arXiv:2304.02899v1 [stat.ML] 6 Apr 2023

用于贝叶斯变量选择的可变复杂度加权温度化吉布斯采样器

Lan V. Truong

英国剑桥郡剑桥特朗平顿剑桥大学工程系，CB21PZ。

贡献作者： lt407@cam.ac.uk;

摘要

Jankowiak 最近引入了子集加权退火吉布斯采样器（wTGS），用于降低高维应用中每次MCMC 迭代的计算复杂度，此类应用不要求精确计算后验纳入概率（PIP）。然而，当信号维度与条件纳入概率估计次数之比较大时，与该采样器相关联的Rao-Blackwell化估计量具有较高的方差。在本文中，我们设计了一种新的子集加权退火吉布斯采样器（wTGS），其每次 MCMC 迭代中条件纳入概率的预期计算次数可以远小于信号维度。不同于子集wTGS和 wTGS，我们的采样器每次MCMC迭代的复杂度是可变的。对于该采样器在有限迭代次数 T下的相关Rao-Blackwell化估计量，我们给出了其方差的上界，并证明对于给定数据集，其方差为 $O\left(\left(\frac{p}{s}\right)^2 \log T / T\right)$，其中 S 是每次 MCMC 迭代中条件纳入概率计算次数的期望值。实验表明，我们的 Rao-Blackwell 化估计量的方差可能小于与子集 wTGS 相关联的对应估计量。

关键词：贝叶斯可变选择、退火吉布斯采样器、计算复杂度、线性回归、线性模型。

1 引言

马尔可夫链蒙特卡罗（MCMC）方法是一类用于从已知函数中抽样的算法。MCMC 方法主要用于计算多维积分的数值近似值，例如在贝叶斯方法中。

统计学、计算物理学 [12]，计算生物学， [9]，以及线性模型 [18–20]。过去十年来，蒙特卡罗算法一直广受欢迎[10, 17]。统计信号处理、机器学习和统计学中的许多实际问题，都需要快速且准确的方法，从具有任意非标准形式的概率分布中抽取样本 [1, 8, 16]。最受欢迎的蒙特卡罗方法之一是马尔可夫链蒙特卡罗（MCMC）算法系列 [1, 17] 和粒子滤波器 [5]。MCMC 技术生成一个马尔可夫链，其不变密度为预先设定的目标概率密度函数[13]。

吉布斯采样器（GS）是一种马尔可夫链蒙特卡罗（MCMC）算法，用于获得一系列从特定多元概率分布中近似抽取的观测值。这一序列可用于近似联合分布、某个可变的边缘分布，或某些可变的子集。它还可用于计算某个可变的期望值（积分）[2, 3]。当联合分布未知或难以直接采样，但每个可变的条件分布已知且易于（或至少相对容易）采样时，GS 便适用。GS 算法依次从每个可变的分布中生成一个样本，并以其他可变的当前值为条件。可以证明，这一采样序列构成一条马尔可夫链，而该马尔可夫链的平稳分布正是所求的联合分布。

GS 通常用作统计推断的手段，尤其是贝叶斯推断。然而，纯粹基于马尔可夫链的方案（即直接从精确的目标分布中进行模拟、无需后续重要性采样校正的方案）取得了更大的成功。这是因为 MCMC 方法通常更能扩展到高维情形，而重要性权重的方差往往会随维度增加（通常呈指数增长）。 [23] 提出了一种自然的方法，将 MCMC 与重要性采样的优势结合起来，使其在高维情形下具有稳健性，并缓解许多基于马尔可夫链的方案所面临的混合缓慢问题。该方案称为退火吉布斯采样器（ TGS），它采用了类似吉布斯采样（GS）的逐分量更新规则，同时具有更好的混合性质，且相关的重要性权重在维度增加时仍保持稳定。借助精心设计的退火机制，TGS 克服了标准 GS 的主要局限，例如强烈的后验相关性导致的混合缓慢。它还不要求按顺序遍历所有坐标，而是根据状态信息迭代地决定接下来应更新哪个坐标。

TGS 已应用于贝叶斯变量选择（BVS）问题，与其他蒙特卡罗方案相比，性能提升了多个数量级 [23]。由于 TGS 以相同频率更新每个坐标，在 BVS 场景中，这可能效率不高，因为采样器的大多数迭代都将用于更新后验纳入概率较低或可忽略的可变，尤其是在信号维度 P 较大时。一种更好的方案称为加权退火吉布斯采样（wTGS） [23]，，它会更频繁地更新纳入概率较大的分量，从而更集中地分配计算资源。然而，尽管

尽管这种方法对贝叶斯变量选择问题具有直观吸引力，但近似所得的后验分布在计算上可能颇具挑战性。其主要原因是，只要协变量超过几十个，模型空间的规模就会大得惊人。为了扩展到高维情形， [11] 最近引入了一种高效的MCMC 方案，与 wTGS 相比，该方案每次迭代的成本可以显著降低。其核心思想是引入一个辅助可变，用于控制在给定 MCMC 迭代中计算哪些条件后验纳入概率（后验纳入概率）。然而，该方案也存在一些缺点，例如，与该采样器相关的 Rao-Blackwell 化估计量具有很高的方差，因为 P/S在少量 MCMC 迭代中就会变得很大， T此外，在子集 wTGS 算法中生成一个在$\binom{P}{S}$ 子集上均匀分布的辅助随机集合 [11] 需要非常长的运行时间。

本文设计了一种新的子集 wTGS，称为可变复杂度wTGS（VC-wTGS）。我们证明，与VC-wTGS 相关联的 Rao-Blackwell 化估计量的方差为 $O\left(\log_T{}^T \left(\frac{P}{S}\right)^2\right)$，其中 P、 S和 T 分别表示信号维度、每次 MCMC 迭代的 PIP 计算次数以及 MCMC迭代次数。实验表明，对于模拟数据集，我们的方案能够非常快速地收敛到后验纳入概率；对于 MNIST数据集，当 P/S 非常高时，Rao-Blackwell 化估计量的方差可能远小于子集 wTGS [11] 。更具体地说，我们的贡献包括：

- 我们提出了一种新的子集 wTGS，称为 VC-wTGS，其中每次MCMC 所需进行的条件 PIP 计算的期望次数可能远小于信号维度。
- 我们分析了在每个有限数量的 MCMC 迭代下相关 Rao-Blackwell 化估计量的方差。我们证明，对于给定的数据集，该方差为 $O\left(\log_T{}^T \left(\frac{P}{S}\right)^2\right)$。
- 我们在模拟数据集（多元高斯数据集）和真实数据集（MNIST）上进行了一些实验。实验表明，在相同数量的 MCMC 迭代 T下，我们的估计量在较高的 P/S 处可以比基于子集 wTGS 的估计量具有更小的方差 [11] 。

2 预备知识

2.1 数学背景

设一个定义在状态空间 $\mathcal{S}$ 上的马尔可夫链 $\{X_n\}_{n=1}^{\infty}$ ，其转移核为 $Q(x, dy)$，初始状态为 $X_1 \sim \nu$，其中 $\mathcal{S}$ 是 $\mathbb{R}$中的波兰空间。本文考虑不可约且正常返的马尔可夫链，因此可以保证平稳分布 π 的存在。在无限状态空间上的不可约递归马尔可夫链称为 Harris链 [21]。如果满足以下详细平衡条件，则称马尔可夫链是可逆的：

$$\pi(dx)Q(x, dy) = \pi(dy)Q(y, dx), \qquad \forall x, y \in \mathcal{S}. \tag{1}$$

定义

$$d(t) = \sup_{x \in \mathcal{S}} d_{\mathrm{TV}}(Q^t(x, \cdot), \pi), \qquad t_{\mathrm{mix}}(\varepsilon) := \min\{t : d(t) \le \varepsilon\}, \tag{2}$$

and

$$\tau_{\min} := \inf_{0 \le \varepsilon \le 1} t_{\mathrm{mix}}(\varepsilon) \left(\frac{2-\varepsilon}{1-\varepsilon}\right)^2, \qquad t_{\mathrm{mix}} := t_{\mathrm{mix}}(1/4). \tag{3}$$

令 $L_2(\pi)$ 表示定义在 $\mathcal{S}$ 上、关于 π 平方可积的复值可测函数所构成的希尔伯特空间。我们在 $L_2(\pi)$ 上定义内积 $\langle f, g\rangle := \int f g^* d\pi$，以及范数 $\|f\|_{2,\,\pi} := \langle f$。令 E_π 为相应的平均算子，定义为 $(E_\pi)(x, y) = \pi(y)\ \forall x, y \in \mathcal{S}$，并且

$$\lambda = \|Q - E_\pi\|_{L_2(\pi) \to L_2(\pi)}, \tag{4}$$

其中 $\|B\|_{L_2(\pi)\to L_2(\pi)} = \max_{v:\|v\|_{2,\,\pi}=1} \|Bv\|_{2,\,\pi}$. Q 可视为作用于 $L_2(\pi)$ 上的线性算子（无穷小生成元），记作 Q，定义为 $(\mathsf{Q}f)(x) := \mathbb{E}_{Q(x,\cdot)}(f)$，且可逆性等价于 Q 的自伴性。算子 Q 从左侧作用于测度，生成测度 $\mu\mathsf{Q}$；也就是说，对于 $\mathcal{S}$ 的任意可测子集 A，$\mu\mathsf{Q}(A) := \int_{x \in \mathcal{S}} Q(x, A)\mu(dx)$。对于具有平稳分布$\pi$的马尔可夫链，我们将该链的谱定义为

$$S_2 := \big\{\xi \in \mathbb{C} : (\xi \mathbf{I} - \mathbf{Q}) \text{ is not invertible on } L_2(\pi)\big\}. \tag{5}$$

众所周知，$\lambda = 1 - \gamma^*$ [14]其中

$$\gamma^* := \begin{cases} 1 - \sup\{|\xi| : \xi \in \mathcal{S}_2, \xi \ne 1\}, \\ \qquad\qquad \text{if eigenvalue 1 has multiplicity 1}, \\ 0, \qquad\quad \text{otherwise} \end{cases}$$

是该马尔可夫链的绝对谱隙。绝对谱隙可以通过马尔可夫链的混合时间 t_{mix} 由以下表达式进行界定：

$$\left(\frac{1}{\gamma^*} - 1\right) \log 2 \le t_{\mathrm{mix}} \le \frac{\log(4/\pi_*)}{\gamma_*}, \tag{6}$$

其中，$\pi_* = \min_{x \in \mathcal{S}} \pi_x$ 是最小平稳概率；如果对某个 $k \ge 1$有$Q^k > 0$ （逐项为正），则该概率为正。有关更详细的讨论，请参见[22] 。在 [6, 22], 中，作者给出了从单条轨迹估计 t_{mix} 和 γ^* 的算法。

定义

$$\mathcal{M}_2 := \left\{\nu \in \mathcal{M}(\mathcal{S}) : \left\|\frac{d\nu}{d\pi}\right\|_2 < \infty\right\}, \tag{7}$$

其中， $\|\cdot\|_2$ 是 $\mathcal{S}$上复值可测函数的希尔伯特空间中的标准 L_2 范数。

2.2 问题设定

考虑带有 $X \in \mathbb{R}^{N\times P}$ 和 $Y \in \mathbb{R}^N$ 的线性回归，并定义以下模型空间：

- 纳入变量：γ_i ∼Bern(h)
- $\sigma_\gamma^2 \in$ $\left(\frac{1}{2}\nu_0\ \ \frac{1}{2}\nu_0\lambda_0\right)$ 噪声
- 方差：InvGamma，系数： $\beta_\gamma \sim \mathcal{N}(0, \sigma_\gamma^2\tau^{-1}\mathbb{I}_{|\gamma|})$
- 响应变量： $Y_n \sim \mathcal{N}(\beta_\gamma.X_{n\gamma}, \sigma_\gamma^2)$

