The Flag Manifold as a Tool for Analyzing and Comparing Sets of Data Sets

Xiaofeng Ma
Department of Mathematics
Colorado State University
Fort Collins, CO, USA 80523
markma@rams.colostate.edu

Michael Kirby
Department of Mathematics
Colorado State University
Fort Collins, CO, USA 80523
michael.kirby@colostate.edu

Chris Peterson
Department of Mathematics
Colorado State University
Fort Collins, CO, USA 80523
peterson@math.colostate.edu

Abstract

The shape and orientation of data clouds reflect variability in observations that can confound pattern recognition systems. Subspace methods, utilizing Grassmann manifolds, have been a great aid in dealing with such variability. However, this usefulness begins to falter when the data cloud contains sufficiently many outliers corresponding to stray elements from another class or when the number of data points is larger than the number of features. We illustrate how nested subspace methods, utilizing flag manifolds, can help to deal with such additional confounding factors. Flag manifolds, which are parameter spaces for nested sequences of subspaces, are a natural geometric generalization of Grassmann manifolds. We utilize and extend known algorithms for determining the minimal length geodesic, the initial direction generating the minimal length geodesic, and the distance between any pair of points on a flag manifold. The approach is illustrated in the context of (hyper) spectral imagery showing the impact of ambient dimension, sample dimension, and flag structure.

1. Introduction

Variability in data observations due, for example, to image lighting, data noise, or batch effects, is typically viewed as a challenge to pattern recognition. In this paper we propose to use the framework of flag manifolds to recast the variation as useful additional structure for classification, differentiation between similar objects, and anomaly detection. The geometry underlying a variation of state can be exploited by observing (via a sensor or collection of sensors) the range of corresponding measurements in sensor state space. In essence, an object provides a map from a variation of state space to a sensor state space. In practice, one is only able to observe a noisy sampling of the map. An underlying assumption is that characteristics of the map provide a signature for the object being sensed. This motivates the robust modeling of a set of data, i.e., modeling specifically to capture the variability of different realizations of a data class. Practically, one can often exploit this variability by considering a collection of observations abstractly as a single point in an appropriate parameter space and algorithmically exploiting the geometry of the parameter space.

Ideas from geometry and topology have shown considerable promise for the analysis of large, and or complex, data sets given their ability to encode this variability. For example, the mathematical framework of the Grassmannian has proven to be effective at capturing many of the pattern variations that so often confound pattern recognition systems. In this setting data is encoded as subspaces and distances are measured using angles between subspaces. The Grassmann manifold is a suitable tool for analyzing many data sets but it requires the dimension of the subspace used to represent the data to be less than half of the ambient dimension.

Initially explored in the setting of subspace packing problems [28, 5, 15], the application of Stiefel and Grassmann manifolds has become widespread in computer vision and pattern recognition. Examples include: video processing [11], classification [10, 4, 31, 32], action recognition [2], expression analysis [29, 30, 16], domain adaptation [14, 26], regression [27, 12], pattern recognition [17], and computation of subspace means [3, 20]. More recently, Grassmannians have also been explored in the deep neural network literature [13]. Much of this progress has hinged on the development of efficient algorithms [7, 9, 1] allowing procedures developed in other settings to be transported to analogous procedures on Grassmann manifolds. A collection of papers by Nishimori et al introduced flag manifolds in the context of independent component analysis and optimization [24, 23, 22, 25]. Later work by others used and extended some of these ideas in a variety of contexts [8, 6, 20, 21, 19]. Very recent work of Ye, Wong, and Lim gives an expanded view of the local differential geometry of flag manifolds with a very practical viewpoint [33]. Two features of [19] are an iterative algorithm for determining the distances between given points $[A], [B]$ on a flag manifold (where A and B are arbitrary orthogonal matrix representatives for $[A]$ and $[B]$) and algorithms for determining how to move from $[A]$ to $[B]$ along a *minimal* length geodesic. These algorithms allow practical computations to be made on flag manifolds. In this paper we utilize such algorithms and illustrate their effectiveness in several sample problems in data analysis. In addition, we develop a modification that allows for more efficient computations to be made in the setting where the final dimension jump in the flag manifold signature is greater than one half of the ambient dimension.

From the data analysis perspective, points on a Grassmann manifold $Gr(k,n)$ parameterize k-dimensional linear subspaces $V \subset \mathbb{R}^n$. Points on a flag manifold $FL(a_1, a_2, \dots, a_d)$ parameterize sequences of nested linear subspaces $V_1 \subset V_2 \subset \cdots \subset V_d = \mathbb{R}^n$ with $a_i = dim(V_i)$ (sometimes the notation $FL(n_1, n_2, \dots, n_d)$ is used where $n_1 = a_1$ and $n_i = dim(V_i) - dim(V_{i-1})$ for $i > 1$). Flag manifolds can be viewed as generalizations or refinements of Grassmannians and have the ability to encode more subtle relationships than are capable with Grassmannians. In practice, the Grassmannian seems to be well suited for data sets where the ambient dimension is much larger than the number of data points (tall matrices) and where the data set is relatively pure. While applicable in this setting, the flag manifold approach is also suitable to the analysis of some data sets where the data dimension may be small relative to the number of observations (wide matrices), where the data set may consist of a mixture of classes, and where the data has been collected under multiple variations of state.

As described above, flag manifolds constitute a refinement of Grassmann manifolds that enable the measurement of the distance between nested spaces. One setting that illustrates their advantage is the problem of comparing mixed data sets. An example of what is meant by this is the following: suppose that one data set has 80 percent of its samples drawn from class A and 20 percent from class B and a second data set has the reverse mixture. The different concentrations lead to different basis vectors in the singular value decomposition. Grassmann methods typically consider the span of the first few basis vectors and utilize the resulting subspace as a representative of the data set. Flag methods refine this approach by utilizing a nested sequence of vector spaces as a representative. The extra structure, i.e. information, in the nested sequence of subspaces increases the ability to distinguish between these data sets as compared to using the information in a single subspace.

Mathematically, as is demonstrated in this paper, the tools for measuring geodesic distances between data represented by tall versus wide matrices are utilized in a different manner. Here we utilize the algorithms of [19] for computing distances between wide matrices and show their use for solving pattern recognition and computer vision problems. The work is in the same spirit as Grassmannian data processing but extends these tools to a distinct yet important application. We argue that in many cases where data is subject to wide variability, the distances measured between large sets of small feature spaces captures more fidelity than algorithms on Euclidean space.

