引文格式：侯璨,李志飞. 基于HySime算法的端元个数估计[J]. 地理空间信息, 2024, 22(4): 92-95. doi:10. 3969/j.issn. 1672-4623.2024.04.021

基于HySime算法的端元个数估计

侯　璨[1]，李志飞[2]
（1. 上海飞未信息技术有限公司湖南分公司，湖南 长沙 410005；2. 湖南星图空间信息技术有限公司，湖南 长沙 410004）

摘　要：基于最小误差的高光谱信号识别（hyperspectral signal identification by minimum error，HySime）是一种估计高光谱端元个数的算法，该算法首先使用多元回归估计信号和噪声相关矩阵，然后使用信号相关矩阵的特征向量子集来表示信号子空间。为了科学评估HySime算法，分别对不同信噪比的高斯白噪声、高斯有色噪声模拟高光谱数据以及马蹄湾村真实高光谱数据的端元个数进行估计。实验表明HySime算法自适应性强，稳定性好，在解算过程中不需要输入任何参数，就能准确估计高光谱数据的端元个数。

关键词：HySime算法；高光谱影像；混合像元分解；端元个数估计
中图分类号：P237　　**文献标志码：**B　　**文章编号：**1672-4623（2024）04-0092-04

Endmember Number Estimation Based on Hyperspectral Signal Identification by Minimum Error

HOU Can[1], LI Zhifei[2]
(1. Hunan Branch of Shanghai Feiwei Information Technology Co., Ltd., Changsha 410005, China;
2. Hunan Satellite Image Information Technology Co., Ltd., Changsha 410004, China)

Abstract: Hyperspectral signal identification by minimum error (HySime) is an algorithm for estimating the number of hyperspectral endmembers. The algorithm uses multiple regression to estimate the signal and noise correlation matrix, and then uses a eigenvector subset of signal correlation matrix to represent the signal subspace. In order to scientifically evaluate HySime algorithm, we estimated the endmember number of Gaussian white noise, Gaussian colored noise simulated hyperspectral data with different signal-to-noise ratios and the real hyperspectral data of Matiwan Village. Experimental results show that HySime algorithm has strong adaptability and good stability, does not need to input any parameters during the calculation process, and can accurately estimate the endmember number of hyperspectral data.

Key words: HySime, hyperspectral image, decomposition of mixed pixel, endmember number estimation

高光谱遥感是用很窄而连续的光谱通道对地物持续遥感成像的技术。在可见光到短波红外波段其光谱分辨率高达纳米数量级，通常具有波段多的特点，光谱通道数多达数十甚至数百个以上，而且各光谱通道间往往是连续的[1]。目前高光谱遥感已经在自然资源、农业、林业等领域得到广泛应用，与高光谱有关的处理技术也迅猛发展。由于高光谱影像空间分辨率相对较低，在影像中普遍存在混合像元，混合像元的存在严重影响遥感分类和面积计算精度。为提高高光谱遥感应用的精度，首先需要解决混合像元分解的问题。混合像元分解的关键问题有2个：①如何准确地估计端元的数目；②如何更有效地寻找出端元[2-3]。目前端元数目估计算法根据原理可以大致分为2类：第①类是基于信息理论准则法，包括Akaike信息准则（AIC，akaike’s information criterion）、最小描述长度（MDL, minimum description length）、贝叶斯信息准则（BIC, bayesian information criterion）等。第②类是基于特征值阈值法，常用的有HFC算法（Harsanyi-Farrand-Chang）、噪声白化HFC法（NWHFC, Noise whitened HFC）[4]、基于最小误差的高光谱信号识别算法（hyperspectral signal identification by minimum error，HySime）[5]等。AIC、MDL、BIC算法都建立在似然函数对数的基础上，主要缺点之一是它们基于高斯噪声的假设。众所周知，高光谱遥感图像中的噪声一般不是高斯分布的。HFC、NWHFC算法来源于涅曼皮尔森的探测理论，通过计算得到的相关特征值和协方差特征值分别表示由样本相关矩阵和样本协方差矩阵产生的特征值，HFC和NWHFC算法由于受误警率设置的影响，不同的误警率会导致不同的端元个数估计值。如何设定合适的误警率通常是算法的关键，基于高光谱数据

收稿日期：2022-12-13。
第一作者简介：侯璨（1990—），硕士，高级工程师，主要从事地理信息应用、自然资源信息化建设等工作，E-mail：canning000@126.com。

