arXiv:2601.01442v1 [stat.ML] 4 Jan 2026

缺失观测下贝叶斯隐马尔可夫模型的快速吉布斯抽样

Li, Dongrong统计与数据科学系，香港中文大学以及 Yu, Tianwei数据科学学院，香港中文大学（深圳）（CUHK-Shenzhen）以及 Fan, Xiaodan*统计与数据科学系，香港中文大学

2026年1月6日

摘要

隐马尔可夫模型（HMM）是一种广泛用于处理序列数据的统计模型。然而，真实世界数据集中的缺失观测往往使模型的应用变得复杂。EM算法和吉布斯采样器可用于估计模型，但会面临非凸性、计算复杂度高以及混合速度慢等各种问题。本文提出了一种折叠吉布斯采样器，通过同时积分掉缺失观测及其对应的潜在状态，高效地从HMM的后验分布中进行采样。该采样器凭借三个优势实现了快速采样。首先，其估计精度与现有方法相当。其次，它能够在每次迭代中产生更大的有效样本量（ESS），这一点在理论和数值上均得到了验证。第三，当缺失条目数量较多时，

*通讯联系：xfan@cuhk.edu.hk

与其他方法相比，该采样器每次迭代的计算复杂度显著更低，因此在计算上更快。总之，所提出的采样算法在计算和理论方面都具有较高速度，尤其适用于缺失条目较多的情况。最后，基于数值模拟和真实数据分析的实证评估表明，在时间复杂度和采样效率（以有效样本量（ESS）衡量）方面，所提出的算法始终优于现有算法。

关键词：序贯建模、马尔可夫链蒙特卡罗（MCMC）、贝叶斯推断、缺失数据、可扩展计算

1 引言

隐马尔可夫模型（HMM）是一种用于序列数据分析的强大统计工具。它假设我们观测到一系列输出$\mathbf{y}$，这些输出由未观测到的潜变量$\mathbf{z}$决定，而潜变量构成一个马尔可夫链。由于它可以被视为一种对观测值进行建模的方法，这些观测值源自未观测到的潜在序列经过变换或破坏后的结果，因此它已广泛应用于与序列分析相关的多个领域，例如DNA 测序（Yoon（2009））、文本分析（Perikos等（2019））、语音识别（Rabiner（1989年）；Bahl等（1983））、音乐分析（Yanchenko和Mukherjee（2017））以及医疗健康记录分析（Kawamoto等（2013））。尽管已有大量文献探讨隐马尔可夫模型中的估计与预测问题（Rabiner（1989年）；Bishop和Nasrabadi（2006）），但相对而言，关注含缺失观测的 HMM 的研究较少。此类缺失观测常见于医疗与健康时间序列，例如电子健康记录（EHR）（Kawamoto等（2013）；Yeh 等人（2012年））；缺失条目源于患者就诊时间不规律或医院记录不规律。

尽管已有多种方法被提出以解决这一问题（我们将在后文介绍），但它们通常具有较高的计算复杂度。本文提出了一种计算高效的吉布斯抽样方法，用于对存在缺失观测的隐马尔可夫模型进行后验推断。所提出的采样器采用了一种新颖的前向后向算法，以解析地边缘化缺失数据及其对应的潜在状态，从而在理论收敛速率和时间复杂度方面提高后验采样的效率。此外，我们证明，在高缺失率场景下，由于该算法具有显著更快的收敛速度，因此表现出尤为明显的优势。

为了提供一个 为后续讨论奠定基础，我们首先回顾相关方法

和这一领域的研究思路。针对不完整序列的马尔可夫建模最早见于 Deltour 等人（1999年）、Albert（2000年）和 Yeh 等人（2010年）的开创性研究。然而，这些方法无法直接应用于隐马尔可夫模型。进入 21世纪以来，针对使用类隐马尔可夫模型对不完整序列进行建模的各种方法陆续涌现。

Yu 和 Kobayashi（2003）研究了具有分类结果的隐半马尔可夫模型，以应对缺失数据问题。Chassan 和 Concordet（2023）系统研究了隐马尔可夫模型中的不同缺失机制，并提出了一种统计检验来区分这些机制。Cooke 等人（2001年）探讨了如何使用众数插补等简单插补策略，对不完整序列中的 HMM 进行估计。然而，这种方法仍相对缺乏系统性，因为对于存在缺失数据的 HMM，众数插补/前向插补的性能尚无理论保证。

Lee 等人（2008年）更系统地探讨了因子隐马尔可夫模型（Ghahramani 和 Jordan（1995年））中的缺失值插补问题。不过，这项工作主要关注在真实参数已知的模型下插补缺失观测值。

Yeh 等人（2012年）和 Speekenbrink 和 Visser（2021）正式研究了存在缺失观测的隐马尔可夫模型。在 Yeh 等人（2012年）的研究中，他们提出了一种用于参数估计的期望最大化算法，并通过模拟实验评估了该算法在不同缺失机制下的表现。Speekenbrink 和 Visser（2021）考察了几种不可忽略缺失机制（Little 和 Rubin（2019年）），并讨论了这些框架下的参数估计。

这两项研究都针对缺失观测已通过解析积分消去的似然函数，采用期望最大化算法（Ng 等人（2012）；Dempster 等人（1977））进行

参数估计，并通过前向-后向概率计算进行迭代更新（Rabiner（1989））。由于期望最大化算法和吉布斯抽样都属于坐标下降法（Dias和Wedel（2004）），因此可以推导出一个针对相同目标函数的吉布斯采样器（Robert等人（2010）；Gelfand and Smith（1990）），作为 Yu 和 Kobayashi（2003）、Yeh 等人（2012）及 Speekenbrink 和 Visser（2021）所述方法的随机化变体。这种随机化形式表明，它不易陷入参数空间中的局部极小值或平坦区域（Dias和Wedel（2004））。Ryd´en（2008）对隐马尔可夫模型的贝叶斯吉布斯抽样作了全面介绍。

上述所有研究均采用迭代算法，通过前向-后向概率计算进行参数更新；对于平均长度为 T 的 n 个序列，其计算复杂度为 $O(nT)$。正如本文所展示的，我们提出了一种新颖的吉布斯抽样算法，随着缺失条目数量的增加，该算法可显著降低计算时间复杂度，从而实现更高效的后验估计。所提出的采样器通过对多余的潜在状态进行边缘化，因而在缩减后的参数空间中运行，并获得更快的理论收敛速率。

除了离散时间类别型隐马尔可夫模型之外，大量研究也致力于使用连续时间隐马尔可夫模型对部分观测序列进行建模（Liu等人（2015年）；Lu（2017年）；James等人（1996年））。然而，将连续时间隐马尔可夫模型应用于离散时间估计场景存在固有局限。这一挑战源于连续时间框架要求估计停留时间；在分析离散时间序列时，这一过程会给参数估计和状态预测带来额外的不确定性。

基于现有文献，我们将本文的贡献总结如下：

首先，我们通过对缺失观测及其对应的潜在状态进行解析边缘化，提出了一种加速吉布斯采样器。我们基于优化后的前向后向算法，提出了一种新颖的折叠后验采样方法。其次，我们通过严格的复杂度分析和收敛速率刻画，确立了所提算法的计算效率。此外，我们证明，该方法在高缺失率情形下具有显著优势：处理缺失观测占比较大的数据集时，其运行速度明显快于现有方法。最后，我们通过全面的数值实验验证了这些结论，结果表明，该方法在计算速度和有效样本量（ESS）方面均表现出色。

本文其余部分安排如下：首先回顾隐马尔可夫模型和缺失数据机制的基本概念。接着详细介绍所提出的算法，并对计算复杂度和收敛速率进行理论分析，突出其相较于现有方法的优势。最后，通过模拟研究和真实世界数据实验验证该算法的有效性，并对运行效率和采样性能进行定量比较。

1.1 隐马尔可夫模型

设$\mathbf{y} = (y_1, ..., y_T)$为长度为 T的观测序列。参数为 $\boldsymbol{\theta}$ 的隐马尔可夫模型是一种生成模型，它假设$\mathbf{y}$ 是通过以下过程生成的：

$$p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}) = p(\mathbf{z}|\boldsymbol{\theta})p(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}),$$

其中$\mathbf{z}$ 是由未观测到的潜在状态组成的序列。在 HMM 框架中， $p(\mathbf{z}|\boldsymbol{\theta})$通常定义为马尔可夫链：

$$p(\mathbf{z}|\boldsymbol{\theta}) = p(z_1|\boldsymbol{\theta}) \prod_{t=1}^{T-1} p(z_{t+1}|z_t, \boldsymbol{\theta}),$$

在给定潜在状态的条件下，观测值彼此独立：

$$p(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) = \prod_{t=1}^{T} p(y_t|z_t, \boldsymbol{\theta}).$$

当 z_t 和 y_t 定义在有限状态空间上时，标准参数化形式为 $\boldsymbol{\theta} = (\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$，其中 $\boldsymbol{\pi}$ 表示初始状态分布，$\mathbf{A}$ 是状态转移矩阵，而 $\mathbf{B}$ 包含发射概率， $B_{ij} = P(y = j|z = i)$对应有效状态 i 和 j。

从贝叶斯视角来看，层次化表述变为：

$$p(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta}) = p(\boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})p(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}),$$

其中 $p(\boldsymbol{\theta})$表示模型参数的先验分布。

在隐马尔可夫建模中，我们主要关注模型参数和潜在序列的估计（Rabiner（1989））。一种常见方法是将$\mathbf{z}$ 视为潜在变量，并应用期望最大化算法（Dempster 等人（1977年）；Ng 等人（2012年））来获得最大似然估计。对于潜在序列估计，通常采用最大后验（Bishop和Nasrabadi（2006））或边缘化最大后验估计量。维特比算法（Forney（1973））提供了一种有效的动态规划方法，用于进行序列级最大后验估计（Bellman（1966））。另一种方法是采用带有数据增广的吉布斯抽样（Liu和Wu（1999）），联合估计参数和潜在状态；Ryd´en（2008年）详细介绍了具体实现。由于吉布斯抽样可被视为随机坐标下降的一种变体，这两种方法在理论上表现相近（Dias和Wedel（2004））。

