arXiv:2302.03026v2 [stat.ML] 2 Jun 2023

基于采样的 一般推断中后验估计器的准确性检验 e

Pablo Lemos [1 2 3 4 *] Adam Coogan [1 2 3 *] Yashar Hezaveh [1 2 3] Laurence Perreault-Levasseur [1 2 3]

摘要

参数推断，即给定一些数据后推断统计模型参数的后验分布，是许多科学学科中的核心问题。生成模型可用于替代马尔可夫链蒙特卡罗方法，以进行后验推断，适用于基于似然和基于模拟的问题。然而，评估生成模型所编码的后验分布的准确性并非易事。本文介绍了“随机点准确性测试”（TARP）覆盖率检验，将其作为一种估计生成式后验估计器覆盖概率的方法。我们的方法不同于以往基于覆盖率的方法，后者需要对后验分布进行评估。我们证明，该方法是检验后验估计器是否准确的必要且充分条件。我们通过多种合成示例展示了这一方法，并表明 TARP 可用于检验高维空间中的后验推断分析结果。我们还表明，在现有方法失效的情况下，我们的方法能够检测出不准确的推断结果。

1. 引言

参数推断的任务，即确定未知参数的取值 θ 在与观测数据一致的统计模型中 x，是科学分析中普遍存在的任务。尽管已经有多种成熟的方法，例如马尔可夫链蒙特卡罗（MCMC）、变分推断（VI）和嵌套抽样（Skilling，2006），近年来人们开始转向应用机器学习，对不同观测（例如Zhu 与 Zabaras，2018；Papernot 与 Mc-Daniel，2018；Charnock 等，2020；Wilson 与 Izmailov，2020；Zuo 等，2020）。这种方法涉及训练一个模型（通常是神经网络），使其将真实后验分布近似为观测的函数。其目标是高效地为新数据推断后验分布，从而无需针对每个新观测执行成本高昂的 MCMC 运行。

基于模拟的推断（SBI，例如 Cranmer 等人，2020），也称为无似然推断（LFI）或隐式似然推断（ILI），近年来广受关注（例如 Ong 等人，2018；Perreault Levasseur 等人，2017；Gonc.alves 等人，2020；Dax 等人，2021；Als-ing 等人，2019；Wagner-Carena 等人，2021；Legin 等人，2021；Brehmer，2021；Coogan 等人，2020；Montel 等人，2022；Coogan 等人，2022；Brehmer 等人，2019；Chen 等人，2020；Mishra-Sharma 与 Cranmer，2022；Karchev 等人，2022b；Hermans 等人，2021a；Anau Montel 与 Weniger，2022；de Witt 等人，2020；Marlier 等人，2021；Karchev 等人，2022a；Ramesh 等人，2022）。SBI 不需要似然的显式表达式，只需借助模拟器生成训练数据即可。基于模拟的推断框架能够处理复杂的高维数据，以及难以或无法使用传统基于似然的方法进行分析的模型。

基于模拟的推断（SBI）的早期发展包括拒绝式近似贝叶斯计算（近似贝叶斯计算）（Rubin，1984；Pritchard 等人，1999；Beaumont 等人，2002；Marjoram 等人，2003；Fearnhead 和 Prangle，2012）的提出。但如今，SBI 已发展出更强大的、由神经网络驱动的摊销方法，例如神经比率估计（NRE）（Cranmer 等人，2015；Thomas 等人，2022；Hermans 等人，2020；Durkan 等人，2020；Miller 等人，2022b）；神经后验估计（NPE）（Rezende 和 Mohamed，2015；Papamakarios 和 Murray，2016；Lueckmann 等人，2018；Lueckmann 等人，2017；Greenberg 等人，2019）以及神经似然估计（NLE）（Price 等人，2018；Papamakarios 等人，2019；Frazier 等人，2022）。近年来，将 SBI 应用于高维参数空间引起了广泛关注。生成对抗网络（GANs）（Goodfellow 等人，2014）、归一化流（Dinh 等人，2014；Rezende 和 Mohamed，2015；Papamakarios 等人，2021）、变分自编码器（Kingma 和 Welling，2013）等生成模型，以及

[1]Mila – Quebec AI Institute，蒙特利尔，魁北克，加拿大[2]蒙特利尔大学，蒙特利尔，魁北克，加拿大 [3]CIELA Institute，蒙特利尔，魁北克，加拿大 [4]Flatiron Institute 计算数学中心，美国纽约州纽约市第五大道162号3楼，邮编10010。通信作者：Pablo Lemos<pablo.lemos@umontreal.ca>。

论文集： 40^{th} 国际机器学习会议，美国夏威夷州檀香山。PMLR 202，2023年。版权所有 © 2023，作者所有。

基于分数的模型/扩散模型（Song 等人，2020；Ho 等人，2020；Sohl-Dickstein 等人，2015），是在此类情形下编码近似后验分布的强大方法。

MCMC 方法的收敛检验，例如 Gelman-Rubin 统计量（Gelman &Rubin，1992）、有效样本量和积分自相关时间，都是成熟的方法。然而，这些方法评估的是样本的多样性，并不能直接保证后验分布得到了正确采样。对于基于模拟的推断，通常通过覆盖概率检验估计后验分布的准确性（另见 Guo 等人（2017）），这需要评估后验分布的密度。（Schall，2012；Prangle 等人，2013；Cranmer 等人，2020；Hermans 等人，2021b）。覆盖概率衡量某个区间包含真实参数值的比例。然而，对于无法评估其密度的生成模型（如生成对抗网络和扩散模型）生成的样本，基于已学习后验分布评估结果计算覆盖概率并不适用。此外，更重要的是，这些覆盖概率检验只是评估估计后验分布准确性的必要条件，而非充分的诊断方法。

人们还提出了其他方法，作为基于模拟的推断的替代验证方法（Lueckmann 等人，2021年；Dalmasso 等人，2020；Deistler 等人，2022年）。例如，基于模拟的校准（SBC）（Talts 等人，2018）提出了一种有趣的技术，只需使用样本，但仅适用于一维后验分布，无法用于全维空间。由Linhart 等人（2022年）提出的另一种方法，能够高效评估后验分布的准确性，但专为归一化流设计，无法应用于其他推断场景。这些方法都无法用于评估高维变量推断的准确性。

本文旨在介绍一种框架，仅使用来自数据 x 与感兴趣参数 θ, p 的真实联合分布的样本x， θ，以及来自估计后验分布 $\hat{}p(\theta|x)$ 的样本，来检验参数推断的准确性。

我们的创新贡献包括证明通过覆盖率检验验证后验估计器准确性的充要条件（定理 3），以及一种能够在实践中实现该定理的方法（§3.2）。我们首先在 §2 中介绍所有必要的符号。接着在 §3 中介绍我们的方法，在 § 4 中展示实验，并在 §5 中总结研究结果。我们的代码可在以下地址获取：
https://github.com/Ciela-Institute/tarp。

2. 形式化

本节将介绍一些基本概念，并逐步引出我们的关键理论结果（定理 3）。覆盖率下一节介绍的检验程序本质上是该定理的实际应用。

2.1. 符号

如引言所述，我们关注取值连续的参数 $\theta \in U \subset \mathbb{R}^n$ 和观测 $x \in V \subset \mathbb{R}^m$，它们取自欧几里得空间的（子集），其联合密度为 $p(\theta, x)$。我们将后验估计器记为 $\hat{}p(\theta|x)$（例如，它可以是神经网络或 MCMC 采样器），并假设我们也可以用它来生成 θ 的样本。

有了这些预备知识，我们给出两个基本定义：

定义 1。 如果后验估计器 $\hat{}p(\theta|x)$ 满足以下条件，则称其**准确**

$$\hat{p}(\theta|x) = p(\theta|x) \quad \forall (x, \theta) \sim p(x, \theta)\,. \tag{1}$$

定义 2。 对于给定的可信水平 $1-\alpha$ 和观测x，可信区域生成器 $\mathcal{G}: \hat{}p, \alpha, x \mapsto W \subset U$ 是满足以下条件的函数

$$\int_{\mathcal{G}(\hat{p},\alpha,x)} \mathrm{d}\theta\, \hat{p}(\theta|x) = 1 - \alpha\,. \tag{2}$$

注意，这类生成器有无穷多个。一种常用的生成器是最高后验密度区域生成器，其定义见 §3.1。

接下来，我们介绍本文的两个核心定义，它们改编自 Hermans 等人(2021年b)（下文简称 H21）。

定义 3。 生成器 $\mathcal{G}$ 在可信水平 $1-\alpha$ 和数据 x 下的覆盖概率 为

$$\mathrm{CP}(\hat{p}, \alpha, x, \mathcal{G}) = \mathbb{E}_{p(\theta|x)}\left[\mathbb{1}\left(\theta \in \mathcal{G}(\hat{p}, \alpha, x)\right)\right]\,. \tag{3}$$

定义 4。 生成器 $\mathcal{G}$ 在可信水平 α 下的期望覆盖概率是对数据分布取平均后的覆盖概率：

$$\mathrm{ECP}(\hat{p}, \alpha, \mathcal{G}) = \mathbb{E}_{p(x)}\left[\mathrm{CP}(\hat{p}, \alpha, x, \mathcal{G})\right]\,. \tag{4}$$

2.2. 覆盖概率

现在，我们来说明具有正确覆盖概率的估计器的一些基本事实。首先给出一个直接的结果：

定理 1。 对于 x 的所有取值以及任意可信区域生成器 $\mathcal{G}(p, \alpha, x)$，后验分布的覆盖概率为$\mathrm{CP}(p, \alpha, x, \mathcal{G}) = 1 - \alpha$。

证明 代入 $\hat{}p(\theta|x) = p(\theta|x)$ 后，覆盖概率的定义变为：

$$\begin{aligned}\mathrm{CP}(p, \alpha, x, \mathcal{G}) &= \mathbb{E}_{p(\theta|x)}\left[\mathbb{1}(\theta \in \mathcal{G}(p, \alpha, x)\right] \\ &= \int_{\mathcal{G}(p,\alpha,x)} \mathrm{d}\theta\, p(\theta|x) \\ &= 1 - \alpha\,,\end{aligned} \tag{5}$$

最后一行由可信区域的定义得出。 ■

由此可直接得出，后验分布也具有 ECP(p, α, x, $\mathcal{G}$) $= 1-\alpha$。

接下来，我们证明该定理中更有意思的逆向结论，这需要引入另一种可信区域生成器。

定义 5. 可定位可信区域生成器$\mathcal{P}_{\theta_r}(\hat{p}, \alpha, x)$ 生成以 θ_r 为中心的可信区域，其中 θ_r 是参数空间中任意选定的点，其含义是

$$\lim_{\alpha \to 1} \mathcal{P}_{\theta_r}(\hat{p}, \alpha, x) = \{\theta_r\} \tag{6}$$

