arXiv:2301.01677v2 [stat.AP] 6 Jan 2023

贝叶斯混合模型捕捉纵向公投数据中投票集团的时间和空间结构

约翰·D·奥布莱恩

鲍登学院数学系

2023年1月9日

摘要

对投票集团进行估计是政治学中的一项重要统计研究。然而，这类分析通常局限于点名表决数据，因为其中可以直接观察到个人投票。在此，我们考察一种由狄利克雷—多项式分量构成的贝叶斯混合模型，用于从市级汇总的纵向公投数据中推断投票集团。作为案例研究，我们分析了美国缅因州423个市镇的投票总数，这些数据涉及2008年至2019年间付诸表决的54项公投问题。利用该模型，我们估计了投票集团数量的后验分布、各集团对每项问题的支持度，以及各市镇中不同集团的混合情况。我们发现，这些投票集团具有地理结构，并且在整个研究期间基本保持一致。对后验分布的进一步分析还得出三项发现：投票集团的整体结构既呈现渐变，也存在断层，这与地理和文化因素相符；少数问题与全州范围的集团结构不一致，且这些问题的内容与特定地区有关；研究期间还出现了集团间极化加剧的迹象。最后，我们概述了该模型在统计学上的扩展方向、它与政治学中其他统计框架（如民意调查）的联系，并详细介绍了适合后续应用该模型的候选地点。

1 引言

1.1 投票集团统计模型

只要选民群体中有不止少数几个人，就必然会混合着不同的政治观点。哪些因素能够解释这种倾向差异的结构及其持续性？这些因素又如何与投票行为相联系？这是政治行为研究中的一个根本问题，既关乎政治学，也关乎整个社会科学领域 [May，1973年，Gunn，1995年，Leeper 和 Slothuus，2014年，Weidlich，1994年，Weakliem 和 Biggert，1999年，Converse，1987年，Gallup 和 Rae，1940年，Kinder， 1998]。因此，大量统计学文献致力于探明可能的解释变量与总票数之间的关联。这一总体思路——从解释变量出发，说明聚合投票行为的各个方面——体现在政治学中常用的统计框架之中，例如广义回归、生态回归和潜在类别分析 [King 等，2004年，Schuessler，1999年，Beck 和 杰克曼，1998年，King，1988年，McCutcheon，1987年，Magidson 等， 2020] ；这些框架都孕育了丰富的统计创新与讨论 [Gelman 等，2001年，Lanza 等，2013年，King，1990年，Beck，2000]。

在此，我们采用另一种方法来理解汇总投票总数背后的结构：使用贝叶斯混合模型，在投票数据本身中识别内部一致的关联模式 [Fr¨uhwirth-Schnatter，2006年，McLachlan和Basford，1988年，McLachlan等，2019年，Fruhwirth-Schnatter等， 2019]。就此而言，这种方法与遗传学、宏基因组学和主题建模领域的模型更为接近；在这些领域中，混合建模被用作理论解释框架的经验性补充 [Falush等，2003年，Hellenthal等，2014年，Holmes等，2012年，Blei等， 2003]。在政治学领域，我们的方法承接了Gormley和Murphy [2008a,b]，的研究，他们开发了一组混合模型，用于从排序选择选举数据中识别投票集团——即投票偏好相似的选民群体[Gormley和Murphy，2008年a、b，2006年，Gormley和Fr¨uhwirth-Schnatter， 2019]。在排序选择投票（RCV）中，选民会对候选人（最多包括所有候选人）进行排序。利用这些数据，Gormley和Murphy采用了一种利用这些偏好之间协方差结构的方法，从而推断出爱尔兰总统选民群体中的四个投票集团、各集团的投票偏好，以及它们在总体中的比例。在后续研究中，他们展示了如何使用混合专家模型将外部协变量与投票集团关联起来 [Gormley和Murphy， 2010]。遗憾的是，这种方法在政治学中的广泛应用可能受到限制，因为RCV选举相对少见，因此这类数据也较难获取。本文旨在

将他们的方法扩展到两个方面：（1）使混合模型适用于公投投票总数这一类更常见的选举数据；（2）捕捉这些数据中此前未加考察的空间和时间变异。为探究该模型阐明选民行为结构的能力，我们对美国缅因州在2008年至2019年间投票表决的54项公投问题的市级投票总数进行了案例研究。

尽管直接根据汇总投票总数推断投票集团等潜在结构的模型并不常见，但关于如何利用唱名表决及类似数据估计投票集团，以及更广义的潜在空间结构，已有大量研究；在这类数据中，可以观察到每位参与者的投票，例如议会、司法合议庭和公司董事会中的投票 [Hix等，2006年；Clinton等，2004年；Katz和King，1999年；Bailey等，2017年；D'Angelo等，2019年；Reuning等， 2019]。关于唱名表决中所用模型的详细综述，参见McAlister [2020]。这些研究也经常采用本文所用混合模型中的一些元素，不过还存在许多其他方法，包括主成分分析、层次聚类、潜在类别分析和拓扑数据分析 [Amelio和Pizzuti，2012年；Bakk等，2014年；Holloway，1990年；De Leeuw，2006年；Vejdemo-Johansson等， 2012]。这些研究与本文的关键区别在于，它们分析的是能够直接观测到个体投票的情境；而在公投选举数据中，投票总数必然反映了个体投票在某一级政府辖区（例如本文中的市镇）内的汇总。利用潜在空间模型及其相关方法来理解公投数据并非全无研究：近期有两项研究试图理解瑞士公投结果，分别采用了基于网络的方法和潜在类别模型 [Mantegazzi，2021年；Koseki， 2018] 此外，一篇较早但未发表的关于加利福尼亚州公投结果的论文，也采用了包含与Gormley和Murphy的方法以及本文所考虑方法相似元素的分析方案 [Dubin和Gerber，1992]。

为从按市镇层级汇总的公投投票总数中揭示投票集团的结构，我们采用贝叶斯混合模型。遵循Gormley和Murphy的做法，我们用"投票集团"来描述投票模式相似的选民。投票模式是一个向量，其中包含对每个问题的支持度分布。在构建模型时，我们对数据作出以下假设：州内所有市镇共有有限但数量未知的投票集团；市级投票数据具有组成数据性质，且混合比例固定；投票模式能够充分表征各市镇中观察到的投票总数[Aitchison，1982]。我们还假设各市镇内的投票集团相互独立地混合，并且每种投票模式均由一个狄利克雷-多项分布定义，用以描述对

每个问题的支持度及其离散程度。在此案例研究中，我们仅关注是/否票数，因此狄利克雷-多项分布转变为Beta-二项（BB）分布。最后，我们假设投票模式内各问题对应的BB分布彼此独立。随后，对每个市镇内投票集团的混合比例进行积分，即可得到该市镇中各问题观察到的支持度。由此，推断目标是恢复投票集团数量的分布、其投票模式的结构（即对每个问题的支持度分布），以及各市镇内每个投票集团的混合比例。

确定投票集团的总体数量本身就是一个值得关注的问题，同时也是对投票集团数量的不确定性进行积分的一种方式。我们改进了一种出生—死亡马尔可夫链蒙特卡洛（BD-MCMC）技术，以推断集团数量的后验分布。该方法构建了一条连续时间马尔可夫链，在每次迭代中，它会在不同的投票集团数量上“等待”一段时间，等待时长与后验密度成正比。通过根据各集团数量对应的等待时间加权，即可恢复最终的后验密度。在此情境下，与其他替代方法相比，BD-MCMC 具有显著优势：相较于可逆跳跃方法，混合速度更快；不同运行结果之间的一致性更强；其后验分布也易于解释。

1.2 缅因州公投数据

公投是最直接的选举民主参与形式之一，要求符合资格的选民就特定治理问题进行全民投票，议题范围从债券到宪法修正案不等[Mendelsohn 和 Parkin，2001年，Bowler 等人， 1998]，如今，政府越来越常通过公投来决定存在公众争议的问题。至少有 20 个国家、美国 26 个州以及数百个城市会在国家层面举行公投选举[1] 本稿将两种最常见的公投形式——公民倡议，即公民绕过立法机构，直接向选民征询意见；以及公民公投，即民众对立法机构通过的法案予以批准或否决——视为同一类，尽管在此考察的数据中，前一种形式的出现频率远高于后一种（54 个问题中有 51 个）。我们仅关注赞成票和反对票，因为原始数据中空白选票的记录方式不一 [缅因州公司、选举与委员会局， 2021]。

The 来自美国缅因州的2008—2019年11月公投数据具备若干 l

[1]这些数字是根据维基百科和 Ballotpedia 得出的。我们没有找到更全面考察这一问题的学术来源。

从建模角度看颇具吸引力的特征：选票问题数量相对较多；研究期间市政边界和投票程序保持一致；人口结构总体稳定。缅因州是美国东部第一个采用公民发起公投的州，并拥有活跃的此类选举传统[Scontras, 2016]。自1980年以来，该州已举行了200多项公投问题的投票；直至2020年新冠疫情相关限制措施妨碍选举程序之前，每年11月都至少有一项选票问题。从市镇组织来看，缅因州采用新英格兰城镇模式，以市镇作为地方治理的主要架构；与美国其他地区常见的紧凑型聚居地或未建制领地相比，这两类地区在缅因州的人口都微不足道。该州几乎所有城镇都在20世纪初之前建立，研究期间仅出现了一个新城镇（一个岛屿从其所在的大陆社区分离出来）。缅因州选举局在整个研究期间基本保持了投票记录的一致性，而质量相近的纸质记录最早可追溯至 19th 世纪 [Archives, 2022]。我们于 1, 2020 [June 2021]从缅因州公司、选举与委员会局网站下载了研究样本中每次11月选举的公投数据。我们按照下文所述流程处理这些原始文件，并将清理后的文件放在本项目的 GitHub 页面上：
https://github.com/cascobayesian/maineelections。

缅因州包含721个市镇，分为城市、城镇、种植园[2], 和乡镇，此外还有一组人口数量可忽略的非建制地区。每个数据集所包含的市镇数量各不相同，这主要取决于在某次选举中是否将乡镇或种植园并入邻近市镇。为统一各年份的市镇数量，如果某个市镇与邻近市镇在研究期间的某次选举中合并，我们便将二者合并。对于非建制地区的投票总数，以及在数据集内投票站设置不一致的乡镇或种植园，我们也予以剔除。任何对某个问题未登记反对票的剩余市镇，均被视为错误，并从分析中删除。经过这一剔除过程，我们得到了一份包含423个市镇的数据集。

公投最常在11月选举中进行投票；届时还会举行国会、州长和总统选举，但公投也会在6月的党派初选期间举行。在2008年—2019年研究期间，共有61个问题进行投票表决，其中7个在6月，54个在11月。从历史上看，6月选举既有较低的

[2]“种植园”这一称谓是缅因州特有的一种独特历史用法，指的是一种市镇自治单位，其权力小于城镇，但大于乡镇。

相较于十一月选举，投票率的变异性很大，因此我们将注意力限制在十一月的问题上。表1提供了有关研究中所纳入问题的更多信息。

Year	Num.	总票数	总票数中位数	内容
2008	3	707,915	805	医疗保健税 ，赌场 ，水资源债券
2009	7	562,985	673	同性婚姻、汽车能效返利， 学校重新划分选区、政府支出 医用大麻、交通债券、选票改革
2010	3	554,783	639	赌场 ，牙科保健债券 ，环境保护债券
2011	4	387,881	440	当日选民登记、赌场， 老虎机设施、选区重划变更
2012	5	697,423	788	同性婚姻、教育债券、下水道债券， 交通债券、水务债券
2013	5	214,507	213	国民警卫队债券、两项大学债券， 交通债券、社区学院债券
2014	7	597,224	694	熊类狩猎、农业债券、小企业债券， 科学债券 ，水务债券 ，企业孵化器债券
2015	3	216,248	230	廉洁选举、住房债券、交通债券
2016	6	748,337	848	大麻合法化、所得税增加、 枪支背景审查、最低工资
2017	4	341,139	363	赌场、医疗补助扩展、交通债券、 政府金融改革
2018	5	626,470	715	居家护理项目、下水道债券、交通债券、 大学债券、社区学院债券
2019	2	186,207	167	交通债券、无效签名

表1：研究期间每张选票上公投的摘要。总票数按每张选票上的最高总票数计算。总票数中位数指数据集中各市镇总票数的中位数。

1.3 论文结构

本文旨在开发一种贝叶斯混合模型，以推断十二年市政公投数据中的投票集团，并对结果进行解读。我们首先介绍一些符号说明，然后对模型进行生成式描述。在讨论推断时，我们介绍一种马尔可夫链蒙特卡罗 (MCMC) 方法，该方法采用双重替换方案高效地抽取参数；我们还开发了一种 BD-MCMC，用于推断投票集团数量及其参数的后验分布。作为案例研究的补充，我们开展模拟，以了解该算法在市镇数量、问题数量和投票总数方面的推断要求。

我们首先展示所有问题合并分析的案例研究结果，然后按选举周期分别展示。我们发现，投票集团在地理上呈现结构，并且在不同选举周期中高度一致；这些投票集团既独立组织，也沿梯度分布。作为一致性检验，我们表明这些投票集团与数据的信息论投影相一致。通过对模型进行后验检验，我们按市镇和问题考察模型的表现，并发现少数“局部”问题不符合模型的假设。最后，我们讨论模型的若干后续发展方向、与其他理解投票数据的统计方法之间的联系，以及其他适用地区和数据集。

2 数据与模型

在接下来的两个小节中，我们将介绍数据的符号说明，以及用于分析这些数据的统计模型。这两个小节涵盖了本文后续大部分所使用的符号。为便于查阅，我们在表2中汇总了这些符号。

