独立成分分析在化学战剂混叠峰识别中的应用

陈媛媛[1,2],王 芳[2],王志斌[1,2,3],李文军[4]

(1. 电子测试技术重点实验室,山西 太原 030051; 2. 山西省光电信息与仪器工程技术研究中心,
山西 太原 030051; 3. 仪器科学与动态测试教育部重点实验室,山西 太原 030051;
4. 天津津航技术物理研究所, 天津 300308)

摘 要: 在战场等复杂环境得到的混合气体的红外光谱主次吸收峰交错重叠, 因此对其定性识别的特征提取方法就显得尤为重要。采集到的各种化学战剂和有机气体的红外光谱数据都是高维度数据,首先采用中心化后降维进行特征提取来尽可能多地捕获到它所包含的本质信息, 由于混合气体的红外光谱是非线性、非高斯性信号,把非高斯性作为独立性度量将各成分作为独立分量分离出来,为了满足实时需求,在传统快速独立成分分析(FastICA)算法的基础上对其迭代过程进行优化,并应用极限学习机(ELM)建立模型进行定量分析。实验结果表明:改进算法的迭代次数较传统算法减少,定量分析均方差 E=2.392 6×10^{-4},回归系数 R=0.999,说明该方法在不影响分离精度的前提下提高了混合物质中纯物质光谱分离出来的效率。

关键词: 混叠峰识别; 红外光谱; 非高斯性; 快速独立成分分析

中图分类号: TN219 **文献标志码:** A **DOI:** 10.3788/IRLA201645.0423001

Application of independent component analysis in aliasing peak identification of chemical warfare agents

Chen Yuanyuan[1,2], Wang Fang[2], Wang Zhibin[1,2,3], Li Wenjun[4]

(1. State Key Laboratory for Electronic Measurenent Technology, Taiyuan 030051, China;
2. Engineering Technology Research Center of Shanxi Province for Opto-Electronic Information and Instrument, Taiyuan 030051, China;
3. Key Laboratory of Instrumentation Science & Dynamic Measurement, Ministry of Education, Taiyuan 030051, China;
4. Tianjin Jinhang Institute of Technical Physics, Tianjin 300308, China)

Abstract: The infrared spectrum of mixed gas got in the battlefield and complex environment results in overlapping and stagger of the primary and secondary peaks, so its feature extraction of qualitative recognition is particularly important. The infrared spectral data collected from a variety of chemical warfare agents and organic gases are high-dimensional data. Centralizing before reducing dimension was used for feature extraction to capture the essence of more information it contained. Since the infrared spectrum of the mixed gas was non-linear and non-Gaussian signal, this method regarded non-Gaussian as independence measure to separate each component as independent component. In order to meet real-

收稿日期: 2015-08-05; **修订日期:** 2015-09-03

基金项目: 国家自然科学基金科学仪器基础研究专款 (61127015);国家国际科技合作专项 (2012DFA10680,2013DFR10150);
山西省青年科技研究基金(2013021028-1)

作者简介: 陈媛媛(1980-),女,副教授,硕士生导师,博士,主要从事光谱信号处理技术、智能算法方面的研究。
Email:chenyy@nuc.edu.cn

time requirements, its iterative process was optimized based on the traditional fast independent component analysis (FastICA) algorithm and extreme learning machine (ELM) model was applied to quantitative analysis. Experiment results show that the iterations of optimized method reduces compared with the traditional method and mean square error of quantitative analysis is $E=2.392\ 6\times10^{-4}$ and regression coefficient is $R=0.999$. And the optimized method improves the isolated efficiency of separating pure substances spectra from mixture substances without affecting the separate accuracy.

