Robust Principal Component Analysis: a Review

Subhabrata Majumdar [*], Snigdhansu Chatterjee [†]

November 4, 2017

Article Type:

Advanced Review

Abstract

Principal Component Analysis (PCA) is widely used many scientific domains. Accurately estimating the underlying low-rank structure in a data matrix in presence of corrupted entries is a problem that has achieved considerable attention in statistical literature. In this paper we review techinques that deal with this robust estimation problem. These span statistical methods useful for accurate estimation of principal components in presence of outlying samples, as well as the recently proposed Principal Component Pursuit approach that is effective when the data matrix contains sparse noise. We summarize the research in these two domains, and present data examples for their comparative evaluation. Finally, we also review methods that perform robust versions of kernel PCA and functional PCA.

Keywords

Principal Component Analysis; Robustness; ROBPCA; Spatial signs; data depth; Principal Component Pursuit

[*]University of Florida Informatics Institute, 432 Newell Drive, CISE Bldg E251, Gainesville, FL 32611
[†]School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church St SE, Minneapolis, MN 55455

INTRODUCTION

Principal component Analysis (PCA) is one of the oldest, yet most widely used methods of unsupervised multivariate analysis. Given a p-dimensional random variable $\mathbb{X}$ with mean vector $\mathbf{0}_p$ and covariance matrix $\mathbf{\Sigma}$, the principal component transformation is defined as:

$$\mathbb{X} \mapsto \mathbb{Y} = \mathbf{\Gamma}^T \mathbb{X} \tag{1}$$

where $\mathbf{\Gamma}$ is orthogonal, and $\mathbf{\Gamma}^T \mathbf{\Sigma} \mathbf{\Gamma} = \mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p); \lambda_1 \geq \ldots \geq \lambda_p \geq 0$. This induces a set of linear transformations on the random vector $\mathbb{X}$ so that the transformed random variables are uncorrelated with each other, and their variances are ordered from highest to smallest (Mardia et al., 1979). When $\mathbf{\Sigma}$ is positive definite, coefficients for these transformations, i.e. the columns of $\mathbf{\Gamma}$, are given by the eigenvectors of $\mathbf{\Sigma}$ following its spectral decomposition. For a size-n sample from the distribution of $\mathbb{X}$, say $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, the first principal component gives the linear combination of columns of $\mathbf{X}$ which maximizes sample variance:

$$\mathbf{w}_1 = \underset{\|\mathbf{w}\|=1}{\arg\max}\, Var(\mathbf{X}\mathbf{w}) = \underset{\|\mathbf{w}\|=1}{\arg\max}\, \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} \tag{2}$$

and the subsequent principal components are defined as

$$\mathbf{w}_k = \underset{\|\mathbf{w}\|=1}{\arg\max}\, \mathbf{w}^T \mathbf{R}_k^T \mathbf{R}_k \mathbf{w}; \quad \mathbf{R}_k = \mathbf{X} - \sum_{s=1}^{k-1} \mathbf{X} \mathbf{w}_s \mathbf{w}_s^T \quad \text{for } 1 < k \leq p \tag{3}$$

Following a lagrange multiplier approach, the eigenvectors of $\mathbf{X}^T\mathbf{X}/n$, equivalently the right singular vectors obtained from the singular value decomposition of $\mathbf{X}$ provide solutions to (2) and (3).

PCA has seen extensive applications in diverse areas, such as image recognition (Alkandari and Aljaber, 2015), finance (Alexander, 2008), climate science (Wilks, 2011) and text mining (Berry and Castellanos, 2007), with the inferential goal being reduction of the intrinsic dimensionality of the feature space without losing information. However, because the objective function to be maximized is quadratic, PCA performs poorly in presence of even a small proportion of corrupted observations (Xu et al., 2013). Based on the domain of application, these corruptions can be the result of data heterogeneity (Saha et al., 2016),

measurement error (Bailey, 2012; Hellton and Thoresen, 2014), or may represent structured noise (Candés et al., 2009).

