随机向量的生成

王璐

在上一章我们介绍了生成一元随机变量的方法，本章将讨论如何生成 $\mathbb{R}^d$ $(d > 1)$ 上的随机向量。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。

对于一元随机变量，我们在上一章介绍了三种主要的抽样方法：CDF 逆变换、A-R 方法和混合抽样。它们都可以推广到多元情形，然而实践中除了几个成功的特例，使用这些方法进行多元抽样的效率通常很低。因此人们又提出了 Markov chain Monte Carlo, Sequential Monte Carlo 等方法，我们后面会介绍。

本章我们将重点关注一些常用的多元分布，比如多元正态分布、多元 t 分布、Dirichlet 分布以及多项分布 (multinomial distribution) 等，对这些多元分布进行抽样已经有非常高效的方法。

除了介绍上述几种多元分布的抽样方法，我们还将介绍一种更一般的多元抽样方法 — copula-marginal 方法。它可以看作一元的 QQ 变换推广到多元的方法，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。当然这一过程是很复杂的，因此人们也将这一方法视为第四种主要抽样方法。

最后我们还会介绍一些随机矩阵的抽样方法。这些随机矩阵可以看作若干有相关结构的随机向量的集合，比如在球面上随机分布的点，Wishart 矩阵，随机图等。

本章我们假设维度 $d < \infty$, 下一章我们将讨论 $d = \infty$ 的情况，即随机过程的产生。

1 一元抽样方法的推广

本节我们将讨论如何将三种主要的一元抽样方法 — CDF 逆变换，A-R 方法，混合抽样 — 推广到多元抽样。

1.1 CDF 逆变换

直接抽取随机向量 $\boldsymbol{X} \in \mathbb{R}^d$ 的方法是依次抽取它的各元素 $X_1, X_2, \ldots X_d$. 首先从 X_1 的边际分布抽样, 然后给定已有元素的样本，从条件分布中抽取下一个元素的样本，原理是

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_1(x_1) f_{2|1}(x_2 \mid x_1) f_{3|1:2}(x_3 \mid x_{1:2}) \cdots f_{d|1:(d-1)}(x_d \mid x_{1:(d-1)}). \tag{1}$$

根据序列生成形式(1)，我们可以构造如下的 CDF 逆变换法在 d 维空间抽样，称为 sequential inversion. 首先抽取 d 个 $\boldsymbol{U}(0,1)$ 的样本 $U_j \overset{iid}{\sim} \boldsymbol{U}(0,1),\ j=1,2,\ldots,d.$ 然后依次令

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

就得到随机向量 $\boldsymbol{X}$ 的一个样本。可以看到，使用 sequential inversion 需要知道 X_1 边际分布的 CDF 以及其它序列条件分布的 CDF. 我们来看一个具体例子。

- **Example**. 随机向量 $\boldsymbol{X}=(X_1,X_2)$ 的 PDF 为

$$f(x_1,x_2)=\begin{cases} x_1+x_2, & (x_1,x_2)\in[0,1]^2 \\ 0, & \text{else.} \end{cases}$$

  下面用 sequential inversion 对 $\boldsymbol{X}$ 进行抽样：

  - X_1 的边际 PDF 为 $f_1(x_1)=\int_0^1 f(x_1,x_2)dx_2 = x_1+1/2$. 因此 X_1 的边际 CDF 为 $F_1(x)=\int_0^x f_1(t)dt=(x^2+x)/2,\ 0\le x\le 1$. 利用二次方程求根公式，可得 F_1 的逆函数:

$$X_1=F_1^{-1}(U_1)=\sqrt{2U_1+1/4}-1/2,\ U_1\sim\boldsymbol{U}(0,1).$$

  - 给定 $X_1=x_1$, X_2 的条件分布的 PDF 为

$$f_{2|1}(x_2\mid x_1)=\frac{f(x_1,x_2)}{f_1(x_1)}=\frac{x_1+x_2}{x_1+1/2},\ 0\le x_2\le 1.$$

    X_2 的条件分布的 CDF 为

$$F_{2|1}(x\mid x_1)=\int_0^x f_{2|1}(t\mid x_1)dt=\frac{x_1x+x^2/2}{x_1+1/2}.$$

    再次利用二次方程求根公式，可得 $F_{2|1}$ 的逆变换为:

$$X_2=F_{2|1}^{-1}(U_2\mid X_1)=\sqrt{X_1^2+(2X_1+1)U_2}-X_1,\ U_2\sim\boldsymbol{U}(0,1).$$

Remarks

1. 使用 sequential inversion 进行多元抽样在实践中经常面临的问题是：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。而且如果每个条件分布的逆函数都需要重新计算，不能利用前面的结果，使用 sequential inversion 抽样会很慢。

2. 也有一些分布使用 sequential inversion 抽样很容易，比如后面会介绍的多项分布。

1.2 Acceptance-rejection (A-R)

A-R 方法也很容易推广到多元。如果想从 $\mathbb{R}^d$ 上的分布 $f(\boldsymbol{x})$ (PDF) 中抽样，可以先从另一个容易抽样的分布 $g(\boldsymbol{x})$ (PDF) 中抽样，只要保证存在常数 c 使得 $f(\boldsymbol{x}) \le cg(\boldsymbol{x})$. 容易证明在多元情形下，A-R 中来自 g 的样本总体被接受的概率也是 $1/c$. 即平均从 g 中抽取 c 个样本，才有一个被接受作为 f 的样本。

A-R 的几何解释在多元情形下依然成立。令

$$S_c(f) = \left\{(\boldsymbol{x}, z) \mid 0 \le z \le cf(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^d\right\}$$

表示一个 $(d+1)$ 维的闭集。如果 $(\boldsymbol{X}, Z) \sim \boldsymbol{U}(S_c(f))$, 则 $\boldsymbol{X} \sim f$. 反过来，如果随机向量 $\boldsymbol{X} \sim f$ 且 $Z \mid \boldsymbol{X} = \boldsymbol{x} \sim \boldsymbol{U}(0, cf(\boldsymbol{x}))$, 则 $(\boldsymbol{X}, Z) \sim \boldsymbol{U}(S_c(f))$.

A-R 的几何解释保证了我们可以使用 f 和 g 未归一化的形式 — $\tilde{f}$ 和 $\tilde{g}$ — 计算来自 g 的样本被接受的概率:

$$\boldsymbol{Y} \sim g,\ A(\boldsymbol{Y}) = \frac{\tilde{f}(\boldsymbol{Y})}{\tilde{c}\tilde{g}(\boldsymbol{Y})}$$

只要保证 $\tilde{f}(\boldsymbol{y}) \le \tilde{c}\tilde{g}(\boldsymbol{y}), \forall \boldsymbol{y}$.

- **Example**. 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. 它们的 unnormalized PDF 为 $\tilde{f}(\boldsymbol{x}) = \mathbf{1}(\boldsymbol{x} \in B_d)$ 和 $\tilde{g}(\boldsymbol{x}) = \mathbf{1}(\boldsymbol{x} \in [-1,1]^d)$, 因此在 A-R 中选取 $\tilde{c} = 1$ 即可。此时抽取 $\boldsymbol{Y} \sim g$ 后只保留 $\|\boldsymbol{Y}\| \le 1$ 的样本, 则来自 g 的样本总体被接受的概率为
$$\frac{\mathrm{vol}(B_d)}{2^d} = \frac{\pi^{d/2}}{2^d \Gamma(1 + d/2)}.$$
  - $d = 2$ 时，上述接受概率为 $\pi/4 \approx 0.785$, 比较高。
  - $d = 9$ 时，上述接受概率 <1%; $d = 23$ 时，接受概率 $< 10^{-9}$.

- **Example**. 假设 f 和 g 都可写为 d 个一元 PDF 的乘积 (各分量独立), 且存在 $c_j = \sup_x f_j(x)/g_j(x)$, $j = 1, 2, \ldots, d$. 则在 A-R 中可选取常数 c 为
$$c = \sup_{\boldsymbol{x}} f(\boldsymbol{x})/g(\boldsymbol{x}) = \prod_{j=1}^{d} c_j$$
  如果每个 $c_j > 1 + \epsilon$, 则 c 将随着 d 指数增长。

Remark

- 通过上述例子，可以看到 A-R 方法在多元抽样中经常面临的问题是:
  - 在高维情形下一般很难找到较小的 c，抽样效率很低。
  - 计算 $c = \sup_{\boldsymbol{x}} f(\boldsymbol{x})/g(\boldsymbol{x})$ 很复杂，一般需要解一个 d 维优化。

1.3 混合抽样

混合抽样（mixture sampling）很容易推广到高维，有时也能使多元抽样变得简单。如果多元分布的 PDF 可以写为如下的连续混合形式

$$f(\boldsymbol{x}) = \int f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x} \mid \boldsymbol{y}) g(\boldsymbol{y}) d\boldsymbol{y},$$

则可先抽取 $\boldsymbol{Y} \sim g$, 给定 $\boldsymbol{Y}$ 再从条件分布抽样 $\boldsymbol{X} \mid \boldsymbol{Y} \sim f_{\boldsymbol{X}|\boldsymbol{Y}}$，即得 $\boldsymbol{X} \sim f$ 的样本。

如果 $f(\boldsymbol{x})$ 可写为如下的离散混合形式

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x})$$

其中 $\pi_k \geq 0$ 且 $\sum_{k=1}^{K} \pi_k = 1$; 每个 f_k 都是 $\mathbb{R}^d$ 上的一个 PDF. 则可如下对 f 抽样：先对一个离散的随机变量 Z 抽样，$P(Z = k) = \pi_k$, $k = 1, 2, \ldots, K$; 给定 $Z = k$, 再从 f_k 抽样 $X \mid Z = k \sim f_k$.

2 多元正态分布 (Multivariate normal)

多元正态分布是最重要的多元分布之一。$\mathbb{R}^d$ 上的多元正态分布由一个期望向量 $\boldsymbol{\mu} \in \mathbb{R}^d$ 和一个半正定的协方差矩阵 $\Sigma \in \mathbb{R}^{d \times d}$ 决定，记为 $N_d(\boldsymbol{\mu}, \Sigma)$. 如果 Σ 可逆，$N_d(\boldsymbol{\mu}, \Sigma)$ 的 PDF 为

$$\phi(\boldsymbol{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})' \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right\}}{(2\pi)^{d/2} |\Sigma|^{1/2}}, \ \boldsymbol{x} \in \mathbb{R}^d$$

如果 Σ 不可逆，说明有些成分是多余的，即存在 $k \in \{1, \ldots, d\}$ 满足

$$P(X_k = \alpha_0 + \sum_{j \neq k} \alpha_j X_j) = 1.$$

以下我们只讨论没有多余成分的正态分布，即 Σ 是可逆的情形。

多元正态分布有很多重要性质（后面会经常用到）：

1. 如果 $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$，则 $A\boldsymbol{X} + \boldsymbol{b} \sim N_d(A\boldsymbol{\mu} + \boldsymbol{b}, A\Sigma A^\top)$.

2. 将 $\boldsymbol{X}$ 分为不相交的两个子向量 $\boldsymbol{X}_1 = (X_1, \ldots, X_r)$ 和 $\boldsymbol{X}_2 = (X_{r+1}, \ldots, X_d)$, 对参数也做相应地划分

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

   则它们各自的边际分布为 $\boldsymbol{X}_j \sim N(\boldsymbol{\mu}_j, \Sigma_{jj})$, $j = 1, 2$.

3. 上述 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$ 独立当且仅当 Σ_{12} 是零矩阵。

4. 给定 $\boldsymbol{X}_2 = \boldsymbol{x}_2$, $\boldsymbol{X}_1$ 的条件分布为

$$\boldsymbol{X}_1 \mid \boldsymbol{X}_2 = \boldsymbol{x}_2 \sim N_d\left(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

从 $N_d(\mathbf{0}, I_d)$ 中抽样很容易，因为此时各分量的相关性都为 0，在多元正态分布中，这意味着各分量都是独立的。因此可以使用 Box-Muller 或 CDF 逆变换独立地从 $N(0,1)$ 中抽样, $Z_j \overset{iid}{\sim} N(0,1)$, $j = 1, \ldots, d$, 则 $\boldsymbol{Z} = (Z_1, \ldots, Z_d)^\top \sim N_d(\mathbf{0}, I_d)$.

对一般的多元正态分布 $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ 抽样, 只需找到矩阵 C 使得 $\Sigma = CC^\top$, 然后对 $\boldsymbol{Z}$ 做线性变换即可:

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

上述矩阵 C 总可以通过特征值分解获得。由于 Σ 是对称的半正定矩阵，因此存在特征值分解

$$\Sigma = P\Lambda P^\top$$

其中 Λ 是对角阵且对角线元素非负 $\Lambda_{ii} \geq 0$. 因此可令 $C = P\Lambda^{1/2}$. 矩阵 C 的选择并不唯一，对于任意正交矩阵 Q, 令 $\tilde{C} = CQ$, 则 $\tilde{C}\tilde{C}^\top = CQQ^\top C^\top = CC^\top = \Sigma$.

由于 Σ 是半正定的，人们也经常使用 Cholesky 分解 $\Sigma = LL^\top$，然后令 $C = L$，其中 L 是下三角矩阵。当 Σ 正定时，Cholesky 分解是唯一的，此时 L 的对角线元素全部为正。

特征值分解和 Cholesky 分解的计算量都是 $O(d^3)$.

3 多元 t 分布

多元 t 分布有三部分参数，center $\boldsymbol{\mu}$, scale matrix Σ 和自由度 ν, 记为 $t_d(\boldsymbol{\mu}, \Sigma, \nu)$. 当 $\nu = 1$ 时，多元 t 分布也称多元 Cauchy 分布；当 $\nu \to \infty$ 时，$t_d(\boldsymbol{\mu}, \Sigma, \nu)$ 收敛到 $N_d(\boldsymbol{\mu}, \Sigma)$.