Key words: aliasing peak identification; infrared spectrum; non-Gaussian; FastICA

0 引 言

红外光谱分析作为一种光谱测量技术与化学计量学的有机结合的新兴分析技术,因其分析速度快、无损检测、效率高、成本低和易于实现在线分析等特点已应用于诸多领域。红外光谱吸收峰的频率、强度和形状是各物质所特有的, 因此红外光谱可以用来对某些单纯环境下的样品或者某些特殊的复杂环境和场合下的样品进行定性定量分析[1]。战时,当我军遭受化学战剂袭击时, 需要尽快获得敌袭击使用的化学战剂类型, 战场上复杂的环境使得我们得到的红外光谱的透过率光谱包括几种神经性毒气和一些有机气体(烟气成分)的混合气体,组成成分相似和分子结构类似的各种有机气体的红外光谱由于特征谱带的重叠或部分重叠而给谱峰的归属辨认带来极大的困难,面对现今愈来愈复杂的混合物体系,尤其是复杂的有机混合物体系, 化学计量学为试样没有验前信息的黑色体系提供了多种波谱的解析方法,常用的有主成分分析(PCA),实际信号的大部分重要信息往往包含在高阶统计特性中,而 PCA 方法利用协方差矩阵参与实际计算时只涉及输入数据的二阶统计特性,容易造成信息丢失[2]。

独立成分分析[3](ICA)是信号处理领域在 20 世纪 90 年代后期发展起来的一项基于信号高阶统计特性的分析方法,ICA 方法已经在特征提取、生物医学信号处理、语音信号处理、图像处理及人脸识别等方面得到了广泛的应用[4]。由于红外光谱的多峰性和重叠性, 使得许多波谱分辨方法无法直接应用于红外光谱,特别是化合物红外光谱的定性分析,ICA 可以从非高斯信号中找到一个使组分变成统计独立或者尽可能独立的非线性表达, 可广泛应用于特征提取和信号分离, 近年来部分研究人员已经将 ICA 结合光信号用于混合光谱中分离出独立组分的光谱, 基于红外光谱无损检测黄花梨可溶性固形物含量[5]和舰船气泡尾流所产生的后向散射光信号[6]。它已经成为一种从混合体系中分离出独立组分的强有力算法, 并逐渐显示了在分析化学领域的强大作用,ICA 是从混合物谱中分离出独立组分的红外光谱, 这种方法使得被分析信号各成分之间的统计依赖性最小,突出了源信号的本质结构,将 ICA 用于混合物的红外光谱进行解析[7],则提供了一种将吸收峰重叠的光谱分离出来的途径, 方便后续建立定量分析模型[8]。

ICA 的实现算法根据目标函数的不同有最大非高斯性法、极大似然估计法和最小互信息法等, FastICA 是以负熵作为衡量非高斯性指标的一种固定点迭代算法,它使用简单、收敛速度快稳定性好,是一种能对大量采样点进行批处理的快速算法,文中采用 FastICA 算法完成主次吸收峰混叠的红外光谱特征提取, 并对传统的 FastICA 算法进行迭代优化, 经过实验验证该优化算法在保证分离精度的前提下提升分离速度的能力。

1 ICA 算法

1.1 基于 ICA 的红外模型

ICA 的数学模型可以简单地概括为多导观察 $X=(x_1,x_2,\cdots,x_m)$是多个信源 S 经混合矩阵 A 组合而成:

$$X=AS=\sum_{i=1}^{n}a_is_i \tag{1}$$

式中:$S=(s_1,s_2,\cdots,s_n)^{\mathrm{T}}$ 为分量彼此统计独立的 n 维源信号;A 为未知的 $m\times n$ 混合矩阵,用来表示信号源到接收阵的传递函数, 由此知道 ICA 的两个主要方面是优化判据(目标函数)和寻优算法。现在的任务是:在 S 与 A 均为未知的条件下,求取一个解混矩阵

B,使得 X 通过它后所得输出 S 的最优逼迫 Y。

$$Y=BX=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}\approx(s_1,s_2,\cdots,s_n)^{\mathrm{T}} \tag{2}$$

