Intrinsic Grassmann Averages for Online Linear and Robust Subspace Learning*

Rudrasis Chakraborty[1], Søren Hauberg[2] and Baba C. Vemuri[1]
[1]Department of CISE, University of Florida, FL 32611, USA
[2]Technical University of Denmark, Richard Petersens Plads, Denmark
[1]{rudrasischa, baba.vemuri}@gmail.com [2]sohau@dtu.dk

Abstract

Principal Component Analysis (PCA) is a fundamental method for estimating a linear subspace approximation to high-dimensional data. Many algorithms exist in literature to achieve a statistically robust version of PCA called RPCA. In this paper, we present a geometric framework for computing the principal linear subspaces in both situations that amounts to computing the intrinsic average on the space of all subspaces (the Grassmann manifold). Points on this manifold are defined as the subspaces spanned by K-tuples of observations. We show that the intrinsic Grassmann average of these subspaces coincide with the principal components of the observations when they are drawn from a Gaussian distribution. Similar results are also shown to hold for the RPCA. Further, we propose an efficient online algorithm to do subspace averaging which is of linear complexity in terms of number of samples and has a linear convergence rate. When the data has outliers, our proposed online robust subspace averaging algorithm shows significant performance (accuracy and computation time) gain over a recently published RPCA methods with publicly accessible code. We have demonstrated competitive performance of our proposed online subspace algorithm method on one synthetic and two real data sets. Experimental results depicting stability of our proposed method are also presented. Furthermore, on two real outlier corrupted datasets, we present comparison experiments showing lower reconstruction error using our online RPCA algorithm. In terms of reconstruction error and time required, both our algorithms outperform the competition.

1. Introduction

Principal component analysis (PCA), a key work-horse of machine learning, can be derived in many ways: Pearson [29] proposed to find the subspace that minimizes the projection error of the observed data; Hotelling [20] instead sought the subspace in which the projected data has maximal variance; and Tipping & Bishop [34] consider a probabilistic formulation where the covariance of normally distributed data is predominantly given by a low-rank matrix. All these derivations lead to the same algorithm. Recently, Hauberg et al. [18] noted that the average of all one-dimensional subspaces spanned by normally distributed data coincides with the leading principal component. Here the average is computed over the Grassmann manifold of one-dimensional subspaces (cf. Sec. 2). This average can be computed very efficiently, but unfortunately their formulation does not generalize to higher-dimensional subspaces.

In this paper, we provide a formulation for estimating the average K-dimensional subspace spanned by the observed data, and present a very simple, parameter-free online algorithm for computing this average. When the data is normally distributed, we show that this average subspace coincides with that spanned by the leading K principal components. We further show that our online algorithm has a linear convergence rate. Moreover, since our algorithm is online, it has a linear complexity in terms of the number of samples. Furthermore, we propose an online robust subspace averaging algorithm which can be used to get the leading K robust principal components. Analogous to its non-robust counterpart, it has a linear time complexity in terms of the number of samples.

1.1. Related Work

In this paper we consider a simple linear dimensionality reduction algorithm that works in an online setting, i.e. only uses each data point once. There are several existing approaches in literature that tackle the online PCA and the online Robust PCA problems and we discuss some of these approaches here:

Oja's rule [28] is a classic online estimator for the leading principal components of a dataset. Given a basis $V_{t-1} \in \mathbf{R}^{D\times K}$ this is updated recursively via $V_t = V_{t-1} + \gamma_t X_t (X_t^T V_{t-1})$ upon receiving the observation X_t. Here γ_t is the step-size (learning rate) parameter that must be

*This research was funded in part by the NSF grant IIS-1525431 to BCV. SH was supported by a research grant (15334) from VILLUM FONDEN.

set manually; small values gives slow-but-sure convergence, while larger values may give fast-but-unstable convergence.

EM-PCA [33] is usually derived for probabilistic PCA, but is easily be adapted to the online setting [9]. Here, the E- and M-steps are given by:

$$\textbf{(E-step)}\ Y_t = (V_{t-1}^T V_{t-1})^{-1}(V_{t-1}^T X_t) \quad (1)$$

$$\textbf{(M-step)}\ \tilde{V}_t = (X_t Y_t^T)(Y_t Y_t^T)^{-1}. \quad (2)$$

The basis is updated recursively via the recursion, $V_t = (1-\gamma_t)V_{t-1} + \gamma_t \tilde{V}_t$, where γ_t is a step-size.

GROUSE and GRASTA [4, 19] are online PCA and matrix completion algorithms. GRASTA can be applied to estimate principal subspaces incrementally on subsampled data. Both of these methods are online and use rank-one updation of the principal subspace at each iteration. We have compared our online subspace estimation algorithm with GROUSE. GRASTA is an online robust subspace tracking algorithm and can be applied on subsampled data and specifically matrix completion problems. The authors proposed an ℓ_1-norm based fidelity term that measures the error between the subspace estimate and the outlier corrupted observations. The robustness of GRASTA is attributed to this ℓ_1-norm based cost. Their formulation of the subspace estimation involves the minimization of a non-convex function in an augmented Lagrangian framework. This optimization is carried out in an alternating fashion using the well known ADMM [7] for estimating a set of parameters involving the weights, the sparse outlier vector and the dual vector in the augmented Lagrangian framework. For fixed estimated values of these parameters, they employ an incremental gradient descent to solve for the low dimensional subspace. Note that the solution obtained is not the optimum of the combined non-convex function of GRASTA. In the experimental results section, we will present comparisons between GRASTA and our recursive robust PCA algorithm.

Recursive covariance estimation [6] is straight-forward, and the principal components can be extracted via standard eigen-decompositions. Boutsidis et al. [6] consider efficient variants of this idea, and provide elegant performance bounds. The approach does not however scale to high-dimensional data as the covariance cannot practically be stored in memory for situations involving very large data sets as those considered in our work.

