An Preliminary Introduction to Copula Theory

Ny *

nymath@163.com

2022 年 11 月 8 日

摘要

Copula 的一些简要介绍**czado2019analyzing**

目录

插图

*我会不定期更新笔记，如果感兴趣的话，可以前往https://github.com/nymath/notes4master获取最新版本。

1 导论

设想我们想要生成一个具有分布函数 F 的随机变量 X，应该如何做呢?

1.1 定理 随机数生成

如果 $U \sim U(0,1)$，即 U 是一个均匀分布。F 为随机变量 X 的分布函数 (单增且右连续), 如果 F 的存在反函数

$$F^{-1} : [0,1] \mapsto \mathbb{R}$$

则 $F^{-1}(U)$ 与 X 同分布。

Remark: X 是一个随机变量, F 是他的 cdf, 则 $F \circ X = F(X)$ 服从 $Uniform(0,1)$。

上述定理告诉我们，想要得到具有分布函数 F 的随机变量的一个样本，我们只需要先模拟一个均匀分布的样本，然后带入 F^{-1} 计算即可。

随机变量容易模拟，那随机向量呢? 模拟随机向量不仅需要模拟边缘分布，还需要模拟相关性结构。从定理 1.1 我们发现，模拟一个随机变量只需要关注这个均匀随机变量 U 即可。类似的，如果我们想模拟随机向量，我们得知道边缘分布函数，以及这些随机变量的依赖性。这使得我们把工作重心转移到了刻画均匀随机向量 $(U_1, \cdots, U_p)$ 的相关性结构上。而Copula，正是用于研究这种相关性结构。

Copula 英文含义为连系动词，目前没有中文翻译（可以把它叫做链接函数，但和 link function 有点冲突），与量化投资中常常提到的 alpha 类似，*Copula* 在不同的场景下有不同的含义，目前我遇到的主要有两种。

1.2 定义 *Copula* 定义 *1*

Copula 是一个累积分布函数，他定义在 $[0,1]^p$ 上，即

$$C : [0,1]^p \to [0,1]$$

Example 1.3 *(常见 Copula)*:

1. 独立 Copula: $C(u_1, \cdots, u_d) = \Pi_{k=1}^{p} u_k$
2. 共单调 (Comonotonicity)Copula: $C(u_1, \cdots, u_d) = \min\{u_1, \cdots, u_d\}$
3. 反单调 Copula():
4. Gaussian Copula:

1.4 定义 *Copula 定义 2*

Copula 还是一个随机向量 $(U_1, \cdots, U_p)$，这个随机向量的联合分布函数是我们上边定义的那个 C，即满足

$$\Pr(U_1 \le u_1, \cdots, U_p \le u_p) = C(u_1, \cdots, u_p)$$

1.5 定理 两种定义的等价性

一方面，给定均匀向量 $(U_1, \cdots, U_p)$，通过下式可以定义一个函数 C

$$C(u_1, \cdots, u_p) = \Pr(U_1 \le u_1, \cdots, U_p \le u_p) = C(u_1, \cdots, u_p)$$

另一方面，如果给定一个联合累积分布函数 C，我们也知道了均匀随机向量的相关性信息。

Remark: 上述两个定义是等价的，以后我们提到 Copula 这个词，需要根据具体场景判断是一个联合累积分布函数，还是一个均匀随机向量。

2 Copula 的一些重要结论

2.1 一般随机向量和 Copula 的关系

可能读到这里会稍有疑惑，我们想模拟一般的随机向量 $(X_1, \cdots, X_p)$, 但现在怎么去模拟均匀向量 $(U_1, \cdots, U_p)$ 了呢？这里我用一个不太严谨的图来表示这种关系

