Robust Principal Component Analysis: A Median of Means Approach
Debolina Paul, Saptarshi Chakraborty, and Swagatam Das

Abstract—Principal component analysis (PCA) is a fundamental tool for data visualization, denoising, and dimensionality reduction. It is widely popular in statistics, machine learning, computer vision, and related fields. However, PCA is well-known to fall prey to outliers and often fails to detect the true underlying low-dimensional structure within the dataset. Following the Median of Means (MoM) philosophy, recent supervised learning methods have shown great success in dealing with outlying observations without much compromise to their large sample theoretical properties. This article proposes a PCA procedure based on the MoM principle. Called the MoMPCA, the proposed method is not only computationally appealing but also achieves optimal convergence rates under minimal assumptions. In particular, we explore the nonasymptotic error bounds of the obtained solution via the aid of the Rademacher complexities while granting absolutely no assumption on the outlying observations. The derived concentration results are not dependent on the dimension because the analysis is conducted in a separable Hilbert space, and the results only depend on the fourth moment of the underlying distribution in the corresponding norm. The proposal's efficacy is also thoroughly showcased through simulations and real data applications.

Index Terms—Median of Means (MoM), nonasymptotic error bounds, principal component analysis (PCA), robust dimensionality reduction, robust PCA.

I. INTRODUCTION

Principal component analysis (PCA) [1], [2] is perhaps the most well-known statistical method for linear dimensionality reduction [3]. Given a set of (mean-centered) points \( \mathcal{X} = \{x_1, \ldots, x_N\} \subset \mathbb{R}^p \), the PCA computes a small number of orthonormal basis vectors, which characterize most of the variability within the data cloud. Mathematically, to find a \( d \)-dimensional \( (d \leq p) \) representation of \( \mathcal{X} \), one projects these \( N \) points on to a \( d \)-dimensional subspace as \( \{Qx_1, \ldots, Qx_N\} \), where \( Q \) is a \( p \times p \) projection matrix of rank \( d \). Denoting the set of all \( p \times p \) projection matrices of rank \( d \) as \( Q_d \), the PCA minimizes the following objective:

\[
\frac{1}{N} \sum_{i=1}^{N} \|x_i - Qx_i\|^2 = \frac{1}{N} \sum_{i=1}^{N} x_i^\top (I - Q)x_i,
\]

where \( Q \in Q_d \).

Though efficient in many real-life scenarios, PCA is well documented to suffer from many disadvantages such as information loss, poor performance when the data lie in a nonaffine manifold, poor performance for high-dimensional data, etc. Researchers continue to tackle these issues through modifications to the original versions, such as probabilistic PCA [4], kernel PCA [5], sparse PCA [6], and robust PCA [7], [8]. Among many disadvantages of PCA, like information loss and poor interpretation, one of the major concerns is that it is well known to fail in the presence of a single outlying observation. Its fragility against severely corrupted data points often puts its reliability at risk. Popular approaches for making the PCA more robust to these outlying and corrupted observations attempt to represent the original data matrix \( x \) as a sum of a low-rank matrix \( L \) and a sparse matrix \( S \) [7], [9], [10], [11], [12], [13], [14]. However, many methods mentioned earlier only allow recovery guarantees under very restrictive assumptions. For example, Candès et al. [7] restricts the structure of \( S \) to a Bernoulli model.

Robust PCA has recently gained popularity, primarily due to the rise of Big Data. These big datasets often contain outliers and hence require delicate handling, usually done by implementing robust statistics against outlying observations. Recently, there has been much interest in the robust dimensionality reduction community on proposing efficient frameworks for PCA [15], [16], [17]. There has also been an array of work based on geometric median-based approaches [18], [19], [20], [21], which focuses on a robust estimation of the covariance matrix and finding the eigendecomposition of the same to find the proper subspace to project. Although efficient in practice, many techniques mentioned above do not come with finite-sample theoretical guarantees. Those with asymptotic rates often assume that the data points are independent and identically distributed (i.i.d.), which does not hold when the data are contaminated with outliers.

To bridge this methodological gap, the Median of Means (MoM) literature provides a promising and attractive framework to adapt PCA to become outlier robust, as well as help us preserve our theoretical understanding of finite-sample error bounds. As opposed to the classical Vapnik–Chervonenkis
empirical risk minimization (ERM) [22], the MoM philosophy provides a more robust framework for efficiently finding estimates of the actual underlying parameter. Although MoM estimators have been in the literature for quite a long time, it has recently been introduced to the machine learning community [23], [24], [25], [26]. Besides being insensitive to outliers, the MoM estimators also possess a solid theoretical backbone comprising exponential concentration results under the mild restriction of finite variance [23], [24], [25], [27], [28]. Recently, several near-optimal results were established from this perspective concerning regression [23], [29], bandits [30], mean estimation [31], clustering [32], [33], [34], classification [26], and optimal transport [35].

In this article, we develop a PCA in the framework of the MoM principle. The proposal is not only computationally efficient but also theoretically appealing. It is well known that the theoretical understanding of many classical ERM [22] such as PCA hinges on the assumptions that the data should be i.i.d. and have sub-Gaussian behavior. However, real data that may be corrupted with outliers do not offer us the luxury to make such simplifying assumptions. Toward our theoretical investigation, we assume that the dataset can be split into two categories: the set of inliers (I) and the set of outliers (O), i.e., \{1, \ldots, N\} = I \cup O. The data points in I are assumed to be independently distributed according to the distribution \(P\), which has a finite fourth moment. We make no assumptions on the points in O; thus, allowing them to be dependent, unboundedly large, having distributions that are entirely dissimilar to \(P\), allowing them to be heavy-tailed, etc. Our theoretical analysis hinges on the application of Rademacher complexity [36], and symmetrization arguments [22], [37]. From a theoretical viewpoint, we emphasize that our analyses are derived under more general and interpretable conditions compared to the literature [26]. To further generalize the proposed setup, we conduct all our theoretical analyses in a separable Hilbert space. This allows us to derive dimension-free bounds that only depend on the data distribution through the fourth moment of the inliers. We quickly recover the data results in a finite-dimensional vector space showing that the derived rates match the state-of-the-art [24], [34].

The main contributions of this article can be summarized as follows.

1) This article proposes a simple yet efficient framework for robust PCA under the MoM paradigm. Apart from being practically efficient, the method comes with a strong theoretical backbone of finite-sample error rates under the mild assumption of the existence of a finite fourth moment of the underlying data distribution. Such consistency guarantees and error rates that are derived subsequently, ensure that the results are reliable and accurate under mild assumptions.

2) Furthermore, the derived generalization bounds are dimension-free meaning that the error rate is also valid for infinite-dimensional Hilbert spaces.

3) We emphasize that as opposed to many prominent works [7], [38], we only assume that the number of outliers is \(o(N)\), which is the natural definition of outliers. We do not make any assumptions on the distribution of the outliers allowing them to be dependent, unboundedly large, having distributions that are entirely dissimilar to the inlier distribution, etc.

4) The detailed experimental analysis via simulation and real-life datasets on background modeling in video and anomaly detection demonstrates the efficacy of Median of Means Principal Component Analysis (MoMPCA) compared to the state-of-the-art in different experimental settings, indicating that the proposed method is highly effective in practice.

The rest of this article is organized as follows. In Section II, we formulate the MoMPCA, followed by a detailed theoretical analysis under minimal and interpretable assumptions in Section III. The experimental results are discussed in Section IV, followed by concluding remarks in Section V.

Notations: Before we proceed further, we discuss a few notations used in this article. Vectors are denoted with bold lower-case letters, and matrices are denoted with bold upper-case letters. \(\langle A, B \rangle = \text{trace}(A^T B)\) denotes the Frobenious inner product between two matrices. \(\|A\| = (\langle A, A \rangle)^{1/2}\) denotes the Frobenious norm of the matrix \(A\). \(\mathbb{P}(E)\) denotes the probability of the event \(E\) and \(\mathbb{E}_Z(\cdot)\) denotes the expectation with respect to the random vector \(Z\). A random variable \(\sigma\) is said to be Rademacher if it takes values in \(+1, -1\) with equal probability. \([N]\) denotes the set \{1, \ldots, N\}, for any \(N \geq 0\) and \(2^A\) denotes the power set of \(A\). For any two sets \(A\) and \(B\), \(A^B\) denotes the set of all functions from \(B\) to \(A\). \(\mu_k = \int \|x\|^k dP\) denotes the kth moment of \(P\). \(Q_d\) denotes the set of all \(p \times p\) real projection matrices of rank \(d\). \(\mathbb{E}_{x \sim P}(\cdot)\) denotes the indicator function.

