Image patch analysis of sunspots: A dimensionality reduction approach

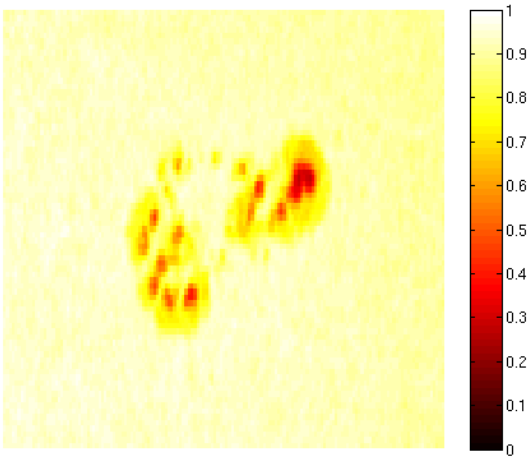
Kevin R. Moon[1], Jimmy J. Li[1], Véronique Delouille[2], Fraser Watson[3], Alfred O. Hero[1]

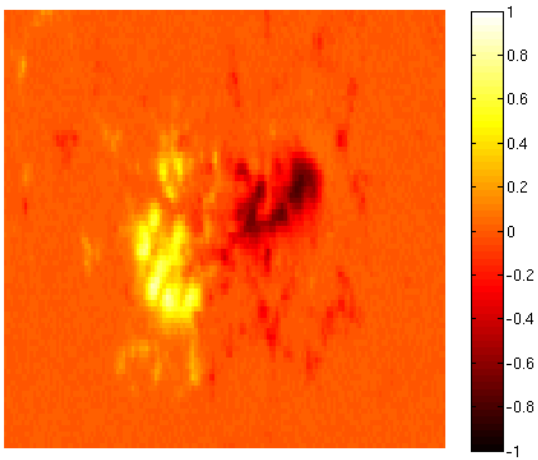
[1]University of Michigan, USA; [2]Royal Observatory of Belgium, Belgium; [3]National Solar Observatory, USA

August 20, 2014

Flares, sunspots, and active regions

- Solar flares and CME's can disrupt technology on Earth
  - Prediction is desirable
- Morphology of sunspot groups and active regions is correlated with solar flare incidence
- Sunspots and active regions are visible in continuum (left) and magnetogram (right) images, respectively

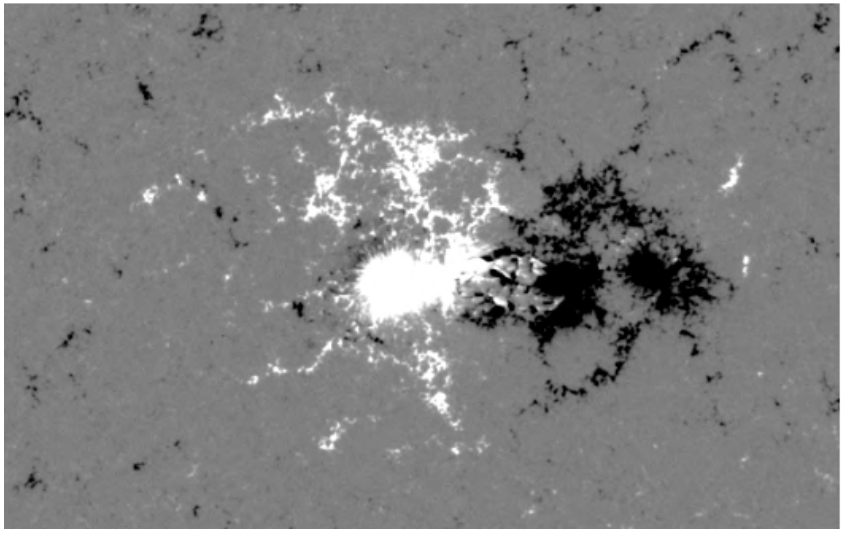




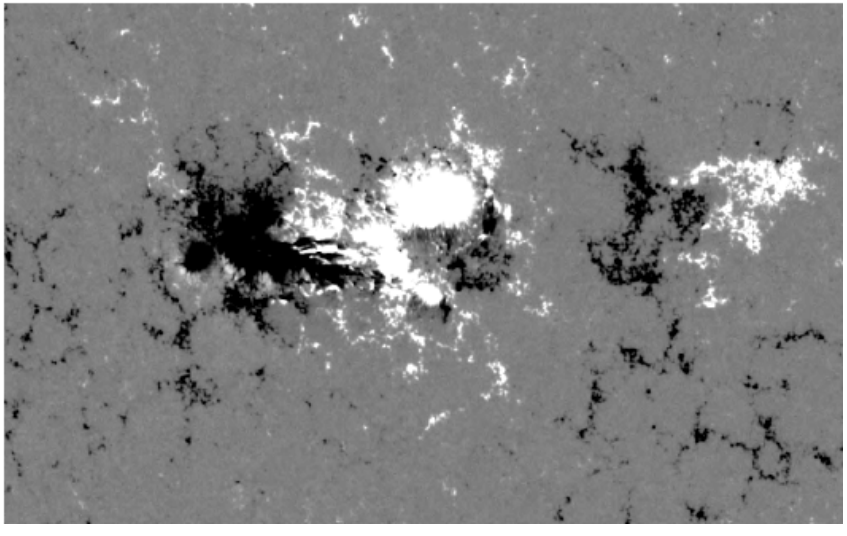
Mt Wilson Classification

- Visual classification system based on **global** features of sunspot/active region configuration
- Alpha–unipolar
- Beta–bipolar and simply divisible
- Beta-Gamma–bipolar but not simply divisible
- Delta–complex; opposite polarity umbrae within the same penumbra
- Gamma–too irregular to be classified as bipolar

Beta



Beta-Gamma



Previous work and our contributions

- Previous work automated the Mount Wilson classification (Stenning et al, 2013) or related multiresolution analysis to Mount Wilson classification (Ireland et al, 2008)
  - Took a supervised classification approach
  - Reduced human bias
  - Still based on a potentially suboptimal classification scheme for solar flare prediction

