SCHOOL OF STATISTICS
UNIVERSITY OF THE PHILIPPINES DILIMAN

WORKING PAPER SERIES

Robust Simultaneous Confidence Interval Estimation of Principal Component Loadings

by

Martin Augustine B. Borlongan
Erniel B. Barrios
Joseph Ryan G. Lansangan

UPSS Working Paper No. 2016-04
September 2016

School of Statistics
Ramon Magsaysay Avenue
U.P. Diliman, Quezon City
Telefax: 928-08-81
Email: updstat@yahoo.com

Abstract

Outlying observations cause perturbations in the eigenvectors that consequently influence the principal components (PCs). We propose a method that integrates bootstrap into the forward search algorithm in the construction of robust confidence intervals for elements of the eigenvectors of the correlation matrix in the presence of outliers. Coverage probability of the bootstrap simultaneous confidence intervals was compared to the coverage probabilities of regular asymptotic confidence region and asymptotic confidence region based on the minimum covariance determinant (MCD) approach through a simulation study. The method produced more stable coverage probabilities for datasets with or without outliers and across several sample sizes compared to approaches based on asymptotic confidence regions.

Keywords: principal component analysis, bootstrap, forward search algorithm, outliers.

MSC Codes: 62F35, 62H12, 62H25

1. Introduction

Principal Component Analysis (PCA) has been extensively used in many applications across various fields but is often used as a descriptive tool and inference is set aside due to complex assumptions it requires on the data structure. There are only a few recent developments on the theories available on probability distributions related to PCA, illustrating the fact that there is little remaining to do and discover in this area (Jolliffe, 2002). There are work that has relaxed assumptions about the data generating process using bootstrap in the construction of empirical distributions of the eigenvalues and eigenvectors, see for example Efron and Tibshirani (1993), Yu et al. (1998), Timmerman et al. (2007)). Furthermore, asymptotic results and bootstrap method assume that the data is free of outlying observation (Jolliffe, 2002). We propose a method that leverages on the implementation of bootstrap incorporated within an automated forward search

algorithm, robustly estimating simultaneous confidence intervals for the elements of the eigenvectors of the correlation matrix from data with outliers.

To extract a robust principal component, it is crucial to identify the outliers individually using a distance measure, e.g., Mahalanobis distance, and to exclude them from the analysis. However, because of masking effect and the distance measure based on the entire data, these observations may not appear to be outliers. Also, cutoff values of Mahalanobis distances are based on quantiles of a chi-square distribution, hence, it necessarily depends on multivariate normality of the data. If n is sufficiently large, then one can invoke the central limit theorem (CLT). However, for smaller sample sizes that are not multivariate normal, these cutoff values may not necessarily provide accurate quantiles. Should removal of outliers to robustify (Jolliffe, 2002) PCA is desired, a different approach is needed to overcome the problem of masking effect or non-normal data (or both).

Jolliffe (2002) described robust estimation as an automatic way of mitigating the effect of extreme or influential observations. PCA can be made robust to outliers without actually identifying them, e.g., use multivariate trimming using robust Mahalanobis distances (Jolliffe, 2002). The location vector and scale matrix are robustly estimated and used in computing Mahalanobis distances instead of the classical mean and covariance matrix, and a specified number of observations with the largest value of these robust Mahalanobis distances are discarded. Alternatively, these outliers can be downweighted, instead of discarding them (M-estimators) (Maronna, 1976). However, since these are all based on robust Mahalanobis distances, influential observations that are not outliers based on this distance will be missed by this approach (Jolliffe, 2002). Jolliffe (2002) also noted that downweighting of an observation can be done based on the observations' influences on parameter estimates instead of using distances.

We focus on estimation of PC loadings that are robust to outlying and influential observations. The forward search algorithm part of the hybrid method aims to robustify PCA.

Forward search algorithm (see for example Atkinson et al. (2004) and Atkinson and Riani (2000)) relies heavily on forward plots of parameter estimates and/or other statistics, and manually monitors these forward plots in search for jumps or dips. The manual search can be very tedious and inefficient and this tends to be subjective on the part of the analyst since there is no clear definition of what a significant jump or dip is in the forward plot. Hadi (1992) suggested the use critical values but was quick to dismiss this idea since it is hard to derive the distribution (if not impossible) of the distance measure used in the forward search algorithm. Without identifying the distribution, there is no statistical method of determining how large should the distance measure be for an observation to be flagged as outlying. Instead, the forward search of Hadi (1992) stops when a fixed sample size h is reached, where $h = (n + p + 1)/2$. This is similar to the smallest possible sample size used in a minimum covariance determinant (MCD) estimator (Rousseeuw and Driessen, 1999). Chosen value of h is a trade-off between accuracy and robustness of the estimates. Considering $h = (n + p + 1)/2$ leads to the highest possible breakdown value of the MCD estimator (Rousseeuw and Driessen, 1999) but compromises the accuracy of the estimates, since the estimates will be based on a smaller sample size. Increasing h, on the other hand, results to declining robustness of the estimates. A balance between accuracy and robustness is achieved by selecting $h = [0.75n]$, the greatest integer in the function (Rousseeuw and Driessen, 1999).

In summary, we developed a method of construction of confidence interval of principal component loadings in the presence of outliers. We also propose a measure of influence in the context of PC Loadings and formulate an alternative method of robust PCA using an automated forward search algorithm. We then construct simultaneous confidence intervals of the loadings of a set of retained PCs using bootstrap.

