E值：具有指数族的 k样本检验

Yunda Hao PeterGr¨unwald† Tyron Lardy Long Long CWI* CWI和莱顿大学 莱顿大学 Yiyang

Reuben Adams 伦敦大学学院

2024年1月9日

arXiv:2303.00471v2 [stat.ME] 8 Jan 2024

摘要

我们构建并比较用于检验 k 个数据样本是否来自相同分布的e变量，备择假设是它们来自指数族中的不同元素。我们考察以下三类e变量：针对（1）同一指数族中的“较小”零假设，以及（2）“较大”的非参数零假设的GRO（增长率最优）e变量；此外还考察（3）通过对充分统计量之和进行条件化而得到的e变量。（2）和（3）均可高效计算，并分别将Turner等人[2021]和Wald [1947] 的思想从伯努利分布推广到一般指数族。我们从对数增长率的角度，对这些e变量进行了理论和基于模拟的比较，发现对于较小的效应，全部四个e变量的表现出人意料地相似；对于高斯位置族和泊松分布，e变量（1）和（3）重合；对于伯努利分布，（1）和（2）重合；但一般而言，在备择假设下，（2）还是（3）增长得更快取决于具体的分布族。我们还讨论了对（1）进行数值近似的算法。

1 引言

E变量（及其取值，即*e*值）是 p 值的一种替代方案，其本身更适用于可选停止和延续情形下的检验，也是任意时刻有效置信区间的基础，这类区间可以持续监测[Gr¨unwald 等，2023，Vovk 和 Wang，2021，Shafer，2021，Ramdas 等，2022，Henzi 和 Ziegel，2022，Gr¨unwald，2023]。尽管它们源自 H. Robbins 及其学生关于任意时刻有效检验的研究（例如 [Darling 和 Robbins，1967]），但直到最近，针对复合零假设的 E 变量才开始得到深入研究。E变量可以对应一种自然的最优性概念，称为 GRO（增长率最优性），由 Gr¨unwald 等人引入并作了详细研究 [2023]。在可选停止的背景下，GRO 可视为一致最强检验的类比。本文针对一个经典统计问题——参数k样本检验——构造 GRO 和近 GROe变量。Wald 的开创性工作已涉及这一方向 [1947]：正如我们

*数学与计算机科学中心（CWI）是荷兰国家数学与计算机科学研究所，位于阿姆斯特丹。

†通讯作者：pdg@cwi.nl

将在示例 1 中解释的那样，他针对两个比例的序贯检验所提出的序贯概率比检验，可以重新诠释为针对伯努利流的 E 值。Wald 的 E 值在 GRO 意义下并非最优——直到最近，Turner 等人才针对伯努利流推导出 GRO 版本 [2021], Turner 和 Gr¨unwald [2022a], ，但研究对象仍仅限于伯努利流。本文针对备择假设与一个任意但固定的指数族$\mathcal{M}$相关联的情形，构造 E 变量；其中每个 k 组的数据均依次从该指数族中的不同分布抽取。我们主要考虑针对零假设的检验，该零假设记为 $\mathcal{H}_0(\mathcal{M})$，其内容是所有组中的结果均独立同分布，且来自 $\mathcal{M}$中的同一个成员。我们为这一零假设构造 GRO E 变量 $S_{\mathrm{GRO}(\mathcal{M})}$ ，但一般而言，该变量无法高效计算。因此，我们另外引入两个更易处理的 E 变量： $S_{\mathrm{GRO}(\mathrm{IID})}$ 和 S_{COND}。前者定义为针对一个范围大得多的零假设的 GRO E 变量；该零假设认为 k 组独立同分布，且来自一个任意分布，记为 $\mathcal{H}_0(\mathrm{IID})$：由于相对于某个零假设 $\mathcal{H}_0$ 的 E 变量，自动也是相对于其任意子集零假设 $\mathcal{H}_0$的 E 变量，$S_{\mathrm{GRO}(\mathrm{IID})}$ 自然也是相对于 $\mathcal{H}_0(\mathcal{M})$的 E 变量。下文提到“零假设”时，均指较小的 $\mathcal{H}_0(\mathcal{M})$。对于这一零假设，使用 $S_{\mathrm{GRO}(\mathrm{IID})}$ 而非 $S_{\mathrm{GRO}(\mathcal{M})}$ （前者并非 GRO）是合理的，因为它更易计算，且能稳健应对模型 $\mathcal{M}$设定错误。然而，正是这种稳健性，也可能导致当 $\mathcal{M}$ 设定正确时，该变量过于保守。我们考虑的第三个 E 变量 S_{COND}不具备任何 GRO 属性，但它专门针对 $\mathcal{H}_0(\mathcal{M})$设计，因此在实践中仍可能优于 $S_{\mathrm{GRO}(\mathrm{IID})}$ 。最后，我们引入一个伪 E 变量 $S_{\mathrm{PSEUDO}(\mathcal{M})}$；当后者易于计算时，它与 $S_{\mathrm{GRO}(\mathcal{M})}$ 一致；在其他情况下，它并非真正的 E 变量，但对我们的理论分析仍然非常有用。

