Linear Algebra and its Applications 451 (2014) 15–32

Contents lists available at ScienceDirect

Linear Algebra and its Applications

www.elsevier.com/locate/laa

A flag representation for finite collections of subspaces of mixed dimensions

Bruce Draper[a], Michael Kirby[b], Justin Marks[c], Tim Marrinan[b,*], Chris Peterson[b]

[a] Department of Computer Science, Colorado State University, Fort Collins, CO 80523, United States
[b] Department of Mathematics, Colorado State University, Fort Collins, CO 80523, United States
[c] Department of Mathematics, Bowdoin College, Brunswick, ME 04011, United States

ARTICLE INFO

Article history:
Received 6 September 2013
Accepted 14 March 2014
Available online 1 April 2014
Submitted by M. Benzi

MSC:
14
15
53
14M15
15A18
53B20

Keywords:
Subspace average
Grassmann manifold
Flag manifold
SVD
Flag mean
Multiset Canonical Correlation Analysis
Extrinsic manifold mean

ABSTRACT

Given a finite set of subspaces of $\mathbb{R}^n$, perhaps of differing dimensions, we describe a flag of vector spaces (i.e. a nested sequence of vector spaces) that best represents the collection based on a natural optimization criterion and we present an algorithm for its computation. The utility of this flag representation lies in its ability to represent a collection of subspaces of differing dimensions. When the set of subspaces all have the same dimension d, the flag mean is related to several commonly used subspace representations. For instance, the d-dimensional subspace in the flag corresponds to the extrinsic manifold mean. When the set of subspaces is both well clustered and equidimensional of dimension d, then the d-dimensional component of the flag provides an approximation to the Karcher mean. An intermediate matrix used to construct the flag can also be used to recover the canonical components at the heart of Multiset Canonical Correlation Analysis. Two examples utilizing the Carnegie Mellon University Pose, Illumination, and Expression Database (CMU-PIE) serve as visual illustrations of the algorithm.

[*] Corresponding author. Tel.: +1 5093011787.
E-mail addresses: draper@cs.colostate.edu (B. Draper), kirby@math.colostate.edu (M. Kirby), jmarks@bowdoin.edu (J. Marks), marrinan@math.colostate.edu (T. Marrinan), peterson@math.colostate.edu (C. Peterson).

http://dx.doi.org/10.1016/j.laa.2014.03.022

1. Introduction

The Grassmann manifold has found use as a setting in which to classify and make comparisons between large data sets. It is particularly effective when aspects of the data can be captured with linear subspaces. A sampling of settings where Grassmann techniques have been applied includes activity modeling and recognition, shape analysis, action classification, face recognition, person detection, subspace tracking, and general manifold clustering [14,10,12,4–6,19,20,18]. Given a cluster of points on a Grassmann manifold, algorithms have been developed to find a point on the manifold which represents the cluster [5,1,7,17]. These cluster representatives play the role of an *average subspace* and can be used to reduce the cost of classification algorithms or to aid in clustering tasks.

In a more general setting, consider data consisting of subspaces of $\mathbb{R}^n$ of differing dimensions, i.e. a data cloud living on a disjoint union of Grassmann manifolds. This paper proposes a *flag mean* representation for such a collection. The flag mean is a nested sequence of vector spaces that best fits the data according to an optimization criterion based on the projection Frobenius norm. The subspaces in the flag can be treated independently as points on Grassmann manifolds, or collectively as a single point on a *flag manifold*.

The layout of the paper is as follows. Section 2 provides background, definitions and motivation for the construction. Section 3 presents the optimization problem, whose solution is the flag mean, and provides an analytical solution by the method of Lagrange multipliers. The result is an ordered set of unit length vectors most central to the collection of subspaces being averaged. Section 4 exploits the singular value decomposition as an efficient computational tool for determining this ordered set of unit length vectors and connects them to multiset canonical correlation analysis. In Section 5 the central vectors are used to construct the flag mean. The construction is then illustrated with two numerical experiments using data drawn from the Carnegie Mellon University Pose, Illumination, and Expression Database (CMU-PIE). Section 6 explores a special case of the flag mean and relates it to alternative subspace means found in the literature. Section 7 discusses conclusions and future work.

2. Background

Many image and video based computer vision systems represent data as a set of linear subspaces of a fixed dimension [17,19,6,4,1]. This structure allows the data to be treated as a collection of points on a single Grassmann manifold. The Grassmann manifold $\mathrm{Gr}(V,p)$ is a manifold whose points parametrize the subspaces of dimension p inside the vector space V. In this paper, we will assume that V is a finite dimensional real vector space and thus we can identify V through its dimension, n. For the remainder of this paper, we denote by $\mathrm{Gr}(n,p)$ the Grassmann manifold of p dimensional subspaces of $\mathbb{R}^n$, $GL(p)$ denotes the general linear group of invertible $p \times p$ matrices and $O(p)$ denotes the orthogonal group of $p \times p$ orthogonal matrices.

Let $\mathbb{R}^{n\times p}$ denote the vector space of $n \times p$ matrices with real entries and let $(\mathbb{R}^{n\times p})^\circ$ denote the open submanifold of full rank $n \times p$ matrices. For each $Y \in (\mathbb{R}^{n\times p})^\circ$, let $[Y]$ denote the column space of Y. There is a natural surjective map $\phi : (\mathbb{R}^{n\times p})^\circ \to \mathrm{Gr}(n,p)$ given by $\phi(Y) = [Y]$ (with $[Y]$ identified with its corresponding point on $\mathrm{Gr}(n,p)$). It is clear that $\phi(X) = \phi(Y)$ if and only if there exists an $A \in GL(p)$ such that $XA = Y$. Thus a point q on $\mathrm{Gr}(n,p)$ corresponds to a p-dimensional subspace V_q of $\mathbb{R}^n$ and can be represented by any element of a $GL(p)$ orbit of a full rank $n \times p$ matrix, Y, whose column space $[Y]$ is equal to V_q. For the purposes of computation, we utilize a representative with orthonormal columns (note that if Y is a representative with orthonormal columns and if $B \in O(p)$ then YB will be another representative with orthonormal columns). Since a matrix representative, with orthonormal columns, for a point on a Grassmann manifold is only unique up to right multiplication by an orthogonal matrix, it is important that the output of any algorithm is invariant to such a multiplication.

