第二章 随机向量的抽样方法

引言

- 一元随机变量的两种主要抽样方法：CDF 逆变换和 A-R 抽样，推广到多元抽样的效率通常很低
- 本章将重点关注一些常用的多元分布的抽样，比如多元正态分布、多元 t 分布、Dirichlet 分布以及多项分布等
- 多元抽样的挑战在于如何给随机向量的元素之间赋予正确的相关结构，本章将介绍 copula-marginal 方法，其基本想法是将一种相关结构已知的多元分布通过边际分布变换得到另一多元分布
- 本章还会介绍一些随机矩阵的抽样方法，比如矩阵正态分布，Wishart 矩阵，随机图等。

一元抽样方法的推广

- CDF 逆变换
  - 对随机向量 $\boldsymbol{X} \in \mathbb{R}^d$ 的抽样过程 (sequential inversion):
    首先抽取 $U_j \overset{iid}{\sim} \boldsymbol{U}(0,1),\ j = 1, 2, \ldots, d$, 然后依次令

$$
\begin{aligned}
X_1 &= F_1^{-1}(U_1) \\
X_j &= F_{j|1:(j-1)}^{-1}(U_j \mid X_{1:(j-1)}),\ j = 2, \ldots, d
\end{aligned}
$$

  - 序列条件分布 $F_{j|1:(j-1)}(x_j \mid x_{1:(j-1)})$ 的逆函数在高维情况下很难计算，且每个条件分布的逆函数都需要重新计算，使用 sequential inversion 抽样会很慢

一元抽样方法的推广

- Acceptance-Rejection
  - A-R 的几何解释在多元情形下依然成立, 仍可使用未归一化的 PDF $\tilde{f}$ 和 $\tilde{g}$ 计算 g 的样本被接受的概率, 只要保证 $\tilde{f}(\boldsymbol{y}) \le \tilde{c}\tilde{g}(\boldsymbol{y}), \forall \boldsymbol{y}$
  - 例如，目标分布 f 是单位球体 $B_d = \{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x}\| \le 1\}$ 内的均匀分布, 令 g 表示 $\boldsymbol{U}[-1,1]^d$ 的 PDF，抽取 $\boldsymbol{Y} \sim g$ 后只保留 $\|\boldsymbol{Y}\| \le 1$ 的样本, 来自 g 的样本期望被接受的概率为

$$\frac{\mathrm{vol}(B_d)}{2^d} = \frac{\pi^{d/2}}{2^d \Gamma(1 + d/2)}.$$

    - ★ $d = 2$ 时，上述接受概率为 $\pi/4 \approx 0.785$
    - ★ $d = 9$ 时，上述接受概率 $<1\%$; $d = 23$ 时，接受概率 $< 10^{-9}$
  - 在高维情形下一般很难找到较小的 c，抽样效率很低

多元正态分布的抽样

- 对 $N_d(\mathbf{0}, I_d)$ 抽样很容易，此时各分量是独立的，可使用 Box-Muller 从 $N(0,1)$ 抽 d 个样本
- 如果 $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$，$A\boldsymbol{X} + \boldsymbol{b} \sim N_d(A\boldsymbol{\mu} + \boldsymbol{b}, A\Sigma A^\top)$
- 对 $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ 抽样, 只需找到矩阵 C 使得 $\Sigma = CC^\top$,

$$\boldsymbol{X} = \boldsymbol{\mu} + C\boldsymbol{Z},\ \boldsymbol{Z} \sim N_d(\mathbf{0}, I_d).$$

- 矩阵 C 的选择并不唯一
  - 特征值分解: $\Sigma = P\Lambda P^\top$, $C = P\Lambda^{1/2}$
  - Cholesky 分解: $\Sigma = LL^\top$, $C = L$，其中 L 是下三角矩阵

多元 t 分布的抽样

- $\mathbb{R}^d$ 上的 $t_d(\boldsymbol{\mu}, \Sigma, \nu)$ 的 PDF 为

$$f(\boldsymbol{x} \mid \boldsymbol{\mu}, \Sigma, \nu) = C_{\boldsymbol{\mu},\Sigma,\nu} \left(1 + (\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right)^{-(\nu+d)/2}$$

  以 $\boldsymbol{\mu}$ 为中心的一系列椭圆等高线, 比多元正态分布的尾厚; 当 $\nu \to \infty$ 时, $t_d(\boldsymbol{\mu}, \Sigma, \nu)$ 收敛到 $N_d(\boldsymbol{\mu}, \Sigma)$

- $t_d(\boldsymbol{\mu}, \Sigma, \nu)$ 各分量的边际分布为

$$\frac{X_j - \mu_j}{\sqrt{\Sigma_{jj}}} \sim t_{(\nu)}$$

- 多元 t 分布可由如下变换生成:

$$\boldsymbol{X} = \boldsymbol{\mu} + \frac{\Sigma^{1/2} \boldsymbol{Z}}{\sqrt{W/\nu}}, \ \boldsymbol{Z} \sim N_d(\boldsymbol{0}, I_d), W \sim \chi^2_{(\nu)}$$

