arXiv:2301.02962v1 [stat.ME] 8 Jan 2023

基于可交换随机划分先验分布的贝叶斯图形实体解析*

Neil G. Marchant 和 Benjamin I. P. Rubinstein 墨尔本大学计算与信息系统学院

Rebecca C. Steorts
统计科学系与计算机科学系
杜克大学

摘要

实体解析（记录链接或去重）是识别并链接数据库中重复记录的过程。本文提出一种用于实体解析的贝叶斯图模型方法，将记录链接到潜在实体，并对链接结构采用可交换的先验表示。首先，我们为链接结构采用一组灵活且易于处理的先验分布，这对应于一类特殊的随机划分模型。其次，我们针对分类/离散记录属性提出一种更符合实际的失真模型，以纠正标准命中-未命中模型中的逻辑不一致。第三，我们引入超先验以提升灵活性。第四，我们采用部分塌缩吉布斯采样器来加快推断。我们使用一系列私有和非私有数据集，考察这些建模贡献的影响，并将模型与另外两种贝叶斯模型进行比较。此外，我们针对住户调查数据开展模拟研究，改变失真程度、重复率和数据集大小。结果表明，在各种情景下，我们的模型都比其他模型表现得更加一致，且通常能取得最高的实体解析准确率（F1 分数）。我们为所提出的方法提供了开源软件，并讨论了本研究及未来方向。

重要性声明

在调查统计中，实体解析（记录链接）用于识别多项调查中由同一个人提交的响应，即使未记录社会

*这是经同行评审后获准在《调查统计与方法学杂志》发表的文章的预先编辑、由作者制作的版本。正式发表版本为 Marchant, N. G., Rubinstein, B. I. P. 和 Steorts, R. C.（2023），《运用可交换随机划分先验的贝叶斯图形实体解析》《调查统计与方法学杂志》现已在线发布，网址：https://doi.org/10.1093/jssam/smac030。

保障号码等唯一标识符。本文通过以下方式推进实体解析的贝叶斯方法：（i）全面评估一类用于描述响应与个人之间链接关系的先验分布；（ii）提出一种更符合实际的模型，用于刻画个人身份识别属性（名称、地址等）中出现的失真。我们分别及联合使用多种数据集对这两项工作进行评估，其中一个数据集来自一项纵向医学调查。结果表明，只要采用宽松的超先验，通用先验分布类别——Ewens-Pitman随机划分——在此前研究过的三个参数区间内都能达到相近的准确度。这一发现意义重大，因为它解答了文献中关于哪个参数区间（如果有的话）最适用于实体解析的疑问。此外，研究发现，所提出的失真模型显著提高了实体解析的准确度，尤其适用于字符串类型属性。我们提供了两项模拟研究，为本文的方法和比较结果提供支持。本文还配套提供了一个 R 软件包，使用经过优化的部分塌缩吉布斯采样器实现所提出的模型。

关键词： 实体解析、记录链接、贝叶斯模型、可交换性、随机划分、*Ewens-Pitman* 随机划分

1 引言

正如文献中普遍所知，实体解析（ER；记录链接或去重）是指对规模庞大且含有噪声（脏数据）的数据库进行处理，并移除其中的重复记录（通常缺少唯一标识符）的过程（Doan 等人， 2012；Elmagarmid 等人， 2007；Naumann 和 Herschel，2010；Getoor 和 Machanavajjhala，2012； Christen， 2012a； Christophides 等人， 2020；Ilyas 和 Chu， 2019；Papadakis 等人， 2021； Binette 和 Steorts，2022）。这一问题在许多领域日益重要，例如调查方法学、官方统计、计算机科学、政治学、医疗保健、人权等。本文受到若干应用场景的启发。例如，我们考察一项纵向医疗保健调查，由于数据集受到隐私限制，其中的信息属于分类变量。这可能会引起调查方法学领域人士的兴趣，因为他们也可能遇到类似问题。此外，我们还考察分类数据集以及基于字符串或文本的数据集（或调查），例如书目／引文文献、餐馆信息，以及一项传统的基准（合成）研究。分析多个数据集的目标，是让调查领域更广泛地了解与实体解析方法相关的数据集。其他总体目标包括：将近期提出的贝叶斯图模型实体解析方法扩展到这些数据集，开展比较分析和模拟研究，并为研究人员提供指导。此外，我们还为学界提供开源软件，供其使用本文提出的方法以及文献中近期提出的两种方法。

实体解析的思想可以追溯到Dunn (1946年），他设想了一本“生命之书”，用于拼接关于某个个体的信息。Newcombe 等人(1959年）提出了首批用于执行实体解析的方法之一，该方法基于启发式统计检验。后来，Fellegi andSunter(1969年）将这一方法形式化。他们建立了一个基于记录对之间一致模式的模型，并采用似然比检验，将记录对分类为已链接（指向同一实体）、可能已链接或未链接。在一些较强的假设下，他们证明了其方法——如今称为Fellegi-Sunter（FS）方法——在统计上是最优的。数十年来，FS 方法在实体解析文献中不断发展，尤其是在调查方法学领域，这是由于其

可扩展性、易用性和简洁性（Winkler，2006；Christen，2012a；Sadinle and Fienberg，2013；Enamorado et al.，2019）。然而，它存在一些固有的局限性：会产生不一致（非传递）的预测，无法自然地处理不确定性，不能利用实体层面的模式，并且与生成式建模方法不兼容（Tancredi and Liseo，2011年）。

这些局限性中的一部分可以通过将 FS 方法调整为贝叶斯框架来解决，同时对记录之间的链接施加一致性约束（Sadinle，2014年，2017）。例如，Sadinle (2014年) 提出了一种 FS 模型的贝叶斯扩展，用于在单个数据库中执行实体解析。该模型通过要求将记录划分为相互链接的组来纳入一致性（传递性）约束。此外，该模型支持多个一致性等级，并纳入关于链接和 m/u 概率（来自 FS 模型）的先验分布。与传统 FS 方法相比，通过计算链接的后验分布，可以量化实体解析的不确定性。然而，尽管该模型具备这些优点，它在文献中仍未得到广泛研究，原因或许部分在于缺少公开可用的实现。本文的目标之一是评估 Sadinle 的模型，将其作为贝叶斯 FS 模型的代表性示例，并将其与另一类贝叶斯实体解析模型进行比较，下面将对此类模型加以介绍。

与贝叶斯 FS 模型的发展同步，其他研究者提出了一类新的生成模型，称为贝叶斯 图模型实体解析模型（Tancredi和Liseo，2011年；Steorts，2015年；Steorts 等人，2016年）。与 FS 方法不同，这些模型并不基于成对记录之间的一致模式进行运算。相反，它们对实体的潜在总体，以及记录由实体生成的过程进行建模。这类模型被称为“图模型”，因为模型中的基本对象构成一个二分图——潜在实体对应一组顶点，记录对应另一组顶点，而链接则是连接这些顶点的边。Tancredi和Liseo (2011年) 提出了一种最早用于跨两个数据库进行实体解析的此类模型。随后，Steorts 等人 (2016年) 将模型扩展至多个数据库，同时允许每个数据库内存在重复记录。不过，Steorts 等人(2016年) 发现，他们的模型以及 Tancredi和Liseo (2011年)提出的模型存在一项局限：链接结构上的均匀先验对数据中实体的数量提供了很强的信息。他们指出，未来工作应考虑更合适的实体解析链接结构先验分布。

Tancredi和Liseo 的模型（2011年）和Steorts 等人（2016年）都假设实体由一组潜在类别属性（例如出生日期、性别）描述，而这些属性在记录中会发生失真。然而，他们对失真过程的建模较为简单，无法捕捉字符串型属性（例如姓名）中现实存在的失真。Steorts（2015年）通过在一个称为blink的模型中提出字符串伪似然和经验驱动先验，解决了这一问题。blink 模型后来被Marchant 等人（2021年）用作基础，用于开发更具可扩展性的贝叶斯图模型实体解析技术。他们提出了一种端到端方法，联合执行分块和实体解析，并且可以在分块层级进行分布式或并行化推断。重要的是，这使得分块不确定性能够传递到实体解析任务中。他们观察到 200× 的加速效果，从而能够将blink 扩展到包含超过一百万条记录的数据集。然而，blink 模型对链接结构采用了与Tancredi和Liseo （2011年）和Steorts 等人

（2016年）相同的均匀先验，因此也存在同样的局限性。

受现有贝叶斯图形实体解析模型的不足所启发，本文提出并评估了若干建模改进。首先，我们为链接结构提出了一组灵活且易于处理的先验分布，即 Ewens-Pitman（EP）随机划分模型族（Pitman，2006年）。这是满足可交换性（一项基本要求）的最一般先验分布族，并且比以往研究采用的均匀先验分布更为灵活。其次，我们在 EP 参数上引入超先验，以进一步提高灵活性并减少调参需求。这一做法的动机在于，以往研究采用的均匀先验分布所包含的信息量较大（Steorts 等人，2016年）。第三，我们为分类／离散属性提出了一种更细致的失真模型，修正了 Tancredi和Liseo (2011年)、Steorts 等人(2016年) 和 Steorts (2015年) 所采用的标准命中-未命中模型中的一处不一致。第四，我们设计了一个部分塌缩吉布斯采样器来拟合模型，并在其中融入了计算优化。

我们在一组私有和非私有数据集上分别及联合评估了我们的建模贡献，并将我们的模型与Steorts(2015年)提出的贝叶斯图形实体解析模型，以及Sadinle (2014年)提出的贝叶斯 FS 模型进行了比较。我们还在一项受控模拟研究中评估了我们的模型（以及其他方案），在该研究中，我们生成了合成住户调查数据，并改变记录数量、失真程度和重复率。总体而言，我们发现，在所测试的各种情境中，我们的模型更具稳健性，且通常能达到更高的实体解析准确率。我们为所有受评估的实体解析方法提供了开源软件，并讨论了我们的贡献及未来工作方向。

本文其余部分安排如下。第2 节介绍实体解析和可交换随机划分的背景，并概述全文使用的符号。第3 节介绍我们提出的贝叶斯图形实体解析模型。第4 节介绍一种部分折叠吉布斯抽样算法，用于近似后验分布，并介绍其他计算加速方法。第5 节展示了对我们提出的失真模型和链接结构先验的实证研究，并将其与近期提出的两种贝叶斯实体解析模型进行比较。第5.5 节概述了一项收录于附录中的受控模拟研究。第6节总结了我们的研究发现。

2 背景与符号说明

本节介绍本文模型中作为先验分布使用的可交换随机划分，并说明相关符号和假设。图 2 列出了全文中使用的符号。

2.1 符号说明与假设

我们回顾本文中使用的符号和假设。我们假设数据（来自一个或多个数据源）是结构化的，也就是说，数据已通过模式对齐技术进行了标准化。在本文中，数据以表格格式表示，其中行对应记录，列对应属性。这与非结构化实体解析不同，后者处理文本描述（段落）或图像。有关这些术语的完整回顾，请参见Papadakis 等人（2021）。

令 $s \in \{1, \ldots, S\}$ 为数据源的索引，令 $i \in \{1, \ldots, N\}$ 为记录的索引，该索引在所有数据源中均唯一。第 i 条记录的来源记为 $\varsigma_i \in \{1, \ldots, S\}$，记录的属性值表示为元组 $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iA})$，其索引为 $a \in \{1, \ldots, A\}$。假设对于所有 i 和 a，$x_{ia} \in \mathrm{D}_a$，其中第 a 个属性的取值域 D_a 是有限字符串集合。假设存在一个（可能无限的）实体总体，其实体以 $e \in \mathrm{N}$ 为索引，并由数据表示。第 i 条记录所指代的实体记为 $\lambda_i \in \mathrm{N}$，并将链接结构定义为 $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_N)$。

我们考虑最一般的情形，即对 $\boldsymbol{\Lambda}$ 不施加任何约束——也就是说，允许同一数据源中存在重复记录，也允许不同数据源之间存在任意链接。链接结构 $\boldsymbol{\Lambda}$ 将记录划分为若干簇。我们根据簇所对应的实体为簇编号，并允许存在空簇。簇 e 的大小记为 $N_e = \sum_i \mathbb{1}[\lambda_i = e]$，非空簇的数量记为 $E = \sum_e \mathbb{1}[N_e > 0]$。

我们关注完全无监督的情形，即对链接结构或实体一无所知。我们的目标是仅根据观察到的记录属性$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$以及数据源标识符$\{\varsigma_1, \ldots \varsigma_N\}$推断链接结构。由于我们采用贝叶斯框架，因此希望获得链接结构的完整后验分布（而非仅仅一个点估计），以便将不确定性传递到实体解析后续任务中，例如回归、多系统估计等（Kaplan 等人，2022年；Tancredi和Liseo，2015年；Steorts 等人，2018；Tancredi 等人，2020；Sadinle，2018）。尽管实体解析后续任务并非本文的目标，但上述文献提出了这类任务的最新方法。

2.2 可交换随机划分

链接结构是实体解析中首要关注的变量，因此在设计模型时，我们会特别关注它。由于我们采用贝叶斯框架，因此必须为链接结构指定先验分布。我们之前指出，链接结构可以解释为记录的划分，即将记录划分为若干子集，其中每个子集中的记录都对应同一个实体。这种解释很方便，因为我们可以借鉴随机划分方面的相关研究来考虑可能的先验分布。本节将回顾一类特殊的随机划分，称为*Ewens-Pitman*（*EP*）族（Pitman，2006，第 62 页），我们在模型中将其用作链接结构的先验分布（见第3节）。

在定义 EP 随机划分族之前，我们先介绍关键概念和符号。考虑一组 N 记录，其中 $[N] = \{1, \ldots\ldots, N\}$ 表示记录标识符。$[N]$ 的划分是由 $[N]$的若干互不相交的非空子集构成的集合。例如，$[N]$ 对于 $N = 3$的划分可以是 $\{1, 2\}, \{3\}$，也可以是 $\{1\}, \{2\}, \{3\}$。我们也可以用链接结构$\boldsymbol{\Lambda} = (\lambda_i)_{i=1\ldots N}$ 来等价地定义划分，其中 λ_i 表示记录 i 被分配到的子集（实体）的标签。[1] 按照这种表示法，$N = 3$ 的上述示例可以写为 $\boldsymbol{\Lambda} = (1, 1, 2)$ 和 $\boldsymbol{\Lambda} = (1, 2, 3)$。

令 $\mathrm{P}[N]$ 表示 $[N]$ 的所有划分组成的集合。$[N]$ 的随机划分是一个取值属于 $\mathrm{P}[N]$ 的随机变量。EP族是满足以下两项理想性质的最一般的随机划分类：

[1]可以使用任何标签来标识实体。唯一重要的是，如果记录 i 和 j 被分配给同一个实体，则 $\lambda_i = \lambda_j$ $\lambda_i \neq \lambda_j$；否则 $\lambda_i \neq \lambda_j$。

以下两项理想属性：

1. 可交换性。 这意味着，划分 P_N 上的分布 P_N 在记录标识符 $[N]$的排列下保持不变。或者等价地， $\mathbf{\Lambda}$ 上的分布具有可交换性。如果记录没有自然的顺序（例如，不知道某条记录是在另一条记录之前还是之后生成的），这是一项合理的要求。

2. 一致性。 这是分布随 N 变化时所具有的一项性质。它确保即使观察到更多记录，分布也不会改变。这是可取的，因为模型可以按顺序以一致的方式进行学习。如果在 $\mathrm{P}_{[1]}, \mathrm{P}_{[2]}, \ldots$上的分布序列 $P_1, P_2, \ldots$满足以下条件，我们就称其为一致的：由 P_M针对 $M > N$诱导的 $\mathrm{P}_{[N]}$上的分布为 P_N。数学上，这意味着

$$P_M(\{\rho : \mathrm{Proj}(\rho, [N]) = \rho'\}) = P_N(\rho'),$$

其中

$$\mathrm{Proj}(\rho, [N]) := \{A_e \cap [N] : A_e \cap [N] \neq \emptyset, A_e \in \rho\}$$

是投影 将 $[M]$的划分 $\rho = \{A_e\}_{e=1\ldots E}$投影到 $[N]$上，而 ρ'是 $[N]$的一个划分。

事实上，EP族保证这些性质在极限意义下成立 $N \to \infty$ (Pitman, 2006, 第 62 页）。

我们可以通过考察随机划分如何逐条记录按顺序生成，直观地理解 EP 族。设 $(\rho_N)_{N=1,\ 2,\ \ldots\ldots}$ 是一列 EP 随机划分，其中 ρ_N 是 $[N]$的一个随机划分。我们从第 1 步开始，此时为 $\rho_1 = \{1\}$——即一条记录被分配给一个实体。后续任一步 $N > 1$ 的随机划分 ρ_N，都是以第 $N-1$步的随机划分 ρ_{N-1} 为条件生成的。设 E 为 ρ_{N-1} 中的子集（已占用实体）数量，N_e 为 ρ_{N-1}中子集（实体） e 的大小。那么， ρ_N 通过将记录 N 分配给以下对象而生成：

- 以概率 $\frac{N_e - \sigma}{N + \alpha}$ 分配给一个已有子集（实体），或
- 以概率 $\frac{\alpha + E\sigma}{N + \alpha}$ 生成一个“新”子集（实体），

其中 σ 和 α 是 EP 参数。这种构造称为双参数中国餐馆过程，并在图 1中进行了可视化。

EP 参数的允许取值根据 σ 的正负分为两种情形：

- $\sigma < 0$ 且 $\alpha = -m\sigma$ 对于某个 $m \in \mathrm{N}$。我们将这种情形称为广义优惠券划分，因为它们与优惠券收集者划分密切相关（Pitman, 2006, 第 46 页）。这些划分通过从大小为 m的有限总体中进行有放回抽样生成，其中混合比例取自浓度参数为 $-\sigma$的对称狄利克雷分布。优惠券收集者划分可在极限$\sigma \to -\infty$下得到。

- $0 \leq \sigma \leq 1$ 与 $\alpha > -\sigma$。这些被称为*Pitman-Yor* 分区，以Pitman和Yor（1997）命名，并通过从无限总体中进行有放回抽样生成。由此产生的

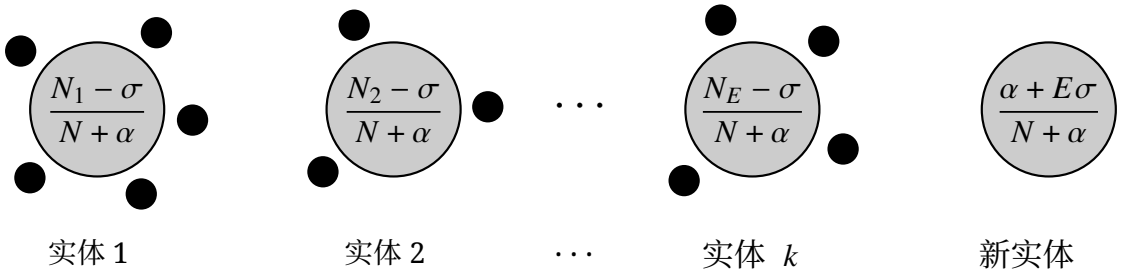


图 1：Ewens–Pitman随机划分的序贯构造示意图。在第N 步，根据先前 $N-1$ 条记录的分配情况，将一条记录（黑色圆点）分配给某个已占用实体或一个新实体（灰色圆点）。分配给各实体的概率标示在灰色圆圈内，并取决于 Ewens-Pitman 参数 σ 和α。

划分表现出优先连接行为。特殊情况 $\sigma = 0$ 对应于Ewens划分（Kingman， 1978）。

为了说明随机划分随 σ变化的行为，我们可以考察划分（实体）中子集的渐近数量 E_N 当 $N \to \infty$时。Pitman (2006，第 70 页）证明了

$$E_N \overset{a.s.}{\asymp} \begin{cases} m, & \sigma < 0 \text{ and } \alpha = -m\sigma \text{ for } m \in \mathbb{N}, \\ \alpha \log N, & \sigma = 0 \text{ and } \alpha > 0, \\ S_\sigma N^\sigma, & 0 < \sigma < 1 \text{ and } \alpha > -\sigma, \end{cases} \tag{1}$$

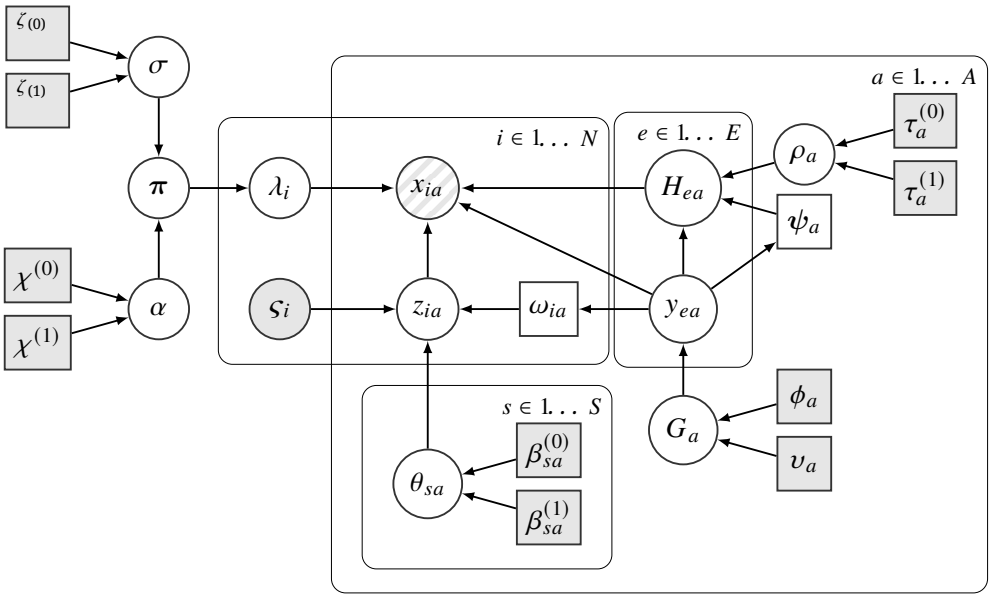
其中， S_σ 是一个严格为正的随机变量。因此，通过改变 σ，我们可以表达这样一种先验信念：实体数量 E_N 渐近而言是常数、对数级增长，或相对于 N呈次线性增长。

3 图模型贝叶斯实体解析

本节中，我们提出一种用于实体解析的生成式模型，该模型纳入了一个由实体构成的潜在总体，其中每个实体都具有一组未知属性。我们的模型对链接结构采用 Ewens-Pitman 类先验分布，并使用一种经过修改的记录失真模型；该模型不同于Tancredi和Liseo (2011年）、Steorts (2015年）和Steorts 等人(2016年采用的常见“命中—未命中”模型。我们在图2中列出了模型变量，并展示了它们的依赖结构。本节最后讨论模型的超参数，并就如何在先验信息有限时设定这些值提出建议。

3.1 模型设定

实体。 我们假设每个实体 $e \in \{1,2,\ldots\}$ 都与一个属性值元组相关联$\boldsymbol{y}_e = (y_{e1}, \ldots, y_{eA})$，这些属性值从一个支撑集为 $\mathrm{D} = \prod_a \mathrm{D}_a$ 的未知分布 G 中独立同分布（i.i.d.）抽取。为提高可处理性，我们假设相关性



i	记录索引
s	数据源索引
a	属性索引
e	实体索引
x_{ia}	记录 i的属性 a
z_{ia}	记录 i的属性 a 的失真指示变量
ω_{ia}	记录 i的属性 a 的失真倾向
ς_i	记录 i的来源
λ_i	记录 i的关联实体

H_{ea}	实体 e的属性 a 的失真分布
ρ_a	H_{ea}的浓度
ψ_a	H_{ea}的基分布
θ_{sa}	来源 s 中属性 a 的失真概率
y_{ea}	实体 e 的属性 a
G_a	属性 a 在取值域上的分布
π	混合比例
σ, α	Ewens-Pitman 参数
$\mathcal{D}_a$	属性 a 的取值域
dist_a	属性 a 的距离度量

图 2：Pitman-Yor 先验下模型的板式图及符号索引。

属性之间的关联可以忽略，并对 $\boldsymbol{G} = (G_1, \ldots, G_A)$ 的每个分量分别设置相互独立的狄利克雷过程（DP）先验分布：

$$\begin{aligned} G_a &\overset{\text{ind.}}{\sim} \mathrm{DP}(\upsilon_a, \phi_a)\,, && \forall a, \\ y_{ea} \mid G_a &\overset{\text{ind.}}{\sim} \mathrm{Discrete}(G_a), && \forall e, a, \end{aligned}$$

其中 $\upsilon_a > 0$是浓度参数， ϕ_a 是取值域为 $\mathrm{D}a$ 的基分布。

链接。 每条记录 $i \in \{1, \ldots, N\}$ 与一个实体 $\lambda_i \in \{1, 2, \ldots\}$ 建立链接，假设这些实体根据未知的混合比例 $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)$从总体中有放回抽取。此过程会根据记录所链接的实体，将记录划分为若干簇。根据第 2.2 节关于可交换性的讨论，我们假设该划分来自参数为 (σ, α) 的 Ewens-Pitman（EP）族。混合比例 $\boldsymbol{\pi}$ 上相应的分布取决于 σ 的符号，或者说，取决于实体总体是有限还是无限。

对于有限情形（广义优惠券划分），令 $\sigma = -\kappa < 0$ 和 $\alpha = m\kappa$，其中某个$m \in \mathrm{N}$。我们对 m 和 κ 设置超先验的模型如下：

$$\begin{aligned} \kappa &\sim \mathrm{Gamma}(\chi^{(0)}, \chi^{(1)}), \\ m &\sim \mathrm{NegativeBinomial}(r, \nu) + 1, \\ \boldsymbol{\pi} \mid \kappa, m &\sim \mathrm{Dirichlet}(\boldsymbol{\kappa}), \\ \lambda_i \mid \boldsymbol{\pi} &\overset{\text{iid.}}{\sim} \mathrm{Categorical}(\boldsymbol{\pi}), && \forall i, \end{aligned} \tag{2}$$

其中 $\chi^{(0)}$、$\chi^{(1)}$、$r > 0$ 和 $0 < \nu \leq 1$ 是超参数，而 $\boldsymbol{\kappa}$ 是一个长度为 m 且各项均为 κ的向量。对 m 设置的超先验是平移负二项分布，其密度定义见附录 A.4.2。

在无限情形（Pitman-Yor 分区）中，混合比例服从二参数泊松-狄利克雷分布（Pitman和Yor，1997）。我们对 σ 和 α 设置超先验的模型如下：

$$\begin{aligned} \sigma &\sim \mathrm{Beta}(\zeta^{(0)}, \zeta^{(1)}), \\ \alpha &\sim \mathrm{Gamma}(\chi^{(0)}, \chi^{(1)}), \\ \boldsymbol{\pi} \mid \sigma, \alpha &\sim \mathrm{PoissonDirichlet}(\sigma, \alpha), \\ \lambda_i \mid \boldsymbol{\pi} &\overset{\text{iid.}}{\sim} \mathrm{Categorical}(\boldsymbol{\pi}), && \forall i, \end{aligned} \tag{3}$$

其中 $\chi^{(0)}$、$\chi^{(1)}$、$\zeta^{(0)}$、$\zeta^{(1)} > 0$ 是超参数。这里我们假设 $\alpha > 0$ 和 $0 < \sigma < 1$，这是可容许参数空间的子集： $0 \leq \sigma \leq 1$ 和 $\alpha > -\sigma$。我们还考虑 $\sigma = 0$的情形，这对应于 Ewens划分。

备注。 通过为 EP 参数设置超先验，我们可以提高模型在面对设定错误的超参数时的稳健性，因为这些超参数难以通过非信息性方式设定。上述先验分布的特殊情形已用于其他实体解析模型中，但超参数是固定的。 Tancredi和Liseo(2011年）、Steorts (2015年）和Steorts 等人(2016年）使用了优惠券收集者划分，其中 $\kappa \to \infty$和

m 固定不变；研究表明，这对观测到的总体规模具有很强的信息性。 Steorts 等人(2018）使用了 Pitman-Yor 划分，其中 σ 和 α 均固定。

数据源。我们假设与记录 i 关联的数据源 $\varsigma_i \in \{1, \ldots\ldots, S\}$ 是从定义在数据源 $\{1, \ldots\ldots, S\}$ 上的离散分布 $\boldsymbol{\xi}$ 中独立同分布抽取的。由于 $\boldsymbol{\xi}$ 与其他模型参数相互独立，因此无需对其进行指定；并且假设数据源指示变量 ς_i 是完全观测到的。

