arXiv:2602.09279v1 [stat.ME] 9 Feb 2026

纵向计数数据零膨胀Beta-二项混合模型的随机EM估计与推断

John Barrera *
数学工程研究所
工程学院
瓦尔帕莱索大学
智利瓦尔帕莱索

Ana Arribas-Gil
统计系
马德里卡洛斯三世大学
赫塔费，西班牙

Dae-Jin Lee
科学与技术学院
IE大学
马德里，西班牙

Cristian Meza
CIMFAV
瓦尔帕莱索大学
瓦尔帕莱索，智利

摘要

分析过度离散、零膨胀的纵向计数数据会带来重大的建模与计算挑战，而标准计数模型（例如泊松模型或负二项混合效应模型）无法充分应对这些挑战。我们提出了一种零膨胀贝塔-二项混合效应回归（ZIBBMR）模型，该模型在贝塔-二项计数模型的基础上增添了零膨胀分量、协变量固定效应和个体特异性随机效应，从而能够处理过多的零值、过度离散以及受试者内相关性。我们通过采用潜变量增广的随机逼近期望最大化（SAEM）算法进行最大似然估计，规避了模型难以处理的似然计算，并实现了高效计算。模拟研究表明，ZIBBMR 的准确性与文献中领先的混合模型方法相当，并且优于更简单的零膨胀计数模型形式，尤其是在小样本情形下。作为案例研究，我们分析了纵向微生物组数据，并将 ZIBBMR 与外部的零膨胀贝塔回归（ZIBR）基准模型进行比较；结果表明，当两种数据类型均可获得时，并行应用计数模型和基于比例的模型能够增强推断稳健性。

关键词 ze ro-膨胀 Beta-二项分布、混合效应、SAEM、《微生物组》、纵向数据、过度离散 n

1 引言

纵向计数数据在生物医学、环境科学和社会科学中十分常见，且通常具有三项关键复杂性：零值过多、过度离散，以及重复测量导致的受试者内相关性。例如，《微生物组》研究中的微生物序列计数（Romero 等，2014年）、流行病学调查中的寄生虫负荷，以及制造过程中的缺陷计数。标准计数模型，例如泊松或负二项混合效应模型，通常无法同时处理所有这些特征，这可能导致估计偏差、推断不可靠以及预测性能下降（McCullagh 和 Nelder，1989；Hall，2005）。

贝塔-二项分布非常适合对过度离散的二项数据进行建模（Skellam，1948；Najera-Zuloaga 等，2019），而零膨胀模型则能够捕捉超出假定计数分布预期的结构性零值（Lambert，1992）。进行重复测量时，混合效应结构可以处理

*通讯作者。电子邮箱：john.barrera@postgrado。 uv.cl

受试者之间未观测到的异质性（Pinheiro 和 Bates，2006）。然而，将零膨胀、贝塔-二项分布建模和个体特异性随机效应结合起来，会产生解析上难以处理的似然函数，使最大似然估计变得复杂。

我们通过引入零膨胀贝塔-二项混合效应回归（ZIBBMR）模型来应对这一挑战。该模型能够同时处理纵向二项型数据中的结构性零值、超二项变异和受试者内相关性。参数通过基于随机逼近期望–最大化（SAEM）算法（Delyon 等，1999）的最大似然估计程序进行估计。该方法结合潜变量增广和随机模拟，以克服计算困难，从而能够对复杂的零膨胀混合模型进行基于似然的推断。具体而言，当期望步难以处理时，SAEM算法可作为 EM 算法的替代方法；它通过模拟近似条件期望，同时保留 EM 算法的收敛性质，从而解决这一问题。本文其余部分的组织结构如下。第 2 节介绍零膨胀 Beta-二项混合效应回归（ZIBBMR）模型、其层次表示，以及通过 SAEM 算法进行的最大似然估计，包括层次增强和实现细节。第 3 节报告模拟研究，将所提出的估计量与广义线性混合模型的一般估计方法（如 glmmTMB）（Brooks 等人，2017年；McGillycuddy 等人，2025年）或广义加性模型的方法（如 gamlss）（Rigby 和 Stasinopoulos，2005；Stasinopoulos 等人，2017年）进行比较，并介绍研究设计、性能指标和结果。第 4 节转向应用部分，介绍纵向阴道微生物组数据集的背景和研究动机，并将 ZIBBMR 与另一种零膨胀 Beta 回归（ZIBR）基准模型进行比较。第 5 节通过讨论和展望未来研究方向作出总结。R 实现和可复现材料总结于“软件”部分和补充材料中。

2 模型定义、估计与推断

2.1 零膨胀 Beta-二项分布混合回归（ZIBBMR）模型的定义

设 Y_{it} 为受试者 i （$1 \le i \le N$）在测量时点 t （$1 \le t \le T_i$）观测到的计数， S_{it} 为已知的试验总次数。为方便起见，将全部观测值记为

$$\mathbf{Y} = \{Y_{it} : 1 \le i \le N,\ 1 \le t \le T_i\},$$

它表示所有受试者在各测量时点的全部响应。

在给定协变量和个体特异性随机效应的条件下，零膨胀 Beta-二项分布混合回归（ZIBBMR）模型假设

$$Y_{it} \sim \begin{cases} 0, & \text{with probability } 1 - p_{it}, \\ \mathrm{BetaBin}(S_{it}, u_{it}\phi, (1-u_{it})\phi), & \text{with probability } p_{it}, \end{cases} \tag{1}$$

其中 $\phi > 0$ 和 $0 < u_{it}$， $p_{it} < 1$。参数 p_{it} （零膨胀概率）和 u_{it} （平均成功概率）通过以下方式与协变量关联:

$$\mathrm{logit}(p_{it}) = a_i + X_{it}^{\top}\alpha, \quad \mathrm{logit}(u_{it}) = b_i + Z_{it}^{\top}\beta, \tag{2}$$

其中， X_{it}和 Z_{it}为协变量向量， α和 β为回归系数向量，$(a_i, b_i)^{\top}$为独立随机截距，且其服从 $a_i \sim N(a, \sigma_1^2)$ 和 $b_i \sim N(b, \sigma_2^2)$。

Beta-二项分布的概率质量函数为

$$\begin{aligned} P\left(Y = Y_{it} | S_{it}, u_{it}, \phi\right) &= f\left(Y_{it}; S_{it}, u_{it}, \phi\right) \\ &= \binom{S_{it}}{Y_{it}} \frac{B\left(Y_{it} + u_{it}\phi, S_{it} - Y_{it} + (1-u_{it})\phi\right)}{B\left(u_{it}\phi, (1-u_{it})\phi\right)}, \end{aligned} \tag{3}$$

其中， $B(\cdot,\cdot)$表示贝塔函数。

由式 (3) 可得均值和方差为

$$E(Y_{it} \mid S_{it}, u_{it}, \phi) = S_{it}u_{it}, \tag{4}$$

and

$$\mathrm{Var}(Y_{it} \mid S_{it}, u_{it}, \phi) = S_{it}u_{it}(1-u_{it})\left[1 + \frac{S_{it}-1}{\phi+1}\right]. \tag{5}$$

与经典二项方差相比，公式（5）中的方差表达式多出一项，使贝塔-二项分布能够捕捉离散数据中的过度离散现象（Skellam，1948）。

若对 Beta 分布的概率密度函数采用以下参数化方式

$$f_{\text{Beta}}(x;\alpha,\beta)=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)},\quad 0<x<1, \tag{6}$$

那么，在 ZIBBMR 参数化中，$\alpha=u_{it}\phi$ 和 $\beta=(1-u_{it})\phi$。
等价的层次表示为

$$\begin{aligned} Y_{it}\mid S_{it},p_{it},w_{it},\phi &\sim \begin{cases}0, & \text{with probability } 1-p_{it},\\ \text{Bin}(S_{it},w_{it}), & \text{with probability } p_{it},\end{cases}\\ w_{it}\mid u_{it},\phi &\sim \text{Beta}(u_{it}\phi,(1-u_{it})\phi), \end{aligned} \tag{7}$$

其中 w_{it} 是一个未观测到的潜变量。该表述将 Beta-二项分布表示为成功概率服从 Beta 分布的二项分布，并便于采用数据增广方法进行推断。
根据 (1) 和 (2)，$\boldsymbol{\theta}=(\phi,a,b,\alpha,\beta,\sigma_1^2,\sigma_2^2)$ 的观测数据似然为

$$\begin{aligned} L(\theta;\mathbf{Y})=\prod_{i=1}^{N}\int_{\mathbb{R}}\int_{\mathbb{R}}\prod_{t=1}^{T_i}&(1-p_{it})^{\mathbb{1}_{\{Y_{it}=0\}}}\\ &\times\left[p_{it}\,f(Y_{it};S_{it},u_{it},\phi)\right]^{\mathbb{1}_{\{Y_{it}>0\}}}g(a_i,b_i)\,da_i\,db_i. \end{aligned} \tag{8}$$

其中 $g(\cdot,\cdot)$ 是 joi　正态随机效应的密度。该积分没有闭式解，因此需要采用随机逼近　c
近似方法　进行最大似然估计。

2.2 通过随机近似 EM 算法进行估计

随机近似期望最大化（SAEM）算法（Delyon 等，1999）是在潜变量模型中进行最大似然估计的一种迭代方法，适用于 EM 算法的期望步无法显式计算的情况。令 $\mathbf{Y}=(Y_{it},1\le i\le N,1\le t\le T_i)$ 和 $\boldsymbol{\varphi}=(\varphi_i,1\le i\le N)$ 分别表示观测数据和非观测数据，因此该模型的完整数据为 $(\mathbf{Y},\boldsymbol{\varphi})$。在这种情况下，SAEM 算法通过随机逼近过程取代 EM 的常规 E 步。给定初始点 $\boldsymbol{\theta}^{(0)}$，算法的第 q 次迭代写为：

- S模拟（S）步：从条件分布 $p\left(\cdot\mid\mathbf{Y};\boldsymbol{\theta}^{(q-1)}\right.$ 中抽取一个实现值 $\boldsymbol{\varphi}^{(q)}$
- 随机逼近（SA）步骤：更新 $s_q(\boldsymbol{\theta})$，即条件期望$\mathbb{E}\left[\log\,p\left(\mathbf{Y},\boldsymbol{\varphi}^{(q)};\boldsymbol{\theta}\right)\mid\mathbf{Y},\boldsymbol{\theta}^{(q-1)}\right]$的近似值：

$$s_q(\boldsymbol{\theta})=s_{q-1}(\boldsymbol{\theta})+\gamma_q\left(\log p\left(\mathbf{Y},\boldsymbol{\varphi}^{(q)};\boldsymbol{\theta}\right)-s_{q-1}(\boldsymbol{\theta})\right)$$

  其中，$\{\gamma_q\}_{q\in\mathbb{N}}$ 是步长递减序列，且满足 $\gamma_1=1$。
- 最大化（M）步骤：根据 $\boldsymbol{\theta}^{(q)}=\arg\max_{\boldsymbol{\theta}}s_q(\boldsymbol{\theta})$ 更新 $\boldsymbol{\theta}^{(q)}$。