The outline of this paper is as follows: In Section 2 we review the geometric framework of the Grassmannian. In Section 3 the theory of the flag manifold is developed along with efficient algorithms to compute geodesic distances. In Section 4 we illustrate the applicability of the method on hyperspectral imagery. In Section 5 we summarize the features of the methodology.

2. The Grassmannian

The Grassmannian, denoted by $Gr(k,n)$, is a geometric object whose points parameterize the k-dimensional subspaces of a fixed n-dimensional vector space. In the context of applications, the fixed n-dimensional vector space is typically taken to be $\mathbb{R}^n$ or $\mathbb{C}^n$ (though vector spaces over other fields can also be considered). For the purposes of this paper, the ambient vector space is taken to be $\mathbb{R}^n$ and we represent $Gr(k,n)$ as a *real matrix manifold*. Each point in $Gr(k,n)$ is identified with an equivalence class of orthogonal matrices leading to the representation of $Gr(k,n)$ as $O(n)/O(k) \times O(n-k)$ or alternatively in terms of special orthogonal matrices as $SO(n)/S(O(k) \times O(n-k))$. In these formulas, $O(n)$ denotes the group of $n \times n$ orthogonal matrices and $O(k) \times O(n-k)$ denotes the subgroup of $O(n)$ consisting of block diagonal matrices with elements from $O(k)$ in the first block and elements from $O(n-k)$ in the second block. The notation $SO(n)$ (resp. $S(O(k) \times O(n-k))$) denotes the subgroup of $O(n)$ (resp. $O(k) \times O(n-k)$) with determinant 1. Thus a point on $Gr(k,n)$ can be identified with an equivalence classes of n-by-n special orthogonal matrices $[Y] \subset SO(n)$ where two elements $Y, Y' \in SO(n)$ are in the same equivalence class, written $Y \sim Y'$, if there exists an M such that $Y' = YM$ where

$$M = \begin{bmatrix} M_k & 0 \\ 0 & M_{n-k} \end{bmatrix} \tag{1}$$

with $M_k \in O(k)$, $M_{n-k} \in O(n-k)$, and $\det(M_k) \cdot \det(M_{n-k}) = 1$. If $Y \sim Y'$ then $[Y] = [Y']$ denote the same point on the Grassmann manifold $Gr(k,n)$. One advantage of this characterization is that we can utilize the well-studied geometry of $SO(n)$ to understand the geometry of $Gr(k,n)$. It is well known that a geodesic path on $SO(n)$, starting at a point $Q \in SO(n)$, is given by a one parameter exponential flow : $t \mapsto Q\exp(tH)$ where H is an n-by-n skew-symmetric matrix. Since $Gr(k,n)$ is a quotient manifold of $SO(n)$ by the subgroup $S(O(k) \times O(n-k))$, it can be readily verified that when representing geodesics on $Gr(k, n)$, one can further restrict H to be a skew symmetric matrix of the form

$$H = \begin{bmatrix} 0_k & -B^T \\ B & 0_{n-k} \end{bmatrix}, B \in \mathbb{R}^{(n-k)\times k} \quad (2)$$

where the size and location of the zero-blocks mirror the size and location of M_k, M_{n-k} in the block diagonal matrix M. A geodesic on $Gr(k, n)$, starting at the point $[Q] \in Gr(k, n)$, can thus be expressed in parameterized form as:

$$Q(t) = Q \exp(t \begin{bmatrix} 0 & -B^T \\ B & 0 \end{bmatrix}). \quad (3)$$

The sub-matrix B specifies the direction and the speed of the geodesic path. More details can be found in [7]. As will be seen later in Section 3.2, an advantage of the characterization of the Grassmannian as an equivalence class of special orthogonal matrices is that this approach allows a straightforward generalization for defining and representing points and geodesics on a *flag manifold* thanks to the underlying Lie theory.

Computations of distances between points on the Grassmannian $Gr(k, n)$ are often performed using an n-by-k orthonormal matrix representative (whose column space corresponds to the point on $Gr(k, n)$). In this setting, a point on $Gr(k, n)$ can be represented as an equivalence class of n-by-k orthonormal matrices where $X \sim X'$ iff $X' = XU$ where $U \in O(k)$. The distance between two points on $Gr(k, n)$ (i.e. two k-dimensional subspaces of $\mathbb{R}^n$) $[X]$ and $[Y]$ can be computed via the compact SVD of X^TY, i.e., $U\Sigma V^T := X^TY$. From the SVD, the geodesic distance between $[X]$ and $[Y]$ is defined as:

$$d_g([X], [Y]) = \sqrt{\lambda_1^2 + \lambda_2^2 + \cdots + \lambda_k^2} \quad (4)$$

where $\lambda_j = \arccos(\sigma_j)$ with σ_j denoting the j^{th} diagonal element of Σ. In the formula $(XU)^TYV = \Sigma$, the columns of XU and YV are the principal vectors between $[X]$ and $[Y]$. The geodesic between $[X]$ and $[Y]$ rotates the columns of XU to the columns of YV while the diagonal elements of Σ encode the cosine of the angles between these corresponding columns.

3. The Flag Manifold

The distinction between geodesics on Grassmannians and flags is captured pictorially in Figure 1. For Grassmannians, one is moving a subspace into another subspace along the shortest trajectory. In the flag setting, this trajectory has to remain faithful to the nesting structure of the subspaces. In Figure 1 (right) we see the required flag alignment of the coordinate directions in the 2D subspace whereas no alignment is required for the Grassmannian (left). The details and ramifications of this difference are elucidated below.

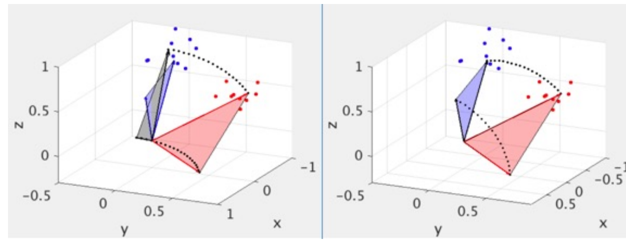

Figure 1: A comparison of geodesics on the Grassmannian (left) and flag (right) manifolds for representing the distance between two data sets. The red subspace is being moved to the blue subspace via the gray subspace in each case.

