On the use of Sparse Principal Component Analysis and Robust: Selection Features of Maize Yield in Rural Tanzania

Article · July 2017

CITATIONS
0

READS
261

3 authors, including:

Justine N Mbukwa
Mzumbe University (MU)
11 PUBLICATIONS **17** CITATIONS

SEE PROFILE

MANTECH Publications

Journal of Engineering Mathematics and Statistics
Volume 1 Issue 1

On the use of Sparse Principal Component Analysis and Robust: Selection Features of Maize Yield in Rural Tanzania

Justine Nkundwe Mbukwa[1], G.V.S.R.Anjaneyulu[2]

Department of Statistics

Acharya Nagarjuna University,

Nagarjuna Nagar Guntur, Andhra Pradesh

Corresponding Author: *mbukwajustine@googlemail.com[1]*

Abstract

This paper has been motivated as a result of an existence of high dimensionality problem in maize yield. This means that an application of the Sparse Principal Component Analysis (SPCA) pattern recognition technique is unknown in selecting few consistent features and easier interpretation as opposed to classical PCA. This paper fulfills the existing knowledge gap in the context of Tanzania. A structure questionnaire was used to collect primary data from Mbozi and Mvomero Districts among small farming household in rural areas. The study was designed on the basis of hierarchical random sampling. The breakdown of facts was made by R-Statistical computing (version 3.3.2) whereas the findings were depicted using graphs and tables. The statistical estimates like percentage, mean and variance were also used. In line with SPCA, PCA and Robust PCA were also fitted for comparison purpose. Results showed 19 variables were condensed to six components explaining 63.7 per cent variations under PCA. Contrary to these findings, there were great improvements of the loadings, consistent and easier to interpret in each PC of the modified model (SPCA). However, the paper discovered that the Robust PCA condensed the p-variable to two PCs such that PC1 explained (81.0 per cent) variances. The study recommends the Sparse and Robustness as the best filtering techniques with reliable results as contrasted to the ordinary PCA.

Keywords: *Classical Principal Components Analysis, Sparse Principal Component Analysis, Dimensionality Reduction, Robustness, Smallholder Farmers and Maize Yield*

1. INTRODUCTION

In multivariate statistical methods, the Principal Component Analysis (PCA) model is responsible for the dimensionality reduction and feature selection from Variance-Covariance matrix p-observed random variables. The new vectors formed with their estimated loadings per variable are called components (Ning-min and Jing (2015)). It is widely used in many fields of studies such as biological studies, engineering, sciences as well as social sciences. It ascertains the maximum variations from the large through transformation of set of interrelated variables (Dominick et al (2012), Mutalib et al(2013)). The PCs are the new set resulting from the linear combinations of the original variables (Skrbi et al (2007), Juneng et al (2009)). The Sparse and Robustness PCA dimensionality models are not known in the context of Tanzania, and of the of the problems have been reported using statistical significance tests. Though, the classical PCA has been documented in some places in the same field, yet it lacks the degree of consistency to both variable estimation and interpretations. Thus, this paper intends to fill the existing knowledge gap as an optimal feature selection of maize yield among smallholder farmers in rural Tanzania.

2. ESTIMATION OF THE CLASSICAL PRINCIPAL COMPONENT ANALYSIS (PCA)

The maximization of variations from the p-random variables through linear combination has been argued by (Hotelling (1933)). Suppose $\mathrm{x}^T = (\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_P)$ is the random variable in p-dimension vector with the Variance –Covariance matrix (**Σ)**, it possible to rewrite this vector in a new vector linear combination $z = e^T(x_i - \bar{x})$. This is in line with finding vector $\underset{\sim}{e}$ such that $\mathrm{Var}(z)$ is at the maximum. If the vector $\underset{\sim}{e}$ is an orthogonal, it can be written as a unit length as $e^T e = 1$. This matrix has eigenvalues-eigenvectors pairs $(\lambda_1, \mathrm{e}_1), (\lambda_2, \mathrm{e}_2), \ldots, (\lambda_p, \mathrm{e}_p)$ and $\lambda_1 \geq \lambda_2 \geq ... \lambda_p > 0$. Therefore, the i^{th} PC can be derived based on:

$$Y_i = e_i^T x = e_{1i}x_1 + e_{2i}x_2 + e_{3i}x_3 + ... + \ell_{pi}x_p \quad for\ i = 1, 2, ... \mathrm{p} \qquad \ldots (2.1)$$

The expression (2.1) can be presented in terms of variation as:

$$\mathrm{Var}(y_i) = Var(e_i^T x) = e_i^T \Sigma e_i = \lambda_i \qquad for \quad i = 1,2,....,p \qquad \ldots (2.2)$$

$$\mathrm{Cov}(Y_i, Y_k) = 0 \qquad for \quad i \neq k \qquad \ldots (2.3)$$

2.1 Retaining the Components

The eigenvalue greater than one criterion has been suggested (Kaiser (1960), Jackson (1991), Johnson and Wichern (992)). The screeplot graphical representation has been recommended. Thus, the graph provides a precise results when the study sample (n>200) (Stevens (1986)). Schmitt and Sass (2011) cited in (Zygmont and Smith (2014)) reported that the loadings of 0.3 or 0.4 are sufficient for variables to form good pattern structure in data.

2.2 The Modified SPCA

Some weaknesses of the ordinary PCA lead to the coefficients to be unstable. This is because each component is a linear combination of all p-variables such that interpretations of them in terms of estimated loadings remain a problematic. This means that the predicted components are associated with unstable loadings. Thus, SPCA has been suggested to rescue this setback by removes all predictors whose coefficients are closely to zero in all PCs with a given values less than some threshold value. This latest approach consisting of a small number of non-zero coefficients have been reported by (Johnstone and Lu(2012), Qi, Luo and Zhao (2013)). This predictive optimal technique takes into account by inserting the absolute values threshold to zero (Cadima and Jolliffe (1995) cited in Zou et al (2006)).

