New Developments and Challenges in Remote Sensing, Z. Bochenek (ed.)
©2007 Millpress, Rotterdam, ISBN 978-90-5966-053-3

Evaluation of feature extraction and reduction methods for hyperspectral images

J.S. Borges & A.R.S. Marçal
Faculdade de Ciências, Universidade do Porto, Portugal

J.M.B. Dias
Instituto de Telecomunicações, Instituto Superior Técnico, Lisboa, Portugal

Keywords: dimensionality reduction, feature selection, hyperspectral images

ABSTRACT: The number of hyperspectral sensors used in remote sensing is rapidly increasing. Although the availability of hyperspectral images is widespread, there is still a lack of efficient algorithms to properly handle the data. The major problem when trying to apply traditional image classification procedures to hyperspectral data is the high dimensionality, which besides increasing computational burden can impair classification due to the Hughes phenomenon. One possible solution to this problem is the dimensionality reduction, or feature reduction, prior to the application of other image processing procedures, such as image classification. It often happens that what was initially a hyperspectral image with several tens or even hundreds of bands is reduced to a 'standard' multi-spectral image with a small number of bands.

The process of feature reduction/selection is nevertheless a delicate task, as ideally one would like to preserve as much information as possible from the original dataset.

In this work some of the most commonly used feature selection and reduction algorithms are tested. The Jeffries-Matusita distance is used to determine the best subset of features and some spectral transformations such as Segmented Principal Component Analysis, Segmented Canonical Analysis, Independent Component Analysis, and Minimum Mean Squared Error are carried out to reduce the data dimensionality. These methods were compared using four test images from the AVIRIS, Hymap and Hysens sensors. Although no straightforward method exists to perform an evaluation, the consistency between the various methods is discussed.

1 INTRODUCTION

Hyperspectral images are characterized by the high number of contiguous spectral bands. The information provided by this type of data allows to characterize and/or discriminate more accurately and precisely different materials than the standard multispectral bands. Unfortunately, the increase of data dimensionality can become a problem due the incapacity of common image processing algorithms to deal with such high volume datasets. The impossibility, or high costs, of having sufficient training samples to keep an acceptable classification error is one of the most common problems – the Hughes phenomenon. Another frequent problem is the high computational cost of dealing with images with hundreds of bands, and often the impossibility of application of simple image processing methods. For these reasons, dimensionality reduction and feature extraction tasks are crucial methods of image pre-processing and analysis of hyperspectral data. These methods will allow the user to work with data more efficiently, preserving the essential information content while making evident features not discernable in the original data, all this with a reduced number of dimensions.

A group of feature reduction methods is based on feature selection. This consists in discarding the features that contribute less to the classes spectral separability using some measurement metric, such as the Jeffries-Matusita (JM) distance. Another group of feature reduction is based on transforming the pixel vectors into a new set of coordinates. One of the most commonly used methods of this type is Principal Components Analysis (PCA) (Duda & Hart 1973). This method projects data in a subspace where the data variance is minimized. Other spectral transformations have been used to reduce data dimensionality. Canonical Analysis (CA) is a technique similar to PCA, with the difference that includes information about the class structure optimizing the separation between classes (Hastie *et al.* 2001). These three methods reveal problems when dealing with high volume datasets such as hyperspectral images. A segmented principal component transformation was proposed for efficient hyperspectral image classification and display (Jia & Richards 1999). This technique is presented in Section 2.1 and it was applied to CA and feature selection with JM distance. Being a multivariate data analysis process, largely used for blind source separation, the Independent Component Analysis (ICA) can also be used as a tool for dimensionality reduction and representation of hyperspectral images (Lennon *et al.* 2001). While PCA and CA are purely second-order statistical methods (assuming the gaussianity of the factors), ICA makes a stronger assumption assuming that the factors are statistically independent and non-gaussian. All these methods give us spectral transformations of the original data, but do not give us information about the dimension of the optimal subspace. Recently, a new mean squared error based approach to determine the signal subspace in hyperspectral imagery was proposed (Dias & Nascimento 2005). This method and other feature reduction and selection methods were applied to a set of four hyperspectral images. The methods are briefly described in Section 2 and the results are presented in Section 3.

