一种用于层级数据的新 Copula 回归模型。

Talagb´e Gabin Akpo[a,b], Louis-Paul Rivest[a,*]

[a]数学与统计系，拉瓦尔大学，*2325*，大学街，*G1V0A6*，加拿大魁北克省魁北克市。[b]国家科学研究院，阿尔芒-弗拉皮耶研究中心健康与生物技术部，*531*号草原大道，*H7V 1B7*，加拿大拉瓦尔市

arXiv:2304.10698v1 [stat.ME] 21 Apr 2023

摘要

本文提出了适用于层级数据的多元 Copula 模型。这些模型考虑了两种类型的相关性：一种是同一观测单元上测量的变量之间的相关性，另一种是同一簇内不同观测单元之间的相关性。该模型用于对层级数据进行 Copula 回归，从而得到特定于簇的预测曲线。在一个简单情形中，每个簇包含两个观测单元，且每个观测单元上测量两个变量，此时新模型是在 D-藤框架内构建的。随后，我们重点研究簇中观测单元数量任意、且每个观测单元上测量两个变量的情形。所提出的 Copula 密度具有显式形式，可用三个 Copula 族表示。我们研究了该模型的性质，将其与线性混合模型进行比较，并以若干特殊情形作为结尾。当三个 Copula 族和边缘分布均服从正态分布时，该模型等价于具有随机簇特定截距的正态线性混合模型。我们提出了选择这三个 Copula 族及估计其参数的方法。我们通过蒙特卡洛研究考察了参数估计量的性质。我们使用一个包含多所学校学生成绩的数据集来实施所提出的模型，并将其性能与标准正态混合线性模型进行比较。关键词：可交换性、异质性、正态线性混合模型、藤 Copula。

1. 引言

简单的二元 Copula 回归使用自变量 y 预测因变量 x。该方法首先选择一个二元 Copula，以概括 x 与 y之间的关系。随后，利用所选 Copula 推导出的给定 x 时 y 的条件分布特征，构建 Copula 回归预测器，参见 [17], [14]和 [5] 展示了该方法的应用，而 [18] 推导了其渐近性质， [1] 则研究了预测误差。该方法可以轻松扩展至多元解释变量 x。

本研究旨在将基本的 Copula 回归模型推广到层级数据：x 和y 是在各个簇中的观测单元上观测到的，并且希望在 Copula 回归预测中纳入簇效应。层级数据的经典回归模型是正态线性混合模型，其中包含特定于簇的随机斜率和截距，参见 [2], [25], [16]和 [10]。

*通讯作者。电子邮箱： Louis-paul.rivest@mat.ulaval.ca

预印本已提交至 *Elsevier* *2023*年*4*月*24*日

在此背景下，有人提出了可交换 Copula 族，用于建模簇内的残余依赖，见 [20] 和[11]； [23] 提供了该方法在生存数据中的应用。

首先构建簇中所有x 和y 变量的联合分布模型。第 2 节给出了模型为产生合适的预测必须满足的条件。模型必须满足可交换性假设：对簇中的观测单元进行置换不会改变变量的联合分布。模型还依赖于部分条件独立假设，以确保对某个观测单元y 的预测不依赖于簇中其他观测单元的x 值。随后，在簇包含两个观测单元这一简单情形下，我们在 D-藤中构建所提出的模型。第 3 节介绍了适用于任意大小簇的一般模型。之后，我们通过证明在特定情形下，条件版本等价于 [2]和 [20]的模型，来研究该模型的性质。我们使用所提出的 Copula 进行簇特异性预测。随后，我们将该 Copula 模型应用于 [10] 的数据集，并与标准正态混合线性模型进行比较。

2. 模型构建：可交换性和条件独立性

本节考虑在一个簇中的所有观测单元上测量 d 个变量；下标 j 表示一个观测单元，$j = 1, \ldots, n$，其中 n 是该簇的大小。观测单元 j 的因变量为 Y_j，而 X_j 是对应的 d 个解释变量构成的向量，$Z_j = (X_j^\top, Y_j)^\top$ 则是观测单元 j 上测量的 d 个变量构成的向量。该模型使用 Copula 构建；因此，有必要定义 $F(x) = (F_1(x_1), \ldots, F_{d-1}(x_{d-1}))$ 和 G(y)，它们分别是 X_j 和 Y_j 的边缘分布，并且对所有观测单元都相同。我们令 $U_j = F(X_j)$ 和 $V_j = G(Y_j)$ 为具有均匀边缘分布的随机变量，并令 $c_{d,1:n}\{(u_1, v_1), \ldots, (u_n, v_n)\}$ 为 $\{(U_j^\top, V_j)^\top$ 的联合分布的 Copula 密度，其中 $j = 1, \ldots, n\}$。现在，我们给出关于联合分布族 $F_{d,1:n}(z_1, \ldots, z_n) : n = 2, 3, \ldots$ 的一些条件，以使其能够构建有用的回归模型。

2.1. 可交换性

我们考虑由下式定义的累积分布函数族

$$\mathcal{F}_d = \left\{F_{d,1:n}(z_1, \ldots, z_n) : z_j \in \mathbb{R}^d,\ j = 1, \ldots, n,\ n = 1, 2 \ldots\right\}.$$

若对于所有$n \geq 2$，族 $\mathcal{F}_d$ d-可交换，则称其满足以下条件：$F_{d,1:n} \in \mathcal{F}_d$

i) **置换不变性：** 对于 $\{1, 2, \ldots, n\}$ 的所有置换 $\{\pi(1), \ldots, \pi(n)\}$，

$$F_{d,1:n}(z_1, \ldots, z_n) = F_{d,1:n}\{z_{\pi(1)}, \ldots, z_{\pi(n)}\}. \tag{1}$$

ii）**边缘化闭包性：** 对于任意$r \leq n$,

$$F_{d,1:r}(z_1, \ldots, z_r) = F_{d,1:n}(z_1, \ldots, z_r, \infty, \ldots, \infty). \tag{2}$$

该定义与 [15] 中给出的经典单变量可交换性定义类似。d-可交换性的一个简单例子是具有随机效应的多变量单因素方差分析模型，$\boldsymbol{Z}_j = \boldsymbol{A} + \boldsymbol{E}_j$，其中 $\boldsymbol{A}$ 和 $\boldsymbol{E}_j : j = 1, \ldots, n$ 是 $d \times 1$ 个相互独立的随机向量，且 $\{\boldsymbol{E}_j : j = 1, \ldots, n\}$ 具有相同的分布。下一个命题给出了d-可交换随机向量的相关矩阵形式。

命题 1. 设 $\{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n\}$ 为一组 $nd \times 1$ 随机向量，满足 (1) 和(2)所给出的 d-可交换性定义。则这 n 个向量的 $nd \times nd$ *Pearson*、*Spearman* 和 *Kendall* 相关矩阵具有如下形式

$$I_n \otimes \Sigma_w + J_n \otimes \Sigma_b = \begin{pmatrix} \Sigma_w + \Sigma_b & \Sigma_b & \ldots & \Sigma_b \\ \Sigma_b & \Sigma_w + \Sigma_b & \ddots & \vdots \\ \vdots & \ddots & \ddots & \Sigma_b \\ \Sigma_b & \ldots & \Sigma_b & \Sigma_w + \Sigma_b \end{pmatrix}, \tag{3}$$

其中，J_n 是一个 $n \times n$ 阶全 *1* 矩阵，$\otimes$ 表示克罗内克积。此外，对于 *Pearson* 和 *Spearman* 相关性，矩阵 Σ_w 和 Σ_b 均为半正定矩阵。

2.2. 部分条件独立性

本节提出一个关于每个簇内依赖关系的假设。

定义 1. 假设随机变量 Y_j 和 $\{\boldsymbol{X}_k : k \neq j\}$ 在给定 $\boldsymbol{X}_j$的条件下相互独立。这是一个部分条件独立假设，可写为

$$Y_j \perp \{\boldsymbol{X}_k : \ k \neq j\} | \boldsymbol{X}_j, \ j = 1, \ldots, n. \tag{4}$$

这一条件弱于标准回归模型所依据的条件独立假设，该假设可表述为

$$Y_j \perp \{(\boldsymbol{X}_k, Y_k) : \ k \neq j\} | \boldsymbol{X}_j, \ j = 1, \ldots, n.$$

现在，我们在一个D-藤中，对簇内两个观测单元的联合分布实施可交换性和独立性假设，其中$d = 2$，即$\boldsymbol{Z} = (X, Y)^\top$。

2.3. 当 $d = n = 2$ 时的 *A-D*-藤模型构造

本节为包含两个观测单元的簇中的随机变量构造 Copula 密度函数，并验证其部分条件独立性和 2-可交换性。4 个随机变量为$U_1 = F(X_1)$、$V_1 = G(Y_1)$、$U_2 = F(X_2)$ 和$V_2 = G(Y_2)$。所提出的D-vine 的四棵树如图 1 所示。

