SPECTRAL CLUSTERING OF POLARIMETRIC SAR DATA WITH WISHART-DERIVED DISTANCE MEASURES

Stian Normann Anfinsen (1), **Robert Jenssen** (1), **Torbjørn Eltoft** (1)

(1) University of Tromsø, Department of Physics and Technology, N-9037 Tromsø, Norway

Email: {stiann,robertj,pcte}@phys.uit.no

ABSTRACT

This paper presents a new spectral clustering algorithm, which is specially tailored for segmentation of polarimetric SAR images. This is accomplished by use of certain pairwise distance measures between pixels. The measures are derived from the complex Wishart distribution, and capture the statistical information contained in the coherency matrix. We demonstrate how the pairwise distances are transformed into an affinity matrix, whose eigendecomposition determines the optimal partitioning of pixels. We further show that the obtained clustering provides an improved initialization of the classical unsupervised Wishart classifier, and that the entire classification can also be performed in a kernel induced feature space. The algorithms are tested on crop classification with promising results.

1 INTRODUCTION

Synthetic aperture radar (SAR) is an important instrument in remote sensing of the earth, providing all-weather measurements that can be used to extract thematic information and geophysical parameters of the imaged surface scene. Image segmentation is one of the application areas, with major importance in mapping of vegetation, urban areas, and sea ice, among others. The capabilities of polarimetric SAR (PolSAR) have long been demonstrated from airborne platforms. With the advent of a number of spaceborne instruments, such as PALSAR and Radarsat-2, it will now be possible to utilise PolSAR data with increased information content and higher detail level for operational purposes.

The literature on PolSAR data classification is extensive, but the Wishart classifier stands out as the preferred general purpose algorithm, both for the supervised [1] and the unsupervised [2] case, of which we focus on the latter. The unsupervised Wishart classifier is based upon a procedure similar to the well established k-means algorithm [3], and inherits attractive properties such as computational efficiency and simple parametrization. Moreover, it produces consistenly good results for various application, and is thus a natural benchmark algorithm. However, despite its simplicity and generally good performance, this classifier suffers from disadvantages that it shares with k-means: The number of classes is fixed by the choice of the user; Convergence is not guaranteed; The clustering result is sensitive to the initialization; Finally, the rate of pixels changing class between iterations may be relatively high, even after several iterations.

Central grouping techniques, such as the k-means algorithm, are characterised by the comparison of individual pixels with class prototypes. They tend to be computationally efficient, but have drawbacks such as restricted flexibility of the resulting discriminant surface, and the implicit requirement that all class members must be similar to a single class prototype. Grouping by pairwise affinities is a different idea, which allows propagation of similarity from pixel to pixel. This allows recovery of clusters that take on more complicated manifold structures in feature space. Spectral clustering is one such technique utilising pairwise affinities between pixels, which is shown to be very promising [4], [5]. It has also been successfully applied to PolSAR data [6], with input features defined from a combination of polarimetric and spatial information.

Our approach is to use the coherency matrix, containing second-order statistics of the scattering coefficients, as input feature. Under the common assumption that the scattering coefficients are jointly complex Gaussian, zero mean and circular [1], the coherency matrix captures all available statistical information of a single pixel. Our motivation is to device a spectral clustering algorithm which can process coherency matrices as input features, while maintaining moderate computational cost. At the same time, we would like to develop an unsupervised segmentation algorithm that limits the number of user specified parameters to a minimal and robust set, and make progress in this direction. The proposed algorithm is tested on a data set from the airborne NASA/JPL AIRSAR instrument, which covers an agricultural area in Flevoland, The Netherlands. Performance measures are defined, and the algorithm is compared with the classical unsupervised Wishart classifier.

The remainder of the paper is organised as follows. Section 2 describes PolSAR data in terms of the feature set used for classification, and provides background theory on the Wishart classifier. The new algorithm is presented in section 3, by reviewing spectral clustering and coherency matrix distance measures, and explaining the detailed implementation. Results are presented in section 4, while conclusions and given in section 5, together with suggestions of further work.

2 THEORY

2.1 Polarimetric SAR Data

The full-polarimetric SAR measures the amplitude and phase of backscattered signals in four combinations of the linear receive and transmit polarizations: horizontal-horizontal (HH), horizontal-vertical (HV), vertical-horizontal (VH), and vertical-vertical (VV). These signals form the complex scattering matrix $\mathbf{S}$, that relates the incident and the scattered electric fields [7]:

$$\mathbf{S} = \begin{bmatrix} S_{HH} & S_{HV} \\ S_{VH} & S_{VV} \end{bmatrix} ,$$

where S_{RT} denotes the complex scattering amplitude for receive polarization $R \in \{H, V\}$ and transmit polarization $T \in \{H, V\}$. It is commonly assumed that natural targets exhibit reciprocity, i.e. S_{HV}=S_{VH}. The polarimetric measurements can then be presented in condensed form as the scattering vector

$$\mathbf{s} = \begin{bmatrix} S_{HH}, & \sqrt{2}S_{HV}, & S_{VV} \end{bmatrix}^T ,$$

where ^{T} denotes matrix transposition. The factor $\sqrt{2}$ in the middle term is defined to preserve the total power of the backscattered signal.

