Approximate Grassmannian Intersections: Subspace-Valued Subspace Learning

Calvin Murdock
Carnegie Mellon University
Machine Learning Department
cmurdock@cs.cmu.edu

Fernando De la Torre
Carnegie Mellon University
Robotics Institute
ftorre@cs.cmu.edu

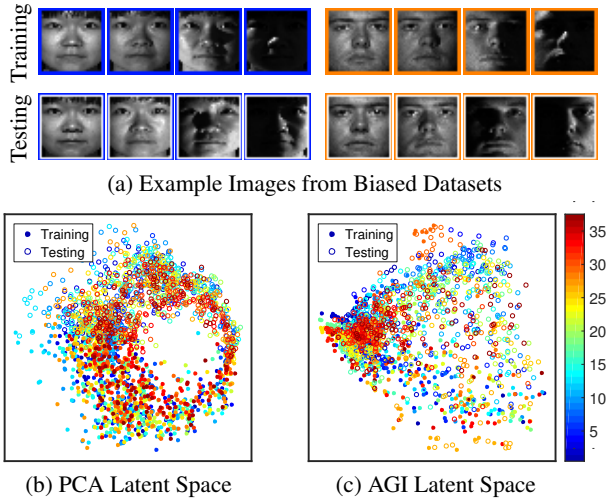
Abstract

Subspace learning is one of the most foundational tasks in computer vision with applications ranging from dimensionality reduction to data denoising. As geometric objects, subspaces have also been successfully used for efficiently representing certain types of invariant data. However, methods for subspace learning from subspace-valued data have been notably absent due to incompatibilities with standard problem formulations. To fill this void, we introduce Approximate Grassmannian Intersections (AGI), a novel geometric interpretation of subspace learning posed as finding the approximate intersection of constraint sets on a Grassmann manifold. Our approach can naturally be applied to input subspaces of varying dimension while reducing to standard subspace learning in the case of vector-valued data. Despite the nonconvexity of our problem, its globally-optimal solution can be found using a singular value decomposition. Furthermore, we also propose an efficient, general optimization approach that can incorporate additional constraints to encourage properties such as robustness. Alongside standard subspace applications, AGI also enables the novel task of transfer learning via subspace completion. We evaluate our approach on a variety of applications, demonstrating improved invariance and generalization over vector-valued alternatives.

1. Introduction

Understanding the structure of data is one of the most fundamental problems in machine learning. Surprisingly, despite their apparent simplicity, models that learn linear subspaces have achieved remarkable success in areas such as dimensionality reduction [11], data denoising [15], collaborative filtering [18], and many others.

Within computer vision, this could partially be attributed to the observation that real data tend to concentrate near lower-dimensional (locally) linear structures. This can arise naturally for a variety of reasons. For example, harmonic analysis of Lambertian reflectance functions has demonstrated that third-order approximations can account for over 99% of the energy, allowing aligned images of convex, Lambertian objects under any illumination to be succinctly summarized by 9-dimensional subspaces [4]. This has led to the practical success of linear component analysis methods for automatic facial analysis [16].



(a) Example Images from Biased Datasets

(b) PCA Latent Space (c) AGI Latent Space

Figure 1: A comparison between PCA and AGI applied to the task of dimensionality reduction. For each method, a subspace was learned from training images with lighting conditions that differed substantially from those in the testing images. Examples are shown (a) alongside two-dimensional Isomap [35] embeddings of the associated latent spaces (b,c). Unlike PCA, AGI provides some invariance to shared variations due to lighting by representing the data as *subspaces*, resulting in discriminative representations that contain more information relating to identity.

Of course, data linearity is often satisfied only in restricted, highly-controlled scenarios. With faces, even slight image misalignment or appearance variability can introduce nonlinearities that adversely affect subspace-based methods [23]. However, subspaces can provide efficient and robust representations of *sets* of data points that span subspaces invariant to certain linear transformations. This is especially useful in scenarios in which sufficient training data is unavailable; subspace representations can generalize a

2380-7504/17 $31.00 © 2017 IEEE
DOI 10.1109/ICCV.2017.462

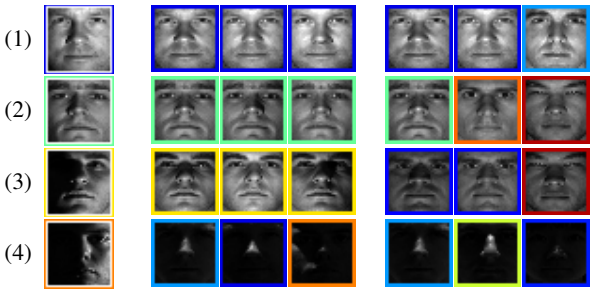


Figure 2: A demonstration of the improved invariance provided by AGI representations. From the example in Fig. 1, a number of testing images are shown (a) beside their 3 nearest neighbors in the latent spaces of AGI (b) and PCA (c). Despite significant lighting variations, the AGI neighbors belong to the same class more consistently in comparison to PCA.

Figure 3: A visualization of subspace completion applied to the task of image relighting. From single images of a novel subject not included during training (b, left), higher-dimensional illumination subspaces are inferred and information from source testing images (a) is transferred to generate images under novel lighting conditions (b, right). The results match closely with the ground truth (c), though harsh shadows can cause blur and artifacts.

small set of training vectors to one that is effectively infinite. An example of this is shown in Fig. 1 where subspace learning is used to reduce the dimensionality of face images with different lighting conditions between the training and testing sets. Image representations from standard vector-valued subspace learning methods like PCA fail to generalize well beyond the training data because a single low-dimensional subspace is insufficient for modeling faces of multiple individuals lit from a variety of directions. However, the data can be well approximated by a *mixture* of low-dimensional subspaces, one for each individual. From this observation, our approach leverages the known linear structure of faces to learn representations that better encode lighting-invariant information related to image identity, as demonstrated in Fig. 2. Instead of learning from vector-valued data in Euclidean space, this suggests learning from subspace-valued data lying on the Grassmannian, a nonlinear manifold that parametrizes the set of all subspaces.

While a number of machine learning techniques for subspace-valued data have been explored in a variety of contexts [19, 36, 20, 24], fundamental tasks such as dimensionality reduction, denoising, and missing data imputation have been relatively unexplored. As these problems are all naturally amenable to solutions using low-rank linear models, we propose an approach for learning subspaces from subspace-valued data. Analogous to standard vector-valued approaches, we aim to learn subspaces that approximately contain all of the training data. However, due to the ambiguity inherent in representing and manipulating subspaces numerically, standard computational machinery like eigendecompositions of sample covariance matrices and low-rank matrix approximations cannot be applied directly.

