Image patch analysis of sunspots and active regions. II. Clustering via dictionary learning

Kevin R. Moon\(^1\), Véronique Delouille\(^2\), Jimmy J. Li\(^1\), Ruben De Visscher\(^2\), Fraser Watson\(^3\), and Alfred O. Hero III\(^1\)

\(^1\) Electrical Engineering and Computer Science Department, University of Michigan
\(^2\) SIDC, Royal Observatory of Belgium
\(^3\) National Solar Observatory

**Abstract**

**Context.** Separating active regions that are quiet from potentially eruptive ones is a key issue in Space Weather applications. Traditional classification schemes such as Mount Wilson and McIntosh have been effective in relating an active region large scale magnetic configuration to its ability to produce eruptive events. However, their qualitative nature prevents systematic studies of an active region’s evolution for example.

**Aims.** We introduce a new clustering of active regions that is based on the local geometry observed in Line of Sight magnetogram and continuum images.

**Methods.** We use a reduced-dimension representation of an active region that is obtained by factorizing (i.e. applying dictionary learning to) the corresponding data matrix comprised of local image patches. Two factorizations can be compared via the definition of appropriate metrics on the resulting factors. The distances obtained from these metrics are then used to cluster the active regions.

**Results.** We find that these metrics result in natural clusterings of active regions. The clusterings are related to large scale descriptors of an active region such as its size, its local magnetic field distribution, and its complexity as measured by the Mount Wilson classification scheme. We also find that including data focused on the neutral line of an active region can result in an increased correspondence between the Mount Wilson classifications and our clustering results.

**Conclusions.** Dictionary learning is a promising new way of characterizing active regions. We provide some recommendations for which metrics and matrix factorization techniques to use to study small, large, complex, or simple active regions.

**Key words.** Sun – active region – sunspot – neutral line – data analysis – classification – clustering – image patches – Hellinger distance – Grassmannian

1. Introduction

Identifying properties of active regions (AR) that are necessary and sufficient for the production of energetic events such as solar flares is one of the key issues in space weather. The Mount Wilson classification (see Table 1 for a brief description of its four main classes) has been effective in
relating a sunspot’s large scale magnetic configuration with its ability to produce flares. Künzel (1960) pointed out the first clear connection between flare productivity and magnetic structure, and introduced a new magnetic classification, $\delta$, to supplement Hale’s $\alpha$, $\beta$ and $\gamma$ classes (Hale et al., 1919). Several studies showed that a large proportion of all major flare events begin with a $\delta$ configuration (Warwick, 1966; Mayfield and Lawrence, 1985; Sammis et al., 2000).

The qualitative nature of the Mount Wilson classification, however, prevents the differentiation between two sunspots with the same classification as well as the study of an AR’s evolution. The last decade has therefore seen many efforts in describing the photospheric magnetic configuration in more details. Typically, a set of scalar properties is derived from line of sight (LOS) or vector magnetogram and analyzed in a supervised classification context to derive which combination of properties is predictive of increased flaring activity (Leka and Barnes, 2004; Guo et al., 2006; Schrijver, 2007; Barnes et al., 2007; Georgoulis and Rust, 2007; Falconer et al., 2008; Song et al., 2009; Huang et al., 2010; Yu et al., 2010; Ahmed et al., 2011; Lee et al., 2012; Bobra and Couvidat, 2015). Examples of scalar properties include: sunspot area, total unsigned magnetic flux, flux imbalance, neutral line length, maximum gradients along the neutral line, or other proxies for magnetic connectivity within AR.

This work introduces a new way of analyzing information contained in magnetogram and continuum data. Instead of focusing on the best set of properties that summarizes the information contained in magnetogram and/or continuum images, we implement a reduced-dimension representation of such images via dictionary learning.

Reduced dimension representation has received a lot of attention in the general context of inverse problems. A signal can be decomposed against a set of basis functions and operations are made on the resulting transform coefficients to keep only a small number of them. For example, orthonormal wavelets (Daubechies, 1988) were proposed about three decades ago to decompose a one-dimensional signal. Reduced dimension representation is achieved in this context through shrinkage of wavelet coefficients (Donoho and Johnstone, 1994). The natural extension of orthonormal wavelets to a two-dimensional signal has been shown to be suboptimal, and various other tailored multiscale transforms were introduced to process images (Donoho, 1999; Candès and Donoho, 2002; Le Pennec and Mallat, 2005; Freeman and Adelson, 1991). In parallel the idea came about two decades ago to consider a set of redundant functions, called a dictionary, rather than an orthogonal (or close to orthogonal) basis in order to have more flexibility in the signal representation. Given a dictionary, it is possible to pick the most appropriate set of elementary functions to approx-

| Class | Classification Rule | Number of AR |
|-------|---------------------|--------------|
| $\alpha$ | A single dominant spot | 50 |
| $\beta$ | A pair of dominant spots of opposite polarity | 192 |
| $\beta\gamma$ | A $\beta$ sunspot where a single north-south polarity inversion line cannot divide the two polarities | 130 |
| $\beta\gamma\delta$ | A $\beta\gamma$ sunspot where umbrae of opposite polarity are together in a single penumbra | 52 |

Table 1. Mount Wilson classification rules and number of each AR per Mt. Wilson class used in this paper.
imate the signal within a given error (Mallat and Zhang, 1993). Finally, in the last decade, methods were proposed to learn the dictionary directly from the data (Elad and Aharon, 2006).

Given a $k \times n$ data matrix $Z$ containing $n$ observations of $k$ different variables, learning a linear dictionary from $Z$ amounts to factoring $Z$ as $Z = AH$ with $A$ the $k \times r$ dictionary and $H$ the $r \times n$ matrix of coefficients where $r < k$. In our application, we consider both the case where $Z$ is formed from a single image and when $Z$ is formed from multiple images. When a single image is used, the data matrix $Z$ is built from a $n$ pixel image by taking overlapping $m \times m$-pixel neighborhoods called patches. Figure 1 presents such a patch and its column representation. The $k$ rows of the $i$–th column of $Z$ are thus given by the $m^2$ pixel values in the neighborhood of pixel $i$. When multiple images are used, the patches are combined into a single data matrix. An example of a factorization of a data matrix containing image patches is given in Figure 2. In this figure, let $z_1$ be the first column of $Z$, containing the intensity values for the first patch. These intensity values are decomposed as a sum of $r$ elements:

$$z_1 = \sum_{j=1}^{r} a_j h_{j,1},$$

where $a_j$ is the $j$–th column of $A$ and $h_{j,1}$ is the $(j, 1)$–th element of $H$. In this representation, the vectors $a_j, j = 1, \ldots, r$ are the elementary building blocks common to all patches, whereas the $h_{j,1}$ are the coefficients specific to the first patch. Said otherwise, the $a_j$ (resp. $h_{j,1}$) are akin to the wavelet functions (resp. wavelet coefficients) in a wavelet decomposition.

To compare ARs and cluster them, some form of distance between ARs is required. To measure the distance between two ARs, we apply some metrics to the corresponding matrices $A$ or $H$ obtained from the factorizations of the two ARs. These distances are further introduced into a clustering algorithm that groups ARs based on the similarity of their patch geometry.

Such patch-based dictionary learning methods investigate the fine scale structures encoded by localized gradients of various directions and amplitudes, or locally smooth areas for example. In contrast, the Mount Wilson classification encodes the relative locations and sizes of concentrations of opposite polarity magnetic flux on a large scale. Although both classification schemes rely on completely different methods, we found some similarities (see Section 5). Moreover the Mount Wilson classification can guide us in the interpretation of dictionary learning results and clusters obtained. The shape of the neutral line separating the two main polarities in an AR is a key element in the Mount Wilson classification scheme, and the magnetic field gradients observed along the
Figure 2. An example of linear dimensionality reduction where the data matrix of AR image patches $Z$ is factored as a product of a dictionary $A$ of representative elements and the corresponding coefficients in the matrix $H$. The $A$ matrix consists of the basic building blocks for the data matrix $Z$ and $H$ contains the corresponding coefficients.

neutral line are important information in the quest for solar activity prediction (Schrijver, 2007; Falconer et al., 2008). We will therefore analyze the effect of including dictionary learning results from pixel patches situated along the neutral line in Section 5.

This paper builds upon results obtained in Moon et al. (2015), which can be summarized as follows:

1. Continuum and magnetogram modalities are correlated, and there may be some advantage in considering both of them in an analysis.

2. A patch size of $m = 3$ includes a significant portion of spatial correlations present in continuum and magnetogram images.

3. Linear methods for learning dictionaries are appropriate to analyze AR observed with continuum images and magnetogram.

With an AR area equal to $n$ pixels and an $m \times m$ patch, the corresponding continuum data matrix $X$ and magnetogram data matrix $Y$ each have size $m^2 \times n$. The full data matrix considered is $Z = \begin{pmatrix} X \\ Y \end{pmatrix}$ with size $2m^2 \times n = 18 \times n$ in our case. An intrinsic dimension analysis in Moon et al. (2015) showed further that a sunspot can be represented accurately with a dictionary containing six elements.

