Improved algorithm for hyperspectral data dimension determination

CHEN Jie[1, 2], DU Lei[1], LI Jing[2]， HAN Yachao[1], GAO Zihong[1]

[1] China Aero Geophysical Survey and Remote Sensing Centre for Land and Resources, Beijing 100083, China
[2] Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100101, China

Corresponding author Chen Jie’s e-mail address: 6592296@qq.com

Abstract. The correlation between adjacent bands of hyperspectral image data is relatively strong. However, signal coexists with noise and the HySime (hyperspectral signal identification by minimum error) algorithm which is based on the principle of least squares is designed to calculate the estimated noise value and the estimated signal correlation matrix value. The algorithm is effective with accurate noise value but ineffective with estimated noise value obtained from spectral dimension reduction and de-correlation process. This paper proposes an improved HySime algorithm based on noise whitening process. It carries out the noise whitening, instead of removing noise pixel by pixel, process on the original data first, obtains the noise covariance matrix estimated value accurately, and uses the HySime algorithm to calculate the signal correlation matrix value in order to improve the precision of results. With simulated as well as real data experiments in this paper, results show that: firstly, the improved HySime algorithm are more accurate and stable than the original HySime algorithm; secondly, the improved HySime algorithm results have better consistency under the different conditions compared with the classic noise subspace projection algorithm (NSP); finally, the improved HySime algorithm improves the adaptability of non-white image noise with noise whitening process.

1. Introduction

Adjacent bands of hyperspectral data acquired with less than 20nm spectral resolution generally have a strong correlation, and their DN value and visual images are often very similar. The hyperspectral image is recorded in the form of a matrix. Each column of the matrix represents the corresponding band spectral response data value. Through a series of matrix calculation, the front rows of the matrix band data can represent most information of the original image, and the back rows of the matrix represents the image noise. Low-dimensional data can be used to represent high-dimensional data with this method, reducing the amount of data processing workload and suppressing the impact of the noise. The quality of dimension reduction results will directly affect the next end-member extraction algorithms' final results. In the data dimension reduction process, how to select the appropriate dimension methods and how to accurately determine the required low-dimensional data are the most important two procedures.

After years of research, there are some relatively mature and widely used methods in hyperspectral data de-correlation transformation. De-correlation transformation is the removal of inter-band correlation of hyperspectral data transformation that converted data bands in different dimensions of space. The principal component analysis (PCA) algorithm and the minimum noise fraction (MNF) algorithm are the most two frequently used de-correlation transformation algorithms. The PCA algorithm is a linear algorithm based on K-L transformation. It sets the original image matrix variables into a set of uncorrelated random variables, arranges them in order of the covariance value to form a new image matrix. Typically, the first principal component matrix contains 80% variance information of all the bands. However, the PCA algorithm is sensitive to image noise. Principal components with large amount of information of all bands do not necessarily mean they have high signal to noise ratio (SNR). When an image noise variance of the principal component is greater than the variance of the signal, the image quality formed with the principal components is poorer. Therefore, Andrew A. Green et al (Andrew A. Green, 1988) proposed the MNF transformation. What is different with the PCA transformation is that the vectors are arranged by their SNR and the impact of noise on image quality is eliminated. Later on, James B. Lee, etc. proposed the noise-adjusted principal component transformation algorithm (NAPC).

However, the data dimension determination is still in the research stage. And for the PCA and MNF transformation algorithms, a key issue is to determine how many dimensions should be chosen for the low-dimensional data to represent high-dimensional data. In practical applications, researchers often select dimensions on their own experiences which make the dimensions determination process tend to be subjective or even blind. The correctness of dimension determination will directly affect the accuracy of the subsequent data processing precision and data analysis results. Therefore, choosing the appropriate rules and methods to determine the optimal number of dimensions plays a significant important role for the subsequent data processing and analysis procedures. Based on the above considerations, this paper analyzes and reviews the commonly used noise-whitening algorithms as well as the subspace dimension determination algorithm, proposes a hyperspectral data subspace dimension determination algorithm based on noise whitening processing, uses the algorithm to calculate the dimension of both the simulated and real hyperspectral data, and finally confirms its stability and accuracy by comparing with different algorithms.

