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见多元分布的抽样，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分量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 $U_j \overset{iid}{\sim} \boldsymbol{U}(0,1),\ j = 1, 2, \ldots, d$, 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，抽样 $\boldsymbol{Y} \sim g$ 后只保留 $\|\boldsymbol{Y}\| \le 1$ 的样本，来自 g 的样本平均被接受的概率为

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

    - ★ $d = 2$ 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

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

  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

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

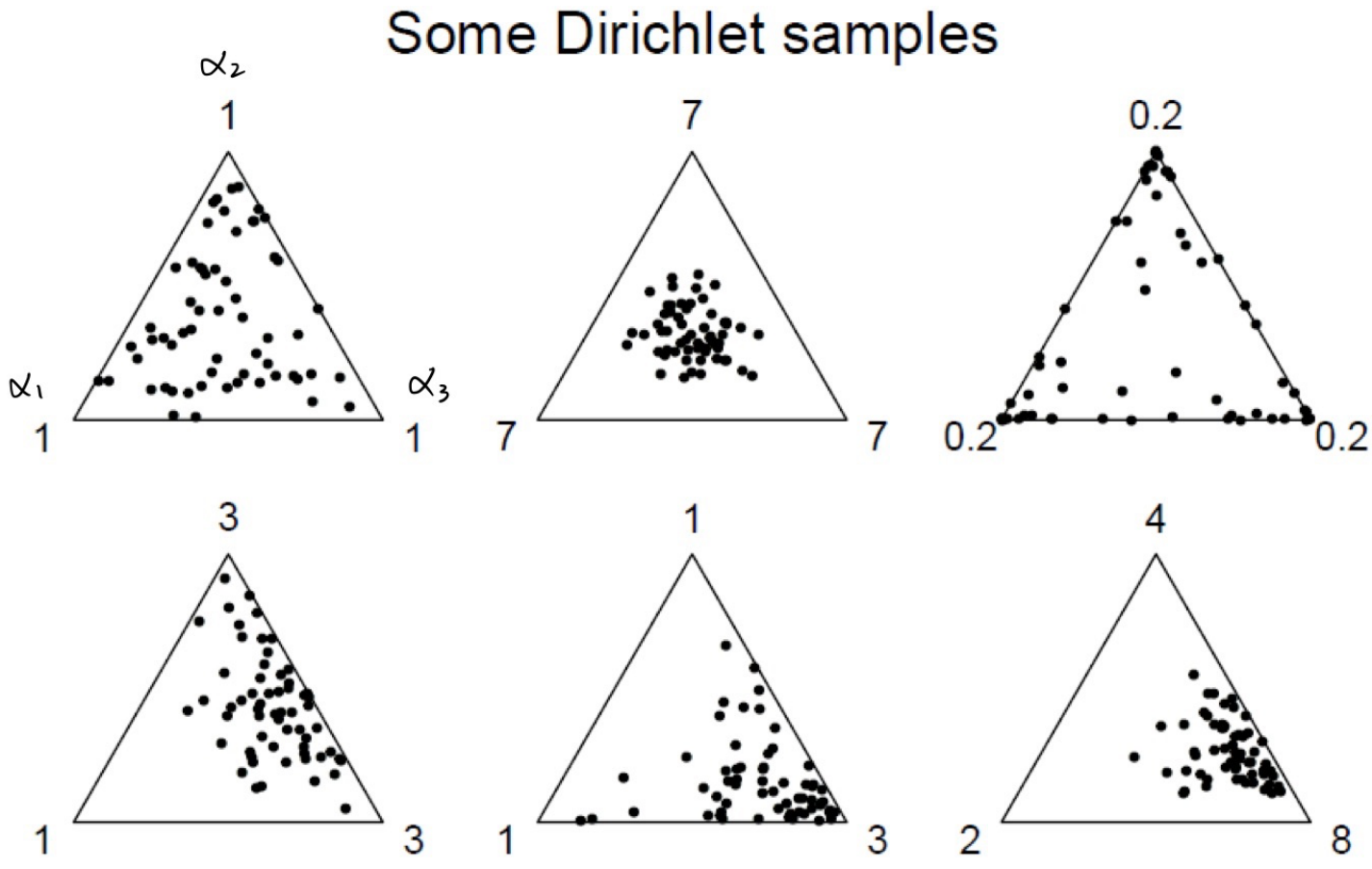
  $\boldsymbol{X} \sim \mathrm{Dir}(\boldsymbol{\alpha})$ 的期望为

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

- $\mathrm{Dir}(\alpha_1, \alpha_2)$ 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

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

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

- $\boldsymbol{U}(\Delta^{d-1})$ 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方法, 可看作一元的 QQ 变换在多元的推广
- $\mathbb{R}^d$ 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:

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

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:
  - 抽样 $\boldsymbol{Z} \sim N_d(\mathbf{0}, I_d)$
  - 令 $\boldsymbol{Y} = R^{1/2}\boldsymbol{Z}$, 则 $\boldsymbol{Y} \sim N_d(\mathbf{0}, R)$ 且 $Y_j \sim N(0,1)$, $j = 1, \ldots, d$
  - 令 $X_j = F_j^{-1}(\Phi(Y_j))$, $j = 1, \ldots, d$.

