蒙特卡洛方法（公选课）课程作业

何显龙
2020102020001

2021 年 5 月 18 日

目录

1 引言

自去年九月入学以来，不断听课题组师兄师姐频繁提及 MCMC 方法和 corner 图；也逐渐了解到我导师朱宗宏教授的引力与天体物理课题组，在传统自然科学的理论、数据、实验方向中，尤其热衷、且擅长数据分析领域。初步调研与了解后，选修了这门《蒙特卡洛方法（公选课)》，以期迅速、系统性的掌握当前主流参数估计的基本框架。

物理学，研究的是物质运动的动力学演化规律。按照研究对象的典型尺寸划分，天体物理、固体物理、粒子物理基本覆盖了大中小三个空间尺度。固体物理或者说凝聚态物理，与人体尺寸相当，或许是人们对其了解的相当深入的一个原因，凝聚态已经成为物理学下最大的二级学科之一；基于牛顿力学的土木、材料、力学等工程学科，也逐渐深入影响和改变了人们的日常生活。

而同粒子物理类似，天体物理学由于其特征尺寸过于大、距离过于遥远，且很容易受到各种大尺度噪声干扰，很难完成高精度的实验测量；数据背后的物理实体不仅看不见摸不着更不可控，对自然科学要求的“实验证据”带来相当大的麻烦。正所谓“宇宙演化只有一次”，从这种意义上讲，“宇宙学”有一丝类似“考古学”，从过去留下的“遗迹”里去还原整个“历史”的原貌。

正是因此，天体物理学中还有相当多的物理机制等待人们用逐渐精细化的模型去理解宇宙。按照自伽利略时期建立的自然科学方法，理想的理解世界的模式是：理论学家基于某些实验事实提出一些公理性假设，基于此推理演绎出一套理论，并给出一些可通过实验检验、甚至区别于其他理论的可观测量；实验工程师将一手观测数据交给数据分析师，后者从中提取出合适的信息，带入理论模型中检验符合程度。在某些理论家一筹莫展的时候，可能还需要借助某些现象学的拟合，去帮助他们打开思路。

数据处理团队末端的一项重要工作，就是对理论学家给出的物理模型中的参数做限制，即所谓“参数估计”。最典型的例子莫过于粒子物理中“朗德（Landé）g 因子”的实验限制与计算，实验技术员测量原子磁矩与角动量，数据分析师据此给出其比值的 g 因子的限制，理论学家依据更基础的量子电动力学原理去计算，与“参数估计”的结果能精确保持一致到小数点后十余位！天体物理中常见参数估计结果如图 1，其描述的是 LIGO 和 Virgo 科学合作组织，基于引力波探测到的 11 个双黑洞并合事件目录 GWTC-1，针对黑洞质量分布函数统计学模型中几个参数绘制的其在参数空间可能的

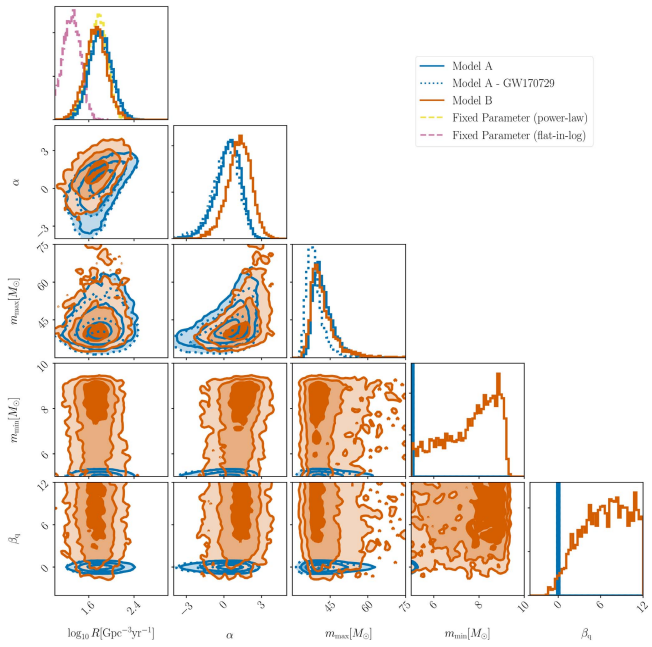


图 1: LIGO 和 Virgo 科学合作组织，针对黑洞质量分布函数模型中几个参数给出的限制。(B. P. Abbott and et al., 2019)

概率分布，进而对参数具体取值范围给出的限制。

参数估计最简单的方法来自传统概率论所代表的“频率学派”，即点估计、区间估计。然而由于经典数理统计将概率解释为频率的渐进行为，因此不可避免依赖重复抽样的大样本进行试验，这对于天体物理学中相对稀少的实验数据样本显得无能为力，例如图 1 仅试图通过 11 个事件分析统计学规律，似乎完全不可行。

基于条件概率中的 Bayes 定理，“贝叶斯学派”将历史经验作为先验分布引入，随后用实验数据对先验分布进行修正得到后验分布，将未知参数视作随机变量而非特定的值，其概率密度分布函数是修正后得到后验分布。这样基于很多小样本实验，也能做出理由充分的参数估计甚至决策行为。

然而为计算参数的后验分布，传统方式是（至少局部）遍历参数空间，然后逐一计算各点的概率密度值，最后再从整体后验分布函数分析参数估计的结果。这仅对于数个参数、简单的测量值与参数依赖关系、少量样本数据的计算量才可行，稍复杂的实际情况就显得无能为力。因此早期“贝叶斯学派”一度被忽视，只能研究可解析计算的“共轭先验分布”等少数案例，

直到计算机的算力大幅超过人类，Bayes 参数估计才逐渐推广流行。

通过遍历参数空间计算后验概率密度的“土”方法，遇到多个参数、高维参数空间情况，哪怕是利用计算机都显得臃肿乏力。能否也只从参数空间抽取少量具有代表性的点，来估计后验分布的轮廓呢？蒙特卡洛（Monte Carlo）方法提供了这样一种思路，其本质核心思想就是通过产生一系列随机数，来模拟真实的概率分布情况。

最后一个问题是，对 Bayes 参数估计给出的多参数、多数据的相当复杂（虽然是解析的）后验分布，如何构造抽样方法才能使得抽取尽可能少的样本（虽然原则上越多越好）、就能尽可能精确的代表后验分布的解析行为？基于转移概率矩阵的马尔科夫链（Markov Chain）给出了一种抽样方法，在细致平衡原理指导的 Metropolis-Hastings 算法下，此种抽样的平稳分布将趋近于复杂解析的后验分布。

