arXiv:2601.01190v1 [stat.ME] 3 Jan 2026

用于混合与层次框架中模型选择的修正贝叶斯准则

Diogenes de Jesus Ramirez*、Anderson Melchor Hernandez、Isabel Cristina Ramíre Z,
和 Luis Raúl Pericchi

摘要。本文提出一种专为混合模型和层次结构设计的修正贝叶斯信息准则（BIC）。该准则纳入对数似然函数海森矩阵的行列式，从而通过考虑似然曲面的曲率，对经典贝叶斯因子加以改进。这类几何信息为模型复杂度提供了更细致的惩罚。所提出的方法改进了模型选择，尤其适用于小样本条件或存在噪声变量的情况。通过理论推导和大量模拟研究（包括线性模型和线性混合模型），我们表明，该准则的表现始终优于贝叶斯信息准则（BIC）、赤池信息准则（AIC）等传统方法及其相关变体。结果表明，整合来自似然景观的曲率信息，能够在复杂数据环境中实现更稳健、准确的模型判别。

目录

1. 引言 2. 背景与相关工作 2.1. 贝叶斯因子 3. 贝叶斯因子的扰动 3.1. 所提出的准则 4. 示例：贝叶斯信息准则与修正准则的比较 5. 模拟 5.1. 线性模型的模拟研究 5.2. 线性混合模型的模拟 5.3. 二项与泊松模型的模拟 6. 证明 7. 结论 致谢 道德标准合规声明 参考文献

1
4
4
4
5
6
8
8
8
10
12
15
19
20
20
20

1. 引言

信息存储需求的不断增长及其分析所面临的挑战，给知识领域的各类主体及其专业领域带来了新的预测挑战。因此，生物统计学、流行病学、经济学、心理学、社会学以及其他许多人类知识分支，每天都在各自组织中面对挑战，不仅要处理分类信息，还要利用大量已存储的数据开展并推进研究，以发现行为模式。通常，已存储的信息对社会中收集和存储

日期：2026年1月6日。2020数学主题分类。62F15,62J12,6208,62H12。关键词与短语。贝叶斯因子、模型选择、混合模型、海森矩阵、贝叶斯信息准则、拉普拉斯近似。*通讯作者：名：Diogenes De Jesus；姓：Ramirez Ramirez；电子邮箱：ddramirezra@unal.edu.co。

这些信息的各类主体都具有重要价值，因为可以根据一组变量构建数学模型，使其尽可能贴近所观察到的现实。通过实施广泛的假设检验，可以评估这些模型；这些检验在很大程度上有助于分类、个体化，并选择数学背景能够使研究者与模拟现实建立连贯联系的模型。然而，任何统计检验都会失去意义和重要性，尤其是在我们不了解自然生成一组数据的确切方式，也不了解模型所包含的特征或变量之间现有的交互作用机制时。接续上述讨论，一旦与现代信息与通信技术相结合，统计模型的需求和应用便愈发重要。这些模型不仅能让学术界人士，也能让商业主体在不确定性日益受到控制的环境中做出决策，而这正是科学与学术发挥作用之处。鉴于以上所述，统计学及其通过实施众所周知的方法和技术来使用一组数据的方式，不仅能够借助特定数据样本进行推断，还能促使人们思考特定总体、样本或数据集的行为，并以概率衡量其误差风险。另一方面，根据 [29], 作者的观点，统计模型在日常生活中的应用，是一种基于数学上未知真相的现实近似。作者认为，不存在能够呈现完美且完整现实的真实模型。正确指定模型对于进行一致且高效的推断至关重要。遗憾的是，要判断模型是否正确并非总是容易，因此需要比较模型并确定其中哪一个最佳，也就是最受数据支持的模型。如果模型是嵌套的，任务就尤其简单，因此模型选择等同于参数检验 [16]。近年来，线性混合模型已被广泛用于医学、心理学及其他领域的纵向数据分析[11, 14, 26]。这类统计模型包含固定效应和随机效应，从而赋予模型层级结构。当数据不独立时，通常会使用这类模型。例如，可以设想从不同教室中抽取学生样本。混合模型的核心在于纳入固定效应和随机效应。固定效应是不会变化的参数。例如，我们可以假设总体中存在某条真实回归线。相比之下，随机效应本身是随机变量。近年来，贝叶斯线性混合模型受到广泛关注，尤其是贝叶斯因子假设检验 [8, 22, 30]。然而，尽管有关贝叶斯程序的文献浩如烟海，对于最适合用于量化支持实验效应所需变量数量的贝叶斯因子假设，仍存在相当大的模糊性；尤其是在考虑模型中不同随机项时，更是如此。

在这项工作中，我们提出了一种扰动贝叶斯规则，可帮助我们决定如何从数据中排除对于特定模型可视为有用的变量。在科学界，科学家多次接受经典零假设显著性检验（简称 NHST）的训练。举例来说，我们考虑需要检验某个具体检验假设的情形（例如，疫苗对某个古代社群的效应）。在经典方法中，NHST 通过考虑一个零假设（例如，疫苗已经显示出效果）来进行检验，因此需要利用样本数据验证这一假设要求。这类统计检验通常会与零假设成立时可能产生的分布样本进行比较。然而，这种方法存在一些局限，例如，我们可以拒绝零假设，却不能直接接受零假设。为克服这一局限，我们可以有效利用所谓的贝叶斯因子指标，简称 BIC [7]。因此，使用这一准则需要指定一个零模型（即假设疫苗没有效果）和一个备择模型。在频率学派世界中，指定完整数据的备择混合效应模型的一种有效方法，是选择一个包含实验所需全部固定效应和随机效应的模型（即考虑最大模型）[25]。尽管最大模型可以与多个零模型进行比较，但明确指定零模型非常恰当，因为它决定了我们要探究的效应问题。例如，在检验实验中的操纵效应、量化效应以及设定先验分布时，都需要作出不同的决定。拟合贝叶斯模型后，人们通常会希望衡量其预测准确性。近年来，交叉验证和信息技术逐渐被用于拟合模型 [2, 18, 23] 。采用交叉验证方法时，需要重新拟合

使用不同训练集的模型。渡边-赤池信息准则（WAIC）是另一种统计方法，可视为对贝叶斯模型偏差信息准则（DIC）的改进。WAIC 是一种贝叶斯方法，使用参数的整个后验分布。此外，它在重新参数化下具有不变性 [10, 24, 30]。

我们的贡献
在本研究中，我们提出一种经过修正的贝叶斯因子假设检验，以处理混合效应模型的贝叶斯推断。具体而言，我们通过加入一项额外项来扩展通常的贝叶斯信息准则（BIC）；该项涉及在对数似然最大值处计算的海森矩阵的行列式。作出这一修正的原因是，某些模型中的对数似然函数可能特别难以优化，甚至无法处理。加入行列式项，能够将似然曲面的局部曲率纳入考量，从而反映模型所包含的信息量。形式上，对于给定的模型假设 H，我们定义扰动 *BIC* 为

$$\text{BIC_HES}(H) := \text{BIC}(H) + \log\left(\det(P)\right), \tag{1.1}$$

其中， P 表示费 Fisher 信息矩阵，$\det(\cdot)$表示其行列式，而$\log(\cdot)$表示自然对数。加入额外的$\log(\det(P))$项并非随意之举；除了其数学推导（见推论3.1 和 [12, 第 7.2 节。2]）之外，该项还刻画了标准 BIC 所忽略的层级或异质数据的结构特征。为说明这一点，请看以下简单的混合效应模型：

$$X_{il} = \mu_i + \epsilon_{il}, \quad \text{for } i = 1, \ldots, p \text{ groups and } l = 1, \ldots, r \text{ observations per group,}$$

其中 $\epsilon_{il} \sim N(0,\ \sigma^2)$ 是独立同分布随机变量。在这一设定下计算 $\log(\det(P))$（见第 4节）可得

$$\log\left(\det(P)\right) = (p+1)\log(r) + \log(p) - \log(2) - (p+2)\log(\hat{\sigma}^2).$$

在大样本情境下，$\log(\det(P))$ 通常会被忽略，因为其贡献渐近可忽略。然而，在小样本情境中，这一项可能相当显著，并通过经验估计量$\hat{\sigma}^2$ 引入明确的数据依赖性。值得注意的是，当模型对数据拟合得极好时（即$\hat{\sigma}^2$ 很小时），项 $-(p+2)\log(\hat{\sigma}^2)$ 会产生很强的正向惩罚。这种行为通过惩罚那些解释随机噪声而非数据中真实结构的模型，来抑制过拟合。因此，显然，额外的 $\log(\det(P))$ 项引入了一种取决于数据层级结构的自适应惩罚。事实上，$\det(P)$ 反映了模型参数中的总体信息量；在这一简单设定下，它取决于组数 p、每组的观测值数量 r 以及残差方差$\hat{\sigma}^2$。因此，该准则不仅像经典BIC那样根据模型维数进行惩罚，还会考虑模型内部的信息结构 [3]。这带来了更灵活的调整，尤其适用于不平衡或小样本的混合效应模型，在这类模型中，各参数的费 Fisher 信息并不均匀。最后，类比于 [26], 中讨论的留一法，我们考虑通过计算得到的逐点对数预测密度来对贝叶斯准则进行预测近似，记为 $\widehat{\text{lpd}}$。

本文的结构安排如下。
第 2节回顾与贝叶斯因子相关的关键概念。第 3 节介绍我们提出的准则，并给出推论 3.1 和定理 3.2，为我们提出的修正贝叶斯准则提供理论依据。此外，定理 3.3 保证了该准则的一致性。

第 4 节通过分析一个含随机效应的模型，展示该准则在实际应用中的有效性，并明确计算系数 $\log(\det(P))$。在第 4小节中，我们将所提出的方法与经典BIC进行比较，并着重指出其优势，尤其是在小样本情形下。

第 5节介绍了涉及线性、混合、二项和泊松模型的模拟研究，以评估我们的准则相较于其他成熟信息准则的表现。第 5.1 小节重点讨论线性情形，而第 5.2 小节讨论混合模型。第 5.3 节介绍二项与泊松模型。

