专题综述

统计相关性分析方法研究进展

樊嵘[1],孟大志[2],徐大舜[1]

(1. 南伊利诺伊州立大学 数学系,伊利诺伊州 卡本代尔 62901,美国;2. 北京工业大学 应用数理学院,北京 100022)

摘 要:系统综述了自 19 世纪开始至今常用的统计相关性的方法,例如 Pearson 和 Spearman 相关系数,CorGc 和 CovGc 相关性及距离相关性方法。重点介绍了 2011 年提出的 MIC 方法以及由此引发的毁誉参半的大量评述,旨在揭示这一热点领域的研究面貌。该领域不仅受到统计学家的关注,而且受到了分析大样本和异质数据的应用研究领域的学者们的追捧,例如基因组生物学家和网络信息研究者。这些研究者期望在众多已有方法的理解和剖析中更恰当地付诸应用,并提出新的应用问题来推动新的分析方法的创造。

关键词:相关分析;Pearson 相关系数;Spearman 相关系数;Kendall 相关系数;互信息;距离相关;最大信息系数

中图分类号:O212 **文献标志码**:A **文章编号**:2095-3070(2014)01-0001-12

0 引言

不同事物间的相互关联是大自然成为一个整体的基础,集合中元素之间关联性的全体构成系统的结构,是系统作为整体的基础。因此,研究相互关联是科学最基本的内容,特别是在信息时代的今天,海量数据成为当今世界最显著的特征,研究事物大量数据之间的关联性成为科学研究的热点。例如,从海量基因组数据中发掘基因之间的相互关系,是现代生物学的一个重要研究课题。

事物之间的关联性十分复杂,有些是确定的,有些则是不确定的,于是,用于描述事物关联性的数量特征大致可以分为确定性的和随机性的。相应地,把研究对象的特征或属性用变量表示,变量之间的关系可以分为 2 类:确定性函数关系和统计关系。事实上,函数是变量之间的对应关系,但在现实中,变量之间的关系往往并不那么简单。比如,子女身高和父母身高、家庭收入和支出、一个人所受的教育程度与其收入等,它们之间确实存在某种关系,但这些关系无法像函数关系那样,能够用一个确定的公式来描述。当一个变量 X 取一定值时,另一个变量 Y 的值可能有几个,并且以不同的概率出现,即一个变量的值不能由另一个变量唯一确定,这种关系称为统计关系。而且,统计关系中有的关系强,有的关系弱,程度各有差异。如何度量事物间统计关系的强弱也是人们关注的问题。度量变量之间的相关程度,并用适当的统计指标表示出来,这个过程就是统计相关性分析。迄今为止,已经提出了很多相关性度量的指标,如 Spearman 秩相关系数(Spearman rank correlation coefficient)、互信息估计(mutual information estimators)[1-3]、最大相关系数(maximum correlation coefficient)[4-5]、基于曲线原理的方法(principle curve-based methods)[6-10]、距离相关(distance correlation)[10]以及最大信息系数(maximal information coefficient, MIC)[11-12]等。这些衡量变量间相关性的统计量都需要满足一定的条件。

在统计相关性度量的研究中,一些系统的理论方法逐渐建立起来,推动了这一研究领域的发展。1959 年,Renyi[4] 认为,在同一个概率空间上度量 2 个随机变量间的关联程度时,必须满足 7 条性质,后称为公理,是度量相关性的统计量应该满足的基本性质。也就是说,如果记 $\delta(X,Y)$ 为随机变量 X 和 Y 之间的相关性度量的统计量,那么它应该满足以下公理:

1) $\delta(X,Y)$ 是对成对随机变量 X 和 Y 之间相关性的度量,X 和 Y 都不能是以 1 为概率的常数;

收稿日期:2014-01-31

通讯作者:樊嵘,E-mail:rongfan@siu.edu

2)$\delta(X,Y)=\delta(Y,X)$;

3)$0\leqslant\delta(X,Y)\leqslant 1$;

4)$\delta(X,Y)=0$ 当且仅当 X 和 Y 相互独立;

5) 如果 X 和 Y 之间有一个严格的依赖关系,即如果 $X=g(Y)$ 或者 $Y=f(X)$,其中,$g(\cdot)$ 和 $f(\cdot)$ 都是 Borel 可测函数,那么 $\delta(X,Y)=1$;

6) 如果 Borel 可测函数 $g(\cdot)$ 和 $f(\cdot)$ 将实数一一映射到自身,那么 $\delta(f(X),g(Y))=\delta(X,Y)$;

7) 如果 X 和 Y 的联合密度函数是正态分布,那么 $\delta(X,Y)=|R(X,Y)|$,其中,$R(X,Y)$ 是 X 和 Y 之间的 Pearson 相关系数。

围绕这组公理掀起了讨论的热潮,Bell[13],Schweizer 和 Wolff[14],Granger[15] 以及 Nelsen[16] 等对这些公理进行了进一步的修正和完善。有些统计量满足所有的这 7 条公理,如 Shannon 互信息[17],而有的统计量虽然具有很好的性质,但是会违背其中的某些公理,如 Bjerve 和 Doksum[18] 提出的相关曲线(correlation curve),它是非对称的,因此违背了公理。本文将综述统计相关性的主要发展过程,特别介绍 2011 年发表在《Science》上的 MIC 方法,以及对这种方法的发展和评论性意见。

1 常用相关性

1.1 Pearson 相关系数

1885 年,英国著名生物学家和统计学家高尔顿在研究人类遗传问题时第一次提出了“回归”的概念[19]。他搜集了 1 078 对父子的身高数据,发现这些数据的散点图大致呈直线分布,也就是说,总的趋势是父亲身高增加时儿子的身高也倾向于增加;但是当父亲高于平均身高时,他们的儿子比他们更高的概率要小于比他们更矮的概率;当父亲矮于平均身高时,他们儿子的身高比他们更矮的概率要小于比他们更高的概率。这揭示了一个规律,即儿子的身高有向他们父辈的平均身高回归的趋势,使得人类身高的分布相对稳定,而不会产生两极分化,这就是所谓的回归效应。在文章中,高尔顿完成了关于 2 个变量相关性的理论。10 年以后,Karl Pearson[20] 提出了至今仍在使用的 Pearson 相关系数:$r=\dfrac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}}$。从不同的角度看 r,它会被赋予不同的意义。Rodgers 和 Nicewander[21] 展示了 Pearson 相关系数 r 是如何被看作一类特殊的均值、一类特殊的方差、两个均值的比率、两个方差的比率、直线的斜率、一个角度的余弦等的。r 通常被看作 2 个随机变量间线性相关性强弱的指标,取值在 $-1\sim 1$。r 的值越接近 1,表示 2 个变量正相关,线性相关性越强;越接近 -1,表示负相关;接近或者等于 0,表示 2 个变量之间的线性关系很弱或不是线性关系。