To address this issue, we introduce Approximate Grassmannian Intersections (AGI), a novel geometric framework for subspace learning posed as finding the approximate intersection of sets on a Grassmann manifold. Specifically, AGI treats each training example as a constraint set for the learned subspace, essentially enforcing that the data be contained within it. To account for nonlinearities, we approximate these constraints by minimizing their average distance to the learned subspace. Though nonconvex, the resulting optimization problem admits a globally-optimal solution that can be efficiently found using a singular value decomposition, reducing to standard principal subspace learning in the special case of vector-valued input data represented as weighted one-dimensional subspaces. This approach naturally supports standard subspace modeling applications like linear dimensionality reduction, but can be applied to input data given as subspaces of variable dimension. Furthermore, our geometric interpretation enables subspace completion, the novel task of transfer learning to increase the dimensionality of an input subspace. It can be applied, for example, in appearance-based image relighting by generating illumination subspaces from single input images, as demonstrated in Fig. 3. Our formulation easily supports additional constraints for incorporating prior knowledge and can be solved with a simple optimization procedure based on the theory of best approximation with iterative projections [5]. To demonstrate the wide applicability of AGI, we evaluate its effectiveness in a variety of applications, including dimensionality reduction, visualization, transfer learning, and classification.

2. Related Work

In this section, we provide a brief overview of previous research on subspace learning and matrix factorization, representations of subspace-valued data on the Grassmann

manifold, and iterative projection algorithms for finding approximate intersections of sets.

Subspace learning has a long history in the fields of statistics, signal processing, and computer vision. Originally introduced by Karl Pearson [33] in 1901, the prototypical subspace learning method of principal component analysis (PCA) [26] began an influential sequence of algorithms motivated by fundamental applications like regression [33], classification [34], clustering [37], and dimensionality reduction [27]. To encourage learning more meaningful low-dimensional representations without supervision, the task of low-rank matrix approximation has been extended to more general matrix factorization techniques in which additional constraints are imposed on the basis vectors or coefficients. Examples include sparse PCA [40], sparse dictionary learning [28], non-negative matrix factorization [7], and robust PCA [8]. Often implemented using alternating optimization algorithms due to the typical biconvexity of the objective functions, convergence can be slow and prone to getting trapped in poor local minima.

Thus, recent approaches have considered techniques that explicitly consider the geometry of the underlying manifold. A survey of linear dimensionality methods solved via optimization programs over matrix manifolds can be found in [11]. Other example include GROUSE [3] and GRASTA [22], online algorithms for subspace identification and tracking derived from incremental gradient descent on the Grassmannian. Despite the non-convexity of subspace estimation, this approach has theoretically attractive convergence properties with global optimality guarantees in the case of noiseless data [39].

Interpreting subspaces as points on a Grassmann manifold has also allowed for more effective representations of data that naturally exhibit invariance to certain linear transformations. Example applications include the classification of linear dynamical systems represented as subspaces defined through their observability matrices [25] and affine-invariant regression of shape data [24]. Furthermore, in [36] the authors present a statistical framework for inference on the Grassmannian by considering the underlying manifold geometry and describe a variety of computer vision applications that can be viewed from this paradigm, demonstrating improved performance in both supervised and unsupervised learning. Similar ideas have been used in designing kernels for the pairwise comparison of subspaces, leading to Grassmannian extensions of kernel discriminant analysis [19], extrinsic dictionary learning [20], and classifiers such as support vector machines, logistic regression, and partial least squares [31].

Because the Grassmannian is not a vector space, arithmetic operations such as addition and subtraction are not well defined between subspaces. Thus, standard least-squares approaches to subspace learning and component analysis [12] are not directly applicable. While Principal Geodesic Analysis [17] addresses this by linearizing the underlying manifold around an intrinsic mean element, this can be computationally expensive and could only be used with subspace data of the same dimension. Our approach overcomes this issue by introducing a novel formulation for learning based on approximate constraint satisfaction with a solution given by the method of averaged projections [6].

Projection algorithms [5] like this have been applied with great success in the fields of signal processing and optics [9] due to their practical advantages and well-studied theoretical properties [5]. Traditionally applied to convex feasibility problems, they have also been used in applications of best approximation [6] and extended to certain regular non-convex sets such as manifolds [30].

3. Grassmannian Geometry Preliminaries

In this section, we provide a brief overview of Grassmannian geometry. More thorough reviews can be found in [14] and [32], which emphasizes computer vision applications.

The Grassmannian $\mathcal{G}_{k,d}$ is the set of all k-dimensional subspaces of $\mathbb{R}^d$. These subspaces can be identified by any set of k non-coincident vectors contained within the subspace, which makes their numerical representation ambiguous. Thus, the Grassmannian is typically defined as in Eq. 1, where $\mathrm{col}(\cdot)$ denotes the range or column space of a matrix and $\mathcal{V}_{k,d}$ is the Stiefel manifold, which parameterizes the set of all orthogonal matrices in $\mathbb{R}^{d\times k}$.

$$\mathcal{G}_{k,d} = \{\mathrm{col}(\mathbf{B}) : \mathbf{B} \in \mathcal{V}_{k,d}\}, \mathcal{V}_{k,d} = \left\{\mathbf{B} : \mathbf{B}^\mathsf{T}\mathbf{B} = \mathbf{I}\right\} \quad (1)$$

As a compact, smooth manifold, this set is imbued with geometric structure. Specifically for our purposes, this allows for the definition of distances between two points on the Grassmannian. The natural geodesic distance $d_\mathcal{G}$ can be expressed as the Euclidean norm of the vector of principal angles between the corresponding subspaces:

$$d_\mathcal{G}(A, B) = \|\boldsymbol{\theta}\|_2, \quad \mathbf{A}^\mathsf{T}\mathbf{B} = \mathbf{U}\mathrm{diag}(\cos\boldsymbol{\theta})\mathbf{V}^\mathsf{T} \quad (2)$$

Here, $\mathbf{A}$ and $\mathbf{B}$ are orthogonal matrices with columns that span the subspaces A and B respectively. The vector $\boldsymbol{\theta}$ containing the principal angles can by found through a singular value decomposition. However, due to the computational expense required in its evaluation, we instead use the projection F-norm d_P, an alternative distance metric that relies on a bijective, isometric embedding Π of the Grassmann manifold in Euclidean space [14]. This is accomplished by identifying a subspace B with the unique linear operator $\Pi(B)$ that projects any point in $\mathbb{R}^d$ orthogonally onto it:

$$\Pi : \mathcal{G}_{k,d} \to \mathbb{R}^{d\times d}, \quad \Pi(B) = \mathbf{B}\mathbf{B}^\mathsf{T} \quad (3)$$