2.1 数据符号说明

423个始终一致的市镇的投票表（是/否）构成了本研究使用的完整数据集（表1）。该完整数据集还按四年选举周期细分，起始于2008—2011年，结束于2016—2019年。我们将城镇编号为 i 至$i = 1$、$\cdots$、$N = 423$。我们将问题从 q $q = 1$、$\cdots$、54进行索引，并用 $r \in \{0, 1\}$ 表示我们指的是“是”票计数还是“否”票计数。对于每个城镇 i，观测到的计数表示如下：

$$\mathbf{c}_{iqr} = (y_{iqr}, n_{iqr}),$$

符号说明	解释
Q	公投问题数量
N	市镇数量
K	投票集团数量（VB）
$q = 1, \cdots, Q$	公投问题索引
$i = 1, \cdots, N$	市镇索引
$k = 1, \cdots, K$	投票集团索引
$\mathbf{c}_{iq} = (c_{iq0}, c_{iq1})$	市镇 i 和问题 q的赞成票与反对票总数
$\mathbf{c} = [\mathbf{c}^{iq}\colon\ i = 1, \cdots, N;\ q = 1, \cdots, Q]$	所有市镇和问题的数据
$\boldsymbol{\alpha}_{kq} = (\alpha_{kq0}, \alpha_{kq1})$	投票群体 k 和问题 q的贝塔-二项分布参数
$\boldsymbol{\alpha} = [\boldsymbol{\alpha}_{kq}\colon\ k = 1,\ \cdots,\ K;\ q = 1,\ \cdots,\ Q]$	所有贝塔-二项分布参数
$\boldsymbol{z} = [z_1, \cdots, z_N],\ z_i \in \{1, \cdots, K\}$	用于将市镇划分至投票集团的辅助变量
η_k	市镇属于投票群体 k的先验概率
$\kappa,\ \theta$	每个 $\boldsymbol{\alpha}_{\boldsymbol{kq}}$的先验参数

表2：数据统计模型参数的符号说明与解释。

其中， $y.$ 表示选票问题的“是”票数， $n.$ 表示对应于 q的“否”票数。我们用 $\boldsymbol{c}$ 表示所有数据的汇总；随后，我们进一步用 $\boldsymbol{c}_{all}$ 表示完整数据，用 $\boldsymbol{c}_i$ 表示从该年份开始的四年周期（例如 $\boldsymbol{c}_{2008}$），适用于每个四年周期。

2.2 贝塔-二项混合模型

现在我们开始介绍用于数据的贝叶斯模型，首先从数据似然入手。如上文所述，我们采用有限混合模型来刻画市政公投数据中的投票集团。该模型假设存在由对所有问题的支持度分布所定义的 $k = 1, \cdots, K$ 个投票集团。为便于初步说明，我们将 K 视为固定值，随后再扩展模型，将其作为随机变量处理。每个市镇的混合比例由 λ_{ik} 指定，并受自然约束 $\sum_{k=1}^{K} \lambda_{ik} = 1$ 的限制。因此，推断建模的目标是恢复投票集团的数量（K）、每个投票集团所属投票模式的参数，以及各市镇内的混合比例。

对于每个 k 投票集团中的每个问题 q，我们使用贝塔-二项分布

由参数$\boldsymbol{\alpha}_{k,\ q}=(\alpha_{kq0},\ \alpha_{k,\ q,\ 1})$指定，以描述 c_{iq}的分布。贝塔-二项分布通过对贝塔分布上的支持比例 p 进行积分，对更常见的二项分布进行了推广。该贝塔分布由 α_{kq0} 和$\alpha_{k,\ q,\ 1}$参数化，从而得到观测计数的概率：

$$\begin{aligned}\mathbb{P}(\boldsymbol{c}_{iq}|\boldsymbol{\alpha}_{kq}) &= \int_0^1 \text{BINOMIAL}(c_{iq0}|c_{iq0}+c_{iq1},p)\cdot\text{BETA}(p|\alpha_{kq0},\alpha_{kq1})dp \\ &= \frac{1}{B(\alpha_{kq0},\alpha_{kq1})}\int_0^1 \binom{c_{iq0}+c_{iq1}}{c_{iq0}} p^{c_{iq0}}\cdot(1-p)^{c_{iq1}}\cdot p^{\alpha_{kq0}-1}\cdot(1-p)^{\alpha_{kq1}-1}dp \\ &= \binom{c_{iq0}+c_{iq1}}{c_{iq0}}\frac{B(c_{ik0}+\alpha_{kq0},c_{ik0}+\alpha_{kq1})}{B(\alpha_{kq0},\alpha_{kq1})},\end{aligned}$$

其中， $B(\alpha,\beta)$ 是贝塔函数，BETA(α,β) 是贝塔分布。该分布的优点是能够灵活地刻画平均支持度及其离散程度，同时允许存在有限的双峰性。

由于有 K 个投票集团，因此市镇 i 和问题 q 的数据概率是每个 k 对应的贝塔-二项密度之和，并以 λ_{ik} 加权：

$$\mathbb{P}(\mathbf{c}_{iq}|\boldsymbol{\alpha}_{\cdot q})=\sum_{k=1}^{K}\lambda_{ik}\mathbb{P}(\mathbf{c}_{iq}|\boldsymbol{\alpha}_{kq}). \tag{1}$$

然而，由于该概率中的求和带来了显著的推断挑战，我们引入标准变量 $z_i\in\{1,\cdots,K\}$，用于 $i=1,\cdots,N$，以指定每个市镇 i所属的投票阵营。随后，我们旨在恢复每个 i对应的 z_i 的后验分布，并据此通过注意到 $\mathbb{P}(z_i=k|\mathbf{c}\ \boldsymbol{\alpha})=\lambda_{ik}$，恢复方程 1 中的成分比例。这一表述要求我们指定市镇属于成分 k的先验概率，我们将其记为 η_k，并要求$\sum_{k=1}^{K}\eta_k=1$。

辅助变量形式使得完全数据似然可以采用高效形式表示：

$$\begin{aligned}\mathbb{P}(\mathbf{c}_{iq}|\boldsymbol{\alpha},z_i) &= \mathbb{P}(z_i=k)\cdot\mathbb{P}(\mathbf{c}_{iq}|\alpha_{z_i,q}) \\ &= \eta_{z_i}\mathbb{P}(\mathbf{c}_{iq}|\alpha_{z_iq\cdot}).\end{aligned}$$

由于假设在给定辅助变量 z_i所指定的贝塔-二项分布后，每个市镇针对每个问题的投票分布均条件独立，因此，给定 $\boldsymbol{\alpha}$ 参数和辅助变量 $\boldsymbol{z}=(z_1,\ \cdots,\ z_N)$后，完整数据似然便成为所有问题和市镇上的乘积。由于 z_i 将市镇与一个投票阵营关联起来，且这一关联适用于所有问题，因此似然变为

$$\begin{aligned}\mathcal{L}(\mathbf{c}|\boldsymbol{\alpha},\boldsymbol{z}) = \mathbb{P}(\mathbf{c}|\boldsymbol{\alpha},\boldsymbol{z}) &= \prod_{i=1}^{N}\mathbb{P}(\mathbf{c}_{iq}|\boldsymbol{\alpha}_{z_i q},z_i)\\ &= \prod_{i=1}^{N}\eta_{z_i}\left\{\prod_{q=1}^{Q}\mathbb{P}(\mathbf{c}_{iq}|\boldsymbol{\alpha}_{z_i q})\right\}\\ &= \prod_{i=1}^{N}\eta_{z_i}\left\{\prod_{q=1}^{Q}\binom{c_{iq0}+c_{iq1}}{c_{iq0}}\frac{B(c_{iz_i0}+\alpha_{kq0},c_{iz_i1}+\alpha_{z_i q1})}{B(\alpha_{z_i q0},\alpha_{z_i q1})}\right\}.\end{aligned} \tag{2}$$

这一似然既可以理解为有限混合模型，其中将问题视为相互独立（但并非同分布）的数据重复样本；也可以理解为退化隐马尔可夫模型，其中类别在各年份保持不变。我们将在讨论可能的模型扩展时，回到这两种视角。

先验设定与生成模型

要完成贝叶斯模型，需要为每个参数设定先验分布。由于 $\boldsymbol{\alpha}$ 是 $\boldsymbol{\alpha}_{kq} = (\alpha_{kq0},\ \alpha_{kq1})$ 的聚合，对于 $k = 1,\ \cdots,\ K$ 和 $q = 1,\ \cdots,\ Q$，我们假设这些参数分别独立地从形状参数为 κ 、尺度参数为 θ 的伽马分布中抽取。由于 κ 和 θ 对于 α_{kq0} 和 α_{kq1} 是相同的，这相当于对某个问题的支持比例设置了一个集中于 $_{1\ 2}$ 的弱信息先验。对于此处的所有运行，$\kappa = 1$ 和 $\theta = 10$。

–

如上文所述，z_i 取值为 k 的先验概率为 η_k，由此完成先验设定。向量 $\boldsymbol{\eta} = (\eta_1,\ \cdots,\ \eta_K)$ 是一组比例之和，因此总和为一。在缺乏关于各组分相对权重的任何信息时，以 $\mathbf{1}_K$（一个由 K 个 1 组成的向量）为参数的对称狄利克雷分布，提供了合理的弱信息先验。我们通过假设 K（即投票模式数量）服从均值为 λ 的泊松随机变量分布，完成模型的先验设定。对于此处考虑的所有运行，我们设定 $\lambda = 10$。该模型可简洁地概括为以下生成过程：

$$\begin{aligned}K &\sim \text{POISSON}(\lambda)\\ \boldsymbol{\eta}\mid K &\sim \text{DIRICHLET}(\gamma_K)\\ z_i\mid\boldsymbol{\eta} &\sim \text{CATEGORICAL}(\boldsymbol{\eta})\\ \alpha_{kqs} &\sim \text{GAMMA}(\kappa,\theta)\\ \boldsymbol{c}_{iq}|\boldsymbol{\alpha},z_i &\sim \text{BETA-BINOMIAL}(\alpha_{z_i,q,0},\alpha_{z_i,q,1})\ ,\end{aligned}$$

其中，γ_K 表示一个 K γs的向量。

2.3 推断

我们采用马尔可夫链蒙特卡罗（MCMC）方法从模型的后验分布中生成样本。除下文详述的标准吉布斯更新步骤外，我们还使用了两种更为复杂的技术，以克服在这一问题中遇到的具体挑战：采用出生-死亡马尔可夫链蒙特卡罗（BDMCMC）方法求取投票集团数量 K 的后验分布，以及采用定制的变量增广方案推断贝塔-二项分布的参数。我们首先解释其他参数的吉布斯更新。

2.3.1 吉布斯更新

$\boldsymbol{\eta}$：生成模型表明

$$\begin{aligned}\mathbb{P}(\eta|K,\boldsymbol{z},\boldsymbol{\alpha},\boldsymbol{c},\boldsymbol{\gamma}_K) &\propto \left\{\prod_{i=1}^{N}\eta_{z_i}\right\}\text{DIRICHLET}(\eta|\gamma_K)\\ &\propto \left\{\prod_{k=1}^{K}\eta_k^{d_k}\right\}\text{DIRICHLET}(\eta|\gamma_K)\\ &\propto \text{DIRICHLET}(\eta \mid \boldsymbol{d}_K+\gamma_K)\end{aligned}$$

其中，$d_k = \sum_{i=1}^{N}\mathbf{1}_{z_i=k}$ 是投票阵营 k 中市镇的数量，$\mathbf{1}.$ 是指示函数，而 $\boldsymbol{d}_K$ 是 d_k 的向量。

$\mathbf{z}$：我们对潜变量 z_i采用标准的吉布斯更新。检查生成过程后，我们观察到

$$\begin{aligned}\mathbb{P}(z_i=k|\boldsymbol{z}_{-i},\boldsymbol{\alpha},\boldsymbol{c},\boldsymbol{\gamma}_K,\eta,K) &\propto \mathbb{P}(z_i=k|\boldsymbol{z}_{-i},\boldsymbol{\alpha},\boldsymbol{c},\eta)\\ &\propto \mathbb{P}(\boldsymbol{c}_i|\boldsymbol{\alpha}_k,\eta)\\ &\propto \eta_k\cdot\text{BETA-BINOMIAL}(c_i|\boldsymbol{\alpha}_k)\\ &= \frac{\eta_k\cdot\text{BETA-BINOMIAL}(c_i|\boldsymbol{\alpha}_k)}{\sum_{j=1}^{K}\eta_j\cdot\text{BETA-BINOMIAL}(c_i|\boldsymbol{\alpha}_j)}.\end{aligned}$$

2.3.2 变量增广与针对 α 的 Metropolis-Hastings 抽样器

贝塔-二项分布是狄利克雷-多项分布（DM）的二类别限定形式。DM 分布在宏基因组学、文本分析、生态学和主题建模中的广泛应用，推动了针对该分布参数的推断方法的发展；对于通用的马尔可夫链蒙特卡洛抽样器而言，这些参数通常具有较慢的收敛特性。通常，推断方法依赖于结合了

期望最大化算法的某个版本 [Mimno 和 McCallum，2008；Li 等人，2019；Holmes 等人，2012]。在本文中，我们改用 Stein 和 Meng [Stein 和 Meng， 2013] 提出的变量增广方案，该方案能够对狄利克雷-多项式参数进行高效的马尔可夫链蒙特卡洛抽样，从而可以更充分地探索后验分布，而不是围绕后验众数进行近似。这种抽样方法还可以轻松结合下面所述的、针对 K 的 BD-MCMC 推断方法。

Stein 和 Meng（SM）针对一组狄利克雷-多项式参数提出了他们的框架。不过，将其扩展到混合模型的情境中相当直接：我们将他们的变量增广方案分别应用于每个混合成分，并独立更新每个 $\boldsymbol{\alpha}_k$ 。为便于说明，假设市镇属于投票群体 k，即$z_i = k$。SM框架为每个观测数据计数 c_{iqr} 增广一个值 r_{iqr}，该值是一个由非同分布伯努利随机变量组成的向量之和，$\nu_{iqr,\ m}$，这些随机变量与狄利克雷-多项式参数及数据相关：

$$r_{iqr} = \sum_{m=1}^{c_{iqr}} \text{BERNOULLI}\left(\frac{\alpha_{kqr}}{\alpha_{qkr} + m - 1}\right) = \sum_{m=1}^{c_{iqr}} \nu_{iq,m}. \tag{3}$$

