arXiv:2602.06301v1 [stat.ME] 6 Feb 2026

狄利克雷过程混合的设计条件先验 elicitation：用于簇数与权重控制的统一框架

JoonHo Lee

摘要。狄利克雷过程混合模型（DPM）广泛用于教育与行为研究中的半参数贝叶斯分析，但浓度参数的设定仍是一项关键障碍。默认超先验往往对聚类施加强烈的非预期假设，而现有基于簇数的校准方法则计算效率低下，且无法控制混合权重的分布。本文提出设计条件先验 elicitation（DCE），这一统一框架可将从业者对聚类结构的信念转化为针对固定设计规模 J 的一致伽马超先验。DCE 有三项贡献。首先，它采用两阶段矩匹配（TSMM）解决计算瓶颈：结合闭式近似与精确的牛顿法细化，无需网格搜索即可校准超参数。其次，为应对“非预期先验”现象，DCE 引入双锚点协议，用于诊断并酌情约束权重主导的风险，同时透明报告由此产生的权衡及其对簇数量保真度的影响。第三，完整工作流程通过开源 DPprior R 包实现，并提供可复现的诊断工具和报告清单。模拟研究表明，无论真实簇结构如何，Gamma(1，1) 等常见默认值都会导致后验坍缩率超过 60%；而 DCE校准先验则能显著降低偏差，并在不同数据信息量水平下提高恢复效果。

关键词：狄利克雷过程混合模型；浓度参数；先验 elicitation；贝叶斯非参数方法；设计条件推断；多中心试验；超先验校准

日期：2026年2月9日。

Lee：阿拉巴马大学，jlee296@ua.edu。

本研究由美国教育部教育科学研究所通过拨付给阿拉巴马大学的 R305D240078 号资助项目提供支持。文中所表达的观点仅代表作者本人，并不代表该研究所或美国教育部的观点。

1. 引言

狄利克雷过程混合模型（DPM）已成为教育与行为科学中半参数贝叶斯分析不可或缺的工具。DPM 不再采用传统层级模型中限制性较强的参数假设（例如随机效应服从正态分布），而是让数据揭示标准方法可能掩盖的潜在结构、偏度和多峰性（Antonelli 等人，2016年；Jackson 和 White，2018年）。无论是估计多中心试验和临床环境中的治疗效应异质性（Cardoso 等人，2024年；Daniels 等人，2023年；Lee 等人，2025年）、综合荟萃分析中异质性研究的证据（Burr 和 Doss，2005；Cao 等人，2024 年；Dunson 等人，2008；Muthukumarana 和 Tiwari，2016），还是捕捉心理测量学和发展研究中的潜在变异（Bonafos 等人，2024年；Finch 和 Edwards，2016；Mignemi 和 Manolopoulou，2025年；Navarro 等人，2006年；Paganin 等人，2023年；Qiu 等人，2025年），DPM 都是稳健的默认选择，可保护推断免受模型设定错误的影响。

然而，DPM的方法灵活性伴随着一项特定代价：难以设定浓度参数 α。该参数支配模型的聚类行为——既控制在 J 个可交换单位中倾向于出现多少个不同的簇，也控制概率质量如何在这些簇之间分布。当数据信息量很强时，α 的后验分布主要由似然决定，先验设定似乎只是次要问题。但在许多应用场景中——尤其是在样本量中等且观测含噪的情形下——对 α 的先验会显著影响后验推断（Dorazio，2009年；Giordano 等人，2023年；Matza & Bistritz，2021年；Murugiah & Sweeting，2012年）。在此类情境中，看似无害的“默认值”或“模糊”超先验可能会编码关于聚类的强烈隐含假设，而这些假设可能与实质性领域知识完全相悖（Gill & Casella，2009年；Greve 等人，2022年；Iwashige & Hashimoto，2025年；Kyung 等人，2010年）。

根本的困难在于如何转译。从业者通常无法直观理解浓度参数 α 本身——问研究者“你对 α有什么先验信念？”很难得到有意义的回答。相比之下，从业者可以 常常清晰表达自己对与研究设计相关的量的看法：“在我们研究的 100 个站点中，我们预计会有多少种实质上不同的治疗效应模式？”或者“多数站点基本上具有相同效应，这种情况是否合理？还是说我们预计会存在相当大的异质性？”这些问题涉及预期的

已占据簇的数量 K_J 以及簇之间的集中程度——这些量可以直接从研究现象的实质性假设来解释（McCullagh &Yang，2008）。

从业者能够理解的量与统计软件所需的技术性超参数之间存在脱节，这造成了我们所称的转换鸿沟。要使用伽马超先验 $\alpha \sim \text{Gamma}(a, b)$来实现狄利克雷过程混合模型——这是实践中最常见的选择（Neal，2000）——研究者必须指定形状参数和率参数 (a, b)。然而，从对聚类结构的认知映射到 (a, b) 既不直接，也不直观。因此，研究者往往只能采用两种不尽如人意的方法之一：要么直接采用常见的默认值，例如 $\text{Gamma}(1,1)$ （例如，Cao 等人，2024年；Ma 和 Chen，2020；Mignemi 和 Manolopoulou，2025年），却不探究其影响；要么在缺乏原则性框架来解读结果的情况下进行临时性的敏感性分析（例如，参见 Kottas，2006年；Leslie 等人，2007）。

这一转换鸿沟影响重大。Lee 等人（2025年）表明，在信息性中等的多中心试验中（以中心间方差与全方差之比衡量），关于 α 的先验可能显著改变中心特异性效应及其分布的后验估计。当研究者假定“数据会自行说明一切”而采用模糊超先验时，可能会无意中对汇聚行为作出强假设——例如，伽马分布(1，1) 先验在 $\alpha = 0$ 附近赋予了显著的概率质量，这意味着先验上强烈预期簇数较少且汇聚程度较高。

文献已认识到，需要根据 α 对 K_J 的诱导分布（即已占据簇数）校准（Greve 等人，2022；Paganin 等人，2023）。Dorazio（2009年）提出了一种原则性方法，通过最小化诱导边际分布 $p(K_J \mid a, b)$ 与目标分布（例如在合理簇数范围内的离散均匀分布）之间的 Kullback–Leibler 散度，来校准伽马超先验。Murugiah 和 Sweeting（2012）提出了一种可扩展的设定策略（SCAL），为占据簇数设定易于解释的尾部概率，并根据设计规模求解 (a, b)。这些基于K_J的方法与从业者直觉十分契合，因为簇数易于解释：研究者可以根据领域知识，判断预期“约有 5 个不同子群”还是“约有 15 个”是否合理（Bonafos 等人，2024年）。

然而，现有基于 K_J 的校准方法存在三个显著局限。首先，存在转换鸿沟：即使从业者能够清楚表达他们对

K_J的看法，也没有标准的闭式映射能将这些看法转换为(a, b)，使得先验设定过程显得晦涩难懂。其次，存在计算鸿沟：正如 Dorazio（2009年）明确指出的那样，最优超参数不存在“闭式解”，因此需要在二维参数空间中进行网格搜索；而其他经验贝叶斯策略也无济于事，往往需要反复运行马尔可夫链蒙特卡洛（Antonelli 等，2016；Dorazio 等，2008），这使得系统性敏感性分析的计算负担难以承受（Giordano 等，2023）。第三，也许是最重要的一点，存在连贯性鸿沟：仅根据对簇数的看法校准先验，并不能保证混合权重所隐含的行为符合实质性期望。对于个单位而言，一个校准后能使期望簇数约为 10 的超先验，可能隐含着某个簇主导整个混合的显著先验概率——这意味着一种强汇聚假设，与原本“十个规模适中的组”这一解释相矛盾。Vicentini 和 Jermyn（2025年）将这种不匹配称为非预期先验现象。

Vicentini 和 Jermyn（2025年）从概念层面进一步阐明了这一连贯性问题。他们指出，基于 K_J的方法具有样本量依赖性（SSD）：校准超先验取决于 J，从而形成他们所称的“先验序列”，并随样本量变化而改变。在多项样本量不同的分析需要共享一个连贯的共同先验，或在 J 随时间增长的流式情境中，这可能并不理想。更根本地，Vicentini 和 Jermyn（2025年）证明，即使某个先验相对于簇数看似“弥散”，仍可能隐含关于权重集中度的强假设。他们提倡采用样本量独立（SSI）方法，以权重泛函而非簇数为锚点来设定先验。

我们采取一种不同但互补的观点。在许多教育与行为应用中——包括站点集合固定的多中心试验、针对特定研究集合的荟萃分析，以及样本明确的心理测量评估——样本量 J 并非任意设定，而是由研究设计决定的。在这类固定-J 情境中，通过 K_J 进行 elicitation 并非概念上的缺陷；它明确表达了一个具有实质意义的问题：“在这些 J 单位中，我们预计会有多少种不同模式？”相关的判断标准不是先验在对 J作假设性更改后是否仍然连贯，而是其在实际设计规模下所隐含的先验预测行为是否具有实质上的合理性。同时，我们也认真看待这样的批评：仅进行簇数量匹配，可能会掩盖不直观的权重影响。我们的解决方案是将基于 K_J的 elicitation 作为

固定-J 应用的主要锚点，同时要求 在报告划分摘要时一并报告权重诊断。当隐含的权重行为在实质上不可接受时，我们会提供一种明确机制，以纳入关于权重集中度的第二项约束，同时透明地报告由此产生的权衡。

本文提出设计条件式 *elicitation* （DCE），一个将从业者对聚类结构的信念转化为狄利克雷过程浓度参数伽马超先验的统一框架。该框架通过三项贡献，弥补了转化、计算和连贯性方面的缺口。首先，我们提出两阶段矩匹配 （TSMM），该方法提供闭式初始化，可将引出的关于 K_J 期望值和方差的信念直接映射为伽马超参数，随后通过精确的牛顿细化，在有限-J 狄利克雷过程诱导分布下确保目标矩得以满足。这种两阶段方法消除了现有方法所需的网格搜索：校准约需 50 毫秒即可完成，且达到机器精度；与差异优化方法相比，速度约提升 900 倍。其次，我们提出双锚点 诊断与细化方案，用于量化任何经 K_J校准的先验所隐含的主导风险。当诊断结果表明权重集中度不可接受时，可选细化步骤会限制主导性，同时明确报告与簇数量目标之间的权衡——将 SSI 批评背后的实质性关切纳入考量，同时保留设计条件式 elicitation 的可解释性。第三，完整工作流程——包括 elicitation、校准、诊断和报告——已在 DPprior R 包中实现，并附有可复现的示例文档和报告清单。

本文其余部分安排如下。第2节介绍 DPM 的基本背景，重点阐述关于 α的划分视角和权重视角。第3节介绍 DCE 方法。第4节报告将 DCE 与现有方法进行比较的模拟结果。第5节讨论其影响及未来方向。技术细节见在线补充材料（OSM）。

2. 理论框架

本节介绍狄利克雷过程（DP）混合模型中狄利克雷过程（DP）浓度参数 α 先验设定所需的最基本背景。我们的目的不是全面回顾狄利克雷过程混合模型，而是阐明对 α 设定先验在应用层面意味着什么，以及为何单一的解释目标可能造成误导。

对 α 设定先验会以两种不同但互补的方式影响聚类。首先，它会塑造诱导出的划分，将 J 可交换单位划分为若干组，并以占用簇数 K_J概括。其次，它会塑造诱导出的混合权重（w_1、 w_2，.....），这些权重决定了混合分布的集中程度，以及某个簇占据主导地位的可能性。这两种视角通过 α相互关联，但不能互换：将先验校准为使 K_J 的分布看起来合理，并不意味着权重行为也会自动变得合理。 图 1 展示了这种“非预期先验”现象，并为第3节中的诊断量（以及可选的双锚点细化）提供了动机。

全文采用形状–率参数化 $\alpha \sim$ 伽马分布(a， b)，因此$\mathbb{E}(\alpha) = a/b$ ，且方差(α) $= a/b^2$。

2.1. **最简狄利克雷过程混合设定与符号。** 我们考虑 J 由以下索引的可交换单位：$j = 1$、.....、 J。在多中心试验、荟萃分析及相关的“多单位”应用中，每个单位都有一个潜在参数 τ_j （例如，中心特异性治疗效应），并伴有估计噪声（例如，Cao 等人，2024年）：

$$\hat{\tau}_j \mid \tau_j \sim \mathsf{N}\big(\tau_j,\ \widehat{se}_j^{\,2}\big)\,, \qquad \tau_j \mid G \overset{\text{i.i.d.}}{\sim} G, \qquad G \sim \mathrm{DP}(\alpha, G_0). \tag{2.1}$$

这里， G_0 是中心化分布，而 $\alpha > 0$控制先验上创建新簇的倾向（Escobar 与 West， 1995； Ferguson， 1973年）。由于从狄利克雷过程中抽取的样本几乎必然是离散的，潜在取值 τ_1、.....、 τ_J会出现相同值，从而对 $\{1, \ldots, J\}$ 诱导出一个随机分区（Hanson 等人， 2005； Sethuraman， 1994）。令 z_j表示单位 j的簇标签；关键的设计条件汇总是已占据簇的数量，即在 J个单位中， $K_J = \#\{z_1, \ldots, z_J\}$。

一种能清晰呈现 α 作用的构造性表示是棍棒折断表示（Ishwaran 与 Zarepour, 2000； Sethuraman, 1994）。对于$G \sim$ 狄利克雷过程(α, G_0)，

$$G = \sum_{h=1}^{\infty} w_h\, \delta_{\theta_h}, \qquad \theta_h \overset{\text{i.i.d.}}{\sim} G_0, \tag{2.2}$$

其中

$$v_h \overset{\text{i.i.d.}}{\sim} \mathrm{Beta}(1, \alpha), \qquad w_1 = v_1, \qquad w_h = v_h \prod_{\ell<h} (1 - v_\ell), \quad h \geq 2. \tag{2.3}$$

权重序列遵循 GEM(α)，并采用按大小加权顺序（而非按从大到小的顺序排序）。我们使用 K_J 和（w_h）来阐释关于 α的两种解释视角。

2.2. **关于 α的两种解释视角：划分与权重。划分视角：通过 K_J理解粒度。**从从业者的角度来看， K_J 回答了一个具体的设计条件问题：在这 J 个单位中，我们预计会有多少个实质上不同的模式？更大的 α 会提高新单位启动新簇的概率，因此也会提高 K_J 的平均值（Congdon，2019；Escobar 与 West，1995）。在给定 α的条件下， K_J 遵循 Antoniak 分布（Antoniak，1974）：

$$\Pr(K_J = k \mid \alpha) = |s(J,k)|\, \alpha^k \, \frac{\Gamma(\alpha)}{\Gamma(\alpha + J)}, \qquad k = 1, \ldots, J, \tag{2.4}$$

其中 $|s(J,\ k)|$ 是第一类无符号斯特林数。在 第3节中，这一划分视角构成了我们的设计条件引出目标的基础：从业者指定J 以及对 $\mathbb{E}(K_J)$ 和 方差(K_J)的信念，我们据此校准 α上的伽马超先验，以匹配这些目标。

权重视角：通过 (w_h)考察集中度和主导性。 权重回答的是另一个问题： 即使存在多个簇，概率质量在这些簇之间是如何分布的——大致均匀，还是先验概率显著倾向于某个主导簇？一个便捷的诊断量是第一个断棒权重（Navarro 等，2006年）：

$$w_1 \mid \alpha \sim \mathrm{Beta}(1, \alpha), \tag{2.5}$$

因此 概率$(w_1 > t \mid \alpha) = (1-t)^{\alpha}$。诸如 概率$(w_1 > 0.5)$和 概率$(w_1 > 0.9)$这样的尾部概率，可对中等和极端主导风险提供易于解释的摘要（Miller & Harrison，2018）。在 $\alpha \sim$ 伽马分布$(a,\ b)$下，这些尾部概率有简单的闭式表达；参见第 3.4 节。

可解释性方面的注意事项。 由于$\mathrm{GEM}(\alpha)$采用大小偏倚顺序， w_1 不一定是最大的权重（Sethuraman，1994；Vicentini与Jermyn，2025年）。准确的解释是： w_1 是包含随机抽取单位的簇的渐近质量。因此，概率$(w_1 > t)$ 表示随机单位属于概率质量超过 t的簇的概率，而不是最大簇超过t的概率。

作为另一项汇总指标，我们考虑共聚类（辛普森）指数

$$\rho = \sum_{h \geq 1} w_h^2, \tag{2.6}$$

其等于随机选取的两个单位属于同一簇的概率。结合起来， K_J、 w_1和 ρ 提供了一组小型的解释性泛函，可以

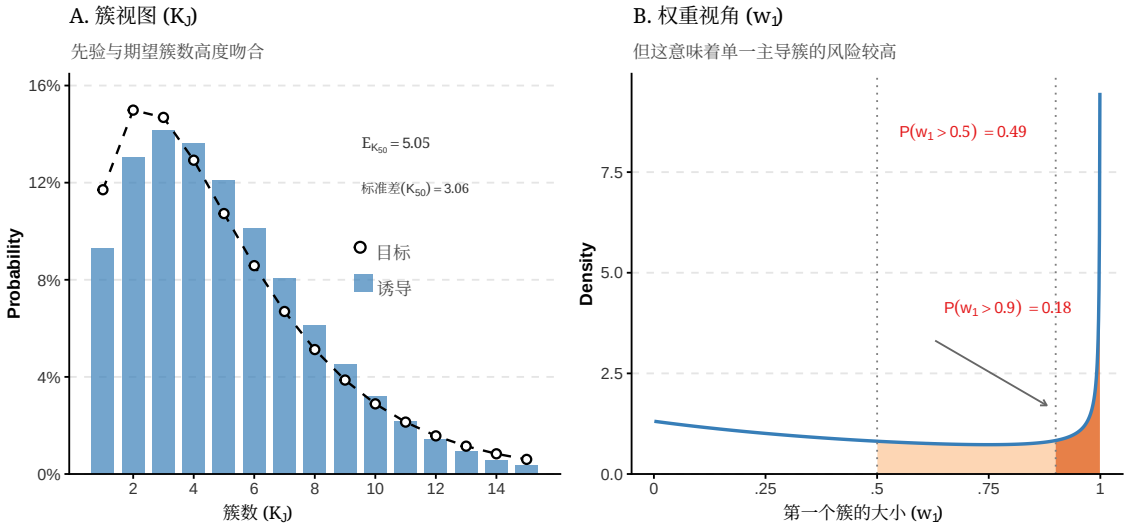
表1。从业者沟通的解释性泛函

量	统计学含义	从业者问题	如何报告
$\mathbb{E}[K_J]$	期望数量的已占据簇数（在 J 个单位	“我预计会有多少个不同的子群我期望什么？	“我们预计大约为 X 个簇。”
$\mathrm{Var}(K_J)$	关于已占用簇的不确定性	“我对这个数字有多大把握？”	“中心区间（例如，80% 或 95%）：$[L, U]$。”
概率$(K_J = k)$	数量边缘概率恰好 k 个簇出现	“恰好……的概率是多少 k 组？”	“众数位于 k^*；大部分概率质量集中在 $[a, b]$。”
概率$(w_1 > 0.5)$	一个随机单位的概率随机单位属于一个占比超过 50% 概率质量	“是否可能有一个组主导整个样本？”	“有 X% 的概率多数簇主导性。”
概率$(w_1 > 0.9)$	极端集中极端度（> 90% 概率质量）	“几乎所有单位是否都可能属于同一个组？”	“X% 极端主导风险。”
$\rho = \sum_{h \geq 1} w_h^2$	共同聚类（辛普森）指数；概率为二随机单元属于同一个簇	“如果我随机挑选两个单位，它们会属于同一个簇？”	“X% 的概率是同一聚类成员关系。”

注。 对于 $\alpha \sim$ 伽马分布(a, b)，所有量均有闭式表达式或可通过高效数值计算求得（精确的矩推导见 OSM 中的附录 A，权重分布理论见附录 E）。

向从业者进行说明，并以透明的方式报告。表1总结了这些量。

2.3.固定设计引导与非预期先验现象。 在许多教育与行为应用中——多中心试验、荟萃分析以及心理测量研究——可交换单位 J 的数量由设计固定，研究人员往往以有限总体估计目标为目标（例如，“我们研究中的 J 个站点”）。在此类情形下，以条件于设计的表述引出信念（“在这 J 个单位中，有多少种不同模式？”）是自然的，从而使 K_J 成为自洽的主要锚点。与此同时，近期研究强调，以 K_J为锚点的先验可能隐藏



注：J = 50。先验 alpha ~ 伽马分布(1.60, 1.22)。w1 是按大小加权（GEM–顺序）的第一个簇比例。

图 1。非预期先验现象：经 K_J 校准的先验，在权重视角下可能意味着显著的主导风险。面板 A： K_{50} 的目标分布（圆点）与诱导分布（柱状图）。面板 B： w_1 的边际密度及主导尾部概率。$J = 50$; α ~伽马分布(1.60, 1.22)。

对权重集中度不直观的含义（Vicentini与Jermyn，2025年）。我们的方法保留了 K_J 的可解释性，同时要求将权重行为与划分摘要一并报告（并在需要时加以控制）。

图 1 给出了一个典型示例。经校准以匹配合理目标分布的伽马超先验，对于 K_{50} 所对应的簇数量行为，能够很好地复现预期结果（面板 A），但在权重视角下却意味着显著的主导风险（面板 B）。具体而言，取 $J = 50$ 和 $\alpha \sim$ 伽马分布(1.60,1.22)时，先验与簇数量目标相当吻合（$\mathbb{E}[K_{50}] = 5.05$，标准差($K_{50}$) $= 3.06$），但意味着概率($w_1 > 0.5$) $= 0.49$ 以及 概率($w_1 > 0.9$) $= 0.18$。由于 w_1 是按大小加权的，这些结果意味着，随机选取的一个单位属于多数簇的先验概率接近 50%，属于近乎全体（$> 90\%$）簇的先验概率则为 18%。

这种反差往往令人意外：“预计约有五个组”常常会被（隐含地）理解为“有五个规模适中的组”，但当 α 在超先验下不确定时，仅靠 K_J 校准并不能保证这种直觉成立。对实践而言，关键启示并非应当放弃基于 K_J 的 elicitation，而是

应将其与透明的权重诊断结合起来，并在适当时明确权衡。

3. 方法：设计条件 elicitation（DCE）

本节介绍设计条件 *elicitation*（*DCE*），这是一套工作流程，可将从业者对聚类结构的信念转化为狄利克雷过程浓度参数的伽马超先验。DCE 之所以是“设计条件式”的，是因为从业者通常会围绕一组固定 的 J 单位（例如站点、学校、教师、研究、受访者）来思考异质性；而对 α 采用相同先验，在不同的 J下可能会隐含实质上不同的划分行为。

elicitation 的核心目标是占据簇数 K_J，其定义为：在狄利克雷过程先验下，J 个可交换单位中所包含的不同簇的数量。DCE 以这一基于划分的锚点为基础，并辅以诊断量（以及在需要时进行额外校准），以评估其隐含的混合权重行为。具体而言，我们采用基于 GEM(α) 表示中的第一个断棒权重 w_1 的摘要，并可选择使用共同聚类概率 $\rho = \sum_{h\geq 1} w_h^2$。

3.1. 概述与工作流程。 DCE 从一个实际观察出发：从业者通常很难直接把握 α，但许多人能够说明 (i) 他们预计在其设计中的 J 个单位里会观察到多少个实质上不同的簇，以及 (ii) 他们对这一预期有多大把握。具体而言，所需输入包括设计规模 J、目标均值 $\mu_K = \mathbb{E}(K_J)$，以及不确定性的摘要，例如 $\sigma_K^2 = \mathrm{Var}(K_J)$（或映射到方差的定性置信度标签；默认的置信度到方差映射见 OSM 中的附录G.4）。作为可选的第四项输入，从业者还可以给出支配容差，例如要求满足 $\Pr(w_1 > 0.9) \leq \delta$。

根据这些输入，DCE 会返回为 $\alpha \sim \Gamma(a, b)$ 校准的超参数 (a, b)，以及一份简明的诊断报告，概述诱导的先验预测分布 $p(K_J)$ 和隐含的权重集中行为，包括在默认阈值 $t \in \{0.5, 0.9\}$ 以及（可选的） ρ下的尾部概率 $\Pr(w_1 > t)$。工作流程如图2所示，由两个部分组成：两阶段矩匹配（TSMM），用于校准 (a, b) 以匹配 (μ_K, σ_K^2)；以及双锚点协议，用于诊断（并在需要时约束）权重主导。

设计条件框架纳入了促成这一与样本量无关的批评的关键警示：仅通过 K_J 校准先验，并不会自动控制诱导的权重（Vicentini与Jermyn，2025年）。因此，DCE使

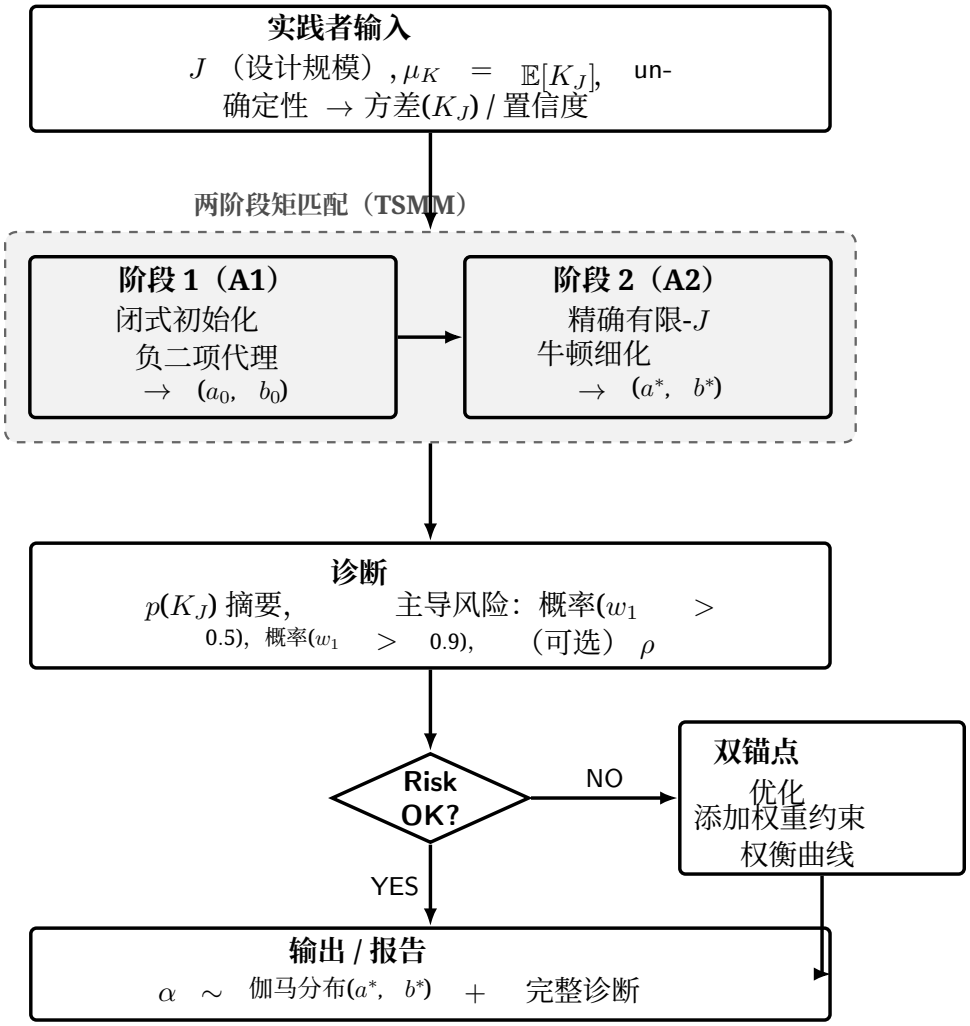


图2. 设计条件式 elicitation（DCE）工作流程。实践者指定 J、目标 $\mu_K = \mathbb{E}(K_J)$和不确定性 σ_K^2。TSMM校准(a, b)；诊断报告 $p(K_J)$和 Pr($w_1 > t$)。当主导风险超过容差时，可选的对偶锚点会进一步细化。

权重影响通过诊断量得以明确，并在主导风险不可接受时提供可选的对偶锚点细化。

3.2. **TSMM 阶段 1：闭式初始化。** 关于 K_J 在狄利克雷过程先验下的分布，存在众所周知的精确形式（公式 (2.4)；Antoniak，1974），但它并不能给出从 （μ_K， σ_K^2）到 （a， b）的简单闭式反演；现有校准通常依赖数值搜索（例如，Dorazio，2009年）。因此，TSMM 从一个足够精确、能够提供有效初始值的大J 代理开始。

算法1 带诊断报告的两阶段矩匹配（TSMM）输入：J；K_J的目标矩（μ_K, σ_K^2）；可选的支配容差（t, δ）。输出：（$a^\star, b^\star$），用于定义 $\alpha \sim \Gamma(a^\star, b^\star)$ 以及一份诊断报告。1：阶段 1（闭式初始化）。设置 $\mu_0 = \mu_K - 1$ 并将 $c_J =$取对数 J。计算 $a_0 = \mu_0^2/(\sigma_K^2 - \mu_0)$ 和 $b_0 = \mu_0\, c_J/(\sigma_K^2 - \mu_0)$。2：阶段 2（精确有限-$J$ 细化）。从（a_0, b_0）出发，在（log a, log b）中应用牛顿法更新，直到精确混合矩满足$\mathbb{E}_{a,b}(K_J) = \mu_K$且 $\mathrm{Var}_{a,b}(K_J) = \sigma_K^2$（关于通过高斯–拉盖尔求积进行矩评估和解析雅可比推导，参见 C节）。3：诊断。报告 $p(K_J)$的摘要（例如，中位数和中心区间）、Pr($w_1 > 0.5$)、Pr($w_1 > 0.9$)，以及可选的 $\mathbb{E}(\rho)$。4：升级（可选）。如果提供了支配容差且 Pr($w_1 > t$) $> \delta$，则应用算法2。

我们采用一种遵循支撑约束的移位近似 $K_J \geq 1$（Arratia等人，2000；Miller & Harrison，2018；Murugiah与Sweeting，2012；Vicentini与Jermyn，2025年）：

$$K_J - 1 \mid \alpha \approx \mathrm{Poisson}(\alpha\, c_J), \qquad c_J = \log J. \tag{3.1}$$

在 $\alpha \sim \Gamma(a,\ b)$ 下，泊松-伽马混合意味着 $K_J - 1$ 可用负二项代理表示，反演过程则通过匹配该代理的均值 $\mu_0 = \mu_K - 1$ 和方差 σ_K^2 得出：

$$a_0 = \frac{\mu_0^2}{\sigma_K^2 - \mu_0}, \qquad b_0 = \frac{\mu_0\, c_J}{\sigma_K^2 - \mu_0}. \tag{3.2}$$

B 节给出了基于伽马分布–泊松共轭关系推导该闭式映射的完整过程，并记录了有限 J 下的近似质量。

当 $\mu_0 > 0$ 且 $\sigma_K^2 > \mu_0$ 时，阶段 1 映射定义良好；从业者表达非退化不确定性时，通常正处于这种情形。当输入接近边界时（例如，对 K_J 抱有近确定性信念），将采用数值防护措施（可行性投影策略见 OSM 中的附录 B.3.3）。阶段 1 用作初始值；阶段 2 则消除有限-J 偏差。

3.3. **TSMM 阶段 2：精确矩细化。** 阶段 1 有意设计得较为简单，但对于中等 J 而言可能存在偏差，$K_J \mid \alpha$因为其条件分布 $K_J \mid \alpha$并非泊松分布：它是伯努利指示变量之和，相较于泊松代理通常呈欠离散状态（Arratia等人，2000）。因此，阶段 2 会细化 (a, b) 使得 K_J 在 $\alpha \sim \Gamma(a, b)$ 下的先验预测（混合）矩与征询目标值完全吻合。

令 $\kappa_J(\alpha) = \mathbb{E}(K_J \mid \alpha)$ 和 $v_J(\alpha) = \mathrm{Var}(K_J \mid \alpha)$。这些条件矩可以用双伽马函数 $\psi(\cdot)$ 和三伽马函数$\psi_1(\cdot)$表示为闭式（Antoniak, 1974）：

$$\kappa_J(\alpha) = \alpha\big\{\psi(\alpha + J) - \psi(\alpha)\big\}, \tag{3.3}$$

$$v_J(\alpha) = \kappa_J(\alpha) - \alpha^2\big\{\psi_1(\alpha) - \psi_1(\alpha + J)\big\}. \tag{3.4}$$

这些条件矩由泊松–二项表示推导而来，其关于 α 的导数详见A.2.3 节。

阶段 2 通过对伽马分布的密度进行数值积分，计算精确的混合矩 $\mathbb{E}_{a,b}(K_J) = \mathbb{E}_{a,b}\{\kappa_J(\alpha)\}$ 以及方差 $_{a,b}(K_J) = \mathbb{E}_{a,b}\{v_J(\alpha)\} + $方差 $_{a,b}\{\kappa_J(\alpha)\}$，然后应用牛顿更新求解两个矩匹配方程。我们在 (log a, log b) 空间中执行更新，以保持正值性，并使用解析导数确保稳定性。在典型应用中，从第 1 阶段初始值出发，经过少数几次迭代即可收敛。所有推导、通过得分函数恒等式得到的雅可比表达式，以及收敛性保障措施，均详见 C节。

3.4. 双锚点协议：诊断并酌情细化。 对 K_J 个锚点进行矩匹配，可以为划分结构设定先验，但无法唯一确定权重集中度。因此，DCE 会报告明确的权重诊断，并在主导风险不可接受时提供可选的细化。

3.4.1. 权重与共同聚类诊断量。 在截断棒表示下，第一个权重服从 $w_1 \mid \alpha \sim$ 贝塔分布(1, α)，因此 概率$(w_1 > t \mid \alpha) = (1 - t)^{\alpha}$。若$\alpha \sim \Gamma(a, b)$，则尾概率可通过闭式积分求得（Vicentini与Jermyn, 2025年）：

$$\Pr(w_1 > t \mid a, b) = \left(\frac{b}{b - \log(1 - t)}\right)^a, \qquad 0 < t < 1. \tag{3.5}$$

我们建议将概率$(w_1 > 0.5)$和概率$(w_1 > 0.9)$作为中等和极端主导风险的默认摘要。通过拉普拉斯变换积分推导该闭式尾概率的方法，以及用于优化的梯度，见 OSM 的附录 E.3。

作为额外的单位级摘要，我们还可选择报告共同聚类概率 $\rho = \sum_{h \geq 1} w_h^2$，即随机选取两个单位属于同一簇的概率。在狄利克雷过程下，$\mathbb{E}(\rho \mid \alpha) = 1/(1 + \alpha)$，因此，$\alpha \sim \Gamma(a, b)$ 下的 $\mathbb{E}(\rho)$ 可通过一维积分计算（闭式表达式和方差计算见 OSM 中的附录 E.4）。

Alg算法 2 双锚点细化 (Opional

) 输入：TSMM 解 (a_K, b_K)；主导性容差 $\Pr(w_1 > t) \le \delta$（可选）；权衡权重 $\lambda \in (0, 1]$（可选）。输出：精细化后的 $(a^\star, b^\star)$ 及权衡报告。1：使用公式（3.5）计算 (a_K, b_K) 处的主导性诊断量。若 $\Pr(w_1 > t) \le \delta$，则停止并保留 (a_K, b_K)。2：否则，定义一个权衡准则，对以下情况施加惩罚：（i）偏离目标矩；（ii）违反主导性容差；λ 控制两者的相对权重。3：以 (a_K, b_K) 为初始值，在 (a, b) 上优化该准则。4：报告精细化后的 $(a^\star, b^\star)$，以及 $\mathbb{E}(K_J)$、$\mathrm{Var}(K_J)$ 和 $\Pr(w_1 > t)$ 的变化结果，使权衡一目了然。

3.4.2. 通过双锚点进行可选精化。 当从业者给出主导性容差（例如 $\Pr(w_1 > 0.9) \le \delta$），而 TSMM 校准先验不满足该容差时，我们提供一种双锚点细化方法，将隐含的权衡明确呈现出来。该细化通过对 (a, b) 求解一个受约束（或带惩罚项的）优化问题，在贴合 K_J 目标与所需的支配控制之间取得平衡。改变权衡权重，可以得到簇数量拟合与主导风险之间的帕累托前沿。由于此步骤为可选步骤，且取决于具体应用，其目标函数、约束处理和数值优化细节见 OSM 的附录 E.5，其中还包括用于权衡分析的帕累托前沿可视化。

3.5. **示例与报告指南**。 为说明这一工作流程，考虑一个包含 $J = 50$ 个单位的设计：从业者预计其中约有五个实质上不同的簇，且不确定性为中等。假设征询得到的目标值为$\mu_K = \mathbb{E}(K_{50}) = 5$ 和 $\sigma_K^2 = \text{方差}(K_{50}) = 10$。

