A NOISELESS CODE LENGTH METHOD (NCLM) TO ESTIMATE DIMENSIONALITY OF HYPERSPECTRAL DATA

Masoud Farzam, Soosan Beheshti

Ryerson University, Department of Electrical Engineering, Toronto, Canada

ABSTRACT

Hyperspectral image analysis has been subjected to many improvements made in past decade. Yet the accurate estimation of dimensionality is still a challenge. Since dimension estimation of the Hyperspectral data is the first step in analysis of an image, the accuracy of analysis results highly depends on the accuracy of the dimension estimation step. Mostly, existing methods isolate the process of dimension estimation and process of denoising which leads to an inaccurate estimation of constituent components in the signal. In this paper, the problem of estimating the dimensionality of Hyperspectral data using the concept of "noiseless code length" is addressed. In our proposed method, NCLM, a set of nested subsets including the Hyperspectral data is generated first and then an error comparison approach is utilized by estimating the noiseless data error rather than noisy data error used by the existing methods to find the optimum subset. It has been shown that the estimated noiseless error has a minimum that represents the accurate estimation of the dimensionality of Hyperspectral data. The comparison of NCLM to other methods shows a substantial improvement in estimation of dimensionality in Hyperspectral imagery.

Index Terms— Hyperspectral imaging, Dimension estimation, Subspace selection, Denoising

1. INTRODUCTION

Estimation of the number of signal sources in a Hyperdata cube is very challenging. According to the definition, the effective dimensionality, is the minimum number of parameters required to account for the observed properties of the data. It is difficult to determine the effective dimensionality of Hyperspectral data in practice. This is mainly because effective dimensionality cannot be simply determined by the dimensionality of a data sample vector, referred to as component dimensionality(which is defined by the number of components in a data vector). For Hyperspectral data, the effective dimensionality is in general much smaller than the component dimensionality due to the high-dimensional structure of data cube. Several methods have been proposed such as principal components analysis (PCA) [5] and factor analysis [6] which make use of the eigenvalue distribution to determine the effective dimensionality. These approaches were basically developed for multispectral imagery with small limited number of bands where component dimensionality is comparable to the effective dimensionality. Also the application of sample correlation matrix is questionable knowing that hyperspectral data is not necessarily stationary with spacial variation. Some other SVD based dimension estimation methods are also in some cases inefficient since the noise present in most hyperspectral data sets is not $i.i.d.$ and, thus, the signal subspace is no longer given by the span of the first "p" singular vectors nor by any other set of eigenvalues. Harsanyi, Farrand, and Chang [4] developed a Neyman-Pearson detection theory-based thresholding method (HFC) to determine the number of spectral endmembers in hyperspectral data (referred to as virtual dimensionality in [3]). The HFC method uses the eigenvalues to measure signal energies in the detection model. All these methods are again using either the eigenvalue distribution concept or the sample correlation matrix calculation to reach the dimensionality. In both cases the base theory and assumptions is not quite valid for Hyperspectral data. In this paper we propose a method based on the concept of "noiseless code length" and reconstruction error. We will show that this method is able to overcome the inaccuracy resulted from the preliminary assumptions made in most existing methods. Also we will show that NCML method is able to simultaneously denoise the noisy Hyperspectral data.

1.1. Mathematical model

We consider a Linear mixture model due to its effectiveness and simplicity. Observation data in this model for each pixel is formulated as:

$$\bar{y}_i(n) = A\, s_i \tag{1}$$

$$y_i(n) = \bar{y}_i(n) + w_i \tag{2}$$

where the elements of $y_i(n)$ and $\bar{y}_i(n)$ both $\in R^N$ are the noisy measured solar radiation signal and the original noiseless solar signal at different spectral bands respectively.
$A = [a_1\ a_2\ ...\ a_c]$ is a $N \times c$ source matrix (or material signature matrix) with each column a_j being the spectral signature of Endmember j. The abundance vector $s \in R^c$ consists of the mixing coefficients satisfying two physical constraints

