arXiv:2303.02438v2 [stat.ME] 1 Jun 2024

通过潜在排斥混合模型对高维数据进行贝叶斯聚类

Lorenzo Ghilotti[1], Mario Beraha[2],和 Alessandra Guglielmi[2]

[1]意大利米兰比可卡大学经济、管理与统计系[2]意大利米兰理工大学数学系

摘要

对中等维度或高维数据进行基于模型的聚类向来十分困难。我们提出一种模型，通过假设一组潜在得分服从混合模型，并利用高斯潜在因子模型将其与观测值联系起来，从而同时实现降维和聚类。Chandra 等人（2023）近期研究了这一方法。作者采用因子解析表示，并假设潜在因子服从混合模型。然而，在存在模型设定错误时，模型性能可能会下降。研究表明，为潜在得分混合模型中的成分特异性均值设置排斥点过程先验，可得到更稳健的模型；在若干模拟场景中，该模型优于针对潜在因子的标准混合模型。排斥点过程必须具有各向异性，才能促使数据簇充分分离；同时，其密度应易于计算，以便进行高效的后验推断。为解决这些问题，我们提出了一种构造各向异性行列式点过程的一般方法。我们通过模拟以及一个植物物种共现数据集展示了该模型的应用。

关键词：各向异性点过程；行列式点过程；高斯因子模型；马尔可夫链蒙特卡洛；基于模型的聚类。

1 引言

本文探讨中等或高维数据的贝叶斯聚类分析。具体而言，我们关注观测值 y_1，……，$y_n \in \mathbb{R}^p$。尽管文献中已提出多种模型（Neal，2003年；Teh 等人，2007年；Duan 和 Dunson，2021年；Natarajan 等人，2023年），贝叶斯混合模型是基于模型的聚类的一种直接方法；近期综述参见 Fruhwirth-Schnatter 等人（2019年）。在混合模型中，假设数据由 m （随机或固定的）同质总体生成。通常，假设每个总体都可以通过某个参数 $\theta \in \Theta$ 对应的参数化密度 $f_\theta(\cdot)$ 进行恰当建模。权重 $\boldsymbol{w} = (w_1, \ldots\ldots, w_m)$（$w_h \geq 0$,

$\sum_{h=1}^{m} w_h = 1$）表示各总体的相对频率。总之，在混合模型下，给定参数时数据的条件分布形式为

$$y_1, \ldots, y_n \,|\, \boldsymbol{w}, \boldsymbol{\theta} \overset{\text{iid}}{\sim} \sum_{h=1}^{m} w_h f_{\theta_h}(\cdot) \tag{1}$$

在贝叶斯方法下，需要为 $\boldsymbol{w}$、 $\boldsymbol{\theta}$ = （θ_1、.....、 θ_m）和m设定合适的先验。

当数据维度 p 大于一个适中的数值（例如 $p > 10$）时，贝叶斯混合模型表现不佳，这是一个长期存在的问题。这部分是由于用于后验推断的算法可扩展性较差（参见 Malsiner-Walli 等人，2016年；Celeux 等人，2019年），另一部分则是由于 p 较大时模型的渐近性质所致。具体而言，Chandra 等人（2023）中的定理1以及推论1和推论2表明，当(1)中的 f_θ 为高斯分布这一事实上的标准分布时，聚类估计会退化。对于固定的 n，随着 $p \to +\infty$，如果协方差矩阵是特定于簇的，则以（后验）概率1，所有观测值都会被聚类为单元素簇；而如果所有簇共享同一个协方差矩阵，则只能检测到一个簇。

从业者最常采用的高维数据聚类方法遵循两步法：首先拟合潜因子模型（Lopes，2014年），为每个观测值关联一个 d维得分 η_i，其中 $d \ll p$。然后，对这些 η_i应用传统聚类算法。然而，这种两步法无法将降维步骤所引入的不确定性传递到聚类估计中。要克服这一关键局限，自然的选择是建立一个同时进行降维和聚类的模型：假设潜在得分服从混合模型，再通过高斯潜在因子模型将其与观测值关联起来。Chandra 等人（2023）最近在贝叶斯非参数框架下研究了这一方法。

然而，混合模型可能会高估为恢复“真实”抽样模型所需的簇数。事实上，贝叶斯混合模型的典型假设是，式（1）中的参数 θ_h 独立同分布地来自某个基分布。在这一假设下，如果真实的数据生成密度与式（1）不一致，那么随着 n 增加，后验簇数将趋于无穷（Cai 等人，2021年）。高斯潜在因子模型与潜在得分的标准混合模型组合时，存在两种模型设定错误：首先，因子分析表示意味着数据接近一个 d维超平面；其次，数据偏离该超平面的部分服从高斯分布。这两项假设都值得质疑，在实践中也不太可能成立。此外，Chandra 等人（2023）假设潜在得分服从狄利克雷过程（DP）混合先验，即使对于设定正确的模型，也可能产生不一致估计（Miller 与 Harrison，2014年）。Ascolani 等人（2022年）在狄利克雷过程混合模型的一致性方面取得了令人鼓舞的进展，但其假设并不适用于 Chandra 等人（2023）的模型。

本文提出 APPLAM：各向异性（排斥性）点过程潜在混合模型。我们考虑一种同时实现降维和聚类的模型，与上述模型类似，但假设以排斥点过程作为先验，用于潜在得分混合分布的成分特定位置参数。由 Petralia等（2012年）开创的排斥混合模型，为混合模型中独立同分布的成分参数缺乏稳健性这一问题提供了实用的解决方案

。另见 Xu 等（2016年）；Xie 和 Xu（2019年）；Fúquene 等（2019年）；Bianchini 等（2020年）；Beraha 等（2022年）。在这些研究中，排斥先验用于对成分特定位置参数进行联合建模，其中一些研究还对其数量进行建模。然而，在潜在混合模型的背景下，我们认为，要使数据簇充分分离，仅在潜在层级形成充分分离的簇并不足够；排斥作用还应考虑将潜变量与观测值联系起来的因子分析模型。为此，我们提出以各向异性行列式点过程（DPP）作为成分特定参数的先验，其中因子载荷矩阵决定各向异性。我们推导出一种各向异性 DPP 的一般构造方法，可产生所需的排斥作用。

1.1 基于潜在混合分布的贝叶斯聚类

令 $\boldsymbol{y} = (y_1, \ldots\ldots, y_n)$、$y_i \in \mathbb{R}^p$、$\Lambda \in \mathbb{R}^{p \times d}$ 为因子载荷矩阵，$\eta_1, \ldots\ldots, \eta_n \in \mathbb{R}^d$ 为一组潜在因子，$\Sigma = \text{diag}(\sigma_1^2, \ldots\ldots, \sigma_p^2)$ $(\sigma_j > 0)$为对角协方差矩阵。令 $\mathcal{N}_p(\boldsymbol{a}, B)$表示均值为 $\boldsymbol{a}$ 、协方差矩阵为 B的 p维高斯分布。潜在因子混合模型假设

$$
\begin{aligned}
y_i \,|\, \eta_i, \Lambda, \Sigma &\overset{\text{ind}}{\sim} \mathcal{N}_p(\Lambda \eta_i, \Sigma), \\
\eta_i \,|\, \boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\Delta} &\overset{\text{iid}}{\sim} \sum_{h=1}^{m} w_h \mathcal{N}_d(\mu_h, \Delta_h),
\end{aligned}
\tag{2}
$$

对于 $i = 1, \ldots\ldots, n$。将潜在因子和观测值的混合模型扩展到其他核分布也很直接。为 $\boldsymbol{w} = (w_1, \ldots\ldots, w_m)$、$\boldsymbol{\mu} = (\mu_1, \ldots\ldots, \mu_m)$、$\boldsymbol{\Delta} = (\Delta_1, \ldots\ldots, \Delta_m)$、$\Lambda$ 以及 σ_j^2赋予先验后，混合模型便完整了。对于我们在此提出的排斥混合模型，m 也是随机的，因此混合成分的数量由数据学习得到。先验的完整描述见第 2.2 节。

引入一组潜在簇指示变量 c_i，使得 $P(c_i = h \,|\, \boldsymbol{w}) = w_h$，我们可以将式（2）中 η_i 的先验等价地表述为 $\eta_i \,|\, c_i = h,\ \boldsymbol{\mu},\ \boldsymbol{\Delta} \overset{\text{ind}}{\sim} \mathcal{N}_d(\mu_h,\ \Delta_h)$，$i = 1, \ldots\ldots, n$。因此，正如贝叶斯混合模型中的标准做法，我们通过潜变量 η_i 对数据 y_i 进行聚类；也就是说，如果 $c_i = c_j$，则 y_i 和 y_j 属于同一个簇。通过 c_i，我们可以将这些簇与已分配分量对应起来，即对于那些满足存在 i 使得 $c_i = h$ 的成分 $h \in \{1, \ldots, m\}$。

当属于不同簇的观测值彼此充分分离时，簇估计具有可解释性，因此也很有用。排斥混合模型通过为簇中心 μ_h's 假设一种偏好规则（即彼此充分分离的）点配置的先验，促使形成彼此分离的簇。定义此类先验的一种直接方法，是假设一个排斥点过程，同时控制成分的基数 m 以及其位置 μ_h's。具体而言，可以通过指定相对于泊松点过程的密度来定义此类过程。例如，Quinlan 等（2021）；Xie 和 Xu（2019）假设一种成对相互作用过程，其相对于一个

适当定义的泊松过程，是

$$p(\{\mu_1,\ldots,\mu_m\}) = \frac{1}{Z}\prod_{j=1}^{m}\phi_1(\mu_j)\prod_{1\le h<k\le m}\phi_2(\|\mu_h-\mu_k\|) \tag{3}$$

其中，ϕ_1 是有界函数，ϕ_2 是非递减函数，而 Z 是通常难以求解的归一化常数。关于相对于泊松点过程的密度定义，参见 Daley & Vere-Jones（2008年）和 Møller & Waagepetersen（2003年）。

这一先验选择可以确保不同簇与彼此分离良好的潜在得分 η_i' s 相关联，但这足以确保数据y_i's形成彼此分离良好的簇吗？根据高斯分布的性质，我们有

$$\{y_i : c_i = h\} \mid \boldsymbol{c}, \boldsymbol{\mu}, \boldsymbol{\Delta}, \boldsymbol{\Lambda} \overset{\text{iid}}{\sim} \mathcal{N}_p(\Lambda\mu_h, \Lambda\Delta_h\Lambda^\top + \Sigma).$$

因此，显然，仅仅在先验上鼓励 $\|\mu_h - \mu_k\|$ 较大，并不足以使数据点形成彼此分离良好的簇，因为簇中心之间的距离为$\|\Lambda\mu_h - \Lambda\mu_k\|$。正如附录 S5.3 所述，我们的数值实验也证实了这一点：若未能在簇中心之间引入适当的排斥，就会导致后验聚类估计极差。当然，若要在成对相互作用点过程中使 $\Lambda\mu_h$'s之间产生排斥，我们可以轻易地修改~(3)，考虑 $\phi_2(x) \equiv \phi_2(\|\Lambda x\|)$；此时归一化常数 Z 取决于ϕ_1，ϕ_2，最重要的是 Λ。式 (3) 中归一化常数的不可解性给后验模拟带来了重大挑战。Beraha 等（2022）讨论了如何使用交换算法从式 (3) 中涉及的参数的后验分布进行采样（Møller 等人，2006年；Murray 等人，2006年），该算法需要从式 (3) 中进行完美采样。不过，Beraha 等（2022）只考察了该算法在对单个实值参数进行采样时的效率。而在本文中，点过程密度确实依赖于 Λ；我们的初步研究表明，通过交换算法更新 Λ 会导致极差的混合，这是因为矩阵Λ 本身具有高维性。

1.2 我们的贡献

我们在第 2.1 节提出一种各向异性行列式点过程（DPP），作为分量特定的潜在因子中心 $\{\mu_1,\ldots,\mu_m\}$ 的先验，其中因子载荷矩阵决定各向异性。作为首要贡献之一，我们推导出各向异性 DPP 的一般构造方法，并为这一类 DPP 提出新的存在条件，这些条件与 Lavancier 等（2015）提出的条件相似，而且相应的密度不涉及难以处理的项。我们从解析上证明，各向异性 DPP 会产生所需的排斥效应。事实上，我们证明，对于位于由 Λ 的列张成的超平面上的点 $\{\Lambda\mu_1,\ldots,\Lambda\mu_m\}$，我们的过程等价于一个标准（各向同性）DPP 分布。原则上，虽然可以直接在这样的超平面上定义标准 DPP，但由此得到的过程在解析上难以处理（需要对定义在随机超平面上的核函数进行特征分解），因此实际用途有限。我们还给出了 DPP 谱密度的显式表达式，这对模拟至关重要。

这一步还需要对 DPP 密度表达式中的两个无穷级数进行标准近似（例如参见 Lavancier 等，2015）。我们通过实证表明，近似误差可以忽略不计。

我们的第二项贡献是将各向异性 DPP 纳入潜在因子混合模型 (2)，作为位置 μ_h 及其基数的先验。我们讨论了所提出模型的若干方面，尤其关注控制该过程排斥性的参数所发挥的作用。我们还提供了通过依赖数据的程序设定此类参数的指南。后验推断带来了不小的计算挑战，我们通过提出一种 Metropolis-within-Gibbs 算法来应对。具体而言，Λ 的完整条件分布采用广为人知的 Metropolis 调整的 Langevin 算法进行更新；为提高计算效率，该算法需要 DPP 密度（对数）的梯度解析表达式，我们在理论结果中给出了这一表达式。我们将模型应用于真实数据集和模拟数据集，以展示其适用性。具体来说，我们考察了博日自然区域公园（法国）不同地点的植物物种出现情况数据（$p = 123$）。我们的目标是推断这些地点的聚类结构（$n = 1139$），使得属于同一簇的地点在植物物种出现模式上表现相似。我们还通过针对高维（p 最高达 1000，这对于传统聚类方法而言极为庞大）和大型（$n = 1000$）数据集开展广泛的模拟研究，进一步深入考察了模型。除了说明采用各向异性排斥先验的必要性之外，我们还将 APPLAM 模型与 Chandra 等人（2023）提出的 LAMB 模型进行了比较。我们发现，在存在模型错设时，APPLAM 能够产生更稳健的聚类估计。

2 方法论

2.1 各向异性 DPP 的一般构造

定义在 Φ（$\mathbb{R}^d$，$\mathcal{B}$（$\mathbb{R}^d$））上的行列式点过程，其中 $\mathcal{B}$（$\mathbb{R}^d$）是博雷尔σ代数，是一个随机子集 $\{\mu_1, \ldots, \mu_m\} \subset \mathbb{R}^d$。参见 Macchi（1975）；Lavancier 等（2015）；Baccelli 等人（2020）。DPP 的概率分布完全由连续的复值协方差核 K 刻画：$\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$，其形式取决于$m$阶乘矩测度。对于不熟悉这一概念的读者而言，这样的一般定义可能显得较为繁琐，相关内容见附录 S1.1。为使讨论更加简洁，本文重点考虑限制在紧区域 $R \subset \mathbb{R}^d$上的 DPP，从而可以写出其相对于泊松过程的密度。不过，我们的结果同样适用于定义在整个 $\mathbb{R}^d$上的一般 DPP。根据 Mercer 定理，将 K 限制在 $R \times R$ 上后，可得到如下谱表示：$K(x, y) = \sum_{j \geq 1} \gamma_j \xi_j(x) \xi_j(y)$，其中 ξ_j构成复值函数 $L^2(R; \mathbb{C})$ 的正交归一基（特征函数），而 γ_j是一个可求和的非负序列，即特征值；ξ_j 表示ξ_j的共轭值。特别地，当限制在 R上时，对于所有 $j = 1$，2、.....，满足 $\gamma_j < 1$ 的 DPP 具有

相对于 R 上的单位速率泊松过程的密度如下:

$$p(\{\mu_1,\ldots,\mu_m\}) = \mathrm{e}^{|R|-D}\det\{C(\mu_h,\mu_k)\}_{h,k=1,\ldots,m},\ \mu_1,\ldots,\mu_m \in R,\ m=0,1,2,\ldots \tag{4}$$

其中 $C(x,\ y) = \sum_{j\geq 1}\gamma_j/(1-\gamma_j)\_\xi_j(x)\ \xi_j(y)$，$|R| = \int_R \mathrm{d}x$、$D = -\sum_{j\geq 1}\log(1-\gamma_j)$，并约定当 $m=0$时，$\det\{C(\mu_h,\ \mu_k)\}_{h,\ ,\ \ldots\ldots\ m} = 1$。此类结果的证明见 Lavancier 等（2015）。正如 Lavancier 等（2015）所指出的，C的连续性意味着，当欧几里得距离 $\|\mu_h-\mu_k\|$趋于零时，对于某个 $h\neq k$，行列式（4）趋于零，这表明 DPP 是一种排斥点过程。DPP 的概率分布由其在 $\mathbb{R}^d$上的核 $K(\cdot)$ 唯一确定，或者等价地由其谱表示唯一确定。与表达式（3）相比，DPP 的密度具有解析可处理的归一化常数 $\mathrm{e}^{|R|-D}$，这显著简化了对该过程超参数的后验推断；例如，Beraha 等（2022）讨论了不可处理归一化常数所带来的计算复杂度。另一方面，要使用 DPP 密度（4），就需要将点过程限制在紧区域 R上，而这是一个成对相互作用点过程所不具有的不便之处。然而，正如第2.3节所讨论的，R 的选择在我们的模型中所起的作用有限。

特征值 γ_j' 的解析表达式对于推断至关重要。沿用 Lavancier 等（2015）、Bianchini 等（2020）和 Beraha 等（2022）提出的所谓“谱方法”，他们假设 $K(x,\ y) = K_0(x-y) = K_0(\|x-y\|)$，即 K 是一个平稳且各向同性的函数。他们不对 K 建模，而是将 ξ_j' 固定为傅里叶基，并假设 γ_j' 服从某个参数模型。这种方法确保了 K 的正定性以及 DPP 密度的存在，但对于混合模型（2）而言不够灵活。尤其是，DPP 核 K 的各向同性与我们希望在各个 $\Lambda\mu_h$ 之间施加排斥的目标相冲突。接下来，我们将给出一种适用于平稳各向异性 DPP 的一般构造、其核 K_0 的傅里叶变换的显式表达式，以及保证 DPP 存在的易于检验的条件。

定理 2.1。设 Λ 为一个（固定的）$p\times d$ 满秩实矩阵。设 W 为一个严格正的随机变量，并令 $h(y)$ 为随机变量 Y的边缘密度，其定义如下:

$$Y \mid W \sim \mathcal{N}_d(0, W(\Lambda^T\Lambda)^{-1}) \tag{5}$$

若 $K_0(x) = \rho h(x)/h(0)$ 对 $x\in\mathbb{R}^d$ 和 $\rho>0$成立，则存在一个定义在 $\mathbb{R}^d$ 上的行列式点过程 Φ，其核为 $K(x,y) = K_0(x-y)$，其中 $\rho\leq\rho_{\max}$ 定义如下:

$$\rho_{\max} = |\Lambda^T\Lambda|^{\frac{1}{2}}(2\pi)^{-d/2}\,\mathbb{E}\big[W^{-\frac{d}{2}}\big], \tag{6}$$

and

$$K_0(x) = \frac{\rho}{\mathbb{E}\big[W^{-\frac{d}{2}}\big]}\,\mathbb{E}\left[W^{-\frac{d}{2}}\exp\left(-\frac{||\Lambda x||^2}{2W}\right)\right], \qquad x\in\mathbb{R}^d. \tag{7}$$

如果 $\varphi(x) = \mathcal{F}(K_0)(x)$ 表示 K_0 的傅里叶变换，则有

$$\varphi(x) = \frac{\rho}{h(0)}\,\mathbb{E}\big[\exp\big(-2\pi^2 W x^T(\Lambda^T\Lambda)^{-1}x\big)\big], \qquad x\in\mathbb{R}^d. \tag{8}$$

此外，对于任意紧致集 $R \subset \mathbb{R}^d$， Φ在 R 上的限制相对于 R 上的单位强度泊松点过程具有密度，条件是 $\rho < \rho_{\max}$。

定理 2.1 中刻画行列式点过程的核 K_0 并非各向同性。参数 ρ 是该过程的强度，即它控制过程中的点数分布。特别地， R 中的点的总数期望等于 $\rho|R|$。我们还要强调，如定理 2.1 所述，明确知道傅里叶变换对于模拟至关重要，因为通常需要按照第 3.1 节所述，使用 φ 近似行列式点过程的密度。

以下结果给出了上述各向异性行列式点过程的一个等价刻画。

定理 2.2。 设 Φ 是定义在 $\mathbb{R}^d$ 上的各向异性行列式点过程，其定义如定理 *2.1* 所述。则 $\tilde{\Phi} = \{\Lambda\mu : \quad \mu \in \Phi\}$ 是定义在 $B = \Lambda\mathbb{R}^d$ 上的平稳且各向同性的行列式点过程，其核为

$$\tilde{K}_0(y) = \frac{\rho|\Lambda^\top\Lambda|^{-1/2}}{\mathbb{E}[W^{-d/2}]}\mathbb{E}\left[W^{-d/2}\exp\left(-\frac{\|y\|^2}{2W}\right)\right].$$

定理 2.2 揭示了我们的各向异性行列式点过程所产生的排斥性，以及它为何适用于潜在混合模型的背景。正如第 1.1 节所讨论的，如果点 $\{\Lambda\mu_1 \ldots \Lambda\mu_m\}$ 彼此充分分离，该模型便会产生可解释簇。由上述定理可知，定理 2.1 中的新构造等价于假设点 $\Lambda\mu_h's$服从各向同性行列式点过程先验，这与我们促使数据簇彼此分离的目标相符。不过，我们要强调，定理 2.2 中的构造并非“可操作的”。事实上，若要处理定义在 $\Lambda\mathbb{R}^d$上的过程，˜就需要在马尔可夫链蒙特卡洛算法的每一步都对 K_0 进行特征分解，也就是说，每当更新 Λ 时都要进行这一操作。除了求出 K_0的特征分解˜在解析上存在困难之外，这还带来了计算负担，而定理 2.1 中的（等价）定义则没有这一负担。

排斥的类型或排斥强度由定理 2.1 中的随机变量 $W \sim p(w)$ 控制。在本文其余部分，我们考虑定理 2.1 中随机变量 W 的一种特定选择，由此得到 Lavancier 等（2015）所讨论的高斯行列式点过程的各向异性对应物，以下简称类高斯 DPP。此外，为证明我们方法的普适性，我们在附录 S3 中展示了如何构造 Whittle-Matérn DPP 的各向异性对应物，尽管在示例中不会使用它。

推论 2.3。 沿用定理 *2.1* 中的记号，令 W 为退化随机变量，定义为 $W = |\Lambda^T\Lambda|^{\frac{1}{d}}c^{-2/d}$，其中 $c > 0$，且 Λ 为固定值。则核 K_0、其傅里叶变换 $\varphi = \mathcal{F}(K_0)$ 以及 $\rho_{\max}$ 如下：

$$\begin{aligned}
K_0(x) &= \rho\exp\left(-\frac{||\Lambda x||^2}{2|\Lambda^T\Lambda|^{\frac{1}{d}}\,c^{-\frac{2}{d}}}\right), \qquad x \in \mathbb{R}^d \\
\varphi(x) &= \rho\,\frac{(2\pi)^{d/2}}{c}\exp\left(-2\pi^2|\Lambda^T\Lambda|^{\frac{1}{d}}c^{-\frac{2}{d}}x^T(\Lambda^T\Lambda)^{-1}x\right), \qquad x \in \mathbb{R}^d \qquad (9) \\
\rho_{\max} &= c(2\pi)^{-d/2}.
\end{aligned}$$