在 ZIBBMR模型中，令 $\varphi_i=(a_i,b_i)$，$1\le i\le N$ 表示未观测到的随机效应，其中 $\boldsymbol{\varphi}=(\varphi_i)_{i=1}^{N}$。根据构造，$\varphi_i\sim N(\boldsymbol{\mu},\mathbf{G})$，其中 $\boldsymbol{\mu}=(a,b)$，且 $\mathbf{G}=\text{diag}(\sigma_1^2,\sigma_2^2)$。
完全数据似然可分解为

$$\begin{aligned} p(\mathbf{Y},\boldsymbol{\varphi};\boldsymbol{\theta})&=p(\mathbf{Y}\mid\boldsymbol{\varphi};\alpha,\beta,\phi)\;p(\boldsymbol{\varphi}\mid\boldsymbol{\mu},\mathbf{G})\\ &\propto|\mathbf{G}|^{-N/2}\prod_{i=1}^{N}\exp\left\{-\frac{1}{2}(\varphi_i-\boldsymbol{\mu})^{\top}\mathbf{G}^{-1}(\varphi_i-\boldsymbol{\mu})\right\}\\ &\quad\times\prod_{i=1}^{N}\prod_{t=1}^{T_i}(1-p_{it})^{\mathbb{1}_{\{Y_{it}=0\}}}\;p_{it}^{\mathbb{1}_{\{Y_{it}>0\}}}\;f(Y_{it};S_{it},u_{it},\phi)^{\mathbb{1}_{\{Y_{it}\ge0\}}}. \end{aligned} \tag{9}$$

SAEM更新方案 给定初始点 $\boldsymbol{\theta}^{(0)}$ 和迭代索引 $q\ge1$，ZIBBMR 的 SAEM算法如下：

1. 模拟步骤：从条件分布 $p(\cdot \mid \mathrm{Y};\, \boldsymbol{\theta}^{(q-1)})$ 中抽取 $\varphi_i^{(q)}, i = 1, \ldots, N$（通过 Metropolis–Hastings 近似；见下文）。
2. 随机逼近步骤：更新充分统计量

$$\begin{aligned} F_1^{(q)} &= F_1^{(q-1)} + \gamma_q \left(\sum_{i=1}^{N} \varphi_i^{(q)} - F_1^{(q-1)} \right), \\ F_2^{(q)} &= F_2^{(q-1)} + \gamma_q \left(\sum_{i=1}^{N} \varphi_i^{(q)} \varphi_i^{(q)\top} - F_2^{(q-1)} \right), \end{aligned} \tag{10}$$

其中，$\{\gamma_q\}$ 是递减的步长序列，且满足 $\gamma_1 = 1$。

3. 最大化步骤：更新

$$\begin{aligned} \boldsymbol{\mu}^{(q)} &= \frac{F_1^{(q)}}{N}, \\ \mathbf{G}^{(q)} &= \frac{F_2^{(q)}}{N} - \frac{F_1^{(q)} F_1^{(q)\top}}{N^2}. \end{aligned} \tag{11}$$

由于存在Beta-二项组件，(β, ϕ)和 α 分别进行更新：

$$\left(\tilde{\beta}^{(q)}, \tilde{\phi}^{(q)} \right) = \arg\max_{\beta,\phi} \sum_{i,t} \mathbb{1}_{\{Y_{it}>0\}} \log \left[\frac{B\left(Y_{it} + u_{it}^{(q)}\phi,\ S_{it} - Y_{it} + (1 - u_{it}^{(q)})\phi \right)}{B\left(u_{it}^{(q)}\phi,\ (1 - u_{it}^{(q)})\phi \right)} \right], \tag{12}$$

$$\tilde{\alpha}^{(q)} = \arg\max_{\alpha} \sum_{i,t} \left[\mathbb{1}_{\{Y_{it}>0\}} \log p_{it}^{(q)} + \mathbb{1}_{\{Y_{it}=0\}} \log(1 - p_{it}^{(q)}) \right], \tag{13}$$

其中，$u_{it}^{(q)} = u_{it}(b_i^{(q)}, \beta)$和 $p_{it}^{(q)} = p_{it}(a_i^{(q)}, \alpha)$通过公式（2）获得。最后，

$$\begin{aligned} \phi^{(q)} &= \phi^{(q-1)} + \gamma_q \left(\tilde{\phi}^{(q)} - \phi^{(q-1)} \right), \\ \alpha^{(q)} &= \alpha^{(q-1)} + \gamma_q \left(\tilde{\alpha}^{(q)} - \alpha^{(q-1)} \right), \\ \beta^{(q)} &= \beta^{(q-1)} + \gamma_q \left(\tilde{\beta}^{(q)} - \beta^{(q-1)} \right). \end{aligned} \tag{14}$$

值得注意的是，SAEM算法的模拟步骤可以采用补充材料A节所述的Metropolis-within-Gibbs方案来实现。然而，这种方法计算量较大，因此并不适用于微生物群数据，因为此类数据集通常规模较大。因此，本文报告的结果采用Metropolis–Hastings算法获得。下一节将介绍SAEM算法模拟步骤的实现。

2.2.1 实现细节

在实践中，SAEM的模拟步骤需要进行仔细调试，以确保对增广参数空间进行高效探索并实现稳定收敛。在使用未经增广的原始设定时，随机效应 $\boldsymbol{\varphi}$ 通过Metropolis–Hastings进行更新，采用三种提议核：(i) 用于全局探索的先验提议 $N(\boldsymbol{\mu}^{(q)}, \mathrm{G}^{(q)})$；(ii) 用于相关局部移动的多元随机游走 $N(\varphi_{i,p-1}, \Omega^{(q)})$；以及(iii) 在 $\varphi_{i,p-1}$ 的单个随机选定分量上进行的单变量随机游走，用于微调。在实践中，每次SAEM迭代中，使用核($kern_1$)运行 m_1 次迭代，使用核($kern_2$)运行 m_2 次迭代，并使用核($kern_3$)运行 m_3 次迭代。这种全局移动与局部移动相结合的方式，使马尔可夫链既能访问后验分布中相距较远的区域，又能在高概率区域细化估计。

通过在每次迭代中模拟多个实现值或独立的马尔可夫链，并采用蒙特卡罗近似，可以提升算法的收敛性。更具体地说，在模拟步骤中，针对 $1 \le l \le m$ 生成 m samples$\varphi_i^{(q,l)} \sim p(\cdot \mid \mathrm{Y}; \boldsymbol{\theta}^{(q-1)})$，而在随机近似步骤中，充分

统计量按照以下方式更新

$$\begin{aligned} F_1^{(q)} &= F_1^{(q-1)} + \gamma_q \left(\sum_{l=1}^{m} \sum_{i=1}^{N} \varphi_i^{(q,l)} - F_1^{(q-1)} \right), \\ F_2^{(q)} &= F_2^{(q-1)} + \gamma_q \left(\sum_{l=1}^{m} \sum_{i=1}^{N} \varphi_i^{(q,l)} \varphi_i^{(q,l)^\top} - F_2^{(q-1)} \right), \end{aligned} \tag{15}$$

使用多条序列或链的优势之一是，它们可用于提供$\mathbb{E}(\varphi_i|Y_i;\ \boldsymbol{\theta})$和 $\mathrm{Var}(\varphi_i|Y_i;\ \boldsymbol{\theta})$ 的经验估计。这些量是通过重要性抽样近似观测对数似然所必需的，进而可以计算似然比检验（LRT），如下一小节所述。

2.2.2 使用重要性抽样近似对数似然

对于复杂的混合效应模型，观测数据的对数似然通常没有闭式表达式。然而，对其进行评估对于实施似然比检验以及计算与给定模型相关的信息准则至关重要。一种可能的近似策略是采用重要性抽样技术（Kloek 和 van Dijk，1978）。

令 $\mathcal{LL}_Y(\boldsymbol{\theta})$ 表示在估计的总体参数处求得的对数似然，即 $\mathcal{LL}_{\boldsymbol{Y}}(\boldsymbol{\theta}) = \log\, p(\boldsymbol{Y};\ \boldsymbol{\theta})$，其中 $p(\boldsymbol{Y};\boldsymbol{\theta}) = L(\boldsymbol{\theta};\ \boldsymbol{Y})$ 表示在给定 $\boldsymbol{\theta}$ 条件下观测数据的联合密度。注意，

$$\mathcal{LL}_{\boldsymbol{Y}}(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} \log p(Y_i; \hat{\boldsymbol{\theta}}),$$

并且，对于给定的提议分布 $\tilde{p}_{\varphi_i}$，其相对于 p_{φ_i} 绝对连续时，Y_i 的边缘密度可写为

$$p(Y_i; \hat{\boldsymbol{\theta}}) = \int p(Y_i, \varphi_i; \hat{\boldsymbol{\theta}})\, d\varphi_i = \int p(Y_i|\varphi_i; \hat{\boldsymbol{\theta}}) \frac{p(\varphi_i; \hat{\boldsymbol{\theta}})}{\tilde{p}(\varphi_i; \hat{\boldsymbol{\theta}})}\, \tilde{p}(\varphi_i; \hat{\boldsymbol{\theta}})\, d\varphi_i.$$

因此，

$$p(Y_i; \hat{\boldsymbol{\theta}}) = \mathbb{E}_{\tilde{p}} \left[p(Y_i|\varphi_i; \hat{\boldsymbol{\theta}}) \frac{p(\varphi_i; \hat{\boldsymbol{\theta}})}{\tilde{p}(\varphi_i; \hat{\boldsymbol{\theta}})} \right].$$

该期望可以通过数值方法近似为：

1. Drawing a sample $\varphi_i^{(1)}, \ldots, \boldsymbol{\varphi}_i^{(K)}$ from the proposal distribution $\tilde{p}_{\varphi_i}$;
2. 计算蒙特卡罗平均

$$\hat{p}_{(i,K)} = \frac{1}{K} \sum_{k=1}^{K} p(Y_i|\varphi_i^{(k)}; \hat{\boldsymbol{\theta}}) \frac{p(\varphi_i^{(k)}; \hat{\boldsymbol{\theta}})}{\tilde{p}(\varphi_i^{(k)}; \hat{\boldsymbol{\theta}})}.$$

提议分布的最优选择应为条件密度 $p_{\varphi_i|Y_i}$，因为它能对该期望产生零方差估计量。然而，由于该条件分布无法以闭式形式获得，我们改用一个与其高度近似的提议分布。该提议分布利用 SAEM算法模拟步骤中获得的随机效应的经验估计条件均值和方差构造，即 $\mu_i = \hat{\mathbb{E}}\left(\varphi_i|Y_i;\ \hat{\boldsymbol{\theta}}\right)$ 和 $\sigma_i^2 = \hat{\mathrm{Var}}\left(\varphi_i|Y_i;\ \hat{\boldsymbol{\theta}}\right)$。随后，根据如下形式的非中心 Student t 分布生成样本 $\varphi_i^{(k)}$、$k = 1$、...、K，

$$\varphi_i^{(k)} = \mu_i + \sigma_i\, T_{i,k},$$

其中 $T_{i,\ k\ \text{独立同分布}\sim}\ t_\nu$。除非另有说明，本文中的对数似然近似是使用 $\nu = 5$ 和$K = 500$计算的。