基于这些基础，研究者开发了基于吉布斯抽样的马尔可夫链蒙特卡罗方法（Ryd´en（2008年）），通过交替更新参数和潜在状态，执行随机坐标下降过程。实证证据表明，马尔可夫链蒙特卡罗算法能够跳出局部众数，其收敛行为优于期望最大化算法（Ryd´en（2008年））。其他计算方法还包括变分推断（Blei 等人（2017年）；Foti 等人（2014）），但这种方法通过变分分布近似后验分布，可能引入估计偏差和不稳定性。

上述所有算法都需要计算前向概率 $\alpha_t(i) = P(y_1、\ldots\ldots、y_t,\ z_t = i \mid \boldsymbol{\theta})$ 和后向概率 $\beta_t(i) = P(y_{t+1}、\ldots\ldots、y_T \mid z_t = i,\ \boldsymbol{\theta})$，以便对 $t=1$、……、T 执行迭代更新。由于这些概率没有封闭形式的解，必须通过动态规划递归计算，通常每次迭代需要 $O(T)$ 次运算。Rabiner（1989年）详细介绍了它们的递归计算方法及其在参数更新中的应用。

在接下来的章节中，我们将提出一种折叠吉布斯采样器。在这种新的采样方案下，我们采用了一种新的策略来计算前向概率和后向概率；当存在缺失观测时，该策略显著快于现有算法。

1.2 符号表示

在下一节中，我们将开始正式分析。为便于理解，我们简要概述本文中常用的符号表示。

来自特定集合的单个数值或元素用小写字母表示，例如 z 或 y。元素的向量或序列用粗体英文字母表示，例如$\mathbf{z}$ 或$\mathbf{y}$。$\mathbf{z}_{t_1:t_2}$ 用于表示序列$\mathbf{z}$中从位置 t_1 到位置 t_2 （包括两端）的值。我们也使用希腊字母（如 π）表示向量或向量集合。我们特别指出，θ 表示所有参数的集合

。对于矩阵或矩阵集合，我们使用粗体字母，例如$\mathbf{A}$ 或$\mathbf{B}$。记号 $\mathbf{\Delta}^k$ 表示定义在（$k-1$）维单纯形中的向量，而 $\mathbf{\Delta}^{r\times k}$ 表示一个 $r \times k$ 矩阵，其每一行都是定义在（$k-1$）维单纯形中的向量。

集合用花体字母或大写字母表示（ $\mathbf{\Delta}$除外），例如 $\mathcal{Z}$， $\mathcal{Y}$ 和 Ω。集合 $\mathcal{Z}$ 的基数表示为 $|\mathcal{Z}|$。我们用 Ω 表示样本空间，用 P 或 μ 表示特定的概率测度。因此，按照测度 μ 对函数 f 积分，表示为 $\int_\Omega f(x)d\mu(x)$。

我们用 T 表示序列长度，并用 n 表示整个数据集的样本量。我们用 p 表示缺失概率（稍后将给出定义）。

2 方法

2.1 问题表述

本节将对不完整序列的隐马尔可夫模型贝叶斯估计问题进行形式化表述。我们系统地回顾现有方法，并提出一种计算效率更高的新算法。

令 $\mathbf{y}$ 表示理论完整观测序列。在我们的框架中，假设 $\mathbf{y}$ 仅部分可观测，即 $\mathbf{y} = (\mathbf{y}_o, \mathbf{y}_m)$，其中 $\mathbf{y}_o$ 表示观测分量，$\mathbf{y}_m$ 表示缺失分量。在整个分析过程中，我们采用可忽略缺失假设（Rubin（1976））。

在这种表述下，联合分布可分解为:

$$p(\boldsymbol{\theta}, \mathbf{y}, \mathbf{z}) = p(\boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})p(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) = p(\boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})p(\mathbf{y}_o|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{y}_m|\mathbf{z}, \boldsymbol{\theta}),$$

其中最后的 e　　质量源于其内在的条件独立结构　　n

隐马尔可夫模型。

一种直接进行后验推断的方法涉及数据增强（Gelfand 和 Smith（1990）），需要从以下条件分布中进行迭代采样（Ryd´en（2008））：

$$\begin{aligned} \mathbf{y}_m|\mathbf{z},\boldsymbol{\theta},\\ \boldsymbol{\theta}|\mathbf{y}_o,\mathbf{z},\\ \mathbf{z}|\boldsymbol{\theta},\mathbf{y}_o. \end{aligned} \tag{1}$$

然而，这种方法在计算上效率低下。已有研究通过将缺失值编码为发射概率固定为 1 的特殊标记，提出了一种替代方法（Yeh 等人（2012）；Speekenbrink 和 Visser（2021））。这种公式在数学上等价于从完整数据似然中对 $\mathbf{y}_m$ 进行解析边缘化。

在此框架下，可以递归计算前向概率$\alpha_t(i) = P(\mathbf{y}_{1:t}, z_t = i \mid \boldsymbol{\theta})$ 和后向概率 $\beta_t(i) = P(\mathbf{y}_{t+1:T} \mid z_t = i, \boldsymbol{\theta})$，从而实现迭代参数更新。已有文献（Yeh 等人（2012））严格推导了这些概率递推公式及其相应的 EM算法实现。

鉴于 EM算法和吉布斯采样均属于协调下降变体，我们推导出一种相应的吉布斯采样器，其性能与之相当。该方法通过从以下针对积分联合分布的条件分布中迭代采样来实现：

$$\begin{aligned} \boldsymbol{\theta}|\mathbf{y}_o,\mathbf{z},\\ \mathbf{z}|\boldsymbol{\theta},\mathbf{y}_o. \end{aligned} \tag{2}$$

从潜在状态条件分布中采样$\mathbf{z}|\boldsymbol{\theta},\mathbf{y}_o$ 需要递归计算

前向-后向概率。完整的计算过程记录于 Ryd´en (2008年)。

然而，不难发现，上述方法需要对每个位置 t计算前向和后向概率，这意味着每次更新的计算复杂度为 $O(nT)$（其中 n 表示样本量）。当序列较长时，该算法会变得很慢。此外，将缺失率（或缺失概率）定义为

$$p = \frac{|\mathbf{y}_m|}{|\mathbf{y}_m| + |\mathbf{y}_o|}, \tag{3}$$

表示缺失观测所占的比例。当缺失率很高时，该算法会变得不必要地复杂。为此，我们提出了一种后验抽样算法，其每次迭代的时间复杂度为 $O((1-p)nT)$。这意味着，当缺失率 p 较高时，所提出的算法将明显快于文献 1 中所述的方法。此外，与所提出的方法相比，文献 1 中的方法也具有更大的探索空间，这会降低吉布斯抽样的收敛速度。正如我们将在后续章节中论证的那样，在存在缺失数据时，可以显著缩小该参数空间，从而提高收敛速度。因此，我们提出的算法不仅在计算复杂度方面具有优势，在收敛率方面也同样如此。

2.2 折叠模型的描述

本节通过对冗余潜在状态进行解析边缘化，推导折叠后的联合分布，并将在后续讨论后验抽样技术。

我们的推导从完整数据似然开始:

$$p(\boldsymbol{\theta}, \mathbf{y}, \mathbf{z}) = p(\boldsymbol{\theta}) p(\mathbf{z}|\boldsymbol{\theta}) p(\mathbf{y}_o|\boldsymbol{\theta}, \mathbf{z}) p(\mathbf{y}_m|\boldsymbol{\theta}, \mathbf{z}),$$

其中，等式成立是因为条件独立结构。

通过对缺失观测进行积分，我们得到：

$$p(\boldsymbol{\theta}, \mathbf{y}_m, \mathbf{z}) = \int p(\boldsymbol{\theta}) p(\mathbf{z}|\boldsymbol{\theta}) p(\mathbf{y}_o|\boldsymbol{\theta}, \mathbf{z}) p(\mathbf{y}_m|\boldsymbol{\theta}, \mathbf{z}) d\mu(\mathbf{y}_m).$$

这一表述构成了 Yeh 等人（2012）；Speekenbrink 和 Visser（2021）所用方法的基础。

关键的是，我们证明了还可以通过将潜在状态分解为 $\mathbf{z} = (\mathbf{z}_m, \mathbf{z}_o)$，进一步降低复杂度。其中，$\mathbf{z}_m$ 对应缺失观测$\mathbf{y}_m$，而$\mathbf{z}_o$ 对应观测数据 $\mathbf{y}_o$。

由此得到改进后的联合分布：

$$p(\boldsymbol{\theta}, \mathbf{y}_o, \mathbf{z}) = p(\boldsymbol{\theta})\, p(\mathbf{z}_m, \mathbf{z}_o|\boldsymbol{\theta})\, p(\mathbf{y}_o|\mathbf{z}_o, \boldsymbol{\theta}).$$

因此，我们可以进一步对对应于缺失观测的潜在状态进行积分（即 $\mathbf{z}_m$ 外：

$$\begin{aligned} p(\boldsymbol{\theta}, \mathbf{y}_o, \mathbf{z}_o) &= \int p(\boldsymbol{\theta}, \mathbf{y}_o, \mathbf{z}_o, \mathbf{z}_m)\, d\mu(\mathbf{z}_m) \\ &= p(\boldsymbol{\theta}) \int p(\mathbf{z}_o, \mathbf{z}_m|\boldsymbol{\theta})\, p(\mathbf{y}_o|\mathbf{z}_o, \boldsymbol{\theta})\, d\mu(\mathbf{z}_m) \\ &= p(\boldsymbol{\theta})\, p(\mathbf{y}_o|\mathbf{z}_o, \boldsymbol{\theta}) \int p(\mathbf{z}_o, \mathbf{z}_m|\boldsymbol{\theta})\, d\mu(\mathbf{z}_m). \end{aligned}$$