失真。我们假设记录 i 的属性 $\boldsymbol{x}_i$ 是通过对关联实体的实体属性 $\boldsymbol{y}_{\lambda_i}$ 进行失真处理而生成的。为简单起见，我们假设失真过程对每个属性独立进行。为判断第 a 个属性是否发生失真，会抽取一个二元指示变量 z_{ia}，其取值取决于失真倾向 ω_{ia} 与数据源／属性级因子 $\theta_{\varsigma_i a}$ 的缩放结果。我们对 $\theta_{\varsigma_i a}$ 设定贝塔先验，并假设在给定真实属性值 $y_{\lambda_i a}$ 时，失真倾向 ω_{ia} 是确定的。具体而言，我们有

$$\theta_{sa} \overset{\text{ind.}}{\sim} \text{Beta}\left(\beta_{sa}^{(0)}, \beta_{sa}^{(1)}\right) \qquad \forall s, a \qquad (4)$$

$$\omega_{ia} \mid y_{\lambda_i a} = \text{propensity}\left(\min_{x \in \mathcal{D}_a \setminus \{y_{\lambda_i a}\}} \text{dist}_a(y_{\lambda_i a}, x), \max_{x,y \in \mathcal{D}_a} \text{dist}_a(y, x)\right) \qquad \forall i, a$$

$$z_{ia} \mid \theta_{\varsigma_i a}, \omega_{ia} \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\theta_{\varsigma_i a} \omega_{ia}\right) \qquad \forall i, a \qquad (5)$$

其中，$\beta_{sa}^{(0)}$，$\beta_{sa}^{(0)} > 0$ 是超参数，并且

$$\text{propensity}(d_{\min}, d_{\max}) := \begin{cases} 0, & d_{\min} = \infty \text{ and } d_{\max} = \infty, \\ 1, & d_{\min} = 0 \text{ and } d_{\max} = 0, \\ \mathrm{e}^{-\frac{d_{\min}}{2 d_{\max}}}, & \text{otherwise.} \end{cases} \qquad (6)$$

失真倾向 ω_{ia} 考虑到某些实体属性值 $y_{\lambda_i a} \in \mathcal{D}_a$ 比其他值更容易发生失真。它利用属性距离度量中的先验信息$\text{dist}_a(y, x)$（见第3.2节）。如果 $y_{\lambda_i a}$ 与取值域中的其他任何值都不接近，那么它不太可能发生失真，且 ω_{ia} 趋近于零。另一方面，如果 $y_{\lambda_i a}$ 与取值域中至少一个其他值接近，就可能发生失真，且 ω_{ia} 趋近于一。Steorts（2015）提出的类似模型并未纳入这一逻辑，该模型实际上假设 $\omega_{ia} = 1$。

抽取失真指示变量 z_{ia} 后，记录属性 x_{ia} 由复制关联实体属性 $y_{\lambda_i a}$ 直接生成（若 $z_{ia} = 0$），或经过失真处理（若 $z_{ia} = 1$）。若 $z_{ia} = 1$，则从与关联实体 λ_i相关联的失真分布 $H_{\lambda_i a}$ 中抽取失真后的值。我们假设$H_{\lambda_i a}$ 本身是从狄利克雷过程中抽取的：

$$\rho_a \sim \text{Gamma}\left(\tau_a^{(0)}, \tau_a^{(1)}\right) \qquad \forall a, \qquad (7)$$

$$H_{ea} \mid y_{ea}, \rho_a \overset{\text{ind.}}{\sim} \text{DP}(\rho_a; \psi_a(y_{ea})) \qquad \forall e, a, \qquad (8)$$

其中 $\tau_a^{(0)}$, $\tau_a^{(1)} > 0$ 是超参数，而 $\psi_a(y_{ea})$ 是一个先验基分布，其支撑集是 $\mathcal{D}_a \setminus \{y_{\lambda_i a}\}$ 的子集。这与 Tancredi 和 Liseo (2011 年) 、 Steorts (2015 年) 以及 Steorts 等人 (2016 年、 2018) 的模型不同；这些模型假设 H_{ea} 在给定 的条件下是确定的 y_{ea}。

用符号概括如下：

$$x_{ia} \mid z_{ia}, y_{\lambda_i a}, H_{\lambda_i a} \overset{\text{ind.}}{\sim} \begin{cases} \delta(y_{\lambda_i a}), & \text{if } z_{ia} = 0, \\ H_{\lambda_i a}, & \text{if } z_{ia} = 1, \end{cases} \qquad \forall i, a \tag{9}$$

其中 $\delta(y)$ 表示位于 y 处的点质量。这类似于 命中-未命中模型 (Copas 和 Hilton, 1990)。然而，我们的构造有所不同：如果记录属性发生失真（$z_{ia} = 1$），则禁止记录属性与实体属性相匹配（$x_{ia} \neq y_{\lambda_i a}$）。

注。 Copas 和 Hilton 的命中-未命中模型 (1990)旨在对连续属性的失真进行建模。对于连续属性，假设 $H_{\lambda_i a}$ 由连续密度函数描述，则从未命中分量 $H_{\lambda_i a}$ 中抽取到未失真值（$y_{\lambda_i a}$）的概率为零。这确保了当 $z_{ia} = 1$ 时，记录值 x_{ia} 总是发生失真（$x_{ia} \neq y_{\lambda_i a}$）。我们的方案通过确保 H_{ea} 在 y_{ea} 上没有质量，从而为离散属性复现相同行为。如果 H_{ea} 在 y_{ea} 上具有显著质量，这一点尤其重要，因为失真值与未失真值之间的界限会变得模糊。除了建模方面的优势，将 y_{ea} 排除在 H_{ea} 的支撑集之外，也使推断更易处理，因为我们可以将 H_{ea} 消去（见附录A.2 和A.3）。

3.2 超参数的选择

本节将为如何设置模型中的超参数提供建议。

距离度量。 我们提出的失真模型由一组距离度量 $\{\text{dist}_a\}$ 参数化，每个属性 $a \in \{1, \ldots, A\}$ 对应一个距离度量。它们编码了有关记录属性值 x 作为实体属性值 y 的失真替代值出现的可能性的先验知识。距离 $\text{dist}_a(y, x)$ 越大， x 是 y 的失真形式的可能性就越低。由于将 x 失真为 y 的可能性未必与将 y 失真为 x 的可能性相同，因此我们不要求距离度量具有对称性。我们建议仔细选择距离度量，并尽可能利用有关失真过程的先验知识。例如，可以选择编辑距离来对通用字符串类型属性中的排版失真进行建模。对于分类属性，可以选择常数距离函数 $\text{dist}_a(y, x) \equiv 0$，它编码了这样的先验信念：取值域中的所有值作为 y 的失真替代值的可能性都相同。

失真基分布。 我们建议使用距离度量来设置基分布 $\psi_a(y_{ea})$ 在方程（8）中。具体而言，我们建议使用 Softmax 分布

$$\psi_a(x \mid y_{ea}) \propto \mathbb{1}[x \neq y_{ea}] \exp(-\text{dist}_a(y_{ea}, x)), \tag{10}$$

其中，温度参数已并入距离度量的定义，而指示函数会将 y_{ea} 排除在支撑集之外。这使取值域中

更接近 y_{ea} 的值获得更大的权重，距离较远的值则权重较小。与 Steorts (2015年) 不同，我们不纳入与 x的经验频率成比例的因子，因为对于我们所考虑的应用，失真值（例如，排版错误）往往并不常见。对于具有 $\text{dist}_a(y, x) \equiv 0$的分类属性，方程 (10) 会简化为均匀分布。在这种情况下，可以通过设置 $\psi_a(x \mid y_{ea}) \propto \mathbb{1}[x \neq y_{ea}] \sum_{i=1}^{N} \mathbb{1}[x_{ia} = x]$.

其他超参数。 在缺乏先验知识的情况下，我们建议将其余超参数设置为产生模糊先验——即，相对于实验而言仅提供少量信息的先验分布（Tiao 和 Box，1973；Bernardo 和 Smith，2009）。我们注意到，贝叶斯学界对于如何指定模糊和/或无信息先验存在不同观点。关于这一点，参见 Syversveen（1998）以及 Irony 和 Singpurwalla （1997）。撇开这些争论，我们的建议如下：

- 对于 m的平移负二项先验，我们设置 r 和 ν，使得先验均值为 N，先验方差为 N^2。
- 对于 α 的伽马先验，我们将 $\chi_{(0)} = 1$和 $\chi_{(1)}$ 设置为较小的值（例如， 10^{-2}）。
- 对于 σ 的贝塔先验，我们设置 $\zeta_{(0)} = \zeta_{(1)} = 1$ 以产生平坦先验。
- 对于实体属性分布的狄利克雷先验，我们建议将 $\upsilon_a = 1$ 设为 ϕ_a，并对所有 a 使用均匀基分布。
- 对于浓度参数 ρ_a 的伽马先验，我们建议对所有 a 将 $\tau_{(0)} = 2$设为小值（例如 10^{-4}），并将 $\tau_{(1)}$ 设为小值。
- 对于 $\theta_{s,\ a}$ 上的贝塔先验，我们通过将所有 s， a 的 $\beta_{(0)sa} = 1$设为 $\beta_{(1)sa} = 4$，来表达对低失真的弱先验信念。

备注。可以调整超参数，以促进过度链接（将并不对应同一实体的记录链接起来）或链接不足 （未能将对应同一实体的记录链接起来）。由于完美链接并非总是可行，从业者可能需要根据具体应用决定是优先避免过度链接还是链接不足。我们可以通过如下设置在模型中促进过度链接：

- $\beta_{(0)sa} \gg \beta_{(1)sa}$ （对高失真的先验信念），
- $\upsilon_a \ll 1$（对实体间属性 a 多样性较低的先验信念），
- $\zeta_{(0)} \ll \zeta_{(1)}$ 以及 $\chi_{(0)} \ll \chi_{(1)}$ （对 Pitman-Yor 先验会产生更多链接的先验信念），或者
- r 接近 0，且 ν 接近 1（对广义优惠券先验会产生更多链接的先验信念）。

同样，我们可以通过反转上述不等式来鼓励链接不足。在实证评估（第5节）中，我们使用方程（12）和（13）中定义的精确率和召回率指标，衡量模型过度链接或链接不足的程度。

4 推断

要使用我们的模型进行实体解析，我们必须求出给定观测到的记录属性及其数据源时，链接结构的后验分布。由于该后验分布无法通过解析方法求得，我们提出了一种基于马尔可夫链蒙特卡罗（MCMC）的近似推断方案。

马尔可夫链蒙特卡罗通过构造平衡分布与后验分布相匹配的马尔可夫链，从而生成后验分布的近似样本。马尔可夫链蒙特卡罗生成的样本之所以是近似的，是因为它们存在自相关，而且可能只有渐近地才与平衡（后验）分布相匹配。马尔可夫链蒙特卡罗框架下有多种算法——有关该领域的介绍，请参阅 Gamerman 和 Lopes (2006年) 或 Brooks 等人 (2011年)。

本文采用一种称为 部分坍缩吉布斯（*PCG*）抽样 (van Dyk 和 Park，2008年）的马尔可夫链蒙特卡罗算法。它是吉布斯抽样的一种推广，通过坍缩（边缘化掉）变量和／或成组更新变量，减少变量更新时的条件化程度。只要遵循规定的规则以确保马尔可夫链的平衡分布得以保持，这就能显著改善收敛性并降低自相关。

理想情况下，我们希望尽可能降低条件化程度，但这必须兼顾计算和数学方面的约束。在我们提出的采样方案中，我们完全坍缩 实体混合比例 π 和失真分布 H_{ea}。我们部分坍缩 失真指示变量 z_{ia}，在更新实体属性时联合更新 y_{ea} 以及失真分布集中度 ρ_a。通过坍缩混合比例，我们获得了一种用于更新链接结构的基于瓮的方案，类似于非参数混合模型中采用的方案（Neal，2000年）。本节余下部分将重点介绍推断中一些不太显然的方面——完整细节见附录 A。

4.1 非共轭性

虽然我们尝试在模型中保持共轭性，但在某些情况下仍无法避免使用非共轭先验分布。这使推断变得复杂，因为吉布斯抽样中使用的后验条件分布不再具有标准形式。已有多种成熟的方法可用于处理非共轭性，包括 Metropolis-Hastings算法（Chib 和 Greenberg，1995）、拒绝抽样（Gilks 和 Wild，1992）以及辅助变量方法（Damlen 等，1999）。我们选择使用辅助变量方法，因为它简单易行，无需设计提议分布，也无需监控接受率。

模型中有三组参数涉及非共轭性问题：

1. 方程（ θ_{sa} （4）中定义的失真概率 ω_{ia}，其中纳入失真倾向 θ_{sa} 会破坏贝塔先验的共轭性。我们在附录A.1中提出一种辅助变量抽样方案来更新 θ_{sa} 。

2. EP 参数　　参数： κ 和 m 定义于方程（2）中，或 α 和 σ 定义于方程（3　　），

具体取决于所处情形。我们采用Teh (2006）提出的辅助变量方案，在伽马先验和贝塔先验下更新 α 和 σ，相关内容总结于附录A.4.1。我们在附录 κ 和 m 中设计了辅助变量更新方法，在伽马先验和平移负二项先验下对其进行更新 A.4.2。

3. 失真分布集中度 ρ_a 在方程 (7) 中定义。我们在附录 ρ_a A.5中为 ρ_a 设计了一种辅助变量更新。

4.2 折叠失真指示变量

Marchant 等人 (2021) 展示了折叠失真指示变量 $\{z_{ia}\}$ 对于改善与方程 (9) 类似的命中-未命中模型的收敛性/混合效果的重要性。涉及 z_{ia} 的后验分布因子可分解为 i 和 a，因此折叠 z_{ia} 可得：

$$\begin{aligned} P(x_{ia} \mid \theta_{\varsigma_i a}, \omega_{ia}, y_{\lambda_i a}, H_{\lambda_i a}) &\propto \sum_{z_{ia}=0}^{1} P(x_{ia} \mid z_{ia}, y_{\lambda_i a}, H_{\lambda_i a}) P(z_{ia} \mid \theta_{\varsigma_i a}, \omega_{ia}) \\ &\propto (1 - \theta_{\varsigma_i a}\omega_{ia})\mathbb{1}[x_{ia} = y_{\lambda_i a}] + \theta_{\varsigma_i a}\omega_{ia} H_{\lambda_i a}(x_{ia}). \end{aligned} \tag{11}$$

我们利用这一结果，对实体属性 $\{y_{ea}\}$实施折叠更新。虽然也可以对链接结构 $\{\lambda_i\}$实施折叠更新，但我们选择不这样做，因为在给定失真指示变量的条件下，可以通过索引降低计算复杂度（见第 4.3节）。从经验上看，这样似乎更有效率（Marchant 等人，2021），只要失真程度不是太高。

4.3 计算方面的考量

现在我们讨论改进计算复杂度的方法。主要瓶颈在于链接结构的更新，其朴素实现的复杂度为 $O(N \cdot E)$，其中 E 是实例化实体的数量。对于取值域 D_a 较大的情况，实体属性的更新也可能存在问题，因为第 a 个属性的复杂度为 $O(E \cdot |\mathrm{D}_a|)$。

我们可以利用失真模型施加的约束，降低链接结构更新的计算复杂度。仔细检查与记录 i 关联的实体的更新（见附录 A.3）可发现，有些实体可以立即排除在考虑范围之外。具体来说，只有属性与相应的未失真记录属性匹配的实体（x_{ia} 其中 $z_{ia} = 0$）才可能与记录 i关联。为了高效查询这组实体，我们维护倒排索引，将属性值 $x \in \mathcal{D}_a$ 映射到以该值实例化的实体集合$\{e : \ x = y_{ea}\}$。只要失真程度相对较低，这种方法就比依次遍历所有实体高效得多。不过需要注意的是，该方法的关键前提是不折叠失真指示变量。

为了改进 co　　实体属性更新的复杂度，我们可以对距离度量设置截断值

[2]在本节陈述时间复杂度时，我们假设可以在 $\Theta(C)$ 时间内抽取一个分类随机变量，其中 C 是类别数。由 Vose (1991) 提出的算法满足这一约束。

。具体而言，对于属性 a，我们将“原始”距离度量 dist_a 替换为

$$\underline{\text{dist}}_a(y,x) = \begin{cases} \text{dist}_a(y,x), & \text{if } \text{dist}_a(y,x) \le d_a^{(\text{cut})}, \\ \infty, & \text{otherwise}, \end{cases}$$

其中，$d_a^{(\text{cut})} \in (0,\infty)$ 是一个可配置的截断值。该近似消除了考虑从实体属性 y 到记录属性 x 的不太可能出现的失真的需要，即 $\text{dist}(y,x) > d_a^{(\text{cut})}$。它在记录链接文献（Christen，2012b）中发挥着与分块类似的作用，并且类似于 Marchant 等人（2021）提出的一种方法。为了利用这一近似，我们构建了能够高效回答范围查询的索引——每个属性对应一个索引。第 a 个属性的索引接收查询值 $x \in \text{D}a$，并返回低于截断值的实体属性值集合：$\{y \in \text{D}a : \text{dist}_a(y,x) \le d_a^{(\text{cut})}\}$。

5 模型比较

我们使用真实链接结构已知的数据集，对实体解析模型开展实证研究。第 5.1 节介绍了研究中使用的数据集，这些数据集的设计动机来自隐私场景和非隐私场景下的实体解析应用。我们在第 5.2节中说明如何评估我们的模型（以及基线模型），方法是计算相关指标，以评估后验预测与真实链接结构的一致程度。第 5.3 节通过改变失真模型和链接结构的先验分布，评估我们的建模贡献所产生的影响。第5.4 节将我们的实体解析模型与Sadinle (2014年）和Steorts (2015年）提出的基线方法进行比较。最后，第 5.5节总结了附录 C中介绍的受控模拟研究。

5.1 数据集

我们研究隐私场景和非隐私场景下的实体解析，这两种场景都是从业者会遇到的。我们在研究中使用的数据集汇总于表1。

隐私场景。在此场景中，从业者可以访问去标识化数据，其中姓名、地址、电话号码等敏感属性均已移除。这会使实体解析颇具挑战性，因为剩余的非敏感属性可能无法提供足够的信息来确定记录的身份。为研究此场景下的实体解析，我们使用从全国长期护理调查（Manton，2010）中提取的数据，并将其称为nltcs。

我们提取了美国阿拉巴马州该调查1982年、1989年和1994年各轮次中的去标识化受访者记录。我们使用所有可用属性进行实体解析，包括出生日期（出生年份、出生月份、出生日期）、登记办公室（登记办公室）和性别（性别）。由于数据经过精心整理，可能出现的唯一失真是将有效属性值替换为另一个有效属性值。因此，我们采用常数距离函数，将这些属性建模为分类变量。

非隐私场景。 在此设置中，实践者可以访问包含敏感属性（如姓名）的数据。我们假设无法获得社会保障号码等唯一标识符，否则

数据集	设置	实体类型	记录数（N）	实体数
nltcs	私有	人员	5,359	3,307
RLdata	非私有	人员	10,000	9,000
cora	非私有	引用	1,295	125
rest	非私有	餐馆	864	752

表 1：数据集摘要。

实体解析就会变得轻而易举。获取带有真实标注、适用于非隐私场景的真实调查数据颇具挑战，因此我们使用了实体解析文献中的三个公开数据集。尽管这些数据集涉及其他取值域，但它们呈现出的特征与真实调查数据中可能遇到的特征相符。具体而言，我们观察到存在多个类似“姓名”的属性，以及不同程度的变化和失真。下面简要介绍每个数据集及用于实体解析的属性：

- RLdata 是一个合成人员数据集，其中 10% 的记录是含有随机错误的重复记录（Sariyar 和 Borg，2010）。我们使用归一化莱文斯坦距离度量对姓名属性（`fname_c1` 和`lname_c1`）进行建模。出生日期相关属性——`bd`，`bm` 和`by` ——则建模为采用常数距离度量的分类属性。[3]

- cora 是一个计算机科学引文记录集合，托管于 RIDDLE 仓库（Bilenko，2003）。在我们考虑的所有数据集中，它是“最杂乱”的一个，因为这些记录提取自多个采用不同引文格式的在线数据源。作为预处理步骤，我们将带连字符的词拆开并移除标点符号。我们还更正了若干错误的真实标注标签。标题、发表场所和作者属性通常包含多个词，且存在语义和字符级变化，因此采用附录 B中描述的混合词元／编辑距离度量进行建模。年份属性则使用归一化莱文斯坦距离进行建模。

- rest 是一个餐厅记录集合，收录了来自 Fodor 和 Zagat 餐厅指南的记录，并托管于 RIDDLE 仓库（Bilenko，2003）。它不像 cora 那么“脏”，因为数据源较少，彼此之间的变化也较小。我们采用了与 cora 相同的预处理步骤。名称和地址属性通常包含多个词，因此采用与 cora 相同的混合距离度量进行建模。城市和类型（菜系）属性则作为分类变量建模，并使用常数距离度量。

5.2 模型评估

ˆ 我们通过将推断出的链接结构 Λ 与真实的链接结构 Λ进行比较，在数据集上评估实体解析模型。回顾一下，$\Lambda = (\lambda_1, \ldots, \lambda_N)$ 指定了数据集中每条ˆ 记录 i 所对应的实体 λ_i。可以使用成对精确率ˆ 和召回率来衡量 Λ 与 Λ 之间的一致程度。成对精确率 是指在 Λ 中链接的记录对中，同时也被链接

[3]这是一个在文献中广泛使用的基准数据集。

在 Λ 中的比例：

$$\mathrm{Pr}(\hat{\Lambda},\Lambda)=\frac{\sum_{i\neq j=1}^{N}\mathbb{1}[\hat{\lambda}_i=\hat{\lambda}_j]\mathbb{1}[\lambda_i=\lambda_j]}{\sum_{i\neq j=1}^{N}\mathbb{1}[\hat{\lambda}_i=\hat{\lambda}_j]}. \tag{12}$$

其取值范围为 0 到 1，数值越大表示假阳性错误越少。成对˜召回率 是指在 Λ 中链接的记录对同时也在 Λ 中链接的比例：

$$\mathrm{Re}(\hat{\Lambda},\Lambda)=\frac{\sum_{i\neq j=1}^{N}\mathbb{1}[\hat{\lambda}_i=\hat{\lambda}_j]\mathbb{1}[\lambda_i=\lambda_j]}{\sum_{i\neq j=1}^{N}\mathbb{1}[\lambda_i=\lambda_j]}. \tag{13}$$

其取值范围为 0 到 1，数值越大表示假阴性错误越少。在实际应用中，通常很难同时获得高精确率和高召回率——必须根据特定应用中哪类错误代价更高来进行权衡。如果精确率和召回率同等˜重要，则可以使用成对 *F1* 分数 来衡量 Λ 与 Λ 之间的一致程度，该分数是精确率和召回率的调和平均数：

$$\mathrm{F1}(\hat{\Lambda},\Lambda)=2\frac{\mathrm{Pr}(\hat{\Lambda},\Lambda)\cdot\mathrm{Re}(\hat{\Lambda},\Lambda)}{\mathrm{Pr}(\hat{\Lambda},\Lambda)+\mathrm{Re}(\hat{\Lambda},\Lambda)}. \tag{14}$$

˜由于我们研究中的模型是贝叶斯模型，因此推断得到的（后验）链接结构 Λ 是一个随机变量，方程（12）–（14）中的指标也可以视为随机变量。我们使用通过 MCMC 生成的样本来估计这些指标在后验分布下的分布。这样，我们便能够在评估中考虑后验不确定性。对于每个指标，我们使用中位数报告其点估计，并同时给出 95% 等尾可信区间。有关每个模型的 MCMC 实现和配置的更多细节见附录 E，MCMC诊断见附录 G。

5.3 链接结构先验与失真模型研究

在本节中，我们研究第 3.1 节提出的两项建模贡献的影响——Ewens-Pitman（EP）链接结构先验和改进的失真模型。我们的目标是以 blink 模型为基准，确定每项建模贡献单独产生的影响。我们在此总结结果，完整结果涵盖链接结构先验与失真模型的八种组合，详见附录 D 。

链接结构先验。 我们考虑链接结构的四种先验分布，它们对应不同的 EP 参数机制（参见第 2.2 节）：

1. PY：Pitman-Yor 制度，取 $\sigma\in(0,1)$，并对 σ,α 设置超先验，详见方程 (3)。
2. Ewens：Ewens 制度，取 $\sigma=0$，并对 α 设置超先验，详见方程 (3)。
3. GenCoupon：带有 $\sigma=-\kappa<0$ 且对 κ、m，详见方程（2）。

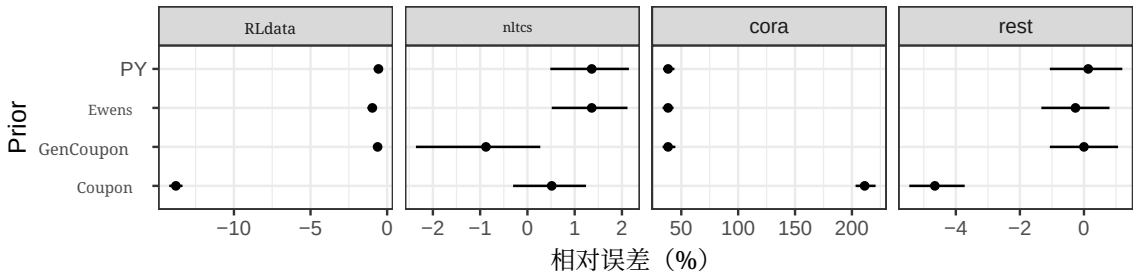


图 3：所有数据集和链接结构先验下，预测实体数量的后验相对误差。对于 cora，观察到链接不足，这可能是由于存在显著噪声，而失真模型未能充分捕捉这些噪声。

4. Coupon：使用优惠券收集者划分，并带有 $\kappa \to \infty$和 $m = N$。

前三种先验具有灵活性，因为它们为 EP 参数设置了超先验。最后一种先验是 GenCoupon 的一个特例，其中 EP 参数是固定的。它用于 Tancredi和Liseo (2011年)、Steorts (2015年) 以及 Steorts 等人 (2016年) 的模型中，并在此处作为基线。

表 2 展示了四种链接结构先验下的实体解析评估指标，其中假设模型的其余部分遵循第 3.1 节中的设定。图 3 从另一角度展示了实体解析准确率，其中绘出了推断所得实体数量的相对误差。两项结果均表明，为 EP 参数设置超先验（如 PY、Ewens 和 GenCoupon 所采用的做法）具有优势。与 Coupon 相比，这些链接结构先验取得了更高的 F1 分数；在 Coupon 中，EP 参数是固定的。图 S8(附录 D) 与这一发现一致，表明使用超先验时，每个数据集推断出的 EP 参数值差异很大。另一个有趣的现象是，PY、Ewens 和 GenCoupon 的实体解析准确率相近。起初这出乎意料，因为已知这三种参数区间具有不同的渐近行为（见方程 1）。这表明，这三种区间或许都足够灵活，能够对我们在此考察的数据集的链接结构建模。值得进一步探究的是，这些观察结果是否也适用于大得多的数据集；在这类数据集中，三种区间的渐近行为会更加明显。

失真模型。 我们将提出的失真模型（在第 3.1 节后半部分中指定）与 Steorts (2015年) 提出的失真模型进行比较。为简洁起见，我们将自己的失真模型称为 我们的方法，将 Steorts(2015年) 提出的模型称为 blink。此处报告的是 GenCoupon 链接结构先验下的结果；其他链接结构先验的结果见附录 D，且呈现出相似的趋势。