Depending on the modelling goals, robust PCA aims to estimate principal components or the underlying low dimensional subspace in presence of corrupted entries in the data matrix. Historically, the instrumental factors behind the evolution of robust PCA methods have been the size and complexity of datasets, availability of computational resources, as well as the nature of corruptions present in the data. Early methods of robust PCA were focused on robustly estimating the population covariance matrix from datasets that are small to moderate in size, and comprised of independent samples: some of which were outliers, i.e. contained corrupted entries. Later on, computational and statistical challenges that surfaced with the advent of high-dimensional datasets having a large number of features were tackled by methods like projection pursuit (Li and Chen, 1985), ROBPCA (Hubert et al., 2005) and M-estimation (Locantore et al., 1999; Majumdar and Chatterjee, 2015).

The theoretical discussion in this review is composed of two sections, and we discuss the above broad approaches of robust PCA on independent data in further detail in the first of those sections. We devote the other section to Principal Component Pursuit (PCP), which, even though introduced very recently (Candés et al., 2009), has motivated a substantial amount of research on the problem of recovering an underlying low-rank structure in the data matrix, rather than the principal components *per se*, in presence of noise. We illustrate the relevance and relative performance of these two types of methods using two real data examples. The first dataset is available in the R package `rrcov`, and consists of the measurements of 18 image features for 218 buses. In the second example we take the pixel matrices from four image files: the Lenna image, and three images from the extended Yale Face Database B (Georghiades et al., 2001; Lee et al., 2005) (fifth images for individuals 1, 2 and 28), add noise to some pixels and attempt to recover the original images using robust PCA techniques. The original images and those with added noise are given in Figure 1. Following this we review the methods of robust PCA in domains that are not multivariate real, for example reproducing kernel hilbert spaces or the space of square-integrable functions, in the section *Robust PCA in Other Spaces*. We finish the review with a concluding section that summarizes the paper and identifies focus areas for future research.

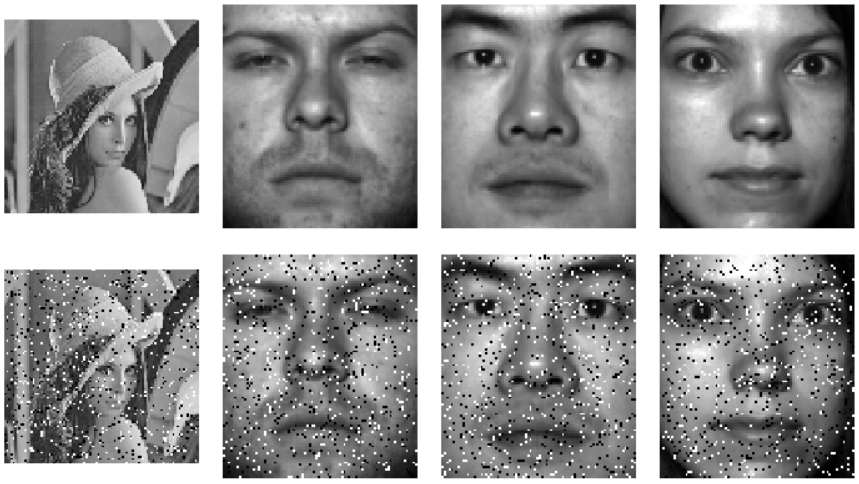

Figure 1: Original and noisy versions of the Lenna image (left) and Yale face images

ROBUST PCA WITH INDEPENDENT SAMPLES

Robust covariance matrices, projection pursuit

The early and by now well-established 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 a projection-based estimator by Maronna et al. (1976), the Minimum Volume Ellipsoid estimator (Rousseeuw, 1984), the Minimum Covariance Determinant (MCD) estimator (Rousseeuw, 1985) and the Stael-Donoho estimators (Maronna and Yohai, 1995; Zuo and Cui, 2005). Although these methods have high breakdown points, they are not entirely suitable for many modern applications where one or more of n and p can be large, and $n < p$ is not uncommon. They typically require computation of the entire covariance matrix, which is not possible when $n < p$. Even when $n > p$, these methods become computationally intensive with large data dimensions.