3.1. Flags and their appearance in data analysis

A *flag* of subspaces in $\mathbb{R}^n$ is a nested sequence of subspaces $\mathbf{V_1} \subset \mathbf{V_2} \subset \cdots \subset \mathbf{V_d} = \mathbb{R}^n$. The signature or type of the flag is the sequence $(\dim \mathbf{V}_1, \dim \mathbf{V}_2, \ldots, \dim \mathbf{V}_d)$. This dimension information can also be encoded as the sequence $(\dim \mathbf{V}_1, \dim \mathbf{V}_2 - \dim \mathbf{V}_1, \dim \mathbf{V}_3 - \dim \mathbf{V}_2, \ldots, \dim \mathbf{V}_d - \dim \mathbf{V}_{d-1})$. In this paper, we will use this second type of encoding for the signature of a flag, thus we will identify the type of a flag in $\mathbb{R}^n$ by the sequence of positive integers $(n_1, n_2, \ldots, n_d)$ where $\dim V_j = \Sigma_{i=1}^{j} n_i$ and $n_1 + n_2 + \cdots + n_d = n$. We let $FL(n_1, n_2, \ldots, n_d)$ denote the *flag manifold* whose points parameterize all flags of type $(n_1, n_2, \ldots, n_d)$. As a special case, a flag of type $(k, n - k)$ is simply a $k-$dimensional subspace of $\mathbb{R}^n$ (which can be considered as a point on the Grassmann manifold $Gr(k, n)$). Hence $FL(k, n - k) = Gr(k, n)$. The idea that the flag manifold is a generalization of the Grassmann manifold will be utilized in Section 3.2 to introduce the geodesic formula on the flag manifold (see [33] for a nice expanded development of the geodesic formula). The nested structure inherent in a flag appears naturally in the context of data analysis.

In [18] a pictorial illustration of the geodesic between discrete Daubechies2 and Daubechies4 wavelets was presented by observing the action, of points along the geodesic between the wavelets on $FL(4, 4, 8, 16)$, on a photograph. In the context of data analysis, one can attach points on a flag manifold to a data set as follows. Let $X \in \mathbb{R}^{n\times p}$ be a data matrix of p samples in $\mathbb{R}^n$. The left singular vectors U obtained from the compact SVD, $X = U\Sigma V^T$, determine an ordered basis for the column span of X. The ordering is based on the magnitude of the corresponding singular values and provides a straightforward way to associate a flag to U. For example, to $U = [u_1|u_2| \ldots |u_k]$, construct the nested sequence of subspaces $\text{span}([u_1]) \subset \text{span}([u_1|u_2]) \subset \cdots \subset \text{span}([u_1| \cdots |u_k]) \subset \mathbb{R}^n$. This is a flag of type $(1, 1, \ldots, 1, n - k)$ in $\mathbb{R}^n$ and corresponds to a point $[U]$ on $FL(1, 1, \ldots, 1, n - k)$. In order to produce a point on $FL(3, 4, n - 7)$ one could consider span($[u_1|\cdots|u_3]$) $\subset$ span($[u_1|\cdots|u_7]$) $\subset \mathbb{R}^n$. As will be discussed in Section 4, using an SVD basis of a data set to produce a flag with a given signature can provide additional information when comparing data sets.

3.2. Representation of the flag manifold

The flag manifold $FL(n_1, n_2, \ldots, n_d)$ parametrizes all flags of type $(n_1, n_2, \ldots, n_d)$. The presentation in [7] gives a representation of the Grassmann manifold $Gr(k, n)$ as the quotient manifold $O(n)/O(k) \times O(n-k)$. Similarly, we can view a flag manifold as a quotient manifold constructed from $O(n)$. In particular, $FL(n_1, n_2, \cdots, n_d) \cong O(n)/O(n_1) \times O(n_2) \times \cdots \times O(n_d)$ where $n_1 + n_2 + \cdots + n_d = n$. In this definition, $O(n_1) \times O(n_2) \times \cdots \times O(n_d)$ denotes the subgroup of $O(n)$ consisting of block diagonal matrices with elements from $O(n_k)$ in the k^{th} block. Although it is common to represent a flag manifold as a quotient manifold of $O(n)$, it is more convenient to represent a flag manifold as a quotient manifold of $SO(n)$ for the purposes of computations involving the exp map (since $\exp(H) \in SO(n)$ for any skew-symmetric matrix H). Hence for the computations in this paper, we make the representation $FL(n_1, n_2, \cdots, n_d) \cong SO(n)/S(O(n_1) \times \cdots \times O(n_d))$. Let $Q \in SO(n)$ be an n-by-n orthogonal matrix, the equivalence class $[Q]$, representing a point on the flag manifold, is the set of orthogonal matrices

$$[Q] = \left\{ Q \begin{bmatrix} M_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & M_d \end{bmatrix} \right\}$$

where $\sum_{i=1}^{d} n_i = n$, $M_i \in O(n_i)$, and $\prod_{i=1}^{d} \det(M_i) = 1$.

3.2.1 Example: $FL(1, 1, 1)$

As a special case, a flag of type $(1, 1, \cdots, 1)$ is called a full flag and $FL(1, 1, \cdots, 1)$ is the full flag manifold in $\mathbb{R}^n$. One way to visualize a full flag in $\mathbb{R}^3$ is to picture a 1-dimensional line living in a 2-dimensional plane living in $\mathbb{R}^3$. The set of all such flags is $FL(1, 1, 1) \cong O(3)/O(1) \times O(1) \times O(1)$. From the perspective of comparing data sets, Figure 2 shows that the SVD basis of ellipsoidal data points corresponds to a flag on $FL(1, 1, 1)$. Let $[u_1, u_2, u_3] \in O(3)$ be the SVD basis of some ellipsoid ordered by the corresponding singular values, here u_1,u_2,u_3 are simply the major, median and minor axis respectively and $[u_1, u_2, u_3]$ is a flag representation of the ellipsoid data set. Comparing two ellipsoids amounts to measuring the geodesic distance between the two corresponding flags on $FL(1, 1, 1)$.

