




Robust Orthogonal Complement Principal Component Analysis

Yiyuan She, Shijie Li & Dapeng Wu


To cite this article: Yiyuan She, Shijie Li & Dapeng Wu (2016) Robust Orthogonal Complement Principal Component Analysis, Journal of the American Statistical Association, 111:514, 763-771, DOI: [10.1080/01621459.2015.1042107](https://doi.org/10.1080/01621459.2015.1042107)

To link to this article: <https://doi.org/10.1080/01621459.2015.1042107>

 View supplementary material [↗](#)

 Accepted author version posted online: 24 Jun 2015.
Published online: 18 Aug 2016.

 Submit your article to this journal [↗](#)

 Article views: 942

 View related articles [↗](#)

 View Crossmark data [↗](#)

 Citing articles: 4 View citing articles [↗](#)

Robust Orthogonal Complement Principal Component Analysis

Yiyuan She, Shijie Li, and Dapeng Wu

ABSTRACT

Recently, the robustification of principal component analysis (PCA) has attracted lots of attention from statisticians, engineers, and computer scientists. In this work, we study the type of outliers that are not necessarily apparent in the original observation space but can seriously affect the principal subspace estimation. Based on a mathematical formulation of such transformed outliers, a novel robust orthogonal complement principal component analysis (ROC-PCA) is proposed. The framework combines the popular sparsity-enforcing and low-rank regularization techniques to deal with row-wise outliers as well as element-wise outliers. A nonasymptotic oracle inequality guarantees the accuracy and high breakdown performance of ROC-PCA in finite samples. To tackle the computational challenges, an efficient algorithm is developed on the basis of Stiefel manifold optimization and iterative thresholding. Furthermore, a batch variant is proposed to significantly reduce the cost in ultra high dimensions. The article also points out a pitfall of a common practice of singular value decomposition (SVD) reduction in robust PCA. Experiments show the effectiveness and efficiency of ROC-PCA in both synthetic and real data. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received October 2014
Revised March 2015

KEYWORDS

Low-rank approximation;
Manifold optimization;
Oracle inequalities; Outliers;
Sparsity.

1. Introduction

During the past few years, big data arising in machine learning, signal processing, genetics, and many other fields pose a dimensionality challenge in statistical computation and analysis. To uncover low-dimensional structures underlying such high-dimensional data, the principal component analysis (PCA) is one of the most popularly used multivariate dimension reduction tools. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a data matrix with n observations in p -dimensional space. PCA can be characterized by finding a low-rank data approximation, that is, $\min_{\mathbf{B}} \|\mathbf{X} - \mathbf{B}\|_F^2$ subject to $\text{rank}(\mathbf{B}) \leq r$. The solution is given by a truncated singular value decomposition (SVD) of \mathbf{X} : $\hat{\mathbf{B}} = \mathbf{U} \text{diag}\{\sigma_1, \dots, \sigma_r\} \mathbf{V}^T = \mathbf{X} \mathbf{V} \mathbf{V}^T$, where \mathbf{V} consists of the first r right singular vectors of \mathbf{X} with $\mathbf{V} \mathbf{V}^T$ defining the rank- r principal subspace. The squared error loss function is reasonable under a Gaussian noise model $\mathbf{X} = \mathbf{B} + \mathbf{E}$, but is notoriously known to be nonrobust and sensitive to atypical observations or the so-called *outliers*. Outliers typically refer to extreme observations far away from the majority of the data, and occur ubiquitously in real life data (Maronna, Martin, and Yohai 2006; Hampel et al. 2011). They may seriously affect statistical estimation and inference—in fact, a single outlier can break down the PCA completely and result in a misleading subspace estimate.

The robustification of PCA has been extensively studied in robust statistics, for example, Rousseeuw and Van Driessen (1999), Locantore et al. (1999), Hubert, Rousseeuw, and Branden (2005), among many others. The recent renowned *principal component pursuit* (PCP) due to Candès et al. (2011) has drawn a lot of attention from researchers even beyond the statistics community. PCP decomposes \mathbf{X} into a low-rank component \mathbf{B} and a sparse gross outlier component \mathbf{S} . The recovery

problem can be formulated by $\min_{\mathbf{B}, \mathbf{S}} \text{rank}(\mathbf{B}) + \lambda \|\mathbf{S}\|_0$ subject to $\mathbf{X} = \mathbf{B} + \mathbf{S}$, where $\|\cdot\|_0$ denotes the element-wise ℓ_0 norm, that is, the number of all nonzeros. PCP applies a convex relaxation to facilitate computation and analysis: $\min_{\mathbf{B}, \mathbf{S}} \|\mathbf{B}\|_* + \lambda \|\mathbf{S}\|_1$ subject to $\mathbf{X} = \mathbf{B} + \mathbf{S}$, where $\|\cdot\|_*$ denotes the matrix nuclear norm (sum of all singular values), and $\|\cdot\|_1$ denotes the element-wise ℓ_1 norm. PCP has various extensions and variants (Zhou et al. 2010; Xu, Caramanis, and Sanghavi 2010; Wright et al. 2013), and has widespread applications in image and video analysis, for example, Wright et al. (2009), Peng et al. (2012), Zhang et al. (2012). Although PCP can effectively deal with additive outliers in the original observation space, it may fail in the presence of another important type of outliers, the so-called *OC outliers*, which is the major concern of this work.

In robust principal component analysis, the outliers worthy of attention must affect the principal subspace estimation. Figure 1 gives some toy examples to illustrate how outliers could interfere with principal subspace identification. In the left panel, some outlying samples exist in the principal component subspace (PC subspace), which we call *pure* PC outliers, or PC outliers, for short. Interestingly, they do not affect the detection of PC subspace. PC outliers are not that harmful (though possibly affecting the order of PC directions), and thus, can be handled in a later stage. However, if one only checks the raw (x, y) coordinates in the observed space, these samples might be labeled as outliers.

The right panel contains some samples atypical in the orthogonal complement subspace (OC subspace), which we call OC outliers. We emphasize that *any* points showing outlyingness in OC coordinates are referred to as OC outliers in this article, such as Point A in Figure 1(b), whether or not they show PC outlyingness. It is the OC outliers that can skew

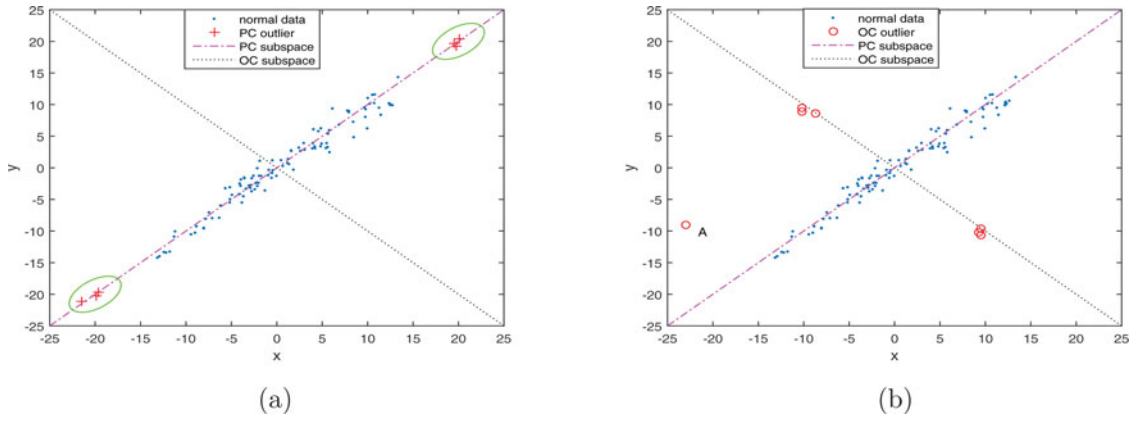


Figure 1. PC outliers and OC outliers. (a) Normal data ("•") and PC outliers ("×"). (b) Normal data ("•") and OC outliers ("o").

the PC subspace. Unfortunately, checking their coordinates in the observation space offers little help, thereby making PCP possibly fail in recovering the genuine PC subspace. However, if one could project the data points onto the ideal OC subspace, such outliers would be easily revealed and detected.

This article proposes a novel robust orthogonal complement principal component analysis (ROC-PCA) to address such OC outliers in principal subspace recovery. In contrast to the existing robust PCA approaches, ROC-PCA explicitly deals with OC outliers, and aims at simultaneous outlier identification and robust principal subspace recovery. Both *row-wise* (r-Type) outliers and *element-wise* (e-Type) outliers are discussed. Our computation algorithm involves Stiefel manifold optimization and allows for all popular sparsity-enforcing penalties to be used. We also establish a nonasymptotic oracle inequality to provide a theoretical guarantee of ROC-PCA from the predictive learning perspective.

