一种度量生物性状非线性相关性的广义相关系数

董晓萌[1],曹彬婕[2],罗凤娟[1],郭满才[1],袁志发[1]
(1 西北农林科技大学 理学院,陕西 杨凌 712100;2 中国农业大学 理学院,北京 100094)

[摘 要] 【目的】克服以往的相关性指标所存在的线性相关性和信息损失的问题,为实际育种工作提供一种新的多性状间的非线性相关分析方法。【方法】采用四阶矩法,提出一种用于描述变量间或向量间相关性程度的广义相关系数。【结果】将该广义相关系数应用于生物性状团的相关分析,可使性状团之间的相关程度介于 0 与 1。【结论】该广义相关系数计算简单,无信息损失,并能度量变量之间的非线性相关性。

[关键词] 广义相关系数;非线性相关;信息损失;性状团
[中图分类号] O212.4;S11+4 [文献标识码] A [文章编号] 1671-9387(2008)05-0191-05

Non-linear generalized correlation coefficient of biology trait

DONG Xiao-meng[1], CAO Bin-jie[2], LUO Feng-juan[1], GUO Man-cai[1], YUAN Zhi-fa[1]
(1 *College of Science, Northwest A & F University, Yangling, Shaanxi* 712100, *China*;
2 *College of Science, China Agricultural University, Beijing* 100094, *China*)

Abstract: 【Objective】 The study was to overcome the problem of linear correlation and information loss, and provide a new way of non-linear correlation analysis in different traits for breeding operator. 【Method】 A kind of new generalized correlation coefficient based on the four rank moments was built, which could describe the correlation of different variables and vectors. 【Result】 It was applied in the analysis of trait group, and got the correlation measurement between 0 and 1. 【Conclusion】 The calculation method was simple, without loss of information and its correlation was non-linear.

Key words: generalized correlation coefficient; non-linear correlation; information loss; trait group

性状的相关性度量是进行系统结构与功能分析的基础。但以往的研究如简单相关、典范相关[1]和广义相关等,均是建立在线性相关的基础上,它们反映了两个(组)变量之间线性相关的程度,但不能反映两个(组)变量之间非线性相关的程度,更不能反映在通常意义下两个(组)变量是否有关,因而这种度量是片面的。将张尧庭[2]、胡永宏[3]和 Nelsen[4]提出的广义相关系数,用于回归模型的检验可导出相应的统计量,但该广义相关系数只利用了其线性关联阵的特征根,造成了原有变异信息的损失。张尧庭[5]和 Kullback[6]提出的建立在申农信息熵基础上的广义相关系数,是对统计中变量之间相关性度量方法给出了说明和统一的方法,包括定量指标和定性指标,但该广义相关系数的缺点是其与变量的分布相关联,必须针对不同的联合密度来考虑,没有一个适用于各类分布的量。黄彩玉等[7]从数性积的几何意义出发,定义了一种多个随机变量样本统计相关性的度量指标。一般在研究多对多的相关中,

* [收稿日期] 2007-07-25
[基金项目] 国家自然科学基金项目(30571072)
[作者简介] 董晓萌(1982—),女,陕西渭南人,在读硕士,主要从事生物数学研究。E-mail:mathmandy@sohu.com
[通讯作者] 郭满才(1963—),男,陕西宝鸡人,教授,硕士生导师,主要从事生物数学研究。

多用典范相关分析来构成广义相关,而在典范相关分析中,需要计算其线性关联阵的特征根,计算过程比较复杂,并且计算出的特征根有多个,不同的人会得出不同的结果,无法统一。而且现实中多对多的相关未必均是线性相关的关系。本研究在前人研究基础上,提出了一种新的多对多的性状之间的广义相关系数,以期为实际育种工作提供一种新的多性状间的非线性相关分析方法。

1 一种新广义相关系数的定义

假设两个多维变量分别为:

