Copula函数

作者：凯鲁嘎吉 - 博客园 http://www.cnblogs.com/kailugaji/

1．Copula介绍

Copula函数把边缘分布函数与联合分布函数联系起来，是研究变量间相依性的一种有效工具。

Copula 函数能够把随机变量之间的相关关系与变量的边际分布分开进行研究，这种思想方法在多元统计分析中非常重要。直观地来看，可以将任意维的联合分布 $H(x_1,\ldots x_n)=P(X_1\le x_1,\ldots X_n\le x_n)$ 分成两步来处理。第一是，对所有的单变量随机变量 X_i，通过累积分布函数(cdf) F_i，我们可以得到随机变量 $U_i=F_i(X_i)$，这是一个均匀随机变量；第二是，随机变量间的相依结构能够通过直接连接这些均匀变量的 n 元 Copula 函数 $C(u_1,u_2,\ldots u_n)$ 来描述。

在一般情形下，n 元 Copula 函数 $C:[0,1]^n\to[0,1]$ 是多元联合分布

$$C(u_1,u_2,\ldots u_n)=P(U_1\le u_1,U_2\le u_2,\ldots U_n\le u_n) \tag{2.3.1}$$

其中 $U_1,U_2,\ldots,U_n$ 是标准均匀变量。

Sklar 使用以下定理给出了 Copula 函数的精确表达式：

定理 2.1 (Sklar 定理)[10] 我们假设 H 是 n 维随机变量 $(X_1,X_2,\ldots,X_n)$ 的联合分布函数，与其对应的边际分布分别是 $F_1,F_2,\ldots,F_n$，那么就存在一个 n 元 Copula 函数 C 使得对于全部 $(x_1,x_2,\ldots,x_n)\in[-\infty,+\infty]^n$，有

$$H(x_1,x_2,\ldots,x_n)=C(F_1(x_1),F_2(x_2),\ldots,F_n(x_n))=C(u_1,u_2,\ldots,u_n) \tag{2.3.2}$$

若 $F_1,F_2,\ldots,F_n$ 是连续的，则 C 唯一；否则 C 仅在 $Ran(F_1)\times\cdots\times Ran(F_n)$ 上唯一。反之，若 C 是一个 Copula 函数，$F_1,F_2,\ldots,F_n$ 是单变量分布函数。则(2.4.2)式定义的 $H(x_1,x_2,\ldots,x_n)$ 是边缘分布为 $F_1,F_2,\ldots,F_n$ 的联合分布函数。

定理给出了一种利用边际分布对多元联合分布建模的方法：

(1)构建各变量的边际分布；

(2)找到一个恰当的 Copula 函数，确定它的参数作为刻画各个变量之间相关关系的工具。

我们假设一个 n 维向量$(X_1, X_2, \ldots, X_n)$的边缘密度函数是 $f(x_i)$，那么联合密度函数 f 就是

$$f(x_1, x_2, \ldots, x_n) = \frac{\partial^n F(x_1, x_2, \ldots, x_n)}{\partial x_1 \partial x_2 \cdots \partial x_n}$$

Copula 函数 C 的密度为

$$c(u_1, u_2, \ldots u_n) = \frac{\partial^n C(u_1, u_2, \ldots u_n)}{\partial u_1 \partial u_2 \cdots \partial u_n}$$

则有

$$f(x_1, x_2, \ldots, x_n) = c(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)) \times \prod_{i=1}^{n} f_i(x_i) \tag{2.3.3}$$

推论： 若 $F_1(x_1), \ldots, F_n(x_n)$均为连续分布函数，设$U_1 = F_1(x_1), \ldots, U_n = F_n(x_n)$，它们是$[0,1]$上的均匀分布，则随机变量$U_1, \ldots, U_n$的联合分布为

$$\begin{aligned} C(u_1, u_2, \ldots u_n) &= H\left(F_1^{(-1)}(u_1), F_2^{(-1)}(u_2), \ldots, F_n^{(-1)}(u_n)\right) \\ &= P(U_1 \le u_1, U_2 \le u_2, \ldots U_n \le u_n) \end{aligned} \tag{2.3.4}$$

其中 $F_i^{(-1)}(u_i)$称为 F_i 的伪逆函数，定义为$F_i(u) = \inf\{x : F(x) \ge u\}$。

实际上，假设 $F_1, F_2, \ldots, F_n$ 是连续的，联合分布函数也已经知道，用(2.4.4)式完全可以构造出相应的 Copula 函数，并且构造出的 Copula 函数完全刻画了变量间的相依结构。

