随机过程的生成

王璐

随机向量是有限个随机变量的集合，随机过程涉及无限个随机变量的集合。比如研究一个粒子随机运动时位置随时间的变化，我们既可以研究粒子在离散时间点 $t \in \{0, 1, 2, \dots\}$ 的位置，也可以研究粒子在一段时间 $[0, T]$ 内连续变化的位置，这两种方式都涉及无限个随机变量。很多非参数 Bayesian 模型的计算涉及对随机过程的抽样，本章我们将介绍如何对一些常见的随机过程进行抽样，比如随机游走 (random walk), Gaussian process, Poisson process 以及 Dirichlet process.

1 随机过程的一些基本概念

一个随机过程一般记为 $\{X(t) \mid t \in \mathcal{T}\}$, 指标集 (index set) $\mathcal{T}$ 可以是离散的集合, 如 $\mathcal{T} = \{1, 2, \dots\}$，或者连续的集合, 如 $\mathcal{T} = [0, \infty)$. 对于离散的随机过程，有时将 $X(t)$ 简记为 X_t. 如果有两个随机过程，一般记为 $X_1(t)$ 和 $X_2(t)$.

有些涉及空间的随机过程，指标集 $\mathcal{T}$ 可以是 $\mathbb{R}^d$ 上的一个区域, 比如 $X(t)$ 可能表示某个地点 t 的温度。这种定义在 $\mathbb{R}^d$ $(d > 1)$ 子集上的随机过程也被称为 **random field**.

随机过程 $\{X(t) \mid t \in \mathcal{T}\}$ 的一次实现定义了一个从 $\mathcal{T}$ 到 $\mathbb{R}$ 的随机函数 $f(\cdot)$. 随机函数 $f(\cdot)$ 也被称为该过程的一条**样本路径** (sample path).

实际应用中，虽然对随机过程的一次抽样只会产生有限个值，但会遇到与随机向量抽样不同的问题。比如，不断生成一个粒子在新时刻 t_j 的位置直到粒子离开某一特定区域。假设粒子的一条样本路径为 $(X(t_1), \dots, X(t_m))$, 终止时刻对应的 m 可以看作一个随机整数 M 的样本，即该向量的维度是随机的。虽然 $P(M < \infty) = 1$，但在抽样前我们对维度 M 并没有有界的预期。在有些随机过程中，我们选择抽样的时刻 t_j 还与之前抽到的某个 $X(t_k)$ 的取值有关。因此随机过程抽样的挑战在于如何用高效的方法前后一致地产生各部分的值。

随机过程主要通过它的任意有限维分布来描述。从指标集 $\mathcal{T}$ 选取任意有限个点 $t_1, \dots, t_m \in \mathcal{T}$, 观察随机过程在这些点的分布，称 $(X(t_1), \dots, X(t_m))$ 的联合分布为随机过程 $X(t)$ 的一个**有限维分布**。Kolmogorov's extension theorem 告诉我们如果一组有限维分布是相容的 (compatible), 即没有矛盾 (contradictions), 则一定存在一个随机过程具有这样的有限维分布。

但是有限维分布不能唯一确定一个随机过程，即两个不同的随机过程可能有完全相同的有限维分布，比如，令 $X(t)$ 为定义在指标集 $\mathcal{T}=[0,1]$ 上的一个随机过程, 随机抽取 $s\sim \boldsymbol{U}[0,1]$, 定义另一个随机过程 $Y(t)$ 如下：

$$Y(t)=\begin{cases} X(s)+1, & t=s \\ X(t), & t\neq s \end{cases}$$

由于 $Y(t)$ 与 $X(t)$ 只在一个零测集上不同, 因此

$$P\left(X(t_1)\leq x_1,\ldots,X(t_m)\leq x_m\right)\equiv P\left(Y(t_1)\leq x_1,\ldots,Y(t_m,)\leq x_m\right),\ \forall t_1,\ldots,t_m\in\mathcal{T}$$

即它们的任意有限维分布都相同。

如果需要研究的性质涉及随机过程 $X(t)$ 在无限个点 t 处的值，比如计算 $\mu=E\left[g\left(X(\cdot)\right)\right]$, 可以使用 Monte Carlo 方法做近似估计。此时需要从随机过程中生成 n 条样本路径 $X_i(t_{ij})$, $i=1,\ldots,n$. 假设第 i 条路径有 M_i 个点 (Monte Carlo 方法只能产生样本路径上有限个点)，则 μ 可估计为

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n} g\left(X_i(t_{i1}),\ldots,X_i(t_{iM_i})\right).$$

2 随机游走

随机游走一般具有以下形式

$$X_t=X_{t-1}+Z_t,\quad t=1,2,\ldots \tag{1}$$

其中 Z_t 是 iid 的随机向量。初始点 X_0 通常取为 $\mathbf{0}$. 如果我们知道如何对 Z_t 抽样，就很容易根据(1)生成 X_t 的路径。图1展示了离散和连续随机游走的一些样本路径。在这两个例子中, $E(Z_t)=0$. 如果 $E(Z_t)=\mu$, 称随机游走具有 drift μ. 如果 Z_t 的协方差矩阵 Σ 有限, 根据中心极限定理，

$$\frac{1}{\sqrt{t}}(X_t-t\boldsymbol{\mu})\to N(\mathbf{0},\Sigma),\quad t\to\infty$$

我们可以对随机游走进行扩展，使 Z_t 的分布随 t 变化, 比如 Pólya's urn process.

- **Pólya's urn process**. 在初始时刻，桶 (urn) 里有一个黑球和一个红球。在随后的每步，我们从桶里随机取出一个球，将该球和一个与它同色的球放回桶中。令 $X_t=(R_t,B_t)$, 其中 R_t 代表 t 时刻的红球数，B_t 代表黑球数。初始时刻 $X_0=(1,1)$, 每一步的增量 Z_t 服从如下分布：

$$Z_t=\begin{cases} (1,0), & \text{概率}=R_t/(R_t+B_t) \\ (0,1), & \text{概率}=B_t/(R_t+B_t) \end{cases}$$

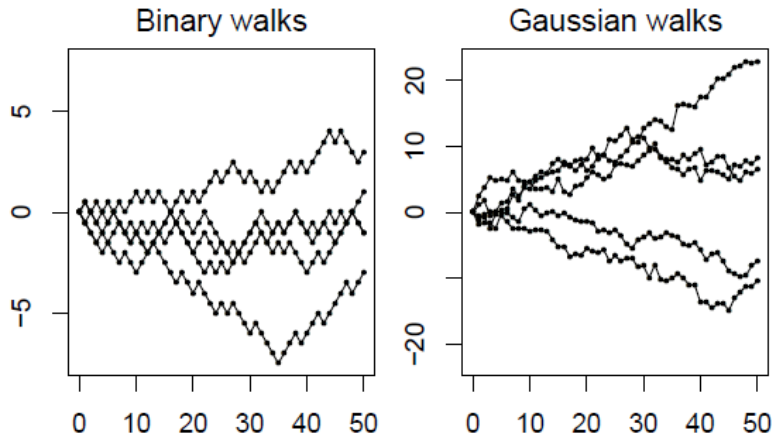

Figure 1: 离散和连续随机游走的 5 条样本路径。每条路径从 $X_0 = 0$ 开始, 持续 50 步。左图中的随机游走，增量 Z_t 取 ± 1 的概率各为 0.5; 右图 $Z_t \sim N(0, 1)$. Picture source: Art B. Owen

在这一过程中，我们感兴趣的变量是 $Y_t = R_t/(R_t + B_t)$, $t \to \infty$. 即足够长时间后桶中红球所占的比例。数学家 Pólya 证明上述过程 Y_t 的每条样本路径都会收敛到一个值 Y_∞, 但 Y_∞ 本身也是随机的，服从 $\boldsymbol{U}(0, 1)$. 我们可以用 Monte carlo 方法检验该结论及 Y_t 的收敛速度。图2展示了 Y_t 的 25 条样本路径，每条路径持续 1000 步。可以看到每条路径都收敛了，但是收敛到不同的值。

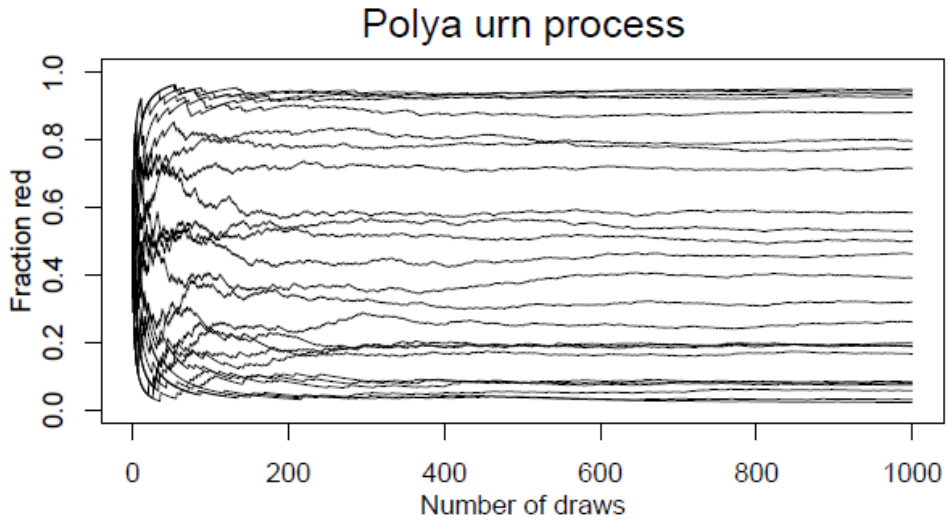

Figure 2: Pólya urn process 的 25 条样本路径。Picture source: Art B. Owen

- **习题**. 对 Pólya's urn process 稍做修改可以用来描述市场竞争中赢家通吃 (winner-take-all) 的现象。比如用 (R_t, B_t) 代表两家公司的用户数，即使它们提供的产品完全相同，如果新用户

