Robust Stochastic Principal Component Analysis

John Goes
University of Minnesota

Teng Zhang
Princeton University

Raman Arora
Johns Hopkins University

Gilad Lerman
University of Minnesota

Abstract

We consider the problem of finding lower dimensional subspaces in the presence of outliers and noise in the online setting. In particular, we extend previous batch formulations of robust PCA to the stochastic setting with minimal storage requirements and runtime complexity. We introduce three novel stochastic approximation algorithms for robust PCA that are extensions of standard algorithms for PCA – the stochastic power method, incremental PCA and online PCA using matrix-exponentiated-gradient (MEG) updates. For robust online PCA we also give a sub-linear convergence guarantee. Our numerical results demonstrate the superiority of the the robust online method over the other robust stochastic methods and the advantage of robust methods over their non-robust counterparts in the presence of outliers in artificial and real scenarios.

1 Introduction

A classic problem in data analysis is the modeling of high-dimensional data by a lower dimensional subspace. The classic method here is Principal Component Analysis (PCA), which seeks the d-dimensional subspace maximizing the projected empirical variance based on i.i.d. draws from an unknown source distribution $\mathcal{D}$ [24].

The solution may be obtained directly by computing the sample covariance matrix of our batch and computing the top d singular vectors; we refer to it as the "batch" approach. If the number of draws is N, and the ambient dimension is D, merely computing the sample covariance matrix will require $\mathcal{O}(ND^2 + D^3)$ flops to run, and $\mathcal{O}(D^2)$ units of memory.

However, in practice these runtime and storage requirements may be unacceptable. Additionally, there are settings in which data is received sequentially and may even change over time, while one has no access to the source distribution. It is essential therefore to develop algorithms that are able to perform in a "streaming setting" as well as on data sets which are too large for batch methods. Furthermore, stochastic approximation algorithms have been shown, both theoretically and empirically, to be computationally preferable on various machine learning problems [4, 10, 34, 36, 37, 38].

Following the success of stochastic approximation algorithms in the "big data" setting, Arora et al. [1, 2] studied PCA in a stochastic optimization framework and reviewed and extended common approaches for stochastic PCA. They formulated the PCA objective as a stochastic optimization problem of seeking the $d-$dimensional subspace (parametrized by $U \in \mathbb{R}^{D\times d}$) that maximizes the variance *over the distribution* $\mathcal{D}$. Formally, they seek to solve the problem:

$$\max_{U\in\mathbb{R}^{D\times d}} \mathbb{E}_x[\mathrm{tr}(U^T xx^T U)] \quad (1)$$
$$\text{subject to } U^T U \preccurlyeq I,$$

where $\mathbb{E}_x$ denotes the expectation w.r.t. $x \sim \mathcal{D}$ and the constraint inequality is introduced to convexify the problem. The optimum will occur with $U^T U = I$. Various stochastic approximation algorithms were studied in [1] for solving (1) where they were categorized into three different types: stochastic gradient descent (SGD), incremental and online algorithms. These algorithms are computationally efficient and have good empirical performance.

However, despite its ubiquitous nature, PCA (and in particular, stochastic PCA) has a major weakness – it is extremely sensitive to outliers. Corrupted data points, which we refer to as outliers, can completely throw off the estimate of the principal subspace even with a single outlier [22]. In practice, we may encounter a high percentage of corruption (see e.g., the discussion at the end of Section 3.1 of [44]) and in

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33. Copyright 2014 by the authors.

theory (under some assumptions) the percentage of outliers tolerated by robust PCA algorithms can be significantly higher than the common 50% breakdown point of point estimators [44, 27, 20]. In such cases, the inliers may still be viewed as arising from $\mathcal{D}$, but the outliers are likely to be generated by a different distribution or may be even hard to model. The presence of these outliers, whose proportion may be significant, can completely distort the estimate of the expected variance and therefore the PCA subspace. There have been several attempts to endow PCA with resilience against outliers or other forms of gross corruptions (see e.g., [12, 17, 18, 19, 22, 23, 25, 32, 35, 41]). Following [9], Candès et al. [7] established a convex de-convolution method for extracting low dimensional subspace structure in the presence of gross but sparse uniformly distributed element-wise corruptions. This inspired the development of many other convex methods for robust PCA, but in the presence of outliers (instead of element-wise corruptions) [42, 33, 44, 27, 14, 15].

However, all of these methods work in the batch setting and therefore do not scale to big data. Some researchers have considered online algorithms for robust PCA [43, 31, 21, 45], which often fall within the first two stochastic approximation categories discussed in [1]. Indeed, [43], [21] and [45] all apply stochastic gradient descent (SGD) approaches. However, [21] applies to gross elementwise matrix corruption, instead of the setting of outliers considered here (they modify a Grassmannian SGD algorithm by incorporating the augmented Lagrangian of ℓ_1 norm which alleviates sparse corruption). Li [31] builds on the incremental PCA approach [5] where each new sample is re-weighted by a fixed influence function. Unfortunately, no performance guarantees are given for any of these methods, which is not surprising as there is little known for stochastic approximation algorithms for the original PCA formulation [1]. Furthermore, we are unaware of any robust version that belongs to the third category of online algorithms.

