

The Application of Different Maize Varieties Classification and Character Evaluation by Multivariate Statistical Method

Jing Liu, Siyu Li, Xin Liu, Fengjiang Li, Mingzhi Lu

College of Chemistry and Life Science, Qinghai University for Nationalities, Xining Qinghai
Email: jjajing333@163.com

Received: Jan. 20th, 2015; accepted: Jan. 31st, 2015; published: Feb. 6th, 2015

Copyright © 2014 by authors and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Background: Corn is cultivated all over the country. Because it has short growth cycle, during the growing season it requires warm and rainy weather. Hebei province is a province with large production of maize yield. To improve the yield, the increase plan of summer maize production is essential. **Methods:** 10 agronomic characters of 14 maize varieties are analyzed by R software for correlation analysis, clustering, principal component analysis and factor analysis. It provides a theoretical basis for evaluation and utilization of the Hebei summer maize premature group of High-Yielding Maize Varieties breeding and varieties (combination). **Conclusions:** 1) Jingshi96-3, Zhongyu NO. 11, Cang9005 × Ji815, 150-4 × H21, and other comprehensive properties are a good combination; 2) Yield factor, ear coarse factor, growth factor and density factor can affect the output of corn, and thus the economic yield.

Keywords

Corn, Agronomic Traits, Clustering Analysis, Principal Component Analysis, Factor Analysis

多元统计分析在不同玉米品种分类和性状评估中的应用

刘 晶, 李思宇, 刘 欣, 李丰江, 芦明智

青海民族大学化学与生命科学学院, 青海 西宁

Email: jiajing333@163.com

收稿日期: 2015年1月20日; 录用日期: 2015年1月31日; 发布日期: 2015年2月6日

摘 要

背景: 玉米在我国各地均有栽培, 因为其生长周期短, 生长期要求温暖多雨。河北省是玉米产量大省, 为了提高产量每年的夏玉米增产计划是必不可少的, 方法: 对14个玉米品种的10个农艺性状应用R软件进行相关分析、聚类分析, 主成分分析和因子分析。为河北夏玉米早熟组高产玉米品种选育和品种(组合)评价利用提供理论依据。结论: 1) 京试96-3, 中玉11号, 沧9005 × 冀815, 150-4 × H21等几个综合性状是较为优良的组合。2) 产量因子, 穗粗因子, 生长因子, 密度因子都能影响玉米的产量, 从而影响经济产量。

关键词

玉米, 农艺性状, 聚类分析, 主成分分析, 因子分析

1. 引言

《本草纲目》曰: 玉米, 原名玉蜀黍, 又名苞米或玉麦。作为我国第二大粮食作物, 提高玉米产量的意义是非常重要的。在遗传育种中, 以多变量分析为基础的多元统计分析在研究数量性状间关系和性状间相互关系的遗传规律研究中得到广泛应用[1]-[3]。基于此基础, 本文对 2006 年阜城原种场试验站河北夏玉米早熟组原始数据进行相关分析, 聚类分析, 主成分分析因子分析。

2. 方法

2.1. 数据来源

如表 1。

2.2. 分析方法[4] [5]

应用 R 软件对不同组合与各个性状间的关系进行相关分析, 聚类分析, 主成分分析、因子分析。

相关分析: 相关分析是对客观现象具有的相关关系进行的研究分析。其目的在于帮助我们对关系的密切程度和变化的规律性有一个具体的数量上的认识, 作出判断, 并且用于推算和预测。其主要内容包括: 1) 确定现象之间有无关系; 2) 确定现象之间关系的密切程度; 3) 测定两个变量之间的一般关系值; 4) 测定因变量估计值和实际值之间的差异。

聚类分析: 聚类分析(Cluster analysis)是根据事物本身的特性研究个体分类的方法, 其原则是同一类中的个体有较大的相似性, 不同类的个体差别比较大。根据分类对象的不同分为样品聚类和变量聚类。

主成分分析: 是一种分析、简化数据集的技术。主成分分析经常用于减少数据集的维数, 同时保持数据集中的对方差贡献最大的特征。这是通过保留低阶主成分, 忽略高阶主成分做到的。这样低阶成分往往能够保留住数据的最重要方面。但是, 这也不是一定的, 要视具体应用而定。由于主成分分析依赖所给数据, 所以数据的准确性对分析结果影响很大。