2 METHODS

Before starting introducing the methods, some notation definitions ought to be made. Capital bold letters denote matrices, small bold letters vectors, and small and capital letters scalars are used throughout. *N* is the number of pixels, *p* the number of spectral bands (columns) and the number of spectral classes is represented by *C*.

2.1 *Segmented principal component analysis*

Principal Component Analysis (PCA) is a well-known method that maps image data into a new, uncorrelated co-ordinated system or vector space (Richards & Jia 2006). The new feature space **z** is given by:

$$\mathbf{z} = \mathbf{E}^T\mathbf{x} \qquad (1)$$

where $\mathbf{E}^T$ is the matrix of normalized eigenvectors of the image covariance matrix ordered by the respective singular eigenvalues. The first k components give the maximum data variance.

Although PCA is a powerful technique, its application to hyperspectral data transformation is often inefficient to transform the complete dataset. A scheme that makes use of block structure of the correlation matrix can be applied so that PCA is conducted on data of smaller dimensionality (Jia & Richards 1999). The elements $\mathbf{R}_{ij}$ of the correlation matrix $\mathbf{R}$ are determined by:

$$\mathbf{R}_{ij} = \upsilon_{ij}/\sqrt{\upsilon_{ii}\upsilon_{jj}} \tag{2}$$

where υ_{ij} are elements of the covariance matrix and υ_{ii} and υ_{jj} are the variances of the *i*th and *j*th bands of data. The scheme proposed by Jia starts by partioning the complete data into *K* subgroups. Each subgroup is composed by highly correlated bands. The PCA is then conducted separately on each subgroup of data. The features selected can be regrouped and transformed again to compress the data further.

2.2 *Segmented canonical analysis*

Canonical Analysis (CA) is a procedure similar to PCA, in the sense that it generates a set of feature axes where the class separation is optimized. This new feature space is given by the linear transformation $\mathbf{V}^{\mathrm{T}}\mathbf{x}$. When the classes are known, CA is more powerful than PCA since it maximizes the ratio between the dispersion among classes and the dispersion within classes. The criterion to be maximized is:

$$J(\mathbf{V}) = \frac{|\mathbf{V}^T\mathbf{S}_A\mathbf{V}|}{|\mathbf{V}^T\mathbf{S}_W\mathbf{V}|} \tag{3}$$

where $\mathbf{S}_A$ is the dispersion matrix among classes and $\mathbf{S}_W$ the dispersion matrix within classes (Duda & Hart 1973). This technique projects the data into a new feature space of dimension *C-1*.

Like PCA, the CA approach does not handle hyperspectral data easily. So the scheme proposed by Jia for PCA can be also used in CA. The canonical coefficients are estimated in each of the *K* subgroups of highly correlated bands.

2.3 *Jeffries-Matusita distance for feature selection*

The Jeffries-Matusita (JM) distance is widely used to measure the class separability or distances between distributions. For multivariate Gaussian distributions the JM distance between class *i* and *j* is given by:

$$J_{ij} = \sqrt{2(1 - e^{-B_{ij}})} \tag{4}$$

where B_{ij} is the Battacharyya distance is given by:

$$B_{ij} = \frac{1}{8}\left(\mathbf{m}_i - \mathbf{m}_j\right)^T \left(\frac{\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j}{2}\right)(\mathbf{m}_i - \mathbf{m}_k) + \frac{1}{2}\ln\left[\frac{\left|\left(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j\right)/2\right|}{\left|\boldsymbol{\Sigma}_i\right|^{1/2}\left|\boldsymbol{\Sigma}_j\right|^{1/2}}\right] \tag{5}$$

in which $\mathbf{m}_i$ and $\mathbf{m}_j$ are the class mean vectors, and $\boldsymbol{\Sigma}_i$, $\boldsymbol{\Sigma}_j$ are the class covariances.

To proceed to a feature selection, one needs to find the subset of features that gives the largest average JM distance. The average pairwise distance is given by:

$$d_{ave} = \sum_{i=1}^{C}\sum_{j=1}^{C} p(\omega_i)p(\omega_k)J_{ij} \tag{6}$$

where $p(\omega_i), p(\omega_j)$ are the class prior probabilities.

In general, for C spectral classes, p total features, and a target of n features, a total of ${}^pC_n \cdot {}^CC_2$ measures of pairwise distances need to be calculated, besides the need of compute the average distance for each subset. This approach turns out to be unfeasible with the complete dataset. So the scheme of creating K subgroups of highly correlated bands can be used in this feature selection procedure.

