arXiv:2304.07897v1 [stat.ME] 16 Apr 2023

信息性抽样下纵向调查数据的贝叶斯单位层级模型：利用家庭脉搏调查分析预期失业

Daniel Vedensky*、Paul A. Parker†和Scott H. Holan‡§,

摘要

美国人口普查局近期发布的家庭脉搏调查（HPS）收集有关冠状病毒社会影响和经济影响的最新信息。调查第一阶段在冠状病毒大流行开始一个月后迅速启动，持续了12周。为追踪大流行病的即时影响，该阶段对个体受访者进行了重新抽样，最长连续三周。受大流行期间预期失业问题的启发，本文利用公开使用微观数据，提出了单位层级的基于模型的估计量，将响应层级和领域层级的纵向依赖纳入考量。具体而言，我们采用伪似然，针对信息性抽样下的高斯响应数据和二元响应数据，研究一种贝叶斯分层单位层级、基于模型的方法。为便于构建这些基于模型的估计，我们开发了一种高效的吉布斯采样器。我们开展了一项实证模拟研究，将所提方法与未考虑单位层级纵向相关性的模型进行比较。最后，我们利用HPS公开使用微观数据，对“预期失业”进行分析，比较基于设计的估计量和基于模型的估计量，并证明所提基于模型的方法表现更优。

关键词： 贝叶斯层次模型；信息性样本；伪似然；小区域估计；时空

*（通信请联系）密苏里大学统计系，密苏里州哥伦比亚市 Middlebush Hall 146号，邮编 65211-6100，dvedensky@mail.missouri.edu

†加利福尼亚大学圣克鲁兹分校统计系，Santa Cruz, CA 95064，High St 1156号，paulparker@ucsc.edu

‡密苏里大学统计系，密苏里州哥伦比亚市 Middlebush Hall 146号，邮编 65211-6100，holans@missouri.edu§美国人口普查局，华盛顿特区 Silver Hill Road 4600号，邮编 20233-9100，scott.holan@census.gov

1 引言

2020年COVID-19疫情暴发后不久，美国人口普查局便推出了家庭脉搏调查（HPS），将其作为一项实验性数据产品。鉴于疫情暴发造成了前所未有的影响，且需要迅速做出决策，HPS旨在缩短从数据收集到数据发布之间的周转时间。为追踪这些影响，该调查涵盖了种类繁多的问题，内容从经济影响（例如申请失业保险以及受访者是否收到经济刺激支票）到健康影响（例如焦虑症状）不等。

本文重点研究的 HPS 第一阶段持续了12周，从2020年4月23日至2020年7月21日，并采用纵向设计，受访者可以在短时间内接受多次访谈。如果某个家庭在某一周被抽中并作出响应，该家庭便会保留在下一周的抽样框中。如果该家庭随后再次作出响应，则会继续保留在抽样框中，进入第三周也是最后一周，之后便会被移出。单位无应答的调整通过调查权重纳入。[1]

这种调查设计形成了一种结构：家庭最多可能连续三次出现在数据中。我们将由此产生的受访者内相关性称为“纵向相关”，以区别于调查每周开展所产生的“时间相关”。

HPS 已被证明有助于调查疫情暴发初期的影响。美国疾病控制与预防中心利用 HPS 衡量大流行病对抑郁症和焦虑等心理健康状况的影响 (Vahratian 等人，2021年)。美国劳工统计局

[1] https://www2.census.gov/programs-surveys/demo/technical-documentation/hhp/2020_HPS_背景.pdf

利用这项调查追踪政府刺激补助的领取和使用情况。[2] 与此同时，其他研究人员利用 HPS 分析了各种不同的结果，范围从粮食不安全 (Schanzenbach 和 Pitts，2020年) 到教育和经济不平等 (Francis 和 Weller，2021年)。然而，这些研究都将数据视为横截面数据，因此忽略了不同周之间的时间依赖性，以及纵向研究设计所提供的宝贵信息。

此外，HPS 收集的丰富人口统计数据自然带来了适合采用单位层级建模的小域问题。然而，关于纵向数据单位层级建模的现有文献往往仅处理高斯响应的情形，尽管调查数据中二元响应十分普遍（例如 Guadarrama 等人 （2020年））。当前允许使用二元响应的方法通常忽略复杂抽样设计和调查权重（例如 Hobza 和 Morales （2016年）；Hobza 等人 （2017）），不过Kunihama 等人 （2019）考虑了一种混合模型，该模型利用报告的调查权重调整混合成分概率。由于调查样本不太可能具有代表性，或可能存在信息性抽样，忽略调查设计和权重的模型容易产生偏差和不精确估计。尽管近年来，在信息性抽样条件下将单位层级模型拟合于复杂调查数据的相关文献取得了许多进展（例如，参见 Parker 等人 （2019）），但迄今为止其主要关注横截面数据。因此，现有的单位层级方法并不直接适用于纵向数据或时间序列数据，例如 HPS。

受 HPS 以及 COVID-19疫情初期预期失业影响的启发，我们提出一种单位层级建模方法，同时考虑纵向结构和复杂抽样设计。具体而言，我们提出了适用于

[2] https://www.bls.gov/opub/btn/volume-9/receipt-and-use-of-stimulus-payments-in-the-time-of-the-covid-19-pan htm

高斯数据和二元数据的模型，使相关领域的研究人员能够回答广泛的科学和政策相关问题。

本文结构如下。第 2节介绍本文的符号和方法。第3节基于 HPS 的现有数据开展实证模拟研究。第 4节利用完整的 HPS 第一阶段数据集进行启发性数据分析，以考察美国 COVID-19 初期的预期失业情况。最后，第 5 节给出讨论和一些总结性意见。

2 方法

2.1 符号与背景

沿用 Parker 等人（2022）的符号，我们用 $\mathcal{U} = \{1, \ldots, N\}$ 表示有限总体的一个枚举。令 $\mathcal{U}$为互不相交的子总体 $\mathcal{U}_j$、 $j = 1$、.....、 k的并集，每个子总体的基数均为 $|\mathcal{U}_j| = N_j$，因此 $\sum_{j=1}^{k} N_j = N$。在每个时间点 $t = 1$、.....、 T，抽取一个大小为 n_t 的样本S_t ，并以 $\pi_{it} = P(i \in \mathcal{S}_t)$表示各单位被抽中的给定（已知）概率。于是，时间点 t 区域 j 中被抽样单位的索引集为 $\mathcal{S}_{jt} = \mathcal{U}_j \cap \mathcal{S}_t$。我们用 $\mathcal{S} = \bigcup_{t=1}^{T} \mathcal{S}_t$ 表示整个样本，其中 $|\mathcal{S}| = \sum_{t=1}^{T} n_t = n$。

设 y_{ijt} 为时间 t 区域 j 中与单位 i 相关的特征。通常，我们希望估计时间 t 区域 j 中该特征的总体均值（若 y_{ijt} 为二元变量，则估计其比例）

$$\overline{y}_{\cdot jt} = \frac{1}{N_j} \sum_{i \in \mathcal{U}_j} y_{ijt}.$$

通常会采用基于抽样设计的估计量，例如 Horvitz-Thompson（HT）估计量（Horvitz 和Thompson，1952），

$$\widehat{\overline{y}}_{\cdot jt} = \frac{1}{N_j} \sum_{i \in \mathcal{S}_{jt}} w_{it} y_{ijt},$$

其中 $w_{it} = \frac{1}{\pi_{it}}$. 所谓直接估计量（如 HT 估计量）仅使用观测到的调查数据，并确保具有设计无偏性和设计一致性等理想性质。然而，当总体被细分为更精细的分区时，问题便会出现，因为某些子总体的样本量必然很小（甚至为零）。这类子总体往往会产生不可靠的直接估计，且标准误较大；通常称之为“小区域”，即使这一称谓并非严格指地理区域。此外，如果某个小区域的样本量为零，就无法计算直接估计。在这些情况下，基于模型的方法更为合适，因为借助协变量和随机效应，小区域可以从其他区域“借用”信息（Rao 和Molina，2015年）。

小区域模型可以在区域层级或单位层级上实施。前者将各区域的直接估计量作为模型中的响应值，并通过区域层级协变量将这些响应值关联起来。由于依赖直接估计量，小样本量对于区域层级模型而言仍可能造成问题。因此，如果能够获得各个单位的响应数据，采用无需直接估计的单位层级模型可能更为合适。

此外，研究表明，与区域层级模型相比，单位层级模型能产生更低的标准误和更好的区间覆盖率（Hidiroglou和You，2016年）。如果数据按一种分辨率制表，而所需估计结果采用另一种分辨率，单位层级模型还可免去临时基准校准的需要。在这种情况下，区域层级模型可能违反内部一致性，而单位层级模型则允许对单位进行任意聚合，从而生成任意分辨率下的估计结果。

基本单位级模型最初由Battese 等人(1988年) 提出，其形式为一个频率学派的混合效应模型

$$y_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + v_j + e_{ij},$$

其中未知参数为 $\boldsymbol{\beta}$，区域层级随机效应为 $v_j \overset{iid}{\sim} N(0,\ \sigma_v^2)$，且与单位层级随机误差 $e_{ij} \overset{iid}{\sim} N(0,\ \sigma_e^2)$相互独立。与区域层级模型不同，有限总体参数不会直接在单位层级模型中进行参数化。相反，在拟合单位层级模型后，可以对总体中的每个单位进行预测。这样便可得到一个合成总体，并可将其汇总至任何所需估计的分辨率。在贝叶斯框架下，可以对每次马尔可夫链蒙特卡罗迭代执行此操作，从而得到估计量的后验预测分布（例如，详情参见Parker 等人(2022年）。