与上文在潜变量混合模型增广中采用的方法类似，这种增广的作用是将求和转化为乘积，具体而言，就是处理嵌入狄利克雷-多项式似然中的伽马函数内的求和（例如参见方程 2，并注意贝塔函数的定义）。对于有兴趣了解这一因式分解如何实现的读者，请参阅 SM 论文中围绕方程（1-3）和（7）的讨论。变量增广与狄利克雷-多项式似然相乘后，得到一个能将伽马函数内的求和因式分解的完全数据似然：

$$\begin{aligned}\mathbb{P}(\boldsymbol{c}|\boldsymbol{\alpha}, \boldsymbol{z}_i = k) \cdot \mathbb{P}(\boldsymbol{\nu}|\boldsymbol{c}, \boldsymbol{\alpha}, \boldsymbol{z}_i = k) &\propto \mathbb{P}(\boldsymbol{c}, \boldsymbol{\nu}|\alpha, \boldsymbol{z}_i = k) \\ &= \prod_{q=1}^{Q} \left\{ \prod_{z_i=k} \frac{\Gamma(\alpha_{kq})}{\Gamma(\alpha_{kq} + c_{iq0} + c_{iq1})} \prod_{r=1}^{2} \left\{ \prod_{m=1}^{c_{iq}} (\alpha_{kqr})^{\nu_{kqr,m}} \cdot (m-1)^{1-\nu_{kqr,m}} \right\} \right\}.\end{aligned}$$

这种增广方法提供了两种互补的抽样器，分别更新贝塔-二项分布的两种参数化形式，从而加快向平稳分布的收敛。若定义 $\alpha_{kq} = \alpha_{kq0} + \alpha_{kq1}$ 和 $p_{kq} = \frac{\alpha_{kq0}}{\alpha_{kq0}+\alpha_{kq1}}$，第一种抽样器独立抽取 α_{kq} 和 p_{kq} 参数，而第二种则直接对贝塔-二项参数 α_{kqs}进行抽样。第一种抽样器利用变量增广可以将 α_{kq} 的后验密度写成与以下表达式相差一个常数因子的形式：

$$\begin{aligned}\mathbb{P}(\alpha_{kq}, p_{qk}|\boldsymbol{r}, \boldsymbol{c}) \propto \pi(\alpha_{kq}, p_{kq}) \cdot \alpha_{kq}^{\sum_{z_i=k} r_{iq0} + r_{iq1}} &\left\{ \prod_{z_i=k} \frac{\Gamma(\alpha_{kq})}{\Gamma(\alpha_{kq} + c_{iq0} + c_{iq1})} \right\} \\ &\cdot \left\{ p_{kq}^{\sum_{z_i=k} r_{iq0}} \cdot (1 - p_{kq})^{\sum_{z_i=k} r_{iq1}} \right\}, \end{aligned} \tag{4}$$

其中 $\pi(\alpha'_{kq}, p_{kq})$ 是这些参数的先验密度。在此模型中，我们假设每个 α_{kqr} 都独立地从 GAMMA(κ, θ) 中抽取，从而确保 α_{kq} 从 GAMMA$(2\kappa, \theta)$ 中抽取，且 p_{kq} 从 BETA(κ, κ) 中抽取。随后，我们利用这一观察构造两步马尔可夫链蒙特卡洛更新，分别针对 α_{kq} 和 p_{kq}。对于 α_{kq}，我们采用随机游走 Metropolis-Hastings 抽样器，根据 GAMMA$(1, 1/20)$ 提议一个新的 α'_{kq}，再依据以方程 4 为目标密度的比率决定是否接受。由于 p_{kq} 的先验是均匀分布，我们采用标准的贝塔/二项共轭更新。随后直接按照方程 3 更新各个 r_{iqr}。

假设 α_{kqr} 服从 GAMMA(κ, θ)，第二个采样器通过注意到以下事实来运行：

$$\mathbb{P}(\alpha_{kqs}|\alpha_{kq(-s)}, \boldsymbol{\nu}, \boldsymbol{z}, \boldsymbol{c}) \propto \mathrm{GAMMA}(\alpha_{kqs}|\kappa + \sum_{z_i=k} r_{iqs}, \theta) \cdot \frac{N_k \cdot \Gamma(\alpha_{kr})}{\prod_{z_i=k} \Gamma(\alpha_{kq} + c_{iq})}$$

其中，s 为 0 或 1，而 $(-s)$ 为另一类别。我们再次采用 Metropolis-Hastings 更新：根据 GAMMA$(1, 20)$ 提出一个新的 α_{kqs}，然后以条件分布作为目标密度来计算接受率。随后，我们将该更新应用于 $s = 0, 1$ 以及后续的 k 和 q。

2.3.3 出生-死亡马尔可夫链蒙特卡洛

Stephens 提出的用于推断模型组分数量的 BD-MCMC 方法，是针对有限混合模型中未知组分数量进行完整后验推断的可逆跳跃方法的重要替代方案 [Stephens, 2000, 1997]（Green, 1995, Capp´e 等, 2003]）。这里我们沿用 Shi、Murray-Smith 和 Titterington 的研究，他们将 Stephens 的原始方法扩展至具有上述潜变量表述的混合模型 [Shi 等, 2002]。

沿用 Stephens 的方法，我们定义一种生灭跳跃过程，使该过程的平稳分布就是混合模型的后验分布。通常，这个连续时间过程会等待一段服从指数分布的时间，然后执行出生步骤（向模型中添加一个新的混合组分）或死亡步骤（移除一个已有组分）。出生和死亡步骤中的等待时间及概率（如下所述）经过设定，以确保得到所需的平稳分布。为简化符号说明，当模型包含 K 个组分时，我们将所有参数合并表示为 $\Theta_K = \{\boldsymbol{z}, \boldsymbol{\eta}, \boldsymbol{\alpha}, K\}$。该状态不包含变量扩增参数 $\boldsymbol{\nu}$，因为这些参数仅用于辅助抽样，不影响边缘分布。我们将移除一个组分后到达的死亡状态记为 $\Theta_{K|k}$。从出生率 β 和初始状态 Θ_K 开始，我们

按照以下步骤从生灭过程进行抽样：

1. 计算每个 ξ_k $k=1$、$\cdots$、K组分的死亡率。2. 从均值为 $\frac{1}{\beta+\sum_{k=1}^{K}\xi_k}$的指数分布中进行模拟。3. 根据伯努利分布$\left(\frac{\beta}{\beta+\sum_{k=1}^{K}\xi_k}\right)$模拟是否发生出生或死亡。(a) 如果发生出生：根据上述先验分布模拟一个新的组分 α_k ，并从 $[0,1]$中均匀抽取η'；对于每个市镇 i, 抽取一个伯努利分布($\frac{1}{K+1}$)，如果成功则重新分配 $z_i=K+1$；设定 $\eta_{K+1}=(\eta_1\cdot(1-\eta')\ \cdots、\ \eta_K\cdot(1-\eta')\ \eta')$。(b) 如果发生死亡：选择要删除的 k^{th} 组分，并通过移除α_k来构造 $\Theta_{K|k}$ ，根据在1、$\cdots$、$K-1$上的均匀类别分布重新分配 $z_i=k$，并重新缩放 $\eta^*_{K|k}=\left(\frac{\eta_1}{1-\eta_k}、\ \cdots、\frac{\eta_{k-1}}{1-\eta_k}、\ \frac{\eta_{k+1}}{1-\eta_K}、\ \cdots、\ \frac{\eta_K}{1-\eta_k}\right)$。

Shi 等人 [2002] 表明，对于上述定义的潜在混合模型，其死亡率为

$$\gamma_k=\frac{\text{POISSON}(K-1|\lambda)}{K\cdot\text{POISSON}(K-1|\lambda)}\cdot\frac{\mathbb{P}(\boldsymbol{c}|\Theta_{K|k})}{\mathbb{P}(\boldsymbol{c}|\Theta_K)}\cdot\frac{\prod_{i=1}^{K|k}\eta_k^{*d_k}}{\prod_{i=1}^{K}\eta_k^{d_k}}$$

其中 $d_k=\sum_{i=1}^{N}\mathbf{1}_{z_i=k}$。所得出生过程的平稳分布为 $\mathbb{P}(K,\ \boldsymbol{\eta},\ \boldsymbol{\alpha},\ \boldsymbol{z}|\boldsymbol{c})$（其手稿中的定理 2）。对于实现的 BD-MCMC 算法，我们将上述生灭马尔可夫链蒙特卡洛的混合版本与离散时间马尔可夫链蒙特卡洛相结合，从而显著提高了 K 中的混合性。第 l 次迭代时的状态记为 $(K,\ \eta_K,\ \theta_K)_l$。下一次迭代中，$l+1$按以下方式抽样：

1. 对上述连续时间生灭过程运行时间 s，从 $(K,\ \eta,\ z,\ \alpha)_t$ 转移至 $(K,\ \eta,\ z,\ \theta)_{t+s}$；2. 设置离散时间 $(K,\ \eta_K,\ \theta_K)_l=(K,\ \eta_K,\ \theta_K)_{t+s}$；以及 3. 根据上述马尔可夫链蒙特卡洛方法，在固定 K 的条件下，对 $\boldsymbol{z}$、$\boldsymbol{\eta}$、$\boldsymbol{\alpha}|K_{l+1}$ 进行抽样。

3 结果

我们以两种互补的方式将模型应用于该数据集：首先，将所有问题汇总到一个数据集中（$Q=54$）；其次，分别使用九个四年选举周期（例如 2008—2011年）的数据。对于这些子集，Q 的范围为 16 至 21。

3.1 完整分析表明投票集团的地理结构 e

图1总结了模型应用于完整数据集后的结果，展示了 K的后验分布、后验样本中的混淆矩阵，以及投影

到州地图上的混淆矩阵。 K 的后验分布（图1(a)）是对样本的 K 值进行加权平均得到的，权重为这些值在 BD-MCMC 中对应的跳跃等待时间。观察到大多数后验样本中的暂态投票集团簇数量较少后，我们筛选出至少包含一定数量市镇的投票集团，从而缩小后验值的范围。分布的众数位于 $K = 6$，在 $K = 7$处也有显著概率质量。我们观察到，不同运行中 K的分布变化很小，不过有两次迭代的后验众数为 7，而不是 6。在介绍其余结果时，我们将使用 $K = 6$。

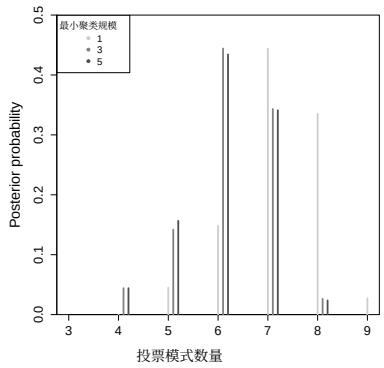
在图1(b)中，我们使用每对市镇的成对共同占据频率矩阵来概括变维聚类结构，并根据 BD-MCMC 中的等待时间对其加权。随后，我们将共同占据矩阵应用于原始矩阵的 k-mediod 聚类（使用 $K = 6$），并据此对行和列进行排序。与其他方法相比，这一程序有两项显著优势：由于所有计算仅基于市镇成员身份的成对关系，因此不受标签切换问题影响；同时，它能有效呈现整体聚类结构，并选取单个 K 作为代表。我们将此称为代表性聚类。

为考察投票集团分布中可能存在的空间结构，我们将共同占据矩阵的结果投影到缅因州地图上，得到图1(c)。对于每个市镇，我们取其在共同占据矩阵中的中位数占据值，再将其归一化，以确定地图上的混合比例。我们发现，与模型直接估算集团成员身份相比，这一统计量稳健得多；后者同样受到标签切换问题和 K 变异性的影响。我们按由南至北的顺序为投票集团着色，并以该投票集团占据比例过半的最南端市镇来确定其位置。

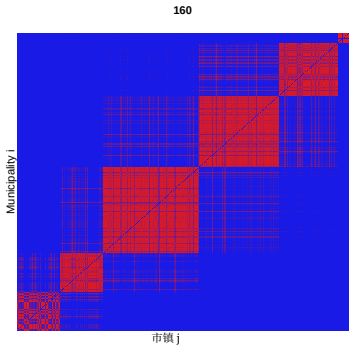
在具有较强后验支持度的 K 值中，地图上有若干特征始终保持不变，这些特征将与下文的讨论相关。最值得注意的是，投票集团在该州呈现出空间结构：投票集团1沿着南部海岸分布，同时覆盖最大的内陆城市；投票集团2沿着中部海岸分布，同时覆盖较小的内陆城市和南部内陆地区；投票集团3覆盖西部州域更靠内陆的部分、东部海岸以及遥远北部的部分地区；投票集团4和5覆盖其余主要为农村的内陆社区。投票集团3、4和5覆盖的市镇大多呈现至少两个集团的某种混合状态，从西南部以集团3为主，逐渐过渡到北部几乎完全由集团5构成。加拿大边境上主要讲法语的社区呈现出它们自己的投票集团（6）。这种强烈的空间结构在不同的马尔可夫链蒙特卡洛运行及 K值中都持续存在，尽管其