Let $d : \mathrm{Gr}(n,p) \times \mathrm{Gr}(n,p) \to \mathbb{R}$ be a metric. The metric, d, is said to be *orthogonally invariant* if for every $[X], [Y] \in \mathrm{Gr}(n,p)$ and every $A \in O(n)$, $d([X],[Y]) = d([AX],[AY])$. One commonly used distance measure on Grassmannians is the metric induced from the projection Frobenius norm (denoted by d_{pF}). It is an elementary exercise to show that d_{pF} is an orthogonally invariant metric on $\mathrm{Gr}(n,p)$. The projection Frobenius norm arises from the identification of points in $\mathrm{Gr}(n,p)$ with $n \times n$ projection matrices of rank p. If X, Y are full rank $n \times p$ matrices with orthonormal columns, then the distance between $[X], [Y] \in \mathrm{Gr}(n,p)$ is computed as a constant times the Frobenius norm of the difference between the projection matrix representations of the points: $d_{pF}([X],[Y]) = 2^{-\frac{1}{2}} \|XX^T - YY^T\|_F$. As shown by Edelman et al., this distance can also be computed as the ℓ_2-norm of the vector of the sines of the principal angles between $[X]$ and $[Y]$ [5]. If X (resp. Y) are orthonormal matrix representatives for $[X]$ (resp. $[Y]$) then the cosines of the principal angles between $[X]$ and $[Y]$ are the singular values of $X^T Y$ [3]. Note that if $A, B \in O(p)$, then the singular values of $X^T Y$ are the same as the singular values of $(XA)^T (YB)$.

In many applications, it can be natural and advantageous to represent aspects of data through subspaces lying in a fixed ambient space that are of differing dimensions. In such applications, a set of subspaces live naturally on a collection of Grassmann manifolds rather than on a single Grassmann manifold. Suppose that $[X] \in \mathrm{Gr}(n,p_1)$ and $[Y] \in \mathrm{Gr}(n,p_2)$ for $p_1 < p_2$. As illustrated in Bjork and Golub's foundational paper, there will be p_1 principal angles between $[X]$ and $[Y]$ [3] and we define $d_{pF}([X],[Y])$ as the ℓ_2-norm of the vector of the sines of the p_1 principal angles between $[X]$ and $[Y]$. Note that d_{pF} is no longer a metric due to the possibility of $d_{pF}([X],[Y]) = 0$ while $[X] \neq [Y]$ (for instance, if $[X]$ is a proper subspace of $[Y]$).

The scenario in which data are represented as subspaces of differing dimensions is of central interest to this paper. Finding an average representation for objects of this type is a very practical problem. For example, an “action subspace” can be approximated from the span of the frames of a video clip. The ambient dimension of such a subspace would be the number of pixels in each frame, but the subspace dimension would vary depending on what type action was being represented. Similarly, an “illumination subspace” might be computed from images of an object under a variety of lighting conditions. The dimension of this subspace would depend on the number of unique surface normals, as explained by Belhumeur and Kriegman [2]. Thus, different objects would need subspaces of different sizes to fully capture lighting information. In either scenario, finding an average or prototype for a collection of these subspaces has benefits when computing statistics, clustering, or classifying data.

Existing methods for representing subspaces cannot handle this type of variation, because they require that all points live on a single Grassmann manifold. In contrast, this paper works with collections of subspaces such as these, that live on a disjoint union of Grassmannians. It proposes that the whole collection of subspaces can be well represented by a flag, and that a subset of the flag serves as a natural representative on each of the Grassmannians. The flag avoids some of the undesirable side-effects of standardizing the subspaces to a single manifold, and contains more subtle information than existing subspace averages.

3. The flag mean optimization problem

Let $\mathcal{D} = \{[X_i]\}_{i=1}^{N}$ be a finite collection of subspaces of $\mathbb{R}^n$. Consider the set of positive integers $\mathcal{P} = \{\dim([X_i]) \mid [X_i] \in \mathcal{D}\}$. We can consider $\mathcal{D}$ as a collection of points lying on the disjoint union of Grassmannians, $\coprod_{p_i \in \mathcal{P}} \mathrm{Gr}(n, p_i)$. We wish to find the one-dimensional subspace $[u^{(1)}] \in \mathrm{Gr}(n,1)$ that lies closest to the elements in $\mathcal{D}$ as measured by the sum of the squares of the sine of the principal angle between $[u^{(1)}]$ and the elements of $\mathcal{D}$. Thus we define:

$$\left[u^{(1)}\right] := \operatorname*{arg\,min}_{[u]\in \mathrm{Gr}(n,1)} \sum_{[X_i]\in\mathcal{D}} d_{pF}\big([u],[X_i]\big)^2 \tag{1}$$

While this optimization problem has some similarities to the Riemannian center of mass, there are important differences in that it uses d_{pF} (instead of the geodesic distance based on arc length), the elements of $\mathcal{D}$ are not restricted to live on a single Grassmannian, and $[u^{(1)}] \in \mathrm{Gr}(n,1)$. We can extend the problem to find a sequence of optimizers, $[u^{(1)}], [u^{(2)}], \dots$ by adding orthogonality constraints. In particular, we define:

$$\begin{aligned} \left[u^{(j)}\right] &:= \operatorname*{arg\,min}_{[u]\in \mathrm{Gr}(n,1)} \sum_{[X_i]\in\mathcal{D}} d_{pF}\big([u],[X_i]\big)^2 \\ &\text{subject to} \quad [u] \perp \left[u^{(l)}\right] \quad \text{for } l < j, \end{aligned} \tag{2}$$

This leads to the set $\{[u^{(1)}], [u^{(2)}], \ldots, [u^{(r)}]\}$ where r denotes the dimension of the span of the elements in $\mathcal{D}$.

Recalling that $d_{pF}([X],[Y]) = \|\sin\Theta\|_2$ (where $\sin\Theta$ denotes the vector whose entries are the sines of the principal angles between $[X]$ and $[Y]$), the sequence of optimizers can be found analytically. Let θ_i be the lone principal angle between $[u]$ and $[X_i]$ and let $u, X_1, \ldots, X_N$ be orthonormal matrix representatives for $[u], [X_1], \ldots, [X_N]$. For $j = 1, \ldots, r$ we can rewrite the cost function of Eq. (2) as,

$$\begin{aligned}
\big[u^{(j)}\big] &= \operatorname*{arg\,min}_{[u]\in \mathrm{Gr}(n,1)} \sum_{[X_i]\in\mathcal{D}} d_{pF}\big([u],[X_i]\big)^2 && (3)\\
&= \operatorname*{arg\,min}_{[u]\in \mathrm{Gr}(n,1)} \sum_{[X_i]\in\mathcal{D}} \|\sin\theta_i\|_2^2 && (4)\\
&= \operatorname*{arg\,max}_{[u]\in \mathrm{Gr}(n,1)} \sum_{[X_i]\in\mathcal{D}} \|\cos\theta_i\|_2^2, && (5)\\
&= \operatorname*{arg\,max}_{[u]\in \mathrm{Gr}(n,1)} \sum_{[X_i]\in\mathcal{D}} \cos^2\theta_i && (6)\\
&= \operatorname*{arg\,max}_{[u]\in \mathrm{Gr}(n,1)} \sum_{[X_i]\in\mathcal{D}} u^T X_i X_i^T u && (7)\\
&= \operatorname*{arg\,max}_{[u]\in \mathrm{Gr}(n,1)} u^T\left(\sum_{i=1}^{N} X_i X_i^T\right)u && (8)
\end{aligned}$$

