DOI:10.19789/j.1004-9398.2015.02.002

基于R环境下的多维变量组合间的相关系数实现*

董晓萌
(渭南师范学院 数学与信息科学学院,陕西 渭南 714000)

摘 要

为了完善多维随机向量各分量不同组合间的广义相关系数理论,以期为实际应用工作提供一种新的相关分析方法. 本文借助于一元线性回归模型,采用矩法,提出一种用于描述多维随机向量各分量不同组合间的广义相关系数,并用R语言编写该函数包,便于应用工作者使用,且丰富了R语言中的函数包.

关键词: 多维随机向量,向量组合,广义相关系数,R语言.
中图分类号: O212.4; S11+4

变量间的相关性度量是进行系统结构与功能分析的基础. 霍特林(Hotelling)、张尧庭[1,3,4]、胡永宏[2]、刘建州等人所研究的广义相关系数均是建立在线性相关的基础之上的,它们反映了两个(组)变量之间线性相关的程度,不能反映两个(组)变量之间非线性的相关程度[5,6],更不能反映在通常意义下两个(组)变量是否有关[7],因而这种度量是片面的. 笔者在前人研究基础之上提出了多维随机向量各分量不同组合之间的广义相关系数[8],该广义相关系数只对多维随机向量各分量在"等权"之下的线性组合及平方和组合的相关性进行了度量,但对各分量不同方次组合的相关性度量并未给出. 本文就作者定义的各分量不同方次组合的广义相关系数的计算过程,使用R语言编写了该广义相关系数的函数包,便于应用工作者使用,并且丰富了多元统计学的知识和R语言中的函数包.

1 多维随机向量各分量不同组合间的相关系数

假设两个多维随机向量[9]分别为: $X=(x_1,x_2,\cdots,x_m)^T\sim N_m(\mu_x,\sum_x)$,

$Y=(y_1,y_2,\cdots,y_p)^T\sim N_p(\mu_y,\sum_y)$,其中$\mu_x=(\mu_{x_1},\mu_{x_2},\cdots,\mu_{x_m})^T$是$X$的均值向量,$\sum_x$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,$\sum_y$为$Y$的离差阵;$\sum_y$为$Y$的离差阵,$\sum_x$和$\sum_y$均是对称正定矩阵. 设$\sum_{XY}$是$X,Y$的协差阵. X,Y的相关阵分别记为: $R_X=(\rho_{Xij})_{m\times m}$,$R_Y=(\rho_{Yij})_{p\times p}$. 令$Z=\begin{bmatrix}X\\Y\end{bmatrix}\sim N_{m+p}(u_Z,\sum_Z)$,其中$\mu_Z=\begin{bmatrix}\mu_X\\\mu_Y\end{bmatrix}$,$\sum_Z=\begin{bmatrix}\sum_X & \sum_{XY}\\ \sum_{YX} & \sum_Y\end{bmatrix}_{(m+p)\times(m+p)}=(\sigma_{Zij})_{(m+p)\times(m+p)}$.

定义1: 令$X'=X_1+X_2+\cdots+X_m$,$Y'=Y_1+Y_2+\cdots+Y_p+Y_1^2+Y_2^2+\cdots+Y_p^2$,则$X'$与$Y'$的广义相关系数为:

$$\rho_{X'Y'}^{(1)}=\frac{\sum_{i=1}^{m}\sum_{j=m+1}^{m+p}\sigma_{Zij}}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}\sigma_{Zij}\sum_{i=m+1}^{m+p}\sum_{j=m+1}^{m+p}(\sigma_{Zij}+3\sigma_{Zij}^2)}}\quad(1)$$

这是因为: 正态分布的三阶中心距等于零,且X'的方差$V(X')=\sum_{i=1}^{m}\sum_{j=1}^{m}\sigma_{Zij}$,$Y'$的方差$V(Y')=\sum_{i=m+1}^{m+p}\sum_{j=m+1}^{m+p}(\sigma_{Zij}+3\sigma_{Zij}^2)$,$X'$与$Y'$的协方差为$COV(X',Y')=\sum_{i=1}^{m}\sum_{j=m+1}^{m+p}\sigma_{Zij}$.

定义2: 令$X'=\sum_{i=1}^{m}(X_i+X_i^2)$,$Y'=\sum_{j=1}^{p}(Y_j+$

收稿日期: 2014-01-15
* 国家统计局项目《基于多元统计分析的我国环保支出研究(2012LY051)》

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Y_j^2),则 X' 与 Y' 的广义相关系数为:

$$\rho_{X'Y'}^{(2)}=\frac{\sum_{i=1}^{m}\sum_{j=m+1}^{m+p}(\sigma_{Zij}+3\sigma_{Zij}^2)}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}(\sigma_{Zij}+3\sigma_{Zij}^2)\cdot\sum_{i=m+1}^{m+P}\sum_{j=m+1}^{m+p}(\sigma_{Zij}+3\sigma_{Zij}^2)}} \tag{2}$$

这是因为:在正态分布下,三阶中心矩等于 0,四阶混合中心矩等于 $3\sigma_{Zij}^2$,且 X' 的方差 $V(X')=\sum_{i=1}^{m}\sum_{j=1}^{m}(\sigma_{Zij}+3\sigma_{Zij}^2)$,$Y'$ 的方差 $V(Y')=\sum_{i=m+1}^{m+P}\sum_{j=m+1}^{m+p}(\sigma_{Zij}+3\sigma_{Zij}^2)$,$X'$ 与 Y' 的协方差为 $\mathrm{COV}(X',Y')=\sum_{i=1}^{m}\sum_{j=m+1}^{m+p}(\sigma_{Zij}+3\sigma_{Zij}^2)$.