因子分析: 因子分析是将多个实测变量转换为少数几个不相关的综合指标的多元统计分析方法。在

Table 1. The summer of 2006 Fucheng seed stock station experimental station in Hebei corn premature set of raw data
表 1. 2006 年阜城原种场试验站河北夏玉米早熟组原始数据

	亩产量 y	穗位 x1	千粒重 x2	全生育期 x3	穗长 x4	倒伏 x5	穗粗 x6	出粒率 x7	株高 x8	秃尖 x9
冀单 17 号	411.33	108	334	92	13.9	3	4.5	83.1	242	1.5
冀 161 × 35-66	457.79	111	211	99	15.7	2	4.2	86.4	258	0.5
12 × 15	360.67	116	252	93	13.1	12.5	3.5	12	255	1.1
150-4 × H21	461.11	104	276	93	16.7	1.4	4.1	83.2	271	1.4
196 × 原 92	369.16	95	235	93	13.5	2.4	4.1	75.7	230	1.2
京 218 × 京 594	450.67	93	283	98	19.1	1.6	3.7	82.3	240	1.5
沧 9005 × 冀 815	474.29	77	274	93	14.3	1.2	4	89.9	217	0.9
京试 96-3	468.98	79	260	98	17	0.4	3.8	88.4	218	0.8
远征 808	420.44	125	239	95	13.9	3.7	4.3	83.4	246	0.7
廊玉 6 号	426.55	107	317	95	13.7	1.2	4.8	89	222	0.8
中玉 11 号	432.21	82	240	94	16.5	1.8	4	84.2	241	1.4
八苏 × 90110	366.82	121	232	93	15.9	4.3	3.8	82	249	1.2
95101	435.26	110	235	102	13	3.9	4.4	83.4	254	1.2
冀单 301	464.12	95	268	95	13	1.9	4.4	83.4	241	1.2

各个领域的科学研究中往往需要对反映事物的多个变量进行大量的观测，收集大量数据以便进行分析寻找规律。

3. 结果

3.1. 性状间相关关系

从表 2 和假设检验可以知道亩产量与穗位呈显著的负相关，亩产量与倒伏呈极显著负相关，亩产量与出粒率呈显著的正相关。穗位与株高呈极显著正相关，倒伏与出粒率呈极显著负相关。其他变量间没有显著的关系

3.2. 聚类分析

3.2.1. R 型分析

从图 1 中可得，对于不同品种的 10 个变量进行分类，可以分为 4 类。第一类与产量有紧密正相关的变量(全生长期，穗长，出粒率，和亩产量)第二类与产量有不太紧密正相关的变量(千粒重，穗重)，第三类与产量有负相关的变量(到伏，穗位，株高)和第四类与产量无关的变量(秃尖)。

3.2.2. Q 型分析

用四种不同方法进行聚类分析结果基本保持一致从图 2 中可以得到以下结论：

第一类：品种 1(冀单 17 号)品种 10(廊玉 6 号)品种 14(冀单 301)

特点：亩产量较高，千粒重较高，全生长期较短，穗长较短，倒伏较少，穗粗较粗，株高较低。

第二类：品种 5(196 × 原 92)品种 12(八苏 × 90110)品种 13(95101)品种 2 冀(161 × 35-66)品种 9(远征 808)

特点：千粒重较低，全生长期较长，倒伏较多，穗粗较粗。

Table 2. The relationship between different corn varieties characters
表 2. 不同玉米品种性状间相关关系

r	y	x1	x2	x3	x4	x5	x6	x7	x8	x9
y	1									
x1	-0.55*	1								
x2	0.16	-0.15	1							
x3	0.41	-0.01	-0.34	1						
x4	0.34	-0.34	-0.03	0.16	1					
x5	-0.68**	0.53	-0.19	-0.19	-0.4	1				
x6	0.22	0.18	0.38	0.09	-0.5	-0.42	1			
x7	0.60**	-0.31	0.12	0.24	0.29	-0.94**	0.52	1		
x8	-0.16	0.61**	-0.31	0.07	0.05	0.39	-0.11	-0.30	1	
x9	-0.19	-0.15	0.34	-0.29	0.22	0.05	-0.17	-0.09	0.23	1