In Eq. 4, the distance d_P is then defined simply as the Euclidean distance between the subspaces' projection matrices, though it can also be equivalently expressed as the

Euclidean norm of the sine of the principal angles. This demonstrates the equivalence between $d_{\mathcal{G}}$ and d_P for small distances due to the isometry of Π.

$$d_P(A,B) = \|\sin\boldsymbol{\theta}\|_2 = \frac{1}{\sqrt{2}}\|\Pi(A) - \Pi(B)\|_F \quad (4)$$

This provides a convenient method to represent subspaces in Euclidean space as elements of the set of projection matrices $\mathcal{P}_{k,d} = \{\mathbf{B}\mathbf{B}^\mathsf{T} : \mathbf{B} \in \mathcal{V}_{k,d}\}$. This set is equivalently expressed in Eq. 5, showing the symmetry, idempotency, and trace constraints of projection matrices.

$$\mathcal{P}_{k,d} = \{\mathbf{P} \in \mathbb{R}^{d\times d} : \mathbf{P}^\mathsf{T} = \mathbf{P}, \mathbf{P}^2 = \mathbf{P}, \mathrm{tr}(\mathbf{P}) = k\} \quad (5)$$

While this set is non-convex, projection can be easily accomplished using a truncated singular value decomposition. However, for the theoretical optimality guarantees discussed in Sec. 4.4, we will also consider the convex hull of this set, denoted as $\mathcal{F}_{k,d} = \mathrm{conv}(\mathcal{P}_{k,d})$, or equivalently:

$$\mathcal{F}_{k,d} = \{\mathbf{Q} \in \mathbb{R}^{d\times d} : \mathbf{0} \preceq \mathbf{Q} \preceq \mathbf{I}, \mathrm{tr}(\mathbf{Q}) = k\} \quad (6)$$

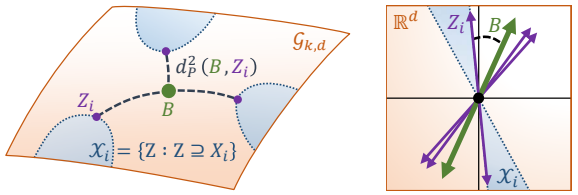
This set, called the Fantope, has been effectively used in applications like sparse PCA [38]. Projection onto it can again be easily accomplished via singular value thresholding.

4. Approximate Grassmannian Intersections

From the perspective of Grassmannian geometry, we now introduce our novel framework for subspace-valued subspace learning. Consider a dataset consisting of subspaces $X_i \in \mathcal{G}_{p_i,d}$ of potentially varying dimension p_i for $i = 1,\ldots,n$. Our goal is to learn a k-dimensional subspace $B \in \mathcal{G}_{k,d}$, where $p_i \leq k < d$ for all i, such that all of the training data X_i are approximately contained within it. To accomplish this, we introduce constraints $\mathcal{X}_i$ for each data point that enforce the data subspaces X_i be contained exactly within a local subspace Z, i.e. $\mathcal{X}_i = \{Z \in \mathcal{G}_{k,d} : Z \supseteq X_i\}$. For example, consider the set of two-dimensional planes containing a fixed, one-dimensional line. While one of the plane's basis vectors must be fixed to be coincident with the line, the other can rotate freely around the line, resulting in a constraint set with one degree of freedom.

AGI attempts to align these local subspaces with a learned global subspace B, but can only satisfy all of these constraints if the input data are exactly contained within some k-dimensional subspace. Due to noise and data nonlinearitites, however, this will likely never be the case. Thus, instead of finding an exact intersection of these constraints, we aim to find an *approximate* intersection by minimizing the average squared distances between B and its projection onto the constraint sets $\mathcal{X}_i$ as in Eq. 7. An illustrative visualization is also shown in Fig. 4.

$$\underset{B,Z_i\in\mathcal{G}_{k,d}}{\arg\min} \sum_{i=1}^{n} d_P^2(B,Z_i) \text{ s.t. } Z_i \in \mathcal{X}_i \quad (7)$$



(a) Grassmannian View (b) Euclidean View

Figure 4: An illustrative overview of our method, comparing a Grassmannian view (a), in which subspaces are represented as points on a manifold, and a Euclidean view (b), in which they are lower-dimensional vector space subsets of $\mathbb{R}^d$. Our goal is to learn a k-dimensional subspace $B \in \mathcal{G}_{k,d}$ (shown in green) that is closest to each local subspace $Z_i \in \mathcal{G}_{k,d}$ (shown in purple), which is constrained to lie within the set $\mathcal{X}_i$ (shown in blue) of all k-dimensional subspaces containing the training data $X_i \in \mathcal{G}_{p_i,d}$.

Note that we approximate the geodesic distance $d_{\mathcal{G}}$ with the more computationally efficient projection F-norm d_P from Eq. 4. As that this relies on an isometric embedding of the Grassmannian in Euclidean space, we also represent subspaces in this same space. Thus, instead of learning a subspace $B \in \mathcal{G}_{k,d}$, we will learn a projection matrix $\mathbf{P} = \mathbf{B}\mathbf{B}^\mathsf{T} \in \mathcal{P}_{k,d}$, where $\mathbf{B} \in \mathcal{V}_{k,d}$ is an orthogonal basis matrix for B. This allows the constraint sets $\mathcal{X}_i$ to be written as:

$$\mathcal{C}_i = \{\mathbf{Q} : \mathbf{Q}\mathbf{X}_i = \mathbf{X}_i\}, \forall i = 1,\ldots,n. \quad (8)$$

To understand these constraints, recall that for the columns of a matrix $\mathbf{X}_i$ to all be contained within some subspace Q, pre-multiplication of the corresponding projection matrix $\mathbf{Q}$ will return the original matrix unchanged. Note that these constraints are now affine with respect to the matrix $\mathbf{Q}$ and our optimization problem in Eq. 7 can be written as:

$$\underset{\mathbf{P},\mathbf{Q}_i\in\mathcal{P}_{k,d}}{\arg\min} \sum_{i=1}^{n} \|\mathbf{P} - \mathbf{Q}_i\|_F^2 \text{ s.t. } \mathbf{Q}_i \in \mathcal{C}_i \quad (9)$$

This can be interpreted as finding the approximate intersection of affine subspaces of *subspaces* mapped to the set of projection matrices in Euclidean space. However, due to the constraints $\mathbf{P}, \mathbf{Q}_i \in \mathcal{P}_{k,d}$, this problem is nonconvex, which typically makes optimization difficult due to the possibility of becoming trapped in poor local minima. However, in our case, the unique global optimum can efficiently be found using a simple singular value decomposition.

