【文章编号】 1004-1540(2008)02-0137-05

近红外光谱数据处理的独立分量分析方法研究

方利民，林 敏
（中国计量学院 计量测试工程学院，浙江 杭州 310018）

【摘 要】 从数学的角度分析比较了主成分分析(PCA)与独立分量分析(ICA)的原理和特点，给出光谱矩阵在两种不同分析方法下的不同分解；同时结合线性回归和神经网络回归，提出“两步法”来确定不同成分含量测定的最优模型. 进而采用 PCA 与 ICA 对实际测得的玉米近红外光谱进行了处理，比较分析了两种不同分解所得矩阵的化学含义，以及 PCA 与 ICA 两种不同分解对玉米光谱分析结果的影响. 仿真结果表明，ICA 从独立性角度对光谱数据矩阵进行分解，所得结果更接近实际光谱. 最后，利用“两步法”对玉米三种主要成分水、淀粉、蛋白质分别建立了各自最优含量测定模型. 结果表明，所建模型符合快速测定要求，具有一定的实用价值.

【关键词】 独立分量分析；主成分分析；神经网络；近红外光谱；玉米样品
【中图分类号】 O657.33 【文献标识码】 A

Research on independent component analysis in near-infrared spectroscopy processing

FANG Li-min, LIN Min
(College of Metrology Measurement and Engineering, China Jiliang University, Hangzhou 310018, China)

Abstract: The principles and characteristics of PCA and ICA were analyzed from the point of view of mathematics, and two different decompositions of spectral matrix were given by using those two methods. Combining with linear regression or neural network, a “two-step method”, that was used to build the different components' optimal model was proposed. ICA and PCA was used to process the measured near-infrared spectra of corn respectively; and the chemical meaning of both matrixes derived from the two decompositions and the influence of the two methods on results were analyzed. The experimental results show that ICA does the matrix decomposition from the perspective of independence, and the results are closer to the actual spectrum. The “two-step method” was applied to build the optimal model of the three major components (moisture, starch, protein) of corn samples. The results show that the models with rapid prediction requirement have practical value.

Key words: independent component analysis; principal component analysis; neural network; near infrared spectroscopy; corn sample

【收稿日期】 2008-04-03

【作者简介】 方利民(1983－)，男，浙江建德人，硕士研究生. 主要研究方向为化学计算学方法.

近红外光谱(Near Infrared Spectroscopy, NIR)被誉为20世纪90年代以来发展最快的光谱分析技术[1],是光谱测量技术与化学计量学的有机结合,被誉为分析的巨人[2].它利用物质的近红外吸收光谱信息,采用化学计量学方法分析处理实验数据,从而对样品进行定性、定量分析测定,是一种快速、无损的新型检测技术.在测定农副产品(如谷物、饲料、水果、蔬菜、肉、蛋、奶等)的品质(如水分、蛋白、油脂含量等)方面已得到广泛使用[3].化学计量方法是NIR在定量定性分析中有效应用的保证,数据处理主要包括多元线性回归法(MLR)、偏最小二乘法(PLS)、主成分回归(PCR)、人工神经网络法(ANN)等等.鉴于近红外光谱本身的解析难点,如包含信息强度低等,越来越多新兴的化学计量方法正被应用于光谱的解析、建模[4-8].

主成分分析(Principal Component Analysis, PCA)是一种古老的多元统计分析技术;以它为核心的主成分回归分析在近红外光谱也得到了相当广泛的应用.独立分量分析(Independent Component Analysis, ICA)是近年发展起来的一种全新的数据分析工具,是解决盲源分离问题的一种有效的方法.从20世纪90年代出现,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.在分析化学领域也逐渐显示了它的强大作用[9-12].

本文从数学的角度将PCA与ICA进行对比,给出光谱矩阵在两种不同分析方法下的不同分解.并通过对实际测得的玉米近红外光谱的处理,提出"两步法"确定不同成分含量测定的最优模型,并给出PCA与ICA两种不同分解对玉米光谱分析结果的影响.

1 PCA与ICA的方法比较

PCA的目的是在尽可能保持原始变量更多信息的前提下,导出一组零均值随机变量相对少的不相关线性组分(主成分).它是通过计算数据协方差矩阵的特征值和特征向量来实现的.保留相关的具有最大特征值的特征向量,PCA就可以用作一个降低数据维数,对数据进行白化处理以保留主变量、提取信号特征的工具.通过计算原始输入数据的协方差矩阵的特征向量,PCA可以将高维数的输入数据矢量线性地变换为低维数的、且各分量之间均不相关的矢量,在矢量空间里,PCA确定了主方向,以及相应的数据变量的长度.

