Journal of the Nigerian Association of Mathematical Physics
Volume 52 (*July & Sept., 2019 Issue*), pp37 – 42
© *J. of NAMP*

A FURTHER MODIFICATION OF GUTTMANN'S RULE IN DETERMINING THE NUMBER OF COMPONENTS IN PRINCIPAL COMPONENT ANALYSIS

[1]Mohammed A., [2]Zakari Y. and [3]Muhammad I.

[1]Division of Academic Planning, Federal Polytechnic, Bida, Nigeria.
[2]Department of Statistics, Ahmadu Bello University, Zaria, Nigeria.
[3]Department of Statistics, Binyaminu Usman Polytechnic, Hadejia, Nigeria.

Abstract

Principal components analysis is a widely used multivariate technique where by researchers attempt to reduce the dimension of a large number of interrelated variables into few non related variables and retaining as much as possible the variation present in the data set. To decide how many components to be retained in principal components analysis remained a very challenging task to researchers. A decision which if made wrongly has drastic effect. Several methods were introduced or modified in many studies. In this study, two new modifications were introduced. The first method looked at confidence intervals around each eigenvalue and components are retained, if the square root of the entire eigenvalue is greater than one (1.0). Method of Monte Carlo was used to simulate multivariate normal data for the analysis. Three different levels of sample size, components loading strengths and numbers of components were used to perform principal components analysis based on correlation matrix. Results from this first method (HGR_1) shows that the method is better than the traditional Guttmann rule (GR) and has the same bias as MGR. The second method (HGR_2) uses the square root of eigenvalues, and then confidence intervals are constructed around these eigenvalues. Components are therefore selected if the entire confidence interval is greater than one (1.0). HGR_2 was the best overall method in all conditions.

***Keywords*:** Gutmann's rule, principal component, eigenvalue, confidence interval

1. Introduction

The origin of Principal Components Analysis techniques are often difficult to trace. However, it is generally accepted that the earliest descriptions of the technique now known as Principal Components Analysis(PCA) were first given by [1],[2]. However, Pearson's comments regarding computations for over 50 years before the widespread availability of computers are interesting. He states that his methods 'can be easily applied to numerical problems,' and although he says that the calculations become 'cumbersome' for four or more variables, he suggests that they are still quite feasible. Thirty two (32) years between Pearson's and Hotelling's papers, very little relevant material seems to have been published, although in [3] indicates that [4] adopted a similar approach to that of Pearson. Also, a footnote in [2] suggests that while [5] was working along similar lines to Hotelling. Hotelling chooses his 'components' so as to maximize their successive contributions to the total of the variances of the original variables, and calls the components that are derived in this way the 'principal components. Hotelling's derivation of PCs uses Lagrange multipliers and ending up with an eigenvalue/eigenvector problem, but it differs in three respects. First, he works with a correlation, rather than covariance matrix, second, he looks at the original variables expressed as linear functions of the components rather than components expressed in terms of the original variables; and third, he does not use matrix notationin [6].

In PCA, selecting the right number of components to retain is very challenging. Also, in [7], a method called Guttmann rule (GR) was introduced, all components whose eigenvalue is greater than one (1.0) were selected. Also, in[8], a modification of

Corresponding Author**:** Zakari Y., Email: yahzaksta@gmail.com, Tel: +2348169766242

the Guttmann rule was proposed. In their work, confidence intervals were computed around each eigenvalue and components were selected when the entire confidence interval of the eigenvalue is greater than one (1.0). The modification shows a better performance than the Guttmann rule. Their new method tends to over select the components to be retained.

A further modification to the Guttmann rule which is called Hybrid Guttmann rule I (HGR_1) will be proposed in this work. Firstly, Confidence interval will be computed around all eigenvalues, and then components will be selected if the square root of the absolute values of the entire confidence interval is greater than one (1.0).

Secondly, square root of all absolute values of eigenvalues will be computed and confidence intervals will be built around each eigenvalue, then number of components to be retained will be selected if the entire confidence interval is greater than one (1.0) and this modification will be known as Hybrid Guttmann rule II (HGR_2).