The rest of the article is organized as follows. Section 2 proposes the ROC-PCA and formulates a useful framework to generalize the M -estimators to robust subspace estimation. In Section 3, a class of computational algorithms involving Stiefel manifold optimization and iterative thresholding is proposed. Section 4 theoretically analyzes the performance of ROC-PCA in finite samples. In Section 5, we point out a pitfall of applying a popular SVD reduction in high-dimensional robust PCA and propose a *batch* variant of ROC-PCA for big data computation. Section 6 presents real data analysis. We conclude in Section 7. A survey of robust PCA methods and models, an algorithm summary, simulation studies, and all technical details are left to the supplementary material.

2. Mathematical Formulation

2.1 Motivation and Model Description

The nonrobustness issue of PCA has been noticed long before and extensively investigated in robust statistics. See, for example, Maronna, Martin, and Yohai (2006) for a comprehensive introduction. In this work, we classify the robust PCA approaches into five classes: (i) *robust covariance matrix*-based methods, such as Maronna (1976) and Rousseeuw (1985); (ii) *projection*-based methods, for example, Li and Chen (1985) and Hubert,

Rousseeuw, and Verboven (2002); (iii) *hybrid* projection-covariance estimation-based methods (Hubert, Rousseeuw, and Branden 2005); (iv) *spherical/elliptical* PCA (Locantore et al. 1999); and (v) *low-rank matrix approximation*-based methods, Croux et al. (2003), Candès et al. (2011), and Zhou et al. (2010), among others. Due to space limitations, we give a more detailed literature review in the supplementary material.

Let X be an $n \times p$ data matrix with n observations in p -dimensional space. Assume no outliers exist for now and $V^o \in \mathbb{R}^{p \times r}$ consists of the top r ideal PC loading vectors. Then $P_{V^o} = V^o V^{oT}$ defines the r -dimensional PC subspace. Recall the characterization of PCA via low-rank matrix approximation: $\min_B \|X - B\|_F^2$ s.t. $\text{rank}(B) \leq r$. The optimal B must lie in the PC subspace, that is, $BP_{V^o} = B$. Decomposing X into XP_{V^o} and $X(I - P_{V^o})$ (the projections of X onto the PC subspace and OC subspace, respectively), the objective function becomes

$$\|X - B\|_F^2 = \|XP_{V^o} - B\|_F^2 + \|X(I - P_{V^o})\|_F^2. \quad (1)$$

Now suppose outliers do exist, in the situation of which the above may result in a misleading B -estimate, due to the nonrobust nature of the ℓ_2 loss function. The robustification of the first term $\|XP_{V^o} - B\|_F^2$ is related to the PC outliers (see Figure 1(a)), and no matter what robust loss is chosen, one can always set $B^o = (XV^o)V^{oT}$ to satisfy the low-rank constraint. Thus, the first term always vanishes in optimization, regardless of the choice of the loss. In contrast, the second term $\|X(I - P_{V^o})\|_F^2$ is independent of B , and its robustification is to address pertinent outliers in the OC subspace (see Figure 1(b)). Therefore, the crux in robust PC estimation lies in incorporating OC outliers into the second term.

Motivated by this, we introduce a *projected* mean-shift outlier model:

$$XV_{\perp}^* = \mathbf{1}\mu^{*T} + S^* + E. \quad (2)$$

We use $d := p - r$ throughout this article. V_{\perp}^* is a $p \times d$ matrix satisfying $V_{\perp}^{*T} V_{\perp}^* = I$, and XV_{\perp}^* gives the coordinates after projecting the data onto the OC subspace. In model (2), XV_{\perp}^* is decomposed into three parts: mean, outlier, and noise. Concretely, (i) $\mathbf{1}\mu^{*T}$ stands for the mean term, where $\mathbf{1} = [1, 1, \dots, 1]^T \in \mathbb{R}^n$ and μ^* is a d -dimensional mean vector for the transformed observations; (ii) $S^* = [s_1^*, \dots, s_n^*]^T = [s_{i,j}^*] \in \mathbb{R}^{n \times d}$ is the outlier matrix, describing the outlyingness of each observation or entry; and (iii) finally, the noise term E has iid

$\mathcal{N}(0, \sigma^2)$ entries (or sub-Gaussian entries that may be dependent). The goal is to recover μ^* , S^* , and V_\perp^* jointly. The problem for small r is seemingly more challenging, because of the increased dimensionality of the subspace where “bad” OC outliers can occur.

Assume that S is sparse (because outliers should not be the norm), then shrinkage estimation can be used:

$$\min_{V_\perp, \mu, S} \ell(V_\perp, \mu, S) = \frac{1}{2} \|\mathbf{XV}_\perp - \mathbf{1}\mu^T - S\|_F^2 + P(S; \lambda) \quad \text{s.t.} \quad V_\perp^T V_\perp = \mathbf{I}, \quad (3)$$

where $P(S; \lambda)$ stands for a general sparsity-promoting penalty (or constraint) with λ as the regularization parameter. Hereinafter, the study of (3) is referred to as *robust orthogonal complement principal component analysis* (ROC-PCA). Different from PCP, where the sparsity is pursued in the raw observation space, ROC-PCA introduces sparsity after projecting the data points onto the OC subspace, and \mathbf{SV}_\perp^T is not necessarily sparse. As will be shown later, ROC-PCA provides a robust guarantee for estimating the OC subspace (and therefore the PC subspace) and can identify the outliers simultaneously. The regularizations through rank reduction and sparsity make ROC-PCA applicable to $p \gg n$ datasets.

There are two main ways of enforcing sparsity in S , corresponding to element-wise (e-Type) outliers and row-wise (r-Type) outliers, respectively. The e-Type ROC-PCA is defined as

$$\min_{(V_\perp, \mu, S)} \frac{1}{2} \|\mathbf{XV}_\perp - \mathbf{1}\mu^T - S\|_F^2 + \sum_{ij} P(|s_{ij}|; \lambda_{ij}) \quad \text{s.t.} \quad V_\perp^T V_\perp = \mathbf{I}. \quad (4)$$

P can take various forms (possibly nonconvex), such as $P(S; \lambda) = \sum_{ij} \lambda_{ij} |s_{ij}|$ or $\lambda \|S\|_1$ when $\lambda = \lambda_{ij}$. This popular ℓ_1 penalty (Tibshirani 1996) is however well known to suffer from biased estimation and inconsistent selection (Zou and Hastie 2005; Zhao and Yu 2006). Moreover, convex penalties have limited power in dealing with multiple gross outliers with high leverage values (She and Owen 2011). One nonconvex alternative is the ℓ_0 penalty $P(S; \lambda) = \sum_{ij} (\lambda_{ij}^2/2) 1_{s_{ij} \neq 0}$ or $(\lambda^2/2) \|S\|_0$ when $\lambda = \lambda_{ij}$. Some fusion penalties, such as SCAD (Fan and Li 2001) and Hard-Ridge (She 2012), can also be applied. Similarly, to address outliers in a row-wise manner, we introduce the r-Type ROC-PCA

$$\min_{(V_\perp, \mu, S)} \frac{1}{2} \|\mathbf{XV}_\perp - \mathbf{1}\mu^T - S\|_F^2 + \sum_i P(\|s_i\|_2; \lambda_i) \quad \text{s.t.} \quad V_\perp^T V_\perp = \mathbf{I}, \quad (5)$$

where s_i^T is the i th row vector of S . All element-wise penalties can be adapted to promote group sparsity, for example, $\lambda \|S\|_{2,1}$ with $\|S\|_{2,1} \triangleq \sum_i \|s_i\|_2$, and $(\lambda^2/2) \|S\|_{2,0}$ with $\|S\|_{2,0} \triangleq \sum_i 1_{s_i \neq 0}$. Classic robust statistics pays special attention to r-Type outliers, while e-Type outliers may arise from a fully independent multivariate contamination model—interested readers can refer to Alqallaf et al. (2009) for more details.

2.2 ROC-PCA as Generalized M-Estimators

ROC-PCA is derived by the use of the *additive* robustification scheme of She and Owen (2011). The conventional way to achieve robust estimation is through modifying the Frobenius norm loss, or using the M -estimators. In this subsection, we first generalize the M -estimators to robust PC subspace estimation (a manifold setting), and then build a universal connection between ROC-PCA and such generalized M -estimators.

We begin by reviewing the definition of the M -estimators in linear regression $y = \mathbf{X}\beta + \epsilon$ with $y = [y_1, \dots, y_n]^T$ and $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times p}$. The ρ -type M -estimator is defined to be a stationary point of $\sum_{i=1}^n \rho(y_i - \mathbf{x}_i^T \beta)$, and the (more general) ψ -type M -estimator is defined to be a solution to the equation $\mathbf{X}^T \psi(y - \mathbf{X}\beta) = \mathbf{0}$, where ψ , not necessarily a derivative function, is applied componentwise. In our ROC-PCA setting, replacing the ℓ_2 loss function with a robust loss ρ leads to robust PC subspace recovery:

$$\min_{(\mu, V_\perp)} \sum_{i=1}^n \sum_{j=1}^d \rho((\mathbf{XV}_\perp - \mathbf{1}\mu^T)_{ij}; \lambda) \quad \text{s.t.} \quad V_\perp^T V_\perp = \mathbf{I}, \quad (6)$$

where λ is a parameter of the loss function. For a general ρ , the optimization with respect to μ and V_\perp could be difficult.