但是,按PCA原理做出的分解只能保证分解出来的各分量不相关,却不能保证这些分量互相独立.这就使得这样的分解缺少实际意义,因而降低了所提取特征的典型性.在很多情况下,观测值实际上是由若干相对独立的信源的加权和组成的,若采用ICA来分解独立分量,再从各独立分量中提取有关特征,就可能更有实际意义,有助于进一步的模式识别[13,14].

记 $X=[x_1,x_2,\ldots,x_l]^T$, $S=[s_1,s_2,\ldots,s_m]^T$, $\boldsymbol{M}$ 是 $l\times m$ 混合矩阵,其元素为 m_{ij} ($i=1,\ldots,l;j=1,\ldots,m$),则ICA的基本模型可写为(要求 $l\geqslant m$)

$$X=\boldsymbol{M}S \tag{1}$$

ICA的目的是在假设源信号相互统计独立的条件下,仅通过可获得的 l 个传感器信号来估计混合矩阵 $\boldsymbol{M}$ 以及源信号 S,即要求找到分离矩阵 $\boldsymbol{W}$ 使其满足

$$\tilde{S}=\boldsymbol{W}X \tag{2}$$

其中 $\tilde{S}$ 是 S 的估计,$\boldsymbol{W}$ 就是混合矩阵 $\boldsymbol{M}$ 的Moore-Penrose逆.ICA属于盲信号处理(Blind Signal Processing, BSP)问题,"盲"意味着对混合信号体系的信息知之甚少,在对源信号和传输通道几乎没有可利用信息的情况下,仅从观测到的混合信号中提取或恢复出源信号.

进行ICA可以概括为[14]:

1)原理一:非线性去相关 求解混阵 $\boldsymbol{W}$ 使其任意两输出 $\tilde{s}_i,\tilde{s}_j$ ($i\neq j$)不但本身不相关,而且经非线性变换后的分量 $g(\tilde{s}_i),h(\tilde{s}_j)$ 也不相关.函数 g,h 自然要适当选择.

2)原理二:使输出尽可能非高斯化 在输出某分量 $\tilde{s}$ 的方差恒定的条件下,将输入 X 各分量作线性组合 $\tilde{s}=\sum_i b_i x_i$.优化选择各权重 b_i,使 $\tilde{s}$ 尽可能非高斯化,则 $\tilde{s}$ 的非高斯性的每一个局部极大值给出一个独立分量.

值得注意的是,ICA与PCA不同,后者是按照能量大小排序进行的分解,而且只保证分解出

的分量互相正交；前者则要求各分量在能量（方差）相等的条件下尽可能独立. 一般的 ICA 算法都要对输入信号进行预白化，也即原理一所说的非线性去相关，但与 PCA 的不相关又有一定的区别. 如图 1 所示，直观地表示了对两个均匀分布混合信号 x_1，x_2 的三种基本变换结果（白化处理、PCA 和 ICA）. ICA 把信号分离为两个独立的成分，其散点图呈正方形，如图 1(d). 白化处理使各成分不相关，其散点图也呈正方形，但相对 ICA 的散点图旋转了个角度，如图 1(b)，说明白化后的信号仍是两个独立源信号的线性组合（独立性不强）. 而 PCA 仅能从能量的角度分离主要和次要成分，对信号的相关性和独立性没有根本的改观，如图 1(c)中，其散点图呈为菱形. 从对信号分离的角度看，ICA 效果最好，白化处理次之、PCA 最差.