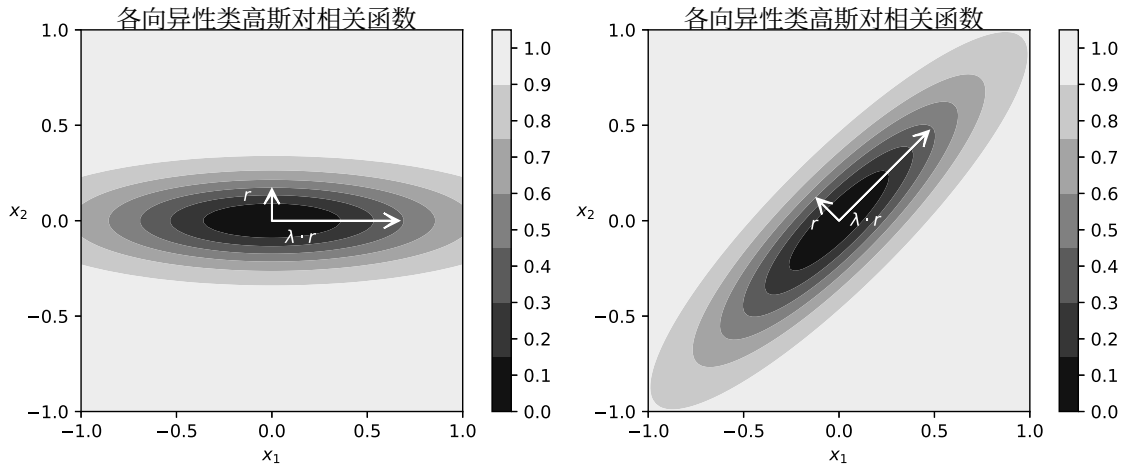

图1：两个类高斯 DPP 的成对相关函数，其中一个沿水平方向呈现强烈排斥（左图），另一个沿第一象限角平分线呈现强烈排斥（右图）。

最后，我们通过考察 Λ 对行列式点过程排斥性的影响来结束本节。首先，显然 Λ 会引入各向异性。为直观展示这一点，我们考察成对相关函数 $g(x)$（成对相关函数，Lavancier 等，2015）；对于类高斯 DPP，我们有：

$$g(x) = 1 - \frac{K_0(x)}{K_0(0)} = 1 - \exp\left(-\frac{||\Lambda x||^2}{2|\Lambda^T\Lambda|^{\frac{1}{d}}\, c^{-\frac{2}{d}}}\right)^2, \quad x \in \mathbb{R}^d \tag{10}$$

并出于可视化目的设定 $p = d = 2$。图1展示了具有不同 $\Lambda \in \mathbb{R}^{2\times 2}$ 的两个类高斯 DPP 的成对相关函数。左图中，Λ 的特征向量为 $e_1 = (1,0)^T$、$e_2 = (0,1)^T$，特征值为 $\lambda_1 = 1$、$\lambda_2 = \lambda(= 4)$，因此沿水平轴的排斥强于沿垂直方向的排斥。右图中，Λ 的特征向量为 $e_1 \equiv \sqrt{2}/2(1,1)^T$、$e_2 = \sqrt{2}/2(-1,1)^T$，特征值为 $\lambda_1 = 1$、$\lambda_2 = \lambda(= 4)$，因此沿第一象限角平分线的排斥强于沿正交方向的排斥。

尽管成对相关函数有助于直观展示排斥性的各向异性，但它几乎无法说明行列式点过程带来了多大程度的排斥。针对空间点过程，人们提出了若干排斥性度量。例如，参见 Lavancier 等（2015）；Biscio 和 Lavancier（2016），其中介绍了一些定量指标。本文尤其关注 Møller 和 O'Reilly（2021）提出的系数 p_0，它是一个全局排斥指标，用于量化一个点对该过程期望点数的影响。对于平稳行列式点过程，$p_0 = \rho^{-1}\int |K_0(x)|^2 \mathrm{d}x$，其中 $\rho := K_0(0)$。p_0越大，行列式点过程的排斥性就越强。对于推论 2.3 中的类高斯 DPP，标准计算可得 $p_0 = \rho\pi^{d/2}c^{-1}$，这表明该过程中排斥的程度不依赖于 Λ。

2.2 APPLAM 模型

本文提出的各向异性排斥点过程潜在混合模型（如前所述，记为 APPLAM）假设似然为（2）。我们按如下方式补充先验设定。首先，遵循混合模型中的惯例，在给定$\{\mu_1, \ldots, \mu_m\}$的条件下，实际上仅在给定 m的条件下，我们假设

$$\begin{aligned} w_1, \ldots, w_m \,|\, m &\sim \text{Dirichlet}(\alpha, \ldots, \alpha) \\ \Delta_1, \ldots, \Delta_m \,|\, m &\overset{\text{iid}}{\sim} \text{IW}_d(\nu_0, \Psi_0) \end{aligned} \tag{11}$$

其中，$\text{IW}_d(\nu_0, \Psi_0)$表示 d维逆Wishart分布，其自由度为 $\nu_0 > d-1$，均值为 $\Psi_0/(\nu_0 - d - 1)$。此外，与潜在因子模型中的常见做法一致，我们假设式（2）中的 $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$为对角矩阵，且

$$\sigma_j^2 \overset{\text{iid}}{\sim} \text{inv-Gamma}(a_\sigma, b_\sigma), \qquad j = 1, \ldots, p, \tag{12}$$

即，其边际先验分布为逆伽马分布，均值为 $b_\sigma/(a_\sigma - 1)$。对于矩阵 Λ，我们通过先验施加稀疏性。事实上，为避免过度参数化，潜在因子模型中常常采用稀疏先验。具体而言，我们假设以参数 a 为参数的狄利克雷-拉普拉斯先验（Bhattacharya等人，2015）作为 Λ的先验，即，将 Λ的元素记为 λ_{jh}，

$$\begin{gathered} \lambda_{jh} \,|\, \phi, \tau, \psi \overset{\text{ind}}{\sim} \mathcal{N}(0, \psi_{jh}\phi_{jh}^2\tau^2), \quad j = 1, \ldots, p,\ h = 1, \ldots, d \\ vec(\phi) \sim \text{狄利克雷分布}(a, \ldots, a) \quad \psi_{jh} \overset{\text{独立同分布}}{\sim} \text{指数分布}(1/2) \quad \tau \sim \text{伽马分布}(pda, 1/2) \end{gathered} \tag{13}$$

其中，对于任意 $p \times d$ 矩阵 A，$vec(A)$ 表示维度为 $p \times d$ 的向量，使得$vec(A)_{p(h-1)+j} = (A)_{j,\,h}$。由于易于抽样且具有良好的频率学派性质，这一先验在贝叶斯因子模型文献中非常常见（Bhattacharya等人，2015年）。另请注意，我们假设上述各参数块之间先验独立。

我们模型的关键在于对潜在因子中心$\{\mu_1, \ldots, \mu_m\}$的先验设定。为使点$\{\Lambda\mu_h\}_{h=1}^m$之间具有排斥性，我们首先固定一个紧致集 $R \subset \mathbb{R}^d$，并假设位置$\{\mu_1, \ldots, \mu_m\}$ 在给定 $m \geq 1$的条件下服从定义在 R上的各向异性 DPP。如以下内容所示，R 的选择作用很有限，我们将在第2.3节讨论如何确定它。由于若 $m = 0$，式（2）便没有定义，因此在我们的模型设定中，必须以 DPP 非空为条件。于是，我们假设位置 $\{\mu_1, \ldots, \mu_m\}$ 及其基数服从以下密度：

$$\begin{aligned} p(\{\mu_1, \ldots, \mu_m\} \,|\, \Lambda) &= f_{\text{DPP}}(\boldsymbol{\mu} \,|\, \rho, \Lambda, K_0; R) \\ &= \frac{\mathrm{e}^{|R|-D}}{1 - \mathrm{e}^{-D}} \det\{C(\mu_h, \mu_k)\}_{h,k=1,\ldots,m}, \qquad m \geq 1,\ \mu_1, \ldots, \mu_m \in R \end{aligned} \tag{14}$$

以及 $p(\emptyset \,|\, \Lambda) = 0$，其中 D和 C在 (4) 后面定义。上述记号 $f_{\text{DPP}}(\boldsymbol{\mu} \,|\, \rho,\ \Lambda,\ K_0;\ R)$ 明确表明，密度取决于 R、强度 ρ、平稳核 K_0，而各向异性由 Λ驱动。

2.3 各向异性行列式点过程的先验 elicitation

Σ、 Δ_h 和 Λ 的先验分布在 LAMB 和 APPLAM 中相同，我们采用 Chandra 等人（2023）建议的默认值。更多详情请参见附录 S6.2。这里，我们重点关注 (14) 中定义 DPP 密度先验的那些超参数，其中 K_0 是推论 2.3 中的各向异性类高斯核，即R、 ρ 和 c。

首先，我们讨论 R 的紧集选择问题，该紧集是 μ_h 的支撑集。请注意，从边际分布来看，每个 μ_h 在 R 上均匀分布。我们认为，一个合理的选择是假设 $R = [-r, r]^d$，即取一个超立方体。事实上，没有理由先验地假设潜在因子沿某一特定轴具有更高的方差。此外，与传统的潜因子模型一样，这样的超立方体应以原点为中心，在传统潜因子模型中，潜在因子服从标准高斯先验。显然， Λ 与 r 之间存在不可识别性。事实上，令 $t > 0$ 和 $\tilde{\Lambda} = t\Lambda$， $\tilde{R} = [-r/t, r/t]^d$ 会使数据的似然保持不变。因此， r 的选择在模型定义中所起的作用有限。特别是，根据我们的经验，当 r 足够大，即 $r \geq 5$ 时，后验推断具有稳健性。因此，在本文其余部分，我们设定 $R = [-10, 10]^d$ 。

参数 ρ 和 c 控制该过程的排斥性和强度。如第 2.1 节所述，在 DPP 中，这些量密切相关。尽管 $\rho|R|$ 是 DPP 的 R 中的期望点数，但请注意，在式（14）中，我们以过程非空为条件，因此 ρ 的作用不像在式（4）中那样明确。因此，我们建议仅考察 DPP 所诱导的排斥性来固定这些参数。在附录 S5.4 中，我们利用定理 2.2 所给出的各向异性 DPP 的等价表示，以深入了解此类参数如何决定所谓的排斥范围（即两个簇中心应相距多远才会相互影响），以及相近点之间的排斥强度。随后，我们提出一种基于观测数据的启发式方法，为超参数 ρ， c设定合理的范围。具体而言，我们遵循 Lavancier 等（2015）；Bianchini 等（2020），并通过 (ρ， s) 对 DPP 进行参数化，使得 $\rho = s\rho_{\max} \equiv sc(2\pi)^{-d/2}$， $s \in (0, 1)$，参见推论 2.3。如此一来，对于固定的 ρ， s 可被解释为排斥强度。

3 后验模拟

3.1 DPP密度的近似

由于表达式中 D 和 C 含有无穷级数，无法以闭式计算式（14）中的点过程密度。我们遵循 Lavancier 等（2015）的方法，使用 $f^{\text{app}}{}_{\text{行列式点过程}}$ 对其进行近似；该近似与式（14）相同，只是将 D 替换为 $D^{\text{app}} = -\sum_{k \in \mathbb{Z}_N^d} \log(1 - \varphi(k))$，并将 C 替换为

$$C^{\text{app}}(x, y) = \frac{1}{|R|} \sum_{k \in \mathbb{Z}_N^d} \frac{\varphi(k)}{1 - \varphi(k)} \exp\left(\frac{2\pi i}{|R|^{1/d}} \langle k, x - y \rangle \right) \tag{15}$$

其中 $\mathbb{Z}_N^d = \{-N, \ldots, N\}^d$ 和 $\langle \cdot, \cdot \rangle$ 表示标量积。请注意，$\boldsymbol{\mu} \sim f_{\text{近似行列式点过程}}(\rho, \Lambda, C^{\text{app}}; R)$ 在条件为非空时，仍是定义在 R 上的行列式点过程。截断级别 N 控制密度近似的精度，并为行列式点过程中的点数设定上限，即 $(2N+1)^d$（Lavancier 等，2015）。这种近似类似于贝叶斯非参数混合模型中常用的截棍先验截断方法（Ishwaran 和 James，2001）。对于狄利克雷过程，通常将先验截断至 50 个支撑点。类似地，当 $d = 2$（我们所考虑的最小值）且 $N = 3$ 时，此处将行列式点过程中的点数上限设为 49；该上限随维度呈指数增长。附录 S5.1 报告了一项模拟结果，表明当 Λ 固定为单位矩阵，或为随机变量且服从先验（13）或后验分布时，截断引入的近似误差实际上可以忽略不计。

3.2 吉布斯抽样算法

将混合权重 $\boldsymbol{w}$ 纳入状态空间并令其服从有限狄利克雷分布（见（11）的第一行）并不方便，因为对 $\boldsymbol{w}$ 的和为一约束会导致复杂的分裂—合并可逆跳跃移动，并使链的混合效果较差。与 Beraha 等（2022）一样，我们考虑先验的以下等价刻画：$\boldsymbol{w}$

$$\boldsymbol{w} = \left(\frac{S_1}{T}, \ldots, \frac{S_m}{T} \right), \quad T = \sum_{h=1}^{m} S_h, \quad S_h \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, 1). \tag{16}$$

在给定 $\boldsymbol{c}$ 的条件下，我们考虑 $\boldsymbol{\mu} = \boldsymbol{\mu}^{(a)} \cup \boldsymbol{\mu}^{(na)}$、$\boldsymbol{S} = [\boldsymbol{S}^{(a)}, \boldsymbol{S}^{(na)}]$ 和 $\boldsymbol{\Delta} = [\boldsymbol{\Delta}^{(a)}, \boldsymbol{\Delta}^{(na)}]$，将其划分为已分配分量和未分配分量，并分别用上标（a）和（na）表示。也就是说，$\boldsymbol{\mu}^{(a)} = \{\mu_{c_i}, i = 1, \ldots, n\}$、$\boldsymbol{\mu}^{(na)} = \boldsymbol{\mu} \setminus \boldsymbol{\mu}^{(a)}$，涉及 S_h 和 Δ_h 的项也同理。

将似然（2）与先验（11）—（14）结合，可以看出，数据与参数的联合分布相对于适当空间上的某个支配测度具有密度；该密度及相应的支配测度见附录 S2。对权重进行归一化会在联合密度的表达式中产生一项 $T^{-n} = (\sum S_{h_1}^{(a)} + \sum S_{h_2}^{(na)})^{-n}$，从而无法按照已分配分量和未分配分量对密度进行因子分解。为解决这一问题，与 Beraha 等（2022）一样，我们引入一个辅助随机变量 $u \,|\, T \sim$，其服从伽马分布 Gamma(n, T)。数据与参数的联合密度见附录 S2。

我们提出了一种吉布斯采样算法；当相应的全条件分布无法直接采样时，则采用梅特罗波利斯步骤。大多数更新都很直接，以下两个步骤除外。算法的完整描述见附录 S4.1。

为更新未分配变量（$\boldsymbol{\mu}^{(na)}$、$\boldsymbol{S}^{(na)}$、$\boldsymbol{\Delta}^{(na)}$），我们将未分配变量的联合全条件分布分解为

$$p(\boldsymbol{\mu}^{(na)}, \boldsymbol{s}^{(na)}, \boldsymbol{\Delta}^{(na)} \,|\, \text{rest}) = p(\boldsymbol{\mu}^{(na)} \,|\, \text{rest}) p(\boldsymbol{s}^{(na)} \,|\, \boldsymbol{\mu}^{(na)}, \text{rest}) p(\boldsymbol{\Delta}^{(na)} \,|\, \boldsymbol{\mu}^{(na)}, \text{rest}),$$

whe其中“其余”指除（$\boldsymbol{\mu}^{(na)}$、$\boldsymbol{s}^{(na)}$、$\boldsymbol{\Delta}^{(na)}$）之外的所有变量。然后 $\boldsymbol{\mu}^{(na)}$ |re st

是一个密度为

$$p(\{\mu_1^{(na)},\dots,\mu_\ell^{(na)}\}\,|\,\text{rest}) \propto f_{\text{DPP}}^{\text{app}}(\{\mu_1^{(na)},\dots,\mu_\ell^{(na)}\}\cup\boldsymbol{\mu}^{(a)}\,|\,\rho,\Lambda,K_0;R)\psi(u)^\ell$$

其中 $\psi(u) = \mathbb{E}[\mathrm{e}^{-uS}]$。另请参见 Beraha 等（2022）。我们采用 Geyer 与 Møller（1994）中的生灭 Metropolis-Hastings算法来从该点过程密度中抽样。给定 $\boldsymbol{\mu}^{(na)}$，不难证明

$$\begin{aligned}\Delta_1^{(na)},\dots,\Delta_\ell^{(na)}\,|\,\text{rest} &\overset{\text{iid}}{\sim} \text{IW}_d(\nu_0,\Psi_0)\\ S_1^{(na)},\dots,S_\ell^{(na)}\,|\,\text{rest} &\overset{\text{iid}}{\sim} \text{Gamma}(\alpha,1+u)\,.\end{aligned}$$

为更新潜在分配变量 $\boldsymbol{c}$，我们发现，对 η_i 求边缘化有助于改善混合链。因此，我们可以从 $\{1,\dots,k+\ell\}$ 上的离散分布中独立地抽取每个 c_i，其中 k 是已分配分量的数量，权重为 ω_{ih}：

$$\begin{aligned}\omega_{ih} &\propto S_h^{(a)}\mathcal{N}_p(y_i\,|\,\Lambda\mu_h^{(a)},\Sigma+\Lambda\Delta_h^{(a)}\Lambda^\top), && h=1,\dots,k\\ \omega_{ik+h} &\propto S_h^{(na)}\mathcal{N}_p(y_i\,|\,\Lambda\mu_h^{(na)},\Sigma+\Lambda\Delta_h^{(na)}\Lambda^\top), && h=1,\dots,\ell\,.\end{aligned}$$

若不加以注意，每次计算 p 维高斯密度都需要 $\mathcal{O}(p^3)$ 次运算。不过，我们利用了协方差矩阵的特殊结构。根据 Woodbury 矩阵恒等式，有

$$\left(\Sigma+\Lambda\Delta\Lambda^\top\right)^{-1} = \Sigma^{-1}-\Sigma^{-1}\Lambda\left(\Delta^{-1}+\Lambda^\top\Sigma^{-1}\Lambda\right)^{-1}\Lambda^\top\Sigma^{-1},$$

因此，现在我们只需计算一个 $d\times d$ 矩阵的逆。所以，计算指数项中的二次型只需 $\mathcal{O}(p)$ 的计算开销。此外，根据矩阵行列式引理，协方差矩阵的行列式可计算为

$$\det(\Sigma+\Lambda\Delta\Lambda^\top) = \det(\Delta^{-1}+\Lambda^\top\Sigma^{-1}\Lambda)\det(\Delta)\det(\Sigma).$$

通过缓存矩阵求逆过程中的运算，即可计算该值，无需额外开销。

3.3 使用基于梯度的 MCMC 算法更新 Λ

如第 1.1 节所述，从 Λ 的完整条件分布中采样并非易事。具体而言，我们发现，随机游走 Metropolis-Hastings 算法会导致 MCMC 链的混合效果很差，而 Haario 等（2001）提出的自适应 Metropolis-Hastings 算法则因 Λ 的维度过高而不适用于此处。相较之下，我们发现 Metropolis 调整的 Langevin 算法（MALA；Roberts 和 Tweedie，1996年）更适合此处。目标密度为

$$p(\Lambda\,|\,\cdots)\propto p(\boldsymbol{y}\,|\,\Lambda,\boldsymbol{\eta},\Sigma)p(\Lambda\,|\,\phi,\tau,\psi)\frac{\mathrm{e}^{1-D^{\text{app}}}}{1-\mathrm{e}^{-D^{\text{app}}}}\det[C^{\text{app}}](\mu_1,\dots,\mu_m), \tag{17}$$

其中， $\cdots$ 表示除 Λ 以外的所有变量。尽管文中未明确说明， D^{app} 和 C^{app} 都依赖于 Λ。

为使用 MALA 从 (17) 中抽样，我们必须计算 $\nabla \log(p(\Lambda \mid \cdots))$。在一项初步研究中，我们尝试使用自动微分（AD，Griewank 等人，1989）来获得 $\nabla \log(p(\Lambda \mid \cdots))$，因为这只需要实现一个计算 $\log(p(\Lambda \mid \cdots))$ 的函数。遗憾的是，我们发现该策略仅在简单情形下可行，即由于 RAM 内存需求，最多只能处理 $p = 50$和 $d = 3$；见附录 S5.2 中的图 S7。因此，在以下定理中，当相关的行列式点过程为类高斯 DPP 时，我们给出 $\nabla \log(p(\Lambda \mid \cdots))$ 的解析表达式。有关 Whittle-Matérn DPP 的情形，见附录 S3。

定理 3.1。在类高斯 *DPP* 先验下， Λ 的对数全条件密度的梯度等于

$$\nabla \log p(\Lambda \mid \cdots) = \Sigma^{-1} \sum_{i=1}^{n} (y_i - \Lambda \eta_i) \eta_i^{\top} - \frac{1}{(\psi \odot \phi^2)\tau^2} \odot \Lambda + \\ + (2\pi^2 c^{-\frac{2}{d}}) \sum_{k \in \mathbb{Z}_N^d} g^{(k)} \frac{\varphi(k)}{(1-\varphi(k))^2} \Big[\frac{1-\varphi(k)}{1-\mathrm{e}^{-D^{\text{app}}}} - v_k^T \left(C^{\text{app}}\right)^{-1} u_k \Big]$$

其中 $\varphi(k)$ 在 (9) 中定义， $\odot$ 表示逐元素（*Hadamard*）乘积，

$$g^{(k)} = 2|\Lambda^T \Lambda|^{\frac{1}{d}} \Lambda (\Lambda^T \Lambda)^{-1} \Big[\frac{1}{d} k^T (\Lambda^T \Lambda)^{-1} k \mathbb{1}_d - k((\Lambda^T \Lambda)^{-1} k)^T \Big],$$

u_k 和 v_k 对于每个 $k \in \mathbb{Z}^d$ 都是 m 维列向量，其元素为

$$(u_k)_j = \mathrm{e}^{2\pi i \langle k, T\mu_j \rangle}, \quad (v_k)_j = \mathrm{e}^{-2\pi i \langle k, T\mu_j \rangle}, \quad j = 1, \ldots, m$$

和 $C^{\text{app}} := C^{\text{app}}$（$\mu_1$、……、$\mu_n$）。

附录 S5.2 中的图 S7 比较了使用自动微分梯度或我们的解析表达式时的内存需求和运行时间。具体而言，使用自动微分梯度所需的内存大约多出 100 倍；由于这是我们算法的瓶颈，因此在实践中影响显著。使用自动微分时，运行时间也会增加 10 倍。通过运行较短的初步链来调整 MALA 算法的步长参数，使接受率达到约 20%。具体来说，我们发现，取值在 10^{-8} 和 10^{-10} 之间通常能使 MCMC 链具有良好的混合性。

4 模拟研究

附录 S5 展示了几个模拟示例，旨在呈现我们模型的不同方面。除了考察截断 DPP 密度的影响（附录 S5.1）以及定理 3.1 中梯度解析表达式的实用性（附录 S5.2）之外，我们还说明了在潜在混合模型的背景下，有必要采用各向异性 DPP，而非

各向同性 DPP 先验。附录 S5.3 展示了这一点。附录 S5.4 说明了一种确定超参数的流程。此外，在附录 S5.5 中，我们证明了潜在簇中心的各向异性 DPP 先验的超参数变化不会显著影响后验推断，体现了后验推断的稳健性。

附录 S6 聚焦于我们 APPLAM 模型的自然竞争对手，即 Chandra 等人（2023）提出的 Lamb 模型。请注意，相较于基于两步法的更具启发性的做法，Chandra 等人（2023）已经有力地论证了潜在混合模型的优势。因此，我们仅展示 APPLAM 对模型错设的稳健性，以及 Lamb 模型在这方面的不足。具体而言，我们考虑了两种模拟情形：模型错设分别发生在潜在层级，即潜在得分 η_i 是从 t 个分布的混合分布中模拟得到的；或发生在似然层面，即在给定所有参数的条件下，观测值服从 t 分布。在这两种情形下，对于潜在维度 $d = \{4, 8\}$、数据维度 $p = \{500, 1000\}$ 的各种取值，以及数据集的 100 个独立同分布重复样本，我们展示了 Lamb 模型如何显著高估簇数，而我们的 APPLAM 模型则不会出现这一问题。此外，我们还表明，APPLAM 比 Lamb 能够产生更可靠的聚类估计。

总之，我们的模拟从经验上证实了这一观点：在存在模型设定错误时，排斥混合模型比非排斥混合模型更稳健。此外，模拟结果也证明了本文引入各向异性 DPP 的合理性，因为显而易见，强制簇中心之间存在各向同性排斥，并不能使后验的数据簇彼此分离。相反，各向同性模型可能会大大高估簇数。