倾向于购买他们朋友购买的产品，公司的用户增量 Z_t 可能服从如下分布：

$$Z_t = \begin{cases} (1,0), & \text{概率} = R_t^\alpha/(R_t^\alpha + B_t^\alpha) \\ (0,1), & \text{概率} = B_t^\alpha/(R_t^\alpha + B_t^\alpha) \end{cases} \tag{2}$$

其中 $\alpha > 1$. 这种情况下两家公司最终不会平分市场份额，而是由一家公司占领全部市场。最终的结果与早期的一些优势或运气有很大关系。选择不同的 $\alpha > 1$ 的值, 基于(2)生成若干条 $Y_t = R_t/(R_t + B_t)$ 的样本路径 (初始时刻 $R_0 = 1, B_0 = 1$)，观察是否出现赢家通吃的现象以及 α 的取值对路径收敛速度的影响。

3 Gaussian processes

Gaussian process 的任意有限维分布都是一个多元正态分布。由于多元正态分布只取决于期望和协方差矩阵，因此定义一个 Gaussian process $\{X(t) \mid t \in \mathcal{T}\}$ 只需要确定一个**期望函数**

$$\mu(t) = E\left[X(t)\right],\ t \in \mathcal{T}$$

和一个**协方差函数**

$$\Sigma(t,s) = \mathrm{Cov}\left(X(t), X(s)\right),\ \forall t, s \in \mathcal{T}.$$

显然协方差函数 $\Sigma(\cdot,\cdot)$ 需满足对称性 $\Sigma(t,s) = \Sigma(s,t)$. 此时 Gaussian process 的任意有限维分布可写为

$$\begin{pmatrix} X(t_1) \\ \vdots \\ X(t_m) \end{pmatrix} \sim N_m \left(\begin{pmatrix} \mu(t_1) \\ \vdots \\ \mu(t_m) \end{pmatrix}, \begin{pmatrix} \Sigma(t_1,t_1) & \cdots & \Sigma(t_1,t_m) \\ \vdots & \ddots & \vdots \\ \Sigma(t_m,t_1) & \cdots & \Sigma(t_m,t_m) \end{pmatrix} \right)$$

Gaussian process 的期望函数可以是任意函数 $\mu : \mathcal{T} \to \mathbb{R}$，而协方差函数还需要再满足一个限制条件：由于多元正态分布的协方差矩阵是 (半) 正定的，一个有效的协方差函数 $\Sigma(\cdot,\cdot)$ 需满足

$$\sum_{i=1}^{m}\sum_{j=1}^{m} x_i x_j \Sigma(t_i,t_j) \geq 0,\ \forall m \geq 1, t_i \in \mathcal{T}, x_i \in \mathbb{R}.$$

Gaussian process 的一个重要应用是为函数插值提供不确定性估计 (uncertainty quantification), 它也是 nonparametric Bayesian model 常用的 prior. 假设 $f(\cdot)$ 是 Gaussian process 的一条样本路径，且已知该 Gaussian process 的期望函数 $\mu(\cdot)$ 和协方差函数 $\Sigma(\cdot,\cdot)$, 我们对 $f(\cdot)$ 的任意有限维分布就有了先验信息 (prior information). 当观察到该样本路径上 k 个点的值 $f(t_1), \ldots, f(t_k)$ 后，利用 $(k+1)$ 维正态分布的条件分布公式，我们可以计算样本路径上任一点 $f(t)$ 的条件期望和方

差: $E\left(f(t) \mid f(t_1), \ldots, f(t_k)\right)$, $Var(f(t) \mid f(t_1), \ldots, f(t_k))$, 即 $f(t)$ 的 posterior mean 和 posterior variance.

用每一点的条件期望 (posterior mean) 定义一个预测函数:

$$\hat{f}(t) = E\left(f(t) \mid f(t_1), \ldots, f(t_k)\right),\ t \in \mathcal{T}.$$

显然 $\hat{f}(t_j) = f(t_j)$, $j = 1, \ldots, k$. 即 $\hat{f}$ 是对观察值的一个插值函数。同时我们知道 $f(t)$ 在每一点的条件方差, 因此可以给出 f 的置信区间 (由每一点的置信区间组成)。由于给定 $f(t_1), \ldots, f(t_k)$ 后, $f(t)$ 的条件分布 (posterior distribution) 也是一个正态分布, 因此可以在每一点对 $f(t)$ 抽样, 生成一条通过已知点的样本路径。

如果对任意间隔 Δ, $\forall t \in \mathcal{T}$, $X(t)$ 和 $X(t+\Delta)$ 都是同分布, 称随机过程 $X(t)$ 是平稳的 (stationary). 对于 Gaussian process, stationarity 等价于

$$\mu(t+\Delta) = \mu(t), \quad \Sigma(t+\Delta, s+\Delta) = \Sigma(t, s), \quad \forall \Delta, \forall t, s \in \mathcal{T}.$$

通常 $\mathcal{T}$ 包含 0, 因此对于 Gaussian process, stationarity 意味着

$$\mu(t) \equiv \mu(0), \quad \Sigma(t, s) = \Sigma(t-s, 0), \quad \forall t, s \in \mathcal{T}.$$

下面列举一些常见的 stationary Gaussian processes.

- **Exponential covariance**. Gaussian process with exponential covariance 的期望函数 $\mu(t) \equiv 0$, 协方差函数为

  $$\Sigma(t, s) = \sigma^2 \exp\left(-\theta|t-s|\right), \quad \theta > 0.$$

  该过程的样本路径是连续的但不可导。

- **Gaussian covariance**. Gaussian process with Gaussian covariance 的期望函数 $\mu(t) \equiv 0$, 协方差函数为

  $$\Sigma(t, s) = \sigma^2 \exp\left(-\theta(t-s)^2\right), \quad \theta > 0.$$

  Gaussian covariance 也被称为 squared exponential covariance. 它的样本路径是任意阶可导的。

  图3展示了分别使用 Gaussian process with exponential covariance 和 Gaussian covariance 对三个已知点 $f(0) = 1$, $f(0.4) = 3$ 及 $f(1) = 2$ 插值的结果。在画预测函数 $\hat{f}$ 时, 从指标集 $\mathcal{T}$

选取的节点为 -0.25 到 1.25 之间间隔为 0.01 的一列点。上述插值得到的预测函数 $\hat{f}$ 与 σ 的取值无关，因为 σ 在计算条件期望时被消掉了。但是 $\hat{f}$ 与 θ 有关：当 θ 很大时，节点之间的相关性随着距离 $|t-s|$ 增加迅速下降，观察点附近节点的预测值与观察值的相关性变得很小，节点的预测值被迅速拉向整体的期望函数 $\mu(t) \equiv 0$; 当 θ 较小时，节点之间的相关性随距离 $|t-s|$ 增加下降地较慢，观察点附近节点的预测值与观察值的相关性很高。可以看到 exponential covariance 在 θ 较小时的插值预测函数似乎是分段线性的 (piecewise linear), 而 Gaussian covariance 在 θ 较小时的预测函数非常平滑。

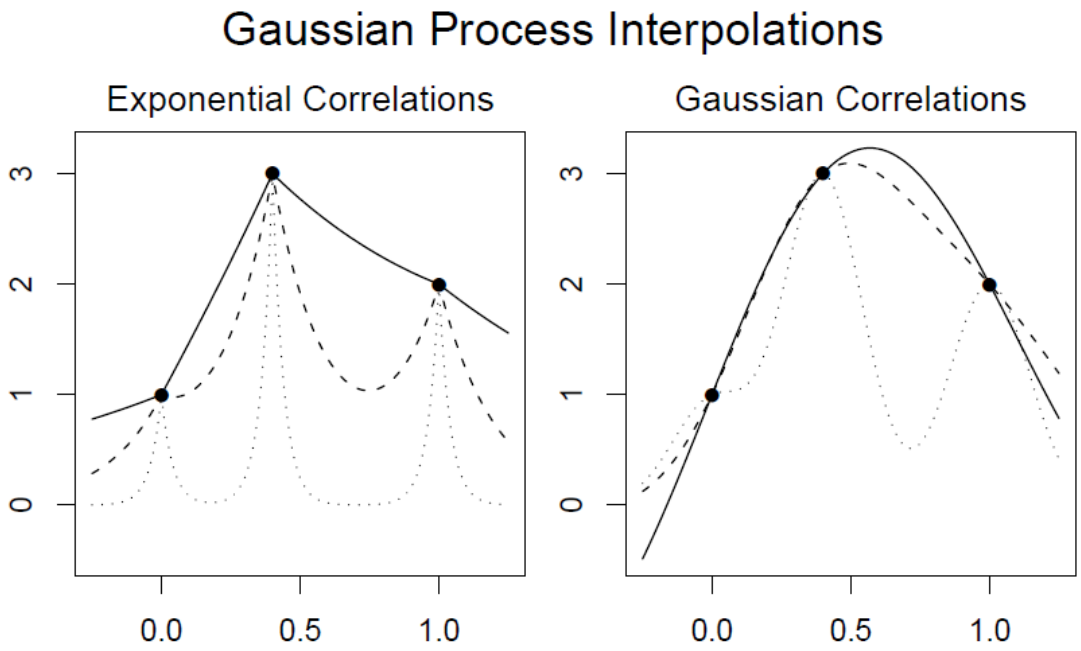

Figure 3: 使用 Gaussian process 对三个观察点插值。左图使用 exponential covariance, 右图使用 Gaussian covariance; 图中的实线，虚线，点线分别对应 $\theta = 1, 5, 25$. Picture source: Art B. Owen

图4展示了 Gaussian process with $\mu(t) = 0$, $\Sigma(t,s) = \exp(-(t-s)^2)$ 生成的若干条通过已知点 $f(0) = 1$, $f(0.4) = 3$ 和 $f(1) = 2$ 的样本路径。从这些模拟中，我们可以近似得到 $f(\cdot)$ 的最大值点 $t^\star$ 的 (posterior) 分布。

Gaussian covariance 产生的样本路径有时过于平滑，Matérn covariances 可以提供介于 exponential 和 Gaussian covariance 之间的平滑度。

- **Matérn covariances**. Matérn class of covariances 由一个平滑度 (smoothness) 系数 ν 控制。对一般的 $\nu > 0$, 协方差函数 $\Sigma(t,s;\nu)$ 通过 Bessel function 定义。当 $\nu = m + 1/2$ 且 m 是非