图 4 绘出了两种失真模型下每个属性推断出的失真程度。结果显示，blink 倾向于鼓励较高的失真，尤其是对于采用非常数距离度量的属性。[4] 例如，预测 RLdata 中的 fname_c1 和 lname_c1 属性

[4]采用非常数距离度量建模的属性包括：cora中的所有属性、rest中的名称和地址，以及RLdata中的 fname_c1 和 lname_c1。

数据集	EP制度	评估指标		
		精确率	召回率	F1 分数
RLdata	PY	0.896 (0.879, 0.917)	0.961 (0.952, 0.972)	0.928 (0.918, 0.939)
	Ewens	0.870 (0.853, 0.893)	0.970 (0.961, 0.978)	0.917 (0.908, 0.931)
	GenCoupon	0.903 (0.886, 0.920)	0.966 (0.955, 0.975)	0.933 (0.923, 0.941)
	Coupon	0.402 (0.396, 0.410)	0.987 (0.982, 0.993)	0.572 (0.565, 0.580)
nltcs	PY	0.921 (0.908, 0.933)	0.924 (0.915, 0.934)	0.923 (0.915, 0.930)
	Ewens	0.921 (0.910, 0.932)	0.925 (0.915, 0.934)	0.923 (0.916, 0.930)
	GenCoupon	0.902 (0.879, 0.918)	0.935 (0.926, 0.944)	0.918 (0.906, 0.927)
	Coupon	0.919 (0.908, 0.930)	0.926 (0.916, 0.935)	0.923 (0.915, 0.930)
cora	PY	0.971 (0.963, 0.979)	0.671 (0.647, 0.696)	0.794 (0.776, 0.813)
	Ewens	0.974 (0.965, 0.981)	0.673 (0.645, 0.697)	0.796 (0.775, 0.813)
	GenCoupon	0.973 (0.965, 0.981)	0.657 (0.632, 0.683)	0.784 (0.766, 0.804)
	Coupon	0.978 (0.971, 0.986)	0.173 (0.164, 0.181)	0.294 (0.281, 0.306)
rest	PY	0.770 (0.735, 0.824)	0.812 (0.759, 0.884)	0.795 (0.755, 0.828)
	Ewens	0.770 (0.711, 0.823)	0.830 (0.781, 0.875)	0.798 (0.760, 0.838)
	GenCoupon	0.794 (0.742, 0.850)	0.821 (0.777, 0.875)	0.807 (0.773, 0.849)
	Coupon	0.637 (0.602, 0.674)	0.911 (0.893, 0.938)	0.750 (0.722, 0.781)

表2：在对应于不同 Ewens-Pitman（EP）参数区间的四种链接结构先验分布下，我们模型的后验评估指标。每项评估指标均报告基于中位数的点估计及其 95% 等尾可信区间。在允许 EP 参数变化的三种参数区间（PY、Ewens 和GenCoupon）中，模型表现相近。在 RLdata 和cora数据集上，Coupon 参数区间的表现则显著下降。

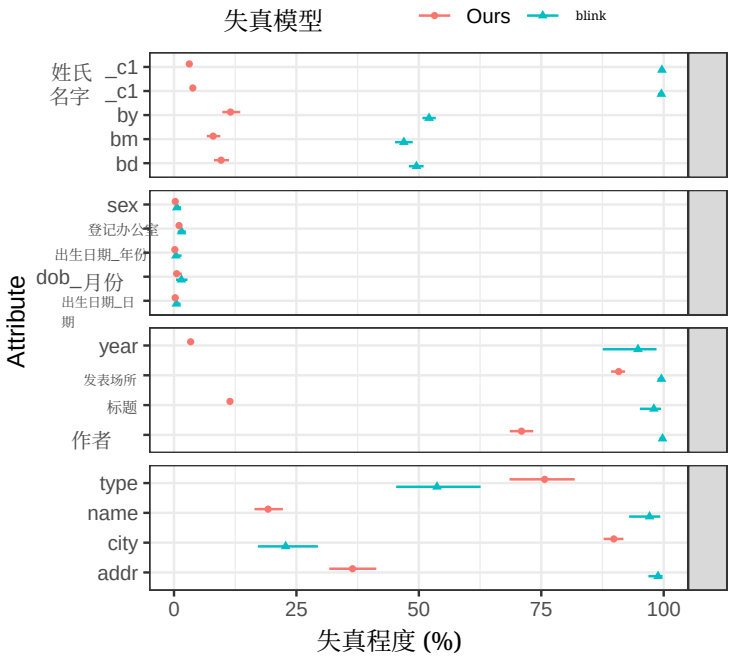


图 4：两种失真模型下后验属性级失真的比较：我们的方法（最上方的红色区间）和 blink（最下方的青绿色区间）。blink 失真模型倾向于偏好较高的失真程度——某些情况下接近 100%—— 这与预期不符。

在 blink 失真模型下几乎达到 100% 的失真程度，这与对该数据集的预期不符。我们的失真模型似乎没有这个问题，因为它要求实体属性与记录属性之间存在不一致，才会将其判定为“失真”。由于高失真会使可靠链接更加困难，我们预计在实际应用中，我们的失真模型可能表现更好。事实上，在完整结果中（见附录S7 D），它在精确率和召回率之间实现了更好的平衡。

总结。 我们回到本节最初的目标，总结本研究中的发现。首先，我们发现，由于采用了超先验，我们提出的链接结构先验通常更加稳健。此外，我们的推断对 EP 参数机制相对不敏感，这可能是因为数据集规模相对较小。将 blink 失真模型与我们提出的链接结构先验结合使用时，也呈现出这一特性。其次，在考察失真模型的表现时（blink 与我们提出的失真模型），我们发现，我们的方法能够预测更合理的扰动率，并提升以 F1 分数衡量的链接准确率。因此，根据本研究的结果，对于与我们所研究的数据集类似的数据集，我们建议今后采用我们的失真模型和链接结构先验。不过，我们强调，还需要进一步探索，才能针对其他数据集提出更具普适性的建议。

5.4 与基线模型的比较

本节中，我们研究实体解析模型与 Steorts (2015年) 和 (Sadinle， 2014年) 提出的模型相比表现如何。由 blink 模型 Steorts (2015年) 提出，是一个值得考虑的自然基线，因为它为我们的模型提供了灵感。与我们的模型相比，blink 在设计上贝叶斯特征较弱，因为许多参数是根据经验或任意设定的。我们的模型和blink都是实体解析的直接建模方法示例——也就是说，它们对观测记录的生成方式进行建模，并将链接结构作为潜变量纳入模型。相比之下，Sadinle (2014年) 提出的模型（我们称之为Sadinle）采用基于比较的实体解析方法。它不直接对数据建模，而是对记录对之间的属性级比较进行建模，并将记录对之间是否存在链接作为潜变量纳入模型。 Sadinle (2018) 从方法论视角比较了直接建模方法与基于比较的方法，但据我们所知，文献中尚无实证比较。本节的目标是首次在多种数据集上进行比较；我们并不强烈主张这些结果能够推广至所有应用或数据集。

为了尽可能公平地进行比较，我们在自己的模型、blink和Sadinle中使用相同的距离函数来建模失真。例如，如果我们在自己的模型和blink中使用编辑距离对名称属性的失真进行建模，那么我们也会在Sadinle中使用编辑距离来比较相同的名称属性。我们设置自己模型和 blink （见第4.3节）的距离截断值，使其与Sadinle所采用的阻塞设计相匹配。有关实验设置的更多信息见附录E。

三个模型的实体解析评估指标见表3 。为简便起见，我们仅给出采用GenCoupon 先验分布的模型结果，在表中记作我们的方法。我们的模型在四个数据集上均取得最高（或并列最高）的 F1 分数。我们认为，blink 表现较差，可能是因为它采用了主观（缺乏灵活性）的先验分布和失真模型，后者倾向于偏好高失真和过度链接。Sadinle 在非隐私场景（ RLdata、cora、rest）中的 F1 分数排名第二，在隐私场景（nltcs）中的 F1 分数则最低。nltcs 上的表现较差，部分原因可能与阻塞方案有关：该方案较为宽松，使模型更容易出现过度链接。另一个重要因素是Sadinle对先验分布中 m概率截断点的敏感性。我们在附录F 中对截断点进行了粗粒度调参，但细粒度调参可能会进一步提升性能。

摘要。 本研究提供了证据，表明与blink 和Sadinle相比，我们的模型在精确率和召回率之间实现了更好的平衡。我们强调，我们的结果基于四个数据集——还需进一步实验，以确定这些结果是否适用于其他数据集和应用。我们推测，我们的模型表现更佳，主要是因为加入先验分布和超先验后灵活性有所提升；这可以视为执行模型选择。

数据集	模型	评估指标		
		精确率	召回率	F1 分数
RLdata	Ours	0.917 (0.902, 0.932)	0.966 (0.953, 0.973)	0.934 (0.922, 0.942)
	blink	0.336 (0.327, 0.344)	0.992 (0.988, 0.996)	0.502 (0.492, 0.511)
	Sadinle	0.534 (0.524, 0.546)	0.964 (0.962, 0.966)	0.687 (0.679, 0.697)
nltcs	Ours	0.901 (0.878, 0.917)	0.934 (0.923, 0.943)	0.916 (0.903, 0.925)
	blink	0.904 (0.890, 0.918)	0.918 (0.904, 0.925)	0.910 (0.903, 0.918)
	Sadinle	0.312 (0.304, 0.319)	0.975 (0.969, 0.979)	0.473 (0.464, 0.480)
cora	Ours	0.973 (0.965, 0.980)	0.657 (0.627, 0.683)	0.784 (0.764, 0.803)
	blink	0.978 (0.970, 0.985)	0.219 (0.207, 0.234)	0.358 (0.341, 0.378)
	Sadinle	0.982 (0.981, 0.983)	0.359 (0.357, 0.362)	0.526 (0.524, 0.529)
rest	Ours	0.795 (0.749, 0.836)	0.830 (0.776, 0.871)	0.811 (0.774, 0.848)
	blink	0.635 (0.586, 0.671)	0.920 (0.893, 0.946)	0.751 (0.713, 0.780)
	Sadinle	0.993 (0.985, 1.000)	0.603 (0.598, 0.607)	0.750 (0.744, 0.756)

表3：我们的模型与两种基线方法的后验性能： blink (Steorts，2015年）和 Sadinle(Sadinle，2014年）。每项评估指标均基于中位数报告点估计，并给出95% 等尾可信区间。对于所有数据集，在可信区间内，我们模型的F1 分数均为最高（或并列最高）。

5.5 受控模拟研究

我们开展了一项模拟研究，在受控条件下评估模型，并改变数据集规模、失真程度和重复程度。由于篇幅所限，我们在此概述这项研究——完整细节见附录C。我们设计了一个家庭调查数据集模拟器，其中收集家庭内个人的响应。由于同一家庭中个人的属性彼此相关——例如，地址相同，家庭成员可能拥有相同的姓氏，个人年龄也可能相关——模拟数据遵循的生成过程比我们的实体解析模型更复杂。这是有意为之，因为这样我们便能在模型设定错误 的情况下，在更真实的设置中评估模型。该数据集模拟器还纳入了一个对我们的模型而言设定错误的记录生成过程。我们的数据集模拟器并非从总体中抽样个人，而是遍历所有个人，随机决定是否纳入某个人，以及若纳入，要为其创建多少条失真记录。

我们使用模型对 16 个模拟数据集进行实体解析，并以 blink 和 Sadinle模型作为基线方法。下面我们根据三个因素汇总结果：

- 重复程度。重复程度对我们模型的性能影响甚微。blink 在重复程度中等至高时表现良好，但在重复程度较低时会出现严重的过度链接。随着重复程度变化，Sadinle 的性能似乎没有一致的趋势；在所有情况下，其 F1 分数都低于我们的模型和blink 。

- 失真程度。我们发现，失真程度越高的数据集越难以进行链接。具体而言，与低失真数据集相比，我们的模型和blink 的召回率下降了约 10 个百分点。Sadinle的召回率降幅更大，为 15–20 个百分点。

- 数据集大小。 我们发现，对于两种数据集大小（1000 条和 10000 条记录），我们的模型表现相近。blink 在大多数情景下的表现也相近；不过，当重复程度较低时，我们观察到它在较大数据集上的精确率有所下降。Sadinle 在较大数据集上的精确率明显较低，但召回率相对稳定。

总之，模拟研究表明，我们的模型在所有测试情景中表现最为稳定。blink 也具有竞争力，但在低重复情景下表现较差。Sadinle 在重复程度不可忽略时 F1 分数最低，而在失真程度较低时则具有一定竞争力。

6 讨论

在本节中，我们总结本文的贡献，并讨论未来工作。我们提出了一种用于实体解析的贝叶斯模型，解决了以往研究的局限性（Steorts 等人，2016年；Steorts，2015年；Marchant 等人，2021）。我们的模型可视为执行图模型实体解析，即将观测到的记录聚类到（未观测到的）潜在实体中。为提高以往研究的可扩展性，我们设计了一个部分塌缩吉布斯采样器，并对其实现进行了优化，使其能够处理约 10,000 条记录的数据集。这使我们能够将模型与 Steorts （2015年）和Sadinle （2014年）提出的模型进行比较；此前，这类比较仅能在玩具规模的数据集上进行。我们将模型与真实数据集和合成数据集进行了比较，并开展了受控模拟研究。我们发现，对于本文考察的数据集，我们的模型对超参数变化的敏感度往往低于竞争模型。要想超越本文考察的数据集和模拟，得出更普遍的结论，还需要进一步分析。

未来工作有许多值得探索的方向。首先，值得研究如何扩展我们提出的模型以及Sadinle (2014年)提出的模型。实现这一点的方法包括设计并行／分布式推断算法、探索更高效的马尔可夫链蒙特卡罗算法，或采用分块技术。另一个值得关注的方向是探索更多样化的数据集，以了解各方法在实际应用中的优势与不足。尽管我们基于四个数据集和一项模拟研究提出了建议，进行更多比较有助于减轻数据集选择方面的偏差，并为用户提供指引。最后，未来工作可考虑微聚类先验分布（Miller 等人，2015年），评估其相较于本文所考察的无限可交换链接结构先验分布的有效性。这需要修改采样方案，并选择一类合适的微聚类先验分布。

致谢

本研究得到美国国家科学基金会 [CAREER-1652431], 阿尔弗雷德·斯隆基金会、澳大利亚研究理事会 [DP220102269] 以及澳大利亚政府的支持

研究培训计划奖学金。

参考文献

Bernardo, J. M. 和 Smith, A. F.（2009），贝叶斯理论，第405卷，John Wiley & Sons。Bilenko, M.（2003），“重复检测、记录链接与身份不确定性：数据集”，http://www.cs.utexas.edu/users/ml/riddle/data.html，访问日期：2019-10-01。Binette, O. 和 Steorts, R. C.（2022），“（几乎）所有的实体解析”，《科学进展》，8，eabi8021。Brooks, S.、Gelman, A.、Jones, G. 和 Meng, X.-L.（2011），马尔可夫链蒙特卡洛手册，Chapman and Hall/CRC，第1版。Chib, S. 和 Greenberg, E.（1995），“理解 Metropolis-Hastings算法”，《美国统计学家》，49，327–335。Christen, P.（2012a），数据匹配：记录链接、实体解析与重复检测的概念与技术，以数据为中心的系统与应用，柏林、海德堡：Springer-Verlag。—（2012b），“可扩展记录链接与去重的索引技术综述”，*IEEE*知识与数据工程汇刊，24，1537–1555。Christophides, V.、Efthymiou, V.、Palpanas, T.、Papadakis, G. 和 Stefanidis, K.（2020），“大数据端到端实体解析概述”，*ACM*计算综述，53。Copas, J. B. 和 Hilton, F. J.（1990），“记录链接：计算机记录匹配的统计模型”，英国皇家统计学会会刊：*A*辑（社会统计学），153，287–320。Crouse, D. F.（2016），“二维矩形指派算法的实现”，*IEEE*航空航天与电子系统汇刊，52，1679–1696。Damlen, P.、Wakefield, J. 和 Walker, S.（1999），“使用辅助变量对贝叶斯非共轭层次模型进行吉布斯抽样”，英国皇家统计学会会刊：*B*辑（统计方法学），61，331–344。Doan, A.、Halevy, A. 和 Ives, Z.（2012），数据集成原理，波士顿：Morgan Kaufmann。Dunn, H. L.（1946），“记录链接”，《美国公共卫生杂志》和《国民健康》，1412–1416。Eddelbuettel, D. 和 François, R.（2011），“Rcpp：无缝集成 R 和 C++”，《统计软件杂志》，40，1–18。Elmagarmid, A. K.、Ipeirotis, P. G. 和 Verykios, V. S.（2007），“重复记录检测：综述”，*IEEE*知识与数据工程汇刊，19，1–16。

Enamorado, T., Fifield, B., and Imai, K. (2019), “使用概率模型辅助合并大规模行政记录”, 《美国政治科学评论》, 113, 353–371. Fellegi, I. P. and Sunter, A. B. (1969), “记录链接理论”, 《美国统计学会杂志》, 64, 1183–1210. Gamerman, D. and Lopes, H. F. (2006), 马尔可夫链蒙特卡罗：用于贝叶斯推断的随机模拟, Chapman and Hall/CRC, 第2版. Getoor, L. and Machanavajjhala, A. (2012), “实体解析：理论、实践与开放挑战”, *Proc. VLDB Endow.*, 5, 2018–2019. Geweke, J. (1992), “评估基于抽样的方法计算后验矩的准确性”, 贝叶斯统计, 4, 641–649. Gilks, W. R. and Wild, P. (1992), “用于吉布斯抽样的自适应拒绝抽样”, 《皇家统计学会杂志：*C*辑（应用统计）》, 41, 337–348. Goldstein, H., Harron, K., and Cortina-Borja, M. (2017), “记录链接的尺度化方法”, 《医学统计学》, 36, 2514–2521. Ilyas, I. F. and Chu, X. (2019), 数据清洗, 美国纽约州纽约市：计算机协会. Irony, T. Z. and Singpurwalla, N. D. (1997), “非信息先验并不存在：与José M. Bernardo的对话”, 《统计规划与推断杂志》, 65, 159–177. Kaplan, A., Betancourt, B., and Steorts, R. C. (2022), “使用链接数据进行恰当推断的实用方法”, 《美国统计学家》, 1–10. Kingman, J. F. C. (1978), “群体遗传学中的随机划分”, 《伦敦皇家学会会刊：*A*辑，数学与物理科学》, 361, 1–20. Manton, K. G. (2010), “全国长期护理调查：1982年、1984年、1989年、1994年、1999年和2004年”, . Marchant, N. G., Kaplan, A., Elazar, D. N., Rubinstein, B. I., and Steorts, R. C. (2021), “d-blink：分布式端到端贝叶斯实体解析”, 《计算与图形统计学杂志》, 30, 406–421. Miller, J. W., Betancourt, B., Zaidi, A., Wallach, H., and Steorts, R. C. (2015年), “微簇：当簇大小随数据集大小呈次线性增长时”, 载于贝叶斯非参数方法：下一代研讨会. Monge, A. E. and Elkan, C. P. (1996), “字段匹配问题：算法与应用”, 载于第二届知识发现与数据挖掘国际会议论文集，编者 Simoudis, E., Han, J., and Fayyad, U., AAAI出版社, KDD’96, 第267–270页. Naumann, F. and Herschel, M. (2010), 重复检测导论, Morgan and Claypool出版社.

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, 249–265. Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959), “生命记录自动链接：计算机可用于从常规记录档案中提取家庭的‘随访’统计数据,”《科学》, 130, 954–959. Papadakis, G., Ioannou, E., Thanos, E., and Palpanas, T. (2021), “实体解析的四代发展,”数据管理综合讲座, 16, 1–170. Pitman, J. (2006), 可交换随机划分, 柏林、海德堡：施普林格柏林海德堡, pp. 37–53. Pitman, J. and Yor, M. (1997), “由稳定子ordinator导出的二参数泊松-狄利克雷分布,”《概率年刊》, 25, 855–900. Sadinle, M. (2014), “使用贝叶斯划分方法检测凶杀案件登记册中的重复记录,”《应用统计年刊》, 8, 2404–2434. — (2017), “用于记录链接的二分匹配的贝叶斯估计,”《美国统计学会杂志》, 112, 600–612. — (2018), “将记录链接不确定性通过贝叶斯方法传播至人权侵害总体规模估计,”《应用统计年刊》, 12, 1013–1038. Sadinle, M. and Fienberg, S. E. (2013), “适用于多记录链接的广义 Fellegi–Sunter 框架及其在凶杀记录系统中的应用,”《美国统计学会杂志》, 108, 385–397. Sariyar, M. and Borg, A. (2010), “RecordLinkage 软件包：检测数据中的错误,”《*R* 期刊》, 2, 61–67. Steorts, R. C. (2015), “采用经验驱动先验的实体解析,”《贝叶斯分析》, 10, 849–875. Steorts, R. C., Hall, R., and Fienberg, S. E. (2016), “图结构记录链接与去重的贝叶斯方法,”《美国统计学会杂志》, 111, 1660–1672. Steorts, R. C., Tancredi, A., and Liseo, B. (2018), “具有精确误差传播的广义贝叶斯记录链接与回归,” 载于统计数据库中的隐私, 编者 Domingo- Ferrer, J. 和 Montes, F., 尚贝里：施普林格国际出版, pp. 297–313. Syversveen, A. R. (1998), “非信息性贝叶斯先验：对其解释以及构造与应用问题的探讨,”统计学预印本, 3, 1–11. Tancredi, A. and Liseo, B. (2011), “记录链接与总体规模问题的层次贝叶斯方法,”《应用统计年刊》, 5, 1553–1585. — (2015), “链接数据的回归分析：问题与可能的解决方案,”《统计学》, 75, 1 –

9 35.

Tancredi, A., Steorts, R., and Liseo, B. (2020), “去重与总体规模估计的统一框架（附讨论）,”《贝叶斯分析》, 15, 633–682. Teh, Y. W. (2006), “插值 Kneser-Ney 的贝叶斯解释,” 技术报告，新加坡国立大学。Tiao, G. C. and Box, G. E. (1973), “ 关于 “ 贝叶斯 ” 估计量的一些评论,” 《美国统计学家》, 27, 12–14. 美国人口普查局 (2016), “当前人口调查, 2016 年度社会与经济补充调查,” https://www.census.gov/data/tables/2016/demo/families/cps-2016.html，访问日期: 2022 年 7 月 1 日。van Dyk, D. A. and Park, T. (2008), “部分折叠吉布斯采样器,” 《美国统计学会杂志》, 103, 790–796. Vose, M. D. (1991), “生成服从给定分布的随机数的线性算法,”*IEEE* 软件工程汇刊,17,972–975. Winkler, W. E. (2006), “记录链接概述与当前研究方向,” 技术报告，统计学 #2006-2，美国人口普查局。

A 吉布斯更新

在本附录中，我们推导了用于对第3节中介绍的实体解析模型进行近似推断的部分塌缩吉布斯采样器的更新。由于所提出模型的非共轭性，其中一些更新并不简单。

A.1 失真概率的更新

本节给出失真概率 θ_{sa}（来源 s 和属性 a）的更新。由于存在失真倾向变量 ω_{ia}，贝塔先验不再具有共轭性，因此此更新较为复杂。为解决这一问题，我们引入以下辅助变量：

$$q_{ia} \mid \omega_{ia} \sim \text{Bernoulli}(\omega_{ia}) \qquad \forall i, a$$

并按如下方式修改失真指示变量的条件分布：

$$z_{ia} \mid \theta_{\varsigma_i a}, q_{ia} \sim \text{Bernoulli}(\theta_{\varsigma_i a} q_{ia}) \qquad \forall i, a.$$

容易证明，将辅助变量边缘化后，即可恢复方程（5）中的原始模型。

注意，后验分布中涉及 q_{ia} 的部分为

$$(\theta_{\varsigma_i a} q_{ia})^{z_{ia}} (1 - \theta_{\varsigma_i a} q_{ia})^{1-z_{ia}} \omega_{ia}^{q_{ia}} (1-\omega_{ia})^{1-q_{ia}} = \left[\theta_{\varsigma_i a}^{z_{ia}} (1-\theta_{\varsigma_i a})^{1-z_{ia}} \omega_{ia}\right]^{q_{ia}} [1-\omega_{ia}]^{1-q_{ia}}.$$

因此，在给定其他变量的条件下，q_{ia} 的分布为：

$$q_{ia} \mid \omega_{ia}, \theta_{\varsigma_i a}, z_{ia} \sim \text{Bernoulli}\left(\frac{\omega_{ia} \theta_{\varsigma_i a}^{z_{ia}} (1-\theta_{\varsigma_i a})^{1-z_{ia}}}{\omega_{ia} \theta_{\varsigma_i a}^{z_{ia}} (1-\theta_{\varsigma_i a})^{1-z_{ia}} + 1 - \omega_{ia}}\right) \qquad \forall i, a. \tag{S15}$$

接下来，注意后验分布中涉及 θ_{sa} 的部分为

$$\theta_{sa}^{\beta_{sa}^{(0)}-1} (1-\theta_{sa})^{\beta_{sa}^{(1)}-1} \prod_{i:\varsigma_i=s} (\theta_{sa} q_{ia})^{z_{ia}} (1-\theta_{sa} q_{ia})^{1-z_{ia}}.$$

因此，在给定其他变量的条件下，θ_{sa} 的分布为：

$$\theta_{sa} \mid \boldsymbol{Q}, \boldsymbol{Z}, \boldsymbol{S} \sim \text{Beta}\left(\beta_{sa}^{(0)} + \sum_{i:\varsigma_i=s} z_{ia},\ \beta_{sa}^{(1)} + \sum_{i:\varsigma_i=s} q_{ia}(1-z_{ia})\right) \qquad \forall s, a. \tag{S16}$$

同样不难看出，给定其他变量时，z_{ia} 的分布是点质量分布。具体而言，我们有

$$z_{ia} \mid x_{ia}, \lambda_i, \boldsymbol{Y} = \begin{cases} 1, & \text{if } x_{ia} \neq y_{\lambda_i a} \\ 0, & \text{otherwise} \end{cases} \tag{S17}$$

总之，要更新失真概率，首先应使用方程（$\{z_{ia}\}$ S17）计算失真指示变量。然后，给定其他变量，使用方程（$\{q_{ia}\}$S15）抽取辅助变量。最后，使用方程（$\{\theta_{sa}\}$S16）更新失真概率。引入辅助变量$\{q_{ia}\}$不会影响其他模型参数的更新。

A.2 实体属性的更新

本节给出实体属性的更新。更新实体属性 y_{ea} 时，我们将基分布 H_{ea} 和失真指示变量 $\boldsymbol{Z}$ 积分消去。

将 H_{ea} 积分消去后，涉及 y_{ea} 的后验分布因子如下：

$$
\begin{aligned}
P(y_{ea} \mid \boldsymbol{Z}, \boldsymbol{\Omega}, \boldsymbol{\Theta}, \boldsymbol{S}, G_a, \rho_a) &\propto P(y_{ea} \mid G_a) \\
&\quad \times \int \prod_{i:\lambda_i=e} P(x_{ia} \mid \theta_{\varsigma_i a}, \omega_{ia}, y_{ea}, H_{ea}) P(H_{ea} \mid y_{ea}, \rho_a)\, \mathrm{d}H_{ea} \\
&\propto P(y_{ea} \mid G_a) \prod_{\substack{i:\lambda_i=e \\ x_{ia}=y_{ea}}} (1-\theta_{\varsigma_i a}\omega_{ia}) \prod_{\substack{i:\lambda_i=e \\ x_{ia}\neq y_{ea}}} (\theta_{\varsigma_i a}\omega_{ia}) \\
&\quad \times \int \prod_{\substack{i:\lambda_i=e \\ x_{ia}\neq y_{ea}}} H_{ea}(x_{ia}) \frac{\prod_{v\in\mathcal{D}_a\setminus\{y_{ea}\}} H_{ea}(v)^{\rho_a\psi_a(v|y_{ea})-1}}{\mathrm{B}(\rho_a\psi_a(y_{ea}))}\, \mathrm{d}H_{ea} \\
&\propto G_a(y_{ea}) \prod_{\substack{i:\lambda_i=e \\ x_{ia}=y_{ea}}} (1-\theta_{\varsigma_i a}\omega_{ia}) \prod_{\substack{i:\lambda_i=e \\ x_{ia}\neq y_{ea}}} (\theta_{\varsigma_i a}\omega_{ia}) \\
&\quad \times \frac{\Gamma(\rho_a)}{\Gamma(\bar{n}_{ea}(y_{ea})+\rho_a)} \prod_{v\in\mathcal{V}_{ea}} \frac{\Gamma(n_{ea}(v)+\rho_a\psi_a(v \mid y_{ea}))}{\Gamma(\rho_a\psi_a(v \mid y_{ea}))},
\end{aligned}
$$