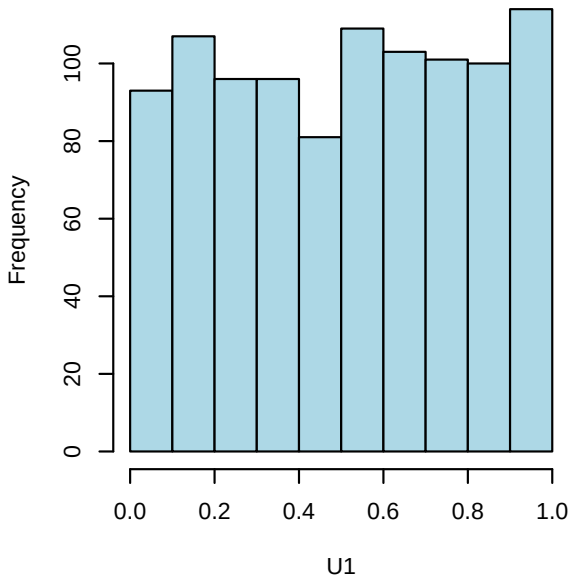
- Gaussian-copula 方法可以将多元正态分布的相关结构和一些边际 CDF 结合产生新的分布，因此也被称为 NORTA 方法 (normal to anything)

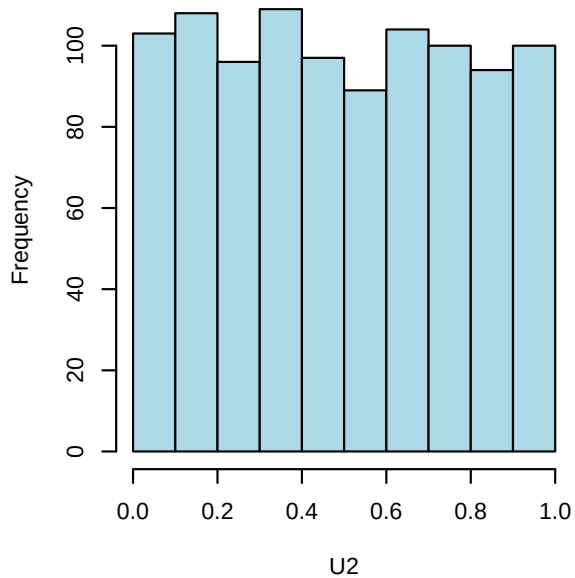
- 一个将 Gaussian-copula 与 Gamma 边际分布结合的例子

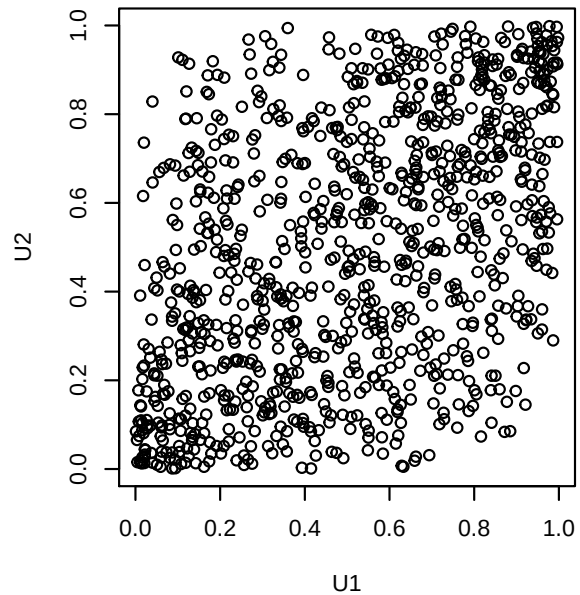
Gaussian-copula 与 Gamma 边际分布结合

```
## -- 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

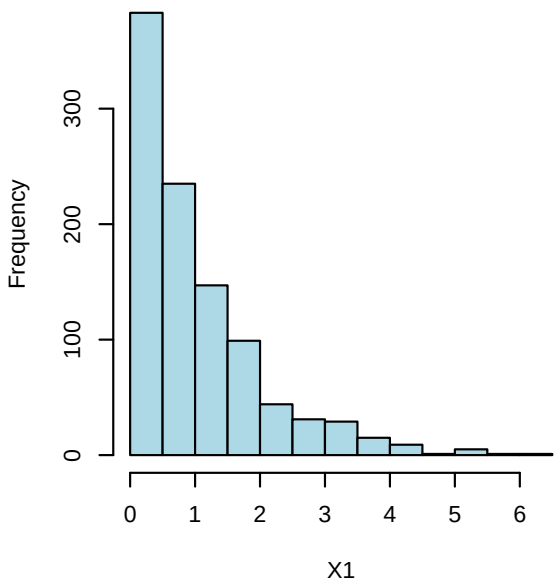


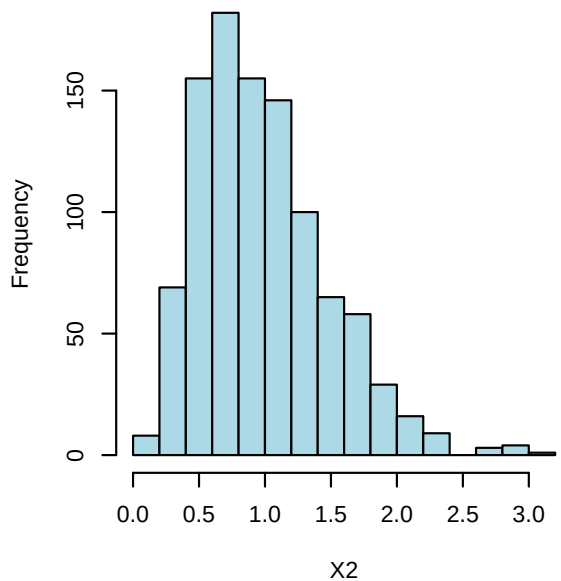


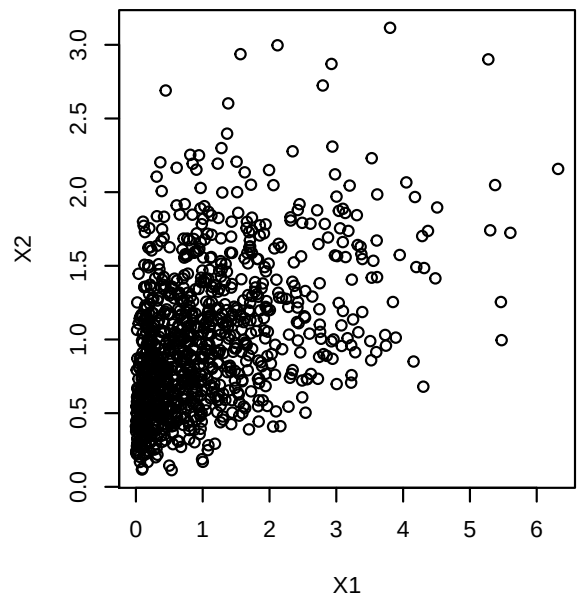


Gaussian-copula 与 Gamma 边际分布结合

