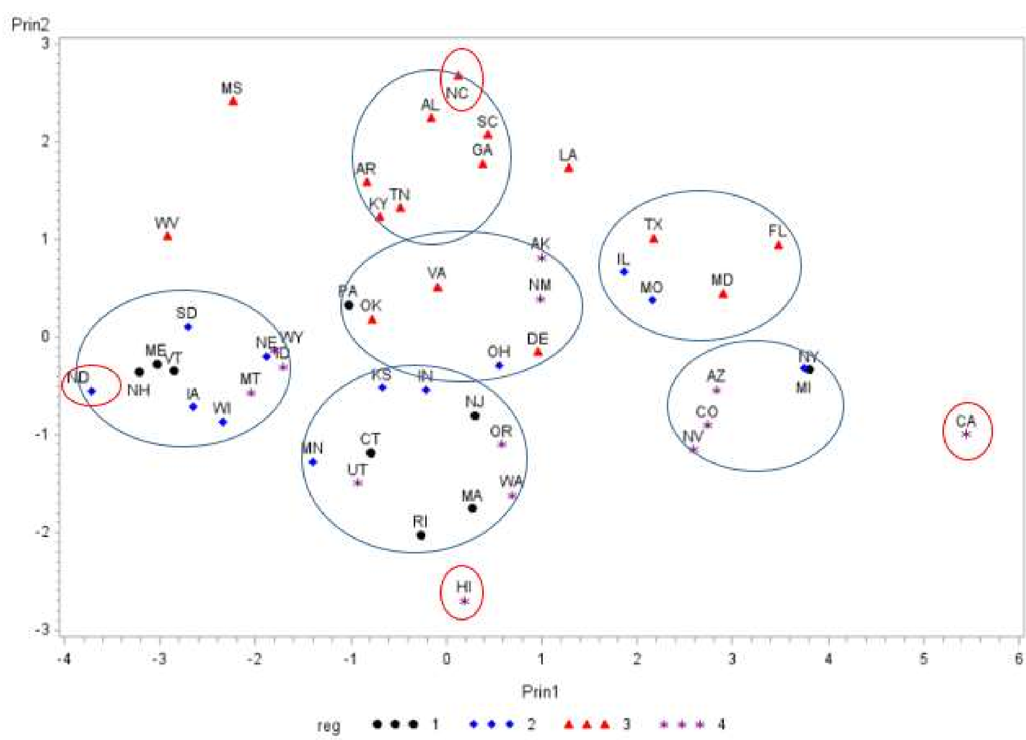
4.2.2 Scatter plot and groups

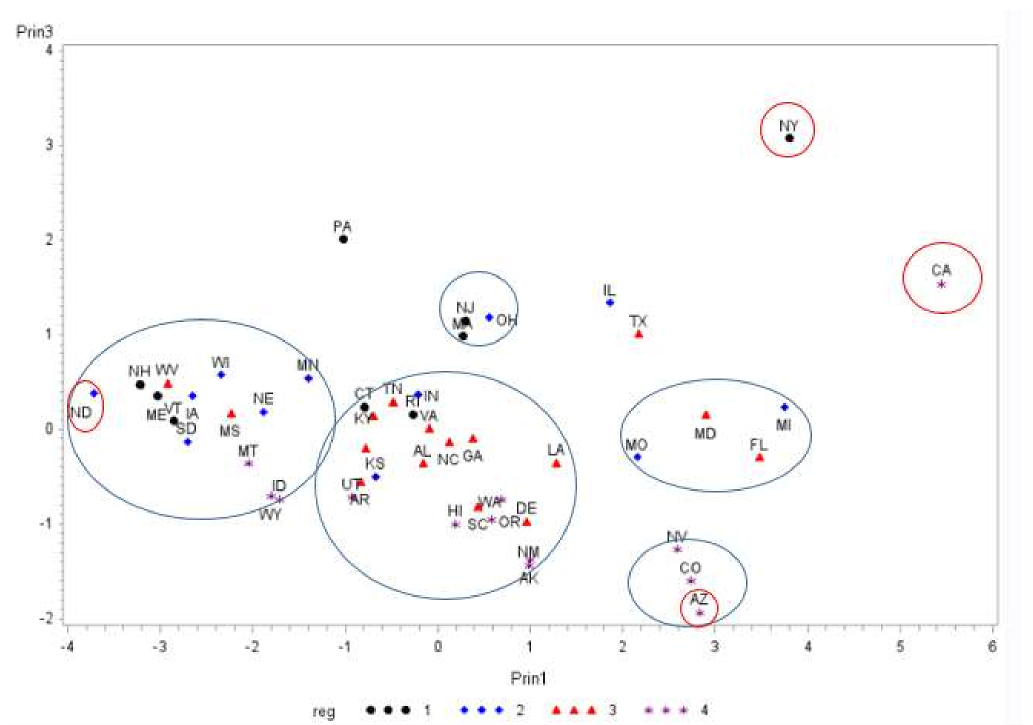
From Figure 4.2.2, we can find out that they can grouped in blue circles and that extrem values in red circles. On prin1 CA is the largest and ND is the smallest. On prin2 NC is the largest and HI is the smallest. On prin3 NY is the largest and AZ is the smallest. CA may be a potential outlier in prin1, because its value is too far away in prin1, lareger than 5. Region 3 is more likely to be in one group. Other regions are mix in groups.