2.2.3 标准误的随机逼近

除了点估计之外，估计程序通常还需要提供模型参数的标准误。在最大似然估计下，可以渐近地通过费舍尔信息

矩阵获得这些标准误，但对于复杂模型，该矩阵无法通过解析方法求得。根据 Louis缺失信息原理（Louis，1982），以下恒等式成立：

$$\partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{y};\boldsymbol{\theta}) = \mathbb{E}\left(\partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{Y},\boldsymbol{\varphi};\boldsymbol{\theta})|\boldsymbol{Y};\boldsymbol{\theta}\right) + \operatorname{Cov}\left(\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y},\boldsymbol{\varphi};\boldsymbol{\theta})|\boldsymbol{y};\boldsymbol{\theta}\right),$$

其中

$$\begin{aligned}\operatorname{Cov}\left(\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y},\boldsymbol{\varphi};\boldsymbol{\theta})|\boldsymbol{Y};\boldsymbol{\theta}\right) =& \mathbb{E}\left(\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y},\boldsymbol{\varphi};\boldsymbol{\theta}) \partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y},\boldsymbol{\varphi};\boldsymbol{\theta})^\top|\boldsymbol{Y};\boldsymbol{\theta}\right) \\ &- \mathbb{E}\left(\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y},\boldsymbol{\varphi};\boldsymbol{\theta})|\boldsymbol{Y};\boldsymbol{\theta}\right) \mathbb{E}\left(\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y},\boldsymbol{\varphi};\boldsymbol{\theta})|\boldsymbol{Y};\boldsymbol{\theta}\right)^\top\end{aligned}$$

因此，$\partial^2_{\boldsymbol{\theta}} \log\ p(\boldsymbol{y};\boldsymbol{\theta})$ 可通过随机序列 $\{H_q\}_{q\in\mathbb{N}}$ 近似，该序列按如下方式更新：

$$\begin{aligned} D_q &= D_{q-1} + \gamma_q \left(\partial_{\boldsymbol{\theta}} \log p(\boldsymbol{Y},\boldsymbol{\varphi}^{(q)};\boldsymbol{\theta}^{(q)}) - D_{q-1}\right), \\ G_q &= G_{q-1} + \gamma_q \Big(\partial^2_{\boldsymbol{\theta}} \log p(\boldsymbol{y},\boldsymbol{\varphi}^{(q)};\boldsymbol{\theta}^{(q)}) \\ &\qquad + \partial_{\boldsymbol{\theta}} \log p(\boldsymbol{y},\boldsymbol{\varphi}^{(q)};\boldsymbol{\theta}^{(q)}) \partial_{\boldsymbol{\theta}} \log p(\boldsymbol{y},\boldsymbol{\varphi}^{(q)};\boldsymbol{\theta}^{(q)})^\top - G_{q-1}\Big), \end{aligned}$$

其中 $H_q = G_q - D_q D_q^\top$。收敛时，$-H_q^{-1}$ 可近似参数估计值的协方差矩阵（Zhu 和 Lee，2002；Cai，2010），随后可用于构建针对模型参数的假设检验程序。

3 模拟研究

本节评估基于 SAEM 的所提 ZIBBMR 估计量在有限样本下的表现，并将其与两种广泛使用的替代方法 glmmTMB 和 gamlss 进行比较。我们在不同样本量、零膨胀水平和离散度下，评估：（i）点估计的准确性；（ii）Wald 置信区间的覆盖率；（iii）第一类错误控制；以及（iv）计算效率。

3.1 替代估计方法

虽然我们主要关注第 2 节所述的 SAEM 程序，但对于带随机效应的零膨胀贝塔-二项回归，还有其他估计策略可供选择：

- glmmTMB（Brooks 等人，2017年；McGillycuddy 等人，2025年）通过 Template Model Builder (TMB) 实现广义线性混合模型的最大似然估计，并使用拉普拉斯近似对随机效应进行积分。它支持贝塔-二项分布，并可在不同的线性预测量中纳入零膨胀项。即使随机效应结构相当复杂，优化器通常也快速且稳定。
- gamlss（Rigby 和 Stasinopoulos，2005；Stasinopoulos 等人，2017年）通过惩罚似然拟合位置、尺度和形状广义加性模型，允许每个分布参数依赖于协变量。它支持零膨胀贝塔-二项分布族，并可通过惩罚准似然纳入随机效应。该方法灵活，但惩罚可能影响似然比检验的有效性（Breslow 和 Clayton，1993年）。

在模拟中，所有方法均拟合相同的数据集，并采用完全相同的协变量、随机效应结构和分布假设。

3.2 设计

除非另有说明，我们考虑 $N \in \{30, 50, 100\}$ 名受试者，每名受试者有 $T_i = T \in \{5, 10, 15\}$ 次重复测量。对于受试者 i 和时间 t，生成协变量 X_{it} 和 Z_{it}，并且

$$\operatorname{logit}(p_{it}) = a_i + X_{it}^\top \alpha^\star, \qquad \operatorname{logit}(u_{it}) = b_i + Z_{it}^\top \beta^\star,$$

其中 $(a_i, b_i)^\top \sim N(\mu^\star, G^\star)$，$\mu^\star = (a^\star, b^\star)$，且 $G^\star = \operatorname{diag}(\sigma_1^{\star 2}, \sigma_2^{\star 2})$。在给定 (p_{it}, u_{it}) 的条件分布下，

$$Y_{it} \sim \begin{cases} 0, & \text{with probability } 1 - p_{it}, \\ \operatorname{BetaBin}\big(S_{it},\, u_{it}\phi^\star,\, (1-u_{it})\phi^\star\big), & \text{with probability } p_{it}, \end{cases}$$

与 $S_{it} \overset{\text{i.i.d.}}{\sim}$ 均匀分布 $\{200, \ldots, 800\}$。

3.3 数据生成机制

我们考察四种主要情形:

1. 设定1：（同号效应，低方差），参数为 $a = b = -0.5$、 $\alpha = \beta = 0.5$、 $\sigma_1 = 0.7$、 $\sigma_2 = 0.5$、 $\phi = 6.4$。2. 设定2：（异号效应，高方差），参数为 $a = -0.5$、 $b = 0.5$、 $\alpha = 0.5$、 $\beta = -0.5$、 $\sigma_1 = 1.4$、$\sigma_2 = 0.8$、 $\phi = 10.4$。3. 设置3：（纳入协变量，高零值比例），参数为 $a = -1.8$、 $b = -0.9$、 $\alpha =$ (0.8, -0.7)、$\beta =$ (0.6, -0.9)、 $\sigma_1 = 1.35$、 $\sigma_2 = 1.28$、 $\phi = 12.3$；两个协变量 $X_1 = Z_1$ （二元，按受试者半数取0/1）和$X_2 = Z_2 \sim N$(2, 1)。4. 设置 4：（I类错误校准），参数为 $a = 0.5$、 $b = -0.5$、 $\alpha =$ (0, -0.5)、 $\beta =$ (0, 0.5)、 $\sigma_1 = 0.7$、$\sigma_2 = 0.5$、 $\phi \sim$ Unif(2, 10)；两个协变量 $X_1 = Z_1$（二元，按受试者半数取0/1）和 $X_2 = Z_2 \sim N$(1, 1)。

设置 1–2 使用单个二元协变量 $X = Z$ （前 $N/2$ 名受试者取值为 0，其余取值为 1）。设置 3 保留该二元协变量，并加入一个连续协变量，以便在更完整的情境下测试估计效果。我们还使用此设置来检验两种可比较方法（ZIBBMR-SAEM 和 glmmTMB）提出的对数似然计算。设置 4 用于测试 $H_0:\ \alpha_1 = 0$、 $H_0:\ \beta_1 = 0$ 和 $H_0:\ \alpha_1 = \beta_1 = 0$。

3.4 估计量比较

我们评估并比较以下估计方法:

- 提出的方法：ZIBBMR–SAEM，如第 2.2 节所述。
- glmmTMB：使用 glmmTMB 软件包（Brooks 等人，2017年）实现的零膨胀贝塔-二项混合效应模型，该方法依靠拉普拉斯近似对随机效应进行积分。
- gamlss：使用 gamlss 框架实现（Rigby 和 Stasinopoulos，2005；Stasinopoulos 等人，2017），该框架支持通过惩罚最大似然估计拟合零膨胀的贝塔-二项分布。

3.5 性能指标

对于每个参数 θ_j，使用以下指标评估其在 $R = 1000$ 次蒙特卡洛重复模拟中的表现:

- 偏差:

$$\mathrm{Bias} = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_j^{(r)} - \theta_j^{\star}\right)$$

- 均方根误差（RMSE）:

$$\mathrm{RMSE} = \left[\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_j^{(r)} - \theta_j^{\star}\right)^2\right]^{1/2}$$

- 平均绝对误差（MAE）:

$$\mathrm{MAE} = \frac{1}{R}\sum_{r=1}^{R}\left|\hat{\theta}_j^{(r)} - \theta_j^{\star}\right|$$

此外，我们还计算估计量的 Wald检验 p值，以及每次模型拟合的平均计算时间。

3.6 实现细节

所有设计均使用 5 条并行链运行 SAEM，并进行 1000 次迭代；设置 4 除外，该设置使用 10 条链，以确保第一类错误率的估计稳定。
初始参数向量的选择与真实值保持适度接近，同时允许充分探索参数空间:

- *Settings 1–2:* $\theta^{(0)} = (18,\ -0.3,\ 0.2,\ 0.8,\ 0.1,\ 0.48,\ 0.72)$

- *Setting 3–4*: $\theta^{(0)} = (6,\ 0.4,\ -0.7,\ 0.3,\ -0.2,\ 0.2,\ 0.1,\ 0.28,\ 0.61)$

这些对应于 $(\phi, a, b, \alpha, \beta, \sigma_1, \sigma_2)$。设置3和4中的附加项对应于额外的连续协变量。

鉴于 SAEM 算法具有随机性，其收敛性本质上与系数γ_q的选择有关，而这些系数只需满足要求不高的条件（Kuhn和Lavielle，2005）。在我们的实现中，我们将这些系数定义为：若 $q \le K_1$，则为 $\gamma_q = 1$；若 $K_1 < q \le K_1 + K_2$，则为 $\gamma_q = {}_{q-K_1}^{1}$，其中 $K_1 + K_2$ 为迭代总次数。关于 K_1 和 K_2的选择，我们借鉴了saemix包中现有的 SAEM 实现（Comets等人，2017年），将其设为 $K_1 = 750$ 和 $K_2 = 250$。这使算法在最初几次迭代中可以自由波动，并在接近某个（局部）最优点时促使其收敛。

glmmTMB 和 gamlss 使用各自的默认优化设置进行比较；对于大多数设计，两者都能快速收敛，但它们依赖的是确定性优化，而非 SAEM 所采用的随机近似方案，这使得 SAEM 在小样本或高度零膨胀情形下更加稳健。

