doi: 10.11845/sxjz.2016006a

Dirichlet 过程及其研究进展

夏业茂 [1,2,*], 刘应安 [1]

(1. 南京林业大学理学院, 南京, 江苏, 210037; 2. 南京林业大学经济与管理学院, 南京, 江苏, 210037)

摘要: 非参数贝叶斯分析主要是将兴趣参数或潜变量的分布视为随机的并赋予一个先验分布. 作为分布函数的分布, Dirichlet 过程是目前非参数贝叶斯分析中最受欢迎的先验分布, 并受到广泛的关注. 本文对近几十年来 Dirichlet 过程的发展作了一下回顾和总结, 并就 Dirichlet 过程在潜变量模型中的应用做了介绍.

关键词: 非参数贝叶斯; Dirichlet 过程; Pólya urn 抽样; Sethurmann 表示; Dirichlet 混合过程; 相依 Dirichlet 过程; 马尔可夫链蒙特卡洛; 分块 Gibbs 抽样器

MR(2010) 主题分类: 62F15; 62F35 / **中图分类号:** O212.8; O212.4

文献标识码: A **文章编号:** 1000-0917(2017)05-0641-26

0 引言

考虑如下贝叶斯推断问题:

$$\begin{aligned} &Y_1, Y_2, \cdots, Y_n|\theta, \text{ iid.} \sim g(y|\theta), \\ &\theta \sim P_0, \end{aligned}$$

其中 $Y_1, Y_2, \cdots, Y_n$ 为观测样本, $g(y|\theta)$ 为 Y_i 的抽样分布, θ 是未知随机参数, P_0 为 θ 的一个确定的先验分布. 现在的问题是如何选择一个恰当的 P_0, 使得基于观测样本 $\{Y_1, Y_2, \cdots, Y_n\}$ 对 θ 的后验推断具有一定的稳健性. 从贝叶斯角度, 一个自然的做法是视 P_0 为未知且随机的, 依据一定的 "先验", 从分布函数空间 (集合) 提取 P, 即

$$\begin{aligned} &Y_1, Y_2, \cdots, Y_n|\theta_1, \theta_2, \cdots, \theta_n, \text{ ind.} \sim g(y_i|\theta_i), \\ &\theta_1, \theta_2, \cdots, \theta_n|P, \text{ iid.} \sim P, \\ &P \sim \mathscr{P}, \end{aligned} \tag{1}$$

其中 $\mathscr{P}$ 为 P 的一个先验或概率分布, 即所谓的 "分布的分布". 无疑, (1) 的做法类似于频率统计的非参数统计问题, 故该类问题又称为非参数或半参数贝叶斯问题. 如果 $\mathscr{P}$ 将概率集中于某个特定的 P_0, 即 $\mathscr{P} = \delta_{P_0}$, 那么 (1) 退化为参数贝叶斯问题. 对 $\mathscr{P}$ 的一般性要求是其支撑集足够大, 能够尽量选取足够多的分布函数, 同时其后验分布又便于处理. 1973 年, Ferguson 在《一些非参数问题的贝叶斯分析》一文中, 从随机过程角度对 P 建立 Dirichlet 过程 (DP)[28], 以其诱导的分布作为 P 先验 (称之为 DP 先验). 这种先验不仅支撑集足够大, 而且后验具有良好的共轭性. 自此, DP 作为一种分布函数的先验, 在贝叶斯框架内得到了快速的展开和应用. 应该说,

收稿日期: 2016-10-27. 修改稿收到日期: 2016-12-26.
基金项目: 国家自然科学基金 (No. 11471161) 和南京市科技创新择优资助项目 (No. 013101001).
E-mail: * ymxia@njfu.edu.cn

它是目前应用最为充分和成功的一种非参数贝叶斯先验. 特别是近 20 年来, 由于计算技术的改进和计算水平的提高, 以及贝叶斯方法的重新兴起, 对 DP 的研究和应用达到了前所未有的广度和深度.

DP 的研究和发展大体上经历三个阶段: 一是 1973 年至 1995 年. 这个阶段对 DP 的研究主要集中于理论性质的探索, 譬如样本的抽样性质、大样本理论等方面. Ferguson[28−29], Blackwell 和 MacQueen[5−6], Antoniak[2], Korwar 和 Hollander[54], Lo[63], Doss[16−17], Pitman[80−81], Sethuraman[87−88] 等在这方面作出了重要贡献. 一些重要的计算方法如独立抽样法、重要抽样法 [18]、序贯抽样法 [53, 55, 67] 等开始应用于贝叶斯后验推断; 二是 1995 年至 1999 年, 这个阶段的主要工作是集中于统计建模和算法的设计和改进. 伴随马尔可夫链蒙特卡洛 (MCMC) 技术在统计领域的兴起, 基于 DP 的非参数贝叶斯分析得到了充分的展开和广泛的应用. 这主要得益于 Escobar[23−24], Escobar 和 West[25], West 等 [98], MacEachern[64], Bush 和 MacEachern[8], Neal[77], Dalal[14] 等人的努力和工作. 第三阶段是 1999 年至目前, 这个阶段主要集中于多重相依 DP 的研究和应用. 相依 DP 研究可以追溯到 Cifarelli 和 Regazzini[11] 的工作, 但真正完全展开应该得益于 MacEachern[65−66] 的努力. 与单个过程相比, 多重 (个) DP 过程注重对一族兴趣分布进行 DP 拟合, 而且更多是关注这些 DP 的内在关联. 这对于更加细致和深入地洞察观测数据的概率发生机制是十分必要的.

本文主要对 DP 先验作一个回顾和梳理: 从过程的产生、其抽样性质、DP 各种延伸、有关 DP 过程的计算方法以及在潜变量模型中的应用来进行归纳, 力求展现其发展轨迹和进一步应用前景. 我们侧重于 DP 的近十几年来的最新发展. 本文的符号记述如下: 对于一个概率分布函数 F, 我们往往也用它来表示对应的概率分布测度; δ_x 表示概率集中于 x 的 Dirac 测度; 记 P 的样本空间为 $\mathcal{X}$; 除非特别说明即为有限维欧氏空间. 为了突出主要结果, 我们以定理的形式将其表现出来.

1 Dirichlet 过程 (DP)

如果视 (1) 中的 P 为随机化分布函数, 那么 P 等同于一个取值于 $[0,1]$ 空间上的随机过程. 显然, 如果样本空间 $\mathcal{X}=\{x_1,x_2,\cdots,x_m\}$ 为有限集, 那么其上的任意概率分布 P 完全由 $\mathbf{p}=(p_1,p_2,\cdots,p_m)=(P(\{x_1\}),P(\{x_2\}),\cdots,P(\{x_m\}))$ 来确定. 此时, P 的先验分布等价于随机向量 $\mathbf{p}$ 的先验分布. 对 $\mathbf{p}$ 的先验的一个自然选择是有限维 Dirichlet 分布:

$$\mathbf{p}\sim \mathrm{Dir}_m(\alpha_1,\alpha_2,\cdots,\alpha_m)\quad(\alpha_j>0).$$

其结果是, P 完全由 $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_m)$ 及有限维 Dirichlet 分布来确定. 如果 $\mathcal{X}$ 可列或不可数, 那么指定 $\mathscr{P}$ 取决于其有限维分布族的形式. 记 $\mathscr{B}$ 为 $\mathcal{X}$ 的 σ- 域. $\alpha(\cdot)$ 为 $(\mathcal{X},\mathscr{B})$ 上一个有限非零的有限可加测度. 称随机过程 $\{P(A):A\in\mathscr{B}\}$ 为一 Dirichlet 过程 (记为 $P\sim D(\alpha)$), 若对于 $\mathcal{X}$ 的任意有限可测划分 $\{A_1,A_2,\cdots,A_m\}$ $(m\in\mathbb{N})$,

$$(P(A_1),P(A_2),\cdots,P(A_m))\sim \mathrm{Dir}_m(\alpha(A_1),\alpha(A_2),\cdots,\alpha(A_m)). \tag{2}$$

如果令 $c=\alpha(\mathcal{X})$ (>0), $F_0(\cdot)=\frac{\alpha(\cdot)}{\alpha(\mathcal{X})}$, 则也记为 $P\sim D(cF_0)$. 显然, (2) 指明了 DP 先验的有限维分布族. Ferguson[28] 证明了在一定正则条件下, $(\mathcal{X},\mathscr{B})$ 上所有概率分布测度集上存在一个概率分布测度 $\mathscr{P}$, 使得其有限维恰好是上述 (2) 的有限维分布族. 因此, 本质上说, $D(cF_0)$ 是无限维 Dirichlet 分布 (见 [28], 可列情形参见定理 2.9).

容易证明: 如果 $P\sim D(cF_0)$, 则对于任何 $A,B\in\mathscr{B}$,

$$\mathbb{E}[P(A)]=F_0(A),\quad \mathbb{D}[P(A)]=\frac{F_0(A)(1-F_0(A))}{c+1},$$
$$\mathrm{Cov}(P(A),P(B))=\frac{F_0(AB)-F_0(A)F_0(B)}{c+1}.$$

这表明, 在 DP 先验下, P 的中心位置位于 F_0. 因此, F_0 可以看作是对 P 的事先 "猜测" 或认识, 又称为底分布; 而 c 反映了 P 围绕其中心 F_0 的波动程度 (c 称为精度或浓度参数): 当 $c\to+\infty$ 时, P 退化为 F_0.

选择 $D(cF_0)$ 作为 P 的先验, 一个巨大优势就是其具有共轭性.

定理 1.1[28] 设 $P\sim D(cF_0)$, 给定 P, $X_1,X_2,\cdots,X_n$ 为来自 P 的独立同分布样本, 则 P 的后验分布为

$$P|X_1,X_2,\cdots,X_n\sim D(c^*F_0^*),\tag{3}$$

其中

$$c^*(\,\cdot\,)=c+n,\quad F_0^*(\,\cdot\,)=\frac{c}{c+n}F_0(\,\cdot\,)+\frac{n}{c+n}F_n(\,\cdot\,),$$

F_n 为 $X_1,X_2,\cdots,X_n$ 的经验分布函数.

令 $F(x)=P((-\infty,x])$ 为 P 对应的分布函数. 定理 1.1 直接给出了均方损失下 F 的后验估计: $\forall x\in\mathcal{X}$,

$$\widehat{F}(x)=\mathbb{E}[F(x)|X_1,X_2,\cdots,X_n]=\frac{c}{c+n}F_0(x)+\frac{n}{c+n}F_n(x).\tag{4}$$

对 (4) 的一种直观解释是: 当 $n\to\infty$ 时, $\widehat{F}(x)$ 由 $F_n(x)$ 主导, 而后者是 (依据一定的拓扑) 收敛于 X_i 的真实分布; 另外当 $c\to\infty$ 时, $\widehat{F}(x)=F_0$, 这与前面讨论 $\mathscr{P}$ 为退化时的情形是一致的; 而 $c\to0$ 时, DP 先验退化为无信息先验, 此时 $\widehat{F}(x)=F_n(x)$, 纳入经典的频率非参数分析框架内. 形成定理即是:

定理 1.2[37] 设 $X_1,X_2,\cdots,X_n|P\sim P$, 且 $P\sim D(cF_0)$. 若 X_i 的真实分布为 G_0, 则

1) 若 $n\to\infty$, 则 $D(c^*F_0^*)\to_w\delta_{G_0}$,

2) 若 $c\to0$, 则 $D(c^*F_0^*)\to_w D(F_n)$.

在有限样本情形下, c 的选择必须慎重, 因为它直接关系到 $\widehat{F}$ 的光滑程度. 对此有各种不同的建议. 例如, Escobar[24] 建议将 c 随机化并赋以均匀离散分布 $U(n^{-1},n^0,n,n^2)$; Ishwaran 和 Zarepour[48] 强调给予 Gamma 分布; Kleinman 和 Ibrahim[51] 建议作出敏感度分析; 宋心远等 [92] 建议从模型选择角度利用信息准则来确定 c 的值.

如果将 $(\mathcal{X},\mathcal{B})$ 上所有概率分布测度看做一个集合 $\mathcal{M}$, 那么 Dirichlet 过程所诱导的先验的支撑集有多大呢? Ferguson[28] 指出, 在逐点收敛意义下, $D(cF_0)$ 的支撑集为所有关于 F_0 绝对连续 (被 F_0 控制) 的概率分布测度所形成的集合. 特别地, 当 $(\mathcal{X},\mathscr{B})$ 为欧氏空间且集合 $\mathcal{M}$ 的拓扑由分布函数弱收敛来刻画时, 则有 $\mathrm{supp}(D(cF_0)=\{Q\in\mathcal{M}:\mathrm{supp}(Q)\subseteq\mathrm{supp}(F_0)\}$. 因此, 当 $\mathrm{supp}(F_0)=\mathbb{R}^d$ 时, $D(cF_0)$ 支撑集为整个 $\mathcal{M}$. 这表明, 以 DP 先验作为未知分布函数的先验是恰当的.

尽管如此, 但 DP 依然展现出不足的一面. 事实上, DP 先验下的 P 的实现值是离散概率分布测度 (见本文的定理 1.3 和定理 1.4). 因此可知, 直接用 DP 来拟合一个绝对连续分布函数是

不合适的. 然而这种离散特性却可以用来挖掘样本中潜藏的 “类”, 这对于解释样本的异质性提供了一种新的角度和方法.

1.1 DP 的两种表示

众所周知, 有限维 Dirichlet 分布可以表示成若干个独立的 Gamma 随机变量与它们和的比. 即, 若 $Y_j \sim \mathrm{Gamma}(\alpha_j, 1)$ $(j = 1, 2, \cdots, m)$ 且独立 (α_j 为形状参数), 则

$$\frac{(Y_1, Y_2, \cdots, Y_m)}{\sum_{j=1}^{m} Y_j} \sim \mathrm{Dir}_m(\alpha_1, \alpha_2, \cdots, \alpha_m).$$

无限维 DP 先验也有类似的特性: 可视为独立增量的 Gamma 过程与其相应的 “和” 的比. 具体地说, 令

$$N(x) = -c \int_x^{\infty} \exp(-y) y^{-1} \mathrm{d}y \quad (0 < x < \infty). \tag{5}$$

定义随机序列 $\{J_k : k = 1, 2, \cdots\}$:

$$\mathbb{P}(J_1 \le x_1) = \exp\{N(x_1)\},$$
$$\mathbb{P}(J_n \le x_n | J_{n-1} = x_{n-1}, \cdots, J_1 = x_1) = \exp(N(x_n) - N(x_{n-1})) \quad (0 < x_n < x_{n-1}).$$

显然, $J_1 > J_2 > \cdots$, a.s.