A. Related Works

This section discusses some of the works related to robust clustering. The most noted work in this direction is arguably the work by Candès et al. [7], which appeals to the philosophy of writing the data matrix as a sum of a low rank and a sparse matrix and minimizing the subsequent approximation error. Tang and Nehorai [15] extended this idea by developing a convex program for a low-rank and block-sparse matrix decomposition. A survey on these papers can be found in [16]. Liu et al. [39] developed a new subspace selection method called angle linear discriminant embedding (ALDE) for dimensionality reduction in a supervised learning setting. Angle PCA proposed by Wang et al. [17] develops an iterative algorithm to minimize an \(\ell_2\) norm-based error and maximize the summation of the ratio between the variance and the reconstruction error to preserve rotational invariance while tackling outliers. Liu and Yu [40] further extended these ideas to develop angular embedding for conducting robust PCA. Hauberg et al. [41], [42] introduced Grassman averages to express dimensionality reduction as an average of the subspaces spanned by the data. The authors further exploit its properties to develop a robust Grassman average as a form of robust PCA. Bouwmans et al. [43] showed the application of robust PCA in image and video processing. Other than the median of the means PCA method, there are other types of robust PCA methods as well, such as the optimal
mean methods [44], [45] and avoid mean calculation [46]. There has also been an array of work based on geometric median-based approaches [18], [19], [20], [21], which focuses on a robust estimation of the covariance matrix and finding the eigendecomposition of the same to find the proper subspace to project.

B. Motivation and a Proof of Concept Result

Robust PCA is commonly formulated by assuming that the data matrix can be decomposed into a low-rank signal component and a noise component. However, this framework is limited by the assumption that the effect of outliers is mild and cannot grow unboundedly large. To address this issue, the authors propose using the MoM framework, which is commonly used in robust mean estimation, to view traditional PCA as an ERM problem. Because the objective cannot be solved using the eigen-decomposition trick, the authors propose an alternative method called projected Adagrad. The authors note that the most challenging aspect of the project was extending the theoretical results to an infinite-dimensional Hilbert space and deriving dimension-free bounds under only the existence of a finite fourth moment of the data distribution. This result not only provides meaningful conclusions with minimal assumptions but also paves the way for future research on robust kernel-based methods.

As a motivating example, in Fig. 1, we show the results of classical PCA and MoMPCA on a toy dataset (available at https://github.com/SaptarshiC98/MOMPCA). The dataset has 1000 observations, out of which 10 of them, i.e., 1% are outliers. The inliers are generated from a 2-D Gaussian random variable with variance 10 in the first dimension, 1 in the second dimension, and covariance 8. The outliers are generated from another 2-D Gaussian random variable with mean vector (15, 50), variance 5 in each dimension with zero covariance between them. Finally, we run the vanilla PCA (shown in red) and the proposed MoMPCA (shown in blue) and plot their first principal components. From Fig. 1, it is clear that a mere 1% outlying observations are enough to render the vanilla PCA ineffective, whereas, MoMPCA correctly identifies the direction of maximum variation. In terms of excess risk (details provided in Section IV-A), we calculate that the excess risk for vanilla PCA is 20.6850 where, in comparison, the extra risk for MoMPCA is only 0.3104, proving its efficacy.

II. PCA WITH MEDIAN OF MEANS

A. Problem Setup

We will follow the same notations as described in Section I. One should note that the PCA objective function (1) can be written as $$\int \xi_v(x) P_N(dx) = (1/N) \sum_{i=1}^{N} \xi_v(x_i)$$, with $$\xi_v(x) = x^\top (I - Q)x$$. Here, $$P_N = (1/N) \sum_{i=1}^{N} \delta_{x_i}$$ denotes the empirical measure, based on $$X$$, i.e., $$P_N(A) = (1/N) \sum_{i=1}^{N} \mathbb{1}_{\{x_i \in A\}}$$ for any set $$A$$.

MoM does not directly minimize the empirical risk (1). Instead, it starts with a partitioning of the data into $$L$$ groups $$B_1, \ldots, B_L$$ (i.e., $$B_l \in 2^L$$, $$\bigcup_{l=1}^{L} B_l = [n]$$ and $$B_l \cap B_{l'} = \emptyset$$ for $$l \neq l'$$), each of which contains exactly $$B$$ many elements (repudiating a few observations when $$L$$ does not divide $$n$$).

B. Optimization

Optimizing (2) is tractable via gradient-based methods. It is easy to see that we can write $$Q = V V^\top$$, where $$V$$ is a $$p \times d$$ real column-orthonormal matrix. Thus, making a variable transform, we can write, (2) as

$$\text{MoM}_{\text{L}}^V(g_V) = \frac{1}{B} \text{Median} \left( \sum_{i \in B_l} g_V(x_i), \ldots, \sum_{i \in B_L} g_V(x_i) \right)$$

where $$g_V(x) = x^\top (I - V V^\top) x$$. At the $$r$$th iteration, let the median partition be $$\ell^{med} = \ell_r$$, i.e., $$\ell_r \in [L]$$ be such that $$\text{MoM}^r_{\text{L}}(g_{V^{(r)}}) = \frac{1}{B} \sum_{i \in B_{\ell^{med}}} g_{V^{(r)}}(x_i)$$. According to [24], $$\nabla \text{MoM}^r_{\text{L}}(g_{V^{(r)}}) = \frac{2}{B} \sum_{i \in B_{\ell^{med}}} x_i x_i^\top V^{(r)}$$. However, after taking a gradient step, $$V^{(r)} = \eta (1/B) \sum_{i \in B_{\ell^{med}}} x_i x_i^\top V^{(r)}$$, the resultant matrix may not be column orthonormal. Following [49], we apply a Gram–Schmidt algorithm to make the columns of this resultant matrix orthonormal. Let $$P_{\text{orth}}(V)$$ denote the resultant matrix after orthonormalizing the columns.
Algorithm 1 MoMPCA

Input: The dataset $\mathcal{X} = \{x_1, \ldots, x_n\}$, $L$, $d$, $\eta$.
Output: $\hat{Q}_{N,L}$.

Initialization: Randomly partition $[N]$ into $L$ many partitions of equal length.
Compute the feature-wise median $\mu$ of $\mathcal{X}$. Replace $x_i$ by $x_i - \mu$.
Compute the first $d$ eigenvectors of $\sum_{i=1}^{n} x_i x_i^T$ and stack them as columns of $V$.
repeat
Compute $v_t \leftarrow \sum_{i \in B_t} x_i (I - V^{(t)}(V^{(t)})^T) x_i$.
Compute the median class $\ell_{\text{med}}$, such that $v_{\ell_{\text{med}}} = \text{median}(\{v_t\}_{t=1}^{L})$.

$V^{(t+1)} \leftarrow P_{\text{orth}} \left( V^{(t)} - \frac{\eta}{B} \sum_{i \in B(t)_{\text{med}}} x_i x_i^T V^{(t)} \right)$.

until objective $V^{(t)}$ converges
Compute $\hat{Q}_{N,L} \leftarrow V^{(0)}(V^{(0)})^T$.