终于，作为数据分析师，我们基于现有计算能力，完成了理论学家公式中未知参数的估计值。若他们还能从更基础的第一性原理出发、用其他更精确的物理常数计算值，与我们从实验数据出发给出的参数估计值一致，将标志着理论、数据、实验的三赢，物理学理解世界的台阶又将螺旋式上升一大步。本文希望概括性阐述，面对算力限制，在参数估计这一具体问题种人们遇到的种种困难，和更重要的当前主流解决方式。重点将关注与物理思绪与脉络的整理回顾，以期整理自己对这一问题提纲挈领的系统性理解，略去具体的数学证明并推荐至参考教材。

2 频率学派

人们最初认识到概率，就是通过事件发生的频率，来估计某件事发生的概率，例如掷硬币的正反面。普通本科教学的概率论，也通常只涉及此，被称为经典统计或“频率学派”。经典统计基于一系列定义：随机现象（试验，Experiment），样本空间和样本点，随机事件。随机事件发生的频率，被称作概率的统计学定义。参照频率的三条基本性质，柯尔莫哥洛夫于 1933 年将其类比作为概率的公理化定义，进而将概率论建立为完整成熟的科学理论，如等可能概型（古典概型、几何概型）、**条件概率**、独立性与伯努利试验等。

通常的本科概率论教材，如本章的主要参考书冯敬海，王晓光，鲁大伟 (2012)，重点会放在（一维、二维）随机变量的分布及其数字特征上，课时量不够的话或许会以大数定律、中心极限定理结尾，略去后半册数理统计部

分。而参数估计正是（经典）数理统计中的重点，所幸基于上半册概率论，有参数估计的需求时自学后半册数理统计并不算难，笔者花费约一周时间基本粗略掌握其思路，以下可视作读书笔记。

2.1 数理统计基本概念

概率论问题中的随机变量分布，由于其参数往往已知（或者假定），因而对随机现象的统计规律完全可以描述或预言。但应用于更多的实际问题，往往并不知道其参数、甚至未知其分布函数形式，尤其对复杂的天体物理背景的现象，很多时候早期研究只能借助所谓“现象学”的简单函数复合。因而参数估计问题，是数理统计最重要的问题之一。

统计问题中研究对象全体，被称作**总体**X，其中每个成员称作**个体**。为推断总体分布等特征，几乎不可能测量所有个体直接分析，因为很多测量是破坏性的，尤其是量子力学原理 2：测量导致波函数坍缩至本征态。因此必须按一定法则从样本中完成**抽样**，被抽取的第 i **个体**被称作**样本** X_i，最常用的是**简单随机抽样**。**统计量**指的是不依赖其他参数的**样本**的某个函数 $T = T(X_1, X_2, \ldots, X_n)$。最简单最常用的参数估计，就是用统计量估计样本均值和方差：

$$\bar{X}(\vec{X}) := \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_i^n X_i \tag{1}$$

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}^2) \tag{2}$$

基于标准正态分布，且相互独立的样本，常用解析函数式的统计量分布有三个：

- 若 $X_i \sim N(0,1)$，可定义卡方分布 $\chi^2 = X_1^2 + X_2^2 + \ldots + X_n^2 \sim \chi^2(n)$
- 若 $X \sim N(0,1), Y \sim \chi^2(n)$，可定义 t 分布 $t = \frac{X}{\sqrt{Y/n}} \sim t(n)$
- 若 $X \sim \chi^2(n), Y \sim \chi^2(m)$，可定义 F 分布 $F = \frac{X/n}{Y/n} \sim F(n,m)$

基于这些常用统计量的解析分布，原则上可对处于概率论中心地位的正态分布的组合分布进行计算。另一方面，也可以借助如下**单正态总体的抽样分布定理**，反过来用统计量估计正态分布的参数：

1. $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$，包含参数 μ、σ

2. $\frac{n-1}{\sigma^2}S^2 \sim \chi^2(n-1)$，仅含参数 σ

3. $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$，仅含参数 μ

2.2 参数的点估计

参数估计的基本前提是，总体 X 的分布函数形式 $F(x;\theta)$ 已知（或假定，如正态分布），但含有未知参数。基本思路构造合适的**统计量**作为参数 θ 的**估计量** $\hat{\theta}(X_1, X_2, \ldots, X_n)$，再带入抽取样本 X_i 的观测值 x_i，以得到参数的**估计值** $\hat{\theta}(x_1, x_2, \ldots, x_n)$。估计量和估计值被统称为**点估计**，两种常用方式是**矩估计法**和**最大似然估计**。

矩估计法由英国数学家 K. 皮尔逊于 19 世纪末 20 世纪初提出，其基本原理是观测数据的样本矩依概率收敛域带参计算的同阶总体矩，令其相等然后从中反解出参数。矩估计法相当直观且简单，尤其对于多参数，只需要从依次计算到足够方程数量的高阶矩即可。需要注意，不同阶矩对一个参数的估计结果可能不一致，因而为便于计算、通常从尽可能低阶（如一阶矩均值、二阶矩方差）开始。

最大（极大）似然估计，最早由德国数学家 Gauss 于 1821 年提出，英国统计学家 Fisher 于 1822 年重提并发展壮大。其基础是**最大似然原理**，即“概率最大的事件最可能出现”。其数学提法是，对来自总体 X 的简单随机样本 X_i 及其观测值 x_i，单次抽样概率 $p(x;\theta)$，因此在**独立同分布**的条件下，抽样结果为此的概率即为：

$$L(\theta) = \prod_{i=1}^{n} p(x_i;\theta) \tag{3}$$

可将此定义为关于参数 θ 的**似然函数（likelihood）**。因此按照**最大似然原理**，参数 θ 的**最大似然估计值** $\hat{\theta}(x_1, x_2, \ldots, x_n)$ 由下式决定：

$$L(\hat{\theta}) = \max L(\theta) \tag{4}$$

某些特殊情况，例如对单正态总体均值和方差的估计，最大似然估计量与矩估计量精确一致，但一般情况不一定相等。

最大似然估计的最经典案例之一，来自著名的**麦克斯韦速率分布**、即经典极限条件下理想气体遵从的**麦克斯韦-玻尔兹曼（速度）分布**。**玻尔兹曼系统**是统计物理早期的经典描述，粒子可分辨、且同一状态可容纳粒子数不受限制。基于平衡状态的孤立系统，对应的微观状态数最多的分布，被称作

最概然分布。参阅汪志诚 (2013)，计算结果显示，关键在于宏观系统的微观粒子 $N \approx 10^{23}$ 过于大，以至于即使对最概然分布仅微小偏离，其状态数与最概然分布状态数相比：$exp(-10^{13})$ 几乎是零，即这种分布形式非常接近于 δ 函数！这就是说最概然分布的微观状态数，非常接近于所有可能的围观状态数，因此最大似然估计完全可以认为是实际参数情况。