2. Simultaneous Confidence Interval for PC Loadings

Consider a standard multivariate data where $n > p$, i.e., in matrix form, the data is represented as:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \quad (1)$$

where $\mathbf{x}_i' = (x_{i1}\ x_{i2}\ x_{i3}\ \ldots\ x_{ip})$ is the row vector corresponding to the observed values of the p variables for the i^{th} observation, $i = 1, 2, \ldots, n$.

Robust distance of the observations based on their PC score is used in selecting the best initial subset to be used in the forward search algorithm (clean data). Forward search algorithm sieves excluded observations using an influence measure, and allows only those that satisfy a specified threshold to be part of the clean data. Asymptotic confidence region and bootstrap confidence interval estimates of the PC loadings (simultaneous) are calculated and compared based on their coverage probability.

2.1 Selection of Initial Subset

Riani and Atkinson (2001) emphasized the importance of outlier-free initial data in the forward search algorithm. We present in the subsequent algorithm the process of selecting initial data.

Algorithm 1

1. Perform PCA on the correlation matrix from the entire sample and compute PC scores, denoted by z_{ik}, $i = 1, 2, \ldots, n$ of each of the n observations on the kth PC. Let r be the number of retained PCs. The PC score of observation i on the k^{th} PC can be expressed as $z_{ik} = \mathbf{a}_k'\mathbf{x}_i$, where $k = 1, 2, \ldots, r$. Let the matrix of PC scores be

$$\boldsymbol{Z} = [z_{ik}] = \begin{bmatrix} z_{11} & \cdots & z_{1r} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nr} \end{bmatrix}.$$

2. Compute robust means $\bar{z}_k$ of the PC scores on the retained PCs and the covariance matrix $\boldsymbol{S}_z$ of the PCs, $k = 1, 2, \dots, r$, e.g., Donoho-Stahel estimator, MCD estimator, or orthogonalized quadrant correlation estimator.
3. Calculate the robust Mahalanobis distances from the PC means $\bar{\boldsymbol{z}}_k$, $k = 1, 2, \dots, r$ of the n observations,

$$\text{Mahalanobis}_i = (\boldsymbol{z}_i - \bar{\boldsymbol{z}}_k)' \boldsymbol{S}_z^{-1} (\boldsymbol{z}_i - \bar{\boldsymbol{z}}_k) \quad (2)$$

where $\bar{\boldsymbol{z}}_k$ is a robust estimate of the mean PC score vector and $\boldsymbol{S}_z$ is a robust estimate of the covariance matrix of the k retained PCs.
4. Sort the observations in increasing order based on the computed robust Mahalanobis distances from the previous step.
5. Determine $m = 0.25 * n$, number of observations with the smallest robust Mahalanobis distances from the best initial subset, where a non-integer m is rounded up and $m < n$.

Algorithm 1 initiates the forward search algorithm with clean data along with a protocol on the process of identifying an observation to be added to the clean data in the next iteration.

2.2 The Forward Search Algorithm

2.2.1 A New Influence Measure

To measure the influence of an observation on the estimates of the loadings of the variables on a set of PCs, consider the influence measure denoted by M defined as follows:

$$M_{(-j)} = \sum_{k=1}^{r} \sum_{i=1}^{p} \frac{\left(\hat{\alpha}_{ki} - \hat{\alpha}_{ki(-j)}\right)^2}{r * p * MSC} \quad (3)$$

r is the number of retained PC, p is the number of variables, $\hat{\alpha}_{ki}$ is the estimate of the loading on the k^{th} PC of the i^{th} variable based on all observations and $\hat{\alpha}_{ki(-j)}$ is the estimate of the loading on the k^{th} PC of the i^{th} variable when observation j is excluded. Moreover, the mean squared change (MSC) is defined as:

$$MSC = \sum_{j=1}^{m} \sum_{k=1}^{r} \sum_{i=1}^{p} \frac{\left(\hat{\alpha}_{ki} - \hat{\alpha}_{ki(-j)}\right)^2}{r * p * m} \quad (4)$$

MSC is calculated from the initial dataset of size $m < n$ (outlier-free) and remains constant throughout the forward search.

Jolliffe (2002) used the sum of squared changes in the coefficients of an eigenvector to monitor the effect of omitting an observation from a dataset. The proposed influence measure $M_{(-j)}$ can be decomposed into two parts: MSC and average squared change in the loadings of the variables on the retained PCs resulting from the exclusion of observation j. MSC is the change on each of the loadings of the p variables on a set of r retained PCs based on m observations that are considered "good" observations. If these m observations constitute a clean and outlier-free dataset, then the average change brought about by individual exclusion on the PC loadings represents the maximum allowable change that an observation is expected to cause. Hence, if $M_{(-j)} > 1$, exclusion of observation j caused a change in the PC loadings that is greater than what is expected for a "good" observation.

Similar to the forward version of Cook's D (Atkinson and Riani, 2000), the forward search version of the influence measure M is given by:

$$M_{(+j)} = \sum_{k=1}^{r} \sum_{i=1}^{p} \frac{\left(\hat{\alpha}_{ki} - \hat{\alpha}_{ki(+j)}\right)^2}{r * p * MSC} \quad (5)$$

where the subscript $+j$ indicates the inclusion of one of the excluded observations to the clean dataset. If $M_{(+j)} > 1$, then inclusion of observation j caused a change in the PC loadings that is greater than what is expected for a "good" observation based on the initial dataset.