阶段 1 得到 $\mu_0 = 4$ 和 $c_{50} = \log(50)$，从而产生闭式初始值 $(a_0, b_0) \approx (2.667, 2.608)$。阶段 2 将其细化为精确矩解 $(a^\star, b^\star) \approx (1.407, 1.076)$，根据构造与 $\mathbb{E}(K_{50}) = 5$ 以及 $\mathrm{Var}(K_{50}) = 10$ 相匹配。在此校准先验下，诱导先验预测分布 $p(K_{50})$ 的中位数约为 4，中央 90% 区间约为 $[1, 11]$。权重诊断给出 $\Pr(w_1 > 0.5) \approx 0.50$ 和 $\Pr(w_1 > 0.9) \approx 0.20$，可选的共同聚类汇总为 $\mathbb{E}(\rho) \approx 0.52$。

如果从业者不愿接受随机抽取的单位落入概率质量超过 90% 的簇中的概率为 20%（即，如果他们要求 $\Pr(w_1 > 0.9) \le 0.10$），DCE 会保留 TSMM 解作为以 K 为锚点的基线，然后

采用可选的双锚点细化，明确权衡矩忠实度与支配控制。

报告指南。 为促进透明度和可重复性，建议使用 DCE校准先验的研究者报告以下核心要素：

(1) 设计规范： J 以及将其视为固定值的原因。(2) *elicitation*目标： $\mathbb{E}(K_J)$ 和 Var(K_J)，并简要说明其实际依据。(3) 校准超先验： (a, b)，用于 $\alpha \sim \Gamma(a, b)$（形状–率参数化）。(4) 权重诊断： Pr($w_1 > 0.5$) 和 Pr($w_1 > 0.9$)——这些是必需项，而非可选项。(5) 双锚点摘要（如适用）： 权衡参数 λ以及由此导致的矩和支配概率变化。

权重诊断应作为常规报告内容：即使不采用双锚点，报告 Pr($w_1 > 0.5$) 和 Pr($w_1 > 0.9$) 也能使隐含的汇聚几何结构一目了然，并直接回应 SSI批评的核心关切。附录 E.6 的 OSM 中提供了详细的报告检查清单、报告陈述示例和扩展指南。

4. 模拟研究

第 2 节和第 3 节确立了设计条件先验 elicitation 的理论框架，并介绍了可选用双锚点细化的两阶段矩匹配（TSMM）算法。本节评估这些方法能否兑现其承诺：相较于传统默认设定和现有校准方法，改善对已占据簇数量 K_J 的后验推断。

模拟研究探讨三个实质性问题。第一，后验聚类对先验设定有多敏感？在什么条件下这种敏感性最为重要？第二，与模糊默认设定和 Dorazio (2009) 的差异优化方法相比，TSMM 校准能否显著改善对真实簇结构的后验恢复？第三，何时双锚点扩展（该扩展纳入权重集中度约束）能在仅针对 K的校准之外带来额外益处？

4.1. 模拟设计。数据生成过程。

我们采用一个层级模型，以反映第 1 节中描述的多中心试验和荟萃分析情境。按照 Lee等人（2025年）的做法，设 J 个可交换单位具有潜在参数

τ_j 观测值带有估计噪声：

$$\hat{\tau}_j \mid \tau_j \sim N(\tau_j,\, \widehat{se}_j^2), \quad j = 1, \ldots, J. \tag{4.1}$$

真实效应 τ_j 来自一个具有相等权重 $\pi_k = 1/K^*$ 的离散 K^*成分混合，其成分均值在对称网格 $\mu_k \in [-2.5, 2.5]$上等间距排列。 $\sigma_{\text{within}} = 0.15$ 的簇内变异会产生可区分但相互重叠的簇。每个模拟数据集都确保所有 K^* 簇均有表示，从而使“真实”簇数具有有意义的参考价值。我们固定$J = 100$ 个单位，样本量取教育多地点试验中常见的规模（Weiss 等，2017），并设定 $\sigma_\tau = 0.25$。

解释注意事项。 在具有重叠混合成分的有限样本中，既不预期也不要求精确恢复 K^* 。我们将 K^* 用作比较方法的参考线，而不是精确恢复的金标准。下面报告的指标用于在受控情境下诊断先验诱导的收缩和合并；它们并非关于结构性真实值的频率学覆盖率的声明。

析因设计与信息量校准。 该模拟采用一个 3×3 析因设计，交叉两个因子：

真实簇数 ($K^* \in \{5, 10, 30\}$)。这些取值涵盖了从“少数实质性子群体”到“许多不同模式”的范围，覆盖了典型的应用情景。

数据信息量 ($I \in \{0.2, 0.5, 0.8\}$)。遵循 Lee等人（2025），信息量定义为

$$I = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \exp\left(\frac{1}{J}\sum_{j=1}^{J} \log \widehat{se}_j^2\right)}, \tag{4.2}$$

其中，分母将典型的单位内方差（ $\widehat{se}_j^2$ 的几何均值）与单位间方差 σ_τ^2相加。低信息量（$I = 0.2$）代表先验至关重要的高噪声情形；高信息量（$I = 0.8$）代表数据丰富的情形，此时从业者可能会认为“数据会自行说明一切”。

我们通过 $\widehat{se}_j^2 = \kappa / n_j$ 生成标准误，采用 $\kappa = 4$ ，并且站点规模 n_j 取自截断对数正态分布（变异系数为 0.5，最小值为 5）。在每个情景的 $R = 200$ 次重复实验中，目标值分别为 0.2、0.5 和 0.8 时，实际信息量分别为 0.194（标准差 $= 0.008$）、0.500（标准差 $= 0.014$）和 0.793（标准差 $= 0.009$）。

比较的方法。 我们比较了针对 $\alpha \sim \text{Gamma}(a, b)$ 的四种超先验设定：

(1) 模糊：Gamma(1, 1)，一种常见的默认值，意味着 $\mathbb{E}(\alpha) = 1$，并且在较小的 α 取值上具有相当大的先验概率质量（例如，Cao 等，2024；Dunson 等，2007；Ma 和 Chen，2020；Mignemi 和 Manolopoulou，2025）。(2) DORO-Unif：Dorazio (2009) 提出的 KL 散度最小化方法，经校准后在 $K_J \in \{1, \ldots, m\}$（其中 $m = \min\{2\mu_K - 1, J\}$）上产生近似均匀的先验。(3) TSMM：两阶段矩匹配（第 3.3 节），经校准后匹配 elicitation目标 $\mathbb{E}(K_J) = K^*$以及 $\text{Var}(K_J) = \text{VIF} \cdot (\mu_K - 1)$，其中 $\text{VIF} = 2.5$（中等置信度）。(4) 双锚点：带有可选权重约束的 TSMM（第 3.4 节）；当 TSMM 解下 $\Pr(w_1 > 0.5) > 0.40$ 时触发，以 $\Pr(w_1 > 0.$

$5) \leq$

0.25，权衡权重为 $\lambda = 0.70$。

对于每次重复试验，我们通过马尔可夫链蒙特卡洛拟合狄利克雷过程混合模型（两条链各运行 1,000 次迭代，其中 500 次为预烧），并记录后验平均簇数 $\mathbb{E}[K_J \mid \text{数据}]$,坍缩概率 概率$(K_J = 1 \mid \text{数据})$、 K_J的 95% 可信区间，以及 α的后验均值。完整的 MCMC 诊断、额外基线（包括 χ^2-DORO）以及完整的超先验参数表见 F 节 的 OSM。

4.2. 默认先验的危害：簇坍塌。 贝叶斯分析的一个核心前提是，随着数据不断累积，先验的影响会逐渐减弱。然而，我们的结果表明，在有限样本混合模型中（$J = 100$），即使数据看起来信息量充足，对 α 的超先验的选择仍会产生重要影响（Matza 和 Bistritz，2021）。

模糊默认先验的系统性失效。 图 3展示了核心发现：无论真实簇数或数据信息量如何，模糊先验都会产生后验估计 $\mathbb{E}[K_J \mid \text{数据}] \approx 2$。这种系统性坍缩的原因是伽马分布(1，1) 在 $\alpha = 0$附近赋予了显著的先验概率质量，从而对复杂性施加强烈惩罚，而似然难以轻易克服这种惩罚。即使在 $K^* = 30$ 和$I = 0.8$的情况下——即真实簇数为 30 且测量相对精确——模糊先验仍会得出 $\mathbb{E}[K_J \mid \text{数据}] = 1.83$ (标准差 $= 0.65$)。

我们将这一现象称为簇坍塌：后验分布赋予 $K_J = 1$显著的概率，这意味着所有单位都属于单个簇。模糊先验在所有情景下的坍缩率约为 60–67%——也就是说，在超过一半的分析中，无论真实结构包含 5、10 还是 30 个簇，后验分布都会集中于单簇解（表 2）。

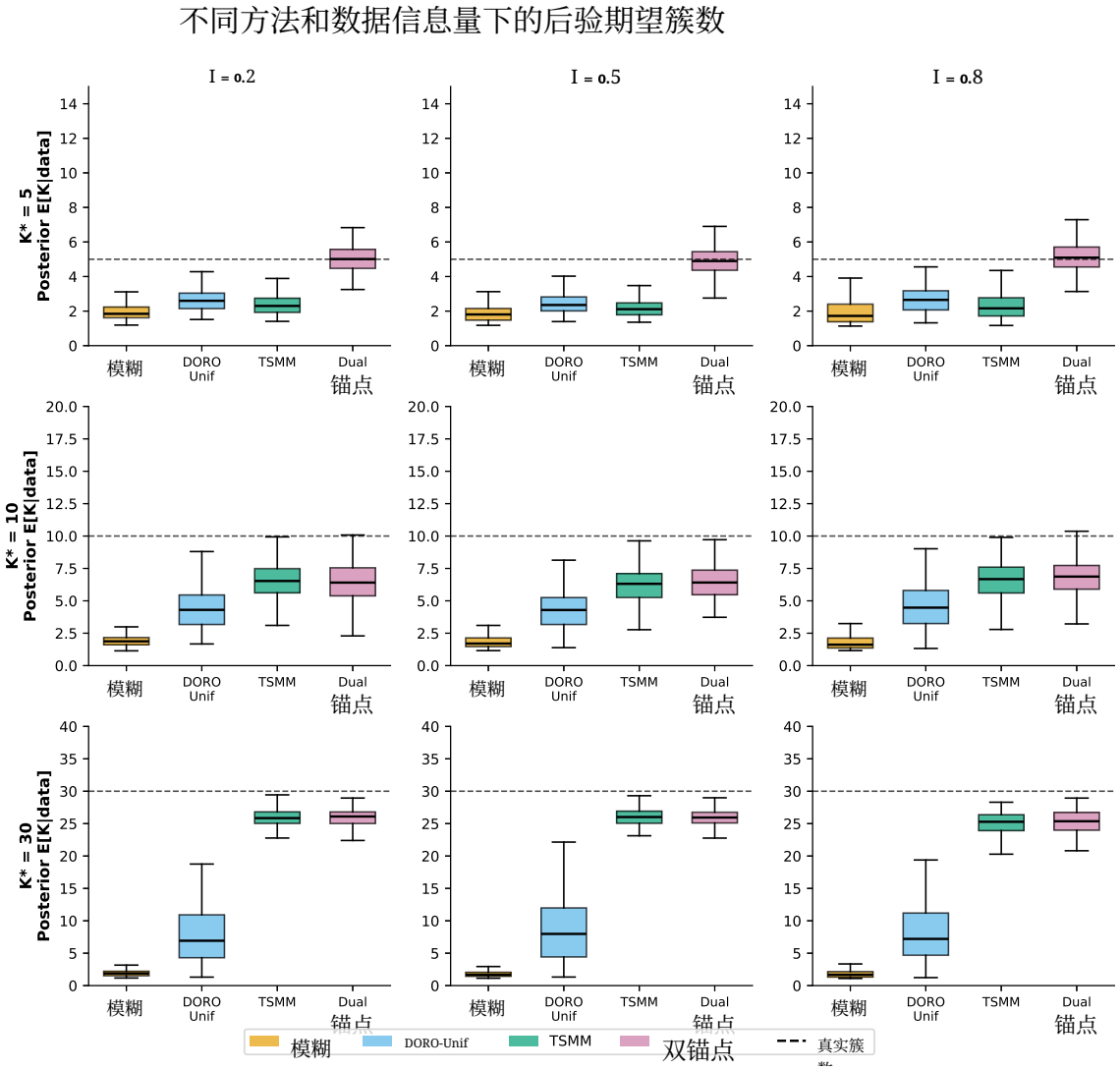


图 3. 按方法和情景展示 $\mathbb{E}[K_J$ |data] 的后验期望簇数。行： $K^* \in \{5, 10, 30\}$; 列： $I \in \{0.2, 0.5, 0.8\}$。200 次重复实验的箱线图；虚线表示真实 K^*。

TSMM 显著改善了恢复效果。TSMM校准先验通过明确编码设计条件期望，实现了定性上的改进。对于 $K^* = 30$ 且 $I = 0.2$ 的情况，TSMM 得到 $\mathbb{E}[K_J \mid$ 数据 $] = 25.89$ （标准差 $= 1.30$），而模糊先验下为 1.96——这意味着对真实聚类结构的恢复率从 7% 提高到 86%。对于 $K^* = 30$，95% 可信区间的覆盖率达到 100%，坍缩概率则降至 0%。

当征询目标值与模糊先验诱导的后验分布之间的差距最大时，改进最为显著。对于 $K^* = 30$，其中 Gamma(1,1) 意味着

$\mathbb{E}(K_J) \approx 5$ 先验地，二者严重不匹配，因此 TSMM 的修正效果显著。对于 $K^* = 5$，模糊先验更接近目标，TSMM 带来的改进较为有限；而且，正如我们接下来将展示的，它还存在一个局限，这促使我们提出双锚点扩展。

DORO-Unif：有所改善，但仍不够。 Dorazio (2009) 方法优于模糊先验，但与 TSMM 相比仍有显著差距。当 $K^* = 30$ 为 I =0.2 时，DORO-Unif 在 $\mathbb{E}[K_J \mid \text{data}] = 8.04$ 上的结果为（标准差 $= 4.96$），仅恢复了约 25% 的真实簇数。由于其目标是在整个范围$\{1, \ldots, J\}$上采用均匀分布，因此概率质量分布得过于稀疏，未能提供支持复杂划分结构所需的正则化。

4.3. 权重控制与双锚点的优势：TSMM 在较小 下的脆弱性。 对于 $K^* = 5$，一个关键规律逐渐显现：与模糊先验相比，TSMM 降低了坍缩概率，但降幅有限（从 $\sim$60%降至 $\sim$48–54%）。这一现象与 Vicentini and Jermyn (2025) 对“非预期先验”的批评一致：仅为匹配 K_J 的矩而校准的先验，可能会无意中暗示主导簇权重具有较高概率。

回顾第 3.4 节，对于 $K^* = 5$，中等置信度下的 TSMM 解给出$\Pr(w_1 > 0.5) \approx 0.56$——随机选取的单位有 56% 的先验概率属于概率质量超过总质量一半的簇。这种权重集中会导致后验脆弱性：当数据含有噪声且似然略微偏向简约模型时，先验对主导性的容忍会使后验分布集中于单簇解。

双锚点提供针对性保护。双锚点扩展通过加入权重主导性约束来应对这种脆弱性。当 TSMM 解超过支配容差（$\Pr(w_1 > 0.5) > 0.40$）时，算法会调整超参数，以约束权重集中度，同时尽可能保持K匹配。

对于 $K^* = 5$，双锚点将坍缩概率从 TSMM 的 $\sim$48%降至约 10–14%——减少至五分之一。这种保护显著改善了估计：$\mathbb{E}[K_J \mid \text{数据}] = 5.04$（基本无偏），均方根误差仅为 0.76，而 TSMM 的均方根误差为 2.70。95% 可信区间的覆盖率达到 100%（表 2）。

关键在于，对于 $K^* \geq 10$，双锚点机制不会激活，因为 TSMM 的解已经满足支配容差。结果与 TSMM 几乎完全相同，这证实该约束是自适应的——它仅在

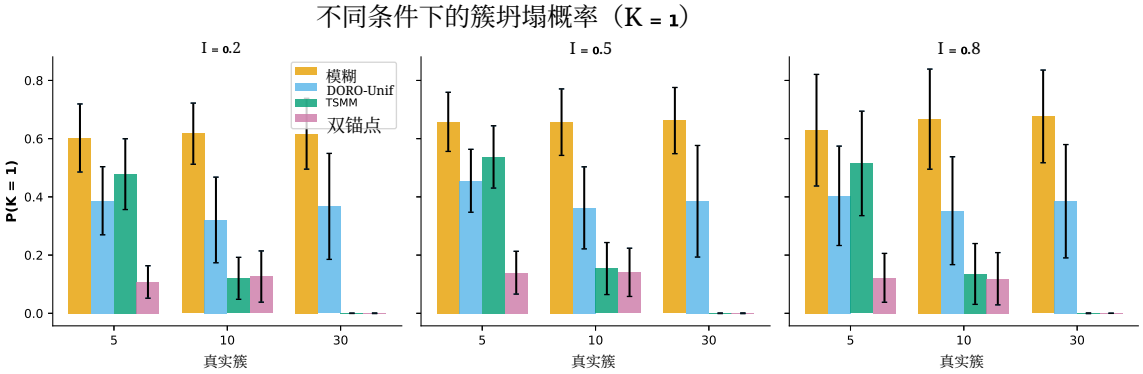


图4. 不同条件下的簇坍塌概率 $\Pr(K_J = 1 \mid \text{data})$。柱形：坍塌概率均值；误差条：200 次重复试验的标准差。

当 K 校准先验与稳健聚类所需的权重稳定性发生冲突时。

4.4. **信息量无法弥补不当先验**。 一个或许令人惊讶的发现是，将数据信息量从 $I = 0.2$ 提高到 $I = 0.8$ ，对设定不当的先验仅能带来有限改善。 图 3 显示，在不同信息量水平下，各方法的排名基本不变：无论 I如何，模糊先验都会失效；即使数据“信息丰富”，TSMM 和双锚点的相对优势依然存在。

这一模式挑战了一种直觉：当信息充足时，“数据会压倒先验”。在 $J = 100$个单位的样本量下，考虑到数据生成过程中的重叠聚类结构，即使 $I = 0.8$ 也无法提供足够的似然信息，来克服一个将概率质量集中在 $\alpha \approx 0$附近的先验。 K_J 的后验分布仍然受先验的结构性假设主导。

这一结果的实际启示是，即使在看似数据丰富的情境中，原则合理的先验设定仍然至关重要。研究者不能指望大样本量或精确测量能够弥补编码了不恰当聚类假设的默认先验。

4.5. **总结与实践启示**。 表 2汇总了低信息情景（$I = 0.2$）下的主要发现；对于多中心试验和荟萃分析而言，这是最具挑战性、也可以说与政策最为相关的情景。

表 2. 低信息设置下的方法性能汇总（$I = 0.2$）

K^*	方法	$\mathbb{E}[K$ |数据]	Bias	均方根误差	概率(K=1)	覆盖率	置信区间宽度
5	模糊	1.99 (0.55)	−3.01	3.06	0.602	0.775	5.84
5	DORO-Unif	2.64（0.63）	−2.36	2.44	0.387	0.955	6.73
5	TSMM	2.37（0.61）	−2.63	2.70	0.478	0.900	6.58
5	双锚点	5.04（0.76）	+0.04	0.76	0.108	1.000	10.83
10	模糊	1.95 (0.49)	−8.05	8.07	0.617	0.175	5.97
10	DORO-Unif	4.51 (1.75)	−5.49	5.76	0.321	0.880	14.37
10	TSMM	6.54 (1.27)	−3.46	3.69	0.120	1.000	16.01
10	双锚点	6.37 (1.47)	−3.63	3.91	0.126	0.990	15.49
30	模糊	1.96 (0.55)	−28.04	28.04	0.617	0.000	5.82
30	DORO-Unif	8.04 (4.96)	−21.96	22.51	0.367	0.570	30.08
30	TSMM	25.89 (1.30)	−4.11	4.31	0.000	1.000	31.18
30	双锚点	25.97 (1.39)	−4.03	4.27	0.000	1.000	31.47

注。 取值为 200 次重复实验的均值（标准差）。覆盖率指 K^*是否位于 K_J的 95% 后验区间内。所有信息量水平、置信水平敏感性分析和计算基准测试的完整结果见 F 节OSM（表 F.4–F.8）。

综合来看，这些模拟支持三项实践结论。第一，在有限样本中，默认设置可能会产生系统性误导。即使数据具有中等程度的信息量，模糊超先验也可能导致后验分布坍缩至一个或两个簇。Gamma(1，1) 的失效并非保守推断所致，而是无法通过增加数据克服的系统性偏差。第二，当分析人员能够根据其设计说出已占据簇数量的合理范围时，基于矩的 elicitation（TSMM）是一种有效的一线补救方法。在各种情景下，与模糊先验相比，TSMM 可将均方误差降低 76–98%。第三，使用小规模μ_K elicitation 时，权重诊断并非可有可无。恰恰是在从业者最自然地以少数亚型（例如“几种效应模式”）来思考的情景中，隐含的大小偏倚权重仍可能高度集中，而双锚点细化可以显著改变后验行为。

按照第3.5节中的报告清单，我们建议应用分析公开征询目标值 (μ_K, σ_K^2)、由此得到的超参数 (a, b)以及至少一种支配风险诊断量（例如 $\Pr(w_1 > 0.5)$）。基于这些发现，我们建议将双锚点作为设计条件式 elicitation的默认方法

在设计规模 J 固定且防止簇坍塌至关重要的应用中。

5. 讨论

本文介绍了设计条件式 elicitation（DCE），这是一种将从业者对聚类结构的信念转化为狄利克雷过程浓度参数的伽马超先验的框架。DCE 解决了现有实践中的三个鸿沟：实质性直觉与正式超参数之间的转换鸿沟、使原则性校准负担沉重的计算鸿沟，以及可能导致 K_J 校准先验产生非预期权重行为的连贯性鸿沟。DCE 将 elicitation 的重点从抽象参数 α 转向具体且依设计而定的问题——在这些 J 单位中，我们预计会出现多少种不同模式？——同时确保对权重影响进行诊断和控制，从而提供一套完整的工作流程，以实现有原则的先验设定。

5.1. 贡献总结。DCE 有三个特点使其有别于现有方法。首先，两阶段矩匹配（TSMM）以极低的额外开销实现机器精度的校准。先采用闭式初始化，再进行牛顿细化，可快速收敛；与基于网格的校准相比，速度约提升 $900\times$ 倍。这种高效性使有原则的先验设计成为应用分析中切实可行的一部分，而非仅仅是理想化的附加项。

其次，该框架明确揭示了通常隐含的一点：匹配 K_J 的目标分布，并不会自动产生对权重集中程度“中性”的先验。图 1 表明，为 $\mathbb{E}[K_{50}] \approx 5$ 校准的超先验可能意味着随机单位属于多数簇的概率接近 50%。通过要求提供诊断量——具体而言是 概率($w_1 > 0.5$)和 概率($w_1 > 0.9$)——并在需要时提供双锚点细化，DCE 确保权重影响清晰透明。

第三，模拟研究表明，默认超先验可能导致系统性误导的推断。在 Gamma(1,1) 下，无论真实簇数或信息性如何，坍塌概率都超过 60%——将信息性从 $I = 0.2$提高到 $I = 0.8$，也无法弥补设定不当的先验。TSMM 大幅降低了均方根误差（例如，$K^* = 30$从 28.04 降至 4.31），而双锚点则在主导风险较高的小型K^* 情景中提供了额外保护（将 $K^* = 5$的坍塌率从 48% 降至 10–14%）。这些结果支持一项实用结论：在中等J 问题中，模糊默认设定可能表现得像强汇聚

先验，而且提高数据信息量并不会自动消除这种影响。

5.2. **设计条件式 elicitation 与 SSD/SSI 之争**。 Vicentini 和 Jermyn 最近的一篇评论（2025年）指出了先验设定的两种方法之间的矛盾： α： 样本量依赖（*SSD*）方法会校准先验，以便在特定 J 下实现目标行为；而样本量独立（*SSI*）方法则将先验锚定在不随 J变化的量上，例如棍棒折断权重。

这项评论有其道理。当样本量发生变化（如在流式场景中），或多个数据集共享 α 上的共同先验（如在层次贝叶斯非参数模型中）时，SSD 先验会产生一系列依赖于设计背景的“先验序列”。Vicen- tini 和 Jermyn（2025年）认为，在这类背景下，SSI 视角——将先验锚定在权重摘要而非簇数上——提供了更具原则性的基础。

然而，本文的应用动机——多中心试验、荟萃分析和固定设计研究——具有一个共同的结构特征： J 由设计固定，而非随时间累积。在这类场景中，分析者通常关注有限总体估计目标，因此“当 $J \to \infty$时会发生什么”并不那么切题。正如 Vicentini 和 Jermyn（2025年）自己也承认的那样：“在 J 始终固定，且先验信念能够自然地用 K_J来表达的应用中，SSD 先验仍可能有其用武之地。”在这种情形下，通过 K_J 进行设计条件式 elicitation 并非病态依赖，而是对一个具有实质意义的问题作出明确编码：在这些 J 单位中，我们预期会出现多少种不同模式？

DCE 将这一固定的J 视角付诸实践，同时吸收了 SSI 批评的核心洞见：仅根据 K_J 进行校准可能导致意料之外的权重行为。我们的解决方案将 K_J 视为主要锚点，因为它在认知上更容易理解，随后将权重诊断作为一项强制检查。双锚点协议体现了一种综合方案：它保留了基于 K_J的可解释性，同时明确呈现划分视角与权重视角之间的权衡。这种综合方案由三项保障措施支持：任何经 K校准的先验都必须进行权重诊断（而非将其作为可选项）；当违反支配容差时，双锚点会生成明确的权衡报告；从业者还可以反转侧重点，将权重信念作为主要锚点，并将 K_J 的影响作为诊断量。

总而言之，SSD 和 SSI 最好理解为针对不同的推断目标——SSD 用于有限-J 设计条件下的问题，SSI 用于渐近或跨-J连贯性。对于促成本论文研究的固定-J 设定，基于 K_J 的 elicitation 配合

权重透明性提供了一套实用且站得住脚的工作流程。划分视角与权重视角之间的权衡并非需要掩盖的缺陷，而是一项应当报告的实质性选择。

5.3. **局限性与未来方向**。 有必要承认若干局限。首先，这些方法假设 α 服从伽马分布；划分层级的斯特林–伽马先验（Zito 等人，2024年）等替代结构在某些情境下可能更具优势。将 DCE 扩展至非伽马族是未来研究的自然方向。其次，该框架是为固定的J 情境而设计的；对于流式数据、序贯更新，或具有共享 α 的层级模型，则需要采用能让 SSI 视角更直接发挥作用的方法（Vicentini与Jermyn，2025年）。第三，校准先验的质量取决于所 elicited 的输入 $(\mu_K,\ \sigma_K^2)$。当从业者对聚类粒度缺乏直观判断时，建议针对合理的输入值进行敏感性分析，而 TS MM 的计算效率使此类分析切实可行（Giordano 等人，2023）。最后，尽管模拟研究聚焦于 $J = 100$ 及特定的数据生成条件，但在更极端的情境下，性能可能有所不同；OSM 的F 节提供了针对置信水平敏感性和替代基线的扩展模拟，不过从业者在外推时仍应谨慎判断。

若干扩展值得探究。将 DCE 扩展到其他贝叶斯非参数先验——尤其是 Pitman–Yor 过程（Pitman 与 Yor，1997）和层级 DP 模型——将拓宽其适用性。初步结果表明，双参数泊松分布–狄利克雷过程可获得类似的矩匹配校准（Ishwaran 与 Zarepour，2000）；将其扩展到 HDP，其中 α 控制顶层原子，γ 控制组层级聚类，这将服务于多层建模学界。此外，将 DCE 与诸如 `bayesplot` 和 `loo` 等贝叶斯工作流工具包相集成（Gelman 等，2020；Roos 等，2015），将降低有原则的先验设定门槛，并促进端到端可复现性。

5.4. **结论**。 随着贝叶斯非参数方法从一种小众统计工具逐渐成熟为社会科学中用于灵活推断的标准工具，先验设定的“黑箱”必须被打开。问题不在于是否 使用信息性先验——因为在低到中等信息量情境下，所有先验都具有信息性——而在于如何使这些先验透明、可解释，并与研究设计保持一致（Richardson & Green，1997）。设计条件式 elicitation 框架提供了必要的工具：从业者可以用簇

粒度的语言阐明信念，以机器精度将这些信念转换为经过校准的超参数，诊断非预期的权重影响，并报告完整的先验设定，从而支持审查与重复试验。配套的 DPprior R 包通过可复现的诊断和报告模板将这一工作流程付诸实践（有关安装、API 参考和计算示例，请参见 OSM 中的附录 G）。在利用狄利克雷过程灵活性的同时，不放弃对底层假设的控制，研究者便能推动教育与行为研究中的贝叶斯非参数建模走向更加严谨且可复现的实践。

参考文献

Abramowitz, M., & Stegun, I. A. (1964).《数学函数手册：含公式、图表和数学表格》。美国国家标准局。 [36,60, 69]。Antonelli, J., Trippa, L., & Haneuse, S. (2016). 缓解广义线性混合模型中的偏差：贝叶斯非参数方法的理由。《统计科学》，31(1), 80–95. https://doi.org/10.1214/15-STS533 [2, 4]。Antoniak, C. E. (1974). 狄利克雷过程混合及其在贝叶斯非参数问题中的应用。《统计学年鉴》， 2(6), 1152–1174. https : //doi.org/10.1214/aos/1176342871 [7, 11, 13, 32, 35, 42]。Arratia, R., Barbour, A. D., & Tavaré, S. (2000). 对数组合结构中的成分数量。《应用概率年刊》， 10(2), 331–361. https://doi.org/10.1214/aoap/1019487347 [12, 34, 38, 41, 42]。Arratia, R., Barbour, A. D., & Tavaré, S. (2003). 对数组合结构：概率方法。欧洲数学会。https://doi. org/10.4171/000 [38, 47, 59]。Barbour, A. D., & Hall, P. (1984). 关于泊松收敛率。《剑桥哲学学会数学汇刊》，95(3), 473–480. https : //doi.org/10.1017/S0305004100061806 [59]。Barbour, A. D., Holst, L., & Janson, S. (1992). 泊松近似。牛津大学出版社。https://doi.org/10.1093/oso/9780198522355.001.0001 [59]。Barbour, A. D., & Xia, A. (1999). 泊松扰动。《ESAIM：概率与统计》， 3, 131–150. https://doi.org/10.1051/ps:1999106 [42]。

Blackwell, D., & MacQueen, J. B. (1973). 通过 Pólya urn方案构造 Ferguson分布。《统计学年鉴》，1 (2), 353–355. https : / / doi . org / 10 . 1214 / aos / 1176342372 [32, 33]。Bonafos, G., Bourot, C., Pudlo, P., Freyermuth, J.-M., Reboul, L., Tronçon, S., & Rey, A. (2024). 基于拓扑增强信号表示的狄利克雷过程混合模型，用于婴儿发声聚类。 [2, 3]。Burr, D., & Doss, H. (2005). 随机效应荟萃分析的贝叶斯半参数模型。《美国统计学会杂志》，100(469), 242–251. https://doi.org/10.1198/016214504000001024 [2]。Cao, W., Chu, H., Hanson, T., & Siegel, L. (2024). 用于估计参考区间的贝叶斯非参数荟萃分析模型。《医学统计学》，43(10), 1905–1919. https://doi.org/10.1002/sim.10001 [2, 3, 6]。Cardoso, P., Dennis, J. M., Bowden, J., Shields, B. M., McKinley, T. J., & MASTER MIND Consortium. (2024). 狄利克雷过程混合模型在反事实预测模型中插补缺失的预测变量数据：预测最佳2型糖尿病治疗的应用。BMC医学信息学与决策制定，24(1), 12. https : / / doi . org / 10 . 1186 / s12911 - 023 - 02400 - 3[2]。Casella, G., & Berger, R. L. (2002). 统计推断（第2版）。Duxbury。 [49]。Congdon, P. D. (2019). 贝叶斯层次模型：使用R的应用（第2版）。Chapman; Hall/CRC. [7]。Daniels, M. J., Linero, A., & Roy, J. (2023). 用于因果推断和缺失数据的贝叶斯非参数方法。Chapman; Hall/CRC. [2]。Dennis, J. E., Jr., & Schnabel, R. B. (1996). 无约束优化和非线性方程的数值方法 [最初由 Prentice-Hall 出版，1983]。SIAM。https://doi.org/10.1137/1.9781611971200 [49, 53]。Dorazio, R. M. (2009). 为狄利克雷过程混合模型的精度参数选择先验。《统计规划与推断杂志》，139(9), 3384–3390. https://doi.org/10.1016/j.jspi.2009.03.009 [2–4, 11, 35, 37, 56, 79,84]。Dorazio, R. M., Mukherjee, B., Zhang, L., Ghosh, M., Jelks, H. L., & Jordan, F. (2008). 使用狄利克雷过程先验对动物丰度中的未观测异质性来源建模。《生物统计学》，64(2), 635–644. https://doi.org/ 10.1111/j.1541-0420.2007.00873.x [4]。

Dunson, D. B., Xue, Y., & Carin, L. (2008). 矩阵棍折过程：灵活的贝叶斯荟萃分析。《美国统计学会杂志》， 103 (481), 317–327. https://doi.org/10.1198/016214507000001364 [2]. Escobar, M. D., & West, M. (1995). 使用混合分布进行贝叶斯密度估计与推断。《美国统计学会杂志》， 90 (430), 577–588. https://doi.org/10.1080/01621459.1995.10476550 [6, 7]. Ferguson, T. S. (1973). 一些非参数问题的贝叶斯分析。《统计学年鉴》， 1(2), 209–230. https://doi.org/10.1214/aos/1176342360[6, 33]. Finch, H., & Edwards, J. M. (2016). 在潜在特质非正态的情况下，使用非参数贝叶斯方法估计 Rasch 模型参数。《教育与心理测量》， 76(4), 662–684. https://doi.org/ 10.1177/0013164415608418 [2]. Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). 贝叶斯工作流程。arXiv 预印本 arXiv:2011.01808. https://doi.org/10.48550/arXiv.2011.01808 [24]. Gill, J., & Casella, G. (2009). 有序贝叶斯社会科学模型的非参数先验：设定与估计。《美国统计学会杂志》， 104(486), 453–464. https://doi.org/10.1198/jasa.2009.0039 [2]. Giordano, R., Liu, R., Jordan, M. I., & Broderick, T. (2023). 评估贝叶斯非参数方法中对棍折先验的敏感性（附讨论）。《贝叶斯分析》， 18(1), 287–366. https://doi.org/10.1214/22-BA1309 [2, 4,24]. Greve, J., Grün, B., Malsiner-Walli, G., & Frühwirth-Schnatter, S. (2022). 在贝叶斯聚类分析中探究数据簇数量的先验及划分分布。《澳大利亚与新西兰统计学杂志》， 64(2), 205–229. https://doi.org/10.1111/anzs.12350 [2, 3]. Hanson, T. E., Branscum, A. J., & Johnson, W. O. (2005). 贝叶斯非参数建模与数据分析：导论。载于《统计学手册》（第25卷，第245–278页）。爱思唯尔. [6]. Ishwaran, H., & Zarepour, M. (2000). 近似狄利克雷过程和贝塔二参数过程层级模型中的马尔可夫链蒙特卡罗。《生物统计学》， 87(2), 371–390. https://doi.org/10.1093/biomet/87.2.371 [6, 24].

