文章编号:1007 - 2985(2013)04 - 0023 - 03

线性相关系数的一种稳健形式*

甘胜进,林　娟

(福建师范大学福清分校数学与计算机科学系,福建 福清　350300)

摘　要:相关系数是反应 2 个随机变量之间线性关系紧密程度的一个量,容易受异常值干扰.提出一种稳健的形式,它在正态分布条件下的性质与相关系数类似,但是抵御异常值远远优于相关系数,具有很好的应用价值.

关键词:相关系数;中位数;中位数绝对偏差;稳健性

中图分类号:O213　　**文献标志码**:A　　**DOI**:10.3969/j.issn.1007 - 2985.2013.04.006

二维随机变量(X,Y)之间的线性相关系数为

$$\rho_{XY}=\frac{E[(X-E(X))(Y-E(Y))]}{\sqrt{E(X-E(X))^2}\sqrt{E(Y-E(Y))^2}}.$$

由于期望、方差抗离群点(outlier)较差,因此在稳健性统计当中,常以中位数(median)med 代替期望,以中位数绝对偏差(median absolute deviation)MAD来代替方差,其均能抵御 50% 离群点[1],抗异常值干扰能力极强,从而 ρ_{XY} 的一个稳健表达形式为

$$\delta_{XY}=\frac{\mathrm{med}[(X-\mathrm{med}(X))(Y-\mathrm{med}(Y))]}{\mathrm{MAD}(X)\mathrm{MAD}(Y)}.$$

一般情况下,δ_{XY} 的取值并不是像 ρ_{XY} 那样介于$[-1,1]$之间[2],但是对于二维正态分布,δ_{XY} 与 ρ_{XY} 具有相似的性质,笔者主要探讨二维正态分布条件下 δ_{XY} 与 ρ_{XY} 的关系.

1　中位数的性质

命题 1　中位数 $\mathrm{med}(X)=\inf\left\{t \mid p(X\geqslant t)=\frac{1}{2}\right\}$[3],当 X 为离散型随机变量时,$\mathrm{med}(X)$ 可能不存在,但当 X 为连续型随机变量时,$\mathrm{med}(X)$ 存在而且唯一.中位数绝对偏差为 $\mathrm{MAD}(X)=\mathrm{med}(|X-\mathrm{med}(X)|)$.中位数具有如下 3 条性质:

(ⅰ) $\mathrm{med}(aX+b)=a\,\mathrm{med}(X)+b$;

(ⅱ) 当 X 与 Y 相互独立时,$\mathrm{med}(XY)=\mathrm{med}(X)\mathrm{med}(Y)$;

(ⅲ) $\mathrm{med}(|X|)=[\mathrm{med}(X^2)]^{\frac{1}{2}}$.

其中 X 与 Y 均为一维随机变量,a,b 均为任意实数.

证明　(ⅰ) $aX+b\geqslant a\,\mathrm{med}(X)+b\Leftrightarrow X\geqslant\mathrm{med}(X)$,或 $X\leqslant\mathrm{med}(X)$,而 $p(X\geqslant\mathrm{med}(X))=p(X$

* **收稿日期**:2013 - 01 - 18

基金项目:福建师范大学福清分校科研项目(KY2012025);福建省教育厅 A 类科技项目(JA12353)

作者简介:甘胜进(1982 -),男,湖北黄冈人,福建师范大学福清分校数学与计算机科学系助教,硕士,主要从事应用统计研究.

$\leqslant \mathrm{med}(X))=\frac{1}{2}$,即证.

(ⅱ) 当 X 与 Y 相互独立时,有

$$p((X-\mathrm{med}(X))(Y-\mathrm{med}(Y))\geqslant 0)=p(X-\mathrm{med}(X)\geqslant 0,Y-\mathrm{med}(Y)\geqslant 0)+$$
$$p(X-\mathrm{med}(X)\leqslant 0,Y-\mathrm{med}(Y)\leqslant 0)=p(X\geqslant \mathrm{med}(X))p(Y\geqslant \mathrm{med}(Y))+$$
$$p(X\leqslant \mathrm{med}(X))p(Y\leqslant \mathrm{med}(Y))=\frac{1}{2}\times\frac{1}{2}+\frac{1}{2}\times\frac{1}{2}=\frac{1}{2},$$

故 $\mathrm{med}[(X-\mathrm{med}(X))(Y-\mathrm{med}(Y))]=0$. 而由性质(ⅰ)可知,

$$\mathrm{med}[(X-\mathrm{med}(X))(Y-\mathrm{med}(Y))]=\mathrm{med}(XY)-\mathrm{med}(X)\mathrm{med}(Y),$$

故性质(ⅱ)成立.

