Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

# Clustering-based hyperspectral band selection using sparse nonnegative matrix factorization*

Ji-ming LI†[1,2], Yun-tao QIAN[1]

(*[1]School of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*)
(*[2]Zhejiang Police College, Hangzhou 310053, China*)
†E-mail: ljming@zju.edu.cn

Received Aug. 30, 2010; Revision accepted Nov. 30, 2010; Crosschecked May 31, 2011

**Abstract:** Hyperspectral imagery generally contains a very large amount of data due to hundreds of spectral bands. Band selection is often applied firstly to reduce computational cost and facilitate subsequent tasks such as land-cover classification and higher level image analysis. In this paper, we propose a new band selection algorithm using sparse nonnegative matrix factorization (sparse NMF). Though acting as a clustering method for band selection, sparse NMF need not consider the distance metric between different spectral bands, which is often the key step for most common clustering-based band selection methods. By imposing sparsity on the coefficient matrix, the bands' clustering assignments can be easily indicated through the largest entry in each column of the matrix. Experimental results showed that sparse NMF provides considerable insight into the clustering-based band selection problem and the selected bands are good for land-cover classification.

**Key words:** Hyperspectral, Band selection, Clustering, Sparse nonnegative matrix factorization
**doi:**10.1631/jzus.C1000304 **Document code:** A **CLC number:** TP75

## 1 Introduction

Hyperspectral sensors collect imagery simultaneously in hundreds of narrow and continuous spectral bands, with a much finer spectral resolution (e.g., 0.01 μm) compared to traditional multispectral techniques. As a result, the three-dimensional (3D) image cube obtained usually contains a large amount of information for computer processing, with the third dimension specifying the spectral bands. It is often time consuming to process these high dimensional data in tasks like land-cover classification or other high level image analysis if all spectral bands are included. Moreover, due to the high correlation between the contiguously spaced spectral bands, redundancies that do not contribute to the model's discriminative power should also be removed. Hence, band selection, or feature selection in hyperspectral data, which is the procedure of selecting the relevant wavelengths in the range of spectrum while keeping the classification accuracy for land-cover discrimination or material identification tasks, is often an essential preprocessing step for hyperspectral classification. The band selection in a hyperspectral data classification problem should be aimed at improving the classification accuracy of the classifier, making the classification procedure more cost-effective and faster, and achieving a better data compression for the original hyperspectral data cube.

Much research has been done on hyperspectral band selection during the past decade. Cheng *et al.* (2006) used the logistic regression model for both band selection and classification, and performed band selection with sequential forward selection. Keshava (2004) developed a method called 'band add-on' that incrementally selects bands to

* Project (No. 60872071) supported by the National Natural Science Foundation of China

increase the angular separation between two spectra. Chang *et al.* (1999) presented an algorithm that comprises band prioritization and band decorrelation, and showed that the algorithm is very effective in eliminating the insignificant bands. Wang and Chang (2007) gave a variable-number variable-band selection method, which determines the number of selected bands through hyperspectral signature's spectral shapes. Archibald and Fann (2007) used a support vector machine (SVM) with embedded-feature-selection (EFS) to achieve a representative subset of bands. Most of the above methods have contributed to the band selection problem more or less. However, none of these studies has been shown to be a superior method, which can be independent of any premise or hypothesis. Therefore, it may be valuable to view the band selection problem from a new perspective and experiment with new methods. An exceptional set of feature selection methods are called feature construction methods (Guyon and Elisseeff, 2003). Feature construction begins with the careful consideration of appropriate data representations. Performance is often enhanced with features derived from the original input. There are many feature construction methods, such as the basic linear transforms of the input features, including principal component analysis (PCA) (Jolliffe, 2002), independent component analysis (ICA) (Hyvärinen and Oja, 2000), and Fisher linear discriminant analysis (LDA) (Fisher, 1936), and clustering methods. Many of these methods are often related to feature extraction concepts.