$X=(x_1,x_2,\cdots,x_m)^T\sim N_m(\mu_x,\boldsymbol{\Sigma_x})$,

$Y=(y_1,y_2,\cdots,y_P)^T\sim N_P(\mu_y,\boldsymbol{\Sigma_y})$。

式中:$\mu_x=(\mu_{x_1},\mu_{x_2},\cdots,\mu_{x_m})^T$ 是 X 的均值向量,$\boldsymbol{\Sigma_x}$ 为 X 的离差阵,$\mu_y(\mu_{y_1},\mu_{y_2},\cdots,\mu_{y_P})^T$ 是 Y 的均值向量,$\boldsymbol{\Sigma_y}$ 为 Y 的离差阵,$\boldsymbol{\Sigma_x}$ 和 $\boldsymbol{\Sigma_y}$ 均是对称正定矩阵。

设 $\boldsymbol{\Sigma_{xy}}$ 是 X、Y 的协差阵。X、Y 的相关阵分别记为:$\boldsymbol{R}_X=(\rho_{Xij})_{m\times m}$,$\boldsymbol{R}_Y=(\rho_{Yij})_{p\times p}$。

令

$$Z=\begin{bmatrix}X\\Y\end{bmatrix}\sim N_{m+p}(u_Z,\boldsymbol{\Sigma_Z}),$$

式中:

$$\mu_Z=\begin{bmatrix}\mu_X\\\mu_Y\end{bmatrix},$$

$$\boldsymbol{\Sigma_Z}=\begin{bmatrix}\boldsymbol{\Sigma_X} & \boldsymbol{\Sigma_{XY}}\\\boldsymbol{\Sigma_{YX}} & \boldsymbol{\Sigma_Y}\end{bmatrix}_{(m+p)\times(m+p)}=(\sigma_{Zij})_{(m+p)\times(m+p)}。$$

将 X 与 Y 的线性组合分别记为:$X'=x_1+x_2+\cdots+x_m$;$Y'=y_1+y_2+\cdots+y_p$,则 X'的方差为:

$$V(X')=\sum_{i=1}^{m}\sum_{j=1}^{m}\sigma_{Zij},$$

Y'的方差为:

$$V(Y')=\sum_{i=m+1}^{m+p}\sum_{j=m+1}^{m+p}\sigma_{Zij},$$

X'与 Y'的协方差为:

$$COV(X',Y')=\sum_{i=1}^{m}\sum_{j=m+1}^{m+p}\sigma_{Zij}。$$

由上述分析可作如下定义:

定义 1:X 与 Y 的广义相关系数 $r_{(1)XY}$ 为:

$$r_{(1)XY}=\frac{COV(X',Y')}{\sqrt{V(X')V(Y')}}=\frac{\sum_{i=1}^{m}\sum_{j=m+1}^{m+p}\sigma_{Zij}}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}\sigma_{Zij}}\sqrt{\sum_{i=m+1}^{m+p}\sum_{j=m+1}^{m+p}\sigma_{Zij}}}。$$

显然,广义相关系数 $r_{(1)XY}$ 满足相关系数的所有性质。其本质上也是一种线性相关,即将原来的变量赋以权重 1,然后线性组合为两个综合变量,这两个综合变量间的简单相关系数就是两个多维变量间的广义相关系数。但该广义相关系数与典范相关系数又有区别,典范相关分析反映了两组随机变量之间的线性相关情况,不同典范变量对之间的典范相关强弱程度有差异。在实际中,往往着重研究的是相关关系较大的几对典范变量,这样就造成了原有变异信息的损失。

现实中的相关关系并非都是线性的,故需研究变量之间的非线性相关性。分析步骤如下:

第一步:由 $\boldsymbol{\Sigma_Z}=\begin{bmatrix}\boldsymbol{\Sigma_X} & \boldsymbol{\Sigma_{XY}}\\\boldsymbol{\Sigma_{YX}} & \boldsymbol{\Sigma_Y}\end{bmatrix}_{(m+p)\times(m+p)}\Rightarrow\boldsymbol{Z}$ 的相关阵 $\boldsymbol{R}_Z$,

$$\boldsymbol{R}_Z=\begin{bmatrix}\boldsymbol{R}_X & \boldsymbol{R}_{XY}\\\boldsymbol{R}_{YX} & \boldsymbol{R}_Y\end{bmatrix}_{(m+p)\times(m+p)}=(\rho_{Zij})_{(m+p)\times(m+p)}。$$