2.2.2 Inclusion of Observations into the Clean Dataset

Given that there are m observations in clean dataset, denote the matrix corresponding to the dataset formed by adding one unit from the excluded observations to the clean dataset as $\boldsymbol{X}_{(+1)}$ and the matrix corresponding to the clean dataset as $\boldsymbol{X}$. The forward search algorithm is implemented as follows:

Algorithm 2

1. Perform PCA using $Cor(\boldsymbol{X})$, the correlation matrix from the clean dataset, and take note of the eigenvectors $\widehat{\boldsymbol{\alpha}}_k$ corresponding to the r retained PCs.
2. Include one unit from the excluded observations into the clean dataset.
3. Perform PCA using $Cor(\boldsymbol{X}_{(+1)})$, the correlation matrix from the dataset formed by the clean dataset and the inclusion of one unit from the excluded observations. Take note of the eigenvectors $\widehat{\boldsymbol{\alpha}}_{k(+j)}$ corresponding to the r retained PCs, $k = 1, \dots, r$.
4. Calculate $M_{(+j)}$ defined in (5).
5. Repeat steps 2 to 4 until all excluded observations have their corresponding $M_{(+j)}$, $j \in$ {the set of indices of the excluded observations}
6. Arrange the $M_{(+j)}$ in ascending order.
7. The excluded observation with the smallest $M_{(+j)}$ is included in the clean dataset if and only if $M_{(+j)} \leq 1$. This is the stopping criterion that automates the forward search algorithm.
8. Repeat 1 to 7 until: (i) an observation fails to enter the clean dataset based on the criterion in step 7, or (ii) all excluded observations have entered the clean dataset.

Denote the clean dataset from the forward search as $\boldsymbol{X}_c$, an $n_c \times p$ matrix where n_c is the number of observations that are able to enter the clean dataset. This clean dataset is used in the subsequent bootstrap method discussed in the next section.

2.3 Bootstrap Simultaneous Confidence Intervals using Bonferroni-Type Adjustment

The bootstrap method implemented in Algorithm 3 is based on Efron and Tibshirani (1993).

Algorithm 3

1. Given the $n_c \times p$ data matrix $\boldsymbol{X}_c$ from the forward search algorithm, generate bootstrap datasets $\boldsymbol{X}_{cj}^*, j = 1, \dots, B$ which are also $n_c \times p$ matrices, and B is the number of bootstrap samples. The rows $\mathbf{x}_i^*$ of $\boldsymbol{X}_{cj}^*$ are a random sample (with replacement) of size n_c from the rows of the original data matrix $\boldsymbol{X}_c$.
2. Perform PCA on each bootstrap sample. For a given set of retained PCs, and for each bootstrap data set, denote the bootstrap replicate of $\mathbf{a}_k$ as $\mathbf{a}_k^*$.
3. For the elements a_{kj} of $\mathbf{a}_k$, we can construct a simultaneous confidence interval by using Bonferroni-type adjustment to the significance level α and getting the $\left(\frac{\alpha}{2p}\right)$ percentile and $\left(1 - \frac{\alpha}{2p}\right)$ percentile of the bootstrap replicates of each α_{kj}.

Solving the Problem of Axis Reflection

Correction for axis reflection is needed in bootstrapped eigenvectors, see for example Mehlman et al. (1995), Yu et al. (1998), Peres-Neto et al. (2003). Axis reflection occurs when the sign of the loadings on a PC changes from one bootstrap sample to another, resulting to wider confidence intervals that do not reflect the true empirical distribution of the loadings. A method similar to the solution suggested by Mehlman et al. (1995) and Yu et al. (1998) is given in Algorithm 4.

Algorithm 4

1. Given data matrix $\boldsymbol{X}_c$, perform PCA on the correlation matrix and take note of the sign of the loadings on each PC.

2. For a bootstrap sample, perform PCA on the correlation matrix and take note of the sign of the loadings on each PC.
3. The sign of the loadings from step 1 is compared to the sign of the loadings from step 2. If more than 50% of the variables have changed the sign of their loadings on a particular PC, then axis reflection occurred on that PC. The loadings from the bootstrap sample are corrected by multiplying all loadings with -1. Note that comparison and adjustment is done individually for each PC.
4. Steps 2 and 3 are done for all bootstrap samples.

This method of correcting the sign of the loadings is slightly different from the one used by Mehlman et al. (1995) and Yu et al. (1998) where they noted the variable with the highest loading (in magnitude) on each PC from a PCA done on the actual dataset, and the sign of the loading is used as reference for the bootstrap replicates. The loadings of these variables on their corresponding PC are tracked across all bootstrap replicates. Axis reflection on a PC occurred if the sign of the loading of the variable being tracked is not the same as the reference sign, and adjustment is done by multiplying -1 to the eigenvector of this PC. This process is also done for all PCs.

3. Simulation Studies

The following factors are considered in the simulation scenarios:

1. The proportion of "important" variables of the $p = 20$ variables (75% and 50%).
2. The number of clean or outlier-free observations n_0 ($n_0 = 50$, $n_0 = 100$ and $n_0 = 300$ representing small, moderate and large sample size, respectively)
3. The proportion of outliers induced (5% and 10% for $n_0 = 100$ and $n_0 = 300$; 10% for $n_0 = 50$).
4. The proportion of variables containing the outliers (100%, 75%, 50% and 25%)

5. Location of outlying values, i.e., set of affected variables by outlying observations: fixed or random.
6. The degree of deviation of the outlying values of an induced outlier (near or far based on interquartile range (IQR) of each of the variables).
7. The group of variables where outlying values of an induced outlier is located (important group or nuisance group).