在第 6 节中，我们将详细证明推论 3.1、定理 3.2 和 定理 3.3。最后，第 7 节概述了未来研究的潜在方向。

2. 背景与相关研究

接下来，我们将紧密遵循 [7]，回顾贝叶斯因子的主要特征。

2.1. 贝叶斯因子。 贝叶斯推断是一种基于计算 $\mathbb{P}(H|D)$ 的测量方法，它表示给定数据 D 时假设 H 的后验概率。贝叶斯定理将该概率表示为

$$\mathbb{P}(H|D) = \frac{\mathbb{P}(D|H) \cdot \mathbb{P}(H)}{\mathbb{P}(D)} \tag{2.1}$$

其中，$\mathbb{P}(H)$ 通常称为假设（或样本）H 的先验概率。利用贝叶斯定理，很容易得出

$$\frac{\mathbb{P}(H_0|D)}{\mathbb{P}(H_1|D)} = \frac{\mathbb{P}(D|H_0)}{\mathbb{P}(D|H_1)} \cdot \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)}. \tag{2.2}$$

在这个等式中，著名的贝叶斯因子为

$$\mathrm{BF}_{0,1} := \frac{\mathbb{P}(D|H_0)}{\mathbb{P}(D|H_1)}. \tag{2.3}$$

注意下标所体现的依赖关系很重要。同样，我们可以看到

$$\mathrm{BF}_{1,0} = \frac{1}{\mathrm{BF}_{0,1}},$$

更一般地，对于每个合适的假设，这一关系为 H_i 其中 i 是一个适当的索引。虽然公式（2.3）看起来似乎相当简单，但在实际示例中计算起来要困难得多。这是因为要计算分子或分母，需要引入多个参数化表示。更确切地说，上述公式（2.3）通常写作

$$\mathrm{BF}_{0,1} = \frac{\int_{\Theta_0} \mathbb{P}(D|H_0, \theta) g_0(\theta) d\theta}{\int_{\Theta_1} \mathbb{P}(D|H_1, \theta) g_1(\theta) d\theta} \tag{2.4}$$

其中，Θ_0 和 Θ_1 分别是模型 H_0 和 H_1的参数空间，g_0 和 g_1 分别是模型 H_0 和 H_1的参数先验概率密度函数。请注意，要计算 $\mathrm{BF}_{0,1}$，必须指定先验 g_0 和 g_1。然而，这会导致估计方面的理论问题，因此，如果缺乏这类先验信息，$\mathrm{BF}_{0,1}$ 便无法获得。基于这一原因，文献中提出了各种分类方案。在 [20] 中，采用了一种分类方法：BIC 在 1 到 3 之间被视为弱证据，3 到 20 之间可视为正面证据，而处于范围 [20, 150] 内的值构成强证据，最后，超过 150 则构成非常强的证据。从某种意义上说，当数据集没有大幅改变变量数量时，BIC 仅提供了一个偏好某个假设而非另一个假设的指标。当所处理的模型与数据集存在不匹配时，有时需要评估其预测行为，从而估计其误差 [27]。

3. A perturbation of Bayes factor

对于给定的模型 H_i，贝叶斯信息准则定义为

$$\mathrm{BIC}(H_i) := -2\log(L_i) + p_i \cdot \log(n),$$

其中，n 是观测值的数量，p_i 是模型 H_i 的自由参数数量，L_i 是模型 H_i 的最大似然。在 [28] 中提出

$$\mathrm{BF}_{0,1} \approx \exp\left(\mathrm{BIC}\left(\Delta_{1,0}/2\right)\right)$$

其中，$\mathrm{BIC}(\Delta_{1,0}) := \mathrm{BIC}(H_1) - \mathrm{BIC}(H_0)$。然而，这并不能让我们将 H_0 和 H_1 视为同一随机过程的产物，尽管从数学上看它们是不同的对象。假设 $\hat{\theta}_1$ 和 $\hat{\theta}_0$

分别是使 L_1 和 L_0 最大化的参数。为解决这一技术问题，我们现在尝试提出一种不同的方法来完善贝叶斯因子。

3.1. 所提出的准则。 在此，我们提出采用如下形式：

$$\text{BIC_HES}(H) := \text{BIC}(H) + log\left(\det(P)\right) \tag{3.1}$$

其中 P是一个大小为 $p \times p$ 的费 Fisher 信息矩阵，该矩阵由统计模型 H 确定，并依赖于 p 参数。正如引言中所预期的，项 $log\left(\det(P)\right)$ 可通过考察似然函数的二阶展开来论证。我们回顾一下，从分析的角度来看，费 Fisher 信息可视为对数似然函数的负海森矩阵。因此，确定其特征值为正还是为负，可以据此对其驻点进行分类。因此，方程（3.1）的本质在于，它考虑了参数所在空间的"几何结构"。不过，为此付出的代价是，该准则会额外增加一项 $p\log(n)$，其中 n 表示模型 H的样本量。我们将这一观察总结为以下推论：

推论 3.1. 设 Θ_i 是 $\mathbb{R}^d$ 中的一个样本集，其取自一个参数数量为 p 的参数族，并假设该样本集在旋转下保持不变（即，对于所有满足 $P^2 = P$的方阵 P，都有 $P\Theta = \Theta$）。假设 $\tilde{\theta}_i \in \Theta$ 是使似然函数 $\log\left(f_i(\cdot|\theta)g_i(\theta)\right)$ 最大化的参数向量，且该函数足够光滑。那么存在一个大小为 $p \times p$ 的正定矩阵 P_i，使得

$$\mathbb{P}(y|H_i) = f_i(y|\tilde{\theta}_i)g_i(\tilde{\theta}_i)\frac{(2\pi)^{p/2}}{\det(P_i)^{1/2}}.$$

利用这一推论，我们可以按如下方式比较不同模型。

定理 3.2。 假设 Θ_i 在旋转下保持不变，也就是说，对于任意正交矩阵 U，$U\ \Theta_i\ = \Theta_i$ $f_i\ |H_i\ g_i$ C^2 Θ_i () 。此外，假设 $\log(())$ 在所有 上均满足，并且在 $i = 0{,}1$处存在唯一的驻点。那么

$$\log(\text{BF}_{0,1}) = \log\left(\frac{f_0(y|\tilde{\theta}_0)g_0(\tilde{\theta}_0)}{f_1(y|\tilde{\theta}_1)g_1(\tilde{\theta}_1)}\right) + \frac{1}{2}\log(2\pi)\left(|\tilde{\theta}_0| - |\tilde{\theta}_1|\right) + \frac{1}{2}\log\left(\frac{\det(P_1)}{\det(P_0)}\right) \tag{3.2}$$

其中 P_i 是由以下公式给出的费 *Fisher* 信息矩阵 $-\nabla^2\log\left(f_i(y|\theta)g_i(\theta)\right)$ 在 $\tilde{\theta}_i$处，且 $|\tilde{\theta}_i| = \text{length}(\tilde{\theta}_i)$ 对于 $i = 0{,}1$。此外，若假设 $|\tilde{\theta}_1| = |\tilde{\theta}_0|$，则 (3.2) 可写成

$$2\log(\text{BF}_{0,1}) = \text{BIC_HES}(\Delta_{1,0}). \tag{3.3}$$

作为 定理 3.2的一个应用，我们考虑数据 $y_1, \ldots, y_n$，在给定参数 θ的条件下视为相互独立。为了与给定数据集保持可比性，并便于解释（3.2），我们需要分别明确表示 $\tilde{\theta}_0$和 $\tilde{\theta}_1$。假设我们有先验分布 $\mathbb{P}(\theta)$，由此得到后验分布 $\mathbb{P}(\theta|y)$ 以及后验预测分布 $\mathbb{P}(\tilde{y}|y) = \int \mathbb{P}(\tilde{y}|\theta)\mathbb{P}(\theta|y)dy$。有时使用所谓的逐点对数预测密度会很有用，记作 lpd，其定义如下：

$$\text{lpd} := \sum_{i=1}^{n}\log(\mathbb{P}(y_i|y)) = \sum_{i=1}^{n}\log\left(\int \mathbb{P}(y_i|\theta)\mathbb{P}(\theta|y)d\theta\right). \tag{3.4}$$

这是对新数据集的期望对数逐点预测密度的一种高估度量。请注意，在实际计算 lpd 时，我们需要从后验分布中抽取样本，记为 θ^j, $j = 1, \ldots p$, $p \in \mathbb{N}$。因此，lpd 可以近似为

$$\widehat{\text{lpd}} = \sum_{i=1}^{n}\log\left(\frac{1}{p}\sum_{j=1}^{p}\mathbb{P}(y_i|\theta^j)\right). \tag{3.5}$$

因此，可以将（3.2）估计为

$$\begin{aligned}\widehat{\log(\mathrm{BF}_{0,1})} = &\sum_{i=1}^{n} \log\left(\frac{\frac{1}{p}\sum_{j=1}^{p} f_0(y_i|\theta_0^j)}{\frac{1}{p}\sum_{j=1}^{p} f_1(y_i|\theta_1^j)}\right) + n\log\left(\frac{\frac{1}{p}\sum_{j=1}^{p} g_0(\theta_0^j)}{\frac{1}{p}\sum_{j=1}^{p} g_1(\theta_1^j)}\right) \\ &+ \frac{1}{2}\log(2\pi)\left(|\tilde{\theta}_0| - |\tilde{\theta}_1|\right) + \frac{1}{2}\log\left(\frac{\det(P_1)}{\det(P_0)}\right).\end{aligned} \tag{3.6}$$

在实际计算 $\widehat{\log(\mathrm{BF}_{0,1})}$ 时，我们可以使用计算得到的逐点对数预测密度，因此

$$\begin{aligned}\widehat{\log(\mathrm{BF}_{0,1})} = &\widehat{\mathrm{lpd}}_0 - \widehat{\mathrm{lpd}}_1 + n\log\left(\frac{\frac{1}{p}\sum_{j=1}^{p} g_0(\theta_0^j)}{\frac{1}{p}\sum_{j=1}^{p} g_1(\theta_1^j)}\right) \\ &+ \frac{1}{2}\log(2\pi)\left(|\tilde{\theta}_0| - |\tilde{\theta}_1|\right) + \frac{1}{2}\log\left(\frac{\det(P_1)}{\det(P_0)}\right).\end{aligned} \tag{3.7}$$