集团的数量和边界会有所变化。

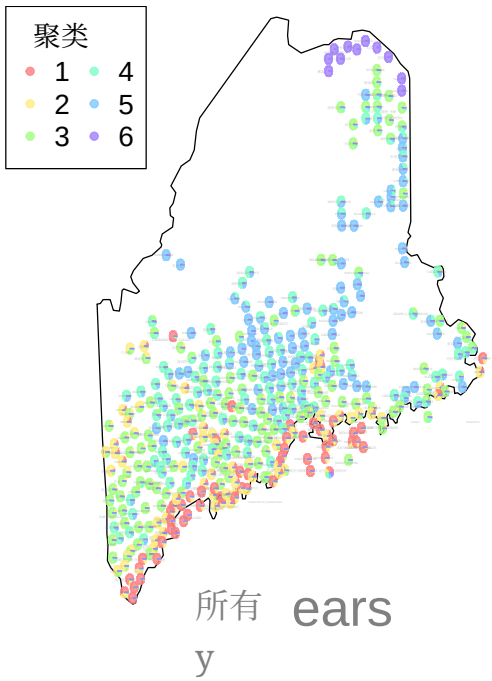
我们注意到，这些模式存在若干例外，而且无论选择 K 还是选举周期如何，这些例外都保持一致。伊斯特波特市（人口1,331）尽管地处极东部且规模较小，但其居民主要来自投票集团1；然而，伊斯特波特既是该地区最大的定居点，也是欧洲人最早建立的定居点之一。东部海岸其他规模较大的社区——MACHIAS、LUBEC和MACHIASPORT——主要呈现投票集团2的特征，并且也相对较早地在殖民时期建立。州西部弗赖堡县治周边的几个社区，尽管远离南部海岸和内陆城市，却显示出对投票集团2的强烈归属。我们注意到，滑雪度假社区Carabasset Valley尽管极其偏远，却主要来自投票集团1；而邻近市镇几乎完全来自投票集团3至5。最后，我们观察到，岛屿自治市Isle au Haut无法很好地由任何投票集团代表。



（a）完整数据集的 K 后验分布。由于小簇通常是短暂的，因此还呈现了过滤这些簇后的结果。



(b) 完整数据中各投票集团的共同占据矩阵，使用 $K = 6$ 的代表性聚类。色标在 0 处为蓝色，在 0.96 处为红色。



(c) 将投票集团共同占据矩阵投影到缅因州地图上。投票集团按从南到北的顺序编号。根据代表性聚类得出的聚类（投票集团）比例。

图1

3.1.1 九个四年选举周期中投票集团的一致性

为了探究数据中的时间模式，我们将模型应用于按九个四年选举周期划分的数据。每个周期中的选票问题数量从16个到21个不等，中位数为18个。如下文的模拟研究所述，这一数量略低于实现一致推断所需的问题数量，因此结果预计会出现更多变异。图2展示了每个选举周期的代表性聚类。

没有一个周期的结果与完整数据集完全相同，但大多数周期都保留了其中发现的关键模式：一组沿海投票集团也涵盖内陆城市；一组投票集团覆盖除城市以外的内陆市镇；以及遥远北部一个独立的投票集团，覆盖法语社区。与完整数据的结果相比，主要差异在于覆盖沿海和内陆市镇的集团数量。例如，在2016年周期中，南部沿海市镇和内陆城市由三个投票集团覆盖，而完整数据中则由两个投票集团覆盖。如下文的模拟研究所示，我们预计，这种差异部分源于选举周期数据集中问题数量较少。

选举周期数据与完整数据的模型结果存在两处显著差异，这些差异似乎是纳入或排除特定问题所致。第一处差异是，在2008年和2009年周期中出现了一个独立的集团，覆盖中部、东部和北部讲英语的市镇（2008年周期中的集团5，2009年周期中的集团6）。与完整数据模型相比，这使内陆集团的范围更加集中。值得注意的是，这两个周期包含大量赌场相关议题（2008年16个问题中有4个，2009年18个问题中有3个），其特殊属性将在下文探讨。第二处差异是，在2013年和2014年选举周期中，没有单独的投票集团将法语社区聚为一组；这些社区转而被归入沿海投票集团。这一差异或许可以用这些周期中非预算议题相对较少来解释：这些周期中唯一的非债券议题是大麻合法化、加强枪支销售法规以及建立排序选择投票制度。如第3.1.2节所述，法语社区在文化议题上的投票模式与农村投票集团相似，而在财务议题上的投票模式则与沿海投票集团相似。

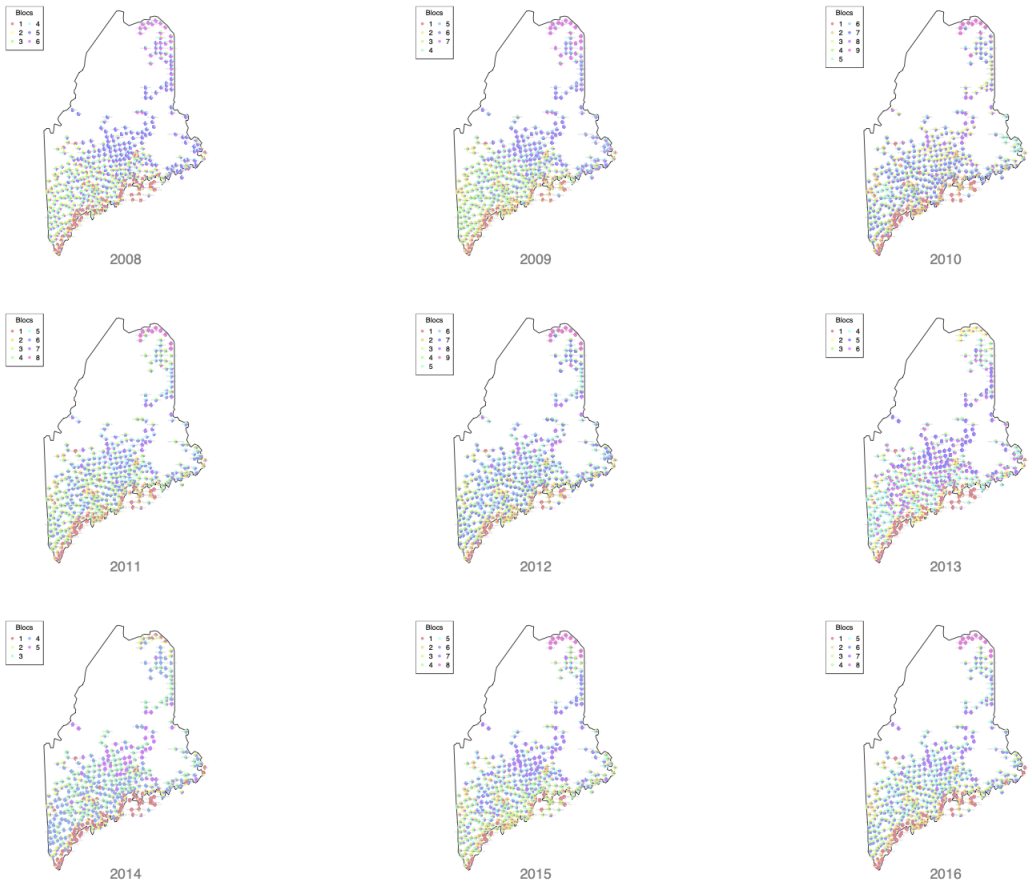


图2：展示从2008—2011年到2016—2019年各四年选举周期的代表性聚类结果。地图以周期的起始年份标注下标。集团数量由 K 的后验众数确定，计算前已过滤掉规模小于5个市镇的簇。代表性聚类的确定方法与图1(c)相同。

3.2 问题支持度的分布表明，大多数投票集团沿梯度分布，法语社区则相对孤立

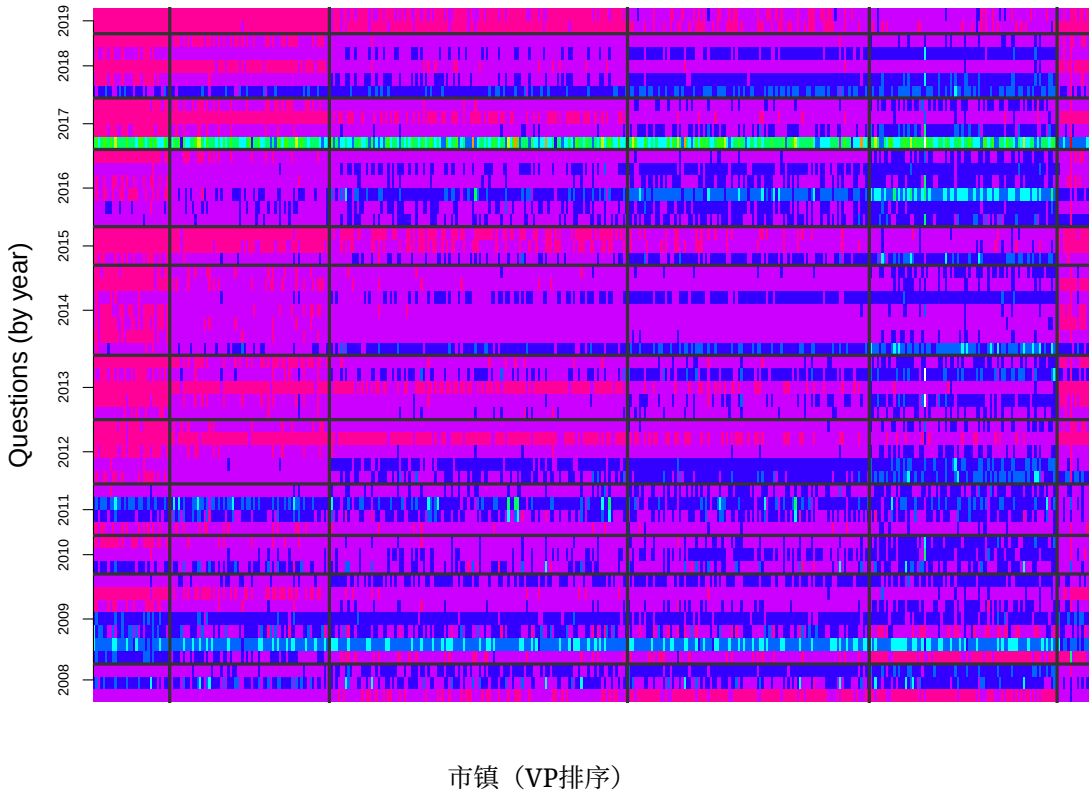


图3：条形码图概述了各市镇对每个问题的支持度。问题按年份纵向排列，市镇则按照与图1相同的投票集团代表性聚类结果横向排列。绿色表示最低支持度（0.15），粉色表示最高支持度（0.9）。灰线表示依据代表性聚类划分的投票集团边界。

图3展示了各投票集团内部对每个公投问题的支持度结构，集团边界由图1所采用的代表性聚类确定。对于几乎所有问题，从投票集团1到投票集团5，问题支持度都呈现梯度变化，且支持度往往在集团边界处逐步改变。这使得各集团中的问题支持度呈现出典型的“条形码”模式。五个问题（2、5、7、16、44）在各投票集团中的支持度一致。投票集团6对应着该州北部边界沿线以法语社区为主的地区，是这一模式中明显的例外。该集团表现出

对各问题的支持度呈现聚类，其模式与其他五个集团不一致：一些问题与投票集团5相符（主要是社会／文化问题，如同性婚姻、狩猎限制和大麻合法化），另一些则与投票集团1和2相似（主要涉及资金，例如债券、学校和医疗保健增税）。

为了解这些结果如何与数据本身更直接的呈现方式相吻合，我们使用一种 t分布式随机邻域嵌入（tSNE）方法。这是机器学习、神经科学和基因组学中常用的技术，参见 Van der Maaten 和 Hinton [2008],Li 等人 [2017]。该方法计算数据条目之间的成对距离（此处为每个市镇按问题细分的投票总数），然后采用随机梯度下降方法，在约束其置于少数几个维度（此处为二维）的条件下，最小化这一高维矩阵的应变。由于数据具有组合性，我们使用中心对数比变换来构造问题之间的距离，然后将各问题的结果相加 [Aitchison， 1982]。最终，每个市镇都被呈现在一个二维平面中，该平面反映了其在所有问题上的支持度相似性。随后，我们根据图1所确定的混合比例，为每个点着色。

图4为图3中的条形码图提供了互补的视角。前五个投票集团沿着一个（大致）一维的梯度依次排列，其中投票集团1和5位于两端，集团2、3和4则填补中间的空间。这种呈现方式表明，正是数据中的相似性促成了集团1和5中相对未混合的社区。值得注意的是，t-SNE将集团5中未混合的市镇与集团4共享的混合社区区分开来，尽管这些社区在地理位置上往往相邻。同样，被识别为在集团3和4之间混合的市镇，在t-SNE图中形成了一条边界。最值得注意的是，该州北部的法语社区在t-SNE表示中形成了一个独立、分隔开的区域，混合模型也显示了这一点。