Two different proportions of important variables mimic the selection of variables when building an index, where it is presumed that most of the variables belong to a single dimension. There are also instances where some variables are included in the analysis but do not contribute any information regarding that dimension.

Factors 4 to 7 pertain to the nature of the outliers present in the dataset. Mean slippage model is commonly used in generating outliers. This means that "good" observations are generated from a certain distribution with mean $\boldsymbol{\mu}$ while outliers are generated from the same but translated distribution with mean $\boldsymbol{\mu} + \boldsymbol{\delta}$. Hadi (1994) used $\boldsymbol{\delta} = 5$ to generate outliers. Similar scenario on the mean slippage model is achieved in the simulation study when all variables of the outliers have outlying values.

Anscombe and Guttman (1960) noted that aside from inherent variability in the data, there are two more sources of variation: measurement errors which arises from incorrectly calibrated instruments or measurements done using other scales; and execution error such as including elements that do not belong to the target population and measuring the wrong variables. Hence, there will be instances when a group of observations might be outlying on certain variables because of measurement error. Moreover, it is possible that for each outlier, different variables are affected by measurement error. Depending on the error in the calibration of the measuring tool, different degree of deviation might be encountered. This is the reason why the proposed method is evaluated across the combinations of these factors related to the nature of the induced outliers.

3.1 Robust Estimates of the Mean and Covariance Matrix of PC Scores

In the simulation study, we calculate the robust distance of each of the observations (Wang et al., 2014) in selection of a clean, initial subset. The Donoho-Stahel estimator is used if the sample size is less than 1000 and there are less than 10 variables or the sample size is 5000 but there are less than 5 variables. If the sample size is less than 50000 and there are less than 20 variables, then the MCD is used. Otherwise, the Orthoginalized Quadrant Correlation estimator is used.

Simultaneous Confidence Interval Estimates of the PC Loadings

Three simultaneous confidence interval estimates are computed. Namely:

- Asymptotic Confidence Region from the entire dataset prior to the forward search (*Asymptotic-Full*).
- Asymptotic Confidence Region from the subset that gives the MCD estimator of the population correlation matrix, where the subset size is chosen to be the integer part $h = \frac{n+p+1}{2}$ to maximize robustness of the estimate (*Asymptotic-MCD*).
- Bootstrap Simultaneous Confidence Intervals Estimate with Bonferroni Adjustment (*Proposed Method*).

Asymptotic Confidence Region for $\boldsymbol{\alpha}_k$

An approximate confidence region for $\boldsymbol{\alpha}_k$, with confidence coefficient $(1-\alpha)$, has the following form (Jolliffe, 2002):

$$\{\boldsymbol{\alpha}_k | (n-1)\boldsymbol{\alpha}_k'\left(l_k\boldsymbol{S^{-1}} + l_k^{-1}\boldsymbol{S} - 2\boldsymbol{I}_p\right)\boldsymbol{\alpha}_k \leq \chi^2_{(p-1);\alpha} \qquad (6)$$

Selecting the MCD Subset for the Asymptotic-MCD

To determine the MCD subset, (Rousseew et al., 2015), alpha function parameter is set to 0.5 to attain a subset of size h = (n + p + 1)/2, similar to the value of h used by Rousseeuw and Driessen (1999) in their simulation study to maximize robustness of the MCD estimator. The function then returns the subset that provides the correlation matrix with the smallest determinant.

The correlation matrix from this subset is then used in PCA and the subsequent asymptotic confidence region is evaluated based on coverage probability.

3.2 Simulated Data

To generate a population of size $1,000,000$, given the group of correlated variables, a starting $1{,}000{,}000 \times 1$ column vector $\mathbf{x_1}$ is generated from $U(1,10)$ that will serve as the basis vector for the rest of the variables in the important variables group, i.e. $\mathbf{x}_j = a * \mathbf{x}_1 + b$ for some constant a and b. For the nuisance variables, their corresponding columns are generated from $N(10,49)$. This will contain at least 60% explained variance by the first principal component or have a scree plot indicating the retention of the first PC. This procedure in generating a dataset results in the retention of a single PC (Salagubang , 2011).

Generating Samples and Induced Outliers

Simple random sampling without replacement is used to generate outlier-free samples of size n_0 from the population. Then, t outliers are generated based on the specified proportion of variables where outlying values are present, the group of variables where outlying values are located, the degree of deviation of the outlying values, and whether the location of outlying values varies from one outlier to another. The value of t is equal to the specified proportion of outliers multiplied by n_0. Hence, the total number of observations in the *full* dataset (unprocessed) is $n = n_0 + t$. Outlying values on a variable is generated using the definition of outliers in exploratory data analysis. An observation is considered to be a near outlier if its value is either less than the first quartile of the observed values minus $1.5 *$ InterQuartile Range (IQR) or greater than the third quartile of the observed values plus $1.5 *$ IQR. On the other hand, if an observation is less than the first quartile of the observed values minus $3.0 *$ IQR or if it is greater than the third quartile of the observed values plus $3.0 *$ IQR, then it is considered to be a far outlier. Specifically, outliers are generated as follows:

- Upper Tail Outliers: $Q_3 + IQR(observed\ values\ of\ \boldsymbol{X}_k) * z$, where $z \sim U(1.5, 3.0)$ or $z \sim U(3.0, 5.0)$, $k = 1, \ldots, p$ where p is the number of variables
- Lower Tail Outliers: $Q_1 - IQR(observed\ values\ of\ \boldsymbol{X}_k) * z$, where $z \sim U(1.5, 3.0)$ or $z \sim U(3.0, 5.0)$, $k = 1, \ldots, p$ where p is the number of variables
- Each sample will have a combination of Upper Tail and Lower Tail Outliers.