在某些特定情况下（见第 4 节），我们可以假设在方程（3.2）中，自由参数的维数 θ_0 和 θ_1相同，且方程（3.2）右端的第一项可以忽略。因此，可以得到以下决策方法：

$$\log(\mathrm{BF}_{0,1}) = \frac{1}{2}\log\left(\frac{\det(P_1)}{\det(P_0)}\right).$$

因此，尽管我们的估计量乍看之下似乎是一个人为添加的项，但它实际上能够揭示参数数量相同的复杂数据集之间的差异。例如，模型仅在一个组成部分上有所不同的情形便可能出现这种情况，如下文方程（4.5）所研究的案例所示。关于 贝叶斯信息准则_HES 的一个重要性质是其一致性。以下结果阐述了这一点。

定理 3.3。 考虑指示准则贝叶斯信息准则_HES，其定义为

$$\mathrm{BIC_HES} = -2\log L(\hat{\theta}) + p\log n + \log|J|$$

其中，对数 $L(\hat{\theta})$ 为最大化后的对数似然， p 为模型中的参数数量， n 为样本量（观测值数量）， J 为观测到的费舍尔信息矩阵（负海森矩阵），在最大似然估计量$\hat{\theta}$处求值。

因此，该准则具有一致性。更准确地说，给定真实数据生成模型 M_0 （具有 p_0 个参数）和过拟合竞争模型 M_1 （具有 p_1 个参数，且 $p_1 > p_0$ 满足 $M_0 \subset M_1$），随着样本量增加，选择真实模型的概率趋近于 *1*：

$$\lim_{n\to\infty} \mathbb{P}(\Delta\mathrm{BIC_HES} > 0) = 1,$$

其中 Δ贝叶斯信息准则_HES $=$ 贝叶斯信息准则_HES(M_1) − 贝叶斯信息准则_HES(M_0)。

4. Examples

本节分析此前在 [17, 第 4], 节研究过的若干模型，并在此基础上考察相关研究对所谓缩放 Beta-2 分布稳健性质的观察。这类分布构成了一个庞大的分布族，可用于处理层级和非层级设定中的尺度参数建模。由于我们感兴趣的是分析公式（3.2），下面考虑[17, 9]中讨论的模型：

$$X_{il} = \mu_i + \epsilon_{il} \quad \text{for } i = 1,\ldots,p \text{ groups and } l = 1,\ldots,r \text{ observations per group.} \tag{4.1}$$

这是一种正态层级模型，定义如下：数据 x_{il} l——第 i 组中的观测值，其中误差 $\epsilon_{il} \sim N(0,\ \sigma^2)$ 是均值为 null、方差为 σ^2 的独立同分布实值高斯变量。请注意，每组具有相同的均值，且该均值是确定性变量，但也可以选择将 μ_i 设为正态随机变量。参数向量为 $\theta = (\mu_1,\ \ldots\ldots,\ \mu_p,\ \sigma^2)$，维数为 $p+1$。对数似然函数为：

$$l(\theta) = -\frac{pr}{2}\log(2\pi) - \frac{pr}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{p}\sum_{l=1}^{r}(X_{il}-\mu_i)^2.$$

最大似然估计量（MLE）为：

- $\hat{\mu}_i = \bar{X}_{i\cdot} = \frac{1}{r}\sum_{l=1}^{r} X_{il}$
- $\hat{\sigma}^2 = \frac{1}{pr}\sum_{i=1}^{p}\sum_{l=1}^{r}(X_{il}-\hat{\mu}_i)^2$

我们先针对一般参数计算海森矩阵$\mathbf{H}(\theta)$，然后再在最大似然估计处求值。因此，可得

$$\begin{aligned}
&\frac{\partial l}{\partial \mu_j} = \frac{1}{\sigma^2}\sum_{l=1}^{r}(X_{jl}-\mu_j),\\
&\frac{\partial l}{\partial(\sigma^2)} = -\frac{pr}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{p}\sum_{l=1}^{r}(X_{il}-\mu_i)^2,\\
&\frac{\partial^2 l}{\partial \mu_j^2} = -\frac{r}{\sigma^2}, \quad \frac{\partial^2 l}{\partial \mu_j \partial \mu_k} = 0 \quad (\text{for } j \neq k),\\
&\frac{\partial^2 l}{\partial(\sigma^2)^2} = \frac{pr}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}\sum_{i=1}^{p}\sum_{l=1}^{r}(X_{il}-\mu_i)^2,\\
&\frac{\partial^2 l}{\partial \mu_j \partial(\sigma^2)} = -\frac{1}{(\sigma^2)^2}\sum_{l=1}^{r}(X_{jl}-\mu_j). \qquad (4.2)
\end{aligned}$$

我们注意到，对于一般参数 μ_j，(4.2) 不为零，而且一般 θ对应的海森矩阵不是分块对角矩阵。然而，在最大似然估计处对其求值时，$\hat{}\theta = (\hat{}\mu_1,..,\hat{}\mu_p,\hat{\sigma}^2)$，非对角项消失。具体而言，交叉导数项变为：

$$\left.\frac{\partial^2 l}{\partial \mu_j \partial(\sigma^2)}\right|_{\theta=\hat{\theta}} = -\frac{1}{(\hat{\sigma}^2)^2}\sum_{l=1}^{r}(X_{jl}-\hat{\mu}_j) = -\frac{1}{(\hat{\sigma}^2)^2}\times 0 = 0$$

右下角项则简化为：

$$\left.\frac{\partial^2 l}{\partial(\sigma^2)^2}\right|_{\theta=\hat{\theta}} = \frac{pr}{2(\hat{\sigma}^2)^2} - \frac{pr\hat{\sigma}^2}{(\hat{\sigma}^2)^3} = -\frac{pr}{2(\hat{\sigma}^2)^2}$$

因此，在最大似然估计处求值的海森矩阵是分块对角矩阵的：

$$\mathbf{H}(\hat{\theta}) = \begin{pmatrix} -\frac{r}{\hat{\sigma}^2}\mathbf{I}_{p\times p} & \mathbf{0} \\ \mathbf{0}^T & -\frac{pr}{2(\hat{\sigma}^2)^2} \end{pmatrix}$$

根据定义，观测费 Fisher 信息矩阵为 $\mathbf{I}(\hat{}\theta) = -\mathbf{H}(\hat{}\theta)$：

$$\mathbf{I}(\hat{\theta}) = \begin{pmatrix} \frac{r}{\hat{\sigma}^2}\mathbf{I}_{p\times p} & \mathbf{0} \\ \mathbf{0}^T & \frac{pr}{2(\hat{\sigma}^2)^2}. \end{pmatrix}$$

行列式等于对角块行列式的乘积：

$$\det(\mathbf{I}(\hat{\theta})) = \left(\frac{r}{\hat{\sigma}^2}\right)^p \times \frac{pr}{2(\hat{\sigma}^2)^2}$$

而对数行列式为：

$$\begin{aligned}
\log\det(\mathbf{I}(\hat{\theta})) &= p\log\left(\frac{r}{\hat{\sigma}^2}\right) + \log\left(\frac{pr}{2(\hat{\sigma}^2)^2}\right)\\
&= p\log(r) - p\log(\hat{\sigma}^2) + \log(p) + \log(r) - \log(2) - 2\log(\hat{\sigma}^2)\\
&= (p+1)\log(r) + \log(p) - \log(2) - (p+2)\log(\hat{\sigma}^2).
\end{aligned}$$

贝叶斯信息准则与修正准则的比较。 标准贝叶斯信息准则与所提出的准则之间的关键区别在于其惩罚项的结构。对于此示例（4.1），参数总数为 $p_{\text{total}} = p + 1$。假设 $n = r$ 为有效样本量（每个 μ_i 根据 r 个观测值进行估计），经典BIC惩罚项为

$$\text{Penalty}_{\text{BIC}} = p_{\text{total}} \log n = (p + 1) \log(r). \tag{4.3}$$

这种表述忽略了模型的内部结构以及每个参数可用的信息量。相比之下，纳入费 Fisher 信息矩阵的对数行列式可得

$$\text{Penalty}_{\text{HES}} \coloneqq (p + 1) \log(r) + \log(p) - \log(2) - (p + 2) \log(\hat{\sigma}^2), \tag{4.4}$$

该准则保留了 BIC 项作为第一部分，同时增加了三个有意义的调整项：

(1) $+\log(p)$：引入组依赖惩罚，以反映模型的层次维度；(2) $-\log(2)$：常数缩放调整；(3) $-(p+2)\log(\hat{\sigma}^2)$：数据依赖项，用于反映模型拟合精度。

当 $\hat{\sigma}^2$ 较小时（表明拟合紧密），$\log(\hat{\sigma}^2)$ 会变得很大且为负值，因此惩罚项 $-(p+2)\log(\hat{\sigma}^2)$ 会变得显著且为较大的正值。这会自适应地惩罚那些可能通过解释随机噪声而非真实信号来过拟合的模型。例如，考虑两个模型 H_0和 H_1，如式（4.1）所示，其每组的观测值数分别为 r_0和 r_1。假设两个模型的参数数量完全相同，$p+1$，且仅在方差项 σ_0^2和 σ_1^2上有所不同。为简单起见，假设式（3.2）右侧的第一项可以忽略。因此，

$$\log(\text{BF}_{0,1}) = (p + 1) \log(r_1/r_0) + (p + 2) \log(\hat{\sigma}_0^2/\hat{\sigma}_1^2), \tag{4.5}$$

其中，$\hat{\sigma}_0^2$，$\hat{\sigma}_1^2$ 分别是 σ_0^2和 σ_1^2的经验估计量。假设 $0 < \hat{\sigma}_0^2 < \hat{\sigma}_1^2$，且$r_1 < r_0$。那么模型 H_0 包含更多噪声变量，并且在所提出的BIC_HES准则下会受到有效惩罚。相比之下，标准BIC无法保证有效区分包含冗余噪声变量的模型。