结果 除了定义 $S_{\mathrm{GRO}(\mathcal{M})}$、 $S_{\mathrm{GRO}(\mathrm{IID})}$ 和 S_{COND} 并证明它们确实达到了其宣称的效果外，我们还通过理论分析和模拟研究了它们的行为。我们的主要理论结果——定理 2 和定理 3——揭示了一些令人意外的事实：对于任意指数族，只要零假设与备择假设之间的“距离”足够小，四种类型的（伪）e变量在备择假设下几乎具有相同的增长率，因此效果几乎同样好。具体而言，设备择假设的k 维参数与零假设的参数空间之间的（最短） ℓ_2距离为 δ。那么，对于上述任意两个e变量 S, S'，都有 $\mathbb{E}[\log S - \log S'] = O(\delta^4)$，其中期望是在备择假设下计算的。这里， $\mathbb{E}[\log S]$ 可解释为 S的增长率，详见第 1.1 节。

尽管对于我们所考虑的分布族， $S_{\mathrm{GRO}(\mathrm{IID})}$ 和 S_{COND} 都可以高效计算，但 $S_{\mathrm{GRO}(\mathcal{M})}$通常并非如此，因为要计算它，我们需要能够得到固定简单备择假设到集合 $\mathcal{H}_0(\mathcal{M})$的反向信息投影 （RIPr； [Li，1999年，Gr¨unwald 等， 2023]）。一般来说，这是 $\mathcal{H}_0(\mathcal{M})$中各元素的凸组合，只能通过数值方法找到。有趣的是，我们发现，对于三个分布族——固定方差高斯分布、伯努利分布和泊松分布——RIPr 在单个点处取得（即一种将全部质量集中在该点上的混合），且该点可以高效计算。此外，在这些情况下， $S_{\mathrm{GRO}(\mathcal{M})}$ 与其他e变量中的一个相一致（伯努利分布对应$S_{\mathrm{GRO}(\mathrm{IID})}$ ，高斯分布和泊松分布对应 S_{COND}）。对于其他指数族，以 $k = 2$为例，我们分别使用 Li（1999年）提出的算法和暴力法来近似 RIPr，进而近似 $S_{\mathrm{GRO}(\mathcal{M})}$ 。我们发现，仅使用由两个 分量构成的混合，就已经可以获得 RIPr 的极高精度近似。这使我们猜想，偏离 RIPr 或许仅仅是由数值误差造成的，而实际的 RIPr 确实可以表示为

仅包含两个分量。尽管这种发展的理论意义不容忽视，但在实际应用中，只要 RIPr 需要多个分量，我们建议使用 S_{COND} 或 $S_{\mathrm{GRO(IID)}}$，而不要使用 $S_{\mathrm{GRO}(\mathcal{M})}$，因为它们的增长率并不会差太多，而且计算起来要容易得多。此外，如果还要求对零假设错设具有稳健性，那么 $S_{\mathrm{GRO(IID)}}$ 是最合理的选择。

方法：限制于单块和简单备择假设e变量主要用于分析序贯任意时刻有效设置：数据以数据流的形式到达，对应于各个组；我们可能希望随时停止或继续抽样（可选停止）；例如，只有当数据看起来足够理想时才停止；或者因为资金用尽、无法继续收集新数据而意外停止。尽管如此，本文关注的是单个数据块中的情形，即一个向量 $X^k = (X_1, \ldots\ldots, X_k)$，其中每个$X_j$ 表示第 j个数据流中的单个结果。目前已有多篇论文（例如 Gr¨unwald等人 [2023], Ramdas等人 [2022], Turner等人 [2021]）阐述了如何将针对此类单个数据块定义的e变量相乘组合，得到 e过程（在本文语境下，即非负超鞅），从而在数据块依次到达——也就是说，每次观测每个样本的一个结果——时，通过可选停止来检验零假设。简言之，将e变量相乘；在任何打算停止的时刻，如果迄今观测到的e值的乘积超过 $1/\alpha$（预先指定的显著性水平为 α），就拒绝零假设。这就构成了一个显著性水平为 α的任意时刻有效检验：无论采用何种停止规则，第一类错误都保证低于 α。类似地，可以通过反演此类检验来扩展该方法，以构造任意时刻有效置信区间，具体描述见Ramdas等人 [2022]。Turner和Gr¨unwald [2022a]针对伯努利数据的双样本检验采用了这一做法；他们的反演方法可扩展至本文讨论的一般指数族情形。因此，更多细节请参阅上述论文，本文则将讨论范围限定于单块情形。此外，Turner等人 [2021],Turner和Gr¨unwald [2022b] 描述了如何将针对以数据块形式到达的数据的 e过程调整为适用于一般数据流的情形，在这种情形下，各个 k 数据流产生数据点的速率并不相同；我们在附录A中简要地将他们的说明扩展至本文的设置。最后，我们主要将讨论范围限定于简单备择假设的情形，即所考虑指数族中的一个成员。虽然这看起来似乎是一个很大的限制，但从简单备择假设扩展到复合备择假设（例如，所考虑的整个族）很容易，只需使用混合方法（即随时间对备择假设进行贝叶斯学习）和/或代入法即可。详细说明同样请参阅Gr¨unwald等人 [2023], Ramdas等人 [2022] ；Turner等人 [2021]则说明了双样本伯努利情形。本文将讨论范围限定于简单备择假设情形：所有“真正的”难点都在于处理复合零假设，而本文对此进行了明确且全面的讨论。