Li and Chen (1985) introduced the idea of Projection Pursuit (PP) in robust PCA to alleviate these problems. They proposed to replace the variances in (2) and (3) by a robust

univariate scale estimator s_n (e.g. median, MCD), and obtain the robust PCs subsequently:

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

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

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:

1. 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 $\mathbf{X}$:

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

2. 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. The set of vectors B is taken as the vectors passing through all possible pairs of sample points if $\binom{n}{2} < 250$, and 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:

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

3. 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 classical PCA, one of their previously proposed methods (Hubert et al., 2002), as well as spherical and ellipsoidal PCA (Locantore et al., 1999).

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. 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:

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

where $\mathbb{I}(.)$ is the 0/1 indicator function. 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) have 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

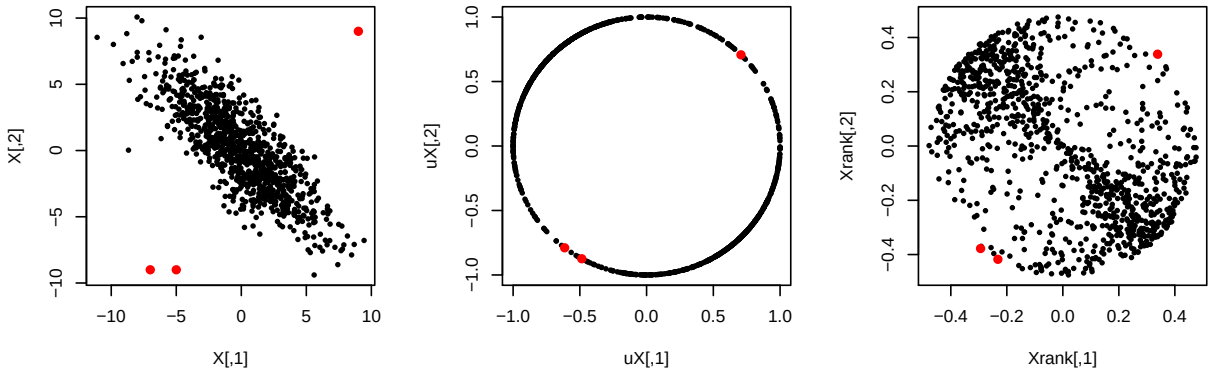

Figure 2: (Left) 1000 points randomly drawn from $\mathcal{N}_2\left((0,0)^T, \left(\begin{smallmatrix} 5 & -4 \\ -4 & 5 \end{smallmatrix}\right)\right)$ (in black), and three outliers (in red);
(Center) their spatial signs: all points reside on the surface of a finite ball in $\mathbb{R}^2$;
(Right) their multivariate ranks based on halfspace depth: retains the shape of the data.

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 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 signs (Majumdar and Chatterjee, 2015) are explicitly constructed as:

$$\tilde{\mathbf{x}} = \left[\sup_{\mathbf{z}} D(\mathbf{z}, \mathbf{X}) - D(\mathbf{x}, \mathbf{X})\right] \mathbf{S}(\mathbf{x} - \bar{\mathbf{X}}) \tag{4}$$

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 2). Consequently, 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).