2.4 *Independent component analysis*

Independent Component Analysis assumes the model (Hyvärinen *et al.* 2001)

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{7}$$

where $\mathbf{x}$ is the vector of observed signals, $\mathbf{A}$ is the matrix of mixing coefficients and $\mathbf{s}$ the vector of source signals. This model makes the following two assumptions: (a) the independent components are assumed statistically independent; (b) the independent components must have nongaussian distributions. In the case of dimensionality reduction of hyperspectral data, the physical sources do not exist and then the model (7) does not hold and assumption (a) is not relevant. ICA is only used to find the projection where all the projected components are "the most independent" in the sense of negentropy.

The estimation of independent components consists in finding the matrix $\mathbf{W} = \mathbf{A}^{-1}$ so that the sources $\hat{\mathbf{s}} = \mathbf{W}\mathbf{x}$ can be estimated from the vector $\mathbf{x}$ of the observed signals by optimizing a statistical independence criterion. The negentropy is estimated by (Hyvärinen *et al.* 2001):

$$J(\mathbf{x}) = \left\{E[G(\mathbf{x})] - E\left[G(\mathbf{x}_g)\right]\right\}^2 \tag{8}$$

where G is a non quadratic function and $\mathbf{x}_g$ is a gaussian random vector of the same covariance matrix as $\boldsymbol{x}$.

FastICA (Hyvärinen 1999) is a method that uses a fixed-point algorithm to estimate the independent components from given multidimensional signals by maximizing equation (8). The basic fixed-point iteration in FastICA is

$$\mathbf{w}_{new} = E\left\{\mathbf{x}g(\mathbf{w}^{\mathbf{T}}\mathbf{x}\right\} - E\left\{\mathbf{x}g'(\mathbf{w}^{\mathbf{T}}\mathbf{x}\right\}\mathbf{w} \tag{9}$$

where $g(u) = \tanh(u)$. The FastICA algorithm starts by choosing an initial (e.g random) vector **w**, then uses the iteration (9) followed by normalization. This process is repeated until it converges, i.e., the old and new values of **w** point in the same direction.

To avoid a direction to be estimated several times and to prevent privileging a vector amongst others, the matrix $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)^T$ is symmetrically decorrelated after every iteration.

2.5 *Minimum mean squared error*

The Minimum Mean Squared Error (Dias & Nascimento, 2005) is a technique to determine the signal subspace in hyperspectral imagery. The method first estimates the signal and noise correlations matrices, then it selects the subset of eigenvalues that best represents the signal subspace.

Let $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2 \ldots \mathbf{Y}_N]$ be a $C \times N$ matrix of spectral vectors. Assuming a linear mixing scenario, each observed pixel is given by

$$\mathbf{y} = \mathbf{x} + \mathbf{n} \tag{10}$$

where **n** models system additive noise. An estimate of the signal correlation matrix is given by

$$\hat{\mathbf{R}}_x = \hat{\mathbf{R}}_y - \hat{\mathbf{R}}_n \tag{11}$$

where $\hat{\mathbf{R}}_y = \mathbf{Y}\mathbf{Y}^T/N$ is the sample correlation matrix of **Y**, and $\hat{\mathbf{R}}_n$ is an estimate of noise correlation matrix which is inferred based on multiple regression theory.

The signal subspace can be identified by finding the subset of eigenvalues that best represents, in the least square sense, the mean value of the data set.
The sample mean value of **Y** is

$$\bar{\mathbf{y}} = \mathbf{c} + \mathbf{w} \tag{12}$$

where **c** is the signal subspace and **w** is a normal random variable with zero mean and covariance $\mathbf{R}_n/N$. Let $\mathbf{c}_k$ be the projection of **c** onto the subspace $\langle E_k \rangle$, spanned by the first *k* singular vectors ordered by the respective singular eigenvalues. The estimation of $\mathbf{c}_k$ can be obtained by projecting $\bar{\mathbf{y}}$ onto the signal subspace $\langle E_k \rangle$, i.e., $\hat{\mathbf{c}}_k = \mathbf{P}_k\hat{\mathbf{y}}$, where $\mathbf{P}_k = \mathbf{E}_k\mathbf{E}_k^T$ is the projection matrix onto $\langle E_k \rangle$.