用上述方法可以把 X' 和 Y' 构造成各 X 和 Y 的不同方次的组合,可定义 X' 和 Y' 之间的广义相关系数 $\rho_{X'Y'}$,在计算中应注意:在正态分布下,奇次中心矩皆等于 0,而偶次中心矩等于 $(k-1)!!\sigma_{ij}^{k/2}$,$k=2,4,6,\cdots$ 符号 $(k-1)!!$ 表示 1 到 $k-1$ 的一切奇数的乘积.

因为变量的四阶中心矩与其四阶混合矩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. 所以前面所定义的广义相关系数也可用其相关阵进行分析,得到与定义 1、定义 2 中等价的式子为(3)式和(4)式:

$$\rho_{X'Y'}^{(1)}=\frac{\sum_{i=1}^{m}\sum_{j=m+1}^{m+p}\sigma_{Zij}}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}\sigma_{Zij}\sum_{i=m+1}^{m+p}\sum_{j=m+1}^{m+p}(\sigma_{Zij}+3\sigma_{Zij}^2)}}\Leftrightarrow \frac{\sum_{i=1}^{m}\sum_{j=m+1}^{m+p}\rho_{Zij}}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}\rho_{Zij}\sum_{i=m+1}^{m+p}\sum_{j=m+1}^{m+p}(\rho_{Zij}+3\rho_{Zij}^2)}} \tag{3}$$

$$\rho_{XY'}^{(2)}=\frac{\sum_{i=1}^{m}\sum_{j=m+1}^{m+p}(\sigma_{Zij}+3\sigma_{Zij}^2)}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}(\sigma_{Zij}+3\sigma_{Zij}^2)\cdot\sum_{i=m+1}^{m+P}\sum_{j=m+1}^{m+p}(\sigma_{Zij}+3\sigma_{Zij}^2)}}\Leftrightarrow \frac{\sum_{i=1}^{m}\sum_{j=m+1}^{m+p}(\rho_{Zij}+3\rho_{Zij}^2)}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}(\rho_{Zij}+3\rho_{Zij}^2)\cdot\sum_{i=m+1}^{m+P}\sum_{j=m+1}^{m+p}(\rho_{Zij}+3\rho_{Zij}^2)}} \tag{4}$$

2 多维随机向量各分量不同组合间相关系数的 R 语言实现

将两个多维随机向量 X 和 Y 中的各分量不同组合之间的广义相关系数,用 R 语言[9,10]编写的函数程序包 generalized. correlation(X,Y)如下:

```
generalized. correlation < - function(X,Y)
 {
   if(is. matrix(X)! =TRUE) X < - as. matrix(X)
   if(is. matrix(Y)! =TRUE) Y < - as. matrix(Y)
     Z < - cbind(X,Y)
       #### 将矩阵 X 和 Y 按列组合成矩阵 Z
   if(is. matrix(Z)! =TRUE) Z < - as. matrix(Z)
   Rx < - cor(X)
       #### 计算变量 X 的相关阵,记: Rx
   Rx2 < - cor(X) * cor(X)
       #### 对变量 X 的相关阵中的各元素平方,记: Rx2
   Srx < - sum(Rx)
       #### 对矩阵 Rx2 中的元素求和,记: Srx
   Ry < - cor(Y); Ry2 < - cor(Y) * cor(Y)
   Rz < - cor(Z); Rz2 < - cor(Z) * cor(Z)
   nx < - nrow(Rx2); cx < - ncol(Rx2)
   nz < - nrow(Rz2); cz < - ncol(Rz2)
   SRxy < - 0
       ####  对矩阵 Rz2 中的第 1 到第 n 行中的
   for(i in 1: nx)
       ####第(cx + 1)列到 cz 列求和,记: Srz
     {  for(j in (nx + 1): cz)
          {  SRxy < - SRxy + Rz[i,j]  }
     }
   SRxy
       Sry2 < - sum(Ry2); Sry1 < - sum(Ry); Sry < - Sry1 + 3 * Sry2
       Rxy < - SRxy/sqrt(Srx * Sry)
       Rxy
       #### Rxy 即为变量 X 和变量 Y 之间的非线性广义相关系数(定义 1)
       }
generalized. correlation < - function(X,Y)
 {
   if(is. matrix(X)! =TRUE) X < - as. matrix(X)
   if(is. matrix(Y)! =TRUE) Y < - as. matrix(Y)
     Z < - cbind(X,Y)
       #### 将矩阵 X 和 Y 按列组合成矩阵 Z
   if(is. matrix(Z)! =TRUE) Z < - as. matrix(Z)
   Rx < - cor(X)
```