3.3. Tangent space at $[Q]$ to $FL(n_1, n_2, \cdots, n_d)$

Let Q be an element of $SO(n)$ and let $(n_1, n_2, \ldots, n_d)$ be any sequence of positive integers which add up to n. We

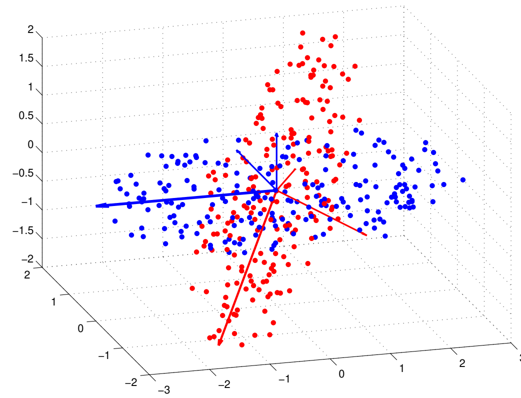

Figure 2: Two sets of ellipsoid shaped data points in $\mathbb{R}^3$. Each SVD basis can be viewed as a point on $FL(1, 1, 1)$

can use Q to build a flag with signature $(n_1, n_2, \ldots, n_d)$. In doing this, we can consider Q as a representative for a point $[Q]$ in $FL(n_1, n_2, \cdots, n_d)$. A tangent vector at $Q \in SO(n)$ can be decomposed uniquely as a component in a direction that does not modify the nested sequence of subspaces and a component in an orthogonal direction that does. The latter represent a tangent vector to $FL(n_1, n_2, \cdots, n_d)$ at $[Q]$. It can be readily computed that tangent vectors in directions that preserve the flag $[Q]$ correspond to n-by-n block diagonal skew-symmetric matrices of the form:

$$G = \begin{bmatrix} G_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & G_d \end{bmatrix}, \tag{5}$$

where G_i is an n_i-by-n_i skew-symmetric matrix. The span of matrices of this form is sometimes called the vertical space of the quotient manifold. The horizontal space is defined to be the orthogonal complement to the vertical space with respect to the standard inner product on matrices. Thus, the horizontal space consists of matrices of the form:

$$H = \begin{bmatrix} \mathbf{0}_{n_1} & & * \\ \vdots & \ddots & \vdots \\ -*^T & \cdots & \mathbf{0}_{n_d} \end{bmatrix} \tag{6}$$

where H is skew symmetric with blocks of zeros down the diagonal where $\mathbf{0}_{n_i}$ denotes an $n_i \times n_i$ matrix of zeros. Elements in the horizontal space correspond to elements in the tangent space to $FL(n_1, n_2, \cdots, n_d)$ at $[Q]$, i.e. to elements in $T_{[Q]}FL(n_1, n_2, \cdots, n_d)$.

3.4. Geodesic and distance: exp and log map

We now describe the exponential map and logarithmic map in the setting of flag manifolds.

3.4.1 Exponential map

As is mentioned earlier, a geodesic path on $SO(n)$ starting at a point Q is given by an exponential flow $Q(t) = Q\exp(tX)$ where $X \in \mathbb{R}^{n\times n}$ is any skew-symmetric matrix. Viewing $FL(n_1, n_2, \ldots, n_d)$ as a quotient manifold of $SO(n)$, one can show that a geodesic on $SO(n)$ is also a geodesic on $FL(n_1, n_2, \ldots, n_d)$ as long as the skew symmetric matrix X points in a direction that is perpendicular to the orbit determined by $S(O(n_1) \times O(n_2) \times \cdots \times O(n_d))$. This leads one to conclude that a geodesic path on $FL(n_1, n_2, \ldots, n_d)$ at $[Q]$ is an exponential flow of the form $Q(t) = Q\exp(tH)$ where H takes the form in (6).

Since each flag is an equivalence class of matrices, $Q(t)$ is just one of the possible representations of a given geodesic flow. Each geodesic flow emanating from $[Q] \in FL(n_1, n_2, \cdots, n_d)$ has the form

$$[Q(t)] = \left\{ Q\exp(tH) \begin{bmatrix} M_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & M_d \end{bmatrix} \right\} \quad (7)$$

where $M_i \in O(n_i)$ and $\prod_{i=1}^{d} \det(M_i) = 1$. Equipped with the metric induced by the inner product $< A, B > = \frac{1}{2} Tr(A^T B)$, we can compute the length of the path between $[Q(0)]$ and $[Q(1)]$ along the geodesic determined by H:

$$Length = \sqrt{\frac{1}{2}\mathrm{Tr}(H^T H)} = \sqrt{\frac{1}{2}\sum_{j=1}^{l} \lambda_j^2} \quad (8)$$

where $\{\pm i\lambda_j\}$ are the eigenvalues of H. This mapping of a tangent vector (based at $[Q]$) to the flag manifold is referred to as the exponential map which in this paper is found by applying the matrix exponential.

3.4.2 Logarithmic map

In data analysis, it is often the case that one is given data sets or representations of data sets (e.g. through an SVD basis) and one wants to measure their similarity. If the representation of the data is given as an orthonormal matrix, M, one can consider the columns of M as an ordered basis and use this ordering to consider M as a representative for a point $[M]$ on a flag manifold. An interesting feature of flag manifolds is that there are typically many geodesics between points. In order to measure the distance between two points on a flag manifold, one needs to find the length of the *shortest* geodesic between their representations. In order to do this, one needs to find a tangent vector, H, that achieves the smallest value for $< H, H >$ among all tangent vectors determining a geodesic between the points. This tangent vector is found via the inverse operation of the exponential map (referred to as the logarithmic map). In [19] there is a description of an iterative algorithm which takes as input two orthogonal matrices Q_0, Q_1 and a flag signature $(n_1, \ldots, n_d)$ and produces as output skew symmetric matrices, G, H, of the forms given in (5) and (6) which satisfy $Q_1 = Q_0 exp(H) exp(G)$ and with $< H, H >$ minimal among all such expressions of this form. In other words, the iterative algorithm approximates the tangent vector H that determines a minimal length geodesic (and determines the distance) between $[Q_0]$ and $[Q_1]$ on $FL(n_1, n_2, \cdots, n_d)$.