而对于其他情况尤其是简单案例中，远偏离 δ 函数、方差相对较大的分布，仅看最大似然估计值，其实不太能代表参数所有可能的情况，即“概率最大的事件”是“最可能”而不是“一定”出现！即参数估计值也只是“最可能”接近真实值，更贴近实际的应该是**区间估计**。

2.3 点估计优良性评定标准

容易看出，虽然总体分布的参数形式和样本观测值是给定的，但同一参数采用不同的估计方法会得到不同的估计量甚至不同的估计值，例如用不同阶矩估计均值显然不总是一样。指导做出选择的标准通常有三个：无偏性、有效性、一致性。

作为随机变量样本的函数，参数估计量 $\hat{\theta}(x_1, x_2, \ldots, x_n)$ 也是随机变量，因此会期待这个随机变量在多次试验中的结果，依概率趋近于待估计参数，即**无偏估计量**的定义：

$$E(\hat{\theta}(X_1, X_2, \ldots, X_n)) = \theta \tag{5}$$

无偏估计量的重要例子是方差估计量，即式 (2) 中等式右边分布是 $n-1$ 而不是方差定义的 n。主要原因在于：$E(\bar{X}^2) := D(\bar{X}) + E^2(\bar{X})$，而其中的 $D(\bar{X}) := \frac{1}{n^2} D(\sum X_i) = \frac{1}{n^2} \sum D(X_i) = \frac{\sigma^2}{n} \neq 0 = D(\mu)$

无偏性作为一个基本要求，描述的是参数估计量的（统计学）均值恰等于待估计参数，即一阶矩。显然下一步对多个无偏估计量的优劣程度判别，需要引入方差比较，即**有效性**的定义：若满足

$$D(\hat{\theta_1}) < D(\hat{\theta_2})$$

就称估计量 $\hat{\theta_1}$ 比 $\hat{\theta_2}$ 有效。参考齐民友，等 (2011)，单参下界由 Cramer-Rao 不等式给出：

$$D(\hat{\theta}) \geq \frac{1}{n \cdot I(\theta)}$$

$$\textbf{Fisher Information} : I(\theta) := E[(\frac{\partial}{\partial \theta} ln f(X;\theta))^2]$$

一致性（相合性），要求样本容量 n 充分大的时候，估计量 $\hat{\theta_n} = \hat{\theta_n}(X_1, X_2, \ldots, X_n)$ 需满足

$$\lim_{n \to +\infty} P[|\hat{\theta_n} - \theta| < \epsilon] = 1$$

由于一致性检验相对繁琐，故通常这一要求很少实际使用；可以验证，几乎所有的常见统计量都是一致性估计量。

2.4 区间估计与假设检验

上节2.3指出，参数的点估计量 $\hat{\theta}(x_1, x_2, \ldots, x_n)$ 也是随机变量，因此为了弥补只能知道某次点估计具体值，最好还需要知道估计值与参数真值之间的误差，即所谓**区间估计**。具体的说，就是依据节2.1给出的总体分布的已知解析形式，结合置信度的要求，反解出参数的置信区间。即，利用已知分布构造两个统计量 $T_1(X_1, X_2, \ldots, X_n), T_2(X_1, X_2, \ldots, X_n)$，使之满足

$$P\{T_1 \leq \theta \leq T_2\} = 1 - \alpha \tag{6}$$

则称**随机区间** $[T_1, T_2]$ 是参数 θ 的置信度为 $1 - \alpha$ 的置信区间。带入某次具体的样本观测值 x_i，就可得到这次具体的置信区间上下限。容易看出，置信区间的意义是，（这次观测）区间能覆盖到参数 θ 的概率为 $1 - \alpha$。

假设检验与参数估计类似，只不过后者是利用样本信息正向推断参数估计量，而前者是预先假设参数真值、再利用样本信息反向推断预设真值的正确与否。常用于判断在已知总体的前提下，判断某次试验、抽样结果是否正常、合理，细节从略。

3 贝叶斯学派

本章主要参考韩明 (2015)。

第 2 章提及的经典统计，是指 20 世纪初由 Pearson 等人开始、经 Fisher 的发展、到 Neyman 完成理论的一系列成果，占在目前国内外已出版的多种《概率论与数理统计》教材中的绝大部分篇幅，但随着实际问题和数据的日渐复杂，经典统计开始在一些领域显得捉襟见肘。

詹姆斯 · 伯努利（James Bernoulli）提出过一个著名的问题：演绎逻辑如何能帮助处理归纳逻辑的推断。托马斯 · 贝叶斯（Thomas Bayes, 1702 一 1761）于 1763 年（去世后两年）发表了一篇文章来回答这个问题，其中提

出的一个定理后来被以他的名字命名——贝叶斯定理。1812 年，Laplace 在他的概率论教科书第一版中首次将贝叶斯思想以贝叶斯定理的现代形式展示给世人，他本人不仅重新发现了贝叶斯定理、阐述得远比贝叶斯更为清晰，还用它来解决天体力学、医学统计、甚至法学问题。在 19 世纪，由于贝叶斯方法在理论和实际应用中存在不完善之处，更重要的是计算能力的限制，其并未得到普遍认可。20 世纪后，随着统计学广泛应用于自然科学、经济研究、心理学、市场研究等领域，贝叶斯统计的研究与应用逐渐受到国际统计学界的关注，形成 "**贝叶斯学派**"。尤其是近几十年 MCMC (Markov Chain Monte Carlo) 方法的研究，使贝叶斯统计的研究与应用得到了再度复兴。特别的，贝叶斯统计是当今大火的人工智能的核心，是使计算机具有智能的根本途径，其主要使用归纳、综合而不是演绎方法。本文主要关注贝叶斯推断应用于参数估计。

经典学派认为，总体母体分布中的未知参数 θ 是常数而非变数，尽管人们暂时还不知道它的值，但可以利用样本来对它进行估计。而贝叶斯学派则把 θ 看成随机变量，其分布可视 θ 的经验情况而预先假定它符合某一先验分布，然后再结合样本观测值的信息得出参数的后验分布，进而对母体分布进行统计推断。两个学派的核心差别是对于概率的不同定义。经典学派认为概率可以用频率来进行解释，估计和假设检验可以通过重复抽样来加以实现。而贝叶斯学派认为概率是一种信念，一切观测值、参数值都遵循某种分布函数。