不失一般性，假设观测到了 y_1，因此 $z_1 \in \mathbf{z}_o$。进一步的

计算结果表明:

$$\begin{aligned}\int p(\mathbf{z}_o, \mathbf{z}_m|\boldsymbol{\theta})d\mu(\mathbf{z}_m) &= \int p(z_1|\boldsymbol{\theta})[\prod_{t=1}^{T-1} p(z_t|z_{t-1}, \boldsymbol{\theta})]d\mu(\mathbf{z}_m) \\ &= \int p(z_1|\boldsymbol{\theta})[\prod_{t=1}^{T-1} p(z_t|z_{t-1}, \boldsymbol{\theta})]dz_{t_1}...dz_{t_k} \\ &= p(z_1|\boldsymbol{\theta}) \prod_{t_i \in o} p(z_{t_{i+1}}|z_{t_i}, \boldsymbol{\theta}) \\ &= p(z_1|\boldsymbol{\theta}) \prod_{t_i \in o} \mathbf{1}_{z_{t_{i+1}}}^T \mathbf{A}^{t_{i+1}-t_i} \mathbf{1}_{z_{t_i}},\end{aligned}$$

其中， z_{t_k} 表示$\mathbf{z}_m$中的潜在状态， o 表示与观测状态对应的索引，$\mathbf{A}$ 表示转移矩阵。上述计算表明，与缺失观测对应的潜在状态实际上是冗余的，可以通过解析方法将其积分消去。

这一塌缩分布使得吉布斯采样器能够根据以下条件分布进行迭代更新:

$$\begin{aligned}&\boldsymbol{\theta}|\mathbf{y}_o, \mathbf{z}_o, \\ &\mathbf{z}_o|\boldsymbol{\theta}, \mathbf{y}_o.\end{aligned} \tag{4}$$

这些条件分布降低了潜在状态的参数空间维数，从而加快收敛速率（详见后续章节）。此外，前向/后向概率计算和$\mathbf{z}_o$ 采样的时间复杂度均有所改善，尤其适用于缺失概率较高的区间（$p \to 1$）。

2.3 后验采样

在本节中，我们提出一个以塌缩分布 $p(\boldsymbol{\theta}, \mathbf{y}_o, \mathbf{z}_o)$ 为目标的吉布斯采样器。我们特别关注$\mathbf{z}_o$ 状态采样，与传统方法相比，它具有更低的时间复杂度。

2.3.1 从 $\mathbf{z}_o$ 的条件分布中采样

首先，我们注意到$\mathbf{z}_o$ 构成一条非齐次马尔可夫链，其中 z_{t_k} 与 $z_{t_{k+1}}$ 之间的转移概率（对 z_{t_k}，$z_{t_{k+1}} \in \mathbf{z}_o$）由$\mathbf{A}^{t_{k+1}-t_k}$参数化。前向概率 $\alpha_{t_k}(z_{t_k}) := p((\mathbf{y}_o)_{1:t_k}, z_{t_k}|\boldsymbol{\theta})$ 可递归计算如下：

$$\begin{aligned}
\alpha_{t_k}(z_{t_k}) &= p((\mathbf{y}_o)_{1:t_k}, z_{t_k}|\boldsymbol{\theta}) \\
&= \sum_{z_{t_{k-1}}} p(z_{t_k}, z_{t_{k-1}}, (\mathbf{y}_o)_{1:t_k-1}|\boldsymbol{\theta}) \\
&= \sum_{z_{t_{k-1}}} p(z_{t_k}, z_{t_{k-1}}, (\mathbf{y}_o)_{1:t_k}|\boldsymbol{\theta}) \\
&= p(y_{t_k}|z_{t_k}, \boldsymbol{\theta}) \sum_{z_{t_{k-1}}} \alpha_{t_{k-1}}(z_{t_{k-1}}) p(z_{t_k}|z_{t_{k-1}}, \boldsymbol{\theta}) \\
&= p(y_{t_k}|z_{t_k}, \boldsymbol{\theta}) \sum_{z_{t_{k-1}}} \alpha_{t_{k-1}}(z_{t_{k-1}}) \mathbf{1}_{z_{t_k}}^T \mathbf{A}^{t_k - t_{k-1}} \mathbf{1}_{z_{t_{k-1}}}.
\end{aligned} \tag{5}$$

计算 $\mathbf{z}_o$ 所有状态的前向概率需要 $O((1-p)NT)$ 次运算，其中包含 N 条长度为 T 的链，缺失率为 p。尽管涉及矩阵指数运算，但借助数值线性代数库可以高效执行这些运算。此外，这些矩阵指数的结果可以预先计算并缓存。因此，后续每次矩阵指数求值只需通过缓存结果耗费 $O(1)$ 的时间复杂度。在实际应用中，当样本量 N 较大时，矩阵运算带来的计算开销变得可以忽略不计。

计算前向概率有助于我们向后采样潜在状态 $\mathbf{z}_o$ 。设 $o = \{t_1, ..., t_K\}$，$\mathbf{z}_o$ 的条件分布具有如下向后分解：

$$p(\mathbf{z}_o|\mathbf{y}_o, \boldsymbol{\theta}) = p(z_{t_K}|\mathbf{y}_o, \boldsymbol{\theta}) \prod_{t_k \in o, k \neq K} p(z_{t_k}|\mathbf{z}_{t_{k+1}:t_K}, \mathbf{y}_o, \boldsymbol{\theta}).$$

对于第一项，很容易看出 $p(z_{t_K}|\mathbf{y}_o, \boldsymbol{\theta}) \propto \alpha_{t_k}(z_{t_k})$。

对于第二项，存在如下分解：

$$
\begin{aligned}
p(z_{t_k}|\mathbf{z}_{t_{k+1}:t_K}, \mathbf{y}_o, \boldsymbol{\theta}) &= p(z_{t_k}|z_{t_{k+1}}, (\mathbf{y}_o)_{1:t_k}, (\mathbf{y}_o)_{t_{k+1}:t_K}, \boldsymbol{\theta}) \\
&= p(z_{t_k}|z_{t_{k+1}}, (\mathbf{y}_o)_{1:t_k}, \boldsymbol{\theta}) \\
&\propto \frac{p(z_{t_{k+1}}|z_{t_k}, \boldsymbol{\theta}) p(z_{t_k}|\mathbf{y}_o)_{1:t_k}, \boldsymbol{\theta})}{\sum_{z\in\mathcal{Z}} p(z_{t_{k+1}}|z_{t_k}=z, \boldsymbol{\theta}) p(z_{t_k}=z|\mathbf{y}_o)_{1:t_k}, \boldsymbol{\theta})} \\
&= \frac{p(z_{t_{k+1}}|z_{t_k}, \boldsymbol{\theta}) p(z_{t_k}|\mathbf{y}_o)_{1:t_k}, \boldsymbol{\theta})}{\sum_{z\in\mathcal{Z}} p(z_{t_{k+1}}|z_{t_k}=z, \boldsymbol{\theta}) \alpha_{t_k}(z)} \\
&= \frac{(\mathbf{1}_{z_{t_{k+1}}}^T \mathbf{A}^{t_{k+1}-t_k} \mathbf{1}_{z_{t_k}}) \alpha_{t_k}(z_{t_k})}{\sum_{z\in\mathcal{Z}} (\mathbf{1}_{z_{t_{k+1}}}^T \mathbf{A}^{t_{k+1}-t_k} \mathbf{1}_z) \alpha_{t_k}(z)},
\end{aligned}
\tag{6}
$$

这为我们提供了一种明确的方法，用于计算观测索引 o 中每个潜在状态的条件概率。利用这一分解，我们可以先对 $t_k \in o$ 进行前向计算$\boldsymbol{\alpha}_{t_k}$，然后依次进行后向采样 z_{t_K}, $z_{t_{K-1}}$..... z_{t_1}。与前向递推过程类似，后向采样方法的时间复杂度也是 $O((1-p)NT)$，矩阵乘法的计算成本可以忽略不计。采样算法的完整描述见算法 1。

2.3.2 参数的条件分布采样

本节讨论参数的条件分布采样问题，包括转移矩阵$\mathbf{A}$、发射矩阵$\mathbf{B}$ 和初始分布 $\boldsymbol{\pi}$。

对于发射矩阵$\mathbf{B}$，我们可以依次对其第 i 行进行采样。如果$\mathbf{B}$ 服从狄利克雷先验 $Dir(\boldsymbol{\eta})$，那么它的条件分布为：

$$
\mathbf{B}_i|\mathbf{B}_{-i}, \mathbf{y}_o, \mathbf{z}_o, \mathbf{A}, \boldsymbol{\pi} \sim Dir(\eta_{ij} + n_{ij}),
$$

其中 n_{ij} 表示　隐状态为 i 且观测状态为 j的观测值数量。

算法 1 折叠吉布斯采样器的前向-后向采样 1: 输入：观测值 $\mathbf{y}_o$、转移矩阵 A、其他参数 $\boldsymbol{\theta}$（包括发射分布和初始概率） 2: 输出：隐状态 $\mathbf{z}_o$的样本 3: 前向递推 4: 对 $k = 1$ 到 K 执行 5: 对每个状态 z_t 执行 6: $\alpha_t(z_t) = p(y_{t_k}|z_{t_k}, \boldsymbol{\theta}) \sum_{z_{t_{k-1}}} \alpha_{t_{k-1}}(z_{t_{k-1}}) \mathbf{1}^T_{z_{t_{k+1}}} \mathbf{A}^{t_{k+1}-t_k} \mathbf{1}_{z_{t_k}}$ 7: 结束循环 8: 结束循环 9: 后向采样 10: 对 $t = K$ 递减至 1 执行 11: $p(z_{t_k}|\mathbf{z}_{t_{k+1}:t_K}, \mathbf{y}_o, \boldsymbol{\theta}) \propto (\mathbf{1}^T_{z_{t_{k+1}}} \mathbf{A}^{t_{k+1}-t_k} \mathbf{1}_z) \alpha_{t_k}(z_{t_k}) \sum_{z \in \mathcal{Z}} (\mathbf{1}^T_{z_{t_{k+1}}} \mathbf{A}^{t_{k+1}-t_k} \mathbf{1}_z) \alpha_{t_k}(z)$ 12: 从 $p(z_{t_k}|\mathbf{z}_{t_{k+1}:t_K}, \mathbf{y}_o, \boldsymbol{\theta})$ 中采样 z_{t_k} 13: 结束循环 14: 返回 $\mathbf{z}_{1:T}$