然而，基本单位级模型做出了一个关键假设：抽中单位的分布与总体分布相匹配。这一假设在复杂抽样设计中很少成立。个体的响应值通常与其被纳入样本的概率相关，这种现象称为信息性抽样（IS）。Parker 等人(2019描述了忽略 IS 所产生的问题，并概述了单位层级模型中最常见的校正方法。

一种调整单位层级模型以适应信息性样本的直接方法是采用伪似然（PL）（Skinner，1989年；Binder，1983年）。在伪似然中，每个抽中单位对似然的贡献会按其调查权重进行指数加权，

$$\mathrm{PL}(\boldsymbol{y}|\theta) = \prod_{i \in \mathcal{S}} f(y_i|\theta)^{w_i}.$$

Savitsky 和 Toth (2016年) 表明，此类似然可以纳入贝叶斯模型；与先验密度 $\pi(\boldsymbol{\theta})$ 相乘后，会得到有效的伪后验分布，

$$\widehat{\pi}(\boldsymbol{\theta}|\boldsymbol{y}, \widetilde{\boldsymbol{w}}) \propto \left\{ \prod_{i \in \mathcal{S}} f(y_i|\theta)^{w_i} \right\} \pi(\boldsymbol{\theta}),$$

前提是权重经过缩放，使其在每个时间点的总和等于样本量，n_t，

$$\widetilde{w}_{it} = n_t \frac{w_{it}}{\sum_{\ell \in S_t} w_{\ell t}}.$$

在横截面情境下，Zhang 等人(2014) 拟合了此类伪似然模型，并用其生成总体估计，尽管采用的是频率学派框架。

这种单位层级建模方法已被用于横截面数据：由Parker等人 (2022年)用于分类数据，以及由Parker等人(2023)用于高斯数据；他们还提供了高效的抽样算法。Sun等人(2022年)使用这一框架分析HPS，但未考虑HPS数据的纵向方面。

2.2 高斯响应模型

对于高斯情形，我们采用AR(1)结构对受访者内纵向相关性进行建模。由此，首次作答和后续作答的伪似然贡献具有不同的函数形式，因为后续作答必须以先前的响应为条件。为区分样本中的这两组受访者，我们将$\mathcal{S}_{t_1} \subset \mathcal{S}_t$ 设为在时间 t 首次填写调查的受访者索引集，将$\mathcal{S}_{t_2} \subset \mathcal{S}_t$ 设为在时间 t. 进行后续调查的受访者索引集——即第二次或第三次填写调查的受访者。单位 $i \in S_{t_1}$的伪似然贡献为

$$y_{it}|\mu_{it}, \sigma^2 \propto \mathrm{N}(y_{it}|\mu_{it}, \sigma^2)^{\widetilde{w}_{it}},$$

其中，我们用 $N(y|\mu,\ \sigma^2)$ 表示均值为 μ、方差为σ^2 的高斯密度函数，其在 y 处的取值。对于 $i \in S_{t_2}$，其贡献为

$$y_{it}|y_{it-1}, \mu_{it}, \sigma^2, \rho \propto \mathrm{N}(y_{it}|\mu_{it} + \rho\widetilde{y}_{i(t-1)}, \sigma^2(1-\rho^2))^{\widetilde{w}_{it}}.$$

这里，$\mu_{it} = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{\psi}_i'\boldsymbol{\eta}_t$，其中 $\boldsymbol{x}_i$ 表示单位 i 已知协变量的 p–维向量，$\boldsymbol{\beta}$ 表示相应的固定效应参数向量。m–维向量 $\boldsymbol{\psi}_i$ 用于标示单位 i 所属的地理区域，而 $\boldsymbol{\eta}_t$ 则是时间 t对应的区域层级随机效应向量。（注意，$\boldsymbol{\psi}_i$ 可以替换为更

复杂的空间基函数集合。）此外， $\widetilde{y}_{it} = y_{it} - \mu_{it}$。自回归参数$\rho \in$ （-1，1）反映了同一受访者当前响应与先前响应之间的相关性。

随后，我们将此伪似然纳入以下层次模型，并将其称为高斯时间依赖单元层级模型（G-TULM），

$$
\begin{aligned}
\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\eta}, \rho, \sigma^2 &\propto \prod_{t=1}^{T} \prod_{i \in S_{t_1}} \mathrm{N}(y_{it}|\mu_{it}, \sigma^2)^{\widetilde{w}_{it}} \\
&\quad \times \prod_{i \in S_{t_2}} \mathrm{N}(y_{it}|\mu_{it} + \rho\widetilde{y}_{i(t-1)}, \sigma^2(1-\rho^2))^{\widetilde{w}_{it}} \\
\mu_{it} &= \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{\psi}_i\boldsymbol{\eta}_t \\
\boldsymbol{\beta} &\sim \mathrm{N}_p(0, \sigma_\beta^2 I_p) \\
\boldsymbol{\eta}_1|\sigma_{\eta_1}^2 &\sim \mathrm{N}_m(\mathbf{0}, \sigma_{\eta_1}^2 I_m) \\
\boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1}, \phi, \sigma_\eta^2 &\sim \mathrm{N}_m(\phi\boldsymbol{\eta}_{t-1}, \sigma_\eta^2 I_m) \\
\rho, \phi &\overset{ind.}{\sim} \mathrm{Unif}(-1, 1) \\
\sigma^2, \sigma_{\eta_1}^2, \sigma_\eta^2 &\overset{ind.}{\sim} \mathrm{IG}(a, b) \\
\sigma_\beta^2, a, b &> 0.
\end{aligned}
$$

ϕ 的值是与随机效应相关的滞后 1 阶自回归参数。因此，该模型中有两个层次的AR(1)结构：一个由 ϕ控制时间随机效应，另一个由 ρ控制纵向相关。

由于除 ρ 之外所有参数的全条件分布均为共轭分布，我们使用吉布斯采样器估计该模型。估计 ρ 时需要执行吉布斯采样中的 Metropolis 步骤。附录提供了采样算法及其推导的详细信息。

2.3 二项响应模型

对于二项情形，我们提出如下模型，并将其称为二项时间依赖单位层级模型（B-TULM），

$$
\begin{aligned}
\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\eta} &\propto \prod_{t=1}^{T}\prod_{i\in\mathcal{S}_t}\text{Binomial}(y_{it}|n_{it},\xi_{it})^{\widetilde{w}_{it}} \\
\text{logit}(\xi_{it}) &= \boldsymbol{x}_i'\boldsymbol{\beta}+\boldsymbol{\psi}_i\boldsymbol{\eta}_t \\
\boldsymbol{\beta} &\sim \text{N}_p(\mathbf{0},\sigma_\beta^2 I_p) \\
\boldsymbol{\eta}_1|\sigma_{\eta_1}^2 &\sim \text{N}_m(\mathbf{0},\sigma_{\eta_1}^2 I_m) \\
\boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1},\phi,\sigma_\eta^2 &\sim \text{N}_m(\phi\boldsymbol{\eta}_{t-1},\sigma_\eta^2 I_m), \qquad \text{for } t=2,\ldots,T \\
\phi &\sim \text{Unif}(-1,1) \\
\sigma_{\eta_1}^2,\sigma_\eta^2 &\overset{ind.}{\sim} IG(a,b) \\
\sigma_\beta, a, b &> 0,
\end{aligned}
$$

其中，我们用 Binomial$(y|n,\xi)$ 表示一个二项概率质量函数，其试验次数为n ，成功概率为 ξ ，并在 y 处取值。请注意，对于我们的示例应用，数据为二元数据，因此对于所有 i 和 t， $n_{it}=1$ 。

与高斯情形不同，我们无法轻易地在似然中纳入 AR(1)结构，因为二项伪似然没有观测误差分量。因此，不存在与前一个模型中的 ρ 等价的自相关参数。为了考虑受访者内相关性，我们改为向设计矩阵中添加一个合成的分类协变量。该协变量有三个水平，用于索引受访者是否属于以下情况之一：(1) 在前一个时间点被抽中，并对问题回答“是”；(2) 在前一个时间点被抽中，并回答“否”；或 (3) 在前一个时间点未被调查抽中。

我们使用 P ́olya-Gamma 用于二项似然的数据增广方案由

通过 Polson 等人 (2013)提出，用于实现完全吉布斯采样器。Polson 等人 (2013)定义随机变量 X 服从参数为 $b > 0$ 和 $c \in \mathbb{R}$ 的 P´olya-Gamma 分布，记作 $X \sim \text{PG}(b, c)$，若

$$X \overset{d}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k-1/2)^2 + c^2/(4\pi^2)},$$

其中 $g_k \overset{ind}{\sim}$ 伽马分布$(b,\ 1)$。他们为 P´olya-Gamma 随机变量提供了一种高效的采样算法，并证明了以下针对以对数优势 $\lambda = \text{logit}(\xi_{it}) = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{\psi}_i'\boldsymbol{\eta}_t$ 参数化的二项似然的积分恒等式，

$$\frac{(e^{\lambda})^a}{(1+e^{\lambda})^b} = 2^{-b} e^{\kappa\lambda} \int_0^{\infty} e^{-\omega\lambda^2/2} p(\omega) d\omega,$$

其中 $\kappa = a - b/2$ 且 $\omega \sim \text{PG}(b,\ 0)$。右侧可以视为 λ 在给定 ω条件下的条件分布，此时被积函数与高斯密度成正比。另一方面，将该表达式视为 ω 在给定λ 条件下的条件分布，则意味着它服从 P´olya-Gamma 分布。这启发了一种简单的吉布斯采样方案，交替抽取高斯随机变量和 P´olya-Gamma 随机变量。

与此同时，积分恒等式的左侧很容易调整为适用于我们的二项伪似然。在 B-TULM 模型中，我们有

$$\xi_{it} = \frac{\exp\{\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{\psi}_i\boldsymbol{\eta}_t\}}{1 + \exp\{\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{\psi}_i\boldsymbol{\eta}_t\}},$$

这使得单位 i 对二项伪似然的贡献可以写为

$$\begin{aligned} p(y_{it}|n_{it}, \xi_{it}) &\propto [\xi_{it}^{y_{it}}(1-\xi_{it})^{n_{it}-y_{it}}]^{\widetilde{w}_{it}} \\ &= \frac{(\exp\{\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{\psi}_i\boldsymbol{\eta}_t\})^{y_{it}\widetilde{w}_{it}}}{(1+\exp\{\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{\psi}_i\boldsymbol{\eta}_t\})^{n_{it}\widetilde{w}_{it}}}. \end{aligned}$$

这正是积分恒等式的左边，其中 $a = y_{it}\widetilde{w}_{it}$ 和 $b = n_{it}\widetilde{w}_{it}$.