经典学派招致批评的重要原因之一，是对式 (6) 区间估计的模糊解释。按经典统计的解释，重复多次抽样，置信区间 $[T_1, T_2]$ 能盖住参数 θ 真实值的频率是置信度 $1-\alpha$。然而人们关心的是参数在什么范围内的概率有多大? 或者说，我们能有多大的把握判断参数在某一个区间内? 因此经典统计中区间估计问题的提法与解释并不能令人满意。而贝叶斯统计从一开始就将参数视作随机变量，因此虽然结果与经典统计相同，但对其的理解正是符合常识和预期需求的，参数落于置信区间的概率为置信度。经典统计对于点估计的解释，也是一种长期使用时 "平均 "地考察结果的优良性，这种说明好坏的标准同样不十分妥当。

但贝叶斯统计也并非十全十美，选择**先验分布**时具有很强的 "主观性"，而（至多）只对个人决策有用；而基于频率渐进概率的经典统计学是更 "客观的" 工具，相对更符合科学的要求。另一方面，如果某个问题存在很强的先验信息或者相当复杂的数据结构，才适合热情地推荐贝叶斯方法。而若有

大批的数据和相对较弱的先验信息，而且一目了然的数据结构能导致已知合适的经典方法（即近似于弱先验信息时的贝叶斯分析），则没有理由去过分强调贝叶斯方法的优质性。

3.1 贝叶斯定理

贝叶斯学派奠基性的工作是贝叶斯定理（或贝叶斯公式），其可以分为：（离散）事件形式和（连续）随机变量形式。通常的《概率论与数理统计》教材中的**贝叶斯定理**定义为：设试验 E 的样本空间为 Ω，A 为 E 的事件，$B_1, B_2, ..., B_n$ 为样本空间 Ω 的一个划分，则

$$P\left(B_i \mid A\right)=\frac{P\left(A \mid B_i\right) P\left(B_i\right)}{\sum_{j=1}^{n} P(A \mid B_j) P\left(B_j\right)}, \quad i=1,2, \cdots, n \tag{7}$$

为理解贝叶斯定理（也称**全概率公式**），参照节 2 中提及的条件概率背后的定义，等式右边分子表示事件 A 和 B_i 同时发生（A 且 B_i）的概率，即 $P(A, B_i) := P\left(A \mid B_i\right) P\left(B_i\right)$ 。而由于 B_i 是对样本空间“不重不漏”的完整划分，显然 $P(A) := \sum_j P(A, B_j) := \sum_{j=1}^{n} P(A \mid B_j) P\left(B_j\right)$。而条件概率定义中事件 A 和 B_i “同时”发生，地位等价当然可以交换顺序：$P\left(A \mid B_i\right) P\left(B_i\right) := P(A, B_i) = P(B_i, A) := P\left(B_i \mid A\right) P\left(A\right)$。理解了贝叶斯定理的完整证明，就能向连续形式进行推广。

从贝叶斯学派的角度去理解，等式右边分子项 $P\left(B_i\right)$ 代表对样本空间划分的**先验分布**，$P(A \mid B_i)$ 表示事件 A 在划分 B_i 里的条件概率、即**抽样（样本）信息**；等式左边表示抽样结果为 A 的条件下，属于划分 B_i 的**后验概率**，其对 B_i 求和（积分）结果为 1，因为 B_i 是对样本空间 Ω 的“不重不漏”完整划分；因而等式右边分母项可视作“归一化因子”。

接下来终于可以正式定义（连续）随机变量形式的贝叶斯定理。对总体 X 的样本 X_i 及其观测值 x_i，其**联合密度函数**形式已知 $f(\overrightarrow{x}; \theta) = f(x_1, x_2, \cdots, x_n; \theta)$，未知参数作为（视作）随机变量（假设）具有特定取值范围 $\theta \in \Theta$。经典统计中的式 (3) 将这一联合密度函数定义为**似然函数**，其参数虽然未知、但是确定的数。而贝叶斯学派将未知参数 θ 也视作与 x_i 平权的随机变量、尽管其不是由抽样得到的观测值决定，因此经典统计中针对样本 X_i 联合的密度函数，从未知参数的角度看只不过是在在给定 θ 下的条

件密度函数，沿用**似然函数**的定义（但推广理解）：

$$L(x \mid \theta) = f\left(x_1, x_2, \cdots, x_n; \theta\right) = \prod_{i=1}^{n} p(x_i \mid \theta) \tag{8}$$

未知参数 θ 作为**随机变量**，理应遵从某种分布 $\pi(\theta)$，这通常是根据经验“假设”的**先验分布**。因而样本**与参数**共同的联合分布应该是 $h(x, \theta) = L(x \mid \theta)\pi(\theta)$，对应经典（离散事件形式）的 $P(A, B_i)$。接下来与式 (7) 离散型贝叶斯定理完全类似：

$$\begin{aligned} &P(A, B_i) \longrightarrow L(x \mid \theta)\pi(\theta) = h(x, \theta) = \pi(\theta \mid x)m(x) \\ &P(A) \longrightarrow m(x) = \textstyle\int_\Theta h(x, \theta)\mathrm{d}\theta = \int_\Theta L(x \mid \theta)\pi(\theta)\mathrm{d}\theta \\ &P(B_i \mid A) \longrightarrow \pi(\theta \mid x) = \frac{L(x|\theta)\pi(\theta)}{\int_\Theta L(x|\theta)\pi(\theta)\mathrm{d}\theta} \propto L(x \mid \theta)\pi(\theta) \end{aligned} \tag{9}$$

由于未知参数也是随机变量，上式定义的样本边缘密度 $m(x) = \int_\theta L(x \mid \theta)\pi(\theta)\mathrm{d}\theta$ 自然可以理解为在样本 X_i 被观测到之前，观测值 x_i 的分布。当样本一经观测得到值 x 后，就可根据预设先验分布推断出更有把握的未知参数的**后验分布** $\pi(\theta \mid x)$，进而估计再次观测的样本值 $\tilde{x}$，即**后验预测分布**：

$$\begin{aligned} \pi(\tilde{x} \mid x) &= \int_\Theta \pi(\tilde{x}, \theta \mid x)\mathrm{d}\theta \\ &= \int_\Theta \frac{\pi(\tilde{x}, \theta, x)}{\pi(x)}\mathrm{d}\theta = \int_\Theta \pi(\tilde{x} \mid \theta, x)\frac{\pi(\theta, x)}{\pi(x)}\mathrm{d}\theta \\ &= \int_\Theta \pi(\tilde{x} \mid \theta, x)\pi(\theta \mid x)\mathrm{d}\theta = \int_\Theta \pi(\tilde{x} \mid \theta)\pi(\theta \mid x)\mathrm{d}\theta \end{aligned} \tag{10}$$

其中第一行是对参数空间的积分，第二行是两次使用条件概率的定义重组，第三行 $\pi(\tilde{x} \mid \theta)$ 等于式 (3) 定义的似然函数。也就是说，首次样本观测值 x 指导计算出未知参数的分布 $\pi(\theta \mid x)$，进而再推断二次样本观测值 $\tilde{x}$ 的**后验预测分布**。

