文章编号：1672-8785(2013)06-0039-06

多指标混合逐次投影高光谱图像降维算法

郑思远　李智勇　周石琳　王亮亮

(国防科技大学电子科学与工程学院，湖南 长沙 410073)

摘　要：降维对于高光谱图像解译具有重要意义。基于二阶统计量分析的经典主成分分析方法在降维过程中会丢失小目标信号。为解决这一问题，本文中引入高阶统计量作为投影指标对主成分分析方法进行拓展，提出了一种基于不同统计量描述的混合逐次投影的高光谱图像降维算法。该方法在保持主成分分析优点的同时，有效结合了非正交向量投影的特点，可以在降维后的低维空间中保留异常信号成分。真实高光谱图像数据的实验结果证明，该方法相对于主成分分析可以提取更加完整的低维信号子空间。

关键词：高光谱图像；降维；主成分分析；投影追踪；高阶统计量

中图分类号：TP751.1　　**文献标识码**：A　　**DOI**：10.3969/j.issn.1672-8785.2013.06.08

Hyperspectral Image Dimension Reduction Algorithm Based on Multi–index Successive Projection

ZHENG Si-yuan, LI Zhi-yong, ZHOU Shi-lin, WANG Liang-liang

(College of Electronic Science and Engineering, National University of Defense Technology, Changsha 410073, China)

Abstract: The dimension reduction is of significance to the interpretation of hyperspectral images. The classical Principal Component Analysis (PCA) method based on Second-order Statistics (SOS) may lose the signals of small targets during dimension reduction. To solve this problem, the PCA method is extended by using high-order statistics as projection indexes and a hyperspectral image dimension reduction method based on multi-index successive projection is proposed. The method incorporates the features of the non-orthogonal projection while retaining the advantages of the PCA method. It can keep the abnormal signal components in the dimension-reduced space. The experimental result of the real hyperspectral image data shows that this method can extract more complete signal subspace than the PCA method.

Key words: hyperspectral image; dimension reduction; principal component analysis; projection pursuit; high–order statistics

收稿日期：2013–05–03

基金项目：国家自然科学基金项目 (40901216)

作者简介：郑思远 (1989-)，男，新疆哈密人，硕士研究生，主要研究方向为模式识别和高光谱图像处理。

E-mail: nick.zsy@gmail.com

0 引言

高光谱遥感技术是当前遥感领域的前沿技术。与传统遥感图像不同，高光谱图像由地物几十个甚至上百个相邻光谱波段的图像构成，包含丰富的光谱信息。然而其高维数据空间内部大部分是空的，数据常集中在低维结构中 [1]。而且高光谱图像的高维数据特性也带来了数据量大、数据冗余度高、计算复杂等问题。因此，通过合适的降维方法将高维数据映射到低维空间对后续的高光谱图像智能解译具有重要意义。主成分分析 (principal components analysis,PCA)[2] 是一种基于数据二阶统计特性分析的线性变换降维算法，在高光谱图像处理中得到广泛应用。然而，在实际的高光谱图像处理中， PCA 方法对于一些出现概率较低的异常像元、小目标不敏感，在降维过程中会丢失这部分信息 [3−5]。近些年基于偏度 (skewness) 和峰度 (kurtosis) 等高阶统计量的研究工作在高光谱图像处理中越来越受到重视 [6−8]。信号高阶统计特性与高光谱数据特点紧密结合，反映了该图像的数据分布与高斯分布的偏离程度。其值越大，含有的异常信息就越多，越有利于小目标、小类别信息的保留。

本文提出一种结合方差与高阶统计量作为优化指标的混合逐次投影降维方法，以解决使用 PCA 降维时容易丢失小目标信号的问题。第二节从逐次投影的角度对 PCA 进行分析；第三节介绍采用高阶统计量进行投影的方法，并从逐次投影的角度对 PCA 方法进行扩展，提出多指标混合逐次投影 (Hybrid Indexes for Iterative Projection,HIIP) 算法，并在第四节中结合一种基于多元数据样本偏度的正态性检验方法来自动确定子空间维数。最后采用真实高光谱数据对本文算法进行验证，并与 PCA 方法进行比较分析。