In this paper, we view the band selection problem as a feature construction problem. We use an unsupervised clustering method for our task. Clustering has been used for feature construction for a fairly long time, and has also been experimented for band selection in recent research (Martinez-Uso *et al.*, 2007). The main idea is to represent a group of 'similar' features by a cluster center or representative exemplar (Qian *et al.*, 2009), which often forms an effective feature reduction in replacing the original feature group. The most popular algorithms include $k$-means and hierarchical clustering. Martinez-Uso *et al.* (2007) had proved the clustering method's high efficiency and pointed out the importance for selecting suitable distance measures between different bands. In this paper, we use the sparse nonnegative matrix factorization (NMF) algorithm in Kim and Park (2008) for band clustering. Though acting as a clustering method for band selection, sparse NMF need not consider distance measures between different spectral bands, which is the key step for most common clustering methods. By imposing sparsity on the coefficient matrix, it easily indicates the clustering membership through the largest entry in each column of the matrix.

## 2 Sparse NMF for band selection

For simplicity, we describe the hyperspectral data with the following notations. As shown in Fig. 1, hyperspectral imagery is a 3D data cube with the width and length corresponding to spatial dimensions and the third dimension corresponding to the spectral domain, which are denoted by $M$, $N$, and $L$ in sequence. $\boldsymbol{R}$ is the image cube with each band $\boldsymbol{R}_l \in \mathbb{R}^{M\times N}$ being a gray-scale image matrix.

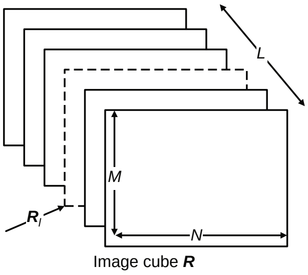

**Fig. 1 The sketch map of hyperspectral imagery**

### 2.1 Nonnegative matrix factorization

NMF was first proposed for finding part-based, linear representations of nonnegative data (Lee and Seung, 1999; 2001; Hoyer, 2004). It has been proved useful for modeling nonnegative data such as images. Given a set of data samples represented in a matrix form, which has only nonnegative entries, NMF aims to find a lower rank factor analysis while approximating the data matrix, with the factors also required to be negative. Let us represent the input data matrix with $\boldsymbol{V} \in \mathbb{R}^{m\times n}$, where each column represents a sample and each row represents a feature. For a given integer $k$ such that $k < \min\{m, n\}$, NMF seeks to find an approximate factorization $\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H}^{\mathrm{T}}$ into nonnegative factors $\boldsymbol{W} \in \mathbb{R}^{m\times k}$ and $\boldsymbol{H} \in \mathbb{R}^{n\times k}$.

The solving problem of $\boldsymbol{W}$ and $\boldsymbol{H}$ is illustrated as

$$\min_{\boldsymbol{W},\boldsymbol{H}} f_k(\boldsymbol{W},\boldsymbol{H}) \equiv \frac{1}{2} \parallel \boldsymbol{V}-\boldsymbol{W}\boldsymbol{H}^{\mathrm{T}} \parallel_{\mathrm{F}}^2 \ \text{s.t.}\ \boldsymbol{W},\boldsymbol{H} \geq \boldsymbol{0}, \tag{1}$$

where the subscript $k$ in $f_k$ denotes the desired low rank $k$, $\boldsymbol{W}$ is often named the basis matrix, and $\boldsymbol{H}$ is named the coefficient matrix.

### 2.2 NMF for clustering

The main purpose of most matrix factorization type methods (e.g., PCA, ICA, and NMF) is to find the latent and meaningful structure hiding in the data, which could be much more effective than the original data representation for the postprocessing task such as classification or clustering, while the goal of many clustering methods is to find a prototype for representing each cluster. Some researchers have found that NMF can be interpreted as a clustering scheme through matrix operation techniques (Xu *et al.*, 2003; Ding *et al.*, 2005; Shahnaz *et al.*, 2006). Kim and Park (2008) showed how $k$-means can be formulated as NMF and built a connection with the clustering method for NMF.