We consider the same dataset as in Moon et al. (2015). It is obtained from the Michelson Doppler Imager (MDI) instrument (Scherrer et al., 1995) on board the SOHO Spacecraft. SOHO-MDI provides two to four times per day a white-light continuum image observed in the vicinity of the Ni I 676.7 nm photospheric absorption line. MDI LOS magnetograms are recorded with a higher nominal cadence of 96 minutes. We selected 424 sunspot images within the time range of 1996-2010. They span the various Mount Wilson classifications (see Table 1), are located within 30° of central meridian, and have corresponding observations in both MDI continuum and MDI LOS magnetogram. We use level 1.8 data for both modalities. The Sunspot Tracking and Recognition Algorithm (STARA) (Watson et al., 2011) provides sunspot maps from continuum images. These masks encompass regions of highest variation observed in both continuum and magnetogram images. This is the reason why we choose to focus primarily on the portion of a continuum and magnetogram images delimited by the STARA mask for our dictionary learning analysis. Figure 3 provides some...
Figure 3. MDI continuum and magnetogram from NOAA 9097 on July 23, 2000 (top) and from NOAA 10045 on July 25, 2002 (bottom) overlaid with the corresponding STARA masks in green.

examples of AR images overlaid with their respective STARA masks. When specifically analyzing patches located along the neutral lines however, we will use the boundary of the magnetic AR as determined by the Solar Monitor Active Region Tracker (SMART) masks (Higgins et al., 2011).

Section 2 describes two linear dictionary learning methods, which are akin to matrix factorization. We use the singular value decomposition (SVD) and nonnegative matrix factorization (NMF). To compare the results from this factorization we need some metrics, and so we use the Grassmanian projection metric and the Hellinger distance for this purpose. To obtain some insight on how these factorizations separate the data, we make some general comparisons in Section 3. In particular, with the defined metrics, we compute the pairwise distances between Mount Wilson classes to identify which classes are most similar or dissimilar according to the dictionary learning results.
Section 4 describes the clustering procedures that take the metrics’ output as input. The method called ‘Evidence Accumulating Clustering with Dual rooted Prim tree Cuts’ (EAC-DC) was introduced by Galluccio et al. (2013) and is used to cluster the ARs. By combining the two dictionary learning methods and two metrics, a total of four procedures are used to analyze the data. Besides analyzing the whole sunspot data, we also look at information contained in patches situated along the neutral lines. The results of the clustering analyses are provided in Section 5.

This paper improves and extends the work in Moon et al. (2014), where fixed size square pixel regions centered on the sunspot group were used as input to the dictionary learning methods prior to clustering. Moreover, here we are using more appropriate metrics to compare the dictionary results.

2. Dictionary Learning

The intrinsic dimension analysis in Moon et al. (2015) showed that linear methods are sufficient to represent the data, and hence we focus on those. Linear dictionary learning methods aim at finding a set of basis vectors or dictionary elements such that each data point (in our case, pair of pixel patches) can be accurately expressed as a linear combination of the dictionary elements. Mathematically, if we use \( m \times m \) patches then this can be expressed as \( Z \approx AH \), where \( Z \) is the \( 2m^2 \times n \) data matrix with \( n \) data points being considered, \( A \) is the \( 2m^2 \times r \) dictionary with the columns corresponding to the dictionary elements, and \( H \) is the \( r \times n \) matrix of coefficients. The goal is to find matrices \( A \) and \( H \) whose product nearly approximates \( Z \). The degree of approximation is typically measured by the squared error \( \|Z - AH\|_F^2 \), where \( \| \cdot \|_F \) denotes the Frobenius norm (Yaghoobi et al., 2009). Additional assumptions on the structure of the matrices \( A \) and \( H \) can be applied in dictionary learning depending on the application. Examples include assumptions of orthonormality of the columns of the dictionary \( A \), sparsity of the coefficient matrix \( H \) (Ramírez and Sapiro, 2012), and nonnegativity on \( A \) and \( H \) (Lin, 2007).

We consider two popular dictionary learning methods: the singular value decomposition (SVD) and nonnegative matrix factorization (NMF).

2.1. Factorization using SVD

To perform dictionary learning using SVD, we take the singular value decomposition of the data matrix \( Z = U\Sigma V^T \) where \( U \) is the matrix of the left singular vectors, \( \Sigma \) is a diagonal matrix containing the singular values, and \( V \) is a matrix of the right singular vectors. If the size of the dictionary \( r \) is fixed and is less than \( 2m^2 \), then the matrix of rank \( r \) that is closest to \( Z \) in terms of the Frobenius norm is the matrix product \( U_r\Sigma_r V_r^T \), where \( U_r \) and \( V_r \) are matrices containing only the first \( r \) singular vectors and \( \Sigma_r \) contains only the first \( r \) singular values (Moon and Stirling, 2000). Thus for SVD, the dictionary and coefficient matrices are \( A = U_r \) and \( H = \Sigma_r V_r^T \), respectively. Note that SVD enforces orthonormality on the columns of \( U_r \). Further details are included in Appendix A.1.

The intrinsic dimension estimated in Moon et al. (2015) determines the number of parameters required to accurately represent the data. It is used to provide an initial estimate for the size of the dictionaries \( r \). For SVD, we choose \( r \) to be one standard deviation above the mean intrinsic dimension estimate, that is, \( r \approx 5 \) or 6. The choice of \( r \) is then further refined by a comparison of dictionaries in Section 3. See Section 4.2 for more on selecting the dictionary size.
Figure 4. Learned dictionary elements using SVD. Dictionary elements are constrained to be orthonormal. The patches consist of uniform patches and gradients in varied directions. The magnetogram patches are essentially zero when the continuum components are nonzero and vice versa. The dictionary size \( r \) is first chosen based on the intrinsic dimension estimates in Moon et al. (2015) and then refined by comparing dictionaries of various sizes in Section 3. Section 4.2 contains more details on choosing \( r \).

Figure 4 shows the learned dictionaries using SVD on the entire data set of 424 image pairs of ARs. Interestingly, the SVD seems to consider the continuum and magnetogram separately as the magnetogram elements are essentially zero when the continuum elements are not and vice versa. This is likely caused by the orthonormality constraint. The dictionary patches largely consist of a mix of uniform patches and patches with gradients in varied directions. The second dictionary element is associated with the average magnetic field value of a patch.

2.2. Factorization using NMF

Non-negative matrix factorization (NMF) (Lee and Seung, 2001) solves the problem of minimizing \( \|Z - AH\|^2_F \) while constraining \( A \) and \( H \) to have nonnegative values. Thus NMF is a good choice for dictionary learning when the data is nonnegative. For our problem, the continuum data is nonnegative while the magnetogram data is not. Therefore we use a modified version of NMF using projected gradient where we only constrain the parts of \( A \) corresponding to the continuum to be nonnegative. An effect of using this modified version of NMF is that since the coefficient matrix \( H \) is still constrained to be nonnegative, we require separate dictionary elements that are either positive or negative in the magnetogram component. Thus we use approximately 1.5 times more dictionary elements for NMF than SVD.

Figure 5 shows the learned dictionary elements using NMF on the entire dataset. For NMF, the modalities are not treated separately as in the SVD results. But as for SVD, the patches largely consist of a mix of uniform patches and patches with gradients in varied directions.

2.3. SVD vs. NMF

There are advantages to both SVD and NMF dictionary learning methods which are summarized in Table 2. SVD produces the optimal rank \( r \) approximation of \( Z \), is fast and unique, and results in orthogonal elements (Langville et al., 2006). NMF has the advantages of nonnegativity, sparsity, and interpretability. The interpretability comes from the additive parts-based representation inherent in NMF (Langville et al., 2006). In contrast, the SVD results are not sparse which can make interpretation more difficult. However, NMF is not as robust as SVD as the NMF algorithm is a numerical approximation to a non-convex optimization problem having local minima. Thus the
Learned dictionary elements using NMF where the continuum dictionary elements are constrained to be nonnegative. All the dictionary patches consist of uniform patches or gradients in varied directions. The order of the elements is not significant. The dictionary size \( r \) is chosen to be approximately 1.5 times larger than the SVD dictionary size, which is chosen based on the intrinsic dimension estimates in Moon et al. (2015) and then refined using the results in Section 3.

**Table 2.** Summary of the advantages of SVD and NMF dictionary learning methods. The advantages of one method complement the disadvantages of the other (Langville et al., 2006). For example, the NMF optimization problem is nonconvex with local minima resulting in solutions that depend on the initialization of the algorithm.

| SVD Advantages | NMF Advantages |
|----------------|----------------|
| Optimal rank \( r \) approximation | Results are nonnegative |
| Fast to compute | Results are sparse |
| Unique | Sparsity and nonnegativity lead to improved interpretability |

solution provided by the NMF algorithm depends on the initialization. More details on dictionary learning using NMF and SVD are included in Appendix A.1.

2.4. Methods for Comparing Dictionary Learning Results

To compare the results from dictionary learning, we may either seek a difference between the dictionaries \( \mathbf{A} \) estimated on different ARs, or between the coefficients from \( \mathbf{H} \).

In Moon et al. (2014), the Frobenius norm was used to compare the dictionaries in the clustering step. However, this fails to take into account the fact that two dictionaries may have the same elements but in a different order. In this case, the Frobenius norm of the difference between two dictionaries may be high even though the dictionaries span the same subspace. A better way to measure the difference would be to compare the subspaces spanned by the dictionaries. The Grassmannian \( \text{Gr}(r, V) \) is a space which parameterizes all linear subspaces with dimension \( r \) of a vector space \( V \). As an example, the Grassmannian \( \text{Gr}(2, \mathbb{R}^n) \) is the space of planes through the origin of the standard Euclidean vector space in \( \mathbb{R}^n \). In our case, we are concerned with the Grassmannian \( \text{Gr}(r, \mathbb{R}^{18}) \), where \( r \) is the size of the dictionary. The space spanned by a given dictionary \( \mathbf{A} \) is then a single point in \( \text{Gr}(r, \mathbb{R}^{18}) \). Several metrics have been defined on this space including the Grassmannian projection metric (Stewart, 1973; Edelman et al., 1998). It can be defined as

\[
d_G(\mathbf{A}, \mathbf{A}') = \| \mathbf{P}_\mathbf{A} - \mathbf{P}_{\mathbf{A}'} \|,
\]
where $P_A = A (A^T A)^{-1} A^T$ is the projection matrix of $A$ and $\|\cdot\|$ is the $\ell_2$ norm. This metric is invariant to the order of the dictionary elements and compares the subspaces spanned by the dictionaries. This metric has a maximum value of 1.