A more useful ψ -type M -estimator for PC subspace recovery is defined as follows. To motivate the definition, we assume $\psi = \rho'$, and view (6) as an *unconstrained* optimization problem on the manifold $\Omega := \mathbb{R}^d \times \mathbb{O}^{p \times d}$, where \mathbb{R}^d denotes a d -dimensional Euclidean manifold, and $\mathbb{O}^{p \times d}$ represents a Stiefel manifold, the set of all $p \times d$ matrices V_\perp satisfying the orthogonality constraint $V_\perp^T V_\perp = \mathbf{I}$. The derivative of the loss with respect to μ is given by $\mathbf{1}^T \psi(\mathbf{XV}_\perp - \mathbf{1}\mu^T; \lambda)$. The trickier part is to define the gradient on the Stiefel manifold $\mathbb{O}^{p \times d}$. Equipped with the *canonical metric* (see Section 3.1), we calculate the Riemannian gradient of $\sum_{i=1}^n \sum_{j=1}^d \rho((\mathbf{XV}_\perp - \mathbf{1}\mu^T)_{ij}; \lambda)$ with respect to V_\perp (details given in the supplement): $\mathbf{X}^T \psi(\mathbf{XV}_\perp - \mathbf{1}\mu^T; \lambda) - V_\perp (\psi(\mathbf{XV}_\perp - \mathbf{1}\mu^T; \lambda))^T \mathbf{XV}_\perp$. Now, given any ψ function, the generalized (ψ -type) M -estimator $(\hat{\mu}, \hat{V}_\perp)$ for ROC-PCA is defined as a solution to the following equations:

$$\begin{cases} \mathbf{1}^T \psi(\mathbf{XV}_\perp - \mathbf{1}\mu^T; \lambda) = \mathbf{0}, \\ \mathbf{X}^T \psi(\mathbf{XV}_\perp - \mathbf{1}\mu^T; \lambda) - V_\perp (\psi(\mathbf{XV}_\perp - \mathbf{1}\mu^T; \lambda))^T \mathbf{XV}_\perp = \mathbf{0}. \end{cases} \quad (7)$$

Interestingly, there is a universal connection between ROC-PCA and the generalized M -estimation.

Theorem 1. (i) Let $\Theta(\cdot; \lambda)$ be an arbitrarily given thresholding rule (see Section 3.2), and P be any penalty associated with Θ such that

$$P(t; \lambda) - P(0; \lambda) = \int_0^{|t|} (\sup\{s : \Theta(s; \lambda) \leq u\} - u) du + q(t; \lambda), \quad (8)$$

for some nonnegative $q(\cdot; \lambda)$ satisfying $q(\Theta(s; \lambda); \lambda) = 0 \forall s \in \mathbb{R}$. Suppose $(\hat{V}_\perp, \hat{\mu}, \hat{S})$ is a coordinate-wise minimum of (4) and Θ is continuous at $(\mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^T) \mathbf{X} \hat{V}_\perp + \frac{1}{n} \mathbf{1}\mathbf{1}^T \hat{S}$. Then $(\hat{V}_\perp, \hat{\mu})$ is necessarily a robust generalized M -estimator associated with ψ , where $\psi(t; \lambda) = t - \Theta(t; \lambda)$, $\forall t$. (ii) Given $0 \leq q < nd$, let $(\hat{V}_\perp, \hat{\mu}, \hat{S})$ be any coordinate-wise minimum of $\min_{(V_\perp, \mu, S)} \frac{1}{2} \|\mathbf{XV}_\perp - \mathbf{1}\mu^T - S\|_F^2$ s.t. $V_\perp^T V_\perp = \mathbf{I}$, $\|S\|_0 \leq q$.

Then, after dropping \hat{S} , $(\hat{V}_\perp, \hat{\mu})$ is also a minimizer of $\frac{1}{2} \sum_{k=1}^{nd-q} r_{(k)}^2$ s.t. $V_\perp^T V_\perp = I$, $R = XV_\perp - \mathbf{1}\mu^T$, where $r_{(1)}, \dots, r_{(nd)}$ are the order statistics of the elements of R satisfying $|r_{(1)}| \leq \dots \leq |r_{(nd)}|$.

The theorem shows the correspondence between the e-Type ROC-PCA estimator and the element-wise form of the generalized M -estimator (7). This universal connection provides some guidance in choosing P , too. For example, redescending ψ -functions are recommended in robust statistics to deal with gross outliers. They correspond to nonconvex penalties. The conclusion can be easily extended to the r-Type outlier case (5), on the basis of She (2012).

On the other hand, ROC-PCA differs from the generalized M -estimation in some significant ways. First, without the explicit introduction of S , the generalized M -estimators cannot reveal outliers inherently. A cut-off value for the residuals (which are not independent) has to be chosen. In contrast, based on the sparsity pattern of \hat{S} , ROC-PCA explicitly labels all outliers. Second, the λ in the generalized M -estimation is a loss parameter, the tuning of which is usually based on large- n asymptotics or worst-case studies (e.g., the breakdown point); in (3), λ is a regularization parameter to control the bias-variance trade-off, and is easy to be tuned in a data-dependent manner. Third, the ψ function in (7) may be nonsmooth or even discontinuous, but the quadratic objective function in (3) is smooth in V_\perp . Therefore, the optimization of V_\perp in ROC-PCA can be much easier and less computationally expensive. (For example, second-order derivative information can be possibly used to develop faster algorithms.) Finally, the design of ρ is most suitable and effective in robustifying the squared error loss, while the additive robustification, by introducing a sparse shift outlier term, naturally extends to other loss functions, such as Bernoulli, Poisson, hinge loss, and others.

2.3 Estimation of PC Directions

Once \hat{V}_\perp is obtained, the PC subspace estimate is given by $\hat{P} = I - \hat{V}_\perp \hat{V}_\perp^T$. This suffices in many PCA applications, such as data visualization. Sometimes, one may want to obtain each individual PC direction ordered in terms of importance. Under the assumption that only pure OC outliers exist, simply applying SVD to $X\hat{P}$ completes the task.

On the other hand, if one suspects that OC outliers and PC outliers coexist, a robust PCA method can be further applied to $X\hat{P}$. Here, ROC-PCA also offers some computational benefits. Indeed, because $r \ll p$ and $X\hat{P}$ is free of OC outliers, a rank- r SVD reduction (Section 5) can be safely performed before running robust PCA, to reduce time and space complexity. Alternatively, one can adopt a *sequential* ROC-PCA scheme to extract the most important PC directions. First, apply ROC-PCA to $X_1 := X$ with the resultant robust rank-1 PC subspace denoted by \hat{P} . A spectral decomposition on \hat{P} yields \hat{v}_1 . Then ROC-PCA can be repeatedly applied to the deflated matrix $X_k = X_{k-1} - X_{k-1}\hat{v}_{k-1}\hat{v}_{k-1}^T$ to get the rest PC directions \hat{v}_k ($2 \leq k \leq r$).

3. Computation

The computation of ROC-PCA defined in (3) is challenging due to the orthogonality constraint, in addition to the nonsmooth and possibly nonconvex P . In this section, we develop an alternating optimization algorithm based on Stiefel manifold optimization and iterative nonlinear thresholdings.

3.1 V_\perp -Optimization

Given μ and S , minimizing ℓ (see (3)) with respect to V_\perp reduces to

$$\min_{V_\perp} f(V_\perp) = \frac{1}{2} \|\mathbf{X}V_\perp - \mathbf{1}\mu^T - S\|_F^2 \quad \text{s.t.} \quad V_\perp^T V_\perp = I. \quad (9)$$

There are many ways of solving the problem. Our goal is to design a fast algorithm even in high dimensions.

Instead of treating (9) as a constrained optimization problem by introducing a few Lagrangian multipliers, we view it as an *unconstrained* optimization problem on the Stiefel manifold $\mathbb{O}^{p \times d} := \{V_\perp \in \mathbb{R}^{p \times d} : V_\perp^T V_\perp = I\}$, to take advantage of the smoothness of f in V_\perp . Optimization on the Stiefel manifold requires preserving the orthogonality constraint in updating V_\perp . Our updating scheme is based on the idea of *retraction*, which smoothly maps the tangent space $\mathcal{T}_{V_\perp}(\mathbb{O}^{p \times d}) := \{\Delta \in \mathbb{R}^{p \times d} : V_\perp^T \Delta + \Delta^T V_\perp = 0\}$ (\mathcal{T}_{V_\perp} for notational simplicity) onto the Stiefel manifold $\mathbb{O}^{p \times d}$, see, for example, Absil, Mahony, and Sepulchre (2008).