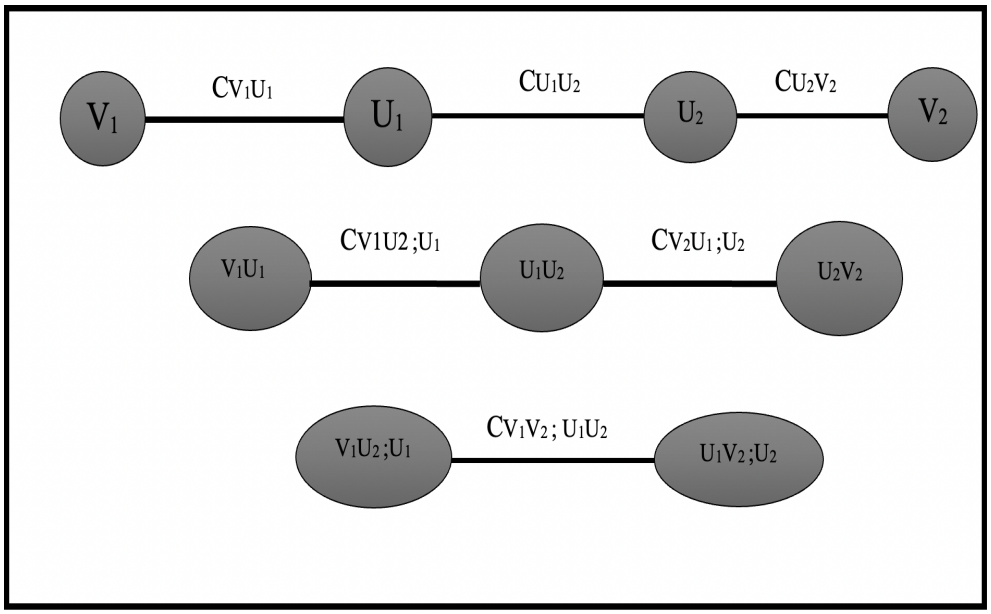


图 1： 所提出的*D*-藤的图示。

该分解涉及 6 个二元 copula：$C_{U_1V_1}$、$C_{U_1U_2}$、$C_{U_2V_2}$、$C_{V_1U_2;U_1}$、$C_{U_1V_2;U_2}$ 以及$C_{V_1V_2;U_1U_2}$。根据 Joe [13, 第 108] 页，可以推导出与图 1 所示树相对应的多元 copula 的密度。其表达式为

$$\begin{aligned} c\{(u_1,v_1),(u_2,v_2)\} &= c_{U_1V_1}(u_1,v_1)c_{U_2V_2}(u_2,v_2)c_{U_1U_2}(u_1,u_2)c_{V_1U_2;U_1}\left\{C_{V_1|U_1}(v_1|u_1),C_{U_2|U_1}(u_2|u_1)\right\} \\ &\times c_{U_1V_2;U_2}\left\{C_{U_1|U_2}(u_1|u_2),C_{V_2|U_2}(v_2|u_2)\right\}c_{V_1V_2;U_1U_2}\left\{C_{V_1|U_1U_2}(v_1|u_1,u_2),C_{V_2|U_1U_2}(v_2|u_1,u_2)\right\}, \end{aligned} \tag{5}$$

其中，条件分布函数$C_{V|U}$ 和$C_{W|UV}$ 定义为

$$C_{V|U}(v|u)=\frac{C_{UV}(u,v)}{\partial u},\quad C_{W|UV}(w|u,v)=\frac{\partial C_{WV;U}\{C_{V|U}(v|u),C_{W|U}(w|u)\}}{\partial C_{V|U}(v|u)}. \tag{6}$$

假设该藤满足简化假设：与第二棵和第三棵树相关的 Copula 不依赖于条件变量。关于这种藤分解的讨论见 [19], [8], 和 [7]。我们希望密度 (5) 满足 2-可交换性的条件，见 (1) 和 (2)，以及部分条件独立性的条件，见 (4)。

若 c {(u_1,v_1), (u_2,v_2)} = c {(u_2,v_2), (u_1,v_1)}，则 Copula 的密度具有 2-可交换性。这要求 U 与 V 之间的依赖关系具有唯一的 Copula 密度，即对于 u, v $\in$ (0, 1)，c_{U_1}V1(u, v) = c_{U_2}V2(u, v)。此外，U_1 与 U_2 之间依赖关系的 Copula 需要是对称的，即 c_{U_1}U2(u, v) = c_{U_1}U2(v, u)，且 c_{V_1}V2;U1U2(u, v) = c_{V_1}V2;U1U2(v, u)。此定义还对条件 Copula 密度 $c_{V_1U_2|}$U1 和 $c_{U_1V_2|}$U2 施加了限制。然而，要使 (5) 满足部分条件独立性条件 (4)，这两个 Copula 必须等于独立 Copula。由此得到以下 Copula 密度

$$c\{(u_1,v_1),(u_2,v_2)\} = c_{U_1U_2}(u_1,u_2)\times\left\{c_{U_1V_1}(u_1,v_1)c_{U_1V_1}(u_2,v_2)\right\}c_{V_1V_2;U_1U_2}\{C_{V_1|U_1}(v_1|u_1),C_{V_1|U_1}(v_2|u_2)\}. \tag{7}$$

其中，U 与 V 之间的关系由任意 Copula 密度 c_{UV} 描述，另有两个对称 copula 密度，$c_{U_1U_2}$ 和$c_{V_1V_2;U_1U_2}$。

考虑式（7）中的三个 Copula 均为正态 Copula 的特殊情形很有意思。令 ρ_1 表示U_1 与U_2之间的相关性，ρ_2 表示U 与V 之间的相关性，ρ_3 为残差相关性。(U_1,U_2,V_1,V_2) 的联合 copula 密度（7）为正态分布，参见 Joe [13,第 119] 页。其相关矩阵为

$$\begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_1\rho_2 \\ \rho_1 & 1 & \rho_1\rho_2 & \rho_2 \\ \rho_2 & \rho_1\rho_2 & 1 & \rho_1\rho_2^2+\rho_3(1-\rho_2^2) \\ \rho_1\rho_2 & \rho_2 & \rho_1\rho_2^2+\rho_3(1-\rho_2^2) & 1 \end{pmatrix}. \tag{8}$$

在（8）中，通过对U_2 进行条件化，得到 $\mathbb{E}\left\{\Phi^{-1}(U_1)\Phi^{-1}(V_2)\right\} = \rho_1\rho_2$；而利用已知（$U_1, U_2$）时$\left\{\Phi^{-1}(V_1),\ \Phi^{-1}(V_2)\right\}$的方差-协方差矩阵，可得 $\rho_3 = \left\{\mathbb{E}\left\{\Phi^{-1}(V_1)\Phi^{-1}(V_2)\right\} - \rho_1\rho_2^2\right\}/(1-\rho_2^2)$】【。

已知（U_1,U_2）服从相关性为 ρ_1 的正态 copula，则当 V_1 和 V_2 定义如下时，(U_1, U_2,V_1,V_2) 的联合分布服从相关矩阵为 (8) 的正态 copula：

$$\Phi^{-1}(V_j) = \rho_2\Phi^{-1}(U_j) + \sqrt{\rho_3}(1-\rho_2^2)^{1/2}A + (1-\rho_2^2)^{1/2}(1-\rho_3)^{1/2}E_j,\ j = 1, 2, \tag{9}$$

其中，A、E_1, E_2 相互独立，且服从 $N(0, 1)$ 分布。模型 (9) 与 [2] 的线性混合模型类似。项 $\rho_2\Phi^{-1}(U_j)$ 表示解释变量对回归的贡献；给定 (U_1, U_2) 时，该项为固定项。簇特异性随机截距 $\sqrt{\rho_3}(1-\rho_2^2)^{1/2}A$ 与实验误差 $(1-\rho_2^2)^{1/2}(1-\rho_3)^{1/2}E_j$ 相互独立。容易证明，在 (9) 中的 $\{\Phi^{-1}(U_1)$、$\Phi^{-1}(U_2)$、$\Phi^{-1}(V_1)$、$\Phi^{-1}(V_2)\}$ 的相关矩阵由 (8) 给出。还要注意，(9) 很容易推广到大小为$n > 2$的簇。一般模型 (7) 也同样如此。这便导出了下一节提出的一般 2-可交换模型。

3. 多变量 2-可交换 Copula 模型

本节提出一种方法，用于构造 $Z_j = (X_j, Y_j)^\top$（$j = 1, \ldots, n$）的联合分布密度，使其满足第 2 节中提出的 2-可交换性和部分条件独立性约束。$\{Z_j : j = 1, \ldots, n\}$的联合累积分布函数记为 $F_{2,1:n}$，其中第一个索引 2 指 Z 的维度，而索引 1 : n 表示该分布涉及从 1 到 n 标记的 n 个观测单元。该联合分布包含 X 和 Y 的边缘分布 F(x) 与 G(y)，以及 Copula 密度 $c_{2,1:n}\{(u_1, v_1), \ldots, (u_n, v_n)\}$。该密度取决于 $c^{(2)}$（描述 X 与 Y 之间关系的 Copula 密度），以及两个 1-可交换 Copula 密度族：$\{c^{(1)}_{1:n}(u_1, \ldots, u_n) : n \geq 2\}$ 和 $\{c^{(3)}_{1:n}(u_1, \ldots, u_n) : n \geq 2\}$，它们分别描述簇内 X 变量与 copula回归残差之间的依赖关系