其中 $i = 1$、2、$\cdots$、P，以及 $n = 1$、2、$\cdots$、N。这里，每个 $\gamma_i \in \{0, 1\}$ 控制系数 β_i 和第 i 个协变量是纳入（$\gamma_i = 1$）还是排除（$\gamma_i = 0$）模型。在下文中，我们用 γ指代向量（γ_1、γ_2、$\cdots$、γ_P）。超参数 $h \in (0, 1)$ 控制总体稀疏性水平；特别是， hP 是先验纳入的协变量数量的期望值。$|\gamma|$ 系数 $\beta_\gamma \in \mathbb{R}^{|\gamma|}$ 服从标准高斯先验，其精度与 $\tau > 0$成正比。这里，$|\gamma| \in \{0, 1, 2, \cdots, P\}$ 表示纳入的协变量总数。响应变量Y_n 服从方差由逆伽马分布先验控制的高斯分布。请注意，我们没有包含偏置项，但在实践中这样做可能更合适。该模型的一个优点是，它会明确地对变量纳入情况进行推断，并允许我们定义后验纳入概率（PIPs），其中

$$\texttt{PIP}(i) := p(\gamma_i = 1|\mathcal{D}) \in [0, 1] \tag{8}$$

且 $\mathcal{D} = \{X, Y\}$ 是观测数据集。

3 主要结果

3.1 Subset wTGS 简介

在本小节中，我们回顾由 [11]提出的子集 wTGS。考虑以下（未归一化的）目标分布：

$$f(\gamma, i, \mathcal{S}) := p(\gamma|\mathcal{D})\frac{\frac{1}{2}\eta(\gamma_{-i})}{p(\gamma_i|\gamma_{-i}, \mathcal{D})}\mathcal{U}(\mathcal{S}|i, \mathcal{A}). \tag{9}$$

这里， $\mathcal{S}$遍历 $\{1, 2, \cdots, P\}$ 中所有大小为 S 且包含一个大小固定为 $A < S$的“锚点”集合 $\mathcal{A} \subset \{1, 2, \cdots, P\}$ 的子集，而 $\eta(\cdot)$ 是某种加权函数。此外， $U(\mathcal{S}|i, \mathcal{A})$ 是在 $\{1, 2, \cdots, P\}$ 的所有大小为 S 且同时包含 i 和 $\mathcal{A}$ 的子集上的均匀分布。

在实践中，该集合 $\mathcal{A}$ 可以在预热期内选定。Subset wTGS 通过为目标分布（9）定义一种采样方案来运行，该方案利用相对于 i 和 $\mathcal{S}$ 的吉布斯更新，以及相对于 γ_i的梅特罗波利化吉布斯更新。

- i-更新：边缘化 i 来自 (9) 得到

$$f(\gamma, \mathcal{S}) = p(\gamma|\mathcal{D})\phi(\gamma, \mathcal{S}) \tag{10}$$

其中我们定义

$$\phi(\gamma, \mathcal{S}) := \sum_{i \in \mathcal{S}} \frac{\frac{1}{2}\eta(\gamma_{-i})}{p(\gamma_i|\gamma_{-i}, \mathcal{D})} \mathcal{U}(\mathcal{S}|i, \mathcal{A}) \tag{11}$$

并利用了以下性质：$\mathcal{U}(\mathcal{S}|i, \mathcal{A}) = 0$ if $i \notin \mathcal{S}$。关键在于，计算 $\phi(\gamma, \mathcal{S})$ 的复杂度为 $\Theta(S)$，而非 $\Theta(P)$。我们可以依据分布，对 i 进行吉布斯更新

$$f(i|\gamma, \mathcal{S}) \sim \frac{\eta(\gamma_{-i})}{p(\gamma_i|\gamma_{-i}, \mathcal{D})} \mathcal{U}(\mathcal{S}|i, \mathcal{A}). \tag{12}$$

- γ更新：与 $wTGS$ 类似，我们对 γ_i 采用 Metropolized 吉布斯更新，从而产生确定性翻转 $\gamma_i \to 1-\gamma_i$。同样，边缘分布 $f(i)$ 与 PIP$(i)+\frac{\varepsilon}{P}$ 成正比，因此采样器会将计算资源集中于 PIP 较大的协变量。
- $\mathcal{S}$更新：通过吉布斯移动更新 $\mathcal{S}$，$\mathcal{S} \sim \mathcal{U}(\cdot|i, \mathcal{A})$。完整算法见算法1。

该算法的详细内容见算法1。此采样器对应的估计量定义为[11]:

$$\texttt{PIP}(i) \approx \sum_{t=1}^{T} \rho^{(t)} \big(\mathbf{1}\{i \in \mathcal{S}^{(t)}\} p(\gamma_i^{(t)} = 1|\gamma_{-i}^{(t)}, \mathcal{D}) + \mathbf{1}\{i \notin \mathcal{S}^{(t)}\} \gamma_i^{(t)} \big). \tag{13}$$

3.2 可变复杂度 wTGS 方案

在第 3.1 节的子集 wTGS中，每次 MCMC 迭代的条件 PIP 计算次数是固定的，即等于 S。下面，我们提出一种基于可变计算复杂度的 wTGS 方案（VC-wTGS），记为 ALG 2，其唯一要求是每次 MCMC 迭代的条件 PIP 计算次数的期望为 S。这意味着 $\mathbb{E}[S_t] = S$，其中 S_t 表示第 t次 MCMC 迭代中的条件 PIP 计算次数。

与 ALG 1相比，ALG 2 允许我们在 MCMC 迭代中使用不同的子集大小。通过 ALG 2，每次 MCMC 迭代中条件 PIP 计算次数的期望为 $P \times (S/P) + 0 \times (1 - S/P) = S$。由于我们的目标是界定每次有限迭代中的方差 T，因此我们不在 ALG 2 中提及 T_{burn}。实际上，我们通常会移除一些初始样本。我们还使用以下 Rao-Blackwell 化估计量的新版本：

$$\texttt{PIP}(i) \approx \sum_{t=1}^{T} \rho^{(t)} p(\gamma_i^{(t)} = 1|\gamma_{-i}^{(t)}, \mathcal{D}). \tag{14}$$

Alg算法 1 子集 S-wTGS 算法 g算法

输入：数据集 $\mathcal{D}=\{X,Y\}$，包含 P 个协变量；先验纳入概率 h；先验精度 τ；子集大小 S；锚点集大小 A；MCMC迭代总数 T；预热迭代数 T_{burn}。输出：近似加权后验样本 $\{\rho^{(t)},\gamma^{(t)}\}_{t=T_{\text{burn}}+1}^{T}$。初始化：$\gamma^{(0)}=(0,0,\cdots,0)$，并选择 $\mathcal{A}$ 作为与 Y 表现出最大相关性的 A 个协变量中的$\underbrace{\qquad}_{P}$。从 $\{1,2,\cdots,P\}$ 中随机选择 $i^{(0)}$，并令 $\mathcal{S}^{(0)}\sim\mathcal{U}(\cdot|i^{(0)},\mathcal{A})$。对于 $t=1,\ 2,\cdots,T$，执行：估计 $f(j|\gamma^{(t-1)})\leftarrow\phi_{t-1}(\gamma)_{-1}$ $\frac{1}{2}\eta(\gamma_{-j}^{(t-1)})p(\gamma_j^{(t-1)}|\gamma_{-j}^{(t-1)},\mathcal{D})$，对于所有 $j\in[P]$。采样 $i^{(t)}\sim f(\cdot|\gamma^{(t-1)})\gamma^{(t)}\leftarrow$ 翻转$(\gamma^{(t-1)}|i^{(t)})$，其中翻转$(\gamma|i)$翻转 i 的 γ 坐标：$\gamma_i\leftarrow 1-\gamma_i$。采样 $\mathcal{S}^{(t)}\sim\mathcal{U}(\cdot|i^{(t)},\mathcal{A})$。估计所有 $j\in\mathcal{S}^{(t)}\phi_t(\gamma)$的条件纳入概率 $p(\gamma_j^{(t)}|\gamma_{-j}^{(t)},\mathcal{D})$。$\leftarrow\sum_{j\in\mathcal{S}^{(t)}}$ $\frac{1}{2}\eta(\gamma_{-j}^{(t)})\ p(\gamma_j^{(t)}|\gamma_{-j}^{(t)},\mathcal{D})$。计算未归一化权重 $\tilde{\rho}^{(t)}\leftarrow\phi^{-1}(\gamma^{(t)})$；如果 $t\le T_{\text{burn}}$，则使用某种自适应方案调整 $\mathcal{A}$。结束if。结束 for。对于 $t=1,\ 2,\cdots,T$，执行 $\rho^{(t)}\leftarrow$ $\sum_{s>T}^{T}$ 预热$\tilde{\rho}^{(s)}$。结束 for。输出：$\{\rho^{(t)},\gamma^{(t)}\}_{t=1}^{T}$。

注1。 在算法 2中，使用伯努利随机变量 $\{Q^{(t)}\}_{t=1}^{T}$ 替代算法 $\mathcal{S}$ 1中的随机集合 $\mathcal{S}$。进行以下替换主要有两个原因：（1）从 $\binom{P}{S}$ 的子集 $[P]$ 生成随机集合 $(P,\ S)$通常需要很长时间；（2）在较高 时，算法 2 的 Rao-Blackwell 化估计量通常比算法 *1* 具有更小的方差 P/S。模拟结果见第 *4* 节。

3.3 算法2的理论界

首先，我们证明以下结果。证明见附录 A。

引理 1. 设 U 和 V 为两个正随机变量，使得对于某个常数 M，$U/V\le M$ 几乎处处成立。此外，假设在概率至少为$1-\alpha$的集合 D 上，我们有

$$|U-\mathbb{E}[U]|\le\varepsilon\mathbb{E}[U], \tag{15}$$
$$|V-\mathbb{E}[V]|\le\varepsilon\mathbb{E}[V], \tag{16}$$

Algorithm 2 A Variable-Complexity Based wTGS Algorithm

Input: Dataset $\mathcal{D} = \{X, Y\}$ with P covariates; prior inclusion probability h; prior precision τ; total number of MCMC iterations T; subset size S.
Output: Approximate weighted posterior samples $\{\rho^{(t)}, \gamma^{(t)}\}_{t=1}^{T}$
Initializations: $\gamma^{(0)} = (\gamma_1, \gamma_2, \cdots, \gamma_P)$ where $\gamma_j \sim \texttt{Bern}(h)$ for all $j \in [P]$.
for $t = 1, 2, \cdots, T$ **do**
  Set $Q^{(1)} = 1$. Sample a Bernoulli random variable $Q^{(t)} \sim \texttt{Bern}(\frac{S}{P})$ if $t \geq 2$.
  if $Q^{(t)} = 1$ **then**
    Estimate $f(j|\gamma^{(t-1)}) \leftarrow \phi_{t-1}(\gamma)^{-1} \frac{\frac{1}{2}\eta(\gamma_{-j}^{(t-1)})}{p(\gamma_j^{(t-1)}|\gamma_{-j}^{(t-1)}, \mathcal{D})}$ for all $j \in [P]$.
    Sample $i^{(t)} \sim f(\cdot|\gamma^{(t-1)})$
    $\gamma^{(t)} \leftarrow \texttt{flip}(\gamma^{(t-1)}|i^{(t)})$ where $\texttt{flip}(\gamma|i)$ flips the i-th coordinate of $\gamma : \gamma_i \leftarrow 1 - \gamma_i$.
    Estimate P conditional PIPs $p(\gamma_j^{(t)}|\gamma_{-j}^{(t)}, \mathcal{D})$ for all $j \in [P]$
    $\phi(\gamma^{(t)}) \leftarrow \sum_{j \in [P]} \frac{\frac{1}{2}\eta(\gamma_{-j}^{(t)})}{p(\gamma_j^{(t)}|\gamma_{-j}^{(t)}, \mathcal{D})}$
    Compute the unnormalized weights $\tilde{\rho}^{(t)} \leftarrow \phi^{-1}(\gamma^{(t)})$
  else
    $\gamma^{(t)} \leftarrow \gamma^{(t-1)}$
    $\tilde{\rho}^{(t)} \leftarrow 1$
  end if
end for
for $t = 1, 2, \cdots, T$ **do**
  $\rho^{(t)} \leftarrow \frac{\tilde{\rho}^{(t)} Q^{(t)}}{\sum_{s=1}^{T} \tilde{\rho}^{(s)} Q^{(s)}}$
end for
Output: $\{\rho^{(t)}, \gamma^{(t)}\}_{t=1}^{T}$.