5. 模拟

本节介绍为评估修正贝叶斯信息准则（BIC_HES）性能而开展的模拟实验。该准则纳入了参数海森矩阵行列式的对数。主要目标是分析这种修正如何影响不同情境下的模型选择，并将其与标准BIC、赤池信息准则（AIC）、修正赤池信息准则（AICc）、修正赤池信息因子（CAIF）和信息复杂度准则（ICOMP）[1, 13, 4, 5]等传统准则进行比较。

下文中，在图中我们使用“指标”一词来指代用于数据分析的特定类型的贝叶斯因子。缩写 PVCSVM 表示给定指标选择正确模型的次数百分比。该缩写源自西班牙语短语“porcentaje de veces que un criterio selecciona el verdadero modelo”（某一准则选出真实模型的次数百分比）。此外，我们将 BIC_HES_SP 定义为 BIC_HES 的修改版本，即减去项 $p\log(n)$，其中 n 表示样本量，p 表示模型中的参数数量。进行这一调整的原因是，我们观察到 BIC_HES 在某些情况下可能会将惩罚项 $p\log(n)$ 计入两次。

5.1. 线性模型的模拟研究。 本节介绍线性模型的模拟设计。所有模拟均使用 R [19], 版本 4.3.2，在一台配备 1 TB 内存和 32 个 2.7 GHz Intel Xeon 核心的 Linux 工作站上进行。模拟方法遵循 [15], ：先定义真实数据生成模型，再引入额外的噪声变量以重新拟合模型，并评估各种选择准则的性能。总共进行了 200 次重复试验，生成了 1,096,800 个拟合模型。我们根据各准则识别出真实数据生成模型的经验频率来评估其性能。

一般模型定义如下：

$$Y_i = \beta_0 + \sum_{k=2}^{10} \beta_k X_{ik} + \sum_{l=1}^{10} \beta_l X_{il} + \varepsilon_i, \quad i = 1, \ldots, n,\ k = 2, 4, \ldots, 10,\ l = 1, \ldots, 10. \tag{5.1}$$

其中，噪声变量对应的索引为 $l = 1,$,10 $\ldots$.

模拟过程如下实施：

a) 生成 k 个相互独立的回归变量，其中 $k = 2$、4、6、8、10。每个变量均服从均值为 10、标准差为 1 的正态分布。b) 生成 k 个回归参数，每个参数均取自均值为 10、标准差为 2 的正态分布。c) 生成均值为 0、标准差为 1 的正态分布误差。d) 使用公式（5.1）计算样本量为 $n = 10$、15、20、50、100、500 的响应变量，并设置不同数量的回归变量 X_k，不包括噪声变量。e) 从均值为 3、标准差为 3 的正态分布中生成 $l = 1$、.....、10 个额外的噪声回归变量。f) 将噪声变量添加到步骤 d) 中的模型中，确保其数量不超过原始回归变量的数量。g) 计算信息准则（贝叶斯信息准则、赤池信息准则、修正赤池信息准则、BIC_HES、BIC_HES_SP 和 CAIF）。

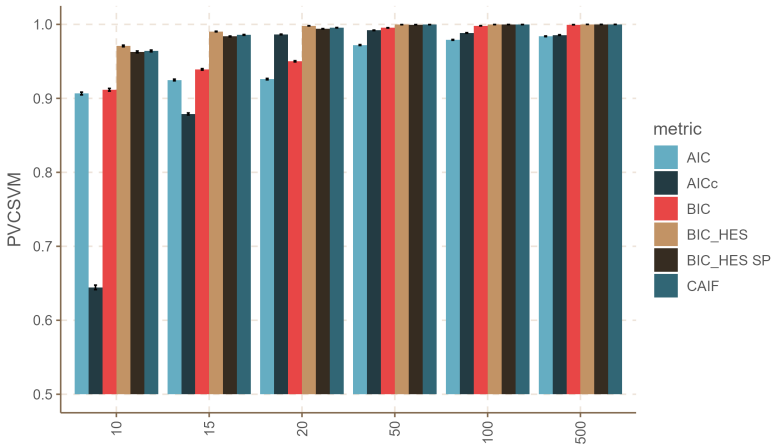


图1. 信息准则与样本量

在图 1中，BIC_HES 和 BIC_HES_SP 准则在识别真实模型方面表现更优，尤其是在样本量较小时。这表明，在这种情况下纳入海森矩阵可以引入更合适的惩罚项。尽管赤池信息准则和修正赤池信息准则广受欢迎，但其表现较差。随着样本量增加，所有准则趋于收敛，这表明当数据充足时，信息准则的选择变得不那么重要。在小样本条件下，BIC_HES 和 BIC_HES_SP 似乎能提供最可靠的结果。

图 2 展示了不同样本量下 BIC 与 BIC_HES 的直接比较。横轴表示 BIC 值，纵轴表示 BIC_HES 值。这两种准则之间强烈的线性关联表明，纳入海森矩阵的对数行列式为选择过程提供了有意义的信息。

图 3 展示了类似的比较，只是将贝叶斯信息准则替换为赤池信息准则。观察到的贝叶斯信息准则_HES 与赤池信息准则之间的关系进一步表明，贝叶斯信息准则_HES 中的附加项有助于区分竞争模型。

图4 比较了不同准则的选择频率随新增噪声变量数量变化的情况。贝叶斯信息准则_在引入噪声变量时，HES 比贝叶斯信息准则更常选择真实模型。然而，当纳入十个噪声变量时，两种准则得出的结果几乎相同

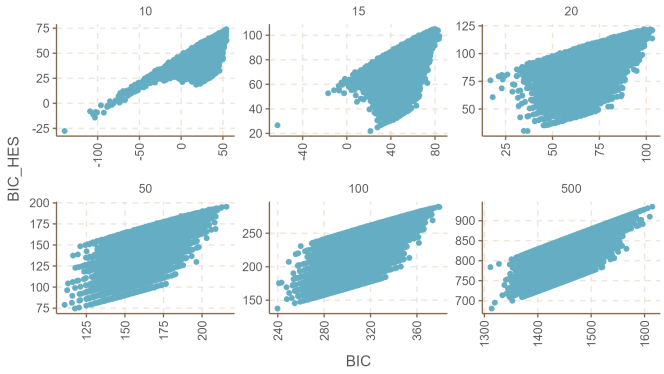


图2。BIC_HES 与 BIC 的比较。

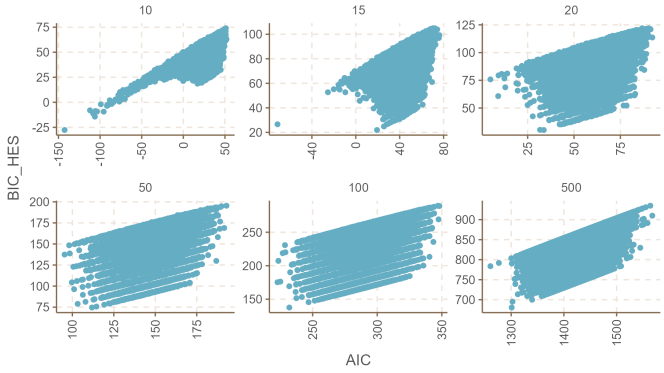


图3。BIC_HES 与赤池信息准则的比较。

表现。图5 按新增噪声变量数量和样本量共同概述了这些结果。再次发现，贝叶斯信息准则_HES 选择真实数据生成模型的概率更高。

总体而言，结果表明，贝叶斯信息准则与参数海森矩阵的对数行列式之间存在明确关联。因此，可以将这一部分纳入贝叶斯信息准则的表达式，以增强多元线性回归框架中的模型判别能力。

5.2. 线性混合模型的模拟。 本节报告了线性混合模型的模拟，其定义如下：

$$y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + u_{0i} + u_{1i} Z_{1ij} + u_{2i} Z_{2ij} + u_{3i} Z_{3ij} + \varepsilon_{ij}. \tag{5.2}$$

其中：

- y_{ij} 表示组 i内观测值 j 的响应变量；
- β_0 是固定截距；
- β_1， β_2， β_3 是固定效应系数；
- $u_{0i} \sim N(0,\ \sigma^2_{u0})$表示组 i 的随机截距；
- $u_{1i},\ u_{2i},\ u_{3i} \sim N(0,\ \sigma^2_{bk})$表示随机效应系数；• $\varepsilon_{ij} \sim N(0,\ \sigma^2_{\varepsilon})$ 表示残差项。