5 JointSpecies 建模

我们将 APPLAM 模型应用于法国博日自然区域公园不同地点植物物种出现情况的数据。这些数据由阿尔卑斯植物保护中心（CBNA）提供，Bystrova 等人（2021）和 Thuiller 等人（2018）此前曾对其进行研究。具体而言，我们分析的数据集涉及 $n = 1139$ 个地点，并记录了每个地点 $p = 123$ 种不同植物物种的存在-缺失情况。我们的目标是推断这些地点的聚类结构，使属于同一簇的地点在植物物种出现模式方面表现相似。请注意，这些数据为二元数据；具体而言，令 $z_i \in \{0, 1\}^p, i = 1, \ldots, n$，使得 $z_{i,j} = 1$当植物物种 j 出现在地点 i, $z_{i,j} = 0$否则。我们假设$z_{i,j} = \mathbb{1}_{[0,\infty)}(y_{i,j})$，并假定采用类高斯 DPP，将 APPLAM 模型应用于 y_i。为避免不可识别性，我们将式（11）中的 Δ_h 固定为 $d \times d$ 单位矩阵。后验模拟只需对吉布斯抽样算法做少量修改，相关讨论见附录 S4.2。

我们通过模型选择来选取超参数。具体而言，我们共考察了 144 个模型，通过调整 (ρ, c) 以及其他超参数来改变排斥性程度。更多详情请参见附录 S7.1。从图 S17 和 S18 可以清楚看出，当 $\rho|R|$ 较大和/或排斥强度 s 较弱时，估计簇数会增加，这证实了我们从模拟研究中得到的见解

研究。具体而言，估计簇数介于 3 到 15 之间；另见表 S4 和表 S5。我们仅关注簇数后验均值小于八的模型，并通过最大化 WAIC 指数（Watanabe，2013）选出其中最佳模型；这在模型选择中是标准做法。由此得到以下超参数取值： $d = 3$、$\rho|R| = 0.1$、 $s = 0.9$、 $\alpha = 10^{-3}$ 和 $a_\sigma = 2$， $b_\sigma = 1$。请注意， α 较小（如本例所示）意味着混合权重 $\boldsymbol{w}$ 在式 (11) 中具有稀疏的边际先验（Rousseau & Mengersen，2011）。

我们在此讨论通过最小化 Binder 损失函数的后验期望（Binder，1978年）得到的估计簇，采用的是最后一种超参数选择。估计得到五个簇，基数介于 200 至 250 之间。为了解释这些估计簇，我们考察各簇中物种存在-缺失的模式。具体而言，对于 $j = 1$，……， p，令 $\bar{z}_j$和 $\bar{z}_{c,\ j}$分别表示整个样本和第 c 个簇中第 j 个物种的经验频率。对于每个簇，我们通过选择使 $\delta_{c,\ j} = (\bar{z}_{c,\ j} - \bar{z}_j)$最大（或最小）的三个物种，选出六个最能代表该簇的物种。总计，我们找到了 20 个最能描述不同簇的物种。图 2 展示了相应的 $\delta_{c,\ j}$。簇 3 与样本平均水平并无显著差异，所有选定物种的 $|\delta_{3,\ j}| < 0.1$ 均如此，但物种存在的总体比例低于平均水平。与其他簇相比，簇 4 的偏离均值程度最为极端，尤其是高于平均水平的偏离。事实上，物种 27（普通岩蔷薇）、46（黄龙胆）和 97（黄花茅）的 $\delta_{4,\ j}$ 约为 0.2 或更高。不过，该簇中低于平均水平的偏离也最为极端：物种 45（欧洲山毛榉）的存在率比其对应的样本平均值低于 0.15，物种 112（欧洲冷杉）则低于 0.1。簇 5 与其他簇的主要区别在于物种 6（常春藤）、45（欧洲山毛榉）、83（香车叶草）、112（欧洲冷杉）、117（田野枫）和 121（一种黑莓）的存在率较高。欧洲山毛榉有助于解释簇 4 与簇 5 之间的差异。最后，簇 2 主要表现为物种 72（圣安东尼萝卜）的存在率较高，而簇 1 中物种 28（欧洲榛）的存在率较高。图 2 右下角的面板展示了针对解释这五个估计簇的 20 个物种所限定的 $\Lambda\Lambda^\top$ 后验期望。物种 72（圣安东尼萝卜）与物种 117（田野枫）之间估计出不可忽略的负相关；从这两个物种在五个估计簇中的经验比例可以观察到这一点。

6 讨论

对中等或高维数据进行基于模型的聚类一直以难度大而著称。我们提出了一种同时实现降维与聚类的模型：假设潜在得分服从混合模型，再通过高斯潜在因子模型将其与观测值联系起来。Chandra 等人（2023）近期研究了这一方法。作者采用因子分析表示，并假设潜在因子服从混合模型。然而，模型错设可能导致性能下降。若为潜在得分混合模型中的成分特异性

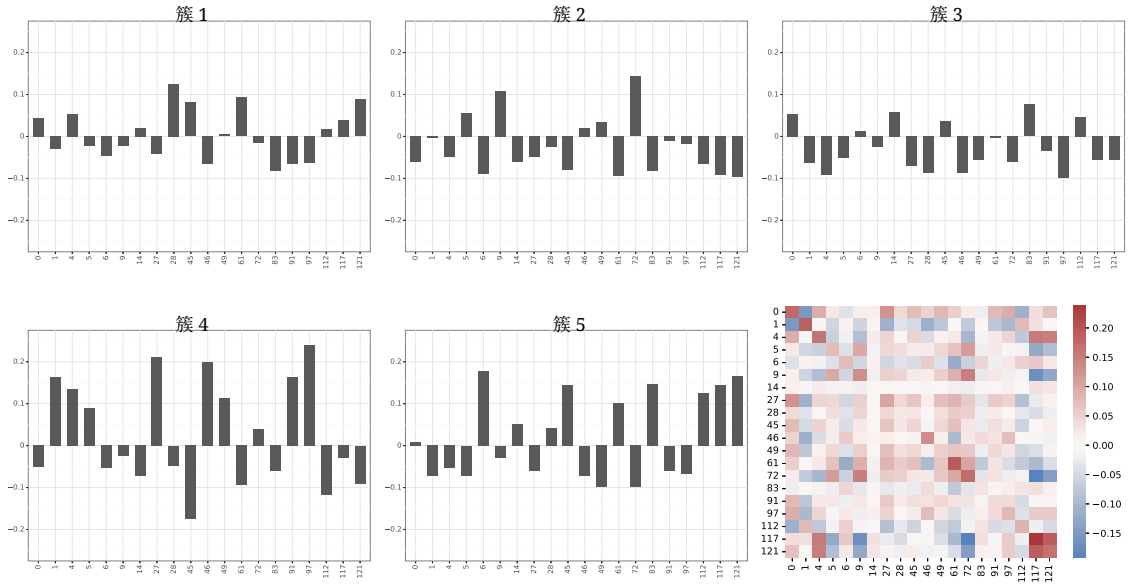

图 2：$\delta_{c,\ j}$ 用于识别五个簇的 20 个物种的取值，以及（右下）限制在这 20 个物种上的 $\Lambda\Lambda^{\top}$ 后验估计。

均值设定排斥点过程先验，则可得到更稳健的模型，并且在若干模拟场景中优于标准Lamb。为了使数据簇充分分离，排斥点过程必须具有各向异性，且其密度应便于计算，以支持高效的后验推断。为解决这些问题，我们提出了一种构造各向异性行列式点过程的通用方法。

我们的采样算法的瓶颈在于对DPP密度进行谱近似，其计算成本会随 d（即潜在因子的维度）呈指数级增长。在潜在因子模型中，通常会将 d 设为较小的值。不过，当 d 取中等或较大值时，Bardenet & Titsias（2015）提出的方法会更高效，但代价是模型中的参数更难解释。Poinas & Lavancier（2021）还提出了另一种DPP密度近似方法，该方法也适用于非（超）矩形域。对于投影DPP，即谱分解的特征值非零即一的情况，我们可以采用 Lavancier & Rubak（2023）中的方法，从未分配分量的全条件分布中精确采样，而不必使用生灭Metropolis-Hastings移动。

我们尚未研究因子载荷矩阵 Λ 的估计问题，而这可能有助于解释数据的相关结构。我们的模型继承了经典因子模型的旋转不变性，但可以通过事后程序处理这一问题，以估计 Λ。例如，可参见 Papastamoulis & Ntzoufras（2022）及其中的参考文献。

致谢

M.B. 和 A.G. 感谢 MUR 提供的资助，资助项目为卓越系 2023-2027。M.B. 获得了欧洲研究委员会（ERC）通过欧盟地平线2020研究与创新计划提供的资助，资助协议编号为 817257。A.G. 获得了 MUR - Prin 2022 - 资助项目编号 2022CLTYP4 的部分支持，该项目由欧盟下一代欧盟计划资助。我们衷心感谢 DataCloud实验室（https://datacloud.polimi.it）：部分数值实验得以完成，离不开 DataCloud实验室提供的云资源。

参考文献

ASCOLANI, F., LIJOI, A., REBAUDO, G. & ZANELLA, G.（2022年）。基于狄利克雷过程混合模型的聚类一致性。*Biometrika*。BACCELLI, F., BŁASZCZYSZYN, B. & KARRAY, M.（2020年）。随机测度、点过程与随机几何。*HAL*预印本，见 *https://hal.inria.fr/hal- 02460214/* 。BARDENET, R. & TITSIAS, M.（2015年）。在缺乏谱信息的情况下对行列式点过程进行推断。载于神经信息处理系统进展，C. Cortes、N. Lawrence、D. Lee、M. Sugiyama & R. Garnett 编，第28卷。Curran Associates, Inc. BERAHA, M., ARGIENTO, R., MØLLER, J. & GUGLIELMI, A.（2022年）。使用排斥点过程进行贝叶斯混合模型的马尔可夫链蒙特卡洛计算。*J. Comp. Graph.Stat.*，1–14。BHATTACHARYA, A., PATI, D., PILLAI, N. S. & DUNSON, D. B.（2015年）。用于最优收缩的狄利克雷–拉普拉斯先验。《美国统计学会会刊》**110**，1479–1490。BIANCHINI, I., GUGLIELMI, A. & QUINTANA, F. A.（2020年）。通过谱密度方法构建行列式点过程混合模型。《贝叶斯分析》**15**，187–214。BINDER, D. A.（1978年）。贝叶斯聚类分析。*Biometrika***65**，31–38。BISCIO, C. A. N. & LAVANCIER, F.（2016年）。量化行列式点过程的排斥性。*Bernoulli***22**，2001–2028。BYSTROVA, D., POGGIATO, G., BEKTAŞ, B., ARBEL, J., CLARK, J. S., GUGLIELMI, A. & THUILLER, W.（2021年）。在联合物种分布模型中，利用残差协方差矩阵对物种进行聚类。《生态学与进化前沿》**9**。CAI, D., CAMPBELL, T. & BRODERICK, T.（2021年）。有限混合模型无法可靠地学习成分数量。载于国际机器学习会议。PMLR。

CELEUX, G., KAMARY, K., MALSINER-WALLI, G., MARIN, J.-M. & ROBERT, C. P.（2019年）。混合模型贝叶斯推断的计算方法。混合分析手册，73–96。CHANDRA, N. K., CANALE, A. & DUNSON, D. B.（2023年）。在贝叶斯模型聚类中摆脱维数灾难。机器学习研究杂志**24**，144–1。DALEY, D. J. & VERE-JONES, D.（2008年）。点过程理论导论。第二卷。《概率及其应用》（纽约）。纽约：施普林格，第2版。一般理论与结构。DUAN, L. L. & DUNSON, D. B.（2021年）。贝叶斯距离聚类。机器学习研究杂志**22**，1–27。FRUHWIRTH-SCHNATTER, S., CELEUX, G. & ROBERT, C. P.（2019年）。混合分析手册。CRC出版社。FÚQUENE, J., STEEL, M. & ROSSELL, D.（2019年）。通过非局部先验选择混合成分。英国皇家统计学会会刊：*B*辑（统计方法学）**81**，809–837。GEYER, C. J. & MØLLER, J.（1994年）。空间点过程的模拟方法与似然推断。斯堪的纳维亚统计学期刊**21**，359–373。GRIEWANK, A. et al.（1989年）。关于自动微分。数学规划：最新进展与应用**6**，83–107。HAARIO, H., SAKSMAN, E. & TAMMINEN, J.（2001年）。一种自适应 Metropolis 算法。*Bernoulli***7**，223–242。HUBERT, L. & ARABIE, P.（1985年）。比较划分。分类学期刊**2**，193–218。ISHWARAN, H. & JAMES, L. F.（2001年）。截棍先验的吉布斯抽样方法。美国统计协会期刊**96**，161–173。LAVANCIER, F., MØLLER, J. & RUBAK, E.（2015年）。行列式点过程模型与统计推断。*J. R. Stat. Soc. Ser. B***77**，853–877。LAVANCIER, F. & RUBAK, E.（2023年）。关于连续行列式点过程的模拟。*arXiv*预印本 *arXiv:2301.11081*。LOPES, H. F.（2014年）。现代贝叶斯因子分析。社会科学中的贝叶斯推断，115–153。MACCHI, O.（1975年）。随机点过程的重合方法。*Adv. App.Prob.* **7**，83–122。MALSINER-WALLI, G., FRÜHWIRTH-SCHNATTER, S. & GRÜN, B.（2016年）。基于稀疏有限高斯混合的模型聚类。*Stat. Comp.***26**，303–324。

MILLER, J. W. & HARRISON, M. T.（2014年）。Pitman–Yor过程混合模型在成分数量方面的不一致性。机器学习研究杂志**15**，3333–3370。MØLLER, J. & O'REILLY, E.（2021年）。行列式点过程及其约化Palm分布的耦合：旨在量化排斥性。应用概率杂志**58**，469–483。MØLLER, J., PETTITT, A. N., REEVES, R. & BERTHELSEN, K. K.（2006年）。一种适用于具有难处理归一化常数的分布的高效马尔可夫链蒙特卡洛方法。*Biometrika***93**，451–458。MØLLER, J. & WAAGEPETERSEN, R. P.（2003年）。空间点过程的统计推断与模拟。CRC出版社。MURRAY, I., GHAHRAMANI, Z. & MACKAY, D. J. C.（2006年）。双重难处理分布的马尔可夫链蒙特卡洛。在第*22*届人工智能不确定性会议论文集中。美国弗吉尼亚州阿灵顿。NATARAJAN, A., DE IORIO, M., HEINECKE, A., MAYER, E. & GLENN, S.（2023年）。贝叶斯距离聚类中的凝聚性与排斥性。美国统计学会杂志，1–11。NEAL, R. M.（2003年）。使用狄利克雷扩散树进行密度建模与聚类。在贝叶斯统计第7卷中。PAPASTAMOULIS, P. & NTZOUFRAS, I.（2022年）。关于贝叶斯因子分析模型的可辨识性。统计与计算**32**，23。PETERSEN, K. B. & PEDERSEN, M. S.（2008年）。矩阵手册。版本20081110。PETRALIA, F., RAO, V. & DUNSON, D.（2012年）。排斥混合模型。在神经信息处理系统进展第25卷中。POINAS, A. & LAVANCIER, F.（2021年）。平稳行列式点过程似然的渐近近似。斯堪的纳维亚统计学期刊。QUINLAN, J. J., QUINTANA, F. A. & PAGE, G. L.（2021年）。使用排斥分布进行简约层次建模。*Test* **30**，445–461。ROBERTS, G. O. & TWEEDIE, R. L.（1996年）。朗之万分布及其离散近似的指数收敛。*Bernoulli*，341–363。ROUSSEAU, J. & MENGERSEN, K.（2011年）。过拟合混合模型中后验分布的渐近行为。皇家统计学会期刊*B*辑：统计方法学**73**，689–710。TEH, Y., DAUME III, H. & ROY, D. M.（2007年）。带合并过程的贝叶斯凝聚层次聚类。神经信息处理系统会议**20**。

THUILLER, W., GUÉGUEN, M., BISON, M. & DUPARC, E. A.（2018年）。结合点过程与景观植被模型，预测时空中的大型食草动物分布——以羚羊为例。《多样性与分布》 **24**，352–362。

WATANABE, S.（2013年）。一种广泛适用的贝叶斯信息准则。《机器学习研究杂志》 **14**，867–897。

XIE, F. & XU, Y.（2019年）。贝叶斯排斥高斯混合模型。《美国统计学会会刊》，187–203。

XU, Y.、MÜLLER, P. & TELESCA, D.（2016年）。利用行列式点过程（DPP）对潜在生物结构进行贝叶斯推断。《生物统计学》**72**，955–964。

S1 进一步的符号说明与证明

S1.1 DPP 的一般构造

m 阶的 Φ 阶乘测度定义为

$$\Phi^{(m)}(B_1 \times \cdots \times B_m) = \sum_{\mu_1,\ldots,\mu_m \in \Phi}^{\neq} \mathbb{1}[\mu_1 \in B_1] \cdots \mathbb{1}[\mu_m \in B_m]$$

对于可测的 B_1、……、$B_m \subset \mathbb{R}^d$。求和是对 Φ 中所有两两不同点组成的 m 元组进行的，这是标准记法。第 m 阶乘矩测度定义为 $M_{\Phi^{(m)}}(B_1 \times \cdots \times B_m) = \mathbb{E}\left[\Phi^{(m)}(B_1 \times \cdots \times B_m)\right]$，其中期望是相对于 Φ 取的。

为了定义行列式点过程（DPP），令 $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ 为连续协方差核。若对所有 $m = 1$、2、……，其第 m 阶乘矩测度相对于 d 维勒贝格测度的 m 重积 $\rho^{(m)}$ 存在密度，则称 Φ 是定义在 $\mathbb{R}^d$ 上的行列式点过程：

$$\rho^{(m)}(\mu_1, \ldots, \mu_m) = \det\{K(\mu_h, \mu_k)\}_{h,k=1,\ldots,m}, \qquad \mu_1, \ldots, \mu_m \in \mathbb{R}^d.$$

根据 Mercer 定理，$K(x, y) = \sum_{j \geq 1} \gamma_j \xi_j(x) \xi_j(y)$，其中 ξ_j 构成复值函数空间 $L^2(\mathbb{R}^d; \mathbb{C})$ 的正交归一基，而 γ_j 是一个可求和的非负序列。于是，存在一个以 K 为核的行列式点过程，当且仅当对所有 j 都有 $\gamma_j \leq 1$；参见 Macchi（1975）。限制在紧致集 $R \subset \mathbb{R}^d$ 上时，Φ 仍是一个行列式点过程，其核为 K 在 $R \times R$ 上的限制。特别地，若对所有 j 都有 $\gamma_j < 1$，则 Φ 关于定义在 R 上的单位强度泊松点过程的密度为

$$p(\{\mu_1, \ldots, \mu_m\}) = \mathrm{e}^{|R|-D} \det\{C(\mu_h, \mu_k)\}_{h,k=1,\ldots,m}, \qquad \mu_1, \ldots, \mu_m \in R,$$

其中 $C(x, y) = \sum_{j \geq 1} \gamma_j/(1-\gamma_j) \xi_j(x) \xi_j(y)$，$|R| = \int_R \mathrm{d}x$，且 $D = -\sum_{j \geq 1} \log(1-\gamma_j)$。此类结果的证明见 Lavancier 等（2015）。

S1.2 定理 2.1 的证明

证明。 由于 Λ 为满秩， $\Lambda^T\Lambda$ 为正定且可逆，因此 Y 的条件分布 (5) 定义良好。记 $|\Lambda^T\Lambda|$ 为 $\Lambda^T\Lambda$ 的行列式，我们显式计算 Y 的边缘密度 $h(x)$。有：

$$h(x)=\int_0^{+\infty} p(x\,|\,w)\cdot p(w)dw=\int_0^{+\infty}\frac{|\Lambda^T\Lambda|^{\frac12}}{(2\pi)^{\frac d2}w^{\frac d2}}\exp\!\left(-\frac{x^T\Lambda^T\Lambda x}{2w}\right)\cdot p(w)dw$$

因此，我们得到

$$h(x)=\frac{|\Lambda^T\Lambda|^{\frac12}}{(2\pi)^{\frac d2}}\,\mathbb{E}\!\left[W^{-\frac d2}\exp\!\left(-\frac{||\Lambda x||^2}{2W}\right)\right],\qquad x\in\mathbb{R}^d \tag{S1.18}$$

因此，由于 $K_0(x)=\rho h(x)/h(0)$，我们有

$$K_0(x)=\frac{\rho}{\mathbb{E}\big[W^{-\frac d2}\big]}\,\mathbb{E}\!\left[W^{-\frac d2}\exp\!\left(-\frac{||\Lambda x||^2}{2W}\right)\right],\qquad x\in\mathbb{R}^d$$

并且，由于 $\varphi=\mathcal{F}(K_0)$，我们计算得

$$\begin{aligned}
\varphi(x)&=\int_{\mathbb{R}^d}\mathrm{e}^{-2\pi i x^Ty}\,K_0(y)dy=\\
&=\frac{\rho}{h(0)}\int_{\mathbb{R}^d}\mathrm{e}^{-2\pi i x^Ty}\,\frac{|\Lambda^T\Lambda|^{\frac12}}{(2\pi)^{\frac d2}}\,\mathbb{E}\!\left[W^{-\frac d2}\exp\!\left(-\frac{||\Lambda y||^2}{2W}\right)\right]dy=\\
&=\frac{\rho\,|\Lambda^T\Lambda|^{\frac12}}{h(0)\,(2\pi)^{\frac d2}}\int_{\mathbb{R}^d}\mathrm{e}^{-2\pi i x^Ty}\int_0^{\infty}w^{-\frac d2}\exp\!\left(-\frac{||\Lambda y||^2}{2w}\right)p(w)dw\,dy=\\
&=\frac{\rho\,|\Lambda^T\Lambda|^{\frac12}}{h(0)\,(2\pi)^{\frac d2}}\int_0^{\infty}\int_{\mathbb{R}^d}w^{-\frac d2}\exp\!\left(-2\pi i x^Ty-\frac{y^T\Lambda^T\Lambda y}{2w}\right)dy\,p(w)dw=\\
&=\frac{\rho\,|\Lambda^T\Lambda|^{\frac12}}{h(0)\,(2\pi)^{\frac d2}}\int_0^{\infty}\int_{\mathbb{R}^d}w^{-\frac d2}\exp\!\left(-\frac12\left[y^T\frac{\Lambda^T\Lambda}{w}y+4\pi i x^Ty\right]\right)dy\,p(w)dw
\end{aligned}$$

上面方括号中的项可以写成

$$\begin{aligned}
\big[...\big]=&\big(y-(-2\pi i w(\Lambda^T\Lambda)^{-1}x)\big)^T\,\frac{\Lambda^T\Lambda}{w}\,\big(y-(-2\pi i w(\Lambda^T\Lambda)^{-1}x)\big)\\
&+4\pi^2x^Tw(\Lambda^T\Lambda)^{-1}x\,.
\end{aligned}$$

将其代入 $\varphi(x)$的最后一个表达式，可得：

$$\varphi(x)=\frac{\rho}{h(0)}\int_0^{+\infty}\exp(-2\pi^2wx^T(\Lambda^T\Lambda)^{-1}x)\,p(w)dw\,.$$

综上，我们推导出

$$\varphi(x)=\frac{\rho}{h(0)}\,\mathbb{E}\!\left[\exp\big(-2\pi^2Wx^T(\Lambda^T\Lambda)^{-1}x\big)\right],\qquad x\in\mathbb{R}^d\,.$$

注意，由于 $h(x)$是一个实值随机变量的密度，利用傅里叶变换的性质，可得

$$K_0(x) = \rho \frac{h(x)}{h(0)} \in L^1(\mathbb{R}^d), \qquad \varphi = \mathcal{F}(K_0) \in L^1(\mathbb{R}^d)\,.$$

因此，为了保证存在核为 K_0 且谱密度为 $\varphi = \mathcal{F}(K_0)$ 的各向异性行列式点过程，我们参考 Lavancier 等（ 2015 ）的推论（ 3.3 ）：由于 $K_0 \in L^1(\mathbb{R}^d)$ $\varphi = \mathcal{F}(K_0)$ 且 $\varphi \in L^1(\mathbb{R}^d)$, 我们只需确保

$$\varphi(x) \leq 1, \qquad \forall x \in \mathbb{R}^d\,.$$

因此，我们需要假设

$$\max_{x\in\mathbb{R}^d} \varphi(x) = \varphi(0) = \frac{\rho}{h(0)} \leq 1\,.$$

该假设等价于

$$\rho \leq \rho_{max}$$

其中

$$\rho_{max} = h(0) = \frac{|\Lambda^T\Lambda|^{\frac{1}{2}}}{(2\pi)^{\frac{d}{2}}}\,\mathbb{E}\left[W^{-\frac{d}{2}}\right].$$

□

S1.3 定理 2.2 的证明

设 $B := \Lambda\mathbb{R}^d$。那么 $\Lambda : \mathbb{R}^d \to B$ 是一个双射。我们使用由 $\Lambda = U\Omega V^\top$ 的奇异值分解给出的正交归一基，对 B 进行坐标变换，其中 U 是一个 $p \times d$ 矩阵，其列构成 B 的正交归一基，Ω 是一个 $d \times d$ 对角矩阵，而 V 是一个 $d \times d$ 正交矩阵。令 $\Lambda^* := \Omega V^\top$，它将 $\mathbb{R}^d$ 映射到用 U 给出的坐标表示的 B。