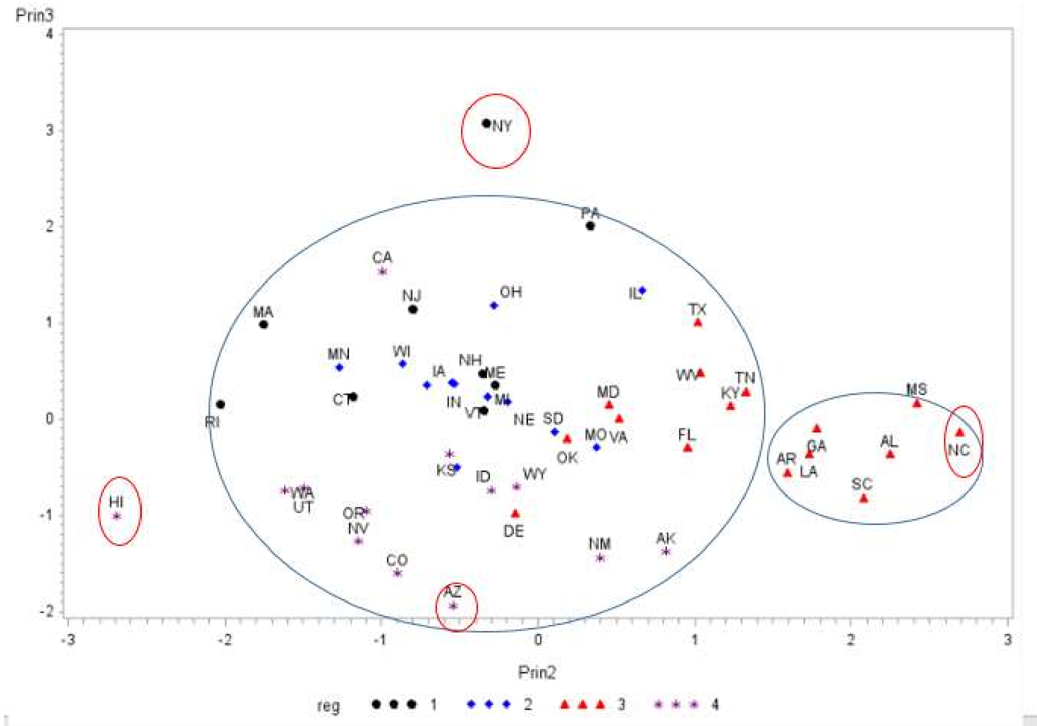
4.2.3 Check normal assumption

From Figure 4.2.3, qqplots show all principal component are approximately in a straight line. So we can conlude that the data are normal in the 3 principal components.

Figure 4.2.2

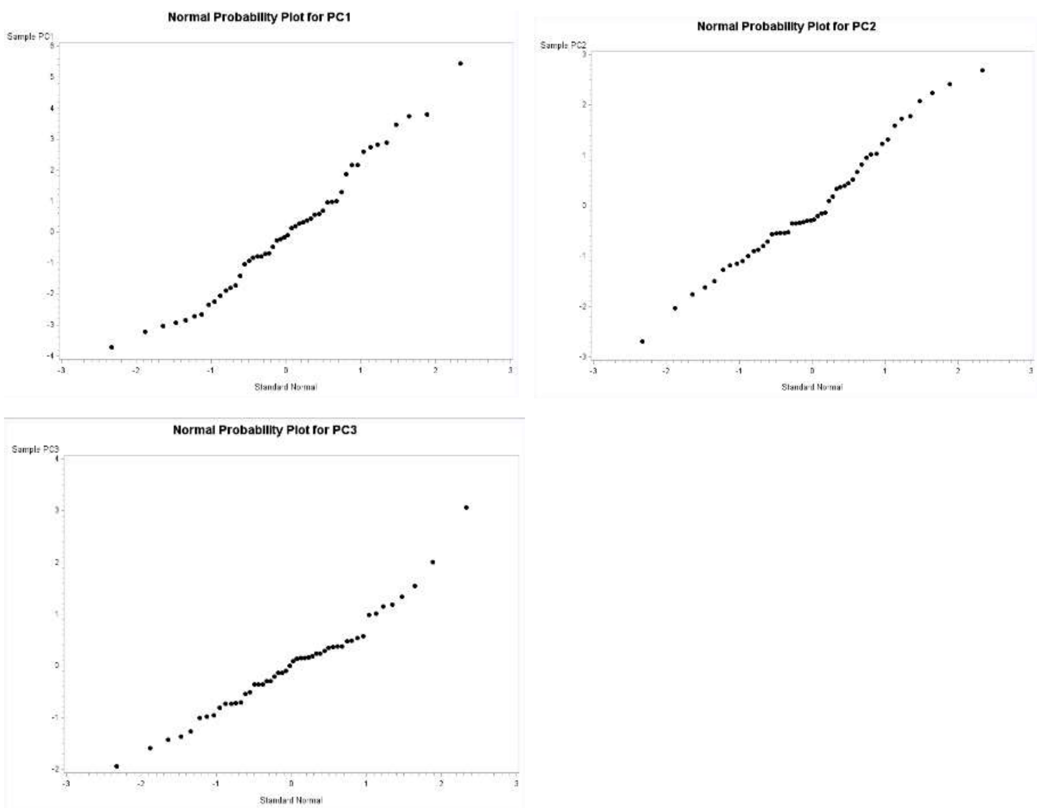






Prin3
Prin2
reg 1 2 3 4
NY
PA
CA
NJ
OH
IL
TX
MA
MN
WI
NH
ME
IA
WV
TN
MD
KY
MS
CT
IN
VT
MI
RI
NE
SD
MO
VA
FL
AL
GA
NC
OK
AR
LA
KS
SC
ID
WY
HI
WA
UT
OR
NV
DE
NM
AK
CO
AZ

Figure 4.2.3



Normal Probability Plot for PC1
Sample PC1
Standard Normal
Normal Probability Plot for PC2
Sample PC2
Standard Normal
Normal Probability Plot for PC3
Sample PC3
Standard Normal

4.3 Factor Analysis

4.3.1 Number of factor

Run a complet factor analysis with principal component method, the SAS software suggust the first two factor should be remained based on the criteria that eigenvalue exceeds 1. However, when we look at the table 4.3.1, we find out that the third eiganvalue is closed to 1 and the third factor explains 10.97% of the correlation. So we should take factor 3 into consideration in this case. The first 3 factors explain 84.02% of the the correlation in total. The fourth factor only explain 6.02% of the correlation and the remaining factors contribute neglegiblely small. So we use the cutoff point at factor 3.

Figure 4.3.1 cree plot

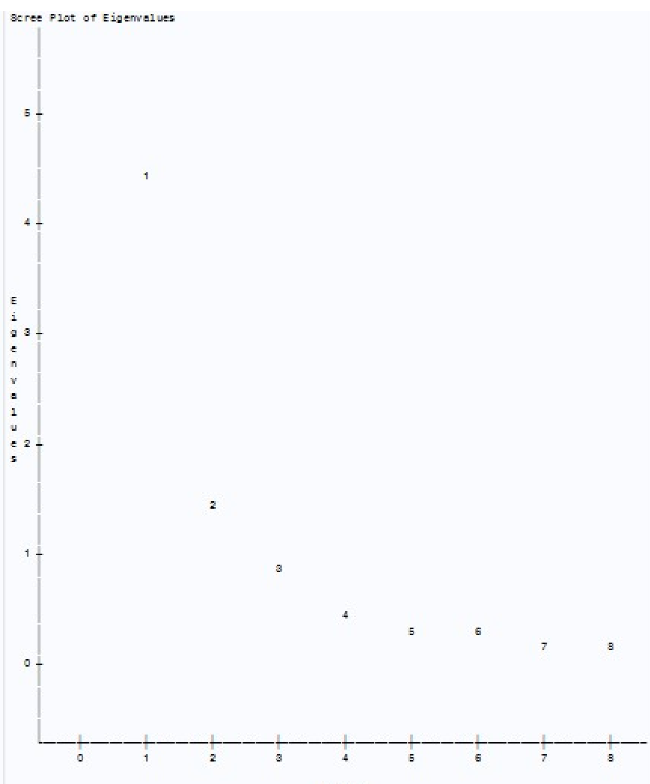


Table 4.3.1

Eigenvalues of the Correlation Matrix: Total = 8 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.40900906	2.97421466	0.5511	0.5511
2	1.43479441	0.55724268	0.1793	0.7305
3	0.87755173	0.39557198	0.1097	0.8402
4	0.48197976	0.20086410	0.0602	0.9004
5	0.28111566	0.03286929	0.0351	0.9356
6	0.24824637	0.11304429	0.0310	0.9666
7	0.13520207	0.00310114	0.0169	0.9835
8	0.13210094		0.0165	1.0000