$s_j \geq 0$ (non-negative) and $\sum_{j=1}^{c} s_j = 1$ (sum-to-one), and c is the number of Endmembers. The last term w_i takes into account possible errors and sensor noises. We will considered an additive white Gaussian noise(AWGN). The noisy data $y(n)$ of length N is available from sensors output. The additive noise $w(n)$ is a sample of the zero mean random variable $W(n)$ with variance σ_w^2. The goal is to estimate the number of constituent independent spectral signals, c. Assume that the noiseless data belongs to the space S_N, $\bar{y}^N \in S_N$ (for example if the elements of $\bar{y}$ are real, one choice is $S_N = R^N$). The orthonormal basis vectors $s_1, s_2, \cdots, s_N$ span the space S_N. The noiseless data is represented by this basis as follows:

$$\bar{y}^N = \sum_{i=1}^{N} \theta^*(i) s_i \tag{3}$$

where $\bar{y}^N = [\bar{y}(1), \bar{y}(2), \cdots, \bar{y}(N)]^T$ and $\theta^*(i)$ is the ith coefficient of the noiseless data.

The least square error estimate of the ith basis coefficient using the observed data is :

$$\lambda(i) = < s_i, y^N > \tag{4}$$

$$= \theta^*(i) + < s_i, w^N > \tag{5}$$

where $y^N = [y(1), y(2), \cdots, y(N)]^T$ and $w^N = [w(1), w(2), \cdots, w(N)]^T$. The main challenge of signal denoising is how to decide which of the estimated coefficients, $\lambda(i)$s, should be ignored (set to zero) and which of them should be used to represent the noiseless data. Consider S_m, a subspace of S_N which is spanned by m elements of the basis. The estimate of noiseless data in this subspace is

$$\hat{y}^N_{S_m} = \sum_{i=1}^{N} \hat{\theta}_{S_m}(i) s_i \tag{6}$$

where for the estimated coefficient $\hat{\theta}_{S_m}(i)$ we have

$$\hat{\theta}_{S_m}(i) = \begin{cases} \lambda(i) & \text{if } s_i \in S_m \\ 0 & \text{otherwise} \end{cases}. \tag{7}$$

The denoising question and the dimensionality estimation is the process of finding a subspace S_m(and therefore which $\hat{y}^N_{S_m} \in S_m$) that best represents the noiseless data. Index m then would be the dimension of the Hyperspectral data. Two important elements in analyzing the denoising problem are the following errors

$$\textbf{Data error}: \quad x_{S_m} = \frac{1}{N} ||y^N - \hat{y}^N_{S_m}||_2^2, \tag{8}$$

$$\textbf{Reconstruction error}: \quad z_{S_m} = \frac{1}{N} ||\bar{y}^N - \hat{y}^N_{S_m}||_2^2. \tag{9}$$

The (noisy) data error, x_{S_m}, is the distance between the noisy observed data and its projection on subspace S_m. This error is available for each subspace. However, the noiseless data error, z_{S_m} (reconstruction error) is not available since it is a function of the unknown noiseless data.

2. NCLM ANALYSIS

The probability density function of the noisy data in Eq.2 is:

$$f(y^N; \bar{y}^N) = \frac{1}{\left(\sqrt{2\pi\sigma_w^2}\right)^N} e^{-\frac{||y^N - \bar{y}^N||_2^2}{2\sigma_w^2}} \tag{10}$$

where y^N is a sample of random variable Y^N. For g^N, any sample of random variable Y^N, the Shannon code is used. Therefore, the codelength of the binary prefix code is:

$$\mathrm{DL}(g^N; \bar{y}^N) = -\frac{1}{N} \log_2(f(g^N; \bar{y}^N)) \tag{11}$$

$$= \log_2 \sqrt{2\pi\sigma_w^2} + \frac{||g^N - \bar{y}^N||_2^2}{2\sigma_w^2 N} \log_2 e. \tag{12}$$

