Classification of Hyperspectral Images using Mixture of Probabilistic PCA Models

Sezer Kutluk
Dept. of Electrical-Electronics Engineering
Istanbul University
Istanbul, Turkey
Email: sezer.kutluk@gmail.com

Koray Kayabol
Dept. of Electronics Engineering
Gebze Technical University
Kocaeli, Turkey
Email: koray.kayabol@gtu.edu.tr

Aydin Akan
Dept. of Electrical-Electronics Engineering
Istanbul University
Istanbul, Turkey
Email: akan@istanbul.edu.tr

***Abstract*—We propose a supervised classification and dimensionality reduction method for hyperspectral images. The proposed method contains a mixture of probabilistic principal component analysis (PPCA) models. Using the PPCA in the mixture model inherently provides a dimensionality reduction. Defining the mixture model to be spatially varying, we are also able to include spatial information into the classification process. In this way, the proposed mixture model allows dimensionality reduction and spectral-spatial classification of hyperspectral image at the same time. The experimental results obtained on real hyperspectral data show that the proposed method yields better classification performance compared to state of the art methods.**

***Index Terms*—hyperspectral image, probabilistic principal component analysis, dimensionality reduction, mixture models.**

I. Introduction

Hyperspectral imaging has become one of the main research topics of remote sensing in recent years. A significant application of hyperspectral imaging is the identification of land cover areas via classification. The rich content of hyperspectral data allows the recognition and classification of forests, urban areas, crop species and water resources.

In this study, the aim is to predict the class labels of the pixels in a hyperspectral image. The proposed method consists of a Gaussian mixture model (GMM) whose parameters are defined by probabilistic PCA model. Considering that a land cover area consists of several pixels, it is natural that the neighboring pixels are most likely in the same class. Along with the mixture model, we use a random field model to include the neighborhood information by defining the mixture proportions to be spatially varying. Thus a segmentation map is also obtained when the pixels in the image are classified using our proposed mixture model. The proposed mixture model allows dimensionality reduction and spectral-spatial classification of hyperspectral image at the same time.

A common problem encountered in the supervised classification of hyperspectral images is that the length of the feature vectors, which is actually the number of spectral bands, is big, while the size of samples in the training set is small which causes an under-determined parameter estimation problem for GMMs. The most common approach is the application of PCA before classification to reduce the dimension of spectral bands as proposed in [12]. In [1] a dimensionality reduction step is applied before training a GMM. In [2] the number of parameters to be estimated is decreased by defining constraints on the covariance matrices of GMM. In [3] a Bayesian GMM is proposed in order to overcome the under-determined problem and this model also includes a spatial regression model over the class labels that was used in [4] and [5]. This regression model is based on the assumption that the mixture model is not stationary over space, i.e. that its mixture proportions are varying over space.

Non-stationary mixture models have been used in several works for image segmentation and classification. For instance, in [6], [7], [8] and [9] the non-stationary mixture model is obtained using a Markov Random Field prior. In [10] and [11] a latent Gaussian random field was proposed where the mixture proportions were related to the class labels using a multinomial logistic function.

One of the most common methods for dimensionality reduction is the principal component analysis (PCA). In [12] the spectral bands in the data obtained by two different sensors were analyzed using PCA, and the classification performance using only the significant bands was investigated. In [13] a method was proposed in which the hyperspectral image is divided into pieces spectrally or spatially, and PCA is applied to each piece separately, and pieces are patched up again in order to go under supervised classification.

Although PCA is not essentially a probabilistic approach, in [14], it is shown that its probabilistic derivation is possible. The approach proposed in [14] is called PPCA and is based on a latent variable model. In this work we use PPCA for dimensionality reduction.

In [15] PPCA is used for feature extraction of hyperspectral images in a supervised (SPPCA) and semi-supervised (S^2PPCA) setting. In SPPCA only the labeled samples are used, and in S^2PPCA the unlabeled samples are utilized as well. The extracted features are used for classification with nearest neighbor and support vector machine classifiers. In these classifications only the spectral information is used.