第二步:由 $\boldsymbol{Z}$ 的相关阵 $\boldsymbol{R_Z}$ 可得到 $\boldsymbol{Z}$ 的相关信息总量 S_{R_Z} 为[8]:

$$S_{R_Z}=S_{R_X}+S_{R_Y}+2S_{R_{XY}}=\sum_{i=1}^{m+p}\sum_{j=1}^{m+p}\rho_{Zij}^2。$$

式中:S_{R_X} 为 X 的相关信息总量,即 $S_{R_X}=\sum_{i=1}^{m}\sum_{j=1}^{m}\rho_{Zij}^2$;$S_{R_Y}$ 为 Y 的相关信息总量,即 $S_{R_y}=\sum_{i=m+1}^{m+p}\sum_{j=m+1}^{m+p}\rho_{Zij}^2$;$S_{R_{XY}}$ 为 X 与 Y 的相关信息总量,即:

$$S_{R_{XY}}=\sum_{i=1}^{m}\sum_{j=m+1}^{m+p}\rho_{Zij}^2=\sum_{i=m+1}^{m+p}\sum_{j=1}^{m}\rho_{Zij}^2。$$

实对称矩阵 $\boldsymbol{R}_Z$ 经过正交矩阵 $\boldsymbol{L}$ 的相似变换而得到的矩阵 $\boldsymbol{P}=L^{-1}R_ZL$ 仍是对称的,两个相似矩阵 $\boldsymbol{R}_Z$ 及 $\boldsymbol{P}$ 的各元素总平方和仍保持为同一常数,即总相关信息保持不变,且 $\boldsymbol{R}_Z$ 与 $\boldsymbol{P}$ 的对角线元素之和相等,即迹不变。

将 $\boldsymbol{Z}=\begin{bmatrix}\boldsymbol{X}\\\boldsymbol{Y}\end{bmatrix}$的协差阵 $\boldsymbol{\Sigma_Z}=(\sigma_{Zij})_{(m+p)\times(m+p)}$ 中的每个元素分别平方后得矩阵 $\boldsymbol{\Sigma_Z}^{(2)}=(\sigma_{Zij}{}^2)_{(m+p)\times(m+p)}$。由 X 和 Y 的四阶中心矩和其四阶混合矩组成的矩阵记为 $\boldsymbol{M}$,则 $\boldsymbol{M}=(\sigma_{Mij})_{(m+p)\times(m+p)}$,其中 $\sigma_{Mij}=E(X_i-\mu_i)^2(X_j-\mu_j)^2$,$(i,j=1,2,\cdots,m+p)$。即矩阵 $\boldsymbol{M}$ 中,对角线上的元素为各个变量的四阶中心矩,非对角线上的元素为对应两个变量之间的四阶混合矩。

可以证明 $\sigma_{Mij}=3\sigma_{Zij}{}^2(i,j=1,2,\cdots,m+p)$,即矩阵 $\boldsymbol{M}$ 中的元素与矩阵 $\sum_Z{}^{(2)}$ 中的对应元素均相差 3 倍。

同理,对于 X、Y 的相关阵分析所得结果是一样的,即将相关阵 $\boldsymbol{R_Z}$ 中的每个元素平方后得到矩阵 $\boldsymbol{R_Z}^{(2)}=(\rho_{Zij}{}^2)_{(m+p)(m+p)}$,该矩阵的每个元素与由标准化后的四阶中心距及其四阶混合矩组成矩阵中的

每个元素均相差 3 倍。

定义 2:假设变量 X 的四阶中心矩为 $(\mu_4)_i$[9],四阶混合矩为 $(\mu_4)_{ij}$(i、$j=1,2\cdots,m$;且 $i\neq j$),变量 Y 的四阶中心矩为 $(\nu_4)_i$,四阶混合矩为 $(\nu_4)_{ij}$,(i、$j=1,2,\cdots,p$;$i\neq j$),变量 X、Y 的四阶混合矩为 $(\mu\nu)_{4ij}$($i=1,2,\cdots,m$;$j=1,2,\cdots,p$),并假设:

$$0\leqslant\sum_{i=1j=1}^{m}\sum^{p}(\mu\nu)_{4ij}\leqslant\sqrt{\sum_{i=1j=1}^{m}\sum^{m}(\mu_4)_{ij}}\sqrt{\sum_{i=1j=1}^{p}\sum^{p}(\nu_4)_{ij}},$$