2. Hyperspectral image Noise estimation and whitening

Under the influence of various factors, it is inevitable to bring in image noise while acquiring hyperspectral images. Currently, there are three main methods for remote sensing image noise estimation: laboratory method, dark current method and image method. It is hard for both the laboratory and the dark current method to apply in practical application because prior knowledge of the image is needed and the calculation requires a complex series of accurate measurement. The image method uses remote sensing images to directly analyze and estimate image noise, so this method is widely used to estimate the hyperspectral image noise (Zhang B, 2011). For hyperspectral images, the image noise estimation method mainly utilizes three features: the type and nature of hyperspectral images, pixel spatial correlation and hyperspectral band correlation.

2.1. Hyperspectral Image Noise Estimation

Since hyperspectral image between adjacent bands have a strong correlation, there are two commonly used algorithms to estimate hyperspectral image noise. One is the de-correlation algorithm with whole spectral dimensions proposed by Roger (Roger, 1996), and the other one is the spectral and spatial de-correlation method (SSDC) proposed by Arnold (Arnold, 1996). The advantage of SSDS algorithm is that it makes use of hyperspectral image spatial and spectral correlation between adjacent bands. And it is little affected by ground covering vegetation types, has a high degree of automotive calculation and does not need any human intervention. So, the SSDC algorithm is one of the stable algorithms to estimate hyperspectral image noise.

2.2. Hyperspectral Image Noise Whitening

Noise whitening is a process which first does the data de-correlation processing, and then does the noise covariance unitization processing. Researches confirm that, for a random vector, the transformation matrix which can do the whitening process is not unique. Meanwhile, when adopting different whitening matrix, the whitening results are also different. Generally, when doing the hyperspectral image processing, this paper assumes that the hyperspectal image information and the noise is irrelevant. This means the covariance matrix of the noise is a diagonal matrix.

3. The HySime algorithm

Based on the Principle of Least Squares, José M. Bioucas-Dias （José M. Bioucas-Dias, 2008) proposes the hyperspectral signal identification by minimum error algorithm (HySime). The HySime algorithm is used to estimate the dimension of hyperspectral subspace data. It estimates the signal as well as the noise correlation matrix first, and then best represents the feature vector subset of signal subspace in the form of minimum mean square error.

Suppose the observation value y consists of the signal x as well as the noise n, shown in the form of vector formula that y=x+n. The signal correlation matrix estimation value is referred as $\hat{R}_x$. Suppose the mean value of noise is 0, and $\hat{R}_x$ is the covariance. Then the signal correlation matrix can be separated as:

$$\hat{R}_x = E\Sigma E^T \tag{1}$$

The space of L formed from $E = [e_1, \cdots, e_L]$ can be separated into two subspace $\langle E_k \rangle$ and $\langle E_k \rangle^{\perp}$ which are mutually orthogonal. And the corresponding feature vectors $E_k = [e_{i_1}, \cdots, e_{i_k}]$, $E_{\underset{k}{\perp}} = [e_{i_{k+1}}, \cdots, e_{i_L}]$and $\pi = \{i_1, i_2, \cdots, i_L\}$ are used to record the sequence of feature matrix.

Suppose the projection matrix of subspace $\langle E_k \rangle$ is referred as $U_k = E_k E_k^T$, and the projection of y on $\langle E_k \rangle$ is referred as $\hat{x}_k = U_k y$.Try to find the best feature vector sequence $\pi = \{i_1, i_2, \cdots, i_L\}$ and k to make $mse(k)$ minimized. The calculation formula is as follows:

$$(k, \pi) = \arg \min_{k, \pi} \left\{ tr\left(U_k^{\perp} R_y\right) + 2tr\left(U_k \hat{R}_n\right) \right\} \tag{2}$$

Since $U_k = E_k E_k^T$, $U_k^{\perp} = I - U_k$, formula 2 can be put as:

$$(k, \pi) = \arg \min_{k, \pi} \left\{ c + \sum_{j=1}^{k} \left(\underbrace{- p_{i_j} + 2\sigma_{i_j}^2}_{\delta_{i_j}} \right) \right\} \tag{3}$$

In the formula above, c is a constant, and p_{i_j} 、 $\sigma_{i_j}^2$ refers to the binomial observation signal and noise correlation matrix respectively.

$$\begin{aligned} p_{i_j} &= e_{i_j}^T R_y e_{i_j} \\ \sigma_{i_j}^2 &= e_{i_j}^T \hat{R}_n e_{i_j} \end{aligned} \tag{4}$$

To minimize the right of the equal sign of formula 3, it is necessary to find all the negative value of δ_i, and the corresponding (k, π) is the calculation result .