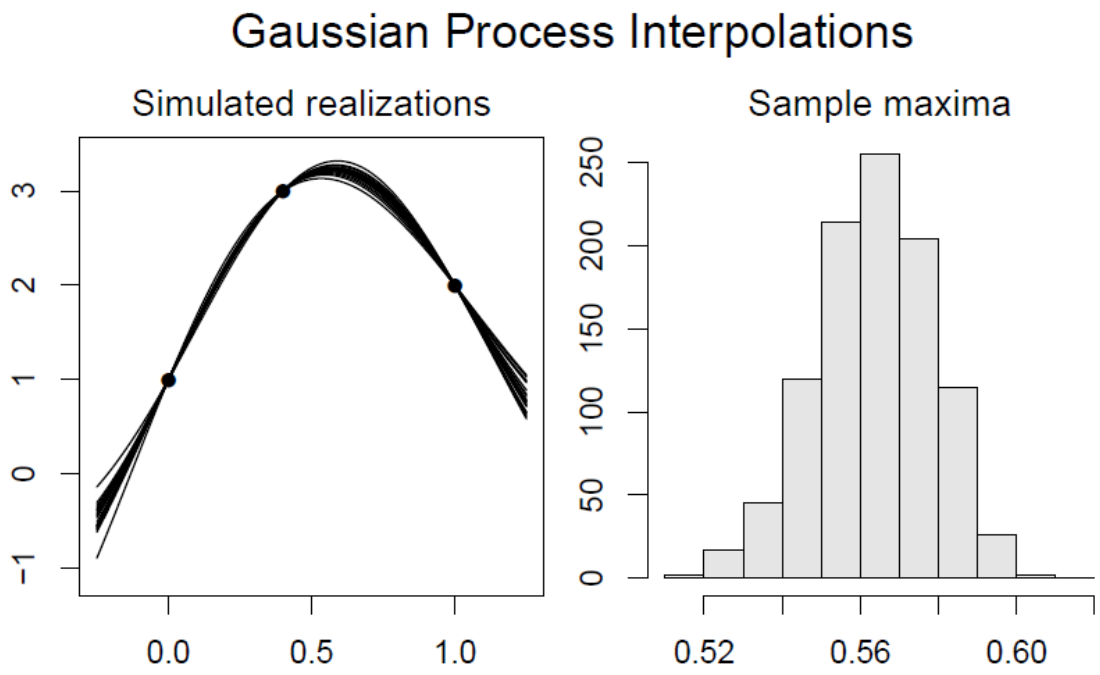

Figure 4: Gaussian process with Gaussian covariance 生成的 20 条通过 3 个已知点的样本路径，$\theta = 1$, $\sigma^2 = 1$ (左图). 1000 条通过左图方式生成的样本路径上最大值点的分布 (右图)。Picture source: Art B. Owen

负整数时, 协方差函数 $\Sigma(t, s; \nu)$ 可以极大简化，比如前 4 个特例为

$$\Sigma\left(t, s; \frac{1}{2}\right) = \sigma^2 \exp\left(-\theta|t-s|\right)$$
$$\Sigma\left(t, s; \frac{3}{2}\right) = \sigma^2 \left(1+\theta|t-s|\right) \exp\left(-\theta|t-s|\right)$$
$$\Sigma\left(t, s; \frac{5}{2}\right) = \sigma^2 \left(1+\theta|t-s|+\frac{1}{3}\theta^2|t-s|^2\right) \exp\left(-\theta|t-s|\right)$$
$$\Sigma\left(t, s; \frac{7}{2}\right) = \sigma^2 \left(1+\theta|t-s|+\frac{2}{5}\theta^2|t-s|^2+\frac{1}{15}\theta^3|t-s|^3\right) \exp\left(-\theta|t-s|\right)$$

其中 $\theta > 0$.

显然 exponential covariance 是 Matérn class 的一个特例 ($\nu = 1/2$). Matérn covariances 在 $\nu \to \infty$ 时收敛到 Gaussian covariance. Matérn covariance with $\nu = m + 1/2$ 生成的样本路径有 m 阶导数。图5展示了 Matérn process 生成的一些样本路径，可以看到 ν 越大，对应的样本路径越光滑；θ 越大，样本路径的振荡越多。

从 Gaussian process 生成一条有 m 个点的样本路径 f，等价于对一个 m 维的多元正态分

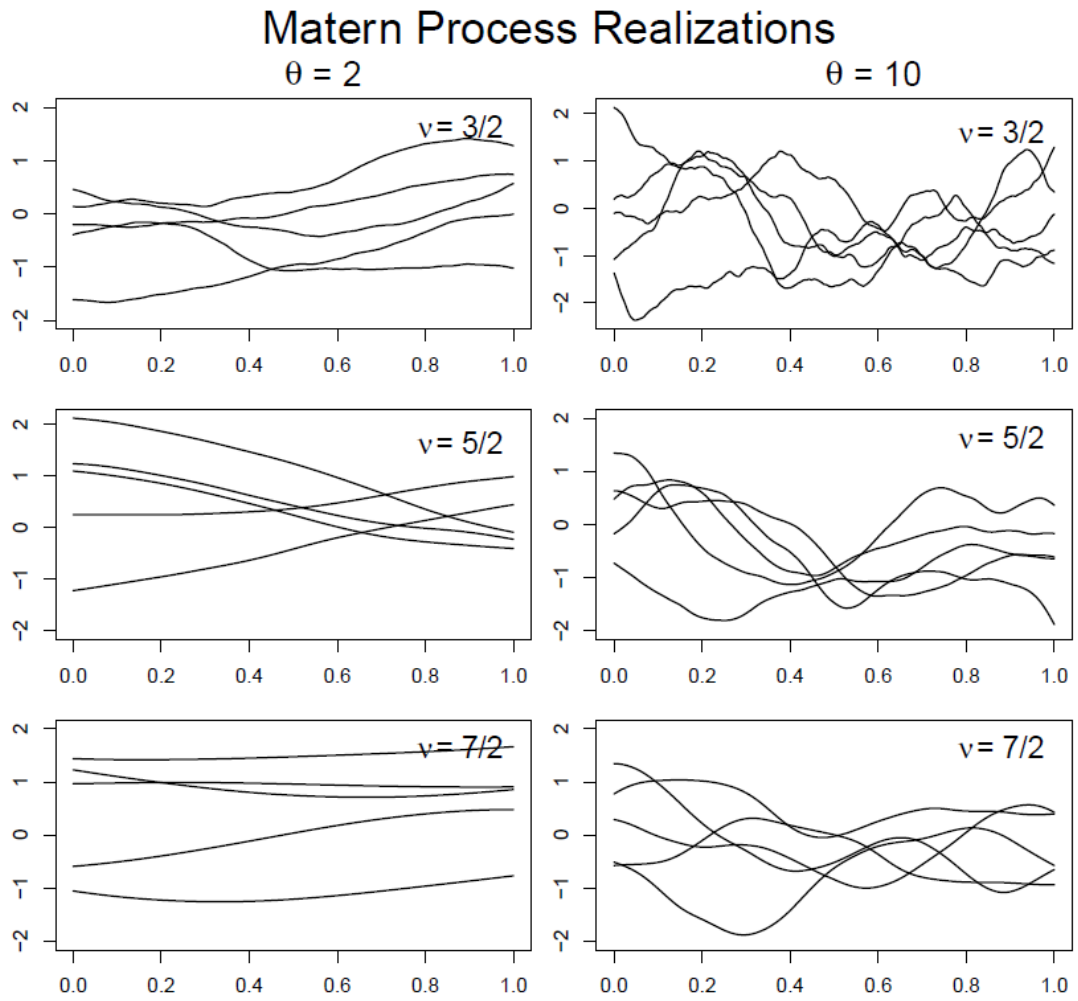

Figure 5: Matérn process 在 $\nu = 3/2, 5/2, 7/2$, $\theta = 2, 10$ 下分别产生的 5 条样本路径 ($\sigma^2 = 1$). Picture source: Art B. Owen

布进行一次抽样。取定 $t_1, \ldots, t_m$, 通过期望函数可以得到随机向量 $(f(t_1), \ldots, f(t_m))$ 的期望 $\boldsymbol{\mu}$, 通过协方差函数可以计算出它的协方差矩阵 Σ. 上一章我们介绍过从 $N_m(\boldsymbol{\mu}, \Sigma)$ 抽样需要计算一个矩阵 C 使得 $\Sigma = CC^\top$, 这一过程的计算量约为 $O(m^3)$. 抽样时先抽 $\boldsymbol{Z} \sim N_m(\mathbf{0}, I_m)$, 则 $\boldsymbol{\mu} + C\boldsymbol{Z} \sim N_m(\boldsymbol{\mu}, \Sigma)$. 如果抽样时使用 Gaussian covariance 且选取的 θ 很小, 会得到非常平滑的样本路径, 但对应的协方差矩阵 Σ 可能非常接近 singular (节点间相关系数几乎为 1), 此时推荐使用特征值分解计算 C, 即做分解 $\Sigma = P\Lambda P^\top$, 然后令 $C = P\Lambda^{1/2}$. 另一个办法是给 Σ 加一个 nugget effect, 即用 $\Sigma_\epsilon = \Sigma + \epsilon I_m$ 替代 Σ, 其中 ϵ 是很小的正数。如果 Σ 是一个有效的协方差矩阵 (半正定), Σ_ϵ 也是有效的协方差矩阵且可逆。此时相当于将模型修改为 $\Sigma_\epsilon(t, s) = Cov\left(X(t) + \varepsilon_t, X(s) + \varepsilon_s\right)$, 其中 ε_t's iid 服从 $N(0, \epsilon)$, 它们可以看作加在原过程 $X(t)$ 上的一些 “扰动”(jitter) 或测量误差。

3.1 Brownian motion

Brownian motion 可能是最重要的一个 Gaussian process, 本节我们讨论如何对 Brownian motion 抽样。Standard Brownian motion 是定义在 $\mathcal{T} = [0, \infty)$ 上的 Gaussian process, 记为 $B(t)$, 它有三条性质:

1. $B(0) = 0$.
2. 对于任意的 $0 = t_0 < t_1 < \cdots < t_m$, $B(t_i) - B(t_{i-1}) \overset{ind}{\sim} N(0, t_i - t_{i-1})$, $i = 1, \ldots, m$.
3. $B(t)$ 的样本路径在 $[0, \infty)$ 上以概率 1 连续。

Standard Brownian motion 也被称为 Wiener process, 以纪念数学家 Norbert Wiener, 他在 1923 年证明了满足上述 3 条性质的随机过程是存在的。尽管 $B(t)$ 的样本路径是连续的，但它也以概率 1 处处不可导。易证 $B(t)$ 的期望函数 $\mu(t) = 0$, 协方差函数 $\Sigma(t, s) = \min(t, s)$, 因此 Brownian motion 不是平稳的.