变量以及簇内的 copula回归残差。这两个密度族分别假定满足条件 (1) 和 (2)，适用于 $d = 1$。

2-可交换copula密度的一般形式为

$$c_{2,1:n}\{(u_1,v_1),\ldots,(u_n,v_n)\} = c^{(1)}_{1,1:n}(u_1,\ldots,u_n)\times\prod_{j=1}^{n}\left\{c^{(2)}(u_j,v_j)\right\}\times c^{(3)}_{1,1:n}\{C_{2|1}(v_1|u_1),\ldots,C_{2|1}(v_n|u_n)\}, \tag{10}$$

其中，u_j、$v_j \in [0,1]$、$j = 1$、.....、n 以及条件分布 $C_{2|1}$ 是通过将式（6）中的 C_{UV} 替换为 $C^{(2)}$推导得到的。对于$n = 2$，式（10）简化为D-藤 Copula 密度（7）。

为了证明 (10) 满足 2-可交换性的条件 (1) 和 (2)，以及条件独立性的条件 (4)，我们对 (10) 中提出的联合密度 $\{(U_1,V_1),\ldots,(U_n,V_n)\}$ 关于 $v_n \in (0,1)$ 进行积分。为此，采用变量替换$w_n = C_{2|1}(v_n|u_n)$较为方便。雅可比矩阵为$dw_n = c^{(2)}(u_n,v_n)dv_n$。利用 Copula 族$C^{(3)}_{1:1,:n}$的边缘化闭包性，该积分等于

$$c^{(1)}_{1,1:n}(u_1,\ldots,u_n)\times\prod_{j=1}^{n-1}c^{(2)}(u_j,v_j)\times c^{(3)}_{1,1:(n-1)}\{C_{2|1}(v_1|u_1),\ldots,C_{2|1}(v_{n-1}|u_{n-1})\}\cdot \tag{11}$$

变量 u_n 仅出现在 $c^{(1)}_{1:1,:n}(u_1,\ldots,u_n)$ 中。利用 Copula族 $C^{(1)}_{1:1,:n}$ 的边缘化闭包性，对 u_n 积分可得到 $(n-1)$ 对 $(U_1,V_1),\ldots,(U_{n-1},V_{n-1})$ 的密度 (10)。因此，(10) 定义了一个适当的 Copula 密度，满足 2- 可交换性的要求 (1) 和 (2)。为证明部分条件独立假设 (4)，对 $v_{n-1},v_{n-2},\ldots,v_2 \in (0,1)$ 对 (11) 积分。通过变量变换 $w_j = C_{2|1}(v_j|u_j)$, $j = 2,\ldots,n-1$即可轻松完成。 $(V_1,U_1,\ldots,U_n)$ 的联合密度为 $c^{(1)}_{1:1,:n}(u_1,\ldots,u_n)c^{(2)}(u_1,v_1)$；因此，给定 U_1,V_1 后，$(U_2,\ldots,U_n)$ 相互独立，且 (4) 成立。

前一段中的推导突显了所提出模型的一项关键性质。若$\{(U_1,V_1),\ldots,(U_n,V_n)\}$的密度为 (10)，则两个向量 $(U_1,\ldots,U_n)$ 和 $\{W_1 = C_{2|1}(V_1|U_1),\ldots,W_n = C_{2|1}(V_n|U_n)\}$ 相互独立，其各自的密度分别为 $c^{(1)}_{1:1,:n}(u_1,\ldots,u_n)$ 和 $c^{(3)}_{1:1,:n}(w_1,\ldots,w_n)$。以下命题对此作了概述。

命题 *2.* 设 $(U_1,V_1),\ldots,(U_n,V_n)$ 是一组由 n 个随机向量组成的集合，其联合密度由 *(10)* 给出。若 $W_j = C_{2|1}(V_j|U_j)$, $j = 1,\ldots,n$，其中分布函数 $C_{2|1}$ 来自 *(6)*，则随机向量 $(U_1,\ldots,U_n)$ 和 $(W_1,\ldots,W_n)$ 相互独立，其各自的密度为 $c^{(1)}_{1:1,:n}$ 和 $c^{(3)}_{1:1,:n}$。

结果　命题 2 的结果给出了以下算法，用于模拟具有式 (10) 所给密度的随机向量：

- 步骤 1：根据可交换 Copula $C_{(1)\,1,1:n}(u_1,\ldots,u_n)$ 模拟 $(U_1,\ldots,U_n)$；
- 步骤 2：根据可交换 Copula W_1、...、W_n，模拟 $(W_1,\ldots,W_n)$，其分布为 $C_{(3)\,1,1:n}(w_1,\ldots,w_n)$；
- 步骤 3：求解由 V_j 定义的方程，其中 $W_j = C_{2|1}(V_j|U_j)$，$j = 1,\ldots,n$，其中 $C_{2|1}$ 由方程 (6) 得出。$(U_1,V_1,\ldots,U_n,V_n)$ 的联合密度由 (10) 给出。

此算法不同于 Czado [6, 第 136] 页中提出的从 D-藤模拟的方法，后者按顺序遍历该藤。

从式（10）构造 $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ 的联合密度是很有意思的。这涉及 $(X_1, \ldots, X_n)$ 的联合边缘密度，

$$g^{(1)}_{1,1:n}(x_1, \ldots, x_n) = c^{(1)}_{1,1:n}\{F(x_1), \ldots, F(x_n)\} \prod_{j=1}^{n} f(x_j) \cdot$$

给定X_j时Y_j的条件密度，适用于$j = 1, \ldots, n$,

$$\prod_{j=1}^{n} \left[g(y_j) c^{(2)} \left\{ G(y_j), F(x_j) \right\} \right],$$

以及簇内残余依赖的一个项:

$$c^{(3)}_{1,1:n}\left[C_{2|1}\{G(y_1)|F(x_1)\}, \ldots, C_{2|1}\{G(y_n)|F(x_n)\} \right] \cdot$$

这突显了 2-可交换模型的逐步构建过程：首先指定X的边缘分布，然后指定给定X 时Y 的条件分布，最后再加入一个分量，以表示簇内的残余依赖。

现在我们假设 Copula$C^{(2)}$ 服从相关性为 ρ_2 的正态分布。(10) 中的条件分布及其逆函数为

$$C_{2|1}(v|u) = \Phi\left\{ \frac{\Phi^{-1}(v) - \rho_2 \Phi^{-1}(u)}{\sqrt{1-\rho_2^2}} \right\} \text{ and } C^{-1}_{2|1}(t|u) = \Phi\left\{ \Phi^{-1}(t)\sqrt{1-\rho_2^2} + \rho_2 \Phi^{-1}(u) \right\},$$

对于 u、v、$t \in [0,1]$，类似结果请参见 [3] 。利用命题 2，簇中一个大小为 n 的簇内单位的因变量 Y_j ，在解释变量 x_j 下，可以表示为 $Y_j = G^{-1}[C^{-1}_{2|1}\{W_j|F(x_j)\}]$，其中 W_j 是一个服从分布 $C^{(3)}_{1\,1\,:n}$ 的向量中的一个分量。由此得到以下线性模型:

$$\Phi^{-1}\{G(Y_j)\} = \rho_2 \Phi^{-1}\{F(x_j)\} + \sqrt{1-\rho_2^2}\Phi^{-1}(W_j) \cdot$$

此外，若X 和Y 的边缘分布为正态分布：$F(y) = \Phi\{(x - \mu_1)/\sigma_1\}$ 和$G(y) = \Phi\{(y - \mu_2)/\sigma_2\}$ ，其中μ_1,μ_2、σ_1^2 、σ_2^2 分别为边缘均值和方差，则模型变为:

$$Y_j = \beta_0 + \beta_1 x_j + \sigma_e \Phi^{-1}(W_j), \tag{12}$$

其中 $\beta_0 = \mu_2 + \beta_1 \mu_1$、$\beta_1 = \rho_2 \sigma_2/\sigma_1$，以及 $\sigma_e^2 = \sigma_2^2(1-\rho_2^2)$。在（12）中，$Y_j$ 的条件边缘分布为正态分布。

联合分布取决于 Copula$C^{(3)}_{1\,1\,,\,:\,n}$。这些模型在 [20]中进行了研究。若该 Copula 为正态且具有可交换性，相关性为 ρ_3，则（12）简化为 [2]的正态混合模型。最后请注意，给定（$X_1 = x_1, \ldots, X x_n$）时，（$Y_1, \ldots, Y_n$）在（12）中的条件密度为

$$f(y_1, \ldots, y_n|x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\sigma_e^n} \exp\left\{ -\frac{1}{2\sigma_e^2} \sum_{j=1}^{n} (y_j - \beta_0 - \beta_1 x_j)^2 \right\} c^{(3)}_{1\,1\,:n}\left[\Phi\{(y_1 - \beta_0 - \beta_1 x_1)/\sigma_e\}, \ldots, \Phi\{(y_n - \beta_0 - \beta_1 x_n)/\sigma_e\} \right] \cdot$$

