



Self-organizing mappings on the Grassmannian with applications to data analysis in high dimensions

Xiaofeng Ma¹ · Michael Kirby¹ · Chris Peterson¹ · Louis Scharf²

Received: 26 July 2018 / Accepted: 8 August 2019
© Springer-Verlag London Ltd., part of Springer Nature 2019

Abstract

We propose a method for extending Kohonen's self-organizing mapping to the geometric framework of the Grassmannian. The resulting algorithm serves as a prototype of the extension of the SOM to the setting of abstract manifolds. The ingredients required for this are a means to measure distance between two points, and a method to move one point in the direction of another. In practice, the data are not required to have a representation in Euclidean space. We discuss in detail how a point on a Grassmannian is moved in the direction of another along a geodesic path. We demonstrate the implementation of the algorithm on several illustrative data sets, hyperspectral images and gene expression data sets.

Keywords Grassmannian · Self-organizing mappings · Geodesic · Data visualization · Gene expression

1 Introduction

In the visualization of data, we often resort to the computation of means or centroids of data, computed from labeled or unlabeled training sets, or from nearest neighbors within a single data set. It is then natural to visualize these centroids, and the associated neighborhood data, using dimensionality reduction techniques such as self-organizing mappings (SOMs) [16–19]. This approach has proven to be a valuable tool for the low-dimensional visualization of data, see, e.g., [20, 27] in addition to an extended bibliography indicating the widespread applications of this methodology [14]. Specifically, in [19] Kohonen shows some of the most extensive applications of

SOM including management of massive textual databases and bioinformatics, the latter of which will also be explored under the context of pathway analysis in this paper. The key ingredient of this idea is that points that are neighbors in a high-dimensional space are represented as neighbors in a low-dimensional index space, a feature that arises through the self-organizing properties of the algorithm. When the data consist of a collection of k -dimensional subspaces of a common n -dimensional vector space, then the appropriate setting for analysis is the Grassmannian manifold, $\text{Gr}(k, n)$, whose points parameterize the k -dimensional subspaces of $\mathbb{R}^n$.

There is now considerable evidence that subspaces may serve as statistics for data analysis. The idea is that data may be organized into distinguishing subspaces. Data residing in the same subspace are invariant to confounding variations in the data, but data residing in separate subspaces are covariant to the variations that count for pattern recognition, or classification, see, e.g., [12, 23–25, 29]. The basic idea behind the approach is to compare an unlabeled observation to a collection of subspaces, each of which is constructed from several patterns of a given class. Each subspace represents a set of data which is labeled as similar, even though it may have been recorded in conditions undergoing a variable change of state. As an example, a vector space could be associated with a set of digital images of an object collected under a variety of different

✉ Michael Kirby
Michael.Kirby@Colostate.Edu

Xiaofeng Ma
Ma@math.colostate.edu

Chris Peterson
peterson@math.colostate.edu

Louis Scharf
Louis.Scharf@Colostate.Edu

¹ Department of Mathematics, Colorado State University, Fort Collins, CO, USA

² Department of Mathematics and Statistics, Colorado State University, Fort Collins, CO, USA

illumination conditions [2, 4, 5, 7]. It has been observed that the pattern set framework, based on subspaces, can enhance the robustness of pattern recognition algorithms. For example, in the case of using the subspace approach to capture variations in illumination conditions, the resolution of images can be reduced without sacrificing classification accuracy [6].

One attractive tool for set-to-set pattern analysis is the geometric framework of the Grassmannian. More formally, the Grassmannian $\text{Gr}(k, n)$ is a manifold whose points parameterize the k -dimensional subspaces of a fixed n -dimensional vector space. As such, it provides a setting for comparing distances between subspaces using a variety of metrics. The most widely used class of metrics are functions of the principal angles between the subspaces as these lead to comparisons that are invariant under orthogonal transformations. Further, it is possible to transition from one subspace to another along a shortest path, or geodesic (typically represented as a path consisting of orthonormal matrices). These ingredients make it possible to convert the SOM algorithm on vector spaces to an analogous algorithm on Grassmannians.

In this paper we develop the Grassmannian SOM algorithm by modifying the standard SOM algorithm to operate in the setting of the Grassmann framework. We use the resulting approach to visualize subspaces of data in high-dimensional spaces on low-dimensional index sets.

2 The mathematics of the Grassmannian

It is conventional to view a datum as a point in a high-dimensional ambient space such as Euclidean, and a collection of data as a set of points in this ambient space. Data from a common source, or of a common characteristic, are then viewed as data that reside in a common neighborhood of the ambient space. But how is neighborhood to be defined? In many applications a neighborhood consists of data that are composed of a linear combination of a common set of orthogonal basis vectors, organized into the k frame $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_k]$. Sometimes the k -frame itself is descriptive of the data, in which case we are interested in the set of all such k -frames. The manifold whose points parameterize the k -frames of an n -dimensional vector space, equipped with the standard inner product, is the Stiefel manifold $\text{St}(k, n)$. It may be represented as $O(n)/O(n-k)$, which codes for the idea that a sampling of the Stiefel manifold for a k -frame $\mathbf{V}$ is a sampling of the orthogonal group for an orthogonal matrix $\mathbf{Q}$, followed by a moding out of the basis for the orthogonal subspace spanned by the trailing $n-k$ orthogonal vectors in the orthogonal matrix $\mathbf{Q}$. The leading k orthogonal vectors of $\mathbf{Q}$ are then an ordered basis for the k -frame $\mathbf{V}$.

But in many applications the k -frame is too fine-grained. In other words, it is not the k -frame itself that matters, but rather the subspace of vectors in the ambient space that can be constructed from the frame. It is the range of $\mathbf{V}$ that counts, not $\mathbf{V}$ itself. This subspace we denote $\langle \mathbf{V} \rangle$. The manifold whose points parameterize all such subspaces is the Grassmann manifold $\text{Gr}(k, n)$ and we represent it as $O(n)/O(k) \times O(n-k)$, which codes for the idea that a sampling from this manifold for a subspace $\langle \mathbf{V} \rangle$ amounts to sampling from $O(n)$, followed by a moding out of the basis for the signal subspace spanned by the leading k orthogonal vectors and a moding out by the basis from the orthogonal subspace spanned by the trailing $n-k$ orthogonal vectors of $\mathbf{Q}$. What is left is a sampling of the Grassmann manifold for a k -dimensional subspace of $\mathbb{R}^n$. This discussion may be generalized to orthogonal groups, k -frames, and k -dimensional subspaces of arbitrary inner product spaces.

In some applications, one is interested in *uniform* samples from $O(n)$, $\text{St}(k, n)$, and $\text{Gr}(k, n)$. This may be achieved by constructing an $n \times n$ matrix of independent and identically distributed normal $[0, 1]$ random variables, and QR factoring it for $\mathbf{Q}$. The matrix $\mathbf{Q}$ obtained in this manner is drawn from the uniform distribution on the orthogonal group with respect to Haar measure, which is to say the distribution of $\mathbf{Q}$ is invariant to left orthogonal transformation. The corresponding samplings for $\mathbf{V}$ and $\langle \mathbf{V} \rangle$ are uniform with respect to Haar measure. It is not claimed that real data, organized into subspaces, produce uniformly distributed subspaces. Rather, it is claimed that the extent to which data and its subspaces produce *non-uniformly* distributed subspaces is the extent to which data bring information about the geometric structure of the underlying mechanisms that generated the data. This raises the question of distances and angles between subspaces of the Grassmannian. In particular, the distances between points on the Grassmannian are measured in terms of the principle angles between their subspaces.