定理 1.3[28] 在以上记号下,

$$P(\,\cdot\,) = \sum_{k=1}^{\infty} q_k \delta_{Z_k}(\,\cdot\,) \sim D(cF_0), \tag{6}$$

其中 Z_k, iid. 服从 $F_0(\,\cdot\,)$, q_k 为随机权满足 $q_k = \frac{J_k}{\sum_{k=1}^{\infty} J_k}$.

如果令 $F(t)$ 为 P 对应的分布函数, 定理 1.3 无疑是说 $F(t) = \frac{Z_t}{Z_\infty} \sim D(cF_0)$, 其中 Z_t 为独立增量的 Gamma 过程: $Z_t \sim \mathrm{Gamma}(cF_0(t), 1)$, $Z_\infty = \lim_{t\to\infty} Z_t \sim \mathrm{Gamma}(c, 1)$. 这恰好是有限维 Dirichlet 分布的无限维化版本.

Sethuraman[87] 给出了 DP 先验另外一种富有创造性的构造, 文献中称为 stick-breaking 构成†. 这种表示法对 DP 先验的研究和应用, 特别是对 DP 的延伸起到了不可替代的作用.

定理 1.4 (Sethuraman[87]) 设 $\{Z_m\}_{m=1}^{\infty}, \text{iid.} \sim F_0$, $\{V_m\}_{m=1}^{\infty}, \text{iid.} \sim \mathrm{Beta}(1, c)$, 且两者独立. 令

$$\pi_1 = V_1, \quad \pi_m = V_m \prod_{l=1}^{m-1} (1 - V_l), \quad \cdots. \tag{7}$$

则

$$P(\,\cdot\,) = \sum_{m=1}^{\infty} \pi_m \delta_{Z_m}(\,\cdot\,) \sim D(cF_0). \tag{8}$$

随机权 (7) 有一个形象的比喻: 若将 V_m 看作许多筷子, 那么随机权 π_m 相当于将首尾相连的筷子从某处折断. 不难发现, $\forall N \in \mathbb{N}$,

$$1 - \sum_{m=1}^{N} \pi_m = \prod_{m=1}^{N} (1 - V_m).$$

† 严格地说, 这种方式可以追溯到 Halmos[40], Freedman[31], Fabius[27], Connor 和 Mosimann[12], Kingman[50] 等人的工作.

根据三级数收敛性定理, 容易证明 $\sum_{m=1}^{\infty}\pi_m=1,$ a.s.

随机权 (7) 和 (8) 的构成区别在于前者的随机权按照递减的方式排列, 而后者却不然. 构成 (8) 中 $\{q_m\}$ 的分布被称为带有参数 c 的 Poisson-Dirichlet 分布 (见 Kingman [50]), 而构成 (9) 中随机权的分布由 Ewens [26] 以 Griffiths, Engen 和 McCloskey 的姓进行命名: $\pi=(\pi_k)\sim$GEM(c). 尽管两者形式不同, Pitman[81] 基于容量置换方式 (size-based permutation, 见 McCloskey [70]) 建立了两者的联系, 详细情形可进一步参见 Pitman [81].

无论何种方式, P 的先验分布完全由随机权集 $\{q_m\}$ 或 $\{V_m\}$ 和原子集 $\{Z_m\}$ 的联合分布来确定. 这种无限维性对 P 的有关泛函性质如分位数的分析、后验抽样造成不便. 一种策略是采用有限维分布近似, 见 Sethuraman 和 Tiwari [88], Doss [18], Muliere 和 Tardella [74], Gelfand 和 Kottas [32], Ishwaran 和 James [44], Ishwaran 和 Zarepour [48] 等.

由 DP 的表示法立即可得 P 的泛函性质:

定理 1.5 (Ferguson[28]) 设 $P\sim D(cF_0)$, 且 $\psi(x),\varphi(x)$ 为 $(\mathcal{X},\mathscr{B})$ 上 F_0- 可积函数, 则

1) $\psi(P)=\int_{\mathcal{X}}\psi(x)P(\mathrm{d}x)<\infty,$ a.s.;

2) $\mathbb{E}[\psi(P)]=\mathbb{E}_{F_0}\psi(X)$;

3) $\mathbb{D}[\psi(P)]=\frac{\mathbb{D}_{F_0}\psi(X)}{c+1}$;

4) $\mathrm{Cov}(\psi(P),\varphi(P))=\frac{\mathrm{Cov}_{F_0}(\psi(X),\varphi(X))}{c+1}$.

1.2 抽样性质

DP 的样本性质, 是指来自随机概率分布测度 P 的随机样本的边际分布 (对 P 积分) 性质. 由于在 DP 先验下, P 的实现值为离散分布, 这导致来自 P 的样本中存在 "结", 其联合分布测度关于相应的勒贝格测度奇异, 致使样本的联合概率密度或质量函数退化.

定理 1.6 (Polya prediction) 若 $P\sim D(cF_0)$, 且 $X_1,X_2,\cdots,X_n|P,$ iid. $\sim P$, 则

$$\begin{cases}X_1\sim F_0,\\ X_2|X_1\sim\dfrac{cF_0(\,\cdot\,)+\delta_{X_1}(\,\cdot\,)}{c+1},\\ \vdots\\ X_n|X_1,X_2,\cdots,X_{n-1}\sim\dfrac{cF_0+\sum_{i=1}^{n-1}\delta_{X_i}}{c+n-1}.\end{cases}\tag{9}$$

Blackwell 和 MacQueen[6] 称序列 (9) 为 Polya urn 序列. 它类似于概率论熟知的 Polya urn 模型. 进一步地, MacQueen 和 Blackwell[5] 证明了如果一个序列 $\{X_n\}$ 是 Polya urn 序列, 那么

$$\frac{cF_0(\,\cdot\,)+\sum_{i=1}^{n-1}\delta_{X_i}(\,\cdot\,)}{c+n-1}\overset{n\to\infty}{\longrightarrow}P(\,\cdot\,),\quad\text{a.s.},$$

且 P 是个 DP. 这无疑是从另外一个角度刻画 DP.

式 (9) 确定的 $\{X_1,X_2,\cdots,X_n,\cdots\}$ 具有可交换性, 因此,

$$X_i|\{X_j,j\neq i\}\sim\frac{cF_0+\sum_{k=1:k\neq i}^{n}\delta_{X_k}}{c+n-1}.\tag{10}$$

样本的条件分布 (10) 在非参数贝叶斯分析的蒙特卡洛抽样特别是 MCMC 抽样中起着关键性作用, 见 Kuo[55], Liu[62], Escobar[24], Ecosbar 和 West[25] 等人的工作.

定理 1.6 可以用所谓的 “中餐馆规则” (Chinese restaurant rule, 见 Aldous [1], Ishwaran 等 [45]) 来描述: 假设一个餐馆可以摆放无限多张不同的圆桌, 且每张桌子容纳的人数不限. 现在顾客 (X_i) 序贯进入. 第一个客人 (X_1) 随机地分配一张圆桌, 第二个客人 (X_2) 以概率 $\frac{1}{c+1}$ 坐到一个客人的圆桌, 而以概率 $\frac{c}{c+1}$ 重新分配一桌, 如此下去. 不同圆桌代表 $\{X_1, X_2, \cdots, X_n\}$ 中的不同值, 同一个圆桌内 X_i 是相同的. 这种比喻有助于了解样本的 “结”: 一个圆桌代表一个类 (cluster).

定义随机变量 $\{D_n : n \in \mathbb{N}\}$: $D_1 = 1$; 对 $i = 2, 3, \cdots$,

$$D_i = \begin{cases} 0, & \text{如果 } X_i = X_j \text{ 对某个 } j \ (j = 1, 2, \cdots, i-1); \\ 1, & \text{否则.} \end{cases}$$

显然, D_i 是用来标示分配新桌子的情况. 令 $D(n) = \sum_{i=1}^{n} D_i$ 表示前 n 个观测值 $\{X_1, X_2, \cdots, X_n\}$ 中不同值的个数 (坐有客人的桌子数). 记 $X_1^*, X_2^*, \cdots, X_{D(n)}^*$ 表示 $X_1, X_2, \cdots, X_n$ 不同值之集 (有客人的桌子).

定理 1.7 (Korwar 和 Hollander[54]) 在上述记号下, 假设 F_0 不含有原子, 则

(i)

$$\mathbb{P}(D_i = 1) = \frac{c}{c+i-1}, \quad i = 1, 2, \cdots, n,$$

且 D_i 相互独立;

(ii) $\mathbb{P}(D_n = 1, \text{i.o.}) = 1$, 且 $D(n) \to \infty$, a.s., $\frac{D(n)}{\log(n)} \to c$, a.s.;

(iii) 给定 $D(n)$, $X_1^*, X_2^*, \cdots, X_{D(n)}^* \overset{\text{iid}}{\sim} F_0$,

(iv)

$$\frac{\sum_{i=1}^{D(n)} X_i^*}{D(n)} \to \mu_0 = E_{F_0}(X), \quad \text{a.s.;}$$

(v) (Antoniak[2])

$$\mathbb{P}(D(n) = k | c, n) = \frac{{}_n a_k c^k}{c^{[n]}}, \tag{11}$$

其中 ${}_n a_k$ 为 $A_n(x) = x(x+1)\cdots(x+n-1) = {}_n a_1 x + {}_n a_2 x^2 + \cdots + {}_n a_n x^n$ 中 x^k 的系数 $({}_n a_1 = (n-1)!, \ {}_n a_n = 1)$,

$$c^{[n]} = \frac{\Gamma(c+n)}{\Gamma(c)} = \begin{cases} c(c+1)\cdots(c+n-1), & n = 1, 2, \cdots; \\ 1, & n = 0. \end{cases}$$

该定理指出, 在 DP 先验下, 尽管 P 是离散分布, 但随着样本量趋于无穷, 样本中不同值的个数 (坐有客人的桌子数, 也就是类数) 趋于无穷 (尽管速度很慢). 另外, 该定理也揭示出

$$\mathbb{E}D(n) = \sum_{i=1}^{n} \frac{c}{c+i-1} \approx c \log\left(\frac{n+c}{c}\right);$$

以及 $\frac{D(n)}{\log(n)}$ 是 c 的强相合估计. 这些都为 c 的估计 (假设 X_i 为观测数据) 提供了一些渠道, 见 Liu [62].

另外, 如果将 (11) 式改写为

$$\mathbb{P}(D(n) = k) \propto c^k \frac{\Gamma(n)\Gamma(c)}{\Gamma(c+n)} = c^k \int_0^1 \eta^{c-1}(1-\eta)^{n-1} \mathrm{d}\eta,$$

并且设 $c\sim \mathrm{Gamma}(a,b)$ $(a,b>0)$, 且定义随机变量 $\eta\sim U[0,1]$, 则可以证明

$$c|\eta,k\sim \mathrm{Gamma}(a+k,b-\log\eta),\quad \eta|c,k\sim\beta(c,n).$$

这对于精度参数 c 在 MCMC 抽样中起到至关重要的作用 (见 [25]).

Antonika[2] 对来自 P 的样本作了更进一步的考察, 指明了有限样本下各种类发生的概率机制. 特别地, 令

$$\{(X_1,X_2,\cdots,X_n)\in C(m_1,m_2,\cdots,m_n)\}=\begin{cases} m_1 \text{ 个不同值恰出现 } 1 \text{ 次},\\ m_2 \text{ 个不同值恰出现 } 2 \text{ 次},\\ \vdots\\ m_n \text{ 个不同值恰出现 } n \text{ 次}\end{cases}$$

来表示样本中 "类" 的模式. 如果用餐桌来比喻则为: 只有 1 个客人的桌子数为 m_1 个, 只有 2 个客人的桌子数为 m_2 个, 依次类推 (注意: 桌子代表不同的 X_i 值). 显然, $\sum_{i=1}^n im_i=n$, $D(n)=\sum_{i=1}^n m_i$.

定理 1.8[2] 设 $P\sim D(cF_0)$, 且 F_0 非原子. 若 $X_1,X_2,\cdots,X_n|P\sim P$, 则

$$\mathbb{P}\big(\{X_1,X_2,\cdots,X_n\}\in C(m_1,m_2,\cdots,m_n)|c\big)=\frac{n!}{\prod_{i=1}^n i^{m_i}(m_i!)}\frac{c^{m_1+m_2+\cdots+m_n}}{c^{[n]}}. \tag{12}$$

特别地, $\mathbb{P}(D(n)=k|c)$ 则是 (12) 对满足 $\sum_{i=1}^n m_i=k$ 的所有 "类" 的概率和.

如果记 $Q(x_1,x_2,\cdots,x_n)$ 为 DP 下样本 $X_1,X_2,\cdots,X_n$ 的联合分布函数, 则由定理 1.6 可知, Q 关于 $\mathbb{R}^n$ 上的勒贝格测度 λ^n 奇异. 然而限制在低维空间 $\mathbb{R}^k$ $(k\le n)$, 则在 F_0 有密度函数 $f_0(x)$ 的条件下, Q 的概率密度函数有解析形式.

引入用以标示样本中类 (有客人的桌子) 的符号: 令 $\mathbf{P}=\{\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_m\}$ 是 $\{1,2,\cdots,n\}$ 的一个划分, $e_k=\#\mathbf{p}_k$ 为集 $\mathbf{p}_k$ 的容度, $m=n(\mathbf{P})=\#\mathbf{P}$. 很显然, 一个 $\mathbf{P}$ 代表了 n 个顾客的一个具体的分配方案, $m=n(\mathbf{P})$ 表示有客人的圆桌数, e_k 代表该方案下每个圆桌的客人数. $X_i=X_j$ 当且仅当 $i,j\in\mathbf{p}_k$ 对某个 $k=1,2,\cdots,m$ 成立.

定理 1.9 (Antoniak[2]) 设 $X_1,X_2,\cdots,X_n|P\sim P$, $P\sim D(cF_0)$, 且 $\mathrm{d}F_0(x)=f_0(x)\mathrm{d}x$ 且非原子. 则在 $R_{\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_m}$ 上 $(X_1,X_2,\cdots,X_n)$ 的联合密度函数为

$$p(x_1,x_2,\cdots,x_n)=\frac{c^m\prod_{j=1}^m(e_j-1)!f_0(x_j^*)}{c^{[n]}}, \tag{13}$$

其中 $\{x_j^*\}$ 为 $x_1,x_2,\cdots,x_n$ 的不同元素.