  其中 $\boldsymbol{Z}$ 和 W 独立, $\Sigma^{1/2}$ 是任何满足 $CC^\top = \Sigma$ 的矩阵 C

- 参数 Σ 是尺度矩阵 (scale matrix), 不是协方差矩阵, $\Sigma = I_d$ 的多元 t 分布的各分量并不独立

多项分布的抽样

- 如果向 d 个格子独立地抛 m 个球，每个球落入格子 j 的概率为 p_j, $j = 1, \ldots, d$. 则落入每个格子 j 的球数 X_j 组成的向量 $\boldsymbol{X} = (X_1, \ldots, X_d)$ 服从多项分布 $\text{Mult}(m, p_1, \ldots, p_d)$, PMF 为

$$P(X_1 = x_1, \ldots, X_d = x_d) = \frac{m!}{x_1! x_2! \ldots x_d!} \prod_{j=1}^{d} p_j^{x_j}$$

- 对多项分布抽样可以按如下序列条件分布的形式依次对每个分量抽样

$$P(X_1, \ldots, X_d) = P(X_1) P(X_2 \mid X_1) \cdots P(X_d \mid X_1, \ldots, X_{d-1})$$

  - $X_1 \sim \text{Bin}(m, p_1)$
  - 给定 $\{X_1, \ldots, X_{j-1}\}$, X_j 的条件分布也是一个二项分布:

$$X_j \mid X_1, \ldots, X_{j-1} \sim \text{Bin}\left(m - \sum_{s=1}^{j-1} X_s, p_j / \sum_{k=j}^{d} p_k\right)$$

Dirichlet 分布的抽样

- Dirichlet 分布的一个样本是一组随机概率，可用于描述多项分布中参数向量 $(p_1, \ldots, p_d)$ 的分布，样本空间是 $\mathbb{R}^d$ 上的 unit simplex:

$$\Delta^{d-1} = \left\{ (x_1, \ldots, x_d) \mid x_j \geq 0, \sum_{j=1}^{d} x_j = 1 \right\}$$

- $\mathrm{Dir}(\alpha_1, \ldots, \alpha_d)$ 有 d 个参数, $\alpha_j > 0$, $j = 1, \ldots, d$, PDF 为

$$f(\boldsymbol{x}) = \frac{1}{D(\boldsymbol{\alpha})} \prod_{j=1}^{d} x_j^{\alpha_j - 1}, \quad \boldsymbol{x} \in \Delta^{d-1}$$

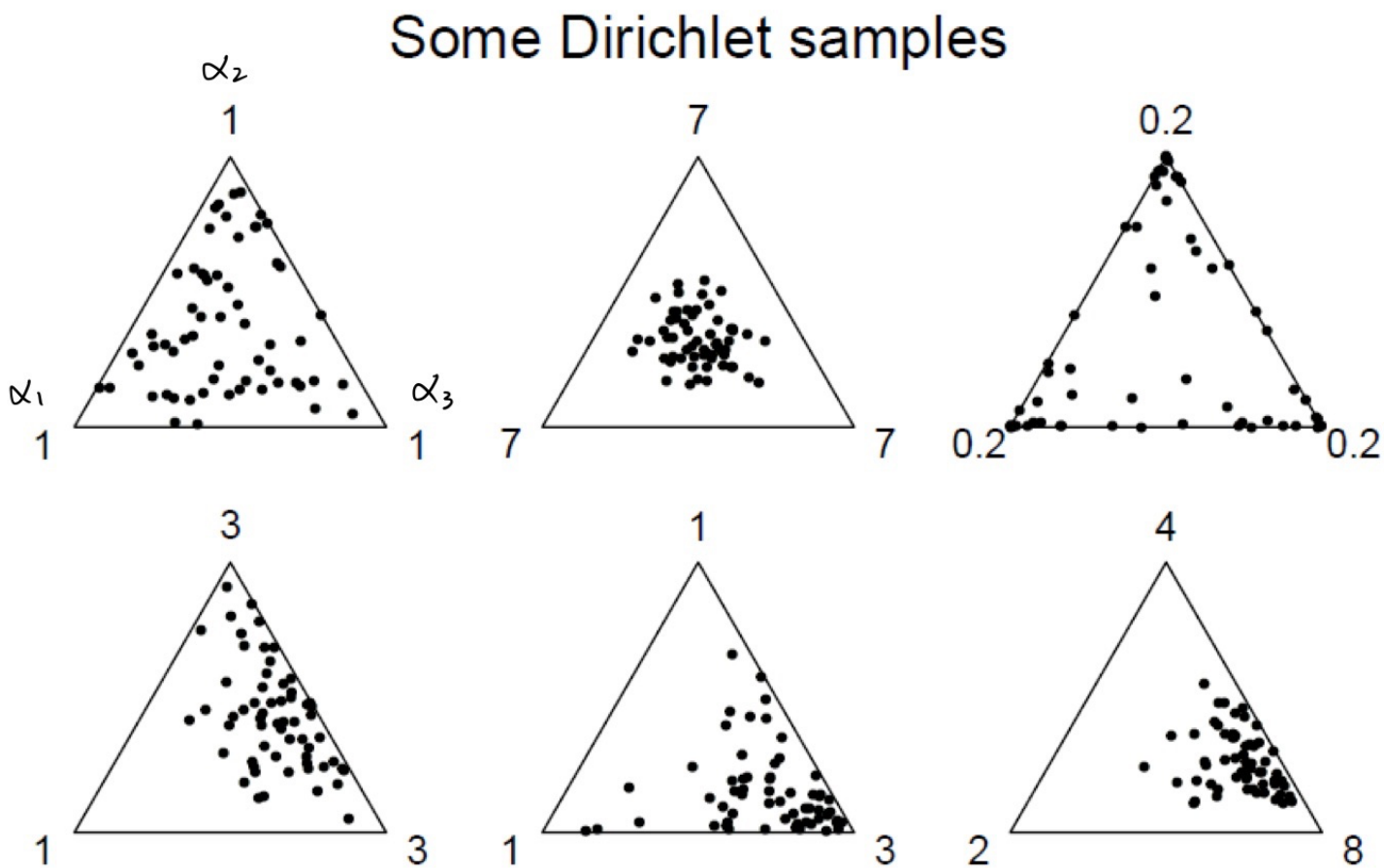
  $\boldsymbol{X} \sim \mathrm{Dir}(\boldsymbol{\alpha})$ 的期望为

