多系统估计模型的神经方法

Joseph Marsh* [1], Nathan A. Judd †[2]、Lax Chan[1,3], 和 Rowland G. Seymour [1]

[1]数学学院，伯明翰大学，英国[2]冲绳科学技术大学院大学，日本[3]皮埃蒙特东方大学经济与企业研究系`，意大利

arXiv:2601.05859v1 [stat.AP] 9 Jan 2026

摘要

利用多系统估计（MSE）估计隐匿总体规模，是定量社会学中的一项关键任务；然而，实际应用往往受到不完善的行政数据和计算资源限制的阻碍。出于隐私考虑，现实世界的数据集经常存在删失和缺失；而标准推断方法（如最大似然估计（MLE）和马尔可夫链蒙特卡罗（MCMC））在数据稀疏时可能会变得难以计算，或无法收敛。为解决这些局限，我们提出一种新颖的基于模拟的贝叶斯推断框架，采用神经贝叶斯估计器（NBE）和神经后验估计器（NPE）。这些神经方法经过摊销：训练完成后，便能即时提供计算高效的后验估计，因此非常适用于计算资源有限的安全研究环境。通过广泛的模拟研究，我们证明，神经估计器的准确度与MCMC相当，速度却快出数个数量级，并且能够稳健应对稀疏情形下传统采样器常见的收敛失败问题。我们将该方法应用于两个真实案例，估计英国现代奴隶制和英格兰东北部女性吸毒的流行率。

1 引言

对隐匿或难以接触群体的人口规模进行估计，一直是定量社会学面临的长期方法论挑战。隐匿总体是指因边缘化地位、参与受污名化或非法行为，或缺乏社会可见性而没有全面抽样框的群体。采用传统调查和人口普查方法估计隐匿总体规模面临重大障碍。准确的规模估计用途广泛，涵盖众多分支领域，并具有重要的政策和理论意义。例如，研究无家可归者的研究人员估算了与无家可归者相关的住院入院人数，以了解医院服务

*通讯作者

†此项工作是在英国伯明翰大学数学学院完成的。

供给情况（Luchenski 等人，2025）。在理解健康的社会决定因素时，准确估算非法药物使用者人数有助于制定干预计划（King 等人，2008）。同样，致力于了解现代奴隶制或人口贩运问题的社会学家和人口学家，也依靠人口规模估计来量化侵犯人权的规模并设计有效的干预措施（Silverman，2020）。其他应用还包括估算无证移民人口和无家可归者的规模，以及统计秘密政治运动的参与者人数；这些工作对于建立准确的社会理论和制定循证政策都至关重要。

多系统估计（MSE）是一种广泛采用的捕获—再捕获技术，通过比较一组不完整的列表，或称“系统”，来估计隐匿人口的规模；这些列表或系统所涵盖的对象均来自该总体（Bird and King, 2018）。这些列表可能包括来自警察、医院、社会服务机构或非政府组织的行政记录，不必是随机样本，通常属于便利样本。MSE的核心逻辑是分析列表之间的重叠程度，即出现在多个列表中的人数。通过观察各列表间的这些重叠模式，研究人员可以对系统之间的依赖关系进行统计建模。随后利用该模型预测唯一未观测到的数量：未被任何系统捕获的人数。最后，将这部分未观测个体的“隐形人数”加到所有列表中观测到的独立个体总数上，从而估计总体人口规模。

在统计学上，对MSE模型进行推断主要有两种范式。传统方法依赖最大似然估计，在给定一组模型假设（通常采用对数线性模型框架）的情况下，对未知人口规模（N）进行点估计。该方法旨在找到使观测数据最有可能出现的单一 N值。另一种方法是使用贝叶斯方法进行推断，这种方法已日益普及。它通常采用马尔可夫链蒙特卡罗（MCMC）技术来近似人口规模的完整后验分布。贝叶斯范式能够更全面地呈现不确定性，给出完整的后验分布形态，而不只是渐近置信区间。它尤其受到重视，因为能够以正式方式纳入先验信息，并且在处理模型不确定性和复杂情形（例如数据稀疏的情形）时具有灵活性。

尽管多系统估计很有用，但在社会学研究中实际应用时往往会受到两项重大挑战的阻碍。首先，社会学家能够获取的行政数据和观察数据往往并不完善。为避免识别出数据集中的个体，个体数据和计数可能会被删失或隐去。此外，数据还可能存在缺失观测，即未能收集到完整数据；使用冲突数据或现成数据时，这种情况尤为常见。这些数据质量问题可能导致难以可靠地估计列表依赖关系和交互项，而准确预测未观测总体又离不开这两者。

其次，这类敏感行政数据中，越来越多的数据只能通过安全或可信研究环境访问。尽管这些环境对于保护受访者机密性至关重要，但它们往往会限制可用的计算资源和软件（Kavianpour 等，2022）。对于前文所述的贝叶斯MCMC方法而言，这一计算瓶颈尤其棘手。虽然马尔可夫链蒙特卡罗在处理不确定性方面具有统计学上的吸引力，但其高昂的计算成本和漫长的运行时间，可能导致它在可信研究环境通常施加的限制下不切实际，甚至无法实现。King 等（2021）提出了唯一一种能够处理删失计数的MSE模型，该模型在开发最大似然估计时考虑了删失值的所有可能取值。然而，在

计算资源有限的环境中，这只能通过频率学派方法实现。

本文提出了两种基于模拟的技术——神经贝叶斯估计器（NBE）和神经后验估计器（NPE）——用于MSE模型的新型贝叶斯推断方法。基于模拟的推断为传统方法提供了一种强有力的替代方案，尤其适用于处理社会学数据所需的复杂模型。这些神经方法不依赖显式的似然计算（在存在审查数据或缺失数据时，似然计算可能难以执行或评估），而是通过在不同参数设定下模拟数据，并将模拟结果与观测数据进行比较，学习近似感兴趣的后验量。这些估计器的一项主要优势在于其摊销特性。神经网络训练完成后（训练是一个计算密集型步骤，可在不受限制的硬件上“离线”完成），便能针对任何给定的数据集几乎即时生成后验估计。这种“一次训练，多次推断”的能力与马尔可夫链蒙特卡罗（MCMC）形成鲜明对比，后者每次分析都需要重新进行耗时的计算。因此，NBE和NPE非常适合安全研究环境，因为它们能够在访问端仅需极少计算资源的情况下，提供稳健的贝叶斯推断。

这两种方法针对不同的推断目标。神经后验估计器利用神经网络，从模拟数据中直接学习整个后验分布的近似值。相比之下，NBE利用神经网络直接学习点估计（例如后验均值或中位数）。这些方法能够在高维或非规则情境下实现高效且灵活的推断，因此非常适用于存在缺失数据或删失数据的多系统估计模型等情况。通过使用根据模型生成的合成数据集训练网络，即使标准解析方法失效，也有可能获得后验估计。

此外，这种无似然方法解决了某些多系统估计情境中固有的一个根本性估计问题。最大似然估计等标准推断方法可能无法得出有限的人口规模估计。例如，如果在某些列表的重叠部分中没有观察到任何个体，就无法确定模型中的交互项，且似然函数不存在定义良好的最大值（更多详情请参见 Chan等人（2020））。相比之下，我们基于模拟的框架不依赖于最大化似然函数，而是直接学习从数据到所关注后验量的映射（例如，通过NBE获得点估计，或通过神经后验估计获得完整分布），从而确保即使在这些具有挑战性的稀疏数据情境下也能生成有效估计。

NBE 和神经后验估计（NPE）的摊销特性使其在实际应用中极为实用。只需承担一次性的初始训练成本，训练好的神经贝叶斯估计器便能几乎即时地对任何新数据进行推断。这非常适合数据频繁更新的多系统估计（MSE）研究，因为它无需重新运行计算密集型算法（如马尔可夫链蒙特卡罗），即可快速重新估计。这让更多人能够使用先进的贝叶斯方法，也免去了用户自行实现复杂推断工具的需要。对于在英国 SafePods 等安全环境中工作的研究人员而言，这一点尤为重要，因为计算能力有限和软件使用限制可能构成重大障碍。我们的工具提供了一种快速、高效且安全的方法，帮助人们从敏感数据中获取洞见。

本文组织如下。我们首先介绍MSE模型，然后提出MSE神经估计器，并说明其如何应用于存在删失数据的情形。接着，我们通过模拟研究评估方法的效果，然后将其应用于真实数据。最后，我们对方法及其局限性进行讨论。

2 多系统估计

考虑一个规模为 N的总体，其中我们无法观测整个总体，并希望估计 N的值。令 K 表示列表的数量，并假设总体中的每个个体都独立地出现在每个列表上或不出现在列表上。令 $\mathcal{X} = \{0,1\}^K \setminus \{(0,\ 0,\ \ldots\ldots,\ 0)\}$ 表示总体中个体在 K 个列表上所有可能的非空捕获模式的集合。对于每个捕获模式 $\boldsymbol{x} \in \mathcal{X}$，令$N_{\boldsymbol{x}}$ 表示具有该模式的个体的观测计数。未观测计数 N_0对应于未被任何列表捕获的个体数量（即捕获模式$\{0,0,\ldots,0\}$），这正是我们希望估计的数量。

A coMSE 的常见模型是对数线性模型，其中我们对计数 $N_{\boldsymbol{x}}$ b 建模 y

$$N_{\boldsymbol{x}} \sim \text{Poisson}(\lambda_{\boldsymbol{x}}), \quad \text{for all } \boldsymbol{x} \in \mathcal{X},$$

并假设各计数彼此独立。该分布的率参数通过每个名单的主效应以及名单之间的交互作用来建模。其表示为

$$\log \lambda_{\boldsymbol{x}} = \alpha + \sum_{k \in \boldsymbol{x}} \beta_k + \sum_{\substack{k,\, l \in \boldsymbol{x} \\ k<l}} \gamma_{k,l},$$

其中 α是截距，$\boldsymbol{\beta} = (\beta_1,\ \ldots\ldots,\ \beta_K)$ 是每个名单的主效应向量，而$\boldsymbol{\gamma} = (\gamma_{1,\ 2},\ \ldots\ldots,\ \gamma_{K-1,\ K})$ 是每对名单的交互项向量。也可以纳入高阶交互作用，但在实际应用中，由于存在可识别性问题和过拟合，交互作用通常仅限于两两效应。因此，未被任何名单捕获的人数可通过以下模型估计

$$N_0 \sim \text{Poisson}(\lambda_0),$$

其中 $\log \lambda_0 = \alpha$ 且 $\mathbb{E}(N_0) = \exp(\alpha)$。在此模型下，预期总体规模由观测计数与未观测计数的估计值相加得出：

$$\mathbb{E}(N) = \sum_{\boldsymbol{x} \in \mathcal{X}} N_{\boldsymbol{x}} + \mathbb{E}(N_0).$$

尽管参数 α 是我们主要关注的对象，但其他参数有助于理解捕获过程的运作方式。正交互项表明，各列表更有可能捕获到相同的个体（例如，由于检测机制相同），而负交互项则意味着个体会回避某些列表，或列表之间缺乏重叠。

捕获模式 $\boldsymbol{x}$ 对似然函数的贡献为

$$f_{\boldsymbol{x}}(N_{\boldsymbol{x}} = n_{\boldsymbol{x}} \mid \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{e^{-\lambda_{\boldsymbol{x}}} \lambda_{\boldsymbol{x}}^{n_{\boldsymbol{x}}}}{n_{\boldsymbol{x}}!},$$

观测数据似然函数则是所有可能的非空捕获模式对应的这些贡献的乘积，以及观测数据向量 $\boldsymbol{n}$

$$f(\boldsymbol{n} \mid \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{\boldsymbol{x} \in \mathcal{X}} f_{\boldsymbol{x}}(N_{\boldsymbol{x}} = n_{\boldsymbol{x}} \mid \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}). \tag{1}$$

为模型参数设定相互独立的先验分布，根据贝叶斯定理，后验分布为

$$\pi(\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{n}) \propto f(\boldsymbol{n} \mid \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}) \pi(\alpha) \pi(\boldsymbol{\beta}) \pi(\boldsymbol{\gamma}).$$