Previous work and our contributions

- Previous work automated the Mount Wilson classification (Stenning et al, 2013) or related multiresolution analysis to Mount Wilson classification (Ireland et al, 2008)
  - Took a supervised classification approach
  - Reduced human bias
  - Still based on a potentially suboptimal classification scheme for solar flare prediction
- Our goal: build a spatially adaptive descriptive model of the image modalities that can be used for flare prediction
  - I.e. perform unsupervised classification on sunspot images
  - Use both global **and** local image features

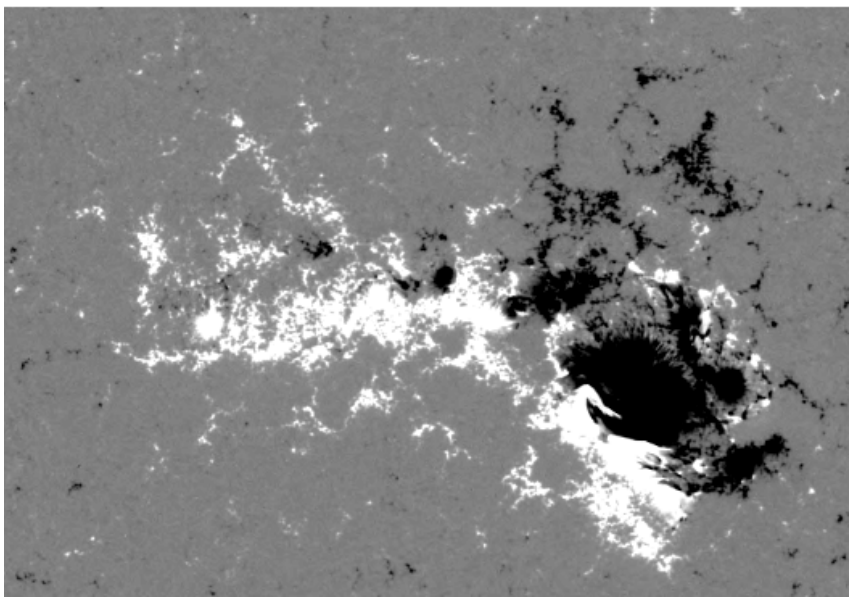
Previous work and our contributions

- Previous work automated the Mount Wilson classification (Stenning et al, 2013) or related multiresolution analysis to Mount Wilson classification (Ireland et al, 2008)
  - Took a supervised classification approach
  - Reduced human bias
  - Still based on a potentially suboptimal classification scheme for solar flare prediction
- Our goal: build a spatially adaptive descriptive model of the image modalities that can be used for flare prediction
  - I.e. perform unsupervised classification on sunspot images
  - Use both global **and** local image features
- Our current contributions focus on **local** features
  - An intrinsic dimension analysis of sunspot images
  - Preliminary image clustering results

Intrinsic dimension motivation

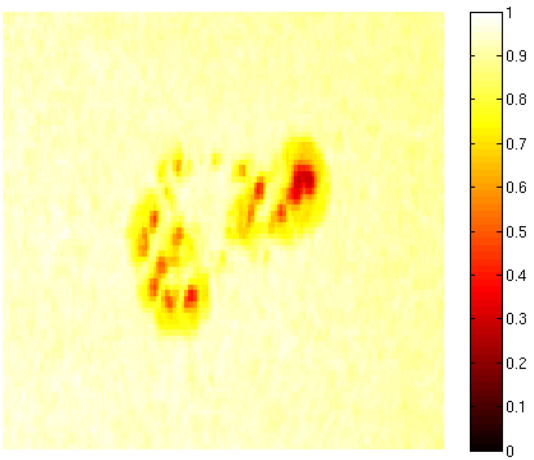
- 3 global parameters required for Mt. Wilson classification
  - Polarity (unipolar, bipolar, irregular)
  - Are opposing polarities separable by a continuous line?
  - Are there opposite polarity umbrae within the same penumbra?
- Adding other parameters (e.g. # of sunspots) would not aid in classification
- So intrinsic dimension is 3

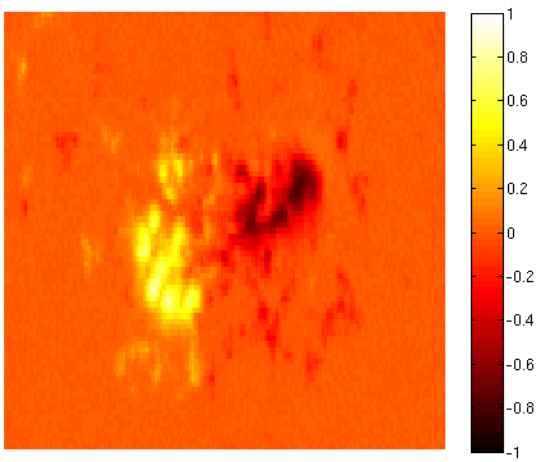
Beta-Gamma-Delta



Intrinsic dimension motivation

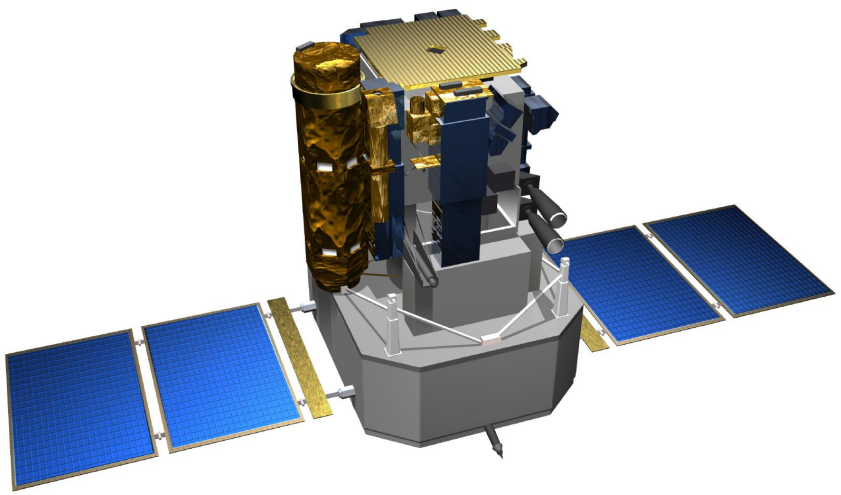
- Consider, for example, a pair of 200×200 pixel images
- Extrinsic dimension is $2 \times 40,000$ (the total number of pixels)
- Question: Can we reduce this without suffering much loss?
  - Apparent spatial and modal dependencies suggest yes
- Specifically,
  1. How many parameters are required to accurately describe/reconstruct the magnetogram and continuum images?
     - I.e. what is the intrinsic dimension?
  2. What about linear vs. nonlinear methods?