We begin by defining a Riemannian gradient of f with respect to V_\perp , denoted by ∇f . Following Edelman, Arias, and Smith (1998), we adopt the *canonical metric* $g_c(\Delta, \Delta) := \text{tr}(\Delta^T (I - \frac{1}{2} V_\perp V_\perp^T) \Delta)$. The Riemannian gradient ∇f is then defined as the unique element in \mathcal{T}_{V_\perp} such that $g_c(\nabla f, \Delta) = \text{tr}(G^T \Delta)$ for any $\Delta \in \mathcal{T}_{V_\perp}$, where G denotes the Euclidean gradient of f with respect to V_\perp , that is, $G_{ij} = \frac{\partial f(V_\perp)}{\partial V_{\perp ij}}$. It is not difficult to show that

$$\begin{aligned} \nabla f &= W V_\perp, \quad \text{with } W = G V_\perp^T - V_\perp G^T, \\ G &= X^T (X V_\perp - \mathbf{1}\mu^T - S). \end{aligned} \quad (10)$$

A valid updating scheme should guarantee that the new trial point lies on the manifold. Let $V_\perp(\tau)$ be a function determining the new trial point with τ as the step size. We use a Cayley transformation-based update due to Wen and Yin (2010):

$$V_\perp(\tau) = \left(I + \frac{\tau}{2} W\right)^{-1} \left(I - \frac{\tau}{2} W\right) V_\perp. \quad (11)$$

It can be verified that the curve generated by (11) always lies on the manifold for any τ , and $V_\perp(\tau)$ is a descent curve passing the point $V_\perp(0) = V_\perp$. Yet the inversion of the $p \times p$ matrix $(I + \frac{\tau}{2} W)$ in (11) may be expensive when p is large. When $d < p/2$, one can write $W = A_1 A_2^T$ with $A_1 = [G, V_\perp]$ and $A_2 = [V_\perp, -G]$, and apply the matrix inversion formula to get $V_\perp(\tau) = V_\perp - \tau A_1 (I + \tau A_2^T A_1 / 2)^{-1} A_2^T V_\perp$ (see Wen and Yin 2010, Lemma 4). This fast-update formula involves the inversion of a $2d \times 2d$ matrix, and turns out to be pretty useful in the design of batch ROC-PCA in Section 5. In the case of $d \geq p/2$,

one possible idea is to approximate \mathbf{W} by the product of two low-rank matrices (Wen and Yin 2010, Lemma 5).

It remains to specify a proper step size τ to guarantee the convergence and efficiency in large problems. We use a *nonmonotone* line search scheme together with Barzilai–Borwein step-size (BB), (Barzilai and Borwein 1988) and (Raydan 1997). In comparison with other commonly used inexact line searches, BB does not guarantee descent in function value at each step, but results in quick convergence and performs well in large-scale nonlinear optimization (Zhang and Hager 2004; Dai and Fletcher 2005; Zhou, Gao, and Dai 2006). In addition, the nonmonotone search scheme only performs backtracking occasionally, and thus saves a lot of computational time. (Be aware that cost of generating a trial point on the manifold is not cheap.)

In more detail, the BB calculation at the k th iteration requires solving $\min_{\tau_k} \|\tau_k^{-1} \delta_k(\mathbf{V}_\perp) - \delta_k(\nabla f)\|_F^2$ and $\min_{\tau_k} \|\delta_k(\mathbf{V}_\perp) - \tau_k \delta_k(\nabla f)\|_F^2$, with $\delta_k(\mathbf{V}_\perp) = \mathbf{V}_\perp^{(k)} - \mathbf{V}_\perp^{(k-1)}$ and $\delta_k(\nabla f) = \nabla f(\mathbf{V}_\perp^{(k)}) - \nabla f(\mathbf{V}_\perp^{(k-1)})$. This leads to $\tau_k^0 = \frac{\text{tr}(\delta_k(\mathbf{V}_\perp)^T \delta_k(\mathbf{V}_\perp))}{|\text{tr}(\delta_k(\mathbf{V}_\perp)^T \delta_k(\nabla f))|}$ and $\tau_k^1 = \frac{|\text{tr}(\delta_k(\mathbf{V}_\perp)^T \delta_k(\nabla f))|}{\text{tr}(\delta_k(\nabla f)^T \delta_k(\nabla f))}$, respectively. The two solutions are used alternatively in odd and even numbered iterations. Because of the nonmonotonic behavior of BB, Raydan's adaptive nonmonotone search scheme is applied to ensure global convergence. That is, compute the stepsize $\tau^{(k)} = \kappa^m \tau_k^i$ ($i = 0$ for even k and $i = 1$ otherwise), where $\kappa \in (0, 1)$ and m_k is the smallest integer satisfying

$$f(\mathbf{V}_\perp^{(k)}(\tau^{(k)})) \leq \max_{0 \leq j \leq \min(k, T)} f(\mathbf{V}_\perp^{(k-j)}) + \rho \tau_k f'(\mathbf{V}_\perp^{(k)}(0)). \quad (12)$$

This criterion uses T most recent function values. It is easy to get $f'(\mathbf{V}_\perp(0)) := \frac{\partial f(\mathbf{V}_\perp(\tau))}{\partial \tau} \Big|_{\tau=0} = \text{tr}\{(\frac{\partial f(\mathbf{V}_\perp(\tau))}{\partial \mathbf{V}_\perp(\tau)})^T (\frac{\partial \mathbf{V}_\perp(\tau)}{\partial \tau})\} \Big|_{\tau=0} = -\text{tr}\{\mathbf{G}^T (\mathbf{G} \mathbf{V}_\perp^T - \mathbf{V}_\perp \mathbf{G}^T) \mathbf{V}_\perp\} = -\frac{1}{2} \|\mathbf{W}\|_F^2$, where \mathbf{W} is calculated according to (10). In practice, we recommend $\kappa = 0.1$, $T = 10$, and $\rho = 1e - 3$.

How to choose the starting point is important. Our initialization of $\mathbf{V}_\perp^{(0)}$ uses the multi-start strategy by Rousseeuw and Van Driessen (1999). First generate m_0 candidate $\mathbf{V}_\perp^{(0)}$ at random; starting with each, run the computational algorithm for n_0 iterations; pick the best m_1 candidates (evaluated by the cost function value) and continue the algorithm till convergence. The final estimate $\hat{\mathbf{V}}_\perp$ is the one that delivers the minimal cost function value. For example, in implementation of r-Type ROC-PCA, we use $m_0 = 10$, $n_0 = 2$, and $m_1 = 2$.

3.2 (μ, \mathbf{S}) -Optimization

Fixing \mathbf{V}_\perp , (3) reduces to $\min_{\mu, \mathbf{S}} g(\mu, \mathbf{S}) = \frac{1}{2} \|\mathbf{X} \mathbf{V}_\perp - \mathbf{1} \mu^T - \mathbf{S}\|_F^2 + P(\mathbf{S}; \lambda)$. The optimization for μ is an ordinary least square (OLS) problem with the solution given by $\mu^o = \frac{1}{n} (\mathbf{X} \mathbf{V}_\perp - \mathbf{S})^T \mathbf{1}$. The \mathbf{S} -optimization involves sparsity-inducing penalties (element-wise or row-wise), which are nondifferentiable and possibly nonconvex (corresponding to redescending ψ -type M -estimators).

To give a general algorithmic framework, we solve the problem from the viewpoint of thresholding rules. A thresholding rule, denoted by Θ , is defined to be an *odd monotone*

unbounded shrinkage function (She 2009). Given any Θ , its vector or matrix version (still denoted by Θ) is defined componentwise. For any $\mathbf{s} \in \mathbb{R}^d$, the *multivariate* version of Θ , denoted by $\vec{\Theta}(\mathbf{s}; \lambda)$, is defined to be $\frac{\mathbf{s}}{\|\mathbf{s}\|_2} \Theta(\|\mathbf{s}\|_2; \lambda)$ if $\mathbf{s} \neq \mathbf{0}$ and otherwise $\mathbf{0}$ (see She 2012). For any $\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_n]^T \in \mathbb{R}^{n \times d}$, $\vec{\Theta}(\mathbf{S}; \lambda) = [\vec{\Theta}(\mathbf{s}_1; \lambda), \dots, \vec{\Theta}(\mathbf{s}_n; \lambda)]^T$.

