arXiv:2302.06031v1 [stat.ME] 13 Feb 2023

错误设定模型中的可靠贝叶斯推断

DAVID T. FRAZIER、ROBERT KOHN、CHRISTOPHER DROVANDI 和 DAVID GUNAWAN

摘要。我们从频率学派视角出发，为贝叶斯推断中的一个基本开放问题——设定错误的模型下贝叶斯方法的不确定性量化效果不佳——提供了一般性解决方案。现有解决方案主要基于对后验分布进行显式高斯近似，或采用计算开销较大的后处理程序；而我们证明，只需用一种直观的近似后验分布替代通常的后验分布，即可实现正确的不确定性量化。重要的是，我们的解决方案不仅适用于基于似然的后验分布和广义贝叶斯后验分布，也适用于似然难以处理、必须进行估计的情形。我们从形式上证明了所提出方法能够实现可靠的不确定性量化，并表明有效的不确定性量化并非渐近结果，即使在小样本下也能实现。我们通过一系列示例说明这一方法，包括线性混合效应模型和广义混合效应模型。关键词：贝叶斯推断、广义贝叶斯推断、模型错设、估计似然

, Pseudo-marginal

1. 引言

贝叶斯方法因其能够处理复杂模型、灵活应对潜变量，并能对涉及所有模型未知量的不确定性进行整体性（尽管是主观的）处理而广受赞誉。尽管这种方法本质上具有主观性，但在广为人知的正则条件下，使用贝叶斯方法获得的不确定性表达式在渐近地意义上与频率学派方法一致。然而，当（贝叶斯）后验分布更新所依据的模型设定错误时，贝叶斯与频率学派统计范式之间的一致性便会丧失，使用贝叶斯方法获得的不确定性表达式从频率学派视角来看也不再“有效”。

Frazier：莫纳什大学计量经济学与商业统计系，克莱顿，维多利亚州3800，澳大利亚。Kohn：新南威尔士大学澳大利亚商学院经济学院，悉尼，新南威尔士州2052，澳大利亚。Drovandi：昆士兰科技大学数学科学学院、数据科学中心，布里斯班，昆士兰州4000，澳大利亚。Gunawan：伍伦贡大学数学与应用统计学院。通讯作者：david.frazier@monash.edu。

为纠正这一问题，已有研究提出了若干后验分布后处理方法（例如，参见 Müller，2013，Syring 和 Martin，2019，以及 Matsubara 等人，2021）。总体而言，据我们所知，所有这类程序都相当于将事后频率学派原则应用于贝叶斯学习算法的输出。在本文中，我们提出一种内在的贝叶斯解决方案，以解决生成从频率学派视角来看校准良好的贝叶斯推断这一根本问题。

受广义贝叶斯推断文献（Bissiri 等人，2016；Chernozhukov 和 Hong，2003年）的启发，我们设计了一种新型后验近似方法。在基于似然的情形下使用该方法时，如果模型设定正确，其表现与“精确”的贝叶斯方法等价；但无论模型设定如何，其生成的后验分布的可信集从频率学派的角度来看都是“有效的”。该方法不依赖于对后验抽样值进行任何临时校正，也不依赖其他外部校准技术；即使必须估计似然，也同样适用。我们从理论上证明，在较弱的正则条件下，即使似然需要估计，我们提出的方法也能无论模型设定如何，都提供有效的频率学派不确定性量化。因此，我们首次给出了一个在可能存在模型设定错误的模型中进行贝叶斯推断，同时还能提供有效（频率学派）不确定性量化的一般框架。

然而，与频率学派方法不同，我们的贝叶斯方法无需计算二阶导数信息即可提供有效的不确定性量化。因此，在难以获得二阶导数的情况下，例如包含许多潜变量的模型中，该方法为必须显式计算这些导数才能正确量化不确定性的频率学派方法提供了一种有用的替代方案。

当模型设定错误时，贝叶斯方法可能无法提供“适用”的推断。为规避这一问题，一些研究者建议采用广义贝叶斯程序，根据针对具体任务的损失函数生成后验分布；例如，参见 Syring 和 Martin (2020)、 Matsubara 等人 (2021)、Jewson 和 Rossell (2021)，以及 Loaiza-Maya 等人 (2021）等例子。不幸的是，正如

一些作者所讨论的，例如参见 Miller (2021）和 Syring 和 Martin (2019），广义贝叶斯后验分布并未校准良好：对于关注量，后验概率为 $(1-\alpha)$ 的可信集包含真实关注量的实际概率（根据真实数据生成过程计算）可能小于或大于 $(1-\alpha)$。已有若干方法被提出以解决这一问题；例如，参见 Holmes 和 Walker (2017），Syring 和 Martin (2019），另可参见 Wu 和 Martin (2020）对这些方法的综述。然而，据我们所知，现有方法无非是在贝叶斯学习算法的输出上施加外部且临时的修正。

我们证明，我们的方法可以直接适用于广义贝叶斯推断：在这一情境下，似然函数被通用损失函数所取代，或者被拟似然所取代，而拟似然至多只是似然的近似。由此得到一类用于广义贝叶斯推断的后验分布，能够产生校准良好的基于损失的可信集。因此，我们首次给出了一种广义贝叶斯后验分布，其可信集能够保证具有正确的宽度，而且不依赖于昂贵的后处理技术。

本文其余部分安排如下。第2 节讨论基于似然的贝叶斯推断中的模型错设这一一般性问题，并说明在这种情形下，某种特定的广义后验方法如何克服贝叶斯推断中已知的问题。第3 节将这一后验方法扩展到似然依赖不可观测潜变量的情形，并证明即使必须估计似然，我们提出的方法仍能给出校准良好的推断。第4节表明，在基于似然的贝叶斯框架中使用的广义后验构造，也适用于基于一般损失函数构建后验分布的情形（如Bissiri 等人，2016）。第5 节总结全文。所有结果的证明均见补充材料。

2. 模型设定错误下的贝叶斯推断

观测数据 $y_{1:n} = (y_1, \ldots, y_n)^\top$ 由某个真实但未知的分布 $P_0^{(n)}$ 生成，其中 $y_i \in \mathcal{Y} \subseteq \mathbb{R}^{d_y} (i = 1, \ldots, n)$。由于 $P_{(n)\,0}$ 未知，我们使用一类模型 $\{P_\theta^{(n)} : \theta \in \Theta\}$ 来近似 $y_{1:n}$ 的分布，该模型依赖于未知参数 $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$。我们假设存在一个测度，同时支配联合分布 $P_0^{(n)}$ 和 $P_\theta^{(n)}$，因此对所有 $n \geq 1$，联合密度 $p_0^{(n)}$ 和 $p_\theta^{(n)}$ 均存在。我们对 θ 的先验信念由概率密度函数 $\pi(\theta)$ 表示。观察到 $y_{1:n}$ 后，通过贝叶斯法则更新先验信念，得到精确后验分布

$$\pi(\theta \mid y_{1:n}) \propto \pi(\theta) \exp\{-\ell_n(\theta)\}, \text{ where } \ell_n(\theta) := -\log p_\theta^{(n)}(y_{1:n}). \tag{1}$$

当 $p_\theta^{(n)}(y_{1:n})$ 可以通过解析方法求得时，有许多方法可以从 $\pi(\theta \mid y_{1:n})$ 中获得样本。然而，在贝叶斯推断中，许多有趣的情形下，$p_\theta^{(n)}(y_{1:n})$ 难以通过解析方法求得。这通常发生在依赖不可观测变量（或称潜变量） $\alpha_{1:n} = (\alpha_1^\top, \ldots, \alpha_n^\top)^\top$ 的模型中，其中 $\alpha_{1:n} \sim p(\alpha_{1:n} \mid \theta)$，且 $\alpha_i \in \mathcal{A} \subseteq \mathbb{R}^{d_\alpha} (i = 1, \ldots, n)$。此时，观测数据似然 $p_\theta^{(n)}(y_{1:n})$ 可通过对完整数据似然 $p_\theta(y_{1:n}, \alpha_{1:n})$ 中的 $\alpha_{1:n}$ 进行积分得到：

$$p_\theta^{(n)}(y_{1:n}) = \int_{\mathcal{A}} p_\theta(y_{1:n}, \alpha_{1:n}) \mathrm{d}\alpha_{1:n}.$$

如果无法进行解析积分，通常仍可获得一个估计量 $\widehat{p}_\theta(y_{1:n} \mid z)$，该估计量依赖于模拟随机变量 $z = (z_1, \ldots, z_N)^\top \in \mathcal{Z}$，其中 N 表示用于构造 $\widehat{p}_\theta(y_{1:n} \mid z)$ 的“模拟抽样值”的数量。在马尔可夫链蒙特卡洛算法中使用 $\widehat{p}_\theta(y_{1:n} \mid z)$ 会得到一个伪边际算法（Andrieu and Roberts, 2009），其目标是关于 $(\theta^\top, z^\top)^\top$ 的联合后验分布：

$$\widehat{\pi}_N(\theta, z \mid y_{1:n}) = \widehat{p}_\theta(y_{1:n} \mid z) \pi(\theta) h(z \mid \theta),$$

其中， $h(z \mid \theta)$ 表示模拟数据 z 的条件密度。如果 $\int \widehat{p}_\theta(y_{1:n} \mid z)h(z \mid \theta)\mathrm{d}z = p_\theta^{(n)}(y_{1:n})$，边际后验 $\widehat{\pi}_N(\theta \mid y_{1:n}) = \int_{\mathcal{Z}} \widehat{\pi}_N(\theta, z \mid y_{1:n})\mathrm{d}z$ agrees with the exact posterior $\pi(\theta \mid y_{1:n})$（Andrieu and Roberts，2009）。

无论是否必须估计 $p_\theta^{(n)}$，当 $P_0^{(n)} \notin \{P_\theta^{(n)}) : \theta \in \Theta\}$ 后验分布推断通常都不是“校准良好”的。为了准确说明校准良好的推断的含义，首先定义 $\theta_\star := \mathrm{argmax}_{\theta\in\Theta} \lim_n \mathbb{E} \log\{p_\theta^{(n)}/p_0^{(n)}\}$，将其作为使从 $P_0^{(n)}$ 到 $\{P_\theta^{(n)}$ 的（极限）Kullback-Liebler散度最小的 $\theta \in \Theta$ 的值： $\theta \in \Theta\}$，其中 $\mathbb{E}(\cdot)$ 表示在 $P_0^{(n)}$ 下的期望算子。基于 $\pi(\theta \mid y_{1:n})$、针对 $\theta_\star$ 且具有后验概率 $(1-\alpha)$ 的可信集，如果该集合以 $(1-\alpha)$ 的 $P_0^{(n)}$-概率渐近地包含 $\theta_\star$，则称其为 校准良好 。这一定义可直接推广到 $\theta_\star$ 的任意函数。

为了说明为什么一般而言 $\pi(\theta \mid y_{1:n})$ 未能校准良好，我们需要先给出一些额外定义。对于二阶可微函数 $f : \Theta \to \mathbb{R}$，令 $\nabla_\theta f(\theta)$ 表示 $f(\theta)$ 的梯度， $\nabla^2_{\theta\theta} f(\theta)$ 表示其海森矩阵。对于式 (1) 中定义的 $\ell_n(\theta)$，令 $m_n(\theta) := -\nabla_\theta \ell_n(\theta)$ 和 $\mathcal{H}_n(\theta) := n^{-1}\nabla^2_{\theta\theta}\ell_n(\theta)$ 分别表示其梯度和（负的）海森矩阵；而 $\mathcal{H}(\theta) := \lim_n \mathbb{E}[\mathcal{H}_n(\theta_\star)]$ 表示期望观测信息，$\mathcal{I}(\theta) := \lim_n \mathrm{Cov}\{m_n(\theta)/\sqrt{n}\}$ 表示费希尔信息。

ˆ对于在经典正则条件下使 $0 = m_n(\theta)$ 成立的 $\theta \in \Theta$ 的取值，即最大似然估计量，参见 White（1982）或 Kleijn and van der Vaart（2012）。已知 $\sqrt{n}(\hat{\theta}_n - \theta_\star) \Rightarrow N(0, \Sigma_\star^{-1})$，其中 $\Sigma_\star^{-1} := \mathcal{H}(\theta_\star)^{-1}\mathcal{I}(\theta_\star)\mathcal{H}(\theta_\star)^{-1}$， $N(\mu, \Sigma)$ 表示均值为 μ、方差为 Σ 的高斯累积分布函数（CDF），而 $\Rightarrow$ 表示在 $P_0^{(n)}$ 下的弱收敛。此外，后验分布满足伯恩斯坦-冯·米塞斯结果

$$P_0^{(n)} \sup_{B\in\Theta} \left| \int_B \left[\pi(\vartheta \mid y_{1:n}) - N\{\vartheta; \hat{\theta}_n, (n\mathcal{H}_\star)^{-1}\} \right] \mathrm{d}\vartheta \right| \to 0,$$

其中 $N\{x;\ \mu,\ \Sigma\}$ 表示均值为 μ、方差为 Σ 的高斯概率密度函数（PDF）；例如，参见 Kleijnand van der Vaart (2012) 中的定理 2.1。伯恩斯坦-冯·米塞斯结果表明，基于 $\pi(\theta \mid y_{1:n})$ 构造的可信集的“宽度”由

$\mathcal{H}(\theta_\star)^{-1}$ 决定，而第一个结果表明，（渐近地）有效的频率学置信集的“宽度”由三明治协方差 $\Sigma_\star^{-1}$ 决定。因此，针对 $\theta_\star$ 的后验概率为（$1-\alpha$）的可信集，一般而言不会以$1-\alpha$概率（ $P_0^{(n)}$ ）包含 $\theta_\star$，除非 $\mathcal{H}(\theta_\star) = \mathcal{I}(\theta_\star)$。因此，后验可信集校准不良，贝叶斯不确定性量化可能过于乐观，以至于在实践中缺乏实用价值。

在设定错误的模型中，缺乏校准良好的可信集，促使研究人员考虑许多不同的方法来“纠正”这一问题。然而，据我们所知，所有被提出的方法要么依赖计算负担沉重的自助法程序，要么相当于明确假设后验分布具有高斯性，并通过事后校正 $\pi(\theta \mid y_{1:n})$ 来纠正覆盖率。第5.1 节讨论了几种此类方法。

2.1. 可靠的贝叶斯不确定性量化。如果我们愿意不再使用“精确”后验进行推断，而改用基于我们将要提出的某种近似似然的后验，那么就能得到渐近校准良好的贝叶斯可信集。为说明这一方法，考虑一个刻意简化的例子：观测数据为 Y_1，……，$Y_n \overset{iid}{\sim} P_0$，其中 $Y_i \in \mathbb{R}^d$，我们希望对未知均值 $\mathbb{E}[Y_i] = \theta_\star$进行推断。我们对未知的 $\theta_\star$有有意义的先验信念 $\pi(\theta)$，目标是在给定 $\pi(\theta)$以及 Y_1，……，Y_n的情况下进行后验推断。

虽然可以对 θ进行非参数贝叶斯推断，但对于如此简单的任务来说，这种做法似乎过于复杂。更简单的做法是注意到，尽管 P_0 未知，但样本均值$\overline{Y} = n^{-1}\sum_{i=1}^n Y_i$ 可以作为统计量，用于对 θ进行贝叶斯推断。遵循Wood (2010提出的合成似然（SL）方法，另见 Price 等人 (2018年）和Frazier 等人 (2022。尽管 P_0 未知，我们仍可使用一个已知分布来近似 $\overline{Y} \mid \theta$ 的分布，并将这一近似作为似然来进行后验推断。

对于一般的汇总统计量 S_n，Wood（2010）建议用 $N\{S_n$ 近似分布$S_n \mid \theta$； $b(\theta)$ $W_n(\theta)/n\}$ 其中 $b(\theta)$，而 $W_n(\theta)/n$ 分别是根据假定模型 $P_\theta^{(n)}$ 得到的 $S_n \mid \theta$ 的均值和（经缩放的）方差。[1] 尽管贝叶斯合成似然（BSL）最常用于处理不可处理的似然问题，但凡是希望对汇总量而非完整数据集进行后验推断的情形，都可以应用 BSL；关于这一点的进一步讨论，参见 Drovandi 等人（2021）。正如 Lewis 等人（2021）所指出的，在模型设定错误时，后验推断尤其适合以对模型错设具有“稳健性”的汇总统计量为条件，而非以整个样本（由 $\ell_n(\theta)$ 概括）为条件。事实上，当模型设定错误时，从不确定性量化的角度来看，精确后验并不具有“稳健性”；此时，或许可以基于一组汇总量进行贝叶斯推断，从而确保针对 θ 的后验推断能够对这种特定形式的模型错设具有“稳健性”。

注意到，平均得分 $n^{-1}m_n(\theta)$ 只是一个汇总统计量，在假定模型下其均值为零（在较弱的正则条件下），[2] 我们可以通过构造矩阵 $W_n(\theta)$（它是 Cov 的相合估计量）$\{m_n(\theta)/\sqrt{n}\}$，将合成似然逻辑应用于这一情形。具体而言，将 $n^{-1}m_n(\theta)$ 视为中心化汇总统计量后，我们可以遵循 Wood(2010) 的做法，使用方差为 $n^{-1}m_n(\theta) \mid \theta$ 的零均值高斯分布来近似 $n^{-1}W_n(\theta)$的分布，从而得到如下贝叶斯合成似然后验

$$\pi_n^Q(\theta) = \frac{M_n(\theta)^{-1/2}\exp\{-Q_n(\theta)\}\pi(\theta)}{\int_\Theta M_n(\theta)^{-1/2}\exp\{-Q_n(\theta)\}\pi(\theta)\mathrm{d}\theta}, \; Q_n(\theta) = \frac{1}{2}\frac{m_n(\theta)^\top}{\sqrt{n}}W_n(\theta)^{-1}\frac{m_n(\theta)}{\sqrt{n}}, \tag{2}$$

[1]对于汇总量 $\overline{Y}$，SL 建议使用近似似然 $N\{\overline{Y}$对 θ 进行贝叶斯推断；θ， $W_n(\theta)/n\}$ 其中 $W_n(\theta)$ 是 Cov($\sqrt{n}\overline{Y}$) 的估计量。[2]对于表示假定模型的 $P(n)_\theta$， $P_\theta^{(n)}m_n(\theta) = 0$ 对所有 $\theta \in \Theta$均成立，只要 $\nabla_\theta P(n)_\theta$ $\ell_n(\theta) = P(n)_\theta$ $\nabla_\theta \ell_n(\theta)$，对所有 $\theta \in \Theta$均成立。我们回顾一下，在以下条件下，积分与微分可以交换：1)$\ell_n(\theta)$ 是关于 $P_\theta^{(n)}$ 的 $y_{1:n}$ 的可积函数；2) 对于几乎所有的 $y_{1:n} \in \mathcal{Y}^{(n)}$，$\nabla_\theta \ell_n(\theta)$ 对所有 $\theta \in \Theta$均存在；3) 存在一个可积函数 $g(y_{1:n})$，使得 $|\nabla_\theta \ell_n(\theta)| \leq g(y_{1:n})$ 对所有$\theta \in \Theta$ 以及几乎所有 $y_{1:n} \in \mathcal{Y}^{(n)}$均成立。

其中 $M_n(\theta) = |W_n(\theta)|$，且符号 $\pi_n^Q(\theta)$ 表示后验分布对 $Q_n(\theta)$的依赖。[3] 关键在于，与标准 BSL 方法不同，式（2）中的后验分布不会导致信息损失，因为渐近地，$\ell_n(\theta)$ 中的信息与 $m_n(\theta)$中的信息相同。

虽然贝叶斯合成似然常用于不可处理模型，但在我们的情形中，精确后验分布 $\pi(\theta \mid y_{1:n})$ 是可处理的；不过，我们用统计量$W_n(\theta)^{-1/2}m_n(\theta)$中的信息替代完整样本中的信息，希望该统计量能够提供对模型设定错误具有稳健性的后验推断。后验分布 $\pi_n^Q(\theta)$ 相当于用 $Q_n(\theta)$中的信息替代通常的似然$\ell_n(\theta)$；该信息类似于得分函数 $m_n(\theta)$的某种（随机）二次型。因此，为了区分通常的贝叶斯合成似然后验与式（2）中的后验分布，在下文中我们将 $\pi_n^Q(\theta)$称为 Q-后验。