2.1 定理 一些简要解释

我们认为知道联合分布函数的信息后，就能知道随机向量 X 的信息，反之亦然。可以认为

$$\text{联合分布函数的信息} = \text{各个边缘分布的信息} + \text{变量间的相依性结构}$$

而 Copula 正是刻画这种相关性信息。因此只要给定了边缘分布后，我们就能模拟出一般的随机向量 X。不严谨的，我们可以认为

$$\text{联合分布函数} = \text{各个边缘分布函数} + \text{Copula}$$

2.2 定理 *Sklar's Theorem*

给定 Copula(均匀向量)，$(U_1, \cdots, U_p)$，以及边缘分布函数 F_i，则随机向量可以表示为

$$(X_1, \cdots, X_p) = (F_1^{-1}(U_1), \cdots, F_p^{-1}(U_p))$$

而且 $(X_1, \cdots, X_p)$ 的联合累积分布函数 F 等于

$$F(x_1, \cdots, x_p) = C(F_1(x_1), \cdots, F_p(x_p))$$

上述定理用一个严格一些的论述，便是著名的 Sklar's Theorem，他是 Copula 理论的核心，请多加思考这个定理。先给出一个通俗的叫法，

2.3 定义 一些通俗叫法

我们称随机向量 (X, Y) 具有 copula C，如果

$$(F_1(X), F_2(Y))\text{具有累积分布函数 C}$$

2.4 定理 *Sklar's Theorem*

如果随机向量 X 具有联合 cdf，F 以及边缘分布函数 F_i，则 X 具有 Copula C which is define as follows:

$$C(u_1, \cdots, u_p) = F(F_1^{-1}(u_1), \cdots, F_p^{-1}(u_p))$$

2.5 定理 *Copula* 的单调不变性

假如 f, g 是单调递增函数 (事实上单调映射即可), 且 (X, Y) 具有 copula C, 则

$$f(X), g(Y)\text{也具有 Copula C.}$$

Remark: 特别的，当这个单调函数取作随机变量 X, Y 的累积分布函数时，我们得到

$$(U_1, U_2) = (F_X(X), F_Y(Y))$$

也具有 copula C.

上述几个定理告诉我们，只要把均匀随机向量 U 的结构弄清楚了，要模拟出一般的随机向量 X，自然是手到擒来。所以从现在开始，所有的讨论全部集中于均匀随机向量了。

3 Dependence Measures

这里的 Measure 并不是测度论中的测度，就是一个度量指标罢了，对于相依性的度量，我们从相关系数，尾部相依指数展开。

3.1 相关系数

3.1.1 Pearson rho

3.1 定义 总体相关系数

Suppose X, Y are random variables on a probability space $(\Omega, \mathcal{F}, P)$, then the Pearson Rho coefficient of X, Y is defined by

$$\rho_p(X,Y) = \frac{E((X-E(X)(Y-E(Y))}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{\langle E-E(X), Y-E(Y)\rangle}{\|X-E(X)\|_2\|Y-E(Y)\|_2}$$

Remark: 通过 Pearson rho 的定义，我们不难发现，随机变量 X, Y 的相关系数其实是离差变量 X-E(X) 与 Y-E(Y) 之间的夹角。（另外值得注意的是，随机变量 X 的方差其实就是在 L^2 范数诱导的距离意义下，X 到均值 E(X) 距离的平方）

3.2 定义 样本相关系数

Suppose X_i and Y_i are random samples of X and Y, then the sample Pearson rho coefficient is defined by

$$\hat{\rho}_p(X,Y) = \frac{\sum_{i=1}^n (X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^n (X_i-\bar{X})^2}\sqrt{\sum_{i=1}^n (X_i-\bar{X})^2}}$$

Remark: 相当于我们用样本对总体方差和总体协方差进行了一个估计，这个估计应该是一致的。

3.1.2 Spearman rho

3.3 定义 总体相关系数

假如随机变量 X, Y 的边缘分布函数是 F_X, F_Y，那么 X Y 之间的总体 Spearman rho 相关系数定义为 X, Y 诱导的 Copula 的 Pearson rho 相关系数，即 $F_X(X), F_Y(Y)$ 的 Pearson rho 相关系数。

$$\rho_s(X,Y) = \rho_p(F_X(X), F_Y(Y))$$

Remark: 通过总体 Spearman rho 相关系数的定义我们不难发现，由于单调变换不改变 (X,Y) 的 Copula，即 f(X),g(Y) 与 X，Y 有相同的 Copula，自然 f(X), g(Y) 的 spearman rho 系数不变。这说明了 Spearman rho 度量是随机变量生成的 copula 的相关性，并不 Care 他们的边缘分布，也就是说它的确可以度量一些非线性关系。