2.1 Overview of angles

We provide a short summary of the computation of angles between subspaces initially described in [3]. Let X and Y be two vector subspaces of $\mathbb{R}^n$ such that

$$p = \dim(X) \geq \dim(Y) = q \geq 1,$$

then the *principal angles* $\theta_k \in [0, \frac{\pi}{2}]$, $1 \leq k \leq q$ between X and Y are defined recursively by

$$\cos(\theta_k) = \max_{u \in X} \max_{v \in Y} u^T v = u_k^T v_k, \quad k = 1 \dots q \quad (1)$$

subject to $\|u\| = \|v\| = 1$, $u^T u_i = 0$ and $v^T v_i = 0$ for $i = 1 \dots k-1$. Clearly, the principal angles satisfy

$0 \leq \theta_1 \leq \theta_2 \leq \dots \leq \theta_q \leq \frac{\pi}{2}$. Henceforth, $\theta = (\theta_1, \dots, \theta_q)$ will denote the principal angle vector. Note that we have abused notation somewhat in using X to represent both a subspace and an orthonormal matrix whose columns span this space. For additional details related to algorithms for the computation of principal angles see [3].

2.2 Metrics on the Grassmannian

Let A, B be two points on the Grassmannian $\text{Gr}(k, n)$. Again, we are thinking of these points as subspaces though they are represented by orthonormal matrices whose columns span the subspaces. The geodesic distance between these two points is given by

$$d_g(A, B) = \|(\theta_1, \dots, \theta_k)\|_2 \quad (2)$$

Other metrics are possible, e.g., the chordal distance

$$d_g(A, B) = \|(\sin(\theta_1), \dots, \sin(\theta_k))\|_2 \quad (3)$$

We note that it is possible to show that the Grassmannian may be isometrically embedded into Euclidean space when the chordal metric is employed and this is not the case for the geodesic metric [11]; see also [9].

Principal angles between subspaces are defined regardless of the dimensions of the subspaces, denoted, e.g., $\dim A$. Thus, inspired by the Riemannian geometry of the Grassmannian, we may define, for any vector subspaces A, B of $\mathbb{R}^n$ the *geodesic* distance

$$d_g(A, B) = \|(\theta_1, \dots, \theta_\ell)\|_2,$$

for any $\ell \leq \min\{\dim A, \dim B\}$. While d_ℓ is not, strictly speaking, a metric (for example, if $\dim A \cap B \geq \ell$, then $d_\ell(A, B) = 0$), it nevertheless provides an efficient and useful tool for analyzing configurations in $\cup_{k \geq \ell} \text{Gr}(k, n)$. The geometry driving these distance measures is captured by the notion of a Schubert variety $\bar{\Omega}_\ell(W) \subseteq \text{Gr}(k, n)$. Let W be a subspace of $\mathbb{R}^n$, then we define

$$\bar{\Omega}_\ell(W) = \{E \in \text{Gr}(k, n) \mid \dim(E \cap W) \geq \ell\}.$$

With this notation, $d_\ell(A, B)$ simply measures the distance between A and $\bar{\Omega}_\ell(B)$, i.e. $d(A, \bar{\Omega}_\ell(B)) = \min\{d_k(A, C) \mid C \in \bar{\Omega}_\ell(B)\}$ (it is worth noting that under this interpretation, $d_\ell(A, B) = d_\ell(B, A)$).

2.3 Geodesics

In this section we introduce the geodesic formula between two points on the Grassmann manifold proposed in [1]. Our presentation of this result is intended to be self-contained and contains additional details and examples not found in [1]. This section is intended to provide the background

necessary to facilitate the understanding of how this tool is integrated into the SOM algorithm.

Here it is convenient to view the Grassmannian $\text{Gr}(k, n)$ as the quotient manifold $O(n)/O(k) \times O(n-k)$. Let $Q \in O(n)$ be an n -by- n orthogonal matrix. The equivalence class $[Q]$ is the set of all orthogonal matrices whose first k columns span the same subspace as the one spanned by the first k columns of Q . A point on the Grassmann manifold is the equivalence class,

$$[Q] = \left\{ Q \begin{pmatrix} Q_k & 0 \\ 0 & Q_{n-k} \end{pmatrix} : Q_k \in O(k), Q_{n-k} \in O(n-k) \right\}$$

One advantage of this representation is that we may utilize the orthogonal group geodesic and the quotient geometry of the Grassmann manifold.

Suppose Q is an element in $O(n)$ (thus Q is an n -by- n orthogonal matrix). The tangent space to $O(n)$ at Q , denoted by $T_Q O(n)$, can be computed by considering curves on $O(n)$ which pass through Q . Let $X(t)$ be any smooth curve on $O(n)$ which goes through Q at $t = 0$, i.e. $X(0) = Q$. Since $X(t)^T X(t) = I$ for all t , differentiating both sides of the equation with respect to t yields $\dot{X}(t)^T X(t) + X(t)^T \dot{X}(t) = 0$. At $t = 0$, we have $\dot{X}(0)^T Q + Q^T \dot{X}(0) = 0$. By computing the dimension of $O(n)$ and the dimension of the set of matrices Δ such that $\Delta^T Q + Q^T \Delta = 0$, one verifies that $T_Q O(n)$ is exactly the set of matrices Δ where $Q^T \Delta$ is any n -by- n skew symmetric matrix. Therefore, the set of tangent vectors in $T_Q O(n)$ is the set of matrices Δ which has the form $\Delta = Q \mathbf{A}$ where $\mathbf{A}$ is any n -by- n skew symmetric matrix. It is further shown in [13] that a geodesic path on $O(n)$ is given by the exponential flow: $Q(t) = Q \exp(t \mathbf{A})$ where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a skew symmetric matrix and $Q(0) = Q$.

$\text{Gr}(k, n)$ is a quotient space of $O(n)$. The tangent space to $O(n)$ at Q , $T_Q O(n)$, can be decomposed into a vertical space V_Q and a horizontal space H_Q . The vertical space is the set of vectors in the tangent space corresponding to motions flowing along the equivalence class $[Q]$ at Q . The horizontal space is defined as the orthogonal (with respect to the Euclidean metric) complement of the vertical space in $T_Q O(n)$. Here the Euclidean metric is defined as a function $d : T_Q O(n) \times T_Q O(n) \rightarrow \mathbb{R}$:

$$\begin{aligned} d(U, V) &= \text{Tr}(U^T V) \\ &= \text{vec}(U)^T \text{vec}(V) \end{aligned}$$

Intuitively, the vectors in the vertical space can be thought of as the set of velocity vectors which preserve the equivalence class while the vectors in the horizontal space modify the equivalence class. Therefore, tangent vectors to geodesics must be restricted to the horizontal space. Moreover, if V is a tangent vector to $\text{Gr}(k, n)$ at $[Q]$, then

there is a horizontal vector $\bar{V} \in H_Q$ which represents V uniquely. Please note that $\text{Gr}(k, n)$ is an abstract manifold as is a tangent space to $\text{Gr}(k, n)$. The horizontal space provides us with a way to represent abstract tangent vectors with matrices. The idea of the vertical space and the horizontal space is illustrated in Fig. 1.