因此，模型（12）可以通过最大似然轻松拟合。

3.1. 使用 *2*-可交换 *Copula* 模型进行预测

假设在一个簇中观测到 $(n-1)$ 个观测单元 $\{(x_1,y_1),\ldots,(x_{n-1},y_{n-1})\}$。本节研究该簇中给定 Y_n 的条件分布。本文推导了给定 x_n 和$\{(x_1,y_1),\ldots,(x_{n-1},y_{n-1})\}$时$Y_n$ 的条件期望的闭式表达式。文中展示了 Copula$C^{(2)}$的不同设定下的预测曲线示例。

给定x_n和 $\{(x_1,y_1),\ ...,\ (x_{n-1},y_{n-1})\}$时，$Y_n$的条件密度 $g_P(y_n)$ 用$w_j = C_{2|1}\left\{G(y_j)|F(x_j)\right\},\ j=1,\ ...,\ n-1$表示。它等于式（10）与式（11）之比，因此得到

$$g_P(y_n)=g(y_n)c^{(2)}\{F(x_n),G(y_n)\}\frac{c^{(3)}_{1,1:n}\left[w_1,\ldots,w_{n-1},C_{2|1}\{G(y_n)|F(x_n)\}\right]}{c^{(3)}_{1,1:(n-1)}(w_1,\ldots,w_{n-1})}. \tag{13}$$

注意到，给定 $\{(x_1,y_1),\ ...,\ (x_{n-1},y_{n-1})\}$时，$W=C_{2|1}\{G(Y_n)|F(x_n)\}$的条件密度就是 $c^{(3)}_{1:1},:n(w_1,\ldots,w_{n-1},w)/c^{(3)}_{1:1},:(n-1)(w_1,\ldots,w_{n-1})$。这是 $(W_n|W_1,\ldots,W_{n-1})$ 的条件分布，其中 $(W_1,\ldots,W_n)$ 服从 Copula $C^{(3)}_{1:1},:n$。因此，可以很容易地根据式（13）进行模拟。从方程（13）出发，并进行变量替换 $w=C_{2|1}\{G(y_n)|F(x_n)\}$，即可轻松得到该结果。

未知量Y_n的最佳预测量是其条件期望，即给定 $\{x_n、(x_1,y_1)、.....、(x_{n-1},y_{n-1})\}$时的条件期望，其中 $w_j=\left[C_{2|1}\left\{G(y_j)|F(x_j)\right\}\right]$ 可表示为

$$\mathbb{E}_P(Y_n)\quad=\quad\int_0^1 G^{-1}\left[C^{-1}_{2|1}\{w|F(x_n)\}\right]\frac{c^{(3)}_{1,1:n}(w_1,\ldots,w_{n-1},w)}{c^{(3)}_{1,1:(n-1)}(w_1,\ldots,w_{n-1})}dw. \tag{14}$$

如果 Copula $C^{(3)}_{1:1,:n}$ 是独立 Copula，则可简化为$C^{(2)}$的标准无条件 Copula 回归，参见 [14], [5] 或[18]。对 $\mathbb{E}_P(Y_n)$ 关于 $\{(X_1,Y_1)、.....、(X_{n-1},Y_{n-1})\}$ 的分布取期望，也可得到$C^{(2)}$的无条件 Copula 回归曲线。因此，所提出的模型给出了随簇而变化的回归曲线，其期望等于基于$C^{(2)}$的边际 Copula 回归。

我们假设 Copula $C^{(3)}_{1:1},:n$ 是相关性为 ρ_3 的可交换正态分布。因此，若 $W_j=C_{2|1}\left\{G(Y_j)|F(x_j)\right\},\ j=1,\ldots,n$，则随机向量 $\left\{\Phi^{-1}(W_1),\ldots,\Phi^{-1}(W_n)\right\}$ 的累积分布函数是一个具有可交换相关矩阵 $\Sigma(n,\rho_3)$ 的多元正态分布，该矩阵的对角线元素为 1，非对角线元素为 ρ_3。利用多元正态分布的标准性质，给定 $\left\{\Phi^{-1}(W_1),\ldots,\Phi^{-1}(W_{n-1})\right\}$时，$\Phi^{-1}(W_n)$ 的条件分布是均值为 μ_0、方差为σ_0^2 的单变量正态分布，其定义如下

$$\mu_0=\frac{(n-1)\rho_3\sum_{j=1}^{n-1}\Phi^{-1}(W_j)/(n-1)}{1+(n-2)\rho_3},\qquad \sigma_0^2=\frac{(1-\rho_3)\{1+(n-1)\rho_3\}}{1+(n-2)\rho_3}, \tag{15}$$

例如，见 [20]。因此，随机变量 $\Phi^{-1}(W_n)=\Phi^{-1}[C_{2|1}\{G(Y_n)|F(x_n)\}$ 的条件分布为 $N(\mu_0,\sigma_0^2)$。因此，当$C^{(3)}_{1:1,:n}$ 为正态 copula 时，(14) 的最终形式为

$$\mathbb{E}_P(Y_n)=\int_{\mathbb{R}}G^{-1}\left[C^{-1}_{2|1}\{\Phi(\mu_0+\sigma_0 z)|F(x_n)\}\right]\phi(z)dz, \tag{16}$$

其中， $\phi(z) = \exp(-z^2/2)/\sqrt{2\pi}$。利用高斯-厄米特求积法可以很容易地计算式（16）的条件期望，参见 [22]。还需注意，密度函数（13）可以表示为：

$$g_P(y_n) = g(y_n)\frac{c^{(2)}\{F(x_n), G(y_n)\}}{\sigma_0} \times \exp\left\{-\frac{1}{2\sigma_0^2}\left(\Phi^{-1}\left[C_{2|1}\{G(y_n)|F(x_n)\}\right] - \mu_0\right)^2 + \frac{1}{2}\Phi^{-1}\left[C_{2|1}\{G(y_n)|F(x_n)\}\right]^2\right\}. \quad (17)$$

g_P 对应的分位数函数形式简单，即

$$G_P^{-1}(w) = G^{-1}\left(C_{2|1}^{-1}\left[\Phi\{\mu_0 + \sigma_0\Phi^{-1}(w)\}|F(x_n)\cdot\right]\right),$$

因此，Y_n 的条件分布的中位数和各类分位数都很容易求得。

当 n 较大时，$\mu_0 \approx \sum^{n-1} j_{=1} \Phi^{-1}(W_j)/(n - 1)$ 在不同簇之间按照 $N(0, \rho_3)$ 分布变化。这定义了簇特异性回归曲线的可能取值范围。利用公式 (16)，当边缘分布 F 和 G 为标准正态分布时，我们构建了三个 Copula $C^{(2)}$的预测曲线。这些曲线展示在图2中。对于每个 Copula $C^{(2)}$，它们分别对应肯德尔相关系数 τ_2 等于 (0.3, 0.6, 0.9) 的情况。使用了三个 Copula 族 $C^{(2)}$：正态 Copula、Frank Copula 和 Clayton Copula。依赖参数根据 Joe [13, 第 58、166 和 168] 页中的内容推导得出，并使用 R 包 copula 中的 iTau 函数，参见 Ivan Kojadinovic 和 Jun Yan [12]。我们还固定了样本量 $n = 21$ 以及 $C_{1\,1}^{(3)}$ 的 Kendall tau 值：n 为$\tau_3 = 0$，1 对应于 $\rho_3 = 0$，16。针对 $\sum^{n-1} j_{=1} \Phi^{-1}(W_j)/(n - 1)$ 的分位数 $q \in \{1/10、5/10、9/10\}$绘制了预测曲线，其近似分布为 $N(0, \rho_3)$。