In [8], Candes et al. formulated Robust PCA (RPCA) as separating a matrix into a low rank (L) and a sparse matrix (S), i.e., data matrix $X \approx L + S$. They proposed Principal Component Pursuit (PCP) method to robustly find the principal subspace by decomposing into L and S. They showed that both L and S can be computed by optimizing an objective function which is a linear combination of nuclear norm on L and ℓ_1 norm on S. Recently, Lois and Vaswani [25] proposed an online RPCA problem to solve two interrelated problems, matrix completion and online robust subspace estimation. The authors have some assumptions including a good estimate of the initial subspace and that the basis of the subspace is dense. Though the authors have shown correctness of their algorithm under these assumptions, these assumptions are often not practical. In another recent work, Ha and Barber [17] proposed an online RPCA algorithm when $X = (L+S)C$ where C is a data compression matrix. They proposed an algorithm to extract L and S when the data X are compress sensed. This problem is quite interesting in its own right but not something pursued in our work presented here. Feng et al. [13] solved RPCA using a stochastic optimization approach. The authors have shown that if each observation is bounded, then their solution converges to the batch mode RPCA solution, i.e., their sequence of robust subspaces converges to the "true" subspace. Hence, they claimed that as the "true subspace" (subspace recovered by RPCA) is robust, so is their online estimate. Though their algorithm is online, the optimization steps (Algorithm 1 in [13]) are expensive for high-dimensional data. In an earlier paper, Feng et al. [12] proposed a deterministic approach to solve RPCA (dubbed DHR-PCA) for high-dimensional data. They also showed that they can achieve maximal robustness, i.e., a breakdown point of 50%. They proposed a robust computation of the variance matrix and then performed PCA on this matrix to get robust PCs. This algorithm is suitable for very high dimensional data. As most of our real applications in this paper are in very high dimensions, we find DHR-PCA to be well suited to carry out comparisons with. Finally, we would like to refer the readers to an excellent source of references on RPCA in a recent MS thesis [36].

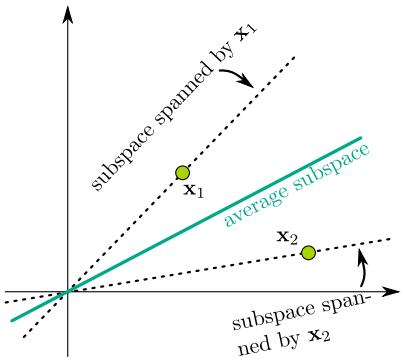

Figure 1. The average of two subspaces.

Motivation for our work: Our work is motivated by the work presented by Hauberg et al. [18], who recently showed that for a data set drawn from a zero-mean multivariate Gaussian distribution, the average subspace spanned by the data coincides with the leading principal component. This idea is sketched in Fig. 1. Given, $\{\mathbf{x}_i\}_{i=1}^N \subset \mathbf{R}^D$, the 1-dimensional subspace spanned by each $\mathbf{x}_i$ is a point on the Grassmann manifold (Sec. 2). Hauberg et al. then compute the average of these subspaces on the Grassmannian using an "extrinsic" metric, i.e. the Euclidean, distance. Besides the theoretical insight, this formulation gave rise to highly

efficient algorithms. Unfortunately, the extrinsic approach is limited to one-dimensional subspaces, and Hauberg et al. resort to deflation methods to estimate higher dimensional subspaces. We overcome this limitation by using an intrinsic metric, extend the theoretical analysis of Hauberg et al., and provide an efficient online algorithm for subspace estimation. We further propose an online robust subspace averaging algorithm akin to online RPCA and proved that in the limit our proposed method returns the first K robust principal components. Moreover, we provide a proof of statistical robustness of our recursive PC estimator.

2. An Online Linear Subspace Learning Algorithm

In this section, we propose an efficient online linear subspace learning algorithm for finding the principal components of a data set. We first briefly discuss the geometry of the Riemannian manifold of K-dimensional linear subspaces in $\mathbf{R}^D$. Then, we will present an online algorithm to get the first K principal components of the D-dimensional data vectors.

2.1. The Geometry of Subspaces

The Grassmann manifold (or the Grassmannian) is defined as the set of all K-dimensional linear subspaces in $\mathbf{R}^D$ and is denoted $Gr(K, D)$, where $D \geq K$. A special case of the Grassmannian is when $K = 1$, i.e., the space of one-dimensional subspaces of $\mathbf{R}^D$, which is known as the *real projective space* (denoted by $\mathbf{R}P^D$). A point $\mathcal{X} \in Gr(K, D)$ can be specified by a basis, X, i.e., a set of K linearly independent vectors in $\mathbf{R}^D$ (the columns of X) that spans $\mathcal{X}$. We say $\mathcal{X} = Col(X)$ if X is a basis of $\mathcal{X}$, where $Col(.)$ is the column span operator. We have included a brief note on the geometry of the Grassmannian in the supplementary material. As the Grassmannian is geodesically complete, one can extend the geodesics on the Grassmannian indefinitely [2, 11]. Given $\mathcal{X}, \mathcal{Y} \in Gr(K, D)$, with their respective orthonormal basis X and Y, the unique geodesic $\Gamma_{\mathcal{X}}^{\mathcal{Y}} : [0, 1] \to Gr(K, D)$ between $\mathcal{X}$ and $\mathcal{Y}$ is given by:

$$\Gamma_{\mathcal{X}}^{\mathcal{Y}}(t) = \text{span}\big(X\hat{V}\cos(\Theta t) + \hat{U}\sin(\Theta t)\big) \tag{3}$$

with $\Gamma_{\mathcal{X}}^{\mathcal{Y}}(0) = \mathcal{X}$ and $\Gamma_{\mathcal{X}}^{\mathcal{Y}}(1) = \mathcal{Y}$, where, $\hat{U}\hat{\Sigma}\hat{V}^T = (I - XX^T)Y(X^TY)^{-1}$ is the "thin" Singular value decomposition (SVD), and $\Theta = \arctan\hat{\Sigma}$. The length of the geodesic constitutes the geodesic distance on $Gr(K, D)$, $d : Gr(K, D) \times Gr(K, D) \to \mathbf{R}^+ \cup \{0\}$ which is as follows: Given $\mathcal{X}, \mathcal{Y}$ with respective orthonormal bases X and Y,

$$d^2(\mathcal{X}, \mathcal{Y}) \triangleq \sqrt{\sum_{i=1}^{K} (\arccos(\sigma_i))^2}, \tag{4}$$

where $\bar{U}\Sigma\bar{V}^T = X^TY$ be the SVD of X^TY, and, $[\sigma_1, \ldots, \sigma_K] = \text{diag}(\Sigma)$. Here $\arccos(\sigma_i)$ is known as the i^{th} *principal angle* between subspace $\mathcal{X}$ and $\mathcal{Y}$.

2.2. The Intrinsic Grassmann Average (IGA)

We now consider *intrinsic averages*[1] (IGA) on the Grassmannian. For the existence and uniqueness of IGA, we need to define an open ball on the Grassmannian. Using the geodesic distance (4) we define an open ball of radius r centered at $\mathcal{X} \in Gr(K, D)$ as $\mathcal{B}(\mathcal{X}, r) = \{\mathcal{Y} \in Gr(K, D) | d(\mathcal{X}, \mathcal{Y}) < r\}$. Let κ be the maximum of the sectional curvature in the ball. Then, we call this ball "regular" [24] if $2r\sqrt{\kappa} < \pi$. Using the results in [35], we know that, for $\mathbf{R}P^D$ with $D \geq 2$, $\kappa = 1$, while for general $Gr(K, D)$ with $\min(K, D) \geq 2$, $0 \leq \kappa \leq 2$. So, on $Gr(K, D)$ the radius of a "regular geodesic ball" is $< \pi/2\sqrt{2}$, for $\min(K, D) \geq 2$ and on $\mathbf{R}P^D$, $D \geq 2$, the radius is $< \pi/2$.