Given an arbitrary thresholding rule Θ , and let P be any function satisfying (8) with $q(t; \lambda)$ nonnegative and $q(\Theta(\mathbf{s}; \lambda); \lambda) = 0$ for all $\mathbf{s} \in \mathbb{R}$. Then the minimization problem $\min_{\mathbf{S}} \frac{1}{2} \|\mathbf{Y} - \mathbf{S}\|_F^2 + \sum_i P(\|\mathbf{s}_i\|_2; \lambda)$ has a globally optimal solution $\mathbf{S}^o = \vec{\Theta}(\mathbf{Y}; \lambda)$. Similarly, a globally optimal solution to $\min_{\mathbf{S}} \frac{1}{2} \|\mathbf{Y} - \mathbf{S}\|_F^2 + \sum_{ij} P(|s_{ij}|; \lambda)$ is $\mathbf{S}^o = \Theta(\mathbf{Y}; \lambda)$. See Lemma 1 in She (2012) for a justification (where the “continuity assumption” of Θ at \mathbf{Y} is not needed because we do not require the uniqueness of \mathbf{S}^o). Starting with various thresholding rules, (8) covers all commonly used penalties, including ℓ_1 , ℓ_0 , ℓ_p ($0 < p < 1$), “ $\ell_0 + \ell_2$ ” (She 2012) and so on. Based on such Θ - P coupling, a general iterative algorithm for updating μ and \mathbf{S} can be designed, which is illustrated below for the r-Type problem:

repeat

$$\mu^{(k)} \leftarrow \frac{1}{n} (\mathbf{X} \mathbf{V}_\perp - \mathbf{S}^{(k)})^T \mathbf{1}$$

$$\mathbf{S}^{(k+1)} \leftarrow \vec{\Theta}(\mathbf{X} \mathbf{V}_\perp - \mathbf{1}(\mu^{(k)})^T; \lambda)$$

$$k \leftarrow k + 1$$

until $\|\mathbf{S}^{(k)} - \mathbf{S}^{(k-1)}\|$ is small enough

Clearly, $\mu^{(k)}$ does not have to be explicitly calculated: $\mathbf{S}^{(k+1)} \leftarrow \vec{\Theta}((\mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}^T) \mathbf{X} \mathbf{V}_\perp + \frac{1}{n} \mathbf{1} \mathbf{1}^T \mathbf{S}^{(k)}; \lambda)$. A summary of the complete algorithm for the r-Type ROC-PCA is shown in the supplementary material. Simply replacing $\vec{\Theta}$ with its componentwise version Θ gives the e-Type ROC-PCA algorithm. This alternating optimization guarantees the function-value decreasing property at each step, $g(\mu^{(k)}, \mathbf{S}^{(k)}) \geq g(\mu^{(k)}, \mathbf{S}^{(k+1)}) \geq g(\mu^{(k+1)}, \mathbf{S}^{(k+1)})$ for any $k \geq 0$.

Progressive quantile thresholding-based iterative screening. The penalty parameter λ in (3) adjusts the bias-variance trade-off, but is inconvenient if one wants to specify its value directly. Here, we propose some constrained forms of ROC-PCA to address the issue. For the r-Type outliers, consider

$$\begin{aligned} \min_{(\mathbf{V}_\perp, \mu, \mathbf{S})} & \frac{1}{2} \|\mathbf{X} \mathbf{V}_\perp - \mathbf{1} \mu^T - \mathbf{S}\|_F^2 + \frac{\eta}{2} \|\mathbf{S}\|_F^2 \\ \text{s.t.} & \mathbf{V}_\perp^T \mathbf{V}_\perp = \mathbf{I}, \|\mathbf{S}\|_{2,0} \leq q, \end{aligned} \quad (13)$$

where, in addition to the ridge penalty to account for large noise and clustered outliers (collinearity), the group ℓ_0 constraint is imposed on \mathbf{S} rather than a penalty. Similarly, $\|\mathbf{S}\|_0 \leq q^e$, gives the constrained form of the e-Type ROC-PCA. They extend the least trimmed squares (LTS; Rousseeuw and Leroy 1987) due to Part (ii) of Theorem 1. Unless otherwise specified, we use the constrained forms of ROC-PCA in computer experiments. Compared with the penalty parameter λ in (5) or (4), q (or q^e), as an upper bound of the number of outliers, is both meaningful and intuitive in robust analysis. Nicely, q is not a sensitive parameter to subspace recovery, as long as it is within a reasonable range (see Section 3.1 of the supplement). The ridge shrinkage parameter η is even more insensitive and its search grid can be small—in implementation, we simply fix η at a small value, say, $1e-3$.

The constrained ROC-PCA shares the same \mathbf{V}_\perp -optimization with the penalized form. As for the \mathbf{S} -optimization, fortunately,

we can adapt the Θ -estimators to this subproblem via a *quantile* thresholding rule $\Theta^\#(\cdot; q^e, \eta)$. For any $\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_n] \in \mathbb{R}^{n \times d}$ and $1 \leq q^e \leq nd$, $\Theta^\#(\mathbf{S}; q^e, \eta)$ shrinks the top q^e largest entries (in absolute value) of \mathbf{S} by a factor of $1 + \eta$, and sets all remaining entries to be 0. A *multivariate* version $\tilde{\Theta}^\#$ to be used for the constrained r-Type ROC-PCA problem is defined as $\tilde{\Theta}^\#(\mathbf{S}; q, \eta) = \text{diag}\{\Theta^\#(g(\mathbf{S}); q, \eta)\} \mathbf{S}^o$, where $g(\mathbf{S}) = [\|\mathbf{s}_i\|_2]_{n \times 1}$ and $\mathbf{S}^o = (\text{diag}\{g(\mathbf{S})\})^+ \mathbf{S}$ with $^+$ standing for the Moore-Penrose pseudoinverse. Now, for the r-Type problem, we can use $\tilde{\Theta}^\#$ in place of $\tilde{\Theta}$ and run $\tilde{\mathbf{S}}^{(k+1)} \leftarrow \tilde{\Theta}^\#((\mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}^T) \mathbf{X} \mathbf{V} + \frac{1}{n} \mathbf{1} \mathbf{1}^T \tilde{\mathbf{S}}^{(k)}; q, \eta)$. The \mathbf{S} -update not only guarantees the nonincreasing of the function value but satisfies $\|\mathbf{S}\|_{2,0} \leq q$ (She, Wang, and Wu 2013). To lessen greediness, we advocate *progressive* quantile-thresholding-based iterative screening. Define a monotone sequence of integers $\{q(k)\}$ decreasing from n to the target value q . At the k th iteration, the above quantile parameter q is now replaced by $q(k)$. Empirically, $q(k) = \max(q, 2n/(1 + e^{\nu k}))$ with $\nu = 0.05$ gives a fast and accurate cooling scheme.

4. Nonasymptotic Analysis of ROC-PCA

The finite-sample performance of ROC-PCA is of great theoretical interest. Due to the equivalence established in Theorem 1, it is not difficult for one to show some asymptotics under the classic setting where $n \rightarrow \infty$ and r, p are fixed, as well as the (nonstochastic) breakdown point properties of ROC-PCA. Nevertheless, we wish to perform large- p or even nonasymptotic robust analysis to meet the challenge of modern statistical applications. Our tool for such theoretical studies is the *oracle inequalities* (Donoho and Johnstone 1994). We take a predictive learning perspective and study the data approximation power of ROC-PCA. Let the model be $\mathbf{X} = \mathbf{A}^* \mathbf{V}^{*T} + \mathbf{S}^* \mathbf{V}_\perp^{*T} + \mathbf{E}$, where $\mathbf{A}^* \in \mathbb{R}^{n \times r}$, $\mathbf{S}^* \in \mathbb{R}^{n \times d}$ with $d = p - r$, $[\mathbf{V}^*, \mathbf{V}_\perp^*] \in \mathbb{O}^{p \times p}$. Assume all entries of \mathbf{E} are iid Gaussian $\sim \mathcal{N}(0, \sigma^2)$ (or sub-Gaussian, as in the supplement). In this section, we ignore the intercept term for simplicity and suppose the outlier matrix \mathbf{S}^* is row-wise sparse. The problem of interest can be formulated by

$$\begin{aligned} \min_{\mathbf{S}, \mathbf{V}_\perp, \mathbf{A}, \mathbf{V}} \|\mathbf{X} - \mathbf{A} \mathbf{V}^T - \mathbf{S} \mathbf{V}_\perp^T\|_F^2 \text{ s.t.} \\ \|\mathbf{S}\|_{2,0} \leq q, [\mathbf{V}, \mathbf{V}_\perp] \in \mathbb{O}^{p \times p}. \end{aligned} \quad (14)$$

This is a rephrasing of ROC-PCA. Indeed, the loss can be written in a separable form $\|\mathbf{X} \mathbf{V} - \mathbf{A}\|_F^2 + \|\mathbf{X} \mathbf{V}_\perp - \mathbf{S}\|_F^2$ and so the optimization with respect to \mathbf{S} and \mathbf{V}_\perp corresponds to (13). On the other hand, with $(\hat{\mathbf{S}}, \hat{\mathbf{V}}_\perp)$ available, the optimal $\hat{\mathbf{A}} \hat{\mathbf{V}} = \mathbf{X}(\mathbf{I} - \hat{\mathbf{V}}_\perp \hat{\mathbf{V}}_\perp^T)$ can be obtained afterward.