其中，$\mathrm{B}(\cdot)$ 是多元贝塔函数，$\mathrm{V}_{ea} = (\bigcup_{i:\ \lambda_i=e}\{x_{ia}\})\setminus\{y_{ea}\}$ 是与实体 e 的第 a 个属性相关联的失真记录值，$n_{ea}(v) = \sum_{i:\ \lambda_i=e} \mathbb{1}[x_{ia}=v]$ 是与实体 e 关联且第 a 个属性等于 v 的记录数量，$\bar{n}_{ea}(v) = \sum_{i:\ \lambda_i=e} \mathbb{1}[x_{ia}\neq v]$ 是与实体 e 关联且第 a 个属性不等于 v 的记录数量。

我们可以通过反复……将上述表达式改写为更便于计算的形式。

应用伽马函数的递推关系[5] 可得：

$$P(y_{ea} \mid \boldsymbol{Z}, \boldsymbol{\Omega}, \boldsymbol{\Theta}, \boldsymbol{S}, G_a, \rho_a) \propto G_a(y_{ea}) \frac{\prod_{v \in \mathcal{V}_{ea}} \prod_{j=1}^{n_{ea}(v)} \{j - 1 + \rho_a \psi_a(v \mid y_{ea})\}}{\prod_{j=1}^{\bar{n}_{ea}(y_{ea})} \{j - 1 + \rho_a\}} \times \prod_{\substack{i:\lambda_i = e \\ x_{ia} = y_{ea}}} (1 - \theta_{\varsigma_i a} \omega_{ia}) \prod_{\substack{i:\lambda_i = e \\ x_{ia} \neq y_{ea}}} (\theta_{\varsigma_i a} \omega_{ia}).$$

请注意，当应用距离阈值时，如第$\mathcal{D}_a$ 节所述，上述分布的支撑集可能只是完整取值域 4.3的一个子集。具体而言，可以证明其支撑集是以下集合的子集：

$$\bigcap_{i:\lambda_i = e} \{y \in \mathcal{D}_a : \mathrm{dist}_a(y, x_{ia}) \leq d_a^{(\mathrm{cut})}\}.$$

利用这一事实可以更高效地实现更新，因为无需在整个取值域 D_a上构造概率质量函数。

A.3 链接结构的更新

本节介绍链接结构的更新方法。更新链接结构时，我们采用Neal (2000)所述的基于瓮的方案。为此，我们只需跟踪总体中与记录关联的实体；任何未与记录关联的孤立实体均予以忽略。这一点很重要，因为在某些 Ewens-Pitman 参数区间内，总体的规模可能是无限的（当 $\sigma \geq 0$）。

要更新记录 i所关联的实体 λ_i，我们先移除当前链接，然后允许该记录加入某个剩余的实例化实体（该实体至少还关联一条其他记录），或实例化一个“新”实体。其条件分布形式如下：

$$P(\lambda_i = e \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\Lambda}_{-i}, \{\rho_a\}) \propto \begin{cases} C \frac{|e| - \sigma}{\alpha + N - 1} \prod_a \int P(x_{ia} \mid z_{ia}, y_{ea}, H_{ea}) \mathrm{d}\mathbb{H}_{-i,ea}, \\ \qquad \text{if } e \text{ is instantiated and } |e| > 0, \\ C \frac{\alpha + \sigma E}{\alpha + N - 1} \prod_a \sum_{y_{ea} \in \mathcal{D}_a} P(y_{ea} \mid G_a) \int P(x_{ia} \mid z_{ia}, y_{ea}, H_{ea}) \mathrm{d}\mathbb{H}_{0,ea}, \\ \qquad \text{if } e \text{ is “new”}, \end{cases} \tag{S18}$$

其中

- C 是归一化常数；
- $\boldsymbol{\Lambda}_{-i}$ =（λ_1、.....、 λ_{i-1}、 λ_{i+1}、.....、 λ_N）是除 i 之外所有记录对应的关联实体；

[5]反复应用伽马函数的递推关系可得

$$\Gamma(z) = \frac{\Gamma(z + n + 1)}{z(z + 1) \cdots (z + n)}$$

对于复数 z（不包括零和负整数）以及非负整数 n。

- $|e| = \sum_{i' \neq i} \mathbb{1}[\lambda_{i'} = e]$ 是与实体 e 关联的记录数量（不包括 i）；
- $E = \sum_{e' \neq e} \mathbb{1}[|e| > 0]$ 是至少包含一条关联记录的实例化实体数量；
- $\mathrm{H}_{0,\,ea}$ 是 H_{ea}的先验分布；并且
- $\mathrm{H}_{-i,\,ea}$ 是在给定观测到的失真记录属性 $x_{i'a}$ 后， H_{ea} 的后验分布，其中$i' \neq i$ 和 $\lambda_{i'} = e$（也以 y_{ea} 和 ρ_a为条件）。

回顾一下，在给定 y_{ea}和 ρ_a的条件下， H_{ea}的先验分布$\mathrm{H}_{0,\,ea}$服从狄利克雷分布$(\rho_a \psi_a(y_{ea}))$。由于当 $z_{ia} = 1$ 时， x_{ia}服从分类变量分布(H_{ea})（而当 $z_{ia} = 0$时， x_{ia}在 y_{ea} 处具有点质量），根据共轭性，后验分布$\mathrm{H}_{-i,\,ea}$ 也服从狄利克雷分布。具体而言，可以证明$\mathrm{H}_{-i,\,ea}$ 服从狄利克雷分布$(\boldsymbol{\alpha}_{-i,\,ea})$，其中

$$\alpha_{-i,ea}(v) = \rho_a \psi_a(v \mid y_{ea}) + \sum_{i' \neq i : \lambda_{i'} = e} z_{i'a} \mathbb{1}[x_{i'a} = v]$$

对于 $v \in \mathcal{D}_a \setminus \{y_{ea}\}$。因此，我们可以将关于 H 的方程 (S18)积分，ea简化如下：

$$\begin{aligned}
&\int P(x_{ia} \mid z_{ia}, y_{ea}, H_{ea}) \, d\mathbb{H}_{-i,ea} \\
&= \mathbb{1}[x_{ia} = y_{ea}]^{1 - z_{ia}} \int H_{ea}(x_{ia})^{z_{ia}} \Gamma(\rho_a) \prod_{v \in \mathcal{D}_a \setminus \{y_{ea}\}} \frac{H_{ea}(v)^{\rho_a \psi_a(v \mid y_{ea}) - 1}}{\Gamma(\rho_a \psi_a(v \mid y_{ea}))} \, \mathrm{d}H_{ea} \\
&= \begin{cases} \frac{\alpha_{-i,ea}(x_{ia})}{\sum_{v \in \mathcal{D}_a \setminus \{y_{ea}\}} \alpha_{-i,ea}(v)}, & z_{ia} = 1 \\ \mathbb{1}[x_{ia} = y_{ea}], & z_{ia} = 0 \end{cases}
\end{aligned} \tag{S19}$$

通过类似的论证，可以将关于 $\mathrm{H}_{0,\,ea}$ 的方程 (S18) 中的积分简化为：

$$\int P(x_{ia} \mid z_{ia}, y_{ea}, H_{ea}) \, d\mathbb{H}_{0,ea} = \begin{cases} \psi_a(x_{ia} \mid y_{ea}), & z_{ia} = 1, \\ \mathbb{1}[x_{ia} = y_{ea}], & z_{ia} = 0. \end{cases} \tag{S20}$$

将方程 (S19) 和 (S20) 代入 (S18)，得到：

$$P(\lambda_i = e \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\Lambda}_{-i}, \{\rho_a\}) \propto \begin{cases} C \frac{|e| - \sigma}{\alpha + N - 1} \prod_{a : z_{ia} = 1} \frac{\alpha_{-i,ea}(x_{ia})}{\sum_{v \in \mathcal{D}_a \setminus \{y_{ea}\}} \alpha_{-i,ea}(v)}, \\ \qquad \text{if } e \text{ is instantiated and } |e| > 0, \\ C \frac{\alpha + \sigma E}{\alpha + N - 1} \prod_{a : z_{ia} = 1} \sum_{y \in \mathcal{D}_a} G_a(y) \psi_a(x_{ia} \mid y)^{z_{ia}} \mathbb{1}[x_{ia} = y]^{1 - z_{ia}}, \\ \qquad \text{if } e \text{ is “new”}. \end{cases}$$

A.4 Ewens-Pitman 参数的更新

本节介绍 Ewens-Pitman 参数的更新。由于 Ewens-Pitman 参数 α 和 σ 的先验分布是非共轭的，因此无法直接进行吉布斯更新。因此，我们介绍需要引入辅助变量的可行更新方法。更新方法（以及先验分布）会因 σ的取值范围而异。Teh (2006) 针对 $0 \leq \sigma < 1$且 $\alpha > 0$的情形，提出了适用于贝塔/伽马先验的方案，概要见第A.4.1节。第A.4.2节提出了一种类似方案，适用于 $\sigma < 0$时的伽马/平移负二项先验。

A.4.1 情况 $0 \le \sigma < 1$和 $\alpha > 0$

Teh (2006) 针对以下情形提出了一种辅助变量方案： $0 \le \sigma < 1$ 且 $\alpha > 0$ ， 使得先验分布

$$\sigma \sim \text{Beta}\left(\zeta^{(0)}, \zeta^{(1)}\right) \quad \text{and} \quad \alpha \sim \text{Gamma}\left(\chi^{(0)}, \chi^{(1)}\right)$$

具有共轭性。下面概述该方案，详情请参阅（Teh， 2006）。该方案在给定两个参数 α 和 σ 的条件下，引入以下几组辅助变量：

$$\begin{aligned} w \mid N, \alpha &\sim \text{Beta}(\alpha + 1, N - 1), \\ u_k \mid \sigma, \alpha, E &\sim \text{Bernoulli}\left(\frac{\alpha}{\alpha + \sigma k}\right), \qquad k \in \{1, \ldots, E-1\} \\ v_{ej} \mid \sigma, \mathbf{\Lambda} &\sim \text{Bernoulli}\left(\frac{j-1}{j-\sigma}\right), \quad \forall e, j \in \{1, \ldots, N_e - 1\}. \end{aligned} \tag{S21}$$

其中， $N_e = |\{i : \lambda_i = e\}|$ 表示与实体 e链接的记录数量， $E = \sum_e \mathbb{1}[N_e > 1]$表示至少与一条记录链接的实体数量。

由此可得，在给定辅助变量和其他模型参数的条件下， α 和 σ 的后验分布为：

$$\begin{aligned} \sigma \mid \{u_k\}, \{v_{ej}\}, \mathbf{\Lambda} &\sim \text{Beta}\left(\zeta^{(0)} + \sum_{k=1}^{E-1}(1-u_k), \zeta^{(1)} + \sum_{e:N_e>1}\sum_{j=1}^{N_e-1}(1-v_{ej})\right), \\ \alpha \mid \{u_k\}, w, \mathbf{\Lambda} &\sim \text{Gamma}\left(\chi^{(0)} + \sum_{k=1}^{E-1} u_k, \chi^{(1)} - \log w\right). \end{aligned} \tag{S22}$$

因此，要更新 α和 σ，首先根据方程（S21）， 在给定链接结构 $\mathbf{\Lambda}$ 以及 α 和 σ 的旧值的条件下，抽取辅助变量 w、$\{u_k\}$和 $\{v_{ej}\}$。然后，在给定辅助变量和链接结构的条件下，根据方程（S22）抽取 α 和 σ 的新值。

A.4.2 $\sigma < 0$情形， $\alpha = m\kappa$ 且 $m \in \mathbb{N}$

我们描述一种辅助变量方案，用于在以下范围内更新 Ewens-Pitman 参数： $\sigma < 0$ 且 $\alpha = m\kappa$，其中 $m \in \mathbb{N}$ 且 $\kappa > 0$。该方案受到 Teh(2006) 的启发。将 N 条记录划分为 E 个实体所对应的似然因子如下（Pitman， 2006）：

$$P(\text{partition config}) = \frac{(m)_{E\downarrow}}{(m\kappa)_{N\uparrow}} \prod_{e=1}^{E} (\kappa)_{N_e\uparrow} = \frac{\kappa^{E-1}(m-1)_{E-1\downarrow}}{(m\kappa-1)_{N-1\uparrow}} \prod_{e=1}^{E} (\kappa-1)_{N_e-1\uparrow}, \tag{S23}$$

其中，“划分配置”是将链接结构 $\mathbf{\Lambda}$ 表示为一个划分[6]， N_e 是链接到第 e 个实体的记录数量，$(x)_{n\uparrow} = \prod_{i=0}^{n-1}(x+i)$ 是上升阶乘，而 $(x)_{n\downarrow} = \prod_{i=0}^{n-1}(x-i)$

[6]记录 i 和 若 $\lambda_i = \lambda_j$，则属于划分中的同一子集；否则属于不同子集 s.

是下降阶乘。我们首先将该方程中的分母表示为

$$\frac{1}{(m\kappa-1)_{N-1\uparrow}} = \frac{\Gamma(m\kappa+1)}{\Gamma(m\kappa+N)} = \frac{\mathrm{B}(m\kappa+1, N-1)}{\Gamma(N-1)}$$
$$= \frac{1}{\Gamma(N-1)} \int_0^1 w^{m\kappa} (1-w)^{N-2} \, \mathrm{d}w,$$

这使我们能够引入以下辅助变量：

$$w \mid m, \kappa, N \sim \mathrm{Beta}(m\kappa+1, N-1). \tag{S24}$$

将方程 (S23) 中的每个后一因子表示为

$$(\kappa-1)_{N_e-1\uparrow} = \prod_{j=1}^{N_e-1} (\kappa+j) = \prod_{j=1}^{N_e-1} \sum_{v_{ej}\in\{0,1\}} \kappa^{v_{ej}} j^{1-v_{ej}}$$

由此可引入以下额外的辅助变量：

$$v_{ej} \mid \kappa \sim \mathrm{Bernoulli}\left(\frac{\kappa}{\kappa+j}\right), \qquad \forall e, j \in \{1, \ldots, N_e - 1\}. \tag{S25}$$

采用这种表示形式后，我们可以为 κ 和 m指定共轭先验，即：

$$\kappa \sim \mathrm{Gamma}(\chi^{(0)}, \chi^{(1)}) \text{ and } m \sim \mathrm{NegativeBinomial}(r, \nu) + 1. \tag{S26}$$

m 服从移位负二项分布，其支撑集为正整数。我们采用的负二项分布参数化方式，是以试验序列中在达到给定的成功次数 $r > 0$ 之前的失败次数 $x \in$**{0，1，2，.....}**为参数。每次试验都是从成功概率为 ν的伯努利分布中进行的一次独立同分布抽样。 x 的密度为

$$P(x \mid r, \nu) = \frac{(x+r-1)!}{(r-1)!x!} \nu^r (1-\nu)^x.$$

最后，我们将方程 (S26) 中的先验分布与似然因子结合，得到以下后验分布，其中 m 和 κ以其他模型参数为条件：

$$m \mid w, \kappa, \mathbf{\Lambda} \sim \mathrm{NegBinomial}(r+E-1, 1-(1-\nu)w^\kappa) + E,$$
$$\kappa \mid \{v_{ej}\}, w, m, \mathbf{\Lambda} \sim \mathrm{Gamma}\left(\chi^{(0)} + E - 1 + \sum_{e=1}^{E} \sum_{j=1}^{N_e-1} v_{ej},\ \chi^{(1)} - m \log w\right). \tag{S27}$$

因此，要更新 κ和 m，首先应在给定链接结构 $\mathbf{\Lambda}$ 以及 α 和 σ 的旧值的条件下，使用方程 (S24) 和 (S25) 抽取辅助变量 w和 $\{v_{ej}\}$。然后，在给定辅助变量和链接结构的条件下，使用方程 (S27) 为 κ和 m 抽取新值。

A.5 更新失真分布的浓度

本节介绍失真分布集中度 ρ_a 的更新。由于更新时无法依赖共轭性，我们提出一种辅助变量方案。更新ρ_a时，我们以实体属性值 $\{y_{ea}\}_{e=1\ldots E}$、记录属性值 $\{x_{ia}\}_{i=1\ldots N}$以及链接 $\boldsymbol{\Lambda} = \{\lambda_i\}_{i=1\ldots N}$ 为条件。我们将失真分布 $\{H_{ea}\}_{e=1\ldots E}$ 积分消去。似然中涉及 ρ_a 的部分为：

$$\begin{aligned}
&\prod_e \int \prod_{i:\lambda_i=e} P(x_{ia} \mid \theta_{\varsigma_i a}, \omega_{ia}, y_{ea}, H_{ea}) P(H_{ea} \mid \rho_a)\, \mathrm{d}H_{ea} \\
&\propto \prod_e \frac{\Gamma(\rho_a)}{\Gamma(\bar{n}_{ea}(y_{ea}) + \rho_a)} \prod_{v \in \mathcal{V}_{ea}} \frac{\Gamma(n_{ea}(v) + \rho_a \psi_a(v \mid y_{ea}))}{\Gamma(\rho_a \psi_a(v \mid y_{ea}))} \\
&= \prod_e \frac{\mathrm{B}(\rho_a, \bar{n}_{ea}(y_{ea}))}{\Gamma(\bar{n}_{ea}(y_{ea}))} \prod_{v \in \mathcal{V}_{ea}} \prod_{j=1}^{n_{ea}(v)} \{j - 1 + \rho_a \psi_a(v \mid y_{ea})\}
\end{aligned} \tag{S28}$$

其中 $\mathrm{B}(\cdot,\ \cdot)$ 是贝塔函数，$\mathrm{V}_{ea} := (\bigcup_{i:\ \lambda_i=e} \{x_{ia}\}) \setminus \{y_{ea}\}$，$n_{ea}(v) = \sum_{i:\ \lambda_i=e} \mathbb{1}[x_{ia} = v]$和$\bar{n}_{ea}(v) = \sum_{i:\ \lambda_i=e} \mathbb{1}[x_{ia} \neq v]$。

将方程（S28）中的贝塔函数表示为

$$\mathrm{B}(\rho_a, \bar{n}_{ea}(y_{ea})) = \int_0^1 w_e^{\rho_a - 1} (1 - w_e)^{\bar{n}_{ea}(y_{ea}) - 1}\, \mathrm{d}w_e$$

使我们能够引入以下辅助变量：

$$w_e \mid \rho_a, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\Lambda} \sim \mathrm{Beta}\left(\rho_a, \sum_e \sum_i \mathbb{1}[x_{ia} \neq y_{\lambda_i a}]\right),$$

对所有 e。我们还可以将方程（S28）中的每个后续因子表示为

$$j - 1 + \rho_a \psi_a(v \mid y_{ea}) = \sum_{u_{evj}=0}^{1} \rho_a \psi_a(v \mid y_{ea})^{u_{evj}} (j - 1)^{1 - u_{evj}}$$

这使我们能够引入以下辅助变量：

$$u_{evj} \mid \rho_a, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\Lambda} \sim \mathrm{Bernoulli}\left(\frac{\rho_a \psi_a(v \mid y_{ea})}{j - 1 + \rho_a \psi_a(v \mid y_{ea})}\right) \tag{S29}$$

对所有 e、$v \in \mathrm{V}_{ea}$，以及 $j \in \{1, \ldots, n_{ea}(v)\}$。

现在，由于 ρ_a 的先验分布为 $\mathrm{Gamma}(\tau_a^{(0)}, \tau_a^{(1)})$，我们得到在给定其他参数的条件下，$\rho_a$ 的后验分布如下：

$$\rho_a \mid \{w_e\}, \{u_{evj}\}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\Lambda} \sim \mathrm{Gamma}\left(\tau_a^{(0)} + \sum_e \sum_{v \in \mathcal{V}_{ea}} \sum_{j=1}^{n_{ea}(v)} u_{evj}, \tau_a^{(1)} - \sum_e \log w_e\right). \tag{S30}$$

因此，要更新 ρ_a，首先根据记录属性 $\boldsymbol{X}$、实体属性 $\boldsymbol{Y}$、链接结构 $\boldsymbol{\Lambda}$以及 ρ_a 的前一个值，使用方程 (A.5) 和 (S29)，抽取辅助变量 $\{w_e\}$和 $\{u_{evj}\}$。然后，在给定辅助变量以及 $\boldsymbol{X}$、 $\boldsymbol{Y}$、 $\boldsymbol{\Lambda}$的条件下，使用方程 (S30) 为 ρ_a 抽取一个新值。

B 混合距离度量

在本附录中，我们介绍一种混合距离度量，可用于比较包含多个标记（单词）的文本字符串，其中各个标记可能会发生失真。本文使用该度量比较cora 和rest 数据集中的名称和地址属性（参见第5.1节），不过它在本文之外也可能有更广泛的应用。我们借鉴了Monge 和 Elkan(1996）提出的一种混合相似度度量。不过，与Monge 和 Elkan不同的是，我们尝试匹配每个字符串中的标记，同时对其中一个字符串缺失的标记施加惩罚。

假设我们要比较一对多词元字符串 x 和 y。作为贯穿全文的示例，我们考虑 x = “University of California, San Diego”和 y = “Univ. Calif., San Diego”。给定一个分隔符（例如空格），我们可以将每个字符串映射为一组标记。例如，贯穿示例中的字符串 x将映射为

$$X = \{\text{“California,”}, \text{“Diego”}, \text{“of”}, \text{“San”}, \text{“University”}\}.$$

请注意，我们使用大写字母 X 表示词元集合[7] 字符串 x 的表示形式——这是我们在本附录中始终采用的惯例。另请注意， X 是 x的一种有损表示形式，因为它会丢弃词元顺序信息。对于我们在姓名和地址方面的应用，这样做是合适的[8]，因为词元的排列不会显著改变字符串的含义。

我们提出通过对词元集合 X 和 Y使用广义编辑距离，来衡量 x 到 y 的距离。我们考虑三种基本编辑操作：

- 词元插入，即将一个词元 b 添加到输入集合中；
- 词元删除，即从输入集合中移除一个词元 a ；以及
- 词元替换，即将输入集合中的一个词元 a 替换为词元 $b \neq a$。

每个基本操作都将输入集合 Q 转换为输出集合 Q'，记作 $Q \to Q'$，并具有相应的代价 $c(Q \to Q') \geq 0$。我们令

$$c(Q \to Q') = \begin{cases} d_i \ \mathrm{dist}_{\mathrm{inner}}(\lambda, b), & \text{if } Q = Q' \setminus \{b\} \text{ (insertion)}, \\ d_d \ \mathrm{dist}_{\mathrm{inner}}(a, \lambda), & \text{if } Q \setminus \{a\} = Q' \text{ (deletion)}, \\ d_s \ \mathrm{dist}_{\mathrm{inner}}(a, b), & \text{if } Q \setminus \{a\} = Q' \setminus \{b\} \text{ (substitution)}, \end{cases}$$

[7]严格来说，我们考虑的是多重集，因为允许词元多次出现。[8]具体而言，cora 中的标题、发表场所和作者属性，以及 rest 中的姓名和地址属性。

其中 d_i， d_d 和 d_s 是非负权重； λ 是空字符串；$\text{dist}_{\text{inner}}(\cdot, \cdot)$ 是标记（字符串）上的 内部距离度量。接着，我们将 x 与 y 之间的混合距离定义为通过一系列基本编辑操作 $T_{X,Y} = (X \to Q_1, Q_1 \to Q_2, \ldots, Q_{l-1} \to Y)$ 将 X 转换为 Y 的最小平均代价。符号表示为

$$\text{dist}_{\text{hybrid}}(x, y) = \min_{T_{X,Y}} \frac{1}{|T_{X,Y}|} \sum_{(Q \to Q') \in T_{X,Y}} c(Q \to Q').$$

我们可以使用现成的线性和指派问题（LSAP）求解器来计算混合距离（Crouse，2016年）。为此，我们需要向 X 和 Y 中添加空字符串标记，以涵盖所有可能的插入和删除操作。具体来说，我们向 $|Y|$ 中添加 X 个空标记，以允许插入；并向 $|X|$ 中添加 Y 个空标记，以允许删除。然后，我们对（修正后的）$\text{dist}_{\text{inner}}$ 中的所有标记对应用 X 和 Y 构造成对代价矩阵。随后将所得矩阵传递给 LSAP求解器，后者会返回最优编辑操作集及其代价。

回到我们一直使用的示例，若将 $\text{dist}_{\text{inner}}$ 设为 Levenshtein距离，则 LSAP 的解为

$$\{(\text{“University”} \leftrightarrow \text{“Univ.”}, 5), (\text{“of”} \leftrightarrow \lambda, 2), (\text{“California,”} \leftrightarrow \text{“Calif.,”}, 6),$$
$$(\text{“San”} \leftrightarrow \text{“San”}, 0), (\text{“Diego”} \leftrightarrow \text{“Diego”}, 0), (\lambda \leftrightarrow \lambda, 0), (\lambda \leftrightarrow \lambda, 0),$$
$$(\lambda \leftrightarrow \lambda, 0), (\lambda \leftrightarrow \lambda, 0)\}.$$

因此我们得出结论：$\text{dist}^{\text{hybrid}}(x, y) = \frac{5+2+6+0+0}{5} = 2。6。$ 与直接对 x 和 y计算时得到较大的值 14 的 Levenshtein距离相比，该距离更能体现 x 和 y 之间的语义接近度。

C 模拟研究

在本附录中，我们开展了一项模拟研究，以了解模型在受控情境下的表现。具体而言，我们模拟实体解析数据集，并改变记录数量、失真程度和重复程度。由于我们的模型是生成式的，我们可以用它来模拟数据；然而，由此生成的数据对于我们的模型而言设定误差微乎其微，这并不现实。因此，我们通过在实体与实体属性之间添加额外的依赖关系，并采用不同的过程从实体生成记录，来模拟特意设定为不符合我们模型的数据。我们未能找到能够生成此类数据的现有数据集模拟器，因此自行实现了一个。

C.1 数据集模拟器

我们概述了模拟器的工作方式：它生成描述家庭总体的个人记录。为简洁起见，此处省略低层级的细节，请读者参阅随附的 Python 脚本。模拟器分两个阶段运行：第一阶段生成一个家庭总体；第二阶段则遍历所有家庭中的个人，为每个人生成随机数量的失真记录。第一阶段通过生成家庭而非个人，

阶段，我们能够纳入个人（实体）之间的其他依赖关系，而这些依赖关系在我们的实体解析模型中并不存在（参见第 3.1节）。

生成家庭。现在描述第一阶段如何生成家庭。在我们的简化模型中，一个家庭可以由一对夫妻、一个单身者、一对夫妻或一名单身者与子女，或一群无亲属关系的成年人组成。家庭中的个人由以下属性描述：姓名（名字_和姓氏_）、出生日期（出生_年份、出生_月份、出生_日期）、性别和邮政编码。家庭中所有个人的邮政编码都相同；名字则取决于性别，而其他属性可能会按下文所述有所不同。属性的随机值由 Faker Python库[9]生成，该库旨在模拟现实世界频率分布。我们会让名称和邮政编码属性的分布更集中，以确保实体不会过于独特（否则实体解析将过于简单）。