对于某个 $0 \leq \varepsilon < 1$。那么，有

$$\mathbb{E}\left[\left|\frac{U}{V} - \frac{\mathbb{E}[U]}{\mathbb{E}[V]}\right|^2\right] \leq \frac{4\varepsilon^2}{(1-\varepsilon)^2}\left(\frac{\mathbb{E}[U]}{\mathbb{E}[V]}\right)^2 + \left[\max\left(M, \frac{\mathbb{E}[U]}{\mathbb{E}[V]}\right)\right]^2 \alpha. \tag{17}$$

我们还回顾以下适用于马尔可夫链的Hoeffding不等式：
引理 2。 *[15, 定理1。1]* 设 $\{Y_i\}_{i=1}^{\infty}$ 为一个平稳马尔可夫链，其状态空间为 $[N]$，转移矩阵为 A，平稳概率测度为 π，平均算子为 E_π，使得 Y_1 服从 π的分布。令 $\lambda = \|A - E_\pi\|_{L_2(\pi) \to L_2(\pi)}$ 并令f_1、f_2、$\cdots$、f_n：$[N] \to \mathbb{R}$ 使得 $\mathbb{E}[f_i(Y_i)] = 0$对所有 i成立，且 $|f_i(\nu)| \leq a_i$对所有 $\nu \in [N]$及所有 i成立。则对于 $u \geq 0$,

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} f_i(Y_i)\right| \geq u\left(\sum_{i=1}^{n} a_i^2\right)^{1/2}\right] \leq 2\exp\left(-\frac{u^2(1-\lambda)}{64e}\right). \tag{18}$$

现在，可以证明以下结果。

引理 3。设

$$\phi(\gamma) := \sum_{j\in[P]} \frac{\frac{1}{2}\eta(\gamma_{-j})}{p(\gamma_j|\gamma_{-j},\mathcal{D})} \tag{19}$$

并定义

$$f(\gamma) := \phi(\gamma)p(\gamma|\mathcal{D}). \tag{20}$$

然后，根据算法 2，序列 $\{\gamma^{(t)}, Q^{(t)}\}_{t=1}^{T}$ 构成一个可逆马尔可夫链，其平稳分布正比于 $f(\gamma)q(Q)$，其中 q 是伯努利（S/P）分布。该马尔可夫链的转移核为 $K((\gamma,Q) \to (\gamma',Q')) = K^*(\gamma \to \gamma')q(Q')$，其中

$$K^*(\gamma \to \gamma') = \frac{S}{P}\sum_{j=1}^{P} f(j|\gamma)\delta(\gamma' - \texttt{flip}(\gamma|j)) + \left(1 - \frac{S}{P}\right)\delta(\gamma' - \gamma). \tag{21}$$

在经典的 wTGS 中 [23], 马尔可夫链 $\{\gamma^{(t)}\}_{t=1}^{T}$ 也构成一个马尔可夫链。然而，该马尔可夫链不同于引理3中的马尔可夫链。不过，这两个马尔可夫链仍具有相同的平稳分布，其正比于 $f(\gamma)$。引理3 的详细证明见附录B。

引理 4。对于应用于(14)算法 $\{\rho^{(t)}, \gamma^{(t)}\}_{t=1}^{T}$ 2的输出序列的 Rao-Blackwell 化估计量，有

$$E_{i,T} := \sum_{t=1}^{T} \rho^{(t)} p(\gamma_i^{(t)} = 1|\gamma_{-i}^{(t)}, \mathcal{D}) \to \texttt{PIP}(i), \quad \texttt{as} \quad T \to \infty. \tag{22}$$

证明。 由引理 3， $\{\gamma^{(t)}, Q^{(t)}\}_{t=1}^{T}$ 构成一个可逆马尔可夫链，其平稳分布为 $f(\gamma)/Z_f q(Q)$，其中 $Z_f = \sum_\gamma f(\gamma)$。因此，根据马尔可夫链的强大数定律 [4],，对于任意有界函数 h，我们有

$$\frac{1}{T}\sum_{t=1}^{T} \phi^{-1}(\gamma^{(t)})Q^{(t)}h(\gamma^{(t)}) \to \mathbb{E}_{qf(\cdot)/Z_f}\left[\phi^{-1}(\gamma)h(\gamma)Q\right] \tag{23}$$

$$= \sum_Q q(Q)\sum_\gamma \frac{f(\gamma)}{Z_f}\phi^{-1}(\gamma)h(\gamma)Q \tag{24}$$

$$= \left(\sum_Q q(Q)Q\right)\left(\sum_\gamma \frac{f(\gamma)}{Z_f}\phi^{-1}(\gamma)h(\gamma)\right) \tag{25}$$

$$= \mathbb{E}_q[Q]\frac{1}{Z_f}\sum_\gamma p(\gamma|\mathcal{D})h(\gamma) \tag{26}$$

$$= \frac{S}{P}\frac{1}{Z_f}\sum_\gamma p(\gamma|\mathcal{D})h(\gamma), \tag{27}$$

其中（26）由 $f(\gamma) = p(\gamma|\mathcal{D})\phi(\gamma)$ 得出。

类似地，我们有

$$\frac{1}{T}\sum_{t=1}^{T} Q^{(t)}\phi^{-1}(\gamma^{(t)}) \to \mathbb{E}_{qf(\cdot)/Z_f}\left[\phi^{-1}(\gamma)Q\right] \tag{28}$$

$$= \sum_{Q} q(Q)Q \sum_{\gamma} \frac{f(\gamma)}{Z_f}\phi^{-1}(\gamma) \tag{29}$$

$$= \mathbb{E}_q[Q] \sum_{\gamma} \frac{1}{Z_f} p(\gamma|\mathcal{D}) \tag{30}$$

$$= \frac{S}{P}\frac{1}{Z_f}, \tag{31}$$

其中（30）也可由 $f(\gamma) = p(\gamma|D)\phi(\gamma)$. □

由（27）和（31），我们得到

$$\frac{\frac{1}{T}\sum_{t=1}^{T} \phi^{-1}(\gamma^{(t)})Q^{(t)}h(\gamma^{(t)})}{\frac{1}{T}\sum_{t=1}^{T} Q^{(t)}\phi^{-1}(\gamma^{(t)})} \to \sum_{\gamma} p(\gamma|\mathcal{D})h(\gamma), \tag{32}$$

或者等价地

$$\sum_{t=1}^{T} \rho^{(t)} h(\gamma^{(t)}) \to \sum_{\gamma} p(\gamma|\mathcal{D})h(\gamma) \tag{33}$$

作为 $T \to \infty$。

现在，通过设定 $h(\gamma) = p(\gamma_i = 1|\gamma_{-i}, \mathcal{D})$，由（33）可得

$$\sum_{t=1}^{T} \rho^{(t)} p(\gamma_i^{(t)} = 1|\gamma_{-i}^{(t)}, \mathcal{D}) \to \texttt{PIP}(i) \tag{34}$$

对于所有 $i \in [P]$。

下面的结果给出了有限 T 下 PIP 估计量的方差界。

引理 5. 对任意 $\varepsilon \in [0, 1]$，设 ν 和 π 分别为可逆马尔可夫序列 $\{(\gamma^{(t)}, Q^{(t)})\}$的初始分布和平稳分布。定义

$$\hat{\phi}(\gamma) := \frac{\phi^{-1}(\gamma)}{\max_{\gamma} \phi^{-1}(\gamma)}, \tag{35}$$

and

$$\varepsilon_0 = \frac{P}{\mathit{PIP}(i)\mathbb{E}_{\pi}[\hat{\phi}(\gamma)]S}\sqrt{\frac{64e\log T}{(1-\lambda_{\gamma,Q})T}}. \tag{36}$$

则有

$$\mathbb{E}\left[\left|\sum_{t=1}^{T}\rho^{(t)}p(\gamma_i^{(t)}=1|\gamma_{-i}^{(t)},\mathcal{D})-\mathit{PIP}(i)\right|^2\right]\leq\frac{4\varepsilon_0^2}{(1-\varepsilon_0)^2}\mathit{PIP}^2(i)+\frac{4P}{S}\frac{1}{\min_\gamma\pi(\gamma)T}\to 0, \tag{37}$$

当 $T\to\infty$ 、 P 和数据集保持固定时为 S 。这里， $\pi(\gamma)$ 是 $\pi(\gamma,Q)$ 的边缘分布。

证明。 见附录 C。 □

接下来，我们给出 $1-\lambda_{\gamma,\ Q}$ 的下界。首先，回顾以下关于谱隙的狄利克雷形式。

定义 1. 设 f、 $g:\Omega\to\mathbb{R}$。与 Ω 上的可逆马尔可夫链 Q 相关联的狄利克雷形式定义为

$$\begin{aligned}\mathcal{E}(f,g)&=\langle(\mathbf{I}-\mathbf{Q})f,g\rangle_\pi &(38)\\&=\sum_{x\in\Omega}\pi(x)[f(x)-\mathbf{Q}f(x)]g(x) &(39)\\&=\sum_{x\in\Omega}\pi(x)\left[\sum_y Q(x,y)(f(x)-f(y))\right]g(x) &(40)\\&=\sum_{x,y\in\Omega\times\Omega}\pi(x)Q(x,y)g(x)(f(x)-f(y)). &(41)\end{aligned}$$

引理 6. *[7]* （变分刻画）对于状态空间为 Ω 且平稳分布为 π 的可逆马尔可夫链 Q ，有

$$1-\lambda=\inf_{\substack{g:\Omega\to\mathbb{R},\\ \mathbb{E}_\pi[g]=0,\mathbb{E}_\pi[g^2]=1}}\mathcal{E}(g,g), \tag{42}$$

其中 $\mathcal{E}(g,g):=\langle(\mathrm{I}-\mathrm{Q})g,g\rangle_\pi$。
引理 7.谱隙$1-\lambda_{\gamma,Q}$ 可逆马尔可夫链 $\{\gamma^{(t)},Q^{(t)}\}$满足

$$1-\lambda_{\gamma,Q}\geq 1-\frac{S}{P}\lambda_P\geq 1-\frac{S}{P} \tag{43}$$

其中， $1-\lambda_P$是 wTGS 算法的可逆马尔可夫链 $\{\gamma^{(t)}\}$ 的谱隙（即 $S=P$）。

证明。 由引理 6 以及事实，即 $\{\gamma^{(t)},Q^{(t)}\}$ 构成具有转移核 $K((\gamma,Q)\to(\gamma',Q'))=K^*(\gamma\to\gamma')q(Q')$ 的可逆马尔可夫链，我们有

$$\begin{aligned}&1-\lambda_{\gamma,Q}\\&=\inf_{g(\gamma,Q):\mathbb{E}_\pi[g]=0,\mathbb{E}_\pi[g^2]=1}\langle g,g\rangle_\pi-\langle\mathbf{K}g,g\rangle &(44)\end{aligned}$$

$$= 1 - \sup_{g(\gamma,Q):\mathbb{E}_\pi[g]=0,\mathbb{E}_\pi[g^2]=1} \langle \mathbf{K}g, g\rangle \tag{45}$$

$$= 1 - \sup_{g(\gamma,Q):\mathbb{E}_\pi[g]=0,\mathbb{E}_\pi[g^2]=1} \sum_{\gamma,Q} \mathbf{K}g(\gamma,Q)g(\gamma,Q)\pi(\gamma,Q) \tag{46}$$

$$= 1 - \sup_{g(\gamma,Q):\mathbb{E}_\pi[g]=0,\mathbb{E}_\pi[g^2]=1} \sum_{\gamma,Q}\sum_{\gamma',Q'} K((\gamma,Q)\to(\gamma',Q'))g(\gamma',Q')g(\gamma,Q)\pi(\gamma,Q) \tag{47}$$