根据朗伯比尔定律，通常认为在未知混合体系中测得的红外光谱是一些纯物质(主要成分)光谱的线性组合。根据上述 ICA 数学模型,对应于红外光谱数据矩阵可建模为 $X_{m\times n}$ 表示各成分的光谱信号与其贡献度乘积的和：

$$X=AS \tag{3}$$

式中：$X_{m\times n}$ 为 m 个样品在 n 个波长处的红外光谱数据矩阵；$S_{k\times n}$ 为独立成分矩阵,在理想的分解状态下相当于纯物质的光谱数据矩阵；$A_{m\times k}$ 为混合矩阵,它与纯物质在混合样品中的浓度矩阵 C 之间存在一定的函数关系,光谱矩阵 $X_{m\times n}$ 分解后,所得 $S_{k\times n}$ 的每一行相当于一种统计独立成分的光谱信息，在混合矩阵 $A_{m\times k}$ 中可以体现出独立成分在混合光谱中的相对浓度信息,$A_{m\times k}$ 的每一列可以被认为是某一独立成分(IC) 光谱在混合光谱中的权重大小,代表该 IC 对整个采样样品红外光谱的贡献。

1.2 基于负熵最大的 FastICA 迭代算法

根据信息论理论,在所有等方差的随机变量中,高斯随机变量具有最大的熵值，因而可以利用熵来度量分离结果的非高斯性,常用熵的修正形式负熵。在信噪分离过程中，可通过对分离结果的非高斯性度量来表示分离结果间的相互独立性，当非高斯性度量达到最大时，则表明已完成对各独立分量的分离[9]。在具有相同方差的随机变量中,高斯分布的随机变量具有最大的负熵,FastICA 算法通过最大化负熵得到 $W(W=B^{\mathrm{T}})$的目标函数可定义为：

$$N_G(W)=\{E[G(w^{\mathrm{T}}X)]-E[G(y_{\mathrm{gauss}})]^2\} \tag{4}$$

式中：$E[\cdot]$为均值运算；$G[\cdot]$为非线性函数,这时的问题就转化为求分离矩阵 W，使分离出的估计信号 $y=w^{\mathrm{T}}X$ (y 为其中一个独立成分,w 为分离矩阵 W 的其中一行,X 为混合信号矩阵)使函数 $N_G(W)$达到最大,且 $E\{(w^{\mathrm{T}}x)^2\}=1$,根据 Kuhn-Tucker 条件[10],$E\{XG(w^{\mathrm{T}}x)\}$的最优值能在满足下式的点上获得：

$$F(W)=\{E[XG(w^{\mathrm{T}}x)]\}+\gamma W=0 \tag{5}$$

式中：w 为权值向量；γ 为常数。用牛顿迭代法求解该方程式,可得 F 的雅克比矩阵：

$$JF(W)=E\{XX^{\mathrm{T}}g'(w^{\mathrm{T}}X)\}-\gamma I \tag{6}$$

由于数据被球化,$E\{x^{\mathrm{T}}x\}=1$,雅克比矩阵变成了对角阵,由此得到牛顿迭代公式为：

$$w^+=w-[E\{XG(w^{\mathrm{T}}X)\}-\gamma w]/[E\{G'(w^{\mathrm{T}}X)\}-\gamma] \tag{7}$$

归一化

$$w=w^+/\|w^+\| \tag{8}$$

规格简化后得到 FastICA 算法的迭代式为：

$$w(k+1)=E\{XG(w^{\mathrm{T}}(k)X)\}-\mathrm{E}\{G'(w^{\mathrm{T}}(k)X)\}w(k) \tag{9}$$

归一化

$$w(k+1)=w(k+1)/\|w(k+1)\| \tag{10}$$

FastICA 算法描述如下：

(1) 令 $k=0$,初始化权值向量 $w(0)$；

(2) $k=k+1$,由公式(9)更新权值 w；

(3) 归一化 $w(k+1)=w(k+1)/\|w(k+1)\|$；

(4) 如果$|w(k+1)-w(k)|>\varepsilon$,没有收敛，转到(2)；

(5) 算法收敛,求出一个独立成分,$y_1=s_1=wX$；

(6) 去掉已经抽取的独立成分,$w_{k+1}=w_{k+1}-\sum_{j=1}^{k}w_{k+1}^{\mathrm{T}}w_jw_j$,$w_{k+1}=w_{k+1}/\sqrt{w_{k+1}^{\mathrm{T}}w_{k+1}}$。