To compare the coefficient matrices, we assume that the 18-dimensional pixel patches within an AR are samples from an 18-dimensional probability distribution, and that each AR has a corresponding (potentially unique) probability distribution of pixel patches. We project these samples onto a lower dimensional space by dictionary learning. In other words, we learn a dictionary $A$ and the coefficient matrix $H$ for the entire dataset $Z$, and then separate the coefficients in $H$ according to the $K$ different ARs (or groups of ARs) considered: $Z = A \left( H_1 \ H_2 \ldots H_K \right)$. The columns of $H_i$ are a collection of projected, low-dimensionality, samples from the $i$th AR (or group), and we let $f_i$ denote the corresponding probability density function. Given two such collections, we can estimate the difference between their probability distributions by estimating the information divergence. Many kinds of divergences exist such as the popular Kullback-Leibler divergence (Kullback and Leibler, 1951). We use the Hellinger distance which is defined as (Hellinger, 1909; Bhattacharyya, 1946; Csiszar, 1967)

$$H(f_i, f_j) = 1 - \int \sqrt{f_i(x)f_j(x)} \, dx,$$

where $f_i$ and $f_j$ are the two probability densities being compared. The Hellinger distance has the advantage over other divergences of being a metric which is not true of divergences in general. To estimate the Hellinger distance, we use the nonparametric estimator derived in Moon and Hero III (2014a,b) that is based on the $k$-nearest neighbor density estimators for the densities $f_i$ and $f_j$. This estimator is simple to implement and achieves the parametric convergence rate when the densities are sufficiently smooth.

3. Comparisons of General Dictionary Learning Results

We apply the metrics mentioned in Section 2.4 and compare the local features as extracted by dictionary learning per Mount Wilson class. One motivation for these comparisons is to investigate differences between the Mount Wilson classes based on the Grassmannian metric and the Hellinger distance. Another motivation is to further refine our choice of dictionary size $r$ in preparation for clustering the ARs. When using the Grassmannian metric to compare dictionaries, we want the dictionaries to be different to reveal differences between ARs. Otherwise, if the dictionaries all look the same as measured by the Grassmannian metric, then the most natural clustering would be a single cluster containing all ARs which would be uninformative. In contrast, when comparing the dictionary coefficients using the Hellinger distance, we want a single, representative dictionary that is able to accurately reconstruct all of the images. Then the ARs will be differentiated based on their respective distributions of dictionary coefficients instead of the accuracy of their reconstructions. The coefficient distributions can then be compared to interpret the clustering results as is done in Section 5.3.

3.1. Grassmannian Metric Comparisons

For the Grassmannian metric, we first learn dictionaries for each of the Mount Wilson types by applying dictionary learning to all of the patches corresponding to sunspot groups of the respective
type. We then use the Grassmannian metric to compare the dictionaries. For example, if we want to compare the $\alpha$ and $\beta$ groups, we collect the patches from all AR designated as $\alpha$ groups into a single data matrix $Z_{\alpha}$. We then factor this matrix with the chosen method to obtain $Z_{\alpha} = A_{\alpha}H_{\alpha}$. Similarly, we obtain $Z_{\beta} = A_{\beta}H_{\beta}$ and then calculate $d_G(A_{\alpha}, A_{\beta})$.

Table 3 shows the corresponding Grassmannian distance metrics when using SVD and NMF and for different sizes of dictionaries $r$. For SVD, the results are very sensitive to $r$. Choosing $r = 5$ results in large differences between the different dictionaries but for $r = 6$, the dictionaries are very similar. This suggests that for SVD, 6 principal components are sufficient to accurately represent the subspace upon which the sunspot patches lie. This is consistent with the results of Moon et al. (2015) where the intrinsic dimension is found to be less than 6 for most patches.

The NMF results are less sensitive to $r$. The difference between the dictionaries is larger when $r = 9$ compared to when $r = 8$. However, the relative differences are approximately the same for both sizes. For example, the $\alpha$ and $\beta\gamma$ groups are most dissimilar while the $\beta$ and $\beta\gamma\delta$ groups are most similar for both $r = 8$ or 9. Pooling the image patches and then comparing the NMF dictionaries in this manner also seems to have some correlation with complexity as the $\alpha$ dictionary is more similar to the $\beta$ dictionary than the others. Interestingly, the $\beta\gamma$ dictionary is most dissimilar from the others.

The presence or lack of dissimilarity between classes when considering the patches in aggregate format is not necessarily predictive of clustering performance. For clustering, we compare the individual ARs to each other. To do this using the Grassmannian metric, we first form the data matrices $Z_i$ for $i = 1, 2, \ldots, 424$ from each AR’s respective patches. We then factor each of these matrices as $Z_i = A_iH_i$, and finally in order to compare the $i$th and $j$th AR using the Grassmannian, we calculate $d_G(A_i, A_j)$.

Table 4 shows the average pairwise Grassmannian distance between the different Mount Wilson types. In all cases, the average distances are high (> 0.7) and are very similar using either SVD or NMF. From the table, we see that the average distance is smallest when comparing the $\beta\gamma\delta$ groups to all others. A $\beta\gamma\delta$ AR contains complex local features in addition to features that are common to simpler sunspots. Thus a dictionary that can accurately reconstruct a $\beta\gamma\delta$ AR may be able to better reconstruct a simpler AR than vice versa. However, the average distance between the dictionaries of $\beta\gamma\delta$ groups and the simple groups ($\alpha$ and $\beta$) is not that much lower than the average difference between the simple groups’ dictionaries. Thus the advantage is small in this respect. In contrast, the average distance between dictionaries of complex groups ($\beta\gamma$ and $\beta\gamma\delta$) is smaller than
Table 4. Average pairwise difference between dictionaries learned for each AR from different Mount Wilson types as measured by the Grassmannian metric $d_G$. The size of the dictionaries is $r = 5$ and $8$ for SVD and NMF, respectively. The $\alpha$ ARs are most dissimilar to each other and the other classes while the $\beta\gamma\delta$ are most similar.

|       | SVD, Average Grassmannian | NMF, Average Grassmannian |
|-------|---------------------------|---------------------------|
| $\alpha$ | 0.87 0.88 0.85 0.85       | 0.90 0.91 0.86 0.85       |
| $\beta$  | 0.88 0.85 0.84            | $\beta$ 0.91 0.86 0.85    |
| $\beta\gamma$ | 0.78 0.76                | $\beta\gamma$ 0.78 0.76   |
| $\beta\gamma\delta$ | 0.75                     | $\beta\gamma\delta$ 0.75   |

the average distance between dictionaries of simple groups. Thus the Grassmannian metric appears to distinguish simple groups more than complex groups. This is explored further in Section 5.

The class that is most dissimilar to itself and to the other types is the $\alpha$ class. This makes sense considering the polarity of the class. Each $\alpha$ AR is dominated strongly by a single polarity in the magnetogram. Thus a group with a single polarity will have a different dictionary than a group that has only the opposite polarity or both polarities. Some of the differences may also be influenced by size since $\alpha$ ARs tend to be smaller than the other ARs. This is explored more in Section 5.2.

3.2. Hellinger Distance Comparisons

For the Hellinger distance, we learn a dictionary $A$ and the coefficient matrix $H$ for the entire data set $Z$. We then separate the coefficients in $H$ according to the Mount Wilson type and compare the coefficients’ distributions using the Hellinger distance. For example, suppose that the data matrix is arranged as $Z = (z_{\alpha}, z_{\beta}, z_{\beta\gamma}, z_{\beta\gamma\delta})$. This is factored as $Z = A (H_{\alpha}, H_{\beta}, H_{\beta\gamma}, H_{\beta\gamma\delta})$. To compare the $\alpha$ and $\beta$ groups, we assume that the columns in $H_{\alpha}$ are samples from the distribution $f_{\alpha}$ and similarly $H_{\beta}$ contains samples from the distribution $f_{\beta}$. We then estimate the Hellinger distance $H(f_{\alpha}, f_{\beta})$ using the divergence estimator in Moon and Hero III (2014a).

When the Hellinger distance is used to compare the collections of dictionary coefficients within the sunspots, the groups are very similar, especially when using SVD (Table 5). This indicates that when the coefficients of all ARs from one class are grouped together, the distribution looks similar to the distribution of the other classes. However, there are some small differences. For both dictionary learning methods, the $\beta\gamma\delta$ groups are the most dissimilar. This could be due to the presence of a $\delta$ spot configuration, where umbrae of opposite polarities are within a single penumbra. Such a configuration may require specific linear combinations of the dictionary elements as compared to the other classes. The presence and absence of these linear combinations in two Mount Wilson types would result in a higher Hellinger distance between them.

