文章编号：1672-6987(2022)01-0111-09；DOI：10.16351/j.1672－6987.2022.01.016

经典相关系数及统计功效对比研究

刘 辉[1]，邵福波[2,3]，宫 响[1]*

(1.青岛科技大学 数理学院，山东 青岛 266061；2.北京交通大学 轨道交通控制与安全国家重点实验室，北京 100044；
3.中车工业研究院有限公司 技术部，北京 100070)

摘 要：本工作选取多种经典相关系数进行了对比研究，如Pearson相关系数、Spearman相关系数、距离相关系数、最大信息系数及HHG相关系数。具体地，在不同数据规模及噪声水平下，对线性、非线性单调、非单调、非函数等不同类型变量的相关性分别进行研究，得到各相关系数的统计功效。通过分析发现，Pearson相关系数、Spearman相关系数更适合衡量线性、非线性单调相关关系，最大信息系数则更适合衡量含有周期性的相关关系，HHG则更适合衡量非函数相关关系。本研究可为挖掘不同相关关系，提供相关系数选取依据。

关键词：相关关系；Pearson相关系数；Spearman相关系数；距离相关系数；最大信息系数；HHG；统计功效

中图分类号：O 221.5 **文献标志码：**A

引用格式：刘辉，邵福波，宫响.经典相关系数及统计功效对比研究[J]. 青岛科技大学学报(自然科学版)，2022，43(1)：111-119.

LIU Hui，SHAO Fubo，GONG Xiang. Comparison of classical correlation coefficients and statistical power[J]. Journal of Qingdao University of Science and Technology(Natural Science Edition)，2022，43(1)：111-119.

Comparison of Classical Correlation Coefficients and Statistical Power

LIU Hui[1]，SHAO Fubo[2,3]，GONG Xiang[1]

(1.College of Mathematics and Physics，Qingdao University of Science and Technology，Qingdao 266061，China；
2.State Key Laboratory of Rail Traffic Control and Safety，Beijing Jiaotong University，Beijing 100044，China；
3.Technical Department，CRRC Academy Co.，Ltd.，Beijing 100070，China)

Abstract：This paper makes a comparison of several classical correlation coefficients，such as Pearson product-moment correlation coefficient，Spearman correlation coefficient，Distance correlation coefficient，Maximum information coefficient and HHG. Practically，under different data scale and noise level，the association of linear，nonlinear and non-function variables is studied respectively，and the statistical power of each correlation coefficient is obtained. It is found that Pearson product-moment correlation coefficient and Spearman correlation coefficient is more suitable to measure the linear and nonlinear monotonic association，and the Maximum Information Coefficient is more suitable to measure the association with periodicity，HHG is a better measure of non-functional association. The research of this paper can provide the choice for mining different correlation coefficient.

收稿日期：2021-02-10

基金项目：国家自然科学基金-山东省联合基金项目(U1906215)；中国科学院海岸带环境过程与生态修复重点实验室(烟台海岸带研究所)开放基金项目(2020KFJJ04).

作者简介：刘 辉(1996—)，男，硕士研究生. *通信联系人.

Key words: correlation; Pearson correlation coefficient; Spearman correlation coefficient; distance correlation coefficient; maximum information coefficient; HHG; statistical power

随着互联网、物联网、云计算等信息技术的迅猛发展,信息技术与人类世界的各个方面相互交融,大数据时代应运而生。人类的数据采集能力不断提升,数据量每年增长约50%,呈爆炸式增长,对数据进行有效地分析与挖掘,将推动国家、企业乃至整个社会的高效、可持续发展[1]。大数据时代的一个重要的特点是数据量大、数据维数高,如何从海量的、高维的数据中快速发掘数据的相关关系是一个重要问题[2]。

数据间的关系可分为:确定性关系,即把特征或者属性用变量表示,变量之间存在一一对应的映射关系,该类关系为函数关系;不确定性关系,即一个变量取一定值时,另一个变量由于受到随机因素的影响,对应的值可能是几个,并且都是以不同的概率出现,该类关系为相关关系。现实生活中,变量之间的相关关系往往是非线性的,相关程度各有差异,如何度量这样关系的强弱是人们关注的问题。

相关系数是衡量变量间相关关系强弱的重要指标。这里的相关系数是总称,不按统计指标的名称区分线性、非线性及复相关系数等,文中提到的具体相关系数均采用特定名称。1888年,GALTON从人类遗传学中提出了“相关”的概念;1920年,PEARSON提出了沿用至今的Pearson相关系数[3]。至2000年前,相关系数研究进展较慢,主要适用于衡量两个变量间的线性或非线性单调相关关系,例如Spearman相关系数[4]、Kendall相关系数[5]、Hoeffding's D统计量[6]以及RÉNYI在1959年提出的最大相关系数[7]等。2000年之后,随着数据量的增长,维数的增多,相关系数的研究得到了快速发展,大量的相关系数的计算方法被提出,可适用于衡量更复杂的相关关系,例如2004年的基于互信息的相关系数[8]、2007年的距离相关系数[9]、2011年的最大信息系数[10]以及2013年的Heller-Heller-Gorfine(HHG)方法[11]等。

对于高维数据间的相关性,目前常用的衡量方法是距离相关系数和HHG方法,可度量任意维度上的相关系数。此外,由于高维数据可看作是一个样品含有多个属性,对具有高维特征的两个变量的相关性进行衡量就相当于对两大类样品间的相关性的衡量,因此也可采用遍历的方法分别计算。