To reduce the effect of inherent speckle noise, PolSAR images are often spatially averaged, and data is represented by the sample covariance matrix

$$\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{s}_i \mathbf{s}_i^H , \tag{1}$$

where ^{H} denotes the Hermitian transpose, and n is the sample size used in the average. Alternatively, the scattering vector is replaced by the linear transformation

$$\mathbf{k} = \frac{1}{\sqrt{2}} \begin{bmatrix} S_{HH} + S_{VV} \\ S_{HH} - S_{VV} \\ 2S_{HV} \end{bmatrix} ,$$

known as the Pauli representation of the scattering vector. The resulting sample covariance matrix

$$\mathbf{T} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{k}_i \mathbf{k}_i^H \tag{2}$$

has been termed the sample coherency matrix. This form is often preferred to the socalled lexicographic version in Eq. (1), because the elements of $\mathbf{T}$ have an intuitive physical interpretation, which is easily related to the underlying scattering mechanisms of the resultion cell.

2.2 The Wishart Classifier

The Wishart classifier [1], [2] is based on the assumption that the elements of the scattering matrix are complex multivariate Gaussian, circular and zero mean. Hence, the density of the scattering vector in Pauli basis is

$$p_{\mathbf{k}}(\mathbf{k}) = \frac{1}{\pi^q |\mathbf{\Sigma}|} \exp\{-\mathbf{k}^H \mathbf{\Sigma}^{-1} \mathbf{k}\} , \tag{3}$$

where $\mathbf{\Sigma} = E\{\mathbf{T}\}$ is the true coherency matrix, $|\cdot|$ denotes the determinant, and q is the dimension of $\mathbf{k}$ (with $q = 3$). It follows that the coherency matrix is complex Wishart distributed with n degrees of freedom around the expectation value $\mathbf{\Sigma}$. We write this as $\mathbf{T} \sim \mathcal{W}(n, \mathbf{\Sigma})$. The density of $\mathbf{T}$ is given by [1]

$$p_{n,\mathbf{\Sigma}}(\mathbf{T}) = \frac{n^{nq} |\mathbf{T}|^{n-q} \exp\{-n \operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{T})\}}{K(n,q)|\mathbf{\Sigma}|^n} , \tag{4}$$

where $\operatorname{tr}(\cdot)$ denotes the trace operation and

$$K(n,q) = \pi^{q(q-1)/2} \prod_{k=1}^{q} \Gamma(n-k+1) \tag{5}$$

is a constant with $\Gamma(\cdot)$ defined as the standard gamma function.

The unsupervised Wishart classifier assigns pixels to classes in much the same way as the well known k-means algorithm [3]. It requires an initial data partitioning, which determines the number of classes, denoted k. The initialization is used to estimate prototype coherency matrices for each class, denoted $\hat{\mathbf{\Sigma}}_i$, i=$1, \ldots, k$. We next measure the distance between the coherency matrix representing each pixel and the class prototypes, and assign the pixel to the class which is closest with respect to the chosen distance measure. After all pixels have been classified, the class prototypes are recalculated, and the procedure repeated until some stop criterion is met.

Whereas the original k-means uses a vector distance to cluster multivariate data, the Wishart classifier requires a matrix distance measure, which is derived from the complex Wishart density [1]. The relation is

$$\begin{aligned} d_W(\mathbf{T}, \mathbf{\Sigma}) &= -1/n \ln p_{n,\mathbf{\Sigma}}(\mathbf{T}) - c \\ &= \ln |\mathbf{\Sigma}| + \operatorname{tr}(\mathbf{\Sigma}^{-1}\mathbf{T}) , \end{aligned} \tag{6}$$

where c is independent of $\mathbf{\Sigma}$. The unsupervised classification rule becomes as follows: Assign the pixel to class $\omega_i, i \in \{1, \ldots, k\}$, if

$$d_W(\mathbf{T}, \hat{\mathbf{\Sigma}}_i) \leq d_W(\mathbf{T}, \hat{\mathbf{\Sigma}}_j), \ \forall \omega_j \neq \omega_i . \tag{7}$$

This is easily seen to be a maximum likelihood (ML) approach, as minimisation of $d_W(\mathbf{T}, \mathbf{\Sigma})$ is equivalent to maximisation of the posterior likelihood

$$\begin{aligned} L(\omega_i|\mathbf{T}) &= \ln p(\omega_i|\mathbf{T}) \\ &\propto \ln p_{n,\mathbf{\Sigma}}(\mathbf{T}|\omega_i) P(\omega_i) \\ &\propto \ln p_{n,\mathbf{\Sigma}_i}(\mathbf{T}) , \end{aligned}$$

under the assumption that $P(\omega_i) = 1/k$, $\forall i$.

The unsupervised Wishart classifier is commonly initialized by a partitioning obtained in H/A/α space, based on the eigendecomposition theory of Cloude and Pottier [2]. It is then referred to as the Cloude-Pottier-Wishart (CPW) classifier. While this is a practical solution when no prior information exists about the natural classes, we argue that this initialization is a detour that causes slow convergence towards the ML solution, and that an alternative exists.

3 The New Algorithm

3.1 Spectral Clustering of PolSAR data

Spectral clustering algorithms are based on eigendecomposition of a matrix storing the pairwise affinities (i.e. similarities) between all data points. We start by calculating pairwise distances between all pixels using a distance measure of choice. We define $d(\mathbf{T}_i, \mathbf{T}_j)$ as a general distance measure between two coherency matrices. Distances are transformed into affinities, e.g. by using the Gaussian kernel function

$$g(\mathbf{T}_i, \mathbf{T}_j) = \exp\left\{-\frac{d^2(\mathbf{T}_i, \mathbf{T}_j)}{2\sigma^2}\right\} \tag{8}$$

or the Laplacian kernel function

$$g(\mathbf{T}_i, \mathbf{T}_j) = \exp\left\{-\frac{d(\mathbf{T}_i, \mathbf{T}_j)}{b}\right\} , \tag{9}$$

where the kernel bandwidths σ and b are smoothing parameters, which determines the size of the neighbourhood of affinity. All affinities are stored in an affinity matrix $\mathbf{G}$ with entries $\mathbf{G}_{ij} = g(\mathbf{T}_i, \mathbf{T}_j)$. We also define $\mathbf{D}$ as a diagonal matrix with entries $\mathbf{D}_{ii} = \sum_j \mathbf{G}_{ij}$, and the Laplacian matrix as

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{G}\mathbf{D}^{-1/2} ,$$

where $\mathbf{I}$ is the identity matrix. A variety of algorithms exists (See pp. 63–64 of [8] for references), that differ by the way they use one or more eigenvectors of $\mathbf{G}$ or $\mathbf{L}$ to project data points into a new feature space induced by the eigendecomposition, and the objective function used to cluster data in this feature space.