注 如果 F_0 是含有原子的概率分布测度, 则见定理 2.5.

如果

$$\pi(\mathbf{P})=\frac{c^{n(\mathbf{P})}\prod_{j=1}^{n(\mathbf{P})}(e_j-1)!}{c^{[n]}}$$

用来表示 n 个顾客一个圆桌的分配机会, 那么

$$p(x_1,x_2,\cdots,x_n|\mathbf{P})=\prod_{j=1}^{n(\mathbf{P})}f_0(x_j^*)$$

则意味着 n 个顾客坐到指定圆桌的概率. 此时, 边际联合密度函数可以表示为

$$p(x_1,x_2,\cdots,x_n)=\sum_{\mathbf{P}}\pi(\mathbf{P})p(x_1,x_2,\cdots,x_n|\mathbf{P}),$$

这里求和是对 $\{1,2,\cdots,n\}$ 的所有划分而言. 进一步的情况见定理 2.4 和定理 2.5.

2 DP 的几种延伸

2.1 混合 DP (MDP)

DP 中的底分布主要是对样本真实分布的一种事先假定或认识. 由于真实分布未知, 这种假定可以允许底分布含有未知参数, 这便得到混合 DP. Antoniak[2] 于 1974 年将 DP 中的 c 和 F_0 推广为带有未知参数的 c_θ 或 F_θ, 并且给予 θ 一个先验 $H(\theta)$, 建立了混合 DP; 并在该框架内研究了 DP 下样本的抽样性质和聚类行为, 得到了一些重要结果.

严格地来说, 设 $(\mathcal{X},\mathscr{B})$ 为样本空间, $(\Theta,\mathscr{B}_\Theta,H)$ 为概率空间. $\alpha_\theta(\cdot)$ 为 $\Theta\times\mathscr{B}$ 上有限非零的有限可加转移测度. 称 $(\mathcal{X},\mathscr{B})$ 上的随机概率分布测度 P 为带有参数 α_θ 和混合分布 H 的 MDP, 记为 $P\sim\int_\Theta D(\alpha_\theta)H(\mathrm{d}\theta)$, 若对于 $\mathcal{X}$ 的可测划分 $\{B_1,B_2,\cdots,B_m\}\subseteq\mathscr{B}$, 及对于任何的 $0\leq y_1,y_2,\cdots,y_m\leq 1$,

$$\mathbb{P}(P(B_1)\leq y_1,\cdots,P(B_m)\leq y_m)=\int_\Theta F(y_1,\cdots,y_m|\alpha_\theta(B_1),\cdots,\alpha_\theta(B_m))H(\mathrm{d}\theta),$$

其中 $F(y_1,\cdots,y_m|\alpha_\theta(B_1),\cdots,\alpha_\theta(B_m))$ 为 $\mathrm{Dir}(\alpha_\theta(B_1),\alpha_\theta(B_2),\cdots,\alpha_\theta(B_m))$ 的累积分布函数.

类似地, 令 $c_\theta=\alpha_\theta(\mathcal{X})$, $F_\theta(\cdot)=\frac{\alpha_\theta(\cdot)}{c_\theta}$, 则可记为 $P\sim\int_\Theta D(c_\theta F_\theta)H(\mathrm{d}\theta)$, 或者写成熟知的分层贝叶斯形式: $P|\theta\sim D(c_\theta F_\theta)$, $\theta\sim H$.

如果说 Ferguson 给出了单个 DP, 那么 Antoniak 实际上定义了一族 DP: $\{D(\alpha_\theta):\theta\in\Theta\}$, 并按照 H 加权平均给出 P 的先验. 将 Ferguson 先验的有关性质稍微修改, 即可得到 MDP 类似的结果. 例如, 若 $P\sim\int_\Theta D(c_\theta F_\theta)H(\mathrm{d}\theta)$, 且 $X|P\sim P$, 则 $\forall A\in\mathscr{B}$,

$$\mathbb{P}(X\in A)=\int_\Theta F_\theta(A)H(\mathrm{d}\theta).$$

这表明 X 的边际分布是 $\{F_\theta:\theta\in\Theta\}$ 按照 H 加权平均 (混合分布).

MDP 在非参数贝叶斯分析中有着广泛的应用. 下面结果给出了分层贝叶斯分析中 P 的后验性质, 它表明 MDP 先验也具有共轭性.

定理 2.1 (Antoniak[2]) 设 $(\Theta,\mathscr{B}_\Theta,H(\cdot))$ 为指标测度空间, P 为 $(\mathcal{X},\mathscr{B})$ 上带有混合分布 H 和转移测度 $\alpha_\theta(\cdot)$ 的 MDP. $(\mathcal{Y},\mathscr{B}_Y)$ 为完备可分空间. $F(\cdot|x)$ 是 $\mathcal{X}\times\mathscr{B}_Y$ 转移概率测度. 如果 X 为来自 P 的样本容量为 1 的样本 (即 $[X|P,\theta]\sim P$, 且 $[Y|P,X,\theta]\sim F(\cdot|x)$), 则

$$[P|Y=y]\sim\int_{\mathcal{X}\times\Theta}D(\alpha_\theta+\delta_x)H(\mathrm{d}x,\mathrm{d}\theta|y).$$

其中 $H(\mathrm{d}x,\mathrm{d}\theta|y)$ 为由 $F(y|x),\alpha_\theta,H$ 共同决定的转移概率分布, $\alpha_\theta(\cdot)+\delta_x(\cdot)$ 为 $(\mathcal{X}\times\Theta)\times\mathscr{B}$ 的转移测度.

定理 2.1 有两种特殊情形: 一种情形就是 H 为退化分布, 另外一种情形就是不出现 Y. 对于情形一, 则有

$$\begin{cases} Y|X,P\sim F(y|x) \\ X|P\sim P \\ P\sim D(cF_0) \end{cases} \Rightarrow [P|Y=y]\sim \int_{\mathcal{X}} D(cF_0+\delta_x)H(\mathrm{d}x|y),$$

其中 $H(\mathrm{d}x|y)$ 是由 $F(x|y)$ 和 H 决定的条件分布函数; 而对于情形二,

$$\begin{cases} X|P,\theta\sim P \\ P|\theta\sim D(\alpha_\theta) \\ \theta\sim H(\,\cdot\,) \end{cases} \Rightarrow [P|X=x]\sim \int_{\Theta} D(\alpha_\theta+\delta_x)H(\mathrm{d}\theta|x),$$

其中 $H(\mathrm{d}\theta|x)$ 是由 α_θ 和 H 决定的条件概率分布.

上述结果推广到 n 个观测值也有类似的结果, 但会有一些细节上的麻烦, 主要原因是由于来自 P 的样本中存在着 "结", $X_1,X_2,\cdots,X_n$ 的边际联合分布 $Q(x_1,x_2,\cdots,x_n,\theta)$ 在 $\mathbb{R}^n$ 上退化, 其直接导致后验分布 $H(\mathrm{d}\theta|x_1,x_2,\cdots,x_n)$ 的解析形式复杂.

定理 2.2 设 $(\Theta,\mathscr{B}_\Theta,H(\,\cdot\,))$ 为指标测度空间, P 为 $(\mathcal{X},\mathscr{B})$ 上带有混合分布 H 和转移测度 $\alpha_\theta(\,\cdot\,)$ 的 MDP. 若 α_θ 满足下面的定理 2.5 的条件, 且 $X_1,X_2,\cdots,X_n|P,\ \text{iid.}\sim P$, 则

$$[P|X_1,X_2,\cdots,X_n]\sim \int_{\Theta} D\bigg(c_\theta F_\theta+\sum_{i=1}^{n}\delta_{x_i}\bigg)H(\mathrm{d}\theta|x_1,x_2,\cdots,x_n),$$

其中 $H(\mathrm{d}\theta|x_1,x_2,\cdots,x_n)$ 是由 H 和 $X_1,X_2,\cdots,X_n$ 的边际联合分布 $Q(x_1,x_2,\cdots,x_n,\theta)$ 决定的条件分布, 见定理 2.5.

下面这个结果给出 MDP 下 $H(\theta|X_1,X_2,\cdots,X_n)$ 的后验分布.

定理 2.3 (Antoniak[2]) 设 $P\in\int_\Theta D(\alpha_\theta)H(\mathrm{d}\theta)$, 且 $\alpha_\theta(\,\cdot\,)=c_\theta F_\theta$; $X_1,X_2,\cdots,X_n|P,\ \text{iid.}\sim P$. 假设在 $(\mathcal{X},\mathscr{B})$ 存在一个 σ- 有限的 σ- 可加测度 $\lambda(\,\cdot\,)$, 使得 (i) $F_\theta(\mathrm{d}x)=f_0(x|\theta)\lambda(\mathrm{d}x)$, (ii) $\lambda(\,\cdot\,)$ 在 $F_\theta(\,\cdot\,)$ 的每个原子上的质量为 1. 则

$$H(\mathrm{d}\theta|X_1,X_2,\cdots,X_n)=\frac{\frac{c_\theta^k}{c_\theta^{[n]}}\prod_{j=1}^{k}f_0(x_j^*|\theta)(m_\theta(x_j^*)+1)^{[e_j^*-1]}H(\mathrm{d}\theta)}{\int_\Theta\frac{c_\theta^k}{c_\theta^{[n]}}\prod_{j=1}^{k}f_0(x_j^*|\theta)(m_\theta(x_j^*)+1)^{[e_j^*-1]}H(\mathrm{d}\theta)}, \tag{14}$$

其中 $\{x_1^*,x_2^*,\cdots,x_k^*\}$ 为 $\{x_1,x_2,\cdots,x_n\}$ 的不同值集, $e_j^*=\#\{i:x_i=x_j^*\}$,

$$m_\theta(x_j^*)=c_\theta F_\theta(\{x_j^*\})=\begin{cases} c_\theta f_0(x_j^*|\theta), & \text{如果 } x_j^* \text{ 是 } F_\theta \text{ 的原子;} \\ 0, & \text{否则.} \end{cases}$$

特别地, 如果 $F_\theta(\,\cdot\,)$ 为非原子的且 $c_\theta=c$ 与 θ 无关, 则

$$H(\mathrm{d}\theta|X_1,X_2,\cdots,X_n)=\frac{\prod_{j=1}^{k}f_0(x_j^*|\theta)H(\mathrm{d}\theta)}{\int_\Theta\prod_{j=1}^{k}f_0(x_j^*|\theta)H(\mathrm{d}\theta)}. \tag{15}$$

Doss[18] 用另一种方式给出了 (15) 的推导, 并通过重要抽样法来获得 P 的后验分布的抽样. 这里也可以应用本文第 3 节给出的 MCMC 抽样技术来完成从 $H(\mathrm{d}\theta|X_1,X_2,\cdots,X_n)$ 的抽样.

2.2 DP 混合 (DPM)

Lo[63] 在考虑密度估计时使用了核函数的 DP 加权形式, 并给出了估计的解析形式. 这项工作可以看作是对 Antoniak 工作 (定理 2.1) 的正式化和规范化.

设 $(\mathcal{Y},\mathscr{B}_Y)$ 和 $(\mathcal{X},\mathscr{B})$ 为欧氏 Borel 空间. $K(y,x)$ 为 $\mathcal{X}\times\mathscr{B}_Y$ 上的转移核.

设 $(\mathcal{X},\mathscr{B})$ 上的随机概率分布测度 $P\sim D(cF_0)$. 对于 $y\in\mathcal{Y}$, 称

$$f(y|P)=\int_{\mathcal{X}}K(y,x)P(\mathrm{d}x) \tag{16}$$

为带有核函数的 Dirichlet 混合过程.

显见, $\forall y\in\mathcal{Y}$:

$$\mathbb{E}[f(y|P)]=\mathbb{E}\left[\int_{\mathcal{X}}K(y,x)P(\mathrm{d}x)\right]=\int_{\mathcal{X}}K(y,x)F_0(\mathrm{d}x)=\mathbb{E}_{F_0}K(y,X).$$

这表明若 $Y\sim f(y|P)$, 则 Y 的边缘分布为 $K*F_0$.

设 $Y_1,Y_2\cdots,Y_n|P,\ \text{iid.}\sim f(y|P)$, 且 $P\sim D(cF_0)$. 主要任务是基于 $Y_1,Y_2\cdots,Y_n$ 来推断 $f(y|P)$. 引入随机变量 $X_1,X_2,\cdots,X_n$, 则该模型等价于如下的分层模型:

$$\begin{cases}Y_i|X_i,P\overset{\text{ind}}{\sim}K(y_i,x_i),\\ X_i|P\overset{\text{iid}}{\sim}P,\\ P\sim D(cF_0).\end{cases} \tag{17}$$

这显然属于定理 2.1 带有 n 个变量时的特殊情形一. 由于涉及到 P 的后验分布, 一般采用两种方式来处理: 一种是基于蒙特卡洛抽样方法, 另外一种是找出其解析表达式. 这里主要介绍 Lo (1984) 的工作 (见 [63]).

设 $\psi(x)$ 是 $(\mathcal{X},\mathscr{B})$ 非负可测函数且 $E_{F_0}\psi(X)<\infty$, 则 $g(P)=\int_X\psi(x)P(\mathrm{d}x)$ 几乎处处有限. 下列定理给出了分层模型下 P 的泛函的后验性质.