Iwashige, F., & Hashimoto, S. (2025). 使用带有归一化逆高斯权重的有限混合的混合进行贝叶斯混合建模。arXiv 预印本 arXiv:2501.*18854* [2]. Jackson, D., & White, I. R. (2018). 荟萃分析何时应避免作出隐含正态性假设？《生物统计学杂志》，60(6), 1040–1058. https://doi. org/10.1002/bimj.201800071 [2]. Johnson, N. L., Kotz, S., & Kemp, A. W. (2005).《一元离散分布》（第3版）。Wiley. https://doi.org/10.1002/0471715816 [44]. Kingman, J. F. C. (1975). 随机离散分布。《英国皇家学会哲学汇刊B辑（方法论）》， 37(1), 1–22. https://doi.org/10.1111/j.2517-6161.1975.tb01024.x [68]. Kottas, A. (2006). 贝塔分布的狄利克雷过程混合及其在密度估计和强度估计中的应用。非参数贝叶斯方法学习研讨会，第23届国际机器学习会议（ICML） [3]. Kyung, M., Gill, J., & Casella, G. (2010). 狄利克雷随机效应模型中的估计。《统计学年鉴》， 38 (2), 979–1009. https://doi.org/10.1214/ 09-AOS731 [2]. Le Cam, L. (1960). 泊松二项分布的一个逼近定理。《太平洋数学杂志》， 10(4), 1181–1197. https://doi.org/10.2140/pjm.1960.10.1181 [38, 41, 42, 47, 59]. Lee, J., Che, J., Rabe-Hesketh, S., Feller, A., & Miratrix, L. (2025). 改进多中心试验中中心特异性效应及其分布的估计。《教育与行为统计学杂志》， 50(5), 731–764. https://doi.org/ 10.3102/10769986241254286 [2, 3, 76, 78, 79, 82, 84]. Leslie, D. S., Kohn, R., & Nott, D. J. (2007). 异方差线性回归的一种通用方法。《统计与计算》， 17 (2), 131–146 [3]. Ma, Z., & Chen, G. (2020). 使用狄利克雷过程先验进行联合分析的贝叶斯半参数潜变量模型：采用 NIMBLE 实现。《统计建模》， 20(1), 71–95 [3]. Matza, A., & Bistritz, Y. (2021). 改进簇数估计的无限高斯混合建模。AAAI 人工智能会议论文集， 35(10), 8921–8929 [2]. McCullagh, P., & Yang, J. (2008). 有多少个簇？《贝叶斯分析》， 3(1), 101– 120. https://doi.org/10.1214/08-BA304 [3].

Mignemi, G., & Manolopoulou, I. (2025). 多位评分者的贝叶斯非参数模型：一般统计框架 [在线优先发表]。《心理测量学》。https://doi.org/10.1017/psy.2025.10035 [2, 3]. Miller, J. W., & Harrison, M. T. (2018). 对成分数量设定先验的混合模型。《美国统计学会杂志》，113 (521), 340–356。https://doi.org/10.1080/01621459.2016.1255636 [7, 12]. Murugiah, S., & Sweeting, T. (2012). 狄利克雷过程混合模型中精度参数先验的选择。《统计规划与推断杂志》，142 (7), 1947–1959。https : / / doi . org / 10 . 1016 / j . jspi . 2012 . 02 . 013 [2, 3, 12,35, 37]. Muthukumarana, S., & Tiwari, R. C. (2016). 使用狄利克雷过程进行荟萃分析。《医学研究中的统计方法》，25 (1), 352–365。https://doi.org/10. 1177/0962280212453891 [2]. Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). 使用狄利克雷过程建模个体差异。《数学心理学杂志》，50 (2), 101–122。https://doi.org/10.1016/j.jmp.2005.11.006 [2, 7]. Neal, R. M. (2000). 狄利克雷过程混合模型的马尔可夫链抽样方法。《计算与图形统计学杂志》，9 (2), 249–265。https://doi.org/10.1080/10618600.2000.10474879 [3]. Nocedal, J., & Wright, S. J. (2006). 数值优化（第2版）。Springer。https: //doi.org/10.1007/978-0-387-40065-5 [49, 73]. Ortega, J. M., & Rheinboldt, W. C. (1970). 多变量非线性方程的迭代求解 [由 SIAM 再版，2000]。Academic Press。https://doi. org/10.1137/1.9780898719468 [49, 53]. Paganin, S., Paciorek, C. J., Wehrhahn, C., Rodríguez, A., Rabe-Hesketh, S., & de Valpine, P. (2023). 贝叶斯半参数项目反应理论模型的计算策略与估计性能。《教育与行为统计学杂志》，48 (2), 147–188。https://doi.org/10.3102/10769986221136105 [2, 3]. Pitman, J. (1996). 按大小加权置换下不变的随机离散分布。《应用概率进展》，28 (2), 525–539。https://doi.org/10. 2307/1428070 [32, 64, 68]. Pitman, J. (2006). 组合随机过程（第1875卷）。Springer。https: //doi.org/10.1007/b11601500 [33, 63, 64, 68].

Pitman, J., & Yor, M. (1997). 从稳定子协调过程导出的双参数泊松-狄利克雷分布。《概率年刊》，25(2), 855–900。https://doi.org/10.1214/aop/1024404422 [24]. Qiu, M., Paganin, S., Ohn, I., & Lin, L. (2025). 不同题目类型的贝叶斯非参数潜在类别分析 [在线优先发表]。《心理学方法》。https://doi.org/10.1037/met0000728 [2]. Richardson, S., & Green, P. J. (1997). 关于成分数量未知时混合模型的贝叶斯分析（附讨论）。《英国皇家统计学会杂志B辑：统计方法学》，59(4), 731–792。https://doi. org/10.1111/1467-9868.00095 [24]. Roos, M., Martins, T. G., Held, L., & Rue, H. (2015). 贝叶斯层次模型的敏感性分析。《贝叶斯分析》，10 (2), 321–349。https://doi.org/10. 1214/14-BA909 [24]. Rubin, D. B. (1981). 平行随机试验中的估计。《教育统计学杂志》，6 (4), 377–401。https://doi.org/10.3102/10769986006004377[77]. Sethuraman, J. (1994). 狄利克雷先验的构造性定义。《中华统计学志》，4 (2), 639–650 [6, 7, 34, 63, 64]. Vicentini, C., & Jermyn, I. H. (2025). 狄利克雷过程混合的精度参数先验选择。https://arxiv.org/abs/2502.00864 [4, 7, 9, 10, 12, 13, 23, 24, 34, 39, 41, 44, 63, 66, 69, 70, 76]. Watterson, G. A. (1974). 选择性中性等位基因的抽样理论。《应用概率进展》，6 (3), 463–488。https://doi.org/10.2307/1426228 [34]. Zito, A., Rigon, T., & Dunson, D. B. (2024). 通过斯特林-伽马先验对潜在划分进行贝叶斯非参数建模 [提前发表]。《贝叶斯分析》，1–28。https://doi.org/10.1214/24-BA1463 [24, 39, 41, 44].

在线补充材料

狄利克雷过程混合的设计条件先验 elicitation：簇数与权重控制的统一框架

JoonHo Lee

目录

在线补充材料 32 附录 A. 狄利克雷过程/中国餐馆过程预备知识与精确 K_J 分布 32 A.1. 目的与范围 32 A.2. 分布理论 33 附录 B. 第一阶段闭式推导与近似理论 41 B.1. 目的与范围 41 B.2. 平移泊松–伽马代理 41 B.3. A1 闭式反演 44 B.4. 置信度到方差的映射 46 B.5. 近似质量：何时第一阶段已足够？ 47 附录 C. 第二阶段牛顿细化：完整算法与收敛性 49 C.1. 目的与范围 49 C.2. 问题陈述与目标方程 50 C.3. 基于得分恒等式的雅可比推导 50 C.4. 算法 C.1：A2-MN（精确矩牛顿法细化）52 C.5. 收敛性：哪些有保证（哪些没有）53 C.6. 计算示例与典型迭代次数 54 C.7. 可选：A2-KL 细化（ K_J上的分布目标）55 附录 D. 近似误差界与有限-J 分析 56 D.1. 目的与范围 56 D.2. 三路误差分解 56 D.3. 条件（固定-α）界 58 D.4. 边缘（伽马混合）界 61 D.5. 何时需要 A2？实用阈值条件 62 附录 E. 权重分布理论与双锚框架 63 E.1. 目的与范围 63

E.2. 棍棒折断权重与大小偏倚解释 63 E.3. w_1 65
的闭式边缘分布 E.4. 共同聚类概率
ρ 68 E.5. 双锚定校准框架 70 E.6.
面向从业者的解释指南 74 附录 F. 扩展模拟研究结果 76 F.1 目的与范围 76
F.2. 全因子设计规范 77 F.3. 超先验方法与精确参数 79 F.4. 推断模型与马尔可夫
链蒙特卡洛配置 80 F.5. 结果指标 81 F.6. 中等置信度下的完整结果（主要设置）
81 F.7. 额外基线： χ^2-DORO 82 F.8. 敏感性分析：置信度设定（TSMM / 双锚
点）82 F.9. 浓度参数 α 83的后验行为 F.10.
K^* 83的基于众数的恢复 F.11. 均方
误差的偏差–方差分解（低信息设置）83 F.12. 计算基准测试（校准成本）84 附
录 G. 软件实现 85 G.1. 目的与范围 85 G.2. 软件包概述 86 G.3. 映射：论文组成
部分与 DPprior 函数 87 G.4. 核心工作流程 88 G.5. 计算示例：端到端信息
elicitation 90 G.6. 可复现性清单 97 G.7. 代码库与资源 98

附录A. 狄利克雷过程/中国餐馆过程预备知识与精确的 K_J 分布

A.1. **目的与范围**。 本附录为设计条件 elicitation（DCE）框架奠定作为参考标准的计算基础。此处给出的结果——尽管大多是 DP/CRP文献中的经典内容（例如，Antoniak，1974；Blackwell 和 MacQueen，1973年；Pitman，1996年）——经过综合整理，旨在提供：(i) 用于近似验证的真实基准；以及 (ii) 两阶段矩匹配（TSMM）算法的稳定评估层。

全文中，我们沿用与正文 §2.1 一致的符号：

- J：固定设计规模（例如多中心试验或荟萃分析中的站点/研究/单位数量）。
- $\alpha > 0$：狄利克雷过程浓度参数。
- K_J：J 次狄利克雷过程抽样中的已占据簇（不同取值）数量。
- 超先验：$\alpha \sim$ 伽马分布(a, b)，采用形状–率参数化：

$$p(\alpha \mid a, b) = \frac{b^a}{\Gamma(a)} \alpha^{a-1} e^{-b\alpha}, \qquad a > 0,\ b > 0. \tag{A.1}$$

A.2.1–A.2.3 节确立了 $K_J \mid \alpha$ 的精确条件分布及其矩。A.2.4–A.2.6 节介绍了伽马混合（边缘）分布、为阶段 1 提供依据的渐近近似，以及数值稳定计算。

A.2. 分布理论。

A.2.1. 中国餐馆过程与划分结构。 设 $G \sim$ 狄利克雷过程(α, G_0) 其基测度 G_0非原子。抽取 $\theta_1, \ldots, \theta_J \mid G \overset{iid}{\sim} G$。由于狄利克雷过程的抽样 G 几乎必然是离散的（Ferguson，1973年），因此 (θ_j) 中会出现相同值，从而在 $\{1, \ldots, J\}$上诱导出可交换随机划分。定义

$$K_J := \#\{\theta_1, \ldots, \theta_J\}. \tag{A.2}$$

一种便于理解的构造性描述是中国餐馆过程（CRP；Blackwell 和 MacQueen，1973年；Pitman，2006年）。将抽样 θ_j 解释为依次到来的“顾客”：

- 顾客 1 总是开启一张新桌子。
- 给定 $i-1$ 位顾客就座后的就座安排，顾客 i 可以坐在一张已有桌子 k 旁，其概率与该桌当前人数成正比 $n_{(i-1)\,k}$；也可以开启一张新桌子，其概率与 α 成正比：

$$\Pr(z_i = k \mid z_{1:i-1}, \alpha) = \begin{cases} \dfrac{n_k^{(i-1)}}{\alpha + i - 1}, & k = 1, \ldots, K_{i-1}, \\[2ex] \dfrac{\alpha}{\alpha + i - 1}, & k = K_{i-1} + 1. \end{cases} \tag{A.3}$$

这里，z_i 是顾客 i对应的桌子标签，K_{i-1} 是 $i-1$ 位顾客就座后已占用桌子的数量，$n_k^{(i-1)}$ 是坐在桌子 k旁的人数。

新桌子的概率 $\alpha/(\alpha + i - 1)$ 仅取决于 α 和步长索引 i，由此立即得到 $K_J \mid \alpha$ (定理 A.1) 中的伯努利和表示。

这种类调和结构解释了为什么精确条件矩中会出现双伽马函数和三伽马函数（命题 A.1 ）。

与截断棒表示的联系。 CRP 分区律等价于截断棒表示（Sethuraman, 1994）：

$$G = \sum_{h=1}^{\infty} w_h \, \delta_{\theta_h^*}, \quad w_h = v_h \prod_{\ell<h}(1-v_\ell), \quad v_h \overset{iid}{\sim} \mathrm{Beta}(1,\alpha). \tag{A.4}$$

权重 $(w_1,\ w_2,\ \ldots\ldots) \sim \mathrm{GEM}(\alpha)$ 表示渐近相对簇大小。在 J 次从 G 抽取所诱导的随机分区具有 CRP 预测规则（A.3），即在对(w_h)边缘化之后。

A.2.2. 两种关于 $K_J \mid \alpha$ 的精确表示。 我们给出两种等价描述：（i）适用于矩与集中不等式的泊松–二项（独立伯努利变量之和）表示；（ii）适用于精确概率计算的斯特林数概率质量函数（Antoniak 分布）。

定理 A.1 （泊松–二项表示）。 在 α条件下，

$$K_J \overset{d}{=} \sum_{i=1}^{J} B_i, \qquad B_i \ \textit{independent}, \quad B_i \sim \mathrm{Bernoulli}\left(\frac{\alpha}{\alpha+i-1}\right). \tag{A.5}$$

等价地，$K_J - 1$服从成功概率为 $\{\alpha/(\alpha+i-1)\}_{i=2}^{J}$ 的泊松–二项分布。

出处说明。 这一表示在 Ewens/CRP 文献中由来已久；关于其在先验 elicitation 中的应用，参见 Watterson（1974）、Arratia等人（2000）以及 Vicentini 和 Jermyn（2025年）。

证明。 定义新表指示变量

$$I_i := \mathbf{1}\{\text{customer } i \text{ starts a new table}\}, \qquad i = 1, \ldots, J, \tag{A.6}$$

使得 $K_J = \sum_{i=1}^{J} I_i$ 且 $I_1 \equiv 1$。令 $\mathcal{F}_{i-1}$ 为由截至步长 $i-1$的就座安排生成的 σ-域。从中国餐馆过程规则（A.3）可知，对于 $i \geq 2$,

$$\Pr(I_i = 1 \mid \mathcal{F}_{i-1}, \alpha) = \frac{\alpha}{\alpha+i-1} =: p_i(\alpha), \tag{A.7}$$

这不依赖于 $\mathcal{F}_{i-1}$。因此，给定 α 时，I_i 与 $\mathcal{F}_{i-1}$ 条件独立。由归纳法可知，给定 α 时，I_1、……、I_J 相互独立，其中对于 $i \geq 2$ 和 $I_1 \equiv 1$，$I_i \sim$ 服从伯努利分布（p_i（α））。□

边界行为。 根据这一表示：

- 当 $\alpha \to 0$: p_i（α）$\to 0$ 对于 $i \geq 2$，因此 $K_J \mid \alpha \Rightarrow \delta_1$（所有单位都属于同一个簇）。

- 当 $\alpha \to \infty$：$p_i(\alpha) \to 1$，因此 $K_J \mid \alpha \Rightarrow \delta_J$（每个单位各自形成一个簇）。

定理 A.2（Antoniak 分布）。对于 $k = 1, \ldots, J$，

$$\Pr(K_J = k \mid \alpha) = |s(J,k)| \cdot \alpha^k \cdot \frac{\Gamma(\alpha)}{\Gamma(\alpha + J)} = |s(J,k)| \cdot \frac{\alpha^k}{(\alpha)_J}, \tag{A.8}$$

其中 $|s(J, k)|$ 是第一类无符号斯特林数，而 $(\alpha)_J := \alpha(\alpha+1)\cdots(\alpha+J-1)$ 是上升阶乘。

出处说明。公式（A.8）是经典的 Antoniak 分布（Antoniak, 1974）。它也出现在狄利克雷过程混合模型的先验 elicitation 工作中，是其中的计算瓶颈（Dorazio, 2009年；Murugiah & Sweeting, 2012）。

无符号斯特林数。组合量 $|s(J, k)|$ 表示由 J 个元素构成且恰有 k 个循环的排列数。它们满足递推关系：

$$|s(J,k)| = |s(J-1,k-1)| + (J-1)\cdot|s(J-1,k)|, \tag{A.9}$$

边界条件为 $|s(0, 0)| = 1$，$|s(J, 0)| = 0$，当 $J \geq 1$ 时，以及 $|s(J, J)| = 1$。

证明。令 $p_{n,k}(\alpha) :=$ 概率$(K_n = k \mid \alpha)$。根据中国餐馆过程，加入第 n 位顾客时，要么创建一张新桌子（概率为 $\alpha/(\alpha+n-1)$），桌子数量增加 1；要么选择一张已有桌子（概率为$(n-1)/(\alpha+n-1)$），桌子数量保持不变。因此，对于 $n \geq 2$ 和 $1 \leq k \leq n$，

$$p_{n,k}(\alpha) = p_{n-1,k-1}(\alpha) \cdot \frac{\alpha}{\alpha+n-1} + p_{n-1,k}(\alpha) \cdot \frac{n-1}{\alpha+n-1}. \tag{A.10}$$

定义候选解$\tilde{p}_{n,k}(\alpha) := |s(n,k)| \cdot \alpha^k \cdot \Gamma(\alpha)/\Gamma(\alpha+n)$。利用 $\Gamma(\alpha+n) = (\alpha+n-1)\Gamma(\alpha+n-1)$和斯特林递推关系（A.9），可直接验证$\tilde{p}_{n,k}(\alpha)$满足相同的递推关系（A.10）。由于 $K_1 \equiv 1$，因此$p_{1,1}(\alpha) = 1$，且$\tilde{p}_{1,1}(\alpha) = |s(1,1)| \cdot \alpha \cdot \Gamma(\alpha)/\Gamma(\alpha+1) = 1$，故基例成立。由归纳法可得，对于所有 n 和 k，$p_{n,k}(\alpha) = \tilde{p}_{n,k}(\alpha)$。□

A.2.3. 精确条件矩。定义

$$\kappa_J(\alpha) := \mathbb{E}[K_J \mid \alpha], \qquad v_J(\alpha) := \mathrm{Var}(K_J \mid \alpha). \tag{A.11}$$

命题 A.1（条件矩）。对于 $\alpha > 0$ 和 $J \geq 1$，

$$\kappa_J(\alpha) = \alpha\big[\psi(\alpha+J) - \psi(\alpha)\big], \tag{A.12}$$

$$v_J(\alpha) = \kappa_J(\alpha) - \alpha^2\big[\psi_1(\alpha) - \psi_1(\alpha+J)\big], \tag{A.13}$$

其中 $\psi(\cdot)$ 是双伽马函数，$\psi_1(\cdot)$ 是三伽马函数。

证明。（逐步推导。）根据 定理 A.1，$K_J = \sum_{i=1}^{J} I_i$ 其中 $\mathbb{E}[I_i \mid \alpha] = p_i(\alpha) = \alpha/(\alpha+i-1)$。因此

$$\kappa_J(\alpha) = \sum_{i=1}^{J} \frac{\alpha}{\alpha+i-1} = \alpha \sum_{r=0}^{J-1} \frac{1}{\alpha+r}. \tag{A.14}$$

利用恒等式 $\sum_{r=0}^{J-1}(\alpha+r)^{-1} = \psi(\alpha+J) - \psi(\alpha)$ (Abramowitz & Stegun, 1964, §6.3)，我们得到（A.12）。

对于方差，由条件独立性可得

$$\mathrm{Var}(K_J \mid \alpha) = \sum_{i=1}^{J} p_i(\alpha)\{1-p_i(\alpha)\} = \sum_{r=0}^{J-1} \frac{\alpha}{\alpha+r}\left(1 - \frac{\alpha}{\alpha+r}\right) = \sum_{r=0}^{J-1} \frac{\alpha\, r}{(\alpha+r)^2}. \tag{A.15}$$

将每一项改写为

$$\frac{\alpha\, r}{(\alpha+r)^2} = \frac{\alpha}{\alpha+r} - \frac{\alpha^2}{(\alpha+r)^2}, \tag{A.16}$$

我们得到

$$v_J(\alpha) = \alpha \sum_{r=0}^{J-1} \frac{1}{\alpha+r} - \alpha^2 \sum_{r=0}^{J-1} \frac{1}{(\alpha+r)^2} = \kappa_J(\alpha) - \alpha^2\big[\psi_1(\alpha) - \psi_1(\alpha+J)\big], \tag{A.17}$$

利用三伽马函数恒等式 $\sum_{r=0}^{J-1}(\alpha+r)^{-2} = \psi_1(\alpha) - \psi_1(\alpha+J)$。 □

推论 A.1（条件欠离散）。对于 $\alpha > 0$ 和 $J \geq 2$，$0 < v_J(\alpha) < \kappa_J(\alpha)$。

证明。对于 $i \geq 2$, $p_i(\alpha) \in (0,1)$，因此 $p_i(1-p_i) < p_i$。此外，$p_1(1-p_1) = 0$，因为 $p_1 = 1$。因此，$v_J(\alpha) = \sum_{i=1}^{J} p_i(1-p_i) < \sum_{i=1}^{J} p_i = \kappa_J(\alpha)$，并且由于当 $J \geq 2$ 时至少有一个 $p_i \in (0,1)$，故其严格为正。□

这种欠离散性特征——条件方差严格小于条件均值——使 $K_J \mid \alpha$ 区别于泊松分布，并对 elicitation 中可接受的目标矩施加限制（见 B 节）。

关于 α 的导数。 C 节中的牛顿细化 使用条件矩的导数。记 $r = i-1 \in \{0,\dots,J-1\}$：

均值导数：

$$\frac{\partial \kappa_J}{\partial \alpha} = \sum_{r=1}^{J-1} \frac{r}{(\alpha+r)^2} = \big[\psi(\alpha+J) - \psi(\alpha)\big] + \alpha\big[\psi_1(\alpha+J) - \psi_1(\alpha)\big]. \tag{A.18}$$

方差导数（求和形式）：

$$\frac{\partial v_J}{\partial \alpha} = \sum_{r=1}^{J-1} \frac{r(r-\alpha)}{(\alpha+r)^3}. \tag{A.19}$$

方差导数（特殊函数形式）：直接对A.13求导，可得到一个使用四伽马函数的等价表达式 $\psi_2(\cdot)$：

$$\frac{\partial v_J}{\partial \alpha} = \frac{\partial \kappa_J}{\partial \alpha} - 2\alpha\big[\psi_1(\alpha) - \psi_1(\alpha+J)\big] - \alpha^2\big[\psi_2(\alpha) - \psi_2(\alpha+J)\big]. \tag{A.20}$$

计算说明。通过特殊函数计算 $\kappa_J(\alpha)$和 $v_J(\alpha)$（见A.12– (A.13）所需的计算量为 $O(1)$（每次 α）；而直接求和（见A.14–(A.15）所需的计算量为 $O(J)$。在目标情形 $J \lesssim 300$下，两种方法都可行；为提高速度和数值光滑性，我们更倾向于使用特殊函数形式。

A.2.4. 伽马超先验下的边缘分布。设 $\alpha \sim$ 伽马分布(a, b)的密度为（A.1）。将 Antoniak 概率质量函数（A.8）对 α 积分，可得 K_J的先验预测分布。

定理 A.3（边缘概率质量函数）。对于 $k = 1, \ldots, J$,

$$p(K_J = k \mid a, b) = \frac{b^a \, |s(J,k)|}{\Gamma(a)} \int_0^\infty \alpha^{k+a-1} e^{-b\alpha} \frac{\Gamma(\alpha)}{\Gamma(\alpha+J)} \, d\alpha. \tag{A.21}$$

一般而言，(A.21) 中的积分没有闭式解，必须通过数值计算求值。

出处说明。 (A.21) 中的“Stirling–Gamma”积分是 Dorazio（2009年）指出的计算瓶颈，影响着 α 超先验的选择。由于该积分没有闭式表达式，因此需要渐近近似（A.2.5 节）来提供初始值，还需要精确求积层（A.2.6 节）作为验证基准。Murugiah 和 Sweeting（2012）给出了簇数的等价 Gamma混合边缘分布，其中涉及一维积分。

证明。 根据全概率公式， $p(K_J = k \mid a, b) = \int p(K_J = k \mid \alpha)\, p(\alpha \mid a, b)\, d\alpha$。代入（A.8）和（A.1），并合并各项，得到（A.21）。□

命题A.2 (利用全期望/全方差定律计算边缘矩)。 在 $\alpha \sim$伽马分布(a, b),

$$\mathbb{E}[K_J \mid a, b] = \mathbb{E}_\alpha[\kappa_J(\alpha)], \tag{A.22}$$

$$\mathrm{Var}(K_J \mid a, b) = \mathbb{E}_\alpha[v_J(\alpha)] + \mathrm{Var}_\alpha(\kappa_J(\alpha)). \tag{A.23}$$

证明。 公式（A.22）是全期望定律。对于（A.23），应用全方差定律：方差(K_J) $=\mathbb{E}\{$方差$(K_J \mid \alpha)\}+$方差$\{\mathbb{E}(K_J \mid \alpha)\}$。 □

分解命题A.2 在计算上具有优势：相比于从完整的概率质量函数（PMF）计算边缘矩（这要求对每个 k进行斯特林–伽马积分），可以通过对 κ_J（α）和 v_J（α）进行一维数值积分来计算边缘矩。这正是 TSMM 阶段 2 牛顿迭代所采用的方法。

A.2.5.渐近近似。 本节记录了两个近似结果，它们为B 节中的阶段 1 闭式初始值提供了依据。它们并不用作最终答案——阶段 2 会以精确的有限J 矩匹配取代这些近似。

定理 A.4 （泊松近似）。 设 $\lambda_J(\alpha)$: $=\kappa_J(\alpha)$ $=\sum_{i=1}^{J}\alpha/(\alpha+i-1)$。则，当 $J\to\infty$ 时，

$$d_{\mathrm{TV}}(\mathcal{L}(K_J \mid \alpha),\ \mathrm{Poisson}(\lambda_J(\alpha)))=O\left(\frac{1}{\log J}\right),\quad \textit{for fixed } \alpha>0. \tag{A.24}$$

出处说明。 对数型组合结构的全变差泊松近似由 Arratia等人（2000）发展而来。

证明。 根据 定理 A.1， $K_J \mid \alpha$ 是由概率为 $p_i(\alpha)=\alpha/(\alpha+i-1)$的独立伯努利变量构成的泊松–二项和。标准的泊松近似界（Le Cam，1960年）给出

$$d_{\mathrm{TV}}(\mathcal{L}(K_J \mid \alpha),\ \mathrm{Poisson}(\lambda_J(\alpha)))\le\frac{\sum_{i=1}^{J}p_i(\alpha)^2}{\lambda_J(\alpha)}. \tag{A.25}$$

对于分子，注意到 $p_1(\alpha)=1$，且对于 $i\ge 2$，其中 $r=i-1\ge 1$ ，

$$p_i(\alpha)^2=\left(\frac{\alpha}{\alpha+r}\right)^2\le\frac{\alpha^2}{r^2},\quad\Rightarrow\quad\sum_{i=1}^{J}p_i(\alpha)^2\le 1+\alpha^2\sum_{r=1}^{\infty}\frac{1}{r^2}<\infty. \tag{A.26}$$

同时， $\lambda_J(\alpha)=\alpha\sum_{r=0}^{J-1}(\alpha+r)^{-1}\sim\alpha\log J$，根据 (A.12) 以及渐近关系 $\psi(\alpha+J)$ $=\log J+O(1)$。因此， $\lambda_J(\alpha)$ 以 $\alpha\log J$ 的速度发散，而 $\sum_i p_i^2=O(1)$，且 (A.25) 右侧为 $O(1/\log J)$，由此证明 (A.24)。□

注 A.1。 Arratia等人（2003）强调，使用精确均值 κ_J（α）而非渐近近似 α 对数 J，对于有限的 J，可以提高准确性，因为 κ_J（α） $=\alpha$ 对数 $J+O$（1），其中常数依赖于 α。

定理 A.5 （负二项近似）。 设 $\alpha \sim$ 伽马分布$(a,\ b)$ 并定义 $c_J :=$对数 J。当 J 较大时，

$$K_J \mid (a,b) \overset{d}{\approx} 1 + \mathrm{NegBin}\left(a,\ \frac{b}{b+c_J}\right), \tag{A.27}$$

其中 负二项分布(r,p) 的均值为 $r(1-p)/p$，方差为 $r(1-p)/p^2$。

出处说明。 随机 J 下的大样本负二项分布极限 α 见 Vicentini 和 Jermyn（2025年）以及 Zito 等人（2024年）。

证明。 根据 定理 A.4，对于每个固定的 α，$K_J - 1 \mid \alpha$ 在全变差意义下可由 泊松分布(αc_J) 很好地近似，当 J较大时。如果 $Y \mid \alpha \sim$泊松分布(αc_J) 且 $\alpha \sim$伽马分布(a,b)，则伽马–泊松混合恰好服从负二项分布：

$$Y \sim \mathrm{NegBin}\left(a,\ \frac{b}{b+c_J}\right). \tag{A.28}$$

事实上，将泊松概率质量函数对伽马密度积分可得 概率$(Y=y)$ $=\binom{a+y-1}{y} p^a (1-p)^y$，其中 $p = b/(b+c_J)$。结合 $K_J \approx 1 + Y$即可得到（A.27）。

□

近似边际矩。 根据（A.27）：

$$\mathbb{E}[K_J \mid a,b] \approx 1 + \frac{a \log J}{b}, \qquad \mathrm{Var}(K_J \mid a,b) \approx \frac{a \log J}{b} + \frac{a(\log J)^2}{b^2}. \tag{A.29}$$

初始值的闭式反演。 给定目标边缘矩（μ_K，σ_K^2），其中 $\sigma_K^2 > \mu_K - 1$，负二项近似可得闭式初始值。令 $m := \mu_K - 1$:

$$a^{(0)} \approx \frac{m^2}{\sigma_K^2 - m}, \qquad b^{(0)} \approx \frac{a^{(0)} \log J}{m}. \tag{A.30}$$

这些值为 TSMM 算法阶段 1提供起点，随后在阶段 2通过精确矩匹配进行细化。

A.2.6. 数值稳定计算。 本节说明如何稳定计算 (i) Antoniak 概率质量函数（A.8）以及 (ii)（A.22）–（A.23）中的伽马混合期望。

A.2.6.1 对数斯特林递推。 无符号斯特林数 $|s(J,\ k)|$ 增长迅速；例如，$|s(50,\ 25)| \approx 1.4 \times 10^{54}$，对于中等$J$便超出双精度范围。因此，我们预先计算 $L_{n,\ k} :=$ 在对数空间中计算 $\log\ |s(n,\ k)|$。

递推式（A.9）在对数空间中，其中 $\log 0 := -\infty$，变为：

$$L_{n,k} = \mathrm{logaddexp}(L_{n-1,k-1},\ \log(n-1) + L_{n-1,k}), \tag{A.31}$$

其中对数加法指数函数$(u, v) := \max(u, v) + \log\{1 + \exp(-|u - v|)\}$。

算法 A.1（对数斯特林数计算）。

输入：最大样本量 $J_{\max}$。

输出：对数斯特林数表 $L[n, k]$，适用于 $1 \le n \le J_{\max}, 1 \le k \le n$。

1. 初始化： $L[1, 1] \leftarrow 0$（因为 $\log |s(1,1)| = \log 1 = 0$）。**2. 对于** $n = 2$ **至** $J_{\max}$：（a） $L[n, 1] \leftarrow L[n - 1, 1] + \log(n - 1)$（因为 $|s(n, 1)| = (n - 1)!$）。（b）**对于** $k = 2$ **至** $n - 1$：（c） $L[n, n] \leftarrow 0$（因为 $|s(n, n)| = 1$）。

$$L[n, k] \leftarrow \operatorname{logaddexp}\big(L[n - 1, k - 1],\ \log(n - 1) + L[n - 1, k]\big).$$

A.2.6.2 对数概率质量函数计算。给定 $\{L_{J,k}\}_{k=1}^{J}$，计算 $k = 1$、.....、J 的 Antoniak 对数概率质量函数：

$$\log p_{J,k}(\alpha) = L_{J,k} + k \log \alpha + \log \Gamma(\alpha) - \log \Gamma(\alpha + J). \tag{A.32}$$

为避免下溢并强制执行归一化，在对数空间中计算 $\log Z = \operatorname{logsumexp}\{\log p_{J,\,k}(\alpha) : k = 1,\ \ldots\ldots,\ J\}$ 并返回 $\log \pi_{J,\,k}(\alpha) = \log p_{J,\,k}(\alpha) - \log Z$。

验证检查：

1. 行和恒等式：$\sum_{k=1}^{J} |s(J,\ k)| = J!$（在对数空间中验证）。**2.** 小-$J$ 合理性检查：将 $|s(J,\ k)|$ 与 $J \le 10$的已知表进行比较。**3.** 概率质量函数检查：$\sum_{k=1}^{J}$ 概率$(K_J = k \mid \alpha) = 1$ 以及非负性。**4.** 边界检查：$\alpha \to 0 \Rightarrow$ 概率$(K_J = 1 \mid \alpha) \to 1$；$\alpha \to \infty \Rightarrow$ 概率$(K_J = J \mid \alpha) \to 1$。

A.2.6.3 伽马混合期望的高斯–拉盖尔求积。对于任意可测的 $g(\alpha)$ 满足 $\mathbb{E}|g(\alpha)| < \infty$,

$$\mathbb{E}_{a,b}\{g(\alpha)\} = \int_0^\infty g(\alpha)\, p(\alpha \mid a, b)\, d\alpha = \frac{1}{\Gamma(a)} \int_0^\infty g(x/b)\, x^{a-1} e^{-x}\, dx, \tag{A.33}$$

使用变量替换 $x = b\alpha$。参数为 $a - 1$ 的广义高斯–拉盖尔求积提供节点 $\{x_m\}_{m=1}^{M}$ 和权重 $\{w_m\}_{m=1}^{M}$，使得

$$\frac{1}{\Gamma(a)} \int_0^\infty g(x/b)\, x^{a-1} e^{-x}\, dx \approx \sum_{m=1}^{M} \tilde{w}_m\, g(\alpha_m), \qquad \alpha_m := x_m / b, \tag{A.34}$$

其中归一化权重为 $\tilde{w}_m := w_m / \Gamma(a)$。

默认设定。我们使用 $M = 80$ 节点，在我们的目标范围内（$a,\ b \in [0.1, 20]$, $J \le 300$），混合矩的相对精度达到 10^{-10} 量级

及其导数。相同的求积节点还可重复用于计算 $\mathbb{E}\{\kappa_J(\alpha)\}$、$\mathbb{E}\{v_J(\alpha)\}$以及 C节 中的雅可比矩阵分量。

A.2.6.4 复杂度摘要。对于固定的最大设计规模 $J_{\max}$:

表A.1。计算复杂度摘要

操作	复杂度
对数斯特林数表预计算	$O(J_{\max}^2)$ 时间/内存，缓存一次
所有 k 的条件概率质量函数 $\Pr(K_J = \cdot \mid \alpha)$	每个 α 对应的 $O(J)$
条件矩 $\kappa_J(\alpha)$ $v_J(\alpha)$	通过特殊函数计算的 $O(1)$
通过求积计算的边缘矩 $\mathbb{E}[K_J]$, $\mathrm{Var}(K_J)$	每 $(a,\ b)$ 对应的 $O(M)$

附录 B. 第一阶段闭式推导与近似理论

B.1. 目的与范围。 本附录推导两阶段矩匹配（TSMM）流程中使用的阶段 1（A1）闭式初始值。阶段 1 将簇数的 elicited 矩（μ_K, σ_K^2），即设计规模为 J 时针对 K_J 的矩，映射为伽马超先验 $\alpha \sim \mathrm{Gamma}(a_0, b_0)$（形状 – 率参数化）。

核心思路是用一个 $(a, b) \mapsto (\mathbb{E}[K_J], \text{方差}(K_J))$ 来替代狄利克雷过程混合模型下精确（且通常不可反演）的矩关系；该代理模型是一个平移泊松–伽马代理。这一代理模型给出了可解析反演的映射，其精度足以作为阶段 2 的稳健初始值；阶段 2 中将通过数值方法匹配精确矩（C节）。

归属说明。 关于 K_J 的代理分布近似（泊松化及其在伽马混合下的负二项推论）是标准方法，并已明确见于近期的狄利克雷过程精度先验研究（例如 Vicentini与Jermyn，2025年；Zito 等人，2024年），其基础是经典的泊松近似界（Arratia等人，2000；Le Cam，1960年）。下文的闭式反演是一个基础的矩匹配步骤；本文将其用作 TSMM 中的阶段 1 初始值。

本节内容安排如下。B.2节介绍移位近似，并推导由此诱导的负二项边缘分布。B.3节给出闭式反演（定理 B.6）及可行性条件。B.4节提供面向从业者的置信度到方差的映射。B.5节量化近似质量，并提出操作建议，说明何时仅阶段 1 就已足够，何时需要阶段 2 细化。