的先验知识难以获取，因此我们很难确定合适的误警率。HySime算法的自适应性强，不需要输入任何参数，因此广泛应用于估计端元个数。

本文采用HySime算法对不同信噪比的高斯白噪声、高斯有色噪声模拟高光谱数据以及马蹄湾村真实高光谱数据的端元个数进行估计，科学评估算法的有效性、稳健性，为真实高光谱遥感影像端元个数估计提供参考依据。

1　HySime算法

HySime算法首先是Bioucas-Dias和Nascimento提出来的，其基本原理是使用多元线性回归估计信号和噪声的相关矩阵。算法采用信号相关矩阵的特征向量子集来代替信号子空间，通过最小化投影误差功率之和与噪声功率来推断子空间，即通过增加或减少子空间的维数来估计端元个数。因此，如果子空间维数估计过高，噪声占支配因素；反之如果子空间维数估计过低，投影误差占支配因素。该算法的优点是计算效率高、非监督及全自动化。HySime算法主要包括噪声估计（Noise Estimation）和信号子空间估计（Signal Subspace Estimation）2个部分。

1.1　噪声估计

在数据分析特别是高光谱遥感数据处理中，噪声估计是一个常见的典型问题。移动差值法（Shift Difference）是最简单的噪声估计法，也被称为邻近像素差值法（nearest neighbor difference[6]，NND）。该方法首先假设邻近像素之间的噪声独立并且具有相同的统计量，同时其光谱信息也假定基本相同。为了获取有效的噪声估计值，移动差分法必须采用均质区域而不是整个影像区域。该方法存在2种不足：①该方法假定邻近像素间有相同的信号信息，而在高光谱数据中该假设不一定满足；②为了改善噪声估计效果，必须采用监督方法选择均质区域。

HySime算法中噪声估计采用多重回归理论（multiple regression theory）。首先假设 $\boldsymbol{Y}$ 为 N 个 L 维光谱观测向量，定义 $\boldsymbol{Z}=\boldsymbol{Y}^{\mathrm{T}}, z_i=[\boldsymbol{Z}]_{:,i}$，其中 $[\boldsymbol{Z}]_{:,i}$ 为矩阵 $\boldsymbol{Z}$ 的第 i 列（z_i 为第 i 个波段的所有图像像素数据）。假定 z_i 为其余 $L-1$ 波段的线性组合，即

$$z_i=\boldsymbol{Z}_{\partial_i}\beta_i+\xi_i \tag{1}$$

式中，$\boldsymbol{Z}_{\partial_i}=[z_1,\cdots,z_{i-1},z_{i+1},\cdots,z_L]\in\boldsymbol{R}^{N\times(L-1)}$；$\beta_i\in\boldsymbol{R}^{L-1}$ 为回归向量；$\varepsilon_i\in R^N$ 为模型误差；利用最小二乘法求得 β_i 的估计值[7]：

$$\hat{\beta}_i=(\boldsymbol{Z}_{\partial_i}^{\mathrm{T}}\boldsymbol{Z}_{\partial_i})^{-1}\boldsymbol{Z}_{\partial_i}^{\mathrm{T}}z_i \tag{2}$$

噪声估计值为：

$$\hat{\xi}_i=z_i-\boldsymbol{Z}_{\partial_i}\hat{\beta}_i \tag{3}$$

噪声相差矩阵 $\hat{\boldsymbol{R}}_n=[\hat{\xi}_1,\cdots,\hat{\xi}_N]^{\mathrm{T}}[\hat{\xi}_1,\cdots,\hat{\xi}_N]/N$。为了减小计算量，经过推导发现 $(\boldsymbol{Z}_{\partial_i}^{\mathrm{T}}\boldsymbol{Z}_{\partial_i})^{-1}$ 可以通过去掉 $(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z})^{-1}$ 的第 i 行、第 i 列后经过变换得到。具体算法如下：

1）输入高光谱数据 $\boldsymbol{Y}=[y_1,y_2,\cdots,y_N]$；

2）令 $\boldsymbol{Z}=Y^{\mathrm{T}},\hat{\boldsymbol{R}}=\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$；