(ⅲ) 由 $p(|X|\geqslant \mathrm{med}(|X|))=\frac{1}{2}$ 可知,$p(X^2\geqslant[\mathrm{med}(|X|)]^2)=\frac{1}{2}$,故 $[\mathrm{med}(|X|)]^2=\mathrm{med}(X^2)$.

性质(ⅰ)表明中位数具有线性性质,也称为仿射同变性质;性质(ⅱ)表明 2 个相互独立的随机变量的乘积的中位数等于各自中位数的乘积,这与期望的性质是一样的;一般情况下,$\mathrm{med}(|X|)$ 极为难求,而性质(ⅲ)揭示了 $\mathrm{med}(|X|)$ 与 $\mathrm{med}(X^2)$ 之间简单的平方关系.

2 二维正态分布下 δ_{XY} 的性质

根据 δ_{XY} 定义以及上述中位数性质,得到 δ_{XY} 另一种表达式:

$$\delta_{XY}=\frac{\mathrm{med}\left[\left(\frac{X-\mathrm{med}(X)}{\sigma_X}\right)\left(\frac{Y-\mathrm{med}(Y)}{\sigma_Y}\right)\right]}{\mathrm{MAD}\left(\frac{X-\mathrm{med}(X)}{\sigma_X}\right)\mathrm{MAD}\left(\frac{Y-\mathrm{med}(Y)}{\sigma_Y}\right)},$$

其中 σ_X,σ_Y 分别为 X 与 Y 的标准差. 对于二维正态分布随机变量 X 和 Y,有

$$E(X)=\mathrm{med}(X),E(Y)=\mathrm{med}(Y),$$

因此不妨设 $\begin{pmatrix}X\\Y\end{pmatrix}\sim N_2(0,\boldsymbol{\Sigma}),\boldsymbol{\Sigma}=\begin{pmatrix}1&\rho\\\rho&1\end{pmatrix}$. 易证 $\mathrm{MAD}(X)=\mathrm{MAD}(Y)=\varphi^{-1}\left(\frac{3}{4}\right)$,其中 $\varphi(\cdot)$ 为标准正态的分布函数,此时只须计算 $\mathrm{med}(XY)$ 便可得到 δ_{XY} 的值,然而计算 $\mathrm{med}(XY)$ 并非易事[4-5],在此只讨论 δ_{XY} 与相关系数 ρ 之间的关系.

定理 1 在二维正态分布条件下,δ_{XY} 与 ρ 形成一一对应关系,即 $\delta_{XY}=\delta(\rho)$,并且 $\delta(\rho)$ 是 ρ 的增函数,$\delta(-1)=-1,\delta(0)=0,\delta(1)=1$.

证明 $\begin{pmatrix}X\\Y\end{pmatrix}\overset{D}{=}\boldsymbol{\Sigma}^{\frac{1}{2}}\begin{pmatrix}U\\V\end{pmatrix}$[6],其中

$$\boldsymbol{\Sigma}^{\frac{1}{2}}=\begin{pmatrix}\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}&-\frac{1}{\sqrt{2}}\end{pmatrix}\begin{pmatrix}\sqrt{1+\rho}&0\\0&\sqrt{1-\rho}\end{pmatrix}\begin{pmatrix}\frac{1}{\sqrt{2}}&\frac{1}{\sqrt{2}}\\\frac{1}{\sqrt{2}}&-\frac{1}{\sqrt{2}}\end{pmatrix}.$$

U,V 为相互独立的标准正态分布,$\overset{D}{=}$ 表示同分布,则

$$X\overset{D}{=}\left(\frac{\sqrt{1+\rho}}{2}+\frac{\sqrt{1-\rho}}{2}\right)U+\left(\frac{\sqrt{1+\rho}}{2}-\frac{\sqrt{1-\rho}}{2}\right)V,$$
$$Y\overset{D}{=}\left(\frac{\sqrt{1+\rho}}{2}-\frac{\sqrt{1-\rho}}{2}\right)U+\left(\frac{\sqrt{1+\rho}}{2}+\frac{\sqrt{1-\rho}}{2}\right)V,$$
$$XY\overset{D}{=}\frac{\rho}{2}U^2+\frac{\rho}{2}V^2+UV\overset{D}{=}\left(\frac{\rho}{2}+\frac{1}{2}\right)U^2+\left(\frac{\rho}{2}-\frac{1}{2}\right)V^2. \tag{1}$$