虽然 $Q_n(\theta)$类似于对数似然的二次近似，但这并不意味着 $\pi_n^Q(\theta)$ 会类似于高斯分布。更一般地说，$\pi_n^Q(\theta)$ 在参数定义于受限空间、且后验分布在支撑集边界附近具有较大概率质量时，仍然有意义。在此类情况下，对后验分布作高斯近似并无特别意义。由此，Q-后验有别于现有的、用于获得校准良好的贝叶斯推断的方法，因为它不依赖于对接受的抽样值进行后处理，也不依赖于额外重复执行抽样算法。正如第5节将更明确论证的那样，后处理步骤通常基于隐含或明确的正态性假设，因此在参数支撑集受限的情况下并无意义。此外，与旨在正确量化不确定性的频率学派方法不同，Q-后验无需估计二阶导数矩阵 $\mathcal{H}_n(\theta)$，即可正确量化不确定性。

以下评论将 Q-后验与文献中已有的几种方法联系起来。

注释 1. 后验分布 $\pi_n^Q(\theta)$ 与 Bissiri 等人(2016) 在（加权）二次损失函数下的一种广义后验形式相似。特别地，若有 $m_i(\theta) = \nabla_\theta$ 对数 $p(y_i \mid$

[3]这一分布近似将$n^{-1}m_n(\theta)|\theta$ 视为零均值高斯分布，其方差为$n^{-1}W_n(\theta)$。(2) 中的项 $Q_n(\theta)$ 是通过重新整理 $\{n^{-1}m_n(\theta)\}^\top\{n^{-1}W_n(\theta)\}^{-1}\{n^{-1}m_n(\theta)\} = \{n^{-1/2}m_n(\theta)\}^\top W_n(\theta)^{-1}\{n^{-1/2}m_n(\theta)\}$得到的。

$\theta)$, $m_{n,i}(\theta) = \{W_n(\theta)/n\}^{-1/2} m_i(\theta)$, and $q_{n,i}(\theta) = -\frac{1}{2}\|m_{n,i}(\theta)\|^2$, 我们有 $\exp\{-Q_n(\theta)\} = \exp\{\sum_{i=1}^n q_{n,i}(\theta)\}$，这与基于（加权）二次损失函数的广义后验相似。然而，我们的方法旨在生成具有适当不确定性量化的后验分布，而不是在贝叶斯推断中使用其他损失函数（讨论见第4节）。此外，与现有的广义贝叶斯方法不同，Q-后验无需艰难地选择调优常数；该常数可能会显著影响通过广义贝叶斯方法得到的后验不确定性量化结果。

注 2。Chernozhukov 和 Hong (2003年) 提出了一种广义贝叶斯推断方法，它基于一组针对 θ 的过度识别估计方程（方程数量多于未知参数），通过对样本矩向量取二次型来实现（另见Chib 等人，2018年）；或者完全用一个 M 估计量准则取代似然；后者也被 Bissiri 等人 (2016) 用于决策理论框架，以构造广义后验。然而，这两种方法都没有涵盖这样一种情况：定义后验分布的损失函数基于似然的得分方程。从理念上看，Chernozhukov 和 Hong (2003年) 和 Bissiri 等人 (2016) 的方法都是通过绕过似然来进行某种形式的贝叶斯推断，而不是在模型设定错误时利用似然进行贝叶斯推断。因此，从理念上说，两者截然不同。在结构上，两者也有所不同，因为式（2）中的后验分布与上述作者提出的后验分布形式不同，但彼此相关。

例1 (线性回归)。 考虑标准线性回归模型

$$y_i = x_i^\top \beta + \sigma\epsilon_i, \ (i = 1, \ldots, n),$$

其中误差分布被假定为 $\epsilon_i \overset{iid}{\sim} N(0, 1)$，且 x_i 和 β 是 d_β维向量。为简单起见，我们对这些参数采用弥散先验： $\pi(\beta) \propto 1$以及 $\pi(\sigma) \propto (\sigma^2)^{-2}$。虽然我们认为回归组成部分的模型设定是正确的，但我们对误差设定并不确定。真实数据生成

过程根据以下误差项设定生成观测数据： $\epsilon_i \overset{iid}{\sim} N(0, \sqrt{1 + |x_{1,\ i}|^{\gamma}})$，其中 $\gamma \geq 0$，而 $x_{1,\ i}$ 表示向量 ~~x_i 的第~~一个元素。当$\gamma = 0$ 时，具有同方差误差的假定模型被正确设定；而如果 $\gamma \neq 0$，则假定模型设定有误，精确后验很可能会产生过窄的可信集。

我们考虑在 $\gamma = 0$ 和 $\gamma = 2$ 的设计下，从模型生成 $n = 100$ 个观测值； x_i 生成为三元独立标准高斯随机向量，并将真实值设为 $\beta = (1, 1, 1)^{\top}$ 和 $\sigma = 1$。当模型设定错误时，可以证明 β 的伪真值保持不变，因此在本例中，我们仅关注对 β 的推断。我们将此设计重复 1000 次，并针对每个数据集使用吉布斯抽样获取精确后验分布，再使用具有高斯提议核的随机游走 Metropolis-Hastings（RWMH）对 Q-后验进行抽样；两种后验分布均使用 10000 个样本进行近似，并舍弃前 5000 个样本作为预热期。我们将 Q-后验中的矩阵 $W_n(\theta)$ 设为 $W_n(\theta) = n^{-1} \sum_{i=1}^{n} \{m_i(\theta) - n^{-1} m_n(\theta)\}\{m_i(\theta) - n^{-1} m_n(\theta)\}^{\top}$。对于每个数据集和方法，我们比较回归系数的后验偏差、方差和边际覆盖率。

表1 总结了结果，并表明 Q-后验得出的结果与精确贝叶斯的结果位置相近，但在两种情形下，其后验方差都更大。较大的后验方差使覆盖率接近名义上的 95%水平，且比精确后验的覆盖率接近得多。即使在同方差情形（$\gamma = 0$）下，精确后验的覆盖率也只有约 85%；而在异方差情形（$\gamma = 2$）下，最低覆盖率还不到 70%。

$\gamma=0$	Q-后验			精确		
	Bias	Var	Cov	Bias	Var	Cov
β_1	-0.0009	0.0395	0.9720	-0.0110	0.0104	0.8380
β_2	0.0188	0.0388	0.9680	-0.0049	0.0103	0.8370
β_3	-0.0212	0.0408	0.9710	-0.0103	0.0105	0.8540
$\gamma=2$	Q-后验			精确		
	Bias	Var	Cov	Bias	Var	Cov
β_1	-0.0072	0.0520	0.9590	-0.0215	0.0104	0.6860
β_2	-0.0265	0.0343	0.9660	-0.0196	0.0104	0.8380
β_3	-0.0027	0.0277	0.9680	-0.0082	0.0104	0.8440

表1. 精确方法和Q-后验在正态线性回归模型中的后验准确度结果。Bias表示重复试验中后验均值的偏差。Var表示重复试验中后验方差偏离的平均值。Cov表示后验覆盖率。实验的名义水平设为95%。

2.2. 自然形式下指数族模型的共轭性。考虑以下情形：Y_1、……、$Y_n \overset{iid}{\sim} p_\theta(y) = \exp\{\eta(\theta)^\top S(y) - A(\theta)\}h(y)$，其中 S： $\mathcal{Y} \to \mathbb{R}^{d_\theta}$ 是充分统计量向量，h： $\mathcal{Y} \to \mathbb{R}$ 是定义在 $\mathcal{Y}$ 上的参考测度，A： $\Theta \to \mathbb{R}$ 是对数配分函数。那么，联合密度 $p_\theta^{(n)}(y_{1:n})$ 的形式为

$$p_\theta^{(n)}(y_{1:n}) = \exp\left\{\eta(\theta)^\top \sum_{i=1}^n S(y_i) - nA(\theta)\right\} \prod_{i=1}^n h(y_i),$$

其中 $A(\theta) = \log\left[\int \exp\left\{\eta(\theta)^\top S(x)\right\} h(x)\mathrm{d}\mu(x)\right]$。

考虑我们的目标是对自然参数 $\eta = \eta(\theta)$ 进行推断。然后可以证明，只要我们对 η 的先验分布是共轭的，Q-后验就具有闭式表达式。具体而言，很容易看出，上述模型的平均得分为 $n^{-1}m_n(\eta) = \nabla_\eta A(\eta) - n^{-1}S_n$，其中 $S_n = \sum_{i=1}^n S(y_i)$。由于 $A(\eta)$ 是非随机的，因此可以使用样本方差 W_n 来估计 $n^{-1}m_n(\eta)$ 的方差：$= \frac{1}{n}\sum_{i=1}^n \{S(y_i) - n^{-1}S_n\}\{S(y_i) - n^{-1}S_n\}^\top$，该估计不依赖于 η。[4] 随后可以考虑使用 $\pi_n^Q(\eta) \propto \exp\{-Q_n(\eta)\}\pi(\eta)$ 对 η 进行推断，其中 $Q_n(\eta) = \frac{n}{2}\{\nabla_\eta A(\eta) - n^{-1}S_n\}^\top W_n^{-1}\{\nabla_\eta A(\eta) - n^{-1}S_n\}$.

通过函数 $\mu = g(\eta) = \nabla_\eta A(\eta)$ 定义均值形式参数 μ。在正则模型中，对于所有 η，函数 $g(\eta)$ 都存在且可逆。参数 $\mu = \mu(\eta)$ 称为模型的均值参数化，并满足 $\mu = \mathbb{E}_{Y\sim p_\theta^{(n)}(y_{1:n})}[S_n]$。其形式

[4] 更一般地，$\mathrm{Var}\{S(y_i)\}$ 的任何相合估计量也都可以使用。

因此， η 的 Q-后验可通过求出 μ 的 Q-后验并使用变量变换公式得到。

引理1。 考虑 $\mathcal{Y} = \mathbb{R}^d$。如果我们关于均值参数的（变换后的）先验信念$\mu = g(\eta)$ 可写为 $\pi(\mu) \propto \exp\{-\frac{1}{2}(\mu - \mu_0)^\top W_0^{-1}(\mu - \mu_0)\}$，那么$\eta$ 的 *Q*-后验为 $\pi_n^Q(\eta) = N\{g(\eta);\, b_n, \Sigma_n^{-1}\}|\nabla_\eta^2 A(\eta)|$，其中

$$\Sigma_n^{-1} = n^{-1}W_0\left[n^{-1}W_n + W_0\right]^{-1}W_n,$$

$$b_n = W_0\left[n^{-1}W_n + W_0\right]^{-1}\frac{1}{n}\sum_{i=1}^{n} S(y_i) + W_0\left[n^{-1}W_n + W_0\right]^{-1}\frac{\mu_0}{n}.$$

引理 1 表明，对于标准形式的指数族，自然参数的 Q-后验是一个变换后的高斯密度。[5] 有趣的是， $|\nabla_\eta^2 A(\eta)|$的计算可以完全避免：先对 $\widetilde{\mu} \sim N\{\mu;\, b_n, \Sigma_n^{-1}\}$进行抽样，然后（数值地）对方程 $\widetilde{\mu} = g(\eta)$ 求逆，从而得到抽样值 $\widetilde{\eta}$。在任何能够可靠计算$g(\eta) = \nabla_\eta A(\eta)$ 的情况下，都可以进行后一步。

3. 估计似然

Q-后验 $\pi_n^Q(\theta)$ 在 (2) 中要求 $m_n(\theta) = \nabla_\theta \ell_n(\theta)$，且 $W_n(\theta)$ 已知到θ。然而，在许多有趣的情形中，观测数据似然只能表示为

$$p_\theta^{(n)}(y_{1:n}) = \int_{\mathcal{A}} p_\theta^{(n)}(y_{1:n}, \alpha_{1:n})\mathrm{d}\alpha_{1:n},$$

其中， $p_\theta^{(n)}(y_{1:n}, \alpha_{1:n})$ 称为完整数据似然， $\alpha_{1:n} \sim p(\alpha_{1:n} \mid \theta)$ 和 $\alpha_i \in \mathcal{A} \subseteq \mathbb{R}^{d_\alpha}(i = 1, \ldots, n)$ 均不可观测。在这种情况下， $p_\theta^{(n)}(y_{1:n})$ 通常难以处理，因此我们无法得到 $m_n(\theta)$ 和 $W_n(\theta)$ 的解析形式或计算上易于处理的形式。此外，尽管可能可以得到似然的一个估计量，例如 $\widehat{p}_\theta(y_{1:n} \mid z)$，但对 $\widehat{p}_\theta(y_{1:n} \mid z)$ 求导以获得 $m_n(\theta)$ 的估计量可能并不可取；在许多情况下， $\widehat{p}_\theta(y_{1:n} \mid z)$ 关于 θ 不可微，粒子滤波方法就是如此。

[5]只要均值参数化下的（变换后的）先验分布 $\pi(\mu)$ 可以写成高斯核即可。

因此，我们必须以某种方式近似 $m_n(\theta)$，以便将 Q-后验推广到必须估计 $p_\theta^{(n)}(y_{1:n})$ 的情形。

Fisher 恒等式（Cappé 等人，2005年）描述了观测对数似然的梯度 $\log p_\theta^{(n)}(y_{1:n})$与完整数据对数似然 $p_\theta(y_{1:n}, \alpha_{1:n})$之间的关系，因此我们可以将 $m_n(\theta) = \nabla_\theta \log p_\theta^{(n)}(y_{1:n})$ 表示为

$$m_n(\theta) = -\nabla_\theta \ell_n(\theta) = \int \{\nabla_\theta \log p_\theta(y_{1:n}, \alpha_{1:n})\} p_\theta(\alpha_{1:n} \mid y_{1:n}) \mathrm{d}\alpha_{1:n},$$

其中，$p_\theta(\alpha_{1:n} \mid y_{1:n})$ 是在给定 θ 的条件下 $\alpha_{1:n}$ 的后验分布。给定抽样值 $\alpha_1^{(j)}{}_{:n} \overset{iid}{\sim} p_\theta(\alpha_{1:n} \mid y_{1:n})$ $(j = 1, \ldots, N)$，我们可以使用简单的蒙特卡洛估计量来估计 $m_n(\theta)$

$$\widehat{m}_n(\theta; z) := \frac{1}{N} \sum_{j=1}^{N} \nabla_\theta \log p_\theta(y_{1:n}, \alpha_{1:n}^{(j)}), \quad \alpha_{1:n}^{(j)} \sim p_\theta(\alpha_{1:n} \mid y_{1:n}),$$

其中，z 再次表示构造 $\widehat{m}_n(\theta; z)$ 所需的所有模拟变量。要获得 $\widehat{m}_n(\theta; z)$，进而得到 $W_n(\theta)$ 的估计量，我们只需能够从 $p_\theta(\alpha_{1:n} \mid y_{1:n})$ 中生成样本。[7]

从 $p_\theta(\alpha_{1:n} \mid y_{1:n})$生成抽样值的能力并无特别严格的限制，因为我们不要求 $p_\theta(\alpha_{1:n} \mid y_{1:n})$具有闭式表达式。正如以下示例所示，在许多情况下都可行，例如状态空间模型和广义随机效应模型。

示例 2. 在状态空间模型中，完整数据似然 $p_\theta(\alpha_{1:n}, y_{1:n})$通常可以写成

$$p_\theta(\alpha_{1:n}, y_{1:n}) = \mu(\alpha_1) \prod_{t=2}^{n} f_\theta(\alpha_t \mid \alpha_{t-1}) \prod_{t=2}^{n} g_\theta(y_t \mid \alpha_t),$$

其中，$f_\theta(\alpha \mid \alpha')$ 是某个不可观测（状态）过程 $\{\alpha_t : t \le n\}$ 的转移密度；$g_\theta(y_t \mid \alpha_t)$ 是 $y_t \mid \alpha_t$ 的条件密度，而 $\mu(\alpha)$ 是状态的不变测度；并且，给定 $\{\alpha_t : t \le n\}$ 时，过程 $\{y_t : t \le n\}$ 条件独立。

[6]为简化记号，我们省略估计量对 N的依赖，因为稍后会将 N 视为 n的函数。[7]虽然可以得到更有效的 $m_n(\theta)$ 估计量，例如基于重要性抽样的估计量，但简单的蒙特卡罗估计量既有效，又便于进行理论分析。我们推测可以使用更有效的估计量，但将对此类估计量的正式研究留待未来开展。

状态的马尔可夫性意味着可以使用滤波和平滑方法，从 $p_\theta(\alpha_{1:n} \mid y_{1:n})$ 中生成抽样值。此外，在许多情况下，例如线性高斯模型和某些随机波动率模型，可以使用 Gibbs 或 MH采样算法从$p_\theta(\alpha_{1:n} \mid y_{1:n})$ 中获得抽样值。无论如何获得 $p_\theta(\alpha_{1:n} \mid y_{1:n})$ 的抽样值，都可以使用 $m_n(\theta)$ 进行估计

$$\widehat{m}_n(\theta; z) = N^{-1} \sum_{j=1}^{N} \sum_{t=2}^{n} \nabla_\theta \log g_\theta(y_t \mid \alpha_t^{(j)}) + N^{-1} \sum_{j=1}^{N} \sum_{t=2}^{n} \nabla_\theta \log f_\theta(\alpha_t^{(j)} \mid \alpha_{t-1}^{(j)}),$$

对于 $\alpha_{1:n}^{(j)} \sim p_\theta(\alpha_{1:n} \mid y_{1:n})$。此外，给定 α 和 α'的取值，通常可以用闭式计算 $\nabla \log f_\theta(\alpha \mid \alpha')$ 和 $\nabla_\theta \log g_\theta(y \mid \alpha)$。

例 3. 随机效应模型在聚类数据和纵向数据分析中无处不在。在这类模型中，结果变量 $y_{i,j}$、$i = 1, \ldots, n$ 以及 $j = 1, \ldots, J$受协变量向量 $x_{i,j}$ 和不可观测随机效应 α_i的影响。在大多数情况下，给定尺度参数 τ后，假设随机效应独立同分布，因此其联合分布为 $f_\tau(\alpha_{1:n}) = \prod_{i=1}^{n} f_\tau(\alpha_i)$，其中 $f_\tau(\cdot)$ 是均值为零的分布，其分布形式已知，但参数 τ未知。给定 $\alpha_{1:n}$ 和 $x_{1:n,j}$后，j的结果分布为$g_\theta(y_{1:n,j} \mid \alpha_{1:n}, x_{1:n,j}) = \prod_{i=1}^{n} g_\theta(y_{i,j} \mid \alpha_i, x_{i,j})$，其中 $g_\theta(y_{i,j} \mid \alpha_i, x_{i,j})$ 属于指数族，且取决于线性指标 $\eta_{i,j} = \alpha_i + x_{i,j}^\top \beta$；$\beta$是回归参数向量，因此 $\theta = (\beta^\top, \tau)^\top$。完整数据似然同样具有乘积形式；给定来自 $\alpha_{1:n} \mid y_{1:n,1:J}, x_{1:n,1:J}, \theta$的后验抽样值后，可像之前一样构造估计量 $\widehat{m}_n(\theta; z)$。

Fisher 恒等式和简单蒙特卡罗方法使我们能够通过对来自 $p_\theta(\alpha_{1:n} \mid y_{1:n})$ 的N个抽样值取平均，近似计算 $m_n(\theta)$，其中，对于每个 $j = 1, \ldots, N$，第 j 个抽样值取决于随机变量 z_j。将所有这些随机变量串联为 z 后，我们将得分方程的模拟估计量 $m_n(\theta)$ 记为 $\widehat{m}_n(\theta; z)$，并将 $\widehat{m}_n(\theta; z)$ 的方差估计量记为 $W_n(\theta; z)$。这些函数对 z 的依赖方式与伪边际文献类似；在该文献中，似然估计量 $\widehat{p}_\theta(y_{1:n} \mid z)$ 也被视为依赖于模拟变量 z。

给定 $\widehat{m}_n(\theta; z)$ 和 $W_n(\theta; z)$ 后，可以使用例如 RWMH-马尔可夫链蒙特卡洛方案对 θ 进行后验推断，该方案用估计版本替代难处理的 $Q_n(\theta)$

$$\widehat{Q}_n(\theta; z) = \frac{1}{2}\widehat{m}_n(\theta; z)^\top \widehat{W}_n(\theta; z)^{-1}\widehat{m}_n(\theta; z).$$

参见算法 1。

算法 1 使用 Q 函数的伪边际 Metropolis–Hastings 马尔可夫链蒙特卡罗（MH-MCMC-Q）

初始化 $\theta^{(0)}$ 以及 $z^{(0)}$。
对于 $i = 1, \ldots, M$ **do**
  抽取 $\theta^* \sim q(\theta|\theta^{i-1})$
  从提议分布 $h(z \mid \theta^*)$中抽取 z^*
  计算 $V^* = \widehat{g}_n(\theta^*; z^*)$:= $\widehat{M}_n(\theta; z)^{-1/2} \exp\{-\widehat{Q}_n(\theta^*; z^*)\}$ 和 $V^{i-1} = \widehat{g}_n(\theta^{i-1}; z^{i-1})$
  计算 Metropolis-Hastings 比率： $r = \alpha(\theta^*, \theta^{(i-1)}) = \frac{V^*\pi(\theta^*)q(\theta^{i-1}|\theta^*)}{V^{i-1}\pi(\theta^{i-1})q(\theta^*|\theta^{i-1})}$
  如果 $\mathcal{U}(0,1) < r$ **则**
    设定 $\theta^i = \theta^*$, $z^i = z^*$
  else
    设定 $\theta^i = \theta^{i-1}$, $z^i = z^{i-1}$
  结束如果
结束循环

