# Set-to-Set Pattern Recognition on Grassmann Manifolds

*Sofya Chepushtanova* and Michael Kirby*

*Department of Mathematics, Colorado State University, Fort Collins, CO*

## Grassmann Manifold Framework: Motivation

- **Set-to-set pattern recognition:** a set of points from a class characterizes the variability of the class information.
- **Grassmann manifolds** $G(k,n)$ (collections of $k$-dimensional subspaces of $\mathbb{R}^n$) provide a geometric framework for characterizing sets of points.
- Subspaces can be realized as points in Euclidean space via **multidimensional scaling**.
- **Sparse support vector machine** identifies optimal dimensions of embedded subspaces.

## Constructing Points on Grassmann Manifold $G(k,n)$

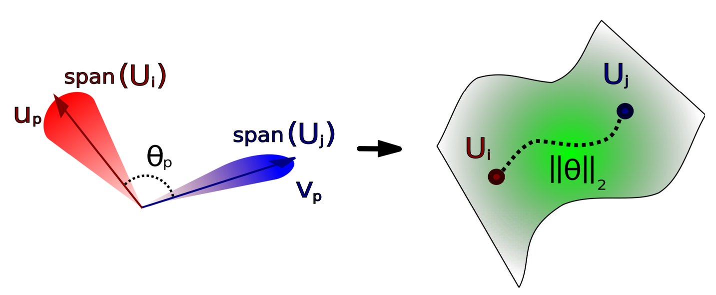
A point on $G(k,n)$ can be represented by an $n \times k$ orthogonal matrix $U$ ($U^TU = I_k$):

Data points in $\mathbb{R}^n$, dimension $k$ → (sample data) → Form $n \times k$ "tall-skinny" matrices $Y_1, Y_2, \ldots, Y_N$ → (SVD: $Y_i = U_i \Sigma_i V_i^T$) → $U_1, U_2, \ldots, U_N$ represent points on $G(k,n)$.

## Embedding $G(k,n)$ in $\mathbb{R}^d$ via Multidimensional Scaling (MDS)

### Classical MDS (Mardia):

- Obtain distance matrix $D \in \mathbb{R}^{N\times N}$ : $D_{ij} = 0, D_{ij} \geq 0$
- Find $B = HAH$, where $H = I - \frac{1}{N}ee^T$ and $A_{ij} = -\frac{1}{2}D_{ij}^2$
- Find the spectral decomposition $B = \Gamma\Lambda\Gamma^T$.
- $X = \Gamma\Lambda^{\frac{1}{2}}$ gives points in $\mathbb{R}^d$, where $d = rank(B) = rank(X) \leq N-1$ (Note: $Be = 0e$.)




- **The arc length distance** is chosen for $D$ on $G(k,n)$: $D_{ij} = d_G(U_i, U_j) = \|\theta\|_2 = \sqrt{\sum_{p=1}^k \theta_p^2}$, where $\theta_p$ are principal angles between subspaces.

**Note:** if $B$ is positive semidefinite (i.e. the eigenvalues are nonnegative), the resulting configuration preserves the geodesic distances (otherwise this is the best approximation to preserving distances).

## Sparse Support Vector Machine (SSVM)

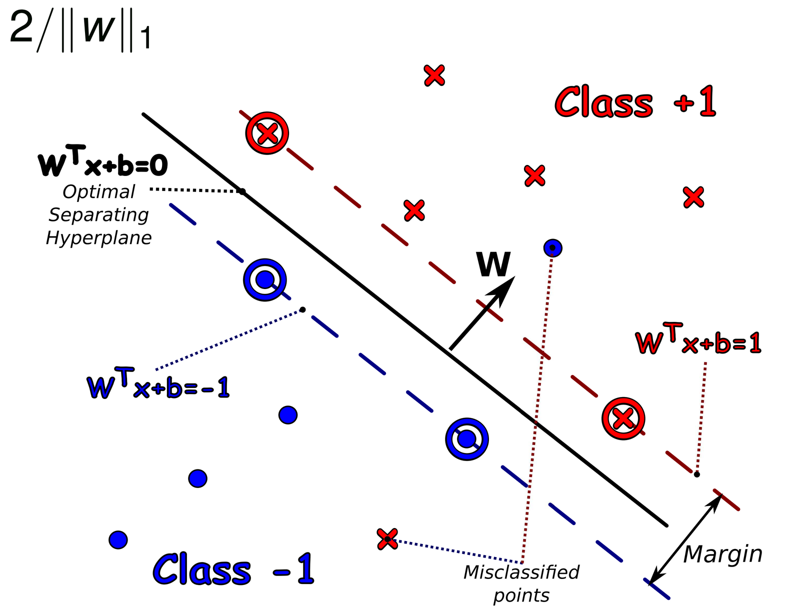
- 2-class data points $x_i \in \mathbb{R}^d$ with labels $y_i \in \{-1,+1\}, i = 1,\ldots,N$
- **Separating hyperplane** $P = \{x : w^Tx + b = 0\}$, $w \in \mathbb{R}^d$ is normal to $P$
- Points on $w^Tx + b = \pm 1$ are **support vectors**
- The optimal $P$ has the largest **margin** $2/\|w\|_1$