1 PCA

PCA 主要是通过对协方差矩阵进行特征分解，得出数据的主成分 (即特征向量) 与它们的权值 (即特征值) 。设高光谱图像中 L 维像元光谱矢量观测样本集为 X=x_i,i=1,L,N， N 为像元样本数量。利用 PCA 方法对高光谱图像进行降维，首先对其进行中心化 (去除均值) 处理：

$$Y = X - \mu \cdot l^T \tag{1}$$

式中， $\mu = \frac{1}{N}\sum_{i=1}^{n} x_i$ 为样本均值， l 为分量均为 1 的 L 维列向量。由 PCA 方法进行降维的具体步骤如下：

(1) 计算的协方差矩阵 $\hat{\sum} = \frac{1}{N-1}YY^T$；

(2) 对 $\hat{\sum}$ 进行特征值分解，将其全部特征值按 $\lambda_1 \geq \lambda_2 \geq L \geq \lambda_L$ 的顺序排列，各特征值对应的特征向量为 e_i；

(3) 选出前 k 个最大特征值所对应的特征向量 {e_l,L,e_k} 作为图像主成分，构成低维子空间的基向量组 E_k，对观测数据进行正交变换，得

$$Z = E_k^T Y \tag{2}$$

则可求得变换后样本数据集的协方差阵为

$$\hat{\sum}_z = \frac{1}{N-1}\sum_{i=1}^{n} ZZ^T$$

$$= E_k^T \hat{\sum} E_k^T = \begin{pmatrix} \lambda_1 & L & 0 \\ M & O & M \\ 0 & L & \lambda_k \end{pmatrix} \tag{3}$$

从结果可以看出，变换后低维空间中的数据各分量是不相关的，而且各波段分量的方差是其对应的特征值 λ_i。通过去掉最小特征值所对应的成分，可达到降维的目的。从投影追踪 [9] 的观点来看，变换矩阵中 E_k 的特征向量都是根据方差最大化的欧氏空间投影方向确定的。主成分分析的过程实际上也可以被看成是以二阶统计量方差为投影指标的逐次投影寻踪过程，其第一主成分的估计为

$$w_1 = \arg\max_{||w||=1} E\{(w^T Y)^2\}$$

$$= \arg\max_{||w||=1} E\{w^T YY^T w\} \tag{4}$$

由瑞利商定理可知， w_1 即为 $E[YY^T]$ 的最大特征值所对应的特征向量。这与之前 PCA 的结果是一致的，并且其各主成分投影方向是相互正交的。假设已得到前 k-1 个主成分 W_{k-1} =

$[w_1, Lw_{k-1}]$ ，为了得到第 k 个主成分，必须先从 Y 中减去前面的 k-1 个主成分：

$$Y_{k-1} = Y - \sum_{i=1}^{k-1} w_i w_i^T Y = Y - U_{k-1} Y \qquad (5)$$

式中， U_{k-1}=$W_{k-1}W_{k-1}^T$ 为投影矩阵。把求得的第 k 个主成分带入数据集，得到新的数据集，继续寻找主成分。

$$w_k = \arg \max_{||w||=1} var\{w^T \hat{Y}_{k-1}\} \qquad (6)$$

2　混合逐次投影降维

2.1　基于高阶统计量的最佳投影方向

从逐次投影过程理解 PCA，其核心就是以方差为指标的相互正交投影矢量寻找过程。实际上利用数据更高阶的统计信息进行投影可得到具有不同性质的投影方向。

偏度与峰度由随机变量的三阶与四阶中心矩定义，在统计学中分别被用来衡量概率分布相对于正态分布的偏斜度以及平坦度。对于中心化数据 $Y = \{y_i\}_{i=1}^N$ ，令 $z = w^T Y = (w^T y_1, L, w^T y_N)^T = (z_1, L, z_N)^T$ ，其三阶、四阶统计量偏度与峰度可表示为

$$k_3(z) = E[(w^T Y)^3] = \frac{1}{N}\sum_{i=1}^{N} z_i^3 \qquad (7)$$

$$k_4(z) = E[(w^T Y)^4] = \frac{1}{N}\sum_{i=1}^{N} z_i^4$$

可类似定义更高阶统计量：

$$k_k(z) = E[(w^T Y)^k] = \frac{1}{N}\sum_{i=1}^{N} z_i^k, k \geq 3 \qquad (8)$$

相对于二阶统计量方差来讲，高阶统计量对异常数据更为敏感。在高光谱图像处理过程中，高阶统计量可以用来寻找大面积背景下的小物体。与使用方差 $E[w^T Y)^2]$ 作为投影指标的 PCA 不同，当将投影指标换作高阶统计量 $k_k(z)$ 时，该最优化问题就可表示为

$$w^* = \max_{||w||=1} \{\frac{1}{N}\sum_{i=1}^{N} z_i^k\}$$

$$= \max_{||w||=1} \{\frac{1}{N}\sum_{i=1}^{N} w^T y_i (y_i^T w)^{k-2} y_i^T w\} \qquad (9)$$

