Copula 小结及后续

黄冬勃

2015 年 11 月 13 日

1 copula 基础

两个随机变量 X,Y, 其概率密度函数分别为：$F(x)=P[X\le x], G(y)=P[Y\le y]$, F,G 是递增的且 $F(-\infty)=G(-\infty)=0, F(\infty)=G(\infty)=1$。联合概率密度函数为 $H(x,y)=P[X\le x, Y\le y]$。$F(x),G(y),H(x,y)\in[0,1]$, 一对关于 X,Y 的观测值 (x,y), 其各自概率密度函数组成一对数值 $(F(x),G(y))$, 可看做是在单位矩形平面 $[0,1]\times[0,1]$ 上的一个点。$H(x,y)$ 具有以下性质：

1. $H(x,y)$ 是二维递增的；
2. $H(x,-\infty)=H(-\infty,y)=0$, $H(\infty,\infty)=1$；
3. $F(x)=H(x,\infty)$, $G(y)=H(\infty,y)$。

Sklar 定理阐明了,Copula 函数如何描述多变量联合分布和其各自单独边缘分布之间的关系 [1]：

Theorem 1.0.1 (Sklar's theorem)**.** *H 为两个边缘分布分别为 F 和 G 的随机变量的联合分布函数, 那么存在一个 copula 函数 C, 使 $\mathbf{I}^2,(\mathbf{I}=[0,1])$ 中所有 x,y 满足,*

$$H(x,y)=C(F(x),G(y)) \tag{1.0.1}$$

如果 F 和 G 是连续的, 则 C 是唯一的; 否则, C 则被确定在 $RanF\times RanG$ 中。反之, 如果 C 是一个 copula,F 和 G 是分布函数, 那么, 由 eq. (1.0.1) 定义 H 是一个边缘分布为 F 和 G 的联合分布函数。

Copula $C(F(x),G(y))$ 可以将多变量各自边缘分布 $F(x),G(y)$ 与联合分布 $H(x,y)$ 连接起来, 定义域为 $[0,1]\times[0,1]$, 另 $u=F(x), v=G(y)$, 其具有以下性质：

1. C 是 “零基面的 (grounded)”$C(0,v)=C(u,0)=0$;
2. $C(1,v)=v$, $C(u,1)=u$；
3. C 是二维递增 (扩展到 N 个变量则是 N 维递增) 的, 即对于 $u_1,u_2,v_1,v_2\in[0,1]$ 且 $u_1\le u_2, v_1\le v_2$, 有:
$$C(u_2,v_2)-C(u_2,v_1)-C(u_1,v_2)+C(u_1,v_1)\ge 0$$
4. 对于每一个对 (u,v), 另 $W(u,v)=\max(u+v-1,0), M(u,v)=\min(u,v)$, 有
$$W(u,v)\le C(u,v)\le M(u,v). \tag{1.0.2}$$

从 eq. (1.0.1) 可得：

$$C(x,y)=H(F^{-1}(x),G^{-1}(y)) \tag{1.0.3}$$

用例子将 F,G,H,C 关系描述, 已知联合分布函数：

$$H(x,y)=\begin{cases}1-e^{-x}-e^{-y}+e^{-(x+y+xy)}, & x\ge 0, y\ge 0,\\ 0, & \text{otherwise}\end{cases}$$

可求得 $F(x)=H(x,\infty)=1-e^{-x}, G(y)=H(\infty,y)=1-e^{-y}$, 则 $F^{-1}(u)=-\ln(1-u), G^{-1}(v)=-\ln(1-v)$。$C(u,v)=H(F^{-1}(u),G^{-1}(v))$, 可得 $C(u,v)=u+v-1+(1-u)(1-v)e^{-\ln(1-u)\ln(1-v)}$。

2 相关性

对于线性关系的双变量，可以用相关系数

$$\rho(x,y)=\frac{cov(x,y)}{\sqrt{(\mathbf{D})x\times\mathbf{D}y}}$$

来测量，但当两个变量是非线性相关关系时，如 $y=x^2$，协方差 $cov(x,y)=\mathbf{E}(x-\mathbf{E}x)(x^2-\mathbf{E}x^2)=0$，$\rho=0$，则无法测量其相关性。