在传统的多系统估计应用中，拟合该模型需要列表重叠的完整数据。然而，在许多现实场景中，某些捕获模式 $\boldsymbol{x}$ 可能存在区间删失或缺失的情况，这可能是随机发生的，也可能是隐私约束所致，因此似然函数的计算可能更加复杂。正如我们将在下一节中所述，这促使我们采用基于模拟的方法。

3 多系统估计模型的神经贝叶斯估计

我们提出一种摊销贝叶斯方法，以取代传统的马尔可夫链蒙特卡罗或优化方法。我们训练神经网络，以近似特定的后验汇总统计量（神经贝叶斯估计）或完整的后验密度（神经后验估计）。具体做法是从先验分布中抽取大量样本，并基于这些先验抽样模拟各列表的捕获计数。我们首先说明如何将神经方法应用于多系统估计模型，然后再介绍删失数据。

3.1 神经贝叶斯估计器

$\hat{}$ 贝叶斯估计量 $\boldsymbol{\theta} = \boldsymbol{\theta}(\mathbf{n})$，正式定义为在后验分布 $\pi(\boldsymbol{\theta} \mid \mathbf{n})$ 上，使所选损失函数 $L(\boldsymbol{\theta}, \boldsymbol{\theta}(\mathbf{n}))$ 的期望最小的估计器。损失函数的选择决定了我们要估计的后验分布的性质。

对于稳健的点估计，我们的目标是后验中位数；在绝对误差（L_1）损失函数下，它是最优的贝叶斯估计量。因此，最优估计器可表示为

$$\hat{\boldsymbol{\theta}}^{*}(\mathbf{n}) = \arg\min_{\hat{\boldsymbol{\theta}}(\mathbf{n})} \mathbb{E}_{\boldsymbol{\theta}\sim\pi(\boldsymbol{\theta}|\boldsymbol{n})}[|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\mathbf{n})|]. \tag{2}$$

在似然复杂或涉及删失计数数据的应用中，我们无法直接计算完成此优化所需的后验期望。因此，我们使用 NBE 近似贝叶斯估计量。NBE 是一个由权重 ϕ 参数化的神经网络，经过训练后可近似从样本空间到参数空间的映射。

$L\boldsymbol{\theta}_{\hat{\phi}}(\hat{\mathbf{n}})$ 是一个权重为 ϕ 的神经网络。Sainsbury-Dale 等人（2024）证明，神经网络 $\boldsymbol{\theta}_{\phi^*}(\mathbf{n})$ 可以近似贝叶斯估计量 $\boldsymbol{\theta}(\mathbf{n})$，其中权重定义如下：

$$\phi^* \equiv \arg\min_{\phi} \frac{1}{M} \sum_{m=1}^{M} |\boldsymbol{\theta}^{(m)} - \hat{\boldsymbol{\theta}}_{\phi}(\boldsymbol{n}^{(m)})|,$$

其中，M 是模拟训练样本的数量。每个样本包含一个从先验 $\pi(\boldsymbol{\theta}^{(m)})$ 中抽取的参数向量 $\boldsymbol{\theta}^{(m)} = (\alpha^{(m)}, \boldsymbol{\beta}^{(m)}, \boldsymbol{\gamma}^{(m)})$，以及一个根据方程（1）中的概率质量函数抽取的相应模拟数据集 $\mathbf{n}^{(m)}$ 。

通过在足够大且多样化的模拟数据集上最小化 L_1 损失，网络便能学会针对任意给定的 $\mathbf{n}$近似后验中位数。训练完成后，推断过程即可摊销：将估计器应用于新的观测数据集$\mathbf{n}_{\text{obs}}$ 时，只需对网络进行一次快速的前向传播，即可得到 $\boldsymbol{\theta} = \boldsymbol{\theta}_{\phi^*}(\mathbf{n}_{\text{obs}})$。

此外，该框架的一项关键优势在于，它能够自然地扩展到不确定性量化。通过将绝对误差损失替换为分位数损失函数，L_τ，我们可以训练神经网络，使其针对任意给定的 τ阶后验分位数。对于任意单个参数 $\theta \in \boldsymbol{\theta}$，分位数损失定义为

$$L_\tau(\theta, \hat{\theta}) = (\hat{\theta} - \theta)\left(\mathbb{1}_{\{\hat{\theta} > \theta\}} - \tau\right), \quad \tau \in (0, 1),$$

ˆ，其中 1 是标准指示函数：当 $\theta > \theta$ 时等于 1，否则为 0（Sainsbury- Dale 等，2025）。已有研究表明，在联合损失 $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \sum_{k=1}^{K} L_\tau(\theta_k, \hat{\theta}_k)$下，估计器是边缘后验量的向量。因此，我们可以通过训练 NBE 来近似 $\tau = 0.025$ 和 $\tau = 0.975$ 分位数，从而构建 95% 可信区间。

3.2 神经后验估计器

NBE 通过最小化选定的损失函数，学习近似特定的后验量，例如中位数或分位数；而另一种更全面的方法是 NPE。NPE 的目标不是学习点估计，而是直接近似整个后验分布 $\pi(\boldsymbol{\theta} \mid \mathbf{n})$。

在此框架中，我们使用条件密度估计器 $q_\phi(\boldsymbol{\theta} \mid \mathbf{n})$ 来建模后验分布，其中 ϕ仍表示网络权重。该架构由两个主要部分组成。首先，将观测数据 $\mathbf{n}$输入一个权重为 ϕ_1 的多层感知机（MLP），生成固定大小、维度为 d^* 的嵌入向量 $\mathbf{s} = h_{\phi_1}(\mathbf{n})$。这个 MLP 充当摘要网络，旨在从原始且可能维度很高的计数数据中提取并学习有信息量的摘要统计量。

其次，使用权重为 ϕ_2 的条件归一化流来建模参数 $\boldsymbol{\theta}$ 与简单基分布（例如标准多元高斯分布）之间的密度 $q_{\phi_2}(\boldsymbol{\theta} \mid \mathbf{s})$，其中映射以摘要向量$\mathbf{s}$为条件。该映射通过堆叠一系列可逆仿射耦合块构建而成。

与 NBE 方法类似，流模型使用从先验分布和生成模型中抽取的 M 参数与模拟数据对 $\{\boldsymbol{\theta}^{(m)}, \mathbf{n}^{(m)}\}_{m=1}^{M}$ 进行训练。不过，网络的组合权重 $\phi = (\phi_1, \phi_2)$ 并非通过最小化类似 L_1 的损失来优化，而是通过最大化给定模拟数据时真实参数的对数概率来优化。这相当于在训练批次上最小化负对数似然损失：

$$\phi^* = \arg\min_{\phi} -\frac{1}{M} \sum_{m=1}^{M} \log q_\phi(\boldsymbol{\theta}^{(m)} \mid \mathbf{n}^{(m)}).$$

网络训练完成后，我们将观测数据$\mathbf{n}_{\text{obs}}$ 输入摘要网络，以计算嵌入 $\mathbf{s}_{\text{obs}} = h_{\phi_1^*}(\mathbf{n}_{\text{obs}})$。随后使用该摘要向量对归一化流进行条件化，从而得到完全确定的近似后验分布 $q_{\phi^*}(\boldsymbol{\theta} \mid \mathbf{s}_{\text{obs}})$。

通过这一单一的训练后近似模型，我们可以完整地刻画后验分布。近似后验分布通常是复杂的高维对象，而任何感兴趣的后验量都需要在参数空间上进行积分，这通常无法通过解析方法求解。尽管如此，根据构造，只需先从基分布 $\mathbf{z} \sim p(\mathbf{z})$ 中采样，再应用逆变换 $f_{\phi^*}^{-1}(\mathbf{z} \mid \mathbf{s}_{\text{obs}})$，即可轻松从这些分布中采样。重复这一过程大量次后，我们便能从近似后验中获得样本，进而计算均值、均值、分位数等量的蒙特卡洛估计。

神经后验估计的另一个优点是，可以在推断后通过后验预测检验进行模型评估。更确切地说，后验预测分布定义为

$$f(\mathbf{n}_{\text{rep}} \mid \mathbf{n}) = \int f(\mathbf{n}_{\text{rep}} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid \mathbf{n}) d\boldsymbol{\theta}.$$

因此，给定观测数据 $\mathbf{n}_{\text{obs}}$，我们可以从近似后验分布 $q_{\phi^*}(\boldsymbol{\theta} \mid \mathbf{s}_{\text{obs}}(\mathbf{n}_{\text{obs}}))$ 中生成 S 个样本（$\boldsymbol{\theta}^{(1)}$, ……, $\boldsymbol{\theta}^{(S)}$），然后利用这些样本从模型中生成新的复制数据集 $\mathbf{n}_{\text{rep}}^{(i)} \sim f(\mathbf{n}^{(i)} \mid \boldsymbol{\theta}^{(i)})$，其中 $i = 1$, ……, S。

最后，复制数据集的集合 $\{\mathbf{n}_{\text{rep}}^{(1)}, \ldots\ldots, \mathbf{n}_{\text{rep}}^{(S)}\}$ 是近似后验预测分布的样本。在这两种情况下，神经估计器一旦训练完成，便可提供摊销推断；将其应用于新的数据集时速度极快，因为这只需进行一次网络的前向传播。

3.3 方法的实现

对于 MSE 数据（由固定长度的非负计数组成的向量），多层感知机（MLP）架构是标准选择，因为这类数据不具备需要采用卷积神经网络、循环神经网络或 DeepSets 等专用架构的特定空间、时间或置换不变结构。

对于 NBE 网络和 NPE 编码器网络，我们均采用了全连接架构，由三个隐藏层组成，每层包含 256 个单元，并使用 ReLU 激活函数。对于 NBE 网络，我们谨慎选择输出层的激活函数，使其与各模型参数的后验分布支持集相匹配。具体而言，实值参数使用恒等激活函数，而取值范围受限于 $[a, b]$ 的参数则使用平移 Sigmoid 函数：

$$f(z) = a + (b - a) \cdot \frac{1}{1 + e^{-z}}.$$

对于 NPE 编码器网络，输出层使用恒等激活函数，宽度为 $d^* = 128$。为确保从近似后验生成的样本符合后验分布的支持集，我们采用拒绝采样方案。

我们使用 Julia 中的 NeuralEstimators.jl 软件包实现了这两种框架（Sainsbury-Dale 等人，2024 年）；算法 1 详细说明了完整的模拟和训练流程。该算法具有通用性，训练后的网络之间最终的差异取决于底层架构和损失函数 $L(\boldsymbol{\theta}, \boldsymbol{\theta})$，对于 NBE 方法和负对数似然方法，分别为 L_1 和 L_τ。

训练采用在线模拟的方式进行，以避免存储海量数据，并通过在每个训练轮次中从先验中模拟新的参数、从模型中模拟新的数据，增强估计器的稳健性。

每个训练轮次都会使用固定的验证集评估性能，以计算验证风险。随后采用早停策略：当验证风险趋于平稳或上升时停止训练，从而确保最终的估计器能够有效泛化而不过拟合。网络权重通过 ADAM（自适应矩估计）优化器进行更新，该优化器会根据先前的梯度为每个权重单独高效地调整学习率。

在实践中，泊松数据模拟得到的计数数据可能非常大。这是因为总体规模较大时，泊松分布的方差也较大。为了提高神经后验估计的稳定性和准确性，我们在训练前对观测到的计数数据应用了对数变换。具体而言，我们使用了变换 $\log(1+n_{\boldsymbol{x}})$，既能稳定计数数据固有的方差，又能压缩较大的数值，从而减轻极端观测值的影响。若不进行这一调整，估计器很可能会退化为在整个参数空间中预测几乎恒定的数值。

此外，这种变换对于网络的反向传播过程也很重要。通过均衡输入特征的尺度，它可以防止较大的数据值导致梯度爆炸，从而确保训练期间的权重更新平稳可靠。这种稳定性使网络的优化器能够更快、更有效地收敛。