**Optimization problem:**

$$\min_{w,b,\xi} \|w\|_1 + Ce^T\xi$$
$$\text{s. t. } y_i(w^Tx_i + be) \geq 1 - \xi_i, \quad (1)$$
$$\xi_i \geq 0, i = 1,\ldots,N$$

**Decision function:**

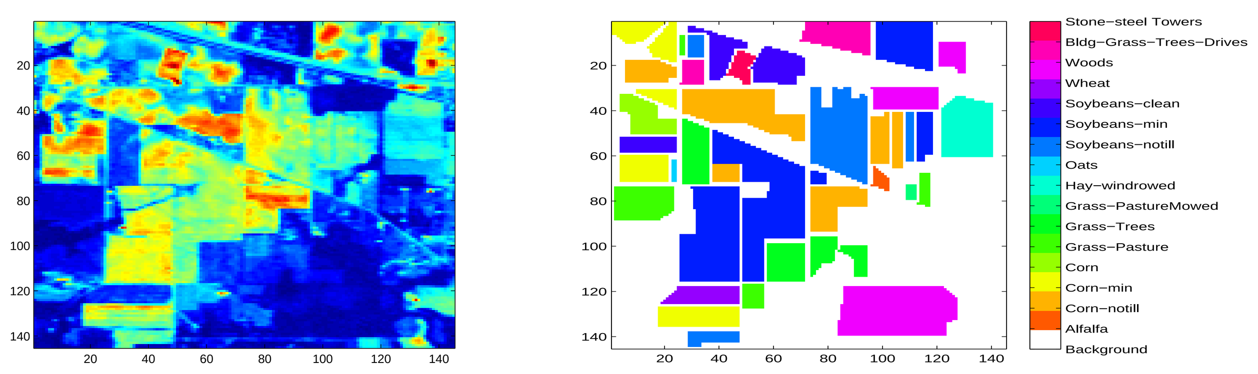
$f(x) = sgn(w^Tx + b)$




## Algorithm Summary

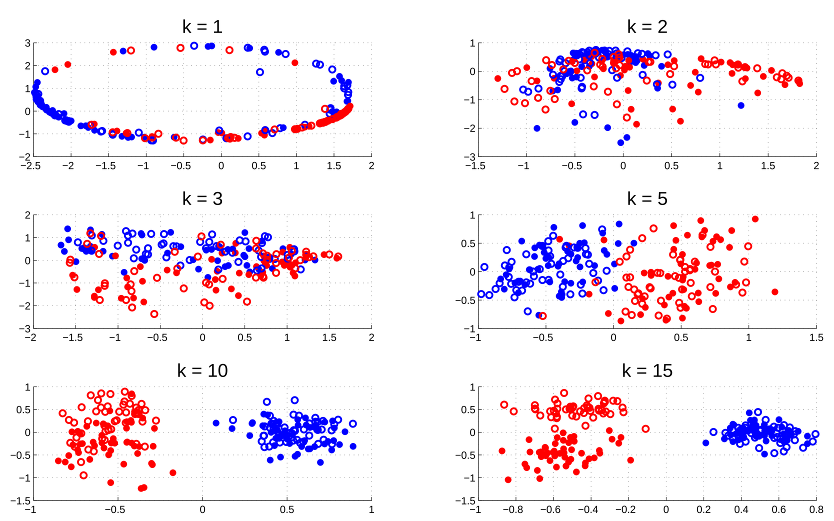
Compute points and distance matrix $D$ on $G(k,n)$
↓
Embed subspaces in Euclidean space (MDS) preserving distances
↓
Feature (dimension) selection and classification via SSVMs