将 P´olya-Gamma 潜变量引入层次模型后，便可得到标准形式的全条件分布，并可直接从中进行抽样。采样算法的更多细节见附录。

3 实证模拟研究

为评估所提出方法的性能，我们针对连续响应和二元响应开展实证模拟研究。我们使用美国人口普查局网站上提供的 HPS 前 12周相关公共使用微观数据文件。[3]我们将来自美国本土相连地区（包括哥伦比亚特区）的受访者作为总体。在两项模拟中，我们均剔除了未提供所考察响应值答案的受访者。

在高斯情形下，我们将响应变量设为受访者报告的“过去7天在家准备和食用的食物支出金额”。这些回答的取值为1至900之间的整数，且呈右偏分布。探索性分析表明，进行 Box-Cox 变换后，数据近似服从高斯分布。[4] 剔除项目无应答后，我们得到729,219个数值。对于二元响应，我们以作为研究动机的应用为依据，考察了受访者对“由于冠状病毒大流行，预计未来四周内家庭会失业”这一问题给出的819,509个回答。

两项模拟均运行 100 个样本。为保留后续响应率的自然模式，我们不会分别对每一周进行抽样，而是从总体中进行一次无放回抽样。若某个单位被抽中，则将其提供的所有每周回答都纳入样本。为了使样本具有信息性，我们采用 Brewer 等人提出的（1984年）泊松法进行与规模成比例的概率抽样。我们将预期样本量设为构建总体的 2%，并对每个

[3] https://www2.census.gov/programs-surveys/demo/datasets/hhp/2020/ [4] Box-Cox 变换使用 `forecast` 软件包完成，参数 $\lambda = -0.0863$由系统自动选择（Hyndman 等人，2022年）。

单位 i 按规模变量的比例

$$s_i = \exp(.1\overline{y}_i + .2w_i^*),$$

其中， w_i^* 是与单位 i的第一个回答相关联的未缩放原始调查权重，y_i 是单位 i各回答的平均值。

随后，我们拟合模型，将性别以及年龄的一次项和二次项作为协变量。如前所述，二元响应模型还包含一个额外的合成分类协变量，有三个水平，用于指示前一周记录的回答类型。对于高斯响应，吉布斯采样器运行 2,000 次迭代，并舍弃前 500 次作为预热期。对于二元响应，采样器运行 8,000 次迭代，并舍弃前 1,000 次作为预热期。我们通过目视检查随机选取的一部分参数的采样链轨迹图来评估收敛情况，未发现不收敛现象。

在采样器的每次迭代 k 中，我们使用总体协变量值，为每个区域在每个时间点生成后验预测。在高斯情形下，我们生成

$$\widehat{y}_{it}^{(k)}|\cdot \sim N(\boldsymbol{x}_i'\boldsymbol{\beta}^{(k)} + \boldsymbol{\psi}_i'\boldsymbol{\eta}_t{}^{(k)}, (\sigma^2)^{(k)})$$

适用于总体中的所有单位 i 和时间点 t 。在二元情形下，过程相同，但使用后验预测分布

$$\widehat{y}_{it}^{(k)}|\cdot \sim \text{Bernoulli}(\text{logit}^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{\psi}_i'\boldsymbol{\eta}_t)).$$

然后，我们汇总这些估计值，得到每个州 j和周 t组合对应的均值，

$$\widehat{\overline{y}}_{jt}^{(k)} = \frac{1}{|\mathcal{U}_j|}\sum_{i\in\mathcal{U}_j}\widehat{y}_{it}^{(k)}.$$

由于这是对采样器的所有迭代执行的，因此我们能够获得域级估计值的后验分布，并据此计算点估计值和可信区间。

作为比较基线，我们计算每个区域在每一周的直接估计，并针对每一周分别拟合横截面单位层级模型。在高斯情形下，我们将其与Parker 等人(2023）所使用的伪似然模型进行比较。该模型是对Battese 等人(1988年）提出的基本单位层级模型的贝叶斯扩展

$$
\begin{aligned}
\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\eta} &\propto \prod_{i\in S_t} \mathrm{N}(y_i|\mu_i,\sigma^2)^{\widetilde{w}_i} \\
\mu_i &= \boldsymbol{x}_i'\boldsymbol{\beta}+\boldsymbol{\psi}_i'\boldsymbol{\eta} \\
\boldsymbol{\eta}|\sigma_\eta^2 &\sim N_m(\mathbf{0}_m,\sigma_\eta^2\boldsymbol{I}_m) \\
\boldsymbol{\beta} &\sim N_p(\mathbf{0}_p,\sigma_\beta^2\boldsymbol{I}_p) \\
\sigma_\eta^2 &\sim \mathrm{IG}(a,b) \\
\sigma_\beta^2, a, b &> 0.
\end{aligned}
$$

我们将其称为高斯基线单位级模型（G-BULM）。

在二元情形下，我们使用Parker 等人（2022年提出的二元单位层级模型，并称之为二元基线单位级模型（B-BULM）

$$
\begin{aligned}
\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\eta} &\propto \prod_{i\in S_t} \mathrm{Binomial}(y_i|n_i,\xi_i)^{\widetilde{w}_i} \\
\mathrm{logit}(\xi_i) &= \boldsymbol{x}_i'\boldsymbol{\beta}+\boldsymbol{\psi}_i'\boldsymbol{\eta} \\
\boldsymbol{\eta}|\sigma_\eta^2 &\sim N_m(\mathbf{0}_m,\sigma_\eta^2\boldsymbol{I}_m) \\
\boldsymbol{\beta} &\sim N_p(\mathbf{0}_p,\sigma_\beta^2\boldsymbol{I}_p) \\
\sigma_\eta^2 &\sim \mathrm{IG}(a,b) \\
\sigma_\beta^2, a, b &> 0.
\end{aligned}
$$

在拟合全部四个模型时，我们设定 $a = b = 1$ 和 $\sigma_\beta^2 = 10^4$ ，以便为 $\boldsymbol{\beta}$、 $\sigma_{\eta_1}^2$和 σ_η^2指定宽松的先验分布。对于每个子样本和每个总体量，我们使用后验均值作为点估计，并构造 95% 可信区间，以量化点估计的不确定性。我们以均方误差（MSE）评估点估计的质量，因为它兼顾了偏差和方差。我们使用区间评分（Gneiting 和 Raftery，2007）评估可信区间的质量。对于给定的（$1-\alpha$） $\times$ 100% 置信区间、 $[l, u]$以及真实值 x，其计算方式如下：

$$S_\alpha(l, u; x) = (u - l) + \frac{2}{\alpha}(l - x)\mathbb{1}\{x < l\} + \frac{2}{\alpha}(x - u)\mathbb{1}\{x > u\},$$

其中，$\mathbb{1}$ 表示示性函数。请注意，区间评分越低越理想。区间评分既奖励较窄的区间，也会惩罚未包含真实值的区间，从而实现与均方误差类似的平衡。高斯响应模拟的结果，包括均方误差、绝对偏差、区间评分和95%可信区间（CI）覆盖率，见表1。所有指标均按区域、周以及全部100个子样本取平均值计算。

在均方误差和区间评分方面，G-TULM优于G-BULM和直接估计量。值得注意的是，直接估计量的偏差最低，这符合预期，因为直接估计量渐近设计无偏。尽管如此，直接估计量的均方误差仍远高于任一模型。

图1 展示了每个时间点两种模型和直接估计量的均方误差（在各州间取平均值）。每周，两种模型的均方误差都显著低于直接估计量。为了更清楚地了解 G-TULM 相对于 G-BULM 带来的均方误差下降，我们还在图 2 中绘制了所有区域的两种模型均方误差之比，并对时间取平均。除一个州外，该比值在所有州都小于 1，这表明与基线

表1：高斯响应的直接法和基于模型的估计的实证模拟研究结果。数值按每周、每个区域计算，再对时间、区域以及100次模拟取平均值。置信区间和区间评分的计算对象为 $\alpha = .05$.