3.1. 后验目标与大样本中的行为。由于需要通过 $Q_n(\theta; z)$ 估计 $Q_n(\theta)$，并在马尔可夫链蒙特卡洛算法中用 $\exp\{-Q_n(\theta; z)\}$ 替代难处理的“似然项” $\exp\{-Q_n(\theta)\}$（位于 $\widehat{\ }$ 内），所得算法并不会从 $\pi_n^Q(\theta)$ 中生成抽样值；若 $m_n(\theta)$ 易于处理，这正是我们希望作为目标的后验分布。根据 Andrieu 和 Roberts（2009）的结果，算法 1 可视为在 $\Theta \times \mathcal{Z}$ 上运行的精确 Metropolis–Hastings算法，其平稳分布为

$$\overline{\pi}_n^Q(\theta, z) \propto \widehat{g}_n(\theta; z)\pi(\theta)h(z \mid \theta), \quad \widehat{g}_n(\theta; z) := \widehat{M}_n(\theta; z)^{-1/2} \exp\{-\widehat{Q}_n(\theta; z)\},$$

其中 $M_n(\theta; z) = |W_n(\theta; z)|$。因此，算法 1 所得到的边际后验分布为

$$\overline{\pi}_n^Q(\theta) \propto \int_{\mathcal{Z}} \widehat{g}_n(\theta; z)\pi(\theta)h(z \mid \theta)\mathrm{d}z.$$

不同于标准伪边际方法，$\widehat{g}_n(\theta; z)$ 是后验核 $M_n(\theta)^{-1/2}\exp\{-Q_n(\theta)\}$的有偏估计量。因此，在马尔可夫链蒙特卡洛算法中使用 $\widehat{g}_n(\theta; z)$ 会导致所得后验分布与式（2）中的 $\pi_n^Q(\theta)$ 不一致。

后验分布 $\overline{\pi}_n^Q(\theta)$ 只能表示为一个难处理的积分，因此，似乎无法对固定的 n 和 N 下其行为得出任何具体结论。不过，只要 $N \to +\infty$，我们就能确保 $\overline{\pi}_n^Q(\theta)$ 与 $\pi_n^Q(\theta)$ 一致地足够接近，其中 $N \to +\infty$ 对任意 $n \geq 1$ 均成立。

3.1.1. 主要结果。 要控制 $\overline{\pi}_n^Q(\theta)$ 的行为，我们必须控制 $\widehat{Q}_n(\theta; z)$ 的行为。$\widehat{\ }$在对 $\widehat{m}_n(\theta; z)$ 和 $W_n(\theta; z)$ 施加某些矩假设的条件下，我们可以实现这种控制。

假设 1。 存在一个函数 $\sigma_n(\theta)$：$\Theta \to \mathbb{R}_+$，使得对于所有 $n \geq 1$ 以及 $j = 2, 4$，以下条件均成立。

(1) $\mathbb{E}_z[\|\{\widehat{m}_n(\theta; z) - m_n(\theta)\}/\sqrt{n}\|^j] \leq \sigma_n^j(\theta)/N$。*(2)* $\mathbb{E}_z[\|\widehat{W}_n(\theta; z) - W_n(\theta)\|^j] \leq \sigma_n^j(\theta)/N$。*(3)* 对于 $\|\cdot\|$，取适当的矩阵范数，令某些 $C > 0$、N 足够大，并对所有 $\theta \in \Theta$、$\|W_n(\theta)^{-1}\{\widehat{W}_n(\theta; z) - W_n(\theta)\}\| \leq C/N$，且 $\sup_{\theta\in\Theta} \|\sigma_n^2(\theta)W_n(\theta)^{-1}\| < C$。

假设 1 对用于 $m_n(\theta)$和 $W_n(\theta)$的估计量的行为作出了限制。假设 1(1-2)要求基于模拟的估计量在 z 中具有二阶矩，且这些矩由某个函数 $\sigma_n^2(\theta)$所界定。假设 1(3)给出了不可行的 $W_n(\theta)$（即得分函数的协方差）与界定其矩的函数 $\sigma_n^2(\theta)$之间的关系。这类假设类似于求解线性系统时进行扰动分析所需的假设（Horn和Johnson，2012，第 5.8 章）。这些$\widehat{\ }$限制共同使我们能够对 $Q_n(\theta; z)$的均值和方差给出上界。由于我们尚未发现能对随机加权矩阵下二次型的均值和方差给出界的一般性结果，因此我们在附录 A 中给出了这样一个一般性结果，即引理 6；该结果本身也可能具有独立的研究价值。

定理1。 对于所有 $n \geq 1$，假设 $\pi_n^Q(\theta)$ 存在，并且对于某个已知函数 φ： $\Theta \to \mathbb{R}^{d_\varphi}$，假设 $\int_\Theta \|\varphi(\theta)\|\pi_n^Q(\theta)\mathrm{d}\theta < \infty$，且 $\int_\Theta \|\varphi(\theta)\|\sigma_n^4(\theta)\pi(\theta)\mathrm{d}\theta < \infty$。若假设 *1*

成立，则当 $N \to +\infty$

$$\int_{\Theta} |\pi_n^Q(\theta) - \overline{\pi}_n^Q(\theta)| \mathrm{d}\theta = O(1/N), \ and \ \left\| \int_{\Theta} \varphi(\theta)\pi_n^Q(\theta)\mathrm{d}\theta - \int_{\Theta} \varphi(\theta)\overline{\pi}_n^Q(\theta)\mathrm{d}\theta \right\| = O(1/N).$$

定理1表明，对于 $N \to +\infty$，难处理的后验分布 $\pi_n^Q(\theta)$ 与其“估计版本” $\overline{\pi}_n^Q(\theta)$ 之间的差异以 $1/N$ 的速率收敛于零。这并非一个“依概率收敛”的结果，并且对于任意固定的 n 都成立，如 $N \to +\infty$ 所示。此外，对于满足 $\| \int_{\Theta} \varphi(\theta)\pi_n^Q(\theta)\mathrm{d}\theta \| < \infty$ 的 θ 的任意函数，这种等价性都成立：根据 $\overline{\pi}_n^Q(\theta)$ 计算的后验矩，与基于难处理的 $\pi_n^Q(\theta)$ 计算的后验矩，在 $O(1/N)$ 阶上等价。因此，当 N 较大时，定理1意味着 $\overline{\pi}_n^Q(\theta)$ 及其矩 $\int_{\Theta} \varphi(\theta)\overline{\pi}_n^Q(\theta)\mathrm{d}\theta$ 的表现恰如 $m_n(\theta)$ 和 $W_n(\theta)$ 已知一般。

直观地说，定理1 意味着，当 N 较大时，基于可行后验分布$\overline{\pi}_n^Q(\theta)$的推断与不可行后验分布 $\pi_n^Q(\theta)$ 的推断一致接近；如果似然在解析上易于处理，我们本希望以该不可行后验分布为目标。定理1的一个直接推论是，如果 $\pi_n^Q(\theta)$ 的可信集校准良好，那么基于 $\overline{\pi}_n^Q(\theta)$的可信集也将校准良好，即使为了进行可行的推断，必须对似然进行估计。

3.1.2. 不确定性量化。如果我们愿意对 $m_n(\theta)$和 $W_n(\theta)$的行为作出额外假设，就可以正式证明，无论模型设定正确与否，$\overline{\pi}_n^Q(\theta)$和 $\pi_n^Q(\theta)$的可信集都渐近地校准良好。重要的是，与针对含有潜变量的模型的频率学派方法不同，基于$\overline{\pi}_n^Q(\theta)$得到可信集不需要计算二阶导数信息。

在介绍正则条件和结果之前，我们先列出一些用于简化结果表述的记号。对于正数序列 $a_n \to +\infty$ as $n \to +\infty$，如果序列 $a_n X_n$ 依概率收敛至零，我们记作 $X_n = o_p(a_n^{-1})$；而记号 $X_n = O_p(a_n^{-1})$表示 $a_n X_n$ 依概率有界。对于集合 $A \subseteq \mathbb{R}^d$，令 Int(A) 表示 A的内部。记号 $\Rightarrow$ 表示在 $P_0^{(n)}$ 下的弱收敛。

假设 2. 存在一个函数 $m : \Theta \to \mathbb{R}^{d_\theta}$，使得：

(1) $\sup_{\theta\in\Theta} \|n^{-1}m_n(\theta) - m(\theta)\| = o_p(1)$。 *(2)* 存在唯一的 $\theta_\star \in Int(\Theta)$，使得 $m(\theta) = 0$ 当且仅当 $\theta = \theta_\star$。 *(3)* 对于某个 $\delta > 0$， $m(\theta)$ 在 $\|\theta - \theta_\star\| \le \delta$ 上连续可微，且 $\mathcal{H}(\theta_\star) := -\nabla_\theta m(\theta_\star)$ 可逆。 *(4)* 存在一个半正定矩阵 $W_\star$，使得 $m_n(\theta_\star)/\sqrt{n} \Rightarrow N(0, W_\star)$。 *(5)* 对于任意 $\epsilon > 0$，都存在一个 $\delta > 0$，使得

$$P_0^{(n)}\left(\sup_{\|\theta-\theta'\|<\delta} \frac{\|n^{-1}\{m_n(\theta) - m_n(\theta')\} - \{m(\theta) - m(\theta')\}\|}{1 + \sqrt{n}\|\theta - \theta'\|} > \epsilon\right) = o(1).$$

我们要求加权矩阵满足以下假设。

假设*3*。对于某个 $\delta > 0$，满足以下条件：*(i)* 当 n 足够大时，矩阵 $W_n(\theta)$ 在 Θ 上一致地为半正定且对称，并且对所有 $\|\theta - \theta_\star\| \le \delta$ 均为正定；*(ii)* 存在某个矩阵 $W(\theta)$，其在 Θ 上一致地为半正定且对称，并且满足 $\sup_{\theta\in\Theta} \|W_n(\theta) - W(\theta)\| = o_p(1)$，且对于所有 $\|\theta - \theta_\star\| \le \delta$， $W(\theta)$ 连续且正定；*(iv)* $W(\theta_\star) = W_\star$；*(v)* 对任意 $\epsilon > 0$， $\sup_{\|\theta-\theta_\star\|\ge\epsilon} -m(\theta)^\top W(\theta)^{-1} m(\theta) < 0$。

注释 3. 假设 2和 3 共同保证不可行准则 $Q_n(\theta)$的光滑性和识别条件。 这些条件允许 $Q_n(\theta) - Q_n(\theta_\star)$在 θ附近 $\theta_\star$存在二次展开，且该展开在该处光滑，余项也能得到适当控制。具体而言，我们的假设不要求损失函数处处可微，只要求其等度连续。后者是一个重要区别，因为许多损失函数仅几乎处处可微，例如基于绝对值函数的损失函数。

以下假设要求某些先验矩存在 .

假设4。 对于 $\theta_\star$ 如假设 2中所定义， $\pi(\theta_\star) > 0$ 和 $\pi(\theta)$ 在 Θ上连续。对于某个 $p \ge 1$， $\int_\Theta \|\theta\|^p \pi(\theta) \mathrm{d}\theta$，且对于所有足够大的 n $\int_\Theta |W_n(\theta)|^{-1/2} \|\theta - \theta_\star\|^p \pi(\theta) \,\mathrm{d}\theta < +\infty$。

以下结果表明，$\overline{\pi}_n^Q(\theta)$ 能够提供有效的（频率学派）不确定性量化。在陈述该结果之前，回顾一下 $\Sigma_\star := \mathcal{H}(\theta_\star)W_\star^{-1}\mathcal{H}(\theta_\star)$，并定义

$$Z_n := \Sigma_\star^{-1}\mathcal{H}(\theta_\star)^\top W_\star^{-1} m_n(\theta_\star), \quad \vartheta := \sqrt{n}(\theta-\theta_\star) - Z_n/\sqrt{n}$$

以及 $\mathcal{T}_n := \{\vartheta = \sqrt{n}(\theta-\theta_\star) - Z_n/\sqrt{n} : \ \theta\in\Theta\}$。

引理2。 在假设 *1-4* 下，当 $n,\ N\to+\infty$，$\int_{\mathcal{T}_n}\|\vartheta\|\,|\overline{\pi}_n^Q(\vartheta) - N\{\vartheta;\ 0,\ \Sigma_\star^{-1}\}|\mathrm{d}\vartheta = o_p(1)$。

以下结果是引理 2 的推论。

引理 3。 在假设 *1-4* 下，若 $\sqrt{n}/N\to 0$，对于 $\overline{\theta} = \int_\theta \theta\overline{\pi}_n^Q(\theta)\mathrm{d}\theta$，$\sqrt{n}\{\overline{\theta}-\theta_\star\} \Rightarrow N\{0,\Sigma_\star^{-1}\}$.

引理 2-3 表明，即使必须估计似然，只要能够从 $p_\theta(\alpha_{1:n}\mid y_{1:n})$ 中模拟抽样值，我们就能利用$\overline{\pi}_n^Q(\theta)$ (对于 $N\to+\infty$)，准确量化不确定性。然而，如果我们希望后验均值不会渐近有偏，由于需要模拟数据，我们要求 $\sqrt{n}/N = o(1)$。后一条件与模拟最大似然估计中渐近高斯点估计所要求的模拟抽样值数量条件相同（早期研究参见Shephard和Pitt, 1997 ）。因此，只要抽样值数量适当增长，在任何可应用滤波方法的情形下，都可以使用 Q-后验得出校准良好的贝叶斯推断。

3.1.3. 方差估计。 为了使 Q-后验校准良好，需要一个一致的估计量来估计 $W_\star = \lim_n$ 协方差 $\{m_n(\theta_\star)/\sqrt{n}\}$。在 $m_n(\theta)$ 具有如下形式的情况下：$m_n(\theta) = \sum_{i=1}^n m_i(\theta)$，其中某些已知函数 $m_i(\theta)$ 仅取决于第 i 个样本单位，矩阵 $W_n(\theta)$ 可取为

$$W_n(\theta) = n^{-1}\sum_{i=1}^n \left\{m_i(\theta) - n^{-1}m_n(\theta)\right\}\left\{m_i(\theta) - n^{-1}m_n(\theta)\right\}^\top. \tag{3}$$

如果担心 $W_n(\theta)$ 在 (3) 中可能不是 $W_\star$的一致估计量，我们建议用户采用自助法技术来估计协方差矩阵；这通常可以通过某种变体的估计方程自助法实现（Hu 和 Kalbfleisch, 2000年 和Chatterjee 和 Bose，2005年）。这种方法既可以直接嵌入
¯ ¯马尔可夫链蒙特卡洛算法中，也可以使用基于后验分布高概率密度区域内某个 $W_n(\theta_n)$的初步估计量 θ_n ，并在整个马尔可夫链蒙特卡洛算法中使用。[8]

对于 Q-后验， $\overline{\pi}_n^Q(\theta)$，选择 $\widehat{W}_n(\theta; z)$ 作为（3）中的样本方差，一般可能无法得到校准良好的 Q-后验。即使在$p_\theta(y_{1:n}, \alpha_{1:n}) = \prod_{i=1}^n g_\theta(y_i|\alpha_i) f_\theta(\alpha_i)$的情况下，也很容易看出这一问题，因此 $\widehat{m}_n(\theta; z)$ 呈现为

$$\widehat{m}_n(\theta; z) = \sum_{i=1}^n N^{-1} \sum_{j=1}^N \nabla_\theta \log g_\theta(y_i|\alpha_i^{(j)}) f_\theta(\alpha_i^{(j)}), \quad \alpha_{1:n}^{(j)} \sim p_\theta(\alpha_{1:n} \mid y_{1:n}),$$

我们注意到，它在 y_i 和 $\alpha_i^{(j)}$ 两方面都存在变异。为了说明这两部分变异都很重要，令 $m_{ij}(\theta) = \nabla_\theta \log g_\theta(y_i \mid \alpha_i^{(j)}) f_\theta(\alpha_i^{(j)})$， $\overline{m}_{iN}(\theta) = \frac{1}{N}\sum_{j=1}^N m_{ij}(\theta)$， $z_i = (\alpha_i^{(1)}, \ldots, \alpha_i^{(N)})$，并考虑

$$\mathrm{Var}\{\overline{m}_{iN}(\theta)\} = \mathbb{E}_Y\left[\mathrm{Var}_{Z|Y}\{\overline{m}_{iN}(\theta)|Y\}\right] + \mathrm{Var}_Y\left[\mathbb{E}_{Z|Y}\{\overline{m}_{iN}(\theta)|Y\}\right].$$

由于 $\alpha_{1:n} \overset{iid}{\sim} p_\theta(\alpha_{1:n} \mid y_{1:n})$， $\mathbb{E}_{Z|Y}\{\overline{m}_{iN}(\theta)|Y\} = \mathbb{E}_{Z|Y}\{m_{ij}(\theta)|Y\}$，而这个期望可以用 $\overline{m}_{iN}(\theta) = \frac{1}{N}\sum_{j=1}^N m_{ij}(\theta)$ 估计。然而，对 $\overline{m}_{iN}(\theta)$ 应用通常的样本协方差估计量可得

$$\widehat{W}_{1n}(\theta; z) := n^{-1} \sum_{i=1}^n \left\{\overline{m}_{iN}(\theta) - n^{-1}\widehat{m}_n(\theta; z)\right\}\left\{\overline{m}_{iN}(\theta; z_i) - n^{-1}\widehat{m}_n(\theta; z)\right\}^\top.$$

直观而言， $\widehat{W}_{1n}(\theta; z)$只是 $\mathrm{Var}_Y\left[\mathbb{E}_{Z|Y}\{m_{i,j}(\theta)|Y\}\right]$ 的一个估计量，因为模拟的 $\alpha_1^{(j)}$: $_n \sim p_\theta(\alpha_{1:n} \mid y_{1:n})$ 的影响已通过对 $\overline{m}_{iN}(\theta)$ 求平均而被“积分消去”。因此，除非对所有 $i = 1, \ldots, N$， $\mathrm{Var}_{Z|Y}\{\overline{m}_{iN}(\theta)|Y\} = 0$，否则通常的样本协方差

[8]后者成立，是因为假设3 实际上比保证ˆ ˆ引理2有效性所需的条件更强。可以证明，若将矩阵 $W_n(\theta)$ 替换为 $W_n(\theta_n)$，其中 θ_n 是 $\theta_\star$的任意初步相合估计量。

用于 $\widehat{m}_n(\theta; z)$ 的估计量会低估方差，并给出过窄的可信集。

然而，鉴于 $\widehat{W}_{1n}(\theta;\ z)$ 是 $\text{Var}_Y\left[\mathbb{E}_{Z|Y}\{\overline{m}_{iN}(\theta)|Y\}\right]$ 的估计量，剩下的就是求出 $\mathbb{E}_Y\left[\text{Var}_{Z|Y}\{\overline{m}_{iN}(\theta)|Y\}\right]$ 的估计量。由于 $\alpha_1^{(j)}:{}_n \overset{iid}{\sim} p_\theta(\alpha_{1:n} \mid y_{1:n})$，一个显而易见的选择是

$$\widehat{W}_{2n}(\theta; z) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{N}\sum_{j=1}^{N}\{m_{ij}(\theta) - \overline{m}_{iN}(\theta)\}\{m_{ij}(\theta) - \overline{m}_{iN}(\theta)\}^\top\right],$$

这里我们注意到，方括号中的项是 $\text{Var}_{Z|Y}\{m_{ij}(\theta)|Y\}$ 的估计量。关键在于，由于$\alpha_1^{(j)}:{}_n \overset{iid}{\sim} p_\theta(\alpha_{1:n}|y_{1:n})$ $(j = 1, \ldots, N)$，根据通常的大数定律，$\widehat{W}_{2n}(\theta;\ z)$ 是相合的。即使序列 $\{\overline{m}_{iN}(\theta):\ i \le n\}$ 存在依赖性，这一后者的性质仍然成立。