In our approach, we compute the eigenvectors of $\mathbf{G}$ corresponding to the U largest eigenvalues: $\{\mathbf{v}_1, \ldots, \mathbf{v}_U\}$. These are stacked as rows in the $U \times N$ matrix

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_U^T \end{bmatrix} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N] , \tag{10}$$

where the columns $\{\boldsymbol{\phi}_i\}_{i=1}^N$, become new feature vectors is an eigenspace with dimension U. The choice of U is not obvious. To simplify the parametrization, we have chosen to fix U to the number of classes, k, which must be selected by the user.

We next perform clustering in eigenspace with an objective function proposed in [9], which leads to a minimum angular distance classifier. We assign a data point $\mathbf{T}_i$ to the class ω_j for which the angle is smallest:

$$\mathbf{T}_i \to \omega_j \ : \quad \max_j \cos\angle(\boldsymbol{\phi}_i, \mathbf{m}_j) ,$$

where the $\{\mathbf{m}_j\}_{j=1}^k$ are class mean vectors in eigenspace. The classification rule reduces to

$$\mathbf{T}_i \to \omega_j \ : \quad \max_j \frac{\boldsymbol{\phi}_i^T \mathbf{m}_j}{||\mathbf{m}_j||} . \tag{11}$$

Classification is done iteratively, with the class mean vectors initialized to unit vectors in eigenspace. The motivation is that, assuming separability, clusters should theoretically be localised along orthogonal axes in eigenspace. When data does not contain k separable clusters, the chosen initialization may lead to empty classes, as certain regions of eigenspace are not populated. Hence, the effective number of clusters is reduced from k to k_E during iterations. We refer to this as a data-adaptive property of the classifier with respect to selection of k.

Another key parameter that must be determined is the kernel bandwidth σ or b from Eqs. (8) and (9). Ersahin et al. [6] report good experience with use of a locally adaptive bandwidth, which is an attractive solution when optimal values based on the data distribution and affinity function cannot be derived. Robust methods for bandwidth selection are under investigation, but in the experiments we resort to a manually chosen value.

Due to computational cost, we cannot spectrally cluster all pixels in a PolSAR image. We therefore use random sampling to extract a sample $\{\mathbf{T}_1, \ldots, \mathbf{T}_N\}$ of size N. Following spectral clustering of the sample, the complete data set can be classified in eigenspace using Eq. (11), provided that the remaining out-of-sample data points can be mapped into eigenspace. This is achived by an approximation known as the Nyström method [10], which defines the jth element of the eigenspace representation for an arbitrary data point $\mathbf{T}_t$ as

$$\boldsymbol{\phi}_t(j) \approx \frac{\sqrt{N}}{\lambda_j} \sum_{i=1}^{N} \mathbf{v}_{ji} d(\mathbf{T}_t, \mathbf{T}_i), \ j = 1, \ldots, U , \tag{12}$$

where $\mathbf{v}_{ji}$ is the ith element of the eigenvector corresponding to the jth largest eigenvalue of $\mathbf{G}$, denoted λ_j. Alternatively, we may use the partitioning obtained by spectral clustering to calculate mean class coherency matrices, and use this to initialize the Wishart classifier. This reduces the computational cost significantly, and is shown in practice to produce very similar classification results. It is therefore our preferred solution.

3.2 Coherency Matrix Distance Measures

In order to apply spectral clustering to PolSAR data, we need to define an appropriate matrix distance measure. We want to measure the pairwise distance between a set of matrices, and it is desirable that all conditions of a general metric apply, so that

1) $d(\mathbf{A},\mathbf{B}) \geq d_0$ [generalised non-negativity]
2) $d(\mathbf{A},\mathbf{B}) = d_0 \Leftrightarrow \mathbf{A}=\mathbf{B}$ [identity of discernibles]
3) $d(\mathbf{A},\mathbf{B}) = d(\mathbf{B},\mathbf{A})$ [symmetry]
4) $d(\mathbf{A},\mathbf{C}) \leq d(\mathbf{A},\mathbf{B}) + d(\mathbf{B},\mathbf{C})$ [subadditivity]

for any arbitrary Hermitian, positive definite, random matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ with dimension $p \times p$.

Two common matrix distance measures that meet these requirements are the Frobenius distance and the Euclidean distance (or ℓ_2 metric). However, these difference measures cannot be related to the Wishart distribution, and the clustering wil not be explicitly related to the statistical information of the data.

It is not obvious how we can derive a metric from the Wishart distribution, so some of the conditions must be relaxed. The distance measure $d_W(\mathbf{T},\mathbf{\Sigma})$ in Eq. (6) is neither symmetric nor subadditive, and the minimum value is not constant. A possible modification is

$$\begin{aligned} d'_W(\mathbf{A},\mathbf{B}) &= \frac{1}{2}\big(d_W(\mathbf{A},\mathbf{B}) + d_W(\mathbf{B},\mathbf{A})\big) \\ &= \frac{1}{2}\big(\ln|\mathbf{AB}| + \mathrm{tr}(\mathbf{AB}^{-1}+\mathbf{BA}^{-1})\big)\ . \end{aligned} \tag{13}$$

This distance measure is symmetric, but $d'_W(\mathbf{A},\mathbf{A}) = \ln|\mathbf{A}| + p$ still depends on the input argument through $\mathbf{A}$.