我们首先生成家庭户主：为一对男性和女性组成的夫妻（为简化起见），或一名单身男性或女性。如果生成的是夫妻，他们很可能姓氏相同，且出生年份相差不大。接下来，我们随机决定是否生成子女。如果生成子女，他们与家庭户主（父母）姓氏相同，且他们的出生年份与父母的出生年份之间会有适当的间隔。如果没有生成子女，我们就随机决定是否生成与家庭户主同住的无亲属关系的成年人。这些成年人年龄须与家庭户主相近。模拟家庭构成时，我们努力遵循当前人口调查的汇总统计数据（美国人口普查局，2016年）。

生成记录。 在第二阶段，为所有家庭中的个人生成记录。我们通过以概率 $p_{\text{inc}} = 0$。9 纳入某个个人，并按照速率参数为 μ、截断至区间 $[1, 4]$ 的泊松分布抽取记录数量，来模拟包含重复记录的单个数据库/文件。每条记录都是通过复制实体属性并施加失真得到的。具体做法是按随机顺序遍历各项属性，并以因属性而异的概率决定是否启动失真过程。对于出生日期中的日、月份、性别和邮政编码，失真过程会根据第一阶段使用的分布抽取一个替代值。对于出生年份，失真过程会在真实出生年份上添加离散高斯噪声。对于姓名，失真过程可能以三种方式之一进行：（1）随机造成排版错误（字符插入、删除、替换或换位）；（2）从变体姓名中均匀随机抽取一个来替换原姓名；或（3）根据第一阶段使用的分布生成替代值。（2）中的变体姓名取自 WeRelate.org变体姓名项目[10]。

C.2 结果

我们使用模拟器，为以下变量的每种组合生成16个数据集：

[9] https://github.com/joke2k/faker [10] https://www.werelate.org/wiki/WeRelate:Variant_names_project

属性	失真概率	
	低失真	高失真
第一_name	10%	40%
last_name	10%	40%
性别	1%	1%
邮政编码	5%	10%
出生_年份	1%	10%
出生_月份	1%	10%
出生_日期	1%	10%

表 S4：低和高两种失真程度下的属性级失真概率。

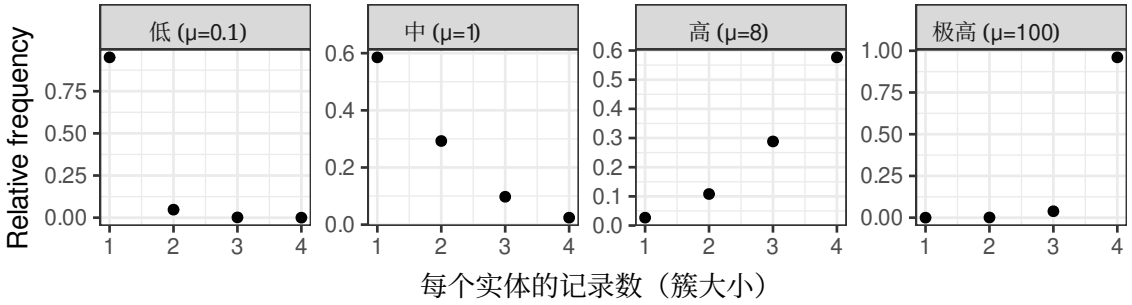


图 S5：低、中、高和极高重复程度下每个实体的记录分布。括号中给出了各重复程度对应的泊松率参数 μ。

• 记录数量。我们考虑期望记录数量为1000和10000的数据集。记录数量是随机的，取决于个人数量和泊松率参数。由于泊松率参数固定（见下文），我们通过改变第一阶段生成的个人数量来控制记录数量。

• 失真程度。我们考虑两种记录失真程度，分别称为“低”和“高”。这对应于不同的失真过程激活概率选择，详见表S4。

• 重复程度。我们考虑四种重复程度，分别称为“低”、“中”、“高”和“极高”。这些程度分别对应泊松率参数 0.1、1、8 和 100。当重复程度为“低”（$\mu = 0。1$）时，数据中表示的实体有超过 95% 仅出现一次；而当重复程度为“极高”（$\mu = 100$）时，数据中表示的实体有超过 95% 出现四次。图中绘出了各个重复程度下每个实体的记录分布S5。

我们对模型、blink和Sadinle 在 16 个模拟数据集上的表现进行了比较评估。模型评估流程见第5.2节，blink 和Sadinle 模型则在第5.4节介绍。图S6展示了各数据集的实体解析评估指标。下面我们对结果作几点观察。

首先，我们发现，当重复程度为中、高或极高时，我们的模型与blink 表现相近。在这些重复程度下，当失真程度较高时，blink 在召回率方面略占优势。差异最大的是中等重复／高失真情形，此时blink 的召回率比我们的模型高出约 10 个百分点。这一差异源于我们模型中对浓度参数 ρ_a 设置的先验分布偏向较高的浓度。这相当于先验信念认为失真总是以相同的方式发生，而非以多种不同的方式发生。然而，模拟数据中的失真并非如此——例如，名字为“JONATHON”的某个人，其记录中的名字可能有四种不同写法：“JOHN”、“JOJN”、“JONATHON”和“ALEX”。如果想利用这一知识，我们可以调高我们模型中的 $\tau_a^{(0)}$，使其倾向于较低的浓度。

其次，我们发现，在重复程度较低时，我们的模型在精确率方面显著优于blink 。我们认为，这是因为blink 使用的链接结构先验信息量很大——它采用优惠券先验，并将 m 固定为记录数量 N 和 $\kappa \to \infty$。然而，在广义优惠券先验下，我们的模型会选择约为 $8 \times N$ 的 m 以及约为 100 的 κ，从而能够更准确地对低重复情形建模。

第三，我们观察到，当重复程度为中、高或极高时，Sadinle 的 F1 分数最低。在这些重复程度下，失真较高时，F1 分数的性能差距最大，约为 20 个百分点；失真较低时，差距则不那么显著，约为 5 个百分点。F1 分数的这些差异主要是由于多数情况下召回率较低所致，而精确率总体上与其他模型相当。

第四，我们讨论数据集大小的影响，该大小以预期记录数量（1000 或 10000）衡量。我们观察到，三个模型呈现出相似的趋势：较小数据集的精确率往往较高，而较大数据集的召回率往往较高（但也有例外）。这种性能差异在低重复设置下最为明显：对于较大的数据集，blink 和Sadinle 的精确率显著下降，而Sadinle的召回率也显著下降。

D 链接结构先验与失真模型研究

在本附录中，我们补充了第5.3节中关于链接结构先验和失真模型的研究结果。我们的目标是分别研究各项建模贡献的影响，确定每项贡献本身是否有益，以及／或者其中一项是否比另一项更有益。

回顾第5.3 节，我们考虑了链接结构先验的四种参数区间——PY、Ewens、GenCoupon 和 Coupon ——以及两种失真模型——我们的方法 （如第3.1节所述）和blink （由Steorts 于（2015年）提出）。因此，共有八种模型变体需要测试，每种链接结构先验和失真模型各对应一种。图S7 在同一张图中展示了八种模型变体在四个数据集上的成对评估指标（F1 分数、精确率和召回率）。下面我们将分别解读各项建模贡献的结果。

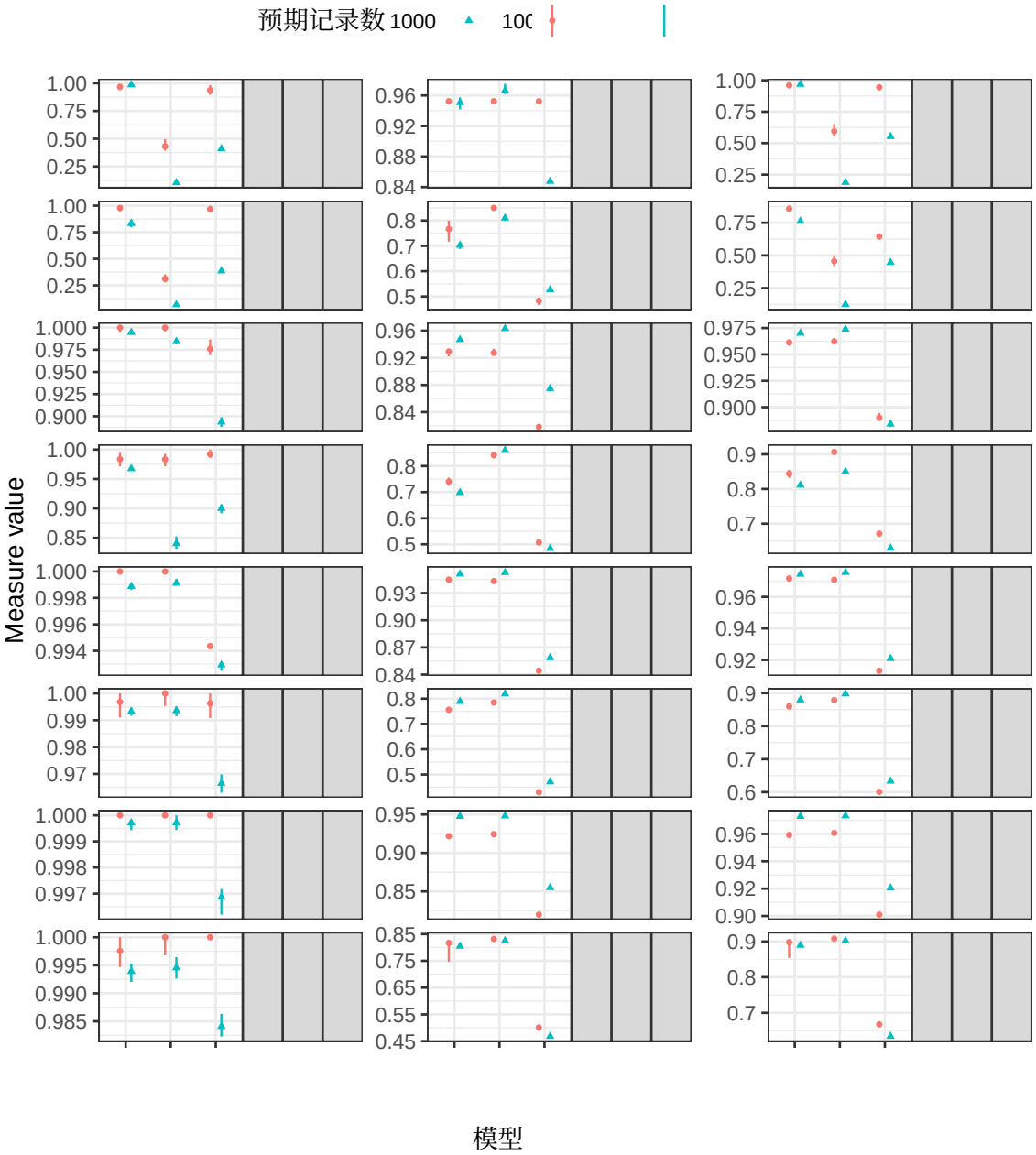


图 S6：我们的模型、blink和Sadinle 在失真程度和重复程度各异的模拟数据集上拟合后的后验评估指标。低失真和高失真分别缩写为“L. Dist.”和“H. Dist.”。低、中、高和极高重复程度分别缩写为“L. Dup.”、“M. Dup.”、“H. Dup.”和“V. H. Dup.”。预期记录数（1000 或 10000）由标记的颜色和形状表示。每项评估指标均报告基于中位数的点估计，并以点估计周围的区间表示 95% 等尾可信区间。

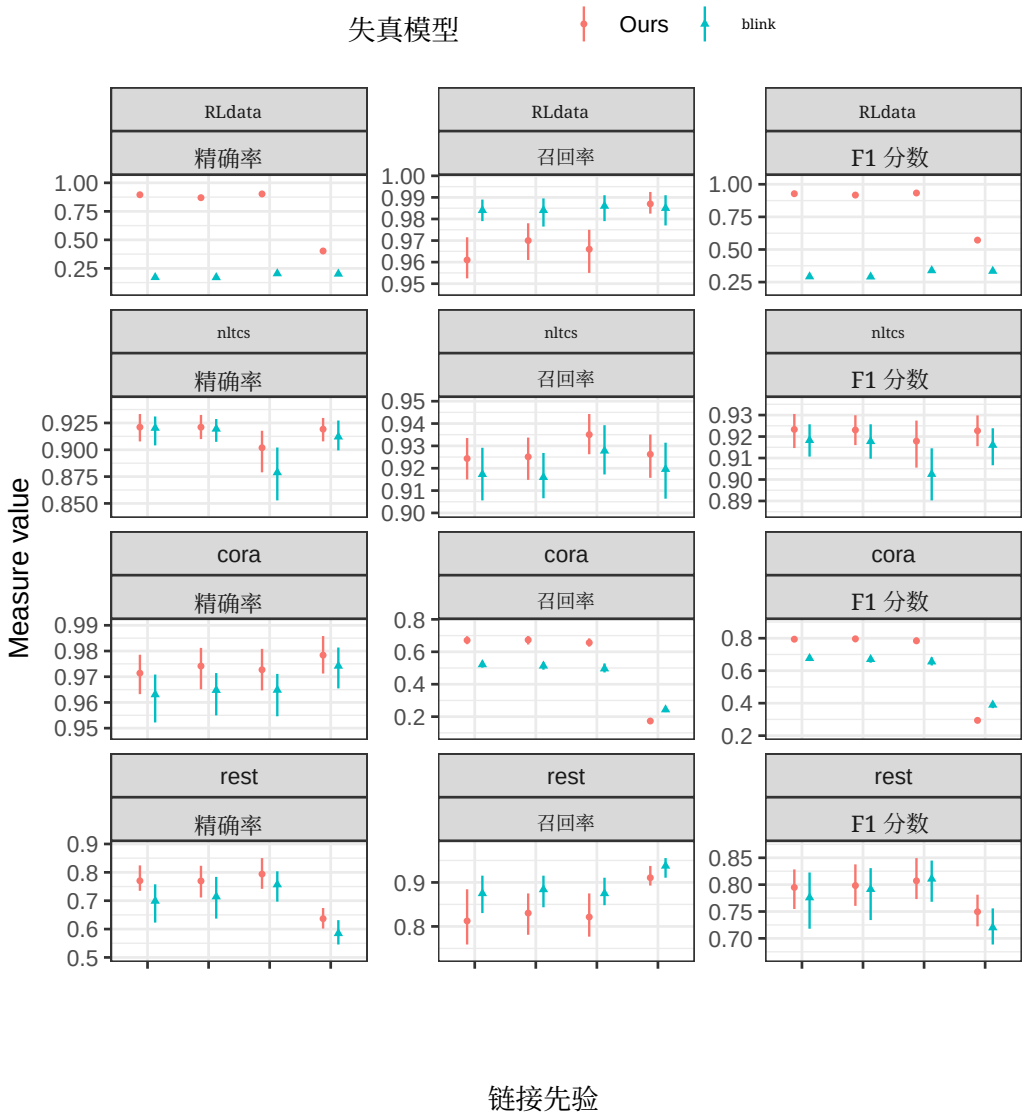


图 S7：实体解析质量随链接结构先验（绘制于x轴上）和失真模型（以线条颜色表示）变化的评估。图中展示了四个数据集（按行分组）的三种成对评估指标（按列分组）。

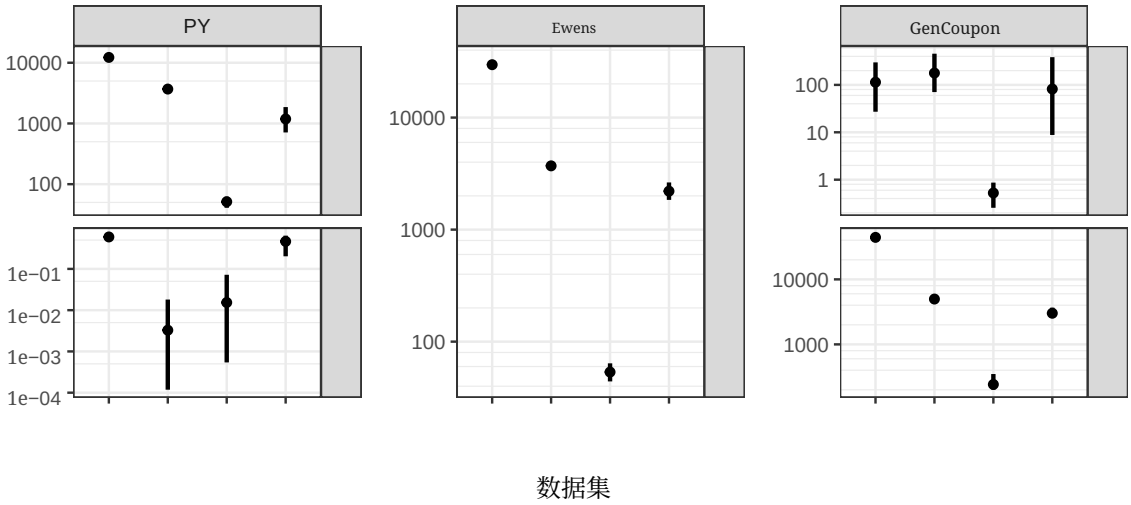


图 S8：三种机制下的 Ewens-Pitman 参数后验分布：PY、Ewens和GenCoupon 在我们的失真模型（我们的方法）下的参数。请注意，参数值以对数刻度呈现。

链接结构先验。 虽然我们在第 5.3 节讨论了链接结构先验的影响，但由于篇幅限制，我们只展示了我们的失真模型（我们的方法）的结果。在第 5.3 节中，我们从表 2 的结果中得出两个主要结论，这些结论在图 S7（以朱红色圆形标记表示）中得到了重复验证：

1. 我们提出对 EP 参数（对于 PY、Ewens 和 GenCoupon）设置宽松超先验的方法，能够提高稳健性，并在三个数据集中取得最高的实体解析准确率（以成对 F1 分数衡量；nltcs 是例外）。不使用超先验时，我们观察到 F1 分数显著较低（参见图 S7 中的 Coupon），尤其是在 cora 和 RLdata 数据集中。图 S8 进一步印证了这一观点，它显示，在宽松超先验的作用下，每个数据集所选取的 EP 参数值差异很大。

2. 尽管已知每种机制都呈现出不同的渐近行为（见第 PY、Ewens 或 GenCoupon）参数机制2.2），我们的推断对 EP 参数机制相对不敏感。

图S7 表明，这些结论同样适用于blink 失真模型（以青绿色三角形标记表示）。我们预计，nltcs 在Coupon 链接先验下表现出色可能只是巧合，因为该先验下的总体规模为 3,387，恰好非常接近真实值 3,307（见表1）。

失真模型。在第5.3节中，我们讨论了GenCoupon链接结构先验下失真先验的影响。现在我们将讨论扩展到其他三种链接结构先验（PY、Ewens 和Coupon）的结果，如图S7所示。

我们发现，除一组数据集和链接结构先验之外，我们的失真模型在所有数据集和链接结构先验下都取得了最高的 F1 分数。例外是，在 cora 数据集上使用 Coupon 链接结构先验时，

该blink 失真模型略胜一筹。第 5.3 节解释了我们的模型性能提升的原因，下面在此概述。blink 失真模型容易进入高失真模式，尤其是对于使用非常数距离度量的属性。这是因为，即使记录属性值实际上并未失真，该模型也允许将其标记为“已失真”。我们的模型修正了这一不一致之处，因此表现得更加稳健。总体而言，我们预期blink 失真模型会导致过度链接（高召回率、低精确率），而我们的模型则预期会更加均衡。图S7 支持这一观点，其中差异在RLdata数据集上最为明显，该数据集的 F1 分数相差 ~0.7。

E 实验设置的更多细节

在本　ppe附录中，我们将进一步介绍第 5 节所述实验的细节。

实现与硬件。 所有实验均使用 R 3.4.4版运行，运行环境为一台配备两颗 28 核 Intel Xeon Platinum 8180M处理器和 12 TB 内存的本地服务器。[11] 我们开发了一个名为 `exchanger`[12] 的开源 R 软件包，除 blink 模型（Steorts，2015年）外，还实现了我们模型的多个变体（采用不同的链接结构先验和失真模型）。由于由Sadinle（2014年）提出的模型没有公开的实现，我们自行开发了实现，并将其作为名为 `BDD`[13] 的开源 R 软件包发布。出于效率考虑，我们使用 ++ Rcpp接口实现了所有模型的推断（Eddelbuettel 和 François，2011年）。附录 C.1 中所述的数据集模拟器以 Python脚本实现。我们提供了一个 Pipfile，用于指定运行该脚本时所需的依赖关系。

超参数设置。我们在设置模型的超参数时遵循了第3.2 节中的建议。在设置两个基线模型的超参数时，我们尽量遵循作者的建议。对于blink，我们将 $m = N$ 设为优惠券收集者先验，并将 $\beta_{sa}^{(0)} = N/1000$ 和 $\beta_{sa}^{(1)} = N/10$ 设为失真概率的贝塔先验（此处 N 为记录总数）。对于Sadinle，我们通过检查每个属性的距离分布来设定一致性水平。按照作者的建议，我们对m概率使用截断均匀先验，对 u概率使用均匀先验。根据附录F中介绍的调参实验，我们将 m概率的下截断点设为 0.95。

初始化与马尔可夫链蒙特卡罗。 对于我们的模型和blink，我们将每条记录链接到一个唯一实体，并将记录属性复制到实体属性中，假设不存在失真，以此初始化链接结构 $\boldsymbol{\Lambda}$、实体属性 $\boldsymbol{Y}$和失真指示变量 $\boldsymbol{Z}$。失真概率 $\boldsymbol{\Theta}$和实体属性分布 $\boldsymbol{G}$通过从其条件分布中抽样进行初始化。Ewens-Pitman 参数和失真分布浓度参数则使用其先验均值进行初始化。

[11] R 脚本发布于github.com/cleanzr/exchanger-experiments。[12] 软件包源代码发布于github.com/cleanzr/exchanger。[13] 软件包源代码发布于github.com/cleanzr/BDD。

Sadinle 模型也采用了类似的初始化方法。我们将每条记录分配给一个唯一实体（簇）。 m-概率和 u-概率通过从其条件分布中抽样进行初始化。

拟合每个模型时，我们运行马尔可夫链蒙特卡罗 2×10^5 次迭代，并将前 10^5 次迭代作为预热期丢弃，同时以 10 为间隔进行稀疏抽样。[14] 由此得到 10^4 个近似后验样本。

F Sadinle（2014）的超参数调参

本附录旨在为 Sadinle（2014）的实体解析模型确定合理的超参数取值，以下简称 Sadinle 模型。我们假设距离函数（用于比较属性）和一致性水平（将实值距离映射为离散等级）均已固定，并且采用平坦先验，正如 Sadinle 所建议的那样。在这些假设下，尚未指定的唯一超参数是 m-概率的下截断点。

m-概率是一组参数 $\{m_{al}\}$，其中 m_{al} 表示在记录对未在等级 0、1、. . .、 $l-1$ 达成一致的条件下，指向同一实体的记录对在属性 a 上达到等级 l 一致的概率。Sadinle 建议对 m_{al} 使用截断平坦先验，使其允许取值落在区间 $[\lambda_{al}, 1]$ 内，其中 $\lambda_{al} \in [0, 1]$ 是一个通常接近 1 的超参数。更具体地说，他建议：若属性 a 是“近乎准确”的准标识符，则设置 $\lambda_{al} = 0.95$；若属性 a 是“不准确”的准标识符，则设置 $\lambda_{al} = 0.85$。由于不清楚哪种 λ_{al} 设置最适合我们的数据集，我们针对四个不同的取值进行了第 5.4 节所述的实验： $\lambda_{al} = 0$、0.5、0.85、0.95。为简单起见，设置 λ_{al} 时，我们对所有属性 a 和一致性水平 l 使用相同的取值。结果见图 S9 和表 S5。

图 S9 绘制了每个截断点 m-概率（位于 x轴上）的后验值λ_{al} （对应于各水平面板）。我们观察到， m_{al} 的后验值往往接近 λ_{al}，尤其是在一致程度为 $l = 0$ 和 $\lambda_{al} \geq 0$。5时。这表明，尽管预计 m_{al}接近 1，模型仍倾向于较小的 m_{al} 值。因此，模型容易“过度链接”——将并非指向同一实体的记录链接起来。要理解模型为何表现出这种行为，还需要进一步探究，这超出了本文的范围。不过，我们推测，这可能与使用平坦先验有关，也可能与 Fellegi-Sunter 型模型已知的稳定性问题有关（Goldstein 等人，2017）。因此，我们得出结论， m_{al} 的后验值对 λ_{al}的选择高度敏感。

表 S5 展示了截断点 λ_{al} 对实体解析性能的影响。结果表明，对于 cora 和 rest ，性能随 λ_{as} 变化相对稳定，尽管 m-概率存在一定变化（如图 S9所示）。这种稳定性可能与这些数据集所采用的阻塞方案有关，该方案排除了相当多的潜在链接，从而避免了过度链接。另一方面， RLdata 和 nltcs 的性能稳定性则差得多。我们发现，随着 λ_{al} 降低，精确率显著下降，

[14] 对于cora 数据集，该链的收敛速度较慢，因此我们将迭代次数增加到 2。5×10^5 并将预热区间增加到 1。5×10^5。

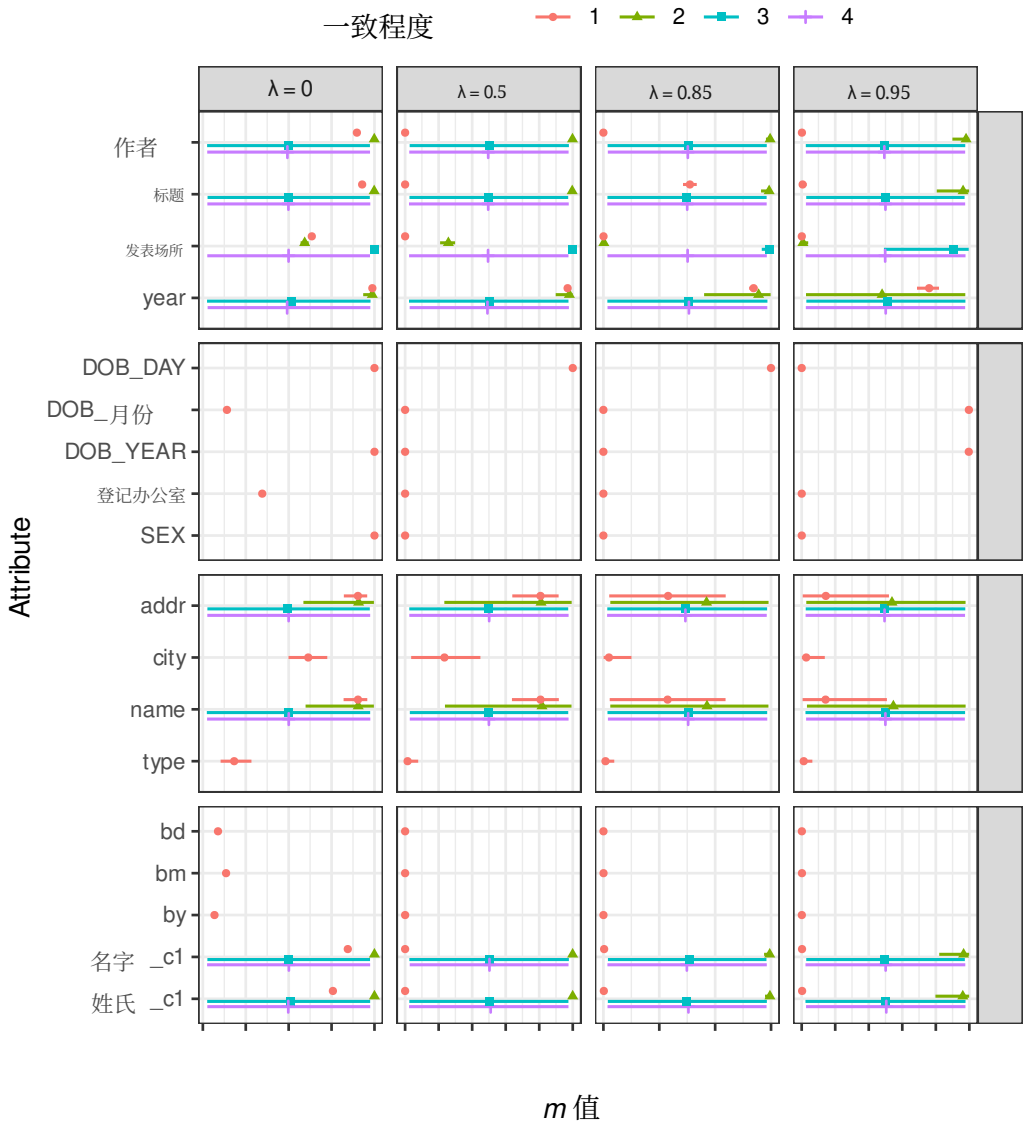


图 S9：在 Sadinle 模型中，m概率的后验分布估计（绘制于 x 轴），以数据集（纵向面板）、属性（y 轴）、一致程度（颜色/标记）和下截断点（横向面板）为变量。点估计基于中位数给出，并附有 95% 等尾可信区间。该模型倾向于为m概率选择较小的值，使其接近下截断点，尤其是在一致程度为 1 时。