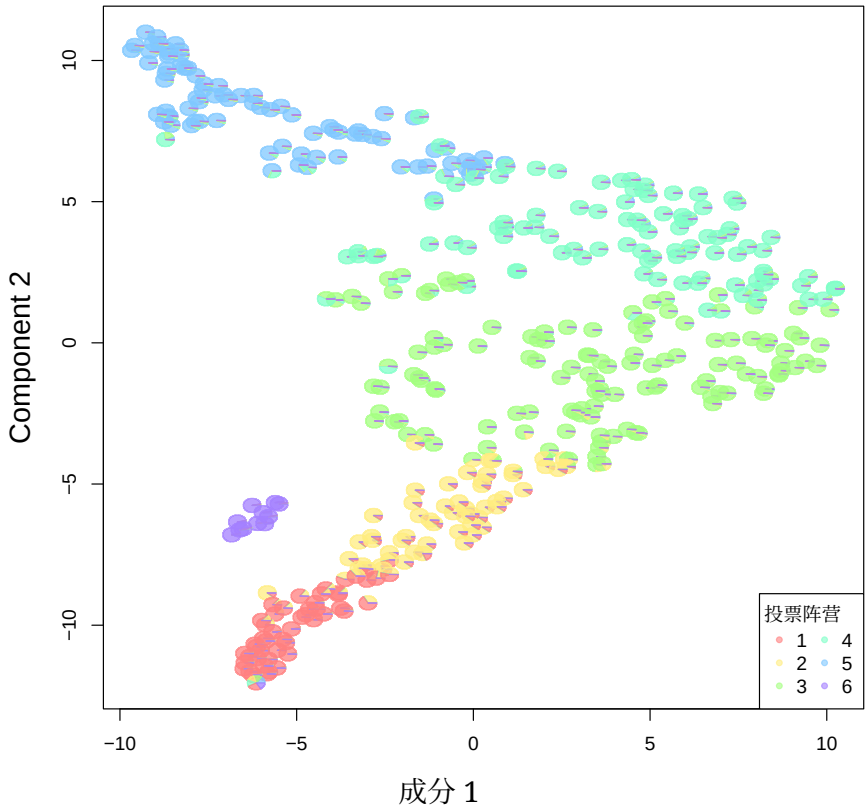


图4：二维 t-SNE 嵌入。市镇根据图1中使用的混合比例着色。颜色与图1c相同。

3.3 模型拟合度量揭示少数“局部”问题

为考察针对单个问题的模型拟合质量与完整数据集的契合程度，我们比较每个市镇中各问题的后验预测分布与数据中支持度的观测值。对于每个问题 q 和每个市镇 i，我们使用$\alpha_{z_{iq}}$ 的后验样本构建支持度的后验预测分布，并将所得分布的中位数 – $\tilde{p}_{iq}$ – 与对命题 $\hat{p}_{iq}$的观测支持度进行比较。对于稀疏化马尔可夫链中的每次迭代，我们根据贝塔分布抽取100个值，参数为（α_{z_iq0}，α_{z_iq1}）。我们将该分布的中位数与观测比例（$\hat{p}_{iq} = \frac{c_{iq0}}{c_{iq0}+c_{iq1}}$）进行比较。尽管这一统计量忽略了由样本量导致的方差，而贝塔-二项分布可以对其进行估计，但它有助于

便于比较不同规模的市镇。我们将 $\widetilde{p}_{iq} - \widehat{p}_{iq}$ 称为
估计问题拟合度.

图5展示了在所有市镇中计算得到的估计问题拟合度的中位数和标准差。几乎所有问题在这两项统计量上都呈现相似数值，表明模型拟合在范围广泛的市镇中保持一致。值得注意的是，有五个问题的标准差明显较高（图5中以红色标示）。其中四个问题涉及在该州创建或监管特定赌场。最后一个问题涉及废除一项学校合并提案，该提案尤其影响缅因州的农村学校。这些“局部”问题很可能违反模型的非局部性假设，因为选民可能受到特定地方关切的驱动，而这些关切压过了更广泛的文化立场。我们还在补充图1—3中进一步考察了各问题在所有市镇中的表现，其中按投票集团对各市镇的估计问题拟合度进行了分组绘图。在同一问题内，我们观察到各市镇之间的变化具有很强的一致性：“局部”问题在所有市镇中都表现出较高的变化程度，同时这种变化在各投票集团之间保持一致。

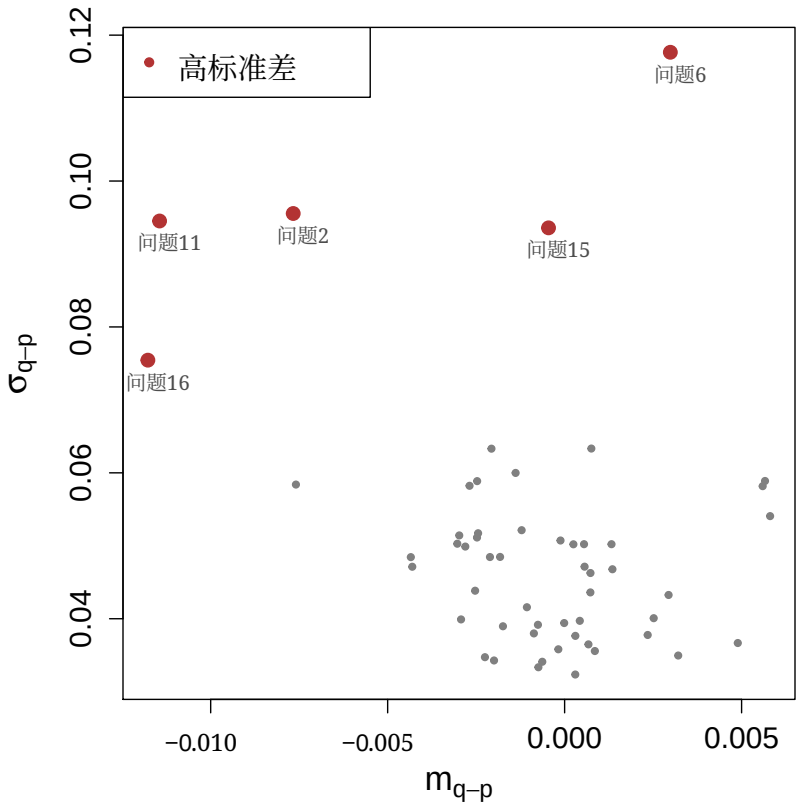


图5：按问题列出的所有市镇的估计问题拟合度（$q-p$）的中位数和标准差。五个问题表现出较高的标准差，以红色标记。每个市镇按问题列出的估计问题拟合度图见补充图1—3。

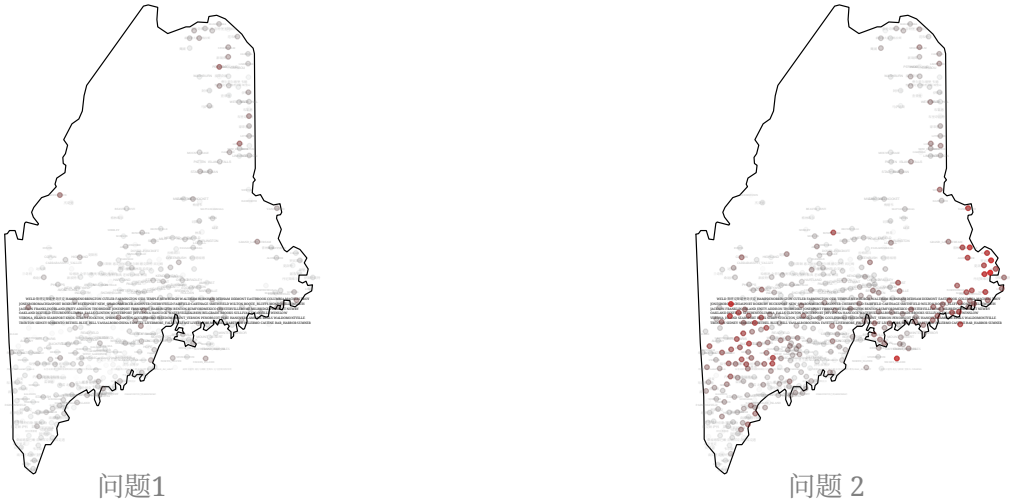


图6：典型问题（左）和“局部”问题（右）的模型拟合。拟合质量以灰色至红色的色阶表示，白色表示完美拟合（0），红色表示观测到的最差拟合（0.06）。

为探究这些问题的局部性，我们将每个市镇对每个问题的估计问题拟合度绘制到缅因州地图上（补充图4）。对于49个问题，除了农村市镇的估计问题拟合度略高（这可能是由于样本量较小），其余几乎没有明显的地理结构。正如图6所示，五个“局部”问题呈现出显著的地理结构。问题2涉及在该州西部的牛津县建造一座赌场，当地支持度相较于模型预期有所提升。在该州最东部，支持度则低于模型预期。这种地理结构证明了该问题具有“局部”性质：模型假设投票集团在该州所有市镇之间是共享的。因此，对于与特定地区相关的问题，模型无法捕捉这种差异，导致拟合度较差，并呈现出具有地理结构的拟合模式。

为进一步理解“局部”问题与投票集团结构之间的相互关系，我们按照代表性聚类结果，绘制了各市镇对每个问题的观测支持度之间的詹森-香农距离矩阵 [Bri¨et 和 Harremo¨es, 2009]（例如，补充图4）。这衡量了两个概率分布之间的距离。我们观察到，几乎所有问题都呈现出明显的分组结构：支持度相近的市镇会按照代表性聚类结果聚集在一起。相比之下，五个高方差问题没有呈现这种分组结构，而是

显示同一集团内各市镇之间的距离持续存在差异。少数低方差问题也呈现出较弱的分组结构：问题7、10、23和38。这些问题与政府运作有关。就支持度而言，这些问题在不同投票集团之间的区分度较低，也就是说，各市镇的支持度更为相近，因此表面上呈现的分组结构较弱。

3.3.1 模型聚类显示极化的时间变化

过去二十年来，美国选举研究的一个焦点是极化的选民群体的出现，其表现包括在广泛议题上的立场日益分化、对其他政治派别的信任下降、媒体隔离，以及依据政治派别迁徙并形成群体 [Liu 等人，2019年；Fiorina 等人，2008年；Iyengar 和 Westwood，2015年；Prior， 2013]。混合模型产生的聚类进一步证明了这一现象，尽管底层投票集团的结构在整个研究期间似乎保持稳定。在图7中，我们绘制了三对投票集团在各问题上的平均支持度差异。这些集团是根据 t-SNE 分析选出的，用于衡量两端集团（集团1和集团5）与一个中间投票集团（集团3）之间的支持度差异。

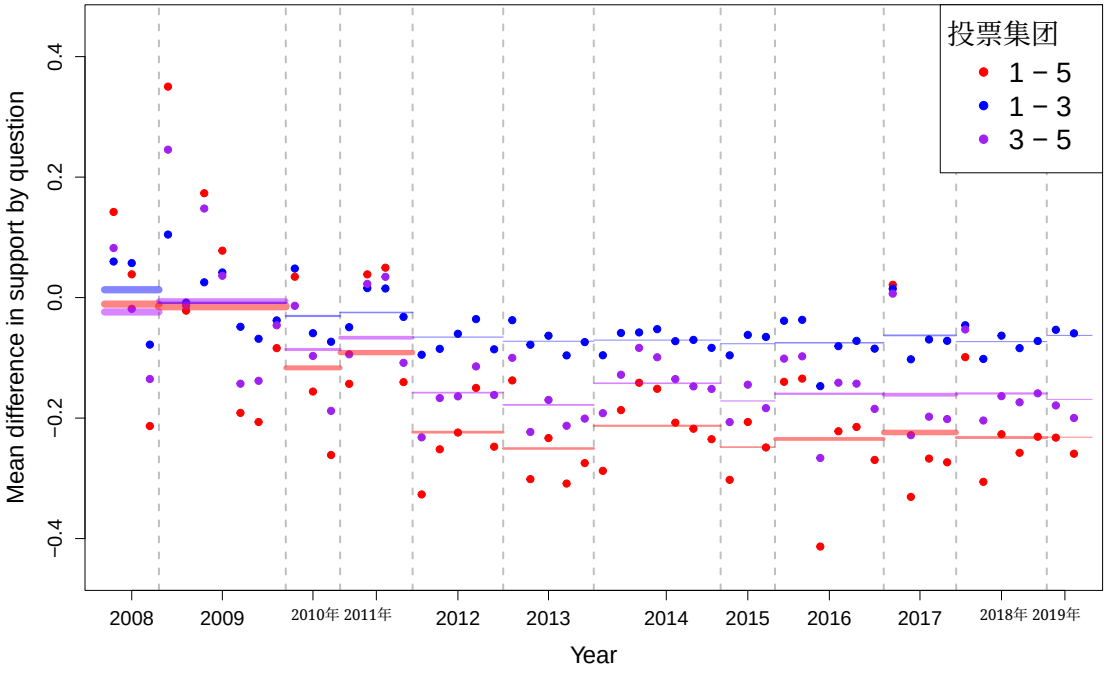


图7: $K = 6$中三对投票集团在各个问题上的支持度差异。彩色线条表示同一年内各问题支持度差异的平均值；线条宽度与该年支持度差异的标准差成正比。年份间距根据各年包含的问题数量确定。

结合图3和图4以及投票集团的选择，可以预期各组集团之间的支持度差异会呈现梯度，而这一现象几乎出现在所有年份的所有问题中。还有两个较不符合预期的观察结果：随着时间推移，支持度差异微不足道的问题消失了；每年纳入考察的问题的方向性也发生了变化。在2008—2011年这段时期，17个问题中有5个问题的支持度差异极小；而在2012—2019年，只有2个问题呈现这一结果。在2008年和2009年，支持度差异正负各半，因此全年平均支持度接近于零（彩色直线）。随后两年，方向性支持度不断增强，并在2012—2019年趋于稳定，这表明各投票集团在投票支持度的相对强弱上出现了分化。方向性的单边性或许可以由选票上出现的问题内容来解释，而这些内容本身可能受到日益加剧的极化影响。

模拟研究

参数	N	Q	C	δ
值	10, 50, 200, 500	1, 5, 10, 20	100, 500, 1000, 3000	0.01, 0.1, 0.3, 1

表 3：模拟研究中使用的值。 δ 是对称狄利克雷分布的浓度参数。

需要多少个问题才能解析数据中的潜在结构，这一数值无法事先确定。通过在理想化条件下研究这一问题，我们仍然可以评估模型的表现，从而了解数据特征（例如选民数量、市镇数量和混合程度）对簇的无偏推断所作的相对贡献。我们将其中若干特征理想化，以开展模拟研究，从而对这一问题获得更清晰的认识。