算法 1 用于多系统估计的神经估计器

1: **初始化：**

- M：训练模拟次数
- V：验证模拟次数
- T：最大训练轮次数
- ϕ：初始网络参数

• $\pi(\alpha, \beta, \gamma)$：**先验分布 2: 对于** $v = 1$ **到** V **执行 3: 抽取** $(\alpha(v)\, \beta(v)\, \gamma(v)) \sim \pi(\alpha, \beta, \gamma)$ **4: 生成计数** $n(v)_{\boldsymbol{x}} \sim f_{\boldsymbol{x}}(N_{\boldsymbol{x}} \mid \alpha(v)\beta(v)\gamma(v))$**，针对每个** $\boldsymbol{x} \in \mathcal{X}$ **生成，并将其存储为验证集 5: 结束循环 6: 对于** $t = 1$ **到** T **执行 7: 抽取** $(\alpha(m)\, \beta(m)\, \gamma(m)) \sim \pi(\alpha, \beta, \gamma)$**，针对** $m = 1, \ldots, M$ **8: 生成计数** $n(m)_{\boldsymbol{x}} \sim f_{\boldsymbol{x}}(N_{\boldsymbol{x}} \mid \alpha(m)\, \beta(m)\, \gamma(m))$**，针对每个** $\boldsymbol{x} \in \mathcal{X}, m = 1, \ldots, M$。ˆ **9: 评估风险** $L(\theta, \theta)$ **10: 如果满足早停标准，则 11: 跳出循环 12: 结束 if 13: 更新网络参数** ϕ **14: 结束循环**

3.4 使用删失计数数据进行估计

在完全观测情形下，数据$\mathbf{n}$包含所有非空捕获模式 $\boldsymbol{x} \in \mathcal{X}$的完整计数。对于存在删失数据的情形，我们假设某个区间 $[a, b]$内的计数被删失，以保护个人身份。因而，删失计数 $n'_{\boldsymbol{x}}$包含结构性零值，其中

$$n'_{\boldsymbol{x}} = \begin{cases} n_{\boldsymbol{x}} & \text{if } n_{\boldsymbol{x}} \notin [a, b] \\ -1 & \text{if } n_{\boldsymbol{x}} \in [a, b] \end{cases}.$$

为区分结构性零值和非结构性零值，我们引入删失（或掩码）向量 $\boldsymbol{c} \in \{0, 1\}^{K(K+1)/2}$；如果值 $n'_{\boldsymbol{x}}$ 已被删失，则 $c_{\boldsymbol{x}} = 1$ ，否则为 0。我们使用删失计数 $\boldsymbol{n}'$ 和删失向量 $\boldsymbol{c}$训练神经网络。实际上，该过程与算法 1 完全相同：在生成数据后应用掩码，而神经网络的输入现在是一个拼接向量（$\boldsymbol{n}'_{\boldsymbol{x}}$， $\boldsymbol{c}$）。

4 模拟研究

为了证明我们的方法能够恢复模型参数并估计未被任何名单捕获的人数，我们开展了多项模拟研究和敏感性分析。在本文余下部分，对于包含 K 个列表的模型，我们先验假设：

$$\alpha \sim U[1, 10]$$
$$\beta_k, \gamma_{lk} \sim N(0, 4^2),$$

对于 $k = 1, ..., K$ 和 $l = 1,..., k$。 β 和 γ 的先验分布均为无信息先验，而 α 的先验则设定为使隐藏总体规模的范围从 $e \approx 2.72$ 到 $e^{10} \approx 22{,}000$；对于多系统估计而言，这被认为是合理的上限。

为评估估计器的性能和稳健性，我们开展了一系列模拟研究。我们首先为数据生成和网络架构设定了基准情景。对于数据，我们假设 $K = 5$个列表，删失区间为$[0, 10]$。相应的基准 NBE 是一个全连接多层感知机，包含三个隐藏层，每层有 256 个神经元。NPE 的基准架构采用一个全连接多层感知机，其编码器网络包含三个隐藏层，每层有 256 个神经元，输出维度为 $d^* = 128$ 的摘要统计量，随后将其输入归一化流。

在此基准情景的基础上，我们通过逐一系统地改变这些参数进行了敏感性分析。我们探究了以下因素对性能的影响：(i) 列表数量，(ii) 删失水平，(iii) 隐藏层数量，以及 (iv) 每个隐藏层中的神经元数量。

对于每种配置，我们都训练了一个新的 NBE 和 NPE，并根据其准确恢复真实模拟截距值（α）的能力评估其性能。为了训练每个 NBE，我们在每个训练轮次从先验分布中模拟 $M = 10,\ 000$ 次抽样，每次抽样生成一个数据集。

为公平评估泛化能力，所有情景均使用同一组固定测试集，其中包含新的参数—数据实现。我们主要采用感兴趣的主要参数 α 的绝对百分比误差（APE）作为指标，以衡量训练后的估计器恢复真实模拟参数值的效果。具体而言，令 exp(ˆα) 表示推断的隐藏总体规模，则 APE 定义为

$$\text{APE} = 100 \left| \frac{e^{\alpha} - e^{\hat{\alpha}}}{e^{\alpha}} \right|.$$

对于所考察的每种列表数量 $K = \{3, 4, 5, 6, 10, 15\}$，我们从先验中抽取 10,000 个参数，并根据模型 (1) 生成相应的实现，将其用作每项敏感性分析的测试数据。

最后，为验证估计器的统计准确性，我们将这些估计值与通过马尔可夫链蒙特卡罗（MCMC）得到的估计值进行了基准比较。

4.1 对人口规模的敏感性

我们首先分析估计器对真实隐藏总体规模的敏感性，然后再改变任何超参数。我们使用针对 $K = 5$ 的估计器，在图1中绘制对数绝对百分比误差（log APE）随真实隐藏总体规模 exp(α) 变化的曲线。

对于两种估计器，我们观察到，在总体规模较小时，对数绝对百分比误差较高且波动更大 $N \leq 300$。这是预期结果，因为低计数模拟会生成稀疏数据，无法为精确估计提供足够的信息。然而，一旦隐匿总体超过约 300 人，神经后验估计的误差便趋于稳定，并且随着总体规模进一步增加而保持一致。

两种方法之间的主要差异在更大规模下显现出来。对于所有不超过约 10,000 人的隐匿总体，NBE 的绝对百分比误差与神经后验估计相近。超过这一阈值后，NBE 的表现开始下降，而神经后验估计成为更准确的方法。值得注意的是，在所测试的整个总体规模范围内，神经后验估计的误差表现都非常稳定，这表明它对隐匿总体的规模具有极强的稳健性。

这部分原因可能在于 NBE 能够对训练数据进行泛化。根据构造方式，数据由从先验中采样得到的参数集生成

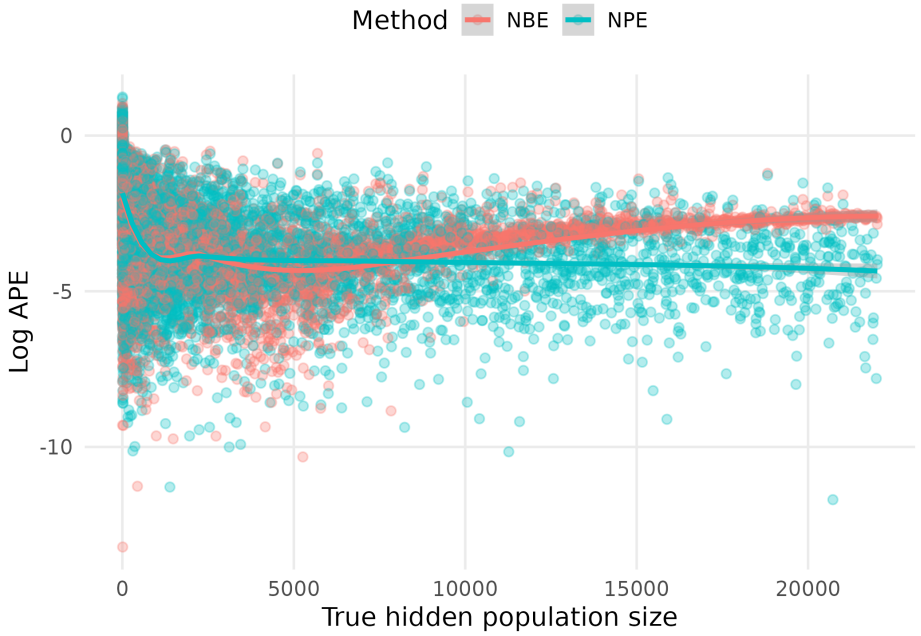

图1：神经贝叶斯估计器（NBE）和神经后验估计（NPE）的对数绝对百分比误差（Log APE）随真实隐藏总体规模变化的情况。趋势线表示每种估计器的平滑平均表现。

因此， α 从区间 [1,10]中均匀随机采样。然而，就由 $N_0 = \exp(\alpha)$ 定义的隐匿总体而言，可以证明， N 服从对数均匀分布，其概率密度函数为 $f_{N_0}(n_0) = \frac{1}{9y}$ （当 $n_0 \in [e, e^{10}]$ 时），其他情况下为零。因此，包含较高隐匿总体的参数集将少得多，估计器可能难以利用较少的训练样本有效学习从参数空间到样本空间的映射。

4.2 改变列表数量

接下来，我们研究在列表数量为 K 时恢复模型参数的能力。我们考察了方法在使用 $K = \{3, 4, 5, 6, 10\}$ 个列表时的表现。前四个数值对应于多系统估计（MSE）研究中列表数量的典型取值。最大值则不具典型性，因为很少会出现列表数量达到这一水平的情况，而且这还会导致需要估计大量参数。

图 2 展示了每个列表数量对应的 APE 箱线图。对于两种估计器，所有列表数量下的 APE 都较为相似；随着列表数量增加，APE 略有下降，但列表数量为 $K = 10$ 时的 NPE 模型除外。直观而言，这是因为更多列表提供了更多数据点，因此有更多信息可用于估计出现在每个线性预测器中的 α 参数。NPE 中 $K = 10$ 的例外情况可能是由于此时数据向量的维度为 $\dim(\mathbf{n}) = 2^{10} - 1$，因此神经网络的输入维度为$2 \times \dim(\mathbf{n})$；由于我们明确纳入了大小相同的删失向量，这对于编码网络而言可能是过大的输入。NPE 通过编码器网络对数据进行汇总，其输出向量的维度为 $d^* = 128$，显著小于输入维度，因此可能不足以学习数据的低维表示。

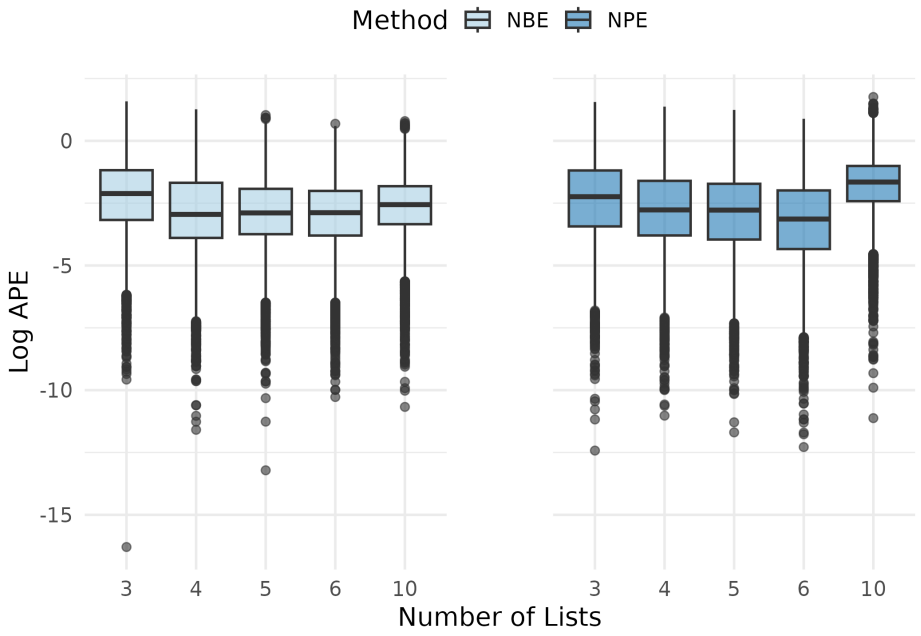

图 2：不同列表大小下 NBE 和 NPE 性能的比较。各面板分别展示了神经贝叶斯估计器（左）和神经后验估计（右）的对数绝对百分比误差，该误差根据测试数据集估计值（n=10,000）计算得出。