Another approach taken in [11] is to consider two sample coherency matrices $\mathbf{A} \sim \mathcal{W}(n,\mathbf{\Sigma}_A)$ and $\mathbf{B} \sim \mathcal{W}(n,\mathbf{\Sigma}_B)$, generated from unknown and potentially different Wishart densities, and the likelihood ratio hypothesis test of $H_0 : \mathbf{\Sigma}_A = \mathbf{\Sigma}_B$ versus $H_1 : \mathbf{\Sigma}_A \neq \mathbf{\Sigma}_B$. Under the null hypothesis, the distributions of $\mathbf{A}$ and $\mathbf{B}$ are equal, and it follows from the properties of the Wishart density that $\mathbf{A}+\mathbf{B} \sim \mathcal{W}(2n,\mathbf{\Sigma})$, where $\mathbf{\Sigma} = \mathbf{\Sigma}_A = \mathbf{\Sigma}_B$. The likelihood function under H_0 is therefore

$$L_{H_0}(\mathbf{\Sigma}|\mathbf{A},\mathbf{B}) = p_{n,\mathbf{\Sigma}}(\mathbf{A})\ p_{n,\mathbf{\Sigma}}(\mathbf{B})\ , \tag{14}$$

which is maximised by the maximum likelihood estimator (MLE) $\hat{\mathbf{\Sigma}} = (\mathbf{A}+\mathbf{B})/2$. The likelihood function under H_1 is

$$L_{H_1}(\mathbf{\Sigma}_A,\mathbf{\Sigma}_B|\mathbf{A},\mathbf{B}) = p_{n,\mathbf{\Sigma}_A}(\mathbf{A})\ p_{n,\mathbf{\Sigma}_B}(\mathbf{B})\ , \tag{15}$$

which attains its maximum value for the MLEs $\hat{\mathbf{\Sigma}}_A = \mathbf{A}$ and $\hat{\mathbf{\Sigma}}_B = \mathbf{B}$. The likelihood ratio test statistic thus becomes (See the appendix of [11] for details):

$$\begin{aligned} Q_1 &= \frac{\sup_{\mathbf{\Sigma}}\{L_{H_0}(\mathbf{\Sigma}|\mathbf{A},\mathbf{B})\}}{\sup_{\mathbf{\Sigma}_A,\mathbf{\Sigma}_B}\{L_{H_1}(\mathbf{\Sigma}_A,\mathbf{\Sigma}_B|\mathbf{A},\mathbf{B})\}} \\ &= \frac{p_{n,\hat{\mathbf{\Sigma}}}(\mathbf{A})\ p_{n,\hat{\mathbf{\Sigma}}}(\mathbf{B})}{p_{n,\hat{\mathbf{\Sigma}}_A}(\mathbf{A})\ p_{n,\hat{\mathbf{\Sigma}}_B}(\mathbf{B})} \\ &= 2^{2nq}\frac{|\mathbf{A}|^n|\mathbf{B}|^n}{|\mathbf{A}+\mathbf{B}|^{2n}}\ . \end{aligned} \tag{16}$$

This expression is turned into a distance measure by the transformation

$$\begin{aligned} d_B(\mathbf{A},\mathbf{B}) &= -\frac{\ln Q_1}{n} \\ &= \ln\left(\frac{|\mathbf{A}+\mathbf{B}|^2}{|\mathbf{A}||\mathbf{B}|}\right) - 2q\ln 2\ . \end{aligned} \tag{17}$$

This distance measure (with the constant term removed) is called the Bartlett distance [12]. We see that $d_B(\mathbf{A},\mathbf{A}) = 0$, and can easily prove identity of discernibles and symmetry. However, subadditivity does not hold, so the Bartless distance is a semimetric only.

Yet another distance measure can be derived from the likelihood ratio test, if we assume the same hypotheses H_0 and H_1 given $\mathbf{\Sigma}_B$ is known. The test statistic then becomes

$$\begin{aligned} Q_2 &= \frac{p_{n,\hat{\mathbf{\Sigma}}_B}(\mathbf{A})\ p_{n,\hat{\mathbf{\Sigma}}_B}(\mathbf{B})}{p_{n,\hat{\mathbf{\Sigma}}_A}(\mathbf{A})\ p_{n,\hat{\mathbf{\Sigma}}_B}(\mathbf{B})} \\ &= \frac{|\mathbf{A}|^n}{|\mathbf{B}|^n}\exp\{-n(\mathrm{tr}(\mathbf{B}^{-1}\mathbf{A})-q)\}\ . \end{aligned} \tag{18}$$

The resulting distance measure has been named the revised Wishart distance [12]:

$$\begin{aligned} d_{RW}(\mathbf{A},\mathbf{B}) &= -\frac{\ln Q_2}{n} \\ &= \ln\frac{|\mathbf{B}|}{|\mathbf{A}|} + \mathrm{tr}(\mathbf{B}^{-1}\mathbf{A}) - q\ , \end{aligned} \tag{19}$$

It satisfies $d_{RW}(\mathbf{A},\mathbf{B}) = 0$, but is not symmetric. A symmetric measure is obtained as

$$\begin{aligned} d'_{RW}(\mathbf{A},\mathbf{B}) &= \frac{1}{2}\big(d_{RW}(\mathbf{A},\mathbf{B}) + d_{RW}(\mathbf{B},\mathbf{A})\big) \\ &= \frac{1}{2}\mathrm{tr}(\mathbf{AB}^{-1}+\mathbf{BA}^{-1}) - p\ . \end{aligned} \tag{20}$$