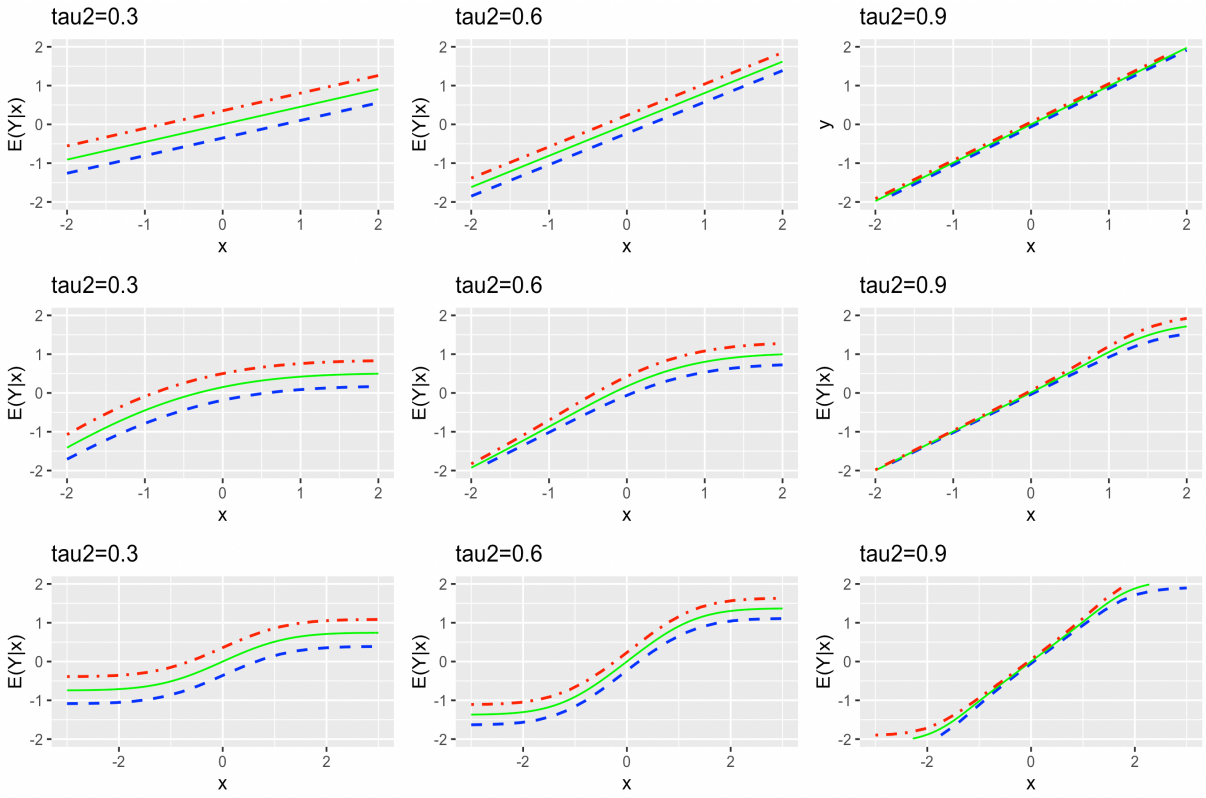


图2： 对于簇间效应的分位数 $q = 1/10$（点线）、$q = 5/10$（实线）和 $q = 9/10$（点划线），采用正态 copula（第一行）、Clayton Copula（第二行）和 Frank Copula（第三行）时的 $\mathbb{E}(Y|x = x_n, (x_1, y_1), \ldots, (x_{n-1}, y_{n-1}))$ 曲线。

这些图展示了$C^{(2)}$ 中的依赖关系对预测曲线的影响。事实上，当 Copula$C^{(2)}$ 为正态 copula 时，如图2所示，我们得到与 [2] 的混合线性回归模型相对应的平行回归直线。对于 Clayton copula，当x取值较小时，回归曲线彼此接近；这可能与该 Copula族的下尾依赖性有关。当相关性 τ_2 增大时，回归曲线趋于一条唯一的直线。这与方程（9）和（12）中给出的模型条件形式一致：随着 ρ_2 增大到 1，模型的随机部分趋于 0。在图2中，我们为$C^{(3)}_{1\,1\,,:n}$ 中的误差依赖参数采用了一个相对较小的值；在许多应用中，例如第5节所述的分析，也会发现这种较小的残余依赖。

4. Copula 选择与 2-可交换模型的参数估计

待分析的数据集为 $\{(X_{ij}, Y_{ij}): i = 1, \ldots, m; j = 1, \ldots, n_i\}$，其中索引$i$表示簇，$j$表示簇内单元。我们假设簇$i$ 中 $2n_i$ 个变量的联合分布由两个边缘分布$F(x|\alpha)$、$G(y|\beta)$以及公式 (10) 中的 Copula 密度决定，该公式涉及一个用于表示 X与在线视频精品$c^{(2)}(u, v; \delta_2)$之间边缘关系的二元Copula密度。

Y 以及两个具有密度函数的 1-可交换 copula 族，$c^{(1)}_{1\,1,\,:n}(u_1,\ldots,u_n;\,\delta_1)$ 和 $c^{(3)}_{1\,1,:n}(w_1,\ldots,w_n;\,\delta_3)$。本节的目标是为该模型的五个分量确定参数族，并估计其参数，这些参数组成向量$\theta=(\alpha,\beta,\delta_1,\delta_2,\delta_3)$。本节遵循 Joe [13] 第 5 章以及 Czado [6] 第 7 章和第 8 章中介绍的模型构建流程。模型的各个组成部分将按照顺序进行选择，并采用本节介绍的技术。

4.1. 模型组成部分的确定

两个边际分布的确定分别独立进行。根据假设观测单元相互独立时计算的赤池信息准则，对竞争模型进行比较。初步估计量，即IFM估计量，见Joe [13, 第5节5],（$\tilde{\alpha}$，$\tilde{\beta}$），通过最大化获得

$$\mathcal{L}_F=\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log\left\{f(x_{ij}|\alpha)\right\}\ \text{ and }\ \mathcal{L}_G=\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log\left\{g(y_{ij}|\beta)\right\}.$$

下一步是计算伪观测值，其定义为

$$\tilde{u}_{ij}=F(x_{ij}|\tilde{\alpha}),\qquad \tilde{v}_{ij}=G(y_{ij}|\tilde{\beta}).\tag{18}$$

随后，使用这$N=\sum n_i$对样本（$\tilde{u}_{ij},\tilde{v}_{ij}$）为（$X$，$Y$）之间的依赖关系选择一个 Copula $C^{(2)}$。需要注意的是，当簇的数量m 趋于无穷大，且簇大小有界时，伪观测值（$\tilde{u}_{ij},\tilde{v}_{ij}$）的二元经验分布函数是$C^{(2)}$ 的一致估计量，其中每个i对应的$n_i<n_{max}$。然而，簇内依赖性会影响估计量的方差，但在这一阶段暂不考虑这一因素，而是采用标准方法，使用 Joe [13] 和 Czado [6], 提出的方法，为二元样本选择 Copula 族$C^{(2)}$。根据 AIC 对竞争模型进行比较，并通过最大化选定 Copula 族的参数的 IFM 估计量 $\tilde{\delta}_2$获得。

$$\mathcal{L}_2=\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log\left[c^{(2)}(\tilde{u}_{ij},\tilde{v}_{ij};\delta_2)\right].$$

为了评估 Copula 族$C^{(1)}_{1\,1\,,:n}$ 和$C^{(3)}_{1\,1\,,:n}$所对应的簇内依赖性，我们采用 Romdhani 等人 [21] 提出的可交换 Kendall's tau 的单位层面版本。该指标通过计算来自不同簇的有序观测值中一致对的比例来评估：$\{(x_{ij},x_{i\ell}),(x_{kr},x_{ks})\}$，这类观测值共有 $\sum_{i>k}n_i(n_i-1)n_k(n_k-1)$ 个可能的配对。Rivest 等人 [20]提出了用于选择可交换 Copula 族的图形方法。还可以根据赤池信息准则比较模型，并通过最大化 δ_1 得到其 IFM 估计量

$$\mathcal{L}_1=\sum_{i=1}^{m}\log\left\{c^{(1)}_{1,1:n_i}(\tilde{u}_{i1},\ldots,\tilde{u}_{in_i};\delta_1)\right\}.$$

Copula族$C^{(3)}_{1\,1\,,\,:\,n}$ 的选择基于伪观测值 $\tilde{w}_{ij}=C^{(2)}_{2|1}(\tilde{v}_{ij}|\tilde{u}_{ij};\,\tilde{\delta}_2)$。在 $\{\tilde{w}_{ij}:i=1,\ldots,m;j=1,\ldots,n_i\}$ 上计算的可交换 Kendall's tau 可用于评估残差的簇内依赖性。IFM

˜估计量 δ_3 通过最大化获得

$$\mathcal{L}_3 = \sum_{i=1}^{m} \log\left\{c^{(3)}_{1,1:n_i}(\tilde{w}_{i1},\ldots,\tilde{w}_{in_i};\delta_3)\right\}.$$

正如 Joe [13, 第5节5] 中所指出的，与 θ 的 IFM 估计相关的五个估计方程可以合并为一个多元估计方程，从而得到 IFM 估计量 $\tilde{\theta}$。Tsiatis [24, 第30], 页中提出的标准渐近理论适用于此处。该理论表明，随着簇的数量 m 趋于无穷大，$\tilde{\theta} - \theta$的联合渐近分布服从一个以零为中心、具有夹心协方差矩阵的多元正态分布。正如预期的那样，该夹心方差估计量考虑了簇内的相关性。

4.2. 参数的最大似然估计

一旦确定了五个模型组成部分的参数族，就可通过最大似然获得参数的最优估计量。本节讨论参数向量 θ的最大似然估计量的性质。θ 的对数似然等于

$$\begin{aligned}
\mathcal{L}(\theta) &= \sum_{i=1}^{m}\sum_{j=1}^{n_i} \log\left\{f(x_{ij}|\alpha)\right\} + \sum_{i=1}^{m}\sum_{j=1}^{n_i} \log\left[c^{(2)}\left\{F(x_{ij}|\alpha), G(y_{ij}|\beta);\delta_2\right\}\right] \\
&+ \sum_{i=1}^{m}\sum_{j=1}^{n_i} \log\left\{g(y_{ij}|\beta)\right\} + \sum_{i=1}^{m} \log\left[c^{(1)}_{1,1:n_i}\left\{F(x_{i1}|\alpha),\ldots,F(x_{in_i}|\alpha);\delta_1\right\}\right] \\
&+ \sum_{i=1}^{m} \log\left[c^{(3)}_{1,1:n_i}\left[C_{2|1}\left\{G(y_{i1}|\beta)|F(x_{i1}|\alpha)\right\},\ldots,C_{2|1}\left\{G(y_{in_i}|\beta)|F(x_{in_i}|\alpha)\right\};\delta_3\right]\right].
\end{aligned}$$