The rest of this paper is organized as follows: In Section II, GMM, PPCA and the spatial regression model are described.

This work is supported by Scientific and Technological Research Council of Turkey (TUBITAK) under Project No. 114E535.

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

In Section III test results obtained using various hyperspectral image datasets are given. Section IV consists of discussions and conclusion.

II. Proposed Method

A pixel is denoted by the vector $\mathbf{s}_n$ of length L in a hyperspectral image which has N pixels and L spectral bands. Each element of a pixel vector comes from a spectral band, therefore a hyperspectral image can be considered as a collection of L different images.

In this study, the hyperspectral image is modeled as a mixture of Gaussians, therefore each pixel is assumed to be generated from a different multivariate Gaussian distribution. As opposed to the similar works that use GMM, we use a PPCA model for each class in order to estimate the parameters of Gauss distribution. This model also involves dimensionality reduction for which PCA is mainly used. In addition to spectral modeling, an auto-logistic regression model is used in order to take advantage of pixel neighborhoods for classification.

A. Gaussian Mixture Model

We assume that a pixel vector is generated from one of K multivariate Gaussian distributions each of which represents a class. Probability of a pixel can be written as follows:

$$p(\mathbf{s}_n) = \sum_{k=1}^{K} \omega_{n,k} \mathcal{N}(\mathbf{s}_n | \mathbf{m}_k, \mathbf{\Sigma}_k) \tag{1}$$

In this equation $\mathcal{N}(\cdot)$ is the Gaussian distribution, $\mathbf{m}_k$ is the mean of the kth class, and $\mathbf{\Sigma}_k$ is its covariance matrix. The parameter $\omega_{n,k}$ is the mixture proportion of kth class. In this model mixture proportions are different for each pixel, thus spatial information can be used in the model.

Parameters $\mathbf{m}_k$ and $\mathbf{\Sigma}_k$ must be estimated for each class using the training data. In this study, we use PPCA [14] for parameter estimation. This method allows dimensionality reduction and parameter estimation of a Gaussian distribution by a latent variable model. Thus, a GMM is trained by applying PPCA to each class in the training set.

B. Probabilistic Principal Component Analysis

PCA is a widely used statistical method in data processing. It is an orthogonal transform that projects the data into a space where the variance is maximum. PCA is generally used as a dimensionality reduction method since only the first few principal components contain most of the variance in the data.

In [14], it was shown that PCA can be performed with a probabilistic approach. In this approach a latent variable model is used and model parameters are estimated using the maximum likelihood estimation method in order to obtain the principal axes. This latent variable model aims at finding the relation between an observation vector $\mathbf{s}_n$ of length L and a latent variable vector $\mathbf{x}_n$ of length q while $q < L$. With the assumption that this relation is linear, we can write the following equation:

$$\mathbf{s}_n = \mathbf{W}_k \mathbf{x}_n + \mathbf{m}_k + \boldsymbol{\epsilon}_k \tag{2}$$

where $\boldsymbol{\epsilon}_k$ is a zero-mean Gaussian random vector with covariance matrix $\sigma^2 \mathbf{I}_q$. Here matrix $\mathbf{W}_k$ of size $L \times q$ is the relation between the observed and latent variables. $\mathbf{m}_k$ is the mean, and $\boldsymbol{\epsilon}_k$ is the error, or additive noise.

Prior probability of $\mathbf{x}_n$ is a zero-mean Gaussian that is $p(\mathbf{x}_n) = \mathcal{N}(0, \mathbf{I}_q)$, and $\mathbf{I}_q$ is the identity matrix of size q.