Ren 采用拉格朗日乘数法对上述问题进行了分析 [10]，并构造了目标函数：

$$J(w) = E[w^T Y (Y^T w)^{k-2} Y^T w] - \lambda(w^T w - 1) \qquad (10)$$

令 $\frac{\partial J(w)}{\partial w} = 0$，可得到最优解表示：

$$(E[Y(Y^T w)^{k-2} Y^T] - \lambda' I)w = 0 \qquad (11)$$

$E[Y(Y^T w)^{k-2} Y^T]$ 的最大特征值所对应的特征向量即为所求的最佳投影方向 w^* 。这里给出一种求解 w^* 的迭代算法：

(1) 初始化随机投影矢量 $w^{(0)}$， n=0；

(2) 计算 $E[Y(Y^T w)^{k-2} Y^T]$ 及其最大特征值所对应的特征向量 $v^{(n)}$；

(3) 令 $w^{(n+1)} = v^{(n)}$，计算 $w^{(n+1)}$ 与 $w^{(n)}$ 之间的夹角 θ 。如果 $\theta < \varepsilon$，即 $w^{(n+1)}$ 与 $w^{(n)}$ 趋于一致，那么 $w^* = v^{(n)}$；否则， $n \longleftarrow n+1$，转 b 继续执行。

2.2　HIIP

不同的投影指标，其投影方向有不同的统计含义。如果使用方差作为投影指标，则得到的投影方向上的数据分布信息量最大，像元能量分布最集中。如果以偏度作为衡量准则，投影方向上的数据直方图分布的偏斜性最大，这在一定程度上说明该方向上可能含有的异常信息较多。而如果选择峰度，则会导致投影方向上的数据分布具有长拖尾性质。

使用 PCA 对高光谱图像进行降维时，只需对其二阶统计量信息进行优化，可以较好地保存图像中的大部分背景信息，但是容易丢失一些出现概率较小的异常或者小目标信号。文献 [9] 提出了以某一种高阶统计量作为优化指标的迭代投影矢量生成算法。该方法可以较好地提取小目标信号，但是会丢失一部分区域面积中的较大目标的信息。在很多情况下，我们希望在对图像进行降维的同时能保存好原始数据中的背景以及异常信息。为了达到上述目的，一种可行的方案就是结合不同优化指标下的投影方向生成投影空间，以便充分利用图像中的信息。我

们提出一种不同优化指标混合的逐次投影算法 HIIP 生成信号子空间，其具体过程如下：

(1) 对高光谱数据样本集 X 进行去均值处理，得到中心化数据 Y 。

(2) 采用不同阶数的统计量 $\kappa_k(z)$ 作为投影指标，在 Y 上求解其最佳投影方向；本文选取 k=2 、 3 、 4 ，即分别为方差、偏度以及峰度，所得到的最佳投影方向分别为 w_1 、 w_2 、 w_3 。记为 $W_{(1)}=[w_1,w_2,w_3]$ 。

(3) 令 $P_{W_{(1)}}=W_{(1)}(W_{(1)}^TW_{(1)})^{-1}W_{(1)}^T$ ， $P_{W_{(1)}}^{\perp}=I-P_{W_{(1)}}$ ，可得到去除 $W_{(1)}$ 投影成分的新数据集 Y^1

$$Y^1=P_{W_{(1)}}^{\perp}Y \tag{12}$$

(4) 用 Y^1 替换 Y ，执行步骤 2 ，产生新的投影方向组 $W_{(2)}$ ，则 $Y^2=P_{W_{(2)}}^{\perp}Y$ ；假设已经得到前 k-1 组投影方向组 $W_{(1)}$ ， $LW_{(k-1)}$ ，则数据第 k 级成分 $Y^k=P_{W_{(k-1)}}^{\perp}Y^{k-1}$ ，之后在 Y^k 上执行 2 ，得到 $W_{(k)}$ 。