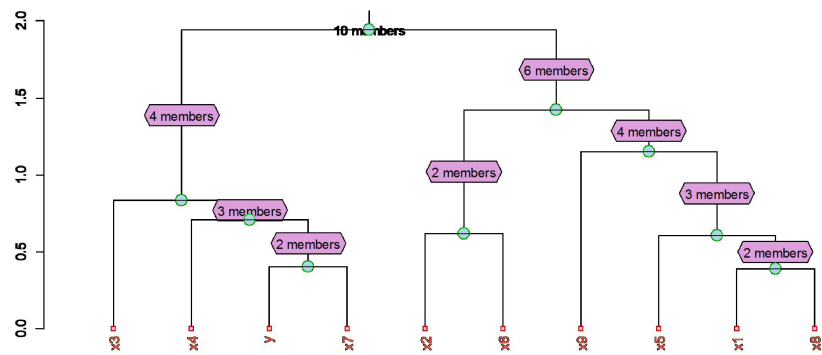


Figure 1. Different varieties of maize variable cluster analysis

图 1. 不同品种玉米变量聚类分析

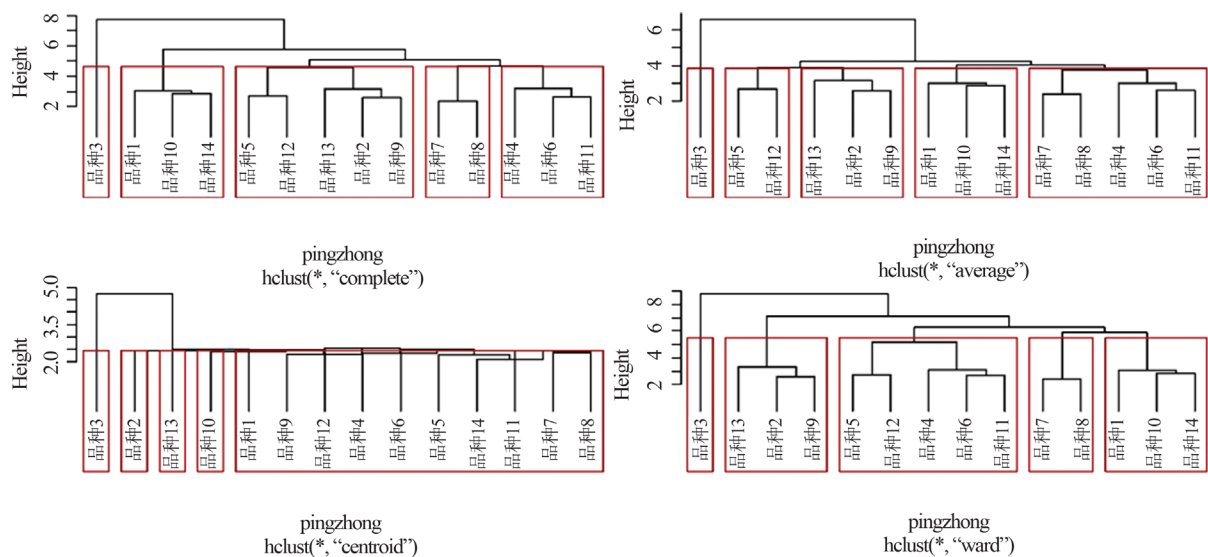


Figure 2. Different corn varieties cluster analysis

图 2. 不同玉米品种聚类分析

第三类：品种 7(京 218 × 京 594)品种 8(沧 9005 × 冀 815)
特点：亩产量高，穗位偏低，倒伏较少，出粒率较高，株高较低，秃尖较少。
第四类：品种 4(150-4 × H21)品种 6(沧 9005 × 冀 815)品种 11(中玉 11 号)
特点：穗长最长，倒伏较少，秃尖较长。
第五类：品种 3(12 × 15)
特点：亩产量最低，全生长期较短，穗长较短，穗粗最细，出粒率最低。
综合评价，第三类群为最优良群体。具有产量高、倒伏少，出粒率高，株高低、穗位低的特点。