3.7 结果

下文的图 1–3（以及补充材料附录 B 中的表 1 和表 2）总结了所提出的 ZIBBMR–SAEM 方法（以下简称 SAEM）的有限样本表现，并将其与 glmmTMB 和 gamlss 进行了比较。

零膨胀分量参数 对于零膨胀分量的截距（a）和斜率（α），三种方法（SAEM、glmmTMB 和 gamlss）在各种情景下的表现总体相当。在中到大型 T 设置下，gamlss 通常能取得最低的均方根误差和平均绝对误差（MAE）；而 SAEM 往往偏差最小，尤其是在每位受试者的重复测量次数较少时。

均值分量参数 对于均值分量的截距（b）和斜率（β），glmmTMB 通常偏差和均方根误差最低，尤其是在 T 较大时。然而，在小型 T 设计中，SAEM 在偏差和均方根误差方面与 glmmTMB 不相上下，甚至表现更优，这表明在纵向信息有限的情形下，SAEM 更为稳定。

离散参数和方差分量参数 对于离散参数（ϕ）和方差分量（σ_1^2、σ_2^2），SAEM 在各种设置下通常具有最小的偏差和误差指标。当 ϕ 较大时，glmmTMB 的表现也很有竞争力，但在小样本量或低离散度设计中，gamlss 对 ϕ 的估计存在明显偏差。

对数似然计算 如上所述，我们采用设置3，不仅验证了连续协变量下的估计，也比较了 SAEM 和 glmmTMB 给出的对数似然估计。由于使用了惩罚似然，因此无法将 gamlss 纳入比较。表1显示，在这一模拟场景下，两种方法在这方面的结果相近。不过，值得注意的是，glmmTMB 在多达 10% 的数据集中未能给出对数似然值；而 SAEM 采用稳健的重要性抽样方法，且无需对该值进行数值近似，因此总能给出模型对数似然的估计。

第一类错误评估 在设置 4（表1）下，对于单参数假设的 Wald 检验和联合假设的似然比检验（LRT），SAEM 和 glmmTMB 的经验第一类错误率均接近名义水平。由于 gamlss 采用惩罚似然框架，因此无法进行 LRT。相比之下，gamlss 的 Wald 检验第一类错误率偏高，这可能是惩罚引入偏差所致。

平均计算时间 关于各程序的运行时间，设定1下每种方法均执行1000次，并据此计算平均值；其他设定的表现相似。对于每位个体有$T_i = 5$ (10, 15)个观测值的情况，SAEM 的平均耗时为13.29（23.14、36.01）秒，而 glmmTMB 完成估计平均需要1.05（1.56、2.03）秒，但不一定能达到收敛，这一点可从对数似然结果中看出。值得注意的是，对于所有 T_i取值，gamlss 的表现都很稳定，平均耗时为13.22（13.16、13.79）秒。

总体而言，这些结果表明，在统计功效充足的设计中，所有方法对零膨胀和均值分量的估计效果都足够好；而基于 SAEM 的估计程序始终能为方差和离散参数提供最可靠的估计，并在具有挑战性的情形下带来切实的计算优势。

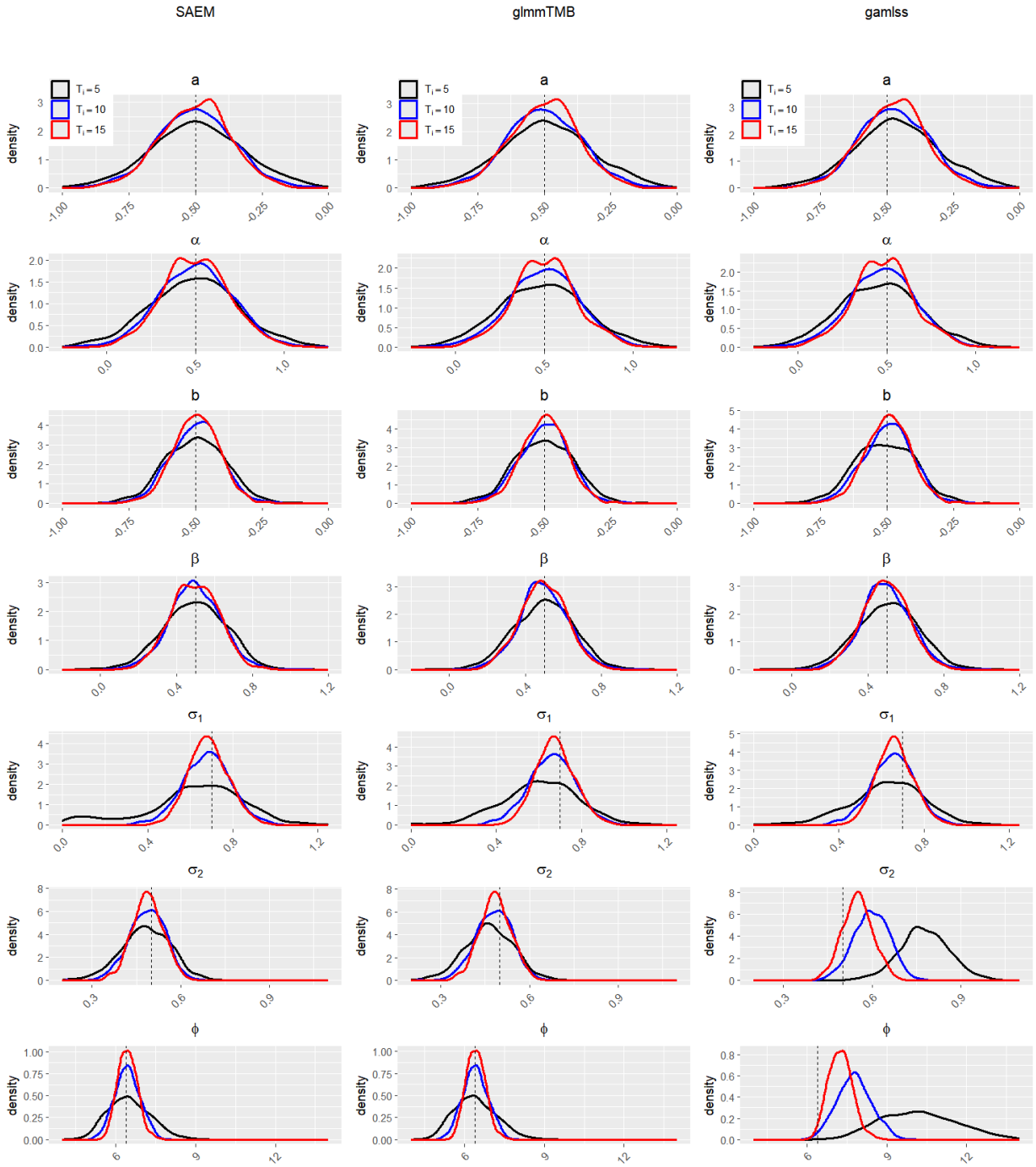


图1：所提出的 SAEM 程序、glmmTMB 和 gamlss 在设定1中的参数估计值经验密度。参数分为：零膨胀分量（a, α）、均值分量（b, β）以及方差和离散度部分的参数（σ_1^2, σ_2^2, ϕ）。垂直点线表示真实参数值。