This denotes the description length of g^N when it is described by the noiseless data $\bar{y}^N$. In each subspace S_m the best representative of the noiseless data is $\hat{y}^N_{S_m}$ in Eq. 6. For the random variable generated by this representative of $\bar{y}^N$ Shannon code is used. Therefore, the codelength of g^N, using this representation is

$$\mathrm{DL}(g^N; \hat{y}^N_{S_m}) = \log_2 \sqrt{2\pi\sigma_w^2} + \frac{||g^N - \hat{y}^N_{S_m}||_2^2}{2\sigma_w^2 N} \log_2 e. \tag{13}$$

The codelength of the noisy data using the estimate $\hat{y}^N_{S_m}$ in different subspaces is:

$$\mathrm{DL}(y^N; \hat{y}^N_{S_m}) = \log_2 \sqrt{2\pi\sigma_w^2} + \frac{\log_2 e}{2\sigma_w^2} x_{S_m}. \tag{14}$$

For nested subspaces of different order, the data error, x_{S_m}, is a decreasing function of order m and is zero in S_N. Therefore, comparison and minimization of this codelength for a set of nested subspaces always leads to choosing the subspace with highest order, S_N. The comparison of this error fails since the noisy data is used to provide the estimate $\hat{y}^N_{S_m}$, and then the estimate is used to describe the same noisy data. However, it is reasonable to use the noisy data once to provide the estimate $\hat{y}^N_{S_m}$ and then use this estimate to describe the "noiseless data". The description length of the noiseless data in subspace S_m, using $\hat{y}^N_{S_m}$, is:

$$\mathrm{DL}(\bar{y}^N; \hat{y}^N_{S_m}) = \log_2 \sqrt{2\pi\sigma_w^2} + \frac{\log_2 e}{2\sigma_w^2} z_{S_m}. \tag{15}$$

The new minimum description length is obtained for S_{m^*} when the following holds :

$$S_{m^*} = \arg\min_{S_m} \mathrm{DL}(\bar{y}^N; \hat{y}^N_{S_m}). \tag{16}$$

In order to compare the new description lengths, the noise variance and the reconstruction errors z_{S_m} are needed. Our

goal is to minimize the reconstruction error. As proved in [1], Z_{S_m} and X_{S_m} have the following expected values:

$$E(X_{S_m}) = (1-\frac{m}{N})\sigma_w^2 + \frac{1}{N}||\Delta_{S_m}||_2^2 \quad (17)$$

$$E(Z_{S_m}) = \frac{m}{N}\sigma_w^2 + \frac{1}{N}||\Delta_{S_m}||_2^2 \quad (18)$$

where $||\Delta_{S_m}||_2$ is the l_2-norm of the vector of discarded coefficients in subspace S_m. Given the noisy data x_{S_m}, one sample of the random variable X_{S_m} is available. The variance of this random variable is of order $\frac{1}{N}$ of its expected value. Therefore, if the length of data is long enough, the variance of this random variable is close to zero. In this case, one method of estimating $||\Delta_{S_m}||_2^2$ is to assume that the available sample x_{S_m} is a good estimate of its expected value in Eq. 17:

$$\frac{1}{N}||\hat{\Delta}_{S_m}||_2^2 \approx x_{S_m} - (1-\frac{m}{N})\sigma_w^2. \quad (19)$$

When m increases, first and second terms in Eq. 18 grow in different directions which results in an optimum point:

$$c = m^* = \arg\min_m \{\frac{m}{N}\sigma_w^2 + \frac{1}{N}||\Delta_{S_m}||_2^2\} \quad (20)$$

c is the number of constituent Endmembers or in another word the key to the dimensionality of the Hyperspectral data. The data subset with index m^* is actually the denoised version of the noisy signal.