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

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

其中归一化常数为

$$C_{\boldsymbol{\mu},\Sigma,\nu} = \frac{\Gamma\left((\nu + d)/2\right)}{|\Sigma|^{1/2}(\nu\pi)^{d/2}\Gamma(\nu/2)}.$$

对于标准的多元 t 分布，$\boldsymbol{\mu} = \mathbf{0}$, $\Sigma = I$, 此时 $f(\boldsymbol{x}) \propto \left(1 + \|\boldsymbol{x}\|^2\right)^{-(\nu+d)/2}$. 但是 $\Sigma = I_d$ 的多元 t 分布的各分量并不独立。

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

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

与多元正态分布相似，多元 t 分布 $t_d(\boldsymbol{\mu}, \Sigma, \nu)$ 的 PDF 的形状是以 $\boldsymbol{\mu}$ 为中心的一系列椭圆等高线，但多元 t 分布依然比多元正态分布的尾厚。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.

4 多项分布 (Multinomial)

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 $\mathrm{Mult}(m, p_1, \ldots, p_d)$. 它的 PMF 为

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

其中 x_j 为非负整数且满足 $\sum_{j=1}^d x_j = m$, 概率 $p_j \geq 0$ 且 $\sum_{j=1}^d p_j = 1$. 因此参数向量 $\boldsymbol{p} = (p_1, \ldots, p_d)$ 可取值的集合为

$$\Delta^{d-1} = \left\{ (p_1, \ldots, p_d) \mid p_j \geq 0, \sum_{j=1}^{d} p_j = 1 \right\}$$

Δ^{d-1} 被称为 $\mathbb{R}^d$ 上的 unit simplex. Δ 的上标 $d-1$ 表示该集合的真实维度是 $d-1$.

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

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

其中 X_1 的边际分布是一个二项分布 $X_1 \sim \mathrm{Bin}(m, p_1)$; 给定 $\{X_1, \ldots, X_{j-1}\}$, X_j 的条件分布也是一个二项分布：此时可能落入格子 j 的球数变为 $m - \sum_{s=1}^{j-1} X_s$, 且这些球只能落入格子 $j, \ldots, d$, 因此每个球落入格子 j 的概率增大为 $p_j / \sum_{k=j}^d p_k$, 所以

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

上述抽样方法可以用算法1实现。

Algorithm 1 Sample $\boldsymbol{X} \sim \text{Mult}(m, p_1, \ldots, p_d)$

Input $m \in \mathbb{N}$, $d \in \mathbb{N}$, $\boldsymbol{p} = (p_1, \ldots, p_d) \in \Delta^{d-1}$.

Let $n = m$ and $S = 1$.

for $j = 1$ to d **do**

$X_j = \text{Bin}(n, p_j/S)$

$n = n - X_j$ ▷ 如果在某步迭代中发现 $n = 0$, 则可直接将后面的分量取为 0.

$S = S - p_j$

return $\boldsymbol{X} = (X_1, \ldots, X_d)$

5 Dirichlet 分布

有时要抽样的随机向量可能是一组随机概率，比如从多项分布的参数空间抽样，此时抽样的样本空间是一个 unit simplex

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

Dirichlet 分布是定义在 unit simplex Δ^{d-1} $(d \geq 2)$ 上最简单的分布之一，它有 d 个参数：$\alpha_j > 0$, $j = 1, \ldots, d$, 记为 $\text{Dir}(\alpha_1, \ldots, \alpha_d)$. 它的 PDF 为

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

其中归一化常数 $D(\boldsymbol{\alpha}) = \prod_{j=1}^{d} \Gamma(\alpha_j) / \Gamma(\sum_{j=1}^{d} \alpha_j)$. $\boldsymbol{X} \sim \text{Dir}(\boldsymbol{\alpha})$ 的期望为

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

Dirichlet 分布有两个特例值得说明：

1. $d = 2$ 时的 Dirichlet 分布 $\text{Dir}(\alpha_1, \alpha_2)$ 等价于 $\text{Beta}(\alpha_1, \alpha_2)$ 分布，即

$$(X_1, X_2) \sim \text{Dir}(\alpha_1, \alpha_2) \ \Leftrightarrow \ X_1 \sim \text{Beta}(\alpha_1, \alpha_2), X_2 = 1 - X_1$$

   且此时 $X_2 \sim \text{Beta}(\alpha_2, \alpha_1)$.

2. $\alpha_j \equiv 1$, $j = 1, \ldots, d$ 对应的 Dirichlet 分布是 Δ^{d-1} 上的均匀分布 $\boldsymbol{U}(\Delta^{d-1})$.

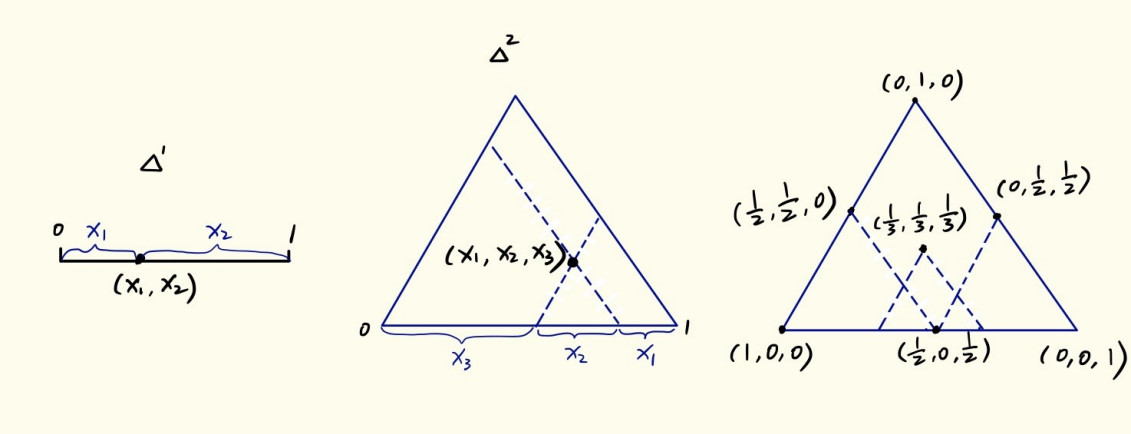


Figure 1: Δ^1 和 Δ^2 空间

$d = 2$ 时的样本空间 Δ^1 对应一个长度为 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，如图1所示。

图2展示了 6 组不同参数向量 $\boldsymbol{\alpha} \in \mathbb{R}^3_+$ 对应的 Dir($\boldsymbol{\alpha}$) 的样本。可以看到，样本倾向于分布在最大的 α_j 对应的角附近。比较 Dir(1,1,1), Dir(7,7,7) 和 Dir(0.2,0.2,0.2) 的样本分布，虽然这三个分布的期望相同，但样本的表现却很不同：$\alpha_j \equiv 1$ 对应 Δ^2 上的均匀分布；较大的 α_j's 倾向于让样本更靠近中心，即分布的期望；较小的 α_j's 倾向于让样本更靠近边界，边界上的点会有某个分量为 0。

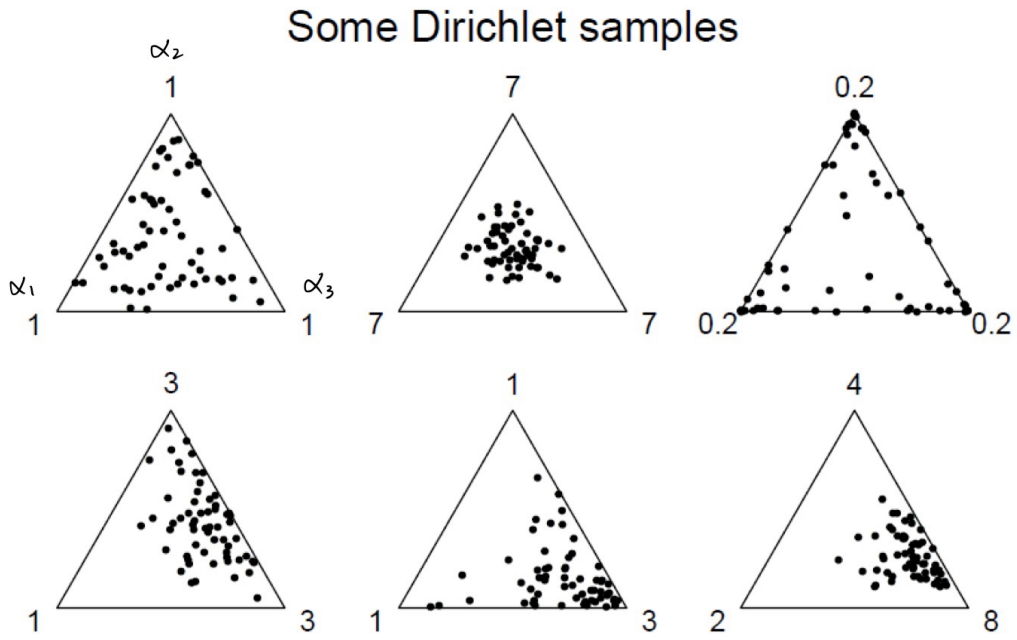


Figure 2: 6 组不同 $\boldsymbol{\alpha} \in \mathbb{R}^3_+$ 对应的 Dir($\boldsymbol{\alpha}$) 样本 (每组 60 个)。$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ 的取值标在三角形的各角上。Picture source: Art B. Owen.

我们在上一章介绍了用 Gamma 分布生成 Beta 分布的方法。类似地，Dirichlet 分布也可以由 Gamma 分布生成，方法如下：

$$\begin{aligned} &Y_j \overset{ind}{\sim} \text{Gam}(\alpha_j),\ j=1,\ldots,d, \quad \text{then} \\ &X_j = \frac{Y_j}{\sum_{k=1}^d Y_k},\ j=1,\ldots,d. \end{aligned} \tag{2}$$

则 $\boldsymbol{X} \sim \text{Dir}(\boldsymbol{\alpha})$，其中 $\text{Gam}(\alpha_j) = \text{Gam}(\alpha_j,1)$. 由(2)可得 $\text{Dir}(\boldsymbol{\alpha})$ 的边际分布为

$$X_j \sim \text{Beta}(\alpha_j, \sum_{k\neq j} \alpha_k),\ j=1,\ldots,d. \tag{3}$$

因为根据 Gamma 分布的性质，$Y_{-j} = \sum_{k\neq j} Y_k \sim \text{Gam}(\sum_{k\neq j} \alpha_k)$ 且与 Y_j 独立, 而 $X_j = Y_j/(Y_j + Y_{-j})$, 所以得到(3). 因此 Dirichlet 分布也可以看作多元 Beta 分布。

- 对于 $\alpha_j \equiv 1,\ j=1,\ldots,d$ 的 Dirichlet 分布，即 $\boldsymbol{U}(\Delta^{d-1})$，此时 Gam(1) 即为 Exp(1) 分布，而 Exp(1) 的样本可由变换 $Y_j = -\log(U_j),\ U_j \sim \boldsymbol{U}(0,1)$ 得到，则可令

$$X_j = \frac{\log(U_j)}{\sum_{k=1}^d \log(U_k)},\ U_j \overset{iid}{\sim} \boldsymbol{U}(0,1).$$

- $\boldsymbol{U}(\Delta^{d-1})$ 还可以使用 **uniform spacings** 方法抽样。令

$$U_j \overset{iid}{\sim} \boldsymbol{U}(0,1),\ j=1,\ldots,d-1$$

它们对应的 order statistics 为 $U_{(1)} \le U_{(2)} \le \cdots \le U_{(d-1)}$. 再扩展两个点 $U_{(0)} = 0,\ U_{(d)} = 1$, 然后令

$$X_j = U_{(j)} - U_{(j-1)},\ j=1,\ldots,d$$

则 $\boldsymbol{X} \sim \boldsymbol{U}(\Delta^{d-1})$. 该方法只需产生 $d-1$ 个随机变量且避免了对数运算，但是排序的计算量为 $O(d\log(d))$, 因此对很大的 d, uniform spacings 可能比从指数分布抽样慢。

Remarks

1. Dirichlet 分布不是一个很灵活的分布，它只有 d 个参数，而期望 $E(\boldsymbol{X}) = \boldsymbol{\alpha}/\sum_{j=1}^d \alpha_j$ 用掉了 $d-1$ 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。因此没有足够的参数让 $\boldsymbol{X}$ 各分量的方差自由变化，更不用说它们之间的 $d(d-1)/2$ 对相关关系。

2. 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。因此不能用 Dirichlet 分布产生 Δ^{d-1} 上分量间有正相关的样本，或是分量间存在很大负相关的样本。

6 Copula-marginal 方法

不是所有的一元分布都可以像正态分布或 t 分布那么容易地推广到多元，Kotz et al. (2000) 就给出了 12 种二元 Gamma 分布。很多一元分布的多元推广形式不唯一的原因是：不能保证一元分布的所有性质在推广到多元时都是相容的，即一种性质推广到多元时不能保证另一种性质还存在。本节将介绍一种较通用的多元抽样方法，或者一种多元分布的构造方法 — copula-marginal 方法，它可以看作一元的 CDF 逆变换在多元的推广。

对于 $\mathbb{R}^d$ 上的随机向量 $\boldsymbol{X} \sim F$，用一元函数 $F_j(x)$ 表示它的分量 X_j 的边际 CDF. 为了简化讨论，这里假设每个 F_j 都是连续函数，因此 $F_j(X_j) \sim \boldsymbol{U}(0,1)$, $j = 1, \ldots, d$. 它们组成的随机向量记为 $\boldsymbol{U} = (F_1(X_1), \ldots, F_d(X_d))$，注意 $\boldsymbol{U}$ 的各分量间一般不独立。我们将 $\boldsymbol{U}$ 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：

$$\begin{aligned} &\text{Sample } \boldsymbol{U} \sim C, \text{ then} \\ &X_j = F_j^{-1}(U_j),\ j = 1, \ldots, d. \end{aligned} \tag{4}$$

Definition 6.1 (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 方法的理论基础是 Sklar 定理 (Sklar, 1959).