方法	MSE	绝对偏差	置信区间覆盖率	区间评分
直接法	11.14×10^{-3}	8.921×10^{-3}	89%	5.280×10^{-1}
G-BULM	4.518×10^{-3}	1.379×10^{-2}	96%	3.352×10^{-1}
G-TULM	$\mathbf{3.086 \times 10^{-3}}$	1.759×10^{-2}	95%	$\mathbf{2.628 \times 10^{-1}}$

模型。

二元响应的结果见表2。图3 和4 分别展示了区域级均方误差图和均方误差时间序列图。在本次模拟中，B-TULM 相较于 B-BULM 在均方误差方面的改进更加显著，而且所提出的 B-TULM 在每个时间点和所有区域均有所改进。尽管直接估计量的置信区间覆盖率比两个模型更接近名义水平，但其区间过宽，导致区间评分较高。两种基于模型的方法的区间评分均显著低于直接估计量，表明其预测覆盖率更佳。此外，与 B-BULM 相比，B-TULM 大幅降低了均方误差和区间评分。

4 家庭脉搏调查分析

本节中，我们使用完整的 HPS 数据集，采用 B-TULM 模型对 COVID-19 第一阶段的“预期失业”进行启发性（二元）分析。上一节中的模拟表明，在

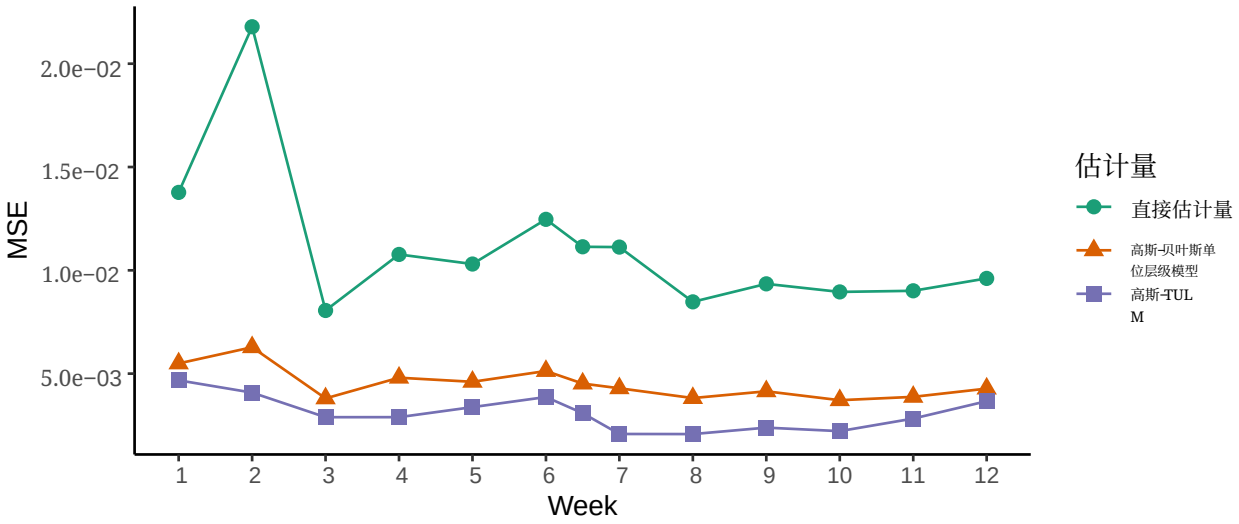

图1：基于经验模拟数据拟合的高斯响应模型与直接法估计量按周比较的均方误差。对于每个子样本，在每周计算各区域的均方误差，然后对区域和子样本取平均。

领域特定样本量较小的情况下，基于模型的估计具有优势。即使样本量大得多，通常也仍需要更细致的分类估计。在这种情况下，相较于基于设计的估计（直接估计量），基于模型的估计具有显著优势：对于样本量较小的区域，前者的标准误会极高；对于没有抽中任何样本值的区域，则无法计算估计值。

同样，在本文中，我们考察美国本土 HPS 第一阶段 12 周内的所有回答，并分析“预期失业”。该变量与我们在实证模拟研究（第 3 节）中使用的变量相同，因此我们使用相同的一组协变量。剔除因题项无应答而缺失的 7,000 条回答后，剩下 1,065,430 条回答和 844,547 名独立受访者。

我们将年龄划分为 13 个类别：一个类别涵盖 18 至 25 岁的受访者，11 个类别按五岁为间隔，涵盖 26 至 80 岁，最后一个类别涵盖 81 岁以上的人群。加上性别的两个类别、12周和 49 个区域，

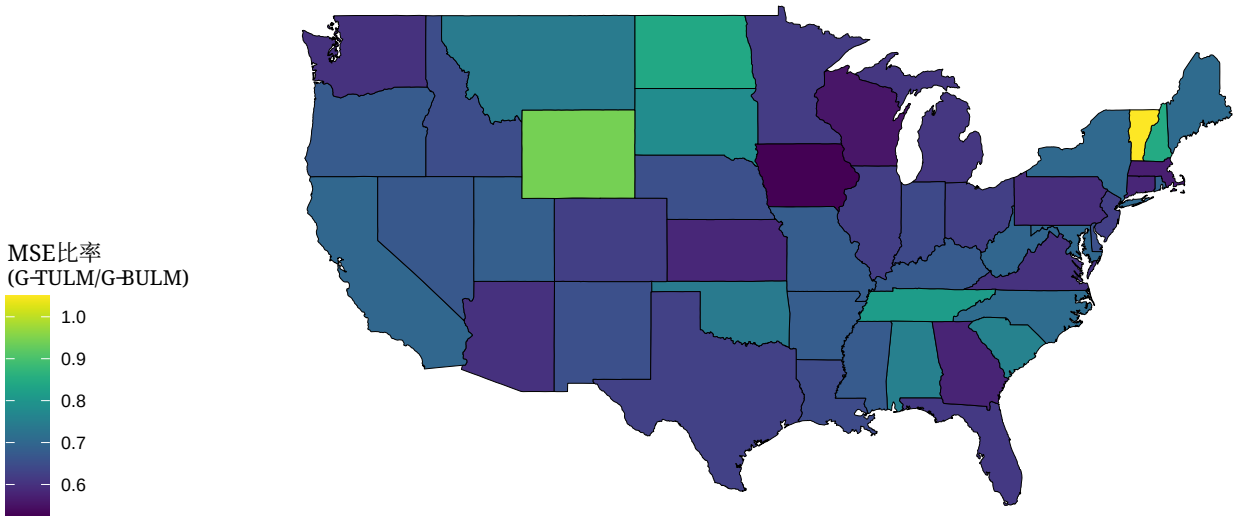

图2：将 G-TULM 模型的均方误差与 G-BULM 模型的均方误差进行比值计算，所用模型均拟合于经验模拟数据。对于每个子样本，先计算每周各区域的均方误差比率，再对各周及子样本取平均。

我们需要估计总体中的 15,288 个域（交叉分类单元）。在这种细分程度下，有些域在样本中仅由一个回答代表。这种情况在更精细的地理区域层级上也很常见。例如，考虑一项按人口普查区层级汇总数据的调查，有些人口普查区可能只有少数几个回答。

为了生成总体层面估计，我们从当前人口调查中获取每个单元的总体规模估计，并通过后验预测分布为总体中的每个个体生成合成回答值。由于同一单元的所有成员使用相同的协变量和随机效应值，因此我们只需为每个单元生成“是”回答数的后验预测。为在后分层过程中考虑受访者内相关性，我们将总体中的每个个体都视为第一周开始前尚未被抽样。在之后的所有周中，我们将单元计数拆分为回答“是”和回答“否”的人数

表 2：高斯响应的直接法估计和基于模型的估计的实证模拟研究结果。针对每个子样本的每一周和每个区域对计算所有指标。随后对这些数值按时间、区域和子样本取平均值。置信区间和区间评分针对 $\alpha = .05$. 计算

方法	MSE	绝对偏差	置信区间覆盖率	区间评分
直接法	6.332×10^{-3}	6.530×10^{-3}	96%	4.127×10^{-1}
B-BULM	1.586×10^{-3}	1.865×10^{-2}	99%	2.103×10^{-1}
B-TULM	$\mathbf{7.397 \times 10^{-4}}$	1.588×10^{-2}	98%	$\mathbf{1.324 \times 10^{-1}}$

前一　周，因为这些群体具有不同的合成协变量值。

我们还计算了直接估计，并在图5中绘出了所有周和区域的估计结果。图6 展示了 B-TULM 估计值的相同图形。由于直接估计量产生的许多估计值接近零或一（这是样本量较小所致），这两幅图的刻度范围有所不同。相比之下，基于模型的估计值分布范围窄得多。请注意，直接估计中存在相当大的噪声。例如，内布拉斯加州17至25岁男性在调查第二周预期家庭失业的估计比例接近一，而邻近州的估计比例则接近零。到了下一周，估计比例较高的州，其数值会急剧下降。这种在相邻（大概相似）区域之间或短时间尺度内的剧烈波动并不现实。与此相对，B-TULM的基于模型的估计在空间和时间上的变化平缓得多，而且通常不会超过0.5。

图7 展示了 B-TULM 标准误与直接估计值的标准误之比。对于任何区域或时间点，该比值均不超过 0。3，表明标准误大幅降低，且这种降低在各个时间点和所有区域均保持一致。