(5) 在达到可满足要求的低维子空间维数后停止迭代，记录子空间的基向量组为 $W=[W_{(1)},LW_{(k)}]$ ，所得低维信号空间即为 $\hat{S}_k=range\{W_{(1)},LW_{(k)}\}$ 降维数据，为

$$Z=W^TY \tag{13}$$

另外需要注意的是，在步骤 4 中生成 Y^k 还可以通过用原数据 Y 直接投影获得，即 $Y^k=P_{[W_{(1)},LW_{(k-1)}]}^{\perp}Y$ 。上述过程可以看作是对主成分分析的一种推广，实际上就是在 PCA 所寻找的最大方差投影的基础上增加高阶投影方向的信息。从几何角度的观点来看， PCA 生成的各级主成分投影方向是相互正交的。这种相互正交的投影方式可以在最大程度上保存数据原有的主要信息。但在高光谱图像中存在的一些异常信号，完全正交的投影并不是最佳投影方式。而基于高阶统计量的投影方向并不一定与主成分方向正交，这种非正交方向在一定程度上弥补了各级正交主成分的不足。

3 子空间维数确定

对于高光谱图像降维，需要考虑的另一个问题就是如何确定最终子空间的维数。在采用逐次投影的方法时，子空间维数的确定实际上就是迭代终止条件的选择。考虑到图像由信号 S 与噪声 N 构成，即

$$Y=S+N \tag{14}$$

如果认为提取的子空间完全包含了原始图像的全部信息，则可认为投影到其补空间的数据为不含任何信息的正态性噪声，即

$$Y^k=P_{\hat{S}_k}^{\perp}Y\approx P_{\hat{S}_k}^{\perp}N \tag{15}$$

因此，我们考虑以下的二元假设检验：

$$\begin{aligned}H_0&:\{Y^k \text{ is normal}\}\\ H_1&:\{Y^k \text{ is non–normal}\}\end{aligned} \tag{16}$$

当假设 H_0 成立时，即可认为降维子空间已满足要求，停止迭代过程。

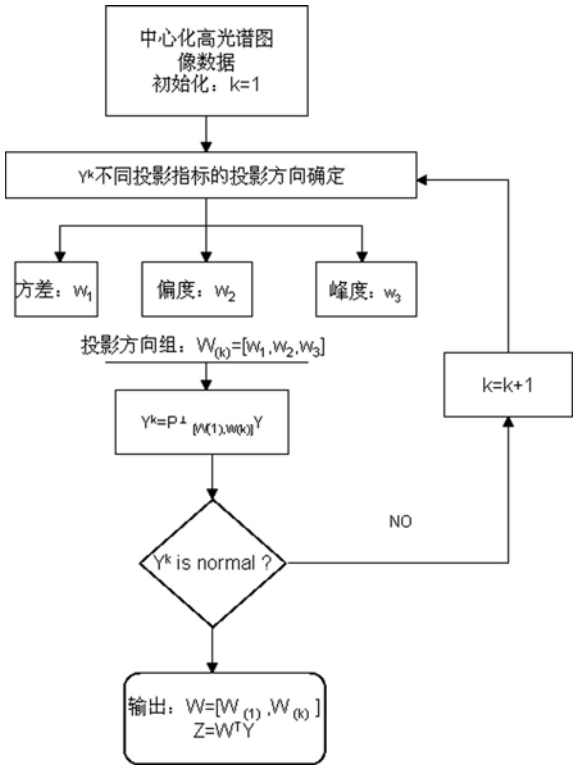

图 1 多指标逐次混合投影算法的流程图

对于 (16) 中的正态性检验问题， Móri 定义了一种用来检验多元数据正态性分布的多元随机变量的偏度形式 [11]：

$$\beta_{1,d}^{\%}(y) = E(||x||^2 x) \tag{17}$$

式中，$x = \sum^{-1/2}(z-\mu)$。对 Y^k 进行白化之后，$\beta_{1,d}^{\%}$ 的数据样本形式为

$$b_{1,d}^{\%} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}||y_i||^2 y_i^T y_j||y_j||^2 \tag{18}$$