Theorem 1 (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 $F_1, \ldots, F_d$. 总可以找到一个 copula C 使得*

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

如果 F_j's 都是连续的，则 copula C 是唯一的；否则 C 只在 F_j's 的取值范围上唯一确定。

Sklar 定理告诉我们：对任意多元分布 F, 存在一个 copula C 使得通过变换(4)可以得到 $\boldsymbol{X} \sim F$. 但使用(4)的困难在于如何确定 $\boldsymbol{U}$ 中各分量的相关性。假设分布 F 与另一多元分布 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

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

其中 G_j 是 Y_j 的边际 CDF, 然后令

$$X_j = F_j^{-1}\left(G_j(Y_j)\right),\ j = 1, \ldots, d$$

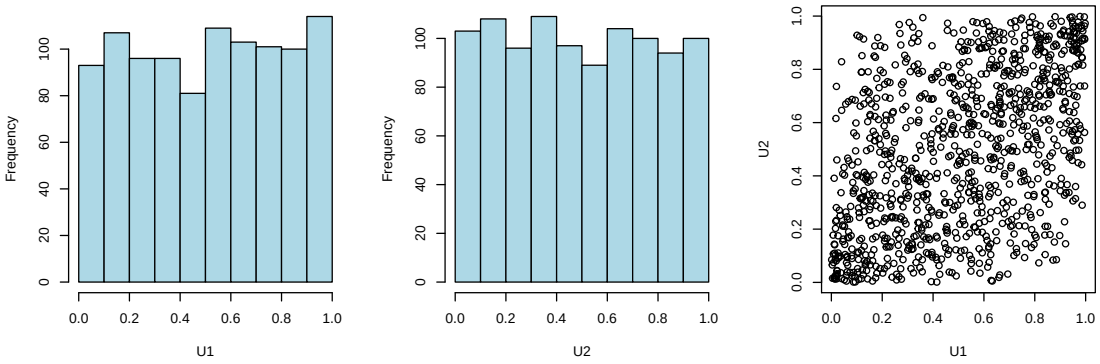
则 $\boldsymbol{X} \sim F$. 最常选取的 G 是多元正态分布。

- **Gaussian copula**. 给定一个相关系数（correlation）矩阵 $R \in \mathbb{R}^{d\times d}$ (对角线元素为 1), 以及 d 个边际 CDFs $F_1, \ldots, F_d$, Gaussian copula 抽样方法如下：
  - 首先独立抽取 $U_j \overset{iid}{\sim} \boldsymbol{U}(0,1),\ j = 1, \ldots, d.$
  - 令 $\boldsymbol{Y} = R^{1/2}\left(\Phi^{-1}(U_1), \ldots, \Phi^{-1}(U_d)\right)^\top$, 则 $\boldsymbol{Y} \sim N_d(\boldsymbol{0}, R)$ 且 $Y_j \sim N(0,1),\ j = 1, \ldots, d.$
  - 令 $X_j = F_j^{-1}\left(\Phi(Y_j)\right),\ j = 1, \ldots, d.$

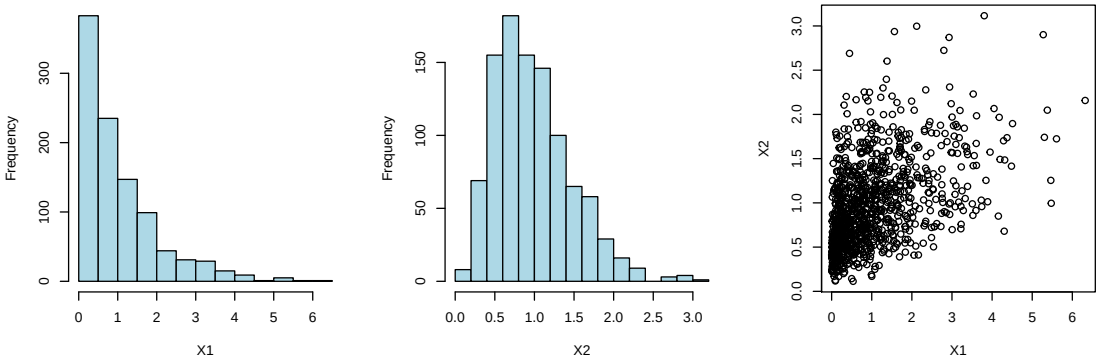
其中 $\Phi()$ 是 N(0,1) 的 CDF. 基于多元正态分布抽样的便利性，Gaussian copula 方法非常流行，但是它隐含的假设是分布 F 的 copula 非常接近于一个正态分布的 copula. 如果实际数据不支持上述假设，则 Gaussian copula 方法不适用。

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). 下面来看一个将 Gaussian-copula 与 Gamma 边际分布结合的例子。

```
## -- 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 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="")
```



虽然上述正态随机向量 $\boldsymbol{Y} \sim N_d(\mathbf{0}, R)$, 但经过变换的 $\boldsymbol{X}$ 的协方差矩阵一般不是 R.

```
> cov(X)
[,1]       [,2]
[1,] 0.9911238 0.2180027
[2,] 0.2180027 0.2374729
> cor(X)
```

```
[,1]       [,2]
[1,] 1.0000000 0.4493564
[2,] 0.4493564 1.0000000
```

有时上述边际分布 F_j 可能没有有限的方差，此时 $Cov(\boldsymbol{X})$ 或 $Corr(\boldsymbol{X})$ 无法定义，需要引入一个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. 注意到当每个 F_j 都连续时，$F_j(X_j) = \Phi(Y_j)$，因此上述 $\boldsymbol{X}$ 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：

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

如果我们希望 X_j 和 X_k 的 rank correlation 为 ρ_{rank}, 对应的 Y_j 和 Y_k 的 rank correlation 也为 ρ_{rank}，则可以令 $R_{jk} = \rho_{jk} = \sin(\pi\rho_{rank}/2)$. 因此，如果给定随机向量 $\boldsymbol{X}$ 的各边际分布和各分量间的 rank correlation matrix, 就可以用 Gaussian copula 方法生成满足上述条件的样本。

我们回顾一下描述随机变量之间相关性的一些常用指标。

Definition 6.2 (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

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

Remarks

1. 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：如果 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的样本均值都为 0，样本方差都为 1，则 $\hat{\rho}_{X,Y} = \boldsymbol{x}^\top \boldsymbol{y}$.

2. 因此对数据 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 做相同的线性变换，比如平移或线性缩放，不会改变它们之间的 Pearson correlation. 但是如果做非线性变换，即使是单调变换，Pearson correlation 一般也会改变。

如果我们不希望随机变量间的相关性受到数据测量单位的影响 (有些数据可能经过非线性单调变换得到)，或者对一些离散数据，比如 0-1 结果、计数值（counts）或者有序分类数据（ordered categories）等 Pearson correlation 不太适用的情况，可以使用下面两种相关性指标。

Definition 6.3 (Spearman's ρ)**.** 令 rx_i 表示 x_i 在 $\boldsymbol{x}$ 中的 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

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

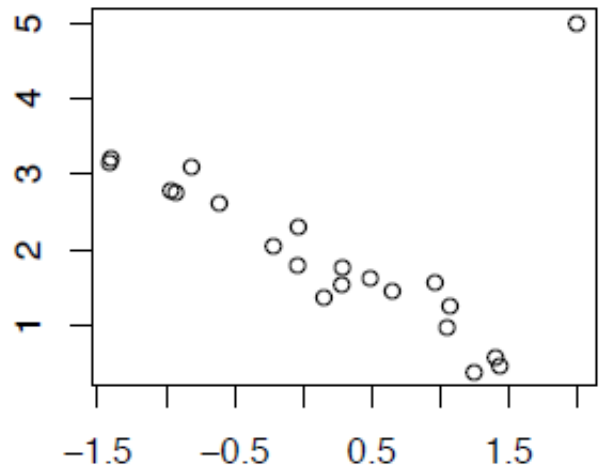
Definition 6.4 (Kendall's τ)**.** 对于 (X, Y) 的任意两对观察值 (x_i, y_i) 和 (x_j, y_j), $i < j$, 如果 $x_i < x_j$ 且 $y_i < y_j$, 或者 $x_i > x_j$ 且 $y_i > y_j$, 称 (x_i, y_i) 和 (x_j, y_j) 是一致的 (concordant), 否则是不一致的 (disconcordant). 如果 $x_i = x_j$ 或者 $y_i = y_j$, 则认为 (x_i, y_i) 和 (x_j, y_j) 既不是一致的也不是不一致的. $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的 Kendall correlation 定义为

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

Remarks

1. Spearman's ρ 和 Kendall's τ 的取值范围都是 [-1,1]. 它们都只取决于数据的 rank, 因此对数据做单调变换不会改变它们之间的 Spearman 或 Kendall correlation，称这种性质为 scale-free.

2. Pearson correlation 很容易受到数据中异常值 (outliers) 的影响, 但 Spearman's ρ 和 Kendall's τ 几乎不会受影响。讨论使用上述三种指标测量以下数据的相关性会有什么不同：



保险金融领域的研究者很早就发现了 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

$$\lim_{u \to 1-} P\left(X_j > F_j^{-1}(u) \mid X_k > F_k^{-1}(u)\right) = 0. \tag{5}$$

(5)表明这两个随机变量在极端事件下是渐近独立的。如果我们用 X_j 和 X_k 表示两个证券的损失，(5)表明当证券 k 遭受巨大损失时，证券 j 遭受巨大损失的概率几乎为 0. 但是在金融危机中，很多证券的价格都是同时暴跌，因此对一些金融数据使用 Gaussian copula 模型会给人一种错误的安全感。

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

- **t copula**. 给定一个 correlation matrix $R \in \mathbb{R}^{d\times d}$, 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：

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

  其中 $T_\nu()$ 是一元 $t_{(\nu)}$ 分布的 CDF，因为 $Y_j \sim t_{(\nu)}$, $j = 1, \ldots, d$.

Remark

- t copula 使较大的 X_j 和 X_k 具有了尾部相关性 (tail dependence), 如图3所示。但由于 t 分布的对称性，当 X_j 和 X_k 都为很小的负数时也存在相同的相关性。而在金融市场中，两只股票大涨和大跌时的尾部相关性一般是不同的。

作业：二元的 Clayton copula 为

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

其中参数 $\theta > 0$, 其对应的 PDF 为

$$c(u_1, u_2 \mid \theta) = (\theta + 1)(u_1 u_2)^{-(\theta+1)}(u_1^{-\theta} + u_2^{-\theta} - 1)^{-(2\theta+1)/\theta}$$

Clayton copula 具有 lower tail dependence 的特性，即当 U_1, U_2 都很小时，它们的相关性大于给定它们都很大时的相关性。显然，当我们需要一个具有 higher tail dependence 的 copula 时，可以做变量变换 $(\tilde{U}_1, \tilde{U}_2) = (1 - U_1, 1 - U_2)$. 使用二元 Clayton copula ($\theta = 2$) 和 N(0,1) 边际分布产生随机向量，并用散点图考察这些样本的特点。

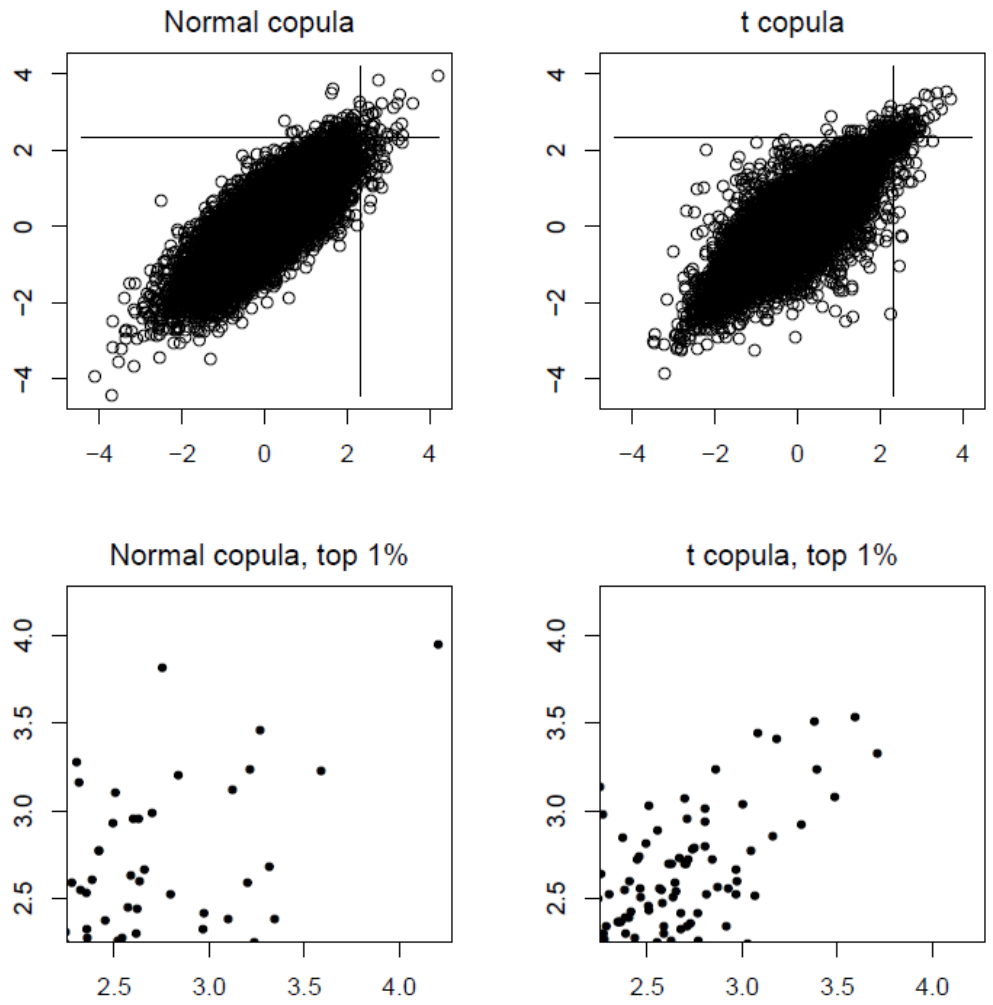


Figure 3: 左上角的图展示了来自二元正态分布的 10000 个样本，$X1$ 和 X_2 的相关系数为 0.8; 右上角的图展示了用 $\nu = 5$, 相关系数也为 0.8 的二元 t 分布对应的 copula 搭配正态边际分布生成的 10000 个样本; 下侧的图是将上侧图中前 1% 的样本放大的效果。Picture source: Art B. Owen.

