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.

Li and Chen (1985) first introduced the idea of Projection Pursuit (PP) in robust PCA to alleviate these problems. Notice that the case for $k = 1$ in (1.1) can be rewritten as

$$\mathbf{w}_1 = \arg\max_{\|\mathbf{w}\|=1} Var(\mathbf{X}\mathbf{w})$$

The proposal of Li and Chen (1985) was to simply use a robust univariate scale estimator s_n (e.g. median, MCD) to obtain the robust PCs from a size n sample $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$:

$$\hat{\mathbf{w}}_1^{\text{PP}} = \arg\max_{\|\mathbf{w}\|=1} s_n(\mathbf{w}^T\mathbf{x}_1, \ldots, \mathbf{w}^T\mathbf{x}_n);$$

$$\hat{\mathbf{w}}_k^{\text{PP}} = \arg\max_{\|\mathbf{w}\|=1; \mathbf{w}\perp\mathbf{w}_s, s<k} s_n(\mathbf{w}^T\mathbf{x}_1, \ldots, \mathbf{w}^T\mathbf{x}_n) \quad \text{for } 1 < k \leq p$$

Aside from not having any restrictions for high-dimensional data, the PP approach allowed the flexibility of using any robust univariate scale estimator, and sequential estimation of the principal components. PP-based robust PCA became a popular method for chemometric data analysis in the 1990-s and early 2000-s, mainly due to the algorithmic developments by Xie et al. (1993), Hubert et al. (2002) and Croux et al. (2007).

The ROBPCA method of Hubert et al. (2005) combines the above two approaches. Specifically, ROBPCA consists of the following steps:

- Do an initial dimension reduction of the data matrix: $\mathbf{X}_{n\times p} \mapsto \mathbf{Z}_{n\times r}, r \leq p$ by projecting all data points on the subspace formed by the right singular vectors of $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}_{n\times r}\mathbf{D}_{r\times r}\mathbf{V}_{r\times p}; \quad \mathbf{Z} = \mathbf{U}\mathbf{D}$$

- Calculate the outlyingness of all samples:

$$O(\mathbf{z}_i) = \max_{\mathbf{v}\in B} \frac{|\mathbf{z}_i^T\mathbf{v} - m_n(\mathbf{z}_j^T\mathbf{v})|}{s_n(\mathbf{z}_j^T\mathbf{v})}$$

  where m_n and s_n are robust location and scale estimators, respectively. When $\binom{n}{2} < 250$, B is the set of vectors passing through all pairs of sample points, and is a collection of 250 randomly chosen non-zero vectors in $\mathbb{R}^r$ otherwise.

  Use the top k PCs calculated from the h least outlying points ($k \leq r; k, h$ suitably chosen) to

transform the data again:

$$\mathbf{Z}^* = \mathbf{Z}\mathbf{P}_0; \quad \mathbf{P}_0 \in \mathbb{R}_{r\times k}$$

- Robustly estimate the scatter matrix of $\mathbf{Z}^*$, take its eigenvectors as the estimated robust PCs.

As Hubert et al. (2005) showed through application on simulated and real data, the combination of a PP approach (first step) and robust scatter matrix estimation (third step) used by ROBPCA results in efficiency gains in estimating population principal components, as well as better detection of outlying points, compared to either of the previous types of methods for robust PCA.

2.2 Data transformation and M-estimation

A parallel approach towards robust PCA has also been developed by researchers, that is focused on the usage of robust transformations on the data, specifically multivariate signs and ranks, and related M-estimates of scatter. First introduced by Möttönen and Oja (1995), the multivariate sign or *spatial sign* of a vector $\mathbf{x} \in \mathbb{R}^p$ with respect to a location parameter $\boldsymbol{\mu} \in \mathbb{R}^p$ is defined as:

$$\mathbf{S}(\mathbf{x}) = \frac{\mathbf{x}-\boldsymbol{\mu}}{\|\mathbf{x}-\boldsymbol{\mu}\|}\mathbb{I}_{\mathbf{x}\neq\boldsymbol{\mu}}$$

When $\mathbf{x}$ is a random sample from an elliptical distribution, the sign transformation keeps the population eigenvectors constant. Since all vectors in the same direction get mapped to the same spatial sign regardless of their magnitude, eigenvector estimates calculated from the Sign Covariance Matrix (SCM), or equivalently the SVD of a sign transformed data matrix can act as robust PCs (Locantore et al., 1999; Visuri et al., 2000).