本研究在总结相关系数计算方法的基础上,选取五种经典的主流相关系数:Pearson相关系数、Spearman相关系数、距离相关系数、最大信息系数和HHG方法,通过对比分析不同高度复杂的数据关系,给出了不同相关系数适用范围。

1 相关系数的定义与计算方法

1.1 相关系数类型

总体上,按计算方法,相关系数可以大致分为4类[12-13]。

1)秩统计量法,即计算两个变量中每个观测值的秩,对比两个变量秩统计量之间的共同变化趋势。Spearman相关系数是历史最悠久的、也是普遍应用的秩相关系数。1938年KENDALL引入协同的概念,提出了τ相关系数。1948年,HOEFFDING提出的D统计量,是通过计算变量的联合秩统计量与其各变量间边际秩统计量乘积的差异来衡量变量间是否独立,即经样本计算所得的统计量大于某一阈值,则拒绝两个随机变量是独立的假设,但是该检验方式不对总体分布进行假设,因此是有偏的。

2)基于距离与核方法,这种方法是Pearson相关系数的扩展,即仍然采用Pearson相关系数的计算方式,将其度量线性相关关系扩展到非线性相关关系。如,2005年GRETTON等[14]提出的希尔伯特-施密特独立性准则(HSIC)方法,在计算互协方差时引入核函数,通过计算协方差矩阵的特征值平方和来衡量相关性,选取不同的核函数效果会有些不同,但是能够保证$HSIC(X,Y)=0$时,X和Y是独立的。这一方法的一个重要进展是SZÉKELY等[15]分别于2007年和2009年通过定义新型方差计算方法,提出了距离相关系数。

3)分箱网格方法,即通过将X和Y离散划分为多个区域,在每个区域内应用经典统计方法或信息论方法。2004年,KRASKOV等[8]提出基于K-近邻距离算法划分网格的熵估计,使得互信息具有自适应性和最小偏差;RESHEF等[16]在2011年、2015年提出最大信息系数,是通过对双变量的散点图进行最优分区,并取最大的信息熵作为相关系数;2013年,SUGIYAMA等[17]提出利用互信息维数衡量随机变量间的相关性,这种方法可以看作是对最

大信息系数的扩展;同年,HELLER 等通过对数据进行分区,形成多个 2X2 列联表,引入置换检验,以提高相关关系衡量能力;2014 年,WANG 等[18]通过计算局部相等的秩统计量来挖掘双变量间的相关关系;2016 年,ZHANG[12] 将相关性与 Hadamard 变换相结合,提出了二元扩展统计量和二元扩展检验来衡量变量间的相关性;2017 年,WANG 等[19]提出广义 R^2,这是对使用距离和划分网格方法的折中;2018 年,ROMANO 等[13]提出随机信息系数,是通过随机网格估计信息熵。

4)K-样本检验方法,用于检验样本是来源于某个分布,同时,也可以应用到相关性检验。2012 年,GRETTON 等[20]基于最大平均差异提出了核两样本检验;2015 年,JIANG 等[21]提出最优离散化的非参数 K-样本检验;2016 年,HELLER 等[22]基于互信息理论提出的一致无分布 K-样本检验。

秩统计量法以及基于距离与核的方法,具有明确的理论推导式,经常用于独立成分分析中,提取独立变量成分;分箱网格方法,能更直观通过对散点图划分网格呈现两个变量间的相关性,但是网格的划分方式、划分数量都会影响到计算方式的时间复杂度;K-样本检验方法,通过检验变量间的分布是否相等来确相关性,更适用于检验分类型变量和连续型变量之间的相关性[23-26]。

1.2　经典相关系数计算方法与检验

1.2.1　Pearson 相关系数

Pearson 相关系数是最经典的线性相关系数,也是应用最广泛的相关系数。其计算方式是将协方差除以标准差,剔除了两个变量量纲的影响,缩小到了 0 到 1 之间,就得到了 Pearson 相关系数(式 1),可以将其理解为标准化后的特殊协方差。

$$\rho_P=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}}。\quad(1)$$

对 Pearson 相关系数进行显著性检验,

$H_0:\rho=0$, $H_1:\rho\neq0$。

检验统计量为:

$$t_P=|\rho_P|\sqrt{\frac{n-2}{1-\rho_P^2}}\sim t(n-2)。\quad(2)$$

在给定的显著性水平 α 下,若拒绝原假设,则可认为总体的两个变量存在线性相关关系,其中 $|\rho_P|$ 越接近 1,线性相关性越强。

1.2.2　Spearman 相关系数

Spearman 相关系数可看作是 Pearson 相关系数衍生出的一种度量方法,该方法基于秩的理论,不需要假设变量之间是线性关系,也不是对原始数据直接进行计算,而是将原始数据的秩作为变量,计算 Spearman 相关系数。常用于推荐系统、经济分析、公共管理、生物医疗等领域。

假设两个随机变量分别为 X、Y(也可以看做两个集合),它们的元素个数均为 n,两个随机变量取的第 $i(1\leqslant i\leqslant n)$个值分别用 X_i、Y_i 表示。对 X、Y 中的元素进行排序,得到两个元素排序后集合 x、y,将排序后集合 x、y 中的元素对应相减得到一个排序差分集合 d。已知样本数据,Spearman 相关系数的计算方式:

$$\rho_S=1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}。\quad(3)$$

其中 $d_i=x_i-y_i$,$1\leqslant i\leqslant n$,元素 x_i、y_i 分别为 X_i 在 X 中的排序以及 Y_i 在 Y 中的排序。

