迈向格点量子色动力学中的完全贝叶斯分析

Julien Frison*John von Neumann-Institut für Computing NIC,
Deutsches Elektronen-Synchrotron DESY, Platanenallee 6, 15738 Zeuthen, Germany

我们提出了一种有前景的方法，通过贝叶斯推断学习物理参数，并利用现代工具取代传统拟合，以及误差的计算和传播方式。我们构建了若干模型，以格点量子色动力学数据中的一个现实案例为例进行说明；结果表明，这些模型能够提取大量信息，且稳定性良好。我们讨论了如何采用完全贝叶斯方法或信息准则评估这些模型，以及仍待解决的模型构建难题。

arXiv:2302.06550v1 [hep-lat] 13 Feb 2023

I. 引言

贝叶斯推断为解释统计分布的数据并学习物理参数提供了一种定义明确的方法；它能够灵活应用于多种模型，而无需任何特定假设。它通常与频率学派方法相对，但可以说更胜一筹，因为任何频率学派陈述都可以重新诠释为带有隐藏先验的贝叶斯陈述。贝叶斯方法的主要缺点是需要更多计算资源。在这方面已有一些进展，也出现了一些功能强大的软件，如今这些软件已广泛应用于学术界的许多领域，甚至工业界，并作为机器学习工具之一。这一趋势未来很可能进一步发展，同时汇聚统计学界及许多其他学科的力量。本文将展示如何在格点量子场论（LQFT）的典型分析中使用这些通用工具。

截至目前，LQFT 的最先进方法包括非线性加权最小二乘法（WLS）或广义最小二乘法（GLS），通常还会结合一些重采样技术[1] ，例如非重叠分块自助法。有时，重采样会被误差线性传播[2][3]所取代。最后，伪贝叶斯模型平均（pBMA）也日益普及[4]，并作为第三层方法发挥作用。我们的工作旨在以一个统一框架取代上述所有方法，在单个步骤中完成全部处理，并提供定义明确的解释，即一种完全且明确的贝叶斯分析。

我们将在第 II 节首先介绍符号约定，以及贝叶斯分析基础知识的一些入门内容。已经熟悉贝叶斯分析的读者可以选择跳过本节，以及第 IV 节，因为我们将在该节中考虑一些基础的玩具模型来对此进行说明。在第 III 节中，我们将回到传统方法，并将其应用于实际的 LQFT 数据。在第 V 节中，我们将针对同一数据介绍贝叶斯模型；随后在第 VI 节中，我们将讨论如何选择并对所构建的模型进行平均；最后，在第 VII 节中，我们将讨论设定错误的模型这一情形。

II. 基础知识

A. 符号约定

在本文中，我们将考虑一组数据 y，它由 n 个向量样本 y_i 组成，每个向量的维度为 p；以及一个参数为 a、维度 $\dim(a)=k$ 的模型 M 。经验均值 $\sum_i y_i/n$ 记为 $\bar{y}$。我们所说的模型，是指一个函数 $a \mapsto M(a)$，其中 $M(a)$ 是一个分布。我们将真实模型记为 M^*，即用于生成数据 y 的模型；在数据量有限时，我们无法真正获知它。可能存在也可能不存在一个真实参数 a^*，使得 $M(a^*) = M^*$。若存在，我们称模型 M^* 可由 M 参数化；若不存在，则称模型 M 设定错误。本文使用一般记号 $P(X|Y)$ 表示在预先已知某些信息 Y 的条件下，X 的条件概率；类似地，使用 $E(\ldots)$ 表示期望值，或使用 $V(\ldots)$ 表示方差。均值为 μ、标准误为 σ 的正态分布记为 $\mathcal{N}(\mu,\ \sigma)$。

* julien.frison@desy.de

B. 贝叶斯公式

虽然我们无法直接获取 a^*，但可以用贝叶斯公式来表达我们的认知

$$P(a|y,M) = \frac{P(y|a,M)P(a|M)}{P(y|M)}. \tag{1}$$

等式左边称为后验分布，而与数据无关的 $P(a|M)$ 是一种半任意的分布，称为先验。它们通过似然 $P(y|a,\ M)$ 联系起来；似然编码了模型的核心信息，而分母 $P(y|M)$ 称为边缘分布，因为它可以通过对分子中的 a 进行边缘化（积分）得到。边缘分布通常可以视为一个常数归一化因子，实际应用中无需计算。

如果在观测任何数据之前，a 的所有取值具有相同的概率，就称先验是平坦的。如果先验相当接近平坦（相对于数据中包含的信息量而言），使得后验分布的行为主要受似然的影响，则称其为无信息先验。

我们还将使用另一种分布——后验预测分布（PPD），它由后验分布构造而成，如下所示：

$$P(y'|y,M) = \int P(y'|a,M)P(a|y,M)da. \tag{2}$$

这个 PPD 的存在使贝叶斯模型成为生成式机器学习模型家族的一员。

C. 最大似然与最大后验

我们将使用两种密切相关的点估计量。
频率学派分析通常使用最大似然估计量（MLE），其定义为

$$a_{MLE} = \underset{a}{\operatorname{argmax}}\, P(y|a,M). \tag{3}$$

一种类似但在我们的情形下更实用的定义是最大后验估计（MAP）

$$a_{MAP} = \underset{a}{\operatorname{argmax}}\left[P(y|a,M)P(a|M)\right], \tag{4}$$

它是后验分布的众数。这个量无法提供参数不确定性的任何信息，但可以利用海森矩阵构建局部高斯近似，这正是 δ 和 Γ方法[2] 所做的。

这可能是一个危险的操作对象，因为原则上，似然在尖峰或漏斗状区域可能存在局部极大值，而这些区域实际上并不代表概率体积中的很大部分。发生这种情况时，适当地对模型进行重新参数化会有所帮助，因为与后验概率的体积不同，MLE 和最大后验估计不会在重新参数化下保持不变。

先验为平坦先验时，MLE 与最大后验估计相同。有趣的是，平坦先验在理论上并非更优的选择。事实上，已有研究证明，在高维情况下，MLE（因此也包括平坦先验）并非明智之选[5]，因为总能找到一个更有效的估计量，而且不必在偏差-协方差权衡上作出任何牺牲。再一次，平坦 的含义并不在重新参数化下保持不变。

D. 最小二乘估计

最小二乘法包括将一个函数 c 最小化，即 $a \mapsto \sum_i ||y_i - f(a)||$；对于广义最小二乘法 :

$$\chi^2_C(a) = \sum_{i=1}^{n} \left[y_i - f(a)\right]^\dagger C^{-1} \left[y_i - f(a)\right], \tag{5}$$

其中，$p \times p$ 矩阵 C 是预先施加的某种输入，并在最小化过程中保持不变。
这相当于计算高斯似然的最大似然估计

$$P(y|a,M_C) \propto e^{-\chi^2_C(a)/2} \propto e^{-\frac{n}{2}[\bar{y}-f(a)]^\dagger C^{-1}[\bar{y}-f(a)]}, \tag{6}$$

其中我们注意到， M_C 一个模型将 C 固定为其固有属性的一部分，同时保持 a 自由，因此进行最小化的所有参数都作用于高斯分布的均值。

请注意，无论 y的真实分布如何，公式 6 都可以写成以下两种形式之一： y （拟合n 个带协方差 C的数据点）或 $\bar{y}$ （拟合 1 个带协方差 C/n的数据点）。这意味着，一旦选定 C ，对每个 i 分别强制采用高斯模型，还是仅对均值采用高斯模型，都不会影响后验分布或最大似然估计。但这会影响第 VI 节中对拟合质量的讨论。在 C 选为 y的经验协方差这一特定情形下，这与中心极限定理[6]有关： $\bar{y}$ 的渐近分布仅取决于 y 的（自）协方差，而与 y的高阶矩无关。TIC 方法将这一性质用于主要可观测量[1]。

在标准格点分析中， C 要么设为数据的经验协方差，要么仅取其对角部分。然而，即使真实分布完全服从高斯分布，数据量有限时，我们也无法获得其真实协方差参数。这是本文将要探讨的问题之一。

有时会添加一项 L_2 正则化[7] ，并将其并入扩展后的 χ^2。这种做法常被不恰当地称为贝叶斯先验，因为它等价于带有高斯分布先验的贝叶斯模型的最大后验估计。此类正则化仍允许将待最小化的函数写成最小二乘形式，并使用 Levenberg-Marquardt 算法[8], ；而采用任意先验的最大后验估计则可能需要使用更通用的最小化器，这类最小化器通常速度较慢且稳定性较差。如今，随着机器学习领域强大优化方法的发展，这一限制已不那么重要。

最小二乘方法本身不包含误差估计，它只提供点估计，而传统的格点方法需要额外增加一层分析，才能将误差纳入考虑。

E. 使用 PyMC 进行抽样

要进行完整的贝叶斯分析，不仅需要得到点估计，还需要得到所有参数的完整后验分布 a。虽然有少数情况下可以通过解析方式写出该分布，但大多数模型并不支持。不过，对于 a 的任意取值，我们都能够计算相应的似然 $P(y|a)$，因此也能在一个常数因子的意义下得到后验分布 $P(a|y)$。蒙特卡洛技术可以按照该概率对 a 进行抽样，而不受该常数因子取值的影响。

市面上有一些用于这类贝叶斯推断的“标准”程序和软件包。我们选择以 PyMC[9], 为基础开展这项工作。PyMC 是一个 Python 软件包，只需几行代码即可构建并拟合模型。这些工具通常基于格点领域开发的一种算法：混合蒙特卡洛 [10]。

参数 a 与规范构型类似，而模型的对数似然类似于晶格作用量，边际似然则类似于配分函数。只要 a由连续参数构成，就能根据后验概率高效地对 a的取值进行采样，即使维度很高也不例外。事实上，即便是我们将要介绍的最复杂模型，以及那些难以进行最小化的模型，其维度也远低于 HMC 所针对的晶格计算中通常遇到的维度。力通过 Theano[11] 的自动微分计算得出。蒙特卡洛历史中 a 的所有取值集合称为轨迹。轨迹通常包含多个并行的蒙特卡洛链，用作遍历性的探针。

本文所示结果是在英特尔至强 Gold 6130 中央处理器的几个核心上获得的。实际上，这些计算也可以在笔记本电脑上运行；而对于最复杂的模型，图形处理器会是一个特别有趣的选择，这些库已经支持图形处理器。用于生成本文结果和图表的所有 Jupyter 笔记本均可在 [12] 获取。

III. 传统自助法拟合示例

在深入探讨贝叶斯方法的应用之前，我们先建立一个传统自助法最小二乘拟合的参考实现，采用如今通常使用的做法。为保持一致，我们展示通过 PyMC 实现获得的结果。不过，我们使用经过充分测试的基于 GSL 的 C++代码也得到了定性上相似的结果，因此最小化器的技术选择似乎并非关键。在本文中，我们的数据始终是来自 CLS 系综 H101 的 1009 个构型的真实 π介子关联函数。为保持一致，我们沿用这个特定示例，但我们的论点适用于任何类型的拟合。