相关工作与实际意义 如前所述，本文是对 Turner等人 [2021], Turner 和 Gr¨unwald [2022a] 关于伯努利流双样本检验的论文，以及 Wald 的（1947年）比例序贯双样本检验的直接（但影响深远的）扩展，将研究对象拓展到来自指数族的数据流。此外，还有非参数序贯 [Lh´eritier 和 Cazals 2018]提出的方法，以及任意时刻有效的双样本检验 [Balsubramani 和 Ramdas（2016年）、Pandeva 等人 2022]提出的方法，它们处理的是一个略有不同的问题。这些方法对备择假设的要求弱得多（在某些版本中，样本可以是任意高维对象，例如图片等）。代价是

在检测到差异之前，它们需要大得多的样本量。事实上，尽管我们的主要兴趣在于理论问题（不同的e值如何比较？它们在何种意义上是最优的？），但在数据成本高昂的场景中，例如随机临床试验，我们在此介绍的方法在实践中可能非常有用：它们是精确的（现有方法通常基于卡方检验，在小样本量下无法提供精确的第一类错误率保证），允许可选停止，而且由于对备择假设作出了较强的参数假设，所需数据量较少。为简单说明这些性质在实践中的重要性，我们提及近期的 SWEPIS研究 [Wennerholm等人 2019]，该研究因发现有害影响而提前终止。正如 Turner等人所示， [2021], 如果该研究采用了任意时刻有效的双样本检验，就能得出有力得多的结论。

我们还要指出， k-样本检验可以视为独立性检验（结果是否独立于其所属的组？）。因此，本文也与近期关于e值及条件独立性检验任意时刻有效检验的论文 [Gr¨unwald等人，2022年，Shaer等人，2022年，Duan等人， 2022]有关。不过，这些论文研究的情形与本文相当不同，因为它们假设协变量（即指示数据所属 k 组的变量）是独立同分布的。

目录 在本引言余下部分，我们将确定一般框架和符号，并简要回顾e变量如何用于任意时刻有效检验／可选停止情形。第2节将详细介绍我们的四种（伪）e变量，并给出刻画其增长率行为的初步结果。第3节将给出主要理论结果，表明对于所有正则指数族，当备择假设的参数与零假设的参数空间相距 δ 个 ℓ_2距离时，四类e变量的期望增长都小得出奇，量级为 δ^4。第4节将对大量标准指数族（高斯分布、伯努利分布、泊松分布、指数分布、几何分布、贝塔分布）进行更详细的比较，并通过模拟展示 δ增大时会发生什么。第 5 节将提供关于 RIPr 的一些补充模拟。所有证明及一些补充模拟均见附录。

1.1 正式设定

考虑一个采用均值参数化的一维正则指数族 $\mathcal{M} = \{P_\mu : \mu \in \mathrm{M}\}$ （有关定义以及下文将要使用的所有关于指数族的标准结果的证明，参见例如 [Barndorff-Nielsen 1978]）。该族中的每个成员都是某个随机变量 U 的分布，该随机变量取值于某个集合 $\mathcal{U}$，并且相对于某个基础测度 $\rho_{[U]}$ 具有密度 $p_{\mu;[U]}$ ；不失一般性地，该基础测度可以取为概率测度。对于正则指数族，M 是 $\mathbb{R}$ 中的一个开区间，而 $p_{\mu;[U]}$ 可以写成：

$$p_{\mu;[U]}(U) = \exp\left(\lambda(\mu) \cdot t(U) - A(\lambda(\mu))\right), \tag{1.1}$$