The equality between Eq. (6) and Eq. (7) follows from the thin singular value decomposition of $u^T X_i$. Let $p_i = \dim([X_i])$. We know that X_i is an $n \times p_i$ matrix whose columns form an orthonormal basis for $[X_i]$ and u is restricted to be a unit vector in $\mathbb{R}^n$ whose span is $[u]$. If the thin SVD of $u^T X_i$ is written as $U\Sigma V^T$, then U is a 1×1 matrix whose only entry is ± 1, Σ is a $1 \times p_i$ matrix whose first entry is $\cos\theta_i$ with the other entries equal to zero, and V is a $p_i \times p_i$ orthonormal matrix [3]. Thus

$$\begin{aligned}
u^T X_i X_i^T u &= U\Sigma V^T V \Sigma^T U^T\\
&= U^2 \cos^2\theta_i\\
&= \cos^2\theta_i && (9)
\end{aligned}$$

Substituting Eq. (8) in as the new cost function transforms the optimization problem into

$$\begin{aligned}
\big[u^{(j)}\big] &:= \operatorname*{arg\,max}_{[u]\in \mathrm{Gr}(n,1)} u^T\left(\sum_{i=1}^{N} X_i X_i^T\right)u \\
&\text{subject to} \quad [u] \perp \big[u^{(l)}\big] \quad \text{for } l < j.
\end{aligned} \qquad (10)$$

Define $\mathbf{A} = \sum_{i=1}^{N} X_i X_i^T$. Remember that $u^{(1)}$ is a unit length vector whose span is $[u^{(1)}]$. To find such a $u^{(1)}$, we consider the Lagrangian

$$L(u, \lambda) = u^T \mathbf{A} u - \lambda\bigl(u^T u - 1\bigr). \tag{11}$$

The partial derivatives of $L(u, \lambda)$ lead to the first order necessary conditions for optimality that are satisfied when

$$\begin{aligned} \mathbf{A}u &= \lambda u \quad \text{and} \\ u^T u &= 1. \end{aligned} \tag{12}$$

Thus we solve the eigenvector problem, $\mathbf{A}u = \lambda u$, and the cost function is maximized when u is the eigenvector associated with the largest eigenvalue of A. We set $u^{(1)}$ equal to this eigenvector. Once $u^{(1)}, \ldots, u^{(j-1)}$ has been found, we can find $u^{(j)}$ by considering Lagrangians that incorporate the condition that $u^{(j)}$ has unit length and incorporate orthogonality constraints in relation to $u^{(l)}$ for $l < j$. This leads to a Lagrangian of the form

$$L(u, \lambda, \lambda_1, \ldots, \lambda_{j-1}) = u^T \mathbf{A} u - \lambda\bigl(u^T u - 1\bigr) - \sum_{l=1}^{j-1} \lambda_l\bigl(u^T u^{(l)} - 0\bigr) \tag{13}$$

The partial derivatives of $L(u, \lambda, \lambda_1, \ldots, \lambda_{j-1})$ lead to the first order necessary conditions for optimality that are satisfied when

$$\begin{aligned} \mathbf{A}u &= \lambda u \quad \text{and} \\ u^T u &= 1 \quad \text{and} \\ u^T u^{(l)} &= 0 \quad \text{for } l = 1 \ldots j-1. \end{aligned} \tag{14}$$

In other words, we seek an eigenvector of $\mathbf{A}$ that is orthogonal to previously found eigenvectors of $\mathbf{A}$. Since $\mathbf{A}$ is a real, symmetric, positive semi-definite matrix, there are $r = \operatorname{rank}(\mathbf{A})$ mutually orthogonal eigenvectors associated with positive eigenvalues. If these eigenvectors are ordered by their associated eigenvalues in descending order, the resulting sequence is the set of sequential optimizers of Eq. (10), $\{[u^{(1)}], \ldots, [u^{(r)}]\}$.

4. Solution via the SVD

Finding the r mutually orthogonal eigenvectors of $\mathbf{A} = \sum_{i=1}^{N} X_i X_i^T$ can be completed in $O(n^3)$ flops with standard eigenvector solvers. In this section, we describe how computations can be carried out more efficiently in cases involving a relatively small number of very tall matrices.

If $P = \sum_{i=1}^{N} \dim([X_i])$ (with X_i an $n \times \dim([X_i])$ matrix with orthonormal columns whose column space is $[X_i]$) and if

$$\mathbf{X} = [X_1, X_2, \dots, X_N], \tag{15}$$

then $\mathbf{X} \in \mathbb{R}^{n \times P}$. Note that $\mathbf{X}\mathbf{X}^T = \sum_{i=1}^{N} X_i X_i^T = \mathbf{A}$ and that $\operatorname{rank}(\mathbf{X}) = \operatorname{rank}(\mathbf{A}) = r$. If the Singular Value Decomposition (SVD) of $\mathbf{X}$ is

$$\mathbf{X} = U\Sigma V^T \tag{16}$$

then

$$\mathbf{X}\mathbf{X}^T = U\Sigma\Sigma^T U^T. \tag{17}$$

Thus the columns of U are the eigenvectors of $\mathbf{X}\mathbf{X}^T$ and the first r left singular vectors of $\mathbf{X}$ are exactly the solutions to the optimization problem in Eq. (10), $\{[u^{(1)}], \dots, [u^{(r)}]\}$. By solving for the singular value decomposition of $\mathbf{X}$ instead of an eigenvalue decomposition of $\mathbf{A}$ the complexity changes to $O(nP^2)$ flops (which is less than $O(n^3)$ when $P < n$).

In a related problem, we can find the eigenvectors of the matrix

$$\mathbf{X}^T\mathbf{X} = V\Sigma^T\Sigma V^T, \tag{18}$$

as the columns of V. These right singular vectors of $\mathbf{X}$ are related to Multiset Canonical Correlation Analysis (MCCA) (see Section 6.3).

5. The flag mean

Let $(q_1, q_2, \dots, q_M)$ be an ordered set of integers such that $q_1 < q_2 < \cdots < q_M$. A flag in $\mathbb{R}^n$ of type $(q_1, q_2, \dots, q_M)$ is a nested sequence of subspaces $S_1 \subset S_2 \subset \cdots \subset S_M$ such that $\dim(S_i) = q_i$. The flag manifold, $\mathrm{FL}(n; q_1, q_2, \dots, q_M)$, is a manifold whose points correspond to the set of all flags of type $(q_1, q_2, \dots, q_M)$. Note that the flag manifold $\mathrm{FL}(n; q_1)$ is equivalent to $\mathrm{Gr}(n, q_1)$ and the points on either correspond to the q_1-dimensional subspaces of $\mathbb{R}^n$. For more details about the geometry of flag manifolds, refer to [13].