4.3 对神经元数量的敏感性

接下来，我们研究估计器对网络架构的敏感性，具体而言是对每个隐藏层中神经元数量的敏感性。我们通过训练使用每层 $\{8, 16, 32, 64, 128, 252\}$ 个神经元的不同架构模型来开展这项分析。需要注意的是，对于 NPE，该参数定义了其编码器网络中的神经元数量。

图 3 展示了每种配置对应的测试集绝对百分比误差（APE）。对于两个估计器，都呈现出明显趋势：通常而言，增加每层的神经元数量会降低 APE。这是预期结果，因为更宽的层能够提供更灵活、更具表达能力的网络，从而捕捉复杂的数据关系。这一发现支持我们最终选择 256 个神经元，这代表了一种切实可行的折中方案。虽然使用更多神经元（例如 512 个）可能会带来边际精度提升，但也会增加计算成本和过拟合风险，因此收益会逐渐递减。

4.4 对隐藏层数量的敏感性

我们探究了网络深度的影响，分别使用 $\{1, 2, 3, 4\}$ 个隐藏层训练估计器。图 4 所示结果表明，这些架构的性能总体保持一致，说明该问题并不需要异常深的模型。具有三个隐藏层的网络取得了最佳性能。由于不同隐藏层数量之间的对数绝对百分比误差似乎差异极小，因此采用三层架构能够在模型简约性与表达能力之间实现有效平衡，同时降低计算成本和过拟合风险。

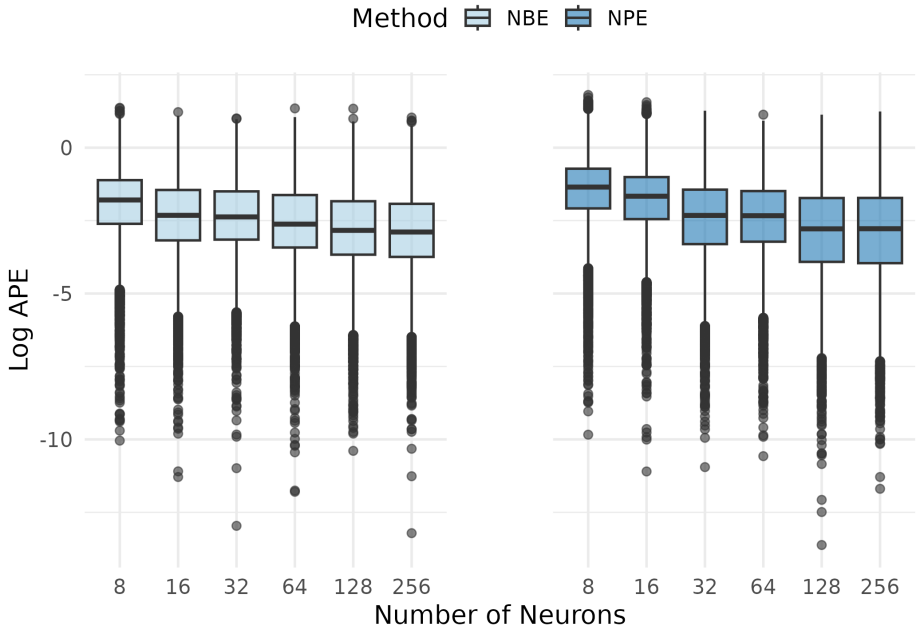

图 3：比较不同神经元数量下 NBE 与 NPE 的性能。各面板分别展示根据测试数据集估计结果（n =10,000）计算得出的神经贝叶斯估计器（左）和神经后验估计（右）的对数绝对百分比误差。

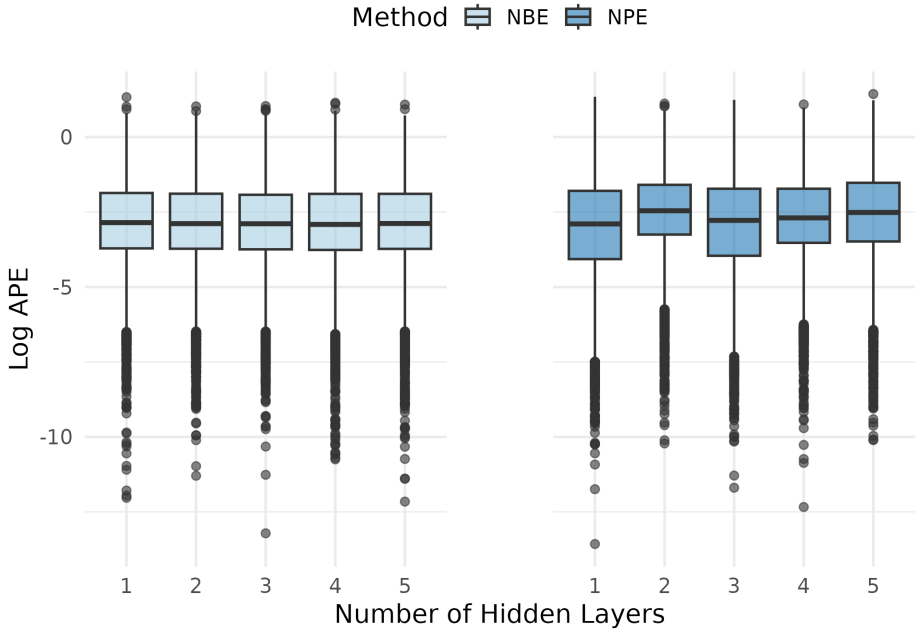

图 4：比较不同隐藏层数量下 NBE 与 NPE 的性能。各面板分别展示根据测试数据集估计结果（n =10,000）计算得出的神经贝叶斯估计器（左）和神经后验估计（右）的对数绝对百分比误差。

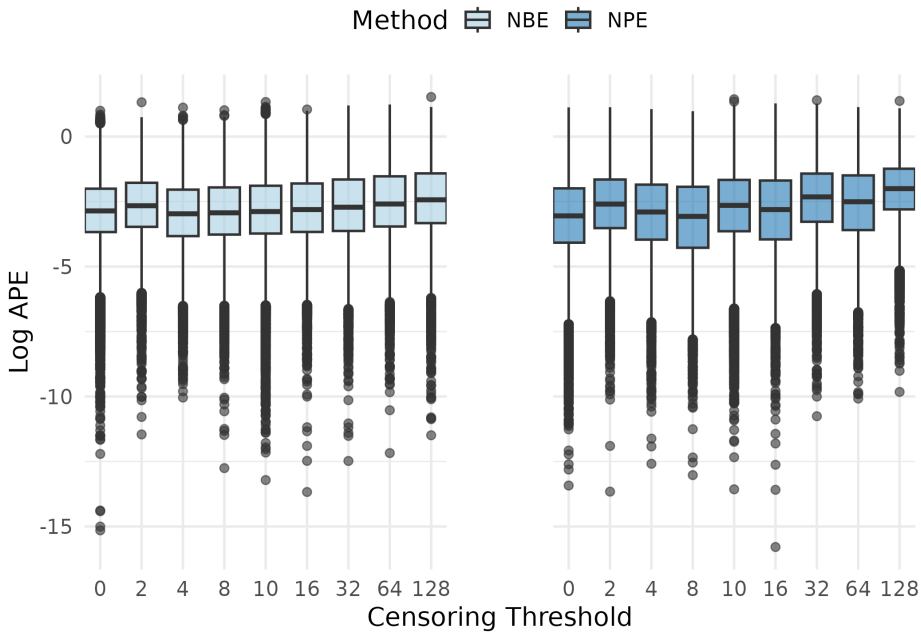

图5：NBE 和 NPE 在不同上限删失水平下的性能比较。各面板分别展示了基于测试数据集估计值（n=10,000）计算的神经贝叶斯估计器（左）和神经后验估计（右）的对数绝对百分比误差。

4.5 对删失水平的敏感性

现在，我们将评估我们的方法在不同删失水平下恢复模型参数的能力。在此分析中，我们使用了前文所述的同一测试数据，但在 $t = \{0, 2, 4, 8, 16, 32, 64, 128\}$ 处应用了不同的删失阈值。针对每个不同的删失水平，分别训练了一个估计器。 $t = 0$ 的情况对应于完全观测数据。在实际场景中，删失阈值可设为最高 20，但我们还纳入了更高的阈值，以确定该方法在何种情况下不再适用。

图5展示了估计器在不同删失阈值下的性能。我们发现，结果没有明显趋势；无论对数据采用何种删失水平，绝对百分比误差都保持稳定。尽管在较高阈值下绝对百分比误差略有上升，但这是预期结果，因为此时数据所含信息较少。关键的是，这种影响可以忽略不计，表明该方法在数据可观测性范围广泛的各种情况下都具有稳健性。

4.6 将估计器与马尔可夫链蒙特卡罗进行比较

为全面比较通过 MCMC采样器获得的 NBE、神经后验估计和后验分布，我们评估了各种特征。请回顾一下，这些方法生成的输出具有不同的粒度：

1. NBE：生成预先定义的一组后验分位数（例如第 2.5、50 和 97.5 百分位数），直接估计中位数和可信区间。2. NPE：从近似后验分布 $q(\boldsymbol{\theta} | \mathbf{n})$ 中生成样本，据此可计算任意统计量（均值、中位数、分位数）。3. 马尔可夫链蒙特卡罗：从金标准后验分布 $\pi(\boldsymbol{\theta} | \mathbf{n})$ 中生成样本，同样可以据此计算任意统计量。

因此，我们的评估框架旨在评估三个方面的表现：（1）点估计准确性，（2）不确定性量化的质量，以及（3）计算成本。在这项模拟研究中，我们有两个“真实值”：模拟的真实隐藏总体（$N_0^{\text{true}} = e^{\alpha}$），它作为准确性的基准；以及 MCMC 后验分布 $\pi(\boldsymbol{\theta} \mid \mathbf{n})$，在假设各链均已收敛的前提下，它作为不确定性的金标准基准。

我们考虑此前描述的、来自 $K = 5$ 列表且删失区间为 $[0, 10]$ 的相同测试数据，并使用上述标准比较各方法。与 King 等（2021）类似，令 $\mathcal{Y}$ 表示包含删失观测的捕获模式集合，即真实观测计数位于区间 $[0, 10]$内。在这种情况下，数据的概率质量函数为

$$f(\boldsymbol{n} \mid \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{\boldsymbol{x} \in \mathcal{X}} f_{\boldsymbol{x}}(N_{\boldsymbol{x}} = n_{\boldsymbol{x}} \mid \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}) \prod_{\boldsymbol{x} \in \mathcal{Y}} F_{\boldsymbol{x}}(N_{\boldsymbol{x}} = n_{\boldsymbol{x}} \mid \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}),$$

其中 $F_{\boldsymbol{x}}$ 表示累积分布函数。请注意，一般而言，对于任意满足 $0 < a < b$的删失区间 $[a, b]$，删失捕获模式的贡献对于所有 $\boldsymbol{x} \in \mathcal{Y}$均变为$\mathbb{P}(a \leq N_{\boldsymbol{x}} \leq b)$。

MCMC 分析使用 Julia 中灵活且功能强大的概率编程软件包 Turing.jl 实现（Fjelde 等人，2025）。我们采用无转向采样器（NUTS）（Homan 和 Gelman，2014）这一高效的哈密顿蒙特卡洛算法进行采样。

对于 10,000 个模拟数据集中的每一个，我们都运行了 4 条相互独立的 MCMC 链，每条链迭代 5,000 次，并丢弃前 1,000 次迭代作为预热期。我们使用 Gelman-Rubin 统计量（R）（Gelman 和 Rubin，1992）评估收敛情况。只有当所有模型参数均满足 $R \leq 1.01$ 时，我们才认为链已成功收敛。

4.7 点估计准确性

为公平比较这三种方法，我们首先为每种方法选取一个点估计。我们使用后验中位数，因为它能稳健地估计可能偏斜的后验分布，而且 NBE 会直接输出该值。

我们的主要关注点是模型恢复真实隐藏总体规模的能力，因此，对于每个测试数据集，我们将估计值 $\hat{N}^{\text{NBE0}}$、$\hat{N}_0^{\text{NPE}}$ 和 $\hat{N}_0^{\text{MCMC}}$ 与真实值 N_{true0} 进行比较。对于每个估计值 $\hat{N}_0$，我们使用原始误差（定义为 $\hat{N} - N^{\text{true0}}$）、绝对误差（AE），$|\hat{N} - N^{\text{true}_0}|$，以及 APE 来量化其表现。