In the $k$-means algorithm, the objective function to be minimized is the sum of squared Euclidean distances from each data sample to its cluster centroid. With $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_n] \in \mathbb{R}^{m\times n}$, the objective function $J_k$ for a given cluster number $k$ can be written as

$$J_k = \sum_{j=1}^{k} \sum_{\boldsymbol{v}_i \in \boldsymbol{C}_j} \parallel \boldsymbol{v}_i - \boldsymbol{c}_j \parallel^2 = \parallel \boldsymbol{V} - \boldsymbol{C}\boldsymbol{B}^{\mathrm{T}} \parallel_{\mathrm{F}}^2, \tag{2}$$

where $\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_k] \in \mathbb{R}^{m\times k}$ is the cluster centroid matrix and $\boldsymbol{c}_j$ is the cluster centroid of the $j$th cluster, $\boldsymbol{B} \in \mathbb{R}^{n\times k}$ denotes clustering assignment, and $B_{ij} = 1$ if the $i$th data sample belongs to the $j$th cluster. Define a diagonal matrix

$$\boldsymbol{D}^{-1} = \mathrm{diag}\left\{\frac{1}{|\boldsymbol{N}_1|}, \frac{1}{|\boldsymbol{N}_2|}, \cdots, \frac{1}{|\boldsymbol{N}_k|}\right\} \in \mathbb{R}^{k\times k}, \tag{3}$$

where $|\boldsymbol{N}_j|$ is the number of data samples in cluster $j$. $\boldsymbol{C}$ is then written as $\boldsymbol{C} = \boldsymbol{V}\boldsymbol{B}\boldsymbol{D}^{-1}$. Hence,

$$J_k = \parallel \boldsymbol{V} - \boldsymbol{V}\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{B}^{\mathrm{T}} \parallel_{\mathrm{F}}^2, \tag{4}$$

and now the $k$-means' target is to seek the $\boldsymbol{B}$ that minimizes $J_k$, where each row of $\boldsymbol{B}$ has only one 1, with all remaining entries being zero. Given any two diagonal matrices $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ that fulfill the constraint $\boldsymbol{D}^{-1} = \boldsymbol{D}_1\boldsymbol{D}_2$, and representing $\boldsymbol{F} = \boldsymbol{B}\boldsymbol{D}_1$ and $\boldsymbol{H} = \boldsymbol{B}\boldsymbol{D}_2$, Eq. (4) can be rewritten as

$$\min_{\boldsymbol{F},\boldsymbol{H}} J_k = \parallel \boldsymbol{V} - \boldsymbol{V}\boldsymbol{F}\boldsymbol{H}^{\mathrm{T}} \parallel_{\mathrm{F}}^2, \tag{5}$$

where $\boldsymbol{F}$ and $\boldsymbol{H}$ have exactly one positive entry in each row, with the remaining entries being zeros. If we set $\boldsymbol{W} = \boldsymbol{V}\boldsymbol{F}$, this objective function is similar to NMF formulation as shown in Eq. (1). In $k$-means, the factor $\boldsymbol{W} = \boldsymbol{V}\boldsymbol{F}$ is the centroid matrix and the factor $\boldsymbol{H}$ has exactly one nonzero entry for each row. Thus, the rows of $\boldsymbol{H}$ represent hard clustering results of corresponding data samples. NMF relaxes these constraints, and the basis vectors of NMF need not be the centroids of the clusters, which could be more flexible than hard clustering. This indicates that each sample can be represented by only a few basis vectors through imposing the sparsity constraint on $\boldsymbol{H}$ in NMF. When a basis vector is close to a cluster center, samples belonging to that cluster can be easily identified by the largest entry in $\boldsymbol{H}$, which corresponds to the basis vector's contribution only. Cluster assignment of samples can be determined in this way.

### 2.3 Sparse NMF for band selection

There are two important points to be cleared in the band selection scheme. The first is that the hyperspectral data is a 3D image cube, which needs to be preprocessed for applying sparse NMF. The second point is when clusters by sparse NMF are obtained, which bands should be chosen to represent the clusters.

#### 2.3.1 Data reformulation

To fulfill the needs of a sparse NMF scheme, we must change the original 3D image and cube to a 2D data matrix. Each band of the image can be reshaped and viewed as a multivariate random variable vector with $M\times N$ length. Each band of the vectored $\boldsymbol{R}_l$ can be denoted as $\boldsymbol{V}(l,\cdot)$. Hence, the 3D hyperspectral cube $\boldsymbol{R} \in \mathbb{R}^{L\times M\times N}$ is reshaped to a 2D matrix $\boldsymbol{V} \in \mathbb{R}^{L\times P}$, where $P = M \times N$.