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

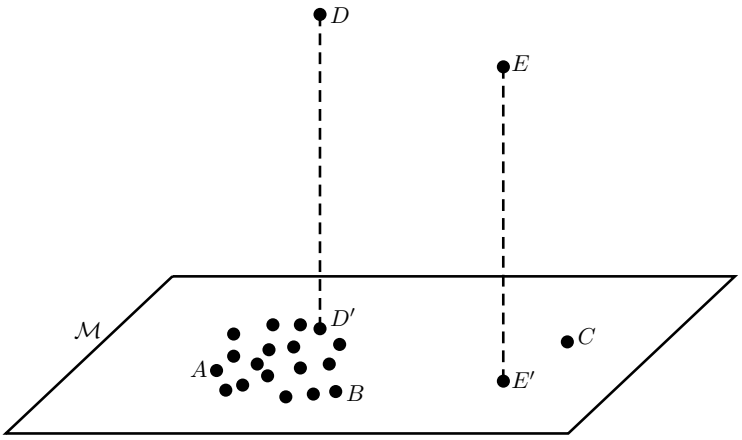

Figure 3: Four different types of points in 3-dimensional data with respect to a 2-PC subspace: A and B are regular observations, C is a good leverage point, D is an orthogonal outlier, and E is a bad leverage point.

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 (see Hubert et al. (2005)), six compounds are flagged as outliers, and these turn out to be the only samples containing alcohol. Over the past two decades, multiple approaches of using robust principal components for detecting anomalous samples have been proposed (Shyu et al., 2003; Jackson and Chen, 2004; Brown et al., 2010; Pascoal et al., 2010). In their 2005 paper, aside from the popular ROBPCA method 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 Figure 3. Here we consider data in 3 dimensions, and consider the relative position of 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 categories. 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.

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 of a pixel takes corresponds to an entry in the data matrix, with noisy pixels denoting corrupted measurements. Although the underlying low-rank structure is still of interest in such situations, for example the face of a person or a handwritten digit, this

problem is fundamentally different because of the inherent structure present in the data (Alkandari and Aljaber, 2015). On the other hand, data from video surveillance consists of moving objects in front of a background that is largely static across frames. In such a situation, accurately decomposing a frame in real-time into the low-rank background image and the localized foreground pixels is of practical interest (Bouwmans et al., 2014; Bouwmans and Zahzah, 2014).

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

$$\mathbf{X} = \mathbf{L}_0 + \mathbf{S}_0 \tag{5}$$

with $\text{rank}(\mathbf{L}_0) = r < p$ and $\mathbf{S}_0$ sparse. The low-rank and sparse structures are recovered using the following optimization setup:

$$\text{minimize } \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1; \quad \text{subject to } \mathbf{L} + \mathbf{S} = \mathbf{X} \tag{6}$$

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, and λ is a tuning parameter that determines the amount of sparsity permitted in $\mathbf{S}$. Candés et al. (2009) proved that given the true underlying structure is indeed low-rank-plus-sparse, i.e. adheres to the decomposition in (5), a polynomial time algorithm based on convex programming can exactly recover these matrices, and this is possible for arbitrary magnitudes of entries in the sparse component.

PCP and matrix completion

Both the polynomial time algorithm and the ability to handle corrupted entries of arbitrary magnitude 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. Matrix completion is the problem of filling in missing entries in a data matrix, and has several real-world applications: most prominently in recommender systems (Candes and Tao, 2010) and also in genomic data integration (Cai et al., 2016). When only

a subset $\Omega \subset \{1, \ldots, n\} \times \{1, \ldots, p\}$ of the entries in the data matrix $\mathbf{Y}$ are observed, the matrix completion algorithm seeks to find out a completed matrix through nuclear norm minimization. Formally stated, this 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 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:

$$\text{minimize } \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1; \quad \text{subject to } \mathbf{P}_\Omega(\mathbf{L} + \mathbf{S}) = \mathbf{Y} \tag{7}$$

Candés et al. (2009) showed that it is possible to solve this problem with minimal modifications to their original PCP algorithm that solves (6). Multiple further studies provided improvements on 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.