Let $\mathcal{D} = \{[X_i]\}_{i=1}^{N}$ be a finite collection of subspaces of $\mathbb{R}^n$ and let $r = \dim(\operatorname{span}(\bigcup_{i=1}^{N}[X_i]))$. The flag mean of $\mathcal{D}$, denoted $[\![\mu_{pF}]\!]$, is (typically) a point on the flag manifold $\mathrm{FL}(n; 1, 2, \dots, r)$ (the notation $[\![\mu_{pF}]\!]$ comes from the use of the projection F-norm in the cost function). Each subspace in the flag acts as an “average” for $\{[X_i]\}_{i=1}^{N}$. The flag is built from the 1-dimensional subspaces, $\{[u^{(1)}], \dots, [u^{(r)}]\}$, arising as the left singular vectors of the matrix, $\mathbf{X}$, built from $\mathcal{D}$. If the non-zero singular values are all distinct, then the flag is:

$$[\![\mu_{pF}]\!](\mathcal{D}) := \left[u^{(1)}\right] \subset \left[u^{(1)}\middle|u^{(2)}\right] \subset \cdots \subset \left[u^{(1)}\middle|\dots\middle|u^{(r)}\right] \tag{19}$$

and we get a point on $\mathrm{FL}(n; 1, 2, \dots, r)$. Algorithm 1 describes this formulation explicitly from the eigenvector decomposition of $\mathbf{A} = \sum_{i=1}^{N} X_i X_i^T$. We note that since $\mathbf{A}$ is a symmetric positive semi-definite matrix, these r eigenvectors form an orthonormal set and the associated eigenvalues are positive.

Algorithm 1 Calculate the flag mean, $[\![\mu_{pF}]\!]([X_1], \dots, [X_N])$

```
Ensure: $X_i^T X_i = I$ for $i = 1, \dots, N$
  Let $\mathbf{A} = \sum_{i=1}^{N} X_i X_i^T$
  Let $r = \dim(\mathrm{span}(\cup_{i=1}^{N} [X_i]))$
  Find $u^{(1)}, \dots, u^{(r)}$ as the eigenvectors of $\mathbf{A}$ ordered based on their associated eigenvalues from largest to smallest
  Let $[\![\mu_{pF}]\!] = \{[u^{(1)}], [u^{(1)}|u^{(2)}], \dots, [u^{(1)}| \dots |u^{(r)}]\}$
```

When $P < n$, computing the thin SVD of $\mathbf{X}$ is faster than computing the eigenvectors of $\mathbf{A}$. It is possible that one of the first r singular values has a multiplicity greater than 1. In this case, the singular vectors associated with that singular value have equal weight and they should be treated as a single subspace with dimension equal to the multiplicity of the singular value (rather than as separate 1-dimensional subspaces). For example, if the second singular value has multiplicity 2, the flag mean would be defined as

$$[\![\mu_{pF}]\!] = \left[u^{(1)}\right] \subset \left[u^{(1)} \middle| u^{(2)} \middle| u^{(3)}\right] \subset \cdots \subset \left[u^{(1)} \middle| \dots \middle| u^{(r)}\right]. \tag{20}$$

In this case, the flag would not contain a 2-dimensional subspace and determines a point on $\mathrm{FL}(n; 1, 3, 4, \dots, r)$.

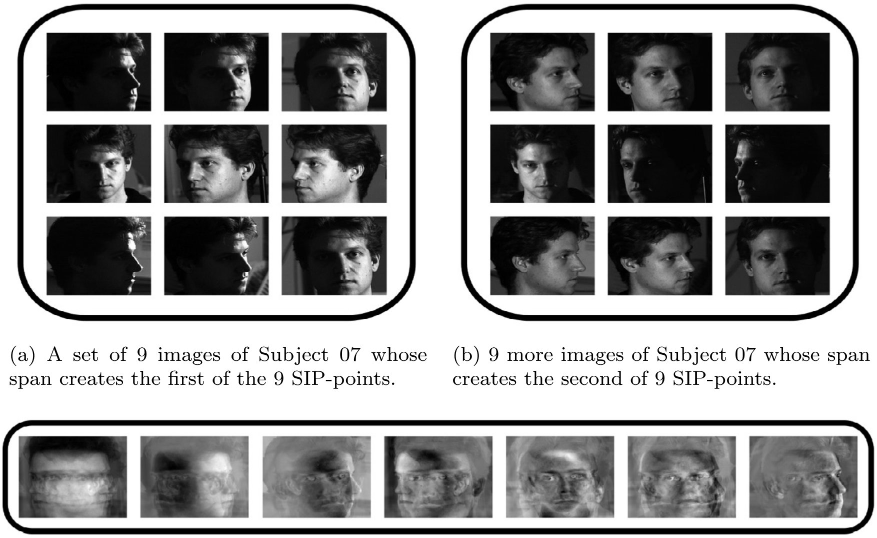
5.1. Illustrations of the flag mean

The following examples are built from a collection of images from the Carnegie-Mellon University Pose, Illumination, and Expression (CMU-PIE) database [16]. The database contains black and white images of 68 subjects from the shoulders up. The images include all combinations of 13 poses, 42 lighting conditions, and 4 expressions for each subject. The images have been registered and cropped, with a resulting resolution of 277 × 299 pixels. Each image can be considered as a 277 × 299 matrix with entries between 0 and 255 thus as a vector in $\mathbb{R}^{277 \times 299} \simeq \mathbb{R}^{82\,823}$.

5.1.1. Pose and illumination subspaces

From the CMU-PIE database, image subspaces can be created in a variety of ways. For example, consider a subset of the images that consists of 9 illumination conditions, 9 poses, and 1 expression for a single subject. Such a subset contains 81 unique images of a single subject. In the following two examples, we partition the 81 images into 9 groups of 9 images in two different ways.

In the first example, we determine an orthonormal basis for the 9 images of the subject that contain all 9 poses in the subset under a single illumination condition. This defines



(a) A set of 9 images of Subject 07 whose span creates the first of the 9 SIP-points.

(b) 9 more images of Subject 07 whose span creates the second of 9 SIP-points.

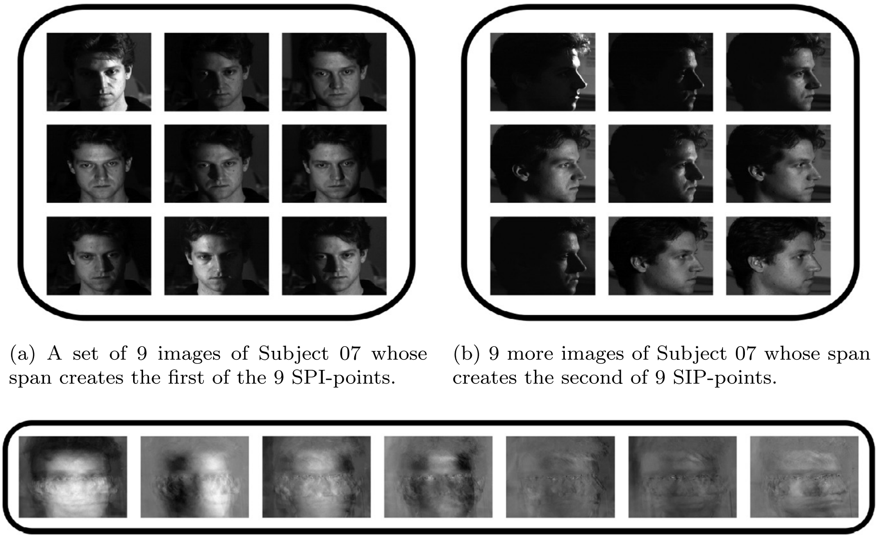
(c) The first 7 images in the flag mean computed from 9 SIP-points of Subject 07. The first image shows the average pose while each of the remaining images appears to be the superposition of a small number of the 9 poses, forming a basis for pose.