In this paper we study stochastic algorithms for robust PCA in a principled framework and propose an online algorithm for robust PCA with good theoretical guarantees and excellent empirical performance. We build on ideas of two recent works on robust PCA [44, 27] since they both adapt well to the stochastic formulation of (2). We present robust analogues for the three categories of stochastic approximation algorithms presented in Arora et al. [1, 2]. However, we emphasize the robust extension of the robust online approach due to the novelty of the approach and especially due to its competitive theoretical guarantees as well as empirical performance.

2 Review of Relevant Work

This work is based on two different research directions. The first one is about robust PCA via convex relaxation of absolute subspace deviations; it includes both the Geometric Median Subspace (GMS) algorithm [44] and the REAPER algorithm [27]. The second one is the study by Arora et al. [1, 2] of PCA in a stochastic optimization framework.

2.1 GMS and REAPER

The GMS [44] and the REAPER [27] paradigms assume a given dataset $\mathcal{X}$ in $\mathbb{R}^D$ and a target dimension $d \in \{1, 2, \dots, D-1\}$ and propose a convex optimization problem solving for a matrix Q from which a d-dimensional subspace is determined; this subspace is a robust approximation of the data.

In GMS, Q is the solution of

$$\min \sum_{x \in \mathcal{X}} \|Qx\|_2 \text{ subject to } Q = Q^T \text{ and } \operatorname{tr}(Q) = 1. \tag{2}$$

In this case, Q is interpreted as a robust inverse covariance (initial dimensionality reduction may be needed to assure that the covariance has full rank). The dimension d can be sometimes estimated from the eigenvalues of Q.

In REAPER, Q is the solution of

$$\min \sum_{x \in \mathcal{X}} \|Qx\|_2 \text{ subject to } Q \preccurlyeq I \text{ and } \operatorname{tr}(Q) = D - d. \tag{3}$$

In this case, Q is interpreted as a tight approximation to the orthogonal projector onto the orthogonal complement of the d-dimensional subspace minimizing the least absolute deviations w.r.t. $\mathcal{X}$.

Both algorithms output a subspace obtained by the span of the bottom d eigenvectors of Q, or equivalently, the top d eigenvectors of $I - Q$. Both GMS and REAPER suggest iteratively re-weighted least squares strategies converging to the solutions of (2) and (3) respectively.

2.2 Stochastic Approaches for PCA

Arora et al. [1] reviewed and extended common approaches for stochastic PCA, while categorizing these approaches into the following three classes.

2.2.1 Stochastic Gradient Descent

Assume that the covariance of $\mathcal{D}$ is known to be Σ. The gradient with respect to U of the PCA objective function $\operatorname{tr}\left(U^T \Sigma U\right)$ is $2\Sigma U$. The observation that $\Sigma =$

$\mathbb{E}_x[xx^T]$, leads to the update

$$U^{(t+1)} = \mathcal{P}_{\text{orth}}\left(U^{(t)} + \eta x_t x_t^T U^{(t)}\right), \qquad (4)$$

where $\mathcal{P}_{\text{orth}}$ is a pseudo-projection with respect to the spectral norm of UU^T onto the set of $D \times D$ matrices with d eigenvalues equal to 1 and the rest zero. This can be obtained by simply taking the SVD of U, which is symmetric, and culling the top d eigenvectors.

It is shown that the cost of performing T iterations costs $\mathcal{O}(TDd)$ flops with $\mathcal{O}(Dd)$ units of memory.

2.2.2 Incremental PCA

The second algorithm considered in Arora et al. [1] is based on the incremental SVD algorithm [5], which computes the SVD of the matrix $X = [x_1, x_2, \ldots, x_T]$ iteratively. If storage and runtime were not an issue we could use this algorithm to incrementally compute the second-moment matrix directly.

Arora et al. [1] extended this to the online setting where we would seek the eigendecomposition of the second moment matrix updated iteratively. This led them to the update

$$C^{(t)} = \mathcal{P}_{\text{rank-}d}\left(C^{(t-1)} + x_t x_t^T\right), \qquad (5)$$

where $\mathcal{P}_{\text{rank-}d}$ denotes the retaining of just the top d eigenvectors and eigenvalues. Following Brand [5] this update can be performed *efficiently* since it is a rank-one symmetric update. The run-time for this algorithm is $\mathcal{O}(Dd^2)$ per iteration, with storage requirements of $\mathcal{O}(Dd)$.

2.2.3 Online PCA

The third algorithm considered in Arora et al. [1] is based on the Randomized Online PCA algorithm of Warmuth and Kuzmin [40], which can be interpreted as solving the following minimization for PCA [2]:

$$\begin{gathered} \min_M \mathbb{E}_x[\operatorname{tr}(Mxx^T)] \qquad (6) \\ \text{subject to } M \succcurlyeq 0,\ ||M||_2 \leq \tfrac{1}{D-d}\ ,\ \operatorname{tr}(M) = 1. \end{gathered}$$

The algorithm is involved and one should consult Warmuth and Kuzmin [40] for the details. It was shown to be an instance of the mirror-descent algorithm by Arora et al. [2] with the distance generating function being a shifted-and-scaled version of the negative von Neumann entropy: $\Psi(M) = \frac{1}{4}(\operatorname{tr}(M \ln M) + \ln D)$. The update rule in this case is

$$M^{(t+1)} = \Pi\left(\exp\left(\ln M^{(t)} - \eta_t x_t x_t^T\right)\right), \qquad (7)$$

where Π is the projection with respect to the quantum relative entropy.

