

STAT3003: 统计算法基础

随机变量的产生 (2)

曾靖

中国科学技术大学管理学院

大纲

- 4. 混合模型
- 5. 特殊变换
- 6. Copula 模型

本节内容来自于《统计计算使用 R 》第 3 章,《统计计算》第 2 章

4 混合模型

4.1 混合模型

- 有时候被称为复合法。
- 假设随机变量 X 的分布函数 $F(x)$ 满足：

$$F(x) = \sum_{i=1} p_i F_i(x),$$

其中对所有的 i 都有 $p_i > 0$ 以及 $\sum_i p_i = 1$ 。随机变量 X 被称为服从一个混合分布模型。

- 为了产生随机样本 X ，我们可以
 - 产生一个独立的潜变量 (latent variable) I 使得其分布满足 $P(I = i) = p_i$ ；
 - 若 $I = i$ ，生成样本 $X_i \sim F_i(x)$ ，并令 $X = X_i$ 。

例子 4.1

随机变量 X 的分布律如下:

$$P(X = j) = \begin{cases} 0.05 & j = 1, 2, 3, 4, 5, \\ 0.15 & j = 6, 7, 8, 9, 10. \end{cases}$$

X 可以被看成随机变量 Y_1 和 Y_2 的混合分布, 其中

- Y_1 是来自于 $\{1, 2, 3, 4, 5\}$ 的离散均匀分布;
- Y_2 是来自于 $\{6, 7, 8, 9, 10\}$ 的离散均匀分布;
- 假设潜变量 I 的分布为 $P(I = 0) = 0.25, P(I = 1) = 0.75$ 。
如果 $I = 0$, 则 $X = Y_1$, 否则 $X = Y_2$ 。

思考: 验证该混合分布与 X 同分布

模拟代码

```
options(width = 40)
n <- 10000
p <- c(0.25, 0.75)
# I 记录分组情况
I <- sample(0:1, size = n, replace = TRUE, prob = p)
y1 <- sample(1:5, size = n, replace = TRUE)
y2 <- sample(6:10, size = n, replace = TRUE)
x <- (1-I)*y1 + I*y2
table(x)/n
```

```
## x
##      1      2      3      4      5
## 0.0516 0.0527 0.0515 0.0475 0.0457
##      6      7      8      9     10
## 0.1472 0.1480 0.1600 0.1456 0.1502
```

例子 4.2: 混合正态模型

随机变量 X 是两个不同的正态分布的混合分布:

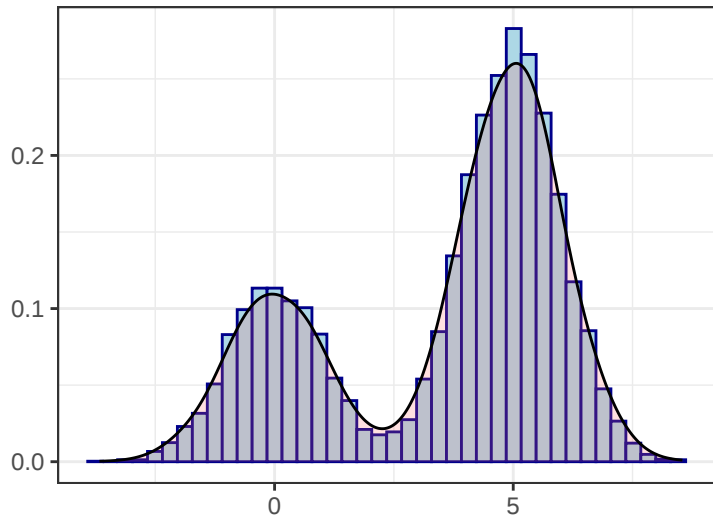
$$X \mid (I = 1) \sim N(0, 1),$$

$$X \mid (I = 2) \sim N(5, 1).$$

- 假设潜变量 I 的分布为 $P(I = 1) = 0.3, P(I = 2) = 0.7$ 。
如果 $I = 1$, 则 $X \sim N(0, 1)$, 否则 $X \sim N(5, 1)$ 。

模拟代码

```
n <- 10000
p <- c(0.3, 0.7)
I <- sample(0:1, size = n, replace = TRUE, prob = p)
y1 <- rnorm(n)
y2 <- rnorm(n, mean = 5, sd = 1)
x <- (1-I)*y1 + I*y2
data <- data.frame(x = x)
ggplot(data, aes(x = x))+
  geom_histogram(aes(y=after_stat(density)),
                 color="darkblue", fill="lightblue",
                 bins = 40)+
  geom_density(alpha=.2, fill="#FF6666")+
  labs(x = "", y = "")
```



例子 4.3：混合线性回归模型

令 $X = (1, x)^\top$ ，潜变量 $W \in \{1, 2\}$ 表示观测所属的回归成分。
混合线性回归模型可以写成

$$Y = \sum_{k=1}^2 I(W = k) X^\top \beta_k^* + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2),$$