It satisfies all conditions, except the triangle inequality, an is thus a semimetric. We note that this distance can be written as

$$d_{RW}(\mathbf{A},\mathbf{B}) = -\frac{1}{2n}\ln\left(\frac{p_{n,\mathbf{A}}(\mathbf{B})}{p_{n,\mathbf{A}}(\mathbf{A})}\frac{p_{n,\mathbf{B}}(\mathbf{A})}{p_{n,\mathbf{B}}(\mathbf{B})}\right)\ . \tag{21}$$

Thus, this distance measure can also be seen as the symmetric version of

$$\begin{aligned} d_{NLL}(\mathbf{A},\mathbf{B}) &= -\frac{1}{n}\ln\frac{p_{n,\mathbf{B}}(\mathbf{A})}{p_{n,\mathbf{B}}(\mathbf{B})} \\ &= \frac{n-p}{n}\ln\frac{|\mathbf{B}|}{|\mathbf{A}|} + \mathrm{tr}(\mathbf{B}^{-1}\mathbf{A}) - q\ . \end{aligned} \tag{22}$$

The motivation of $d_{NLL}(\mathbf{A},\mathbf{B})$ is different from that of $d_{RW}(\mathbf{A},\mathbf{B})$, as $d_{NLL}(\mathbf{A},\mathbf{B})$ can be interpreted as the normalised log-likelihood of $\mathbf{A}$ given $\mathbf{B}$ is true. The likelihood function is normalised to $0 \leq p_{n,\mathbf{B}}(\mathbf{A})/p_{n,\mathbf{B}}(\mathbf{B}) \leq 1$ by the maximum likelihood value. This yields the property: $0 \leq d_{NLL}(\mathbf{A},\mathbf{B}) \leq \infty$

The distance measure is an essential part of a spectral clustering algorithm, as it incorporates all prior knowledge about the data. In the result section, we will use the Bartlett distance, $d_B(\mathbf{A},\mathbf{B})$, and also $d'_{RW}(\mathbf{A},\mathbf{B})$, which is coined as the symmetrized normalised log-likelihood (SNLL) distance. These have the most solid theoretical foundation, and have also shown to be most successful in practice.

3.3 Implementation

We have outlined a scheme which can be used to cluster PolSAR data using pairwise affinities. This is a summary of the steps, as we have implemented it:

- A multi-look complex PolSAR image represented by coherency matrices is filtered by a polarimetric speckle filter. We have used a 7×7 refined Lee filter [13].
- Extract a data sample $\{\mathbf{T}_i\}_{i=1}^N$ by random sampling.
- Compute the affinity matrix $\mathbf{G}$ from Eq. (9), using one of the distances defined in Eq. (17) and (20), and a selected bandwidth b. This gives rise to the Bartlett Spectral Wishart (BSW) classifier and the SNLL Spectral Wishart (SSW) classifier, respectively.
- Eigendecompose $\mathbf{G}$ and form the eigenspace data representation $\mathbf{\Phi}$ from Eq. (10) for a selected dimension $U = k$.
- Cluster the eigenspace vectors $\{\boldsymbol{\phi}_i\}_{i=1}^N$ using Eq. (11).
- Use the spectral clustering result to estimate class mean coherency matrices $\{\hat{\mathbf{\Sigma}}_i\}_{i=1}^{k_E}$, and initialize the Wishart classifier in Eq. (7).
- Classify the whole image with the Wishart classifier.

4 Results

The algorithms are tested on a size 200×320 subset of a PolSAR image covering an agricultural area in Flevoland, The Netherlands, acquired by the NASA/JPL AIRSAR L-band sensor in August 1989 [14]. Although ground truth data exists for the Flevoland data set, the detail level is not seen as sufficient for our evaluation. We have therefore extracted samples of areas that should, based on their individual homogeneity and mutual distinctness, in our opinion be distinguished as separate classes. The extraction was supported by existing ground truth data as well as available C-band and P-band data [14]. The image subset is shown in figure 1, with the 10 selected ground truth classes indicated.

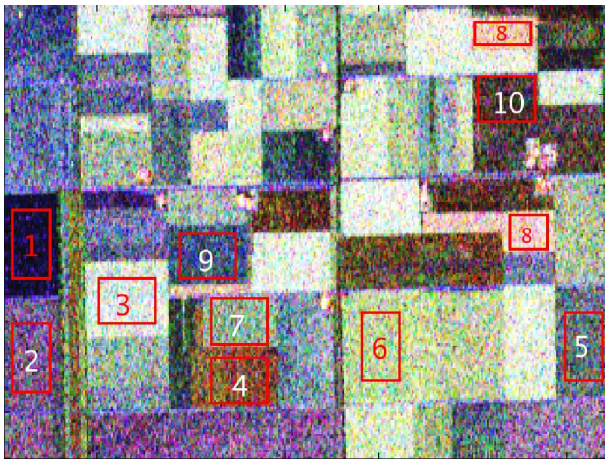

Fig. 1. Flevoland test data with extracted ground truth areas.

Evaluation of unsupervised classification is not trivial, since it may not be possible to relate the resulting partitioning directly and unambiguously to physical classes in the ground truth data. One ground truth class may be covered by two or more clusters, or one cluster may cover several ground truth classes. In order to obtain performance measures that reflect these problems, we adopt an evaluation procedure similar to the one found in [15].