则称

$$r_{(2)XY}{}^2=\frac{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{p}(\mu\nu)_{4ij}}{\sqrt{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{m}(\mu_4)_{ij}}\sqrt{\frac{1}{3}\sum_{i=1j=1}^{p}\sum^{p}(\nu_4)_{ij}}}=\frac{\sum_{i=1j=1}^{m}\sum^{p}(\mu\nu)_{4ij}}{\sqrt{\sum_{i=1j=1}^{m}\sum^{m}(\mu_4)_{ij}}\sqrt{\sum_{i=1j=1}^{p}\sum^{p}(\nu_4)_{ij}}}$$

为变量 X、Y 的广义相关系数 $r_{(2)XY}$ 的平方。

式中:$(\mu_4)_{ij}=E(X_i-\mu_{x_i})^2(X_j-\mu_{x_j})^2$,$i$、$j=1,2,\cdots,m$;

$(\nu_4)_{ij}=E(Y_i-\mu_{y_i})^2(Y_j-\mu_{y_j})^2$,$i$、$j=1,2,\cdots,p$;

$(\mu\nu)_{4ij}=E(X_i-\mu_{x_i})^2(Y_j-\mu_{y_j})^2$,$i=1,2,\cdots,m$;$j=1,2,\cdots,p$。

因为变量的四阶中心矩及其四阶混合矩所组成矩阵中的每个元素,与相关阵的每个元素平方后所得的矩阵中每个元素相差 3 倍,故定义 2 中的式子与下式是等价的,即:

当 $0\leqslant S_{R_{XY}}\leqslant\sqrt{S_{R_X}}\sqrt{S_{R_Y}}$,

$$r_{(2)XY}{}^2=\frac{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{p}(\mu\nu)_{4ij}}{\sqrt{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{m}(\mu_4)_{ij}}\sqrt{\frac{1}{3}\sum_{i=1j=1}^{p}\sum^{p}(\nu_4)_{ij}}}\Leftrightarrow\frac{S_{R_Z}-S_{R_X}-S_{R_Y}}{2\sqrt{S_{R_X}}\sqrt{S_{R_Y}}}=\frac{S_{R_{XY}}}{\sqrt{S_{R_X}}\sqrt{S_{R_Y}}}。$$

显然,当 $m=p=1$ 时,X、Y 的相关系数是其简单相关系数,即:$r_{(2)XY}{}^2=\rho^2$。

2　广义相关系数的性质

对于任意的变量 X、Y 有如下性质:

① 对称性,即:$r_{(2)XY}=r_{(2)YX}$;

②相同变量之间的相关系数是 1,即:$r_{(2)XX}=r_{(2)YY}=1$;

③广义相关系数值介于 0 与 1,即:$0\leqslant r_{(2)XY}{}^2\leqslant 1$,其中当 X 与 Y 相互独立时,$r_{(2)XY}=0$;

④广义相关系数 $r_{(2)XY}$ 是非线性的。

证明:①因为

$$r_{(2)XY}{}^2=\frac{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{p}(\mu\nu)_{4ij}}{\sqrt{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{m}(\mu_4)_{ij}}\sqrt{\frac{1}{3}\sum_{i=1j=1}^{p}\sum^{p}(\nu_4)_{ij}}},$$

又因为

$$r_{(2)YX}{}^2=\frac{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{p}(\mu\nu)_{4ij}}{\sqrt{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{m}(\nu_4)_{ij}}\sqrt{\frac{1}{3}\sum_{i=1j=1}^{p}\sum^{p}(\mu_4)_{ij}}},$$

其中 $(\nu\mu)_{4ij}$ 是变量 X、Y 的四阶混合矩,且 $(\mu\nu)_{4ij}=(\nu\mu)_{4ij}$($i=1,2,\cdots,m$;$j=1,2,\cdots,p$)。