Modifications

In a paper subsequent to Candés et al. (2009), Zhou et al. (2010) added an entrywise noise component $\mathbf{Z}$ to the objective functions in (6):

$$\text{minimize } \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1; \quad \text{subject to } \mathbf{L} + \mathbf{Z} + \mathbf{S} = \mathbf{X} \tag{8}$$

and (7):

$$\text{minimize } \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1; \quad \text{subject to } \mathbf{P}_\Omega(\mathbf{L} + \mathbf{Z} + \mathbf{S}) = \mathbf{Y} \tag{9}$$

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, with the advantage that here the magnitude and structure of corrupted entries can be arbitrary. Further modifications of PCP include a dual formulation of the problem

(Becker et al., 2011), adding an ℓ_1/ℓ_2-penalization term on $\mathbf{L}$ (Tang and Nehorai, 2011), the case when the low dimension component is a union of multiple lower dimensional subspaces (Wohlberg et al., 2012), 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. Further details on modifications of the PCP, algorithmic developments, and its applications in video surveillance can be found in Bouwmans and Zahzah (2014).

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 discussed until now.

Bus data

Following the 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 (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. We use projection depth as our choice of 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

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

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, and there is probably enough noise in the data apart from the low-rank and sparse components estimated by PCP.

Image denoising

As mentioned before, a major area of application of robust PCA, and PCA in general, is to extract the underlying low-rank structure in image recognition problems that are often high-dimensional in nature. Although PCP was designed keeping this very generative model (i.e. (5)) in mind, Zhao et al. (2014) showed that even the classical PCA performs fairly

Figure 4: Denoising results of four images. Left to right in a row indicates the original image, image with noise added, and denoising by DPCA, ROBPCA and PCP, respectively.

well in comparison of low-rank-plus-sparse methods to remove certain types of noise from an image, as well as background subtraction of videos.

We first resize our images: the lenna image from 128×128 to 96×96 and the Yale images from 192×168 to 96×84. After this we randomly select 10% of the pixels from each image and turn their values to 0 or 1 with probability 0.5. Such noise occurs naturally in image recognition problems, and degrades image quality as well as the performance of image classification algorithms (Qiu et al., 2004). Following this we center and scale each of these matrices of pixel values and apply DPCA, ROBPCA and PCP on them. We take the top 10 estimated PCs for DPCA and ROBPCA to reconstruct the images. Figure 4 gives the results obtained from each of these methods. The first and second images in each row denote the

original image with and without speckle noise, while the others give denoised versions from the three methods. PCP has the best performance: it recovers the faces almost perfectly, and does better than the other two methods for the Lenna image. The structure in the data seem to have affected the performance of DPCA and ROBPCA, which retain some of the noise in the reconstructed versions. Using a larger number of PCs retains the noises as well. Finally, the performance of all methods suffer when they are applied on the Lenna image, which is a relatively complex image.

ROBUST PCA IN OTHER SPACES

Kernel PCA

Kernel PCA is prominent among nonlinear methods of dimension reduction, i.e. methods that aim to reduce the dimension of a nonlinear transformation of the data matrix $\mathbf{X}$, instead of a linear mapping of it. This is because even though kernel PCA performs linear PCA on the transformed feature space, it is able to do so without using the transformed random variables. Specifically, suppose the transformation map is $\phi : \mathbb{R}^p \mapsto \mathcal{H}$, a reproducing kernel Hilbert space (RKHS). Then there exists a *kernel function* $K : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$ that gives the inner product in the transformed space:

$$K(\mathbf{a}, \mathbf{b}) = \langle \phi(\mathbf{a}), \phi(\mathbf{b}) \rangle \tag{10}$$

This 'kernel trick' enables us to perform PCA on the transformed data $\{\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)\}$ just by replacing inner products with the kernel function in (10) (Schölkopf et al., 1999). As a consequence, it is possible to extend methods of linear PCA to their kernel versions. This holds for robust methods as well. Indeed, Debruyne and Verdonck (2010); Debruyne et al. (2010) provided algorithms that extend SPCA, projection pursuit and ROBPCA into kernel spaces, and Yang et al. (1999); Huang et al. (2009) and Huang and Reh (2011) used robust loss functions in the kernel version of a reformulation of the classical PCA problem in (2) and (3). Wang et al. (2007); Deng et al. (2007) and Pang et al. (2010) proposed more methods that connect robust classical PCA methods to kernel PCA.