Warmuth and Kuzmin's algorithm includes a regret analysis that leads to convergence guarantees. Arora et al. [1] showed that if we take a constant step size $\eta_t = \eta$, assume that $\|x_t\|_2 \leq 1$ uniformly in t, and perform N iterations such that

$$N \geq \mathcal{O}\left(\left(\frac{(D-d)\operatorname{tr}(M^*\Sigma) + \epsilon}{\epsilon}\right)\frac{d \log D/d}{\epsilon}\right), \qquad (8)$$

where M^* is the optimal solution, then the iterates $M^{(t)}$ will satisfy

$$\mathbb{E}_x\left[\frac{1}{T}\sum_{t=1}^{T} \operatorname{tr}\left((D-d)M^{(t)}\Sigma\right)\right] - \operatorname{tr}((D-d)M^*\Sigma) \leq \epsilon, \qquad (9)$$

where the iterates and optimal solution are scaled appropriately relative to the objective function.

3 Robust PCA Stochastic Algorithms

In this section we use ideas from the batch robust-PCA methods of GMS [44] and REAPER [27] to extend the three types of online algorithms considered in Arora et al. [1] to the robust setting.

Here $\mathbb{E}_x$ denotes the expectation with respect to the random variable x, which obeys a certain distribution $\mathcal{D}'$. We may assume that $\mathcal{D}'$ is a mixture of two components, which represent inliers (with distribution $\mathcal{D}$) and outliers. The aim is to apply a sufficiently robust objective function so that the minimizers when $x \sim \mathcal{D}'$ and $x \sim \mathcal{D}$ are sufficiently close. The analysis in [44, 27] shows that under some strict assumptions on the underlying distributions the minimizers are the same for the objective functions of both GMS and REAPER and with weaker assumptions they are nearby.

3.1 Robust SGD

We suggest two robust modifications of the minimization problem in (2) and apply SGD to solve them. Here we do not use directly the formulations of [44, 27], but similar ones.

3.1.1 Method One

We introduce the following analogue to the PCA objective:

$$\max_{U \in \mathbb{R}^{D \times d}} \mathbb{E}_x\|U^t x\|_2 \text{ subject to } \|U\|_2 \leq 1. \qquad (10)$$

Indeed, by replacing $\mathbb{E}_x\|U^t x\|_2$ with $\mathbb{E}_x\|U^t x\|_2^2$ we obtain the PCA objective. One may also use $\mathbb{E}_x\|U^t x\|_2^p$ with $0 < p < 2$. We have chosen $p = 1$ since then the optimization problem is still convex (which is true for any $p \geq 1$), while more robust to outliers (we expect robustness to strengthen when lowering p). If $d = 1$, then the solution of (10) is the well-known projection-pursuit PCA [30, 33], whose robustness is established in [30, 11]. We are unaware though of references establishing the robustness of the minimizer of (10) when $d > 1$.

The gradient of the new objective function with respect to U is given as $\|Ux\|_2^{-1} \cdot xx^TU$. This leads us to consider updates of the form

$$U^{(t+1)} = \mathcal{P}_{\text{orth}}\left(U^{(t)} + \eta_t x_t x_t^T U^{(t)}/\|U^{(t)T}x\|_2^{2-p}\right). \tag{11}$$

In fact, per the argument presented in Arora et al. [1], this projection needs only be done infrequently for numerical reasons. The number of operations remains $\mathcal{O}(NDd)$, where N is the number of data points processed, and with memory $\mathcal{O}(Dd)$.

3.1.2 Method Two

We consider the formulation

$$\min_{U\in\mathbb{R}^{D\times(D-d)}} \mathbb{E}_x\|U^t x\|_2 \text{ subject to } U^TU = I. \tag{12}$$

Again, we may replace $\mathbb{E}_x\|U^t x\|_2$ with $\mathbb{E}_x\|U^t x\|_2^p$, where $p > 0$ and $p = 2$ results in the PCA solution. While the objective function $\mathbb{E}_x\|U^t x\|_2$ of (12) is convex, the constraint, and thus the optimization problem, is not convex. Robustness of the solution of (12) under special conditions was studied in [28, 29].

This leads to the update formula

$$U^{(t+1)} = \mathcal{P}_{\text{orth}}\left(U^{(t)} - \eta_t x_t x_t^T U^{(t)}\|U^{(t)T}x\|_2^{-1}\right). \tag{13}$$

We can parametrize the $U \in \mathbb{R}^{D\times(D-d)}$ matrices by their complements and preserve the complexity achievements of Method One above. We remark that this second method is very similar to [45], when applied to a single subspace modeling (instead of hybrid linear modeling).