我们使用以下过程模拟数据。对于每次模拟，我们首先设定投票集团数量（K）、市镇数量 N、问题数量（Q）、所有城镇中的选民数量（C），以及一个参数 $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_K)$；该参数根据 DIRICHLET 控制市镇内部的混合程度。我们根据 DIRICHLET(δ)，为每个城镇独立抽取其混合分布。对于每个问题，我们独立抽取一对来自 GAMMA(1，20) 的 α 值。然后，我们模拟每个问题和每个城镇中的$\lfloor C \cdot \delta_k \rceil$ 名选民，并将各投票集团中的赞成票和反对票相加，从而得到每个城镇的总票数。汇总各 N个市镇后，即得到最终数据集。每次模拟所使用的值列于表 3 中。每组值均进行了十次迭代。

我们发现，混合程度会显著影响算法推断出正确投票集团数量的能力。如果混合程度很高（几乎所有市镇都由各投票集团均匀混合构成），算法推断正确投票集团数量的能力就会受到限制。然而，如果混合程度较不均匀（例如在 $\delta = 0.3$或更少时），算法在问题数量和市镇数量不同的情况下，表现似乎都较为稳定。选民数量对结果的影响似乎较弱。在市镇数量适中且至少存在一些未混合的社区（$\delta < 0.3$）时，$Q = 20$ 可确保后验众数接近模拟值。

4 讨论

4.1 模型及其可能的扩展

本研究的主要目的是将投票集团模型的统计应用范围扩展至按市级汇总的纵向公投票数总计。由于个人公投选票所含信息有限，该模型需要多张选票才能得出稳定的推断，因此其适用范围仅限于定期举行公投的地区。然而，许多政体频繁举行公投，为该模型提供了丰富的应用场景。美国许多州经常举行此类选举，其情形与本文考察的案例研究相似。瑞士独自贡献了全球三分之一的公投，因此似乎尤其适合作为研究地点。

与本文提出的模型相似的模型广泛应用于遗传学（潜在狄利克雷分配模型的变体 [Blei等，2003，Falush等， 2003]）、微生物组分析（狄利克雷-多项式混合模型和系统发育混合模型 [Holmes等，2012，Mao和Ma，2022]）以及文本分析 [DiMaggio等，2013，Bohr和Dunlap， 2018]。在所有这些领域中，混合模型首先用于捕捉数据中存在的经验结构。接下来，更广泛的分析工作是根据外部变量解释观察到的聚类结构（例如，基因组数据中的迁移、选择和漂变，以及微生物组数据中的生态和环境变化）。本文提出的模型以及Gormley和Murphy（及其他学者）的方法，为政治分析构建了类似的框架：通过识别结构相似的投票行为模式，理论分析便能据此进行校准，以解释纵向投票中的总体趋势。这种类比不仅停留在概念层面：关于如何建模复杂潜在过程与观察到的混合结构之间联系的丰富文献，为后续研究提供了许多途径 [Lawson等，2018，Hellenthal等， 2014]。

本文的另一个目的，是展示如何利用该模型揭示数据集中的空间和时间模式。在案例研究中，选举周期分析表明，在整个研究期间，投票阵营结构总体上保持一致。同样，无论是在完整数据中还是在选举周期数据中，各市镇之间混合比例的相关性都显示出显著的空间结构。这种空间结构与tSNE数据投影高度吻合，进一步印证了该模型的推断能力。然而，这两类变异都未在该方法中直接建模。发展能够明确纳入这些协方差结构的技术，为未来的模型开发提供了许多自然的研究方向。沿用Gormley和Murphy [2008a], a专家混合方法似乎是纳入协变量信息的一种直接扩展。在

时间变异的背景下，由于同一组选民必然会在每张选票上进行投票，按选票对数据进行分块的混合模型将能够更灵活地捕捉数据中的时间结构。这样的模型可以根据某一投票阵营前往投票站的概率所偏向的潜在混合结构，按比例对隐藏状态（投票阵营）进行积分。这样便可以直接对某个市镇中表面上的投票阵营比例随时间发生的变化进行建模。

进一步开展统计研究的另一个重要方向，是对选票问题之间的相关性进行建模。不出所料，交通债券问题会引发相似的投票模式。涉及各类债券的选票通常比同性婚姻或大麻合法化等文化问题更加相似。虽然理想情况下，该模型应考虑完整的 Q^2 协方差矩阵，但狄利克雷树混合模型等方法 [Mao和Ma， 2022]可能更具可行性。然而，与当前使用狄利克雷树混合模型、假设固定系统发育树的方法不同，这种扩展还需要对潜在树进行估计。在这方面，该问题与微生物群落中的统计估计问题表现出显著相似性 [O'Brien等人， 2014]。另一个独立但相关的考量是，问题本身在多大程度上遵循投票集团假设。虽然绝大多数问题（49个）似乎与模型的预期一致，但对于结果部分确定的五个问题，我们观察到了空间聚集偏差（例如补充图4）。这些问题异常"局部化"的特征，可能与问题内容有关：其中五个问题中的四个旨在将赌场设置在这些地区。调整模型，以便分别推断并建模这些问题，将减少对投票阵营结构估计的方差。

除了上述用途外，该模型还将一种 BD-MCMC 方法引入政治学文献，用于获得投票集团数量的后验分布。与 AIC 或 BIC 等常用决策准则不同，这些准则会选出单一的"最佳"成分数量，而 BD-MCMC 能提供成分数量的后验分布，从而更连贯地呈现模型的不确定性。这与可逆跳跃 MCMC 或热力学积分方法类似，但计算负担更低。BD-MCMC 算法具有相对较高的灵活性，这也意味着该技术很可能可以重新适用于明确建模空间、时间和协变量结构的模型。

从更广泛的背景来看，公投数据也可以被视为多视图问题的一个异常规整的例子。这是机器学习、医学和基因组学中常见的一项统计挑战，发生在试图协调对单个对象进行的多种不同类型检测（或"视图"）时 Carmichael [2020], Chao et al. [2017],Yi et al. [2005]。在这一背景下

，视图由不同的选票问题构成，这些问题可能沿着不同维度区分选民（例如交通债券与大麻合法化）。补充图4中观察到的可变阻塞结构至少部分反映了这一点。与许多视图具有不同分布结构的多视图情境不同，公投数据（以及更广泛而言的投票数据）在不同视图之间具有一致的底层分布。未来的模型开发可以利用这一比较，将为更异质情境开发的框架引入其中。从互补的角度来看，公投数据也可以作为多视图算法开发的有用训练集。

4.2 与政治学中其他方法的联系

对于政治学家而言，一个自然的问题是，如何可信地解释混合模型的结果：应在多大程度上将该模型的组分理解为对应于界定明确的政治文化？解释上的这一难题是混合模型的常见特征，文本分析人员、遗传学家和微生物生态学家也面临类似的一系列问题。这些潜在组分是否代表了此前未被发现、但具有重要意义且可验证其特性的亚群体？还是说，它们纯粹是用于帮助对数据分布进行经验刻画的概率成分？混合模型与 t-SNE 分析之间的紧密关联表明，可以从这些数据中提取有意义的潜在结构；但政治学仍需探究的问题是，这些结构的持久性如何对应其他分析维度。即便承认混合模型识别出了潜在空间中的某些稳定特征，更广泛的解释问题仍涉及三个方面：（1）潜在空间的地形结构是什么？（2）这种地形结构会发生多大变化（例如，随时间变化）？（3）社区与这种地形结构之间的关系有多稳定？在当前分析阶段，这些问题尚无定论，而且答案很可能因经验背景而异。在缅因州这样人口稳定的地区，地形结构本身以及社区的混合比例可能都相对固定。在加利福尼亚州这样内部和外部迁移规模都较大的地区，潜在空间的结构和社区的构成则可能迅速变化。

本文提出的方法与政治学中许多现有技术存在联系，包括生态回归、民意调查和潜在类别分析。生态回归解决了聚合后回归能够在多大程度上得以保留的问题，并且已被证明在特定条件下具有一致性 [King 等人， 2004]。本文提出的方法为考察汇总投票数据和回归方法（专家混合模型）提供了一种新方式。于是，一个尚待解决的问题是如何调和这两种方法，

具体而言，就是如何将生态回归的一致性条件转化到这一新语境中。

所介绍的方法也可能对民意调查方法产生影响。从统计学角度看，民意调查实践主要关注两点：(1) 获取足够大的随机样本；(2) 对结果进行偏差校正，以纳入样本中代表不足或代表过度的人口统计类别。混合模型识别出的潜在结构，可能为抽样的初始程序以及偏差校正程序提供一种新方法。为了获得具有代表性的样本，利用推断出的投票集团的空间配置（例如，未混合的市镇）来指导抽样程序，可能比随机抽样更有效，因为意见结构在很大程度上可能已经体现在投票集团之间的分离中。同样，对投票集团的推断也允许采用一种新型的偏差校正程序：如果记录了受访者所在的市镇，就可以通过重新加权样本中各投票集团的相对占比来校正民意调查中的偏差。

在大多数社会科学领域，有限混合模型通常作为潜在类别分析（LCA）中的一个步骤出现。潜在类别分析是政治学、经济学和社会学中广泛使用的工具，用于识别具有共同人口统计、调查、选票或经济特征的行动者隐含类别，从而理解这些特征如何影响特定的响应变量。根据数据的不同，这些分析可能呈现出明显不同的形式，但通常包括三个步骤：使用期望最大化算法或马尔可夫链蒙特卡洛算法推断混合模型的参数；根据这些结果将个体分配到潜在类别；以及使用这些变量进行逻辑回归或概率单位回归，以建立政治参与某些方面的模型。虽然也可以采用统一的框架，但这些分析通常按这种分步方式进行。与本文所介绍的方法相比，一个关键区别在于，这些分析通常使用不同类型的个体层面数据来识别隐含类别，而本文模型则依靠对单一数据类型的重复观测来识别集团。分析潜在类别分析与专家混合模型之间的比较，可以成为一项有价值的研究，以了解这些方法在具体情境中的相对优势。

4.3 缅因州公投数据的未来分析

缅因州公投数据的分析得出了两个值得进一步考察的显著结论。第一，投票集团的分布与该州的定居模式惊人地相似 [Fobes, 1944]。投票集团1对应最早建立的社区（通常在1650年之前定居），而最新建立的社区则属于投票集团5（在1850年至1910年间定居）。尽管这种相似性十分显著，但它可能只是巧合

，源于其他经济或社会过程；这一点需要借助协变量信息加以检验。如果这一判断成立——也就是说，定居过程促成了当今的投票集团——那么这一观察对政治学可能具有重大意义：虽然投票集团的观点可能发生变化并存在差异，但其总体结构或许能在很长的时间尺度上保持一致。这引发了一些问题：定居与殖民是如何形成稳定的投票集团的？这些集团又如何在数据集中持续存在或逐渐衰退？例如，缅因州在过去一个世纪经历了数轮移民潮：如果定居模式决定了投票集团，新移民会如何影响这些集团？

另一个相关研究方向是构建一个能够进一步追溯历史的数据集。缅因州的选举记录从20世纪中叶起就较为持续地保存下来，而从1880年起则具备现代水平的精确度。自1910年起便有公投数据，但在20世纪期间，其使用的一致性有所不同。自20世纪70年代末以来，公投的使用频率不断提高，因此至少对于1970—2022年这一时期而言，这些数据可以用于考察投票集团在同一时期内的稳定性。尽管投票集团在此处研究的时间段内显然保持一致，但这一时期也缺乏大规模迁移。观察更长时间跨度内的一致性（或变化），将进一步说明在稳定或变化的人口统计背景下，时间集团的使用情况。

4.4 将模型扩展至其他选举背景

最后一个值得考虑的问题是，如何扩展类似模型，以便从其他类型的选举数据中进行推断，尤其是候选人选举。在单选区相对多数制背景下，候选人选举很可能可以直接借鉴公投模型，因为两名候选人之间的竞争类似于公投中的是/否投票。议会制中常用的其他形式的候选人选举，与排序选择投票数据更为相似（尽管并不完全相同），因此可能需要仔细的分布分析，才能构建出足够完善的模型。不过，其前景十分可观：这些选票中存在的强协方差信息，可以为潜在投票集团的结构提供信息。

将这类模型扩展到多种选举类型，开辟了政治分析的新前沿，使人们能够同时并互为补充地考察多种选举类型——公投、候选人选举和RCV。重要的是，这些分析可以涵盖不同的汇总尺度——市镇、选区、州——从而帮助理解投票集团及其在时间和空间上的稳定性。将多种空间尺度与选举类型跨时间综合起来，可以构建一个新的框架，用以结合政治观点的历史演变和当前状况进行分析。

即使不进行这些扩展，该模型也可应用于种类繁多的公投数据。除缅因州外，美国还有许多州定期举行公投。例如，在研究期间，加利福尼亚州的选票问题数量约为缅因州的三倍（不过，其市政边界不像缅因州那样稳定，人口也不如缅因州稳定）。自1980年以来，瑞士在社区层级汇总的公投数量相近，且有数字记录；更早至 19th 世纪的数据则以模拟形式保存。这些数据日益便于获取，标志着政治学家接触"大数据"的新篇章。如此丰富的数据使研究者能够把握政治观点的结构、维系与演变中的更广泛模式，而该模型正是迈出的第一步。

参考文献

John Aitchison。《成分数据的统计分析》。*Journal of the Royal StatisticalSociety: Series B (Methodological)*，44(2):139–160，1982年。

Alessia Amelio 和 Clara Pizzuti。《分析意大利议会的投票行为：群体凝聚力与演变》。载于*2012*年 *IEEE/ACM*社会网络分析与挖掘国际会议，第140–146页。IEEE，2012年。

缅因州档案馆。历史选举记录，2022年。网址：https://www.maine.gov/sos/arc/research/index.html。

Michael A Bailey、Anton Strezhnev 和 Erik Voeten。《基于联合国投票数据估计动态州偏好》。*Journal of Conflict Resolution*，61(2):430–456，2017年。

Zsuzsa Bakk、Daniel L Oberski 和 Jeroen K Vermunt。《将潜在类别分配与外部变量关联：正确推断的标准误》。《政治分析》，22(4)，2014年。