一旦选定模型中两个边际分布和三个 Copula 的参数族，就可以轻松地最大化该函数。由此可得到参数向量的最大似然估计量 $\hat{\theta}$。这一最大化过程通过优化器完成，例如 R 函数 *optim* 或 *nlminb*。在 $\hat{\theta}$ 处求值的 $\mathcal{L}(\theta)$ 的海森矩阵的相反数，即为该模型的观测费舍尔信息。其逆矩阵是 $\hat{\theta} - \theta$ 的渐近协方差矩阵。它可用于计算所有已估计参数的标准误估计。在此阶段，还可以通过似然比检验比较 $C^{(2)}$ 的嵌套候选参数族，从而验证 IFM 模型选择步骤的合理性，该步骤忽略了簇内依赖性。

当边缘分布$F(x)$、$G(y)$和 Copula$C^{(2)}$ 均为正态分布时，似然可拆分为$F(x)$和$C^{(1)}_{1\,1,:\,n}$ 个观测的参数边缘似然，以及回归参数（β_0、β_1、σ_e^2）和$C^{(3)}_{1\,1,:\,n}$个观测的参数条件似然，见公式（12）。 [2] 的标准正态混合线性模型属于这一类，因为其参数是通过条件似然估计的。一般而言，这些参数以复杂的方式相互交织，其估计依赖于对数似然$\mathcal{L}(\theta)$。

4.3 IFM 与 *ML* 估计量抽样性质的蒙特卡洛研究

对于参数为 θ的 2-可交换 Copula 模型，$\hat{\theta}$ 和 $\tilde{\theta}$ 分别是最大似然估计量和 IFM 估计量。本节研究当m 有限时，它们的抽样性质。具体而言，通过蒙特卡洛模拟，将估计量 $\hat{\psi}$ 的期望和方差近似为

$$\mathbb{E}_B(\hat{\psi}) = \frac{1}{B}\sum_{b=1}^{B}\hat{\psi}_b, \quad \mathbb{V}_B(\hat{\psi}) = \frac{1}{B-1}\sum_{i=1}^{B}\left\{\hat{\psi}_b - \mathbb{E}_B(\hat{\psi})\right\}^2,$$

其中，b 表示在$B = 1000$ 次蒙特卡罗模拟中获得的估计值。IFM 估计量 $\tilde{\psi}$ 的期望和方差也采用类似的方法进行评估。

在这项针对 Copula 2-可交换模型的蒙特卡洛研究中，边缘分布F 和G 均为均值为 $\mu_1 = \mu_2 = 0$ 、方差为 $\sigma_1^2 = \sigma_2^2 = 1$ 的正态分布。可交换 Copula$C^{(1)}_{1\,1,:n}$ 和 $C^{(3)}_{1\,1,:n}$ 属于正态族，其相关性分别为 $\rho_1 = 0.31$ 和 $\rho_3 = 0.16$，对应的 Kendall 秩相关系数分别为 0.2 和 0.1。在实际应用中，通常会发现如此低水平的簇内关联。在这项研究中，相关性采用其 logit 值进行参数化，

$$\eta = \log\left(\frac{\rho}{1-\rho}\right). \tag{19}$$

研究了三种 Copula$C^{(2)}$：正态 copula、Clayton Copula 和一种具有两个参数的 Khoudraji copula [9]，该 copula 描述了U 与V之间的非对称关系。其定义为

$$C^{(2)}(u, v; \rho, \kappa) = u^{1-\kappa}C_\rho(u^\kappa, v), \quad u, v \in [0, 1], \tag{20}$$

其中，C_ρ 是一个相关性为 $\rho \in (0, 1)$ 的正态 copula，$\kappa \in (0, 1)$ 是非对称参数。在模拟中，κ 通过其 logit 进行参数化，参见（19）。考虑了 copula$C^{(2)}$的 Kendall's tau τ_2 的两个取值：0.4 和 0.6。研究了两种样本量，即 $m = 10$ 和 50；其对应的簇大小分别为 $n = 30$ 和 $n = 18$ 。表 1 至表 2 展示了 copula 参数估计量的期望和方差。

表 1：当$C^{(2)}$为正态 copula 时，估计量的期望值及其方差（括号内数字乘以 10）。

$\tau_2(\eta_2)$	m	方法	$\eta_1 = -0.80$	η_2	$\eta_3 = -1.69$
0.4(0.35)	10	MV	-0.96(1.90)	0.35(0.37)	-1.81(2.29)
		IFM	-1.00(2.20)	0.34(0.62)	-1.89(2.41)
	50	MV	-0.85(0.58)	0.35(0.12)	-1.74(0.75)
		IFM	-0.86(0.63)	0.35(0.16)	-1.76(0.80)
0.6(1.44)	10	MV	-0.90(1.97)	1.45(0.31)	-1.83(2.24)
		IFM	-0.95(2.30)	1.44(0.41)	-1.93(2.50)
	50	MV	-0.84(0.53)	1.44（0.09）	-1.73(0.70)
		IFM	-0.85(0.60)	1.44(0.11)	-1.75(0.76)

表 2：估计量的期望值及其方差乘以 10 后的数值（括号内），当$C^{(2)}$为 Clayton copula 时 la.

$\tau_2(\delta_2)$	m	方法	$\eta_1 = -0.80$	δ_2	$\eta_3 = -1.69$
0.4(1.33)	10	MV	-0.88(1.63)	1.35（0.66）	-1.82(2.36)
		IFM	-0.98(2.18)	1.33(0.84)	-1.91(2.68)
	50	MV	-0.83(0.46)	1.33(0.15)	-1.73(0.67)
		IFM	-0.85(0.60)	1.32(0.19)	-1.75(0.73)
0.6(3)	10	MV	-0.88(1.48)	3.02(2.39)	-1.80(2.48)
		IFM	-0.98(2.11)	2.96(3.13)	-1.91(2.56)
	50	MV	-0.82(0.45)	3.01(0.61)	-1.73(0.71)
		IFM	-0.86(0.59)	2.99(0.77)	-1.75(0.77)

表 3： 当$C^{(2)}$为 Khoudraji copula 时，估计量的期望及其方差（乘以 10 后，列于括号内）。

τ_2	m	方法	$\eta_1 = -0.80$	$\eta_\rho = 0.75$	$\eta_\kappa = 1.52$	$\eta_3 = -1.69$
0.4	10	MV	-0.90(1.83)	0.80(0.55)	1.51(2.42)	-1.85(2.13)
		IFM	-0.95(2.08)	0.75(0.65)	1.49(3.25)	-1.89(2.49)
	50	MV	-0.82(0.51)	0.77(0.12)	1.51(0.98)	-1.73(0.63)
		IFM	-0.83(0.58)	0.76(0.18)	1.51(1.44)	-1.74(0.79)
τ_2	m	方法	$\eta_1 = -0.80$	$\eta_\rho = 1.45$	$\eta_\kappa = 3.48$	$\eta_3 = -1.69$
0.6	10	MV	-0.86(1.77)	1.49(0.36)	3.39(5.30)	-1.81(2.20)
		IFM	-0.94(2.09)	1.45(0.38)	3.62（6.34）	-1.88(2.74)
	50	MV	-0.81(0.49)	1.47(0.11)	3.48(2.14)	-1.77(0.73)
		IFM	-0.83(0.63)	1.47(0.14)	3.50(2.79)	-1.79(0.86)

在表1至表3中，所有估计量的偏倚都可以忽略不计。讨论重点在于方差。正如预期，U-V 关联在$C^{(2)}$ 中的强度不会影响两个估计量对于 η_1 的精度。对于 Copula$C^{(2)}$ 的参数而言，IFM估计量的精度损失大于其他 2 个 Copula 族参数的精度损失。最大似然估计量在$m = 10$ 处的效率高于在$m = 50$处的效率。补充材料提供了簇内不等样本量 $\{n_i\}$ 的更多模拟结果。不等样本量会导致所有估计量的精度略有损失，尤其是在$m = 10$处。一般而言，这种损失对于 IFM 估计量而言比对于最大似然估计量而言更为显著。总体而言，两种估计方法给出了相似的结果；这支持第 4 节提出的建议，即使用 IFM 估计量来选择所提出 Copula 模型的分量。详细的模拟结果，包括对 $k = 1$ 为2时 (μ_k, σ_k) 的估计量抽样性质的介绍，见补充材料。

5. 使用 2-可交换 Copula 模型对数学成绩建模

本节重新考察 [10] 中讨论的一个数据集。该数据集涉及四年级和七年级学生的数学成绩，共测量了$N = \sum n_i = 728$ 名学生，来自$m = 48$ 所小学。簇样本量n_i 在 4 到 40 之间。四年级成绩（X）和七年级成绩（Y）的取值范围均为 0 到 40。我们通过变换$M \to (M + 1/2)/41$ 将其映射到 (0,1) 区间。为打破成绩中的并列值，我们给每个成绩都加上了一个微小的随机扰动。