The computation of the vertical space at a point $Q \in O(n)$ is similar to the computation of the tangent space to $O(n)$. Let

$$V(t) = Q \begin{pmatrix} Q_k(t) & 0 \\ 0 & Q_{n-k}(t) \end{pmatrix}$$

be any smooth curve on the elements of the equivalence class $[Q]$ with $Q_k(0) = I_k$ and $Q_{n-k}(0) = I_{n-k}$, i.e. $V(0) = Q$. We observe that $V(t)^T V(t) = I$. Differentiating both sides and evaluating at $t = 0$ yields

$$\dot{Q}_k(0)^T + \dot{Q}_k(0) = 0$$

$$\dot{Q}_{n-k}(0)^T + \dot{Q}_{n-k}(0) = 0$$

Utilizing this computation along with a dimension count, we show that the vertical space at a point Q is the set of matrices

$$V_Q = \left\{ Q \begin{pmatrix} C & 0 \\ 0 & D \end{pmatrix} \right\},$$

where C is a k -by- k skew symmetric matrix and D is a $(n-k)$ -by- $(n-k)$ skew symmetric matrix. The horizontal space H_Q is the set of matrices which are orthogonal to the vertical space and living in $T_Q O(n)$. Consider the following set of equations

$$\begin{aligned} \text{Tr}(\Delta^T Q \begin{pmatrix} C & 0 \\ 0 & D \end{pmatrix}) &= 0 \\ \Delta &= Q\mathbf{A} \end{aligned}$$

where $\mathbf{A} \in \mathcal{R}^{n \times n}$, $C \in \mathcal{R}^{k \times k}$ and $D \in \mathcal{R}^{(n-k) \times (n-k)}$ are skew symmetric matrices. The solution set to the above system, i.e. the horizontal space at Q is the set of matrices

$$H_Q = \left\{ Q \begin{pmatrix} 0 & -B^T \\ B & 0 \end{pmatrix} \right\}.$$

We observe that the orthogonal group geodesic

$$Q(t) = Q \exp \left(t \begin{pmatrix} 0 & -B^T \\ B & 0 \end{pmatrix} \right)$$

has horizontal tangent vector

$$\dot{Q}(t) = Q(t) \begin{pmatrix} 0 & -B^T \\ B & 0 \end{pmatrix}$$

for all t along $Q(t)$. Therefore, $Q(t)$ is still a shortest path on the quotient space $\text{Gr}(k, n)$, i.e. by further restricting $\mathbf{A}$ to be of the form

$$\tilde{\mathbf{A}} = \begin{pmatrix} 0 & -B^T \\ B & 0 \end{pmatrix}, \quad B \in \mathcal{R}^{(n-k) \times (k)}$$

we obtain a representative of the geodesic path on $\text{Gr}(k, n)$

$$Q(t) = Q \exp(t\tilde{\mathbf{A}}).$$

The $(n-k)$ -by- (k) submatrix B specifies the velocity of the geodesic flow. This approach provides us an easy method to compute the geodesic formula on the Grassmann manifold using n -by- k matrices.

In numerical applications, what matters is the span of the first k columns of $Q(t)$; hence, the geodesic formula can be rewritten as,

$$\Phi(t) = Q \exp(t\tilde{\mathbf{A}})J$$

where $J = \begin{bmatrix} I_k \\ 0_{n-k,k} \end{bmatrix} \in \mathcal{R}^{n \times k}$ and $\Phi(t)$ is a n -by- k orthonormal matrix for any $t \in [0, 1]$. For numerical computation we introduce the following way to represent points on the Stiefel manifold. Each point on $St(k, n)$ can be expressed as a n -by- k orthonormal matrix Y such that $Y^T Y = I_k$. In this way, we can view the Grassmann

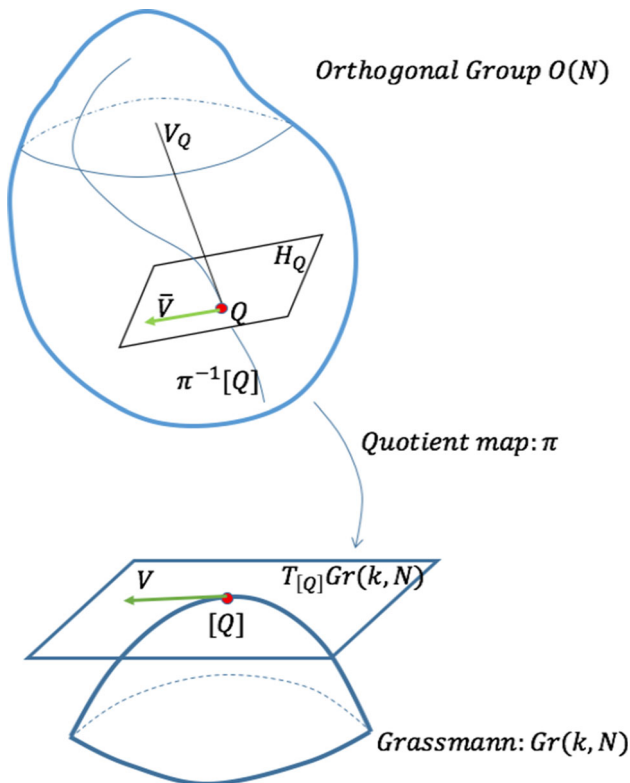


Fig. 1 This figure illustrates the idea of the vertical space V_Q and horizontal space H_Q at a point Q . $T_{[Q]}\text{Gr}(k, n)$ is the tangent space to $\text{Gr}(k, n)$ at $[Q]$. It can be shown that a tangent vector $V \in T_{[Q]}\text{Gr}(k, n)$ can be uniquely represented by a tangent vector $\bar{V} \in H_Q$. Hence, we can use matrices to represent tangent vectors to points on $\text{Gr}(k, n)$

manifold $\text{Gr}(k, n)$ as a quotient manifold of the Stiefel Manifold $St(k, n)$. The equivalence class $[Y]$ can be defined as

$$[Y] = \{YD : D \in O(k)\}.$$

By viewing the Grassmann manifold as a quotient manifold of the Stiefel manifold, we can compute the corresponding vertical space and horizontal space. With the computation illustrated above, one can verify the vertical space at Y is the set of matrices,

$$V_Y = \{YA : A \in \mathbb{R}^{k \times k}, A^T + A = 0\}.$$

And the horizontal space is

$$H_Y = \{N : N \in \mathbb{R}^{n \times k}, N^T Y = 0\}.$$

Therefore, a tangent vector to the Grassmann manifold at $[Y]$ can be uniquely represented by a n -by- k matrix N where $Y^T N = 0$. With this in mind, we can use the following theorem to compute the geodesic flow $\Phi(t)$ numerically in a convenient way.

Theorem 1 If $\Phi(t) = Q \exp(t\tilde{A})J$, with $\Phi(0) = X$ and $\dot{\Phi}(0) = H$ where $\tilde{A} = \begin{bmatrix} 0 & -B^T \\ B & 0 \end{bmatrix}$, then

$$\Phi(t) = XV \cos(\Sigma t) V^T + U \sin(\Sigma t) V^T \quad (4)$$

where $U\Sigma V^T$ is the compact singular value decomposition (SVD) of H . Here U is an n -by- k orthonormal matrix, Σ is a k -by- k diagonal matrix and V is a k -by- k orthogonal matrix.