其中， $\lambda(\mu)$ 将均值参数 μ 映射到规范参数 β。于是我们有 $\mu = \mathbb{E}_{P_\mu}[t(U)]$, 其中$t(U)$ 是 U 的可测函数，而 $A(\beta)$ 是对数归一化因子。测度 $\rho_{[U]}$诱导出相应的（边缘）测度 $\rho := \rho_{[X]}$，定义在充分统计量 $X := t(U)$ 上；同样，密度 (1.1) 也诱导出相应的密度 $p_\mu := p_{\mu;[X]}$，定义在 X 上，即我们有

$$p_\mu(X) := p_{\mu;[X]}(X) = \exp\left(\lambda(\mu) \cdot X - A(\lambda(\mu))\right). \tag{1.2}$$

我们将定义的所有e-变量都可以用 X的充分统计量所诱导的测度和密度来表示；换言之，不失一般性，我们可以假定我们的族是自然的。因此，从现在起，我们只假定观测数据以充分统计量 X 的形式给出，而非可能更细粒度的 U，并且不再提及 U；为简单起见，我们将 $p_{\mu;[X]}$ 简写为 p_μ ，将 $\rho_{[X]}$ 简写为 ρ。注意，指数族通常用载体函数 $h(X)$ 来定义，并将 ρ 设为勒贝格测度或计数测度；我们通过将 h 吸收到 ρ中来涵盖这种情形，而我们不要求后者是勒贝格测度或计数测度。

数据以数据块 $X^k = (X_1, \ldots\ldots, X_k) \in \mathcal{X}^k$的形式给出，其中 $\mathcal{X}$是 ρ的支集。要计算我们的e值，我们只需要知道 $X^k \in \mathcal{X}^k$；在备择假设下，所有 X_j、 $j = 1 \ldots k$都根据 $\mathcal{M}$中的某个元素 P_{μ_j}进行分布。在我们的主要结果中，我们将备择假设设为简单的，即假设 $\boldsymbol{\mu} = (\mu_1, \ldots\ldots, \mu_k) \in \mathrm{M}^k$ 预先固定。因此，备择假设可表示为

$$\text{simple } \mathcal{H}_1 : X_1 \sim P_{\mu_1}, X_2 \sim P_{\mu_2}, \ldots, X_k \sim P_{\mu_k} \text{ independent.}$$

请注意，在本节余下部分及第2节中，我们将始终固定 $\boldsymbol{\mu}$ 。这不失一般性，因为 $\boldsymbol{\mu}$ 被定义为 M^k中的任意元素，因此，针对 $\boldsymbol{\mu}$ 陈述的所有结果都适用于 M^k中的任意元素。通过混合方法或代入法扩展到复合备择假设是直接的，其处理方式已成为e值检验中的标准做法 [Ramdas等人， 2022]。

我们的零假设直接设为复合假设，因为就零假设而言，复合情形与简单情形本质上截然不同 [Ramdas 等，2022，Gr¨unwald 等人，2023]。它表示 X^k 同分布。我们将考虑这一零假设的若干不同变体，它们均为复合假设：设 $\mathcal{P}$ 是 $\mathcal{X}$上的一个分布集合，那么相对于 $\mathcal{P}$ 的零假设 *relative to* ，记作 $\mathcal{H}_0(\mathcal{P})$，定义为

$$\text{composite } \mathcal{H}_0(\mathcal{P}) : X_1 \sim P, X_2 \sim P, \ldots, X_k \sim P \text{ i.i.d. for some } P \in \mathcal{P}.$$

我们最重要的零假设实例是针对同一个指数族 $\mathcal{M}$ 的 $\mathcal{H}_0 = \mathcal{H}_0(\mathcal{M})$，备择假设即取自该指数族；此时 $\mathcal{H}_0(\mathcal{M})$是一个一维参数族，表示 X_i 独立同分布地来自P_{μ_0} ，其中 $\mu_0 \in \mathrm{M}$。不过，我们也会考虑 $\mathcal{H}_0 = \mathcal{H}_0(\mathcal{P})$，其中 $\mathcal{P}$ 是 $\mathcal{X}$上*all* 分布构成的更大集合。此时，零假设仅表示 X^k 独立同分布；我们将这一零假设简记为 $\mathcal{H}_0(\mathrm{IID})$。最后，我们有时还会考虑 $\mathcal{H}_0 = \mathcal{H}_0(\mathcal{M}')$，其中 $\mathcal{M}' \subset \mathcal{M}$ 是 $P_\mu \in \mathcal{M}$ 的一个子集，且对某个子区间 $\mathrm{M}' \subset \mathrm{M}$， $\mu \in \mathrm{M}'$。我们用来获取反对这些零假设的证据的统计量是e变量。