We first identify the dominant predicted class label in each ground truth class, and calculate descriptivity D_i for i=1, . . . , 10, as the fraction of dominant predicted class labels to the total number of class labels. We next compute the matching matrix $\mathbf{M}$ (equivalent to the confusion matrix in supervised classification) with entries M_{ij}, indexed by row i and column j, defined as the fraction of the dominant predicted class labels in ground truth class i found in ground truth class j. We thus have $M_{ii} = D_i$. Furthermore, we define compactness

$$C_i = D_i - \sum_{j \neq i} M_{ij},\ i,j = 1,\ldots,10 \qquad (23)$$

and representivity

$$R_j = D_j - \sum_{i \neq j} M_{ij},\ i,j = 1,\ldots,10\ . \qquad (24)$$

We now have measures that can be interpreted as the homogeneity of the classification of a ground truth class (D_i), the uniqueness of a dominant class label within the whole set of dominant class labels (C_i), and the confinement of a dominant class label to the corresponding ground truth class (R_i). Negative values of C_i and R_i are set to zero, thus all measures range from zero to one. We shall see that R_i and C_i are highly correlated, and that all performance measures are affected by the number of effective classes, k_E.

We have tested and compared three different algorithms:

(i) the CPW classifier, (ii) the BSW classifier, (iii) and the SSW classifier. For a fair comparison, given the assumption of an unknown number of classes, all algorithms are evaluated for k=16, which equals the number of predetermined zones in H/A/α space [2]. All classifiers are evaluated for 1 and 10 iterations, respectively. The spectral clustering algorithms use a kernel bandwidth of σ=0.42, determined empirically. They are evaluated for a spectral clustering sample size of 1% to 10% of the total number of pixels in the subset, which is 64,000. The matching matrices and derived merits for the CPW, BSW and SSW classifier are summarised in table I. Here we have used 10 iterations of the Wishart classifier and N=6,400 in spectral clustering.

Fig. 2, 3, and 4 are plots of mean descriptability, mean compactness and mean representivity, averaged over all ground truth classes, as a function of spectral clustering sample size N. Fig. 2 shows that the CPW classifier gives the most homogeneous classification of individual ground truth classes. Variability for the other algorithms must be seen in relation to k_E, which is plotted against N in Fig. 5. All results are averages of 10 Monte Carlo runs. Fig. 3 and 4 show correlated results. The BSW classifier attains highest values of compactness and representivity, and the SSW classifier also outperforms the CPW classifier, both requiring that N exceeds a certain level. Fig. 5 shows that only 9 zones in H/A/α space contain data samples, thus k_E=9 for the CPW classifier. For the spectral clustering algorithms, classification is stochastic with respect to the required random sampling. Thus, k_E is stochastic, and variation is illustrated by standard deviation bars. Fig. 5 further demonstrates the data adaptive property of the BSW classifier, for which k is reduced to k_E during clustering in kernel space. We observe that the number of clusters that can be resolved from data increases, but non-uniformly, with N. This is well known behaviour in non-parametric estimation of multimodal functions. It is not known why the adaptiveness is different for the BSD and the SSW classifier.

Fig. 6 shows examples of classification results for the CPW, BSW, and SSW classifiers. The colour map of the figures have been rendered to make comparison of the results as easy as possible. However, correspondance to the same physical class is not guaranteed for between areas of equal colour in the different classification images. Visual inspection of the results are consistent with the results of table I and Fig. 2–4; The CPW classifier produces more homogeneous ground truth classes, but the other classifiers detect more true boundaries and delineate better between apparently distinct regions.

Finally, we have assessed the convergence properties of the algorithms. For all classifiers, we calculate the swap rate for each iteration, defined as the fraction of pixels

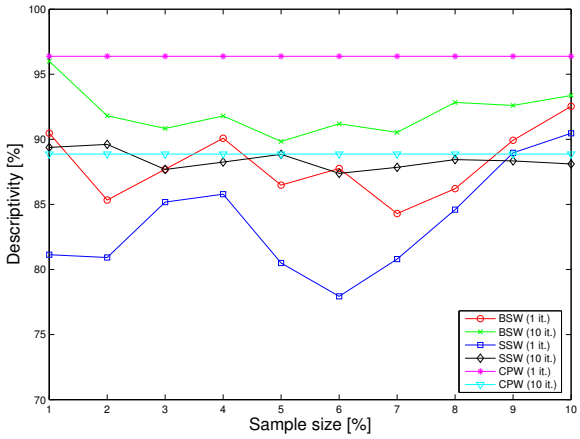

Fig. 2. Comparison of descriptivity for tested algorithms.

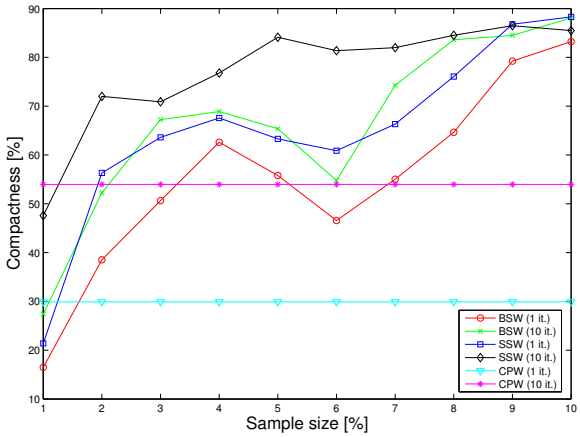

Fig. 3. Comparison of compactness for tested algorithms.