B.2. 平移泊松–伽马代理。

B.2.1. 移位近似。 回顾一下，中国餐馆过程（CRP）表示下的簇数满足

$$K_J = 1 + \sum_{i=2}^{J} Z_i, \qquad Z_i \mid \alpha \sim \text{Bernoulli}\left(\frac{\alpha}{\alpha + i - 1}\right),$$

其中 (Z_i) 在给定 α 的条件下相互独立（见 A 节，定理 A.1）。因此$K_J - 1$ 是一个泊松–二项和。经典泊松近似理论（Arratia等人，2000；Barbour 与 Xia，1999；Le Cam，1960年）启发我们用具有相同均值的泊松随机变量替代 $K_J - 1$ 。

定义 B.1 （平移泊松代理）。对于设计规模为 $J \geq 2$ 的情形，我们采用以下代理

$$K_J - 1 \mid \alpha \approx \text{Poisson}(\alpha c_J), \qquad c_J := \log J. \tag{B.1}$$

平移 1 可精确保证支撑约束 $K_J \geq 1$ ，同时为“超额”簇保留简单的计数模型。

为何 c_J = 对数 J**?** 精确的条件均值具有涉及双伽马函数的闭式表达式，

$$\kappa_J(\alpha) := \mathbb{E}[K_J \mid \alpha] = \sum_{i=1}^{J} \frac{\alpha}{\alpha + i - 1} = \alpha\{\psi(\alpha + J) - \psi(\alpha)\}, \tag{B.2}$$

并满足大 J渐近展开 $\kappa_J(\alpha) = \alpha$ 对数 $J + O(\alpha)$ 其中 α （Antoniak，1974）。最简单的代理将精确的、依赖于 α 的均值（B.2）替换为首项 α 对数 J，得到（B.1）。

关于线性化误差的重要提醒。展开式 $\kappa_J(\alpha) = \alpha \log J + O(\alpha)$ 隐含了一个依赖于 α 的常数：更准确地说， $\kappa_J(\alpha) = \alpha \log J + \alpha\gamma - \alpha\psi(\alpha) + O(\alpha/J)$，其中 γ 是欧拉-马斯克罗尼常数，且 $\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$。对于常数阶的 α （例如 $\alpha \in [0.5, 5]$） ，项 $-\alpha\psi(\alpha)$ 会带来不会随 $J \to \infty$ 增大而消失的偏差。因此，最好将 A1 视为阶段 2 的初始值，而不是独立方法：即使在 J 很大时，线性化 $\kappa_J(\alpha) \approx 1 + \alpha \log J$ 在边际均值中仍可能留下 $O(\mathbb{E}[\alpha])$ 阶的残差误差。

一种常见的替代方法是采用调和数缩放 H_{J-1}，因为

$$H_{J-1} = \sum_{i=1}^{J-1} \frac{1}{i} = \psi(J) + \gamma, \qquad H_{J-1} = \log J + \gamma + O(1/J), \tag{B.3}$$

其中 $\gamma \approx 0.5772$ 是欧拉-马斯克罗尼常数。

引理 B.1 (调和–双伽玛等价)。 对于整数 $J \geq 1$, $H_{J-1} = \psi(J) + \gamma$，其中 γ 是欧拉-马斯克罗尼常数。

因此，对于整数设计规模 J，只有两个不同的 α-无关竞争项：$\log J$和 H_{J-1}。使用 H_{J-1} 代替 $\log J$只会改变一个 $O(1)$ 项，因此两种标度渐近等价。

表 B.1. 标度常数选择对比

J	$\log J$	H_{J-1}	$\psi(J) + \gamma$	差异
10	2.303	2.829	2.829	0.526
25	3.219	3.776	3.776	0.557
50	3.912	4.499	4.499	0.587
100	4.605	5.187	5.187	0.582
300	5.704	6.283	6.283	0.579

注。 对于整数 J， $\psi(J) + \gamma \equiv H_{J-1}$ 恰好成立。差值 $H_{J-1} - \log J$在 $J \to \infty$时收敛至 $\gamma \approx 0.577$。

数值分析表明，对于 $\alpha > 0.2$的实际取值， $c_J = \log J$ 带来的 $K_J - 1$ 平均误差始终更小；对于移位近似，交叉点（即调和选项产生更小误差时的临界点）约为 $\alpha^* \approx 0.20$ ，低于典型引出信念的范围。因此，我们采用 c_J =$\log J$ 作为默认值。

实践说明。 TheDPprior 实现还提供调和选项（以及可选的插入式双伽玛校正），适用于 J 非常小或隐含的 α非常小的情况；详情请参阅软件包简介。

B.2.2. 伽马混合移位泊松。 阶段 1 反演利用了移位泊松代理与伽马先验共轭这一性质。

引理 B.2 (伽马混合移位泊松)。 设 $X \mid \alpha \sim$ 泊松(αc_J) 且 $c_J >0$，并设 $\alpha \sim$ 伽马分布(a, b) ，其形状参数为 $a > 0$ ，率参数为 $b > 0$。则 X 的边缘分布为负二项分布：

$$X \sim \mathrm{NegBin}\left(a,\, p = \frac{b}{b + c_J}\right), \tag{B.4}$$

其中我们采用如下参数化： p 为“成功”概率， $X \in \{0, 1, 2, \ldots\}$ 表示在 a 次成功之前的失败次数（当 a为非整数时，通过伽马函数延拓定义）。其前两个矩为

$$\mathbb{E}[X] = \frac{a c_J}{b}, \qquad \mathrm{Var}(X) = \frac{a c_J}{b} + \frac{a c_J^2}{b^2}. \tag{B.5}$$

证明。 通过直接边缘化，

$$\begin{aligned}\Pr(X=x) &= \int_0^\infty \underbrace{\frac{e^{-\alpha c_J}(\alpha c_J)^x}{x!}}_{\text{Poisson}(\alpha c_J)} \underbrace{\frac{b^a}{\Gamma(a)}\alpha^{a-1}e^{-b\alpha}}_{\text{Gamma}(a,b)}\, d\alpha \\ &= \frac{b^a c_J^x}{\Gamma(a)\,x!}\int_0^\infty \alpha^{a+x-1}e^{-(b+c_J)\alpha}\, d\alpha \\ &= \frac{\Gamma(a+x)}{\Gamma(a)\,x!}\left(\frac{b}{b+c_J}\right)^a\left(\frac{c_J}{b+c_J}\right)^x, \end{aligned} \tag{B.6}$$

这就是负二项分布的概率质量函数（B.4）。矩（B.5）由 负二项分布(a,p) 的标准公式得出。□

出处说明。泊松–伽马混合恒等式是经典结果（例如，参见 Johnson 等人，2005）。该恒等式在大规模 J情形下对狄利克雷过程浓度参数先验的应用，见 Vicentini 和 Jermyn（2025年，§3.1）以及 Zito 等人（2024年）。

与狄利克雷过程簇数的联系。令 $X = K_J - 1$，引理 B.2 给出近似量

$$K_J - 1 \mid (a,b) \approx \text{NegBin}\left(a,\ \frac{b}{b+\log J}\right), \tag{B.7}$$

这就是用于推导下文第 1 阶段映射的近似。

B.3. **A1 闭式反演。**

B.3.1. 闭式映射。设 (μ_K, σ_K^2) 表示对K_J的均值和方差进行引出先验，并定义平移后的均值

$$\mu_0 := \mu_K - 1. \tag{B.8}$$

在移位负二项近似（B.7）下， μ_0 和 σ_K^2 与（B.5）相匹配。

定理 B.6 （第 1 阶段 / A1 映射）。固定 $J \geq 2$和目标矩（μ_K, σ_K^2）。令 $\mu_0 = \mu_K - 1$和 $c_J =$对数 J。若

$$\mu_0 > 0 \quad \textit{and} \quad \sigma_K^2 > \mu_0, \tag{B.9}$$

则移位负二项近似有闭式解 n

$$\boxed{a_0 = \frac{\mu_0^2}{\sigma_K^2 - \mu_0}, \qquad b_0 = \frac{\mu_0\, c_J}{\sigma_K^2 - \mu_0}} \tag{B.10}$$

我们将其用作 *TSMM* 中的阶段 *1* 初始值 $\alpha \sim$ 伽马分布（a_0, b_0）。

该映射在正文中以公式（8）–（9）的形式给出。

证明。 根据 引理 B.2 ，在 $X = K_J - 1$ 下，代理矩为

$$\mu_0 = \mathbb{E}[X] = \frac{ac_J}{b}, \qquad \sigma_K^2 = \mathrm{Var}(X) = \frac{ac_J}{b} + \frac{ac_J^2}{b^2}. \tag{B.11}$$

第一个方程给出 $b = ac_J/\mu_0$。代入第二个方程可得 $\sigma_K^2 = \mu_0 + \mu_0^2/a$，因此 $a = \mu_0^2/(\sigma_K^2 - \mu_0)$。将其代入 $b = ac_J/\mu_0$ 得到（B.10）。□

B.3.2. 隐含的集中度参数摘要。

推论 B.2 (α 的隐含统计量**)**。在采用第 *1* 阶段初始值 $\alpha \sim$ 伽马分布(a_0, b_0)时，

$$\mathbb{E}[\alpha] = \frac{a_0}{b_0} = \frac{\mu_0}{c_J}, \qquad \mathrm{CV}(\alpha) = \frac{1}{\sqrt{a_0}} = \frac{\sqrt{\sigma_K^2 - \mu_0}}{\mu_0}. \tag{B.12}$$

解释。在 A1 下， μ_K 通过简单规则$\mathbb{E}[\alpha] \approx (\mu_K - 1)/\log J$设定先验均值 $\mathbb{E}[\alpha]$。形状参数 a_0 控制不确定性：随着 $\sigma_K^2 \downarrow \mu_0$，$a_0 \to \infty$，且先验在 $\mu_0/\log J$ 处变为准退化。相反，随着 σ_K^2增大， a_0 减小，这意味着对于 α，尾部更重且不确定性更大。

B.3.3. 可行区域与投影策略。

命题 B.3 (A1 的可行域**)**。 闭式映射 (B.10) 在以下条件下可得正值 $(a_0,\ b_0)$:

1. $\mu_K > 1$ （等价地， $\mu_0 > 0$）；*2.* $\mu_K \le J$ （因为 $1 \le K_J \le J$）；
3. $\sigma_K^2 > \mu_K - 1$ （相对于平移泊松代理存在过度离散）。

证明。 条件 (1) 和 (3) 确保 $\mu_0 > 0$ 和 $\sigma_K^2 - \mu_0 > 0$，因此（B.10）有定义，且满足 $a_0,\ b_0 > 0$。条件 (2) 是一项支撑约束，由 $K_J \le J$推出。 □

重要细节。 条件 (3) 是一项代理模型约束：在 A1 下，方差$(K_J - 1)$ $= \mu_0 + \mu_0^2/a \ge \mu_0$。在精确 DP 模型中， $K_J \mid \alpha$ 相对于泊松分布是条件欠离散的（A 节，命题 A.1），因此使用精确矩后，目标值满足 $\sigma_K^2 \le \mu_0$ 的情形仍可能可行。这正是 A1 作为初始值、并且存在 A2 的原因。

可行性投影（仅阶段 *1*）。当用户提供的 (μ_K, σ_K^2) 违反条件 (3) 时，阶段 1 会进行最小方差膨胀:

$$\sigma_{K,\mathrm{eff}}^2 := \max\{\sigma_K^2,\ \mu_0 + \varepsilon\}, \qquad \varepsilon = \max\{10^{-8},\ 10^{-6}\mu_0\}. \tag{B.13}$$

该投影会产生一个以μ_0/c_J为中心的准退化伽马先验（较大的 a_0）。该投影仅用于获得数值稳定的阶段 2 初始值;

阶段 2 求解器使用精确的矩方程，重新校准至原始 elicited 矩。

B.3.4. 计算示例。 考虑一个包含 $J = 50$ 个站点的设计，从业者预计会有约5个不同的簇，且不确定性中等： $\mu_K = 5$, $\sigma_K^2 = 10$。
步骤 1。 计算平移后的均值和缩放常数：

$$\mu_0 = 5 - 1 = 4, \qquad c_{50} = \log(50) \approx 3.912.$$

步骤 2。 验证可行性：$\sigma_K^2 = 10 > \mu_0 = 4$。
步骤 3。 应用闭式映射（B.10）：

$$a_0 = \frac{4^2}{10 - 4} = \frac{16}{6} \approx 2.667, \qquad b_0 = \frac{4 \times 3.912}{6} \approx 2.608.$$

步骤 4。 解读隐含的 α 先验：

$$\mathbb{E}[\alpha] = \frac{4}{3.912} \approx 1.02, \qquad \mathrm{CV}(\alpha) = \frac{1}{\sqrt{2.667}} \approx 0.612.$$

第一阶段结果 （a_0， b_0） ≈ （2.667，2.608）作为第二阶段牛顿细化的初始值，由此得到精确矩解（$a^\star$， $b^\star$） ≈ （1.407，1.076）——详见C节。

B.4. **置信度到方差的映射**。 从业者可能会发现，以定性的方式表达对 K_J 的不确定性，比直接设定 σ_K^2 更容易。我们提供了一个方差膨胀因子（VIF）映射，将定性置信水平转换为方差设定。

定义 B.2 （方差膨胀因子）。VIF 将定性置信度转换为方差：

$$\sigma_K^2 = \mathrm{VIF} \cdot (\mu_K - 1) = \mathrm{VIF} \cdot \mu_0. \tag{B.14}$$

由于 A1 要求 $\sigma_K^2 > \mu_K - 1$，因此必须满足方差膨胀因子 > 1。边界方差膨胀因子 $= 1$对应平移泊松方差。

表 B.2。VIF默认值建议

置信水平	VIF	σ_K^2 当 $\mu_K = 5$	解释
High	1.5	6.0	对簇数相当确定
中等（默认值）	2.5	10.0	中等不确定性
Low	5.0	20.0	显著不确定性

此映射在 DPprior_fit(confidence = "medium") 中实现，主要用于阶段 1 初始化；如果需要精确匹配矩，则可以调用阶段 2。

其他置信度接口。该DPprior软件包还支持：

1. 变异系数：$\mathrm{CV}_K = \sigma_K/\mu_K$，通过 $\sigma_K^2 = (\mathrm{CV}_K \cdot \mu_K)^2$ 转换。2. 中心区间：如果从业者指定 $K_J \in [k_L, k_U]$ 以概率 q （例如 $q = 0.80$），则通过正态代理近似方差：

$$\sigma_K^2 \approx \left(\frac{k_U - k_L}{2z_{(1+q)/2}} \right)^2, \tag{B.15}$$

其中， z_p 表示标准正态分布的 p 分位数。

B.5. **近似质量：阶段 1 何时足够？** 阶段 1 有意采用近似方法。这里涉及两种不同的近似：

1. 泊松化误差。 的条件分布是泊松–二项分布。泊松近似界（Arratia等人，2003；Le Cam，1960年）控制其与具有匹配均值的泊松分布之间的全变差距离，且该界会随着 J 增大而（缓慢）减小（A 节）。2. 均值线性化误差。即使 $K_J - 1$ 恰好服从均值为$\kappa_J(\alpha)-1$的泊松分布，阶段 1 仍会进一步将精确均值（B.2）替换为 αc_J，其中 $c_J =$对数 J。当 J较大时，首项是正确的，但被省略的(α) 项在中等 J时可能不可忽略。

这些近似适用于初始化，但若用作最终赋值结果，则可能过于粗略。

B.5.1. 近似行为与局限。

命题B.4 (A1的近似行为)。 设 (a_0, b_0) 为目标(μ_K, σ_K^2)对应的阶段 *1* 解。记 $M_1(a_0, b_0) = \mathbb{E}_{a_0,\ b_0}[K_J]$和 $V(a_0, b_0)$ =方差$_{a_0,\ b_0}(K_J)$ 为 $\alpha \sim$伽马分布(a_0, b_0)下的精确混合矩。近似误差很大程度上取决于诱导先验下 α 的大小：

1. 非消失误差区域。当 $\mathbb{E}[\alpha] = O(1)$ （即浓度参数为常数量级）时，线性化 $\kappa_J(\alpha) \approx 1 + \alpha \log J$会引入 $O(\mathbb{E}[\alpha])$ 阶的偏差，该偏差不会随着 $J \to \infty$而消失。在此区域，即使 J 很大，误差$|M_1(a_0, b_0) - \mu_K|$和 $|V(a_0, b_0) - \sigma_K^2|$也可能仍与零保持正距离。

2. 消失误差机制。若 $\mathbb{E}[\alpha] = O(1/\log J)$（等价地，$\mu_K - 1 = O(1)$当 $J \to \infty$ 时），则线性化偏差的相对贡献会减小，且近似误差随 J 减小。

说明。 精确条件均值满足 $\kappa_J(\alpha) = \alpha\{\psi(\alpha + J) - \psi(\alpha)\}$，展开后为

$$\kappa_J(\alpha) = \alpha \log J + \alpha\gamma - \alpha\psi(\alpha) + O(\alpha/J).$$

项 $-\alpha\psi(\alpha)$ 是一个与 J 无关的 α 函数，在伽马混合后会给边缘矩带来 $O(\mathbb{E}[\alpha])$ 偏差。当 α 为常数阶时，无论 J 如何，该偏差都会持续存在。泊松近似误差（全变差意义下为 $O(1/\log J)$）则是另一种阶数更小的效应。

注释 B.2。 命题B.4 阐明了 A1 被设计为一种初始化方法，而非独立的 elicitation 程序。闭式表达式（B.10）可立即提供计算起始值，但在 $\mathbb{E}[\alpha] = O(1)$ 所示的典型情形下，第二阶段牛顿细化对于准确进行矩匹配至关重要。

表B.3。代表性阶段1矩误差

J	目标 μ_K	目标 $\sigma^2{}_K$	A1 均值误差 $\lvert\Delta\mu\rvert$	A1 方差误差 $\lvert\Delta\sigma^2\rvert$
25	5	10	0.84	5.55
50	5	10	0.58	4.38
50	10	22.5	2.43	12.03
100	10	20	1.86	10.02
300	15	30	2.75	13.75

注。 对于每一行，我们从 A1 计算 (a_0, b_0)，然后在 $\mathbb{E}[K_J]$ 和方差(K_J)的条件下，使用高斯–拉盖尔求积（A 节，§A.2.6）计算其精确隐含值。桌子报告绝对误差 $|\Delta\mu| = |M_1(a_0, b_0) - \mu_K|$ 和 $|\Delta\sigma^2| = |V(a_0, b_0) - \sigma_K^2|$。

B.5.2. 有限-J 误差幅度。 桌子揭示了两种规律：

1. 对于固定的 elicited 矩，误差可能随着 J 减小；但正如 命题B.4所阐明的，这种下降并不保证会一直趋近于零——线性化偏差可能留下非零残差。2. 对于固定的 μ_K，误差会随着 J增大，因为较大的簇数对应较大的 $\mathbb{E}[\alpha]$，从而放大线性化偏差项 $-\alpha\psi(\alpha)$。

B.5.3. 操作建议。 阶段 1 闭式映射提供了计算上即时的初始化（亚毫秒级），其精度足以用于

定性敏感性探索。然而，对于需要精确矩匹配的应用——包括正式的先验预测校准、模拟研究，以及任何将报告先验诱导的 $p(K_J)$ 的分析——阶段 2 牛顿法细化至关重要。

当 J 处于几十或一两百的范围时，最好将 A1 视为一种快速的闭式初始值，而非独立的征询方法。在这种情况下，第二阶段的矩匹配（C节）能显著缩小矩差异，同时保留所征询输入的可解释性。

附录C. 第二阶段牛顿细化：完整算法与收敛性

C.1. 目的与范围。 本附录完整说明 TSMM 第二阶段（A2）的设定：通过精确矩牛顿法细化，调整伽马超参数（a， b）

$$\alpha \sim \mathrm{Gamma}(a, b) \qquad (\text{shape } a > 0,\ \text{rate } b > 0)$$

使位点分区数 K_J 的精确狄利克雷过程诱导混合矩与征询目标值（μ_K， σ_K^2）相匹配（正文算法框 1）。

第一阶段（B 节）提供一个快速的闭式初始值 $(a_0,\ b_0)$，该初始值基于大J 近似。第二阶段通过牛顿法求解精确混合矩方程，以消除剩余的有限J 偏差。该方法使用由得分恒等式计算的解析雅可比矩阵，并采用与矩函数相同的高斯–拉盖尔求积节点进行计算。

归属说明（哪些是经典内容，哪些是本文特有内容）。 牛顿法及其局部二次收敛理论是经典的（例如，Dennis 与 Schnabel，1996年；Nocedal 与 Wright，2006年；Ortega 与 Rheinboldt，1970年）。用于对期望求导的得分函数恒等式同样是标准的（例如，Casella 与 Berger，2002年）。本文特有的是如下校准流程，它将以下内容结合起来：（i） $K_J \mid \alpha$ 的精确条件矩（A 节，命题 A.1），（ii）阶段 1 的闭式映射 $(\mu_K,\ \sigma_K^2,\ J) \mapsto (a_0,\ b_0)$（B 节，定理 B.6），以及（iii）精确矩牛顿求解器，其雅可比矩阵以解析方式计算，并通过求积高效求值。

C.2. **问题陈述与目标方程。** 我们采用形状-速率伽马密度

$$g_{a,b}(\alpha) = \frac{b^a}{\Gamma(a)}\alpha^{a-1}e^{-b\alpha}, \qquad \alpha > 0,\ a > 0,\ b > 0. \tag{C.1}$$

Let

$$\kappa_J(\alpha) := \mathbb{E}(K_J \mid \alpha), \qquad v_J(\alpha) := \mathrm{Var}(K_J \mid \alpha), \tag{C.2}$$

表示精确条件矩（A 节，命题 A.1）。

C.2.1. 伽马混合下的混合矩。 定义 $\alpha \sim$ 伽马分布(a, b) 下的混合（先验预测）均值和方差:

$$M_1(a,b) := \mathbb{E}_{a,b}[K_J] = \mathbb{E}_{\alpha\sim g_{a,b}}\{\kappa_J(\alpha)\}, \tag{C.3}$$

并且，根据全方差定律，

$$V(a,b) := \operatorname*{Var}_{a,b}(K_J) = \mathbb{E}_{\alpha\sim g_{a,b}}\{v_J(\alpha)\} + \operatorname*{Var}_{\alpha\sim g_{a,b}}\{\kappa_J(\alpha)\}. \tag{C.4}$$

等价地，令 $m_r(a,b) := \mathbb{E}\{\kappa_J(\alpha)^r\}$（对 $r \in \{1,2\}$），以及 $v_1(a,b) := \mathbb{E}\{v_J(\alpha)\}$,

$$M_1(a,b) = m_1(a,b), \qquad V(a,b) = v_1(a,b) + m_2(a,b) - m_1(a,b)^2. \tag{C.5}$$

上述所有期望均使用广义高斯–拉盖尔求积进行计算（A 节，§A.2.6.3）。重要的是，在每次牛顿迭代中，我们都使用相同的求积节点和权重来计算矩及其雅可比矩阵元素，针对当前的 (a, b)。

C.2.2. 求根表述。

定义 C.3（A2 矩匹配问题）。给定 $(J,\ \mu_K,\ \sigma_K^2)$，求 $(a^\star,\ b^\star) \in (0,\ \infty)^2$，使得

$$F(a,b) := \begin{pmatrix} M_1(a,b) - \mu_K \\ V(a,b) - \sigma_K^2 \end{pmatrix} = \mathbf{0}. \tag{C.6}$$

阶段 2 从第 1 阶段初始值 C.6 $(a_0,\ b_0)$ 开始求解 (B 节)。

C.3. **通过得分恒等式推导雅可比矩阵。**

C.3.1. 得分恒等式。

引理 C.3（得分函数恒等式）。 设 $\alpha \sim g_{a,\ b}$ 且令 h（α）为任意函数，使得积分号下求导是合理的。对于 $\theta \in \{a, b\}$,

$$\frac{\partial}{\partial\theta}\,\mathbb{E}_{a,b}\{h(\alpha)\} = \mathbb{E}_{a,b}\{h(\alpha)\,s_\theta(\alpha)\}, \qquad s_\theta(\alpha) := \frac{\partial}{\partial\theta}\log g_{a,b}(\alpha). \tag{C.7}$$

对于 伽马分布（a, b）密度 （C.1），

$$s_a(\alpha) = \log b - \psi(a) + \log \alpha, \qquad s_b(\alpha) = \frac{a}{b} - \alpha. \tag{C.8}$$

证明。 将 $\mathbb{E}_{a,b}\{h(\alpha)\} = \int_0^\infty h(\alpha) g_{a,b}(\alpha)\, d\alpha$ 写出，并在积分号下求导：

$$\frac{\partial}{\partial\theta} \int h(\alpha)\, g_{a,b}(\alpha)\, d\alpha = \int h(\alpha) \frac{\partial}{\partial\theta} g_{a,b}(\alpha)\, d\alpha = \int h(\alpha)\, g_{a,b}(\alpha)\, \frac{\partial}{\partial\theta} \log g_{a,b}(\alpha)\, d\alpha,$$

由此得到（C.7）。 （C.8）中的得分公式可通过对 $\log g_{a,b}(\alpha) = a \log b - \log \Gamma(a) + (a-1) \log \alpha - b\alpha$. □ 求导得到

C.3.2. *A2* 系统的雅可比矩阵元素。

推论 C.3 ($F(a,b)$的雅可比矩阵。 令 F定义见 (C.6)。其雅可比矩阵$\mathbf{J}_F(a,b) = \nabla F(a,b)$ 可写为

$$\mathbf{J}_F(a,b) = \begin{pmatrix} \partial_a M_1 & \partial_b M_1 \\ \partial_a V & \partial_b V \end{pmatrix}, \tag{C.9}$$

其中对于 $\theta \in \{a,b\}$,

$$\partial_\theta M_1 = \partial_\theta m_1 = \mathbb{E}\{\kappa_J(\alpha)\, s_\theta(\alpha)\}, \tag{C.10}$$

and

$$\partial_\theta V = \partial_\theta v_1 + \partial_\theta m_2 - 2 m_1\, \partial_\theta m_1, \tag{C.11}$$

with

$$\partial_\theta v_1 = \mathbb{E}\{v_J(\alpha)\, s_\theta(\alpha)\}, \qquad \partial_\theta m_2 = \mathbb{E}\{\kappa_J(\alpha)^2\, s_\theta(\alpha)\}. \tag{C.12}$$

证明。 将引理 C.3 应用于 m_1, m_2 和 v_1，然后对（C.5）求导。 □

实现说明（效率）。在每次牛顿迭代中，求积会生成节点 $\{\alpha_m\}$和权重 $\{\tilde{w}_m\}$ (A 节，§A.2.6.3）。我们对每个节点仅计算一次 $\kappa_J(\alpha_m)$、 $v_J(\alpha_m)$以及得分 $s_a(\alpha_m)$ $s_b(\alpha_m)$ ，然后通过加权求和计算（C.10）–（C.12）中的所有项。这样可以避免有限差分雅可比矩阵，并使每次迭代的计算成本保持为 $O(M)$。

C.3.3. 对数参数化（推荐）。为确保$(a,\ b) > 0$ 并提高数值稳定性，我们使用对数参数求解（C.6）

$$\eta = (\eta_1, \eta_2) = (\log a, \log b) \in \mathbb{R}^2, \qquad (a,b) = (e^{\eta_1}, e^{\eta_2}). \tag{C.13}$$

令 $G(\eta) := F(e^{\eta_1},\ e^{\eta_2})$。根据链式法则，

$$\mathbf{J}_G(\eta) = \nabla_\eta G(\eta) = \mathbf{J}_F(a,b)\, \mathrm{diag}(a,b). \tag{C.14}$$

C.4.**算法 C.1：A2-MN（精确矩牛顿法细化）**。完整的 A2-MN 流程如下所述。

A算法 C.1（A2-MN：用于精确矩匹配的牛顿细化）.

输入： 设计规模 J；目标矩（μ_K， σ_K^2）；可选初始猜测（a_0， b_0）；求积规模 M；容差 $\texttt{tol}_F$， $\texttt{tol}_\eta$；最大迭代次数 $\texttt{max_iter}$。

输出： 细化后的（$a^\star$， $b^\star$）；诊断量（迭代次数、最终残差、终止代码）。

0. **可行性预筛选（低成本必要检查）**。要求

$$1 \le \mu_K \le J, \qquad 0 \le \sigma_K^2 \le \frac{(J-1)^2}{4}. \tag{C.15}$$

（方差界是对取值于$\{1,\dots,J\}$上的变量应用的波波维丘不等式。）若不满足，则在数值优化前返回错误。

1. **阶段 1 初始化（如有需要）**。如果（a_0， b_0）未提供，则通过阶段 1 闭式映射（B 节，定理 B.6）计算。如果 $\sigma_K^2 \le \mu_0 : = \mu_K - 1$，应用阶段 1 投影策略（B 节，§B.3.3）以获得数值稳定的初始值。

2. **初始化**。设置 $\eta_0 = (\log\, a_0,\ \log\, b_0)$ 和 $t \leftarrow 0$。

3. **迭代** 对 $t = 0, 1, \dots, \texttt{max_iter} - 1$:

(a) **计算矩和残差**。令 $(a_t, b_t) = (e^{\eta_{t,1}}, e^{\eta_{t,2}})$。使用高斯–拉盖尔求积计算 $\big(M_1(a_t,b_t)\ \ V(a_t,b_t)\big)$（A 节，§A.2.6.3），并设置

$$f_t = G(\eta_t) = \begin{pmatrix} M_1(a_t,b_t) - \mu_K \\ V(a_t,b_t) - \sigma_K^2 \end{pmatrix}. \tag{C.16}$$

(b) **停止准则（残差）**。如果 $\|f_t\|_\infty \le \texttt{tol}_F$，返回 $(a^\star, b^\star) = (a_t, b_t)$.

(c) **雅可比矩阵（基于得分的）**。计算 $\mathbf{J}_F(a_t,b_t)$ 使用 推论 C.3（使用与步骤 3a 相同的求积节点）。构造

$$\mathbf{J}_G(\eta_t) = \mathbf{J}_F(a_t,b_t)\,\mathrm{diag}(a_t,b_t). \tag{C.17}$$

如果 $|\det(\mathbf{J}_G)| < 10^{-12}$，则对 $\mathbf{J}_G \leftarrow \mathbf{J}_G + \varepsilon I_2$ 使用 $\varepsilon = 10^{-8}$ 进行正则化。

（d）**牛顿方向**。求解 2×2 线性方程组

$$\mathbf{J}_G(\eta_t)\,\Delta\eta_t = -f_t. \tag{C.18}$$

(e) **停止准则（步长）**。若 $\|\Delta\eta_t\|_\infty \le \texttt{tol}_\eta$，则返回 $(a_t,\ b_t)$。

(f) 回溯线搜索（Armijo）。设置 $\lambda \leftarrow 1$ 并重复：$\eta_{\text{cand}} \leftarrow \eta_t + \lambda\,\Delta\eta_t$，$f_{\text{cand}} \leftarrow G(\eta_{\text{cand}})$。满足以下条件时接受该步长：

$$\|f_{\text{cand}}\|_2 \le (1 - c\lambda)\,\|f_t\|_2, \qquad c = 0.5. \tag{C.19}$$

否则，设置 $\lambda \leftarrow \lambda/2$ 并重试。如果 $\lambda < 10^{-8}$，则以警告标志（无进展）终止。

(g) 更新。设置 $\eta_{t+1} \leftarrow \eta_t + \lambda\,\Delta\eta_t$。

4. **不收敛时返回**。如果 $t = \texttt{max_iter}$ 且未满足停止准则，则返回最后一次迭代结果，并附上警告标志。

推荐默认值。 在我们的参考设置中，使用 $M = 80$ 个求积节点，$\texttt{tol}_F = 10^{-8}$，$\texttt{tol}_\eta = 10^{-10}$，以及 $\texttt{max_iter} = 20$。采用阶段 1 初始化时，通常只需少量牛顿步即可收敛至接近机器精度（见 §C.6）。

C.5. 收敛性：能保证什么（以及不能保证什么）。

定理 C.7 (牛顿细化的局部二次收敛)。考虑求根问题 $G(\eta) = 0$，其中 G 由 (C.13)定义。假设存在 $\eta^\star$使得 $G(\eta^\star) = 0$，并假设：

1. G在 $\eta^\star$的某个邻域内连续可微，且 $\mathbf{J}_G$ 在该邻域内利普希茨连续；*2.* $\mathbf{J}_G(\eta^\star)$ 非奇异。

那么存在 $\eta^\star$ 的一个邻域 $\mathcal{N}_0$，使得如果 $\eta_0 \in \mathcal{N}_0$，无阻尼牛顿迭代

$$\eta_{t+1} = \eta_t - \mathbf{J}_G(\eta_t)^{-1} G(\eta_t) \tag{C.20}$$

有定义、始终留在该邻域内、收敛到 $\eta^\star$，并且满足

$$\|\eta_{t+1} - \eta^\star\| \le C\,\|\eta_t - \eta^\star\|^2 \tag{C.21}$$

其中某个常数为 $C > 0$。

证明概要。这是非线性系统牛顿–坎托罗维奇理论的标准推论；参见 Ortega 和 Rheinboldt（1970年，第 10 章）或 Dennis 和 Schnabel（1996年，第 5 章）。□

局部与全局行为。 定理 C.7是局部的，并不保证从任意初始点出发都能收敛。TSMM 依靠阶段 1 将 (a_0, b_0) 置于中等到较大 J 情形下的解附近（B 节，命题 B.4）。对于较小的

J，算法采用标准的全局化保护措施（对数参数化、线搜索、雅可比正则化），以防止发散，并在目标数值上存在问题时平稳地终止。

C.6. **计算示例与典型迭代次数。**

C.6.1. 计算示例（附录 *B.3.4* 的延续）。 B 节，§B.3.4 考虑 $J = 50$ 个站点，其目标为 $(\mu_K, \sigma_K^2) = (5, 10)$。阶段 1 给出

$$(a_0, b_0) \approx (2.667,\ 2.608).$$

在 (a_0, b_0) 处计算精确混合矩，表明存在有限-J 偏差（阶段 1 低估了均值和方差）。阶段 2 通过求解 (C.6) 对此加以校正。

表 C.1. 阶段 1 → 阶段 2 校正（附录 B.3.4 的示例，$J = 50$）

阶段	(a, b)	$\mathbb{E}_{a,\ b}(K_J)$	方差$_{a,\ b}(K_J)$	相对误差（均值、方差）
阶段 1（A1 初始化）	(2.667, 2。608)	4.415	5.618	−11.7%, −43.8%
阶段 2（A2 最终结果）	(1.408, 1.077)	5.000	10.000	≈ 0, ≈ 0

解释。 阶段 1 使用泊松–伽马代理，忽略了 $K_J \mid \alpha$ （A 节，推论 A.1）中的条件欠离散，因此当 J 不大时，可能会低估 $\mathrm{Var}(K_J)$。阶段 2 调整伽马超先验，以恢复精确的混合矩。

C.6.2. 典型迭代次数基准。 下表报告了使用参考设置 $M = 80$、$\texttt{tol}_F = 10^{-8}$、$\texttt{tol}_\eta = 10^{-10}$ 和 Armijo 阻尼时 A2-MN 的代表性表现。

表 C.2。来自阶段 1 初始化的代表性 A2-MN 迭代

J	目标 $(\mu_K,\ \sigma_K^2)$	阶段 1 $(a_0,\ b_0)$	A2-MN 步长	阶段 2 $(a^\star,\ b^\star)$
25	(5, 10)	(2.667, 2.146)	6	(1。035, 0.531)
50	(5, 10)	(2.667, 2。608)	5	(1.408, 1。077)
50	(10, 22。5)	(6.000, 2。608)	5	(2.240, 0.579)
100	(10, 20)	(7.364, 3.768)	4	(3.578, 1。327)
300	(15, 30)	(12.250, 4。991)	5	(6.772, 2。091)

这些示例体现了更广泛的规律：阶段 1 已足够接近，可使牛顿法处于稳定状态；少量迭代即可实现近乎机器精度的匹配。

计算扩展性。 每次 A2-MN 迭代只需计算固定数量的求积期望，因此每次迭代的复杂度为 $O(M)$（默认采用 $M = 80$）。相比之下，分布级细化（例如下文的 A2-KL）需要计算 $k = 1, \ldots, J$ 上的完整诱导概率质量函数，因此每次迭代的复杂度为 $O(JM)$。

C.7. **可选：A2-KL 细化（以 K_J 为分布目标）。** 当从业者信念可用（μ_K, σ_K^2）概括时，矩匹配是合适的。然而，在某些情况下，从业者（或旧版基线）会指定 $k \in \{1, \ldots, J\}$ 上的完整目标分布 $p^\star(k)$。