然而，我们确实注意到，如果 $\{\overline{m}_{iN}(\theta):\ i \le n\}$ 是相关的，那么就需要为 $\widehat{W}_{1n}(\theta;\ z)$ 使用另一种估计量。可以通过将 $W_{1n}(\theta; z)$ 替换为某种异方差和自相关一致（HAC）协方差估计量（Newey 和 West，1987），或对序列 $\{\overline{m}_{iN}(\theta):\ i \le n\}$ 使用自助法来获得此类估计量。例如，对于相关数据，可以对“数据” $\{\overline{m}_{iN}(\theta):\ i \le n\}$ 使用野生自助法（Shao，2010）或 Politis 和 Romano（1994）的平稳自助法，以获得方差的相合估计量。[9]

3.2. **示例：随机效应模型。**

3.2.1. 线性随机效应模型。

为了展示我们针对含潜变量的贝叶斯推断所提出的广义方法的优势，我们重新考察示例 1中的简单线性回归模型，并证明在含潜变量的情形下，我们的方法同样表现出色。为此，我们考虑带有随机效应的标准线性回归模型

$$y_i = \alpha_i + x_i^\top\beta + \sigma\epsilon_i,\ (i = 1, \ldots, n),$$

[9]或者，可以使用无浪费序贯蒙特卡罗采样器 $W(\theta)$ 估计矩阵Dau 和 Chopin (2022)，这样只需运行一次算法，就能得到粒子估计量 $\widehat{m}_n(\theta; z)$渐近方差的相合估计量。

其中我们假设误差分布为 $\epsilon_i \overset{iid}{\sim} N(0,1)$，$x_i$和 β是 d_β维向量，而 α_i 是一个不可观测的随机效应参数，其假设分布为 $\mathcal{N}(0,\sigma_\alpha^2)$。与前文给出的线性回归示例类似，我们的主要推断目标是对固定效应参数 β进行推断。

β 和 σ 的先验分布与前面给出的相同，现在我们再为 $\alpha_i|\sigma_\alpha^2 \overset{iid}{\sim} N$补充一个先验分布（$0,\sigma_\alpha^2$），其中 $\pi(\sigma_\alpha^2) \propto (\sigma_\alpha^2)^{-1}$。在此先验选择下，且以 $\theta = (\beta^\top,\sigma^2)^\top$和 σ_α^2为条件时，随机效应的条件后验分布具有已知的闭式形式：令 $x_{1:n}$ 表示一个 $n \times$ 维（β）矩阵，其列为$x_{1,\ 1:n} =(x_{1,\ i},\ldots,x_{1,\ n})^\top$，并令 I_n为一个 $n \times n$维单位矩阵，

$$p(\alpha_{1:n} \mid y_{1:n},x_{1:n},\beta,\sigma,\sigma_\alpha^2) = N\left[\alpha_{1:n};\left\{I_n\left(1+\frac{\sigma^2}{\sigma_\alpha^2}\right)\right\}^{-1}(y_{1:n}-x_{1:n}\beta),\sigma^2\left\{I_n\left(1+\frac{\sigma^2}{\sigma_\alpha^2}\right)\right\}^{-1}\right].$$

更一般　　地，β 和 σ^2 的条件后验分布也具有已知的闭式形式。

与线性回归示例类似，假设我们再次希望确保针对 β 的推断能够抵御误差项 ϵ_i的建模假设可能出现的偏离，尤其是异方差性的可能形式及其是否存在。鉴于在此示例中，条件后验分布 $p(\alpha_{1:n} \mid y_{1:n},x_{1:n},\beta,\sigma,\sigma_\alpha^2)$具有已知的闭式形式，我们可以通过使用来自条件后验分布 $p(\alpha_{1:n} \mid y_{1:n},x_{1:n},\beta,\sigma,\sigma_\alpha^2)$的 m 组独立抽样值，估计 $m_n(\theta)$和 $W_n(\theta)$来实现这一点。

因此，可以使用（伪边际）Metropolis-within-Gibbs 方案对 $\theta = (\beta^\top,\sigma^2)^\top$ 进行推断：对于 θ 的每个 Metropolis 步骤，我们首先从 $p(\alpha_{1:n} \mid y_{1:n},x_{1:n},\beta,\sigma,\sigma_\alpha^2)$ 中抽样，然后构造 $\widehat{Q}_n(\theta;\ z)$，再根据通常的判据决定接受或拒绝。现在，我们使用来自条件后验分布的 $m = 5$ 次模拟，在算法的每一步估计梯度及其方差，从而实现这种抽样方法。

与例 1类似，我们根据以下扰动项设定，从线性回归模型中生成观测数据：$\epsilon_i \sim N(0,\sqrt{1+|x_{1,\ i}|^\gamma})$，其中 $\gamma \geq 0$，且 $x_{1,\ i}$ 表示向量 x_i的第一个元素。显然，当 $\gamma = 0$ 时，数据生成过程是同方差的；而当 $\gamma \neq 0$时，精确后验很可能会产生过窄的可信集，导致

覆盖率不可靠。另一方面，引理 2-3 表明，无论 γ取何值，我们提出的广义后验都能正确量化不确定性。

为此，我们生成 $n = 100$ 来自假定数据生成过程的数据，设计方案为 $\gamma = 0$，以及 $\gamma = 2$，其中 x_i 生成为三元独立标准高斯随机向量，并将真实值设为 $\beta = (0。5，1。5，1，1)^\top$，以及 $\sigma = 1$。我们将该设计重复1000次，并对每个数据集使用吉布斯抽样抽取精确后验，同时使用算法1中的随机游走样本抽取广义后验；两个后验分布均使用10000个样本进行近似，其中前5000个样本因预热期而被舍弃。对于每个数据集和方法，我们随后比较回归系数的后验偏差、方差和覆盖率。表2总结了重复抽样结果。

$\gamma = 0$	Q-后验			精确		
	Bias	Var	Cov	Bias	Var	Cov
β_1	-0.0519	0.0310	0.9350	-0.0804	0.0555	0.9820
β_2	-0.0017	0.0220	0.9320	-0.0015	0.0154	0.8830
β_3	-0.0132	0.0299	0.9450	-0.0136	0.0208	0.8880
β_4	0.0079	0.0295	0.9200	0.0076	0.0191	0.8490
$\gamma = 2$	Q-后验			精确		
	Bias	Var	Cov	Bias	Var	Cov
β_1	-0.0572	0.0308	0.9300	-0.0844	0.0536	0.9800
β_2	-0.0047	0.0474	0.9150	-0.0060	0.0206	0.7620
β_3	-0.0034	0.0312	0.9290	-0.0034	0.0194	0.8740
β_4	0.0000	0.0296	0.9290	0.0001	0.0191	0.8480

表 2. 正态线性回归模型中精确贝叶斯和 Q-后验的后验准确度结果。Bias 是重复试验中后验均值的偏差。Var 是重复试验中平均后验方差偏差。Cov 是后验覆盖率。实验的名义水平设为 95% 。

与简单线性回归示例类似，表 2 中的结果表明，当模型被正确设定时，基于估计的似然梯度的广义贝叶斯方法所产生的结果与精确贝叶斯相似。然而，尽管在正确设定 ($\gamma = 0$) 和模型设定错误 ($\gamma = 1$)下，精确贝叶斯的覆盖率都相对明显偏离名义水平，但我们的方法在这两种情况下产生的覆盖率都非常接近名义 95% 水平。

3.2.2. 广义线性随机效应模型。 我们观测来自组 j、 y_{ij} 的个体 i 的二元结果，并令 x_{ij} 表示严格外生回归变量向量。在

具体而言，我们认为这些结果是根据以下形式的二元随机效应 Probit 模型生成的

$$y_{ij} \sim \text{Bernoulli}(p_{ij}), \quad p_{ij} = \Phi\left(x_{ij}^\top \beta + \alpha_i\right), \tag{4}$$

对于个体 $i = 1, \ldots, n$，每个个体有 $j = 1, \ldots, J$ 个观测值，其中 α_i 表示个体 i 的不可观测随机效应，而 $\Phi(x)$ 表示标准高斯累积分布函数。与线性随机效应情形类似，我们再次对 β 采用弥散先验分布，$\pi(\beta) \propto 1$。关于随机效应参数的标准假设是 $\alpha_i \overset{iid}{\sim} N(0, \sigma_\alpha^2)$。然而，随机效应项本身的性质使其精确的参数表示方式并不明确。

本节展示如何在该模型中应用 Q-后验，并比较在随机效应设定正确和错误时，我们的方法与精确后验的表现。为简单起见，我们取 $J = 1$，并仅用 i 下标表示各量；不过，$J > 1$ 情形的处理方式类似。

给定 θ，完整数据似然为

$$p\left(y_{1:n}, \alpha_{1:n} \mid \theta\right) = \prod_{i=1}^{n} \left\{\Phi\left(x_i^\top \beta + \alpha_i\right)\right\}^{y_i} \left\{1 - \Phi\left(x_i^\top \beta + \alpha_i\right)\right\}^{1-y_i} N(\alpha_i; 0, \sigma_\alpha^2);$$

注意，σ_α^2 唯一出现的位置来自密度分量 $N(\alpha_i; 0, \sigma_\alpha^2)$ 中的贡献。根据 Fisher 恒等式，相对于 β 的对数梯度为

$$\nabla_\beta \log p(y_{1:n}, \alpha_{1:n} \mid \theta) = \sum_{i=1}^{n} x_i^\top \int_{\mathbb{R}} u_i(\theta, \alpha_i) p_\theta(\alpha_i \mid y_{1:n}) \mathrm{d}\alpha_i,$$

其中

$$u_i(\theta, \alpha_i) = \frac{\left[y_i - \Phi\left(x_i^\top \beta + \alpha_i\right)\right] \varphi\left(x_i^\top \beta + \alpha_i\right)}{\Phi\left(x_i^\top \beta + \alpha_i\right)\left[1 - \Phi\left(x_i^\top \beta + \alpha_i\right)\right]},$$

而关于 σ_α^2 的梯度为

$$\nabla_{\sigma_\alpha^2} \log p_\theta(y_{1:n}, \alpha_{1:n}) = -\frac{n}{2\sigma_\alpha^2} + \frac{1}{2}\left(\frac{1}{\sigma_\alpha^2}\right)^2 \sum_{i=1}^{n} \int_{\mathbb{R}} \alpha_i^2 p_\theta(\alpha_i \mid y_{1:n}) \mathrm{d}\alpha_i.$$

与线性情形不同，$p_\theta(\alpha_{1:n} \mid y_{1:n})$ 在此广义线性模型中无法通过解析方法求得。然而，按照Zeger 和 Karim (1991) 和McCulloch (1997)的方法，可以使用基于提议分布 $p_\theta(\alpha_{1:n} \mid y_{1:n})$ 的 Metropolis-Hastings 算法，从 $N(0,\ \sigma_\alpha^2)$中可靠地生成样本。

随后，可以使用 Metropolis-within-Gibbs 方法评估 Q-后验，即在第 k 步抽取 N 条路径$\alpha_{1:n,\ j}(k) \sim p_\theta(\alpha_{1:n} \mid y_{1:n}),\ j = 1,\ \ldots\ldots,\ N$，然后利用这些抽样值估计梯度。梯度估计完成后，可以使用第 3.1.3 节所讨论的方法估计其方差。

现在，我们比较随机效应 Probit 模型中 Q-后验与精确后验的表现。我们从随机效应 Probit 模型中生成 $n = 100$个样本观测值，并设定随机效应分布的两种不同情形：第一个数据生成过程正确地假定随机效应分布为 $N(0, \sigma_\alpha^2)$；而在第二个数据生成过程中，随机效应的实际分布为具有 4 个自由度的学生 t 分布，因此在第二种设计下，假定模型存在设定错误。回归变量 x_i 再次生成为三元独立标准高斯随机向量，并将真实值设为 $\beta = (1, 1, 1, 1)^\top$。在这两种情形下，随机效应参数的尺度均设为 1。

我们重复这两种设计1000次，并对每个数据集使用类似于Zeger 和 Karim (1991)的方法，通过 Gibbs-within-Metropolis 抽样方案对精确后验分布进行抽样。我们还使用上文讨论的 Gibbs-within-Metropolis 方法对 Q-后验进行抽样，并在所有实验中使用 $N = 5$条件后验分布的抽样值。[10] 两种后验分布均使用10000个样本值进行近似，其中前5000个样本值作为预热期而舍弃。随后，我们针对每个数据集和每种方法，比较回归系数的后验偏差、方差和覆盖率。表3总结了重复抽样结果。

与其他示例类似，Q 后验分布给出的后验位置与精确贝叶斯的结果相近，但无论模型设定正确与否，其不确定性量化都仍接近名义水平。相比之下，由于方差

[10]由于这是一个 Gibbs-within-Metropolis 抽样方案，我们运行采样器以获得 $2 * N$ α_i的抽样值，然后舍弃前 N 个样本值作为预热期。

精确贝叶斯的方差（平均而言）小于 Q-后验给出的方差，因此，精确贝叶斯后验分布的覆盖率通常比 Q-后验的覆盖率偏离名义水平更多。对于相应的后验分布本身，这种差异尤其明显；我们在图 1中绘出了设定错误的数据生成过程下某一次重复实验的结果。在这种情况下，显然，尽管后验分布的位置相近，但其尺度却大不相同。正是这种不确定性的增加，使得即使模型设定错误，不确定性量化仍然可靠。

	Q-后验			精确		
$N(0, 1)$	Bias	Var	Cov	Bias	Var	Cov
β_1	-0.0438	0.0358	0.9760	-0.1016	0.0266	0.9060
β_2	-0.0414	0.0375	0.9740	-0.0841	0.0306	0.9420
β_3	-0.0360	0.0339	0.9680	-0.0955	0.0344	0.9160
β_4	-0.0352	0.0396	0.9780	-0.0863	0.0298	0.9280
	Q-后验			精确		
t_4	Bias	Var	Cov	Bias	Var	Cov
β_1	-0.1255	0.0365	0.9440	-0.1638	0.0249	0.8350
β_2	-0.1053	0.0401	0.9480	-0.1509	0.0273	0.8450
β_3	-0.1208	0.0391	0.9440	-0.1607	0.0277	0.8450
β_4	-0.1171	0.0405	0.9490	-0.1595	0.0282	0.8520

表 3. 精确后验分布和 Q- 后验分布在 probit随机效应模型中的后验准确度结果。Bias 表示各次重复试验中后验均值的偏差。Var 表示各次重复试验中后验方差偏离值的平均值。Cov 表示后验覆盖率（95% 名义覆盖率）。第一个设计（$N(0, 1)$）表示随机效应分布设定正确的情形，而设计 t_4 表示随机效应分布设定不正确的情形。

4. 广义损失函数后验

当模型设定错误时，没有理由假设 KL极小化解 $\theta_\star$所产生的推断适合当前目的。解决这一问题的一种方法是改用广义后验进行推断，这类后验使用针对当前问题的损失函数来进行贝叶斯推断。为确保贝叶斯推断在设定错误的模型中仍然适合当前目的，Bissiri 等人（2016）建议将似然后验更新替换为基于损失函数的更新 q： $\mathcal{Y} \times \Theta \to \mathbb{R}_+$ 该损失函数对于当前具体的推断任务至关重要。如果随后考虑 $q_n(\theta) = \sum_{i=1}^{n} q(y_i, \theta)$ 作为关于

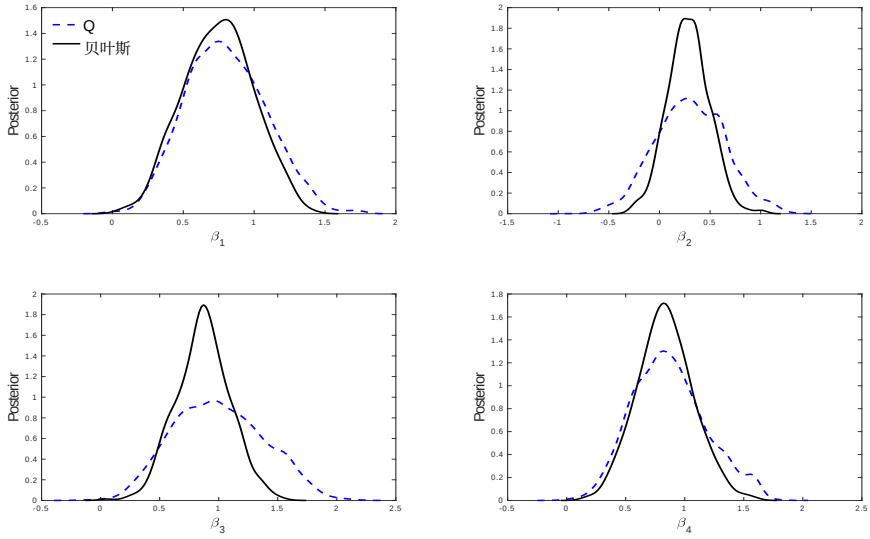

图1. 随机效应 Probit 模型中随机效应分布设定错误时，代表性数据集上的精确后验分布与 Q-后验比较。

观测样本的累积损失函数，则后验更新可以基于广义后验

$$\pi_n^q(\theta) = \frac{\exp\{-\omega q_n(\theta)\}\,\pi(\theta)}{\int_\Theta \exp\{-\omega q_n(\theta)\}\,\pi(\theta)},$$

对于某个学习率 $\omega \geq 0$。

尽管广义后验分布很有用，但其通常并未校准良好，也无法正确量化不确定性；关于这一点的讨论，以及规避该问题的现有方法，可参见 Miller（2021）的第4节及其中的参考文献。在本节中，我们将说明，Q-后验同样可以应用于广义贝叶斯后验，以实现有效的不确定性量化。给定 $q_n(\theta)$，我们可以用由 d_θ 维估计方程 $\psi_n(\theta) = \nabla_\theta q_n(\theta)$ 构成的集合来表示该损失函数中的信息，其中我们假设这些导数对所有 $\theta \in \Theta$ 都存在，并且每个 $i \geq 1$。[11] 令 $V_n(\theta)$ 为 $\mathrm{Cov}\{\psi_n(\theta)/\sqrt{n}\}$ 的估计量，我们可以将先验信念 $\pi(\theta)$ 更新为后验信念，得到

[11]一般而言，我们只要求该导数几乎处处存在；这允许导数在可数多个点上未定义，例如，当我们的损失函数为中位数损失时就是如此。

以下 Q-后验

$$\pi_n^{\Psi}(\theta) = \frac{\exp\{-\Psi_n(\theta)\}\pi(\theta)}{\int_\Theta \exp\{-\Psi_n(\theta)\}\pi(\theta)\mathrm{d}\theta}, \quad \Psi_n(\theta) := \frac{1}{2}\frac{\psi_n(\theta)^\top}{\sqrt{n}} V_n(\theta)^{-1}\frac{\psi_n(\theta)}{\sqrt{n}}. \tag{5}$$

根据第 3.1 节的结果，Q-后验 $\pi_n^{\Psi}(\theta)$ 应当能够给出一个正确量化不确定性的广义贝叶斯后验，而无需选择任何调节常数；详情见定理 2 。在陈述这一结果之前，我们先通过一个简单示例展示 Q-后验 $\pi_n^{\Psi}(\theta)$ 的适用性，说明在基于一般损失函数对未知参数进行贝叶斯推断时， $\pi_n^{\Psi}(\theta)$ 能够正确量化不确定性。

4.1. 稳健分位数推断。 考虑从模型中观测到一个独立同分布序列 $y_1, \ldots, y_n$

$$y_i = \theta + \epsilon_i, \quad i = 1, \ldots, u, \quad \epsilon_i \overset{iid}{\sim} F(\cdot)$$

其中，未知参数 θ 具有先验密度 $\pi(\theta)$，且与 ϵ相互独立。遵循Doksum 和 Lo (1990) 的做法，出于稳健性考虑，我们不希望根据θ从 $\{y_i$估计其后验分布： $i \geq 1\}$ 而是根据样本中位数 $T_n = \mathrm{med}(y_1, \ldots, y_n)$；相关方法见 Lewis 等人 (2021)。当 F 具有密度 f时，由此得到精确后验分布

$$\pi(\theta|T_n) = \pi(\theta)\exp\left\{\frac{1}{2}(n-1)\log F(T_n-\theta)\{1-F(T_n-\theta)\} + \log f_\theta(T_n)(1-F))\right\}$$

当 n 为奇数时。

鉴于上述后验分布的形式，Miller (2021) 建议改为对 θ进行广义贝叶斯推断，使用更简单的损失函数 $-\frac{1}{2}\log F(T_n-\theta)\{1-F(T_n-\theta)\}$。尽管 Miller(2021) 证明了这种后验分布在大样本下表现良好，但即使模型设定正确，所得后验分布的覆盖率仍不正确。相比之下，我们建议使用 Q-后验进行贝叶斯推断，其将