原始误差分布用于评估估计量偏差。绝对误差用于量化估计误差的大小。APE 可以说是本研究中最关键的指标，因为它以真实总体规模为基准对误差进行标准化，从而能够在不同规模的模拟之间有意义地比较相对准确性。

对原始误差的分析（左图）证实，三种估计器的中位数均无偏，其分布都紧密集中在零附近。所有方法都存在极端离群值，不过 NBE 似乎最稳健，尾部较短。相比之下，NPE 呈现出轻微的正向中位数偏差（完整范围图见补充材料中的图1）。

三种方法的绝对误差（中间图）分布看起来相似，大多数误差都集中在零附近。NBE 略占优势，其分布在较低误差值处稍微更为集中。

绝对百分比误差（右图）揭示了最显著的性能差异。马尔可夫链蒙特卡罗分布明显更宽，尾部也长得多。相比之下，两种神经方法表现都更优，其绝对百分比误差分布集中在零附近，反映出其中位数绝对百分比误差更低，方差也显著更小。这表明

神经估计器具有更强的稳健性，并且在测试的总体规模范围内能够提供更准确的相对估计。最后，表1通过偏差、平均绝对百分比误差（MAPE）和均方根误差（RMSE）汇总了这些指标。

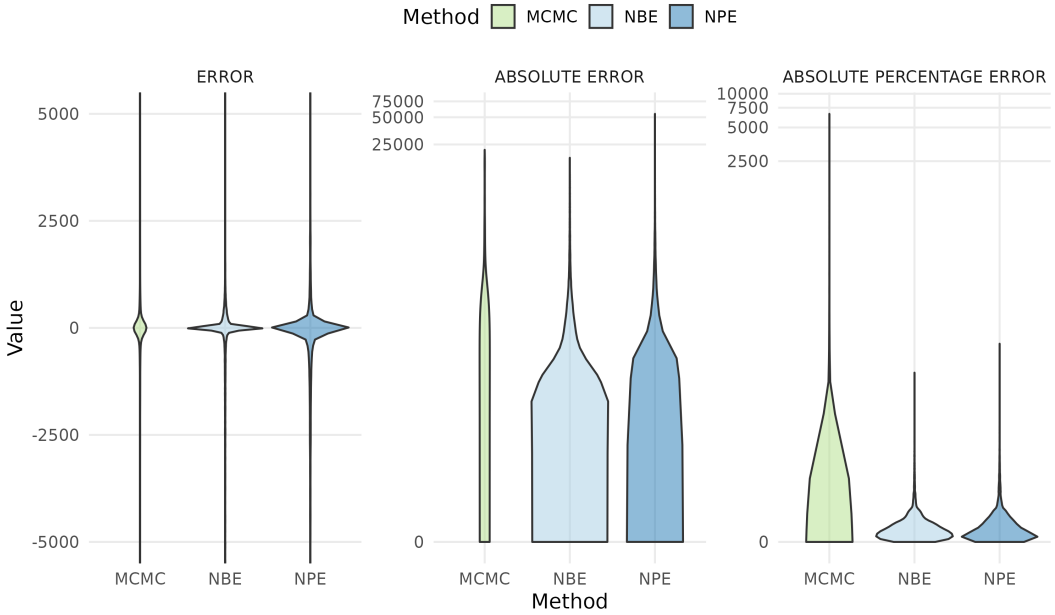

图6：展示每次模拟中估计的隐匿人口（N_0）误差指标分布的小提琴图。每幅图比较 MCMC（绿色）、NBE（浅蓝色）和 NPE（深蓝色）的参数估计值。指标分别为：（左）原始误差、（中）绝对误差和（右）绝对百分比误差。请注意，为便于查看，误差图的 y 轴范围经过限制；绝对误差和绝对百分比误差则以对数尺度显示。

4.8 不确定性量化

接下来，我们评估各方法的不确定性量化质量。评估分为两部分：一是直接比较 95% 可信区间（CI）的覆盖率，三种方法均可计算；二是更详细地分析完整的后验校准，但这仅适用于 NPE 和 MCMC 方法。

ˆ 不确定性量化准确性的主要指标是校准度。如果估计器 θ 的 95% 可信区间在 95% 的重复模拟中包含真实参数 θ，则该估计器校准良好（Little，2011）。我们首先计算三种方法的 95% 可信区间的经验覆盖概率（ECP）（表1）。对于 NBE，95% 的结果直接取自网络输出。对于 MCMC 和 NPE，我们从后验分布中生成样本，然后计算分位数。

此外，由于 NPE 和马尔可夫链蒙特卡罗都能生成完整的后验样本，我们可以评估它们在整个可信区间范围内的校准情况，而不仅仅是 95% 的可信区间。我们通过计算常规区间的经验覆盖概率（图7），生成完整的校准图。该图表明，NPE 具有良好的校准效果，表现为其紧密跟随 $y = x$，这说明后验分布在不确定性方面是诚实的。在 95% 水平下，NPE 的经验覆盖率为 0.961，接近理想值，并略微偏于保守。

另一方面，马尔可夫链蒙特卡罗表现出明显的校准失准。经验覆盖率曲线明显低于 $y = x$ 线，这表明后验分布总体上

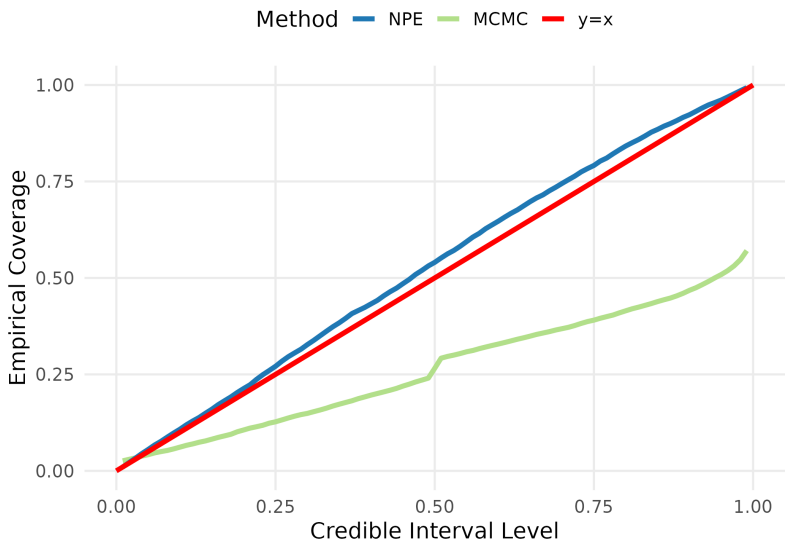

图7：校准图比较了 NPE（蓝色）和 MCMC（绿色）的经验覆盖率（纵轴）与名义可信区间水平（横轴）。红色的 y=x 线表示完美校准。

过窄且过于自信。此外，在 95% 水平下，马尔可夫链蒙特卡罗的可信区间仅有 50.9% 的时间包含真实参数，这表明采样器未能充分探索整个后验分布。

最后，NBE 的 95% 可信区间略大于 NPE 的经验覆盖概率，这表明二者都能提供较为准确且可靠校准的不确定性。

4.9 计算速度

神经方法的一项关键实际优势在于其摊销计算成本。这一成本分为两个阶段：一次性的“训练”成本和“每次推断”成本。

MCMC 采样器没有训练成本，但其推断成本很高，每个新数据集都需要数分钟到数小时才能收敛。如图8所示，MCMC 的运行时间（绿色）与迭代次数直接成正比，因此计算成本高昂。

相比之下，NBE 和 NPE 需要投入大量的一次性训练成本。训练完成后，对新数据集进行推断的成本很低，只需对神经网络进行前向传播即可。

NBE（浅蓝色）和 NPE（深蓝色）的推断时间比 MCMC 快几个数量级。这种摊销效率使它们在大规模模拟研究或实时应用中具有显著优势。

通过对数据拟合线性模型，我们可以观察其缩放特性。MCMC 和 NPE 的成本都随迭代次数 N 线性增长，即 $\mathcal{O}(N)$。相比之下，NBE 是 $\mathcal{O}(1)$ 操作，其运行时间与 N 无关。

MCMC 采样器（绿色）的成本最高，每次迭代约为 46.8 毫秒。NPE（深蓝色）的效率显著更高，每次迭代仅需 0.032 毫秒。NBE（浅蓝色）通常最快，单次前向传播耗时恒定，为 0.118 毫秒。这意味着在推断时，NPE 的速度平均约为 MCMC 采样器的 1,450 倍，凸显了摊销神经估计器在实际应用中的优势。

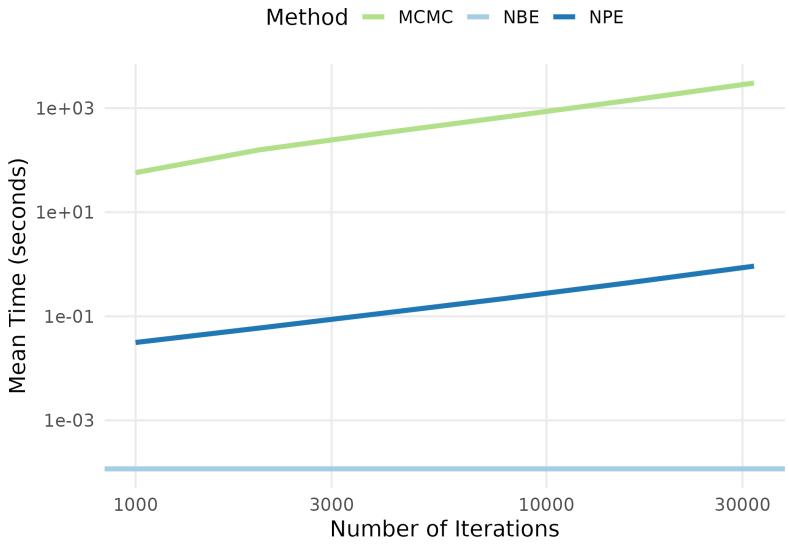

图8：MCMC（绿色）、NBE（浅蓝色）和 NPE（深蓝色）的平均推断时间（秒，对数尺度）随迭代次数的变化。每个点表示在 100 个测试数据集上的平均推断时间，每次运行重复 10 次。该图凸显了摊销推断在效率方面的显著提升：NBE 的耗时几乎恒定且极短，不受迭代次数影响；NPE 的推断速度则明显快于 MCMC。

	NBE	NPE	MCMC
Bias	-636	-194	2061
MAPE	0.53	0.669	52.1
RMSE	2190	2032	6593
95%经验覆盖概率	0.966	0.960	0.493
训练时间（秒）	1051	1577	0
推断时间（N，毫秒）	0.117	$0.0285 \times N$	$93 \times N$

表1：各方法的平均点估计误差、经验覆盖概率及计算成本（毫秒）汇总。

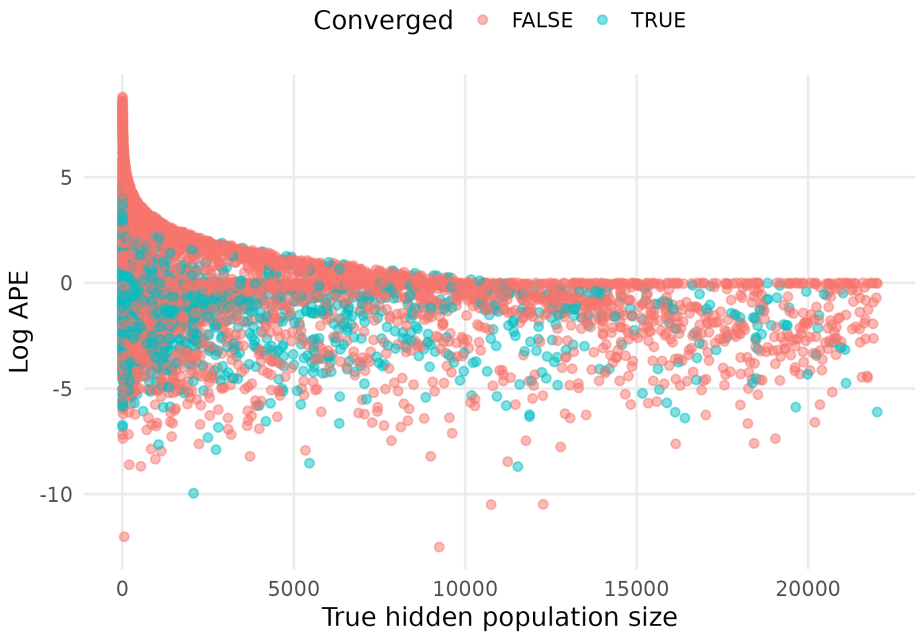

图 9：散点图展示了真实隐藏总体规模与基于马尔可夫链蒙特卡罗获得的后验样本计算出的对数绝对百分比误差之间的关系，其中样本（蓝色）已收敛，（红色）未收敛；收敛的定义为 Gelman-Rubin（GR）统计量 ≤ 1.01。