2.0.1 Kendall's (τ), Pearson's ϱ, and Spearman's rho (ρ)

两个随机变量 X,Y。

Kendall's τ可以对变量之间这种联合的一致性进行采样量化。它描述了两个随机独立均匀分布向量的一致性与非一致性概率的差值 [2]:

$$\begin{aligned}\tau=\tau_{X,Y}&=P[(X_1-X_2)(Y_1-Y_2)>0]-P[(X_1-X_2)(Y_1-Y_2)<0]\\ \tau_{X,Y}&=\tau(C_1,C_2)=4\iint_{\mathbf{I}^2}C_2(u,v)dC_1(u,v)-1\end{aligned}\tag{2.0.1}$$

而若 $u,v=[0,1]$ 且均匀分布，则 $\tau_{X,Y}=4\mathbf{E}[C(u,v)]-1$。

Person's ϱ:

$$\varrho_{X,Y}=\mathrm{cov}(X,Y)/\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}\tag{2.0.2}$$

$\varrho_{j_1j_2}$ 仅适用线性相关的变量，但可以将其经过变换，使 $U=F(X),V=G(Y)$，[3]，U,V 为 $[0,1]$ 的均匀分布，公式2.0.2可变为Spearman's rho (ρ)：

$$\begin{aligned}\rho_{X,Y}&=corr(U,V)\\ \rho_{X,Y}&=\mathrm{cov}(U,V)/(\mathrm{var}(U)\mathrm{var}(Y))^{1/2}\\ &=12E\left[\left(U-\frac{1}{2}\right)\left(V-\frac{1}{2}\right)\right]\\ &=12E[UV]-3\end{aligned}\tag{2.0.3}$$

Spearman's ρ是变换后变量 $F(X)$ 和 $G(Y)$ 的 Pearson 系数 ϱ [2]，它不是根据原始变量来计算，而是通过变量的秩次来计算 ϱ，即得到 ρ。通过关联性测量，当 X,Y 是正（负）关联时，τ,ρ 也是正（负）关联的。当 X,Y 相互独立时，它们为 0。当 X,Y 通过非线性变换后是严格递增的，它们的值保持不变。

τ 和 ρ 是基于秩次的关联性测量值。从文章 [4] 关于相关性结构的统计推断应该总是基于秩次的，因为它们在 X,Y 的单增变换条件下是不变统计量。

在文章 [4]，作者给出了 Spearman's ρ 优于 pearson's ϱ 的几个方面：

1. 当且仅当随机变量 X，Y 是函数相关时，$E(\rho)=\pm1$;
2. 当且仅当 X，Y 是线性相关时，才满足 $E(\varrho)=\pm1$，相比 ρ 更受限制，且
3. ρ 较为通用，对于不同分布都可以估算出一个意义明确的总体参数，然而当 ϱ 应用在“重尾分布”的情形时 (heavy-tailed distributions)，例如柯西分布，是无法得到一个理论上的相关参数的。

3 后续

研究清楚我们所要测量变量的分布特征，获得相关性，并且利用相关性进行整体模型构建。

从二维变量拓展到三维变量情况。

参考文献

[1] B. Ravens, “An introduction to copulas,” *Technometrics*, vol. 42, no. 3, 2000.

[2] R. Montes-Iturrizaga and E. Heredia-Zavoni, “Environmental contours using copulas,” *Applied Ocean Research*, vol. 52, pp. 125–139, 2015. [Online]. Available: <GotoISI>://WOS:000360419100012

[3] M. S. Smith and P. J. Danaher, “Modeling multivariate distributions using copulas: Applications in marketing,” *Marketing Science*, vol. 30, no. 1, pp. 4–21, 2011.

[4] C. Genest and A.-C. Favre, “Everything you always wanted to know about copula modeling but were afraid to ask,” *JOURNAL OF HYDROLOGIC ENGINEERING*, vol. 12, no. 4, pp. 347–368, JUL-AUG 2007, Conference on Copula Modeling in Hydrology, Quebec City, CANADA, MAY, 2004.