D. Time Complexity

It is a well-known fact that in a dataset with $n$ number of datapoints and $p$ many features, the time complexity of vanilla PCA is $O(\min[n^2, p^2])$. For our method, we note that the computation of $v_t$’s requires $O(kp^2(p \lor b))$ time. This is because, $v_t = \text{trace}(V^{(0)}(V^{(0)})^T \sum_{i \in B_t} x_i x_i^T)$ and requires $O(p^2b+p^2d+p^3+p) = O(p^2(b \lor p))$, taking into account all the matrix multiplications and trace computation. The median computation takes $O(L)$ time. For the computation of $V^{(t+1)}$, we note that, the gradient computation takes $O(bp^2+p^2d+p^3) = O(p^2b)$ time and the Gram–Schmidt orthogonalization takes $O(p^2d^2)$ time, making a total of $O(p^2b + p^2d^2)$ time. Thus, the per iteration complexity of the loop is $O(p^2(b \lor p) + p^2b + p^2d + L)$ time. In general, $b > p$ and the per iteration complexity becomes $O(p^2b + p^2d + L)$.

III. Theoretical Properties

A. Notation and Setup

Instead of performing the analysis on a real vector space as was done in our recent MoM-based robust clustering work [34], here we develop our proofs for a (real) Hilbert space in context to robust PCA. Suppose the dataset $H$ is a separable (real) Hilbert space. Let the norm in $H$ be denoted by $\| \cdot \|$. A linear operator $L : H \rightarrow H$ is said to be Hilbert–Schmidt if $\sum_{i \geq 1} \| L e_i \|^2 < \infty$ and the sum is independent of the chosen orthonormal basis $\{e_i\}_{i \geq 1}$ of $H$. This sets $HS(H)$ to be a Hilbert–Schmidt linear operators. An inner product on $HS(H)$ is defined as $\langle L_1, L_2 \rangle = \sum_{i \geq 1} \langle L_1 e_i, L_2 e_i \rangle$. It can be shown that $HS(H)$ is also a separable (real) Hilbert space. The trace of a linear operator is defined as $\text{tr}(L) = \sum_{i \geq 1} \langle L e_i, e_i \rangle$.

For any $x, y \in H \setminus \{0\}$, the outer product operator $x \otimes y$ is defined as $\langle x \otimes y \rangle(z) = \langle x, y \rangle$. This outer product satisfies the following properties.

1) $\| x \otimes y \|_{HS(H)} = \| x \| \| y \|$. 
2) $\text{tr}(x \otimes y) = \langle x, y \rangle$. 
3) $\langle L, x \otimes y \rangle_{HS(H)} = \langle Ly, x \rangle$.

Recall that if $U$ is an orthogonal projector operator in $H$, then $U^2 = U$. Moreover, $\| U x \|^2 = \langle x, U x \rangle \leq \| x \|^2$ and
\[ (x \otimes y, U)_{\text{HS}(\mathcal{H})} = (Ux, Uy). \] U has rank \( d < \infty \), iff it is Hilbert–Schmidt and
\[
\|U\|^2_{\text{HS}(\mathcal{H})} = d; \quad \text{tr}(U) = d.
\]

Suppose \( V \) is a closed subspace of \( \mathcal{H} \), then \( \Pi_V \) denotes the (unique) orthogonal projector onto \( V \). \( V^\perp \) represents the orthogonal complement of \( V \).

We aim to find the “best” subspace of dimension \( d \) in this Hilbert space that minimizes the reconstruction error. Hence we consider the following objective, similar to (2).

We assume that the dataset can be split into two categories: the set of inliers (\( I \)) and the set of outliers (\( O \)), i.e., \( |N| = I \cup O \). We will assume the following in the data generation process.

A1: \( \{x_i\}_{i \in I} \) are i.i.d. according to the distribution \( P \).
A2: \( \mathbb{E}_{x \sim P}(x) = 0 \) and \( \mu_d = \mathbb{E}_{x \sim P}(\|x\|^4) < \infty \).
A3: \( \exists \eta > 0 \), such that \( L > (2 + \eta)|O| \).

We emphasize that we do not make any assumptions about the distributions of the outlying observations. They are even allowed to be dependent and may even come from some heavy-tailed distributions. Assumptions A1 and A2 state that the inlying observations are independently and identically generated from some distribution with zero mean and a finite fourth moment. To see the significance of Assumption A3, one should note that by the pigeonhole principle, (strictly) less than half of the partitions may contain an outlying observation. More than half of the partitions, thus, do not contain an outlier. Since the median is driven by the majority half of the partitions, naturally by A3, since the median is only affected by the objective function value of the majority half of these partitions, the MoM estimates can be expected only to be based on the inlying observations. We note that our analysis does not require “\( L > 4|O| \)” as in [26] but only requires that \( L > (2 + \eta)|O| \), which is a much weaker condition. Moreover, opposed to [26], A3 has nice practical interpretations.

Let
\[ \xi_V(x) = \|x - \Pi_V x\|^2 = \|\Pi_V^\perp x\|^2 = (\Pi_V^\perp x, x)_{\text{HS}(\mathcal{H})}. \]

For notational simplicity, for any function \( g \), We define the operator, \( \text{MoM}^N_{\text{L}}(\cdot) \) as follows:
\[
\text{MoM}^N_{\text{L}}(g) = \frac{1}{B} \text{Median} \left( \sum_{i \in B_1} g(x_i), \ldots, \sum_{i \in B_L} g(x_i) \right).
\]

It is easy to observe that the MoMPCA problem in Section II can be restated as
\[
\min_{V \in \mathcal{V}_d} \text{MoM}^N_{\text{L}}(\xi_V).
\]

Here, \( \mathcal{V}_d \) denotes the set of all \( d \)-dimensional subspaces of \( \mathcal{H} \). Let the unique minimizer to the above problem be \( \hat{V}_d = \text{argmin}_{V \in \mathcal{V}_d} \text{MoM}^N_{\text{L}}(\xi_V) \). Let \( V^*_d = \text{argmin}_{V \in \mathcal{V}_d} P\xi_V \) be the global population minimizer. The excess risk of the estimate, \( \hat{V}_d \) is given by
\[
\mathcal{R}(\hat{V}_d) = P\xi_{\hat{V}_d} - P\xi_{V^*_d} = \int \xi_{\hat{V}_d} dP - \inf_{V \in \mathcal{V}_d} \int \xi_V dP. \quad (4)
\]

We note that \( P\xi_{V^*_d} \leq P\xi_V \) and \( \text{MoM}^N_{\text{L}}(\xi_{V^*_d}) \leq \text{MoM}^N_{\text{L}}(\xi_V) \), for all \( V \in \mathcal{V}_d \). Thus,
\[
|P\xi_{\hat{V}_d} - P\xi_{V^*_d}| \leq (P\xi_{\hat{V}_d} - \text{MoM}^N_{\text{L}}(\xi_{V^*_d})) + (\text{MoM}^N_{\text{L}}(\xi_{V^*_d}) - \text{MoM}^N_{\text{L}}(\xi_V)) - \text{MoM}^N_{\text{L}}(\xi_{V^*_d}) \leq (P\xi_{\hat{V}_d} - \text{MoM}^N_{\text{L}}(\xi_{V^*_d})) + (\text{MoM}^N_{\text{L}}(\xi_{V^*_d}) - P\xi_{V^*_d}) \leq 2 \sup_{V \in \mathcal{V}_d} |\text{MoM}^N_{\text{L}}(\xi_V) - P\xi_V|.
\]

It is thus, enough to prove bounds on the uniform deviation \( \sup_{V \in \mathcal{V}_d} |\text{MoM}^N_{\text{L}}(\xi_V) - P\xi_V| \). Toward that, we will first derive bounds on the Rademacher complexity in Section III-B, followed by our main results.

B. Bounds on the Rademacher Complexity

For our theoretical understanding of MoMPCA, we need to compute the Rademacher complexity of the function class \( \Xi_d = \{\xi_V : V \in \mathcal{V}_d\} \). We first recall the definition of Rademacher complexity. Let \( y_1, \ldots, y_m \overset{i.i.d}{\sim} P \), be random variables in some space \( S \). Suppose \( F \subseteq \mathbb{R}^d \) be a class of functions from \( S \) to \( \mathbb{R} \) and let \( \sigma_1, \ldots, \sigma_m \) be i.i.d. Rademacher random variables. The empirical Rademacher complexity, based on \( Y = \{y_1, \ldots, y_m\} \), is defined as
\[
\hat{R}_Y(F) = \frac{1}{m} \mathbb{E} \left( \sup_{f \in F} \sum_{i=1}^m \sigma_i f(y_i) \right).
\]

Similarly, the population Rademacher complexity is defined as
\[
R_m(F) = \mathbb{E}(R_Y(F)) = \frac{1}{m} \mathbb{E} \sup_{f \in F} \sum_{i=1}^m \sigma_i f(y_i).
\]

We now compute the Rademacher complexity of the function class \( \{\xi_V : V \in \mathcal{V}_d\} \) in Theorem 1.