To be able solve this research issue, the Least Absolute Shrinkage Selection Operator (LASSO) was recommended (Tibshirani (1996)). It rationalizes the computation, stable parameter estimates and higher precision by shrinking some loadings to zero. The LASSO technique has been adapted to PCA in connection to standard regression:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i \qquad i = 1,2,...,n;\ j = 1,2,...,p \qquad \ldots. (2.2.4)$$

$y_1, y_2, ..., y_n$ are the values on response variable y

$x_{ij}\ (i = 1,2,...,n;\ j = 1,2,...,\mathrm{p})$ Correspond to the values of p; $e_1, \mathrm{e}_2, ..., \mathrm{e}_n$ are error term and $\beta_0, \beta_1, \beta_2, ..., \beta_n$ the parameters of the regression equation. The factor estimates are through the partial derivative total square error as shown as:

$$\sum_{i=1}^{n} e_i^{\,2} = \sum_{j=1}^{n} (\mathrm{y}_i - \beta_0 - \sum_{j}^{p} \beta_j x_{ij})^2 \qquad \ldots (2.2.5)$$

To find the estimated parameters $\beta = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_n)$ in equation (3.13), the partial derivatives are used.

Basically, the LASSO forces the constraints on coefficients denoted as:

$$\sum_{j=1}^{p} || \beta_j || \le k \qquad \ldots (2.2.6)$$

For the suitable choice of parameter t, this imposed control forces some of the loadings in regression to zero. The LASSO is derived through minimization of the sum of the squares residual with addition penalty function $\sum_{j=1}^{p} || \beta_j ||$. Thus, minimize:

$$\sum_{i=1}^{n} e_i^{\,2} = \sum_{j=1}^{n} (\mathrm{y}_i - \beta_0 - \sum_{j}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} | \beta_j | \qquad \ldots (2.2.7)$$

For given certain multiplier λ

In this regard, $\lambda \ge 0$ is a turning point constraint that uses to manage the restriction such that as $\lambda \to 0$, the linear estimate is computed and $\lambda \to \infty$, $\hat{\beta}_{lasso}$ estimated. Therefore, λ should be in between where both a linear model is fitted and coefficients are shrunk to zero where the lambda is very large.

3. LITERATURE REVIEW ON BOTH MODIFIED AND ORDINARY PCA

Apart from the LASSO as proposed by (Tibshirani (1986), some other scholars like Vines (2000) and Anaya-Izquierdo et al (2001) cited Croux et al(2013)) proposed to use the ridges or constraints on the coefficients. This modified model has been also noted in the field of bioinformatics (Chen(2011)), recognizing optimal groups

and traits choice (Lounici (2013)), multivariate in the direction of time series analysis (Wang, Han and Liu(2013)), text form of big data analysis (Zhang and Ghaoui (2011)) as well as cancer investigative research to select the relevant attributes that speed up the degree of the problem (Hsu (2014)) cited in (Ning-Min and Jing (2015)).

The ordinary PCA has been reported fundamentally as a method of data reduction. Ong'ala et al (2016) used the technique to discover the knowledge and pattern recognition of the sugar yield representative traits. It was disclosed that about 80.6 per cent total variations were explained by ten components.

The question of attentiveness of small farming household on risk problems facing them has been disclosed by (Kasaka-Lwayo and Obi (2012)). About 66.13 per cent of the variances were predicted by seven PCs. It has further revealed that the identified sources of farmers risks are statistically significant associated with age, gender, education and information access. However, it known that in order to perform the PCA, there must be several variables that are measuring the same parameter of interest in the following section.

3.1 Empirical Studies on Correlated variables

In order to perform the dimensionality reduction via the Principal Component Analysis, the p-variables need to relate and measuring alike one thing in common. In line with this paper, here are the reviewed related research works: Baha et al (2013) fitted stochastic frontier modeling to measure the sources of maize inefficiency among farmers in Tanzania. Thus the farm size, formal education, number of plots owned by a farmer, number of times a farmer contacts with extension officers, use of insecticides and use of hand hoes were found to be substantial in order to increase efficiency maize production in the study area. This indicated about 37.70 per cent inefficient of maize.

Msuya et al (2008) reported the level of maize efficiency in Tanzania. The results indicated the average productivity was 1.19tone per hectare. Technically, the maize proficient is sixty one on average. The low levels of education, lack of extension services, limited capital, land fragmentation and in access as well as high input prices contributed negatively on technical efficiency of maize in the study area. In Malawi, Chirwa et al(2007) argued that an

improved seeds or varieties often have higher output oriented of beans to farmers than indigenous seeds per given piece of land. Yengoh (2012) come up with the tangible findings from the factors that predict the yield differences among the selected small farming household in Cameroon. The farmers who apply improved seeds had a great chance of realizing more productivity than those that do not apply completely. However, it is very common for the smallholder farmers use the indigenous skills to select the superior seeds emanating from their past produce for the coming season.

Birachi et al (2011) highlighted the factors facing the amount of production and sales by smallholder farmers in Burundi. Using systematic random sampling a total sample size of 380 farmers was estimated. The results indicated the considerable change in beans productivity a result of using, improved seed, fertilizers and market demand. The level of adoption and spatial distribution of improved common bean varieties in Southern Highlands of Tanzania has been reported (Letaa et al (2014)). In this study, it was discovered that the quantity of beans produced tremendously influenced the quantity marketed. The small farming households, who realized high productivity, are also likely to supply large proportion of their beans to the market place. Samiee et al (2009) disclosed some aspects of correlation between crop yield and level of education of the smallholder producers.

Therefore, the education of a person may increase in the level of awareness on how to use all types of agricultural inputs such as animal droppings, improved seeds and fertilizers. Bumb et al(2011) revealed that an application fertilizer is very substantial to uplift considerable crop yield in Cameroon. In line with this inputs it likely to improve income from farming sector and producers become less vulnerable to other expected risks like crop failures and food shortages. FAO (2011 a) highlighted that the use of inorganic fertilizers should be concurrent practiced hand by hand with the soil management practices to land sustainability to crop produce.

In this practice, it is possible to increase crop production by conserving the soil and producing more. Ariga et al (2008) noted in Western highlands of Kenya. In a small scale subsistence farming system, maize consumes more amounts of fertilizers as than other staple food like cassava, yams

and grain legumes (Crawford et al., 2006). URT (2013) revealed that only 10kg/ha of fertilizer is used in Tanzania, more than 50kg per ha in South Africa while in SADC average is 16kg/ha is consumed and Vietnam is 365kg/ha. Goyari (2014) conducted the study in Udalguri District of Assam in India to uncover the seasonal dissimilarities to cultivators who implemented high yielding varieties (HYV) and Traditional Varieties (TV).