Let $\mathcal{X}_1, \ldots, \mathcal{X}_N$ be independent samples on $Gr(K, D)$ drawn from a distribution $P(\mathcal{X})$, then we can define an *intrinsic average* $\mathcal{M}^*$ as:

$$\mathcal{M}^* = \underset{\mathcal{M} \in Gr(K,D)}{\text{argmin}} \sum_{i=1}^{N} d^2\Big(\mathcal{M}, \mathcal{X}_i\Big) \tag{5}$$

On $Gr(K, D)$, IGA exists and is unique if the support of $P(\mathcal{X})$ is within a "regular geodesic ball" of radius $< \pi/2\sqrt{2}$ [3]. Note that for $\mathbf{R}P^D$, we can choose this bound to be $\pi/2$. In the rest of the paper, we have assumed that data points on $Gr(K, D)$ are within a "regular geodesic ball" of radius $< \pi/2\sqrt{2}$ unless otherwise specified. With this assumption, the IGA is unique. *Note that this assumption is needed for proving the theorem presented below.*

The IGA may be computed using a Riemannian steepest descent, but this is computationally expensive and requires selecting a suitable step-size [30]. Recently Chakraborty et al. [10] proposed a simple and efficient inductive (intrinsic) mean estimator:

$$\mathcal{M}_1 = \mathcal{X}_1\,, \quad (\forall k \geq 1)\ \left(\mathcal{M}_{k+1} = \Gamma_{\mathcal{M}_k}^{\mathcal{X}_{k+1}}\left(\frac{1}{k+1}\right)\right) \tag{6}$$

This approach only needs a single pass over the data set to estimate the IGA. Consequently, Eq. 6 has linear complexity in the number of observations. Furthermore, it is a truly online algorithm as each iteration only needs one new observation.

Equation 6 merely performs repeated geodesic interpolation, which is analogous to standard recursive estimators of Euclidean averages: Consider observations $\mathbf{x}_k \in \mathbf{R}^D, k = 1, \ldots, N$. Then the Euclidean average can be computed recursively by moving an appropriate distance away from the

[1] These are also known as Fréchet means [23, 15].

k^{th} estimator $\mathbf{m}_k$ towards $\mathbf{x}_{k+1}$ on the straight line joining $\mathbf{x}_{k+1}$ and $\mathbf{m}_k$. The inductive algorithm (6) for computing the IGA works in the same way and is entirely based on traversing geodesics in $Gr(K, D)$ and without requiring any optimization.

Theorem 1. *(**Weak Consistency** [10]) Let $\mathcal{X}_1, \ldots, \mathcal{X}_N$ be samples on $Gr(K, D)$ drawn from a distribution $P(\mathcal{X})$. Then $\mathcal{M}_N$ (6) converges to the IGA of $\{\mathcal{X}_i\}_{i=1}^N$ in probability as $N \to \infty$.*

Theorem 2. *(**Convergence rate**) Let $\mathcal{X}_1, \ldots, \mathcal{X}_N$ be samples on $Gr(K, D)$ drawn from a distribution $P(\mathcal{X})$. Then Eq. 6 has a linear convergence rate.*

Proof. See the supplementary material. ■

2.3. Principal Components as Grassmann Averages

Following Hauberg et al. [18] we pose the linear dimensionality reduction as an averaging problem on the Grassmannian. We consider an intrinsic Grassmann average (IGA), i.e. an average using the geodesic distance, which allow us to consider $K > 1$ dimensional subspaces. We then propose an *online linear subspace learning* and show that for the zero-mean Gaussian data, the expected IGA on $Gr(K, D)$, i.e., expected K-dimensional linear subspace, coincides with the first K principal components.

Given $\{\mathbf{x}_i\}_{i=1}^N$, the algorithm to compute the IGA to get the leading K-dimensional principal subspace is sketched in Algorithm 1.

Algorithm 1: The IGA algorithm to compute PCs

Input: $\{\mathbf{x}_i\}_{i=1}^N \subset \mathbf{R}^D, K > 0$
Output: $\{\mathbf{v}_1, \ldots, \mathbf{v}_K\} \subset \mathbf{R}^D$
1 Partition the data $\{\mathbf{x}_j\}_{j=1}^N$ into blocks of size $D \times K$;
2 Let the i^{th} block be denoted by, $X_i = [\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iK}]$;
3 Orthogonalize each block and let the orthogonalized block be denoted by X_i ;
4 Let the subspace spanned by each X_i be denoted by $\mathcal{X}_i \in Gr(K, D)$;
5 Compute IGA, $\mathcal{M}^*$, of $\{\mathcal{X}_i\}$;
6 Return the K columns of an orthogonal basis of $\mathcal{M}^*$; these span the principal K-subspace.

Let $\{\mathcal{X}_i\}$ be the set of K-dimensional subspaces as constructed by IGA in Algorithm 1. Moreover, assume that the maximum principal angle between $\mathcal{X}_i$ and $\mathcal{X}_j$ is $< \pi/2\sqrt{2}$, for all $i \neq j$. This condition is needed to ensure that the IGA exists and is unique on $Gr(K, D)$. The condition can be ensured if the angle between $\mathbf{x}_l$ and $\mathbf{x}_k$ is $< \pi/2\sqrt{2}$, for all $\mathbf{x}_l, \mathbf{x}_k$ belonging to different blocks. For $\mathbf{x}_l, \mathbf{x}_k$ in the same block, the angle must be $< \pi/2$. *Note that, this assumption is needed to prove Theorem 3. In practice, even if IGA is not unique, we find a local minimizer of Eq. 5 [23], which serves as the principal subspace.*

Theorem 3. *(**Relation between IGA and PCA**) Let us assume that $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \Sigma)$, for all i. Using the same notations as above, the j^{th} column of M converges to the j^{th} principal vector of $\{\mathbf{x}_i\}_{i=1}^N$, $j = 1, \ldots, K$ as $N \to \infty$, i.e., in the limit, M spans the principal K-subspace, $\mathcal{M}^*$, where $\mathcal{M}^*$ is defined as in Eq. 5.*