在 standard Brownian motion 的基础上，可以生成更复杂的 Brownian motions. 将 standard Brownian motion $B(t)$ 记为 $B(\cdot) \sim \text{BM(0,1)}$. 定义一个新的随机过程

$$X(t) = \delta t + \sigma B(t)$$

容易证明 $X(t)$ 的期望函数 $\mu(t) = \delta t$, 协方差函数 $\Sigma(t, s) = \sigma^2 \min(t, s)$. 我们称 $X(t)$ 是 drift δ, 方差 σ^2 的 Brownian motion, 记为 $X(\cdot) \sim \text{BM}(\delta, \sigma^2)$.

如果想得到 $X(\cdot) \sim \text{BM}(\delta, \sigma^2)$ 在 $[0, T]$ 上的样本路径，只需先抽取 $B(\cdot) \sim \text{BM(0,1)}$ 在 $[0, 1]$ 上的样本路径，然后令 $X(t) = \delta t + \sigma\sqrt{T} B(t/T)$ 即可. 因此我们只需关注如何对 standard Brownian motion 在 [0,1] 上抽样。对于 [0,1] 上的任意一列点，$0 < t_1 < t_2 < \cdots < t_m \leq 1$, 根据定义可以如下得到 $B(\cdot)$ 在这些点的样本:

$$\begin{aligned} B(t_1) &= \sqrt{t_1} Z_1, \\ B(t_j) &= B(t_{j-1}) + \sqrt{t_j - t_{j-1}} Z_j, \quad j = 2, \ldots, m \end{aligned} \tag{3}$$

其中 $Z_j \overset{ind}{\sim} N(0, 1)$, $j = 1, \ldots, m$. 也可以将上述过程写为矩阵形式

$$\begin{pmatrix} B(t_1) \\ B(t_2) \\ \vdots \\ B(t_m) \end{pmatrix} = \begin{pmatrix} \sqrt{t_1} & 0 & \cdots & 0 \\ \sqrt{t_1} & \sqrt{t_2 - t_1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{t_1} & \sqrt{t_2 - t_1} & \cdots & \sqrt{t_m - t_{m-1}} \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_m \end{pmatrix}. \tag{4}$$

等式(4)中的系数矩阵恰好是该随机向量协方差矩阵的 Cholesky 分解矩阵：

$$Var\left(\begin{pmatrix} B(t_1) \\ B(t_2) \\ \vdots \\ B(t_m) \end{pmatrix}\right) = (\min(t_j, t_k))_{1\le j,k\le m} = \begin{pmatrix} t_1 & t_1 & \cdots & t_1 \\ t_1 & t_2 & \cdots & t_2 \\ \vdots & \vdots & \ddots & \vdots \\ t_1 & t_2 & \cdots & t_m \end{pmatrix}.$$

3.2 Brownian bridge

对于 $[l, r]$ 上的 Brownian motion $B(t)$, 如果给定两端的值 $B(l)$ 和 $B(r)$, 称 $B(t)$ 在 $[l, r]$ 上的条件分布为一个 Brownian bridge.

Standard Brownian bridge 是 $[0, 1]$ 上给定 $B(0) = B(1) = 0$ 的 standard Brownian motion, 记为 BB(0,1). 对 Brownian bridge 抽样可以通过 Brownian motion 的路径得到。如果 $B(\cdot) \sim$ BM(0,1), 令

$$\tilde{B}(t) = B(t) - tB(1), \quad t \in [0, 1]$$

则 $\tilde{B}(\cdot) \sim$ BB(0,1). $\tilde{B}(t)$ 也是一个 Gaussian process, 期望函数 $\mu(t) = 0$, 协方差函数 $\Sigma(t, s) = \min(t, s)\,(1 - \max(t, s))$.

知道如何对 BB(0,1) 抽样就可以生成任意两点之间的一条 Brownian motion 的路径。比如，给定路径的起点 $B(l)$ 和终点 $B(r)$，可以按如下方式产生 BM(δ, σ^2) 在 $[l, r]$ 上的一条路径：

$$B(t) = B(l) + \frac{t-l}{r-l}\,(B(r) - B(l)) + \sigma\sqrt{r-l}\tilde{B}\left(\frac{t-l}{r-l}\right),\ l \le t \le r$$

其中 $\tilde{B}(\cdot) \sim$ BB(0,1). 注意该条件分布与 drift δ 无关。

3.3 Geometric Brownian motion

Brownian motion 最初是一个描述物体受周围粒子的碰撞在空间中做随机运动的模型。根据中心极限定理，很多次微小碰撞的累加效应会趋于一个正态分布。与之类似的一个过程是股票价格受到各种市场信息的影响不断波动。但股价变化经常被描述为一种相乘效应或对数尺度 (log scale) 上的累加效应，其极限分布是对数正态分布，对应的随机过程被称为 geometric Brownian motion.

用 S_t 表示某支股票在 t 时刻的价格。与股价的绝对变化相比，金融中人们更关心的是股票的收益率，即 S_t 在一个很小的区间 Δ 上的相对变化：

$$\frac{S_{t+\Delta} - S_t}{S_t} = \frac{\Delta S_t}{S_t} \approx \frac{dS_t}{S_t}.$$

经典的金融模型 (Hull, 2003) 这样描述 S_t 的相对变化：

$$\frac{dS_t}{S_t} = \delta dt + \sigma dB_t \tag{5}$$

其中 $B \sim \text{BM}(0,1)$. 该模型假设 S_t 在一个很小的区间 Δ 上的相对变化

$$\frac{\Delta S_t}{S_t} \sim N(\delta\Delta, \sigma^2\Delta).$$

注意，由于等式(5)的右边有一个随机微分项 dB_t，等式的左边 $\neq d\log(S_t)$. 常将(5)写为以下形式:

$$dS_t = \delta S_t dt + \sigma S_t dB_t. \tag{6}$$

股价的初始值 S_0 一般是给定的，我们称满足(6)的随机过程 S_t 是一个 **geometric Brownian motion**, 记为 $S \sim \text{GBM}(S_0, \delta, \sigma^2)$. 称 σ 为波动率 (volatility) 参数，δ 为 drift.

方程(6)是一个随机微分方程 (stochastic differential equation, SDE), 它是少数几个有解析解的 SDE. 其解的形式为

$$S_t = S_0 \exp\left\{(\delta - \sigma^2/2)t + \sigma B_t\right\} \tag{7}$$

其中 t 前的系数 $(\delta - \sigma^2/2)$ 是根据 **Itô's formula** 得到的。

Theorem 1 (Itô's formula)**.** 如果 $dS_t = a(S_t)dt + b(S_t)dB_t$ 且 $f(\cdot)$ 是一个二阶连续可导的函数，则

$$df(S_t) = \left(f'(S_t)a(S_t) + \frac{1}{2}f''(S_t)b^2(S_t)\right)dt + f'(S_t)b(S_t)dB_t.$$

令 $X_t = f(S_t) = \log(S_t)$, 根据 Itô's formula 和(6)可得

$$dX_t = \left(\delta - \frac{1}{2}\sigma^2\right)dt + \sigma dB_t. \tag{8}$$

为了对(8)求积分，将 $[0,t]$ 分成 N 个时间间隔为 $\Delta = t/N$ 的小区间，当 $N \to \infty$ 时，根据(8)，X_t 可写为如下增量累加的形式:

$$\begin{aligned}
X_t &= X_0 + \sum_{j=1}^{N} \left[X(j\Delta) - X\left((j-1)\Delta\right)\right] \\
&= X_0 + \sum_{j=1}^{N} \left[\left(\delta - \frac{1}{2}\sigma^2\right)\Delta + \sigma\left[B(j\Delta) - B\left((j-1)\Delta\right)\right]\right] \\
&= X_0 + \left(\delta - \frac{1}{2}\sigma^2\right)N\Delta + \sigma\left[B(N\Delta) - B(0)\right] \\
&= X_0 + \left(\delta - \frac{1}{2}\sigma^2\right)t + \sigma B_t
\end{aligned}$$

代入 $S_t = \exp(X_t)$ 可得(7).

基于 geometric Brownian motion 的 Monte Carlo 方法在路径依赖 (path dependent) 的金融期权定价中有广泛应用。期权是一种套期保值的金融工具，与某种资产挂钩；**路径依赖**是指期权的价格不仅取决于到期日资产的价格，还与到期日前的资产价格有关。

- **亚式看涨期权的定价**。航空公司最怕遇到油价大幅上涨。用 S_t 表示 t 时刻的油价，假设当前时刻的油价为 $S_0 = 1$. 如果价格 $S_t > 1.1$，航空公司就会面临亏损。有一种亚式看涨期权可以帮助航空公司对冲油价上涨的风险。如果航空公司购买了该期权，就会在一年之后收到以下金额

$$f\left(S(\cdot)\right) = \max\left(0, \frac{1}{12}(\sum_{j=1}^{12} S_{j/12}) - K\right).$$

即如果这一年 12 个月的平均油价高于 K (e.g. $K = 1.1$), 航空公司就会从期权中得到高出部分的补偿; 如果这 12 个月平均油价低于 K, 航空公司就不会得到补偿。该看涨期权可以保证航空公司的油价成本不超过 K，亚式不代表该期权只在亚洲出售，而是指期权到期日的收益与有效期内标的资产的平均价格有关，而不是只与某个时刻的价格有关。

那么这样一份期权的售价是多少呢? 理论上该期权在当前时刻的合理价格为 $e^{-rT}E\left(f(S)\right)$ (Hull, 2003), 其中 T 是距离到期日的时间，r 是无风险利率 (risk-free interest rate). 假设油价的波动 S_t 是一个 geometric Brownian motion, 则可以根据(7)生成大量 S_t 的样本路径, 每条路径都可以计算一个 f 的值。根据大数定律，这些 f 的独立观察值的平均值会收敛到 $E\left(f(S)\right)$.