However, the purpose of this work is to introduce a new method which may help in determining the correct number of components to be retained with the smallest bias in PCA.

As such, in [8], a modification of Guttmann rule in determining the number of factors in Exploratory Factor Analysis was proposed. In their study, confidence intervals were created around eigenvalues and factors were retained if the entire confidence interval is greater than 1.0. Simulated data were used at different levels of sample size, loading strength and number of factors. The new method outperforms the traditional Guttmann rule but does not outperform the Mean Average Partial and Parallel Analysis.

Also, in [9] is another researcher who worked in similar area, his study comprised of two phases. The first phase explore the Guttmann rule, Scree plot, Bartlett's chi-square test, Parallel analysis on normal data using the estimation method of maximum likelihood. Single outlier was introduced in generating sample correlation matrix for different conditions (sample size, number of variables and estimation method). The second phase explored the Guttmann rule, Scree plot, Mean Average Partial and Parallel Analysis of data also containing outlier at different levels of sample size, number of variables and estimation method. The performance of Parallel Analysis and Guttmann rule were generally best across all conditions.

2. Methodology

It is the intention of this work to introduce a further modification of Guttmann rule known as Hybrid Guttmann rule (HGR) and to compare between the traditional Guttmann rule (GR), modified Guttmann rule (MGR) and the Hybrid Guttmann rules.

Simulation

Simulated data will be the consideration of this study. Simulated data with sample sizes of 30, 150 and 240 were used. The number of components considered are 15, 30 and45. Also, 0.3, 0.5 and 0.7 are used as our components loading strengths.

Monte Carlo Simulation

Method of Monte Carlos with R was used to simulate data having similar characteristics (sample size, loading strengths and number of components) with the data used in previous study. There are twenty four different conditions for simulations.

However, simulation of multivariate random numbers with P=15, 30, 45, loading strengths=0.3, 0.5, 0.7 and Sample sizes of 30, 150 and 240.

Guttmann's Rule

This traditional method introduced by [7], was used to determine the relevant components in PCA as one of the methods, eigenvalues were computed from correlation matrix using R. All components with eigenvalues greater than one (1.0) are retained as the most relevant components. Below is the command that computed eigenvalues from correlation matrix when p= 0.3, n=30 and loading strength =0.3;

```
> x=rnorm(450,0:0.3)
> x=matrix(data=x,nrow=30)
>princomp(x)
Call:
princomp(x = x)
Standard deviations:
Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8
1.6463562 1.4838177 1.4188540 1.3405194 1.2006445 1.1208342 1.0782320 0.9434674
Comp.9     Comp.10    Comp.11  Comp.12   Comp.13   Comp.14    Comp.15
0.9051163 0.7385262 0.7276066 0.6153441 0.5652622 0.4307268 0.3334864
 15 variables and 30 observations.
> R<-cor(x)
>eigen(R)
$values
[1] 2.3247442 2.0194171 1.8733817 1.4648847 1.3919827 1.1310513 1.0633430
```

[8] 0.9143492 0.6627826 0.5776558 0.5036371 0.4294573 0.3412821 0.1834562
[15] 0.1185748
The number of components with eigenvalues greater than one are seven, therefore original Guttmann rule suggests that the number of relevant components to be selected is seven ($\widehat{M} = 7$)

Bartlett's Test for n=30, p=15, and loading strengths =0.3
$H_0: l_{14} = l_{15}$ (The 14th and 15th components contributed the same amount of variation)
$H_1: l_{14} \neq l_{15}$ (The 14th and 15th components contributed different amount of variation)
This means that the true number of relevant components is fourteen (M=14).