[1] 公式（11） [2] 本质上就是 [6]的公式（27.20）

这个示例将仅说明最小二乘拟合可能失败的诸多情形之一。对于非线性函数，原则上可能发生四种情况。第一种失败发生在最大似然估计的峰值位于全局最大值处，但其邻域内的概率体积很小。第二种情况从理论角度来看几乎同样令人头疼：似然存在局部极大值，且这些局部极大值离最小化器的起始点相对较近。第三种情况则纯粹是数值不稳定性问题，取决于初始猜测、参数化和正则化项。最后，第四种情况是拟合结果看似合理，却无法提供可靠的拟合优度。本节的示例主要会遭遇第二种失败。

我们首先为数据增加一个新维度，用以表示自助法：

$$z_{b,i} = y_{\beta_b(i)} \tag{7}$$

其中，函数 β_b 是在 $[1,n]^{[1,n]}$ 中针对每个 b 随机抽取的（我们抽取了 $N_b = 200$ 个）。

使用带有 $a =$ （v， E）的公式 6，我们将拟合公式写为[2]

$$P(z|v,E) \propto \prod_{b=1}^{N_b} e^{-\frac{n}{2}\sum_t [\bar{z_b}(t) - f_t(v_b,E_b)]^2/\sigma^2(t)} \tag{8}$$

$$f_t(v_b,E_b) = \sum_{j=1}^{N_{\exp}} v_{bj} e^{-E_{bj}t}. \tag{9}$$

E 和 v 的最大似然估计现在都是一个 $N_b \times N_{\exp}$ 矩阵。我们称之为线性模型（关于 v 呈线性）；对于 $N_{\exp} = 3$，从图1中可以看到，这会给出非常可疑的结果。事实上，自助法声称能够以很高的精度确定 E_1 和 E_2，但它们实际上是退化的；与此同时， v 的方差却非常大。

在这种情况下，仔细考察便能理解为什么结果会被吸引到这一退化状态：退化线在似然中是一条长而平坦的方向，因为一个分量中的任何正贡献都可以由另一个分量中的负贡献抵消。我们甚至可以针对每个 E，解析地求出线性变量 v，并发现当趋于退化极限时，待求解的线性方程组会变成奇异的，因此局部极小值会被推向 v 的较大值。这条退化线可能非常狭窄，代表的积分概率也可能很小，但最小化器对此并不敏感：它只考察无穷小的体积。

因此，我们可以提出一种重新参数化方案，大体上解决我们的问题，尽管仍留下数量不可忽略的离群值：

$$g_t(w_b,E_b) = \sum_{j=1}^{N_{\exp}} w_{bj}^2 e^{-E_{bj}t}. \tag{10}$$

这种重新参数化等价于从半平坦先验（乘以一个雅可比行列式）中获得最大后验估计，并要求 v 为正；但幸运的是，在这个特定情形下，我们可以避免加入信息量更强的先验。如图1所示，这种重新参数化模型与先前的线性模型在自助法误差方面存在明显的不一致。不过，每自由度的 χ^2 要好得多，结果也更合理。如果现在以重新参数化模型所得结果给出的非常精确的估计作为起点，对线性模型逐次自助样本地重复进行最小化，结果将显著改善（但与贝叶斯结果之间仍会存在一些轻微的张力）。

总而言之，自助法拟合不能被当作处理复杂非线性模型的黑箱。最小化器的收敛本身并不能保证自助法误差可靠。必须针对具体情况评估结果，并据此采用临时解决方案，而这些方案可能难以令人信服。对第五节 B 中同一模型应用贝叶斯推断，将不太容易受到初始值以及（隐式或显式）先验的影响。

IV. 几个非常基础的玩具模型

A. 一维高斯

我们先从高斯模型 $p = 1$ 的情况开始。在这种情况下， y_i 只是实数，而 $a = (\mu, \sigma^2)$。例如， y_i可以是每个构型 i对应的 plaquette 值。

[2] 我们的 PyMC 模型还包含一个　　对能量进行排序的隐藏式重新参数化。也可以像 C++ code 中那样在之后排序

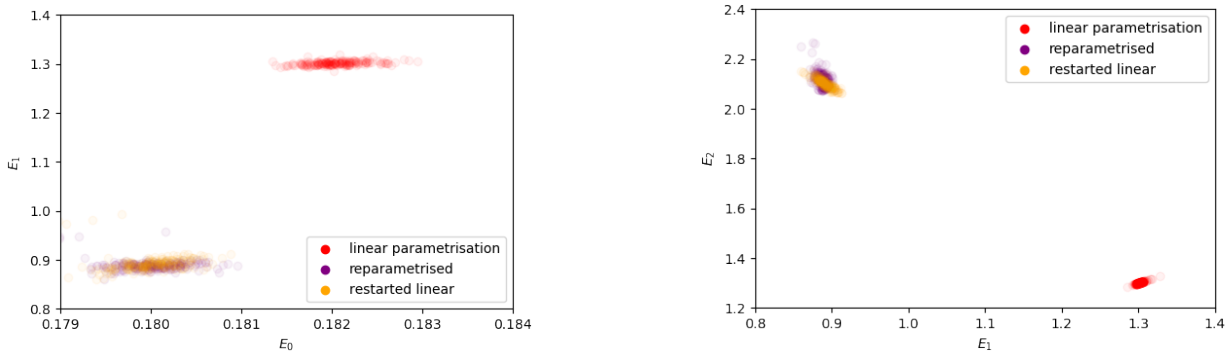


图1. 我们展示了三态非相关最小二乘拟合中两个状态的能量之间的相关性。每个点代表一个自助样本。重新启动后的（黄色）结果对应于线性（红色）模型，但其最小化过程从一个经过微调的初始猜测开始；该初始猜测来自更稳定的重新参数化（紫色）模型。结果可能会因实现细节和超参数的选择而发生显著变化。

我们对 μ, σ^2施加平坦先验。原则上，这只能通过让某个先验族趋近于非信息极限来实现；这里我们具体考虑的是，让 μ 对应的高斯分布越来越宽，以及为 σ^2采用非正规分布 $\Gamma(0,0)$（我们稍后会介绍这一对象）。实践中通常无需写出这些极限，在 PyMC 中我们只需设置

$$\log P(\mu|M) = 0 \tag{11}$$
$$\log P(\sigma^2|M) = 0. \tag{12}$$

定义我们模型的似然为

$$P(y|M,\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(y_i-\mu)^2}. \tag{13}$$

这是一个特例，其后验分布（公式 1）可以通过解析方法计算，并由众所周知的正态逆伽马分布表示为

$$P(\mu,\sigma^2|y,M) = NIG(\bar{y},\lambda,\alpha,\beta) \quad \text{where} \tag{14}$$
$$\lambda = n \tag{15}$$
$$\alpha = n/2 \tag{16}$$
$$\beta = \frac{1}{2}\sum_{i=1}^{n}(y_i-\bar{y})^2 \tag{17}$$
$$NIG(\bar{y},\lambda,\alpha,\beta) \equiv \frac{\sqrt{\lambda}}{\sqrt{2\pi\sigma^2}}\frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\left(-\frac{2\beta+\lambda(\mu-\bar{y})^2}{2\sigma^2}\right). \tag{18}$$

这是关于 μ 的一个高斯分布，其均值和最大后验估计为 $\bar{y}$。在 σ 方向上，情况则更复杂；在此阶段，我们不需要完全理解该分布的行为，但可以指出，我们知道它的均值和众数：

$$E(\sigma^2) = \frac{\beta}{\alpha-1} \to \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 \tag{19}$$
$$\sigma^2_{MLE} = \frac{\beta}{\alpha+3/2} \to \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2. \tag{20}$$

这个后验分布是我们进行贝叶斯推断的最终结果。我们不仅掌握了最大似然估计 ($\bar{y}$, σ^2_{MLE}) 的信息，包括在固定 σ^2 的情况下进行最小二乘拟合所得的 $\bar{y}$，还能够给出这些参数取值的置信水平区间，并了解它们彼此之间的相关性。

B. 平凡模型

让我们回到 σ 固定的情形，就像进行 χ^2 拟合时那样，不过现在有了 $p \neq 1$。具体来说，先来看这样一种情形：每个数据分量的均值各有一个拟合参数，即 $k = p$ 和 $f_i(a) = a_i$，采用公式 6 中的似然，并对 a 使用平坦先验。这就是我们所说的平凡模型，因为自由度为零，最大似然估计只不过是对数据的重复。

这里， a 的后验分布实际上是一个多元高斯分布，其均值为 $\bar{y}$，协方差为 C/n；其中， C 是进入似然函数的任意矩阵。在 n 较大的极限下，无论模型允许单个数据点的 y_i 波动幅度有多大， a 都将被完全确定。

在讨论信息准则或对拟合区间应用截断时，该模型将再次变得有趣。

C. 重新诠释最小二乘法

在公式 6 中，将 GLS 写成高斯似然时，我们使用参数 a 描述其均值，但协方差 C 被预先视为已知，即在执行贝叶斯推断或最小化之前就已知。然而，实际上我们对 C 的了解并不完备； C 是一个模型参数，与 a 中的其他参数一样，都是根据同一组数据估计得到的。在贝叶斯模型中强制将 C 视为完全已知，相当于设置一个狄拉克先验

$$P(C|M) = \delta(C - \langle (y - \bar{y})\,(y - \bar{y})^T \rangle) \tag{21}$$

这并不是一种很贝叶斯的思考方式：首先，在贝叶斯分析中，模型、先验和数据理想情况下应该是彼此独立的三项。其次， δ 函数作为先验所包含的信息实在过多：一旦先验设定，贝叶斯推断就被一个不正确 的猜测束缚住了，无论再输入多少新数据，都无法使 C 发生改变。当然，如果我们获得了一些新数据，就会改变先验；但这会发生在我们所建立的优美统计框架之外，因此会影响结果的概率解释。

对于平凡模型这一特例，经验数据协方差几乎就是 C 的最大似然估计（略去 $1/n$ 偏差）。在这种情况下，冻结 C 只会影响我们对误差条的估计。然而，一般情况下并不能保证这一点。特别是对于高维度的相关拟合，在机器精度范围内，数据协方差很容易成为非正定矩阵。如果维度大于样本数，基本代数甚至可以证明 它是奇异的，因此使用其广义逆的模型会导致零似然。

V. 贝叶斯模型

A. 威沙特分布

当我们转向先验较为宽松的完全贝叶斯分析时，必须决定先验的函数形式。事实证明，威沙特分布

$$\mathcal{W}(C^{-1}|V,\nu) = \frac{|C^{-1}|^{(\nu-p-1)/2} e^{-\operatorname{Tr}(V^{-1}C^{-1})/2}}{2^{\frac{\nu p}{2}} |V|^{\nu/2} \Gamma_p(\frac{\nu}{2})} \tag{22}$$

是一种特别有趣的选择，具有清晰的解释，且计算得以简化。这里注意， p 尺度矩阵的维度为 V，而 Γ_p 是多元 Γ 函数（不要与 Γ分布混淆）。

事实上，这是均值已知时高斯似然的共轭先验。这意味着，如果我们从威沙特先验开始，就会得到威沙特后验。如果从一个信息量极低的威沙特先验（$\nu \sim p$）开始并加入一些数据，就会得到信息量更高的威沙特分布，其中 $\nu' = \nu + n$ 会增加本轮推断所用数据点的数量 n，而 $\nu' V'$ 会趋近于根据这些数据计算出的经验逆协方差 C^{-1}。

在标量情形 $p = 1$ （例如第 IV A 节）中，威沙特分布称为 Γ 分布，其表达式同样为

$$\Gamma(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}. \tag{23}$$

它是一个推广 用于非整数自由度的 χ^2 分布（ α 或 ν承担这一作用） .