The results indicated that there was more yield for chemical users than non-users. Amos (2007) reported a frontier estimated model that there is a positive relationship between age and number of years a smallholder farmer attended to school with the level of inefficient carried a study determine the technical efficiency of cocoa production in Nigeria. It has been reported by Mushunje (2005) argued that the more the farm sizes possessed by the smallholder farmers the higher crop yield than those with small pieces of land. Cousins (1989), Abel and Blaikie (1989) discovered that there was an inversely relationship between the herd size and yield. This is due to the fact that the more the cattle the more the arable land are required.

4. DESCRIPTION OF THE STUDY AREA AND METHODS

The study was conducted in Tanzania, in East Africa which lies along the Latitude 60 00' S and Longitude of 350 00' East of Greenwich. It covered two districts namely Mvomero (eastern zone) and Mbozi (southern highland zone). They are found along two distinct agro-ecological zone with two varied rainfall distributions (bimodal and unimodal) respectively. They receive an average annual rainfall (935 mm) and (955 mm) correspondingly (Magehema, Chan'a and Mkoma(2014)). The multistage random sampling with five hierarchies was used. Initially the agricultural zone was treated as a single cluster, followed by district, division, ward, village and households or individual smallholder farmer as lower level of analysis.

In this paper, a total of 430 sample size was estimated using the systematic random sampling. The relevant personal primary data were collected using structured questionnaire preceded by pilot study. During the field work, some variable's measurements such as quantity of maize harvests were first recorded in terms of number of buckets or containers. Data analysis was done using R statistical

package (3.3.2). This was preceded by the descriptive measures followed by the classical PCA and ultimately sparse modeling. The actual field activities took place from October, 2015 to February, 2016. During the field, each target respondent was requested to provide the quantities harvested in his/her farm for previous farming season within six months in terms of bags/number of tins/buckets for easier conversion into kilograms along with other information on related variables from stipulated standard questionnaire.

In this paper, the maize yield estimation was made via recall approach (Casley and Kumar (1988), Howard et al (1995), Lekas et al (2001), Erenstein, Malik and Singh (2007)).The exercise was done at farmer's house or close to store to enable both enumerators and the respondents to be comfortable in the course of providing the accurate of the estimates followed by physical verifications.

To ensure the precision of the research findings, the ethical issue was highly observed along with clearance form. This was in line with ensuring the question of validity. In this regard, the present study included soley smallholder farmers who have been registered to the Mtandao wa Vikundi Vya Wakuma Wadogo Tanzania (MVIWATA) synonymously referred as small farmers group network. Table 4.1 indicates the detailed definitions of the correlated variables that have been used in understanding the patterns from high dimensionality of maize yield in the study area.

Table 4.1: Definition of the variables used the current study

Variable name	Descriptions of the variable
age (x1)	age of the household head
nsyear(x2)	number of year a responded attended at school
hsize (x3)	household size
ncow (x4)	number of cows

ngoat (x5)	number of goats
nsheep (x6)	number of sheep
nchcken (x7)	number of chicken
qtylocalseedmaizerain(x8)	quantity of local seed maize per acre in kg
nhiredlabor (x9)	number of hired labor used in farming activities
nhomelabor (x10)	number of home labor used in farming activities
qtyimprseedmz(x11)	quantity of improved seeds of maize in kg per acre
qtypestmzrain (x12)	quantity of pesticides used for maize per acre
qtyfertmaize(x13)	quantity of fertilizers in Kg per acre
Mzhar2(x14)	Total quantity of maize harvested
fsmzr2(x15)	Total farm sizes of maize in acre
qtysamz-y2r(x16)	quantity of sales (marketed) maize in kg
price_imprmaizer (x17)	price of improve seed maize
price_fertmaizer (x18)	price of fertilizer for the quantity in kg used
price_Lpestmaizer(x19)	price of pesticides for maize per liter in Tanzanian shillings

5. PRESENTATIONS OF THE FINDINGS AND DISCUSSION

5.1 Descriptive Measures

The survey study via structured questionnaire aimed to cover the sample of 430 farming households. For the purpose of analysis, only 97.9 per cent response rate was employed. The research results showed that 265(63 per cent) of the respondents were from Mbozi district (Table 5.1.2).

Table 5.1.2: Distribution of the Respondents by District

>transform(as.data.frame(table(anu$district)),percentage_column=Freq/nrow(mpb1)*100)

Var1	Freq	Percentage_column
1 Mvomero	156	37
2 Mbozi	265	63
Total	421	100

Table 5.1.3: Distribution of the responds by gender

>transform(as.data.frame(table(anu$gender)),percentage_column=Freq/nrow(mpb1)*100)

Var1	Freq	Percentage_column
1 Male	239	57
2 Female	182	43
Total	421	100

Table 5.1.4: Respondent's Summary for the Maize Yield, Age and Farm size

```
> summary(yieldmz)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     2.0     7.5     9.3    16.0    35.0
> summary(age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.00   33.00   41.00   42.04   50.00   83.00
> summary(fsmzr2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   2.000   2.091   3.000  18.000
```

Table **5.1.3** indicates the gender participation rate in the study area. It was revealed that, there were a great proportion of male in the field place (57 per cent). Yet, these study finding conquer with that of (FAO (2011)) report indicating that in developing countries, more than 40 per cent are female.

Table **5.1.4** depicts the some descriptive measures of the respondents. On average there were 9.3 bags maize yield/acre per respond, 42 of their ages and 2 acres being cultivated by and individual (Table 5.1.4).