Again, for clustering, we compute the pairwise Hellinger distance between each AR’s collection of coefficients. This is done by forming the data matrix from the 424 ARs as $Z = (Z_1, Z_2, \ldots, Z_{424})$ and factoring it as $Z = A (H_1, H_2, \ldots, H_{424})$. The columns of $H_i$ are samples from a distribution $f_i$ and the distributions $f_i$ and $f_j$ are compared by estimating $H(f_i, f_j)$. The corresponding dictionaries for the two methods are shown in Figures 4 and 5.
**4. Clustering of Active Regions: Methods**

**4.1. Clustering Algorithm**

The clustering algorithm we use is the Evidence Accumulating Clustering with Dual rooted Prim tree Cuts (EAC-DC) method in Galluccio et al. (2013) which scales well for clustering in high dimensions. EAC-DC clusters the data by defining a metric based on the growth of two minimal spanning trees (MST) grown sequentially from a pair of points. To grow the MSTs, a base dissimilarity measure is required as input such as those described in Section 2.4. From the new metric defined using the MSTs, a similarity measure between inputs is created. It is fed into a spectral clustering algorithm that groups together inputs which are most similar. The similarity measure based on the MSTs adapts to the geometry of the data, and this results in a clustering method that is robust and competitive with other algorithms (Galluccio et al., 2013). See Appendix A.2 for more details.
Table 7. Summary of the dissimilarity and similarity measures used.

| Measure                  | Type     | Properties                                                                 |
|--------------------------|----------|-----------------------------------------------------------------------------|
| Grassmannian metric      | Dissimilarity | Compares dictionaries by comparing the subspace spanned by the dictionary elements |
| Hellinger distance       | Dissimilarity | Compares the underlying distributions of dictionary coefficients; estimated using Moon and Hero III (2014a) |
| EAC-DC based measure     | Similarity | Based on sequentially grown MSTs of the data; requires a base dissimilarity measure as input |

4.2. Clustering Input: Dictionary Sizes

As input to the clustering algorithm, we use the dictionary learning results as described in Section 2.4. For the first approach, we learn a dictionary $A_i$ for each AR as in Moon et al. (2014). The dictionaries are the inputs in this method and the base dissimilarity measure is the Grassmannian metric. In the second approach, we learn a single dictionary from the entire dataset. We then project the data onto a lower dimensional space, i.e. we learn the coefficient matrices $H_i$. These matrices are the inputs in this method and the base dissimilarity measure is the Hellinger distance estimated using each AR’s respective coefficients. Table 7 provides a summary of the various dissimilarity and similarity measures that we use.

As mentioned in Section 2, the estimated intrinsic dimension from Moon et al. (2015) is used to provide an initial estimate for the size of the dictionaries $r$. The choice of $r$ is further refined by the dictionary comparison results from Section 3. For SVD, we choose $r$ to be one standard deviation above the mean intrinsic dimension estimate which is approximately 5 or 6. When comparing the dictionaries with the Grassmannian, we want the dictionaries to be different to reveal differences between ARs. In contrast, when comparing the dictionary coefficients, we want the single dictionary to accurately represent the images. In both cases, the dictionaries should not be too large as this may add spurious dictionary elements due to the noise. The results in Table 3 suggest that for SVD, the dictionaries are essentially identical for $r = 6$. This means that 6 dictionary elements are sufficient to accurately reconstruct most of the images. This is consistent with the intrinsic dimension estimates in Moon et al. (2015). Thus we choose $r = 6$ when using the Hellinger distance for the SVD dictionary coefficients. For the Grassmannian metric where the SVD dictionaries are the input, we choose $r = 5$ to maintain differences between the dictionaries.

Since our mixed version of NMF requires approximately 1.5 times the number of dictionary elements as SVD (see Section 2.1), we choose $r = 8$ within the sunspots. This strikes a balance between accurate representation of the data, difference between the dictionaries, and limiting the effects of noise.

4.3. Clustering Input: Patches within Sunspots and Along the Neutral Line

Our main focus in this paper is on data matrices $Z$ containing the patches within the STARA masks, that is, within sunspots. The clustering based on these patches is discussed in Sections 5.1-5.4.

It is well-known, however, that the shape of the neutral line separating the main polarities plays an important role in the Mt Wilson classification. For this reason, we conduct in Section 5.5 an
Table 8. The labels used to compare with the clustering results when analyzing the effects of including the neutral line.

| # of clusters | Mount Wilson Comparison |
|----------------|-------------------------|
| 2              | Simple ($\alpha$ and $\beta$), complex ($\beta\gamma$ and $\beta\gamma\delta$) |
| 3              | $\alpha$, $\beta$, and complex |
| 4              | Mount Wilson ($\alpha$, $\beta$, $\beta\gamma$, and $\beta\gamma\delta$) |

experiment where we perform dictionary learning on a data matrix containing only the patches situated along the neutral line and belonging to the SMART masks.

To compute the location of a neutral line, we assume it is situated in the middle between regions of opposite polarity, and proceed as follows. First, we determine regions of high magnetic flux of each polarity using an absolute threshold at 50 Gauss. Second, we compute for each pixel the distance to the closest high flux region in each polarity. We do this with the Fast Marching method (Sethian, 1995). Once the two distance fields (one for each polarity) are calculated, the neutral line can be obtained by finding the pixels that lie on or are close to the zero level set of the difference of these two distance fields. In this paper, we choose a maximum distance of 10 pixels to determine the neutral line region.

We extract the patches that lie in the neutral line region and within the SMART mask associated to the AR. Call the resulting data matrix $Z_N$. We then apply SVD and NMF dictionary learning as before and calculate the pairwise distance between each AR neutral line using either the Hellinger distance or the Grassmannian metric on the results. Define the resulting $424 \times 424$ (Grassmannian or Hellinger) dissimilarity matrix as $D_N$ for whatever configuration we are currently using, e.g. the Grassmannian metric on SVD dictionaries. Similarly, define $Z_S$ and $D_S$ as the respective data matrix and dissimilarity matrix of the data from within the sunspots using the same configuration. The base dissimilarity measure $D$ inputted in the clustering algorithm is now a weighted average of the (Grassmannian or Hellinger) distances computed within the neutral line regions and within the sunspots: $D = wD_N + (1 - w)D_S$ where $0 \leq w \leq 1$. Using a variety of weights, we then compare the clustering output to different labeling schemes based on the Mount Wilson labels as shown in Table 8.

5. Clustering of Active Regions: Results

Given the choices of dictionary learning techniques (NMF and SVD) and base dissimilarity measures (Grassmannian metric and Hellinger distance), we have four different clustering results on the data within the sunspots. These results are the main focus of this paper. Section 5.1 provides some general observations on these four clustering schemes. Sections 5.2 and 5.3 focus on the Grassmannian and Hellinger distance based clusterings, respectively, and Section 5.4 provides some recommendations for which metrics and matrix factorization techniques to use to study different ARs. The neutral line clustering results are then given in Section 5.5.
Table 9. Adjusted Rand Index between different clustering results when using the patches contained within the STARA masks. Results from SVD and NMF are most similar when using the same metrics.

5.1. Generalities

The EAC-DC algorithm does not automatically choose the number of clusters. We use the mean silhouette heuristic to determine the most natural number of clusters as a function of the data (Rousseeuw, 1987). The silhouette is a measure of how well a data point belongs to its assigned cluster. The heuristic chooses the number of clusters that results in the maximum mean silhouette. In three of the four clustering configurations, the number of clusters that maximize the mean silhouette is 2. Under the SVD Grassmannian scheme, ten clusters maximize the mean silhouette although it is still high when two clusters are chosen. For this reason, we focus on the two clustering case for all clustering schemes throughout.

To compare the clustering results to the labels, we use the adjusted Rand index (ARI). The ARI is a measure of similarity between two clusterings (or a clustering and labels) and takes values between -1 and 1. A 1 indicates perfect agreement between the clusterings and a 0 corresponds to the agreement from a random assignment (Rand, 1971). Thus a positive ARI indicates that the clustering correspondence is better than a random clustering while a negative ARI indicates it is worse.

Based on the ARI, the clustering results are all generally different from each other as seen in Table 9. The most similar results are obtained when comparing SVD and NMF with the Grassmannian metric (ARI= 0.20) or with the Hellinger distance (ARI= 0.27). All other clustering pairs have less correspondence.

Visualizing the clusters in lower dimensions is done with multidimensional scaling (MDS) as in Moon et al. (2014). Let $S$ be the $424 \times 424$ symmetric matrix that contains the AR pair similarities as created by EAC-DC algorithm. MDS projects the similarity matrix $S$ onto the eigenvectors of the normalized Laplacian of $S$ (Kruskal and Wish, 1978). Let $c_i \in \mathbb{R}^{424}$ be the projection onto the $i$th eigenvector of $S$ derived under the NMF Grassmannian scheme. The first eigenvector represents the direction of highest variability in the matrix $S$, and hence a high value of the $k$–th element of $c_1$ indicates that the $k$–th AR is dissimilar to other ARs.