Given any $(\hat{\mathbf{A}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}, \hat{\mathbf{V}}_\perp)$, define its *mean* approximation error by $M(\hat{\mathbf{A}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}, \hat{\mathbf{V}}_\perp; \mathbf{A}^*, \mathbf{S}^*, \mathbf{V}^*, \mathbf{V}_\perp^*) = \frac{1}{np} \|\hat{\mathbf{A}} \hat{\mathbf{V}}^T + \hat{\mathbf{S}} \hat{\mathbf{V}}_\perp^T - \mathbf{A}^* \mathbf{V}^{*T} - \mathbf{S}^* \mathbf{V}_\perp^{*T}\|_F^2$. or $M(\hat{\mathbf{A}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}, \hat{\mathbf{V}}_\perp)$ when there is no ambiguity. The approximation error is always meaningful in evaluating the performance of an estimator, without requiring any signal strength assumption.

Theorem 2. Let $(\hat{\mathbf{A}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}, \hat{\mathbf{V}}_\perp)$ be any globally optimal point of (14). Then, the following oracle inequality holds for any $(\mathbf{A}, \mathbf{S}, \mathbf{V}, \mathbf{V}_\perp)$ satisfying $\|\mathbf{S}\|_{2,0} \leq q$, $\mathbf{A} \in \mathbb{R}^{n \times r}$, $[\mathbf{V}, \mathbf{V}_\perp] \in$

$\mathbb{O}^{p \times p}$:

$$\begin{aligned} \mathbb{E} \left[M(\hat{\mathbf{A}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}, \hat{\mathbf{V}}_\perp) \right] \lesssim M(\mathbf{A}, \mathbf{S}, \mathbf{V}, \mathbf{V}_\perp) \\ + P_o(q, r) + \frac{\sigma^2}{np}, \end{aligned} \quad (15)$$

where $P_o(q, r)$ is short for $P_o(q, r; n, p, \sigma^2) = \sigma^2 \{qp + rn + rp + q \log(en/q)\}/(np)$, and \lesssim denotes an inequality that holds up to a multiplicative numerical constant.

The theorem applies to \mathbf{E} with $\text{vec}(\mathbf{E})$ being sub-Gaussian (which includes bounded random matrices, and allows for column and/or row dependencies). Equation (15) is in expectation form; a high probability result with the same error rate P_o can be obtained as well (without the last additive term $\sigma^2/(np)$). See the supplementary material for proof details. The result is nonasymptotic in nature and applies to any r, q, n, p . Note that it does not require any incoherence condition that is commonly assumed in the literature.

According to (15), a sharp risk upper bound is obtained by taking the infimum of the right-hand side over the set of all valid reference signals $(\mathbf{A}, \mathbf{S}, \mathbf{V}, \mathbf{V}_\perp)$. First, with $\mathbf{S} = \mathbf{S}^*$, $\mathbf{A} = \mathbf{A}^*$, $\mathbf{V} = \mathbf{V}^*$, such that the first term $M(\cdot)$ vanishes, we can get an error rate of order $\sigma^2 \{q^* p + r^* n + r^* p + q^* \log(en/q^*)\}/(np)$ (which is optimal in a minimax sense). But our conclusion holds more generally—in particular, \mathbf{S}^* does not have to be exactly sparse. Indeed, when \mathbf{S}^* contains many small but nonzero entries, a reference \mathbf{S} with much reduced support can benefit from the bias-variance trade-off to attain a lower bound than simply taking $\mathbf{S} = \mathbf{S}^*$. In other words, the obtained oracle inequality ensures the ability of ROC-PCA in dealing with mild outliers, which is of great practical interest.

A by-product is the finite-sample breakdown property. First, define the finite-sample breakdown point for an arbitrary estimator $(\hat{\mathbf{A}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}, \hat{\mathbf{V}}_\perp)$ in terms of its **risk**: Given a finite data matrix \mathbf{X} and an estimator $(\hat{\mathbf{A}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}, \hat{\mathbf{V}}_\perp)(\mathbf{X})$, abbreviated as $\hat{\mathbf{B}}$, its breakdown point is $\epsilon^*(\hat{\mathbf{B}}) = \frac{1}{n} \cdot \min\{q \in \mathbb{Z}^+ : \sup_{\mathbf{X} \in \mathcal{B}(q)} \mathbb{E}[M(\hat{\mathbf{B}}; \mathbf{B})] = +\infty\}$, where $\mathbb{Z}^+ = \mathbb{N} \cup \{0\}$, $\mathcal{B}(q) = \{\mathbf{X} \in \mathbb{R}^{n \times p} : \mathbf{X} = \mathbf{B} + \mathbf{E}, \text{ where } \mathbf{B} = \mathbf{A} \mathbf{V}^T + \mathbf{S} \mathbf{V}_\perp^T, \text{ vec}(\mathbf{E}) \text{ is sub-Gaussian, } \mathbf{A} \in \mathbb{R}^{n \times r}, \|\mathbf{S}\|_{2,0} \leq q, [\mathbf{V}, \mathbf{V}_\perp] \in \mathbb{O}^{p \times p}\}$. Note that the randomness of $\hat{\mathbf{B}}$ is accounted by taking the expectation. It follows from (15) that $\epsilon^*(\hat{\mathbf{B}}) \geq (q + 1)/n$.

Furthermore, we show that in a minimax sense, the error rate obtained in Theorem 2 is essentially optimal. Consider the following signal class

$$\begin{aligned} \mathcal{S}(r, q) = \{(\mathbf{A}^*, \mathbf{S}^*, \mathbf{V}^*, \mathbf{V}_\perp^*) : \mathbf{A}^* \in \mathbb{R}^{n \times r}, \\ [\mathbf{V}^*, \mathbf{V}_\perp^*] \in \mathbb{O}^{p \times p}, \|\mathbf{S}^*\|_{2,0} \leq q\}, \end{aligned} \quad (16)$$

where $1 \leq q \leq n$, $1 \leq r \leq n \wedge p$. Let $\ell(\cdot)$ be a nondecreasing loss function with $\ell(0) = 0$, $\ell \not\equiv 0$.

Theorem 3. Assume $\mathbf{X} = \mathbf{A}^* \mathbf{V}^{*T} + \mathbf{S}^* \mathbf{V}_\perp^{*T} + \mathbf{E}$ with $e_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, $n \geq 2$, $1 \leq q \leq n$, $1 \leq r \leq n \wedge p$, $r(n + p - r) \geq 8$, $qd \geq 8$. Then there exist positive constants C, c (depending

on $\ell(\cdot)$ only) such that

$$\inf_{(\hat{A}, \hat{S}, \hat{V}, \hat{V}_\perp)} \sup_{(A^*, S^*, V^*, V_\perp^*) \in \mathcal{S}(r, q)} \mathbb{E}[\ell(M(\hat{A}, \hat{S}, \hat{V}, \hat{V}_\perp; A^*, S^*, V^*, V_\perp^*) / (CP_o(q, r)))] \geq c > 0, \quad (17)$$

where $(\hat{A}, \hat{S}, \hat{V}, \hat{V}_\perp)$ denotes an arbitrary estimator of $(A^*, S^*, V^*, V_\perp^*)$ and $P_o(q, r) = P_o(q, r; n, p, \sigma^2) = \sigma^2\{qd + rn + rp + q \log(en/q)\} / (np)$.

We give some examples of ℓ to illustrate the conclusion. Using the indicator function $\ell(u) = 1_{u \geq 1}$, we learn that for any estimator $(\hat{A}, \hat{S}, \hat{V}, \hat{V}_\perp)$, $M(\hat{A}, \hat{S}, \hat{V}, \hat{V}_\perp; A^*, S^*, V^*, V_\perp^*) \gtrsim \sigma^2\{r(p+n) + qd + q \log(en/q)\} / (np)$ occurs with positive probability. For $\ell(u) = u$, the theorem shows that the risk $\mathbb{E}[M(\hat{A}, \hat{S}, \hat{V}, \hat{V}_\perp; A^*, S^*, V^*, V_\perp^*)]$ is bounded from below by the same rate up to some multiplicative constant. Therefore, ROC-PCA can essentially achieve the minimax optimal rate nonasymptotically.

Various asymptotic results can be obtained from the finite-sample bound. In fact, as long as $np \gg qp + rn + rp + q \log(en/q)$, the approximation error tends to zero. In real-life applications, the values of q (the number of outliers) and r (the number of principal components) of interest are typically very small (even as small as 2 or 1), in which case the proposed ROC-PCA, exploiting the parsimony offered by low rankness and sparsity, has guaranteed small error in theory.

5. Batch ROC-PCA

Modern applications call for the need of scalable algorithms in high dimensions. Unfortunately, most methods reviewed in the supplementary material suffer from heavy computational burden when directly applied to $p \gg n$ datasets (and some may fail in principle). A widely acknowledged practice in the robust PCA literature is to perform an SVD reduction beforehand (see, e.g., Hubert, Rousseeuw, and Verboven 2002). Nevertheless, we found that such a preprocessing may be unreliable and nonrobust when p is very high. In this section, we discuss it in details and propose a batch variant of ROC-PCA to meet the challenge.