Compared to robust linear PCA, robust kernel PCA has an additional step of reverse mapping the principal component projections on the feature space $\mathcal{H}$ to the input space $\mathbb{R}^p$. This is important because the areas of application for robust kernel PCA are concerned with the recognition of shapes that are mostly nonlinear in presense of additional noise: for example in computer vision (Lampart, 2008), environmental science (Hsieh, 2009), and neuroimaging (Mwangi et al., 2014). An exact pre-image, however, may not always exist (Mika et al., 1999), and obtaining an approximate reconstruction is challenging because $\mathcal{H}$ can be infinite dimensional. Mika et al. (1999) first provided a solution of this problem using an iterative algorithm. Later on, Kwok and Tsang (2004) provided another method of pre-image reconstruction using multidimensional scaling that results in better recovery in the original data space but does not suffer from the instability issues of Mika et al. (1999).

Functional PCA

For functional data, the matrix $\mathbf{X}$ shall correspond to the realizations of a random function, say f_X, that takes values in $L^2(\mathcal{I})$: $\mathcal{I}$ being a real interval. Following the definition of inner products in the functional space: $\langle a, b\rangle = \int_{\mathcal{I}} a(y)b(y)dy$ for $a, b \in L^2(\mathcal{I})$, one can replace the dot products in (2) and (3) by these inner products to define functional principal components. Analogous to the real setting, the covariance *operator* of f_X is defined as $\gamma_X = (f_X - \mathbb{E}f_X) \otimes (f_X - \mathbb{E}f_X); a \otimes b : \mathcal{H} \to \mathcal{H}$ so that $(a \otimes b)c = \langle b, c\rangle a$. Following this, f_X has the Karhunen-Loéven expansion:

$$f_X = \mathbb{E}f_X + \sum_{l=1}^{\infty} \lambda_l^{1/2} c_l \phi_l \tag{11}$$

where $\{\lambda_l, l \geq 1, \lambda_l \geq \lambda_{l+1}\}$ and $\{\phi_l, l \geq 1\}$ are eigenvalues and orthonormal eigenfunctions of γ_X, respectively, and $c_l = \lambda_l^{-1/2}\langle f_X - \mathbb{E}f_X, \phi_l\rangle$ are real-valued random variables. The top eigenfunctions are able to provide a finite dimensional approximation of f_X, and turn out to be its principal components.

It is possible to reduce the robust functional PCA problem to robust PCA on real domain by mapping the original data onto a finite set of orthogonal basis functions and working on the matrix of the corresponding coefficients. This approach was taken by Locantore et al. (1999) and Boente and Salibian-Barrera (2015). However, the smoothing approximations

can produce bias that can be avoided using a fully functional approach (Zhang and Chen, 2007). Similar to PCA on the real domain, this can be done by robustly estimating γ_X or another covariance operator, or directly estimating the top eigenfunctions in (11) in a robust manner. Gervini (2008) took the first approach by formulating a functional version of SPCA, while the work of Bali et al. (2011) that performs projection pursuit in the functional domain is prominent in the second category. Theoretical details of these methods, functional outlier detection methods, as well as their comparative performance, can be found in Bali and Boente (2014).

CONCLUSION

In the above sections we have reviewed several methods available in the literature concerned with recovery of an underlying low rank structure in the data matrix in presence of atypical data points. The data examples illustrate that robust PCA is not a 'one-method-fits-all' problem, and care should be exercised on what techinque should be applied on what data. Most of these literature is devoted towards robustness in presence of outlying samples. We believe robustness towards other factors, like model misspecification and missing data, needs to be further explored. In functional data, many of the methods proposed are robust against outlying curves that do not conform to the shape of the other curves, but not many that can detect outlying points in an otherwise typical curve. As mentioned in Bali and Boente (2014), this is a challenging problem and needs more attention.