高斯 Copula 是多元正态分布形式的 Copula 函数：

(1)二元高斯 Copula $C(u, v) = \Phi_{\rho}\left(\Phi^{[-1]}(u), \Phi^{[-1]}(v)\right)$ (2.3.5)

(2) n 元高斯 Copula：　$X=(X_1,\ldots,X_p)\sim N_p(0,\Sigma)$

$$C(u_1,\ldots,u_p)=\Phi_\Sigma\left(\Phi^{(-1)}(u_1),\ldots,\Phi^{(-1)}(u_p)\right)$$

$$=\int_{-\infty}^{\Phi^{(-1)}(u_1)}\cdots\int_{-\infty}^{\Phi^{(-1)}(u_p)}\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\{-\frac{1}{2}(x_1,\ldots,x_p)\Sigma^{-1}\begin{pmatrix}x_1\\ \vdots\\ x_p\end{pmatrix}\}dx_1\cdots dx_p$$

其 Copula 密度为

$$c(u_1,\ldots u_p)=\frac{\partial C(u_1,\ldots u_p)}{\partial u_1\cdots\partial u_p}$$

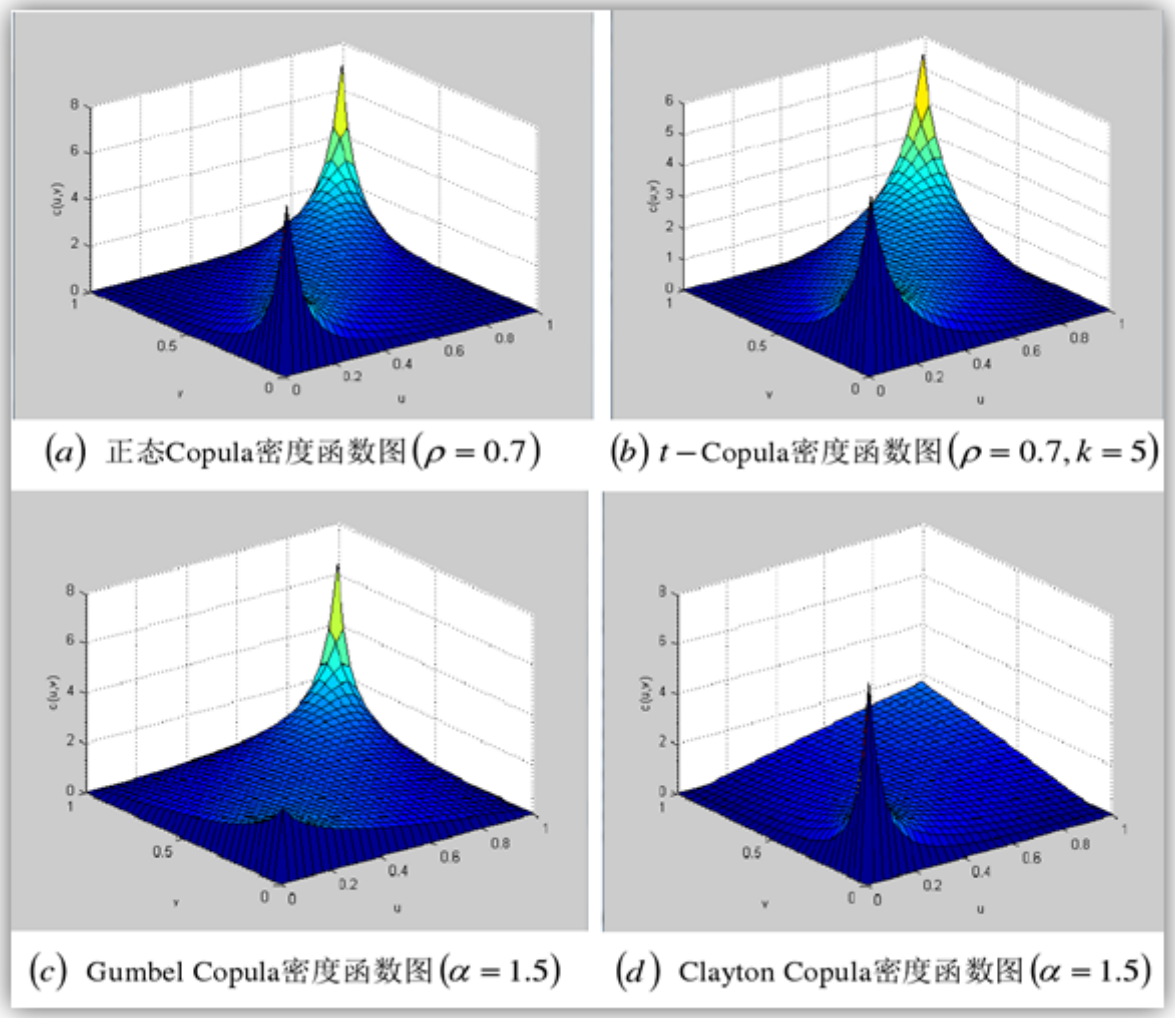
$$=\frac{1}{|\Sigma|^{1/2}}\exp\{-\frac{1}{2}\left[\Phi^{(-1)}(u_1),\ldots,\Phi^{(-1)}(u_p)\right]\Sigma^{-1}\begin{pmatrix}\Phi^{(-1)}(u_1)\\ \vdots\\ \Phi^{(-1)}(u_p)\end{pmatrix}+\frac{1}{2}\sum_{i=1}^{p}\left(\Phi^{(-1)}(u_i)\right)^2\}$$