#### 2.3.2 Sparse NMF

The $L_1$ norm penalty has been widely recognized in recent years, and has been used successfully for achieving sparse solutions (Tibshirani, 1996). By imposing the $L_1$ norm constraint on rows of the $\boldsymbol{H}$ factor, we can achieve the sparse $\boldsymbol{H}$ factor to indicate the clustering membership. The formulation is given below:

$$\min_{\boldsymbol{W},\boldsymbol{H}} \frac{1}{2}\Big[ \parallel \boldsymbol{V} - \boldsymbol{W}\boldsymbol{H}^{\mathrm{T}} \parallel_{\mathrm{F}}^2 + \eta \parallel \boldsymbol{W} \parallel_{\mathrm{F}}^2 + \beta \sum_{j=1}^{n} \parallel \boldsymbol{H}(j,\cdot) \parallel_1^2 \Big] \quad \text{s.t.} \quad \boldsymbol{W},\boldsymbol{H} \geq \boldsymbol{0}, \tag{6}$$

where $\boldsymbol{H}(j,\cdot)$ is the $j$th row vector of $\boldsymbol{H}$. Parameter $\eta > 0$ controls the size of the entries of $\boldsymbol{W}$ to avoid a very large value, which may cause unstable results, and $\beta > 0$ controls the sparseness in rows of $\boldsymbol{H}$. A larger value of $\beta$ means more sparsity. Small values of $\beta$ and $\eta$ incline to better approximation results. The sparse NMF objective function in Eq. (6) can be solved by iterating the following nonnegativity constrained least square problems until the convergence condition is fulfilled (Kim and Park, 2007):

$$\min_{\boldsymbol{H}} \left\| \begin{pmatrix} \boldsymbol{W} \\ \sqrt{\beta}\boldsymbol{e}_{1\times k} \end{pmatrix} \boldsymbol{H}^{\mathrm{T}} - \begin{pmatrix} \boldsymbol{V} \\ \boldsymbol{0}_{1\times n} \end{pmatrix} \right\|_{\mathrm{F}}^2 \ \text{s.t.}\ \boldsymbol{H} \geq \boldsymbol{0}, \tag{7}$$

where $\boldsymbol{e}_{1\times k} \in \mathbb{R}^{1\times k}$ is a row vector having every entry as one, and $\boldsymbol{0}_{1\times n}$ is a zero vector, and

$$\min_{\boldsymbol{W}} \left\| \begin{pmatrix} \boldsymbol{H} \\ \sqrt{\eta}\boldsymbol{I}_{k} \end{pmatrix} \boldsymbol{W}^{\mathrm{T}} - \begin{pmatrix} \boldsymbol{V}^{\mathrm{T}} \\ \boldsymbol{0}_{k\times m} \end{pmatrix} \right\|_{\mathrm{F}}^2 \ \text{s.t.}\ \boldsymbol{W} \geq \boldsymbol{0}, \tag{8}$$

where $\boldsymbol{I}_k$ is an identity matrix of size $k \times k$ and $\boldsymbol{0}_{k\times m}$ is a zero matrix of size $k \times m$.

Through the sparsity, the matrix factorization procedure has a new function of interpretability for the generation process of data samples. It is obvious that sparsity on $\boldsymbol{H}^{\mathrm{T}}$ means that each sample is represented by a small number of basis vectors as to respective cluster centers. When a basis vector is close to a cluster center, data samples in that cluster can be identified easily. As a result, clustering assignment can be determined by the largest entry of each row in $\boldsymbol{H}$. We can use an example to demonstrate the clustering procedure. Here, we have a small data matrix $\boldsymbol{V} \in \mathbb{R}^{2\times 5}$. The five samples are generated from three different separated bivariate Gaussians. The first two columns of $\boldsymbol{V}$ belong to the same cluster, the third and fourth columns of $\boldsymbol{V}$ belong to another cluster, and the fifth column of $\boldsymbol{V}$ is generated from the other cluster.