对所有 x 和 θ_r 均成立。区域的形状并不重要：例如，它们可以是球或超立方体。

最后，我们将函数 $f(\cdot)$ 在以 θ_r 为中心的可信区域 Θ 上的平均值记作

$$\overline{f(\theta_r)}(\Theta) := \frac{1}{\mathrm{vol}[\Theta]} \int_\Theta \mathrm{d}\theta \, f(\theta) \,. \tag{7}$$

当 $f(\cdot)$ 是概率密度函数时，$f(\cdot)(\Theta)$ 也同样是概率密度函数，因为它是 $f(\cdot)$ 与密度 $\mathbf{1}(\theta \in \Theta)$ 的卷积。

定理 2。 假设对于可定位可信区域生成器 $\mathcal{P}_{\theta_r}$，后验估计器的覆盖概率对所有 θ_r、x 和 α 均等于 $1-\alpha$。进一步假设，$\hat{}p(\cdot|x)$ 和 $p(\cdot|x)$ 在其定义域上均连续。则 $\hat{}p(\cdot|x) = p(\cdot|x)$。

证明 为简便起见，定义 $\Theta := \mathcal{P}_{\theta_r}(\hat{}p, \alpha, x)$。

覆盖概率定义中的积分可以写成

$$\begin{aligned} \mathrm{CP}(\hat{p}, \alpha, x, \mathcal{P}_{\theta_r}) &= 1-\alpha \\ &= \int_\Theta \mathrm{d}\theta \, p(\theta|x) \\ &= \mathrm{vol}[\Theta] \, \overline{p(\cdot|x)}(\Theta) \,, \end{aligned} \tag{8}$$

其中第一个等式由假设得出。由于我们假设 $\hat{}p(\cdot|x)$ 的支撑集包含 $p(\cdot|x)$ 的全部支撑集，因此可信区域的体积为正。根据可信区域的定义，我们还有

$$1-\alpha = \int_\Theta \mathrm{d}\theta \, \hat{p}(\theta|x) = \mathrm{vol}[\Theta] \, \overline{\hat{p}(\cdot|x)}(\Theta) \,. \tag{9}$$

令其等于前一个表达式，得到 $\hat{}p(\cdot|x)(\Theta) = p(\cdot|x)(\Theta)$；根据假设，该式对所有 θ_r 和 x 都成立。取 $\alpha \to 1$（即令 Θ 足够小），即可得到所需结果。 ■

2.3. 期望覆盖概率

前一个结果仍然不太实用，因为计算后验估计器的覆盖概率在计算上代价很高。实际上，这需要对 $p(\theta, x)$ 在 x 中的样本绘制直方图，而 x 可能具有很高的维度。然而，正如 H21 中所指出的，计算期望覆盖概率要简单得多。

下一个定理是我们的主要理论结果：只要对于定义可信区域位置的任意函数 $\theta_r(x)$ 都成立，正确的期望覆盖率就足以验证后验估计器是准确的。

定理 3。 假设对于所有 α、x 和 $\theta_r(\cdot)$，可定位可信区域生成器 $\mathcal{P}_{\theta_r}$ 的 $1-\alpha$ 的期望覆盖概率等于 $1-\alpha$，其中该生成器根据 x 为可信区域分配位置。进一步假设$\hat{}p(\cdot|x)$ 的支撑集包含 $p(\cdot|x)$ 的支撑集，并且两个函数在其定义域上都是连续的。那么$\hat{}p(\cdot|x) = p(\cdot|x)$。

证明 同样地，为清晰起见，令 $\Theta := \mathcal{P}_{\theta_r}(\hat{}p, \alpha, x)$。

首先，我们利用可信区域的定义，求出 Θ 的体积表达式：

$$1-\alpha = \int_\Theta \mathrm{d}\theta \, \hat{p}(\theta|x) = \mathrm{vol}[\Theta] \, \overline{p(\cdot|x)}(\Theta) \,, \tag{10}$$

这意味着

$$\mathrm{vol}[\Theta] = \frac{1-\alpha}{\overline{p(\cdot|x)}(\Theta)} \,. \tag{11}$$

这使我们能够展开并简化期望覆盖率的表达式：

$$\begin{aligned} \mathrm{ECP}(\hat{p}, \alpha, \mathcal{P}_{\theta_r}) &= 1-\alpha \\ &= \int \mathrm{d}x \, p(x) \int_\Theta \mathrm{d}\theta \, p(\theta|x) \\ &= \int \mathrm{d}x \, p(x) \; \mathrm{vol}[\Theta] \, \overline{p(\cdot|x)}(\Theta) \\ &= (1-\alpha) \int \mathrm{d}x \, p(x) \, \frac{\overline{p(\cdot|x)}(\Theta)}{\overline{\hat{p}(\cdot|x)}(\Theta)} \,. \end{aligned} \tag{12}$$

约去因子 $1-\alpha$ 后可知，最后一行中的积分等于 $\mathbf{1}$。

根据假设，这对任意位置函数$\theta_r(x)$的选择都成立。因此，我们可以对积分关于 $\theta_r(x)$取泛函导数。回想一下，这些平均值

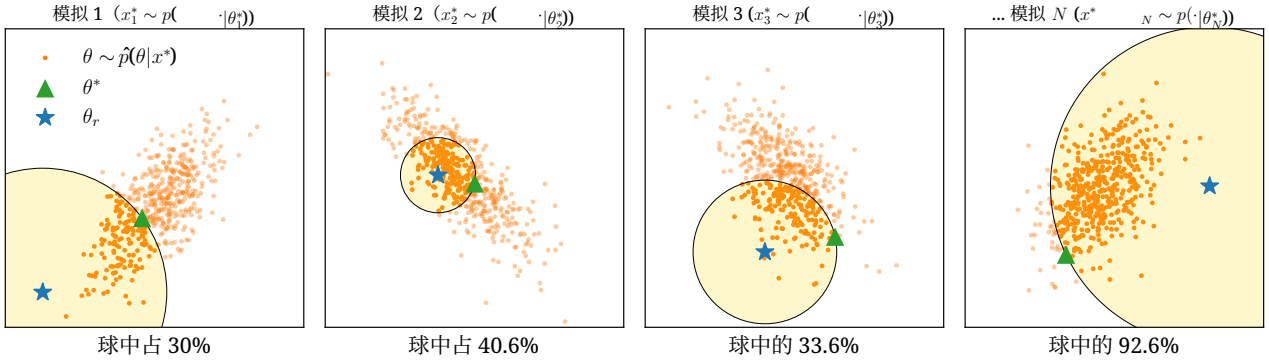


图 *1*。用于评估后验估计器 $\hat{}p$ 质量的所提覆盖率检验示意图。给定一组模拟（各面板），我们从后验估计器中抽取样本（橙色点）。我们抽取一个参考参数点 θ_r，并确定落在以 θ_r 为中心、延伸至用于生成模拟的真实参数点 θ^* 的球内的点 f 的比例（球以黄色标示，f 标注于各面板下方）。我们的覆盖率检验汇总 f 的统计量，从而提供一种保证 $\hat{}p$ 准确性的必要且充分的方法。

取决于被积函数中的 θ_r，我们得到

$$0 = \frac{\delta}{\delta\theta_r(x)} \int \mathrm{d}x\, p(x)\, \frac{\overline{p(\cdot|x)}(\Theta)}{\overline{\hat{p}(\cdot|x)}(\Theta)} \tag{13}$$

$$= \int \mathrm{d}x\, \delta\theta_{r,i}(x)\, p(x) \frac{\partial}{\partial\theta_{r,i}} \left(\frac{\overline{p(\cdot|x)}(\Theta)}{\overline{\hat{p}(\cdot|x)}(\Theta)} \right) \tag{14}$$

$$= \int \mathrm{d}x\, \delta\theta_{r,i}(x) \frac{\overline{p(\cdot|x)}(\Theta)\, p(x)}{\overline{\hat{p}(\cdot|x)}(\Theta)} \times \left[\frac{\partial \log \overline{p(\cdot|x)}(\Theta)}{\partial\theta_{r,i}} - \frac{\partial \log \overline{\hat{p}(\cdot|x)}(\Theta)}{\partial\theta_{r,i}} \right], \tag{15}$$

其中，下标 i 表示 θ_r 的各个分量。由于该表达式必须对所有变分 $\delta\theta_{r,\,i}$ 都成立，被积函数必须等于零（即必须满足欧拉-拉格朗日方程）。根据假设，被积函数中花括号外的因子非零，因此

$$\frac{\partial \log \overline{p(\cdot|x)}(\Theta)}{\partial\theta_{r,i}} = \frac{\partial \log \overline{\hat{p}(\cdot|x)}(\Theta)}{\partial\theta_{r,i}}\,. \tag{16}$$

这意味着，对于某个依赖于 x 的积分常数 c，有 $\log p(\cdot|x)(\Theta) = \log \hat{}p(\cdot|x)(\Theta) + c(x)$。但是，由于对数中的函数本身就是密度，我们有$c(\cdot) = 0$。取极限 $\alpha \to 1$，得到 $\hat{}p(\theta|x) = p(\theta|x)$。 ■

我们将在下一节介绍的覆盖率检验方法，实际上是对这一结果的实用实现。

3. 我们的方法

在证明了我们的主要理论结果（参见定理 3）之后，本节首先利用该结果解释常见的覆盖概率计算中存在的盲点，然后介绍我们新的覆盖率检验程序。

3.1. 高后验密度覆盖率检验

在介绍所提出的方法之前，我们先讨论 HPD 覆盖率。

定义 6. 我们将 *HPD* 可信区域生成器$\mathcal{H}(\hat{}p, \alpha, x)$ 定义为这样一种生成器：它生成质量为 $1-\alpha$ 的区域，且该区域在 U[1] 中所占体积尽可能小。

请注意，这并不是一个可定位的可信区域生成器。它可以与定义4结合使用，以计算最高后验密度期望覆盖概率（HPD期望覆盖概率）。HPD期望覆盖概率常用于评估覆盖率（Hermans 等人，2021b；Rozet 等人，2021年；Miller 等人，2022a；Deistler 等人，2022年；Tejero-Cantero 等人，2020）。

使用最高后验密度区域计算期望覆盖概率，最直观的方法或许是针对 α、[2] 的所有可能取值计算这样的区域，然后利用（3）计算期望覆盖率。不过在实践中，有一种更高效的期望覆盖概率计算方法，该方法源自以下结果：