3.4 定义 样本相关系数

类似的，我们可以利用样本对 X，Y 的 spearman rho 进行估计，

$$\hat{\rho}_s(X,Y) := \frac{\sum_{i=1}^n (r_{i1}-\bar{r}_1)(r_{i2}-\bar{r}_2)}{\sqrt{\sum_{i=1}^n (r_{i1}-\bar{r}_1)^2}\sqrt{\sum_{i=1}^n (r_{i2}-\bar{r}_2)^2}},$$

其中 r_{i1} 代表 X_i 在样本中的 rank。

Remark: 如何理解这个公式是总体 spearman rho 的一个估计？我们可以分子分母同时除以样本容量 n 的平方，把 $\frac{r_{i1}}{n}$ 看作均匀变量 U_1 的一次观测，自然上式就成为了 U_1, U_2 的 Pearson rho 的一个估计。

3.1.3 Kendall tau

3.5 定义 总体相关系数-*kendall-tau*

The Kendall's τ between the continuous random variables X_1 and X_2 is defined as

$$\tau(X_1,X_2) = P((X_{11}-X_{21})(X_{12}-X_{22})>0) - P((X_{11}-X_{21})(X_{12}-X_{22})<0),$$

where (X_{11},X_{12}) and (X_{21},X_{22}) are independent and identically distributed copies of (X_1,X_2).

kendall tau 的系数估计有点复杂，在实际中，由于部分 Copula 的参数和 Kendall 相关系数有一个函数，所有我们可以通过估计 Kendall tau 系数进而得到 Copula 参数的估计值。

3.2 各个相关系数之间的关系

3.6 定理 关联

假设 X,Y 服从二维正态分布，则

$$\rho_p = 2\sin\left(\frac{\pi}{6}\rho_s\right) \text{ and } \tau = \frac{2}{\pi}\arcsin(\rho)$$

Family	Kendall's τ	Range of τ
Gaussian	$\tau = \frac{2}{\pi}\arcsin(\rho)$	$[-1,1]$
t	$\tau = \frac{2}{\pi}\arcsin(\rho)$	$[-1,1]$
Gumbel	$\tau = 1-\frac{1}{\delta}$	$[0,1]$
Clayton	$\tau = \frac{\delta}{\delta+2}$	$[0,1]$
Frank	$\tau = 1-\frac{4}{\delta}+4\frac{D_1(\delta)}{\delta}$ with $D_1(\delta)=\int_0^\delta \frac{x/\delta}{e^x-1}dx$ (Debye function)	$[-1,1]$

3.3 尾部相依性

3.7 定义 上尾相依系数

The upper tail dependence coefficient of a bivariate distribution with copula C is defined as

$$\lambda_U = \lim_{t\to 1^-} P\left(X_2 > F_2^{-1}(t) \mid X_1 > F_1^{-1}(t)\right) = \lim_{t\to 1^-} \frac{1-2t+C(t,t)}{1-t},$$

3.8 定义 下尾相依系数

while the lower tail dependence coefficient is

$$\lambda_L = \lim_{t\to 0^+} P\left(X_2 \le F_2^{-1}(t) \mid X_1 \le F_1^{-1}(t)\right) = \lim_{t\to 0^+} \frac{C(t,t)}{t}.$$

Family	Upper tail dependence	Lower tail dependence
Gaussian	–	–
t	$2t_{\nu+1}\left(-\sqrt{\nu+1}\sqrt{\frac{1-\rho}{1+\rho}}\right)$	$2t_{\nu+1}\left(-\sqrt{\nu+1}\sqrt{\frac{1-\rho}{1+\rho}}\right)$
Gumbel	$2-2^{1/\delta}$	–
Clayton	–	$2^{-1/\delta}$
Frank	–	–
Joe	$2-2^{1/\delta}$	–
BB1	$2-2^{1/\delta}$	$2^{-1/(\delta\theta)}$
BB7	$2-2^{1/\theta}$	$2^{-1/\delta}$
Galambos	$2^{-1/\delta}$	–
BB5	$2-\left(2-2^{-1/\delta}\right)^{1/\theta}$	–
Tawn	$(\psi_1+\psi_2)-\left(\psi_1^\theta+\psi_2^\theta\right)^{1/\theta}$	–
t-EV	$2\left[1-T_{\nu+1}\left(z_{1/2}\right)\right]$	–
Hüsler-Reiss	$2\left[1-\Phi\left(\frac{1}{\lambda}\right)\right]$	–
Marshall-Olkin	$\min\{\alpha_1,\alpha_2\}$	–