It can be seen that $p(\mathbf{s}_n|\mathbf{x}_n)$ is also Gaussian under the assumption that the additive noise $\boldsymbol{\epsilon}_k$ is Gaussian:

$$p(\mathbf{s}_n|\mathbf{x}_n) = \mathcal{N}(\mathbf{W}_k \mathbf{x}_n + \mathbf{m}_k, \sigma_k^2 \mathbf{I}_L) \tag{3}$$

As shown in [14], marginal probability of $\mathbf{s}_n$ is also a Gaussian:

$$p(\mathbf{s}_n) = \int p(\mathbf{s}_n|\mathbf{x}_n) p(\mathbf{x}_n) d\mathbf{x}_n = \mathcal{N}(\mathbf{m}_k, \mathbf{\Sigma}_k). \tag{4}$$

The Gaussian distribution in the above equation constitutes a class in the mixture model. In [14] the covariance matrix is defined as $\mathbf{\Sigma}_k = \mathbf{W}_k \mathbf{W}_k^T + \sigma_k^2 \mathbf{I}_L$. By these definitions, the log-likelihood function can be written as follows:

$$\mathcal{L} = -\frac{N_k}{2} \{ d \ln(2\pi) + \ln|\mathbf{\Sigma}_k| + tr(\mathbf{\Sigma}_k^{-1} \mathbf{S}_k) \} \tag{5}$$

where N_k is the number of pixels in class k, and $\mathbf{S}_k$ is the sample covariance matrix which can be calculated as follows:

$$\mathbf{S}_k = \frac{1}{N_k} \sum_{n=1}^{N_k} (\mathbf{s}_n - \hat{\mathbf{m}}_k)(\mathbf{s}_n - \hat{\mathbf{m}}_k)^T \tag{6}$$

Here $\hat{\mathbf{m}}_k$ is the maximum likelihood estimation of the mean and can be calculated as follows:

$$\hat{\mathbf{m}}_k = \frac{1}{N_k} \sum_{n=1}^{N_k} \mathbf{s}_n \tag{7}$$

It can be shown that the matrix $\mathbf{W}_k$ that maximizes the log-likelihood function in (5) is as follows [14]:

$$\hat{\mathbf{W}}_k = \mathbf{U}_q (\mathbf{\Lambda}_q - \sigma_k^2 \mathbf{I})^{1/2} \mathbf{R} \tag{8}$$

Here the columns of the matrix $\mathbf{U}_q$ of size $L \times q$ are the eigenvectors of the matrix $\mathbf{S}_k$ and the corresponding eigenvalues $\lambda_{k,1}, \cdots \lambda_{k,q}$ constitute the $q \times q$ diagonal matrix $\mathbf{\Lambda}_q$. $\mathbf{R}$ is an arbitrary orthogonal rotation matrix of size $q \times q$ which is choosen as the identity matrix in this study.

Rather than using the same approach with [14], in this study we assume that the variance is distributed by an Inverse-Gamma distribution:

$$\sigma_k^2 \sim \text{Inv} - \text{Gamma}(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma_k^2)^{-\alpha-1} \exp(-\beta/\sigma_k^2) \tag{9}$$

Maximum-a-posteriori estimation of σ_k^2 is calculated as follows:

$$\hat{\sigma_k^2} = \frac{N_k \sum_{i=q+1}^{L} \lambda_{k,i} + 2\beta}{N_k(L-q) + 2(\alpha+1)} \tag{10}$$

C. Spatial Model

We define a label vector $\mathbf{z}_n \in \{0,1\}^K$ of length K for each pixel in a hyperspectral image that contains K land cover classes. We assume that the binary label vector $\mathbf{z}_n$ has the property that $\sum_{k=1}^{K} z_{n,k} = 1$. Thus $\mathbf{z}_n \in \{[1,0,\ldots,0],[0,1,\ldots,0],\ldots,[0,0,\ldots,1]\}$. We assume that $\mathbf{s}_n$ vectors are conditionally independent given class labels $\mathbf{z}_n$, but $\mathbf{z}_n$ vectors are spatially dependent. The joint probability of $\mathbf{s}_n$ and $\mathbf{z}_n$ can be written as follows:

$$p(\mathbf{s}_n, \mathbf{z}_n|\theta_{1:K}, \beta) \propto p(\mathbf{s}_n|\mathbf{z}_n, \theta_{1:K}) p(\mathbf{z}_n|\mathbf{z}_{\mathcal{M}(n)}, \beta) \quad (11)$$

where $p(\mathbf{z}_n|\mathbf{z}_{\mathcal{M}(n)}, \beta)$ are the prior probabilities of class labels, $\mathcal{M}(n)$ shows the pixels around the pixel n, and β is the smoothing parameter. We define an auto-logistic model over class labels which are the latent variables. According to this model, conditional probability of a class label can be written as follows [16]:

$$p(z_{n,k}|z_{\mathcal{M}(n),k}, \beta) \propto e^{\beta\left(z_{n,k} + z_{n,k}\sum_{m\in\mathcal{M}(n)} z_{m,k}\right)} \quad (12)$$