3.2 Robust Incremental Method

This method is based on the GMS formulation. We recall that the GMS algorithm defines the "robust precision" matrix

$$\hat{Q} = \underset{Q\in\mathbb{R}^{d\times d},\text{tr}(Q)=1}{\arg\min} \mathbb{E}_x\|Qx\|_2, \tag{14}$$

giving $\hat{\Sigma} = \hat{Q}^{-1}$ as the "robust covariance". Differentiating we see that

$$\hat{Q} = c \cdot \left(\mathbb{E}_x \frac{xx^T}{\|\hat{Q}x\|_2}\right)^{-1}, \tag{15}$$

where c is a constant. Equivalently,

$$\hat{\Sigma} = c^{-1} \cdot \mathbb{E}_x \frac{xx^T}{\|(\hat{\Sigma})^{-1}x\|_2}. \tag{16}$$

This leads us to consider the update

$$\Sigma^{(t)} = \mathcal{P}_{\text{rank-}d}\left(\Sigma^{(t-1)} + \frac{x_t x_t^T}{\|(\Sigma^{(t-1)})^{-1}x_t\|_2}\right). \tag{17}$$

However, after taking the rank-d projection, the inverse will not be defined.

By Sherman-Morrison-Woodbury formula,

$$(\Sigma + xx^T)^{-1} = \Sigma^{-1} - \frac{\Sigma^{-1}xx^T\Sigma^{-1}}{1 + x^T\Sigma^{-1}x}.$$

In order to compute and store this update efficiently, we consider the inverse update

$$(\Sigma^{(t)})^{-1} = (\Sigma^{(t-1)})^{-1} - \frac{(\Sigma^{(t-1)})^{-1}xx^T(\Sigma^{(t-1)})^{-1}}{1 + x^T(\Sigma^{(t-1)})^{-1}x}. \tag{18}$$

Combining the rank-d update with the inverse update yields the result. The inverse update can be computed in $4D + 5Dd$ flops, with the subsequent rank-one update remaining as efficient as in Arora et al. [1]. Therefore the robust incremental PCA still has runtime $\mathcal{O}(NDd^2)$ and space complexity $\mathcal{O}(Dd)$.

3.3 Robust Online PCA

This method is based on the REAPER formulation. Equation (3), which formulates the REAPER's minimization, suggests the following stochastic analogue:

$$\min\nolimits_Q \mathbb{E}_x||Qx||_2 \tag{19}$$
$$\text{subject to } \text{tr}(Q) = 1 \text{ and } ||Q||_2 \leq \tfrac{1}{D-d}.$$

We remark that the minimizer of (19) is semi-definite positive, so the constraint $Q \succcurlyeq 0$ can be added to (19) (in analogy to (6)).

Applying the mirror-descent algorithm to (19) with the distance generating function being a shifted-and-scaled version of the negative von Neumann entropy: $\Psi(M) = \frac{1}{4}(\text{trace}M \ln M + \ln D)$, we get the following MEG updates:

$$Q^{(t+1)} = \Pi\left(\exp\left(\log Q^{(t)} - \eta_t \frac{(x_t x_t^T Q^{(t)} + Q^{(t)} x_t x_t^T)}{2||Q^{(t)}x_t||_2}\right)\right). \tag{20}$$

For the robust online PCA update above in (20), we prove the following convergence guarantee.

Lemma 3.1. *If we perform T iterations with step size $\eta_t = \frac{2}{\sqrt{T}}\sqrt{\frac{d}{D-d}}$, then*

$$(D-d)\mathbb{E}_x\|\hat{Q}x\|_2 \leq (D-d)\inf_Q \mathbb{E}_x\|Qx\|_2 + \sqrt{\frac{d(D-d)}{T}},$$

where $\hat{Q}$ is sampled uniformly from the set $\{Q^{(1)}, Q^{(2)}, \dots Q^{(T)}\}$.

Proof. Using Lemma 2 of [39], and an online-to-batch conversion [8], we have that

$$(D-d)\mathbb{E}_x\|\hat{Q}x\|_2 \leq (D-d)\inf_Q \mathbb{E}_x\|Qx\|_2 + \sqrt{\frac{2B}{T}},$$

where $B = \sup_Q \Psi(Q)$, and we used a step size of $\eta'_t = \sqrt{\frac{B}{T}}$. It is easy to check that Ψ is nonnegative and 1-strongly convex with respect to $\|\cdot\|_1$. Furthermore, the maximum value of Ψ is achieved on the corners of the matrix simplex constraints:

$$\sup_Q \Psi(Q) = \frac{1}{4}\left(\ln\frac{1}{D-d} + \ln D\right) \leq \frac{1}{4}\frac{d}{D-d}.$$

Using this bound and a step-size of $\eta_t = 4\eta'_t = \frac{2}{\sqrt{T}}\sqrt{\frac{d}{D-d}}$, we get the desired bound. □

The efficiency of the online updates depends on the ranks of the intermediate iterates. If k_t denotes the rank of the intermediate iterate $Q^{(t)}$, the online algorithm requires $O(N\bar{k}^2D)$ operations to process N data points, where $\bar{k}^2 = \sum_{t=1}^N k_t^2$.