现在考虑变换后的点 $\tilde{\mu}_h = \Lambda\mu_h$。显然，它们与点 $\mu_h^* = \Lambda^*\mu_h$ 之间存在一一对应关系，由 $\tilde{\mu}_h = U\mu_h^*$ 给出。因此，只需证明 $\Phi^* = \{\mu_1^*, \ldots, \mu_m^*\}$ 是 $B^* = \Lambda^*\mathbb{R}^d = \mathbb{R}^d$ 上的各向同性行列式点过程。考虑可测函数 h： $B^{*k} \to \mathbb{R}_+$,

$$\mathbb{E}\sum_{\boldsymbol{y}_k^*\in\Phi^*}^{\neq} h(y_1^*, \ldots, y_k^*) = \mathbb{E}\sum_{\boldsymbol{x}_k\in\Phi}^{\neq} h(\Lambda^* x_1, \ldots, \Lambda^* x_k),$$

其中求和是针对点过程支撑集内所有两两不相交的 k-元点组进行的。根据行列式点过程的定义

$$\begin{aligned}
\mathbb{E}\sum_{\boldsymbol{y}_k^*\in\Phi^*}^{\neq} h(y_1^*, \ldots, y_k^*) &= \mathbb{E}\sum_{\boldsymbol{x}_k\in\Phi}^{\neq} h(\Lambda^* x_1, \ldots, \Lambda^* x_k) \\
&= \int_{(\mathbb{R}^d)^k} \det\{K(x_i, x_j)\}_{i,j=1,\ldots,k}\, h(\Lambda^* x_1, \ldots, \Lambda^* x_k)\mathrm{d}x_1 \cdots \mathrm{d}x_k \\
&= \int_{B^{*k}} \det\{K((\Lambda^*)^{-1}y_i^*, (\Lambda^*)^{-1}y_j^*)\}_{i,j=1,\ldots,k}\, h(y_1^*, \ldots, y_k^*)\, \det(\Lambda^*)^{-k}\mathrm{d}y_1^* \cdots \mathrm{d}y_k^*
\end{aligned}$$

这表明 $\Lambda^*\mu$ 是 B^* 上的行列式点过程，其核为 $\det(\Lambda^*)^{-1}K((\Lambda^*)^{-1}y_i^*,\ (\Lambda^*)^{-1}y_j^*)$，且 B^* 上的支配测度为勒贝格测度。通过按定理 2.1 选取 $K \equiv K_0$，即可得到平稳性和各向同性。事实上，

$$\begin{aligned} K((\Lambda^*)^{-1}y_i^*, (\Lambda^*)^{-1}y_j^*) &= K_0((\Lambda^*)^{-1}(y_i^* - y_j^*)) \\ &= \frac{\rho}{\mathbb{E}[W^{-d/2}]}\mathbb{E}\left[W^{-d/2}\exp\left(-\frac{\|y_i^* - y_j^*\|^2}{2W}\right)\right] \end{aligned}$$

注意到 U 是正交的（即 $\tilde{\mu}_j$ 与 μ_j^*之间的映射是等距双射），且 $\det(\Lambda^*) = \det(\Lambda^\top\Lambda)^{1/2}$，即可得出该结果。

S1.4 推论 2.3 的证明

证明。 令 $c > 0$。在定理 2.1 中，令

$$W = |\Lambda^T\Lambda|^{\frac{1}{d}} \cdot c^{-\frac{2}{d}}$$

因此， $W^{-\frac{d}{2}} = |\Lambda^T\Lambda|^{-\frac{1}{2}} \cdot c$，并由 (S1.18) 得

$$h(x) = \frac{c}{(2\pi)^{\frac{d}{2}}} \cdot \exp\left(-\frac{||\Lambda x||^2}{2|\Lambda^T\Lambda|^{\frac{1}{d}}\, c^{-\frac{2}{d}}}\right), \qquad x \in \mathbb{R}^d .$$

由 (7) 和 (8) 得

$$K_0(x) = \rho \cdot \exp\left(-\frac{||\Lambda x||^2}{2|\Lambda^T\Lambda|^{\frac{1}{d}}\, c^{-\frac{2}{d}}}\right), \qquad x \in \mathbb{R}^d$$

$$\varphi(x) = \rho\,\frac{(2\pi)^{\frac{d}{2}}}{c} \cdot \exp\left(-2\pi^2|\Lambda^T\Lambda|^{\frac{1}{d}}c^{-\frac{2}{d}}x^T(\Lambda^T\Lambda)^{-1}x\right), \qquad x \in \mathbb{R}^d .$$

由 (6)，存在条件要求 $\rho \le \rho_{max}$，其中

$$\rho_{max} = \frac{c}{(2\pi)^{d/2}} .$$

□

S1.5 定理 3.1 的证明

证明。考虑

$$\log p(\Lambda \mid \cdots) \propto \log p(\boldsymbol{y} \mid \Lambda, \boldsymbol{\eta}, \Sigma) + \log f_{\text{DPP}}^{\text{app}}(\boldsymbol{\mu} \mid \Lambda) + \log p(\Lambda \mid \phi, \tau, \psi) \qquad \text{(S1.19)}$$

我们通过将上述三项的梯度相加，计算 Λ 的对数全条件密度的梯度。请注意，只有第二项依赖于各向异性 DPP。

由于它是最复杂的一项，我们使用下面两个引理对其进行推导。对于 (S1.19) 的第一项，我们有

$$\begin{aligned}\nabla \log p(\boldsymbol{y} \mid \Lambda, \boldsymbol{\eta}, \Sigma) &= \nabla \left(-\frac{1}{2} \sum_{i=1}^{n} (y_i - \Lambda\eta_i)^\top \Sigma^{-1} (y_i - \Lambda\eta_i) \right) \\ &= \Sigma^{-1} \cdot \sum_{i=1}^{n} (y_i - \Lambda\eta_i)\eta_i^T \,.\end{aligned}$$

最后一项的梯度为

$$\nabla \log p(\Lambda \mid \phi, \tau, \psi) = -\frac{1}{(\psi \odot \phi^2)\tau^2} \odot \Lambda \,.$$

(S1.19) 中第二项的梯度为,

$$\begin{aligned}\nabla \log f_{\text{DPP}}^{\text{app}}(\boldsymbol{\mu} \mid \Lambda) &= \nabla \left[-D^{\text{app}} - \log(1 - \mathrm{e}^{-D^{\text{app}}}) + \log\det[C^{\text{app}}](T\mu_1, \ldots, T\mu_n) \right] \\ &= -\nabla D^{\text{app}} - \nabla \log(1 - \mathrm{e}^{-D^{\text{app}}}) + \nabla \log\det[C^{\text{app}}](T\mu_1, \ldots, T\mu_n) \\ &= -\frac{1}{1 - \mathrm{e}^{-D^{\text{app}}}} \nabla D^{\text{app}} + \nabla \log\det[C^{\text{app}}](T\mu_1, \ldots, T\mu_n) \qquad \text{(S1.20)}\end{aligned}$$

为处理 (S1.20)，需要计算项 ∇D^{app} 和 $\nabla \log\det[C^{\text{app}}](T\mu_1, \ldots, T\mu_n)$。

引理 S1.1。 对于类高斯 *DPP* 先验,

$$\nabla D^{\text{app}} = \sum_{k \in \mathbb{Z}_N^d} \frac{\varphi(k)}{1 - \varphi(k)} (-2\pi^2 c^{-\frac{2}{d}}) g^{(k)} \qquad \text{(S1.21)}$$

其中 $\varphi(k)$指的是 (9)。

证明。记

$$\nabla D^{\text{app}} = -\sum_{k \in \mathbb{Z}_N^d} \nabla \log(1 - \varphi(k)) = \sum_{k \in \mathbb{Z}_N^d} \frac{1}{1 - \varphi(k)} \nabla\varphi(k).$$

对于类高斯 DPP 先验，由式（9）可知，对于 $k \in \mathbb{Z}^d$

$$\begin{aligned}\nabla\varphi(k) &= \frac{\rho}{c} (2\pi)^{\frac{d}{2}} \nabla \exp\left(-2\pi^2 |\Lambda^T\Lambda|^{\frac{1}{d}} c^{-\frac{2}{d}} k^T (\Lambda^T\Lambda)^{-1} k \right) \\ &= \varphi(k) \left(-2\pi^2 c^{-\frac{2}{d}} \right) \nabla \left(|\Lambda^T\Lambda|^{\frac{1}{d}} k^T (\Lambda^T\Lambda)^{-1} k \right) \,.\end{aligned}$$

注意,

$$\begin{aligned}\nabla \left(|\Lambda^T\Lambda|^{\frac{1}{d}} k^T (\Lambda^T\Lambda)^{-1} k \right) =& \nabla \left(|\Lambda^T\Lambda|^{\frac{1}{d}} \right) k^T (\Lambda^T\Lambda)^{-1} k + \\ &+ |\Lambda^T\Lambda|^{\frac{1}{d}} \nabla \left(k^T (\Lambda^T\Lambda)^{-1} k \right) = \\ =& \frac{1}{d} |\Lambda^T\Lambda|^{\frac{1}{d}-1} 2|\Lambda^T\Lambda| \Lambda (\Lambda^T\Lambda)^{-1} k^T (\Lambda^T\Lambda)^{-1} k + \\ &+ |\Lambda^T\Lambda|^{\frac{1}{d}} \nabla \left(k^T (\Lambda^T\Lambda)^{-1} k \right)\end{aligned}$$

其中，最后一步使用了 Petersen 与 Pedersen（2008年）的公式（53）。于是，

$$\begin{aligned}\nabla\left(k^T(\Lambda^T\Lambda)^{-1}k\right) &= \nabla\mathrm{tr}\left(k^T(\Lambda^T\Lambda)^{-1}k\right) = \\ &= \nabla\mathrm{tr}\left((\Lambda^T\Lambda)^{-1}kk^T\right) = -\Lambda(\Lambda^T\Lambda)^{-1}(2kk^T)(\Lambda^T\Lambda)^{-1}\end{aligned}$$

其中，在最后一步中，我们应用了 Petersen 与 Pedersen（2008） 的公式（125）。对于类高斯 DPP 先验，这得出

$$\nabla\varphi(k) = \varphi(k)(-4\pi^2c^{-\frac{2}{d}})|\Lambda^T\Lambda|^{\frac{1}{d}}\Lambda(\Lambda^T\Lambda)^{-1}\Big[\frac{1}{d}k^T(\Lambda^T\Lambda)^{-1}k\mathbb{1}_d - kk^T(\Lambda^T\Lambda)^{-1}\Big]$$

其中， $\mathbb{1}_d$ 是由 1 构成的 $d\times d$ 矩阵。这就完成了引理 S1.1 的证明。 □

引理 S1.2。 对于类高斯 *DPP* 先验，

$$\nabla\log\det[C^{\mathrm{app}}] = (-2\pi^2c^{-\frac{2}{d}})\sum_{k\in\mathbb{Z}_N^d}\frac{\varphi(k)}{(1-\varphi(k))^2}g^{(k)}v_k^T(C^{\mathrm{app}})^{-1}u_k \tag{S1.22}$$

其中，φ（k）指的是（9）。

证明。我们有

$$\begin{aligned}\frac{\partial}{\partial\Lambda_{ij}}\log\det[C^{\mathrm{app}}] &= \mathrm{tr}\left(\left(\frac{\partial}{\partial U}\log\det U\right)\Big|_{U=C^{\mathrm{app}}}\cdot\frac{\partial}{\partial\Lambda_{ij}}C^{\mathrm{app}}\right) &\text{(S1.23)}\\ &= \mathrm{tr}\left(\frac{1}{\det[C^{\mathrm{app}}]}\left(\frac{\partial}{\partial U}\det U\right)\Big|_{U=C^{\mathrm{app}}}\cdot\frac{\partial}{\partial\Lambda_{ij}}C^{\mathrm{app}}\right)\\ &= \mathrm{tr}\left((C^{\mathrm{app}})^{-1}\cdot\frac{\partial}{\partial\Lambda_{ij}}C^{\mathrm{app}}\right) &\text{(S1.24)}\end{aligned}$$

其中，使用 Petersen 与 Pedersen（2008） 的公式（137）得到（S1.23），并使用 Petersen 与 Pedersen（2008） 的公式（49）得到（S1.24）。现在，注意到

$$C^{\mathrm{app}} = \sum_{k\in\mathbb{Z}_N^d}\frac{\varphi(k)}{1-\varphi(k)}u_kv_k^T$$

其中 u_k, v_k 在该命题中定义。因此，对于类高斯 DPP 先验，

$$\frac{\partial}{\partial\Lambda_{ij}}C^{\mathrm{app}} = \sum_{k\in\mathbb{Z}_N^d}\frac{\varphi(k)}{(1-\varphi(k))^2}(-2\pi^2c^{-\frac{2}{d}})g_{ij}^{(k)}u_kv_k^T$$

其中 $g_{ij}^{(k)} = (g^{(k)})_{ij}$。接下来，回到（S1.24），对于类高斯 DPP 先验，我们有 e

$$\begin{aligned}
\frac{\partial}{\partial \Lambda_{ij}} \log\det[C^{\mathrm{app}}] &= \operatorname{tr}\left((C^{\mathrm{app}})^{-1} \cdot \frac{\partial}{\partial \Lambda_{ij}} C^{\mathrm{app}}\right) \\
&= \operatorname{tr}\left(\sum_{k\in\mathbb{Z}_N^d} \frac{\varphi(k)}{(1-\varphi(k))^2}(-2\pi^2 c^{-\frac{2}{d}}) g_{ij}^{(k)} (C^{\mathrm{app}})^{-1} u_k v_k^T\right) \\
&= \sum_{k\in\mathbb{Z}_N^d} \frac{\varphi(k)}{(1-\varphi(k))^2}(-2\pi^2 c^{-\frac{2}{d}}) g_{ij}^{(k)} \operatorname{tr}\left((C^{\mathrm{app}})^{-1} u_k v_k^T\right) \\
&= (-2\pi^2 c^{-\frac{2}{d}}) \sum_{k\in\mathbb{Z}_N^d} \frac{\varphi(k)}{(1-\varphi(k))^2} g_{ij}^{(k)} v_k^T (C^{\mathrm{app}})^{-1} u_k \,.
\end{aligned}$$

由此完成引理 S1.2 的证明。 □

回到定理 3.1 的证明，在类高斯 DPP 先验的情形下，由 (S1.21) 和 (S1.22)，公式 (S1.20) 可得

$$\nabla \log f_{\mathrm{DPP}}^{\mathrm{app}}(\boldsymbol{\mu} \,|\, \Lambda) = (2\pi^2 c^{-\frac{2}{d}}) \sum_{k\in\mathbb{Z}_N^d} g^{(k)} \frac{\varphi(k)}{(1-\varphi(k))^2}\Big[\frac{1-\varphi(k)}{1-\mathrm{e}^{-D^{\mathrm{app}}}} - v_k^T (C^{\mathrm{app}})^{-1} u_k\Big]$$

由此完成定理 3.1 的证明。

□

S2 测度论细节

根据正文中的讨论，我们有 $\boldsymbol{y} \in \mathbb{R}^{p\times n}$, $\boldsymbol{\eta} \in \mathbb{R}^{d\times n}$, $\Sigma \in \mathbb{R}_+^p$, $\Lambda \in \mathbb{R}^{p\times d}$, $\boldsymbol{\psi} \in \mathbb{R}_+^{p\times d}$, $\boldsymbol{\phi} \in \mathbb{S}^{p\times d-1}$（$pd-1$ 维单纯形），以及 $\tau \in \mathbb{R}_+$。此外，我们将 $\boldsymbol{\mu}$ 视为随机点配置，其取值于 $\Omega = \cup_{m=0}^{\infty}\Omega_m$，其中 Ω_m 表示由 $\mathbb{R}^d$ 中（两两不同的）m-元组构成的空间。我们在每个 Ω_m上赋予使下列映射可测的最小 σ-代数

$$(\mu_1, \ldots, \mu_m) \mapsto \{\mu_1, \ldots, \mu_m\}\,,$$

其中，左侧将 $\mu_1, \ldots, \mu_m$ 视为 R^m 中的有序向量，右侧则将其视为 R中的无序点集；其中，$R \subset \mathbb{R}^d$ 是定义 μ 的超立方体。于是，Ω 上的 σ-代数是包含各个 Ω_m 上所有 σ-代数之并的最小 σ-代数。由此可得 $(\boldsymbol{\mu}, \boldsymbol{s}, \boldsymbol{\Delta}, \boldsymbol{c}) \in \cup_{m=0}^{\infty}\{\Omega_m \times \mathbb{R}_+^m \times \mathcal{SP}_d^m \times \{1, \ldots, m\}^n\}$，其中 $\mathcal{SP}_d$ 表示 $d\times d$ 对称正定矩阵空间。

现在考虑集合 $\mathfrak{Y} \subset \mathbb{R}^{p\times n}$, $\mathfrak{N} \subset R^{d\times n}$, $\Xi \subset \mathbb{R}_+^p$, $\Lambda \subset \mathbb{R}^{p\times d}$, $\Psi \subset \mathbb{R}_+^{p\times d}$, $\Phi \subset \mathbb{S}^{p\times d-1}$ $\mathfrak{T} \subset \mathbb{R}_+$, $\mathfrak{O}_m \subset \Omega_m$, $\mathfrak{S}_m \subset \mathbb{R}_+^m$, $\mathfrak{D}_m \subset \mathcal{SP}_d^m$, $\mathfrak{C}_m \subset \{1, \ldots, m\}^n$。该

数据与参数的联合分布的支配测度为

$$\begin{aligned}
&\nu\left(\mathfrak{Y}\times\mathfrak{N}\times\times\Xi\times\Lambda\times\Psi\times\Phi\times\mathfrak{T}\times\cup_{m\geq 0}\left\{\mathfrak{O}_m\times\mathfrak{S}_m\times\mathfrak{D}_m\times\mathfrak{C}_m\right\}\right)=\\
&\quad\int_{\mathfrak{Y}}\mathrm{d}\boldsymbol{y}\times\int_{\mathfrak{n}}\mathrm{d}\boldsymbol{\eta}\times\int_{\Xi}\mathrm{d}\boldsymbol{\Sigma}\times\int_{\Lambda}\mathrm{d}\Lambda\times\int_{\Psi}\mathrm{d}\boldsymbol{\psi}\times\int_{\Phi}\mathrm{d}\boldsymbol{\phi}\times\int_{\mathfrak{T}}\mathrm{d}\tau\times\\
&\qquad\times\sum_{m=0}^{\infty}\frac{\mathrm{e}^{-|R|}}{m!}\int_{\mathfrak{O}_m}\mathrm{d}\mu_m\times\int_{\mathfrak{S}_m}\mathrm{d}s_m\times\int_{\mathfrak{D}_m}\mathrm{d}\Delta_m\times\sum_{c_1,\ldots,c_n=1}^{M}\mathbb{1}[\boldsymbol{c}\in\mathfrak{C}_m]\,.
\end{aligned}$$

数据和参数关于 ν 的密度为:

$$\begin{aligned}
&p(\boldsymbol{y},\boldsymbol{c},\boldsymbol{\eta},\boldsymbol{\mu},\boldsymbol{S},\boldsymbol{\Delta},\Sigma,\Lambda,\boldsymbol{\psi},\boldsymbol{\phi},\tau)=\\
&\quad\frac{1}{T^n}\left[\prod_{i=1}^{n}\mathcal{N}_p(y_i\,|\,\Lambda\eta_i,\Sigma)\right]\left[\prod_{h=1}^{k}(S_h^{(a)})^{n_h}\mathrm{Ga}(S_h^{(a)}\,|\,\alpha,1)\mathrm{IW}(\Delta_h^{(a)}\,|\,\nu_0,\Psi_0)\prod_{i:c_i=h}\mathcal{N}_d(\eta_i\,|\,\mu_h^{(a)},\Delta_h^{(a)})\right]\\
&\quad\left[\prod_{h=1}^{\ell}\mathrm{Ga}(S_h^{(na)}\,|\,\alpha,1)\mathrm{IW}(\Delta_h^{(na)}\,|\,\nu_0,\Psi_0)\right]f_{\text{DPP}}^{\text{app}}(\boldsymbol{\mu}^{(a)}\cup\boldsymbol{\mu}^{(na)}\,|\,\rho,\Lambda,K_0;R)\\
&\quad\left[\prod_{j=1}^{p}\text{inv-Ga}(\sigma_j^2\,|\,a_\sigma,b_\sigma)\prod_{h=1}^{d}\mathcal{N}(\lambda_{jh}\,|\,0,\psi_{jh}\phi_{jh}^2\tau)\mathrm{Exp}(\Psi_{jw}\,|\,1/2)\right]\mathrm{Dir}(vec(\phi)\,|\,a)\mathrm{Ga}(\tau\,|\,pda,1/2)
\end{aligned}$$

现引入辅助变量 u，使得 $u\,|\,T\sim\mathrm{Ga}(n,t)$，并考虑包含 $u\in\mathbb{R}_+$ 的扩展参数空间。此外，在以 $\boldsymbol{c}$ 为条件时，我们将 $\boldsymbol{\mu}=\mu^{(a)}\cup\mu^{(na)}$、$\boldsymbol{S}=[\boldsymbol{S}^{(a)},\boldsymbol{S}^{(na)}]$ 和 $\boldsymbol{\Delta}=[\boldsymbol{\Delta}^{(a)},\boldsymbol{\Delta}^{(na)}]$ 划分为已分配分量和未分配分量（分别以上标 (a) 和 (na) 表示）。扩展空间上的支配测度 ν' 可以直接推导得到。例如，参见 Beraha 等（2022）中的公式（17）。相对于 ν'，数据和参数的联合密度则为

$$\begin{aligned}
&p(\boldsymbol{y},\boldsymbol{c},\boldsymbol{\eta},\boldsymbol{\mu}^{(a)},\boldsymbol{\mu}^{(na)},\boldsymbol{S}^{(a)},\boldsymbol{S}^{(na)}\boldsymbol{\Delta}^{(a)},\boldsymbol{\Delta}^{(na)},\Sigma,\Lambda,\boldsymbol{\psi},\boldsymbol{\phi},\tau,u)\\
&\quad=\frac{u^{n-1}}{\Gamma(n)}\left[\prod_{i=1}^{n}\mathcal{N}_p(y_i\,|\,\Lambda\eta_i,\Sigma)\right]\left[\prod_{h=1}^{k}\mathrm{e}^{-uS_h^{(a)}}(S_h^{(a)})^{n_h}\mathrm{Ga}(S_h^{(a)}\,|\,\alpha,1)\mathrm{IW}(\Delta_h^{(a)}\,|\,\nu_0,\Psi_0)\right.\\
&\qquad\left.\times\prod_{i:c_i=h}\mathcal{N}_d(\eta_i\,|\,\mu_h^{(a)},\Delta_h^{(a)})\right]\left[\prod_{h=1}^{\ell}\mathrm{e}^{-uS_h^{(na)}}\mathrm{Ga}(S_h^{(na)}\,|\,\alpha,1)\mathrm{IW}(\Delta_h^{(na)}\,|\,\nu_0,\Psi_0)\right]\\
&\quad f_{\text{DPP}}^{\text{app}}(\boldsymbol{\mu}^{(a)}\cup\boldsymbol{\mu}^{(na)}\,|\,\rho,\Lambda,K_0;R)\prod_{j=1}^{p}\left[\text{inv-Ga}(\sigma_j^2\,|\,a_\sigma,b_\sigma)\prod_{h=1}^{d}\mathcal{N}(\lambda_{jh}\,|\,0,\psi_{jh}\phi_{jh}^2\tau)\right.\\
&\qquad\left.\times\prod_{h=1}^{d}\mathrm{Exp}(\Psi_{jd}\,|\,1/2)\right]\mathrm{Dir}(vec(\phi)\,|\,a)\mathrm{Ga}(\tau\,|\,pda,1/2)\,.
\end{aligned}$$

S3 各向异性 Whittle-Matérn DPP

推论 S3.1。 沿用定理 *2.1* 中的记号，令

$$W \sim \mathrm{Gamma}\left(\nu + \frac{d}{2}, \frac{1}{2|\Lambda^T\Lambda|^{\frac{1}{d}}\alpha^2}\right), \qquad \nu, \alpha > 0,$$