5.2 Descriptive Measures for classical PCA for Related Variables

Best feature selection is in line with standardizing 19 observable variables via scale () function in R-statistical computing. To see the visible components, the variables must be highly correlated to explain one thing in common under factor map as shown in the first and second quadrant (Figure 5.2.1).

```
> fviz_pca_var(pca, col.var ="cos2")+
+  scale_color_gradient2(low="red",
mid="blue",
+ high="black", midpoint=0.6)
```

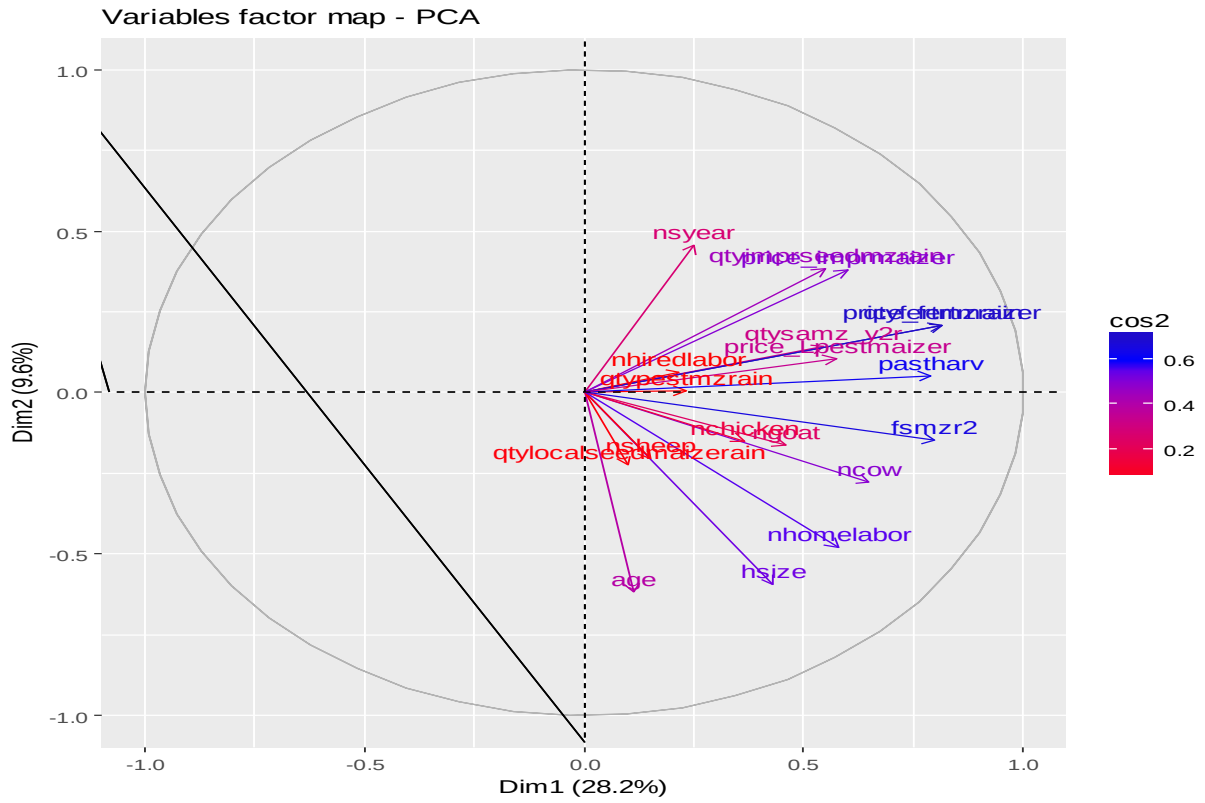

Figure 5.2.1: Visualization of the variables on the factor map:

Figure 5.2.2, indicates the highly correlated variables visualization and predictive power in the first estimated component. This has been facilitated by the function fviz_pca_contrib()[factoextra] package in R. Since there many variables, only top contribution variables can easily be estimated R code from pc1 to 6 (Figure 5.2.2).

If of all random variables factors could have been contributing uniformly, the joint contribution would anticipated to be one per it length(p-predictors) = 0.1 = 10 per cent as shown by a red line to the figure below. Thus, for a given component, a variable whose contribution exceeds this cutoff could be considered as significant contributing to the component.

> fviz_contrib(pca, choice = "var", axes = 1)