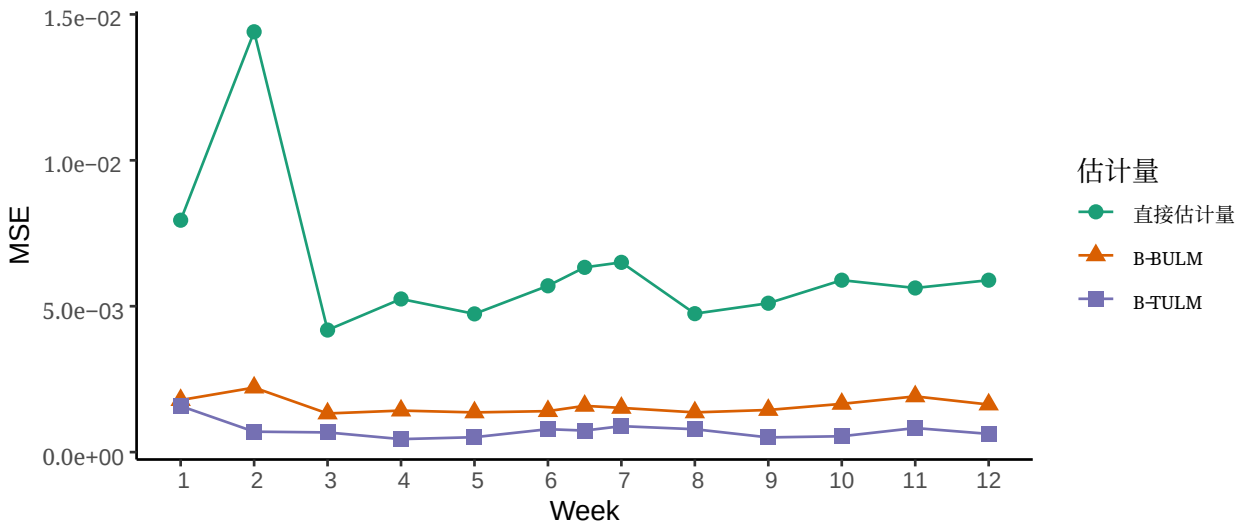

图3：基于经验模拟数据拟合的二元响应模型和直接估计量按周比较均方误差。对于每个子样本，计算每周各区域的均方误差，然后对各区域和子样本取平均。

进一步考察分析结果，还发现了其他一些有趣的发现。加利福尼亚州、佛罗里达州、佐治亚州、印第安纳州、内华达州和得克萨斯州在第12周的预期失业率均较高。回想一下，HPS 第12周的数据收集于2020年7月16日至2020年7月21日进行。因此，第11周的数据收集于7月14日结束。与此同时，加利福尼亚州于7月13日实施了一项重大停业令。[5] 这项停业令可能是该州第12周预期失业增加的原因之一。此外，这一日期大约在美国独立日假期周末一周之后，也与我们预计感染发展为病例和住院人数的时间相吻合。此外，佛罗里达州在7月26日创下 COVID-19 单日死亡人数纪录，而得克萨斯州则在当周早些时候创下死亡人数纪录。

内华达州的情况呈现出另一种值得注意的模式。具体而言，第五周和第六周的预期失业似乎有所下降，这大致与.....相吻合

[5] https://www.reuters.com/article/us-health-coronavirus-usa/california-shuts-down-businesses-schools

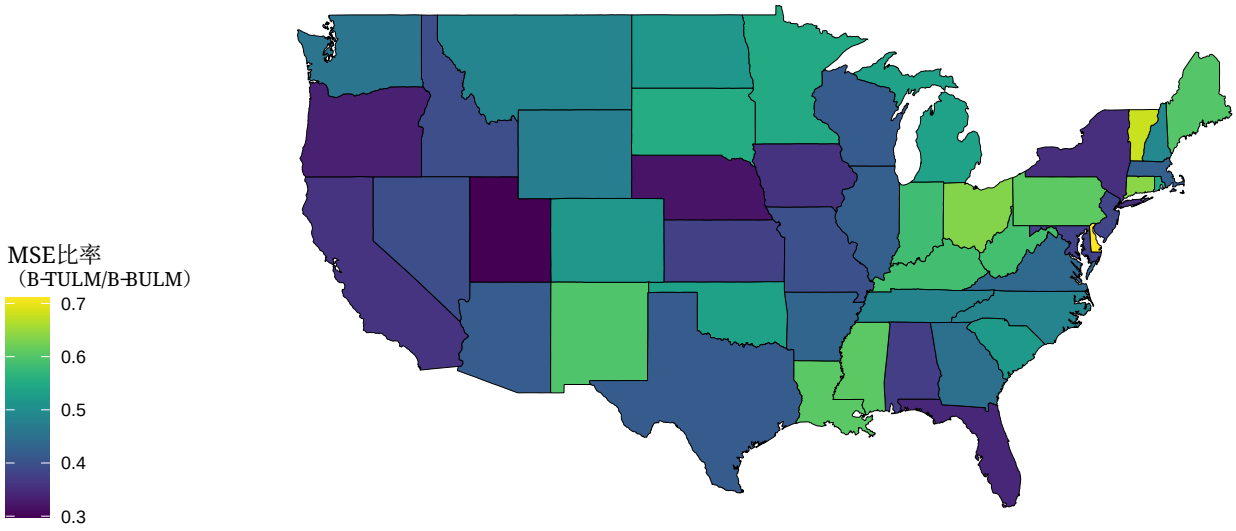

图 4：根据经验模拟数据拟合得到的 B-TULM 模型均方误差与 B-BULM 模型均方误差之比。对于每个子样本，先计算各周每个区域的均方误差比率，再对各周和子样本取平均。

随着赌场酒店在首轮限制措施解除后重新开放。[6]

为更好地展示这些趋势，图8 展示了各州每10万人中的病例数，并叠加了预期失业的TULM估计值。HPS数据收集每周结束时绘制的竖条以不同深浅着色，反映该周预期失业比例的TULM估计值。

5 讨论

本文提出了一种适用于具有纵向结构的调查数据的新型贝叶斯单元级方法。具体而言，我们针对高斯响应提出了一种多尺度模型，该模型既考虑了同一单位的一组回答随时间变化的相关性，也考虑了不同地理域之间随时间变化的相关性。此外，我们还提出了一种类似的模型，可用于

[6] https://www.rgj.com/story/news/2020/07/10/nevada-bars-closed-las-vegas-casinos-open-governor-sisolak5414497002/

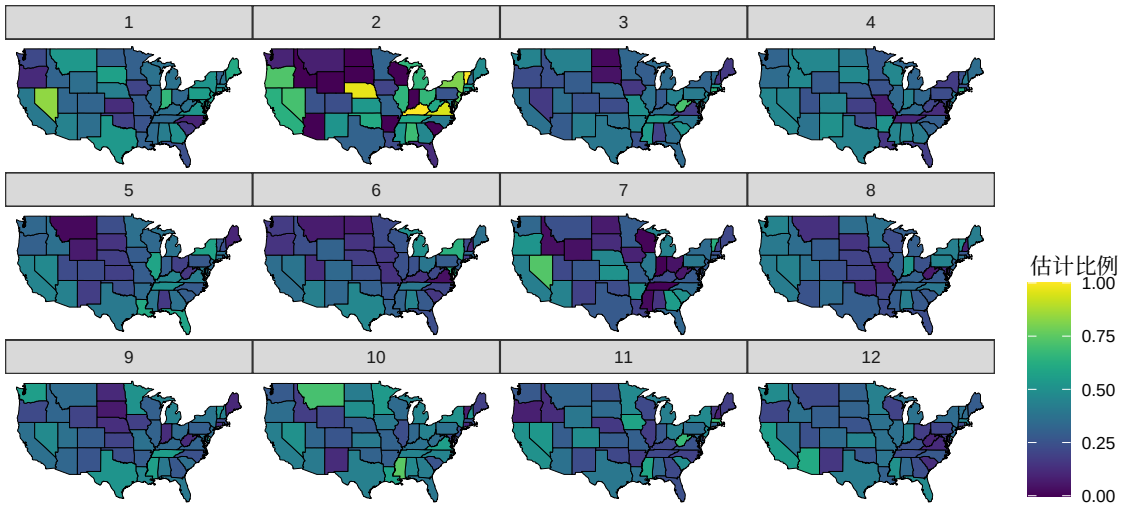

图5：基于全部 HPS第一阶段数据计算的、预期家庭失业的17至25岁男性比例的直接估计区域时间序列图。该图按周分面，左上角为第一个时间点，右下角为最后一个时间点。

二项数据，因为许多调查变量在单位层级上往往是二元的。在这两种情况下，我们都强调计算效率，通过构建具有条件共轭性的吉布斯采样器来实现这一点。值得注意的是，尽管我们在高性能计算集群上开展模拟研究以缩短运行时间，但本文提出的模型计算效率高，使用计算资源适中的独立笔记本电脑或工作站即可拟合。

通过模拟研究，我们表明，与传统直接估计量以及缺乏纵向或时间成分的现有基于模型的估计量相比，所提出的方法能够得到更精确的点估计和区间估计。此外，我们还运用所提出的方法，利用 HPS 数据估计 COVID-19疫情初期几周的预期失业情况。所得估计比直接估计更精确，因而可能具有重要价值

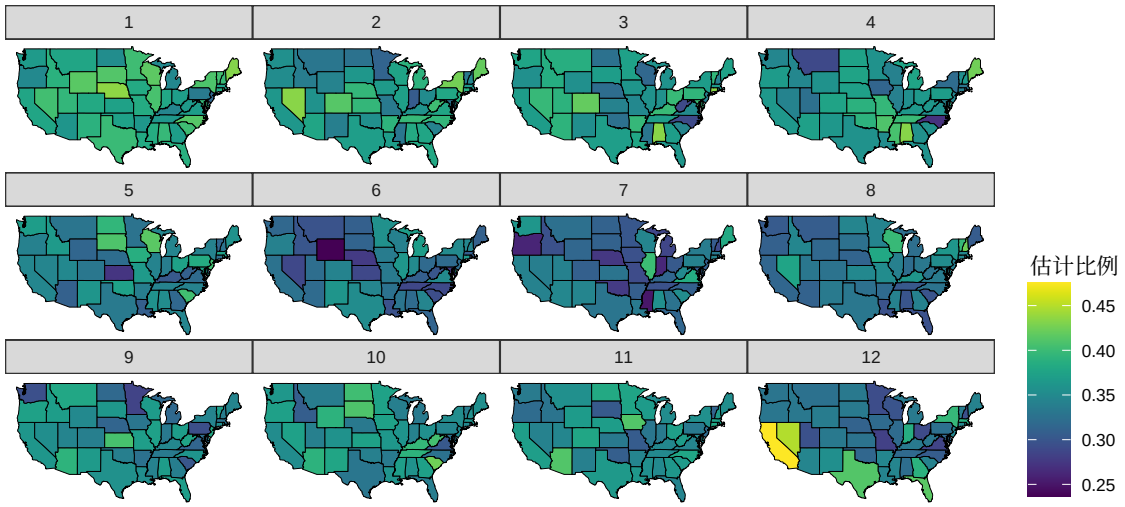

图6：B-TULM 对预期家庭失业的17至25岁男性比例的估计所绘制的区域时间序列图，模型基于 HPS第一阶段的全部数据进行拟合。该图按周分面，左上角为第一个时间点，右下角为最后一个时间点。请注意，颜色刻度与图5不同。