```
# 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，无法使用 $\mathrm{Cov}(\boldsymbol{X})$ 或 $\mathrm{Corr}(\boldsymbol{X})$ 考察 $\boldsymbol{X}$ 各分量间的相关结构，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)

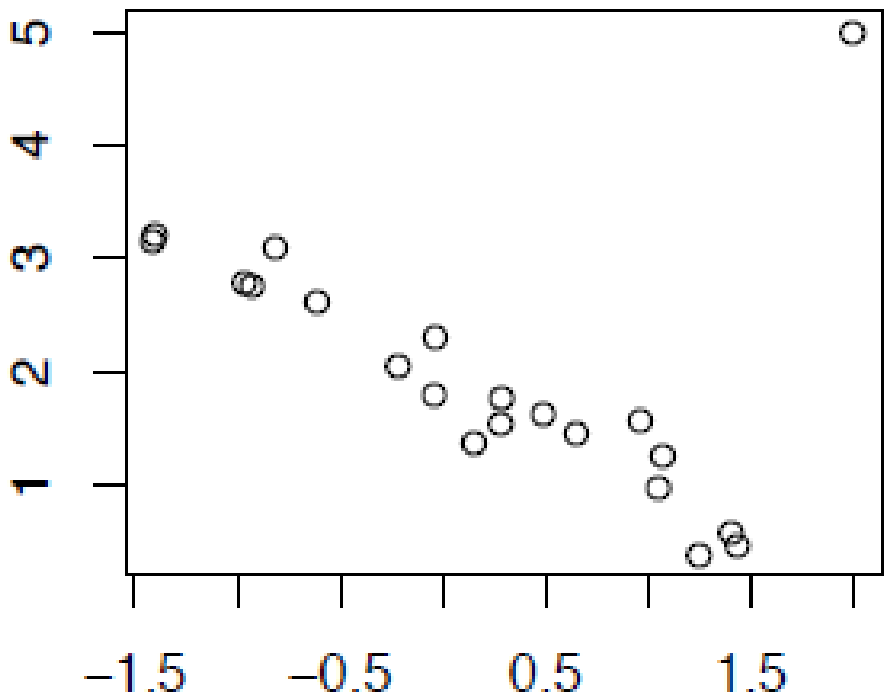
对于 (X, Y) 的任意两对观察值 (x_i, y_i) 和 (x_j, y_j), $i < j$, 如果 $x_i < x_j$ 且 $y_i < y_j$, 或者 $x_i > x_j$ 且 $y_i > y_j$, 称 (x_i, y_i) 和 (x_j, y_j) 是一致的 (concordant), 否则是不一致的 (disconcordant). 如果 $x_i = x_j$ 或者 $y_i = y_j$, 则认为 (x_i, y_i) 和 (x_j, y_j) 既不是一致的也不是不一致的. $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的 Kendall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

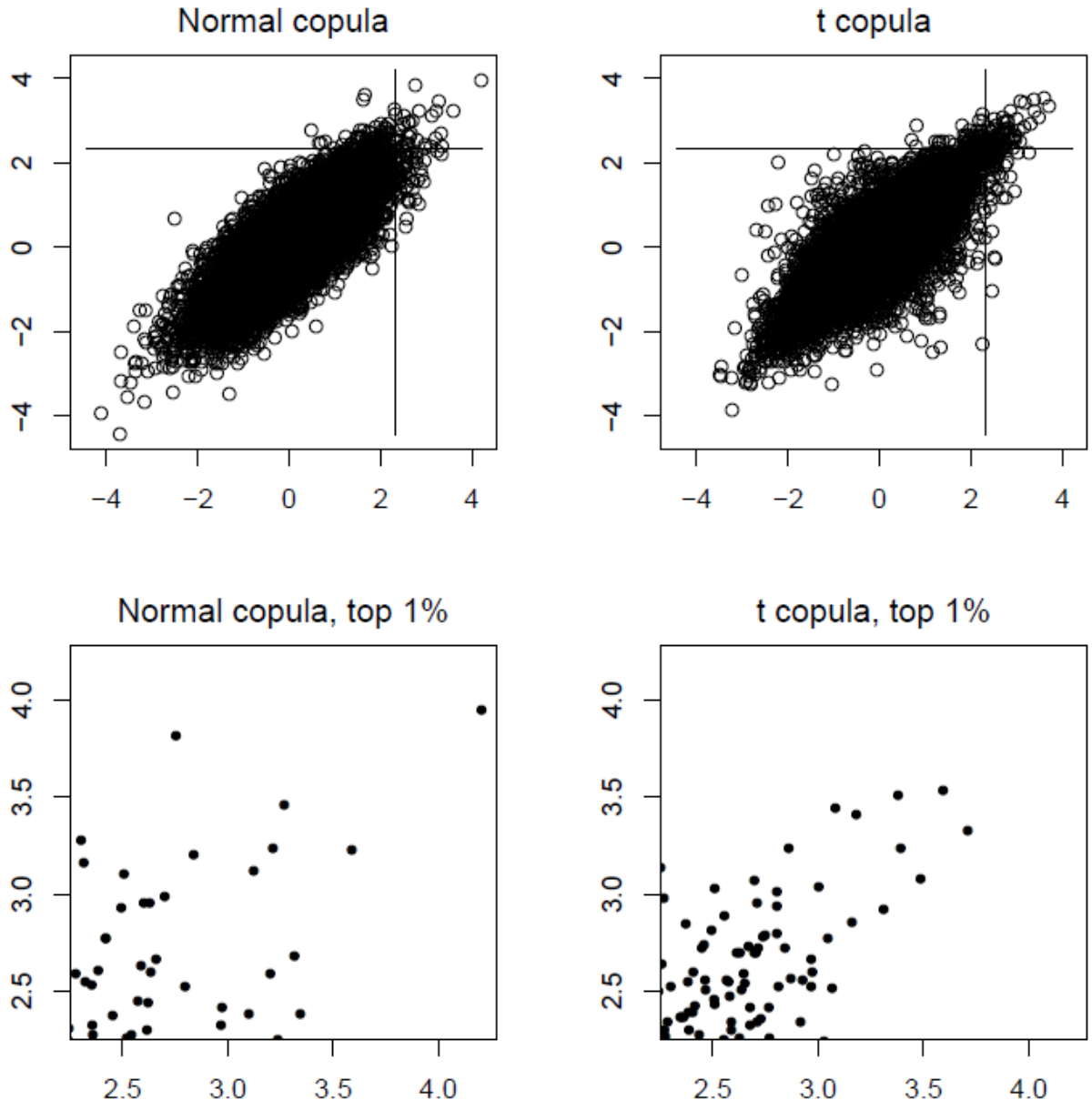
- Gaussian copula 方法的一个缺点 — 尾部独立性(tail independence)，即如果 $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$，则任意两个分量 X_j 和 X_k 有如下性质