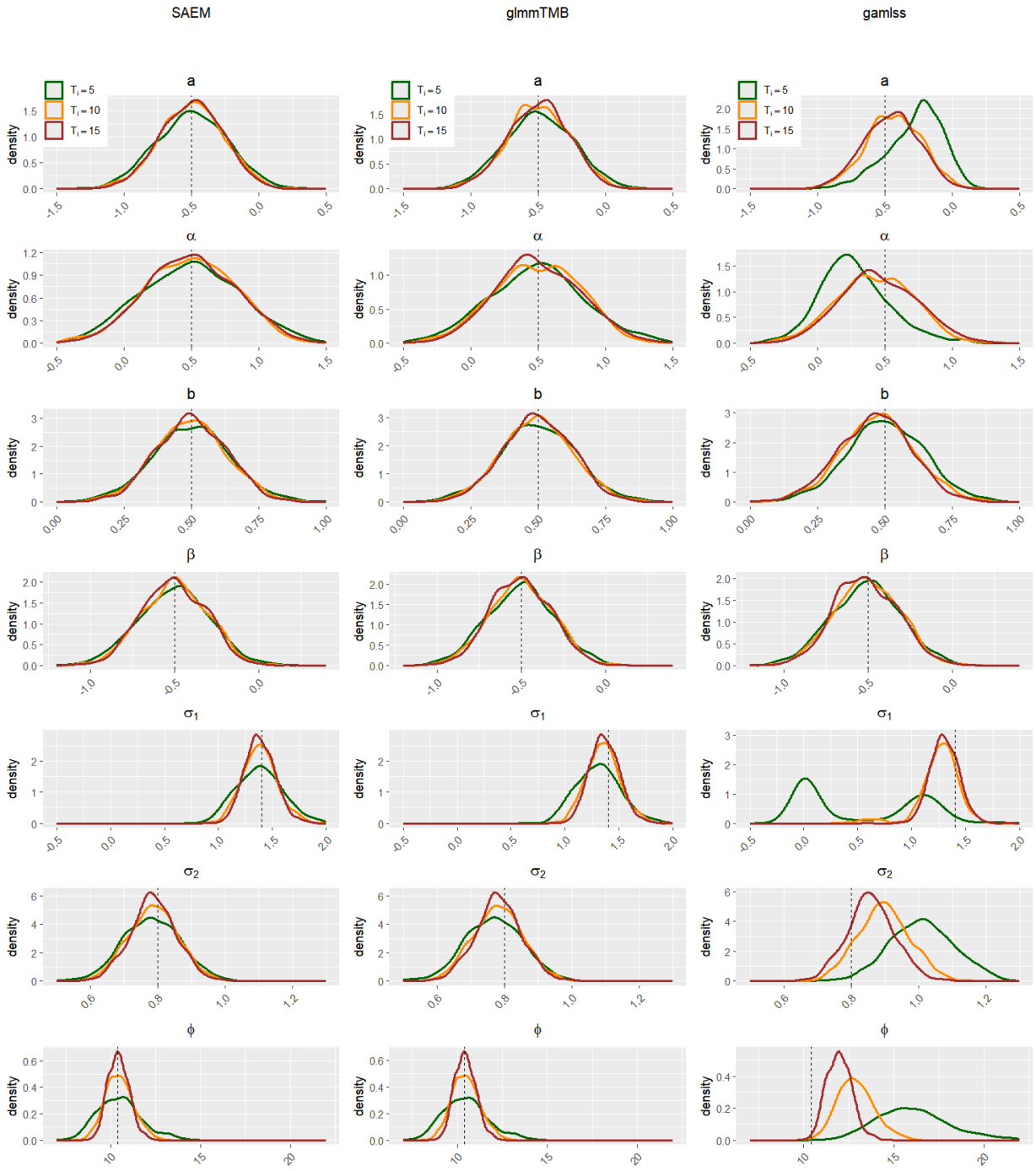


图2：所提出的 SAEM 程序、glmmTMB 和 gamlss 在设定2中的参数估计值的经验密度。参数的分组和解释与图1一致。

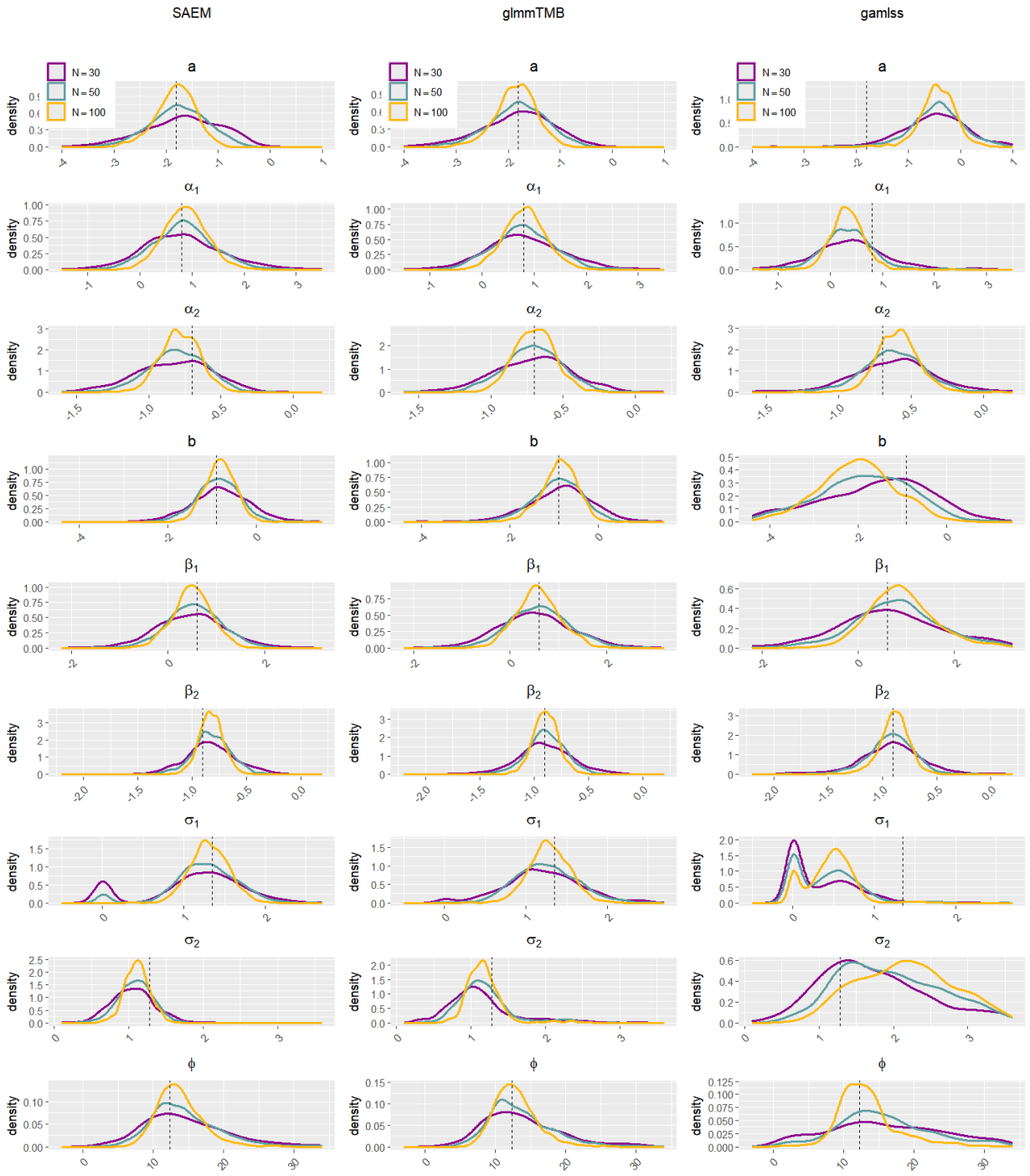


图3：针对所提出的 SAEM 程序、glmmTMB 和 gamlss，在设置3中参数估计值的经验密度。参数的分组和解释遵循图1。

表1：在设置3下，使用 SAEM、glmmTMB 和 gamlss 检验 $H_0:\ \alpha_1 = 0$、 $H_0:\ \beta_1 = 0$ 和 $H_0:\ \alpha_1 = \beta_1 = 0$ 的经验第一类错误率。前两个假设使用 Wald 检验；联合零假设使用似然比检验（LRT）（gamlss 不支持）。名义水平为 0.05 和 0.01。

		SAEM		glmmTMB		gamlss	
		显著性水平		显著性水平		显著性水平	
零假设		0.05	0.01	0.05	0.01	0.05	0.01
$H_0:\alpha_1 = 0$	$N = 30$	0.078	0.025	0.076	0.015	0.284	0.170
	$N = 50$	0.060	0.019	0.061	0.018	0.328	0.195
$H_0:\ \beta_1 = 0$	$N = 30$	0.077	0.022	0.081	0.019	0.168	0.067
	$N = 50$	0.080	0.022	0.085	0.023	0.159	0.058
$H_0:\ \alpha_1 = \beta_1 = 0$	$N = 30$	0.063	0.012	0.069	0.014		
	$N = 50$	0.063	0.017	0.070	0.018		

4 《微生物组》计数数据应用

4.1 背景与动机

人类阴道微生物群是一个复杂的生态系统，在女性生殖健康中发挥着至关重要的作用。妊娠期间其组成和稳定性尤为重要，因为它们会显著影响围产结局，包括早产和感染。16S rRNA基因测序的出现，使细菌多样性和时间动态得以详细表征，克服了培养法的局限性。

在Romero等（2014）的研究中，对22名健康孕妇（均足月分娩）的阴道微生物群进行了纵向比较，并与32名非孕女性进行对照。通过高通量测序和纵向计数数据统计模型，研究发现，妊娠期间的阴道微生物群具有更高的稳定性，且乳杆菌属物种（卷曲乳杆菌、加氏乳杆菌、阴道乳杆菌和詹氏乳杆菌）的丰度更高。相比之下，非孕女性表现出更高的多样性，且与菌群失调相关的分类单元（如加德纳菌属、阿托波菌属和普雷沃菌属）更为常见。

这项研究是最早分析孕妇阴道微生物组纵向变化的研究之一，而此前的大多数研究都聚焦于年轻的孕妇和非孕妇（Srinivasan and Fredricks, 2010；Ravel 等人, 2011），并采用基于培养的方法。这项研究因其方法上的创新性及研究结论而被广泛引用（Florova et al., 2021; Kroon et al., 2018; Zhu et al., 2022），其原始数据至今仍可公开获取（Yi, 2024）。

表 2：Romero 等人的研究中孕妇和非孕妇的描述（2014）.

特征[1]	非孕妇 N = 32	孕妇 N = 22
Age	37 (31–43)	24 (20–29)
观测次数	24 (21–29)	6（6–7）
族裔来源		
黑人	16（50%）	19（86%）
西班牙裔及其他族裔	3 (9.4%)	0 (0%)
白人	13 (41%)	3 (14%)
Nugent评分[2]		
< 7	16（50%）	19（86%）
≥ 7	16（50%）	3（14%）

[1] 年龄：均值（Q1–Q3）；族群和 Nugent评分：n（%）。[2] 至少有一个样本的 Nugent评分 ≥ 7 （细菌性阴道病诊断（Nugent等人，1991年））。

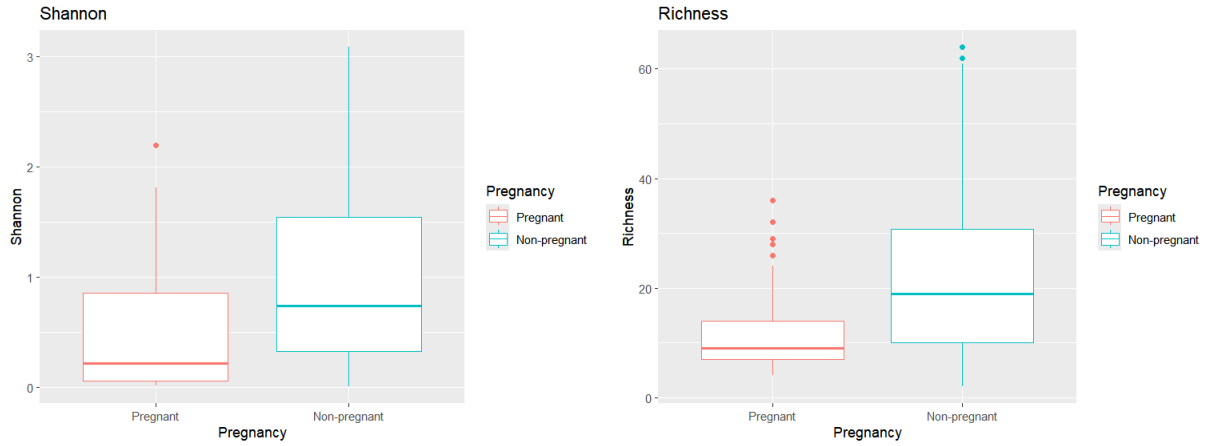
尽管本研究具有诸多优势，但仍存在局限性。如表 2所示，大多数参与者为非裔美国人，这一因素通过遗传和环境影响与微生物群组成相关（Serrano等人，2019年）。样本量较小也限制了研究结果的普适性。此外，原始混合模型仅将妊娠作为协变量，未纳入年龄等其他可能相关的因素。对于零值较多的分类单元，研究使用了零膨胀模型，但零值部分未设置随机效应。

后来，Zhang等人（Zhang等人，2020年）对这些数据应用了几种混合模型设定，包括不含零膨胀的线性模型和负二项模型，以及用于对数转换后的丰度的高斯零膨胀模型。

然而，对于计数数据，不建议进行这种转换（McMurdie 和 Holmes，2014年），而且该分析仅关注丰度效应（通过 Wald 检验），忽略了协变量可能对零膨胀产生的影响。

4.2 《微生物组》模型拟合与比较

我们分析了 Romero 等（2014）的数据集。由于表 2 呈现了该数据集中一些重要的协变量特征，图 4 现在展示了孕妇和非孕妇各自观测值的 α 多样性指数。众所周知，α 多样性是一种生态学测量指标，用于量化单个样本或生物群落中的物种多样性；其可通过偏重多样性较高群落的指数（丰富度、Chao1），以及关注群落均匀度的指数（Shannon、Simpson）来表示（Thukral，2017年）。如图所示，从产次和物种丰富度两方面来看，所有指数均显示非孕妇的多样性更高。这与 Ravel 等（2011）的研究结果一致，该研究指出，无论族裔来源如何，阴道微生物群的稳定性都是孕妇的常见特征。