$$Ck = -\left[n - \frac{2p+11}{6}\right] \sum_{K=14}^{15} \ln(l_k) \quad (1)$$

$$= -\left[30 - \frac{41}{6}\right] \sum (\ln 0.1186 + \ln 0.1835)$$

=88.6715 and

$$\chi^2_{1-\alpha,\ P\left(\frac{P-1}{2}\right)} \quad (2)$$

$$= \chi^2_{0.05,105}$$

= 124.342
Since $c_k < \chi^2_{0.05,105}$ we accept H_0 and continue to test further until H_0 is rejected.
Testing further, we have;
$H_0: l_{13} = l_{14} = l_{15}$ (The 13th, 14th and 15th components contributed the same amount of variation)
$H_1: l_{13} \neq l_{14} \neq l_{15}$ (The 13th to 15th components contributed different amount of variation)
This means that the true number of relevant components is thirteen (M=13).

$$c_k = -\left[30 - \frac{41}{6}\right] \sum (\ln 0.1186 + \ln 0.1835 + \ln 0.3413)$$

= 113.5755
$H_0: l_{12} = l_{13} = l_{14} = l_{15}$ (The 12th to 15th components contributed the same amount of variation)
$H_1: l_{12} \neq l_{13} \neq l_{14} \neq l_{15}$ (The 12th to 15th components contributed different amount of variation). This means that the true number of relevant components is twelve (M=12).

$$c_k = -(23.1667)(-5.7476) = 133.15$$

H_0 is now rejected with M=12, therefore GR bias = $M - \widehat{M} = 12 - 7 = 5$

This was how bias was computed for twenty four (24) different conditions (three levels of sample size, loading strengths and number of observed components).

Modified Guttmann Rule
From the modification of Guttmann rule proposed by [8], confidence intervals (CI) were created around each eigenvalue and components are retain if the entire eigenvalue is greater than one (1.0). The equation for calculating confidence interval of eigenvalues is

$$l_i \pm Z_{1-\frac{\alpha}{2}} \left[\sqrt{\frac{2l_i^2}{n}}\right] \quad (3)$$

Where, l_i represents the observed eigenvalue, $Z_{1-\frac{\alpha}{2}}$ represents the appropriate Z- value for the confidence interval, n is the sample size, and α is the level of confidence.
From Bartlett's test, M=12
$l_{11} = 0.5036$
CI = $0.5036 \pm 1.96(0.13)$
MGR= 0.7585+0.2487= 1.0072 which is > 1
for $l_{12} = 0.4295$
CI = $0.4295 \pm 1.96(0.1109)$
MGR= 0.6494 + 0.2121= 0.8615 which is < 1
So $\widehat{M} = 11$
MGR bias = 12-11= 1

Hybrid Guttmann Rule I (HGR_1)

This is one of the new methods that is introduced in this study, eigenvalues from correlation matrix were used for all the different data sets.

$HGR_1 = \sqrt{MGR}$ (4)

From Bartlett'stest , M=12

Taking $l_{11} = 0.5036$

$HGR_1 = \sqrt{MGR}$

$= \sqrt{1.007}$

=1.0036 and is > 1

For $l_{12} = 0.4295$

$HGR_1 = \sqrt{0.8615} = 0.9282$ Which is < 1

So,$\widehat{M} = 11$and $HGR_1 bias = 12 - 11 = 1$

Hybrid Guttmann Rule II (HGR_2)

This is another approach also introduced in this work;

$L_i = \sqrt{|l_i|}$ (5)

Where l_i is the observed eigenvalue. Confidence interval for L_i are computed and all components with entire eigenvalue (L) greater than one (1.0) will be retained.

Confidence interval for L_i is

$$L_i \pm Z_{1-\frac{\alpha}{2}}\left[\sqrt{\frac{2L_i^2}{n}}\right]$$

From P=15, n=30 and loading strength = 0.3, we have M=12

With $l_{13} = 0.3413$,

$L_i = \sqrt{0.3413}$ = 0.5842

CI = $0.5842 \pm 1.96(0.1508)$

HGR_2 = 0.8798+0.2886=1.1684 this is > 1

$HGR_2 bias = 11 - 12 = -1$

This was how bias are computed for all the different levels of p, n and loading strengths that were used in this study.