The joint probability of $z_{n,k}$, $n = 1,\ldots,N$ can be obtained by multiplying the conditional probabilities. As opposed to [16], in this model a larger area is used as $\mathcal{M}(n)$. β is chosen to be homogeneous for regression, therefore we can find a proper joint probability for the binary image. This auto-logistic model was used before in [4] and [5].

Probability of $\mathbf{s}_n$ can be written as the marginal probability of the joint probability $p(\mathbf{s}_n, \mathbf{z}_n|\theta_{1:K}, \mathbf{z}_{\mathcal{M}(n)}, \beta) = p(\mathbf{s}_n|\mathbf{z}_n, \theta_{1:K}) p(\mathbf{z}_n|\mathbf{z}_{\mathcal{M}(n)}, \beta)$ with respect to the latent variable vector $\mathbf{z}_n$ as follows:

$$p(\mathbf{s}_n|\theta_{1:K}, \mathbf{z}_{\mathcal{M}(n)}, \beta) = \sum_{\mathbf{z}_n} \prod_{k=1}^{K} [p(\mathbf{s}_n|\theta_k)\omega_{n,k}]^{z_{n,k}} \quad (13)$$

where $\omega_{n,k}$ shows the non-stationary mixture proportions that can be found using (12) as follows:

$$\omega_{n,k} = p(z_{n,k} = 1|z_{\mathcal{M}(n),k}, \beta) = \frac{e^{\beta v_{n,k}}}{\sum_{j=1}^{K} e^{\beta v_{n,j}}} \quad (14)$$

where $v_{n,k} = 1 + \sum_{m\in\mathcal{M}(n)} z_{m,k}$.

D. Classification

After the parameters $\hat{\theta}_{1:K} = \{\hat{\mathbf{m}}_{1:K}, \hat{\Sigma}_{1:K}\}$ of the mixture are estimated, new observations can be classified using the model. Classification is performed by maximizing the posterior probability of class labels $\mathbf{z}_{test}$. This posterior probability can be factorized as follows:

$$p(\mathbf{z}_{1:N}, \beta|\mathbf{s}_{1:N}, \hat{\theta}_{1:K}) \propto p(\mathbf{s}_{1:N}|\mathbf{z}_{1:N}, \hat{\theta}_{1:K}) p(\mathbf{z}_{1:N}|\beta) \quad (15)$$

Here we can write the likelihood as follows:

$$p(\mathbf{s}_{1:N}|\mathbf{z}_{1:N}, \hat{\theta}_{1:K}) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(\mathbf{s}_n|\hat{\theta}_k)^{z_{n,k}}$$

Using the conditionals $p(\mathbf{z}_n|\mathbf{z}_{\mathcal{M}(n)}, \beta)$ given in (12), we can write the joint probability of class labels $p(\mathbf{z}_{1:N}|\beta)$ given in (15).

With the conditional independence assumption, that is $p(\mathbf{z}_{n'}|\mathbf{z}_{\{1:N\}\setminus n'}, \beta) = p(\mathbf{z}_{n'}|\mathbf{z}_{\mathcal{M}(n')}, \beta)$, joint probability of the random field can be written as follows:

$$p(\mathbf{z}_{1:N}|\beta) = \frac{\prod_{k=1}^{K} \exp\left\{\beta \sum_{n=1}^{N} z_{n,k}\left(1 + \frac{1}{2}\sum_{m\in\mathcal{M}(n)} z_{m,k}\right)\right\}}{\mathcal{Z}(\beta)} \quad (16)$$

where $\mathcal{Z}(\beta)$ is the normalization term.

