Robust Principal Components with Fast and Robust Bootstrap: An application to populated areas in Santa Fe

Javier Bussi
IITAE, Univ. Nacional de Rosario, Rosario, Argentina - jbussi@fceecon.unr.edu.ar

Gonzalo Mari
IITAE, Univ. Nacional de Rosario, Rosario, Argentina - mari.gonzalo@gmail.com

Fernanda Mendez
IITAE, Univ. Nacional de Rosario, Rosario, Argentina - nandixx@hotmail.com

Abstract

Principal components analysis (PCA) is a widely used technique within multivariate statistical methods. The purpose of this technique is adequately representing a set of $n$ observations and $p$ variables through fewer variables constructed as linear combinations of the original ones. It is based on the calculation of eigenvalues and eigenvectors of the covariance (or correlation) matrix. The presence of outliers in the data can distort the sample covariance matrix. Therefore various ways have been proposed to deal with this difficulty using robust techniques. In this paper we considered a related type of PCA based on robust estimates of shape. In particular we use the eigenvectors and eigenvalues of multivariate MM-estimators of shape. As in classical PCA, results based on asymptotic normality can be used to construct confidence intervals or to estimate standard errors under the assumption of some underlying elliptical distribution. Such assumptions are often not appropriate in those cases where robust estimation is most recommended, therefore alternative techniques should be used, among which is the Bootstrap method. The Bootstrap inference applied to robust estimators such as the MM-estimator requires fewer assumptions but involves high computational cost and a loss of robustness in the presence of outliers. An alternative computationally simpler and more resistant to the presence of outliers is the Fast and Robust Bootstrap (FRB). FRB can be used to obtain many recalculations of the MM-shape or scatter matrix. As with the classical bootstrap, we will base our inference on the eigenvalues and eigenvectors of these recomputed matrices. The use of the robust principal components method and the robust bootstrap inference is illustrated on a dataset of indicators of critical needs of populated areas of the province of Santa Fe, Argentina, from the National Census Population and Housing 2010, in order to achieve a stratification of the populated areas for a future sampling frame. The first robust component provides a summary of three indicators of social needs (education, basic services and housing spaces) into a single index that can be used in order to sort the observation units according to their social needs. We use the geometric method to stratify the populated areas since the distribution of the critical needs index is asymmetrical. It is observed that the needs increase from the south to the north of the province.

Keywords: Principal components analysis; MM-estimators; Fast and Robust Bootstrap.

1. Introduction

Principal component analysis (PCA) is a widely used technique in multivariate statistics. The purpose of this technique is to adequately represent a set of $n$ observations and $p$ variables through fewer variables constructed as linear combinations of the original ones. It is based on the eigenvectors and eigenvalues of the sample covariance (or correlation) matrix. The sample covariance matrix however is notorious for being sensitive to outliers. Therefore various ways have been proposed to deal with this difficulty using robust techniques. In this paper we considered a related type of PCA based on robust estimates of shape. In particular we use the eigenvectors and eigenvalues of multivariate MM-estimators of shape. MM-estimators are designed to be both highly robust against outliers and highly efficient for normal data.

We are primarily interested in the inference part of the PCA method based on the MM-estimator. As in classical PCA, results based on asymptotic normality can be used to construct confidence intervals or to
estimate standard errors under the assumption of some underlying elliptical distribution. Such assumptions are often not appropriate in those cases where robust estimation is most recommended, therefore alternative techniques should be used, among which is the Bootstrap method.

The Bootstrap inference applied to robust estimators such as the MM-estimator requires fewer assumptions but involves high computational cost and a loss of robustness in the presence of outliers. An alternative computationally simpler and more resistant to the presence of outliers is the Fast and Robust Bootstrap (FRB). FRB can be used to obtain many recalculation of the MM-shape or scatter matrix. As with the classical bootstrap, we will base our inference on the eigenvalues and eigenvectors of these recomputed matrices.

In Section 2 some theoretical aspects of PCA based on the MM-estimates are discussed. In Section 3 the FRB method is described. In Section 4 the use of the robust principal components method and the robust bootstrap inference is illustrated on a dataset of indicators of critical needs of populated areas of the province of Santa Fe from the National Census Population and Housing 2010. Section 5 contains some concluding remarks.

2. PCA based on the MM-estimator

MM-estimators of multivariate location and shape are based on two loss functions that must verify the following regularity conditions:

(R1) \( \rho \) is real, symmetric, twice continuously differentiable and \( \rho(0) = 0 \).

(R2) \( \rho \) is strictly increasing on \([0, c] \) and constant on \([c, \infty) \) for some finite constant \( c \).