其中 Λ 为固定值。则核 K_0、其傅里叶变换 $\varphi = \mathcal{F}(K_0)$ 以及 $\rho_{\max}$ 如下：

$$\begin{aligned} K_0(x) &= \rho\,\frac{2^{1-\nu}}{\Gamma(\nu)}\left\|\frac{\Lambda x}{\alpha\,|\Lambda^T\Lambda|^{\frac{1}{2d}}}\right\|^{\nu} K_\nu\left(\left\|\frac{\Lambda x}{\alpha\,|\Lambda^T\Lambda|^{\frac{1}{2d}}}\right\|\right), \qquad x \in \mathbb{R}^d \\ \varphi(x) &= \rho\,\frac{\Gamma(\nu+\frac{d}{2})}{\Gamma(\nu)}\,\frac{(2\sqrt{\pi}\alpha)^d}{\left(1+4\pi^2\alpha^2|\Lambda^T\Lambda|^{\frac{1}{d}}\,x^T(\Lambda^T\Lambda)^{-1}x\right)^{\nu+\frac{d}{2}}}, \qquad x \in \mathbb{R}^d \\ \rho_{\max} &= \frac{\Gamma(\nu)}{\Gamma(\nu+\frac{d}{2})\,(2\sqrt{\pi}\alpha)^d} \end{aligned} \tag{S3.25}$$

其中 K_ν 是第二类修正贝塞尔函数。

证明。 令 $\nu > 0$ 和 $\alpha > 0$。在定理 2.1 中，设

$$W \sim \mathrm{Gamma}\left(\nu + \frac{d}{2}, \frac{1}{2|\Lambda^T\Lambda|^{\frac{1}{d}}\alpha^2}\right).$$

应用 (S1.18)，我们显式计算 $h(x)$，得

$$\begin{aligned} h(x) &= \frac{|\Lambda^T\Lambda|^{\frac{1}{2}}}{(2\pi)^{\frac{d}{2}}}\int_0^{+\infty}\left[w^{-\frac{d}{2}}\exp\left(-\frac{||\Lambda x||^2}{2w}\right)\left(\frac{1}{2|\Lambda^T\Lambda|^{\frac{1}{d}}\alpha^2}\right)^{\nu+\frac{d}{2}}\frac{w^{\nu+\frac{d}{2}-1}}{\Gamma\left(\nu+\frac{d}{2}\right)}\right. \\ &\qquad\left.\times\exp\left(-\frac{w}{2|\Lambda^T\Lambda|^{\frac{1}{d}}\alpha^2}\right)\right]dw \\ &= \frac{|\Lambda^T\Lambda|^{-\frac{\nu}{d}}}{(2\pi)^{\frac{d}{2}}(2\alpha^2)^{\nu+\frac{d}{2}}\Gamma\left(\nu+\frac{d}{2}\right)}\int_0^{+\infty} w^{\nu-1}\exp\left[-\frac{1}{2}\left(\frac{1}{|\Lambda^T\Lambda|^{\frac{1}{d}}\alpha^2}w + \frac{||\Lambda x||^2}{w}\right)\right]dw\,. \end{aligned}$$

本文中，参数为 (p, a, b) 的 $[0, +\infty)$ 上的广义逆高斯分布（GIG），指的是密度为

$$f(x) = \frac{(a/b)^{\frac{p}{2}}}{2K_p(\sqrt{ab})}x^{p-1}\exp\left(-\frac{1}{2}\left[ax + \frac{b}{x}\right]\right),\ x > 0 \tag{S3.26}$$

其中 $p \in \mathbb{R}$, $a > 0$ 和 $b > 0$。在上述表达式中， K_p 是第二类修正贝塞尔函数。我们用 GIG(p, a, b)表示这种分布。

回到 $h(x)$ 的计算，我们从最后一个积分中看到，被积函数是一个参数为以下数值的广义逆高斯分布的密度

$$p=\nu, \qquad a=\frac{1}{|\Lambda^T\Lambda|^{\frac{1}{d}}\alpha^2}, \qquad b=||\Lambda x||^2.$$

因此,

$$\begin{aligned} h(x) &= \frac{|\Lambda^T\Lambda|^{-\frac{\nu}{d}}}{(2\pi)^{\frac{d}{2}}(2\alpha^2)^{\nu+\frac{d}{2}}\,\Gamma\big(\nu+\frac{d}{2}\big)}\cdot\left(|\Lambda^T\Lambda|^{\frac{1}{d}}\alpha^2\,||\Lambda x||^2\right)^{\frac{\nu}{2}}\cdot 2\,K_\nu\left(\frac{||\Lambda x||}{|\Lambda^T\Lambda|^{\frac{1}{2d}}\,\alpha}\right) \\ &= \frac{1}{(\sqrt{\pi}\,\alpha)^d\;2^{\nu+d-1}\;\Gamma\big(\nu+\frac{d}{2}\big)}\cdot\left\|\frac{\Lambda x}{\alpha\,|\Lambda^T\Lambda|^{\frac{1}{2d}}}\right\|^{\nu}\cdot K_\nu\left(\left\|\frac{\Lambda x}{\alpha\,|\Lambda^T\Lambda|^{\frac{1}{2d}}}\right\|\right). \end{aligned}$$

注意，由于 $x\to 0$，因此 $x^\nu K_\nu(x)\to 2^{\nu-1}\Gamma(\nu)$。因此，核 K_0 是

$$K_0(x)=\rho\,\frac{2^{1-\nu}}{\Gamma(\nu)}\cdot\left\|\frac{\Lambda x}{\alpha\,|\Lambda^T\Lambda|^{\frac{1}{2d}}}\right\|^{\nu}\cdot K_\nu\left(\left\|\frac{\Lambda x}{\alpha\,|\Lambda^T\Lambda|^{\frac{1}{2d}}}\right\|\right).$$

为推导谱密度 $\varphi=\mathcal{F}(K_0)$，由式（8）可得

$$\begin{aligned} \varphi(x) &= \frac{\rho}{h(0)}\int_0^\infty \exp\big(-2\pi^2 w x^T(\Lambda^T\Lambda)^{-1}x\big)\left(\frac{1}{2|\Lambda^T\Lambda|^{\frac{1}{d}}\alpha^2}\right)^{\nu+\frac{d}{2}}\frac{w^{\nu+\frac{d}{2}-1}}{\Gamma\big(\nu+\frac{d}{2}\big)}\exp\left(-\frac{w}{2|\Lambda^T\Lambda|^{\frac{1}{d}}\alpha^2}\right)dw \\ &= \frac{\rho}{h(0)\left(2|\Lambda^T\Lambda|^{\frac{1}{d}}\alpha^2\right)^{\nu+\frac{d}{2}}}\int_0^\infty\frac{w^{\nu+\frac{d}{2}-1}}{\Gamma\big(\nu+\frac{d}{2}\big)}\,\exp\left[-\left(2\pi^2 x^T(\Lambda^T\Lambda)^{-1}x+\frac{1}{2|\Lambda^T\Lambda|^{\frac{1}{d}}\alpha^2}\right)w\right]dw \\ &= \frac{\rho}{h(0)\left(2|\Lambda^T\Lambda|^{\frac{1}{d}}\alpha^2\right)^{\nu+\frac{d}{2}}}\cdot\frac{1}{\left(2\pi^2 x^T(\Lambda^T\Lambda)^{-1}x+\frac{1}{2|\Lambda^T\Lambda|^{1/d}\,\alpha^2}\right)^{\nu+\frac{d}{2}}} \\ &= \frac{\rho}{h(0)}\cdot\frac{1}{\left(1+4\pi^2\alpha^2|\Lambda^T\Lambda|^{\frac{1}{d}}\,x^T(\Lambda^T\Lambda)^{-1}x\right)^{\nu+\frac{d}{2}}}, \end{aligned}$$

因此，谱密度 $\varphi=\mathcal{F}(K_0)$为

$$\varphi(x)=\rho\,\frac{\Gamma\big(\nu+\frac{d}{2}\big)}{\Gamma(\nu)}\,\frac{(2\sqrt{\pi}\alpha)^d}{\left(1+4\pi^2\alpha^2|\Lambda^T\Lambda|^{\frac{1}{d}}\,x^T(\Lambda^T\Lambda)^{-1}x\right)^{\nu+\frac{d}{2}}}.$$

总而言之，我们报告了该模型的核 K_0 和谱密度 φ

$$K_0(x)=\rho\,\frac{2^{1-\nu}}{\Gamma(\nu)}\cdot\left\|\frac{\Lambda x}{\alpha\,|\Lambda^T\Lambda|^{\frac{1}{2d}}}\right\|^{\nu}\cdot K_\nu\left(\left\|\frac{\Lambda x}{\alpha\,|\Lambda^T\Lambda|^{\frac{1}{2d}}}\right\|\right)$$

$$\varphi(x)=\rho\,\frac{\Gamma\big(\nu+\frac{d}{2}\big)}{\Gamma(\nu)}\,\frac{(2\sqrt{\pi}\alpha)^d}{\left(1+4\pi^2\alpha^2|\Lambda^T\Lambda|^{\frac{1}{d}}\,x^T(\Lambda^T\Lambda)^{-1}x\right)^{\nu+\frac{d}{2}}}.$$

根据公式（6），存在条件要求 $\rho \leq \rho_{max}$，其中

$$\rho_{max} = \frac{\Gamma(\nu)}{\Gamma\left(\nu + \frac{d}{2}\right)(2\sqrt{\pi}\alpha)^d}.$$

□

现在我们说明如何在 Whittle-Matérn DPP 先验的情况下，也解析地计算 Λ 全条件的梯度。

定理 S3.2. 在 *Whittle-Matérn* 类 *DPP* 先验下，Λ 的对数全条件密度的梯度等于

$$\begin{aligned}\nabla \log p(\Lambda \mid \cdots) = \Sigma^{-1} \sum_{i=1}^{n} (y_i - \Lambda\eta_i)\eta_i^\top + \\ + 4\pi^2\alpha^2 \left(\nu + \frac{d}{2}\right) \sum_{k \in \mathbb{Z}_N^d} \frac{g^{(k)}}{a^{(k)}} \frac{\varphi(k)}{(1-\varphi(k))^2} \Big[\frac{1-\varphi(k)}{1-\mathrm{e}^{-D^{\mathrm{app}}}} - v_k^T \left(C^{\mathrm{app}}\right)^{-1} u_k\Big] + \\ - \frac{1}{(\psi \odot \phi^2)\tau^2} \odot \Lambda\end{aligned}$$

其中 $\varphi(k)$ 定义见 *(S3.25)*，$\odot$ 表示逐元素（*Hadamard*）乘积，

$$g^{(k)} = 2|\Lambda^T\Lambda|^{\frac{1}{d}} \Lambda(\Lambda^T\Lambda)^{-1} \Big[\frac{1}{d} k^T(\Lambda^T\Lambda)^{-1}k \mathbb{1}_d - k((\Lambda^T\Lambda)^{-1}k)^T\Big],$$

u_k 和 v_k 是对每个 $k \in \mathbb{Z}^d$ 而言的 m 维列向量，其元素为

$$(u_k)_j = \mathrm{e}^{2\pi i k^T T\mu_j}, \quad (v_k)_j = \mathrm{e}^{-2\pi i k^T T\mu_j}, \quad j = 1, \ldots, m$$

$C^{\mathrm{app}} := C^{\mathrm{app}}(T\mu_1, \ldots, T\mu_n)$，以及 $a^{(k)} = 1 + 4\pi^2\alpha^2|\Lambda^T\Lambda|^{\frac{1}{d}} k^T(\Lambda^T\Lambda)^{-1}k$。

证明。 证明过程与定理 3.1 的证明基本相同。我们用以下两个结果替代引理 S1.1 和引理 S1.2

引理 S3.3. 对于 *Whittle-Matérn* 类 *DPP* 先验，

$$\nabla D^{\mathrm{app}} = \sum_{k \in \mathbb{Z}_N^d} \frac{\varphi(k)}{1-\varphi(k)} 4\pi^2\alpha^2 \left(-\nu - \frac{d}{2}\right) \frac{g^{(k)}}{a^{(k)}} \tag{S3.27}$$

其中 $\varphi(k)$指的是 *(S3.25)*。

证明。写出

$$\nabla D^{\mathrm{app}} = -\sum_{k \in \mathbb{Z}_N^d} \nabla \log(1-\varphi(k)) = \sum_{k \in \mathbb{Z}_N^d} \frac{1}{1-\varphi(k)} \nabla\varphi(k).$$

对于 Whittle-Matérn 类 DPP 先验，由式 (S3.25) 可知，对于 $k \in \mathbb{Z}^d$

$$\begin{aligned}\nabla\varphi(k) &= \rho\,\frac{\Gamma(\nu+\frac{d}{2})}{\Gamma(\nu)}\,(2\sqrt{\pi}\alpha)^d\nabla\left(\left(1+4\pi^2\alpha^2|\Lambda^T\Lambda|^{\frac{1}{d}}\,k^T(\Lambda^T\Lambda)^{-1}k\right)^{-\nu-\frac{d}{2}}\right)\\ &= \frac{\varphi(k)}{a^{(k)}}4\pi^2\alpha^2\left(-\nu-\frac{d}{2}\right)\nabla\left(|\Lambda^T\Lambda|^{\frac{1}{d}}k^T(\Lambda^T\Lambda)^{-1}k\right).\end{aligned}$$

注意

$$\begin{aligned}\nabla\left(|\Lambda^T\Lambda|^{\frac{1}{d}}k^T(\Lambda^T\Lambda)^{-1}k\right) =&\nabla\left(|\Lambda^T\Lambda|^{\frac{1}{d}}\right)k^T(\Lambda^T\Lambda)^{-1}k+\\ &+|\Lambda^T\Lambda|^{\frac{1}{d}}\nabla\left(k^T(\Lambda^T\Lambda)^{-1}k\right)=\\ =&\frac{1}{d}|\Lambda^T\Lambda|^{\frac{1}{d}-1}2|\Lambda^T\Lambda|\Lambda(\Lambda^T\Lambda)^{-1}k^T(\Lambda^T\Lambda)^{-1}k+\\ &+|\Lambda^T\Lambda|^{\frac{1}{d}}\nabla\left(k^T(\Lambda^T\Lambda)^{-1}k\right)\end{aligned}$$

其中，在最后一步中使用了 Petersen 与 Pedersen（2008）的公式（53）。于是，

$$\begin{aligned}\nabla\left(k^T(\Lambda^T\Lambda)^{-1}k\right) &= \nabla\mathrm{tr}\left(k^T(\Lambda^T\Lambda)^{-1}k\right)=\\ &= \nabla\mathrm{tr}\left((\Lambda^T\Lambda)^{-1}kk^T\right) = -\Lambda(\Lambda^T\Lambda)^{-1}(2kk^T)(\Lambda^T\Lambda)^{-1}\end{aligned}$$

其中，在最后一步中应用了 Petersen 与 Pedersen（2008）的公式（125）。对于 Whittle-Matérn 类 DPP 先验，可得

$$\nabla\varphi(k) = \frac{\varphi(k)}{a^{(k)}}\,8\pi^2\alpha^2\left(-\nu-\frac{d}{2}\right)|\Lambda^T\Lambda|^{\frac{1}{d}}\Lambda(\Lambda^T\Lambda)^{-1}\left[\frac{1}{d}k^T(\Lambda^T\Lambda)^{-1}k\mathbb{1}_d - kk^T(\Lambda^T\Lambda)^{-1}\right]$$

其中 $\mathbb{1}_d$ 是由 1 组成的 $d\times d$ 矩阵。引理 S3.3 得证。 □

引理 S3.4。 对于 *Whittle-Matérn* 类 *DPP* 先验，

$$\nabla\log\det[C^{\mathrm{app}}] = 4\pi^2\alpha^2\left(-\nu-\frac{d}{2}\right)\sum_{k\in\mathbb{Z}_N^d}\frac{\varphi(k)}{(1-\varphi(k))^2}\frac{g^{(k)}}{a^{(k)}}v_k^T(C^{\mathrm{app}})^{-1}u_k \tag{S3.28}$$

其中 $\varphi(k)$ 指的是（*S3.25*）。

证明。 按照引理 S3.4 的证明过程，我们有

$$\begin{aligned}\frac{\partial}{\partial\Lambda_{ij}}\log\det[C^{\mathrm{app}}] &= \mathrm{tr}\left(\left(\frac{\partial}{\partial U}\log\det U\right)\Big|_{U=C^{\mathrm{app}}}\cdot\frac{\partial}{\partial\Lambda_{ij}}C^{\mathrm{app}}\right) &\text{(S3.29)}\\ &= \mathrm{tr}\left(\frac{1}{\det[C^{\mathrm{app}}]}\left(\frac{\partial}{\partial U}\det U\right)\Big|_{U=C^{\mathrm{app}}}\cdot\frac{\partial}{\partial\Lambda_{ij}}C^{\mathrm{app}}\right)\\ &= \mathrm{tr}\left((C^{\mathrm{app}})^{-1}\cdot\frac{\partial}{\partial\Lambda_{ij}}C^{\mathrm{app}}\right) &\text{(S3.30)}\end{aligned}$$

其中，利用 Petersen 与 Pedersen（2008）的公式（137）得到（S3.29），并利用 Petersen 与 Pedersen（2008）的公式（49）得到（S3.30）。现在注意到，

$$C^{\text{app}} = \sum_{k \in \mathbb{Z}_N^d} \frac{\varphi(k)}{1-\varphi(k)} u_k v_k^T$$

其中 u_k， v_k 在陈述中定义。因此，对于 Whittle-Matérn 类 DPP 先验，

$$\frac{\partial}{\partial \Lambda_{ij}} C^{\text{app}} = \sum_{k \in \mathbb{Z}_N^d} \frac{\varphi(k)}{(1-\varphi(k))^2} 4\pi^2 \alpha^2 \left(-\nu - \frac{d}{2}\right) \frac{g_{ij}^{(k)}}{a^{(k)}} u_k v_k^T .$$

接下来，回到公式 (S3.30)，对于 Whittle-Matérn 类 DPP 先验，

$$\frac{\partial}{\partial \Lambda_{ij}} \log\det[C^{\text{app}}] = 4\pi^2 \alpha^2 \left(-\nu - \frac{d}{2}\right) \sum_{k \in \mathbb{Z}_N^d} \frac{\varphi(k)}{(1-\varphi(k))^2} \frac{g_{ij}^{(k)}}{a^{(k)}} v_k^T (C^{\text{app}})^{-1} u_k$$

由此完成引理 S3.4 的证明 □

回到定理 S3.2 的证明，对于 Whittle-Matérn 类 DPP 先验的情形，由 (S3.27) 和 (S3.28)，公式 (S1.20) 可得

$$\nabla \log f_{\text{DPP}}^{\text{app}}(\boldsymbol{\mu} \,|\, \Lambda) = 4\pi^2 \alpha^2 \left(\nu + \frac{d}{2}\right) \sum_{k \in \mathbb{Z}_N^d} \frac{g^{(k)}}{a^{(k)}} \frac{\varphi(k)}{(1-\varphi(k))^2} \Big[\frac{1-\varphi(k)}{1-\mathrm{e}^{-D^{\text{app}}}} - v_k^T (C^{\text{app}})^{-1} u_k \Big]$$

由此完成定理 S3.2 的证明。 □

S4 马尔可夫链蒙特卡洛算法的更多细节

S4.1 吉布斯采样器

下面我们概述吉布斯采样器的步骤；若全条件分布的计算较为直接，则省略相关计算。我们用 $X|\cdots$ 表示参数块 X 的全条件分布，即在给定所有其他参数的条件下， X 的条件分布。

1. 更新 (ψ, τ, ϕ)。按照 Bhattacharya 等人（2015年），抽取样本

$$\psi_{jh} \,|\, \lambda_{jh}, \phi_{jh}, \tau \overset{\text{ind}}{\sim} \text{giG}\left(\frac{1}{2}, 1, \frac{\lambda_{jh}^2}{\phi_{jh}^2 \tau^2}\right)$$

$$\tau \,|\, \phi, \Lambda \sim \text{giG}\left(p \cdot d \cdot (a-1), 1, 2 \sum_{\substack{j=1:p \\ h=1:d}} \frac{|\lambda_{jh}|}{\phi_{jh}}\right)$$

$$\phi_{jh} = \frac{T_{jh}}{T}, \quad T_{jh} \overset{\text{ind}}{\sim} \text{giG}(\, a-1, 1, 2|\lambda_{jh}|\,), \quad T := \sum_{j,h} T_{jh},$$

对于 $j = 1, \ldots, p$, $h = 1, \ldots, d$，其中 giG 表示广义逆高斯分布（见 (S3.26)）。

2. 更新 Λ。从全条件密度中抽取样本

$$p(\Lambda \mid \cdots) \propto p(\boldsymbol{y} \mid \Lambda, \boldsymbol{\eta}, \Sigma) p(\Lambda \mid \phi, \tau, \psi) f_{\mathrm{DPP}}^{\mathrm{app}}(\boldsymbol{\mu} \mid \rho, \Lambda, K_0; R)$$

使用 Metropolis-Hastings步骤；更多细节请参见第3.3节。

3. 的更新。从以下分布中独立抽取每个 σ_j^2

$$\sigma_j^2 \mid y^{(j)}, \boldsymbol{\eta}, \lambda^{(j)} \overset{\text{ind}}{\sim} \text{inv} - \text{Gamma}\left(\frac{n}{2} + a_\sigma, \frac{1}{2}\sum_{i=1}^{n}\left(y_i^{(j)} - {\lambda^{(j)}}^T \eta_i\right)^2 + b_\sigma\right)$$

其中， ${\lambda^{(j)}}^\top$ 是 Λ 的第 j行，而 $y^{(j)} = (y_{1j}、\ \ldots\ldots、\ y_{nj})^\top$。

4. 未分配变量的更新($\boldsymbol{\mu}^{(na)}$、 $\boldsymbol{s}^{(na)}$、 $\boldsymbol{\Delta}^{(na)}$)。按照 Beraha 等（2022）的做法，我们将未分配变量的联合全条件分布分解为

$$p(\boldsymbol{\mu}^{(na)}, \boldsymbol{s}^{(na)}, \boldsymbol{\Delta}^{(na)} \mid \text{rest}) = p(\boldsymbol{\mu}^{(na)} \mid \text{rest}) p(\boldsymbol{s}^{(na)} \mid \boldsymbol{\mu}^{(na)}, \text{rest}) p(\boldsymbol{\Delta}^{(na)} \mid \boldsymbol{\mu}^{(na)}, \text{rest}),$$

其中，“rest”表示除($\boldsymbol{\mu}^{(na)}$、 $\boldsymbol{s}^{(na)}$、 $\boldsymbol{\Delta}^{(na)}$)以外的所有变量。于是， $\boldsymbol{\mu}^{(na)} \mid$ rest是一个密度为以下形式的Gibbs点过程

$$p(\{\mu_1^{(na)}, \ldots, \mu_\ell^{(na)}\} \mid \text{rest}) \propto f_{\mathrm{DPP}}^{\mathrm{app}}(\{\mu_1^{(na)}, \ldots, \mu_\ell^{(na)}\} \cup \boldsymbol{\mu}^{(a)} \mid \rho, \Lambda, K_0; R) \psi(u)^\ell$$

其中 $\psi(u) = \mathbb{E}[\mathrm{e}^{-uS}]$。我们采用 Geyer 与 Møller（1994）中的生灭 Metropolis-Hastings 算法，从该点过程密度中抽取样本。给定 $\boldsymbol{\mu}^{(na)}$ 后，可以直接证明

$$\begin{aligned}
\Delta_1^{(na)}, \ldots, \Delta_\ell^{(na)} \mid \cdots &\overset{\text{iid}}{\sim} \mathrm{IW}_d(\nu_0, \Psi_0) \\
S_1^{(na)}, \ldots, S_\ell^{(na)} \mid \cdots &\overset{\text{iid}}{\sim} \text{Gamma}(\alpha, 1+u)
\end{aligned}$$

5. 已分配变量的更新($\boldsymbol{\mu}^{(a)}$， $\boldsymbol{s}^{(a)}$， $\boldsymbol{\Delta}^{(a)}$)。令 k 表示分配向量 $\boldsymbol{c}$ 中唯一值的数量，并假设活跃组分为前k个。我们可以使用 Gibbs扫描对已分配变量进行采样。显然,

$$\begin{aligned}
S_h^{(a)} \mid \cdots &\overset{\text{ind}}{\sim} \text{Gamma}(\alpha + n_h, 1+u) \\
\Delta_h^{(a)} \mid \cdots &\overset{\text{ind}}{\sim} \mathrm{IW}_d\left(\nu_0 + n_h, \Psi_0 + \sum_{i:c_i=h} (\eta_i - \mu_h^{(a)})(\eta_i - \mu_h^{(a)})^\top\right).
\end{aligned}$$