数据集	截断点	评估指标		
		精确率	召回率	F1 分数
RLdata	0.00	0.068 (0.067, 0.068)	0.909 (0.906, 0.912)	0.126 (0.125, 0.126)
	0.50	0.069 (0.069, 0.069)	0.921 (0.919, 0.924)	0.128 (0.128, 0.129)
	0.85	0.315 (0.308, 0.321)	0.965 (0.962, 0.967)	0.475 (0.467, 0.481)
	0.95	0.534 (0.524, 0.546)	0.964 (0.962, 0.966)	0.687 (0.679, 0.697)
nltcs	0.00	0.114 (0.113, 0.114)	0.983 (0.978, 0.987)	0.204 (0.203, 0.205)
	0.50	0.111 (0.110, 0.111)	0.964 (0.958, 0.970)	0.199 (0.197, 0.200)
	0.85	0.162 (0.160, 0.164)	0.972 (0.968, 0.976)	0.278 (0.275, 0.281)
	0.95	0.312 (0.304, 0.319)	0.975 (0.969, 0.979)	0.473 (0.464, 0.480)
cora	0.00	0.980 (0.979, 0.981)	0.378 (0.377, 0.378)	0.545 (0.544, 0.546)
	0.50	0.981 (0.979, 0.984)	0.390 (0.389, 0.391)	0.558 (0.557, 0.560)
	0.85	0.984 (0.983, 0.984)	0.383 (0.382, 0.384)	0.552 (0.551, 0.553)
	0.95	0.982 (0.981, 0.983)	0.359 (0.357, 0.362)	0.526 (0.524, 0.529)
rest	0.00	1.000 (0.985, 1.000)	0.607 (0.598, 0.607)	0.756 (0.744, 0.756)
	0.50	0.985 (0.985, 1.000)	0.598 (0.598, 0.607)	0.744 (0.744, 0.756)
	0.85	0.985 (0.985, 1.000)	0.598 (0.598, 0.607)	0.744 (0.744, 0.756)
	0.95	0.993 (0.985, 1.000)	0.603 (0.598, 0.607)	0.750 (0.744, 0.756)

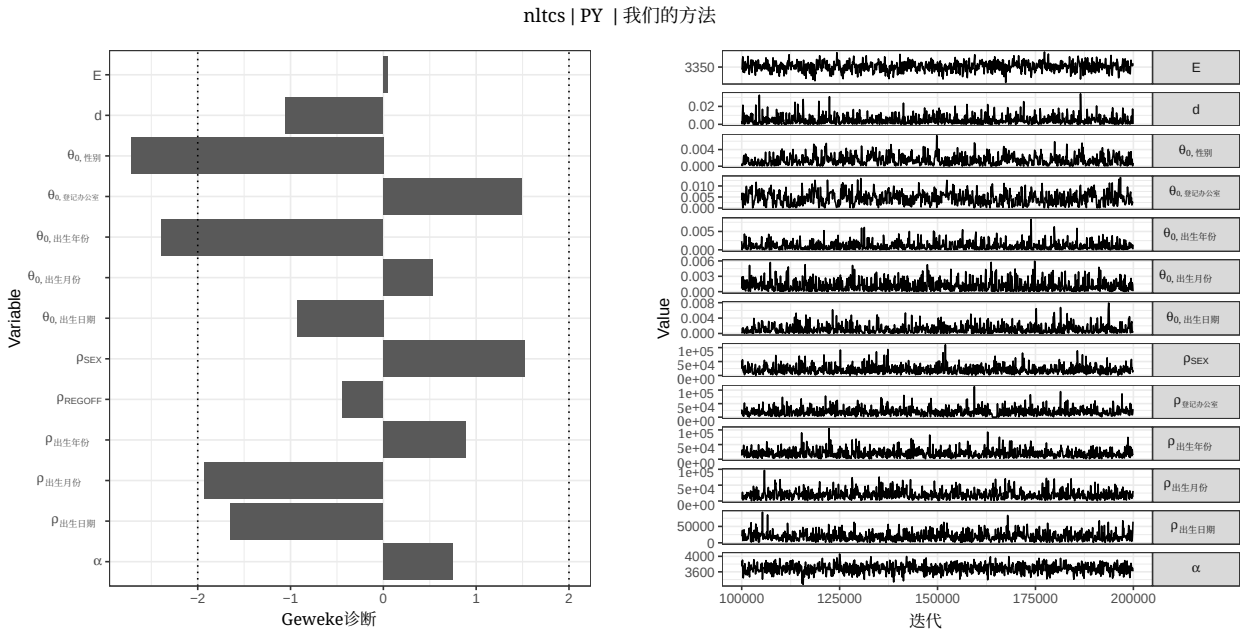
表 S5：Sadinle 模型的后验性能随下截断点 λ_{al}在 m概率上的变化。根据中位数报告各评估指标的点估计，并给出 95% 等尾可信区间。

虽然召回率仍相对稳定。这表明存在过度链接，这也是预期之中的，因为后验 m概率显著更小。基于这些结果，我们将 $\lambda_{al} = 0.95$设为其他实验的默认值，因为它似乎能实现平衡性能。

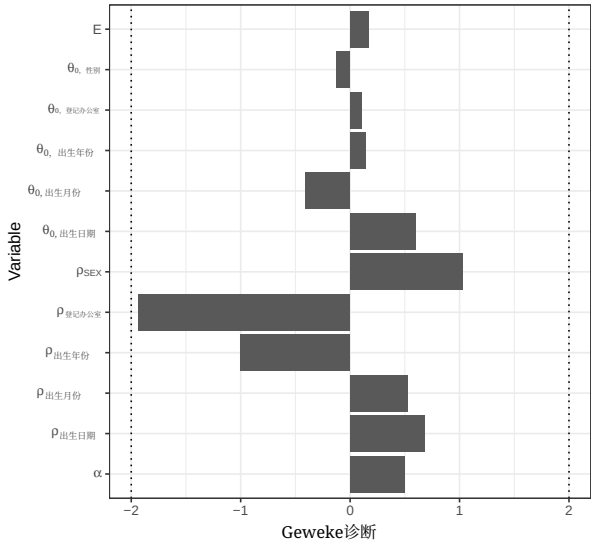
G 马尔可夫链蒙特卡罗诊断

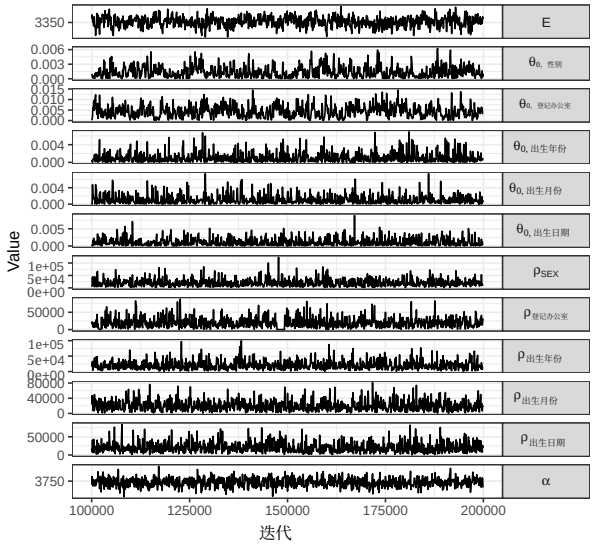
G.1 链接结构先验分布的研究

此处我们展示了第5.3节中拟合模型的收敛诊断。针对每个数据集、链接结构先验和失真模型，我们展示了部分模型变量的 Geweke诊断图和轨迹图。每组图前都带有一个标题，格式为“数据集 | 链接结构先验 | 失真模型”。Geweke诊断图（左侧）展示每个变量（纵轴）对应的 Z分数（横轴）。Z分数用于检验马尔可夫链前 10% 与最后 50% 部分的均值是否相等，通常预期其范围为 $[-2, 2]$ (Geweke,1992)。轨迹图（右侧）展示链中每一步（横轴）对应的变量值（标注在右侧面板中）。请注意，变量 E 表示实例化实体的数量。我们将属性的整数索引替换为命名索引。例如，$\theta_{0,\ \text{city}}$ 表示名为“城市”的属性在来源 0 中的失真概率。