3）令 $\boldsymbol{R}'=\hat{\boldsymbol{R}}^{-1}$；

4）for i=1: L

$$\hat{\beta}_i=([\boldsymbol{R}']_{\partial_i,\partial_i}-[\boldsymbol{R}']_{\partial_i,i}[\boldsymbol{R}']_{i,\partial_i}/[\boldsymbol{R}']_{i,i})[\hat{\boldsymbol{R}}]_{\partial_i,i}$$

$$\hat{\xi}_i=z_i-\boldsymbol{Z}_{\partial_i}\hat{\beta}_i$$

5）输出 $\hat{\xi}$。

1.2　信号子空间估计

信号子空间估计是HySime算法的核心。首先根据前节提到的噪声估计法确定信号空间的正交方向子集，然后通过寻找原始信号 x 和由 $y=x+n$ 得到的 x 的噪声投影两者之间的最小均方根误差确定子集。假设噪声服从均值为零，协方差为 $\hat{\boldsymbol{R}}_n$ 的高斯分布，即 $n\sim N(0,\hat{\boldsymbol{R}}_n)$。假定信号样本相关矩阵为 $\hat{\boldsymbol{R}}_x=[\hat{x}_1,\cdots,\hat{x}_N][\hat{x}_1,\cdots,\hat{x}_N]^T/N$，特征分解为：

$$\hat{\boldsymbol{R}}_x=\boldsymbol{E}\Sigma\boldsymbol{E}^{\mathrm{T}} \tag{4}$$

式中，$\boldsymbol{E}=[e_1,\cdots,e_L]$ 为 $\hat{\boldsymbol{R}}_x$ 的特征向量按特征值大小排列的特征向量矩阵。将空间 R^L 分解为由 $\boldsymbol{E}_k=[e_{i_1},\cdots,e_{i_k}]$ 确定的 k 维子空间 $\langle\boldsymbol{E}_k\rangle$ 和由 $\boldsymbol{E}_k^{\perp}=[e_{i_{k+1}},\cdots,e_{i_L}]$ 确定的子空间 $\langle\boldsymbol{E}_k\rangle^{\perp}$。假定 $U_k=\boldsymbol{E}_k\boldsymbol{E}_k^{\mathrm{T}}$ 为投影矩阵，经过推导可知子空间参数 k 的计算公式为[7]：

$$k=\arg\min_k\left\{\mathrm{tr}(U_k^{\perp}\hat{\boldsymbol{R}}_y)+2\,\mathrm{tr}(U_k\hat{\boldsymbol{R}}_n)\right\} \tag{5}$$

具体算法如下：

1）输入高光谱图像数据 $\boldsymbol{Y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_N]$，并计算 $\hat{\boldsymbol{R}}_y=(\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}})/N$；

2）计算估计的噪声相关矩阵 $\hat{\boldsymbol{R}}_n=\frac{1}{N}\sum_i(\hat{\xi}_i\hat{\xi}_i^{\mathrm{T}})$；

3）计算估计的信号相关矩阵 $\hat{\boldsymbol{R}}_x=\frac{1}{N}\sum_i((y_i-\hat{\xi}_i)(y_i-\hat{\xi}_i)^{\mathrm{T}})$；

4）计算 $U_k=\boldsymbol{E}_k\boldsymbol{E}_k^{\mathrm{T}}$；

5）计算 $k=\arg\min_k\left\{\mathrm{tr}(U_k^{\perp}\hat{\boldsymbol{R}}_y)+2\,\mathrm{tr}(U_k\hat{\boldsymbol{R}}_n)\right\}$。

2 实验与结果分析

2.1 实验数据

为了更好地评价算法有效性，实验采用模拟数据和真实高光谱数据2种数据。模拟数据采用USGS光谱库[8]中明矾石、水铵长石、方解石、高岭石、云母5种光谱作为端元构建大小200×200共224个波段的模拟图像，光谱范围0.38~2.5 μm，光谱分辨率为10 nm，其光谱曲线如图1所示。模拟图像的背影为5种光谱的均值，并构建25个模块，每一列的模块有相同的尺寸，如图2所示[9]。

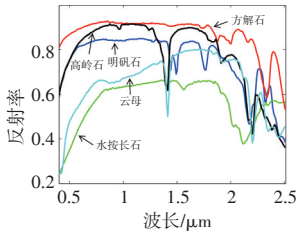


图1 5种光谱曲线

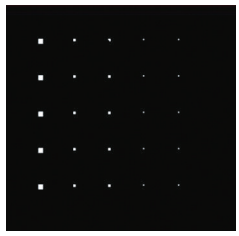


图2 25个模块

真实数据为雄安（马蹄湾村）高光谱数据集，该数据由中国科学院上海技术物理研究所研制的高分专项航空系统全谱段多模态成像光谱仪采集，光谱范围为400~1 000 nm，波段数为256个，影像大小为3 750×1 580，空间分辨率为0.5 m，数据的真彩色影像如图3所示。