4.1. Generalization of Principal Subspace Learning

To show this, we first draw connections to standard vector-valued subspace learning. While the objective in Eq. 9 differs substantially from standard approaches that explicitly minimize data approximation error, it turns out that for the special case of vector-valued data represented as weighted one-dimensional subspaces, AGI is equivalent to principal subspace learning.

Specifically, consider a dataset consisting of vectors $\boldsymbol{x}_i \in \mathbb{R}^d$ for $i = 1, \ldots, n$. Principal subspace learning with PCA can be posed as finding the k-dimensional subspace spanned by the columns of $\mathbf{B} \in \mathcal{V}_{k,d}$ that minimizes the average approximation error:

$$\arg\min_{\mathbf{B} \in \mathcal{V}_{k,d}} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \mathbf{P}\boldsymbol{x}_i\|_2^2 \text{ s.t. } \mathbf{P} = \mathbf{B}\mathbf{B}^\mathsf{T} \tag{10}$$

The solution subspace can then be found as the span of the top k left singular vectors of the matrix $[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$.

If we represent the input data $\boldsymbol{x}_i$ as one-dimensional subspaces spanned by the basis $\mathbf{X}_i = \|\boldsymbol{x}_i\|_2^{-1} \boldsymbol{x}_i \in \mathcal{V}_{1,d}$ associated with weights $a_i^2 = \|\boldsymbol{x}_i\|_2^2$, then:

$$\|\boldsymbol{x}_i - \mathbf{P}\boldsymbol{x}_i\|_2^2 = \min_{\mathbf{Q}_i \in \mathcal{P}_{k,d}} \frac{a_i^2}{2} \|\mathbf{P} - \mathbf{Q}_i\|_F^2 \text{ s.t. } \mathbf{Q}_i \in \mathcal{C}_i \tag{11}$$

To see why this is the case, note that in order to satisfy the equality constraint in Eq. 11, the local subspace projection matrix $\mathbf{Q}_i$ must be decomposed as $\mathbf{Q}_i = \mathbf{X}_i\mathbf{X}_i^\mathsf{T} + \bar{\mathbf{X}}_i\bar{\mathbf{X}}_i^\mathsf{T}$ where $\bar{\mathbf{X}}_i \in \mathcal{V}_{k-1,d}$ and $\bar{\mathbf{X}}_i^\mathsf{T}\mathbf{X}_i = \mathbf{0}$. Furthermore, in order to minimize its distance to $\mathbf{P}$, $\bar{\mathbf{X}}_i$ must be contained within the subspace corresponding to $\mathbf{P}$, i.e. $\mathbf{P}\bar{\mathbf{X}}_i = \bar{\mathbf{X}}_i$. Since projection matrices are idempotent and their trace is equal to the dimensionality of their corresponding subspaces, both sides of Eq. 11 can then be shown to equal to $\boldsymbol{x}_i^\mathsf{T}\boldsymbol{x}_i - \mathrm{Tr}(\mathbf{P}\boldsymbol{x}_i\boldsymbol{x}_i^\mathsf{T})$, demonstrating the equivalence between the principal subspace objective in Eq. 10 and a weighted version of the AGI objective in Eq. 9.

4.2. Globally-Optimal Solution

More generally, this reasoning can be extended to show that the globally-optimal solution of the nonconvex problem in Eq. 9 is given by a singular value decomposition. Specifically, let $\mathbf{X}_i \in \mathcal{V}_{p_i,d}$ be an orthogonal matrix with columns spanning the p_i-dimensional subspace X_i. Then, following a similar argument from above:

$$\|\mathbf{X}_i - \mathbf{P}\mathbf{X}_i\|_F^2 = \min_{\mathbf{Q}_i \in \mathcal{P}_{k,d}} \frac{1}{2} \|\mathbf{P} - \mathbf{Q}_i\|_F^2 \text{ s.t. } \mathbf{Q}_i \in \mathcal{C}_i \tag{12}$$

In this case, both sides of Eq. 12 are equal to $p_i - \mathrm{Tr}(\mathbf{P}\mathbf{X}_i\mathbf{X}_i^\mathsf{T})$. This suggests that AGI is optimizing the average Euclidean reconstruction error of projecting each basis vector from each data subspace X_i onto the learned subspace B. Thus, as in Eq. 10, the globally-optimal solution can be found as the span of the top k left singular vectors of the matrix $[\mathbf{X}_1, \ldots, \mathbf{X}_n]$ where $\mathbf{X}_i$ is *any* orthogonal matrix with columns spanning X_i. Subspace-valued subspace learning can thus be solved using standard vector-valued subspace learning algorithms such as PCA by first transforming the input data into sets of basis vectors. However, the novel formulation of AGI can also naturally incorporate additional constraints, as described in Sec. 4.4.

4.3. Inference and Subspace Completion

The solution subspace can be applied towards a variety of applications through inference of latent variables that relate data with their locations on the subspace. Recall that this is accomplished in standard vector-valued subspace learning by fixing a basis matrix $\mathbf{B}$ for the learned subspace and then finding a lower-dimensional latent representation of the vector $\boldsymbol{x}_i$ as $\boldsymbol{w}_i = \mathbf{B}^\mathsf{T}\boldsymbol{x}_i$. However, because subspace-valued data are essentially infinite sets of vectors, inference in AGI must proceed in a way that preserves their inherent invariances. Furthermore, in order to map subspace data of varying dimension to the same lower-dimensional latent space, *subspace completion* must first be used to ensure consistent dimensionality and enable direct comparisons between them.