注释1。一对（θ^*，x^*）和一个后验估计器$\hat{}p$（$\theta|x$）唯一确定一个最高后验密度置信区域，如下：

$$\{\theta \in U \mid \hat{p}(\theta|x^*) \geq \hat{p}(\theta^*|x^*)\}\,. \tag{17}$$

依此可定义相应的**HPD置信水平** $1 - \tilde{\alpha}_{\mathrm{HPD}}(\hat{}p, \theta^*, x^*)$，即对该区域上的$\hat{}p(\theta|x)$ 进行积分所得。

[1]注意，这对于均匀密度函数来说定义不明确。[2]请注意，先前的研究，例如 Perreault Levasseur 等人(2017)，曾尝试基于少数几个 α值进行准确性检验。与通常用于覆盖率检验的遍历 α 所有可能取值相比，这项检验的限制性要小得多。

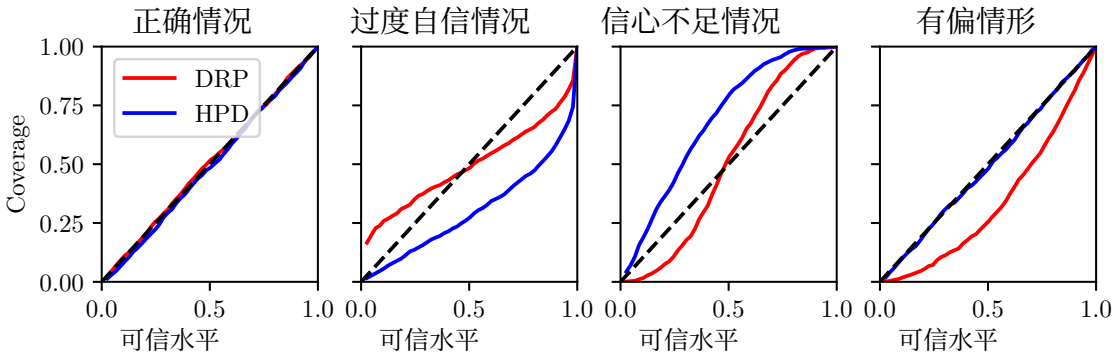


图2。第4.1节所述四种情形下高斯玩具模型的结果。红线表示本文提出的方法，蓝色曲线表示最高后验密度区域。

算法1 根据一组模拟 $\{\theta_i, x_i\}$,$i \in [1, N_{\text{sims}}]$，使用最高后验密度区域计算 ECP($\hat{p}, \alpha, \mathcal{H}$)

为每次模拟生成 n 个样本 $\{\theta_{ij}\} \sim \hat{p}(\theta|x_i)$ x_i。**对于** $i \leftarrow 1$ 至 N_{sims} **执行**$f_i = (1/\text{n}) \cdot \sum_{j=1}^{n} \mathbb{1}\left[\hat{p}(\theta_{ij}|x_i) < \hat{p}(\theta_i^*|x_i)\right]$**结束循环**期望覆盖概率($\hat{p}$, α, $\mathcal{H}$) $= (1/N_{\text{sims}}) \sum_{i=1}^{N_{\text{sims}}} \mathbb{1}(f_i < 1-\alpha)$

随后，我们可以重新推导出关于这一最高后验密度可信水平的一项重要结果：

引理 1. 我们可以计算 $1-\alpha$ 最高后验密度区域的期望覆盖概率：

$$\text{ECP}(\hat{p}, \alpha, \mathcal{H}) = \mathbb{E}_{p(\theta,x)}\left[\mathbb{1}\left(\tilde{\alpha}_{\text{HPD}}(\hat{p}, \theta, x) \geq \alpha\right)\right]. \quad (18)$$

证明 首先，我们注意到：

$$\theta^* \in \mathcal{H}\left(\hat{p}, \alpha, x^*\right) \Leftrightarrow \tilde{\alpha}_{\text{HPD}}(\hat{p}, \theta^*, x^*) \geq \alpha. \quad (19)$$

这是因为，如果 $\theta^* \in \mathcal{H}\left(\hat{p}, \alpha, x^*\right)$，则由 (θ^*, x^*) 定义的 HPD 置信区域包含于 $\mathcal{H}\left(\hat{p}, \alpha, x^*\right)$。

然后，根据式 (4)，可知式 (18) 成立。 ■

这一结果在实践中可用于根据真实联合分布 $p(\theta, x)$ 的样本计算 HPD 期望覆盖概率，如算法1所示。如前所述，该算法需要对后验估计器进行显式求值。我们尝试在 §A 中进一步直观地说明这两个定义之间的联系。

正如文献中广为人知的，仅用最高后验密度区域估计期望覆盖概率，并不足以证明后验估计器是准确的。定理 3 揭示了原因：根据定义，最高后验密度区域生成器不具备可定位性。可定位性对定理的证明至关重要，因为证明要求改变位置函数 $\theta_r(x)$。

为了具体说明仅考虑最高后验密度覆盖率为何可能失效，我们考虑 H21 中讨论的有趣情形：$\hat{p}(\theta|x) = p(\theta)$。根据期望覆盖概率的定义，

$$\begin{aligned} \text{ECP}(\hat{p}, \alpha, \mathcal{H}) &= \mathbb{E}_{p(x,\theta)}[\mathbb{1}(\theta \in \mathcal{H}(\hat{p}, \alpha)] \\ &= \mathbb{E}_{p(\theta)}[\mathbb{1}(\theta \in \mathcal{H}(\hat{p}, \alpha)] \\ &= \int_{\mathcal{H}(\hat{p},\alpha)} \mathrm{d}\theta\, p(\theta) \\ &= 1-\alpha\,. \end{aligned} \quad (20)$$

在第二行中，我们利用了一个事实：在此情形下，HPD 生成器与 x 无关，即 $\mathcal{H}(\hat{p}, \alpha, x) = \mathcal{H}(\hat{p}, \alpha)$。我们可以将第三行识别为先验的可信区域定义，由此得到第四行。这意味着在此情形下，$\hat{p}(\theta|x)$ 具有完美的 HPD 期望覆盖概率。

现在，我们介绍一种能够弥补这些盲点的覆盖率检验方法。

3.2. 随机点距离覆盖率检验

这里提出的方法围绕位置 θ_r生成球形可信区域：

定义 7。 给定距离度量 $d: \; U \times U \to R$，我们将 *TARP* 区域生成器 $\mathcal{D}_{\theta_r}(\hat{p}, \alpha, x, d)$ 定义为一种可定位生成器，它能够生成可信水平为 $1-\alpha$ 的可信区域：

$$\mathcal{D}_{\theta_r}(\hat{p}, \alpha, x, d) := \{\theta \in U \mid d(\theta, \theta_r) \leq R(\hat{p}, \alpha, x)\}\,, \quad (21)$$

其中，$R(\hat{p}, \alpha, x)$ 满足(2) 。

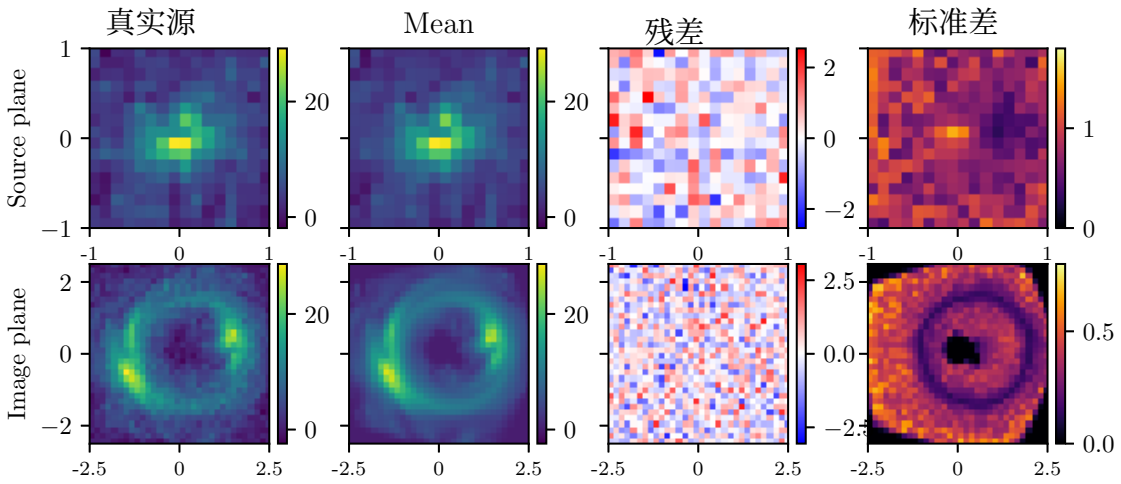


图 3. 所执行的一次引力透镜模拟示例。顶部面板显示我们试图推断的（潜在）源平面，底部面板则显示畸变图像。从左到右，图中依次显示后验估计器样本的真实值、均值和标准差（对于本图，使用的是“精确”估计器），以及残差。此处所示颜色刻度中的观测噪声设为 1。

算法2 使用 TARP 方法计算 ECP($\hat{p}, \alpha, \mathcal{D}_{\theta_r}$)，使用一组模拟 $\{\theta_i, x_i\}, i \in \{1, \ldots, N_{\text{sims}}\}$、参数距离度量 d：$U \times U \to \mathbb{R}^{\geq 0}$ 以及参考点采样分布 $\tilde{p}(\cdot|x)$。

为每次模拟 x_i 生成 n 个样本 $\{\theta_{ij}\} \sim \hat{p}(\theta|x_i)$，其中 $j = \{1, \ldots, n\}$。**对** $i \leftarrow 1$ 至 N_{sims} **执行**$\theta_r \sim \tilde{p}(\theta_r|x)$ {生成参考点}$f_i = (1/\mathrm{n}) \cdot \sum_{j=1}^{n} \mathbb{1}\left[d(\theta_{ij}, \theta_r) < d(\theta_i^*, \theta_r)\right]$**循环结束**期望覆盖概率($\hat{p}$, α, $\mathcal{D}_{\theta_r}$) $= (1/N_{\text{sims}}) \sum_{i=1}^{N_{\text{sims}}} \mathbb{1}($ $f_i < 1 - \alpha$)

由此结果可得出一个关键结论，与上一节类似：

注 2. 一对 (θ^*, x^*) 和一个后验估计器 $\hat{p}(\theta|x)$唯一确定给定 d 和θ_r时的***TARP***[3] 可信区域：

$$\{\theta \in U \mid d(\theta, \theta_r) \leq d(\theta^*, \theta_r)\} \tag{22}$$

由此又可定义相应的**TARP 置信水平** $1 - \tilde{\alpha}_{\text{TARP}}(\hat{p}, \theta^*, \theta_r, d)$，即$\hat{p}(\theta|x)$ 在该区域上的积分。