定义 C.4（KL 细化目标）。 给定 $\{1, \ldots, J\}$ 上的目标概率质量函数 $p^\star$，定义

$$(a^\dagger, b^\dagger) \in \arg\min_{a>0,\, b>0} D_{\mathrm{KL}}(p^\star \,\|\, p_{a,b}), \qquad D_{\mathrm{KL}}(p^\star \| p_{a,b}) := \sum_{k=1}^{J} p^\star(k) \log \frac{p^\star(k)}{p_{a,b}(k)}, \tag{C.22}$$

其中 $p_{a,b}(k) = \Pr(K_J = k \mid a, b)$ 是诱导边缘概率质量函数（A 节，定理 A.3）。

C.7.1. 通过得分恒等式求梯度。 记$p_{a、b}(k) = \mathbb{E}_{a、b}\{p(K_J = k \mid \alpha)\}$，其中$p(K_J = k \mid \alpha)$是 Antoniak 概率质量函数（A 节，公式（A.8））。 引理 C.3 给出，对于$\theta \in \{a, b\}$,

$$\frac{\partial}{\partial \theta} p_{a,b}(k) = \mathbb{E}_{a,b}\{p(K_J = k \mid \alpha)\, s_\theta(\alpha)\}. \tag{C.23}$$

因此,

$$\frac{\partial}{\partial \theta} D_{\mathrm{KL}}(p^\star \| p_{a,b}) = -\sum_{k=1}^{J} \frac{p^\star(k)}{p_{a,b}(k)} \frac{\partial}{\partial \theta} p_{a,b}(k). \tag{C.24}$$

（C.23）中的所有期望都可以使用与 A2-MN 中相同的求积工具进行计算；额外的成本来自遍历 $k = 1$、……、 J。

C.7.2. 算法概述。 下面概述 A2-KL 优化过程。

算法 C.2（A2-KL：分布细化；概述）。

输入： J；目标概率质量函数 $p^\star(1{:}J)$；初始值 $(a_{\text{init}}, b_{\text{init}})$（推荐使用 A2-MN 解）；求积规模 M；优化器设置（例如 L-BFGS 容差）。

输出： $(a^\dagger,\ b^\dagger)$ 最小化（C.22）。

1. 初始化 $(a, b) \leftarrow (a_{\text{init}}, b_{\text{init}})$，并在 (log a,log b) 中进行优化。2. 在每次优化器迭代中：

- 计算 $p_{a,b}(k)$ 对于 $k = 1,\ldots, J$ (A 节, 定理 A.3；通过 A 节, §A.2.6.1) 进行稳定求值，
- 计算 KL 目标函数 (C.22)，
- 通过求积，根据 (C.23)–(C.24) 计算梯度。

3. 返回收敛后的解。

C.7.3. *DORO*-均匀基线（可选目标）。在 DORO 式校准中用作基线的一种简单分布目标，是在 $\{1,\ldots,m\}$上的均匀概率质量函数，其中 m 选择该值以匹配目标均值尺度（Dorazio，2009年）：

$$p^\star(k) = \frac{1}{m}\,\mathbf{1}\{1 \le k \le m\}, \qquad m := \min\{2\mu_K - 1,\ J\}. \tag{C.25}$$

该基线可通过 A2-KL 进一步改进，而 A2-MN 可提供快速且稳定的初始值。

A附录D. 近似误差界与有限-J 分析

D.1. **目 的目的与范围。** B 节介绍了阶段 1（A1）近似

$$K_J - 1 \mid \alpha \approx \text{Pois}(\alpha c_J), \qquad c_J = \log J, \tag{D.1}$$

这意味着，在对伽马超先验 K_J 进行混合后，$\alpha \sim$ 伽马分布(a, b)服从平移负二项分布。C节 随后介绍了阶段 2（A2），其将(a, b) 调整为匹配精确的有限-J 矩。

本附录针对有限 J 时的 A1，提供严格的分布层面误差量化，具体有两个目标：

1. 误差核算。识别并界定小型/中等规模 J 下的主要近似误差来源。 2. 决策支持。将这些界限转化为可执行的指导意见，说明何时建议使用 A2（或精确枚举）。

全文中，我们研究固定设计规模 J 下狄利克雷过程／中国餐馆过程（CRP）的簇数 $K_J \in \{1,\ldots,J\}$。

D.2. **三路误差分解。**

D.2.1. 条件律与边缘律。设

$$S_J := K_J - 1 \in \{0, 1, \ldots, J-1\}. \tag{D.2}$$

由于总变差（TV）距离在确定性平移下保持不变，因此对 S_J 的 A1 进行界定，等价于对 K_J 的 A1 进行界定。

根据 A 节（定理 A.1），给定 α 时，我们有泊松–二项表示

$$S_J = \sum_{j=2}^{J} B_j, \qquad B_j \mid \alpha \text{ independent}, \quad B_j \sim \text{Bernoulli}(p_j(\alpha)), \quad p_j(\alpha) = \frac{\alpha}{\alpha + j - 1}. \tag{D.3}$$

定义精确条件概率质量函数和 A1 代理条件概率质量函数

$$p_J(k \mid \alpha) := \mathbb{P}(S_J = k \mid \alpha), \qquad \tilde{p}_J(k \mid \alpha) := \mathbb{P}\big(\text{Pois}(\alpha c_J) = k\big), \qquad k = 0, \ldots, J-1. \tag{D.4}$$

令 $\pi_{a,\,b}$ 表示伽马分布$(a,\ b)$密度（形状–率），并定义诱导的边际概率质量函数

$$p_{a,b}(k) := \int_0^\infty p_J(k \mid \alpha)\, \pi_{a,b}(\alpha)\, d\alpha, \qquad \tilde{p}_{a,b}(k) := \int_0^\infty \tilde{p}_J(k \mid \alpha)\, \pi_{a,b}(\alpha)\, d\alpha. \tag{D.5}$$

在 A1 下，泊松–伽马混合恰好是负二项分布，因此 $\tilde{p}_{a,\,b}$ 可用闭式表示（B 节）。

D.2.2. 误差来源与分解。

定义 D.5（误差来源）。固定 $J \geq 2$ 和 $\alpha > 0$。记作

$$\lambda_J(\alpha) := \mathbb{E}[S_J \mid \alpha] = \sum_{j=2}^{J} p_j(\alpha) = \kappa_J(\alpha) - 1 = \alpha\{\psi(\alpha + J) - \psi(\alpha + 1)\}, \tag{D.6}$$

其中 $\kappa_J(\alpha) = \mathbb{E}[K_J \mid \alpha]$ 和 $\psi(\cdot)$ 是双伽马函数。将条件 A1 误差分解为：

1. *E1*（泊松化）：

$$E_1(\alpha) := d_{\text{TV}}(\mathcal{L}(S_J \mid \alpha),\ \text{Pois}(\lambda_J(\alpha)))\,; \tag{D.7}$$

2. *E2*（均值线性化）：

$$E_2(\alpha) := d_{\text{TV}}(\text{Pois}(\lambda_J(\alpha)),\ \text{Pois}(\alpha c_J))\,, \qquad c_J = \log J; \tag{D.8}$$

3. *E3*（伽马混合）：超先验 $\alpha \sim$ 伽马分布(a, b) 并未引入第三个近似步长，而是将 $E_1(\alpha)$ 和 $E_2(\alpha)$ 传递到边缘$p_{a,b}$。

定理 D.8（混合下的误差分解）。令$p_{a,\ b}$ 为在 $\alpha \sim$ 伽马分布(a, b)下 S_J 的精确边缘概率质量函数；令 $\tilde{p}_{a,\ b}$ 为 *A1*（平移负二项分布）近似，其 $c_J =$对数 J。则

$$d_{\mathrm{TV}}(p_{a,b}, \tilde{p}_{a,b}) \le \mathbb{E}_{\alpha\sim\mathrm{Gamma}(a,b)}[E_1(\alpha)] + \mathbb{E}_{\alpha\sim\mathrm{Gamma}(a,b)}[E_2(\alpha)]\,. \tag{D.9}$$

证明。根据定义和詹森不等式，

$$\begin{aligned} d_{\mathrm{TV}}(p_{a,b}, \tilde{p}_{a,b}) &= \frac{1}{2}\sum_{k=0}^{J-1}\left|\int \{p_J(k \mid \alpha) - \tilde{p}_J(k \mid \alpha)\}\,\pi_{a,b}(\alpha)\,d\alpha\right| \\ &\le \int \left\{\frac{1}{2}\sum_{k=0}^{J-1}|p_J(k \mid \alpha) - \tilde{p}_J(k \mid \alpha)|\right\}\pi_{a,b}(\alpha)\,d\alpha \\ &= \mathbb{E}_\alpha\big[d_{\mathrm{TV}}\big(p_J(\cdot \mid \alpha), \tilde{p}_J(\cdot \mid \alpha)\big)\big]\,. \end{aligned}$$

对固定的 α，应用三角不等式

$$d_{\mathrm{TV}}\big(p_J(\cdot \mid \alpha), \tilde{p}_J(\cdot \mid \alpha)\big) \le d_{\mathrm{TV}}(\mathcal{L}(S_J \mid \alpha), \mathrm{Pois}(\lambda_J(\alpha))) + d_{\mathrm{TV}}(\mathrm{Pois}(\lambda_J(\alpha)), \mathrm{Pois}(\alpha c_J))\,,$$

这等于 $E_1(\alpha) + E_2(\alpha)$。积分即完成证明。 □

出处说明。共同混合分布下总变差的收缩是标准结论；定理 D.8 的作用是明确指出，A1 误差由两个条件项控制（E_1，E_2），这两个项可以分别界定和解释。

D.3. 条件（固定-α）界。

D.3.1. 泊松化误差 $E_1(\alpha)$。

定理 D.9 (泊松近似总变差界)。固定 $J \ge 2$ 和 $\alpha > 0$。令$\lambda_J(\alpha) = \sum_{j=2}^{J} p_j(\alpha)$ 如 (D.6) 所示，则

$$E_1(\alpha) = d_{\mathrm{TV}}(\mathcal{L}(S_J \mid \alpha),\ \mathrm{Pois}(\lambda_J(\alpha))) \le \min\left(1, \frac{1}{\lambda_J(\alpha)}\right)\sum_{j=2}^{J} p_j(\alpha)^2. \tag{D.10}$$

此外，

$$\sum_{j=2}^{J} p_j(\alpha)^2 = \lambda_J(\alpha) - v_J(\alpha) = \alpha^2\{\psi_1(\alpha+1) - \psi_1(\alpha+J)\}, \tag{D.11}$$

其中 $v_J(\alpha) =$ 方差$(S_J \mid \alpha)$，且 $\psi_1(\cdot)$ 是三伽马函数。

证明。 界（D.10）是针对独立伯努利指示变量之和的标准陈–斯坦/Le Cam 泊松近似不等式（Barbour & Hall，1984年；Bar- bour 等，1992年；Le Cam，1960年）。此处 $S_J = \sum_{j=2}^{J} B_j$ 服从 $B_j \sim$ 伯努利分布$(p_j(\alpha))$在给定 α的条件下相互独立，且 $\lambda_J(\alpha) = \sum_{j=2}^{J} p_j(\alpha)$。

对于（D.11），注意

$$v_J(\alpha) = \sum_{j=2}^{J} p_j(\alpha)\{1 - p_j(\alpha)\} = \sum_{j=2}^{J} p_j(\alpha) - \sum_{j=2}^{J} p_j(\alpha)^2 = \lambda_J(\alpha) - \sum_{j=2}^{J} p_j(\alpha)^2,$$

因此 $\sum_{j=2}^{J} p_j(\alpha)^2 = \lambda_J(\alpha) - v_J(\alpha)$。最后，

$$\sum_{j=2}^{J} p_j(\alpha)^2 = \alpha^2 \sum_{m=1}^{J-1} \frac{1}{(\alpha+m)^2} = \alpha^2\{\psi_1(\alpha+1) - \psi_1(\alpha+J)\}$$

根据级数表示，$\psi_1(x) = \sum_{m=0}^{\infty}(x+m)^{-2}$。 □

出处说明。（D.10）是经典结果（Barbour & Hall，1984年；Barbour等人，1992年；Le Cam，1960年）。特殊函数恒等式（D.11）可由标准三伽马函数级数展开得到。

解释（为何率参数较慢）。 对于固定的 α，$\lambda_J(\alpha) \asymp \alpha \log J$，而 $\sum p_j(\alpha)^2$在 $J \to \infty$时收敛到一个有限常数，因此（D.10）给出 $E_1(\alpha) = O(1/\log J)$。这是对数缓慢的率参数：例如，使用自然对数时，

$$\frac{1}{\log 20} \approx 0.33, \qquad \frac{1}{\log 50} \approx 0.26, \qquad \frac{1}{\log 100} \approx 0.22.$$

关于相关“对数组合结构”的尖锐结果表明，$1/\log J$这一标度在许多情形下并不仅仅是该证明的人为产物（Arratia等人，2003）。

D.3.2. 均值线性化误差 $E_2(\alpha)$ 对于 $c_J = \log J$。 A1 使用 $c_J = \log J$对精确的条件均值 $\lambda_J(\alpha)$ 进行线性化。

定理 D.10 （$\lambda_J(\alpha)$ 的均值线性化界）。 固定 $\alpha > 0$ 和 $J \geq 2$。令$\lambda_J(\alpha) = \alpha\{\psi(\alpha+J) - \psi(\alpha+1)\}$，

$$|\lambda_J(\alpha) - \alpha \log J + \alpha\,\psi(\alpha+1)| \leq \frac{\alpha^2}{J} + \frac{\alpha}{2J} + \frac{\alpha}{12J^2}. \tag{D.12}$$

特别地，

$$|\lambda_J(\alpha) - \alpha \log J| \leq \alpha\,|\psi(\alpha+1)| + \frac{\alpha^2}{J} + \frac{\alpha}{2J} + \frac{\alpha}{12J^2}. \tag{D.13}$$

证明。写出

$$\lambda_J(\alpha) - \alpha \log J = \alpha\{\psi(\alpha+J) - \log J\} - \alpha\psi(\alpha+1).$$

Thus

$$|\lambda_J(\alpha) - \alpha \log J + \alpha\psi(\alpha+1)| = \alpha\,|\psi(\alpha+J) - \log J|.$$

界

$$|\psi(\alpha+J) - \log J| \le |\psi(\alpha+J) - \log(\alpha+J)| + |\log(\alpha+J) - \log J|.$$

对于第一项，标准双伽马不等式给出

$$\left|\psi(x) - \log x + \frac{1}{2x}\right| \le \frac{1}{12x^2} \qquad (x>0),$$

因此

$$|\psi(\alpha+J) - \log(\alpha+J)| \le \frac{1}{2(\alpha+J)} + \frac{1}{12(\alpha+J)^2} \le \frac{1}{2J} + \frac{1}{12J^2}.$$

对于第二项，$\log(\alpha+J) - \log J = \log(1+\alpha/J) \le \alpha/J$。合并并乘以 α 得到（D.12），且（D.13）由三角不等式得出。□

出处说明。双伽马近似不等式是标准结果（Abramowitz & Stegun, 1964, §6.3）。

解释。（D.12）中的主导“偏差”项是与 J 无关的量 $-\alpha\,\psi(\alpha+1)$；其余项为 $O(\alpha^2/J)$。因此，即使泊松化误差 $E_1(\alpha)$ 已经很小，使用 $c_J = \log J$ 在中等 J 时仍可能留下不可忽略的均值失配。

D.3.3. 组合条件界（*A1*）。

推论 D.4（组合条件总变差界）。对于 $c_J = \log J$，

$$d_{\mathrm{TV}}(\mathcal{L}(S_J \mid \alpha),\ \mathrm{Pois}(\alpha \log J)) \le E_1(\alpha) + E_2(\alpha), \tag{D.14}$$

其中 $E_1(\alpha)$ 满足（D.10）且

$$E_2(\alpha) = d_{\mathrm{TV}}(\mathrm{Pois}(\lambda_J(\alpha)),\ \mathrm{Pois}(\alpha\log J)) \le \sqrt{\tfrac{1}{2}\,\mathrm{KL}(\mathrm{Pois}(\lambda_J(\alpha)) \,\|\, \mathrm{Pois}(\alpha\log J))}. \tag{D.15}$$

泊松 *KL* 散度有明确表达式：

$$\mathrm{KL}(\mathrm{Pois}(\lambda) \,\|\, \mathrm{Pois}(\lambda')) = \lambda \log\left(\frac{\lambda}{\lambda'}\right) + \lambda' - \lambda, \qquad \lambda, \lambda' > 0. \tag{D.16}$$

证明。(D.14) 是使用中间项 $\mathrm{Pois}(\lambda_J(\alpha))$ 的三角不等式。(D.15) 是将 Pinsker 不等式应用于两个泊松分布，而 (D.16) 是泊松分布的标准 KL 公式。□

率参数解释。对于固定的 α 以及中等/较大的 J：定理 D.9 意味着$E_1(\alpha) = O(1/\log J)$。定理 D.10 意味着 $\lambda_J(\alpha) = \alpha \log J + O(\alpha)$，因此 (D.16) 中的泊松分布–泊松分布 KL 通常为 $O(\alpha/\log J)$，从而得到 $E_2(\alpha) = O\left(\sqrt{\alpha}/\log J\right)$。因此，当 J 处于几十这一量级时，均值线性化项通常会主导分布误差，除非 α 极其小。

D.4. **边缘（伽马混合）界。**

D.4.1. 边缘总变差界。

定理 D.11 （在 $\alpha \sim$ 伽马分布(a, b) 下的边缘总变差界）。 设 $p_{a,b}$ 和 $\tilde{p}_{a,b}$分别为 $S_J = K_J - 1$ 在 $\alpha \sim$ 伽马分布(a, b) 下的精确边缘概率质量函数和 *A1* 边缘概率质量函数，其中$c_J =$对数 J。则

$$d_{\mathrm{TV}}(p_{a,b}, \tilde{p}_{a,b}) \le \mathbb{E}_\alpha[E_1(\alpha)] + \mathbb{E}_\alpha[E_2(\alpha)], \tag{D.17}$$

其中 $E_1(\alpha)$ 和 $E_2(\alpha)$ 在定义 *D.5* 中给出，并由(D.10) 和 (D.15)界定。此外，若 $\mathbb{E}[\sqrt{\alpha}] < \infty$ （等价地， $a > 1/2$），则

$$\mathbb{E}_\alpha[E_1(\alpha)] = O\left(\frac{1}{\log J}\right), \qquad \mathbb{E}_\alpha[E_2(\alpha)] = O\left(\frac{\mathbb{E}[\sqrt{\alpha}]}{\sqrt{\log J}}\right) \quad (J \to \infty). \tag{D.18}$$

对于 $\alpha \sim$ 伽马分布(a, b),

$$\mathbb{E}[\sqrt{\alpha}] = \frac{\Gamma(a + 1/2)}{\Gamma(a)} \cdot \frac{1}{\sqrt{b}}. \tag{D.19}$$

证明。 (D.17) 正是 定理 D.8。(D.18) 中的缩放关系由 推论 D.4 中的逐点率结论以及 $\lambda_J(\alpha) \asymp \alpha$ 对数 J（当 J足够大时）和 $\sqrt{\alpha}$的可积性推出。最后，(D.19) 是伽马分布的标准半阶矩。 □

解释。 对 α 进行混合并不能消除有限-J 线性化效应：如果超先验在中等 α 上赋予不可忽略的概率质量，那么对于 $J \in [10, 100]$，均值线性化分量仍可能是分布差异的主要来源。

D.4.2. 矩层面的影响。

推论 D.5 (通过总变差耦合得到的矩推论)。 设 P 和 Q 是定义在 $\{1, 2, \ldots, J\}$上的分布。那么，对于任意有界函数 f,

$$|\mathbb{E}_P[f(K)] - \mathbb{E}_Q[f(K)]| \le 2\|f\|_\infty \, d_{\mathrm{TV}}(P, Q). \tag{D.20}$$

*App*将其与 $f(K) = K$和 $f(K) = K^2$结合，可得到保守界

$$|\mathbb{E}[K_J \mid a, b] - \tilde{\mu}_K| \le 2J\, d_{\mathrm{TV}}(p_{a,b}, \tilde{p}_{a,b}), \qquad \left|\mathrm{Var}(K_J \mid a, b) - \tilde{\sigma}_K^2\right| \le 4J^2\, d_{\mathrm{TV}}(p_{a,b}, \tilde{p}_{a,b}), \tag{D.21}$$

其中 $(\tilde{\mu}_K$、 $\tilde{\sigma}_K^2)$是 *A1*（负二项分布）的代理矩。

备注。(D.21) 是有意采用的最坏情况界；在典型的多站点应用中， K_J集中在较小的取值上，因此实际矩差异通常远小于这些界所暗示的程度。推论 D.5 的主要用途是将分布容差转换为矩误差的保守界。

D.5. **何时需要 A2? 实用阈值条件**。上述界形式化地说明了为什么 A1 在 J 较小或中等时可能不可靠：泊松化项（$\sim 1/\log J$）和均值线性化项（通常为$\sim 1/\sqrt{\log J}$）都会随 J 缓慢衰减。

命题 D.5 (经验阈值)。假设我们希望 *A1* 近似中的主导（对数）误差项达到较小的容差水平 $\epsilon > 0$ 。一个简单的粗略充分条件可写为

$$\frac{a/b}{\log J} \lesssim \epsilon \qquad \Longleftrightarrow \qquad J \gtrsim \exp\!\left(\frac{a}{b\,\epsilon}\right) \tag{D.22}$$

这表明，为了使阶段 *1* 在统一的分布层面上足够准确，可能需要非常大的 J 。

备注。 命题 D.5 有意采用保守设定，并省略常数以及 E_2 项的贡献；列出该命题仅为提供一种易于理解的缩放规律。实际决策应基于有限 J 情形，具体总结如下。

表 D.1. 阶段 1 与阶段 2 的实用指南

情形	J 范围	建议
较大 J，用于探索	$J \ge 100$	单独使用 A1 作为快速初始值设定器
中等 J，可发表的分析	$J \in [30, 100]$	建议使用 A1 +A2
较小 J，低信息设置	$J \in [10, 30]$	A2 必需
非常小 J	$J < 10$	优先采用精确枚举或分布层级校准（例如，A2-KL）

实用决策规则。 默认推荐阶段 2（A2）；单独使用阶段 1（A1）主要适用于快速初始化，或在 J 明显较大时进行粗略探索。

附录 附录 E. 权重分布理论和双锚框架

E.1. **目的与范围。** 本附录提出基于权重的锚点，与章节 A 至C中确立的簇数校准相辅相成。两阶段矩匹配（TSMM）算法能够高效地将从业者对 K_J 的信念转化为伽马超参数，而 Vicentini 和 Jermyn（2025年）证明，仅依据簇数校准的先验可能会对截断棒表示的权重造成“意料之外”的影响——例如，在 K_J中看似“弥散”的先验，仍可能隐含着某个簇主导混合的强假设。

本附录通过以下方式回应这一问题：

1. 推导第一个断棒权重的闭式分布 w_1（在伽马超先验下，定理 E.12），从而能够快速计算主导性诊断；2. 建立共同聚类概率的矩结构 $\rho = \sum_{h=1}^{\infty} w_h^2$（命题 E.7和E.8），为权重泛函提供一种具有直接应用解释的替代方案；3. 形式化双锚校准方案（定义 E.8、算法 E.1

)

使从业者能够在诊断表明其行为在实质上不可接受时，约束权重主导。

归属说明（经典与新颖内容）。Sethuraman 断棒表示和 GEM（按大小加权）排序属于经典内容（Pitman，2006年；Sethuraman，1994）。在伽马混合下，w_1 的闭式边际分布（定理 E.12）见 Vicentini 和 Jermyn（2025年，附录 A），此处转载是因为它能够实现 O（1）主导性诊断。本文特有的是条件设计校准工作流，该工作流将 K_J 锚点（章节 A至C）与这些权重诊断相结合，并由此形成双锚细化选项（§E.5）。

全文沿用正文及前面附录中确立的记号： J 表示固定设计规模， $\alpha \sim$ 伽马分布(a， b)采用形状–率参数化， K_J表示已占据簇的数量，而(w_1， w_2，.....)表示按 GEM 顺序排列的棍棒折断权重。

E.2. **断棒权重与按大小加权解释。**

E.2.1. ***GEM*分布及其构造。** 我们回顾 Sethuraman 的截断棒表示。狄利克雷过程 $G \sim \mathrm{DP}(\alpha,\ G_0)$ 具有如下表示（Sethuraman，1994）：

$$G = \sum_{h=1}^{\infty} w_h\, \delta_{\theta_h^*}, \qquad \theta_h^* \overset{iid}{\sim} G_0, \tag{E.1}$$

其中权重 $(w_1, w_2, \ldots)$ 按如下方式构造。

定义 E.6 (GEM分布). 设 $v_h \overset{iid}{\sim}$服从贝塔分布$(1, \alpha)$，$h = 1, 2, \ldots$。定义

$$w_1 := v_1, \qquad w_h := v_h \prod_{\ell=1}^{h-1}(1 - v_\ell) \quad \text{for } h \geq 2. \tag{E.2}$$

序列 $(w_1, w_2, \ldots)$ 服从 $\mathrm{GEM}(\alpha)$ 分布（Griffiths–Engen–McCloskey；参见 Pitman，2006年）。

式（E.2）中的构造具有“截断棒表示”解释：从一根单位棒开始，我们截取比例 v_1 以形成 w_1，再从剩余部分截取比例 v_2 以形成 w_2，依此类推。由构造可知，$\sum_{h=1}^{\infty} w_h = 1$ 几乎必然成立。

E.2.2. 关键注意事项：大小偏倚顺序。可解释性警告。 序列 $(w_1, w_2, \ldots)$ 处于大小偏倚（GEM）顺序，而非从大到小排序：

$$(w_1, w_2, \ldots) \neq (w_{(1)}, w_{(2)}, \ldots) \quad \text{where } w_{(1)} \geq w_{(2)} \geq \cdots. \tag{E.3}$$

权重 w_1 表示随机选取的单位所属簇的渐近比例——等价地，即从泊松分布–Dirichlet分布中按大小加权抽取的结果（Pitman，1996年）——而不必然是最大簇份额。

这一区分对于应用中的 elicitation 至关重要。当从业者表达关于“主导簇比例”的信念时，他们通常指的是 $\max_h w_h = w_{(1)}$，而非按大小加权的 w_1。关系概率 $(\max_h w_h > t) \geq$ 概率$(w_1 > t)$成立，因此基于 w_1 的诊断为主导风险提供了一个保守界：若概率$(w_1 > t)$较小，则概率$\big(w_{(1)} > t\big)$ 也同样较小。

Despite这种大小偏倚解释，w_1 仍是一个有用的诊断量，因为：

1. 在伽马超先验下，它具有闭式分布（定理 E.12），无需模拟即可快速计算；
2. 它为主导风险提供了具有数学依据的下界；3. 关键的是，若 $w_1 > 0.5$，则 w_1 必然是最大的权重（即$w_1 = w_{(1)}$），因为当一个权重已达到0.5时，其他权重都不可能超过它。因此，尾概率概率$(w_1 > 0.5)$ 是衡量多数主导风险的精确指标。

E.2.3. 的条件分布。 根据棍棒折断构造（E.2），$w_1 = v_1$ 其中 $v_1 \sim$ 贝塔分布(1, α)。

命题 E.6 （ w_1的条件分布）。 给定 $\alpha > 0$,

$$w_1 \mid \alpha \sim \mathrm{Beta}(1, \alpha). \tag{E.4}$$

密度、累积分布函数和矩为:

$$p(w_1 \mid \alpha) = \alpha(1 - w_1)^{\alpha-1}, \qquad \Pr(w_1 \le x \mid \alpha) = 1 - (1 - x)^{\alpha}, \tag{E.5}$$

$$\mathbb{E}[w_1 \mid \alpha] = \frac{1}{1+\alpha}, \qquad \mathrm{Var}(w_1 \mid \alpha) = \frac{\alpha}{(1+\alpha)^2(2+\alpha)}. \tag{E.6}$$

证明。 由 $w_1 = v_1$ 和 $v_1 \sim$ 贝塔分布(1, α)可立即得到。 □

解释。 当 $\alpha \to 0$, $\mathbb{E}[w_1 \mid \alpha] \to 1$ 时：第一个簇几乎吸收全部概率质量。当 $\alpha \to \infty$, $\mathbb{E}[w_1 \mid \alpha] \to 0$时：概率质量在许多簇之间变得分散。这种 α 与期望支配度之间的单调关系，是基于权重的诊断的基础。

E.3. 的闭式边缘分布 w_1。 当 α 本身服从带有伽马超先验的随机分布时， w_1 的无条件（边缘）分布具有闭式表达式。这一关键结果无需进行蒙特卡洛模拟即可快速计算主导性诊断。

E.3.1. 主要定理。

定理 E.12 (w_1 的无条件分布)。设 $\alpha \sim$ 伽马分布(a, b) 的密度为 $g_{a,\ b}(\alpha) = \frac{b^a}{\Gamma\ a\ ()}$ $\alpha^{a-1}e^{-b\alpha}$ ，其中 $\alpha \to 0$。则 w_1 具有以下闭式分布:

(i) 密度:

$$p(w_1 \mid a, b) = \frac{a\, b^a}{(1 - w_1)[b - \log(1 - w_1)]^{a+1}}, \qquad w_1 \in (0, 1). \tag{E.7}$$

(ii) 累积分布函数:

$$F_{w_1}(x \mid a, b) = \Pr(w_1 \le x \mid a, b) = 1 - \left(\frac{b}{b - \log(1 - x)}\right)^a, \qquad x \in [0, 1)。 \tag{E.8}$$

(iii) 生存（尾部）概率:

$$\boxed{\Pr(w_1 > t \mid a, b) = \left(\frac{b}{b - \log(1 - t)}\right)^a} \tag{E.9}$$

(iv) 分位数函数：对于 $u \in (0,1)$**，**

$$Q_{w_1}(u \mid a, b) = 1 - \exp\!\big(b\,[1-(1-u)^{-1/a}]\big)\,. \tag{E.10}$$

出处说明。 密度和累积分布函数公式见 Vicentini 和 Jermyn（2025年，附录 A），他们在样本量无关的先验设定背景下推导了这些公式。为求完整起见，我们给出一个自成一体的证明。

证明。 我们对条件分布中的 α 积分消去：

$$\begin{aligned} p(w_1 \mid a, b) &= \int_0^\infty p(w_1 \mid \alpha)\, g_{a,b}(\alpha)\, d\alpha \\ &= \int_0^\infty \alpha(1-w_1)^{\alpha-1} \cdot \frac{b^a}{\Gamma(a)} \alpha^{a-1} e^{-b\alpha}\, d\alpha \\ &= \frac{b^a}{\Gamma(a)(1-w_1)} \int_0^\infty \alpha^a \exp\!\big[-\alpha\big(b - \log(1-w_1)\big)\big]\, d\alpha. \end{aligned} \tag{E.11}$$

定义 $\beta := b - \log(1-w_1)$。由于 $\log(1-w_1) < 0$ 当 $w_1 \in (0,1)$时，我们有$\beta > b > 0$。(E.11) 中的积分是一个伽马积分：

$$\int_0^\infty \alpha^a e^{-\beta\alpha}\, d\alpha = \frac{\Gamma(a+1)}{\beta^{a+1}}. \tag{E.12}$$

将 (E.12) 代入 (E.11)，并利用 $\Gamma(a+1) = a\,\Gamma(a)$：

$$p(w_1 \mid a, b) = \frac{b^a}{\Gamma(a)(1-w_1)} \cdot \frac{a\,\Gamma(a)}{[b - \log(1-w_1)]^{a+1}} = \frac{a\,b^a}{(1-w_1)[b-\log(1-w_1)]^{a+1}}.$$

由此得到 (E.7)。

对于累积分布函数，我们对密度积分。令 $u = b - \log(1-w)$，则 $du = dw/(1-w)$：

$$\begin{aligned} F_{w_1}(x \mid a, b) &= \int_0^x \frac{a\,b^a}{(1-w)[b-\log(1-w)]^{a+1}}\, dw \\ &= a\,b^a \int_b^{b-\log(1-x)} u^{-a-1}\, du \\ &= a\,b^a \left[-\frac{u^{-a}}{a}\right]_b^{b-\log(1-x)} \\ &= b^a \left[b^{-a} - [b-\log(1-x)]^{-a}\right] = 1 - \left(\frac{b}{b-\log(1-x)}\right)^a. \end{aligned}$$

生存函数（E.9）为 $1 - F_{w_1}(t)$，分位数函数（E.10）通过求解 $F_{w_1}(Q) = u$ 得到，其中 Q。 □

通过拉普拉斯变换进行另一种推导。生存概率（E.9）可以更直接地推导得出。根据 命题 E.6，概率$(w_1 > t \mid \alpha) = (1-t)^\alpha = \exp(-c_t\alpha)$其中 $c_t := -\log(1-t) > 0$。然后：

$$\Pr(w_1 > t \mid a, b) = \mathbb{E}_\alpha[\exp(-c_t\alpha)] = \left(\frac{b}{b+c_t}\right)^a = \left(\frac{b}{b-\log(1-t)}\right)^a,$$

利用伽马分布$(a,\ b)$随机变量的拉普拉斯变换。 □

E.3.2. 极限行为与一致性检查。生存概率公式（E.9）满足预期的边界条件。

As $t \downarrow 0$: 概率$(w_1 > t) = \big(b/[b-\log(1-t)]\big)^a \to (b/b)^a = 1.$

当 $t \uparrow 1$: $\log(1-t) \to -\infty$，因此 $b - \log(1-t) \to +\infty$，所以 $\Pr(w_1 > t) \to 0$。

这些极限证实，在伽马超先验下，w_1 几乎必然具有支撑集 $(0,\ 1)$。

E.3.3. 基于梯度的优化所需的导数。对于双锚点细化（§E.5），$\Pr(w_1 > t \mid a, b)$ 关于 (a, b) 的梯度很有用。定义 $c_t := -\log(1-t) > 0$ 以及 $S(t;\ a, b) := \Pr(w_1 > t \mid a,\ b) = (b/(b+c_t))^a$。于是：

$$\frac{\partial S}{\partial a} = S(t; a, b) \cdot \log\left(\frac{b}{b+c_t}\right), \tag{E.13}$$

$$\frac{\partial S}{\partial b} = S(t; a, b) \cdot \frac{a\, c_t}{b(b+c_t)}. \tag{E.14}$$

由于 $\log(b/(b+c_t)) < 0$，我们有 $\partial S/\partial a < 0$：增大形状参数 a（使α先验更加集中）会降低支配概率。同样，$\partial S/\partial b > 0$：增大率参数 b（使 α 向较小取值移动）会提高支配概率。

E.3.4. 默认诊断阈值。

推论 E.6 (默认诊断阈值)**.** 以下阈值提供了易于解读的支配性基准：

- 概率$(w_1 > 0.5 \mid a, b)$：多数主导的概率（按大小偏置的簇包含超过一半的单位）。由于 $w_1 > 0.5$ 蕴含 $w_1 = w_{(1)}$，这也就是最大簇超过50%的概率。

- 概率$(w_1 > 0.9 \mid a, b)$：近乎普遍支配的概率（某个簇包含至少90%的单位）。

这些阈值可通过以下方式即时计算(E.9)，并提供正文算法框 *1* 中建议的诊断量。

E.4. 共同聚类概率 ρ。 第一个权重 w_1 刻画单个按大小加权的簇。另一个用于概括整体权重集中度的函数量是共同聚类概率，它衡量随机选取的两个单位属于同一簇的概率。

E.4.1. 定义与应用解释。

定义 E.7 （共同聚类概率／辛普森指数）。共同聚类概率为

$$\rho := \sum_{h=1}^{\infty} w_h^2. \tag{E.15}$$

等价地，在给定权重 (w_h) 的条件下，若 Z_i, Z_j 是独立的簇分配，且概率 $\Pr(Z = h \mid \mathbf{w})$ 为 $= w_h$，则

$$\rho = \Pr(Z_i = Z_j \mid \mathbf{w}). \tag{E.16}$$

应用解释。 量 ρ 直接回答了这个问题：“随机抽取两个单位，它们属于同一个潜在簇的概率是多少？”对于从业者而言，与询问单个权重相关的问题相比，这种表述通常更直观。

共同聚类概率有时在生态学中称为辛普森指数，用来衡量随机选取的两个个体属于同一物种的概率。

E.4.2. 的条件矩 ρ。 GEM 结构使得可以递归计算 ρ的矩。

命题 E.7 （ ρ 的条件矩）。 对于 $\alpha > 0$:

(i) 一阶矩:

$$\mathbb{E}[\rho \mid \alpha] = \frac{1}{1+\alpha}. \tag{E.17}$$

(ii) 二阶矩:

$$\mathbb{E}[\rho^2 \mid \alpha] = \frac{\alpha+6}{(\alpha+1)(\alpha+2)(\alpha+3)}. \tag{E.18}$$