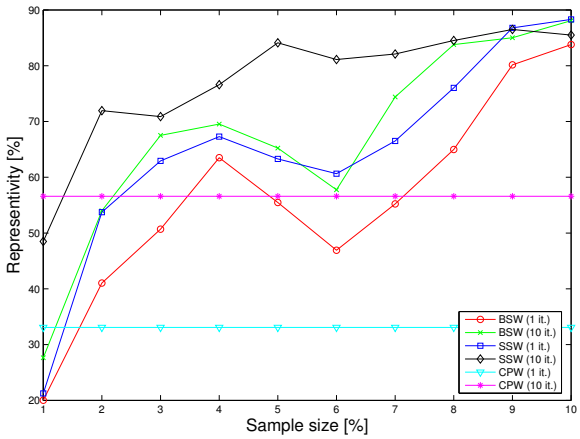

Fig. 4. Comparison of representivity for tested algorithms.

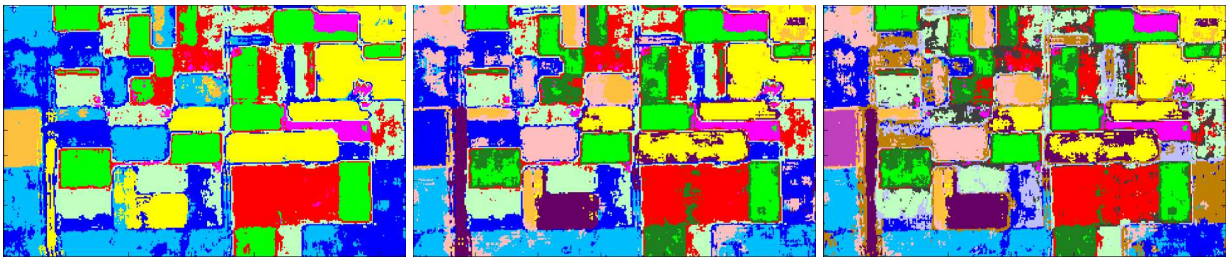

Fig. 6. Classification result for CPW classifier (left, 9 classes), BSW classifier (middle, 11 classes), and SSW classifier (right, 14 classes).

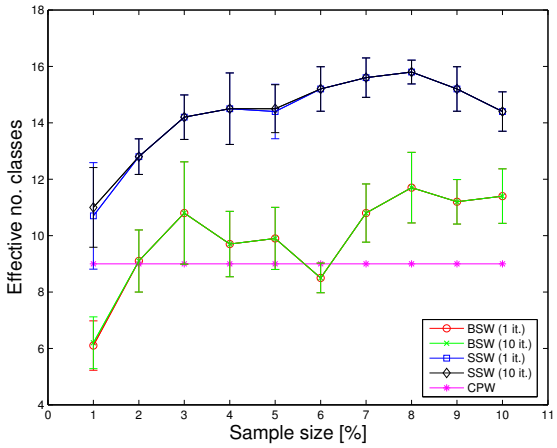

Fig. 5. Comparison of effective number of classes for tested algorithms.

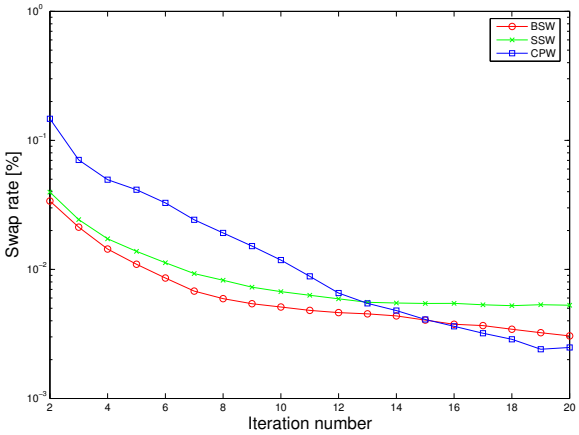

Fig. 7. Convergence of Wishart classifier for different initializations.

swapping class label from one iteration to the next. The result in Fig. 7 shows that the algorithms based on spectral clustering provide better initialization of the Wishart classifier than the H/A/α decomposition, assuming we use less than 13 iterations. We believe that the superior swap rates of the CPW classifier at high iteration numbers can be explained by the lower effective number of classes.

5 Conclusion

We have defined two distance measures that are suited for calculation of pairwise affinities for PolSAR data coherency matrices. We have further demonstrated how PolSAR data can be segmented by spectral clustering with coherency matrices as input features. Comparison with the Cloude-Pottier-Wishart classifier shows that spectral clustering provides a good initialization of the iterative minimum distance classification procedure, both in terms of convergence and classification accuracy. Future work will concentrate on methods for robust selection of the kernel bandwidth used in affinity calculation, and studies of the data adaptive selection of the number of classes, in order to develop and verify a fully automatic segmentation algorithm.

Acknowledgment

The authors would like to thank the Data Fusion Comittee of the IEEE Geoscience and Remote Sensing Society (GRSS-DFC) and Anthony Freeman at NASA/JPL for providing the Flevoland data set.

6 References

[1] J.-S. Lee, M. R. Grunes, and R. Kwok, "Classification of multi-look polarimetric SAR imagery based on complex Wishart distribution," *Int. J. Remote Sensing*, vol. 15, no. 11, pp. 229–231, 1994.

[2] E. Pottier and J.-S. Lee, "Unsupervised classification of POLSAR images based on the complex Wishart distribution and the H/A/α polarimetric decomposition theorem," in *Proc. 3rd EUSAR 2000 Conf.*, Munich, Germany, May 2000, pp. 265–268.

[3] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. and Prob.*, L. L. Cam and J. Neyman, Eds., vol. 1. Berkeley, USA: University of California Press, 2000, pp. 281–297.

[4] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral algorithms: Analysis and an algorithm," in *Advances in Neural Information Processing Systems 14*. Cambridge, USA: MIT Press, 2001, pp. 849–856.

[5] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the Nyström method," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 2, pp. 214–225, 2004.