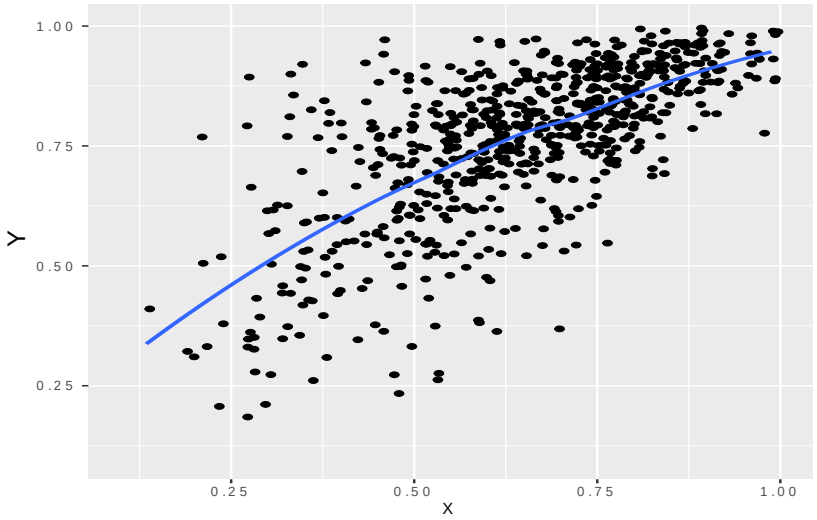


图3： Y 与 X 的散点图及其平滑曲线。

本分析的一个目标是构建给定X 时Y 的预测模型，其支撑集为 $(0,1)$。另一个目标是研究 2-可交换 Copula 模型能否捕捉图3中呈现的非线性；图3展示了$N = 728$ 个数据点的散点图及一条平滑曲线。我们的目标是将基于 Copula 开展的分析与 [10] 中报告的分析进行对比，后者基于标准正态线性混合模型。

5.1. X 和 Y 的边缘分布选择

F 和G 的候选分布族为贝塔分布（记作 $\mathcal{B}$）和广义贝塔分布（记作 $\mathcal{GB}3$），见 [4]。事实上，根据该分布直方图的构造，我们得到一个非对称分布。若X 服从 $\mathcal{B}(\alpha,\beta)$ 分布，则$Y = X/\{\lambda + (1-\lambda)X\}$ 在 $\lambda \in (0,1)$ 上服从一个 $\mathcal{GB}3(\alpha,\beta,\lambda)$ 分布，其密度函数为

$$\frac{\lambda^{\alpha}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{y^{\alpha-1}(1-y)^{\beta-1}}{\{1-(1-\lambda)y\}^{\alpha+\beta}}, \quad 0 \le y \le 1.$$

表4比较了这两个分布对两个边际分布的拟合情况。如第4.1节所述，这项初步分析未考虑簇内依赖性。因此，*pse*（伪标准误）忽略了课堂内依赖性。

表4： $\mathcal{B}$和 $\mathcal{GB}3$ 分布对两个边际分布的拟合情况。

变量	模型	$\tilde{\theta}$	pse	AIC
X	$\mathcal{B}$	(4.280,2.332)	(0.222,0.115)	-542.88
	$\mathcal{GB}3$	(4.280,2.330,1)	(0.222,0.115,NA)	-540.80
Y	$\mathcal{B}$	5.24,1.79	(0.28,0.08)	-789.57
	$\mathcal{GB}3$	(2.616,2.319,0.29)	(0.303,0.241,0.069)	-826.96

拟合效果最佳的模型分别是适用于X 的贝塔分布和适用于Y的广义贝塔分布。

5.2. 双变量 *Copula* 的选择$C^{(2)}$

第一步是计算式（18）中定义的 $\tilde{u}_{ij}$ et $\tilde{v}_{ij}$ 在 $j = 1, \ldots, n_i$ 和 $i = 1, \ldots, m$ 对应的伪观测值。Kendall 秩相关系数为 0.49（$pse = 0.03$），因此两个变量之间存在较强的关联。根据 [13, chap. 1], ，计算单位正方形四个象限中各子样本的 Kendall 秩相关系数。结果表明，较高等级之间的关联强于较低等级之间的关联。此外，左上象限与下方象限的 Kendall 秩相关系数相差 0.1，这表明在剔除边缘分布的影响后，图3中观察到的部分不对称性仍然存在。

使用 R 包 VineCopula 中的 BiCopEst 函数和 copula 包中的 fitCopula 函数，对该双变量样本拟合了若干种 Copula。为了捕捉数据中的非对称性，我们使用 Khoudraji 装置（见公式（20））构造了非对称替代模型。表5中拟合效果最好的 Copula 是由以下形式给出的生存型 Khoudraji 正态 copula。

$$C^{(2)}(u, v|\rho_2, \kappa_1, \kappa_2) = u + v - 1 + (1-u)^{1-\kappa_1}(1-v)^{1-\kappa_2} C_{\rho_2}\{(1-u)^{\kappa_1}, (1-v)^{\kappa_2}\},$$

其中，C_{ρ_2} 是相关系数为 ρ_2 的二元正态 copula。该三参数族中 copula 的密度可以通过 copula 函数进行计算。条件分布 $w = \partial C^{(2)}(u,v|\rho_2,\kappa_1,\kappa_2)/\partial u$ 具有以下显式形式。

$$w = 1 - (1-\kappa_1)(1-u)^{-\kappa_1}(1-v)^{1-\kappa_2} C_{\rho_2}\{(1-u)^{\kappa_1}, (1-v)^{\kappa_2}\} - \kappa_1(1-v)^{1-\kappa_2}\Phi\left[\frac{\Phi^{-1}\{(1-v)^{\kappa_2}\} - \rho_2\Phi^{-1}\{(1-u)^{\kappa_1}\}}{\sqrt{1-\rho_2{}^2}}\right]. \quad (21)$$

在表5中，Survival Khoudraji-Normal2 指的是通过设置 $\kappa_2 = 1$得到的该 Copula 的双参数版本。

表5：（*u*,*v*）关系中若干 Copula 的参数估计值、伪标准误 *pse* 及 AIC。

Copula $C^{(2)}$	$\tilde{\theta}$	*pse*	AIC
正态	0.682	0.016	-454.32
生存-Gumbel	1.892	0.057	-459.5
生存 Khoudraji-正态 copula	(0.791, 0.822, 0.960)	(0.023, 0.043, 0.029)	-474.10
生存 Khoudraji-Normal2	(0.762, 0.837)	(0.021, 0.045)	-468.64

5.3. 可交换 *Copula* 族 $C^{(1)}_{1\,1,:n}$ 和 $C^{(3)}_{1\,1,:n}$

X 的可交换 Kendall's tau 和 Y 的可交换 Kendall's tau 分别为 0.046（标准误 = .017）和 0.095（标准误 = 0.035），表明第七学年的学校效应更强。通过最大化伪对数似然 $\mathcal{L}_1$ 和 $\mathcal{L}_3$，比较了采用式（21）评估的若干 Copula 族对 $\{\tilde{u}_{ij}\}$ 和 $\{\tilde{w}_{ij}\}$的联合分布的拟合效果。结果见表 6。对于$C^{(1)}_{1\,1,:n}$和$C^{(3)}_{1\,1,:n}$，正态族均为最佳选择。

表 6：三类 $C^{(1)}_{1\,1,:n}$ 和 $C^{(3)}_{1\,1,:n}$ 族的参数估计值、伪标准误 *pse* 及赤池信息准则（AIC）。

	$C^{(1)}_{1,1:n}$			$C^{(3)}_{1,1:n}$		
族	$\tilde{\theta}$	*pse*	AIC	$\tilde{\theta}$	*pse*	AIC
Frank	0.251	0.138	-5.69	0.935	0.179	-69.76
Gumbel	1.040	0.022	-4.64	1.109	0.024	-66017
正态	0.064	0.025	-12.76	0.167	0.035	-80.03

为完成分析，我们对所提出模型中各分量的 10 个参数进行完整的最大似然估计。κ_2 的参数估计值非常接近 1。我们首先使用完整似然对H_0：$\kappa_2 = 1$ 进行似然比检验。结果为 $\chi^2_{1,\,obs} = 8.2$，对应的 p 值为 0.4%。因此，包含 10 个参数的完整模型是确定的；参数估计值及其标准误（*se*）列于表 7。

表 7： 完整模型参数的最大似然估计量。

分量	$\hat{\theta}$	*se*
F ($\mathcal{B}$)	$(\hat{\alpha}_1, \hat{\beta}_1)$ = (4.271, 2.359)	(0.235, 0.124)
G ($\mathcal{GB}3$)	$(\hat{\alpha}_2, \hat{\beta}_2, \hat{\lambda})$ = (2.457, 2.470, 0.248)	(0.245, 0.254, 0.052)
$C^{(1)}_{1,1:n}$ (正态分布)	$\hat{\rho}_1 = 0.063$	0.026
$C^{(2)}$ （生存 Khoudraji-正态）	$(\hat{\rho}, \hat{\kappa}_1, \hat{\kappa}_2)$ = (0.795, 0.822, 0.959)	(0.024, 0.046, 0.029)
$C^{(3)}_{1,1:n}$ （正态）	$\hat{\rho}_3 = 0.161$	0.040