在评估大流行病的社会和经济影响时。

我们使用了 HPS 第一阶段的数据，但自那时以来，HPS 的设计和结构已经修订。例如，从第二阶段开始，受访者在后续周中不会再次被选中。因此，若要使用第一阶段之后的数据进行估计，可能需要对我们所提出的方法做一些调整。尽管如此，使用所提出的方法联合分析第一阶段数据与第一阶段之后的数据，可能会有所助益。

我们构建了适用于高斯数据和二项数据的模型，因此可以利用 HPS 中种类繁多的回答进行估计。不过，还有其他一些响应类型无法直接纳入我们当前的框架。例如，高阶分类变量在许多调查中十分常见。因此，将该方法推广至多项或有序响应类型，是未来的研究方向。

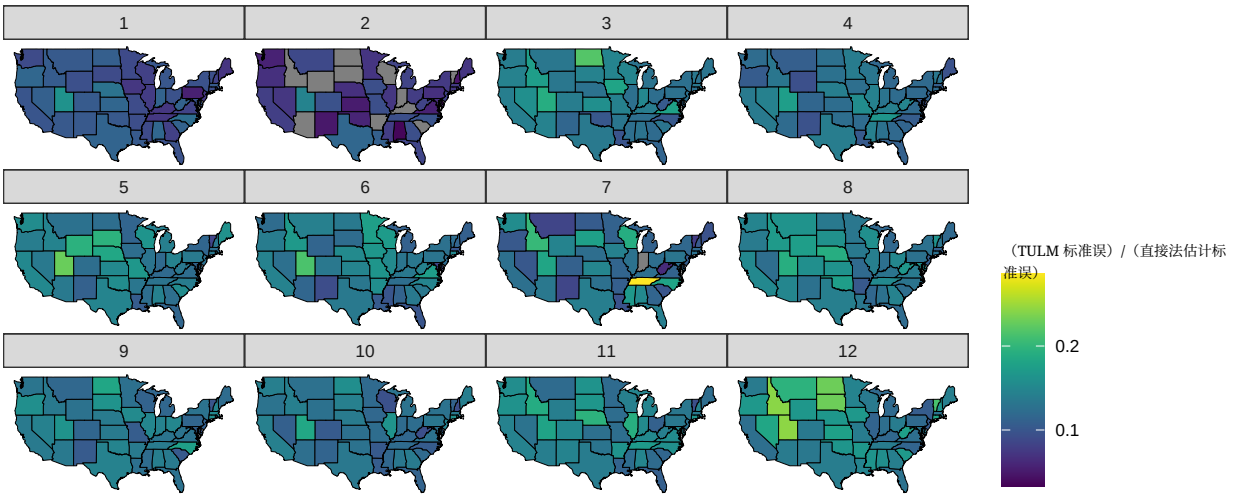

图7：基于完整的 HPS第一阶段数据拟合得到的区域层级时间序列图，展示 B-TULM 标准误与直接法估计标准误之比，用于估计预期家庭失业的 17至25岁男性比例。图按周分面，左上角为第一个时间点，右下角为最后一个时间点。灰色区域表示样本中的回答数量不足以计算直接法估计。

最后，尽管本研究主要受 HPS 启发，但也存在其他具有纵向结构的调查，例如收入与项目参与调查（SIPP）、当前人口调查（CPS）和月度零售贸易调查等。在某些情况下，这些调查采用的面板结构比 HPS 第一阶段所用的结构更为复杂。因此，本研究本身也具有独立价值，而且我们的方法很可能有助于利用其他调查的数据生成估计结果，或许只需稍作调整。分析其他具有纵向结构的调查数据，是未来研究的另一方向。

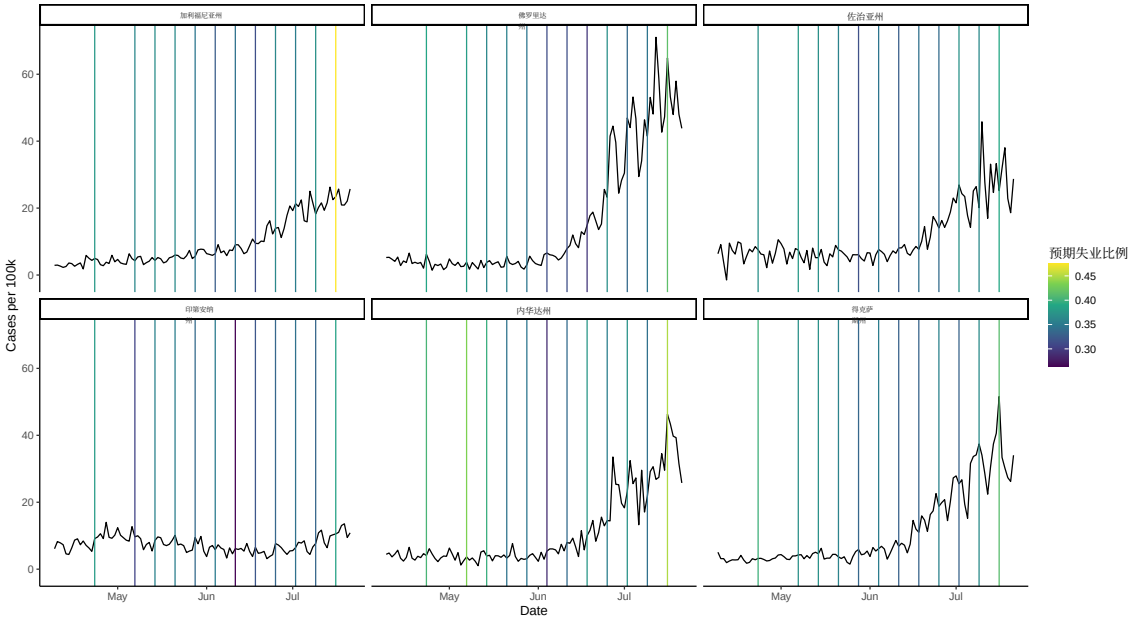

图8：HPS第一阶段部分州的COVID-19病例数（每10万人），以及TULM预期失业估计值。病例数数据来自《纽约时报》，通过covdata R 库（Healy，2020年）。

致谢

本文发布旨在向相关人士介绍正在进行的研究，并鼓励开展讨论。关于统计问题的观点仅代表作者本人，不代表美国国家科学基金会或美国人口普查局。

资助

本研究部分由美国国家科学基金会（NSF）通过资助项目 SES-1853096、NCSE-2215168 和 NCSE-2215169 提供支持。

参考文献

Battese, G. E.、Harter, R. M. 和 Fuller, W. A.（1988年）。“利用调查和卫星数据预测县级作物面积的误差分量模型。”《美国统计学会会刊》，83，401，28–36。

Binder, D. A.（1983年）。“复杂调查中渐近正态估计量的方差。”《国际统计评论》，51，3，279。

Brewer, K.、Early, L. 和 Hanif, M.（1984年）。“泊松抽样、修正泊松抽样与共址抽样。”《统计规划与推断杂志》，10，1，15–30。

Francis, D. V. 和 Weller, C. E.（2021）。“COVID-19期间的经济不平等、数字鸿沟与远程学习。”《黑人政治经济学评论》，49，1，41–60。

Gneiting, T. 和 Raftery, A. E.（2007）。“严格适当评分规则、预测与估计。”《美国统计学会会刊》，102，477，359–378。

Guadarrama, M.、Morales, D. 和 Molina, I.（2020）。“单位层级模型下一般参数的时间稳定小区域估计。”卢森堡社会经济研究所（LISER）工作论文系列，10。

Healy, K.（2020年）。covdata：COVID-19 病例与死亡时间序列。R软件包版本0.5.2。

Hidiroglou, M. A. 和 You, Y.（2016年）。“单位层级与区域层级小区域估计量的比较。”《调查方法学》，42，41–61。

Hobza, T. 和 Morales, D.（2016年）。“单位层级Logit混合模型下的经验最佳预测”。《官方统计学杂志》，32，3，661–692。

Hobza, T.、Morales, D. 和 Santamar´ıa, L.（2017）。“单位级时序二项逻辑混合模型下贫困比例的小区域估计”。TEST，27，2，270–294。

Horvitz, D. G. 和 Thompson, D. J. (1952)。“有限总体不放回抽样的一种推广。”《美国统计学会会刊》，47，260，663–685。

Hyndman, R.、Athanasopoulos, G.、Bergmeir, C.、Caceres, G.、Chhay, L.、O’Hara-Wild, M.、Petropoulos, F.、Razbash, S.、Wang, E. 和 Yasmeen, F. (2022年)。forecast：用于时间序列和线性模型的预测函数。R软件包，版本 8.16。

Kunihama, T.、Halpern, C. T. 和 Herring, A. H. (2019)。“混合尺度纵向调查的非参数贝叶斯模型。”《英国皇家统计学会会刊C辑：应用统计》，68，4，1091–1109。

Parker, P. A.、Holan, S. H. 和 Janicki, R. (2022年)。“信息性抽样下用于非高斯数据的计算高效贝叶斯单位级模型：在医疗保险覆盖率估计中的应用。”《应用统计年刊》，16，2，887–904。