Proof. Let X_i be the corresponding orthonormal basis of $\mathcal{X}_i$, i.e., X_i spans $\mathcal{X}_i$,for all i. The IGA, $\mathcal{M}^*$ can be computed using Eq. 5. Let, $X_i = [\mathbf{x}_{i1} \ldots \mathbf{x}_{iK}]$ where and let $\mathbf{x}_{ij}$ be samples drawn from $N(\mathbf{0}, \Sigma)$. Let, $M = [M_1 \ldots M_K]$ be an orthonormal basis of $\mathcal{M}^*$. The distance between $\mathcal{X}_i$ and $\mathcal{M}^*$ is defined as $d^2(\mathcal{X}_i, \mathcal{M}^*) = \sum_{j=1}^K (\arccos((S_i)_{jj}))^2$, where $\bar{U}_i S_i \bar{V}_i^T = M^T X_i$ be the SVD, and $(S_i)_{jj} \geq 0$ (we use $(A)_{lm}$to denote $(l, m)^{th}$ entry of matrix A). As arccos is a decreasing function and a bijection on $[0, 1]$, we can write an alternative form of Eq. 5 as follows:

$$\mathcal{M}^* = \underset{\mathcal{M}}{\operatorname{argmax}} \sum_{i=1}^N \sum_{j=1}^K ((S_i)_{jj})^2 \tag{7}$$

In fact the above alternative form can also be derived using a Taylor expansion of the RHS of Eq. 5. Note that, in the above equation S_i is a function of M. It is easy to see that $(M^T X_i)_{lm} \sim \mathcal{N}(0, \sigma^2_{M_l})$, $l = 1, \ldots, K$, $m = 1, \ldots, K$. Also, $(M^T X_i \bar{V}_i)_{lm} \sim \mathcal{N}(0, \sigma^2_{M_l})$, $l = 1, \ldots, K, m = 1, \ldots, K$ as $\bar{V}_i$ is orthogonal. Thus, $(S_i)_{ll} = (\bar{U}_i^T M^T X_i \bar{V}_i)_{ll} \sim \mathcal{N}(0, \sigma^2_{\bar{U}_{il} M_l})$. So, $\sum_{j=1}^K (S_i)_{jj}^2 \sim \Gamma(\frac{1}{2} \sum_{j=1}^K \sigma^2_{\bar{U}_{ij} M_j}, 2)$ and $E[\sum_{j=1}^K (S_i)_{jj}^2] = \sum_{j=1}^K \sigma^2_{\bar{U}_{ij} M_j}$. Now, as $N \to \infty$, RHS of Eq. 7 becomes $E[\sum_{j=1}^K (S_i)_{jj}^2]$. In order to maximize $E[\sum_{j=1}^K (S_i)_{jj}^2] = \sum_{j=1}^K \sigma^2_{\bar{U}_{ij} M_j}$, $\bar{U}_{ij}$ should be the left singular vectors of $M^T X_i$, and M_j should be the j^{th} eigenvector of Σ, for all $j = 1, \ldots, K$. Hence, M spans the principal subspace, $\mathcal{M}^*$. ■

Now, using Theorem 1 and Theorem 3, we replace the line 5 of the IGA Algorithm 1 by Eq. 6 to get an *online subspace learning* algorithm that we call, *Recursive IGA (RIGA)*, to compute leading K principal components, $K \geq 1$.

3. A Robust Online Linear Subspace Learning Algorithm

Let $\{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_N\} \subset Gr(K, D), K < D$ be inside a regular geodesic ball of radius $< \pi/2\sqrt{2}$ s.t., the Fréchet Median (FMe) exists and is unique. Let $X_1, X_2, \cdots, X_N$ be the corresponding orthonormal bases, i.e., X_i spans $\mathcal{X}_i$, for all i. The FMe can be computed via the following minimization:

$$\mathcal{M}^* = \arg\min_{\mathcal{M}} \sum_{i=1}^N d(\mathcal{X}_i, \mathcal{M}) \tag{8}$$

With a slight abuse of notation, we use the notation $\mathcal{M}^*$ (M) to denote both the FM and the FMe (and their orthonormal basis). The FMe is robust as was shown in [14], hence we call our estimator *Robust IGA* (RoIGA). In the following theorem, we will prove that RoIGA leads to the robust PCA in the limit as the number of the data samples goes to infinity. An algorithm to compute RoIGA is obtained by simply replacing Step 5 of Algorithm 1 by computation of RoIGA via minimization of Eq. 8 instead of Eq. 5. This minimization can be achieved using the Riemannian steepest descent, but instead, here we use the stochastic gradient descent of batch size 5 to compute RoIGA. As at each iteration, we need to store only 5 samples, the algorithm is online. The update step for each iteration of the online algorithm to compute RoIGA (we refer to our online RoIGA algorithm as Recursive RoIGA (RRIGA)) is as follows:

$$\mathcal{M}_1 = \mathcal{X}_1\ , \quad \mathcal{M}_{k+1} = \mathrm{Exp}_{\mathcal{M}_k}\Big(\frac{\mathrm{Exp}^{-1}_{\mathcal{M}_k}(\mathcal{X}_{k+1})}{(k+1)d(\mathcal{M}_k, \mathcal{X}_{k+1})}\Big) \tag{9}$$

where, $k \geq 1$, Exp and Exp^{-1} are Riemannian exponential and inverse exponential functions (see supplementary section for the definition of these maps). We refer the readers to [5] for the consistency proof of the estimator.

Theorem 4. ***(Robustness of RoIGA)*** *Assuming the above hypotheses and notations, as $N \to \infty$, the columns of M converge to the robust principal vectors of the $\{\mathbf{x}_i\}_{i=1}^N$, where M is the orthonormal basis of $\mathcal{M}^*$ as defined in Eq. 8.*

Proof. Let, $X_i = [\mathbf{x}_{i1} \cdots \mathbf{x}_{iK}]$ and $\mathbf{x}_{ij}$ be i.i.d. samples drawn from $N(\mathbf{0}, \Sigma)$. Let, $M = [M_1 \cdots M_K]$ be an orthonormal basis of $\mathcal{M}$. Define the distance between $\mathcal{X}_i$ and $\mathcal{M}$ by $d(\mathcal{X}_i, \mathcal{M}) = \sqrt{\sum_{j=1}^K (\arccos((S_i)_{jj}))^2}$, where $\bar{U}_i S_i V_i^T = M^T X_i$ be the SVD, and $(S_i)_{jj} \geq 0$. Since $\arccos$ is a decreasing function and is a bijection on $[0, 1]$, we can rewrite Eq. 8 alternatively as follows:

$$\mathcal{M}^* = \arg\max_{\mathcal{M}} \sum_{i=1}^N \sqrt{\sum_{j=1}^K ((S_i)_{jj})^2} \tag{10}$$

In fact the above alternative form can also be derived using a Taylor expansion of the RHS of Eq. 8.