1.3 优化的 FastICA 迭代算法

在实际应用过程中，为了应对某些实时检测的要求，需要在不影响分离效果保证精确度的同时减少迭代次数，加快算法迭代速度，在此基础上对 FastICA 迭代过程进行优化。为了在牛顿迭代法求解过程中减少求雅可比矩阵的次数，一般将牛顿迭代法改为对所有迭代过程的雅克比矩阵都取为 $J(w_0)$,得到的迭代公式如下：

$$w_{k+1}=w_k-F(w_k)/J(w_0) \tag{11}$$

对迭代方程进行改进,假定已经求得 $w(k)$,则 $w(k+1)$可通过下面的过程获得：

$$\begin{cases} w_{k+1}^{(0)}=w_k \\ w_k^{(i)}=w_k^{(i-1)}-F(w_k^{(i-1)})/J(w_k^{(0)}),i=1,2,\cdots,m \\ w_{k+1}=w_k^{(0)} \end{cases} \tag{12}$$

该式收敛阶为 $m+1$,且每 m 次迭代只需要计算一次 $J(w)$,该优化方法在不增加雅可比矩阵次数的情况下，大幅度减少收敛时的迭代次数，从而减少计算量,提高算法速度。当 $m=2$ 时,公式(12)可简化为：

$$w_{k+1}=w_k-(F(w_k)+F(w_k-F(w_k)-F(w_k)/J(w_k))) \tag{13}$$

优化 FastICA 算法描述如下：

(1) 初始化权值矢量 w_0；

(2) 由公式(12)更新权值 w_{k+1}；

(3) 归一化 $w_{k+1}=w_{k+1}/\|w_{k+1}\|$；

(4) 如果$|w_{k+1}-w_k|>\varepsilon$,算法不收敛，则返回步骤(2)；

(5) 算法收敛,求出一个独立成分,$y_1=s_1=wX$;

(6) 去掉已经抽取的独立成分,$w_{k+1}=w_{k+1}-\sum_{j=1}^{k} w_{k+1}^{\mathrm{T}} w_j w_j$,$w_{k+1}=w_{k+1}/\sqrt{w_{k+1}^{\mathrm{T}} w_{k+1}}$。

2 实验分析

2.1 混合光谱的分离

实际情况中往往得到的是混合气体的透过率光谱,其中包括某几种神经性毒气和有机气体(烟气成分),用毒性较小的沙林模拟剂 DMMP 和反 2 丁烯(烟气的一种)作为实验的研究对象,实验采用扩散法配制 DMMP 蒸气,扩散管内加入 DMMP 液体置于广口瓶中,扩散管通道半径很细,可以近似认为其底部腔内的 DMMP 蒸汽处于饱和状态, 使用德国 Bruker 公司的 VERTEX70 型红外光谱仪,EGOLD-A 型长光程气体吸收池,光程长 20 m,容积 500 ml,实验温度为 296 K,气压为 101 325 Pa,不同浓度的混合气体充入密闭气室中,采集到样品的透过率光谱 400 条,其中光谱范围为大气窗口 8~14 μm(1 300~700 cm^{-1}),分辨率为 4 cm^{-1}。实验装置如图 1 所示。