Spearman 相关系数的显著性检验与 Pearson 相关系数类似,在原假设成立的条件下检验统计量为 t_s 近似服从自由度为 $n-2$ 的 t 分布:

$$t_s=|\rho_S|\sqrt{\frac{n-2}{1-\rho_S^2}}\sim t(n-2)。\quad(4)$$

在给定的显著性水平下,若拒绝原假设,则可认为总体的两个变量之间存在相关关系,Spearman 相关系数 $|\rho_S|$ 越接近 1,两个变量间的相关性越强。

1.2.3　距离相关系数

距离相关,顾名思义,是基于范数(距离的度量方式之一)的理论提出的,又类似于积矩协方差和相关系数,是对经典的双变量相关性度量方法进行的推广和扩展,在很大程度上克服了 Pearson 相关系数不能度量非线性关系的弱点,常用于机器学习、特征工程等领域。该方法从随机变量的特征函数出发,定义了一个新的类似于加权 $2-L$ 的范数,则两个随机变量 X、Y 的协方差称为距离协方差,记为 $\mathrm{dcov}(X,Y)$,距离标准差分别为 $\mathrm{dcov}(X)$、$\mathrm{dcov}(Y)$。其距离相关系数 $\mathrm{dcor}(X,Y)$是对距离协方差 $\mathrm{dcov}(X,Y)$的标准化。

在样本数据中,分别计算 X、Y 的欧几里得距离矩阵,记为 $a_{k,l}=\|x_k-x_l\|$,$b_{k,l}=\|y_k-y_l\|$,其中 $k,l=1,2,\cdots,n$;并记 $\bar{a}_{k\cdot}$ 为距离矩阵 $a_{k,l}$ 的第 k 行平均;记 $\bar{a}_{\cdot l}$ 为距离矩阵 $a_{k,l}$ 的第 l 列平均;记 $\bar{a}$

为距离矩阵$a_{k,l}$的全平均;同理,可得$\bar{b}_{k\cdot}$、$\bar{b}_{\cdot l}$以及$\bar{b}$。

通过上述定义,利用样本数据计算得到的距离相关系数为

$$\mathrm{dcor}(X,Y)=\frac{\upsilon(X,Y)}{\sqrt{\upsilon(X,X)\cdot\upsilon(Y,Y)}}。\tag{5}$$

其中,$A_{k,l}=a_{k,l}-\bar{a}_{k\cdot}-\bar{a}_{\cdot l}+\bar{a}$,$B_{k,l}=b_{k,l}-\bar{b}_{k\cdot}-\bar{b}_{\cdot l}+\bar{b}$,样本的距离协方差为$\upsilon(X,Y)=\sqrt{\frac{1}{n^2}\sum_{k,l=1}^{n}A_{k,l}B_{k,l}}$。

距离相关系数的取值范围为0~1,当距离相关系数等于1时,两个随机变量间存在完全相关关系;当距离相关系数为0时,两个随机变量间不存在相关关系,即相互独立。

使用距离相关系数对两个随机变量进行相关检验,检验统计量为$\upsilon(X,Y)$,使用置换检验来计算在原假设成立的条件下的P值。

利用距离相关系数对两个随机变量X、Y间的独立性检验所提出的假设为

$H_0:F_{XY}(x,y)=F_X(x)\cdot F_Y(y)$,

$H_1:F_{XY}(x,y)\neq F_X(x)\cdot F_Y(y)$。

对随机变量X、Y之间的相关关系进行检验,置换检验过程如下:

1)在原假设的条件下,构造排列后的数据集$(x_1,y_1^*),(x_2,y_2^*),\cdots,(x_n,y_n^*)$,其中$x_i$是原数据列,$y_i^*$是对$y_i$随机排列后的数据列;

2)对排列后的数据集(x_i,y_i^*),计算其检验统计量$\upsilon(X,Y)$;

3)重复步骤1、步骤2多次(例如999次),分别计算出每次排列后的检验统计量。

置换检验的P值为:重复多次计算得出的检验统计量$\upsilon(x,y^*)$中大于等于原始数据的检验统计量$\upsilon(x,y)$的个数与重复次数的比值。

1.2.4 最大信息系数

最大信息系数(maximal information coefficient,MIC)于2011年提出,是用于检测变量之间非线性相关性的最新方法。其思想为:如果两个随机变量之间存在某种关系,那么可以在两个随机变量的散点图上划分出多个网格,对数据进行分区以封装这种关系。因此,最大信息系数计算的关键有两个方面:1)网格划分的数目,即在给定数据的散点图上要划分成多少个分区;2)网格划分的位置,即若在X轴上划分a次,那么这a次划分点是如何设置在x轴上的。最大信息系数常用于生物信息、医学等领域。

若已设定划分网格数和划分间隔点,则给定了一种划分,计算该划分方式下的信息熵为

$$I(D,a,b)=\sum_{a,b}f(x,y)\lg\frac{f(x,y)}{f(x)f(y)}。\tag{6}$$

其中,D为给定的数据集;a、b是对这个数据集的划分;$f(x,y)$是该区域内的联合概率密度,$f(x)$、$f(y)$分别为边际概率密度。

若确定了划分网格的数目,则通过改变网格的划分间隔点的位置,就会得到不同的信息熵,记其中最大的信息熵为$\max I(D,a,b)$。为了方便在不同维数之间进行比较,将其标准化,使其取值范围设置在0到1之间。那么,最大信息系数定义为

$$MIC(D)=\max_{ab<n^{0.6}}\left\{\frac{\max I(D,a,b)}{\lg(\min\{a,b\})}\right\}。\tag{7}$$