**Theorem 1:** Suppose \( \Xi_d = \{\xi_V : V \in \mathcal{V}_d\} \). Let \( y_1, \ldots, y_m \overset{i.i.d}{\sim} P \), with \( P \) satisfying Assumption 2. Then, \( R_m(\Xi_d) \leq \sqrt{\frac{d}{m}} \).

**Proof:** Suppose \( y_1, \ldots, y_m \overset{i.i.d}{\sim} P \) and let \( \mathcal{Y} = \{y_1, \ldots, y_m\} \). The empirical Rademacher complexity is thus, given by
\[
\hat{R}_{\Xi_d}(Q) = \frac{1}{m} \mathbb{E}_\sigma \sup_{V \in \mathcal{V}_d} \sum_{i=1}^m \sigma_i \xi_V(y_i) = \frac{1}{m} \mathbb{E}_\sigma \sup_{V \in \mathcal{V}_d} \sum_{i=1}^m \sigma_i \|y_i\|_2^2 - (\Pi_V, y_i \otimes y_i)_{\text{HS}(\mathcal{H})} = \frac{1}{m} \mathbb{E}_\sigma \sup_{V \in \mathcal{V}_d} \left( |\Pi_V, \sum_{i=1}^m \sigma_i y_i \otimes y_i|_{\text{HS}(\mathcal{H})} \right) \leq \frac{1}{m} \mathbb{E}_\sigma \sup_{V \in \mathcal{V}_d} \left( |\Pi_V|_{\text{HS}(\mathcal{H})} \left( \sum_{i=1}^m \sigma_i y_i \otimes y_i \right) \right) \leq \frac{\sqrt{d}}{m} \mathbb{E}_\sigma \left( \sum_{i=1}^m \sigma_i y_i \otimes y_i \right)_{\text{HS}(\mathcal{H})}.
\]
We observe that

\[
\frac{d}{m} \sum_{i=1}^{m} \left\| y_i \right\|_2^2 = \frac{d}{m} \sum_{i=1}^{m} \left\| y_i \right\|_2^2 \leq \frac{d}{m} \sum_{i=1}^{m} \left\| y_i \right\|_4^4.
\]

Thus, inequalities (6) and (7) follow from appealing to Cauchy–Schwarz and Jensen’s inequalities, respectively. Thus,

\[
R_m(\mathcal{E}_d) = \mathbb{E}_y \hat{R}_X(\mathcal{E}_d) = \frac{1}{m} \sqrt{d} \mathbb{E} \left[ \sum_{i=1}^{m} \left\| y_i \right\|_4^4 \right] \leq \frac{1}{m} \sqrt{d} \mathbb{E} \left[ \sum_{i=1}^{m} \left\| y_i \right\|_4^4 \right] = \frac{d}{m} \mu_4.
\]

Inequality (8) follows from Jensen’s inequality. \hfill \Box

C. Uniform Concentration Bounds

We are now ready to state and prove the uniform concentration bound result in Theorem 2. Theorem 2 asserts that with a high probability, sup \(\mathcal{V}_d\) \(|\text{MoM}_{\xi}^L(\xi) - P_{\xi}^d| \leq \left( (L/N) + \delta \right)^2 / \mathbb{E}[\phi((2P-P_B)_{\xi}^d)] \)

\[
\sup_{\mathcal{V}_d} \left| \text{MoM}_{\xi}^L(\xi) - P_{\xi}^d \right| \leq C \mathbb{E}[\phi(2P-P_B)_{\xi}^d] \left( \sqrt{L/N} + \delta \left[(L/N) \right] \right)
\]

with \(C = 2 \max\left( (8(4 + \eta) / \eta) )^{1/2}, (16(4 + \eta) / \eta) \right) \). \(\mathbb{E}[\phi(2P-P_B)_{\xi}^d] \)

\[\mathbb{E}\left[ \left( \frac{2(P-P_B)_{\xi}^d}{\epsilon} \right) \right] \]

\[\mathbb{E}\left( \frac{2(P-P_B)_{\xi}^d}{\epsilon} \right) \leq \mathbb{E}\left[ \left( \frac{2(P-P_B)_{\xi}^d}{\epsilon} \right) \right] \leq 2 \mathbb{E}\left[ \left( \frac{2(P-P_B)_{\xi}^d}{\epsilon} \right) \right] \leq 4 \mathbb{E}\left[ \left( \frac{2(P-P_B)_{\xi}^d}{\epsilon} \right) \right]
\]

To bound sup \(\mathcal{V}_d\) \(\sum_{i=1}^{L} \left| \text{MoM}_{\xi}^L(\xi) - P_{\xi}^d \right| \), we first bound the quantity, \(\mathbb{E}[\phi(2(P-P_B)_{\xi}^d)/\epsilon) \). We observe that

\[
\mathbb{E}[\phi(2(P-P_B)_{\xi}^d)/\epsilon) \leq \mathbb{E}\left[ \left( \frac{2(P-P_B)_{\xi}^d}{\epsilon} \right) \right] \leq \frac{4}{\mathbb{E}\left[ \left( \frac{2(P-P_B)_{\xi}^d}{\epsilon} \right) \right]}
\]

Here (11) follows from Chebyshev’s inequality. We now concentrate on bounding the term sup \(\mathcal{V}_d\) \(\sum_{i=1}^{L} \left| \text{MoM}_{\xi}^L(\xi) - P_{\xi}^d \right| \), appealing to [50, Th. 26.5] we observe that, with probability at least \(1 - e^{-2L/4 + \eta} \), \(\forall \mathcal{V} \in \mathcal{V}_d \)

\[
\frac{1}{L} \sum_{i=1}^{L} \left| \text{MoM}_{\xi}^L(\xi) - P_{\xi}^d \right| \leq \frac{1}{L} \sum_{i=1}^{L} \left( \frac{2(P-P_B)_{\xi}^d}{\epsilon} \right) \leq 2 \mathbb{E}\left[ \left( \frac{2(P-P_B)_{\xi}^d}{\epsilon} \right) \right] + \delta.
\]

Here \(\{\sigma_t\}_{t \in \mathcal{L}}\) are independent Rademacher variables. Suppose that \(\{\xi_t\}_{t \in \mathcal{L}}\) are independent Rademacher random variables and independent of \(\{\sigma_t\}_{t \in \mathcal{L}}\). From (12), we get

\[
\frac{1}{L} \sum_{i=1}^{L} \left| \text{MoM}_{\xi}^L(\xi) - P_{\xi}^d \right| \leq \frac{1}{L} \sum_{i=1}^{L} \left( \frac{2(P-P_B)_{\xi}^d}{\epsilon} \right) \leq 2 \mathbb{E}\left[ \left( \frac{2(P-P_B)_{\xi}^d}{\epsilon} \right) \right] + \delta.
\]