Nathaniel Beck 和 Simon Jackman。超越默认线性：广义加性模型。《美国政治学杂志》，第596–627页，1998年。

Nathaniel L Beck。政治方法论：一门热情欢迎新人的学科。*Journal of the AmericanStatistical Association*，95(450):651–654，2000年。

David M Blei、Andrew Y Ng 和 Michael I Jordan。潜在狄利克雷分配。《机器学习研究杂志》，3(1月):993–1022，2003年。

Jeremiah Bohr 和 Riley E Dunlap。《环境社会学的关键议题，1990年–2014年：计算文本分析的结果》。环境社会学，4(2):181–195，2018年。

Shaun Bowler、Todd Donovan 和 Caroline J Tolbert。《公民作为立法者：美国的直接民主》。俄亥俄州立大学出版社，1998年。Jop Bri¨et 和 Peter Harremo¨es。经典与量子詹森-香农散度的性质。*Physical Review A*，79(5):052311，2009年。

Olivier Capp´e、Christian P Robert 和 Tobias Ryd´en。可逆跳跃、生灭过程及更一般的连续时间马尔可夫链蒙特卡洛抽样器。*Journal of the RoyalStatistical Society: Series B (Statistical Methodology)*，65(3):679–700，2003年。

Iain Carmichael。在多视图混合模型中学习稀疏性和块对角结构。*arXiv* 预印本 *arXiv:2012.15313*，2020年。

Guoqing Chao、Shiliang Sun 和 Jinbo Bi。多视图聚类综述。*arXiv* 预印本 *arXiv:1712.06246*，2017年。

Joshua Clinton、Simon Jackman 和 Douglas Rivers。点名表决数据的统计分析。*American Political Science Review*，98(2):355–370，2004年。

Philip E Converse。政治过程中公众舆论观念的变化。*The PublicOpinion Quarterly*，51:S12–S24，1987年。

Jan De Leeuw。参议院投票模式的主成分分析。*Real Data Analysis*，第 405–411 页，2006年。

Paul DiMaggio、Manish Nag 和 David Blei。利用主题建模与文化社会学视角之间的关联：以美国政府艺术资助的报纸报道为例。*Poetics*，41(6):570–606，2013年。

Jeffrey A Dubin 和 Elisabeth R Gerber。投票提案的投票模式：选民类型的混合模型。技术报告，1992年。

Silvia D'Angelo、Thomas Brendan Murphy 和 Marco Alf`o。多维网络的潜在空间建模：在欧洲歌唱大赛投票交换中的应用。*The Annals of Applied Statistics*，13(2):900–930，2019年。

Daniel Falush、Matthew Stephens 和 Jonathan K Pritchard。使用多位点基因型数据推断群体结构：连锁位点与相关等位基因频率。遗传学，164（4）：1567–1587，2003年。

Morris P Fiorina、Samuel J Abrams 等。美国公众的政治极化。*AnnualReview of Political Science*，11：563，2008年。Charles B Fobes。缅因州的定居路径与人口分布。*EconomicGeography*，20（1）：65–69，1944年。Sylvia Fr¨uhwirth-Schnatter。有限混合和马尔可夫切换模型，第425卷。Springer，2006年。Sylvia Fruhwirth-Schnatter、Gilles Celeux 和 Christian P Robert。混合分析手册。CRC出版社，2019年。George Gallup 和 Saul Forbes Rae。民主脉搏：民意调查及其运作方式。1940年。Andrew Gelman、David K Park、Stephen Ansolabehere、Phillip N Price 和 Lorraine C Minnite。生态回归中的模型、假设与模型检验。*Journal of the RoyalStatistical Society: Series A (Statistics in Society)*，164（1）：101–118，2001年。

Isobel Claire Gormley 和 Sylvia Fr¨uhwirth-Schnatter。专家混合模型。载于混合分析手册，第 271–307 页。Chapman and Hall/CRC，2019年。

Isobel Claire Gormley 和 Thomas Brendan Murphy。《秩数据的潜在空间模型》。载于*ICML*统计网络分析研讨会，第90–102页。Springer，2006年。Isobel Claire Gormley 和 Thomas Brendan Murphy。《用于秩数据的专家混合模型及其在选举研究中的应用》。《style id='3'>The Annals of Applied Statistics，2(4):1452–1477，2008年。Isobel Claire Gormley 和 Thomas Brendan Murphy。《爱尔兰选民群体中的投票集团探索：一种混合建模方法》。《style id='5'>Journal of the American Statistical Association，103(483):1014–1027，2008年。Isobel Claire Gormley 和 Thomas Brendan Murphy。《利用社会人口统计学协变量对排序偏好数据进行聚类》。载于《选择建模：理论与实践的最新进展》。Emerald Group Publishing Limited，2010年。Peter J Green。《可逆跳跃马尔可夫链蒙特卡罗计算与贝叶斯模型判定》。《style id='9'>Biometrika，82(4):711–732，1995年。John AW Gunn。《现代政治学中的公众舆论》。《style id='11'>Political Science in History: Research Programs and Political Traditions，第99–122页，1995年。

Garrett Hellenthal、George BJ Busby、Gavin Band、James F Wilson、Cristian Capelli、Daniel Falush 和 Simon Myers。《人类混合历史遗传图谱》。《style id='1'>《科学》，343(6172):747–751，2014年。Simon Hix、Abdul Noury 和 Gerard Roland。《欧洲议会中的政治维度》。《style id='3'>《美国政治学杂志》，50(2):494–520，2006年。Steven Holloway。《联合国大会投票的四十年》。《style id='5'>《加拿大政治学杂志》/《加拿大政治学评论》，23(2):279–296，1990年。Ian Holmes、Keith Harris 和 Christopher Quince。《狄利克雷-多项式混合模型：微生物宏基因组学的生成模型》。《style id='7'>《公共科学图书馆·综合》，7(2):e30126，2012年。Shanto Iyengar 和 Sean J Westwood。《党派分歧间的恐惧与厌恶：群体极化的新证据》。《style id='9'>American Journal of Political Science，59(3):690–707，2015年。Jonathan N Katz 和 Gary King。《多党选举数据的统计模型》。《style id='11'>American Political Science Review，93(1):15–32，1999年。

Donald R Kinder. 传播与观点。《Annual Review of Political Science》，1(1): 167–197，1998年。

Gary King. 政治学事件计数的统计模型：常规程序中的偏差及指数泊松回归模型的证据。《美国政治学杂志》，第838–863页，1988年。

Gary King。《政治方法论》。《政治分析》，2:1–29，1990年。Gary King、Martin A Tanner 和 Ori Rosen。《style id='3'>生态推断：新的方法论策略。剑桥大学出版社，2004年。Shin Alexandre Koseki。瑞士政治分歧的地理演变：一种评估地方人口之间极化程度与动态的网络方法。《公共科学图书馆·综合》，13(11):e0208227，2018年。Stephanie T Lanza、Donna L Coffman 和 Shu Xu。潜在类别分析中的因果推断。《结构方程建模：多学科期刊》，20(3):361–383，2013年。Daniel J Lawson、Lucy Van Dorp 和 Daniel Falush。关于如何避免过度解读 STRUCTURE 和 ADMIXTURE 条形图的教程。《自然·通讯》，9(1):1–11，2018年。

Thomas J Leeper 和 Rune Slothuus。政党、动机性推理与公众舆论形成。政治心理学，35:129–156，2014年。Wentian Li、Jane E Cerise、Yaning Yang 和 Henry Han。t分布随机邻域嵌入在人类遗传数据中的应用。生物信息学与计算生物学杂志，15(04):1750017，2017年。Ximing Li、Jiaojiao Zhang 和 Jihong Ouyang。带变分流形正则化的狄利克雷多项式混合模型：短文本上的主题建模。载于*AAAI*人工智能会议论文集，第33卷，第7884–7891页，2019年。Xi Liu、Clio Andris 和 Bruce A Desmarais。美国的迁移与政治极化：县级人口迁移网络分析。《公共科学图书馆·综合》，14(11):e0225405，2019年。Jay Magidson、Jeroen K Vermunt 和 John P Madura。潜在类别分析。SAGE出版社，千橡市，加利福尼亚州，美国：2020年。缅因州公司、选举与委员会局。选举结果，2021年。URL https://www.maine.gov/sos/cec/elec/。

Daniele Mantegazzi。瑞士政治意识形态的地理分布随时间的变化。《空间经济分析》，16(3)：378–396，2021年。

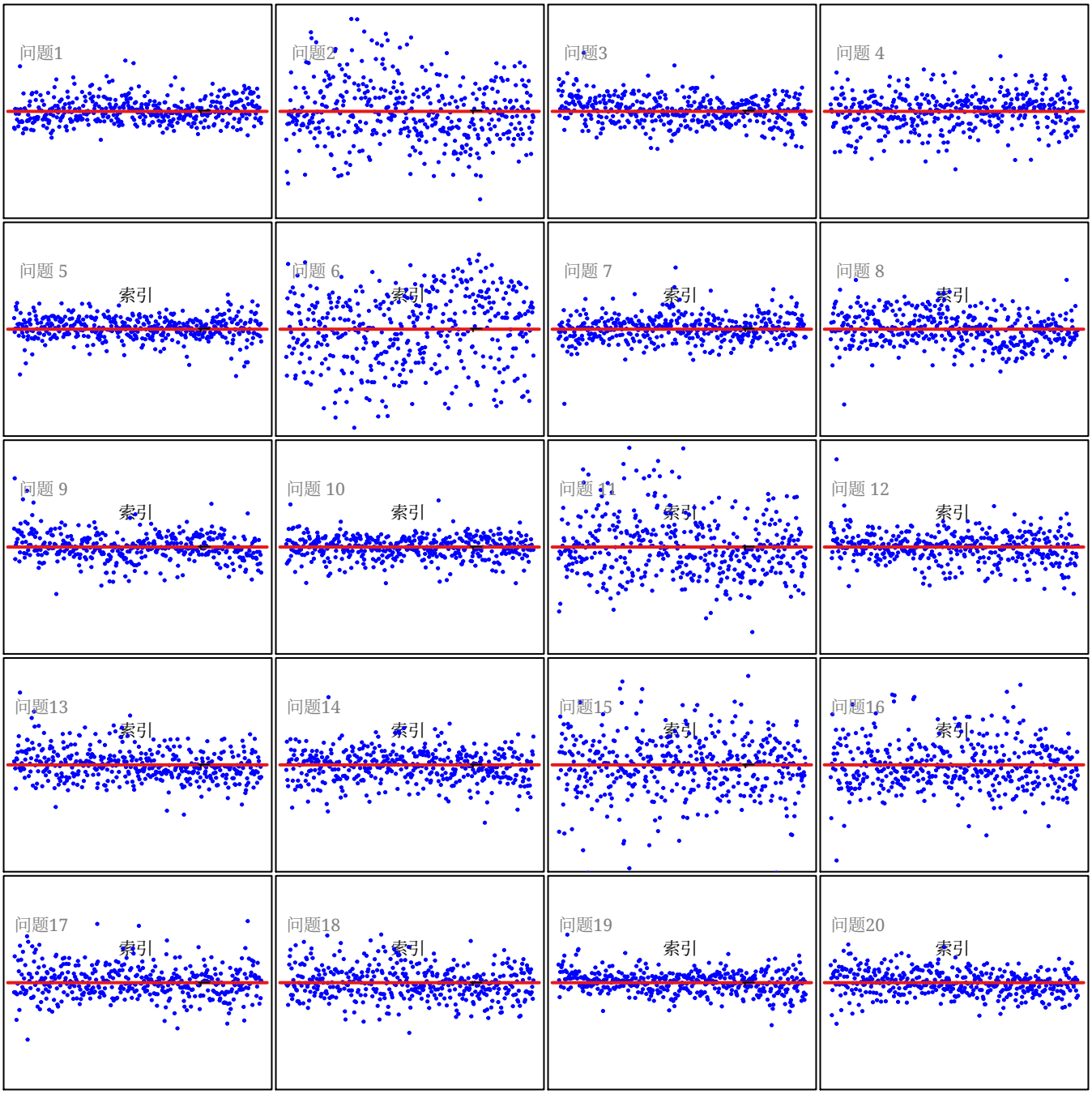
Jialiang Mao 和 Li Ma。用于微生物组组成聚类的狄利克雷树多项式混合模型。*The Annals of Applied Statistics*，16(3):1476–1499，2022年。John D May。政党的意见结构：曲线差异的特殊定律。*Political Studies*，21(2):135–151，1973年。Kevin McAlister。潜变量模型与唱名表决量尺分析论文集。博士论文，密歇根大学，2020年。HTML 链接。Allan L McCutcheon。潜在类别分析。第64卷。SAGE出版社，1987年。Geoffrey J McLachlan 和 Kaye E Basford。混合模型：用于推断和聚类的应用，第38卷。M. Dekker，纽约，1988年。Geoffrey J McLachlan、Sharon X Lee 和 Suren I Rathnayake。有限混合模型。统计学及其应用年鉴，6:355–378，2019年。Matthew Mendelsohn 和 Andrew Parkin。导言：公投民主。载于 公投民主，第1–22页。Springer，2001年。