4 Poisson point process

Point process 是指某个集合 $S \subset \mathbb{R}^d$ 内的一列随机点 $\{\boldsymbol{P}_1, \boldsymbol{P}_2, \dots\}$. S 通常被称为状态空间 (state space). 定义在 $S = [0, \infty)$ 上的一维 point process 可以用来描述随机来电的时间、网站访问高峰的时间、台风登陆的时间等；定义在二维或高维空间的 point process 可以描述地震的位置、森林中树的位置、星云中星系的位置等。

Point process 中点的个数可以是固定的或随机的、有限的或无限的 (countably infinite), 记为 $N(S)$. 对于集合 $A \subset S$, 用 $N(A)$ 表示落在 A 中的点的个数，即

$$N(A) = \sum_{i=1}^{N(S)} \mathbf{1}(\boldsymbol{P}_i \in A).$$

以下我们主要关注 non-explosive point processes 的抽样方法, 即对于任意体积 (volumn) 有限的集合 A, $P\left(N(A) < \infty\right) = 1$.

Point process 的有限维分布对应的是 S 上任意 J ($J \geq 1$) 个不相交的子集中点的个数的联合分布, 即 $(N(A_1), \dots, N(A_J))$ 的分布，其中 $A_1, \dots, A_J \subset S$ 且互不相交。

Definition 4.1 (Homogeneous Poisson process)**.** 如果对 S 上任意 J 个不相交的子集 $A_j \subset S$ 且

$\text{vol}(A_j) < \infty$, $j = 1 \ldots, J$, 点列 $\{\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots\}$ 满足

$$N(A_j) \overset{ind}{\sim} \text{Po}\left(\lambda \cdot \text{vol}(A_j)\right), \ j = 1 \ldots, J,$$

称该点列为 S 上一个强度 (intensity) 为 λ ($\lambda > 0$) 的 homogeneous Poisson process, 记为 $\{\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots\} \sim \text{PP}(S, \lambda)$.

现实世界中的很多 point processes 都不 homogeneous, 比如台风登陆会集中在一年的某一段时间，某些区域地震发生的频率很高。因此我们需要在 Poisson process 中加入一些非均匀性：将常数强度 λ 替换为一个随空间改变的强度函数 $\lambda(\boldsymbol{s}) \geq 0$, $\boldsymbol{s} \in S$. 一般要求该强度函数满足

$$\int_A \lambda(\boldsymbol{s}) d\boldsymbol{s} < \infty, \ \text{vol}(A) < \infty.$$

注意满足该条件的强度函数可以是无界的，比如 $\lambda(t) = t$, $t \in [0, \infty)$.

Definition 4.2 (Non-homogeneous Poisson process)**.** 如果对 S 上任意 J 个不相交的子集 $A_j \subset S$ 且 $\text{vol}(A_j) < \infty$, $j = 1 \ldots, J$, 点列 $\{\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots\}$ 满足

$$N(A_j) \overset{ind}{\sim} \text{Po}\left(\int_{A_j} \lambda(\boldsymbol{s}) d\boldsymbol{s}\right), \ j = 1 \ldots, J,$$

其中强度函数 $\lambda(\boldsymbol{s}) \geq 0$, 称该点列为 S 上的一个 non-homogeneous Poisson process, 记为 $\{\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots\} \sim \text{NHPP}(S, \lambda)$.

对 non-homogeneous Poisson process 抽样基于以下定理。

Theorem 2. *$\lambda(\boldsymbol{s}) \geq 0$ 是 S 上的一个强度函数且 $\Lambda(S) = \int_S \lambda(\boldsymbol{s}) d\boldsymbol{s} < \infty$. 如果 S 上的点列 $\{\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots\}$ 满足*

$$N(S) \sim Po(\Lambda(S)),$$

且给定 $N(S) = n$,

$$P(\boldsymbol{P}_i \in A) = \frac{1}{\Lambda(S)} \int_A \lambda(\boldsymbol{s}) d\boldsymbol{s}, \quad \forall A \subset S, \ i = 1, \ldots, n,$$

则点列 $\{\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots\} \sim NHPP(S, \lambda)$.

Proof. 令 $\Lambda(A) = \int_A \lambda(\boldsymbol{s}) d\boldsymbol{s}$, $\forall A \subset S$. 对 S 上任意不相交的 J ($J \geq 1$) 个子集 $A_1, \ldots, A_J$, 令

$$A_0 = \{\boldsymbol{s} \in S \mid \boldsymbol{s} \notin \cup_{j=1}^J A_j\}.$$

则对任意正整数 $n_j \geq 0$, $j = 1, \ldots, J$,

$$P^\star = P\left(N(A_1) = n_1, \ldots, N(A_J) = n_J\right) = \sum_{n_0=0}^{\infty} P\left(N(A_0) = n_0, N(A_1) = n_1, \ldots, N(A_J) = n_J\right). \tag{9}$$

令 $n = n_0 + n_1 + \cdots + n_J$. 根据定理条件，给定 n, 每个点 $\boldsymbol{P}_i$ 落在子集 A_j 内的概率为 $\Lambda(A_j)/\Lambda(S)$, $j = 0, 1, \ldots, J$. 因此这 n 个点在不相交的子集 $A_0, A_1, \ldots, A_J$ 中的分布是一个多项分布。所以

$$\begin{aligned} P\left(N(A_0) = n_0, N(A_1) = n_1, \ldots, N(A_J) = n_J\right) &= P\left(N(S) = n\right) \frac{n!}{n_0! n_1! \cdots n_J!} \prod_{j=0}^{J} \left(\frac{\Lambda(A_j)}{\Lambda(S)}\right)^{n_j} \\ &= \frac{\Lambda(S)^n e^{-\Lambda(S)}}{n!} \frac{n!}{\Lambda(S)^n} \prod_{j=0}^{J} \frac{\Lambda(A_j)^{n_j}}{n_j!} \\ &= \prod_{j=0}^{J} \frac{\Lambda(A_j)^{n_j} e^{-\Lambda(A_j)}}{n_j!}. \end{aligned}$$

代入(9)得

$$\begin{aligned} P^\star &= \sum_{n_0=0}^{\infty} \prod_{j=0}^{J} \frac{\Lambda(A_j)^{n_j} e^{-\Lambda(A_j)}}{n_j!} = \left(\sum_{n_0=0}^{\infty} \frac{\Lambda(A_0)^{n_0} e^{-\Lambda(A_0)}}{n_0!}\right) \prod_{j=1}^{J} \frac{\Lambda(A_j)^{n_j} e^{-\Lambda(A_j)}}{n_j!} \\ &= \prod_{j=1}^{J} \frac{\Lambda(A_j)^{n_j} e^{-\Lambda(A_j)}}{n_j!}. \end{aligned}$$

上式表明

$$N(A_j) \overset{ind}{\sim} \mathrm{Po}\left(\Lambda(A_j)\right), \ j = 1 \ldots, J.$$

根据定义，点列 $\{\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots\} \sim \mathrm{NHPP}(S, \lambda)$. □

Remark

1. 定理2表明，如果能从 PDF 为 $\rho(\boldsymbol{s}) \propto \lambda(\boldsymbol{s})$ 的分布抽样，就可以对强度为 $\lambda(\boldsymbol{s})$ 的 NHPP 抽样。

2. 如果 $\Lambda(S) = \infty$, 则无法使用定理2对 NHPP 抽样，因为 Monte Carlo 方法不能产生无限个点 ($N(S) \sim \mathrm{Po}(\infty)$). 实践中为保证 $\Lambda(S) < \infty$，一般将 S 选为一个很大的有界集合，基本覆盖我们感兴趣的区域。

3. 定理2允许 S 是一个无界的集合，只要满足 $\Lambda(S) < \infty$.

如果能在 S 上均匀取点，就可以使用定理2的如下推论对 S 上的 homogeneous Poisson process 抽样。

Corollary 1. *对于 S 上的一个强度为 λ 的 homogeneous Poisson process, 如果 $vol(S) < \infty$, 可以如下对其抽样：首先抽取*

$$N(S) \sim Po\left(\lambda \cdot vol(S)\right),$$

然后在 S 上独立均匀地抽取 $N(S)$ 个点 $\boldsymbol{P}_i \sim \boldsymbol{U}(S)$, $i = 1, \ldots, N(S)$.

Proof. 令定理2中的 $\lambda(\boldsymbol{s}) \equiv \lambda$, 此时

$$P(\boldsymbol{P}_i \in A) = \frac{1}{\Lambda(S)} \int_A \lambda d\boldsymbol{s} = \frac{\mathrm{vol}(A)}{\mathrm{vol}(S)}, \quad \forall A \subset S,$$

因此 $\boldsymbol{P}_i \sim \boldsymbol{U}(S)$, $i = 1, \ldots, N(S)$. □

练习. 如何在圆盘 $D = \{\boldsymbol{x} \in \mathbb{R}^2 \mid \boldsymbol{x}^\top \boldsymbol{x} \leq 1\}$ 上抽取一列点 $\{\boldsymbol{P}_1, \boldsymbol{P}_2, \ldots\} \sim PP(D, \lambda)$?

4.1 $[0, \infty)$ 上的 Poisson process

Poisson process 的很多应用都是描述事件发生的时刻，因此本节专门讨论状态空间为 $\mathcal{T} = [0, \infty)$ 上的 Poisson process. 以下我们假设 $\mathcal{T}$ 上的点列 (事件发生的时刻) 是按顺序产生的: $T_1 < T_2 < \cdots$.