当 N → ∞ 时， $b_{1,d}^{\%}$ 的极限分布趋近于正态分布 [12]：

$$Nb_{1,d}^{\%} \longrightarrow 2(d+2)\chi_d^2 \tag{19}$$

在给定显著性水平 α 下，当 $Nb_{1,d}^{\%}/2(d+2)$ 的值超过 χ_d^2 的 $1-\alpha$ 分位数 $\chi_{d,1-\alpha}^2$ 时，则拒绝式 (16) 中的原假设 H_0；当接受 H_0 时，整个迭代过程结束，即可得到逐次混合投影的子空间。结合前面投影过程的介绍，逐次混合投影算法的流程图如图 1。

4　混合逐次投影降维

图 2　PHI 数据的第 7 波段的场景图

我们采用由上海技物所研制的推扫式成像光谱仪生成的高光谱图像对算法进行验证。PHI 数据由机载平台在实际飞行中获得，剔除部分信噪比较低、图像模糊的波段。实际参与处理的波段有 123 个，图像大小为 220×90。图 2 为主要场景示意图，主要地物包括机场跑道、沙地以及 7 台测试用车辆。

本文采用 Matlab 作为实验平台实现 HIIP 算法，并将实验结果与 PCA 方法进行对比。图 3 为 PCA 以及 HIIP 的投影成分分量图。实验中，HIIP 算法正态性检验的显著性水平参数设置为 $\alpha = 10^{-4}$，用 HIIP 得到 3 个基向量组，共 9 个分量。图 3 (a)~(e) 分别是 HIIP 得到的第 1、3、5、7、9 成分的影像，(f)~(j) 为 PCA 得到的与之相对应的第 1、3、5、7、9 成分影像。从投影成分的结果中可以看出，PCA 中前几个主成分包含了图像的大部分背景信息，其后面几个成分中几乎没有有效信息。这说明在通过主成分变换后，大部分数据都集中在前几个主成分方向上，而其他与主成分正交的方向上的信息量趋于 0。但是从 (c)~(e) 的结果来看，以偏度与峰度作为投影指标时，在非正交投影方向上仍含有一些重要的小目标信息。