最后，在均值已知的高斯似然与 Wishart先验的情形下，后验预测分布也是已知的：它由多元 Student-t 分布给出

$$t_{\nu+n-p+1}\left(y \mid \mu, \frac{\left(V^{-1}+\sum_{i=1}^{n}(y_i-\mu)(y_i-\mu)^T\right)^{-1}}{\nu+n-p+1}\right). \tag{24}$$

对于 n 非常大的情况，Student-t 分布会收敛于高斯分布，这为第 IV C 节的模型提供了渐近理论上的依据。然而，两者存在两点差异：首先，尺度矩阵是根据拟合参数μ 计算的，而不是根据数据平均值 $\langle y\rangle$计算的。其次，对于有限的 n 值，该分布具有更宽的尾部。因此，由于 $\langle(y_i-\mu)(y_i-\mu)^T\rangle$ 并非完全等同于真实的 协方差矩阵，对离群值进行较弱的惩罚是合理的。

B. 不相关模型

现在我们将展示一个简单不相关模型的一些初步结果（即对角 C）。该模型应用于与第 III 节相同的数据，即一个 π介子关联函数。我们还尝试了其他系综和通道，结果相似。需要强调的是，我们的数据 y 按构型编号标记，似然是对各构型的乘积。这不同于自助法分析：在自助法分析中，我们首先对每个自助样本内的数据进行平均，然后在平均值空间中进行处理；但它与 Γ 方法中的涨落有些相似。

为求简便，我们选择高斯分布作为似然。不过，这并非理论上的限制：我们用若干任意的推广形式交叉检验了部分结果，甚至在第 B 节说明了如何构建非参数推断。只要我们持续研究关联函数（而非派生量），中心极限定理在一定程度上适用，并且拓扑扇区数量足够多，这种高斯近似就很可能效果良好。这绝不意味着后验分布必须是高斯分布[3]。无论如何，我们在此模型构建中所做的任何近似，都可以在根据第 VI 节比较模型时进行事后评估。

选定一些超参数 t_{min} 和 N_{exp}后，我们为标准差采用如下形式的先验

$$P(1/\sigma_t^2)=\Gamma(\alpha,\beta) \tag{25}$$

在定义域内，于公式 23 中任意选取 (α, β)

$$\alpha \ll n, \qquad \beta/\alpha=\frac{1}{n}\sum_{i=1}^{n}\left(y_i(t)-\bar{y}(t)\right)^2. \tag{26}$$

对于 $\alpha=\beta=0$，这是平坦的；对于如此简单的模型，实际上并不需要非零值。我们需要设定的第二部分是似然

$$P(y_{t>t_{min}}|v_j,E_j,\sigma_t)=\mathcal{N}(y_{t>t_{min}}|\sum_{j=1}^{N_{exp}}v_j e^{-E_j t},\sigma_t), \tag{27}$$

其中，对 v_j 和 E_j 设置任意的无信息先验，而似然还包含 E_j 按顺序排列这一约束（这是通过 PyMC 中的重新参数化实现的）。

定义模型后，我们可以根据分析需要，选择任意复杂程度的方法，从简单的最大后验估计到对后验分布进行完整抽样。抽样结果如图2所示。图3将这两种方法的结果与更传统的 χ^2 拟合进行了比较。基态的精度达到千分之一，而另外两个激发态似乎也得到了良好控制，尽管使用的是极其无信息的先验。轨迹对于确认抽样是否充分非常有用；与最大似然估计技术的收敛指标相比，它能提供大量信息。这里可以看到，较高激发态的抽样难度更大：链之间的一致性变差，自相关时间也变长。不过，重要的是，只需延长轨迹，就能系统地改善这一点。

---

[3] 如果我们的所有拟合参数都足够精确，可以进行泰勒展开，这通常意味着我们本应选择更复杂的模型，并且高阶项中会有信号

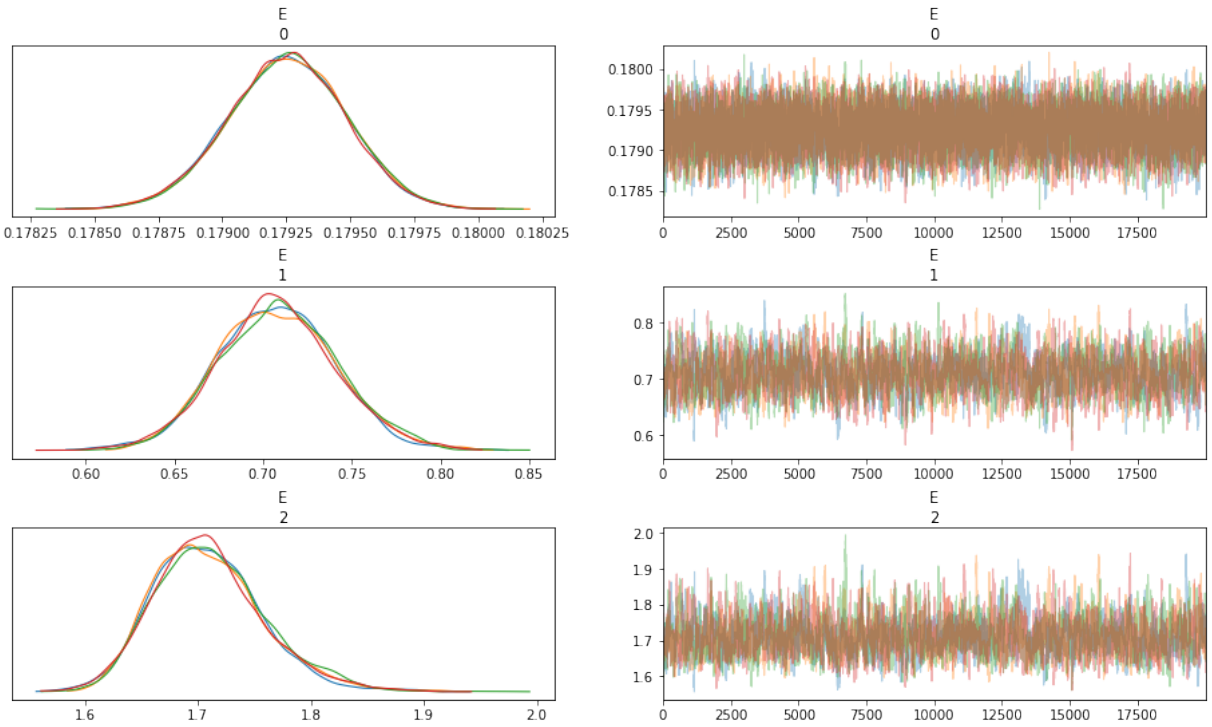


图2. 我们展示了能量变量不相关模型的 HMC 轨迹，每条链使用一种颜色。左侧表示每个参数的后验分布（横轴为参数值），右侧表示蒙特卡洛历史（横轴为轨迹编号）。不同链之间后验分布宽度的微小差异，可以看作“误差的误差”。

C. 具有时间相关性的模型

与第五节 B 类似，我们可以利用多元分布和第五节 A 中的 Wishart先验定义一个相关模型。PyMC 通过 Bartlett参数化实现了这一模型，但会使抽样维度增加 $O(p^2)$，从而带来速度变慢和不稳定的风险。因此，我们提出一种轻量级的解决方案：我们选择的先验允许对 C 变量进行解析边缘化。也就是说，对于给定的（v， E）取值，我们需要将所有（v， E， C）组合的后验概率相加，但 C 可以取任意值：

$$P(v, E|y, M) = \int \frac{P(y|v, E, C, M)P(v, E|M)P(C|M)}{P(y|M)} dC. \tag{28}$$

在对 C 进行积分时，描述均值的所有参数实际上都已作为似然的参数给出，因此已知，因此边缘化似然对应于式 24 的后验预测分布：

$$P(y_{t>t_{min}}|v, E) = t_{\nu+n-p+1}\left(y \mid \mu, \frac{\left(V^{-1} + \sum_{i=1}^{n}(y_i - \mu)(y_i - \mu)^T\right)^{-1}}{\nu + n - p + 1}\right). \tag{29}$$

其中，已知均值由其参数化形式代替

$$\mu(t) = \sum_{j=1}^{N_{exp}} v_j e^{-E_j t}. \tag{30}$$

图 4 展示了边缘化模型的一些抽样结果，图 5 则展示了极低统计量下的相同结果。

遗憾的是，不进行边缘化的模型无法以足够的遍历性和较低的自相关进行采样，[4]