- 首先注意到 U_1 的边际分布为 $\boldsymbol{U}(0,1)$:

$$P(U_1 \leq u_1) = P(U_1 \leq u_1, U_2 \leq 1) = C(u_1, u_2 = 1 \mid \theta) = u_1.$$

  所以可以抽取 $U_1 \sim \boldsymbol{U}(0,1)$

- 给定 $U_1 = u_1$, 考虑从条件分布抽取 $U_2 \sim C_{U_2|U_1}$. 考察条件分布的 CDF

$$C_{2|1}(u_2 \mid u_1) = P(U_2 \leq u_2 \mid U_1 = u_1) = \int_0^{u_2} c_{2|1}(U_2 = t \mid U_1 = u_1)dt \tag{6}$$

其中条件分布的 PDF 为

$$c_{2|1}(t \mid u_1) = \frac{c(u_1, t \mid \theta)}{\underbrace{c_1(u_1)}_{=1}} = c(u_1, t \mid \theta)$$

则(6)可写为

$$\begin{aligned} C_{2|1}(u_2 \mid u_1) &= (\theta+1)u_1^{-(\theta+1)} \int_0^{u_2} t^{-\theta-1}(u_1^{-\theta} + t^{-\theta} - 1)^{-1-(\theta+1)/\theta} dt \\ &= -\frac{\theta+1}{\theta} u_1^{-(\theta+1)} \int_0^{u_2} (u_1^{-\theta} + t^{-\theta} - 1)^{-(\theta+1)/\theta-1} dt^{-\theta} \\ &= u_1^{-(\theta+1)}(u_1^{-\theta} + u_2^{-\theta} - 1)^{-(\theta+1)/\theta} \end{aligned}$$

另一种更简便的求法为

$$\begin{aligned} P(U_2 \le u_2 \mid U_1 = u_1) &= \lim_{\Delta \to 0+} P(U_2 \le u_2 \mid u1 - \Delta < U_1 \le u_1) \\ &= \lim_{\Delta \to 0} \frac{P(U_2 \le u_2, u1 - \Delta < U_1 \le u_1)/\Delta}{P(u1 - \Delta < U_1 \le u_1)/\Delta} \\ &= \frac{\partial C(u_1, u_2 \mid \theta)/\partial u_1}{\underbrace{c_1(U_1 = u_1)}_{=1}} = \frac{\partial C(u_2, u_1 \mid \theta)}{\partial u_1} \\ &= u_1^{-(\theta+1)}(u_1^{-\theta} + u_2^{-\theta} - 1)^{-(\theta+1)/\theta} \end{aligned}$$

两种方法得到的结果一致。

- $C_{2|1}(\cdot \mid u_1)$ 的逆函数为

$$C_{2|1}^{-1}(w \mid u_1) = \left[(w^{-\theta/(\theta+1)} - 1)u_1^{-\theta} + 1\right]^{-1/\theta}, \ w \in (0,1)$$

因此给定 U_1, 可令 $U_2 = C_{2|1}^{-1}(W \mid U_1)$, 其中 $W \sim \boldsymbol{U}(0,1)$.

- 最后令 $X_1 = \Phi^{-1}(U_1)$, $X_2 = \Phi^{-1}(U_2)$, 则 (X_1, X_2) 即是 Clayton copula 搭配 N(0,1) 边际分布产生的随机向量。

7 球面上的随机点

本节讨论如何从 d 维空间的超球面抽样，以及与此相关的、从球对称或椭球对称分布抽样的问题。我们先介绍如何对超球面上的均匀分布抽样。

定义 d 维空间的单位球为

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

$d=2$ 和 $d=3$ 的 S^{d-1} 分别对应单位圆周和单位球面，此时利用坐标变换很容易实现从均匀分布 $\boldsymbol{U}(S^{d-1})$ 中抽样，且只需使用 $(d-1)$ 个随机变量:

- $d=2$ 时，令 $\boldsymbol{X}=(\cos(2\pi U),\sin(2\pi U))$, $U\sim\boldsymbol{U}(0,1)$.
- $d=3$ 时，

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

  - 这里用到了 Archimedes' hat box theorem(帽盒定理)：如果球面上的点 $(X,Y,Z)\sim\boldsymbol{U}(S^2)$，则 $Z\sim\boldsymbol{U}(-1,1)$.
  - 如果使用球坐标 $Z=\cos(\varphi)$, $\varphi\sim\boldsymbol{U}[0,\pi]$ 产生单位球面上的点，这些点会过于集中在球的上下极点附近，如图4所示。此时我们相当于先在长方形区域 $[0,2\pi]\times[0,\pi]$ 上对 (θ,φ) 均匀取点，然后通过球坐标变换映射到球面上。这导致 $\varphi\approx0$ 或 $\varphi\approx\pi$ 的条状区域被压缩到球的极点附近很小的区域内，如图5的左图所示。

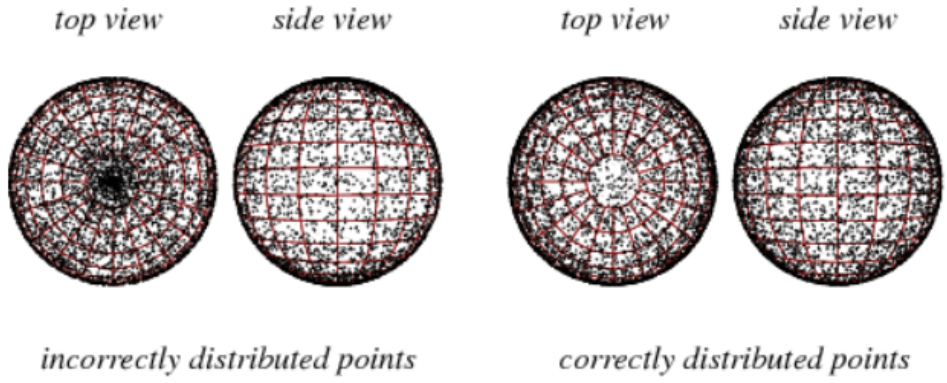


Figure 4: 使用 $\varphi\sim\boldsymbol{U}[0,\pi]$ 产生单位球面上的点 (左)；使用 $Z\sim\boldsymbol{U}(-1,1)$ 产生单位球面上的点 (右)。

  - 如果先在长方形区域 $[0,2\pi]\times[0,1]$ 上对 (θ,Z) 均匀取点，然后映射到球面上，由于在 $Z=1$ 和 $Z=0$ 处变化相同的 Δ 对应极角 φ 不同幅度的变化，因此 $Z\approx1$ 和 $Z\approx0$ 的条状区域映射到球面上的面积可能是相等的，如图5的右图所示。由帽盒定理可知它们确实是严格相等的。

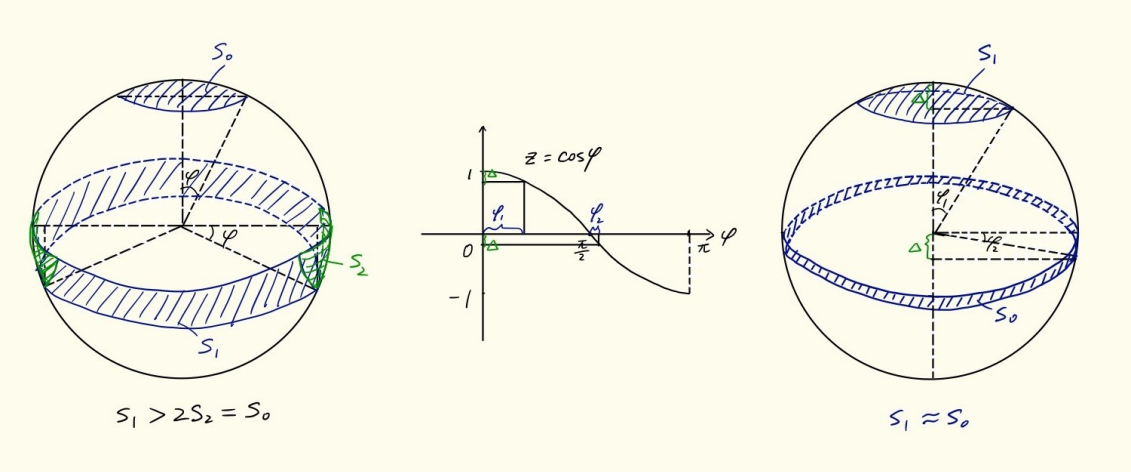


Figure 5: 左: 极角 0 和 $\pi/2$ 的 φ 邻域对应的球面面积; 中: $z = \cos(\varphi)$; 右: z 坐标 1 和 0 的 Δ 邻域对应的球面面积。

– 当 $Z \sim \boldsymbol{U}(-1, 1)$, 对应的极角 φ 的 PDF 为

$$f(\varphi) = \frac{1}{2}\sin(\varphi),\ \varphi \in [0, \pi]$$

显然此时 φ 并不服从均匀分布。

当 $d > 3$ 时，从球面均匀抽样 $\boldsymbol{X} \sim \boldsymbol{U}(S^{d-1})$ 的一种简便做法是令

$$\boldsymbol{X} = \boldsymbol{Z}/\|\boldsymbol{Z}\|,\ \boldsymbol{Z} \sim N_d(0, I_d). \tag{7}$$

原因是 $N_d(0, I_d)$ 的 PDF

$$\phi(\boldsymbol{z}) = (2\pi)^{-d/2}\exp\left(-\|\boldsymbol{z}\|^2/2\right)$$

在等高超球面 (contour) 上是常数。在上一章介绍的 Box-Muller 方法中，为了得到 $N_2(0, I_2)$ 的样本，我们先抽一个 $\sqrt{\chi^2_{(2)}}$ 半径，然后在其对应的圆周上均匀取点。(7)则将这一过程反过来，使用 d 个正态变量产生 d 维超球面上均匀分布的样本。

当我们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变量的半径 $R = \|\boldsymbol{X}\|$.

- **Example**. 如果 $\boldsymbol{X} \sim f(\boldsymbol{x}) \propto \exp(-\|\boldsymbol{x}\|)$，如何得到 $\boldsymbol{X}$ 的样本?

  – 注意到

$$P(r \le \|\boldsymbol{X}\| \le r + dr) \propto \exp(-r)r^{d-1}dr$$

其中 $r^{d-1}dr$ 正比于球面 $S^{d-1}(r)=\{\boldsymbol{x}\in\mathbb{R}^d:\|\boldsymbol{x}\|=r\}$ 和球面 $S^{d-1}(r+dr)$ 之间所夹的体积。因此半径 $R=\|\boldsymbol{X}\|$ 的 PDF 为

$$f_R(r)\propto r^{d-1}\exp(-r)$$

所以 $R\sim\mathrm{Gam}(d,1)$.

- 则对 $\boldsymbol{X}$ 的抽样可以如下进行:

$$\boldsymbol{Z}\sim N_d(0,I_d),\ R\sim\mathrm{Gam}(d),\ \text{then } \boldsymbol{X}=R\frac{\boldsymbol{Z}}{\|\boldsymbol{Z}\|}.$$

- *练习*. 如果 $\boldsymbol{X}\sim f(\boldsymbol{x})\propto\|\boldsymbol{x}\|^k\,\mathbf{1}\{\|\boldsymbol{x}\|\le 1\}$, $k>-d$. 如何得到 $\boldsymbol{X}$ 的样本?

  - 此时

$$P(r\le\|\boldsymbol{X}\|\le r+dr)\propto r^k\mathbf{1}\{0<r<1\}\,r^{d-1}dr$$

则半径 $R=\|\boldsymbol{X}\|$ 的 PDF 为

$$f_R(r)\propto r^{k+d-1}\mathbf{1}\{0<r<1\}$$

对应 $\mathrm{Beta}(k+d,1)$ 分布。

  - 因此对 $\boldsymbol{X}$ 的抽样可以如下进行:

$$\boldsymbol{Z}\sim N_d(0,I_d),\ R\sim\mathrm{Beta}(k+d,1),\ \text{then } \boldsymbol{X}=R\frac{\boldsymbol{Z}}{\|\boldsymbol{Z}\|}.$$

  - 本例中，$k=0$ 对应单位球 $B_d=\{\boldsymbol{x}\in\mathbb{R}^d:\|\boldsymbol{x}\|\le 1\}$ 内的均匀分布；$-d<k<0$ 对应的分布在球心 $\boldsymbol{x}=\mathbf{0}$ 处的概率密度无限大。

在这两个例子中，$f(\boldsymbol{x})\propto h(\|\boldsymbol{x}\|)$ 且我们能识别 PDF $\propto r^{d-1}h(r)$ 的分布。如果不能识别半径 R 的分布，可以尝试 A-R 方法。此时我们需要找到一个在 $[0,\infty)$ 上较容易抽样的 PDF $g(r)\propto\tilde{g}(r)$, 且能找到一个常数 c 使得

$$r^{d-1}h(r)\le c\tilde{g}(r).$$

对球对称分布做线性变换可以得到椭球对称分布，我们从多元正态和多元 t 分布抽样时使用过该方法。如果 $\boldsymbol{X}\sim\boldsymbol{U}(B_d)$，令 $\boldsymbol{\mu}\in\mathbb{R}^d$, $C\in\mathbb{R}^{d\times d}$ 是一个可逆矩阵，则 $\boldsymbol{Y}=\boldsymbol{\mu}+C\boldsymbol{X}$ 均匀地分布在椭球 $\mathcal{E}(\boldsymbol{\mu},\Sigma)$ 内,

$$\mathcal{E}(\boldsymbol{\mu},\Sigma)=\left\{\boldsymbol{y}\in\mathbb{R}^d\mid(\boldsymbol{y}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\le 1\right\}$$

其中 $\Sigma=CC^\top$. 注意如果 $\boldsymbol{X}\sim\boldsymbol{U}(S^{d-1})$, $\boldsymbol{\mu}+C\boldsymbol{X}$ 不一定在椭球 $\mathcal{E}(\boldsymbol{\mu},\Sigma)$ 表面上均匀分布。

除了均匀分布，我们也需要一些球面上的非均匀分布，因为有些现象在特定方向上发生地更频繁。一个常用的球面 S^{d-1} 上的非均匀分布是 **von Mises-Fisher** 分布，它的 PDF 为

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

其中参数 $\kappa \geq 0$, 向量 $\boldsymbol{\mu} \in S^{d-1}$. 当 $\kappa > 0$ 时，von Mises-Fisher 分布在点或方向 $\boldsymbol{\mu}$ 处的概率密度最大；κ 越大，von Mises-Fisher 分布越集中在 $\boldsymbol{\mu}$ 附近.