模拟设计如下：

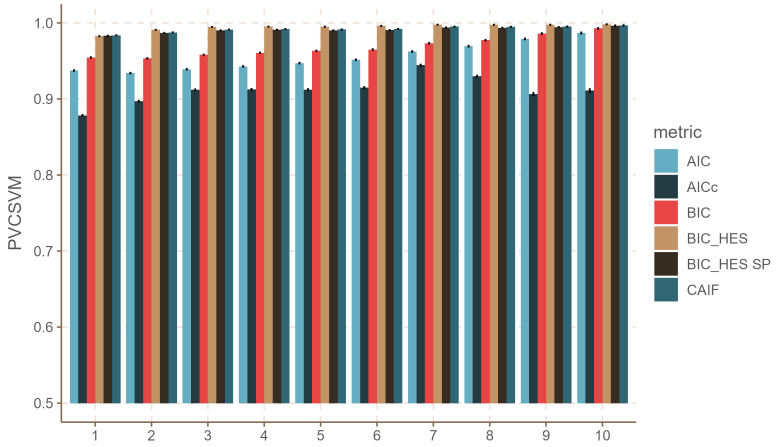


图4. 信息准则与噪声变量数量的关系

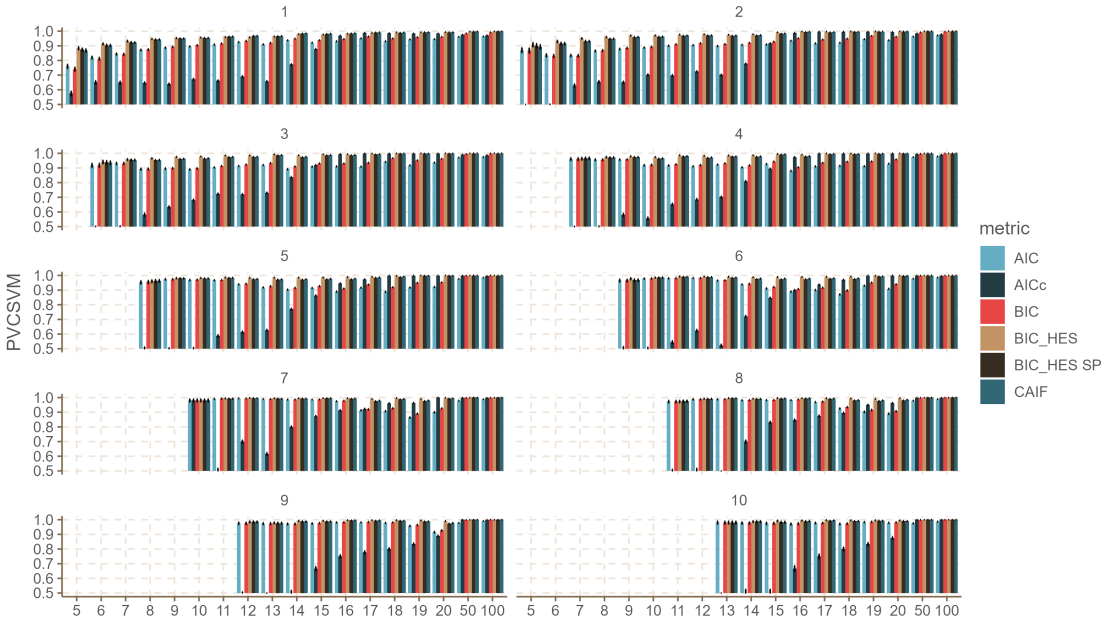


图5. 信息准则与噪声变量数量及样本量的关系

a) 从均值为 2、标准差为 3 的正态分布中生成固定效应的回归变量。b) 从均值为 2、标准差为 3 的正态分布中生成固定效应参数。c) 从均值为 0、标准差为 1 的正态分布中生成全局误差。d) 从均值为 3、标准差为 3 的正态分布中生成随机效应的回归变量。e) 从均值为 3、标准差为 1 的正态分布中生成随机效应参数。

f) 从均值为 0、标准差为 2 的正态分布中生成组级随机误差。g) 使用方程（5.2）计算样本量为 $n = 20$、50、100、500 且组大小为 $N = 5$、10、20 的响应变量。h) 为固定效应添加噪声回归变量（均值 3，标准差 3）。i) 为随机效应添加噪声回归变量（均值 3，标准差 3）。j) 将噪声变量引入模型，确保其数量不超过原始回归变量的数量。k) 计算信息准则（贝叶斯信息准则、赤池信息准则、修正赤池信息准则、BIC_HES、BIC_HES_SP 和 CAIF）。

模拟在 R [19] 中执行，使用的版本为 4.3.2，运行于配备 1 TB 内存和 32 个 2.7 GHz Intel Xeon 核心的 Linux 系统上。模型估计采用了 `GLMMadaptive` package [21] 。该软件包使用自适应高斯-埃尔米特求积，对无法通过解析方法求解随机效应积分的聚类数据拟合混合效应模型。它还提供估计参数的海森矩阵，这对本分析至关重要。总共进行了五次重复试验，生成了 8,825 个模型。

图6 表明，贝叶斯信息准则_HES 比其他准则更频繁地识别出真实模型，并且在较大样本量下达到相近的表现。对于小样本，贝叶斯信息准则_HES 仍优于竞争准则。图7 至11 展示了固定效应和随机效应下，按组大小及所添加噪声变量的数量划分的结果。在所有情形中，贝叶斯信息准则_HES 始终比传统贝叶斯信息准则更频繁地识别出真实数据生成模型，尤其是在样本量较小或存在噪声变量时。这些发现凸显了修正准则在混合效应模型框架中的稳健性。

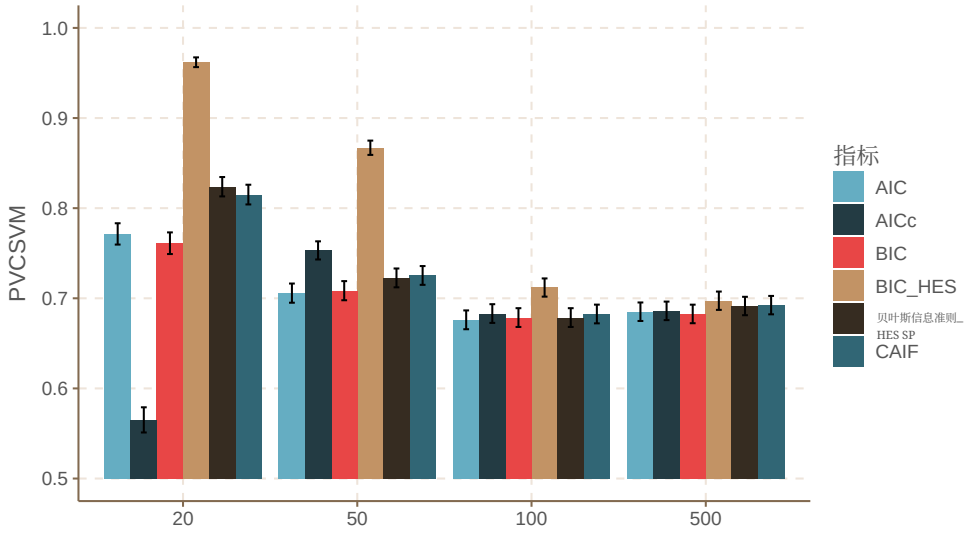


图6. 信息准则与样本量

5.3. **二项与泊松模型的模拟**。本节介绍旨在评估多种信息准则有效性的模拟，包括赤池信息准则、贝叶斯信息准则、修正赤池信息准则、贝叶斯信息准则_HES、贝叶斯信息准则_HES_SP、CAIF 和 ICOMP。该过程基于 [6], ，涉及拟合三个理论模型：广义高斯混合模型、逻辑斯蒂模型和泊松模型。各模型的参数定义如下：

- 受试者数量：20、50、100；
- 组内刺激数量：2、5、10；

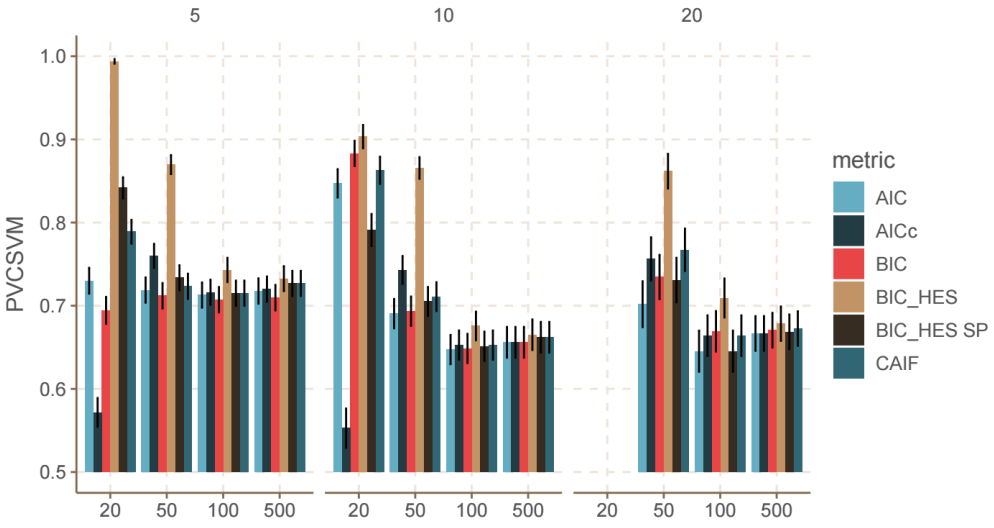


图7. 按组大小和样本量划分的信息准则

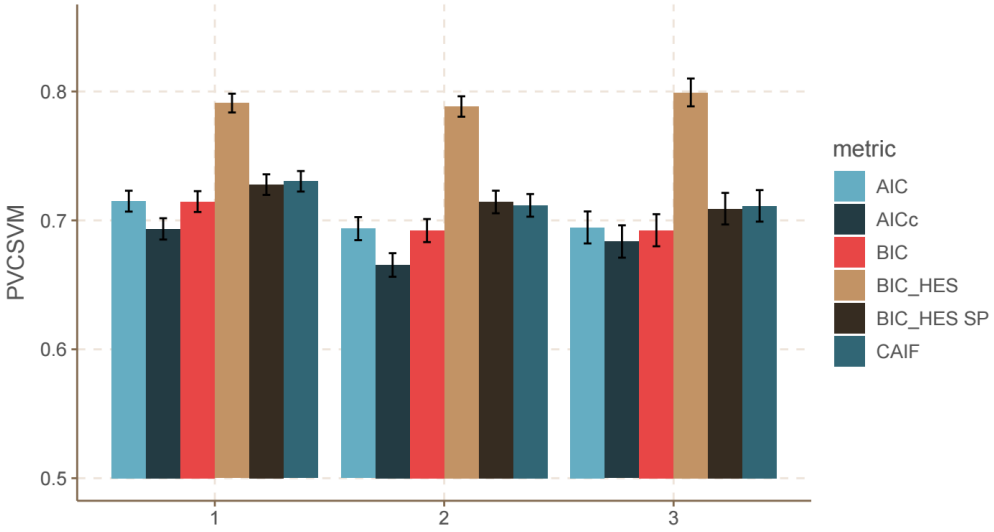


图8. 按新增固定效应变量的数量划分的信息准则

• 外群刺激数量：2、5、10；• 截距与斜率之间的相关性： -0.7, -0.5, 0, 0。5, 0。7。

模型设定：

$$\begin{aligned} \text{Gaussian mixed model:} \quad & \beta_0 = 800,\ \beta_1 = 50, \\ \text{Logistic mixed model:} \quad & \beta_0 = 0.5,\ \beta_1 = 1, \\ \text{Poisson mixed model:} \quad & \beta_0 = 5,\ \beta_1 = 4. \end{aligned}$$