[4] PyMC 中的 Wishart 和 Wishart-Bartlett 实现已知效率较低，文档建议改用 LKJ 先验。较新版本或其他软件可能效率更高。

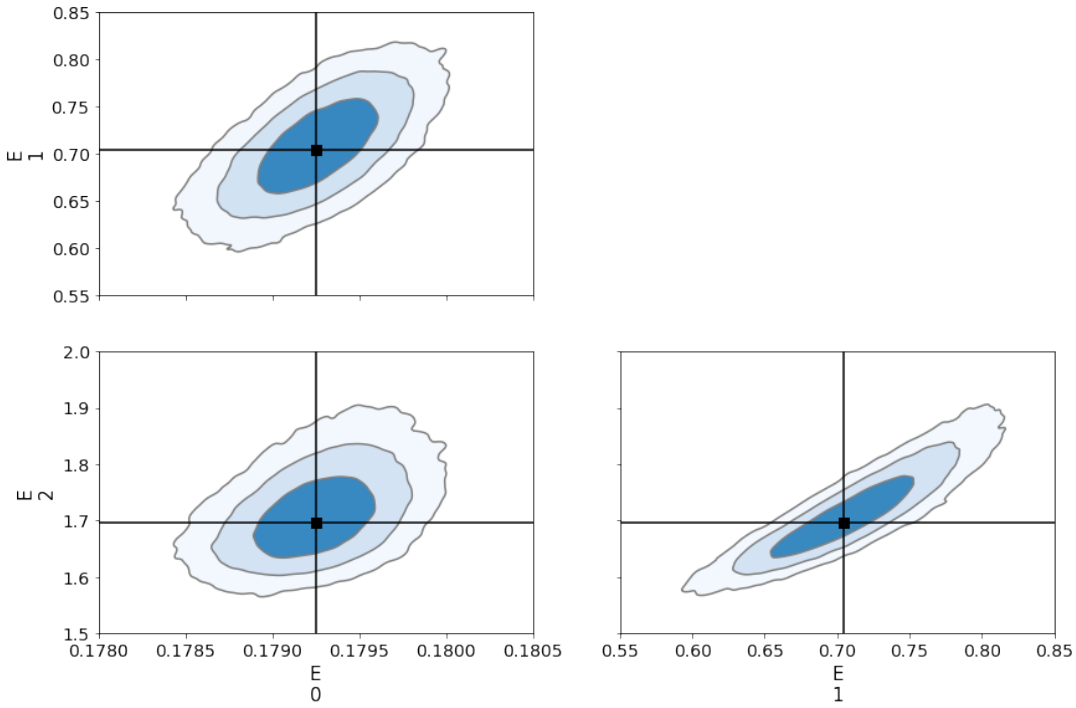


图3. 我们展示了在协方差未知且先验非常平坦的不相关模型中，各能量概率密度之间的相关性。黑色方块表示通过抽样确定的后验分布众数，而直接最小化则存在第三节中讨论的问题。图中显示了 68%、95% 和 99.7% 的等高线，分别类比于 1、2 和 3σ。

因此，我们无法明确验证这一直觉：使用完整似然（包括 χ^2 拟合未考虑的归一化因子）能够让我们的推断选择条件数更优、且 χ^2 更有意义的协方差矩阵。

D. 具有自相关的模型

如果我们希望贝叶斯分析能够取代自助法和 Γ 方法，就需要在我们的模型中为自相关数据提供一定支持。这类问题称为时间序列分析。

分箱始终是一种可行的选择，我们将在 E 节讨论，但本节提出一种更直接的解决方案。从理论上说，我们对 HMC 作为马尔可夫过程的谱知之甚少，因此很难构建出有效的模型。不过，要求模型在较长距离上具有一定的自相关指数衰减，并包含若干模态以区分 τ_{int} 和 τ_{exp}，这是合理的。自回归 AR(r)模型正好能够做到这一点：给定一个正态分布的创新项 $\xi = \mathcal{N}(0, \sigma)$，我们定义

$$y(\tau) = \rho_0 + \sum_{i=1}^{r} \rho_i y(\tau - i) + \xi, \tag{31}$$

其中，τ 表示 HMC 时间，以区别于欧几里得时间 t。

该模型中 y 的协方差可以写成 ρ 元素的函数，而将这一递推关系截断到给定阶数，类似于为 Γ 方法选择窗口。这需要计算带状对角矩阵的逆，因此即使 r 很小，τ中仍存在长程相关性。

图6展示了使用 $AR(4)$ 模型描述单个时间切片上关联函数的结果。图7比较了数据的自相关与后验预测分布。

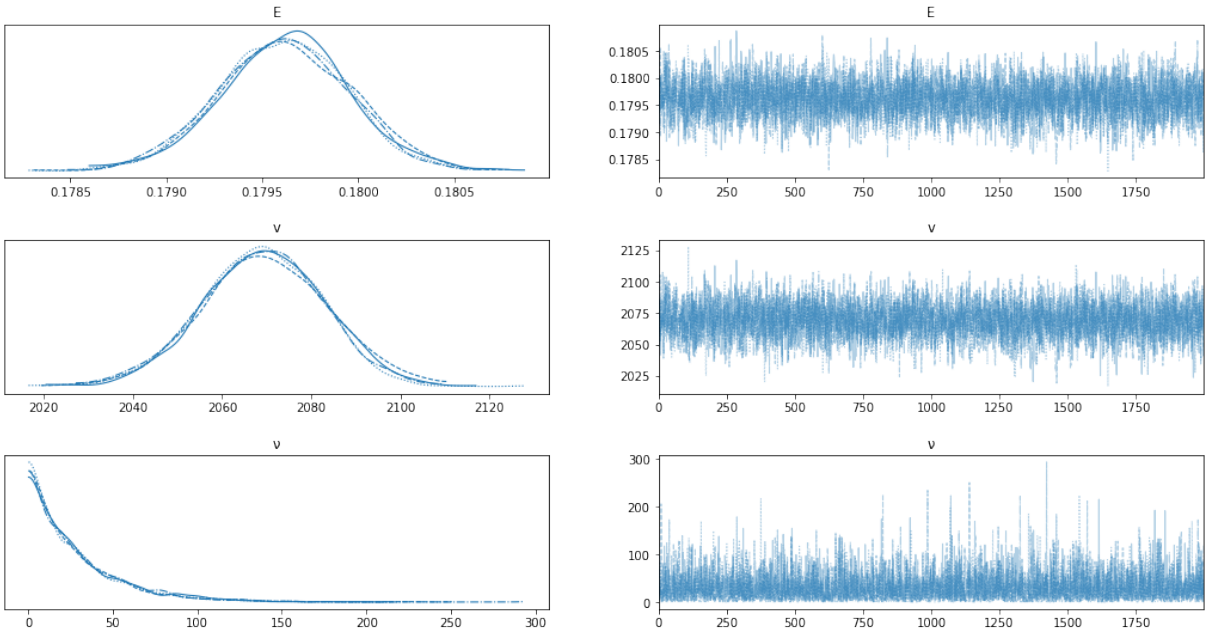


图 4. 展示了对采用边缘化协方差矩阵的 $N_{\text{exp}} = 1$ 相关模型中所有参数的抽样。不同线型分别表示 4 条链。 ν 的典型值远小于配置数量，因此选用了非常宽松的先验。请注意，可以通过这些抽样确定最大后验估计，但对于同一模型，我们直接计算最大后验估计却失败了。不进行边缘化的同一模型也往往会失败。

E. 对完整模型的展望

由于我们希望同时考虑 HMC 时间和欧几里得时间上的相关性，因此面临维数灾难。向量自回归（VAR）模型是一种合理的描述方式，其中 ξ 是多元的，且每个 ρ 参数都是一个 $p \times p$ 变量，但这种方式需要 $O(p^2 r)$ 个参数，不太可能符合实际情况。

因此，需要进行一定程度的近似，这实质上就是欠拟合与过拟合的权衡，并应根据第 VI 节的方法进行评估。最合适的模型取决于各个数据集，因此这里不再详细讨论。不过，我们在图 8 中展示了一个非常简单的向量 AR(1) 模型的一些结果，该模型假设其矩阵为常数。

请注意，类似的问题也存 在于 Γ 方法中，其中计算包含自相关的协方差需要
将.....考虑在内（对每个 矩阵元采用不同处理）很困难，因此人们普遍采用非相关拟合。

VI. 信息准则与模型平均

A. 从赤池到渡边-赤池

赤池准则[13]

$$AIC = -2 \log \mathcal{L}(a_{MLE}) + 2k \tag{32}$$

在格点领域中广受欢迎[4], ，因为它可以与相关的 χ^2联系起来。随后将其转化为权重

$$w(M) = \frac{e^{-AIC/2}}{\sum_{M'} e^{-AIC(M')/2}} \tag{33}$$

其渐近地描述了模型 M 的边缘概率。所有模型共有的参数 a_c 将按照以下方式分布：

$$P(a_c) = \sum_M w(m) P(a_c|M). \tag{34}$$

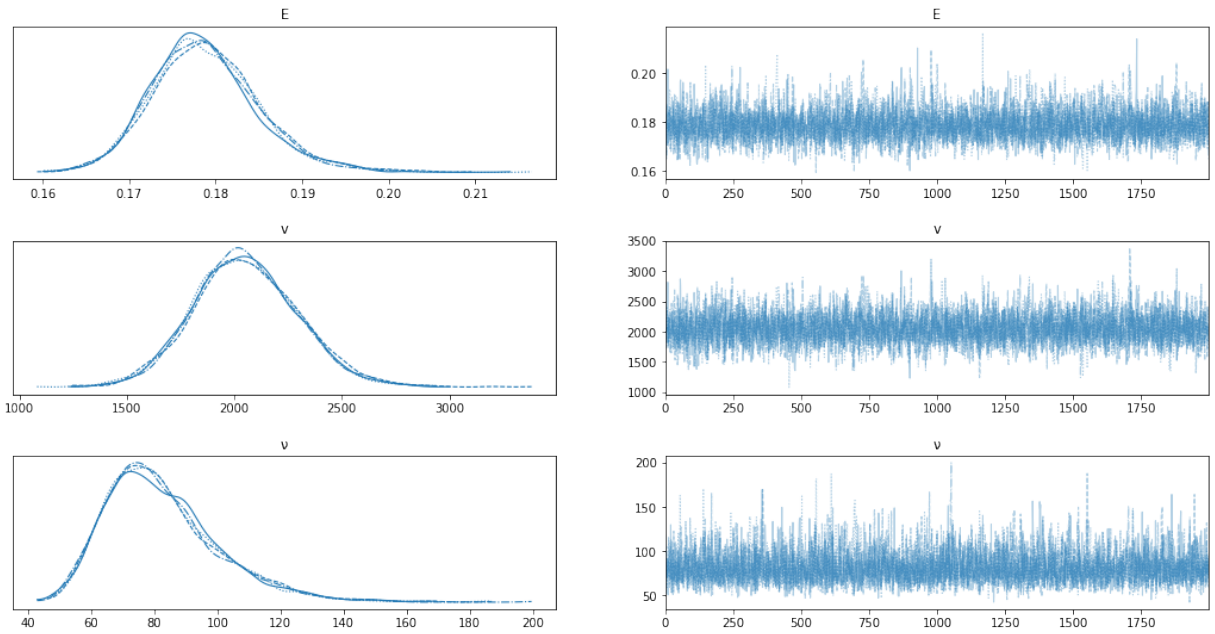


图 5. 有趣的是，仅使用 9 个构型进行抽样也完全可行。在这种情况下，我们甚至可以计算最大后验估计。相比之下，众所周知，对于如此小的统计样本，数据经验协方差矩阵的特征值至少有一半在代数上为零，因此无法进行 χ^2 拟合。

这有时被称为伪贝叶斯模型平均，因为它使用了围绕最大似然估计的近似，而真正的贝叶斯模型平均应当使用某些边缘概率作为权重。

渡边-赤池[14]或广泛适用信息准则（WAIC）是一种贝叶斯推广，它将完整的后验分布纳入考量，如下所示

$$WAIC(M) = -\sum_{i=1}^{n} \log P(y_i'|y, M) + k_{WAIC}, \tag{35}$$

其中，有效参数个数[5]

$$k_{WAIC} = \sum_{i=1}^{n} \left\{ \int \left(\log P(y_i'|a, M)\right)^2 P(a|y, M)da - \left(\int \log P(y_i'|a, M) P(a|y, M)da \right)^2 \right\} \tag{36}$$