3. SIMULATION RESULT

The spectral reflectance used in the subsequent experiments are selected from the USGS digital spectral library which contains 224 spectral bands covering wavelengths ranging from 0.38 m to 2.5 m. A set of spectral profiles are selected as the Endmembers to create the mixture. To create linear mixtures, we randomly selected positive abundance vectors followed by multiplying them to spectral Endmembers and adding a Gaussian noise with SNR=10. The resulting image is then degraded by a spatial $k \times k$ average filter to produce mixed pixels(k controls the degree of mixing). With a small k, only the pixels close to the block boundary are mixed, so the mixture data are very likely to contain pure pixels. We consider 10 Endmembers from USGS library to create the signal. Figure 1 shows both created noiseless and noisy signals corresponding to the first spectral band for 3136 pixels in the image. To start the subset analysis, we created the first subset by assuming only two Endmembers in the subset and used an existing projection method [2] to estimate these two Endmembers. Further we calculated the data error and reconstruction error, z_{S_m}, using the extracted Endmembers. Second subset was created using these two Endmembers plus a

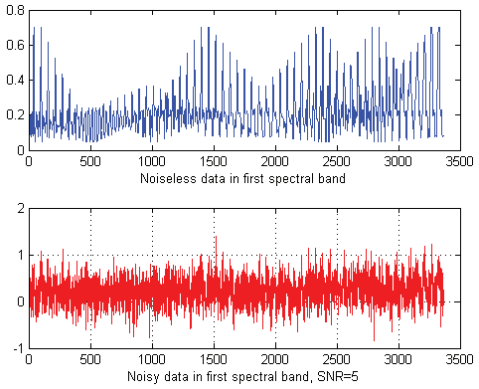

Fig. 1. Noiseless and noisy signal corresponding to the first spectral band

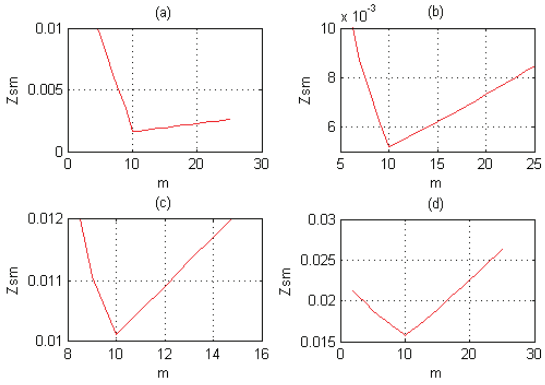

Fig. 2. Estimation of dimensionality a)SNR=15, b)SNR=10 , c)SNR=7 , d)SNR=5

new Endmember that will be estimated again using the projection method. In each step reconstruction error is calculated and plotted. Figure 2(a) shows how number of Endmembers, c, changes with the subset index m. It is clear that z_{S_m} has a minimum in $m = 10$ which is compatible with the original number of Endmembers we selected at the beginning. The descending part of the z_{S_m} graph is mainly due to the term $||\Delta_{S_m}||_2^2$ and the ascending part is due to the term $\frac{m}{N}\sigma_w^2$. As a result, NCLM for this case is able to predict the dimensionality of the Hyperspectral data precisely. To check the robustness of NCLM to the noise power, we have created the nosiy signal with different SNR values and in each step we have calculated the minimum point of reconstruction error. It can be seen in Figure 2 that for all SNR values of 15, 10, 7 and 5, NCLM is accurately estimating $m^* = c = 10$ which is the original dimensionality of the data.

Further to investigate the robustness of the system to the number of constituent components in the signal, four different