1.2 Spearman 相关系数

Spearman 相关系数又称秩相关系数、等级相关系数或顺序相关系数,是利用 2 个变量的秩做线性相关分析,用来衡量 2 个变量间是否单调相关,定义如下。

定义 1 Spearman 相关系数 ρ 被定义为 2 个 n 维随机变量 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$ 和 $\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_n)$ 的秩之间的 Pearson 相关系数:

$$\rho=\frac{\sum_{i=1}^{n}(r_i-\bar{r})(s_i-\bar{s})}{\sqrt{\sum_{i=1}^{n}(r_i-\bar{r})^2}\sqrt{\sum_{i=1}^{n}(s_i-\bar{s})^2}}。$$

其中,r_i 和 s_i 分别是 x_i 和 y_i 的秩,$i=1,2,\cdots,n$。当变量里出现相等值的时候(秩结),该值对应的秩为这几个值对应的秩的平均值。ρ 的取值范围为 $[-1,1]$。当一个变量随另一个变量单调递增的时候,$\rho=1$,反之,$\rho=-1$。

Spearman 相关系数与变量的分布和样本容量都无关,只要 2 个变量的观测值是成对的等级评定资料,或者是由连续变量观测资料转化得到的等级资料,就可以用 Spearman 相关系数进行研究。图 1[22] 表现了

Pearson 相关系数和 Spearman 相关系数的联系和区别。图中的 2 个变量 X 和 Y 之间的 Pearson 线性相关系数为 0.88，表示它们的线性相关程度为 0.88；Spearman 相关系数为 1，表示它们的单调相关程度等于 1，也就是说，这 2 个变量间的单调性很强。

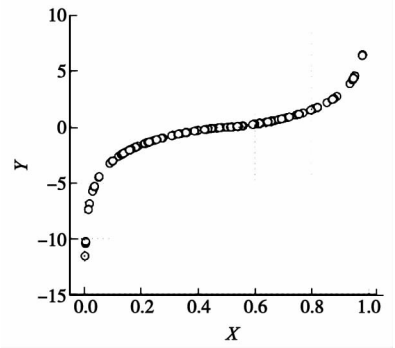


图 1　两个随机变量的散点图

1.3　Kendall 相关系数

Kendall 相关系数(Kendall coefficient of concordance)[23] 是衡量等级变量相关程度的一个统计量，它的主要思想是根据 2 个变量间序对的一致性来判断其相关性。

设 $\boldsymbol{X}$，$\boldsymbol{Y}$ 分别是 n 维随机变量，其中 X_i，Y_i 分别表示 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的第 i 个分量。记 (X_i, Y_i) 为一个序对。当 (X_i, Y_i) 与 (X_j, Y_j) 的排行相同时，即 $X_i > X_j$ 且 $Y_i > Y_j$，或者 $X_i < X_j$ 且 $Y_i < Y_j$ 时，称这个序对是一致的；当 $X_i > X_j$ 且 $Y_i < Y_j$，或者 $X_i < X_j$ 且 $Y_i > Y_j$ 时，称这个序对是不一致的；当 $X_i = X_j$ 或者 $Y_i = Y_j$ 时，这个序对既不是一致的也不是不一致的。

定义 2　设 2 个 n 维随机变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$，它们之间的 Kendall 系数 τ 定义为：

$$\tau = \frac{2P}{\frac{1}{2}n(n-1)} - 1 = \frac{4P}{n(n-1)} - 1。$$

其中，P 表示一致的序对个数。Kendall 系数 τ 的取值范围是 $-1 \sim 1$，当 $\tau = 1$ 时，表示 2 个随机变量拥有一致的等级相关性；当 $\tau = -1$ 时，表示 2 个随机变量拥有完全相反的等级相关性；当 $\tau = 0$ 时，表示 2 个随机变量是相互独立的。

1.4　最大相关系数

设 X_1 和 X_2 为定义在概率空间 (X, A, P) 上，分别在 (X_1, B_1) 和 (X_2, B_2) 上取值的随机变量。映射 X_i：$(X, A) \to (X_i, B_i)$ 生成 A 中的一个子代数 $A_i = X_i^{-1}(B_i)$，$i = 1, 2$。记 P_i 为测度 P 在 A_i 上的一个限制，$i = 1, 2$。设函数 φ 具有有限二阶矩 $E|\varphi| = \int |\varphi| \mathrm{d}P$ 和内积 $(\varphi_1, \varphi_2) = E(\varphi_1, \varphi_2)$，$L^2 = L^2(P)$ 是 A 可测函数 φ 的 Hilbert 空间，$L_i^2 = L_i^2(P)$ 是 A_i 可测函数 φ 的 Hilbert 空间，$i = 1, 2$。Hirschfeld[24] 和 Gebelein[25] 定义了 2 个随机变量 X_1, X_2 之间的最大相关系数

$$R(X_1, X_2) = \sup\{r(\varphi(X_1), \psi(X_2))\}。$$

其中，$r(X_1, X_2)$ 是 X_1 和 X_2 的 Pearson 相关系数，$\varphi(X_1) \in L_1^2$，$\psi(x_2) \in L_2^2$ 为任意函数。

一般情况下得不到最大相关系数的一个精确值，如何选择合适的函数 φ, ψ，使得 $r(\varphi, \psi)$ 可以达到上确界，是很多统计学家们一直在探索的问题。但是只有在某些特殊的情形下，才可以取到相关系数的上确界。比如，若 2 个变量间具有线性关系，那么最大相关系数退化为 Pearson 相关系数；若 X_1 和 X_2 相互独立，则当且仅当 $R(X_1, X_2) = 0$，或者等价地，当且仅当 2 个子空间 L_1^2 与 L_2^2 正交。

1959 年，Renyi[4] 指出，如果

$$R(X_1, X_2) = r(\varphi(X_1), \psi(X_2)) = \rho,$$

其中，$\varphi(X_1)$，$\psi(X_2)$ 满足

$$E\varphi(X_1) = E\psi(X_2) = 0, E\varphi(X_1)^2 = E\psi(X_2)^2 = 1,$$

那么，$E(\varphi(X_1) \mid X_2) = \rho\psi(X_2)$，$E(\psi(X_2) \mid X_1) = \rho\varphi(X_1)$。

基于这个结论，1985 年，Breiman 和 Frideman[26] 给出另外一种条件期望的算法来寻找 φ 和 ψ，使得 $r(\varphi, \psi)$ 达到最大。该文中也提出了当样本观测值 (X_1, X_2) 满足一定条件时，最大化 φ, ψ 是如何被估计的。如果 (X_1, X_2) 服从二维高斯分布，Pearson 相关系数等于 r，那么它们的最大相关系数等于 r 的绝对值，即

$$R(X_1, X_2) = |r|。$$

Sethuraman[27] 在 X_1 和 X_2 相互独立且只有有限个取值的情形下，讨论了最大相关系数的近似分布问题。Dembo 等[28] 对独立同分布随机变量部分和之间的最大相关系数和 Pearson 相关系数之间的关系做了研究。Czaki 与 Fisher[29] 从几何角度研究了最大相关系数的性质，例如，$R(X_1, X_2)$ 可以看作是 2 个子空间 L_1^2 和 L_2^2 的夹角余弦，即 $R(X_1, X_2) = \cos(L_1^2, L_2^2)$。