- 其中 $I(W = k)$ 是指示函数，潜变量 W 决定观测来自哪一条回归线。
- 因此，我们有

$$Y \mid (X, W = 1) \sim N(X^\top \beta_1^*, \sigma^2),$$

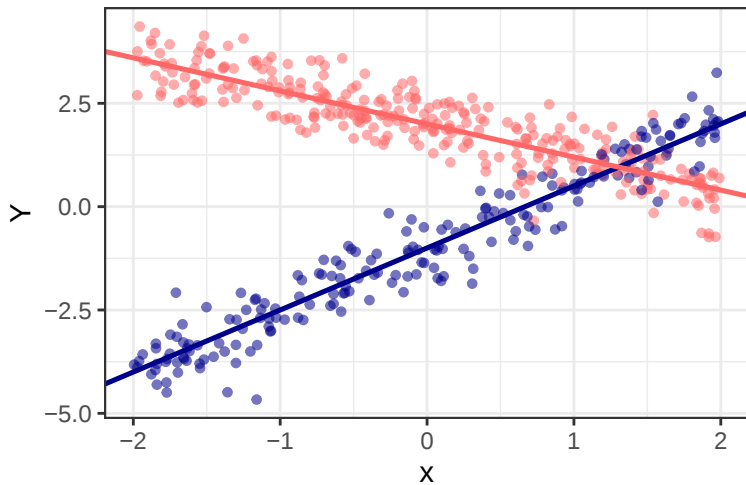
$$Y \mid (X, W = 2) \sim N(X^\top \beta_2^*, \sigma^2).$$

- 下面例子中取 $P(W = 1) = 0.4$, $P(W = 2) = 0.6$,
 $\beta_1^* = (-1, 1.5)^\top$, $\beta_2^* = (2, -0.8)^\top$, $\sigma = 0.5$ 。

模拟代码

```
set.seed(3003)
n <- 500
p <- c(0.4, 0.6)
beta1 <- c(-1, 1.5)
beta2 <- c(2, -0.8)
sigma <- 0.5

x <- runif(n, min = -2, max = 2)
X <- cbind(1, x)
W <- sample(1:2, size = n,
            replace = TRUE, prob = p)
mu1 <- drop(X %*% beta1)
mu2 <- drop(X %*% beta2)
mu <- ifelse(W == 1, mu1, mu2)
y <- rnorm(n, mean = mu, sd = sigma)
data43 <- data.frame(x, y, W = factor(W))
```



W • $W = 1$ • $W = 2$

5 特殊变换

5.1 特殊方法

- 若 $X_1, \dots, X_r \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(p)$, 那么 $\sum_{i=1}^r X_i \sim \text{Bin}(r, p)$;
- 若 $X_1, \dots, X_r \stackrel{\text{i.i.d.}}{\sim} \text{Geom}(p)$, 那么 $\sum_{i=1}^r X_i \sim \text{NegBin}(r, p)$;
- 若 $X_1, \dots, X_r \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$, 那么 $\sum_{i=1}^r X_i^2 \sim \chi^2(r)$;
- 若 $X_1, \dots, X_r \stackrel{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$, 因为 $\text{Gamma}(1, \lambda) \sim \text{Exp}(\lambda)$, 因此 $\sum_{i=1}^r X_i \sim \text{Gamma}(r, \lambda)$.

- 若 $X \sim \chi^2(n)$ 和 $Y \sim \chi^2(m)$ 相互独立, 那么 $F = \frac{X/n}{Y/m} \sim F(n, m)$.
- 若 $Z \sim \mathcal{N}(0, 1)$ 和 $Y \sim \chi^2(n)$ 相互独立, 那么 $T = \frac{Z}{\sqrt{Y/n}} \sim t(n)$.
- 若 $U_1, U_2 \sim U(0, 1)$ 且相互独立, 令 $Z_1 = \sqrt{-2 \log(U_1)} \cos(2\pi U_2)$ 和 $Z_2 = \sqrt{-2 \log(U_1)} \sin(2\pi U_2)$ 。则 Z_1 和 Z_2 独立且均来自于 $N(0, 1)$.

例子 5.1: 对数分布

- 若 $U, V \sim \text{Unif}(0, 1)$ 是相互独立的, 那么

$$X = \left\lfloor 1 + \frac{\log(V)}{\log(1 - (1 - \theta)^U)} \right\rfloor$$

服从参数为 θ 的对数分布 (Logarithmic distribution)。

```
rlogarithmic2 <- function(n, theta){  
  th <- rep(theta, length = n)  
  u <- runif(n)  
  v <- runif(n)  
  x <- floor(1 + log(v) / log(1 - (1-th)^u))  
  return(x)  
}
```

```
n <- 1000
theta <- 0.5
x <- rlogarithmic2(n, theta)
k <- sort(unique(x))
p <- -1 / log(1-theta) * theta^k / k
round(rbind(table(x)/n, p), 3)
```

```
##          1          2          3          4          5          6          7
## 0.750 0.154 0.06 0.023 0.006 0.004 0.003
## p 0.721 0.180 0.06 0.023 0.009 0.004 0.002
```