[6] K. Ersahin, I. G. Cumming, and M. J. Yedlin, "Classification of polarimetric SAR data using spectral graph partitioning," in *Int. Geosc. Remote Sensing Symp., IGARSS'06*, Denver, USA, aug 2006, in press.

TABLE I

MATCHING MATRICES SHOWING CORRESPONDENCE OF GROUND TRUTH AND PREDICTED CLASSES

CPW classifier	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	C_i
GT1	99.8	0.2	0	0	0	0	0	0	0.2	0	99.5
GT2	0	99.5	0	0	0.5	0	0	0	99.5	0	0
GT3	0	0	99.7	0	0	0.3	0	0	0	0	99.3
GT4	0	0	0	92.8	7.2	0	0	0	0	92.8	0
GT5	0	20.2	0	0	79.8	0	0	0	20.2	0	39.5
GT6	0	0	0	0	0	100	0	0	0	0	100
GT7	0	0	0	0	0	0.3	99.7	0	0	0	99.3
GT8	0	0	3.0	0	0	1.5	0	95.5	0	0	91.0
GT9	0	100	0	0	0	0	0	0	100	0	0
GT10	0	0	0	100	0	0	0	0	0	100	0
R_i	99.8	0	96.7	0	72.2	97.8	99.7	95.5	0	7.1667	$\nwarrow D_i$
BSW classifier	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	C_i
GT1	99.8	0.03	0	0	0	0	0	0	0.1	0	99.7
GT2	0	98.9	0	0	19.5	0	0	0	1.1	0	78.9
GT3	0	0	72.6	0	0	0.1	0	4.8	0	0	67.7
GT4	0	0	0	98.5	1.0	0	0	0	0	0.5	97.0
GT5	0	18.3	0	2.5	73.2	0	0	0	18.0	0	38.3
GT6	0	0	0	0	0	99.6	0	10.1	0	0	89.5
GT7	0	0	0	0	0.2	0.5	98.7	0	0	0	98.1
GT8	0	0	6.6	0	0	7.2	0.1	88.6	0	0	79.4
GT9	0	0.4	0	0	10.2	0	0	0	99.1	0	88.6
GT10	0	0	0	0.8	0.03	0	0	0	0.2	96.3	95.3
R_i	99.8	80.2	66.1	95.2	52.4	91.8	98.6	78.3	79.7	95.8	$\nwarrow D_i$
SSW classifier	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	C_i
GT1	99.8	0	0	0	0	0	0	0	0.2	0	99.7
GT2	0	98.9	0	0	0.02	0	0	0	0.6	0	98.3
GT3	0	0	60.9	0	0	0.03	0	6.1	0	0	54.7
GT4	0	0	0	94.2	0.6	0	0	0	0	0	93.6
GT5	0	1.2	0	0.5	82.1	0	0.02	0	6.1	0	74.4
GT6	0	0	0	0	0	99.6	0	0	0	0	99.6
GT7	0	0	0	0	0.1	0.1	86.3	0	0	0	86.1
GT8	0	0	1.9	0	0	3.4	0	92.4	0	0	87.2
GT9	0	0.1	0	0	0.3	0	0	0	99.3	0	98.9
GT10	0	0	0	0.6	0.1	0	0	0	0.2	91.3	90.5
R_i	99.8	97.7	59.0	93.1	81.1	96.2	86.3	86.3	92.3	91.3	$\nwarrow D_i$

[7] J. J. van Zyl and F. T. Ulaby, "Scattering matrix representation for simple targets," in *Radar Polarimetry for Geoscience Applications*, F. T. Ulaby and C. Elachi, Eds. Norwood, USA: Artech House, 1990, ch. 2, pp. 17–52.

[8] R. Jenssen, "An information theoretic approach in machine learning," Ph.D. dissertation, Univ. of Tromsø, Tromsø, Norway, May 2005. [Online]. Available: http://www.phys.uit.no/ robertj/ROBERT_NOBS/Thesis.html

[9] R. Jenssen, D. Erdogmus, J. C. Principe, and T. rn Eltoft, "Information theoretic angle-based spectral clustering: A theoretic analysis and an algorithms," in *Proc. Int. Joint Conf. on Neural Networks (IJCNN2006)*, Vancouver, Canada, July 2006, pp. 4904–4911.

[10] C. K. Williams and M. Seeger, "Using the Nyström method to speed up kernel machines," in *Advances in Neural Information Processing Systems 14*. MIT Press, 2001, pp. 682–688.

[11] K. Conradsen, A. A. Nielsen, J. Schou, and H. Skriver, "A test statistic in the complex Wishart distribution and its application to change detection in polarimetric SAR data," *IEEE Trans. Geosci. Remote Sensing*, vol. 41, no. 1, pp. 4–19, 2003.

[12] P. R. Kersten, J.-S. Lee, and T. L. Ainsworth, "Unsupervised classification of polarimetric synthetic aperture radar images using fuzzy clustering and EM clustering," *IEEE Trans. Geosci. Remote Sensing*, vol. 43, no. 3, pp. 519–527, 2005.

[13] J.-S. Lee, M. R. Grunes, and G. de Grandi, "Polarimetric SAR speckle filtering and its implication for classification," *IEEE Trans. Geosci. Remote Sensing*, vol. 37, no. 5, pp. 2363–2373, 1999.

[14] Data set GRSS_DFC_0004, IEEE GRSS Data Fusion reference database (http://www.dfc-grss.org/), 2000.

[15] L. Ferro-Famil, E. Pottier, and J.-S. Lee, "Unsupervised classification of multifrequency and fully polarimetric SAR images," *IEEE Trans. Geosci. Remote Sensing*, vol. 39, no. 11, pp. 2332–2342, 2001.