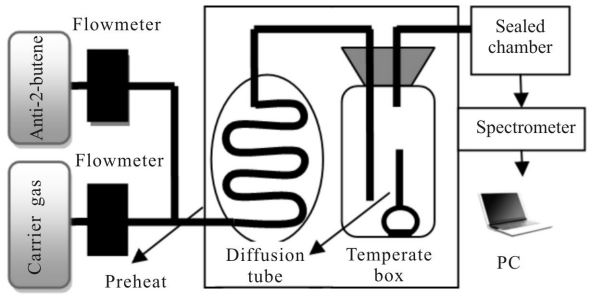

图 1 实验装置框图

Fig.1 Block diagram of experiment device

将得到的混合气体光谱数据进行 ICA 解析,并且分别采用普通快速独立分量分析 FastICA 算法和优化后的 FastICA 算法对吸收峰位置互相交错重叠的特征光谱进行识别, 也对两种算法的识别率和迭代速度进行了对比, 分别选取不同的数据量来对两种算法的识别率和迭代速度进行比较, 结果如图 2 和表 1 所示。可以看出:就该实验而言,FastICA 算法和优化后的 FastICA 算法的分离能力基本一致,应用优化后的 FastICA 算法能够减少 FastICA 的迭代次数,进一步加快收敛速度,也充分说明了该优化算法能够在提高算法性能的情况下保持 FastICA 算法在特征提取方面的良好性能。

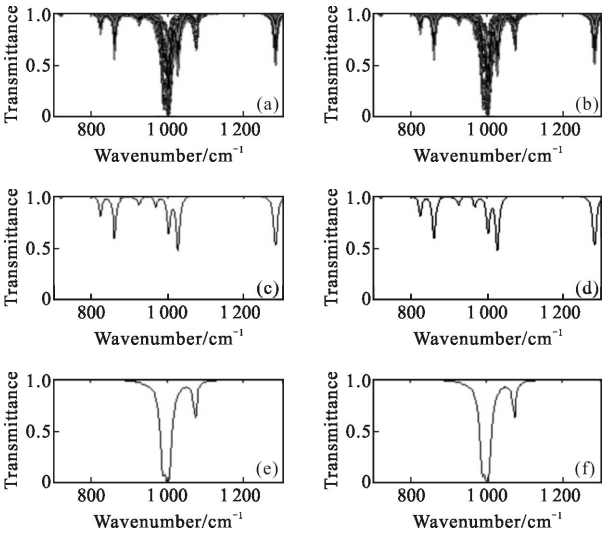

图 2 两种算法分离出的 DMMP 和反 2 丁烯的谱图

Fig.2 Isolated DMMP and anti-2-butene spectra of two algorithms

表 1 FastICA 和优化后的 FastICA 迭代次数和时间

Tab.1 Iterations and time of FastICA and optimized FastICA

Algorithm	Sample number			
	100	200	300	400
FastICA (Frequency)	16	20	26	32
FastICA(Time/s)	0.104 4	0.235 8	0.327 5	0.430 2
Optimized FastICA (Frequency)	14	15	18	21
Optimized FastICA (Time/s)	0.100 2	0.200 5	0.301 0	0.411 3

图 3 为 FastICA 和 PCA 处理后的谱图。由图 3 可以看出,PCA 处理后由于一些信息的丢失导致吸收峰消失或者合并在强吸收峰内, 使某一吸收峰的峰宽变窄或增大, 消失的峰位很可能影响到对特征光谱的解析,也可能会影响后续的定量分析。为进一步检验 FastICA 算法在混合情况下对于特征峰交错重叠中特征提取的优越性, 分别用 PCA 算法和 FastICA 算法对光谱数据处理后进行定量分析。