4 Experimental Results

We run both artificial and natural experiments to verify that these algorithms reliably detect linear structure in the presence of outliers. We also try to measure how well the algorithms perform in various regimes of $D, d, N_{\text{out}}, N_{\text{in}}$, where N_{out} and N_{in} represent the number of outliers and inliers respectively.

4.1 Artificial Data

We consider a data regime which consists of a low-dimensional linear structure in the presence of high-dimensional outliers (we follow [27] but allow asymmetric covariance for the outliers). We summarize its components in Table 1, while explaining them in details below.

Table 1: A model for sampling a low-dimensional linear structure in the presence of high-dimensional outliers

D	Ambient dimension
L	A d-dim. subspace of $\mathbb{R}^D$ containing inliers
N_{in}	Number of inliers
N_{out}	Number of outliers
ρ_{in}	Inlier number-per-dimension ratio: N_{in}/d
ρ_{out}	Outlier number-per-dimension ratio: N_{out}/D
Σ_{out}	Asymmetric outlier covariance matrix
Σ_{in}	Symmetric inlier covariance matrix
λ	knob to control outlier magnitude

The model for the inliers' component fixes a d-dimensional space L and the $D \times D$ matrix Σ_{in}, which is the identity on L and zero elsewhere. We draw inliers from $\mathcal{N}(0, (1/d)\Sigma_{\text{in}})$.

The model for the outliers fixes the $D \times D$ diagonal matrix Σ_{out}, whose eigenvalues are $1, 2, \dots, D$. We draw outliers from $\mathcal{N}(0, \Sigma_{\text{out}}(\frac{\lambda}{(D*\text{mean}(\Sigma_{\text{out}})}))$, where we normalize by the mean of the eigenvalues of Σ_{out} (which we denote by mean(Σ_{out}) to give outliers a comparable magnitude to inliers. The multiplicative factor λ is used as a knob to control the outlier magnitude.

Figures 1, 2 and 3 show the outcome of experiments in various data regimes at extremes. The robust online PCA clearly outperforms all algorithms consistently in all these regimes and converges quickly to the ground truth. The second stochastic formulation of robust PCA is also doing well.

Following the discussion in [27], we introduce the useful statistics ρ_{in}, ρ_{out} (see Table 1). Using (batch) REAPER, recovery guarantees are seen to be linear with respect to ρ_{in} and ρ_{out}. This may be visualized in Figure 4, which is the result of the robust online PCA algorithm running after 3000 iterations. Note that because our initializations are random, some noise is introduced into this graph that indicates some sensitivity to initialization.

4.2 Astronomical Data

An interesting test of our streaming algorithms are astronomical data sets. As telescopes gather more and more astronomical data, it is critical to find ways of processing these data sets quickly without sacrificing much in accuracy.

A particularly interesting data set is the VIMOS Very Large Telescope (VIMOS-VLT) Deep Survey [16], which is aimed at understanding the evolution of galaxies. For details on the problem and data set, consult [16, 6].

Batch-PCA does not work well in this context because

Table 2: The subspace recovery error $\|P_{L_*} - P_{\hat{L}}\|_F/3\sqrt{2}$ for the different algorithms and the face experiment

SGD	R-SGD1	R-SGD2	Inc	R-Inc	MD	R-MD
0.64	0.84	0.74	0.62	0.60	0.52	0.45

of the high noise-to-signal ratio. We compare the top 4 eigenspectra of our robust online method with the top 4 given by [6].

We seek to represent the spectrum of a galaxy by a few continuous parameters that account for the best-fit spectrum. Thus good methods are roughly continuous. Figure 5 compares the robust online PCA algorithm with the robust PCA algorithm employed in the state-of-the art eigenspectra extraction and the non-robust online PCA algorithm. Visually the eigenpectra uncovered by robust online PCA are a close and relatively smooth fit to ground truth (closely approximated by the eigenspectra uncovered by [6]), while online PCA is completely failing to discover the eigenspectra.

4.3 Face Data

We compose a data set containing images of faces under different illuminating conditions, which serve as inliers concentrated around a low-dimensional subspace, and random natural images, which serve as outliers.

The inliers are 640 face images from the Extended Yale Face Database [26], where there are 10 faces and for each face 64 images were taken under different illuminating conditions (the images are cropped to contain only the region of the face). In theory, images of a face under varying illuminating conditions are well-approximated by a 3-dimensional linear subspace if there are no shadows [26] and well-approximated by a 9-dimensional subspace when they contain shadows [3]. In practice, 5-dimensional subspaces approximate well each one of these faces. Here we consider a low-rank model to the set of all faces. We set $d = 9$ since we were still able to recover faces after projecting them onto a 9-dimensional subspace.

The outliers are 400 random images from the BACKGROUND/Google folder of the Caltech101 database [13] as outliers. All images (both inliers and outliers) are converted to grayscale and downsampled to 20×20 pixels.