对两个随机变量进行的独立性检验,提出假设:

$H_0:MIC=0$,

$H_1:MIC\neq0$。

最大信息系数的检验统计量为$MIC(D)$,其置换检验与上文中提到的距离相关系数的置换检验是相同的。

1.2.5 HHG

HELLER等[22]提出了一个新的相关关系检验方法,该方法基于秩的理论,依据距离的大小对原始数据进行分区,从而形成多个2×2列联表,再进行置换检验以确定数据间的相关关系。对于样本数据,首先分别计算样本内各个个体间的距离$d(x_i,x_j)$,$d(y_i,y_j)$,其中$i,j\in\{1,2,\cdots,n\}$。假设随机变量X、Y是独立的并且存在连续的联合密度函数,那么在样本(X,Y)空间中存在一个点(x_i,y_i),分别在该点周围有个半径为r的空间,如果数据间存在相关关系,那么在该空间的界限处X、Y的联合分布是不等于边际分布的笛卡尔积。HHG常用于遗传学等领域。

相关关系显著性检验过程如下,定义:

$$A_{11}(i,j)=\sum_{k=1,k\neq i,j}^{n}I\{d(x_i,x_k)\leqslant d(x_i,x_j)\}\cdot I\{d(y_i,y_k)\leqslant d(y_i,y_j)\},\tag{8}$$

$$A_{12}(i,j)=\sum_{k=1,k\neq i,j}^{n}I\{d(x_i,x_k)\leqslant d(x_i,x_j)\}\cdot I\{d(y_i,y_k)>d(y_i,y_j)\},\tag{9}$$

$$A_{21}(i,j)=\sum_{k=1,k\neq i,j}^{n} I\{d(x_i,x_k)>d(x_i,x_j)\}\cdot I\{d(y_i,y_k)\leqslant d(y_i,y_j)\}, \tag{10}$$

$$A_{22}(i,j)=\sum_{k=1,k\neq i,j}^{n} I\{d(x_i,x_k)>d(x_i,x_j)\}\cdot I\{d(y_i,y_k)>d(y_i,y_j)\}。 \tag{11}$$

其中,$I\{\cdot\}$为示性函数。

$$S(i,j)=\frac{(n-2)\{A_{12}(i,j)A_{21}(i,j)-A_{11}(i,j)A_{22}(i,j)\}^2}{A_{1\cdot}(i,j)A_{2\cdot}(i,j)A_{\cdot 1}(i,j)A_{\cdot 2}(i,j)}。 \tag{12}$$

其中,$A_{m\cdot}(i,j)=A_{m1}(i,j)+A_{m2}(i,j)$,$A_{\cdot m}(i,j)=A_{1m}(i,j)+A_{2m}(i,j)$,$m=1,2$,$n$ 为样本数量。

为检验随机变量 X、Y 之间的相关性,提出假设:

$$H_0:F_{XY}(x,y)=F_X(x)\cdot F_Y(y),$$

$$H_1:F_{XY}(x,y)\neq F_X(x)\cdot F_Y(y)。$$

其中,F 为随机变量的分布函数。

检验统计量为

$$T=\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} S(i,j)。 \tag{13}$$

对两个随机变量进行的独立性检验,HHG 的置换检验与上文中提到的距离相关法的置换检验是相同的。HHG 可以采用列联表 φ 相关系数衡量变量间的相关程度:

$$\varphi=\sqrt{\frac{\max S(i,j)}{n-2}}。 \tag{14}$$

2　统计功效分析

2.1　统计功效

统计功效(statistical power)是指在假设检验的问题中,当原假设错误时,拒绝原假设的概率。其计算公式为

$$power=P(\text{reject } H_0\,|\,\text{False}(H_0))=1-\beta, \tag{15}$$

其中,$\text{False}(H_0)$表示原假设是错误的,β 表示第二类错误。

统计功效是检验某项实验有效性的一个很有用的指标,功效越大,说明犯第二型错误的概率越小。在实际研究工作中,功效值越大说明拒绝零假设越有利,研究结果也越可靠。统计功效的设定一般为 0.8,将它作为计算的阈值。当假设检验中的 P 值小于 0.05 且功效大于 0.8 时认为是有显著差异的。

2.2　统计功效的蒙特卡洛模拟

蒙特卡洛模拟,又称为统计模拟方法,是一类随机方法的统称。这类方法的特点是,可以在随机采样上计算得到近似结果,随着采样的次数增多,得到的结果是正确结果的概率逐渐加大,最终会收敛于实际值。本工作利用蒙特卡洛模拟计算统计功效,是通过大量模拟次数中,原假设发生的概率小于给定值(如 0.01,0.05)的次数占比。

比较不同相关系数的衡量能力,本工作选取了不同的样本量(10、20、30、50、100、200、500)、数据类型(线性、非线性单调、非单调、非函数)及噪声水平等情景,比较不同相关系数的衡量能力。按照表 1 所示的数学表达式随机生成模拟数据,图 1 展示本文所选取数据类型的散点图。