Data

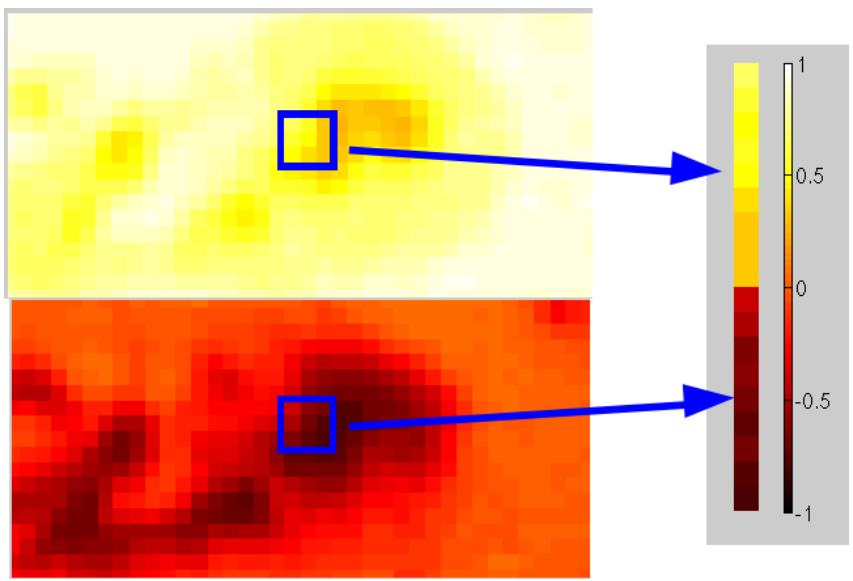
- Images come from the MDI instrument on board the SOHO Spacecraft
  - Continuum images are derived from visible light intensity
  - Magnetogram images measure the intensity and polarity of the longitudinal component of the magnetic field
- Expertly generated masks mark the location of the umbra and penumbra



From NASA's SOHO web page

Image patches as features

- 3×3 patches for continuum (top) and magnetogram (bottom) mapped to a single vector
  - Captures spatial and modal dependencies



- To reconstruct images, use the center pixel location
  - Useful for constructing images of local intrinsic dimension
- Full data matrix is $2d \times N$ where d is the number of pixels in each patch and N is the number of image pixels

Images

Three kinds of images chosen to illustrate our methods. L to R: background, single, multiple sunspots. Continuum (top) and magnetogram (bottom).