4.10 MCMC 收敛

模拟研究中的一个重要发现是，马尔可夫链蒙特卡罗的失败率很高。10,000 个测试数据集中，只有 7,141 个（71.41%）实现了收敛；可以看出（图 9），未能收敛至平稳分布的后验分布，其绝对百分比误差显著更大。

检查后发现，马尔可夫链蒙特卡罗的失败集中在非典型数据集中，尤其是那些生成了异常高名单计数的模拟（图10）。

这些数据可能会导致复杂的后验几何结构，使 NUTS 采样器难以处理；这通常表现为采样器无法在预热阶段找到正确的调优参数。虽然通过专家实施和逐个案例仔细调优或许可以弥补这些失败，但这凸显出马尔可夫链蒙特卡罗解决方案并非对所有数据都能“开箱即用”。

对于 7,141 个已收敛的数据集，我们将 MCMC 得到的截距参数后验中位数与神经方法的估计值进行了比较（图11）。这一比较验证了 NBE 在复现计算量更大的 MCMC 结果方面的准确性。

我们使用绝对百分比误差评估性能，发现对于较小的人口规模，NBE 的平均绝对百分比误差低于 MCMC。然而，对于非常大的人口规模，MCMC 的表现更好，这可能得益于信息量更大的数据。还可以看出，NBE 的绝对百分比误差在各种人口规模下都相当稳定，表明无论真实隐匿人口规模如何，其性能都稳健且稳定。

MCMC 出现了显著的收敛失败，即使是在实践中可能较为罕见的模拟情形下也是如此，这凸显了该估计器的稳健性。摊销估计器能够在广泛的数据生成参数范围内提供可靠的估计，包括传统采样器无法应对的复杂情形。

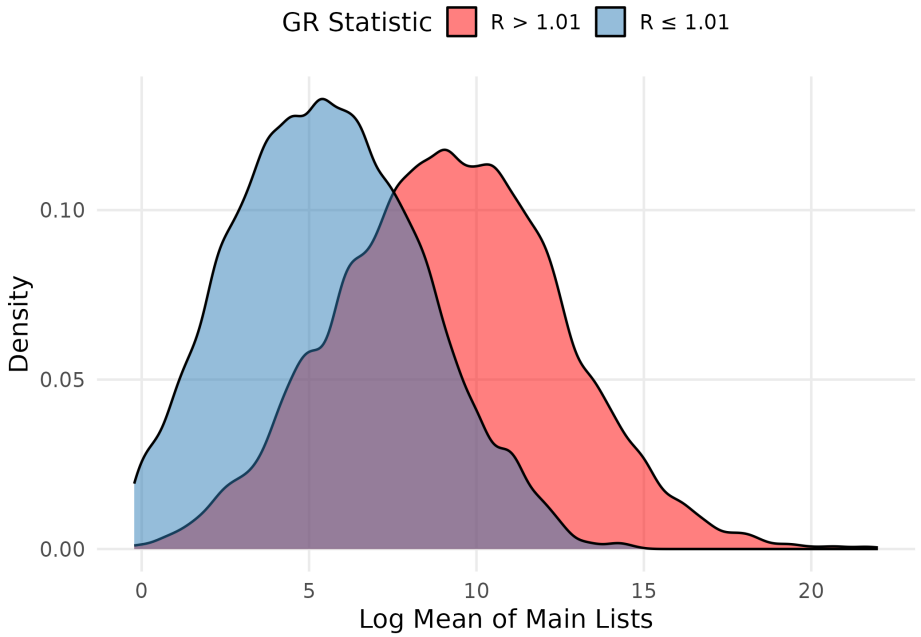

图10：五个主要列表中每个列表的对数平均计数密度，已收敛的链（蓝色）与未收敛的链（红色）叠加显示，其中收敛定义为 Gelman-Rubin（GR）统计量 ≤ 1.01。

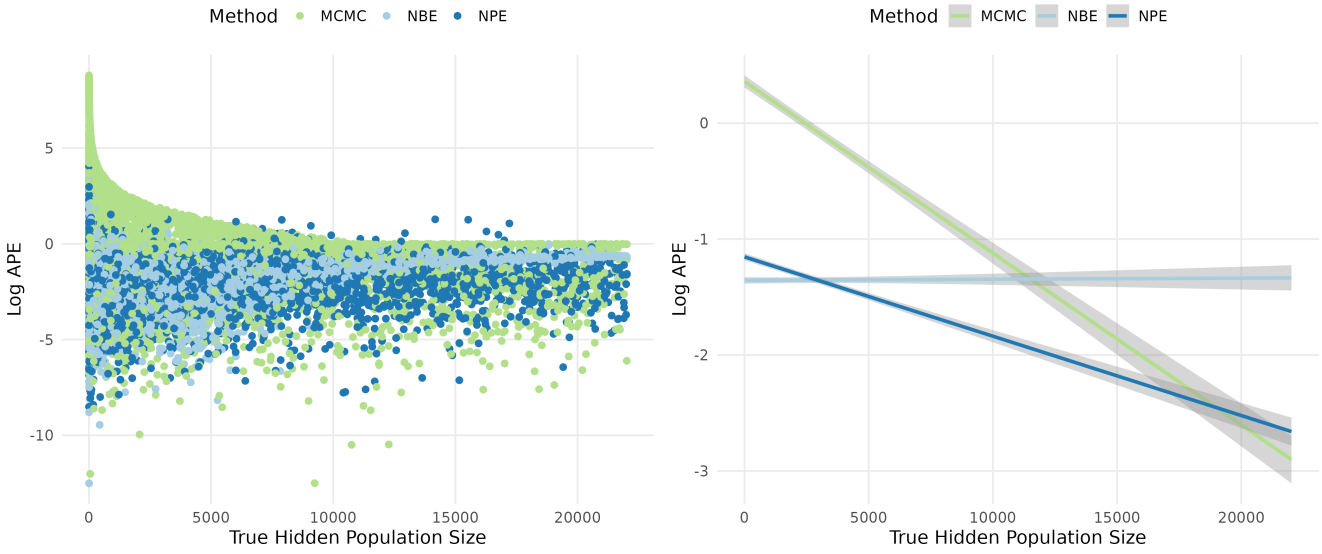

图11：左：NBE（浅蓝色）、神经后验估计（深蓝色）和马尔可夫链蒙特卡罗（绿色）推断的隐藏总体规模相对于真实总体规模的对数绝对百分比误差。右：各方法的线性拟合及其95%置信区间。

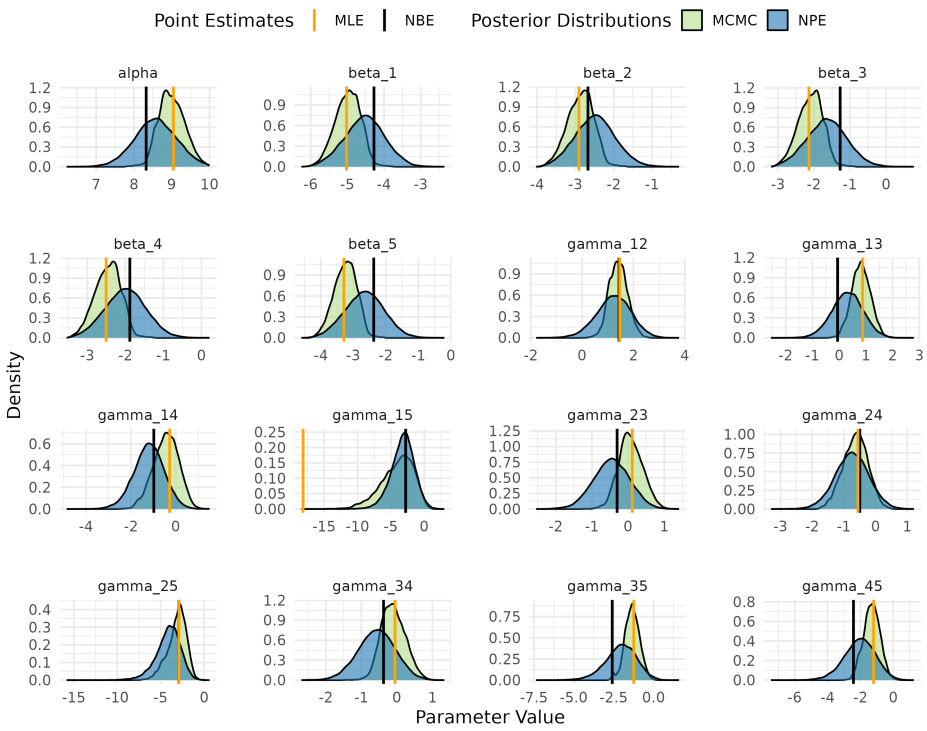

图12：现代奴隶制数据集的参数估计值和后验分布对比。竖线表示点估计（最大似然估计和NBE），填充曲线表示后验密度（马尔可夫链蒙特卡罗和神经后验估计）。

5 真实数据分析

现在我们使用 我们的方法分析两个真实数据集，并估计隐匿人口的规模 .

5.1 英国现代奴隶制

我们将方法应用于英国现代奴隶制受害者人数的数据，这些数据由 Silverman（2020）分析。该数据集包含不同报告机制记录的受害者计数，包括地方政府、非政府组织和警察部队。这是一个值得关注的多系统估计数据集示例，因为原始研究的发现促成了英国新立法的起草。

我们针对一个包含 $K = 5$ 个列表且没有删失的系统训练了多种架构。通过最小化未见测试数据上 α 的平均绝对百分比误差（MAPE），选出了最终模型。选定的 NBE 模型采用四个隐藏层，每层包含 256 个神经元；而最优 NPE 模型采用一个宽度为 128 的隐藏层作为编码器网络，摘要网络的输出维度也为 128。

推断的后验分布和点估计的比较见图12。为求完整，我们还将通过神经方法推断出的估计值与马尔可夫链蒙特卡罗和最大似然估计进行了比较。我们的模型估计总体上与 Silverman（2020）报告的结果一致；不过，我们使用了所有两两交互项，未进行正式的模型选择。值得注意的是，交互项 γ_{15} 的最大似然估计显著比贝叶斯估计更负。这是因为列表1和列表5之间没有任何重叠，导致该参数在似然框架中无法识别。由于先验带来的正则化，贝叶斯方法和神经方法能够更稳健地处理这一情况。

方法	隐匿人口	95%置信区间
MCMC	7987	(4462, 16154)
NPE	5291	(1784, 15294)
NBE	4123	(889, 18746)
MLE	8489	(3761, 19159)

表2：现代奴隶制数据集中各方法推断的隐匿人口及其不确定性。

ˆ 表2比较了通过似然函数（最大似然估计和后验均值）以及我们的两种神经估计器，针对两种估计所得的隐匿人口估计值 $N_0 = \exp(\hat{\alpha})$。对于贝叶斯方法，我们报告后验中位数和 95% 可信区间；对于频率学派方法，我们报告最大似然估计和 95% 置信区间。

我们发现，所有估计值总体上都相近，但各方法对截距 α 的估计存在显著差异。如图12所示， α 的 MCMC 后验分布中心约为 9.0，与最大似然估计非常接近；而 NPE 和 NBE 的分布则向左偏移，中心更接近 8.5。由于指数链接函数 $\hat{N}_0 = \exp(\hat{\alpha})$，这种在对数空间中的小幅偏移会导致总体规模上的巨大差异（约为 8,000 对 5,000）。