We use the iterated conditional mode (ICM) algorithm in order to predict class labels. Variables are updated iteratively as follows:

$$\begin{aligned} \mathbf{z}_n^t &\leftarrow \max_{\mathbf{z}_n} p(\mathbf{s}_n|\mathbf{z}_n, \phi_{1:K}^{t-1}) p(\mathbf{z}_n|\mathbf{z}_{\mathcal{M}(n)}^{t-1}, \beta^{t-1}) \\ \beta^t &\leftarrow \max_{\beta} \prod_{n=1}^{N} p(\mathbf{z}_n^t|\mathbf{z}_{\mathcal{M}(n)}^t, \beta) \end{aligned} \quad (17)$$

In the update rules above $n = 1,..,N$, $k = 1,..,K$, and t is the time index. For the estimation of β we use the pseudo-likelihood approach given in [16].

III. Test Results

In order to evaluate the performance of the proposed model, we performed tests using the hyperspectral image datasets named Indian Pines, Pavia Centre, Pavia University, and Salinas, which are widely used in the literature. We construct the training set by randomly choosing 10 pixels from each class, and use the rest of the pixels as test sets.

Ground truth maps and class names of Indian Pines, Pavia Centre, Pavia University and Salinas datasets are given in Figures 1, 2, 3 and 4, respectively.

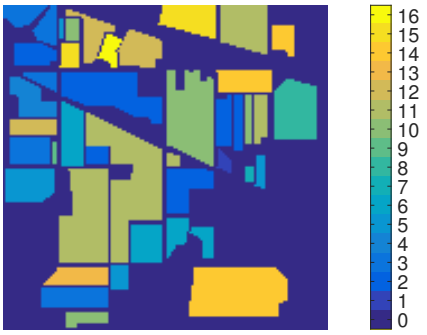

Fig. 1. Indian Pines dataset with classes (0) Background, (1) Alfalfa, (2) Corn-notill, (3) Corn-mintill, (4) Corn, (5) Grass-pasture, (6) Grass-trees, (7) Grass-pasture-mowed, (8) Hay-windrowed, (9) Oats, (10) Soybean-notill, (11) Soybean-mintill, (12) Soybean-clean, (13) Wheat, (14) Woods, (15) Buildings-Grass-Trees-Drives, and (16) Stone-Steel-Towers.

Test results are given in Table I. This table shows the average classification accuracies of 20 different runs. In this table, we can see the results of PPCA with different cases that the number of principal components are chosen to be 5, 10,

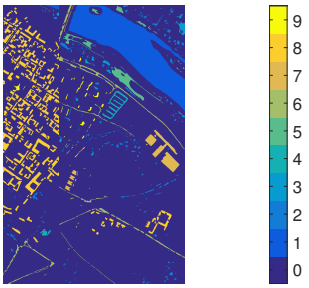

Fig. 2. Pavia Centre dataset with classes (0) Background, (1) Water, (2) Trees, (3) Asphalt, (4) Self-Blocking Bricks, (5) Bitumen , (6) Tiles, (7) Shadows, (8) Meadows, (9) Bare Soil

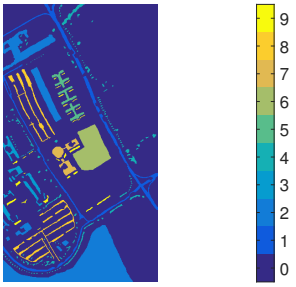

Fig. 3. Pavia University dataset with classes (0) Background, (1) Asphalt, (2) Meadows, (3) Gravel, (4) Trees, (5) Painted metal sheets, (6) Bare Soil, (7) Bitumen, (8) Self-Blocking Bricks, (9) Shadows.