## Application to Hyperspectral Imagery

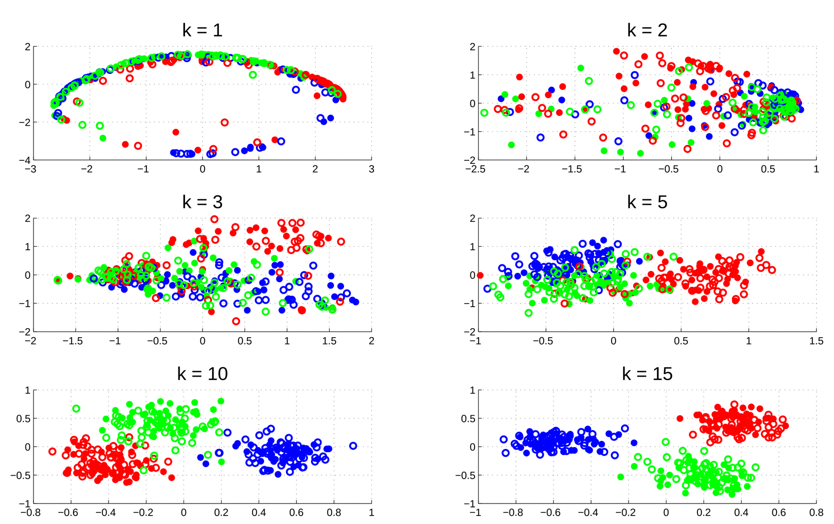


- AVIRIS Indian Pines data set collected in Indiana in 1992
- $145 \times 145$ images, 220 spectral bands (ranging from 0.4 to 2.5$\mu m$).
- The scene contains two-thirds agriculture and one-third other natural vegetation.

**Examples of embedding points on $G(k,220)$ in Euclidean space:**




**2 Classes:**
Corn-notill (blue) and Grass/Pasture (red).
*Dimensions correspond to the two top eigenvalues of B (MDS).*
Solid dots - training set, hollow dots - testing set

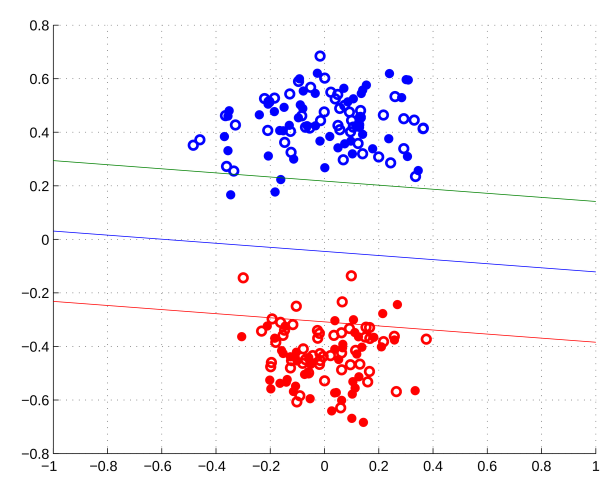



**3-Classes:**
Corn-notill (blue), Grass/Pasture (red) and Grass/Trees (green).
*Dimensions correspond to the two top eigenvalues of B (MDS).*
(Solid dots - training set, hollow dots - testing set)

**SSVM** applied to configuration of points on $G(15,220)$ embedded in Euclidean space:

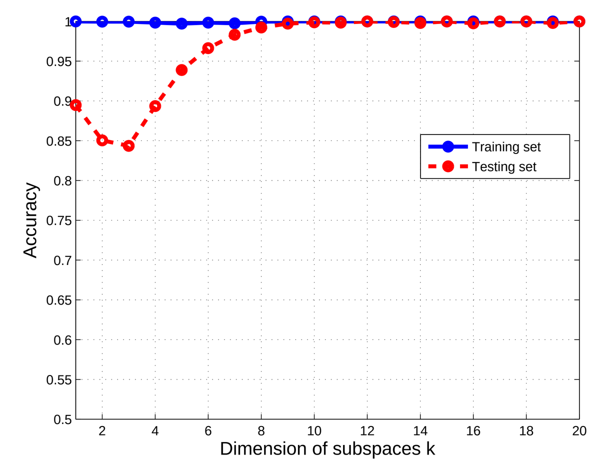
Corn-notill (blue) and Grass/Pasture (red)