Figure 6 displays the scatter plot of $c_1$ vs. $c_2$ (top) and $c_1$ vs. $c_3$ (bottom). In this representation, the clusters under the NMF Grassmannian scheme are separable. Comparing them with the Mount Wilson classification we see a concentration of simple ARs in the region with highest $c_1$ values (most dissimilar AR), and a concentration of complex ARs in the region with lowest $c_1$ (more similar AR). We can show this more precisely by computing the mean similarity of the $i$th AR to all other ARs as the mean of the $i$th row (or column) of $S$. The value from $c_i$ is then inversely related to this mean similarity as seen in Figure 7.
Figure 6. Scatter plot of MDS variables $c_1$ vs. $c_2$ (top) and $c_1$ vs. $c_3$ (bottom) where $c_i \in \mathbb{R}^{424}$ is the projection of the similarity matrix onto the $i$th eigenvector of the normalized Laplacian of the similarity matrix when using the Grassmannian distance between the NMF dictionaries. Each point corresponds to one AR and they are labeled according to the clustering (left) and the Mount Wilson labels (right). In this space, the clusters data appear to be separable and there are concentrations of complex ARs in the region with lowest $c_1$ values. Other patterns are described in the text.

The similarity defined under this clustering scheme gathers in Cluster 1 ‘similar’ AR that are for a large part of the type $\beta\gamma$ and $\beta\gamma\delta$, whereas Cluster 2 contains AR that are more ‘dissimilar’ to each other, with a large part of $\alpha$ or $\beta$ active regions. The other clustering configurations have similar relationships between the first MDS coefficient and the mean similarity.

Table 10 makes this clearer by showing the mean similarity measure within each cluster and between the two clusters. This can be calculated in the following manner. Suppose that the similarity matrix is organized in block form where the upper left block corresponds to Cluster 1 and the lower right block corresponds to Cluster 2. The mean similarity of Cluster 1 is calculated by taking the mean of all the values in the upper left block of this reorganized similarity matrix. The mean similarity of Cluster 2 is found similarly from the lower right block and the mean similarity between the clusters is found from either the lower left or upper right blocks. These means show that under the NMF Grassmannian clustering scheme, ARs in Cluster 1 are very similar to each other while ARs in Cluster 2 are not very similar to each other on average. In fact, the ARs in Cluster 2 are more similar to the ARs in Cluster 1 on average than to each other. The other clustering configurations have similar relationships between cluster assignment and mean similarities although it is flipped for the Hellinger based configurations.
Moon et al: Clustering of active regions

Figure 7. Mean similarity of an AR with all other ARs as a function of its MDS variable \( c_1 \) under the NMF Grassmannian scheme. Cluster 1 is associated with those ARs that are most similar to all other ARs while Cluster 2 contains those that are least similar to all others.

|          | 1 vs. 1 | 1 vs. 2 | 2 vs. 2 |
|----------|---------|---------|---------|
| SVD, Grass. | 0.85    | 0.40    | 0.28    |
| NMF, Grass. | 0.88    | 0.42    | 0.29    |
| SVD, Hellinger | 0.29    | 0.42    | 0.87    |
| NMF, Hellinger | 0.30    | 0.42    | 0.88    |

Table 10. Mean similarity of ARs to other ARs either in the same cluster (1 vs. 1 or 2 vs. 2) or in the other cluster (1 vs. 2) under the different schemes. For the Grassmannian schemes, Cluster 1 contains ARs that are very similar to each other while Cluster 2 contains ARs that are very dissimilar to each other. Under the Hellinger distance schemes, these relationships are switched.

5.2. Clustering Within Sunspots: Grassmannian metric

All clustering configurations result in clusters containing ARs from different Mount Wilson classes. Figure 8 compares histograms of the Mt. Wilson classes divided by cluster assignment for the Grassmannian configurations. With SVD dictionaries, Cluster 2 contains more \( \beta \) and \( \alpha \) groups but fewer \( \beta\gamma \) and \( \beta\gamma\delta \) groups than Cluster 1. With NMF dictionaries, the trends are similar except the differences are slightly more pronounced. These patterns can also be observed in the MDS space for the NMF case in Figure 6. The ARI for the clustering results compared to a simple vs. complex labeling scheme is 0.10 and 0.13 for SVD and NMF, respectively. This suggests that the correspondence is better than a random clustering but not particularly strong.

Table 11 reveals that total sunspot size is correlated with the clustering under both configurations. Cluster 2 contains most of the small sunspot groups while Cluster 1 contains most of the large sunspot groups. This split is especially strong for the simple groups and is less strong for the complex groups.

When comparing the projection matrices \( P_A \) of the dictionaries for all ARs assigned to Cluster 1, we see that the portion of the \( P_A \)'s corresponding to the magnetogram data all have a similar structure. The continuum portion of the \( P_A \)'s is also structured but several different patterns exist within Cluster 1. In contrast, the projection matrices of ARs assigned to Cluster 2 tend to be less
Figure 8. Histograms of the Mt. Wilson classes divided by clustering assignments using the Grassmannian metric. Cluster 1 contains more of the complex groups while Cluster 2 contains more of the simple groups.

Table 11. Mean and median number of pixels of the ARs in each cluster under the Grassmannian clustering schemes. Cluster 2 contains more of the small ARs using both SVD and NMF dictionaries.

| Cluster                 | Number of Pixels | SVD, Grassmannian | NMF, Grassmannian |
|-------------------------|------------------|-------------------|-------------------|
|                         | $\alpha$ | $\beta$ | $\beta\gamma$ | $\beta\gamma\delta$ | All |
| Mean, SVD Grassmannian | 465     | 190     | 754           | 177               | 634  | 1395 | 1043 | 854 | 340 |
| Mean, NMF Grassmannian | 416     | 215     | 656           | 252               | 815  | 575  | 1346 | 1056 | 838 | 355 |
| Median, SVD Grassmannian | 310   | 115     | 550           | 56                | 597  | 464  | 1145 | 817 | 631 | 122 |
| Median, NMF Grassmannian | 252   | 127     | 475           | 64                | 586  | 510  | 1172 | 659 | 614 | 119 |

structured and differ significantly in the magnetogram portion. This is consistent with the results in Table 10 where the ARs in Cluster 1 are similar to each other based on the Grassmannian scheme and the Cluster 2 ARs are not.

Comparing the corresponding magnetogram dictionary elements of some example projection matrices in Figure 9 shed further light on clustering interpretation. The Cluster 1 dictionary has elements 1 and 5 that are nearly neutral; element 2 has a negative polarity and is uniform; elements 3 and 4 show diagonal gradients. Any directional gradient can be formed from a linear combination of these two gradient elements. Thus this projection matrix and dictionary appears to correspond to an AR that is fairly regular in the magnetogram: most spatial magnetogram features can be represented by simple gradients, that is, most $3 \times 3$ neighborhoods exhibit a strong gradient in magnetic field in one given direction. In contrast, the dictionary from Cluster 2 has an irregular component in addition to the two weaker gradients in elements 3 and 4.

These results are consistent with Table 11 since smaller sunspots are generally less structured and have weaker magnetic field values than larger sunspots. Thus more of the smaller sunspots are assigned to Cluster 2 than Cluster 1.

The NMF Grassmannian results are similar. The ARs assigned to Cluster 1 all have projection matrices with the continuum and magnetogram components that are similar to each other, while the ARs in Cluster 2 generally have less structured projection matrices.
Figure 9. Magnetogram dictionary elements corresponding to example projection matrices from the two clusters using the SVD Grassmannian approach. The Cluster 1 dictionary contains simple gradients and uniform patches. It is estimated from the NOAA 9097 shown in Figure 3 (top) while Cluster 2 dictionary is estimated from the NOAA 10045 shown in Figure 3 (bottom).

Figure 10. Histograms of the Mount Wilson classes divided by clustering assignment using the Hellinger distance. As for the Grassmannian configurations, Cluster 1 contains more of the complex ARs while Cluster 2 contains more of the simple ARs.

All these observations indicate that although the correspondence with the Mt Wilson classification is somewhat weak, both clustering configurations do take into account some form of complexity of the ARs.

5.3. Clustering Within the Sunspots: Hellinger Distance

We now analyze the results from the Hellinger-based clusterings, for which the corresponding dictionary elements are represented in Figure 4 (for the SVD factorization) and in Figure 5 (for the NMF). Figure 10 shows clear patterns between the Hellinger-based clusterings and Mount Wilson type distribution, similarly to the Grassmannian results in Figure 8. Thus the Hellinger distance clustering approach appears to separate out the complex sunspots suggesting that these configurations are also clustering based on some measure of AR complexity.

The Hellinger-based clusterings appear to be correlated with sunspot size for some of the Mount Wilson classes, see Table 12. Based on the mean and median number of pixels, the Hellinger distance on the NMF coefficients tends to gather in Cluster 2 the smallest AR from classes $\alpha$, $\beta$, and $\beta\gamma$. Similarly, the Hellinger distance on the SVD coefficients separates the $\beta$ and $\beta\gamma$ AR by size with Cluster 1 containing the largest and Cluster 2 containing the smallest AR.
Table 12. Mean and median number of pixels of the ARs in each cluster under the Hellinger clustering schemes. Cluster 1 contains the larger sunspots for all groups when using NMF and for some of the groups when using SVD.

| Cluster                  | $\alpha$ 1 | $\alpha$ 2 | $\beta$ 1 | $\beta$ 2 | $\beta\gamma$ 1 | $\beta\gamma$ 2 | $\beta\gamma\delta$ 1 | $\beta\gamma\delta$ 2 | All 1 | All 2 |
|--------------------------|------------|------------|-----------|-----------|------------------|------------------|-----------------------|-----------------------|-------|-------|
| Mean, SVD Hellinger      | 278        | 260        | 582       | 270       | 823              | 577              | 1234                  | 1354                  | 756   | 422   |
| Mean, NMF Hellinger      | 384        | 183        | 788       | 156       | 847              | 515              | 1418                  | 880                   | 882   | 283   |
| Median, SVD Hellinger    | 148        | 128        | 477       | 70        | 612              | 472              | 1012                  | 1172                  | 588   | 174   |
| Median, NMF Hellinger    | 265        | 121        | 580       | 56        | 631              | 393              | 1157                  | 665                   | 677   | 105   |

Figure 11. Histograms of the marginal distributions of the coefficients corresponding to the mean magnetic field value (dictionary element 2 in Figure 4) for Cluster 1 (left) and Cluster 2 (right) using the SVD coefficients. Cluster 1 ARs contain more patches with near neutral magnetic field values.