The criteria for the signal subspace order determination is given by:

$$\hat{k} = \operatorname{argmin}_k \left(\bar{\mathbf{y}}^T \mathbf{P}_k^{\perp} \bar{\mathbf{y}} + 2tr(\mathbf{P}_k \mathbf{R}_n / N) \right) \tag{13}$$

3 RESULTS

This section presents the results of the application of methods described to four hyperspectral images acquired by different sensor systems.

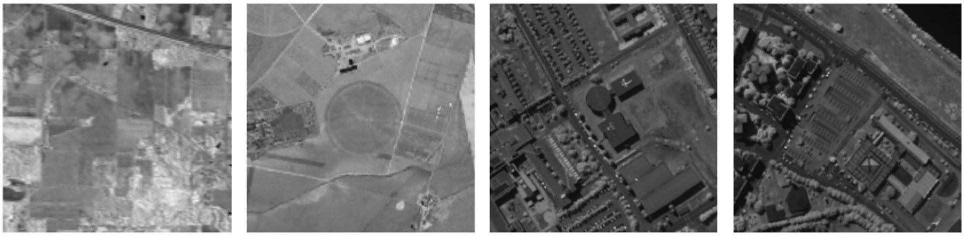

Figure 1. Test images (from left to right): A (band 50), B (band 43), C (band 80) and D (band 87).

3.1 *Test datasets*

The datasets used in this work were collected by three different sensors.

Image A was collected by the AVIRIS airborne sensor system. This image is of a small area (145 × 145 pixels) gathered over the Indian Pines test site in North-Western Indiana, containing 220 bands from 0.4µm to 2.4µm (Landgrebe 2003). A total of 16 classes were identified, but due to an insufficient number of training samples, only 9 classes were considered.

Image B is a section of 255 × 255 pixels from an Hymap image acquired over Barrax test site under the Daisex campaign in 1999 (Berger et al 2001). The Barrax test site is a well-described agricultural site close to the town of Albacete in Spain. The section used in this work has 128 spectral bands and 5 classes.

Images C and D are from Pavia University and Pavia centre, respectively. These images were acquired from ROSIS sensor in the HySens Pavia campaign in 2002 (Dell' Acqua *et al.* 2004). A section of 255 × 255 pixels was considered on both images. Image C has 103 bands and 9 cover classes, whereas image D has 102 bands and 8 cover classes.

3.2 *Band segmentation*

The first task performed on the test images was band segmentation. The segmentation was based on the results obtained by searching the edges in the correlation matrix (Equation 2). Figure 2 presents the correlations matrices of the 4 test images. As a result, the complete set of bands of images A and B was divided into 4 subsets and images C and D in 2 subsets. The exact band segmentation is present in Table 1.

This procedure of segmentation was used to perform PCA, CA and feature selection with JM distance. Each of these methods was applied in each of the subgroups generated by the correlation matrices of each image.

3.3 *Feature selection/reduction*

The feature reduction/selection methods presented in Section 2 were applied to the 4 test images, using MATLAB software (Mathworks 2002). Initially, a PCA procedure

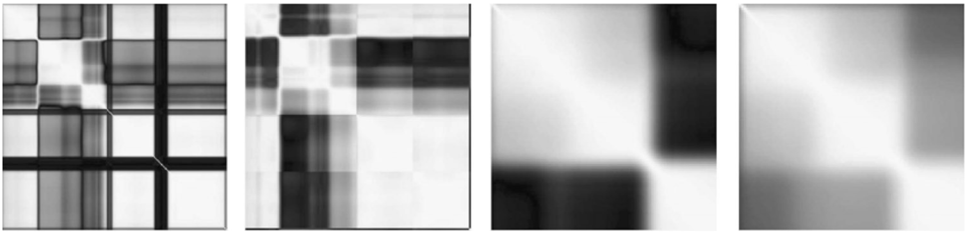

Figure 2. Correlation matrices of images A, B, C and D (from left to right). White corresponds maximum correlation (+ 1 or −1) and black to minimum (0).

Table 1. Segmentation of the complete set of bands.