4.3.2 Check number of factor adequacy

Run factor analysis with maximum likelyhood method. Here we do not discuss factor analysis with manximum likelyhood in detail, because there are too many solutions and some of their propotion explain the correlation or covariance are negetive, which is not reliable. But Maximum likelyhood method offer two chisquare tests are useful to check the number of factor adequacy.

The first test is for H0: No common factors. The Null hypothesis assumes that there are no common factors can explain the correlation of the variables. The p-value is less than 0.0001, which shows that we reject the null hypothesis and conclude that at least one commom factor can explain the correlation of the variables.

The second test is for H0: 3 factors are sufficient. The alterative hypothesis is that the model should need more than 3 factors to explain the correlation of the variables. The p-value of this Chisqure test is 0.7272. So we fail to reject the null hypothesis statistically and conclude that 3 factors are enough to explain all the variables.

Table 4.3.2 significant test

Significance Tests Based on 50 Observations			
Test	DF	Chi-Square	Pr > ChiSq
H0: No common factors	28	259.5000	<.0001
HA: At least one common factor			
H0: 3 Factors are sufficient	7	4.4457	0.7272
HA: More factors are needed			

Chi-Square without Bartlett's Correction	5.007818
Akaike's Information Criterion	-8.992182
Schwarz's Bayesian Criterion	-22.376343
Tucker and Lewis's Reliability Coefficient	1.044134

4.3.3 Interpretaion of factor loadings

4.3.3.1 Unrotated factor analysis

Three factors groups the variables in a understandable way. See the table 4.3.3.1. All crimes are responded to x3(population). The number of all crimes are rated to x3(population) in its state. If the population in a state is large, then burglary, rape, robbery, assault, auto theft, larceny, and murder will happen more frequently. All the variables contribute prevailently to Factor 1. The highest coefficient is 0.8675 form x8 (burglary). We can see this group visally in Figure 4.3.3.1.

B(x4 murder), E(x7 assault) and G(x9 larceny) contribute to Factor 2. The number of murder and the one of assault are positive rated together, The more murder the more assault and the less larceny. The highest contribution to Factor 2 is x4 (murder). Its coefficient shows that x4(muder) is 0.7586 correlated to Factor 2.

Factor 3 explained A(x3 population), D(x6 robbery) and G(larceny). The number of X3(population) and the one of X6(robbery) are positive rated, and they are negtive related to x9(larceny). The structure coefficients for these variables shows that A(x3 population) is the highest, 0.6639, correlated to Factor 3.

As we can see that the highest correlated to a factor. Names for the factors are below: Factor 1, x8 (burglary); Factor 2, x4 (muder); Factor 3, x3 (population). Factor 1 explains 4.409 of the variance. Factor 2 explains 1.435 of the variance. Factor explains 0.878 of the variance.

Compare to the preliminary analysis, pearson correlation. We can find out that only x5, x6, x7, and x8, these crimes are group together. This groups is similar to the group related to Factor 1, but still has some difference. Groups by factor analysis can expose more deeper correlation information.

Table 4.3.3.1

Factor Pattern				
		Factor1	Factor2	Factor3
x8	burg	0.86753	-0.32326	-0.16878
x5	rape	0.83904	0.12168	-0.26081
x6	robb	0.81042	-0.04847	0.37894
x7	assa	0.78106	0.53281	-0.11883
x10	auto	0.70925	-0.44418	0.12205
x9	larc	0.69626	-0.50166	-0.37462
x4	murd	0.55177	0.75798	-0.16519
x3	popu	0.62714	0.07583	0.66393

