江西农业大学学报 2012,34(4):828 - 831 http://xuebao.jxau.edu.cn
Acta Agriculturae Universitatis Jiangxiensis E - mail: ndxb7775@sina.com

独立分量分析在近红外光谱定量分析中的应用

谢秀娟[1],赵龙莲[2]

(1. 福建农林大学 计算机与信息学院,福建 福州 350002; 2. 中国农业大学 信息与电气工程学院,北京 100094)

摘要: 采用独立分量分析(ICA)方法,对玉米样品的近红外光谱进行分解,得到统计上独立的各成分光谱;然后用多元回归方法建立基于ICA成分的玉米粗蛋白质、粗淀粉和粗脂肪含量的定量分析模型,3种成分建模集和预测集的化学值和近红外预测值之间的相关系数都较高,且平均相对误差都较低。结果表明,ICA方法建立的玉米样品3个主要成分的近红外模型预测准确度都较高,可应用于玉米育种中大批样品的快速品质分析。

关键词: 独立分量分析;近红外光谱;定量分析
中图分类号: O657.33 **文献标志码:** A **文章编号:** 1000 - 2286(2012)04 - 0828 - 04

The Application of ICA to Quantitative Analysis by Near Infrared Reflectance Spectroscopy

XIE Xiu-juan[1], ZHAO Long-lian[2]

(1. Institute of Computer and Information, Fujian Agriculture and Forestry University, Fuzhou 350002, China; 2. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China)

Abstract: Independent component analysis was applied to decompose the independent components in near infrared spectral of corn powder samples. Each component spectrum obtained was statistically independent. And quantitative analysis model based on ICA constituents of corn protein, starch and crude fat content was established by means of multiple regressions. The correlation coefficients between the prediction value by NIRS and chemical analysis results of the three components in modeling set and prediction set were relatively high and the mean relative errors were lower. The result shows that the accuracy of the NIR model based on ICA to analysis three main components of the corn samples is high comparatively, and the model can be used in quality analysis in large quantities of maize breeding samples.

Key words: independent component analysis (ICA); near infrared reflectance spectroscopy (NIRS); quantitative analysis

独立分量分析(independent component analysis,ICA)是20世纪90年代后期发展起来的一种盲信源分解的方法[1],它利用数据的高阶统计性质,把信号分解成若干个互相独立或尽可能独立的成分,可广泛应用于信号的分离和特征提取[2]。传统的信源分解技术主要是建立在主成分分析(principal component analysis,PCA)的基础上,它根据方差极大原理,去除向量间的线性相关,找出原始信号中隐含的内在信息,目的在于降低向量维数,且分解出的成分都是按照能量的大小排列的。但按照PCA原理分解出

收稿日期:2012 - 12 - 21 修回日期:2012 - 04 - 15
基金项目:福建省自然科学基金项目(2008J0210)和福建省教育厅科技项目(JA08062)
作者简介:谢秀娟(1972—),女,讲师,硕士,主要从事信号与信息处理研究,E - mail: xjuan1126@126.com。

来的各成分只能保证不相关,却不能保证这些成分互相独立,这就使得这样的分解缺少实际的物理或生理意义,因而降低了所提取特征的典型性。而采用 ICA 来分解独立成分,再从独立成分中提取有关特征,就可能会更有实际意义,有助于进一步的模式识别[1-2]。

近红外光谱分析(near infrared spectroscopy,NIRS)技术具有分析速度快、无污染、低消耗、非破坏性,可以实现多组分同时测定等优点[3-4],经过 50 多年的发展,近红外光谱分析技术已广泛应用于农业、食品、药品、生物、化妆品、纺织、多聚物、有机物生产等领域。近红外光谱法作为一种快速分析方法已经在众多领域中被得到应用[5-6]。

独立分量分析作为一种盲信号分离的有效方法[7-8],在语音识别、图像处理、生物医学信号处理等领域已经被得到广泛应用,如文献[9]利用 ICA 方法从高分辨率训练图像中提取出独立分量进行处理,重建结果提高了人脸辨识;文献[10]采用独立分量分析和小波变换结合,可更好地降低膈肌肌电信号中的心电干扰;也有应用于光谱数据分析的报道,邵咏妮等[11]研究了用 ICA 和 BP 神经网络法对稻谷的可见/近红外光谱进行分析,实现了对稻谷年份的鉴别;毕贤等[12]将 ICA 用于红外光谱定性分析,从混合光谱中分离出独立组分的光谱。本文以玉米粉末样品为例,研究 ICA 方法在近红外光谱定量分析中的应用。