(a) Shannon

(b) 丰富度

(c) Pielou

(d) Simpson

图 4：孕妇和非孕妇的 α 多样性指数。

我们考虑本研究的以下设定：

- 模型1：妊娠、时间和年龄作为协变量，并将妊娠作为关注因素进行检验。
- 模型2：妊娠、时间、年龄以及时间与妊娠之间的交互作用作为协变量，检验妊娠的效应及其交互作用。

鉴于所考虑模型的特性，我们仅选取零值比例介于 0.1 至 0.9 之间的分类单元，并剔除了只出现在两组女性中其中一组里的细菌。因此，共选取了 57 个分类单元，以上所述模型将应用于这些分类单元。最初，我们考虑使用 glmmTMB 得到的 ZIBBMR 估计值

作为比较（由于 gamlss 无法近似相同的对数似然函数，因此将其舍弃）；然而，该估计值的结果不一致，导致在所考虑的 57 个分类单元中，有 54 个无法计算对数似然。因此，我们将采用另一种用于微生物群数据的模型——零膨胀 Beta 回归（ZIBR）模型（Chen 和 Li，2016年；Barrera 等人，2025年）。该模型针对比例数据，对零膨胀和条件均值进行建模，并使用 Beta 分布描述非零部分。ZIBR 仅作为外部比较模型；它在数据输入（比例与计数）和分布假设（Beta 与 Beta-二项分布）方面均不同于本文提出的 ZIBBMR。

使用显著性水平 0.05 构建相应的似然比检验后，SAEM-ZIBR 在模型1中检测到 29 个分类单元，其中妊娠具有显著性，与 SAEM-ZIBBMR 检测到的数量相同。对于模型2，SAEM-ZIBR 分别检测到 9 个与妊娠相关、15 个与交互作用相关的分类单元，而 SAEM-ZIBBMR 分别检测到 9 个和 11 个。尽管结果看起来相似，但应注意，两种模型检测到的细菌物种并不一定相同，如表3所示。

表3：在所考虑的两种设定中，使用 SAEM 算法针对感兴趣的协变量，仅由 ZIBR 或 ZIBBMR 检测到的细菌分类单元。

仅由 ZIBBMR 检测到			仅由 ZIBR 检测到		
模型1	模型2		模型1	模型2	
妊娠	妊娠	交互作用	妊娠	妊娠	交互作用
毛螺菌科 阴道乳杆菌	毛螺菌科 葡萄球菌属	普雷沃菌基因群3 血斯尼思菌	气球菌属 拟杆菌目	气球菌属 乳酸杆菌	气球菌属 梭菌目 大分支尔德菌 嗜胨菌属 阿托波菌属 变形菌门

结果表明，对于两种方法和模型设定，检测到的大多数细菌都是相同的。因此，分析各个拟议模型特有的细菌具有一定价值。对于 ZIBR，最值得注意的特有细菌是气球菌属，它在三个感兴趣的协变量中均被检测到。Wang等人 Wang等人（2020年）提到，当阴道微生物群中的乳酸杆菌因各种原因（如需氧性阴道炎或妊娠）而增加时，气球菌属会作为补偿性反应而减少。然而，Huang等人 Huang等人（2023年）对此说法提出异议，指出妊娠对气球菌属影响的证据仅限于Romero的数据（本研究对此进行了考察），而其他类似研究并未发现显著影响。

ZIBBMR检测到的最值得注意的分类单元是毛螺菌科。一项研究 Yang等人（2023年）比较了早产女性与正常分娩女性的肠道微生物群，发现正常分娩女性的毛螺菌科丰度更高，二者存在显著相关性。根据Sorbara和Pamer（2020年）的说法，这可能是由于毛螺菌科在人体内能够产生抗炎代谢物。不过，这一作用是由不同于本研究所分析的微生物菌落发挥的，但这一发现仍值得关注。ZIBBMR还检测到普雷沃菌基因群3和血斯尼思菌。

5 结论

我们提出了一种针对纵向微生物组计数数据的零膨胀贝塔-二项混合回归（ZIBBMR）模型，并提出了一种基于 SAEM 算法的估计与推断策略。该模型在纳入零膨胀分量和均值分量中的个体特异性随机效应的同时，兼顾了零膨胀和过度离散。模拟研究表明，该方法能够提供准确的参数估计值、具有竞争力的推断性能，并且在小样本情形下表现稳健；与 glmmTMB 和 gamlss 等替代实现相比，该方法还具有显著优势。

将 ZIBBMR 模型应用于一组纵向人类阴道微生物组数据后，研究发现，在调整胎龄及其他人口统计学协变量后，部分分类单元的丰度和零膨胀模式与妊娠状态相关。与外部基准方法 ZIBR（Barrera等人，2025年）比较后发现，两者识别出的显著分类单元既有重叠，也各不相同，这体现了基于计数数据的 ZIBBMR 和基于比例数据的 ZIBR 各自的优势。鉴于两种模型的数据输入和分布假设不同，在计数数据和相对丰度数据均可用时并行应用这两种模型，可以提高微生物组研究中推断的稳健性。

该方法还被证明比 glmmTMB 更可靠；对于该数据集中的大多数分类单元，glmmTMB 都无法给出有效的似然值或标准误。这些结果凸显了所提出方法在微生物组研究中的适用性，尤其适用于高零膨胀和过度离散与重复测量并存的情形。

ZIBBMR框架的一项显著优势是直接使用原始测序计数，避免了与基于比例的变换相关的统计局限；这类变换容易产生虚假相关（McMurdie 和 Holmes，2014年；Gloor等人，2016年）。通过利用贝塔-二项分布，该模型能够自然地适应微生物组计数数据特有的过度离散和相关结构，并为未来的方法学扩展提供灵活的基础。

若干有前景的方向值得进一步探索。从建模角度看，纳入相关随机斜率或跨分类单元相关结构，可用于捕捉分类单元之间的生态依赖关系。扩展该框架以处理多变量响应或整合系统发育信息，将有助于提升生物学可解释性和泛化能力。在计算方面，开发自适应SAEM方案或混合EM–变分方法，或可提高大规模高通量测序数据集的可扩展性并加快收敛。从应用角度看，评估联合使用ZIBBMR和ZIBR作为互补工具，或能实现更稳健的分类单元检测并获得更丰富的生物学见解。最后，除微生物组研究外，该框架还可应用于其他普遍存在零膨胀纵向计数数据的领域，例如流行病学和生态学。

软件

ZIBBMR模型及其相关SAEM估计程序的R实现（包括用于模拟、模型拟合、推断和诊断的函数）可在 https://github.com/jbarrera232/saem-zibbmr 获取。该代码复现了本文第 4 节中呈现的数据分析结果，并附有示例脚本，展示如何将ZIBBMR和ZIBR模型拟合到微生物组数据集。该软件包依赖用于混合效应建模和数值优化的标准R库，并且设计上易于扩展，以适用于相关的零膨胀混合模型设定。

补充材料

与本文相关的补充材料包括模型增广情形的推导，以及第 3 节的其他表格。

资助

Ana Arribas-Gil谨此感谢MCIN/AEI/10.13039 /501100011033和欧洲区域发展基金（ERDF）“建设欧洲之路”资助的项目 PID2021-123592OB-I00所提供的支持。Dae-Jin Lee的研究得到西班牙国家研究署（AEI）SPHERES项目的支持，项目编号为 PID2023-153222OB-I00。John Barerra和Cristian Meza获得了ANID MATH-AmSud AMSUD 230032-SMILE项目的支持。

参考文献

Barrera, J., Meza, C. 和 Arribas-Gil, A.（2025）。一种用于估计人类微生物组数据零膨胀两部分混合模型的随机方法。arXiv 预印本 arXiv:2504.15411。Breslow, N. E. 和 Clayton, D. G.（1993）。广义线性混合模型中的近似推断。《美国统计学会会刊》，88(421):9–25。Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M. 和 Bolker, B. M.（2017）。glmmTMB 在用于零膨胀广义线性混合建模的软件包中兼顾速度与灵活性。《R期刊》，9(2):378–400。Cai, L.（2010）。利用 Metropolis-Hastings Robbins-Monro算法进行高维探索性项目因子分析。《心理测量学》，75(1):33–57。Chen, E. Z. 和 Li, H.（2016）。用于分析纵向微生物组组成数据的两部分混合效应模型。《生物信息学》，32(17):2611–2617。Comets, E., Lavenu, A. 和 Lavielle, M.（2017）。利用 saemix（一种 SAEM算法的 R 实现）估计非线性混合效应模型中的参数。《统计软件杂志》，80(3):1–41。