(1) 式中最后一个等式是通过$\begin{pmatrix}U\\V\end{pmatrix}\overset{D}{=}\begin{pmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}}\end{pmatrix}\begin{pmatrix}U\\V\end{pmatrix}$得到. 由(1) 式,$-1\leqslant\rho_1<\rho_2\leqslant 1$,则

$$\left(\frac{\rho_1}{2}+\frac{1}{2}\right)U^2+\left(\frac{\rho_1}{2}-\frac{1}{2}\right)V^2\overset{r}{<}\left(\frac{\rho_2}{2}+\frac{1}{2}\right)U^2+\left(\frac{\rho_2}{2}-\frac{1}{2}\right)V^2,$$

$$\operatorname{med}\left[\left(\frac{\rho_1}{2}+\frac{1}{2}\right)U^2+\left(\frac{\rho_1}{2}-\frac{1}{2}\right)V^2\right]<\operatorname{med}\left[\left(\frac{\rho_2}{2}+\frac{1}{2}\right)U^2+\left(\frac{\rho_2}{2}-\frac{1}{2}\right)V^2\right],$$

故 $\delta(\rho_1)<\delta(\rho_2)$. 由(1) 式可知,当 $\rho=-1$ 时,

$$\operatorname{med}(XY)=-\operatorname{med}(V^2)=-(\operatorname{med}(|V|))^2=-\left(\varphi^{-1}\left(\frac{3}{4}\right)\right)^2,$$

故 $\delta(-1)=-1$;当 $\rho=0$ 时,$\frac{1}{2}U^2-\frac{1}{2}V^2\overset{D}{=}\frac{1}{2}V^2-\frac{1}{2}U^2$,即 XY 关于零点为一对称分布,$\operatorname{med}(XY)=0$,$\delta(0)=0$;$\rho=1$ 类似 $\rho=-1$ 的情形讨论.

从证明过程可知,δ_{XY} 跟 ρ_{XY} 关系非常密切,而且性质比较类似. 在实际应用当中,采用样本

$$\hat{\delta}_{XY}=\frac{\operatorname{med}[(X_i-\operatorname{med}(X_i))(Y_i-\operatorname{med}(Y_i))]}{\operatorname{med}(|X_i-\operatorname{med}(X_i)|)\operatorname{med}(|Y_i-\operatorname{med}(Y_i)|)}$$

来估计 δ_{XY},其中 $\{X_i,Y_i\}_{i=1}^n$ 为独立同分布于二维正态的样本,$\hat{\delta}_{XY}$ 抗干扰能力比 $\hat{\rho}=\dfrac{\frac{1}{n}\sum_{i=1}^n(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^n(X_i-\bar{X})^2}\sqrt{\frac{1}{n}\sum_{i=1}^n(Y_i-\bar{Y})^2}}$ 强,因此使用 $\hat{\delta}_{XY}$ 代替 $\hat{\rho}$ 具有更好的实际效果.

参考文献:

[1] PETER J ROUSSEEUW, CHRISTOPHE CROUX. Alternatives to the Median Absolute Deviation [J]. J. Amer. Statist. Assoc., 1993, 88: 1 273 - 1 283.

[2] MICHAEL FALK. On MAD and Comedians [J]. Annals of the Institute of Statistical Mathematics, 1997, 49: 615 - 644.

[3] 孙山泽. 非参数统计讲义 [M]. 北京:北京大学出版社, 2000.

[4] MICHAEL FALK. A Note on the Comedian for Elliptical Distributions [J]. Journal of Multivariate Analysis, 1998, 67: 306 - 317.

[5] STAMATIS CAMBANE, STEEL HUANG, GORDON SIMONS. On the Theory of Elliptically Contoured Distributions [J]. Journal of Multivariate Analysis, 1981, 11: 368 - 385.

[6] MICHAEL FALK. The Sample Covariance is Not Efficient for Elliptical Distributions [J]. Journal of Multivariate Analysis, 2002, 80: 358 - 377.

A Robust Form of the Linear Correlation Coefficient

GAN Sheng-jin, LIN Juan

(Department of Mathematics & Computer Science, Fuqing Branch of Fujian Normal University, Fuqing 350300, Fujian China)

Abstract: Correlation coefficient reflects the closeness of the linear relationship between two random variables, which is susceptible to the interference of abnormal values. This paper presents a robust form, which has the same nature as the correlation coefficient under conditon of the normal distribution, but has higher point breaking down than that of correlation coefficient, so it has good application value.

Key words: correlation coefficient; median; median absolute deviation; robustness

(责任编辑　向阳洁)