图3 马蹄湾村高光谱数据真彩色影像

2.2 模拟数据实验结果及分析

考虑到算法的随机性，本次实验共分为5组，每组实验采用HySime算法分别对不同信噪比的高斯白噪声（η =0）和高斯有色噪声（η =1/18）高光谱影像进行端元个数估计，实验结果如图4所示。

采用HySime算法估计高斯白噪声（η =0）高光谱模拟影像的端元个数，5组实验估计端元个数的结果均相同，当信噪比SNR=2 dB时，估计端元个数为1；

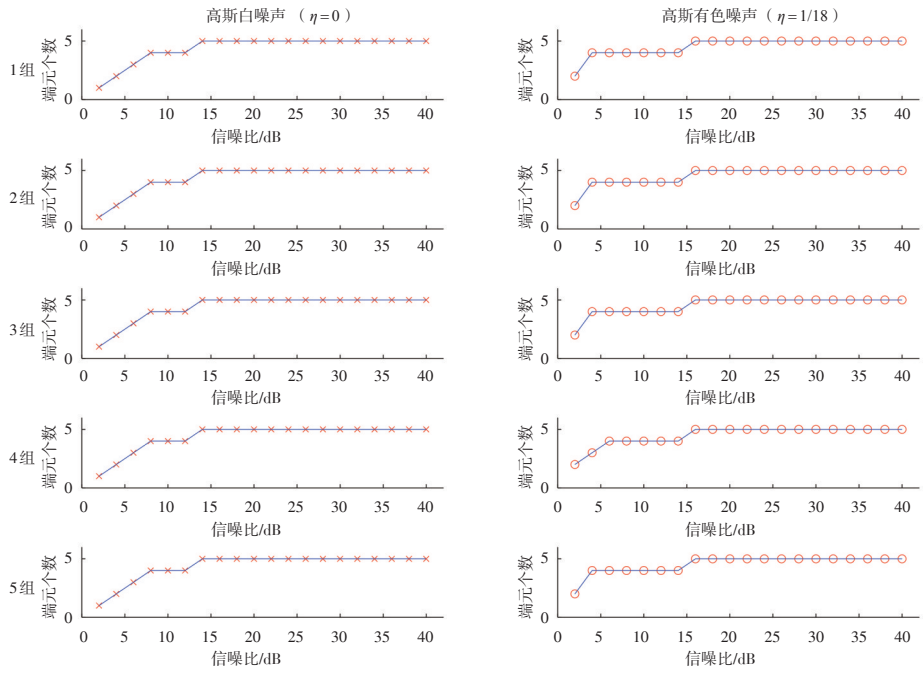


图4 模拟数据实验结果

当信噪比SNR=4 dB时，估计端元个数为2；当信噪比SNR=6 dB时，估计端元个数为3；当信噪比SNR=8、10、12 dB时，估计端元个数为4；当信噪比SNR≥14 dB时，估计端元个数为5。采用HySime算法估计高斯有色噪声（η =1/18）高光谱模拟影像的端元个数，除第4组实验信噪比SNR=4 dB时估计的端元个数为3外，其余估计的结果均相同，即当信噪比SNR=2 dB时，估计端元个数为2；当信噪比SNR=4、6、8、10、12、14 dB时，估计端元个数为4；当信噪比SNR≥16 dB时，估计端元个数为5。

综上所述，当信噪比SNR≥16 dB时，HySime算法能很好地估计法出端元个数，当信噪比低于16 dB时，HySime算法估计的端元个数略低于实际端元个数。通过实验表明在高信噪比高光谱遥感数据中该算

法可靠性高、稳定性好。

2.3　真实数据实验结果及分析

经过实地考察发现，该地区存在地物类别共计19类，包括水稻茬、草地、榆树、白蜡、国槐、菜地、杨树、大豆、刺槐、水稻、水体、柳树、复叶槭、栾树、桃树、玉米、梨树、荷叶、建筑。为了科学评价HySime算法，本次实验分别采用HFC、NWHFC、HySime算法对真实数据进行端元个数估计，并通过计算信号分量投影在信号子空间的相对功率来评价算法的精度，计算式为 $e=1-P_E/P_N$，其中 P_E 为信号分量投影在信号子空间的功率，P_N 为总功率。其端元估计的实验结果如表1所示，HySime算法的均方误差分布如图5所示。