3. Results

Table 1: Summary of Barttlet test with Loading Strength = 0.3

N	**P=15** **M**	**P=30** **M**	**P=45** **M**
30	12	28	-
150	14	29	44
240	14	29	44

α= 0.05

Table 1 showsthe required number of relevant components (M) that should be retained when loading strength is 0.3 for different samples sizes and numbers of components.

Table 2: Summary of Barttlet test with Loading Strength = 0.5

N	**P=15** **M**	**P=30** **M**	**P=45** **M**
30	11	29	-
150	14	29	44
240	14	29	44

α= 0.05

Table 4.02 showsthe actual number of relevant components (M) that should be retained when loading strength is 0.5 for different number of samples (n) and different number of components (P).

Table 3 summary of Barttlet test when Loading Strength = 0.7

N	P=15 M	P=30 M	P=45 M
30	12	29	-
150	14	29	44
240	14	29	44

α= 0.05

Table 4.03 showsthe actual number of relevant components (M) that should be retained when loading strength is 0.7 for different number of samples (n) and different number of components (P).

Table 4: Descriptive Statistics of Bias for the Four Methods

Statistic	GR Bias	MGR Bias	HGR_1 Bias	HGR_2 Bias
Mean	14.33	3.58	3.58	**0.17**
Median	15.00	1.50	1.50	-1.00
Std. Deviation	7.26	4.54	4.54	**2.93**
Minimum	5.00	-1.00	-1.00	-2.00
Maximum	25.00	12.00	12.00	8.00

95% confidence intervals were used for MGR, HGR_1 and HGR_2.

Table 4 shows the descriptive statistics for the bias of all the four methods averaged across all number of observed components, sample sizes and loading strengths. As clearly shown, HGR_2 method generally performs better than the other methods with a mean and standard deviation bias of 0.17 and respectively.The second least bias method in determining the number of relevant components in PCA is HGR_1 and MGR with mean bias of 3.58 and median bias of 1.50.

Table 5: Mean Bias across Number of components.

Number of components	GR(SD)	MGR(SD)	HGR_1(SD)	HGR_2(SD)
15	6.33(1.32)	-0.33(1.00)	-0.33(1.00)	-1.11(0.33)
30	15.56(0.88)	5.11(4.99)	5.11(4.99)	**1.89(4.34)**
45	24.33(0.52)	7.17(2.48)	7.17(2.48)	**-0.5(0.55)**

95% confidence intervals were used for MGR, HGR_1 and HGR_2.

Table 5 shows the mean bias across number of observed components for all the four methods. The HGR_2 method outperforms the rest of the methods with a mean bias of -0.5 when number of components is 45. Surprisingly, MGR and HGR_1 methods exhibit the same bias all through.

Table 6: Mean Bias across components loading strengths

components loading strength	GR (SD)	MGR (SD)	HGR_1 (SD)	HGR_2(SD)
0.3	14.25(7.6298)	3.75(4.62)	3.75(4.62)	0.125(2.80)
0.5	14.375(7.67)	3.75(4.92)	3.75(4.92)	0.125(3.23)
0.7	14.25(7.48)	3.35(4.68)	3.35(4.68)	0.25(3.15)

95% confidence intervals were used for MGR, HGR_1 and HGR_2.

Table 4.06 is mean bias across three different loading strengths. From the above table it is seen that HGR_2 is the best with a mean bias of 0.25 when loading strength is 0.7.

Table 7: Mean Bias across different sample sizes

N	GR (SD)	MGR (SD)	HGR_1 (SD)	HGR_2(SD)
30	10.83(6.4)	6.33(5.83)	6.33(5.83)	3.17(4.95)
150	15.33(8.23)	3.56(4.64)	3.56(4.64)	-0.67(0.50)
240	15.67(6.95)	1.78(2.64)	1.78(2.64)	-1.00(0.00)

95% confidence intervals were used for MGR, HGR_1 and HGR_2.

From the above table, HGR_2 outperformed the remaining methods with mean bias of -1.00 and standard deviation of 0.00.