4 Bivariate Copula Classes

本节主要讨论两变量的 copula，当然部分结论也适用于多变量的情形。

4.1 Gaussian Copula

本文介绍正态 Copula 的模拟，首先给出正态 Copula 的定义，

4.1 定义 正态 *copula*

假设 Φ 具有协方差矩阵 Σ(对角线为 1) 的 p 维正态分布的联合累积分布函数，φ 则是标准正态分布的累积分布函数，则正态 Copula(Σ) 定义为

$$C(u_1, \cdots, u_p) = \Phi(\varphi^{-1}(u_1), \cdots, \varphi^{-1}(u_p))$$

接下来是正态 Copula 的模拟，事实上，我们只需要先模拟一个具有协方差矩阵 Σ 的多元正态向量，$(Z_1, \cdots, Z_p)$，然后利用

$$(U_1, \cdots, U_p) = (\varphi(Z_1), \cdots, \varphi(Z_p)).$$

即可得到具有协方差矩阵 Σ 的正态 Copula，这里有一个代码可以参考一下，见附件。这里选用了 000651.SZ 和 601318.SH 的简单收益率进行分析，我们绘制了散点图，左图是正态 Copula，边缘分布也为正态 (实际上就是多元正态分布)，右图是正态 Copula，边缘分布为 Gamma 分布，可以发现，利用正态 Copula+Gamma 分布能够更好的拟合极端情形。以然后我用正态 Copula 模拟

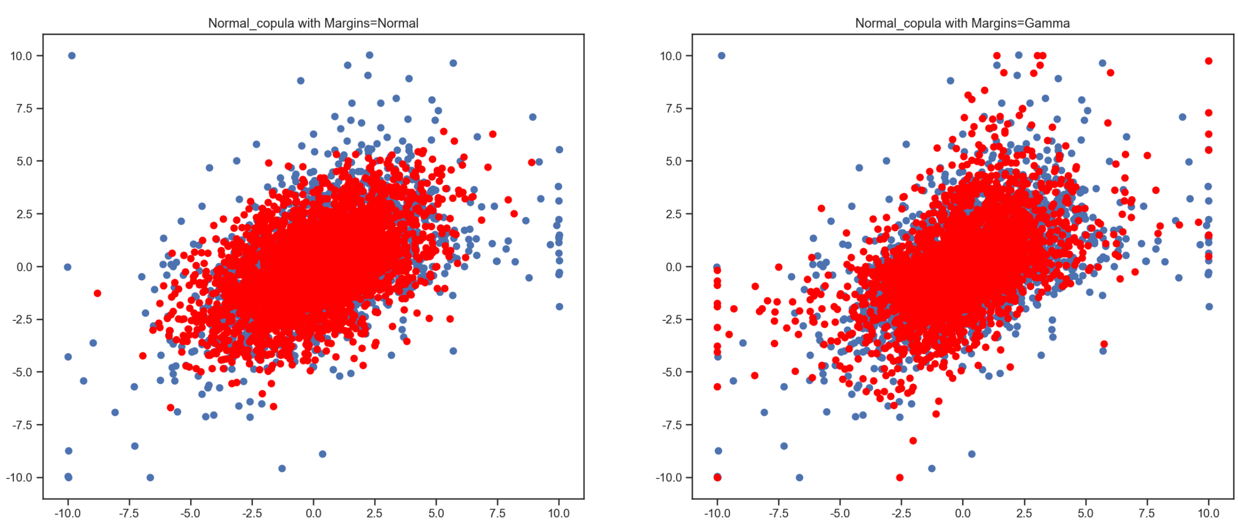

图 1: 正态 Copula 下不同边缘分布的对比

4.2 t Copula

5 Archimedean Copulas

5.1 定义 *Archimedean Copula*

Suppose $\Omega = \{\varphi : [0,1] \to [0,\infty) | \varphi$ is a continuous, strictly monotone decreasing, , and convex function.$\}$ Let $\varphi \in \Omega$, then

$$C(u_1, \cdots, u_p) = \varphi^{[-1]}(\varphi(u_1) + \cdots + \varphi(u_p))$$

is indeed a copula, where $\varphi^{[-1]}$ is φ's pseudo-inverse defined by

$$\varphi^{[-1]}(t) = \varphi^{-1}(t)\chi_{[0,\varphi(0)]}(t)$$