定义 1. 我们将样本空间 Ω （本文中该空间始终为 $\Omega = \mathcal{X}^k$）上的任意非负随机变量 S 称为相对于 $\mathcal{H}_0$ 的 ***e*变量** ，如果它满足

$$\textit{for all } P \in \mathcal{H}_0 : \qquad \mathbb{E}_P[S] \leq 1. \tag{1.3}$$

1.2 一般 $\mathcal{H}_0$ 的 GROE-变量

一般来说，对于上述任何一种零假设，都存在许多e变量可用于检验。每个e变量 S 又可以与一个增长率相关联，该增长率定义为 $\mathbb{E}_{P_\mu}[\log S]$。粗略地说，如果简单备择假设成立，那么在连续的独立实验中使用 S 并将结果相乘，可以将此增长率解释为所能达到的（渐近）指数增长率

（参见例如 [Gr¨unwald 等，2023，第2节.1] 或 [Kelly,1956]）。增长率最优（GRO）e变量，便是在所有e变量中增长率最高的e变量。Gr¨unwald 等人 [2023]的核心结果（定理1）指出，在非常弱的条件下，GROe变量的形式是备择假设与该备择假设在零假设上的反向信息投影 [Li 1999]之间的似然比。我们通过给出引理1和引理2，将他们的定理1应用于我们的情形；这两个引理都是其定理1的特例。在陈述这些引理之前，我们需要引入一些额外的符号和定义。对于 $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$，我们采用以下记号：

$$p_{\boldsymbol{\mu}}(X^k) := \prod_{i=1}^{k} p_{\mu_i}(X_i).$$

在本文中，每当我们提到 KL 散度 $D(Q\|R)$ 时，指的都是测度 Q 和 R 在$\mathcal{X}^k$上的 KL 散度。这里，Q 必须是概率测度，而 R 可以是次概率测度，正如 Gr¨unwald 等人 2023]所述。R 在 $\mathcal{X}^k$ 上的次概率测度，是指总质量不超过 1 的测度，即 $\int_{x\in\mathcal{X}} dR(x) \leq 1$。

以下引理是 Gr¨unwald 等人 [2023], 的定理1（最简版本）在应用于我们的 k-样本检验设定时的一个非常特殊的情形：

引理1。设 $\mathcal{P}$ 是 $\mathcal{X}^k$ 上的一组概率分布，$\text{CONV}(\mathcal{P})$ 是其凸包。那么，存在一个次概率测度 P_0^*，其密度为 p_0^*，使得

$$D(P_{\boldsymbol{\mu}}\|P_0^*) = \inf_{P\in\text{CONV}(\mathcal{P})} D(P_{\boldsymbol{\mu}}\|P). \tag{1.4}$$

P_0^* 称为 $P_{\boldsymbol{\mu}}$ 在 $\text{CONV}(\mathcal{P})$上的反向信息投影（RIPr）。

显然，如果 $P_0^* \in \text{CONV}(\mathcal{P})$（即最小值能够达到），那么 P_0^* 是一个概率测度，也就是说，其积分恰好为一。我们将在第4节中证明，某些特定的指数族确实具有这一性质。然而，一般而言，我们既不能期望最小值能够达到，也不能期望 RIPr 的积分为一。下面的引理2（同样是 [Gr¨unwald 等人 2023 年定理 1], 的一个特例）表明，RIPr 刻画了 GRO e变量，并解释了下文定义中使用术语 GRO 的原因。

定义2。$S_{\text{GRO}(\mathcal{P})}$ 定义为

$$S_{\text{GRO}(\mathcal{P})} := \frac{p_{\boldsymbol{\mu}}(X^k)}{p_0^*(X^k)} \tag{1.5}$$

其中，p_0^* 是 *RIPr* 从 $P_{\boldsymbol{\mu}}$ 到 $\text{CONV}(\mathcal{P})$ 的密度。

引理2。对于 $\mathcal{X}$ 上的每个分布集合 $\mathcal{P}$，$S_{\text{GRO}(\mathcal{P})}$ 都是 $\mathcal{H}_0(\mathcal{P})$ 的 ***e*变量**。此外，它还是 $\mathcal{H}_0(\mathcal{P})$ 的 ***GRO*（增长率最优）*e*变量**，即它本质上唯一地实现了

$$\sup_S \mathbb{E}_{P_{\boldsymbol{\mu}}}[\log S]$$

其中，上确界取遍所有适用于 $\mathcal{H}_0(\mathcal{P})$ 的 ***e*变量**。

这里，本质唯一性是指，任何其他 GROe变量在 $P_{\boldsymbol{\mu}}$ 下都必须以概率 1 等于 $S_{\text{GRO}(\mathcal{P})}$。这进而意味着测度 P_0^* 实际上是唯一的，因为正则指数族的成员必须具有全支撑。因此，一旦我们固定了备择假设，并将零假设定义为 $\mathcal{H}_0(\mathcal{P})$，其中 $\mathcal{P}$ 是 $\mathcal{X}$ 上的某个分布集合，那么要使用的最优（按 GRO 意义）e变量就是上文定义的 $S_{\text{GRO}(\mathcal{P})}$ e变量。