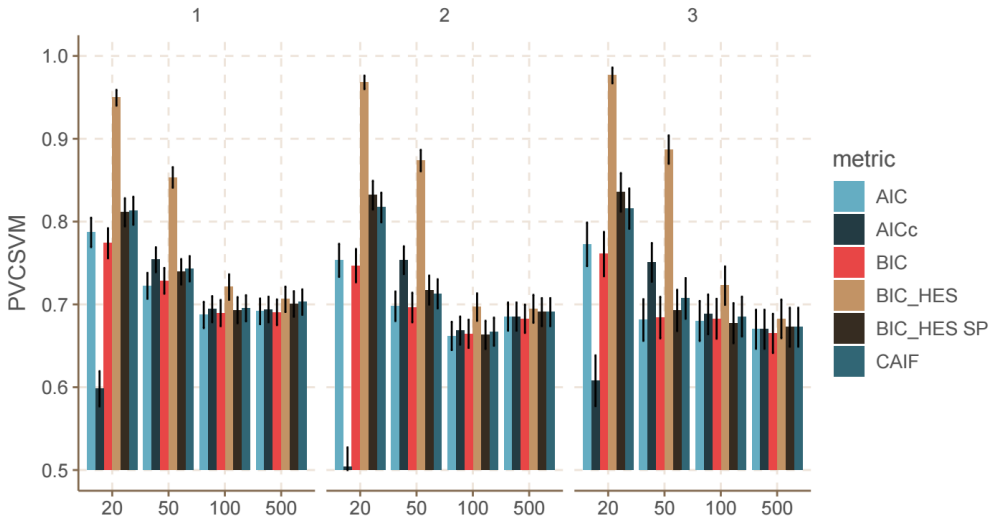


图9. 按样本量和新增固定效应变量数量划分的信息准则。

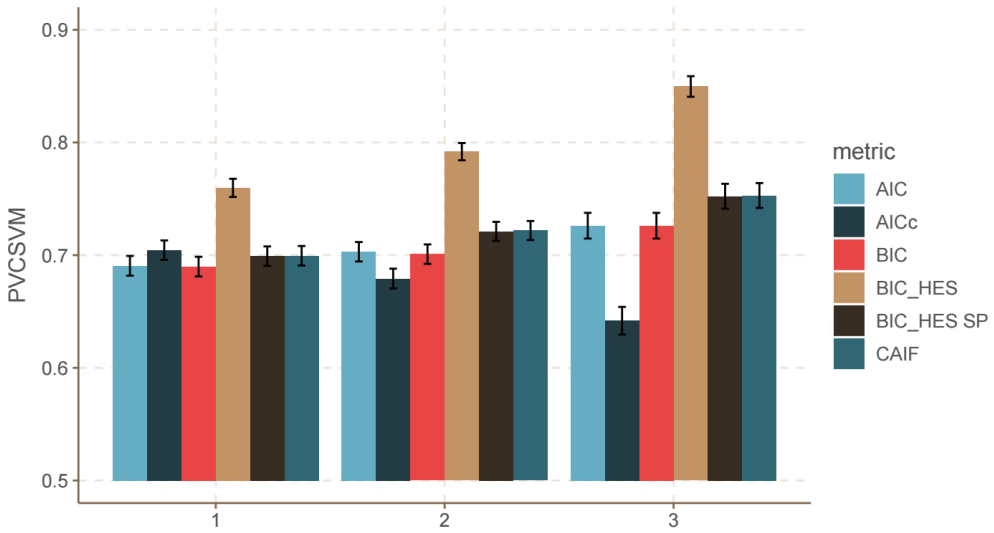


图10. 按新增随机效应噪声变量数量划分的信息准则。

模拟方案的结构如下：

(1) 数据模拟：针对每种模型类型，按照预先设定的参数组合生成合成数据集，从而为评估选择准则提供可控环境。**(2) 拟合真实模型：**对理论模型（高斯、逻辑斯蒂、泊松）进行拟合，并应用多种准则：贝叶斯信息准则、赤池信息准则、修正赤池信息准则、BIC_HES 和 BIC_HES_SP。**(3) 拟合设定错误的模型：**同时拟合备择模型，以考察这些准则能否区分正确设定与错误设定。

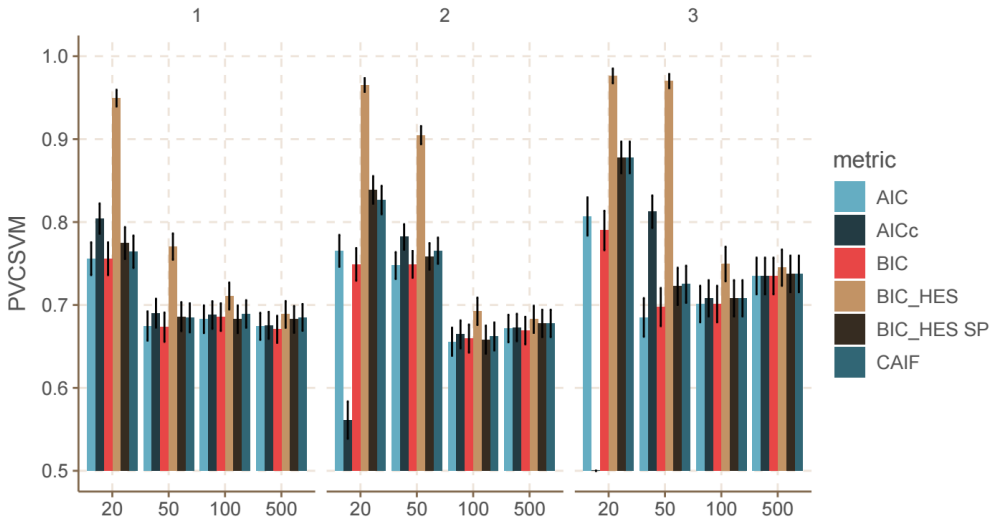


图11. 按样本量和新增随机效应噪声变量数量划分的信息准则

(4) **准则比较：**比较正确模型（M1）和错误模型（M2）的 BIC_HES 值，以确定该准则识别正确模型的能力。(5) **判定规则：**• 如果 BIC_HES$_{M2}$ >BIC_HES$_{M1}$，则该准则未能检测出真实模型。• 如果 BIC_HES$_{M2}$ ≤BIC_HES$_{M1}$，则该准则成功识别出真实模型。(6) **重复实验：**针对全部三种模型族，在多种参数组合下重复该过程，以确保结果在不同配置下均具有稳健性。

使用 R 4.3 版本进行了模拟。2 [19] 实验在一台配备 512 GB 内存、搭载 16 个主频为 2.2 GHz 的 Intel Xeon Gold 5320 处理器的 Linux 服务器上进行。分析中使用了 `GLMMadaptive` 软件包 [21]，该软件包用于估计分组数据的混合效应模型，并提供本分析所用参数的海森矩阵。共进行了 20 次重复试验，生成并评估了 124,273 个模型。

图12 展示了各准则选择真实模型的比例。BIC、BIC_HESS 和 BIC_HESS_SP 在识别二项和泊松混合模型中的真实模型方面效果最佳，这归功于其恰当的复杂度惩罚以及对随机效应的有效处理。相比之下，赤池信息准则、修正赤池信息准则、CAIF 和 ICOMP 的表现波动较大，且往往较差，这凸显了在这些情境下使用能够恰当平衡拟合优度与模型复杂度的准则的必要性。

图 13 表明，在高斯、二项和泊松混合模型模拟中，无论随机效应之间的相关性如何，BIC 和 BIC_HES 都是最稳健的准则。尽管 AIC 和修正赤池信息准则在某些设定下表现出竞争力，但它们对极端相关性值更为敏感（ρ）。CAIF 和 ICOMP 得出的结果一致性较低，这表明当随机效应之间的相关性较强时，它们的适用性有限。这些发现强调，在混合效应建模中，应采用能够有效应对相关随机效应所带来额外复杂性的信息准则。

6. 证明

本节给出推论 3.1和定理3.2的证明。

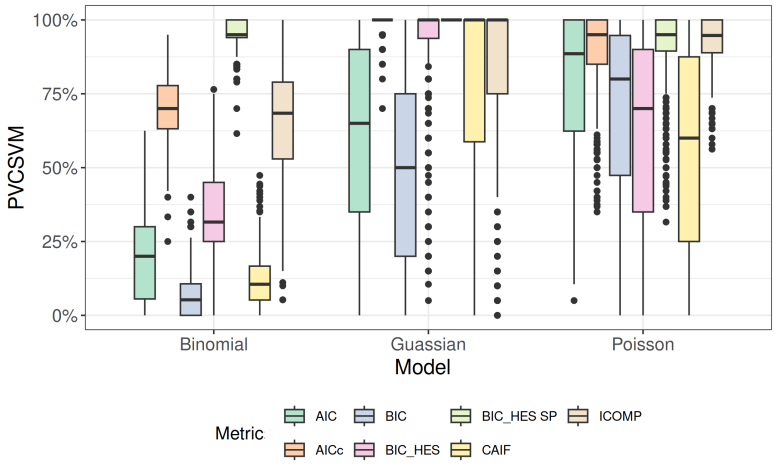


图 12。二项与泊松模型的信息准则

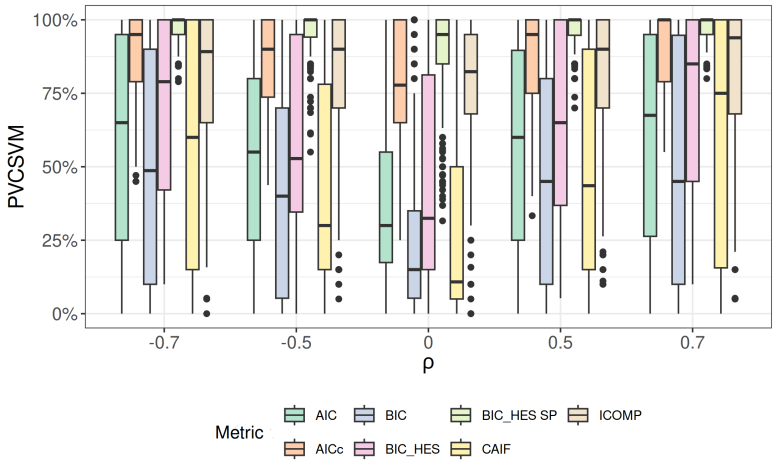


图 13。随机效应之间不同相关性下的选择百分比

推论 *3.1*的证明。

$$\begin{aligned}\mathbb{P}(y|H_i) &= \int_{\Theta_i} \mathbb{P}(y|H_i,\theta)g_i(\theta)d\theta \\ &= \int_{\Theta_i} f_i(y|\theta)g_i(\theta)d\theta \\ &= \int_{\Theta_i} \exp\left(\log\left(f_i(y|\theta)g_i(\theta)\right)\right)d\theta\end{aligned}$$