1　材料与方法

1.1　玉米样品的近红外光谱

玉米样品的粗蛋白质、粗淀粉和粗脂肪含量是衡量玉米营养品质的重要指标,而这些品质指标的常规测定方法速度慢、费用高,不适于品质育种工作中大批量育种材料的鉴定筛选。而近红外光谱分析技术的特点使得它特别适合于育种工作中大批样品的快速品质分析。

玉米粉末样品共 90 个(过 40 目筛),由中国农科院品种资源所提供。在 Bruker Vector 22/N 傅里叶变换近红外光谱仪上采集其漫反射光谱,光谱范围为 4 000 ~ 12 000 cm^{-1},分辨率为 8 cm^{-1},得到的 90 个玉米样品的近红外光谱如图 1 所示。

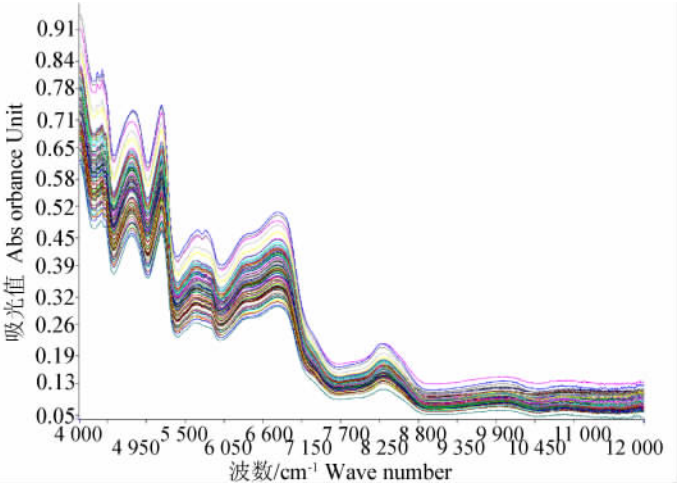


图 1　玉米粉末样品的近红外光谱

Fig. 1　NIRS of corn powder samples

1.2　玉米样品化学值的测定

常规化学法测得每个样品粗淀粉、粗蛋白质和粗脂肪的化学值含量(单位样品中各成分所占的百分比),其中粗蛋白质含量采用国标 GB5511—1985 测定,粗脂肪含量采用国标 GB 5512—1985 测定,粗淀粉含量采用国标 GB 5006—1985 测定。

1.3　近红外光谱数据处理

1.3.1　ICA 理论简介　设 $X=(x_1,x_2,\cdots x_m)$ 是 m 维观测信号,则 ICA 的数学模型表示为:

$$X=AS=\sum_{i=1}^{n}\alpha_1 S_1 \tag{1}$$

(1)式中,A 是未知的 $m\times n$ 混合矩阵,用来表示信号源到接收阵的传递函数;$S=(s_1,s_2,\cdots,s_n)^T$ 是分量彼此统计独立的 n 维源信号。

ICA 理论认为用来观测的混合数据阵 X 是由独立源 S 经 A 线性加权获得。利用观测信号 x_i($i=1,2,\cdots,n$)的信息来估计混合矩阵 A 和独立成分 s_i,需求得一个分离矩阵 W,使之得到最佳分离。

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

(2)式中 W 作用在 X 上所获得的信号 Y 是独立源 S 的最优逼近,该分离矩阵为:

$$W=(A^TA)^{-1}A^T \tag{3}$$

因分离后的信号 Y 与源信息 S 之间的比例因子以及排列对应顺序无法确定，所以，若分离后的信号之间是相互独立的，即认为已正确实现了信号分离[13]。基于负熵的快速定点迭代 FastICA[8] 算法如下：