表 1　模拟数据数学表示

Table 1　Mathematical representation of simulated data

相关关系	数学表达式
Linear	$y=x$
Parabolic	$y=4\cdot(x-1/2)^2$
Cubic	$y=128\cdot(x-1/3)^3-48\cdot(x-1/3)^2-12\cdot(x-1/3)+2$
Exp	$y=e^{5x}$
Fourth power	$y=4\cdot(x^2-1/2)^2$
Cos	$y=\cos(12x)$
Linear/Periodic	$y=\sin(10\pi\cdot x)+x$
Sin (Fourier Frequency)	$y=\sin(16\pi\cdot x)$
Sin (non-Fourier Frequency)	$y=\sin(13\pi\cdot x)$
Sin (Varying Frequency)	$y=\sin(7\pi\cdot x\cdot(1+x))$
Hyperbola	$y=\sqrt{1+x^2}$
Circle	$y=\sqrt{1-x^2}$

图 1(a)表示两个变量之间存在线性单调相关关系,图 1(e)表示两个变量之间存在非线性单调相关关系,图 1(b)、(c)、(d)、(f)、(g)、(h)、(i)、(j)表示两个变量之间存在非单调相关关系,图 1(k)、(l)表示两个变量之间存在非函数关系。对每个相关关系在相同的噪声水平下,选取的样本量为 10、20、30、50、100、200、300、500,通过蒙特卡洛模拟,计算得出 5 个相关系数的统计功效,结果如图 2 所示。

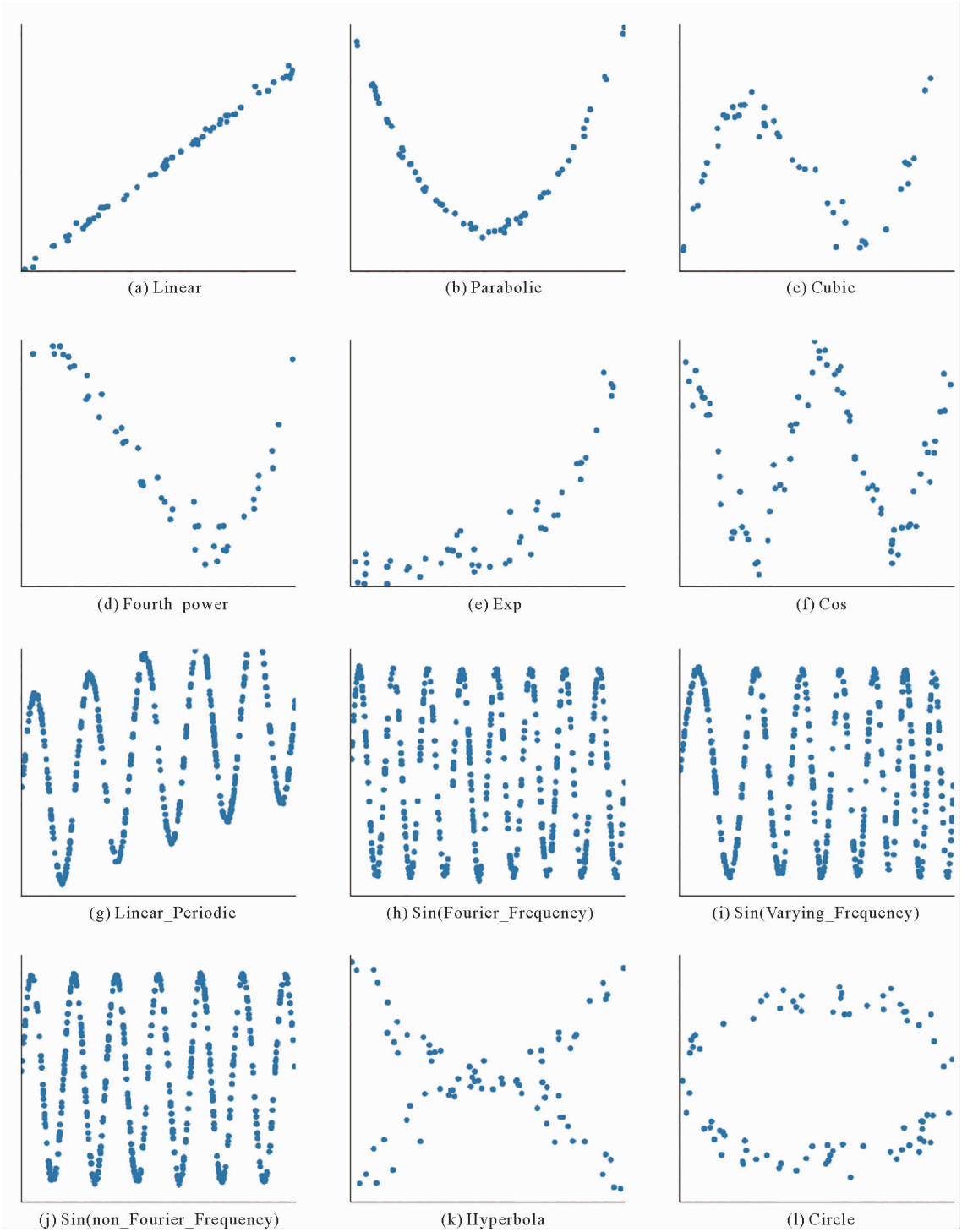

图1 基于蒙特卡洛方法随机生成的不同相关关系数据

Fig.1 Generated data with different correlations randomly based on Monte Carlo method