故 $r_{(2)XY}=r_{(2)YX}$。

②因为 $r_{(2)XX}{}^2=\frac{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{m}(\mu\mu)_{4ij}}{\sqrt{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{m}(\mu_4)_{ij}}\sqrt{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{m}(\mu_4)_{ij}}}=\frac{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{m}(\mu_4)_{ij}}{\sqrt{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{m}(\mu_4)_{ij}}\sqrt{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{m}(\mu_4)_{ij}}}$,即:$r_{(2)XX}{}^2=1$。

同理可证

$$r_{(2)YY}{}^2=\frac{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{m}(\nu_4)_{ij}}{\sqrt{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{m}(\nu_4)_{ij}}\sqrt{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{m}(\nu_4)_{ij}}}=1。$$

即:$r_{(2)XX}=r_{(2)YY}=1$。

③ 先假设当变量 X、Y 相互独立时,则 X 与 Y 之间的四阶混合矩之和为 0,

即:$$\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{p}(\mu\nu)_{4ij}=0。$$

故其广义相关系数为:$r_{(2)XY}{}^2=\frac{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{p}(\mu\nu)_{4ij}}{\sqrt{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{m}(\mu_4)_{ij}}\sqrt{\frac{1}{3}\sum_{i=1j=1}^{p}\sum^{p}(\nu_4)_{ij}}}=0$,

即:$r_{(2)XY}=0$。

当变量 X、Y 相互不独立时,则 X 与 Y 的广义相关系数 $r_{(2)XY}{}^2=\frac{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{p}(\mu\nu)_{4ij}}{\sqrt{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{m}(\mu_4)_{ij}}\sqrt{\frac{1}{3}\sum_{i=1j=1}^{p}\sum^{p}(\nu_4)_{ij}}}$。

因为 $0\leqslant\sum_{i=1}^{m}\sum_{j=1}^{p}(\mu\nu)_{4ij}\leqslant\sqrt{\sum_{i=1j=1}^{m}\sum^{m}(\mu_4)_{ij}}\times\sqrt{\sum_{i=1j=1}^{p}\sum^{p}(\nu_4)_{ij}}$,

则 $$0\leqslant\frac{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{p}(\mu\nu)_{4ij}}{\sqrt{\frac{1}{3}\sum_{i=1j=1}^{m}\sum^{m}(\mu_4)_{ij}}\sqrt{\frac{1}{3}\sum_{i=1j=1}^{p}\sum^{p}(\nu_4)_{ij}}}\leqslant 1,$$

从而 $0 \leqslant r_{(2)XY}{}^2 \leqslant 1$。

④ 因为该广义相关系数的本质是变量之间的四阶中心矩与其四阶混合矩之间的关系,又因为对于任意的变量 X_i,假设其均值为 μ_i,方差为 σ_i^2,则其四阶中心矩为:

$3\sigma_i^4 = E(X_i-\mu_i)^4 = E(X^4 - 4X^3\mu_i + 6X^2\mu_i{}^2 - 4X\mu_i^3 + \mu_i^4) = E(X^4) - 4\mu_i E(X^3) + 6\mu_i^2 E(X^2) - 4\mu_i^3 E(X) + \mu_i^4 = E(X^4) - 4\mu_i E(X^3) + 6\mu_i^2(\mu_i^2+\sigma_i^2) - 3\mu_i^4$。

可知四阶中心矩 σ_i^4 不仅与变量的一阶原点矩和二阶原点矩有关,而且还与变量的三阶原点矩和四阶原点矩有关。即广义相关系数 $r_{(2)XY}$ 也与变量的一阶原点矩、二阶原点矩、三阶原点矩和四阶原点矩有关,故广义相关系数是非线性的。

对任意的变量 X_j,假设其均值为 μ_j,方差为 σ_j^2,且 $Y=X_i+X_j$,Y 的均值为 μ_Y,则 Y 的四阶中心矩与 X_i 和 Y_j 的四阶中心矩及其四阶混合矩有如下关系:

$E(Y-\mu_y)^4 = E[(X_i+X_j)-(E(X_i)+E(X_j))]^4 = E[(X_i-\mu_i)+(X_j-\mu_j)]^4 = E(X_i-\mu_i)^4 + 4E(X_i-\mu_i)^3(X_j-\mu_j) + 6E(X_i-\mu_i)^2(X_j-\mu_j)^2 + 4E(X_i-\mu_i)(X_j-\mu_j)^3 + E(X_j-\mu_j)^4$。