(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

```
    #### 计算变量 X 的相关阵,记: Rx
  Rx2 < - cor( X) * cor( X)
    #### 对变量 X 的相关阵中的各元素平方,
    记: Rx2
    #### 对矩阵 Rx2 中的元素求和,记: Srx
  Ry < - cor( Y) ; Ry2 < - cor( Y) * cor( Y)
  Rz < - cor( Z) ; Rz2 < - cor( Z) * cor( Z)
  nx < - nrow( Rx2) ; cx < - ncol( Rx2)
  nz < - nrow( Rz2) ; cz < - ncol( Rz2)
  SRxy1 < - 0
    ####  对矩阵 Rz2 中的第 1 到第 n 行中的第
  for( i in 1: nx)
    ####( cx + 1) 列到 cz 列求和,记: Srz
  {  for( j in ( nx + 1) : cz)
    {  SRxy1 < - SRxy1 + Rz [i,j]  }
  }
SRxy1
SRxy2 < - 0
    ####  对矩阵 Rz2 中的第 1 到第 n 行中的第
  for( i in 1: nx)
    ####( cx + 1) 列到 cz 列求和,记: Srz
    {  for( j in ( nx + 1) : cz)
      {  SRxy2 < - SRxy2 + 3* Rz2 [i,j]  }
    }
  SRxy2
  SRxy < - SRxy1 + SRxy2
  Srx2 < - sum( Rx2) ; Srx1 < - sum( Rx) ; Srx <
- Srx1 + 3* Srx2
  Sry2 < - sum( Ry2) ; Sry1 < - sum( Ry) ; Sry <
- Sry1 + 3* Sry2
  Rxy < - SRxy/sqrt( Srx* Sry)
  Rxy
    #### Rxy 即为变量 X 和变量 Y 之间的非线
性广义相关系数( 定义 2)
}
```

在 R 程序中,输入变量 X 和 Y 的原始数据值,再运行 generalized. correlation(X,Y) ,就可直接得到变量 X 和 Y 的非线性广义相关系数值.

3　结　论

本文就作者定义的各分量不同方次组合的广义相关系数的计算过程,使用 R 语言编写了该广义相关系数的函数包,便于应用工作者使用,并且丰富了多元统计学的知识和 R 语言中的函数包.

参　考　文　献

[1] 张尧庭. 广义相关系数及其应用 [J]. 应用数学学报,1978(4) : 33 - 39.
[2] 胡永宏. 一种广义相关系数 [J]. 统计与信息论坛,1997(1) : 20 - 23.
[3] 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.
[4] Nelsen N B. An Introduction to copulas ,lectures notes in statistics [M]. New York : Spring Verlag ,1998 : 139.
[5] Kullback S. Information theory and statistics [M]. [s. n.] : JohnWiley & Sons Inc, 1959.
[6] 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.
[7] 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.
[8] 陈希孺. 概率论与数理统计 [M]. 安徽: 科学出版社,2002: 126 - 140.
[9] 董晓萌,曹彬婕,罗凤娟等. 一种度量生物性状非线性相关性的广义相关系数 [J]. 西北农林科技大学学报: 自然科学版,2008(5) : 191 - 195.
[10] 汤银才. R 语言与统计分析 [M]. 北京: 高等教育出版社,2008: 27 - 67.
[11] 薛毅,陈立萍. 统计建模与 R 软件 [M]. 北京: 清华大学出版社,2006: 423 - 471.

(下转第 10 页)

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[4] 常庚哲,史济怀. 数学分析教程[M]. 北京:中国科学技术大学出版社,2012:8.

A Peculiar Proof on Fermat's Little Theorem

Deng Yong

(Department of Mathematics, Kashgar Teacher's College, Kashgar, Xinjiang 844006)

Abstract

We presented a peculiar method to prove the Fermat's little theorem without using the conception or property of Euclidean algorithm, group theory and binomial coefficients.

Key words: Fermat's little theorem, formal power series, infinite product, logarithmic derivative, inductive approach.

作者简介 邓勇(1967 -),男,四川遂宁人,喀什师范学院数学系,教授,主要从事高等代数教学研究.

(上接第 7 页)

Generalized Correlation Coefficient Among Multi-variable Groups and Its R-program with R-language

Dong Xiaomeng

(Department of Math and Information Science, Weinan Teachers University, Weinan, Shannxi 714000)

Abstract

The recearch was done in order to complete generalized correlation coefficient theory system among multi-variable groups and provide the non-linear correlation method for operator. A kind of new generalized correlation coefficient based on the four rank moments was built with linear regression model, which could describe the correlation of different multi-variable groups, give its function package with R-language, enrich the R-language's function package.

Key words: multi-dimentional random vector; vector groups, generalized correlation coefficient, R-language.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net