对于参数$\mathbf{A}$ 和 $\boldsymbol{\pi}$，其条件分布无法直接进行解析求解。不过，可以采用Metropolis-within-Gibbs方案（Gilks等人（1995年））更新参数，提议分布可以选用随机游走（Gelman等人（1997年））或基于梯度的更新（Patterson和Teh（2013年））。

最后，算法2总结了完整的折叠吉布斯采样器，它交替对参数和潜在序列$\mathbf{z}_o$ 进行采样，并通过算法 1中的前向后向过程完成。

算法2针对具有不完整观测值的HMM的折叠吉布斯抽样

要求： 参数 $\boldsymbol{\theta}$的初始估计值 $\boldsymbol{\theta}_0$ ，从后验分布中抽取的样本数 N

$i \leftarrow 1$

当 $i \leq N$ **时**

从条件分布 $p(\boldsymbol{\theta}|\mathbf{z}_o, \mathbf{y}_o)$中采样 $\boldsymbol{\theta}_k$

$k \leftarrow 1$

当 $k \leq n$ **时**

使用算法 1 对潜在序列（$\mathbf{z}_k$）$_o$ 进行采样

$k \leftarrow k + 1$

结束循环

$i \leftarrow i + 1$

结束循环

丢弃预热期产生的样本，并将剩余样本作为抽样结果保留

，这些样本来自后验分布。

2.4 预测分布

所提出的方法要求对潜在状态进行边缘化，因此需要针对潜在状态和缺失值设计高效的预测程序。本节构建计算框架,

以便通过三项主要任务，在我们的抽样方案下获得预测分布：未来状态预测、潜在序列重构和缺失观测值插补。每项预测任务都需要通过前向-后向递推显式采样潜在轨迹$\mathbf{z}$ 。

预测 注意到，$\mathbf{z}_k$ 本身是一个长度为 T 的马尔可夫链。我们当然可以讨论它在时间窗口 $T+1:\ T+W$ 内的未来行为。例如，在医疗记录分析中，根据患者的不完整观测序列 $\mathbf{y}_o$ 预测其健康状况的未来行为和进展模式是很有意义的。假设我们有兴趣抽取序列 $\tilde{\mathbf{z}} = (z_1, \ldots, z_T, z_{T+1}, \ldots, z_{T+W}) = (\mathbf{z}, \mathbf{z}^W)$。$\tilde{\mathbf{z}}$ 的预测分布可通过以下计算得到：

$$
\begin{aligned}
p(\tilde{\mathbf{z}}|\mathbf{y}_o) &= p(\mathbf{z}, \mathbf{z}^W|\mathbf{y}_o) \\
&= \int_\Omega p(\mathbf{z}, \mathbf{z}^W|\boldsymbol{\theta}, \mathbf{y}_o) p(\boldsymbol{\theta}|\mathbf{y}_o) d\mu(\boldsymbol{\theta}) \\
&= \int_\Omega p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y}_o) p(\mathbf{z}^W|\boldsymbol{\theta}, \mathbf{z}, \mathbf{y}_o) p(\boldsymbol{\theta}|\mathbf{y}_o) d\mu(\boldsymbol{\theta}) \\
&= \int_\Omega p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y}_o) p(\mathbf{z}^W|\boldsymbol{\theta}, \mathbf{z}) p(\boldsymbol{\theta}|\mathbf{y}_o) d\mu(\boldsymbol{\theta}).
\end{aligned}
$$

在计算上，我们首先从其边缘后验分布中抽取 $\boldsymbol{\theta}$ ，以及相应的 $\mathbf{z}$。对于每个抽取的参数和潜在序列，我们让该序列再推进 W 步。这样的构造从 $p(\tilde{\mathbf{z}}|\mathbf{y}_o)$ 中抽取一个样本，从而得到有效的预测分布。根据 Yeh 等人（2012）给出的前向—后向概率，可以采用类似于算法 1 或 Ryd´en（2008）中所定义算法的过程，从 $\mathbf{z}$ 中进行采样，其时间复杂度为 $O(T)$。

latex

Copy

基于新样本的预测分布 贝叶斯建模中另一种更常见的预测分布，是给定一个新序列 $\tilde{\mathbf{y}}_o$，预测相应的潜在序列 $\tilde{\mathbf{z}}$。请注意，

$$p(\tilde{\mathbf{z}}|\tilde{\mathbf{y}}_o, \mathbf{y}_o) \propto \int_\Omega p(\tilde{\mathbf{z}}, \tilde{\mathbf{y}}_o|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}_o) d\mu(\boldsymbol{\theta}),$$

这表明，在计算上，为 $\tilde{\mathbf{z}}$ 提供预测分布，只需从其边缘后验分布中抽取 $\boldsymbol{\theta}$，并从 $\tilde{\mathbf{y}}_o$中对每个潜在序列进行采样。正如上一节所讨论的，该过程的时间复杂度也是 $O(T)$。

缺失观测值填补 在贝叶斯建模中，填补缺失观测 $\mathbf{y}_m$也颇具意义。$\mathbf{y}_m$ 的预测分布可通过以下步骤获得:

$$\begin{aligned} p(\mathbf{y}_m|\mathbf{y}_o) &= \int_\Omega p(\mathbf{y}_m, \mathbf{z}, \boldsymbol{\theta}|\mathbf{y}_o) d\mu(\boldsymbol{\theta}) d\mu(\mathbf{z}) \\ &= \int_\Omega p(\mathbf{y}_m|\mathbf{z}, \boldsymbol{\theta}, \mathbf{y}_o) p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{y}_o) d\mu(\boldsymbol{\theta}) d\mu(\mathbf{z}) \\ &= \int_\Omega p(\mathbf{y}_m|\mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{y}_o) d\mu(\boldsymbol{\theta}) d\mu(\mathbf{z}). \end{aligned}$$

在计算上，我们首先从后验分布中抽取 $\boldsymbol{\theta}$ 和$\mathbf{z}$，并从 $p(\mathbf{y}_m|\boldsymbol{\theta},\mathbf{z})$中对$\mathbf{y}_m$进行采样，该分布由发射分布指定。

我们强调，尽管所提出的方法通常具有 $O(n(1-p)T)$ 的计算复杂度，但对新序列进行预测后验分布采样的时间复杂度与 T 成正比。因此，所提出算法的优势主要在于训练过程中较低的计算复杂度和更快的收敛率（如第 3 节所述）。事实上，所提出的算法之所以快速，是因为它在训练过程中避免了不必要的 $\mathbf{y}_m$ 和 $\mathbf{z}_o$ 预测与填补，从而加快了收敛速度并降低了时间复杂度。

3 复杂度分析

本节将分析所提出采样器的复杂度。我们认为，就收敛率和计算复杂度而言，所提出的折叠采样器比现有方法更快。

3.0.1 收敛率分析

为保证比较公平，我们主要比较式 1、式 2 和式 4 中所述采样器的收敛速度。令 $\mathbf{F}_g$ 为式 1 所述算法诱导的转移核，$\mathbf{F}_m$ 为式 2 所述算法诱导的转移核，$\mathbf{F}_c$ 为式 4 所述算法诱导的转移核。为保证比较公平，假设所有采样器都直接从各自的条件分布中采样。我们用谱隙来描述采样器的收敛率，其定义见 3.1。

定义 3.1 对于马尔可夫转移核$\mathbf{F}$，其谱隙 $Gap(\mathbf{F})$ 定义如下：

$$Gap(\mathbf{F}) = 1 - \lambda_1,$$

其中， λ_1 是其第二大特征值；对于可逆转移核，该值也称为谱半径（*Liu（1994）*）。

粗略地说，马尔可夫转移核的谱隙描述了其向平稳分布的收敛速率。更具体地说，具有谱隙 Gap**(F)** 的转移核 $\mathbf{F}$ 以总变差距离意义下的几何速率收敛（Liu 和 Liu（2001））：

$$||\pi_t - \pi||_{TV} \leq (1 - Gap(\mathbf{F}))^t ||\pi_0 - \pi||_{TV},$$

其中， π 是平稳分布， π_t 表示第 t 步时的分布。上述结果的正向推导可见于 Bardenet 等人（2017）。不难看出，谱隙更大的链往往收敛得更快。

T定理 3.2 三个吉布斯采样器的谱隙按以下顺序排列：*：*

$$Gap(\mathbf{F}_c) \geq Gap(\mathbf{F}_m) \geq Gap(\mathbf{F}_g).$$

证明 1（定理 3.2 的证明）我们遵循 *Liu（1994）*提出的证明思路。三个采样器的访问方案可用以下图示表示：

$$\begin{aligned} &\mathbf{F}_g : \boldsymbol{\theta} \rightarrow \mathbf{y}_m \rightarrow \{\mathbf{z}_m, \mathbf{z}_o\} \\ &\mathbf{F}_m : \boldsymbol{\theta} \rightarrow \{\mathbf{z}_m, \mathbf{z}_o\} \\ &\mathbf{F}_c : \boldsymbol{\theta} \rightarrow \mathbf{z}_o. \end{aligned} \tag{7}$$

根据 *Liu（1994）*的定理 *1*，可立即推导出：

$$||\mathbf{F}_c|| \leq ||\mathbf{F}_m|| \leq ||\mathbf{F}_g||,$$

其中， $||\mathbf{F}||$ 表示转移核$\mathbf{F}$的算子范数。由于这三个算子均为可逆算子，其谱半径等于其算子范数。对于可逆转移核，谱隙等于 *1* 减去谱半径（*Liu* 和 *Liu（2001）*；*Liu（1994）*），因此可立即得到所需结果。 ■

定理 3.2 表明，所提出的采样器收敛速度至少与 Yeh 等人（2012）；Speekenbrink 和 Visser（2021）提出的方法一样快，而这些方法的理论收敛速度超过了经典吉布斯采样器。因此，我们证明了所提出的方法具有较低的计算复杂度。更快的收敛率意味着所提出的采样器能够迅速与目标分布混合，并具有更高的有效样本量（Berger 等人（2014））。