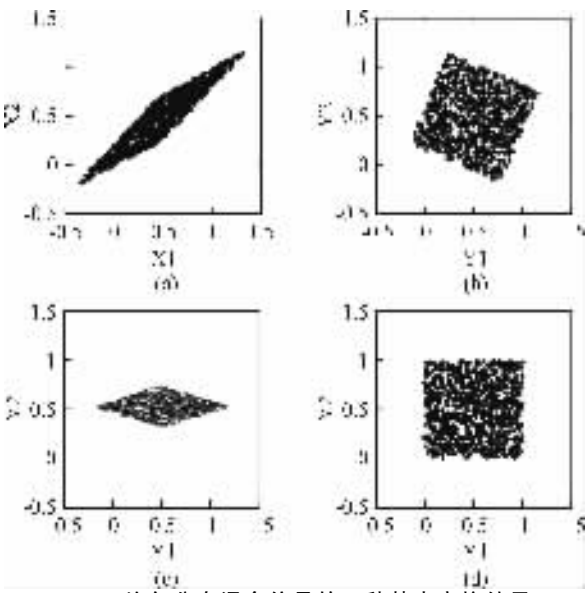

图 1　均匀分布混合信号的三种基本变换结果

Figure 1　Three basic transformations of the mixed signal with uniform distribution

2　NIR 光谱模型

根据 Beer-Lambert 定律，对于未知的混合体系，测得的近红外光谱通常认为是一些纯物质（主要成分）光谱的线性组合. 因此，对于 NIR 光谱数据矩阵 $\boldsymbol{A}_{l\times n}$，可以建模为各成分的光谱信号与其浓度乘积的加和：

$$\boldsymbol{A}=\boldsymbol{MI} \tag{3}$$

其中，$\boldsymbol{A}_{l\times n}$ 是 l 个样品在 n 个波长处的近红外光谱数据矩阵，$\boldsymbol{I}_{m\times n}$ 是独立成分矩阵，在理想的分解状态下相当于纯物质的光谱数据矩阵，$\boldsymbol{M}_{l\times m}$ 是混合矩阵，它与纯物质在混合样品中的浓度有关.

ICA 根据模型(3)，将每个样品的近红外光谱作为 m 个独立成分的线性组合. 光谱矩阵 $\boldsymbol{A}_{l\times n}$ 分解后，所得 $\boldsymbol{I}$ 的每一行相当于一种统计独立成分的光谱信息，该独立成分在混合光谱中的相对浓度信息，在混合矩阵 $\boldsymbol{M}$ 中得以体现，即 $\boldsymbol{M}$ 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 NIR 光谱的贡献. 因此，可以肯定的是，混合矩阵 $\boldsymbol{M}$ 与浓度矩阵 $\boldsymbol{C}$ 之间存在一定的函数关系. 应该指出的是，ICA 分离出的成分是相互独立的，而传统的主成分分析(PCA)和因子分析(FA)分离出的成分则是相互正交的，这是 ICA 与 PCA、FA 的主要区别.

若混合矩阵 $\boldsymbol{M}$ 与浓度矩阵 $\boldsymbol{C}$ 之间的函数关系为线性关系，则可建立类似主成分回归的独立分量回归(Independent Component Regression，ICR)，浓度矩阵 $\boldsymbol{C}$ 和经 ICA 计算得到的混合矩阵 $\boldsymbol{M}$ 符合下列回归方程：

$$\boldsymbol{C}=\boldsymbol{MB} \tag{4}$$

$\boldsymbol{B}$ 为回归系数矩阵.

但是，混合矩阵 $\boldsymbol{M}$ 与浓度矩阵 $\boldsymbol{C}$ 之间的函数关系可能并非线性. 对于这种情况，将使用神经网络模型来建立他们之间的关系. 考虑 BP 神经网络的设计，将混合矩阵 $\boldsymbol{M}$ 作为输入，要求预测的浓度矩阵 $\boldsymbol{C}$ 作为网络的输出，采用具有三层的神经网络. 由于近红外光谱数据矩阵一般比较庞大，首先采用离散小波变换对近红外光谱数据进行压缩，再将压缩后的数据进行 ICA 分解，最后将分解所得的混合矩阵作为网络的输入. 这样，既有效提取特征信息，又极大地减少数据量，从而提高网络的运行速度.

3　数据和性能评价指标

3.1　采样数据

所用数据为 Cargill 公司提供的 80 个玉米样品的近红外光谱数据及其对应的淀粉、水和蛋白质含量值. 对于每一个玉米样品，在 1 100～2 498 nm波长范围内，每隔 2 nm 测定其吸光度，

得到的近红外光谱数据矩阵为80行700列的，如图2的玉米样本的近红外光谱图. 所用算法为自编Matlab7.01软件(Mathworks®)环境下小波压缩算法、PCA算法、FastICA算法和BP神经网络算法.