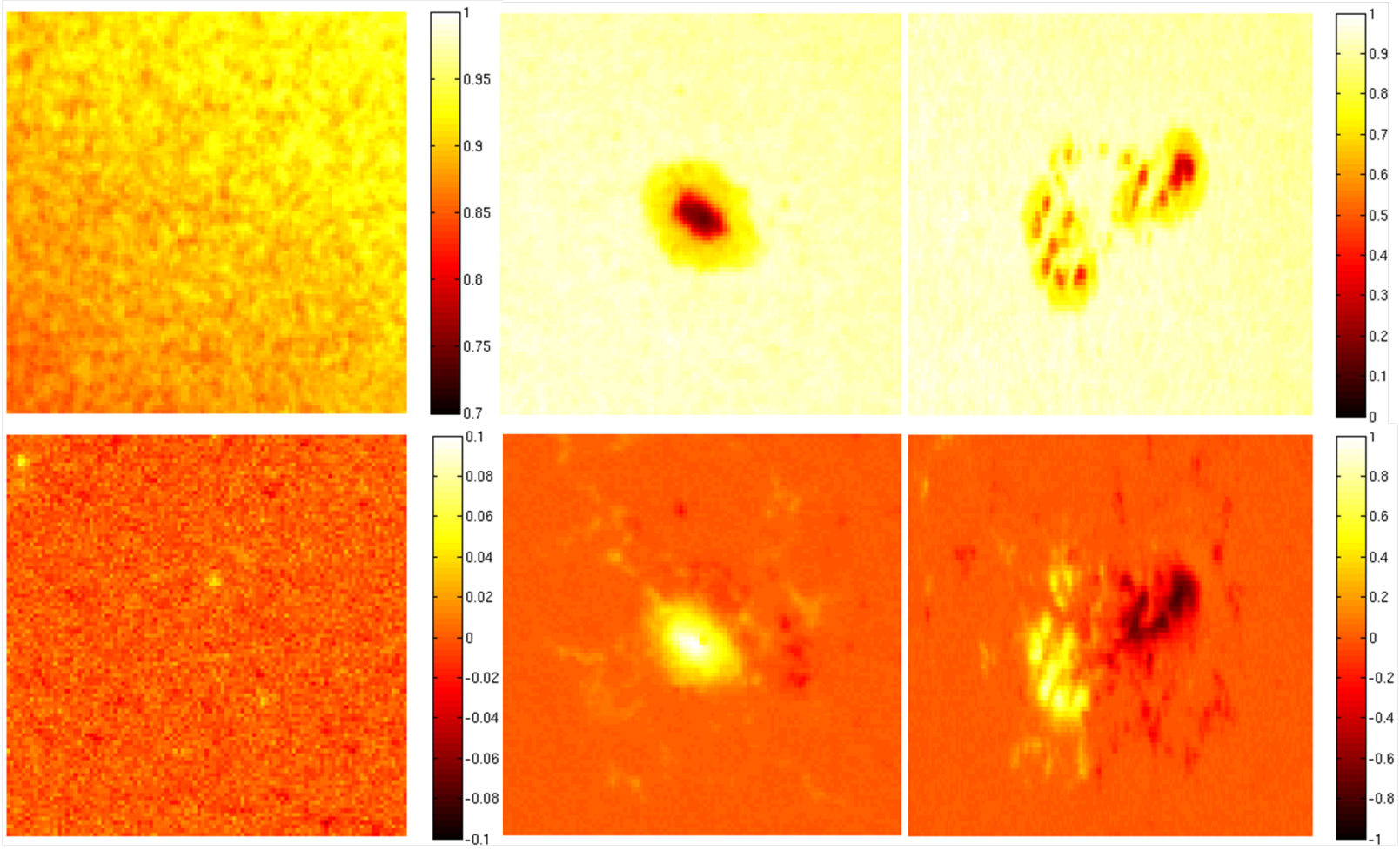
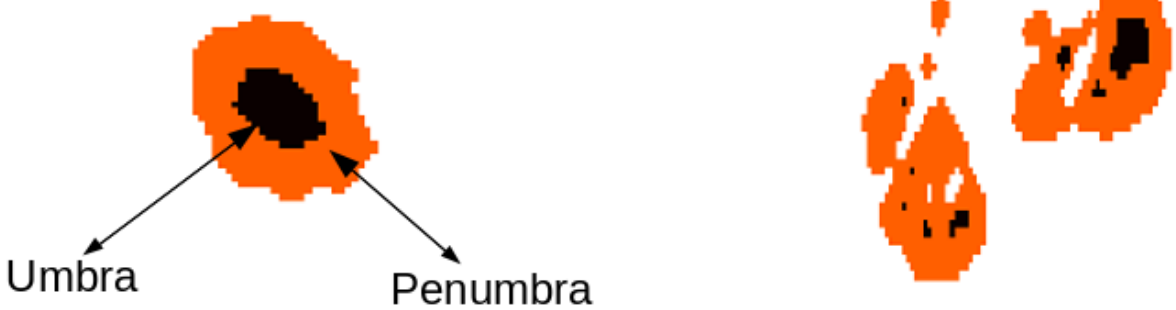
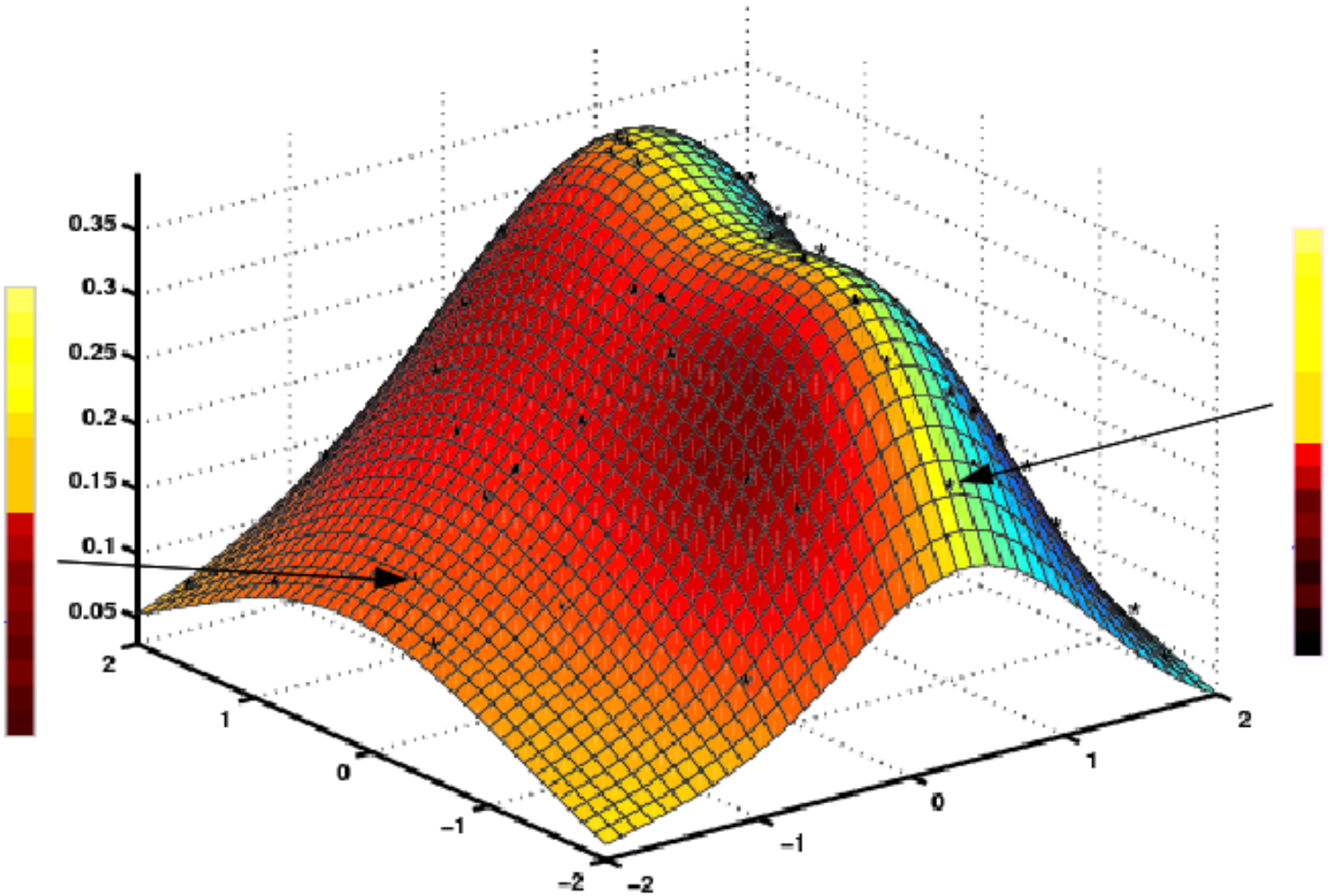


Image masks

Masks for the sunspot images (Watson et al, 2011). The interior is the umbra and the exterior is the penumbra.