We apply the algorithms of Section 3 to this data set to obtain a 9-dimensional subspace. In Figure 6, we visually demonstrate the results by showing the projections of the face images to the learned subspace; clearer images indicate on better performance for this particular images. For these examples the robust methods performs better, while the robust online PCA performs the best.

We also quantitatively measure the performance of these algorithms by the error term $\|P_{L_*} - P_{\hat{L}}\|_F/(3\sqrt{2})$, where $\hat{L}$ is the fitted subspace by the algorithms and L_* is the subspace fitted by PCA to the set of inliers. When $L^* \perp \hat{L}$, $\|P_{L_*} - P_{\hat{L}}\|_F = 3\sqrt{2}$; therefore, we normalize $\|P_{L_*} - P_{\hat{L}}\|_F$ by $3\sqrt{2}$ in order to get an error between 0 and 1. The smaller the error the better the estimated subspace. These errors are recorded in Table 2, which indicates that generally the robust methods work better than the non-robust ones and that the robust online PCA has the smallest subspace recovery error.

5 Conclusion

We have developed stochastic formulations of robust PCA in analogy to common stochastic formulations of PCA. As in the PCA case, there remain no quantitative guarantees for the gradient descent or the pure incremental methods. However we have shown in artificial and real data that our algorithms perform better than stochastic PCA (non-robust) methods on outlier-corrupted data. We extended the previous stochastic formulations without sacrificing computation or speed. We also proved a sub-linear convergence rates for a robust analogue to the online PCA algorithm.

It is worth noting for the practitioner that projecting all data points onto the sphere as a pre-processing step in and of itself introduces quite a bit of robustness even to traditional online PCA. In particular the algorithm of Warmuth and Kuzmin [40] closely parallels the success of robust online PCA after projecting data to the sphere, though even here robust PCA retains a slight edge.

Acknowledgements

The first author was partially supported by the National Science Foundation Graduate Research Fellowship Program (Fellow ID: 2012146015). When preparing this work, he resided at the Statistical and Applied Mathematical Sciences Institute as a Visiting Graduate Fellow and would like to thank the Institute.

The last author was partially supported by NSF award DMS-09-56072 and by the Feinberg Foundation Visiting Faculty Program Fellowship of the Weizmann Institute of Science; he thanks the Department of Com-

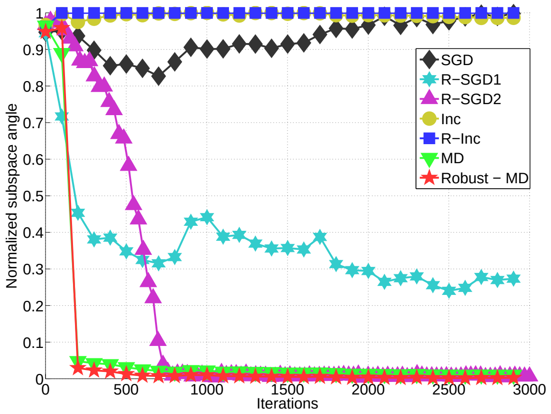

Figure 1: d=1 dimensional subspace with outliers ($\lambda = 10$) in D=100 dimension. 80% of the data are outliers.

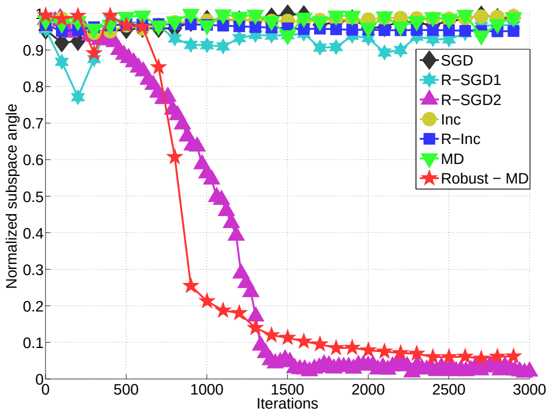

Figure 2: d=5 dimensional subspace with outliers ($\lambda = 10$) in D=100 dimension. 71% of the data are outliers.

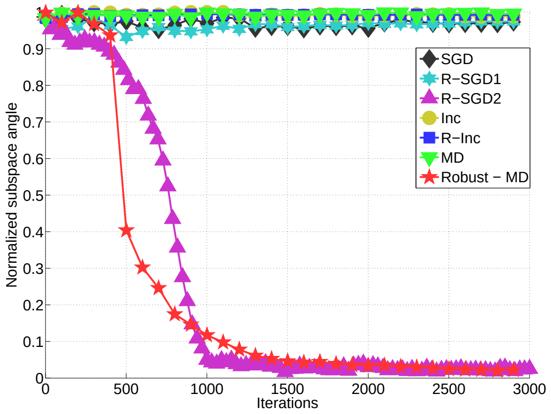

Figure 3: d=10 dimensional subspace with outliers ($\lambda = 10$) in D=100 dimension. 45% of the data are outliers.