从 von Mises-Fisher 分布抽样的关键在于知道如何对随机变量 $W = \boldsymbol{\mu}^\top \boldsymbol{X}$ 进行抽样。综合Ulrich (1984), Wood (1994) 和 Hoff (2009), 对 von Mises-Fisher 分布抽样的一种算法总结如下：

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

其中对 W 可以使用 A-R 方法抽样，比如选取经过变换的 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，可通过 Gram-Schmidt 算法得到；$B\boldsymbol{V}$ 是在与 $\boldsymbol{\mu}$ 垂直的方向上均匀分布的单位向量; $\boldsymbol{X}$ 由两部分构成，与 $\boldsymbol{\mu}$ 平行的部分 $W\boldsymbol{\mu}$ 和与 $\boldsymbol{\mu}$ 垂直的部分 $\sqrt{1-W^2}B\boldsymbol{V}$. R Package `rstiefel` 实现了上述抽样算法。

8 随机矩阵

很多实际问题需要生成随机矩阵 $\mathcal{X} \in \mathbb{R}^{n\times d}$. 有时可以将 $\mathcal{X}$ 看成 $\mathbb{R}^d$ 空间上的 n 个随机向量，或者将 $\mathcal{X}$ 中的元素重新排列成一个 $nd \times 1$ 的向量，这样就可以用随机向量的抽样方法产生 $\mathcal{X}$. 比如，当 $\mathcal{X}$ 的各列 (行) 向量都是独立的，就很适合对 $\mathcal{X}$ 逐列 (行) 独立地生成随机向量。但是在一些问题中，$\mathcal{X}$ 既不是一些独立向量的集合，又没有 $\mathbb{R}^{nd}$ 上随机向量的相关结构复杂。这种情况下，直接生成一个随机矩阵可能要比产生一个高维的随机向量容易。

8.1 矩阵正态分布 (Matrix normal distribution)

矩阵正态分布常用于描述行和列都有相关性的数据矩阵。比如一些基因数据中，数据的列可能对应一些有相关性的基因，数据的行对应的是受试者；受实验室条件的限制，不同受试者的测量值不完全是独立的。又比如记录不同商品在不同时期价格的面板数据，在商品之间和不同时期的价格之间都可能存在相关性。

- $\mathbb{R}^{n\times d}$ 上的矩阵正态分布 $N_{n\times d}(M, \Gamma, \Sigma)$ 有三个参数矩阵 $M \in \mathbb{R}^{n\times d}$, $\Gamma \in \mathbb{R}^{n\times n}$, $\Sigma \in \mathbb{R}^{d\times d}$.

- Γ 和 Σ 都是半正定的对称矩阵。
- 如果 $\mathcal{X} \sim N_{n\times d}(M,\Gamma,\Sigma)$, 则 $\mathcal{X}$ 的元素 $\mathcal{X}_{ij}$ 满足

$$E(\mathcal{X}_{ij}) = M_{ij}, \quad Cov(\mathcal{X}_{ij},\mathcal{X}_{kl}) = \Gamma_{ik}\Sigma_{jl}.$$

- 注意到对任意常数 c, 如果用 $c\Gamma$ 替换 Γ, 同时用 $c^{-1}\Sigma$ 替换 Σ, 并不会改变矩阵正态分布的形状，因此需要再增加一个条件保证参数的可识别性 (identification). 比如可以令 $\mathrm{tr}(\Gamma) = m$, 此时参数是唯一确定的，即模型是可识别的。

Remarks

1. 如果将 $\mathcal{X}$ 按如下方式转化成一个 $nd\times 1$ 的向量

$$vec(\mathcal{X}) = vec(\begin{pmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_d \\ | & & | \end{pmatrix}) = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_d \end{pmatrix} \in \mathbb{R}^{nd}$$

则 $vec(\mathcal{X}) \sim N_{nd}\left(vec(M), \Sigma\otimes\Gamma\right)$, 其中 $\otimes$ 代表 Kronecker product, 定义如下

$$A\otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1q}B \\ \vdots & & \vdots \\ a_{p1}B & \cdots & a_{pq}B \end{pmatrix}, \quad A\in\mathbb{R}^{p\times q},\ B\in\mathbb{R}^{m\times n}$$

2. 对一般的 $\mathbb{R}^{nd}$ 上的正态分布，我们需要 $nd(nd+1)/2$ 个参数来描述协方差矩阵，但是矩阵正态分布只需要 $n(n+1)/2 + d(d+1)/2$ 个参数来描述协方差矩阵。当 n 和 d 很大时，矩阵正态分布可以极大地减少参数个数。

3. 如果 $\mathcal{X} \sim N_{n\times d}(M,\Gamma,\Sigma)$, A 和 B 是非随机矩阵，在满足维数匹配的情况下

$$A\mathcal{X}B^\top \sim N_{n\times d}(AMB^\top, A\Gamma A^\top, B\Sigma B^\top)$$

4. 对 $N_{n\times d}(M,\Gamma,\Sigma)$ 的抽样很简单。如果我们能找到矩阵 A 和 B 满足 $\Gamma = AA^\top$, $\Sigma = BB^\top$, 就可以如下从 $N_{n\times d}(\Theta,\Gamma,\Sigma)$ 分布抽样

$$\mathcal{Z} \sim N_{n\times d}(\mathbf{0}, I_n, I_d), \text{ then } \mathcal{X} = M + A\mathcal{Z}B^\top$$

其中对 $N_{n\times d}(\mathbf{0}, I_n, I_d)$ 抽样相当于独立地从 $N(0,1)$ 中抽取 nd 个样本。

如果我们对 $\mathbb{R}^{n\times d}$ 上的正态随机矩阵 $\mathcal{X}$ 做 SVD 分解

$$\mathcal{X}=\mathcal{U}\mathcal{D}\mathcal{V}^\top$$

假设 $n>d$ 且 rank($\mathcal{X}$)=d, 则 D 是 $d\times d$ 的对角矩阵且对角线元素为正数;

$$\mathcal{U}\in\mathbb{V}_{d,n}=\left\{Y\in\mathbb{R}^{n\times d}:Y^\top Y=I_d\right\},\ \mathbb{R}^n\text{上的 } d \text{ 维 Stiefel manifold}$$
$$\mathcal{V}\in\mathbb{O}_d=\left\{Q\in\mathbb{R}^{d\times d}:Q^\top Q=QQ^\top=I_d\right\},\ d \text{ 阶正交矩阵的集合。}$$

下一小节将讨论 $\mathcal{U}$ 和 $\mathcal{V}$ 服从的分布。

8.2 随机正交矩阵

给若干向量乘以同一正交矩阵，相当于对坐标系做一个旋转，并不改变向量的长度或向量之间的夹角。我们将 d 阶正交矩阵组成的空间记为

$$\mathbb{O}_d=\left\{Q\in\mathbb{R}^{d\times d}:Q^\top Q=QQ^\top=I_d\right\}$$

用 $\boldsymbol{U}(\mathbb{O}_d)$ 表示 $\mathbb{O}_d$ 上的均匀分布，该分布具有以下性质：如果 $\mathcal{Q}\sim\boldsymbol{U}(\mathbb{O}_d)$, $\tilde{Q}\in\mathbb{O}_d$, 则

$$\tilde{Q}\mathcal{Q}\sim\boldsymbol{U}(\mathbb{O}_d)\text{ 且 }\mathcal{Q}\tilde{Q}\sim\boldsymbol{U}(\mathbb{O}_d).$$

这是因为对 $\mathcal{Q}$ 左乘或右乘一个正交矩阵 $\tilde{Q}$ 只是对 $\mathcal{Q}$ 中的向量整体做了一次旋转，并不改变它们的长度或夹角。下面讨论如何从 $\boldsymbol{U}(\mathbb{O}_d)$ 分布抽样。

- 首先研究 $\mathcal{Q}\sim\boldsymbol{U}(\mathbb{O}_d)$ 的第一列 $\mathcal{Q}_{\cdot 1}$ 的边际分布。显然 $\mathcal{Q}_{\cdot 1}$ 是在 $\mathbb{R}^d$ 上均匀分布的一个单位向量，即 $\mathcal{Q}_{\cdot 1}\sim\boldsymbol{U}(S^{d-1})$, 因此可令

$$\mathcal{Q}_{\cdot 1}=\frac{\boldsymbol{Z}_1}{\|\boldsymbol{Z}_1\|},\ \boldsymbol{Z}_1\sim N_d(\mathbf{0},I_d)$$

- 给定 $\mathcal{Q}_{\cdot 1}$, $\mathcal{Q}$ 的第二列 $\mathcal{Q}_{\cdot 2}$ 在与 $\mathcal{Q}_{\cdot 1}$ 垂直的单位圆 (球) 上均匀分布，可以如下产生：

$$\boldsymbol{Z}_2\sim N_d(\mathbf{0},I_d),\ \text{令 }\tilde{\boldsymbol{Z}}_2=\boldsymbol{Z}_2-(\boldsymbol{Z}_2^\top\mathcal{Q}_{\cdot 1})\mathcal{Q}_{\cdot 1}$$

  此时向量 $\tilde{\boldsymbol{Z}}_2$ 与 $\mathcal{Q}_{\cdot 1}$ 垂直，再进一步单位化即可产生 $\mathcal{Q}_{\cdot 2}=\tilde{\boldsymbol{Z}}_2/\left\|\tilde{\boldsymbol{Z}}_2\right\|$.

- 类似地，我们可以继续从 $N_d(\mathbf{0},I_d)$ 抽样，对其进行 Gram-Schmidt 正交化及单位化, 依次产生 $\mathcal{Q}$ 的后面几列向量。

- 上述过程等价于直接从 $N(0,1)$ 多次独立抽样组成 $\mathbb{R}^{d\times d}$ 上的随机矩阵 $\mathcal{Z}$, 再对 $\mathcal{Z}$ 进行 Gram-Schmidt 正交化。

除了 $\boldsymbol{U}(\mathbb{O}_d)$, $\mathbb{O}_d$ 上的一个很重要的非均匀分布是 **Bingham** 分布，在空间统计学和形状分析中有广泛应用。$\mathbb{O}_d$ 上的 Bingham(L,Ψ) 分布有两个参数矩阵，L 是 $d\times d$ 对角矩阵，Ψ 是 $d\times d$ 对称矩阵；为保证参数可识别，一般要求 L 的对角元素递减排列。$\mathcal{Q}\sim$ Bingham(L,Ψ) 的 PDF 为

$$f(Q)\propto \exp\left\{\mathrm{tr}\left(LQ^\top\Psi Q\right)\right\} \tag{8}$$

- 证明 $E(\mathcal{Q})=\mathbf{0}_{d\times d}$.
  - 从(8)可知 Bingham 分布是椭球对称分布，且椭球的中心是 $\mathbf{0}_{d\times d}$.
- 证明 Bingham 分布具有 **antipodal symmetry**, 即如果 $\mathcal{Q}\sim$ Bingham(L,Ψ), S 是 $d\times d$ 的对角矩阵且对角线元素为 1 或-1, 则

$$\mathcal{Q}S\overset{d}{=}\mathcal{Q}.$$

Proof. 此处我们需要用到矩阵 trace 的一个性质：$\mathrm{tr}(AB)=\mathrm{tr}(BA)$.

$$\begin{aligned} f(QS)&\propto \exp\left\{\mathrm{tr}\left(LS^\top Q^\top\Psi QS\right)\right\}\propto \exp\left\{\mathrm{tr}\left(SLS^\top Q^\top\Psi Q\right)\right\}\\ &\propto \exp\left\{\mathrm{tr}\left(LQ^\top\Psi Q\right)\right\}. \end{aligned}$$

□

由于 Bingham 分布的期望不属于 $\mathbb{O}_d$, 因此人们更关心 Bingham 分布的 mode, 即概率密度最大的点 (矩阵)。此时我们只需要找到使 $\mathrm{tr}\left(LQ^\top\Psi Q\right)$ 最大的 Q. 对称矩阵 Ψ 存在特征值分解 $\Psi=B\Lambda B^\top$, 则 $\mathrm{tr}\left(LQ^\top\Psi Q\right)=\mathrm{tr}\left(LQ^\top B\Lambda B^\top Q\right)$. 为计算 trace, 将 $d\times d$ 正交矩阵 Q 和 B 写成列向量形式，令

$$M=\begin{pmatrix}\boldsymbol{q}_1^\top\\ \vdots\\ \boldsymbol{q}_d^\top\end{pmatrix}\begin{pmatrix}\boldsymbol{b}_1 & \cdots & \boldsymbol{b}_d\end{pmatrix}\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_d\end{pmatrix}\begin{pmatrix}\boldsymbol{b}_1^\top\\ \vdots\\ \boldsymbol{b}_d^\top\end{pmatrix}\begin{pmatrix}\boldsymbol{q}_1 & \cdots & \boldsymbol{q}_d\end{pmatrix}$$