Images	Subgroup 1	Subgroup 2	Subgroup 3	Subgroup 4
A	Bands 1:36	Bands 38:103	Bands 111:149	Bands 165:220
B	Bands 1:19	Bands 20:48	Bands 49:63	Bands 64:128
C	Bands 1:73	Bands 74:103	---	---
D	Bands 1:73	Bands 74:102	---	---

using the complete set of bands was performed in each test image. As expected, this task involved a high computational cost, but achieved good results in the sense that about 99% of data variance is explained in the first 4 or 3 principal components for all images.

The segmented PCA procedure using the segmentation presented on Table 1 proved to be much faster than the complete PCA. The time of computation was reduced in approximately 73%, 65%, 41% and 40% for images A, B, C and D, respectively. In order to compare results of complete PCA and segmented PCA procedures, the cumulative ordered eigenvalues of both approaches were plotted (Figure 3) as function of the number of components considered. It can be observed that, although the difference between both methods is considerable in the initial principal components, this difference tends to rapidly vanish. From the 6th component onwards, there is almost no difference between the two methods.

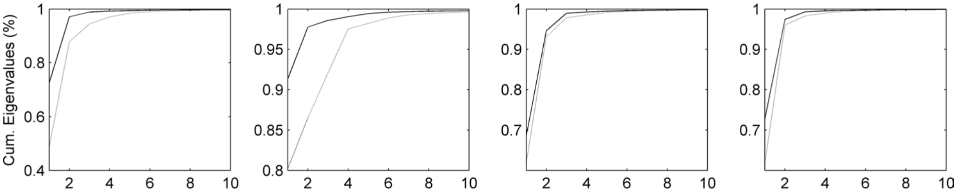

Figure 3. Cumulative eigenvalues of PCA over the complete band dataset (line) and the segmented PCA (doted line) as a function of the number of components, for the 4 test images (A, B, C and D, from left to right).

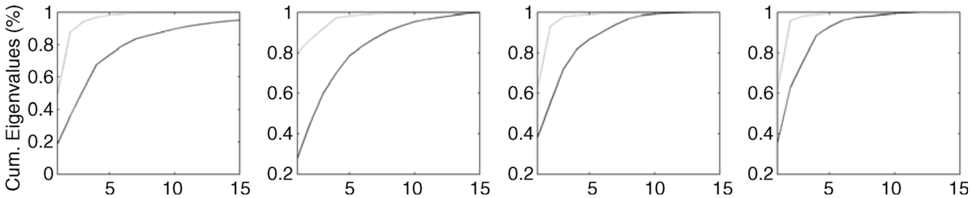

Figure 4. Cumulative eigenvalues of segmented CA (line) and the segmented PCA (doted line) as a function of the number of components, for the 4 test images (A, B, C and D, from left to right).

The application of CA over the complete test sets was not possible because the data dimensionality was too high. Instead, CA was applied to each of the subgroups of high correlated bands similarly to what was done with the segmented PCA approach. The CA reduces the number of features to the number of classes minus one. Using the first two canonical components, it is possible to observe the separation between classes. However it was observed, that a significant class overlapping exists. The difficulty of using CA in separating data is also observed in Figure 4, where the cumulative eigenvalues of CA transformation are compared with the ones from segmented PCA. The cumulative eigenvalues of CA are generally lower than those of PCA, except for a large number of components retained.

The JM distance was used to select the best 2 features in each subgroup of highly correlated bands. First, the pairwise JM distance between each pair of classes was determined for all combinations of two out of the number of bands in each subset; then the average pairwise JM distance was evaluated. The bands selected by this method are presented in Table 2.

FastICA algorithm has been applied to the complete test images, unlike PCA, revealing no difficulty in dealing with hyperspectral images. The only problem of ICA was the lack of convergence when estimating the final independent components. However, this problem is not “severe” since the main information is present in the first few independent components, and so, there is no need to estimate the last independent components.

The results of the application of Minimum Mean Squared Error for estimating the signal subspace was applied to the four images and the estimates were: 10, 24, 103 and 102 for A, B, C and D images respectively. There are respectively, 9, 5, 9 and 8 land cover classes identified in the test images. In image A the estimation (10 instead of 9) was good,

Table 2. Features selected by J-M distance.

Images	Subgroup 1	Subgroup 2	Subgroup 3	Subgroup 4
A	Bands 17 and 28	Bands 74 and 100	Bands 118 and 132	Bands 168 and 181
B	Bands 14 and 19	Bands 20 and 35	Bands 49 and 63	Bands 86 and 103
C	Bands 59 and 72	Bands 74 and 83	---	---
D	Bands 55 and 72	Bands 74 and 98	---	---

but for the remaining images, the predicted values were much more different than the actual number of classes expected. These differences are most likely due to the presence of rare pixels not accounted in truth data and spectral variability of recorded signals.