(iii) 方差:

$$\mathrm{Var}(\rho \mid \alpha) = \frac{2\alpha}{(\alpha+1)^2(\alpha+2)(\alpha+3)}. \tag{E.19}$$

出处说明。 这些结果是经典的；参见 Kingman（1975）和 Pitman（1996年, 2006年）关于泊松–狄利克雷分布的一般理论。

证明概要。 根据棍棒折断构造， ρ 满足该分布递推

$$\rho \overset{d}{=} v^2 + (1-v)^2\rho', \tag{E.20}$$

其中 $v \sim$ 贝塔分布$(1, \alpha)$ 且 ρ' 是 ρ的一个独立副本。取期望并求解不动点方程，可得 $\mathbb{E}[\rho] = 1/(1+\alpha)$。二阶矩可通过对（E.20）两边平方并进行类似运算得到。□

关键观察。 比较（E.17）与（E.6）：

$$\mathbb{E}[\rho \mid \alpha] = \mathbb{E}[w_1 \mid \alpha] = \frac{1}{1+\alpha}. \tag{E.21}$$

因此，预期共同聚类概率等于预期第一个权重。这个恒等式简化了 elicitation：向从业者询问共同聚类情况，得到的信息可直接用于基于 w_1的校准。

E.4.3. 伽马超先验下的边缘矩。

命题 E.8 （ ρ的边际均值）。 在 $\alpha \sim$ 伽马分布(a, b)下：

$$\mathbb{E}[\rho \mid a, b] = \mathbb{E}\left[\frac{1}{1+\alpha}\right] = I_1(a, b), \tag{E.22}$$

其中，函数 $I_c(a, b)$：$= \mathbb{E}[(\alpha + c)^{-1}]$ 对于 $c > 0$ 定义如下：

$$I_c(a, b) = b^a\, c^{a-1}\, e^{bc}\, \Gamma(1-a, bc), \tag{E.23}$$

其中 $\Gamma(s, x) = \int_x^\infty t^{s-1}e^{-t}\, dt$ 表示上不完全伽马函数。

对于特殊情形 $c = 1$:

$$\mathbb{E}[\rho \mid a, b] = \mathbb{E}[w_1 \mid a, b] = b^a e^b \Gamma(1-a, b). \tag{E.24}$$

出处说明。 式（E.23）所依据的积分恒等式是标准结果（Abramowitz & Stegun，1964，§6.5）。该恒等式在狄利克雷过程权重泛函中的应用见 Vicentini 和 Jermyn（2025年）。

证明。 我们计算

$$I_c(a, b) = \int_0^\infty \frac{1}{\alpha + c} \cdot \frac{b^a}{\Gamma(a)} \alpha^{a-1} e^{-b\alpha}\, d\alpha = \frac{b^a}{\Gamma(a)} \int_0^\infty \frac{\alpha^{a-1}e^{-b\alpha}}{\alpha + c}\, d\alpha. \tag{E.25}$$

利用积分恒等式（Abramowitz & Stegun，1964，公式 6.5.3）：

$$\int_0^\infty \frac{x^{a-1}e^{-bx}}{x + c}\, dx = c^{a-1}e^{bc}\Gamma(a)\Gamma(1-a, bc), \tag{E.26}$$

我们得到

$$I_c(a,b) = \frac{b^a}{\Gamma(a)} \cdot c^{a-1} e^{bc} \Gamma(a) \Gamma(1-a, bc) = b^a c^{a-1} e^{bc} \Gamma(1-a, bc). \qquad \square$$

计算说明。 对于 $a < 1$，公式（E.23）可直接使用标准的不完全伽马函数实现进行计算。对于 $a \geq 1$，上不完全伽马函数 $\Gamma(1-a, bc)$ 需要进行解析延拓，数值库会自动处理此过程（例如，R 中的 `pgamma` 、Python 中的 `scipy.special.gammaincc` ）。

实用数值计算方案（基于求积）。 若要稳定地计算 $I_c(a,b)$，而不依赖特殊函数的解析延拓：(1) 构造广义高斯–拉盖尔节点/权重 $\{(x_m, \tilde{w}_m)\}_{m=1}^{M}$，对应权重函数 $x^{a-1}e^{-x}$ (A 节，§A.2.6)；(2) 映射 $\alpha_m = x_m/b$；(3) 近似计算 $I_c(a,b) \approx \sum_{m=1}^{M} \tilde{w}_m/(\alpha_m + c)$。对多个期望（例如 K_J 矩和 ρ 诊断量）使用相同的节点，可显著节省计算开销。

E.4.4. Unc的无条件二阶矩 ρ。 利用部分分式分解

$$\frac{\alpha+6}{(\alpha+1)(\alpha+2)(\alpha+3)} = \frac{5}{2(\alpha+1)} - \frac{4}{\alpha+2} + \frac{3}{2(\alpha+3)}, \tag{E.27}$$

我们得到：

推论 E.7 （ ρ 的无条件二阶矩）。 在 $\alpha \sim$ 伽马分布$(a,\ b)$ ：

$$\mathbb{E}[\rho^2 \mid a,b] = \tfrac{5}{2} I_1(a,b) - 4\, I_2(a,b) + \tfrac{3}{2} I_3(a,b), \tag{E.28}$$

因此

$$\mathrm{Var}(\rho \mid a,b) = \mathbb{E}[\rho^2 \mid a,b] - \left[\mathbb{E}[\rho \mid a,b]\right]^2. \tag{E.29}$$

设计启示。 在伽马分布超先验下，ρ 的分布没有简单的闭式形式，但其均值和方差可通过 I_c 泛函得到闭式结果（“矩闭式”）。这足以用于基于矩的诊断和校准。

E.5. **双锚定校准框架**。

E.5.1. 动机与设计原则。 TSMM 算法（C节）生成超参数 (a_K, b_K)，使其与为簇数 elicited 的矩 (μ_K, σ_K^2) 相匹配K_J。然而，正如 Vicentini 和 Jermyn（2025年）所示，这类 K_J-校准先验可能导致权重行为与实质性预期相悖——例如，即便从业者预期“有几个规模适中的组”，某个簇仍可能以很高的概率主导混合。

双锚框架通过以下方式解决这一问题：(1) 利用 Pr($w_1 > t$)诊断 (a_K, b_K) 所隐含的主导风险；(2) 当诊断量超出从业者指定的容差时，调整超参数；(3) 报告簇数量保真度与权重约束满足度之间的明确权衡。

E.5.2. 形式化问题陈述。

定义 E.8 （双锚点目标函数）。 给定：

- 锚点*1*（簇数）：目标矩（μ_K， σ_K^2），对应于 K_J；或等价地，TSMM 解（a_K， b_K）；
- 锚点*2*（权重约束）：支配阈值 $t \in$ (0，1) 和容差$\delta \in$ (0，1)，使得 Pr($w_1 > t$) $\leq \delta$ 符合预期。

双锚点解在这两个锚点之间取得平衡：

$$(a^*, b^*) = \arg \min_{a>0,\, b>0} \mathcal{L}_\lambda(a, b), \tag{E.30}$$

其中目标函数为

$$\mathcal{L}_\lambda(a, b) = \lambda \cdot D_1(a, b) + (1 - \lambda) \cdot D_2(a, b), \tag{E.31}$$

其中：

- $D_1(a, b) = \left(\frac{\mathbb{E}[K_J|a,b] - \mu_K}{\mu_K}\right)^2 + \left(\frac{\mathrm{Var}(K_J|a,b) - \sigma_K^2}{\sigma_K^2}\right)^2$ (normalized moment discrepancy);
- $D_2(a, b) = \max\big(0,\, \Pr(w_1 > t \mid a, b) - \delta\big)^2$ (dominance penalty);
- $\lambda \in [0, 1]$ controls the relative importance of the two anchors.

边界情况。 当 $\lambda = 1$ 时，问题简化为纯粹的 K_J 校准（TSMM）。当 $\lambda = 0$ 时，问题完全忽略簇数，仅最小化主导性——对于固定-J 应用而言，这种极端做法并不推荐，因为 K_J 信念具有实际意义。

E.5.3. 帕累托前沿可视化。 改变 λ 在（E.31）中的取值，可描绘簇数量保真度与主导风险之间的帕累托前沿：（1）选择网格 $\lambda_1, \ldots, \lambda_G \subset [0, 1]$（例如， $\lambda_g = g/$（$G - 1$））；（2）对每个 λ_g，通过算法 E.1 求解（E.30）；（3）绘制所得点对 $\big(D_1$（$a^*_{\lambda_g}$， $b^*_{\lambda_g}$），概率 $\Big(w_1 > t \mid a^*_{\lambda_g}, b^*_{\lambda_g}\Big)$，并可选择根据诱导的变化为各点添加注释，标明其对 $\mathbb{E}[K_J]$ 和 方差(K_J)的影响。

该帕累托图展示了何时对 K_J 拟合进行小幅放宽即可显著降低主导风险，从而引导从业者做出知情的权衡决策。

图 E.1 展示了计算示例中的这种权衡（$J = 50$， $\mu_K = 5$，中等置信度）。前沿呈现出清晰的“肘部”结构：适度放宽

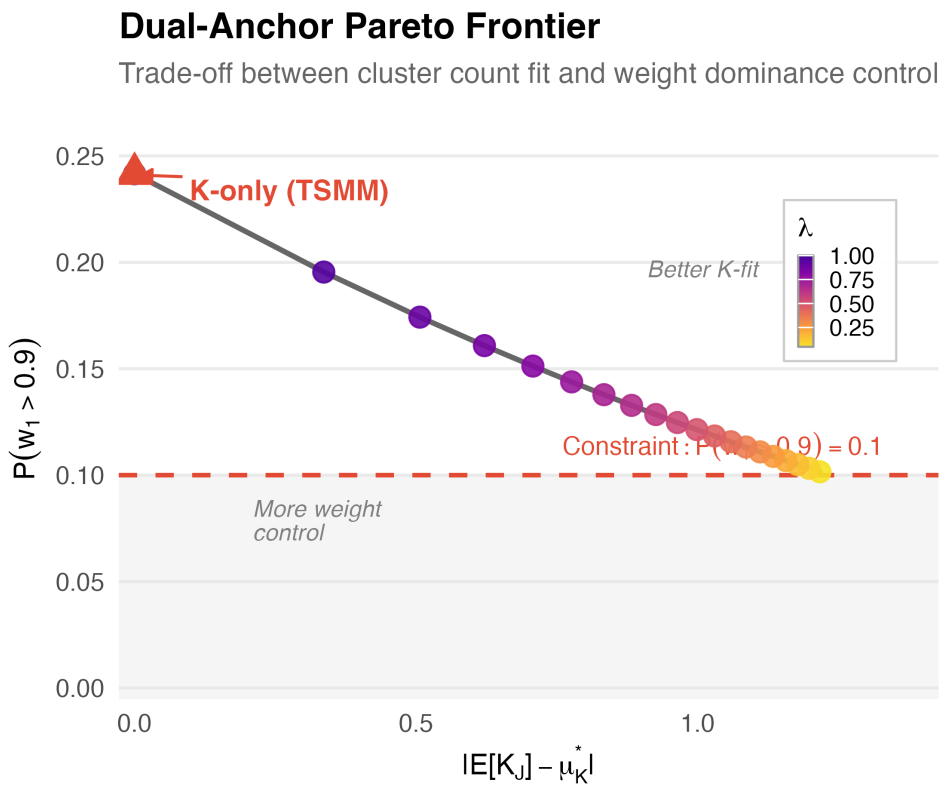


图 E.1. 双锚帕累托前沿（$J = 50$, $\mu_K = 5$）。x轴显示 $|\mathbb{E}[K_J] - \mu_K^*|$; y轴显示 $\Pr(w_1 > 0.9)$。每个点表示在特定 λ 值（颜色刻度）下的一个校准超先验。红色三角形标示仅针对 K的 TSMM 解（$\lambda = 1$）；水平虚线表示约束 $\Pr(w_1 > 0.9) = 0.10$；灰色阴影区域满足此约束。

K_J 保真度（与 μ_K的偏差约为 1 个单位）可显著降低主导性，使 $\Pr(w_1 > 0.9)$ 从 0.24 降至目标约束 0.10。超过这个肘部后，进一步放宽带来的收益会递减。这一可视化有助于从业者选择能够体现其实质性优先事项的 a λ 值——无论是精确匹配簇数量，还是稳健地防止意外出现单簇主导。

E.5.4. 算法。以下描述双锚精化协议。

算法 E.1（双锚细化协议）。

输入： 匹配目标矩（a_K， b_K）的 TSMM 解（μ_K， σ_K^2）；支配阈值 t 和容差 δ；权衡参数 λ；下界（$a_{\min}$， $b_{\min}$）。默认值：$t = 0.5$、 $\delta = 0.25$、 $\lambda = 0.7$、$a_{\min} = b_{\min} = 0.01$。

输出： 细化后的超参数（a^*， b^*）及权衡报告。

1. **在 TSMM 解处计算诊断量：**

$$p_0 = \Pr(w_1 > t \mid a_K, b_K) = \left(\frac{b_K}{b_K - \log(1-t)} \right)^{a_K}.$$

2. 检查约束是否已满足。 如果 $p_0 \le \delta$：返回（a_K， b_K），状态为“在 TSMM 解处已满足主导性约束”。无需细化。

3. **否则，执行约束优化：**

(a) 根据 (E.31) 定义双锚点目标函数 $\mathcal{L}_\lambda(a, b)$。(b) 参数化：优化 $\theta = (\log a, \log b)$，以自动满足正值约束。(c) 计算梯度： $\nabla D_1(a, b)$ 使用 C节中 $(\mathbb{E}[K_J], \mathrm{Var}(K_J))$ 的雅可比矩阵（通过求积计算得分恒等式）；当 $\Pr(w_1 > t) > \delta$ 时， $\nabla D_2(a, b)$ 使用 (E.13)–(E.14)，否则为零。(d) 将优化器初始化为 (a_K, b_K)。(e) 使用盒约束 $a, b \ge a_{\min}$ 运行 L-BFGS-B（Nocedal 与 Wright，2006）。(f) 当 $\|\nabla \mathcal{L}_\lambda\|_\infty < 10^{-6}$ 或 $|\mathcal{L}_\lambda^{(\mathrm{new})} - \mathcal{L}_\lambda^{(\mathrm{old})}| < 10^{-8}$ 时停止。

4. 报告权衡： 期望簇数的变化 $\Delta\mu_K = \mathbb{E}[K_J \mid a^*, b^*] - \mu_K$；簇数方差的变化 $\Delta\sigma_K^2 =$ 方差$(K_J \mid a^*, b^*) - \sigma_K^2$；新的主导性概率 概率$(w_1 > t \mid a^*, b^*)$；状态：若 概率$(w_1 > t \mid a^*, b^*) \le \delta$，则为“满足约束”，否则为“帕累托最优折中”。

桌子 E.1. 推荐的双锚默认值

参数	默认值	理由
阈值 t	0.5	多数主导性——最容易解释的阈值；当 $w_1 > 0.5$，这必然是最大簇
容差 δ	0.25	多数主导性的先验概率不超过 25%- 融资
权衡 λ	0.7	优先保证簇数量保真度，但允许均值-有意义的权重调整

E.5.5. 推荐默认值。 选择这些默认值的依据是：(1) $t = 0.5$代表“多数主导”——这是对从业者而言最容易解读的阈值，也是按大小加权排序与排名排序相一致的点；(2) $\delta = 0.25$ 提供了有意义的约束，同时避免要求过于严格；(3) $\lambda = 0.7$体现了设计条件理念，即簇数仍是主要锚点。

E.5.6. 失效模式和边界情况。不可行性。在某些配置下，不存在任何 $(a,\ b)$ 能够同时精确匹配 $(\mu_K,\ \sigma_K^2)$ 并满足 $\Pr(w_1 > t) \leq \delta$。当征询得到的簇数目标意味着 α 的分布集中在较小取值附近时，就会出现这种情况，而这必然会导致较高的主导风险。在这种情况下，算法 E.1 会返回一个使加权目标函数最小化的帕累托最优解，并在权衡报告中明确量化这一折衷。

边界解。如果优化器将 $a \to 0$ 或 $b \to 0$ 推至边界，盒约束$a,\ b \geq 0.01$ 可防止出现退化解。这种边界行为表明锚点之间存在根本性的矛盾，因此应重新审视征询得到的目标值。

单一权重锚点下的可识别性。 双锚点目标函数使用平均主导概率，而非完整的分布约束。在有两个自由参数（$a,\ b$）和三个有效约束（两个簇数矩加上一个主导界限）的情况下，当主导界限处于激活状态时，问题通常约束过多，因此需要采用加权优化，而非精确满足所有约束。

E.6. **从业者解释指南**。

表 E.2. 主导风险解释

$\Pr(w_1 > 0.5)$	风险等级	解释	建议
< 0.20	Low	权重可能较为均衡	无需采取行动
0.20–0.40	中等	Some 多数 主导性风险	报告诊断；敏感性分析
0.40–0.60	显著	有意义的 主导性概率	双锚点 细化推荐的
> 0.60	High	先验偏向一个主导簇	双锚点至关重要；验证连贯性

E.6.1. 主导风险类别。

E.6.2. 报告清单。 本节为使用设计条件式 elicitation 的出版物提供详细的报告指南。该清单确保先验对划分和权重的影响得到明确说明，避免沿用“默认值”伽马先验时含糊不清。

使用 *DCE/TSMM* 时建议披露的信息。

1. 设计规模与背景。 报告 J，并简要说明为何将其视为固定值（例如“参与多中心试验的 $J = 47$ 个站点”）。2. *elicitation*目标。 报告 elicitation 得到的均值 $\mathbb{E}(K_J)$ 和方差 $\mathrm{Var}(K_J)$，或所采用的置信度到方差的映射。简要说明所选取值的实质性依据。3. 校准方法。 报告使用的仅是阶段 1 还是完整TSMM，并说明细化方法（例如精确矩牛顿法）及收敛容差。4. 所得超先验。 报告校准后的超参数 (a, b)，对应于 $\alpha \sim$伽马分布(a, b)，并注明参数化约定（形状–率）。5. 的先验预测摘要 K_J。 至少报告校准先验下诱导的 $\mathbb{E}(K_J)$、$\mathrm{Var}(K_J)$，以及 K_J 的中心区间（例如 90% 或 95%）。6. 权重诊断（必需）。 报告校准先验下的 $\Pr(w_1 > 0.5)$ 和 $\Pr(w_1 > 0.9)$。如果共同聚类具有实质性意义，也可以报告 $\mathbb{E}(\rho)$。请注意， w_1 按大小偏倚顺序排列（GEM），因此所报告的概率指随机选取的单位所属簇的质量，而非最大簇的质量。7.双锚点用法（如适用）。 报告触发该方法的支配容差、权衡参数 λ，以及 K_J 矩和支配概率的变化。8. 敏感性分析。 至少报告一种敏感性变体，例如针对σ_K^2采用不同置信水平、采用不同支配容差，或与标准基线超先验进行比较，并说明后验结论如何变化或保持稳定。9. 软件版本。 报告所用 DPprior软件包的版本，并提供分析代码的可复现链接（例如 GitHub 代码仓库或 OSM档案）。

示例报告陈述。

> “我们采用设计条件式 elicitation 方法，为浓度参数α 指定了伽马超先验。对于包含$J = 50$ 个站点的样本，我们以期望簇数为 $\mathbb{E}(K_{50}) = 5$为目标，并设定中等置信度，对应于 $\mathrm{Var}(K_{50}) = 10$。两阶段矩匹配（TSMM）得到超参数 $\alpha \sim \mathrm{Gamma}(1.41, 1.08)$（形状–率）。诊断检查显示，簇数的先验预测 95% 区间为 $[1, 11]$，主导概率为 $\Pr(w_1 > 0.5) = 0.50$，且 $\Pr(w_1 > 0.9) = 0.20$。由于主导风险超过了我们设定的 0.40 容差，我们采用了带有$\lambda = 0.70$的双锚点细化，将 $\Pr(w_1 > 0.5)$ 降至 0.25，同时使 $\mathbb{E}(K_{50})$ 变为 6.2。分析使用 DPprior v1.0 完成。”

实践考量。

这份检查清单引出两点评论。首先，权重诊断是常规操作，而非例外。即使分析人员无意采用双锚点，报告$\Pr(w_1 > 0.5)$ 和$\Pr(w_1 > 0.9)$ 也能使隐含的汇聚几何结构一目了然，并直接回应 SSI 批评的核心关切（Vicentini与Jermyn，2025年）。其次，*TSMM* 支持前瞻性设定。TSMM 的计算效率（每次校准约 50 毫秒）使得将先验 elicitation 和敏感性检查纳入预分析计划切实可行，从而为以兼具实质依据和可操作、可复现的方式陈述α 先验提供了路径。

附录 F. 扩展模拟研究结果

F.1. **目的与范围。** 本附录报告完整的模拟研究结果，扩展了正文（§4）所呈现的摘要。具体而言，我们提供：设计和数据生成过程（DGP）的全因子设定（包括信息性校准目标）；各方法所用的确切超先验参数（包括置信水平映射）；正文中未展示的其他指标的完整性能表（例如，95% 区间宽度、 α的后验均值、基于众数的恢复率）；由 Lee 等人（2025年）提出的 χ^2-DORO 基线结果；关于置信度（方差）设定的敏感性分析；以及凸显 TSMM 和双锚点校准实际效率的计算基准测试。

全文关注的目标估计量是使用 $J = 100$ 个站点拟合 DP混合模型时的有限样本占据簇数 K_J。DGP 中使用的“真实”簇数 K^*

被视为校准基准——而非保证能够完美恢复的依据——因为即使模型设定正确，重叠成分和有限抽样变异也可能妨碍精确恢复。

F.2. **全因子设计规范**。

F.2.1. 设计网格。 表 F.1总结了全因子设计。

表 F.1. 全因子设计规范

因子	水平	取值
真实簇数（K^*）	3	5、10、30
信息性（I）	3	0.2、0.5、0.8
先验方法	5	模糊；DORO-Unif；TSMM；对偶-锚点；χ^2-DORO
置信水平	3	高；中；低（TSMM/对偶仅）
每个单元的重复试验次数	—	$R = 200$

置信度因子仅对TSMM和双锚点有意义，因为它决定了 K_J 的目标离散度（通过方差膨胀因子映射确定；见§F.3）。对于模糊、DORO-Unif 和 χ^2-DORO，每个 μ_K使用单一校准。

在实际操作中，每个（K^*， I）情景包含 9 种拟合的超先验变体：模糊（1 种）、DORO-Unif（1 种）、 χ^2-DORO（1 种）、置信度为 $\in$ {低、中、高} 的 TSMM（3 种），以及置信度为 $\in$ {低、中、高} 的双锚点（3 种）。在 9 个情景和$R = 200$次重复实验下，完整运行拟合了 $9 \times 200 \times 9 = 16$，200 个狄利克雷过程混合模型。

F.2.2. 数据生成过程（*DGP*）。我们按照Rubin（1981）的平行实验模型模拟多站点效应估计：

$$\hat{\tau}_j \mid \tau_j, \widehat{se}_j^2 \sim \mathcal{N}(\tau_j, \widehat{se}_j^2), \quad j = 1, \ldots, J. \tag{F.1}$$

真实站点效应由恰好包含 K^* 个成分的有限离散混合生成：

$$\tau_j \mid z_j \sim \mathcal{N}(\mu_{z_j}, s^2_{\text{within}}), \qquad z_j \in \{1, \ldots, K^*\}, \tag{F.2}$$

其中，聚类均值在 $[-d, d]$ 上均匀间隔排列（ $d = 2.5$），随后对实现的$\{\tau_j\}$ 进行中心化和重新缩放，以确保 Var(τ_j) $= \sigma_\tau^2$ 精确成立。具体而言，对于每次重复试验，我们计算 $\tau_{\text{最终}j} = \sigma_\tau \cdot (\tau_{\text{原始}j} - \bar{\tau}^{\text{raw}})/\text{SD}(\tau_{\text{原始}})$。这种标准化确保中心间方差恰好等于 σ_τ^2，从而将信息性

操作与效应大小中的随机变异区分开来。生成簇分配时，确保每次重复试验中所有 K^* 个簇至少出现一次 —— 前 K^* 个站点确定性地分配给不同的簇，其余站点则均匀随机分配。

F.2.3. 固定设计参数。 表 F.2记录了所有情景中保持不变的数据生成过程常数。

表 F.2. 固定的数据生成过程和校准参数

组成部分	参数值		Role
站点数量	J	100	有限样本机制
站点间标准差	σ_τ	0.25	重新缩放后设置 Var(τ)
簇内标准差	s_{within}	0.15	控制组分重叠
簇间分离范围	d	2.5	均值范围为 $[-d, d]$
标准误缩放	κ	4	定义 $\widehat{se}_j^2 = \kappa/n_j$
站点规模变异系数	CV(n_j)	0.5	会导致标准误异质
最小站点规模	$n_{\min}$	5	避免退化标准误
混合权重	π_k	$1/K^*$	等权重簇

F.2.4. 信息量指数与校准目标。 根据 Lee 等人（2025年），信息性定义如下：

$$I \equiv \frac{\sigma_\tau^2}{\sigma_\tau^2 + \mathrm{GM}(\widehat{se}_j^2)}, \qquad \mathrm{GM}(\widehat{se}_j^2) = \exp\left(\frac{1}{J}\sum_{j=1}^{J} \log \widehat{se}_j^2\right). \tag{F.3}$$

我们设定 $\widehat{se}_j^2 = \kappa/n_j$，其中 $\kappa = 4$，并从伽马分布生成站点大小：$n_j \sim$ 伽马分布（形状参数 $= 1/$变异系数^2, 率参数 $=$ 形状参数$/\bar{n}$），在下限 $n_{\min} = 5$ 处截断并四舍五入为整数。对于每个目标信息性水平 $I \in \{0.2, 0.5, 0.8\}$，通过对 $\mathbb{E}[I]$ 的蒙特卡洛估计进行一维求根，校准平均站点大小 $\bar{n}$ 。

表 F.3报告了已完成运行中使用的校准目标。

表 F.3. 信息量校准结果

目标 I	校准后的 $\bar{n}_j$	达成的均值 I	绝对误差
0.2	17.4	0.194	0.006
0.5	72.9	0.500	0.000
0.8	278.4	0.793	0.007

F.2.5. 重复实验与可复现性。所有情景均使用 $R = 200$ 次重复实验。随机种子采用确定性设置，以确保可复现性并避免意外的

跨情景/方法耦合。对于复制索引 r 和情景（K^*, I）：DGP 种子为 $\texttt{dgp_base} + 100000r + 1000K + \lfloor 100I \rfloor$；马尔可夫链蒙特卡洛种子则加入方法和置信度哈希。已完成运行所使用的基础种子为：$\texttt{dgp_base}$ =20260116、$\texttt{mcmc_base} = 98765$和 $\texttt{calibration} = 12345$。

F.3. **超先验方法与精确参数。**

F.3.1. 方法比较。我们比较五种 $\alpha \sim \text{Gamma}(a, b)$ 超先验：(1) 模糊：固定为 Gamma(1, 1)。(2) DORO-Unif：采用 Dorazio 风格的 KL最小化，以 K_J 上的离散均匀分布为目标（Dorazio, 2009）。(3) TSMM：两阶段矩匹配，校准 (a, b) 以匹配用户输入 (μ_K, σ_K^2)。(4) 双锚点：TSMM 加上权重主导控制，在检测到主导风险时强制执行 $\Pr(w_1 > 0.5) \leq 0.25$ （E节）。(5) χ^2-DORO（Lee 等人，2025年）：采用基于KL的匹配，并将离散度固定为 $\text{Var}(K_J) = 2\mu_K$。

一个关键区别在于，TSMM/双锚点允许分析者通过 (μ_K, σ_K^2) 对 K_J 的位置和离散度同时进行编码，而 χ^2-DORO 则将离散度硬编码为 $2\mu_K$。这种固定方差限制了研究者在 K_J 中编码“置信度”的强度。

F.3.2. 置信度到方差的映射（*TSMM* / 双锚点）。我们使用方差膨胀因子（VIF）映射：

$$\sigma_K^2 \equiv \text{Var}(K_J) = \text{VIF} \times (\mu_K - 1), \qquad \text{VIF} \in \{1.5,\ 2.5,\ 5.0\} \tag{F.4}$$

分别对应高、中、低置信度。除非另有说明，正文结果均采用中等置信度。

F.3.3. 精确 $(a,\ b)$ 所用超参数。 表 F.4 列出了模拟运行中使用的超先验，以及诱导先验矩（通过高斯–拉盖尔求积计算）和主导性诊断量 $\Pr(w_1 > 0.5)$。

表 F.4. 模拟中使用的超先验参数（按方法、置信度和目标 μ_K）

方法	置信度	μ_K	σ_K^2 $(a,\ b)$	$\mathbb{E}[K]$	$P(w_1>.5)$ 注
模糊	—	—	— (1.00,1.00)	4.84	0.591 伽马分布(1,1)
TSMM	low	5	20 (0.62,0.56)	5.00	0.606 A2-MN
双锚点低		5	20 (2.44,1.28)	7.81	0.347 λ=0.70
TSMM	med	5	10 (1.67,1.65)	5.00	0.557 A2-MN
双锚点	med	5	10 (5.65,3.44)	7.20	0.354 λ=0.70
TSMM	high	5	6（4。35，4。47）	5.00	0.534 A2-MN
双锚点	high	5	6（21.87，14.53）	6.85	0.361 λ=0.70
TSMM	low	10	45 (1.18,0.40)	10.00	0.306 A2-MN
TSMM	med	10	22.5 (2.99,1.10)	10.00	0.232 A2-MN
TSMM	high	10	13.5 (7。19，2。73)	10.00	0.197 A2-MN
TSMM	low	30	145 (2。20，0。14)	30.00	0.018 A2-MN
TSMM	med	30	72.5 (5.18,0.34)	30.00	0.003 A2-MN
TSMM	high	30	43.5 (10.93，0.75)	30.00	0.001 A2-MN
DORO-Unif	—	5	— (2.43,2.46)	4.99	0.547 Dorazio-uniform
DORO-Unif	—	10	— (1.53,0.54)	9.91	0.284 Dorazio-uniform
DORO-Unif	—	30	— (0.96,0.05)	29.40	0.077 Dorazio-uniform
χ^2-DORO	—	5	10（1.85,1.81）	5.04	0.550 方差 $=2\mu_K$
χ^2-DORO	—	10	20 (3.84,1.43)	10.01	0.219 方差 $=2\mu_K$
χ^2-DORO	—	30	60 (6.94,0.47)	30.02	0.002 方差 $=2\mu_K$

注释。 (1) $\Pr(w_1 > 0.5)$ 是在诱导超先验下计算的（E节）。当 $\Pr(w_1 > 0.5)$ 超过诊断阈值（0.40，实现中采用的阈值）时，将触发双锚点，并强制满足 $\Pr(w_1 > 0.5) \le 0.25$。(2) 对于 $\mu_K \in \{10, 30\}$ 在$J = 100$处，主导性诊断量远低于触发阈值；因此，双锚点会返回 TSMM 参数（在数值容差范围内），性能差异应归因于蒙特卡洛变异，而非先验存在实质性差异。

F.4. 推断模型和马尔可夫链蒙特卡洛配置。 在每次重复试验中，我们使用中国餐馆过程先验对簇分配进行建模，拟合一个狄利克雷过程混合模型：

$$\hat{\tau}_j \sim \mathcal{N}(\tau_j, \widehat{se}_j^2), \qquad \tau_j \sim \mathcal{N}(\theta_{z_j}, \sigma_{z_j}^2), \qquad z \sim \mathrm{CRP}(\alpha), \tag{F.5}$$

使用超先验：

$$\alpha \sim \mathrm{Gamma}(a, b), \quad \tau \sim \mathcal{N}(0, 100), \quad \sigma_\tau^2 \sim \mathrm{Unif}(0, 10), \quad \sigma_k^2 \sim \mathrm{InvGamma}(2, 1). \tag{F.6}$$

马尔可夫链蒙特卡洛设置（NIMBLE）：2 条链，1,000 次迭代，500 次预烧，抽稀间隔为 1，监测 α 和 z。对于每次马尔可夫链蒙特卡洛抽样 t，计算已占用簇数

as

$$K_J^{(t)} = \left|\{z_1^{(t)}, \ldots, z_J^{(t)}\}\right|. \tag{F.7}$$

根据 $\{K_J^{(t)}\}$ 计算 K_J 的后验汇总（均值、众数、分位数）。

F.5. **结果指标。** 对于每个情景-方法单元（以及适用时的置信水平），我们报告：$K \equiv \mathbb{E}[K_J \mid \text{data}]$ 的后验均值；偏差 $\mathbb{E}[K] - K^*$（对重复实验取期望）；均方根误差 $\sqrt{\mathbb{E}[(\widehat{K} - K^*)^2]}$；坍缩概率 $\Pr(K_J = 1 \mid \text{data})$；95% 可信区间覆盖率（95% 后验区间包含 K_J K^* 的重复实验所占比例）；置信区间宽度（95% 后验区间的平均宽度）；以及 α的后验均值。我们还报告基于众数的准确度度量：$\Pr(\text{mode}(K_J) = K^*)$ 和 $\Pr(|\text{mode}(K_J) - K^*| \leq 1)$。

F.6. **中等置信度下的完整结果（主要设置）。** 本节报告四种主要方法（模糊、DORO-Unif、TSMM、双锚点）在完整 3×3 网格（K^*, I）上的中等置信度完整结果。

F.6.1. 后验期望簇数的偏差。 图 F.1使用小提琴图展示了所有 9 个情景中偏差（$\mathbb{E}[K_J \mid \text{数据}] - K^*$）的分布，全面呈现了估计误差的集中趋势和变异性。模糊先验（黄色）在所有情景中均表现出严重的负偏差（约为-3 当 $K^* = 5$, -8 当 $K^* = 10$，以及 -28 当 $K^* = 30$）。双锚点（粉色）在 $K^* = 5$ 时实现了近乎零偏差，而在较大的 K^*时显著降低了偏差。当 $K^* \in \{10, 30\}$ 时，主导风险较低，TSMM 和双锚点的表现相当。

F.6.2. 坍缩概率。 坍缩的后验概率 $\Pr(K_J = 1 \mid \text{数据})$刻画了一种关键的失效模式：DP混合模型退化为单分量模型。 图 F.2 展示了所有情景下的坍缩概率。这幅图为双锚点的贡献提供了最有力的证据：坍缩概率直接刻画了实际中的失效模式，而这正是基于权重的诊断方法旨在解决的问题。

F.6.3. 均方误差。 图 F.3 展示了后验期望簇数相对于真实值的均方误差 K^*，并采用对数刻度，以适应不同方法之间的巨大差异。对数刻度可视化对于呈现 TSMM 和双锚点所达成的数量级改进是必要的。

F.6.4. 相对于模糊先验的改进。 为了更直接地总结改进，图 F.4 给出了热力图，展示相对于模糊基线的均方误差降低百分比。

F.6.5. 区间校准：覆盖率与宽度。 图 F.5展示了所有情景下 95% 可信区间的覆盖率。对于中到大 K^*，模糊先验产生的区间较窄，但覆盖率较差，这反映出后验分布集中在非常小的 K_J附近，虽有信心却并不正确。

图 F.6 展示了 95% 可信区间的宽度。对于较大的 K^*，DORO-Unif 的区间非常宽且变化很大（宽度的标准差较大），表明其诱导的$p(K_J)$ 在有限 J 设定下可能不稳定。

F.6.6. 低信息摘要。 由于 $I = 0.2$ 是先验影响最大的一种设定，图 F.7 针对这一关键情形提供了简洁的四面板摘要。该摘要表明，双锚点在低信息区域的整体表现最佳：对 $K^* = 5$的估计近乎无偏，在所有 K^*情景下均大幅降低均方误差，坍缩概率始终较低，且区间覆盖率恰当。

F.7. **额外基线： χ^2-DORO**。 Lee 等人（2025年）提出了一种 χ^2-DORO 校准方法，该方法在实证中表现良好，并成为本研究的起点（在该论文中称为“信息性先验”）。该方法的目标是：

$$\mathbb{E}[K_J] = \mu_K, \qquad \mathrm{Var}(K_J) = 2\mu_K, \tag{F.8}$$

并且（在典型实现中）使用网格搜索／离散 KL 最小化来处理 $(a,\ b)$，计算开销很大。

两项局限促使我们提出 TSMM/Dual-Anchor 泛化：（1）计算效率低： 网格搜索校准比 TSMM 的基于牛顿法的校准慢几个数量级，使交互式 elicitation 变得繁琐。（2）离散程度固定： 约束 Var(K_J) $= 2\mu_K$ 无法调整，限制了研究者对 K_J 进行置信度（或怀疑程度）编码的能力，而不仅限于均值。