3.3. 主成分分析

用 R 程序主成分分析表格如表 3。
由表 3 可知第一主成分贡献率为 35%，第二主成分贡献率为 18%，前五个主成分累计贡献率已达 90%，因此只需用五个主成分就能很好概括这组数据。除了通过计算累积贡献率判断之外，通常也将碎石图作为直观判断的依据。具体方法是：在碎石图上找一拐点，使得该点之后的点纵坐标对应的值均较小且相差不大，则取拐点的序号为主成分的个数。
从图 3 可知取 5 个主成分很切合实际。
从表 4 可得，以下结果

$$\begin{cases} Z_1 = 0.44y - 0.35X_1 + 0.15X_2 + 0.16X_3 + 0.2X_4 - 0.51X_5 + 0.2X_6 + 0.47X_7 - 0.28X_8 \\ Z_2 = -0.3X_1 - 0.25X_2 + 0.58X_4 - 0.64X_6 - 0.12X_7 + 0.10X_8 + 0.23X_9 \\ Z_3 = -0.16y - 0.25X_1 + 0.55X_2 - 0.56X_3 - 0.26X_8 + 0.48X_9 \\ Z_4 = 0.35X_1 + 0.17X_2 + 0.24X_4 - 0.13X_5 + 0.24X_6 + 0.21X_7 + 0.65X_8 + 0.5X_9 \\ Z_5 = -0.42y + 0.21X_1 - 0.4X_2 - 0.54X_3 + 0.2X_4 - 0.31X_5 + 0.41X_7 - 0.14X_9 \end{cases}$$

由上面式子线性关系可以看出：

Table 3. Principal component extraction analysis table					
表 3. 主成分提取分析表					
	第一主成分	第二主成分	第三主成分	第四主成分	第五主成分
标准差	1.87	1.35	1.34	1.10	0.81
方差占比例	0.35	0.18	0.18	0.12	0.07
累计比例	0.35	0.53	0.71	0.83	0.9

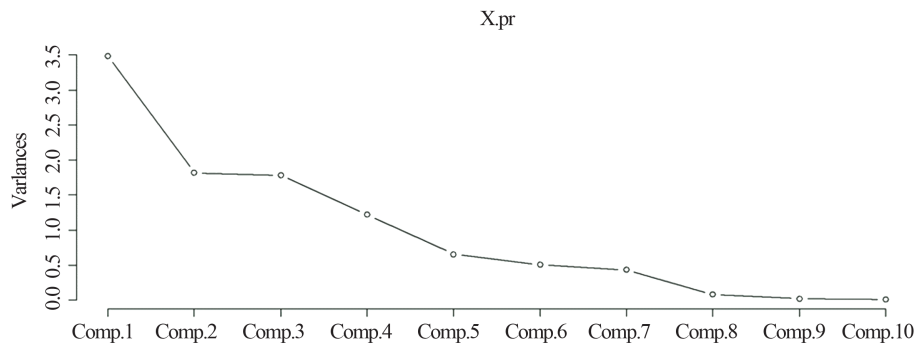


Figure 3. Principal component analysis of gravel
图 3. 主成分分析碎石图

Table 4. Principal component coefficient matrix
表 4. 主成分系数矩阵

	1	2	3	4	5	6	7	8	9	10
y	0.44		-0.16		-0.42	-0.24	0.53	0.34	0.25	-0.28
X1	-0.35	-0.3	-0.25	0.35	0.21	-0.35	-0.29	0.41		-0.42
X2	0.15	-0.25	0.55	0.17	-0.4	-0.48	-0.24		-0.3	0.24
X3	0.16		-0.56		-0.54	0.24	-0.53		-0.15	
X4	0.2	0.58		0.24	0.2	-0.43	-0.3	-0.34	0.36	
X5	-0.51			-0.13	-0.31			0.19	0.61	0.47
X6	0.2	-0.64		0.24		0.12		-0.51	0.45	
X7	0.47	-0.12		0.21	0.41	0.12		0.46		0.55
X8	-0.28	0.1	-0.26	0.65			0.44	-0.21	-0.29	0.32
X9		0.23	0.48	0.5	-0.14	0.56	-0.11	0.2	0.16	-0.24

在第一主成分里, y(亩产量), X2(千粒重), X3(全生育期), X4(穗长), X6(穗粗), X7(出粒率)为正指标, 其中以亩产量和出粒率最为显著。而 X1(穗位), X5(倒伏), X8(株高)为负指标, 其中以倒伏最为显著。可以总结第一主成分主要是对产量多少的一个描述。