最后，贝叶斯范式的一项关键优势在于能够通过后验预测分布（PPD）评估模型拟合情况， $\pi(\mathbf{n}_{\text{new}} | \mathbf{n})$。该分布表示在给定观测数据的条件下，模型所隐含的新数据分布。我们利用从神经后验估计（NPE）获得的后验样本模拟生成新的数据集，从而得到后验预测分布。如图13所示，对于几乎所有交叉分类概况，观测数据计数（以红线表示）均落在边际后验预测分布的95% 可信区间内。值得注意的例外是四重重叠的计数 n_{1234}。除这一离群值外，模型对观测数据的拟合尚属合理。

5.2 英格兰东北部的女性吸毒者

第二个应用涉及英格兰东北部女性吸毒者的数据，该数据最初由 King 等（2021）发表。数据集记录了通过四种不同报告机制识别出的15至34岁女性吸毒者人数：缓刑监管、戒毒干预计划（DIP）监狱评估、戒毒治疗以及DIP 社区评估。

虽然可以获得每个列表及各种列表组合中出现的女性人数，但为防止身份可能被披露，计数在1到4之间的数据已被删失。我们针对这一由 $K = 4$ 个列表组成的系统训练了多种架构，并将删失区间 $[1, 4]$ 纳入网络输入。与上一节一样，我们通过最小化未见测试数据上的平均绝对百分比误差（MAPE）来选择最终模型。最优的NBE模型使用了4个隐藏层，每层包含256个神经元；而选定的NPE模型采用一个包含单个隐藏层的编码器网络，隐藏层宽度为128，摘要网络的输出维度为128。

参数估计值和后验分布如图14所示。与现代奴隶制数据集不同，在该数据集中NBE给出的总体估计较为保守；在这里，NBE对截距 α 的点估计（黑色竖线）明显高于最大似然估计（橙色竖线）。这一变化表明，神经贝叶斯估计器推断出的隐匿人口规模更大。这与以下观察一致：NBE对主效应（例如 β_1， β_2等）的估计通常更为负向

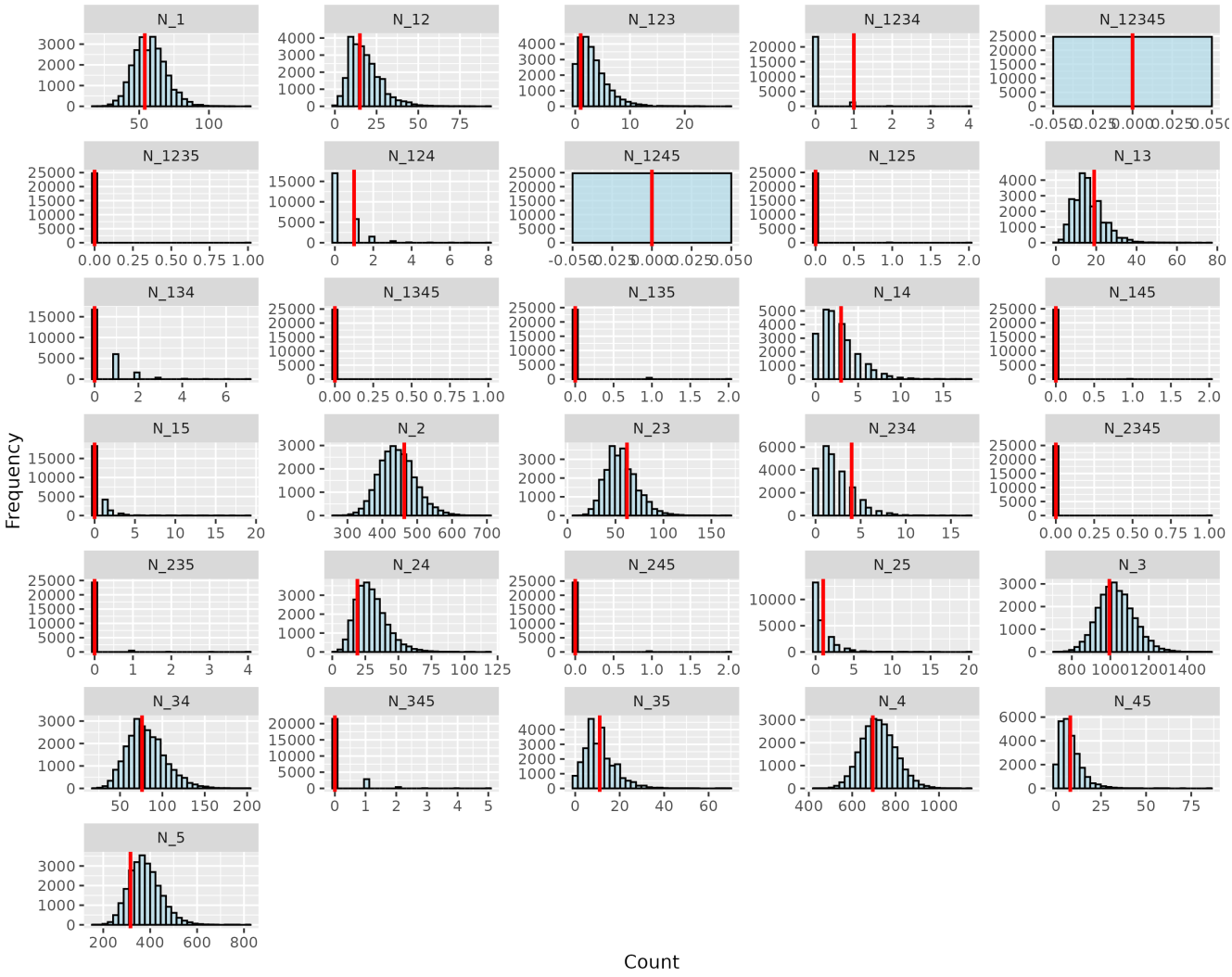

图13：使用神经后验估计（NPE）生成的后验预测分布（PPD）。直方图显示不同列表交集的模拟计数，红色垂直线表示数据集中的实际观测计数。

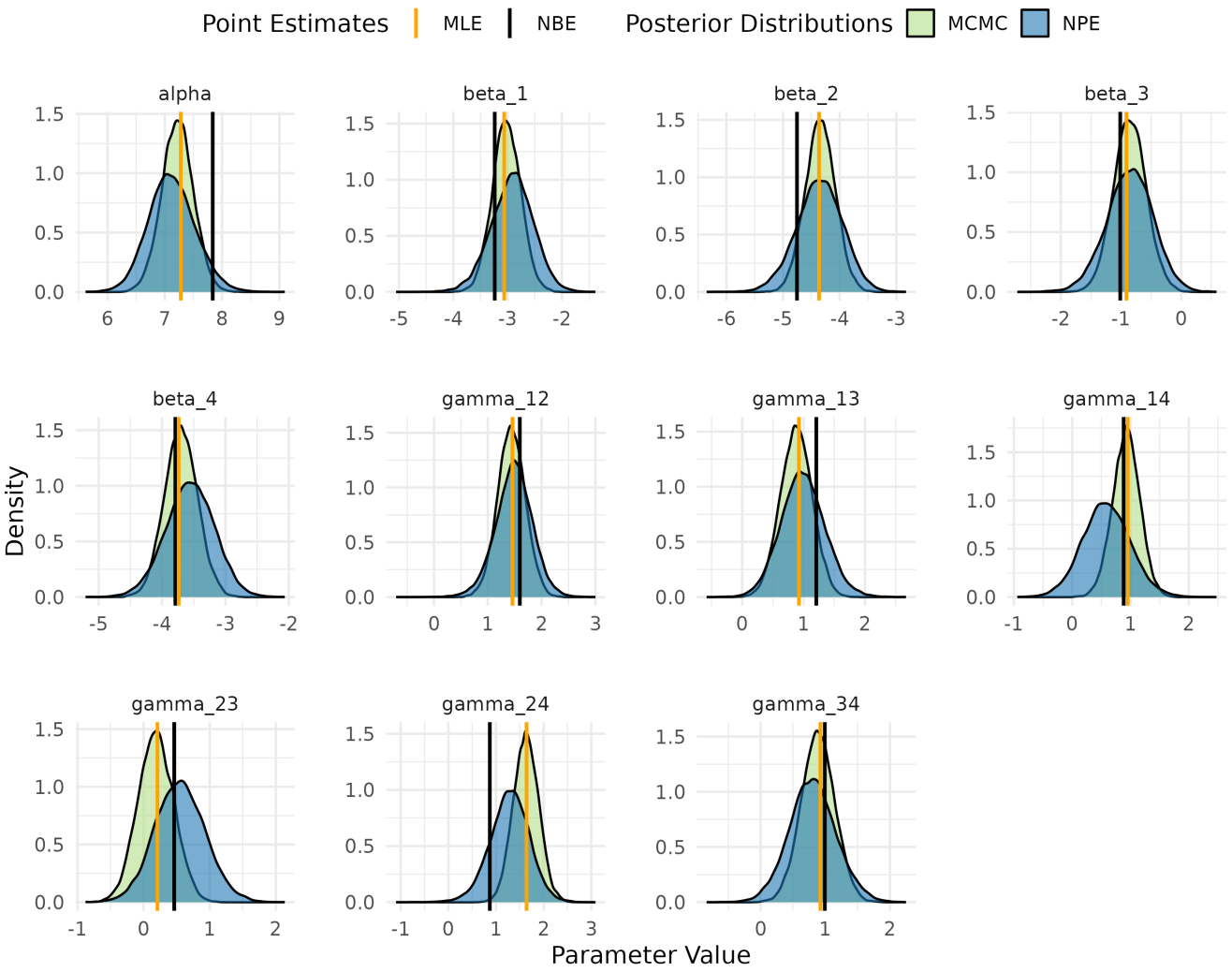

图14：女性吸毒者数据集（$K = 4$）的参数估计值比较。竖线表示点估计（MLE 和 NBE），填充曲线表示后验密度（马尔可夫链蒙特卡罗和 NPE）。

比其 MLE 对应值更低，这意味着捕获概率更低，进而意味着未观测总体更大。

此外，我们观察到，交互项 γ_{14}，γ_{23} 和 γ_{24} 的 NPE 后验分布（蓝色密度曲线）与 MCMC 后验分布（绿色密度曲线）存在差异。这表明，与标准 MCMC 方法相比，神经后验估计器能够检测出列表之间的差异依赖关系。

表3比较了推断的隐匿总体。与参数估计值一致，NBE 得出的估计值高于经典方法。请注意，对于删失数据，似然和后验评估会明确考虑分配给区间 $[1, 4]$ 的概率质量。

我们进一步利用图15所示的后验预测分布评估模型拟合情况。在此图中，红色实线表示精确观测计数，竖直阴影带表示删失区间 $[1, 4]$。如果预测计数的概率质量大部分落在该阴影区域内，则认为模型对删失数据的拟合良好。

方法	隐匿人口	95%置信区间
MCMC	1371	(795, 2346)
NPE	1225	(579, 2852)
NBE	2530	(80, 3942)
MLE	1450	(840, 2505)