现在，我们可以在 $\tilde{\theta}$ 处对 $\log\left(f_i(y|\theta)g_i(\theta)\right)$ 进行展开。因此，可以得到以下关于 $\theta = \theta_i$ 的二阶展开~:

$$\log\left(f_i(y|\theta)g_i(\theta)\right) = \log\left(f_i(y|\tilde{\theta}_i)g_i(\tilde{\theta}_i)\right) + (\theta-\theta_i)\cdot\nabla Q(\tilde{\theta}_i) + \\ + \frac{1}{2}(\theta-\theta_i)\nabla^2 Q(\tilde{\theta}_i)\cdot(\theta-\theta_i)^t$$

其中， ∇Q 和 ∇^2 分别是 Q 在 $\tilde{\theta}_i$ 处的梯度和海森矩阵，且 $Q = \log\left(f_i(y|\theta)g_i(\theta)\right)$。因此，

$$\begin{aligned}\mathbb{P}(y|H_i) &= \int_{\Theta_i} \exp\left(Q(\tilde{\theta}_i) + (\theta-\theta_i)\cdot\nabla Q(\tilde{\theta}_i) + \frac{1}{2}(\theta-\theta_i)\nabla^2 Q(\tilde{\theta}_i)\cdot(\theta-\theta_i)^t\right)d\theta \\ &= \int_{\Theta_i} \exp\left(Q(\tilde{\theta}_i) + \frac{1}{2}(\theta-\theta_i)\nabla^2 Q(\tilde{\theta}_i)\cdot(\theta-\theta_i)^t\right)d\theta\end{aligned}$$

其中，最后一行用到了 $\tilde{\theta}_i$ 是 Q的驻点。因此，

$$\mathbb{P}(y|H_i) = \exp(Q(\tilde{\theta}_i))\int_{\Theta_i}\exp\left(\frac{1}{2}x\nabla^2 Q(\tilde{\theta}_i)\cdot x^t\right)dx.$$

现在引入

$$P_i = -\nabla^2 Q(\tilde{\theta}_i). \tag{6.1}$$

现在写作

$$\mathbb{P}(y|H_i) = \exp(Q(\tilde{\theta}_i))\int_{\Theta_i}\exp\left(-\frac{1}{2}xP_i\cdot x^t\right)dx.$$

˜由于 θ_i 是 Q的极大点，因此 P_i 是正定的，我们可以进行奇异值分解。也就是说，

$$P_i = U\Sigma_i U^t$$

其中 U 是一个正交矩阵。因此，

$$\begin{aligned}\mathbb{P}(y|H_i) &= \exp(Q(\tilde{\theta}_i))\int_{U\Theta_i}\exp\left(-\frac{1}{2}y\Sigma_i\cdot y^t\right)dy \\ &= \exp(Q(\tilde{\theta}_i))\int_{\Theta_i}\exp\left(-\frac{1}{2}y\Sigma_i\cdot y^t\right)dy.\end{aligned}$$

由此，我们最终得到

$$\begin{aligned}\mathbb{P}(y|H_i) &= \exp(Q(\tilde{\theta}_i))\Pi_{j=1}^p\sqrt{\frac{2\pi}{\lambda_j}} \\ &= f_i(y|\tilde{\theta}_i)g_i(\tilde{\theta}_i)\Pi_{j=1}^p\sqrt{\frac{2\pi}{\lambda_j}} \\ &= f_i(y|\tilde{\theta}_i)g_i(\tilde{\theta}_i)\frac{(2\pi)^{p/2}}{\det(P_i)^{1/2}}.\end{aligned}$$

其中（λ_j）是 P_i的特征值。 □

定理证明 *3.2*。 公式（3.2）直接由推论3.1得出，而最后一个公式（3.3）只需进行少量简单计算即可得到。事实上，为简便起见，我们记

$$\log(L_i) = \log(f_i(y|\tilde{\theta}_i)g_i(\tilde{\theta}_i)), \quad \text{for } i = 0, 1.$$

由于 $|\theta_1| = |\theta_0|$ ，我们立即得到

$$\begin{aligned}2\log(\mathrm{BF}_{0,1}) =& 2\log(L_0) - 2\log(L_1) + |\theta_1|\log(m) - |\theta_0|\log(m) \\ &+ \log(\det(P_1)) - \log(\det(P_0))\end{aligned}$$

其中， m 是 $i=0$ 中 Θ_i 的观测值数量，1。由此，根据 BIC_HES 的定义，我们得到

$$\begin{aligned} 2\log(\mathrm{BF}_{0,1}) &= \mathrm{BIC_HES}(H_1) - \mathrm{BIC_HES}(H_0) \\ &= \mathrm{BIC_HES}(\Delta_{1,0}). \end{aligned}$$

□

最后，我们给出定理 3.3的证明。

证明。回顾一下，如果给定一个真实数据生成模型 M_0（具有 p_0 个参数）和一个过拟合竞争模型 M_1（具有 p_1 个参数， $p_1 > p_0$ 且满足$M_0 \subset M_1$），随着样本量增加，选择真实模型的概率趋近于 1，那么模型选择准则 IC 就是一致的:

$$\lim_{n\to\infty} \mathbb{P}(IC(M_0) < IC(M_1)) = 1$$

为了证明一致性，我们必须证明，当 $n \to \infty$时，差值 Δ贝叶斯信息准则_HES $=$ 贝叶斯信息准则_HES(M_1) −贝叶斯信息准则_HES(M_0) 为正。我们考察过拟合模型 M_1与真实模型 M_0之间的准则值之差:

$$\Delta\mathrm{BIC_HES} = \mathrm{BIC_HES}(M_1) - \mathrm{BIC_HES}(M_0).$$

代入贝叶斯信息准则_HES 的定义:

$$\begin{aligned} \Delta\mathrm{BIC_HES} &= [-2\log L_1 + p_1\log n + \log|J_1|] - [-2\log L_0 + p_0\log n + \log|J_0|] \\ &= \underbrace{-[2(\log L_1 - \log L_0)]}_{\text{Likelihood Ratio Term }(\Lambda)} + \underbrace{(p_1-p_0)\log n}_{\text{Penalty Term}} + \underbrace{(\log|J_1| - \log|J_0|)}_{\text{Hessian Difference Term }(\Delta H)} \end{aligned}$$

我们分析各项的渐近阶，当 $n \to \infty$时。似然比统计量定义为

$$\Lambda = 2(\log L_1 - \log L_0).$$

在拉普拉斯近似下，将较大模型的参数写作 $\theta_1 = (\theta_0, \eta)$，其中 $\eta \in \mathbb{R}^{p_1-p_0}$ 为额外参数。对数似然在 $(\theta_0, 0)$ 附近作二阶泰勒展开，得到

$$\log L_1(\hat{\theta}_1) = \log L_1(\theta_0, 0) + (\hat{\eta})^\top \frac{\partial \log L_1}{\partial \eta}\Big|_{(\theta_0,0)} + \frac{1}{2}\hat{\eta}^\top J_{\eta\eta}\hat{\eta} + o_p(\|\hat{\eta}\|_2^2),$$

其中， $J_{\eta\eta}$ 是与额外参数 η 对应的海森矩阵块，余项满足 $o_p(\|\hat{}\eta\|_2^2) \to 0$ 依概率以快于 $\|\hat{}\eta\|_2^2$的速度趋于零，而 $\hat{}\eta$ 是额外参数的最大似然估计量。由于 θ_0使真实模型 M_0的似然达到最大值，因此关于原始参数的得分为零:

$$\frac{\partial \log L_1}{\partial \theta}\Big|_{(\theta_0,0)} = 0.$$

根据 θ_0 和 η对海森矩阵及参数差进行分块，二次项渐近地化简为

$$\Lambda = 2(\log L_1 - \log L_0) = \hat{\eta}^\top J_{\eta\eta}\hat{\eta} + + o_p(\|\hat{\eta}\|_2^2),$$

其中， $J_{\eta\eta} = -\underbrace{\frac{\partial^2}{\partial\eta\,\partial\eta^\top}}_{\text{对数}} L_1\big|_{(\theta_0,\ 0)}$是与额外参数对应的海森矩阵块。注意到

$$\sqrt{n}\hat{\eta} \xrightarrow{d} \mathcal{N}(0, I_{\eta\eta}^{-1}),$$

其中

$$I_{\eta\eta} = \mathbb{E}\Big[-\frac{\partial^2 \log L_1}{\partial\eta\,\partial\eta^\top}\big|_{(\theta_0,0)}\Big]$$

is t额外参数的费 Fisher 信息矩阵。因此，似然比统计量满足 s

$$\Lambda = 2(\log L_1 - \log L_0) = \hat{\eta}^\top J_{\eta\eta}\hat{\eta} + o_p(\|\hat{\eta}\|_2^2) \xrightarrow{d} \chi^2_{p_1-p_0}.$$