References

Alexander, C. (2008). *Market Risk Analysis Volume III, Pricing, Hedging and Trading Financial Instruments*. Wiley.

Alkandari, A. and Aljaber, S. J. (2015). Principle Component Analysis algorithm (PCA) for image recognition. In *2015 Second International Conference on Computing Technology and Information Management (ICCTIM)*, pages 76–80.

Bailey, S. (2012). Principal Component Analysis with Noisy and/or Missing Data. *Publ. Astron. Soc. Pac.*, 124:1015–1023.

Bali, J. L. and Boente, G. (2014). Robust Functional Principal Component Analysis. In *New Advances in Statistical Modeling and Applications*, Studies in Theoretical and Applied Statistics, pages 41–54. Springer.

Bali, J. L., Boente, G., Tyler, D. E., and Wang, J. L. (2011). Robust functional principal components: a projection-pursuit approach. *Ann. Statist.*, 39:2852–2882.

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

Berry, M. W. and Castellanos, M. (2007). *Survey of Text Mining II: Clustering, Classification, and Retrieval*. Springer.

Boente, G. and Salibian-Barrera, M. (2015). S-Estimators for Functional Principal Component Analysis. *J. Amer. Statist. Assoc.*, 110:1100–1111.

Bouwmans, T., Porikli, F., Höferlin, B., and Vacavant, A. (2014). *Background Modeling and Foreground Detection for Video Surveillance*. CRC Press, Boca Raton, FL.

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

Brown, R. J. C., Goddaard, S. L., and Brown, A. S. (2010). Using principal component analysis to detect outliers in ambient air monitoring studies. *Int. J. Environ. An. Ch.*, 90:761–772.

Cai, T., Cai, T. T., and Zhang, A. (2016). Structured Matrix Completion with Applications to Genomic Data Integration. *J. Amer. Statist. Assoc.*, 111:621–633.

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

Candes, E. J. and Tao, T. (2010). The Power of Convex Relaxation: Near-Optimal Matrix Completion. *IEEE Trans. Inf. Theory*, 56:2053–2080.

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

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

Debruyne, M., Hubert, M., and Horebeek, J. V. (2010). Detecting Influential Observations in Kernel PCA. *Comput. Stat. Data Anal.*, 54.

Debruyne, M. and Verdonck, T. (2010). Robust kernel principal component analysis and classifcation. *Adv. Data Anal. Classif.*, 4:151–167.

Deng, X., Yuan, M., and Sudjianto, A. (2007). A note on robust kernel principal component analysis. *Contemp. Math.*, 443:21–34.

Georghiades, A., Belhumeur, P., and Kriegman, D. (2001). From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):643–660.

Gervini, D. (2008). Robust functional estimation using the spatial median and spherical principal components. *Biometrika*, 95:587–600.

Hellton, K. H. and Thoresen, M. (2014). The Impact of Measurement Error on Principal Component Analysis. *Scand. J. Stat.*, 41:1051–1063.

Hsieh, W. W. (2009). *Machine Learning Methods in the Environmental Sciences: Neural Networks and Kernels*. Cambridge University Press.

Huang, H.-H. and Reh, Y.-R. (2011). Neurocomputing. *An iterative algorithm for robust kernel principal component analysis*, 74:3921–3930.

Huang, S. Y., Yeh, Y. R., and Eguchi, S. (2009). Neural comput. *Robust kernel principal component analysis*, 21:3179–3213.

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

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

Jackson, D. A. and Chen, Y. (2004). Robust principal component analysis and outlier detection with ecological data. *Environmetrics*, 15:129–139.