5.2 多元正态分布

设随机向量 $X = (X_1, \dots, X_d)^\top \in \mathbb{R}^p$ 服从多元正态分布 $N(\mu, \Sigma)$ ，联合密度函数为

$$f(x) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{(x - \mu)^\top \Sigma^{-1} (x - \mu)}{2} \right\}, x \in \mathbb{R}^k.$$

有不同的方法可以生成多元正态分布。

(1) R 中现成的命令:

`MASS:mvrnorm(n, mean = mu, sigma = sigma)`

`mvtnorm:rmvnorm(n, mean = mu, sigma = sigma)`

- `n` 样本数
- `mean` = 指定均值, 默认是 0
- `sigma` = 指定方差, 默认是对角阵

(2) 通过矩阵分解生成多元正态分布:

- 产生服从 p 元标准正态分布 $N(0, I)$ 的随机向量 $Z = (Z_1, \dots, Z_p)$;

```
Z <- matrix(rnorm(n*p), nrow = n, ncol = p)
```

- 对正定矩阵 Σ 进行分解 $\Sigma = CC^\top$;
- $X = \mu + CZ \sim \mathcal{N}(\mu, \Sigma)$.

```
X <- Z %*% t(C) + matrix(mu, n, p, byrow = TRUE)
```

Question: 那么做怎样的分解好?

5.3 协方差矩阵分解

(1) 谱分解或特征根分解: eigen

- 令 $\Sigma = P\Lambda P^\top$, 其中 Λ 是由 Σ 的特征根组成的对角阵, P 是由对应的特征向量组成;
- 易知 $\Sigma^{1/2} = P\Lambda^{1/2}P^\top$, 则 $C = \Sigma^{1/2}$ 。

```
rmvn.eigen <- function(n, mu, Sigma){  
  p <- length(mu)  
  ev <- eigen(Sigma, symmetric = TRUE)  
  lambda <- ev$values  
  V <- ev$vectors  
  C <- V %*% diag(sqrt(lambda)) %*% t(V)  
  Z <- matrix(rnorm(n*p), nrow = n, ncol = p)  
  X <- Z %*% t(C) + matrix(mu, n, p, byrow = TRUE)  
  X  
}
```

(2) Choleski 分解: chol

- $\Sigma = CC^\top$, 其中 C 为下三角阵。

```
rmvn.Choleski <- function(n, mu, Sigma){  
  p <- length(mu)  
  Q <- chol(Sigma) # Q is a upper triangular matrix  
  Z <- matrix(rnorm(n*p), nrow = n, ncol = p)  
  X <- Z %*% t(Q) + matrix(mu, n, p, byrow = TRUE)  
  X  
}
```

方法比较

```
library("MASS");  
library("mvtnorm")  
n <- 100           # Sample size  
p <- 30            # Dimension  
N <- 100           # Iterations  
  
# 正态分布的均值和方差  
mu <- numeric(p)  
  
Sigma <- matrix(0, p, p)  
for (i in 1:p){  
  for(j in 1:p){  
    Sigma[i,j] <- 0.5^abs(i-j)  
  }  
}
```

```
library(rbenchmark)
benchmark(MASS::mvrnorm(n, mu, Sigma),
          mvtnorm::rmvnorm(n, mu, Sigma),
          mvtnorm::rmvnorm(n, mu, Sigma,
                           method = "chol"),
          rmvn.eigen(n, mu, Sigma), ## self-defined
          rmvn.Choleski(n, mu, Sigma), ## self-defined
          order = "relative", replications = N)
```

```

##                                     test
## 5                                rmvn.Choleski(n, mu, Sigma)
## 4                                rmvn.eigen(n, mu, Sigma)
## 1                                MASS::mvrnorm(n, mu, Sigma)
## 3 mvtnorm::rmvnorm(n, mu, Sigma, method = "chol")
## 2                                mvtnorm::rmvnorm(n, mu, Sigma)
##  replications elapsed relative user.self sys.self
## 5           100    0.016     1.000     0.015     0.000
## 4           100    0.021     1.313     0.021     0.001
## 1           100    0.022     1.375     0.020     0.001
## 3           100    0.022     1.375     0.021     0.001
## 2           100    0.032     2.000     0.030     0.001
##  user.child sys.child
## 5           0         0
## 4           0         0
## 1           0         0
## 3           0         0
## 2           0         0

```

5.4 多元 Student's t 分布

- 多元 t 分布可以被看作 (一元) t 分布在多元变量上的拓展。
假设 p 元变量 X 可以表示成如下形式

$$X = \mu + \frac{Y}{\sqrt{u/\nu}}, \text{ where } Y \sim N(0, \Sigma), \text{ and } u \sim \chi_{\nu}^2.$$