表1　雄安（马蹄湾村）数据端元个数估计结果

算法	误警率（P_f）			相对功率（均值）
	10^{-3}	10^{-4}	10^{-5}	
HFC	16	15	14	4.884 5e−5
NWHFC	17	17	14	5.148 8e−5
HySime		19		1.673 0e−6

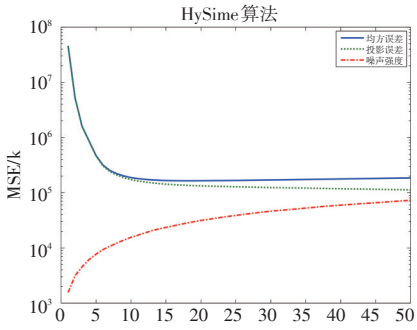


图5　端元个数估计均方误差分布图

在误警率为10^{-3}条件下，HFC和NWHFC算法估算的端元个数分别为16、17；在误警率为10^{-4}条件下，HFC和NWHFC算法估算的端元个数分别为15、17；在误警率为10^{-5}条件下，HFC和NWHFC算法估算的端元个数均为14；通过表1的实验结果分析可知，HFC和NWHFC算法由于受误警率设置的影响，不同的误警率会导致不同的端元个数估计值。其整体趋势为正相关，即误警率越小，端元个数的估计值越少。而使用HySime算法不受误警率设置影响，该实验所得的端元个数估计值为19。估算的端元个数与实地考察结果基本一致。通过相关学者的进一步研究发现，端元个数的差异会严重影响高光谱遥感影像地物分类精度，严重影响地物和光谱特征间的精确数量关系模型，在地物识别方面容易忽略微小地物等问题。

综上所述，由于HySime算法适应性强，稳定性好，不需要输入任何参数等特点，通常采用此算法估算真实高光谱遥感影像估算端元个数。

3　结　语

HySime算法首先使用多元回归估计信号和噪声相关矩阵，然后使用信号相关矩阵的特征向量子集来表示信号子空间。该子空间是通过最小化投影误差功率与噪声功率之和来推断的，它们分别是子空间维度的递减函数和递增函数。因此，如果子空间维度被高估，则噪声功率项占主导地位；如果子空间维度被低估，则投影误差功率项占主导地位。通过一系列模拟数据及真实高光谱数据进行的实验结果表明该算法自适应性强，稳定性好，不需要输入任何参数，是一种准确估计高光谱影像的端元个数的算法。通过相关学者对HySime算法进一步的研究，我们可以了解HySime算法对弱端元的探测能力较差，算法过程中容易忽略弱端元。但是，由于高光谱数据的空间分辨率较低，采用高光谱数据通常是基于宏观上的分析研究，而忽视弱端元的存在对精度影响较小，基本能够满足需求，因此在端元个数不清楚的地区通常使用此算法估计端元的个数。

参考文献

[1]　童庆禧，张兵，郑兰芬. 高光谱遥感[M]. 北京:高等教育出版社，2006

[2]　李志飞，杨敏华. 一种基于N-FINDR的优化端元提取算法[J]. 测绘地理信息，2017，42(5):70-73

[3]　郭诗韵. 一种基于空谱协同的高光谱端元提取方法研究[J]. 地理空间信息，2020，18(2):34-37

[4]　李二森. 高光谱遥感图像混合像元分解的理论与算法研究[D]. 郑州:信息工程大学，2011

[5]　Bioucas-dias J M, Nascimento J M P. Estimation of Signal Subspace on Hyperspectral Data[C]//Remote Sensing. International Society for Optics and Photonics, 2005: 59820L-8

[6]　Green A A, Berman M, Switzer P, et al. A Transformation for Ordering Multispectral Data in Terms of Image Quality with Implications for Noise Removal[J]. IEEE Transactions on Geoscience and Remote Sensing, 1988, 26(1):65-74

[7]　Bioucas-dias J M, Nascimento J M P. Hyperspectral Subspace Identification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2008, 46(8):2 435-2 445

[8]　Chang C I. Hyperspectral Data Processing:Algorithm Design and Analysis[M]. John Wiley & Sons, 2013