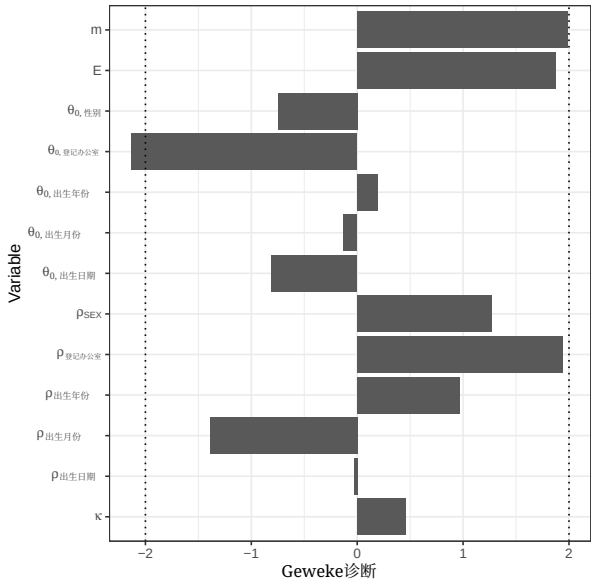


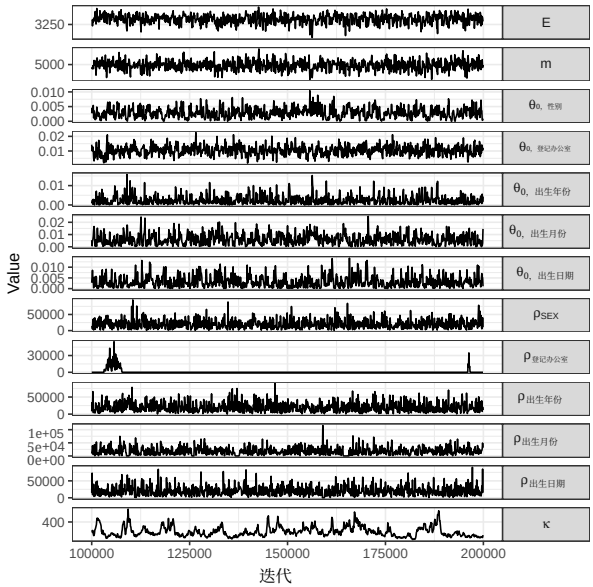
nltcs | Ewens |我们的方法



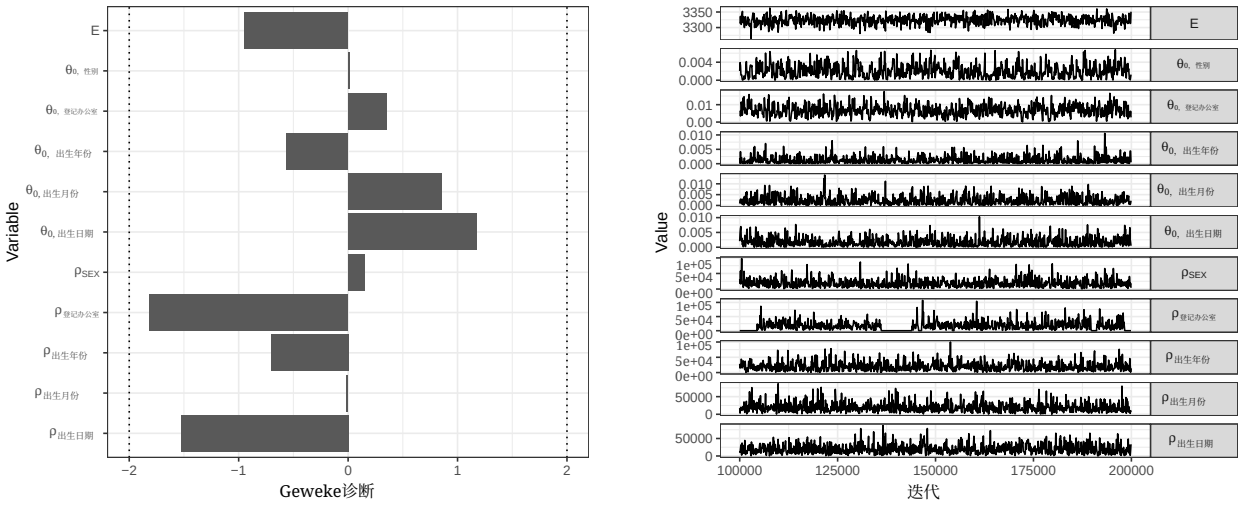


nltcs | GenCoupon |我们的方法

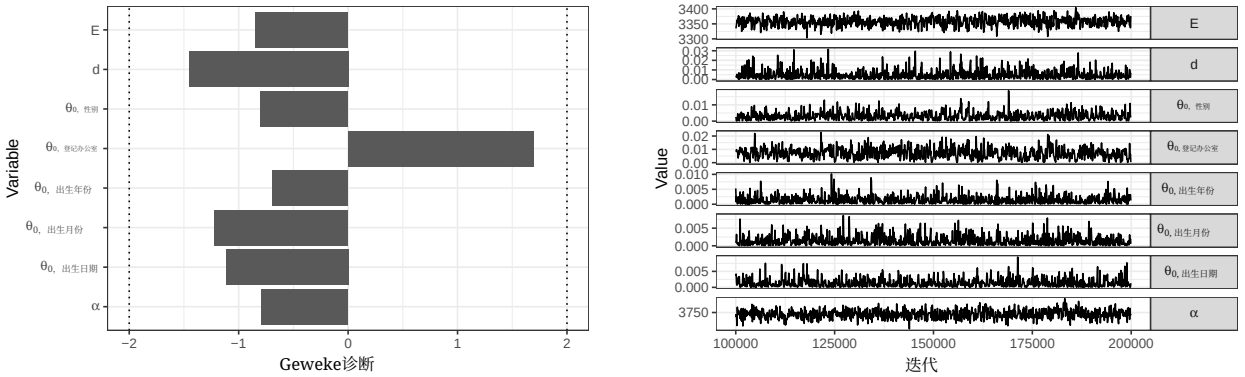




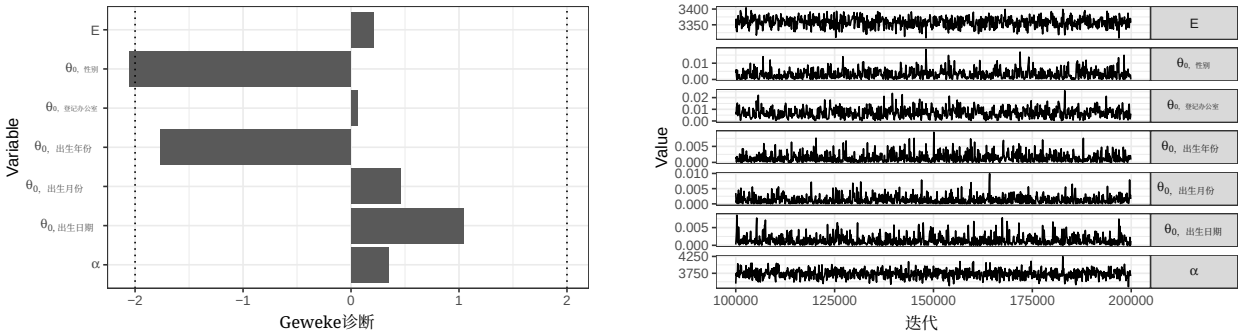
nltcs | Coupon | 我们的方法



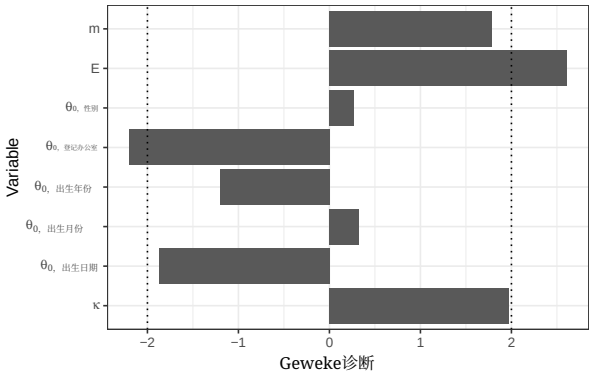
nltcs |PY |blink

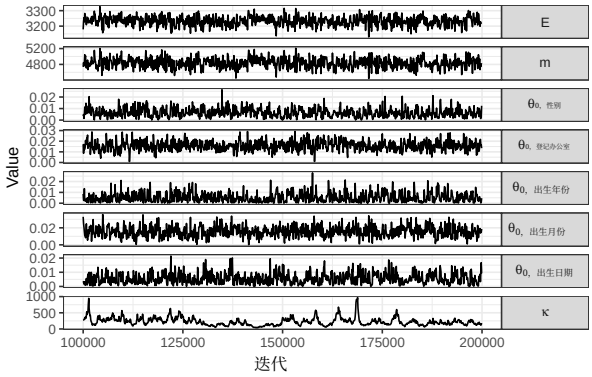


nltcs | Ewens | blink

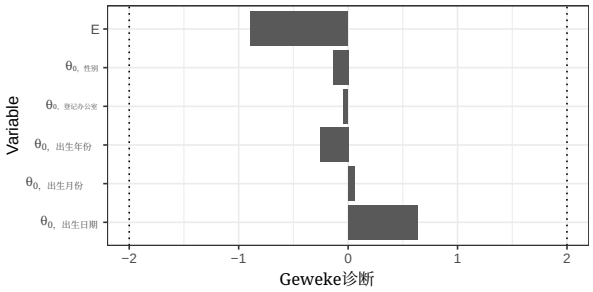


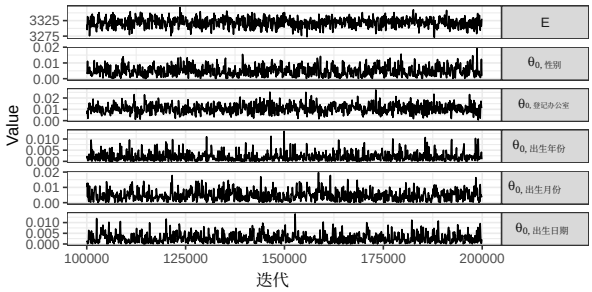
nltcs | GenCoupon | blink



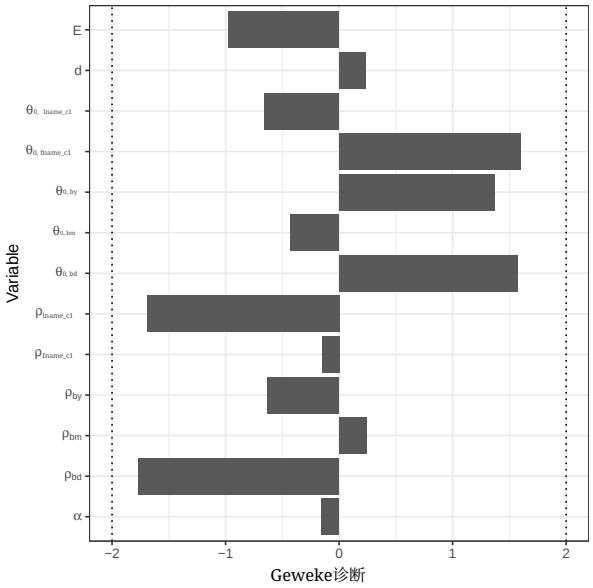


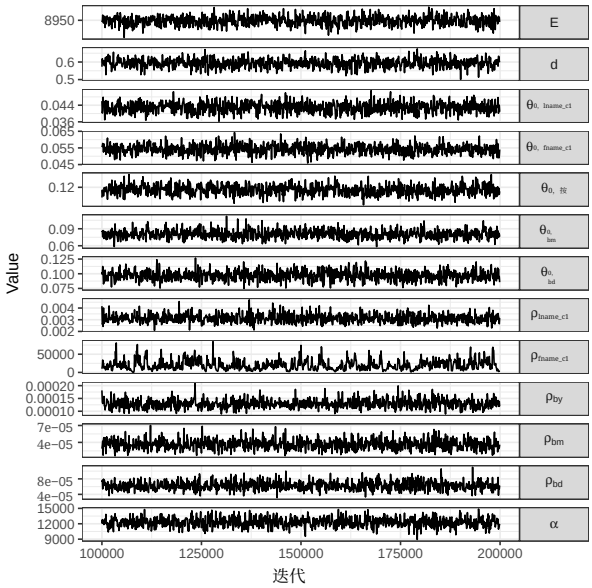
nltcs | Coupon | blink



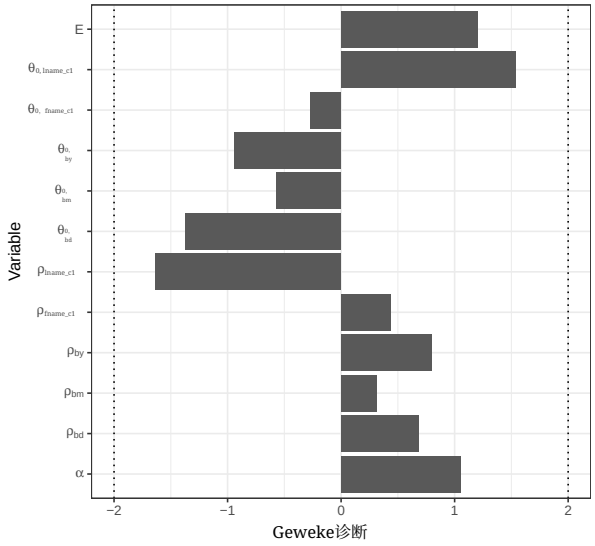


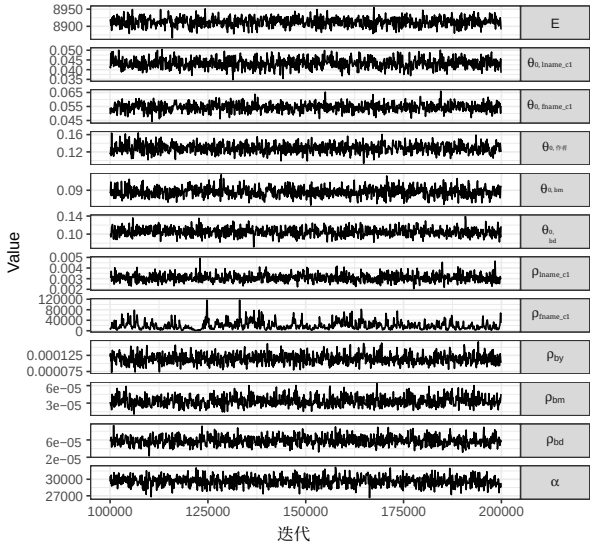
RLdata |PY |我们的方法



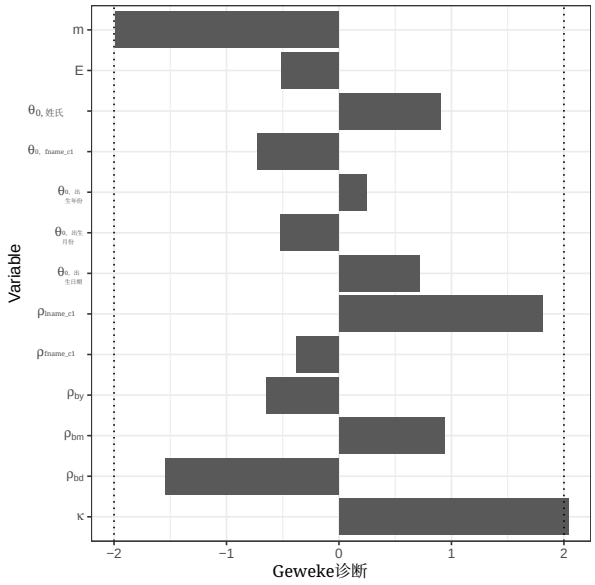


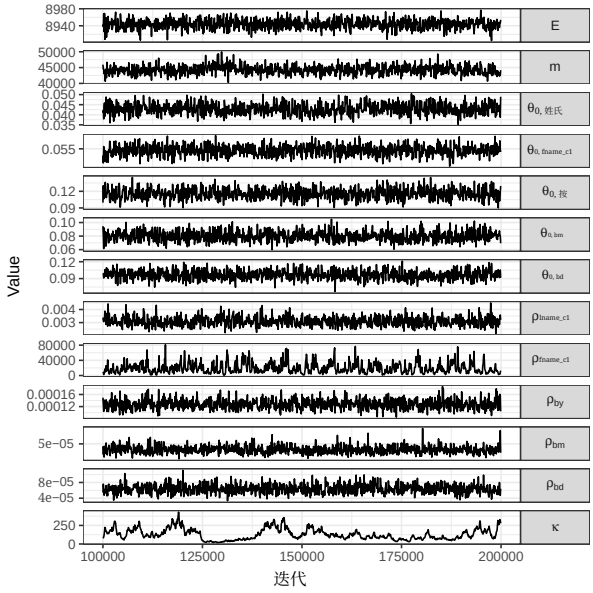
RLdata |Ewens |我们的方法



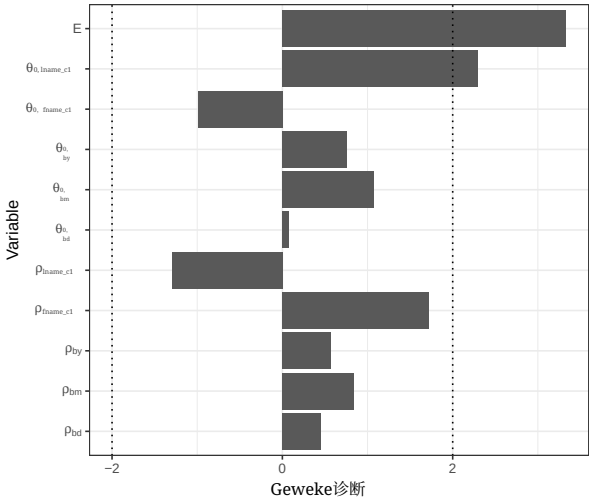


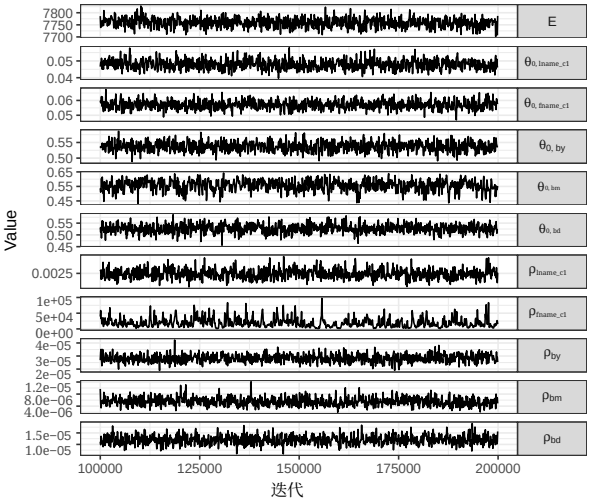
RLdata |GenCoupon | 我们的方法



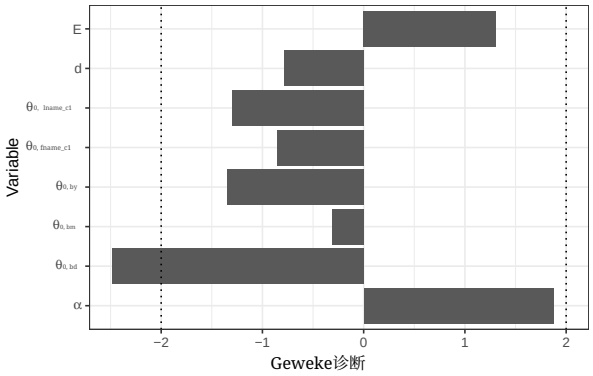


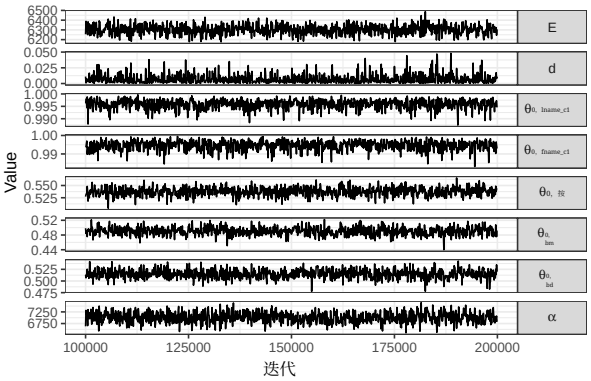
RLdata | Coupon | 我们的方法



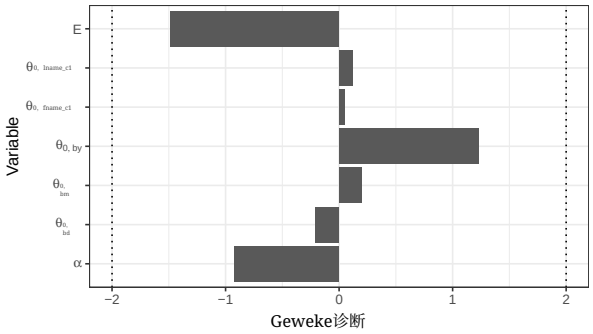


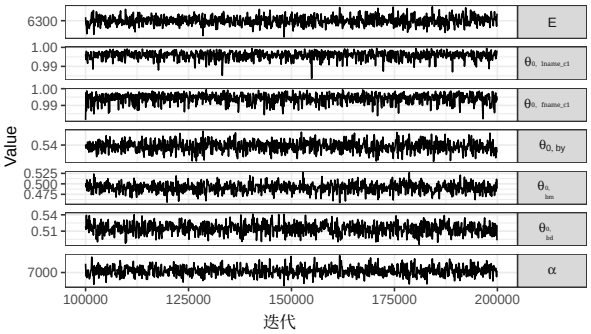
RLdata | PY |blink



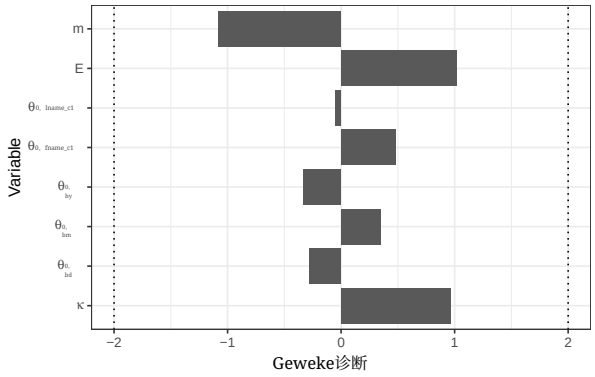


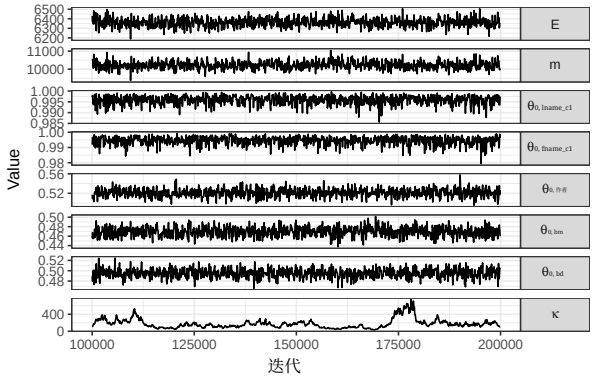
RLdata | Ewens |blink



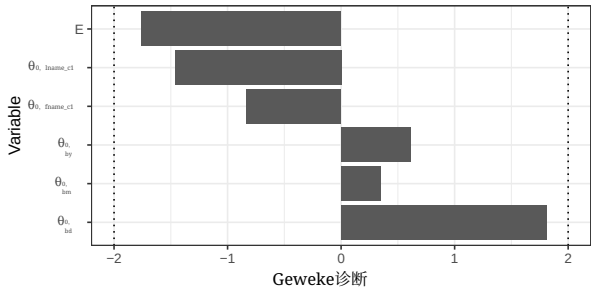


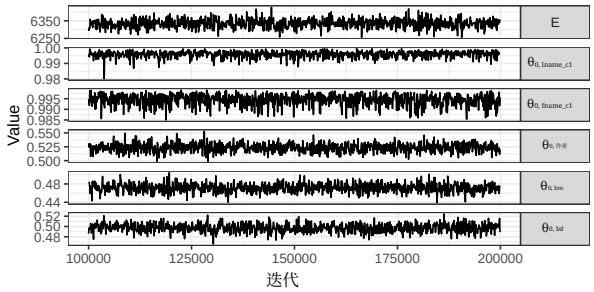
RLdata | GenCoupon | blink



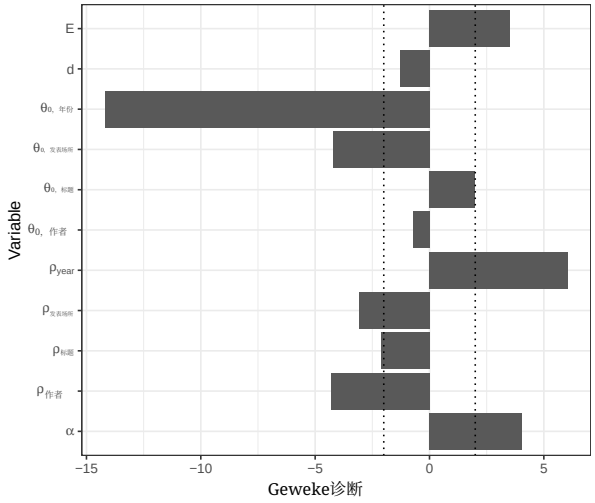


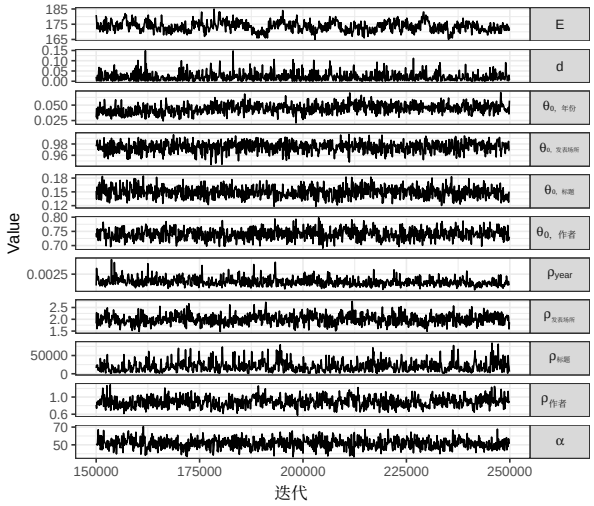
RLdata | Coupon |blink



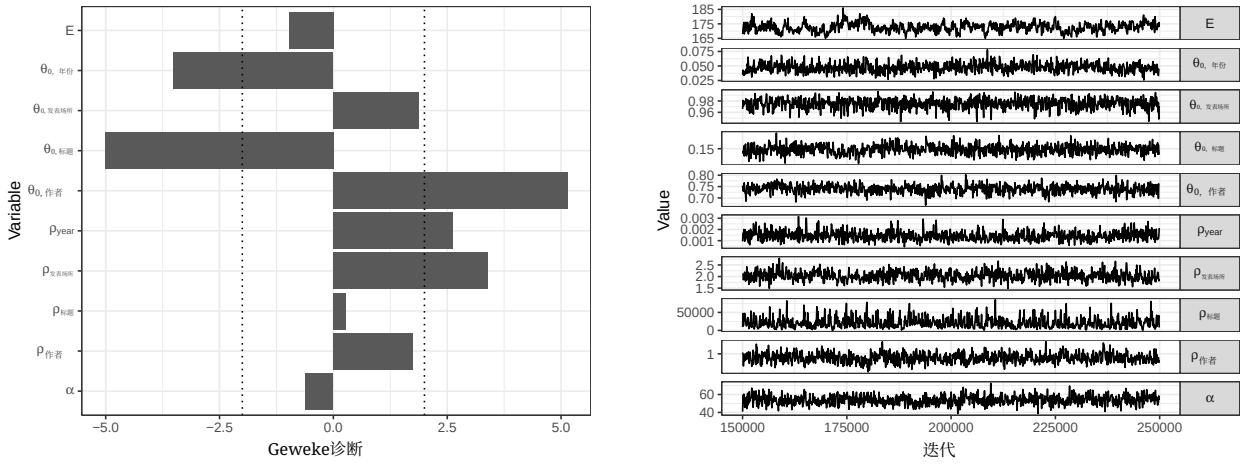


cora |PY |我们的方法

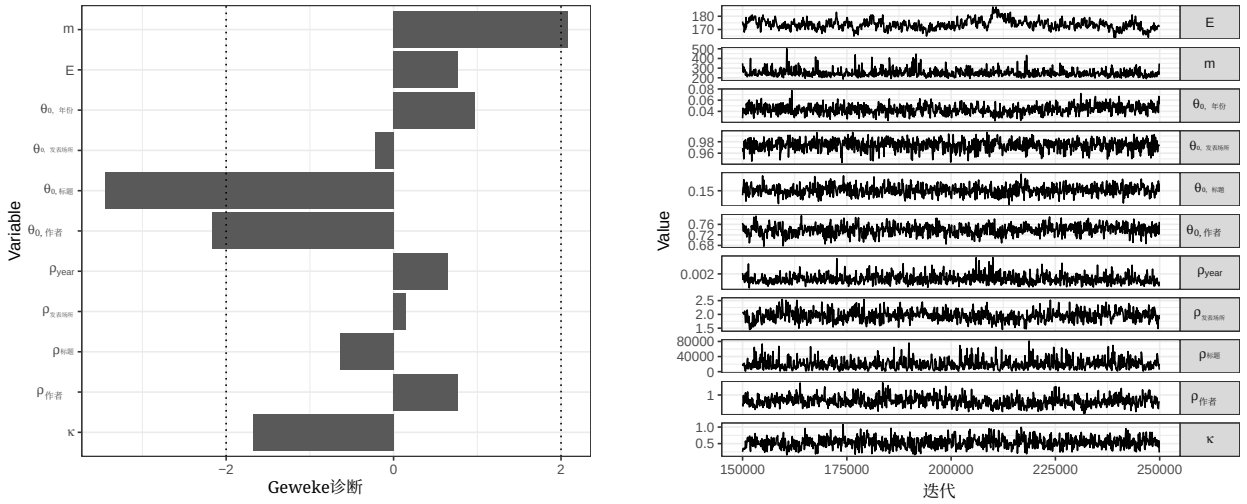




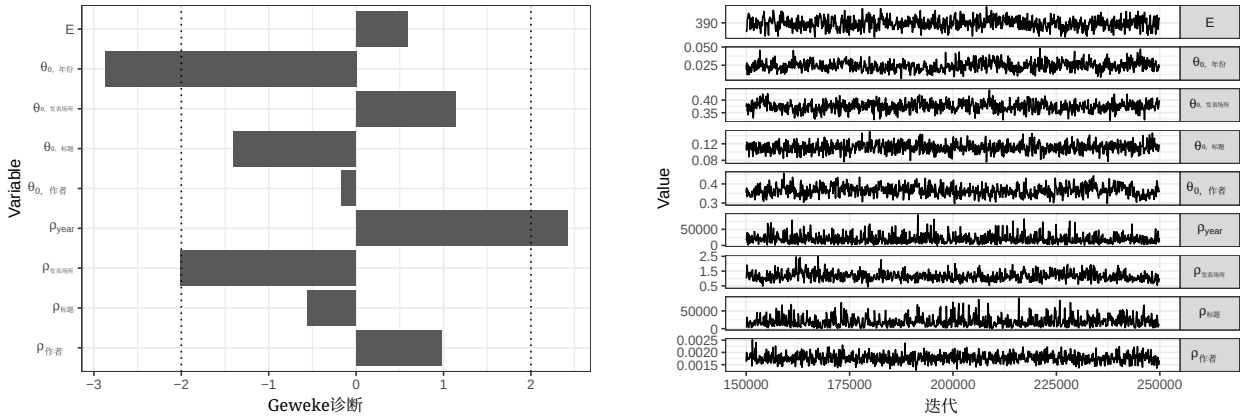
cora | Ewens | 我们的方法



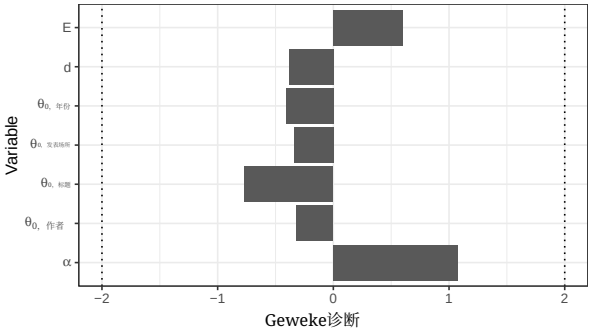
cora | GenCoupon | 我们的方法

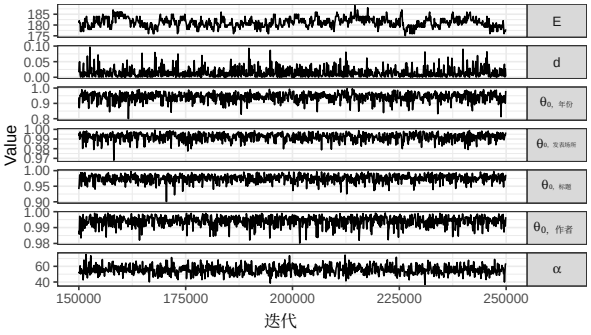


cora | Coupon | 我们的方法

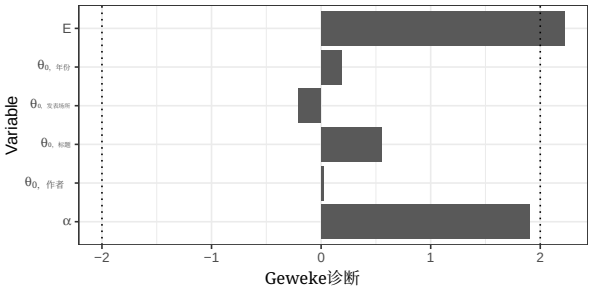


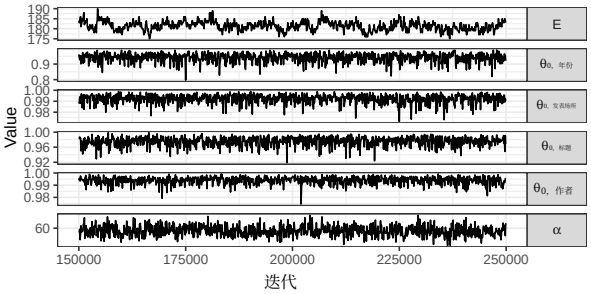
cora |PY |blink



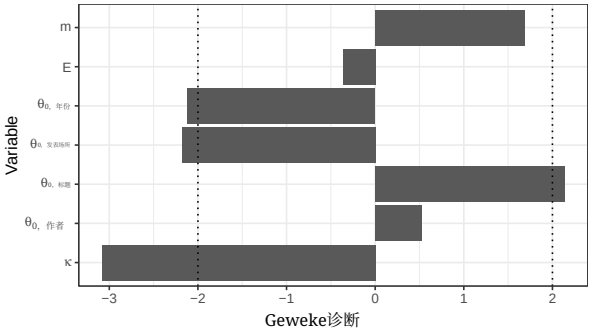