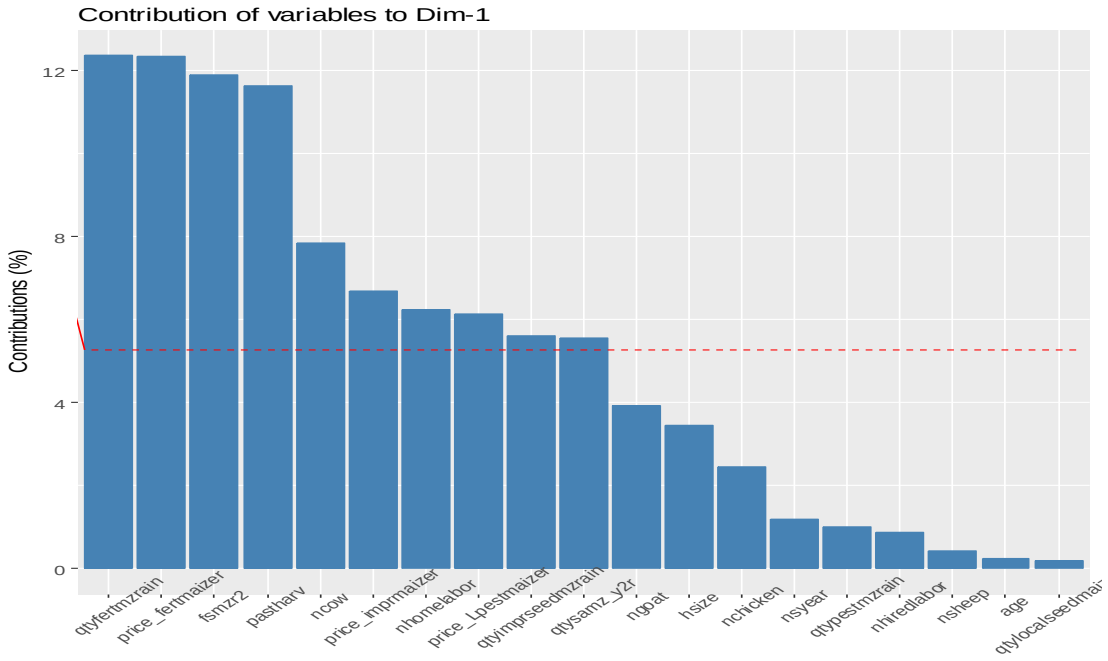

Figure 5.2.2: Contributions of variables on PC1

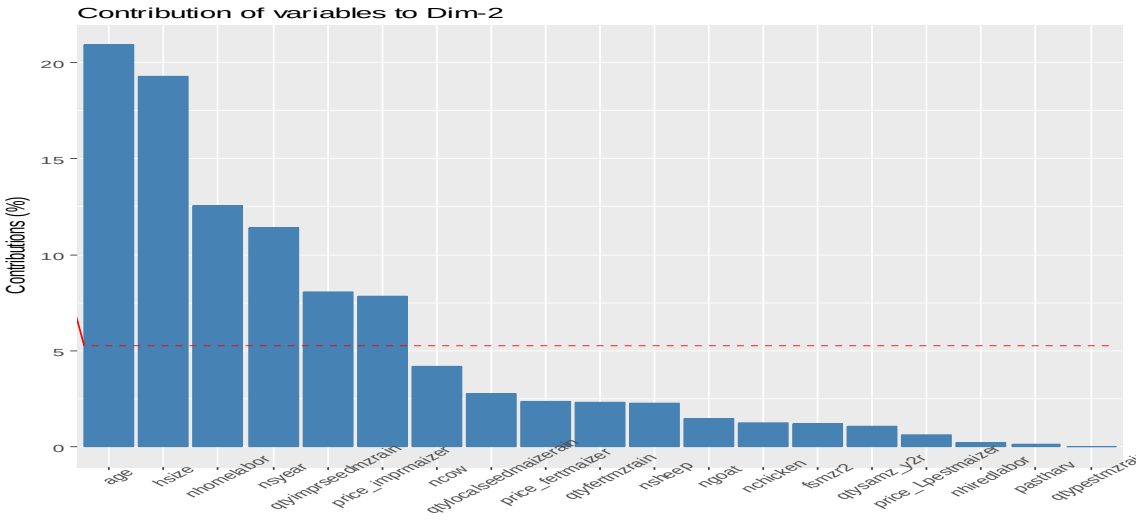

Figure 5.2.3: Contributions of variables on PC2

In figure 5.2.3 indicates the variables that have a predictive power and seems to be significant in the pc2. All variable above the red line are important.

> fviz_contrib(pca, choice = "var", axes = 2)

5.3 Results for Classical PCA

Table 5.3.5 indicates the extent to which each component is regressed over all 19 related variables. The research findings highlighted that about 63.7 percent were explained by the retained six PCs in the study area. In these results, the PC1 is positively related to all nineteen variables. On the other hand, the amount of inorganic fertilizers used per acre, the past maize harvests, the plot size in acres and cost of fertilizers in Tanzanian shillings varied positively with the estimated PC1 as well as their coefficients were above the threshold value (ibid).

In PC2 and the number of years a respondent has been at school were positively related (0.3376). The respondent's age (-0.4575), plot size in acres (-0.4389) and the number of home' manpower (-0.3543) had the same negative coefficient's sign and they move together in different direction with PC2. The PC3 vary in opposite direction with number sheep (-0.3282), number of chicken (-0.3161) as well as the number of hired labor (-0.4679). The PC4 and the quantity of local seed maize (0.6308) were positively related whereas it varied inversely with the age of the farmer (-0.3003), the number of hired labor (-0.3286) and quantity of improved maize seed (-0.3026).

The PC5 were positively correlated with the quantity of pesticides maize (0.6648) and the price in litre of pesticides under maize (0.3093) while number of year the respondents were at school (-0.3028) and number of goats (-0.3256) were opposite related. The PC6 is positively associated with the household sizes (0.3149) and number of sheep (0.4736) .However, it is negatively related with number of chicken (-0.4134). The results in this paper are somehow different from that of ong'ala et al (2016) whose 17 variables were reduced to 10 (80.6 per cent variances), Kisaka-lwayo and Obi (2012) reduced the 20 perceived risk facing smallholder farmers to 7 components explaining 66.13 per cent variance).

Table 5.3.6: Number of Principal Components (PCs) and (19) Related Variables

	PC1	PC2	PC3	PC4	PC5	PC6
age	0.04918477	-**0.457512644**	0.20512846	-**0.300339224**	0.229306051	-0.1851169439
nsyear	0.10858104	**0.337642586**	-0.26561202	0.106800780	-**0.302655010**	0.2551657077
hsize	0.18594284	-**0.438938534**	0.09945967	-0.088241182	-0.185134319	**0.3140873938**
ncow	0.27986696	-0.205147814	-0.10700988	0.212003052	-0.113466686	-0.1833579963
ngoat	0.19785133	-0.120731061	0.12516535	0.129882501	-**0.325648664**	-0.0686217647
nsheep	0.06517389	-0.150304535	-**0.32820641**	-0.153806621	0.157282372	**0.4735664274**
nchicken	0.15684570	-0.110853371	**-0.31610729**	-0.039879311	-0.030689753	-**0.4134012612**
qtylseedz	0.04379812	-0.165841204	-0.11007745	**0.630389451**	0.084061678	-0.0419675499
nhiredlabor	0.09384286	0.047102843	-**0.46792886**	-**0.328556034**	0.289460007	0.1842232690
nhomelabor	0.24978958	-**0.354300853**	0.15893273	-0.115593906	-0.188713118	0.1363125381
qtyimprseedmz	0.23660430	0.284182897	0.20919241	-**0.302604724**	-0.048910728	-0.1072649424
qtypestmz	0.10026208	0.004327353	0.14258566	0.138224303	**0.664759328**	-0.1440847159
qtyfertmz	**0.35156576**	0.152803124	0.19564097	0.106263118	0.046485207	0.1807907491
pastharv	**0.34099223**	0.038800315	-0.19613736	-0.012336258	-0.043290899	-0.2702692944
fsmzr2	**0.34464426**	-0.110395453	-0.24756612	-0.072048829	0.055937882	-0.0007184448
qtysamz_y2r	0.23562426	0.104589932	-0.25233170	-0.005030497	-0.028645574	-0.2591764745
price_imprmz	0.25861340	0.280279680	0.26133278	-0.277443723	-0.001501021	-0.1034022377
price_fertmz	**0.35122830**	0.153914963	0.20595077	0.117689524	0.050263337	0.1920831347
price_Lpestmz	0.24764669	0.078831370	0.09938476	0.258221168	**0.309299277**	0.2478668034
std dev	2.316	1.3495	1.2295	1.0960	1.0815	1.0143
% variance	0.282	0.0958	0.0796	0.0632	0.0616	0.0541
Total variance	0.282	0.3782	0.4578	0.5210	0.5826	**0.6367**