When looking at the SVD coefficients, we see that their marginal distributions are similar across clusters, except for the coefficients that correspond to the second dictionary element of Figure 4. Recall that this second dictionary element is associated with the average magnetic field value of a patch. If the corresponding coefficient is close to zero, it means the average magnetic field in the patch is also close to zero. Figure 11 shows histograms of the coefficients of the second dictionary element. The histograms correspond to patches from all AR separated by cluster assignment. The histograms show that Cluster 1 has a high concentration of patches with near zero average magnetic field. In contrast, the larger peaks for Cluster 2 are centered around $+1$ and $-1$. This suggests that the clustering assignments are influenced somewhat by the amount of patches in an AR that have near zero average magnetic field values. As we are considering only the core (sunspot) part of the AR, having $3 \times 3$ patch with a near zero average magnetic field entails that the corresponding patch is likely to be located along the neutral line separating strong magnetic fields. Thus the local distribution of magnetic field values is related to cluster assignments when using the SVD coefficients.

Checking the individual ARs and their coefficient distributions in each cluster, we see indeed that Cluster 1 does contain more ARs with patches having near zero average magnetic field. This tends
to include more of the complex ARs in Cluster 1 since they are more likely to have a neutral line close to the regions of strong magnetic fields that will therefore be included in the STARA masks.

It should be noted however that the correspondence is not perfect. There are some ARs in Cluster 2 where the regions of opposing polarity are close to each other and some ARs in Cluster 1 where the regions of opposing polarity are far apart. Thus the distribution of average magnetic field values is only one factor in the clustering. As mentioned previously, the size of the AR is another factor, especially for $\beta$ groups.

Investigating the joint histograms of the NMF dictionary elements corresponding to positive and negative magnetic field values reveals that the NMF Hellinger clustering results are also influenced by the local magnetic field distribution.

5.4. Discussion of Sunspot Results

All four clustering configurations have several things in common. Cluster 1 generally contains larger sunspots while Cluster 2 contains smaller sunspots (Tables 11 and 12). The MDS plots are all very similar and are ordered by average similarity (Figs. 6 and 7). One cluster contains AR that are very similar to each other on average while the other cluster contains AR that are not very similar to each other (Table 10). The relative distributions of the Mount Wilson labels for each cluster is similar across configurations (Figs. 8 and 10). Despite these similarities, the correspondence between the results of the different configurations is fairly low as mentioned in Section 5.1 (Table 9). Additionally, there are some subtle differences that may provide insight into the situations for which each configuration is best suited.

Indeed, under the Grassmanian configuration, Cluster 1 contains mostly large or complex sunspots that are all quite similar to each other, while Cluster 2 gathers AR that are mostly dissimilar to each other. Hence this approach allows us to cluster together large or complex ARs that share some regularity pattern, and also to distinguish between different types of small sunspots. In contrast, for the Hellinger distance schemes, the Cluster 2 ARs containing the smallest sunspots are most similar to each other while the Cluster 1 ARs are more dissimilar. This indicates that the Hellinger distance approaches are best for distinguishing between different types of larger or complex ARs.

When NMF is applied on datasets where all values in the dictionary and coefficient matrices are constrained to be nonnegative, its results are generally more interpretable than SVD. In our application however, the magnetogram components can be negative. Hence the NMF results are not particularly sparse and lose some benefits of nonnegativity since the positive and negative magnetogram components can cancel each other. This results in some loss of interpretability. Additionally, the SVD results seem to be more interpretable due to separate treatment of the continuum and magnetogram components. However, there is still some value in the NMF approach as we see that the Hellinger distance clustering on NMF coefficients are better at separating the ARs by size than the SVD approach. Additionally, the NMF approaches tend to agree more strongly with the Mount Wilson labels than their SVD counterparts as is seen in Section 5.5 below. Future work could include using alternate forms of NMF such as in Ding et al. (2010) where sparsity and interpretability is preserved even when the dictionary is no longer constrained to be nonnegative.
5.5. Clustering with Neutral Line Data

As described in Section 4.3, we analyze the effects of including data from the neutral line in the clustering. We proceed by taking a weighted average of the dissimilarities calculated from the sunspots and from the neutral line data matrices. Using the ARI, we compare the results to labels based on the Mount Wilson classification scheme, see Table 8 for the label definition. We use a grid of weights, starting from a weight of 0 for a clustering using only patches within sunspots up to a weight of 1 for a clustering that takes into account only the neutral line data.

Figure 12 plots the ARI for the four different schemes as a function of the weight. In nearly all cases, the ARI is above zero which indicates that the clustering does slightly better than a random assignment. In general, the correspondence of the clustering results with the Mount Wilson based labels decreases as the weight approaches 1. This means that the natural clusterings associated with only the neutral line data do not correspond well with the Mount Wilson based labels. However in several cases, including some information from the neutral line at lower weights appears to increase the correspondence, e.g., the ARI increases for the 2 cluster case for the Grassmannian on SVD dictionaries with a weight around 0.1, and also for the 3 and 4 cluster cases for the Grassmannian on NMF dictionaries with a weight around 0.5.
Another pattern that is evident from Figure 12 is that when using the Grassmannian metric, the ARI is generally higher when using fewer clusters. In contrast, this is not generally the case when using the Hellinger distance.

When using the Grassmannian metric on the NMF dictionaries or the Hellinger distance on the NMF coefficients, the mean silhouette decreases as the number of clusters increases for all weights. So based on the mean silhouette heuristic, 2 is the most natural number of clusters for all weights when using NMF. This is not the case when using SVD (Fig. 13). While the 2 cluster scheme commonly has the highest mean silhouette width, there are certain weights where a larger number of clusters results in a larger mean silhouette width.

The gradients of magnetic field values across the neutral line is a key quantity used in several indicators of potential eruptive activity (Schrijver, 2007). We therefore repeated the neutral line experiment where we focused only on the gradients within the magnetogram as follows. When we applied SVD dictionary learning to the data matrix $Z$ extracted from the neutral line, the resulting dictionary matrix $A$ was very similar to that shown in Figure 4. Note that elements 3 and 4 correspond to the gradient patterns within the magnetogram data. Therefore, after learning $A$ and $H$ from $Z$, we kept only the coefficients corresponding to dictionary elements 3 and 4, i.e. the 3rd and 4th rows of $H$. We then estimated the Hellinger distance between the ARs’ underlying distributions of these two coefficients. In other words, we only included the neutral line coefficients corresponding to magnetogram gradients as input into the clustering algorithm. For the data within the sunspots, we included all coefficients as before.

Figure 14 shows the ARI as a function of the weight for this experiment. For the all cases, the ARI stays fairly constant with a few increases and decreases until the weight increases to 0.9, after which it drops dramatically. Under the original experiment where we used all of the SVD coefficients from the neutral line, we do get a higher ARI for the 3 and 4 cluster cases with a weight of 0.1 (Fig. 12). However for higher weights, the ARI for the case of all coefficients continually decreases. Thus the correspondence with the Mount Wilson labels and the clustering appears to be better preserved when we only include the magnetogram gradient coefficients.

While the correspondence of the clusterings with the Mount Wilson based labels are fairly low, the clear patterns in the ARI indicate that the relationship between the weight and the ARI is un-
Figure 14. Plot of the ARI using the Hellinger distance within the neutral line on only the coefficients corresponding to the SVD dictionary elements associated with magnetogram gradients. The corresponding dictionary elements are similar to the 3rd and 4th elements in Figure 4. Focusing on the gradients results in a higher ARI for higher weights than when all coefficients are used as seen in Figure 12.

Table 13. Summary of features distinguishing the clusters under the various classification schemes tested.

| Class. Scheme | Cluster 1                                                                 | Cluster 2                                                                 |
|---------------|---------------------------------------------------------------------------|---------------------------------------------------------------------------|
| SVD Grassmannian | large, complex sunspots with regular pattern in gradient; small value of Grassmannian metric between ARs | small, simple sunspots with irregular pattern in gradient; large value of Grassmannian metric |
| NMF Grassmannian | idem SVD Grassmannian                                                    | idem SVD Grassmannian                                                    |
| SVD Hellinger  | largest $\beta, \beta \gamma$ sunspots; high concentration of patches with average magnetic field value $\approx 0$; large Hellinger distance between ARs | smallest $\beta, \beta \gamma$ sunspots; high concentration of patches with average magnetic field value close to $+1$ or $-1$; small Hellinger distance between ARs |
| NMF Hellinger  | largest $\alpha, \beta, \beta \gamma$ sunspots; large Hellinger distance between ARs | smallest $\alpha, \beta, \beta \gamma$ sunspots; small Hellinger distance between ARs |

likely to be due entirely to noise. Thus there may be value in including the neutral line when clustering and analyzing the ARs.