To do this, we first consider solving for the constrained local subspace representations $\mathbf{Q}_i$ with the learned $\mathbf{P}$ fixed. Note that this is simply the Euclidean projection of $\mathbf{P}$ onto the intersection of the sets $\mathcal{P}_{k,d}$ and $\mathcal{C}_i$:

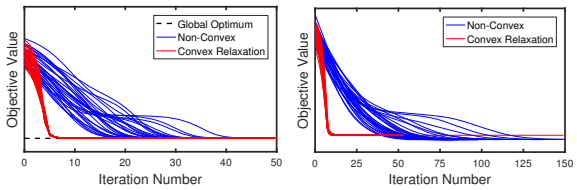
$$\arg\min_{\mathbf{Q}_i} \|\mathbf{P} - \mathbf{Q}_i\|_F^2 \text{ s.t. } \mathbf{Q}_i \in \mathcal{P}_{k,d} \cap \mathcal{C}_i \tag{13}$$

The solution can be decomposed as $\mathbf{Q}_i = \mathbf{X}_i\mathbf{X}_i^\mathsf{T} + \bar{\mathbf{X}}_i\bar{\mathbf{X}}_i^\mathsf{T}$, where $\bar{\mathbf{X}}_i \in \mathcal{V}_{k-p_i,d}$ spans the top $k - p_i$ eigenvectors of $(\mathbf{I} - \mathbf{X}_i\mathbf{X}_i^\mathsf{T})\mathbf{P}$. The result is a k-dimensional subspace Z_i that is close to B, but contains the data subspace X_i. Essentially, instead of projecting our data onto the learned subspace, we are projecting the learned subspace onto the set of subspaces that contain our data.

The same approach can be used for subspace completion to infer m-dimensional subspaces with $p_i \leq m < k$. However, in order to ensure consistency, we must introduce a relative ordering of the dimensions of the learned subspace B. Fortunately, this ordering is naturally given by the singular values associated with the globally-optimal solution described in the previous section. Thus, given an m-dimensional approximation $\hat{B} \in \mathcal{G}_{m,d}$ spanning the top m dimensions of B with the projection matrix $\hat{\mathbf{P}} = \hat{\mathbf{B}}\hat{\mathbf{B}}^\mathsf{T}$ for $\hat{\mathbf{B}} \in \mathcal{V}_{m,d}$, we can find an m-dimensional completed input subspace $\hat{X}_i \in \mathcal{G}_{m,d}$ as the span of the columns of the matrix $\hat{\mathbf{X}}_i = [\mathbf{X}_i, \bar{\mathbf{X}}_i']$ where $\bar{\mathbf{X}}_i' \in \mathcal{V}_{m-p_i,d}$ contains the top $m - p_i$ eigenvectors of $(\mathbf{I} - \mathbf{X}_i\mathbf{X}_i^\mathsf{T})\hat{\mathbf{P}}$ as its columns.

Finally, since all completed subspaces $\hat{X}_i$ are of the same dimension m, consistent lower dimensional representations can now be inferred. As with standard subspace learning, we find an approximation $\hat{X}_i$ that is contained within the learned subspace B. However, instead of minimizing the Euclidean reconstruction error, we minimize the distances between projection matrices, as shown in Eq. 14 below.

$$\arg\min_{\mathbf{W}_i \in \mathcal{V}_{m,k}} \left\| \hat{\mathbf{X}}_i\hat{\mathbf{X}}_i^\mathsf{T} - \mathbf{B}\mathbf{W}_i\mathbf{W}_i^\mathsf{T}\mathbf{B}^\mathsf{T} \right\|_F^2 \tag{14}$$



(a) AGI Convergence (b) Robust AGI Convergence

Figure 5: A visualization of the convergence properties of our optimization algorithm applied with 30 random initializations for unconstrained AGI (a) and robust AGI (b). Without additional constraints, the global optimum from Sec. 4.2 is achieved with both our original formulation and its convex relaxation, which also appears to speed convergence substantially.

The columns of the solution $\mathbf{W}_i$ span the column space of $\mathbf{B}^\mathsf{T}\hat{\mathbf{X}}_i$ and can be found as its left singular vectors. Note that this again reduces to standard inference in vector-valued subspace learning for the special case of weighted one-dimensional subspaces with $m = 1$.

Because any vector within the subspace $\hat{X}_i$ is a linear combination of the columns of $\hat{\mathbf{X}}_i$, the lower-dimensional representation found by projecting it onto $\mathbf{B}$ is also a linear combination of the columns of $\mathbf{W}_i$. Thus, the lower-dimensional latent representation for the subspace X_i is itself a subspace $W_i \in \mathcal{G}_{m,k}$ that can be uniquely represented in terms of its projection matrix $\mathbf{M}_i = \mathbf{W}_i\mathbf{W}_i^\mathsf{T}$, which we treat as the coordinates of the associated latent space.

To enable the transfer of information between the completed subspaces $\hat{X}_i$ as in the image relighting example in Fig. 3, consistency between the choice of their basis matrices must be enforced. Specifically, we choose $\hat{\mathbf{Z}}_i \in \mathcal{V}_{m,d}$ to be the basis matrix for $\hat{X}_i$ that is closest to the fixed m-dimensional basis matrix $\hat{\mathbf{B}}$ in terms of Euclidean distance. Since we know that $\hat{\mathbf{Z}}_i$ must span the columns of $\hat{\mathbf{X}}_i$, this is equivalent to finding the orthogonal linear transformation $\mathbf{R}_i \in \mathcal{V}_{m,m}$ that brings it closest to $\hat{\mathbf{B}}$, as shown in Eq. 15.

$$\underset{\mathbf{R}_i \in \mathcal{V}_{m,m}}{\arg\min} \left\| \hat{\mathbf{B}} - \hat{\mathbf{X}}_i\mathbf{R}_i \right\|_F^2 = \underset{\mathbf{R}_i \in \mathcal{V}_{m,m}}{\arg\max} \operatorname{Tr}(\hat{\mathbf{B}}^\mathsf{T}\hat{\mathbf{X}}_i\mathbf{R}_i) \quad (15)$$

The solution is $\mathbf{R}_i = \mathbf{U}_i\mathbf{V}_i^\mathsf{T}$ from the singular value decomposition $\hat{\mathbf{B}}^\mathsf{T}\hat{\mathbf{X}}_i = \mathbf{U}_i\mathbf{S}_i\mathbf{V}_i^\mathsf{T}$ so that $\hat{\mathbf{Z}}_i = \hat{\mathbf{X}}_i\mathbf{R}_i$. Then, the coefficients found by projecting a data vector lying in another input subspace onto the global basis $\hat{\mathbf{B}}$ can be used directly with the completed local basis $\hat{\mathbf{Z}}$ to generate novel samples extrapolated from X_i.

4.4. Optimization with Additional Constraints

In this section, we demonstrate how prior knowledge can be introduced to improve the quality of the learned subspace through the example of robust AGI, a constrained variation of our method. We also present a simple, unified optimization algorithm that supports this and a wide range of additional constraints that can be expressed as the intersection of sets equipped with efficient projection operators.