1.5 互信息

Shannon 在 1948 年首次定义了互信息(mutual information)[17]，用来度量 2 个变量 X, Y 间的相互依赖程度：

$$I(X,Y) = \sum p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)。$$

其中：$p(x,y)$ 是 X 和 Y 的联合概率密度函数；$p(x)$ 和 $p(y)$ 分别是 X 和 Y 的边际密度函数。

互信息具有很好的性质。首先，它满足 Renyi 提出的所有 7 条公理；其次，互信息的值与衡量联合概率密度函数和边际密度函数乘积之间距离的 Kullback-Leibler 散度的值一致。但是，通常由于随机变量的概率密度函数未知，造成互信息无法或者难以估计。于是针对如何估计互信息值，科学家做了许多努力。比如，Moon 等[1] 介绍了如何用核密度估计(kernel density estimation，KDE) 的方法估计互信息值；Kraskov 等[3] 用 k-最近邻距离 (k-nearest neighbor distances，KNN) 的方法改进了传统的计算互信息的方法等。Walters-Willians[30] 比较了到 2008 年为止不同的互信息估计方法的算法优劣和估计有效性等。比如，他指出，当样本容量 n 趋于无穷大，且方差随之变小，而且 k 随着 N 以适当的方式增长时，KDE 和 KNN 估计的密度函数均收敛于真实的概率密度函数，因此，KDE 和 KNN 比其他估计互信息的方法有其优越性的一面。

1.6 CorGc 和 CovGc 相关性

Delicado[31] 在考虑变量间的相关性的时候，将方差协方差矩阵以及相关系数矩阵作谱分解，并得到这些特征的主成分表示，在此基础上给出了新的度量方法，分别称为 Covariance along a Generating Curve(CovGc) 和 Correlation along a Generating Curve(CorGc)。它们是二维随机变量在 $\mathbf{R}^2$ 上沿着一条一维曲线来定义相关性的方法。在定义 CovGc 和 CorGc 之前，需要一些准备知识。

首先对 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的协方差和相关系数矩阵做谱分解，得到方差、协方差和相关系数的一些基于特征值的表达式，即设 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的方差协方差矩阵为：

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{\boldsymbol{X}}^2 & \sigma_{\boldsymbol{XY}} \\ \sigma_{\boldsymbol{YX}} & \sigma_{\boldsymbol{Y}}^2 \end{bmatrix} = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}^{\mathrm{T}} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}。$$

其中：$\lambda_1 \geqslant \lambda_2$ 是协方差矩阵 $\boldsymbol{\Sigma}$ 的特征值；α 是对应特征值 λ_1 的特征向量和 x 轴的夹角。则 $\boldsymbol{X}, \boldsymbol{Y}$ 的方差、协方差、相关系数分别用特征值和 α 表示为：

$$\begin{aligned} Var(\boldsymbol{X}) &= \sigma_{\boldsymbol{X}}^2 = \lambda_1\cos^2\alpha + \lambda_2\sin^2\alpha, \\ Var(\boldsymbol{Y}) &= \sigma_{\boldsymbol{Y}}^2 = \lambda_1\sin^2\alpha + \lambda_2\cos^2\alpha, \\ Cov(\boldsymbol{X},\boldsymbol{Y}) &= \sigma_{\boldsymbol{XY}} = (\lambda_1 - \lambda_2)\cos\alpha\sin\alpha, \\ \rho_{\boldsymbol{XY}} &= \frac{\sigma_{\boldsymbol{XY}}}{\sigma_{\boldsymbol{X}}\sigma_{\boldsymbol{Y}}} = \frac{(\lambda_1 - \lambda_2)\cos\alpha\sin\alpha}{(\lambda_1\cos^2\alpha + \lambda_2\sin^2\alpha)^{\frac{1}{2}}(\lambda_1\sin^2\alpha + \lambda_2\cos^2\alpha)^{\frac{1}{2}}}。 \end{aligned}$$

然后还需要给出二维随机变量 $(\boldsymbol{X},\boldsymbol{Y})$ 沿着曲线 $c(I)$ 分布的定义。

定义 3 设 $(\boldsymbol{X},\boldsymbol{Y})$ 是从下面公式得到的二维随机变量

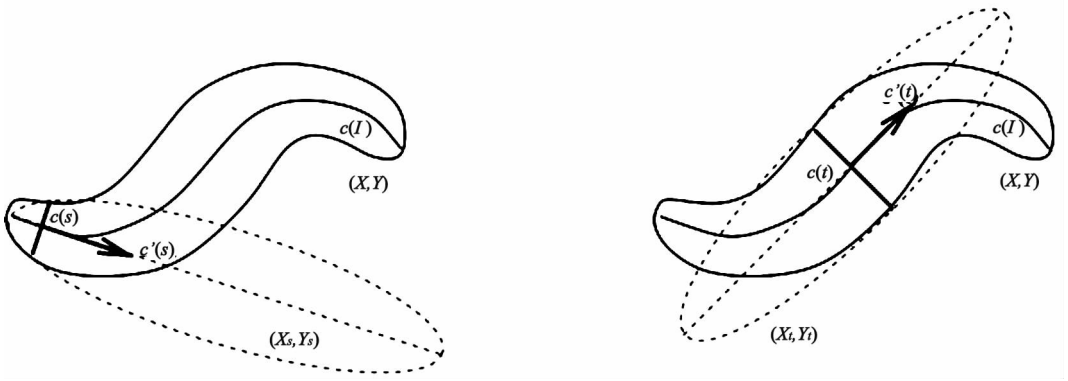
$$(\boldsymbol{X},\boldsymbol{Y}) = \chi_c(S,T) : A \to \mathbf{R}^2。$$

其中：$c(I)$ 是生成曲线；(S,T) 是生成随机变量；$\chi_c : I \times \mathbf{R} \to \mathbf{R}^2$ 定义为 $\chi_c(s,t) = c(s) + tv(s)$，其中 $c(s) : I \subseteq \mathbf{R} \to \mathbf{R}^2$ 是平面上的一维光滑曲线，并且对所有的 $s \in I$ 满足 $\| c'(s) \| = 1$；$v(s)$ 是一个酉向量，并且对所有

的 $s \in I$ 满足 $c'(s)^{\mathrm{T}} v(s) = 0$。则称二维随机变量 $(\boldsymbol{X},\boldsymbol{Y})$ 沿着曲线 $c(I)$ 分布。