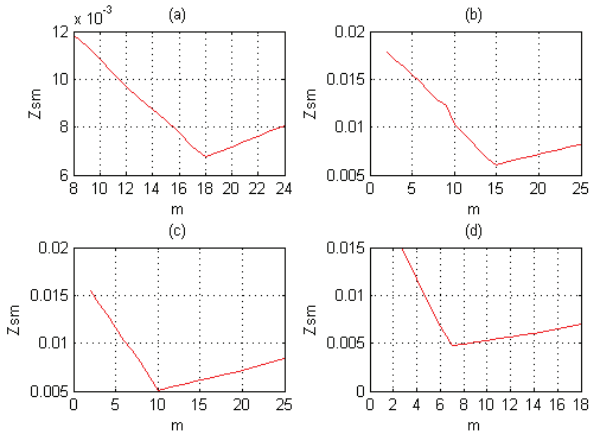

Fig. 3. Estimation of dimensionality a)c=20, b)c=15 , c)c=10 , d)c=5

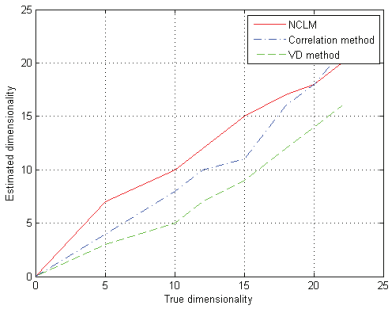

Fig. 4. Estimation of dimensionality a)c=20, b)c=15 , c)c=10 , d)c=5

set of Hyperspectral data were created with original number of Endmembers equal to $c = 20, 15, 10, 5$. In each case the minimum point of z_{S_m} was calculated and matched with the dimensionality of the corresponding data. Figure 3 shows that NCLM is accurately predicting the dimensionality for $c = 15, 10$ but is overestimating the dimensionality for the case of $c = 5$ and underestimating for the case of $c = 20$. This is mainly due to the fact that for these two specific cases the number of original constituent components from both end of the set of 25 Endmembers is too low(in fact either m or $N-m$ is too low). This causes the estimations used in Equation 19 not to be accurate any more. Further calculations showed that the minimum number of constituent components from each end of the set in the signal needs to be at least 7 for NCLM to be accurate.

To compare NCLM with other methods, we have selected the most common used methods, Sample Correlation method and Virtual Dimensionality(VD) method, as reference. First we generated mixed data including specified number of USGS library spectrum data as the constituent Endmembers. We then used each method to estimate the dimensionality of mixed data. Figure 4 shows the results. As it is shown, NCLM is the most accurate among the three methods used. Specifically, as it was discussed before in mid-ranges, NCLM is precisely estimating the dimensions while VD and correlation based method are both underestimating the dimensionality. Correlation method is only showing a better estimation for number of Endmembers $c < 5$.

4. CONCLUSION

A method was proposed to estimate the dimensionality of the Hyperspectral data based on the concept of noiseless code length in different subsets. It was explained that the comparison of the error in each subset based on the noisy signal would fail and therefore it is reasonable to use the concept of estimated "noiseless data error". In other word, the process of denoising and dimension estimation should be implemented in parallel in order to get an accurate result and that is what most of dimension estimation methods are lacking. It was shown that NCLM is highly robust to the noise level and precisely estimates the dimensionality assuming that the number of constituent Endmembers is not quite low. At the same time NCLM does the denoising since it finds the optimum subset that best represents the signal. As a future work, some improvements on the very low dimensional cases using different distributions rather than chi-square is proposed.

5. REFERENCES

[1] S.Beheshti and M.A.Dahleh "A new information-theoretic approach to signal denoising and best basis selection" *IEEE Trans. on Signal processing, vol. 53, NO. 10, October 2005*

[2] Jos M. P. Nascimento, Jos M. Bioucas Dias "Vertex component analysis: A fast algorithm to unmix Hyperspectral data" *IEEE Transactions on Geoscience and Remote Sensing, VOL. 43, NO. 4, April 2005*

[3] Chein-I Chang, Qian Du, "Estimation of number of spectrally distinct signal sources in Hyperspectral imagery" *IEEE Transactions on Geoscience and Remote Sensing, VOL. 42, NO. 3, March 2004*

[4] J. Harsanyi, W. Farrand, and C.-I Chang "Determining the number and identity of spectral Endmembers" *Proc. 9th Thematic Conf. Geologic Remote Sensing, Feb. 1993*

[5] J. A. Richards, "Remote sensing digital image analysis," *2nd ed. New York: Springer-Verlag, 1993*

[6] E. R. Malinowski, Theory of error in factor analysis, *Anal. Chem., vol. 49, pp. 606612, 1977*