For robust AGI, the local affine constraints $\mathcal{C}_i$ are replaced with robust alternatives $\mathcal{C}_i^r$ that allow for sparse errors $\mathbf{E}_i$ in a manor similar to the convex formulation for robust PCA [8]. The local subspace Z_i is then constrained to contain $\mathbf{X}_i + \mathbf{E}_i$. We assume that these errors are present only within some subset of the data features, so we restrict that the $\ell_{2,1}$-norm of $\mathbf{E}_i^\mathsf{T}$ be less than some maximum threshold ε, which constrains the sum of the Euclidean norms of the rows of $\mathbf{E}$. The robust affine constraint sets $\mathcal{C}_i^r$ can be written as:

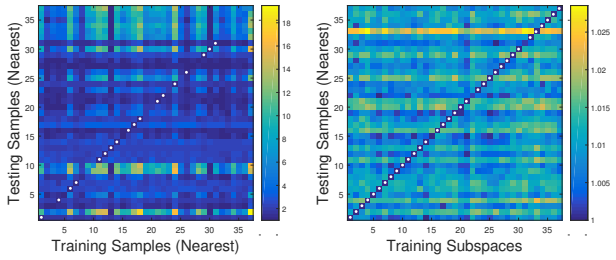
$$\mathcal{C}_i^r = \left\{ \mathbf{Q} : \mathbf{Q}(\mathbf{X}_i + \mathbf{E}_i) = \mathbf{X}_i + \mathbf{E}_i, \left\|\mathbf{E}_i^\mathsf{T}\right\|_{2,1} \leq \varepsilon \right\} \quad (16)$$

While these constraints are convex, inferring $\mathbf{Q}_i$ cannot be accomplished in closed form as with Eq. 13. However, the projection onto this this set *can* be found in closed form with a soft-thresholding operator. This suggests the use of an iterative projection algorithm that finds the projection onto an intersection of sets via sequences of projections onto the individual sets, which can often be computed much more efficiently. Specifically, in this work, we employ the Douglas-Rachford algorithm [6].

To learn the global subspace projection matrix $\mathbf{P}$, we employ an alternating optimization strategy. After random initialization, we fix $\mathbf{P}$ and then solve for it's projection $\mathbf{Q}_i$ onto each of the constraint sets $\mathcal{C}_i^r$. Then, with $\mathbf{Q}_i$ fixed, we solve for $\mathbf{P}$, repeating this process until convergence.

The convex relaxation of our problem replaces the set of projection matrices $\mathcal{P}_{k,d}$ with its convex hull, the Fantope $\mathcal{F}_{k,d}$. From convexity, the resulting solution for $\mathbf{P}$ is the sample average $\mathbf{P} = \frac{1}{n}\sum_i \mathbf{Q}_i$. Thus, this optimization procedure is exactly the averaged projections algorithm for finding the intersection of convex sets, so it is guaranteed to converge to the globally-optimal solution [10]. Then, after training, the learned subspace can be found from the Fantope representation as the span of its top k left singular vectors. This strategy has been effectively used in other convex relaxations of subspace learning [38] and extrinsic methods for learning on Grassmann manifolds [21].

On the other hand, for the original nonconvex formulation in which each $\mathbf{Q}_i$ is constrained to be a projection matrix, the average will generally not be a projection matrix. Instead, the solution $\mathbf{P}$ is its projection onto $\mathcal{P}_{k,d}$. In this case, convergence to the globally-optimal solution is not theoretically guaranteed due to the nonconvexity of the set of projection matrices $\mathcal{P}_{k,d}$. However, similar methods been applied successfully with certain other nonconvex sets [1]. Furthermore, our empirical results were promising, demonstrating consistent convergence (as in Fig. 5) robust to initialization and encouraging further theoretical investigations of this behavior.



(a) PCA Distances (b) AGI Distances

Figure 6: Relative distances between the low-dimensional representations of training and testing data for PCA (a) and AGI (b). For PCA, each pixel shows the distance between the nearest training and testing image of the corresponding classes. For AGI, the smallest distance is shown between the one-dimensional subspace of each testing image and the p_i-dimensional subspace of *all* training images of the corresponding classes. For a particular row, a white circle on the diagonal indicates that the minimum distance was smallest between training elements from the same class.

4.5. Implementation Details

While the optimization strategy described in the previous section relies on the Euclidean representation of subspaces as matrices in $\mathbb{R}^{d\times d}$, this would be very inefficient in high-dimensional settings like those common within computer vision. However, both projection matrices and Fantope elements are low-rank and so we represent them numerically as low-rank matrix factorizations $\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{U}^\mathsf{T}$ where $\mathbf{U} \in \mathcal{V}_{r,d}$ and $\mathbf{D}$ is a diagonal $r \times r$ matrix. For projection matrices, $r = k$ and $\mathbf{D} = \mathbf{I}$. On the other hand, members of the Fantope $\mathcal{F}_{k,d}$ have a fixed trace equal to k, but can have rank $r > k$ with arbitrary non-negative elements in $\mathbf{D}$. Thus, we enforce them to have a maximum rank of $r_{max} = 5k$, which we found to be more than sufficient in our experiments.

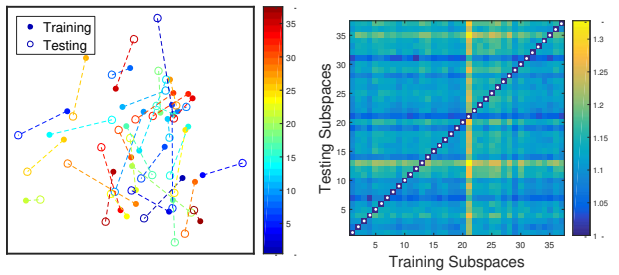
Projection on many constraint sets, including those considered in Sec. 4.4, can preserve the low-rank factorization as shown in Fig. 7. For example, since projection onto the affine subspace defined by the constraint set $\mathcal{C}_i$ in Eq. 8 will not result in a symmetric matrix, we instead employ projection onto its intersection with the set of $d \times d$ symmetric matrices $\mathcal{S}_d$, which can be efficiently computed in closed-form while preserving rank and symmetry. Also, projection of the average of the local representations $\mathbf{Q}_i$ onto the set of projection matrices is computed with an incremental SVD algorithm [2] so that the full matrices never need to be evaluated or stored during optimization.