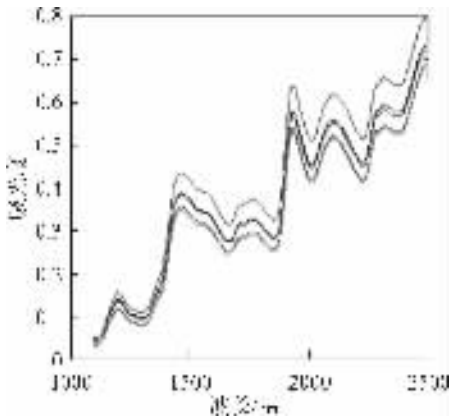

图2 玉米样品近红外光谱图

Figure 2 Near infrared spectra of corn samples

3.2 性能评价指标

模型预测过程中，以均方根误差(RMSEP)和相关系数(R)的大小作为模型预测准确度的评价. RMSEP的数值越小，R越大，模型的预测准确度越高. RMSEP定义如下

$$\mathrm{RMSEP}=\sqrt{\frac{\sum_{i=1}^{n}(\widetilde{y_i}-y_i)^2}{n}} \tag{5}$$

式(5)$\widetilde{y_i}$为y_i的预测值.

4 仿真结果与讨论

4.1 参数说明

线性模型直接由(4)式可得，对于PCA用得分矩阵$\boldsymbol{T}$替代$\boldsymbol{M}$即可.

非线性模型采用3层BP神经网络：输入层、中间隐层和输出层. 传递函数分别用tansig函数和purelin函数，优化学习算法选用的是Levevberg-Marquardt学习算法. 为了提高网络的推广能力，在训练之前将80个玉米样品划分为训练样品集、验证样品集和测试样品集，其中验证样品集和测试样品集从80个样品中均匀选取1/8而生成，余下的60个作为训练样品集. 首先，网络的输入是PCA分解得到的得分矩阵$\boldsymbol{T}$或者经ICA计算得到的混合矩阵$\boldsymbol{M}$. 其次，中间隐层神经元数的选取关系到模型的拟合精度，因此需要通过选取不同的值来得到最优值. 经实验分析，选取最优取值为nods=7. 最后，由于该网络可用于对三种主要成分的含量测定，输出层节点数可以固定为3个，若要单独测定，则可将其固定为1个.

值得注意的是，对于PCA与ICA分解中分量数PCs与ICs的选择不仅关系到各自算法的精度，而且与网络的运行速度以及所建模型的精确度有关. 在实际操作中，将采用不同的分量数来得到最优的数值.

4.2 结果分析

1)将80个玉米样品经预处理后分别进行PCA分解和ICA分解，分别给出在主成分数PCs=4、5的得分矩阵T和独立分量数ICs=4、5的混合矩阵M的比较分析.

(1)PCA：取PCs=4，T为80×4矩阵，80个样本对应的4个主成分的最大得分值分别为2.851 7，0.586 2，0.073 5，0.053 3；平均值为0.726 3，0.110 9，0.027 1，0.018 4；

取PCs=5，T为80×5矩阵，80个样本对应的5个主成分的最大得分值分别为2.851 7，0.586 2，0.073 5，0.053 3，0.039 9；平均值为0.726 3，0.110 9，0.027 1，0.018 4，0.010 6.

可见，PCA分解是按照能量大小排序进行的分解，前两个主分量占了98%的能量. 从均值与最大值的差别可见不同样本经PCA分解所得的得分向量差别较大. 此外，分解成分的数目PCs=5时，前四个成分与PCs=4时是一样的.

(2)ICA：取ICs=4，$\boldsymbol{M}$为80×4矩阵，80个样本对应的4个独立分量的权重值之间的差别最大值为0.023 1，0.008 5，0.020 3，0.009 3；

取ICs=5，$\boldsymbol{M}$为80×5矩阵，80个样本对应的5个独立分量的权重值之间的差别最大值为0.018 1，0.017 8，0.007 1，0.004 9，0.002 8.

ICA分解与PCA有明显的不同，不同样本对应的独立分量的权重值之间的差别很小. 说明分离出的独立分量包含混合光谱潜在的共同信息，也即可能是共同的“纯光谱”，因此ICA所得结果更接近实际光谱，更具有价值. 此外，ICA分解ICs=5时的前4个成分并不是ICs=4时分解所得成分，相比PCA，ICA分解时选取的不同成分数对结果的影响更大.