则称 X 满足多元 t 分布, 并记 $X \sim T_{\nu}(\mu, \Sigma; p)$ 。

如果 $X \sim T_\nu(\mu, \Sigma; p)$, 那么 X 具有如下一些性质:

- 矩性质:
 1. $E(X) = \mu$;
 2. 如果 $\nu > 2$, 协方差矩阵存在且 $\text{Cov}(X) = \frac{\nu}{\nu-2}\Sigma$;
- 与一元 t 分布的关系:
 - $T_\nu(0, 1; 1) \sim t_\nu$, 即退化为一元 t 分布;
 - $\sigma_{jj}^{-1/2}(X_j - \mu_j) \sim t_\nu, \forall j = 1, \dots, p$.
 - 对于任意向量 $a \neq 0, (a^\top \Sigma a)^{-1/2} a^\top (X - \mu) \sim t_\nu$;
- 线性变换:
 - 对于任意向量 a 以及非奇异矩阵 C , 我们有 $CX + a \sim T_\nu(C\mu + a, C\Sigma C^\top; p)$ 。

6 Copula 分布

6.1 Copula 分布

- 问题 1: 考虑一个多元随机变量 $X = (X_1, \dots, X_p)^\top$, 假设每个分量 X_i 是正态分布随机变量, X 是否为多元正态分布?
- 问题 2: 假设每个分量 X_i 是独立正态随机变量, X 是否为多元正态分布?
- 问题 3: 考虑两个独立正态随机变量 X, Y 。令 $Z = |X| \cdot \text{sign}(Y)$, Z 服从什么分布? (X, Z) 和 (Y, Z) 是否为多元正态?
- 问题 4: 如何保证 X 是多元正态分布?
- 边际分布 + Copula $\Rightarrow$ 多元分布

Definition 1 (Copula 分布). 一个 p 维的 Copula 是一个多元变量 $(U_1, \dots, U_p)$ 的累积密度函数 $C : [0, 1]^p \rightarrow [0, 1]$, 其边际分布为均匀分布。

- 假设 $U_i \sim \text{Unif}(0, 1), i = 1, \dots, p$, 则

$$C(u_1, \dots, u_p) = P(U_1 \leq u_1, \dots, U_p \leq u_p).$$

- $C(1, \dots, 1, u_i, 1, \dots, 1) = u_i$ (how to prove?)

存在性定理

Theorem 2 (Sklar's Theorem). 考虑一个多元累计分布函数 F 以及其边际分布 $F_1, \dots, F_p$, 则存在一个 Copula C 使得

$$F(x_1, \dots, x_p) = C(F_1(x_1), \dots, F_p(x_p)), \quad \forall x_i \in \mathbb{R}, i = 1, \dots, d.$$

Question: Copula 如何构造?

构造性定理

Theorem 3. 假设多元分布 $X = (X_1, \dots, X_p)$ 的 CDF 为 F , 边际分布满足 $X_i \sim F_i$, 则该多元分布对应的 **Copula** 为多元变量 $(F_1(X_1), \dots, F_p(X_p))$ 的 CDF, 记为 C 。并且我们有

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

Proof:

$$\begin{aligned} F(x_1, \dots, x_p) &= P(X_1 \leq x_1, \dots, X_p \leq x_p) \\ &= P(F_1(X_1) \leq F_1(x_1), \dots, F_p(X_p) \leq F_p(x_p)) \\ &= C(F_1(x_1), \dots, F_p(x_p)). \end{aligned}$$

Corollary 1. 对于任意一个多元分布 F ，其边际分布为 F_i ， $i = 1, \dots, p$ ，则其 Copula 为

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

思考: *How to prove?*

- 边际分布 + Copula $\Rightarrow$ 多元分布

Theorem 4. 对于任意的一个 Copula $C : [0, 1]^p \rightarrow [0, 1]$ 和 p 个一元累计分布函数 $F_1, \dots, F_p$, 构造多元函数 G

$$G(x_1, \dots, x_p) = C(F_1(x_1), \dots, F_p(x_p)).$$

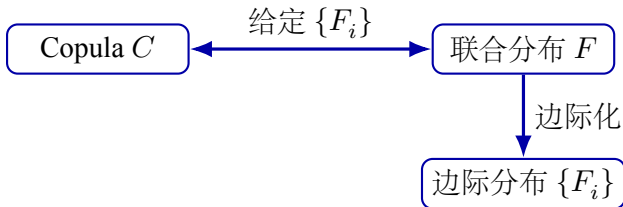
则

- G 是一个多元累计分布函数;
- G 的边际分布函数为 $F_1, \dots, F_p$;
- G 的 Copula 是 C 。

思考: *How to prove?*

结论:

- F 由边际分布 F_i 和 Copula C 完全决定
- Copula C 由 F 和 F_i 完全决定
- F_i 由 F 完全决定



Lemma 1. 假设 $(X_1, \dots, X_p)$ 的 Copula 为 C 。令 $T_i : \mathbb{R} \rightarrow \mathbb{R}$ 为严格单增函数，则 $(T_1(X_1), \dots, T_p(X_p))$ 的 Copula 仍然是 C 。

思考: *How to prove?*

6.2 一些常用的 Copula 模型

- Gaussian Copula: 多元正态分布的 Copula。
 - 记 $(X_1, \dots, X_p) \sim N(0, \Sigma)$ 且 $\sigma_{ii} = 1, i = 1, \dots, p$, 则 $(\Phi(X_1), \dots, \Phi(X_p))$ 的 CDF 被称为 Gaussian Copula, 这里 $\Phi(\cdot)$ 是标准正态分布的 CDF。
- t -Copula: 多元 t 分布的 Copula。
 - 记 $(X_1, \dots, X_p) \sim T_\nu(0, \Sigma; p)$ 且 $\sigma_{ii} = 1, i = 1, \dots, p$, 则 $(F_t(X_1), \dots, F_t(X_p))$ 的 CDF 被称为 t -Copula, 其中 $F_t \sim t_\nu$ 。
- 回顾问题 4: 如何保证 X 是多元正态分布?

6.2 一些常用的 Copula 模型

- Gaussian Copula: 多元正态分布的 Copula。
 - 记 $(X_1, \dots, X_p) \sim N(0, \Sigma)$ 且 $\sigma_{ii} = 1, i = 1, \dots, p$, 则 $(\Phi(X_1), \dots, \Phi(X_p))$ 的 CDF 被称为 Gaussian Copula, 这里 $\Phi(\cdot)$ 是标准正态分布的 CDF。
- t -Copula: 多元 t 分布的 Copula。
 - 记 $(X_1, \dots, X_p) \sim T_\nu(0, \Sigma; p)$ 且 $\sigma_{ii} = 1, i = 1, \dots, p$, 则 $(F_t(X_1), \dots, F_t(X_p))$ 的 CDF 被称为 t -Copula, 其中 $F_t \sim t_\nu$ 。
- 回顾问题 4: 如何保证 X 是多元正态分布?
 - Gaussian 边际 CDF + Gaussian Copula

- 多元正态的 **Copula** 完全由相关系数矩阵 M 决定。
 - 对于多元正态 $X = (X_1, \dots, X_p) \sim N(\mu, \Sigma)$, 我们对每一个变量做变换 $Z_i = \sigma_{ii}^{-1/2}(X_i - \mu_i), i = 1, \dots, p$, 则

$$Z = (Z_1, \dots, Z_p) \sim N(0, M),$$

这里 $M = \{\text{diag}(\Sigma)\}^{-1/2}\Sigma\{\text{diag}(\Sigma)\}^{-1/2}$ 表示 Σ 对应的相关系数矩阵。由 Lemma 1 可知正态分布 Z 和 X 的 **Copula** 一致。

- 因此, **Gaussian Copula** 和多元正态分布 $N(0, M)$ 一一对应, 这里 M 为一个相关系数矩阵。

6.3 生成样本

Question: 假设已知多元分布 $(X_1, \dots, X_p)$ 的边际分布为 $X_i \sim F_i$, Copula 为 C , 如何生成样本 $(X_1, \dots, X_p)$?

Solution:

1. 从 Copula 中生成样本 $(U_1, \dots, U_p)$;
2. 做变换 $(X_1, \dots, X_p) = (F_1^{-1}(U_1), \dots, F_p^{-1}(U_p))$.

Proof: 根据 Theorem 4, 生成的样本 $(X_1, \dots, X_p)$ 的联合密度函数为:

$$C(F_1(x_1), \dots, F_p(x_p)).$$

我们需要证明上述算法得到的随机变量 $(X_1, \dots, X_p)$ 满足该联合密度函数:

$$\begin{aligned} & P(F_1^{-1}(U_1) \leq x_1, \dots, F_p^{-1}(U_p) \leq x_p) \\ &= P(U_1 \leq F_1(x_1), \dots, U_p \leq F_p(x_p)) \\ &= C(F_1(x_1), \dots, F_p(x_p)). \end{aligned}$$

Question: 已知多元分布 $(X_1, \dots, X_p)$ 的联合分布函数 F 以及边际分布 $X_i \sim F_i$, 如何生成满足 Copula 分布的样本?

Solution:

1. 从 F 中生成样本 $(X_1, \dots, X_p)$;
2. 做变换 $(U_1, \dots, U_p) = (F_1(X_1), \dots, F_p(X_p))$.