There are two components of a multivariate data point: its direction and magnitude. Spatial sign discards the magnitude and only uses the direction. Consequently, although the sign transformation provides an intuitively simple way of robustly estimating population eigenvectors, the estimates are not very accurate, in terms of asymptotic and finite-sample efficiencies (Majumdar and Chatterjee, 2015). In fact, Magyar and Tyler (2014) showed that the eigenvectors of the M-estimate of scatter proposed by Tyler (1987) has uniformly lower asymptotic risk than those obtained from the SCM.

Majumdar and Chatterjee (2015) rectified this by weighting the spatial signs by a bounded distance measure from the origin. They used data depth (Zuo and Serfling, 2000) to construct these weight functions. For some $\mathbf{y} \in \mathbb{R}^p$ and a set of points in $\mathbb{R}^p$, say $(\mathbf{y}_1, \ldots, \mathbf{y}_n)^T = \mathbf{Y}$, data depth (denoted

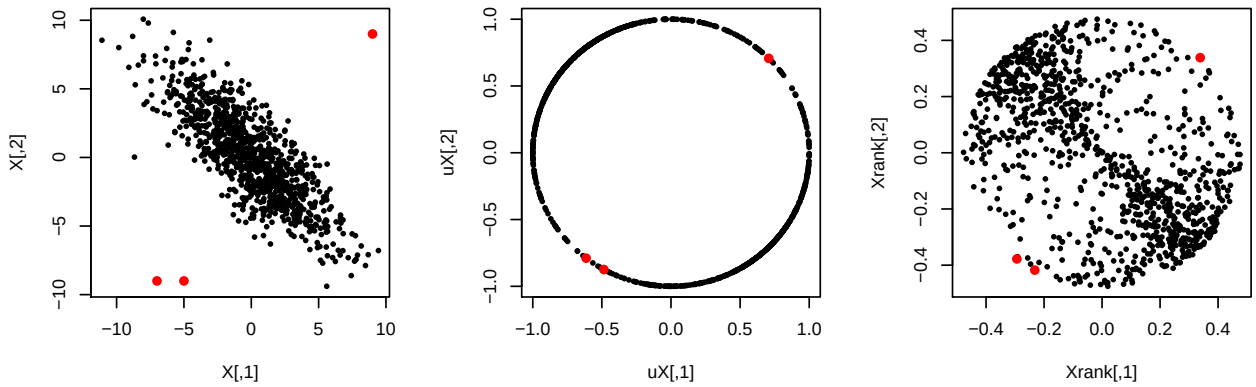

Figure 1: allthree

by $D(\mathbf{y}, \mathbf{Y})$) provides an affine invariant scalar measure of how close $\mathbf{y}$ is to the data cloud. The depth-based weighted spatial signsMajumdar and Chatterjee (2015) are explicitly constructed as:

$$\tilde{\mathbf{x}} = \left[\sup_{\mathbf{z}} D(\mathbf{z}, \mathbf{X}) - D(\mathbf{x}, \mathbf{X})\right] \mathbf{S}(\mathbf{x} - \bar{\mathbf{X}}) \tag{2.1}$$

The transformation $\mathbf{x} \mapsto \tilde{\mathbf{x}}$ preserves the magnitude information of a point: points with the same direction but different magnitudes get mapped further from the origin as the magnitude increases. However, due to the boundedness of data depth, this mapping limits the maximum distance an outlying point can get mapped to (see figure 1). Thus the weighted sign transformation improves upon the sign-based PCA in terms of lower estimation errors for elliptic underlying distributions, while still preserving robustness properties like high breakdown points and bounded influence functions (Majumdar and Chatterjee, 2015).