SVD reduction. People usually conduct an SVD reduction in advance before applying robust PCA to high-dimensional data: given X with all its columns properly centered, obtain its top q right singular vectors in V^o , with $q = \text{rank}(X)$ typically, and form $\tilde{X}_{n \times q} = X_{n \times p} V_{p \times q}^o$. Then, apply robust PCA on \tilde{X} with the resultant estimate denoted by $\tilde{V} \in \mathbb{O}^{q \times r}$. In the end, $V^o \tilde{V}$ is reported as the estimated PC directions. In such a procedure, the computational burden of robust PCA can be significantly reduced.

The SVD reduction is commonly believed to be valid for robust PCA and does not cause any information loss (from which it appears that the challenge of high dimensionality is not that serious). But it just amounts to a rank- q PCA. In fact, even assuming (ideally) that the true column means can be accurately and robustly estimated, the obtained directions may be misleading when $p \gg n$ and/or in the presence of OC outliers.

If the back-transformed estimate $V^o \tilde{V}$ coincides with the authentic loading matrix V^* , then the PC subspace must lie in

the observed row space, namely, $P_{V^*} \subseteq P_{V^o}$. Hence, the belief in the reduction is that as long as Xu is $\mathbf{0}$, or approximately so, u should not contain much information about P_{V^*} (or deserve to be checked for OC outliers).

Let us consider a toy example with the i th row of X^* given by $[a_i, 0, \dots, 0]$, and the i th row of the corrupted matrix X given by $[a_i, \epsilon a_i, \dots, \epsilon a_i]$, where ϵ is set small enough. Then, we have $X\alpha = \mathbf{0}$ for $\alpha = [1, -\frac{1}{\epsilon(p-1)}, \dots, -\frac{1}{\epsilon(p-1)}]^T$. With p ultra-high, α and $[1, 0, \dots, 0]^T$ determine nearly the same projection, indicating that the true PC subspace essentially lies in the orthogonal complement of the observed row space! Accordingly, simply applying the SVD reduction is questionable and the curse of dimensionality is nontrivial. This perhaps surprising finding is closely connected to Johnstone and Lu (2009). It is easy to show that the existence of OC outliers only makes this phenomenon much more severe.

Therefore, we caution against such a plain PCA-based dimension reduction in ultra-high dimensional problems (with possible OC outliers). On the other hand, a reduction can be safely made in the OC subspace with the help of ROC-PCA.

Batch ROC-PCA. We propose a batch ROC-PCA (BROC-PCA) to speed the computation. The basic idea is to estimate V_\perp in a batch fashion. Each time, identify only m ($m < d$) least significant OC loading vectors. By setting $m \ll p/2$, the inversion formula-based update in Section 3.1 can be effectively used. Moreover, a rank- $(p-m)$ SVD reduction can be performed afterward as the reduced m least significant dimensions contain no PC information. This step makes the problem size drop after each batch processing.

Concretely, given $X_1 := X$ and a series of batch sizes m_k ($1 \leq k \leq K$) satisfying $\sum_{k=1}^K m_k = d$, the BROC-PCA procedure is as follows. For each k , apply ROC-PCA to X_k with the resultant estimate denoted by $V_{\perp, k}$, containing m_k least significant OC loading vectors of X_k . Form an intermediate matrix $Z = X_k(I - V_{\perp, k} V_{\perp, k}^T)$, and obtain $X_{k+1} = ZV_k$, where V_k is the top $(p - \sum_{i=1}^k m_i)$ right singular vectors of Z . Finally, the product of all V_1, \dots, V_K is delivered as the PC directions estimate, that is, $\hat{V} := \prod_{k=1}^K V_k$. An attractive feature is that the number of columns of X_k gets smaller as k increases. In implementation, to take advantage of the fast update formula, we recommend choosing m_k satisfying $m_k < (p - \sum_{i=0}^{k-1} m_i)/2$ (assuming $m_0 = 0$) unless the problem size is sufficiently small. A rule of thumb in large- p computation is $30 \leq m_k \leq 100$. To attain further speedup, we adopt a progressive error control scheme—the error tolerance used in the ROC-PCA algorithm is gradually tightened up from the computation of the first batch to the K th batch.

BROC-PCA shares similarity with some sparse PCA algorithms, for example, Shen and Huang (2008). However, instead of repeatedly solving the rank-1 problem, BROC-PCA estimates m_k OC loading vectors at each time; more importantly, the SVD reduction is then employed to reduce the dimensionality by m_k . Accordingly, the overall computational cost can be substantially reduced (by about 70% for $p = 1000$).

6. Numerical Experiments

We performed extensive simulation studies to show the performance of ROC-PCA in the presence of r-Type/e-Type outliers.

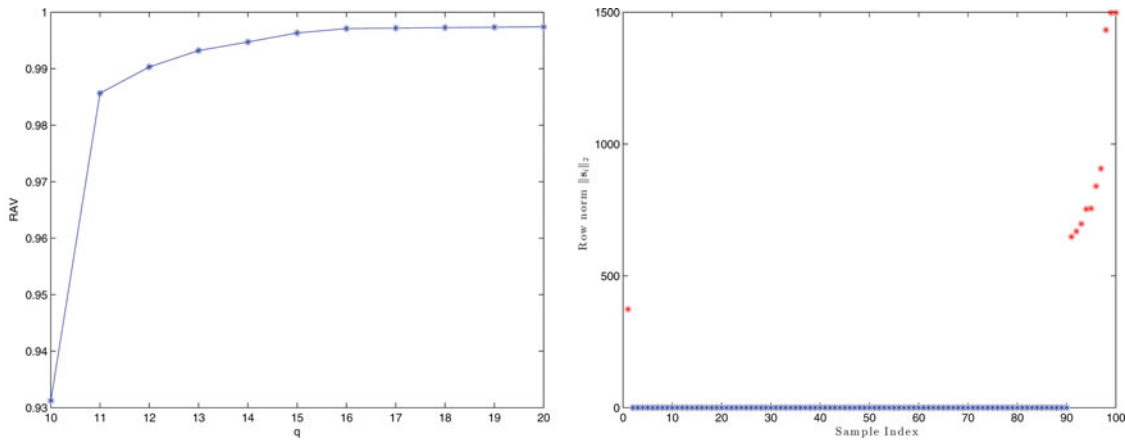


Figure 2. Left: RAV versus q (with $r = 3$) on the segmentation dataset. Right: Outlyingness plot: row norms of \hat{S} . All nonzeros are marked in red. The last 10 correspond to the foliage outliers, while the first one is from the cement class.

Due to page limitations, the results are reported in the supplementary material.

We also applied ROC-PCA to analyze a segmentation dataset collected by Brodley (Lichman 2013), which contains features extracted from seven classes of hand-segmented images (brick-face, sky, foliage, cement, window, path, and grass). Each class has 330 image regions, and for each image region 19 features are provided (e.g., the contrast of vertically or horizontally adjacent pixels). In this experiment, we randomly picked 90 samples from the cement class as normal observations and 10 from the foliage class as outliers.

In PCA applications, the adjusted variance (Shen and Huang 2008) is often used to assess the goodness of fit. We use a robust version, called robust adjusted variance (RAV), to take outliers into account. Let \hat{V}_r be a robust estimate of the top r loading vectors from data matrix X . The RAV explained by \hat{V}_r is then defined as $\|X^0 P_{\hat{V}_r}\|_F^2 / \|X^0\|_F^2$, where X^0 is a submatrix of X containing clean samples only.

We ran the r-Type algorithm with $r = 1, 2, 3$ and q decreasing from 20 to 10. Three principal components seem to be enough from the high RAV percentages in Table 1. The left panel of Figure 2 shows a relatively big drop in RAV when q changes from 11 to 10. ROC-PCA thus yielded 11 outliers rather than 10. To get some intuition, we also plotted in the right panel the outlyingness for each observation, that is, row norms of \hat{S} (with $q = 11$). The last 10 samples in the foliage class were successfully identified and most observations in the cement class have \hat{s}_i exactly zero. Interestingly, the first sample pops up with a large row norm of about 400(!) in the outlyingness plot.

Table 1. Explained variance in terms of $\text{RAV} \times 100$, by the top 1–3 PCs. The RAV evaluation of the first six rows is based on the cement class, and ROC-PCA at the bottom is after removing all its detected outliers.

	1 PC	2 PCs	3 PCs
PCA	5	20	66
PCP	13	54	57
RAPCA	45	71	80
S-PCA	48	75	81
ROBPCA	48	75	82
ROC-PCA ⁰	48	76	82
ROC-PCA	58	91	99

We examined the data carefully and verified this finding—for example, the 7th feature for the first observation takes value 375.1, while the other 89 samples in the same class show an average of only 1.8. The pleasant finding shows the power of ROC-PCA in automatic outlier detection without much human intervention.