Equation (13) follows from the fact that \(\varphi(\cdot)\) is 1-Lipschitz and appealing to [50, Lemma 26.9]. We now introduce a phantom sample \(\mathcal{X}' = \{x'_1, \ldots, x'_n\}\), which are i.i.d. and follows the
law $P$, independent of $\mathcal{X}$. Thus, (13) further equals the following quantity:

$$
\frac{4}{L} \mathbb{E} \left[ \sup_{v \in \mathcal{V}_d} \sum_{i \in L} \sigma_i \mathbb{E} X' \left( (P_{B_i} - P_{B_i}) \xi_v \right) \right] + \delta
\leq \frac{4}{L} \mathbb{E} \left[ \sup_{v \in \mathcal{V}_d} \sum_{i \in L} (P_{B_i}' - P_{B_i}) \xi_v \right] + \delta
\leq \frac{4}{L} \mathbb{E} \left[ \sup_{v \in \mathcal{V}_d} \sum_{i \in L} \frac{1}{B} \sum_{j \in B_i} (\xi_v(x_j) - \xi_v(x_i)) \right] + \delta
\leq \frac{4}{L} \mathbb{E} \left[ \sup_{v \in \mathcal{V}_d} \sum_{i \in L} \xi_i (\xi_v(x_j) - \xi_v(x_i)) \right] + \delta
\leq \frac{4}{L} \mathbb{E} \left[ \sup_{v \in \mathcal{V}_d} \sum_{i \in L} \xi_i (\xi_v(x_j) - \xi_v(x_i)) \right] + \delta
\leq \frac{4}{N} \mathbb{E} \left[ \sup_{v \in \mathcal{V}_d} \sum_{i \in \mathcal{I}} \gamma_i (\xi_v(x_j) - \xi_v(x_i)) \right] + \delta
\leq \frac{8}{N} \mathbb{E} \left[ \sup_{v \in \mathcal{V}_d} \sum_{i \in \mathcal{I}} \gamma_i \xi_v(x_i) \right] + \delta
\leq \frac{8}{N} \sqrt{d \mu_4} |\mathcal{I}| + \delta.
$$

Equation (14) follows from observing that $(\xi_v(x_j) - \xi_v(x_i)) \overset{\text{d}}{=} \xi_i (\xi_v(x_j) - \xi_v(x_i))$. In (15), $\{\gamma_i\}_{i \in \mathcal{I}}$ are independent Rademacher random variables owing to their construction. Equation (16) can be derived from Theorem 1. Thus, combining (12), (13), and (17), we conclude that, with probability of at least $1 - e^{-2Ld^2}$

$$
\sup_{v \in \mathcal{V}_d} \sum_{i=1}^L \mathbb{I} \{ (P - P_{B_i}) \xi_v > \epsilon \}
\leq L \left( \frac{4}{B} \mu_4 + \frac{|\mathcal{I}|}{L} + \frac{8}{N} \sqrt{d \mu_4} |\mathcal{I}| + \delta \right).
$$

We choose $\delta = (2/4 + \eta) - (|\mathcal{I}|/L)$ and $\epsilon = 2 \max \left\{ \sqrt{\frac{8(4 + \eta) \mu_4}{\eta}} \sqrt{\frac{L}{N}}, \frac{16 \sqrt{d \mu_4} (4 + \eta) \sqrt{|\mathcal{I}|}}{\eta} \right\}$.

The right-hand side of (18) thus becomes strictly lesser than $(L/2)$. Thus, we have shown that

$$
\mathbb{P} \left( \sup_{v \in \mathcal{V}_d} (P \xi_v - \mathbb{E} X' (P \xi_v)) > \epsilon \right) \leq e^{-2Ld^2}.
$$

Similarly, we can show that

$$
\mathbb{P} \left( \sup_{v \in \mathcal{V}_d} (\mathbb{E} X' (P \xi_v) - P \xi_v) > \epsilon \right) \leq e^{-2Ld^2}.
$$

Merging the above two inequalities, we obtain

$$
\mathbb{P} \left( \sup_{v \in \mathcal{V}_d} |\mathrm{MoM}^N_L (\xi_v) - P \xi_v| > \epsilon \right) \leq 2e^{-2Ld^2}.
$$

Alternatively, with at least probability $1 - 2e^{-2Ld^2}$

$$
\sup_{v \in \mathcal{V}_d} |\mathrm{MoM}^N_L (\xi_v) - P \xi_v|
\leq 2 \max \left\{ \frac{16(2 + \eta) \mu_4}{\eta} \sqrt{\frac{L}{N}}, \frac{32 \sqrt{d \mu_4} (2 + \eta) \sqrt{|\mathcal{I}|}}{\eta} \right\}
\leq C \sqrt{\mu_4} \left( \sqrt{\frac{L}{N}} + \sqrt{\frac{d |\mathcal{I}|}{N}} \right).
$$

From Theorem 2, we can say that for any $V \in \mathcal{V}_d$, $\mathrm{MoM}^N_L (\xi_v)$, and $P \xi_v$ are close to each other with a high probability. Thus, one can expect their corresponding minimum values to be close enough with a high probability. Theorem 3 affirms this claim. We show that $\mathcal{R}(\hat{\mathcal{V}}_d) \lesssim \sqrt{\mathbb{E} \| x \|^4 ((|\mathcal{I}|/L)^{1/2} + \frac{d |\mathcal{I}|}{N})}$ with a high probability.

**Theorem 3:** Under A1–A3, and $N > L$, with probability at least, $1 - 2e^{-2L((2/4 + \eta) - (|\mathcal{I}|/L))^2}$

$$
\mathcal{R}(\hat{\mathcal{V}}_d) \leq 2C \sqrt{\mathbb{E} \| x \|^4 \left( \frac{L}{N} + \frac{\sqrt{d |\mathcal{I}|}}{N} \right)},
$$

where $C$ is defined as in Theorem 2.

**Proof:** Using the bound derived in (5), we observe the following:

$$
\mathcal{R}(\hat{\mathcal{V}}_d) \leq 2 \sup_{v \in \mathcal{V}_d} |\mathbb{E} \xi_v (P \xi_v) - P \xi_v |
\leq 2C \sqrt{\mathbb{E} \| x \|^4 \left( \frac{L}{N} + \frac{\sqrt{d |\mathcal{I}|}}{N} \right)}.
$$

From Theorem 2, (19) holds with probability at least $1 - 2e^{-2L((2/4 + \eta) - (|\mathcal{I}|/L))^2}$, proving the desired result. 

**D. Inference for Real Vector Spaces**

In this section, we discuss the main implications of Theorem 3 for real vector spaces. We consider the special case of $\mathcal{H} = \mathbb{R}^p$. We observe that $(2/(4 + \eta)) - (|\mathcal{I}|/L) \geq (2/(4 + \eta)) - (1/(2 + \eta)) = \eta/(2 + \eta)(4/4 + \eta) > 0$. Thus, if $L \rightarrow \infty$, $1 - 2e^{-2L((2/4 + \eta) - (|\mathcal{I}|/L))^2} \rightarrow 1$. We note that

$$
\mathcal{R}(\hat{\mathcal{V}}_d) \lesssim \sqrt{\mathbb{E} \| x \|^4 \left( \frac{L}{N} + \frac{d |\mathcal{I}|}{N} \right)}.
$$

If each component of the inliers were generated independently of the same distribution, it is easy to observe that, $\mathbb{E} \| x \|^4 \approx p^2$. Thus, it is only natural to make the following assumption about $P$.

**A4:** $\mathbb{E}_{x \sim P} \| x \|^4 = \Theta (p^2)$.

Thus, with a high probability

$$
\mathcal{R}(\hat{\mathcal{V}}_d) \lesssim p \left( \sqrt{\frac{L}{N}} + \frac{\sqrt{d |\mathcal{I}|}}{N} \right) \leq p \left( \sqrt{\frac{L}{N}} + \frac{d}{N} \right).
$$

In other words, if $L = o((N/p^2))$ and $p^2 d = o(N)$, then $\mathcal{R}(\hat{\mathcal{V}}_d) \rightarrow 0$, with probability tending to 1.
To formally state this result, we make the following two assumptions.

A5: \( L \to \infty \), as \( N \to \infty \).

A6: \( p^3 \log L, p^2 d = o(N) \).

Such conditions apply naturally: as \( n \) increases, so too must \( L \) to preserve a proportion of the outlier-free partitions. Besides, the increase of \( L \) should be slower than \( n \) so that each partition can be assigned with sufficient data points. Note that since \( d \leq p \), if \( p^3 = o(N) \), then the second part of Assumption A6 is satisfied. In other words, \( N \) has to increase faster than the cube of the feature space dimensions. The condition implied by A6, i.e., \( |O| = o(n) \) is intuitively appealing, and widely accepted [26], [35] as the number of outliers should be small by definition.

Before we state our result in Corollary 1, we recall that \( X_n = O_p(a_n) \) if the sequence of random variables \( \{X_n/a_n\}_{n \in \mathbb{N}} \) is tight [51].