$$= 1 - \sup_{g(\gamma,Q):\mathbb{E}_\pi[g]=0,\mathbb{E}_\pi[g^2]=1} \frac{S}{P}\sum_{\gamma,Q}\sum_{\gamma',Q'} K^*(\gamma\to\gamma')q(Q')g(\gamma',Q')g(\gamma,Q)\pi(\gamma,Q) \tag{48}$$

$$= 1 - \frac{S}{P}\sup_{g(\gamma,Q):\mathbb{E}_\pi[g]=0,\mathbb{E}_\pi[g^2]=1} \sum_{\gamma,Q}\sum_{\gamma',Q'} K^*(\gamma\to\gamma')\frac{f(\gamma)}{Z_f}q(Q)g(\gamma',Q')g(\gamma,Q)q(Q') \tag{49}$$

$$= 1 - \frac{S}{P}\sup_{g(\gamma,Q):\mathbb{E}_\pi[g]=0,\mathbb{E}_\pi[g^2]=1} \sum_{\gamma,\gamma'} K^*(\gamma\to\gamma')\frac{f(\gamma)}{Z_f}\sum_{Q,Q'} g(\gamma',Q')g(\gamma,Q)q(Q)q(Q') \tag{50}$$

$$= 1 - \frac{S}{P}\sup_{g(\gamma,Q)\ :\mathbb{E}_\pi[g]=0,\mathbb{E}_\pi[g^2]=1} \sum_{\gamma,\gamma'} K^*(\gamma\to\gamma')\pi(\gamma)\Big(\sum_{Q} g(\gamma,Q)q(Q)\Big)\Big(\sum_{Q'}\pi(\gamma',Q')q(Q')\Big) \tag{51}$$

$$= 1 - \frac{S}{P}\sup_{g(\gamma,Q):\mathbb{E}_\pi[g]=0,\mathbb{E}_\pi[g^2]=1} \sum_{\gamma,\gamma'} K^*(\gamma\to\gamma')\pi(\gamma)h(\gamma)h(\gamma') \tag{52}$$

其中

$$\pi(\gamma) = \frac{f(\gamma)}{Z_f}, \tag{53}$$

$$Z_f = \sum_\gamma f(\gamma), \tag{54}$$

$$h(\gamma) := \sum_Q g(\gamma,Q)q(Q). \tag{55}$$

注意,

$$\mathbb{E}_\pi[h(\gamma)] = \sum_\gamma h(\gamma)\pi(\gamma) \tag{56}$$

$$= \sum_\gamma\sum_Q g(\gamma,Q)q(Q)\pi(\gamma) \tag{57}$$

$$= \sum_{\gamma,Q} g(\gamma,Q)\pi(\gamma,Q) \tag{58}$$

$$= \mathbb{E}_\pi[g(\gamma,Q)] \tag{59}$$

$$= 0. \tag{60}$$

另一方面，我们也有

$$\mathbb{E}_\pi\big[h^2(\gamma)\big] = \sum_\gamma \Big(\sum_Q g(\gamma,Q)q(Q)\Big)^2 \pi(\gamma) \tag{61}$$

$$\leq \sum_\gamma \Big(\sum_Q g(\gamma,Q)^2 q(Q)\Big)\pi(\gamma) \tag{62}$$

$$= \sum_{\gamma,Q} g(\gamma,Q)^2 \pi(\gamma,Q) \tag{63}$$

$$= \mathbb{E}_\pi\big[g(\gamma,Q)^2\big] \tag{64}$$

$$= 1, \tag{65}$$

其中（62）源于函数 x^2 在 $[0,\infty)$。

由（60）、（65）和（52），我们得到

$$1-\lambda_{\gamma,Q} \geq 1 - \frac{S}{P} \sup_{h(\gamma):\mathbb{E}_\pi[h]=0,\mathbb{E}_\pi[h^2]\leq 1} \sum_{\gamma,\gamma'} K^*(\gamma\to\gamma')\pi(\gamma)h(\gamma)h(\gamma'). \tag{66}$$

现在，注意到 $\mathbb{E}_\pi[h]=0$ 等价于 $h \perp_\pi \mathbf{1}$。令 $|\Omega| = 2^{P+1} := n$，并且 h_1、h_2、$\cdots$、h_n 是 $\mathbf{K}^*$ 对应于按递减顺序排列的特征值 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ 的特征函数，并且由于 $\mathbf{K}^*$ 是自伴的，它们相互正交。设 $h_1 = \mathbf{1}$。由于 $\|h\|_{2,\,\pi}=1$ 和 $h \perp_\pi \mathbf{1}$，我们有 $h = \sum_{j=2}^n a_j h_j$，因为它垂直于 h_1，所以只能由这些特征向量表示。对两边取 l_2-范数，我们有 $\sum_{j=2}^n a_j^2 \leq 1$，因为形如 $\langle h_i, h_j\rangle_\pi = 0$ 和 $\langle h_i, h_i\rangle = \|h_i\|_{2,\,\pi}^2 = 1$。因此，

$$\sup_{h:\mathbb{E}_\pi[h]=0,\mathbb{E}_\pi[h^2]\leq 1} \sum_{\gamma,\gamma'} K^*(\gamma\to\gamma')\pi(\gamma)h(\gamma)h(\gamma') \leq \max_{a_2,a_3,\cdots,a_n:\sum_{j=2}^n a_j^2\leq 1} \sum_{j=1}^n a_j^2\lambda_j \tag{67}$$

$$\leq \lambda_2 \sum_{j=2}^n a_j^2 \tag{68}$$

$$= \lambda_2, \tag{69}$$

其中 $\sum_{j=2}^n a_j^2 \leq 1$ 且 $\lambda_j \in$ 满足条件(P)，使得 $\lambda_2 \geq \lambda_3 \cdots \geq \lambda_n$。因此，由式（69），可得

$$1-\lambda_{\gamma,Q} \geq 1 - \frac{S}{P}\lambda_2 \tag{70}$$

$$= 1 - \frac{S}{P}\lambda_P \tag{71}$$

$$= \frac{S}{P}\big(1-\lambda_P\big) + 1 - \frac{S}{P} \tag{72}$$

$$\geq 1 - \frac{S}{P}. \tag{73}$$

□

结合引理 4、引理 5 和引理 7，我们得到以下定理。

定理 1. 对于 (14)中的基于可变复杂度子集 wTGS 的估计量，给定数据集 (X, Y)，有

$$E_{i,T} := \sum_{t=1}^{T} \rho^{(t)} p(\gamma_i^{(t)} = 1 | \gamma_{-i}^{(t)}, \mathcal{D}) \to \texttt{PIP}(i), \quad \texttt{as} \quad T \to \infty \tag{74}$$

and

$$\mathbb{E}\left[\left| \sum_{t=1}^{T} \rho^{(t)} p(\gamma_i^{(t)} | \gamma_{-i}^{(t)}, \mathcal{D}) - \texttt{PIP}(i) \right|^2 \right] = O\left(\frac{\log T}{T} \left(\frac{P}{S} \right)^2 \left(\frac{\max_\gamma \phi(\gamma)}{\min_\gamma \phi(\gamma)} \right)^2 \right), \tag{75}$$

其中

$$\phi(\gamma) = \frac{1}{2} \sum_{j \in [P]} \frac{p(\gamma_j = 1 | \gamma_{-j}, \mathcal{D})}{p(\gamma_j | \gamma_{-j}, \mathcal{D})}. \tag{76}$$

证明。 首先，（74）已在引理 4中证明。现在，我们利用引理 5）和引理 5 以及引理 7来证明（75）。

注意到

$$\begin{aligned} \mathbb{E}_\pi[\hat{\phi}(\gamma)] &= \mathbb{E}_\pi \left[\frac{\phi^{-1}(\gamma)}{\max_\gamma \phi^{-1}(\gamma)} \right] \\ &\geq \frac{\min_\gamma \phi^{-1}(\gamma)}{\max_\gamma \phi^{-1}(\gamma)} \qquad (77) \\ &= \frac{\min_\gamma \phi(\gamma)}{\max_\gamma \phi(\gamma)}. \qquad (78) \end{aligned}$$

此外，我们还

$$\begin{aligned} \phi(\gamma) &= \sum_{j \in [P]} \frac{\frac{1}{2}\eta(\gamma_{-j})}{p(\gamma_j | \gamma_{-j}, \mathcal{D})} \qquad (79) \\ &= \frac{1}{2} \sum_{j \in [P]} \frac{p(\gamma_j = 1 | \gamma_{-j}, \mathcal{D})}{p(\gamma_j | \gamma_{-j}, \mathcal{D})}. \qquad (80) \end{aligned}$$

现在，请注意，

$$\frac{p(\gamma_j = 1 | \gamma_{-j}, \mathcal{D})}{p(\gamma_j | \gamma_{-j}, \mathcal{D})} = \begin{cases} 1, & \gamma_j = 1 \\ \frac{p(\gamma_j = 1 | \gamma_{-j}, \mathcal{D})}{p(\gamma_j = 0 | \gamma_{-j}, \mathcal{D})}, & \gamma_j = 0. \end{cases} \tag{81}$$

在附录D中，我们可以根据数据集估计 $\frac{p(\gamma_j=1|\gamma_{-j},\mathcal{D})}{p(\gamma_j=0|\gamma_{-j},\mathcal{D})}$。具体而言，设 $\tilde{\gamma}_1$ 由 γ_{-i} 给出，其中 $\gamma_i = 1$；$\tilde{\gamma}_0$ 由 γ_{-i} 给出，其中 $\gamma_i = 0$，那么我们可以证明彩票计划

$$\frac{p(\gamma_j = 1|\gamma_{-j}, \mathcal{D})}{p(\gamma_j = 0|\gamma_{-j}, \mathcal{D})} = \left(\frac{h}{1-h}\right)\sqrt{\tau\frac{\det(X_{\tilde{\gamma}_0}^T X_{\tilde{\gamma}_0} + \tau I)}{\det(X_{\tilde{\gamma}_1}^T X_{\tilde{\gamma}_1} + \tau I)}}\left(\frac{\|Y\|^2 - \|\tilde{Y}_{\tilde{\gamma}_0}\|^2 + \nu_0\lambda_0}{\|Y\|^2 - \|\tilde{Y}_{\tilde{\gamma}_1}\|^2 + \nu_0\lambda_0}\right)^{\frac{N+\nu_0}{2}}. \tag{82}$$

这里，

$$\begin{aligned}\|\tilde{Y}_\gamma\|^2 &= \tilde{Y}_\gamma^T \tilde{Y}_\gamma &(83)\\ &= Y^T X_\gamma (X_\gamma^T X_\gamma + \tau I)^{-1} X_\gamma^T Y. &(84)\end{aligned}$$

□

4 实验

在本节中，我们通过模拟展示 PIP 估计量是收敛的，随着 $T \to \infty$。此外，我们在模拟数据集和真实数据集上比较 VC-wTGS 与 Subset wTGS 相关 Rao-Blackwell 化估计量的方差。为了计算 $p(\gamma_i|\gamma_{-i},\ Y)$，对于新的设置，我们采用与 [23, 附录 B1] 中相同的技巧。关于该后验分布的推导，请参见附录 D。

4.1 模拟数据集

首先，我们进行一项模拟实验。设 $X \in \mathbb{R}^{N\times P}$ 是一个多元（随机）高斯矩阵的一个实现。我们考虑 $N = 100$ 和 $P = 200$ 的情形。我们运行 $T = 20000$ 次迭代。

图 1 显示，式（14）中的 Rao-Blackwell 化估计量在 $T \to \infty$ 时收敛于 PIP 的值，其中 S 取不同值。图 2 显示了在 S 取不同值时，对 PIP(0) 和 PIP(1) 进行估计的方差≈ 0 在 T 足够大时，wTGS 的方差非常小，因此，
variable-complexitywTGS 的方差在 T 足够大时也很小。