$$\underbrace{\begin{pmatrix} 9.3891 & 8.0281 \\ 7.8440 & 7.6814 \\ 5.0772 & 2.4629 \\ 2.5532 & 4.2337 \\ 2.5420 & 2.1794 \end{pmatrix}^{\mathrm{T}}}_{\boldsymbol{V}} \approx \underbrace{\begin{pmatrix} 0.0842 & 0 & 0.1022 \\ 0 & 0.0485 & 0.0988 \end{pmatrix}}_{\boldsymbol{W}} \cdot \underbrace{\begin{pmatrix} 11.3023 & 0 & 29.1259 & 55.3683 & 0 \\ 0 & 0 & 0 & 0 & 32.7325 \\ 81.4755 & 76.8569 & 25.0516 & 0.9178 & 25.5586 \end{pmatrix}}_{\boldsymbol{H}^{\mathrm{T}}}. \tag{9}$$

It is one of the sparse NMF results with parameters $k = 3, \eta = 9, \beta = 0.0001$, and the correct cluster assignments of data samples are explicitly indicated by the largest entry of $\boldsymbol{H}$.

### 2.3.3 Band selection scheme

As described in Section 2.3.2, by imposing the sparsity on the $\boldsymbol{H}$ factor, the sparse coefficient factor could indicate the clustering membership. This band selection scheme is as illustrated in Algorithm 1.

**Algorithm 1** Band selection scheme with sparse nonnegative matrix factorization

```
Input: the hyperspectral image cube
Output: the selected bands indexes
1: Data reformulation: the 3D hyperspectral cube $\boldsymbol{R} \in \mathbb{R}^{L\times M\times N}$ is reshaped to a 2D matrix $\boldsymbol{V} \in \mathbb{R}^{L\times P}$, where $P = M \times N$.
2: Sparse nonnegative data matrix factorization: substitute $\boldsymbol{V}$ in the above step back to Eqs. (1) and (6) for $V$. Use Eqs. (7) and (8) to solve Eq. (6).
3: Band selection based on the sparse matrix factor $\boldsymbol{H}$: each band of data $\boldsymbol{V}(l,\cdot)$ can be viewed as a sparse linear combination of the basis vectors in $\boldsymbol{W}$, and matrix $\boldsymbol{H}$ is the sparse coefficient matrix, which determines the cluster assignments for each band. Find and save the band having the largest cluster indicator entry in each cluster.
```

After sparse NMF clustering on the reshaped data matrix, a specific spectral band (e.g., the $l$th band in the hyperspectral cube) belongs to the cluster by the largest entry of its linear representation coefficients corresponding to the $l$th row of the $\boldsymbol{H}$ factor matrix. For the band selection goal, we need to choose a suitable band from each cluster and the band can represent its cluster well for the subsequent task such as classification. There are some ways to solve this problem, such as randomly choosing band or rearranging the bands in a cluster by some distance metrics for choosing. In our algorithm, we use the bands that have the largest indicator entries in their own clusters as the selected band subsets. This strategy is reasonable, as the clustering results of sparse NMF have been considered to be a soft clustering scheme, and we just choose the most confident band to represent its cluster.

## 3 Experiments and results

### 3.1 Dataset description and experimental setup

A real hyperspectral dataset was used in our experiments. The dataset is a section of the subscene taken over Washington D.C. mall (1280 × 307 pixels, 210 bands, and 7 land-cover classes) by the hyperspectral digital imagery collection experiment (HYDICE) sensor (Neher and Srivastava, 2005). Fig. 2 shows the 80th band of the data. It has been widely accepted that, because of atmospheric water absorption, a total of 19 channels can be identified as noisy (1, 108–111, 143–153, 208–210) and safely removed as a preprocessing step.