Intrinsic dimension estimation



D-dimensional observations lie on surface of dimension $m < D$.

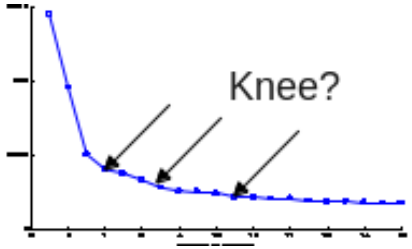
Why intrinsic dimension?

- Knowing the intrinsic dimension and the subspace or manifold can allow us to reduce the dimension of the feature vectors
  - Reduces computational burden
  - Enables more efficient reconstruction and storage
- Can be a measure of feature dependence
- Useful for data interpretation

Intrinsic dimension estimation applied to sunspot images

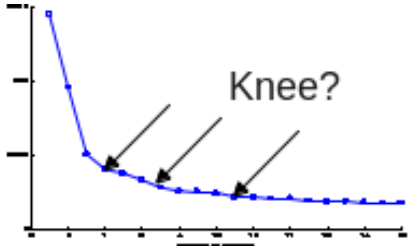
- k-Nearest Neighbors (k-NN) approach (Costa and Hero, 2006; Carter et al, 2010)
  - Appropriate for any smooth manifold
  - Can also estimate local dimension

Intrinsic dimension estimation applied to sunspot images

- k-Nearest Neighbors (k-NN) approach (Costa and Hero, 2006; Carter et al, 2010)
  - Appropriate for any smooth manifold
  - Can also estimate local dimension
- PCA finds a set of linearly uncorrelated vectors (principal components)
  - Principal components are the singular vectors of the data matrix
  - 

  - Only appropriate for linear subspaces

Intrinsic dimension estimation applied to sunspot images

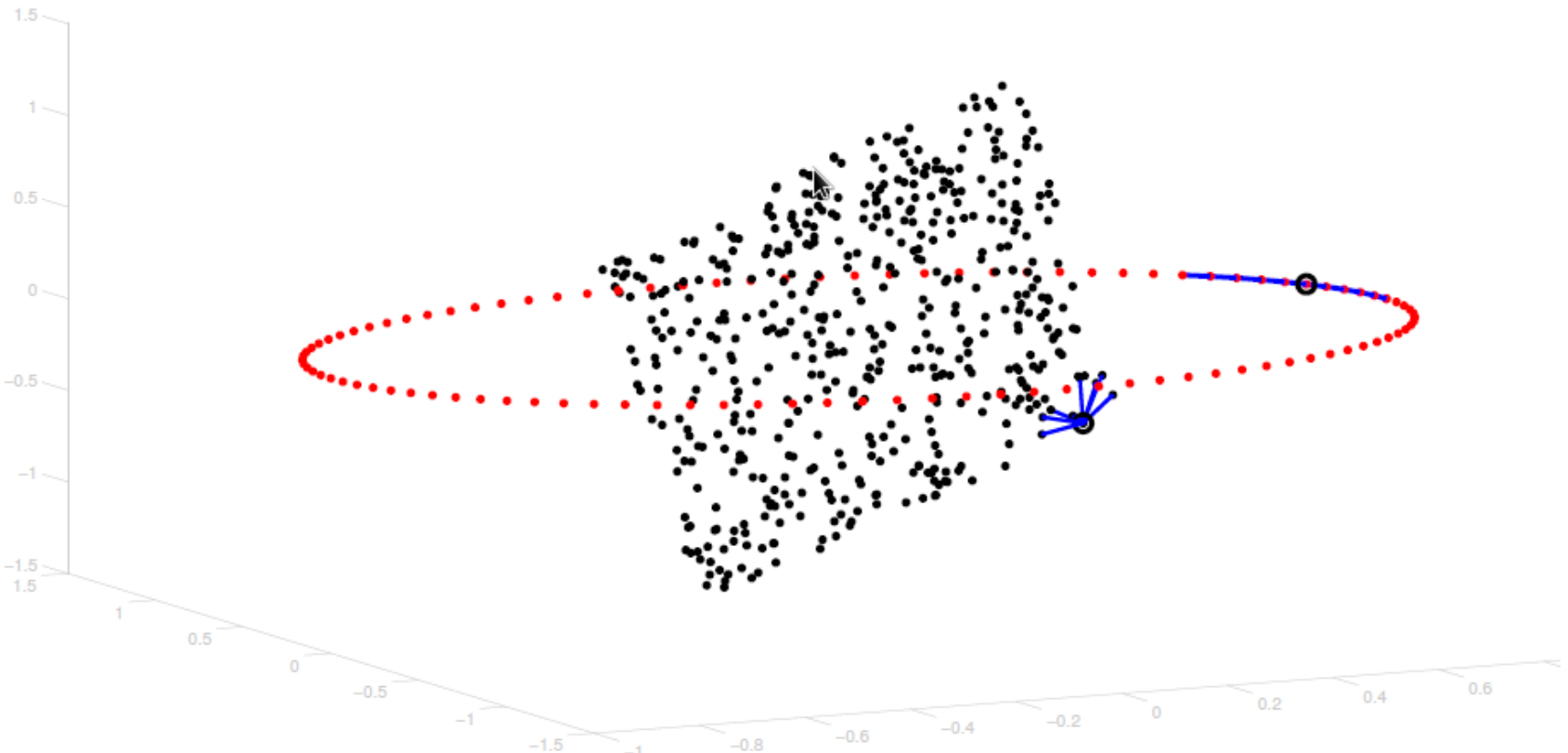
- k-Nearest Neighbors (k-NN) approach (Costa and Hero, 2006; Carter et al, 2010)
  - Appropriate for any smooth manifold
  - Can also estimate local dimension
- PCA finds a set of linearly uncorrelated vectors (principal components)
  - Principal components are the singular vectors of the data matrix
  - 

  - Only appropriate for linear subspaces
- Comparing the two methods enables us to determine if linear decomposition methods are sufficient