2 e变量的四种类型

本节中，我们定义四类e变量；这些定义可应用于任意底层单参数指数族。更确切地说，我们定义三个“实值”e变量 $S_{\mathrm{GRO}(\mathcal{M})}, S_{\mathrm{COND}}, S_{\mathrm{GRO}(\mathrm{IID})}$，以及一个“伪$e$变量” $S_{\mathrm{PSEUDO}(\mathcal{M})}$；它是 $S_{\mathrm{GRO}(\mathcal{M})}$ 的一种变体，对于某些指数族而言是e变量，而对于其他指数族则不是。

2.1 $\mathcal{H}_0$（$\mathcal{M}$）的 GRO e 变量与伪e变量

现在我们考虑我们主要关注的零假设 $\mathcal{H}_0(\mathcal{M})$ 对应的 GRO e变量。实际上，对于某些指数族 $\mathcal{M}$，式（1.4）中关于 $\mathrm{CONV}(\mathcal{M})$ 的下确界确实会在某个 $P_{\mu_0^*} \in \mathcal{M}$ 处取得。在这种简单情形下，我们可以解析地确定 $S_{\mathrm{GRO}(\mathcal{M})}$ （当 $S_{\mathrm{GRO}(\mathcal{M})} = S_{\mathrm{PSEUDO}(\mathcal{M})}$ 时会出现这种情况，见下文）。对于所有其他 $\mathcal{M}$，也就是说，当下确界根本无法取得，或在 $\mathrm{CONV}(\mathcal{M}) \setminus \mathcal{M}$ 中时，我们不知道能否解析地确定 $S_{\mathrm{GRO}(\mathcal{M})}$。在这种困难情形下，我们将按下文定义，通过 $S'_{\mathrm{GRO}(\mathcal{M})}$ 对其进行数值近似。首先，对于固定参数 $\mu_0 \in \mathrm{M}$，我们将向量 $\langle \mu_0 \rangle$ 定义为指示 $\mathcal{X}^k$ 上所有参数均等于 μ_0 时分布的向量：

$$\langle \mu_0 \rangle := (\mu_0, \ldots, \mu_0) \in \mathrm{M}^k \tag{2.1}$$

接下来，给定 W M 上的一个分布，我们定义

$$p_W := \int p_{\langle \mu_0 \rangle}(X^k) dW(\mu_0) \tag{2.2}$$

它是按照 W，对 $\mathcal{H}_0(\mathcal{M})$ 中的分布进行边际化后得到的贝叶斯边际密度。显然，如果 W 的支撑有限，那么相应的分布 P_W 具有$P_W \in \mathrm{CONV}(\mathcal{M})$。现在我们令

$$S'_{\mathrm{GRO}(\mathcal{M})} := \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{W'_0}(X^k)}$$

其中选择 W'_0，使得 $p_{W'_0}$在 (1.4) 中距离取得最小值仅差一个很小的 ϵ，即 $D(P_{\mu_1,\ldots,\mu_k} \| P'_{W_0}) = \inf_{P \in \mathrm{CONV}(\mathcal{M})} D(P_{\mu_1,\ldots,\mu_k} \| P) + \epsilon'$对某个 $0 \le \epsilon' < \epsilon$成立。于是，根据 Gr¨unwald 等人 [2023]的推论2，除非 $\epsilon' = 0$，否则 $S'_{\mathrm{GRO}(\mathcal{M})}$ not 是一个e-变量；不过，在每种情况下（即对于 $\mathcal{M}$ 的每种选择），我们都通过数值计算验证，当 δ 极小时，$\sup_{\mu_0 \in \mathrm{M}} \mathbb{E}_{P_{\mu_0,\ldots,\mu_0}}[S] = 1 + \delta$，也就是说，随着 ϵ' 趋近于 0，δ 会迅速趋近于 0。我们将在第 5 节详述计算过程。