From the proof of Theorem 3, we know that $\sum_{j=1}^K ((S_i)_{jj})^2 \sim \Gamma(\frac{1}{2}\sum_{j=1}^K \sigma^2_{\bar{U}_{ij}M_j}, 2)$. So, $\sqrt{\sum_{j=1}^K ((S_i)_{jj})^2} \sim N_g(\frac{1}{2}\sum_{j=1}^K \sigma^2_{\bar{U}_{ij}M_j}, \sum_{j=1}^K \sigma^2_{\bar{U}_{ij}M_j})$, where N_g is the Nakagami distribution [27]. Now, as $N \to \infty$, the RHS of Eq. 10 becomes $E[\sqrt{\sum_{j=1}^K ((S_i)_{jj})^2}]$. $E[\sqrt{\sum_{j=1}^K ((S_i)_{jj})^2}] = \sqrt{2}\Gamma(\sum_{j=1}^K \sigma^2_{\bar{U}_{ij}M_j} + 0.5)/\Gamma(\sum_{j=1}^K \sigma^2_{\bar{U}_{ij}M_j})$, where Γ is the well known gamma function. It is easy to see that as Γ is an increasing function, $E[\sqrt{\sum_{j=1}^K ((S_i)_{jj})^2}]$ is maximized *iff* $\sum_{j=1}^K \sigma^2_{\bar{U}_{ij}M_j}$ is maximized, i.e., when M spans the principal K-subspace.

Now, if we contrast with the objective function of RIGA in Eq. 7, there we had to maximize $E[\sum_{j=1}^K ((S_i)_{jj})^2] = \sum_{j=1}^K \sigma^2_{\bar{U}_{ij}M_j}$. Thus, $E[\sqrt{\sum_{j=1}^K ((S_i)_{jj})^2}] = \rho(\mathfrak{m}) \triangleq \sqrt{2}\Gamma(\mathfrak{m} + 0.5)/\Gamma(\mathfrak{m})$, where $\mathfrak{m} = \sum_{j=1}^K \sigma^2_{\bar{U}_{ij}M_j}$. Hence, the *influence function* [21] of ρ is proportional to $\psi(\mathfrak{m}) \triangleq \frac{\partial E[\sqrt{\sum_{j=1}^K ((S_i)_{jj})^2}]}{\partial \mathfrak{m}}$ and if we can show that $\lim_{\mathfrak{m}\to\infty} \psi(\mathfrak{m}) = 0$, then we can claim that our objective function in Eq. 10 is robust [21].

Now, $\psi(\mathfrak{m}) = \Gamma(\mathfrak{m})\Gamma(\mathfrak{m} + 0.5)\frac{\phi(\mathfrak{m}+0.5)-\phi(\mathfrak{m})}{\Gamma(\mathfrak{m})^2}$, where ϕ is the polygamma function [1] of order 0. After some simple calculations, we get,

$$\begin{aligned}\lim_{\mathfrak{m}\to\infty} (\phi(\mathfrak{m} + 0.5) - \phi(\mathfrak{m})) &= \lim_{\mathfrak{m}\to\infty} \log(1 + 1/(2\mathfrak{m})) \\ &+ \lim_{\mathfrak{m}\to\infty} \sum_{k=1}^{\infty} \left(B_k \left(\frac{1}{k\mathfrak{m}^k} - \frac{1}{k(\mathfrak{m} + 0.5)^k} \right) \right) \\ &= \lim_{\mathfrak{m}\to\infty} \log(1 + 1/(2\mathfrak{m})) + 0 = 0\end{aligned}$$

Here, $\{B_k\}$ are the Bernoulli numbers of the second kind [32]. So, $\lim_{\mathfrak{m}\to\infty} \psi(\mathfrak{m}) = 0$. ■

We would like to point out that the outlier corrupted data can be modeled using a mixture of independent random variables, Y_1, Y_2, where $Y_1 \sim \mathcal{N}(\mathbf{0}, \Sigma_1)$ (to model non-outlier data samples) and $Y_2 \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma_2)$ (to model outliers), i.e., $(\forall i)$, $\mathbf{x}_i = w_1 Y_1 + (1 - w_1)Y_2$, $w_1 > 0$ is generally large, so that the probability of drawing outliers is low. Then as the mixture components are independent, $(\forall i)$, $\mathbf{x}_i \sim \mathcal{N}((1 - w_1)\boldsymbol{\mu}, w_1^2\Sigma_1 + (1 - w_1)^2\Sigma_2)$. A basic assumption in any online PCA algorithm is that data is centered. So, in case the data is not centered (similar to the model of $\mathbf{x}_i$), the first step of PCA would be to centralize the data. But then the algorithm can not be made online, hence our above assumption that $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \Sigma)$ is a valid assumption in an online scenario. But, in a general case, after centralizing the data as the first step of PCA, the above theorem is valid.

4. Experimental Results

In this section, we present an experimental evaluation of the proposed estimators on both real and synthetic data. Our overall experimental findings are that the RIGA and RRIGA estimators are more accurate and faster than other online linear and robust linear subspace estimators. We believe that the higher accuracy in RIGA and RRIGA can be attributed to the use of intrinsic geometry of the Grassmannian in our geometric formulation. Specifically, finding the full set of PCs is

cast as an intrinsic averaging problem on the Grassmannian achieved using a recursive estimator in both cases. From a computational perspective, in the online PCA case, we attribute the efficiency observed in the experiments to RIGA being an optimization and parameter free method. In the case of RRIGA, the reasons for accuracy and efficiency are much more complicated. At this juncture, we speculate the reason to be that our geometric formulation leads to directly finding the subspaces using a recursive scheme as opposed to methods that incrementally update basis of the subspace in an alternating fashion with no convergence guarantees. In the following, we consider RIGA and RRIGA separately.

4.1. Online Linear Subspace Estimation

Baselines: Here, we present a comparison with Oja's rule and the online version of EM-PCA (Sec. 1.1). For Oja's rule we follow common guidelines and consider step-sizes $\gamma_t = {}^{\alpha}/_{D\sqrt{t}}$ with α-values between 0.005 and 0.2. For EM-PCA we follow the recommendations from Cappé [9] and use step-sizes $\gamma_t = {}^{1}/_{t^{\alpha}}$ with α-values between 0.6 and 0.9 along with Polyak-Ruppert averaging. For GROUSE, we have chosen the stepsize to be 0.1.