如图2所示,5种相关系数度量方法在具有线性相关关系数据下的统计功效都为1,其中最大信息系数在样本量为10时,其统计功效较其他方法低,但仍然高于0.8;具有非线性单调相关关系的数据,5种相关系数度量方法的统计功效也为1;对于非单调关系,如图2(c)、(d)、(g)、(j),Pearson相关系数或Spearman相关系数随着样本量的递增,其统计功效也大于0.8,距离相关系数、最大信息系数和HHG,在大样本情况下,可以度量出本研究中所提到的所有非单调相关关系以及非函数相关关系,对于小样本情况,如果数据中不存在明显的周期性,HHG的统计功效高于其他方法。

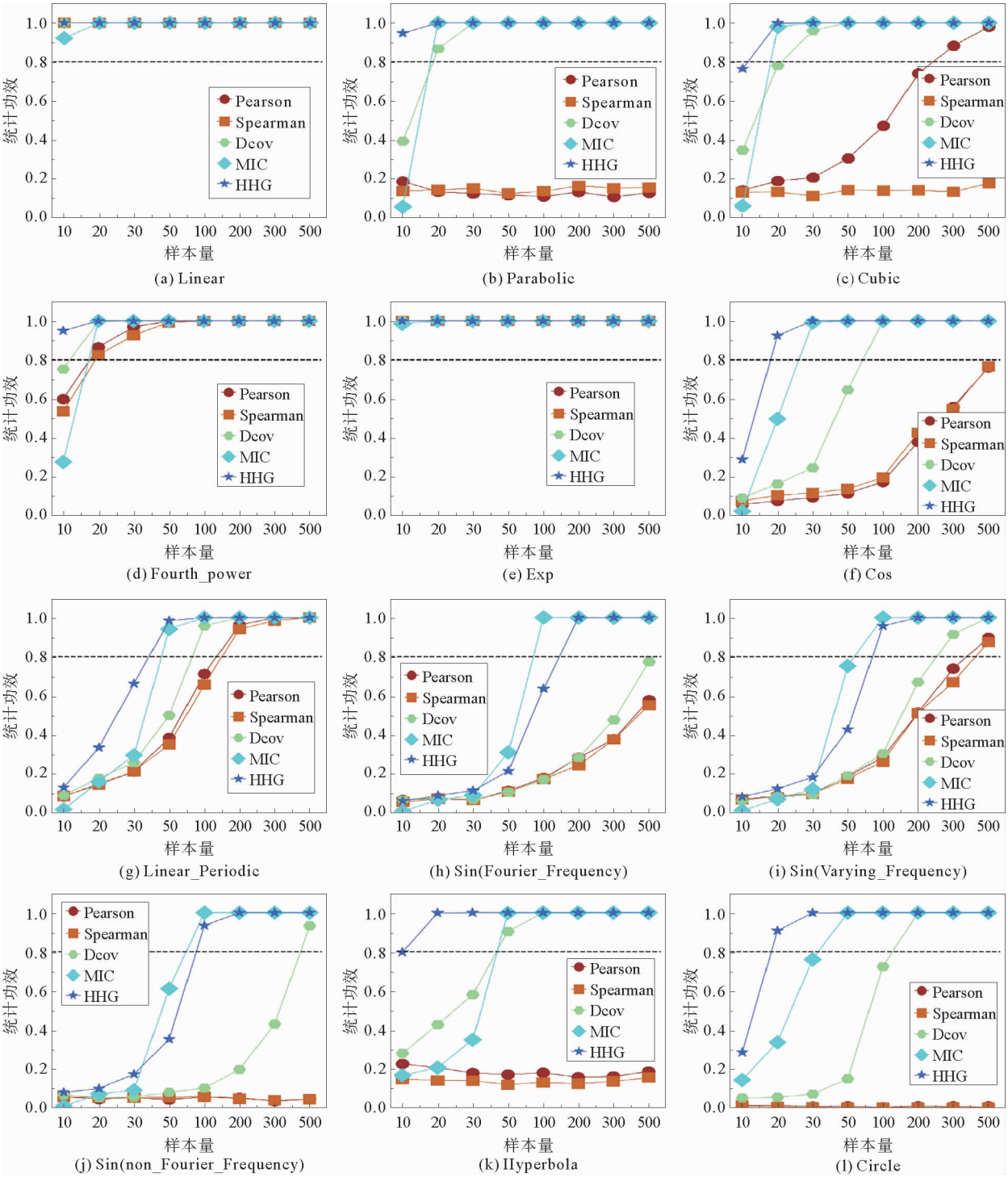

图 2　不同样本量下的统计功效

Fig.2　Statistical power of different sample sizes

在相同的样本量,不同的噪声水平下,如图 3 所示,5 种相关系数的统计功效与噪声水平呈反比;在线性相关关系和非线性单调相关关系中,Pearson 相关系数、Spearman 相关系数和距离相关系数统计功效优于最大信息系数和 HHG 的统计功效;对于非单调相关关系,当数据中存在明显的周期性时,最大信息系数的统计功效最高,HHG 的统计功效次之,当数据中不存在周期性时,HHG 的统计功效高于其他相关系数的统计功效;对于非函数相关关系,HHG 的统计功效最高。