2）对 PCA 与 ICA 方法所建模型的预测能力作比较. 用 PCA、ICA 方法结合线性、非线性回归所建模型对玉米样品分别进行了淀粉、水和蛋白质的含量预测，结果表明，不同的成分的含量测定对模型的线性或者非线性要求不同，比如水在使用非线性模型（神经网络回归）时的预测结果要明显优于线性模型.

因此，在实际操作中，对不同成分含量的测定需选取相应的最优模型. 这里采用“两步法”：首先，固定分解方法，即采用 PCA 还是 ICA，比较线性回归与非线性回归的优劣；然后，固定第一步选取的回归方法，即采用线性还是非线性，比较 PCA 和 ICA 两种分解方法的优劣，最后所得模型即为该成分含量测定的最优模型. 比如对水分含量的测定，通过第一步选取非线性回归，再比较 PCA 与 ICA 方法的优劣，见表 1.

表 1　PCA 与 ICA 非线性模型对水分含量的预测比较

Table 1　Prediction performance index of ICA-ANN method and PCA-ANN method

ICs	ICA-ANN		PCA-ANN	
Or PCs	R	RMSEP	R	RMSEP
3	0.774 2	0.243 4	0.750 5	0.250 2
4	0.865 2	0.189 5	0.788 9	0.233 1
5	0.908 5	0.159 5	0.895 5	0.168 1
6	0.927 1	0.142 5	0.904 4	0.163 7
7	0.985 5	0.067 2	0.937 9	0.133 5

R：correlation coefficient；RMSEP：root mean square error of prediction.

由表 1 可见，ICA 非线性模型要明显优于 PCA 非线性模型的预测结果. 因此，玉米水分含量测定的最优模型可定为 ICA 结合非线性模型. 利用 ICA-ANN 方法，选取参数 ICs＝7，nods＝7，对玉米样品中的水分含量进行预测，预测值与化学分析所得测定值的相关性分析，见图 3. 可以看出，预测结果与试验测定结果之间具有很好的线性相关关系，能够满足玉米样品中水分含量的快速分析.

对淀粉、蛋白质含量的测定同样可以采用上述的“两步法”来确定最优模型. 实验结果表明，对淀粉和蛋白质含量的测定所得最优模型均为 ICA-ANN 模型，在选取最优参数 ICs＝10、nods＝7 后，所得指标分别为：蛋白质 R＝0.976 0，RMSEP＝0.106 6；淀粉 R＝0.971，RMSEP＝0.179 1，模型的预测结果是令人满意的. 文献[15]采用离散余弦变换和 BP 神经网络，建立玉米 NIR 与主要成分之间的关系模型. 将其结果与本文所得结论比较，本文所得的三种主要成分含量测定结果均优于该文献的结果.

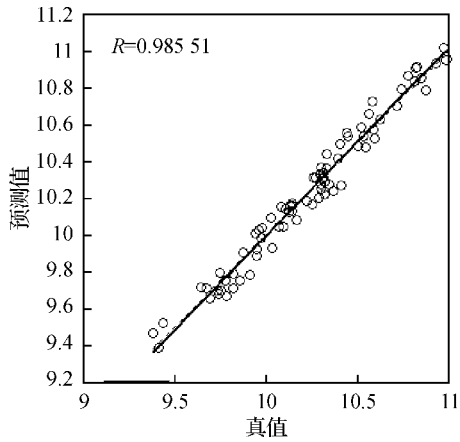

图 3　玉米水分含量真值与 ICA-ANN 方法预测值相关性

Figure 3　Correlation between predicted value and true value of the moisture content in corn

5　结　语

与主成分分析相比，独立分量分析（ICA）方法不仅能够从样品光谱中分解出主要成分的光谱信息，而且能实现样品成分含量的测定. ICA 方法提取出的独立分量与实际光谱之间更为接近，更能体现光谱的真实情况. 用 ICA 方法对近红外光谱进行分解，再与线性或者 ANN 回归相结合，建立的最优分析模型更具有实际意义，而且模型的预测结果也是令人满意的. 本文结合 PCA、ICA 和人工神经网络，提出“两步法”确定不同成分含量测定的最优模型，应用于玉米样品的近红外光谱数据处理，可作为近红外光谱分析的补充，具有良好的应用前景.

【参　考　文　献】

[1]　徐广通，袁洪福，陆婉珍. 现代近红光谱技术及应用进展[J]. 光谱学与光谱分析，2000，20(2)：134-142.