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

  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

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

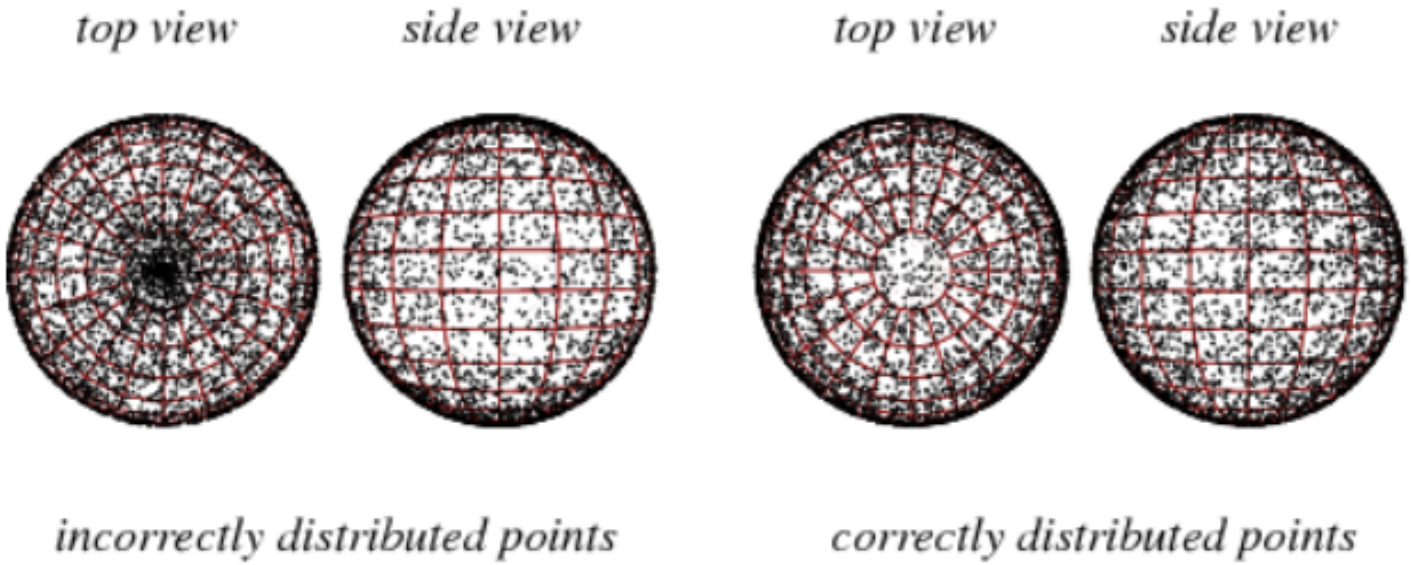
  其中参数 $\theta > 0$

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

矩阵正态分布

- 有些实际问题需要生成的随机矩阵 $\mathcal{X} \in \mathbb{R}^{n\times d}$ 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，直接生成一个随机矩阵可能比生成一个高维的随机向量容易

- 矩阵正态分布常用于描述行和列都有相关性的数据矩阵，如基因数据、不同时期多种商品价格的面板数据

- $\mathbb{R}^{n\times d}$ 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

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

  - 令 $\mathrm{tr}(\Gamma) = m$ 保证参数可识别

矩阵正态分布

- $\mathbb{R}^{nd}$ 上的正态分布一般需要 $nd(nd+1)/2$ 个参数描述协方差矩阵，但是矩阵正态分布只需要 $n(n+1)/2+d(d+1)/2$ 个参数，当 n 和 d 很大时，矩阵正态分布可以极大地减少参数个数
- $\mathcal{X} \sim N_{n\times d}(M,\Gamma,\Sigma)$，如果将 $\mathcal{X}$ 的各列首尾相连排列成一个 $nd\times 1$ 的向量 $vec(\mathcal{X})$，则 $vec(\mathcal{X}) \sim N_{nd}\left(vec(M),\Sigma\otimes\Gamma\right)$，其中 $\otimes$ 为 Kronecker product
- 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

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

- 找到矩阵 A 和 B 满足 $\Gamma = AA^{\top}$，$\Sigma = BB^{\top}$，就可如下对 $N_{n\times d}(M,\Gamma,\Sigma)$ 分布抽样

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

随机正交矩阵

- d 阶正交矩阵组成的空间记为

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

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

- 如何对 $\mathcal{Q} \sim \boldsymbol{U}(\mathbb{O}_d)$ 抽样？
  - $\mathcal{Q}_{\cdot 1} \sim \boldsymbol{U}(S^{d-1})$, 可令 $\mathcal{Q}_{\cdot 1} = \dfrac{\boldsymbol{Z}_1}{\|\boldsymbol{Z}_1\|}$, $\boldsymbol{Z}_1 \sim N_d(\mathbf{0}, I_d)$
  - 给定 $\mathcal{Q}_{\cdot 1}$, $\mathcal{Q}_{\cdot 2}$ 在与 $\mathcal{Q}_{\cdot 1}$ 垂直的单位圆 (球面) 上均匀分布，可以如下产生：$\boldsymbol{Z}_2 \sim N_d(\mathbf{0}, I_d)$, 令 $\tilde{\boldsymbol{Z}}_2 = \boldsymbol{Z}_2 - (\boldsymbol{Z}_2^\top \mathcal{Q}_{\cdot 1})\mathcal{Q}_{\cdot 1}, \mathcal{Q}_{\cdot 2} = \tilde{\boldsymbol{Z}}_2 / \left\|\tilde{\boldsymbol{Z}}_2\right\|$
  - 继续从 $N_d(\mathbf{0}, I_d)$ 抽样，Gram-Schmidt 正交化及单位化，依次产生后面几列向量