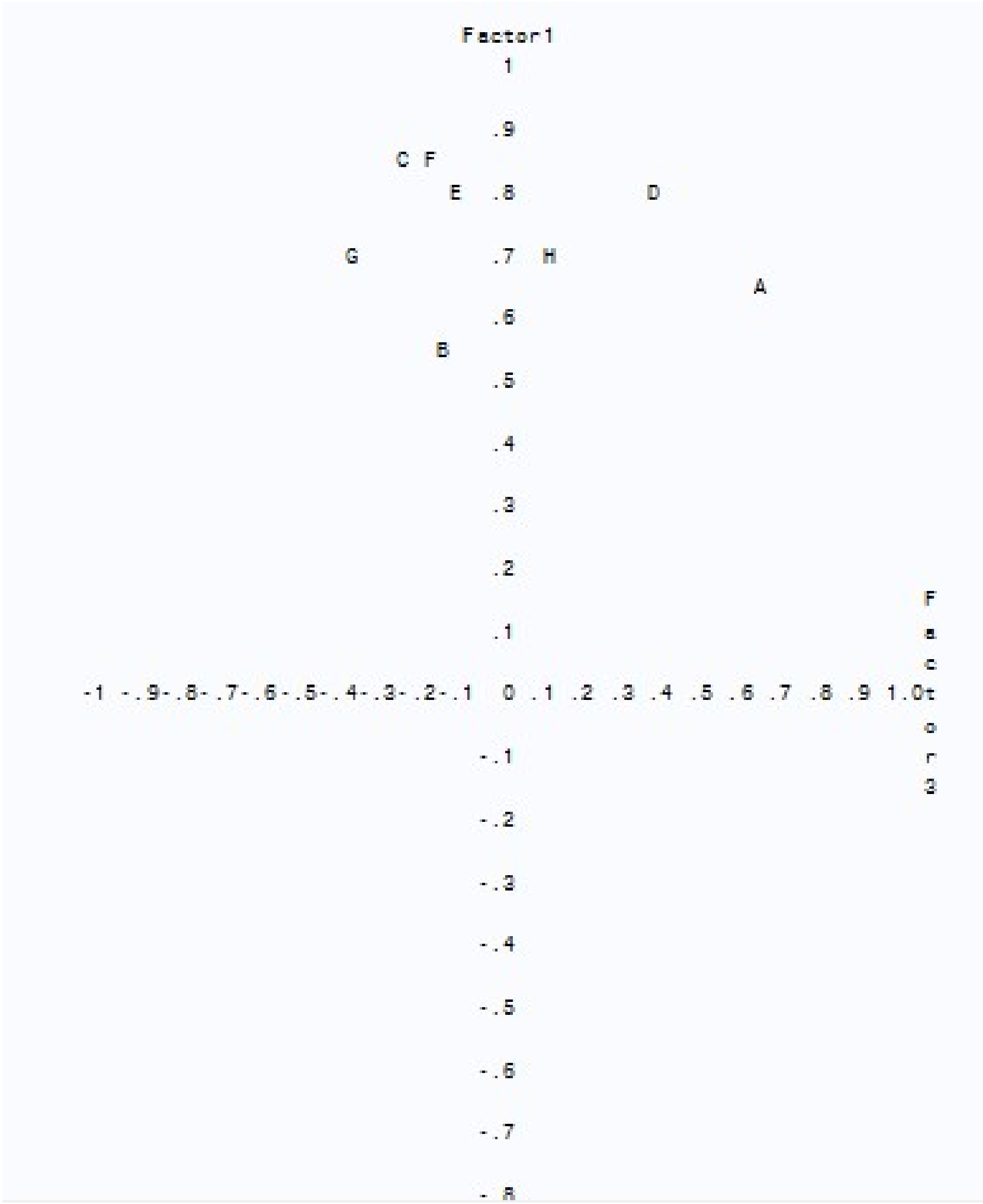
Table4.3.3.2

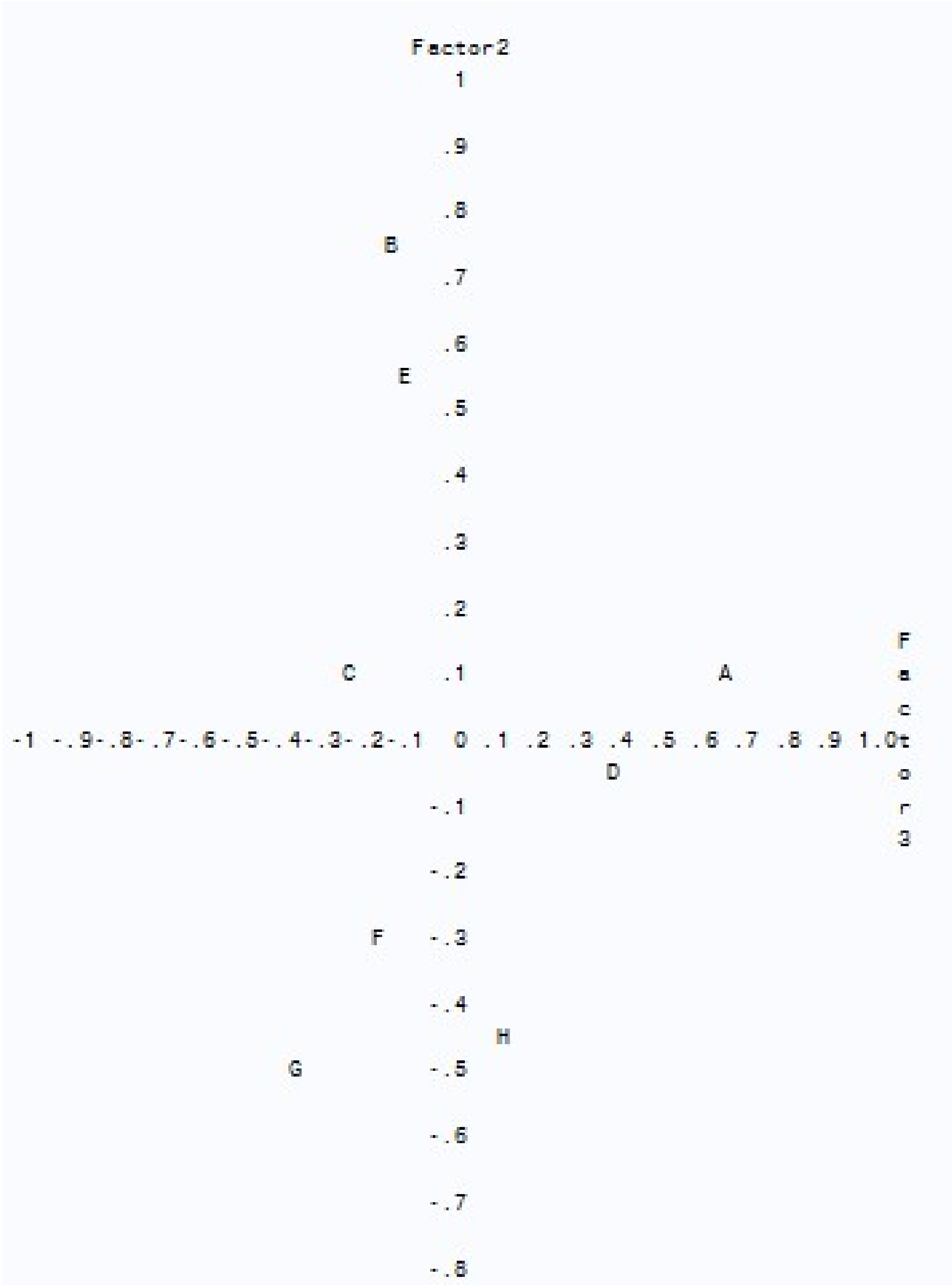
Variance Explained by Each Factor		
Factor1	Factor2	Factor3
4.4090091	1.4347944	0.8775517

Final Communality Estimates: Total = 6.721355							
x3	x4	x5	x6	x7	x8	x9	x10
0.83985450	0.90628196	0.78681218	0.80273116	0.90806602	0.88558917	0.87678744	0.71523276