3.5. 2k Embedding

For many practical applications, the trailing n_d columns are not of interest, e.g. computations on $FL(k, n-k) = Gr(k, n)$ are usually performed using n-by-k orthonormal matrices since only the first k columns are of interest. Here in this section we will prove that the iterative algorithm from [19] can be performed in a lower dimensional space if $k = \sum_{i=1}^{d-1} n_i$ is relatively small, more specifically, if $k < n/2$.

Without loss of generality, the geodesic between two flags of type $(n_1, n_2, \cdots, n_d)$ can always be identified with a geodesic between the identity matrix, I, and some $Q \in SO(n)$ by moving the initial point to I, i.e.,

$$Q = I\exp(\begin{bmatrix} A & -B^T \\ B & 0 \end{bmatrix}) \quad (9)$$

where $k = \sum_{i=1}^{d-1} n_i$, $B \in \mathbb{R}^{(n-k)\times k}$ and A is a k-by-k skew-symmetric matrix of the form

$$A = \begin{bmatrix} \mathbf{0}_{n_1} & -B_{2,1}^T & \cdots & -B_{d-1,1}^T \\ B_{2,1} & \mathbf{0}_{n_2} & & -B_{d-1,2}^T \\ \vdots & & \ddots & \vdots \\ B_{d-1,1} & B_{d-1,2} & \cdots & \mathbf{0}_{n_{d-1}} \end{bmatrix}. \quad (10)$$

$Q(t) = I\exp(t\begin{bmatrix} A & -B^T \\ B & 0 \end{bmatrix}), t \in [0, 1]$ traces an n-by-n representation of the geodesic flow between $[I]$ and $[Q]$. The following theorem and its corollary provides a method to modify the iterative algorithm from [19] with $2k$-by-$2k$ matrices instead of n-by-n matrices.

Theorem 1. *Let $[Q] \in FL(n_1, n_2, \cdots, n_d)$. Suppose $Q(t) = \exp(t\begin{bmatrix} A & -B^T \\ B & 0 \end{bmatrix})$ with $Q(0) = I$, $Q(1) = Q$ is a flag geodesic flow between $[I]$ and $[Q]$. If*

$$q(t) = \exp(t\begin{bmatrix} A & -B^T \\ B & 0 \end{bmatrix}) I_{n,k} \quad (11)$$

and $\mathrm{span}\{q(0)\} \cap \mathrm{span}\{q(1)\} = \{0\}$, then for all $t \in [0, 1]$, $\mathrm{span}\{q(t)\} \subset \mathrm{span}\{[q(0), q(1)]\}$, where $k = \sum_{i=1}^{d-1} n_i$ and $I_{n,k}$ denotes the first k columns of an n-by-n identity matrix.

Note that if $2k \geq n$, Theorem 1 is trivial. So here we assume $2k < n$. Before proving the theorem, we need to introduce some notation. Let $q := QI_{n,k} = q(1)$ be the first k columns of Q. In fact, $q(t)$ defined in Equation (11) can be understood as a geodesic path between $I_{n,k}$ and q by viewing $FL(n_1, n_2, \cdots, n_d)$ as a quotient manifold of the Stiefel manifold $St(k, n)$ (refer to [33] for more details). Further, we write the n-by-k orthonormal matrix q in block matrix form as

$$q = \begin{bmatrix} q_k \\ q_{n-k} \end{bmatrix} \tag{12}$$

where q_k and q_{n-k} denote the first k rows and the trailing $n-k$ rows of q respectively.

Lemma 1. *If $q(t)$ is defined as in Equation* (11)*, such that $q(0) = I_{n,k}$ and $q(1) = q$, then* $\operatorname{span}\{q_{n-k}\} = \operatorname{span}\{B\}$.

Proof. Let $U_B R_B := B$ be the compact QR decomposition of B (U_B: $(n-k)$-by-k, R_B: k-by-k). Define

$$f(t) = (I - U_B U_B^T) J q(t) \tag{13}$$

where $J = \begin{bmatrix} 0 & I_{n-k} \end{bmatrix}$ is the last $n-k$ rows of the n-by-n identity matrix. Hence left multiplication by J on $q(t)$ simply selects the last $n-k$ rows of $q(t)$. By definition $f(0) = 0$. Differentiate $f(t)$ to get:

$$\dot{f}(t) = (I - U_B U_B^T) J \begin{bmatrix} A & -B^T \\ B & 0 \end{bmatrix} q(t) = 0 \tag{14}$$

Therefore, $f(t) \equiv 0$ for $t \in [0,1]$. If we evaluate $f(t)$ at $t = 1$, we get:

$$f(1) = (I - U_B U_B^T) q_{n-k} = 0 \tag{15}$$

By the assumption that $q(0)$ and $q(1)$ do not intersect, we know q_{n-k} is of rank k hence U_B is also of rank k. The conclusion follows. □

Now we present a proof to the theorem.

Proof. Let $UR := [I_{n,k}, q]$ be the thin QR-decomposition of $[q(0), q(1)]$. Consequently, U is an orthonormal basis for $\operatorname{span}\{[q(0), q(1)]\}$. The n-by-k orthonormal matrix U takes the block form

$$U = \begin{bmatrix} I_k & 0 \\ 0 & C \end{bmatrix}. \tag{16}$$

Note that $\operatorname{span}\{C\} = \operatorname{span}\{q_{n-k}\}$ where q_{n-k} is defined in Equation (12). Define

$$g(t) = (I - UU^T) q(t). \tag{17}$$

By definition, $g(0) = (I - UU^T) I_{n-k} = 0$. If we differentiate $g(t)$, we get:

$$\dot{g}(t) = \begin{bmatrix} 0 & 0 \\ (I_{n-k} - CC^T)B & 0 \end{bmatrix} q(t) \tag{18}$$

By Lemma 1, $\operatorname{span}\{B\} = \operatorname{span}\{q_{n-k}\} = \operatorname{span}\{C\}$. We conclude that $\dot{g}(t) \equiv 0$, which implies $g(t) \equiv 0$. Therefore $q(t)$ is always living in the span of $[q(0), q(1)]$. □

The theorem shows that the flag geodesic flow $q(t)$ between $I_{n,k}$ and q never leaves the $2k$-dimensional subspace $\operatorname{span}\{[I_{n,k}, q]\}$, which leads to the conclusion that the logarithmic map computation can be performed within this $2k$ dimensional space without loss of information. Here we introduce the following corollary.