随机正交矩阵

$\mathbb{O}_d$ 上一个重要的非均匀分布是 **Bingham** 分布，Bingham(L, Ψ) 的 PDF 为

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

其中参数矩阵 L 是 $d \times d$ 对角矩阵 (对角线元素递减排列)，Ψ 是 $d \times d$ 对称矩阵

- Bingham 分布的期望是 $\mathbf{0}_{d\times d}$
- Bingham 分布具有 **antipodal symmetry**: 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 $\mathcal{Q}S \overset{d}{=} \mathcal{Q}$
- Bingham 分布的 mode 是什么？

定理

$\mathbb{O}_d$ 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。

随机正交矩阵

- Hoff (2009) 提出一种基于 Gibbs sampler 对 Bingham 分布抽样的方法 (R Package `rstiefel`)
- 每个正交矩阵对应一个旋转变化，将 $\mathbb{R}^n$ 上的向量投影到一个 k 维子空间 ($k < n$) 对应一个 $n \times k$ 的投影矩阵 P，P 属于如下的Stiefel manifold:

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

- 对 $\boldsymbol{U}(\mathbb{V}_{k,n})$ 抽样，可以先抽 $\mathcal{Q} \sim \boldsymbol{U}(\mathbb{O}_n)$，然后只保留 $\mathcal{Q}$ 的前 k 列，或使用以下定理

定理

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

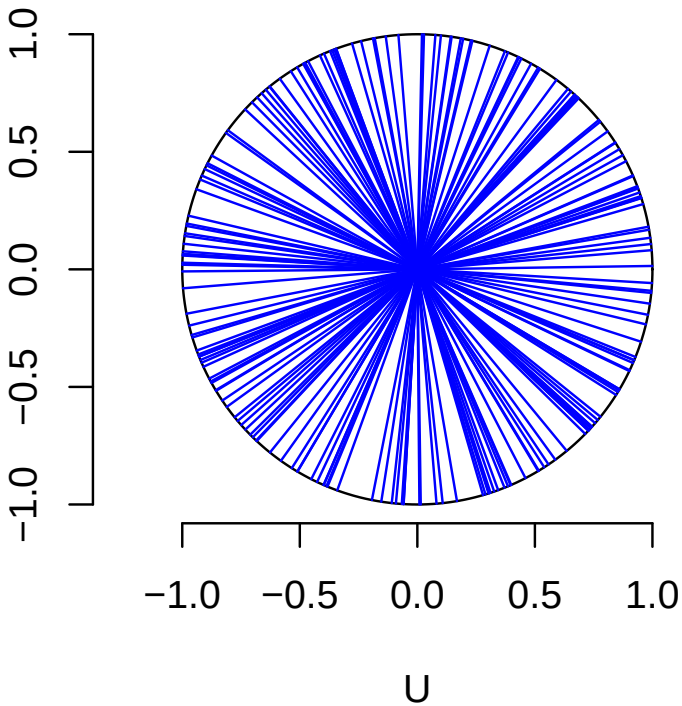
- $\mathcal{U} \sim \boldsymbol{U}(\mathbb{V}_{d,n})$，且 $\mathcal{U}$ 与 $(\mathcal{D}, \mathcal{V})$ 独立;
- $\mathcal{V} \mid \mathcal{D} \sim \mathit{Bingham}(\mathcal{D}^2, -\Sigma^{-1}/2)$;
- $\mathcal{D}^2$ 的对角线元素与 *Wishart* 分布 $W_d(\Sigma, n)$ 的随机矩阵的特征值同分布。

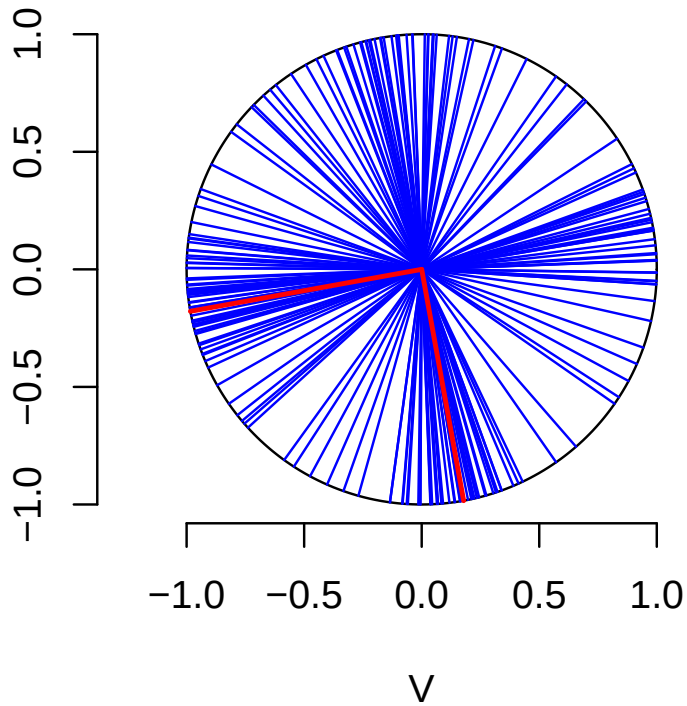
数值实验

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

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

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





Wishart 分布

- 如果 $\mathbb{R}^d$ 上的随机向量 $\boldsymbol{X}_i \overset{iid}{\sim} N_d(\mathbf{0}, \Sigma)$, $i = 1, \ldots, n$ 且 $n \geq d$, 则随机矩阵

$$\mathcal{W} = \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top \in \mathbb{R}^{d \times d} \tag{4}$$

  服从Wishart分布 $W_d(\Sigma, n)$.