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

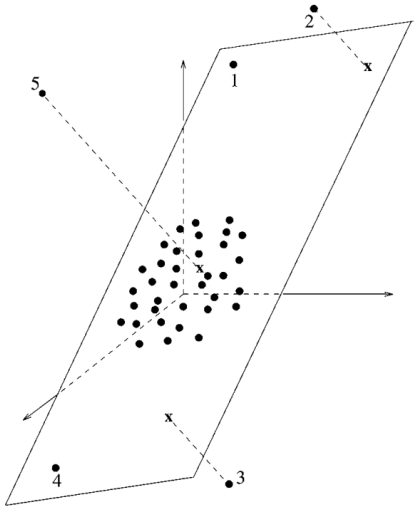

Figure 2: allthree

compare different relevant techniques as well.

We illustrate this in 2. 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 i^{th} observation these distances are defined as:

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

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

The above notion of outliers depends on the fact that the $n \times p$ data matrix $\mathbf{X}$ is composed of observations from several independent samples in its rows, and some of these samples have corrupted observations. However, in many practical situations, rows of $\mathbf{X}$ might not be independent, the corrupted observations can have a pattern across samples, or both. For example in face or handwriting recognition, each individual picture can be taken as a data matrix. The value a pixel takes corresponds to an entry in the data matrix, with noisy pixels denoting corrupted observations. Although the underlying low-rank structure is still of interest in such situations, for example the face of a person or a handwritten digit, the problem is fundamentally different because of the inherent structure present in the data.

Candés et al. (2009) first introduced *Principal Component Pursuit* (PCP), which decomposes the data matrix into low-rank and sparse components to tackle this situation. Formally, PCP considers the following additive model:

$$\mathbf{X} = \mathbf{L}_0 + \mathbf{S}_0 \tag{3.1}$$

with $\text{rank}(\mathbf{L}) = r < p$ and $\mathbf{S}$ sparse. The low-rank and sparse structures are recovered using nuclear norm penalization on the first component and ℓ_1-norm penalization on the second component, respectively:

$$\text{minimize } \|\mathbf{L}\|_* + \|\mathbf{S}\|_1; \quad \text{subject to } \mathbf{L} + \mathbf{S} = \mathbf{X} \tag{3.2}$$

where $\|.\|_*$ denotes the nuclear norm of a matrix, i.e. sum of its singular values, and $\|.\|_1$ denotes ℓ_1-norm, i.e. sum of the absolute values of its entries. Candés et al. (2009) proved that given the true underlying structure is indeed low-rank-plus-sparse, i.e. adheres to the decomposition in (3.1), a polynomial time algorithm based on convex programming can recover these matrices, and this is possible for arbitrary magnitudes of entries in the sparse component.

3.1 PCP and matrix completion

Both the polynomial time algorithm and arbitrary magnitude of corrupted entries are strengths of PCP over traditional methods of robust PCA. Another reason the PCP is attractive by itself is because with slight modifications, it can perform robust matrix completion. The matrix completion problem attempts to fill in a data matrix $\mathbf{Y}$ when only a subset $\Omega \subset \{1, \ldots, n\} \times \{1, \ldots, p\}$ of its entries are observed through nuclear norm minimization. Formally stated, the problem amounts to

$$\text{minimize } \|\mathbf{L}\|_*; \quad \text{subject to } \mathbf{P}_\Omega \mathbf{L} = \mathbf{Y}$$

where $\mathbf{P}_\Omega$ is the known indicator matrix of non-missing entries: $(\mathbf{P}_\Omega)_{ij} = \mathbb{I}_{(i,j)\in\Omega}$. PCP simply assumes there is a sparse noise component in the incomplete data: $\mathbf{Y} = \mathbf{P}_\Omega(\mathbf{L}_0 + \mathbf{S}_0)$, and recovers the low-rank structure:

$$\text{minimize } \|\mathbf{L}\|_* + \|\mathbf{S}\|_1; \quad \text{subject to } \mathbf{P}_\Omega(\mathbf{L} + \mathbf{S}) = \mathbf{Y} \tag{3.3}$$

Candés et al. (2009) showed that it is possible to solve this problem with minimal modifications to their original PCP algorithm that solves (3.2). Multiple further studies provided improvements for several aspects of this basic setup. The work of Chen et al. (2011) is prominent among them. In particular, they assumed the presence of both errors and missing entries, with deterministic or random support for each of them, and provided theoretical performance guarantees when the fraction of observed entries vanishes as $n \to \infty$. They also performed worst-case analysis for the errors-only or missing-only scenarios.

