Robust Principal Component Analysis: a Review

Subhabrata Majumdar, Snigdhansu Chatterjee

Abstract:

Keywords: Data depth; Multivariate ranking; Principal components analysis; robust statistics; functional data; dimension reduction

1   Introduction

Principal component Analysis (PCA) is one of the oldest, yet most widely used methods of unsupervised multivariate analysis. For a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ containing observations in p variables for n samples, each column having mean 0, principal components are defined as p-dimensional vectors $\mathbf{w}_k, 1 \leq k \leq p$ such that

$$\mathbf{w}_k = \begin{cases} \arg\max_{\|\mathbf{w}\|=1} \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} \text{ if } k = 1 \\ \arg\max_{\|\mathbf{w}\|=1} \mathbf{w}^T \mathbf{R}_k^T \mathbf{R}_k \mathbf{w} \text{ if } k > 1; \quad \mathbf{R}_k = \mathbf{X} - \sum_{s=1}^{k-1} \mathbf{X} \mathbf{w}_s \mathbf{w}_s^T \end{cases} \tag{1.1}$$

Following a lagrange multiplier approach, the eigenvectors of $\mathbf{X}^T \mathbf{X}$, equivalently the right singular vectors obtained from the singular value decomposition of $\mathbf{X}$ provide solutions to (1.1).

more stuff

robustness towards outliers

robustness towards corrupted entries

combine?

2   Robust covariance estimation, data transformation, and beyond

2.1   Robust covariance matrices, projection pursuit

The earliest approaches to robust PCA were based on robustly estimating the population covariance matrix, and using eigenvectors of that estimate as principal components. Some methods of robust covariance matrix estimation include the Minimum Volume Ellipsoid estimator (Rousseeuw, 1984), a projection-based estimator by (Maronna et al., 1976), the Minimum Covariance Determinant (MCD) estimator (Rousseeuw, 1985) and the Stael-Donoho estimators (Maronna and Yohai, 1995; Zuo and Cui, 2005). Although these estimators have high breakdown points, they suffered from two severe drawbacks. Firstly the explicit evaluation of the population covariance matrix meant that obtaining principal components were not possible when $n < p$. Secondly, even when $n > p$, these methods become computationally intensive with large data dimensions.

2.2 Data transformation

2.3 Robust PCA and outlier detection

Aside from obtaining a lower dimensional projection of the data matrix $\mathbf{X}$ in spite of outliers that is close enough to the projection of $\mathbf{X}$ by the first few population eigenvectors, detecting the outliers themselves is also closely associated with robust PCA. These samples can be of interest for mechanistic reasons. For example in the analysis of near infra-red absorbance for 39 gasoline samples over 226 wavelengths using ROBPCA (Hubert et al., 2005), six compounds are flagged as outliers, and these turn out to be the only samples containing alcohol. Hubert et al. (2005) also introduced a notion of outlier diagnostics that is applicable to any method of robust PCA and can serve as a means to compare different relevant techniques as well.

We illustrate this in **??**. Here we consider data in 3 dimensions, and consider the relative position of the samples with respect to the two-dimensional principal component subspace $\mathcal{M}$. We can classify such points into four categories:

1. *Regular observations*: points that form a homogeneous group close to $\mathcal{M}$ (A and B in figure);

2. *Good leverage points*: points that lie close to $\mathcal{M}$, but at a distance from the regular observations (C in figure);

3. *Orthogonal outliers*: These points (point D in figure) lie far away from their projections on $\mathcal{M}$ (point D', but the projections themselves are close to the regular observations;

4. *Bad leverage points*: These points are also far away from their projections on $\mathcal{M}$ (E and E' respectively), but the projections are also far away from the regular observations.

Hubert et al. (2005) introduced the concept of *score distance* (SD) and *orthogonal distance* (OD) to distinguish between these four types of points. With our notation, for the ith observation these distances are defined as:

$$SD_i = \sum_{j=1}^{q} \frac{t_{ij}}{\lambda_j}; \quad OD_i = \|(\mathbf{I} - \mathbf{W}_k \mathbf{W}_k^T)(\mathbf{x}_i - \boldsymbol{\mu})\|$$

The SD can be interpreted as the weighted distance of the projection of a point on the hyperplane formed by the first k PCs, while OD is the orthogonal distance of that point and the k-PC hyperplane.

It is now clear from our picture that regular observations have low values of both SD and OD, while bad leverage points have high values of both. An orthogonal outlier has small SD but large OD, whereas a good leverage point has high SD but small OD. To explicitly classify sample points into these 4 categories, Hubert et al. (2005) use $\sqrt{\chi^2_{k,0.975}}$ and $[\hat{\mu}(OD^{2/3}) + \hat{\sigma}(OD^{2/3})\Phi^{-1}(0.975)]^{3/2}$ as upper cutoffs for score distance and orthogonal distance, respectively. Here $\hat{\mu}$ and $\hat{\sigma}$ are univariate MCD estimators, and Φ is the standard normal cumulative distribution function.

3   Principal Component Pursuit

4   Numerical examples

5   Robust PCA in other spaces

5.1   Kernel PCA

5.2   Functional PCA

6   Conclusion

References

M. Hubert, P. J. Rousseeuw, and K. V. Branden. ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics*, 47-1:64–79, 2005.

R. A. Maronna, W. A. Staehl, and V. J. Yohai. Bias-Robust Estimators of Multivariate Scatter Based on Projections. *J. Mult. Anal.*, 42:141–161, 1976.

R.A. Maronna and V.J. Yohai. The behavior of the Stahel-Donoho Robust Multivariate Estimator. *J. Amer. Statist. Assoc.*, 90:329–341, 1995.

P. Rousseeuw. Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications, Volume B (Proc. 4th Pannonian Symp. Math. Statist., Bad Tatzmannsdorf, 1983)*, pages 283–97, Dordrecht, 1985. D. Reidel.

P. J. Rousseeuw. Least Median of Squares Regression. *J. Amer. Statist. Assoc.*, 79:871–880, 1984.

Y. Zuo and M. Cui. Depth weighted scatter estimators. *Ann. Statist.*, 33-1:381–413, 2005.