*Dimensions correspond to the two largest absolute values of the sparse weight vector $w$. Sparsity is induced by $\ell_1$-norm regularization term in the objective function in (1).*

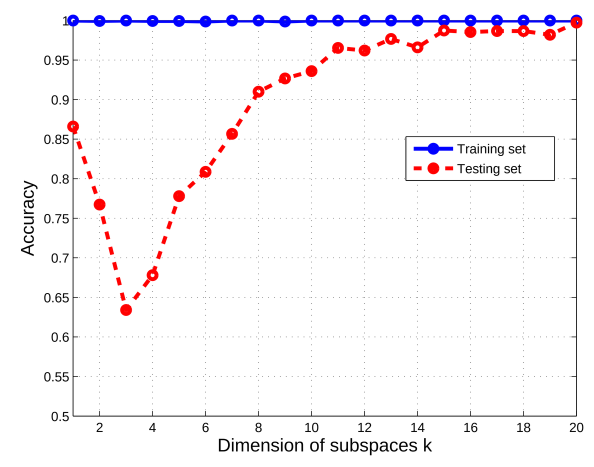


## Application to Hyperspectral Imagery

**SSVM accuracy rates** obtained for Grassmannian $G(k,220)$ as a function of $k$:




*Corn-notill and Grass/Pasture*




*Corn-notill, Grass/Pasture, and Grass/Trees*

**SSVM feature selection** in embedded subspaces of dimension $d$ for $N = 200$ points constructed on $G(k,220)$, for classes Corn-notill and Grass/Pasture:

| dimension of subspaces $k$ | dimension of feature space of embedded points, $d$ | number of negative eigenvalues of $B$ | number of zero eigenvalues of $B$ | features selected | number of features selected |
|---|---|---|---|---|---|
| 1 | 131 | 68 | 1 | 1-3,5-7,10 | 7 |
| 2 | 156 | 43 | 1 | 1-6,8,11 | 8 |
| 3 | 126 | 73 | 1 | 1-6,10-13,16-18,20,23,43,39,47,62,74 | 20 |
| 5 | 147 | 52 | 1 | 1,3,6,9,14,15,18,19,34,37,39,42,52,63 | 14 |
| 10 | 195 | 4 | 1 | 1,4,5,8,15,28,38,65,71 | 9 |
| 20 | 199 | 0 | 1 | 1,3,24,31,63 | 5 |
| 25 | 199 | 0 | 1 | 1,2,8,14 | 4 |

## Contributions/Future Work

- **Novel Components of Algorithm:**
  - Samples of data from a class can be encoded as subspaces that are "abstract" points on a Grassmann manifold, and there are natural metrics for computing distances, e.g. arc length distance.
  - Subspaces are embedded as points in $\mathbb{R}^d$ using MDS.
  - SSVMs capture important low-dimensional feature sets in embedded spaces and can achieve 100% classification accuracy on embedded subspaces generated from test data.
- **Future work:**
  - Determine (computationally) the optimal number of constructed points on $G(k,n)$ for training and testing.
  - Apply the method to other hyperspectral and medical data sets.

## Bibliography

- A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. MATRIX ANAL. APPL, 20(2), pp. 303-353, 1998.
- K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, Academic Press, 1979.
- V. N. Vapnik, *The nature of statistical learning theory*, New York: Springer, 1995.
- O. L. Mangasarian, *Exact 1-norm support vector machines via unconstrained convex differentiable minimization*, Journal of Machine Learning Research, 7, pp. 1517-1530, 2006
- S. Chepushtanova, C. Gittins, and M. Kirby, *Band Selection in Hyperspectral Imagery Using Sparse Support Vector Machines*, preprint 2014

## Acknowledgement

*This material is based upon work supported by the National Science Foundation under Grants No. DMS-1228308 and DMS-1322508. 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.*

 *email: chepusht@math.colostate.edu*