Multivariate MM-estimators of location, shape and covariance are now defined as follows:

**Definition 1** Let \( \mathcal{X}_n = \left\{ x_1, \ldots, x_1 \right\} \subset \mathbb{R}^p \) with \( n \geq p + 1 \). Let \( \rho_0 \) and \( \rho_1 \) satisfy (R1) and (R2) and let \((\hat{\mu}_n, \hat{\Sigma}_n)\) minimize \( \frac{1}{n} \sum_{i=1}^{n} \rho_0 \left( \left[ (x_i - T)^{\prime}C^{-1}(x_i - T) \right]^{\frac{1}{2}} \right) \) among all \((T, C) \in \mathbb{R}^p \times \text{PDS}(p)\). Here \( \text{PDS}(p) \) denotes the set of positive definite symmetric \( p \times p \) matrices. Denote \( \hat{\sigma}_n := |\hat{\Sigma}_n|^{1/(2p)} \). Then the multivariate MM-estimators for location and shape \((\hat{\mu}_n, \hat{\Sigma}_n)\) minimize \( \frac{1}{n} \sum_{i=1}^{n} \rho_1 \left( \left[ (x_i - T)^{\prime}G^{-1}(x_i - T) \right]^{\frac{1}{2}} / \hat{\sigma}_n \right) \) among all \((T, G) \in \mathbb{R}^p \times \text{PDS}(p)\) for which \(|G| = 1\). The MM-estimator for the covariance matrix is \( \hat{\Sigma}_n = \hat{\sigma}_n^2 \hat{\Gamma}_n \).

The idea is to estimate the scale by means of a very robust S-estimator, and then estimate the location and shape using a different \( \rho \) function that yields better efficiency at the central model. In this paper we consider loss functions in the family of Tukey’s biweight functions.

\[
\rho(x) = \begin{cases} 
\frac{x^2}{2c^2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} & |x| \leq c \\
\frac{|x|}{c} & |x| \geq c 
\end{cases}
\]

where \( c > 0 \) is a user-chosen tuning constant.

The robust PCA method based on MM-estimator essentially consists of estimating these eigenvalues and eigenvectors by the eigenvalues and eigenvectors of the MM-estimator of shape \( \hat{\Gamma}_n \).

The class of MM-estimators should be used to improve the efficiency of S-estimators when improvement is desirable, that is, in relatively low dimensions \((p < 15)\). For higher dimensions, the S-estimator is highly efficient itself. Finally, the difference in computational complexity is not a major factor since the extra M-step for the MM-estimator is almost negligible in comparison with the computation of the initial S-estimator.
3. Fast and robust bootstrap for multivariate MM

We are primarily interested in the inference part of the PCA method based on the MM-estimator. As in classical PCA, results based on asymptotic normality can be used to construct confidence intervals or to estimate standard errors. However, these results only hold under the assumption of some underlying elliptical distribution. Such assumptions are often not appropriate in those cases where robust estimation is most recommended. The bootstrap provides a computer-intensive alternative, however applying bootstrap on robust estimators such as the MM-estimator raises some difficulties. One serious problem is the high computational cost of these estimators, since computing the MM-estimator, particularly the initial S-estimator, is a time-consuming task, which is particularly costly for large datasets in high dimensions. Another typical problem that arises is the instability of the classical bootstrap procedure, since the presence of outliers in the original sample can be repeated in large numbers in the bootstrap samples. Therefore, although the MM-estimator on the original data yields a robust PCA solution, it may fail in bootstrap samples with many outliers. That is, the classic bootstrap may be less robust than the MM-estimator itself. To overcome the problems associated with applying the classical bootstrap to robust estimators on potentially contaminated data, we used an adaptation of the Fast and Robust Bootstrap of Salibian-Barrera and Zamar (2002) to multivariate MM-estimators. This extension can be used to obtain many recalculations of the MM-shape or scatter matrix. As with the classical bootstrap, we will base our inference on the eigenvalues and eigenvectors of these recomputed matrices.

4. Example

One of the objectives of any census, is to be the basis to construct sampling frames for inter-census surveys. The possibilities are that the census constitutes itself in a complete frame of the population, or that a master sample is selected from it. The latter is often preferable than the former because of the simplicity of update activities. On the counterpart, a master sample is formed by only a part of the population, and is selected using probabilities sampling designs in order to make proper inferences to the whole population.

There is some history of master samples in Argentina. The National Institute of Statistics and Census (INDEC) developed a urban sampling frame based on a sample after the 1991 Population Census. This frame was extended to the rural population after the 2001 Census. In the last 2010 Census, the INDEC constructed a Urban Master Sample. The frames mentioned allow to make inferences at different levels of disaggregation, but not inside a province.

For this reason, the Provincial Institute of Statistics and Censuses (IPEC) of Santa Fe intended to develop a housing master sample that permits inferences for different levels of disaggregation inside the province. This application attempts to solve the stratification problem of the populated areas of the province, that forms the primary sampling units (PSU) of the sampling design.

The first classification is related to size: cities $^1$ (more than 10000 inhabitants), towns (between 2000 and 10000 inhabitants), and villages (below 2000 inhabitants). Within each group, the objective is to form homogeneous group related to social and demographical variables from the 2010 Census. To do that, we took 12 indicators for identifying critical needs of the populated areas $^2$. Using the principal component analysis (PCA), the aim is to construct an index of critical needs with the first principal component to sort the populated areas in each group.