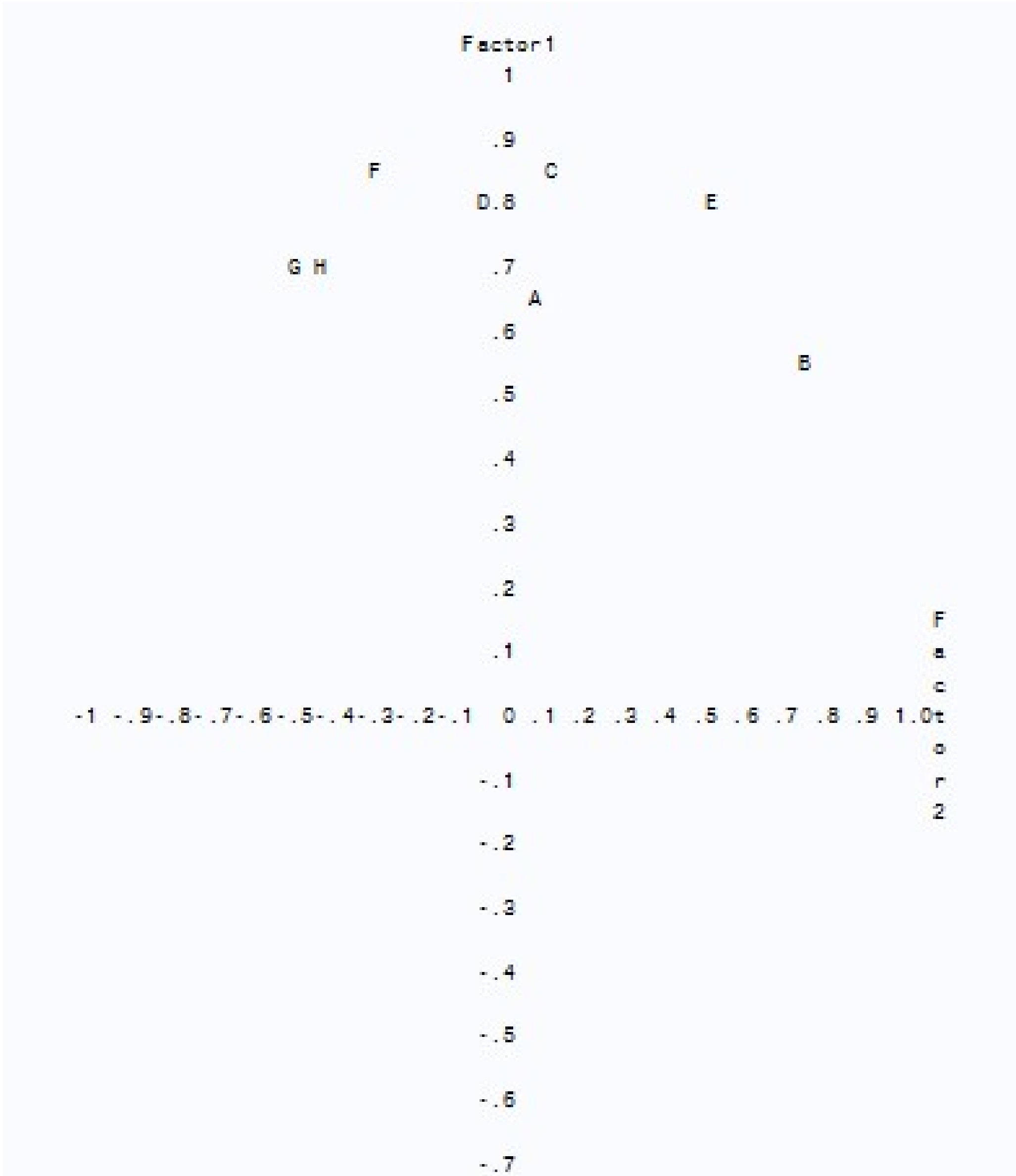
Figure 4.3.3.1

A=x3 B=x4 C=x5 D=X6 E=X7 F=X8 G=X9 H=X10





Factor2
1
.9
.8
B
.7
.6
E
.5
.4
.3
.2
C
.1
A
-1 -.9 -.8 -.7 -.6 -.5 -.4 -.3 -.2 -.1 0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0
Factor3
D
-.1
-.2
F
-.3
-.4
H
G
-.5
-.6
-.7
-.8


Factor1
1
.9
F
C
D.8
E
G H
.7
A
.6
B
.5
.4
.3
.2
.1
-1 -.9-.8-.7-.6-.5-.4-.3-.2-.1 0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0
-.1
-.2
-.3
-.4
-.5
-.6
-.7
Factor2

4.3.3.2 Rotated Factor Analysis

We apply varimax rotation and get table 4.3.3.2. We find out that x5 (rape), x8 (burglary), x9 (larceny) and x10 (auto theft) contribute most to Factor 1. X9 (larceny) is the highest one, 0.9316, correlated to Factor 1. The coefficient sugest that a state with high number of larceny will also has high number of burglary, auto theft and rape. Factor 1 explains the criminal motivation as money and sex, the basic craving of human's need.

X4(murder), X7 (assault) and X5 (rape) contribute most to Factor 2. X4 (murder) is correlated most, 0.9443, to Factor 2. The coefficient struture suggest that a state with high number of murder will also has high number of assault and rape. Factor 2 explains the degree of criminal violance in procession.

Factor 3 explains most in X3 (population) and robbery. X3 (population) is highest correlated, 0.8878, to Factor 3. The coefficient struture shows that a state with large population will have large number of robbery as well.

The names of the highest correlated to a factor are below: Factor 1, x9 (larceny); Factor 2, x4 (muder); Factor 3, x3 (population). Factor 1 explains 2.700 of the variance. Factor 2 explains 2.242 of the variance. Factor explains 1.778 of the variance. Compare to the preliminary analysis, pearson correlation. Its groups are quite different from the groups derived from Rotated Factor Analysis. The Groups by Rotated Factor Analysis are more reasonable.

Table 4.3.3.2.1

Rotated Factor Pattern				
		Factor1	Factor2	Factor3
x9	larc	0.93156	0.08608	0.03978
x8	burg	0.85365	0.25883	0.29980
x10	auto	0.70065	-0.00698	0.47357
x4	murd	-0.02168	0.94429	0.11885
x7	assa	0.25180	0.87544	0.27977
x5	rape	0.60057	0.62270	0.19589
x3	popu	0.10435	0.20183	0.88782
x6	robb	0.42160	0.28523	0.73731

Table 4.3.3.2.2

Variance Explained by Each Factor		
Factor1	Factor2	Factor3
2.7006090	2.2423789	1.7783674

Final Communality Estimates: Total = 6.721355							
x3	x4	x5	x6	x7	x8	x9	x10
0.83985450	0.90628196	0.78681218	0.80273116	0.90806602	0.88558917	0.87678744	0.71523276

4.3.3.3 Final Factor chosen

Compare to the factors between unrotated and rotated. We find out that rotated factors explain better.