我们可以像 HPD 情形那样计算期望覆盖率：

引理 2. 我们可以如下计算 $1 - \alpha$ *TARP* 区域的期望覆盖概率：

$$\mathrm{ECP}(\hat{p}, \alpha, \mathcal{D}_{\theta_r}) = \mathbb{E}_{p(\theta,x)}\left[\mathbb{1}(\tilde{\alpha}_{\text{TARP}}(\hat{p}, \theta^*, \theta_r, x^*, d) \geq \alpha)\right]. \tag{23}$$

证明 设 $\mathcal{D}_{\theta_r}(x^*, \alpha, \hat{p}, d)$ 是一个以 θ_r 为中心、半径为 $R(\hat{p}, \alpha, x)$、可信度为 $1 - \alpha$ 的球。同样，由 (θ^*, x^*) 定义的 TARP区域具有相同的中心、半径$d(\theta^*, \theta_r)$ 和可信度 $1 - \tilde{\alpha}$，其中 $\tilde{\alpha}$。由此可得：

$$\theta^* \in \mathcal{D}_{\theta_r}(x^*, \alpha, \hat{p}, d) \Leftrightarrow d(\theta^*, \theta_r) \leq R(\hat{p}, \alpha, x). \tag{24}$$

由于 R 是 α 的单调函数，且这些区域以同一点为中心，因此我们有

$$d(\theta^*, \theta_r) \leq R(\hat{p}, \alpha, x) \Leftrightarrow \tilde{\alpha} \geq \alpha\,. \tag{25}$$

由此，根据 (4) 可得 (23)。 ■

至此，我们已具备制定算法所需的一切条件，该算法见算法2。虽然它与算法1相似，但有三点关键区别：

- TARP 通过从某个分布 $\tilde{p}(\theta|x)$ 中随机抽取 θ_r，来实现定理 3 的要求：对于可信区域位置的所有可能选取方式，覆盖率都成立；该分布可以依赖于 x。

[3] TARP 是“Test of Accuracy with RandomPoints”（随机点准确性检验）的缩写。本文的早期版本使用了 DRP（“Distance to Random Point”，到随机点的距离）这一名称。

- 随着所测试的后验样本、模拟次数和参考点数量增加，TARP 会探测尺寸更小的可信区域（即更大的 α）。按照定理 3 证明中的逻辑，这意味着它渐近地检验 $\hat{}p(\theta|x)$ 和 $p(\theta|x)$ 的均值在越来越小的球上是否匹配。
- TARP 不需要显式计算后验估计器 $\hat{}p$：只需计算从 $\hat{}p$ 和 θ_r 中采样得到的参数之间的距离。

在下一节中，我们通过一系列实验检验所提出的方法，并将其性能与最高后验密度覆盖概率进行比较。

4. 实验

我们将算法应用于三个不同的实验中，算法如算法2所述。对于所有实验，我们将所有参数 θ 归一化至范围 $[0,1]$，；除非另有说明，否则我们在 D 维超立方体 $x \in [0,1]^D$ 中均匀生成参考点，其中 D 是参数空间的维数。我们使用欧几里得距离或 L2 距离作为度量来计算 TARP 区域。我们将在 §4.2 中探讨参考点分布和距离度量的影响。

4.1. 高斯玩具模型

作为第一个示例，我们可以使用一个简单的高斯玩具模型。在这个模型中，我们假设所有后验分布都是高斯分布。因此，我们可以根据估计的均值和协方差矩阵，为验证模拟生成后验分布的样本。首先，我们通过在参数空间中进行均匀采样来生成“模拟”，$\theta^* \sim \mathcal{U}(-5,\ 5)$。我们还通过从 $\log\ \sigma \sim \mathcal{U}(-5,\ -1)$ 中采样，随机生成后验估计的协方差矩阵 Σ 的对角元素，并将非对角元素设为 0。为了进行验证，我们还需要知道后验分布的均值。我们考虑三种情况：

- 首先，我们从正态分布$\mathcal{N}(\theta^*,\ \Sigma)$ 中抽取这些值。这意味着覆盖率概率应呈均匀分布。我们称之为正确情况。
- 其次，我们从 $\mathcal{N}(\theta^*,\ 0.5\Sigma)$ 和 $(\mathcal{N}(\theta^*,\ 2\Sigma))$ 中抽取真实值。这意味着后验样本来自一个过窄（过宽）的分布，因此会过度自信（信心不足）
- 最后，我们要构造一个有偏情形。为此，我们将均值设为：

$$\theta^* - \text{sign}(\theta^*) \cdot Z\left(1 - \frac{|\theta^*|}{5}\right) \cdot \sigma, \qquad (26)$$

其中，Z 是逆生存函数。这个例子的构想是引入一种位置相关偏差：真实值离原点越远，后验分布的偏差就越大。我们特意设计了这种偏差，使最高后验密度覆盖概率无法察觉到它。不过，这个例子的目的是说明，有些偏差会被最高后验密度（HPD）忽略，而 TARP 的随机性应该能够检测到它们。函数 (26) 绘于附录 §C

对于上述每种情况，我们都希望比较此方法与HPD覆盖概率检验的表现。由于在这个玩具模型中，我们知道正确的后验分布，因此可以轻松计算HPD和TARP覆盖概率。选择TARP参考点时，我们使用先验（$\tilde{p}(\theta_r|x) = p(\theta_r)$）。

高斯玩具模型的结果如图2所示。每个面板中，x轴表示可信水平 $1-\alpha$，而 y轴表示期望覆盖率 ECP($\hat{}p, \alpha, \mathcal{G}$)。对于准确的后验估计器，ECP($\hat{}p, \alpha\mathcal{G}$) $= 1-\alpha$, $\forall\alpha \in$ (0,1)，如第2节所述，因此会得到黑色虚线对角线。我们在第一个面板中看到，“正确”情况确实如此，因为该情况在构造时就是准确的。对于我们测试过的所有 D 值，结果都一致，最大达到 $D = 1000$。

第二个和第三个面板分别展示了信心不足和过度自信的情形。我们可以看到，与 HPD 方法相比，这些情形会导致不同的覆盖率图。这并不完全出乎意料：对于信心不足的估计器，随机选取点得到的 TARP区域更有可能覆盖后验估计器的大约一半$\alpha \sim 0.5$；而对于过度自信的估计器，它们可能覆盖的区域非常小 $\alpha \sim 1$，也可能非常大 $\alpha \sim 0$。我们在§B中进一步阐述了这一直觉，并给出了一些图示。最后，在第四个面板中，我们看到，有偏情形无法通过 HPD区域检测出来，但可以通过 TARP 检测出来。这说明，正如§2中所解释的，ECP($\hat{}p,\ \alpha$) $= 1-\alpha$并不意味着对于 HPD区域而言后验分布是准确的，但对于 TARP区域而言则意味着后验分布是准确的。

我们还针对协方差矩阵包含非对角元素的高斯分布情形重复了这一示例。为此，我们随机生成了大小为$D(D-1)/2$的数组，然后将其转换为下三角矩阵，并将其用作协方差矩阵的乔列斯基分解。我们发现，向协方差矩阵中添加非零元素并不会改变我们的结果。

4.2. 对 θ_r 分布和距离度量的依赖

图2所示的高斯玩具模型的所有结果都依赖于 §3.2 中指定的两个选择：用于抽取参考点的分布 θ_r，以及距离度量 $d(\cdot,\cdot)$。因此，研究我们的方法对这两项选择的依赖性至关重要

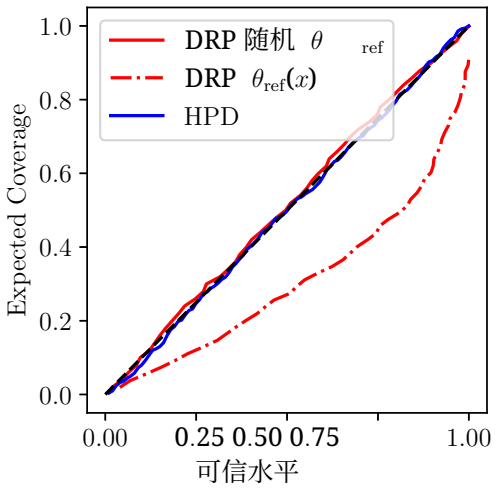


图4. §4.3中描述的无信息后验估计器的期望覆盖率与可信水平的关系。蓝色曲线表示使用HPD区域计算的覆盖率，红线表示使用TARP区域计算的覆盖率。实线使用独立于 x的参考点，而点划线使用依赖于 x的参考点。

是否受这两项选择的不同方案影响。首先，我们通过从各种分布中抽取 θ_r，重复进行了高斯玩具模型实验的全部四个版本：

- 均匀分布，既覆盖较宽的范围$\theta_r \sim \mathcal{U}(0, 1)$，也只覆盖该范围的一部分$\theta_r \sim \mathcal{U}(0, 0。5)$
- 正态分布，中心位于 $\theta = 0.5$，标准差在 0。01 到 0。1 之间变化
- θ_r 固定在某个值上，要么位于参数空间的中心 $\theta_r = 0.5$，要么位于其他位置。

我们还使用曼哈顿距离或 L1 距离代替 L2 距离重复了实验。对于正确、过度自信和信心不足情形，我们发现曲线与图2所示的曲线非常相似。在有偏情形中，不同的 θ_r分布会产生不同的曲线，但它们都清楚地表明存在偏差。§D 中展示了这些图。因此，我们得出结论：对于 θ_r 的不同分布和距离度量的不同选择，所提出的方法都具有稳健性。

4.3. 揭示估计器何时不提供有用信息

作为第二个基准，我们考虑前面提到的情形：学习得到的后验估计器等于先验 $\hat{}p(\theta|x) = p(\theta)$。我们对此示例感兴趣的原因在于，在这种情况下，使用 HPD区域计算的期望覆盖概率对于任意 α 的值都等于$1-\alpha$，如前所述。不过，使用 TARP，我们可以通过根据 x以依赖于它的方式抽取参考点，避免这一盲点。

为使讨论具体化，我们考虑一个一维示例，其高斯先验为 $p(\theta) = \mathcal{N}(\theta; \mu_0, \sigma_0^2)$。在此例中，我们的“前向模型”只是从 $\{x_i\}_{i=1}^{n_x} \sim N(\theta, {\sigma_x}^2)$ 中生成 n_x 个数据点。在这个共轭模型中，我们可以轻松推导出真实后验分布：

$$p\left(\theta|\left\{x_i\right\}_{i=1}^{n_x}\right) = \mathcal{N}(\mu|m, s), \tag{27}$$

$$s = \left(\frac{1}{\sigma_0^2} + \frac{n+x}{\sigma_x + x^2}\right)^{-1}, \tag{28}$$

$$m = s\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_i x_i}{\sigma_x^2}\right) \tag{29}$$

