# 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@fcecon.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.