$$E(X_j) = \frac{\alpha_j}{\sum_{k=1}^{d} \alpha_k}, \ j = 1, \ldots, d$$

- $\mathrm{Dir}(\alpha_1, \alpha_2)$ 等价于 $\mathrm{Beta}(\alpha_1, \alpha_2)$ 分布
  $\mathrm{Dir}(1, \ldots, 1)$ 是 Δ^{d-1} 上的均匀分布 $\boldsymbol{U}(\Delta^{d-1})$

Dirichlet 分布的抽样

- $d = 2$ 时的样本空间 Δ^1 是一个长度为 1 的线段，$d = 3$ 时的样本空间 Δ^2 可以用一个等边三角形表示



Dirichlet 分布的抽样

- $\boldsymbol{X} \sim \text{Dir}(\boldsymbol{\alpha})$ 可以由 Gamma 分布生成：

$$\begin{aligned} Y_j &\overset{ind}{\sim} \text{Gam}(\alpha_j, 1),\ j = 1, \ldots, d, \\ X_j &= \frac{Y_j}{\sum_{k=1}^{d} Y_k},\ j = 1, \ldots, d \end{aligned} \tag{1}$$

- $\boldsymbol{U}(\Delta^{d-1})$ 还可以使用 **uniform spacings** 方法抽样，只需产生 $d-1$ 个 $\boldsymbol{U}(0,1)$ 随机变量且避免了对数运算，但是排序的计算量为 $O(d\log(d))$

- Dirichlet 分布不是一个很灵活的分布，期望 $E(\boldsymbol{X})$ 用掉 $d-1$ 个参数，剩下的归一化参数 $\sum_{j=1}^{d} \alpha_j$ 描述 $\boldsymbol{X}$ 距 $E(\boldsymbol{X})$ 的远近

- Dirichlet 分布的各分量几乎是独立的，由于和为 1 的限制，各分量间有很小的负相关

Copula-marginal 方法

- 一种较通用的多元分布抽样/构造方法
- $\mathbb{R}^d$ 上的随机向量 $\boldsymbol{X} \sim F$，$F_j(x)$ 表示分量 X_j 的边际 CDF，$F_j(X_j) \sim \boldsymbol{U}(0,1)$，将随机向量 $\boldsymbol{U} = (F_1(X_1), \ldots, F_d(X_d))$ 服从的分布称为 F 的 copula，用 C 表示
- 如果 C 已知，copula-marginal 抽样过程如下：

$$\begin{aligned} &\text{抽取 } \boldsymbol{U} \sim C \\ &X_j = F_j^{-1}(U_j),\ j = 1, \ldots, d \end{aligned} \tag{2}$$

定义 (Copula)

如果函数 $C : [0,1]^d \to [0,1]$ 是 d 个边际分布为 $\boldsymbol{U}[0,1]$ 的随机变量的联合 CDF，则函数 C 是一个 copula.

Copula-marginal 方法

定理 (Sklar 定理)

F 是 $\mathbb{R}^d$ 上一个多元分布的 *CDF*，其边际分布的 *CDF* 为 $F_1, \ldots, F_d$. 则存在一个 *copula* C 使得

$$F(x_1, \ldots, x_d) = C\left(F_1(x_1), \ldots, F_d(x_d)\right).$$

如果所有 F_j 都是连续的，则 *copula* C 是唯一的；否则 C 只在 $F_1, \ldots, F_d$ 的取值范围上唯一确定。

- 对任意多元分布 F, 存在一个 copula C 使得通过变换(2)可以得到 $\boldsymbol{X} \sim F$, 困难在于如何确定 $\boldsymbol{U}$ 中各分量的相关性
- 假设多元分布 F 与分布 G 的 copula 相同，都为 C，且从 G 中抽样较容易, 则可以先对 G 抽样 $\boldsymbol{Y} \sim G$, 此时

$$(G_1(Y_1), \ldots, G_d(Y_d)) \sim C$$

然后令 $X_j = F_j^{-1}\left(G_j(Y_j)\right),\ j = 1, \ldots, d$, 则 $\boldsymbol{X} \sim F$

Copula-marginal 方法

- **Gaussian copula**. 给定一个相关系数矩阵 $R \in \mathbb{R}^{d\times d}$, 以及 d 个边际 CDFs $F_1, \ldots, F_d$, Gaussian copula 抽样方法如下:
  - 抽取 $\boldsymbol{Z} \sim N_d(\mathbf{0}, I_d)$
  - 令 $\boldsymbol{Y} = R^{1/2}Z$, 则 $\boldsymbol{Y} \sim N_d(\mathbf{0}, R)$ 且 $Y_j \sim N(0,1)$, $j = 1, \ldots, d$
  - 令 $X_j = F_j^{-1}(\Phi(Y_j))$, $j = 1, \ldots, d$.