4. The Improved HySime Algorithm

In HySime algorithm, it is easy to known from formula 3 and 4 that the estimation of image noise as well as signal matrix are two important procedures. And the calculation of corresponding feature vectors affects the final results directly. Meanwhile, the noise spectrum estimation algorithm adopts the full image spectral de-correlation method to estimate the dimensions, calculating pixel by pixel. Then it uses the original observation data minus the noise estimation value to get near true signal value, and the signal related estimation matrix can be calculated this way.

The original HySime algorithm is applicable when image noise is accurately estimated. However, besides the huge calculation, it is very difficult to accurately estimate a whole image’s noise value. So it is not applicable to estimate the image noise and it cannot be an ideal algorithm to estimate the image noise in order to have a good precision with the original HySime algorithm.

Suppose the observation image is white noise. The noise covariance matrix can be referred as $\hat{R}_n$ $=\sigma^2 I$, and the feature value λ and its corresponding feature vector x of the signal related matrix fits into the formula 5:

$$\left(\hat{R}_x - \lambda I\right)x = \left(R_y - \hat{R}_n - \lambda I\right)x = \left(R_y - \left(\sigma^2 + \lambda\right)I\right)x = 0 \tag{5}$$

It can be known after solving formula 5 with the HySime algorithm that when the noise is white noise, no matter big or small, the calculated feature vectors keep the same. This means the subspace determined by the feature vectors is not affected by image noise. And when the observation image noise is white noise, the impact of the image noise can be eliminated following two steps: first do the whitening transformation to the image noise, and then estimate its image noise with HySime algorithm. This method can improve the adaptability of the algorithm to image noise.

For a certain original observation data y, suppose $\hat{R}_n$ is the calculated noise covariance matrix. Separate $\hat{R}_n$ with its feature value to get its feature value and feature vectors, referred as Λ_n and A. Use matrix $F = A\Lambda_n^{-1/2}$ to do the image noise whitening process and mark the calculated observation data as y_w. Put y_w as the original HySime algorithm input value. Then the noise is white noise and its covariance matrix is a unit matrix, that is $\hat{R}_{nw} = I$. Mark the observation related matrix as R_{yw} and the signal related matrix equals to the observation data subtracting the noise value, that is $\hat{R}_{xw} = R_{yw} - \hat{R}_{nw} = R_{yw} - I$. Then do the calculation with HySime algorithm. Mark the feature vector of $\hat{R}_{xw}$ as $E = \left[e_1, \cdots, e_L\right]$, and mark the sequence of $\pi = \{i_1, i_2, \cdots, i_L\}$. Then it is easy to know that $\sigma_{i_j}^2 = e_{i_j}^T \hat{R}_{nw} e_{i_j} = 1$. So formula 3 can be put as formula 6. The corresponding (k, π) value of all the negative value of $\left(2 - p_{i_j}\right)$ is the subspace dimensions needed to be calculated.

$$\left(k, \pi\right) = \arg\min_{k,\pi}\left\{c + \sum_{j=1}^{k}\left(2 - p_{i_j}\right)\right\} \tag{6}$$

5. Experiments and analysis

5.1. Simulated Data Experiment

The simulated data used in this paper is made by manually adding noise into the real hyperspectral data. Signal data can be calculated by the number of end-member spectra multiplied by their corresponding abundance. And end-member spectra are selected from USGS spectral libraries.