The proof of this Theorem is given in [13]. At this point, given initial conditions i.e. an initial position X and an initial velocity H , we can sample the resulting geodesic at various values of $t \in [0, 1]$. Our task, however, is the inverse operation: Given n -by- k orthonormal matrices X and Y representing equivalence classes $[X]$ and $[Y]$ on $\text{Gr}(k, n)$, find an appropriate velocity matrix H such that a geodesic with velocity H , starting at $[X]$, reaches $[Y]$ in unit time. The idea behind these two problems is illustrated in Fig. 2.

Instead of computing H directly, we assemble H via its compact SVD $H = U\Theta V^T$. By Theorem 1, at $t = 1$, we have

$$YD = XV \cos(\Theta) V^T + U \sin(\Theta) V^T$$

where D is any k -by- k orthogonal matrix (since we are only required to reach a point in the equivalence class, i.e. $YD \in [Y]$). H is in the tangent space; hence, it can be readily verified that $X^T H = 0$ and consequently $X^T U = 0$. Multiplying by X^T on both sides of the equation yields

$$V \cos(\Theta) V^T = X^T Y D$$

$$U \sin(\Theta) V^T = (I - XX^T) Y D$$

Then,

$$\begin{aligned} U \sin(\Theta) V^T (V \cos(\Theta) V^T)^{-1} &= U \tan(\Theta) V^T \\ &= (I - XX^T) Y D (X^T Y D)^{-1} \\ &= (I - XX^T) Y (X^T Y)^{-1} \end{aligned}$$

Therefore, to find the velocity matrix H , it suffices to compute the compact SVD, $(I - XX^T) Y (X^T Y)^{-1} = U \Sigma V^T$, and $H = U \Theta V^T$ where $\Theta = \arctan(\Sigma)$. One subtlety in Eq. (4) is that if V is multiplied from the right on both sides of the equation, we still have a representative of the same equivalence class as $\Phi(t)$, i.e. $\Phi(t)V$ is equivalent to $\Phi(t)$ on $\text{Gr}(k, n)$ for all t .

To summarize the derivation above, we present the formula for computing the geodesic path between two points $X, Y \in \text{Gr}(k, n)$, which can be found in [1].

$$G(t) = XV \cos \Theta t + U \sin \Theta t \quad (5)$$

We observe that

$$[G(0)] = [X]$$

and

$$[G(1)] = [Y]$$

and the trajectory $G(t)$ traces out the path of shortest distance on $\text{Gr}(k, n)$ in terms of the geodesic metric given by Equation (2). The quantities U, Σ and V are found by computing the singular value decomposition of the projection of

$$M = Y(X^T Y)^{-1}$$

onto the orthogonal complement of X , i.e.,

$$U \Sigma V^T = (I - XX^T) Y (X^T Y)^{-1}$$

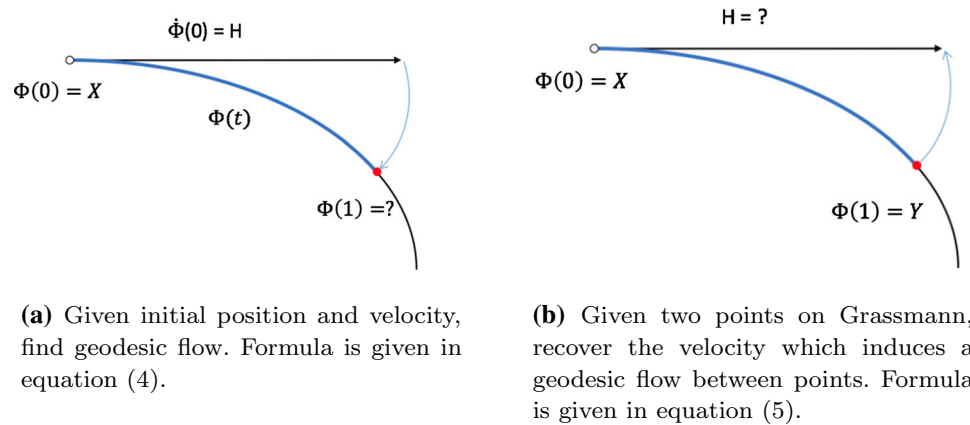
where X and Y are given and the inverse of $X^T Y$ exists. Further, it can be shown that

$$\Theta = \text{atan}(\Sigma)$$

to complete the requirements of computing the geodesic between two subspaces X and Y as prescribed in Eq. (5). This formula is a key ingredient for extending the self-organizing mapping algorithm on vector spaces to Grassmannians.

We present the following example as an illustration of the numerical computation of the geodesic formula between two points.

Fig. 2 This figure illustrates the two problems related to the geodesic formula



Let

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

and

$$Y = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{3}} \\ 0 & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{2}} & 0 \end{pmatrix}$$

be two matrices representing points $[X]$ and $[Y]$ on $\text{Gr}(2, 4)$. First we compute the compact singular value decomposition

$$U\Sigma V^T = (I - XX^T)Y(X^TY)^{-1}.$$

We find

$$\Sigma = \begin{pmatrix} 1.6180 & 0 \\ 0 & 0.6180 \end{pmatrix}.$$

Compute atan along the diagonal to get

$$\Theta = \begin{pmatrix} 1.0172 & 0 \\ 0 & 0.5536 \end{pmatrix}.$$

Hence

$$\cos \Theta t = \begin{pmatrix} \cos(1.0172t) & 0 \\ 0 & \sin(0.5536t) \end{pmatrix}$$

and

$$\sin \Theta t = \begin{pmatrix} \sin(1.0172t) & 0 \\ 0 & \sin(0.5536t) \end{pmatrix}.$$

Now we can sample points along this geodesic.

For $t = 0$,

$$G(0) = \begin{pmatrix} -0.5257 & -0.8507 \\ 0.8508 & -0.5257 \\ 0 & 0 \\ 0 & 0 \end{pmatrix},$$

whose column vectors span the same subspace as the column vectors of X . The same can also be verified for Y and $G(1)$ by computing the principal angles between Y and $G(1)$. If t is sampled uniformly on the interval $[0, 1]$, one can verify that all the distances between any pair of adjacent points are the same. i.e. the geodesic has constant speed. One example is at $t = \frac{1}{2}$, the geodesic distance between X and $G(\frac{1}{2})$ is the same as the distance between $G(\frac{1}{2})$ and Y . i.e. $d_g(X, G(\frac{1}{2})) = d_g(G(\frac{1}{2}), Y)$ (Table 1).

Table 1 The classes of the Indian Pines data

Alfalfa	46
Corn-notill	1428
Corn-mintill	830
Corn	237
Grass-pasture	483
Grass-trees	730
Grass-pasture-mowed	28
Hay-windrowed	478
Oats	20
Soybean-notill	972
Soybean-mintill	2455
Soybean-clean	593
Wheat	205
Woods	1265
Buildings-Grass-Trees-Drives	386
Stone-Steel-Towers	93

The bold denotes the data that were used in our experiment

3 Self-organizing mappings on $\text{Gr}(k, n)$

In this section we present the extension of the SOM algorithm on vector spaces to the setting of the Grassmannian, a manifold whose points parameterize all k -dimensional subspaces of a fixed n -dimensional vector space.