为研究该点列的性质，定义如下计数函数 (counting function)

$$N(t) \equiv N([0, t]) = \sum_{i=1}^{\infty} \mathbf{1}\left(T_i \leq t\right), \quad 0 \leq t < \infty.$$

$\mathcal{T} = [0, \infty)$ 上的 homogeneous Poisson process 具有以下三条性质:

1. $N(0) = 0$.
2. $N(t) - N(s) \sim \mathrm{Po}\left(\lambda(t - s)\right)$, $0 \leq s < t$.
3. 增量独立: 对任意的 $0 = t_0 < t_1 < \cdots < t_m$, $N(t_i) - N(t_{i-1})$, $i = 1, \ldots, m$ 是独立的。

其中增量 $N(t) - N(s)$ 代表点列落在区间 $(s, t]$ 上的个数。将满足上述三条性质的点列记为 $\{T_1, T_2, \ldots\} \sim \mathrm{PP}([0, \infty), \lambda)$, 或简记为 $\mathrm{PP}(\lambda)$. 参数 λ 被称为该过程的速率 (rate) 或频率 (单位时间内出现的点数)。

点列 $\{T_1, T_2, \ldots\} \sim \mathrm{PP}(\lambda)$ 还有另一重要特性:

$$T_i - T_{i-1} \overset{iid}{\sim} \mathrm{Exp}(\lambda), \ i \geq 1, \ T_0 = 0. \tag{10}$$

即相邻点(事件)的时间间隔服从指数分布 $\mathrm{Exp}(\lambda)$ 或 $\mathrm{Exp}(1)/\lambda$, 它的期望是 $1/\lambda$. 严格的证明过程见 (Hoel, 1971). 我们可以简单验证一下: 如果 $T_i - T_{i-1} \sim \mathrm{Exp}(\lambda)$, 则 $P\left(T_i - T_{i-1} > x\right) = \exp(-\lambda x)$; 如果 $T_i - T_{i-1} > x$, 说明区间 $(T_{i-1}, T_{i-1} + x)$ 上没有点出现，对于 $\mathrm{PP}(\lambda)$ 的一个长度为 x 的区间，没有点出现的概率为 $P\left(\mathrm{Po}(\lambda x) = 0\right) = \exp(-\lambda x)$, 结果相符。

根据(10), 可以如下产生 $\mathrm{PP}(\lambda)$ 的点列:

$$T_0 = 0, \ T_i = T_{i-1} + E_i/\lambda, \ i \geq 1 \tag{11}$$

其中 $E_i \overset{iid}{\sim} \text{Exp}(1)$. 该方法被称为 **exponential spacings method**. 实际抽样时可以不断运行(11)直到出现的点数达到目标值或者点发生的时刻超过了窗口期 $[0,T]$.

如果只需要在一个有界区间 $[0,T]$ 上对 PP(λ) 抽样，Corollary 1提供了一个更简单的抽样方法：

$$\begin{aligned} N &= N(T) \sim \text{Po}(\lambda T) \\ S_i &\sim \boldsymbol{U}[0,T],\ i = 1,\dots N \\ T_i &= S_{(i)},\ i = 1,\dots,N. \end{aligned} \tag{12}$$

(12)的最后一步是将 $[0,T]$ 上均匀分布的点列 $\{S_i\}$ 从小到大排序再输出。

现实生活中，non-homogeneous 的现象很多，我们可以类似定义 $\mathcal{T} = [0,\infty)$ 上的 non-homogeneous Poisson process, 它具备以下三条性质：

1. $N(0) = 0$.
2. $N(t) - N(s) \sim \text{Po}\left(\int_s^t \lambda(x)dx\right),\ 0 \le s < t$.
3. $N(t)$ 的增量独立。

其中强度函数 $\lambda(x) \ge 0$ 且 $\int_s^t \lambda(x)dx < \infty$, $0 \le s < t < \infty$. 将满足上述三条性质的点列记为 $\{T_1, T_2, \dots\} \sim \text{NHPP}([0,\infty), \lambda)$, 或简记为 NHPP($\lambda$).

为方便对 NHPP(λ) 抽样，定义如下的 cumulative rate function:

$$\Lambda(t) = \int_0^t \lambda(s)ds.$$

假设 $\lambda(t) > 0$, $\forall t$, 则 $y = \Lambda(t)$ 严格单调递增，因此有逆函数 $t = \Lambda^{-1}(y)$.

对于点列 $\{T_1, T_2, \dots\} \sim \text{NHPP}(\lambda)$, 定义随机变量 $Y_i = \Lambda(T_i)$ 及如下的计数函数

$$M(y) = \sum_{i=1}^{\infty} \mathbf{1}\left(Y_i \le y\right) = \sum_{i=1}^{\infty} \mathbf{1}\left(T_i \le \Lambda^{-1}(y)\right) = N\left(\Lambda^{-1}(y)\right).$$

进一步研究函数 $M(y)$ 的性质。首先，注意到 $\Lambda(0) = 0$, 因此 $\Lambda^{-1}(0) = 0$, 则 $M(0) = N(0) = 0$.

其次

$$\begin{aligned} M(y) - M(x) = N\left(\Lambda^{-1}(y)\right) - N\left(\Lambda^{-1}(x)\right) &\sim \text{Po}\left(\int_{\Lambda^{-1}(x)}^{\Lambda^{-1}(y)} \lambda(t)dt\right) \\ \Leftrightarrow \text{Po}\left(\Lambda\left(\Lambda^{-1}(y)\right) - \Lambda\left(\Lambda^{-1}(x)\right)\right) &\Leftrightarrow \text{Po}(y-x), \quad 0 \le x < y. \end{aligned}$$

最后 $M(y)$ 的增量 $M(y_i) - M(y_{i-1}) = N\left(\Lambda^{-1}(y_i)\right) - N\left(\Lambda^{-1}(y_{i-1})\right)$, 显然也是独立的。所以我们证明了

$$Y_i = \Lambda(T_i) \sim \text{PP}(1).$$

因此，可以如下从 NHPP(λ) 中抽取点列 $T_1, T_2, \ldots$

$$\begin{aligned} Y_i &= Y_{i-1} + E_i,\ i = 1, 2, \ldots \\ T_i &= \Lambda^{-1}(Y_i) \end{aligned} \tag{13}$$

其中 $E_i \overset{iid}{\sim} \text{Exp}(1)$, $Y_0 = 0$. 该方法被称为 **non-homogeneous exponential spacings algorithm**.

Remarks.

1. 虽然我们在推导算法(13)时假设 $\Lambda(t)$ 严格递增，在实践中可以放宽这个要求，与计算 CDF 的逆函数类似，如果 $y = \Lambda(t)$ 有一些跳跃或者在某些区间是常数，可以使用如下的广义逆函数

$$\Lambda^{-1}(y) = \inf\{t \geq 0 \mid \Lambda(t) \geq y\}.$$

   如果 Λ 或 Λ^{-1} 没有解析形式，可以考虑使用数值方法逼近。

2. 如果 $\lim\limits_{t\to\infty} \Lambda(t) = \infty$, 算法(13)可以一直运行下去。如果 $\lim\limits_{t\to\infty} \Lambda(t) = M < \infty$, 当 $y > M$ 时，$\Lambda^{-1}(y)$ 不存在；这种情况下如果(13)运行到某一步 j 出现 $Y_j = Y_{j-1} + E_j > M$, 则点 T_j 无法产生，算法停止，仅输出 $(j-1)$ 个点。

5 Dirichlet process

Dirichlet process (DP) 描述的是分布的分布，它的每一条样本路径都是一个分布。DP 的有限维分布是一个 Dirichlet 分布。Dirichlet process 在 nonparametric Bayesian model 中有广泛应用，常被用作一个未知分布的 prior. DP prior 的共轭性 (conjugacy) 使得计算后验分布变得容易。有了 Dirichlet process, 我们可以将有限维的混合分布 (finite-component mixture model) 推广到无限维 (infinite-component mixture model), 即 Dirichlet process mixture model.

令 F 是 $\Omega \subset \mathbb{R}^d$ 上一个分布的 CDF，对 Ω 做一个分割 (partition): $\Omega = A_1 \cup A_2 \cup \cdots \cup A_m$, 其中 $A_i \cap A_j = \emptyset$, $i \neq j$. 这个分割定义了 unit simplex Δ^{m-1} 上的一个向量：

$$(F(A_1), \ldots, F(A_m)) \in \Delta^{m-1} \equiv \left\{ (p_1, \ldots, p_m) \mid p_j \geq 0, \sum_{j=1}^{m} p_j = 1 \right\} \tag{14}$$

其中 $F(A_j) = P(X \in A_j \mid X \sim F)$, $j = 1, \ldots, m$.

如果分布 F 是随机的，(14)中的向量 $(F(A_1),\ldots,F(A_m))$ 是 Δ^{m-1} 上的一个随机点。如果随机向量 $(F(A_1),\ldots,F(A_m))$ 服从 Dirichlet 分布，如何保证给 Ω 的任意有限分割分配的 Dirichlet 分布都是一致的 (coherent)? 比如，对 Ω 的两种分割: $\Omega = A_1^{(1)} \cup A_2^{(1)} \cup \cdots \cup A_m^{(1)}$ 和 $\Omega = A_1^{(2)} \cup A_2^{(2)} \cup \cdots \cup A_m^{(2)}$，假设 $(F(A_1^{(j)}),\ldots,F(A_m^{(j)})) \sim \mathrm{Dir}(\alpha_1^{(j)},\ldots,\alpha_m^{(j)})$, $j = 1, 2$. 如果 $A_1^{(1)} \subset A_1^{(2)}$，如何保证这两个 Dirichlet 分布中的参数也能体现这种关系? 为了消除分配的 Dirichlet 分布对不同分割的敏感性，我们需要定义一个新的分布描述 F 的概率在整个 Ω 上是如何分布的，这引出了 Dirchlet process.