5. Experimental Results

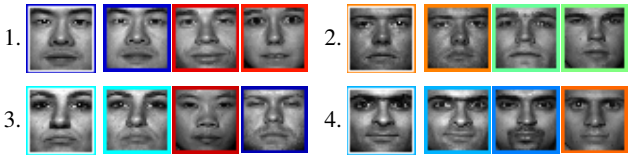
To demonstrate the effectiveness of AGI, we evaluate a variety of applications on both real and synthetic datasets. First, continuing the example shown in Figures 1-3, we demonstrate further results on the on the Extended Yale

$$
\begin{aligned}
P_{\mathcal{P}_{k,d}}\mathbf{M} &= \mathbf{U}^{(k)}\mathbf{D}^{(k)}\mathbf{U}^{(k)\mathsf{T}},\ P_{\mathcal{F}_{k,d}}\mathbf{M} = \mathbf{U}(P_{B_k^1}\mathbf{D})\mathbf{U}^\mathsf{T} \\
P_{C_i\cap\mathcal{S}_d}\mathbf{M} &= \mathbf{X}_i\mathbf{X}_i^\mathsf{T} + (\mathbf{I}-\mathbf{X}_i\mathbf{X}_i^\mathsf{T})\mathbf{U}\mathbf{D}\mathbf{U}^\mathsf{T}(\mathbf{I}-\mathbf{X}_i\mathbf{X}_i^\mathsf{T}) \\
P_{C_i^r\cap\mathcal{S}_d}\mathbf{M} &= \tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i^+ + (\mathbf{I}-\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i^+)\mathbf{U}\mathbf{D}\mathbf{U}^\mathsf{T}(\mathbf{I}-\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i^+), \\
\tilde{\mathbf{X}}_i &= \mathbf{X}_i + P_{B_\epsilon^1}(\mathbf{M}-\mathbf{I})^+(\mathbf{X}_i - \mathbf{M}\mathbf{X}_i)
\end{aligned}
$$

Figure 7: Projections onto some of the constraint sets discussed in the text. Note that $\cdot^+$ denotes the Moore-Penrose pseudoinverse, which admits a low-rank update, and B_k^1 is the ℓ_1 ball of radius k, whose projection can also be computed efficiently [13].



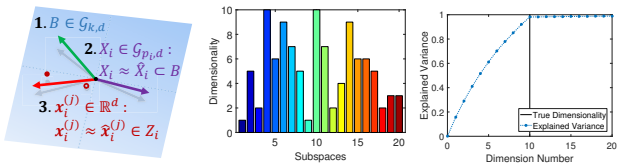
(a) AGI Latent Space (Subspace) (b) AGI Distances (Subspace)



(c) Example Nearest-Neighbor Subspaces

Figure 8: A demonstration of AGI applied to the classification of illumination subspaces constructed from images under a variety of lighting conditions. The latent space embedding (a) shows the representations of the training and testing subspaces belonging to the same class connected by dotted lines while the relative distance matrix (b) shows correct nearest-neighbor classification of all testing subspaces. Visualization of three nearest neighbors (c) shows that the learned representations are able to encode visual similarities invariant to lighting.

Face Database B [29], which contains images of 38 individuals under approximately 64 different lighting conditions. The training images consisted of all images lit from the left with a single subject left out, resulting in substantial bias between the training and testing sets. We represented each subject as a a subspace retaining 90% of the variance and learned a global subspace of dimensionality $k = 200$ with subspace completion of $m = 40$ dimensions used for inference. As discussed previously, AGI representations are more discriminative, even when the testing data are represented individually as one-dimensional subspaces. This is shown quantitatively in Fig. 6, where the nearest distances between training and testing data of the same class is much lower for AGI. Furthermore, if prior knowledge is known about which testing images are of the same individual, groups of testing data for a particular class can be summarized with a single testing subspace that better encodes



(a) Data Overview (b) Dimensionality (c) Eigenspectrum

Figure 9: A summary of the synthetic dataset used in our experiments. A pictorial overview is shown (a) demonstrating the data-generating process. The dimensionalities p_i of the class subspaces are also shown (b) alongside the eigenspectrum of the data covariance matrix (c), which demonstrates that the vast majority of the data variability of the 20 classes is contained within a 10-dimensional subspace, resulting in significant class overlap.

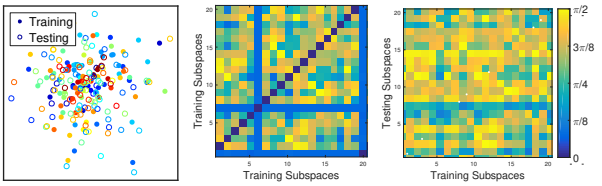
Table 1: Synthetic Data Classification Accuracies

PCA	Subspace Angle	AGI (Vector)	AGI (Subspace)
14.6%	15.5%	96.1%	100%

the data invariance. This shown in Fig. 8, in which the resulting class representations appear to encode more complex appearance variations that are invariant to lighting.

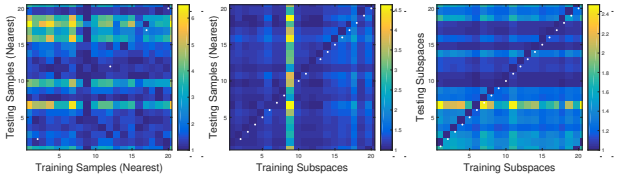
We also construct a difficult synthetic dataset for the task of set-valued classification where groups of data vectors known to belong to the same class are evaluated jointly, an overview of the of which is shown in Fig. 9. The data generating process can be summarized as follows: first, a shared k-dimensional subspace B of $\mathbb{R}^d$ is chosen with $k = 10$ and $d = 1000$. Then, for each of $n = 20$ classes, a lower-dimensional training subspace X_i is chosen with random dimensionality $p_i \leq k$ so that it is close to B along with a noisy version used for testing. Finally, for each class, a total of p_i noisy data vectors $\boldsymbol{x}_i^{(j)}$ are drawn near the subspaces for both training and testing sets, the minimum number required to define the associated subspace, for a total of only 103 samples in each. The difficulty, which is visualized in Fig. 10, arises from the small number of biased training examples–some classes have only a single data point–along with the relatively large number of nearly-overlapping classes. Thus, the discriminative information between classes is essentially contained within the slight discrepancies between their local subspaces X_i and the global subspace B, which account for less than 2% of the total variance in the dataset, as shown in Fig. 9c.