在第二主成分里, X4(穗长), X8(株高), X9(秃尖)为正指标, 其中以穗长最为显著。而 X1(穗位), X2(千粒重), X6(穗粗), X7(出粒率)为负指标, 其中以穗粗最为显著。可以总结第二主成分主要是对玉米穗粗细长短位置等形态的一个描述。

在第三个主成分里, X2(千粒重)和 X9(秃尖)为正指标, 其中二者都达到了 0.5 左右, y(亩产量), X1(穗位), X3(全生育期), X8(株高)为负指标, 其中以全生育期最为显著。可以得到第三主成分主要是对玉米生长时间有关的一个描述。

在第四个主成分里, X1(穗位), X2(千粒重), X4(穗长), X6(穗粗), X7(出粒率), X8(株高), X9(秃尖)为正指标, 其中以 X1(穗位), X8(株高), X9(秃尖)最为显著, X5(倒伏)为负指标, 可以总结第四主成分主要是对倒伏的描述。

在第五个主成分里, X1(穗位), X4(穗长), X7(出粒率)为正指标, 其中以出粒率为最为显著, 而 y(亩产量), X2(千粒重), X3(全生育期), X5(倒伏), X9(秃尖)为负指标, 其中以 y(亩产量), X2(千粒重), X3(全生育期)最为显著。可以看到第五主成分是对生长情况一个描述。

根据表 5 可以得到各个品种综合得分表 6。

从表 6 中可以得到京 218 × 京 594 综合排名第一, 京试 96-3 是综合排名第二, 中玉 11 号综合排名第三, 沧 9005 × 冀 815 综合排名第四, 150-4 × H21 综合排名第五。

3.4. 因子分析

3.4.1. 因子分析公因子的意义

本研究共收集到 14 份玉米品种 10 个性状的数据, 应用 KMO 和 Bartlett 检验表明, 本试验符合因子分析要求, 可对数据集进行因子分析。因子分析采用主成分方法, 按最小特征根大于 1 的原则保留公因子, 共得总累计贡献率为 84% 的 5 个公因子。为简化对因子的解释对因子采用方差极大正交旋转, 各因子旋转后载荷矩阵见表 7。

Table 5. The varieties of the principal component values
表 5. 各个品种的主成分值

	第 1 主成分	第 2 主成分	第 3 主成分	第 4 主成分	第 5 主成分
品种 1	-0.09	-1.59	2.38	1.27	-0.26
品种 2	0.42	0.03	-3.12	0.11	0.26
品种 3	-5.47	0.72	0.50	-1.19	-1.15
品种 4	0.24	0.83	0.47	2.21	0.13
品种 5	-0.74	-0.29	0.73	-1.01	1.136
品种 6	1.21	2.53	0.55	0.96	-0.63
品种 7	1.98	0.06	0.93	-1.83	-0.07
品种 8	2.38	1.44	-0.44	-1.63	-0.34
品种 9	-0.79	-1.33	-1.43	-0.04	0.78
品种 10	1.33	-2.68	0.63	-0.41	-0.16
品种 11	0.74	1.43	0.68	0.09	0.46
品种 12	-1.83	0.63	-0.07	0.48	1.73
品种 13	-0.21	-0.80	-2.03	0.87	-1.27
品种 14	0.83	-0.97	0.24	0.12	-0.62

Table 6. The varieties comprehensive score
表 6. 各个品种综合得分

排名	品种	得分	排名	品种	得分
6	冀单 17 号	0.27	2	京试 96-3	0.88
11	冀 161 × 35-66	-0.42	13	远征 808	-0.80
14	12 × 15	-2.13	8	廊玉 6 号	0.04
5	150-4 × H21	0.66	3	中玉 11 号	0.76
9	196 × 原 92	-0.25	10	八苏 × 90110	-0.40
1	京 218 × 京 594	1.17	12	95101	-0.63
4	沧 9005 × 冀 815	0.72	7	冀单 301	0.14

Table 7. The factor analysis of the characteristic value and contribution
表 7. 因子分析中特征值和贡献率

	Factor1	Factor2	Factor3	Factor4	Factor5
因子对方差贡献	2.37	1.69	1.60	1.40	1.31
方差贡献率	0.24	0.17	0.160	0.14	0.13
累计贡献率	0.24	0.41	0.57	0.71	0.84