估计 I，$c(s)$ 和 $c'(s)$ 的一般方法是用主曲线(principal curves)拟合算法。主曲线是第一主成分的非线性推广，第一主成分是对数据集的一维线性最优描述。主曲线强调寻找通过数据分布的“中间”(middle)并满足“自相合”的光滑一维曲线，其理论基础是寻找嵌入高维空间的非欧氏低维流形。更多关于主曲线的信息可以参见文献[32-34]。Hastie、Stuetzle等和 Delicado 分别引入了 3 种不同的主曲线的概念，并给出了相应的算法，这些算法都已经嵌入到很多软件中，比如 R 和 Matlab。

最后，利用方差协方差及其相关系数的谱分解表示，给出局部线性相关性度量的定义，再将该定义扩展到全局相关性度量。由于 $(\boldsymbol{X},\boldsymbol{Y})$ 可以看作是曲线 $c(I)$ 上的随机点和正交随机噪音生成的，于是 $c(I)$ 可以代表 $(\boldsymbol{X},\boldsymbol{Y})$ 的分布结构。特别当 $c(I)$ 是一条直线的时候，$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 之间的相关性被看作是线性相关，相关程度可以用方差和相关系数矩阵很好地度量。为了获得更一般的线性相关程度的度量，首先在 $c(I)$ 上的一个点 $c(s)$ 周围定义局部的方差和协方差度量。这样定义的思想是将 $(\boldsymbol{X},\boldsymbol{Y})$ 的分布在 $c(s)$ 周围线性化，也就是说，在 $c(s)$ 周围寻找一个随机变量 $(\boldsymbol{X}_s,\boldsymbol{Y}_s)$，使得它沿着一条直线的分布与 $(\boldsymbol{X},\boldsymbol{Y})$ 的分布近似。图 2[33] 显示了在 2 个点 $c(s)$ 和 $c(t)$ 处的线性化过程。



(a) 点 $c(s)$ 处的线性化　　(b) 点 $c(t)$ 处的线性化

图 2　(x,y) 在 $c(s)$ 和 $c(t)$ 处的线性化过程图

定义 4　设 $(\boldsymbol{X},\boldsymbol{Y}) = \chi_c(S,T)$ 是沿着曲线 $c(I)$ 分布的二维随机变量。对 $s \in I$，记 $\alpha(s)$ 为 $c'(s)$ 与 x 轴的夹角，定义 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 在 $c(s)$ 处的局部方差为

$$LV_{\boldsymbol{X}}(S) = V(S)\cos^2\alpha(s) + V(T \mid S = s)\sin^2\alpha(s),$$

$$LV_{\boldsymbol{Y}}(S) = V(S)\sin^2\alpha(s) + V(T \mid S = s)\cos^2\alpha(s),$$

在 $c(s)$ 处的局部协方差为

$$LCov_{(\boldsymbol{X},\boldsymbol{Y})}(S) = (V(S) - V(T \mid S = s))\cos\alpha(s)\sin\alpha(s),$$

在 $c(s)$ 处的局部相关系数为：

$$LCor_{(\boldsymbol{X},\boldsymbol{Y})}(S) = LCov_{(\boldsymbol{X},\boldsymbol{Y})}(S)/(LV_{\boldsymbol{X}}(S)LV_{\boldsymbol{Y}}(S))^{\frac{1}{2}}。$$

由于局部协方差和局部相关系数可能为负数，当扩展为全局协方差和全局相关系数时可能会被消去，于是引入平方再开根号的办法来避免这个问题的发生。

定义 5　沿用定义 4 中的符号，$\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 沿着它们的生成曲线 $c(I)$ 产生的分布的(全局)协方差定义为：

$$CovGc(\boldsymbol{X},\boldsymbol{Y}) = \{E_S[(LCov_{(\boldsymbol{X},\boldsymbol{Y})}(S))^2]\}^{\frac{1}{2}},$$

沿着曲线 $c(I)$ 的(全局)相关系数定义为：

$$CorGc(\boldsymbol{X},\boldsymbol{Y}) = \{E_S[(LCor_{(\boldsymbol{X},\boldsymbol{Y})}(S))^2]\}^{\frac{1}{2}}。$$

这种相关性度量的性质很好地满足经过适当修改的 Renyi 的 7 条公理，其适用范围也非常广泛，可以用来测试变量间的独立性、线性关系以及定义变量间的相似性。比如，当 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的样本数据分布在一个环或者矩形上时，CorGc 能捕捉到这个相关性，给出较高的值；当 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 满足某种特定的关系时，这种度量得到的值

与适用于度量该关系的统计量的值很接近。比如，当 X 和 Y 之间线性相关的时候，CorGc 的值与 Pearson 相关系数几乎相等。另外，这种算法已经嵌入到 R 软件中了，使用起来非常方便。

1.7 距离相关

距离相关(distance correlation，dCor) 是 Szkely 等[35] 于 2007 年提出来的一种新的度量相关性的统计量。它不同于以往的基于协方差矩阵和方差矩阵所定义的相关性，具体来说，总体距离相关等于 0 可以推导出变量间的独立性，但是反之不成立。区别于传统的计算样本矩之间的距离，样本距离相关是通过计算样本本身的欧几里得距离来衡量变量间的相关程度。

设 X 为 p 维随机变量，Y 为 q 维随机变量，X 和 Y 都有有限一阶矩，则 X 和 Y 之间的总体距离协方差(theoretical distance covariance) 定义为：

$$V(\boldsymbol{X},\boldsymbol{Y}) = \sqrt{V^2(\boldsymbol{X},\boldsymbol{Y})} = \sqrt{\| f_{\boldsymbol{X},\boldsymbol{Y}}(t,s) - f_{\boldsymbol{X}}(t) f_{\boldsymbol{Y}}(s) \|^2} = \sqrt{\frac{1}{c_p c_q}\int_{\mathbb{R}^{p+q}} \frac{| f_{\boldsymbol{X},\boldsymbol{Y}}(t,s) - f_{\boldsymbol{X}}(t) f_{\boldsymbol{Y}}(s) |^2}{| t |_p^{1+p} \cdot | s |_q^{1+q}} \mathrm{d}t\mathrm{d}s}。$$

其中：$f_{\boldsymbol{X}}(t)$ 和 $f_{\boldsymbol{Y}}(s)$ 为 X 和 Y 的特征函数；$f_{\boldsymbol{X},\boldsymbol{Y}}(t,s)$ 为 X 和 Y 的联合特征函数；$c_d = \frac{\pi^{1+d/2}}{\Gamma((1+d)/2)}, d = p$ 或者 q。

类似地，总体距离方差(distance variance) 定义为：

$$V(\boldsymbol{X}) = \sqrt{V(\boldsymbol{X},\boldsymbol{X})} = \sqrt{\| f_{\boldsymbol{X},\boldsymbol{Y}}(t,s) - f_{\boldsymbol{X}}(t) f_{\boldsymbol{Y}}(s) \|^2}。$$