$\boldsymbol{\mu}^{(a)}$ 的完整条件分布正比于

$$p(\boldsymbol{\mu}^{(a)} \mid \cdots) \propto f_{\mathrm{DPP}}^{\mathrm{app}}(\boldsymbol{\mu}^{(a)} \cup \boldsymbol{\mu}^{(na)} \mid \rho, \Lambda, K_0; R) \prod_{h=1}^{k} \prod_{i:c_i=h} \mathcal{N}_d(\eta_i \mid \mu_h^{(a)}, \Delta_h^{(a)})$$

并使用 Metropolis-Hastings步骤从中采样。

6. 潜在分配变量的更新 $\boldsymbol{c}$。我们发现，对 η_i 进行边缘化有助于改善混合链。因此，我们可以从定义在 $\{1,\ldots,k+\ell\}$ 上、权重为 ω_{ih} 的离散分布中独立采样每个 c_i：

$$\begin{aligned}\omega_{ih} &\propto S_h^{(a)}\mathcal{N}_p(y_i \mid \Lambda\mu_h^{(a)}, \Sigma + \Lambda\Delta_h^{(a)}\Lambda^\top), && h = 1,\ldots,k\\ \omega_{ik+h} &\propto S_h^{(na)}\mathcal{N}_p(y_i \mid \Lambda\mu_h^{(na)}, \Sigma + \Lambda\Delta_h^{(na)}\Lambda^\top), && h = 1,\ldots,\ell\end{aligned}$$

如果不加注意，对 p 维高斯密度的每次评估将需要 $\mathcal{O}(p^3)$ 次运算。然而，我们利用了协方差矩阵的特殊结构。根据 Woodbury 矩阵恒等式，我们有

$$\left(\Sigma + \Lambda\Delta\Lambda^\top\right)^{-1} = \Sigma^{-1} - \Sigma^{-1}\Lambda\left(\Delta^{-1} + \Lambda^\top\Sigma^{-1}\Lambda\right)^{-1}\Lambda^\top\Sigma^{-1},$$

因此，现在我们需要计算一个 $d \times d$ 矩阵的逆。因此，计算指数项中的二次型仅需 $\mathcal{O}(p)$ 的计算成本。此外，利用矩阵行列式引理，协方差矩阵的行列式可计算为

$$\det(\Sigma + \Lambda\Delta\Lambda^\top) = \det(\Delta^{-1} + \Lambda^\top\Sigma^{-1}\Lambda)\det(\Delta)\det(\Sigma).$$

通过缓存前一次矩阵求逆的运算，这可在无需额外成本的情况下计算得到。

7. 更新辅助变量 $u \sim$ 伽马分布(n, T)。

8. 更新潜在得分 $\boldsymbol{\eta}$。对于 $i = 1, \ldots, n$，从下列分布中独立抽取每个 η_i

$$\eta_i \mid \cdots \overset{\text{ind}}{\sim} \mathcal{N}_d(m_i\,,\, S_i)$$

其中

$$S_i = \left(\Lambda^T\Sigma^{-1}\Lambda + (\Delta_{c_i}^{(a)})^{-1}\right)^{-1}, \quad m_i = S_i\left(\Lambda^T\Sigma^{-1}y_i + (\Delta_{c_i}^{(a)})^{-1}\mu_{c_i}^{(a)}\right)$$

这是通过标准计算得出的。

S4.2 二元数据的吉布斯采样器

这里的吉布斯采样器需要增加一个步骤来处理二元观测值$z_{i,\ j}$。具体而言，对于原始 APPLAM 模型所述的吉布斯更新步骤（即步骤 1-8），除这里固定的参数外，均可应用，因为在给定 $\boldsymbol{y} = (y_1, \ldots, y_n)$ 的条件下，所有涉及的参数均独立于 $\boldsymbol{z} = (z_1, \ldots, z_n)$。然后，按如下方式更新 y_i:

$$y_{i,j} \mid z_{i,j}, \cdots \overset{\text{ind}}{\sim} \begin{cases}\mathcal{N}(y_{i,j} \mid (\Lambda\eta_i)_j, 1)\mathbb{1}_{[0,\infty)}(y_{i,j}), & \text{if } z_{i,j} = 1;\\ \mathcal{N}(y_{i,j} \mid (\Lambda\eta_i)_j, 1)\mathbb{1}_{(-\infty,0)}(y_{i,j}), & \text{if } z_{i,j} = 0,\end{cases}$$

对于 $i = 1,\ldots,n;\ \ j = 1,\ldots,p$，其中 $(\Lambda\eta_i)_j$ 表示 p 维向量 $\Lambda\eta_i$ 的第 j 个元素。也就是说，若 $z_{i,j} = 1$（分别地，若 $z_{i,j} = 0$），则 $y_{i,j}$ 的全条件分布服从支撑集为 $[0,\infty)$（分别为支撑集 $(-\infty, 0)$）的截断高斯分布。

S5 通过数值模拟探索 APPLAM 模型

S5.1 DPP密度近似中的截断级别

此处我们通过数值示例说明第 3.1 节所述近似过程引入的误差大小。这种近似过程在 DPP 文献中十分常见；例如参见 Lavancier 等（2015）。近似效果显然受式（15）中级数展开的截断级别 N 控制。我们讨论这种截断对 D^{app} 和 C^{app} 评估结果的影响；请注意，近似误差取决于 $\varphi(k)$ 趋于 0 的收敛速率，其一般表达式见式（8）。我们考虑两种示例情形，第一种对应于 $d = 2$，第二种取 $d = 4$。对于下文考虑的所有 DPP，我们设定强度使得 $\rho|R| = 1$，并设定 c 使得 $\rho = \rho_{\max}/2$。对于两种情形（$d = 2$,4），我们考虑推论 2.3 中的类高斯 DPP，其中 Λ 或为固定值 I_d（各向同性情形），或为随机变量，其分布要么服从先验（13），要么服从后验分布（针对附录 S5.3 的模拟情景（$d = 2$）以及附录 S6 的模拟研究 A（$d = 4$）中所用的各向异性模型）。当 Λ 为随机变量时，对于下一组图中展示的所有量，我们报告基于从 Λ 的随机分布中抽取的 100 个样本计算出的中位数和 95% 可信区间。

图 S3（左）对应于 $d = 2$ 的情形，并显示了 $\varphi(k) = \varphi(k_1,0)$ 的取值，即 $\varphi(k)$ 沿第一个分量趋于 0 的速度。同样，图 S3（右）显示了 D^{app} 随 N 变化的取值，而图 S4 则展示了不同 N 取值下的 C^{app}。显然，$\varphi(k)$ 的极快衰减意味着，只要 $N \geq 3$，N 的选择对 D^{app} 和 C^{app} 的近似影响很小。

我们对 $d = 4$、$\Lambda = I_4$ 和 Λ 为随机变量的情形重复同样的分析，这些变量服从先验分布，或服从附录 S6 的模拟研究 A 所考察的一个示例中通过马尔可夫链蒙特卡洛得到的后验分布。参见图 S5 和图 S6，它们清楚地表明，N 的选择对 D^{app} 和 C^{app} 的近似影响很小。

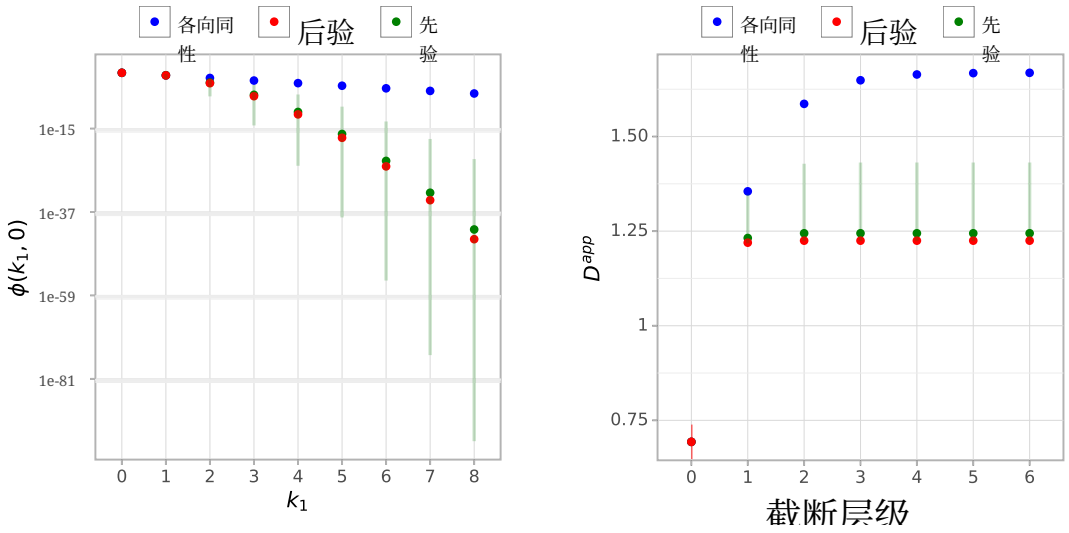

图 S3：情形 $d = 2$。随着 k_1 增大，$\varphi(k_1,0)$ 的取值（左）；D^{app} 随截断级别 N 变化的取值（右）。当 Λ 为随机变量（取自先验或后验分布）时，图中报告了所研究量的中位数和 95% 可信区间（绿色），这些结果由 Λ 的 100 个样本得到。

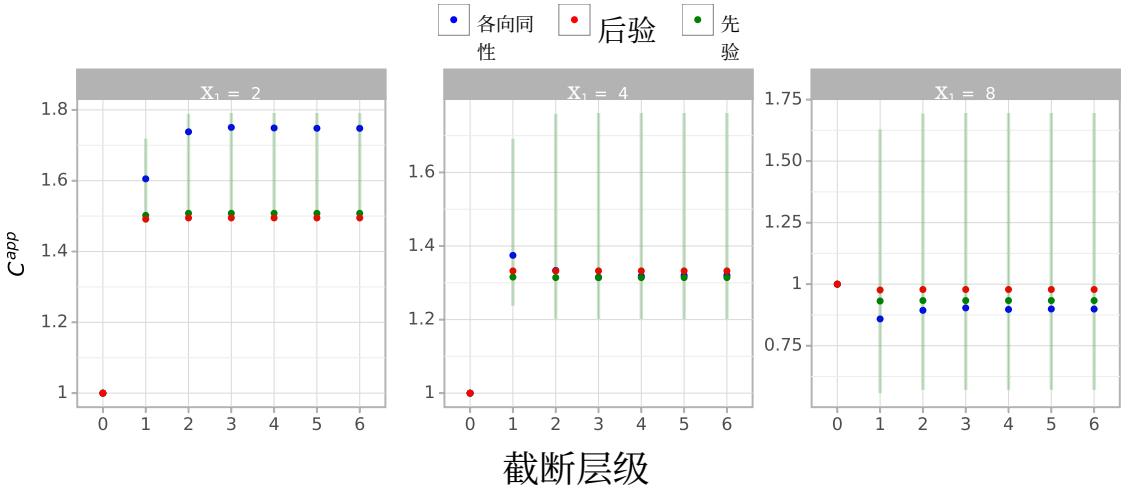

图 S4：情形 $d = 2$。$C^{app}((x_1, 0), (0,0))$ 随截断级别 N 变化的取值，以及随着 x_1 增大时的取值。当 Λ 为随机变量（取自先验或后验分布）时，图中报告了所研究量的中位数和 95% 可信区间（绿色），这些结果由 Λ 的 100 个样本得到。

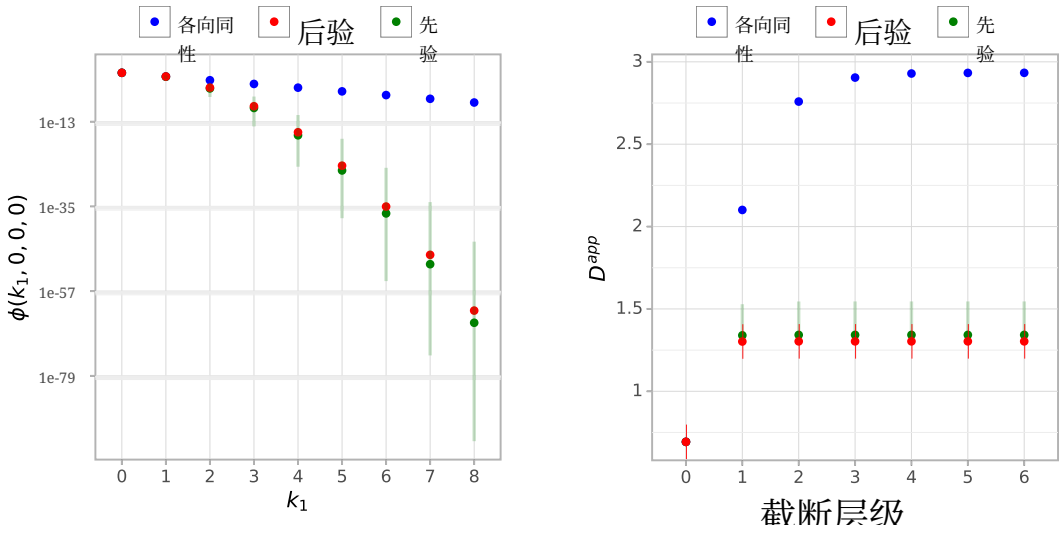

图 S5：情形 $d = 4$。随着 k_1 取值增大时，$\varphi(k_1,0,0,0)$ 的取值（左）；D^{app} 关于截断级别 N 的函数值（右）。当 Λ 为随机变量（取自先验或后验分布）时，图中报告了所讨论量的中位数和 95% 可信区间（绿色），这些结果基于从 Λ 中抽取的 100 个样本。

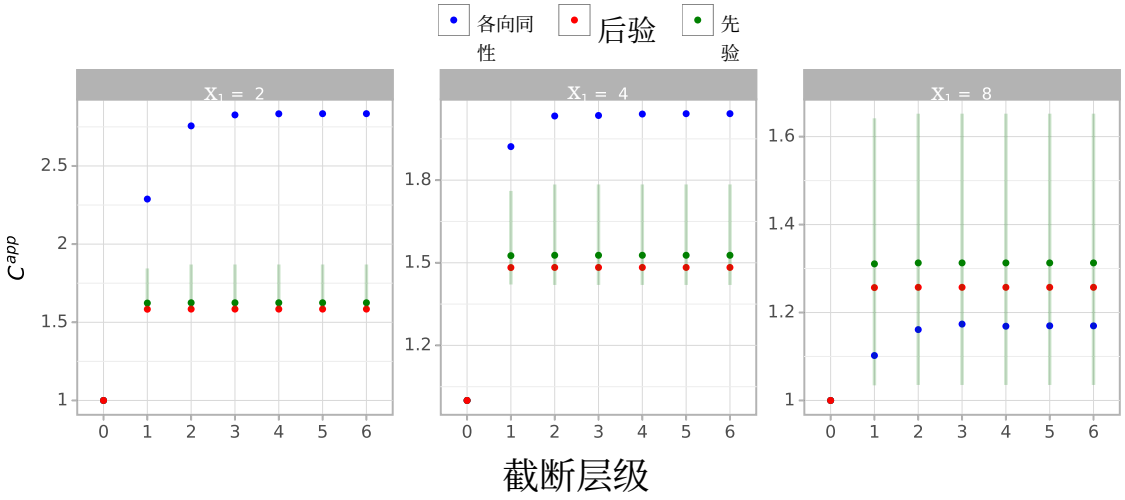

图 S6：情形 $d = 4$。对于不同的截断级别 N，随着 x_1 取值增大时，$C^{app}((x_1,0,0,0),(0,0,0,0))$ 的取值。当 Λ 为随机变量（取自先验或后验分布）时，图中报告了所讨论量的中位数和 95% 可信区间（绿色），这些结果基于从 Λ 中抽取的 100 个样本。

S5.2 解析方法与自动微分

图 S7 比较了使用自动微分方法和解析方法对高维载荷矩阵 Λ 进行抽样时的内存使用量（以字节为单位）和每次迭代的执行时间（以秒为单位）。固定模型的其他所有参数后，我们设定 $N = 4$ 和 6 个成分中心 μ_h；另见 (15)。请注意，N 的取值会影响内存使用量和执行时间。我们仅比较对矩阵 Λ 进行抽样的性能。我们使用从一个 p 维高斯分布中模拟得到的 100 个数据点，其中 $p = 100,\ 200$。

S5.3 各向异性过程与各向同性过程的比较

这里我们着重说明，在潜因子模型中进行聚类时，为什么 (14) 中的各向异性 DPP 比各向同性 DPP 更可取。我们比较了基于推论 2.3 中类高斯 DPP 的 APPLAM 模型及其各向同性改型；在后者中，(14) 中的各向异性点过程被替换为下文定义的各向同性高斯 DPP

$$K_0(x) = \rho \exp\left(-\frac{||x||^2}{2\, c^{-\frac{2}{d}}}\right), \qquad x \in \mathbb{R}^d.$$

对于这样的各向同性高斯 DPP，K_0 的傅里叶变换 $\varphi(x)$等于

$$\varphi(x) = \rho\, \frac{(2\pi)^{d/2}}{c} \exp\left(-2\pi^2 c^{-\frac{2}{d}} ||x||^2\right), \qquad x \in \mathbb{R}^d$$

对于 $\rho < \rho_{\max}$，其中 $\rho_{\max} = c(2\pi)^{-d/2}$。这是 Lavancier 等（2015）方程（3.8）中的高斯（或平方指数）点过程，其中 $\alpha = \sqrt{2}c^{-1/d}$。

数据生成。我们考虑 $p = 100$和 $d = 2$。我们分别从 $m = 2$ 个高斯密度中各模拟 500 个独立同分布的 $\eta_i \in \mathbb{R}^2$，这些高斯密度的均值为 $\mu_1 = (-2,0)$ 和 $\mu_2 = (2,0)$，协方差矩阵均等于 $0.5I_2$。然后我们定义 $\Lambda = V\Lambda'$，其中

$$\Lambda' = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \tag{S5.31}$$

其中 $\lambda_1 = 3$，$\lambda_2 = 1/3$，且 V 的列构成 $\mathbb{R}^p$中一个 d维线性子空间的正交归一基。最后，我们模拟

$$y_i \,|\, \eta_i, \Lambda, \Sigma \overset{\text{ind}}{\sim} \mathcal{N}_p(y_i \,|\, \Lambda\eta_i,\, \Sigma) \qquad i = 1, \ldots, n$$

其中，$n = 1000$和 $\Sigma = 10I_p$。

模型超参数。对于两个模型（APPLAM 和各向同性 模型），我们将 (2) 中的 Σ 固定为 I_p，并将每个 $\Delta_h = I_d$ 固定。对于两个行列式点过程，我们将 $R = [-10, 10]^2$ 固定。为了使两个先验相匹配，我们将 (ρ, c) 的值设定如下。我们令两个过程具有相同的强度，即 $\rho|R| = 0.1$；这意味着混合分布的成分数量具有相同的先验均值。我们为 c 设定两个不同的值，以保证沿潜在空间的 x 轴，对相关函数 (10) 相同。作出这一选择时

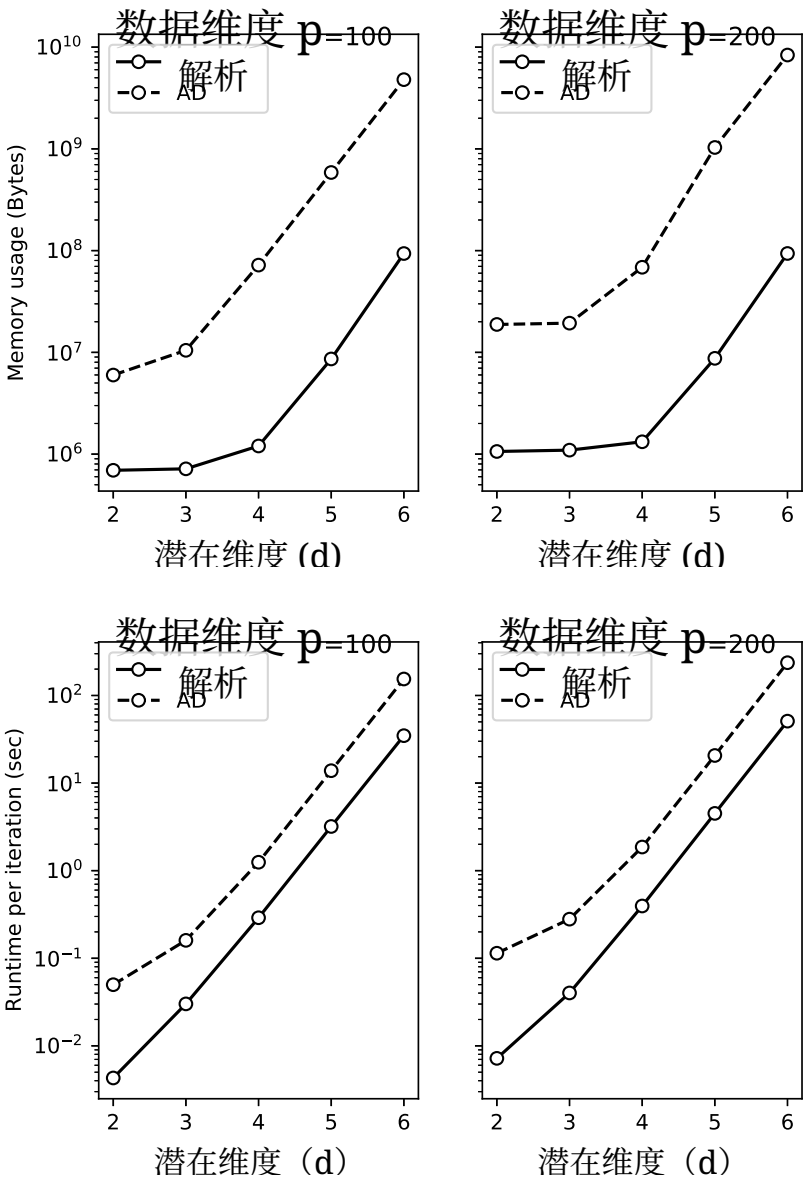

图 S7：当数据维度为 $p = 100$ （左图）和 $p = 200$ （右图）时，随着潜在维度 d 的变化，使用 $n = 100$ 个样本进行马尔可夫链蒙特卡洛的内存需求（上排）和每次迭代的运行时间。

充分了解了数据生成过程，因为潜在簇沿 x 轴分离（见数值 μ_1 和 μ_2），以使两个先验尽可能匹配。这要求 $c_{iso} = c_{aniso} \cdot \lambda_1/\lambda_2$，其中 λ_1 和 λ_2 在下文定义（S5.31）。接着，我们设定 c_{aniso}，使得 $\rho = 0.9\rho_{max}$。遵循 Chandra 等人（2023）的做法，我们将 (13) 中的 $a = 0.5$ 设定为。在两个排斥混合模型中，对于权重的先验分布 $\boldsymbol{w}$，我们假设 (16) 中的 $\alpha = 10^{-3}$。注意，α 较小（如本例所示）意味着 (11) 中 $\boldsymbol{w}$ 的边际先验大多集中于 $\mathbb{R}^{m-1}$ 中单纯形的边界上。

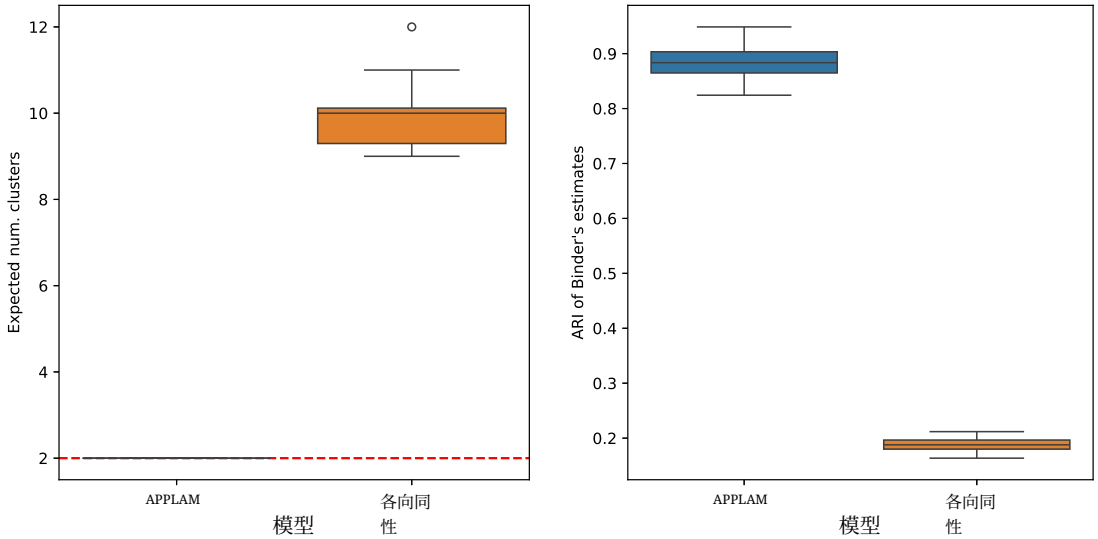

图 S8：簇数后验均值的箱线图（左）；红色虚线表示真实簇数。基于 Binder 损失函数得到的估计簇与真实簇之间 ARI 的箱线图。每个箱线图对应 100 个重复数据集的结果。