ˆ 表3：King 等（2021）数据集的推断的隐藏总体（N_0）和不确定性区间。

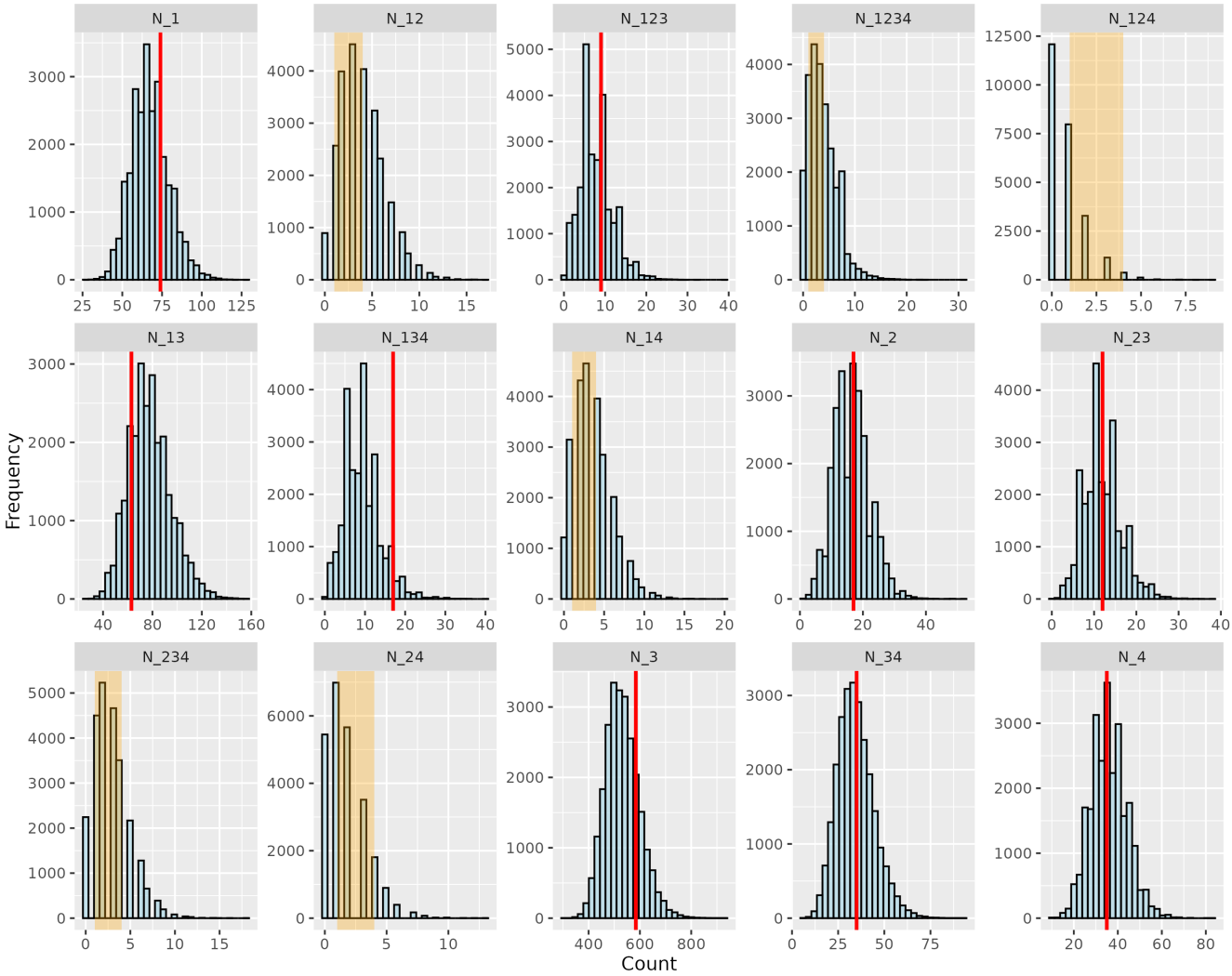

图15：女性吸毒者数据集的后验预测分布（PPD）。红色竖线表示精确观测计数。橙色阴影带表示数据被删失的单元格（计数为1-4）。直方图显示，神经模型能够正确预测删失区间内的计数。

目视检查证实，神经后验估计能够有效捕捉数据结构。对于主列表中的精确计数（例如，n_1、……、n_4），观测值都接近预测直方图的众数。尤为重要的是，对于删失交互项（例如，n_{12}、n_{124}等），后验预测分布在删失区域（橙色带）内赋予了显著的概率质量。这表明，尽管删失带来了歧义，神经方法仍成功学会了为这些低计数交互作用分配概率。

6 讨论

本文提出了一种适用于MSE模型的基于模拟的推断框架。即使计数数据被删失或缺失，这种方法也能实现稳健推断；而在这类情形下，传统的基于似然的方法通常会变得难以处理，或在计算上成本过高。通过大量模拟研究，我们证明了神经方法在完全观测和删失两种情形下，都能准确恢复模型参数和总体规模。此外，我们还表明，神经后验估计提供的不确定性量化具有良好的校准度；在马尔可夫链蒙特卡罗无法收敛的情况下，其表现往往更优。

将我们的方法应用于真实世界的数据集，揭示了两种神经网络方法之间一个重要的区别。对于现代奴隶制数据集，NBE、NPE 和马尔可夫链蒙特卡罗的估计结果总体一致，但由于先验对不可识别交互项的正则化作用，神经方法倾向于给出略为保守的隐藏总体规模估计。

然而，对女性吸毒者数据集的分析显示出显著差异：NBE 估计的隐匿人口规模为 2,530，几乎是 NPE（1,225）和马尔可夫链蒙特卡罗（1,371）估计值的两倍。

NBE 的任务是学习从样本空间到参数空间的直接映射。此处观察到的差异表明，对于这种特定的稀疏数据配置，网络难以学习准确的映射。在极度稀疏、审查程度较高或数据集较为特殊的情况下，这种逆映射会变得高度非线性且复杂。当前的神经贝叶斯估计架构可能缺乏捕捉这种复杂性所需的灵活性或容量。未来的研究可以通过探索更深、更具表达能力的架构（例如 Transformer 或残差网络），或改进先验设定，使网络训练聚焦于参数空间中最相关的区域，从而解决这一问题。因此，我们建议研究人员将 NBE 视为一种快速的初步筛查工具。当 NBE 与 NPE 的估计结果存在显著差异时，应将 NPE 视为更可靠的标准，因为它能够进行后验预测检验（图15），以验证学习到的分布是否能够合理解释观测数据。

这种差异似乎是由截距与主列表效应之间的平衡所驱动的。与马尔可夫链蒙特卡罗相比，神经估计器推断出的 βs 值更大，这意味着个体出现在每个列表中的概率高于马尔可夫链蒙特卡罗的估计值。在多系统估计中，更高的捕获概率通常会导致更低的隐匿人口估计值，因为只需较少的未观测个体即可解释观测计数。因此，尽管所有贝叶斯方法都能有效处理交互项的稀疏性，神经网络最终选择了一个可捕获性更高、从而对隐匿人口作出更保守估计的解。

对比 NBE 在两个数据集中的表现颇有启发。在现代奴隶制案例中，NBE 起到了正则化器的作用，相较于最大似然估计，它缩小了总体估计。然而，在女性吸毒者案例中，NBE 得出的总体估计更大（这是由更大的截距 α所致）。这表明神经网络

估计器能够根据数据进行调整；它们并不会总是朝着同一个方向使结果产生偏差（例如，总是趋于保守）。相反，它们会根据观测列表中稀疏性和相关性的具体结构作出响应。在本例中，模型推断可捕获性较低（β 值更负），因此需要更大的隐匿人口来解释观测计数。

将神经贝叶斯估计器用于多系统估计，是一项重要的实践进展，主要解决了计算和可及性方面的瓶颈，而这些瓶颈往往会妨碍传统推断方法的应用。我们的方法最主要的优势在于其摊销特性。初始的高强度计算训练阶段完成后，便可近乎即时地为新的数据集生成后验估计。这对于以迭代方式收集数据的多系统估计研究尤其有益，使研究人员能够迅速更新研究结果，而无须投入大量时间重新运行完整的马尔可夫链蒙特卡罗分析。

此外，基于模拟的框架天然能够规避病态的估计情形。例如，当列表之间存在空交集，导致有限的最大似然估计不存在时，我们的方法仍能确保始终得到稳定且有限的估计值。

然而，本文所述方法也存在一些局限。估计器使用特定的固定先验分布进行训练。因此，如果研究人员的先验信念与训练时使用的先验分布不同，就无法将这些信念纳入分析，除非从头开始重新训练一个全新的模型。此外，神经贝叶斯估计器仅返回后验中位数和 95% 可信区间；若无法从后验预测分布中生成样本，就无法进行模型评估。

此外，我们当前的框架侧重于对单个预先定义的模型进行参数估计，并未涵盖模型选择这一关键步骤。对列表依赖关系作出不同假设，可能会导致隐匿总体规模估计出现显著差异（Silverman，2020；King 等人，2021）。

最后，整个框架的准确性取决于用于训练的模拟数据质量。如果训练数据未能涵盖现实场景的全部范围，那么估计器在未见数据上的表现必然不佳。最后，网络本身如同一个“黑箱”。虽然我们已证明这些估计器能够高精度地逼近贝叶斯后验量，但其内部决策过程并不透明，这与 MCMC 采样器可解释、循序渐进的逻辑形成鲜明对比。

未来研究的一项重要改进方向，是克服模型当前的僵化性。我们建议开发一种条件神经贝叶斯估计器，使其能够在数据之外接收额外输入。这些输入可以指定先验超参数，甚至定义模型结构，例如应纳入哪些交互项。例如，在 Richards 等人（2024 年）的研究中，作者明确将删失值作为输入，并证明其表现具有可比性。这样便能构建一个灵活性很高的单一估计器，使其在推断时适应研究者的具体信念和假设，从而无需针对每项独特分析都重新训练模型。

代码、材料和数据可用性

本研究中使用的所有代码和数据均可在 https://github.com/HiddenHarmsHub/npe-mse-censored-missing 仓库中找到。

致谢

本研究得到英国研究与创新署未来领袖奖学金 [MR/X034992/1] 的资助。本文所述的计算使用伯明翰大学的 BlueBEAR 高性能计算服务完成，该服务为该校研究群体提供高性能计算服务。详情请见 http://www.birmingham.ac.uk/bear。

参考文献

Bird, S. M. 和 King, R.（2018）。用于为公共政策提供依据的多系统估计（或捕获-再捕获估计）。《统计学及其应用年度评论》，5(1):95–118。

Chan, L., Silverman, B. W., 和 Vincent, K.（2020）。稀疏捕获数据的多系统估计：存在不重叠列表时的推断挑战。《美国统计协会杂志》，116(535):1297–1306。Fjelde, T. E., Xu, K., Widmann, D., Tarek, M., Pfiffer, C., Trapp, M., Axen, S. D., Sun, X., Hauru, M., Yong, P., Tebbutt, W., Ghahramani, Z., 和 Ge, H.（2025 年）。Turing.jl：通用概率编程语言。《*ACM* 概率机器学习汇刊》Gelman, A. 和 Rubin, D. B.（1992）。使用多条序列进行迭代模拟推断。《统计科学》，7(4):457–472。Homan, M. D. 和 Gelman, A.（2014）。无转向采样器：在哈密顿蒙特卡洛中自适应设置路径长度。《机器学习研究杂志》，15(1):1593–1623。Kavianpour, S., Sutherland, J., Mansouri-Benssassi, E., Coull, N., 和 Jefferson, E.（2022）。可信研究环境的下一代能力：访谈研究。《医学互联网研究》，24(9):e33720。King, R., Bird, S. M., Hay, G., 和 Hutchinson, S. J.（2008）。通过贝叶斯捕获-再捕获方法，按性别、地区和年龄组估计苏格兰当前注射毒品者人数及其药物相关死亡率。《医学研究统计方法》，18(4):341–359。King, R., de Rivera Ortega, O. R., 和 McCrea, R.（2021）。存在删失单元格时的多系统估计，章节：短篇论文集，SIS 2021。培生。Little, R.（2011）。校准贝叶斯：一般统计学中的应用，尤其是缺失数据。《统计科学》，26(2)。Luchenski, S. A., B¨ohning, D., Aldridge, R., Stevenson, F., Tariq, S., 和 Hayward, A. C.（2025 年）。估计英格兰无家可归者住院入院规模：一项利用全国住院事件统计数据开展的基于人群的多系统估计研究。《英国医学杂志公共卫生》，3(2):e002978。Richards, J., Sainsbury-Dale, M., Zammit-Mangion, A., 和 Huser, R.（2024 年）。利用超阈值模型对删失数据进行推断的神经贝叶斯估计器。《机器学习研究杂志》，25(390):1–49。Sainsbury-Dale, M., Zammit-Mangion, A., 和 Huser, R.（2024 年）。使用神经贝叶斯估计器进行无似然参数估计。《美国统计学家》，78:1–14。

Sainsbury-Dale, M., Zammit-Mangion, A., Richards, J., 和 Huser, R.（2025 年）。利用图神经网络处理不规则空间数据的神经贝叶斯估计器。计算与图形统计学杂志，34(3):1153–1168。

Silverman, B. W. (2020). 用于量化现代奴役的多系统分析：经典方法与贝叶斯方法。英国皇家统计学会会刊 *A* 辑：社会统计，183(3):691–736。