We consider each testing group to be points on a subspace and use AGI to find invariant lower-dimensional representations with $k = 10$ and $m = 20$. Fig. 11 shows the distance matrices between the low-dimensional representations while Tab. 1 shows the resulting nearest-neighbor prediction accuracies. Also shown is the accuracy from selecting the label of the training subspace that is closest to each testing subspace in terms of angle, which still performs poorly due to the substantial noise in the dataset as shown



(a) Embedding (b) Training Angles (c) Testing Angles

Figure 10: A visualization demonstrating the difficulty of our synthetic dataset. A two-dimensional embedding of the data (a) shows no recognizable structure. This results from the close proximity between the training classes (a), where each element of the matrix corresponds to the angle between the corresponding subspaces. The angles between the training and testing subspaces (b) also demonstrate the large disparity between training and testing data due to the added noise.



(a) PCA (b) AGI (Vector) (c) AGI (Subspace)

Figure 11: Relative distances between the low-dimensional representations of training and testing data for PCA (a), AGI with testing vectors represented as one-dimensional subspaces (b), and AGI with testing vectors of the same class grouped together as p_i-dimensional subspaces.

in Fig. 10c. On the other hand, even when each testing vector is treated as its own subspace, AGI gives perfect accuracy when all testing data of the same class are modeled as higher-dimensional subspaces. This indicates that AGI is also able to leverage the denoising capabilities of subspace learning to improve discriminability.

6. Conclusion

The framework of Approximate Grassmannian Intersections is a novel interpretation of subspace learning as approximate constraint satisfaction on the Grassmann manifold, generalizing standard vector-valued techniques such as PCA to naturally support subspace-valued data. Despite the nonconvexity of our formulation, the globally-optimal solution can be found efficiently and described a simple optimization framework that supports a variety of additional constraints for incorporating prior knowledge. More generally, our approach explicitly leverages the known geometric structure of data to learn representations invariant to certain transformations for improved generalization, especially in cases with extremely limited training data.

Acknowledgments: This research was supported in part by the National Science Foundation under grants RI-1617953 and IIS-1418523.

References

[1] F. J. A. Artacho, J. M. Borwein, and M. K. Tam. Global behavior of the douglas–rachford method for a nonconvex feasibility problem. *Journal of Global Optimization*, 65(2):309–327, 2016. 6

[2] C. G. Baker, K. A. Gallivan, and P. Van Dooren. Low-rank incremental methods for computing dominant singular subspaces. *Linear Algebra and its Applications*, 436(8):2866–2888, 2012. 7

[3] L. Balzano, R. Nowak, and B. Recht. Online identification and tracking of subspaces from highly incomplete information. In *Allerton Conference on Communication, Control, and Computing*, 2010. 3

[4] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003. 1

[5] H. H. Bauschke and J. M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM review*, 38(3):367–426, 1996. 2, 3

[6] H. H. Bauschke, P. L. Combettes, and D. R. Luke. Finding best approximation pairs relative to two closed convex sets in Hilbert spaces. *Journal of Approximation Theory*, 127(2):178–192, 2004. 3, 6

[7] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173, 2007. 3

[8] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011. 3, 6

[9] P. Combettes. The convex feasibility problem in image recovery. *Advances in imaging and electron physics*, 95:155–270, 1996. 3

[10] P. L. Combettes. Inconsistent signal feasibility problems: Least-squares solutions in a product space. *IEEE Transactions on Signal Processing*, 42(11):2955–2966, 1994. 6

[11] J. P. Cunningham and Z. Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research (JMLR)*, 2015. 1, 3

[12] F. De la Torre. A least-squares framework for component analysis. *Pattern Analysis and Machine Intelligence*, 34(6):1041–1055, 2012. 3

[13] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *International Conference on Machine Learning (ICML)*, 2008. 7

[14] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. 3

[15] Y. Ephraim and H. L. Van Trees. A signal subspace approach for speech enhancement. *IEEE Transactions on speech and audio processing*, 3(4):251–266, 1995. 1

[16] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003. 1

[17] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging*, 23(8):995–1005, 2004. 3

[18] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001. 1

[19] J. Hamm and D. D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *International Conference on Machine Learning (ICML)*, 2008. 2, 3

[20] M. Harandi, R. Hartley, C. Shen, B. Lovell, and C. Sanderson. Extrinsic methods for coding and dictionary learning on grassmann manifolds. *International Journal of Computer Vision (IJCV)*, 114(2-3):113–136, 2015. 2, 3

[21] M. Harandi, C. Sanderson, C. Shen, and B. C. Lovell. Dictionary learning and sparse coding on Grassmann manifolds: An extrinsic solution. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 6

[22] J. He, L. Balzano, and A. Szlam. Incremental gradient on the Grassmannian for online foreground and background separation in subsampled video. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 3

[23] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using laplacianfaces. *Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005. 1

[24] Y. Hong, R. Kwitt, N. Singh, N. Vasconcelos, and M. Niethammer. Parametric regression on the grassmannian. *Pattern Analysis and Machine Intelligence*, 38(11):2284–2297, 2016. 2, 3

[25] W. Huang, F. Sun, L. Cao, D. Zhao, H. Liu, and M. Harandi. Sparse coding and dictionary learning with linear dynamical systems. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[26] I. Jolliffe. *Principal component analysis*. Springer, 2002. 3

[27] N. Kambhatla and T. K. Leen. Dimension reduction by local principal component analysis. *Neural computation*, 9(7):1493–1516, 1997. 3

[28] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T. S. Lee, and T. J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural Computation*, 15(2):349–396, 2003. 3

[29] K.-C. Lee, J. Ho, and D. J. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *Pattern Analysis and Machine Intelligence*, 27(5):684–698, 2005. 7

[30] A. S. Lewis and J. Malick. Alternating projections on manifolds. *Mathematics of Operations Research*, 33(1):216–234, 2008. 3

[31] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen. Combining multiple kernel methods on Riemannian manifold for emotion recognition in the wild. In *International Conference on Multimodal Interaction*, 2014. 3

[32] Y. M. Lui. Advances in matrix manifolds for computer vision. *Image and Vision Computing*, 30(6):380–388, 2012. 3

[33] K. Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. 3

[34] C. R. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203, 1948. 3

[35] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. 1

[36] P. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa. Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition. *Pattern Analysis and Machine Intelligence*, 33(11):2273–2286, 2011. 2, 3

[37] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (GPCA). In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003. 3

[38] V. Q. Vu, J. Cho, J. Lei, and K. Rohe. Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In *Advances in Neural Information Processing Systems (NIPS)*, 2013. 4, 6

[39] D. Zhang and L. Balzano. Global convergence of a Grassmannian gradient descent algorithm for subspace estimation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016. 3

[40] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006. 3