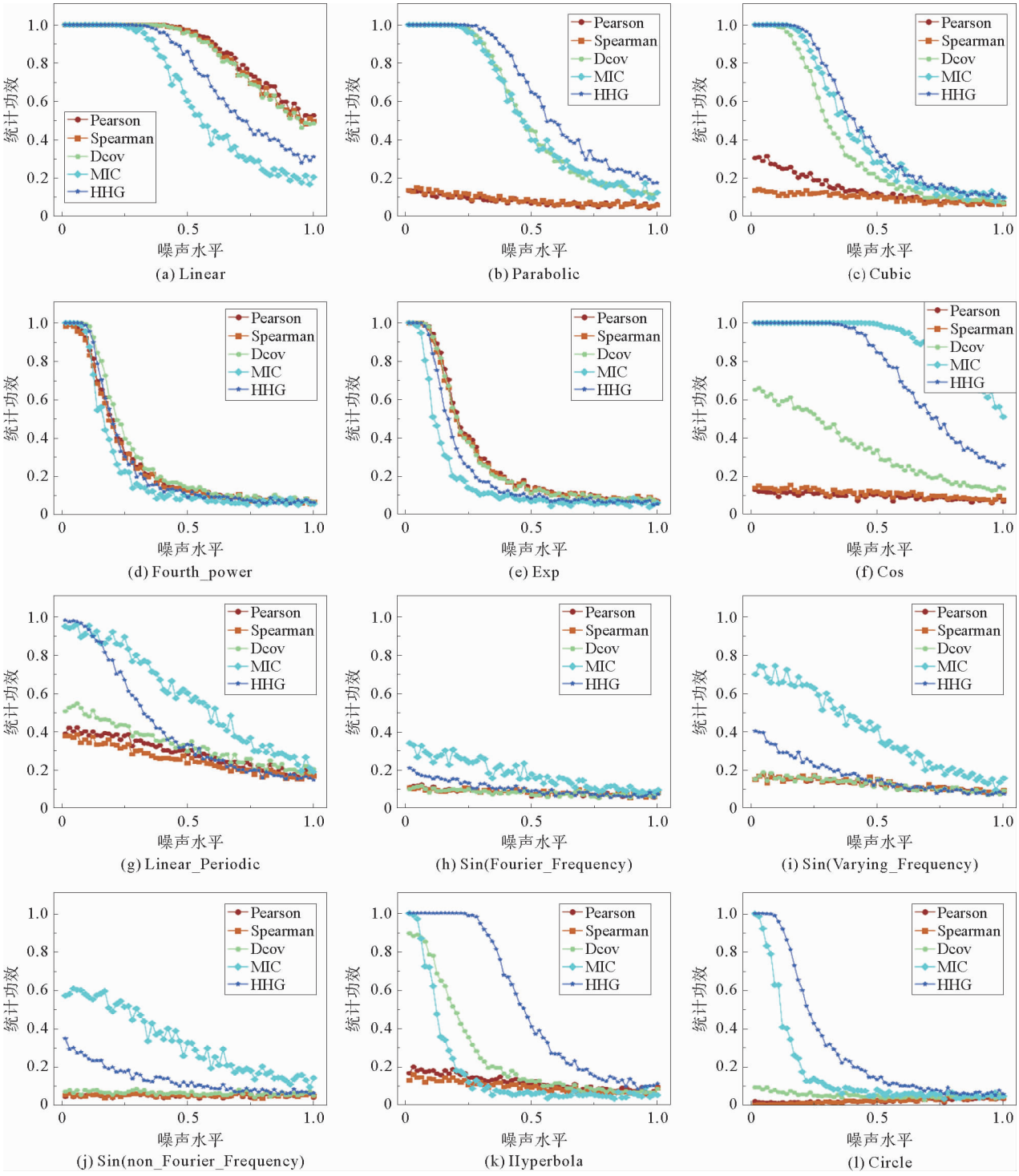

图3 不同噪声水平下的统计功效

Fig.3 Statistical power at different noise levels

由图4所示,可以根据想要挖掘的相关关系选取不同相关系数。当数据量小于50时,使用Pearson相关系数和Spearman相关系数挖掘单调相关关系,使用HHG方法挖掘非单调相关关系;当数据量大于50时,还是使用Pearson相关系数和Spearman相关系数挖掘单调相关关系,使用HHG方法挖掘非单调相关关系,使用最大信息系数挖掘周期性相关关系。由第二节中相关系数的计算方法可知,HHG方法需要提前计算出数据之间的距离,因此当数据量过于庞大时,其计算过程有较高的空间复杂度,同时,HHG方法的检验统计量是通过对数据的全局计算得到的,其时间复杂度也相对较高。在选取不同的相关系数时,也需要将时间复杂度与空间复杂度考虑在内。

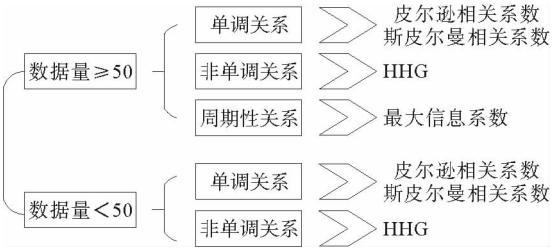

图 4　基于不同数据规模和相关关系的相关数选取树

Fig.4　Correlation coefficient selection tree based on different data sizes and correlations

3　结　语

对比不同度量高度复杂的数据关系的方法,并通过蒙特卡洛模拟得到不同相关系数的统计功效,对不同类型数据关系度量方法的使用做出引导。Pearson 相关系数和 Spearman 相关系数更适合衡量线性、非线性单调相关关系,最大信息系数则更适合衡量含有周期性的相关关系,HHG 方法则更适合衡量非函数相关关系。该研究可为挖掘不同相关关系,提供相关系数选取依据。该工作主要研究的是数值型变量间的相关关系,并未对分类型变量间的相关系数,如 φ 相关系数、V 相关系数、γ 相关系数、λ 相关系数等,进行对比总结。

参　考　文　献

[1] VIKTOR M.大数据时代:生活、工作与思维的大变革[M]. 杭州:浙江人民出版社,2012:20-29.
VIKTOR M. Big Data: A Revolution That Will Transform How We Live, Work and Think [M]. Hangzhou: Zhejiang People's Publishing House, 2012: 20-29.