定理 2.4 (Lo[63])

$$\begin{aligned}\mathbb{E}[g(P)|Y_1,Y_2,\cdots,Y_n]&=\int g(P)\mathscr{P}(\mathrm{d}P|Y_1,Y_2\cdots,Y_n)\\&=\int_{\mathcal{X}^n}\int g(P)\mathscr{P}(\mathrm{d}P|X_1,X_2,\cdots,X_n)\mathrm{d}H(x_1,x_2,\cdots,x_n|Y_1,Y_2,\cdots,Y_n),\end{aligned}$$

其中 $[P|X_1,X_2,\cdots,X_n]\sim D(cF_0+\sum_{i=1}^n\delta_{X_i})$,

$$H(\mathrm{d}x_1,\mathrm{d}x_2,\cdots,\mathrm{d}x_n|Y_1,Y_2,\cdots,Y_n)=\frac{\prod_{i=1}^nK(Y_i,x_i)\prod_{i=1}^n\left(cF_0(\mathrm{d}x_i)+\sum_{j=1}^{i-1}\delta_{x_j}(\mathrm{d}x_i)\right)}{\int_{\mathcal{X}^n}\prod_{i=1}^nK(Y_i,x_i)\prod_{i=1}^n\left(cF_0(\mathrm{d}x_i)+\sum_{j=1}^{i-1}\delta_{x_j}(\mathrm{d}x_i)\right)}. \tag{18}$$

由该定理立得: 给定 $Y_1,Y_2,\cdots,Y_n$, P 的后验分布是带有混合分布 $H(x_1,x_2,\cdots,x_n|Y_1,Y_2,\cdots,Y_n)$ 的 MDP (Antoniak[2]). 严格地说,

$$[P|Y_1,Y_2,\cdots Y_n]\sim\int D\left(cF_0+\sum_{i=1}^n\delta_{x_i}\right)\mathrm{d}H(x_1,x_2,\cdots,x_n|Y_1,Y_2,\cdots,Y_n).$$

这也是定理 2.1 的结果. 另外, 定理 2.4 也给出了拟合分布 F 的一个估计

$$\begin{aligned}\hat{F}(x|Y_1,Y_2,\cdots,Y_n)&=\mathbb{E}[\mathbb{E}[P((-\infty,x])|X_1,X_2,\cdots,X_n]|Y_1,Y_2,\cdots,Y_n]\\&=\frac{n}{n+c}\int_{\mathcal{X}^n}F_n(x|x_1,x_2,\cdots,x_n)H(\mathrm{d}x_1,\mathrm{d}x_2,\cdots,\mathrm{d}x_n|Y_1,Y_2,\cdots,Y_n)\\&\quad+\frac{c}{n+c}F_0(x).\end{aligned}$$

现在问题的关键是求条件分布 (18) 中 $H(\mathrm{d}x_1,\mathrm{d}x_2,\cdots,\mathrm{d}x_n|Y_1,Y_2,\cdots,Y_n)$ 的密度函数.

设 F_0 无原子, 且 $\mathrm{d}F_0=f_0(x)\mathrm{d}x$. 在定理 1.9 的记号下, $(X_1,X_2,\cdots,X_n)$ 的联合密度函数为

$$p(x_1,x_2,\cdots,x_n)=\frac{c^{n(\mathbf{P})}\prod_{j=1}^{n(\mathbf{P})}(e_j-1)!f_0(x_j^*)}{c^{[n]}},\quad (x_1,x_2,\cdots,x_n)\in R_{\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_{n(\mathbf{P})}}$$

其中 $x_1^*,x_2^*,\cdots,x_{n(\mathbf{P})}^*$ 为 $x_1,x_2,\cdots,x_n$ 不同值之集.

从而 $\{Y_1,Y_2,\cdots,Y_n,X_1,X_2,\cdots,X_n\}$ 在 $R_{\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_{n(\mathbf{P})}}$ 的联合概率密度函数为

$$p(y_1,y_2,\cdots,y_n,x_1,x_2,\cdots,x_n)=\frac{c^{n(\mathbf{P})}\prod_{j=1}^{n(\mathbf{P})}(e_j-1)!\big\{\prod_{i\in\mathbf{p}_j}K(y_i,x_j^*)\big\}f_0(x_j^*)}{c^{[n]}}.$$

由此可得, 在 $R_{\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_{n(\mathbf{P})}}$ 上, $(x_1,x_2,\cdots,x_n|y_1,y_2\cdots,y_n)$ 的条件密度函数为

$$\begin{aligned}&H(\mathrm{d}x_1,\mathrm{d}x_2,\cdots,\mathrm{d}x_n|Y_1,Y_2,\cdots,Y_n)\\&\quad=\frac{c^{n(\mathbf{P})}\prod_{j=1}^{n(\mathbf{P})}(e_j-1)!\big\{\prod_{i\in\mathbf{p}_j}K(Y_i,x_j^*)\big\}f_0(x_j^*)\mathrm{d}x_j^*}{\sum_{\mathbf{P}}c^{n(\mathbf{P})}\prod_{j=1}^{n(\mathbf{P})}(e_j-1)!\int_{\mathcal{X}}\big\{\prod_{i\in\mathbf{p}_j}K(Y_i,x_j^*)\big\}f_0(x_j^*)\mathrm{d}x_j^*}.\end{aligned}\tag{19}$$

一个特例是 $n=1$, $H(\mathrm{d}x_1|y_1)\propto K(y_1,x_1)f_0(x_1)\mathrm{d}x_1$, 这恰是定理 2.1 的特殊情形一.

下面的结果给出了在 $P\sim D(cF_0)$ 情形下, $f(y|P)$ 后验分布的解析形式.

定理 2.5 (Lo[63]) 沿用定理 1.9 的记号. 设 $Y_i|X_i,P,\ \text{ind.}\sim K(y_i,x_i)$, 且 $X_i|P,\ \text{iid.}\sim P$. 若 $P\sim D(cF_0)$ 且 F_0 是非原子的, 则

$$\widehat{f}(y|Y_1,Y_2,\cdots,Y_n)=\mathbb{E}[f(y|P)|Y_1,Y_2,\cdots,Y_n]=\frac{c}{n+c}\mathbb{E}_{F_0}K(y,X)+\frac{n}{n+c}(K*F_n)(y),$$

其中

$$(K*F_n)(y)=\sum_{\mathbf{P}}\varpi(\mathbf{P})\prod_{j=1}^{n(\mathbf{P})}\frac{e_j}{n}\frac{\int_{\mathcal{X}}K(y,x)\prod_{i\in\mathbf{p}_j}K(Y_i,x)F_0(\mathrm{d}x)}{\int_{\mathcal{X}}\prod_{i\in\mathbf{p}_j}K(Y_i,x)F_0(\mathrm{d}x)},$$

$\varpi(\mathbf{P})$ 是权函数

$$\varpi(\mathbf{P})\propto c^{n(\mathbf{P})}\prod_{j=1}^{n(\mathbf{P})}(e_j-1)!\int_{\mathcal{X}}\prod_{i\in\mathbf{p}_j}K(y_i,x)F_0(\mathrm{d}x),$$

使得 $\sum_{\mathbf{P}}\varpi(\mathbf{P})=1.0$.

由定理 1.9 可知, 给定 $\mathbf{P}$, $X_1^*, X_2^*, \cdots, X_{n(\mathbf{P})}^*$ 是独立同分布且服从 F_0. 现给定 $Y_1, Y_2, \cdots, Y_n$, X_j^* 独立, 且后验分布为

$$H_j(\mathrm{d}x_j^*|\mathbf{p}_j, Y_1, Y_2, \cdots, Y_n) = \frac{\prod_{i\in\mathbf{p}_j} K(Y_i, x_j^*)F_0(\mathrm{d}x_j^*)}{\int_{\mathcal{X}} \prod_{i\in\mathbf{p}_j} K(Y_i, x_j^*)F_0(\mathrm{d}x_j^*)}, \quad j = 1, 2, \cdots, n(\mathbf{P}). \tag{20}$$

这点从中餐馆法则上不难理解: 第 j 个类出现的概率从原来的 $F_0(x_j^*)$ 现转变为 $H_j(x_j^*|\mathbf{p}_j)$, 这种认识的改变主要取决于圆桌上的观测值的信息 $\{Y_j : j \in \mathbf{p}_j\}$.

从定理 2.5 不难得到, 观测变量 $Y = \{Y_1, Y_2, \cdots, Y_n\}$ 的预测密度 (边际似然) 为

$$m(Y) = \sum_{\mathbf{P}} \pi(\mathbf{P})\pi(Y|\mathbf{P}),$$

其中

$$\pi(\mathbf{P}) = \frac{c^{n(\mathbf{P})}\prod_{j=1}^{n(\mathbf{P})}(e_j - 1)!}{c^{[n]}}, \quad \pi(Y|\mathbf{P}) = \prod_{j=1}^{n(\mathbf{P})} \int_{\mathcal{X}} \prod_{i\in\mathbf{p}_j} K(Y_i, x_j^*)F_0(\mathrm{d}x_j^*).$$

Ishwaran 和 James[45] 设计了一个算法来对 $\mathbf{P}$ 执行序贯抽样. 进一步地, $X_1, X_2, \cdots, X_n$ 中恰好有 k 个不同值的后验概率为

$$\mathbb{P}(D(n) = k|Y) = \frac{\sum_{\{\mathbf{P}:\ n(\mathbf{P})=k\}} \pi(\mathbf{P})\pi(Y|\mathbf{P})}{m(Y)} = \frac{m_k(Y)}{m(Y)},$$

其中 $m_k(Y) = \sum_{\{\mathbf{P}:\ n(\mathbf{P})=k\}} \pi(\mathbf{P})\pi(Y|\mathbf{P})$.

值得指出的是, Basu 和 Chib[3] 利用 MCMC 技术对 DPM 的边际似然 $m(Y)$ 给出了近似计算, 这对于非参数贝叶斯的统计应用特别是贝叶斯因子计算迈出了重要一步.

2.3 DP 有限维近似

DP 的无限项求和形式对于直接研究其后验统计性质相当不便, 这点从 Lo 的结果 [63] 可以看出. Ishwaran 和 Zarepour[48], Ishwaran 和 James[43−44], Ishwaran 和 Takahara[47] 在考虑 DP 的近似形式时, 对 Sethuraman 表示 (见定理 1.4) 进行了有限维近似, 并作了进一步推广:

$$P \sim \mathscr{P}_N = \sum_{k=1}^{N} \pi_k \delta_{Z_k}(\,\cdot\,)^{\dagger\dagger}, \tag{21}$$

其中 N 为折断水平, $\{\pi_1, \pi_2, \cdots, \pi_N\}$ 为随机权, Z_k, iid. $\sim F_0$ 为原子, 且 $\{\pi_k\}$ 与 $\{Z_k\}$ 独立. 并称先验 (23) 为 $\mathscr{P}_N(\mathbf{a}, \mathbf{b})$ 测度, 其中 $\mathbf{a} = (a_k)_{k=1}^{N-1}, \mathbf{b} = (b_k)_{k=1}^{N-1}$.

显然 $\mathscr{P}_N$ 完全由 $\{\pi_k\}_{k=1}^N$ 与 $\{Z_k\}_{k=1}^N$ 的联合分布来决定, 这是有限维分布. 在此先验下, P 的实现值为 $\sum_{k=1}^N \pi_k\delta_{Z_k}$ 的形式, P 的取值空间为含有至多 N 个原子的有限离散分布之集.

自然地, 可以考虑如下两种随机权 (Ishwaran 和 James[43]):

(1) 折断 stick-breaking 随机权

$$\pi_1 = V_1, \quad \cdots, \quad \pi_m = V_m \prod_{k=1}^{m-1}(1 - V_k), \quad \cdots, \quad \pi_N = \prod_{k=1}^{N-1}(1 - V_k), \tag{22}$$

†† 记号含义: $[P|\{Z_k\}, \{\pi_k\}] = \sum_{k=1}^N \pi_k\delta_{Z_k}(\,\cdot\,)$, $\mathscr{P}_N \overset{D}{=} H(z_1, z_2, \cdots, z_N, \pi_1, \pi_2, \cdots, \pi_K)$.

其中 $\{V_k\}_{k=1}^{N-1}$, ind. $\sim \beta(a_k, b_k)$ $(a_k, b_k > 0)$.

(2) Dirichlet 随机权

$$(\pi_1, \pi_2, \cdots, \pi_N) \sim \mathrm{Dir}_m(A_1, A_2, \cdots, a_N). \tag{23}$$

可以证明 (Connor 和 Mosimann[12], Ishwaran 和 James[43]), 由 (22) 和 (23) 给出的 $\{\pi_k\}$ 的分布均为广义 Dirichlet 分布 ‡, 其后验具有共轭性, 这种特性对于后验分析特别是随机权的后验抽样十分方便.

一种特殊情形是: 随机权 (22) 中的 $V_k \sim \beta(1, c)$, (23) 随机权分布取为对称 Dirichlet 分布 $D(\frac{c}{N}, \cdots, \frac{c}{N})$. 前者对应的 $\mathscr{P}_N$ 称为 Dirichlet 过程的折断形式, 而后者 Ishwaran 和 James 称之为有限维 Dirichlet 先验 [45−46].

下面定理揭示了在 N 趋于无穷时, $\mathscr{P}_N$ 弱收敛到 Dirichlet 过程先验.

定理 2.6 (Ishwaran 和 Zarpour[48]) 在有限维 Dirichlet 先验下, 对于 $(\mathcal{X}, \mathscr{B})$ 的 F_0 可积函数 g,

$$\int\int_{\mathcal{X}} g(x)P(\mathrm{d}x)\mathscr{P}_N\left(\frac{c}{N}P\right) \to \int\int_{\mathcal{X}} g(x)P\left(\frac{c}{N}x\right)\mathscr{P}_\infty\left(\frac{c}{N}P\right)$$

其中 $\mathscr{P}_\infty$ 为 DP 先验.

带有随机权 (22) 的近似先验下的样本边际分布比较复杂, 但对于有限维 Dirichlet 先验下来说, 样本仍然具有 Polya 性.

引理 2.1 设 $K_i|\mathbf{p}$, iid. $\sim \sum_{k=1}^N p_k\delta_k$ $(i = 1, 2, \cdots, n)$ 且 $\mathbf{p} = (p_1, p_2, \cdots, p_N) \sim \mathrm{Dir}_N(\frac{c}{N}, \cdots, \frac{c}{N})$; 记 $K^* = \{K_1^*, K_2^*, \cdots, K_m^*\}$ 为 $\{K_1, K_2, \cdots, K_{i-1}\}$ 不同值之集, 且 n_j^* 为 K_j^* 重复的次数, 则

$$\mathbb{P}(K_i = K_j^*|K_1, K_2, \cdots, K_{i-1}) = \frac{n_j^* + c/N}{c + i - 1},$$
$$\mathbb{P}(K_i \bar{\in} K^*|K_1, K_2, \cdots, K_{i-1}) = \frac{c(1 - m/N)}{c + i - 1}.$$

由引理 2.1 立即得到:

定理 2.7 (Polya urn scheme, Pitman[81]) 设 $P \sim \mathscr{P}_N$ 且为有限维 Dirichlet 先验. 若 $X_1, X_2, \cdots, X_n|P$, iid. $\sim P$, 则

$$X_n|X_1, X_2, \cdots, X_{n-1} \sim \frac{c(1 - m_n/N)}{c + n - 1}F_0 + \sum_{j=1}^{m_n}\frac{n_j^* + c/N}{c + n - 1}\delta_{X_j^*}, \tag{24}$$

其中 m_n 为 $X_1, X_2, \cdots, X_{n-1}$ 不同值的个数, $X_1^*, X_2^*, \cdots, X_{m_n}^*$ 为不同值之集, n_j^* 为 X_j^* 重复次数.