Corollary 1. *Suppose $q(t)$ is defined as in Equation* (11) *such that $q(0) = I_{n,k}$ and $q(1) = q$. Let $UR := [I_{n,k}, q]$ be the compact QR-decomposition of $[q(0), q(1)]$, then $\phi(t) = U^T q(t)$ is a geodesic flow between $\phi(0) = U^T q(0)$ and $\phi(1) = U^T q(1)$ on $FL(n_1, n_2, \cdots, n_{d-1}, k)$. Moreover, $d(\phi(0), \phi(1)) = d(q(0), q(1))$ and $q(t) = UU^T \phi(t)$.*

This corollary can be proved by combining the results from Theorem 1 and Corollary 2.2 in [7].

4. Numerical Experiments

4.1. Ellipsoid data

The purpose of this synthetic example is to show the difference between a flag geodesic and a Grassmannian geodesic, as well as their corresponding geodesic distances, when comparing data sets. As can be seen in Figure 2, each ellipsoid data cloud contains 100 data points in $\mathbb{R}^3$. Let $\{r_i\}$ and $\{b_i\}$ denote the data points in the red and blue ellipsoid respectively. Each data set can be written as a short wide data matrix $[r_1, r_2, \cdots, r_{100}] = R \in \mathbb{R}^{3\times 100}$ and $[b_1, b_2, \cdots, b_{100}] = B \in \mathbb{R}^{3\times 100}$. We denote the SVD basis for each ellipsoid data set by $U_R = [u_R^{(1)}, u_R^{(2)}, u_R^{(3)}]$ and $U_B = [u_B^{(1)}, u_B^{(2)}, u_B^{(3)}]$. One can view the SVD basis as giving the major, medium, and minor axes of the corresponding ellipsoid.

The Grassmannian geodesic distance between the two bases is 0 since the columns of U_R or U_B span all of $\mathbb{R}^3$. To compare two ellipsoids via the Grassmannian setting, one would typically represent the data sets with their first principal components namely $u_R^{(1)}$ and $u_B^{(1)}$, and then compute the distance between these two vectors on $Gr(1,3)$. Hence the Grassmannian geodesic between the two ellipsoids is the rotation that moves the major axis of the first ellipsoid to the major axis of the second ellipsoid. The distance between these representations on $Gr(1,3)$ is the angle between the major axes. The information contained in the relationship between the other two axes is lost. Note that this limitation comes from the Grassmannian rather than the data itself.

By representing the two ellipsoids of data points by their SVD bases U_R, U_B such that $[U_R]$, $[U_B] \in FL(1,1,1)$, one retains more information in describing the corresponding ellipsoids since $FL(1,1,1)$ has dimension 3 (while

$Gr(1, 3)$ has dimension 2). The geodesic between the two flag representations correspondingly encodes more information than in the Grassmannian setting.

4.2. MNIST image data set

Here we utilize the well-studied MNIST data set to illustrate the use of the flag manifold for comparing sets of SVD bases of "mixed" digits. We select hand written digits "1" and "5" from the training set of the MNIST data set, where each digit is a 28×28 image. All images are vectorized and centered by subtracting the mean of all images. Then we form a set of mixed digits data sets consisting of two classes, namely "major $1/$minor 5" and "major $5/$minor 1". "major $1/$minor 5" (resp. "major $5/$minor 1") is formed by concatenating m "1"'s (resp. m "5"'s) and p "5"'s (resp. p "1"'s). In general m is assumed to be larger then p. Hence each data set is represented by a $784 \times (m + p)$ matrix. We compute the SVD basis for each $784 \times (m + p)$ matrix and select the first k columns of the SVD basis as a representation for each data set. Thus each data set is represented by a $784 \times k$ orthonormal matrix. For the following experiment $m = 16$, $p = 9$ and $k = 5$. We may consider each 784×5 SVD basis as a data point on $FL(2, 3, 779)$ or $Gr(5, 784)$. The first 5 eigen-digits for both of the two classes in this experiment are presented in Figure 4. One can compute the pairwise flag and Grassmannian geodesic distances and store the data in a distance matrix. We then embed these data points in Euclidean space using multidimensional scaling.

In Figure 3, we see the configurations produced from MDS using Grassmannian and flag distances. We observe that the Grassmannian MDS configuration has significant overlapping between the two classes. This is not surprising since each data point, no matter which class is considered, captures the span of both "1"'s and "5"'s. As can be seen in the flag MDS configuration, there is a much clearer separation between the two classes (except for one point). Note that the input matrices fed to the algorithm are identical for each of the configurations. The improvement comes from the additional structure in the flag.

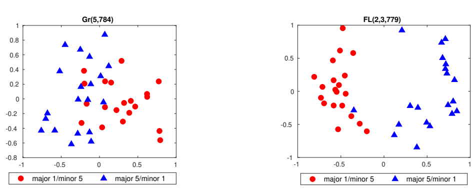

Figure 3: Grassmannian vs flag MDS configurations

4.3. Indian Pines hyperspectral image data

To further illustrate the utility of the flag model, we apply it to the Indian Pines hyperspectral image data set. The

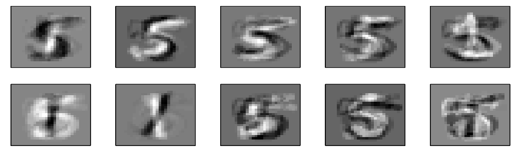

Figure 4: First 5 eigendigits of major 5/minor 1 data set and of major 1/minor 5 data set

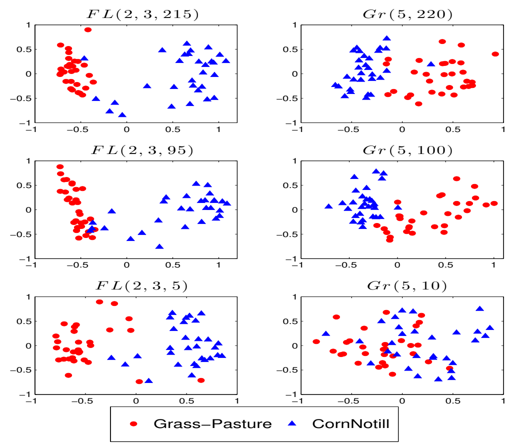

Figure 5: Comparison of the Grassmann and Flag manifolds for representing Grass-Pasture and CornNotill. The subspace dimension is fixed while ambient dimension varies.

hyperspectral images in this data set are 145×145 pixels by 220 spectral bands (from $0.4\mu m$ to $2.4\mu m$). 10366 pixels are labelled and each is assigned to one of 16 classes. Here we will test both the flag model and the Grassmann model on the task of visualizing sets of data sets.