3.2 共轭先验分布

容易看出，贝叶斯理论中的似然函数仅依赖模型，先验分布的选择具有相当大的主观性。Bayes 和 Laplace 最初使用贝叶斯分析时，简便起见对未知参数使用常数（均匀）先验分布，即 $U(a, b)$。这种选取无信息先验分布的方法称为**贝叶斯假设**，形式简洁、使用方便，且符合人们对无信息的直观认识，具有其合理性。显然**在贝叶斯假设下，后验密度“正比于”似然函数**。这种被称为“逆概率”(inverse probability) 的方法对 19 世纪统计学发展早

期产生了巨大的影响，也招致了“主观性选择”的批评，甚至使得 Jeffreys 对贝叶斯理论进行了具有非常重大意义的改进，提出建议使用的“Jeffreys 先验分布”等**无信息先验分布**，促成所谓“客观贝叶斯学派”的建立。

先验分布的选择，在特定样本观测下对正确获取后验分布具有十分重要的意义。尽管式 (9) 从理论上提供了一种，用样本信息修正先验密度函数得到后验分布的方法，但实际问题中往往由于先验和似然并不简单，导致积分困难。一些先驱提议对不同的总体分布（似然函数形式），使用特定的**共轭先验分布（族）**，这样其后验分布可解析且仍属于该共轭分布族，在早期没有计算机辅助计算的年代提供研究的可能性。前文提及的、目前仍广泛使用的**贝叶斯假设**，尤其 $U(0,1)$，就是二项分布 $B(n,\theta)$ 的共轭先验分布族 $Be(a,b)$ 的最低阶情况 $Be(1,1)$。

教材韩明 (2015) 中的例题显示，当样本容量 n 增大时，后验分布的密度函数越来越向样本均值集中，后验方差越来越小，表明这时先验信息对后验分布的影响越来越小。这说明在小样本情况下，先验分布的选取较为重要；但随样本数据信息的增加，先验分布在贝叶斯分析中的敏感性变弱，因此其选择可以考虑从方便计算的角度出发，如选取共扼先验分布等。

另一个很有意思的是韩明 (2015) 的例题 2.3.8。对均值未知、方差已知的正态总体分布 $N(\mu,\sigma_0^2)$ 及其样本观测值 x_i，其共轭先验分布仍是正态分布；若均值 μ 的先验分布为参数已知的 $N(\mu_1,\sigma_1^2)$，则其后验分布必满足 $N(\mu_2,\sigma_2^2)$，其中：

$$\mu_2=\frac{\bar{x}h_0+\mu_1h_1}{h_0+h_1},\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$
$$\sigma_2^2=\frac{1}{h_0+h_1},h_0:=\left(\frac{\sigma_0^2}{n}\right)^{-1},h_1:=\left(\sigma_1^2\right)^{-1}$$

可以看出，如果样本容量 n 较大，h_0 项将占据主导，即后验均值会趋近于样本均值 $\mu_2\approx\bar{x}$、且后验方差趋于 0。另一方面，原则上先验分布的参数 μ_1,σ_1^2 可以随意选择，甚至笔者认为可以选取上次先验计算后的后验参数，即将后验结果作为先验再次重复计算！如此循环 m 次后，容易得到递推关系式：

$$\mu_{m+1}=\frac{\bar{x}h_0+\mu_mh_m}{h_0+h_m},h_{m+1}=(\sigma_{m+1}^2)^{-1}=h_0+h_m,m\geq1$$

容易解得：

$$\frac{1}{\sigma_m^2} = h_m = (m-1)h_0 + h_1 = \frac{m(m-1)}{\sigma_0^2} + \frac{1}{\sigma_1^2}$$
$$\mu_m = \frac{(m-1)\bar{x}h_0 + \mu_1 h_1}{h_m} = \bar{x} + (\mu_1 - \bar{x})\frac{\sigma_m^2}{\sigma_1^2}$$

即迭代多次后验结果视作先验的再次计算，最终方差的分布趋于已知常数、均值分布趋于样本观测均值，与增大样本容量的结果一致。但此法仅限于基于共轭先验分布的简单情况，更实际的情况是，很多时候对后验分布函数图的绘制都有相当大的困难，几乎不可能通过重复迭代提高估计精度，从结果上看也并无必要。

充分统计量是一个相当重要的经典统计和贝叶斯统计中为数不多相一致的观点之一。有定理保证，使用全部样本的分布算得的后验分布与用充分统计量算得的后验分布相同的，因此可以用于简化数据、降低维数，尤其早期贝叶斯统计学发展初期、缺少计算机辅助的年代。

3.3 贝叶斯统计推断

经典统计的点估计，就是寻找一个统计量，用其样本观测值去估计未知参数（的具体取值）。贝叶斯学派观点类似，寻找样本（或其观察值）的函数，以使它尽可能地”接近“未知参数的合理分布。

在试验的样本观测值之前，为寻找最好的估计量，首先需要在参数空间上定义**损失函数**（实操中是预设），表示估计量与真值不同而造成的损失。**风险函数**定义为损失函数的期望，对不同的分布的期望又可分为经典的一**致最小风险估计**、**先验风险**、**后验风险**。为使风险函数最小，解得的点估计量被称作**贝叶斯解**。对三种不同的预设：平方、绝对值、0-1 损失函数的贝叶斯解，分别对应三种常用点估计量的选择：**后验期望**、**后验中位数**、**后验众数估计**，分别对应后验分布的期望值、概率密度积分为 0.5 处的值、概率密度最大值。通常这三者并不相等，对简单的解析情况可能彼此是相互的加权平均，当后验密度函数分布是单峰且对称时三者才恰好相等。

找到合适的点估计量后，我们还需根据试验的样本观测值评定其估计误差，最好用的方式是用参数 θ 对估计量 $\hat{\theta}$ 的**后验均方差**或其平方根来度量。容易证明，取**后验均值估计量** $\hat{\theta_E} = E(\theta|x)$ 时，可使后验均方差 $MSE(\hat{\theta}|x)$ 最小，故实际中常取后验均值作为 θ 的贝叶斯估计。

需要注意的是，虽然前文平方损失选择下的后验风险与后验均值，在计

算公式上高度类似（仅相差后验分布分母的归一化常数），但二者目的并不相同，最好还是使用不同概念名称。后验风险的目的是，在未知观测值 x 的前提下，为使“可能的”损失、风险最小，目的是找到估计量（的函数形式），因此是对函数（量）a 求导（常数不影响求导）。而后验均方差，是已知样本观测值 x 的条件下，目的是计算参数的后验误差；因此比后验风险多出的归一化常数，其他解析形式完全一致。另一方面，高度类似的计算公式，也暗示从方差最小的目的出发，取后验期望（均值）估计、平方损失函数的选择就显得相当必要和合理了。