在赤池信息准则[6]的有效范围内趋于 k 。

这一准则几乎适用于任何模型，对模型与真实模型有多接近的假设要求很弱；而赤池信息准则则要求模型处于真实模型的邻域内。例如，我们可以比较相关拟合与非相关拟合的 WAIC 值。

虽然采用这些信息准则似乎让我们彻底放弃了拟合优度这一概念，但值得注意的是，传统条件 $\chi^2 \leq p - k$ 与不等式 $AIC_{\chi^2} \leq AIC_{trivial}$ 严格等价；该不等式描述了第 IV C 节的 χ^2 模型是否比第 IV B 节的平凡模型做出更好的预测。

[5] 这是在 PyMC 中 p_{WAIC2} 的实现，与 Watanabe 最初提出的版本略有不同 p_{WAIC1} [6] 请注意，AIC 和 WAIC 的定义约定在 [14] 和 [15]中相差一个因子 2 -2

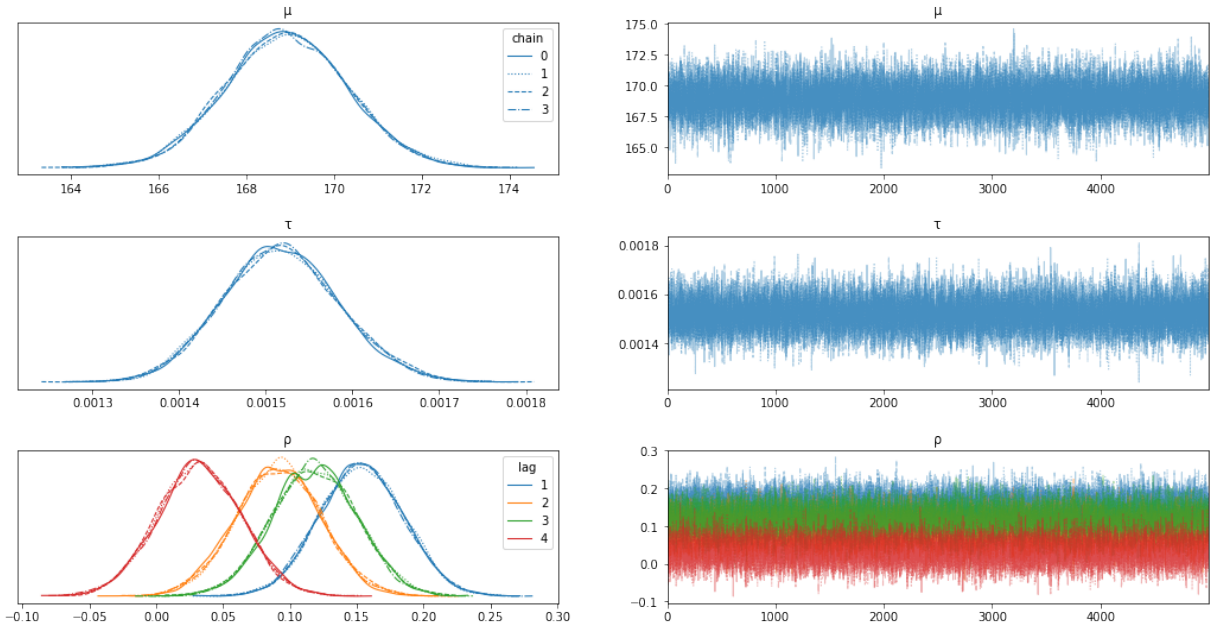


图6. 这个 $AR(4)$ 模型给出了该时间片关联函数期望值的结果 $\mu = 168.9(1.4)$。若忽略自相关，朴素估计会低估误差，将其估为 $\mu = 168.9(8)$。加入高阶项后，信号仍保持稳定，但这些高阶项与零值基本相容。

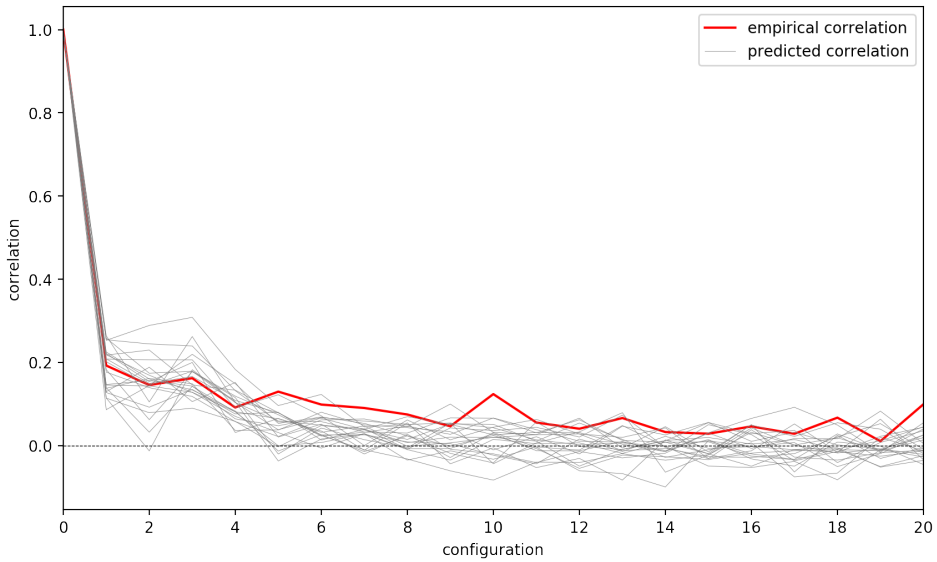


图7. 使用采样参数从这个 $AR(4)$ 模型生成预测值后，我们可以检查这些预测值的自相关是否与数据的自相关非常相似。结合 ρ 的取值，这证实了阶数已经足够高。

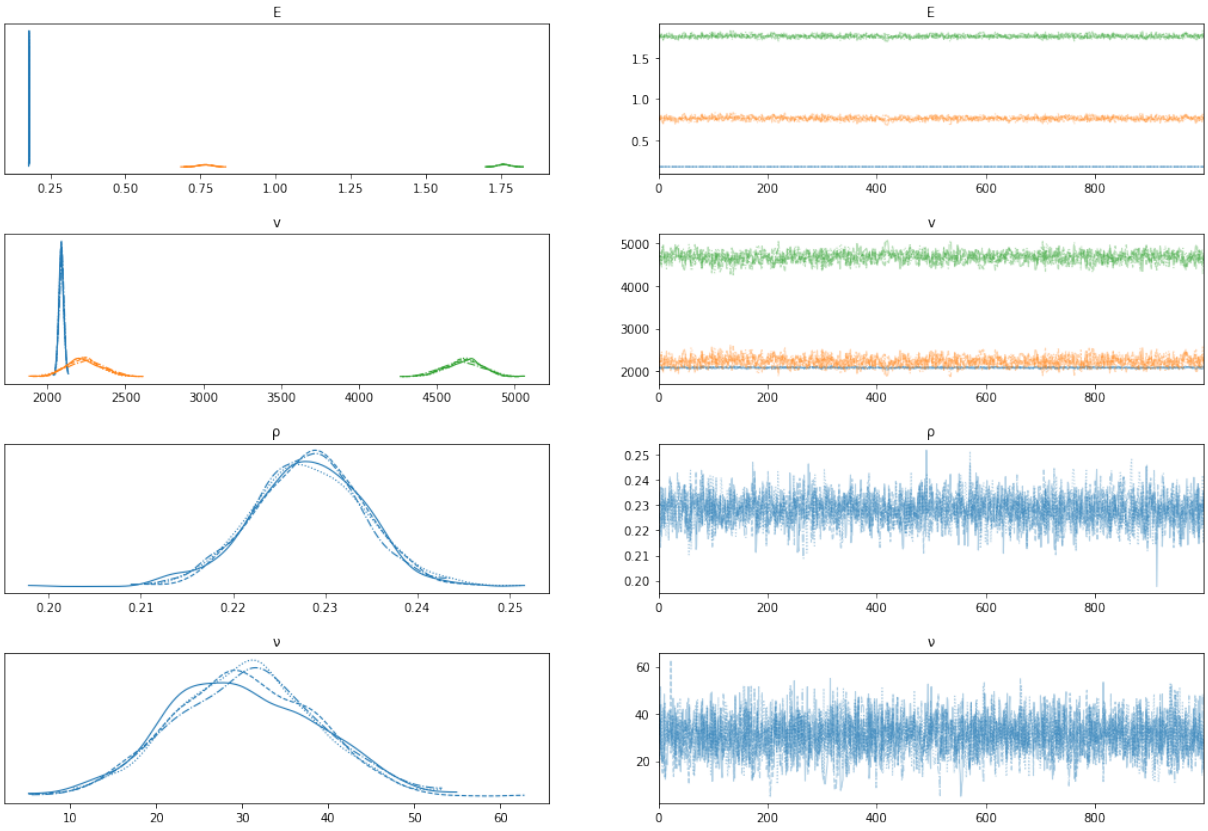


图 8。该模型与图 4 类似，但基于一个考虑了某些自相关的 AR(1) 模型。因此， E_0 上的误差棒保守约 20 %，这与滞后参数 ρ 明显不为零相一致。

B. 交叉验证与期望对数预测密度

广泛适用信息准则恰好与另一个估计量非常接近；该估计量基于留一交叉验证（留一法）[15]构造。实际上，它们都是对期望对数预测分布（期望对数预测密度）的近似

$$elpd = \sum_{i=1}^{n} \int P(y_i'|M^*) \log P(y_i'|y, M) dy_i' \tag{37}$$

其中， M^* 代表生成这些数据的真实 模型。在 PyMC 中，留一法通过对后验分布进行重要性抽样（也称重新加权）来实现，因此无需针对训练数据与测试数据之间的每一种划分都重新生成一条 HMC 链。

期望对数预测密度本身与交叉熵有关，因此，最大化期望对数预测密度也等同于最小化从 M 到 M^*的 Kullback–Leibler 散度

$$D_{KL}(M^*||M) = \sum_{i=1}^{n} \int P(y_i'|M^*) \log \frac{P(y_i'|y, M^*)}{P(y_i'|y, M)} dy_i'. \tag{38}$$

最后，我们来考虑这样一种情况：真实模型已知是一个协方差已知的多元正态分布（例如，因为它是合成数据），而我们的模型采用非相关拟合：此时期望对数预测密度与 [16]的预期 χ^2 非常相似。区别在于，期望对数预测密度是对整个后验（预测）分布取平均，而不是只考察与最大似然估计的差异。

C. 分类变量与混合模型

原则上，贝叶斯模型可以包含任意复杂程度，因此，贝叶斯模型平均的一种有趣等效做法是编写一个足够通用的单一模型，使其涵盖我们所有的模型。一种方法是添加一些分类变量，用来控制采用哪种模型。例如

$$P(t_{min}) = \mathbb{1}_{t_0<t_{min}<t_1} \tag{39}$$
$$P(y_{t<t_{min}}|\mu_t, \sigma_t) = \mathcal{N}(\mu_t, \sigma_t) \tag{40}$$
$$P(y_{t\geq t_{min}}|v, E, \sigma_t) = \mathcal{N}(\sum_{j=1}^{N_{exp}} v_j e^{-E_j t}, \sigma_t). \tag{41}$$

后验概率 在轨迹中得到的 $P(t_{min} \mid y)$，可以直接解释为贝叶斯模型平均权重 .