Dirichlet process 是通过一个常数 $\alpha > 0$ 和 Ω 上的一个确定的分布 G (CDF) 定义的，它的任意有限维分布对应 Ω 的一个有限分割，且满足

$$(F(A_1),\ldots,F(A_m)) \sim \mathrm{Dir}\left(\alpha G(A_1),\ldots,\alpha G(A_m)\right).$$

一般将 Dirichlet process 记为 $F \sim \mathrm{DP}(\alpha, G)$, 或简记为 $\mathrm{DP}(\alpha G)$. 此时随机分布 F 的期望是 G，α 决定了 F 到 G 的平均距离。简单验证如下: 由于 Dirichlet 随机向量的每个元素的边际分布是一个 Beta 分布，即

$$F(A_j) \sim \mathrm{Beta}\left(\alpha G(A_j), \alpha\left(1 - G(A_j)\right)\right)$$

所以

$$\begin{aligned} E\left(F(A_j)\right) &= G(A_j) \\ Var\left(F(A_j)\right) &= \frac{G(A_j)\left(1 - G(A_j)\right)}{\alpha + 1}. \end{aligned}$$

可见 α 越大，F 的分布越集中在 G 附近。

Dirichlet process 常作为 nonparametric Bayesian model 的 prior. 这类模型假设观察值 $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ 独立地来自一个未知的分布 F, F 的 prior 为 $F \sim \mathrm{DP}(\alpha, G)$, 然后推断给定数据 $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ 下 F 的 posterior 分布。以 $n = 1$ 为例，令 $(A_1,\ldots,A_m)$ 是 Ω 的一个分割。给定分布 F, $\boldsymbol{X_1}$ 落在集合 A_j 的概率为 $F(A_j)$, $j = 1,\ldots,m$. 没有观察到数据前，预期 $\boldsymbol{X_1}$ 落在各集合的概率服从分布 (prior):

$$(F(A_1),\ldots,F(A_m)) \sim \mathrm{Dir}\left(\alpha G(A_1),\ldots,\alpha G(A_m)\right).$$

观察到 $\boldsymbol{x}_1 \in A_k$ 后，$(F(A_1),\ldots,F(A_m))$ 的 posterior PDF 为

$$\begin{aligned} p\left(F(A_1),\ldots,F(A_m) \mid \boldsymbol{x}_1 \in A_k\right) &\propto p\left(F(A_1),\ldots,F(A_m)\right) \cdot P\left(\boldsymbol{x}_1 \in A_k \mid F(A_1),\ldots,F(A_m)\right) \\ &\propto \left(\prod_{j=1}^{m} F(A_j)^{\alpha G(A_j)-1}\right) \cdot F(A_k) \end{aligned}$$

$$\propto \left(\prod_{j\neq k} F(A_j)^{\alpha G(A_j)-1}\right)\cdot F(A_k)^{\alpha G(A_k)}$$

因此给定 $\boldsymbol{x}_1 \in A_k$, $(F(A_1),\ldots,F(A_m))$ 的 posterior 为

$$\mathrm{Dir}\left(\alpha G(A_1),\ldots,\alpha G(A_{k-1}),\alpha G(A_k)+1,\alpha G(A_{k+1}),\ldots,\alpha G(A_m)\right). \tag{15}$$

(15)对 Ω 的任意分割都成立, 因此 F 的 posterior 也是一个 Dirichlet process:

$$F \mid \boldsymbol{x}_1 \sim \mathrm{DP}\left(\alpha G+\delta_{\boldsymbol{x}_1}\right)$$

其中 $\delta_{\boldsymbol{x}_1}$ 是一个退化 (degenerate) 分布的 CDF, 该分布的所有概率都集中在点 $\boldsymbol{x}_1$ 处, 此时 $\delta_{\boldsymbol{x}_1}(A)=\mathbf{1}(\boldsymbol{x}_1\in A)$.

依此类推，当有 n 个观察值 $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ 时，F 的 posterior 分布为

$$F \mid \boldsymbol{x}_1,\ldots,\boldsymbol{x}_n \sim \mathrm{DP}\left(\alpha G+\sum_{i=1}^{n}\delta_{\boldsymbol{x}_i}\right).$$

5.1 Stick-breaking process

从 Dirichlet process 的定义我们并不知道该如何对 $\mathrm{DP}(\alpha,G)$ 抽样，特别是对 $\mathrm{DP}(\alpha,G)$ 的一次抽样得到的是一个分布。Sethuraman (1994) 给出了一种直接建立 DP 样本的方法，称为 **stick-breaking construction**.

$F\sim\mathrm{DP}(\alpha,G)$ 可写为以下形式，称为 DP 的 stick-breaking representation:

$$F=\sum_{j=1}^{\infty}\pi_j\delta_{\boldsymbol{X}_j},\ \pi_j=\theta_j\prod_{i<j}(1-\theta_i),\ \theta_i\overset{iid}{\sim}\mathrm{Beta}(1,\alpha),\ \boldsymbol{X}_j\overset{iid}{\sim}G. \tag{16}$$

(16)相当于先从基准分布 (base) G 中 iid 抽取一列点 $\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots$, 然后给每个点 $\boldsymbol{X}_j$ 分配一个概率 π_j. 这些概率从一个 **stick-breaking process** 中产生以保证 $\sum_{j=1}^{\infty}\pi_j=1$. 从 DP 的 stick-breaking representation (16)可以发现：如果 $F\sim\mathrm{DP}(\alpha,G)$, 则 F 一定是一个离散分布，不可能是连续分布。因此在 Bayesian 模型中，DP 不适合作为一个连续分布的 prior.

Stick-breaking process 的抽样过程可以类比如下：开始时有一根长度为 1 的棍子 (stick), 代表要分配给所有 $\{\boldsymbol{X}_j\}_{j=1}^{\infty}$ 的总概率为 1. 首先抽取 $\theta_1\sim\mathrm{Beta}(1,\alpha)$, 将棍子去掉长度 θ_1, 代表分配给 $\boldsymbol{X}_1$ 的概率 $\pi_1=\theta_1$. 此时棍子剩下的长度为 $1-\theta_1$, 接着抽取 $\theta_2\sim\mathrm{Beta}(1,\alpha)$, 再从剩下的棍子中去掉 θ_2 比例，即长度 $\theta_2(1-\theta_1)$, 代表分配给 $\boldsymbol{X}_2$ 的概率 $\pi_2=\theta_2(1-\theta_1)$. 对随后的每个 $\boldsymbol{X}_j$ $(j\geq 3)$, 我们都从剩下的棍子中去掉一个比例 $\theta_j\sim\mathrm{Beta}(1,\alpha)$, 去掉的长度代表分配给 $\boldsymbol{X}_j$ 的概率 π_j.

由于

$$E(\theta_j)=\frac{1}{1+\alpha},\ \theta_j\sim\mathrm{Beta}(1,\alpha)$$

如果 α 很小 (接近 0), 上述过程会倾向于给排在前面的点分配较大的概率，后面的点只能分到很小的概率，如图6所示。图6还表明分配给点 $\{\boldsymbol{X}_j\}$ 的概率 $\{\pi_j\}$ 与基准分布 G 在这些点的概率密度无关，但 G 会影响这些点的位置。

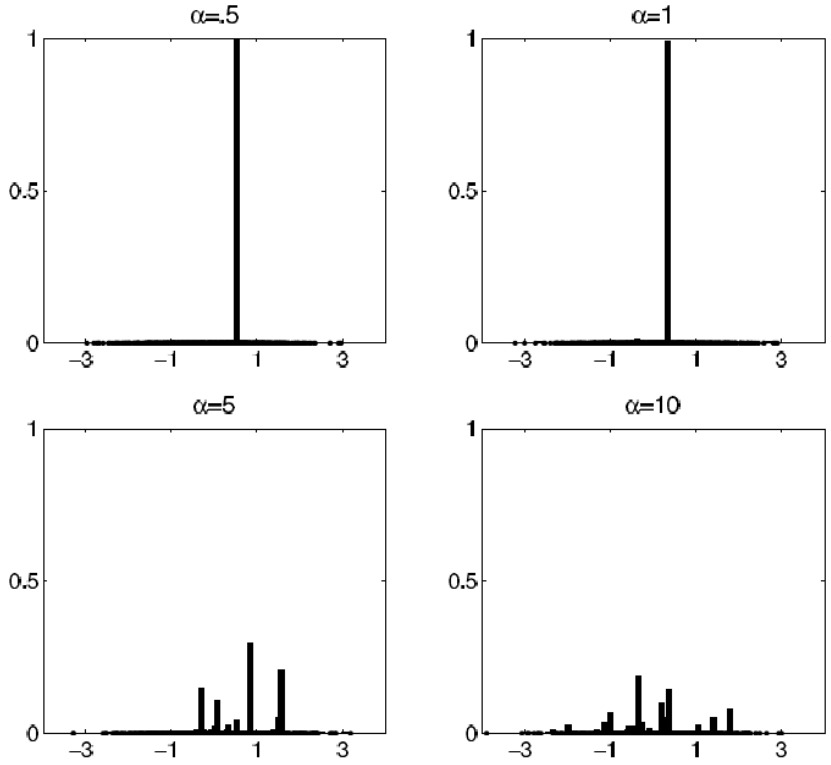

Figure 6: 使用 stick-breaking representation 对 $\mathrm{DP}(\alpha, G)$ 抽样，在不同 α 取值下得到的样本分布。其中 G 是 $N(0,1)$, 图中纵坐标表示点的概率。Picture source: Gelman et al. (2013).

5.2 Chinese restaurant process

本节介绍另一种从 Dirichlet process 抽样的方法。对于任意分布 F, 如果能获取 F 的大量样本 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ (n 很大), 就可以用它们的 empirical distribution

$$F_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\boldsymbol{x}_i} \tag{17}$$

近似 F.

考虑对以下两阶段 (two-stage) 模型抽样:

$$\begin{aligned} F &\sim \mathrm{DP}(\alpha, G) \\ \boldsymbol{X}_i &\sim F,\ i = 1, \ldots, n. \end{aligned} \tag{18}$$

我们先考虑 $n=1$ 的抽样。如果 $F \sim \mathrm{DP}(\alpha, G)$ 且 $\boldsymbol{X}_1 \sim F$, 那么 $\boldsymbol{X}_1$ 的边际分布是什么? 注意到 $\forall A \subset \Omega$ (Ω 是 DP 的 support)

$$P(\boldsymbol{X}_1 \in A) = E[\mathbf{1}(\boldsymbol{X}_1 \in A)] = E[E[\mathbf{1}(\boldsymbol{X}_1 \in A) \mid F]] = E[F(A)] = G(A).$$

因此 $\forall \alpha > 0$, $\boldsymbol{X}_1$ 的边际分布都是 G. 此时我们不需要先从 DP 产生 F, 可以直接抽取 $\boldsymbol{X}_1 \sim G$.