最后一个概念是贝叶斯统计框架下的**区间估计**，目的是希望在参数空间中寻找一个区间，使得其覆盖的后验概率尽可能大、且其区间长度尽可能小。从定义出发，可以通过损失函数计算，结果被称作最高后验密度 (highest posterior density) 可信集，简称 **HPD 可信集**。对简单的单峰后验分布，通常可信集的两端的后验概率密度应相等；而对复杂的多峰后验分布，可能得到的是几个互不相接的区间。因此有不少统计学家建议放弃 HPD 准则、采用相连接的**双侧等尾可信区间**为宜，即正负无穷端按累计密度分别等尾截断，显然单峰分布的等尾就是 HPD。其他概念与经典统计类似，此处从略。受限于篇幅，**假设检验**部分亦从略。

4 蒙特卡洛模拟

容易看出，贝叶斯方法的关键结果是式 (9) 未知参数后验分布函数的计算，其核心是式 (3) 似然函数，但考虑到各种实际情况，通常给定参数 θ 的条件下、$f(X|\theta)$ 作为是样本 X 的（条件概率的）联合密度函数，或者说观测值 x 对未知参数 θ 的依赖是相当复杂的。另一方面，该式的未知参数空间仅为 1 维，稍实际的情况，例如为估计正态分布，其均值和方差就已经 2 维，更**高维的参数空间**（更多未知参数一起估计）原则上几乎不可能遍历某局部参数空间，因为所谓“遍历”将导致**网格数随维数而指数增加**！这些都暗示我们不太可能得到实际后验分布的完整解析信息，只能通过按分布函数抽取样本进行模拟、以期良好的近似分布函数的数值表观形式。

这种随机抽样方法，现在被统称为**蒙特卡洛方法（Monte Carlo，MC）**，也称**随机模拟法**、**统计试验法**。其起源最早可追溯至 17 世纪甚至更早，Monte Carlo 是摩纳哥（Monaco）的一个著名城市，位于地中海之滨，以旅游、赌博闻名。Von Neumann 等把计算机随机模拟方法定名为 Monte Carlo 方法，

反应了其的随机性。至今，MC 方法几乎已经成为随机统计规律模拟的通用基础方法，除了**随机数**和**随机过程**之外，相当多确定性问题也可以被转化为概率统计模型、进而通过随机抽样，来高效逼近确定性问题的真值。

4.1 随机数

到底什么叫**随机性**？或者说，随机事件按概率发生，如何去实现这种概率事件？按照经典统计学派的观点，用实际大样本抽样的频率去近似概率。这对于经典简单的物理体系很容易实现，如著名的**蒲丰投针**、相当常见的掷硬币掷骰子这种古典概型。但对于稍微复杂的抽样分布，古典概型就显得有些捉襟见肘，更不可能使用经典物理方式耗时耗力抽样，例如使用正 20 面体的骰子掷出 0 9 这 10 个数字的均匀分布。计算机的出现，某种角度提供了这种大量重复性（抽样）操作的可能性。

狭义的**随机数**，特指 $[0,1]$ 区间内均匀分布的随机变量，即 $[0,1]$ 区间内的数等可能产生。因为这是最简单、最基本的随机变量，其他复杂概率密度函数模型的随机变量分布，几乎都可以通过变量替换等方式、从 $U(0,1)$ 产生，离散型也可以覆盖。

计算机上其实很难实现现实世界的随机性，即使哪怕后者例如“掷硬币”这种古典概率试验，在牛顿力学逐渐显示出其普适性后，都曾经一度有人认为：只要已知充足的初始、边界条件，原则上都可以通过经典力学方程式计算出完整的硬币运动方式，进而判断最终掷硬币的正反面结果，尽管其数学上可能是复杂的。尤其以开普勒的语录著名“20 世纪物理学大厦已经基本搭建完成！虽然还有相对论和量子力学两朵乌云。”不过这一物理过程对很多条件都过于敏感、轻微扰动都可能会覆盖很多解的不同区间（正面反面正面反面），使得在可视作随机变量的人力不可控的轻微扰动范围内，硬币正反面的结果看起来像随机变量。

当然我们并不关心，事实上也很难从掷硬币的条件出发去严格计算结果。从概率的角度说，我们更关心的其实是其概率分布，即随机性的结果是否优良。这就引出了**随机性统计检验**，常见的是**均匀性**、**独立性**，具体操作被称作频数分布检验。通常随机数的产生方式有**平方取中法**、**乘同余法**、**线性同余法等**，教材Rubinstein and Kroese (2017) 中也提及了许多其他方法。需要注意区别**随机产生（抽样）**与**均匀遍历**，我们的目的是为了随机的产生 $[0,1]$ 区间内的数，而不是产生形如 $0.1, 0.2, \cdots, 0.9, 1.0$ 样式均匀遍历。或者说，例如我们期待先产生 20 个数、使其期望间隔为 0.05，再在此基础上

再产生 30 个数、使这 50 个数期望间隔为 0.02；以此类推，在产生之前并不知道需要多少个数、进而避免等间隔抽样；产生之后不管有多少个数、都能（近似）满足等间隔分布的统计规律，即符合随机性检验。通常随机数的产生方式有**平方取中法、乘同余法、线性同余法等**，教材Rubinstein and Kroese (2017) 中也提及了许多其他方法。

4.2 随机抽样

得到了最简单、最基本的随机变量，我们就可以据此构造其他给定概率密度分布函数的样本抽样结果了，最简单的变量代换是从 $U(0,1)$ 到 $U(a,b)$。随机抽样的一个重要依据，是**累计分布函数（Cumulative Distribution Function）——概率分布函数（Probability Distribution Function）**的积分，满足 $U(0,1)$。由此对离散、连续的概率分布函数，原则上都可以求积分、产生 $U(0,1)$ 随机数、反解目标随机数，这是基本的**反函数抽样**方法。

然而对于某些形式稍微复杂的分布，常常不可解析反解，若每次抽样再涉及数值解方程就显得相当臃肿。某些情况可能可以通过变量代换完成，典型的例子是正态分布（麦克斯韦速度分布）抽样，增加一维分布、再将直角坐标转化为极坐标、进而可解析反解，回忆我们在高等数学中对高斯积分（正态积分）也是用的这种办法。这被称作**变换抽样方法**。