显然, 当 $N \to \infty$ 时, 式 (24) 即为 DP 先验下的预测分布 (9). 这从另一个角度说明有限维 Dirichlet 先验可以看做 DP 一个好的近似.

‡ 广义 Dirichlet 分布 $\mathrm{GD}(A_1, A_2, \cdots, a_{N-1}, b_1, b_2, \cdots, b_{N-1})$ 的密度函数为

$$\left(\prod_{k=1}^{N-1}\frac{\Gamma(a_k + b_k)}{\Gamma(a_k)\Gamma(b_k)}\pi_k^{a_k - 1}\right)\pi_N^{b_{N-1}-1}\prod_{m=1}^{N-2}\left(1 - \sum_{l=1}^{m}\pi_l\right)^{b_m - (a_{m+1} + b_{m+1})}.$$

进一步地, Ishwaran 和 James[43] 基于分层模型的边际似然考虑了这种近似产生的误差. 设

$$\begin{cases} Y_i|X_i,\ \text{ind.} \sim p(y_i|x_i), \\ X_i|P,\ \text{iid.} \sim P, \\ P \sim \mathscr{P}_N, \end{cases} \tag{25}$$

并记

$$m_N(Y) = \int \prod_{i=1}^{n} p(y_i|x_i) P(\mathrm{d}x_i) \mathscr{P}_N(\mathrm{d}P) = \sum_{\mathbf{P}} \pi_N(\mathbf{P}) p(Y_1, Y_2, \cdots, Y_n|\mathbf{P})$$

为 $Y_1, Y_2, \cdots, Y_n$ 的边际分布, 其中

$$\pi(\mathbf{P}) = \frac{(c/N)^{n(\mathbf{P})} N!}{c^{[n]}(N - n(\mathbf{P}))!} \prod_{l=1}^{n(\mathbf{P})} \left(1 + \frac{c}{N}\right)^{[e_l - 1]}.$$

定理 2.8 (Ishwaran 和 Zarepour[48], Ishwaran 和 James[43])

$$\int_{\mathcal{Y}^N} |m_N(Y) - m_\infty(Y)| \mathrm{d}Y \leq 4\left(1 - \mathbb{E}\left[\left(\sum_{k=1}^{N-1} \pi_k\right)^n\right]\right). \tag{26}$$

特别地, 对于 DP 先验, 上界等价于 $4n\exp(-(N-1)c)$.

Ishwaran 和 Zarepour[48] 设计了一种分块 Gibbs 算法用来执行 Bayes 抽样. 这种算法显著性特点是避免了 DP 下样本的 Polya urn 性质的使用, 改为直接对 P 抽样. 详细见本文第 3 节.

2.4 相依 DP (DDP)

在纵向数据分析、分组数据分析、多水平分析、方差分析等研究中, 统计模型通常要涉及到 "组间分布" 或 "组间函数" 的拟合问题. 相依 DP 就是对随机函数族 $\{P_\lambda : \lambda \in \Lambda\}$ 建立具有某种关联性的 DP 构建. 建立相依 DP 来主要是利用过程的离散性来拉拢不同组别的信息, 共享部分原子, 从而达到 "share strength" 的目的. 研究的主要内容是厘清过程族内在的关联性, 至于其样本的抽样性质、后验分析和计算却都是继承 DP 或 MDP.

相依 DP 可以溯源到 Antoniak[2] 的混合 DP 过程, 但这些过程是通过底分布的超参数来建立联系. Cifarelli 和 Regazzini[11] 是首次考虑了通过协变量有关联的 DP. 具体地说, 他们假定过程 $F_x \sim D(cF_{0x})$, 其中 $F_{0x}(\,\cdot\,) = N(\,\cdot\,|\beta x, \sigma^2)$, x 为回归协变量. 这项工作可以视为 Antoniak [2] 工作的一个推广: 允许底分布的均值带有回归协变量, 协变量不同水平下的 DP 通过公共的回归系数来建立联系. 但这种推广只是建立了底分布之间的简单联系, 没有涉及到不同分布的原子和随机权间的联系. 该模型后来被 Muliere 和 Petrone [73] 作出进一步讨论; Mira 和 Petrone[72], Carota 和 Parmigiani[9], Giudici 等 [38] 在不同的场合下都采用类似做法. 有关底分布回归的 DP 拟合可进一步参考 Tomlinson 和 Escobar[96] 和 Gelfand 等 [33].

建立过程关联性的另外一种渠道是将协变量引入到过程的随机权和原子上. MacEachern[65] 首次定义了如下 DP: 设 D 为一指标集, $\theta_k(D) = \{\theta_{kx} : x \in D\}$ 为定义在 D 上的原子序列, 且假定 $\forall x \in D : \theta_{kx}, \text{iid.} \sim F_0$. 考虑分布函数族 $\mathcal{P} = \{P_x : x \in D\}$, 对每个 $x \in D$,

$$P_x(\,\cdot\,) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k(x)}(\,\cdot\,), \tag{27}$$

其中 $\boldsymbol{\pi}=(\pi_k)\sim \mathrm{GEM}(c)$ 与 x 无关. 这种推广主要是刻画协变量不同水平下的 F_x 与 $F_{x'}$ 的关联性.

De Iorio 等 [15] 在考虑双因素方差分析模型的分布拟合时建立了 ANOVA 类型的相依 DP. 设 $x=(v,w)$ $(v=1,2,\cdots,V,\ w=1,2,\cdots,W)$ 为两因素分类向量, 考虑 $F(v,w)=\sum\pi_k\delta_{\theta_k(v,w)}$, 其中 $\theta_k(v,w)=m_k+A_{kv}+B_{kw}$, A_{kv}, B_{kv} 分别为因素 v,w 的随机效应. 这种做法目的是解释因素的水平不同组合间分布的关联.

Müller 等 [75] 在考虑多组数据的交叉分析时, 假定 $Y_j=\{y_{ij},i=1,2,\cdots,n_j\}$ 满足:

$$Y_j|H_j\sim p(y_j|H_j)=\int p(y_j|x_j)H_j(\mathrm{d}x_j).$$

为了建立组间水平分布 $\{H_j\}$ 的关联性, 假定 $H_j=\varepsilon F_0+(1-\varepsilon)F_j$ $(0\leq\varepsilon\leq 1)$, 其中 F_0 代表各个组的共同底分布, F_j 代表 H_j 偏离底线分布的 "异质" 行为. 这是一种折中策略: 因为 $\varepsilon=0$ 和 1 分别对应于独立模型和可交换模型; 而从组间拉拢信息强度来看, 显然独立模型较可交换模型强度大. 为了区分上述两种极端情形, Müller 等 [76] 建议用带有高斯核的 DPM 来拟合 F_j: $H_j\sim\mathrm{DP}(\alpha,G_0)$. 这种做法使各组分布享有共同原子集, 导致组间分布关联性产生. 这些工作可以看做是在组间层面建立过程的相依结构.

Gelfand 等 [33] 将 MacEachern[65] 的工作延伸到空间数据 $Y(D)=\{Y(x):x\in D\}$, 并建立空间 DP (SDP): $P(x)=\sum_{k=1}^{\infty}\pi_k\delta_{\theta_k(x)}$, 其中 D 为欧氏空间的子集, $\boldsymbol{\pi}=(\pi_k)\sim\mathrm{GEM}(c)$, $\theta_k(D)=\{\theta_k(x):x\in D\},\mathrm{iid.}\sim F_0$, 且与 $\boldsymbol{\pi}$ 独立, 这里 F_0 为随机过程 (如高斯过程) 先验, $\theta_k(D)$ 的实现值为空间 D 上的一个 "曲面". 对于观测过程 $Y(D)$, 有限维分布 $Y^n=\{Y(x_1),Y(x_2),\cdots,Y(x_n)\}\sim G^n$, $G^n\sim D(cF_{0n})$, 这里 F_{0n} 为 F_0 的有限维分布 (比如多元正态分布). 空间 DP 的先验完全由 $\boldsymbol{\pi}$ 和 $\{\theta_{kx}:k\in\mathbb{N},x\in D\}$ 的联合分布来刻画. 与单个 DP 下的样本性质类似, 空间 DP 的观测数据 Y_D (也就是曲面) 也呈现出聚 "类" 的特性. 这不仅建立了空间数据因地域变化而产生的关联性, 而且指明这些曲面可能存在 "类" 的差别和联系. Gelfand 等 [33] 进一步讨论了 F_0 选取以及后验 MCMC 抽样的设计. Duan 等 [19] 和 Petrone 等 [79] 将此项工作推广到高维乘积空间, 建立了广义空间 Dirchlet 过程.

为了解决空间点位置远近的相似性问题, Griffin 和 Steel[39] 构建了基于次序的空间 DP (π DDP). 其主要想法是对于空间的每个位置 x 处, 存在一个置换 $\mathbf{p}(x)$, 使得 P_x 的 stick-breaking 构成 (27) 中的随机权与原子被重新标示. 这种置换要保证相近位置的过程要有较强的关联性 (次序性). 具体地说, 设 $\mathbf{p}(x)=(p_1(x),p_2(x),\cdots,p_{n(x)}(x))$ 满足: i) $\mathbf{p}(x)\subseteq\{1,2,\cdots,N\}$ $(n(x)\leq N)$; ii) $p_i(x)=p_j(x)$, 当且仅当 $i=j$ (确定性). 注意, $n(x)$ 意味着置换数目随 x 而异. 定义

$$P_x(\,\cdot\,)=\sum_{m=1}^{n(x)}\pi_m(x)\delta_{\theta_{p_m(x)}}(\,\cdot\,),\tag{28}$$

其中 $\theta_1,\theta_2,\cdots,\theta_N,\ \mathrm{iid.}\sim H$,

$$\pi_1(x)=V_{p_1(x)},\quad\cdots,\quad\pi_m(x)=V_{p_m(x)}\prod_{k=1}^{m-1}(1-V_{p_k(x)}),\quad\cdots,\quad\pi_{n(x)}=\prod_{k=1}^{n(x)-1}(1-V_{p_k(x)}),$$

这里, $V_k,\mathrm{ind.}\sim\beta(a_k,b_k)$ $(k=1,2,\cdots,N-1,a_k,b_k>0)$. 容易证明, 给定 $p(x)$, $\mathbb{E}P_x(B)=H(B)$. 在 $a_k=1,\ b_k=c,\ N=\infty$ 时 $\mathbb{D}P_x(B)=\frac{H(B)(1-H(B))}{c+1}$. 这些结果与单个 DP 是一致的. 特别地,

Griffin 和 Steel 指出了 $P_{x_1}(B)$ 与 $P_{x_2}(B)$ 的关联性, 并使用点过程如 Possion 过程来确定 $\mathbf{p}$. 关于 SDP 的一个应用可参见 Reich 和 Fuentes [82].

Dunson 等 [22] 在考虑贝叶斯密度回归估计时提出了加权 MDP (WMDP)

$$P_x = \sum_{m=1}^{n} b_m(x) P^*_{x_m}, \quad P^*_{x_m}, \text{ iid.} \sim D(cG_0) \tag{29}$$

其中 $\mathbf{b}(x) = (b_1(x), b_2(x), \cdots, b_n(x))$ 为权函数. 这本质上是 n 个 DP 按照 $\mathbf{b}$ 加权混合的过程. Dunson 等 [22] 给出该模型下样本的 Polya urn 性, 并结合 MCMC 技术获得了后验分析. 相依 DP 过程的其他延伸和发展可参见 Dunson 和 Park [21], Rodriguez 等 [85] , Dunson [20], Crandell 和 Dunson [13], Scarpa 和 Dunson [86] 等.

2.5 分层 DP (HDP)

分层 DP 可以看做是 DDP 的一种形式, 与通过引入协变量来建立关联性不同的是, 这里主要是通过共享不同层的原子来建立相依. 假设一组数据分成 J 个组别, 但每个组别可能存在潜在 "类". 一般的做法是用 J 个分离的 DP 来拟合每个组别的分布: $j = 1, 2, \cdots, J,\ i = 1, 2, \cdots, n_j$,

$$\begin{cases} Y_{ij}|\theta_{ij} \sim F(y_{ij}, \theta_{ij}), \\ \theta_{1j}, \cdots, \theta_{n_j,j}|G_i, \text{iid.} \sim G_j, \end{cases} \tag{30}$$

并假定 $G_j \sim D(\alpha_{0j} F_{0j})$. 但这种做法会隔离组间的 "类" 的联系. 为了拉拢不同组别的 "类" 的信息, 一种做法就是将 α_{0j} 或 F_{0i} 联接起来. 如 Cifarelli 和 Regazzini[11], MacEachern[65−66], Tomlinson 等 [96], Müller 等 [75], De Iorio 等 [15], Ishwaran 和 James[46] 等人的工作. 另外一种做法是考虑 $G_j \sim D(\alpha_0 G_0(\tau))$, 并将 τ 视为随机变量. 积分出 τ 会产生相依的 DP (见 Carota 和 Parmigiani [9], Fong 等 [30], Muliere 和 Petrone [73]). 但这种方法一般不能解释组间类可能存在 "共享" 的原子.

Teh 等 [95] 对分组数据进行建模时提出了分层 Dirichlet 过程. 该过程假定:

$$\begin{cases} G_j|\alpha_0, G_0, \text{ iid.} \sim D(\alpha_0 G_0),\ j = 1, 2, \cdots, J, \\ G_0|\gamma, H \sim D(\gamma H), \end{cases} \tag{31}$$

其中 H, G_0 分别为 $(\mathcal{X}, \mathscr{B})$ 的已知分布, $\gamma > 0,\ \alpha_0 > 0$.

根据定理 1.4 可知

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k},$$

其中 $\phi_k, \text{iid.} \sim H$, $\boldsymbol{\beta} = \{\beta_k\} \sim \text{GEM}(\gamma)$, 且两者相互独立. 另外, 由 DP 的离散性可知, G_j 必为如下形式:

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}, \tag{32}$$

且 $\{\pi_{jk}\}$ 与 $\{\pi_{j'l}\}$ 独立.