不过请注意，显式使用分类变量往往会使 HMC 不稳定，因此通常最好对它们进行边缘化（对于有限集合来说，这很简单）。这样还可以确保在每一步抽样中， t_{min} 的所有取值都会有所贡献，即使是最罕见的取值也不例外。此时，模型表示为一个混合

$$P(y|\mu, v, E, \sigma) \propto \sum_{t_{min}} \prod_{i=1}^{n} \left\{ \left[\prod_{t<t_{min}} P(y_{it}|\mu_t, \sigma_t) \right] \left[\prod_{t\geq t_{min}} P(y_{it}|v, E, \sigma_t) \right] \right\}. \tag{42}$$

我们还选择在 我们的层次模型中增加一层，采用任意的（例如 $\alpha = 1$）狄利克雷超先验 r

$$P(w) = \mathcal{D}(\alpha, \cdots, \alpha) \tag{43}$$
$$P(t_{min}) = w_{t_{min}}. \tag{44}$$

一旦 t_{min} 完成边缘化，仍然可以对连续变量 w 进行轨迹分析。

如果不对 t_{min} 或 w进行边缘化，而是先对其他所有参数 μ_t、v_j、E_j、σ_t进行边缘化，则与 wBMA b 变得完全 ifest. The only diff 等价 is th 的表达式 licit 存在 o f a w pr ior instead of an implic 它 fl at pri or。非边缘化版本和边缘化版本的一些结果分别展示在图 9 和图 10 中；考虑到实际问题，两者应给出相同的结果。两种情况下，此处组合的模型均源自图2中的不相关模型，其中引入了 t_{min} 截断。

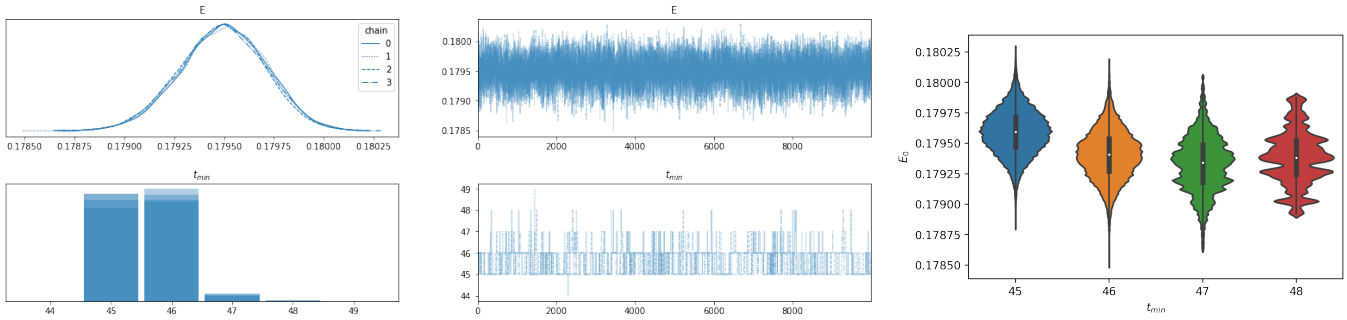


图1。非边缘化模型更难抽样，但能提供直接的解释，其中 t_{min} 的后验概率对应于贝叶斯模型平均权重。如右侧的小提琴图所示，也可以在事后分别查看每个 t_{min}的子样本。我们看到（与事后对单独的抽样或拟合进行模型平均时不同），采样器会花更多时间精确描述实际重要的模型，这可能是好事，也可能是坏事。

D. 关于贝叶斯因子的说明

原则上，可以直接计算模型的边缘概率，即对公式 1 的分母进行积分。这样就无需使用信息准则，尽管已有研究表明，它在渐近意义上与 BIC 等价。两个模型的相对重要性由贝叶斯因子给出

$$K(M_1, M_2) = \frac{P(y|M_1)}{P(y|M_2)} \tag{45}$$

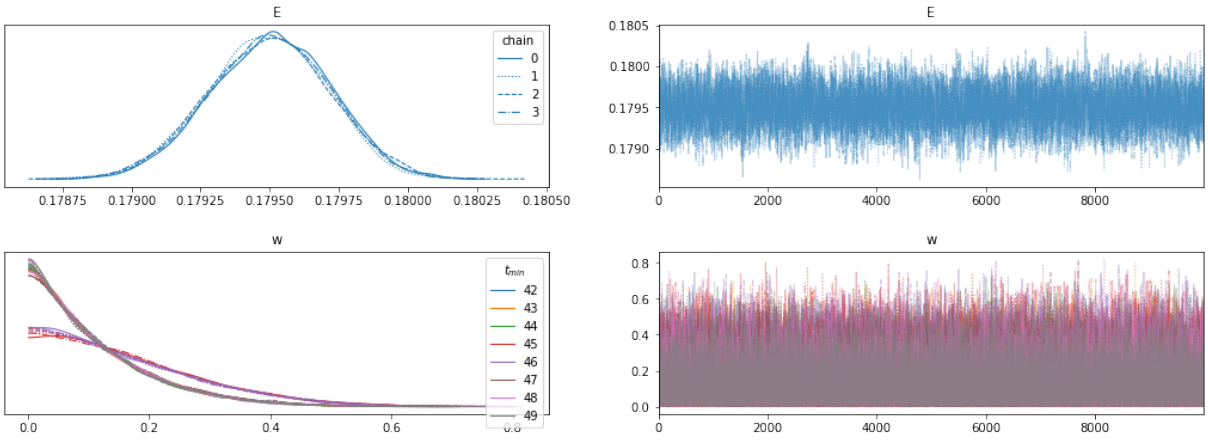


图 10. 这个边缘化模型为我们留下了连续变量 w，它们仍然能提供关于 t_{min} 的信息，但解读起来更加微妙。特别是，我们看到 w 对 $t_{min} = 45$ 和 46 赋予了稍高的权重。无论是就 E 还是 w而言，结果都与图 9 非常相似。

这种方法的计算量可能很大，稳定性也较差，但如图 11 所示，可以在 PyMC 中实现。它没有使用 PyMC 常用的 HMC 变体 NUTS，而是采用一种更复杂的过程——序贯蒙特卡罗（SMC），分多个步骤来解决问题。需要注意的是，与先验预测分布相关的边缘概率原则上可能对先验的选择非常敏感。我们已确认，在这个具体示例中并非如此。

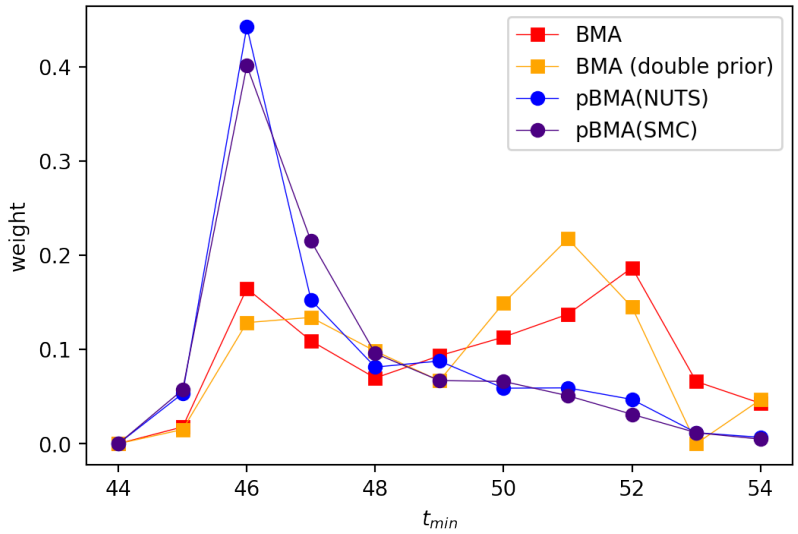


图 11. 对于与图 10 相同的问题，我们现在直接计算由序贯蒙特卡洛（SMC）计算得出的边缘似然所对应的 BMA 权重。我们还展示了与 pBMA 权重（根据广泛适用信息准则计算）的比较。两者是不同的量，因为理论上边际似然渐近地给出 BIC，而广泛适用信息准则渐近地给出赤池信息准则。我们还比较了两种先验选择。

VII. 错设模型

A. 协方差的渐近知识

我们在第 IV C 节讨论了将 χ^2 拟合作为贝叶斯模型时其表现有多差，但在第 V A 节表明，如果对平均值的建模接近真实情况，它在渐近意义上与一个定义良好的模型相似。随后，我们在第 VI 节介绍了一种建立在一些非常弱的假设之上的信息准则。这种广泛适用信息准则甚至可以应用于设定错误的模型，而 AIC 则不能。这使我们能够根据经验检验广泛适用信息准则向 AIC 的收敛情况，并且还可以基于广泛适用信息准则比较 χ^2 拟合与完整的贝叶斯模型。

对于复杂模型，确定 AIC 可能很困难，因为最小化过程比抽样更加不稳定；然而，比较 k_{WAIC}与 k通常更容易。根据我们的经验，对于 χ^2 拟合， k_{WAIC}似乎始终与k 非常接近，即使在统计量较低的情况下也是如此。这可能与第 A 节的讨论有关。因此，按照这一准则， χ^2 模型在渐近意义下的微小错设似乎不会产生很大影响。然而，在将广泛适用信息准则与完整的贝叶斯模型进行比较时，在统计量很大的极限情况下， χ^2 模型会被强烈排除。

在图12（分别在图13）中，我们展示了当改变作用于协方差的先验强度时，信息准则（分别是有效参数个数）如何变化。我们还研究了配置数量变化时的类似情况。在一个模型中，我们定义超先验 $P(\nu) = \exp(-\nu/\nu_{\text{prior}})$，其中 ν 是第五节 B 中模型的 Wishart参数。在另一个模型中，我们将 ν 精确固定为 ν_{prior}。尺度参数 V 的选取使得 Wishart先验以数据的经验协方差为中心。当 ν_{prior} 较小时，先验趋于平坦，因此这一额外的自由度使模型能够参数化真实分布。另一方面，当 $\nu_{\text{prior}} \to \infty$ 时，模型趋向于第 IV C 节中的 χ^2 模型，而该模型设定有误。