Following [16–19], for data $x^{(\mu)}$, $\mu = 1, \dots, P$ in Euclidean space we select an initial set of centers $\{c_i\}$ where the subscript i is the label of the spatial index a_i . Since the algorithm iteratively updates these initial centers, we add a superscript m to identify the value of c_i at the m th iteration. The update equation is given by

$$c_i^{m+1} = c_i^m + \epsilon_m h(d(a_i, a_{i^*}))(x - c_i^m)$$

where i^* is the winning center associated to pattern x , i.e.,

$$i_m^* = \arg \min_i \|x - c_i^m\|_2.$$

Here the distance between the point x and the center c_i^m is given by the standard Euclidean norm. We also take the

$$i^* = \arg \min_i d_g(X, C_i)$$

where the metric d_g is given by Eq. (2). To move the centers toward the pattern subspace X according to the SOM update, we compute the geodesic, as described in detail above, between each subspace center C_i and subspace pattern X

$$U\Sigma V^T = (I - C_i^T)C_i(X^T C_i)^{-1} \text{ and } \Theta = \arctan(\Sigma).$$

Our localization term now becomes

$$t = \epsilon_n h_n(d(a_i, a_{i^*})).$$

We now take

$$h_n(s) = \exp(-s^2/\sigma_n^2)$$

where $\sigma_n = \sigma_0(1 - n/T)$ and $\epsilon_n = \epsilon_0(1 - n/T)$. The centers thus change along the geodesic by moving from $C_i(0)$ to $C_i(t) = C_i V \cos \Theta t + U \sin \Theta t$, where t is adjusted both for the local neighborhoods of the indices and the step size.

Algorithm 1: Grassmannian Self-Organizing Mapping

Input Data: Load class labeled data matrices $\{X_i^j\} \in \mathbb{R}^{n \times d}$ where k is the number of samples in each subspace and n is the dimension of the data, j is the class index, i is the matrix index.

Output Data: Final centers and indices of each data subspace.

Result: Representation of points on $\text{Gr}(k, n)$ as indices of SOM centers.

Initialization: Set the number of samples per subspace k , the number of centers N , initialize centers C_i as random d -dimensional subspaces and select the index set.

Define: Select the (pseudo)-metric on Grassmannian and compute the distance matrix between all pairs of subspaces.

Step 1: Present a random subspace to the network.

Step 2: Move all the centers C_i proportionally towards the presented subspace along the appropriate geodesic.

Step 3: Repeat until convergence.

localization function as the standard

$$h(s) = \exp(-s^2/\sigma^2)$$

and d is a metric that induces the topology on the index set. For simplicity, in this paper we will restrict our attention to the case

$$d(a_i, a_j) = \|a_i - a_j\|_2$$

On the Grassmannian the points are no longer elements of n -dimensional Euclidean space, but points $X, Y \in \text{Gr}(k, n)$, i.e., k -dimensional subspaces of $\mathbb{R}^n$. For a given subspace X we identify the center, i.e., from the set of subspaces that represent centers $\{C_i\}$, that is closest via

4 Numerical results

In this section we apply the Grassmannian SOM algorithm to both synthetic and real-world data sets including hyperspectral images and gene expression data.

4.1 Synthetic data

We begin with an illustrative example concerning a path along random points on the Grassmannian. We randomly select points on $\text{Gr}(2, 10)$ by sampling the interval $[0, 1]$ using the uniform distribution to fill the entries of four 10×2 matrices. These are then orthogonalized using the QR-decomposition to produce the matrices $Q_1, \dots$,

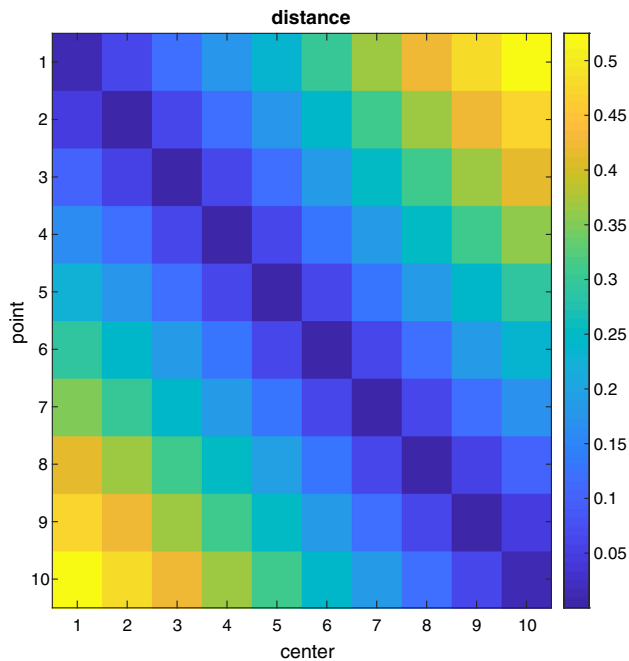


Fig. 3 This matrix has the distances between the point i and center j after convergence. Note that point i is closest to center j when $i = j$, reflecting the ordering mechanism of the Grassmannian SOM

$Q_4 \in \mathbb{R}^{10 \times 2}$. These points serve as representatives of the subspaces $[Q_1], [Q_2], [Q_3], [Q_4] \in \text{Gr}(2, 10)$.

4.1.1 Parameterizing subspaces

Our first application concerns the use of 1D integer index sets to parameterize a path through a set of subspaces on the Grassmannian. When we use clock arithmetic, this is the well-known approximate solution to the traveling salesman problem on a Grassmann manifold. We generate data by sampling ten points on the line segment

$$Z(t) = (1 - t)Q_1 + tQ_2$$

where each element Z is then orthogonalized using the QR-decomposition. Using classical multidimensional scaling, we establish that the distance matrix is non-Euclidean given the eigenvalues $(0.3, 0.0028, -0.0004)$. The Grassmannian SOM algorithm serves to *sort* the points on this segment.¹ We initialize the centers in the Grassmannian SOM algorithm with random points taken from $\text{Gr}(2, 10)$ and use the integer index set $\{1, 2, \dots, 10\}$. The distances between the points and the final ordered centers are shown in Fig. 3. We took $r_0 = 10$, $\epsilon_0 = 0.2$ and the minimum angle distance measure when computing the winning

¹ Of course in this example the ordering of the points on the Grassmannian is available to us. However, in general we can determine a one-dimensional parameterization of a set of points on a Grassmannian that approximately passes through nearest neighbors.

center for a data point; see also [6]. We proceed similarly in the examples that follow. Note that although we are measuring pseudo-distances between points in this way, we are updating the centers along the true Grassmannian geodesics.

4.1.2 A square on $\text{Gr}(2, 10)$

In our second illustrative example, shown in Fig. 4, we add additional segments of points to the curve on the Grassmannian. In one case we connect two segments of ten points each that share one point. Secondly, we make a square by having four segments that share four points. In each case the mapping of the path on $\text{Gr}(2, 10)$ to the square lattice index set captures the geometry of the points. We see an apparently missing point on each lattice where two points on the curve had the same winning center.

4.2 Indian Pines

To illustrate the utility of the proposed method for visualizing real data, we apply it to the well-known Indian Pines hyperspectral image [21]. We have considered this data set before in the context of the band selection problem [8] and the persistent homology for signal detection on Grassmannians [10]. A related visualization application invokes the technique of multidimensional scaling and sparse support vector machines [9]. The classes are shown in Fig. 5.

