Comparing Classical and Robust Sparse PCA

Valentin Todorov and Peter Filzmoser

Abstract. The main drawback of principal component analysis (PCA) especially for applications in high dimensions is that the extracted components are linear combinations of all input variables. To facilitate the interpretability of PCA various sparse methods have been proposed recently. However all these methods might suffer from the influence of outliers present in the data. An algorithm to compute sparse and robust PCA was recently proposed by Croux *et al.* We compare this method to standard (non-sparse) classical and robust PCA and several other sparse methods. The considered methods are illustrated on a real data example and compared in a simulation experiment. It is shown that the robust sparse method preserves the sparsity and at the same time provides protection against contamination.

Keywords: Principcal component analysis, robust statistics.

1 Introduction

Principal component analysis (PCA) is a widely used technique for dimension reduction achieved by finding a smaller number q of linear combinations of the originally observed p variables and retaining most of the variability of the data. It is important to be able to interpret these new variables, referred to

Valentin Todorov
United Nations Industrial Development Organization (UNIDO), Vienna, Austria
e-mail: v.todorov@unido.org

Peter Filzmoser
Department of Statistics and Probability Theory,
Vienna University of Technology, Vienna, Austria
e-mail: p.filzmoser@tuwien.ac.at

R. Kruse et al. (Eds.): Synergies of Soft Computing and Statistics, AISC 190, pp. 283–291.
springerlink.com

as *principal components*, especially when the original variables have physical meaning. The link between the original variables and the principal components is given by the so called *loadings matrix* used for transforming the data and thus it should serve as a means for interpreting the PCs. However, PCA usually tends to provide PCs which are linear combinations of all the original variables (by giving them non-zero loadings). Regarding the interpretability of the results it would be very helpful to reduce not only the dimensionality but also the number of used variables (ideally to relate each PC to only a few variables). It is not surprising that vast research effort was devoted to this issue and various proposals have been introduced in the literature. A straightforward informal method is to set to zeros those PC loadings which have absolute values below a given threshold (*simple thresholding*). In [6] SCoTLASS was proposed which applies a *lasso* penalty on the loadings in a PCA optimization problem. Recently a reformulated PCA as a regression problem has been proposed [13] that uses the *elastic net* to obtain a sparse version (SPCA).

Despite more or less successful in achieving sparsity, all these methods suffer a common drawback - all are based on the classical approach to PCA which measures the variability through the empirical variance and is essentially based on computation of eigenvalues and eigenvectors of the sample covariance or correlation matrix. Therefore the results may be very sensitive to the presence of even a few atypical observations in the data. The outliers could artificially increase the variance in an otherwise uninformative direction and this direction will be determined as a PC direction. To cope with the possible presence of outliers in the data, recently a method has been proposed [1] which is sparse and robust at the same time. It utilizes the *projection pursuit* approach where the PCs are extracted from the data by searching the directions that maximize a robust measure of variance of data projected on it. An efficient computational algorithm was proposed in [2]. Another robust sparse PCA algorithm was proposed by [9] maximizing the L1-norm variance instead of the classical variance but unfortunately no R implementation was available and in the short time we could not include it in the comparison.

The paper [13] defined the (minimal) requirements for a good sparse method as follows: (i) without any penalty constraint the method is equivalent to standard PCA; (ii) the method is computationally efficient for both large n and large p and (iii) it avoids misidentifying important variables. To these requirements we will add one more: (iv) the method should attain the properties (i) to (iii) even in the presence of outliers in the data.

The remainder of the paper is organized as follows. Section 2 presents briefly the sparse and robust methods considered. Section 3 illustrates these methods on real data examples and Section 4 compares them on simulated data sets. The final Section 5 concludes.