可知任意两个变量之和的四阶中心矩,不仅与这两个变量的四阶中心矩有关,还与这两个变量之间的所有四阶混合矩有关,它们之间的关系如上式所示。证毕。

上述构建的广义相关系数,包括了所有的一对一、一对多和多对多的相关关系。故可以将其应用于生物性状中,对不同性状团进行相关性分析,即将性状团与性状团的相关程度用一个数来表示。

3 应用实例

本例数据来自参考文献[10],春播菜用大豆性状按某种生物学意义分成了 5 个性状团(X_1、X_2、X_3、X_4、X_5):X_1,生育期(出苗期、开花期、结荚期、鼓粒期、上市期);X_2,植株形态(株高、结荚高度、茎粗、茎节数、一次分枝数);X_3,产量性状(单株鲜荚数、单株鲜荚重、单株鲜粒数、单株鲜粒重);X_4,荚粒形态(荚长、荚宽、粒长、粒宽);X_5,品质性状(籽粒可溶性蛋白质含量、籽粒淀粉含量、籽粒可溶性糖含量、籽粒异黄酮含量)。

经计算所得春播菜用大豆 5 个性状团及不同两个性状团间的总相关信息量如表 1 所示。春播菜用大豆 5 个性状团的方差及不同的两个性状团之间的协方差如表 2 所示。由春播菜用大豆不同性状团的总相关信息值及其方差与协方差,可计算出各个性状团之间的广义相关系数值,见表 3。

表 1 春播菜用大豆不同性状团及其之间的总相关信息量

Table 1 Total correlation information between trait groups of spring vegetable soybean

性状团 Ttrait group	X_1	X_2	X_3	X_4	X_5
X_1	15.679	35.466	46.637	38.038	28.208
X_2	35.466	12.173	26.149	26.064	24.043
X_3	43.637	26.149	10.738	25.328	20.583
X_4	38.038	26.064	25.329	8.331	16.518
X_5	28.208	24.043	20.583	16.518	4.934

表 2 春播菜用大豆不同性状团的方差及其之间的协方差

Table 2 Variance and covariance between trait groups of spring vegetable soybean

性状团 Ttrait group	X_1	X_2	X_3	X_4	X_5
X_1	120.099	−26.559	−212.575	−0.290	0.204
X_2	−26.559	12.498	30.007	0.723	0.158
X_3	−212.575	30.007	452.835	3.768	−1.158
X_4	−0.290	0.723	3.768	0.083	−0.020
X_5	0.204	0.158	−1.158	−0.020	0.245

表 3 春播菜用大豆不同性状团之间的广义相关系数

Table 3 Generalized correlation coefficient between trait groups of spring vegetable soybean

性状团 Ttrait group	X_1	X_2	X_3	X_4	X_5
X_1	1	−0.686	−0.912	−0.001	0.038
X_2	0.525	1	0.399	0.709	0.090
X_3	0.815	0.376	1	0.614	−0.110
X_4	0.783	0.525	0.575	1	−0.002
X_5	0.657	0.670	0.581	0.504	1

表 3 中对角线右上面的元素是由定义 1 得到的相关系数,对角线左下面的元素是由定义 2 得到的广义相关系数。由定义 1 得出的结果为:在 5 个性状团中,生育期与产量性状的相关性达到最大的负相关;荚粒形态与植株形态的相关性达到最大的正相关,荚粒形态与产量性状的相关系数次之;生育期与品质性状的相关性为最小的正相关。由定义 2 得出的结果为:生育期与产量性状的相关性最强,与荚粒形态的相关性次之;品质性状与植株形态的相关性最强,与生育期的相关性次之;在 5 个性状团中,植株形态与产量性状的相关性最小。

典范相关分析的结果为:5 个性状团中,除了荚粒形态与品质性状的最小典范相关系数为 0.202 5 外,其他性状团之间的最小典范相关系数为 0.999 7～0.999 9,即它们之间的相关性很大[10]。