式 (32) 表明: 不同组间的分布可能共享相同的原子 ϕ_k, 这对于提取和拉拢不同组间的信息、洞悉观测数据内在的概率机理是十分必要的.

定理 2.9 (Teh 等 [95]) 若记 $\boldsymbol{\beta}=(\beta_k)_{k=1}^{\infty}$, $\boldsymbol{\pi}_j=(\pi_{jk})_{k=1}^{\infty}$, 则

$$\boldsymbol{\pi}_j|\alpha_0,\boldsymbol{\beta}\sim D(\alpha_0\boldsymbol{\beta}).$$

该定理揭示了两个随机权之间的内在联系. 更进一步地, 由 Sethurman 构成可知, 存在 $\{V_m\}$, iid. $\beta(1,\gamma)$ $(m=1,2,\cdots)$, 使得

$$\beta_1=V_1,\quad \cdots,\quad \beta_m=V_m\prod_{l=1}^{m-1}(1-V_l),\quad \cdots.$$

则由定理 2.9 可知, 存在 $\{U_{jk}\}$, ind. $\sim\beta\big(\alpha_0\beta_k,\alpha_0(1-\sum_{l=1}^{k}\beta_l)\big)$ 使得

$$\pi_{j1}=U_{j1},\quad \cdots,\quad \pi_{jm}=U_{jm}\prod_{l=1}^{m-1}(1-U_{jl}),\quad \cdots.$$

Teh 等发展了模型 (30)–(31) 的中餐馆法则: 用享有一个共同餐单的多家联营餐馆的顾客分配计划来比喻该过程的概率机制, 给出了其近似模型及 MCMC 的抽样计划. 详细情形可参见 Teh 等 [95].

2.6 嵌套 DP (NDP)

Rodriguez 等 (2008)[84] 在考虑 Teh 等 [95] 的问题 (31) 时提出了一个嵌套式 DP. Rodriguez 等注意到按照 Teh 等的模型 (31), 组间分布 G_j, $j=1,2,\cdots,J$ 只是享有同一个离散分布的原子集 (一个菜单), 尽管每个组内的观测数据可能有不同的 "子类" (原子), 但这些子类都来自 G_0. 但在实际问题中, 可能这些原子本身就来自不同的分布, 它们之间也有类的存在. 用中餐馆规则的菜单来比喻: 可能菜单上存在不同菜系之差别. 为此对 (31) 做了以下推广:

$$\begin{cases}\mathbb{P}(G_j=G_k^*|\{\pi_k^*\},\{G_k^*\})=\pi_k^*, & \text{任意的 } j=1,2,\cdots,J,\\ G_k^*=\sum\limits_{\ell=1}^{\infty}w_{k\ell}\delta_{\theta_{k\ell}^*}, & \text{任意的 } k=1,2,\cdots,\end{cases}\tag{33}$$

其中 $\theta_{k\ell}^*\sim H$, 对所有的 k,ℓ, $w_{k\ell}^*=u_{k\ell}^*\prod_{s=1}^{\ell-1}(1-u_{sk}^*)$, $\pi_k^*=v_k^*\prod_{s=1}^{k-1}(1-v_s^*)$, $v_k^*,\text{iid.}\sim\beta(1,\alpha)$, $u_{k\ell}^*,\text{iid.}\sim\beta(1,\beta)$. H 为 $(\mathcal{X},\mathscr{B})$ 的已知分布. 模型 (33) 表明第 j 组的分布与第 j' 组的分布有关联, 而且它们分别共享原子 $\{\theta_{k\ell}^*\}$. 事实上, Rodriguez 等 [84] 证明了以下结果:

$$\begin{cases}\mathbb{P}(G_j=G_{j'}|H)=\dfrac{1}{1+c},\\ \mathbb{E}(G_j(A)|H)=H(A),\\ \mathbb{D}(G_j(A)|H)=\dfrac{H(A)(1-H(A))}{1+\beta},\\ \mathrm{Corr}(G_j(A),G_{j'}(A)|H)=\dfrac{1}{1+\alpha}.\end{cases}$$

这表明在组间层次上, 当 $j=j'$ 时, $\mathrm{Corr}(X_{ij},X_{i'j'})=\frac{1}{1+\beta}$, 否则为 $\frac{1}{(1+\alpha)(1+\beta)}$. 这不仅揭示同一组内的 "类" 是有关联的, 而且也说明不同组间的 "类" 也存在关联性. Rodriguez 等 [84] 考虑了有限维折断近似, 并设计了 MCMC 抽样算法.

3 有关计算方法

在 DP 框架下, 非参数贝叶斯推断所遇到的一个重要挑战就是后验分布缺乏完整的解析形式. 这不仅表现在 P 无限维形式上, 更为困难的在于样本的联合分布退化. 后验推断往往需要借助于蒙特卡洛方法, 从后验分布抽取随机数来完成. 通常的抽样策略大致分为三种: 第一种就是避开 P, 也就是积分出 P, 直接从样本的联合分布出发, 利用样本的 Polya 性来完成后验抽样; 第二种采取近似的形式, 也就是对 P 进行折断, 用有限维分布来近似无限维分布; 第三种就是直接对 P 抽样, 但会遭受抽要速度慢, 效率低等困境.

3.1 MCMC 抽样

在 MCMC (Geman 和 Geman[36], Hastings[42], Metropolis 等 [71], Tanner 和 Wong[94]) 技术引入统计领域后 (Gelfand 和 Smith[35]), 非参数贝叶斯的统计计算得到大为改善 (见 Gelfand 和 Kuo [34]). 自此针对不同模型下的算法设计和改进成为研究热点和中心. 其核心内容是如何加快 MCMC 抽样的收敛速度, 提高抽样效率. 这方面的工作应该得益于 Escobar[24−25], Escobar 和 West[25], West, Müller 和 Escobar[98], MacEachern[64], Bush 和 MacEachern[8], MacEachern 和 Müller[68−69], MacEachern 等 [67], Liu[62], Neal[77], Dalal[14] 等大批研究者的工作. 为了一般性, 这里主要介绍 Escobar[24], Escobar 和 West[25], MacEachern[68], MacEachern 和 Müller[69] 等人的工作. 更详细的讨论见 Neal [77].

Escobar 在考虑正态分布的均值估计时假定

$$\begin{cases} Y_i|X_i,P,\ \text{ind.} \sim p(y_i|x_i), \\ X_i|P,\ \text{iid.} \sim P,\ P \sim D(cF_0). \end{cases} \tag{34}$$

根据定理 1.6 可知:

$$\mathbb{E}(X_i|Y_1,Y_2,\cdots,Y_n) = \frac{c}{n+c-1}\mathbb{E}_{F_0}X_i + \mathbb{E}\left[\frac{\sum_{j\neq i}\delta_{X_j}(X_i)}{n+c-1}\middle|Y_1,Y_2,\cdots,Y_n\right],$$

其依赖于 $X_1,X_2,\cdots,X_n$ 的后验分布 $H(\mathrm{d}x_1,\mathrm{d}x_2,\cdots,\mathrm{d}x_n|Y_1,Y_2\cdots,Y_n)$. 一种有效的方式是利用 MCMC 技术从该后验分布抽取随机样本: 对 $i=1,2,\cdots,n$,

$$\text{从 } p(X_i|Y_1,Y_2,\cdots,Y_n,X_j,j\neq i) \text{ 抽取 } X_i.$$

下面定理给出了 MCMC 抽样的满条件分布:

定理 3.1 (Escobar[24])

$$[X_i|Y_1,Y_2,\cdots,Y_n,X_j,j\neq i] \propto p(y_i|x_i)\frac{1}{c+n-1}\left[cF_0+\sum_{j\neq i}\delta_{x_j}\right](\mathrm{d}x_i)$$

$$\sim \begin{cases} X_j & \text{以概率 } \dfrac{p(y_i|x_j)}{c(y_i)+\sum_{j\neq i}p(y_i|x_j)}; \\ h(x_i|y_i) & \text{以概率 } \dfrac{c(y_i)}{c(y_i)+\sum_{j\neq i}p(y_i|x_j)}, \end{cases}$$

其中 $p(y_i|x_i)$ 为正态条件密度函数,

$$c(y_i) = c\int_{\mathcal{X}} p(y_i|x_i)F_0(\mathrm{d}x_i),\quad h(x_i|y_i) = \frac{p(y_i|x_i)F_0(\mathrm{d}x_i)}{c(y_i)}.$$

Escobar 和 West[25] 对上述算法做了改进, 其特点就是结合数据的不同值来加快收敛速度, 提高抽样效率. 记 $\{X_j : j \neq i\}$ 中不同值 $\{X_k^*\}$ 的总个数为 n_i^* $(< n)$, 且 X_k^* 的重复次数为 m_k, 则

$$[X_i|Y_1, Y_2, \cdots, Y_n, X_j, j \neq i] \propto p(y_i|x_i)\frac{1}{c+n-1}\left[cG_0 + \sum_{k=1}^{n_i^*} m_k \delta_{x_k^*}\right](\mathrm{d}x_i)$$

$$\sim \begin{cases} X_k^* & \text{以概率 } \dfrac{m_k p(y_i|x_k^*)}{c(y_i) + \sum_{k=1}^{n_i^*} m_k p(y_i|x_k^*)}; \\ h(x_i|y_i) & \text{以概率 } \dfrac{c(y_i)}{c(y_i) + \sum_{k=1}^{n^*} m_k p(y_i|x_k^*)}. \end{cases}$$

上述算法在某些情形下会遭遇以下困难: (1) 计算边际密度 $c(y_i)$ 的积分, 这除了一些标准分布外, 一般难以计算; (2) MCMC 抽样的收敛速度较慢, 这主要表现在 X_i^* 更新较慢, 抽样过程中 X_i^* 会 "粘" 在某些原子处. 对于问题 (1), West 等 [98] 建议用一步蒙特卡洛方法从 F_0 中抽取样本来近似积分; MacEachern 和 Müller[69] 给出了另外一种近似方法; 对于问题 (2), Bush 和 MacEachern[8] 建议每一次抽样后更新 X_k^*.

更有效的抽样涉及到样本类的结构. 引入分类变量 $K_i \in \mathbb{N}$ $(i = 1, 2, \cdots, n)$, 使得: $K_i = j$ 当且仅当 $X_i = X_j^*$. 则 $X_1, X_2, \cdots, X_n$ 完全由 $X^* = \{X_1^*, X_2^*, \cdots, X_{I^*}^*\}$ 以及 $K = \{K_1, K_2, \cdots, K_n\}$ 来确定, 其中 I^* 表示 $X_1, X_2, \cdots, X_n$ 中不同值的总个数. 更新 $X_1, X_2, \cdots, X_n$ 等价于更新 X^* 和 K. 记 $n_j = \#\{i : S_i = j\}$. 这些记号的含义不难从中餐馆规则得到理解.

记 $K_{-i} = \{K_j : j \neq i\}$, C_{-i} 表示 K_{-i} 不同值之集, $n_{-i,j}$ 表示 K_{-i} 中等于 j $(\in C_{-i})$ 的个数. 积分出 $\{\pi_k\}$, 由 Polya urn 规则可知

$$K_i|K_{-i} \sim \frac{c}{c+n-1}\delta_j \mathrm{I}\{j \bar{\in} C_{-i}\} + \frac{\sum_{j \in C_{-i}} n_{-i,j}\delta_j}{c+n-1}. \tag{35}$$

改进 Gibbs 抽样为执行如下:

步骤 1):

$$K_i|Y_1, Y_2 \cdots, Y_n, K_{-i}, X^* \sim \frac{cm(y_i)\delta_j \mathrm{I}\{j \bar{\in} C_{-i}\} + \sum_{j \in C_{-i}} n_{-i,j} p(y_i|x_j^*)\delta_j}{cm(y_i) + \sum_{j \in C_{-i}} n_{-i,j} p(y_i|x_j^*)}$$

其中 $m(y_i) = \int p(y_i|x)F_0(\mathrm{d}x)$.

步骤 2): 从 $p(X^*|Y_1, Y_2, \cdots, Y_n, K)$ 中抽取 X^*.

以上讨论限于将 P 积分出, 利用样本的 Polya urn 预测规则来完成. 那么能否可以从 P (或等价地对随机权和原子) 直接进行抽样? 答案是肯定的. 这方面主要得益 Walker[97] 和 Papaspiliopoulos 和 Roberts[78] 的工作. Walker 在考虑 DP 后验抽样时, 引入均匀随机变量 u_i, 使得

$$\begin{cases} (Y_i|K_i, X^*), \text{ ind.} \sim f(y_i|x_{K_i}^*), \\ ((K_i, u_i)|\{\pi_k\}), \text{ iid.} \sim \sum_{g=1}^{\infty} I\{u_i < \pi_k\}\delta_k, \\ X_k^*, \text{ iid.} \sim F_0, \\ \{\pi_k\} \sim \mathrm{GEM}(c). \end{cases} \tag{36}$$

后验分布的 MCMC 抽样主要是更新 $K=\{K_1,K_2,\cdots,K_n\}$, $X^*=\{X_k^*\}$, $\pi^*=\{\pi_k^*\}$, $U=\{u_1,u_2,\cdots,u_n\}$, 以及 α. 困难主要在于更新 K 及 U; 对于前者, Walker 注意到 (38) 中的无限项和实际上为有限项求和, 因此更新 K 相当直接; 而对于更新 U, 主要是利用 Gibbs 抽样器. Papaspiliopoulos 和 Roberts[78] 设计了一种 Retrospective MCMC 抽样算法, 不必考虑引入辅助变量 u_i, 而直接基于离散分布的追溯方法 (见 [83]) 来完成对 P 中的随机权和原子抽样.

3.2 分块 Gibss 抽样器

分块 Gibbs 抽样器 (blocked Gibbs sampler) 是由 Ishwaran 和 Zarepour[48], Ishwaran 和 James[43] 等在使用近似 DP 先验作为非参数贝叶斯先验时发展起来的一种贝叶斯抽样算法. 它的特点是不需要借助于 DP 先验下样本的 Polya urn 性质来获得 MCMC 抽样中的满条件分布, 而是改为对随机概率分布 P 直接抽样. 这种方法得益于 P 的有限维性.