Fig. 1. Organizing the PIE images into groups where subject and illumination are fixed while pose varies creates what we refer to as SIP-points. The flag mean of these SIP-points appears to capture the common pose information.

a subspace that we will refer to as a Subject-Illumination-Pose-point, or an SIP-point. It is possible to create 9 distinct SIP-points from the subset of 81 images. The images used to create two of these SIP-points for Subject 07 are displayed in Figs. 1a and 1b. Thus, from the 81 images, we create 9 SIP-points, one for each illumination condition. The SIP-points are 9-dimensional subspaces of pixel space and correspond to points on $\mathrm{Gr}(82\,823, 9)$. The first 7 vectors in the flag mean of 9 SIP-points of Subject 07 are displayed in Fig. 1c. With the exception of the first vector, it appears as though the vectors in the flag mean each approximate a small number of poses of Subject 07. This seems reasonable, because the commonality between the SIP-points used to create the flag mean of Subject 07 is that they contain all 9 poses from the subset.

In the second example, we determine an orthonormal basis for the 9 images of the subject that contain all 9 illumination conditions in the subset under a single pose. Subspaces that have a single subject, a single pose, and a range of illumination conditions will be referred to as Subject-Pose-Illumination-points or SPI-points. The images used to create two of these SPI-points for Subject 07 are displayed in Figs. 2a and 2b.

As before, we create 9 SPI-points, one for each pose. The first 7 vectors in the flag mean of the 9 SPI-points of Subject 07 are displayed in Fig. 2c. This time each vector in the flag appears to approximate a single illumination condition of Subject 07, while individual poses are not discernible. It is important to note that the flags displayed in Fig. 1c and Fig. 2c are both created using the same set of 81 images. The differences in the associated flags are a result of how the images were organized into subspaces. For the



(a) A set of 9 images of Subject 07 whose span creates the first of the 9 SPI-points.

(b) 9 more images of Subject 07 whose span creates the second of 9 SIP-points.

(c) The first 7 images in the flag mean computed from 9 SPI-points of Subject 07. The first image shows the average (neutral) illumination while the next three images appear to contain the other the illumination information present in the set, forming a basis for illumination.

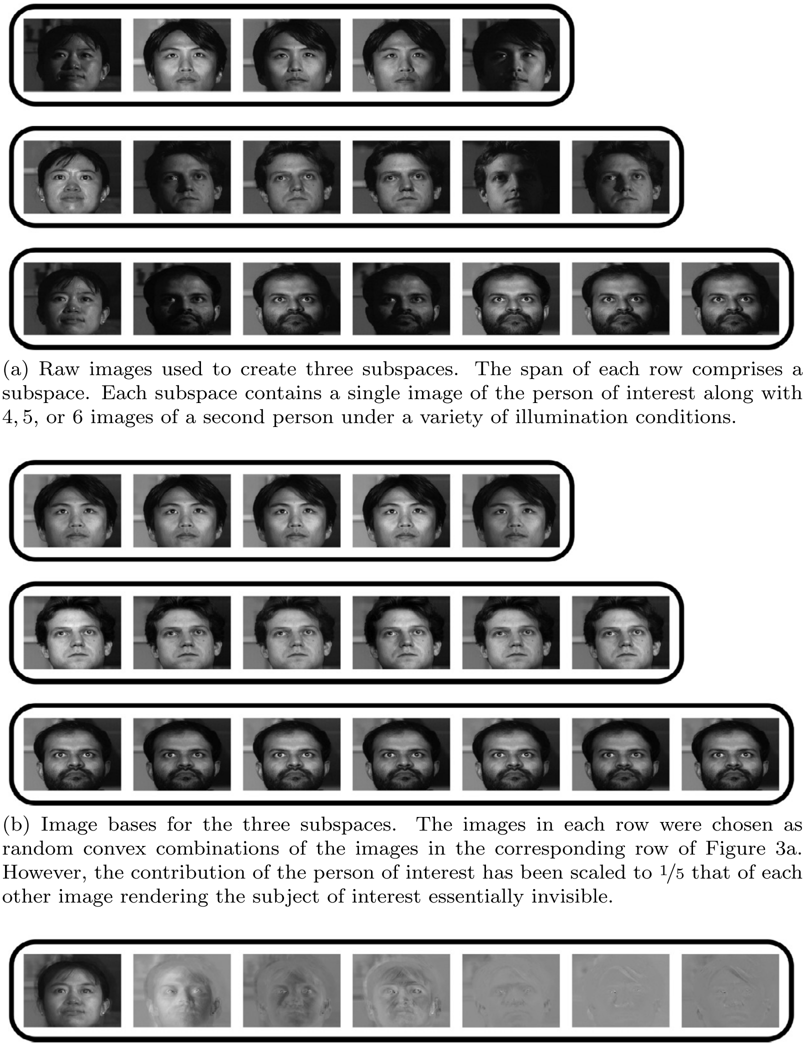
Fig. 2. The PIE images here have been grouped on subject and pose, with illumination running free within a group. The span of one such group forms what we refer to as an SPI-point. The flag mean of these SPI-points appears to capture the illumination information that is common amongst the set of points.

SIP-points, the commonality is that each subspace contains all 9 of the poses, and this appears to be represented in the images that make up the flag. For the SPI-points, the shared attribute is that each subspace contains all 9 illumination conditions, and again this similarity seems to “shine” through in the flag representation.

5.1.2. Hidden signal

The previous illustrations are visually interesting, but there may be too many forms of variation present for us to build a firm understanding of the utility of the flag representation. In this example, we will see how the flag mean can reveal a common attribute from within subspaces of differing dimensions when the common attribute is weakly represented.

One practical issue that the flag mean addresses is how to find an average representation for subspaces of differing dimensions. As we will see in Section 6, existing methods for representing subspaces require that all points live on a single Grassmann manifold. When this is not the case, points are typically up-projected to the size of the largest subspace by finding the closest orthonormal matrix, or down-projected to the size of the smallest subspace using the left singular vectors from a truncated SVD. The problem with these methods is that up-projecting is not unique, and down-projecting discards potentially useful information. The flag mean sidesteps these issues by allowing subspaces of differing dimensions. Furthermore, by encoding the representation as a flag, rather than a subspace, more subtle information can be expressed.