基于上述损失函数相对于 θ 的梯度，

$$m_n(\theta) = \frac{1}{2}\frac{f(T_n-\theta)}{F(T_n-\theta)} - \frac{1}{2}\frac{f(T_n-\theta)}{1-F(T_n-\theta)}.$$

对于任意 $\theta \in \Theta$，$m_n(\theta)$ 的方差均可通过独立同分布自助法一致地估计。令 $W_n(\theta)$ 表示该估计值；随后我们可以使用 Q-后验对 θ 进行推断

$$\pi_n(\theta) \propto \pi(\theta)\exp\left[-\frac{1}{2}\left\{\frac{1}{2}\frac{f(T_n-\theta)}{F(T_n-\theta)} - \frac{1}{2}\frac{f(T_n-\theta)}{1-F(T_n-\theta)}\right\}^2 W_n(\theta)^{-1}\right].$$

在本例中，我们使用估计方程自助法（参见 Hu 和 Kalbfleisch，2000年以及 Chatterjee 和 Bose，2005年的相关讨论）计算矩阵 W_n（θ）在每个 θ取值处的值。

现在，我们比较在模型设定正确和设定错误的模型中，使用广义后验分布和 Q-后验所得到的不确定性量化结果。两种情况下，我们都假设 $F(\cdot)$ 服从标准高斯分布。在第一个实验中（称为 DGP1），观测数据由均值为 $\theta = 1$、方差为 4 的高斯分布生成。在模型设定错误的情形下（称为 DGP2），我们根据参数化形式 $0.9N(\theta = 1, 4) + 0.1N(0, 1)$ 生成观测数据，该数据服从混合高斯分布。

在第一种情况下，真实中位数为 1；而在第二种情况下，数据的真实中位数实际上约为 0.84.[12] 我们从两个数据生成过程中各模拟生成 1000 个观测数据集，并比较广义贝叶斯方法与基于 Q-后验的方法的结果。表 4比较了这两种方法的后验均值、方差和覆盖率。结果表明，Miller (2021) 提出的广义后验无法对真实中位数实现可靠的覆盖，而 Q-后验的覆盖率再次接近名义水平。

[12]T该值必须通过数值地求混合分布的累积分布函数的反函数来获得。

	Q-后验			精确		
DGP1	Bias	Var	Cov	Bias	Var	Cov
θ	-0.0031	0.0654	0.9420	-0.0031	0.0152	0.6700
DGP2	Bias	Var	Cov	Bias	Var	Cov
θ	-0.0115	0.0607	0.9360	-0.0118	0.0157	0.7160

表4. 使用广义后验和Q-后验计算中位数的后验准确度结果。Bias表示重复试验中后验均值的偏差。Var表示重复试验中后验方差偏差的平均值。Cov表示后验覆盖率（95%的名义覆盖率）。

4.2. **不确定性量化**。 如果在基于似然的贝叶斯语境中， $m_n(\theta)$ 和 $W_n(\theta)$ 满足的正则条件在广义贝叶斯语境下对 $\psi_n(\theta)$ 和 $V_n(\theta)$ 也成立，那么后验分布 $\pi_n^{\Psi}(\theta)$ 就能正确量化不确定性。

定理2. 假设假设*4*成立，且存在非随机函数 $\psi : \Theta \to \mathbb{R}^{d_\theta}$ 和 $V : \Theta \to \mathbb{R}^{d_\theta \times d_\theta}$ ，使得假设*2*和*3*满足$m_n(\theta) = \psi_n(\theta)$、 $m(\theta) = \psi(\theta)$以及 $W_n(\theta) = V_n(\theta)$。那么，对于 $\vartheta := \sqrt{n}(\theta - \theta_\star) - Z_n/\sqrt{n}$， $Z_n := \Omega_\star^{-1}\nabla_\theta\psi(\theta_\star)^\top V(\theta_\star)^{-1}\psi_n(\theta)$， $\Omega_\star = \left[\nabla_\theta\psi(\theta_\star)^\top V(\theta_\star)^{-1}\nabla_\theta\psi(\theta_\star)\right]$，以及 $\mathcal{T}_n := \{\vartheta = \sqrt{n}(\theta - \theta_\star) - Z_n/\sqrt{n} : \theta \in \Theta\}$，当 $n \to +\infty$， $\int_{\mathcal{T}_n} \|\vartheta\| |\pi_n^{\Psi}(\vartheta) - N\{\vartheta; 0, \Omega_\star^{-1}\}| d\vartheta = o_p(1)$，且对于 $\overline{\theta} = \int_\theta \theta \pi_n^{\Psi}(\theta) d\theta$， $\sqrt{n}(\overline{\theta} - \theta_\star) \Rightarrow N(0, \Omega_\star^{-1})$。

Bissiri 等（2016）的广义贝叶斯后验的可信集宽度由 $[\nabla_\theta\psi(\theta_\star)]^{-1}$ 决定，而 Q-后验 $\pi_n^{\Psi}(\theta)$ 的可信集宽度则由 $\Omega(\theta_\star)^{-1} = [\nabla_\theta\psi(\theta_\star)]^{-1}V(\theta_\star)[\nabla_\theta\psi(\theta_\star)]^{-1}$ 决定，从而确保 $\pi_n^{\Psi}(\theta)$ 能够正确量化不确定性。现有的广义贝叶斯推断覆盖率校正方法依赖于对后验分布进行事后处理并计算二阶导数信息，例如 Holmes 和 Walker（2017）、Syring 和 Martin（2019）、Giummolè 等（2019）所采用的方法；或者依赖于昂贵的自助法程序，例如 Matsubara 等（2021）所采用的方法。相比之下，我们的方法无需后处理，并针对校准良好的广义后验推断问题，提供了一种内在的（广义）贝叶斯解决方案。

注 4. 一类常见的广义后验分布，是当 $q(y_i, \theta)$ 本身由某种伪似然得到时，其中包括部分似然（Cox，1975）和复合似然（Lindsay，1988，以及 Varin 等人，2011）。例如，

Miller (2021) 给出了若干此类方法及其与广义后验分布之间联系的有用示例。然而，正如 Miller (2021) 所承认的，由这些损失函数构造的广义后验分布“即使在渐近情况下也无法达到正确的覆盖率”。不过，在 Miller (2021) 所讨论的每种情形中，损失函数都足够光滑，因此可以将其表示为式（ $\pi_n^{\Psi}(\theta)$ （5）中的 Q-后验 θ。因此，例如，只要复合似然关于 θ 足够光滑，我们就可以应用 Q-后验，基于该损失函数进行广义贝叶斯推断，并获得正确的覆盖率。

5. 讨论与结论

5.1. **替代方法**。 在似然框架下，当模型设定正确时，我们有 $\Sigma_\star^{-1} = W_\star^{-1} = \mathcal{I}(\theta_\star)^{-1}$，此时 Q-后验与精确后验在渐近意义下一致。然而，如果模型设定错误，Q-后验仍能给出校准良好的可信集；但只有当 $W_n(\theta_\star)$ 是 $W_\star$的相合估计量时，这一结论才成立。否则，Q-后验将无法正确量化不确定性。不过，正如第 3.1.3 节所述，在大多数情况下，都可以利用现有公式或自助法获得 $W(\theta_\star)$ 的可靠估计量。

Q-后验方法与现有的贝叶斯推断方法有显著不同，尤其适用于可能存在模型设定错误的情形。目前受到广泛关注的两种方法是 Müller (2013 提出的“夹心”修正方法，以及 BayesBag 方法，参见 Hugginsand Miller (2019)。

Müller的方法是利用显式的高斯近似 $\theta \sim N\{\bar{\theta}, \mathcal{H}_n(\bar{\theta})^{-1} W_n(\bar{\theta}) \mathcal{H}_n(\bar{\theta})^{-1}\}$，对标准后验分布的抽样值进行校正，其中 $\bar{\theta}$是后验均值。这种校正既可以通过直接从多元正态分布中抽样来实现，也可以对每个后验抽样值 θ 根据以下线性方程进行调整

$$\widetilde{\theta} = \bar{\theta} + \mathcal{H}_n(\bar{\theta})^{-1} W_n(\bar{\theta}) \mathcal{H}_n(\bar{\theta})^{1/2} (\theta - \bar{\theta});$$

另见，Giummolè 等人（2019）关于使用评分规则构造广义后验分布的相关方法。我们认为，这种事后校正在若干方面并非最优

原因如下：首先，从哲学层面看，这相当于将频率学派原则应用于贝叶斯学习算法的输出，因此其本质上并非贝叶斯方法；其次，它需要显式计算二阶导数，这可能很困难，而且数值表现可能不稳定；第三，当后验分布并非近似高斯分布时，这种高斯近似效果不佳，例如参数的支撑集受限，或样本量较小˜时；最后，这种校正很容易产生一个 θ 值，使其落在 $\pi(\theta)$的支撑集之外，例如当 Θ 是 $\mathbb{R}^{d_\theta}$的有界子集时。[13]

另一种方法是采用后验装袋法，正如 BayesBag 方法所提出的；有关 BayesBag 的深入讨论，请参见 Huggins 和 Miller (2019)。该方法试图通过装袋法校正后验覆盖率。令 $b = 1$、.....、 B 表示自助法索引，令 $y_1^{(b)}{}_{:n} = (y_1^{(b)}$、.....、 $y_n^{(b)})$ 表示第 b 个自助样本，其中 $y_i^{(b)}$ 从原始数据集中有放回抽样得到。BayesBag 后验分布定义为

$$\pi^\star(\theta \mid y_{1:n}) \approx B^{-1} \sum_{b=1}^{B} \pi^\star(\theta \mid y_{1:n}^{(b)}).$$

BayesBag 易于使用，且不需要额外的算法工具。不过，它确实需要重新运行马尔可夫链蒙特卡洛抽样算法，以获得 θ 针对每个$\{y_1^{(b)}{}_{n:}: \quad b = 1$、.....、 $B\}$的后验抽样值。如果模型维度很高或包含潜变量，后者的计算开销可能很大。即便如此，Huggins 和 Miller (2019) 仍证明，BayesBag 后验分布的可信集并未得到良好校准。渐近地，BayesBag 后验方差为 $c^{-1}\mathcal{H}_\star^{-1} + c^{-1}\Sigma_\star^{-1}$，其中 $c = \lim_n B/n$。因此，BayesBag 后验分布会根据 B的选择，表现出覆盖率过高或过低。只有当参数 θ 是标量时，才可能选择 c 使后验分布具有有效的频率学派覆盖率；也就是说，使 $\Sigma_\star^{-1} = c(\mathcal{H}_\star^{-1} + \Sigma_\star^{-1})$。 最后，BayesBag 方法本身应用于具有弱相关数据的模型，或似然难处理的情形，似乎并不简单。

[13]虽然可以对抽样值进行变换，使其限制在适当的空间内，但此类变换可能不具备不变性，而且通常没有理论依据来决定应采用哪种变换。

与事后校正方法和 BayesBag 不同，Q-后验无需进行任何事后抽样校正，也不需要采用自助法方法来获得校准良好的后验分布。此外，在需要使用伪边际方法进行贝叶斯推断的情况下，也可以采用 Q-后验。

5.2. **结论**。 我们提出了一种新的贝叶斯推断方法，通常称为 Q-后验，能够提供可靠的不确定性量化。在基于似然的情形下，Q-后验可以被视为一种贝叶斯合成似然（Wood，2010）后验，其中我们用得分方程似然的近似替代观测样本的似然。Q-后验的关键特征在于：无论模型设定如何，它都能保证所得到的可信集渐近地校准良好；而当模型设定正确时，Q-后验与精确后验一致（在大样本下）。重要的是，即使由于存在潜变量而必须估计似然，Q-后验在相当宽松的正则条件下仍能提供校准良好的推断。

当应用于广义贝叶斯后验时（Bissiri 等人，2016），我们证明了 Q-后验仍然校准良好。据我们所知，现有的所有旨在校正广义后验覆盖率的方法，要么对后验抽样值进行事后校正，而这些方法最终基于对所得后验抽样值作出的某种（隐含的）正态性假设；要么可能需要复杂的自助法方法。相比之下，Q-后验无需任何额外调节或对抽样值进行事后校正，即可提供正确的不确定性量化。

当似然难以处理且必须进行估计时，可以利用 Fisher 恒等式和简单的蒙特卡洛估计量构造一种 Q-后验。原则上，也可以采用其他蒙特卡罗方法，例如重要性抽样，来构造 Q-后验。本文仅采用简单的蒙特卡洛估计量，因为这类方法最容易应用，也最容易进行理论分析。在作者正在开展的研究中，我们探讨了使用序贯重要性抽样估计量，其得到的 Q-后验性能相近。不过，在这类情况下，分析后验分布的行为

由于蒙特卡罗估计量具有序贯性，分析会更加困难，要获得类似于定理 1 中的理论结果也会更加艰巨。我们将这一重要问题留待未来研究。

参考文献

Andrieu, C. 和 Roberts, G. O. (2009). 高效蒙特卡罗计算的伪边际方法。《统计学年鉴》，37(2):697–725. 4, 5, 15 Bissiri, P. G., Holmes, C. C. 和 Walker, S. G. (2016). 更新信念分布的一般框架。《皇家统计学会会刊：B辑（统计方法学）》，78(5):1103–1130. 2, 3, 8, 9, 26, 30, 33 Cappé, O., Moulines, E. 和 Rydén, T. (2005). Springer统计学丛书. 13 Chatterjee, S. 和 Bose, A. (2005). 估计方程的广义自助法。《统计学年鉴》，33(1):414–436. 20, 29 Chernozhukov, V. 和 Hong, H. (2003). 一种经典估计的马尔可夫链蒙特卡洛方法。《计量经济学杂志》，115(2):293–346. 2, 9, 41, 45, 46 Chib, S., Shin, M. 和 Simoni, A. (2018). 矩条件模型的贝叶斯估计与比较。《美国统计学会会刊》，113(524):1656–1668. 9 Cox, D. R. (1975). 部分似然。《生物统计学》，62(2):269–276. 30 Dau, H.-D. 和 Chopin, N. (2022). 无浪费序贯蒙特卡洛。《皇家统计学会会刊：B辑（统计方法学）》，84(1). 21 Doksum, K. A. 和 Lo, A. Y. (1990). 基于部分信息的位置一致且稳健的贝叶斯程序。《统计学年鉴》，18(1):443–453. 28 Drovandi, C., Nott, D. J. 和 Frazier, D. T. (2021). 关于贝叶斯受限似然方法的讨论：贝叶斯回归中对非充分统计量进行条件化。《贝叶斯分析》，16(4):1393–1462. 7 Frazier, D. T., Nott, D. J., Drovandi, C. 和 Kohn, R. (2022). 使用合成似然进行贝叶斯推断：渐近理论与调整。《美国统计学会会刊》，（刚录用）：1–28. 6, 41 Giummolè, F., Mameli, V., Ruli, E. 和 Ventura, L. (2019). 使用适当评分规则进行客观贝叶斯推断。《Test》，28(3):728–755. 30, 31 Holmes, C. C. 和 Walker, S. G. (2017). 为一般贝叶斯模型中的幂似然赋值。《生物统计学》，104(2):497–503. 3, 30 Horn, R. A. 和 Johnson, C. R. (2012). 矩阵分析。剑桥大学出版社. 16 Hu, F. 和 Kalbfleisch, J. D. (2000). 估计函数自助法。《加拿大统计学杂志》，28(3):449–481. 20, 29 Huggins, J. H. 和 Miller, J. W. (2019). 使用装袋后验分布进行稳健推断与模型批评。arXiv 预印本 arXiv:1912.07104. 31, 32 Jewson, J. 和 Rossell, D. (2021). 一般贝叶斯损失函数选择及非正规模型的使用。arXiv 预印本 arXiv:2106.01214. 2 Kleijn, B. J. 和 van der Vaart, A. W. (2012). 模型设定错误下的伯恩斯坦-冯·米塞斯定理。《电子统计学杂志》，6:354–381. 5 Lehmann, E. L. 和 Casella, G. (2006). 点估计理论。施普林格科学与商业媒体. 41

Lewis, J. R., MacEachern, S. N. 和 Lee, Y. (2021). 贝叶斯受限似然方法：贝叶斯回归中对非充分统计量进行条件化（附讨论）。《贝叶斯分析》，16(4):1393–1462. 7, 28 Lindsay, B. G. (1988). 复合似然方法。《当代数学》，80(1):221–239. 30 Loaiza-Maya, R., Martin, G. M. 和 Frazier, D. T. (2021). 聚焦贝叶斯预测。《应用计量经济学杂志》，36(5):517–543. 2 Matsubara, T., Knoblauch, J., Briol, F.-X., Oates, C. 等 (2021). 针对难以处理的似然的稳健广义贝叶斯推断。arXiv 预印本 arXiv:2104.07359. 2, 30 McCulloch, C. E. (1997). 广义线性混合模型的最大似然算法。《美国统计学会会刊》，92(437):162–170. 25 Miller, J. W. (2021). 广义后验的渐近正态性、集中性与覆盖率。《机器学习研究杂志》，22:168–1. 3, 27, 28, 29, 31 Müller, U. K. (2013). 设定错误模型下贝叶斯推断的风险及夹心协方差矩阵。《计量经济学》，81(5):1805–1849. 2, 31 Newey, W. K. 和 West, K. D. (1987). 一种简单的半正定异方差与自相关一致协方差矩阵。《计量经济学》，55(3):703–708. 21 Politis, D. N. 和 Romano, J. P. (1994). 平稳自助法。《美国统计学会会刊》，89(428):1303–1313. 21 Price, L. F., Drovandi, C. C., Lee, A. 和 Nott, D. J. (2018). 贝叶斯合成似然。《计算与图形统计学杂志》，27(1):1–11. 6 Shao, X. (2010). 依赖型野生自助法。《美国统计学会会刊》，105(489):218–235. 21 Shephard, N. 和 Pitt, M. K. (1997). 非高斯测量时间序列的似然分析。《生物统计学》，84(3):653–667. 19 Syring, N. 和 Martin, R. (2019). 校准一般后验可信区域。《生物统计学》，106(2):479–486. 2, 3, 30 Syring, N. 和 Martin, R. (2020). 含噪图像边界上的稳健且速率最优的吉布斯后验推断。《统计学年鉴》，48(3):1498–1513. 2 Varin, C., Reid, N. 和 Firth, D. (2011). 复合似然方法综述。《中华统计学志》，21(1):5–42. 30 White, H. (1982). 设定错误模型的最大似然估计。《计量经济学》，50(1):1–25. 5 Wood, S. N. (2010). 含噪非线性生态动力系统的统计推断。《自然》，466(7310):1102–1104. 6, 7, 33 Wu, P.-S. 和 Martin, R. (2020). 广义贝叶斯推断中学习率选择方法的比较。arXiv 预印本 arXiv:2012.11349, 6. 3 Zeger, S. L. 和 Karim, M. R. (1991). 带随机效应的广义线性模型；一种吉布斯抽样方法。《美国统计学会会刊》，86(413):79–86. 25

附录 A. 主要引理

本节包含一些用于证明正文结果的引理，这些引理本身也具有独立价值。我们首先定义以下补充记号，供本附录后续部分使用。对于两个（可能是随机的）序列 a_n, b_n，我们称

$a_n \lesssim b_n$，如果存在足够大的 n'，使得对所有 $n \geq n'$，都存在一个 $C > 0$ 满足 $a_n \leq Cb_n$（几乎处处）；而如果 $a_n \lesssim b_n$ 且 $b_n \lesssim a_n$（几乎处处），我们记作 $a_n \asymp b_n$。对于一个已知函数 $f\colon \Theta \to \mathbb{R}^d$，我们将 f 关于概率测度 H 的期望记为 $\mathbb{E}_H[f(\theta)] = \int_\Theta f(x)\mathrm{d}H(x)$。在全文中，我们滥用记号 $\|\cdot\|$，在向量的情形下表示欧几里得范数，在矩阵的情形下表示一种适当的矩阵范数。后一种记号滥用并无影响，因为我们只处理长度固定的向量和矩阵。

引理 4. 设 $Z \in \mathbb{R}^d$ 满足 $\mu = \mathbb{E}[Z]$，并令 Z_l（分别为 μ_l）、$l = 1, \ldots, d$ 表示 Z（分别为 μ）的坐标。设 M 为半正定矩阵，且对某个矩阵 A 有 $M = A^\top A$。假设存在某个 $\sigma > 0$，使得对每个 $l = 1, \ldots, d$，都有 $|\mu_j| \leq \sigma$；并且对 $j \in \{2, 4\}$、$\mathbb{E}[|Z_l - \mu_l|^j] \leq \sigma^j$，有方差 $(\|AZ\|^2) \leq 2\sigma^4\|M\|^2(d+2)^2$.