3.0.2 计算复杂度分析

我们证明，所提出的采样器在计算上比其他具有竞争力的方法更快。与隐马尔可夫模型（HMM）相关的数值算法，其主要计算瓶颈在于计算前向概率或后向概率。所有与隐马尔可夫模型相关的研究，包括处理缺失观测的研究（Yeh 等人（2012）；Speekenbrink 和 Visser（2021））， 都会从 1 到 T 递归计算这些概率，时间复杂度为 $O(nT)$，其中 n 是样本量， T 是序列长度（不失一般性，我们假设所有序列的长度均为 T）。

然而，式 6 和式 5 表明，如果将缺失观测和潜在状态积分消去，只需针对$\mathbf{y}_o$中的观测值所对应的所有潜在状态，递归计算前向概率和后向概率。该方法的计算复杂度仅为 $O((1-p)nT)$，其中 p 是式 3 中定义的缺失率。由此可见，当 $p \to 0$时，时间复杂度接近 Yeh 等人（2012）；Yu 和 Kobayashi（2003）所提出算法的时间复杂度，这些算法对应完全观测情形。然而，当 $p \to 1$时，意味着存在大量缺失观测，此时计算复杂度会显著降低。这表明，当缺失率较高时，所提出的算法比其他竞争算法计算速度更快，因此特别适用于包含大量缺失条目的序列的数据集。

最后，我们指出，尽管式 6 和式 5 中需要用到矩阵指数，但我们可以提前缓存转移矩阵$\mathbf{A}$ 的乘法结果，并将其存储在字典中。采用这种优化技术后，对于 $k = 1,\ldots,T$，矩阵$\mathbf{A}^k$可以直接从字典中查找，其计算复杂度为 $O(1)$。因此，计算转移矩阵指数所产生的唯一时间复杂度来自预计算与缓存过程，其时间复杂度为

$O(T^2)$（最坏情况下）。因此，更新$\mathbf{z}_o$的整体计算复杂度变为 $O(npT + T^2) \approx O(npT)$，其中 $n \gg T$。由此，矩阵乘法引入的计算复杂度通常可以忽略不计。此外，在分块缺失机制下，只要缺失块的长度在一个固定长度附近变化，计算成本就能进一步降低。

4 模拟研究

本节通过数值模拟评估所提出的采样器的性能。我们将该采样器与三种竞争方法进行比较：EM算法（EM）、标准吉布斯采样采样器（如 1 所述），以及一种针对将$\mathbf{y}_m$ 积分掉后的分布进行采样的吉布斯采样器，如 Yeh 等人（2012）；Speekenbrink 和 Visser（2021）所述（其方案可由 2 描述，简称为部分折叠吉布斯）。我们将所提出的方法简称为折叠吉布斯。

我们通过数值模拟分三步展示所提出采样器的优势：首先，我们表明，所提出的采样器在估计和预测准确率方面表现相当。其次，我们表明，所提出的采样器计算速度更快，因为生成固定数量的样本所需时间更短。第三，我们展示了所提出的采样器在有效样本量（ESS）方面也具有较高的计算效率，即每秒的 ESS 更大。最后，我们表明，所提出的采样器在理论上也更快，因为其每次迭代的 ESS 也大于其他采样器。总之，所提出的方法在估计方面与其他方法表现相当，但在计算效率和采样效率方面更具优势。所有实验均在一台配备 AMD Ryzen 7 5800H CPU 和 32GB 内存的笔记本电脑上进行。

我们考虑具有以下参数化形式的隐马尔可夫模型 :

$$\boldsymbol{\pi} = (0.6, 0.3, 0.1),$$

$$\mathbf{A} = \begin{pmatrix} 0.6 & 0.3 & 0.1 \\ 0.1 & 0.6 & 0.3 \\ 0.3 & 0.1 & 0.6 \end{pmatrix},$$

$$\mathbf{B} = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{pmatrix},$$

如上一节所述， $\boldsymbol{\pi}$ 表示初始分布，$\mathbf{A}$表示转移矩阵，$\mathbf{B}$表示发射分布。

4.1 模拟研究 I：随机缺失

本节中，我们在随机缺失机制下进行模拟。我们生成 $n = 500$ 个序列，每个序列的长度为 $T = 20$。所有观测值均以 p 的概率随机删除。我们将概率 p 设为 0，0.1、0.3、0.5、0.7、0.9。当缺失概率为 $p = 0$ 时，对应于序列完全观测到的情形。

首先，我们比较估计精度。对于参数，估计精度通过真实值与后验均值之间的均方误差（MSE）来评估。对于潜在状态，各算法的性能根据从后验分布汇总得出的多数投票结果，通过潜在状态 $\mathbf{z}$ 的预测准确率来衡量。对于EM算法，我们首先估计参数，然后通过维特比算法（Forney（1973））得出预测结果，以此预测隐状态。我们从后验分布中抽取5000个样本，并将前一半作为预热样本丢弃。

表1显示，不同采样器的估计误差和预测准确率相当；在不同缺失率下，各采样器的估计精度也没有显著差异。这表明，所提出的方法在近似后验分布方面与其他方法同样有效。

基于所有采样器在估计方面表现相近这一观察，我们认为，所提出采样器的优势在于其后验采样效率。具体而言，它能够在给定时间内生成更多样本，每次迭代都能获得更高的有效样本量（ESS），并且在给定时间段内也能取得更大的ESS。图4.1首先展示了每个采样器运行1000次迭代所需时间的结果。当缺失概率设为0%（对应完全观测情形）时，所有方法的表现相近。然而，随着缺失概率增加，所提出的采样器开始比竞争方法运行得更快。当缺失概率达到90%时，所提出的方法几乎比竞争方法快四倍，体现了其在计算复杂度方面的优势。此外，随着缺失概率增加，所提出的方法在时间复杂度方面直接呈现出线性衰减的模式。

接下来，我们通过考察每次迭代产生的有效样本量（ESS），比较各采样器的理论采样效率。鉴于模型中的参数众多，我们仅评估有效样本量的中位数。如表2所示，涉及缺失观测时，所提出的采样器所产生的有效样本量始终优于竞争方法。表2中的结果从数值上验证了所提出的采样器在理论上比现有方法更高效（就收敛率而言）。由于与其他算法相比，所提出的方法每次迭代的有效样本量更高，且每1000次迭代所耗费的时间更少，因此其每秒有效样本量也更高，体现出其在后验探索方面的优越效率。结合运行

表1：不同采样器生成的后验分布在不同缺失概率（随机缺失模式）下的估计性能。在固定的缺失概率水平下，所有采样器在后验近似方面的性能相近。

缺失概率	模型	潜在状态预测准确率	初始误差	转移误差	发射误差
0%	部分折叠吉布斯	0.79 (0.01)	0.0050 (0.00354)	0.00063 (0.00060)	0.0010 (0.00133)
	折叠吉布斯	0.79 (0.011)	0.0050 (0.00352)	0.00088 (0.00090)	0.0012 (0.00139)
	标准吉布斯采样	0.79 (0.01)	0.0026 (0.00184)	0.00065 (0.00099)	0.0007 (0.00064)
	EM	0.76 (0.021)	0.0062 (0.00421)	0.00118 (0.00120)	0.0015（0.00150）
10%	部分折叠吉布斯	0.76（0.0131）	0.0080（0.00568）	0.00115（0.00119）	0.0015（0.00219）
	折叠吉布斯	0.76 (0.023)	0.0009 (0.00062)	0.00035 (0.00049)	0.0005 (0.00091)
	标准吉布斯采样	0.77（0.031）	0.0077（0.00544）	0.00085（0.00065）	0.0016（0.00157）
	EM	0.65（0.030）	0.0118（0.00733）	0.00120 (0.00140)	0.0022 (0.00250)
30%	部分折叠吉布斯	0.71 (0.0121)	0.0014 (0.00098)	0.00062 (0.00058)	0.0008 (0.00071)
	折叠吉布斯	0.70 (0.021)	0.0017 (0.00075)	0.00031 (0.00025)	0.0004 (0.00044)
	标准吉布斯采样	0.70 (0.023)	0.0006 (0.00042)	0.00073 (0.00064)	0.0010 (0.00103)
	EM	0.57 (0.025)	0.0021 (0.00150)	0.0010 (0.00090)	0.0018 (0.00200)
50%	部分折叠吉布斯	0.64 (0.0191)	0.0013 (0.00092)	0.00388 (0.00474)	0.0020 (0.00161)
	折叠吉布斯	0.64 (0.0201)	0.0018 (0.00125)	0.00360 (0.00305)	0.0031 (0.00315)
	标准吉布斯采样	0.64 (0.0212)	0.0137 (0.00969)	0.00294 (0.00317)	0.0029 (0.00260)
	EM	0.47 (0.015)	0.0045 (0.00300)	0.0040 (0.00400)	0.0045 (0.00450)
70%	部分折叠吉布斯	0.56 (0.022)	0.0093 (0.00658)	0.00399 (0.00396)	0.0035 (0.00445)
	折叠吉布斯	0.55 (0.049)	0.0197 (0.01393)	0.01290 (0.01672)	0.0110 (0.01841)
	标准吉布斯采样	0.54 (0.02)	0.0006 (0.00042)	0.00244 (0.00209)	0.0026（0.00282）
	EM	0.45（0.018）	0.0205（0.01512）	0.0130（0.01200）	0.0115（0.01150）
90%	部分折叠吉布斯	0.40 (0.0231)	0.0011 (0.00079)	0.00681 (0.00793)	0.0092 (0.01084)
	折叠吉布斯	0.38 (0.022)	0.0154 (0.01089)	0.02056 (0.01874)	0.1226 (0.09503)
	标准吉布斯采样	0.37 (0.0218)	0.0140 (0.00987)	0.07118 (0.08169)	0.1536 (0.21586)
	EM	0.36 (0.030)	0.0254 (0.02010)	0.0300 (0.02950)	0.1500 (0.14000)