(a) Raw images used to create three subspaces. The span of each row comprises a subspace. Each subspace contains a single image of the person of interest along with 4, 5, or 6 images of a second person under a variety of illumination conditions.

(b) Image bases for the three subspaces. The images in each row were chosen as random convex combinations of the images in the corresponding row of Figure 3a. However, the contribution of the person of interest has been scaled to 1/5 that of each other image rendering the subject of interest essentially invisible.

(c) Images from the flag mean computed for the three subspaces in Figure 3b. The first and fourth images appear to exclusively contain information about the subject of interest, who is common to all three spaces. The second, third, and higher order images contain the info about the subjects that only appear in one subspace.

Fig. 3. In the hidden signal experiment, images of a person of interest, Subject 1 from the CMU-PIE dataset, are mixed into sets of images of other people. Computing the flag mean of the three sets recovers information pertaining to the person of interest because her picture is a common feature in the three sets.

To illustrate these two properties, we look at a different collection of images from the CMU-PIE database and use the images to create three subspaces of $\mathbb{R}^{82\,823}$ with differing dimensions. We start with 3 subjects from the database photographed in the frontal pose. The "weak signal" will be the woman shown in the first image in each row of Fig. 3a. Each of the subspaces consists of the span of one image of the woman and $4, 5$, or 6 frontal images of other subjects. Each image has a unique illumination condition. The resulting subspaces determine points on $\mathrm{Gr}(82\,823, 5)$, $\mathrm{Gr}(82\,823, 6)$, and $\mathrm{Gr}(82\,823, 7)$, and have the shared property that they all contain a one-dimensional space representing the subject of interest.

Images of random vectors taken from these three spaces can be seen in Fig. 3b. Each row in the figure represents a single subspace. In order to emphasize that the flag mean finds the strongest *common* signal, rather than the strongest signal overall, the contribution of the person of interest has been scaled to 1/5 that of each other images in the subspaces. That is, each picture in Fig. 3b is created as a linear combination of the pictures in Fig. 3a with the caveat that the coefficient on the subject of interest is also multiplied by 1/5. Thus, the subject of interest cannot easily be seen in any of these pictures. These spanning sets of images are then used to find an orthonormal basis for each subspace, and the flag mean of the three spaces is computed.

The first 7 images from the flag mean of these subspaces can be seen in Fig. 3c. The first vector, which determines the best 1-dimensional representation of the subspaces according to the flag mean, looks distinctly like the person of interest. The subsequent images contain details that correspond to the other people present in the subspaces. This experiment cannot be recreated with other existing subspace means because the other mean representations cannot accommodate subspaces of differing dimensions. Additionally, note that this result is quite distinct from what we would see if we performed standard Eigenfaces on the entire collection of images [9]. With an Eigenfaces approach we would see the first image as a generic average of all the people present, along with finer details in the subsequent images.

The utility of the flag mean in this example is two-fold. First, its flexibility allows an average to be computed for subspaces of differing dimensions. Second, the flag structure provides a more detailed breakdown of the information shared between the subspaces. If the task at hand was to identify which person was common in the three sets, we could measure the similarity between a probe image to each of the four subjects used in the first subspaces within the flag. Even naive methods would recognize that the first flag image contains information about the person of interest. On the other hand, if we had up or down-projected the subspaces to live on a single Grassmann manifold and computed one of the existing subspace averages, the result would be a space spanning 5 or 7 dimensions (depending on our projection). Comparing images of our four subjects to this object would merely confirm that each person was present in the average. Thus the order of the subspaces within the flag has the ability to reveal the strongest common signal, as desired.

Further elements of the flag beyond the first 1-dimensional subspace may also be of interest. For instance, a related subspace average, the extrinsic manifold mean, is interested in the span of the first p elements of the flag [17]. This connection will be made explicit in Section 6.1.

5.2. Weighted flag mean

The ability to compute a weighted flag mean offers advantages under certain circumstances. For instance, if we have a variable level of confidence in the given subspaces, we may opt to scale the contribution of each subspace according to reliability. Alternatively, we could weight the contribution of each subspace according to the dimension of the subspace, perhaps to increase the impact of larger subspaces.

The flag mean allows for weighted mean calculations. Let $\{c_i\}_{i=1}^N$ be weights that are selected to correspond to the N subspaces. Define $\mathbf{B} = \sum_{i=1}^N c_i X_i X_i^T$ and compute the eigenvectors of $\mathbf{B}$. The weighted flag mean is then the nested sequence of subspaces created from these eigenvectors. Within the flag, the subspace of dimension q is the span of the q eigenvectors corresponding to the q largest eigenvalues. The computation of the weighted flag mean, $[\![\nu_{pF}]\!]$, is the same as in Algorithm 1 except we replace $\mathbf{B}$ for $\mathbf{A}$. If we let $\mathbf{Y} = [\sqrt{c_1}X_1, \sqrt{c_2}X_2, \dots, \sqrt{c_N}X_N]$ then we have $\mathbf{B} = \mathbf{Y}\mathbf{Y}^T$. Therefore, computational tractability for large n can be maintained using the thin SVD of $\mathbf{Y}$.

6. Related work

One main benefit of the flag mean, $[\![\mu_{pF}]\!]$, is that it can be computed for collections of subspaces of $\mathbb{R}^n$ with differing dimensions. However, if the subspaces all have the same dimension, p, then the p-dimensional subspace appearing in the flag, $[\![\mu_{pF}]\!]$, can be compared directly to the commonly used Karcher mean, and the lesser known (but quite useful) extrinsic manifold mean of Srivastava and Klassen [17].

6.1. A generalization of the extrinsic manifold mean

Let $\{[X_1], [X_2], \dots, [X_N]\}$ be a finite collection of points on $\mathrm{Gr}(n,p)$. Srivastava and Klassen define the *extrinsic manifold mean*, $[\mu_E]$, as

$$[\mu_E] = \operatorname*{arg\,min}_{[\mu]\in \mathrm{Gr}(n,p)} \frac{1}{N}\sum_{i=1}^N d_{pF}\big([\mu],[X_i]\big)^2, \tag{21}$$

where $d_{pF}([\mu],[X_i])$ is again the metric on $\mathrm{Gr}(n,p)$ derived from the projection Frobenius norm. In Section 5, the flag mean was defined as

$$[\![\mu_{pF}]\!] = \big\{\big[u^{(1)}\big], \big[u^{(1)}\big|u^{(2)}\big], \dots, \big[u^{(1)}\big|\dots\big|u^{(r)}\big]\big\}, \tag{22}$$

where $u^{(j)}$ was the eigenvector of $\sum_{i=1}^N X_i X_i^T$ corresponding to the jth largest eigenvalue. Following the same line of reasoning as in Section 3 the following relationship can be observed,

$$[\mu_E] = \operatorname*{arg\,min}_{[\mu]} \frac{1}{N}\sum_{i=1}^N d_{pF}\big([\mu],[X_i]\big)^2 \tag{23}$$

$$= \arg\max_{[\mu]} \frac{1}{N} \sum_{j=1}^{p} u^{(j)T} \left(\sum_{i=1}^{N} X_i X_i^T \right) u^{(j)} \tag{24}$$

$$= \left[u^{(1)} \middle| u^{(2)} \middle| \dots \middle| u^{(p)} \right] \tag{25}$$

$$= \text{the } p\text{-dimensional subspace in the flag } [\![\mu_{pF}]\!]. \tag{26}$$

In the restricted case where $[X_i] \in \mathrm{Gr}(n,p)$ for $1 \leqslant i \leqslant N$, the p-dimensional subspace of the flag $[\![\mu_{pF}]\!]$ is equal to the extrinsic manifold mean.