(Synthetic) Gaussian Data: Theorem 3 states that the RIGA estimates coincide in expectation with the leading principal subspace when the data are drawn from a zero-mean Gaussian distribution. We empirically verify this for an increasing number of observations drawn from randomly generated zero-mean Gaussians. We measure the *expressed variance* which is the variance captured by the estimated subspace divided by the variance captured by the true principal subspace:

$$\text{Expressed Variance} = \frac{\sum_{k=1}^{K}\sum_{n=1}^{N}\mathbf{x}_n^T\mathbf{v}_k^{(\text{est})}}{\sum_{k=1}^{K}\sum_{n=1}^{N}\mathbf{x}_n^T\mathbf{v}_k^{(\text{true})}} \in [0,1].$$

An expressed variance of 1 implies that the estimated subspace captures as much variance as the principal subspace. The top panel of Fig. 2 shows the mean ($\pm$ one standard deviation) expressed variance of RIGA over 150 trials. It is evident that for the Gaussian data, the RIGA estimator does indeed converge to the true principal subspace.

A key aspect of any online estimator is that it should be stable and converge fast to a good estimate. Here, we compare RIGA to the above-mentioned baselines. Both Oja's rule and EM-PCA require a step-size to be specified, so we consider a larger selection of such step-sizes. The middle panel of Fig. 2 shows the expressed variance as a function of number of observations for different estimators and step-sizes. EM-PCA was found to be quite stable with respect to the choice of step-size, though it does not seem to converge to a good estimate. Oja's rule, on the other hand, seems to converge to a good estimate, but its practical performance is critically dependent on the step-size. GROUSE is seen to oscillate for small data size however, with a large number of samples, it yields a good estimate. On the other hand, RIGA is parameter-free and is observed to have good convergence properties.

In the bottom panel of Fig. 2, we perform a stability analysis of GROUSE and RIGA. Here, for a fixed value of N, we generate a data matrix and perform 200 independent runs on the data matrix and report the mean ($\pm$ one standard deviation) expressed variance. As can be seen from the figure, RIGA is very stable in comparison to GROUSE.

Human Body Shape: Online algorithms are generally well-suited for solving large-scale problems as by construction, they should have linear time-complexity in the number of observations. As an example, we consider a large collection of three-dimensional scans of human body shape [31]. This dataset contains $N = 21862$ meshes which each consist of 6890 vertices in $\mathbf{R}^3$. Each mesh is, thus, viewed as a $D = 6890 \times 3 = 20670$ vector. We estimate a $K = 10$ dimensional principal subspace using Oja's rule, EM-PCA, GROUSE and RIGA respectively. The average reconstruction error (squared distance between the original data and its estimate) over all meshes are 16.8 mm for Oja's rule, 1.9 mm for EM-PCA, 1.4 mm for GROUSE, and 1.0 mm for RIGA. *Note that both Oja's rule and EM-PCA explicitly minimize the reconstruction error, while RIGA does not but yet outperforms the baseline methods.* We speculate that this is due to RIGA's excellent convergence properties and it being a parameter free algorithm is not bogged down by the hard problem of step-size tuning confronted in the baseline algorithms used here.

Santa Claus Conquers the Martians: We now consider an even larger scale experiment and consider all frames of the motion picture *Santa Claus Conquers the Martians (1964)*[2]. This consist of $N = 145,550$ RGB frames of size 320×240, corresponding to an image dimension of $D = 230,400$. We estimate a $K = 10$ dimensional subspace using Oja's rule, EM-PCA, GROUSE and RIGA respectively. Again, we measure the accuracy of the different estimators via the reconstruction error. Pixel intensities are scaled to be between 0 and 1. Oja's rule gives an average reconstruction error of 0.054, EM-PCA gives 0.025, while RIGA and GROUSE give 0.023. Here RIGA and EM-PCA give roughly equally good results, with a slight advantage to RIGA. GROUSE gives same reconstruction error as RIGA. Oja's rule does not fare as well. As with the shape data, it is interesting to note that RIGA outperforms some of the other baseline methods on the error measure that they optimize even though RIGA optimizes a different measure.

[2] https://archive.org/details/SantaClausConquerstheMartians1964

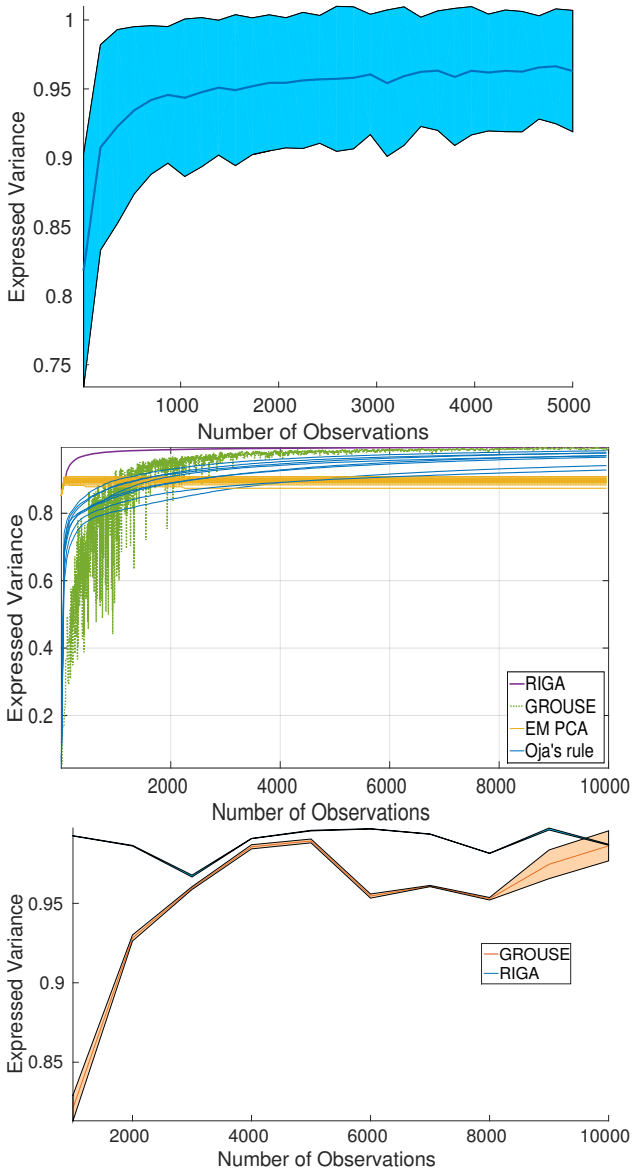

Figure 2. Expressed variance as a function of number of observations. *Top:* The mean and one standard deviation of the RIGA estimator computed over 150 trials. In each trial data are generated in $\mathbf{R}^{50}$ and we estimate a $K = 2$ dimensional subspace. *Middle:* The performance of different estimators for varying step-sizes. Data are generated in $\mathbf{R}^{250}$ and we set $K = 20$. In both experiments, we observe similar trends with other values of D and K. *Bottom:* Stability analysis comparison of GROUSE and RIGA (for a fixed N, we randomly generate a data matrix, X, from a Gaussian distribution on $\mathbf{R}^{250}$. we estimate $K = 20$ dimensional subspace and report the mean and one standard deviation over 200 runs on X.)