更普遍、复杂的情况是只知道分布函数的形式，连积分得到累积分布都相当困难，这时候只能使用 **Von Neumann 舍选抽样方法** (田菲, 2007)。其图景有些类似 “蒲丰投针” 或 “通过方形及其内接圆撒点以估计 π”，用分布函数的最大值归一化至 $[0,1]$ 区间后，先随机产生 x 轴上的数，再产生随机数与 $f(x)$ 比较，若小于就保留 x、即落在分布函数下方、表示被正确覆盖的面积，否则舍去。

但此通法有个较为致命的缺陷。由于我们目标是 x 轴上的数，判定条件是随机数与 $f(x)$ 比较，因而其**接受率**很受 $f(x)$ 形式的影响，特别的对于小方差的正态分布这种高峰值函数，峰值以外的密度值由于过小而很难被抽样获取、导致抽样效率非常低。这使得我们对分布函数的 “归一化” 做改进，不至 $[0,1]$ 区间的水平线，而是构造一个已知的简单形式去覆盖分布函数，例如用峰状三角形去覆盖正态分布，这样就直接砍掉近半无效的抽样空间。当然还有其他舍选方式，例如乘分布抽样、减分布抽样、乘减抽样、积分分布抽样等，此处从略。

MC 方法的应用，主要在物理上的随机问题，和通常方法处理不了的如

高维积分等、或其他要求效率但不要求精度的确定性问题上。相比通常均匀遍历的方式，随机数对简单的过程并不占优势，但对复杂繁琐的数学结构，MC 方法几乎是有限时间唯一可做的方式，例如对本文的关键性目标问题：**高维参数空间下繁琐数学结构的贝叶斯后验概率分布函数**，连分布函数的样式都未知、或者说就是希望知道其轮廓或等高线，几乎只能依靠后文将提及的 MCMC 方法。

另一方面，对向量 $\boldsymbol{X} = (X_1, \cdots, X_n)$ 进行抽样时，若各分量分别独立还相对容易抽样。但当 X_i 不独立时，就变得相对麻烦。并且，随着向量维数增加，抽样效率渐进失效，即所谓“维数灾难”。MCMC 方法可回避维数灾难、Gibbs 抽样可解决分量关联，即先给定种子（先验），按条件概率依次抽一圈，得到一个向量。重复多次以收敛。

5 马尔科夫链

上一章 4 简单介绍了 MCMC 方法中的前者，蒙特卡洛方法（Monte Carlo），笼统地说，只要涉及随机模拟方法、或者使用到 $U(0,1)$ 随机数的都可称为 MC 方法。接下来将介绍 MCMC 中的后者，马尔科夫链（Markov Chains），属于随机过程的一种。同“随机数”一样，**随机过程**也很容易从命名上理解。本章主要参考的Rubinstein and Kroese (2017) 书中从数学上对其的定义是，形如上节末单个随机变量的集合，或者说随机向量 $\boldsymbol{X} = (X_1, \cdots, X_n)$ 的各分量，如果视其为时序依次演化，就被称作随机过程，注意区分随机变量（random variables）与随机过程（stochastic process）的英文原文。

泊松过程是相当常用的一种随机过程，其在一段区间的事件数满足泊松分布，且其参数仅与速率参数（简单的时齐过程为给定常数）、区间长度（例如时间间隔），由于其源自两点分布的微元极限，而能近似描述相当多的实际情况，尤其是如某段时间访问某地的人数。对泊松过程的模拟，本质上就是产生一系列正比于区间长度间隔的泊松分布随机变量，也是比较容易做到。

马尔科夫过程也是一类简单、理想的随机过程，特指下一时刻随机变量的值只受上一时刻的影响，而与此前所有状态都无关；或者更严格的说，停留在下一个可能状态上的概率，只与上一次的状态有关，抹去之前的所有历史记忆。对离散的时刻，其十分类似链式一环套一环，因而被称作**马尔科夫**

链；连续的时刻，用狭义的**马尔科夫过程**指代。可以被认为是某种**随机游走**过程。

由于马氏链描述的是不同状态之间相互转移的方式，彼此间转移概率组成的**转移概率矩阵**就乘了马氏链的数学核心。目前基于这一随机矩阵的应用相当广泛。Google 的各网站节点之间的链接，正是转移概率矩阵的形式。基于大量样本的分析，可以计算出从此网站出发转移向下一个网站的概率排名，再把转移概率高的排在前列，以便于后续网民的资料查阅过程，这正是搜索引擎的核心算法（当然也可以投放广告、人为修改权重）。

作为一个随机变量的序列，我们当然会期望考察其**平稳分布**，即时间序列趋于无穷时，随机变量在各态上可能的分布情况。幸运的是有定理保证，对于**不可约**、**非周期**的马氏链，其平稳分布存在且唯一，还不依赖于初始分布的给定；数学上也就是任给一个初始状态或者初始分布，乘上转移概率矩阵的无穷次幂，总会趋于唯一的平稳状态。其行为十分类似“不动点”，数学上回到线性代数中对转移概率矩阵的“左”特征向量。**非周期性**容易理解，这在马氏链里非常重要。**不可约性**的解释是只存在唯一的所有状态组成的**类**，使得这个**类**中的所有状态都存在某个时刻能彼此**互通**，或从某个状态总能跳跃到其他所有状态，哪怕不止一步，借用了《群论》的名词“不可约”、“等价类”。

从平稳分布的定义出发，可以推导出所谓**细致平衡原理**，其描述的是平衡状态下，从 i 状态转移到 j 状态的所有可能的概率和，与从 j 状态转移到 i 状态的概率和相等，即从 i 流出至 j 的全概率、应该等于从 j 流入 i 的和，从化学反应方程式的角度理解就是正逆反应速率相等。MCMC 方法的 M-H 抽样的核心思想，正是基于此原理。

6 MCMC 方法

如张磊 (2013)、叶纺 (2014)、赵琪 (2007) 所述，在贝叶斯估计中，最后一步我们要把后验均值作为参数的估计。但事实上我们得到的参数后验密度往往非常复杂，哪怕有时候只是解析地表示出后验密度的分布形式的往往都不太可能（需遍历高维参数空间），不得不只从后验密度抽样样本来代替总体。这其中的关键问题就是如何高效、正确的抽样，才能符合真正的后验密度分布形式。同样由于实际的后验密度往往过于复杂，前文所述传统的 MC 舍选抽样方法几乎不太可行，这时我们就需要使用 MCMC 方法抽