Proof: 根据 Corollary 1, 生成的 Copula 形式为

$$F(F_1^{-1}(u_1), \dots, F_p^{-1}(u_p)).$$

因此, 我们需要证明上述算法生成的随机变量 $(U_1, \dots, U_p)$ 具有如上的联合密度函数:

$$\begin{aligned} & P(U_1 \leq u_1, \dots, U_p \leq u_p) \\ &= P(X_1 \leq F_1^{-1}(u_1), \dots, X_p \leq F_p^{-1}(u_p)) \\ &= F(F_1^{-1}(u_1), \dots, F_p^{-1}(u_p)). \end{aligned}$$

例子 6.1

假设已知多元分布 $X = (X_1, X_2)$ 的边际分布为 $X_i \sim F_i$, $i = 1, 2$, Copula 为相关系数矩阵为 M 的 Gaussian Copula, 如何生成样本 $X = (X_1, X_2)$?

Note: 这里 X 不一定是多元正态分布。

1. 从 Gaussian Copula 中生成样本

- 从正态分布 $N(0, M)$ 中生成样本 $Z = (Z_1, Z_2)$ 。
- $(U_1, U_2) = (\Phi(Z_1), \Phi(Z_2))$ 则服从相关矩阵为 M 的 Copula 分布。

2. 令 F_1 和 F_2 表示 X_1 和 X_2 的边际分布, 则 $X = (F_1^{-1}(U_1), F_2^{-1}(U_2))$ 。

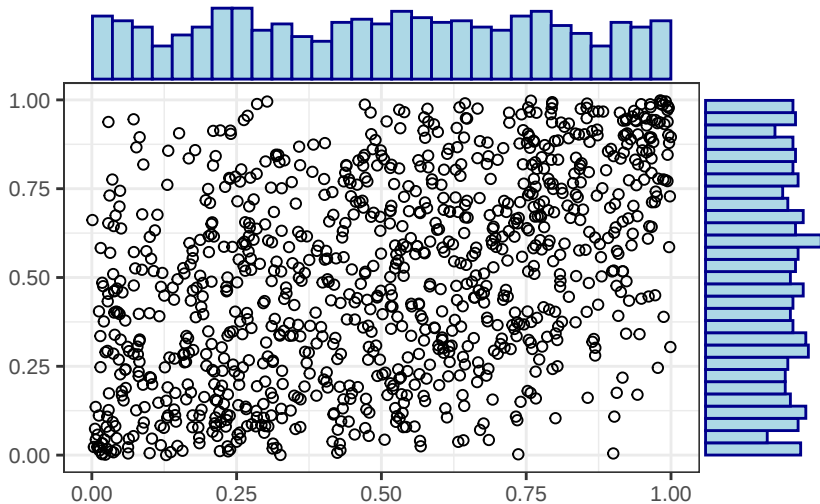
6.4 模拟代码

```
library(ggExtra)
library(latex2exp)
library(MASS)
sim.copula <- function(n, rho, qm1, qm2){
  mu <- rep(0, 2)
  M <- matrix(rho, 2, 2)
  diag(M) <- 1
  dat <- mvrnorm(n, mu, M)
  dat[,1] <- qm1(pnorm(dat[,1]))
  dat[,2] <- qm2(pnorm(dat[,2]))
  return(dat)
}
```

均匀边际分布

```
qm1 <- function(p) qunif(p)
qm2 <- function(p) qunif(p)
n <- 1000; rho <- 0.5
sim <- sim.copula(n, rho, qm1, qm2)
df <- data.frame(x = sim[,1], y = sim[,2])
p <- ggplot(df, aes(x, y)) + geom_point(shape = 1) +
  labs(x = "", y = "", title =
    TeX(sprintf("Uniform margins $(\\rho = %.1f)$",
      rho)))+
  theme_bw()+
  theme(plot.title = element_text(size = 8, hjust = 0.5),
    axis.text = element_text(size = 8),
    axis.title = element_text(size = 8))
ggExtra::ggMarginal(p, type = "histogram", color="darkblue",
  fill="lightblue")
```

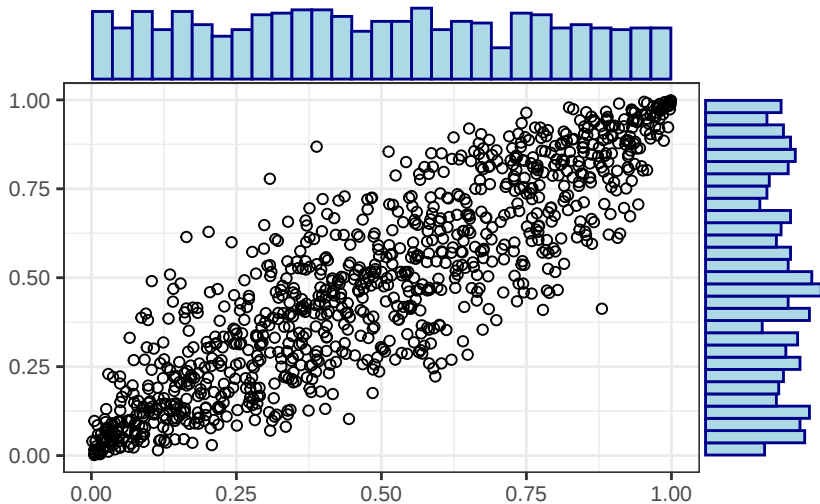
Uniform margins ($\rho = 0.5$)