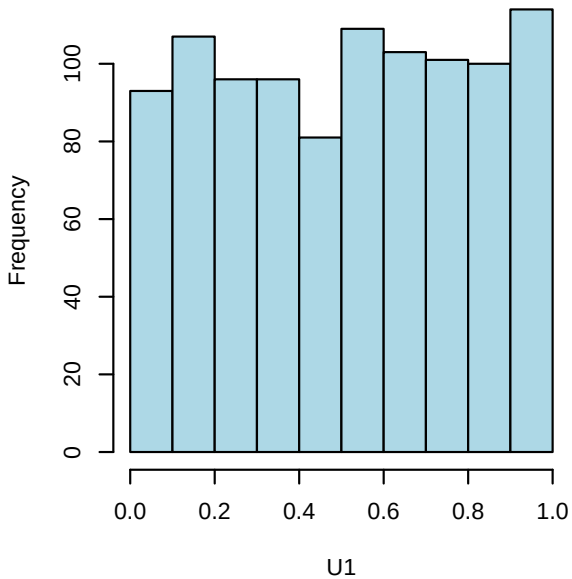
- Gaussian-copula 方法可以将多元正态分布的相关结构和一些边际 CDFs 结合产生新的分布，因此也被称为 NORTA 方法 (normal to anything)

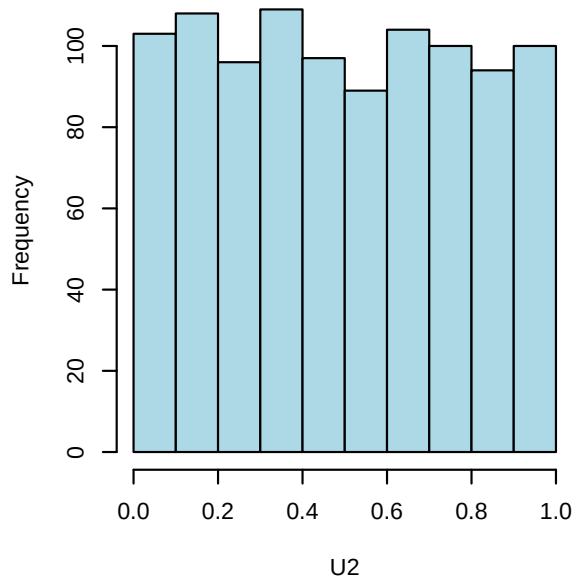
- 一个将 Gaussian-copula 与 Gamma 边际分布结合的例子

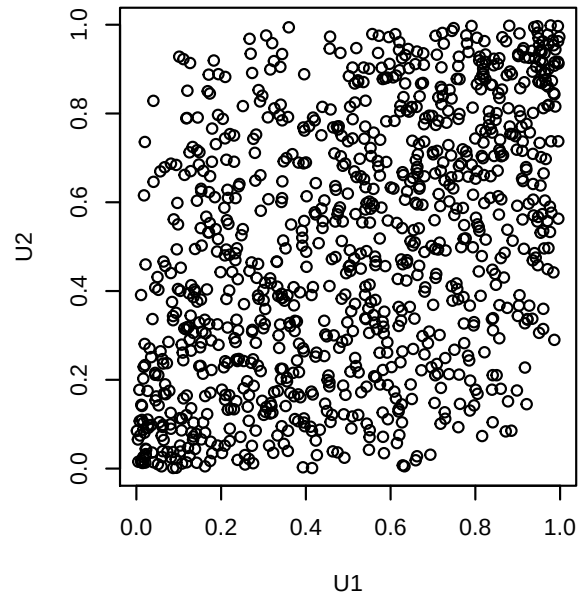
Gaussian-copula 与 Gamma 边际分布结合

```
## -- generate 1000 samples from bivariate normal
n = 1000
rho = 0.5
# compute square root of covariance matrix
ed = eigen(matrix(c(1,rho,rho,1),2,2), symmetric=TRUE)
R = ed$vectors %*% diag(sqrt(ed$values))

Y = matrix(rnorm(n*2),n,2) %*% t(R)
U = pnorm(Y)

par(mfrow = c(1,3))
hist(U[,1], xlab="U1", main="", col="lightblue")
hist(U[,2], xlab="U2", main="", col="lightblue")
plot(U[,1],U[,2], type="p", xlab="U1", ylab="U2", main="")
```