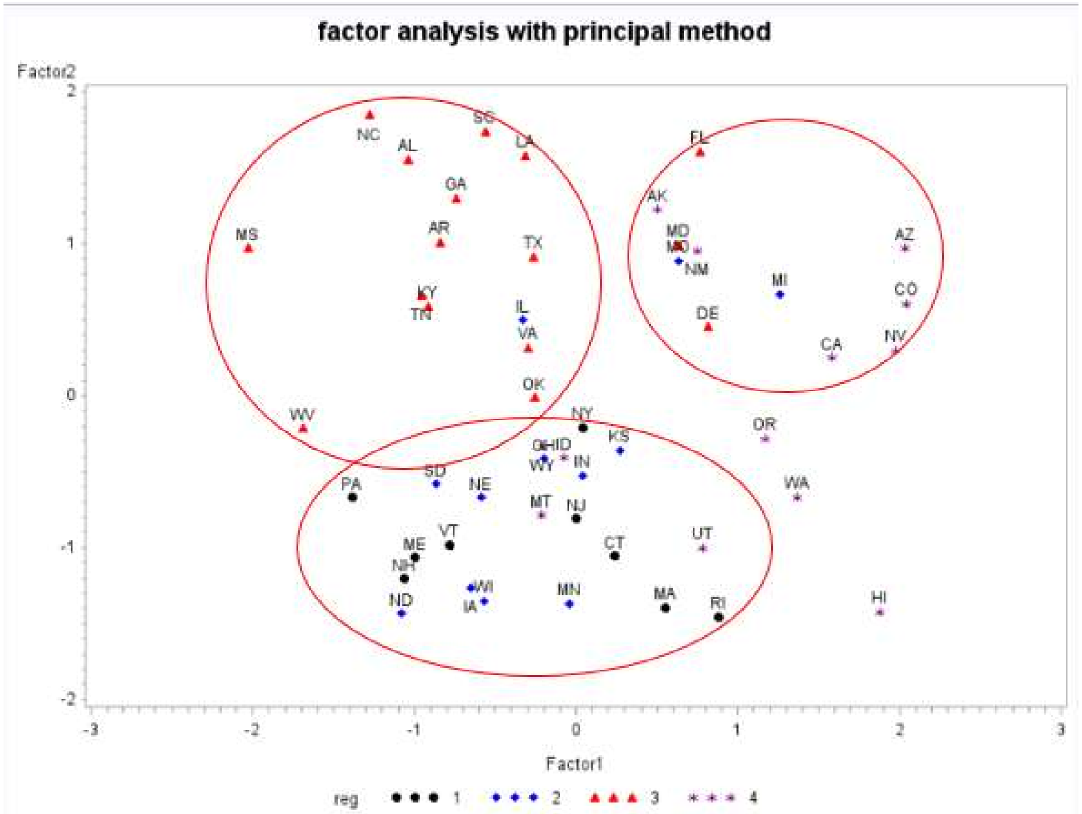
Factor analysis under correlation matrix, communality is equal to the summation of the variance explained by factors. The sprecific variance is equal to 1- commuality.

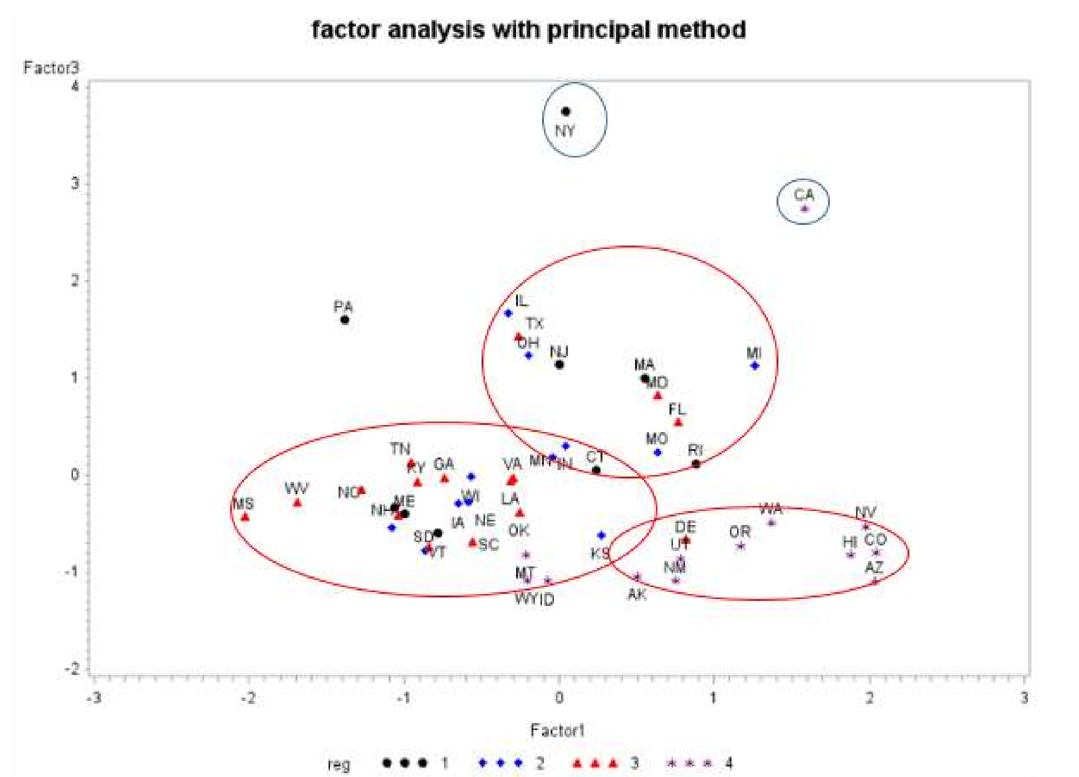
From Table 4.3.3.2.2, We can find out that the communality of x3 is 0.840, and its specific variance is 0.16, which is equal to 1-0.840. The communality of x4 is 0.906, and its specific variance is 0.094. The communality of x5 is 0.787, and its specific variance is 0.213. The communality of x6 is 0.803, and its specific variance is 0.197. The communality of x7 is 0.908, and its specific variance is 0.092. The communality of x8 is 0.886, and its specific variance is 0.114. The communality of x9 is 0.877, and its specific variance is 0.123. The communality of x10 is 0.715, and its specific variance is 0.285.