均匀边际分布

```
qm1 <- function(p) qunif(p)
qm2 <- function(p) qunif(p)
n <- 1000; rho <- 0.9
sim <- sim.copula(n, rho, qm1, qm2)
df <- data.frame(x = sim[,1], y = sim[,2])
p <- ggplot(df, aes(x, y)) + geom_point(shape = 1) +
  labs(x = "", y = "",
        title =
          TeX(sprintf("Uniform margins $(\\rho = %.1f)$",
                      rho))) +
  theme_bw() +
  theme(plot.title = element_text(size = 8, hjust = 0.5),
        axis.text = element_text(size = 8),
        axis.title = element_text(size = 8))
ggExtra::ggMarginal(p, type = "histogram", color="darkblue",
                    fill="lightblue")
```

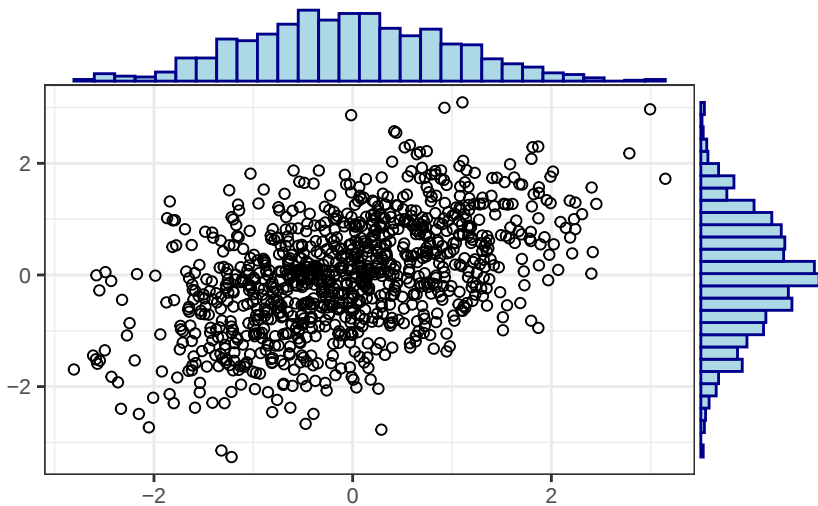
Uniform margins ($\rho = 0.9$)



正态边际分布

```
qm1 <- function(p) qnorm(p)
qm2 <- function(p) qnorm(p)
n <- 1000; rho <- 0.5
sim <- sim.copula(n, rho, qm1, qm2)
df <- data.frame(x = sim[,1], y = sim[,2])
p <- ggplot(df, aes(x, y)) + geom_point(shape = 1) +
  labs(x = "", y = "", title =
    TeX(sprintf("Normal margins $(\\rho = %.1f)$",
      rho)))+
  theme_bw()+
  theme(plot.title = element_text(size = 8, hjust = 0.5),
    axis.text = element_text(size = 8),
    axis.title = element_text(size = 8))
ggExtra::ggMarginal(p, type = "histogram", color="darkblue",
  fill="lightblue")
```

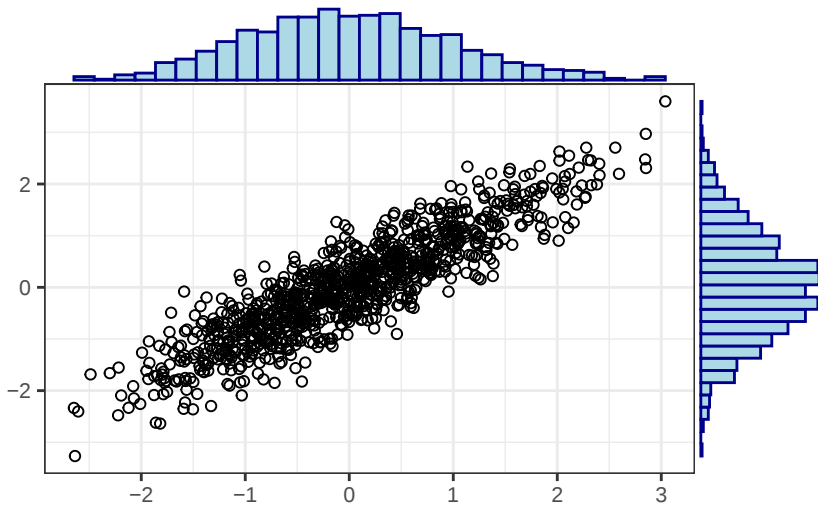
Normal margins ($\rho = 0.5$)