(1) 观测信号 X 做去均值和白化预处理，设白化后的信号为 Z 满足 $E(ZZ^T) = I$。

(2) 选择具有单位方差的初始分离矩阵 W。

(3) 迭代计算 $E[ZG(WZ)] - E[G(W^TZ)]W \Rightarrow W$。

(4) 归一化处理分离矩阵 $W/||W|| \Rightarrow W$。

(5) 判断 W 是否收敛，若收敛则分离出一个独立分量 W^TZ，否则返回步骤(3)。

(6) 判断混合信号中的多个独立分量是否已经全部分离完毕，若没有则返回(2)，否则分离过程结束。

1.3.2 基于 ICA 的定量分析模型的建立 近红外光谱定量分析模型的建立步骤如下：

(1) 随机选择90个玉米样品中的2/3为建模集，剩余1/3为预测集，选取玉米粉末光谱中信息量大且噪声较小的 4 000～8 000 cm^{-1} 波段作为分析谱区。

(2) 为了消除高频随机噪声对分析模型的影响，采用中心化和一阶导数法(15点平滑)对光谱数据进行预处理。

(3) 采用 FastICA 算法提取光谱的独立成分，得到玉米粗蛋白质、粗淀粉和粗脂肪3种主要成分的近红外光谱。

(4) 用多元回归法建立基于 ICA 成分的玉米粗蛋白质、粗淀粉和粗脂肪含量的定量分析模型。

2 结果与分析

根据建模集留一法交叉验证的结果选取9个 ICA 成分代表样品的近红外光谱，即取9个 ICA 成分参与建模，再用所建模型对预测集样品进行预测。表1列出了建模集交叉验证的结果，包括预测集的化学值和近红外预测值之间的相关系数，平均绝对误差和平均相对误差，同时列出了用 PCA 作为特征提取方法的结果。

表1 玉米粉末样品建模集和预测集定量分析结果

Tab.1 Quantitative analysis results to corn powder samples modeling set and prediction

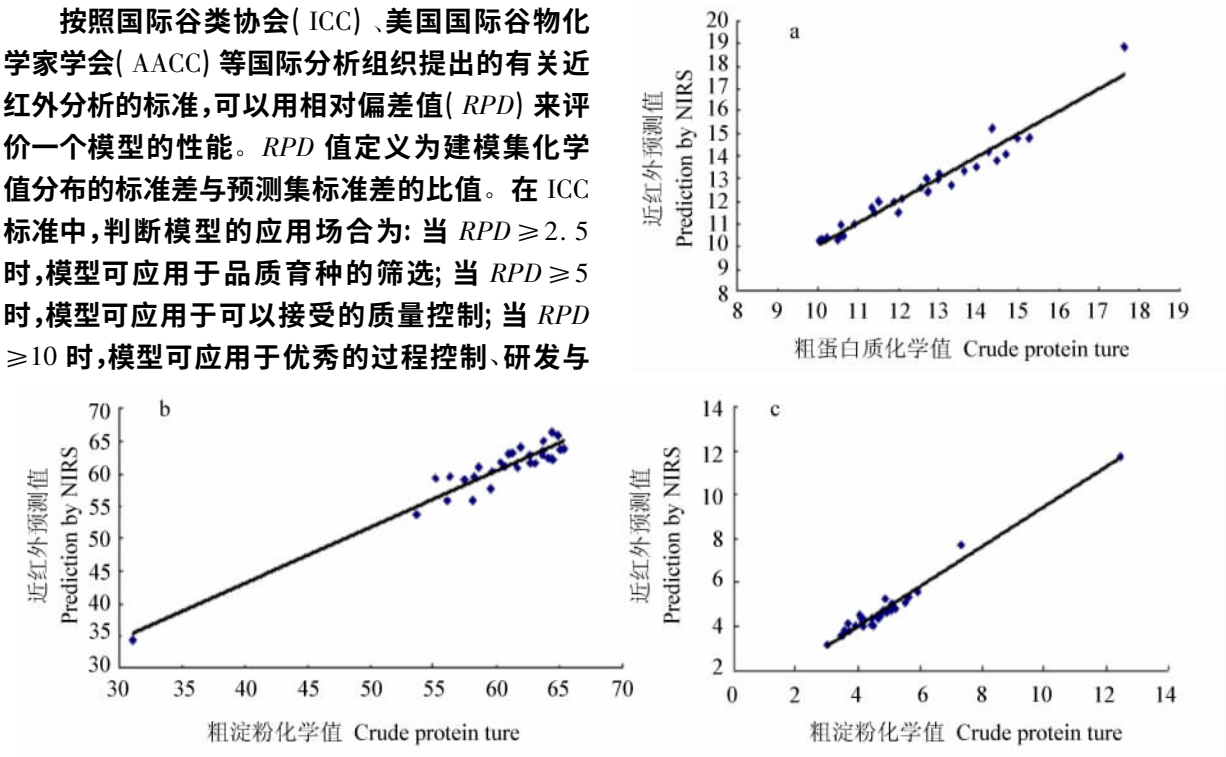
成分 Composition	参数 Parameter	建模集 Modeling set		预测集 Prediction set	
		PCA	ICA	PCA	ICA
粗蛋白质	相关系数	0.953 6	0.955 7	0.977 4	0.977 1
Crude protein	平均绝对误差/%	0.342 7	0.352 8	0.328 4	0.316 2
	平均相对误差/%	2.752 5	2.854 6	2.576 1	2.486 2
粗淀粉	相关系数	0.872 1	0.906 5	0.950 3	0.959 3
Crude starch	平均绝对误差/%	1.750 9	1.718 9	1.534 3	1.575 6
	平均相对误差/%	3.166 1	3.031 5	2.573 5	2.766 9
粗脂肪	相关系数	0.962 6	0.964 0	0.982 7	0.984 4
Crude fat	平均绝对误差/%	0.290 6	0.285 6	0.283 8	0.272 5
	平均相对误差/%	6.578 8	6.524 6	5.864 0	5.629 8