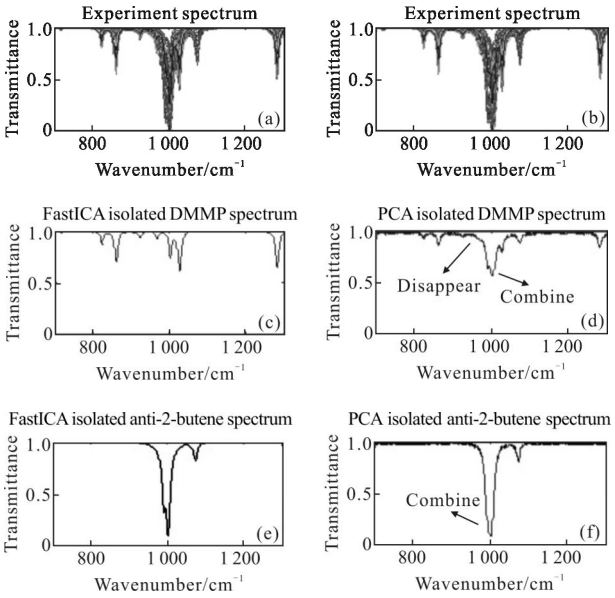

图 3 FastICA 和 PCA 处理后的谱图

Fig.3 Spectra after processed by FastICA and PCA

2.2 定量分析

将优化 FastICA 和 PCA 算法各自分离出来的 400 个 DMMP 样品光谱中的 320 个样品作为训练集,80 个样品作为测试集,用极限学习机(ELM)建立浓度预测模型进行定量分析，其中隐含层神经元个数为 20,预测结果如图 4、5 所示。

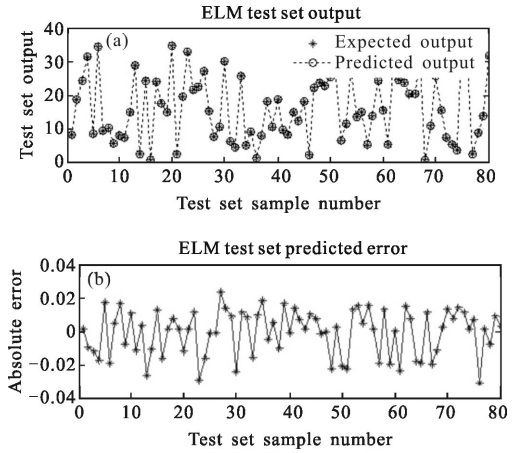

图 4 FastICA 处理后 DMMP 测试集预测结果

Fig.4 DMMP test set predictions after processed by FastICA

优化 FastICA 处理后测试集预测结果中，均方误差 E=2.392 6×10^{-4},回归系数 R=0.999，PCA 处理后测试集预测结果中,均方差 E=2.501 3×10^{-4},回归系数 R=0.989，说明FastICA 算法对于混叠峰的分离效果良好,定量分析结果更精确一点。