表 7 总结的最终模型拟合情况通过两所学校加以说明：编号为 1 的学校，$n_1 = 19$；以及编号为 30 的学校，$n_{30} = 31$。它们各自的 μ 值，见式（15），分别为 −0.464 和 0.810；这意味着，对于相同的 math4 成绩，学校 30 的 math7 成绩高于学校 1。这一点可从图4中看出，图中给出了根据式（16）构造的两条回归曲线；这些曲线使用高斯-厄米特求积法近似每个x值对应的正态积分。

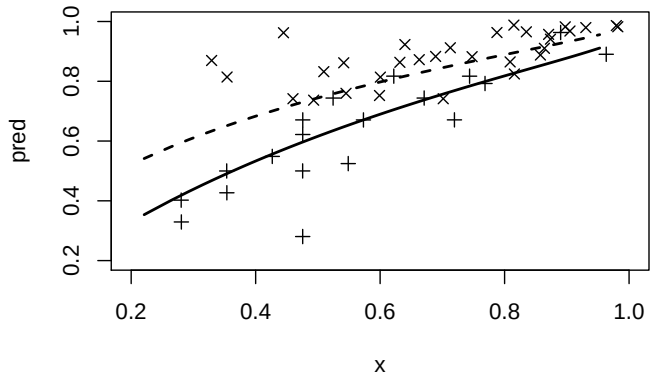


图4： 学校1（+字符和实线）与学校30（× 字符和虚线）的散点图和回归曲线).

拟合的 Copula 模型的条件残差可定义为$y_{ij} - \hat{y}_{ij}$，其中$\hat{y}_{ij}$ 是根据式（16）在x_{ij}处计算的预测值；μ_0 和 σ_0的取值（见式（15））对应于学校i。Copula 模型（*CM*）的条件残差可与混合线性模型的条件残差进行比较。第一个模型*ML*1具有随机学校截距，而第二个模型*ML*2具有可能相关的随机斜率和截距。*CM*、*ML*1和*ML*2的残差均方误差和残差四分位距（IQR）分别为（0.0107，0.0112，0.0101）和（0.1082，0.1186，0.1098）。因此，与分析结果一致

表5中报告的*ML*1拟合效果较差。*CM* 和*ML*2的拟合效果非常相似，因为后者捕捉到了图4中可见的学校间斜率变化。*CM* 的残差均方误差较大；但其残差四分位距较小，并且在53%的数据点上，其绝对残差小于*ML*2。

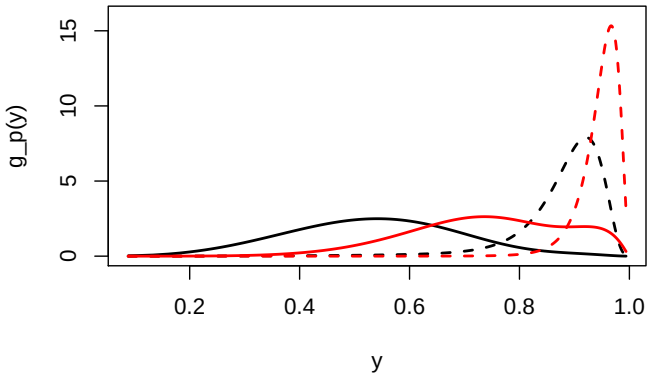


图5： 学校1（黑色实线）和学校30（红色虚线）中，Y 在$X = 0.4$ 和$X = 0.9$ 时的预测密度函数。

2-可交换 Copula 模型的主要优势在于，它使Y 的预测区间能够同时取决于学校和X值。图5对此进行了说明，其中给出了学校1和学校30中$X = 0.4$ 与$X = 0.9$ 时，成绩Y的预测密度（由公式(17)给出）。在$X = 0.4$ 时以及学校1中，Y 的变异性更大。这可以从表8中给出的Y 的相应95%预测区间看出。

表8： 学校1和学校30中$X = 0.4$ 与$X = 0.9$ 时的95%预测区间。

学校	$X = 0.4$	$X = 0.9$
1	(0.23, 0.89)	(0.66,0.97)
30	(0.44,0.97)	(0.86,0.99)

6. 结论

本文提出的 2-可交换 Copula 模型提供了灵活的方法，可在层级数据集已知 Y 的情况下预测变量 X 。第 4 节重点介绍了所提方法的一个关键特征：给定 X 时，Y 的预测密度具有灵活性。其形状和支持集可取决于已知的 X 值、Y 以及簇。一个突出的

问题在于，所提出的模型能否推广到两个或更多连续解释变量的情形。这种推广的关键在于，能否获得灵活的（$d-1$）-可交换族$C^{(1)}_{d-1,1:n}$，用于描述簇中所有单元的解释变量的联合分布。可以使用由（3）给出相关矩阵的椭圆 Copula。这可以部分确定单元内$C^{(2)}$的X变量和Y变量的联合分布。还可以考虑使用藤分解来完成$C^{(2)}$的设定。这些问题将成为未来研究的对象。

补充材料

补充材料包含命题的证明以及蒙特卡洛模拟研究的其他结果。

致谢及其他内容

我们谨此感谢 ´Etienne Marceau 对本研究先前版本所作的认真审阅。

资金支持

谨此致谢加拿大自然科学与工程研究委员会提供的支持。

利益披露声明

作者声明，就本文的研究、撰写和/或发表而言，不存在任何潜在的利益冲突。

数据可用性

如有需要，可提出申请获取数据。

参考文献

[1] E. F. Acar, P. Azimaee, M. E. Hoque，Copula 模型的预测评估，《加拿大统计学杂志》，47（2019年）8–26。[2] G. E. Battese, R. M. Harter, W. A. Fuller，利用调查和卫星数据预测县级作物面积的误差分量模型，《美国统计学会会刊》，83（1988）28–36。[3] C. Bernard, C. Czado，条件分位数与尾部依赖，《多元分析杂志》，138（2015）104–126。[4] W. M. Cockriel, J. B. McDonald，两个多元广义贝塔分布族，《统计学通讯——理论与方法》，47（2018）5688–5701。

[5] G. J. Crane, J. Van der Hoek，Copula 的条件期望公式，《澳大利亚与新西兰统计学杂志》50（2008年）53–67。[6] C. Czado，《使用藤状 Copula 分析相关数据：R 实用指南》，2019年。[7] C. Czado, T. Nagler，基于藤状 Copula 的建模，《统计学及其应用年度评论》，9（2022年）453–477。[8] J. Dissmann, E. C. Brechmann, C. Czado, D. Kurowicka，正则藤状 Copula 的选择与估计及其在金融收益中的应用，《计算统计与数据分析》，59（2013年）52–69。[9] C. Genest, K. Ghoudi, L.-P. Rivest，对“通过 Copula 理解关系”（E. Frees 和 E. Valdez 著）的讨论，《北美精算学杂志》，3（1998年）543–552。[10] H. Goldstein，多层统计模型，Wiley，第2版，2011年。[11] K. Grover, E. F. Acar, M. Torabi，基于 Copula 的小区域估计预测，《加拿大统计学杂志》，48（2020年）685–711。[12] Ivan Kojadinovic, Jun Yan，使用 copula R 软件包对连续边缘分布的多元分布进行建模，《统计软件杂志》，34（2010年）1–20。[13] H. Joe，使用 Copula 进行依赖建模，Chapman and Hall，2014年。[14] P. Kumar, M.Shoukri，基于 Copula 的预测模型：在主动脉瓣反流研究中的应用，《*BMC* 医学研究方法学》，7（2007年）1–9。[15] J. F. Mai, M. Scherer，《模拟 Copula：随机模型、抽样算法与应用》，《定量金融》，2012年。[16] C. E. McCulloch, S. R. Searle，《广义线性混合模型》，John Wiley & Sons，2001年。[17] R. B. Nelsen，《Copula 导论》，Springer，2006年。[18] H. Noh, A. E. Ghouch, T. Bouezmarni，基于 Copula 的回归估计与推断，《美国统计学会会刊》，108（2013年）676–688。[19] A. Panagiotelis, C. Czado, H. Joe，多元离散数据的成对 Copula 构造，《美国统计学会会刊》，107（2012年）1063–1072。[20] L. Rivest, F. Verret, S. Baillargeon，使用 Copula 进行单位层级小区域估计，《加拿大统计学杂志》/《加拿大统计学评论》，44（2016年）397–415。[21] H. Romdhani, L. Lakhal-Chaieb, L.-P. Rivest，聚类数据的可交换 Kendall’s tau，《加拿大统计学杂志》，42（2014年）384–403。[22] L. A. Stefanski, D. D. Boos，M 估计法则，《美国统计学家》，56（2002年）29–38。[23] C.-L. Su, J. G. Neslehova, W. Wang，使用分层 Kendall Copula 对分层聚类删失数据建模，《加拿大统计学杂志》，47（2019年）182–203。[24] A. A. Tsiatis，《半参数理论与缺失数据》，Springer，2006年。[25] G. Verbeke, G. Molenberghs，纵向数据的线性混合模型，Springer，2000年。