用 R 程序得到结果如下表 7。

通过假设检验可知，五个因子可以代表整体样本的信息。

从表 8 可以看出，第 1 公因子载荷阵中出粒率最大，穗粗、亩产量次之，穗长、生长期、千粒重较

Table 8. Factor analysis load matrix
表 8. 因子分析载荷矩阵

	Factor1	Factor2	Factor3	Factor4	Factor5
y(亩产量)	0.42	-0.13	-0.15		0.88
X1(穗位)	-0.23	0.37	0.65		-0.34
X2(千粒重)	0.10	0.18	-0.21	0.95	
X3(生长期)	0.13			-0.42	0.44
X4(穗长)	0.27	-0.85			0.14
X5(倒伏)	-0.91	0.13	0.26		-0.27
X6(穗粗)	0.56	0.78	0.10	0.21	0.10
X7(出粒率)	0.93		-0.15		0.21
X8(株高)	-0.18	-0.12	0.97		
X9(秃尖)		-0.36	0.26	-0.514	-0.24

小。而表现负向较大的有倒伏、穗位、株高。表明出粒率较大，穗粗、亩产量、穗长、生长期、千粒重之间存在相互促进的关系，这些性状都与产量较为直接的关系，故可称第 1 公因子为产量因子。从第 1 公因子来看，产量育种中应注重对出粒率、穗粗、亩产量、穗长、生长期与千粒重性状的选择而倒伏对产量有较高负向作用，育种中应特别注重提高品种的病倒伏性，因为倒伏引起产量的损失是最直接的原因。穗位、株高与产量呈负向作用，说明穗位、株高都是引起倒伏的重要原因。第 2 公因子载荷阵中，穗粗和穗位载荷值最高，两个性状与穗相关，且穗长与第二公因子呈负向相关，因此可称之为玉米穗粗因子。一般来说穗粗、穗位高较高的品种其生物产量也较高。第 3 公因子载荷阵中，穗位和株高值最高，两个性状与玉米植株生物量生产相关，因此可称作生长因子。一般来说株高、穗位高较高的品种其生物产量也较高。第 4 公因子载荷阵中千粒重最高、秃尖穗粗依次之，在生产上，密度和秃尖与千粒重之间有密切的联系，因此这些性状都与密度有关，故第 4 公因子可称之密度因子。第五公因子的载荷阵中，亩产量和生长期最高，而穗位、倒伏呈负向作用，这些性状都与产量有关，为了区别第一公因子，我们命名为产量第二因子。故由于第五因子所占方差比例比较小可与第一因子合并。故主要由四个公因子影响，分别是产量因子，穗粗因子，生长因子，饱满度因子。

3.4.2. 因子分析图形表示

在现代农业科学研究中常需对生物的多个性状进行综合分析，若是对第一个公因子产量进行研究可以从图中看出，品种 10，5，4 得分最高。若是对第二个公因子穗粗因子进行研究，从图中可以看出品种 10，13，9 得分最高。由图 4 可依次类推。

4. 讨论

- 1) 利用多元统计分析使我们能从宏观方面对性状间的关系进行剖析，加深对性状间内在联系规律性认识。利用因子分析对 10 个原始性状变量进行综合，根据基因型的因子得分进行聚类一方面加深对这些性状间内在关系有更深理解，这样可使育种中有针对性地对几个主因子进行选择 and 改良，提高品种选育中效率与鉴定准确性，另一方面通过对性状的聚类，利于对性状进行优缺点评估，为选育更加贴近目标环境品种提供有价值信息。
- 2) 本研究中玉米品种(组合)的 10 个性状里，主要涉及的是产量、发育和抗性性状，经因子分析简化