现在，回顾一下，若对每个 $\varepsilon > 0$ 都存在一个常数 $M > 0$，使得.....则称随机变量序列 X_n 为随机有界，记作 $X_n = O_p(1)$

$$\sup_n \mathbb{P}(|X_n| > M) < \varepsilon.$$

由于 Λ 依分布收敛于一个以概率 1 为有限值的卡方变量，我们可以选择 M 足够大，使得对于所有充分大的 n，$\mathbb{P}(\Lambda > M) < \varepsilon$。因此，根据定义，

$$\Lambda = O_p(1).$$

复杂度惩罚项为 $(p_1 - p_0)\log n$。由于 M_1 是一个过拟合模型，我们有 $p_1 > p_0$。于是

$$\lim_{n\to\infty} (p_1 - p_0)\log n = \infty.$$

该项的阶为 $O(\log n)$。

海森矩阵差项为 $\Delta H = \log|J_1| - \log|J_0|$。基于边际似然的拉普拉斯展开，e
mar项 $\log|J|$ 体现了参数依赖性，并且相对于 n为 $O_p(1)$。o

$$\Delta H = O_p(1).$$

将渐近阶代回 ΔBIC_HES 的表达式中：

$$\Delta\text{BIC_HES} = \underbrace{-O_p(1)}_{\text{Likelihood Gain}} + \underbrace{O(\log n)}_{\text{Complexity Penalty}} + \underbrace{O_p(1)}_{\text{Hessian Difference}}$$

由于 $n \to \infty$，项 $O(\log n)$ 占主导地位，超过 $O_p(1)$ 项。注意，

$$\Delta\text{BIC_HES} \approx (p_1 - p_0)\log n + O_p(1).$$

由于 $p_1 - p_0 > 0$，主导项确保 ΔBIC_HES 趋于正无穷。

因此，较大模型的准则 M_1 大于 e
true 模型 M_0 的概率趋近于 1。也就是说，

$$\lim_{n\to\infty} \mathbb{P}(\Delta\text{BIC_HES} > 0) = 1.$$

□

7. 结论

本文提出了一种扰动贝叶斯信息准则（BIC_HES），通过纳入对数似然函数海森矩阵的对数行列式，改进了传统的模型选择方法。这一修正引入了一种依赖于数据的复杂度惩罚，为模型比较提供了更具适应性且能体现几何信息的方法。通过对线性、混合、二项和泊松模型进行大量模拟，我们证明，BIC_HES 的表现始终优于贝叶斯信息准则、赤池信息准则和修正赤池信息准则等经典准则，尤其是在小样本情形以及存在噪声变量时。该准则具有很强的稳健性，能够更准确地识别真实的数据生成模型。从实践角度来看，BIC_HES 兼容标准统计软件，可利用后验抽样或观测费希尔信息矩阵进行实现。未来研究可探索将其扩展至高维情形、正则化贝叶斯框架和非参数模型。总体而言，BIC_HES 是贝叶斯模型选择领域一项前景广阔的进展，兼具理论严谨性与实践适用性。

致谢

AMH 获得了 PRIN 2022 项目的资助。该项目名为“理解量子神经网络学习过程（LeQun）”，项目申请代码为 2022WHZ5XH – CUP J53D23003890006。作者 AMH 是“国家数学分析、概率及其应用小组（GNAMPA）”的成员，该小组隶属于“弗朗切斯科·塞韦里”国家高等数学研究所（INdAM）。

遵守伦理标准

- 潜在利益冲突披露：作者确认不存在利益冲突。
- 涉及人类参与者和/或动物的研究：作者确认，本文研究不涉及将人类参与者和/或动物作为数据集。

参考文献

[1] *H. Akaike*，信息论与最大似然原理的扩展，载于《第二届国际信息论研讨会论文集》，*B. N. Petrov* 和 *F. Csaki* 编，布达佩斯，*1973* 年，*Akademiai Kiado*，第 *267–281* 页。[2] *H. Akaike*，信息论与最大似然原理的扩展，载于《*Hirotugu Akaike* 论文选集》，施普林格，*1998* 年，第 *199–213* 页。[3] *S.-i. Amari*，《信息几何及其应用》，第 *194* 卷，施普林格，*2016* 年。[4] *H. Bozdogan*，模型选择与赤池信息准则（*AIC*）：一般理论及其解析扩展，《心理计量学》，*52*（*1987*），第 *345–370* 页。[5] ，*ICOMP*：一种新的模型选择准则，载于《分类及相关数据分析方法》，*H. H. Bock* 编，爱思唯尔科学出版社（北荷兰），阿姆斯特丹，*1988* 年，第 *599–608* 页。[6] *L. M. DeBruine* 和 *D. J. Barr*，通过数据模拟理解混合效应模型，《心理科学方法与实践进展》，*4*（*2021*），第 *2515245920965119* 页。[7] *T. J. Faulkenberry*，计算贝叶斯因子以衡量实验所得证据：*BIC* 近似的扩展，*arXiv* 预印本 *arXiv:1803.00360*，（*2018*）。[8] *J. Gabry* 和 *B. Goodrich*，通过 *Stan* 进行贝叶斯应用回归建模，软件包*“rstanarm”*，（*2020*）。[9] *A. Gelman*，层级模型中方差参数的先验分布（对 *Browne* 和 *Draper* 文章的评论），（*2006*）。[10] *A. Gelman*、*J. Hwang* 和 *A. Vehtari*，理解贝叶斯模型的预测信息准则，《统计与计算》，*24*（*2014*），第 *997–1016* 页。[11] *A. Gelman* 和 *A. Vehtari*，过去 *50* 年最重要的统计学思想是什么？《美国统计学会杂志》，*116*（*2021*），第 *2087–2097* 页。[12] *L. Held* 和 *D. S. Bové*，《应用统计推断》，施普林格，柏林、海德堡，*doi*，*10*（*2014*），第 *16* 页。[13] *C. M. Hurvich* 和 *C.-L. Tsai*，小样本中的回归与时间序列模型选择，《生物计量学》，*76*（*1989*），第 *297–307* 页。[14] *M. J. Lindstrom* 和 *D. M. Bates*，用于重复测量数据的非线性混合效应模型，《生物计量学》，（*1990*），第 *673–687* 页。[15] *A. A. Neath* 和 *J. E. Cavanaugh*，使用施瓦茨信息准则变体进行回归与时间序列模型选择，《统计学通讯：理论与方法》，*26*（*1997*），第 *559–580* 页。[16] *Y. Pawitan*，《依循似然：使用似然进行统计建模与推断》，牛津大学出版社，*2001* 年。[17] *M.-E. Pérez*、*L. R. Pericchi* 和 *I. C. Ramírez*，作为尺度参数稳健先验的缩放贝塔平方分布，（*2017*）。[18] *J. Piironen* 和 *A. Vehtari*，贝叶斯预测方法在模型选择中的比较，《统计与计算》，*27*（*2017*），第 *711–735* 页。

[19] *R* 核心团队，*R*：统计计算的语言与环境，*R* 统计计算基金会，维也纳，奥地利，*2022* 年。[20] *A. E. Raftery*，社会研究中的贝叶斯模型选择，《社会学方法论》，（*1995*），第 *111–163* 页。[21] *D. Rizopoulos*，*GLMMadaptive*：使用自适应高斯求积的广义线性混合模型，*2023* 年。*https://drizopoulos.github.io/GLMMadaptive/*，*https://github.com/drizopoulos/GLMMadaptive*。[22] *J. N. Rouder*、*R. D. Morey*、*P. L. Speckman* 和 *J. M. Province*，方差分析设计的默认贝叶斯因子，《数学心理学杂志》，*56*（*2012*），第 *356–374* 页。[23] *D. Spiegelhalter*、*A. Thomas*、*N. Best* 和 *W. Gilks*，*Bugs 0.5*：使用吉布斯抽样进行贝叶斯推断手册（第二版），*MRC* 生物统计学部、公共卫生研究所，剑桥，英国，（*1996*），第 *1–59* 页。[24] *D. J. Spiegelhalter*、*N. G. Best*、*B. P. Carlin* 和 *A. Linde*，偏差信息准则：十二年回顾，英国皇家统计学会会刊 *B* 辑：统计方法学，*76*（*2014*），第 *485–493* 页。[25] *J. van Doorn*、*F. Aust*、*J. M. Haaf*、*A. M. Stefan* 和 *E.-J. Wagenmakers*，混合模型的贝叶斯因子，《计算脑与行为》，（*2021*），第 *1–13* 页。[26] *A. Vehtari*、*A. Gelman* 和 *J. Gabry*，使用留一交叉验证和 *WAIC* 进行实用的贝叶斯模型评估，《统计与计算》，*27*（*2017*），第 *1413–1432* 页。[27] *A. Vehtari* 和 *J. Lampinen*，使用交叉验证预测密度进行贝叶斯模型评估与比较，《神经计算》，*14*（*2002*），第 *2439–2468* 页。[28] *E.-J. Wagenmakers*，普遍存在的 *p* 值问题的实用解决方案，《心理学通报与评论》，*14*（*2007*），第 *779–804* 页。[29] *F. V. P. Wallas S. Nascimento*，量子力学背景下香农熵的研究：自由与受限谐振子的应用，《*Química Nova*》，*39*（*2016*），第 *757–764* 页。[30] *S. Watanabe* 和 *M. Opper*，奇异学习理论中贝叶斯交叉验证与广泛适用信息准则的渐近等价性，《机器学习研究杂志》，*11*（*2010*）。
（*D. D. J. Ramirez Ramirez*）哥伦比亚国立大学数学系，马尼萨莱斯校区，马尼萨莱斯，哥伦比亚。电子邮箱：ddramirezra@unal.edu.co（*A. Melchor Hernandez*）博洛尼亚大学数学系，圣多纳托门广场 *5* 号，*40126*，博洛尼亚（意大利）。电子邮箱：anderson.melchor@unibo.it（*Isabel Cristina Ramírez*）哥伦比亚国立大学数学系，麦德林校区，麦德林，哥伦比亚。电子邮箱：iscramirezgu@unal.edu.co（*Luis Raúl Pericchi*）波多黎各大学，里奥彼德拉斯校区，圣胡安，波多黎各。电子邮箱：luis.pericchi@upr.edu