4.2. Robust Subspace Estimation

We now present the comparative experimental evaluation of robust extension (RRIGA). Here we use DHR-PCA and GRASTA as baseline and measure performance using the *reconstruction error (RE)*. We have used UCSD anomaly detection database [26] and Extended YaleB database [16].

UCSD anomaly detection database: This data contains images of pedestrian movement on walkways captured by a stationary mounted camera. The crowd density on the walkway varies from sparse to very crowded. The anomaly includes bikers, skaters, carts, people in wheelchair etc.. This database is divided in two sets: "Peds1" (people are walking towards the camera) and "Peds2" (people are walking parallel to the camera plane). In "Peds1" there are 36 training and 34 testing videos where each video contains 180 frames of dimension 158×238 ($D = 37604$). In "Peds2" there are 12 training and 16 testing videos containing varying samples of dimension 240×360 ($D = 86400$). The test frames do not have anomalous activities. Some sample frames (with and without outliers) are shown in Fig. 3. We first extract K principal components on the training data (including anomalies) and then compute reconstruction error on the test frames (without anomalies) using the computed principal components. It is expected that if the PC computation technique is robust, the reconstruction error will be good as PCs should not be affected by the anomalies in training samples. In Fig. 4, we compare performance of RRIGA with GRASTA and DHR-PCA in terms of RE and time required by varying K from 1 to 100. In terms of time it is evident that RRIGA is very fast compared to both GRASTA and DHR-PCA. RRIGA also outperforms both DHR-PCA and GRASTA in terms of RE. Moreover, it is evident that RRIGA scales very well both in terms of RE and computation time unlike it's competitors.

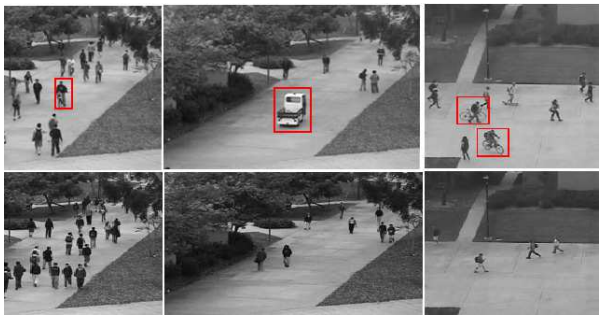

Figure 3. top and bottom row contains outliers (identified in a rectangular box) and non-outliers frames of UCSD anomaly data respectively

Yale ExtendedB database: This data contains 2414 face images of 38 subjects. We crop each image to make a 32×32 images ($D = 1024$). Due to varying lighting condition, some of the face images are shaded/ dark and appeared as outliers (this experimental setup is similar to the one in [22]). In Fig. 5 some sample face images (outlier and non-outlier) are shown. One can see that due to poor lighting condition, though the middle face in top row is a face image, it looks

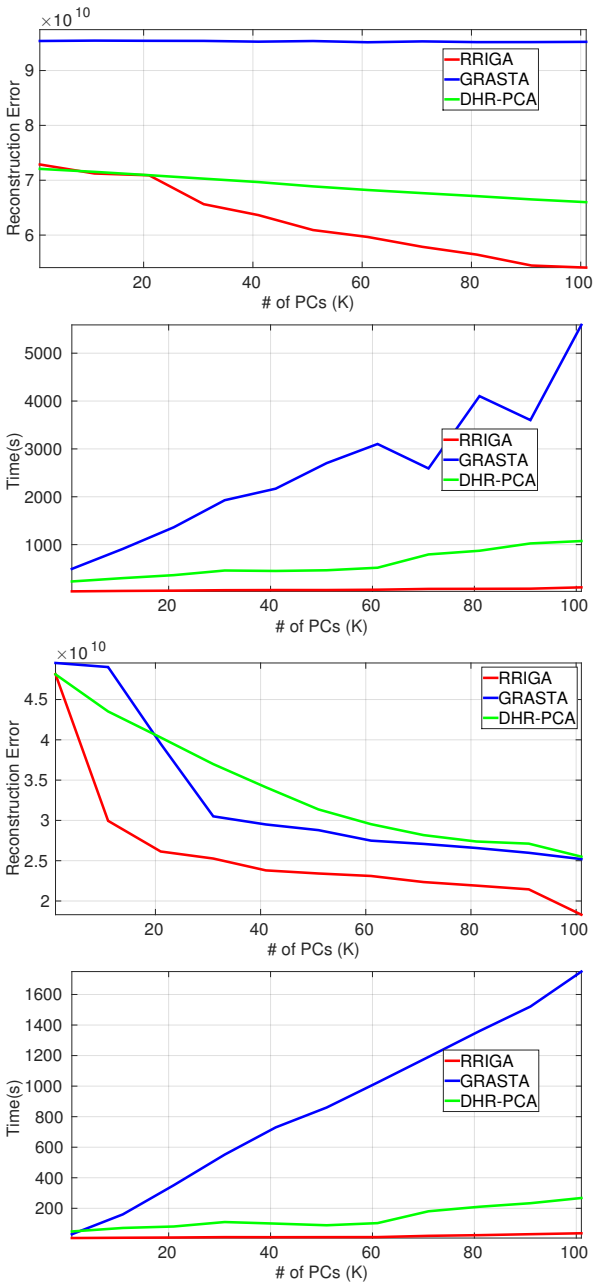

Figure 4. *Top two:* on "Peds1" anomaly data; *Bottom two:* on "Peds2" anomaly data

completely dark and an outlier. For testing, we have used 142 non-outlier face images of 38 subjects and the rest we used to extract PCs. We report RE (with varying K) and time required for both RRIGA, GRASTA and DHR-PCA in Fig. 6. From the figure it is evident that for small number of PCs (i.e., small K) RRIGA performs similar to DHR-PCA, while for larger K values, RRIGA outperforms DHR-PCA and GRASTA. In terms of time required, RRIGA is faster than both DHR-PCA and GRASTA.