—（2023）。“具有异方差结构的小区域估计的共轭建模方法。”《调查统计与方法学杂志》。

Parker, P. A.、Janicki, R. 和 Holan, S. H.（2019）。“具有信息性抽样的调查数据单位层级模型：扩展的综合概述。” arXiv：arXiv 预印本 1908.10488。

Polson, N. G., Scott, J. G., 和 Windle, J. (2013).《逻辑模型的贝叶斯推断

使用 P´olya–Gamma 潜变量》。《美国统计学会会刊》，108，504，1339–1349。

Rao, J. N. K. 和 Molina, I.（2015年）。《小区域估计》。新泽西州霍博肯：Wiley。

Savitsky, T. D. 和 Toth, D.（2016年）。《信息性抽样下的贝叶斯估计》。《电子统计学期刊》，10，1，1677–1708。

Schanzenbach, D. 和 Pitts, A.（2020年）。《粮食不安全程度上升了多少？来自人口普查家庭脉搏调查的证据》。《政策研究所快速研究报告》。

Skinner, C. J.（1989年）。《领域均值、回归与多变量分析》。载于《复杂抽样调查分析》，C. J. Skinner、D. Holt 和 T. M. F. Smith 编，第2章，59–84页。奇切斯特：Wiley。

Sun, A.、Parker, P. A. 和 Holan, S. H.（2022年）。《通过信息性抽样的单位层级模型分析家庭脉搏调查公开使用微观数据》。《统计学》，5，1，139–153。

Vahratian, A., Blumberg, S. J., Terlizzi, E. P., and Schiller, J. S.（2021）。“COVID-19疫情期间美国成年人焦虑或抑郁障碍症状及心理健康护理的使用情况——美国，2020年8月至2021年2月。”《发病率与死亡率周报》，70，13，490–494。

Zhang, X., Holt, J. B., Lu, H., Wheaton, A. G., Ford, E. S., Greenlund, K. J., and Croft, J. B. (2014). “基于多层回归与后分层的小区域人口健康结局估计：慢性阻塞性肺疾病病例研究

患病率：使用行为风险因素监测系统。”《美国流行病学杂志》，179，8，1025–1033。

A 附录

A.1 G-TULM 的采样算法

A.1.1 符号说明

描述全条件分布需要一些额外的符号，我们在此引入。定义 $\boldsymbol{W}$ 为所有时间点上所有受访者的调查权重对角矩阵，其中（$\boldsymbol{W}$）$_{ii} = w_{ii}$，

$$\boldsymbol{W} = (w_{11}, w_{21}, \ldots, w_{n_1 1}, w_{12}, w_{22} \ldots, w_{n_T T})$$

类似地，令 $\boldsymbol{\Psi}$ 为关联向量的 $n \times m$ 矩阵，其中（$\boldsymbol{\Psi}$）$_{ij} = 1$：若第 i 行对应的单位属于区域 j，则该项为相应值，其余项均设为零。令 $\boldsymbol{X}$ 为设计矩阵，其第 i行为协变量 $\boldsymbol{x}_i$ 组成的 p–维向量。最后，我们令$\boldsymbol{\eta}$ 为每个时间点的随机效应向量的拼接，

$$\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \boldsymbol{\eta}_2', \ldots, \boldsymbol{\eta}_T') = (\eta_{11}, \eta_{21}, \ldots, \eta_{n_1 1}, \eta_{12}, \eta_{22} \ldots, \eta_{n_T T}).$$

对于这些矩阵和向量中的每一个，我们用下标 t 表示对应于时间 t收集的回答的条目子集。我们用上标 f表示首次受访者对应的条目子集，用上标 r 表示后续调查受访者对应的条目子集。下标和上标有多种组合方式。例如，

- $\boldsymbol{X}_t^r$ 是时间点抽样的所有后续调查受访者的协变量矩阵 t.

- $\boldsymbol{W}^f$ 是样本中所有时间点的所有后续调查受访者对应的调查权重矩阵

- $\boldsymbol{\Psi}_2$ 是时间 2 所有受访者（包括首次作答者和后续作答者）的指示向量矩阵。

- $\boldsymbol{\eta}$ 表示所有时间的所有回答对应的随机效应向量。

对于响应向量 $\boldsymbol{y}$，我们定义额外的上标 r_1 和 r_2，使得$\boldsymbol{y}^{r_2}$ 为所有后续作答的向量，而 $\boldsymbol{y}^{r_1}$ 为前几周相应回答组成的向量。采用这一记号，来自某个家庭的首次响应 i 记录为条目 y_i^f 。如果该家庭给出第二次响应，则该响应记为 $y_{i'}^{r_2}$，而首次响应则重复记为 $y_{i'}^{r_1}$，其中 i' 不一定等于 i。第三次响应将记录为另一个条目 $y_{i''}^{r_2}$ ，而 $y_{i''}^{r_1} = y_{i'}^{r_2}$ 将重复第二次响应。这种“重复”简化了索引过程，便于我们引用后续作答以及该单位前一周的相应回答。

对于二元情形，我们还有 P´olya-Gamma 数据增广方案中的潜变量向量

$$\boldsymbol{\Omega} = \mathrm{diag}(\omega_{11}, \omega_{21}, \ldots, \omega_{1T}, \omega_{2T}, \ldots, \omega_{n_T T})$$

以及额外的参数向量

$$\boldsymbol{\kappa} = (\widetilde{w}_{11}(y_{11}-1/2)\ \widetilde{w}_{12}(y_{12}-1/2)\ldots, \widetilde{w}_{1T}(y_{1T}-1/2)\ \widetilde{w}_{21}(y_{21}-1/2)\ldots, \widetilde{w}_{n_T T}(y_{n_T T}-1/2))'.$$

与之前一样，对于任意给定的 t，令 $\boldsymbol{\Omega}_t = \mathrm{diag}(\omega_{1t},\ \omega_{2t},\ \ldots\ldots,\ \omega_{n_t t})$；也就是说，$\boldsymbol{\Omega}$中仅对应时间 t 的元素。同样，令 $\boldsymbol{\kappa}_t = (\kappa_{1t},\ \kappa_{2t},\ \ldots\ldots,\ \kappa_{n_t t})$。最后，请注意，$\boldsymbol{\kappa}/\boldsymbol{\omega}$ 表示逐元素除法。

A.1.2 算法

1. 样本 $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}|\cdot \sim N_p\Bigg(\boldsymbol{\mu}_\beta = \sigma_\beta\left(\frac{1}{\sigma^2}(X^f)'W^f\boldsymbol{u}_\beta - \frac{1}{\sigma^2(1+\rho)}(X^r)'W^r\boldsymbol{v}_\beta\right), \quad ,$$
$$\Sigma_\beta = \left(\frac{1}{\sigma_\beta^2}I_p + \frac{1}{\sigma^2}(X^f)'W^fX^f - \frac{1-\rho}{\sigma^2(1+\rho)}(X^r)'W^rX^r\right)^{-1}\Bigg)$$

其中 $\boldsymbol{u}_\beta = \boldsymbol{y}^f - \boldsymbol{\eta}^f$ 和 $\boldsymbol{v}_\beta = \boldsymbol{y}^{r_2} - \boldsymbol{\eta}^{r_2} - \rho(\boldsymbol{y}^{r_1} - \boldsymbol{\eta}^{r_1})$。

2. 样本 $\boldsymbol{\eta}_1$:

$$\boldsymbol{\eta}_1|\cdot \sim N_m\Bigg(\boldsymbol{\mu}_{\eta_1} = \Sigma_{\eta_1}\left(\frac{1}{\sigma^2}(\Psi_1^f)'W_1^f\boldsymbol{u}_{\eta_1} + \frac{\rho}{\sigma^2(1-\rho^2)}(\Psi_2^r)'W_2^r\boldsymbol{v}_{\eta_1} + \frac{\phi}{\sigma_\eta^2}\boldsymbol{\eta}_2\right),$$
$$\Sigma_{\eta_1} = \left(\frac{1}{\sigma^2}(\Psi_1^f)'W_1^f\Psi_1^f + \frac{\rho^2}{\sigma^2(1-\rho^2)}\Psi_2^rW_2^r\Psi_2^r + \left(\frac{1}{100} + \frac{\phi^2}{\sigma_\eta^2}\right)I_m\right)^{-1}\Bigg)$$

其中 $\boldsymbol{u}_{\eta_1} = y^f - X^f\boldsymbol{\beta}$和 $\boldsymbol{v}_{\eta_t} = -(\boldsymbol{y}_2^{r_2} - X_2^r\boldsymbol{\beta} - \Psi_2^r\boldsymbol{\eta}_2 - \rho(\boldsymbol{y}_2^{r_1} - X_2^r\boldsymbol{\beta}))$.