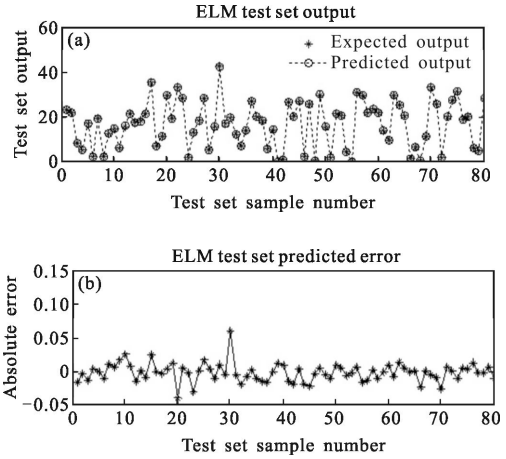

图 5 PCA 处理后测试集 DMMP 预测结果

Fig.5 DMMP test set predictions after processed by PCA

3 结 论

提出了一种以 FastICA 算法为基础的红外光谱主次吸收峰严重混叠的识别方法，它能够提取高阶统计信息，对混合气体中各气体吸收峰重叠的光谱进行识别分离,同时为了契合实时探测的要求,对传统的 FastICA 算法的迭代过程进行优化。利用搭建的实验系统采集多组化学战剂模拟剂 DMMP 和反 2 丁烯混合气体的红外光谱数据，得到的数据分别经过传统 FastICA 和优化 FastICA 在不同数据量下迭代次数的对比和分离精确度的对比,优化的 FastICA 迭代次数减少,分离精度与传统 FastICA 相当。定量分析中分别用 PCA 和优化 FastICA 对数据进行处理，再经过建立 ELM 的定量分析模型,PCA 处理后测试集预测结果中,均方差 E=2.501 3×10^{-4},回归系数 R=0.989，优化 FastICA 处理后测试集预测结果中,均方误差 E=2.392 6×10^{-4},回归系数 R=0.999,该优化方法相比于常用的 PCA 算法精度上有一定的优越性。优化后的 FastICA 算法能够在不影响普通 FastICA 良好分离性能的前提下有效地减少了普通 FastICA 算法的迭代次数，进一步加快收敛速度,这与光谱识别的实时要求是很好的契合，具有普遍的实用价值。

参考文献：

[1] Chu Xiaoli. Molecular Spectroscopy Andlytlcal Technology Combined with Cheemometrics and its Applications [M]. Beijing: Chemical Industry Press, 2011. (in Chinese)

褚小立. 化学计量学方法与分子光谱分析技术[M]. 北京: 化学工业出社, 2011.

[2] Wang Yifan, Tang Zhengning. Dimensionality reduction method based oncombination of PCA and ICA [J].***Optical Technique***, 2014(2): 180-183. (in Chinese)
王一帆, 唐正宁. 基于 PCA 和 ICA 的多光谱数据降维方法 [J]. 光学技术, 2014(2): 180-183.

[3] Hyvarinen A, Oja E. Independent component analysis: algorithms and applications [J]. ***Neural Networks***, 2000, 13: 411-430.

[4] Mei Tiemin. Theory and Algorithms of Blind Source Separation [M]. Xi′an: Xidian University Press, 2013. (in Chinese)
梅铁民. 盲源分离理论与算法[M]. 西安: 西安电子科技大学出版社, 2013.

[5] Xu Wenli, Sun Tong, Hu Tian, et al. Huanghua pear soluble solids contents Vis/NIR spectroscopy by analysis of variables optimization and FICA [J]. ***Spectroscopy and Spectral Analysis***, 2014, (12): 3253-3256. (in Chinese)
许文丽, 孙通, 胡田, 等. 基于变量优选和快速独立成分分析的黄花梨可溶性固形物可见/近红外光谱检测 [J]. 光谱学与光谱分析, 2014, (12): 3253-3256.

[6] Wan Jun, Zhang Xiaohui, Rao Jionghui, et al. Processing of backscattering signal of warship wake flow based on independent component analysis [J]. ***Infrared and Laser Engineering***, 2013, 42(1): 244-250. (in Chinese)
万俊, 张晓晖, 饶炯辉, 等. 基于独立成分分析的舰船气泡尾流后向散射光信号处理 [J]. 红外与激光工程, 2013, 42 (1): 244-250.

[7] Nikos Pasadakis, Kardamakis Andreas A. Identifying constituents in commercial gasoline using Fourier transform infrared spectroscopy and independent component analysis[J]. ***Analytica Chimica Acta***, 2006, 578(2): 250-255.

[8] Cheng Yuanyuan, Wang Zhibin, Wang Zhaoba. Mind evolutionary bat algorithm and its application to feature selection of mixed gases infrared spectrum [J]. ***Infrared and Laser Engineering***, 2015, 44(3): 845-851. (in Chinese)
陈媛媛, 王志斌, 王召巴. 思维进化蝙蝠算法及其在混合气体红外光谱特征选择中的应用 [J]. 红外与激光工程, 2015, 44(3): 845-851.

[9] Yang Fusheng, Hong Bo. Theory and Application of Independent Component Analysis [M]. Beijing: Tsinghua University Press, 2006: 1-88. (in Chinese)
杨福生, 洪波. 独立分量分析的原理与应用 [M]. 北京: 清华大学出版社, 2006: 1-88.

[10] Hyvarinen A. Fast and robust fixed-point algorithm for independent component analysis [J]. ***IEEE Transactions on Neural Networks***, 1999, 10(3): 626-634.