我们固定 $n_x = 50$、$\mu_0 = 0$、$\sigma_0 = 1$ 和 $\sigma_x = 0.1$。我们从前向模型生成 500 个样本，并用“无信息估计器” $\hat{}p(\theta|x) = p(\theta)$以三种方式计算期望覆盖率：1）使用最高后验密度区域；2）使用 TARP 区域，其中 θ_r 从 $\mathcal{U}(0, 1)$ 中随机抽取；3）使用 TARP 区域，其中 $\theta_r = x_0 + u$，且 x_0 是第一个观测，$u \sim \mathcal{U}(0, 1)$。我们预计前两种方法的期望覆盖概率等于 $1-\alpha$，但第三种方法并非如此。

我们在图 4 中展示了结果。首先，我们注意到，使用最高后验密度区域时，即使估计器是错误的，我们仍能得到正确的期望覆盖率（这验证了 § 2 中的理论讨论）。这意味着，在这种情况下，HPD 覆盖率可能会误导我们，让我们以为估计器准确，而实际上它完全没有信息量。有趣的是，当我们使用从先验中随机选取参考点的 TARP 区域时（红线），也会出现同样的情况。这是因为，正如 §2 中所讨论的，只有当区域的选择取决于 x 时，定理 3 才在两个方向上都成立。最后，正如预期的那样，当 θ_r 的抽样分布存在某种 x 依赖时，期望覆盖率并非 $1-\alpha$。因此，我们看到，即使我们对 x 引入少量依赖，TARP 中的 $\tilde{p}(\theta_r|x)$ 也会揭示后验估计器并不准确。我们还在 §D 中进一步探讨 $\tilde{p}(\theta_r|x)$ 对 x 的依赖如何影响结果。

4.4. 引力透镜

为了在更现实的高维情形下测试我们的算法，我们考虑一个简化的天体物理学问题：引力透镜源重建。当遥远星系发出的光线由于另一个中间星系（即“透镜”）的质量作用而沿弯曲路径而非直线传播时，自然界中就会发生引力透镜现象（Treu，2010年）。其结果是背景星系的图像发生严重扭曲，呈环状

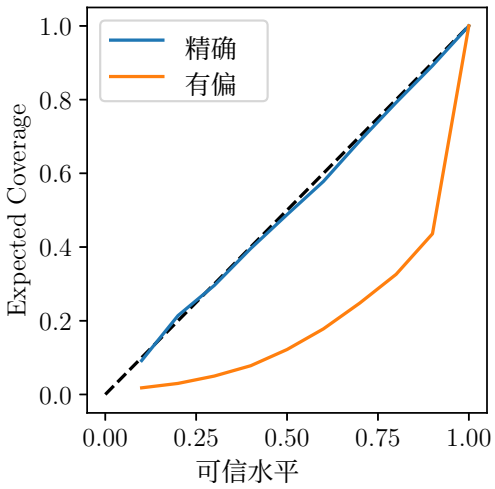


图*5*。对于我们的引力透镜示例，基于最高后验密度覆盖率的检验难以实现。图中显示了期望覆盖概率与可信水平的关系。正如预期，精确的后验估计器（蓝色曲线）能够准确表征后验分布，而有偏估计器（橙色）则不能。

图像。源重建的目标是：在假设畸变的数学形式已完全知晓的情况下，从含噪图像中推断源星系发出的光在没有畸变时的样子。在这种高维情形下，基于后验分布的最高后验密度区域进行覆盖率检验是不可行的。

在此情景中，模拟器从 θ 多元正态分布中抽取源星系的光；该分布是根据星系图像数据集拟合得到的（Stone & Courteau，2019；Stone 等人，2021年）。随后应用一个 A 矩阵来编码引力透镜畸变，并通过添加标准差为 σ_n的高斯像素噪声生成最终观测，因此$x \sim \mathcal{N}(A\theta,\ \sigma_n^2)$。为方便计算，我们使用$16 \times 16$像素的源图像和 32×32像素的观测。

如Adam 等人（2022年）所示，并在 §E 中回顾，θ 的后验样本可以通过扩散建模技术生成。一般而言，这种方法生成的后验样本会带有细微偏差。然而，采用我们对 θ的多元正态先验后，便可以生成无偏的后验样本。在结果中，我们将由此得到的样本称为“有偏”和“精确”样本。

图5展示了精确后验估计器和有偏后验估计器的结果，使用了500次模拟，每次模拟包含1000个后验样本，并从先验中采样 θ_r。正如预期，我们的方法对精确估计器得到了正确的覆盖率。需要特别强调的是，对于这种维度的参数空间（256个参数），必须使用生成模型；此前已有的方法都无法计算经验覆盖概率，以评估此类模型的准确性。另一方面，有偏估计器生成的曲线与图2右下角面板中的曲线相似，这表明它确实存在偏差。

5. 结论

检验估计后验分布的准确性是参数推断的关键环节。对于 MCMC 等成熟采样方法，已有完善的收敛诊断方法；但要直接检验后验推断结果的准确性却很困难，尤其是使用深度学习方法计算的结果。基于似然的推断和基于模拟的推断都存在这一问题。本文提出了 TARP 覆盖概率这一新技术，在无法获得显式后验评估结果时，仅使用后验样本即可检验估计后验分布的准确性。尽管我们重点关注基于生成式机器学习模型的后验估计器，但该方法同样可用于检验 MCMC 样本的正确性，不过计算成本可能很高。

我们证明了，这项检验足以证明推断结果准确，而其他类似检验虽为必要条件，却并不足够。我们还测试了所选 $\hat{p}(\theta_r|x)$ 以及 TARP 方法所用距离度量的影响，发现它们不会显著影响我们的结果。唯一例外是后验估计器等于先验的情况，此时只有当 $\hat{p}(\theta_r|x)$ 依赖于 x 时，TARP 才有效。是否存在这种风险，需由该方法的使用者自行判断。

我们成功地将这项检验应用于各种推断问题，尤其是在其他方法失效的情况下，并表明它能够很好地扩展到高维后验分布。因此，我们提出将 TARP 覆盖概率作为一种工具，用于检验未来基于生成模型的后验推断分析是否准确。

6. 更广泛影响

我们的工作聚焦于检验统计推断的正确性，这是一个尚待解决的重要问题。我们预计，这项工作将通过提升机器学习应用于各领域科学问题时的可信度，产生积极的社会影响。然而，与任何统计方法一样，若不当应用我们的方法（尤其是未能妥善选择 θ_{ref} 的抽样分布），可能会得出无效结论。

参考文献

Adam, A., Coogan, A., Malkin, N., Legin, R., Perreault-Levasseur, L., Hezaveh, Y. 和 Bengio, Y. 基于强引力透镜中源星系的后验样本

采用基于评分的先验。*arXiv* 预印本 *arXiv:2211.03812*，2022年。

Alsing, J., Charnock, T., Feeney, S. 和 Wandelt, B. 利用神经密度估计器和主动学习实现快速无似然宇宙学。《皇家天文学会月刊》，488(3):4440–4458，2019。

Anau Montel, N. 和 Weniger, C. 检测即截断：使用截断边际神经比率估计研究源群体。载于第*36*届神经信息处理系统会议，2022年11月。

Beaumont, M. A.、Zhang, W. 和 Balding, D. J. 群体遗传学中的近似贝叶斯计算。《遗传学》，162(4):2025–2035，2002。

Brehmer, J. 粒子物理学中的基于模拟的推断。《自然评论：物理学》，3(5):305–305，2021年1月。doi: 10.1038/s42254-021-00305-6。

Brehmer, J., Mishra-Sharma, S., Hermans, J., Louppe, G. 和 Cranmer, K. 挖掘暗物质子结构：利用机器学习从强引力透镜推断子晕群体性质。天体物理学杂志，886(1):49，2019。doi: 10.3847/1538-4357/ab4c41。

Charnock, T., Perreault-Levasseur, L. 和 Lanusse, F. 贝叶斯神经网络。*arXiv* 电子印本，文章编号 arXiv:2006.01490，2020年6月。doi: 10.48550/arXiv.2006. 01490。

Chen, Y., Zhang, D., Gutmann, M., Courville, A. 和 Zhu, Z. 隐式模型的神经近似充分统计量。*arXiv* 预印本 *arXiv:2010.10079*，2020年。

Coogan, A., Karchev, K., and Weniger, C. 强引力透镜星系中暗物质子结构的目标式无似然推断。载于第*34*届神经信息处理系统大会，2020年10月。

Coogan, A., Anau Montel, N., Karchev, K., Grootes, M. W., Nattino, F., and Weniger, C. 从不孤身前行：扰动体群体对强引力透镜中子晕测量的影响。2022年9月。

Cranmer, K., Pavez, J., and Louppe, G. 使用经校准的判别分类器近似似然比。*arXiv*预印本 *arXiv:1506.02169*，2015年。

Cranmer, K., Brehmer, J., and Louppe, G. 基于模拟的推断前沿。美国国家科学院院刊，117(48):30055–30062，2020年。

Dalmasso, N., Pospisil, T., Lee, A. B., Izbicki, R., Freeman, P. E. 和 Malz, A. I. 使用 Python 和 R 的条件密度估计工具及其在测光红移和无似然宇宙学推断中的应用。《天文学与计算》，30:100362，2020。

Dax, M., Green, S. R., Gair, J., Macke, J. H., Buonanno, A. 和 Sch¨olkopf, B. 利用神经后验估计进行实时引力波科学研究。《物理评论快报》，127:241103，2021年12月。doi：10.1103/PhysRevLett.127. 241103。
U
R
L
https://link.aps.org/doi/10.1103/PhysRevLett.127.241103。

de Witt, C. S.、Gram-Hansen, B.、Nardelli, N.、Gambardella, A.、Zinkov, R.、Dokania, P.、Siddharth, N.、Espinosa- Gonzalez, A. B.、Darzi, A.、Torr, P. 和 Baydin, A. G. 面向全球卫生决策的基于模拟的推断。2020年。doi：10.48550/ARXIV.2005.07062。
URLhttps://arxiv.org/abs/2005.07062。

Deistler, M., Goncalves, P. J., 和 Macke, J. H. 面向可扩展、便捷的基于模拟的推断的截断提议分布。*arXiv* 预印本 *arXiv:2210.04815*，2022年。

Dinh, L., Krueger, D., 和 Bengio, Y. NICE：非线性独立成分估计。*arXiv* 预印本 *arXiv:1410.8516*，2014。

Durkan, C., Murray, I., 和 Papamakarios, G. 关于无似然推断的对比学习。载于机器学习国际会议，第2771–2781 页。PMLR，2020。