正态边际分布

```
qm1 <- function(p) qnorm(p)
qm2 <- function(p) qnorm(p)
n <- 1000; rho <- 0.9
sim <- sim.copula(n, rho, qm1, qm2)
df <- data.frame(x = sim[,1], y = sim[,2])
p <- ggplot(df, aes(x, y)) + geom_point(shape = 1) +
  labs(x = "", y = "", title =
    TeX(sprintf("Normal margins $(\\rho = %.1f)$",
      rho)))+
  theme_bw()+
  theme(plot.title = element_text(size = 8, hjust = 0.5),
    axis.text = element_text(size = 8),
    axis.title = element_text(size = 8))
ggExtra::ggMarginal(p, type = "histogram", color="darkblue",
  fill="lightblue")
```

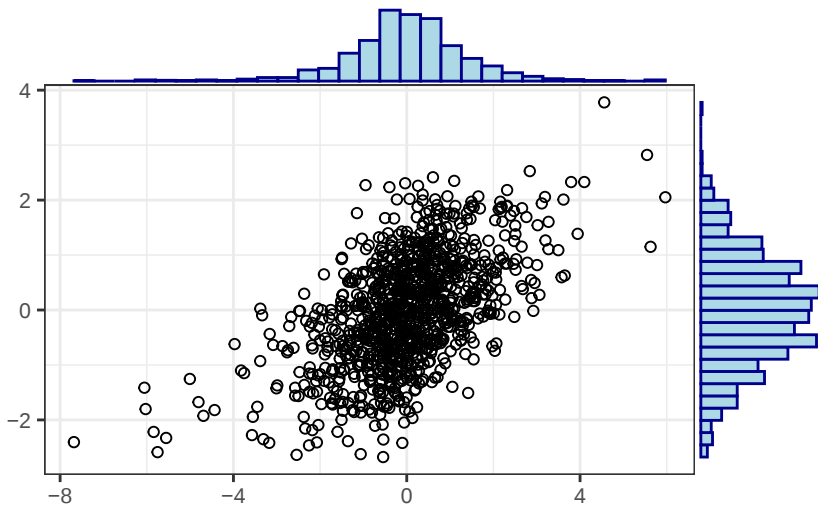
Normal margins ($\rho = 0.9$)



正态和 t 边际

```
qm1 <- function(p) qt(p, df = 4)
qm2 <- function(p) qnorm(p)
n <- 1000; rho <- 0.5
sim <- sim.copula(n, rho, qm1, qm2)
df <- data.frame(x = sim[,1], y = sim[,2])
p <- ggplot(df, aes(x, y)) + geom_point(shape = 1) +
  labs(x = "", y = "", title =
    TeX(sprintf(
      "Normal + Student's t margins $(\\rho = %.1f)$",
      rho))) +
  theme_bw() +
  theme(plot.title = element_text(size = 8, hjust = 0.5),
        axis.text = element_text(size = 8),
        axis.title = element_text(size = 8))
ggExtra::ggMarginal(p, type = "histogram", color="darkblue",
                    fill="lightblue")
```

Normal + Student's t margins ($\rho = 0.5$)



6.5 Copula 的应用场景

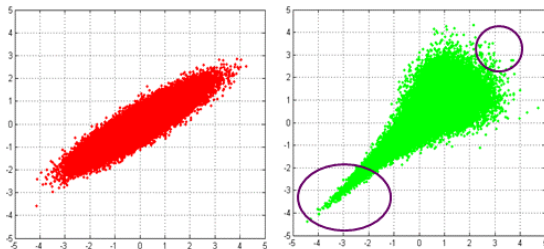
Copula 将每个变量的边际行为与变量之间的依赖结构分开建模，特别适合描述非线性、非对称和尾部相关。

- 金融与信用风险：刻画资产收益的联合损失、违约时间的相关性，并用于投资组合风险和 CDO 定价。
- 保险精算：建模赔款额与理赔费用、不同险种索赔之间的依赖，并计算总损失和再保险价格。
- 生存与医学数据：分析成对或聚类生存时间、复发事件以及竞争风险之间的相关性。
- 工程可靠性：描述相关部件的寿命或退化过程，计算系统失效概率并辅助维护决策。

参考：Frees and Valdez (1998), Li (2000), Shih and Louis (1995), Navarro and Durante (2017)。

金融案例：相同边际，不同尾部风险

下图中的两个二元模型具有相同的正态边际分布，但依赖结构并不相同。



- 左图采用对称的相关结构；右图的 **Clayton Copula** 具有下尾相关，能够描述两项资产同时出现较大损失的情形。
- 因此，相同的边际分布并不保证相同的投资组合尾部风险。

为什么使用 Copula model?

- 边际分布不同：可组合正态、重尾、偏态、寿命等不同边际分布。
- 尾部依赖可能不对称：两个变量可能在正常时期关系较弱，但在极端情况下更容易同时取大值或同时取小值。高斯分布的依赖结构是对称的，难以描述这种现象。例如，两个资产平时的相关性不高，但金融危机时经常同时大幅下跌。普通高斯模型可能低估这种联合损失