M 的对角线元素为

$$\begin{aligned} M_{jj}&=\boldsymbol{q}_j^\top\begin{pmatrix}\lambda_1\boldsymbol{b}_1 & \cdots & \lambda_d\boldsymbol{b}_d\end{pmatrix}\begin{pmatrix}\boldsymbol{b}_1^\top\boldsymbol{q}_j\\ \vdots\\ \boldsymbol{b}_d^\top\boldsymbol{q}_j\end{pmatrix}\\ &=\sum_{k=1}^d\lambda_k\left(\boldsymbol{b}_k^\top\boldsymbol{q}_j\right)^2,\quad j=1,\ldots,d. \end{aligned}$$

则

$$\mathrm{tr}\left(LQ^\top \Psi Q\right) = \mathrm{tr}\left(LQ^\top B\Lambda B^\top Q\right) = \sum_{j=1}^{d}\sum_{k=1}^{d} l_j\lambda_k\left(\boldsymbol{b}_k^\top \boldsymbol{q}_j\right)^2.$$

因此当 Q 的 d 个列向量与 B 的 d 个列向量越接近，Q 的概率密度越大。由于 L 和 Λ 中的对角线元素递减排列，为使 trace 最大，应该让 $\boldsymbol{q}_1$ 与 $\boldsymbol{b}_1$ 一致，$\boldsymbol{q}_2$ 与 $\boldsymbol{b}_2$ 一致，依此类推。上述分析证明了如下定理：

Theorem 2. *$\mathbb{O}_d$ 上的 Bingham(L,Ψ) 的 mode 为 B 和 $\{BS : S = diag(s_1,\ldots,s_d), s_j \in \{-1,1\}\}$，其中 B 为 Ψ 的 d 个特征向量组成的矩阵。*

Bingham(L,Ψ) 在它的 mode 附近的集中度与 L 中对角线元素之间的差距正相关，也与 Λ 中对角线元素的差距正相关，这可以从上述证明过程推断出来。比如

$$l_1 \approx l_2 \;\Rightarrow\; \boldsymbol{q}_1 \overset{d}{\approx} \boldsymbol{q}_2$$

如果 $\lambda_1 \gg \lambda_2$, 则 $\boldsymbol{q}_1$ 和 $\boldsymbol{q}_2$ 有较大概率分布在 $\boldsymbol{b}_1$ 和 $\boldsymbol{b}_2$ 所在的子空间平面。

Hoff (2009) 提出了一种基于 Gibbs sampler 对 Bingham 分布抽样的方法, R Package `rstiefel` 包含该抽样函数。

每个正交矩阵对应一个旋转变化。有时我们需要将 $\mathbb{R}^n$ 上的向量投影到 $\mathbb{R}^k$ 空间 $(k<n)$, 此时需要一个 $n\times k$ 的投影矩阵 P, P 属于如下的 **Stiefel manifold**:

$$\mathbb{V}_{k,n} = \left\{P \in \mathbb{R}^{n\times k} : P^\top P = I_k\right\}.$$

如果想得到 $\boldsymbol{U}(\mathbb{V}_{k,n})$ 的样本，可以先抽取 $\mathcal{Q} \sim \boldsymbol{U}(\mathbb{O}_n)$，然后只保留 $\mathcal{Q}$ 的前 k 列，显然在抽取 $\mathcal{Q}$ 时没有必要生成全部 n 列。另一种方法是使用以下定理。

Theorem 3. *如果 $\mathcal{X} \sim N_{n\times d}(\mathbf{0}, I_n, \Sigma)$，对 $\mathcal{X}$ 做 SVD 分解 $\mathcal{X} = \mathcal{U}\mathcal{D}\mathcal{V}^\top$，则*

- *$\mathcal{U} \sim \boldsymbol{U}(\mathbb{V}_{d,n})$, 且 $\mathcal{U}$ 与 $(\mathcal{D},\mathcal{V})$ 独立;*
- *$\mathcal{V} \mid \mathcal{D} \sim$ Bingham$(\mathcal{D}^2, -\Sigma^{-1}/2)$;*
- *$\mathcal{D}^2$ 的对角线元素与 Wishart 分布 $W_d(\Sigma,n)$ 的随机矩阵的特征值同分布。*

Proof. $\mathcal{X} \sim N_{n\times d}(\mathbf{0}, I_n, \Sigma)$ 意味着 $\mathcal{X}$ 的每一行都独立地服从 $N_d(\mathbf{0},\Sigma)$. 因此 $\mathcal{X}$ 的 PDF 可写为

$$f(X \mid \Sigma) \propto \prod_{i=1}^{n}\exp\left\{-\boldsymbol{x}_i^\top\Sigma^{-1}\boldsymbol{x}_i/2\right\} \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\boldsymbol{x}_i^\top\Sigma^{-1}\boldsymbol{x}_i\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\mathrm{tr}\left(X\Sigma^{-1}X^\top\right)\right\} \propto \exp\left\{-\frac{1}{2}\mathrm{tr}\left(X^\top X\Sigma^{-1}\right)\right\}$$

代入 SVD 分解 $X = UDV^\top$

$$\begin{aligned} f(U,D,V \mid \Sigma) &\propto \exp\left\{-\frac{1}{2}\mathrm{tr}\left(VD\underbrace{U^\top U}_{I_d}DV^\top\Sigma^{-1}\right)\right\}\cdot|\boldsymbol{J}| \\ &\propto \exp\left\{-\frac{1}{2}\mathrm{tr}\left(VD^2V^\top\Sigma^{-1}\right)\right\}\cdot|\boldsymbol{J}| \\ &\propto \exp\left\{-\frac{1}{2}\mathrm{tr}\left(D^2V^\top\Sigma^{-1}V\right)\right\}\cdot|\boldsymbol{J}| \end{aligned}$$

其中 $\boldsymbol{J}$ 是一个 Jacobian matrix, $\boldsymbol{J} = dX/d(U,D,V)$. 可以证明 $|\boldsymbol{J}|$ 不依赖 U 或 V, 只与 D 有关。证明的难点在于如何确定 Jacobian matrix $\boldsymbol{J}$，因为 SVD 的分解形式不唯一 (给 left singular vector 乘以-1，同时给 right singular vector 乘以-1，不影响 SVD 分解)，此处略去证明细节。

因此 $(\mathcal{U},\mathcal{D},\mathcal{V})$ 的联合 PDF 不依赖 U, 即关于 U 是常数，可以推断 $\mathcal{U}$ 在 $\mathbb{V}_{d,n}$ 上均匀分布，且与 $(\mathcal{D},\mathcal{V})$ 独立。

由于 $|\boldsymbol{J}|$ 与 V 无关，所以

$$\begin{aligned} f(V \mid D) &\propto \exp\left\{-\frac{1}{2}\mathrm{tr}\left(D^2V^\top\Sigma^{-1}V\right)\right\} \\ &\propto \exp\left\{\mathrm{tr}\left(D^2V^\top\left(-\Sigma^{-1}/2\right)V\right)\right\} \end{aligned}$$

因此给定 $\mathcal{D} = D$, $\mathcal{V}$ 的条件分布为 $\mathrm{Bingham}(D^2, -\Sigma^{-1}/2)$.

最后可得 $\mathcal{D}$ 的边际分布的 PDF 具有以下形式

$$f(D) \propto |\boldsymbol{J}| \int_{\mathbb{O}_d} \exp\left\{\mathrm{tr}\left(D^2V^\top\left(-\Sigma^{-1}/2\right)V\right)\right\}\mu(dV)$$

其中 μ 是 $\mathbb{O}_d$ 上均匀分布的测度。

□

Example. 从 $\mathcal{X} \sim N_{2\times 2}(\mathbf{0}, I_2, \Sigma)$ 分布随机抽取 100 个样本，其中

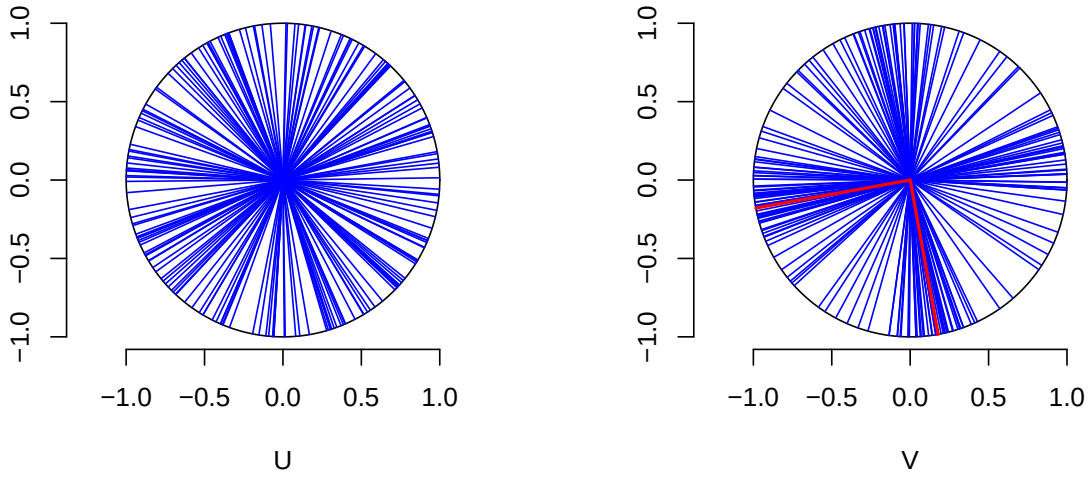
$$\Sigma = \begin{pmatrix} 9 & 1.5 \\ 1.5 & 1 \end{pmatrix}.$$

对 $\mathcal{X}$ 做 SVD 分解 $\mathcal{X} = \mathcal{U}\mathcal{D}\mathcal{V}^\top$, 在单位元上展示 $\mathcal{U}$ 和 $\mathcal{V}$ 的列向量分布。

```
library(plotrix)
```

```
# prepare plot ----------------------
par(mfrow = c(1,2))
xlabt = c('U','V')
for(i in 1:2){
plot(0, 0, asp=1, type = "n", xlim=c(-1,1), ylim=c(-1,1), xlab = xlabt[i],
    ylab="", bty="n")
draw.circle(0, 0, 1, nv = 1000, border = NULL, col = NA, lty = 1, lwd = 1)
}

# specify Sigma and other paras ------------------------------------
d = 2
Sigma = matrix(c(9,1.5,1.5,1),d,d)
ED_sig = eigen(Sigma, symmetric = TRUE)
ED_sig$values # 9.2720019 0.7279981 (eigengap is large)

B = ED_sig$vectors %*% diag(sqrt(ED_sig$values)) # B*t(B) = Sigma

n = 2 # for visualization convenience

nrep = 100 # number of experiments

for(j in 1:nrep){
Z = matrix(rnorm(4),n,d)
X = Z %*% t(B) # X~N(0,I,Sigma)

SVD_X = svd(X)
U = SVD_X$u # R seems to let U(1,1) always be negative
V = SVD_X$v

# randomly change sign of U[,1]
if(runif(1) <0.5){
U[,1] = -U[,1]
V[,1] = -V[,1]
}

# randomly change sign of U[,2]
```

```
if(runif(1) <0.5){
U[,2] = -U[,2]
V[,2] = -V[,2]
}

par(mfg=c(1,1))
lines(rbind(c(0,0),t(U[,1])), col="blue")
lines(rbind(c(0,0),t(U[,2])), col="blue")

par(mfg=c(1,2))
lines(rbind(c(0,0),t(V[,1])), col="blue")
lines(rbind(c(0,0),t(V[,2])), col="blue")
}

Q = ED_sig$vectors
par(mfg=c(1,2))
lines(rbind(c(0,0),t(Q[,1])), col="red",lwd=2)
lines(rbind(c(0,0),t(Q[,2])), col="red",lwd=2)
```



可以看到 $\mathcal{U}$ 服从 $\mathbb{V}_{2,2}$, 也即 $\mathbb{O}_2$ 上的均匀分布，$\mathcal{V}$ 更集中在 Σ 的特征向量 (红色) 所在的轴附近。

8.3 Wishart 分布

Wishart 分布是 Bayesian 方法和多元分析中的一个重要分布。如果 $\mathbb{R}^d$ 上的随机向量

$$\boldsymbol{X}_i \overset{iid}{\sim} N_d(\mathbf{0}, \Sigma),\ i = 1, \ldots, n \tag{9}$$

且 $n \geq d$, 则随机矩阵

$$\mathcal{W} = \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top \in \mathbb{R}^{d \times d}$$

服从 **Wishart** 分布 $W_d(\Sigma, n)$.