6. Conclusion

Current AR classification systems such as the Mount Wilson scheme focus on the largest length scale when describing the geometrical arrangements of the magnetic field. This work has focused on classifying ARs at a fine length scale. We have shown that when we analyze and cluster the AR based on the global statistics of the local properties, there are similarities to the classification based on the large scale characteristics. For example, when clustering using the Hellinger distance, 1 cluster contained most of the complex AR. Other large scale properties such as the size of the AR also influenced the clustering results. Table 13 summarizes the properties that are found to influence the clustering under the four classification schemes.
From our results, we can provide several recommendations for which schemes to use based on the desired outcomes. If we wish to study differences between large, complex ARs (as determined by the Mount Wilson system), then the Hellinger distance-based methods are best. In contrast, if differences between smaller and simpler ARs are to be highlighted, then the Grassmannian based approaches work better. In both cases, the SVD approaches tend to lead to more interpretable results although the NMF approaches had several advantages such as greater correspondence with the Mount Wilson labels. Modified versions of NMF may also lead to improved interpretability.

While the correspondence of the clustering results with the Mount Wilson scheme was not particularly high, this does not prevent our method of using global statistics of local properties from being able to accurately representing the Mount Wilson classes. Unsupervised classification or clustering separates the AR naturally based on the geometry of the input feature space. In contrast, supervised classification uses training labels to build a classifier or predictor within the feature space. Thus supervised classification can always do at least as well as unsupervised learning in reproducing class labels.

A good way of assessing how well a given feature space can do in a supervised setting is to estimate the Bayes error. The Bayes error gives the minimum average probability of error that any classifier can achieve on the given data and can be estimated in the two class setting by estimating upper and lower bounds such as the Chernoff bound using a divergence-based estimator as in Moon and Hero III (2014b). These bounds can be estimated using various schemes and combinations of data (inside sunspots, along the neutral line, etc.) to determine which scheme is best at reproducing the desired labels (e.g., the Mount Wilson labels).

In our studies, we also found that the STARA masks are too restrictive in some cases which led to a mismatch between the Mount Wilson label and the extracted data. For example, there were several cases where an AR was labeled as a $\beta$ class but the STARA mask only extracted magnetic field values of one polarity. For this reason, we expect that including information beyond the STARA masks will lead to improved matching with the Mount Wilson labels. Additionally, going beyond the STARA masks will also be desirable for other tasks such as finding global statistics of local properties that relates to increased eruptive activity.

These methods of comparing AR images can also be adapted to a time series of image pairs. For example, image pairs from a given point in time may be compared to the image pairs from an earlier period to measure how much the ARs have changed. The evolution of an AR may also be studied by defining class labels based on the results from one of the clustering schemes in this paper. From the clustering results, a classifier may be trained that is then used to assign an AR to one of these clusters at each time step. The evolution of the AR’s cluster assignment can then be examined.

Acknowledgements. This work was partially supported by the US National Science Foundation (NSF) under grant CCF-1217880 and a NSF Graduate Research Fellowship to KM under Grant No. F031543. VD acknowledges support from the Belgian Federal Science Policy Office through the ESA-PRODEX program, grant No. 4000103240, while RDV acknowledges support from the BRAIN.be program of the Belgian Federal Science Policy Office, contract No. BR/121/P1/PREDISOL. We thank Dr. Laura Balzano at the University of Michigan, as well as Dr. Laure Lefèvre and Ms. Marie Dominique at the Royal Observatory of Belgium for their feedback on the manuscript. The authors gratefully acknowledge Dr. Paul Shearer’s help in implementing the modified NMF algorithm.
Moon et al: Clustering of active regions

References

Ahmed, O. W., R. Qahwaji, T. Colak, P. A. Higgins, P. T. Gallagher, and D. S. Bloomfield. Solar Flare Prediction Using Advanced Feature Extraction, Machine Learning, and Feature Selection. *Sol. Phys.*, 404, 2011. 10.1007/s11207-011-9896-1. 1

Barnes, G., K. D. Leka, E. A. Schumer, and D. J. Della-Rose. Probabilistic forecasting of solar flares from vector magnetogram data. *Space Weather*, 5, S09002, 2007. 10.1029/2007SW000317. 1

Bhattacharyya, A. On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics*, 401–406, 1946. 2.4

Bobra, M. G., and S. Couvidat. Solar Flare Prediction Using SDO/HMI Vector Magnetic Field Data with a Machine-learning Algorithm. *ApJ*, 798, 2015. 10.1088/0004-637X/798/2/135. 1411.1405. 1

Candès, E. J., and D. L. Donoho. Recovering edges in ill-posed inverse problems: optimality of curvelet frames. *Ann. Statist.*, 30(3), 784–842, 2002. 10.1214/1028674842, URL http://dx.doi.org/10.1214/aos/1028674842. 1

Csiszar, I. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2, 299–318, 1967. 2.4

Daubechies, I. Orthonormal bases of compactly supported wavelets. *Communications on pure and applied mathematics*, 41(7), 909–996, 1988. 1

Ding, C., T. Li, and M. I. Jordan. Convex and semi-nonnegative matrix factorizations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(1), 45–55, 2010. 5.4

Donoho, D. L. Wedgelets: nearly minimax estimation of edges. *Ann. Statist.*, 27(3), 859–897, 1999. 10.1214/aos/1018031261, URL http://dx.doi.org/10.1214/aos/1018031261. 1

Donoho, D. L., and J. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3), 425–455, 1994. 10.1093/biomet/81.3.425. http://biomet.oxfordjournals.org/content/81/3/425.full.pdf+html, URL http://biomet.oxfordjournals.org/content/81/3/425.abstract. 1

Edelman, A., T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2), 303–353, 1998. 2.4

Elad, M., and M. Aharon. Image Denoising Via Sparse and Redundant Representations Over Learned Dictionaries. *Image Processing, IEEE Transactions on*, 15(12), 3736–3745, 2006. 10.1109/TIP.2006.881969. 1

Falconer, D. A., R. L. Moore, and G. A. Gary. Magnetogram Measures of Total Nonpotentiality for Prediction of Solar Coronal Mass Ejections from Active Regions of Any Degree of Magnetic Complexity. *ApJ*, 689, 1433–1442, 2008. 10.1086/591045. 1

Freeman, W., and E. Adelson. The design and use of steerable filters. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 13(9), 891–906, 1991. 10.1109/34.93808. 1

Galluccio, L., O. Michel, P. Comon, M. Kliger, and A. O. Hero III. Clustering with a new distance measure based on a dual-rooted tree. *Information Sciences*, 251, 96–113, 2013. 1, 4.1, A.2, A.2
Moon et al: Clustering of active regions

Georgoulis, M. K., and D. M. Rust. Quantitative Forecasting of Major Solar Flares. *ApJ*, **661**, L109–L112, 2007. 10.1086/518718. 1

Guo, J., H. Zhang, O. V. Chumak, and Y. Liu. A Quantitative Study on Magnetic Configuration for Active Regions. *Sol. Phys.*, **237**, 25–43, 2006. 10.1007/s11207-006-2081-2. 1

Hale, G. E., F. Ellerman, S. B. Nicholson, and A. H. Joy. The Magnetic Polarity of Sun-Spots. *ApJ*, **49**, 153, 1919. 10.1086/142452. 1

Hellinger, E. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *Journal für die reine und angewandte Mathematik*, **136**, 210–271, 1909. 2.4

Higgins, P. A., P. T. Gallagher, R. McAteer, and D. S. Bloomfield. Solar magnetic feature detection and tracking for space weather monitoring. *Advances in Space Research*, **47**(12), 2105–2117, 2011. 1

Huang, X., D. Yu, Q. Hu, H. Wang, and Y. Cui. Short-Term Solar Flare Prediction Using Predictor Teams. *Sol. Phys.*, **263**, 175–184, 2010. 10.1007/s11207-010-9542-3. 1

Kruskal, J. B., and M. Wish. Multidimensional Scaling, vol. 11. Sage, 1978. 5.1

Kullback, S., and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 79–86, 1951. 2.4

Künstel, H. Die Flare-Häufigkeit in Fleckengruppen unterschiedlicher Klasse und magnetischer Struktur. *Astronomische Nachrichten*, **285**, 271, 1960. 1

Langville, A. N., C. D. Meyer, R. Albright, J. Cox, and D. Duling. Initializations for the nonnegative matrix factorization. In Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 23–26. Citeseer, 2006. 2.3, 2

Le Pennec, E., and S. Mallat. Sparse geometric image representations with bandelets. *Image Processing, IEEE Transactions on*, **14**(4), 423–438, 2005. 10.1109/TIP.2005.843753. 1

Lee, D. D., and H. S. Seung. Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing Systems (NIPS), 556–562, 2001. 2.2

Lee, K., Y.-J. Moon, J.-Y. Lee, K.-S. Lee, and H. Na. Solar Flare Occurrence Rate and Probability in Terms of the Sunspot Classification Supplemented with Sunspot Area and Its Changes. *Sol. Phys.*, **281**, 639–650, 2012. 10.1007/s11207-012-0091-9. 1

Leka, K. D., and G. Barnes. Photospheric Magnetic Field Properties of Flaring vs. Flare-Quiet Active Regions III: Discriminant Analysis of a Statistically Significant Database. In American Astronomical Society Meeting Abstracts #204, vol. 36 of *Bulletin of the American Astronomical Society*, 715, 2004. 1

Lin, C.-J. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, **19**(10), 2756–2779, 2007. 2, A.1

Mallat, S., and Z. Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, **41**(12), 3397–3415, 1993. 10.1109/78.258082. 1