The subsequent experimental analysis was organized for one main consideration. It aims at analyzing the effectiveness of selected bands for the classification task. Through changing the cluster number parameter $k$, different selected band subsets are evaluated. This part focuses on comparing the sparse NMF method with relevant techniques from the recent literature using different classifiers. The experiments have to be well designed for this consideration in view of an objective reflection of the problem. Therefore, in our experiments, we set $k$ from 1 to 20, with an increment of 1. For each $k$, sparse NMF was experimented 50 times with corresponding classification tasks for analyzing the average performance. There are two parameters to be determined in sparse NMF. In the experiment, $\eta$ was estimated by the largest entry of the input matrix $\boldsymbol{A}$. The parameter $\beta$ was used to control the degree of sparsity. The general behavior of sparse NMF was not very sensitive to $\beta$; however, too large $\beta$ values might lead to worse approximation. In the experiment, we used $\beta = 0.01$.

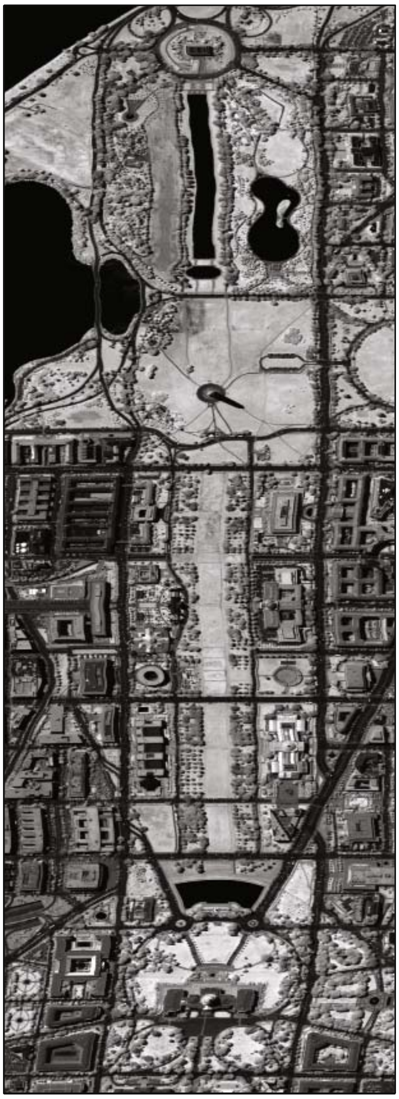

**Fig. 2 Washington D.C. mall HYDICE dataset, band 80**

### 3.2 Experiment result: classification accuracy of selected subsets of bands

To assess the performance of sparse NMF regarding recent relevant band selection techniques, a comparison study was carried out on sparse NMF and three other dimensionality reduction methods, i.e., maximum-variance principal component analysis (MVPCA), $k$-means, and affinity propagation (AP). MVPCA is a joint band-prioritization and band-decorrelation approach to band selection, which was introduced in Chang *et al.* (1999) for hyperspectral image classification and was also used in some comparative works for band selection. The $k$-

means algorithm is a well-known clustering method for cluster analysis; it aims to partition $n$ samples into $k$ clusters in which each sample belongs to the cluster with the nearest mean. In our band selection scheme, we choose the nearest band to each of the $k$ clusters' means to represent its cluster. AP is a new clustering algorithm which operates by simultaneously considering all data points as potential cluster centers (called 'exemplars') and exchanging message between data points until a good set of exemplars and clusters emerges. AP has been used for band selection and proved to be efficient in classification tasks (Qian *et al.*, 2009). A comparison of classification accuracies with different numbers of selected bands between the above four band selection methods was provided, as well as an assessment of these band subsets' effectiveness for representation in the classification test with two different types of classifiers, SVM and $K$-nearest neighbor (KNN). The two classifiers were used to compare the significance of the subsets of selected image bands that are obtained when using different classification schemes. For the evaluation of the selected bands' effectiveness in classification tasks, we used all of the original bands as a baseline for comparison. During classification, we randomly chose 6000 samples from each class. Eventually, a set of 3500 training samples (500 from each class for learning the classifier) and a set of 38 500 test samples (5500 from each class for assessing the accuracy) were obtained.