Delyon, B., Lavielle, M. 和 Moulines, E.（1999）。EM 算法随机逼近版本的收敛性。《统计学年鉴》，27:94–128。Florova, V., Tikhonova, E., Petrova, L., Zakharevich, N., Kovaleva, A. 和 Gladkikh, M.（2021）。利用分子方法评估正常妊娠及产后时期阴道微生物群的动态变化。《微生物学、流行病学与免疫生物学杂志》，98(5):456–466。Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. 和 Egozcue, J. J.（2016）。微生物组数据集属于组成数据：这并非可有可无。《微生物学前沿》，7:322。Hall, D. B.（2005）。含随机效应的零膨胀泊松回归和二项回归：案例研究。《生物统计学》，56:1030–1039。Huang, B., Fettweis, J. M., Brooks, J. P., Jefferson, K. K. 和 Nelson, K. E.（2023）。妊娠期阴道微生物组变异及其与早产的关联：系统综述。《生殖科学》，30(5):1251–1266。Kloek, T. 和 van Dijk, H. K.（1978）。方程组参数的贝叶斯估计：蒙特卡洛积分的应用。《计量经济学》，46(1):1–19。Kroon, S. J., Ravel, J. 和 Huston, W. M.（2018）。宫颈阴道微生物群、女性健康与生殖结局。《生育与不育》，110(3):327–336。Kuhn, E. 和 Lavielle, M.（2005）。非线性混合效应模型中的最大似然估计。《计算统计学与数据分析》，49(4):1020–1038。Lambert, D.（1992）。零膨胀泊松回归及其在制造缺陷中的应用。《技术计量学》，34:1–14。Louis, T.（1982）。使用 EM 算法时求取观测信息矩阵。《皇家统计学会杂志》。B辑（方法论），44(2):226–233。McCullagh, P. 和 Nelder, J. A.（1989）。《广义线性模型》。Chapman and Hall，伦敦，第2版。McGillycuddy, M., Warton, D. I., Popovic, G. 和 Bolker, B. M.（2025）。在 glmmTMB 中简约拟合大型多变量随机效应。《统计软件杂志》，112(1):1–19。McMurdie, P. J. 和 Holmes, S.（2014）。勿弃可用之物：为何对微生物组数据进行稀疏化并不妥当。《PLOS计算生物学》，10(4):e1003531。Najera-Zuloaga, J., Lee, D.-J. 和 Arostegui, I.（2019）。用于分析纵向离散有界结局的贝塔-二项混合效应模型方法。《生物计量学杂志》，61(3):600–615。Nugent, R. P., Krohn, M. A. 和 Hillier, S. L.（1991）。采用标准化革兰染色判读方法可提高细菌性阴道病诊断的可靠性。《临床微生物学杂志》，29(2):297–301。Pinheiro, J. C. 和 Bates, D. M.（2006）。《S和S-PLUS中的混合效应模型》。Springer，纽约。Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S., McCulle, S. L., Karlebach, S., Gorle, R., Russell, J., Tacket, C. O., Brotman, R. M., Davis, C. C., Ault, K., Peralta, L. 和 Forney, L. J.（2011）。育龄女性的阴道微生物组。《美国国家科学院院刊》，108(增刊1):4680–4687。Rigby, R. A. 和 Stasinopoulos, D. M.（2005）。位置、尺度与形状的广义加性模型。《皇家统计学会杂志》：C辑（应用统计学），54(3):507–554。Romero, R., Hassan, S. S., Gajer, P., Tarca, A. L., Fadrosh, D. W., Bieda, J., Chaemsaithong, P., Miranda, J., Chaiworapongsa, T. 和 Ravel, J.（2014）。正常妊娠女性的阴道微生物群组成和稳定性不同于非妊娠女性。《微生物组》，2:4。Serrano, M. G., Parikh, H. I., Brooks, J. P., Edwards, D. J., Arodz, T. J., Edupuganti, L., Huang, B., Girerd, P. H., Bokhari, Y. A., Bradley, S. P., Fettweis, J. M., Sheth, N. U., Jean, S., Lee, V., Bokhari, A., Forney, L. J., Xu, Z. Z., Ma, B., Zhou, X., Gao, H., Bouchet, V., Paudel, S., Holm, J. B., Sodergren, E., Weinstock, G. M., Wylie, T. N., Mitreva, M., Stout, M. J., Jefferson, K. K. 和 Nelson, K. E.（2019）。妊娠期阴道微生物组动态变化的种族和族裔多样性。《自然医学》，25:1001–1011。Skellam, J. G.（1948）。将成功概率视为不同试验组间的变量，由二项分布导出的概率分布。《皇家统计学会杂志》：B辑，10:257–261。Sorbara, M. T. 和 Pamer, E. G.（2020）。基于微生物组的疗法。《自然·微生物学评论》，18(6):345–355。Srinivasan, S. 和 Fredricks, D. N.（2010）。人类阴道细菌群落与细菌性阴道病。《临床微生物学评论》，21(2):213–233。Stasinopoulos, D. M., Rigby, R. A., Heller, G. Z., Voudouris, V. 和 De Bastiani, F.（2017）。《灵活回归与平滑：在 R 中使用 GAMLSS》。Chapman and Hall/CRC，博卡拉顿。

Thukral, A. K.（2017年）。临床微生物学和传染病领域下一代测序应用的全面综述。《印度医学微生物学杂志》，35(3):339–355。Wang, B., Yao, M., Lv, X., Jiang, S., Li, P., Song, F., Zeng, Z., Chen, Q., Wang, S., Li, S., Jiang, S., Qian, X., and Ma, J.（2020年）。人体微生物群与健康和疾病。《Engineering》，6(3):302–315。Yang, L., Zhang, X., Li, X., Ma, J., Gao, Y., Liu, L., Wang, H., and Zhao, L.（2023）。人体健康与疾病中的毛螺菌科：机制与潜在治疗策略。《微生物学重要评论》，49(2):159–176。Yi, N.（2024）。NBZIMM：负二项和零膨胀混合模型。R软件包版本1.0，提交版本5e454376e0a84f62cbb61d32bd1713b19bb9b71a。Zhang, Y., Han, S., Chen, Y., Chen, L., and Li, F.（2020年）。用于微生物组研究差异丰度分析的零膨胀高斯混合模型。《美国统计学会会刊》，115(531):1021–1035。Zhu, H.-T. and Lee, S.-Y.（2002）。通过结合马尔可夫链蒙特卡罗方法的随机逼近算法分析广义线性混合模型。《统计与计算》，12(2):175–183。Zhu, Q., Pan, Y., and Li, F.（2022）。用于微生物组差异丰度分析的结构零建模。《生物信息学》，38(5):1352–1358。

纵向计数数据零膨胀Beta-二项混合模型随机EM估计与推断的补充材料

John Barrera、Ana Arribas-Gil、Dae-Jin Lee 和 Cristian Meza

附录A SAEM模拟步骤的层次增强

模拟步骤的另一种表述利用式(7)中的层次设定，通过用潜变量增强未观测数据

$$\mathbf{w} = \{w_{it} : 1 \le i \le N,\ 1 \le t \le T_i\},$$

其中 w_{it} 表示潜在二项成功概率。在此增强下，完全数据似然可写为

$$\begin{aligned} p(\mathbf{Y}, \mathbf{w}, \boldsymbol{\varphi}; \theta) &= p(\mathbf{Y} \mid \mathbf{w}, \boldsymbol{\varphi}; \alpha)\; p(\mathbf{w} \mid \boldsymbol{\varphi}; \beta, \phi)\; p(\boldsymbol{\varphi} \mid \boldsymbol{\mu}, \mathbf{G}) \\ &\propto \prod_{i,t} (1 - p_{it})^{\mathbb{1}_{\{Y_{it}=0\}}}\; p_{it}^{\mathbb{1}_{\{Y_{it}>0\}}}\; f_1(Y_{it}; S_{it}, w_{it})^{\mathbb{1}_{\{Y_{it}>0\}}} \\ &\quad \times \prod_{i,t} f_2(w_{it}; u_{it}, \phi)^{\mathbb{1}_{\{Y_{it}>0\}}} \\ &\quad \times |\mathbf{G}|^{-N/2} \prod_{i} \exp\!\left\{ -\frac{1}{2} (\varphi_i - \boldsymbol{\mu})^{\top} \mathbf{G}^{-1} (\varphi_i - \boldsymbol{\mu}) \right\}, \end{aligned} \tag{1}$$

其中 f_1 为二项概率质量函数， f_2 为式(6)中的Beta概率密度函数。

Gibbs型模拟 按照（Meza等人，2007），增强的SAEM模拟步骤如下进行：

1. 使用 Metropolis–Hastings算法对 $\boldsymbol{\varphi}^{(q+1)} \sim p(\cdot \mid \mathbf{Y}, \mathbf{w}^{(q)}; \theta^{(q)})$ 进行采样。
2. 抽样 其条件分布为闭式 Beta 条件分布 on:

$$w_{it} \mid \mathbf{Y}, \boldsymbol{\varphi} \;\sim\; \mathrm{Beta}(Y_{it} + u_{it}\phi,\; S_{it} - Y_{it} + (1 - u_{it})\phi)\,.$$

Metropolis–Hastings 核 对 $\boldsymbol{\varphi}$ 进行采样时，使用三个提议核，并在每次 SAEM 迭代中分别循环 m_1、 m_2 和 m_3 次。这种策略兼顾全局探索与局部细化：

1. 先验提议（$kern_1$）： $N(\boldsymbol{\mu}^{(q)}, \mathbf{G}^{(q)})$。
2. 多元随机游走（$kern_2$）： $N(\varphi_{i,\ p-1}, \Omega^{(q)})$，其中 $\Omega^{(q)}$ 经过调优，以获得最佳接受 e.
3. 单变量随机游走（$kern_3$）：将 $\varphi_{i,\ p-1}$ 的一个随机选定分量扰动 $N(0, 1)$ 噪声。

接受率 对于未通过 w 进行增广的原始设定，Metropolis–Hastings 接受率为：

$$\Delta_{i,p}^{(1)} = \sum_{t=1}^{T_i} \mathbb{1}_{\{Y_{it}>0\}} \left[\log \frac{B\big(Y_{it} + \hat{u}_{it}\phi^{(q)},\ S_{it} - Y_{it} + (1-\hat{u}_{it})\phi^{(q)}\big)}{B\big(Y_{it} + u_{it}\phi^{(q)},\ S_{it} - Y_{it} + (1-u_{it})\phi^{(q)}\big)} - \log \frac{B\big(\hat{u}_{it}\phi^{(q)},\ (1-\hat{u}_{it})\phi^{(q)}\big)}{B\big(u_{it}\phi^{(q)},\ (1-u_{it})\phi^{(q)}\big)} \right] + \mathbb{1}_{\{Y_{it}>0\}} \log \frac{\hat{p}_{it}}{p_{it}} + \mathbb{1}_{\{Y_{it}=0\}} \log \frac{1-\hat{p}_{it}}{1-p_{it}}.$$

其中 $\Delta_i^{(2)}{}_{,p}$ 适用于通过加入高斯先验校正得到的对称核。对于带有 w 的增广 Gibbs 设定：

$$\Delta_{i,p}^{(1)} = \sum_{t=1}^{T_i} \mathbb{1}_{\{Y_{it}>0\}} \left[\log \frac{\Gamma(u_{it}\phi^{(q)})\Gamma((1-u_{it})\phi^{(q)})\hat{p}_{it}}{\Gamma(\hat{u}_{it}\phi^{(q)})\Gamma((1-\hat{u}_{it})\phi^{(q)})p_{it}} + \phi^{(q)}(\hat{u}_{it} - u_{it}) \log \frac{w_{it}^{(q)}}{1-w_{it}^{(q)}} \right] + \mathbb{1}_{\{Y_{it}=0\}} \log \frac{1-\hat{p}_{it}}{1-p_{it}},$$

其中 $\Delta_i^{(2)}{}_{,p}$ 与原始情形类似。

附录 B 模拟研究：补充表格

作为对第 3.7 节所呈结果的补充，表 1–2 总结了所提出的 ZIBBMR–SAEM方法的有限样本表现，并与 glmmTMB 和 gamlss 进行比较，同时采用所提出的 ZIBBMR–SAEM方法。

表1：SAEM算法以及 glmmTMB 和 gamlss 软件包在 1000 次模拟运行中所得结果的汇总统计量。对于每个参数值和每个个体的观测数， T_i，粗体数字表示偏差、均方根误差和平均绝对误差各自的最小绝对值。