$n \geq 2$ 时, 可以用序列条件分步法产生 $\boldsymbol{X}_i$, $i \geq 2$. 即依次抽取 $\boldsymbol{X}_i \sim F \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{i-1}$. 如何对该条件分布抽样? 我们从 Bayesian 角度考虑(18), 将 $\mathrm{DP}(\alpha, G)$ 作为 F 的 prior, 则观察到 $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{i-1}$ 后, F 的 posterior 为 $\mathrm{DP}\left(\alpha G + \sum_{j=1}^{i-1} \delta_{\boldsymbol{X}_j}\right)$, 即

$$F \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{i-1} \sim \mathrm{DP}\left(\alpha + i - 1, \frac{\alpha G + \sum_{j=1}^{i-1} \delta_{\boldsymbol{X}_j}}{\alpha + i - 1}\right). \tag{19}$$

因此观察到 $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{i-1}$ 后, 将 F 的分布更新为(19). 此时 $\boldsymbol{X}_i \sim F \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{i-1}$ 的边际分布 (积掉 F 的不确定性) 为

$$\boldsymbol{X}_i \sim (\alpha G + \sum_{j=1}^{i-1} \delta_{\boldsymbol{X}_j})/(\alpha + i - 1). \tag{20}$$

注意到(20)中的分布是一个混合分布, 相当于 $\boldsymbol{X}_i$ 以概率 $\alpha/(\alpha+i-1)$ 来自分布 G, 以概率 $1/(\alpha+i-1)$ 重复之前抽取的样本 $\boldsymbol{X}_j$, $j = 1, \ldots, i-1$. 即(20)中的抽样可如下进行:

$$\boldsymbol{X}_i = \begin{cases} \boldsymbol{Y} \sim G & \text{概率 } \alpha/(\alpha+i-1) \\ \boldsymbol{X}_1 & \text{概率 } 1/(\alpha+i-1) \\ \vdots & \vdots \\ \boldsymbol{X}_{i-1} & \text{概率 } 1/(\alpha+i-1). \end{cases} \tag{21}$$

由(21)生成的随机过程被称为 **Chinese restaurant process**, 因为样本的产生过程可以类比如下: 顾客 $i = 1, 2, \ldots$ 依次到达一个中餐馆, 顾客 1 从分布 G 中抽到桌 $\boldsymbol{X}_1$; 顾客 2 以概率 $\alpha/(\alpha+1)$ 开一个新桌 $\boldsymbol{Y} \sim G$, 或者以概率 $1/(\alpha+1)$ 加入顾客 1 所在的桌 $\boldsymbol{X}_1$; 依次类推, 顾客 i 以概率 $\alpha/(\alpha+i-1)$ 开一个新桌, 或者随机选一个之前到达的顾客, 加入他所在的桌 (每桌可坐的人数无限制)。显然, 越多顾客就坐的桌有更大的概率吸引新顾客的加入, 如图7所示。

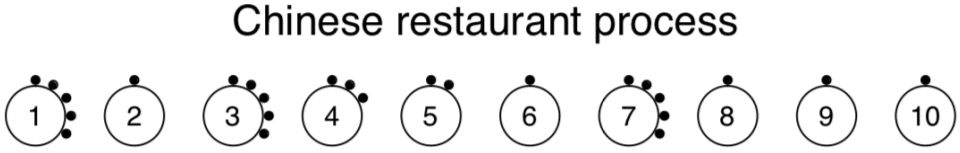

Figure 7: Chinese restaurant process ($\alpha = 4$) 的一次抽样: 前 25 名到达顾客选择桌子的情况, 一共坐了 10 桌。Picture source: Art B. Owen

Remarks

1. α 越小, 新顾客选择开新桌的概率就越小, 这一过程产生的桌数就越少。CRP 有点类似 Pólya urn process, 它们都是强化型的随机游走，不同的是 CRP 总有可能产生新桌，而 Pólya urn process 只能在两种球色中选取。

2. 在 CRP 中，第 n 个顾客到达时期望开设的桌数是

$$\sum_{i=1}^{n} \frac{\alpha}{\alpha+i-1} \leq 1+\int_0^n \frac{1}{1+x/\alpha}dx = 1+\alpha\log(1+n/\alpha) \sim O(\log n)$$

   在一些应用中我们希望桌数的增长比 $O(\log n)$ 快，Pitman-Yor process (Pitman and Yor, 1997) 可以让桌数以 $O(n^\beta)$ $(0<\beta<1)$ 增长。

3. 从 CRP (21) 产生的一条样本路径 $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ (n 很大) 的 empirical distribution(17)可近似看作从 DP(α, G) 中随机生成的一个分布。

5.3 Dirichlet process mixture model

在 CRP 模型(18)中再加一层就得到了 **Dirichlet process (DP) mixture model**:

$$\begin{aligned} F &\sim \mathrm{DP}(\alpha, G) \\ \boldsymbol{X}_1,\ldots,\boldsymbol{X}_n &\sim F \\ \boldsymbol{Y}_i &\overset{ind}{\sim} H(\cdot \mid \boldsymbol{X}_i),\ i=1,\ldots,n. \end{aligned} \tag{22}$$

其中 $\{\boldsymbol{Y}_i\}_{i=1}^n$ 是观察值，$\{\boldsymbol{X}_i\}_{i=1}^n$ 和 F 是待估计的参数。

我们用一个简单的例子考察从模型(22)生成的数据 $\{\boldsymbol{Y}_i\}_{i=1}^n$ 的特点。将 base G 选为 $N_2(\boldsymbol{0}, \sigma_0^2 I)$, 令 $\boldsymbol{Y}_i \overset{ind}{\sim} N_2(\boldsymbol{X}_i, \sigma_1^2 I)$. 由于 $\{\boldsymbol{X}_i\}_{i=1}^n$ 来自一个 CRP, 它们中会出现重复的值，这些重复的值会使 $\{\boldsymbol{Y}_i\}_{i=1}^n$ 出现聚集效应，形成若干 clusters, 如图8所示。图8展示了 α 取 3 个不同值时产生的三组样本，其中 $\sigma_0=3$, $\sigma_1=0.4$. 这些样本 $\{\boldsymbol{y}_i\}$ 形成的 clusters 通常对应 $\{\boldsymbol{X}_i\}$ 中多次重复出现的值，有时也可能是因为很多 $\boldsymbol{X}_i$ 的取值非常接近；样本中那些远离 clusters 的 outliers 通常对应 $\{\boldsymbol{X}_i\}$ 中出现次数很少的值，可能是只出现过一次的值。在 CRP 中，α 越小，顾客选择开新桌的概率就越小，产生的 $\{\boldsymbol{X}_i\}$ 越容易出现重复值，这与图8展示的情况一致: α 越小，样本 $\{\boldsymbol{y}_i\}$ 的聚集效应越明显。

Remarks

1. CRP 模型产生重复值的特点使它很适合描述有 clusters 特征的数据, 而且 DP mixture model (22)不需要提前设定 clusters 的个数或者给这个数目设一个上限。在给定观察值 $\{\boldsymbol{y}_i\}_{i=1}^n$ 后，

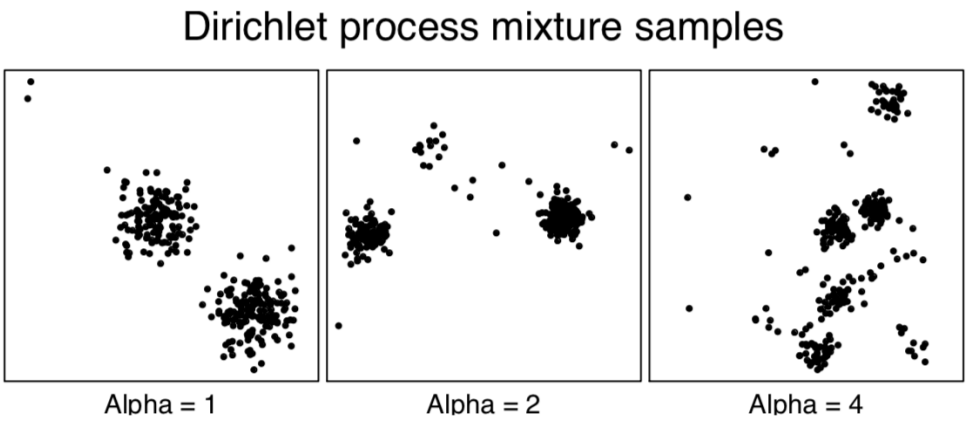

Figure 8: DP mixture model 生成的样本 $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{200}$, 从左到右依次对应 $\alpha = 1, 2$ 和 4. Picture source: Art B. Owen

一般使用 Markov chain Monte Carlo (MCMC) 方法估计模型(22)中 clusters 的位置 ($\{\boldsymbol{X}_i\}$ 取到的不同值) 和个数 ($\{\boldsymbol{X}_i\}$ 取到不同值的个数), 实现以数据驱动 (data-driven) 方式估计 clusters 的信息。

2. DP mixture model 允许 clusters 的个数是无限的，这并不代表我们观察到的有限数据是由无限个 clusters 产生的，而是使模型具有很大的灵活度，可以随着新观察值的加入不断引入新的 clusters.

3. 如果一些先验信息告诉我们数据对应的 clusters 较少，可以在 DP prior 中选取较小的 α, 或者再给 α 加一个 gamma hyperprior 以增加对数据的适应性 (data-adaptivity).

References

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Hoel, P. G. (1971). *Introduction to statistical theory*. Houghton Mifflin, Boston.

Hull, J. C. (2003). *Options futures and other derivatives*. Pearson Education India.

Pitman, J. and Yor, M. (1997). The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650.