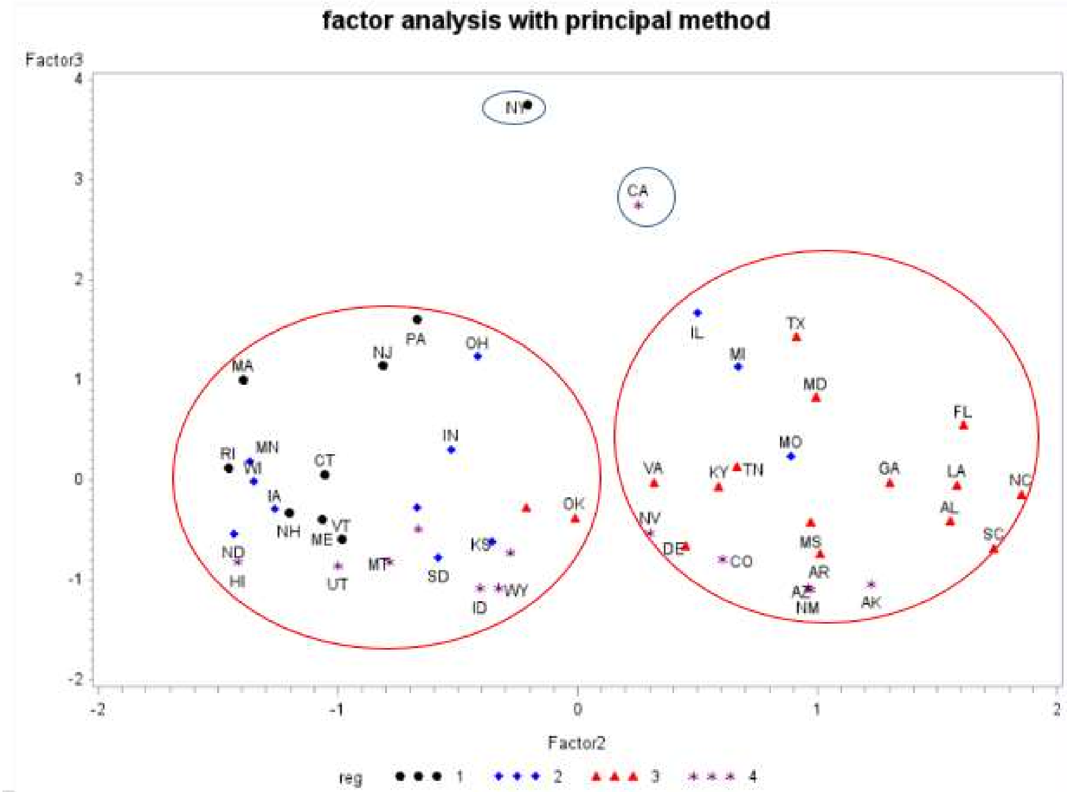
4.3.3.4 Factor Score

Factor scores are estimated values of the factors. Plot the Factor score in 50 states. We can group them in red circles. On Factor 3 axis, we find out that NY and CA is too far away from others, and NY's value is closed to 4 and CA's value is colsed to 3. They seem to be potential outliers and need further analysis.

Figure 4.3.3.4 Factor Score plot







5 Conclusion

The original data have 8 varialbes, x3 to x10. Using Principal Component Analysis or Factor Analysis we can reduce the large data set into the one with only 3 variavles without lossing to much variance. Compare to the pearson correlation matrix, Principal Component Analysis or Factor Analysis exposes the relationship between variables in more reasonable way and groups better. Principal Component Analysis generate a new set of variables. On the other hand, Factor Analysis explores the original variables with unobserved factors. Both methods try to approximate the covariance matrix or correlation matrix in a purpose to reduce variables.