总体距离系数(distance correlation) 定义为：

$$R(\boldsymbol{X},\boldsymbol{Y}) = \sqrt{R^2(\boldsymbol{X},\boldsymbol{Y})} = \begin{cases} \sqrt{\dfrac{V^2(\boldsymbol{X},\boldsymbol{Y})}{\sqrt{V^2(\boldsymbol{X})V^2(\boldsymbol{Y})}}}, & V^2(\boldsymbol{X})V^2(\boldsymbol{Y}) > 0 \\ 0, & V^2(\boldsymbol{X})V^2(\boldsymbol{Y}) = 0 \end{cases}。$$

$R(\boldsymbol{X},\boldsymbol{Y})$ 具有性质：①$0 \leqslant R(\boldsymbol{X},\boldsymbol{Y}) \leqslant 1$；②$R(\boldsymbol{X},\boldsymbol{Y}) = 0$ 表示 X 和 Y 相互独立。这个系数与 Pearson 相关系数的定义非常类似。事实上，在 X 和 Y 都是标准正态分布的条件下，它们之间也存在一定的函数关系，例如 $R(\boldsymbol{X},\boldsymbol{Y}) \leqslant | r(\boldsymbol{X},\boldsymbol{Y}) |$，等号在 $r(\boldsymbol{X},\boldsymbol{Y}) = \pm 1$ 时成立。当 X 和 Y 的联合分布的观测值 $(\boldsymbol{X},\boldsymbol{Y}) = \{(\boldsymbol{X}_k,\boldsymbol{Y}_k), k = 1,2,\cdots,n\}$ 给定的时候，定义

$$a_{kl} = | X_k - X_l |_p, \bar{a}_k = \frac{1}{n}\sum_{l=1}^{n} a_{kl}, \bar{a}_l = \frac{1}{n}\sum_{k=1}^{n} a_{kl},$$

$$\bar{a} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}, A_{kl} = a_{kl} - \bar{a}_k - \bar{a}_l + \bar{a}, k,l = 1,2,\cdots,n。$$

类似地，定义：

$$b_{kl} = | Y_k - Y_l |_q, B_{kl} = b_{kl} - \bar{b}_k - \bar{b}_l + \bar{b}, k,l = 1,2,\cdots,n。$$

于是，样本距离协方差(empirical distance covariance，dCov) 定义为：

$$V_n(\boldsymbol{X},\boldsymbol{Y}) = \sqrt{V_n^2(\boldsymbol{X},\boldsymbol{Y})} = \sqrt{\frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}}。$$

类似地，样本距离方差 $V_n(\boldsymbol{X})$ 定义为

$$V_n(\boldsymbol{X}) = \sqrt{V_n^2(\boldsymbol{X})} = \sqrt{V_n^2(\boldsymbol{X},\boldsymbol{Y})} = \sqrt{\frac{1}{n^2}\sum_{k=1}^{n} A_{kl}^2}。$$

样本距离相关(empirical distance correlation)$R_n(\boldsymbol{X},\boldsymbol{Y})$ 定义为：

$$R_n(\boldsymbol{X},\boldsymbol{Y}) = \sqrt{R_n^2(\boldsymbol{X},\boldsymbol{Y})} = \begin{cases} \sqrt{\dfrac{V_n^2(\boldsymbol{X},\boldsymbol{Y})}{\sqrt{V_n^2(\boldsymbol{X})V_n^2(\boldsymbol{Y})}}}, & V_n^2(\boldsymbol{X})V_n^2(\boldsymbol{Y}) > 0 \\ 0, & V_n^2(\boldsymbol{X})V_n^2(\boldsymbol{Y}) = 0 \end{cases}。$$

$R_n(\boldsymbol{X},\boldsymbol{Y})$ 具有性质：①$0 \leqslant R_n(\boldsymbol{X},\boldsymbol{Y}) \leqslant 1$；②$R_n(\boldsymbol{X},\boldsymbol{Y}) = 0$ 当且仅当 X 和 Y 相互独立。如果 $R_n(\boldsymbol{X},\boldsymbol{Y}) = 1$，那么存在一个非零常数 b 和一个正交矩阵 C，使得 $\boldsymbol{Y} = a + b\boldsymbol{X}\boldsymbol{C}$。

利用上述距离协方差 dCov 做样本独立性检验的时候，它的统计相关性比所有期望有限的备择假设都要好。数值拟合的结果显示，当随机变量间的关联性是非线性的时候，利用 dCov 做的检验比极大似然比检验的

势要高很多。统计量 dCov 能够很好地探测到变量间的非线性或者非单调的关系。

2　MIC 相关性[11]

Reshef 等[11] 在 2011 年 11 月发表在《Science》上的文章掀起了研究相关性新的热潮。文章引入的最大信息系数(maximal information coefficient，MIC) 被用来度量变量间的相关程度。MIC 方法的主要思想基于这样一个认识：如果 2 个变量间存在某种相关，那么在这 2 个变量构成的散点图上进行网格划分后，数据在网格中的分布情况可以反映出它们之间的关联性。MIC 的算法与传统的算法也有很大的区别。传统的 Pearson 相关等的计算都可以写出公式，用计算器计算得出，但是 MIC 没有一个简单的计算公式，也不能通过任何一个计算器计算得到，而必须借助现代化的数字计算机运行一系列程序算法才有可能得到。也正是这个原因，导致 MIC 方法到现在才被发现和提出。纵使如此，得到 MIC 精确解的计算量仍然非常巨大，因此文中给出了一个简化的优化方法，可以得到 MIC 的近似解。这是统计学关于计算机密集型(computer-intensive) 方法的另外一个例子[36]。

2.1　方法的定义

MIC 的算法主要由以下 2 个因素决定：① 网格划分数，即在给定的数据集形成的散点图上，在 x 轴和 y 轴上分别进行多少次的划分；② 网格划分的位置，即如果在 x 轴上划分 a 次，那么这 a 个划分点是等距放置还是以某种其他方式放置在 x 轴上。若给定划分数和划分位置，则给定了一种划分，计算该划分下的互信息值：

$$I(D,X,Y)=\sum_{x\in X,y\in Y}p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)。$$

其中：D 是给定的数据集；X,Y 是对这个数据集的划分，都是随机变量；$p(x,y)$ 是联合概率函数，这里用落入格子中的样本数占样本容量的比例来近似；$p(x)$，$p(y)$ 是边缘概率分布函数，这里分别用落入 $(k,k+1)$ 和 $(l,l+1)$ 区间的样本数占样本容量的比例来近似，其中 $0\leqslant k\leqslant x-1,0\leqslant l\leqslant y-1$。