Fearnhead, P. 和 Prangle, D. 为近似贝叶斯计算构建汇总统计量：半自动近似贝叶斯计算。英国皇家统计学会杂志：*B*辑（统计方法论），74(3):419–474，2012。

Frazier, D. T., Nott, D. J., Drovandi, C., and Kohn, R. 使用合成似然进行贝叶斯推断：渐近性质与调整。美国统计学会杂志，(刚刚接收)：1–28，2022。

Gelman, A. and Rubin, D. B. 使用多个序列从迭代模拟中进行推断。*Statisticalscience*，第 457–472 页，1992。

Gonc¸alves, P.J., Lueckmann, J.-M., Deistler, M., Nonnenmacher, M., Ocal, K., Bassetto, G., Chintaluri, C., Podlaski, W. F., Haddad, S. A., Vogels, T. P., Greenberg, D. S., and Macke, J. H. 训练深度神经密度估计器，以识别神经动力学的机制模型。*eLife*，9:e56261，2020年9月。ISSN 2050-084X。doi: 10.7554/eLife.56261。
URLhttps://doi.org/10.7554/eLife.56261。

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio,Y. 生成对抗网络。*stat*，1050:10，2014。

Greenberg, D., Nonnenmacher, M. 和 Macke, J. 用于无似然推断的自动后验分布变换。发表于国际机器学习会议，第 2404–2414 页。PMLR，2019。

Guo, C., Pleiss, G., Sun, Y. 和 Weinberger, K. Q. 现代神经网络的校准。*arXiv* 电子预印本，文章编号 arXiv:1706.04599，2017 年 6 月。doi：10.48550/arXiv. 1706.04599。

Hermans, J., Begy, V., and Louppe, G. 使用摊销近似比率估计器进行无似然MCMC。载于国际机器学习会议，第4239–4248页。PMLR，2020。

Hermans, J., Banik, N., Weniger, C., Bertone, G., and Louppe, G. 通过神经基于模拟的推断，利用恒星流约束温暗物质。*Mon. Not.Roy. Astron. Soc.*，507(2):1999–2011，2021a。doi: 10.1093/mnras/stab2181。

Hermans, J., Delaunoy, A., Rozet, F., Wehenkel, A., and Louppe, G. 避免基于模拟的推断陷入危机。*arXiv*预印本 *arXiv:2110.06581*，2021b。

Ho, J.、Jain, A. 和 Abbeel, P. 去噪扩散概率模型。*CoRR*，abs/2006.11239，2020年。网址：https://arxiv.org/abs/2006.11239。

Hyv¨arinen, A. 和 Dayan, P. 通过得分匹配估计未归一化统计模型。机器学习研究期刊，6(4)，2005年。

Karchev, K.、Anau Montel, N.、Coogan, A. 和 Weniger, C. 使用去噪扩散复原模型进行强透镜源重建。载于第*36*届神经信息处理系统大会，2022年11月a。

Karchev, K.、Trotta, R. 和 Weniger, C. SICRET：使用截断边缘神经比率估计的Ia型超新星宇宙学。2022年9月b。doi: 10.1093/mnras/stac3785。

Kingma, D. P. 和 Welling, M. 自动编码变分贝叶斯。*arXiv* 预印本 *arXiv:1312.6114*，2013年。

Legin, R.、Hezaveh, Y.、Levasseur, L. P. 和 Wandelt, B. 基于模拟的强引力透镜参数推断。2021年12月。

Linhart, J.、Gramfort, A. 和 Rodrigues, P. L. 基于归一化流的SBI算法验证诊断。*arXiv* 预印本 *arXiv:2211.09602*，2022年。

¨Lueckmann, J.-M.、Goncalves, P. J.、Bassetto, G.、Ocal, K.、Nonnenmacher, M. 和 Macke, J. H. 神经动力学机制模型的灵活统计推断。神经信息处理系统进展，30，2017年。

Lueckmann, J.-M., Bassetto, G., Karaletsos, T. 和 Macke, J. H. 使用仿真器网络进行无似然推断。*arXiv* 电子预印本，文章编号 arXiv:1805.09294，2018年5月。

Lueckmann, J.-M., Boelts, J., Greenberg, D., Goncalves, P. 和 Macke, J. 基于模拟的推断基准测试。载于 国际人工智能与统计学会议，第 343–351 页。PMLR，2021年。Marjoram, P., Molitor, J., Plagnol, V. 和 Tavar´e, S. 无需似然的马尔可夫链蒙特卡罗。 美国国家科学院院刊，100(26):15324–15328，2003年。Marlier, N., Br¨uls, O. 和 Louppe, G. 用于多指机器人抓取的基于模拟的贝叶斯推断，2021年。网址 https://arxiv.org/abs/2109.14275。

Miller, B. K., Cole, A., Weniger, C., Nattino, F., Ku, O., 和 Grootes, M. W. swyft：Python 中的截断边缘神经比率估计。开源软件期刊，7(75):4205，2022年a。doi: 10.21105/joss.04205。URLhttps://doi.org/10.21105/joss.04205。

Miller, B. K., Weniger, C., 和 Forr´e, P. 对比神经比率估计。2022年10月b。

Mishra-Sharma, S. 和 Cranmer, K. 用于表征银河系中心 γ 射线超出的神经模拟式推断方法。*Phys.Rev.D*，105(6):063017，2022年。doi: 10.1103/PhysRevD.105.063017。

Montel, N. A., Coogan, A., Correa, C., Karchev, K., 和 Weniger, C. 使用截断边缘神经比率估计，根据强引力透镜图像估算温暗物质质量。*Mon. Not. Roy. Astron.Soc.*，518(2): 2746–2760，2022年。doi: 10.1093/mnras/stac3215。

Ong, V. M.-H., Nott, D. J., Tran, M.-N., Sisson, S. A. 和 Drovandi, C. C. 利用合成似然在高维空间中进行无似然推断。《计算统计与数据分析》，128:271–291，2018。

Papamakarios, G. 和 Murray, I. 使用贝叶斯条件密度估计快速进行无需 ε的模拟模型推断。神经信息处理系统进展，29，2016。

Papamakarios, G., Sterratt, D. 和 Murray, I. 序贯神经似然：利用自回归流快速进行无似然推断。载于第 *22* 届国际人工智能与统计学会议，第 837–848 页。PMLR，2019。

Papamakarios, G.、Nalisnick, E. T.、Rezende, D. J.、Mohamed, S. 和 Lakshminarayanan, B. 面向概率建模和推断的归一化流。《机器学习研究杂志》，22(57):1–64，2021年。

Papernot, N. 和 McDaniel, P. 深度 k 近邻：迈向自信、可解释且稳健的深度学习。*arXiv* 电子预印本，文章编号 arXiv:1803.04765，2018年3月。doi: 10.48550/arXiv.1803.04765。

Perreault Levasseur, L.、Hezaveh, Y. D. 和 Wechsler, R. H. 神经网络估计参数的不确定性：在强引力透镜中的应用。《天体物理学杂志快报》，850(1):L7，2017。doi: 10.3847/2041-8213/aa9704。Prangle, D.、Blum, M.、Popovic, G. 和 Sisson, S. 使用覆盖率性质的近似贝叶斯计算诊断工具。" arxiv 预印本。*arXiv* 预印本 *arXiv:1301.3166*，412，2013。Price, L. F.、Drovandi, C. C.、Lee, A. 和 Nott, D. J. 贝叶斯合成似然。《计算与图形统计学杂志》，27(1):1–11，2018。Pritchard, J. K.、Seielstad, M. T.、Perez-Lezaun, A. 和 Feldman, M. 人类 Y 染色体的群体增长：Y 染色体微卫星研究。《分子生物学与进化》，16(12):1791–1798，1999。Ramesh, P.、Lueckmann, J.-M.、Boelts, J.、Tejero-Cantero, ´A.、Greenberg, D. S.、Gonc.alves, P. J. 和 Macke, J. H. GATSBI：用于基于模拟的推断的生成对抗训练。*arXiv* 电子预印本，文章编号 arXiv:2203.06481，2022年3月。doi：10.48550/arXiv.2203.06481。Rezende, D. 和 Mohamed, S. 使用归一化流的变分推断。载于国际机器学习会议，第 1530–1538 页。PMLR，2015。Rozet, F. 等。用于无似然推断的任意边缘神经比率估计。2021。Rubin, D. B. 对应用统计学家而言具有贝叶斯依据且相关的频率计算。《统计年刊》，12(4):1151–1172，1984。doi: 10.1214/aos/1176346785。网址 https://doi.org/10.1214/aos/1176346785。Schall, R. 置信区间的经验覆盖率：置信水平的点估计与置信区间。《生物计量学杂志》，54(4):537–551，2012。Skilling, J. 用于一般贝叶斯计算的嵌套抽样。《贝叶斯分析》，1(4):833–859，2006。doi: 10.1214/06-BA127。网址 https://doi.org/10.1214/06-BA127。Sohl-Dickstein, J.、Weiss, E. A.、Maheswaranathan, N. 和 Ganguli, S. 使用非平衡热力学的深度无监督学习。*CoRR*，abs/1503.03585，2015。网址 http://arxiv.org/abs/1503.03585。

Song, Y.、Sohl-Dickstein, J.、Kingma, D. P.、Kumar, A.、Ermon, S. 和 Poole, B. 通过随机微分方程进行基于得分的生成建模。*arXiv* 预印本 *arXiv:2011.13456*，2020。Stone, C. 和 Courteau, S. 径向加速度关系的内禀散布。《天体物理学杂志》，882(1):6，2019年9月。doi: 10.3847/1538-4357/ab3126。Stone, C.、Courteau, S. 和 Arora, N. 星系标度关系的内禀散布。《天体物理学杂志》，912(1):41，2021年5月。doi: 10.3847/1538-4357/abebe4。Talts, S.、Betancourt, M.、Simpson, D.、Vehtari, A. 和 Gelman, A. 使用基于模拟的校准验证贝叶斯推断算法。*arXiv* 预印本 *arXiv:1804.06788*，2018。Tejero-Cantero, A.、Boelts, J.、Deistler, M.、Lueckmann, J.-M.、Durkan, C.、Gonc.alves, P.、Greenberg, D. 和 Macke, J. sbi：用于基于模拟的推断的工具包。《开放源代码软件期刊》，5(52):2505，2020年8月。doi：10.21105/joss.02505。Thomas, O.、Dutta, R.、Corander, J.、Kaski, S. 和 Gutmann, M. U. 通过比率估计进行无似然推断。《贝叶斯分析》，17(1):1–31，2022。Treu, T. 星系的强引力透镜。《天文学与天体物理学年度评论》，48:87–125，2010。Vincent, P. 得分匹配与去噪自动编码器之间的联系。《神经计算》，23(7):1661–1674，2011。Wagner-Carena, S.、Park, J. W.、Birrer, S.、Marshall, P. J.、Roodman, A. 和 Wechsler, R. H. 使用贝叶斯神经网络的层次推断：在强引力透镜中的应用。《天体物理学杂志》，909(2):187，2021。doi：10.3847/1538-4357/abdf59。Wilson, A. G. 和 Izmailov, P. 贝叶斯深度学习与泛化的概率视角。载于 Larochelle, H.、Ranzato, M.、Hadsell, R.、Balcan, M. 和 Lin, H.（编），神经信息处理系统进展，第 33 卷，第 4697–4708 页。Curran Associates, Inc.，2020。网址 https://proceedings.neurips.cc/paper/2020/file/322f62469c5e3c7dc3e58f5a4d1ea399-Paper.pdf。Zhu, Y. 和 Zabaras, N. 用于代理建模和不确定性量化的贝叶斯深度卷积编码器-解码器网络。 计算物理学杂志，366：415–447，2018年8月。doi: 10.1016/j.jcp. 2018.04.018。