估计簇的比较。为了比较两个模型在估计簇方面的表现，我们报告了对 100 个重复数据集进行的推断结果；这些数据集是按照前文所述的方法独立生成的。具体而言，对于每个数据集，我们反复抽取潜在因子 η_i、正交归一矩阵 V，然后生成数据点。图 S8（左）展示了簇数后验均值的箱线图（基于 100 个重复数据集）。图 S8（右）展示了聚类估计与数据集真实划分之间调整兰德指数（ARI）取值的箱线图。聚类估计是通过最小化 Binder 损失的后验期望得到的（Binder, 1978年）。一个数据集的两个划分之间的调整兰德指数（Hubert 和 Arabie, 1985年）可以取负值，其上界为 1，对应两者完全一致的情形。显然，我们的 APPLAM 模型能够在全部 100 个重复数据集中准确识别数据的真实簇（ARI 值接近 1），而各向同性改型则无法做到这一点。

S5.4 超参数选择的实证分析

这里我们介绍一种依赖于数据的方法，用于为推论 2.3 中超参数 ρ 和 c 的取值选择合适的范围；入门性讨论见第2.3节。对 $(\rho,\ c)$ 的这些依赖于数据的估计值，应理解为提示我们所设定的 ρ 和 c 取值的数量级，而非超参数的精确取值。具体而言，我们将 Beraha 等（2022）第 6.1 节中的论证调整到我们的设定中。该文作者提出，通过控制 (i) 过程的排斥范围 以及 (ii) 先验所诱导的排斥性强度，来确定 Strauss 过程先验的超参数。第 (i) 点基于以下假设：排斥性

应当被诱导出来以消除冗余簇，但不应严重影响密度估计。第 (ii) 点则源于这样的论点：排斥性强度不应被视为一个绝对值，而应视为样本量的函数。更多详情请参见 Beraha 等（2022）。

在我们的设定中，我们利用上述（i）为推论 2.3 中的参数$\rho_{\max}$ 选取合适的范围。由于 $\rho_{\max} = c(2\pi)^{-d/2}$，这等价于为 c 选取合适的范围。直接使用定理 2.2 中定义的过程 $\Phi = \{\Lambda\mu\colon\ \mu \in \Phi\}$ 最为简便。特别地，$\tilde{\Phi}$ 定义在 $\mathbb{R}^d$（的适当子集）上，且具有各向同性。主要思路是先选取一个合适的距离 $\hat{r}$，使得我们预期从 Φ 中抽取的任意两个簇中心，即 $\Lambda\mu_h$ 和 $\Lambda\mu_k$，满足 $\|\Lambda\mu_h - \Lambda\mu_k\| > \hat{r}$。为此，我们沿用 Beraha 等（2022）的论证，计算所有数据点 y_i 和 y_j之间的成对距离 $d_{i,\ j}$。若样本量较大，也可以考虑对数据进行子抽样。然后，我们对这些 $d_{i,\ j}$，计算核密度估计，并将该密度的最小局部最小值选为 $\hat{r}$。直观地说，我们将从 0 到 $\hat{r}$ 的核密度解释为同一簇内数据点之间距离的密度。按此方式选择 $\hat{r}$，可以保证我们不会强制分离不同簇的位置。图 S9 展示了示意图。

现在令 $g(r)$ 为 Φ 的对相关函数。由于该过程具有各向同性，因此 $g \equiv g(r)$，其中 $r \in [0, +\infty)$，取值于 $[0, 1)$。固定这样一个值 $\hat{r}$ 后，我们旨在选择 c，使得 $g(\hat{r})$ 取较大的值，例如 0。99。根据 Lavancier 等（2015），$\hat{r}$ 可解释为“排斥范围”，即距离大于 $\hat{r}$ 的两点实际上相互独立。$g(r)$ 的表达式为

$$g(r) = 1 - \exp\left(-\frac{r^2}{2|\Lambda^T\Lambda|^{1/d}c^{-2/d}}\right),$$

它取决于随机变量 Λ。一旦我们估计出上述表达式中的 Λ，便可通过求解 $g(\hat{r}) = 0.99$ 得到 $\hat{c}$。我们可以假设 Λ 的先验即插即用估计值，例如其先验均值。然而，如果后验分布与先验相差甚远，这种做法就会失效。因此，我们采用以下思路。首先注意，项 $|\Lambda^T\Lambda|^{1/d}$ 表示超正方形 R（行列式点过程 Φ 在此定义）与集合 $R = \Lambda R$（Φ 在此定义）之间的体积变化。为了粗略估计 $|R|$，我们再次考察数据的成对距离，并取其最大值，记为 $r_{\max}$。我们近似取 $|\tilde{R}| \approx (r_{\max})^d$。这与 Bianchini 等（2020）和 Beraha 等（2022）提出的方法类似，他们将行列式点过程定义在包含所有数据的最小超正方形上。求解 $|\tilde{R}| = |\Lambda^T\Lambda|^{1/d}|R|$，并根据 $R = [-10, 10]^d$，可得 $|\Lambda^T\Lambda|^{1/d} \approx (r_{\max}/20)^d$。因此，我们固定

$$\hat{c} = (\log 10)^{d/2}\left(\frac{r_{\max}}{10\hat{r}}\right)^d,$$

并且，根据推论 2.3，$\hat{\rho}_{\max} = \hat{c}(2\pi)^{-d/2}$。

将这一论证应用于附录 S6 中稍后分析的数据集，可得出 $\hat{\rho}_{\max} \times |R|$ 介于 0.5 和 20.3 之间的范围，这表明我们应当重点关注强度（非常）小的 DPP。

在得到$\hat{\rho}_{\max}$的估计值后，我们转而讨论上述第（ii）点，即排斥性的强度。由于 DPP 在 $\rho \le \rho_{\max}$ 时存在（且当不等式严格成立时具有密度），

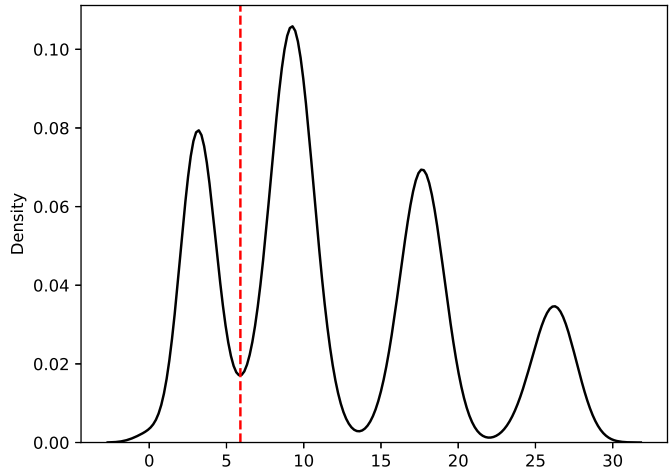

图 S9：从数据估计 $\hat{r}$ 的示意图。黑线表示附录 S5.1 中分析的某个模拟数据集里数据点之间成对距离的核密度估计。红色虚线表示 $\hat{r}$ 的值。

因此可以像 Lavancier 等（2015）、Bianchini 等（2020）和 Beraha 等（2022）那样，直接将 $\rho = s\rho_{\max}, s \in (0, 1)$设定为参数。如此一来，参数 s 便可解释为 DPP 的排斥性强度。我们认为，s 应是样本量的增函数。事实上，采用排斥先验的意义在于：如果某个簇的基数较小，那么排斥性应在该簇中占据主导，而不是似然，因此小簇出现在大簇附近的概率应当较低。然而，“大”和“小”簇的定义取决于样本量；例如，在一个包含 $n = 30$ 个个体的数据集中，大小为 10 的簇算大，但在一个包含 $n = 1000$的数据集中则非常小。

在下一小节中，我们将开展一项模拟研究，以评估模型对其超参数（即 ρ 和 s）的稳健性。我们的模拟证实了此处获得的见解，即强度应当相当小。具体而言，我们提出以下建议。首先，在固定 $R = [-10, 10]^d$ 后，选择一个 ρ 的值，使得 $\rho|R|$ 较小，即小于 5，或可能小于 1。在固定 ρ 后，只需选择排斥性强度 s 的值，该值应随样本量增加而增大。可以针对 $s \in \{0.5, 0.75, 0.9\}$ 的不同取值开展稳健性分析，或根据用户偏好固定 s。

S5.5 对模型超参数的稳健性

这里展示 APPLAM 模型对其超参数的稳健性。为此，我们采用附录 S6.1 中模拟研究 A 使用的合成数据生成机制。具体而言，我们将潜在维度设为 $d = 4$，数据维度设为 $p = 500$。我们从每个均值为 $\mu_1 = (7.5, \ldots, 7.5) \in \mathbb{R}^d$, $\mu_2 = (2.5, \ldots, 2.5)$,$\mu_2 = (-2.5, \ldots, -2.5)$ 和 $\mu_4 = (-7.5, \ldots, -7.5)$、协方差矩阵为单位矩阵的 $m = 4$ 个高斯核中，模拟生成 $n_0 \in \{50, 100, 250\}$ 个独立同分布潜在因子 $\eta_i \in \mathbb{R}^d$。随后，我们独立生成 $n = n_0 \times 4 \in \{200, 400, 1000\}$ p 维数据

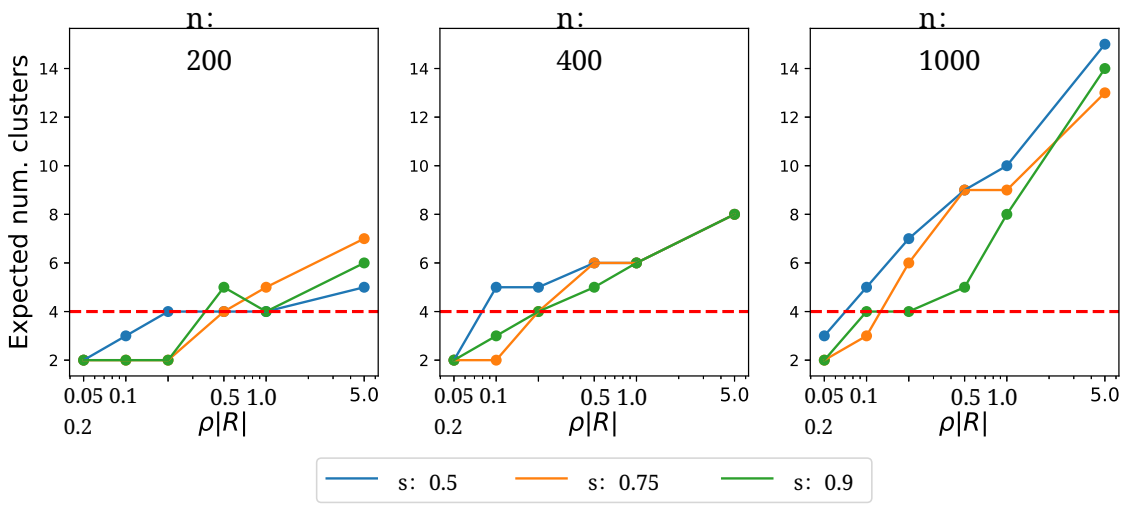

图 S10：不同超参数和样本量下簇数的后验均值。

这些数据来自多元 Student t 分布 $t_p(\Lambda\eta_i, 10I_4, 3)$（多元 Student t 分布的参数化方式请参见附录 S6.1），其中 Λ 从一组矩阵构成的集合中抽取（每个数据集抽取一次）；该集合中的矩阵列构成 d 维线性子空间 $\mathbb{R}^p$的正交归一基。

我们对先验超参数进行细网格探索。具体而言，我们考虑 $\rho|R| \in \{0.05, 0.1, 0.2, 0.5, 1, 5\}$，其中 c满足 $\rho = s\rho_{max}$（见推论 2.3），且 $s \in \{0.5, 0.75, 0.9\}$。请记住，对于固定的 ρ，s 表示排斥强度。我们将式 (16) 中的 $\alpha = 10^{-3}$ 设为（$\boldsymbol{w}$ 的边际先验是稀疏的），将式 (12) 中的 $a_\sigma = 1, b_\sigma = 0.3$ 和式 (13) 中的 $a = 0.5$ 设为 Chandra 等人（2023）建议的值。每次拟合 APPLAM 模型时，我们运行 2500 次预烧迭代，并运行 10^4 次迭代，抽稀间隔为 2，最终样本量为 5000。图 S10 展示了在不同超参数和样本量取值下簇数的后验均值。图 S11 则显示了数据集的聚类估计与真实划分之间的调整兰德指数（ARI）。聚类估计是通过最小化 Binder 损失（Binder, 1978）的后验期望得到的。可以看到，除超参数取值极端的情况外，后验推断具有稳健性；不过，当 $\rho|R|$ 介于 0.1 和 0.5 之间时，估计簇与真实簇的一致性更高（ARI 值更大）。图 S10 还指出，当 $\rho|R|$ 较大且 s 较小时，估计簇数会增加。相反，较大的 s 和较小的 $\rho|R|$ 会带来更强的排斥性，从而使估计簇数减少。这与我们根据先验特性所作的预期完全一致，详见附录 S6.2。

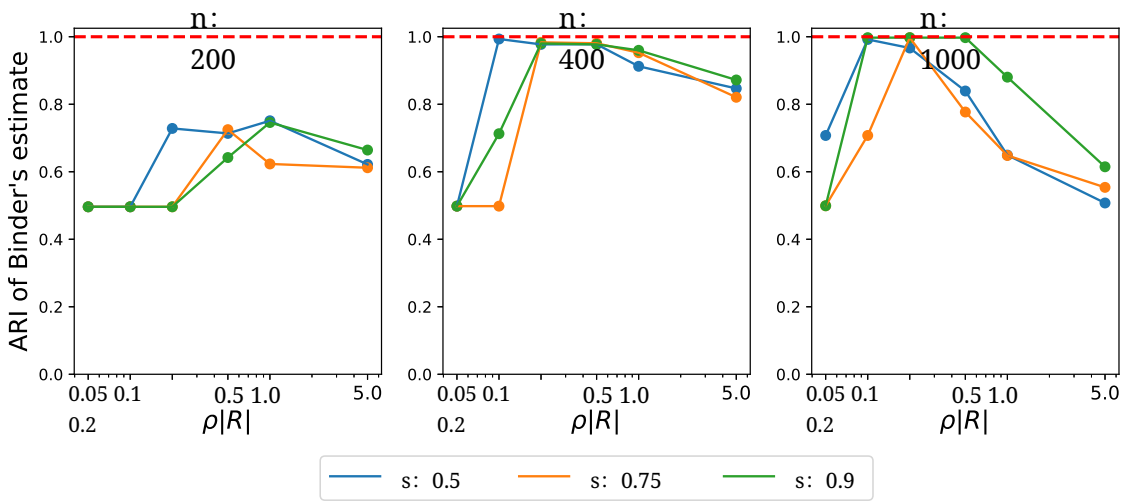

图 S11：在不同超参数和样本量下，Binder 损失估计簇的调整兰德指数（ARI）。

S6 APPLAM 与 LAMB 在模拟数据集上的比较

S6.1 合成数据生成

我们展示了基于推论 2.3 中类高斯 DPP 的 APPLAM 模型与 Chandra 等人（2023）的 Lamb 模型之间的比较。该比较假设似然函数为 (2)，并采用总质量为 α_{DP} 的狄利克雷过程混合模型，其基测度为正态-逆威沙特分布。我们考虑两组不同的模拟数据，分别称为模拟研究 A 和 B。这两种设置都意味着我们假设用于聚类的两个模型存在模型设定错误。令 $t_p(q, \Sigma, \nu)$ 表示具有位置参数 q、尺度矩阵 Σ 和自由度 ν 的多元 p 维 Student t 分布。

在模拟研究 A 中，我们从 $m = 4$ 个均值为 μ_h、$h = 1$、……、m 且协方差矩阵为单位矩阵的高斯核中，各模拟生成 50 个独立同分布潜在因子 $\eta_i \in \mathbb{R}^d$。接着我们模拟

$$y_i \,|\, \eta_i, \Lambda, \Sigma \overset{\text{ind}}{\sim} t_p(y_i \,|\, \Lambda\eta_i,\ \Sigma,\ 3), \qquad i = 1, \ldots, n$$

其中 $\Sigma = 10I_p$。我们固定 $\mu_1 = (7.5, \ldots, 7.5) \in \mathbb{R}^d$，$\mu_2 = (2.5, \ldots, 2.5)$，$\mu_2 = (-2.5, \ldots, -2.5)$ 和 $\mu_4 = (-7.5, \ldots, -7.5)$。因子载荷矩阵 Λ 固定为从由以下矩阵构成的集合中均匀抽取的一次结果：这些 $p \times d$ 矩阵的列构成 $\mathbb{R}^p$ 中某个 d 维线性子空间的正交归一基。对于模拟研究 B，从各个 $m = 4$ 多元 Student t 分布 $t_d(\mu_h, I_d, 3)$、$h = 1, \ldots, m$ 中分别生成 50 个独立同分布潜在因子，其中 μ_h 的取值与模拟研究 A 中相同。在两种模拟设置下，随后从似然函数 (2) 中独立生成 $n = 50 \times 4 = 200$ 个数据点，其中 $\Sigma = 0.1I_p$ 和 Λ 按照模拟研究 A 中的方式生成。对于模拟设置 A，我们考察 $p = \{500, 1000\}$ 和 $d = \{4, 8\}$；对于设置 B，则报告 $p = \{500, 1000\}$ 和 $d = 4$。对于每种模拟设置，我们抽取 100 个相互独立的

数据集及其后验估计的平均值（见附录 S6.3）。

S6.2 超参数的确定与马尔可夫链蒙特卡洛细节

APPLAM 和 LAMB 中共同超参数的设定遵循 Chandra 等人（2023）采用的默认选择。具体而言，我们在 (12) 中设定 $a_\sigma = 1, b_\sigma = 0.3$，并在 (13) 中设定$a = 0.5$。此外，LAMB 假设潜在因子服从高斯密度的狄利克雷过程位置-尺度混合分布。具体来说，采用正态-逆 Wishart 分布作为基测度，其位置为 μ_0，尺度为 k，协方差矩阵为 Ψ_0，自由度为 ν_0。默认设置将 μ_0 设为零向量，$k = 0.001, \Psi_0 = \delta\, I_d$，其中 $\delta = 20$ 和 $\nu_0 = d + 50$。相应地，对于 APPLAM，我们在 (11) 中设定 $\Psi_0 = \delta\, I_d, \delta = 20$ 和 $\nu_0 = d + 50$。我们还将 (16) 中的 $\alpha = 10^{-3}$ 设为稀疏，以便为$\boldsymbol{w}$采用稀疏边际先验。

很难让 APPLAM 和 Lamb 中簇数的先验相匹配：对于 Lamb，它由 α_{DP} 决定；而对于 APPLAM，它由成分数量 m 的先验诱导，该先验受 (ρ, c) 控制。具体而言，$\rho|R|$ 是不以 $m \geq 1$ 为条件时成分数量的先验期望值。因此，在模拟中，我们仅通过实证评估聚类估计对这些参数选择的稳健性。具体而言，我们考虑 $\alpha_{DP} = \{0.1, 0.5, 1\}$，它们分别对应于簇数的先验期望值 1。57、3。63 和 5。88。对于我们的模型，我们则遵循附录 S5.5 中的论述，通过设定 $\rho|R| \in \{0.5, 1, 5\}$ 和 $c = c_\rho = 2\rho(2\pi)^{d/2}$ 来检验不同程度的排斥性，即 $s = 0.5$。

每次拟合 APPLAM 模型时，我们都会运行 2000 次预烧迭代次数和 5000 次迭代，抽稀间隔为 2，最终样本量为 2500。对于 Lamb 模型的每次运行，我们则模拟 10^6 次预烧迭代和 5×10^4 次迭代，抽稀间隔为 10。Lamb 算法在我们的模拟中混合效果不佳（见图 S12），因此需要非常长的预烧阶段。

S6.3 两种模型的比较

这里我们重点考察两个模型的聚类估计。我们考虑簇数的期望，以及由分配变量 c_i's 诱导的受试者随机划分的一个后验汇总统计量。与之前一样，通过最小化 Binder 损失函数的后验均值（Binder，1978年）来获得聚类估计，并分别计算两个模型的估计分区与真实聚类之间的调整兰德指数（ARI）。ARI 等于 1 表示两个划分完全一致。相反，如果 ARI 等于 0，则估计分区等价于对数据进行随机标记。模拟研究 A 的图 S13-S14 以及模拟研究 B 的图 S15，展示了在 100 个数据集上，簇数的后验期望（上排）以及 Binder 损失估计的簇与真实簇之间的调整兰德指数（ARI）（下排）的箱线图。表 S1-S2-S3 展示了各独立数据集汇总后的数值，包括簇数的后验均值和后验众数，以及与之前相同的调整兰德指数值。

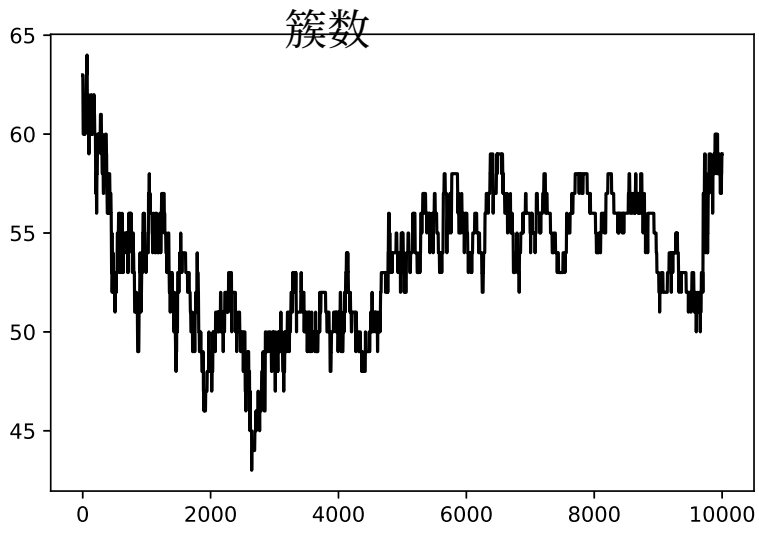

图 S12：模拟研究 B 中，Lamb 模型的马尔可夫链蒙特卡洛算法经过 10^5 次迭代后簇数的轨迹图，$p = 100$、$d = 5$、$\alpha_{DP} = 0.5$。

这些图总体表明，Lamb 模型倾向于高估真实簇数（$m = 4$）。另一方面，APPLAM 对 $\rho|R|$ 的选择表现得相当稳健：估计簇数不仅始终适中（这有利于提高可解释性），而且相应的后验估计也围绕正确的簇数分布。此外，就估计分区与真实分区之间的调整兰德指数值而言，APPLAM 总体上比 Lamb 能够提供更好的聚类估计。

p	模型	参数	簇数后验均值	簇数后验众数	ARI Binder调整兰德指数
500	APPLAM	0.5	4.02	3.90	0.69
		1.0	4.60	4.55	0.65
		5.0	6.23	6.30	0.67
	Lamb	0.1	19.36	19.05	0.41
		0.5	19.48	19.00	0.41
		1.0	19.64	18.80	0.41
1000	APPLAM	0.5	3.70	3.75	0.71
		1.0	4.02	4.05	0.71
		5.0	5.57	5.60	0.67
	Lamb	0.1	19.53	18.00	0.37
		0.5	19.58	17.85	0.37
		1.0	19.47	18.05	0.37

表S1：模拟 A，情形 $d = 4$：簇数后验均值（Mean nclus）、簇数后验众数（Mode nclus），以及基于 Binder 损失的聚类估计与真实聚类数之间的 ARI（ARI Binder's），均在 100 个模拟数据集上取平均。参数对于 APPLAM 模型指 $\rho|R|$，对于 Lamb 模型指 α_{DP}。

p	模型	参数	簇数后验均值	簇数后验众数	Binder调整兰德指数
500	APPLAM	0.5	3.21	3.20	0.64
		1.0	3.29	3.05	0.63
		5.0	3.71	3.60	0.63
	Lamb	0.1	25.28	24.15	0.45
		0.5	25.79	25.10	0.46
		1.0	24.64	22.65	0.47
1000	APPLAM	0.5	2.82	2.85	0.60
		1.0	2.92	2.90	0.60
		5.0	3.50	3.45	0.67
	Lamb	0.1	16.62	15.35	0.42
		0.5	17.49	16.05	0.39
		1.0	17.22	16.25	0.41