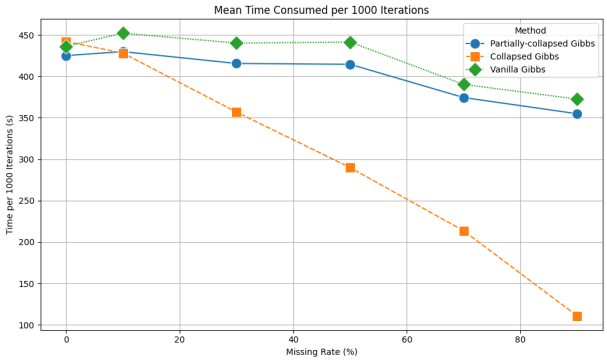

图1：随机缺失机制下，不同缺失概率对应的不同采样器每1000次迭代所耗费的时间。由于计算复杂度是确定的，我们仅展示平均耗时（事实上，不同实验中的计算时间标准差也很小，主要由系统抖动造成）。带圆形标记的线表示部分折叠吉布斯采样器的平均耗时，带菱形标记的线表示原始吉布斯采样器的耗时。带方形标记的线表示所提出的折叠采样器在不同缺失概率下的平均耗时。随着缺失率升高，在随机缺失机制下，就每次迭代的绝对耗时而言，所提出的方法明显快于竞争方法。此外，从图中可以看出，我们的方法的计算时间呈线性衰减，这验证了复杂度分析中给出的结果。

根据图4.1所述的时间结果，我们从理论角度（以每次迭代的有效样本量衡量）和计算角度（以每秒有效样本量及运行时间衡量）实证表明，所提出的方法能够快速探索后验分布。

表2：比较不同缺失概率下各吉布斯抽样方法每次迭代的有效样本量。涉及缺失观测时，所提出的方法始终能在每次迭代中产生更大的有效样本量。

缺失概率	标准吉布斯采样	折叠吉布斯	部分折叠吉布斯
0%	0.0067 (0.0005)	0.0067（0.0005）	**0.0072（0.0004）**
10%	0.0036（0.0003）	**0.0038（0.0003）**	0.0034（0.0002）
30%	0.0019（0.0002）	**0.0026 (0.0004)**	0.0023 (0.0003)
50%	0.0024 (0.0003)	**0.0036 (0.0005)**	0.0026 (0.0003)
70%	0.0016 (0.0002)	**0.0031 (0.0004)**	0.0019 (0.0002)
90%	0.0010 (0.0001)	**0.0047 (0.0006)**	0.0008 (0.0001)

4.2 模拟研究 II：分块缺失

我们继续通过考察分块缺失情形开展模拟研究。在分块缺失设置中，我们不再随机设定缺失条目，而是指定序列中的一个连续区块作为缺失部分。图 4.2 以图示形式展示了分块缺失模式。黑色单元格表示观测条目，白色连续区块则表示观测序列中的缺失块。

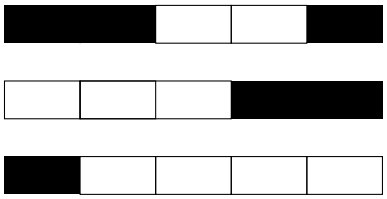

图 2：具有分块缺失模式的序列。白色块表示观测序列中的缺失条目

为生成存在分块缺失的不完整序列，我们按如下步骤进行：对于每个序列，选取一个随机索引集合 $\{i, i+1, \ldots, i+j\}$，其中 $j \approx T \times p$。将这些索引对应的观测值设为缺失。数值实验完全沿用上一节讨论的步骤。表 3 记录了不同缺失数据程度下各算法的估计精度。可以清楚地看到，与现有方法相比，所提出算法的估计精度相近。图 3 对每 1000 次迭代所耗费的时间进行了直接比较，结果表明，当缺失概率较高时，所提出的采样器始终比现有方法耗时更少，体现出其在计算效率方面的优势。此外，随着缺失概率增加，所提出方法的时间复杂度呈现出明显的线性衰减趋势。同样，表 4 比较了每次迭代的有效样本量中位数，展示了理论上的采样效率。总而言之，我们通过实证验证，面对存在分块缺失的数据集时，所提出的方法无论在计算层面还是理论层面的后验探索中，都比现有算法更快。

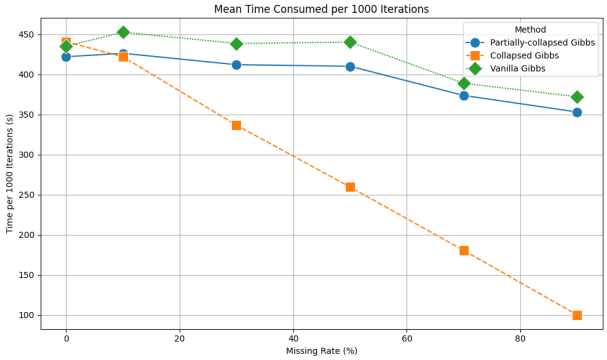

图 3：在块缺失机制下，不同采样器每 1000 次迭代的平均耗时比较。由于计算复杂度是确定的且不确定性很小，我们仅报告均值（事实上，真实标准差也相当小，主要由系统抖动引起）。带圆形标记的线表示部分折叠吉布斯采样器的平均耗时，带菱形标记的线表示原始吉布斯采样器的耗时。所提出的折叠采样器在不同缺失概率下的平均耗时由带方形标记的线表示。在块缺失机制下，当缺失概率较高时，所提出的方法明显快于其他任何现有方法。此外，从图中可以得出，我们的方法的计算时间呈线性衰减，这验证了复杂度分析中给出的结果。

5 真实数据分析

在真实数据分析部分，我们采用所提出的框架分析两个不同的数据集。首先，我们将所提出的算法应用于精神分裂症数据集（Hedeker 和 Gib-bons（1997））以及 CLEAR 数据集（Huang等（2019）），以比较所提出的方法与现有算法之间的采样效率。所有实验均在配备 AMD Ryzen 7 5800H CPU 和 32GB 内存的笔记本电脑上进行。

5.1 真实数据分析 I：精神分裂症研究

本节中，我们将贝叶斯隐马尔可夫模型拟合到精神分裂症数据集上。该数据集是美国国家精神卫生研究所发布的公开疾病进展数据集。我们还将比较各种方法的采样效率。

精神分裂症数据集（Hedeker 和 Gibbons（1997））包含 437 名确诊为精神分裂症的患者的不完整观测序列。其中，108 名患者服用了安慰剂，其余 329 名患者接受了药物治疗。研究人员连续六周监测他们的健康状况，并要求他们每天报告自己的精神状态。精神状态是一个分类变量，分为四个等级：重度、中度、轻度或正态。该数据集于 1997 年首次得到研究，当时重点关注模式混合模型的应用。较新的研究，例如 Yeh 等人（2012年）以及 Speekenbrink 和 Visser（2021年），则采用最大似然估计（MLE）方法，将该数据集建模为隐马尔可夫模型（HMM）。

鉴于精神分裂症患者可能无法准确报告自己的病情，使用隐马尔可夫模型对这一过程进行建模既顺理成章，也有其优势。事实上，将真实病情视为未知隐变量，并将观测到的病情视为从发射过程中抽样得到的变量，这种方法在

医疗记录分析文献中已广为确立（Altman 和 Petkau（2005年）） .

在精神分裂症数据集中，约有50%的条目缺失。具体而言，第2、4和5周的观测值中，缺失比例超过90%。与此同时，第3和6周约有30%的条目缺失。这种缺失模式表明，缺失可能是由研究设计导致的，因为它们与特定日期密切相关，因此可视为可忽略的缺失。

为分析这些数据，我们采用以下方法：分别针对接受安慰剂的患者和接受研究药物治疗的患者，训练两个不同的隐马尔可夫模型。对于每个模型，我们从均匀分布中初始化参数，并运行吉布斯采样器5000步。舍弃前2500个样本作为预热期，剩余2500个样本用于后验抽样。所提方法估计得到的后验均值和标准差列于表6，供参考。每项实验重复十次。补充材料中提供了对所估计后验分布的完整说明。这些图中的红色实线表示估计的后验均值。

除后验采样外，我们还将重点放在采样效率上。由于无法获得真实参数，我们通过随机遮蔽部分已有观测值，并让模型对其进行预测，来评估估计性能。随后，将对缺失观测的预测准确率与真实值进行比较，从而定量评估不同采样器生成的后验分布的估计精度。我们在报告中将这一指标命名为交叉验证预测准确率 。此外，与第4.1节和第4.2节类似，我们还报告了各采样器在平均时间复杂度、收敛率（以每次迭代的有效样本量衡量）以及总体采样效率（以一段时间内的有效样本量数量衡量）方面的比较结果。表5展示了在治疗组和对照

组上针对各种指标评估所得的相应结果。总的来说，所有算法在后验预测方面均取得了相近的结果，而所提出的折叠采样器在计算上更快，并且每次迭代产生更大的有效样本量，因此在10次实验取平均后，在给定时间段内能够获得更高的有效样本量。此外，这进一步验证了在缺失概率较高的情形下，所提出的采样器在时间复杂度方面的优势。因此，表5汇总的结果表明，正如前文所述，所提出的采样器在更低的计算复杂度和更好的收敛率方面具有优势。

5.2 真实数据分析 II：MRSA感染研究

在本节中，我们将所提出的算法应用于 CLEAR 数据集（Huang等（2019）），以开展全面的医疗记录分析，并将其采样效率与现有算法进行比较。CLEAR 数据集追踪了来自不同群体的患者，记录其耐甲氧西林金黄色葡萄球菌（MRSA）感染及定植状态。

MRSA 是一种可能存在于人体内且通常难以根除的危险细菌（Turner等（2019））。严重的 MRSA感染可能导致重度脓毒症，甚至危及生命。在本研究中，研究人员于实验开始时招募了此前感染过 MRSA 的患者，并将其分为治疗组和对照组。随后，对治疗组患者实施了重复去定植方案。两组患者均接受了一般卫生和环境清洁方面的教育。治疗组包括 1058 名患者，而对照组包括 1063 名患者。

在第0、1、3和6天从患者的咽喉部采集样本，并提供检测结果以确定每位患者样本中的MRSA定植。因此，

第2、4、5天的观测值被视为缺失数据。此外，第0、1、3和6天的观测值也存在许多缺失条目。在对照组中，第1天和第3天的观测值约有20%缺失，而第6天的观测值有33%缺失。总体缺失率约为55%。在治疗组中，第1、3和6天的观测值约有30%缺失，总体缺失率约为57%。总体而言，这些序列呈现出分块缺失结构，同时伴有一些随机缺失。