请注意，对于这样的层次模型，其参数具体是什么并非一目了然。对于赤池信息准则，我们只计算经过最小化的自由度，不包括那些已被积分掉的自由度。

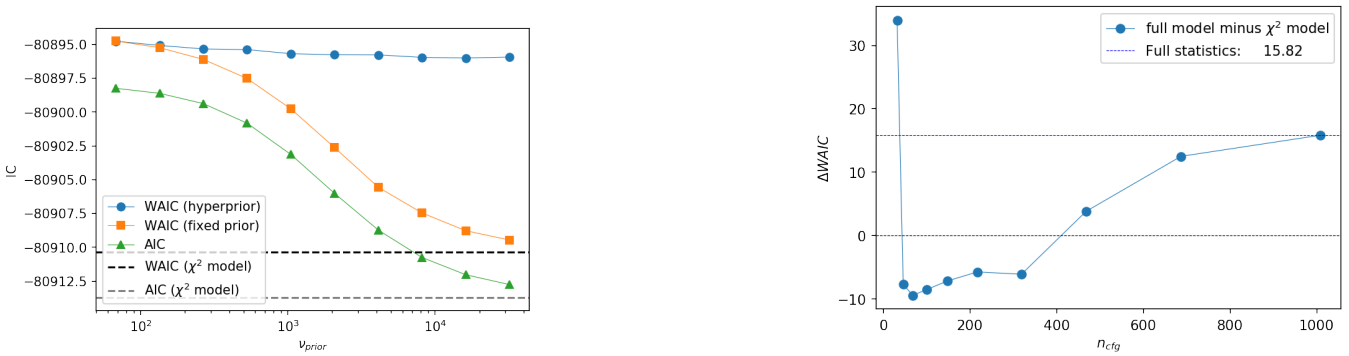


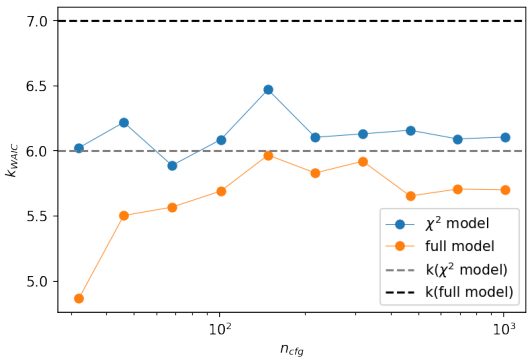
图12。我们展示了贝叶斯模型的信息准则：使用第VA节中的先验时，信息准则随该先验强度的变化（左图，完整统计数据）或随分析中包含的构型数量的变化（右图，带有 $\nu_{\text{prior}} = \sqrt{n_{\text{cfg}}}$的超先验）而变化。先验中的尺度矩阵设为经验协方差。 χ^2 模型（ν 固定为 ∞）作为比较对象一并展示。对于足够大的统计量，WAIC更倾向于较宽松的先验（超先验也会选择这种先验），因此它强烈排除了 χ^2 模型，在该模型中 C 完全由先验固定。

B. 信息准则与有偏协方差

让我们考虑这样一种情况：我们知道实际的协方差矩阵 C，但不希望在模型中使用它，例如因为稳定性问题。我们决定使用带有任意权重W的高斯模型（ χ^2 拟合）来拟合数据。这样一来，模型在 W方面设定错误，但与此同时，我们假设其均值参数化仍然正确，并且其真实值会渐近达到。

This 通常发生在对相关数据执行非相关拟合（对角 W）时 a.

在这种情况下，AIC 不适用。不过，WAIC 适用。本节将介绍竹内信息准则[17] (TIC)，这是另一种适用性居中的准则。它在渐近意义上等价于 WAIC，但和 AIC 一样，可用于最大似然估计。TIC 的另一个优点是更容易推导出简洁的公式。我们从 [14]中用英文给出的公式出发，写出有效参数个数



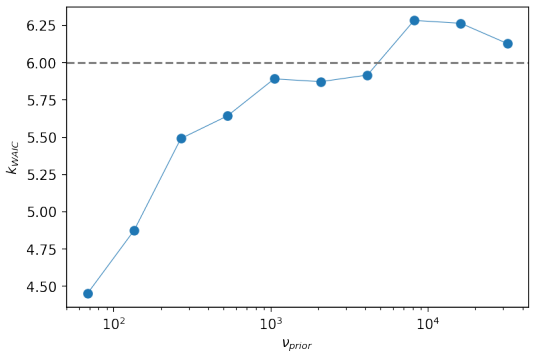


图13. 我们展示了广泛适用信息准则的有效自由度，它随纳入分析的构型数量（左）或先验强度（右，100 个构型）变化。对于较宽松的 ν 先验且统计量较低的情况，有效参数个数往往会减少。这通常意味着，对于某些参数，我们没有信号，除了先验之外也没有学到任何新信息。

的

$$k_{TIC} = \mathrm{Tr}\left(I(a_{MLE})J^{-1}(a_{MLE})\right) \tag{46}$$

$$I(a) = \int \nabla \log P(y \mid a, M) \nabla \log P(y \mid a, M) P(y \mid M^*) dy \tag{47}$$

$$J(a) = -\int \nabla^2 \log P(y \mid a, M) P(y \mid M^*) dy. \tag{48}$$

本节开头列出的假设可转化为

$$P(y \mid a, M) = \frac{\det(W^\dagger W)^{1/2}}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}||W(y - f(a))||^2\right) \tag{49}$$

$$P(y \mid M^*) = \frac{\det(C)^{-1/2}}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}(y-\mu)^T C^{-1}(y-\mu)\right). \tag{50}$$

随后可以解析地求出高斯积分，发现若 $|f(a_{MLE}) - \mu| = O(1/\sqrt{n})$可以忽略，则

$$k_{TIC} \simeq \mathrm{Tr}\left[(G^\dagger C_W G)(G^\dagger G)^{-1}\right] = \mathrm{Tr}\left[\mathcal{P} C_W\right], \tag{51}$$

其中， $C_W = WCW^\dagger$, $F_{i\alpha} = \partial f_i / \partial a_\alpha (i \le p, \ \alpha \le k)$是拟合函数的 $n \times k$ 雅可比矩阵， $G = WF$而 $\mathcal{P} = G(G^\dagger G)^{-1} G^\dagger$ 是一个投影算子。这些量也出现在期望卡方中[16],，并且 TIC 也可以写成以下形式（相差一个与 f无关的常数）：

$$TIC = \chi^2_{MLE} - 2E(\chi^2 | M^*). \tag{52}$$

如果 $W = C^{-1/2}$ ，那么我们的模型设定良好，并且可得出 AIC，即 $k_{TIC} = k$。如果 $W_{ij} = (C_{ii})^{-1/2}\delta_{ij}$, 且数据具有很强的相关性，那么 $k_{TIC} \to p$，这弥补了不相关模型往往会低估进入信息准则第一项的 χ^2 这一事实。我们通过将式（51）与 k_{WAIC} 的完整数值计算结果进行比较，交叉验证了它的一致性（误差在几个百分点以内）。

第三种有趣的情况是 $W = \sqrt{\beta} C^{-1/2}$，其中 β 可以解释为逆温度。温度不会改变 MLE 的位置，但会改变似然在最大值处的取值。它还会给出$k_{TIC} = \beta k$。在零温度下，系统冻结在其 MLE 处，模型会因k_{TIC}增加而受到惩罚：似然中的统计涨落会被放大，从而导致过拟合。在高温下，后验分布无法从拟合参数中学到任何信息，因此对数似然会分散到参数的许多取值上，这同样会导致 TIC（或广泛适用信息准则）的取值不佳。表 I 展示了在 $W_{ij} = \sqrt{\beta}(C_{ii})^{-1/2}\delta_{ij}$ 中取不同 β 值时，AIC 与 WAIC 在一个不相关模型上的比较。

对于带平坦先验的常数拟合（以及可能任何参数呈线性的拟合函数），在 n 的大极限下， k_{WAIC} 也适用同一公式，如 A 节所示。

模型	Δ_{WAIC}	k_{WAIC}	赤池信息准则	广泛适用信息准则	E
$\beta = 1$			16.79	8.2	0.17938(28)
$\beta = 0.5$	1651		8.25	3.9	0.17936(39)
$\beta = 2$	2642		32.62	16.2	0.17938(20)
$\beta = 0.1259$	10269		2.04	0.81	0.17937(78)

表 I。对于从 $t_{src} + 13$ 开始的单指数非相关 χ^2 模型，我们比较了若干模型。 $\beta = 1$ 是标准的贝叶斯推断，尽管强相关数据带来了较大的 k_{WAIC}，但按广泛适用信息准则衡量，它仍明显更受青睐。低温情形 $\beta = 2$ 使结果更接近最大似然估计，误差条也更小，但所得模型的预测能力实际上更低。情形 $\beta = 0.1259$ 经调整后得到 $k_{TIC} = k$ （因此 β 补偿了我们舍弃 C 的非对角元素所造成的影响），但这同样未能得到良好的模型，因为过度保守会受到惩罚。由于这些模型在 C 中设定有误，广泛适用信息准则与赤池信息准则在此并不一致。 E 中中心值的微小差异，主要来自可通过系统方法改进的蒙特卡洛误差。

VIII. 结论

我们提出了一种完全贝叶斯框架，并将其应用于π介子关联函数，以开展完整分析，无需任何额外要素或层次。在这项概念验证中，我们选择了一个相对简单的问题。然而，事实证明，该框架优于经典方法：我们能够以很高的精度可靠地拟合三个态，无需舍弃任何早期时间数据；也能够使用9个构型进行相关拟合。我们新方法的优越性及其灵活性带来的益处，在更复杂的问题中可能会更加显著，而这正是我们计划在未来研究的方向。

不过，这种方法旨在使建模假设更加明确且便于评估，而非自动解决我们面临的所有挑战。因此，应针对具体情况构建并评估模型，以正确描述数据。我们介绍了几种处理格点量子色动力学数据中常见问题的方法，尤其是欧几里得时间相关性和蒙特卡洛时间相关性。当多种这类复杂因素同时出现时，仍需进一步研究，以确定适用于不同情形的良好模型。

我们讨论了评估模型时可用的一些工具，包括基于贝叶斯方法和最大似然估计的信息准则，并介绍了几种进行模型平均的方法。

随着我们编写的代码公开，我们欢迎读者用自己的数据尝试我们的模型，或构建更复杂的模型。一旦从定义明确的贝叶斯问题出发，同样的模型也可以用于多个中间层次的近似。我们使用的 PyMC 套件已经实现了其中许多方法；根据实际限制和问题的复杂程度，选择其中一种或另一种都可能是合理的。将这些方法集成在一个工具箱中当然会很方便：层次模型、高斯似然、变分推断、最大后验估计附近的正态近似，.....