图 F.8 比较了 χ^2-DORO 在低信息场景下与其他方法的表现。
TSMM/Dual-Anchor 可视为一种有理论依据且灵活的扩展 χ^2-DORO：通过允许 σ_K^2 随置信度变化，并加入可选的主导性控制，它们扩展了 χ^2-DORO 的理念，同时保留校准精度并提升实际可用性。

F.8. **敏感性分析：置信度设定（TSMM / 双锚点）**。
图 F.9　　报告 TSMM 和双锚点对置信度（方差）的敏感性

在 VIF $\in \{1.5, 2.5, 5.0\}$ 设定下（分别对应高、中、低置信度）。

关键观察：（1）置信度越高（ σ_K^2越小），通常越能降低中等至较大 K^*情况下的均方根误差，因为先验更强烈地支持较多的簇数。（2）对于 $K^* = 5$, TSMM 仍然容易受到影响：即使置信度较高，估计量仍可能存在向下偏差（而置信度较低时则经常坍缩）。双锚点则稳健得多，在不同置信水平下都能保持较低的均方根误差和较低的坍缩概率。（3）对于 $K^* \in \{10, 30\}$，双锚点通常不会被触发（表 F.4），因此 TSMM 和双锚点的敏感性模式非常接近。

F.9. **浓度参数 α的后验行为**。由于 K_J 是 α的有限样本泛函，考察其后验分布 α 有助于解读方法的表现。图 F.10 展示了不同情景和方法下 α 的后验均值。在模糊先验下，后验分布 α 仍然较小（约为 0.35–0.38），这与其强烈倾向于较少簇的特点一致。DORO-Unif 得到的后验分布 α 高于模糊先验，但在较大的 K^*情况下低于 TSMM/双锚点。当 α较大时，TSMM/双锚点会显著提高 K^* （例如， $\alpha \approx 12$ 对于 $K^* = 30$），使先验与簇丰富情形相匹配。对于 $K^* = 5$，双锚点产生的后验分布 α明显大于 TSMM，反映了避免第一个权重占主导所需的额外概率质量。

图 F.11 展示了 K_J 的后验标准差，用于刻画不确定性校准。

F.10. **基于众数的恢复：** K^*。 对于诸如 K_J 之类的离散泛函，后验众数可以作为一种有用的汇总。 图 F.12 同时展示了精确恢复（概率(众数(K_J) $= K^*$)）与近似恢复（概率(|众数(K_J) $- K^*| \leq 1$)）的比率。在该设计中，精确恢复通常颇具挑战，尤其是对于 $K^* = 30$，其精确匹配概率即便在 TSMM/双锚点下仍低于 5%。当 $K^* = 5$ 时，双锚点显著改善了近似恢复（概率约为 0.32–0.34，而 TSMM/DORO-Unif 为 0.11–0.18），这与其防止坍缩及极端欠拟合的目标相一致。

图 F.13 比较了簇数估计的后验众数与均值。DORO-Unif 和 TSMM（对于较小的 K^*）往往显示许多点聚集在众数 $= 1$，且均值 $>$ 众数，表明这些后验分布将显著概率质量置于 $K_J = 1$，同时对更高取值具有一些支撑。双锚点减轻了这种坍缩模式。

F.11. 均方误差的偏差–方差分解（低信息设置）。为更好地理解误差是由系统性低估还是蒙特

卡洛变异所驱动，我们将均方误差分解为：

$$\mathrm{MSE}(\widehat{K}) = \underbrace{\mathrm{Bias}(\widehat{K})^2}_{\text{systematic}} + \underbrace{\mathrm{Var}(\widehat{K})}_{\text{across replications}} . \tag{F.9}$$

图 F.14 展示了 $I = 0.2$的这一分解。对于模糊、DORO-Unif、TSMM 和 χ^2-DORO，偏差约占均方误差的 90–100%——主要失效模式是系统性低估 K_J。对于 K^* =5处的双锚点，均方误差几乎完全由方差驱动（偏差 ≈ 0），这表明双锚点约束在很大程度上消除了系统性坍缩，仅留下残余采样/马尔可夫链蒙特卡洛变异性。

表 F.5给出了这一分解的数值。

表 F.5. $I = 0.2$的均方误差偏差–方差分解

情景	方法	偏差	标准差($\widehat{K}$)	偏差²	MSE	偏差²/均方误差
K^*=5	模糊	−3.01	0.55	9.06	9.39	96.5%
K^*=5	DORO-Unif	−2.36	0.63	5.57	5.96	93.4%
K^*=5	TSMM	−2.63	0.61	6.92	7.28	95.0%
K^*=5	双锚点	+0.04	0.76	0.00	0.58	0.3%
K^*=5	χ^2-DORO	−2.57	0.59	6.60	6.92	95.5%
K^*=10	模糊	−8.05	0.49	64.80	65.05	99.6%
K^*=10	DORO-Unif	−5.49	1.75	30.14	33.18	90.8%
K^*=10	TSMM	−3.46	1.27	11.97	13.62	87.9%
K^*=10	双锚点	−3.63	1.47	13.18	15.28	86.2%
K^*=10	χ^2-DORO	−2.87	1.08	8.24	9.42	87.4%
K^*=30	模糊	−28.04	0.55	786.24	786.48	100.0%
K^*=30	DORO-Unif	−21.96	4.96	482.24	506.67	95.2%
K^*=30	TSMM	−4.11	1.30	16.89	18.61	90.8%
K^*=30	双锚点	−4.03	1.39	16.24	18.19	89.3%
K^*=30	χ^2-DORO	−3.28	1.11	10.76	11.97	89.9%

F.12. 计算基准测试（校准成本）。尽管所有方法在后验推断上具有相同的马尔可夫链蒙特卡洛成本，但获得（a, b）的校准成本却差异显著。

Lee 等人（2025年）的 χ^2-DORO（以及经典的 DORO 实现）通常依赖基于网格的KL匹配；当网格较精细，或需要针对多个 μ_K 取值重复进行先验 elicitation 时，这种方法可能会很慢（Dorazio，2009年）。相比之下，TSMM 和双锚点通过快速确定性优化计算 $(a,\ b)$ （第 C 节和E节）。

表 F.6 给出了数量级时间比较，与我们实现中观察到的计算扩展规律一致。

表 F.6. 代表性时间比较（每次校准 +每次 MCMC 拟合）

方法	校准（秒）	马尔可夫链蒙特卡洛（秒）	总计（秒）	校准加速比	总加速比
模糊	0.00	2.1	2.1	—	22.5×
DORO-Unif（网格 100×100）	45.2	2.1	47.3	1×	1.0×
χ^2-DORO（网格搜索）	45.2	2.1	47.3	1×	1.0×
TSMM（A2-MN 牛顿法）	0.05	2.1	2.15	904×	22.0×
双锚点（A2-MN +约束）	0.10	2.1	2.20	452×	21.5×

注。 (1) 所列时间仅为代表性数值，旨在说明数量级差异；实际耗时取决于硬件、软件版本以及 DORO 类方法所用网格的分辨率。(2) 标题中的 ≈900× 加速比针对校准步骤（例如，45.2 秒/ 0.05 秒 ≈ 900×），这是各方法间存在实质性差异的部分。

附录G. 软件实现

G.1. **目的与范围**。 本附录介绍正文中提出的方法如何在 DPprior R 包中实现，并提供一个最简计算示例，以复现端到端 elicitation 工作流程：(1) 簇数量信念的设计条件设定；(2) 两阶段矩匹配（TSMM）校准；(3) 对非预期权重行为进行诊断检查；以及 (4)（可选）双锚点细化。

本附录有意不作为完整的软件手册。完整的软件包文档——包括扩展示例、案例研究和完整的 API 参考——在线维护：

- 软件包网站： https://joonho112.github.io/DPprior/
- API 参考： https://joonho112.github.io/DPprior/articles/api-reference.html
- 应用指南：
  https://joonho112.github.io/DPprior/articles/applied-guide.html
- 双锚点专题文档： https://joonho112.github.io/DPprior/articles/dual-anchor.html
- 案例研究： https://joonho112.github.io/DPprior/articles/case-studies.html

希望了解复现论文工作流程所需内容之外的实现细节的读者，请参阅这些资源。

G.2. **软件包概览**。

G.2.1. *DPprior* 实现了什么。该软件包实现了狄利克雷过程（DP）浓度参数的校准和诊断，

$$\alpha \sim \mathrm{Gamma}(a, b) \quad \text{(shape–rate parameterization)},$$

其中，DPprior 返回形状参数 a 和率参数 b。因此，$\mathbb{E}[\alpha] = a/b$ 且 $\mathrm{Var}(\alpha) = a/b^2$。

G.2.2. 计算架构。 DPprior 软件包由三个计算层组成，直接实现了正文第 2–3 节中提出的框架。

第 *1* 层：精确计算。 这一层提供参考级的分布计算工具，涵盖 K_J (A 节)，包括对数斯特林数计算、精确的 Antoniak 概率质量函数，以及通过双伽马函数和三伽马函数计算条件矩和边缘矩。这些函数可作为验证近似的真实基准，并支持阶段 2 中的精确矩匹配。

第 *2* 层：*TSMM* 校准。 这一层实现两阶段矩匹配算法（第 3.2–3.3 节；B 节和 C）：阶段 1（A1）提供快速的闭式初始值 (a_0, b_0)，由大样本J 近似推导而来；阶段 2（A2-MN）采用牛顿细化，使 K_J的精确混合矩相匹配。这是`DPprior_fit()`中的默认方法。

第 *3* 层：权重诊断与双锚点。本层实现权重分布理论（E节），包括 w_1 尾部概率、共同聚类矩的闭式计算，以及用于平衡簇数和权重集中度目标的双锚点优化。

G.2.3. 设计原则。该软件包遵循若干设计原则，以支持应用研究：(1) 面向实践者的接口。主要入口 `DPprior_fit()`接受直观的输入（J、 μ_K，以及方差或置信水平），并返回经校准的伽马超参数及诊断摘要。(2) 渐进式信息呈现。简单用例只需调用一次函数，而高级用户可以使用底层的精确计算和优化机制。(3) 透明报告。每个拟合对象都包含目标矩和达成值、权重诊断以及主导风险评估，支持正文第3.5节中的报告清单。

G.2.4. 安装。 DPprior 托管在 GitHub 上：

```
# 从 GitHub 安装（开发版本）安装。软件包("



# 来自 CRAN（如可用）安装。
```

G.2.5. 版本固定与参数化。由于校准是数值性的（且绘图默认值可能发生变化），可复现分析应记录：

```
packageVersion("DPprior")
sessionInfo()
```

参数化说明。DPprior使用形状–率约定。如果下游软件使用形状–尺度约定（例如 NumPy），请通过 尺度 = 1/ 率 进行转换。

G.3. **映射：论文组成部分到 DPprior 函数**。表 G.1 提供了从论文的方法组成部分到软件包 API 的实现映射。

表 G.1. 论文组成部分 →DPprior 实现

论文组成部分的用途		主要功能	典型用途
TSMM 阶段 1 (A1)	闭式初始值 (a_0, b_0)	DPprior_a1()	调用 by 内部 DPprior_fit()
TSMM 阶段 2 (A2-MN)	精确矩匹配——通过牛顿法	DPprior_拟合(method="A2-MN")默认值	for 可发表的分析
阶段 2 替代方案（A2-KL）	分布层面的 KL 匹配	DPprior_拟合(method="A2-KL")可选；	适用于小型 J 或基线
校准条目	指定 (J, μ_K, σ_K^2)，获得 (a, b)	DPprior_拟合()	大多数用户的主要入口
诊断	K_J、w_1、ρ	DPprior_diagnostics()的摘要	用户之后始终建议进行校准
可视化	先验仪表板 / com- 分量	plot(), plot_K_先验(), plot_w1_先验()	传达先验含义
双锚点	约束 $\Pr(w_1 > t)$	DPprior_对偶()	可选细化步长

表 G.2 总结了支撑校准算法的精确计算函数。

表 G.2. 精确计算函数

函数	用途	输入	输出
compute_log_stirling()	对数斯特林数矩阵	J	矩阵 $L[n,k]$
pmf_K_给定_alpha()	条件概率质量函数 $p(k \mid \alpha)$	J, alpha, logS	概率向量
pmf_K_边缘()	边缘概率质量函数 $p(k \mid a,\ b)$	J, a, b, logS	概率向量
均值_K_在给定_alpha()	条件下的均值 $\kappa_J(\alpha)$	J, alpha	标量
方差_K_在给定_alpha()	条件方差 $v_J(\alpha)$	J, alpha	标量
exact_K_矩()	边缘 $\mathbb{E}[K_J]$, 方差(K_J)	J, a, b	命名列表

表 G.3 总结了支持诊断与双锚点协议的权重分布函数。

桌子 G.3。权重分布函数

函数	用途	输入	输出
prob_w1_exceeds()	尾概率 $\Pr(w_1 > t)$	t,a, b	概率
cdf_w1()	w_1的累积分布函数	x、a、b	$F_{w_1}(x)$
密度_w1()	w_1的密度	x，a，b	$f_{w_1}(x)$
分位数_w1()	w_1的分位数函数	p，a，b	分位数值
均值_w1()	w_1的均值	a，b	$\mathbb{E}[w_1]$
均值_rho()	$\mathbb{E}[\rho]$的共同聚类均值	a，b	标量
方差_rho()	ρ的方差	a，b	Var(ρ)

G.4. **核心工作流程**。

G.4.1. 步骤 *1*：指定设计与簇数量目标。 关键设计输入是 J，即解释簇数时所依据的设计规模（例如，多中心研究中的站点数量）。引出的信念以目标均值和方差表示：$\mu_K = \mathbb{E}[K_J]$和 $\sigma_K^2 =$Var(K_J)。

在DPprior中，目标可以通过以下两种等效方式之一给出：直接提供方差_K，或提供定性的置信度等级，再通过方差膨胀因子（VIF）映射为方差：$\sigma_K^2 =$ VIF $\times(\mu_K - 1)$。

表 G.4. 置信度与 VIF 映射

置信度解释		VIF	σ_K^2 (μ_K=5)
“高”	基于先验证据，有相当把握	1.5	6
“中”	有一定把握，但也承认存在不确定性	2.5	10
“低”	相当不确定；只是粗略估计	5.0	20

G.4.2. 第 *2* 步：通过 *TSMM* 校准伽马超先验。主要校准函数为：

```
拟合 ← DPprior_拟合(J = J, mu_K = mu_K, 置信度 = "中等")
```

默认情况下，`DPprior_fit()` 使用 A1 进行初始化，并使用 A2-MN（牛顿细化），以确保在有限的 J 下实现的矩与目标相匹配。返回的对象包含超参数`fit$a` 和 `fit$b`。

常见的敏感性选项包括：`method = "A1"` （仅近似；速度快，主要用于对较大的 J 进行筛选），`method = "A2-KL"`（基于KL的分布匹配；较慢），`check_diagnostics = TRUE` （在拟合期间计算诊断），以及 `warn_dominance = TRUE`（当主导风险高时发出警告）。

G.4.3. 第*3*步：运行诊断并可视化。拟合之后，计算诊断摘要：

```
诊断量 ←DPprior_诊断(拟合)
```

诊断列表包含 α、K_J、w_1 和 ρ （共同聚类）的摘要，以及警告标志：`diag$alpha` （ α 的均值/标准差/分位数），`diag$K` （ K_J 的摘要：均值、方差、众数、分位数），`diag$weights`（ w_1 的摘要，包括尾部概率），`diag$coclustering` （ $\rho = \sum_h w_h^2$ 的摘要），以及 `diag$warnings` （任何诊断警告）。

解释说明。该软件包将 w_1 用作第一个按大小加权的截断棒权重（GEM 顺序），不一定是实际成分权重中最大的那个。这与手稿中的诊断量定义一致，并支持在 $\alpha \sim$ 伽马分布(a, b) (E节)条件下进行闭式计算。

可视化：

```
绘图(fit)            #完整仪表板绘图_K_先验
```

G.4.4. 第*4*步（可选）：双锚点细化。如果诊断显示主导簇的先验概率较高（例如，$\Pr(w_1 > 0.5) \geq 0.40$），则双锚点程序可以降低主导风险：

```
拟合_对偶 ←DPprior_对偶（拟合 =拟合，w1_目标 = 列表（概率 =
```

```
    lambda  = 0.7, # 优先考虑 K拟合，而非权重拟合损失_类型 =
```

G.5. **计算示例：端到端信息 elicitation**。 此示例对应贯穿 OSM 使用的计算情境：$J = 50$、 $\mu_K = 5$, 以及中等置信度（VIF $= 2.5 \Rightarrow \sigma_K^2 = 10$）。

```
库(DPprior)

#  ============================================================
# 步骤 0：设计  +elicitation目标
#  ============================================================
J    ← 50


# 选项 A（推荐）：定性不确定性
fit  ← DPprior_fit(J  = J, mu_K = mu_K, 置信度  = "中等")

# 选项 B（等价）：直接指定方差(K) # fit  ←DPprior_fit(J


print(fit)
```

预期输出：

```
DPprior 先验启发结果
==========================================

伽马超先验：alpha  ~ 伽马分布(a  = 1.4082, b  = 1.0770)
  期望[α] = 1.308,    标准差[α] = 1.102

目标 (J  = 50):
  E[K_J]    = 5.00
  方差(K_J)  = 10.00
   （来自置信度  = 'medium'）

达成值:
  E[K_J] = 5.000000, 方差(K_J) = 10.000000
```

```
  残差  = 3.94e-10

方法：A2-MN（7 次迭代）

主导风险：高（P(w1 > 0.5) = 50%）
```

该校准在 7 次牛顿迭代中实现了机器精度矩阵匹配（残差 $\approx 10^{-10}$）。

```
#  ============================================================
# 步骤 1：诊断（建议用于每次分析）
#  ============================================================
诊断量  ←DPprior_诊断(fit)

# 手稿中使用的关键量
E_K       ← 诊断量$K$均值
方差_K    ← 诊断量$K$方差
概率_dom05 ← 诊断量$权重$概率_超过["概率_大于_0.5"]
P_dom09 ← diag$权重$概率_超过["概率_大于_0.9"]
E_rho   ← diag$共同聚类$均值

    # 可视化绘
```

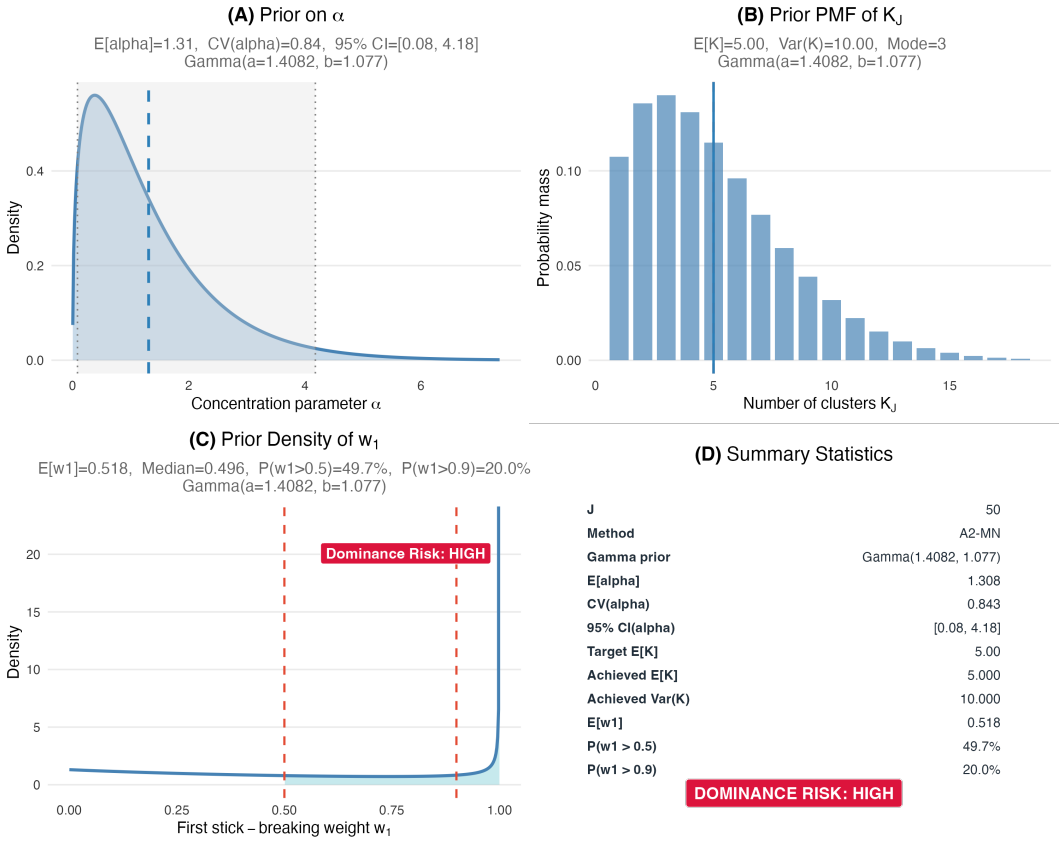


图 G.1. 计算示例的 TSMM 诊断仪表板（$J = 50$, $\mu_K = 5$，置信度为“中”）。面板 (A) 显示浓度参数 α 的先验密度。面板 (B) 显示 K_J 的边缘概率质量函数，并标出目标均值 $\mathbb{E}[K_J] = 5$。面板 (C) 展示第一个断棒权重 w_1 的边缘密度，并在阈值 $t = 0.5$ 和 $t = 0.9$ 处标出参考线。面板 (D) 总结校准结果，包括主导风险评估。

```
#  ============================================================
# 步骤 2（可选）：若主导性较高，则进行双锚点细化
#  ============================================================
如果 (P_主导05 >= 0.40) {

  拟合_对偶 ←DPprior_对偶( 拟合 = 拟合, w1_目标 = 列表(概率 = 列
```

```
    verbose = FALSE ) 诊断量_对偶 ← DPprior_诊
```

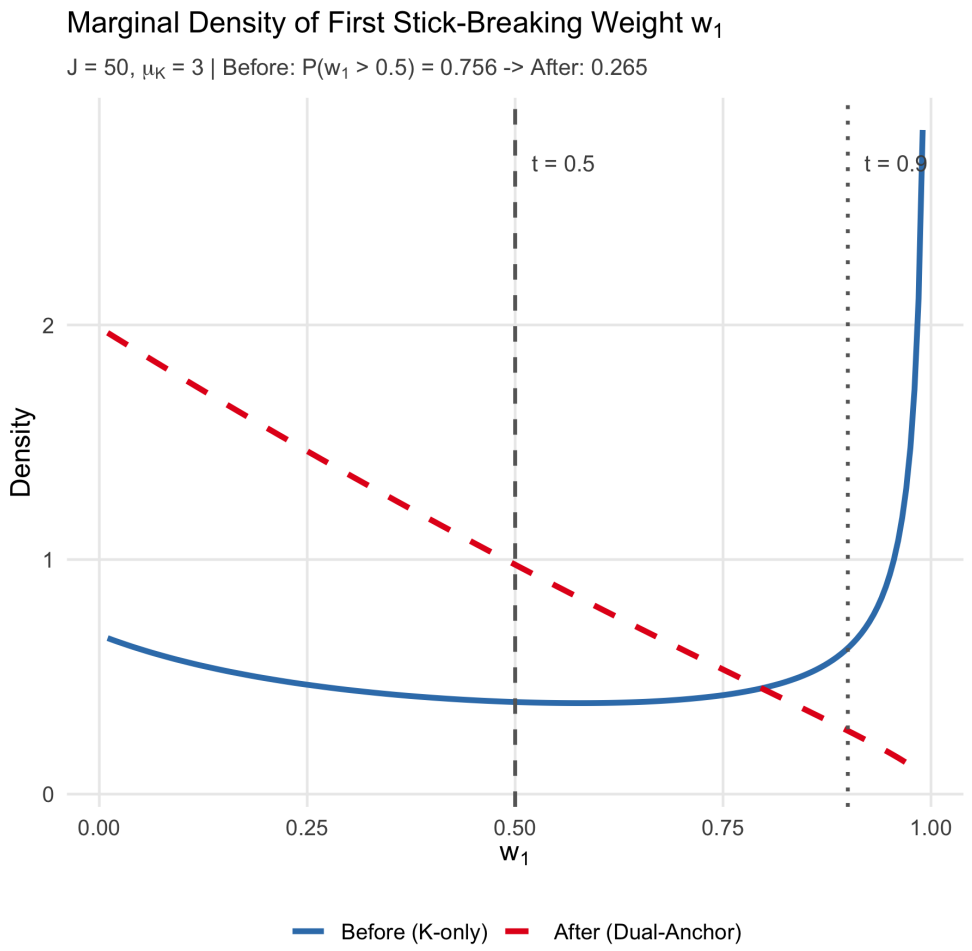


图 G.2. 双锚点细化前后的权重分布（$J = 50$、 $\mu_K = 3$，置信度 = “低”）。蓝色实线显示在仅使用 K 的 TSMM 校准下， w_1 的边际密度，$\Pr(w_1 > 0.5) = 0.756$。红色虚线显示经过双锚点细化后的密度，目标为 $\Pr(w_1 > 0.5) \leq 0.25$ 和 $\lambda = 0.7$，达到 $\Pr(w_1 > 0.5) = 0.265$。竖直参考线表示 $t = 0.5$ 和 $t = 0.9$ 处的诊断阈值。

```
#  ============================================================
# 第 3 步：在你的 DP 混合模型中使用 (a, b)
#  ============================================================

# Stan（形状-率）： #
```

```
# JAGS（形状-率）：# alpha ~ 伽马分布(a, b) # 在 R
```

G.5.1. 敏感性分析：改变不确定性设定。简单的敏感性分析会比较不同的置信水平：

```
拟合 ←list( "高置信度" = DPprior_fit(J = 50, mu_K = 5, 置信度 = "高"),
```

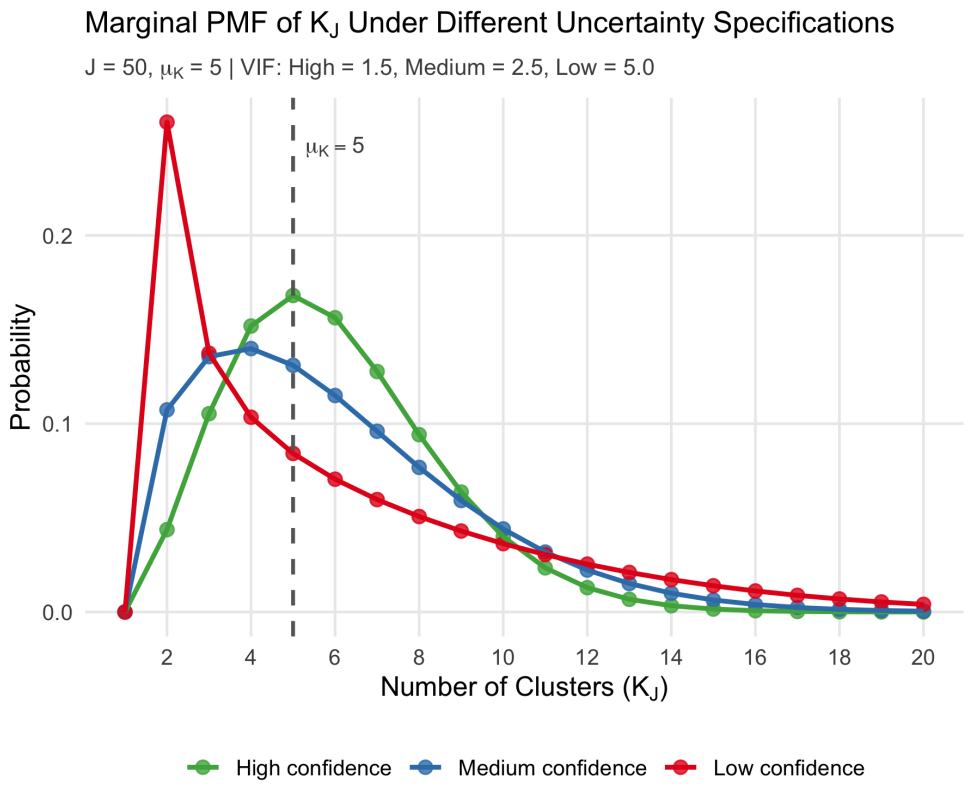


图 G.3。诱导的 K_J 分布对不确定性设定（$J = 50,\ \mu_K = 5$）的敏感性。三条曲线分别对应高置信度（方差膨胀因子 $= 1.5$，绿色）、中等置信度（方差膨胀因子 $= 2.5$，蓝色）和低置信度（方差膨胀因子 $= 5.0$，红色）。不确定性越高（置信度越低），分布越宽、越分散。三个先验的目标均值 $\mu_K = 5$（垂直虚线）相同，但隐含的方差各不相同。

表 G.5 总结了不同置信水平下诱导先验的特征。

表 G.5。敏感性分析摘要（$J = 50,\ \mu_K = 5$）

置信度	VIF	方差(K)	a	b	$P(w_1>0.5)$	$\mathbb{E}[\rho]$
High	1.5	6	3.568	2.900	0.465	0.484
中等	2.5	10	1.408	1.077	0.497	0.518
Low	5.0	20	0.518	0.341	0.563	0.585

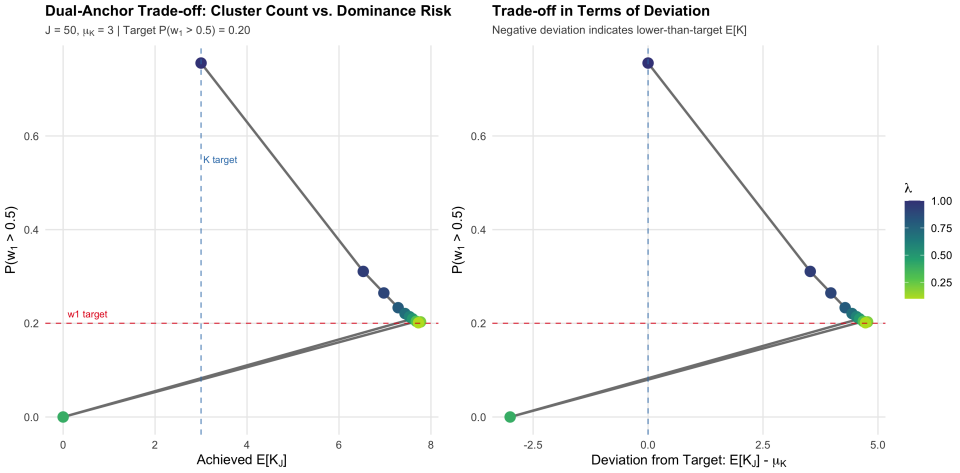


图 G.4. 双锚点权衡曲线（$J = 50$, $\mu_K = 3$）。每个点表示使用不同权衡参数 $\lambda \in [0.1, 1.0]$ 进行的一次校准。左图绘制达成值 $\mathbb{E}[K_J]$ 与 $\Pr(w_1 > 0.5)$ 的关系，显示较低的 λ 取值会优先满足权重约束，但代价是偏离 K 目标。右图则以偏离目标 μ_K 的程度展示相同的权衡。水平红色虚线表示目标 $\Pr(w_1 > 0.5) = 0.20$；垂直蓝色虚线标示 K 目标。

G.6. **可复现性清单**。 为支持透明报告和精确复现，我们建议在使用 TSMM 和/或双锚点校准时，记录 表 G.6 所列项目。

表 G.6. 可复现性与报告清单

需记录的项目	获取方法	重要性
设计规模 J	fit$J	所有目标均以设计为条件
已设定 μ_K 和 σ_K^2	fit$target$mu_K,fit$target$var_K	定义聚类数锚点
置信度映射	置信水平 + 方差膨胀因子	解释“低/中/高”在各个 papers
校准方法	fit$method	A1 与 A2-MN 与 A2-KL
超先验 (a, b)	fita, fitb	供后续马尔可夫链蒙特卡洛使用
实现的矩	fit$fit	确认校准成功
主导性诊断-tics	诊断量$权重$概率_超过[...]	标记意外集中
共同聚类 summary	诊断量$共同聚类$均值	可解释汇总诊断
双锚点（ifused）w1_目标、lambda、损失_类型		完整定义细化步骤
软件版本	packageVersion("DPprior"), sessionInfo()	确保数值可复现性

可直接根据这些对象整理出一段简短的“方法部分”式陈述，并遵循正文第 3.5 节建议的报告规范。

G.7. **代码库与资源**。

*P*软件包源代码。 DPprior R 包的源代码可在以下位置获取：

- GitHub 代码仓库： https://github.com/joonho112/DPprior
- 软件包网站： https://joonho112.github.io/DPprior/

模拟的可复现性。 正文和F 节中所有用于生成模拟结果的代码均存档于单独的可复现性代码库中：

- 可复现性代码库： https://github.com/joonho112/dpprior-elicitation此代码库包含完整的复现材料，组织如下。

表 G.7. 模拟可复现性材料

Item	文件名	描述
模拟设计	design.R	全因子设定
校准	步长1_calibrate.R	所有方法的先验构建
MCMC拟合	步长2_simulate.R	使用并行 pro- 的后验推断处理
聚合	步长3_aggregate.R	汇总统计量计算
图表	步长4_figures.R	所有正文和 OSM 图表

种子策略。 可复现的运行应记录确定性设种方案。OSM 报告了一种在重复试验和情景层面固定种子的设种方法。一种便捷的模式是，根据基础种子加上（K^*、I）、方法和置信水平的哈希值，推导出每次重复试验的种子，然后再从重复试验种子推导出每条链的种子：

```
设置.seed(2026) # 用于可复现性的主种子
```

版本要求： R 版本 $\geq$ 4.0.0；DPprior 版本 1.0.0；依赖项：stats、Rcpp（$\geq$ 1.0.0）。

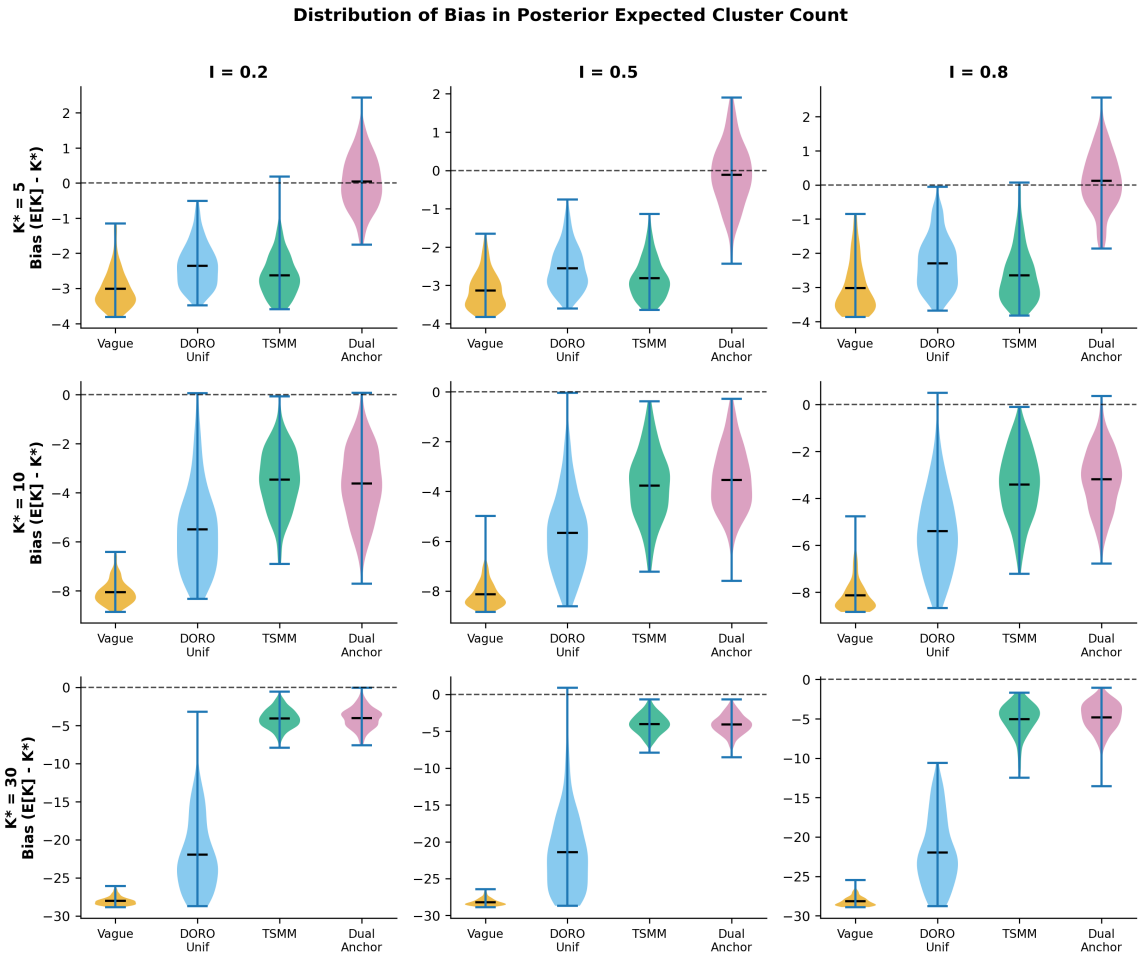


图 F.1。按方法和情景划分的后验期望簇数偏差分布$\mathbb{E}[K_J \mid \text{data}] - K^*$。行：真实簇数$K^* \in \{5, 10, 30\}$；列：信息量水平 $I \in \{0.2, 0.5, 0.8\}$。虚线表示零偏差。小提琴图展示 200 次重复实验中的完整分布，黑色水平线表示中位数。