用 $F(x),G(y)$ 来把 (2.3.5)中的 u,v 换掉，我们可以得到一个二元分布 $H(x,y)$ 且具有高斯相关结构，$H(x,y)$ 的边缘分布为 F,G，不论 F,G 是何种函数，只需要它们是连续的，$H(x,y)$ 都可以独立的看作是高斯 Copula 函数。由此可以看出来，通过 Copula 函数自身就能够了解到随机变量间相关关系。正是如此，Copula 函数又被称为是连接函数[37]。Nelsen 在[10]中继续深入讨论了 Copula 函数。

参考文献：赵梦婷. 高斯Copula过程及其应用[D].华中科技大学,2016.

2．常见的Copula函数（二元）

作为联系边际分布与联合分布的纽带，Copula函数可以选择多种样式，关键取决于随机变量间相关关系符合什么样的类型。Copula函数与边际分布可以分开处理，先通过一定方式获取每一维度上的边际分布，再通过一定方式选取合适的Copula函数，再将两者相乘，即可得到最终的联合分布。



(a) 正态Copula密度函数图$(\rho = 0.7)$ (b) $t-$Copula密度函数图$(\rho = 0.7, k = 5)$

(c) Gumbel Copula密度函数图$(\alpha = 1.5)$ (d) Clayton Copula密度函数图$(\alpha = 1.5)$

3．高斯混合Copula函数

➢ Gaussian Mixture Copula Function

$$F(x_1, x_2, .., x_d; \Theta) = \mathrm{C}(F_1(x_1, \lambda_1), F_2(x_2, \lambda_2), ..., F_d(x_d, \lambda_d); \theta)$$

$$f(x_1, x_2, .., x_d; \Theta) = c(u_1, u_2, ..., u_d; \theta) \cdot \prod_{i=1}^{d} f_i(x_i; \lambda_i)$$

非参数估计：

高斯核密度估计

$$u_j = F_j(x_j)$$

$$c_{gmc}(u_1, u_2, ..., u_d; \Theta) = \frac{\psi(y_1, y_2, ..., y_d; \Theta)}{\prod_{j=1}^{d} \psi_j(y_j)}$$

$$\psi(y_1, y_2, ..., y_d; \Theta) = \sum_{k=1}^{K} \alpha_k \phi(y_1, y_2, ..., y_d; \theta_k)$$

$$= c_{gmc}(u_1, u_2, ..., u_d; \Theta) \cdot \prod_{j=1}^{d} \psi_j(y_j)$$

$$y_j = \Psi_j^{-1}(u_j)$$

$$\max_{\Theta} \; l(\Theta | \{u^{(i)}\}_1^N) = \sum_{i=1}^{N} \log(c_{gmc}(u_1^{(i)}, u_2^{(i)}, ..., u_d^{(i)}; \Theta))$$

参考文献：

[1] Tewari A , Giering M J , Raghunathan A . Parametric Characterization of Multimodal Distributions with Non-gaussian Modes[C]// Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, Vancouver, BC, Canada, December 11, 2011. IEEE, 2011.

[2] 《An Introduction to Copulas》https://files-cdn.cnblogs.com/files/kailugaji/2006_An_Introduction_to_Copulas.rar