证明。记作

$$\begin{aligned}\|AZ\|^2 = Z^\top MZ &= (Z-\mu)^\top M(Z-\mu) - \mu^\top M\mu + 2\mu^\top MZ \\ &= \|A(Z-\mu)\|^2 - \|A\mu\|^2 + 2\mu^\top MZ.\end{aligned}$$

于是，

$$\mathrm{Var}(\|AZ\|^2) = \mathrm{Var}\{\|A(Z-\mu)\|^2\} + 4\mathrm{Var}\{\mu^\top MZ\} + 2\mathrm{Cov}\{2\mu^\top MZ, \|A(Z-\mu)\|^2\}.$$

由柯西–施瓦茨不等式，

$$|\mathrm{Cov}\{\mu^\top MZ, \|A(Z-\mu)\|^2\}| \leq \sqrt{\mathrm{Var}\{\mu^\top MZ\}}\sqrt{\mathrm{Var}\{\|A(Z-\mu)\|^2\}}$$

因此

$$\mathrm{Var}(\|AZ\|^2) \leq \mathrm{Var}\{\|A(Z-\mu)\|^2\} + 4\{\mu^\top M\Sigma M\mu\} + 2\sqrt{4\{\mu^\top M\Sigma M\mu\}}\sqrt{\mathrm{Var}\{\|A(Z-\mu)\|^2\}}.$$

现在，我们利用结果陈述中的矩假设来界定 $\mathrm{Var}[\|A(Z-\mu)\|^2]$。具体而言，

$$\mathrm{Var}[\|A(Z-\mu)\|^2] \leq \|M\|^2\mathrm{Var}[\|Z-\mu\|^2],$$

and

$$\begin{aligned}\mathrm{Var}[\|(Z-\mu)\|^2] &= \sum_{i=1}^d \mathrm{Var}[(Z_i-\mu_i)^2] + 2\sum_{i=1}^d\sum_{j,i\neq j}^d \mathrm{Cov}\left[(Z_i-\mu_i)^2, (Z_i-\mu_j)^2\right] \\ &\leq d\sigma^4 + 2\sum_{i=1}^d\sum_{j,i\neq j}^d \mathrm{Cov}\left[(Z_i-\mu_i)^2, (Z_i-\mu_j)^2\right] \\ &\leq d\sigma^4 + 2\sum_{i=1}^d\sum_{j,i\neq j}^d \left\{\mathrm{Var}[(Z_i-\mu_i)^2]\mathrm{Var}[(Z_j-\mu_j)^2]\right\}^{1/2} \\ &= d\sigma^4 + 2d(d-1)\sigma^4 \\ &\leq 2\sigma^4 d^2.\end{aligned}$$

对于 A, B 维数相容的半正定矩阵，我们有以下不等式

$$\mathrm{Tr}AB \leq \|A\|\mathrm{Tr}(B).$$

利用 Tr(·) 的性质以及上述不等式，我们可以得到

$$\begin{aligned}\mu^\top M\Sigma M\mu = \mathrm{tr}(\mu^\top M\Sigma M\mu) \leq & \sigma^2\mathrm{Tr}(M\mu\mu^\top M)\\ \leq & \sigma^2\|M\|^2\|\mu\|^2\\ \leq & \sigma^4\|M\|^2.\end{aligned}$$

应用所显示的两个方程，便可得到

$$\mathrm{Var}(Z^\top MZ) \leq 2\sigma^4\|M\|^2[d^2+2+2d] = \sigma^2\|M\|^2(d+2)^2.$$

□

下面的引理用于将引理 4 的结果扩展到二次型中的矩阵为随机矩阵的情形。

引理 5 。 设 A 为正定矩阵，且 $B = A + H$ 满足 $\rho := \|A^{-1}H\| < 1$，则 n

$$\|B^{-1} - A^{-1}\| \leq \|A^{-1}\|^2\|H\|/(1-\rho).$$

证明。 由于 $B = A(I + A^{-1}H)$，且 $\rho < 1$, B 是非奇异的。记

$$A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1} = A^{-1}HB^{-1},$$

从而

$$\|A^{-1} - B^{-1}\| = \|A^{-1}HB^{-1}\| \leq \|A^{-1}H\|\|B^{-1}\| \leq \|A^{-1}\|\|H\|\|B^{-1}\|. \tag{6}$$

接下来只需对 $\|B^{-1}\|$ 进行界定。根据 Woodbury 恒等式,

$$B^{-1} = (A+H)^{-1} = A^{-1} - A^{-1}H(I+A^{-1}H)^{-1}A^{-1} = A^{-1} - A^{-1}H(A+H)^{-1} = A^{-1} - A^{-1}HB^{-1},$$

其中倒数第二个不等式成立，是因为 $(I + A^{-1}B)^{-1} = (A + B)^{-1}A$。

根据三角不等式,

$$\|B^{-1}\| \leq \|A^{-1}\| + \|A^{-1}HB^{-1}\| \leq \|A^{-1}\| + \|A^{-1}H\|\|B^{-1}\| = \|A^{-1}\| + \rho\|B^{-1}\|,$$

这意味着

$$\|B^{-1}\| \leq \|A\|^{-1}/(1-\rho).$$

将上述内容代入公式（6）即可得到结果。 □

引理 6 利用引理 4-5 对随机加权矩阵对应二次型的方差进行界定。

引理 6. 设 $Z \in \mathbb{R}^d$ 具有 $\mu = \mathbb{E}[Z]$，方差为 Σ，并设 $Z_1, \ldots, Z_N$ 为与 Z具有相同分布的独立同分布观测。此外，设 $Z_{i,l}$（分别地，$\mu_{i,l}$）、$l = 1, \ldots, d$表示 Z_i（分别地，μ）的坐标$\widehat{\ }$，并将其样本均值记为 $\overline{Z}$。设 Σ 为随机协方差矩阵，满足：对于某个（随机）平方根矩阵 A，有 $\widehat{\Sigma} = A^\top A$。对于某个 $\sigma > 0$，且对每个$j = 1, \ldots, d$，满足以下矩假设：对每个 $l = 1, \ldots, d$，以及$j \in \{2, 4\}$、$\mathbb{E}[|Z_{i,l} - \mu_{i,l}|^j] \leq \sigma^j$、$|\mu_l| \leq \sigma$、$\|\Sigma\| \leq \sigma^2$，并且

$$\mathbb{E}\|\widehat{\Sigma} - \Sigma\| \leq C\sigma^2/N, \quad \mathbb{E}\|\widehat{\Sigma} - \Sigma\|^2 \leq C\sigma^4/N.$$

若 $\|\Sigma(\widehat{\Sigma}^{-1} - \Sigma^{-1})\| < 1$ 如 $N \to +\infty$，则

$$\mathbb{E}[\overline{Z}^\top \widehat{\Sigma}^{-1} \overline{Z}] = \mu^\top \Sigma^{-1} \mu + O(\|\sigma^2 \Sigma^{-1}\|^2/N)$$

and

$$\mathrm{Var}(\overline{Z}^\top \widehat{\Sigma}^{-1} \overline{Z}) = O(\|\sigma^2 \Sigma^{-1}\|^4/N).$$

注 5. 引理 6中关于 Z_i 的矩假设与引理 4中的要求完全相同。条件 $\|\Sigma(\widehat{\Sigma}^{-1} - \Sigma^{-1})\| < 1$，当 $N \to +\infty$ 时，是应用引理 5中的扰动结果所必需的。这个条件很有用，因为它使我们能够利用 $\widehat{\Sigma} - \Sigma$的矩，而不是较难解释的 $\widehat{\Sigma}^{-1} - \Sigma^{-1}$的矩，来界定 $\overline{Z}^\top \widehat{\Sigma}^{-1} \overline{Z}$的矩。结合其他条件，对$\widehat{\Sigma} - \Sigma$ 的矩假设足以控制 $\overline{Z}^\top \widehat{\Sigma}^{-1} \overline{Z}$的前两个矩，它们取决于方差$\overline{Z}$以及 Σ的方差。

证明。 我们首先证明均值结果。根据迭代期望定律以及二次型的性质

$$\begin{aligned}
\mathbb{E}[\overline{Z}^\top \widehat{\Sigma}^{-1} \overline{Z}] &= \mathbb{E}\{\mathbb{E}[\overline{Z}^\top \widehat{\Sigma}^{-1} \overline{Z} \mid \widehat{\Sigma}^{-1}]\} \\
&= \mathbb{E}\left[\mu^\top \widehat{\Sigma}^{-1} \mu + \mathrm{Tr}\{\mathrm{Cov}(\overline{Z}) \widehat{\Sigma}^{-1}\}\right] \\
&= \mu^\top \Sigma^{-1} \mu + N^{-1} \mathrm{Tr}(\Sigma \Sigma^{-1}) + \mathbb{E}\left[\mu^\top (\widehat{\Sigma}^{-1} - \Sigma^{-1}) \mu + N^{-1} \mathrm{Tr}[\Sigma(\widehat{\Sigma}^{-1} - \Sigma^{-1})]\right].
\end{aligned}$$

对于维度相近的实方阵 A、B，以及 A 的弗罗贝尼乌斯范数 $\|A\|_F$，回顾以下不等式

$$|\mathrm{Tr}(AB)| \leq \|A\|_F \|B\|_F$$

在矩阵上 在固定维度下，所有矩阵范数都是等价的，因此我们有 t

$$|\mathrm{Tr}(AB)| \leq c\|A\|\|B\|$$

对于某个有限的 $c > 0$ 以及我们选定的矩阵范数 $\|\cdot\|$。利用三角不等式和上述不等式，可得

$$\begin{aligned}|\mu^\top[\widehat{\Sigma}^{-1} - \Sigma^{-1}]\mu + N^{-1}\mathrm{Tr}\{\Sigma(\widehat{\Sigma}^{-1} - \Sigma^{-1})\}| \leq & |\mu^\top(\widehat{\Sigma}^{-1} - \Sigma^{-1})\mu| + N^{-1}|\mathrm{Tr}\{\Sigma(\widehat{\Sigma}^{-1} - \Sigma^{-1})\}| \\ \lesssim & \|\mu\mu^\top\|\|\widehat{\Sigma}^{-1} - \Sigma^{-1}\| + N^{-1}\|\Sigma\|\|\widehat{\Sigma}^{-1} - \Sigma^{-1}\|.\end{aligned}$$

根据引理5，对于 N 足够大，使得 $\|\Sigma(\widehat{\Sigma}^{-1} - \Sigma^{-1})\| < 1$,

$$\|[\widehat{\Sigma}^{-1} - \Sigma^{-1}]\| \lesssim \|\Sigma^{-1}\|^2\|\widehat{\Sigma} - \Sigma\|. \tag{7}$$

应用等式（7）即可得到

$$\begin{aligned}|\mu^\top[\widehat{\Sigma}^{-1} - \Sigma^{-1}]\mu + N^{-1}\mathrm{Tr}\Sigma[\widehat{\Sigma}^{-1} - \Sigma^{-1}]| \lesssim & \{\|\mu\mu^\top\| + N^{-1}\|\Sigma\|\}\|\widehat{\Sigma}^{-1} - \Sigma^{-1}\| \\ \lesssim & \{\|\mu\mu^\top\| + N^{-1}\|\Sigma\|\}\|\Sigma^{-1}\|^2\|\widehat{\Sigma} - \Sigma\|.\end{aligned}$$

取期望，

$$\begin{aligned}\mathbb{E}|\mu^\top[\widehat{\Sigma}^{-1} - \Sigma^{-1}]\mu + N^{-1}\text{迹}\{\Sigma(\widehat{\Sigma}^{-1} - \Sigma^{-1})\}| \lesssim & \|\mu\mu^\top\|\|\Sigma^{-1}\|^2\sigma^2/N + \|\Sigma\|\|\Sigma^{-1}\|^2\sigma^2/N^2 \\ \lesssim & \sigma^4\|\Sigma^{-1}\|^2/N + \sigma^4\|\Sigma^{-1}\|^2/N^2,\end{aligned}$$

其中我们用到了以下事实：$\|\mu\mu^\top\| \lesssim \|\mu^2\|^2 \lesssim \sigma^2$，以及 $\|\Sigma\| \lesssim \sigma^2$。

为了给方差建立上界，我们使用迭代方差定律：

$$\mathrm{Var}(\overline{Z}^\top\widehat{\Sigma}^{-1}\overline{Z}) = \mathbb{E}\{\mathrm{Var}(\overline{Z}^\top\widehat{\Sigma}^{-1}\overline{Z} \mid \widehat{\Sigma})\} + \mathrm{Var}[\mathbb{E}(\overline{Z}^\top\widehat{\Sigma}^{-1}\overline{Z} \mid \widehat{\Sigma})], \tag{8}$$

并分别控制每一项。对于（8）中的第一项，我们可以直接应用引理4，^在 Σ 和（7）的条件下，得到如下界

$$\begin{aligned}\mathrm{Var}(\overline{Z}^\top\widehat{\Sigma}^{-1}\overline{Z} \mid \widehat{\Sigma}) \lesssim \|\sigma^2\widehat{\Sigma}^{-1}\|^2/N \leq & \frac{\sigma^4}{N}\left\{\|(\widehat{\Sigma}^{-1} - \Sigma^{-1})\|^2 + \|\Sigma^{-1}\|^2\right\} \\ \lesssim & \frac{\sigma^4}{N}\|\Sigma^{-1}\|^2\left(1 + \|\Sigma^{-1}\|^2\|\widehat{\Sigma} - \Sigma\|^2\right);\end{aligned}$$

^并对 $\Sigma - \Sigma$ 应用矩界，我们得到

$$\mathbb{E}\{\mathrm{Var}(\overline{Z}^\top\widehat{\Sigma}^{-1}\overline{Z} \mid \widehat{\Sigma})\} \lesssim \frac{\sigma^4}{N}\|\Sigma^{-1}\|^2(1 + \|\Sigma^{-1}\|^2\sigma^4/N) = O(N^{-1}\|\sigma^2\Sigma^{-1}\| \vee N^{-2}\|\sigma^2\Sigma^{-1}\|^4).$$

由于 $\|\sigma^2\Sigma^{-1}\| < \infty$，（8）中的第一项有上界 O（$N^{-1}\|\sigma^2\Sigma^{-1}\|$），对所有 N充分大的情形均成立。

^为了界定（8）中的第二项，使用第一个期望的证明，并以 Σ 为条件，可得

$$\begin{aligned}\mathbb{E}(\overline{Z}^\top\widehat{\Sigma}^{-1}\overline{Z} \mid \widehat{\Sigma}^{-1}) &= \mu^\top\widehat{\Sigma}^{-1}\mu + \mathrm{Tr}(\Sigma\widehat{\Sigma}^{-1})/N \\ &= \mu^\top\Sigma^{-1}\mu + \mu^\top(\widehat{\Sigma}^{-1} - \Sigma^{-1})\mu + N^{-1}\mathrm{Tr}[\Sigma(\widehat{\Sigma}^{-1} - \Sigma^{-1})] + d/N.\end{aligned}$$

对上述结果取方差，并使用柯西–施瓦茨不等式，可得

$$\begin{aligned}\mathrm{Var}[\mathbb{E}(\overline{Z}^\top\widehat{\Sigma}^{-1}\overline{Z}\mid\widehat{\Sigma}^{-1})] \le \mathrm{Var}[\mathrm{Tr}(\widehat{\Sigma}^{-1}-\Sigma^{-1})\mu\mu^\top] + N^{-2}\mathrm{Var}[\mathrm{Tr}(\widehat{\Sigma}^{-1}-\Sigma^{-1})\Sigma] \\ + 2\sqrt{\mathrm{Var}[\mathrm{Tr}(\widehat{\Sigma}^{-1}-\Sigma^{-1})\mu\mu^\top]}\sqrt{N^{-2}\mathrm{Var}[\mathrm{Tr}(\widehat{\Sigma}^{-1}-\Sigma^{-1})\Sigma]}\end{aligned} \tag{9}$$

考虑方程（9）中的第一项；根据迹不等式

$$|\mathrm{Tr}(\widehat{\Sigma}^{-1}-\Sigma^{-1})\mu\mu^\top| \le \|\widehat{\Sigma}^{-1}-\Sigma^{-1}\|\mathrm{Tr}(\mu\mu^\top) \lesssim \sigma^2\|\widehat{\Sigma}^{-1}-\Sigma^{-1}\|.$$

应用方程（7），可得

$$|\mathrm{Tr}[(\widehat{\Sigma}^{-1}-\Sigma^{-1})\mu\mu^\top]| \lesssim \sigma^2\|\Sigma^{-1}\|^2\|\widehat{\Sigma}-\Sigma\|.$$

因此，利用矩界假设，

$$\mathrm{Var}\{\mathrm{Tr}[(\widehat{\Sigma}^{-1}-\Sigma^{-1})\mu\mu^\top]\} \lesssim \sigma^4\|\Sigma^{-1}\|^4\mathbb{E}\|\widehat{\Sigma}-\Sigma\|^2 \lesssim \|\sigma^2\Sigma^{-1}\|^4/N. \tag{10}$$

对于（9）中的第二项，同样根据迹不等式以及（7）

$$|\mathrm{Tr}\{\Sigma(\widehat{\Sigma}^{-1}-\Sigma^{-1})\}| \lesssim \|\Sigma\|\|\Sigma^{-1}-\Sigma^{-1}\| \lesssim \sigma^2\|\Sigma^{-1}-\Sigma^{-1}\| \lesssim \sigma^2\|\Sigma^{-1}\|\|\widehat{\Sigma}-\Sigma\|,$$

由于 $\|\Sigma\| \le \sigma^2$。因此，与（9）中的第一项类似

$$\mathrm{Var}[\mathrm{Tr}\{\Sigma(\widehat{\Sigma}^{-1}-\Sigma^{-1})\}] \lesssim \sigma^4\|\Sigma^{-1}\|^4\mathbb{E}\|\widehat{\Sigma}-\Sigma\|^2 \lesssim \|\sigma^2\Sigma^{-1}\|^4/N. \tag{11}$$

将式（10）-（11）中的方差界代入式（9），并重新整理各项，可得

$$\begin{aligned}\mathrm{Var}\{\mathbb{E}(\overline{Z}^\top\widehat{\Sigma}^{-1}\overline{Z}\mid\widehat{\Sigma}^{-1})\} &\lesssim \|\sigma^2\Sigma^{-1}\|^4/N + \|\sigma^2\Sigma^{-1}\|^4/N^3 + \|\sigma^2\Sigma^{-1}\|^4/N^2 \\ &= O(\|\sigma^2\Sigma^{-1}\|^4/N).\end{aligned}$$

□

引理 7. 设 Z 是一个均值为 μ 的正标量值随机变量，满足$\mathbb{E}\left[(Z-\mu)^2\right] \le Bb^2$，其中 B，$b>0$。那么，对任意 $0 \le \lambda \le 1/b$，

$$\mathbb{E}\left[\exp(-\lambda Z)\right] \le \exp(-\lambda\mu) + B(\lambda b)^2.$$

*Pr*证明。 对任意 $Z \ge 0$，将 $\exp(-\lambda Z)$ 在 $Z=\mu$ 处展开泰勒展开，可得

$$\exp(-\lambda Z) = \exp(-\lambda\mu)\sum_{n=0}^{\infty}\frac{(-1)^n\lambda^n}{n!}(Z-\mu)^n = \mathcal{P}(Z) + \mathcal{E}(Z),$$

其中 $\mathcal{P}(Z)$ 是该级数的截断

$$\mathcal{P}(Z) := \exp(-\lambda\mu)\left\{1-\lambda(Z-\mu)\right\}$$

且 $\mathcal{E}(Z)$ 是拉格朗日余项，

$$\mathcal{E}(Z) := \exp(-\lambda\zeta)\frac{\lambda^2}{2!}(Z-\mu)^2$$

其中 ζ 是介于 Z 和 μ 之间的常数。

$\exp(-x)$ 在 $x \geq 0$ 上有界，这意味着对于任意 $\lambda \geq 0$ 和 $Z \geq 0$

$$\mathcal{E}(Z) \leq \lambda^2 (Z-\mu)^2.$$

因此，

$$\exp(-\lambda Z) = \mathcal{P}(Z) + \mathcal{E}(Z) \leq \exp(-\lambda\mu)\left\{1 - \lambda(Z-\mu)\right\} + \lambda^2 (Z-\mu)^2.$$

根据假设，对于某些 B、$b > 0$，有 $\mathbb{E}\left[(Z-\mu)^2\right] \leq Bb^2 < \infty$。对两边取期望可得

$$\begin{aligned}\mathbb{E}\left[\exp(-\lambda Z)\right] &\leq \exp(-\lambda\mu)\left\{1 - \lambda\mathbb{E}(Z-\mu)\right\} + \lambda^2 \mathbb{E}(Z-\mu)^2 \\ &\leq \exp(-\lambda\mu) + B(\lambda b)^2.\end{aligned}$$

□

以下结果表明，若假设 2–4 中的正则条件对 $m_n(\theta)$ 和 $W_n(\theta)$ 成立，则（可能不可行的）Q-后验分布 $\pi_n^Q(\theta)$ 渐近地服从高斯分布。当然，若正则条件对 $m_n(\theta) = \psi_n(\theta)$ 和 $W_n(\theta) = V_n(\theta)$ 成立，则该结果也适用于基于损失函数 $q_n(\theta)$ 且满足 $\Sigma_\star = \Omega_\star$ 的 Q-后验分布 $\pi_n^\Psi(\theta)$（详情见第4节）。

引理8。 以下 假设 *1-4*，如 $n \to +\infty$，$\int_{\mathcal{T}_n} \|t\| |\pi_n^Q(\vartheta) - N\{\vartheta;\ 0,\ \Sigma_\star^{-1}\}| \mathrm{d}\vartheta = o_p(1)$.