Gaussian-copula 与 Gamma 边际分布结合

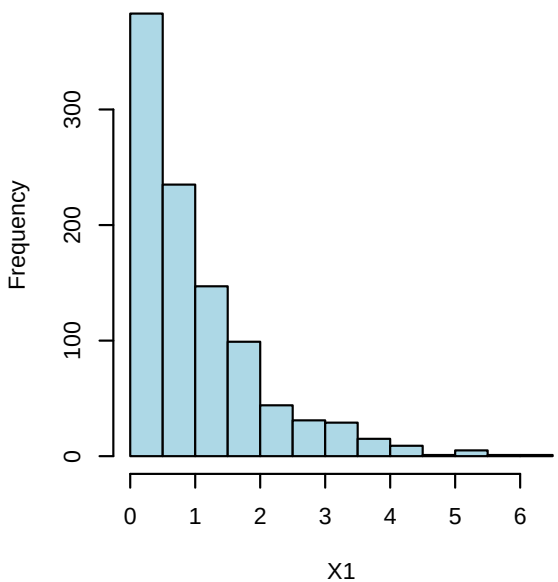


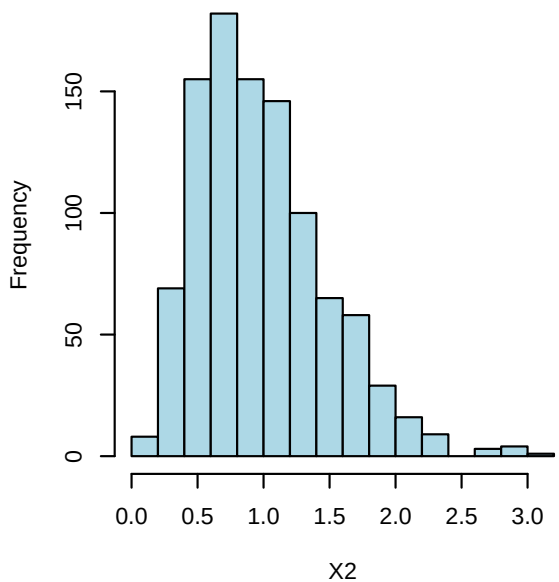


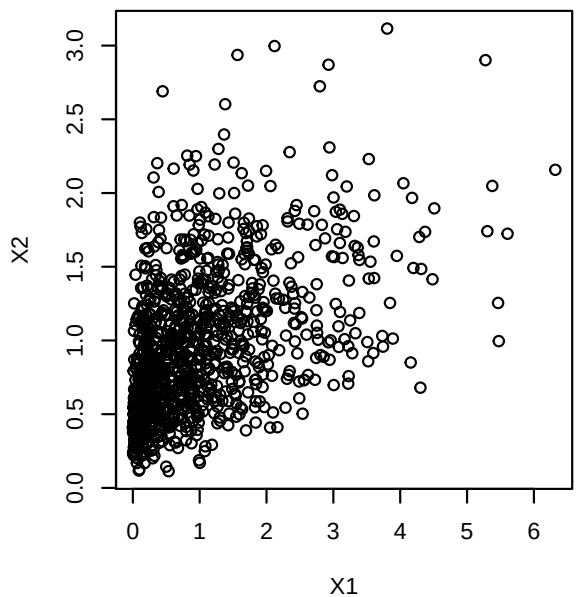


Gaussian-copula 与 Gamma 边际分布结合

```
# Gaussian copula with gamma margins
X = cbind( qexp(U[,1]), qgamma(U[,2],4,4)) # Exp(1), Gamma(4,4)
par(mfrow = c(1,3))
hist(X[,1], xlab="X1", main="", col="lightblue")
hist(X[,2], xlab="X2", main="", col="lightblue")
plot(X[,1],X[,2], type="p", xlab="X1", ylab="X2", main="")
```







Copula-marginal 方法

- Gaussian-copula 中随机向量 $\boldsymbol{X}$ 各分量间的相关性与相关系数矩阵 R 的关系是什么?
  - 如果边际分布 F_j 没有有限的方差，Cov($\boldsymbol{X}$) 或 Corr($\boldsymbol{X}$) 无法定义，需引入新的描述相关性的指标
- 定义 X_j 和 X_k 的rank correlation为 $F_j(X_j)$ 和 $F_k(X_k)$ 的 correlation
  - 由于 $F_j(X_j) = \Phi(Y_j)$，因此 $\boldsymbol{X}$ 的 rank correlation 矩阵和 $\boldsymbol{Y}$ 的相同
- 对于正态随机向量 $\boldsymbol{Y}$, McNeil et al. (2005) 给出了分量 Y_j 和 Y_k 的 rank correlation ρ_{rank} 与 $\rho_{jk} = Corr(Y_j, Y_k)$ 的关系:

$$\rho_{rank}(Y_j, Y_k) = Corr\left(\Phi(Y_j), \Phi(Y_k)\right) = \frac{2}{\pi}\arcsin(\rho_{jk})$$

- 如果希望 X_j 和 X_k 的 rank correlation 为 ρ_{rank}, 可以令 $R_{jk} = \rho_{jk} = \sin(\pi\rho_{rank}/2)$

描述相关性的常用指标

定义 (Pearson correlation)

随机变量 X 和 Y 的 Pearson correlation 定义为

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

如果 (X, Y) 有 n 对观察值 $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, 令 $\boldsymbol{x} = (x_1, \ldots, x_n)$, $\boldsymbol{y} = (y_1, \ldots, y_n)$, 则 $Corr(X, Y)$ 的样本估计量为

$$Corr(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

- Pearson correlation 测量的是两组数据向量 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的线性相关性，主要取决于它们的夹角
- 对数据 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 做相同的线性变换不会改变它们之间的 Pearson correlation, 但非线性变换一般会改变 Pearson correlation