- Wishart 分布 $W_d(\Sigma, n)$ 有两个参数: $d \times d$ 的对称正定矩阵 Σ 和自由度 n.
- Wishart 分布 $W_d(\Sigma, n)$ 的 support 是 $\mathbb{R}^{d \times d}$ 上对称正定矩阵的集合。
- $d = 1$ 时的 Wishart 分布是 $\sigma^2 \chi^2_{(n)}$.
- 对自由度为正整数的 Wishart 分布，(9)揭示了 Wishart 分布的抽样方法。但正如 χ^2 分布可以有非整数的自由度，Wishart 分布的自由度 ν 可以是满足 $\nu > d - 1$ 的任意正数。$\mathcal{W} \sim W_d(\Sigma, \nu)$ 的 PDF 为

  $$f(W) \propto |W|^{(\nu - d - 1)/2} \exp\left\{ -\frac{1}{2} \mathrm{tr}(\Sigma^{-1} W) \right\},\ W \in \mathbb{R}^{d \times d} \text{ 且是对称正定矩阵。}$$

  - 如果我们把 $|W|$ 看作矩阵广义上的 “绝对值”, 把 $\mathrm{tr}((AB))$ 看作是矩阵的 “点积”, 会发现 Wishart 分布的 PDF 跟 Gamma 分布的 PDF 有些相似。这个结果并不意外，因为 χ^2 变量可以通过一些一元正态变量的平方和生成，而 χ^2 分布本身是一个 Gamma 分布。
- 如果 $\mathcal{W} \sim W_d(\Sigma, \nu)$, $E(\mathcal{W}) = \nu\Sigma$.
- 从(9)可得，如果 $\mathcal{W} \sim W_d(\Sigma, \nu)$, 对于任意矩阵 $C \in \mathbb{R}^{k \times d}$ $(k \leq d)$,

  $$C\mathcal{W}C^\top \sim W_k(C\Sigma C^\top, \nu).$$

  注意 $k > d$ 时, 矩阵 $C\Sigma C^\top$ 不可逆。因此对任意 Wishart 分布 $W_d(\Sigma, \nu)$ 抽样只需先抽取 $\mathcal{W} \sim W_d(I, \nu)$，然后找到矩阵 C 使得 $CC^\top = \Sigma$, 则 $C\mathcal{W}C^\top \sim W_d(\Sigma, \nu)$.
- 对任意的 $\nu > d - 1$, 可采用如下的 Bartlett 分解 (Bartlett, 1933) 对 $W_d(I, \nu)$ 抽样: 如果 L 是 $\mathbb{R}^{d \times d}$ 上的下三角矩阵，且各元素独立服从分布

  $$L_{ij} \sim \begin{cases} N(0,1), & i > j \\ \sqrt{\chi^2_{(\nu - i + 1)}} & i = j \\ 0 & i < j \end{cases}$$

则 $LL^\top \sim W_d(I,\nu)$. 对角线元素的抽样可使用 $\chi^2_{(\nu-i+1)} = 2\mathrm{Gam}\left((\nu-i+1)/2\right)$.

- 实践中常用的模型是

$$\boldsymbol{X}_i \overset{iid}{\sim} N_d(\boldsymbol{\mu},\Sigma),\ i=1,\ldots,n$$

则有

$$\mathcal{W} = \sum_{i=1}^{n}(\boldsymbol{X}_i-\bar{\boldsymbol{X}})(\boldsymbol{X}_i^\top-\bar{\boldsymbol{X}})^\top \sim W_d(\Sigma,n-1)$$

其中 $\bar{\boldsymbol{X}} = \sum_{i=1}^n \boldsymbol{X}_i/n$. 减去均值使 Wishart 分布的自由度降为 $n-1$, 因此对 Σ 的一个无偏 (unbiased) 估计量为

$$\hat{\Sigma} = \frac{W}{n-1}.$$

  - $d=1$ 时，上述结果退化为

$$s^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 \sim \chi^2_{(n-1)},\quad x_i \overset{iid}{\sim} N(\mu,\sigma^2)$$

    得到我们非常熟悉的对 σ^2 的无偏估计量 $\hat{\sigma}^2 = s^2/(n-1)$ (样本方差).

以下定理告诉我们，上述估计量 $\hat{\Sigma}$ 不仅无偏而且是极大似然估计量 (MLE).

Theorem 4. *如果 W 是 Wishart 分布 $W_d(\Sigma,\nu)$ 的一个样本，则 Σ 的 MLE 为 $\hat{\Sigma} = W/\nu$.*

Proof. $W_d(\Sigma,\nu)$ 完整的 PDF 为

$$f(W) = \frac{|W|^{(\nu-d-1)/2}}{2^{\nu d/2}\Gamma_d(\nu/2)|\Sigma|^{\nu/2}}\exp\left\{-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}W)\right\}$$

其中 $\Gamma_d(\nu/2)$ 是与 ν, d 有关的二元 gamma 函数。则 Σ 的 log likelihood 为

$$l(\Sigma \mid W) = -\frac{\nu}{2}\log(|\Sigma|) - \frac{1}{2}\mathrm{tr}(\Sigma^{-1}W) + \underbrace{\cdots}_{\text{与}\Sigma\text{无关}}$$

为计算方便，我们将最大化 $l(\Sigma \mid W)$ 转化为最小化以下函数：

$$-2l(\Sigma \mid W) = \nu\log(|\Sigma|) + \mathrm{tr}(\Sigma^{-1}W). \tag{10}$$

Michael Perlman 为最小化(10)提供了以下非常简洁的解法，不需要使用矩阵微积分的知识。

找到矩阵 B 使得 $BB^\top = W$. 引入一个新的矩阵

$$\Psi = B^\top\Sigma^{-1}B$$

则 $\Psi^{-1} = B^{-1}\Sigma(B^\top)^{-1}$, $\Sigma = B\Psi^{-1}B^\top$.

即 Σ 和 Ψ 之间存在一一对应的关系，因此可以将目标函数(10)重新定义为 Ψ 的函数:

$$\min_{\Sigma} \ \nu \log(|\Sigma|) + \mathrm{tr}(\Sigma^{-1}W) \ = \ \min_{\Psi} \ \nu \log(|\Sigma_\Psi|) + \mathrm{tr}(\Sigma_\Psi^{-1}W)$$

其中 $\Sigma_\Psi = B\Psi^{-1}B^\top$. 进一步

$$\begin{aligned}
\nu \log(|\Sigma_\Psi|) + \mathrm{tr}(\Sigma_\Psi^{-1}W) &= \nu \log(|B\Psi^{-1}B^\top|) + \mathrm{tr}\left((B^\top)^{-1}\Psi B^{-1}W\right) \\
&= \nu \log(|B| \cdot |\Psi|^{-1} \cdot |B^\top|) + \mathrm{tr}\left(\Psi B^{-1}BB^\top(B^\top)^{-1}\right) \\
&= \nu \log(|W|) - \nu \log(|\Psi|) + \mathrm{tr}(\Psi). \qquad (11)
\end{aligned}$$

由于 Ψ 也是对称矩阵，存在特征值分解 $\Psi = Q\Omega Q^\top$, 其中 $\Omega = \mathrm{diag}(\omega_1, \ldots, \omega_d)$. 则有

$$|\Psi| = |Q\Omega Q^\top| = \prod_{j=1}^{d} \omega_j$$

$$\mathrm{tr}(\Psi) = \mathrm{tr}(Q\Omega Q^\top) = \sum_{j=1}^{d} \omega_j.$$

此时目标函数(11)可写为

$$\begin{aligned}
\nu \log(|W|) - \nu \log(|\Psi|) + \mathrm{tr}(\Psi) &= \nu \log(|W|) - \nu \sum_{j=1}^{d} \log(\omega_j) + \sum_{j=1}^{d} \omega_j \\
&= \nu \log(|W|) + \sum_{j=1}^{d} (\omega_j - \nu \log(\omega_j))
\end{aligned}$$

注意到每个一元函数 $\omega_j - \nu \log(\omega_j)$ 是 ω_j 的严格凸函数，存在唯一的最小值点 $\hat{\omega}_j = \nu$, $j = 1, \ldots, d$. 因此目标函数(11)的最小值点为

$$\hat{\Psi} = Q(\nu I)Q^\top = \nu I$$

则 Σ 的 MLE 为

$$\hat{\Sigma} = B\hat{\Psi}^{-1}B^\top = W/\nu.$$

□

8.3.1 正态随机矩阵的 polar decomposition

如果 $\mathcal{X} \sim N_{n\times d}(\mathbf{0}, I_n, \Sigma)$, $n \geq d$, 即 $\mathcal{X}$ 的每一行 $\boldsymbol{X}_i \overset{iid}{\sim} N_d(\mathbf{0}, \Sigma)$, $i = 1, \ldots, n$. 此时

$$\mathcal{S} = \mathcal{X}^\top\mathcal{X} = \sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i^\top \sim W_d(\Sigma, n).$$

总可以对 $\mathcal{X}$ 做 SVD 分解 $\mathcal{X}=\mathcal{U}\mathcal{D}\mathcal{V}^\top$, 则有

$$\mathcal{S}=\mathcal{X}^\top\mathcal{X}=\mathcal{V}\mathcal{D}\underbrace{\mathcal{U}^\top\mathcal{U}}_{I_d}\mathcal{D}\mathcal{V}^\top=\mathcal{V}\mathcal{D}^2\mathcal{V}^\top$$

我们已经证明 Σ 的 MLE 为 $\hat{\Sigma}=\mathcal{S}/n$, 可以看到 $\mathcal{X}$ 的行向量的协方差矩阵 Σ 主要与 $\mathcal{D}$ 和 $\mathcal{V}$ 有关。定义 $\mathcal{S}^{1/2}\triangleq\mathcal{V}\mathcal{D}\mathcal{V}^\top$. 由于 $\mathcal{S}$ 是正定矩阵，$\mathcal{S}$ 和 $\mathcal{S}^{1/2}$ 都可逆且有

$$\mathcal{S}^{-1/2}\triangleq\left(\mathcal{S}^{1/2}\right)^{-1}=\mathcal{V}\mathcal{D}^{-1}\mathcal{V}^\top.$$

$\mathcal{X}$ 的 **polar decomposition** 定义为

$$\begin{aligned}\mathcal{X}&=\mathcal{X}(\mathcal{X}^\top\mathcal{X})^{-1/2}(\mathcal{X}^\top\mathcal{X})^{1/2}\\&=\mathcal{H}\mathcal{S}^{1/2}\end{aligned}\tag{12}$$

其中 $\mathcal{H}\triangleq\mathcal{X}(\mathcal{X}^\top\mathcal{X})^{-1/2}=\mathcal{X}\mathcal{S}^{-1/2}$.

Remarks

1. 矩阵的 SVD 分解不唯一，矩阵的 polar decomposition 是唯一的。
2. 代入 $\mathcal{X}$ 的 SVD 分解，则有 $\mathcal{H}=\mathcal{U}\mathcal{D}\mathcal{V}^\top\mathcal{V}\mathcal{D}^{-1}\mathcal{V}^\top=\mathcal{U}\mathcal{V}^\top$, 可见 $\mathcal{H}\in\mathbb{V}_{d,n}$.
3. 当 $\mathcal{X}\sim N_{n\times d}(\mathbf{0},I_n,\Sigma)$, $\mathcal{U}\sim\boldsymbol{U}(\mathbb{V}_{d,n})$. $\mathcal{H}$ 只是对 $\mathcal{U}$ 做了一个正交变换，不改变列向量的长度和夹角，因此 $\mathcal{H}\sim\boldsymbol{U}(\mathbb{V}_{d,n})$. $\mathcal{U}$ 及 $\mathcal{H}=\mathcal{U}\mathcal{V}^\top$ 分布的均匀性与 $\mathcal{X}$ 的行向量间独立有关。
4. $\mathcal{H}\mathcal{H}^\top=\mathcal{X}(\mathcal{X}^\top\mathcal{X})^{-1}\mathcal{X}^\top$ 是 $\mathbb{R}^n$ 上的投影矩阵，可将 $\mathbb{R}^n$ 上的向量投影到 $\mathcal{X}$ 所在的 d 维子空间。比如
   - $\mathcal{H}\mathcal{H}^\top\mathcal{X}=\mathcal{X}$
   - $\mathcal{H}\mathcal{H}^\top\boldsymbol{y}=\mathcal{X}(\mathcal{X}^\top\mathcal{X})^{-1}\mathcal{X}^\top\boldsymbol{y}=\mathcal{X}\hat{\boldsymbol{\beta}}$, 得到的系数向量 $\hat{\boldsymbol{\beta}}$ 与 OLS 估计量一致，这也是 OLS 的几何解释。

正态随机矩阵的 polar decomposition 有以下定理：

Theorem 5. *令 $n\times d$ 随机矩阵 $\mathcal{X}$ 的 polar decomposition 为 $\mathcal{X}=\mathcal{H}\mathcal{S}^{1/2}$. 则 $\mathcal{X}\sim N_{n\times d}(\mathbf{0},I_n,\Sigma)$ 当且仅当*

1. *$\mathcal{H}\sim\boldsymbol{U}(\mathbb{V}_{d,n})$;*
2. *$\mathcal{S}\sim W_d(\Sigma,n)$;*
3. *$\mathcal{H}$ 和 $\mathcal{S}$ 独立。*

当 $\mathcal{H}$ 与 $\mathcal{S}$ 独立时, $\mathcal{H}$ 与 $\mathcal{S}^{1/2}$ 也是独立的。

8.3.2 Inverse Wishart 分布

在 Bayesian 模型中，我们经常会用到 Wishart 随机矩阵的逆。当 $\nu > d-1$ 且 Σ 是 $d \times d$ 对称正定矩阵时, $\mathcal{W} \sim W_d(\Sigma, \nu)$ 以概率 1 可逆，称 $\mathcal{W}^{-1}$ 服从的分布为 inverse Wishart 分布，记为 $\mathcal{W}^{-1} \sim IW_d(\Sigma^{-1}, \nu)$.