- Wishart 分布 $W_d(\Sigma, \nu)$ 有两个参数: $d \times d$ 对称正定矩阵 Σ 和自由度 ν ($\nu > d-1$), PDF 为

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

- 如果 $\mathcal{W} \sim W_d(\Sigma, \nu)$, $E(\mathcal{W}) = \nu\Sigma$, 对任意矩阵 $C \in \mathbb{R}^{k \times d}$ ($k \leq d$),

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

  因此只需对 $W_d(I, \nu)$ 抽样即可得到任意 $W_d(\Sigma, \nu)$ 的样本

Wishart 分布

- 自由度为正整数的 Wishart 分布，可按(4)进行抽样，对任意的 $\nu > d-1$, 可采用如下的 Bartlett 分解对 $W_d(I, \nu)$ 抽样: L 是 $d \times d$ 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 $LL^\top \sim W_d(I, \nu)$

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

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

  因此对 Σ 的一个无偏估计量为 $\hat{\Sigma} = \dfrac{W}{n-1}$

Wishart 分布

定理

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

证明：

$W_d(\Sigma, \nu)$ 完整的 PDF 为

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

则 Σ 的对数似然函数为

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

正态随机矩阵的 polar decomposition

- 矩阵的 SVD 分解不唯一，矩阵的 polar decomposition 是唯一的
- 对 $\mathcal{X}$ 做 SVD 分解 $\mathcal{X} = \mathcal{U}\mathcal{D}\mathcal{V}^\top$，则有 $\mathcal{S} = \mathcal{X}^\top\mathcal{X} = \mathcal{V}\mathcal{D}^2\mathcal{V}^\top$，定义 $\mathcal{S}^{1/2} \triangleq \mathcal{V}\mathcal{D}\mathcal{V}^\top$
- $\mathcal{X}$ 的 polar decomposition 定义为

$$\mathcal{X} = \mathcal{X}(\mathcal{X}^\top\mathcal{X})^{-1/2}(\mathcal{X}^\top\mathcal{X})^{1/2} = \mathcal{H}\mathcal{S}^{1/2} \tag{5}$$

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

定理

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

Inverse Wishart 分布

- Bayesian 模型经常用到 Wishart 随机矩阵的逆
- 如果 $\mathcal{W} \sim W_d(\Sigma, \nu)$, 称 $\mathcal{W}^{-1}$ 服从的分布为 inverse Wishart 分布, 记为 $\mathcal{W}^{-1} \sim IW_d(\Sigma^{-1}, \nu)$
- 令 $\mathcal{M} = \mathcal{W}^{-1}$, $\Psi = \Sigma^{-1}$, $\mathcal{M} \sim IW_d(\Psi, \nu)$ 的 PDF 为

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

- 当 $\nu > d-1$ 时, $\mathcal{M} \sim IW_d(\Psi, \nu)$ 的期望为 $E(\mathcal{M}) = \Psi/(\nu - d - 1)$

Normal-inverse-Wishart 分布

- Bayesian 多元正态模型假设数据 $\boldsymbol{x}_i \overset{iid}{\sim} N_d(\boldsymbol{\mu}, \Sigma)$, $i = 1, \ldots, n$, normal-inverse-Wishart 分布是参数 $(\boldsymbol{\mu}, \Sigma)$ 的共轭先验分布
- 如果

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

  称 $\boldsymbol{\mu} \in \mathbb{R}^d$ 和 $\Sigma \in \mathbb{R}^{d \times d}$ 服从normal-inverse-Wishart分布
  $(\boldsymbol{\mu}, \Sigma) \sim NIW_d(\boldsymbol{\mu}_0, \kappa_0, \Omega_0, \nu_0)$

  - NIW prior 可以理解为: 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
  - 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

Normal-inverse-Wishart 分布

- 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:

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

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

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

$$\nu_n = \nu_0 + n$$

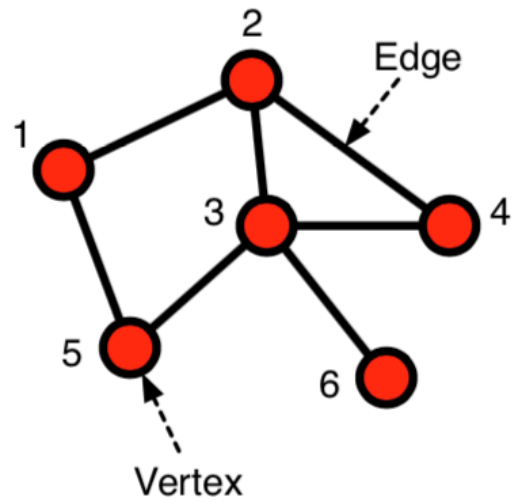
$$\boldsymbol{\mu}_n = \frac{\kappa_0\boldsymbol{\mu}_0 + n\bar{\boldsymbol{x}}}{\kappa_0 + n}$$

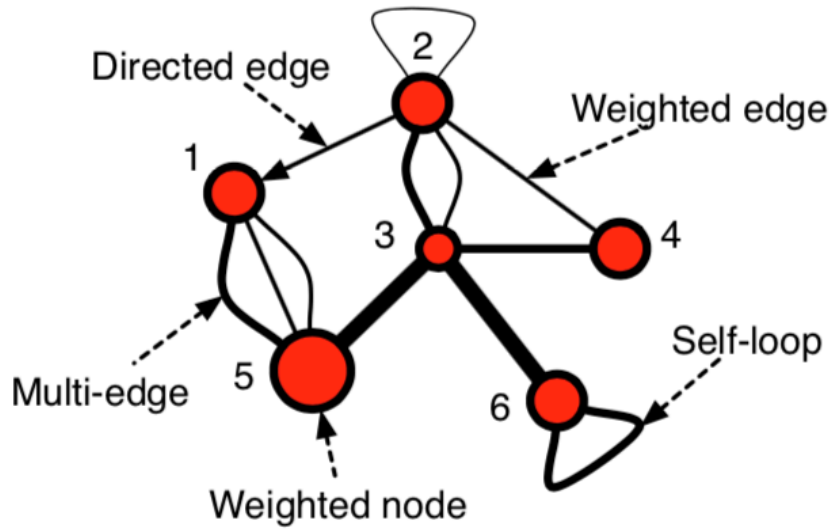
$$\kappa_n = \kappa_0 + n$$

- 每个观察值都会使 $(\boldsymbol{\mu}, \Sigma)$ 后验分布中的 ν 和 κ 增加 1
- $\boldsymbol{\mu}$ 的后验期望 $\boldsymbol{\mu}_n$ 是先验期望 $\boldsymbol{\mu}_0$ 和样本均值 $\bar{\boldsymbol{x}}$ 的加权和, 且权重与各自的样本数有关