Note: the bolded and underlined PCs coefficients/loadings are >|0.3|

5.4 Results for Sparse PCA (SPCA)

To enhance the proper interpretation of the classical PCA, the modified PCA has been addressed (Table 5.4.7). The study findings portrayed that the SPC1 move in the similar direction with the amount of local seed maize (0.258), the size of pesticides maize/acre (0.259), maize fertilizer in kg (0.505), the price of industrial fertilizer per 50 bag (0.544) and the price of pesticides per litre (0.56). The SPC2 is inversely related with the household size (-0.716), number of cows (-0074), number of goats (-0.276) and number of home labor -0.636). Furthermore, the quantity of pesticides per acre (0.033) was positively correlated with sparse PC2. The variables like the household size (-0.019), number of sheep (-0.662), number of hired labor (-0.713) and farm size (-0.216) were likely to move together as contrary to SPC3. The numbers of goat whose loading is (0.076) was directly related with sparse PC3.

The SPC4 indicated a directly correlation with quantity of local maize seed (0.412) while other remaining variables vary inversely related. The SPC5 is positively associated with (age (0.668) and quantity of pesticides for maize (0.283)) while the rest are negatively related. In the SPC6, all variables varied inversely proportional. The modified PCA results from this paper bear a resemblance with that of Tibshiran (1986), Croux et al(2013), Ning-Min and Jing (2015) whose results indicated that the later findings remain more stable in terms of interpretations.

Table 5.4.7: Sparse Principal Component Analysis (SPCA)

```
> sparcePCA6 <- spca(stdmaize, K = 6, type = "predictor", sparse = "varnum",
+                    para = c(5, 5, 5, 5, 5,5), trace = FALSE)
>sparcePCA6
Call:
spca(x = stdmaize, K = 6, para = c(5, 5, 5, 5, 5, 5), type = "predictor", sparse = "varnum", trace = FALSE)
```

6 sparse PCs

Pct. of exp. var. : 12.5 8.3 6.9 7.8 6.8 7.8

Num. of non-zero loadings: 5 5 5 5 5 5

Sparse loadings

	SPC1	SPC2	SPC3	SPC4	SPC5	SPC6
age	0.000	0.000	0.000	0.000	**0.684**	0.000
nsyear	0.000	0.000	0.000	0.000	**-0.672**	0.000
hsize	0.000	**-0.716**	**-0.019**	0.000	0.000	0.000
ncow	0.000	**-0.074**	0.000	0.000	0.000	**-0.402**
ngoat	0.000	**-0.276**	**0.076**	0.000	0.000	0.000
nsheep	0.000	0.000	**-0.662**	0.000	0.000	0.000
nchicken	0.000	0.000	0.000	0.000	0.000	**-0.362**
qtylocalseedmaizerain	**0.258**	0.000	0.000	**0.412**	0.000	0.000
nhiredlabor	0.000	0.000	-0.713	0.000	0.000	0.000
nhomelabor	0.000	**-0.636**	0.000	0.000	0.000	0.000
qtyimprseedmzrain	0.000	0.000	0.000	**-0.612**	0.000	0.000
qtypestmzrain	**0.259**	**0.033**	0.000	0.000	**0.283**	0.000
qtyfertmzrain	**0.505**	0.000	0.000	**-0.164**	**-0.003**	0.000
pastharv	0.000	0.000	0.000	0.000	0.000	**-0.668**
fsmzr2	0.000	0.000	**-0.216**	0.000	0.000	**-0.285**
qtysamz_y2r	0.000	0.000	0.000	0.000	0.000	**-0.424**
price_imprmaizer	0.000	0.000	0.000	**-0.647**	0.000	0.000
price_fertmaizer	**0.544**	0.000	0.000	**-0.101**	**-0.002**	0.000
price_Lpestmaizer	**0.562**	0.000	0.000	0.000	0.000	0.000

5.5 Checking the Outliers via Variable Graph Map in PCA

To avoid the issue model misspecification, the outliers were careful tested. It entails to identify an observation that is far distant from other observations in data set as supported in the literature (Barnet and Lewis, 1995 cited in Walfish (2006), Croux et al (2013). The observations that were deviated from others have been indicated (figure, 5.5.5) which calls for robust principal component analysis to be fitted. >fviz_pca_ind(pca)