David M Mimno 和 Andrew McCallum。基于任意特征并通过狄利克雷-多项式回归进行条件化的主题模型。载于 *UAI*，第24卷，第411–418页。Citeseer，2008年。John D O'Brien、Xavier Didelot、Zamin Iqbal、Lucas Amenga-Etego、Bartu Ahiska 和 Daniel Falush。从宏基因组数据推断社区系统发育结构的贝叶斯方法。*Genetics*，197(3):925–937，2014年。Markus Prior。媒体与政治极化。*Annual Review of Political Science*，16:101–127，2013年。Kevin Reuning、Michael R Kenwick 和 Christopher J Fariss。探索潜变量模型的动态。*Political Analysis*，27(4):503–517，2019年。Alexander A Schuessler。生态推断。*Proceedings of the National Academy of Sciences*，96(19):10578–10581，1999年。Charles A Scontras。倡议与公投：缅因州的奥德赛。*Lewiston Sun Journal*，2016年。JQ Shi、R Murray-Smith 和 DM Titterington。针对成分数量未知的混合模型的出生-死亡马尔可夫链蒙特卡洛方法。技术报告，Citeseer，2002年。Nathan M Stein 和 Xiao-Li Meng。使用复合界链进行实用的完美抽样：狄利克雷-多项式模型。*Biometrika*，100(4):817–830，2013年。Matthew Stephens。正态分布混合模型的贝叶斯方法。博士论文，英国牛津，1997年。Matthew Stephens。成分数量未知的混合模型的贝叶斯分析——可逆跳跃方法的替代方案。*Annals of Statistics*，第40–74页，2000年。Laurens Van der Maaten 和 Geoffrey Hinton。使用 t分布随机邻域嵌入可视化数据。《机器学习研究杂志》，9(11)，2008年。Mikael Vejdemo-Johansson、Gunnar Carlsson、Pek Yee Lum、Alan Lehman、Gurjeet Singh 和 Tigran Ishkhanov。政治的拓扑结构：美国众议院的投票连通性。载于*NIPS 2012 Workshop on Algebraic Topology and Machine Learning*，2012年。David L Weakliem 和 Robert Biggert。当代美国的地区与政治观点。*Social Forces*，77(3):863–886，1999年。

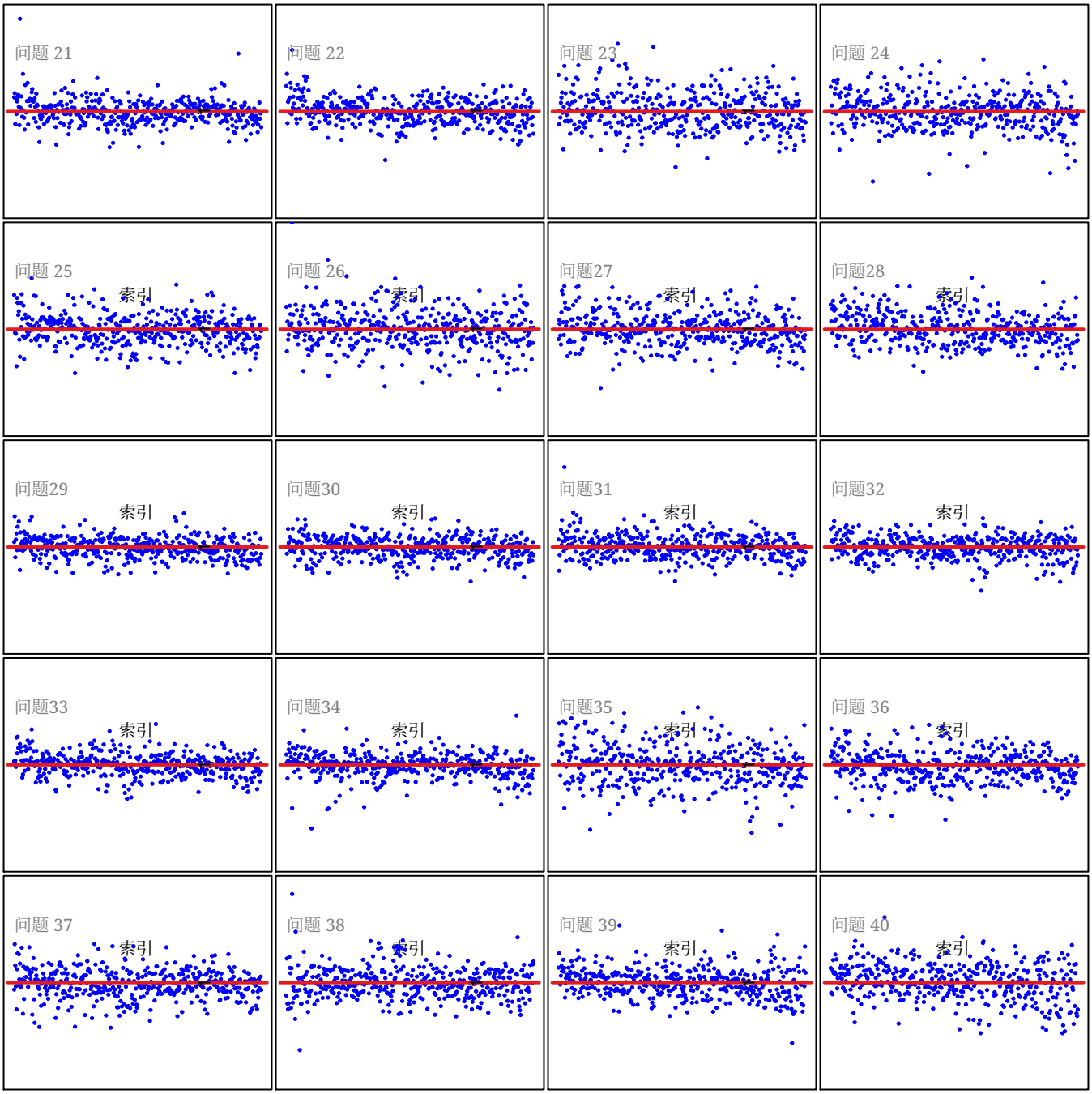
WolfgangWeidlich。《社会动力学的协同建模概念及其在集体政治观点形成中的应用》。《数学社会学杂志》，18(4)：267–291，1994年。

Xing Yi、Yunpeng Xu 和 Changshui Zhang。有限混合模型的多视图EM算法。载于国际模式识别与图像分析会议，第420–425页。Springer，2005年。

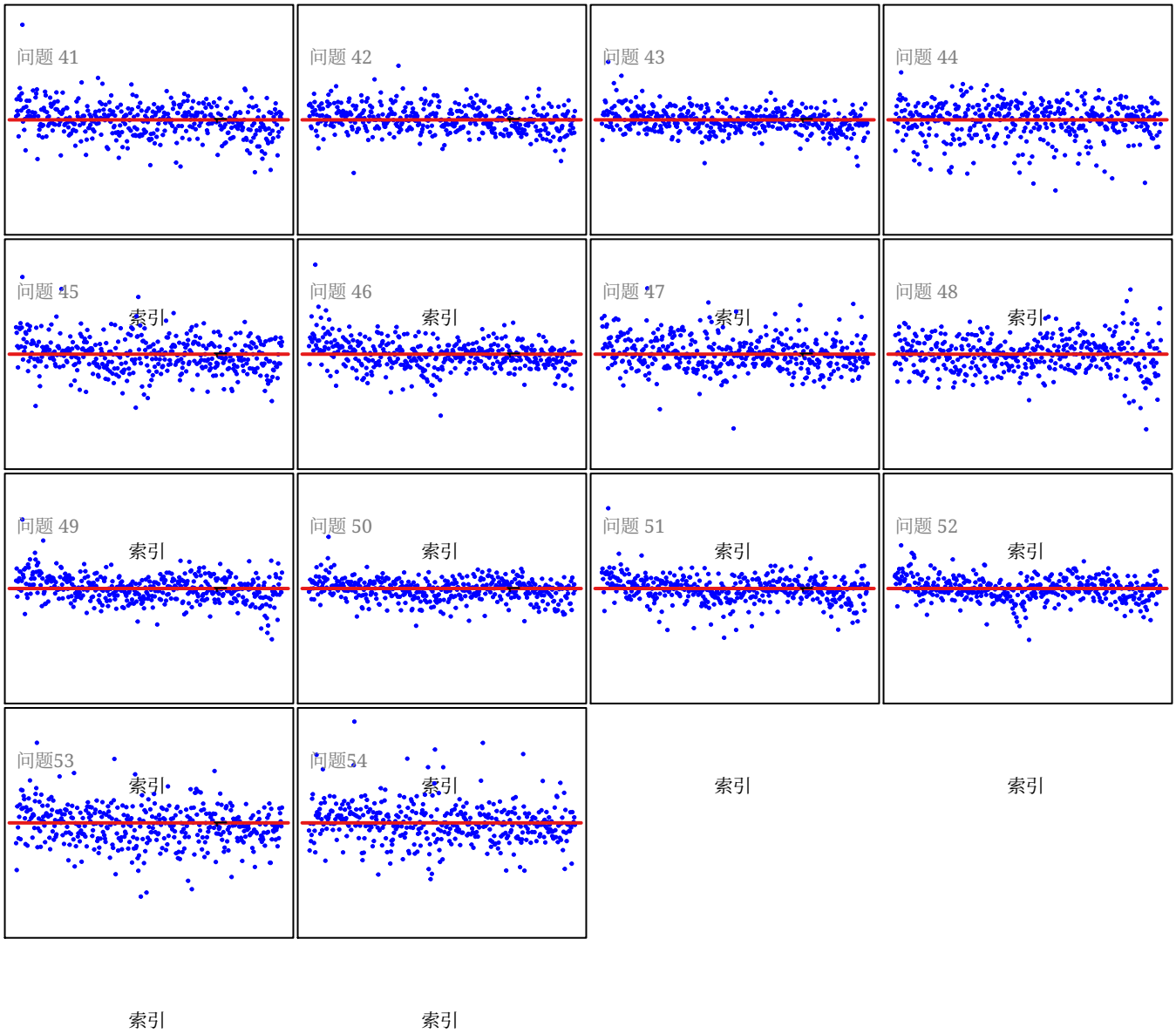
补充图



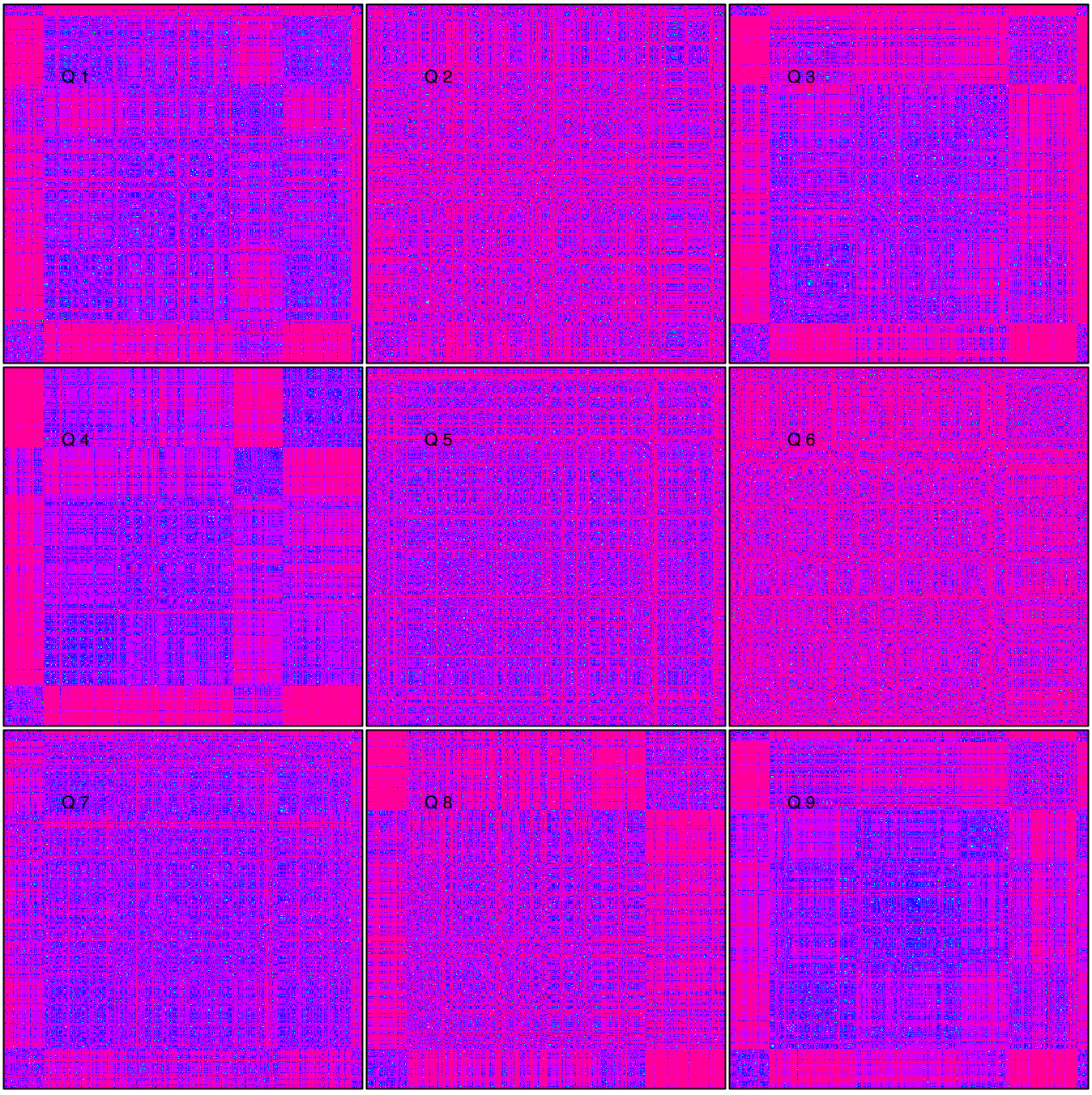
补充图1：按代表性聚类排序的每个市镇的估计问题拟合度。 y轴范围从 -0.03 – -0.03。问题1至20。



补充图2：按代表性聚类排序的各市镇估计问题拟合度。 y轴的刻度范围为 -0.03 – -0.03。问题21–40。



补充图3：按照代表性聚类排序的各市镇估计问题拟合度。 y轴刻度范围为 -0.03 – -0.03。问题 41–54。



补充图4：使用詹森-香农散度衡量各市镇之间支持度的两两相似度。蓝色表示高度相似，红色表示高度不相似。问题1-9。问题10-54可在项目的 GitHub 上获取。