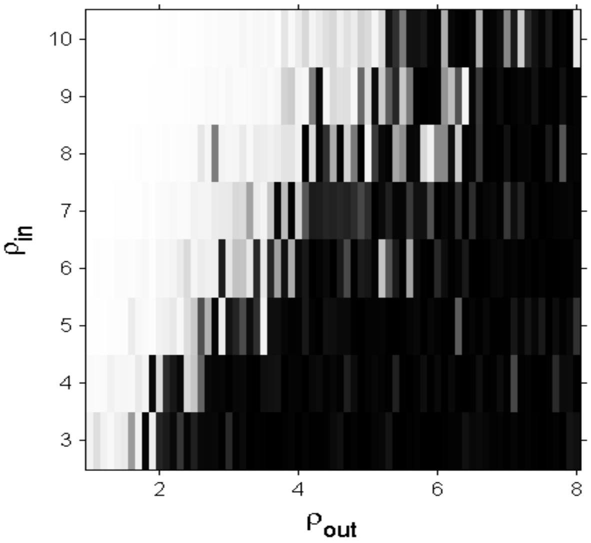

Figure 4: Three Dimensional Subspace in D=100 dimensions. Recovery is measured by the subspace angle of the ground truth space and the recovered space after 2000 iterations, with white signifying a zero subspace angle and black a completely orthogonal angle.

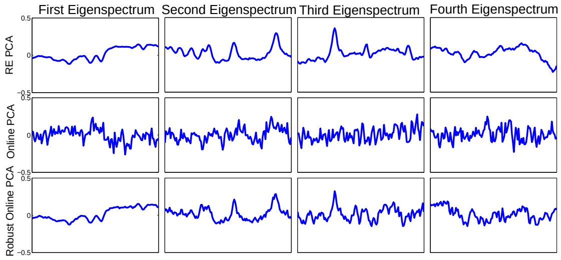

Figure 5: The top four eigenspectra for the VVDS galaxies. The top row is the state-of-the art algorithm of [6]. The middle row is online PCA. The bottom row is the robust online PCA.

puter Science and Applied Mathematics at the Weizmann Institute for their hospitality and inspiration at the time of completing this work. He also thanks Nati Srebro for his hospitality during his visit at TTI on September 2012. In this visit, the last author presented his joint work on robust PCA and the third author and Nati Srebro introduced him to their joint work on online PCA. This meeting triggered this joint work.

We thank Joshua Vogelstein for suggesting the astrophysics data and Tyler Maunu, Tamás Budavári and David Lawlor for their help with processing the data or interpreting the results of the various algorithms on this data. We also thank the VVDS Team for making the data publicly available along with their metadata.

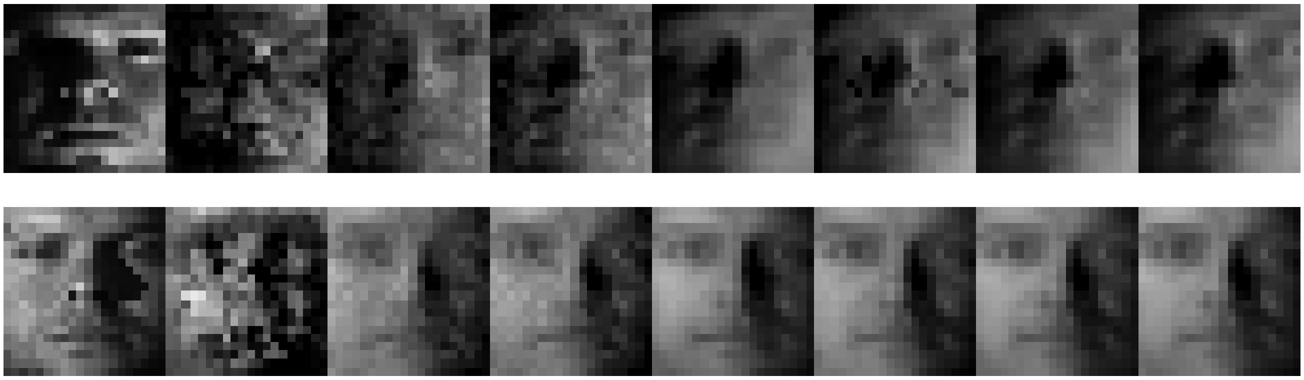

Figure 6: From left to right: the original images, the projection of original images to the fitted subspace by the following methods: SGD, robust SGD (method one), robust SGD (method two), incremental, robust incremental, online PCA and robust online PCA.

References

[1] R. Arora, A. Cotter, K. Livescu, and N. Srebro. Stochastic optimization for PCA and PLS. In *Allerton Conference*, pages 861–868, 2012.

[2] R. Arora, A. Cotter, and N. Srebro. Stochastic optimization of PCA with capped MSG. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[3] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, February 2003.

[4] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *NIPS'07*, pages 161–168, 2007.

[5] M. Brand. Incremental singular value decomposition of uncertain data with missing values. In *ECCV*, pages 707–720, 2002.

[6] T. Budavári, V. Wild, A. S. Szalay, L. Dobos, and C.-W. Yip. Reliable eigenspectra for new generation surveys. *Monthly Notices of the Royal Astronomical Society*, 394(3):1496–1502, 2009.

[7] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[8] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *Information Theory, IEEE Transactions on*, 50(9):2050–2057, 2004.