若固定网格划分数，则通过改变网格划分位置，会得到不同的互信息值，记其中的最大互信息值为 $I(D,x,y)$。进一步，为了方便在不同的维数之间进行比较，将其标准化，使其取值在区间 $[0,1]$：

$$M(D)_{x,y}=\frac{I(D,x,y)}{\log(\min\{x,y\})}。$$

定义 6　设有 2 个随机变量的数据集 D，样本容量为 n，网格划分数小于 $B(n)$。它的极大互信息系数(MIC) 定义为

$$MIC(D)=\max_{xy<B(n)}\{M(D)_{x,y}\}。$$

MIC 的计算图解见图 3，具体解释如下。

(A) 计算每一个 (x,y) 划分对应的最大互信息。最左边一列的 3 个图是 2×2 划分下对应的几个不同的划分位置。最下面的那种划分位置可以得到该划分下的最大互信息。中间 3 个图是 2×3 划分下对应的几个不同的划分位置。其中最下面的那种划分位置得到该划分下的最大互信息。

(B) 由标准化后的互信息得分组成的矩阵。这个矩阵保存了所有的 m_{xy} 及其相应的划分方法。该矩阵中的元素分别对应 A 图各 $x\times y$ 划分下划分位置最好时得到的互信息值。

(C) 特征矩阵 $\boldsymbol{M}=[m_{xy}]$ 可表示为可视化的一个表面，MIC 对应于这个表面上的最大值点。在这个例子中，有很多种划分都可以得到最高得分。(B) 中的星号标示了其中一个样本划分得到的分值，在(C) 中的星号标示了这个得分在表面上的位置。

MIC 旨在发现大数据集中 2 个变量间所有重要的、但未被发现的关系，并且有效地识别这些关系。结果证明，MIC 几乎能探测出所有的函数甚至非函数关系。当 2 个变量间是函数关系时，MIC 与该函数的决定系数 R^2(函数的因变量真实值与预测值之间的 Pearson 相关系数的平方) 几乎相等。特别地，当该函数是线性函数时，MIC 约等于 Pearson 相关系数 r 的平方。MIC 的取值范围是 $0\sim1$。当 MIC 为 0 时，表示 2 个变量完全统计独立；而当 MIC 为 1 时，表示函数没有噪音，这个完全符合统计常识。

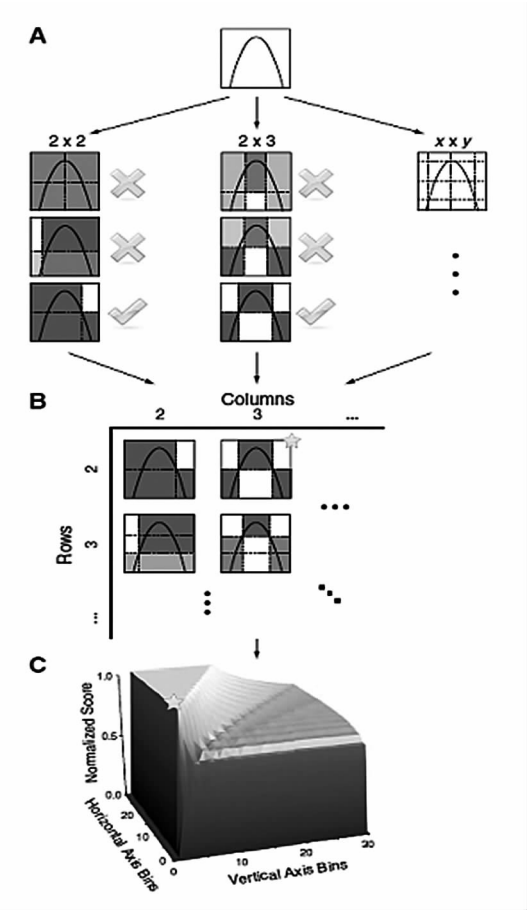


图 3　MIC 计算示意图[11]

2.2　MIC 的性质

MIC 的主要性质有 2 个：普适性和等价性。普适性是指当样本容量足够大的时候，MIC 可以探测到更大范围的相关性。例如周期函数，或者像圆这样的非函数，及由几个函数合成的超函数。等价性是指不论哪种类型的函数，在相同噪音（干扰）条件下，MIC 与该函数相关性（R^2）的得分接近。这 2 个性质可通过图 4 表现出来。

Rolatlonship Type	Thumbnails of Relationships Used ——Increasing Noise→				MIC Score ——Increasing Noise→			
Linear					0.80	0.65	0.50	0.35
Two Lines					0.80	0.65	0.50	0.35
Parabouc					0.80	0.65	0.50	0.35
Line and Parabola					0.80	0.65	0.50	0.35
X					0.77	0.65	0.50	0.35
Ellipse					0.78	0.65	0.50	0.35
Sinusoldal					0.80	0.65	0.50	0.35
Sinusoldal (*Mixture of two signals*)					0.76	0.65	0.50	0.35
Non-coexistence					0.69	0.65	0.50	0.35

(a)

Relationship Type	R^2 w.r.t. Smoothing spline — Increasing Noise →				R^2 w.r.t Locally Weighted Nonparametrk Regression(Loess) — Increasing Noise →			
Linear	0.85	0.70	0.54	0.37	0.85	0.70	0.53	0.36
Two Lines	0.28	0.28	0.27	0.23	0.27	0.27	0.26	0.22
Parabolic	0.87	0.71	0.50	0.28	0.86	0.71	0.52	0.31
Line and Parabola	0.31	0.31	0.31	0.27	0.31	0.31	0.30	0.26
X	0.00	0.01	0.01	0.02	0.00	0.00	0.00	0.00
Ellipse	0.03	0.04	0.04	0.04	0.03	0.03	0.03	0.03
Sinusoidal	0.84	0.65	0.43	0.27	0.85	0.64	0.43	0.26
Sinusoidal (*Mixture of two signals*)	0.49	0.49	0.40	0.29	0.49	0.49	0.40	0.27
Non-coexistence	0.52	0.47	0.41	0.31	0.47	0.46	0.40	0.30

(b)

图 4　MIC 的普适性和等价性示意图[11]

图 4 中，(a) 是随着噪音的增加几种函数关系的变化形态及其得到的 MIC 值；(b) 是 4 种噪音下用 2 种算法计算得到的 R^2。在噪音比较低的情况下，MIC 为各种函数关系都给出了较高的值，表现出它的普适性。同时从图 4 可以看出，MIC 值和 R^2 在很多情形下是近似相等的，表现出它的等价性。

2.3　MIC 与其他统计量的比较

MIC 代表了最广义的一种相关。不论 2 个变量是函数相关、超函数相关、甚至是没有任何函数关系的相关时，MIC 都可以探测到这种相关性，即在无噪音的情况下，给出较高的 MIC 值。数值拟合结果见表 1。从表中也可以看出，MIC 值与专门适用于测量某种函数关系的统计量的得分相当。比如，当 2 个变量为线性关系时，MIC 和 Pearson 相关系数都是 1；当 2 个变量是指数关系时，MIC 与衡量单调性的 Spearman 统计量得分都是 1。这也从一个方面说明，MIC 与其他统计量之间存在一定的等价性。