In this application we selected the 12 classes that were large enough to give 20 subspaces of dimension ten. Since this application is merely intended to illustrate the model, we made no attempt to optimize our parameters. However, our previous work suggests these dimensions are reasonable [9]. Thus, we are visualizing 240 labeled points in 220 dimensions by first constructing sets of ten-dimensional subspaces in 220 dimensions using the SVD.

We initialized the centers for Grassmannian SOM by selecting 900 ten-dimensional subspaces at random, corresponding to a 30×30 integer lattice. This was done by computing the singular value decomposition of matrices of size 220 by 10 from the uniform distribution. In Fig. 6, we see the results of the Grassmannian SOM algorithm where points in the same class have been organized to have similarly valued indices.

In Figs. 7 and 8, we see the results of the Grassmannian SOM when the points reside on $\text{Gr}(2, 220)$ and $\text{Gr}(1, 220)$, respectively. This data set is well-known as a challenging classification problem. For example, there are classes which are inherently very similar such as corn (green), corn-notill (red) and corn-mintill (blue). We see that these three classes are well separated for SOM on $\text{Gr}(10, 220)$ while there is overlap using $\text{Gr}(1, 220)$ and $\text{Gr}(2, 220)$. In

Fig. 4 Mapping of two segments (left) and four connected segments (right) to a 2D lattice

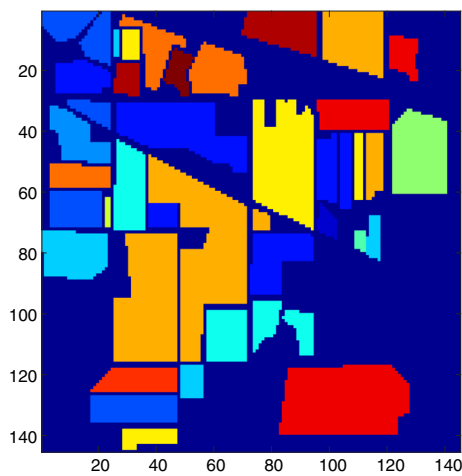
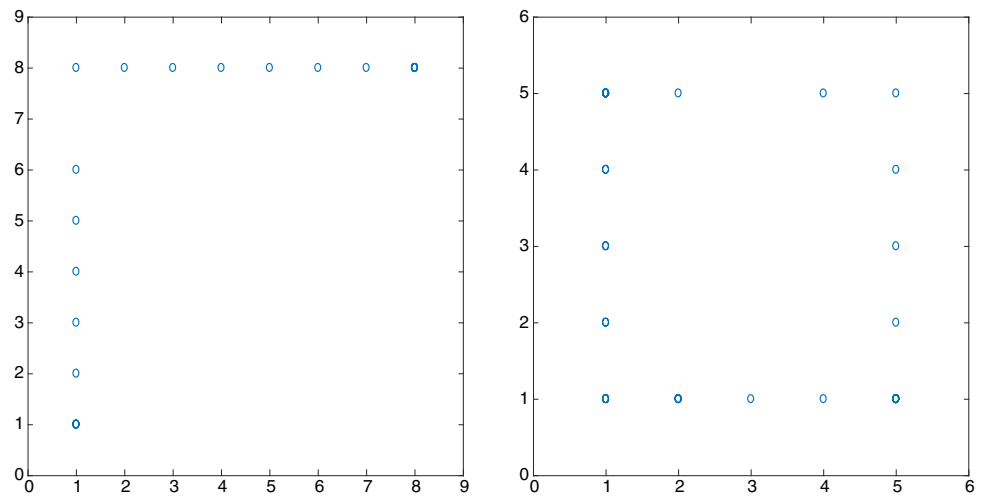


Fig. 5 The class regions of the Indian Pines data set [21]

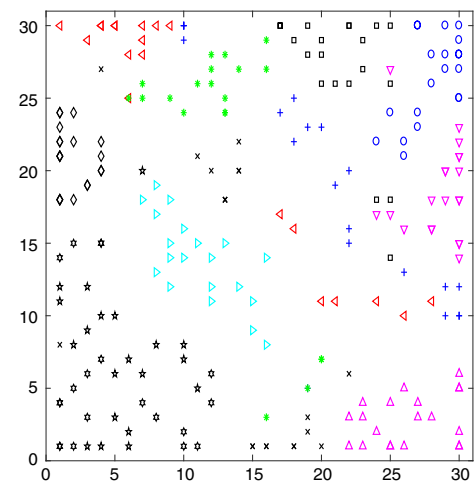


Fig. 7 The converged Grassmannian SOM applied to the Indian Pines classes on $\text{Gr}(2, 220)$ (color figure online)

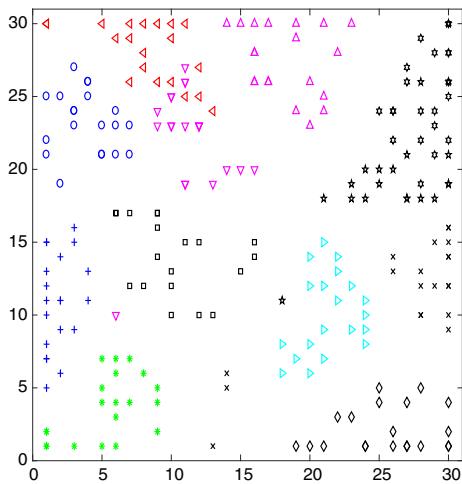


Fig. 6 This figure shows the final configuration of the points as mapped to the 2D index set from $\text{Gr}(10, 220)$

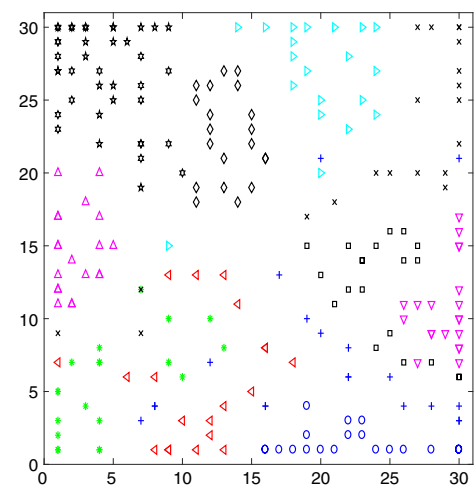


Fig. 8 The converged Grassmannian SOM applied to the Indian Pines classes on $\text{Gr}(1, 220)$ (color figure online)

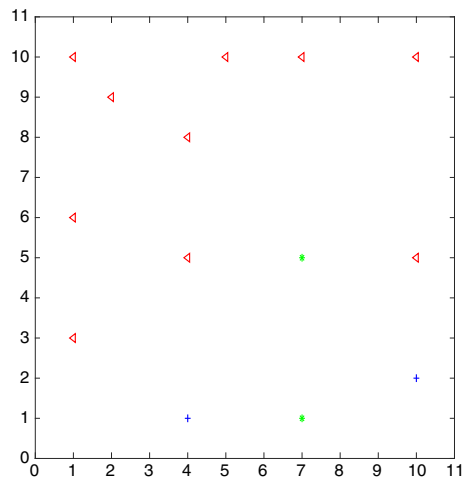


Fig. 9 Mapping of gene expression data on T cell receptor signaling pathway (color figure online)

particular, the corn-mintill (blue) is much less localized on the lower-dimensional Grassmannians. We observe excellent clustering in the majority of classes with the possible exception of green pasture (x) which shows distinct spread suggesting it has significant spectral overlap with the other classes. These results vary the dimension of the Grassmannian and are higher resolution than those presented in the preliminary work [15].