3. 样本 $\boldsymbol{\eta}_t$ 适用于 $t = 2, \ldots, T-1$:

$$\boldsymbol{\eta}_t|\cdot \sim N_m\Bigg(\boldsymbol{\mu}_{\eta_t} = \Sigma_{\eta_t}\Bigg(\frac{\phi}{\sigma_\eta^2}(\boldsymbol{\eta}_{t-1} + \boldsymbol{\eta}_{t+1}) \quad ,$$
$$+ \frac{1}{\sigma^2}\left((\Psi_t^f)'W_t^f\boldsymbol{u}_{\eta_t} + \frac{1}{1-\rho^2}\left((\Psi_t^r)'W_t^r\boldsymbol{v}_{\eta_t} - \rho(\Psi_{t+1}^r)'W_{t+1}^r\boldsymbol{w}_{\eta_t}\right)\right)\Bigg),$$
$$\Sigma_{\eta_t} = \Bigg(\frac{1+\phi^2}{\sigma_\eta^2}I_m$$
$$+ \frac{1}{\sigma^2}\left((\Psi_t^f)'W_t^f\Psi_t^f + \frac{1}{1-\rho^2}\left(\left(\Psi_t^r\right)'W_t^r\Psi_t^r + \rho^2(\Psi_{t+1}^r)'W_{t+1}^r\Psi_{t+1}^r\right)\right)\Bigg)^{-1}\Bigg),$$

其中

- $\boldsymbol{u}_{\eta_t} = \boldsymbol{y}_t^f - X_t^f\boldsymbol{\beta}$,
- $\boldsymbol{v}_{\eta_t} = \boldsymbol{y}_t^{r_2} - (1-\rho)X_t^r\boldsymbol{\beta} - \rho(\boldsymbol{y}_t^{r_1} - \Psi_t^r\boldsymbol{\eta}_{t-1})$,
- $\boldsymbol{w}_{\eta_t} = \Psi_{t+1}^r\boldsymbol{\eta}_{t+1} + (1-\rho)X^{r_2}\boldsymbol{\beta} - (\boldsymbol{y}_{t+1}^{r_2} - \rho\boldsymbol{y}_{t+1}^{r_1})$.

4. 样本 $\boldsymbol{\eta}_T$:

$$\boldsymbol{\eta}_T|\cdot \sim N_m\Bigg(\boldsymbol{\mu}_{\eta_T} = \Sigma_{\eta_T}\left(\frac{\phi}{\sigma_\eta^2}\boldsymbol{\eta}_{T-1} + \frac{1}{\sigma^2}\left((\Psi_T^f)W_T^f\boldsymbol{u}_{\eta_T} + \frac{1}{1-\rho^2}(\Psi_T^r)W_T^r\boldsymbol{v}_{\eta_T}\right)\right),$$
$$\Sigma_{\eta_T} = \left(\frac{1}{\sigma_\eta^2}I_m + \frac{1}{\sigma^2}\left((\Psi_T^f)'W_T^f\Psi_T^f + \frac{1}{1-\rho^2}((\Psi_T^r)'W_T^r\Psi_T^r)\right)\right)^{-1}\Bigg)$$

其中 $\boldsymbol{u}_{\eta_T} = \boldsymbol{y}_T^f - X_T^f\boldsymbol{\beta}$和 $\boldsymbol{v}_{\eta_T} = \boldsymbol{y}_T^{r_2} - X_T^{r_2}\boldsymbol{\beta} - \rho(\boldsymbol{y}_T^{r_1} - X_T^r\boldsymbol{\beta} - \Psi_T^r\boldsymbol{\eta}_{T-1})$。

5. 使用 Metropolis 步骤对 ρ 进行抽样。全条件分布为

$$\begin{aligned} p(\rho|\cdot) &\propto (1-\rho^2)^{-N_r/2} \\ &\times \exp\left\{-\frac{1}{2\sigma^2(1-\rho^2)}\left((\rho^2-\rho)\widetilde{\boldsymbol{y}}^{r_1} - \rho\widetilde{\boldsymbol{y}}^{r_2})'W^r((\rho^2-\rho)\widetilde{\boldsymbol{y}}^{r_1} - \rho\widetilde{\boldsymbol{y}}^{r_2})\right)\right\} \\ &\times \mathbb{I}(-1<\rho<1). \end{aligned}$$

无法直接从该分布中抽样，因此我们使用随机游走 Metropolis 算法，其均匀提议分布以 ρ.的上一次抽样值为中心

6. 对 ϕ进行抽样:

$$\phi|\cdot \sim \text{TruncNorm}\left(\frac{\sum_{t=2}^T \boldsymbol{\eta}_t'\boldsymbol{\eta}_{t-1}}{\sum_{t=2}^T \boldsymbol{\eta}_{t-1}'\boldsymbol{\eta}_{t-1}}, \frac{\sigma_\eta^2}{\sum_{t=2}^T \boldsymbol{\eta}_{t-1}'\boldsymbol{\eta}_{t-1}}, -1, 1\right)$$

7. 对 σ^2进行抽样:

$$\sigma^2|\cdot \sim \text{IG}\left(1+\frac{N}{2}, 1+\frac{1}{2}(\widetilde{\boldsymbol{y}}^f)'W^f\widetilde{\boldsymbol{y}}^f + \frac{1}{2(1-\rho^2)}(\widetilde{\boldsymbol{y}}^{r_2} - \rho\widetilde{\boldsymbol{y}}^{r_1})'W^r(\widetilde{\boldsymbol{y}}^{r_2} - \rho\widetilde{\boldsymbol{y}}^{r_1})\right)$$

8. 对 $\sigma_{\eta_1}^2$进行抽样:

$$\sigma_{\eta_1}^2|\cdot \sim \text{IG}\left(a + \frac{m(T-1)}{2}, b + \frac{1}{2}\boldsymbol{\eta}_1'\boldsymbol{\eta}_1\right)$$

9. 样本 σ_η^2:

$$\sigma_\eta^2|\cdot \sim \text{IG}\left(a + \frac{m(T-1)}{2}, b + \frac{1}{2}\sum_{t=2}^T(\boldsymbol{\eta}_t - \phi\boldsymbol{\eta}_{t-1})'(\boldsymbol{\eta}_t - \phi\boldsymbol{\eta}_{t-1})\right)$$

A.2 B-TULM 的采样算法

1. 样本 ω_{it} 对于 $t=1$、.....、T，对于 $i=1$、.....、n_t:

$$\omega_{it}|\cdot \sim \mathrm{PG}(\widetilde{w}_{it}n_{it}, \boldsymbol{x}_i/\boldsymbol{\beta}+\boldsymbol{\psi}_i'\boldsymbol{\eta}_t).$$

2. 样本 $\boldsymbol{\eta}_1$:

$$\boldsymbol{\eta}_1|\cdot \sim \mathrm{N}_m\left(\boldsymbol{\mu}=\boldsymbol{\Sigma}\left(\boldsymbol{\Psi}_1'\boldsymbol{\Omega}_1(\boldsymbol{\kappa}_1/\boldsymbol{\omega}_1-X_1\boldsymbol{\beta})+\frac{\phi}{\sigma_\eta^2}\boldsymbol{\eta}_2\right), \boldsymbol{\Sigma}=\left(\Psi_1'\boldsymbol{\Omega}_1\Psi_1+\left(\frac{1}{\sigma_{\eta_1}^2}+\frac{\phi^2}{\sigma_\eta^2}\right)I_m\right)^{-1}\right).$$

3. 样本 $\boldsymbol{\eta}_t$ 针对 $t=2,..,T-1$:

$$\boldsymbol{\eta}_t|\cdot \sim \mathrm{N}_p\Bigg(\boldsymbol{\mu}=\boldsymbol{\Sigma}\left(\Psi_t'\boldsymbol{\Omega}_t(\boldsymbol{\kappa}_t/\boldsymbol{\omega}_t-X_t\boldsymbol{\beta})+\frac{\phi}{\sigma_\eta^2}(\boldsymbol{\eta}_{t-1}+\boldsymbol{\eta}_{t+1})\right),$$
$$\boldsymbol{\Sigma}=\left(\Psi_t'\boldsymbol{\Omega}_t\Psi_t+\left(\frac{1}{\sigma_\eta^2}+\frac{\phi^2}{\sigma_\eta^2}\right)I_m\right)^{-1}\Bigg).$$

4. 样本 η_T:

$$\boldsymbol{\eta}_T|\cdot \sim \mathrm{N}_p\left(\boldsymbol{\mu}=\boldsymbol{\Sigma}\left(\Psi_T'\boldsymbol{\Omega}_T(\boldsymbol{\kappa}_T/\boldsymbol{\omega}_T-X_T\boldsymbol{\beta})+\frac{\phi}{\sigma_\eta^2}\boldsymbol{\eta}_{T-1}\right), \boldsymbol{\Sigma}=\left(\Psi_T'\boldsymbol{\Omega}_T\Psi_t+\frac{1}{\sigma_\eta^2}I_m\right)^{-1}\right).$$

5. 样本 $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}|\cdot \sim \mathrm{N}_p\left(\boldsymbol{\mu}=\boldsymbol{\Sigma}\left(\boldsymbol{X}'\boldsymbol{\Omega}(\boldsymbol{\kappa}/\boldsymbol{\omega}-\boldsymbol{\Psi}^f\boldsymbol{\eta}^f\right), \boldsymbol{\Sigma}=\left(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}+\frac{1}{\sigma_\beta^2}I_p\right)^{-1}\right).$$

6. 样本 ϕ:

$$\phi|\cdot \sim \mathrm{TruncNorm}\left(\frac{\sum_{t=2}^T\boldsymbol{\eta}_t'\boldsymbol{\eta}_{t-1}}{\sum_{t=2}^T\boldsymbol{\eta}_{t-1}'\boldsymbol{\eta}_{t-1}}, \frac{\sigma_\eta^2}{\sum_{t=2}^T\boldsymbol{\eta}_{t-1}'\boldsymbol{\eta}_{t-1}}, -1, 1\right).$$

7. 样本$\sigma_{\eta_1}^2$:

$$\sigma_{\eta_1}^2|\cdot \sim \mathrm{IG}\left(a+\frac{m}{2}, b+\frac{\boldsymbol{\eta}_1'\boldsymbol{\eta}_1}{2}\right).$$

8. 样本 σ_η^2:

$$\sigma_\eta^2|\cdot \sim \mathrm{IG}\left(a+\frac{m(T-1)}{2}, b+\frac{1}{2}\sum_{t=2}^T(\boldsymbol{\eta}_t-\phi\boldsymbol{\eta}_{t-1})'(\boldsymbol{\eta}_t-\phi\boldsymbol{\eta}_{t-1})\right).$$