考虑如下的分层模型

$$\begin{cases} Y_i|X_i,P,\text{ind.}\sim p(y_i|X_i),\\ X_i|P,\text{iid.}\sim P,\\ P\sim\sum_{g=1}^{G}\pi_g\delta_{Z_g}(\,\cdot\,). \end{cases} \tag{37}$$

定义指示变量 $K_i\in\{1,2,\cdots,G\}$, 使得 $X_i=Z_{K_i}$, 则模型 (38) 可改写为

$$\begin{cases} Y_i|X_i,K,Z,\text{ind.}\sim p(y_i|Z_{K_i}),\\ K_i|\pi\sim\sum_{g=1}^{G}\pi_g\delta_g(\,\cdot\,),\\ Z=\{Z_g\},\text{iid.},\sim p(Z),\ \boldsymbol{\pi}=(\pi_1,\pi_2,\cdots,\pi_G)\sim p(\boldsymbol{\pi}). \end{cases}$$

将 π,Z 看作参数, 则 (Z,π,K,Y) 的联合分布为

$$p(Y,Z,K,\pi)=\prod_{i=1}^{n}\prod_{g=1}^{G}\left(\pi_g p(y_i|Z_g)^{\{K_i=g\}}\right)p(Z)p(\pi).$$

分块 Gibbs 抽样执行如下:

1) 从 $p(\boldsymbol{\pi},Z|K,Y)$ 抽取 $\boldsymbol{\pi},Z$;

2) 从 $p(K|\boldsymbol{\pi},Z,Y)$ 抽取 K.

记 $K^*=\{K_1^*,K_2^*,\cdots,K_m^*\}$ $(m\le\min\{G,n\})$ 为 $K=\{K_1,K_2,\cdots,K_n\}$ 中的不同值之集. 将 Z 分为 Z_K^* 和 $Z_{(-K^*)}$. 下列定理给出了分块 Gibbs 抽样器各满条件分布.

定理 3.2 (Ishwaran 和 Zarepour[43−44,48]) 分块 Gibbs 抽样中的满条件分布分别为:

1)

$$\begin{cases} p(\boldsymbol{\pi}|K,Y)\propto p(\boldsymbol{\pi})p(K|\boldsymbol{\pi}),\\ p(Z_{K^*}|\boldsymbol{\pi},K,Y)\propto\prod_{j=1}^{m}p(Z_{K_j^*})\prod_{\{i:K_i=K_j^*\}}p(y_i|Z_{K_j^*}),\\ p(Z_{-K^*}|\boldsymbol{\pi},K,Y,\cdots)\propto\prod_{\{g:g\ne K_j^*\text{对某个}j\}}p(z_g); \end{cases}$$

2)

$$p(K|\boldsymbol{\pi},Z,Y)=\prod_{i=1}^{n}\left(\sum_{g=1}^{G}\pi_{ig}^*\delta_g\right),$$

其中 $\pi_{ig}^* = c_{ig}\pi_g p(y_i|z_g)$, c_{ig} 为正则常数使得 $\sum_{g=1}^{G}\pi_{ig}^* = 1$.

由 Dirichlet 分布的共轭性可知, 随机权 (见 [22]) 的后验分布仍然是 Dirichlet 分布. 如果随机权为折断的 stick-breaking 形式, 则其后验分布表现为如下广义 Dirichlet 分布: 令 $m_j = \#\{i : K_i = K_j^*\}$, 则折随机权 (22) 的后验分布为

$$\pi_1^* = V_1^*, \quad \cdots, \quad \pi_m^* = V_m^*\prod_{l=1}^{m-1}(1-V_l^*), \quad \cdots, \quad \pi_N^* = \prod_{l=1}^{N-1}(1-V_l^*),$$

其中 V_k^* $(k = 1, \cdots, N-1), \text{ind.} \sim \beta(1+m_k, c+\sum_{j=k+1}^{N} m_j)$.

4 DP 在潜变量模型中的应用

潜变量模型 (Bollen[7], Jöreskog 和 Sörbom[49], Skrondal 和 Rabe-Hesketh[89], Bentler 和 Wu[4], Kong[53], Lee[58], 宋心远等 [90]) 是以含有不可观测潜变量为特征的一大类多元统计模型. 这类模型包括通常的因子分析模型、结构方程模型、随机系数或截距模型、广义潜变量模型、变量误差模型、多层因子分析模型、潜在类模型等. 该模型通过引入较低维数的潜变量来刻画潜变量与观测变量之间的依赖关系, 并定量分析和解释潜变量之间的相互关系. 这对于探测观测变量变异源、压缩高维数据以及进行典型相关分析具有重要意义. 目前, 这类模型广泛应用于心理计量学、社会行为学、经济学、遗传学、生物医学等领域.

基于 DP 的贝叶斯非/半参数方法在潜变量模型中的研究和应用主要集中于随机参数和随机潜变量的分布拟合上的应用. 在正态随机效应模型方面, Kleinman 和 Ibrahim[51] 为了解决随机效应的异质性问题, 对随机效应的分布使用 DP 先验来建立统计模型:

$$\begin{aligned} &y_{ij}|b_i \sim N(X_{ij}b_i, \sigma^2), \quad i = 1, 2, \cdots, n, \ j = 1, 2, \cdots, n_i, \\ &b_i|P \sim P, \quad P \sim DP(cF_0), \end{aligned}$$

其中 X_{ij} 为协变量. Kleinman 和 Ibrahim[52] 还将上述问题延伸到广义线性随机效应模型; 然而在有限样本情形下, 这种处理方式会对截距参数的估计及估计精度构成影响, 造成估计不稳定. 事实上, $\mathbb{E}(b_i|P) = \sum \pi_k Z_k$, $\text{Var}(b_i|P) = \sum \pi_k Z_k^2 - (\sum \pi_k Z_k)^2$, 这会导致截距参数出现额外异质行为. Li 等 [61] 注意到这种现象, 提出了中心矫正方法; Müller 等 [76] 基于分布空间划分的方法对纵向随机效应模型建立了半参数贝叶斯分析. 在结构方程模型, Lee 等 [59] 使用近似 DP 先验对带有固定协变量的结构方程模型的外生潜变量的分布进行拟合; 宋心远等 [91] 将其推广到带有多项数据的非线性潜变量模型的外生变量的分布拟合, 并对 Lee 等 [59] 的工作作了改进; 宋心远等 [92] 考虑了半参数贝叶斯框架下模型选择和比较问题, 给出了经验证明; Yang 和 Dunson[105] 对结构方程模型半参数贝叶斯分布拟合作了改进; 唐安明和唐年胜 [93] 对纵向生存数据的随机效应建立了中心 DP 拟合; 在动力潜变量模型, Lennox 等 [60] 对潜马尔可夫模型建立了 MDP 分析, Chow 等 [10] 考虑了非线性动力因子模型的自回归参数的分布近似 DP 建模; 对于实证因子分析模型, 夏业茂和潘茂林 [104] 对因子分析模型的相关参数进行了半参数贝叶斯分析. 夏业茂等 [99, 101] 基于有限维 Dirichlet 分布对多水平因子模型、潜马尔可夫因子模型建立了半参数贝叶斯分析, 并进一步地研究了有限维 Dirichlet 过程的相关性质. 夏业茂和勾建伟 [100] 对带有次序数学的因子分析的相关参数建立了 Dirichlet 分析. 夏业茂和刘应安 [103] 还考虑了有限维 Dirichlet 过程下的贝叶斯因子的计算. 这些成果为开展潜变量模型在贝叶斯框架下非参数分析研究提供了良好的开端.

DP 在潜变量模型中的应用才刚刚起步, 大多是关注是单个 DP 过程在潜变量的应用. 这对于多水平、多组别、时序、空间等潜变量模型的分析尚不充分. 比如对不同空间中潜在因素的关联性以及纵向模型中潜在因素时间关联性的研究远不够充分. 这些全新的课题值得进一步的研究.

5 注记

DP 先验作为随机分布函数的一种先验, 在非参数贝叶斯分析领域获得成功的主要原因是由于其后验分布形式简洁、理论完善和使用方便. 特别是近年来 MCMC 技术的应用, 为非参数贝叶斯计算提供了技术支撑, 大大加快了 DP 的研究进展, 拓宽了其应用领域.

作为一个 "分布的分布", DP 先验也有它的局限性. 进行统计模型和分布进行拟合时要注意以下几点:

(1) DP 先验只是对离散分布进行直接拟合. 这从 DP 的样本实现值可以看得出来. 如果需要对连续特别是绝对连续的分布进行拟合, 可以考虑带有核函数的 DP 混合过程先验, 或者使用 Polya tree (Lavine[56−57], Hanson[41]) 先验.

(2) DP 下的样本实现值会呈现聚类特性. 这主要是缘于分布的离散性 (含有原子). 然而这种离散性对于挖掘数据潜在类具有重要作用, 特别是在解释参数或有关潜变量的异质性方面具有重要作用.

(3) 在利用 DP 对诸如残差变量、随机效应、因子变量分布进行拟合时必须谨慎. 因为这些变量的分布往往有特定的限制和要求, 比如要求分布的均值或中位数为 0. 这方面可以参见 Li 等 [61], Yang 和 Dunson[105], 宋心远等 [91], 唐安民和唐年胜 [93], 夏业茂和勾建伟 [102] 等人的工作. 另外, 需要特别指出的是, DP 过程下样本具有可交换性, 因此, 对于具有次序结构的变量的分布一般是不能直接进行 DP 拟合的.

(4) 在统计应用中, 似然函数扮演着重要的角色. 在非参数贝叶斯框架下, 似然函数有着复杂的解析形式. 如何寻求一种有效快速的算法来计算似然函数依然是一个挑战.

参考文献

[1] Aldous, D.J., Exchangeability and related topics, In: École d'Éte de Probabilités de Saint-Flour XIII–1983, Lecture Notes in Math., Vol. 1117, New York: Springer-Verlag, 1985, 23-34.

[2] Antoniak, C.E., Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, *Ann. Statist.*, 1974, 2(6): 1152-1174.

[3] Basu, S. and Chib, S., Marginal likelihood and Bayes factors for Dirichlet process mixture models, *J. Amer. Statist. Assoc.*, 2003, 98(461): 224-235.

[4] Bentler, P.M. and Wu, E.J.C., EQS6: Structural Equations Program Manual, Encino, CA: Multivariate Software, 2006.

[5] Blackwell, D., Discreteness of Ferguson selections, *Ann. Statist.*, 1973, 1(2): 356-358.

[6] Blackwell, D. and MacQueen, J.B., Ferguson distributions via polya urn schemes, *Ann. Statist.*, 1973, 1(2): 353-355.

[7] Bollen, K.A., Structural Equations with Latent Variables, New York: John Wiley & Sons, 1989.

[8] Bush, C.A. and MacEachern, S.N., A semiparametric Bayesian model for randomised block designs, *Biometrika*, 1996, 83(2): 275-285.

[9] Carota, C. and Parmigiani, G., Semiparametric regression for count data, *Biometrika*, 2002, 89(2): 265-281.

[10] Chow, S.M., Tang, N.S., Yuan, Y., Song, X.Y. and Zhu, H.T., Bayesian estimation of semiparametric nonlinear dynamic factor analysis models using the Dirichlet process prior, *Br. J. Math. Stat. Psychol.*, 2011, 64(1): 69-106.

[11] Cifarelli, D. and Regazzini, E., Problemi statistici non parametrici in condizioni di scambialbilità parziale: impiego di medie associative, Technical Report, Quad. Insitit. Mat. Finana. Univ. Torino III, 1978, 1-13 (in Italian).

[12] Connor, R.J. and Mosimann, J.E., Concepts of independence for proportions with a generalization of the Dirichlet distribution, *J. Amer. Statist. Assoc.*, 1969, 64(325): 194-206.

[13] Crandell, L.J. and Dunson, D.B., Posterior simulation across nonparametric models for functional clustering, *Sankhya B*, 2011, 73(1): 42-61.

[14] Dalal, S.R., Dirichlet invariant processes and applications to nonparametric estimation of symmetric distribution functions, *Stochastic Process. Appl.*, 1979, 9(1): 99-107.

[15] De Iorio, M., Müller, P., Rosner, G.L. and MacEacher, S.N., An ANOVA model for dependent random measures, *J. Amer. Statist. Assoc.*, 2004, 99(465): 205-215.

[16] Doss, H., Bayesian nonparametric estimation of the median: Part I. Computation of the estimates, *Ann. Statist.*, 1985, 13(4): 1432-1444.

[17] Doss, H., Bayesian nonparametric estimation of the median: Part II. Asymptotic properties of the estimates, *Ann. Statist.*, 1985, 13(4): 1445-1464.

[18] Doss, H., Bayesian nonparametric estimation for incomplete data via successive substitution sampling, *Ann. Statist.*, 1994, 22(4): 1763-1786.

[19] Duan, J.A., Guindani, M. and Gelfand, A.E., Generalized spatial Dirichlet process models, *Biometrika*, 2007, 94(4): 809-825.

[20] Dunson, D.B., Nonparametric Bayes local partition models for random effects, *Biometrika*, 2009, 96(2): 249-262.

[21] Dunson, D.B. and Park, J.H., Kernel stick-breaking processes, *Biometrika*, 2008, 95(2): 307-323.

[22] Dunson, D.B., Pillai, N. and Park, J.H., Bayesian density regression, *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 2007, 69(2): 163-183.

[23] Escobar, M.D., Estimating the means of several normal populations by estimating the distribution of the means, Ph.D. Thesis, New Haven: Yale Univ., 1988.

[24] Escobar, M.D., Estimating normal means with a Dirichlet process prior, *J. Amer. Statist. Assoc.*, 1994, 89(425): 268-277.

[25] Escobar, M.D. and West, M., Bayesian density estimation and inference using mixtures, *J. Amer. Statist. Assoc.*, 1995, 90(430): 577-588.

[26] Ewens, W.J., Population genetics theory—the past and the future, In: Mathematical and Statistical Developments of Evolutionary Theory, NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., Vol. 299, Dordrecht: Kluwer Acad. Publ., 1990, 177-227.

[27] Fabius, J., Asymptotic behavior of Bayes' estimates, *Ann. Math. Statist.*, 1964, 35(2): 846-856.

[28] Ferguson, T.S., A Bayesian analysis of some nonparametric problems, *Ann. Statist.*, 1973, 1(2): 209-230.

[29] Ferguson, T.S., Prior distributions on spaces of probability measures, *Ann. Statist.*, 1974, 2(4): 615-629.

[30] Fong, D.K.H., Pammer, S.E., Arnold, S.F. and Bolton, G.E., Reanalyzing ultimatum bargaining: comparing nondecreasing curves without shape constraints, *J. Busin. Econom. Statist.*, 2002, 20(3): 423-430.