[2] 梁吉业,冯晨娇,宋鹏.大数据相关分析综述[J].计算机学报:2016,39(1):1-18.
LIANG Jiye, FENG Chenjiao, SONG Peng. A survey on correlation analysis of big data [J]. Chinese Journal of Computers, 2016, 39(1): 1-18.

[3] PEARSON K. Notes on the history of correlation [J]. Biometrika, 1920, 13(1): 25-45.

[4] SPEARMAN C. General intelligence, objectively determined and measured [J]. The American Journal of Psychology, 1904, 15(2): 201-292.

[5] KENDALL M. A new measure of rank correlation [J]. Biometrika, 1938, 30(1): 81-93.

[6] HOEFFDING W. A non-parametric test of independence [J]. Annals of the Institute of Statistical, 1948, 19(4): 546-557.

[7] RENYI A. On measures of dependence [J]. Acta Mathematica Acadamiae Scientiarum Hungaricae, 1959, 10(3): 441-451.

[8] KRASKOV A, STOGBAUER H, GRASSBERGER P. Estimating mutual information [J]. Physical Review E, 2004, 69(6): 066138.

[9] SZEKELY G, RIZZO M, BAKIROV N. Measuring and testing dependence by correlation of distances [J]. Annals of Statistics, 2007, 35(6): 2769-2794.

[10] RESHEF D, RESHEF F, FINUCANE H, et al. Detecting novel associations in large data sets [J]. Science, 2011, 334 (6062): 1518-1524.

[11] HELLER R, HELLER Y, GORFINE M. A consistent multivariate test of association based on ranks of distances [J]. Biometrika, 2013, 100(2): 503-510.

[12] ZHANG K. BET on independence [J]. Journal of the American Statistical Association, 2019, 114(528): 1620-1673.

[13] ROMANO S, VINH N, VERSPOOR K, et al. The randomized information coefficient: assessing dependencies in noisy data [J]. Machine Learning, 2018, 107(3): 509-549.

[14] GREETTON A, BOUSQUET O, SMOLA A. et al. Measuring statistical dependence with Hilbert-Schmidt norms [C]// Algorithmic Learning Theory Proceedings of 16th International Conference, Singapore, 2005: 63-77.

[15] SZEKELY G, RIZZO M. Brownian distance covariance [J]. The Annals of Applied Statistics, 2009, 3(4): 1236-1265.

[16] RESHEF Y, RESHEF D, FINUCANE H, et al. Measuring dependence powerfully and equitably [J]. Journal of Machine Learning Research, 2016, 17(211): 1-63.

[17] SUGIYAMA M, BORGWARDT K. Measuring statistical dependence via the mutual information dimension[C]// Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, 2013: 1692-1698.

[18] WANG Y, WATERMAN M, HUANG H. Gene coexpression measures in large heterogeneous samples using count statistics [J]. Proceedings of the National Academy of Sciences of the United States of America, 2014, 111(46): 16371-16376.

[19] WANG X, JIANG B, LIU J. Generalized R-squared for detecting dependence [J]. Biometrika, 2017, 104(1): 129-139.

[20] GRETTON A, BORGWARDT K, RASCH M, et al. A kernel two-sample test [J]. Journal of Machine Learning Research, 2012, 13(25): 723-773.

[21] JIANG B, YE C, LIU J. Nonparametric K-sample tests via dynamic slicing [J]. Journal of the American Statistical Association, 2015, 110(510): 642-653.

[22] HELLER R, HELLER Y, KAUFMAN S, et al. Consistent distribution-free K-sample and independence tests for univariate random variables [J]. Journal of Machine Learning Research, 2016, 17(29): 1-54.

[23] BREIMAN L, FRIEDMAN J. Estimating optimal transformations for multiple regression and correlation [J]. Journal of the American Statistical Association, 1985, 80(395): 580-598.

[24] LOPEZPAZ D, HENNIG P, SCHOLKOPF B. The randomized dependence coefficient [C]// Proceedings of 26th International Conference on Neural Information Processing Systems, 2013: 1-9.

[25] 吴喜之,赵博娟.非参数统计[M].北京:中国统计出版社,2013:112-130.
WU Xizhi, ZHAO Bojuan. Non-Parametric Statistics [M]. Beijing: China Statistics Press, 2013: 112-130.

[26] 樊嵘,孟大志,徐大舜.统计相关性分析方法研究进展[J].数学建模及其应用,2014,3(1):1-12.
FAN Rong, MENG Dazhi, XU Dashun. Survey of research process on statistical correlation analysis [J]. Mathematical Modeling and Its Applications, 2014, 3(1): 1-12.

(责任编辑　姜丰辉)