For a chosen dimension k (note that $k = \sum_{i=1}^{d-1} n_i$ for $FL(n_1, n_2, \cdots, n_d)$), we assemble 30 $n \times k$ matrices X_i from each class (so there are $p = 60$ data matrices in total). Each data matrix consists of k 200×1 data vectors which belong to one of the two classes. Then for each matrix X_i, a compact SVD is applied to obtain an SVD/PCA basis, hence each data point (subspace) is represented by a $220 \times k$ orthonormal matrix U_i where $U_i \Sigma_i V_i^T = X_i$. The distance between SVD bases, assumed as representatives for points on a given flag manifold, can then be computed to obtain a $p \times p$ distance matrix. We use this distance matrix to embed these flags as points in Euclidean space via Multi-Dimensional Scaling (MDS). The first two coordinates of the optimal Euclidean configuration are selected for visualization in $\mathbb{R}^2$. Figure 5 illustrates the Euclidean embedding configurations for fixed subspace dimension $k = 5$ with various ambient dimensions using both the Grassmannian geodesic distance and flag distance. The ambient space is selected to be the n spectral bands with highest responses for $n = 100, 10, 5$. It is observed in the first two rows that both Grassmannian and flag geodesic distance provide a good separation with relatively large ambient dimension at $n = 220$ and 100. When the ambient dimension is reduced to $n = 10$, the third row of Figure 5 shows that the flag distance MDS embedding separates two classes in $\mathbb{R}^2$ while the Grassmannian MDS embedding shows heavy overlapping. Figure 6 shows the eigenvalues corresponding to the MDS embedding using flag distance on $FL(2, 3, 5)$ (left) and $Gr(5, 10)$ (right). As we can see, the largest eigenvalue on the left panel is dominating which also suggests that flag MDS configurations are separable in lower dimension, which we don't observe in the Grassmannian MDS eigenvalues plot. Figure 7 shows, for fixed ambient dimension $n = 220$, how sets of data sets are pulled apart by increasing the dimension in the flag structure. From top left, we observe that the embedding of data points on $FL(1, 219)$ to $\mathbb{R}^2$ live on a circle and are not separable. As we increase the flag structure dimension, the corresponding MDS configurations start to show more separation and for $FL(1, 4, 215)$, the embedding of two classes is linearly separable.

In Figure 8, we select 6 bands (bands: 3,29,42,61,65,158) and use 20 pixels within the same class to form a data matrix of size 6×30. Each class consists of 20 such short and wide matrices and each matrix is represented by its 6-by-6 SVD basis and assumed to be representatives for points on $FL(2, 2, 2)$. The pairwise distance is computed to obtain MDS configurations on $\mathbb{R}^2$. It is observed that the MDS embeddings of 3 classes are separable in low dimensional space with only 6 bands.

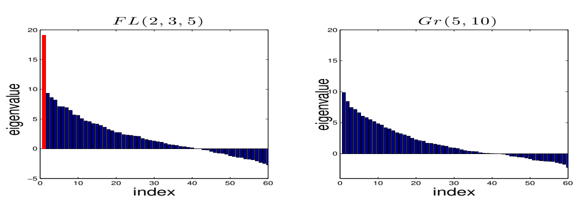

Figure 6: Eigenvalues of MDS for $FL(2, 3, 5)$ and $Gr(5, 10)$ in descending order.

5. Conclusion

We have proposed a geometric framework for comparing distances between nested subspaces, i.e., points on a flag manifold. This approach exploits a mathematical framework that enables the data analyst to gain insight into the way the data resides in its ambient space, both in terms of dimension and distribution. This approach is suitable for the analysis of wide data matrices, e.g., where the number of data features is less than the number of points and for data sets consisting of a mixture of classes.

We have presented the theoretical foundation for computing geodesic distances between two points on a flag manifold. The theory lends itself naturally to numerical algorithms for computing the distance as well as the set of points along the shortest path between the two points. This formulation allows one to move a set of nested subspaces into another set of nested subspaces along the shortest path that respects the intrinsic geometry. These tools provide a mechanism to leverage angles between subspaces where the previous formalism on the Grassmannian may fail.

The flag geodesic algorithms have been demonstrated on mixed MNIST data sets and on the Indian Pines hyperspectral data set where the number of hyperspectral features (each corresponding to a frequency band) and flag structure are varied. In particular, we focus on the transition from tall to wide matrices. We see that the geodesic distance on the flag manifold is able to separate the data for visualization in two dimensions in a setting where the Grassmannian framework failed to do so.

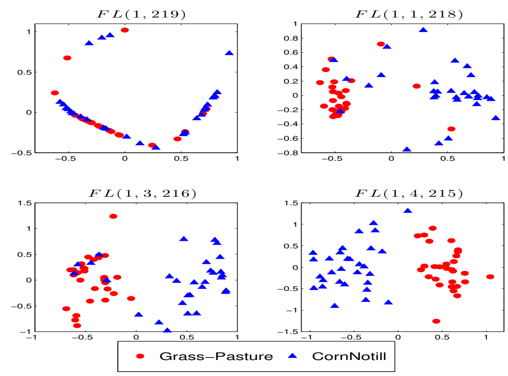

Figure 7: Configuration of points on various flag manifolds embedded in Euclidean space.

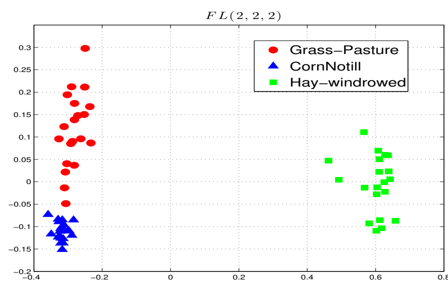

Figure 8: Configuration of points on $FL(2, 2, 2)$ embedded in $\mathbb{R}^2$ for 3 classes.