4.3 Gene expression data

Here we examine the application of Grassmannian SOM to two gene expression data sets. The first is related to the immune response in mice to the Ebola virus while the second explores the human immune response to respiratory infection.

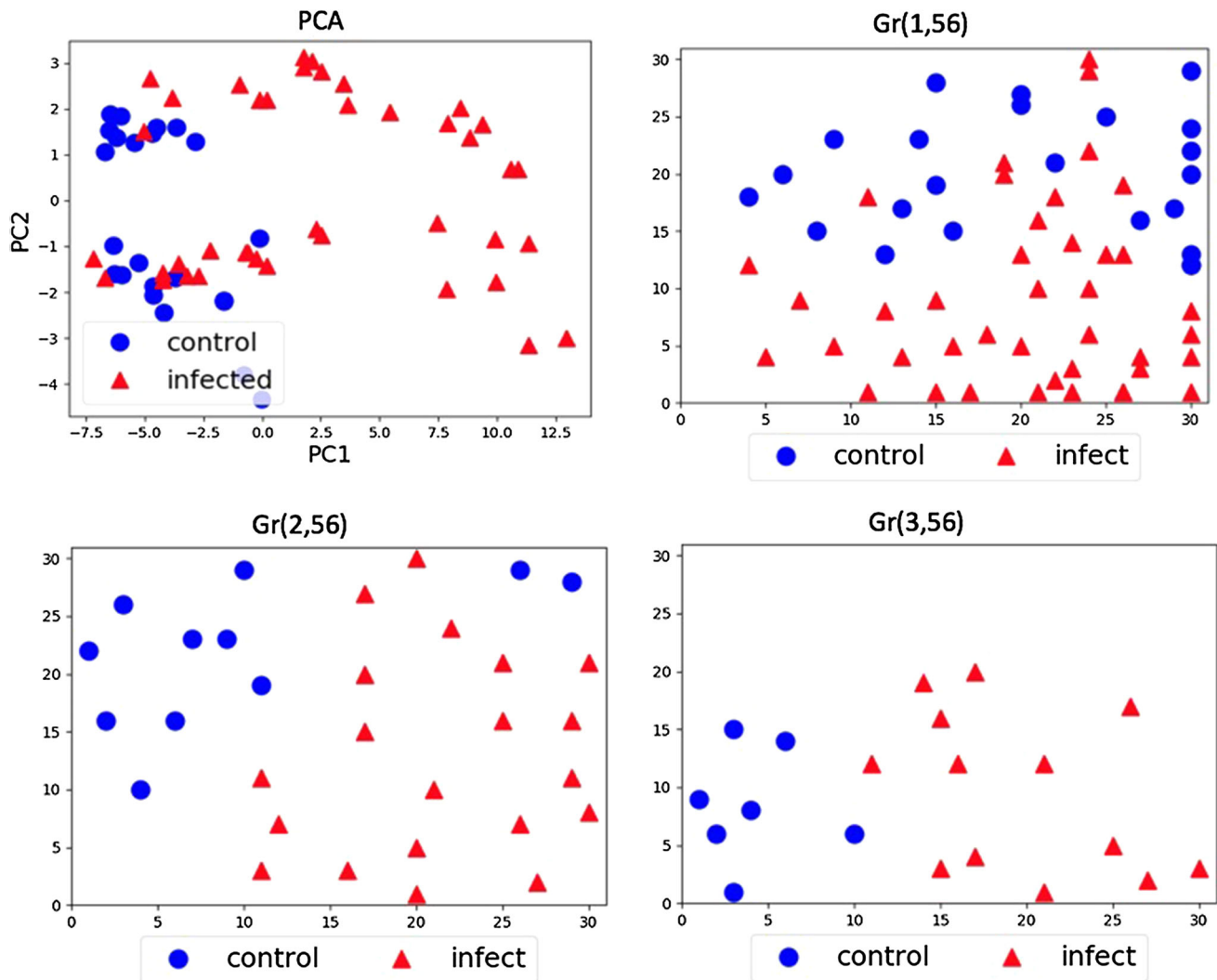


Fig. 10 These four plots are 2D visualizations of Uninfected Control and Infected subjects from hour 30 to 48. Top left: PCA visualization. Top right: Grassmannian SOM on $Gr(1,56)$. Bottom left:

Grassmannian SOM on $Gr(2,56)$. Bottom right: Grassmannian SOM on $Gr(3,56)$ (color figure online)

4.3.1 Ebola mice data

In this example we examine the application of Grassmannian SOM to a gene expression data set collected from mice responding to infection from the Ebola virus [28]. Each raw data point consists of a set of over 12,000 genes. As a preprocessing step, we identify the subset of all discriminatory genes that classify infected versus controls; see [26, 30] for details. Using these genes as the starting point, we identified some 1300 biological pathways potentially of interest in the immune response to infection. Subsequently we applied machine learning techniques to select best pathways for further study. We have selected one of these pathways, i.e., the T cell receptor signaling pathway consisting of 48 genes, as an example to test the Grassmannian SOM algorithm on the Ebola Virus. We pick 3 points at random from each class to construct a single point on the Grassmannian. Hence, each point on the Grassmannian lives on $\text{Gr}(3, 48)$, i.e., it consists of 48 genes and three biological samples. The result of training the Grassmannian SOM algorithm is shown in Fig. 9 where the high-dimensional observations are mapped to the two-dimensional index set in the usual manner. The red points represent control samples of healthy non-human primates. The green * points are the gene expression values at day one, while the blue + samples reflect expression at day 2 after exposure to infection by the Ebola virus. Although we do observe some of the desired clustering with this example, additional data appear to be required to provide a more complete picture. Hence, we present the following example on H3N2 influenza data set.

4.3.2 H3N2 influenza data

The H3N2 gene expression data sets were downloaded from GEO GSE73072 which consists of 7 studies. Two H3N2 challenge studies, i.e. *Dee2* and *Dee5*, are selected for this experiment. See [22] for more details. We used the Reactome interferon alpha beta signaling pathway, which contains 56 genes, to form our data matrix. Hence, each data point resides on $\text{Gr}(k, 56)$. The solid blue circles represent uninfected control data (before inoculation) and red triangles represent infected samples from hour 30 to 48 after inoculation. For each k , we attached 900 randomly generated k -dimensional subspaces to a 30×30 integer lattice, which is done in the same way as is described in Sect. 4.2. In Fig. 10, we see the results of Grassmannian SOM when data points live on $\text{Gr}(1, 56)$ (top right), $\text{Gr}(2, 56)$ (bottom left) and $\text{Gr}(3, 56)$ (right). We observe that two classes are well separated for SOM on $\text{Gr}(3, 56)$ while we start seeing overlaps on $\text{Gr}(2, 56)$ and even more overlaps on $\text{Gr}(1, 56)$. As a comparison, we also included the 2D visualization via PCA(top left) of this dataset, from

which we can also find overlaps between two classes when data are projected onto the first two principal components. This example shows strong clustering performance when Grassmannian SOM is applied to biological gene expression pathway data.