描述相关性的常用指标

定义 (Spearman's ρ)

令 rx_i 表示 x_i 在 $\boldsymbol{x}$ 中的排序 (rank), 令 $\boldsymbol{rx} = (rx_1, \ldots, rx_n)$. 同理可得 $\boldsymbol{ry}$. 则 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的 Spearman correlation 定义为 $\boldsymbol{rx}$ 和 $\boldsymbol{ry}$ 的 Pearson correlation

$$\hat{\rho} = Corr(\boldsymbol{rx}, \boldsymbol{ry}).$$

定义 (Kendall's τ)

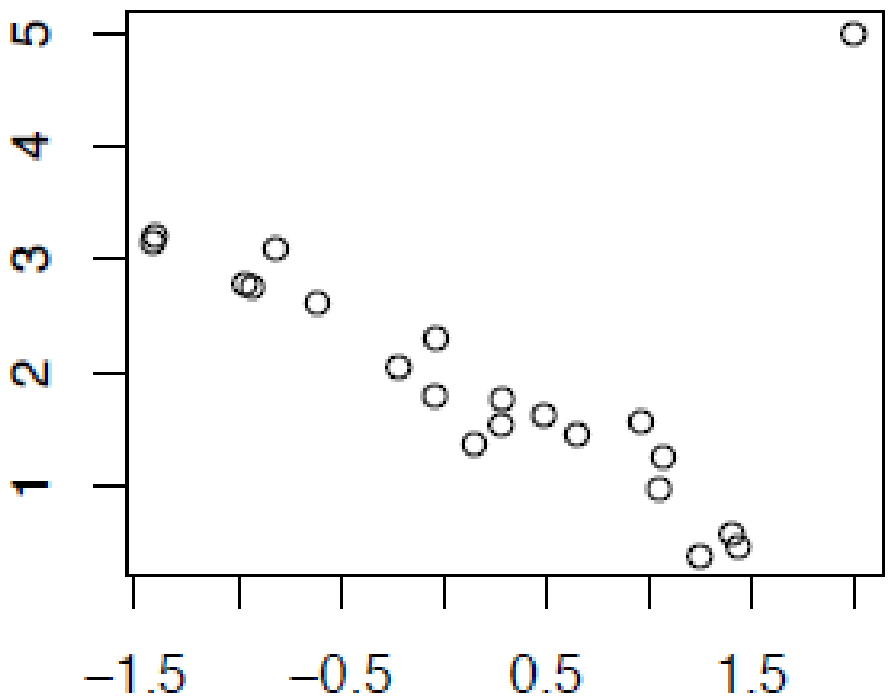
对于 (X, Y) 的任意两对观察值 (x_i, y_i) 和 (x_j, y_j), $i < j$, 如果 $x_i < x_j$ 且 $y_i < y_j$, 或者 $x_i > x_j$ 且 $y_i > y_j$, 称 (x_i, y_i) 和 (x_j, y_j) 是一致的 (concordant), 否则是不一致的 (disconcordant). 如果 $x_i = x_j$ 或者 $y_i = y_j$, 则认为 (x_i, y_i) 和 (x_j, y_j) 既不是一致的也不是不一致的. $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的 Kendall correlation 定义为

$$\tau = \frac{(\#\text{ concordant pairs}) - (\#\text{ disconcordant pairs})}{\binom{n}{2}}$$

描述相关性的常用指标

- Spearman's ρ/Kendall's τ 的取值范围 [-1,1], 只取决于数据的大小排序 (rank)，对数据做单调变换不会改变 Spearman/Kendall correlation，称这种性质为 scale-free
- Pearson correlation 很容易受到数据中异常值 (outliers) 的影响，但 Spearman's ρ/Kendall's τ 几乎不会受影响

使用上述三种指标测量以下数据的相关性会有什么不同?