参数	值		**Bias**	**RMSE**	**MAE**	**Bias**	**RMSE**	**MAE**	**Bias**	**RMSE**	**MAE**
			SAEM			glmmTMB			gamlss		
a	-0.5	$T_i = 5$	**0.0076**	0.1743	0.1381	0.0093	0.1700	0.1351	0.0394	**0.1644**	**0.1301**
		$T_i = 10$	**-0.0001**	0.1419	0.1129	0.0011	0.1389	0.1109	0.0272	**0.1340**	**0.1066**
		$T_i = 15$	-0.0009	0.1296	0.1039	**0.0002**	0.1263	0.1009	0.0223	**0.1224**	**0.0984**
α	0.5	$T_i = 5$	**-0.0093**	0.2474	0.1968	-0.0104	0.2389	0.1921	-0.0399	**0.2273**	**0.1825**
		$T_i = 10$	0.0024	0.2031	0.1628	**0.0013**	0.1953	0.1560	-0.0251	**0.1863**	**0.1483**
		$T_i = 15$	0.0049	0.1837	0.1480	**0.0043**	0.1752	0.1395	-0.0181	**0.1682**	**0.1337**
b	-0.5	$T_i = 5$	0.0005	0.1136	0.0910	**-0.0003**	**0.1103**	**0.0888**	-0.0090	0.1176	0.0932
		$T_i = 10$	0.0012	0.0955	0.0758	**0.0006**	0.0939	0.0743	0.0033	**0.0939**	**0.0743**
		$T_i = 15$	0.0026	0.0857	0.0681	**0.0015**	0.0844	0.0669	0.0053	**0.0837**	**0.0664**
β	0.5	$T_i = 5$	-0.0027	0.1622	0.1303	**-0.0019**	**0.1550**	**0.1238**	0.0053	0.1607	0.1285
		$T_i = 10$	-0.0012	0.1336	0.1054	**-0.0009**	**0.1281**	**0.1011**	-0.0035	0.1287	0.1018
		$T_i = 15$	-0.0044	0.1233	0.1002	**-0.0036**	0.1191	0.0958	-0.0072	**0.1188**	**0.0955**
σ_1	0.7	$T_i = 5$	-0.0771	0.2466	0.1828	-0.0931	0.2046	0.1589	**-0.0718**	**0.1895**	**0.1454**
		$T_i = 10$	**-0.0210**	0.1135	0.0896	-0.0377	0.1147	0.0909	-0.0457	**0.1106**	**0.0877**
		$T_i = 15$	**-0.0135**	0.0928	**0.0743**	-0.0235	0.0933	0.0749	-0.0376	**0.0925**	0.0750
σ_2	0.5	$T_i = 5$	**-0.0199**	**0.0925**	**0.0706**	-0.0375	0.0933	0.0730	0.2804	0.2967	0.2804
		$T_i = 10$	**-0.0136**	**0.0632**	**0.0500**	-0.0207	0.0647	0.0511	0.0987	0.1156	0.1016
		$T_i = 15$	**-0.0113**	**0.0532**	**0.0426**	-0.0150	0.0545	0.0436	0.0518	0.0728	0.0598
ϕ	6.4	$T_i = 5$	**0.1208**	0.8169	0.6489	0.0258	**0.7992**	**0.6377**	3.8619	4.1377	3.8619
		$T_i = 10$	0.0431	0.4741	0.3795	**0.0157**	**0.4731**	**0.3791**	1.3843	1.5273	1.3877
		$T_i = 15$	0.0456	0.3700	0.2961	**0.0309**	**0.3689**	**0.2957**	0.8584	0.9680	0.8618
			SAEM			glmmTMB			gamlss		
a	-0.5	$T_i = 5$	**0.0056**	0.2603	**0.2094**	0.0084	**0.2548**	0.2039	0.2250	0.3043	0.2653
		$T_i = 10$	0.0043	0.2331	0.1867	**0.0034**	0.2255	0.1819	0.0678	**0.2137**	**0.1720**
		$T_i = 15$	**0.0029**	0.2252	0.1820	0.0031	0.2183	0.1766	0.0476	**0.2054**	**0.1671**
α	0.5	$T_i = 5$	**-0.0057**	0.3674	0.2953	-0.0073	0.3543	**0.2802**	-0.2220	**0.3408**	0.2889
		$T_i = 10$	-0.0100	0.3370	0.2706	**-0.0063**	0.3175	0.2592	-0.0719	**0.2910**	**0.2354**
		$T_i = 15$	**0.0001**	0.3299	0.2653	0.0010	0.3105	0.2503	-0.0433	**0.2865**	**0.2329**
b	0.5	$T_i = 5$	0.0015	0.1466	0.1157	**0.0010**	**0.1425**	**0.1127**	0.0059	0.1457	0.1150
		$T_i = 10$	**-0.0029**	0.1358	0.1075	-0.0031	**0.1330**	**0.1050**	-0.0235	0.1403	0.1106
		$T_i = 15$	0.0008	0.1290	0.1028	**-0.0007**	**0.1254**	**0.1005**	-0.0371	0.1402	0.1106
β	-0.5	$T_i = 5$	**-0.0017**	0.2084	0.1667	-0.0024	**0.1976**	**0.1583**	-0.0099	0.2038	0.1631
		$T_i = 10$	0.0020	0.1925	0.1539	**0.0013**	**0.1847**	**0.1478**	-0.0051	0.1910	0.1537
		$T_i = 15$	-0.0050	0.1851	0.1492	**-0.0031**	**0.1754**	**0.1416**	-0.0127	0.1863	0.1505
σ_1	1.4	$T_i = 5$	**-0.0137**	**0.2173**	**0.1737**	-0.0895	0.2238	0.1807	-0.8534	1.0247	0.8755
		$T_i = 10$	**-0.0158**	**0.1601**	**0.1269**	-0.0467	0.1617	0.1289	-0.1458	0.2412	0.1739
		$T_i = 15$	**-0.0144**	**0.1425**	**0.1135**	-0.0333	0.1429	0.1147	-0.0967	0.1637	0.1322
σ_2	0.8	$T_i = 5$	**-0.0182**	**0.0884**	**0.0704**	-0.0310	0.0919	0.0732	0.2122	0.2339	0.2133
		$T_i = 10$	**-0.0122**	**0.0741**	**0.0594**	-0.0185	0.0753	0.0605	0.0980	0.1245	0.1046
		$T_i = 15$	**-0.0122**	**0.0688**	**0.0544**	-0.0163	0.0699	0.0553	0.0599	0.0926	0.0751
ϕ	10.4	$T_i = 5$	0.1150	**1.2357**	**0.9758**	**0.0891**	1.2472	0.9846	5.7136	6.0776	5.7136
		$T_i = 10$	0.0933	**0.7848**	**0.6240**	**0.0917**	0.7878	0.6267	2.4983	2.6995	2.4997
		$T_i = 15$	0.0537	**0.5907**	**0.4705**	**0.0516**	0.5909	0.4708	1.5690	1.7177	1.5714

表 2：SAEM算法以及软件包 glmmTMB 和 gamlss 在 1000 次模拟运行中所得结果的汇总统计。对于每个参数值和个体数量， N，粗体数字表示偏差、均方根误差和平均绝对误差中各自绝对值最低的结果。

参数	值		**Bias**	**RMSE**	**MAE**	**Bias**	**RMSE**	**MAE**	**Bias**	**RMSE**	**MAE**
				SAEM			glmmTMB			gamlss	
a	-1.80	$N = 30$	0.1160	0.7644	0.6135	**-0.0405**	**0.7423**	**0.5641**	2.8743	15.8214	3.9015
		$N = 50$	0.0459	0.5463	0.4365	**-0.0087**	**0.5450**	**0.4303**	2.3195	16.8357	2.3344
		$N = 100$	**-0.0133**	**0.3729**	**0.2930**	-0.0469	0.3820	0.2998	1.5483	2.9205	1.5484
α_1	0.80	$N = 30$	**0.0121**	0.7619	0.5952	0.0169	**0.7410**	**0.5777**	3.2631	71.7294	7.6583
		$N = 50$	0.0442	0.5778	0.4520	**0.0024**	**0.5679**	**0.4481**	-1.5974	30.2205	2.1587
		$N = 100$	0.0763	0.4131	0.3277	**0.0235**	**0.3931**	**0.3115**	-0.6225	2.6617	0.6918
α_2	-0.70	$N = 30$	-0.1262	0.2910	0.2285	**-0.0110**	**0.2760**	**0.2145**	0.2487	5.5525	0.6081
		$N = 50$	-0.1042	0.2259	0.1768	**-0.0112**	**0.2034**	**0.1586**	2.3606	71.9637	2.4620
		$N = 100$	-0.0822	0.1564	0.1250	**0.0024**	**0.1395**	**0.1112**	0.1088	0.1835	0.1417
b	-0.90	$N = 30$	0.1056	**0.6766**	**0.5229**	**0.0517**	0.7610	0.5797	-0.9298	1.9406	1.3621
		$N = 50$	0.0453	**0.4834**	**0.3813**	**-0.0145**	0.5748	0.4438	-1.1801	1.8631	1.3633
		$N = 100$	0.0870	**0.3464**	**0.2765**	**0.0257**	0.4004	0.3144	-1.2317	1.6162	1.3007
β_1	0.60	$N = 30$	-0.1100	**0.7413**	**0.5794**	**-0.0611**	0.8246	0.6292	0.3391	2.7814	1.1914
		$N = 50$	-0.0461	**0.5612**	**0.4432**	**0.0169**	0.6545	0.5079	0.3506	1.3624	0.9027
		$N = 100$	-0.0926	**0.4106**	**0.3235**	**-0.0118**	0.4563	0.3577	0.3600	0.8875	0.6322
β_2	-0.90	$N = 30$	0.0796	**0.2504**	**0.1914**	**0.0059**	0.2613	0.1995	-0.0382	0.5451	0.2462
		$N = 50$	0.0803	0.1880	0.1478	**0.0078**	**0.1797**	**0.1386**	-0.0135	0.2269	0.1647
		$N = 100$	0.0761	0.1324	0.1062	**0.0074**	**0.1139**	**0.0902**	0.0093	0.1384	0.1054
σ_1	1.35	$N = 30$	-0.2125	0.6456	0.4726	**-0.1283**	**0.5296**	**0.3991**	3.2119	46.7426	5.1846
		$N = 50$	-0.1032	0.4363	0.3201	**-0.1000**	**0.3878**	**0.3113**	-0.3045	18.5523	1.5728
		$N = 100$	**-0.0361**	**0.2359**	**0.1883**	-0.0626	0.2557	0.2024	-0.8943	0.9369	0.8992
σ_2	1.28	$N = 30$	-0.2987	**0.5016**	**0.3763**	**-0.1585**	0.5622	0.4284	1.1680	3.8981	1.3919
		$N = 50$	-0.2019	**0.3310**	**0.2547**	**-0.0951**	0.3928	0.2939	1.5041	3.4468	1.5802
		$N = 100$	-0.1671	**0.2373**	**0.1967**	**-0.1009**	0.2804	0.2083	1.6499	3.5445	1.6793
ϕ	12.30	$N = 30$	4.4611	14.3424	6.8313	**3.3669**	**12.6306**	**5.9997**	7.2160	15.4420	10.2447
		$N = 50$	2.4394	5.9903	4.0989	**1.5291**	**5.3352**	**3.6469**	4.5668	9.4448	6.6060
		$N = 100$	1.4718	3.6310	2.6230	**0.5794**	**3.1230**	**2.2986**	1.3951	5.2315	3.3507

参考文献

Meza, C., Jaffr´ezic, F. 和 Foulley, J.-L.（2007）。使用 SAEM 算法对非线性混合效应模型中的方差参数进行REML估计。《生物计量学杂志：生物科学数学方法》， 49(6):876–888。