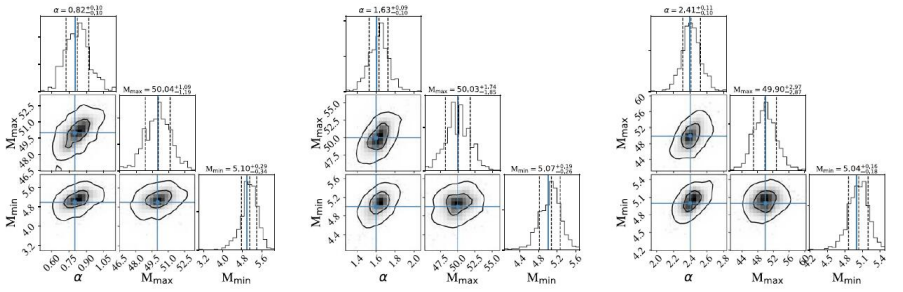
样。

前文所述的模拟过程中不改变抽样分布叫做静态蒙特卡洛方法，只能解决一些较简单的问题。1953 年，统计物理学家Metropolis et al. (1953) 等人首次将马氏链引入蒙特卡洛方法，作为一种动态的蒙特卡洛方法，解开了普通蒙特卡洛方法不改变抽样分布的瓶颈。此后 Hastings 随后对其加以概括和推广，奠定了 MCMC 方法的基石；1984 年 Geman S 和 GeInan D 引入了 Gibbs 算法，目前 MCMC 方法已成为一种标准化的统计计算工具。

如李航 (2019)，具体地说，MCMC 方法的基本想法是: 在随机变量的状态空间上定义一个满足遍历定理的马尔可夫链，使其平稳分布就是抽样的目标分布；然后在这个马尔可夫链上进行随机游走，每个时刻得到一个样本；根据遍历定理，当时间趋于无穷时，样本的分布趋近平稳分布。所以，当时间足够长时 (时刻或角标大于某个正整数 m)，在之后的时间里随机游走得到的样本集合就是目标概率分布的抽样结果。至于这个所谓**预热期**选择剔除达到“平稳分布”前的多少个样本点，很多时候是某种经验选取。

其整体框架是利用马氏链基于上一时刻完成下一时刻的抽样，构造马氏链的**转移概率矩阵（离散）**或**转移核函数（连续）**的过程中，第一步使用 MC 方法随机抽样到下一状态，第二步使用到类似**舍选抽样**的思想、定义所谓**接受概率**以指定是否进行此次跳跃。为使得马氏链抽样平稳，必须使得转移核函数满足前文所述的细致平衡原理，即在添加接受概率这一项因子的情况下保留其他有一定随意性取的函数类型。由于类似接受拒绝方法，需注意此处的接受率（亦即抽样效率）也不能太低，否则抽样很难遍历全部参数空间。接受概率是转移两态目标分布与建议转移核函数（有一定选择随意性）之积的比值，与 1 取较小值，目标分布函数在此、经由细致平衡原理指导最后的平稳分布。更重要的是，由于此处真正起作用的是转移两态目标函数之比，因此可以忽略原本就繁琐的贝叶斯后验分布函数的核的归一化积分常数，只计算其分子项的核项。这种基本方法被称作 **Metropolis-Hastings 算法**。

Gibbs 算法或 **Gibbs 抽样**，某种角度可以视作 **Metropolis-Hastings 算法**的特例，通过选择特殊的建议转移核函数以使其**接受概率**恒为 1。基本思想是针对多参数估计，从一个初始样本（条件）出发，依条件密度对多个参数依次迭代更新产生新的参数向量分量，小循环一次得到样本序列中的下一个，依次进行大循环最终得到全部样本序列。这对高维参数空间尤其适用，直接略去舍选抽样中计算目标分布的步骤，几乎减半了计算量。



(a) assuming $\alpha = 0.8$, $M_{min} = 5M_{\odot}$ and $M_{max} = 50M_{\odot}$.

(b) assuming $\alpha = 1.6$, $M_{min} = 5M_{\odot}$ and $M_{max} = 50M_{\odot}$.

(c) assuming $\alpha = 2.4$, $M_{min} = 5M_{\odot}$ and $M_{max} = 50M_{\odot}$.

图 2: 使用模拟数据，对 Power Law 形式黑洞质量分布函数模型中三个参数给出的限制。Ding et al. (2020)

7 应用：Corner 图

最终，得益于 MCMC 方法和计算机辅助，原则上解决了高维参数空间、复杂参数依赖关系的依分布抽样的问题，避免全参数空间的遍历计算，仅通过抽样少量样本就得到了参数后验概率分布图的等高线形式，结果以 corner 图 1 的形式进行展示，其中每张小图代表高维参数空间的 1D 直方图或 2D 等高线图（边缘分布），进而可以据此对参数进行区间估计。

但需要注意的是，不是所有 corner 图都是基于 MCMC 方法。数据处理分析团队、或数值模拟工作者，更进一步的思路是：当前不关心参数估计的具体值，而关心能以多高的精度估计此参数。例如最近我学习和重现的工作Ding et al. (2020) 及图 2 ，其关注的重点是使用模拟数据、误差引入方式、选择效应等其他实际因素能以多高的精度重现预设的分布函数。因而此例是对例如 1000 次实现（realization），每次实现中用 1000 个样本估计参数，此处仅用（极大似然）点估计（以节省机时），关心 1000 次实现的统计学性质（分布）。而不是针对图形中对某 1 次实现的一个数据点，改用 MCMC 去估计这一次实现下用 1000 个样本估计参数的后验分布结果等高线图。

Bibliography

B. P. Abbott, R. Abbott, T. D. A. and et al. (2019). Binary black hole population properties inferred from the first and second observing runs of advanced LIGO and advanced virgo. *The Astrophysical Journal*, 882(2):L24.

Ding, X., Liao, K., Biesiada, M., and Zhu, Z.-H. (2020). Black hole mass function and its evolution—the first prediction for the einstein telescope. *The Astrophysical Journal*, 891(1):76.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092.

Rubinstein, R. Y. and Kroese, D. P. (2017). *Simulation and the Monte Carlo Method, 3rd Edition*.

冯敬海，王晓光，鲁大伟 (2012). 概率论与数理统计. 高等教育出版社.

叶纺 (2014). 马尔可夫链蒙特卡罗方法及其 *R* 实现. PhD thesis, 南京大学.

张磊 (2013). 随机波动率模型参数估计*:* 贝叶斯和极大似然方法. PhD thesis, 清华大学.

李航 (2019). 统计学习方法第二版. 清华大学出版社.

汪志诚 (2013). 热力学 · 统计物理. 第 *5* 版. 高等教育出版社.

田菲 (2007). 马尔可夫链蒙特卡罗算法. PhD thesis, 湖北大学.

赵琪 (2007). *MCMC* 方法研究. PhD thesis, 山东大学.

韩明 (2015). 贝叶斯统计学及其应用. 同济大学出版社.

齐民友，等 (2011). 概率论与数理统计. 第 *2* 版. 高等教育出版社.