Sarlette and Sepulchre have described a generalization of the extrinsic manifold mean to data lying on a class of manifolds broader than the Grassmann manifold [15]. Their *induced arithmetic mean* is defined to be the set of points that globally minimize the weighted sum of squared Euclidean distances in $\mathbb{R}^m$ to each of the points in question. It requires the assumption that the points come from a connected compact homogeneous manifold, $\mathcal{M}$, smoothly embedded in $\mathbb{R}^m$, such that the Euclidean norm is constant over the points of $\mathcal{M}$; and that the Lie group $\mathcal{G}$ acts as a subgroup of the orthogonal group on $\mathbb{R}^m$. The Grassmann manifold satisfies these assumptions and, in this setting, the induced arithmetic mean is the same as the extrinsic manifold mean. When restricted to $SO(3)$, the induced arithmetic mean is equivalent to the *projected arithmetic mean* defined by Moakher [11].

6.2. Relationship to the Karcher mean

The *Karcher mean* is a commonly used representation for a collection of data on a Grassmann manifold. It can be viewed as a center of mass for a point cloud [7]. The sample Karcher mean is defined as

$$[\mu_K] = \arg\min_{[\mu]} \frac{1}{N} \sum_{i=1}^{N} d\big([\mu],[X_i]\big)^2, \tag{27}$$

where $d([\mu],[X_i])$ is the geodesic distance based on arc length. The geodesic distance based on arc length is computed as the ℓ_2-norm of the vector of principal angles between $[\mu]$ and $[X_i]$, or $d([\mu],[X_i]) = \|\Theta_i\|_2$. It is shown in [5] that this measure of distance is the canonical metric on the Grassmann manifold in the sense that it is equivalent to the Euclidean metric in the tangent space of a single point on a Grassmannian. The Karcher mean finds the point that minimizes the sum of the squared distances between itself and the data (with respect to the geodesic distance), and can be viewed as an analogue of the arithmetic mean in a Euclidean space.

The equivalence between the dominant p-dimensional subspace of $[\![\mu_{pF}]\!]$ and $[\mu_E]$, and the definition of $[\mu_K]$ reveal that the flag mean optimizes a similar cost function to that optimized by the Karcher mean. The difference in the formulation is that $[\![\mu_{pF}]\!]$ and $[\mu_E]$ use a metric derived from the projection F-norm to measure distance, while $[\mu_K]$ uses the canonical metric on $\mathrm{Gr}(n,p)$. These two metrics are asymptotically equivalent for small distances. In general, for $[X] \neq [Y]$ we have $d([X],[Y]) > d_{pF}([X],[Y])$ [5]. Additionally, both metrics are orthogonally invariant as they are based on principal angles. As a result, distances will be independent of the choice of orthonormal representative for a point in $\mathrm{Gr}(n,p)$. For tightly clustered sets of points, $d([X],[Y]) \approx d_{pF}([X],[Y])$, and hence $[\mu_K] \approx [\mu_E]$.

6.3. Right singular vectors of $\mathbf{X}$*: MCCA*

As mentioned in Section 4, there is a connection between the flag mean and multiset extensions to Canonical Correlation Analysis (CCA). Multiset Canonical Correlation Analysis (MCCA) has been considered by Kettenring, and more recently by Via et al. [21,8]. These works examine data matrices, $D_1, \ldots, D_N$, where matrix D_i represents a data set of size $n \times p_i$.

Let $P = \sum_{i=1}^{N} p_i$ and suppose that $X_1, \ldots, X_N$ are orthonormal bases for the data matrices. Using the general multiset formulation, MCCA seeks canonical vectors $k_i \in [X_i]$ in order to solve

$$\begin{aligned} &\underset{k_1,\ldots,k_N}{\arg\max} \quad \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} k_i^T k_j \\ &\text{subject to} \quad \sum_{i=1}^{N} k_i^T k_i = 1, \quad k_i \in [X_i]. \end{aligned} \tag{28}$$

Expressing k_i as a linear combination, we write $k_i = X_i \alpha_i$ for $\alpha_i \in \mathbb{R}^{p_i}$. Next we form the Lagrangian

$$L(\alpha,\lambda) = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j\neq i}}^{N} \alpha_i^T X_i^T X_j \alpha_j - \lambda\left(\sum_{i=1}^{N} \alpha_i^T \alpha_i - 1\right) \tag{29}$$

Recall that $\mathbf{X} = [X_1, \ldots, X_N]$. If we define the block column vector $\alpha \in \mathbb{R}^P$ as $\alpha = [\alpha_1^T, \ldots, \alpha_N^T]^T$, then Eq. (29) takes on the form

$$L(\alpha,\lambda) = \alpha^T\big(\mathbf{X}^T\mathbf{X} - I\big)\alpha - \lambda\big(\alpha^T\alpha - 1\big). \tag{30}$$

Finding the extrema of Eq. (30) via differentiation requires

$$\mathbf{X}^T\mathbf{X}\alpha = (\lambda + 1)\alpha. \tag{31}$$

Thus the set of canonical components of multiset CCA, $k_i = X_i \alpha_i$, is obtained by finding the eigenvectors of $\mathbf{X}^T\mathbf{X}$ via Eq. (31).

It can be shown that the above argument extends to the jth set of canonical vectors, $\{k_i^{(j)}\}_{i=1}^N$. Specifically, we compute

$$k_i^{(j)} = X_i \alpha_i^{(j)} \tag{32}$$

where $\alpha^{(j)} = [\alpha_1^{(j)T}, \ldots, \alpha_N^{(j)T}]^T$ is the eigenvector of $\mathbf{X}^T\mathbf{X}$ associated with the jth largest eigenvalue.

As mentioned in Section 4, computing the right singular vectors of $\mathbf{X}$ is equivalent to computing the eigenvectors of $\mathbf{X}^T\mathbf{X}$. If $\mathbf{X} = \mathcal{U}\Sigma\mathcal{V}$ is the singular value decomposition of $\mathbf{X}$, then $\mathcal{V} = [v^{(1)}|\ldots|v^{(P)}] = [\alpha^{(1)}|\ldots|\alpha^{(P)}]$. Thus the columns of $\mathcal{V}$ can be used to compute the canonical vectors of MCCA.