4. Results and Discussion

With fix number of variables ($p = 20$), we vary the sample size (n), proportion of "important" variables (%Important) and the proportion of outliers induced (%Outliers). A simulation study is designed to evaluate the performance of the proposed method compared to two asymptotic approaches for all scenarios. The asymptotic confidence region from the unprocessed data set is denoted as *Asymptotic-Full* while the asymptotic confidence region from the MCD subset is denoted as *Asymptotic-MCD*. The confidence coefficient is set at 0.95 for all three approaches.

4.1 Outlier-Free Datasets

Table 1 shows the size of the clean data identified using Algorithm 1 and the average coverage probabilities of the three methods. With 75% important variables, *Asymptotic-Full* confidence region outperforms the proposed method and the *Asymptotic-MCD* confidence region. Proposed method and *Asymptotic-MCD* yield average coverage probabilities that increase as the sample size increases. The proposed method is generally at par with the *Asymptotic-Full* for moderate ($n_0 = 100$) and large (($n_0 = 300$) datasets. The same pattern for increasing sample size is noted for datasets with 50% important variables. However, average coverage probabilities of the proposed method and *Asymptotic-MCD* are much lower when there are 75% important variables.

Table 1. Average Size of Clean Data and the Average Coverage Probabilities of the three methods for a clean dataset

			Average Coverage Probabilities		
n_0	%Important	Size of Clean Data	Proposed Method	Asymptotic - Full	Asymptotic - MCD
50	75	49.985	0.9150	0.9800	0.4700
100	75	99.995	0.9600	1.0000	0.8550
300	75	299.995	0.9650	1.0000	1.0000
50	50	50.000	0.3900	0.8500	0.2150
100	50	100.000	0.6100	0.9900	0.4800
300	50	300.000	0.9000	1.0000	0.8200

4.2 Induced Outliers

Starting with n_0 observations from the population, t induced outliers are included in the dataset for a total sample size of $n = n_0 + t$. The value of t depends on n_0 and the proportion of outliers required by the simulation setting.

4.2.1 Effect of Sample Size

In Table 2, coverage probabilities for the proposed method for all sample sizes are comparable when there are 75% important variables. On the other hand, average coverage probabilities significantly drop when only 50% of the variables are important, with the coverage probability improving as the sample size becomes large. Regardless of other factors, the simultaneous confidence intervals from the proposed method provide better coverage probabilities than the asymptotic confidence region from the unprocessed dataset regardless of sample size. In comparison to the asymptotic confidence region from the MCD subset, the simultaneous confidence intervals from the proposed method is better with small sample sizes and 75% of the variables are important. Similar is true for small to moderate sample sizes and 50% of the variables are important.

Table 2. Average Coverage Probabilities by Sample Size

75% Important Variables						
				Average Coverage Probabilities		
n_0	**Outliers In Initial Data**	**Outliers In Clean Data**	**Size Of Clean Data**	**Proposed Method**	**Asymptotic - Full**	**Asymptotic - MCD**
50	0	2.36	52.31	0.8485	0.5211	0.7011
100	0	3.06	103.05	0.8541	0.5755	0.8976
300	0	8.41	308.40	0.8497	0.5248	0.9773
50% Important Variables						
				Average Coverage Probabilities		
n_0	**Outliers In Initial Data**	**Outliers In Clean Data**	**Size Of Clean Data**	**Proposed Method**	**Asymptotic - Full**	**Asymptotic - MCD**
50	0	2.51	52.44	0.5014	0.4748	0.3542
100	0	3.47	103.47	0.6295	0.5693	0.5935
300	0	10.16	310.16	0.7318	0.5400	0.8160

4.2.2 Effect of Number of Induced Outliers in the Dataset

In Table 3, average coverage probabilities of the proposed method are better than the *Asymptotic-Full* for all sample sizes and proportion of outliers present in the data. On the other hand, the proposed method performed better than the *Asymptotic-MCD* when $n_0 = 50$, with the latter obtaining average coverage probabilities that are at least at par or greater than the former for larger sample sizes. Better coverage probabilities are obtained when 5% induced outliers are present compared to scenarios with 10% induced outliers. Moreover, for scenarios with 50% important variables, average coverage probability of the proposed method increases as the sample size increases.