4.2 真实数据集

在本次模拟中，我们在 MNIST 数据集上运行算法 2。图 3 和图 4 分别绘制了不同 S 取值下的 PIP(0) 和 PIP(1)，以及式（14）中 Rao-Blackwell 化估计量的估计方差。这些图展示了估计 PIP(0) 和 PIP(1) 时计算复杂度与估计方差之间的权衡。在运行 T 次 MCMC 迭代时，算法 2 中期望的 PIP 计算次数仅为 ST，而 wTGS 中为 TP。然而，方差会随之增加。根据定理 1，对于给定数据集，方差为 $O\left(\left(\frac{P}{S}\right)^{2\log_T T}\right)$，即最多增加（$P/S$）^{2} 倍。~~对于~~许多应用，我们并不需要精确估计后验纳入概率，因此可以使用 VC-wTGS 来降低计算复杂度，尤其是在 P 非常大（拥有数百万个协变量）时。图 5 显示，在 [11] 取值较高时，VC-wTGS 的表现优于 Subset wTGS P/S。

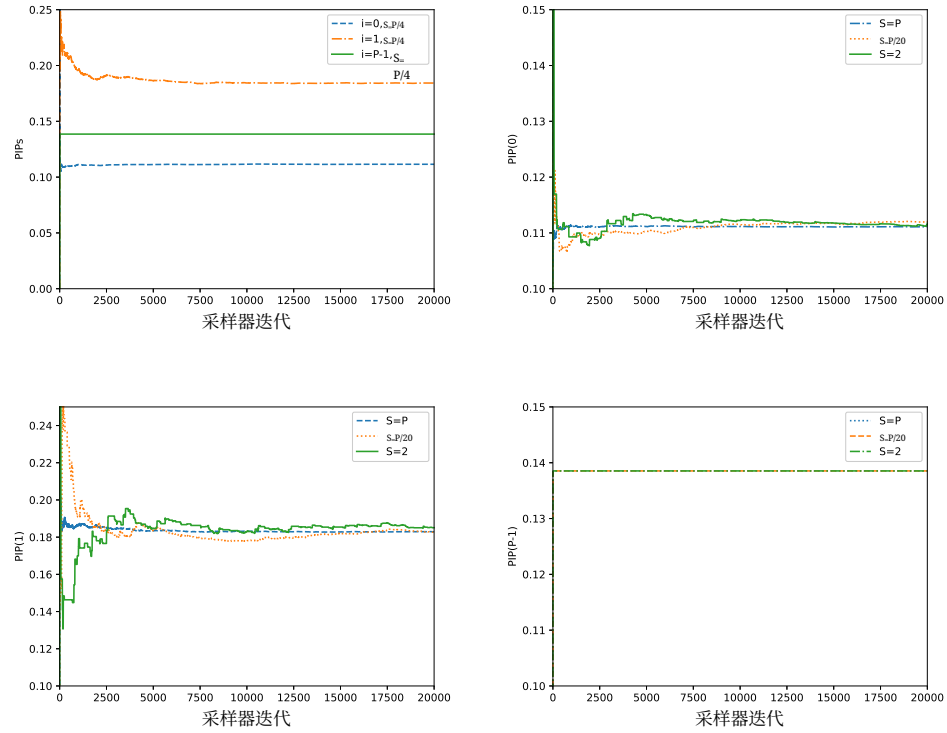

图 1：VC-wTGS Rao-Blackwell化估计量（算法 2）

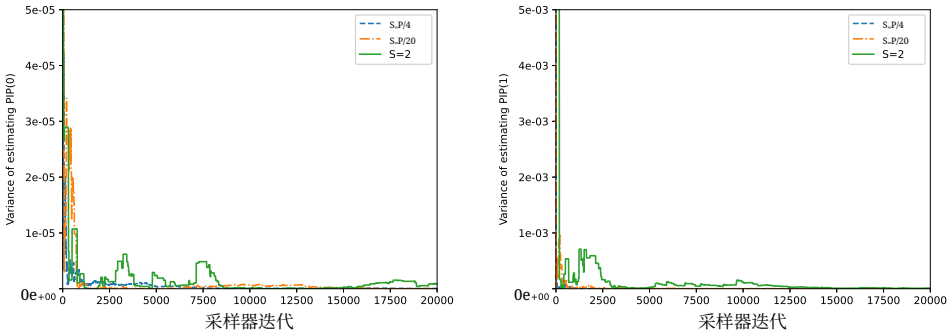

图 2：VC-wTGS Rao-Blackwell 化估计量的方差

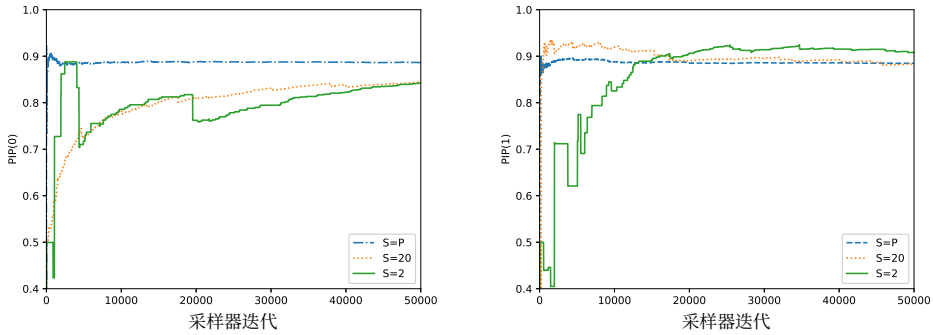

图 3：VC-wTGS Rao-Blackwell化估计量（算法 2）

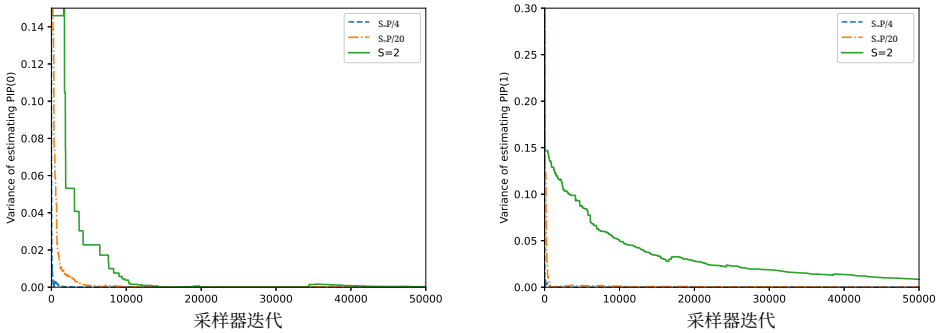

图 4：VC-wTGS Rao-Blackwell 化估计量的方差

附录 A 引理 1的证明

注意，以至少 $1-\alpha$ 的概率，我们有

$$(1-\varepsilon)\mathbb{E}[U] \leq U \leq (1+\varepsilon)\mathbb{E}[U] \tag{A1}$$
$$(1-\varepsilon)\mathbb{E}[V] \leq V \leq (1+\varepsilon)\mathbb{E}[V]. \tag{A2}$$

因此，我们有

$$\left(\frac{1-\varepsilon}{1+\varepsilon}\right)\frac{\mathbb{E}[U]}{\mathbb{E}[V]} \leq \frac{U}{V} \leq \left(\frac{1+\varepsilon}{1-\varepsilon}\right)\frac{\mathbb{E}[U]}{\mathbb{E}[V]}. \tag{A3}$$

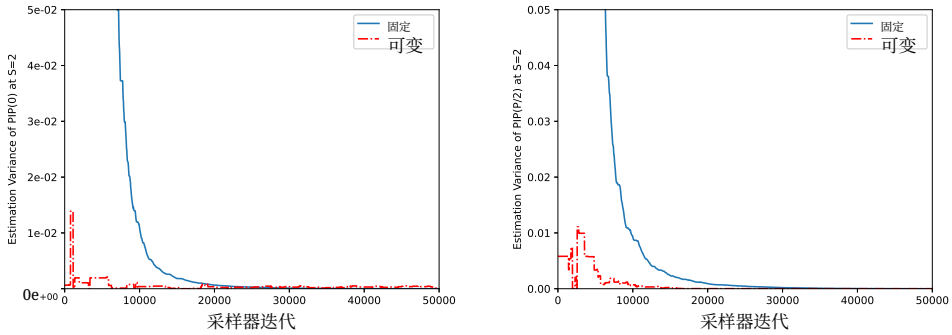

图 5：比较 $S = 2$ 处子集 wTGS 与 VC-wTGS 的方差。

根据（A3），以至少 $1 - \alpha$的概率，有

$$\left|\frac{U}{V} - \frac{\mathbb{E}[U]}{\mathbb{E}[V]}\right| \leq \frac{2\varepsilon}{1-\varepsilon}\left(\frac{\mathbb{E}[U]}{\mathbb{E}[V]}\right). \tag{A4}$$

由（A4）可知

$$\mathbb{E}\left[\left|\frac{U}{V} - \frac{\mathbb{E}[U]}{\mathbb{E}[V]}\right|^2\right] = \mathbb{E}\left[\left|\frac{U}{V} - \frac{\mathbb{E}[U]}{\mathbb{E}[V]}\right|^2 \middle| D\right]\mathbb{P}(D) + \mathbb{E}\left[\left|\frac{U}{V} - \frac{\mathbb{E}[U]}{\mathbb{E}[V]}\right|^2 \middle| D^c\right]\mathbb{P}(D^c) \tag{A5}$$

$$\leq \frac{4\varepsilon^2}{(1-\varepsilon)^2}\left(\frac{\mathbb{E}[U]}{\mathbb{E}[V]}\right)^2 + \left[\max\left(M, \frac{\mathbb{E}[U]}{\mathbb{E}[V]}\right)\right]^2 \alpha. \tag{A6}$$

附录 B 引理 3

序列 $\{\gamma^{(t)}\}$ 的转移核可以写成

$$K^*(\gamma \to \gamma') = \frac{S}{P}\sum_{j=1}^{P} f(j|\gamma)\delta(\gamma' - \texttt{flip}(\gamma|j)) + \left(1 - \frac{S}{P}\right)\delta(\gamma' - \gamma). \tag{B7}$$

这意味着，对于任意一对 (γ, γ')，只要存在某个 $i \in [P]$ 使得 $\gamma' = \texttt{flip}(\gamma|i)$，就有

$$K^*(\gamma \to \gamma') = \frac{S}{P}\sum_{j=1}^{P} f(j|\gamma)\delta(\gamma' - \texttt{flip}(\gamma|j)) \tag{B8}$$

$$= \frac{S}{P} f(i|\gamma). \tag{B9}$$

现在，由 ALG 2，我们也有

$$f(i|\gamma) = \phi^{-1}(\gamma)\frac{\frac{1}{2}\eta(\gamma_{-i})}{p(\gamma_i|\gamma_{-i},\mathcal{D})} \tag{B10}$$

and

$$f(i|\gamma') = \phi^{-1}(\gamma')\frac{\frac{1}{2}\eta(\gamma'_{-i})}{p(\gamma'_i|\gamma'_{-i},\mathcal{D})}. \tag{B11}$$

由（B10）和（B11）以及 $\gamma_{-i} = \gamma'_{-i}$，可得

$$\frac{K^*(\gamma \to \gamma')}{K^*(\gamma' \to \gamma)} = \frac{\frac{S}{P}f(i|\gamma)}{\frac{S}{P}f(i|\gamma')} \tag{B12}$$

$$= \frac{f(i|\gamma)}{f(i|\gamma')} \tag{B13}$$

$$= \frac{\phi(\gamma')p(\gamma'|\mathcal{D})}{\phi(\gamma)p(\gamma|\mathcal{D})} \tag{B14}$$

$$= \frac{f(\gamma')}{f(\gamma)}. \tag{B15}$$

此外，我们还有 $K^*(\gamma \to \gamma') = K^*(\gamma' \to \gamma) = 0$ 如果 $\gamma' \neq \gamma$ 且 $\gamma' \neq \text{flip}(\gamma|i)$ 对任意 $i \in [P]$。此外，$K^*(\gamma \to \gamma') = K^*(\gamma' \to \gamma) = 1 - \frac{S}{P}$ 如果 $\gamma = \gamma'$。

综合所有这些情况，可得出

$$f(\gamma)K^*(\gamma \to \gamma') = f(\gamma')K^*(\gamma' \to \gamma) \tag{B16}$$