表 1　MIC 与其他统计量的比较表

Relationship Type	MIC	Pearson	Spearman	Mutual (*KDE*)	Information (*Kraskov*)	CorGC (Princlpal Curve-Based)	Maximal Correlation
Random	0.18	-0.02	-0.02	0.01	0.03	0.19	0.01
Linear	1.00	1.00	1.00	5.03	3.89	1.00	1.00
Cubic	1.00	0.61	0.69	3.09	3.12	0.98	1.00
Exponential	1.00	0.70	1.00	2.09	3.62	0.94	1.00
Sinusoidal (Founer frequency)	1.00	0.09	-0.09	0.01	-0.11	0.36	0.64
Categorical	1.00	0.53	0.49	2.22	1.65	1.00	1.00
Periodic/Linear	1.00	0.33	0.310	0.69	0.45	0.49	0.91
Paraoolic	1.00	-0.01	-0.01	3.33	3.15	1.00	1.00
Sinusoidal (non-Founor troquoncy)	1.00	0.00	0.00	0.01	0.20	0.40	0.80
Sinusoidal (varyiogfrequency)	1.00	-0.11	-0.11	0.02	0.06	0.38	0.76

2.4 基于 MIC 拓展的统计量

定义 7 最大非对称得分(maximal asymmetry score,MAS)用来度量 2 个变量间的单调性,定义为

$$MAS(D) = \max_{xy<B} | M(D)_{x,y} - M(D)_{y,x} |。$$

定义 8 最大边值(maximum edge value,MEV)用来衡量随机变量间的关系是否是函数关系,或者说与函数关系的接近程度,定义为

$$MEV(D) = \max_{xy<B}\{M(D)_{x,y}: x = 2 \text{ 或 } y = 2\}。$$

定义 9 最小网格单元数(minimum cell number,MCN)用来衡量相关性的复杂程度,也就是说要达到 MIC 得分需要的最小网格单元数,定义为

$$MCN(D,\varepsilon) = \max_{xy<B}\{\log(xy): M(D)_{x,y} \geqslant (1-\varepsilon)MIC(D)\}。$$

这些统计量统称为 maximal information-based non-parametric exploration(MINE)。这几个统计量表达无噪音的相关性得分见表 2。MIC 在实际例子中也表现出良好的性质,详细参见文献[11]。

3 对 MIC 的评论

Reshef 等列举了很多 MIC 的优良性质,但是很快就有人对 MIC 方法提出了批评。其中 Gorfine 等[37],Simon 等[38]以及 Kinney 等[39]提出的批评意见比较有代表性。

表 2 计算某些样本相关性的 MINE 统计量值表

Data	MIC	MAS	MEV	MCN
	1.00	0.00	1.00	2.00
	1.00	0.74	1.00	3.00
	1.00	0.89	1.00	4.00
	1.00	0.69	1.00	2.56
	0.79	0.16	0.70	6.91
	0.71	0.03	0.32	6.87
	0.46	0.19	0.22	6.98

Gorfine 等比较了 MIC 与另外 2 种统计量 dCor 和 HHG 的优劣。文章指出:① 在完全无噪音(不切实际)的函数下,MIC 比 HHG 有微弱的优势,比 dCor 有显著的优势;②如果样本容量适中(30～100),那么在大部分含噪音的函数以及变量间具有非函数关系的情形下,HHG 和 dCor 在统计检验中的势都远高于 MIC;③Reshef 声称 dCor 的方法在寻找随机变量之间的关系时不适用,但是恰恰相反,由于在实际应用中样本容量多集中在 30～100,而不是 MIC 适用的大样本容量(>100),所以 dCor 和 HHG 方法更适用于实际,同时,dCor 和 HHG 都适用于任意维数的随机变量,而 MIC 只适用于一维随机变量;④针对 Reshef 声称的 MIC 具有的相合性,M. Gorgine 等也给出一个反例:当 2 个变量之间是菱形关系(diamond relation)时,随着样本容量的增大,MIC 的势并没有相应地增加,甚至减少了,并不满足检验相合性要求,而好的检验应该满足 2 个基本性质:①要满足相合性,也就是当样本容量增大的时候,势应该增大并趋向于 1;②在有限样本的情况下,检验也应该有较高的势。

Simon 和 Tibshirani 也用数值模拟的方法指出 MIC 方法的势比很多其他统计量的势都低,进一步指出在如此低的势的情况下,定义"等价性"是没有任何意义的。也就是说会出现这样的情况:即使 2 个变量不相关,MIC 也会给出比较高的得分,从而造成 2 个变量相关的假象。

Kinney 和 Atwal[42]等则不客气地指出,MIC"等价性"的定义是错误的。他们用数学方法证明了 Reshef 定义的"等价性"是不可能存在的,即任何一种对非平凡的相关性度量的统计量,包括 MIC 本身在内,均不满足这种等价性的定义。然后,他们通过数值模拟的方法得出与 Reshef 几乎完全相反的结果,声称 Reshef 等提供的数值模拟结果可能是伪造的。

以上综述给出了统计相关性分析的一个发展全貌,特别介绍了 MIC 方法,尽管仍然会有所遗漏,例如,Bjerve 和 Doksum[18]在平面上定义了一种度量相关程度的量,称为相关曲线,它是基于非参数回归的局部估计系数和仅考虑一个变量的局部线性相关性的度量来给出的。可以注意到,无论 MIC 方法如何受到批评,但是它的出现的确引发了关于这个问题的新的研究热潮。事实上,相关性这个概念本身就没有完全统一,以上多数方法都强调了相关的某个性质,从而给出适合某种特定相关意义的计算方法。本文希望应用研究者们在各自应用的数据特点和问题目标中能够选择更适合的分析方法,同时,更希望他们在自己研究的数据对

象的特征中提出新的问题，推动统计相关方法的发展。

参考文献

[1]Moon Y I, Rajagopalan B, Lall U. Estimation of mutual information using kernel density estimators[J]. Physical Review E, 1995, 52(3): 2318-2321.

[2]Darbellay G, Vajda I. Estimation of the Information by an adaptive partitioning of the observation space[J]. IEEE Transaction Information, 1999, 45(4): 1315 -1321.

[3]Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information[J/OL]. (2004-06-23)[2014-01-14]. Physical Review E, 2004. http://journals. aps. org/pre/abstract/10. 1103/PhysRevE. 69. 066138.

[4]Rényi A. On measures of dependence[J]. Acta Mathematica Academiae Scientiarum Hungarica, 1959, 10(3/4): 441-451.

[5]Breiman L, Friedman J H. Estimating optimal transformations for multiple regression and correlation[J]. Journal of the American Statistical Association, 1985, 80(391): 580-598.