Zuo, Y.、Chen, C.、Li, X.、Deng, Z.、Chen, Y.、Behler, J.、Cs´anyi, G.、Shapeev, A. V.、Thompson, A. P.、Wood, M. A. 和 Ong, S. P. 机器学习原子间势的性能和成本评估。《物理化学杂志 *A*》，124(4)：731–745，2020年。doi：10.1021/acs.jpca.9b08723。
URLhttps://doi.org/10.1021/acs.jpca.9b08723。
PMID：31916773。

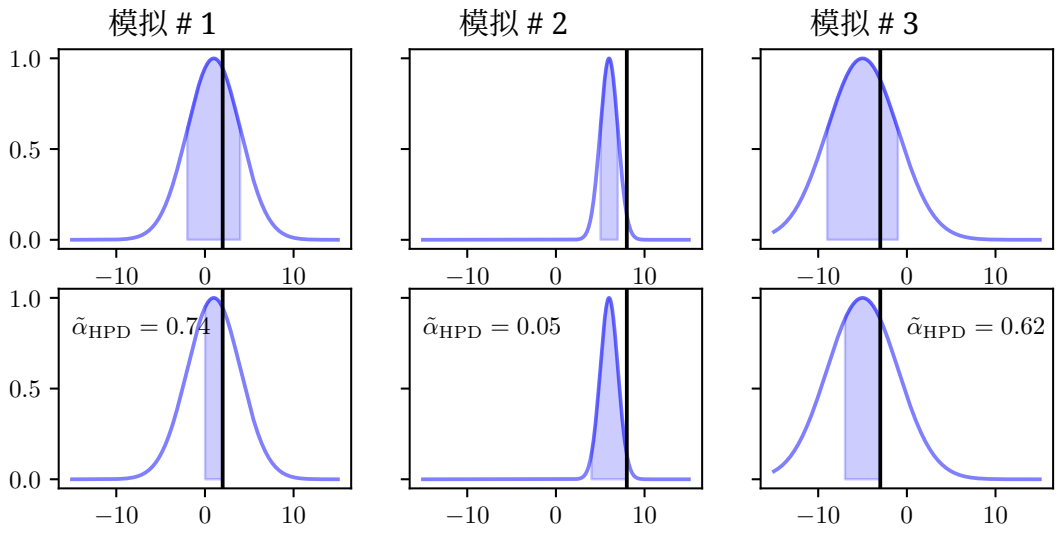


图 *6.* 该图展示了计算高后验密度覆盖率的两种方法背后的直观理解。每一列展示玩具示例中的三次模拟之一。蓝色曲线表示预测后验分布，黑色竖线表示相应的真实值。我们希望计算 68% 可信水平（$\alpha = 0.32$）下的覆盖率。第一种方法如顶行所示，包括计算 $1 - \alpha$ 可信区域，然后检查真实值有多少次位于该区域内（本例中为两次，因此覆盖率为 $2/3$）。另一种方法在 §3.1 中进行了介绍，如底行所示：首先找出由真实值定义的 HPD区域，然后找出 $\tilde{\alpha}_{\text{HPD}} < \alpha$，同样为两次。该图说明了这两种方法完全等价。

A. 两种定义之间的联系

§3.1讨论了计算覆盖概率的两种可能方法之间的差异，这两种方法都适用于HPD区域和TARP区域。本附录将进一步帮助读者直观理解二者之间的联系。首先来看HPD区域的情形，如图6所示。第一种方法或许更直观，但效率低得多：先计算 $1 - \alpha$ 可信区域，然后对每次模拟检查真实值落在该区域内的频率，并针对多个alpha值重复这一过程（注意其中的嵌套循环）。第二种方法则源于十分重要的引理 1，且已用于算法1：对每次模拟，找出由真实值定义的HPD区域及其对应的可信水平 $1 - \tilde{\alpha}_{\text{HPD}}$。然后，我们可以将 $1 - \alpha$ 水平下的覆盖率计算为 $\sum_{i=1}^{N} \tilde{\alpha}_{\text{HPD}} \geq \alpha$，其中 N 是模拟次数。
类似的逻辑也适用于 TARP 可信区域。虽然我们可以从参考点计算半径，使得 α达到某个特定值，但使用由真实值定义的可信区域在计算上高效得多，而且效果等同，如图 7 所示。算法2使用的就是这种方法。

B. 关于过度自信和信心不足图形的直观理解

习惯使用覆盖概率来验证 SBI分析的从业者，通常会看到过度自信和信心不足的曲线，例如图2中的蓝色曲线。然而，同一图也展示了 TARP 方法会为过度自信和信心不足的后验估计器生成不同的曲线。本附录旨在帮助读者直观理解这些差异。

首先，我们关注信心不足的后验分布，如图 8 上图所示。在这种情况下，我们看到 TARP覆盖率往往接近 0。5。这是因为，无论随机参考点位于何处，如果真实值接近后验分布的峰值，TARP面积都很可能覆盖分布的大约一半。另一方面，对于过度自信的后验分布，如图 8 下图所示，我们看到 TARP覆盖率往往接近 0 或 1。这是因为，无论随机参考点位于何处，如果真实值远离后验分布的峰值，TARP面积都很可能覆盖整个分布，或完全不覆盖。

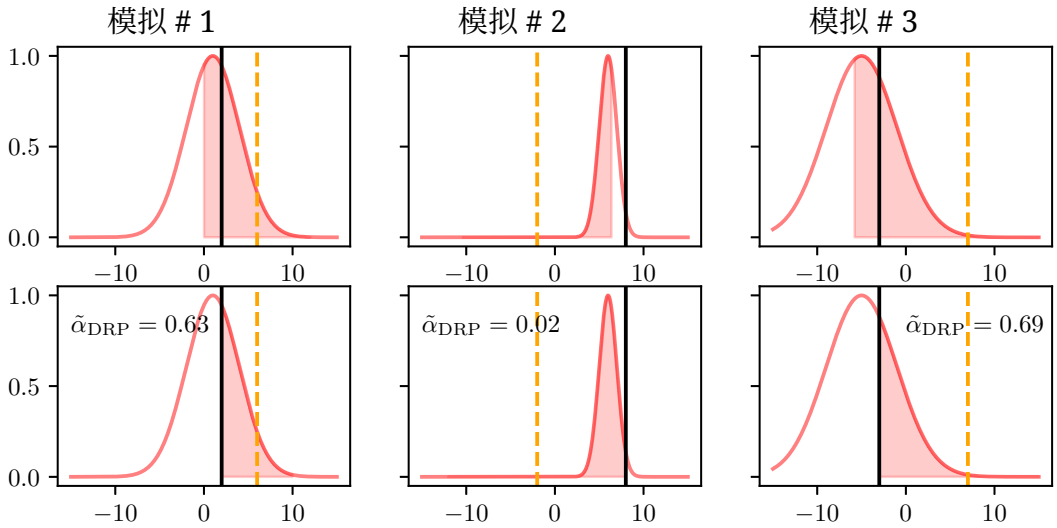


图7. 与图6类似，本图展示了计算“到随机点的距离”覆盖率的两种方法背后的直观原理。每一列展示了玩具示例中的三次模拟之一。红色曲线表示预测后验分布，黑色竖线表示相应的真实值。橙色线表示随机选取的参考点。我们想要计算 68% 可信水平（$\alpha = 0.32$）下的覆盖率。第一种方法如上排所示：找出以参考点为中心的 $1 - \alpha$ 可信区域，然后检查真实值落在该区域内的频率（此例中为两次，因此覆盖率为 $2/3$）。另一种方法如下排所示：找出由参考点和真实值定义的 TARP 区域，然后计算 $\tilde{\alpha}_{\mathrm{HPD}} < \alpha$ 的频率，结果同样为两次。这幅图说明了这两种方法完全等价。

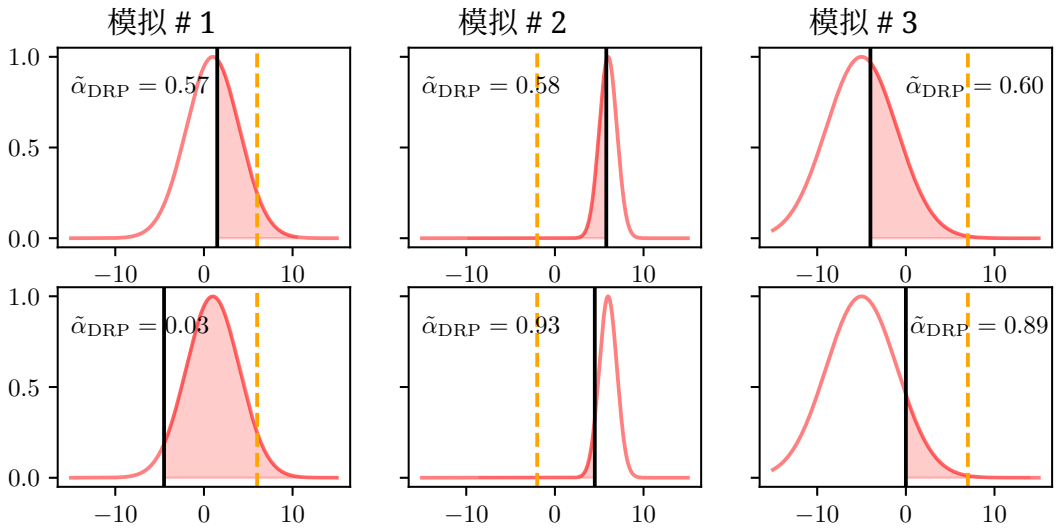


图8。 该图说明了使用 TARP 覆盖率得到过度自信和欠自信曲线形状背后的原因，例如图2中的曲线。上方展示了三个过度自信预测的模拟示例，下方则展示了欠自信预测。该图表明，TARP 覆盖率往往接近 0。5，而对于过度自信区域，它们则往往接近 0 或 1。

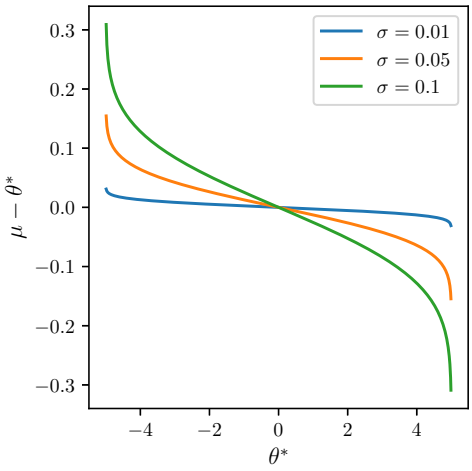


图*9*。该函　数用于 §4.1 有偏情形中作为均值，其依赖于位置，且对应三种不同的 sigma 值　　a.