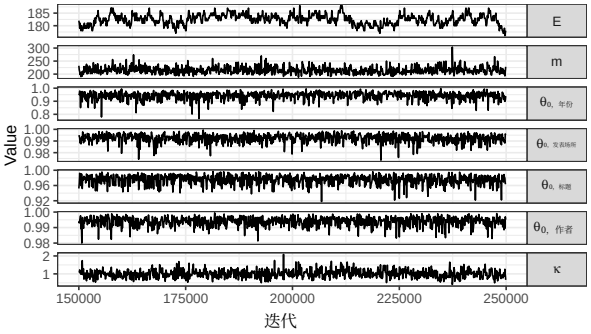
cora |Ewens |blink



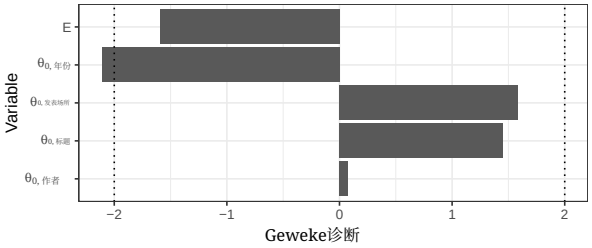


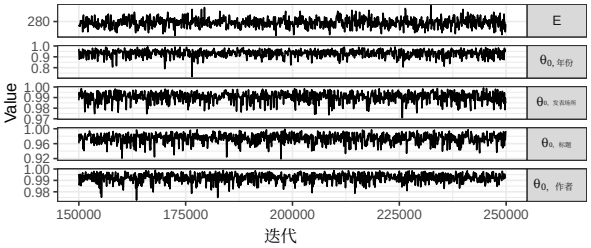
cora |GenCoupon | blink



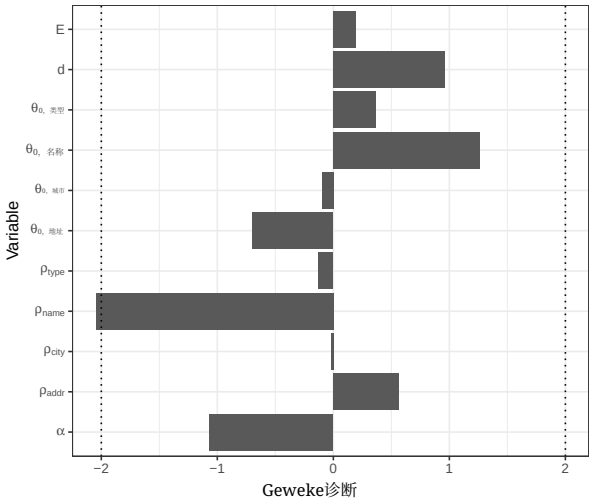


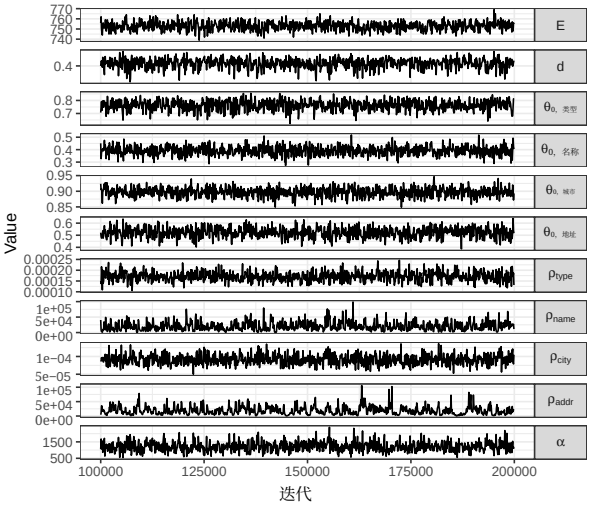
cora | Coupon | blink



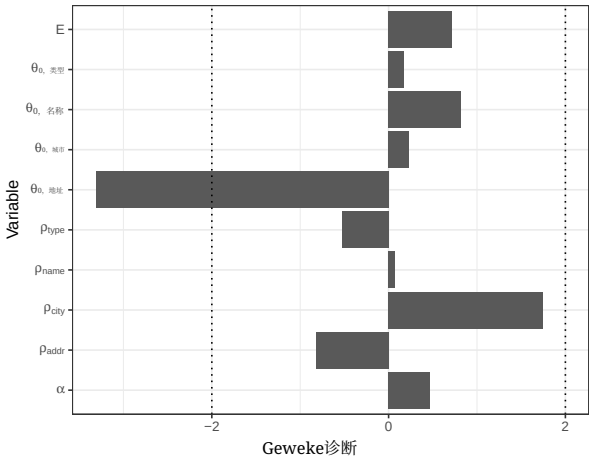


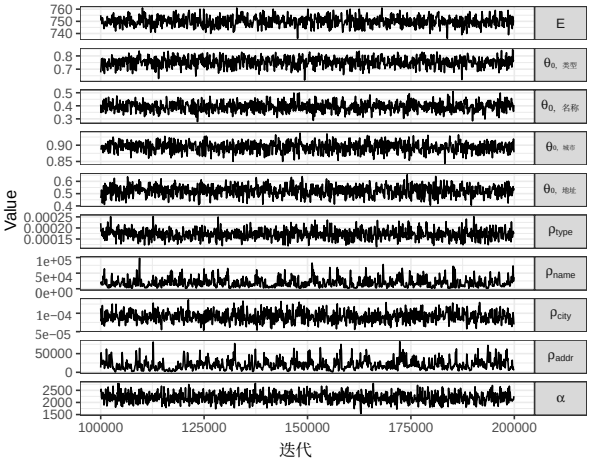
rest |PY |我们的方法



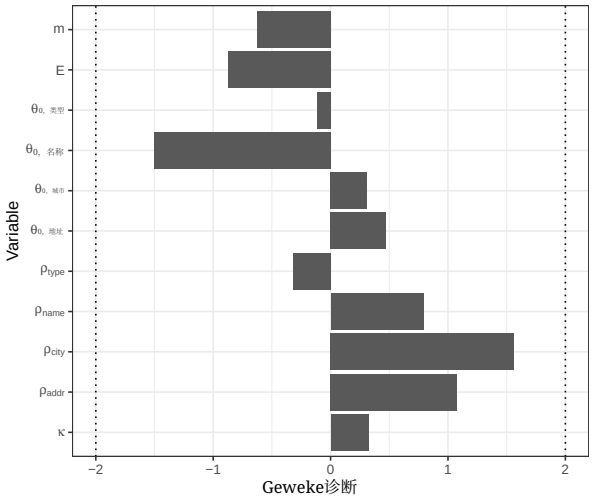


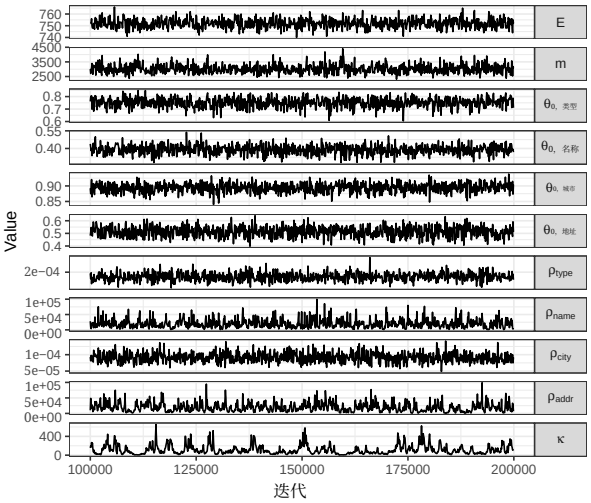
rest |Ewens |我们的方法



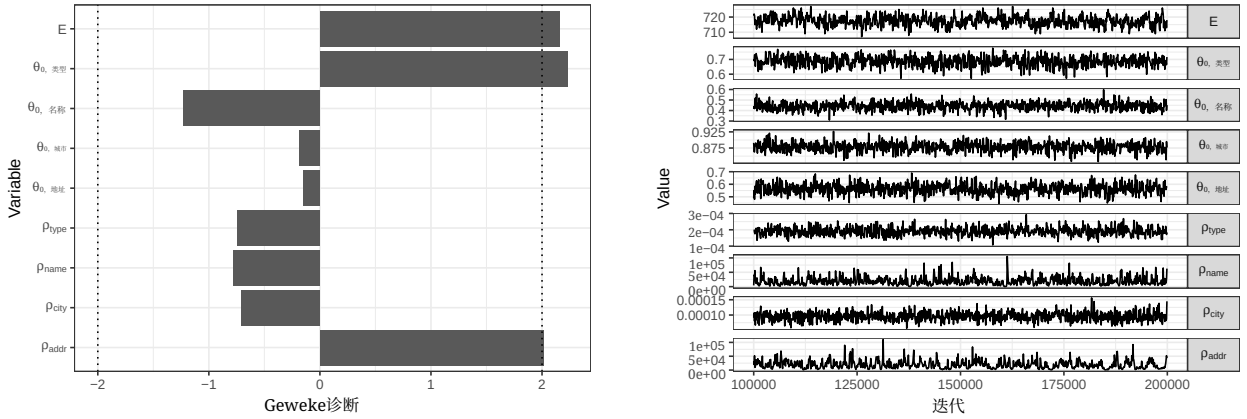


rest | GenCoupon |我们的方法

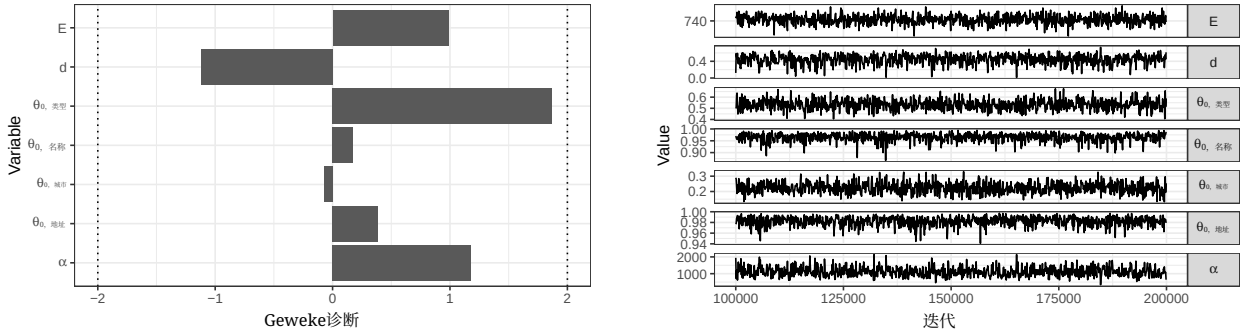




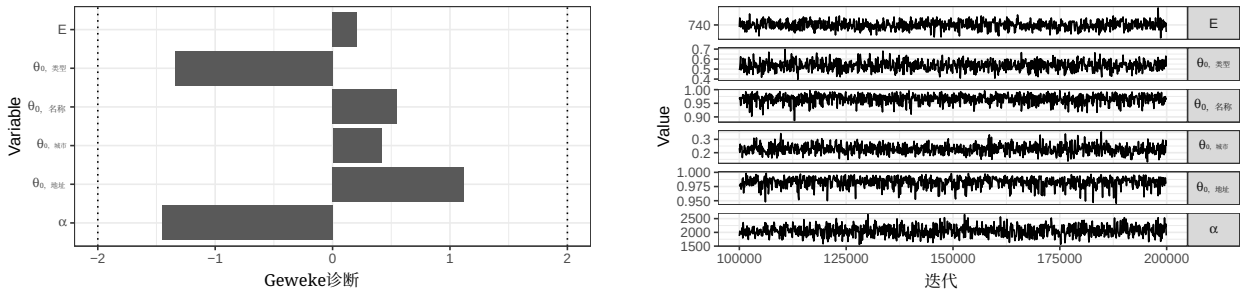
rest | Coupon | 我们的方法



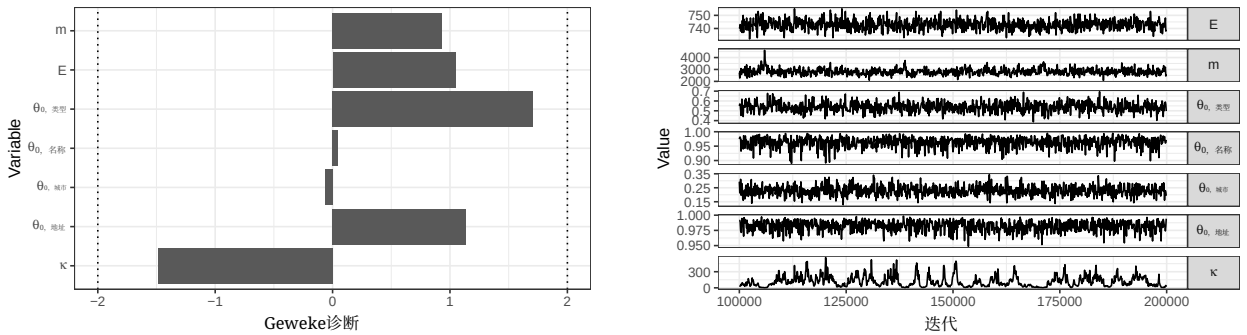
rest |PY |blink



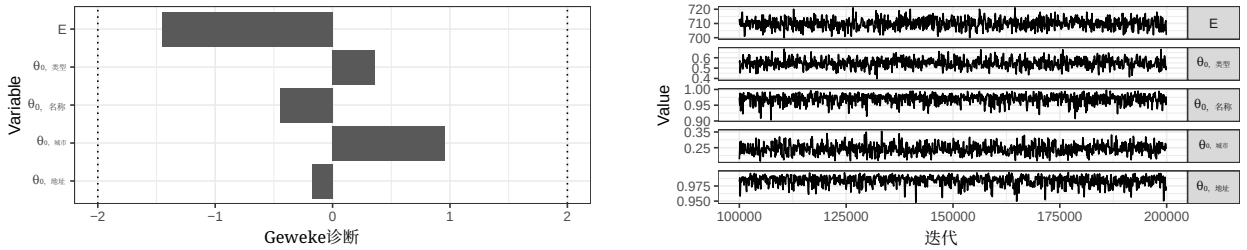
休息 | Ewens | blink



rest | GenCoupon | blink



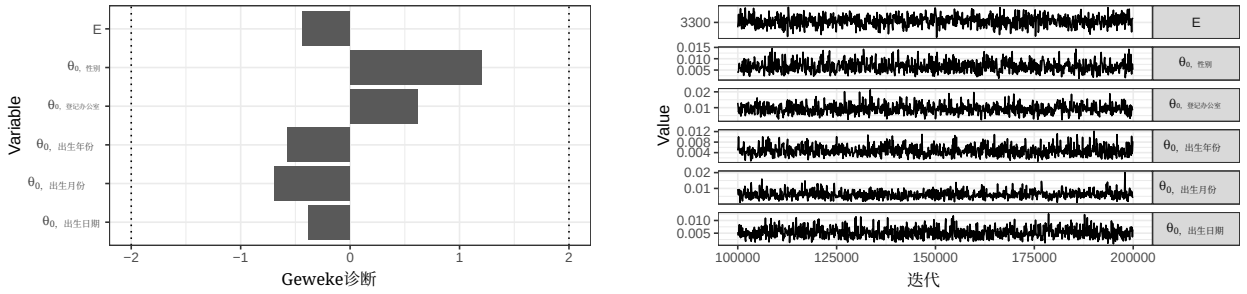
rest | Coupon | blink



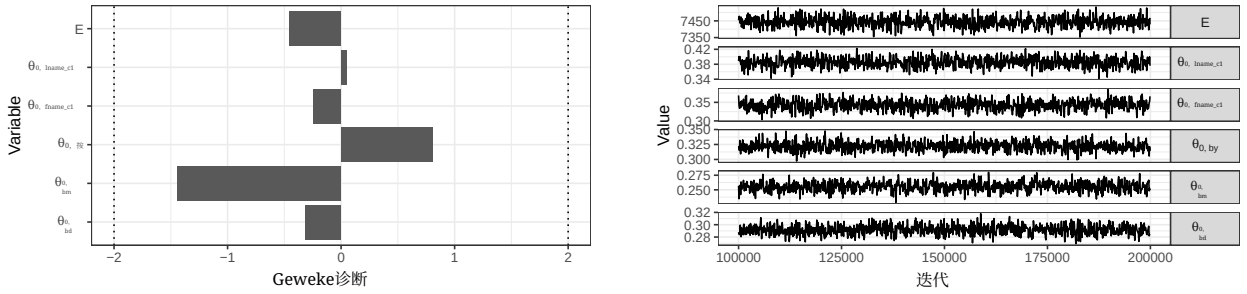
G.2 与基线模型的比较

这里展示第 5.4 节中拟合的模型的收敛诊断。对于每个数据集和模型，我们展示所选模型变量的 Geweke诊断图和轨迹图。每组图前均有一个格式为“数据集 | 模型”的标题。Geweke诊断图（左侧）在横轴上显示每个变量的 Z分数，在纵轴上显示变量。Z分数用于检验马尔可夫链前 10% 与最后 50% 部分的均值是否相等，通常预期处于范围 $[-2, 2]$（Geweke，1992）内。轨迹图（右侧）在横轴上显示链中每一步对应的变量值（变量标注于右侧面板）。请注意，变量 E 表示实例化实体的数量。我们将属性的整数索引替换为命名索引。例如，$\theta_{0,\text{city}}$ 表示名为“城市”的属性在来源 0 中的失真概率。

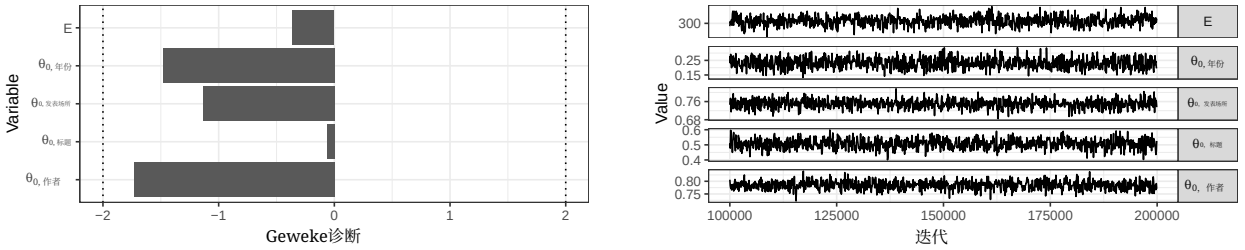
nltcs | blink



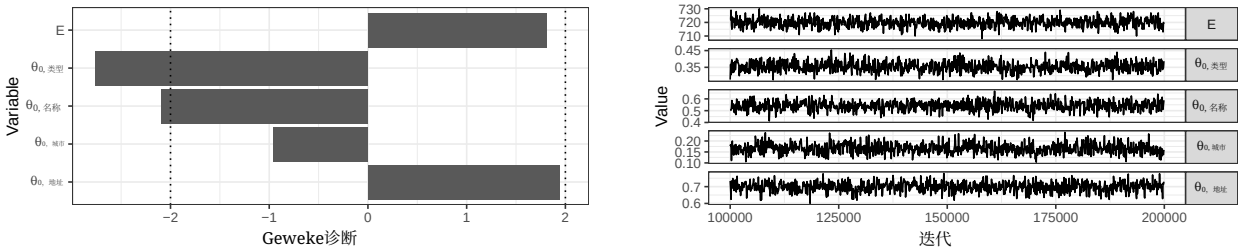
RLdata |blink



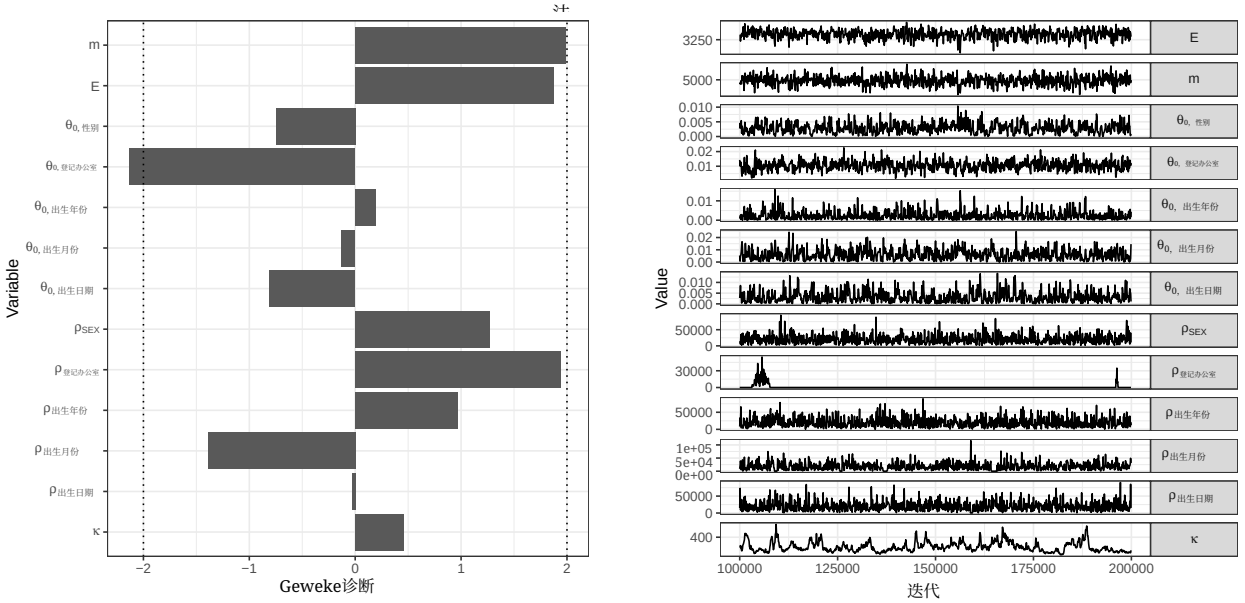
cora |blink



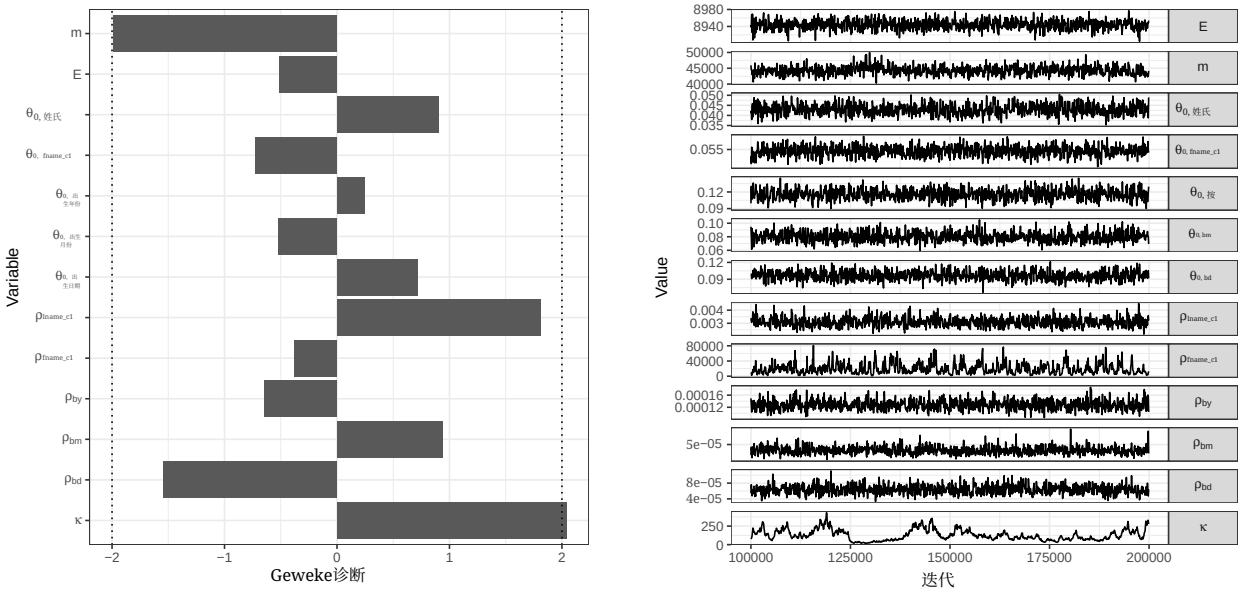
rest |blink



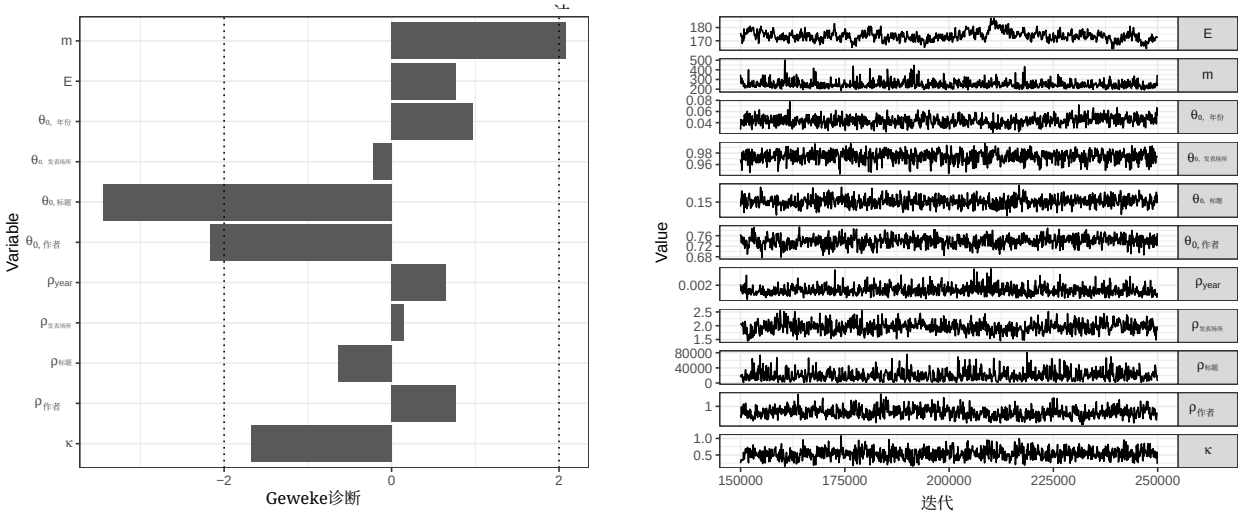
nltcs | 我们的方



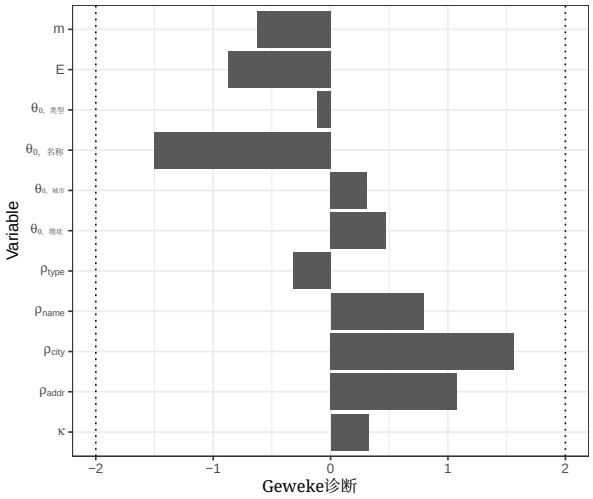
RLdata 我们的方法

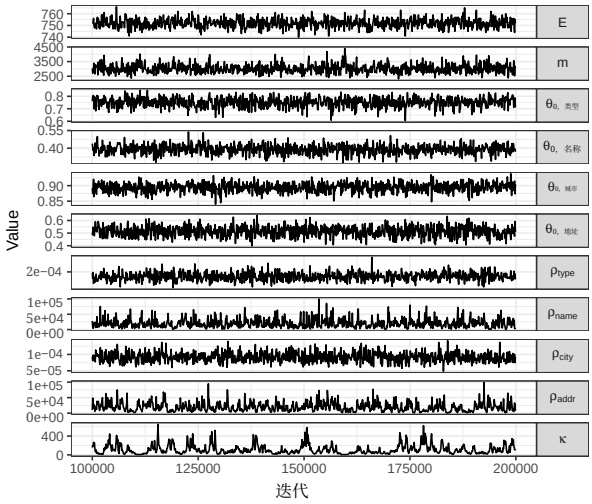


cora 我们的方

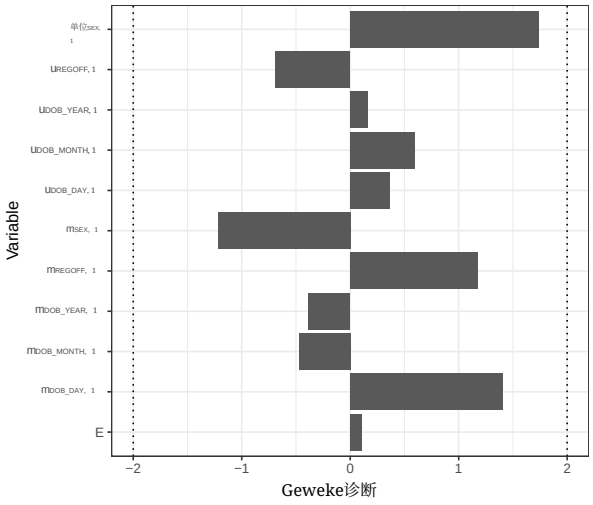


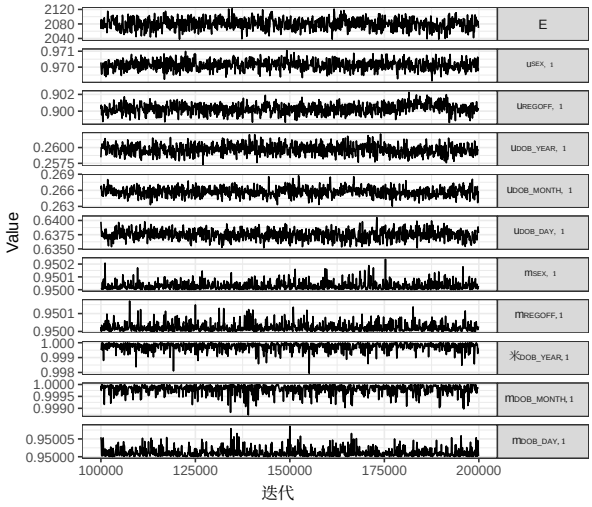
rest |我们的方



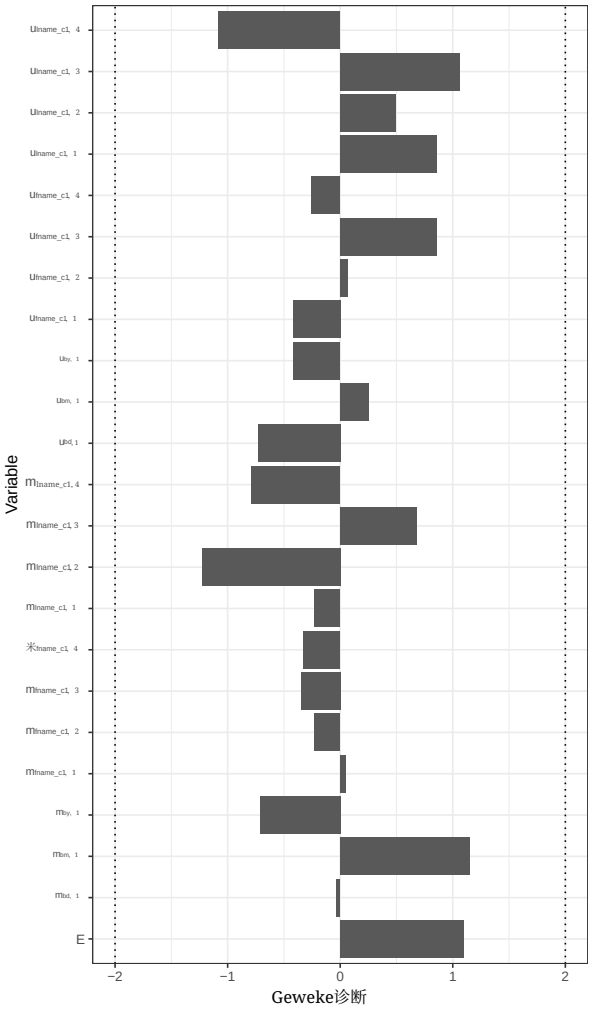


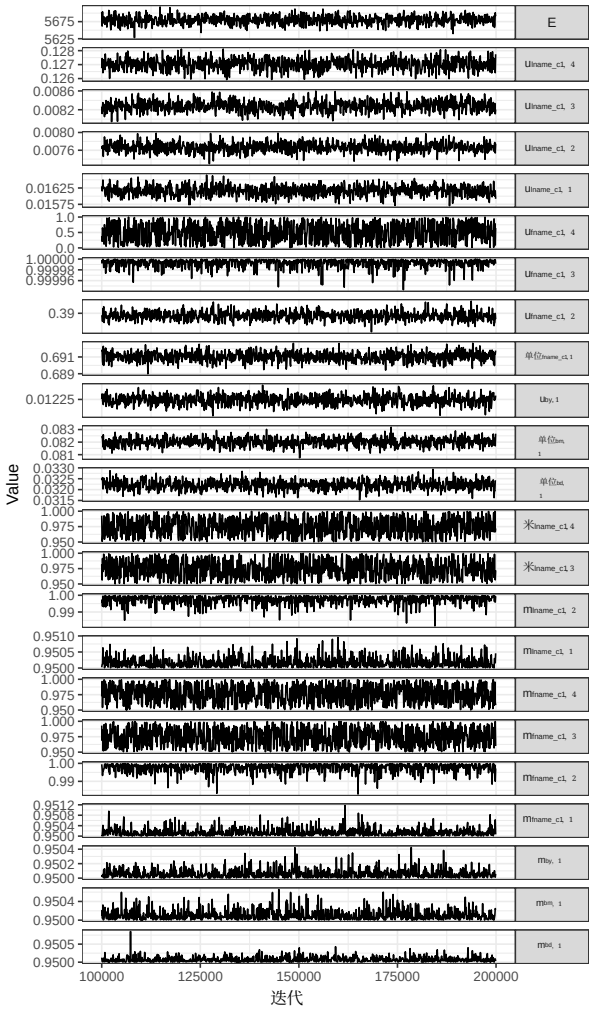
nltcs |Sadinle



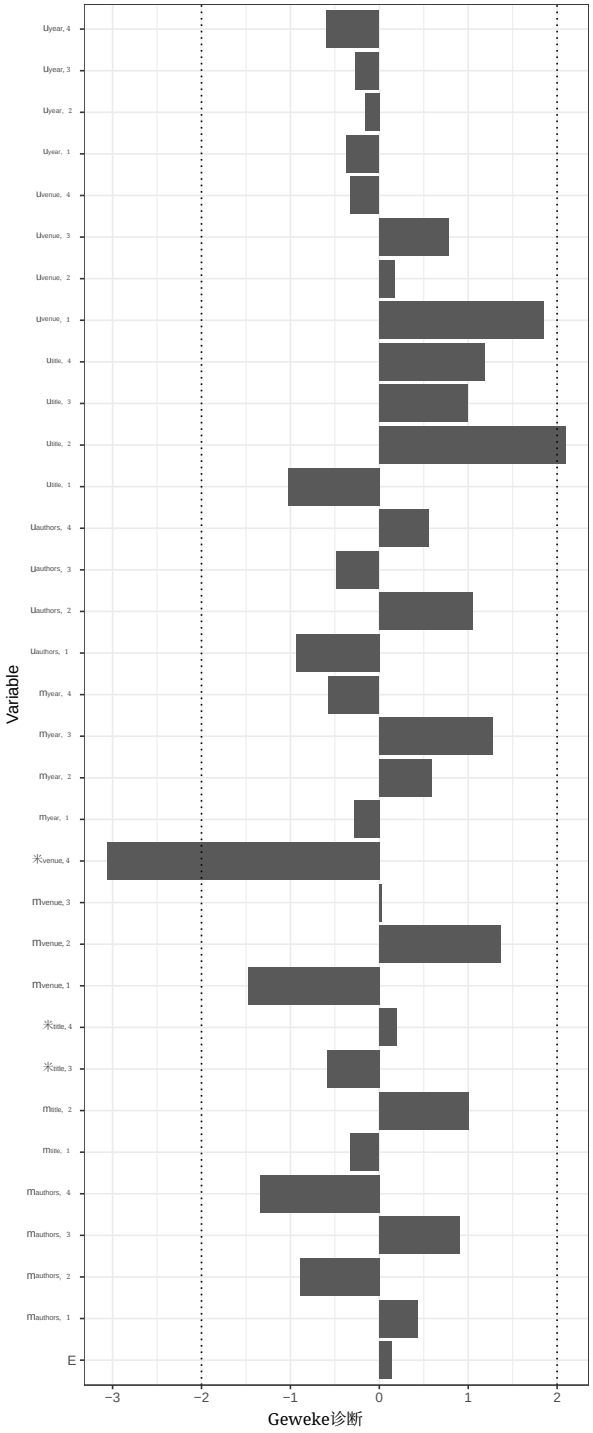


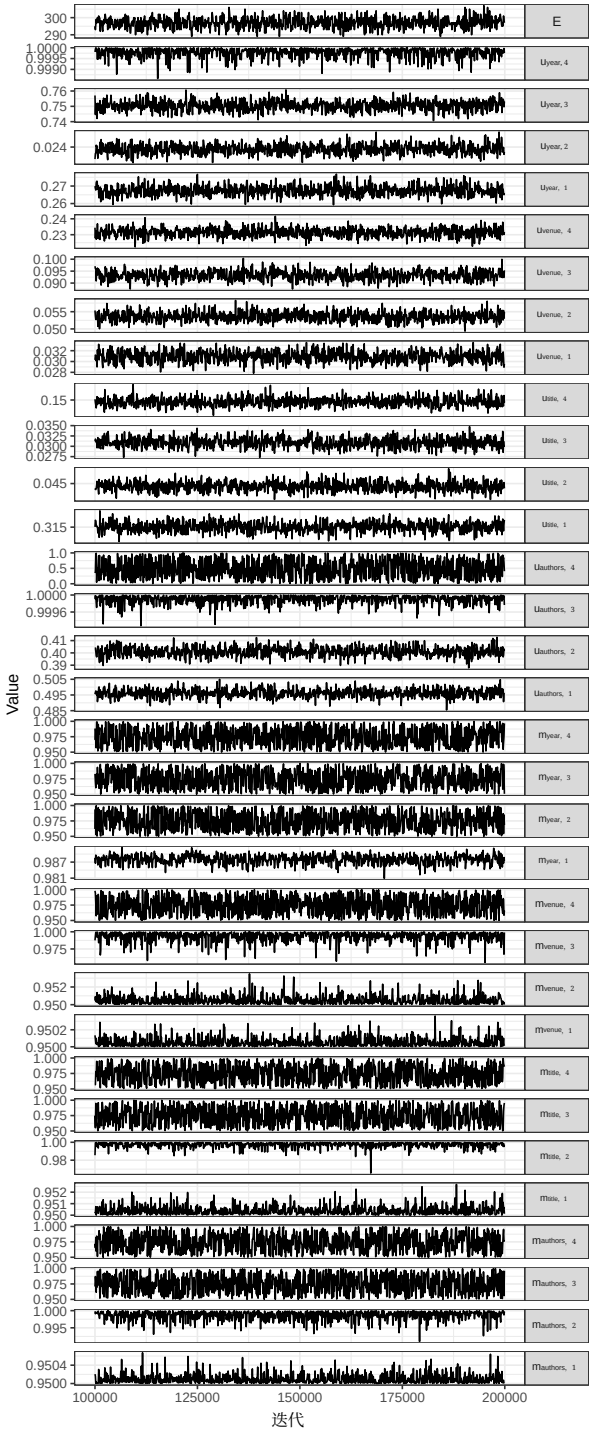
RLdata |Sadinle





cora |Sadinle





rest |Sadinle