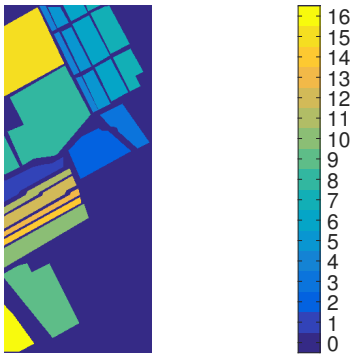

Fig. 4. Salinas dataset with classes (0) Background, (1) Brocoli green weeds 1, (2) Brocoli green weeds 2, (3) Fallow, (4) Fallow rough plow, (5) Fallow smooth, (6) Stubble, (7) Celery, (8) Grapes untrained, (9) Soil vinyard develop, (10) Corn senesced green weeds, (11) Lettuce romaine 4wk, (12) Lettuce romaine 5wk, (13) Lettuce romaine 6wk, (14) Lettuce romaine 7wk, (15) Vinyard untrained, (16) Vinyard vertical trellis.

and 15 which are shown as PPCA-5, PPCA-10, and PPCA-15, respectively. In addition, the table shows the results obtained using the Bayesian GMM approach proposed in [3], and using another approach that PCA dimension reduction is applied before GMM classification (PCA+GMM) for benchmarking.

In PCA+GMM, PCA is applied to the whole image before training and test set separation. Moreover, although the tests were repeated using different number of principal components, the highest classification accuracies for all datasets were obtained using only the first few principal components. The best results of (PCA+GMM) are given in Table I. As seen from this table, the proposed method yields better classification results than those of other two methods.

Example classification maps obtained from the tests are given in Figure 5.

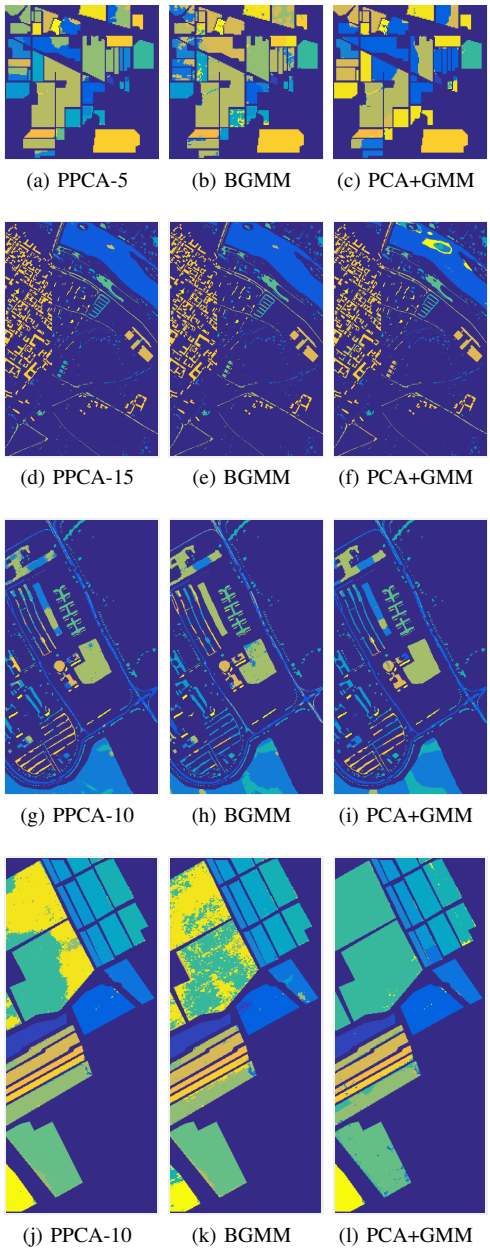

Fig. 5. Classification maps for Indian Pines (5a, 5b, 5c), Pavia Centre (5d, 5e, 5f), Pavia University (5g, 5h, 5i), and Salinas (5j, 5k, 5l) datasets.

	PPCA-5	PPCA-10	PPCA-15	BGMM	PCA+GMM
Indian Pines	**73.18**	69.17	71.62	65.04	57.40
Pavia Centre	96.04	96.13	**96.16**	95.03	88.67
Pavia University	81.26	**82.86**	82.10	68.11	71.97
Salinas	87.00	**88.58**	87.22	82.74	81.56

TABLE I
CLASSIFICATION PERFORMANCE OF THE PROPOSED METHOD COMPARED WITH TWO OTHER APPROACHES

IV. CONCLUSION

In this study a mixture model for dimensionality reduction and pixel-based contextual classification of hyperspectral images is proposed. Dimensionality reduction before classification is a common technique for hyperspectral images. Rather than performing two tasks separately, the PPCA-based method proposed in this study performs both dimensionality reduction and classification in one model with higher accuracies. Future work will be the automatic determination of the number of relevant principle components.