**Corollary 1:** Then under, A1–A6, \( \mathcal{R}(\hat{V}_d) = O_p(p((L/N)^{1/2} + (d/N)^{1/2}) \) and \( \mathcal{R}(\hat{V}_d) \overset{P}{\to} 0 \).

**Proof:** We note that \( \mathcal{R}(\hat{V}_d) \leq p((L/N)^{1/2} + (d/N)^{1/2}) \), with probability at least \( 1 - 2e^{-2L(2^{(d+4)/4})-(d^2(2^{d+1}-d)^2)} \). By A3 and A6, \( 2e^{-2L(2^{(d+4)/4})-(d^2(2^{d+1}-d)^2)} = o(1) \). Thus, \( \mathcal{P}(\mathcal{R}(\hat{V}_d) = O(p((L/N)^{1/2} + (d/N)^{1/2})) > 1 - o(1) \). Hence, \( |P f_{\hat{g}_d} - P f_{\hat{g}}| = O_p(p((L/N)^{1/2} + (d/N)^{1/2}) \).

Under A6, \( p((L/N)^{1/2} + (d/N)^{1/2}) = o(1) \) \( \implies \mathcal{R}(\hat{V}_d) = O_p(1) \).

To conclude that \( \hat{V}_d \overset{P}{\to} V_d^* \), we need to ensure that \( V_d^* \) is identifiable. Toward ensuring that, we make the following identifiability assumption on \( V_d^* \). This type of identifiability conditions is especially popular in clustering literature [52], [53].

A7: For all \( \epsilon > 0 \), there exists \( \delta > 0 \), such that, if \( \|V - V_d^*\| > \epsilon \), \( \mathcal{R}(V) \geq \delta \).

We will say that \( V_b \to V \) if \( \|V_b - V\| \to 0 \). With this notion of convergence in the Frobenious sense, we are now ready to prove that \( \hat{V}_d \) is consistent for \( V_d^* \).

**Corollary 2:** Under A1–A7, \( \hat{V}_d \overset{P}{\to} V_d^* \).

**Proof:** We fix \( \epsilon > 0 \). By A7, there exists \( \delta > 0 \), such that, if \( \|\hat{V}_d - V_d^*\| > \epsilon \), \( \mathcal{R}(\hat{V}_d) \geq \delta \). Thus, \( \mathcal{P}(\|\hat{V}_d - V_d^*\| > \epsilon) \leq \mathcal{P}(\mathcal{R}(\hat{V}_d) \geq \delta) \to 0 \) as \( N \to \infty \), by appealing to Corollary 1. Thus, \( \|\hat{V}_d - V_d^*\| \overset{P}{\to} 0 \implies \hat{V}_d \overset{P}{\to} V_d^* \).

**Remark (Cost of Robustness):** The above results under our paradigm, the MoMPCA estimates admit an excess risk of \( O(p((L/N)^{1/2} + (d/N)^{1/2})) \). We observe that since \( L \geq 1 \), \( p((L/N)^{1/2} + (d/N)^{1/2}) = \Omega(n^{-1/2}) \). Thus, our framework’s convergence rate for MoMPCA is generally slower than its ERM counterpart, which has a rate of \( O(n^{-1/2}) \). This reiterates that there is “no free lunch” when compromising robustness for the convergence rate, which is not unusual given that MoM operates on data contaminated with outliers. However, if the number of partitions \( L \) increases slowly compared to \( n \) [say, \( L = O(\log n) \) and \( |O| = O(\log n) \)], the excess risk for MoMPCA estimates draws closer to its ERM counterpart at a rate of \( \tilde{O}(n^{-1/2}) \).

**Remark (Choice of \( L \)):** If the number of partitions can mitigate the effect of the outliers, under the proposed framework, the excess risk of the robust MoM estimates decreases with the block size at the rate of \( b \) as \( 1/\sqrt{b} = \sqrt{L/n} \). Since \( L \) can identifiability approximately as \( 2|O| \), the corresponding excess risk has the rate of \( O(|O|/n^{1/2}) \). However, the error bound \( O(|O|/n^{1/2}) \) becomes vacuous if \( |O| \propto n \). Thus, it is essential that \( |O| = o(n) \) for our consistency guarantees to hold, as this enables us to select \( L \) satisfying A5 and A6. One should note that one can achieve an error rate of \( O(n^{(b^2-1)/2}) \) if \( |O| = O(n^\beta) \), for some \( 0 < \beta < 1 \).

**Remark (Comparison With MCM-PCA):** The recent array of works on geometric median-based approaches provide an attractive alternative to our proposal [18], [19], [20], [21]. However, the two approaches are significantly different. The geometric median-based approaches focus on finding a robust estimate of the dispersion matrix using a geometric median-based loss function. These approaches then go on to find a suitable subspace for low-dimensional representation of the data by an eigen decomposition of this robust estimate of the covariance matrix. On the other hand, MoMPCA approaches changing the ERM problem and introducing a robust estimate of the projection error. Theoretically, we derive finite-sample error bounds in Hilbert space without imposing any assumption on the outliers, thus relaxing the i.i.d. assumption of the data distribution imposed in [19].

**Remark:** The MoMPCA method proposed in this article can provide a more accurate analysis of data with outliers compared to existing methods. This is because the method is specifically designed to handle outlier data, and does not rely on unrealistic assumptions about the data distribution. The MoMPCA method does not make assumptions about the distribution of the outlier data. This allows for greater flexibility in dealing with different types of outlier behavior, making the method more useful in real-world applications where the nature of outliers may not be known in advance. Furthermore, the MoMPCA method provides a more interpretable understanding of PCA in the presence of outliers. This can be especially useful in applications where the insights provided by the analysis are important for decision-making. We note that MoMPCA does not require the data to have nice tail-conditions such as sub-Gaussian or subexponential behavior as required by many relevant works in this direction [24], [34], which are often unrealistic assumptions for real-world data. This means that the method can be applied to a wider range of datasets, making it more useful in practice.

**IV. EXPERIMENTAL RESULTS**

In this section, we demonstrate the efficacy of MoMPCA on synthetic and natural datasets. We apply MoMPCA to perform different tasks, including low-rank matrix reconstruction in the presence of outliers, background modeling for video data, and anomaly detection for real data benchmarks. The codes and machine specifications are given at https://github.com/SaptarshiC98/MOMPCA.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Parameter Selection: The gradient descent parameter $\eta$ used for the optimization here is typically chosen based on standard methods [55]. In our experiments, we fix $\eta$ at 0.01. Thus, the only parameter to be selected is the number of partitions $L$. By the relationship $LB = n$, the block size $B$ is chosen automatically once $L$ is fixed. For the PCA to be effective, the block size must be greater than $p$, which imposes an upper bound on $L$ to be $(n/p)$. On the other hand, to maintain the parity with the theoretical results, we notice in Assumption A3 that $L$ must be greater than $2|O|$, where $|O|$ is the number of outliers present in the datapoint. This imposes a lower bound on $L$, and hence in ideal cases, $L$ should be between $2|O|$ and $(n/p)$. However, the number of outliers present in a dataset is typically unknown, and $L$ must be greater than $2|O|$ signifies a particular case where every partition contains exactly one outlier. In practice, since the partitions are done randomly, the probability of that happening is very low. Hence, we may select $L$ to be much less than $2|O|$. This relaxation in the choice of $L$ leads to the possibility of using the MoMPCA algorithm for datasets where the number of outliers is quite large and $2|O|$ surpasses the quantity $(n/p)$. Hence, in practice, we have chosen $L$ to be less than $(n/p)$, with regard to being sufficiently large so that the median class can be free of outliers.

A. Simulation Study for Excess Risk Comparison

To empirically verify the efficacy of MoMPCA, we perform a simulation study to compare against vanilla PCA and the recently proposed MCM-PCA [19]. We generate the inlying observations of the dataset independently of $p = 10$ dimensional multivariate Gaussian distribution with mean 0 and covariance matrix $\Sigma = \text{diag}(10, 9, \ldots, 2, 1)$. Thus, the first two dimensions are most important when projecting the entire data in a 2-D subspace. The outliers, which consist of 10% of datasets, are generated independently of multivariate Gaussian as well, with mean vector $(-20, 20, 0, 0, 0, 0, 0, 0, 0, 0)$ and $I_{10}$ as the covariance matrix.