Appendix:Code

```
/* US CRIME DATA
 This data consist of measurements of 50 states for 12 variables.
 It states for 1985 the reported number of crimes in the 50 states
 classified according to 7 categories (X4-X10)

X1: State
X2: land area (land)
X3: population 1985 (popu)
X4: murder (murd)
X5: rape
X6: robbery (robb)
X7: assault (assa)
X8: burglary (burg)
X9: larceny (larc)
X10: auto theft (auto)
X11: US State region number (reg)
X12: US State division number (div)
*/
DATA CRIME;
INPUT x1 $ x2-x12;
label x1="state" x2="land" x3="popu" x4="murd" x5="rape" x6="robb"
     x7="assa" x8="burg" x9="larc" x10="auto" x11="reg" x12="div";
DATALINES;
ME  33265  1164   1.500  7   12.600  62  562      1055   146  1   1
NH  9279   998 2   6   12.100  36  566    929    172  1   1
VT  9614   535 1.300   10.300  7.600   55  731    969    124  1   1
MA  8284   5822   3.500   12 99.500  88  1134   1531   878  1   1
RI  1212   968 3.200   3.600   78.300  120 1019   2186   859  1   1
CT  5018   3174   3.500   9.100   70.400  87  1084   1751   484  1   1
NY  49108  17783  7.900   15.500  443.300 209 1414   2025   682  1   2
NJ  7787   7562   5.700   12.900  169.400 90  1041   1689   557  1   2
PA  45308  11853  5.300   11.300  106   90  594    1001   340  1   2
OH  41330  10744  6.600   16 145.900 116 854    1944   493  2   3
IN  36185  5499   4.800   17.900  107.500 95  860    1791   429  2   3
IL  56345  11535  9.600   20.400  251.100 187 765    2028   518  2   3
MI  58527  9088   9.400   27.100  346.600 193 1571   2897   464  2   3
WI  56153  4775   2   6.700   33.100 44  539    1860   218  2   3
MN  84402  4193   2   9.700   89.100 51  802    1902   346  2   4
IA  56275  2884   1.900   6.200   28.600  48  507    1743   175  2   4
MO  69697  5029   10.700  27.400  200.800 167 1187   2074   538  2   4
ND  70703  685 0.500   6.200   6.500   21  286    1295   91   2   4
SD  77116  708 3.800   11.100  17.100  60  471    1396   94   2   4
```

```
NE  77355   1606    3   9.300    57.300  115  505    1572    292  2    4
KS  82277   2450    4.800    14.500   75.100   108  882    2302    257  2    4
DE  2044    622 7.700    18.600   105.500  196  1056   2320    559  3    5
MD  10460   4392    9.200    23.900   338.600  253  1051   2417    548  3    5
VA  40767   5706    8.400    15.400   92 143  806    1980    297  3    5
WV  24231   1936    6.200    6.700    27.300   84   389    774    92   3    5
NC  52669   6255    11.800   12.900   53 293  766    1338    169  3    5
SC  31113   3347    14.600   18.100   60.100   193  1025   1509    256  3    5
GA  58910   5976    15.300   10.100   95.800   177  900    1869    309  3    5
FL  58664   11366   12.700   22.200   186.100  277  1562   2861    397  3    5
KY  40409   3726    11.100   13.700   72.800   123  704    1212    346  3    6
TN  42144   4762    8.800    15.500   82 169  807    1025    289  3    6
AL  51705   4021    11.700   18.500   50.300   215  763    1125    223  3    6
MS  47689   2613    11.500   8.900    19 140  351    694    78   3    6
AR  53187   2359    10.100   17.100   45.600   150  885    1211    109  3    7
LA  47751   4481    11.700   23.100   140.800  238  890    1628    385  3    7
OK  69956   3301    5.900    15.600   54.900   127  841    1661    280  3    7
TX  266807  16370   11.600   21  134.100  195  1151   2183    394  3    7
MT  147046  826 3.200    10.500   22.300   75   594    1956    222  4    8
ID  83564   1005    4.600    12.300  20.500   86   674    2214    144  4    8
WY  97809   509 5.700    12.300   22 73   646    2049    165  4    8
CO  104091  3231    6.200    36  129.100  185  1381   2992    588  4    8
NM  121593  1450    9.400    21.700   66.100   196  1142   2408    392  4    8
AZ  114000  3187    9.500    27  120.200  214  1493   3550    501  4    8
UT  84899   1645    3.400    10.900   53.100   70   915    2833    316  4    8
NV  110561  936 8.800    19.600   188.400  182  1661   3044    661  4    8
WA  68138   4409    3.500    18  93.500   106  1441   2853    362  4    9
OR  97073   2687    4.600    18  102.500  132  1273   2825    333  4    9
CA  158706  26365   6.900    35.100   206.900  226  1753   3422    689  4    9
AK  591004  521 12.200   26.100   71.800   168  790    2183    551  4    9
HI  6471    1054    3.600    11.800   63.300   43   1456   3106    581  4    9
;
proc corr data=crime;
var x2-x10;
run;
proc princomp data=crime out=pccrime;
var x3-x10;
run;

GOptions Reset=Axis Reset=Symbol;
Proc GPlot Data=pccrime;
    Plot Prin2*Prin1=x11 Prin3*Prin1=x11 Prin3*Prin2=x11/VAxis=Axis1
         HAxis=Axis2 Frame;
    Symbol1 C=Black V=Dot I=None PointLabel=("#x1");
```

```
    symbol2 c=blue v=diamondfilled i=none pointlabel=("#x1");
    Symbol3 C=red V=trianglefilled I=None PointLabel=("#x1");
    Symbol4 C=purple V=star I=None PointLabel=("#x1");
Run;
Quit;     *scatter plot on x11(region), label on states;

GOptions Reset=Axis Reset=Symbol;
Proc GPlot Data=pccrime;
     Plot Prin2*Prin1=x12 Prin3*Prin1=x12 Prin3*Prin2=x12/
         VAxis=Axis1 HAxis=Axis2 Frame;
     Symbol1 C=Black V=1 I=None PointLabel=("#x1");
     symbol2 c=blabck v=2 i=none pointlabel=("#x1");
     Symbol3 C=black V=3 I=None PointLabel=("#x1");
     Symbol4 C=black V=4 I=None PointLabel=("#x1");
     Symbol5 C=black V=5 I=None PointLabel=("#x1");
     Symbol6 C=black V=6 I=None PointLabel=("#x1");
     Symbol7 C=black V=7 I=None PointLabel=("#x1");
     Symbol8 C=black V=8 I=None PointLabel=("#x1");
     Symbol9 C=black V=9 I=None PointLabel=("#x1");
Run;
Quit;    *scatter plot on x12(division), label on states;

Proc Sort data=pccrime out=sort_prin1;
     by Prin1;
run;

Data norm_p;
set sort_prin1;
    NN=(_N_-.5)/50;
    Z=PROBIT(NN);
Run;   *Calculate Normal quantiles;

GOptions Reset=Axis Reset=Symbol;
Proc GPLOT data=norm_p;
    title 'Normal Probability Plot for PC1';
    Plot Prin1*Z / HAXIS=AXIS1 VAXIS=AXIS2;
       AXIS1 Label=('Standard Normal') ;
       AXIS2 Label=('Sample PC1') ;
       Symbol1 C=Black V=Dot I=None;
run;                 *plot qqplot for pc1;

Proc Sort data=pccrime out=sort_prin2;
    by Prin2;
run;
```

```
Data norm_p; set sort_prin2;
    NN=(_N_-.5)/50;
    Z=PROBIT(NN);
Run;    *Calculate Normal quantiles;

Proc GPLOT data=norm_p;
    title 'Normal Probability Plot for PC2';
    Plot Prin2*Z / HAXIS=AXIS1 VAXIS=AXIS2 ;
       AXIS1 Label=('Standard Normal');
       AXIS2 Label=('Sample PC2');
       Symbol1 C=Black V=Dot I=None;
run;                *plot qqplot for pc2;

Proc Sort data=pccrime out=sort_prin3;
     by Prin3;
run;

Data norm_p; set sort_prin3;
    NN=(_N_-.5)/50;
    Z=PROBIT(NN);     *Calculate Normal quantiles;
Run;

Proc GPLOT data=norm_p;
    title 'Normal Probability Plot for PC3';
    Plot Prin3*Z / HAXIS=AXIS1 VAXIS=AXIS2 ;
       AXIS1 Label=('Standard Normal') ;
       AXIS2 Label=('Sample PC3');
       Symbol1 C=Black V=Dot I=None;
run;                *plot qqplot of pc3;


proc factor data=crime method=principal scree;
var x3-x10;
run;
*run a default FA and decide the number of factor being chosen;

proc factor data=crime n=3 method=principal ROTATE=VARIMAX S C EV RES
   REORDER SCORE OUT=SCORES1 PREPLOT PLOT;
title" factor analysis with principal method";
var x3-x10;
run;
```

```
GOptions Reset=Axis Reset=Symbol;
Proc GPlot Data=scores1;
  Plot factor2*factor1=x11 factor3*factor1=x11 factor3*factor2=x11/
VAxis=Axis1 HAxis=Axis2 Frame;
  Symbol1 C=Black V=Dot I=None PointLabel=("#x1");
  symbol2 c=blue v=diamondfilled i=none pointlabel=("#x1");
  Symbol3 C=red V=trianglefilled I=None PointLabel=("#x1");
  Symbol4 C=purple V=star I=None PointLabel=("#x1");
Run;
Quit;  *score plot on x11(region);

Proc GPlot Data=scores1;
   Plot factor2*factor1=x12 factor3*factor1=x12 factor3*factor2=x12/
VAxis=Axis1 HAxis=Axis2 Frame;
   Symbol1 C=Black V=1 I=None PointLabel=("#x1");
   symbol2 c=blabck v=2 i=none pointlabel=("#x1");
   Symbol3 C=black V=3 I=None PointLabel=("#x1");
   Symbol4 C=black V=4 I=None PointLabel=("#x1");
   Symbol5 C=black V=5 I=None PointLabel=("#x1");
   Symbol6 C=black V=6 I=None PointLabel=("#x1");
   Symbol7 C=black V=7 I=None PointLabel=("#x1");
   Symbol8 C=black V=8 I=None PointLabel=("#x1");
   Symbol9 C=black V=9 I=None PointLabel=("#x1");
Run;
Quit;   *score plot on x12(division);

proc factor data=crime n=3 method=ml heywood rotate=varimax;
var x3-x10;
run;
*run maximum likelyhood method to obtain Chisquare Test;
```