4 DISCUSSIONS AND CONCLUSION

The main objective of this work was to present and discuss some methods of pre-processing hyperspectral images. Methods of feature reduction, feature selection and estimation of signal subspace were presented and briefly evaluated on four hyperspectral datasets.

The main difficulty was to evaluate the performance of the methods presented. The absence of an independent criterion for evaluating the effectiveness of the methods is clearly a problem in the image processing field. It is possible to evaluate feature reduction methods based on classification accuracies. However, we must keep in mind that this type of evaluation is always dependent on the classification methods used. In this work we hoped to try to evaluate the performance of the feature selection/reduction methods independently, without being tied to a particular classification method.

Segmented PCA showed to be much faster than traditional PCA and revealed similar results except when only the first few principal components are considered. CA results were not so satisfactory as PCA. ICA, unlike PCA and CA, proved to be able to deal easily with hyperspectral datasets, and at the same time representing the essential information on a smaller space. The determination of signal subspace dimensionality is a difficult task mainly due the presence of rare pixels not accounted in truth data and spectral variability. However, this type of method should be taken in consideration when a feature reduction task is need. Overall, this work clearly indicates that there is great need to develop independent analysis tools for the evaluation of feature selection/reduction methods.

ACKNOWLEDGEMENTS

The first author would like to thank the ‘‘Fundação para a Ciência e a Tecnologia’’ (FCT) for the financial support (PhD grant SFRH/BD/17191/2004) and ‘‘Centro de Investigação em Ciências Geo-Espaciais, Faculdade de Ciências da Universidade do Porto’’ (CICGE-FCUP). The authors would also like to thank David Landgrebe for providing the AVIRIS data, Paolo Gamba for providing images from Pavia and José Nascimento for providing the HySime algorithm.

REFERENCES

Berger, M., *et al.* 2001, The DAISEX Campaigns in Support of a Future Land-Surface-Processes Mission, *ESA Bulletin 105*, 101–111.

Dell’ Acqua, F., Gamba, P., Ferrari, J., Palmason, A., Benediktsson, J. A., Arnason, K. 2004, Exploiting Spectral and Spatial Information in Hyperspectral Urban Data With High Resolution, IEEE *Geoscience And Remote Sensing Letters*, 1(4):322–326.

Dias, J., Nascimento, J. 2005, Signal subspace identification in hyperspectral linear mixtures. *Proc SPIE – Conf. on Image and Signal Processing for Remote Sensing*, Bruges, Belgium, Vol. SPIE-5982, 191–198.

Duda, R.O., Hart, P.E. 1973, Pattern Classification and Scene Analysis, John-Wiley & Sons, Inc.

Hastie, T., Tibshirani, R., Friedman, J. 2001, The Elements of Satistical Learning: Data Mining, Inference, and Prediction, Springer Verlag.

Hyvärinen, A., 1999, Fast and robust fixed-point algorithms for independent component analysis, *IEEE Trans. on Neural Networks*, 10(3):626–634.

Hyvärinen, A., Karhunen, J., Oja, E. 2001, Independent Component Analysis, John-Wiley & Sons, Inc.

Jia, X., Richards, J.A. 1999, Segmented Principal Components Transformation for Efficient Hyperspectral Remote-Sensing Image Display and Classification, *IEEE Trans. On Geoscience and Remote Sensing*, 37(1):538–542.

Landgrebe, D.A. 2003, Signal Theory Methods in Multispectral Remote Sensing, John-Wiley & Sons, Inc.

Lennon, M., Mercier, G., Mouchot, M.C., Hubert-Moy, L. 2001, Independent Component Analysis as a tool for the dimensionality reduction and representation of hyperspectral images, *Geoscience and Remote Sensing Symposium. IGARSS '01*. IEEE 2001 International, vol.6, 2893–2895.

MathWorks 2002, Using Matlab, Version 6.5. The MathWorks, Inc. Natick. MA.

Richards, J.A., Jia, X. 2006, Remote Sensing Digital Image Analysis, Springer-Verlag (4th edition).