Table 8: Mean Bias for different Ratios of Sample Sizes to Number of Observed components by Methods

Ratio	GR(SD)	MGR(SD)	HGR_1 (SD)	HGR_2(SD)
2:1	5(0)	1.00(0)	1.00(0)	-1.33(0.58)
5:1	15(0)	2.33(1.15)	2.33(1.15)	-1.00(0)
8:1	15(0)	1.33(0.58)	1.33(0.58)	-1.00(0)
10:1	6(0)	-1.00(0)	-1.00(0)	-1.00(0)
16:1	8(0)	-1.00(0)	-1.00(0)	-1.00(0)

95% confidence intervals were used for MGR, HGR_1 and HGR_2.
Table 8 shows the mean bias for different ratios of sample sizes to number of observed components by methods. From the results above, HGR_2 method performs better than the rest methods with a mean bias of -1.33.

4. Discussion of Results

Descriptive Statistics

Table 4.04 shows descriptive statistics for the bias of allfour methods through alllevels of observed components, sample sizes and loading strengths. It can be seen that HGR_2 technique is the most favorable with an average bias of 0.17. Next to this are HGR_2 and MGR with exactly the same average bias of 3.58.

Number of observed components

The various levels of number of components used are 15, 30 and 45. The average bias across these levels as shown in Table 4.05 is quite revealing and interesting. The mean bias tends to be directly proportional to the number of components in all the methods considered.HGR_2 almost appear to have a different pattern with others when number of components is 45.the method tends to over factor with large number of observed components. However the method (HGR_2) is better than the rest three methods with average bias of 1.89.

Factor loading strengths

The average bias across loading strengths in Table 4.06 shows that as loading strengths increases, mean bias appears to decrease in MGR and HGR_1 methods, but GR and HGR_2 exhibit opposite behavior. In HGR_2 the mean bias remain constant at 0.3 and 0.5 as loading strengths and suddenly decrease when loading strength moved to 0.7. Average bias across all the loading strength indicates that HGR_2 has the least bias of 0.125 and is chosen as the best method.

Number of sample sizes

From the mean bias across three different sample sizes, HGR_2 is mostly influenced by sample size. The method shows the smallest mean bias of -1.00 when sample size is 240. The mean bias continue to decrease with increase number of sample sizes.

Ratio of sample size to number of observed components

The mean bias as presented in Table 4.08 shows that HGR_2 is the most affected by different ratio of sample sizes. Other methods are fairly sensitive to changes in sample sizes.

5. Conclusion

To identify the most suitable method of selecting the relevant number of components in PCA is very challenging; GR, MGR, HGR_1 and HGR_2 were compared in this study to ascertain the best method using different levels of sample size, loading strengths and number of observed components. HGR_2 was found to be an improvement over MGR that was introduced in previous study by Russell and Ross (2014).

References

[1] Pearson, K. (1901). On Lines and Planes of Closest fit to System of Points in Space. *Philosophical Magazine volume 2, number 11, 559-572.*

[2] Hotelling, H. (1933). Analysis of a Complex Statistical Variable into Principal Components.

[3] Rao, C.R. (1964). The use and Interpretation of Principal Components Analysis in Applied Research.*Sankhya Journal of Educational Psychology, volume 24, 417-441.*

[4] Frisch, R. (1929). Correlation and Scatter in Statistical Variables. *Nordic Statistical Journal.*

[5] Thurstone, L. (1931). The Measurement of Social Attitude.*Journal of Abnormal and Social Psychology.*

[6] Jollife, I. (2002). *Principal Component Analysis.* Berlin: Springer.

[7] Guttmann, L. (1954). Some Necessary Conditions for Common- Factor Analysis. *Psychometrika, volume 19, 149-161.*

[8] Russell, T. andRoss, L. (2014). Evaluating a Proposed Modification of the Guttmann rule forDetermining the Number of Factors in an Exploratory Factor Analysis. *Journal ofPsychological Test and Assessment Modelling, volume 56, 104-123.*

[9] Swain, S. V. (2009). Determining the Number of Factors in Data Containing a SingleOutlier.*An unpublished Ph. D Thesis of Louisiana State University, Department of Educational Theory, Policy, and Practice.*