Resample the spectra for 224 bands following the AVIRIS response functions and center wavelengths. The abundance data is distributed randomly with the Dirichlet method. The experiment is carried out this way: Set the number of different end-members to 5, 10, 15 and 20 respectively. The abundance of end-members follows the Dirichlet distribution which forms the signal data. Add a certain amount of noise to these four signal data to make their signal to noise ratio to be 15dB, 25dB and 35dB. Added noise can be separated into two types which are the white noise and colored noise. Abundance threshold value is set to 1. Each of the above combination can produce 100 sets of simulated data. Use the original Hysime algorithm, the improved Hysime algorithm and the NSP algorithm to do the calculation test respectively. Signal with manually added noise is relatively simple, so noise spectrum estimation adopts the full image spectral de-correlation method to estimate the dimensions, calculating pixel by pixel. The NSP algorithms with false alarm probability value of 10e-3 and 10e-4 are referred as NSP_10e-3 and NSP_10e-4. Signal subspace estimation is carried out for each 100 simulated data sets, and the mean and standard deviation of calculation results are used to be the evaluation indicators.

Conclusions can be drawn from the results of comparison of different algorithms with different parameters:

1) When the noise is white noise, under different parameters, results from the improved HySime algorithm as well as the original HySime algorithm have some consistency. However, when the noise is colored noise, the accuracy and stability of improved algorithm on the end-member estimation is much better than the original HySime algorithm.
2) Under different parameters, end-member estimation results from the improved Hysime algorithm and the NSP algorithm are basically the same.
3) When white noise exists, once the SNR increases, signal subspace estimation accuracy of each algorithm rises. Because the higher the SNR is, the better signal to noise suppression capability it is, and the smaller impact noise on the algorithm it is. This also illustrates the importance of noise estimation and removal.

5.2. Real Data Experiment

The AVIRIS data was collected from the Cuprite mining district, Nevada, in June, 1997. The DN value represents the reflectance ratio. It contains 224 bands between 400 nm to 2500 nm, its radiometric resolution is 10 nm and its spatial resolution is 20 m. To start with, remove water vapor absorption or low SNR bands so as to 192 bands are usable. The Cuprite mining district has a detailed ground survey result and geological background information. Gregg Swayze et al. have used the AVIRIS data which acquired in 1990 to identify 18 kinds of mineral types. They have carried out an X-ray diffraction analysis with field samples to determine their mineral categories and verified the results of the analysis (Swayze, 1992). Thereafter, Gregg Swayze et al. used the AVIRIS data acquired in 1995, analyzed the distribution of 25 kinds of minerals and did the mineral mapping with 23 kinds of minerals (Swayze, 1997).

From previous studies it can be concluded that the dimension estimation value of Cuprite mining zone basically lies between 18 and 30 (Ciobanu C, 1999). This paper uses the same four algorithms with simulated data, including: HySime algorithm, improved HySime algorithm, false alarm probability value with NSP algorithm 10e-3 and 10e-4's. The results are as follows:

Table. 1. Hyperspectral dimension estimation in Cuprite mining area

Dimension Estimation / Noise Estimation	HySime algorithm	Improved HySime algorithm	NSP_10e-3	NSP_10e-4
With Spectral de-correlation algorithm	20	27	27	27
With SSDC algorithm	-	26	26	26

The dimension number estimation value is 26 and 27 in the table which has a good consistency with the estimation value of 18-30 by other researchers (Ciobanu C, 1999）. The MNF transformation is adopted to check the correctness of dimension estimation in Cuprite mining district hyperspectral images. Choose different band combinations to calculate the pure pixel index (PPI), set the projected number of iterations to 20,000, set the threshold value of the coefficient to 2 and count the pixel number when the PPI index is greater than 1 in results. The results are shown in Table 2:

Table. 2. PPI calculation results with different parameters

Bands Number	10	20	26	30	40	50
PPI	639	874	1003	1060	1176	1355

The greater PPI is, the greater possibility pure pixels exist and the greater possibility the spectral space can be separated. Data in table 2 shows that when the number of bands is 26 or 30, the number of pixels is close to each other. The main image information is mostly concentrated in the top 40 bands after the MNF transformation. With the increase of the number of bands, the noise it contains is also rising. In this case, select the first 26 or 30 bands to calculate the PPI, the results are close to each other. This shows that the top 26 and top 30 bands can represent the original hyperspectral images good enough and have a good consistency. On the contrary, when the band number is 10 or 20, they cannot represent the original hyperspectral image completely. To further verify the above analysis, select the intersection of the above six cases and then 475 pixels were obtained. Mark these 475 pixels as pure pixels and label the pixel set as Φ. These pixels are little influenced by image noise and represent the pixels which are most likely to be end-members. Analyze each pixel's independence, not including the 475 pixels of pixel set Φ for the above six cases. The greater the spectral angle is, the more independent each pixel is to pixel set Φ, and the more possibility they belong to pure pixels. For the above six cases, the minimum angle of all spectral angle maximum value is 2.756° and its band number is 10. Meanwhile, the maximum angle of all spectral angle maximum value is 4.118° and its band number is 50. And the spectral angles are basically the same of 4.114° which is significantly different with other spectral angles of different band numbers, when band number is 26 and 30. This suggests that the information subspace obtained by dimension estimation can represent spectral information very well. And data analysis is effective with the help of dimension estimation number.

6. Conclusions

1) The improved HySime dimension estimation algorithm does not need to be performed pixel by pixel to estimate the hyperspectral image noise. It improves the efficiency as a result of reducing the amount of computation. What's more, it improves the accuracy thanks to eliminating errors cause by the noise estimation inaccuracy.

2) The improved HySime algorithm does not depend on the assumption that the mean value of hyperspectral image noise is zero. This makes the original algorithm to be more applicable. For both white and colored noise data, it can estimate the dimension number which is close to the true value.
3) Noise estimation plays an important role in real hyperspectral data processing and analysis. No matter the noise is white or colored, to accurately estimate or eliminate the noise will significantly improve the accuracy of image interpretation.

7. Acknowledgement

This paper is sponsored by two scientific research projects from China Geology Survey: one is the technical research of hyperspectral geological surveying, and its project number is 1212031513012; the other one is the Tianshan-Beishan key mining belt surveying. Its project number is 1212010030001500008.

References

[1] Bioucas Dias, J.M., J.M.P. Nascimento. Hyperspectral subspace identification [J]. IEEE Transactions on Geoscience and Remote Sensing, 2008, 46(4): 2435-2445.

[2] Cawse, K., A. Robin, M. Sears. The effect of noise whitening on methods for determining the intrinsic dimension of a hyperspectral image in 3rd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, WHISPERS 2011, June 6, 2011 - June 9, 2011. 2011. Lisbon, Portugal: IEEE Computer Society.

[3] Chang, C. I. , Q. Du. Estimation of number of spectrally distinct signal sources in hyperspectral imagery [J]. IEEE Transactions on Geoscience and Remote Sensing, 2004, 42(3): 608-619.

[4] Eldar, Y.C., A.V. Oppenheim. MMSE whitening and subspace whitening [J]. IEEE Transactions on Information Theory, 2003, 49(7): 1846-1851.

[5] Green, A.A., M. Berman, P. Switzer, et al. A transformation for ordering multispectral data in terms of image quality with applications for noise removal [J]. IEEE Transactions on Geoscience and Remote Sensing, 1988, 26(1): 65-74.

[6] Lee, J.B., A.S. Woodyatt, M. Berman. Enhancement of high spectral resolution remote-sensing data by noise-adjusted principal components transform [J]. IEEE Transactions on Geoscience and Remote Sensing, 1990, 28(3): 295-304.

[7] Roger, R.E. Principal Components transform with simple, automatic noise adjustment [J]. International Journal of Remote Sensing, 1996, 17(14): 2719-2727.

[8] Roger, R.E., J.F. Arnold. Reliably estimating the noise in AVIRIS hyperspectral images [J]. International Journal of Remote Sensing, 1996, 17(10): 1951-1962.

[9] Swayze, G., R.N. Clark, F. Kruse, et al. Ground-Truthing AVIRIS mineral mapping at Cuprite, Nevada. Summaries of the third annual JPL airborne geosciences workshop. AVIRIS Workshop: JPL Publication 92-14, 1992: 47-49.