ACKNOWLEDGMENT

The authors would like to thank David Landgrebe and Paolo Gamba for providing hyperspectral data sets acquired by AVIRIS and ROSIS sensors, respectively.

REFERENCES

[1] W. Li, S. Prasad, and J. E. Fowler, “Hyperspectral image classification using Gaussian mixture models and Markov random fields", IEEE Geosci. Remote Sens. Lett., vol. 11, no. 1, pp. 153–157, 2014.
[2] A. Berge and A. H. S. Solberg, “Structured Gaussian components for hyperspectral image classification", IEEE Trans. Geosci. Remote Sens., vol. 44, no. 11, pp. 3386–3396, 2006.
[3] K. Kayabol, “Bayesian Gaussian mixture model for spatial-spectral classification of hyperspectral images", 23rd European Signal Processing Conference (EUSIPCO), pp. 1845–1849, 2015.
[4] K. Kayabol, A. Voisin, and J. Zerubia, “SAR image classification with non-stationary multinomial logistic mixture of amplitude and texture densities", Int. Conf. Image Process. ICIP, pp. 173–176, 2011.
[5] K. Kayabol and J. Zerubia, “Unsupervised amplitude and texture classification of SAR images with multinomial latent model", IEEE Trans. Image Process., vol. 22, no. 2, pp. 561–572, 2013.
[6] S. Sanjay-Gopal and T. J. Hebert, “Bayesian pixel classification using spatially variant finite mixtures and the generalized EM algorithm", IEEE Trans. Image Process., vol. 7, no. 7, pp. 1014–1028, 1998.
[7] Y. Zhang, M. Brady, and S. Smith, “Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm", IEEE Trans. Medical Imaging, vol. 20, no. 1, pp. 45–57, 2001.
[8] K. Blekas, A. Likas, N. P. Galatsanos, and I. E. Lagaris, “A spatially constrained mixture model for image segmentation", IEEE Trans. Neural Netw., vol. 16, no. 2, pp. 494–498, 2005.
[9] A. Diplaros, N. A. Vlassis, and T. Gevers, “A spatially constrained generative model and an EM algorithm for image segmentation", IEEE Trans. Neural Netw., vol. 18, no. 3, pp. 798–808, 2007.
[10] C. Fernandez and P. J. Green, “Modelling spatially correlated data via mixtures: A Bayesian approach", J. R. Stat. Soc. B, vol. 64, no. 4, pp. 805–826, 2002.
[11] M. A. T. Figueiredo, “Bayesian image segmentation using Gaussian field priors", Energy Minimization Methods Computer Vision and Pattern Recognition, vol. LNCS 3757, pp. 74–89, 2005.
[12] C. Rodarmel and J. Shan, “Principal component analysis for hyperspectral image classification", Surveying and Land Information Systems, Vol. 62, No. 2, pp.115-000, 2002.
[13] A. Agarwal, T. El-Ghazawi, H. El-Askary and J. Le-Moigne, “Efficient hierarchical-PCA dimension reduction for hyperspectral imagery", IEEE International Symposium on Signal Processing and Information Technology, 2007.
[14] M. E. Tipping and C. M. Bishop, “Probabilistic principal component analysis", Journal of the Royal Statistical Society, Series B, 61, Part 3, pp.611-622, 1999.
[15] J. Xia, J. Chanussot, P. Du and X. He, “(Semi-) supervised probabilistic principal component analysis for hyperspectral remote sensing image classification", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol.7, no.6, pp.2225-2237, 2014.
[16] J. Besag, “Spatial interaction and the statistical analysis of lattice systems", J. R. Stat. Soc. B, vol. 36, no. 2, pp. 192–236, 1974.