Computing $k^{(j)}$ for $j = 1, \ldots, P$ using the right singular vectors of $\mathbf{X}$ presents an interesting connection between MCCA and the set of subspaces $\{[u^{(1)}], \ldots, [u^{(r)}]\}$. In particular,

$$u^{(j)} = \frac{\mathbf{X}v^{(j)}}{\sigma^{(j)}} \tag{33}$$

$$= \frac{1}{\sigma^{(j)}} \sum_{i=1}^N X_i v_i^{(j)} \tag{34}$$

$$= \frac{1}{\sigma^{(j)}} \sum_{i=1}^N k_i^{(j)}. \tag{35}$$

In other words, the 1-dimensional subspaces that are used to construct the flag mean in Section 5 are the average of the corresponding canonical vectors in multiset CCA, scaled to unit length. The utility of this connection is that if we compute the singular value decomposition of $\mathbf{X}$, we find both a collection of vectors in the span of $\mathbf{X}$ that represent its constituent subspaces (the flag mean), as well as a collection of vectors that are restricted to live within each X_i for $i = 1, \ldots, N$ that are conditioned on each of the other subspaces (the canonical components).

6.4. Further connection to CCA

Although the left singular vectors of $\mathbf{X}$ did not appear to be of central interest to the authors of the multiset CCA paper [21], they were of interest to Kettenring [8]. He was interested in finding a vector z that solves an optimization problem that looks very different from the ones in Eqs. (2) and (10), but has a similar goal. The solution to each of these three optimization problems is a vector in the middle of a collection of data matrices. Kettenring shows that the multiset CCA formulation produces the solution as $z = \frac{\mathbf{D}\alpha}{N}$ with $\mathbf{D} = [D_1, D_2, \ldots, D_N]$ the array of data matrices. Thus, if one were to use orthonormal bases, X_i, in place of the general data matrices D_i, Kettenring's method for finding the central vector z would be the same as the computation in Eq. (33) up to scaling.

7. Discussion and future work

We have described an algorithm for constructing a flag mean representation for a finite collection of subspaces of a fixed vector space. The subspaces are allowed to have differing dimensions. The algorithm, based on an optimization criterion derived from the projection Frobenius norm, allows comparisons between collections of subspaces at multiple levels. The flag mean representation has connections with the extrinsic manifold mean but differs in allowing for the input data to be non-equidimensional and differs in the form and interpretation of the output. For well clustered and equidimensional sets of subspaces, both the extrinsic manifold mean and one of the components of the flag mean recover a good approximation for the well known Karcher mean. The flag mean representative for a collection of subspaces was shown to have a strong connection to the set of left singular vectors of an associated data matrix $\mathbf{X}$.

By representing our average as a flag, rather than a point on a specific Grassmannian, we get an ordering from the average representative. This structure can be useful when classifying data with multiple semantically meaningful labels. As a consequence, the flag mean is applicable in settings requiring one to classify data of different dimensions, to classify data with multiple labels, to identify (without supervision) which forms of variation cause classification to fail, and to organize multi-label data sets without supervision. To achieve these tasks further work will be needed to identify useful metrics and similarity scores for flag manifolds.

Acknowledgements

The authors would like to thank the referee for a very careful reading of the original manuscript. The referee caught several errors and made many concrete suggestions that, we feel, improved the exposition and motivation for the paper.

This research was partially supported by awards DMS-1228308 and DMS-1322508 from the National Science Foundation as well as by DARPA (Space and Naval Warfare Systems Center Pacific) under award No. N66001-11-1-4184. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or DARPA.

References

[1] E. Begelfor, M. Werman, Affine invariance revisited, in: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, CVPR '06, IEEE Computer Society, Washington, DC, USA, 2006, pp. 2087–2094.

[2] P.N. Belhumeur, D.J. Kriegman, What is the set of images of an object under all possible illumination conditions?, Int. J. Comput. Vis. 28 (3) (1998) 245–260.

[3] A. Björck, G.H. Golub, Numerical methods for computing angles between linear subspaces, Math. Comp. 27 (1973) 579–594.

[4] J.M. Chang, M. Kirby, H. Kley, C. Peterson, B. Draper, J.R. Beveridge, Recognition of digital images of the human face at ultra low resolution via illumination spaces, in: Proceedings of the 8th Asian Conference on Computer Vision-Volume, Part II, Springer-Verlag, 2007, pp. 733–743.

[5] A. Edelman, T. Arias, S.T. Smith, The geometry of algorithms with orthogonality constraints, SIAM J. Matrix Anal. Appl. 20 (1998) 303–353.

[6] J. Hamm, Subspace-based learning with Grassmann kernels, PhD thesis, University of Pennsylvania, 2008.

[7] H. Karcher, Riemannian center of mass and mollifier smoothing, Comm. Pure Appl. Math. 30 (5) (1977) 509–541.

[8] J.R. Kettenring, Canonical analysis of several sets of variables, Biometrika 58 (3) (1971) 433–451.

[9] Michael Kirby, Lawrence Sirovich, Application of the Karhunen–Loeve procedure for the characterization of human faces, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1) (1990) 103–108.

[10] Y.M. Lui, J.R. Beveridge, M. Kirby, Action classification on product manifolds, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 2010, pp. 833–839.

[11] Maher Moakher, Means and averaging in the group of rotations, SIAM J. Matrix Anal. Appl. 24 (1) (2002) 1–16.

[12] B. Mondal, S. Dutta, R.W. Heath, Quantization on the Grassmann manifold, IEEE Trans. Signal Process. 55 (8) (2007) 4208–4216.

[13] D. Monk, The geometry of flag manifolds, Proc. Lond. Math. Soc. 3 (2) (1959) 253–286.

[14] V. Patrangenaru, K.V. Mardia, Affine shape analysis and image analysis, in: Proc. 22nd Leeds Ann. Statistics Research Workshop, July 2003.

[15] Alain Sarlette, Rodolphe Sepulchre, Consensus optimization on manifolds, SIAM J. Control Optim. 48 (1) (2009) 56–76.

[16] T. Sim, S. Baker, M. Bsat, The CMU pose, illumination, and expression (PIE) database, in: Proceedings of the 5th International Conference on Automatic Face and Gesture Recognition, 2002.

[17] A. Srivastava, E. Klassen, Monte Carlo extrinsic estimators of manifold-valued parameters, IEEE Trans. Signal Process. 50 (2) (2002) 299–308.

[18] A. Srivastava, E. Klassen, Bayesian and geometric subspace tracking, Adv. in Appl. Probab. 36 (1) (2004) 43–56.

[19] P. Turaga, A. Veeraraghavan, A. Srivastava, R. Chellappa, Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition, IEEE Trans. Pattern Anal. Mach. Intell. 33 (11) (2011) 2273–2286.

[20] O. Tuzel, F. Porikli, P. Meer, Human detection via classification on Riemannian manifolds, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2007, pp. 1–8.

[21] J. Via, I. Santamaria, J. Perez, A learning algorithm for adaptive canonical correlation analysis of several data sets, Neural Networks 20 (1) (2007) 139–152.