We then run all three algorithms on the datasets, starting with a total number of datapoints $n = 100$ to $n = 2000$, on a difference of 100. The experiments for each value of $n$ are repeated 500 times. We then report the average value of the excess risk for the inliers, which is defined in (4) takes the form as follows:

$$R(\hat{P}) = \mathbb{E}\left(\|P_0x\|^2 - \|\hat{P}x\|^2\right)$$
$$= \mathbb{E}\left(x^\top(P_0 - \hat{P})x\right)$$
$$= \text{Trace}(x^\top(P_0 - \hat{P})x)$$
$$= \text{Trace}((P_0 - \hat{P})xx^\top)$$
$$= \text{Trace}((P_0 - \hat{P})\Sigma).$$

We calculate the excess risk for each algorithm, where $P_0$ is the actual projection matrix for projecting onto the first two-dimension and $\hat{P}$ is the projection matrix estimated by the respective algorithms. Then, we plot the average value of test error for each of the three algorithms and compare them in Fig. 3. We can clearly see that the proposed MoMPCA algorithm performs the best among its peers in terms of test error. In Fig. 3, we show the excess risk of MoMPCA and MCM-PCA in the same subfigure since both have a similar decay rate, with MoMPCA consistently better. On the other hand, the vanilla PCA falls apart and has an excess risk much more significant than its peers. Additionally, one should note that the rate of decrease in the excess risk resembles the risks derived in Corollary 1, validating the theoretical analysis in the process.

B. Simulation Study on Recovering Low-Rank Matrices

To empirically validate the efficacy of MoMPCA, we performed this dimensionality reduction procedure on various simulated datasets. We have generated the matrix $x_0$ as a product of two matrices $x_{1,0}$ and $x_{2,0}$ of orders $n \times r$ and $r \times p$, respectively, where each entry is simulated from the Gaussian distribution, where $r$ is the rank of $x_0$ and $r < \min\{n, p\}$. To add outliers, we add random noise generated from Unif($-500, 500$) to randomly selected $\sqrt{n}$ many rows of $x_0$.

We then run the MoMPCA algorithm along with the peer algorithms on $x_0$. We calculate the efficacy of each of the relative reconstruction errors as $((\|x - x_0\|)/\|x_0\|)$, where the norm is taken over those rows that do not contain an outlier.
Here \( X \) is the projected matrix in the lower \( d \)-dimensional affine space. Clearly, a lower score represents a better representation in lower-dimensional space. Since the true structure of the matrix lies in a lower-dimensional space, the MoMPCA method successfully plots the dataset in a lower-dimensional space despite having \( \sqrt{n} \) many outliers due to its robustness property.

We compare our method with the baselines as well as state-of-the-art methods such as principal component pursuit (PCP) [7], principal component pursuit with features (PCPF) [12], matrix-factorization-based robust PCA (MFRPCA) [54] and robust weight learning with adaptive neighbors (RWL-AN) [8]. The standard protocols used by each of the original papers of the competing algorithms have been implemented in our experiments. The algorithms providing nondeterministic output have been run 20 times, and the average values obtained have been reported. The results demonstrated in Table I clearly indicate that out of all the methods, the proposed MoMPCA algorithm works best in terms of the reconstruction error.

**Runtime Analysis:** We run each of the algorithms as demonstrated in Section IV-B and compare their total runtime. In Table II, the total time in seconds needed to run the algorithms on an M1 Macbook Pro with 8-GB RAM and 256-GB storage is provided. As evident, the first three algorithms, namely PCP [7], PCPF, and MFRPCA are all matrix-factorization-based algorithms and hence take a huge amount of time to run as compared to the rest. The proposed algorithm MoMPCA is run for 50 iterations since the algorithm mostly converges much before. We can clearly see from the runtime analysis that the time required to run MoMPCA is almost equivalent to that of MCM-PCA while providing a better performance against the outliers.

**C. Background Modeling in Video**

Suppose we are given a video with \( f \) many frames. The objective here is to isolate the moving objects in the video, such as moving cars and pedestrians. For simplicity, we only consider black and white videos (and thus, there is only a single channel) and let each frame be of size \( m \times n \). We construct \( \mathcal{X} = \{x_1, \ldots, x_m\} \subset \mathbb{R}^f \) by computing a \( f \times 1 \) vector whose features represent the value of a particular pixel for each frame. Since consecutive frames are highly correlated, one can expect that the data cloud \( \mathcal{X} \) lies near a low-dimensional affine space in \( \mathbb{R}^f \). Thus, the background of the video can be modeled as the projection of \( \mathcal{X} \) on this low-dimensional affine space. To demonstrate the efficacy of modeling the background through MoMPCA, we take the standard video, called “highway” (available in the github repository), which captures moving cars on a busy highway. The video contains 1700 frames, and each frame is of size \( 240 \times 320 \). We construct a 76800 \( \times \) 1700 data matrix whose rows denote the data points. We run the MoMPCA with \( L = 40 \) and \( d = 5 \). The background is thus given by \( \{Qx_1, \ldots, Qx_{76800}\} \), where \( Q \) is the projection matrix, given by Algorithm 1. The output background for frame no. 863 is shown in Fig. 4 along with the original frame. The object is constructed by taking \( \{|\|x_i - Qx_i\||_2\}_{i=1}^{76800} \). The object conforms to the moving cars in the original frame, and the background only consists of the highway and motionless trees.

**D. Application in Anomaly Detection**

Anomaly or outlier detection is a key problem in machine learning and computer vision. In this section, we will focus on how to detect outliers through MoMPCA. Suppose \( \mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{R}^p \) be \( n \) data points in the \( p \)-dimensional real vector space. To detect outliers within the data, we first project \( \mathcal{X} \) to a \( d \)-dimensional affine space via MoMPCA. We then compute the squared Euclidean distance between the original and projected data points \( \tilde{v}_i = |\|x_i - Qx_i\||_2^2 \) and sort \( \tilde{v}_i \)’s in ascending order. We call the \( i \)th point an outlier if \( v_i \) belongs to the largest \( o \% \) of the \( \{|\|v_i\||_2\|\}_{i=1}^{n} \), where \( o \) is known beforehand.

For our comparative analysis, we take the KDDCUP, Thyroid, and Arrhythmia datasets, available from the UCI machine learning repository [62] and ODDS library [63]. The details of these datasets are reported in Table III. We compared our method with different state-of-the-art techniques for outlier detection including one class support vector machines (OC-SVM) [56]; deep structural energy-based

---

**Table I**

| \( n \) | \( p \) | \( \text{rank}(x_0) \) | PCP [7] | PCPF [12] | MFRPCA [54] | RWL-AN [8] | MCM-PCA [19] | MoMPCA (Proposed) |
|---|---|---|---|---|---|---|---|---|
| 500 | 500 | 10 | \( 1.6 \times 10^{-3} \) | \( 1.5 \times 10^{-3} \) | \( 2.3 \times 10^{-4} \) | \( 8.7 \times 10^{-2} \) | \( 1.8 \times 10^{-3} \) | \( 1.5 \times 10^{-3} \) |
| 1000 | 500 | 10 | \( 7.3 \times 10^{-4} \) | \( 2.8 \times 10^{-4} \) | \( 1.8 \times 10^{-4} \) | \( 2.9 \times 10^{-4} \) | \( 2.7 \times 10^{-4} \) | \( 2.1 \times 10^{-4} \) |
| 2000 | 500 | 10 | \( 8.8 \times 10^{-4} \) | \( 1.1 \times 10^{-4} \) | \( 8.3 \times 10^{-4} \) | \( 3.1 \times 10^{-4} \) | \( 6.4 \times 10^{-5} \) | \( 5.6 \times 10^{-5} \) |
| 5000 | 500 | 10 | \( 2.1 \times 10^{-5} \) | \( 9.6 \times 10^{-5} \) | \( 5.7 \times 10^{-4} \) | \( 3.3 \times 10^{-4} \) | \( 8.1 \times 10^{-6} \) | \( 7.2 \times 10^{-6} \) |
| 10000 | 500 | 10 | \( 7.8 \times 10^{-5} \) | \( 8.2 \times 10^{-5} \) | \( 2.9 \times 10^{-4} \) | \( 9.5 \times 10^{-5} \) | \( 4.3 \times 10^{-7} \) | \( 3.8 \times 10^{-7} \) |