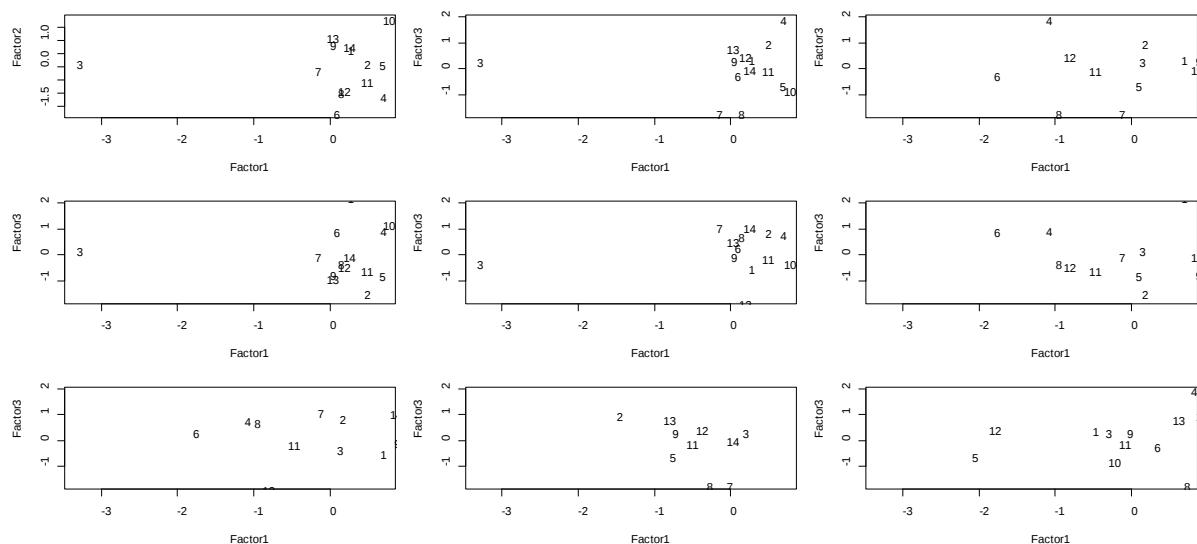


Figure 4. 14 varieties under different two factors of scatter plot

图4. 14个品种在不同两个因子下的散点图

为5个公因子，其中第1公因子、第2公因子、第4公因子、第5公因子对产量有正向作用，而第3公因子都对产量有负向作用。第1公因子与产量最为密切，该因子中与产量同向的性状对产量都有正向的协同作用，提高相对性状值可有效提高产量。第2公因子与产量呈一定负相关，从植物生理学来看，株高更多体现的是生物产量性状。第3公因子与产量因子呈一定的正相关，但却未列入产量因子中，说明穗粗因子性状可能与产量因子性状在遗传基础的差异性，协调好穗粗与穗长可以提高产量。第4公因子密度因子与产量在一定程度呈正相关，但是到达一定高度时又呈负相关。所以合理安排种植密度可以提高产量。

3) 在作物遗传育种中，育种目标的确定具有非常重要的意义，育种目标决定材料的选择和改良方向。由于气候环境、病虫害流行趋势、栽培管理水平不同，不同地区对育种目标的要求是各不相同的。在本研究中，初步得到了一些影响产量因子信息，并筛选到京218×京594，京试96-3，中玉11号，沧9005×冀815，150-4×H21等几个综合性状较为优良的组合，但为获得更准确目标环境对品种要求的信息，应增加测验品种并采用多年多点方式进行研究。

参考文献 (References)

- [1] 谭贤杰 (2011) 应用多元统计分析玉米农艺性状. *种子*, 5, 32-36.
- [2] 林琼 (2008) 福建省杂交早稻主要性状与产量的统计分析. *种子*, 4, 32-36.
- [3] 杨文鹏, 王伟 (2013) 多元统计分析等方法在 02 玉米自交系性状选育上的应用. *贵州农业科学*, 6, 1-4.
- [4] 薛毅, 陈立萍 (2007) 统计建模与 R 软件. 清华大学出版社, 北京.
- [5] 高慧璇 (2011) 应用多元统计分析. 北京大学出版社, 北京.

汉斯出版社为全球科研工作者搭建开放的网络学术中文交流平台。自2011年创办以来，汉斯一直保持着稳健快速发展。随着国内外知名高校学者的陆续加入，汉斯电子期刊已被450多所大中华地区高校图书馆的电子资源采用，并被中国知网全文收录，被学术界广为认同。

汉斯出版社是国内开源（Open Access）电子期刊模式的先行者，其创办的所有期刊全部开放阅读，即读者可以通过互联网免费获取期刊内容，在非商业性使用的前提下，读者不支付任何费用就可引用、复制、传播期刊的部分或全部内容。