KNN is a method for classifying objects based on closest training examples in the feature space. It is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its $k$ nearest neighbors. The best choice of $k$ depends upon the data; in general, larger values of $k$ reduce the effect of noise on the classification, but they make boundaries between classes less distinct. A good $k$ can be selected by various heuristic techniques, for example, cross-validation. In our experiment, as a matter of experience, we chose $k = 5$. Recently, much attention has been put on SVM for the classification of hyperspectral data (Bazi and Melgani, 2006; Munoz-Mari *et al.*, 2007). SVM seeks a high dimensional hyperplane to maximize the margin between two different classes of samples. SVM usually provides high classification accuracies and very good generalization capabilities; it involves only a few control parameters for tuning and choosing. In our experiment, we used the LIBSVM library (Chang and Lin, 2001). The optimal parameters of SVM (radial basis function kernel) were obtained by 10-fold cross validation. The critical parameters ($C$, $\gamma$) were searched on a grid during cross validation. Pairs of ($C$, $\gamma$) were tried and the one with the best cross validation accuracy was picked. We tried exponentially growing sequences of $C$ and $\gamma$ to identify good parameters: in experiment, $C = 2^{-5}, 2^{-4}, \ldots, 2^{10}, \gamma = 2^{-10}, 2^{-9}, \ldots, 2^{5}$. Figs. 3 and 4 show the classification results.

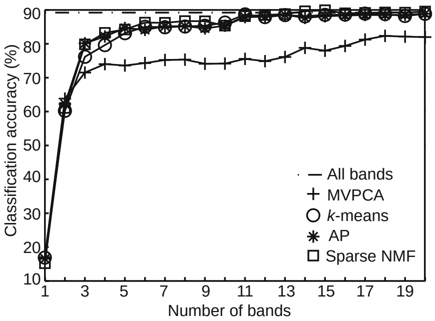

**Fig. 3 Classification performance of selected band subsets (KNN5) for the four band selection methods. The all original bands classification result is also included for comparison**

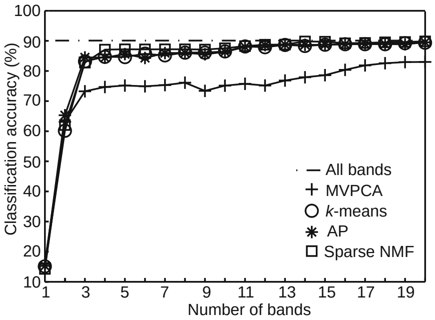

**Fig. 4 Classification performance of selected band subsets (SVM) for the four band selection methods. The all original bands classification result is also included for comparison**

As illustrated in Figs. 3 and 4, it seems that MVPCA gave weaker results compared with the other three methods. The $k$-means and AP seemed to perform similarly well often and gave very good results. Only when the cluster number was small, was $k$-means a little superior to AP in classification accuracy results. The sparse NMF performed very well in band selection, and obtained high classification accuracies which are very close to the accuracy baseline using all original bands. The classification results of the above experiments have proved that sparse NMF performs very well on band selection and data compression for hyperspectral data. Though in our experiments, the data dimension had been largely decreased, there were still very strong classification results, which means that the selected band subsets are very representative. It is interesting to investigate the obtained bases by sparse NMF when classification results are very good. The few nonzero and non-maximal entries in the coefficients can also be helpful to explain the relationship between bases and bands in the same cluster. The main drawback of sparse NMF is that the sparse NMF does not have a unique solution theoretically. Here, we experimented the sparse NMF 50 times for each data set, and used the mean of performance for evaluation. To our knowledge, most of the sparse NMF clustering procedures reached the same results because of the imposed sparsity constraints. However, it is still an issue to be considered for researchers.

## 4 Conclusions

In this paper, we present a band selection algorithm using sparse NMF for the band selection problem of hyperspectral imagery. Different from many of the clustering based band selection methods, sparse NMF does not need considerations on the distance metric between bands. By imposing sparsity on the coefficient matrix, it indicates the clustering membership through the largest entry in each column of the matrix. Experimental results show that the band subsets selected by sparse NMF have very good performance for real applications like land cover classification. In future work, nonnegative tensor factorization with sparse constraints will be exploited; it may be a superior alternative because of its avoiding ruining the spatial structure of the band image with data reshaping.