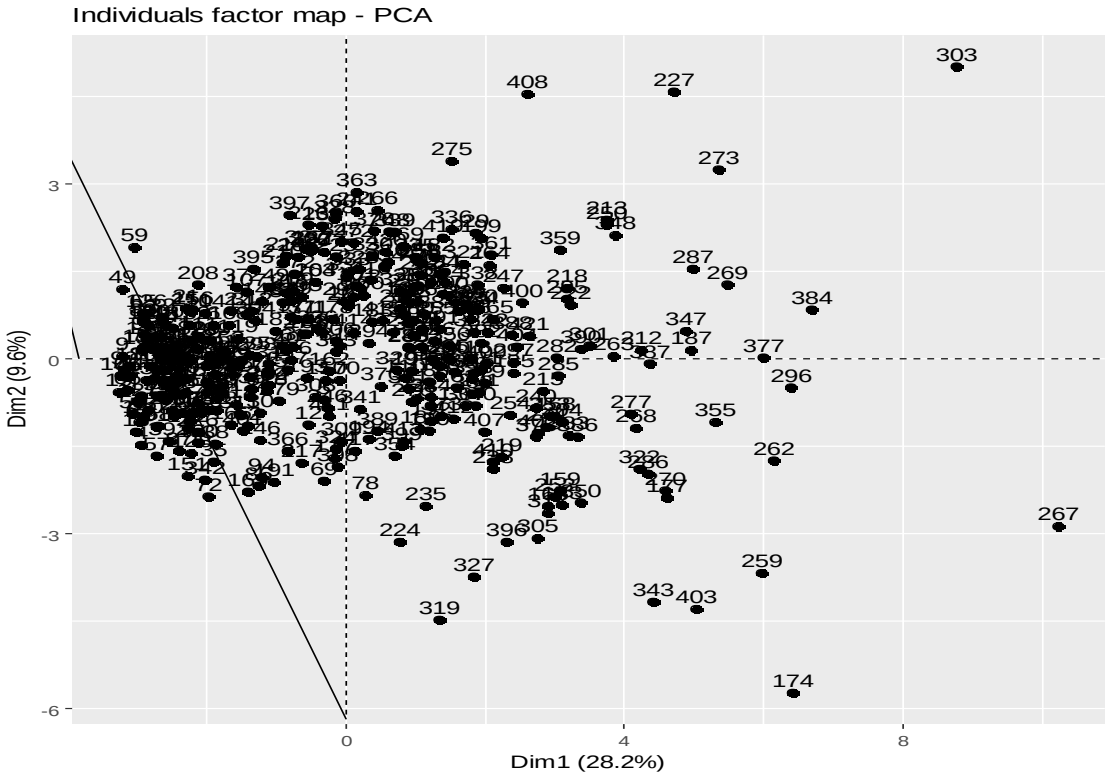

Figure 5.5.5: Visualization of the variables on the factor map

5.6 Robust PCA Fit model

Zhang, Lin, Zhang and Chang (2015) indicated that the classical PCA is sensitive to outliers in such way that the even a single observation may result in poor performance of the model. Thus, both robust for PC1 and PC2 explain about 0.8662 total variations. These findings support the work which stressed to make the estimates stable due to outliers (ibid). The research results have been presented below (Table 5.6.8)

Table 5.6.8: Robust Principal Component Analysis

```
> ropca <- PCAgrid(stdmaize)
> ropca
Call:
PCAgrid(x = stdmaize)
Standard deviations:
[1] 2.7 1.3
19  variables and  421 observations.
```

> ropca2$loadings

Loadings:	Comp.1	Comp.2
age		0.844
nsyear	0.103	
hsize		0.220
ncow	0.250	
ngoat	0.358	0.138
nsheep		
nchicken		0.217
qtylocalseedmaizerain		
nhiredlabor		
nhomelabor		0.374
qtyimprseedmzrain	0.219	

qtypestmzrain	0.102	
qtyfertmzrain	0.471	
pastharv	0.301	
fsmzr2	0.217	
qtysamz_y2r	0.108	
price_imprmaizer	0.147	-0.124
price_fertmaizer	0.461	
price_Lpestmaizer	0.351	
Standard deviation	2.73	1.32
Proportion of Variance	0.81	0.19
Cumulative Proportion	**0.81**	1.00

CONCLUDING REMARKS

In this paper, it was discovered that a total of 63.7 per cent variations were explained by without destroying the original value of the p-variables. In line with this model the sparse principal component was fitted to capture the effect through regressing one to all p- variables. In this modified PCA, the results showed that the quantity of local maize seed, quantity of pesticides, quantity of fertilizer maize, price of fertilizer and price in litre for pesticides have a power to component 1 while household size, number of cows, number of goat and number of home labor load highly in component 2. The model simplified a pattern of variables to be visualized. Furthermore, it has been observed that the robust PCA to be significant over both prior models estimated since it has improved the estimates and number of components from six to three with great variances. Hence both sparse and robust are recommended the best.

ACKNOWLEDGEMENTS

The completion of this work is the results of joint effort from many stakeholders. To mention few, the first and foremost, the authors would like to recognize the financial support from the government of Tanzania via Mzumbe University. Highly appreciation should go to the respondents (farmers), the farmer's network association (MVIWATA) for wonderful cooperation shown, village leaders and the field assistants. This integration work managed to bring the sound base and consolidated filled the research gap for maize yield in Tanzania.

REFERENCES

1) Ahmed, M. H. (2016). Climate change adaptation strategies of maize producers of the Central Rift Valley of Ethiopia. Journal of Agriculture and Rural Development in the Tropics and Subtropics (JARTS), 117(1), pp.175-186.

2) Amos T. T. 2007. 'An Analysis of Productivity and Technical Efficiency of Smallholder Cocoa Farmers in Nigeria', Journal of Social Science, 15(2): 127 - 133

3) Amin, M. R., Zhang, J., & Yang, M. (2015). Effects of climate change on the yield and cropping area of major food crops: A case of Bangladesh. Sustainability, 7(1), pp.898-915.

4) Baha, M., Temu, A and Philip, D. (2013). Sources of Technical Efficiency Among Smallholders Maize Farmers in Babati District, Tanzania. African Journal of Economic Review, 1(2), 1-14.

5) Birachi, E. A., Ochieng, J., Wozemba, D., Ruraduma, C., Niyuhire, M. C., & Ochieng, D. (2011). Factors influencing smallholder farmers' bean production and supply to market in Burundi. African Crop Science Journal, 19 (4), pp.335-342.

6) Bumb, B.L., Johnson, M.E., Fuentes, P.A (2011): Policy Options for Improving Regional Fertilizer Markets in West Africa. Washington DC: IFPRI Discussion Papers 01084; The International Food Policy Research Institute (IFPRI); 2011.

7) Casley, D. J., and Kumar, K. (1988). The Collection, Analysis and Use of Monitoring and Evaluation Data.

Baltimore, MD: Johns Hopkins University Press for the World Bank.

8) [8] Chabala, L. M., Kuntashula, E., Kaluba, P., & Miyanda, M. (2015). Assessment of Maize Yield Variations Due to Climatic Variables of Rainfall and Temperature. Journal of Agricultural Science, 7(11), 143.

9) Chen, X. (2011). Adaptive elastic-net sparse principal component analysis for pathway association testing. Statistical applications in genetics and molecular biology, vol. 10, no. 1, pp. 1-21.

10) Chirwa, R.M., Aggarwal, Vas, D., Phiri, M.A. and Mwenda, R.A. 2007. Experiences in implementing the Bean Seed Strategy in Malawi. Journal of Sustainable Agriculture 29(27):43-69.

11) Croux, C., Filzmoser, P., & Fritz, H. (2013). Robust sparse principal component analysis. Technometrics, 55(2), 202-214.

12) Dominick, D., Juahir, H., Latif, M. T., Zain, S. M., & Aris, A. Z. (2012). Spatial Assessment of air Quality Patterns in Malaysia using Multivariate Analysis. Atmospheric Environment, 60, 172–181. doi:10.1016/j.atmosenv.2012.06.021.

13) FAO. (2011a). Save and Grow. A policy maker guide to the sustainable Intensification of smallholder crop production. Rome

14) FAO 2011: The State of Food and Agriculture Report 2010-2011: "Women in agriculture: Closing the gender gap for development".