Using the MM estimators and a diagnostic plot, we can detect in each group, some populated areas with big robust distance possessing great influence in the estimation of the eigenvectors. This means they could

---

$^1$The capital city, Santa Fe, and the biggest city, Rosario, are not included in the application because they enter into the sample with certainty

$^2$http://www.santafe.gov.ar/index.php/web/Estructura-de-Gobierno/Ministerios/Economia/Secretaria-de-Planificacion-y-Politica-Economica/Direccion-Provincial-del-Instituto-Provincial-de-Estadistica-y-Censos-de-la-Provincia-de-Santa-Fe/Temas-Especificos/Datos-Estadisticos/Poblacion/Censo-Nacional-Poblacion-Hogares-y-Viviendas-2010/Todos-los-Distritos-de-la-Provincia/Estadisticas/CARENCIAS/Indicadores-de-Carencias-Criticas-segun-Censo-Nacional-de-Poblacion-2010.-Provincia-Santa-Fe
greatly influence a classical principal component analysis. Instead, they would have little influence on robust principal component analysis because large values of robust distances correspond with small weights in the MM estimators.

There are several influence points in the cities and towns diagnostic plots that may cause a greater influence in the classic PCA. This can be verified comparing the percentage of explained variance of the first principal component in the classical and the robust PCA.

<table>
<thead>
<tr>
<th>Size Group</th>
<th>Classic PCA</th>
<th>Robust PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cities</td>
<td>80.7%</td>
<td>70.6%</td>
</tr>
<tr>
<td>Towns</td>
<td>89.2%</td>
<td>64.1%</td>
</tr>
<tr>
<td>Villages</td>
<td>71.7%</td>
<td>71.1%</td>
</tr>
</tbody>
</table>

We considered the use of the robust PCA in all groups. In the following figures, the loadings of the first principal component are shown with the 95% BCa Bootstrap confidence intervals.

It can be seen, in all cases, that the loadings of the first principal component have positive sign, which can be viewed as a weighted mean of all variables. Thus constitutes the index of critical needs that resumes gaps in education, basic services, and housing characteristics. The next step considers the computation of the index for each populated area, and the grouping into each stratum. We consider the geometric stratification method of Gunning and Horgan (2004), that is an approximate method suitable for asymmetric populations.

The reason of this choice it can be seen in the following histograms of the index for each type of populated area.
For this application, we consider 5 strata, from a very low deficiency in critical needs (stratum 1) to a very high deficiency (stratum 5). We present for each group and each region into which the province is divided, the number of populated areas that belong to each stratum and region.

### Table 2: Classification of the populated areas according to the region and stratum

<table>
<thead>
<tr>
<th>Region</th>
<th>Cities</th>
<th></th>
<th>Towns</th>
<th></th>
<th>Villages</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Stratum</td>
<td>Total</td>
<td>Stratum</td>
<td>Total</td>
<td>Stratum</td>
<td>Total</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5</td>
<td></td>
<td>1 2 3 4 5</td>
<td></td>
<td>1 2 3 4 5</td>
<td></td>
</tr>
<tr>
<td>Venado Tuerto</td>
<td>1 2</td>
<td>3 6 8 3 17</td>
<td>1 7 6 2</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rosario</td>
<td>3 10 4 1 1 19</td>
<td>5 27 8</td>
<td>40</td>
<td>13 15 7 1 36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Santa Fe</td>
<td>3 1 3 2 1 10</td>
<td>4 9 5 7 25</td>
<td>6 20 23 12</td>
<td>61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rafaela</td>
<td>1 2 3 1 7 2 2 7 4 2 17</td>
<td>13 27 23 9 72</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reconquista</td>
<td>1 2 3 6 3 5 4 12</td>
<td>1 11 7 19</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>8 15 8 8 6 45</td>
<td>17 46 26 16 6 111</td>
<td>20 56 63 49 16 204</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Considering strata 1 and 2 as low needs, stratum 3 as middle needs, and strata 4 and 5 as high needs, in the following plots we show the distribution of the populated areas in each size group and region. It can be seen that the needs increase from the south to the north of the province.

### 5. Conclusions

In this paper we considered robust PCA based on multivariate MM-estimators of shape. We also investigated the fast and robust bootstrap method to estimate confidence intervals for the proportion of variance explained of the components and to calculate confidence limits for the loads of the principal robust components. We developed an application of these methods to data of indicators of critical needs of populated areas in the province of Santa Fe from the National Census Population and Housing 2010, in order to achieve a stratification of the populated areas for a future sampling frame. The first robust component provided a summary of three indicators of social needs (education, basic services and housing spaces) into a single index that can be used in order to sort the observation units according to their social needs. We used the geometric method to stratify the populated areas since the distribution of the critical needs index is asymmetrical. It was observed that the stratum to which the populated areas belong is related to the region where they are located.
finding that the populated areas in the north of the province have a greater social need than the southern ones.

References