随机图

- 随机图是一种常用的描述网络型数据的模型
- 一个图通常由一个节点的集合 $\mathcal{V}$ 和一个边的集合 $\mathcal{E}$ 组成，记为 $G(\mathcal{V}, \mathcal{E})$





- 图 G 中边的出现情况常用**邻接矩阵** A 表示. 对无向图，A 是一个 0-1 矩阵，如果边 $(i, j) \in \mathcal{E}$, 令 $A_{ij} = A_{ji} = 1$

随机图

一些反映图的拓扑特征的常用指标：

- 图密度. 图中出现的边数与可能出现的边数的比，对于有 n 个节点的无向图 $G(\mathcal{V}, \mathcal{E})$

$$\text{图密度} = \frac{\mathcal{E} \text{ 中的边数}}{\binom{n}{2}}$$

- 集聚系数/传递性 (Clustering coefficient / Transitivity). 社交网络中，集聚系数反映“我朋友的朋友还是我的朋友”的程度，因此也被称为传递性
  - 对于图 $G(\mathcal{V}, \mathcal{E})$ 中的三个节点 (i, j, k)，如果边 $(i, j) \in \mathcal{E}$，$(j, k) \in \mathcal{E}$，称 (i, j, k) 为一个 **triplet**
  - 如果边 (k, i) 也在 $\mathcal{E}$ 中，称 (i, j, k) 为一个 **closed triplet**
  - 图 G 的集聚系数定义为：

$$C = \frac{\text{closed triplets 的个数}}{\text{triplets 的个数}} = \frac{3 \times \text{三角形的个数}}{\text{triplets 的个数}}$$

$C \in [0, 1]$，C 越接近 1 表示图 G 中的节点相互连接集结成团的程度越高

随机图

- 平均最短路径长度. 图 G 中任意一对节点间的最短路径长度取平均值就是图 G 的**平均最短路径长度**
  - 图 $G(\mathcal{V}, \mathcal{E})$ 中的一条**路径**对应一系列点 $x \to y \to \cdots \to z$, 且每一对连续的点 $i \to j$ 都有边连接, 路径的长度为路径中边的个数
  - 如果平均最短路径长度以 $O(\log(n))$ 增长，同时图 G 又有较高的集聚系数，称图 G 具有 “小世界 (small world)” 性质
  - **小世界网络**是一种接近真实社会的网络结构，网络中的大部分节点彼此并不相连，但绝大部分节点之间只需经过几条边就可以到达

随机图

- 度数分布. 度数分布描述的是图 G 中节点度数的概率分布：从图 G 中随机抽一个节点，其度数为 k 的概率

$$P(k) = \text{度数为 } k \text{ 的节点在所有节点中所占的比例}$$

  - 节点的度数是图或网络中一个基本的测量指标，节点 i 的度数为所有与 i 相连的边数

$$k_i = \sum_{j \neq i} A_{ij},\ i = 1, \ldots, n$$

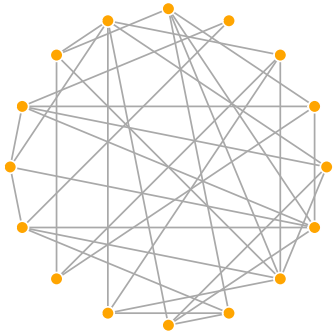
  - 很多真实世界网络的度数分布呈现一种“右偏”或“厚尾”的特点，即只有少数几个节点有很大的度数，有一些节点有中等的度数，绝大多数节点的度数都很小. 这些分布 (或仅其尾部) 遵循某种**幂律分布** (power law)，即度数 k $(k \geq 1)$ 出现的概率随着 k 增大以多项式速度递减

$$P(k) \propto \frac{1}{k^{\alpha}},\ \alpha > 0$$

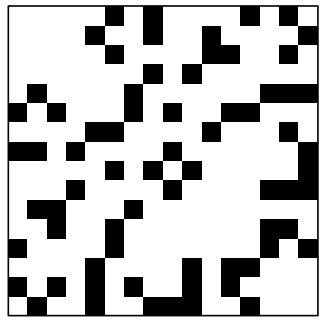
随机图模型

- Erdős-Rényi 模型. 该模型假设图 $G(\mathcal{V}, \mathcal{E})$ 的每条边是独立的且以相同的概率 p 出现
  - Erdős-Rényi 图和真实世界的网络相差很大，比如在边数相同的情况下，它所包含的三角形个数比**小世界图**少很多 (集聚系数较低)
  - 小世界图可以由 **Watts–Strogatz** 模型产生：把节点排成一个圆环，每个节点与左右各 K 个相邻节点连接，然后将每条边以某一概率重新连接 (rewire)
  - Watts–Strogatz 模型的邻居结构可以在图中产生很多三角形，少量的重新连接可以降低平均最短路径长度