本文对性状团之间的最小典范相关系数进行区间估计,其 95%的置信区间如下所示:

r_{12}:[0.996 3,1.000 0],r_{13}:[0.997 6,1.000 0],
r_{14}:[0.996 3,1.000 0],r_{15}:[0.998 8,1.000 0],
r_{23}:[0.998 8,1.000 0],r_{24}:[0.997 6,1.000 0],
r_{25}:[0.997 6,1.000 0],r_{34}:[0.998 8,1.000 0],
r_{35}:[− 0.780 6,0.897 2],r_{45}:[0.996 3,1.000 0]。

由以上分析可知,由定义 2 得出的非线性相关系数中,除了产量性状与品质性状的非线性相关系数在其对应的最小典范相关系数的置信区间 r_{35}:[− 0.780 6,0.897 2]内,其他性状团之间的非线性相关系数均不在其对应的最小典范相关系数的置信区间内。故广义相关系数 $r_{(2)XY}$ 与典范相关系数的本质不同,因为典范相关分析是一种线性相关,而本研究定义 2 中的相关是一种非线性相关。

4　结　论

本研究从生物性状的总相关信息入手,利用四阶矩法定义了一种无信息损失、非线性的广义相关系数,且该广义相关系数计算简单,无信息损失,并能度量变量之间的非线性相关性。最后利用该广义相关系数对生物性状团进行相关性分析,使性状团之间的相关程度介于 0 与 1。但对于该广义相关系数的抽样分布及假设检验的问题还未涉及到,因而本研究方法可继续研究下去。

[参考文献]

[1] 袁志发,周静芋. 多元统计分析 [M]. 北京:科学出版社,2002:172-180,241-256.
Yuan Z F,Zhou J Y. Mutivariate statistics analysis [M]. Beijing:Science Press, 2002:172-180,241-256. (in Chinese)

[2] 张尧庭. 广义相关系数及其应用 [J]. 应用数学学报,1978(4):33-39.
Zhang Y T. Generalized correlation coefficient and its application [J]. Journal of Applied Mathematics,1978(4):33-39. (in Chinese)

[3] 胡永宏. 一种广义相关系数 [J]. 统计与信息论坛,1997(1):20-23.
Hu Y H. A kind of generalized correlation coefficient [J]. Statistics and Information Tribune,1997(1):20-23. (in Chinese)

[4] Nelsen N B. An introduction to copulas,lectures notes in statistics [M]. New York:Spring Verlag,1998:139.

[5] 张尧庭. 关于度量变量之间的相关程度 [J]. 上海财经大学学报,1999(2):60-63.
Zhang Y T. How to measure the correlation among random variables [J]. Journal of Shanghai University of Finance and Economics,1999(2):60-63. (in Chinese)

[6] Kullback S. Information theory and statistics [M]. [s. n.]: John Wiley & Sons Inc,1959.

[7] 黄彩玉,邱中华,唐加山. 多个随机变量样本统计相关性的另一种度量指标 [J]. 南京邮电学院学报,1999(1):87-91.
Huang C Y,Qiu Z H,Tang J S. Another measure indication of mutivariable related coefficient [J]. Journal of Nanjing Institute of Posts and Telecommunications,1999(1):87-91. (in Chinese)

[8] 刘垂玕. 作物数量性状的遗传相关信息及其可加性 [J]. 安徽农业科学,1981(S1):52-57.
Liu C Y. Genetic correlation information and additivity of crop quantity [J]. Journal of Anhui Agricultural Science,1981(S1):52-57. (in Chinese)

[9] 陈希孺. 概率论与数理统计 [M]. 合肥:中国科学技术大学出版社,2002:126-140.
Chen X R. Probability theory and mathematics statistics [M]. Hefei:China Science Technology University Press,2002:126-140. (in Chinese)

[10] 周以飞,周德银. 春播菜用大豆生育期,农艺性状和品质性状的典范相关分析 [J]. 福建农林大学学报,2005(3):11-17.
Zhou Y F,Zhou D Y. Canonical correlation analysis of growing period,agronomic character,yield and quality character in the spring vegetable soybean [J]. Journal of Fujian Agricultural and Forestry University,2005(3):11-17. (in Chinese)