引理*8*的证明。 证明该结果所采用的方法与 Lehmann 和 Casella(2003年)第6章定理8.2中给出的方法，以及 Chernozhukov 和 Hong (2003年)中的定理1类似。不过，我们的证明有所不同，因为矩阵 $W_n(\theta)$ 允许在 $\theta_\star$之外是奇异的，并且 Θ 不假设为紧集。这使得该证明与 Frazier 等人(2022)中的引理1最为相似。

在本证明的其余部分中，我们滥用记号并定义

$$Q_n(\theta) = -\frac{1}{2} \frac{m_n(\theta)^\top}{\sqrt{n}} W_n(\theta)^{-1} \frac{m_n(\theta)}{\sqrt{n}}.$$

对于适当定义的余项 $R_n(\theta)$，我们有恒等式

$$\begin{aligned}Q_n(\theta) - Q_n(\theta_\star) &= -\{m_n(\theta)/\sqrt{n}\}^\top W(\theta_\star)^{-1} \mathcal{H}(\theta_\star)^\top \Sigma_\star^{-1} \Sigma_\star \sqrt{n}(\theta - \theta_\star) - \frac{n}{2}(\theta-\theta_\star)\Sigma_\star(\theta-\theta_\star) + R_n(\theta) \\ &= -\frac{1}{2} t^\top \Sigma_\star t + \frac{1}{2}\frac{1}{n} Z_n^\top \Sigma_\star Z_n + R_n(\theta), \qquad (12)\end{aligned}$$

其中 $Z_n := \Sigma_\star^{-1} \mathcal{H}(\theta_\star)^\top W_\star^{-1} m_n(\theta_\star)$，且 $t := \sqrt{n}(\theta - \theta_\star) - Z_n/\sqrt{n}$。定义 $T_n := \theta_\star + Z_n/n$，

$$\omega(t) := Q_n\left(T_n + t/\sqrt{n}\right) - Q_n(\theta_\star) - \frac{1}{2n} Z_n^\top \Sigma_\star Z_n,$$

并应用（12）可得

$$\omega(t) = -\frac{1}{2} t^\top \Sigma_\star t + R_n(T_n + t/\sqrt{n}).$$

回想一下 $\mathcal{T}_n = \{\sqrt{n}(\theta - \theta_\star) - Z_n/\sqrt{n} : \ \theta \in \Theta\}$，以及对于 $M_n(\theta) = |W_n(\theta)|^{-1/2}$，

$$\begin{aligned}
\pi_n^Q(t) &= \frac{M_n\left(T_n + t/\sqrt{n}\right) \exp\left\{Q_n\left(T_n + t/\sqrt{n}\right)\right\} \pi\left(T_n + t/\sqrt{n}\right)}{\int_{\mathcal{T}_n} M_n\left(T_n + t/\sqrt{n}\right) \exp\left\{Q_n\left(T_n + t/\sqrt{n}\right)\right\} \pi\left(T_n + t/\sqrt{n}\right) \mathrm{d}t} \\
&= \frac{M_n\left(T_n + t/\sqrt{n}\right) \exp\left\{Q_n\left(T_n + t/\sqrt{n}\right) - Q_n(\theta_\star) - \frac{1}{2n} Z_n^\top \Sigma_\star^{-1} Z_n\right\} \pi\left(T_n + t/\sqrt{n}\right)}{\int_{\mathcal{T}_n} M_n\left(T_n + t/\sqrt{n}\right) \exp\left\{Q_n\left(T_n + t/\sqrt{n}\right) - Q_n(\theta_\star) - \frac{1}{2n} Z_n^\top \Sigma_\star^{-1} Z_n\right\} \pi\left(T_n + t/\sqrt{n}\right) \mathrm{d}t} \\
&= M_n\left(T_n + t/\sqrt{n}\right) \exp\{\omega(t)\} \pi\left(T_n + t/\sqrt{n}\right) / C_n,
\end{aligned}$$

其中

$$C_n = \int_{\mathcal{T}_n} M_n\left(T_n + t/\sqrt{n}\right) \exp\{\omega(t)\} \pi\left(T_n + t/\sqrt{n}\right) \mathrm{d}t.$$

如果满足以下条件，则所述结果成立

$$\int_{\mathcal{T}_n} \|t\|^\gamma \left|\pi_n^Q(t) - N\{t; 0, \Sigma_\star^{-1}\}\right| \mathrm{d}t = C_n^{-1} J_n = o_p(1),$$

其中

$$\begin{aligned}
J_n &= \int_{\mathcal{T}_n} \|t\|^\gamma \left| M_n\left(T_n + \frac{t}{\sqrt{n}}\right) \exp\{\omega(t)\} \pi\left(T_n + \frac{t}{\sqrt{n}}\right) - C_n N\{t; 0, \Sigma_\star^{-1}\} \right| \mathrm{d}t \\
&= \int_{\mathcal{T}_n} \|t\|^\gamma \left| M_n\left(T_n + \frac{t}{\sqrt{n}}\right) \exp\{\omega(t)\} \pi\left(T_n + \frac{t}{\sqrt{n}}\right) - C_n \frac{|\Sigma_\star|^{1/2}}{(2\pi)^{d_\theta/2}} \exp\left\{-\frac{1}{2} t^\top \Sigma_\star t\right\} \right| \mathrm{d}t.
\end{aligned}$$

然而，

$$J_n \le J_{1n} + J_{2n},$$

其中

$$\begin{aligned}
J_{1n} &:= \int_{\mathcal{T}_n} \|t\|^\gamma \left| M_n\left(T_n + \frac{t}{\sqrt{n}}\right) \exp\{\omega(t)\} \pi\left(T_n + \frac{t}{\sqrt{n}}\right) - \pi(\theta_\star) |M(\theta_\star)|^{\frac{1}{2}} \exp\left\{-\frac{1}{2} t^\top \Sigma_\star t\right\} \right| \mathrm{d}t \\
J_{2n} &:= \left| C_n \frac{|\Sigma_\star|^{1/2}}{(2\pi)^{d_\theta/2}} - \pi(\theta_\star) M(\theta_\star) \right| \int_{\mathcal{T}_n} \|t\|^\gamma \exp\left\{-\frac{1}{2} t^\top \Sigma_\star t\right\} \mathrm{d}t,
\end{aligned}$$

对于 $M(\theta_\star) := |W(\theta_\star)|^{-1/2}$。

注意，若 $J_{1n} = o_p(1)$，则

$$C_n = \pi(\theta_\star) M(\theta_\star) \int_{\mathbb{R}^{d_\theta}} \exp\left\{-\frac{1}{2} t^\top \Sigma_\star t\right\} \mathrm{d}t + o_p(1) = \pi(\theta_\star) M(\theta_\star) \frac{(2\pi)^{d_\theta/2}}{|\Sigma_\star|^{1/2}} + o_p(1).$$

因此，如果 $J_{1n} = o_p(1)$，

$$J_{2n} = \left| C_n \frac{|\Sigma_\star|^{1/2}}{(2\pi)^{d_\theta/2}} - \pi(\theta_\star) M(\theta_\star) \right| \int_{\mathcal{T}_n} \|t\|^\gamma \exp\left\{-\frac{1}{2} t^\top \Sigma_\star t\right\} \mathrm{d}t = o_p(1),$$

因为对于任意 $0 \leq \gamma < \infty$， $\int_{\mathbb{R}^{d_\theta}} \|t\|^\gamma \exp\left\{-\frac{1}{2}t^\top \Sigma_\star t\right\} \mathrm{d}t < \infty$。

因此，如果我们能够证明 $J_{1n} = o_p(1)$，则结果成立。为此，我们将 $\mathcal{T}_n$ 划分为三个区域，并分别分析每个区域上的 J_{1n}。对于满足 $\delta = o(1)$ 的某些 $0 \leq h < \infty$ 和 $\delta > 0$，这些区域定义如下。区域1： $\|t\| \leq h$；区域2：$h < \|t\| \leq \delta\sqrt{n}$；区域3： $\|t\| \geq \delta\sqrt{n}$。

区域1： 在此区域内，如果

$$\|t\|^\gamma \left| M_n\left(T_n + \frac{t}{\sqrt{n}}\right) \exp\{\omega(t)\} \pi\left(T_n + \frac{t}{\sqrt{n}}\right) - \pi(\theta_\star) M(\theta_\star) \exp\left\{-\frac{1}{2}t^\top \Sigma_\star t\right\} \right| = o_p(1).$$

注意，根据假设 3 和 4，

$$\sup_{\|t\| \leq h} \left\| M_n\left(T_n + t/\sqrt{n}\right) - M(\theta_\star) \right\| = o_p(1), \text{ and } \sup_{\|t\| \leq h} \left| \pi\left(T_n + t/\sqrt{n}\right) - \pi(\theta_\star) \right| = o_p(1),$$

and

$$T_n = \theta_\star + Z_n/n = \theta_\star + o_p(1),$$

由于 $Z_n/\sqrt{n} = O_p(1)$ 根据假设 2。同样，根据假设 2，

$$\sup_{\|t\| \leq h} \left\| T_n + t/\sqrt{n} - \theta_\star \right\| = O_p(1/\sqrt{n})$$

因此根据引理 9

$$\sup_{\|t\| \leq h} |R_n(T_n + t/\sqrt{n})| = o_p(1).$$

因此，根据这些等价关系和控制收敛定理， $J_{1n} = o_p(1)$。

区域 2： 如果 $\delta = o(1)$，那么上确界 $_{h \leq \|t\| \leq \delta\sqrt{n}} \|M_n\left(T_n + t/\sqrt{n}\right) - M(\theta_\star)\| = o_p(1)$ 根据假设 2 和连续映射定理成立。于是，当 $J_{1n} \leq C_{1n} + C_{2n} + C_{3n}$ 对于 h 足够大且 $\delta = o(1)$时，其中

$$C_{1n} := C \int_{h \leq \|t\| \leq \delta\sqrt{n}} \|t\|^\gamma \exp(-t^\top \Sigma_\star t/2) \sup_{h \leq \|t\| \leq \delta\sqrt{n}} \left| \exp\left\{|R_n(T_n + t/\sqrt{n})|\right\} \left\{\pi\left(T_n + t/\sqrt{n}\right) - \pi\left(\theta_\star\right)\right\} \right| \mathrm{d}t$$

$$C_{2n} := C \int_{h \leq \|t\| \leq \delta\sqrt{n}} \|t\|^\gamma \exp(-t^\top \Sigma_\star t/2) \exp\left\{|R_n(T_n + t/\sqrt{n})|\right\} \pi\left(T_n + t/\sqrt{n}\right) \mathrm{d}t$$

$$C_{3n} := C\pi\left(\theta_\star\right) \int_{h \leq \|t\| \leq \delta\sqrt{n}} \|t\|^\gamma \exp(-t^\top \Sigma_\star t/2) \mathrm{d}t.$$

对于任意固定的 h，第一项 $C_{1n} = o_p(1)$，因此根据控制收敛定理，当 $h \to +\infty$ 时， $C_{1n} = o_p(1)$。对于 C_{3n}，我们有，对于任意 $0 \leq \gamma \leq 2$，都存在一个足够大的 h'，使得对所有 $h > h'$，且 $\|t\| \geq h$

$$\|t\|^\gamma \exp\left(-t^\top \Sigma_\star t/2\right) = O(1/h).$$

因此， C_{3n} 可以 通过使 h 足够大且 δ 足够小，可以将其任意减小。

若 $C_{2n} = o_p(1)$，则该结果成立。为此，我们证明，对于某个 $C > 0$ 以及所有 $h \le \|t\| \le \delta\sqrt{n}$，其概率趋近于一（wpc1），

$$\exp(-t^\top \Sigma_\star t/2) \exp\left\{|R_n(T_n + t/\sqrt{n})|\right\} \pi(T_n + t/\sqrt{n}) \le C \exp\left\{-t^\top \Sigma_\star t/4\right\}. \tag{13}$$

如果方程（13）成立，则 C_{2n} 的上界为

$$C_{2n} \le C \int_{h \le \|t\| \le \delta\sqrt{n}} \|t\|^\gamma \exp\left(-t^\top \Sigma_\star t/4\right) \mathrm{d}t,$$

对于某些 h 足够大且 δ 足够小时，该值可以任意小。为证明方程（13），首先注意到，由于 $\pi(\theta)$ 的连续性，假设 4，$\pi(T_n + t/\sqrt{n})$ 在$\{t$ 上有界：$h \le \|t\| \le \delta\sqrt{n}\}$ 因此可以在分析中将其略去。

现在，由于 $\|T_n - \theta_\star\| = o_p(1)$，对于任意 $\delta > 0$，$\|T_n + t/\sqrt{n} - \theta_\star\| < 2\delta$ 对于所有 $\|t\| \le \delta\sqrt{n}$以及 n 足够大时。因此，根据引理 9，存在某个 $\delta' > 0$ 以及 h 足够大，使得（wpc1）

$$\sup_{h \le \|t\| \le \delta'\sqrt{n}} \frac{|R_n(T_n + t/\sqrt{n})|}{1 + \|t + \Sigma_\star^{-1} Z_n/\sqrt{n}\|^2} \le \frac{1}{4}\lambda_{\min}(\Sigma_\star),$$

其中，$\lambda_{\min}(\Sigma_\star)$ 表示 $\Sigma_\star$的最小特征值。由于 $Z_n/\sqrt{n} = O_p(1)$，我们有$n^{-1} Z_n^\top \Sigma_\star^{-1} Z_n = O_p(1)$。因此，在集合 $h \le \|t\| \le \delta'\sqrt{n}$上

$$|R_n(T_n + t/\sqrt{n})| \le \frac{1}{4}\lambda_{\min}(\Sigma_\star) + \frac{1}{4}\|t + \Sigma_\star^{-1} Z_n/\sqrt{n}\|^2 \lambda_{\min}(\Sigma_\star) \le O_p(1) + \frac{1}{4}\|t\|^2 \lambda_{\min}(\Sigma_\star),$$

以及对于某个 $C > 0$

$$\begin{aligned}
\exp\{\omega(t)\} &\le \exp\left\{-\frac{1}{2}t^\top \Sigma_\star t + |R_n(T_n + t/\sqrt{n})|\right\} \\
&\le C \exp\left\{-\frac{1}{2}t^\top \Sigma_\star t + \frac{1}{4}\lambda_{\min}(\Sigma_\star)\|t\|^2\right\} \\
&\le C \exp\left(-\frac{1}{4}t^\top \Sigma_\star t\right).
\end{aligned}$$

结果随即得出。

区域3： 对于 $\delta\sqrt{n}$较大的，

$$\int_{\|t\| \ge \delta\sqrt{n}} \|t\|^\gamma N\{t; 0, \Sigma_\star^{-1}\} \mathrm{d}t$$

可以任意小，因此在分析中将其忽略。利用 $\omega(t)$ 的定义和恒等式 $\theta = T_n + t/\sqrt{n}$，考虑

$$\begin{aligned}
J_{1n} &:= \int_{\|t\| \ge \delta\sqrt{n}} \|t\|^\gamma M_n\left(T_n + t/\sqrt{n}\right) \exp\{\omega(t)\} \pi\left(T_n + t/\sqrt{n}\right) \mathrm{d}t, \\
&= n^{(d_\theta + \gamma)/2} \int_{\|\theta - T_n\| \ge \delta} \|\theta - T_n\|^\gamma M_n(\theta) \exp\left\{Q_n(\theta) - Q_n(\theta_\star) - \frac{1}{2n} Z_n^\top \Sigma_\star^{-1} Z_n\right\} \pi(\theta) \mathrm{d}\theta.
\end{aligned}$$

Now,

$$J_{1n} = \exp\left\{-\frac{1}{2n} Z_n^\top \Sigma_\star^{-1} Z_n\right\} n^{(d_\theta+\gamma)/2} \int_{\|\theta - T_n\| \geq \delta} \|\theta - T_n\|^\gamma M_n(\theta) \exp\left\{Q_n(\theta) - Q_n(\theta_\star)\right\} \pi(\theta)\, \mathrm{d}\theta,$$

$$= O_p(1) n^{(d_\theta+\gamma)/2} \int_{\|\theta - T_n\| \geq \delta} \|\theta - T_n\|^\gamma M_n(\theta) \exp\left\{Q_n(\theta) - Q_n(\theta_\star)\right\} \pi(\theta)\, \mathrm{d}\theta$$

因为 $n^{-1} Z_n^\top \Sigma_\star^{-1} Z_n = O_p(1)$ 根据假设2。

定义 $Q(\theta) := -m(\theta)^\top W(\theta) m(\theta)/2$，并注意到 $Q(\theta_\star) = 0$ 根据假设 2 以及 $W(\theta_\star)$根据假设 3是正定的。因此，对于任意 $\delta > 0$,

$$\sup_{\|\theta - \theta_\star\| \geq \delta} \frac{1}{n} \left\{Q_n(\theta) - Q_n(\theta_\star)\right\} \leq \sup_{\|\theta - \theta_\star\| \geq \delta} 2|n^{-1} Q_n(\theta) - Q(\theta)| + \sup_{\|\theta - \theta_\star\| \geq \delta} \left\{Q(\theta) - Q(\theta_\star)\right\}.$$

由假设 2 和 3，第一项依概率收敛至零。由假设 3，对于任意 $\delta > 0$，存在一个 $\epsilon > 0$，使得

$$\sup_{\|\theta - \theta_\star\| \geq \delta} \left\{Q(\theta) - Q(\theta_\star)\right\} \leq -\epsilon.$$

因此,

$$\lim_{n \to +\infty} P_0^{(n)} \left[\sup_{\|\theta - \theta_\star\| \geq \delta} \exp\left\{Q_n(\theta) - Q_n(\theta_\star)\right\} \leq \exp(-\epsilon n)\right] = 1. \tag{14}$$

使用 $T_n = \theta_\star + o_p(1)$、定义 $M_n(\theta) = |W_n(\theta)^{-1}|^{1/2}$以及方程（14）可得

$$\begin{aligned} J_{1n} &\leq \{1 + o_p(1)\} O_p(1) n^{(d_\theta+\gamma)/2} \int_{\|\theta - \theta_0\| \geq \delta} M_n(\theta) \|\theta - \theta_\star\|^\gamma \pi(\theta) \exp\{Q_n(\theta) - Q_n(\theta_\star)\} \mathrm{d}\theta \\ &\leq O_p(1) \exp(-\epsilon n)\, n^{(d_\theta+\gamma)/2} \int_{\|\theta - \theta_\star\| \geq \delta} M_n(\theta) \|\theta - \theta_\star\|^\gamma \pi(\theta)\, \mathrm{d}\theta \\ &\leq O_p\left\{\exp(-\epsilon n)\, n^{(d_\theta+\gamma)/2}\right\} \\ &= o_p(1); \end{aligned}$$

其中第二个不等式由假设 4中的矩假设得到。 □

以下结果用于引理的证明2，且是Chernozhukov 和 Hong (2003年)中的命题1的推论。

引理9。 在假设 *2*下，且对于 $R_n(\theta)$ 如引理 *2*的证明中所定义，对每个 $\epsilon > 0$ 都存在一个足够小的 $\delta > 0$ 以及 $h > 0$ 足够大的值，使得

$$\limsup_{n \to +\infty} P_0^{(n)} \left[\sup_{h/\sqrt{n} \leq \|\theta - \theta_0\| \leq \delta} \frac{|R_n(\theta)|}{1 + \sqrt{n}^2 \|\theta - \theta_0\|^2} > \epsilon\right] < \epsilon$$

and

$$\limsup_{n \to +\infty} P_0^{(n)} \left[\sup_{\|\theta - \theta_0\| \leq h/\sqrt{n}} |R_n(\theta)| > \epsilon\right] = 0.$$