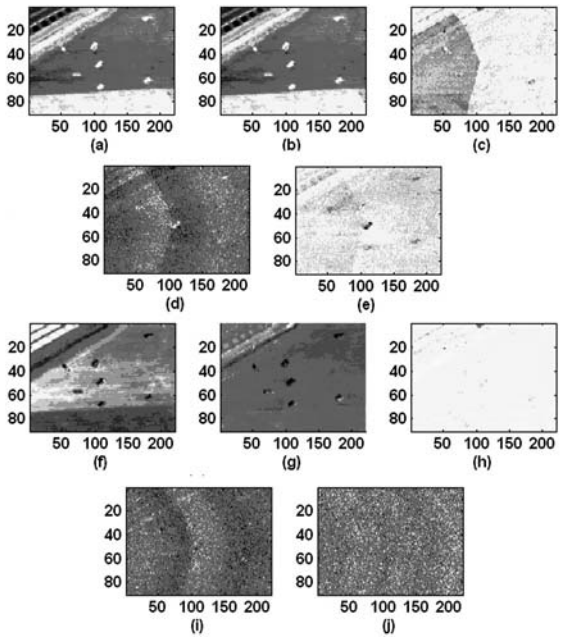

图 3　HIIP 与 PCA 投影成分的影像

为进一步比较 PCA 与 HIIP，我们将原始图像数据分别投影到 HIIP 得到的信号子空间以及由 PCA 前 9 个主成分构成的子空间的补空间中：

$$\bar{Z} = P_{\hat{S}}^{\perp} Y \tag{20}$$

投影到补空间的信号即可认为是降维过程中损失的信号成分。则各像素点损失能量可定义为

$$E_{loss} = ||\bar{z}_i||_2^2 = ||P_{\hat{S}}^{\perp} y_i||^2 \tag{21}$$

图 4 中分别展示了两种算法损失能量的分布图，其中图 4(a) 与 4(b) 分别为 PCA 与 HIIP 的能量

损失。从中可明显看出 PCA 方法中仍有关于车辆目标的信号成分损失。而与 PCA 相比，采用 HIIP 时，小目标的能量损失明显减少了。

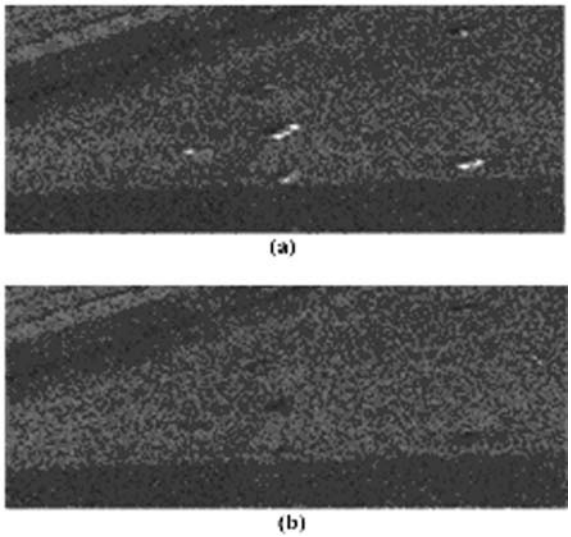

图 4 损失的能量分布示意图。(a) PCA， (b) HIIP

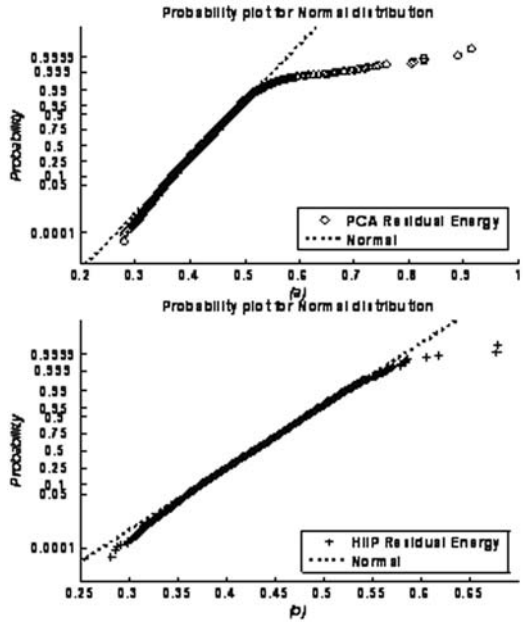

图 5 补空间投影信号能量正态的拟合示意图。(a)PCA， (b)HIIP

理想情况下，通过降维手段得到的子空间应包含原始图像中的全部信号成分，而在其补空间中为不含有任何信息的高斯噪声残余信号。假设补空间投影信号为零均值高斯分布，则其分量平方和形式服从分布。当样本数量足够多时，其均方根近似服从正态分布。因此，通过衡量补空间投影信号能量均方根的正态性，可以从侧面反映低维信号投影空间的准确程度。图 5 中 (a) 为 PCA 方法的残余信号能量正态拟合结果， (b) 为 HIIP 残余能量的拟合结果。从图中可以看出，由 PCA 得到的子空间在其补空间中仍有非正态信号成分，使得残余能量形式明显偏离正态分布。逐次混合投影算法通过引入高阶统计量信息对主成分分析进行有效补充，得到了更完整的信号子空间。

5 结论

在高光谱图像的降维方法中，基于二阶统计量分析的主成分分析方法可以有效保存图像中的背景信息，但会丢失出现概率较低的异常信号成分 。本文从逐次投影追踪的角度对 PCA 方法进行了分析，并在此基础上通过引入基于高阶统计量的投影方法对 PCA 进行了拓展。然后根据此思想设计了一种 HIIP 的高光谱数据降维方法。从实验结果来看，本文算法减少了 PCA 降维过程中的异常信号损失，与 PCA 相比，所得到的信号子空间有效增加了对于小目标信号的表示。今后还需要进一步研究如何将该方法与高光谱图像异常检测、小目标检测以及分类等应用相结合，以提高算法的性能。

参考文献

[1] Landgrebe D. Hyperspectral Image Data Analysis [J]. *IEEE Signal Processing Magazine*, 2002, **19**(1): 17–28.

[2] Bajorski P. Statistical Inference in PCA for Hyperspectral Images [J].*Selected Topics in Signal Processing*,2011,**5**(3): 438–445.

[3] Kuybeda O, Malah D, Barzohar M. Rank estimation and Redundancy Reduction of High-dimensional Noisy Signals with Preservation of rare Vectors [J].*IEEE Transactions on Signal Processing*, 2007, **55**(12): 5579–5592.

[4] Prasad S, Bruce L M. Limitations of Principal Components Analysis for Hyperspectral Target Recognition [J]. *IEEE Geoscience and Remote Sensing Letters*, 2008,**5**(4): 625–629.

（下转第 48 页）