[6]Hastie T, Stuetzle W. Principal curves[J]. Journal of the American Statistical Association, 1989, 84(406): 502-516.

[7]Tibshirani R. Principal curves revisited[J]. Statistics and Computing, 1992, 2(4): 183-190.

[8]Kégl B, Krzyzak A, Linder T, et al. A polygonal line algorithm for constructing principal curves[C]// Advances in Neural Information Processing Systems. Cambridge: MIT Press, 1999(11): 501-507.

[9]Delicado P, Smrekar M. Measuring non-linear dependence for two random variables distributed along a curve[J]. Statistics and Computing, 2009(19): 255-269.

[10]Delicado P. Another look at principal curves and surfaces[J]. Journal of Multivariate Analysis, 2001(77): 84-116.

[11]Reshef D N, Reshef Y A, Finucane H K, et al. Detecting novel associations in large data sets[J]. Science, 2011(334): 1518-1524.

[12]Speed T. A correlation for the 21(upst)century[J]. Science, 2011(334): 1502-1503.

[13]Bell C. Mutual information and maximal correlation as measures of dependence[J]. The Annals of Mathematical Statistics, 1962(33): 587-595.

[14]Schweizer B, Wolff E F. On nonparametric measures of dependence for random variables[J]. The Annals of Mathematical Statistics, 1981, 9(4): 879-885.

[15]Granger C W, Massoumi E, Racine J. A dependence metric for possibly nonlinear processes[J]. Journal of the American Statistical Association, 1989(84): 502-516.

[16]Nelsen R B. An introduction to copulas, 2(upnd)edn, spring series in statistics[M]. New York: Springer, 2006.

[17]Shannon C E, Weaver W. The mathematical theory of communication[M]. Champaign: University of Illinois Press, 1949.

[18]Bjerve S, Doksum K. Correlation curves: measures of association as functions of covariate value[J]. The Annals of Mathematical Statistics, 1993(21): 890-902.

[19]Galton F. Regression towards mediocrity in hereditary stature[J]. Journal of the Anthropological Institute, 1885(15): 246-263.

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

[21]Rodgers J L, Nicewander W A. Thirteen ways to look at the correlation coefficient[J]. The American Statistician, 1988, 42(1): 59-66.

[22]Wikimedia. File: Spearman fig1. svg[EB/OL]. [2014-01-14]. http://commons. wikimedia. org/wiki/File: Spearman_fig1. svg? uselang=zh-cn.

[23]Kendall M G. A new measure of rank correlation[J]. Biometrika, 1938(30): 81-93.

[24]Hirschfeld H O. A connection between correlation and contingency[J]. Proceedings of the Cambridge Philosophical Society, 1935, 31(4): 520-524.

[25]Gebelein H. Das statistische problem der korrelationals variations-und eigenwertproblem undsein Zusammenhangmit der Ausgleichsrechnung[J]. Journal of Applied Mathematics and Mechanics, 1941, 21(6): 364-379.

[26]Breiman L, Friedman J. Estimating optimal transformations for multiple regression andcorrelation(with discussion)[J]. Journal of the American Statistical Association, 1985(80): 580-619.

[27]Sethuraman J. The asymptotic distribution of the renyi maximal correlation[J]. Communications in Statistics, Theory Method, 1990, 19(11): 4291-4298.

[28]Dembo A, Kagan A, Shepp L A. Remarks on the maximum correlation coefficient[J]. Bernoulli, 2001, 7(2): 343-350.

[29]Czaki P, Fisher J. On the general notion of maximum correlation[J]. Publ. Math. Inst. Hung. Acad. Sci, 1963(8): 27-51.
[30]Walters-Williams J. Estimation of mutual information: A survey[J]. Lecture Notes in Computer Science, 2009(5589): 389-396.
[31]Delicado P, Smrekar M. Measuring non-lineat dependence for two random variables distributed along a curve[J]. Statistics and Computing, 2009(19): 255-269.
[32]Hastie T, Stuetzle W. Principal curves[J]. Journal of the American Statistical Association, 1989(84): 502-516.
[33]Kegl B. Learning and design of principal curves[J]. IEEE Trans, Pattern Analysis and Machine Intelligence, 2000(22): 281-297.
[34]Delicado P. Another look at principal curves and surfaces[J]. Journal of Multivariate Analysis, 2001(77): 84-116.
[35]Szekely G, Rizzo M. Measuring and testing independence by correlation distances[J]. The Annals of Statistics, 2007(35): 2769-2794.
[36]Diaconis P, Efron B. Computer-intensive methods in statistics[J]. Scientific American, 1983(248): 116-129.
[37]Gorfine M, Heller R, Heller Y. Comment on 'Detecting Novel Associations in Large Data Sets[EB/OL]. [2014-01-14]. http://www. math. tau. ac. il/~ruheller/Papers/science6. pdf.
[38]Simon N, Tibshirani R. Comment on ' Detecting novel associations in large data sets' by Reshef et, al, Science, Dec 16, 2011 [EB/OL]. [2014-01-14]. http://statweb. stanford. edu/~tibs/reshef/comment. pdf .
[39]Kinney J B, Atwal G S. Equitability, mutual information and the maximal information coefficient[J]. Proceedings of the National Academy of Sciences, 2014, 111(9): 3354-3359.
[40]Heller R, Hellere Y, Gorfine M. A consistent multivariate test of association based on ranks of distances[EB/OL]. [2014-01-14]. http://xxx. tau. ac. il/pdf/1201. 3522v3. pdf .

Survey of Research Process on Statistical Correlation Analysis

Fan Rong[1], Meng Dazhi[2], Xu Dashun[1]
(1. Department of Mathematics, Southern Illinois University, Carbondale, IL 62901, USA;
2. Department of Applied Mathematics, Beijing Poletychnic University, Beijing 100022, China)

Abstract: Correlation analysis is a major research topic in both theoretical statistical study and practical applications. It has been paid more and more attention as the amount of data is increasing significantly. This article reviews several methods that are commonly used, including the Pearson correlation and Spearman correlation developed in 19(upth)century and CorGc and CovGc introduced in 21(upst)century etc. In particular, we include MIC that was proposed in 2011 and its positive and negative comments, aiming at sketching the whole research topic. Methods of correlation analysis themselves play a key role in statistics, especially in analyzing large heterogeneity datasets, such as complex information networks and genome -proteome datasets. This survey tries to provide some understanding of existing methods and their applications. We hope to encourage some new applications, which in turn may promote some new methods developing.

Key words: correlation analysis; Pearson correlation coefficient; Spearman correlation coefficient; Kendall correlation coefficient; mutual information; distance correlation; MIC

作者简介

樊嵘(1979—),女,博士,主要从事概率统计模型及其应用。
徐大舜(1971—),男,副教授,博士,主要从事微分方程理论及其应用。