Local intrinsic dimension

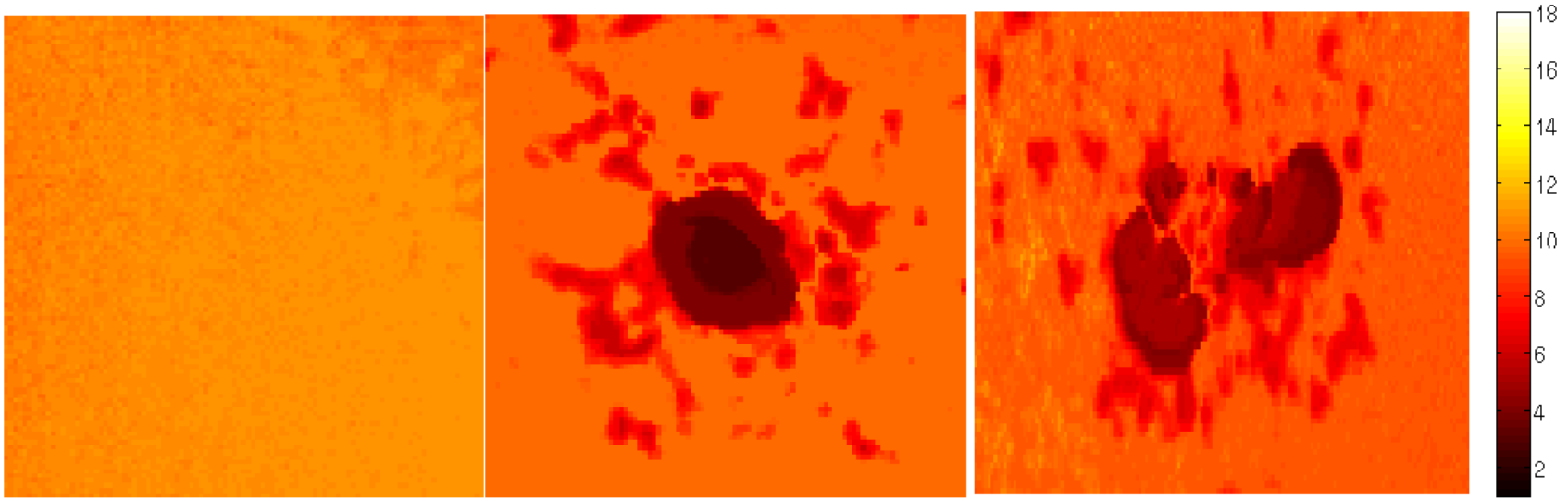
Data points may lie on different manifolds with distinct dimensions

- D=3 but average intrinsic dimension is 1.5
- Local intrinsic dimension is 1 & 2



k-NN results

Local dimension estimate $\hat{m}(i)$ of the three images



- Umbra, penumbra, and magnetic fragments have $\hat{m}(i)$ between 3 and 6
- Background has $\hat{m}(i)$ between 9 and 11
  - Stronger spatial and modal correlations in umbra, penumbra, and magnetic fragments

k-NN and PCA estimates of intrinsic dimension

k-NN Results

	Background		Penumbra		Umbra	
	mean	std	mean	std	mean	std
Single Spot	**8.9**	2.1	**4.5**	1.1	3.4	0.6
Multiple Spots	**8.6**	1.7	**4.8**	0.8	**4.0**	0.6

PCA results

	Background		Penumbra		Umbra	
	97%		97%		97%	
Single Spot	**10.1**		**4.3**		6.3	
Multiple Spots	**8.9**		**4.8**		**3.4**	

- Used twenty similar images for each type (single and multiple)
- 97% PCA threshold results are within 1 std of mean k-NN results for most regions
- Linear methods are likely sufficient

Unsupervised classification of the Images

- Approximately 500 pairs of sunspot group images available
- Unsupervised classification (clustering) attempts to group similar points (in this case, pairs of images) together
- Image Classification Steps:
  1. Form the data matrix for each image from pixel patches
  2. Learn a dictionary from each data matrix
  3. Cluster the learned dictionaries using a method well-suited for high dimensions (e.g. Galluccio et al, 2013)

Learning the Dictionary

- A dictionary is a set of basis vectors (dictionary elements) that can be used to accurately reconstruct the data

$$Z = \begin{bmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \dots & \mathbf{z}_N \end{bmatrix}$$
$$= \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_m \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \dots & \mathbf{h}_N \end{bmatrix} = AH$$

  - Z is $18 \times N$ data matrix, A is $18 \times m$ dictionary, and H is $m \times N$ coefficients matrix

- Example: PCA (linear method)
  - Principal components are the dictionary elements
  - Number of principal components m chosen to form the dictionary based on intrinsic dimension estimates

- The learned dictionaries form the data points to be classified

Clustering Algorithm

- Unsupervised classification groups together a set of objects s.t. objects within a group are more similar to each other than to those in other groups
- Some measure of pointwise similarity/dissimilarity is required
- Many methods exist (e.g. hierarchical clustering, kmeans, spectral clustering)
- We use Gallucio et al's method (2013) which is well adapted for finding nonlinearly separable groups
    - Inspired by the k-NN intrinsic dimension estimator