5 Conclusion

We have presented an extension of the self-organizing mapping algorithm to the geometric setting of the Grassmann manifold. The approach moves centers toward data points presented to the network by moving proportionally along the geodesic, or shortest path, between two elements of $\text{Gr}(k, n)$. We illustrate the method by showing that the algorithm organizes the hyperspectral image data in the index space and separates ten-dimensional subspaces of 220-dimensional space. While lower-dimensional Grassmannians also capture significant structure, the 10-D subspaces captured the most variability consistent with observations made using other algorithms. We also observe that three-dimensional subspaces resolve the H3N2 data into separable control and infected classes while these are clearly non-separable using either a standard PCA projection or Grassmannian SOM with one-dimensional subspaces. Hence, the data subspace perspective is essential to adequately process the data using SOM.

It is not necessary in practice for the points to reside on the same Grassmannian. The distance between a k -dimensional and j -dimensional subspace with $k < j$ is now just a function of the first k angles. It is of course potentially important to experiment with the size of the dimensions, but for this paper we have fixed them to be equal for each subspace.

Note that we have yet to systematically explore the impact of the metric for computing winning centers on the algorithm. One can envision optimizing this metric for improved data visualization. It will be interesting to consider the extension of SOM to other abstract manifolds such as the Stiefel and flag manifolds.

Acknowledgements This work is supported by the National Natural Science Foundation of China under Grant 61771178. We want to thank Shannon Stiverson for discovering the Reactome interferon signal pathway as discriminatory at 48–72 h. This paper is based on research partially supported by the National Science Foundation under Grants No. DMS-1513633, DMS-1322508, and CISE-1712788, as well as DARPA awards N66001-17-2-4020 and D17AP00004.

References

1. Absil P-A, Mahony R, Sepulchre R (2004) Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Acta Appl Math* 80(2):199–220

2. Beveridge JR, Draper B, Chang J-M, Kirby M, Kley H, Peterson C (2008) Principal angles separate subject illumination spaces in YDB and CMU-PIE. *IEEE Trans Pattern Anal Mach Intell* 31:351–363
3. Björck A, Golub GH (1973) Numerical methods for computing angles between linear subspaces. *Math Comput* 27(123):579–594
4. Chang J-M, Beveridge JR, Draper B, Kirby M, Kley H, Peterson C (2006) Illumination face spaces are idiosyncratic. In: *IPCV'06*, vol 2. CSREA Press, pp 390–396
5. Chang J-M, Kirby M, Kley H, Peterson C, Beveridge JR, Draper B (2006) Examples of set-to-set pattern classification. In: *Mathematics in signal processing conference digest*, Royal Agricultural College, Cirencester, UK, The Institute for Mathematics and its Applications, pp 102–105
6. Chang J-M, Kirby M, Kley H, Peterson C, Draper B, Ross BJ (2007) Recognition of digital images of the human face at ultra low resolution via illumination spaces. In: *Computer vision—ACCV 2007*. Springer, pp 733–743
7. Chang J-M, Kirby M, Peterson C (2007) Set-to-set face recognition under variations in pose and illumination. In: *2007 Biometrics symposium*, Baltimore, MD
8. Chepushtanova S, Gittins C, Kirby M (2014) Band selection in hyperspectral imagery using sparse support vector machines. In: Miguel V-R, Fred AK (eds) *Algorithms and technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XX*, volume 9088. International Society for Optics and Photonics, 2014
9. Chepushtanova S, Kirby M (2017) Sparse Grassmannian embeddings for hyperspectral data representation and classification. *IEEE Geosci Remote Sens Lett* 14(3):434–438
10. Chepushtanova S, Kirby M, Peterson C, Ziegelmeier L (2015) An application of persistent homology on Grassmann manifolds for the detection of signals in hyperspectral imagery. In: *Proceedings of the IEEE international geoscience and remote sensing symposium (IGARSS)*, Milan, Italy, 2015
11. Conway JH, Hardin RH, Sloane NJA (1996) Packing lines, planes, etc.: packings in Grassmannian spaces. *Exp Math* 5(2):139–159
12. Draper B, Kirby M, Marks J, Marrinan T, Peterson C (2014) A flag representation for finite collections of subspaces of mixed dimensions. *Linear Algebra Appl* 451:15–32
13. Edelman A, Arias TA, Smith ST (1998) The geometry of algorithms with orthogonality constraints. *SIAM J Matrix Anal Appl* 20(2):303–353
14. Kaski S, Kangas J, Kohonen T (1998) Bibliography of self-organizing map (som) papers: 1981–1997. *Neural Comput Surv* 1(3&4):1–176
15. Kirby M, Peterson C (2017) Visualizing data sets on the grassmannian using self-organizing mappings. In: *2017 12th international workshop on self-organizing maps and learning vector quantization, clustering and data visualization (WSOM)*, pp 1–6
16. Kohonen T (1982) Self-organized formation of topologically correct feature maps. *Biol Cybern* 43:59
17. Kohonen T (1990) The self-organizing map. *Proc IEEE* 78(9):1464–1480
18. Kohonen T (1998) The self-organizing map. *Neurocomputing* 21(1):1–6
19. Kohonen T (2013) Essentials of the self-organizing map. *Neural Netw* 37:52–65
20. Kohonen T, Oja E, Simula O, Visa A, Kangas J (1996) Engineering applications of the self-organizing map. *Proc IEEE* 84(10):1358–1384
21. Landgrebe D (1992) AVIRIS NW Indiana's Indian Pines 1992 data set
22. Liu T-Y, Burke T, Park LP, Woods CW, Zaas AK, Ginsburg GS, Hero AO (2016) An individualized predictor of health and disease using paired reference and target samples. *BMC Bioinf* 17(1):47
23. Marrinan T, Ross BJ, Draper B, Kirby M, Peterson C (2015) Flag manifolds for the characterization of geometric structure in large data sets. In: *Numerical mathematics and advanced applications—ENUMATH 2013*. Springer, pp 457–465
24. Marrinan T, Draper B, Ross BJ, Kirby M, Peterson C (2014) Finding the subspace mean or median to fit your need. In: *2014 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, pp 1082–1089
25. Marrinan T, Ross BJ, Draper B, Kirby M, Peterson C (2016) Flag-based detection of weak gas signatures in long-wave infrared hyperspectral image sequences. In: *Proceedings SPIE defense + security*. International Society for Optics and Photonics, pp 98401N–98401N
26. O'Hara S, Wang K, Slayden RA, Schenkel AR, Huber G, O'Hern CS, Shattuck MD, Kirby M (2013) Iterative feature removal yields highly discriminative pathways. *BMC Genom* 14(1):832
27. Ritter H, Kohonen T (1989) Self-organizing semantic maps. *Biol Cybern* 61(4):241–254
28. Rubins KH, Hensley LE, Wahl-Jensen V, Daddario DiCaprio KM, Young HA, Reed DS, Jahrling PB, Brown PO, Relman DA, Geisbert TW (2007) The temporal program of peripheral blood gene expression in the response of nonhuman primates to Ebola hemorrhagic fever. *Genome Biol* 8(8):R174
29. Santamaria I, Scharf LL, Peterson C, Kirby M, Francos J (2016) An order fitting rule for optimal subspace averaging. In: *Statistical signal processing workshop (SSP), 2016 IEEE*. IEEE, pp 1–4
30. Wang K, Langevin S, O'Hern C, Shattuck M, Ogle S, Forero A, Morrison J, Slayden R, Katze M, Kirby M (2016) Anomaly detection in host signaling pathways for the early prognosis of acute infection. *PloS ONE* 11:e0160919

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.