表 S2：模拟 A，情形 $d = 8$：簇数的后验均值（簇数后验均值）、簇数的后验众数（簇数后验众数），以及 Binder 损失聚类估计（Binder调整兰德指数）与真实聚类数之间的 ARI，均为 100 个模拟数据集的平均值。参数 对于 APPLAM 模型指 $\rho|R|$，对于 Lamb 模型指 α_{DP}。

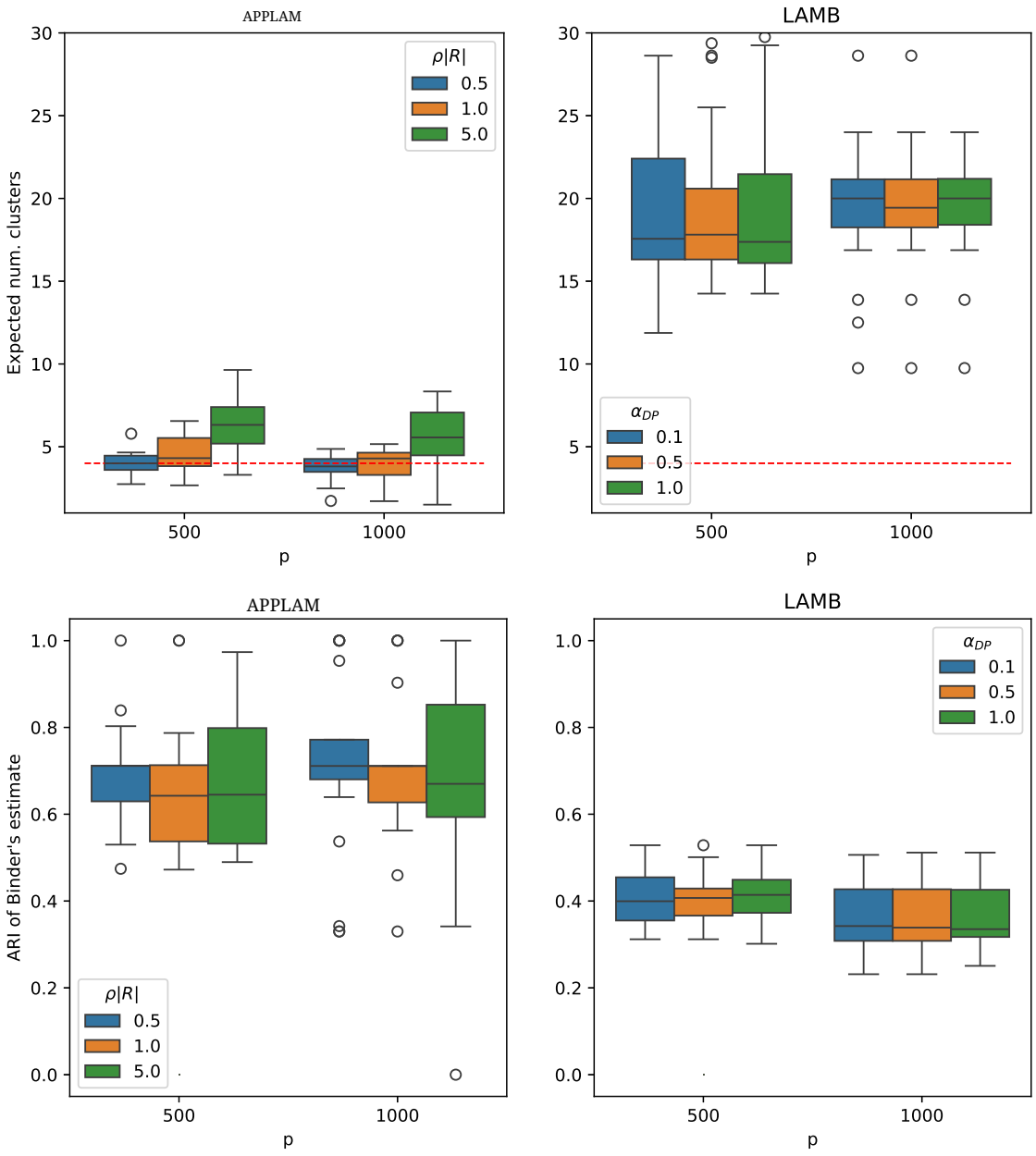

图 S13：模拟 A，情形 $d = 4$：簇数量后验期望的箱线图（上排），以及 Binder 估计的聚类结果的调整兰德指数（ARI）箱线图（下排）。每个箱线图对应 100 个重复数据集的输出。

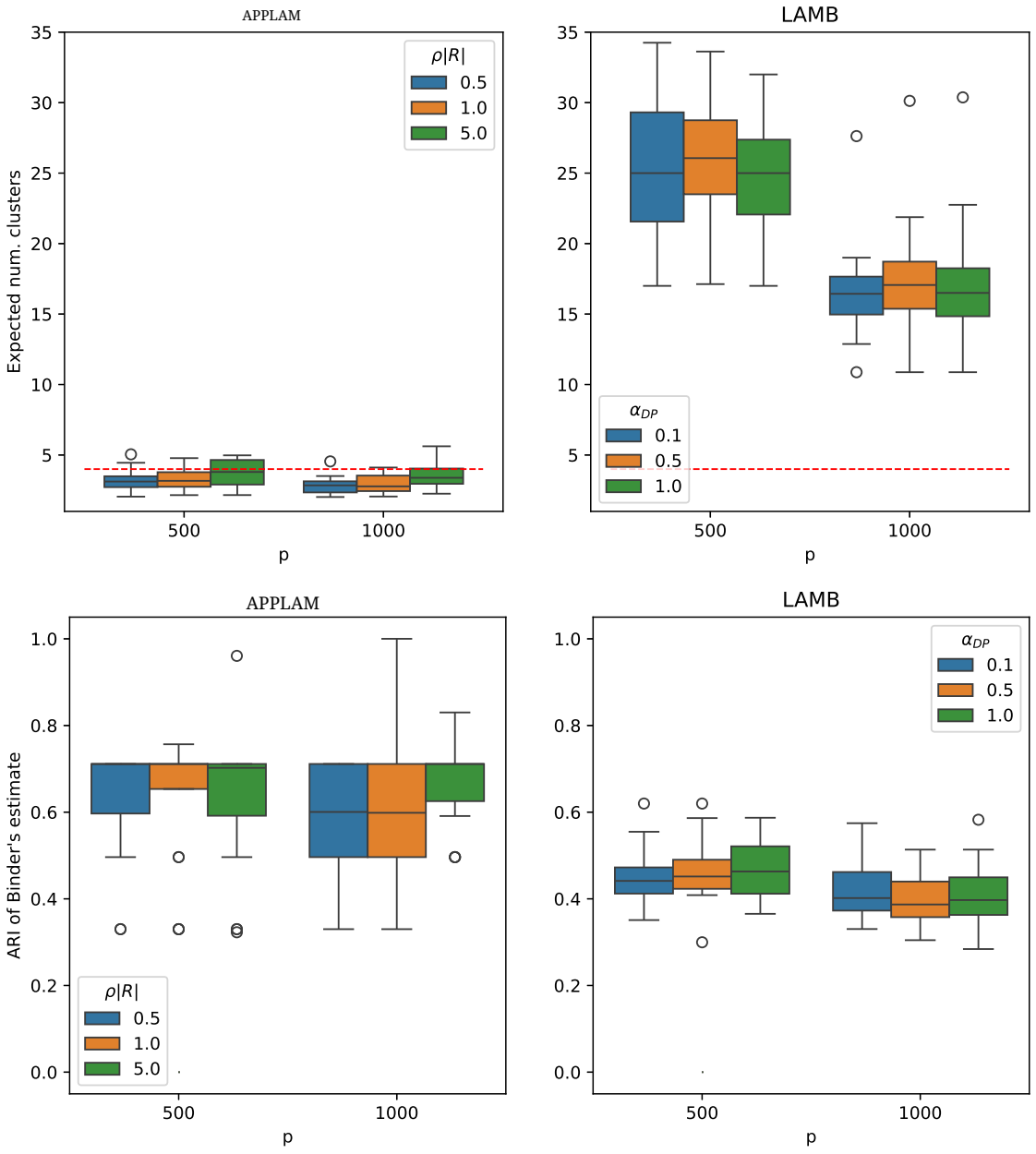

图 S14：模拟 A，情形 $d = 8$：簇数后验期望的箱线图（上排），以及 Binder 估计簇的调整兰德指数（ARI）箱线图（下排）。每个箱线图对应 100 个重复数据集的输出。

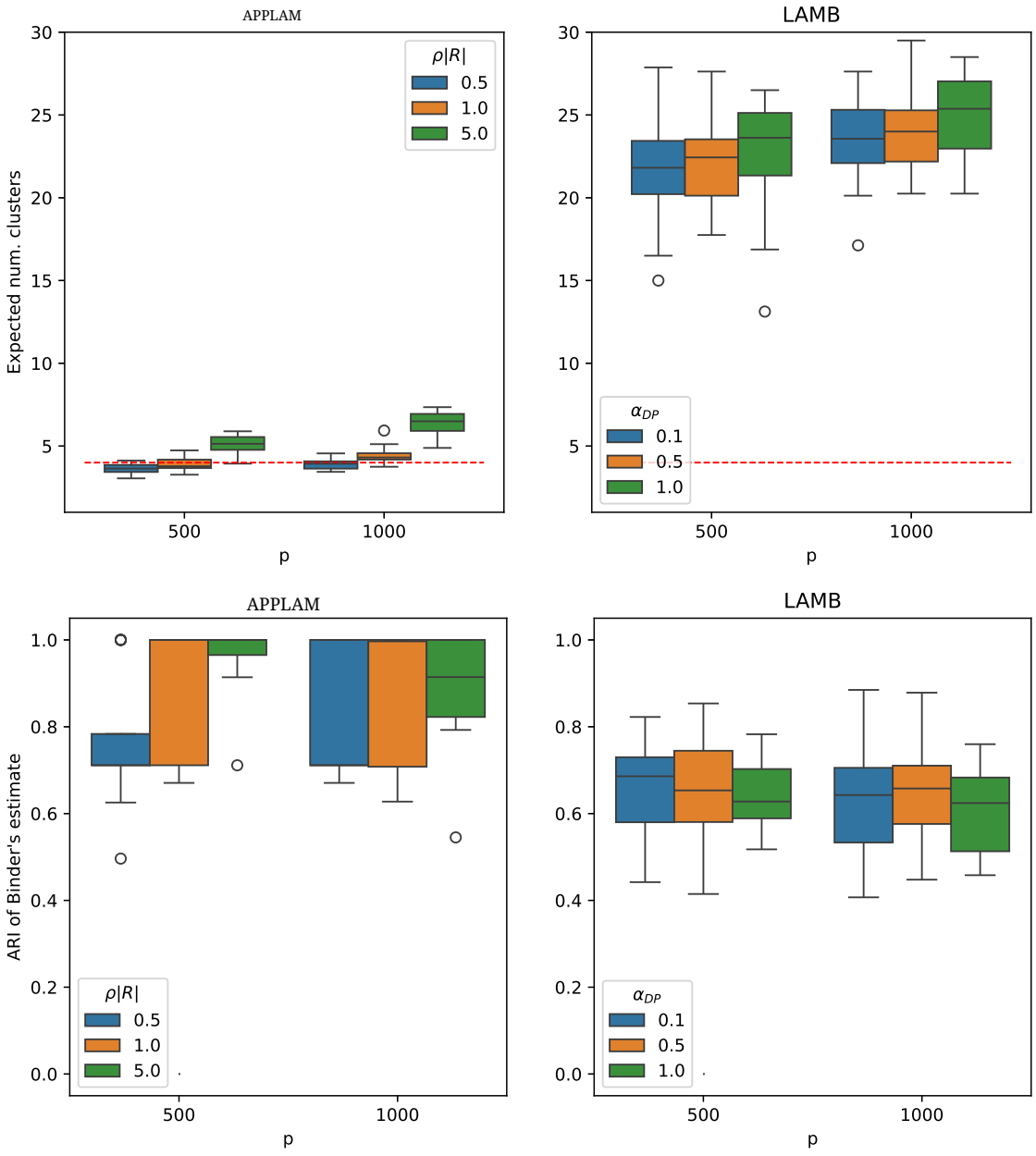

图 S15：模拟 B：簇数后验期望的箱线图（上排），以及 Binder 估计簇的调整兰德指数（ARI）箱线图（下排）。每个箱线图对应 100 个重复数据集的输出。

p	模型	参数	簇数后验均值	簇数后验众数	Binder调整兰德指数
500	APPLAM	0.5	3.61	3.50	0.77
		1.0	3.91	3.95	0.88
		5.0	5.11	5.00	0.97
	Lamb	0.1	21.71	20.10	0.66
		0.5	21.96	20.70	0.65
		1.0	22.72	21.35	0.64
1000	APPLAM	0.5	3.89	3.85	0.79
		1.0	4.44	4.35	0.88
		5.0	6.37	6.30	0.90
	Lamb	0.1	23.40	23.20	0.63
		0.5	24.02	22.45	0.65
		1.0	24.86	23.25	0.60

表S3：模拟 B：簇数后验均值（簇数后验均值）、簇数后验众数（簇数后验众数），以及 Binder 损失聚类估计（Binder调整兰德指数）与真实聚类数之间的 ARI，均基于 100 个模拟数据集取平均值。参数指 APPLAM 模型的 $\rho|R|$ 和 Lamb 模型的 α_{DP}。

S7 Bauges 数据集详情

S7.1 超参数设定

我们考察了广泛的先验超参数。具体而言，我们考虑 $d \in \{3, 4\}$、$\rho|R| \in \{0.1, 0.5, 3, 5\}$，同时固定 c，使得 $\rho = s\rho_{\max}$（见推论 2.3），其中 $s \in \{0.5, 0.75, 0.9\}$。此外，我们考虑式（16）中的 $\alpha \in \{10^{-3}, 10^{-2}, 10^{-1}\}$，以及式（12）中 Σ 的超参数的两种取值，具体为 $a_\sigma = 3$、$b_\sigma = 2$ 和 $a_\sigma = 2$、$b_\sigma = 1$。式（13）中的超参数 a 固定为 0.5，如 Chandra 等人（2023）所建议。我们在前文（特别是附录 S5.4–S5.5）中讨论过，$\rho|R|$ 应较小（例如不大于 5），而 $s \in$（0、1）表示行列式点过程的排斥性强度。超参数 α 控制潜在参数 η_i's 的混合分布的权重 $\boldsymbol{w}$ 的先验的 稀疏性； α 越小，该先验越 稀疏 。为超参数（a_σ、b_σ）所设定的值，是方差参数边际先验的标准选择：二者都对应于 σ_j^2's 的先验均值等于 1，但前者的先验方差等于 1，而后者的先验方差为无穷大。将 3 和 4 作为 d（即潜在因子 η_i's 的维度）的取值，相当于施加降维约束（此处为 $p = 123$）。由于使用 $d = 2$ 会导致 WAIC 方面明显较差的估计结果，因此我们不在此报告其后验推断。我们分别拟合所得的 144 个模型，运行马尔可夫链蒙特卡洛算法，进行 3000 次预烧迭代，并进行 5000 次迭代，抽稀间隔为 5。

S7.2 使用 APPLAM 得到的推断

图 S16 展示了我们拟合的所有模型的 WAIC（Watanabe，2013）值。显然，总体而言，采用 $d = 3$ 的模型比采用 $d = 4$ 的模型拟合效果更好（WAIC 指数更大）。表 S4 和表 S5 列出了 WAIC 的所有数值，以及通过最小化 Binder 损失的后验期望值而得到的估计簇数。

图 S17 和图 S18 分别展示了采用 $d = 3$ 和 $d = 4$ 时，簇数后验均值随 $\rho|R| \in \{0.1, 0.5, 3, 5\}$ 变化的情况。

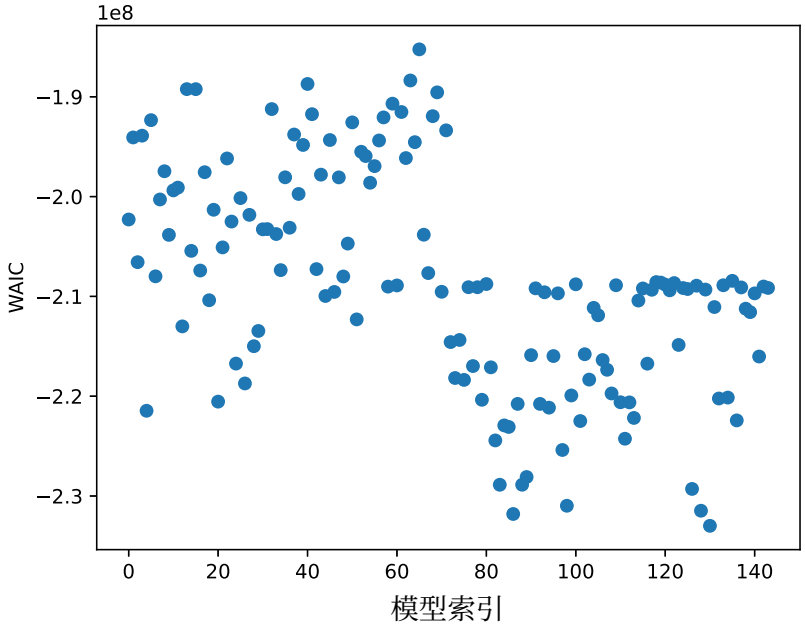

图 S16：不同拟合模型的 WAIC 值：前半部分对应包含 $d = 3$ 的模型，后半部分对应包含 $d = 4$ 的模型。

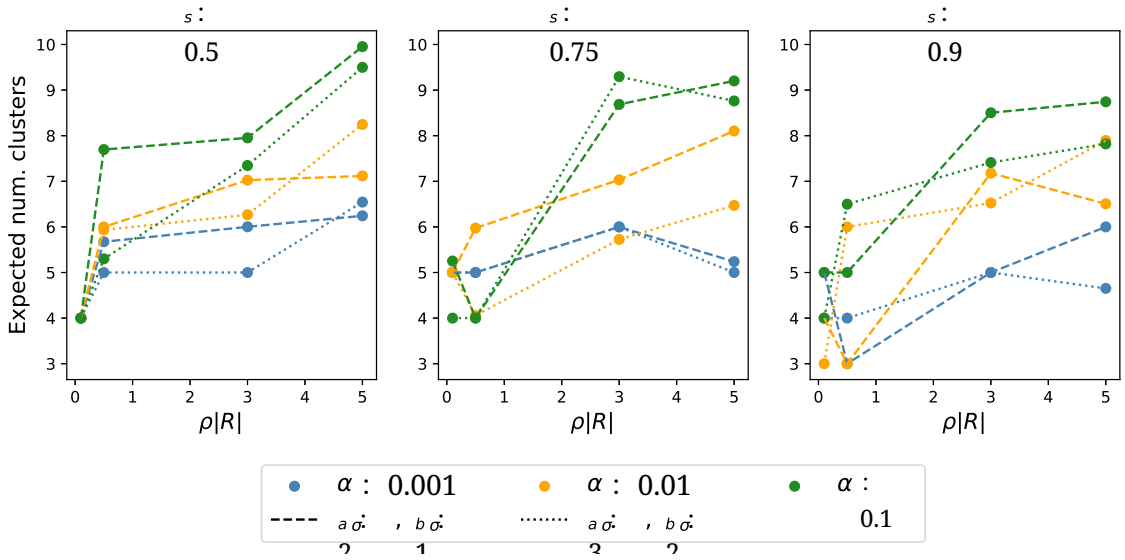

图 S17：对于不同的模型超参数选择，$d = 3$ 的簇数后验均值随 $\rho|R| \in \{0.1, 0.5, 3, 5\}$ 的变化。

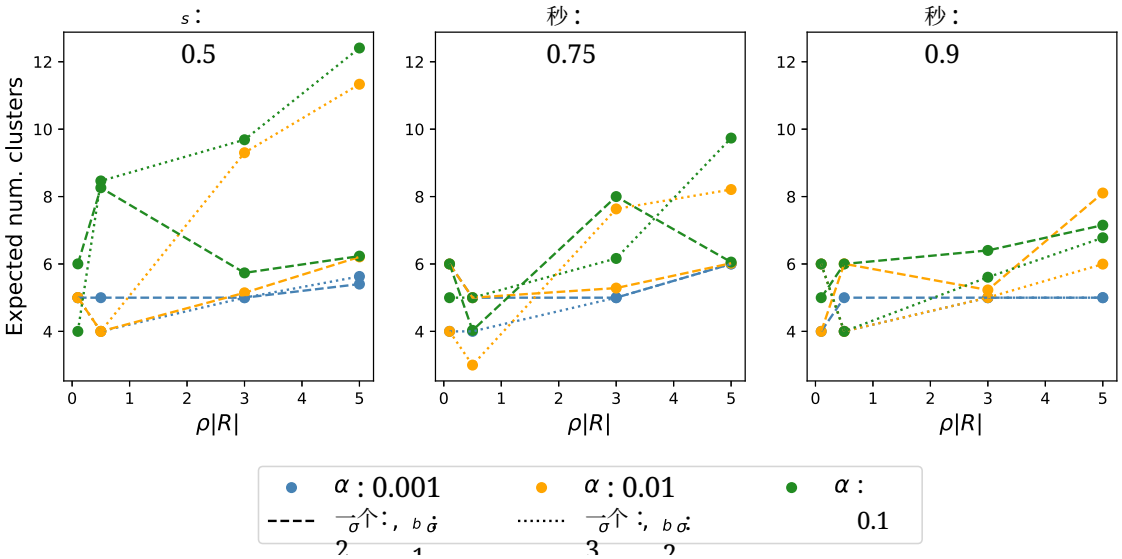

图 S18：对于不同的模型超参数选取，$d = 4$ 关于 $\rho|R| \in \{0.1, 0.5, 3, 5\}$的簇数后验均值。

Clus.	WAIC
5	-2.03e+08
6	-1.94e+08
7	-2.00e+08
8	-1.95e+08
9	-1.89e+08
10	-1.92e+08
6	-2.07e+08
6	-1.98e+08
7	-2.10e+08
8	-1.94e+08
11	-2.10e+08
11	-1.98e+08
5	-2.08e+08
5	-2.05e+08
7	-1.93e+08
8	-2.12e+08
9	-1.96e+08
10	-1.96e+08
8	-1.99e+08
8	-1.97e+08
9	-1.94e+08
8	-1.92e+08
11	-2.09e+08
12	-1.91e+08
5	-2.09e+08
6	-1.92e+08
7	-1.96e+08
9	-1.88e+08
10	-1.95e+08
11	-1.85e+08
5	-2.04e+08
6	-2.08e+08
9	-1.92e+08
8	-1.90e+08
9	-2.10e+08
11	-1.93e+08

ameters. In all cases, the latent
he posterior expectation of the

lus	WAIC
5	-2.20e+08
5	-2.09e+08
10	-2.21e+08
7	-2.24e+08
12	-2.21e+08
7	-2.22e+08
5	-2.10e+08
5	-2.09e+08
9	-2.17e+08
6	-2.09e+08
7	-2.09e+08
8	-2.09e+08
5	-2.09e+08
5	-2.09e+08
5	-2.09e+08
6	-2.15e+08
7	-2.09e+08
7	-2.09e+08
6	-2.29e+08
6	-2.09e+08
13	-2.31e+08
8	-2.09e+08
15	-2.33e+08
8	-2.11e+08
6	-2.20e+08
6	-2.09e+08
9	-2.20e+08
7	-2.08e+08
11	-2.22e+08
7	-2.09e+08
5	-2.11e+08
5	-2.12e+08
7	-2.10e+08
10	-2.16e+08
8	-2.09e+08
8	-2.09e+08

ameters. In all cases, the latent
he posterior expectation of the

S7.3 植物物种的科学名称

我们报告了图 2 中所示物种的科学名称。

- “0”：Acer pseudoplatanus L.
- “1”：Astrantia major L.
- “4”：Laserpitium latifolium L.
- “5”：Pimpinella major (L.) Hudson
- “6”：Hedera helix L.
- “9”：Cirsium palustre (L.) Scop.
- “14”：Prenanthes purpurea L.
- “27”：Helianthemum nummularium (L.) Miller
- “28”：Corylus avellana L.
- “45”：Fagus sylvatica L.
- “46”：Gentiana lutea L.
- “49”：森林老鹳草
- “61”：欧洲白蜡树
- “72”：块根毛茛
- “83”：香车叶草
- “91”：常绿薹草
- “97”：黄香草木樨
- “112”：欧洲冷杉
- “117”：田野槭
- “121”：欧洲黑莓组