证明。 该结果是 Chernozhukov 和 Hong (2003年) 中命题1的一个特例。因此，只需验证其充分条件在我们的情形下成立。

他们结果中的假设(i)-(iii)直接由假设2 和3以及假设 $\sqrt{n}m_n(\theta_\star)$ 中的正态性2推出。他们的假设(iv)表述如下：对于任意$\epsilon > 0$，存在一个 $\delta > 0$ 使得

$$\limsup_{n\to+\infty} P_0^{(n)}\left\{\sup_{\|\theta-\theta'\|\leq\delta}\frac{\sqrt{n}\|\{m_n(\theta)-m_n(\theta')\}-\{\mathbb{E}\left[m_n(\theta)\right]-\mathbb{E}\left[m_n(\theta')\right]\}\|}{1+n\|\theta-\theta'\|}>\epsilon\right\}<\epsilon.$$

在我们的情形下，假设2(iv)满足此条件。因此，该结论得证。 □

Appendix B. Proofs of Main Results

引理*1*的证明。考虑均值参数化 $\mu = g(\eta) = \nabla_\eta A(\eta)$下的 Q-后验分布，其先验信念为 $\pi(\mu) \propto \exp\{-\frac{1}{2}(\mu-\mu_0)^\top W_0^{-1}(\mu-\mu_0)\}$，其中 μ_0和 W_0是已知的先验超参数。记 $\bar{S}_n = n^{-1}S_n$，可得

$$n^{-1}m_n(\eta) = g(\eta) - \bar{S}_n = \mu - \bar{S}_n = n^{-1}m_n(\mu).$$

因此，记为 $\bar{S}_n = n^{-1}S_n$，则 μ 的 Q-后验分布为

$$\pi_n^Q(\mu) \propto \exp\{-\frac{n}{2}\left(\mu-\bar{S}_n\right)^\top W_n^{-1}\left(\mu-\bar{S}_n\right)\}\exp\{-\frac{1}{2}(\mu-\mu_0)^\top W_0^{-1}(\mu-\mu_0)\}.$$

经代数运算可得

$$\begin{aligned}
&\exp\left\{-\frac{1}{2}\mu^\top\left[(W_n/n)^{-1}+W_0^{-1}\right]\mu-\mu^\top\left[(W_n/n)^{-1}\bar{S}_n+\mu_0\right]\right\}\\
&=\exp\left\{-\frac{1}{2}\mu^\top\left[(W_n/n)^{-1}+W_0^{-1}\right]\mu-\mu^\top\left[(W_n/n)^{-1}+W_0^{-1}\right]\left[(W_n/n)^{-1}+W_0^{-1}\right]^{-1}\left[(W_n/n)^{-1}\bar{S}_n+\mu_0\right]\right\}\\
&\propto\exp\left\{-\frac{1}{2}\left(\mu-\Sigma_n^{-1}\left[(W_n/n)^{-1}\bar{S}_n+\mu_0\right]\right)^\top\Sigma_n\left(\mu-\Sigma_n^{-1}\left[(W_n/n)^{-1}\bar{S}_n+\mu_0\right]\right)\right\}\\
&=\exp\left\{-\frac{1}{2}(\mu-b_n)^\top\Sigma_n(\mu-b_n)\right\}
\end{aligned}$$

for

$$\begin{aligned}
\Sigma_n^{-1} &= \left[(W_n/n)^{-1}+W_0^{-1}\right]^{-1}\\
&= W_0\left[n^{-1}W_n+W_0\right]^{-1}W_n\frac{1}{n},
\end{aligned}$$

并且第二个等式由伍德伯里恒等式推出，其中

$$\begin{aligned}
b_n &= \Sigma_n^{-1}\left[(W_n/n)^{-1}\bar{S}_n+\mu_0\right]\\
&= W_0\left[n^{-1}W_n+W_0\right]^{-1}\bar{S}_n+W_0\left[n^{-1}W_n+W_0\right]^{-1}W_n\frac{\mu_0}{n}\mu_0.
\end{aligned}$$

因此，我们看到 $\pi_n^Q(\mu) = N\{\mu;\ b_n, \Sigma_n^{-1}\}$。对于正则指数族，如果模型可识别（在 η 中），则存在从 $\mu \mapsto \eta = g^{-1}(\mu)$出发的参数变换。随后，变量替换 $\mu \mapsto \eta$ 得出

$$\pi_n^Q(\eta) = \pi_n^Q\{g(\eta)\}|\nabla_\eta g(\eta)| = N\{g(\eta); b_n, \Sigma_n^{-1}\}|\nabla_\eta^2 A(\eta)|,$$

其中，第二个等式成立是因为 $g(\eta) = \nabla_\eta A(\eta)$。 □

*Proo*定理*1*。 我们通过证明更一般的结果来证明该定理

$$\int_\Theta \|\varphi(\theta)\||\pi_n^Q(\theta) - \overline{\pi}_n^Q(\theta)|\mathrm{d}\theta = O(1/N).$$

取 $\varphi(\theta) = 1$，可得第一个结果；第二个结果则由以下事实得出：

$$\left\|\int_\Theta \varphi(\theta)\{\pi_n^Q(\theta) - \overline{\pi}_n^Q(\theta)\}\mathrm{d}\theta\right\| \le \int_\Theta \|\varphi(\theta)\||\pi_n^Q(\theta) - \overline{\pi}_n^Q(\theta)|\mathrm{d}\theta.$$

回顾以下定义

$$\widehat{Q}_n(\theta; z) := \frac{1}{2}\frac{\widehat{m}_n(\theta; z)^\top}{\sqrt{n}}\left[n^{-1}\widehat{W}_n(\theta; z)\right]^{-1}\frac{\widehat{m}_n(\theta; z)}{\sqrt{n}}$$

and

$$Q_n(\theta) := \frac{1}{2}\frac{m_n(\theta)^\top}{\sqrt{n}}\left[n^{-1}W_n(\theta)\right]^{-1}\frac{m_n(\theta)}{\sqrt{n}}.$$

令 $\mathbb{E}_z$ 表示在固定的 θ 和 $y_{1:n}$ 下，关于模拟数据 $z = (z_1, \ldots, z_N)$ 的期望，其中为简化记法，省略了 $\mathbb{E}_z$ 对 θ 和 $y_{1:n}$ 的依赖。我们首先证明，对所有 Θ,

$$\mathbb{E}_z[\exp\{-\widehat{Q}_n(\theta; z)\}] = \exp\{-Q_n(\theta)\}\{1 + O(1/N)\} + O\left\{\sigma_n^2(\theta)/N\right\}. \tag{15}$$

^我们首先利用迭代期望定律和引理 6 对 $\mathbb{E}_z[Q_n(\theta; z)]$ 进行界定。具体而言，令 $\overline{Z} = \widehat{m}_n(\theta;\ z)$、 $\widehat{\Sigma} = \widehat{W}_N(\theta;\ z)$、 $\sigma^2 = \sigma_n^2(\theta)$ 和 $\mu = m_n(\theta) + O\{\sigma_n^2(\theta)/N\}$，由引理 6 可得

$$\mathbb{E}_z[\widehat{Q}_n(\theta; z)] = Q_n(\theta) + O\{\|\sigma_n^2(\theta)W_n(\theta)^{-1}\|^2/N\}.$$

由于在假设 1 下，上确界$_{\theta\in\Theta}\|\sigma_n^2(\theta)W_n(\theta)^{-1}\| < \infty$ 对所有 $n \ge 1$ 均成立，因此可得$\|\sigma_n^2(\theta)W_n(\theta)^{-1}\|^2 < \infty$，由此我们得出

$$\mathbb{E}_z[\widehat{Q}_n(\theta; z)] = \frac{1}{2}m_n(\theta)^\top W_n(\theta)^{-1}m_n(\theta) + O(1/N). \tag{16}$$

现在我们对方差进行界定 $\mathbb{E}_z\{\widehat{Q}_n(\theta; z) - \mathbb{E}_z[\widehat{Q}_n(\theta; z)]\}^2$。使用与上文相同的定义，引理 6 表明，当 N 足够大时，存在某个 $C > 0$,

$$\mathbb{E}_z\{\widehat{Q}_n(\theta; z) - \mathbb{E}_z[\widehat{Q}_n(\theta; z)]\}^2 \le C\{\|\sigma_n^2(\theta)W_n(\theta)^{-1}\|^4\}\sigma_n^4(\theta)/N.$$

同样，根据假设 1，对于所有 $n \geq 1$，上确界$_{\theta\in\Theta} \|\sigma_n^2(\theta)W_n(\theta)^{-1}\|^4 < \infty$，因此

$$\mathbb{E}_z\{\widehat{Q}_n(\theta;z) - \mathbb{E}_z[\widehat{Q}_n(\theta;z)]\}^2 \lesssim \sigma_n^4(\theta)/N. \tag{17}$$

利用方程（16）和（17），我们可以给出 $\mathbb{E}_z[\exp\{-\lambda Q_n(\theta)\}]$的上界，其中 $0 \leq \lambda < \sqrt{N}$，根据引理 7，取 $b = 1/\sqrt{N}$，其中 $N \geq 1$由定义可知，并取 B使得$\mathbb{E}_z\{\widehat{Q}_n(\theta) - \mathbb{E}_z[\widehat{Q}_n(\theta)]\}^2 \leq Bb^2$；于是

$$\begin{aligned}\mathbb{E}_z[\exp\{-\lambda\widehat{Q}_n(\theta)\}] &\leq \exp\{-\lambda\mathbb{E}_z[\widehat{Q}_n(\theta)]\} + B(\lambda b)^2\\ &= \exp[-\lambda Q_n(\theta) + O\{\lambda/N\}] + O\{\lambda^2\sigma_n^4(\theta)/N\}\\ &= \exp\{-\lambda Q_n(\theta)\}[1 + O\{\lambda/N\}] + O\{\lambda^2\sigma_n^4(\theta)/N\}.\end{aligned}$$

由于上述结论对所有 $0 \leq \lambda \leq 1$ 都成立，因此不失一般性地，取 $\lambda = 1$，使 N 足够大，从而得到方程（15）。

定义 $\overline{g}_n(\theta) := \mathbb{E}_z \exp\{-Q_n(\theta;\ z)\}$，$g_n(\theta) := \exp\{-Q_n(\theta)\}$，并应用方程（15）得到

$$|\overline{g}_n(\theta) - g_n(\theta)| \leq g_n(\theta)O\{1/N\} + O\{\sigma_n^4(\theta)/N\}. \tag{18}$$

使得

$$\begin{aligned}\left|\int_\Theta \overline{g}_n(\theta)\pi(\theta)\mathrm{d}\theta - \int_\Theta g_n(\theta)\pi(\theta)\mathrm{d}\theta\right| &\leq \int_\Theta |\overline{g}_n(\theta) - g_n(\theta)|\pi(\theta)\mathrm{d}\theta\\ &\lesssim \frac{1}{N}\int_\Theta g_n(\theta)\pi(\theta)\mathrm{d}\theta + N^{-1}\int_\Theta \sigma_n^4(\theta)\pi(\theta)\mathrm{d}\theta.\end{aligned}$$

在该结论的假设下，$\int_\Theta g_n(\theta)\pi(\theta)\mathrm{d}\theta < \infty$ 对所有 $n \geq 1$ 均成立，因此

$$\begin{aligned}\left|\int_\Theta \|\varphi(\theta)\|\overline{g}_n(\theta)\pi(\theta)\mathrm{d}\theta - \int_\Theta \|\theta\|g_n(\theta)\pi(\theta)\mathrm{d}\theta\right| &\lesssim \frac{1}{N}\int_\Theta \|\varphi(\theta)\|\pi_n^Q(\theta)\mathrm{d}\theta + \frac{1}{N}\int_\Theta \|\varphi(\theta)\|\sigma_n^4(\theta)\pi(\theta)\mathrm{d}\theta\\ &= \frac{1}{N}\int_\Theta \|\varphi(\theta)\|\pi_n^Q(\theta)\mathrm{d}\theta + O(1/N),\end{aligned} \tag{19}$$

其中，第一个等式右侧的第一项是将该项乘以并除以 $\int_\Theta g_n(\theta)\pi(\theta)\mathrm{d}\theta$ 得到的；根据假设，对于所有 $n \geq 1$，该值均有限。等号成立是因为根据假设，$\int_\Theta \|\varphi(\theta)\|\sigma_n^4(\theta)\pi(\theta)\mathrm{d}\theta < \infty$。此外，根据假设，对于 $n \geq 1$，$\int_\Theta \|\varphi(\theta)\|\pi_n^Q(\theta)\mathrm{d}\theta < \infty$，并且

$$\begin{aligned}\left|\int_\Theta \|\varphi(\theta)\|\overline{g}_n(\theta)\pi(\theta)\mathrm{d}\theta - \int_\Theta \|\varphi(\theta)\|g_n(\theta)\pi(\theta)\mathrm{d}\theta\right| &\lesssim \frac{1}{N}\int_\Theta \|\varphi(\theta)\|\pi_n^Q(\theta)\mathrm{d}\theta + O(1/N)\\ &= O(1/N).\end{aligned} \tag{20}$$

由公式（20）可进一步推出

$$\frac{\left|\int_\Theta \overline{g}_n(\theta)\pi(\theta)\mathrm{d}\theta - \int_\Theta g_n(\theta)\pi(\theta)\mathrm{d}\theta\right|}{\int_\Theta g_n(\theta)\pi(\theta)\mathrm{d}\theta} = O(1/N), \tag{21}$$

因此

$$\frac{\int_\Theta \overline{g}_n(\theta)\pi(\theta)\mathrm{d}\theta}{\int_\Theta g_n(\theta)\pi(\theta)\mathrm{d}\theta} = 1 + O(1/N);\ \frac{\int_\Theta g_n(\theta)\pi(\theta)\mathrm{d}\theta}{\int_\Theta \overline{g}_n(\theta)\pi(\theta)\mathrm{d}\theta} = 1 + O(1/N).$$

将 $\overline{\pi}_n^Q(\theta) - \pi_n^Q(\theta)$ 写为

$$\begin{aligned}\overline{\pi}_n^Q(\theta) - \pi_n^Q(\theta) =& \frac{\overline{g}_n(\theta)\pi(\theta)}{\int_\Theta \overline{g}_n(\theta)\pi(\theta)\mathrm{d}\theta} - \frac{g_n(\theta)\pi(\theta)}{\int_\Theta g_n(\theta)\pi(\theta)\mathrm{d}\theta} \\ =& \{\overline{g}_n(\theta) - g_n(\theta)\} \frac{\pi(\theta)}{\int_\Theta g_n(\theta)\pi(\theta)\mathrm{d}\theta} \frac{\int_\Theta g_n(\theta)\pi(\theta)\mathrm{d}\theta}{\int_\Theta \overline{g}_n(\theta)\pi(\theta)\mathrm{d}\theta} \\ & - g_n(\theta)\pi(\theta)\left(\frac{1}{\int_\Theta g_n(\theta)\pi(\theta)\mathrm{d}\theta} - \frac{1}{\int_\Theta \overline{g}_n(\theta)\pi(\theta)\mathrm{d}\theta}\right),\end{aligned}$$

并应用三角不等式，得到

$$\begin{aligned}\left|\overline{\pi}_n^Q(\theta) - \pi_n^Q(\theta)\right| \leq& \left|\overline{g}_n(\theta) - g_n(\theta)\right| \frac{\pi(\theta)}{\int_\Theta \overline{g}_n(\theta)\pi(\theta)\mathrm{d}\theta} \\ & + \frac{\left|\int_\Theta \overline{g}_n(\theta)\pi(\theta)\mathrm{d}\theta - \int_\Theta g_n(\theta)\pi(\theta)\mathrm{d}\theta\right|}{\int_\Theta \overline{g}_n(\theta)\pi(\theta)\mathrm{d}\theta}\pi_n^Q(\theta).\end{aligned}$$

两边乘以 $\|\varphi(\theta)\|$，对两边积分并应用公式（20）和（21），

$$\begin{aligned}\int_\Theta \|\varphi(\theta)\| \left|\overline{\pi}_n^Q(\theta) - \pi_n^Q(\theta)\right| \mathrm{d}\theta \leq& \frac{1}{\int_\Theta \overline{g}_n(\theta)\pi(\theta)\mathrm{d}\theta} \int_\Theta \|\varphi(\theta)\| \left|\overline{g}_n(\theta) - g_n(\theta)\right| \pi(\theta)\mathrm{d}\theta \\ & + \frac{\left|\int_\Theta \overline{g}_n(\theta)\pi(\theta)\mathrm{d}\theta - \int_\Theta g_n(\theta)\pi(\theta)\mathrm{d}\theta\right|}{\int_\Theta \overline{g}_n(\theta)\pi(\theta)\mathrm{d}\theta} \int_\Theta \|\varphi(\theta)\|\pi_n^Q(\theta)\mathrm{d}\theta \\ \leq& O\left(1/N\right) + \frac{\int_\Theta g_n(\theta)\pi(\theta)\mathrm{d}\theta}{\int_\Theta \overline{g}_n(\theta)\pi(\theta)\mathrm{d}\theta} O\left(1/N\right) \int_\Theta \|\varphi(\theta)\|\pi_n^Q(\theta)\mathrm{d}\theta \\ =& O\left(1/N\right).\end{aligned}$$

对于 $n \geq 1$, $\int_\Theta \|\varphi(\theta)\|\pi_n^Q(\theta) < \infty$，第二个不等式中的第一项为 $O(1/N)$；第二项也为 $O(1/N)$，因为 $\int_\Theta g_n(\theta)\pi(\theta)\mathrm{d}\theta / \int_\Theta \overline{g}_n(\theta)\pi(\theta)\mathrm{d}\theta = 1 + O(1/N)$。所述结果由此得证。

□

引理 2的证明。根据三角不等式，

$$\int_\Theta \|\vartheta\| |\overline{\pi}_n^Q(\vartheta) - N\{\vartheta; 0, \Sigma_\star^{-1}\}| \mathrm{d}\vartheta \leq \int_\Theta \|\vartheta\| |\overline{\pi}_n^Q(\vartheta) - \pi_n^Q(\vartheta)| \mathrm{d}\vartheta + \int_\Theta \|\vartheta\| |\pi_n^Q(\vartheta) - N\{\vartheta; 0, \Sigma_\star^{-1}\}| \mathrm{d}\vartheta.$$

第一项为 $O(1/N)$，对于 $N \to$ ，由定理 $n \to +\infty$ 1可得。第二项为 $o_p(1)$ 由引理8可得。□

引理*3*的证明。由定理1 和变量替换 $\theta = \theta_\star + \vartheta/\sqrt{n} + Z_n/\sqrt{n}$,

$$\begin{aligned}\overline{\theta} = \int_\Theta \theta\overline{\pi}_n^Q(\theta)\mathrm{d}\theta &= \int_\Theta \theta\left\{\overline{\pi}_n^Q(\theta) - \pi_n^Q(\theta)\right\}\mathrm{d}\theta + \int_\Theta \theta\pi_n^Q(\theta)\mathrm{d}\theta \\ &= O(1/N) + \int_{\mathcal{T}_n}(\theta_\star + Z_n/n + \vartheta/\sqrt{n})\pi_n^Q(\vartheta)\mathrm{d}\vartheta\end{aligned}$$

以便

$$\sqrt{n}(\overline{\theta} - \theta_\star) - Z_n/\sqrt{n} = O(\sqrt{n}/N) + \int_{\mathcal{T}_n}\vartheta\pi_n^Q(\vartheta)\mathrm{d}\vartheta.$$

然而，由于 $\mathcal{T}_n \to \mathbb{R}^{d_\theta}$ 当 $n \to \infty$， $\int_{\mathcal{T}_n}\vartheta N\{\vartheta;\ \mathbf{0},\ \Sigma_\star^{-1}\}\mathrm{d}\vartheta = o(1)$，并根据引理 2

$$\left\|\int_{\mathcal{T}_n}\vartheta\pi_n^Q(\vartheta)\mathrm{d}\vartheta\right\| \le \int_{\mathcal{T}_n}\|\vartheta\|\|\pi_n^Q(\vartheta) - N\{\vartheta; 0, \Sigma_\star^{-1}\}\mathrm{d}\vartheta + o(1) = o_p(1).$$

因此，我们有 $\|\sqrt{n}(\overline{\theta} - \theta_\star) - Z_n/\sqrt{n}\| = o_p(1)$。根据假设 2， $Z_n/\sqrt{n} \Rightarrow N(0,\ \Sigma_\star^{-1})$，结论即得。 □

定理 *2*的证明。 证明采用与证明引理 8相同的技巧，但 $m_n(\theta) = \psi_n(\theta)$ 以及 $W_n(\theta) = V_n(\theta)$，因此省略证明。 □