[2]　MCCLURE W F, CROWELL B, STANFIELD D L, et al. Near infrared technology for precision environmental measurements: I Determination of nitrogen in green and dry-grass tissue[J]. Near Infrared Spectrosc, 2002(10): 177-185.

（下转第 145 页）

【参　考　文　献】

[1] 胡豫杰，张志刚，常　茹. 浅议我国热计量方法的选用与发展方向[J]. 天津城市建设学院学报，2001，7(4)：282-284.

[2] 齐世清，曲秀云，林　平. 微功耗 IC 卡热能表的研制[J]. 仪器仪表学报，2003，24(4)：169-172.

[3] 金海龙，潘　勇. 新型智能热量表的开发研究[J]. 传感技术学报，2005，18(2)：350-352.

[4] 王长军，朱善安. 温度自校正型低功耗热能表[J]. 自动化仪表，2004，25(2)：27-29.

[5] 赵伟国，梁国伟，李文军. 低功耗远传热能表的研究[J]. 电测与仪表，2005，42(480)：31-34.

[6] 梁国伟，王　芳，李长武，等. 基于热传递的铂膜气体流量计实验研究[J]. 中国计量学院学报，2006，17(1)：36-39.

[7] 梁一灵，梁国伟，王雨辰. 组合热膜探头在气体流速测量中的应用[J]. 中国计量学院学报，2007，18(3)：191-194.

[8] 魏小龙. MSP430 系列单片机接口技术及系统设计实例[M]. 北京：北京航空航天大学出版社，2002：134-169.

[9] 袁汶雯，高　峰，陈　隆. 水表集抄系统的低功耗设计[J]. 电子技术应用，2002，28(10)：38-41.

[10] 袁志勇，邹久朋. 过程装备中数据采集系统的低功耗设计[J]. 工业仪表与自动化装置，2003(1)：34-36.

[11] 黄竹霞. 智能系统中的低功耗设计技术[J]. 汽轮机技术，2002，44(5)：269-270.

[12] 李国祥，吕士健，王树铎，等. 中华人民共和国城镇建设行业标准——热能表[S]. 北京：中国标准出版社，2000.

(上接第 141 页)

[3] 陆婉珍. 现代近红外光谱分析技术[M]. 2 版. 北京：中国石化出版社，2007：51-98.

[4] 陈华才，陈星旦. 近红外光谱在药物领域的应用与研究进展[J]. 中国计量学院学报，2003，14(4)：0261-0267.

[5] 陈华才，吕　进，俸春红，等. 近红外光谱法测定茶多酚中总儿茶素含量[J]. 中国计量学院学报，2005，16(1)：17-20.

[6] 刘辉军，吕　进，林　敏，等. 基于 RBF 网络和 NIRS 的绿茶水分含量分析模型[J]. 中国计量学院学报，2005，16(3)：188-190.

[7] 林　敏，毛谦敏，吕　进，等. 最优小波包变换的化学模式特征选择方法[J]. 中国计量学院学报，2005，16(3)：182-187.

[8] 林　敏，吕　进，徐立恒，等. 茶叶近红外光谱数据的离散余弦变换压缩方法[J]. 中国计量学院学报，2003，14(4)：0268-0270.

[9] CHEN J，WANG X Z. A new approach to near-infrared spectral data analysis using independent component analysis[J]. J Chem Inf Comput Sci，2001，41：992-1001.

[10] SANGJOOH H，GILWON Y. Identification of pure component spectra by independent component analysis in glucose prediction based on mid-infrared spectroscopy[J]. Applied Optics，2006，45(32)：8374-8380.

[11] 毕　贤，李通化，吴　亮. 独立组分分析在近红外光谱分析中的应用[J]. 高等学校化学学报，2004，25(6)：1023-1027.

[12] SHAO X G，WANG W，HOU Z Y，et al. A new regression method based on independent component analysis[J]. Talanta，2006，69：676-680.

[13] ANDRZEJ C，SHUN-ICHI A. Adaptive blind signal and image processing[M]. New York：John Wiley & Sons，2002：1-156.

[14] 杨福生，红　波. 独立分量分析的原理与应用[M]. 北京：清华大学出版社，2006：1-25.

[15] 林　敏，吕　进. 基于神经网络与近红外光谱的玉米成分检测方法[J]. 红外技术，2004，26(3)：78-81.