基于上一节所述的治疗方案，我们将检测结果建模为对 MRSA 定植真实状态的不准确观测，而该真实状态是不可观测的潜在状态。这种做法是合理的，因为 MRSA 通常难以根除，且检测结果呈阴性并不一定意味着已彻底根除（Turner等（2019））。我们分别为治疗组和对照组训练两个不同的模型。我们使用 MCMC 生成 5000 个样本，并保留最后 2500 个样本作为后验抽样。与第 5.2 节的设置类似，我们通过八次重复实验，比较不同采样器在平均预测准确率、时间复杂度和收敛率方面的表现。

估计得到的后验均值和标准差记录于表8。该表列出了使用我们的方法框架估计的各参数的后验均值和方差。有关 MCMC 模拟所得后验分布的更详细可视化结果，请参阅补充材料。转移矩阵的后验分布表明，治疗在根除 MRSA感染方面更加有效。这一点可以从以下事实看出：对于感染 MRSA 的患者，治疗组患者从感染转变为治愈的概率通常显著更高。

治疗组和对照组的采样效率比较

见表7。表中结果表明，所有采样器的预测得分相近。然而，所提出的采样器在后验探索方面效率更高，因为运行1000次迭代所需时间更短，且其每次迭代的有效样本量（ESS）显著高于竞争方法，因此每秒的有效样本量更高。这些结果验证了我们在第3节中得出的结论：所提出的采样器具有更低的计算复杂度和更好的收敛率。

6 结论与讨论

本文针对存在缺失数据的贝叶斯隐马尔可夫模型，提出了一种新颖的吉布斯抽样算法。该算法从塌缩分布中进行采样，并将与缺失观测对应的潜在状态积分消去。这种方法降低了潜在状态采样的计算复杂度，并提高了收敛速度，尤其是在缺失数据概率较高时。这些优势通过理论分析以及基于模拟和真实数据分析的实证证据得到了验证。与现有算法相比，在合成数据集和真实世界数据集上的实证评估表明，所提算法在每次迭代和每秒的有效样本量方面具有计算优势。因此，我们得出结论：通过积分消去无关数据，所提算法在计算和理论效率方面均优于现有算法。

补充材料

软件： 我们的代码已公开发布于 https://github.com/lidongrong/PHMM。Python 软件包也将在未来发布。

在补充材料中，我们提供了更多真实数据应用的数值结果，并对后验密度进行了详细的图形描述。

参考文献

Albert, P. S.（2000年）。一种适用于存在不可忽略缺失数据的纵向二元数据的过渡模型。*Biometrics 56*(2)，602–608。

Altman, R. M. 和 A. J. Petkau（2005年）。隐马尔可夫模型在多发性硬化病灶计数数据中的应用。《医学统计学》*24*(15)，2335–2344。

Bahl, L. R.、F. Jelinek 和 R. L. Mercer（1983年）。一种连续语音识别的最大似然方法。*IEEE*《模式分析与机器智能汇刊》(2)，179–190。

Bardenet, R., A. Doucet 和 C. Holmes（2017年）。关于适用于高维数据的马尔可夫链蒙特卡罗方法。《机器学习研究杂志》*18*(47), 1–43. Bellman, R.（1966）。动态规划。《科学》*153*(3731), 34–37. Berger, J., M. Bayarri 和 L. Pericchi（2014）。有效样本量。《计量经济学评论》*33*(1-4), 197–217. Bishop, C. M. 和 N. M. Nasrabadi（2006）。模式识别与机器学习，第4卷。Springer。Blei, D. M., A. Kucukelbir 和 J. D. McAuliffe（2017）。变分推断：面向统计学家的综述。*Journal of the American Statistical Association 112*(518), 859–877。

Chassan, M. 和 D. Concordet（2023年）。如何检验隐马尔可夫模型中的缺失数据机制。*Computational Statistics & Data Analysis 182*, 107723。

Cooke, M., P. Green, L. Josifovski, and A. Vizinho (2001年). 针对缺失和不可靠声学数据的鲁棒自动语音识别。*Speech Communication 34*(3), 267–285.

Deltour, I., S. Richardson, and J.-Y. L. Hesran (1999年). 基于间歇性缺失数据的马尔可夫模型估计随机算法。*Biometrics 55*(2), 565–573.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977年). 通过EM算法从不完整数据中求得最大似然。*Journal of the Royal Statistical Society: Series B(Methodological) 39*(1), 1–22.

Dias, J. G. and M. Wedel (2004). 对于难以处理的高斯混合似然，比较EM、SEM和MCMC的性能。*Statistics and Computing 14*, 323–332.

Forney, G. D.（1973）。维特比算法。*IEEE*会刊 *61*(3), 268–278。Foti, N., J. Xu, D. Laird, and E. Fox（2014）。隐马尔可夫模型的随机变分推断。神经信息处理系统进展 *27*。Gelfand, A. E. and A. F. Smith （1990）。计算边际密度的基于抽样的方法。*Journal of the American statistical Association 85*(410), 398–409。Gelman, A., W. R. Gilks, and G. O. Roberts（1997）。随机游走Metropolis算法的弱收敛与最优尺度。*The Annals of Applied Probability 7*(1), 110–120。

Ghahramani, Z. and M. Jordan（1995年）。因子隐马尔可夫模型。神经信息处理系统进展 *8*。

Gilks, W. R., N. G. Best 和 K. K. Tan（1995年）。用于吉布斯抽样的自适应拒绝 Metropolis 采样。皇家统计学会期刊 *C*辑：应用统计 *44*(4), 455–472。

Hedeker, D. 和 R. D. Gibbons（1997）。随机效应模式混合模型在纵向研究缺失数据中的应用。心理学方法 *2*(1), 64。

Huang, S. S., R. Singh, J. A. McKinnell, S. Park, A. Gombosev, S. J. Eells, D. L. Gillen, D. Kim, S. Rashid, R. Macias-Gil 等（2019年）。通过去定植降低 MRSA 携带者的出院后感染风险。新英格兰医学杂志 *380*(7), 638–650。

James, M. R., V. Krishnamurthy 和 F. Le Gland（1996）。用于隐马尔可夫模型参数估计的连续时间滤波器和平滑器的时间离散化。*IEEE* 信息论汇刊 *42*(2), 593–605。

Kawamoto, R., A. Nazir, A. Kameyama, T. Ichinomiya, K. Yamamoto, S. Ta mura, M. Ya- mamoto, S. Hayamizu 和 Y. Kinosada（2013年）。用于分析时间序列健康检查数据的隐马尔可夫模型。载于*MEDINFO 2013*，第 491–495 页。IOS Press。

Lee, D., D. Kulic 和 Y. Nakamura（2008年）。使用因子隐马尔可夫模型恢复缺失的运动数据。载于*2008*年 *IEEE* 国际机器人与自动化会议，第 1722–1728 页。IEEE。

Little, R. J. 和 D. B. Rubin（2019年）。缺失数据的统计分析，第 793 卷。John Wiley &Sons。

Liu, J. S.（1994年）。贝叶斯计算中折叠吉布斯采样器及其在基因调控问题中的应用。*Journal of the American Statistical Association 89*（427），958–966。

Liu, J. S. and J. S. Liu（2001年）。科学计算中的蒙特卡罗策略，第10卷。Springer。

Liu, J. S. and Y. N. Wu（1999年）。数据增广的参数扩展。*Journal ofthe American Statistical Association 94*(448)，1264–1274。

Liu, Y.-Y.、S. Li、F. Li、L. Song 和 J. M. Rehg（2015年）。用于疾病进展的连续时间隐马尔可夫模型高效学习。神经信息处理系统进展 *28*。

Lu, S.（2017年）。一种用于对地震震级-频度分布建模的连续时间 HMM 方法。应用统计学杂志 *44*(1)，71–88。

Ng, S. K.、T. Krishnan 和 G. J. McLachlan（2012年）。EM算法。计算统计学手册：概念与方法，139–172。

Patterson, S. 和 Y. W. Teh（2013年）。概率单纯形上的随机梯度黎曼朗之万动力学。神经信息处理系统进展 *26*。

Perikos, I.、S. Kardakis、M. Paraskevas 和 I. Hatzilygeroudis（2019年）。社交媒体情感分析中的隐马尔可夫模型。载于*2019*年 *IEEE*大数据、云计算、数据科学与工程国际会议（*BCD*），第 130–135 页。IEEE。

Rabiner, L. R.（1989年）。隐马尔可夫模型及其在语音识别中若干应用的教程。*IEEE* 会刊 *77*(2)，257–286。

Robert, C.、G. Casella、C. P. Robert 和 G. Casella（2010年）。Metropolis–Hastings 算法。《*R*语言蒙特卡罗方法》，167–197。Rubin, D. B.（1976年）。推断与缺失数据。*Biometrika 63*(3)，581–592。Ryd´en, T.（2008年）。期望最大化与马尔可夫链蒙特卡罗在隐马尔可夫模型估计中的比较：计算视角。Speekenbrink, M. 和 I. Visser（2021年）。隐马尔可夫模型中的可忽略缺失数据与不可忽略缺失数据。Turner, N. A.、B. K. Sharma-Kuinkel、S. A. Maskarinec、E. M. Eichenberger、P. P. Shah、M. Carugati、T. L. Holland 和 V. G. Fowler Jr（2019年）。耐甲氧西林金黄色葡萄球菌：基础研究与临床研究概述。*Nature Reviews Microbiology 17*(4)，203–218。Yanchenko, A. K. 和 S. Mukherjee（2017年）。使用状态空间模型进行古典音乐作曲。*arXiv* 预印本 *arXiv:1708.03822*。Yeh, H.-W.、W. Chan 和 E. Symanski（2012年）。离散时间隐马尔可夫模型中的间歇性缺失观测。*Communications in Statistics-Simulation and Computation 41*(2)，167–181。Yeh, H.-W.、W. Chan、E. Symanski 和 B. R. Davis（2010年）。离散时间马尔可夫链中可忽略间歇性缺失数据的转移概率估计。*Communications in Statistics—Simulation and Computation® 39*(2)，433–448。Yoon, B.-J.（2009年）。隐马尔可夫模型及其在生物序列分析中的应用。《当代基因组学》*10*(6)，402–415。