We also compared ROC-PCA with PCA, PCP (Candès et al. 2011), S-PCA (Locantore et al. 1999), RAPCA (Hubert, Rousseeuw, and Verboven 2002), and ROBPCA (Hubert, Rousseeuw, and Branden 2005) on the segmentation dataset. Among these methods only ROC-PCA and PCP labeled outliers explicitly. The outlying entries detected by PCP scatter all over the matrix, and the zero/nonzero pattern provides little help in separating the foliage samples from the majority of the cement data. Table 1 gives the RAV rates explained by the top 1–3 PCs. The first six rows evaluated RAV on the 90 cement samples. The true performance of ROC-PCA, as shown in the last row, was assessed on the 89 clean observations in consideration of its outlier detection. The improvement bought by simultaneous subspace recovery and outlier identification is significant.

7. Conclusion

We showed that PCA is sensitive to a type of outliers that may not be easily revealed in the original observation space; mathematically formulating these projected outliers gave rise to the robust orthogonal complement PCA. We showed that ROC-PCA comes with a robust guarantee as generalized robust M-estimators, and provides ease in computation and regularization. Our theoretical analyses revealed the high breakdown point of ROC-PCA and established a nonasymptotic oracle inequality that can achieve the minimax error rate.

Some future research topics include further studies of the performance of the manifold optimization algorithm, as well as the development of faster algorithms in very high dimensions. Another interesting direction is to jointly investigate observation anomalies, which are of independent interest in many computer vision applications, and OC outliers (which can skew the PC subspace) to provide a reinforced robustification of PCP.

Supplementary Materials

The supplement materials provide a literature survey of robust PCA, an overall summary of the computational algorithm, simulation studies, and all technical details.

Acknowledgment

The authors thank the editor, the associate editor, and two anonymous referees for their careful comments and useful suggestions that significantly improve the quality of the article.

Funding

This work was supported in part by NSF grants CCF-1117012, CCF-1116447, and DMS-1352259.

References

- Absil, P.-A., Mahony, R., and Sepulchre, R. (2008), *Optimization Algorithms on Matrix Manifolds*, Princeton, NJ: Princeton University Press. [766]
- Alqallaf, F., Van Aelst, S., Yohai, V., and Zamar, R. (2009), "Propagation of Outliers in Multivariate Data," *The Annals of Statistics*, 37, 311–331. [765]
- Barzilai, J., and Borwein, J. M. (1988), "Two-Point Step Size Gradient Methods," *IMA Journal of Numerical Analysis*, 8, 141–148. [767]
- Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011), "Robust Principal Component Analysis?" *Journal of the ACM*, 58, 11:1–11:37. [763,764,770]
- Croux, C., Filzmoser, P., Pison, G., and Rousseeuw, P. J. (2003), "Fitting Multiplicative Models by Robust Alternating Regressions," *Statistics and Computing*, 13, 23–36. [764]
- Dai, Y.-H., and Fletcher, R. (2005), "Projected Barzilai-Borwein Methods for Large-Scale Box-Constrained Quadratic Programming," *Numerische Mathematik*, 100, 21–47. [767]
- Donoho, D., and Johnstone, I. (1994), "Ideal Spatial Adaptation Via Wavelet Shrinkages," *Biometrika*, 81, 425–455. [768]
- Edelman, A., Arias, T. A., and Smith, S. T. (1998), "The Geometry of Algorithms With Orthogonality Constraints," *SIAM Journal on Matrix Analysis and Applications*, 20, 303–353. [766]
- Fan, J., and Li, R. (2001), "Variable Selection Via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [765]
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (2011), *Robust Statistics: The Approach Based on Influence Functions* (Vol. 114), New York: Wiley. [763]
- Hubert, M., Rousseeuw, P., and Branden, K. (2005), "Robpca: A New Approach to Robust Principal Component Analysis," *Technometrics*, 47, 64–79. [763,764,770]
- Hubert, M., Rousseeuw, P., and Verboven, S. (2002), "A Fast Method for Robust Principal Components With Applications to Chemometrics," *Chemometrics and Intelligent Laboratory Systems*, 60, 101–111. [764,769,770]
- Johnstone, I. M., and Lu, A. Y. (2009), "On Consistency and Sparsity for Principal Components Analysis in High Dimensions," *Journal of the American Statistical Association*, 104, 682–693. [769]
- Li, G., and Chen, Z. (1985), "Projection-Pursuit Approach to Robust Dispersion Matrices and Principal Components: Primary Theory and Monte Carlo," *Journal of the American Statistical Association*, 80, 759–766. [764]
- Lichman, M. (2013), *UCI Machine Learning Repository*, Irvine, CA: University of California, Irvine, School of Information and Computer Sciences. [770]
- Locantore, N., Marron, J., Simpson, D., Tripoli, N., Zhang, J., Cohen, K., Boente, G., Fraiman, R., Brumback, B., Croux, C., Fan, J., Kneip, A., Marden, J. I., Peñá, D., Prieto, J., Ramsay, J. O., Valderrama, M. J., Aguilera, A. M., Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., and Cohen, K. L. (1999), "Robust Principal Component Analysis for Functional Data," *Test*, 8, 1–73. [763,764,770]
- Maronna, R. A. (1976), "Robust M-Estimators of Multivariate Location and Scatter," *The Annals of Statistics*, 4, 51–67. [764]
- Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006), *Robust Statistics*, Chichester, UK: Wiley. [763,764]
- Peng, Y., Ganesh, A., Wright, J., Xu, W., and Ma, Y. (2012), "Rasl: Robust Alignment by Sparse and Low-Rank Decomposition for Linearly Correlated Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34, 2233–2246. [763]
- Raydan, M. (1997), "The Barzilai and Borwein Gradient Method for the Large Scale Unconstrained Minimization Problem," *SIAM Journal on Optimization*, 7, 26–33. [767]
- Rousseeuw, P., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, 41, 212–223. [763,767]
- Rousseeuw, P. J. (1985), "Multivariate Estimation With High Breakdown Point," *Mathematical Statistics and Applications*, B, 283–297. [764]
- Rousseeuw, P. J., and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*. (Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics), New York: Wiley. [767]
- She, Y. (2009), "Thresholding-Based Iterative Selection Procedures for Model Selection and Shrinkage," *Electronic Journal of Statistics*, 3, 384–415. [767]
- (2012), "An Iterative Algorithm for Fitting Nonconvex Penalized Generalized Linear Models With Grouped Predictors," *Computational Statistics & Data Analysis*, 56, 2976–2990. [765,766,767]
- She, Y., Li, H., Wang, J., and Wu, D. (2013), "Grouped Iterative Spectrum Thresholding for Super-Resolution Sparse Spectrum Selection," *IEEE Transactions on Signal Processing*, 61, 6371–6386. [768]
- She, Y., and Owen, A. (2011), "Outlier Detection Using Nonconvex Penalized Regression," *Journal of the American Statistical Association*, 106, 626–639. [765]
- Shen, H., and Huang, J. Z. (2008), "Sparse Principal Component Analysis Via Regularized Low Rank Matrix Approximation," *Journal of Multivariate Analysis*, 99, 1015–1034. [769,770]
- Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288. [765]
- Wen, Z., and Yin, W. (2010), "A Feasible Method for Optimization With Orthogonality Constraints," *Mathematical Programming*, 142, 1–38. [766]
- Wright, J., Ganesh, A., Min, K., and Ma, Y. (2013), "Compressive Principal Component Pursuit," *Information and Inference*, 2, 32–68. [763]
- Wright, J., Ganesh, A., Rao, S., Peng, Y., and Ma, Y. (2009), "Robust Principal Component Analysis: Exact Recovery Of Corrupted Low-Rank Matrices by Convex Optimization," in *Proceedings of Neural Information Processing Systems* (Vol. 3), pp. 2080–2088. [763]
- Xu, H., Caramanis, C., and Sanghavi, S. (2010), "Robust PCA Via Outlier Pursuit," in *Advances in Neural Information Processing Systems*, pp. 2496–2504. [763]
- Zhang, H., and Hager, W. W. (2004), "A Nonmonotone Line Search Technique and Its Application to Unconstrained Optimization," *SIAM Journal on Optimization*, 14, 1043–1056. [767]
- Zhang, Z., Ganesh, A., Liang, X., and Ma, Y. (2012), "Tilt: Transform Invariant Low-Rank Textures," *International Journal of Computer Vision*, 99, 1–24. [763]
- Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *The Journal of Machine Learning Research*, 7, 2541–2563. [765]
- Zhou, B., Gao, L., and Dai, Y.-H. (2006), "Gradient Methods With Adaptive Step-Sizes," *Computational Optimization and Applications*, 35, 69–86. [767]
- Zhou, Z., Li, X., Wright, J., Candès, E., and Ma, Y. (2010), "Stable Principal Component Pursuit," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*, pp. 1518–1522. [763,764]
- Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society, Series B*, 67, 301–320. [765]