[9] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.*, 21(2):572–596, 2011.

[10] M. Collins, A. Globerson, T. Koo, X. Carreras, and P. L. Bartlett. Exponentiated gradient algorithms for conditional random fields and max-margin markov networks. *J. Mach. Learn. Res.*, 9:1775–1822, June 2008.

[11] C. Croux and A. Ruiz-Gazen. High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95(1):206 – 226, 2005.

[12] F. De La Torre and M. J. Black. A framework for robust subspace learning. *International Journal of Computer Vision*, 54(1-3):117–142, 2003.

[13] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 106(1):59–70, Apr. 2007.

[14] J. Feng, H. Xu, and S. Yan. Robust PCA in high-dimension: A deterministic approach. *International conf. on machine learning (ICML)*, 2012.

[15] J. Feng, H. Xu, and S. Yan. Online robust PCA via stochastic optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[16] O. L. Fèvre et al. The VIMOS VLT deep survey. first epoch VVDS-Deep survey: 11564 spectra with 17.5 <I(AB)< 24, and the redshift distribution over $0 < z < 5$. *Astronomy and Astrophysics*, 439:845, 2005.

[17] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[18] R. Gnanadesikan and J. R. Kettenring. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, pages 81–124, 1972.

[19] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Ap-*

proach Based on Influence Functions. Wiley-Interscience, New York, April 2005.

[20] M. Hardt and A. Moitra. Can we reconcile robustness and efficiency in unsupervised learning? In *Proceedings of the Twenty-sixth Annual Conference on Learning Theory (COLT 2013)*, 2013.

[21] J. He, L. Balzano, and J. Lui. Online robust subspace tracking from partial information. *arXiv preprint arXiv:1109.3827*, 2011.

[22] P. J. Huber and E. M. Ronchetti. *Robust Statistics.* Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, 2nd edition, 2009.

[23] M. Hubert, P. J. Rousseeuw, and K. V. Branden. ROBPCA: a new approach to robust principal component analysis. *Technometrics*, 47(1), 2005.

[24] I. T. Jolliffe. *Principal Component Analysis.* Springer Series in Statistics. Springer, 2nd edition, 2002.

[25] Q. Ke and T. Kanade. Robust L_1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, pages 739–746, June 2005.

[26] K. C. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 27(5):684–698, 2005.

[27] G. Lerman, M. McCoy, J. A. Tropp, and T. Zhang. Robust computation of linear models, or how to find a needle in a haystack. *ArXiv e-prints*, Feb. 2012.

[28] G. Lerman and T. Zhang. ℓ_p-Recovery of the Most Significant Subspace among Multiple Subspaces with Outliers. *ArXiv e-prints*, Dec. 2010.

[29] G. Lerman and T. Zhang. Robust recovery of multiple subspaces by geometric l_p minimization. *Ann. Statist.*, 39(5):2686–2715, 2011.

[30] G. Li and Z. Chen. Projection-Pursuit Approach to Robust Dispersion Matrices and Principal Components: Primary Theory and Monte Carlo. *Journal of the American Statistical Association*, 80(391):759–766, 1985.

[31] Y. Li. On incremental and robust subspace learning. *Pattern recognition*, 37(7):1509–1518, 2004.

[32] R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust statistics: Theory and methods.* Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 2006.

[33] M. McCoy and J. Tropp. Two proposals for robust PCA using semidefinite programming. *Elec. J. Stat.*, 5:1123–1160, 2011.

[34] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, January 2009.

[35] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[36] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal Estimated sub-GrAdient SOlver for SVM. In *ICML'07*, pages 807–814, 2007.

[37] S. Shalev-Shwartz and N. Srebro. SVM optimization: Inverse dependence on training set size. In *ICML'08*, pages 928–935, 2008.

[38] S. Shalev-Shwartz and A. Tewari. Stochastic methods for l_1 regularized loss minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML'09, pages 929–936, 2009.

[39] N. Srebro, K. Sridharan, and A. Tewari. On the universality of online mirror descent. In *Advances in Neural Information Processing Systems 24*, pages 2645–2653, 2011.

[40] M. K. Warmuth and D. Kuzmin. Randomized Online PCA Algorithms with Regret Bounds that are Logarithmic in the Dimension. *Journal of Machine Learning Research*, 9:2287–2320, Oct. 2008.

[41] H. Xu, C. Caramanis, and S. Mannor. Principal component analysis with contaminated data: The high dimensional case. In *COLT*, 2010.

[42] H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via outlier pursuit. *Information Theory, IEEE Transactions on*, PP(99):1, 2012.

[43] L. Xu and A. Yuille. Robust principal component analysis by self-organizing rules based on statistical physics approach. *Neural Networks, IEEE Transactions on*, 6(1):131–143, 1995.

[44] T. Zhang and G. Lerman. A novel m-estimator for robust PCA. Submitted, available at arXiv:1112.4863.

[45] T. Zhang, A. Szlam, and G. Lerman. Median K-flats for hybrid linear modeling with many outliers. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on Computer Vision*, pages 234–241, Kyoto, Japan, 2009.