现在我们来考虑这种“简单”情况：对于某个 $\mu_0^* \in \mathrm{M}$，$P_0^* = P_{\langle \mu_0^* \rangle}$。显然，必有 $\mu_0^* := \arg\min_{\mu_0 \in \mathrm{M}} D(P_{\boldsymbol{\mu}} \| P_{\langle \mu_0 \rangle})$。简单计算可得，此时

$$\mu_0^* = \frac{1}{k} \sum_{i=1}^{k} \mu_i. \tag{2.3}$$

定义 3. $S_{\mathrm{PSEUDO}(\mathcal{M})}$ 定义为

$$S_{\mathrm{PSEUDO}(\mathcal{M})} := \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0^* \rangle}(X^k)}.$$

$S_{\text{PSEUDO}(\mathcal{M})}$ 相对于 $\mathcal{H}_0(\mathcal{M})$ 并不总是一个真正的e-变量，这也解释了“伪”这个名称的由来。尽管如此，它在定理 2 和推导中仍将作为辅助工具发挥很大作用。请注意，如果它是一个e-变量，那么我们知道它等于 $S_{\text{GRO}(\mathcal{M})}$：

命题 1. $S_{\text{PSEUDO}(\mathcal{M})}$ 是关于 $\mathcal{M}$ 的 ***e*变量**，当且仅当 $S_{\text{PSEUDO}(\mathcal{M})} = S_{\text{GRO}(\mathcal{M})}$。

上述命题并未给出任何易于验证的条件，来判断$S_{\text{PSEUDO}(\mathcal{M})}$ 是否为e变量。下面的命题确实提供了一个有时容易检验的条件（我们将在下文中大量使用该条件）。令 μ_0^* 如式 (2.3) 所示，定义

$$f(\mu_0) := \sum_{i=1}^{k} \text{VAR}_{P_{\mu_i+\mu_0-\mu_0^*}}[X] - k\text{VAR}_{P_{\mu_0}}[X].$$

命题 2. 如果 $f(\mu_0^*) > 0$，那么 $S_{\text{PSEUDO}(\mathcal{M})}$ 不是 ***e*变量**。如果 $f(\mu_0^*) < 0$，那么存在一个区间 $\mathtt{M}' \subset \mathtt{M}$，使得 μ_0^* 位于$\mathtt{M}'$ 的内部，从而 $S_{\text{PSEUDO}(\mathcal{M})}$ 是关于$\mathcal{H}_0(\mathcal{M}')$的 ***e*变量**，其中 $\mathcal{M}' = \{P_\mu : \ \mu \in \mathtt{M}'\}$。

2.2 关于 $\mathcal{H}_0(\text{IID})$的 GROE-变量

回顾一下，我们将 $\mathcal{H}_0(\text{IID})$定义为这样的分布集合：在该集合下，$X_j$、$j = 1, \ldots k$独立同分布，且来自 $\mathcal{X}$上的某个任意分布。根据e变量的定义性质，即在零假设下期望值不超过 1（1.3），显然，关于 $\mathcal{H}_0(\text{IID})$的任何$e$变量也是关于 $\mathcal{H}_0(\mathcal{M})$的e变量，因为 $\mathcal{H}_0(\mathcal{M}) \subset \mathcal{H}_0(\text{IID})$。特别地，在指数族的情形下，我们也可以使用关于 $\mathcal{H}_0(\text{IID})$的 GRO$e$变量。事实证明，这个$e$变量（我们记作 $S_{\text{GRO}(\text{IID})}$）形式简单，通常易于计算。现在我们来证明这一点：

定理1。 引理*1*中所述的最小 *KL* 散度下确界$P \in \text{CONV}(\mathcal{H}_0(\text{IID}))\, D(P_{\boldsymbol{\mu}} \| P)$ 由分布 P_0^*在 $\mathcal{X}^k$上的密度达到

$$p_0^*(x^k) = \prod_{j=1}^{k} \frac{1}{k} \sum_{i=1}^{k} p_{\mu_i}(x_j).$$

因此，下面定义的 $S_{\text{GRO}(\text{IID})}$ 是 $\mathcal{H}_0(\text{IID})$的 ***GRO e*** 变量。

定义 4。 $S_{\text{GRO}(\text{IID})}$ 定义为

$$S_{\text{GRO}(\text{IID})} := \frac{p_{\boldsymbol{\mu}}(X^k)}{\prod_{j=1}^{k} \left(\frac{1}{k} \sum_{i=1}^{k} p_{\mu_i}(X_j) \right)}.$$

定理1的证明将 Turner等人在 [2021] 双样本伯努利情形中的论证推广到了 k样本的一般情形。证明中使用的论证实际上并不要求备择假设等于指数族中 k 个相互独立元素的乘积分布——它也可以由 k 任意分布的乘积给出。不过，我们仅针对前一种情形陈述该结果，因为这正是我们在本文中所关注的情形。