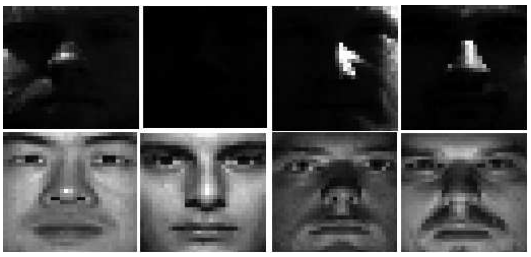

Figure 5. top and bottom row contains outliers and non-outliers images of YaleExtendedB data respectively

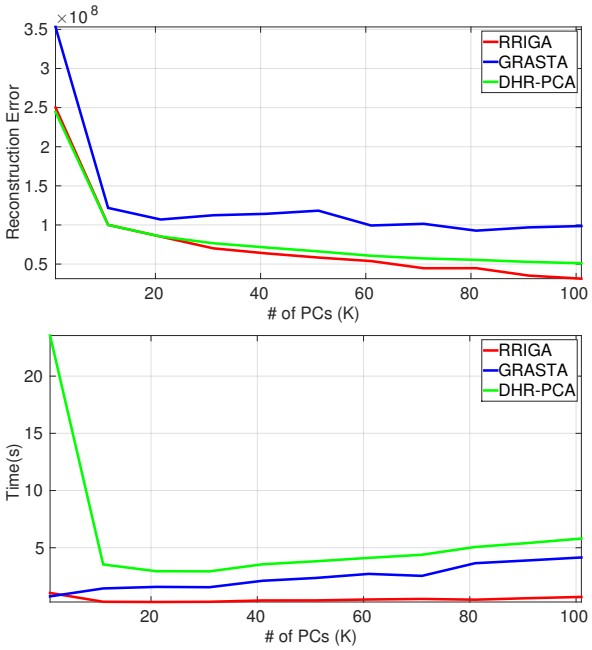

Figure 6. Performance comparison on YaleExtendedB data

5. Conclusions

In this paper, we present a new geometric framework for estimating the full set of principal components from given data. We present two online algorithms, for estimating PCA and RPCA. Since they are inherently online, they are naturally scalable to very large data sets as demonstrated in the experimental results section. The key idea in the geometric framework involves computing an intrinsic Grassmann average as a proxy for the principal linear subspace. We show that the if the samples are drawn from a Gaussian distribution, the intrinsic Grassmann average coincides with the principal subspace in expectation. Further, for our online recursive RPCA algorithm, we proved that the estimated principal components are statistically robust. Our algorithms have a linear time complexity and linear convergence rate. Unlike most other online algorithms there are not step-sizes or other parameters to tune; a most useful property in practical settings. Our future work will focus on application of our geometric approach to the matrix completion problem.

References

[1] M. Abramowitz, I. A. Stegun, et al. Handbook of mathematical functions. *Applied mathematics series*, 55:62, 1966. 5

[2] P.-A. Absil, R. Mahony, and R. Sepulchre. Riemannian geometry of grassmann manifolds with a view on algorithmic computation. *Acta Applicandae Mathematicae*, 80(2):199–220, 2004. 3

[3] B. Afsari. Riemannian lp center of mass: Existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 139(2):655–673, 2011. 3

[4] L. Balzano, R. Nowak, and B. Recht. Online identification and tracking of subspaces from highly incomplete information. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 704–711. IEEE, 2010. 2

[5] S. Bonnabel. Stochastic gradient descent on riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013. 5

[6] C. Boutsidis, D. Garber, Z. Karnin, and E. Liberty. Online principal components analysis. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 887–901. SIAM, 2015. 2

[7] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 2

[8] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011. 2

[9] O. Cappé. Online expectation-maximisation. *Mixtures: Estimation and Applications*, pages 31–53, 2011. 2, 6

[10] R. Chakraborty and B. C. Vemuri. Recursive frechet mean computation on the grassmannian and its applications to computer vision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4229–4237, 2015. 3, 4

[11] Y. Chikuse. *Statistics on Special Manifolds*. Springer, February 2003. 3

[12] J. Feng, H. Xu, and S. Yan. Robust pca in high-dimension: A deterministic approach. *arXiv preprint arXiv:1206.4628*, 2012. 2

[13] J. Feng, H. Xu, and S. Yan. Online robust pca via stochastic optimization. In *Advances in Neural Information Processing Systems*, pages 404–412, 2013. 2

[14] P. T. Fletcher, S. Venkatasubramanian, and S. Joshi. The geometric median on riemannian manifolds with application to robust atlas estimation. *NeuroImage*, 45(1):S143–S152, 2009. 5

[15] M. Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. In *Annales de l'institut Henri Poincaré*, volume 10, pages 215–310. Presses universitaires de France, 1948. 3

[16] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE transactions on pattern analysis and machine intelligence*, 23(6):643–660, 2001. 7

[17] W. Ha and R. F. Barber. Robust pca with compressed data. In *Advances in Neural Information Processing Systems*, pages 1936–1944, 2015. 2

[18] S. Hauberg, A. Feragen, R. Enficiaud, and M. J. Black. Scalable robust principal component analysis using grassmann averages. *TPAMI*, 2015. 1, 2, 4

[19] J. He, L. Balzano, and A. Szlam. Incremental gradient on the grassmannian for online foreground and background separation in subsampled video. In *CVPR*, pages 1568–1575, 2012. 2

[20] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933. 1

[21] P. J. Huber. *Robust statistics*. Springer, 2011. 5

[22] W. Jiang, F. Nie, and H. Huang. Robust dictionary learning with capped l 1-norm. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 3590–3596. AAAI Press, 2015. 7

[23] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977. 3, 4

[24] W. S. Kendall. Probability, convexity, and harmonic maps with small image i: uniqueness and fine existence. *Proceedings of the London Mathematical Society*, 3(2):371–406, 1990. 3

[25] B. Lois and N. Vaswani. Online robust pca and online matrix completion. *arXiv preprint arXiv:1503.03525*. 2

[26] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, volume 249, page 250, 2010. 7

[27] M. Nakagami. The m-distribution-a general formula of intensity distribution of rapid fading. *Statistical Method of Radio Propagation*, 1960. 5

[28] E. Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273, 1982. 1

[29] K. Pearson. On lines and planes of closest fit to system of points in space. *Philosophical Magazine*, 2(11):559–572, 1901. 1

[30] X. Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006. 3

[31] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black. Dyna: A model of dynamic human shape in motion. *SIGGRAPH*, 34(4):120:1–120:14, Aug. 2015. 6

[32] S. Roman. *The umbral calculus*. Springer, 2005. 5

[33] S. Roweis. EM algorithms for pca and spca. *Advances in neural information processing systems*, pages 626–632, 1998. 2

[34] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999. 1

[35] Y.-C. Wong. Sectional curvatures of grassmann manifolds. *Proceedings of the National Academy of Sciences*, 60(1):75–79, 1968. 3

[36] X. Xu. Online robust principal component analysis for background subtraction: A system evaluation on toyota car data. 2014. 2