[31] Freedman, D.A., On the asymptotic behavior of Bayes' estimates in the discrete case II, *Ann. Math. Statist.*, 1963, 34(4): 1386-1403.

[32] Gelfand, A.E. and Kottas, A., A computational approach for full nonparametric Bayesian inference under Dirichlet Process mixture models, *J. Comput. Graph. Stat.*, 2002, 11(2): 289-305.

[33] Gelfand, A.E., Kottas, A. and MacEachern, S.N., Bayesian nonparametric spatial modeling with Dirichlet process mixing, *J. Amer. Statist. Assoc.*, 2005, 100(471): 1021-1035.

[34] Gelfand, A.E. and Kuo, L., Nonparametric Bayesian bioassay including ordered polytomous response, *Biometrika*, 1991, 78(3): 657-666.

[35] Gelfand, A.E. and Smith, A.F.M., Sampling-based approaches to calculating marginal densities, *J. Amer. Statist. Assoc.*, 1990, 85(410): 398-409.

[36] Geman, S. and Geman, D., Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *Trans. Pattern Anal. Mach. Intell.*, 1984, PAMI-6(6): 721-741.

[37] Ghosh, J.K. and Ramamoorthi, R.V., Bayesian Nonparametrics, New York: Springer-Verlag, 2003.

[38] Giudici, P., Mezzetti, M. and Muliere, P., Mixtures of products of Dirichlet processes for variable selection in survival analysis, *J. Statist. Plann. Inference*, 2003, 111(1/2): 101-115.

[39] Griffin, J.E. and Steel, M.F.J., Order-based dependent Dirichlet processes, *J. Amer. Statist. Assoc.*, 2006, 101(473): 179-194.

[40] Halmos, P.R., Random alms, *Ann. Math. Statist.*, 1944, 15(2): 182-189.

[41] Hanson, T.E., Inference for mixtures of finite Polya tree models, *J. Amer. Statist. Assoc.*, 2006, 101(476): 1548-1565.
[42] Hastings, W.K., Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, 1970, 57(1): 97-109.
[43] Ishwaran, H. and James, L.F., Gibbs sampling methods for stick-breaking priors, *J. Amer. Statist. Assoc.*, 2001, 96(453): 161-173.
[44] Ishwaran, H. and James, L.F., Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information, *J. Comput. Graph. Stat.*, 2002, 11(3): 508-532.
[45] Ishwaran, H. and James, L.F., Generalized weighted Chinese restaurant processes for species sampling mixture models, *Statist. Sin.*, 2003, 13(4): 1211-1235.
[46] Ishwaran, H. and James, L.F., Computational methods for multiplicative intensity models using weighted Gamma process: proportional hazards, marked point processes, and panel count data, *J. Amer. Statist. Assoc.*, 2004, 99(465): 175-190.
[47] Ishwaran, H. and Takahara, G., Independent and identically distributed Monte Carlo algorithms for semiparametric linear mixed models, *J. Amer. Statist. Assoc.*, 2002, 97(460): 1154-1166.
[48] Ishwaran, H. and Zarepour, M., Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models, *Biometrika*, 2000, 87(2): 371-390.
[49] Jöreskog, K. and Sörbom, D., LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language, Hove and London: Scientific Software International, 1996.
[50] Kingman, J.F.C., Taylor, S.J., Hawkes, A.G., Walker, A.M., Cox, D.R., Smith, A.F.M., Hill, B.M., Burville, P.J. and Leonard, T., Random discrete distributions, *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 1975, 37: 1-22.
[51] Kleinman, K.P. and Ibrahim, J.G., A semiparametric Bayesian approach to the random effects model, *Biometrics*, 1998, 54(3): 921-938.
[52] Kleinman, K.P. and Ibrahim, J.G., A semi-parametric Bayesian approach to generalized linear mixed models, *Statist. Med.*, 1998, 17(22): 2579-2596.
[53] Kong, A., Liu, J.S. and Wong, W.H., Sequential imputations and Bayesian missing data problems, *J. Amer. Statist. Assoc.*, 1994, 89(425): 278-288.
[54] Korwar, R.M. and Hollander, M., Contributions to the theory of Dirichlet processes, *Ann. Probab.*, 1973, 1(4): 705-711.
[55] Kuo, L., Computations of mixtures of Dirichlet processes, *SIAM J. Sci. Stat. Comput.*, 1986, 7(1): 60-71.
[56] Lavine, M., Some aspects of Polya tree distributions for statistical modelling, *Ann. Statist.*, 1992, 20(3): 1222-1235.
[57] Lavine, M., More aspects of Polya tree distributions for statistical modelling, *Ann. Statist.*, 1994, 22(3): 1161-1176.
[58] Lee, S.Y., Structural Equation Modeling: A Bayesian Approach, Chichester: John Wiley & Sons, 2007.
[59] Lee, S.Y., Lu, B. and Song, X.Y., Semiparametric Bayesian analysis of structural equation models with fixed covariates, *Statist. Med.*, 2008, 27(13): 2341-2360.
[60] Lennox, K.P., Dahl, D.B., Vannucci, M., Day, R. and Tsai, J.W., A Dirichlet process mixture of hidden Markov Models for protein structure prediction, *Ann. Appl. Stat.*, 2010, 4(2): 916-942.
[61] Li, Y.S., Lin, X.H. and Müller, P., Bayesian inference in semiparametric mixed models for longitudinal data, *Biometrics*, 2010, 66(1): 70-78.
[62] Liu, J.S., Nonparametric hierarchical Bayes via sequential imputations, *Ann. Statist.*, 1996, 24(3): 911-930.
[63] Lo, A.Y., On a class of Bayesian nonparametric estimates: I. Density estimates, *Ann. Statist.*, 1984, 12(1): 351-357.
[64] MacEachern, S.N., Estimating normal means with a conjugate style Dirichlet process prior, *Comm. Stat. Simulat. Comput.*, 1994, 23(3): 727-741.
[65] MacEachern, S.N., Dependent Dirichlet processes, In: ASA Proceedings of the Section on Bayesian Statistical Science, Alexandria, VA: Amer. Statist. Assoc., 1999: 50-55.
[66] MacEachern, S.N., Decision theoretic aspects of dependent nonparametric processes, In: Bayesian Methods with Applications to Science, Policy and Official Statistics, Crete: International Society for Bayesian Analysis, 2000: 551-560.
[67] MacEachern, S.N., Clyde, M. and Liu, J.S., Sequential importance sampling for nonparametric Bayes models: The next generation, *Canad. J. Statist.*, 1999, 27(2): 251-267.

[68] MacEachern, S.N. and Müller, P., Estimating mixture of Dirichlet process models, *J. Comput. Graph. Stat.*, 1998, 7(2): 223-238.

[69] MacEachern, S.N. and Müller, P., Efficient MCMC schemes for robust model extensions using encompassing Dirichlet process mixture models, In: Robust Bayesian Analysis, Lecture Notes in Statist., Vol. 152, New York: Springer-Verlag, 2000: 295-315.

[70] McCloskey, J.W., A model for the distribution of individuals by species in an environment, Ph.D. Thesis, East Lansing, MI: Michigan State Univ., 1965.

[71] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E., Equation of state calculations by fast computing machines, *J. Chem. Phys.*, 1953, 21(6): 1087-1092.

[72] Mira, A. and Petrone, S., Bayesian hierarchical non-parametric inference for change-point problems, In: Bayesian Statistics 5, Oxford: Oxford Univ. Press, 1996: 693-703.

[73] Muliere, P. and Petrone, S., A Bayesian predictive approach to sequential search for an optimal dose: parametric and nonparametric models, *J. Ital. Statist. Soc.*, 1993, 2(3): 349-364.

[74] Muliere, P. and Tardella, L., Approximating distributions of random functionals of Ferguson-Dirichlet priors, *Canadian J. Statist.*, 1998, 26(2): 283-297.

[75] Müller, P., Quintana, F. and Rosner, G., A method for combining inference across related nonparametric Bayesian models, *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 2004, 66(3): 735-749.

[76] Müller, P., Quintana, F.A., Rosner, G.L. and Maitland, M.L., Bayesian inference for longitudinal data with non-parametric treatment effects, *Biostatistics*, 2014, 15(2): 341-352.

[77] Neal, R.M., Markov chain sampling methods for Dirichlet process mixture models, *J. Comput. Graph. Statist.*, 2000, 9(2): 249-265.

[78] Papaspiliopoulos, O. and Roberts, G.O., Retrospective Markov Chain Monte Carlo methods for Dirichlet process hierarchical models, *Biometrika*, 2008, 95(1): 169-186.

[79] Petrone, S., Guindani, M. and Gelfand, A.E., Hybrid dirichlet mixture models for functional data, *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 2009, 71(4): 755-782.

[80] Pitman, J., Some developments of the Blackwell-MacQueen urn scheme, In: Statistics, Probability and Game Theory, Papers in honor of David Blackwell, Hayward, CA: IMS, 1996: 245-267.

[81] Pitman, J., Random discrete distributions invariant under size-biased permutation, *Adv. Appl. Probab.*, 1996, 28(2): 525-539.

[82] Reich, B.J. and Fuentes, M., A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields, *Ann. Appl. Stat.*, 2007, 1(1): 249-264.

[83] Ripley, B.D., Stochastic Simulation, Chichester: John Wiley & Sons, 1987.

[84] Rodríguez, A., Dunson, D.B. and Gelfand, A.E., The nested Dirichlet process, *J. Amer. Statist. Assoc.*, 2008, 103(483): 1131-1154.

[85] Rodriguez, A., Dunson, D.B. and Gelfand, A.E., Bayesian nonparametric functional data analysis through density estimation, *Biometrika*, 2009, 96(1): 149-162.

[86] Scarpa, B. and Dunson, D.B., Enriched stick-breaking processes for functional data, *J. Amer. Statist. Assoc.*, 2014, 109(506): 647-660.

[87] Sethuraman, J., A constructive definition of Dirichlet priors, *Statist. Sin.*, 1994, 4(2): 639-650.

[88] Sethuraman, J. and Tiwari, R.C., Convergence of Dirichlet measures and the interpretation of their parameters, In: Statistical Decision Theory and Related Topics III, New York: Academic Press, 1982: 305-316.

[89] Skrondal, A. and Rabe-Hesketh, S., Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models, New York: Chapman & Hall/CRC, 2004.

[90] Song, X.Y. and Lee, S.Y., Basic and Advanced Bayesian Structural Equation Modeling: With Applications in the Medical and Behavioral Sciences, New York: John Wiley & Sons, 2012.

[91] Song, X.Y., Xia, Y.M. and Lee, S.Y., Bayesian semiparametric analysis of structural equation models with mixed continuous and unordered categorical variables, *Statist. Med.*, 2009, 28(17): 2253-2276.

[92] Song, X.Y., Xia, Y.M., Pan, J.H. and Lee, S.Y., Model comparison of Bayesian semiparametric and parametric structural equation models, *Struct. Equat. Model.*, 2011, 18(1): 55-72.

[93] Tang, A.M. and Tang, N.S., Semiparametric Bayesian inference on skew-normal joint modeling of multivariate longitudinal and survival data, *Statist. Med.*, 2015, 34(5): 824-843.

[94] Tanner, M.A. and Wong, W.H., The calculation of posterior distributions by data augmentation, *J. Amer. Statist. Assoc.*, 1987, 82(398): 528-540.

[95] Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M., Hierarchical Dirichlet processes, *J. Amer. Statist. Assoc.*, 2006, 101(476): 1566-1581.
[96] Tomlinson, G. and Escobar, M., Analysis of densities, Technical Report, Toronto: University of Toronto, 1999.
[97] Walker, S.G., Sampling the Dirichlet mixture model with slices, *Comm. Statist. Simulation Comput.*, 2007, 36(1): 45-54.
[98] West, M., Müller, P. and Escobar, M.D., Hierarchical priors and mixtures models, with applications in regression and density estimates, In: Aspects of Uncertainty, A Tribute to D. V. Lindley, London: John Wiley & Sons, 1994: 363-386.
[99] Xia, Y.M. and Gou, J.W., Assessing heterogeneity in multilevel factor analysis model: A semiparametric Bayesian approach, *Acta Math. Sin.*, 2015, 38(4): 751-768 (in Chinese).
[100] Xia, Y.M. and Gou, J.W., A note on the finite-dimensional Dirichlet prior, *Comm. Statist. Theory Methods*, 2017, 46(19): 9388-9396.
[101] Xia, Y.M., Gou, J.W. and Liu, Y.A., Semi-parametric Bayesian analysis for factor analysis model mixed with hidden Markov model, *Appl. Math. J. Chinese Univ. Ser. A*, 2015, 30(1): 17-30 (in Chinese).
[102] Xia, Y.M. and Gou, J.W., Bayesian semiparametric analysis for latent variable models with mixed continuous and ordinal outcomes, *J. Korean Statist. Soc.*, 2016, 45(3): 451-465.
[103] Xia, Y.M. and Liu, Y.A., Bayesian semiparametric analysis and model comparison for confirmatory factor model, *Chinese J. Appl. Probab. Statist.*, 2016, 32(2): 157-183.
[104] Xia, Y.M. and Pan, M.L., Bayesian analysis for confirmatory factor model with finite-dimensional Dirichlet prior mixing, *Comm. Statist. Theory Methods*, 2017, 46(9): 4599-4619.
[105] Yang, M.G. and Dunson, D.B., Bayesian semiparametric structural equation models with latent variables, *Psychometrika*, 2010, 75(4): 675-693.

Dirichlet Process and Its Recent Developments

XIA Yemao[1,2], LIU Ying'an[1]

(*1. School of Science, Nanjing Forestry University, Nanjing, Jiangsu, 210037, P. R. China; 2. College of Economics and Management, Nanjing Forestry University, Nanjing, Jiangsu, 210037, P. R. China*)

Abstract: The core of the nonparametric Bayesian analysis is to treat the distribution of parameters and/or latent variables of interest to be random and assign a prior. As a distribution of distribution, Dirichlet process perhaps is the most popular prior within nonparametric Bayesian analysis and has received wide attention. In this paper, we synthesize and review some developments of Dirichlet process during the past decades, and present some new applications within the framework of latent variable models.

Keywords: nonparametric Bayes; Dirichlet processes; Polya urn sampling; Sethurman representation; mixture of Dirichlet process; dependence Dirichlet process; Markov Chains Monte Carlo; blocked Gibbs sampler