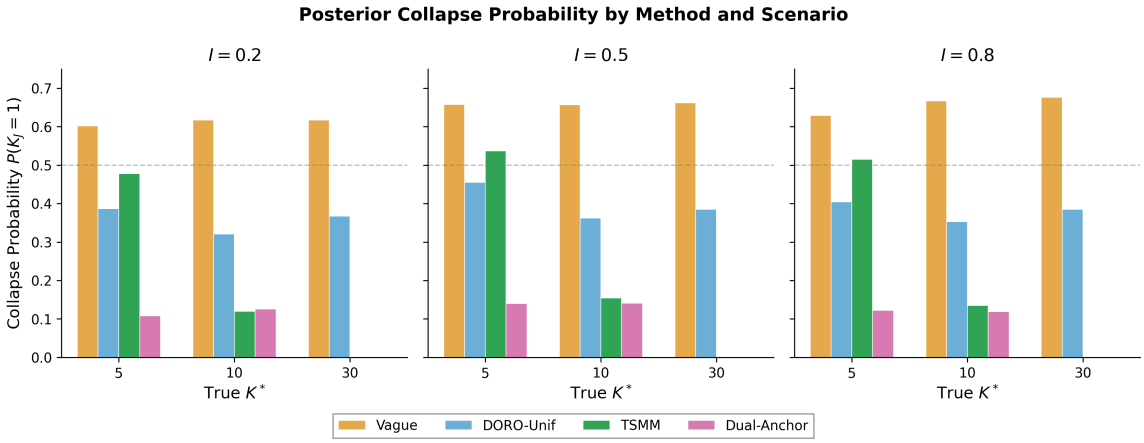


图 F.2. 按方法和情景划分的后验坍缩概率 $P(K_J = 1)$。水平虚线表示 50% 的坍缩概率。无论真实 K^* 如何，模糊先验（黄色）产生的坍缩率均超过 60%。TSMM（绿色）显著降低了 $K^* \geq 10$的坍缩，但在 $K^* = 5$下仍易受影响。双锚点（粉色）在所有情景下均能持续防止坍缩，即使对于 $K^* = 5$，也能将 $P(K_J = 1)$ 降至 15% 以下。

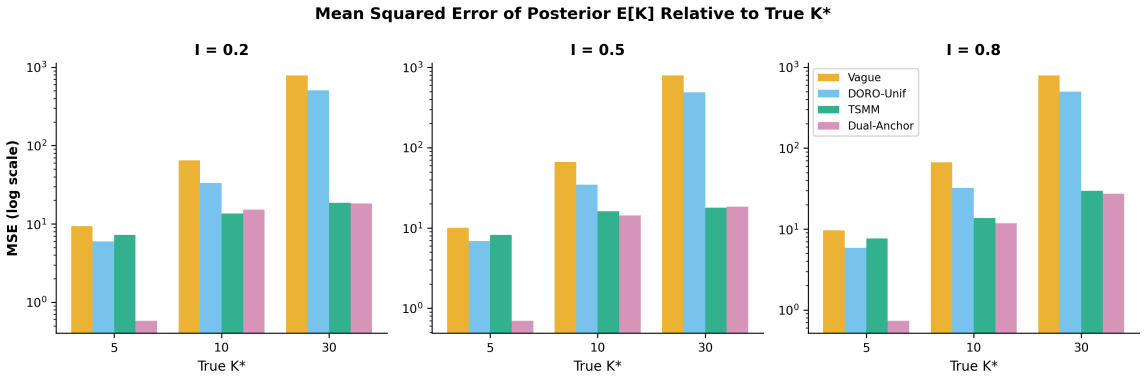


图 F.3. 后验期望簇数$\mathbb{E}[K_J \mid$ 相对于真实 K^* 的均方误差] （对数刻度）。对于 $K^* = 30$，模糊先验的均方误差为≈ 790，DORO-Unif 为 ≈ 500，而 TSMM 和双锚点的均方误差达到$\approx$ 18–30，表示均方误差降低约 97%。对于 $K^* = 5$，双锚点的均方误差为 < 1，相比之下，TSMM 的均方误差为 $\approx$ 7–8。

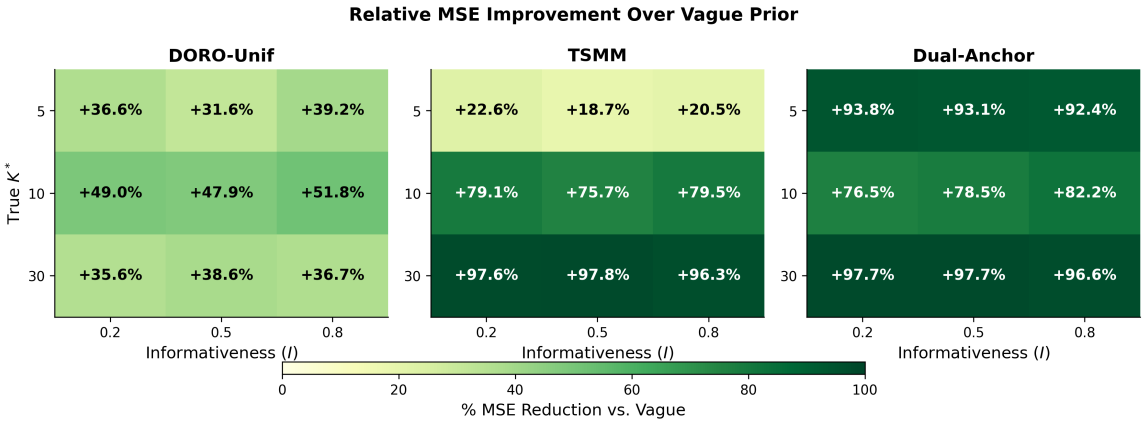


图 F.4. 相较于模糊先验的均方误差相对改善（%）。每个单元格显示 $(MSE_{\text{Vague}} - MSE_{\text{Method}})/MSE_{\text{Vague}} \times 100\%$。当 $K^* = 30$ 时，TSMM 和双锚点的均方误差降低约为 96–98%。对于 $K^* = 5$，双锚点带来的改善最为显著（均方误差降低约为 92–94%），而这正是 TSMM 表现不佳之处（改善仅为 18–23% ）。

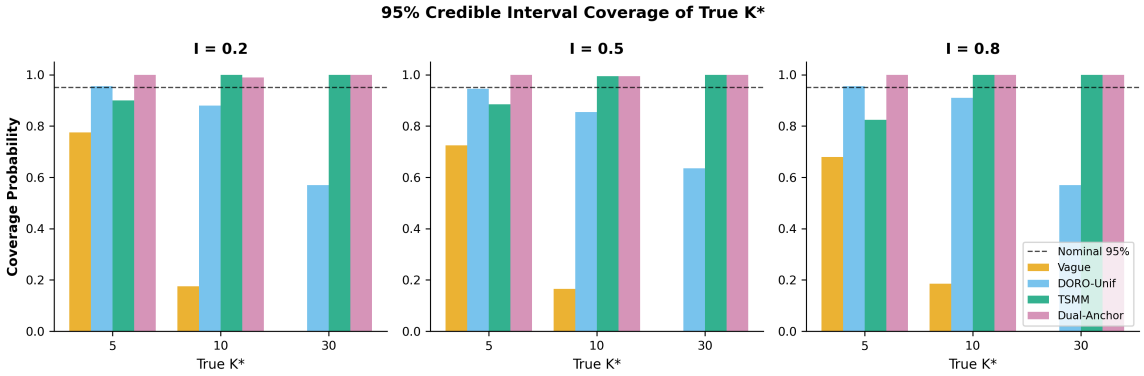


图 F.5. 按方法和情景划分的真实 K^* 的 95% 可信区间覆盖率。水平虚线表示名义 95% 水平。对于 $K^* \geq 10$，模糊先验（黄色）产生的区间覆盖率严重不足；对于 $K^* = 30$，覆盖率实际上为 0%。TSMM 和双锚点（绿色、粉色）在所有情景下均达到接近名义水平或更保守的覆盖率。

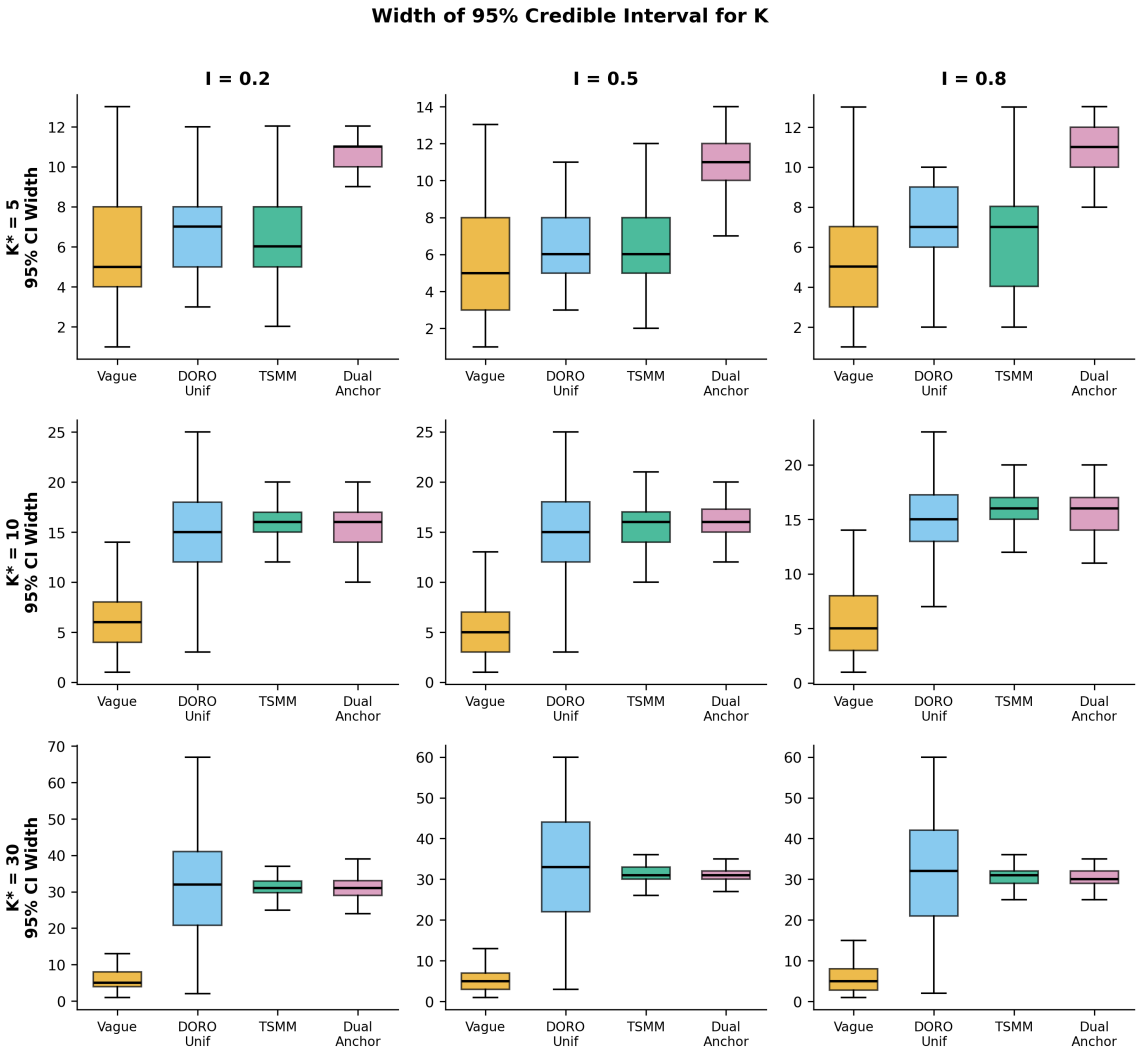


图 F.6. 按方法和情景划分的 K_J 的 95% 可信区间宽度。箱线图展示了 200 次重复实验中区间宽度的分布。无论真实的 K^* 为何，模糊先验都会产生较窄的区间（约为 5–6）；而 TSMM 和双锚点会产生与 K^*相适应、宽度随之变化的区间。

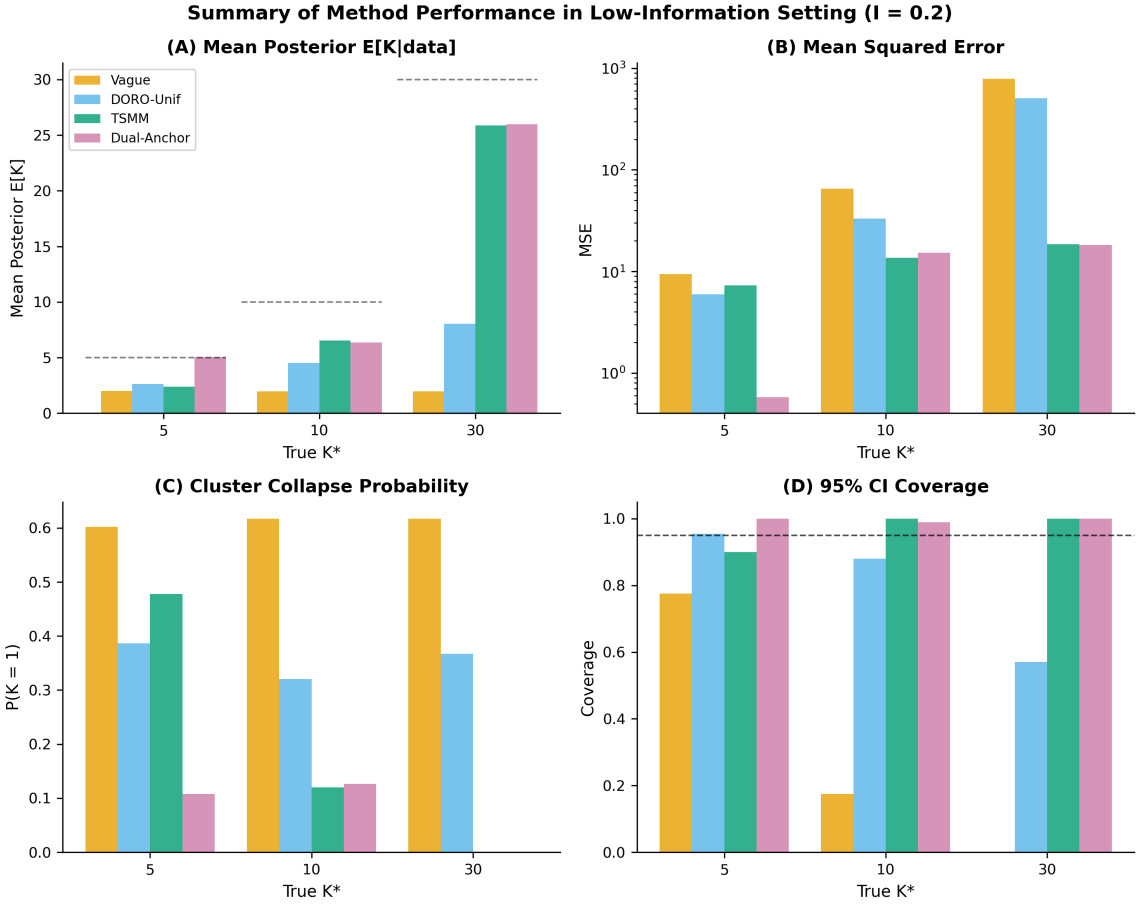


图 F.7. 低信息设置（$I = 0.2$）下的方法性能汇总。面板 (A)：后验 $\mathbb{E}[K_J \mid \text{数据}]$ 均值，虚线表示真实的 K^*。面板 (B)：均方误差（对数刻度）。面板 (C)：簇坍缩概率 $P(K_J = 1)$。面板 (D)：95% 置信区间覆盖率，虚线表示名义 95% 水平。

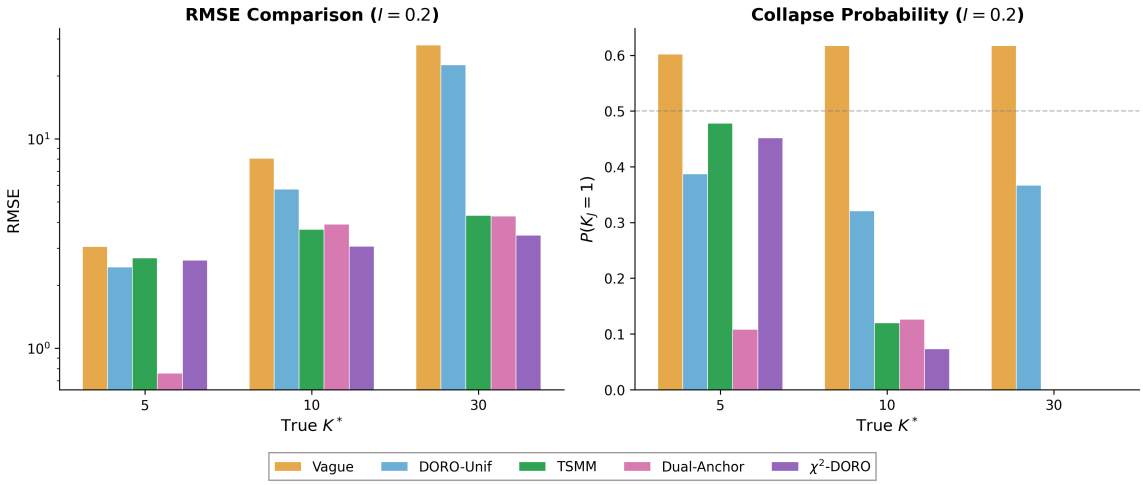


图 F.8. 包含针对 $I = 0.2$ 的 χ^2-DORO 基线的比较。左图：均方根误差（对数刻度）。右图：坍缩概率。对于 $K^* \in \{10, 30\}$，χ^2-DORO（紫色）通常优于 DORO-Unif 和 TSMM，但仍存在显著的低估；当 $K^* = 5$ 时，坍缩概率也不可忽略——这正是双锚点旨在介入的情形。

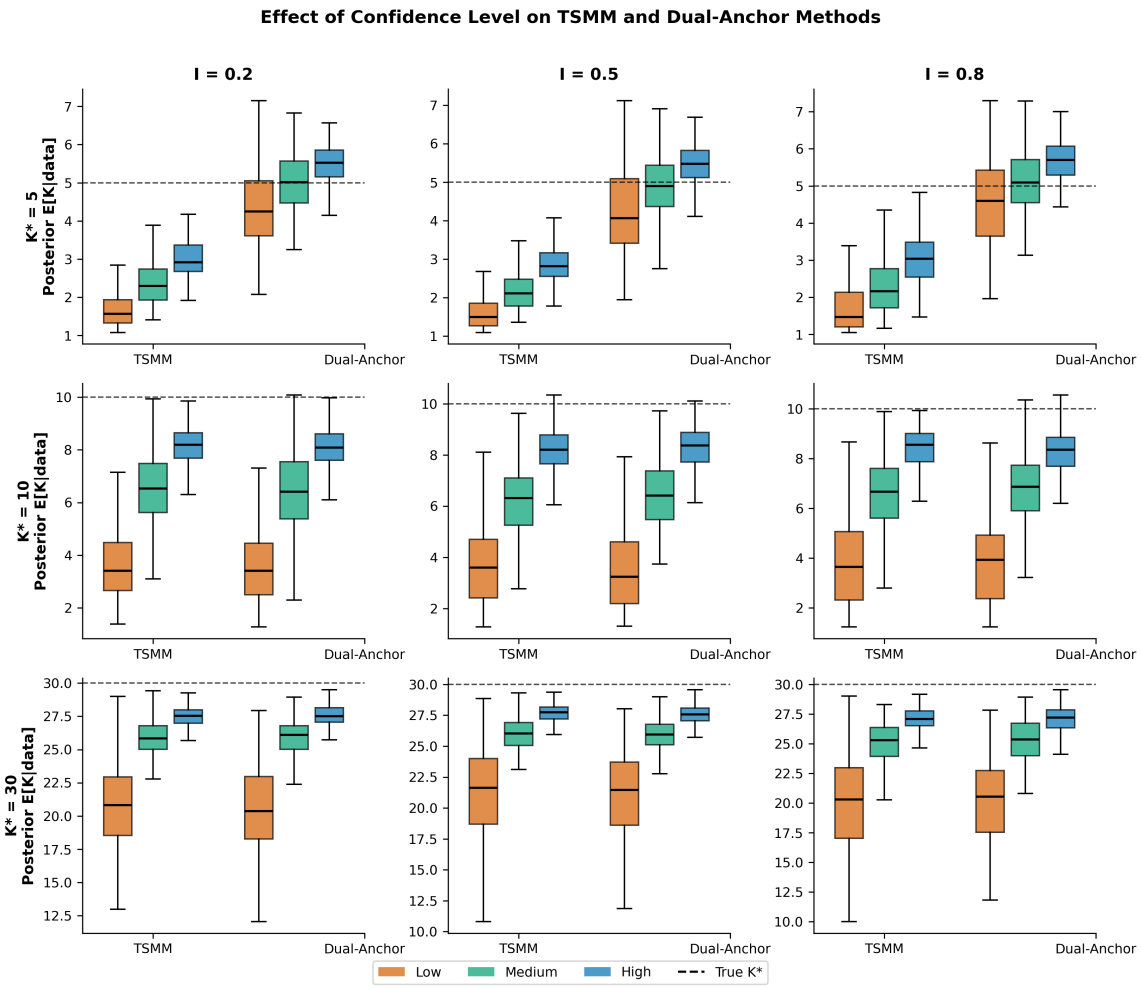


图 F.9. 置信水平对 TSMM 和双锚点方法的影响。行：真实的 $K^* \in \{5, 10, 30\}$；列：信息量水平$I \in \{0.2, 0.5, 0.8\}$。每个面板中，箱线图展示低（橙色）、中（Green）和高（蓝色）置信度设置下的后验分布 $\mathbb{E}[K_J \mid \text{数据}]$ 。虚线表示真实的 K^*。

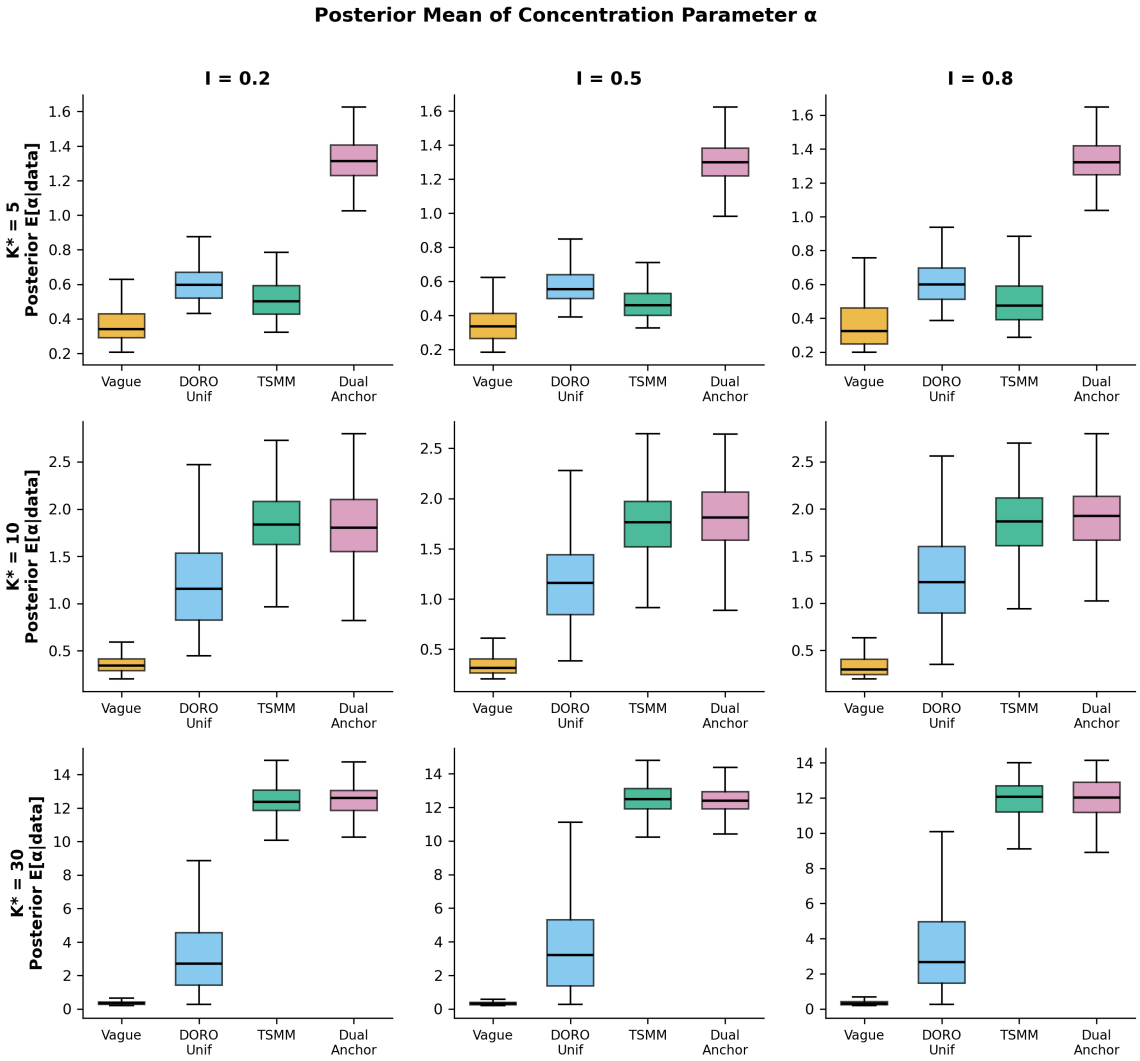


图 F.10. 按方法和情景划分的浓度参数后验均值（中等置信度）。行：真实 $K^* \in \{5, 10, 30\}$；列：信息量水平 $I \in \{0.2, 0.5, 0.8\}$。请注意各行之间不同的 y 轴刻度。

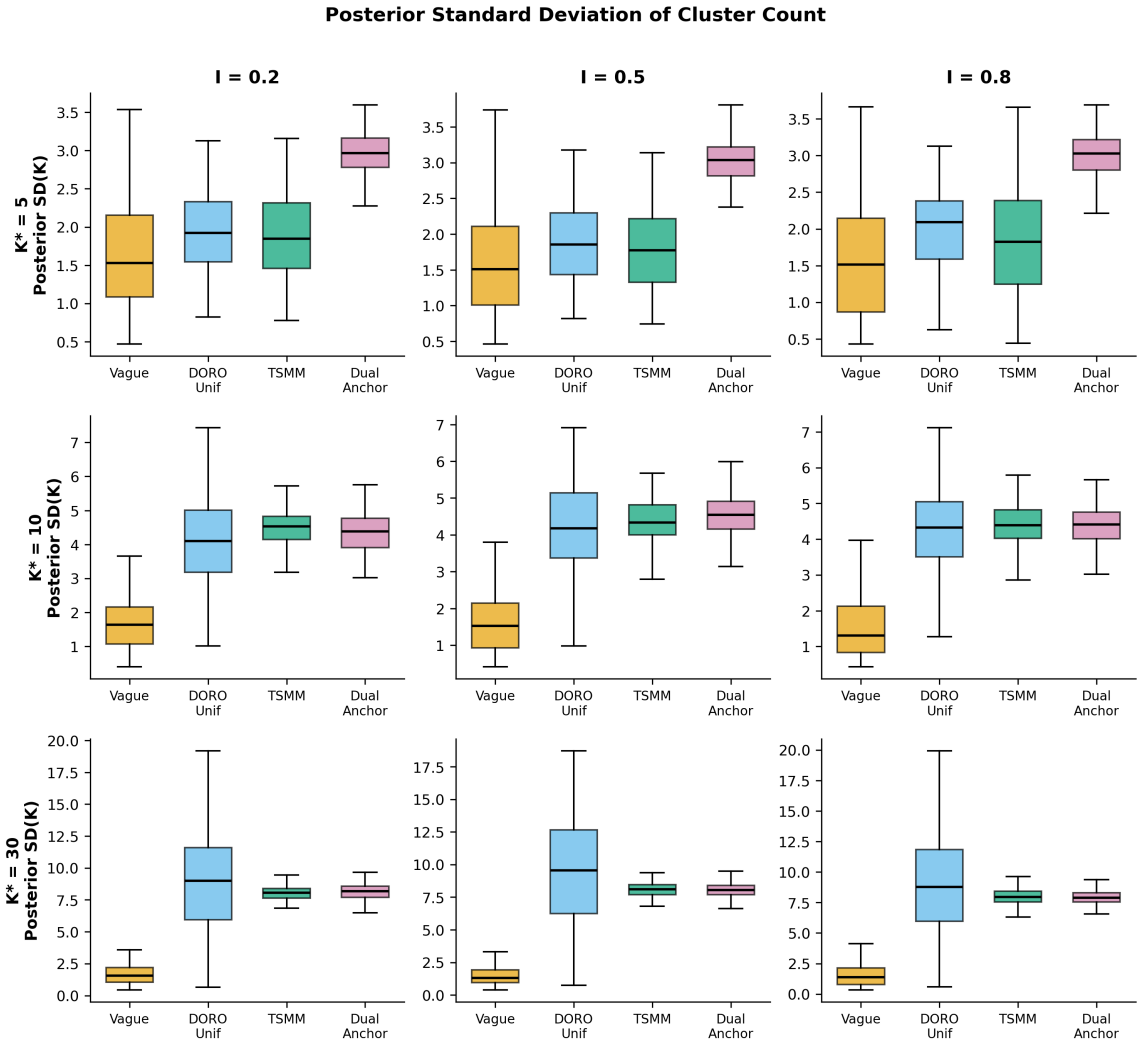


图 F.11. 簇数 K_J 的后验标准差（按方法与情景）。无论 K^*如何，模糊先验都会产生始终较窄的后验分布（低不确定性），而 TSMM 与双锚点则产生随 K^*增大而适当缩放的不确定性。

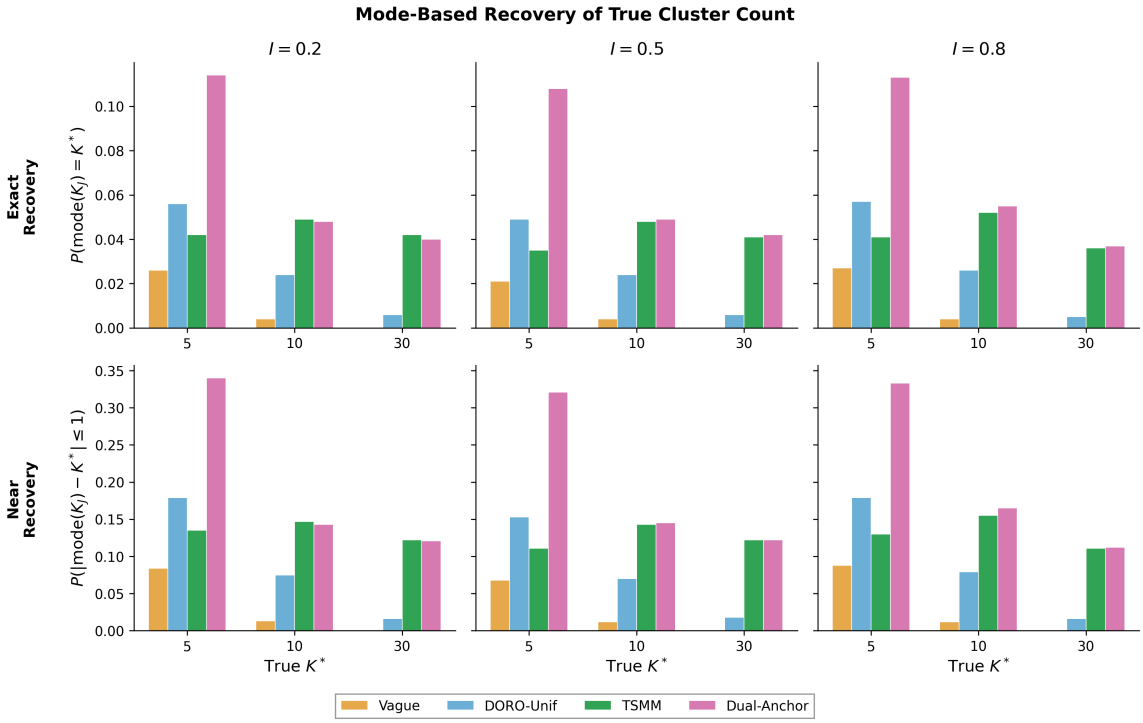


图 F.12. 基于众数的真实簇数恢复。上排：精确恢复率 $P(\text{众数}(K_J) = K^*)$。下排：近似恢复率 $P(|\text{众数}(K_J) - K^*| \leq 1)$。列：信息量水平 $I \in \{0.2, 0.5, 0.8\}$。

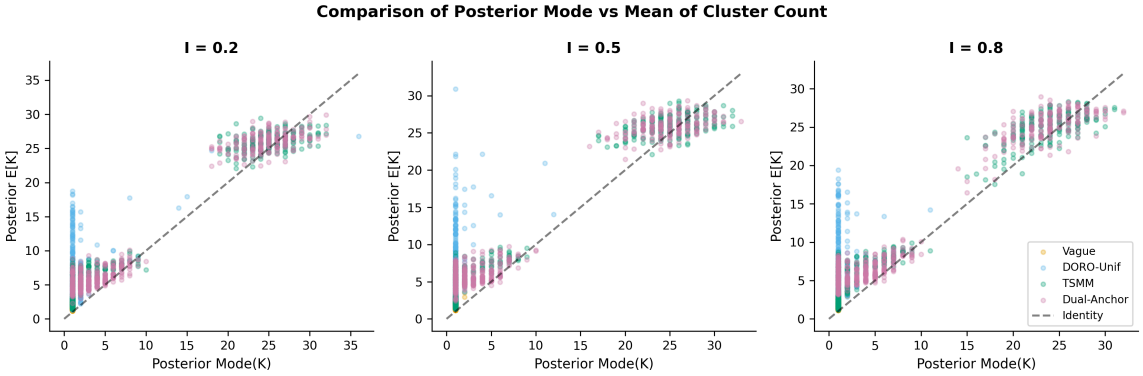


图 F.13. 簇数的后验众数与均值比较 K_J。每个点代表一次重复试验。虚线对角线表示众数等于均值的位置。高于该线的系统性偏差表示后验均值大于众数（后验分布右偏），而低于该线的点则表示相反。

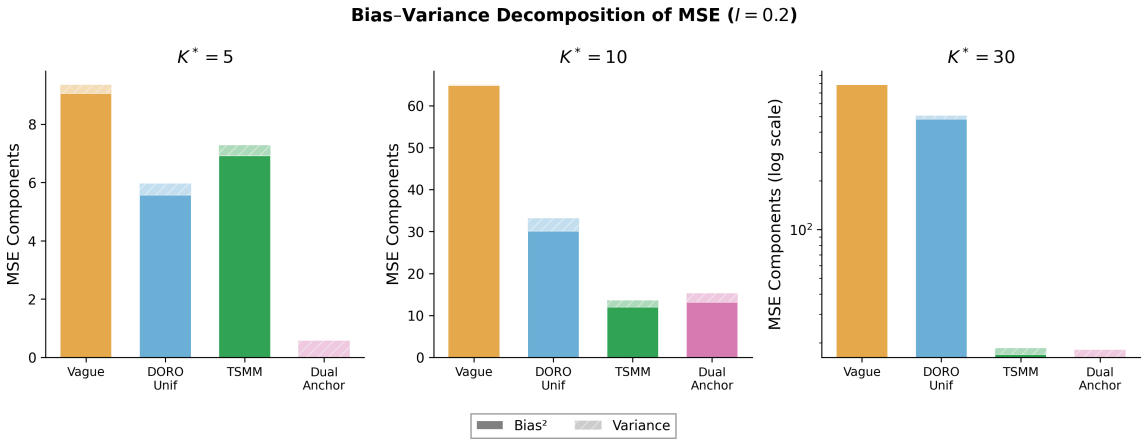


图 F.14。 $I = 0.2$ 的均方误差偏差–方差分解。实心条表示偏差^2；网纹部分表示方差。对于 $K^* = 30$，使用对数刻度。方法颜色：模糊（黄色）、DORO-Unif（蓝色）、TSMM（绿色）、双锚点（粉色）。