Table 3. Average Coverage Probabilities by Sample Size by %Outliers

75% Important Variables						
				Average Coverage Probabilities		
n_0	**%Outliers**	**Outliers In Clean Data**	**Size Of Clean Data**	**Proposed Method**	**Asymptotic - Full**	**Asymptotic - MCD**
50	10	2.36	52.31	0.8485	0.5211	0.7011
100	5	1.95	101.95	0.8896	0.6057	0.8879
100	10	4.17	104.15	0.8186	0.5452	0.9073
300	5	5.12	305.11	0.8596	0.5448	0.9762
300	10	11.71	311.70	0.8398	0.5049	0.9785

50% Important Variables						
				Average Coverage Probabilities		
n_0	%Outliers	Outliers In Clean Data	Size Of Clean Data	Proposed Method	Asymptotic - Full	Asymptotic - MCD
50	10	2.51	52.44	0.5014	0.4748	0.3542
100	5	2.32	102.32	0.6394	0.5833	0.5729
100	10	4.63	104.63	0.6196	0.5552	0.6140
300	5	6.41	306.41	0.7393	0.5537	0.8030
300	10	13.91	313.91	0.7243	0.5263	0.8290

4.2.3 Effect of the Group where Outlying Values of an Induced Outlier are Located

Regardless of whether outliers are in in the important variables group in on both important and nuisance variables groups, the proposed method performed better than the *Asymptotic-Full* approach for all sample sizes (see Table 4). The average coverage probabilities of the confidence region from the unprocessed dataset when outlying values are in the nuisance variables group are at least 0.99 while the average coverage probabilities of the simultaneous confidence intervals from the proposed method are at least 0.91. With the *Asymptotic-MCD* approach, average coverage probabilities increase as the sample size increases regardless of the group where outlying values are located. Moreover, the proposed method is always better than the *Asymptotic-MCD* approach for $n_0 = 50$.

The proposed method performs relatively poorly when outlying values are in the important variables group as compared when outlying values are in the nuisance variables group or in both groups of variables. Even if more induced outliers were able to enter the clean dataset under the scenarios where outlying values are only in the nuisance variables group, the coverage probabilities of the simultaneous confidence intervals from the proposed method are high. The same can be said for the *Asymptotic-Full*, with its coverage probabilities reaching 0.99. Hence, outliers in the nuisance variables may have very little effect on the loadings of the variables. This is not the case however when induced outliers have outlying values only in the important variables group. Inclusion of some of these induced outliers in the clean dataset caused the coverage probabilities

of the proposed method and the *Asymptotic-Full* approach to drop to low values, suggesting that these induced outliers have a significant impact on the PC loadings. On the other hand, when outlying values of the induced outliers are present in both groups of variables, the proposed method was successful in preventing the entry of such observations in the clean dataset. Hence, coverage probabilities under these scenarios are high and are close to the desired coverage probability. Induced outliers in both important variables group and nuisance variables group are the most influential in the computation of the PC loadings, altering the values such that the sample PC loadings are very different from their true population values. This may also be the reason why the coverage probability of the *Asymptotic-Full* approach dropped to values close to 0 when these outliers are included in the dataset.

A different pattern is observed when there are 50% important variables in the dataset, as shown in Table 5. For the proposed method and the *Asymptotic-MCD* approach, average coverage probability improves as the sample size increases. Furthermore, the proposed method consistently obtained higher average coverage probabilities than the Asymptotic- MCD approach for $n_0 = 50$. In comparison to the *Asymptotic-Full* approach, the proposed method did not fare well in terms of coverage probability when outlying values are in the important variables group or in the nuisance variables group.

Table 4. Average Coverage Probabilities by Sample Size by Outlying Values Group for 75% Important Variables

				Average Coverage Probabilities		
n_0	Group	Outliers In Clean Data	Size Of Clean Data	Proposed Method	Asymptotic - Full	Asymptotic - MCD
50	Important Variables	2.54	52.49	0.7311	0.5457	0.6932
50	Nuisance Variables	4.14	54.08	0.9171	0.9900	0.6925
50	Both Groups of Variables	0.40	50.36	0.8971	0.0275	0.7175
100	Important Variables	2.99	102.99	0.7216	0.6641	0.8927
100	Nuisance Variables	5.85	105.85	0.9318	0.9991	0.9063
100	Both Groups of Variables	0.34	100.33	0.9089	0.0632	0.8938
300	Important Variables	7.62	307.61	0.6870	0.5525	0.9746
300	Nuisance Variables	17.14	317.14	0.9379	0.9996	0.9789
300	Both Groups of Variables	0.48	300.46	0.9243	0.0223	0.9784

Table 5. Average Coverage Probabilities by Sample Size by Outlying Values Group for 50% Important Variables

				Average Coverage Probabilities		
n_0	Group	Outliers In Clean Data	Size Of Clean Data	Proposed Method	Asymptotic - Full	Asymptotic - MCD
50	Important Variables	3.47	53.41	0.4754	0.5082	0.3643
50	Nuisance Variables	3.39	53.32	0.5796	0.9057	0.3446
50	Both Groups of Variables	0.66	50.59	0.4493	0.0104	0.3536
100	Important Variables	4.64	104.64	0.5109	0.6855	0.5870
100	Nuisance Variables	4.94	104.93	0.7480	0.9843	0.6129
100	Both Groups of Variables	0.84	100.84	0.6296	0.0380	0.5805
300	Important Variables	13.45	313.45	0.5234	0.6070	0.8127
300	Nuisance Variables	14.69	314.69	0.8925	0.9968	0.8150
300	Both Groups of Variables	2.34	302.34	0.7795	0.0163	0.8204

On the other hand, the proposed method outperformed the *Asymptotic-Full* approach when outlying values are in both the important and nuisance variables group. Note that the proposed method was still able to prevent entry of most of the induced outliers when outlying values are in both important variables group and nuisance variables group.