利用逆变换关系寻找 $IW_d(\Sigma^{-1}, \nu)$ 的 PDF. $\mathbb{R}^{d\times d}$ 上矩阵变换 $M = W^{-1}$ 对应的 Jacobian 行列式为

$$\left|\frac{\partial W}{\partial M}\right| = |M|^{-(d+1)}$$

当 $\mathcal{W} \sim W_d(\Sigma, \nu)$, 令 $\mathcal{M} = \mathcal{W}^{-1}$, $\Psi = \Sigma^{-1}$, 则 $\mathcal{M} \sim IW_d(\Psi, \nu)$ 的 PDF 为

$$\begin{aligned} f_{\mathcal{M}}(M) &= f_{\mathcal{W}}(M^{-1}) \cdot \left|\frac{\partial W}{\partial M}\right| \\ &\propto |M|^{-(\nu-d-1)/2-(d+1)} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Psi M^{-1})\right\} \\ &\propto |M|^{-(\nu+d+1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\Psi M^{-1})\right\} \end{aligned}$$

- 当 $\nu > d-1$ 时，$\mathcal{M} \sim IW_d(\Psi, \nu)$ 的期望存在, $E(\mathcal{M}) = \Psi/(\nu - d - 1)$.
- **Normal-inverse-Wishart 分布**。如果数据

  $$\boldsymbol{x}_i \overset{iid}{\sim} N_d(\boldsymbol{\mu}, \Sigma),\ i = 1, \ldots, n.$$

  使用 Bayesian 模型，我们需要给参数 $\boldsymbol{\mu}$ 和 Σ 设定先验 (prior) 分布，然后计算给定观察值 $\boldsymbol{x}_{1:n}$ 下参数的后验 (posterior) 分布。从后验分布可以估计观察到数据 $\boldsymbol{x}_{1:n}$ 后参数的期望、方差以及置信区间等。在 Bayesian 多元正态模型中，normal-inverse-Wishart 分布是 $(\boldsymbol{\mu}, \Sigma)$ 的共轭 (conjugate) 先验分布，因为由此推导出的 $(\boldsymbol{\mu}, \Sigma)$ 的后验分布也是一个 normal-inverse-Wishart 分布。

  如果

  $$\begin{aligned} \Sigma &\sim IW_d(\Omega_0, \nu_0) \\ \boldsymbol{\mu} \mid \Sigma &\sim N_d(\boldsymbol{\mu}_0, \Sigma/\kappa_0) \end{aligned}$$

  称 $\boldsymbol{\mu} \in \mathbb{R}^d$ 和 $\Sigma \in \mathbb{R}^{d\times d}$ 服从 **normal-inverse-Wishart** 分布 $(\boldsymbol{\mu}, \Sigma) \sim NIW_d(\boldsymbol{\mu}_0, \kappa_0, \Omega_0, \nu_0)$, 其中 $\boldsymbol{\mu}_0, \kappa_0, \Omega_0, \nu_0$ 都是已知的或主观设定的值。Normal-inverse-Wishart prior 可以理解为：在没有看到数据前，根据经验或历史数据猜测 $\Sigma \sim IW_d(\Omega_0, \nu_0)$ (prior); 给定 Σ, 猜测 $\boldsymbol{\mu}$ 是 κ_0 个独立的 $N_d(\boldsymbol{\mu}_0, \Sigma)$ 随机向量的平均值，所以 $\boldsymbol{\mu} \mid \Sigma$ 的 prior 为 $N_d(\boldsymbol{\mu}_0, \Sigma/\kappa_0)$. 注意 $\kappa_0 > 0$ 不一定为整数，κ_0 越大表明我们对 $\boldsymbol{\mu}$ 的先验分布的不确定性越小。

观察到数据 $\boldsymbol{x}_i \overset{iid}{\sim} N_d(\boldsymbol{\mu}, \Sigma)$, $i = 1, \ldots, n$ 后, $(\boldsymbol{\mu}, \Sigma)$ 的后验分布为:

$$
\begin{aligned}
p(\boldsymbol{\mu}, \Sigma \mid \boldsymbol{x}_{1:n}) &= \frac{p(\boldsymbol{x}_{1:n}, \boldsymbol{\mu}, \Sigma)}{p(\boldsymbol{x}_{1:n})} \\
&\propto p(\boldsymbol{x}_{1:n} \mid \boldsymbol{\mu}, \Sigma) p(\boldsymbol{\mu}, \Sigma) \\
&\propto p(\Sigma) p(\boldsymbol{\mu} \mid \Sigma) \prod_{i=1}^{n} p(\boldsymbol{x}_i \mid \boldsymbol{\mu}, \Sigma) \\
&\propto |\Sigma|^{-(\nu_0+d+1)/2} \exp\left[-\frac{1}{2}\mathrm{tr}(\Omega_0 \Sigma^{-1})\right] |\Sigma|^{-1/2} \exp\left[-\frac{\kappa_0}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^\top \Sigma^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)\right] \cdot \\
&\qquad \prod_{i=1}^{n} \left\{ |\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right] \right\} \\
&\propto |\Sigma|^{-(\nu_0+d+1+1+n)/2} \exp\left[-\frac{1}{2}\mathrm{tr}(\Omega_0 \Sigma^{-1})\right] \exp\left[-\frac{\kappa_0}{2}\left(\boldsymbol{\mu}^\top \Sigma^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}_0^\top \Sigma^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}_0^\top \Sigma^{-1}\boldsymbol{\mu}_0\right)\right] \cdot \\
&\qquad \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(\boldsymbol{\mu}^\top \Sigma^{-1}\boldsymbol{\mu} - 2\boldsymbol{x}_i^\top \Sigma^{-1}\boldsymbol{\mu} + \boldsymbol{x}_i^\top \Sigma^{-1}\boldsymbol{x}_i\right) \right\} \\
&\propto |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2}\left[\boldsymbol{\mu}^\top(\kappa_0+n)\Sigma^{-1}\boldsymbol{\mu} - 2\left(\frac{\kappa_0\boldsymbol{\mu}_0 + \sum_{i=1}^n \boldsymbol{x}_i}{\kappa_0+n}\right)^\top (\kappa_0+n)\Sigma^{-1}\boldsymbol{\mu} + \right.\right. \\
&\left.\left. \left(\frac{\kappa_0\boldsymbol{\mu}_0 + \sum_{i=1}^n \boldsymbol{x}_i}{\kappa_0+n}\right)^\top (\kappa_0+n)\Sigma^{-1}\left(\frac{\kappa_0\boldsymbol{\mu}_0 + \sum_{i=1}^n \boldsymbol{x}_i}{\kappa_0+n}\right) \right] \right\} \cdot |\Sigma|^{-(\nu_0+n+d+1)/2} \exp\left\{ -\frac{1}{2}\mathrm{tr}(\Omega_0\Sigma^{-1}) \right\} \cdot \\
&\qquad \exp\left\{ -\frac{1}{2}\left[\kappa_0 \boldsymbol{\mu}_0^\top \Sigma^{-1}\boldsymbol{\mu}_0 + \sum_{i=1}^n \boldsymbol{x}_i^\top \Sigma^{-1}\boldsymbol{x}_i - \frac{1}{\kappa_0+n}(\kappa_0\boldsymbol{\mu}_0 + n\bar{\boldsymbol{x}})^\top \Sigma^{-1}(\kappa_0\boldsymbol{\mu}_0 + n\bar{\boldsymbol{x}}) \right] \right\} \\
&\propto N\left(\boldsymbol{\mu} \mid \frac{\kappa_0\boldsymbol{\mu}_0 + n\bar{\boldsymbol{x}}}{\kappa_0+n}, \frac{\Sigma}{\kappa_0+n}\right) |\Sigma|^{-(\nu_0+n+d+1)/2} \exp\left\{ -\frac{1}{2}\mathrm{tr}(\Omega_0\Sigma^{-1}) \right\} \cdot \\
&\qquad \exp\left\{ -\frac{1}{2(\kappa_0+n)}\left[n\kappa_0 \boldsymbol{\mu}_0^\top \Sigma^{-1}\boldsymbol{\mu}_0 - 2n\kappa_0 \boldsymbol{\mu}_0^\top \Sigma^{-1}\bar{\boldsymbol{x}} + (n+\kappa_0)\sum_{i=1}^n \boldsymbol{x}_i^\top \Sigma^{-1}\boldsymbol{x}_i - n^2 \bar{\boldsymbol{x}}^\top \Sigma^{-1}\bar{\boldsymbol{x}} \right] \right\} \\
&\propto N\left(\boldsymbol{\mu} \mid \frac{\kappa_0\boldsymbol{\mu}_0 + n\bar{\boldsymbol{x}}}{\kappa_0+n}, \frac{\Sigma}{\kappa_0+n}\right) |\Sigma|^{-(\nu_0+n+d+1)/2} \exp\left\{ -\frac{1}{2}\mathrm{tr}(\Omega_0\Sigma^{-1}) \right\} \cdot \\
&\qquad \exp\left\{ -\frac{1}{2(\kappa_0+n)}\left[n\kappa_0 \left(\boldsymbol{\mu}_0^\top \Sigma^{-1}\boldsymbol{\mu}_0 - 2\boldsymbol{\mu}_0^\top \Sigma^{-1}\bar{\boldsymbol{x}} + \bar{\boldsymbol{x}}^\top \Sigma^{-1}\bar{\boldsymbol{x}}\right) + \right.\right. \\
&\qquad \left.\left. (n+\kappa_0)\left(\sum_{i=1}^n \boldsymbol{x}_i^\top \Sigma^{-1}\boldsymbol{x}_i - n\bar{\boldsymbol{x}}^\top \Sigma^{-1}\bar{\boldsymbol{x}}\right) \right] \right\} \\
&\propto N\left(\boldsymbol{\mu} \mid \frac{\kappa_0\boldsymbol{\mu}_0 + n\bar{\boldsymbol{x}}}{\kappa_0+n}, \frac{\Sigma}{\kappa_0+n}\right) |\Sigma|^{-(\nu_0+n+d+1)/2} \exp\left\{ -\frac{1}{2}\mathrm{tr}(\Omega_0\Sigma^{-1}) \right\} \cdot \\
&\qquad \exp\left\{ -\frac{1}{2}\left[\frac{n\kappa_0}{n+\kappa_0}\mathrm{tr}\left((\boldsymbol{\mu}_0-\bar{\boldsymbol{x}})^\top \Sigma^{-1}(\boldsymbol{\mu}_0-\bar{\boldsymbol{x}})\right) + \sum_{i=1}^n \mathrm{tr}\left((\boldsymbol{x}_i-\bar{\boldsymbol{x}})^\top \Sigma^{-1}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})\right) \right] \right\} \\
&\propto N\left(\boldsymbol{\mu} \mid \frac{\kappa_0\boldsymbol{\mu}_0 + n\bar{\boldsymbol{x}}}{\kappa_0+n}, \frac{\Sigma}{\kappa_0+n}\right) |\Sigma|^{-(\nu_0+n+d+1)/2} \cdot \\
&\qquad \exp\left\{ -\frac{1}{2}\mathrm{tr}\left(\left[\Omega_0 + \frac{n\kappa_0}{n+\kappa_0}(\boldsymbol{\mu}_0-\bar{\boldsymbol{x}})(\boldsymbol{\mu}_0-\bar{\boldsymbol{x}})^\top + \sum_{i=1}^n (\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^\top \right] \Sigma^{-1} \right) \right\}
\end{aligned}
$$

由此可得 $(\boldsymbol{\mu}, \Sigma)$ 的 posterior 为

$$\Sigma \mid \boldsymbol{x}_{1:n} \sim IW_d(\Omega_n, \nu_n)$$
$$\boldsymbol{\mu} \mid \Sigma, \boldsymbol{x}_{1:n} \sim N_d(\boldsymbol{\mu}_n, \Sigma/\kappa_n)$$

其中

$$\Omega_n = \Omega_0 + \frac{n\kappa_0}{n+\kappa_0}(\boldsymbol{\mu}_0 - \bar{\boldsymbol{x}})(\boldsymbol{\mu}_0 - \bar{\boldsymbol{x}})^\top + \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$$
$$\nu_n = \nu_0 + n$$
$$\boldsymbol{\mu}_n = \frac{\kappa_0\boldsymbol{\mu}_0 + n\bar{\boldsymbol{x}}}{\kappa_0 + n}$$
$$\kappa_n = \kappa_0 + n$$

即 $(\boldsymbol{\mu}, \Sigma) \mid \boldsymbol{x}_{1:n} \sim NIW_d(\boldsymbol{\mu}_n, \kappa_n, \Omega_n, \nu_n)$.

Remarks

1. 从上述结果可以看出, 每个观察值都会使 $(\boldsymbol{\mu}, \Sigma)$ posterior 中的 ν 和 κ 增加 1.
2. $\boldsymbol{\mu}$ 的 posterior mean $\boldsymbol{\mu}_n$ 是 prior mean $\boldsymbol{\mu}_0$ 和样本均值 $\bar{\boldsymbol{x}}$ 的加权和, 且权重与各自的样本数有关。如果选取了一个很大的 prior 样本数 κ_0, 则实际数据对 $\boldsymbol{\mu}$ 的 posterior 影响很小，因为 $\boldsymbol{\mu}$ 的 prior 占据了很大权重；如果选取的 κ_0 很小，数据很多 $n \gg \kappa_0$, 则 $\boldsymbol{\mu}$ 的 posterior 主要受实际数据影响，此时 $\boldsymbol{\mu}$ 的 posterior mean 会很接近样本均值 $\bar{\boldsymbol{x}}$.

References

Bartlett, M. S. (1933). On the theory of statistical regression. *Proceedings of the Royal Society of Edinburgh*, 53:260–283.

Hoff, P. D. (2009). Simulation of the matrix bingham–von mises–fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2):438–456.

Kotz, S., Balakrishnan, N., and Johnson, N. (2000). *Continuous Multivariate Distributions: Models and Applications*, volume 1. Wiley, New York, 2nd edition.

McNeil, A. J., Frey, R., Embrechts, P., et al. (2005). *Quantitative risk management: Concepts, techniques and tools*, volume 3. Princeton university press Princeton.

Sklar, M. (1959). *Fonctions de Répartition À N Dimensions Et Leurs Marges*. Université Paris 8.

Ulrich, G. (1984). Computer generation of distributions on the m-sphere. *Applied Statistics*, pages 158–163.

Wood, A. T. (1994). Simulation of the von mises fisher distribution. *Communications in statistics-simulation and computation*, 23(1):157–164.