## References

Archibald, R., Fann, G., 2007. Feature selection and classification of hyperspectral images, with support vector machines. *IEEE Geosci. Remote Sens. Lett.*, **4**(4):674-677. [doi:10.1109/LGRS.2007.905116]

Bazi, Y., Melgani, F., 2006. Toward an optimal SVM classification system for hyperspectral remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, **44**(11):3374-3385. [doi:10.1109/TGRS.2006.880628]

Chang, C.C., Lin, C.J., 2001. LIBSVM: a Library for Support Vector Machines. Available from http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

Chang, C.I., Du, Q., Sun, T.L., Althouse, M.L.G., 1999. A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.*, **37**(6):2631-2641. [doi:10.1109/36.803411]

Cheng, Q., Varshney, P.K., Arora, M.K., 2006. Logistic regression for feature selection and soft classification of remote sensing data. *IEEE Geosci. Remote Sens. Lett.*, **3**(4):491-494. [doi:10.1109/LGRS.2006.877949]

Ding, C., He, X., Simon, H.D., 2005. On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering. Proc. SIAM Data Mining Conf., p.606-610.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. *Ann. Hum. Genet.*, **7**(2):179-188. [doi:10.1111/j.1469-1809.1936.tb02137.x]

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, **3**(7-8):1157-1182. [doi:10.1162/153244303322753616]

Hoyer, P.O., 2004. Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.*, **5**:1457-1469.

Hyvärinen, A., Oja, E., 2000. Independent component analysis: algorithms and applications. *Neur. Networks*, **13**(4-5):411-430. [doi:10.1016/S0893-6080(00)00026-5]

Jolliffe, I., 2002. Principal Component Analysis (2nd Ed.). Springer, p.516.

Keshava, N., 2004. Distance metrics and band selection in hyperspectral processing with applications to material identification and spectral libraries. *IEEE Trans. Geosci. Remote Sens.*, **42**(7):1552-1565. [doi:10.1109/TGRS.2004.830549]

Kim, H., Park, H., 2007. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, **23**(12):1495-1502. [doi:10.1093/bioinformatics/btm134]

Kim, J., Park, H., 2008. Sparse Nonnegative Matrix Factorization for Clustering. CSE Technical Report, GT-CSE-08-01, Georgia Institute of Technology.

Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**(6755):788-791. [doi:10.1038/44565]

Lee, D.D., Seung, H.S., 2001. Algorithms for non-negative matrix factorization. *Adv. Neur. Inform. Process. Syst.*, **13**:556-562.

Martinez-Uso, A., Pla, F., Sotoca, J.M., Garcia-Sevilla, P., 2007. Clustering-based hyperspectral band selection using information measures. *IEEE Trans. Geosci. Remote Sens.*, **45**(12):4158-4171. [doi:10.1109/TGRS.2007.904951]

Munoz-Mari, J., Bruzzone, L., Camps-Valls, G., 2007. A support vector domain description approach to supervised classification of remote sensing images. *IEEE*

*Trans. Geosci. Remote Sens.*, **45**(8):2683-2692. [doi:10.1109/TGRS.2007.897425]

Neher, R., Srivastava, A., 2005. A Bayesian MRF framework for labeling terrain using hyperspectral imaging. *IEEE Trans. Geosci. Remote Sens.*, **43**(6):1363-1374. [doi:10.1109/TGRS.2005.846865]

Qian, Y., Yao, F., Jia, S., 2009. Band selection for hyperspectral imagery using affinity propagation. *IET Comput. Vis.*, **3**(4):213-222. [doi:10.1049/iet-cvi.2009.0034]

Shahnaz, F., Berry, M.W., Pauca, V.P., Plemmons, R.J., 2006. Document clustering using nonnegative matrix factorization. *Inform. Process. Manag.*, **42**(2):373-386. [doi:10.1016/j.ipm.2004.11.005]

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, **58**(1):267-288.

Wang, S., Chang, C.I., 2007. Variable-number variable-band selection for feature characterization in hyperspectral signatures. *IEEE Trans. Geosci. Remote Sens.*, **45**(9):2979-2992. [doi:10.1109/TGRS.2007.901051]

Xu, W., Liu, X., Gong, Y., 2003. Document Clustering Based on Non-negative Matrix Factorization. Proc. 26th Annual Int. ACM SIGIR Conf. on Research and Development in Informaion Retrieval, p.267-273. [doi:10.1145/860435.860485]