Hence, in terms of coverage probability, the proposed method is more stable when 75% of the variables are important compared when only 50% of them are important. Asymptotic results with no pre-processing of data is affected when outliers are found in the important variables group or in both important and nuisance variables group, with the latter causing it to have very low coverage probability. As expected, *Asymptotic-MCD* relies on the sample size.

4.2.4 Effect of the Proportion of Variables with Outlying Values

The average coverage probabilities in Table 6 indicates that the simultaneous confidence intervals from the proposed method are better than the asymptotic confidence region based on unprocessed dataset regardless of the proportion of variables with outlying values for each induced outlier. Furthermore, the proposed method performs better than the *Asymptotic-MCD* approach for $n_0 = 50$. For larger sample sizes, the *Asymptotic-MCD* approach obtained average coverage probabilities that are at par or better than the proposed method.

Table 6. Average Coverage Probabilities by Sample Size by Proportion of Variables with Outlying Values for 75% Important Variables

				Average Coverage Probabilities		
n_0	%Var Outlying	Outliers In Clean Data	Size Of Clean Data	Proposed Method	Asymptotic - Full	Asymptotic - MCD
50	25	2.93	52.89	0.8392	0.5996	0.7017
50	50	2.58	52.51	0.8308	0.5342	0.7033
50	75	1.63	51.57	0.8958	0.4075	0.6942
50	100	2.25	52.23	0.8075	0.5650	0.7092
100	25	3.79	103.78	0.8267	0.6867	0.8942
100	50	3.43	103.43	0.8467	0.6048	0.8921
100	75	2.03	102.02	0.9104	0.4300	0.9004
100	100	2.92	102.91	0.8113	0.5854	0.9096
300	25	10.42	310.41	0.8198	0.6035	0.9779
300	50	9.51	309.50	0.8542	0.6013	0.9781
300	75	5.48	305.47	0.9206	0.3958	0.9771
300	100	8.08	308.07	0.7588	0.4725	0.9750

Table 7. Average Coverage Probabilities by Sample Size by Percentage of Variables with Outlying Values for 50% Important Variables

				Average Coverage Probabilities		
n_0	%Var Outlying	Outliers In Clean Data	Size Of Clean Data	Proposed Method	Asymptotic - Full	Asymptotic - MCD
50	25	3.80	53.72	0.4742	0.5346	0.3588
50	50	2.74	52.68	0.5500	0.5050	0.3571
50	75	1.43	51.36	0.5108	0.3725	0.3479
50	100	1.59	51.53	0.4400	0.4992	0.3517
100	25	5.40	105.40	0.5804	0.6552	0.5900
100	50	3.87	103.87	0.6415	0.6160	0.5979
100	75	1.91	101.91	0.6792	0.4250	0.5935
100	100	1.94	101.94	0.6046	0.5925	0.5913
300	25	15.95	315.95	0.6408	0.6023	0.8119
300	50	11.18	311.18	0.7202	0.6246	0.8102
300	75	5.58	305.58	0.8250	0.4208	0.8231
300	100	5.73	305.73	0.7504	0.4846	0.8217

In Table 7, with 50% important variables, the proposed method does not completely outperform the *Asymptotic-Full* approach in terms of coverage probability. The simultaneous confidence intervals from the proposed method have higher coverage probability than the asymptotic confidence region from the unprocessed dataset when 50% or 75% of the variables contain the

induced outliers. On the other hand, the proposed method is still better than the *Asymptotic-MCD* approach for $n_0 = 50$. For larger sample sizes, the asymptotic confidence region from the MCD subset is at par or better than the simultaneous confidence intervals from the proposed method. However, for the same proportion of variables with outlying values, the average coverage probability from the proposed method and the *Asymptotic-MCD* approach improve as the sample size increases.

4.2.9 Robustness of the Simultaneous Confidence Intervals from the Proposed Method

Robustness of the simultaneous confidence intervals from the proposed method has been established for the simulation scenarios considered in the study. Both the simultaneous confidence intervals being proposed, and the asymptotic confidence regions from unprocessed data and from the MCD subset obtained lower average coverage probabilities when only 50% of the variables are important compared when 75% are important. Also, patterns of the performance in terms of average coverage probabilities of the three approaches differ based on the proportion of important and nuisance variables.

With 75% important variables, the simultaneous confidence intervals of the loadings of the variables on the first PC generated by the proposed method is robust to changes in sample size, having average coverage probabilities comparable for all sample sizes. However, it is affected by the number of outliers present in the data. Specifically, average coverage probabilities are slightly lower when there are 10% outliers compared when only 5% outliers are present. Good coverage probabilities are obtained when outliers are present only in the nuisance variables groups or in both important variables group and nuisance variables group. Average coverage probabilities when outlying values are on the important variables group drop to around 0.70. This pattern of performance of the proposed method may be explained by its ability to prevent the entry of influential outliers in the clean dataset. This further explains why there is a significant drop in the coverage probability of the proposed method when outlying values are only in the important variables group.

With 50% important variables, the proposed method is affected by sample size. Average coverage probability of the simultaneous confidence intervals from the proposed method improves as the sample size increases. This is similar to the behavior of the average coverage probability of the asymptotic confidence region based on the MCD subset. Generally, robustness property discussed in the case of 75% important variables also apply to this case. The only difference is that average coverage probabilities are lower. Likewise, the same performance in terms of outlier-detection is also observed.