对所有 γ', γ。

这意味着 $\{\gamma^{(t)}\}_{t=1}^T$ 构成一个具有平稳分布 $f(\gamma)/Z_f$ 的可逆马尔可夫链，其中

$$Z_f = \sum_\gamma f(\gamma). \tag{B17}$$

由于 $\{Q_t\}_{t=1}^T$ 是独立同分布的伯努利序列，且 $q(1) = S/P$ 并与 $\{\gamma^{(t)}\}_{t=1}^T$ 独立，因此 $\{\gamma^{(t)}, Q^{(t)}\}_{t=1}^T$ 构成一个马尔可夫链，其转移核满足：

$$K((\gamma, Q) \to (\gamma', Q')) = q(Q')K^*(\gamma \to \gamma'). \tag{B18}$$

由（B18）可得

$$q(Q)f(\gamma)/Z_f K((\gamma, Q) \to (\gamma', Q')) = \left[K^*(\gamma \to \gamma')f(\gamma)/Z_f\right]q(Q)q(Q') \tag{B19}$$

对于任意一对 (γ, Q) 和 (γ', Q')。

最后，根据 (B16) 和 (B19)，可得

$$q(Q)f(\gamma)/Z_f K((\gamma,Q)\to(\gamma',Q')) = q(Q')f(\gamma)/Z_f K((\gamma',Q')\to(\gamma,Q)). \tag{B20}$$

这意味着 $\{\gamma_t, Q^{(t)}\}_{t=1}^T$ 构成一条可逆马尔可夫链，其平稳分布为 $q(Q)f(\gamma)/Z_f$。

附录C 引理 5

ˆ 首先，根据 $\phi(\gamma)$ 在（35）中的定义，我们有

$$\rho^{(t)} = \frac{\hat{\phi}(\gamma^{(t)})}{\sum_{t=1}^T \hat{\phi}(\gamma^{(t)})}. \tag{C21}$$

此外，注意到

$$0 \le \hat{\phi}(\gamma) \le 1. \tag{C22}$$

现在，令 $g: \{0,1\}^P \to \mathbb{R}_+$，使得对所有 γ，$g(\gamma) \le 1$。于是，应用引理 2并进行测度变换，以概率 $1 - 2\frac{d\nu}{d\pi}\exp(_\ \zeta^2 T(1-\lambda)64e)$，我们有

$$\frac{1}{T}\left|\sum_{t=1}^T \hat{\phi}(\gamma^{(t)})g(\gamma^{(t)})Q^{(t)} - \mathbb{E}_\pi\left[\sum_{t=1}^T \hat{\phi}(\gamma^{(t)})g(\gamma^{(t)})Q^{(t)}\right]\right| \le \zeta \tag{C23}$$

对任意 $\zeta > 0$。

同样地，利用引理2，以至少 $1 - 2\frac{d\nu}{d\pi}\exp(_\ \zeta^2 T(1-\lambda)64e)$ 的概率，有

$$\frac{1}{T}\left|\sum_{t=1}^T \hat{\phi}(\gamma^{(t)})Q^{(t)} - \mathbb{E}_\pi\left[\sum_{t=1}^T \hat{\phi}(\gamma^{(t)})Q^{(t)}\right]\right| \le \zeta. \tag{C24}$$

利用并集界，以至少 $1 - 4\frac{d\nu}{d\pi}\exp(_\ \zeta^2 T(1-\lambda)64e)$ 的概率，有

$$\frac{1}{T}\left|\sum_{t=1}^T \hat{\phi}(\gamma^{(t)})g(\gamma^{(t)})Q^{(t)} - \mathbb{E}_\pi\left[\sum_{t=1}^T \hat{\phi}(\gamma^{(t)})g(\gamma^{(t)})Q^{(t)}\right]\right| \le \zeta, \tag{C25}$$

$$\frac{1}{T}\left|\sum_{t=1}^T \hat{\phi}(\gamma^{(t)}) - \mathbb{E}_\pi\left[\sum_{t=1}^T \hat{\phi}(\gamma^{(t)})\right]\right| \le \zeta. \tag{C26}$$

现在，令 $\zeta = \zeta_0 := \frac{\varepsilon}{T}\min\left\{\mathbb{E}_\pi\left[\sum_{t=1}^T \hat{\phi}(\gamma^{(t)})g(\gamma^{(t)})Q^{(t)}\right], \mathbb{E}_\pi\left[\sum_{t=1}^T \hat{\phi}(\gamma^{(t)})\right]\right\}$，其中 $\varepsilon > 0$（稍后选取），则以至少 $1 - 4\frac{d\nu}{d\pi}\exp(_\ \zeta_0^2 T(1-\lambda)64e)$ 的概率，

有

$$\frac{1}{T}\left|\sum_{t=1}^{T}\hat{\phi}(\gamma^{(t)})g(\gamma^{(t)})Q^{(t)}-\mathbb{E}_{\pi}\left[\sum_{t=1}^{T}\hat{\phi}(\gamma^{(t)})g(\gamma^{(t)})Q^{(t)}\right]\right|$$
$$\leq\frac{\varepsilon}{T}\mathbb{E}_{\pi}\left[\sum_{t=1}^{T}\hat{\phi}(\gamma^{(t)})g(\gamma^{(t)})Q^{(t)}\right], \tag{C27}$$

$$\frac{1}{T}\left|\sum_{t=1}^{T}\hat{\phi}(\gamma^{(t)})Q^{(t)}-\mathbb{E}_{\pi}\left[\sum_{t=1}^{T}\hat{\phi}(\gamma^{(t)})Q^{(t)}\right]\right|\leq\frac{\varepsilon}{T}\mathbb{E}_{\pi}\left[\sum_{t=1}^{T}\hat{\phi}(\gamma^{(t)})Q^{(t)}\right]. \tag{C28}$$

此外，通过设置

$$U:=\frac{1}{T}\sum_{t=1}^{T}\hat{\phi}(\gamma^{(t)})g(\gamma^{(t)})Q^{(t)}, \tag{C29}$$

$$V:=\frac{1}{T}\sum_{t=1}^{T}\hat{\phi}(\gamma^{(t)})Q^{(t)}, \tag{C30}$$

我们有

$$\frac{U}{V}=\frac{\sum_{t=1}^{T}\phi^{-1}(\gamma^{(t)})g(\gamma^{(t)})Q^{(t)}}{\sum_{t=1}^{T}\phi^{-1}(\gamma^{(t)})Q^{(t)}} \tag{C31}$$

$$=\sum_{t=1}^{T}\rho^{(t)}g(\gamma^{(t)}) \tag{C32}$$

and

$$M:=\sup(U/V)\leq 1 \tag{C33}$$

由于对于所有 $\gamma^{(t)}$, $\sum_{t=1}^{T}\rho^{(t)}=1$ 和 $g\left(\gamma^{(t)}\right)\leq 1$。

由（C26）-（C33），根据引理 1，可得

$$\mathbb{E}\left[\left|\sum_{t=1}^{T}\rho^{(t)}g(\gamma^{(t)})Q^{(t)}-\frac{\mathbb{E}_{\pi}[U]}{\mathbb{E}_{\pi}[V]}\right|^{2}\right]\leq\frac{4\varepsilon^{2}}{(1-\varepsilon)^{2}}\left(\frac{\mathbb{E}_{\pi}[U]}{\mathbb{E}_{\pi}[V]}\right)^{2}+\left[\max\left(1,\frac{\mathbb{E}_{\pi}[U]}{\mathbb{E}_{\pi}[V]}\right)\right]^{2}\alpha, \tag{C34}$$

其中 $\alpha:=4\frac{d\nu}{d\pi}\exp\left(-\frac{\varepsilon^{2}T_{(1-\lambda_{\gamma,Q})\text{最小}}\{\mathbb{E}_{\pi}[U],\mathbb{E}_{\pi}[V]\}^{2}}{64e}\right)$，其中 $\lambda_{\gamma,Q}$ 是可逆马尔可夫链 $\{\gamma^{(t)},Q^{(t)}\}$的平稳分布。

现在，通过设定

$$\varepsilon=\varepsilon_{0}=\frac{1}{\min\{\mathbb{E}_{\pi}[U],\mathbb{E}_{\pi}[V]\}}\sqrt{\frac{64e\log T}{(1-\lambda_{\gamma,Q})T}}, \tag{C35}$$

我们得到 $\alpha = 4\frac{d\nu}{d\pi}\frac{1}{T}$。然后，我们得到

$$\mathbb{E}\left[\left|\sum_{t=1}^{T}\rho^{(t)}g(\gamma^{(t)}) - \frac{\mathbb{E}_\pi[U]}{\mathbb{E}_\pi[V]}\right|^2\right] \leq \frac{4\varepsilon_0^2}{(1-\varepsilon_0)^2}\left(\frac{\mathbb{E}_\pi[U]}{\mathbb{E}_\pi[V]}\right)^2 + \left[\max\left(1, \frac{\mathbb{E}_\pi[U]}{\mathbb{E}_\pi[V]}\right)\right]^2\alpha. \tag{C36}$$

现在，注意到

$$\frac{\mathbb{E}_\pi[U]}{\mathbb{E}_\pi[V]} = \frac{\mathbb{E}_\pi\left[g(\gamma)Q\hat{\phi}(\gamma)\right]}{\mathbb{E}_\pi\left[\hat{\phi}(\gamma)Q\right]} \tag{C37}$$

$$= \frac{\mathbb{E}_\pi\left[g(\gamma)Q\phi^{-1}(\gamma)\right]}{\mathbb{E}_\pi\left[\phi^{-1}(\gamma)Q\right]}. \tag{C38}$$

另一方面，由引理 3，我们有 $\pi(\gamma, Q) = {}^{q(Q)f(\gamma)}{}_{Z_f}$，其中 $Z_f := \sum_\gamma f(\gamma)$ 且 $f(\gamma) = p(\gamma|\mathcal{D})\phi(\gamma)$。由此可得

$$\mathbb{E}_\pi\left[g(\gamma)Q\phi^{-1}(\gamma)\right] = \mathbb{E}_{q(Q)f(\gamma)/Z_f}\left[g(\gamma)Q\phi^{-1}(\gamma)\right] \tag{C39}$$

$$= \sum_\gamma\sum_Q g(\gamma)Q\phi^{-1}(\gamma)\frac{f(\gamma)}{Z_f}q(Q) \tag{C40}$$

$$= \frac{1}{Z_f}\sum_\gamma\sum_Q g(\gamma)q(Q)Qp(\gamma|\mathcal{D}) \tag{C41}$$

$$= \frac{1}{Z_f}\mathbb{E}_{p(\gamma|\mathcal{D})}\left[g(\gamma)\right]\mathbb{E}_q[Q]. \tag{C42}$$

同理，我们有

$$\mathbb{E}_\pi\left[\phi^{-1}(\gamma)Q\right] = \mathbb{E}_{q(Q)f(\gamma)/Z_f}\left[\phi^{-1}(\gamma)Q\right] \tag{C43}$$

$$= \sum_Q\sum_\gamma \phi^{-1}(\gamma)Q\frac{f(\gamma)}{Z_f}q(Q) \tag{C44}$$

$$= \frac{1}{Z_f}\left(\sum_\gamma P(\gamma|\mathcal{D})\right)\mathbb{E}_q[Q]. \tag{C45}$$

由（C38）、（C42）和（C45），可得

$$\frac{\mathbb{E}_\pi[U]}{\mathbb{E}_\pi[V]} = \mathbb{E}_{p(\gamma|\mathcal{D})}\left[g(\gamma)\right]. \tag{C46}$$

对于给定的问题，通过令 $g(\gamma) = p(\gamma_i = 1|\gamma_{-i}, \mathcal{D})$，由（C46）可得

$$\frac{\mathbb{E}_\pi[U]}{\mathbb{E}_\pi[V]} = \texttt{PIP}(i). \tag{C47}$$

此外，我们有

$$\mathbb{E}_\pi[V] = \mathbb{E}_\pi\left[\hat{\phi}(\gamma)Q\right] \tag{C48}$$
$$= \sum_{\gamma,Q} \hat{\phi}(\gamma)Q\frac{f(\gamma)}{Z_f}q(Q) \tag{C49}$$
$$= \left(\sum_{\gamma} \hat{\phi}(\gamma)\frac{f(\gamma)}{Z_f}\right)\left(\sum_{Q} Qq(Q)\right) \tag{C50}$$
$$= \mathbb{E}_\pi[\hat{\phi}(\gamma)]\mathbb{E}_Q[Q] \tag{C51}$$
$$= \frac{S}{P}\mathbb{E}_\pi[\hat{\phi}(\gamma)]. \tag{C52}$$