Acknowledgment

This paper is based on research partially supported by the National Science Foundation under Grants No. NSF-1633830, NSF-1830676, and NSF-1712788.

References

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[2] Sherif Azary and Andreas Savakis. Grassmannian sparse representations and motion depth surfaces for 3d action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 492–499, 2013.

[3] Rudrasis Chakraborty and Baba 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.

[4] Jen-Mei Chang. *Classification on the Grassmannians: theory and applications*. Colorado State University, 2008.

[5] John H Conway, Ronald H Hardin, and Neil JA Sloane. Packing lines, planes, etc.: Packings in Grassmannian spaces. *Experimental mathematics*, 5(2):139–159, 1996.

[6] Bruce Draper, Michael Kirby, Justin Marks, Tim Marrinan, and Chris Peterson. A flag representation for finite collections of subspaces of mixed dimensions. *Linear Algebra and its Applications*, 451:15–32, 2014.

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

[8] Simone Fiori. Extended Hamiltonian learning on Riemannian manifolds: Theoretical aspects. *IEEE transactions on neural networks*, 22(5):687–700, 2011.

[9] Kyle A Gallivan, Anuj Srivastava, Xiuwen Liu, and Paul Van Dooren. Efficient algorithms for inferences on Grassmann manifolds. In *IEEE Workshop on Statistical Signal Processing, 2003*, pages 315–318. IEEE, 2003.

[10] Mehrtash T Harandi, Conrad Sanderson, Sareh Shirazi, and Brian C Lovell. Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching. In *CVPR 2011*, pages 2705–2712. IEEE, 2011.

[11] Jun He, Laura Balzano, and Arthur Szlam. Incremental gradient on the Grassmannian for online foreground and background separation in subsampled video. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1568–1575. IEEE, 2012.

[12] Yi Hong, Roland Kwitt, Nikhil Singh, Brad Davis, Nuno Vasconcelos, and Marc Niethammer. Geodesic regression on the Grassmannian. In *European Conference on Computer Vision*, pages 632–646. Springer, 2014.

[13] Zhiwu Huang, Jiqing Wu, and Luc Van Gool. Building deep networks on Grassmann manifolds. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[14] Sriram Kumar and Andreas Savakis. Robust domain adaptation on the l1-Grassmannian manifold. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 103–110, 2016.

[15] Gitta Kutyniok, Ali Pezeshki, Robert Calderbank, and Taotao Liu. Robust dimension reduction, fusion frames, and Grassmannian packings. *Applied and Computational Harmonic Analysis*, 26(1):64–76, 2009.

[16] Mengyi Liu, Ruiping Wang, Zhiwu Huang, Shiguang Shan, and Xilin Chen. Partial least squares regression on Grassmannian manifold for emotion recognition. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 525–530. ACM, 2013.

[17] Bei Ma and Hailin Zhang. Recognition of faces using texture-based principal component analysis and Grassmannian distances analysis. In *International Conference on Graphic and Image Processing (ICGIP 2011)*, volume 8285, page 82856C. International Society for Optics and Photonics, 2011.

[18] Xiaofeng Ma, Michael Kirby, and Peterson Chris. Self-organizing mappings on the flag manifold with applications to hyper-spectral image data analysis. *Neural Computing and Applications*, pages 1–11, 2021.

[19] Xiaofeng Ma, Michael Kirby, and Chris Peterson. Self-organizing mappings on the flag manifold. In *International Workshop on Self-Organizing Maps*, pages 13–22. Springer, 2019.

[20] Tim Marrinan, J Ross Beveridge, Bruce Draper, Michael Kirby, and Chris Peterson. Flag manifolds for the characterization of geometric structure in large data sets. In *Numerical Mathematics and Advanced Applications-ENUMATH 2013*, pages 457–465. Springer, 2015.

[21] Timothy Marrinan, J Ross Beveridge, Bruce Draper, Michael Kirby, and Chris Peterson. Flag-based detection of weak gas signatures in long-wave infrared hyperspectral image sequences. In *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XXII*, volume 9840, page 98401N. International Society for Optics and Photonics, 2016.

[22] Yasunori Nishimori, Shotaro Akaho, Samer Abdallah, and Mark D Plumbley. Flag manifolds for subspace ICA problems. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–1417. IEEE, 2007.

[23] Yasunori Nishimori, Shotaro Akaho, and Mark D Plumbley. Riemannian optimization method on generalized flag manifolds for complex and subspace ICA. In *AIP Conference Proceedings*, volume 872, pages 89–96. American Institute of Physics, 2006.

[24] Yasunori Nishimori, Shotaro Akaho, and Mark D Plumbley. Riemannian optimization method on the flag manifold for independent subspace analysis. In *International Conference on Independent Component Analysis and Signal Separation*, pages 295–302. Springer, 2006.

[25] Yasunori Nishimori, Shotaro Akaho, and Mark D Plumbley. Natural conjugate gradient on complex flag manifolds for complex independent subspace analysis. In *International Conference on Artificial Neural Networks*, pages 165–174. Springer, 2008.

[26] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

[27] David A Shaw and Rama Chellappa. Regression on manifolds using data-dependent regularization with applications in computer vision. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6(6):519–528, 2013.

[28] Thomas Strohmer and Robert W Heath Jr. Grassmannian frames with applications to coding and communication. *Applied and computational harmonic analysis*, 14(3):257–275, 2003.

[29] Sima Taheri, Pavan Turaga, and Rama Chellappa. Towards view-invariant expression analysis using analytic shape manifolds. In *Face and Gesture 2011*, pages 306–313. IEEE, 2011.

[30] Pavan Turaga and Rama Chellappa. Locally time-invariant models of human activities using trajectories on the Grassmannian. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2435–2441. IEEE, 2009.

[31] Tiesheng Wang and Pengfei Shi. Kernel Grassmannian distances and discriminant analysis for face recognition from image sets. *Pattern Recognition Letters*, 30(13):1161–1165, 2009.

[32] Xinchao Wang, Wei Bian, and Dacheng Tao. Grassmannian regularized structured multi-view embedding for image classification. *IEEE Transactions on Image Processing*, 22(7):2646–2660, 2013.

[33] Ke Ye, Ken Sze-Wai Wong, and Lek-Heng Lim. Optimization on flag manifolds. *arXiv e-prints*, page arXiv:1907.00949, Jul 2019.