**Table II**

| \( n \) | \( p \) | \( \text{rank}(x_0) \) | PCP [7] | PCPF [12] | MFRPCA [54] | RWL-AN [8] | MCM-PCA [19] | MoMPCA (Proposed) |
|---|---|---|---|---|---|---|---|---|
| 500 | 500 | 10 | 106.05 | 104.13 | 106.81 | 8.14 | 2.34 | 5.18 |
| 1000 | 500 | 10 | 320.43 | 311.18 | 321.32 | 11.40 | 4.94 | 10.31 |
| 2000 | 500 | 10 | 952.87 | 932.65 | 956.05 | 26.76 | 10.35 | 21.27 |
| 5000 | 500 | 10 | 3203.54 | 3197.43 | 3215.29 | 49.03 | 21.02 | 53.28 |
| 10000 | 500 | 10 | 9608.68 | 9547.83 | 9677.04 | 103.42 | 45.96 | 98.04 |

**Table III**

| Data | # Dimensions | # Instances | Anomaly ratio |
|---|---|---|---|
| KDDCUP | 120 | 494,021 | 0.2 |
| Thyroid | 6 | 3,772 | 0.025 |
| Arrhythmia | 274 | 452 | 0.15 |
models-energy (DSEBM-e) & deep structural energy-based models-reconstruction (DSEBM-r) [57]; deep clustering network (DCN) [58]; GMM-EN, PAE, E2E-AE, PAE-GMM-EM, PAE-GMM, DAGMM-p, DAGMM-NVI, DAGMM [59], robust and sparse PCA (RSPCA) [60], and Stochastic PCA with Barron loss [61]. Many competing algorithms employ a deep neural network to detect anomalies within the data. To measure the performance of the peer algorithms, we take the average precision, recall, and $F_1$ score between the ground truth and the obtained labeling (inlier/outlier) of the data points. These measures of accuracy for all three datasets are reported in Table IV. The performance indicator values for the peer algorithms are quoted from [59]. It is observed from Table IV that the MoMPCA is quite competitive against state-of-the-art anomaly detection methods, including even the ones based on deep neural networks.

V. CONCLUSION

Despite the efficacy, computational simplicity, and ease of visualization of the classical PCA, it often fails to successfully represent the true low-dimensional structure of a dataset in the presence of even a small number of outliers. The traditional approach to formulating robust PCA assumes that the data matrix can be decomposed into a low-rank signal component and a noise component. However, this framework is not suitable when dealing with outliers that follow arbitrary distributions and/or are correlated with each other.

This article proposes an alternative PCA method based on the MoM estimator to circumvent this problem. The eigen-decomposition trick for PCA cannot be applied in this context, so we use an alternative approach involving projected Adagrad. The proposed MoMPCA, equipped with a computationally simple gradient descent-based optimization procedure, exhibits significant robustness to the presence of outliers. Under minimal and interpretable assumptions, we establish the consistency of MoMPCA in a general separable Hilbert space and find the convergence rate using uniform concentration bounds. The paper’s theoretical analysis is carried out with the aid of symmetrization arguments and Rademacher complexities, which, although extensively used in a supervised learning setting, seldom find application in an unsupervised learning scenario. The parametric rates for real vector spaces can easily be recovered by taking the Hilbert space to be $\mathbb{R}^p$. The applicability of our theoretical results to an infinite-dimensional Hilbert space to derive dimension-free bounds with minimal assumptions is not only novel but also opens up exciting avenues for future research in robust kernel-based methods.

Through practical applications in computer vision, MoMPCA is shown to be compelling relative to even some of the recent deep learning models. The robustness of MoM...
estimators comes at the expense of slower convergence rates than their ERM counterparts, as demonstrated in this article. We stress the fact that there is no "wizardry in the MoM estimator" and that the interplay between the partitions and the outliers determines how effective MoM is. A potential future extension of our work could be to make the optimization faster by using adaptive gradient-based optimizers and proving their convergence properties. Future research in this direction might render fruitful avenues in improving the "slow" ERM rates by establishing fast rates under more restrictive assumptions [64], [65] or finding lower bounds on the approximation error.

REFERENCES

[1] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," London, Edinburgh, Dublin Phil. Mag. J. Sci., vol. 2, no. 11, pp. 559–572, Nov. 1901.

[2] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," Chemometrics Intell. Lab. Syst., vol. 2, nos. 1–3, pp. 37–52, Aug. 1987.

[3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Cham, Switzerland: Springer, 2009.

[4] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," J. Roy. Statist. Soc. B, vol. 61, pp. 611–622, Jan. 1999.

[5] B. Schölkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in Proc. Int. Conf. Artif. Neural Netw., Cham, Switzerland: Springer, 1997.

[6] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," J. Comput. Graph. Statist., vol. 15, no. 2, pp. 265–286, 2006.

[7] J. Wright, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" J. ACM, vol. 58, no. 1, pp. 1–37, 2009.

[8] R. Zhang and H. Tong, "Robust principal component analysis with adaptive neighbors," in Proc. Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 6961–6970.

[9] S. Minsker, "Uniform bounds for robust mean estimators," 2018, arXiv:1812.03523.

[10] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi, "Bandits with heavy tail," IEEE Trans. Inf. Theory, vol. 59, no. 11, pp. 7711–7717, Nov. 2013.

[11] C. Brunet-Saumard, E. Genetay, and A. Saumard, "K-bMOM: A robust Lloyd-type clustering algorithm based on bootstrap median-of-means," 2020, arXiv:2002.03899.

[12] D. Paul, S. Chakraborty, S. Das, and J. Xu, "Uniform concentration bounds toward a unified framework for robust clustering," in Proc. Adv. Neural Inf. Process. Syst., vol. 34, 2021, pp. 8307–8319.

[13] G. Stuerma, P. Laforgue, P. Mozharovskyi, and F. Buc, "When OT meets mom: Robust estimation of Wasserstein distance," in Proc. 24th Int. Conf. Artif. Intell., (Cambridge, MA, USA: Morgan & Claypool), 2021, pp. 1–12.

[14] S. Liu, L. Feng, and H. Qiao, "Scatter balance: An angle-based supervised dimensionality reduction," IEEE Trans. Neural Netw. Learn. Syst., vol. 26, no. 2, pp. 277–289, Feb. 2015.

[15] S. Liu and Y. Yu, "Angular embedding: A new angular robust principal component analysis," 2020, arXiv:2011.11013.

[16] S. Haeber, A. Feragen, and M. J. Black, "Grassmann averages for scalable robust PCA," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 3810–3817.

[17] S. Haeber, A. Feragen, R. Enficiaud, and M. J. Black, "Scalable robust principal component analysis using Grassmann averages," IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 11, pp. 2298–2311, Nov. 2016.

[18] T. Bouwhmans, S. Javed, H. Zhang, Z. Lin, and R. Oza, "On the applications of robust PCA in image and video processing," Proc. IEEE, vol. 106, no. 8, pp. 1427–1457, Aug. 2018.

[19] Q. Gao, L. Ma, Y. Liu, X. Gao, and F. Nie, "Angle 2DPCA: A new formulation for 2DPCA," IEEE Trans. Cybern., vol. 48, no. 5, pp. 1672–1678, May 2018.

[20] Q. Gao, P. Zhang, W. Xia, D. Xie, X. Gao, and D. Tao, "Enhanced tensor RPCA and its application," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 6, pp. 2133–2140, Jun. 2021.

[21] S. Liao, J. Li, Y. Liu, Q. Gao, and X. Gao, "Robust formulation for PCA: Avoiding mean calculation with $L_2$-norm maximization," in Proc. AAAI Conf. Artif. Intell., Apr. 2018, vol. 32, no. 1, pp. 1–13.