附录 A：常数拟合的广泛适用信息准则

常数拟合是公式 50 的一种特殊情形，其中 $f(a) = aE$，其中 $E = (1, \cdots, 1)$。有效参数个数 k_{WAIC} 随后由以下量的方差给出

$$\log\left(P(y_i|a)\right) = -\frac{1}{2}||W(y_i - aE)||^2 \tag{A1}$$

根据 a 的后验分布。

采用平坦先验时，可以通过解析计算得到该分布

$$P(a|y) \propto P(y|a) \propto \prod_i \exp\left(-\frac{1}{2}||W(y_i - aE)||^2\right) \tag{A2}$$

$$\propto \exp\left(-\frac{|WE||^2}{2n}\left(a - \frac{\langle W\bar{y} \mid WE\rangle}{||WE||^2}\right)^2\right), \tag{A3}$$

其中， $\bar{y}$ 表示所有向量 y_i的均值。

接着，我们围绕该后验分布的均值展开对数似然，去掉其中关于 a 不具有方差的前导项，以及渐近可忽略的 $(a-\bar{a})^2$ 项：

$$\log\left(P(y_i|a)\right) \simeq \text{cst} + \langle W y_i - \frac{\langle W\bar{y} \mid WE\rangle}{||WE||^2} WE \mid aWE - \frac{\langle W\bar{y} \mid WE\rangle}{||WE||^2} E\rangle. \tag{A4}$$

由于我们知道 a 的方差，这现在归结为

$$V\left(\log\left[P(y_i|a)\right] \mid M, y\right) \simeq \frac{\langle W(y_i-\bar{y}) \mid WE\rangle^2}{n||WE||^2}, \tag{A5}$$

a最后对 i 求和后，如果在本节中 C 是数据的经验协方差 ，

$$k_{WAIC} \simeq \frac{\langle WE \mid C_W \mid WE\rangle}{\langle WE \mid WE\rangle}, \tag{A6}$$

这几乎与公式 51 完全一致。

附录 B：贝叶斯自助法

我们方法的一个明显局限是需要提供一个噪声的参数模型。实际上，这个模型不仅可以任意复杂，而且还存在非参数模型。我们将特别介绍贝叶斯自助法（BB）[18], ，并不是因为我们提倡使用它，而是因为它与格点量子场论（LQFT）领域中通常使用的经典自助法[1] 相似。

让我们设想一下，我们只对一个数感兴趣，例如某个 plaquette，或固定时间处的一个关联函数。我们可以将一个简单的模型写成如下形式：

$$P(w) = \mathcal{D}(1, 1, \cdots, 1) \tag{B1}$$

$$P(y) = \sum_i w_i \delta(y - y_i), \tag{B2}$$

其中， $\mathcal{D}$ 是用作 w先验的狄利克雷分布。这个公式意味着，我们的似然赋予w_i 这样的概率：观测到与数据中已观测值相似的值；而对任何其他值，概率为零。这类似于以概率 $1/n$ 抽取 y_i 来构建经典自助样本，只不过给予 w 更多自由度，从而得到更平滑的分布。

请注意，这个模型可以看作一个与数据无关的模型的边缘化版本：该模型为 y的所有取值赋予权重，并对所有未观测值进行简单积分。

难点在于，如何用一个将贝叶斯自助法与参数模型结合起来的模型，取代我们通常使用的自助法拟合，同时显式获得与物理相关的量。确实，这里存在某种矛盾：一方面，贝叶斯自助法拒绝进行任何泛化，也拒绝在似然中纳入任何已观测值之外的取值；另一方面，我们希望采用一种能够对新系统作出物理预测的参数化。这并非新问题，因为经典自助法的结果通常被用于基于强高斯假设的最小二乘拟合，这本身就多少有些不自然。我们认为，可以选择一个不那么极端且具有一定灵活性的模型来解决这一问题，它介于单一高斯分布与无穷多个狄拉克分布之间；不过，这个问题留待未来研究。

附录 C：全局拟合

在格点量子场论分析中，拟合往往会串联进行。例如，先进行关联函数拟合，再根据这些拟合的结果求连续极限，并结合另一个拟合得到的定标输入。人们已经开发出相应方法，以便在各个可观测量分别拟合、再与其他可观测量一同纳入全局拟合时，妥善传递这条链上的相关性。例如，在自助法分析中，只需固定用于抽取样本的随机种子。而贝叶斯分析则复杂得多，因此我们建议尽可能采用联合拟合。

另一方面，如果可观测量之间不相关（或通过联合拟合得到），那么可以优雅地串联这些拟合：对于每个可观测量，我们都有一条 HMC轨迹，可将其作为某个新数据集的一个组成部分，并在该数据集上进行下一次拟合。与 a χ^2 拟合不同，对于这个新数据集，我们不对高斯性作任何假设；是否强制满足这一假设，取决于模型。这可能很重要，因为对于许多可观测量而言，高斯近似的效果不如对平均初级可观测量的近似那么好。

附录 D：变分推断

我们讨论了如何使用贝叶斯模型进行最大似然估计或最大后验估计，如何在这些估计值附近进行正态近似，以及如何进行完整的蒙特卡洛采样。变分推断（PyMC 中的 ADVI）是一种我们未曾讨论的中间方法，比完整采样略简单。我们在探索中未能通过这种方法获得良好结果，但确实将其用作 HMC 热化的起点。

其思路是用一族基分布来近似后验分布。这些基分布可以是任意形式，但常见的选择是不相关多元高斯分布。尽管这族基分布无法生成整个分布空间，但对于某些用途而言已足够。我们无需将后验分布确定为任意函数，只需优化少数几个参数即可。具体做法是最小化 Kullback–Leibler 散度。

这与围绕最大似然估计进行正态近似不同。考虑服从双峰分布的数据，就能轻易看出这一点：最大似然估计会选择其中一个峰，而变分推断则倾向于用更宽的分布覆盖两个峰。

附录 E：分箱与 K 折

如果我们无法对蒙特卡洛自相关建模，一个自然的备用方法是对数据进行分箱。贝叶斯推断同样可以采用这种方法，就像自助法中那样，原则上甚至 Γ method 也可以采用。与往常一样，这简化了建模过程，但代价是损失部分数据信息。

与分箱有些类似，我们也可以对数据进行折叠，将大小为 n 的向量 y_i 重塑为 $\frac{n}{K} \times K$ 矩阵 y_{ij}—如果我们的模型将这些新数据描述为使用一组共同参数的 K 个相同独立副本，那么实际上什么都没有改变。然而，对折叠后的数据应用留一法，实际上相当于对原始数据进行 K 折交叉验证，这可能更适合用于构建自相关数据的信息准则。事实上，留一法通常不适用于时间序列；针对这类情况，也存在留未来法等替代方法。

这种折叠也使我们能够将分箱重新表述为一种模型选择，而非对数据的修改：未分箱的版本有 K 个相同副本，因此，例如，对于高斯似然的情况，它相当于一个 K 分量的多元高斯分布，其精度矩阵就是单位矩阵；而分箱版本则可以通过选择一个秩一精度矩阵来构建。

附录 F：最大似然估计中的过度收缩

如第 II C 节所述，收缩可以改善最大似然估计。例如，可以通过蒂霍诺夫正则化（有时被不恰当地称为先验）来实现这一点。不过，将其与重采样结合时必须谨慎。设想一次拟合存在一个非常平坦的方向，例如关联函数中高激发态的能量，或者一个仅受幺正性不等式约束的方向。与我们的精度目标相比，仅仅提供较弱的正则化并不足以避免引入偏差：正则化必须比数据提供的信息更弱。否则，对于每个自助样本，似然的最大值都会非常接近先验概率的最大值，而自助法分布可能会严重低估所推断参数的不确定性。可以通过为每个自助样本移动先验来解决这一问题，但在贝叶斯推断中根本不会出现这个问题。

致谢

我们感谢 Gregorio Herdo´ıza、Rainer Sommer、Carlos Pena 和 Lorenzo Barca 就讨论及稿件提出的意见。本研究中的所有计算均在德国电子同步加速器（DESY）完成。我们感谢 CLS 倡议的同事提供用于检验我们方法的系综。

---

[1] B. Efron，《自助法：再看刀切法》，载于 *Breakthroughs in statistics*（施普林格，1992），第 569–593 页。[2] U. Wolff（ALPHA），以更少误差计算蒙特卡罗误差，《计算物理通讯》**156**，143（2004），[勘误：Com- put.Phys.Commun. 176，383（2007）], arXiv:hep-lat/0306017。

[3] A. Ramos，蒙特卡洛数据误差分析的自动微分，计算物理通讯 **238**，19（2019），arXiv:1809.01289 [hep-lat]。[4] W. I. Jay 和 E. T. Neil，格点场论结果分析中的贝叶斯模型平均，Phys. Rev. D **103**，114502（2021），arXiv:2008.01069 [stat.ME]。[5] C. Stein，多元正态分布均值的通常估计量不可容许，载于 *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability: Contributions to the Theory of Statistics* 1，第卷（加利福尼亚大学出版社，1956），第 197 页。[6] P. Billingsley，*Probability and measure*，第 3 版，Wiley-Interscience 出版物（Wiley，纽约 [u.a.]，1995）。[7] A. E. Hoerl 和 R. W. Kennard，岭回归：非正交问题中的有偏估计，Technometrics **12**，55（1970），https://www.tandfonline.com/doi/pdf/10.1080/00401706.1970.10488634。[8] K. Levenberg，最小二乘法中求解某些非线性问题的方法，《应用数学季刊》**2**，164（1944）。[9] J. Salvatier、T. V. Wiecki 和 C. Fonnesbeck，使用 PyMC3 在 Python 中进行概率编程，PeerJ 计算机科学 **2**，e55（2016）。[10] S. Duane、A. Kennedy、B. J. Pendleton 和 D. Roweth，混合蒙特卡洛，《物理快报B》**195**，216（1987）。[11] Theano开发团队，Theano：用于快速计算数学表达式的 Python框架，arXiv 电子预印本 **abs/1605.02688**（2016）。[12] J. Frison，迈向格点量子色动力学的完全贝叶斯分析——配套代码，10.5281/zenodo.7612101（2023）。[13] H. Akaike，时间序列的典型相关分析及信息准则的使用，载于 *Mathematics in science and engineering* 126 1976 27 96，第 卷（Elsevier，），第–页。[14] S. Watanabe，不可实现且正则情形下统计学习的状态方程，《电子信息通信学会电子、通信与计算机科学基础论刊》**E93-A**，617（2010）。[15] A. Vehtari、A. Gelman 和 J. Gabry，使用留一交叉验证和广泛适用信息准则进行实用的贝叶斯模型评估，《统计与计算》**27**，1413（2016）。[16] M. Bruno 和 R. Sommer，相关数据和自相关数据的拟合，《计算物理通讯》**285**，108643（2023），arXiv:2209.14188 [hep-lat]。[17] K. Takeuchi，信息统计量的分布与模型的有效性判据，《数学科学》**153**，12（1976）。[18] D. B. Rubin，贝叶斯自助法，《统计年鉴》**9**，130（1981）。