3.2 Modifications

In a subsequent paper, Zhou et al. (2010) added an entrywise small noise component $\mathbf{Z}$ to the objective functions in (3.2):

$$\text{minimize } \|\mathbf{L}\|_* + \|\mathbf{S}\|_1; \quad \text{subject to } \mathbf{L} + \mathbf{Z} + \mathbf{S} = \mathbf{X} \tag{3.4}$$

and (3.3):

$$\text{minimize } \|\mathbf{L}\|_* + \|\mathbf{S}\|_1; \quad \text{subject to } \mathbf{P}_\Omega(\mathbf{L} + \mathbf{Z} + \mathbf{S}) = \mathbf{Y} \tag{3.5}$$

This brought the PCP formulation closer to the classical robust PCA setup that separates a lower-dimensional component in presence of both data-wide additional noise and corrupted entries, but in this formulation the magnitude and structure of corrupted entries can be arbitrary. Further modifications of PCP include the case when the lower-dimensional component is a union of multiple lower dimensional subspaces (Wohlberg et al., 2012), adding an ℓ_1/ℓ_2-penalization term on $\mathbf{L}$ (Tang and Nehorai, 2011), a dual formulation of the problem (Becker et al., 2011), and non-convex robust matrix completion (Shang et al., 2014). PCP has been an active area of research in the signal and image processing community for the past few years. Please refer to Bouwmans and Zahzah (2014) for a detailed review on further modifications of the method, algorithmic developments, and its applications in video surveillance.

4 Numerical examples

In this section we present two data analytical scenarios to demonstrate the comparative performance and applicability of the different approaches of robust PCA we have discussed.

4.1 Bus data

Available in the R package `rrcov`, this dataset consists of the measurements of 18 image features for 218 buses. Following a similar analysis in Maronna et al. (2006), pp. 213, we set aside variable 9 and scale the other variables by dividing with their respective median absolute deviations (MAD). We do this done because all the variables had much larger standard deviations compared to their MADs, and variable 9 had MAD = 0. Following this, we compare the performances of the classical PCA

Quantile	Method of PCA					
	CPCA	SPCA	ROBPCA	MPCA	DPCA	PCP
10%	1.9	1.2	1.2	1.0	1.2	1.3
20%	2.3	1.6	1.6	1.3	1.6	1.8
30%	2.8	1.8	1.8	1.7	1.9	2.1
40%	3.2	2.2	2.1	2.1	2.3	2.5
50%	3.7	2.6	2.5	3.1	2.6	3.2
60%	4.4	3.1	3.0	5.9	3.2	3.8
70%	5.4	3.8	3.9	25.1	3.9	5.7
80%	6.5	5.2	4.8	86.1	4.8	11.9
90%	8.2	9.0	10.9	298.2	6.9	80.2
Max	24	1037	1055	1037	980	1157

Table 1: Quantiles of squared data-to-projection distances for bus data

(CPCA), PCA based on the eigenvector estimate from the MCD covariance matrix (MPCA), spatial sign-based PCA (SPCA), depth-based weighted sign PCA (DPCA), ROBPCA, and PCP. For the sake of uniformity, we use projection depth as our fixed depth function while doing DPCA.