Yu, S.-Z. 和 H. Kobayashi（2003）。用于移动性跟踪的含缺失数据和多条观测序列的隐半马尔可夫模型。信号处理 *83*(2), 235–250。

表 3：不同模型生成的后验分布在不同缺失概率下（分块缺失模式）的估计性能。在固定的缺失概率水平下，所有模型生成的分布在后验近似方面的性能相近。

缺失概率	模型	潜在状态预测准确率	初始误差	转移误差	发射误差
0%	标准吉布斯采样	0.78 (0.023)	0.0028 (0.0020)	0.0004 (0.0005)	0.0007 (0.0010)
	折叠吉布斯	0.78 (0.020)	0.0083 (0.0059)	0.0015 (0.0014)	0.0029 (0.0035)
	部分折叠吉布斯	0.78 (0.022)	0.0038 (0.0027)	0.0008 (0.0010)	0.0010 (0.0013)
	EM	0.76 (0.030)	0.0120 (0.0090)	0.0025 (0.0026)	0.0040 (0.0041)
10%	标准吉布斯采样	0.76 (0.018)	0.0002 (0.0002)	0.0016 (0.0023)	0.0018 (0.0019)
	折叠吉布斯	0.74 (0.019)	0.0059 (0.0042)	0.0014 (0.0019)	0.0010 (0.0016)
	部分折叠吉布斯	0.76 (0.022)	0.0024 (0.0017)	0.0012 (0.0011)	0.0008 (0.0009)
	EM	0.72 (0.025)	0.0085 (0.0060)	0.0028 (0.0029)	0.0035 (0.0036)
30%	标准吉布斯采样	0.68 (0.015)	0.0043 (0.0031)	0.0008 (0.0011)	0.0014 (0.0013)
	折叠吉布斯	0.65 (0.030)	0.0007 (0.0005)	0.0012 (0.0014)	0.0012 (0.0012)
	部分折叠吉布斯	0.67 (0.014)	0.0003 (0.0002)	0.0015 (0.0024)	0.0021 (0.0022)
	EM	0.57 (0.035)	0.0020 (0.0018)	0.0023 (0.0024)	0.0030 (0.0031)
50%	标准吉布斯采样	0.59 (0.022)	0.0021（0.0015）	0.0033（0.0039）	0.0032（0.0037）
	折叠吉布斯	0.56（0.025）	0.0020（0.0014）	0.0031 (0.0035)	0.0026 (0.0033)
	部分折叠吉布斯	0.58 (0.013)	0.0014 (0.0010)	0.0017 (0.0012)	0.0019 (0.0019)
	EM	0.47 (0.028)	0.0050 (0.0040)	0.0045 (0.0046)	0.0048 (0.0049)
70%	标准吉布斯采样	0.49 (0.027)	0.0034 (0.0024)	0.0024 (0.0039)	0.0035 (0.0046)
	折叠吉布斯	0.47 (0.014)	0.0143 (0.0101)	0.0049 (0.0036)	0.0047 (0.0039)
	部分折叠吉布斯	0.50 (0.020)	0.0039 (0.0028)	0.0031 (0.0028)	0.0020 (0.0023)
	EM	0.45 (0.032)	0.0200 (0.0150)	0.0058 (0.0050)	0.0060 (0.0055)
90%	标准吉布斯采样	0.38 (0.029)	0.0094 (0.0067)	0.1425 (0.1414)	0.0606 (0.0666)
	折叠吉布斯	0.37 (0.021)	0.0052 (0.0037)	0.0095 (0.0099)	0.0152 (0.0160)
	部分折叠吉布斯	0.37 (0.016)	0.0167 (0.0118)	0.0146 (0.0159)	0.0219 (0.0239)
	EM	0.35 (0.035)	0.0250 (0.0200)	0.0150 (0.0140)	0.0200 (0.0190)

表 4：在分块缺失机制下，按每次迭代的有效样本量中位数比较缺失概率不同时的吉布斯抽样方法。涉及分块缺失观测时，所提出的方法始终能在每次迭代中产生更大的有效样本量。

缺失概率	标准吉布斯采样	折叠吉布斯	部分折叠吉布斯
0%	**0.0046 (0.0003)**	0.0039 (0.0002)	0.0044 (0.0003)
10%	0.0029 (0.0004)	**0.0049 (0.0003)**	0.0027 (0.0002)
30%	0.0025 (0.0003)	**0.0037 (0.0002)**	0.0036 (0.0002)
50%	0.0010 (0.0001)	**0.0050 (0.0004)**	0.0016 (0.0001)
70%	0.0013 (0.0001)	**0.0025 (0.0002)**	0.0011 (0.0001)
90%	0.0011 (0.0001)	**0.0107 (0.0005)**	0.0058 (0.0003)

表5：对接受药物和安慰剂的患者组进行吉布斯抽样方法比较。实证评估表明，所提出的方法在计算和理论层面的后验探索中效率更高，因为每次迭代所需时间更少（对于计算时间，我们仅报告均值，因为唯一的不确定性来自计算机的随机抖动），且每次迭代产生的有效样本数高于其他竞争算法。

		指标			
组别	采样器	每 1000 次迭代所需时间	每次迭代的中位数有效样本量	每秒的中位数有效样本量	交叉验证预测准确率
治疗组	标准吉布斯采样	126.95	0.0031 (0.0003)	0.024	**45.21% (2.3%)**
	折叠吉布斯	**82.99**	0.0036 (0.0005)	**0.044**	44.57% (3.1%)
	部分折叠吉布斯	110.51	0.0011 (0.0008)	0.010	**45.34%** (1.7%)
对照组	标准吉布斯采样	127.59	0.0022 (0.0011)	0.017	**47.88%** (3.2%)
	折叠吉布斯	**82.07**	0.016 (0.0036)	**0.200**	47.56% (3.0%)
	部分折叠吉布斯	109.77	0.014（0.0007）	0.130	47.20%（3.2%）

表6：精神分裂症数据集研究中估计的后验均值（标准差）

		安慰剂组				治疗组			
		正态	Mild	中度	重度	正态	Mild	中度	重度
初始分布		0.01(10^{-4})	0.10(0.001)	0.33(0.004)	0.56(0.003)	0.01(10^{-5})	0.09(10^{-4})	0.26(0.0020)	0.62(0.0019)
转移矩阵	**正态**	0.69(0.019)	0.12(0.011)	0.10(0.008)	0.09(0.007)	0.94(0.001)	0.05(0.001)	0.01(0.000)	0.01(0.000)
	Mild	0.04(0.001)	0.90(0.002)	0.05(0.001)	0.02(0.000)	0.20(0.001)	0.78(0.001)	0.02(0.000)	0.01(0.000)
	中度	0.02(0.001)	0.19(0.003)	0.73(0.005)	0.07(0.001)	0.05(0.000)	0.26(0.001)	0.66(0.002)	0.03(0.000)
	重度	0.01(0.000)	0.03(0.000)	0.08(0.001)	0.87(0.001)	0.03(0.000)	0.11(0.001)	0.30(0.002)	0.56(0.001)
发射矩阵	**正态**	0.62(0.026)	0.13(0.014)	0.14(0.014)	0.11(0.010)	0.86(0.005)	0.11(0.005)	0.01(0.000)	0.01(0.000)
	Mild	0.03(0.001)	0.86(0.004)	0.08(0.004)	0.03(0.001)	0.01(0.000)	0.93(0.001)	0.05(0.001)	0.01(0.000)
	中度	0.01(0.000)	0.10(0.004)	0.82(0.005)	0.06(0.002)	0.01(0.000)	0.19(0.003)	0.75(0.003)	0.06(0.002)
	重度	0.01(0.000)	0.02(0.000)	0.06(0.001)	0.91(0.001)	0.00(0.000)	0.01(0.000)	0.07(0.002)	0.91(0.002)

表7：在 CLEAR 数据集中对治疗组和对照组进行吉布斯抽样方法比较。实证评估表明，所提出的方法在计算和理论层面的后验探索中效率更高，因为每次迭代所需时间更少（对于计算时间，我们仅报告均值，因为唯一的不确定性来自计算机的随机抖动），且每次迭代产生的有效样本数高于其他竞争算法。

		指标			
组	采样器	每 1000 次迭代所需时间	每次迭代的中位数有效样本量	每秒的中位数有效样本量	交叉验证预测准确率
治疗组	标准吉布斯采样	343.24	0.00017 (0.00012)	0.00049	**88.21% (1.2%)**
	折叠吉布斯	**232.64**	**0.0026** (0.0003)	**0.011**	**87.57% (1.2%)**
	部分折叠吉布斯	397.91	0.0017 (0.0003)	0.0043	**88.89%** (0.9%)
对照组	标准吉布斯采样	387.92	0.00037 (0.0003)	0.00095	**82.01% (0.9%)**
	折叠吉布斯	**287.85**	**0.0019** (0.0004)	**0.00667**	**82.97%** (1.3%)
	部分折叠吉布斯	320.83	0.0015 (0.0002)	0.00467	**80.13% (1.8%)**

表8：耐甲氧西林金黄色葡萄球菌感染研究中的后验均值估计值（标准差）

	初始分布	转移矩阵		发射矩阵	
状态		**Cure**	**感染**	**Cure**	**感染**
治愈（对照组）	0.71 (0.028)	0.99(0.006)	0.01(0.006)	0.94(0.008)	0.06(0.008)
感染（对照组）	0.29 (0.028)	0.11(0.032)	0.89(0.032)	0.35(0.050)	0.65(0.050)
治愈（治疗组）	0.69 (0.035)	0.98(0.024)	0.03(0.024)	0.99(0.011)	0.01(0.011)
感染（治疗组）	0.31 (0.035)	0.29(0.004)	0.72(0.004)	0.30(0.045)	0.71(0.045)