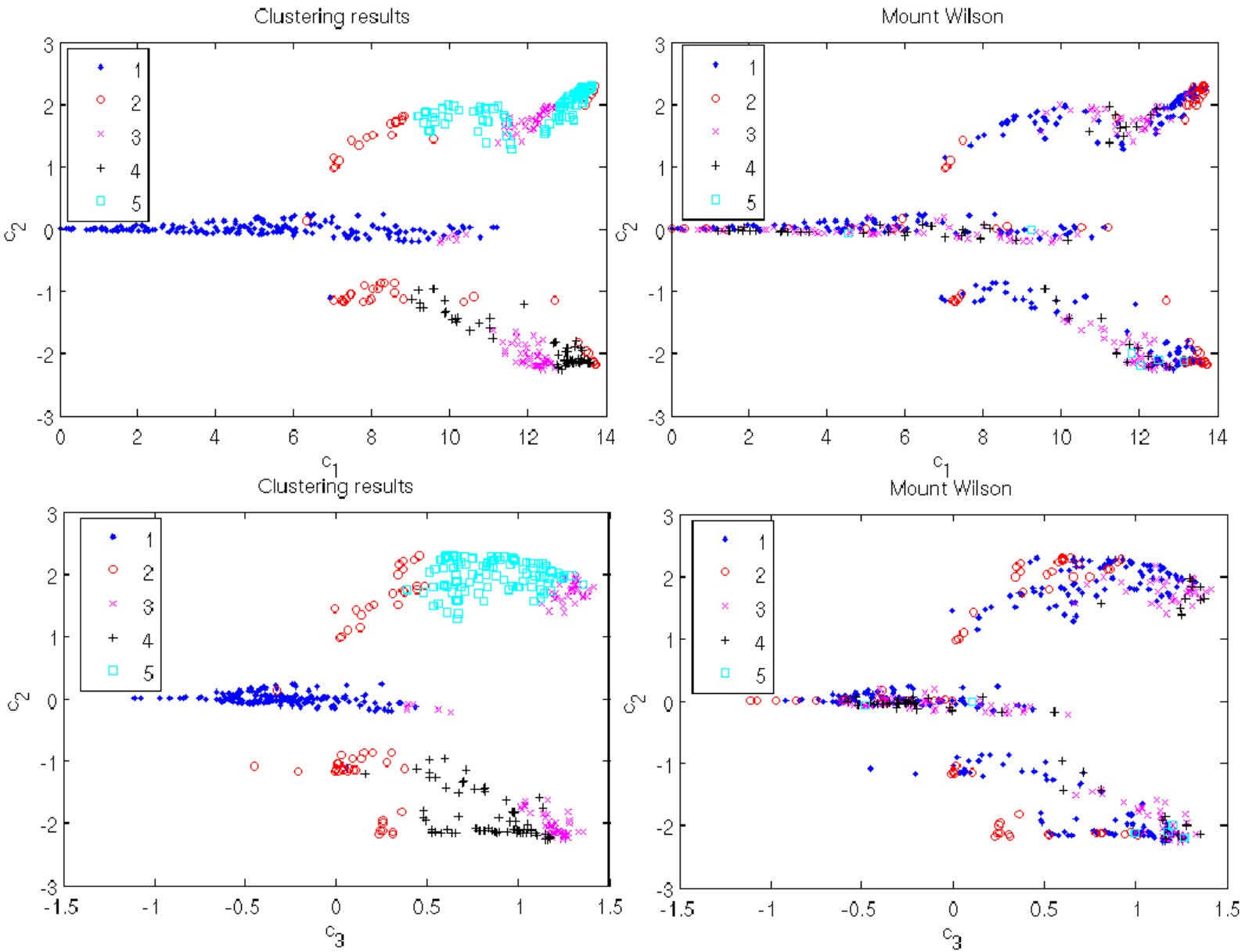
Our Experiment

1. Extracted the Region of Interest (ROI) from each image
   - 300×300 pixel image centered on the sunspot group
2. Chose 3×3 patch sizes
3. Learned the dictionaries using PCA
   - Chose the number of principal components to be 7
4. Clustered the dictionaries using Gallucio et al's method (2013)

Mt. Wilson Comparison

- Mt. Wilson classes: beta (1), alpha (2), beta-gamma (3), beta-gamma-delta (4), and beta-delta (5)
- Compared the clustered results to the Mt. Wilson labels
- Measures of correspondence (closer to 1 => better correspondence)
  - Normalized mutual information (NMI) = 0.11
  - Adjusted Rand Index (ARI) = 0.03
  - Consistent with local vs. global features
- Visualization of images in low dimension using Multidimensional Scaling (MDS, next slide)