For all classical PCA methods, we set the number of PCs at 3. We compare the above methods using the distance of actual data and their projections on the principal component space. For PCP, these are is simply row norms of the sparse matrix $\mathbf{S}_0$ obtained from the procedure, while for other methods this is the orthogonal distances of corresponding samples. Each of its column lists different quantiles of the squared orthogonal distance for a sample point from the hyperplane formed by top 3 PCs estimated by the corresponding method. Table 1 presents the different quantiles of squares of these distances for all the methods. For DPCA, the estimated principal component subspaces are closer to the data than CPCA for more than 90% of samples, and the distance only becomes larger for higher quantiles. This means that for CPCA, estimated basis vectors of the hyperspace get pulled by extreme outlying points, while the influence of these outliers is very low for DPCA. SPCA and ROBPCA perform very closely in this respect, the percentage of points that have less squared distance than CPCA being between 80% and 90% for both of them. This percentage is only 60% for PCP and 50% for MPCA, which suggests that the corresponding 3-dimensional subspace estimated by MCD is possibly not an accurate representation of the truth, also there is probably enough noise in the data apart from the low-rank and sparse components estimated by PCP.

4.2 Image denoising

5 Robust PCA in other spaces

5.1 Kernel PCA

5.2 Functional PCA

6 Conclusion

References

S. Becker, E. J. Candés, and M. Grant. TFOCS: flexible first-order methods for rank minimization. In *Low-rank Matrix Optimization Symposium, SIAM Conference on Optimization*, 2011.

T. Bouwmans and E. H. Zahzah. Robust PCA via Principal Component Pursuit: A review for a comparative evaluation in video surveillance. *Comput. Vis. Image Underst.*, 122:22–34, 2014.

E. J. Candés, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58:11, 2009.

Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis. Low-rank matrix recovery from errors and erasures. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2313–2317, 2011.

C. Croux, P. Flizmoser, and M. R. Oliviera. Algorithms for Projection-Pursuit robust principal component analysis. *Chemom. Intell. Lab. Syst.*, 87:218–225, 2007.

M. Hubert, P. J. Rousseeuw, and S. Verboven. A Fast Method for Robust Principal Components With Applications to Chemometrics. *Chemom. Intell. Lab. Syst.*, 60:101–111, 2002.

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

G. Li and Z. Chen. Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo. *J. Amer. Statist. Assoc.*, 80(759–766), 1985.

N. Locantore, J.S. Marron, D.G. Simpson, N. Tripoli, J.T. Zhang, and K.L. Cohen. Robust principal components of functional data. *TEST*, 8:1–73, 1999.

A.F. Magyar and D.E. Tyler. The asymptotic inadmissibility of the spatial sign covariance matrix for elliptically symmetric distributions. *Biometrika*, 101:673–688, 2014.

S. Majumdar and S. Chatterjee. Robust estimation of principal components from depth-based multivariate rank covariance matrix. `http://arxiv.org/abs/1502.07042`, 2015.

R. Maronna, D. Martin, and V. Y. Yohai. *Robust Statistics: Theory and Methods*. Wiley, New York, NY, 2006.

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.

J. Möttönen and H. Oja. Multivariate spatial sign and rank methods. *J. Nonparametric Stat.*, 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.

F. Shang, Y. Liu, J. Cheng, and H. Cheng. Robust Principal Component Analysis with Missing Data. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1149–1158. ACM, 2014.

G. Tang and A. Nehorai. Robust principal component analysis based on low-rank and block-sparse matrix decomposition. In *2011 45th Annual Conference on Information Sciences and Systems*, pages 1–5, 2011.

D.E. Tyler. A distribution-free M-estimator of multivariate scatter. *Ann. Statist.*, 15:234–251, 1987.

S. Visuri, V. Koivunen, and H. Oja. Sign and rank covariance matrices. *J. Statist. Plan. Inf.*, 91: 557–575, 2000.

B. Wohlberg, R. Chartrand, and J. Theiler. Local principal component pursuit for nonlinear datasets. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3925–3928, 2012.

Y.-L. Xie, J.-H. Wang, Y.-Z. Liang, L.-X. Sun, X.-H. Song, and R.-Q. Yu. Robust principal component analysis by projection pursuit. *J. Chemom.*, 7:527–541, 1993.

Z. Zhou, X. Li, J. Wright, E. J. Candés, and Y. Ma. Stable principal component pursuit. In *2010 IEEE International Symposium on Information Theory Proceedings*, pages 1518–1522, 2010.

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

Y. Zuo and R. Serfling. General notions of statistical depth functions. *Ann. Statist.*, 28-2:461–482, 2000.