由表1结果可以看出，用 PCA 和 ICA 2种方法进行特征提取，然后建立判别模型，所得的结果相当。利用 ICA 法进行特征提取，玉米样品粗蛋白质、粗脂肪和粗淀粉3种组分建模集和预测集化学值和近红外预测值间相关系数都较高，预测集的平均相对误差较低，分别为：2.486 2%，2.766 9%，5.629 8%。

图2所示为预测集样品粗蛋白质、粗脂肪和粗淀粉3种组分的化学值和近红外预测值的散点图。可以看出，各数据点很好地分布在回归线两侧，说明了用常规化学法测得的玉米3种不同成分的化学值和近红外预测值之间的拟合存在较好的线性关系。

为了进一步分析这3个模型的性能，因此将模型建模样品化学值的分布范围，平均值和标准差列于表2。

按照国际谷类协会(ICC) 、美国国际谷物化学家学会(AACC) 等国际分析组织提出的有关近红外分析的标准,可以用相对偏差值(*RPD*) 来评价一个模型的性能。*RPD* 值定义为建模集化学值分布的标准差与预测集标准差的比值。在 ICC 标准中,判断模型的应用场合为: 当 $RPD \geqslant 2.5$ 时,模型可应用于品质育种的筛选; 当 $RPD \geqslant 5$ 时,模型可应用于可以接受的质量控制; 当 $RPD \geqslant 10$ 时,模型可应用于优秀的过程控制、研发与应用的研究。本文中玉米粗蛋白质、粗淀粉和粗脂肪 3 个模型的 *RPD* 值分别为: $RPD_{粗蛋白质} = 1.523/0.3162 = 4.82$; $RPD_{粗淀粉} = 5.145/1.5756 = 3.27$; $RPD_{粗脂肪} = 1.445/0.2725 = 5.30$。三者的 *RPD* 值都大于 2.5,达到 ICC 规定的要求,因此该模型至少可以用于品质育种的筛选。



(a) 玉米粗蛋白质预测集样品化学值和近红外预测值散点图; (b) 玉米粗淀粉预测集样品化学值和近红外预测值散点图; (c) 玉米粗脂肪预测集样品化学值和近红外预测值散点图。

(a) scatter diagram of prediction value by NIRS and the chemical value of corn protein in prediction set; (b) scatter diagram of prediction value by NIRS and the chemical value of corn crude starch in prediction set; (c) scatter diagram of prediction value by NIRS and the chemical value of corn Crude fat in prediction set.