15) Goswami, A. K., R.S. Chauhan.,& D.S. Dalawat, 2005: Revs of Hydroxytriazene. Anal.chem., 24:pp.75-102.

16) Goyari, P. (2014). Irrigation Difference and Productivity Variations in Paddy Cultivation:Field Evidences from Udalguri District of Assam. In. Jn.of Agri.Eco.Vol,69, Issue No.1,pp.90-106

17) Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of educational psychology, 24(6), 417-441.

18) Jackson J. A User's Guide to Principal Components. New York: John Wiley & Sons; 1991.

19) Johnstone, I. M., & Lu, A. Y. (2012). On consistency and sparsity for principal components analysis in high dimensions. Journal of the American Statistical Association. 104(486): 682–693.

20) Juneng, L., Latif, M. T., Tangang, F. T., & Mansor, H. (2009). Spatio–temporal characteristics of PM10 concentration across Malaysia. Atmospheric Environment, 45, 4370–4378.

21) Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20, 141-151.

22) Kisaka-Lwayo, M., & Obi, A. (2012). Risk perceptions and management strategies by smallholder farmers in KwaZulu-Natal Province, South Africa. International Journal of Agricultural Management, 1(3), 28-39.

23) Lekasi, J. K., J. C. Tanner, S. K. Kimani, and J. P. C. Harris. (2001). Managing Manure to Sustain Smallholder Livelihoods in the East African Highlands. Ryton on Dunsmore, UK: Henry Doubleday Research Association.

24) Letaa, E., Kabungo, C., Katungi, E., Ojara, M., & Ndunguru, A. (2014). Farm Level Adoption and Spatial Diffusion of Improved Common Bean Varieties in Southern Highlands of Tanzania. African Crop Science Journal, 23(3), pp.261- 277.

25) Magehema, A. O., Chang'a, L. B., & Mkoma, S. L. (2014). Implication of rainfall variability on maize production in Morogoro, Tanzania. International Journal of Environmental Sciences, 4(5), 1077-1086.

26) Mijinyawa, Y., & Akpenpuun, T. D. (2015). Climate change and its effect on grain crops yields in the middle belt in Nigeria. African Journal of Environmental Science and Technology, 9(7), pp.641-645.

27) Msuya, E. E., Hisano, S., & Nariu, T. (2008). Explaining productivity

variation among smallholder maize farmers in Tanzania. This paper was presented in the XII World Congress of Rural Sociology of the International Rural Sociology Association,Goyang, Korea 2008.

28) Mugi-Ngenga, E. W., Mucheru-Muna, M. W., Mugwe, J. N., Ngetich, F. K., Mairura, F. S., & Mugendi, D. N. (2016). Household's socio-economic factors influencing the level of adaptation to climate variability in the dry zones of Eastern Kenya. Journal of Rural Studies, 43, pp.49-60.

29) Mutalib, S. N. S. A., Juahir, H., Azid, A., Sharif, S. M., Latif, M. T., Aris, A. Z., Zain, S. M., & Dominick, D. (2013). Spatial and temporal air quality pattern recognition using environmetric techniques: a case study in Malaysia. Environmental Science,Processes & Impacts 1-12,doi: 10. 1039/c3em00161j

30) Ning-min, S., & Jing, L. (2015). A Literature Survey on High-Dimensional Sparse Principal Component Analysis. International Journal of Database Theory and Application, 8(6), 57-74.

31) Omoyo, N. N., Wakhungu, J., & Oteng'i, S. (2015). Effects of climate variability on maize yield in the arid and semi arid lands of lower eastern Kenya. Agriculture & Food Security, 4(1), pp.1-13

32) Ong'ala, J., Mwanga, D., & Nuani, F. (2016). On the Use of Principal Component Analysis in Sugarcane Clone Selection. Journal of the Indian Society of Agricultural Statistics,70(1), pp.33-39

33) Owuor, G., Wangia, S.M., Onyuma, S., Mshenga, P. and Gamba, P. (2004). Self-Help Groups, A milk supply chain. Review of Agricultural Economics 31(1):pp.103-121.

34) Qi, X., Luo, R., & Zhao, H. (2013). Sparse principal component analysis by choice of norm. Journal of multivariate analysis, 114, 127-160.

35) Rasul, G., Chaudhry, Q. Z., Mahmood, A., & Hyder, K. W. (2011). Effect of temperature rise on crop growth and productivity. Pak. J. Meteorol, 8, 53-62.

36) Samiee, A., Rezvafar, A., Faham, E.(2009). "Factors influencing the adoption of pest management (IPM) by wheat growers in Varamin Country, Iran". African Journal of Agriculture and research Vol.4 (5).pp.vol (4)5:491-61

37) Stevens, J. (1986). Applied multivariate statistics for the social sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

38) Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), vol.58, issue.1, 267-288.

39) United Republic of Tanzania. Ministry of Agriculture Food Security and Cooperatives Agriculture, Sector Review and Public Expenditure Review (ASP/PER2008/09)

40) URT (2013): National Agricultural Policy. Ministry of agriculture food security and Cooperatives Dar es Salaam, October 2013

41) Yengoh,G.T.(2012). Determinants of yield differences in small-scale food crop Farming systems in Cameroon. Agriculture & Food Security ,1(1), pp.1:17

42) Yesuf, M., and Kohlin, G. (2008). Market imperfections and Farm Technology Adoption Decisions:A case study from the Highlands of Ethiopia. Environment for Development Discussion Paper Series, Environment for development Central America, China, Ethiopia, Kenya, South Africa and Tanzania: pp. 1-2

43) Walfish, S. (2006). A review of statistical outlier methods. Pharmaceutical technology, 30(11), 82.

44) Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. Journal of computational and graphical statistics, 15(2), pp.265-286

45) Zhang, H., Lin, Z., Zhang, C., & Chang, E. Y. (2015, January). Exact Recoverability of Robust PCA via Outlier Pursuit with Tight Recovery Bounds. In AAAI (pp. 3143-3149).

46) Zygmont, C., & Smith, M. R. (2014). Robust factor analysis in the presence of normality, violations, missing data, and outliers: empirical questions and possible solutions. Quant Methods Psychol, 10(1), 40-55.