因此，我们得到

$$\min\{\mathbb{E}_\pi[U], \mathbb{E}_\pi[V]\} = \mathbb{E}_\pi[V]\min\left\{1, \frac{\mathbb{E}_\pi[U]}{\mathbb{E}_\pi[V]}\right\} \tag{C53}$$
$$= \mathbb{E}_\pi[V]\min\left\{1, \texttt{PIP}(i)\right\} \tag{C54}$$
$$= \mathbb{E}_\pi[V]\texttt{PIP}(i) \tag{C55}$$
$$= \frac{S}{P}\mathbb{E}_\pi[\hat{\phi}(\gamma)]\texttt{PIP}(i). \tag{C56}$$

由（C36）、（C47）和（C56），可得

$$\mathbb{E}\left[\left|\sum_{t=1}^{T}\rho^{(t)}p(\gamma_i^{(t)} = 1|\gamma_{-i}^{(t)}, \mathcal{D}) - \texttt{PIP}(i)\right|^2\right] \leq \frac{4\varepsilon_0^2}{(1-\varepsilon_0)^2}\texttt{PIP}^2(i) + 4\frac{d\nu}{d\pi}\frac{1}{T}, \tag{C57}$$

and

$$\varepsilon_0 = \frac{P}{\texttt{PIP}(i)\mathbb{E}_\pi[\hat{\phi}(\gamma)]S}\sqrt{\frac{64e\log T}{(1-\lambda_{\gamma,Q})T}}. \tag{C58}$$

现在，注意到

$$\frac{d\nu}{d\pi}(\gamma, Q) = \frac{p_{\gamma_1,Q_1}(\gamma, Q)}{\pi(\gamma, Q)} \tag{C59}$$
$$\leq \frac{1}{\pi(\gamma, Q)} \tag{C60}$$
$$= \frac{1}{\pi(\gamma)q(Q)} \tag{C61}$$
$$\leq \frac{P}{S}\frac{1}{\min_\gamma \pi(\gamma)}. \tag{C62}$$

结合（C57）和（C62），我们有

$$\mathbb{E}\left[\left|\sum_{t=1}^{T}\rho^{(t)}p(\gamma_i^{(t)}=1|\gamma_{-i}^{(t)},\mathcal{D})-\texttt{PIP}(i)\right|^2\right]\leq\frac{4\varepsilon_0^2}{(1-\varepsilon_0)^2}\texttt{PIP}^2(i)+\frac{4P}{S}\frac{1}{\min_\gamma\pi(\gamma)T}. \tag{C63}$$

附录 D 推导 $p(\gamma_i|\mathbb{D}\ \gamma_{-i})$

注意到

$$p(\gamma_i|\mathcal{D},\gamma_{-i})=\frac{p(\gamma_i|\mathcal{D},\gamma_{-i})}{p(1-\gamma_i|\mathcal{D},\gamma_{-i})}\left(1+\frac{p(\gamma_i|\mathcal{D},\gamma_{-i})}{p(1-\gamma_i|\mathcal{D},\gamma_{-i})}\right)^{-1}. \tag{D64}$$

此外，我们有

$$\begin{aligned}\frac{p(\gamma_i=1|\mathcal{D},\gamma_{-i})}{p(\gamma_i=0|\mathcal{D},\gamma_{-i})}&=\frac{p(\gamma_i=1,\mathcal{D}|\gamma_{-i})}{p(\gamma_i=0,\mathcal{D}|\gamma_{-i})} &\text{(D65)}\\ &=\frac{p(\gamma_i=1|\gamma_{-i},X)}{p(\gamma_i=0|\gamma_{-i},X)}\frac{p(Y|\gamma_i=1,\gamma_{-i},X)}{p(Y|\gamma_i=0,\gamma_{-i},X)} &\text{(D66)}\\ &=\left(\frac{p(\gamma_i=1)}{p(\gamma_i=0)}\right)\left(\frac{p(Y|\gamma_i=1,\gamma_{-i},X)}{p(Y|\gamma_i=0,\gamma_{-i},X)}\right) &\text{(D67)}\\ &=\left(\frac{h}{1-h}\right)\left(\frac{p(Y|\gamma_i=1,\gamma_{-i},X)}{p(Y|\gamma_i=0,\gamma_{-i},X)}\right). &\text{(D68)}\end{aligned}$$

另一方面，对于任意元组 $\gamma=(\gamma_1,\gamma_2,\cdots,\gamma_P)$，使得 $\gamma_i=1$（因此 $|\gamma|\geq1$），我们有

$$p(Y|\gamma_i=1,\gamma_{-i},\beta_\gamma,\sigma_\gamma^2,X)=\frac{1}{\left(\sigma_\gamma\sqrt{2\pi}\right)^N}\exp\left(-\frac{\|Y-X_\gamma\beta_\gamma\|^2}{2\sigma_\gamma^2}\right). \tag{D69}$$

由此可得

$$\begin{aligned}&p(Y|\gamma_i=1,\gamma_{-i},X)\\ &=\int_{\beta_\gamma}\int_{\sigma_\gamma^2=0}^{\infty}\frac{1}{\left(\sigma_\gamma\sqrt{2\pi}\right)^N}\exp\left(-\frac{\|Y-X_\gamma\beta_\gamma\|^2}{2\sigma_\gamma^2}\right)p(\beta_\gamma|\gamma_i=1,\gamma_{-i})p(\sigma_\gamma^2|\gamma_i=1,\gamma_{-i})d\beta_\gamma d\sigma_\gamma^2 &\text{(D70)}\\ &=\int_{\sigma_\gamma^2=0}^{\infty}\texttt{InvGamma}\left(\frac{1}{2}\nu_0,\frac{1}{2}\nu_0\lambda_0\right)\int_{\beta_\gamma}\frac{1}{\left(\sigma_\gamma\sqrt{2\pi}\right)^N}\exp\left(-\frac{\|Y-X_\gamma\beta_\gamma\|^2}{2\sigma_\gamma^2}\right)\\ &\quad\times\frac{1}{\left(\sigma_\gamma\sqrt{2\pi\tau^{-1}}\right)^{|\gamma|}}\exp\left(-\frac{\|\beta_\gamma\|^2}{2\sigma_\gamma^2\tau^{-1}}\right)d\beta_\gamma d\sigma_\gamma^2. &\text{(D71)}\end{aligned}$$

现在，注意到

$$\begin{aligned}
&\|Y - X_\gamma\beta_\gamma\|^2 + \tau\|\beta_\gamma\|^2 \\
&\quad = (Y - X_\gamma\beta_\gamma)^T(Y - X_\gamma\beta_\gamma) + \tau\beta_\gamma^T\beta_\gamma && \text{(D72)}\\
&\quad = Y^TY - 2Y^TX_\gamma\beta_\gamma + \beta_\gamma^TX_\gamma^TX_\gamma\beta_\gamma + \tau\beta_\gamma^T\beta_\gamma && \text{(D73)}\\
&\quad = Y^TY - 2Y^TX_\gamma\beta_\gamma + \beta_\gamma^T\left(X_\gamma^TX_\gamma + \tau I\right)\beta_\gamma. && \text{(D74)}
\end{aligned}$$

现在，考虑正定矩阵$X_\gamma^TX_\gamma + \tau I$的特征值分解（奇异值分解）（注意， $\tau > 0$）：

$$X_\gamma^TX_\gamma + \tau I = U^T\Lambda U \tag{D75}$$

其中， Λ 是由 $X_\gamma^TX_\gamma + \tau I$的所有正特征值组成的对角矩阵。令

$$\begin{aligned}
\tilde{\beta}_\gamma &:= \sqrt{\Lambda}U\beta_\gamma, && \text{(D76)}\\
\tilde{Y}_\gamma &:= \sqrt{\Lambda^{-1}}UX_\gamma^TY. && \text{(D77)}
\end{aligned}$$

于是，我们有

$$\begin{aligned}
&\|Y - X_\gamma\beta_\gamma\|^2 + \tau\|\beta_\gamma\|^2 \\
&\quad = Y^TY - 2Y^TX_\gamma\beta_\gamma + \beta_\gamma^T(X_\gamma^TX_\gamma + \tau I)\beta_\gamma && \text{(D78)}\\
&\quad = Y^TY - 2Y^TX_\gamma\sqrt{\Lambda^{-1}}U^T\tilde{\beta}_\gamma + \tilde{\beta}_\gamma^T\tilde{\beta}_\gamma && \text{(D79)}\\
&\quad = Y^TY - 2\tilde{Y}_\gamma^T\tilde{\beta}_\gamma + \tilde{\beta}_\gamma^T\tilde{\beta}_\gamma && \text{(D80)}\\
&\quad = \left(\|Y\|^2 - \|\tilde{Y}_\gamma|^2\right) + \left(\tilde{Y}_\gamma^T\tilde{Y}_\gamma - 2\tilde{Y}_\gamma^T\tilde{\beta}_\gamma + \tilde{\beta}_\gamma^T\tilde{\beta}_\gamma\right) && \text{(D81)}\\
&\quad = \left(\|Y\|^2 - \|\tilde{Y}_\gamma|^2\right) + \|\tilde{Y}_\gamma - \tilde{\beta}_\gamma\|^2. && \text{(D82)}
\end{aligned}$$

因此，我们有

$$\begin{aligned}
d\beta_\gamma &= \det(U^T\Lambda^{-1/2})d\tilde{\beta}_\gamma && \text{(D83)}\\
&= \det(X_\gamma^TX_\gamma + \tau I)^{-1/2}d\tilde{\beta}_\gamma. && \text{(D84)}
\end{aligned}$$

因此，我们有

$$\begin{aligned}
&\int_{\beta_\gamma} \frac{1}{\left(\sigma_\gamma\sqrt{2\pi}\right)^N}\exp\left(-\frac{\|Y - X_\gamma\beta_\gamma\|^2}{2\sigma_\gamma^2}\right)\frac{1}{\left(\sigma_\gamma\sqrt{2\pi\tau^{-1}}\right)^{|\gamma|}}\exp\left(-\frac{\|\beta_\gamma\|^2}{2\sigma_\gamma^2\tau^{-1}}\right)d\beta_\gamma && \text{(D85)}\\
&\quad = \int_{\tilde{\beta}_\gamma} \frac{1}{\left(\sigma_\gamma\sqrt{2\pi}\right)^N}\exp\left(-\frac{\left(\|Y\|^2 - \|\tilde{Y}_\gamma|^2\right) + \|\tilde{Y}_\gamma - \tilde{\beta}_\gamma\|^2}{2\sigma_\gamma^2}\right)
\end{aligned}$$

$$\times \frac{1}{\left(\sigma_\gamma\sqrt{2\pi\tau^{-1}}\right)^{|\gamma|}} \det(X_\gamma^T X_\gamma + \tau I)^{-1/2} d\tilde{\beta}_\gamma \tag{D86}$$

$$= \frac{1}{\left(\sigma_\gamma\sqrt{2\pi}\right)^{N}} \tau^{|\gamma|/2} \exp\left(- \frac{\left(\|Y\|^2 - \|\tilde{Y}_\gamma|^2\right)}{2\sigma_\gamma^2} \right) \det(X_\gamma^T X_\gamma + \tau I)^{-1/2}. \tag{D87}$$