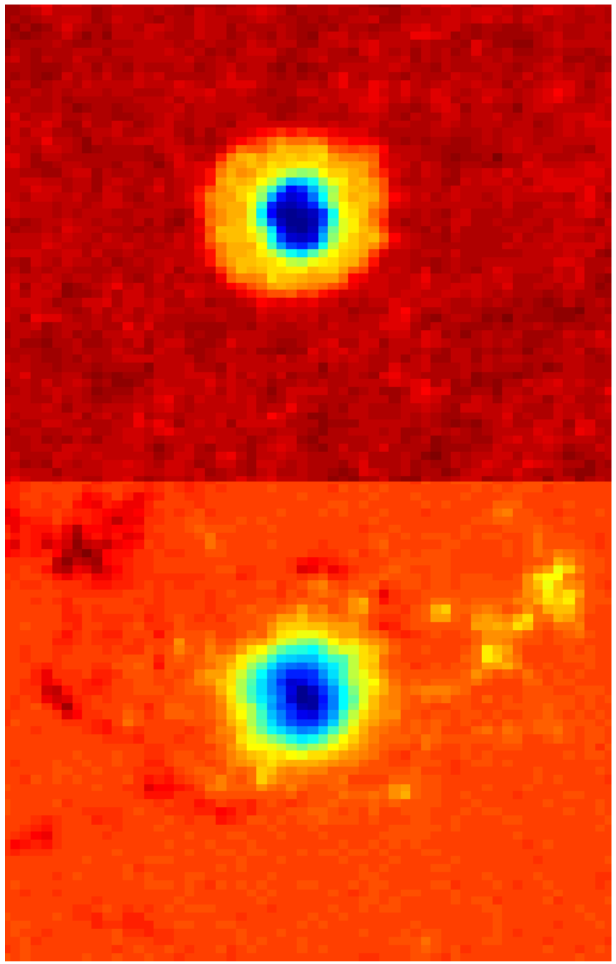
MDS Clustering Results



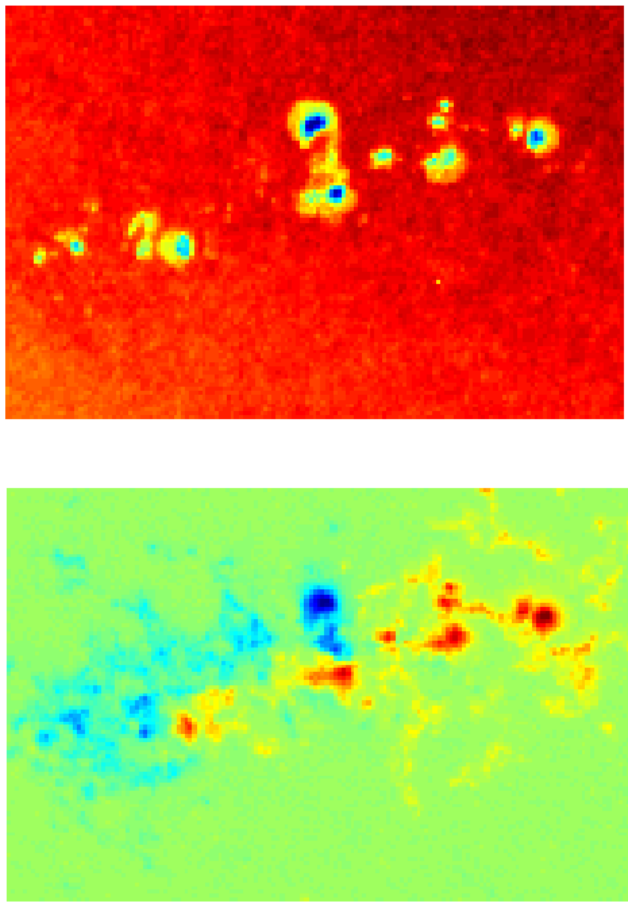
Interpreting the classification results

Some groups are correlated with physical features e.g. longitudinal extent

Cluster 2



Cluster 3



Conclusion

- Intrinsic dimension of the joint continuum and magnetogram patches suggests there are strong spatial and modal correlations in the sunspots and magnetic fragments
  - Suggests stronger spatial and modal correlations in these regions
- Linear decomposition methods (e.g. PCA) are likely sufficient
- At least three linearly separable groups of images result from our unsupervised classification approach based on local features
  - There is some physical interpretability of these clusters

Conclusion

- Intrinsic dimension of the joint continuum and magnetogram patches suggests there are strong spatial and modal correlations in the sunspots and magnetic fragments
  - Suggests stronger spatial and modal correlations in these regions
- Linear decomposition methods (e.g. PCA) are likely sufficient
- At least three linearly separable groups of images result from our unsupervised classification approach based on local features
  - There is some physical interpretability of these clusters
- Other questions answered (not presented today)
  1. What correlation exists between modalities and what spatial patterns produce that correlation?
  2. What phenomena exist at different scales within the images?

Future Directions

- Use the relationships and potential image features we have determined to better predict solar activity (e.g. flares)
  - Requires more data (including time series)
- Analyze the magnetic fragments more systematically
- Refine the image segmentation and feature extraction to better find image clusters
  - Adaptively define the ROI based on SMART masks (Higgins et al, 2011)
  - Include global and long range spatial features
- Anomaly detection approach
  - Treat each image as a distribution of points reconstructed from a common dictionary

For more details...

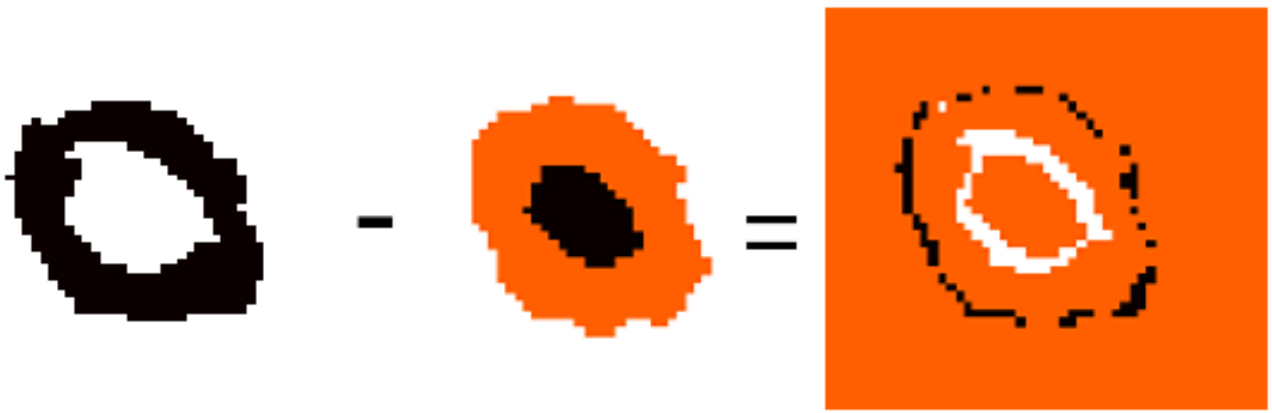
K.R. Moon, J.J. Li, V. Delouille, F. Watson, A.O. Hero, "Image patch analysis of sunspots: A dimensionality reduction approach." Available on arxiv, to appear in *IEEE International Conference of Image Processing* (ICIP) 2014.

k-NN graph length and intrinsic dimension

- $\mathbf{Z}_n$ is a set of n random vectors in $\mathbb{R}^D$, m is intrinsic dimension
- The k-NN graph length is $L_{k,\gamma}(\mathbf{Z}_n)$ =sum of power weighted (γ) lengths of edges
  - $0 < \gamma < m$
- For large n, $L_{k,\gamma}(\mathbf{Z}_n) = n^{\alpha(m)} c + \varepsilon_n$ (Costa and Hero, 2006)
  - $\alpha = (m-\gamma)/m$, $\varepsilon_n \to 0$ a.s. as $n \to \infty$, and c is a constant wrt n that depends on the Rényi entropy
- Intrinsic dimension m is found using non-linear least squares over different values of n

Comparison of k-NN to single sunspot mask (Backup)

Difference (right) between the penumbra mask (middle orange) and the level set of pixels i specified by $\{i : \hat{m}(i) = 4\}$ (left) for the single sunspot image.



- Discrepancy likely due to use of both mag and cont images in dimension estimation