t copula

- 多元正态分布抽样的便利性使 Gaussian copula 方法非常流行，但它隐含的假设是目标分布的 copula 非常接近一个正态分布的 copula

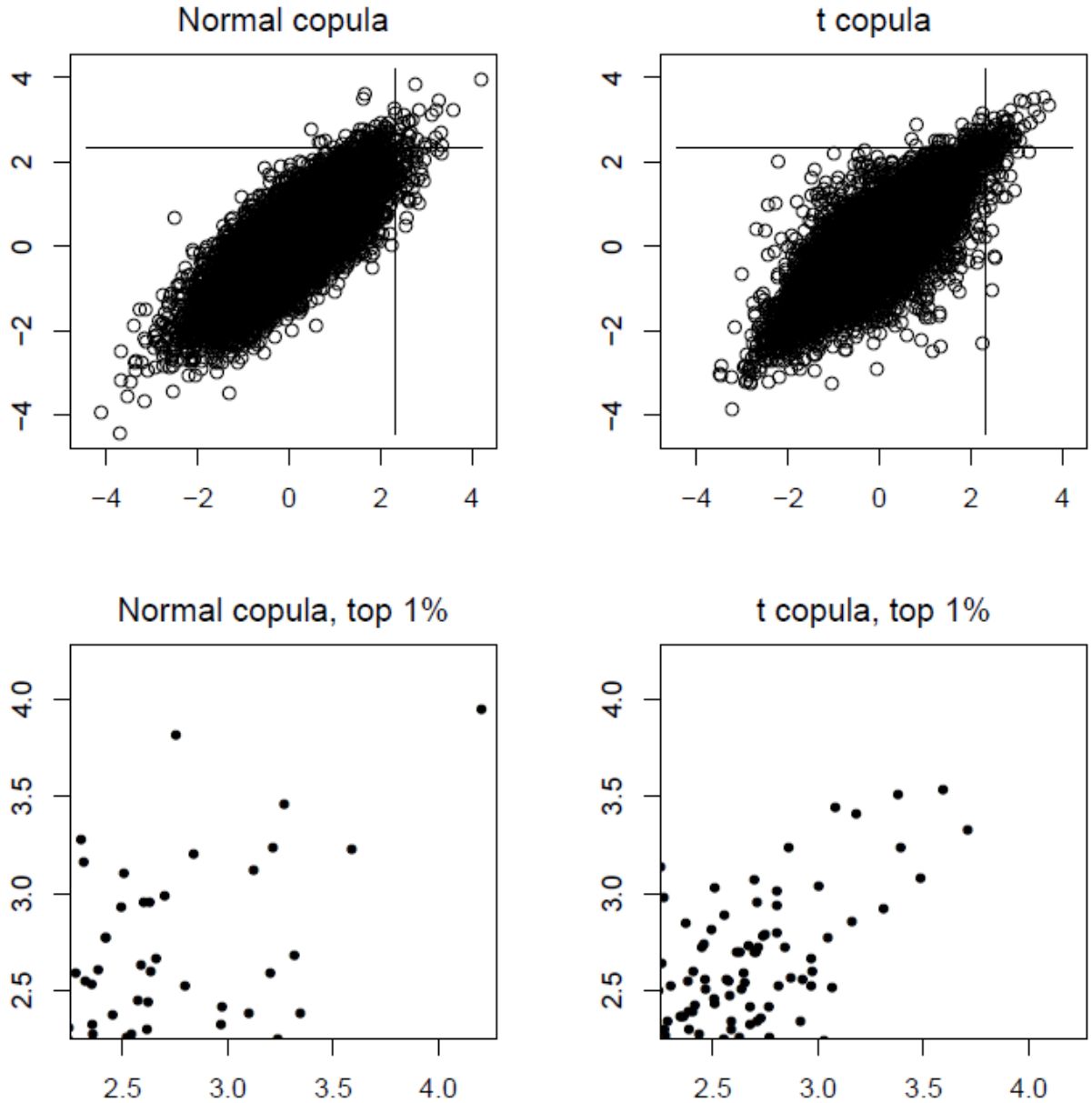
- Gaussian copula 方法的一个缺点 — 尾部独立性(tail independence)，即如果 $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$, 则任意两个分量 X_j 和 X_k 有如下性质

$$\lim_{u \to 1-} P\left(X_j > F_j^{-1}(u) \mid X_k > F_k^{-1}(u)\right) = 0$$

  极端事件下渐近独立

- t copula 可以避免尾部独立性，当数据的边际分布具有长尾时 (有 outliers)，t copula 更有优势

t copula


Normal copula
t copula
Normal copula, top 1%
t copula, top 1%
−4
−2
0
2
4
2.5
3.0
3.5
4.0

t copula

- t **copula**. 给定一个相关系数矩阵 $R \in \mathbb{R}^{d \times d}$, 自由度 $\nu > 0$, 以及 d 个边际 CDFs $F_1, \ldots, F_d$, t copula 抽样过程如下:

$$\boldsymbol{Y} \sim t_d(\mathbf{0}, R, \nu), \text{ 令} X_j = F_j^{-1}\left(T_\nu(Y_j)\right), \ j = 1, \ldots, d$$

  其中 $T_\nu()$ 是一元 $t_{(\nu)}$ 分布的 CDF

- t copula 使较大的 X_j 和 X_k 具有尾部相关性 (tail dependence), 当 X_j 和 X_k 都为很小的负数时也存在相同的相关性, 而金融市场中两只股票大涨和大跌时的尾部相关性一般是不同的

- Clayton copula 具有 lower tail dependence，即当 U_1, U_2 都很小时，它们的相关性大于它们都很大时的相关性

$$C(u_1, u_2 \mid \theta) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}$$