通过结合 (D71) 和 (D87)，我们得到

$$p(Y|\gamma_i = 1, \gamma_{-i}, X)$$

$$= \int_{\beta_\gamma} \int_{\sigma_\gamma^2=0}^{\infty} \frac{1}{\left(\sigma_\gamma\sqrt{2\pi}\right)^{N}} \exp\left(- \frac{\|Y - X_\gamma\beta_\gamma\|^2}{2\sigma_\gamma^2} \right) p(\beta_\gamma|\gamma_i = 1, \gamma_{-i}) p(\sigma_\gamma^2|\gamma_i = 1, \gamma_{-i}) d\beta_\gamma d\sigma_\gamma^2 \tag{D88}$$

$$= \int_{\sigma_\gamma^2=0}^{\infty} \texttt{InvGamma}\left(\frac{1}{2}\nu_0, \frac{1}{2}\nu_0\lambda_0\right) \frac{1}{\left(\sigma_\gamma\sqrt{2\pi}\right)^{N}} \tau^{|\gamma|/2}$$

$$\times \exp\left(- \frac{\left(\|Y\|^2 - \|\tilde{Y}_\gamma|^2\right)}{2\sigma_\gamma^2} \right) \det(X_\gamma^T X_\gamma + \tau I)^{-1/2} d\sigma_\gamma^2 \tag{D89}$$

$$= \det(X_\gamma^T X_\gamma + \tau I)^{-1/2} \tau^{|\gamma|/2} (2\pi)^{-N/2} \int_{\sigma_\gamma^2=0}^{\infty} \texttt{InvGamma}\left(\frac{1}{2}\nu_0, \frac{1}{2}\nu_0\lambda_0\right) (\sigma_\gamma^2)^{-N/2}$$

$$\times \exp\left(- \frac{\left(\|Y\|^2 - \|\tilde{Y}_\gamma\|^2\right)}{2\sigma_\gamma^2} \right) d\sigma_\gamma^2 \tag{D90}$$

$$= \det(X_\gamma^T X_\gamma + \tau I)^{-1/2} \tau^{|\gamma|/2} (2\pi)^{-N/2}$$

$$\times \int_{\sigma_\gamma^2=0}^{\infty} \frac{(1/2\lambda_0\nu_0)^{1/2\nu_0}}{\Gamma(1/2\nu_0)} (1/\sigma_\gamma^2)^{1/2\nu_0+1} \exp\left(- 1/2\nu_0\lambda_0/\sigma_\gamma^2 \right) (\sigma_\gamma^2)^{-N/2}$$

$$\times \exp\left(- \frac{\left(\|Y\|^2 - \|\tilde{Y}_\gamma\|^2\right)}{2\sigma_\gamma^2} \right) d\sigma_\gamma^2 \tag{D91}$$

$$= \det(X_\gamma^T X_\gamma + \tau I)^{-1/2} \tau^{|\gamma|/2} (2\pi)^{-N/2} \frac{(1/2\lambda_0\nu_0)^{1/2\nu_0}}{\Gamma(1/2\nu_0)}$$

$$\times \int_{\sigma_\gamma^2=0}^{\infty} (1/\sigma_\gamma^2)^{1/2\nu_0+1+N/2} \exp\left(- \frac{\left(\|Y\|^2 - \|\tilde{Y}_\gamma\|^2 + \nu_0\lambda_0\right)}{2\sigma_\gamma^2} \right) d\sigma_\gamma^2 \tag{D92}$$

$$= \det(X_\gamma^T X_\gamma + \tau I)^{-1/2} \tau^{|\gamma|/2} (2\pi)^{-N/2} \frac{(1/2\lambda_0\nu_0)^{1/2\nu_0}}{\Gamma(1/2\nu_0)}$$

$$\times \Gamma\left(\frac{N + \nu_0}{2}\right) \left(\frac{\|Y\|^2 - \|\tilde{Y}_\gamma\|^2 + \nu_0\lambda_0}{2}\right)^{-\frac{N+\nu_0}{2}}. \tag{D93}$$

令 $\tilde{\gamma}_1$ 由 γ_{-i} 给出，其中 $\gamma_i = 1$； $\tilde{\gamma}_0$ 由 γ_{-i} 给出，其中 $\gamma_i = 0$。由此可得

$$\frac{p(Y|\gamma_i = 1, \gamma_{-i}, X)}{p(Y|\gamma_i = 0, \gamma_{-i}, X)} = \sqrt{\tau}\sqrt{\frac{\det(X_{\tilde{\gamma}_0}^T X_{\tilde{\gamma}_0} + \tau I)}{\det(X_{\tilde{\gamma}_1}^T X_{\tilde{\gamma}_1} + \tau I)}}\left(\frac{\|Y\|^2 - \|\tilde{Y}_{\tilde{\gamma}_0}\|^2 + \nu_0\lambda_0}{\|Y\|^2 - \|\tilde{Y}_{\tilde{\gamma}_1}\|^2 + \nu_0\lambda_0}\right)^{\frac{N+\nu_0}{2}}. \tag{D94}$$

另一方面，我们有

$$\begin{aligned}\|\tilde{Y}_\gamma\|^2 &= \tilde{Y}_\gamma^T \tilde{Y}_\gamma &\text{(D95)}\\ &= Y^T X_\gamma (X_\gamma^T X_\gamma + \tau I)^{-1} X_\gamma^T Y. &\text{(D96)}\end{aligned}$$

因此，最终得到

$$\frac{p(Y|\gamma_i = 1, \gamma_{-i}, X)}{p(Y|\gamma_i = 0, \gamma_{-i}, X)} = \sqrt{\tau\frac{\det(X_{\tilde{\gamma}_0}^T X_{\tilde{\gamma}_0} + \tau I)}{\det(X_{\tilde{\gamma}_1}^T X_{\tilde{\gamma}_1} + \tau I)}\left(\frac{S_{\tilde{\gamma}_0}}{S_{\tilde{\gamma}_1}}\right)^{N+\nu_0}}, \tag{D97}$$

其中

$$S_\gamma := Y^T Y - Y^T X_\gamma (X_\gamma^T X_\gamma + \tau I)^{-1} X_\gamma^T Y + \nu_0\lambda_0. \tag{D98}$$

基于此，我们可以估计

$$p(\gamma_i|\mathcal{D}, \gamma_{-i}) = \frac{p(\gamma_i|\mathcal{D}, \gamma_{-i})}{p(1-\gamma_i|\mathcal{D}, \gamma_{-i})}\left(1 + \frac{p(\gamma_i|\mathcal{D}, \gamma_{-i})}{p(1-\gamma_i|\mathcal{D}, \gamma_{-i})}\right)^{-1}. \tag{D99}$$

将 $\tilde{\gamma}_0$ 中包含的可变集合记为 $I = \{j : \tilde{\gamma}_{0,\,j} = 1\}$。定义 $F = \left(X_{\tilde{\gamma}_0}^T X_{\tilde{\gamma}_0} + \tau I\right)^{-1}$，$\nu = X^T Y$ 以及 $\nu_{\tilde{\gamma}_0} = (\nu_j)_{\,j\in I}$。另定义 $A = X^T X$ 和 $a_i = (A_{ji})_{\,j\in I}$。然后，利用与 [23, 附录 B1], 中相同的论证，我们可以证明

$$S(\tilde{\gamma}_1) = S(\tilde{\gamma}_0) - d_i\left(\nu_{\tilde{\gamma}_0}^T F a_i - \nu_i\right)^2, \tag{D100}$$

其中 $d_i = (A_{ii} + \tau - a_i^T F a_i)^{-1}$。此外，我们可以通过对 $F = LL^T$ 进行 Cholesky 分解来计算 $a_i^T F a_i$，并

$$\begin{aligned}a_i^T F a_i &= \|a_i^T L\|^2 &\text{(D101)}\\ &= \sum_{j\in I}(BL)_{ij}^2, &\text{(D102)}\end{aligned}$$

其中 B 是由 A 中对应于纳入 γ的变量的列构成的 $p \times |\gamma|$ 矩阵。

此外，我们还有

$$X_{\tilde{\gamma}_1}^T X_{\tilde{\gamma}_1} + \tau I = \begin{pmatrix} X_{\tilde{\gamma}_0}^T X_{\tilde{\gamma}_0} + \tau I & a_i \\ a_i^T & A_{ii} + \tau \end{pmatrix} \tag{D103}$$

因此，利用分块矩阵行列式的舒尔公式，我们很容易看出

$$\frac{\det(X_{\tilde{\gamma}_0}^T X_{\tilde{\gamma}_0} + \tau I)}{\det(X_{\tilde{\gamma}_1}^T X_{\tilde{\gamma}_1} + \tau I)} = d_i. \tag{D104}$$

使用该算法时，如果无法预先计算 $X^T X$ ，则每次 MCMC 迭代的计算复杂度为 $O(N|\gamma|^2 + |\gamma|^3 + P|\gamma|^2)$。否则，如果可以预先计算 $X^T X$ ，则每次 MCMC 迭代的计算复杂度为 $O(|\gamma|^3 + P|\gamma|^2)$。

参考文献

[1] C. Andrieu、N. de Freitas、A. Doucet 和 M. I. Jordan。《机器学习的马尔可夫链蒙特卡洛方法简介》。《机器学习》，50:5–43，2004 年。

[2] C. M. Bishop.《模式识别与机器学习》。Springer，2006 年。[3] W. M. Bolstad.《理解计算贝叶斯统计学》。John Wiley，2010 年。[4] L. Breiman. 一类马尔可夫链的强大数定律。《数理统计年刊》，31:801–803，1960 年。[5] M. F. Bugallo、S. Xu 和 P. M. Djuri´c. 用于机动目标的 EKF 与粒子滤波方法性能比较。《数字信号处理》，17:774–786，2007 年。[6] R. Combes 和 M. Touati. 马尔可夫链谱隙的高效计算估计。《ACM 计算系统测量与分析论文集》，3:1–21，2019 年。[7] P. Diaconis 和 L. Saloff-Coste. 可逆马尔可夫链的比较定理。《应用概率年刊》，3:696–730，1993 年。[8] W. J. Fitzgerald. 应用于信号处理的马尔可夫链蒙特卡罗方法。《信号处理》，81:3–18，2001 年。[9] A. Gupta 和 J. B. Rawlings. 随机化学动力学模型中的参数估计方法比较：系统生物学中的示例。《AIChE 期刊》，美国化学工程师学会，60 4:1253–1268，2014 年。[10] T. Hesterberg. 科学计算中的蒙特卡罗策略。《计量学》，44:403–404，2002 年。[11]M. Jankowiak. 百万维空间中的贝叶斯变量选择。ArXiv，abs/2208.01180，2022 年。

[12] M. F. Kasim、A. F. A. Bott、P. Tzeferacos、D. Q. Lamb、G. Gregori 和 S. M. Vinko。无需源剖面即可从质子射线照相中重建场。Physical Review. E，100 3-1:033208，2019。[13] F. Liang、C. Liu 和 R. J. Carroll。高级马尔可夫链蒙特卡罗方法：从过去的样本中学习。2010。[14] D. Paulin。通过 Marton 耦合和谱方法得到的马尔可夫链集中不等式。Electronic Journal of Probability，20(79):1 – 32，2015。[15] S. Rao。马尔可夫链的 Hoeffding 不等式。Electronic Communications in Probability，2018。[16] J. Read、L. Martino 和 D. Luengo。用于分类器链多维学习的高效蒙特卡罗方法。Pattern Recognit.，47:1535–1546，2012。[17] C. P. Robert 和 G. Casella。蒙特卡罗统计方法。Technometrics，47:243 – 243，2005。[18] L. V. Truong。具有马尔可夫或隐马尔可夫信号先验的线性模型的副本分析。ArXiv，abs/2009.13370，2020。[19] L. V. Truong。通过副本方法处理具有隐马尔可夫源的线性模型。2021 IEEE International Symposium on Information Theory (ISIT)，第 396–401 页，2021。[20] L. V. Truong。关于具有马尔可夫信号先验的线性模型。载于 AISTATS，2022。[21] P. Tuominen 和 R. L. Tweedie。具有连续分量的马尔可夫链。Proceedings of the London Mathematical Society，s3-38(1):89–114，1979 年 1 月。[22] G. Wolfer 和 A. Kontorovich。估计遍历马尔可夫链的混合时间。载于第 32 届年度学习理论会议，2019。[23] G. Zanella 和 G. O. Roberts。可扩展的重要性调温与贝叶斯变量选择。Journal of the Royal Statistical Society: Series B (Statistical Methodology)，81，2019。