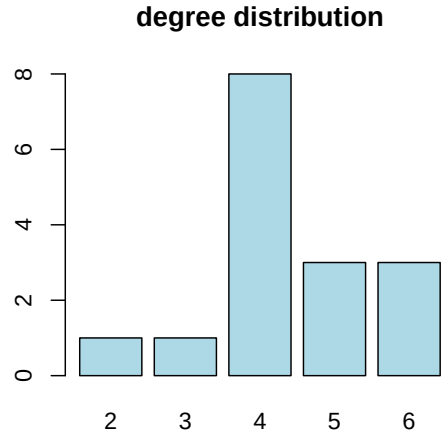
Erdős-Rényi 图 vs 小世界图

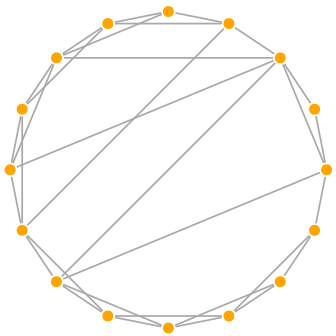


Erdos−Renyi

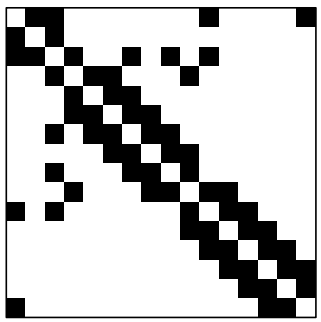


density = 0.29
transitivity = 0.19
ASPL = 1.88

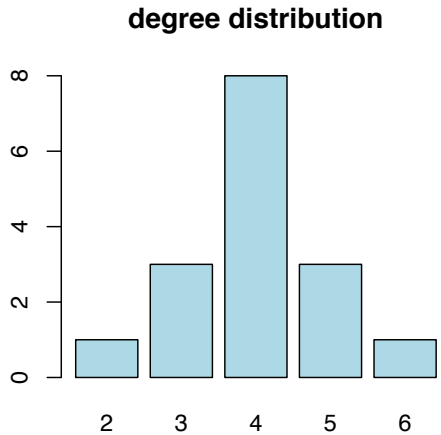




Small world

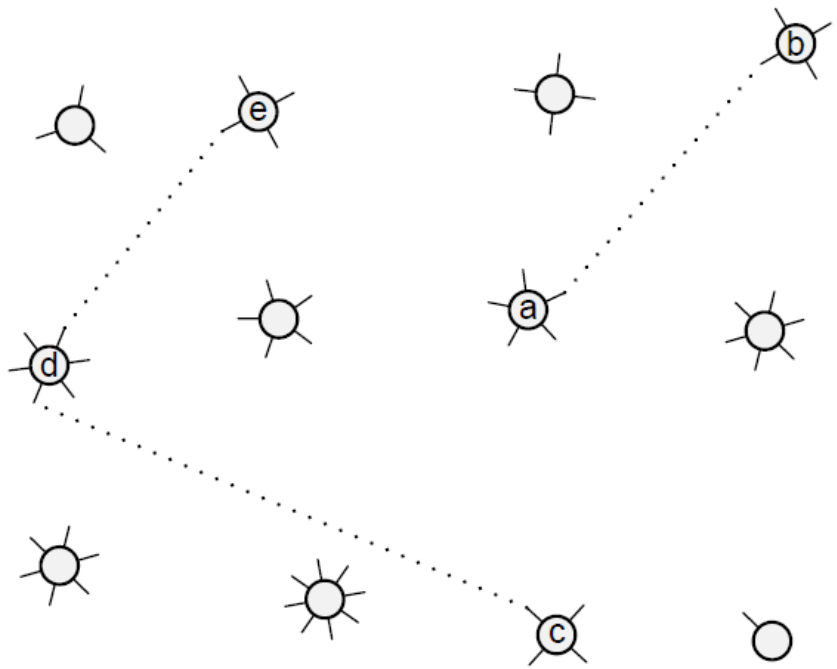


density = 0.27
transitivity = 0.38
ASPL = 2.23



随机图模型

- Configuration model. 该模型可以保证生成的随机图的度数分布与目标度数分布或实际观察的度数分布一致
  - 首先给定图中每个节点的目标度数 $\boldsymbol{k} = \{k_1, \ldots, k_n\}$, $\boldsymbol{k}$ 来自实际观察的图或者从目标度数分布中抽样 (比如，可以让 $\boldsymbol{k}$ 服从幂律分布)
  - 让每个节点 i 生出 k_i 个枝 (half edge) 形成一个 stub，然后将 stubs 的枝随机相连形成边



随机图模型

- Exponential random graph model (ERGM). ERGM 将一个图 G 简化为一个多元特征向量 $(\phi_1(G), \phi_2(G), \ldots, \phi_J(G))$, 并假设图 G 出现的概率密度只取决于这些统计量

$$f(G) \propto \exp\left(\sum_{j=1}^{J} \beta_j \phi_j(G)\right)$$

- Stochastic block model (SBM). 社交网络中同一团体内的成员之间建立联系的频率一般比不同团体的成员之间建立联系的频率高. SBM 假设节点 i 和节点 j 有边连接的概率 P_{ij} 为如下形式:

$$P_{ij} = \begin{cases} \rho_1, & i \text{ 和} j \text{ 在同一团体} \\ \rho_0, & \text{否则} \end{cases} \qquad \text{且 } \rho_1 \geq \rho_0$$

  - 更一般的 SBM 将节点集合 $\mathcal{V}$ 分成 k 个子集 $\mathcal{V}_1, \ldots, \mathcal{V}_k$, 每个子集 $\mathcal{V}_r$ 有 n_r 个节点, 如果节点 $i \in \mathcal{V}_r$, $j \in \mathcal{V}_s$, 则 $P_{ij} = \rho_{rs} = \rho_{sr}$

随机图模型

- Latent space model. 该模型可以用较少的参数产生边之间丰富的相关结构，是一种更灵活的随机图模型
  - 主要想法是给图中的每个节点 i 在低维空间中分配一组坐标 $\boldsymbol{x}_i \in \mathbb{R}^d$ ($d < n$, low-dimensional embedding)
    - ★ 在社交网络中，$\boldsymbol{x}_i$ 可能代表成员 i 的年龄，收入，学历，兴趣爱好等特征
  - 然后令节点 i 和节点 j 有边连接的概率 P_{ij} 为它们在低维空间的 (加权) 距离 $\sum_{r=1}^{d} \beta_r |x_{ir} - x_{jr}|$ 或 (加权) 点积 $\sum_{r=1}^{d} \beta_r x_{ir} x_{jr}$ 的函数
  - 给定节点的坐标，边的出现是独立的，即 $A_{ij} \overset{ind}{\sim} \text{Bern}(P_{ij})$
  - 该模型只需要 $O(nd)$ 个参数描述边的概率矩阵 $P \in \mathbb{R}^{n \times n}$