Kwok, J. T. and Tsang, I. W. (2004). The Pre-Image Problem in Kernel Methods. *IEEE Trans. Neural Net.*, 15(6):1517–1525.

Lampart, C. H. (2008). Kernel Methods in Computer Vision. In Curless, B., Freeman, W. T., and Van Gool, L., editors, *Foundations and Trends in Computer Graphics and Vision*, volume 4 No. 3, pages 193–285. Now Publishers.

Lee, K., Ho, J., and Kriegman, D. (2005). Acquiring Linear Subspaces for Face Recognition under Variable Lighting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(5):684–698.

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

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

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

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

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press Inc, first edition.

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

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

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

Mika, S., Schölkopf, B., Smola, A., Müller, K., Scholz, M., and Rätsch, G. (1999). Kernel PCA and De-Noising in Feature Spaces. *Adv. Neural Inf. Process. Syst.*, pages 536–542.

Möttönen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *J. Nonparametric Stat.*, 5:201–213.

Mwangi, B., Tian, T. S., and Soares, J. C. (2014). A Review of Feature Reduction Techniques in Neuroimaging. *Neuroinformatics*, 12:229–244.

Pang, Y. W., Wang, L., and Yuan, Y. (2010). Generalized KPCA by adaptive rules in feature space. *Int. J. Comput. Math.*, 87:956–968.

Pascoal, C., Oliveira, M. R., Pacheco, A., and Valadas, R. (2010). Detection of outliers using robust principal component analysis: A simulation study. In *Combining Soft Computing and Statistical Methods in Data Analysis*, pages 499–507.

Qiu, F., Berglund, J., Jensen, J. R., Thakkar, P., and Ren, D. (2004). Speckle Noise Reduction in SAR Imagery Using a Local Adaptive Median Filter. *GISci. Remote Sens.*, 41(3):244–266.

Rousseeuw, P. (1985). 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. D. Reidel.

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

Saha, P., Roy, N., Mukherjee, D., and Sarkar, A. K. (2016). Application of Principal Component Analysis for Outlier Detection in Heterogeneous Traffic Data. *Procedia Comput. Sc.*, 83:107–114.

Schölkopf, B., Smola, A., and Müller, K.-R. (1999). *Kernel principal component analysis*, chapter Advances in Kernel Methods Support Vector Learning, pages 327–352. MIT Press.

Shang, F., Liu, Y., Cheng, J., and Cheng, H. (2014). 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.

Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K., and Chang, L. W. (2003). A Novel Anomaly Detection Scheme Based on Principal Component Classifier. In *ICDM Foundation and New Direction of Data Mining workshop*, pages 172–179.

Tang, G. and Nehorai, A. (2011). 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.

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

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

Wang, L., Pang, Y. W., Shen, D. Y., and Yu, N. H. (2007). An iterative algorithm for robust kernel principal component analysis. In *Int. Conf. Mach. Learn. Cybern. 2007*, volume 6, pages 3484–3489.

Wilks, D. (2011). *Statistical Methods in the Atmospheric Sciences*, volume 100. Academic Press, third edition.

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

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

Xu, H., Caramanis, C., and Mannor, S. (2013). Outlier-Robust PCA: The High-Dimensional Case. *IEEE Trans. Inf. Theory*, 59:546–572.

Yang, S., Shen, H., Meng, J., and Shen, Z. (1999). A Fixed-point Iteration Algorithm for Robust Kernel Principal Component Analysis. *J. Comput. Inf. Syst.*, 10.

Zhang, J.-T. and Chen, J. (2007). Statistical inferences for functional data. *Ann. Statist.*, 35:1052–1079.

Zhao, Q., Meng, D., Xu, Z., Zuo, W., and Zhang, L. (2014). Robust principal component analysis with complex noise. In *Proc. of the 31st Int. Conf. on Machine Learning, Beijing, China, 2126 June 2014*.

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

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

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