C. 有偏情形实验函数

图9展示了函数 (26)，该函数在 §4.1 的一维情形中用作正态分布的均值。该函数说明，当 θ^* 为零时，分布以正确值为中心；而随着我们远离零，后验估计器的偏差会越来越大。

D. 对 θ_r 分布和距离度量的依赖

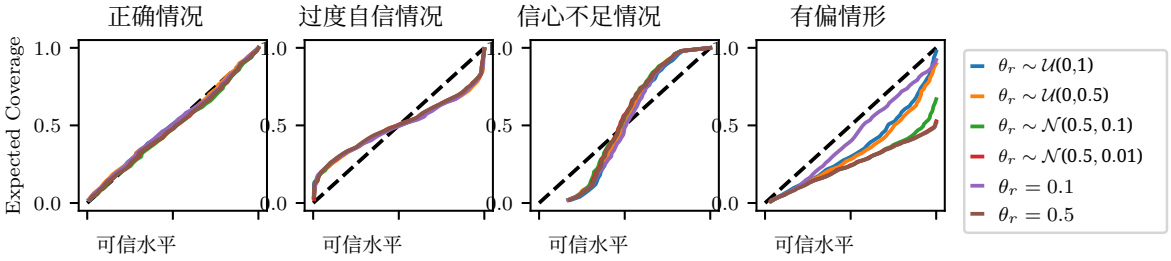


图*10*。 § 4.1 中所述的高斯玩具模型实验，按照 §4.2 中的描述，从不同分布中抽取参考点 θ_r 。

图10展示的内容与图2相同，但改变了用于抽取 θ_r 的分布。我们发现，这只会在有偏情形下产生影响；但即便如此，所有分布都明显呈现出偏差。图11展示了相同的实验，并比较了使用 L1 和 L2 作为距离度量的情况。在这种情形下，我们没有发现明显差异。因此，我们得出结论：我们的方法对分布和距离度量的选择具有稳健性。

在 §4.3 中，我们讨论了当分布 $\tilde{p}(\theta_r|x)$ 对 x 存在一定依赖时，若后验估计器只是简单地恢复先验，TARP 方法会揭示该后验估计器不准确。图12展示了在不同分布下这一实验的结果。我们看到，如预期所示，不依赖于 x 的分布（以实线表示）无法检测出后验估计器不准确。另一方面，强烈依赖于 x 的分布（以虚线表示）则能非常清楚地检测出估计器不准确。最后，我们展示了一种对 x 依赖较弱的分布。如预期所示，TARP 确实偏离了对角线，但偏离程度比其他依赖于 x 的分布小得多。

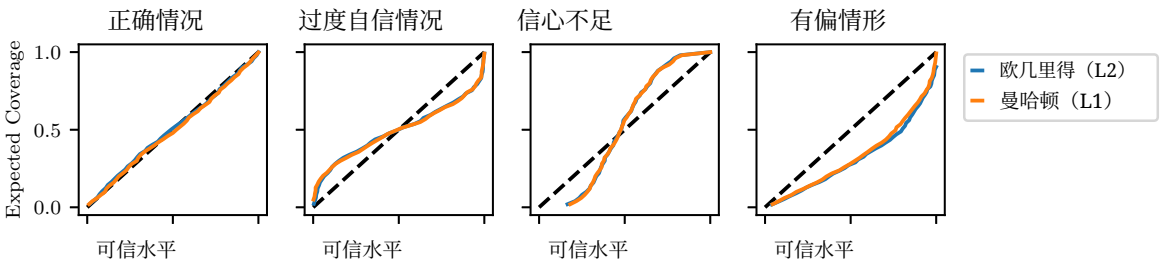


图*11*。§4.1 中介绍的高斯玩具模型实验，采用 §4.2 中介绍的 L1 或 L2 距离度量。

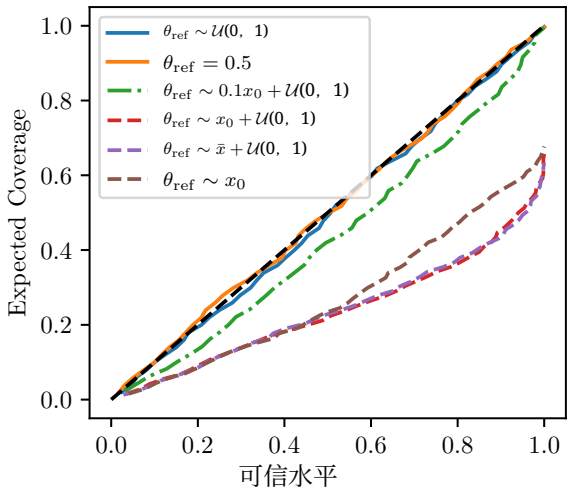


图*12*。与图4中相同的期望覆盖率与可信水平关系图，展示了不同的 $\tilde{p}(\theta_r|x)$ 分布。实线表示不依赖于 x 的分布，点划线表示与 x 弱相关的分布，点线表示相关性更强的分布。

E. 引力透镜实验详情

如Adam 等人 （2022年）所示，引力透镜源重建可以使用基于得分的建模技术来完成。下面我们总结基于得分的建模背后的关键思想，以及如何生成有偏和精确的后验样本。

基于得分的建模通过对从先验 p（θ） 采样得到的训练数据集添加不同尺度的噪声来实现，这些噪声尺度由 $t \in [0, T]$索引。此处， $t = 0$ 对应于未扰动的数据（p_0（θ） $= p$（θ））， 而 $t = T$对应于对数据施加极强扰动，使其淹没在噪声中并服从高斯分布（p_T（θ） $= \mathcal{N}$（$\theta|0,\ \sigma_T^2$））。加噪过程可由随机微分方程（SDE）（Song 等人，2020）描述

$$\mathrm{d}\theta_t = f(t, \theta)\,\mathrm{d}t + g(t)\,\mathrm{d}w\ , \tag{30}$$

其中， w 是标准维纳过程。利用去噪得分匹配（Hyv¨arinen 与 Dayan，2005；Vincent，2011年；Song 等人，2020），可以训练神经网络来近似时间相关的先验得分 $\boldsymbol{\nabla}_\theta p_t$（$\theta$），其中 p_t（θ）是经过加噪过程直到时间 t后得到的、定义在数据上的分布。给定先验得分后，可以从 p_T开始，通过随时间反向求解相应的反向随机微分方程（RSDE）来生成样本：

$$\mathrm{d}\theta = \left[f(t, \theta) - g^2(t)\boldsymbol{\nabla}_\theta \log p_t(\theta)\right]\mathrm{d}t + g(t)\,\mathrm{d}w\ , \tag{31}$$

其中，$\mathrm{d}t$ 是负时间步长。

为简单起见，我们不拟合基于得分的模型，而是对 PROBES 数据集中的星系图像拟合多元高斯分布，将其作为先验，得到 $p(\theta) = \mathcal{N}(\mu_0, \Sigma_0)$。我们采用 Song 等人 (2020) 提出的方差爆炸随机微分方程作为噪声过程。因此，时刻 t 的先验为 $p_t(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0 + \sigma_t^2\mathbb{I})$，其中 σ_t^2是噪声过程在时刻 t的方差。利用该表达式可以解析地计算先验得分。

为了修改采样过程，以便从 $p(\theta|x)$ 中生成样本（给定某个观测 x），我们必须对 RSDE 中的得分进行条件化处理，将先验得分替换为后验得分：

$$\mathrm{d}\theta = \left[f(t, \theta) - g^2(t)\boldsymbol{\nabla}_\theta \log p_t(\theta|x)\right]\mathrm{d}t + g(t)\,\mathrm{d}w\ , \tag{32}$$

根据贝叶斯定理，后验得分为

$$\boldsymbol{\nabla}_\theta \log p_t(\theta|x) = \boldsymbol{\nabla}_\theta \log p_t(x|\theta) + \boldsymbol{\nabla}_\theta \log p_t(\theta)\,, \tag{33}$$

其中，等式右侧的第一项是似然的得分。正如 Adam 等人 (2022年) 所指出的，这个依赖时间的似然通常难以处理，但可以近似为

$$\hat{p}_t(x|\theta) = \mathcal{N}(x|A\theta, \sigma_n^2\mathbb{I} + \sigma_t^2 AA^T) \approx p_t(x|\theta)\,, \tag{34}$$

其中，矩阵 A 编码了引力透镜畸变， σ_n 是观测中噪声的标准差（见 § 4.4）。然而，当 $p(\theta)$ 为多元高斯分布时，随时间变化的似然可以解析求得，其结果为

$$p_t(x|\theta) = \mathcal{N}(x|A\theta_c(\theta), \sigma_n^2\mathbb{I} + A\Sigma_c A^T)\,, \tag{35}$$

其中

$$\Sigma_c := \left(\Sigma_0^{-1} + \sigma_t^{-2}\mathbb{I}\right)^{-1}\ , \qquad \theta_c(\theta) := \sigma_t^{-2}\Sigma_c\theta\,. \tag{36}$$

因此，我们有两种方法对源星系光的后验分布进行采样：使用精确的随时间变化的似然（(35)）求解反向随机微分方程 (32)，或使用近似的、有偏的似然（(34)）求解。我们分别将它们称为“精确”采样器和“有偏”采样器。

最后，我们通过使用欧拉-丸山方法进行离散化，求解精确和有偏的反向随机微分方程（例如，参见 Song 等人(2020)）。我们发现，300 个步长足以确保收敛。