图 2　预测集样品的化学值和近红外预测值的散点图

Fig. 2　Scatter diagram of the chemical value and prediction value by NIRS in prediction set

表 2　建模集样品的化学值分布

Tab. 2　Distribution of chemical value of modeling sample set

成分 Composition	含量分布范围/% Content distribution	平均值/% Average value	标准差 Standard deviation
粗蛋白质 Crude protein	9.163 ~ 16.862	12.260	1.523
粗淀粉 Crude starch	31.210 ~ 67.960	60.677	5.145
粗脂肪 Crude fat	2.286 ~ 12.078	4.642	1.445

3　结论与讨论

近红外光谱分析是一种间接分析技术,其准确性受样品的代表性、样品化学值的准确性等因素的影响[14],它的定标过程复杂,需要选取大量具有代表性的样品进行分析[4]。因此,必需扩大模型样品的覆盖范围,在模型中不断添加更多更复杂的新样品,以不断完善模型,为 ICA 分析提供更准确的光谱数据。

研究结果表明,采用 FastICA 算法提取玉米样品近红外光谱的 ICA 成分,并用多元回归法建立基于 ICA 成分的玉米粗蛋白质、粗淀粉和粗脂肪含量的定量分析模型,3 种组分建模集和预测集的化学值和近红外预测值间相关系数与 PCA 方法分析的结果相比都较高; 进一步分析预测集样品的化学值和近红

(下转第 838 页)

[9] 胡勇有,王鑫,张太平. 用低浓度生活污水筛选适于华南人工湿地的植物[J]. 华南理工大学学报: 自然科学版,2006,34(9):78-82.
[10] Cheng S,Grosse W,Karrenbrock F,et al. Efficiency of constructed wetlands in deconta-mination of water polluted by heavy metals[J]. Ecology Engineering,2001,18(3):317-325.
[11] 廖新娣,骆世明. 人工湿地对猪场废水有机物处理效果的研究[J]. 应用生态学报,2003,26(1):113-117.
[12] Solano M L,Soriano P,Ciria M P. Constructed wetlands as a sustainable solution waste water treatment in small villages[J]. Biosystems Engineering,2004,87(1):109-118.
[13] 杨林,伍斌,赖发英. 7种典型挺水植物净化生活污水中氮磷的研究[J]. 江西农业大学学报,2011,33(3):616-621.

(上接第831页)

外预测值的散点图,表明常规化学法测得的玉米不同成分的化学值含量和近红外预测值拟合存在较好的线性关系。因此,ICA方法建立的玉米样品主要成分的近红外模型具有较高的预测准确度,能满足一般分析的要求,可将该模型应用于玉米育种中大批样品的品质分析中。

参考文献:

[1] 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.
[2] Comom P. Independent component analysis: A new con-cept[J]. Signal Processing,1994,36(3):287-314.
[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,2006: 174-203.
[4] 张灵帅,邢军,王卫东,等. 近红外光谱分析技术进展及其在烟草行业中的应用[J]. 光谱实验室,2009,26(2):197-201.
[5] 严衍禄,赵龙莲,杨曙明,等. 近红外光谱分析基础与应用[M]. 北京: 中国轻工业出版社,2005: 190-260.
[6] 赵龙莲,张录达,李军会,等. 小波包熵和Fisher判别在近红外光谱法鉴别中药大黄真伪中的应用[J]. 光谱学与光谱分析,2008,28(4):817-820.
[7] Hyvarinen A,Oja E. Independent component analysis: Algorithms and application[J]. Neural Networks,2000,13(4/5):411-430.
[8] Hyvarinen A. Fast and robust fixed-point algorithm for independent component analysis[J]. IEEE Trans on Neural Networks,1999,10(3):626-634.
[9] 乔建苹. 基于独立分量分析的人脸超分辨率重建[J]. 计算机工程,2011,37(3):180 - 182.
[10] 伍飞云,杨智,范正平,等. 基于独立分量分析和小波变换的膈肌肌电信号降噪[J]. 信号处理,2010,26(10):1532 - 1538.
[11] 邵咏妮,曹芳,何勇. 基于独立组分分析和BP神经网络的可见/近红外光谱稻谷年份的鉴别[J]. 红外与毫米波学报,2007,26(6):433-436.
[12] 毕贤,李通化,吴亮. 独立组分分析在红外光谱分析中的应用[J]. 高等学校化学学报,2004,32(6):44-48.
[13] 朱佳,袁晓辉. 基于独立分量分析的说话人自动识别方法的研究[J]. 仪器仪表与分析监测,2011(1):13-16.
[14] 孟兆芳,赵龙莲,程奕,等. 近红外光谱法测定玉米品质指标的研究[J]. 华北农学报,2008,23(2):147-150.