Mayfield, E. B., and J. K. Lawrence. The correlation of solar flare production with magnetic energy in active regions. *Sol. Phys.*, **96**, 293–305, 1985. 10.1007/BF00149685. 1
Moon et al: Clustering of active regions

Moon, K. R., and A. O. Hero III. Ensemble estimation of multivariate f-divergence. In Information Theory (ISIT), 2014 IEEE International Symposium on, 356–360. IEEE, 2014a. 2.4, 3.2, 4.2

Moon, K. R., and A. O. Hero III. Multivariate f-Divergence Estimation With Confidence. In Advances in Neural Information Processing Systems, 2420–2428, 2014b. 2.4, 6

Moon, K. R., J. J. Li, V. Delouille, R. De Visscher, F. Watson, and A. O. Hero, III. Image patch analysis of sunspots and active regions. I. Intrinsic dimension and correlation analysis. ArXiv e-prints, 2015. 1503. 04127. 1, 1, 2, 2.1, 4, 5, 3.1, 4.2

Moon, K. R., J. J. Li, V. Delouille, F. Watson, and A. O. Hero III. Image patch analysis and clustering of sunspots: A dimensionality reduction approach. In IEEE International Conference on Image Processing. IEEE, 2014. 1, 2.4, 4.2, 5.1

Moon, T. K., and W. C. Stirling. Mathematical Methods and Algorithms for Signal Processing. Prentice hall New York, 2000. 2.1

Prim, R. C. Shortest connection networks and some generalizations. Bell system technical journal, 36(6), 1389–1401, 1957. A.2

Ramírez, I., and G. Sapiro. An MDL framework for sparse coding and dictionary learning. Signal Processing, IEEE Transactions on, 60(6), 2913–2927, 2012. 2

Rand, W. M. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66(336), 846–850, 1971. 5.1

Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65, 1987. 5.1

Sammis, I., F. Tang, and H. Zirin. The Dependence of Large Flare Occurrence on the Magnetic Structure of Sunspots. ApJ, 540, 583–587, 2000. 10.1086/309303. 1

Scherrer, P. H., R. S. Bogart, R. I. Bush, J. T. Hoeksema, A. G. Kosovichev, et al. The Solar Oscillations Investigation - Michelson Doppler Imager. Solar Physics, 162, 129–188, 1995. 10.1007/BF00733429. 1

Schrijver, C. J. A Characteristic Magnetic Field Pattern Associated with All Major Solar Flares and Its Use in Flare Forecasting. ApJ, 655, L117–L120, 2007. 10.1086/511857. 1, 1, 5.5

Sethian, J. A. A Fast Marching Level Set Method for Monotonically Advancing Fronts. In Proc. Nat. Acad. Sci, 1591–1595, 1995. 4.3

Song, H., C. Tan, J. Jing, H. Wang, V. Yurchyshyn, and V.Abramenko. Statistical Assessment of Photospheric Magnetic Features in Imminent Solar Flare Predictions. Sol. Phys., 254, 101–125, 2009. 10.1007/s11207-008-9288-3. 1

Stewart, G. W. Error and perturbation bounds for subspaces associated with certain eigenvalue problems. SIAM Review, 15(4), 727–764, 1973. 2.4

Warwick, C. S. Sunspot Configurations and Proton Flares. ApJ, 145, 215, 1966. 10.1086/148755. 1

Watson, F. T., L. Fletcher, and S. Marshall. Evolution of sunspot properties during solar cycle 23. Astronomy & Astrophysics, 533, A14, 2011. 10.1051/0004-6361/201116655, 1108. 4285. 1
Appendix A: Method Details

A.1. Dictionary Learning and Matrix Factorization

As mentioned in Section 2, the goal of linear dictionary learning is to accurately decompose the $2m^2 \times n$ data matrix $Z$ into the product of two matrices $A$ (with size $2m^2 \times r$) and $H$ (with size $r \times n$), where $A$ has fewer columns than rows ($r < 2m^2$). The matrix $A$ is the dictionary and the matrix $H$ is the coefficient matrix. The columns of $A$ form a basis for the data in $Z$.

The two dictionary learning methods we use are principal component analysis (SVD) and non-negative matrix factorization (NMF). These two methods can be viewed as solving two different optimization problems where the objective function is the same but the constraints differ. Let $A = [a_1, a_2, \ldots, a_r]$ and $H = [h_1, h_2, \ldots, h_n]$. For SVD, the optimization problem is

$$\min_{A,H} \|Z - AH\|_F^2$$

subject to $a_i^T a_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$

In words, SVD requires the columns of $A$ to be orthonormal.

For standard NMF, the optimization problem is

$$\min_{A,H} \|Z - AH\|_F^2$$

subject to $a_i \geq 0$, $\forall i = 1, \ldots, r$

$h_i \geq 0$, $\forall i = 1, \ldots, n$

where $a \geq 0$ applied to a vector $a$ implies that all of $a$’s entries are greater than or equal to 0. In our problem, only the continuum is nonnegative so we only apply the constraint to the continuum part of the matrix $A$. So if $a_i$ and $b_i$ are both vectors with length $m^2$ corresponding to the continuum and magnetogram parts, respectively, then we have $A = \begin{bmatrix} a_1 & a_2 & \cdots & a_r \\ b_1 & b_2 & \cdots & b_r \end{bmatrix}$. The NMF method we use also constrains the columns of $H$ to lie on a simplex, i.e. $\sum_{j=1}^r h_i(j) = 1$. Thus the optimization problem for our approach to NMF is

$$\min_{A,H} \|Z - AH\|_F^2$$

subject to $a_i \geq 0$, $\forall i = 1, \ldots, r$

$h_i \geq 0$, $\forall i = 1, \ldots, n$

$\sum_{j=1}^r h_i(j) = 1$, $\forall i = 1, \ldots, n$

This problem is not convex and is solved in an alternating manner by fixing $H$, finding the matrix $A$ that solves the problem assuming $H$ is fixed, and then solving for $H$ while $A$ is fixed. This process is repeated until the algorithm converges to a local minimum. See Lin (2007) for more details on the convergence analysis.
A.2. The EAC-DC Clustering Method

Let $V = \{v_1, v_2, \ldots, v_N\}$ be a set of vertices and let $E = \{e_{ij}\}$, where $e_{ij}$ denotes an edge between vertices $v_i$, $v_j$, $i, j \in \{1, \ldots, N\}$, be a set of undirected edges between them. The pair $(V, E) = G$ is the corresponding undirected graph. In our application, $V$ corresponds to the set of AR image pairs being clustered and $E$ contains all possible edges between the vertices. The weight of an edge $e_{ij}$ is defined as $w_{ij}$ and measures the base dissimilarity between two vertices $v_i$ and $v_j$. In many applications, the base dissimilarity is the Euclidean distance. In our case, we use either the Grassmannian metric or the Hellinger distance as the base dissimilarity measure.

A spanning tree $T$ of the graph $G$ is a connected acyclic subgraph that passes through all $N$ vertices of the graph and the weight of $T$ is the sum of all the edge weights used to construct the tree, $\sum_{e_{ij} \in T} w_{ij}$. A minimal spanning tree of $G$ is a spanning tree which has the minimal weight $\min_T \sum_{e_{ij} \in T} w_{ij}$.

Prim’s algorithm (Prim, 1957) is used by Galluccio et al. (2013) to construct the dual rooted MST. In Prim’s algorithm, the MST is grown sequentially where at each step, a single edge is added. This edge corresponds to the edge with minimal weight that connects a previously unconnected vertex to the existing tree. The root of the MST corresponds to the beginning vertex. For the dual rooted MST, we begin with two vertices $v_i$ and $v_j$ and construct the minimal spanning trees $T_i$ and $T_j$. At each step, the two edges that would grow both trees $T_i$ and $T_j$ using Prim’s algorithm are proposed and the edge with minimal weight is added. This continues until $T_i$ and $T_j$ connect. The weight of the final edge added in this algorithm defines a new metric between the vertices $v_i$ and $v_j$. This process is repeated for all pairs of vertices and this new metric is used as input to spectral clustering (Galluccio et al., 2013).

A primary advantage of this metric based on the hitting time of the two MSTs is that it depends on the MST topology of the data. Thus if two vertices belong to the same cluster, then the MST distance between them will be small since cluster points will be close together. This is the case even if the vertices are far away from each other (e.g. on opposite ends of the cluster). However, if the two vertices are in different clusters that are well separated, then the MST distance between them will be large. See Figure A.1 for an example. Thus this method of clustering is very robust to the shape of the clusters. Galluccio et al. (2013) contains many more examples.

The MST based metric can be computationally intensive to compute as Prim’s algorithm must be run as many times as there are pairs of vertices. To counter this, Galluccio et al. (2013) proposed the EAC-DC algorithm which uses the information from only a subset of the dual rooted MSTs. This is done by calculating the dual rooted MSTs for a random pair of vertices. Three clusters are defined for each run: all vertices that are connected to one of the roots in the MSTs form two of the clusters (one for each root) while all points that are not connected to either of the MSTs are assigned to a third “rejection” cluster. A co-association measure for two vertices is then defined as the number of times those vertices are contained in the same non-rejection cluster divided by the total number of runs (dual rooted MSTs). This co-association measure forms a similarity measure to which spectral clustering is applied.
Figure A.1. Dual rooted Prim tree built on a 2-dimensional data set when the roots are chosen from the same class (left) and different classes (right). The X’s mark the roots of the trees and the dashed line is the last connected edge. The length of the last connected edge is greater when the roots belong to clusters that are more separated.