5. Conclusions

A method of constructing robust simultaneous confidence intervals of PC loadings is proposed. Simulation study show that the simultaneous confidence interval from the proposed method yield better coverage probabilities compared to the confidence region based on "unprocessed" datasets specifically for scenarios that include outlying values in both important variables group and nuisance variables group. In comparison to the asymptotic confidence region based on the MCD subset, the proposed method performed better when the sample size is small. Expectedly, in large sample sizes, the *Asymptotic-MCD* approach dominates the proposed method.

The proportion of important variables in the dataset affects the performance of the method. For the case when majority of the variables are important, the coverage probabilities of the simultaneous confidence intervals from the proposed method are essentially similar across different sample sizes. However, for datasets where only half of the variables are important, higher coverage probabilities are achieved as sample size increases.

Cerioli et al. (2014) noted that the forward search algorithm is capable of generating robust estimates of location and scatter for small datasets provided that bulk of the data is included into the clean dataset. However, for large sample sizes, the forward search algorithm still produces robust estimates even if only a little more than 50% of the dataset is included into the clean data. The proposed method is consistent with this result in the context of PCA. Specifically, the method complements the *Asymptotic-MCD* for smaller sample size. The *Asymptotic-MCD* always utilized a sample size of $h = (n + p + 1)/2$ and its average coverage probability increases as the sample size increases. Furthermore, the proposed method consistently outperformed the *Asymptotic-MCD* approach for $n_0 =$ 50. As Cerioli et al. (2014) noted, it is important to be able to use as many observations as possible in the estimation process when dealing with small sample sizes. For a small sample size, the proposed forward search will arrive at a clean dataset that is always larger than the MCD subset. Assuming that the clean dataset from the proposed forward search and the MCD subset are outlier-free, estimates from the proposed forward search will be more accurate than the estimates from the MCD subset resulting to higher coverage probabilities of the subsequent simultaneous confidence intervals. However, for $n_0 =$ 300, the MCD subset already provides accurate estimates and already outperformed the proposed method. For large sample sizes, stopping at a little higher than 50 % of the original number of observations already provides accurate estimates (Cerioli et al., 2014).

References:

Anscombe, F.J. and Guttman, I. (1960). *Rejection of Outliers*. Technometrics, 2(2):123-147.

Atkinson, A. and Riani, M. (2000). *Robust Diagnostic Regression Analysis*. New York: Springer.

Atkinson, A., Riani, A. and Cerioli, A. (2004). *Exploring Multivariate Data with the Forward Search*. New York: Springer.

Cerioli, A., Farcomeni, A. and Riani, M. (2014). *Strong consistency and robustness of the Forward Search estimator of multivariate location and scatter*. Journal of Multivariate Analysis, 126:167-183.

Efron, B. and Tibshirani, R.J. (1993). *Introduction to the Bootstrap*. New York: Chapman and Hall.

Hadi, A.S. (1992). *Identifying Multiple Outliers in Multivariate Data*. Journal of the Royal Statistical Society, 54(3): 761-771.

Hadi, A.S. (1994). *A Modification of a Method for the Detection of Outliers in Multivariate Samples*. Journal of the Royal Statistical Society, 56(2): 393-396.

Jolliffe, I.T. (2002). *Principal Component Analysis (2nd ed.)*. New York: Springer.

Maronna, R.A. (1976). *Robust M-Estimators of Multivariate Location and Scatter*. The Annals of Statistics, 4(1):51-67.

Mehlman, D.W., Shepherd, U.L. and Kelt, D.A. (1995). *Bootstrapping Principal Components Analysis: A Comment*. Ecology, 76(2):640-643.

Peres-Neto, P.R., Jackson, D.A., Somers, K.M. (2003). *Giving Meaningful Interpretation to Ordination Axes: Assessing Loading Significance in Principal Component Analysis*. Ecology, 84(9):2347-2363.

Riani, M. and Atkinson, A.C. (2001). *A Unified Approach to Outliers, Influence, and Transformations in Discriminant Analysis*. Journal of Computational and Graphical Statistics, 10(3): 513-544.

Rousseeuw, P.J. and Driessen, K.V. (1999). *A Fast Algorithm for the Minimum Covariance Determinant Estimator*. Technometrics, 41(3):212-223.

Salagubang, J.C. (2011). *Outlier Detection in High Dimensional Data in the Context of Clustering*. Unpublished manuscript.

Timmerman, M.E., Kiers, H.A.L. and Smilde, A.K. (2007). *Estimating confidence intervals for principal component loadings: A comparison between the bootstrap and asymptotic results*. British Journal of Mathematical and Statistical Psychology, 60:295-314.

Wang, J., Zamar, R., Marazzi, A., Yohai, V., Salibian-Barrera, M., Maronna, R., Zivot, E., Rocke, D., Martin., D., Maechler, M., and Konis, K. (2014). *robust: Robust Library*. R Package version 0.4-16, URL http://CRAN.R-project.org/package=robust

Yu, C., Quinn, J.T., Dufournaud, C.M., Harrington, J.J., Rogers, P.P., and Lohani, B.N. (1998). *Effective dimensionality of environmental indicators: a principal component analysis with bootstrap confidence intervals*. Journal of Environmental Management, 53:101-119.