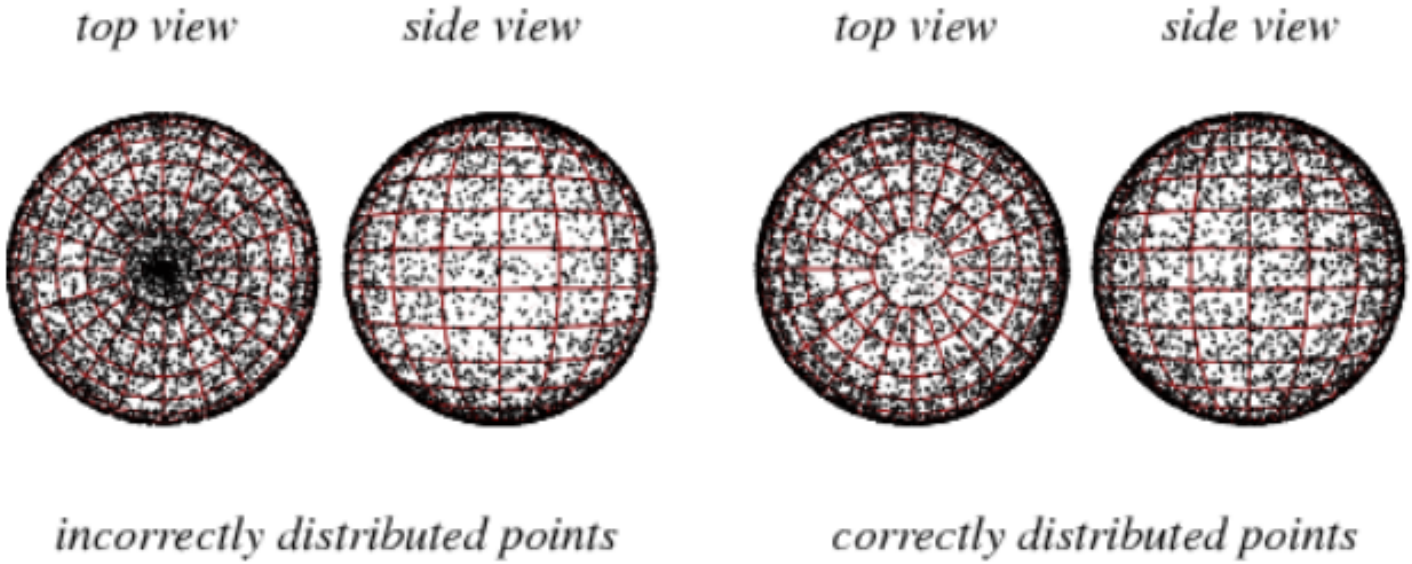
  其中参数 $\theta > 0$

球面上的随机点

- 对 d 维空间的球对称或椭球对称分布抽样一般需要先从单位超球面上均匀取点
- d 维空间的单位超球面

$$S^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\}$$

  - $d = 2$: $\boldsymbol{X} = (\cos(2\pi U), \sin(2\pi U)),\ U \sim \boldsymbol{U}(0,1).$
  - $d = 3$: $U_1, U_2 \overset{iid}{\sim} \boldsymbol{U}(0,1),\ R = \sqrt{U_1(1-U_1)},\ \theta = 2\pi U_2,$
    $\boldsymbol{X} = (2R\cos(\theta), 2R\sin(\theta), 1 - 2U_1)$



球面上的随机点

- $d > 3$ 时，对单位超球面上均匀分布 $\boldsymbol{X} \sim \boldsymbol{U}(S^{d-1})$ 的一种简便抽样方法：

$$\boldsymbol{X} = \boldsymbol{Z} / \|\boldsymbol{Z}\|, \ \boldsymbol{Z} \sim N_d(0, I_d). \tag{3}$$

- 知道如何从球面上均匀取点，就可以从任意一个球对称分布中抽样，只要知道如何抽取目标随机向量的模长 $R = \|\boldsymbol{X}\|$
  - **例**. 如果 $\boldsymbol{X} \sim f(\boldsymbol{x}) \propto \exp(-\|\boldsymbol{x}\|)$，如何得到 $\boldsymbol{X}$ 的样本?
  - 练习. 如果 $\boldsymbol{X} \sim f(\boldsymbol{x}) \propto \|\boldsymbol{x}\|^k \mathbf{1}\{\|\boldsymbol{x}\| \leq 1\}$, $k > -d$. 如何得到 $\boldsymbol{X}$ 的样本?
  - 当 $f(\boldsymbol{x}) \propto h(\|\boldsymbol{x}\|)$ 时，如果不能识别 PDF $\propto r^{d-1} h(r)$ 的分布，可以尝试 A-R 方法
- 对球对称分布做线性变换可以得到椭球对称分布

球面上的非均匀分布

一个常用的球面 S^{d-1} 上的非均匀分布是von Mises-Fisher分布, PDF:

$$f(\boldsymbol{x}) \propto \exp\left(\kappa \boldsymbol{\mu}^\top \boldsymbol{x}\right)$$

- 参数 $\kappa \geq 0$, 向量 $\boldsymbol{\mu} \in S^{d-1}$, $\kappa > 0$ 时 von Mises-Fisher 分布在点 $\boldsymbol{\mu}$ 处的概率密度最大, κ 越大 von Mises-Fisher 分布越集中在 $\boldsymbol{\mu}$ 附近
- R Package `rstiefel` 可对 von Mises-Fisher 分布抽样，抽样的关键在于对随机变量 $W = \boldsymbol{\mu}^\top \boldsymbol{X}$ 的抽样，算法总结如下:

$$W \sim h(w) \propto (1-w^2)^{(d-3)/2} \exp(\kappa w) \mathbf{1}\{w \in (-1,1)\}$$
$$\boldsymbol{V} \sim \boldsymbol{U}(S^{d-2})$$
$$\boldsymbol{X} = W\boldsymbol{\mu} + \sqrt{1-W^2} B\boldsymbol{V}$$

其中对 W 使用 A-R 方法抽样 (选取经过变换的 Beta 分布)，矩阵 $B \in \mathbb{R}^{d\times(d-1)}$ 由与 $\boldsymbol{\mu}$ 垂直的 $(d-1)$ 个单位正交向量组成，$B\boldsymbol{V}$ 是在与 $\boldsymbol{\mu}$ 垂直的方向上均匀分布的单位向量