2.3 条件 e变量 S_{cond}

到目前为止，我们将e-变量定义为 $P_{\boldsymbol{\mu}}$ 与指数族中经过巧妙选择的元素之间的似然比，这些元素分别来自 $\mathcal{H}_0(\mathcal{M})$或 $\mathcal{H}_0(\mathrm{IID})$。现在我们采用不同的方法：不再考虑完整的原始数据 $X_1, \ldots X_k$，而是以充分统计量之和为条件，即$Z = \sum_{i=1}^{k} X_i$。事实证明，这样做实际上会将 $\mathcal{H}_0(\mathcal{M})$压缩为一个单一分布，从而使零假设变得简单。也就是说，在 $\mathcal{H}_0(\mathcal{M})$的所有元素下，$X^k \mid Z$ 的分布都相同，正如我们将在命题3中证明的那样。这意味着，我们不必使用原始数据的似然比，而可以使用以 Z为条件的似然比，这样便会“自动”得到一个e-变量。

定义5. 令 Z 为随机变量 $Z := \sum_{i=1}^{k} X_i$，S_{COND} 定义为

$$S_{\mathrm{COND}} := \frac{p_{\boldsymbol{\mu}}\left(X^{k-1} \mid Z\right)}{p_{\langle \mu_0 \rangle}\left(X^{k-1} \mid Z\right)},$$

其中 $\mu_0 \in \mathrm{M}$ 和 (X) 是式(1.2)中的充分统计量。

命题*3.* 对所有 $\boldsymbol{\mu}' = (\mu'_1, \ldots, \mu'_k) \in M^k$，我们有 $p_{\boldsymbol{\mu}'}(x^{k-1} \mid Z = z)$ 仅通过 $\lambda_j := \lambda(\mu'_j) - \lambda(\mu'_k)$，$j = 1, \ldots k-1$依赖于$\boldsymbol{\mu}'$，也就是说，它可以表示为 $(\lambda_1, \ldots, \lambda_{k-1})$ 的函数。作为一个特例，对于所有 $\mu_0, \mu'_0 \in M$，成立 $p_{\langle \mu_0 \rangle}(x^k \mid Z) = p_{\langle \mu'_0 \rangle}(x^k \mid Z)$。直接可得，$S_{\mathrm{COND}}$ 是 $\mathcal{H}_0(\mathcal{M})$ 的一个 e变量，

示例 1。 [**伯努利模型**] 若 $\mathcal{M}$ 为伯努利模型且 $k = 2$，则条件e变量可化简为 $(X_1, X_2) \in \{0,1\}^2$ 在其和为 $Z \in \{0,1,2\}$条件下的条件概率之比。显然，对于所有 $\mu'_1, \mu'_2 \in \mathrm{M} = (0, 1)$，都有 $p_{\mu'_1, \mu'_2}((0, 0) \mid Z = 0) = p_{\mu'_1, \mu'_2}((1, 1) \mid Z = 2) = 1$，因此当 $Z = 0$ 或 $Z = 2$ 时 $S_{\mathrm{COND}} = 1$，无论备择假设为何，两个组的结果相同的数据实际上都会被忽略。伯努利模型中 S_{COND} 的非序贯版本此前已由 Adams [2020]进行了非常详尽的分析。

此外，对于任意 $c \in \mathbb{R}$，我们有 $\mathrm{M}_c := \{(\mu'_1, \mu'_2) : \lambda(\mu_1) - \lambda(\mu_2) = c\}$ 是 M^2 内具有相同优势比 $\log(\mu_1(1-\mu_2)/((1-\mu_1)\mu_2)) = c$的分布所构成的直线。Wald [1947] 的两个比例的序贯概率比检验基于为备择假设固定一个 c（将其视为“效应量”的一种概念），并通过条件概率的乘积分析配对数据序列 $X_{(1)}, X_{(2)}$、……，其中 $X_{(i)} = (X_{i,1}, X_{i,2})$

$$\frac{p_c(X_{(i)} \mid Z_{(i)})}{p_0(X_{(i)} \mid Z_{(i)})} = S_{\mathrm{COND}}(X_i),$$

因此，实际上使用 S_{COND}（这里我们稍微滥用了记号，当我们指的是$p_{\mu'_1,\mu'_2}(x \mid z)$时，写作 $p_c(x \mid z)$，其中任意 $\mu'_1, \mu'_2 \in \mathrm{M}_c$）。不过，重要的是要注意，这个乘积并非用于任意时刻有效检验，而是用于一种经典序贯检验，该检验采用固定停止规则，且专门针对优化检验功效而设计。

3 我们的e变量的增长率比较

上文我们提供了几种构造 e-变量的方案 $S = S^{\mu}$，其定义隐含地依赖于所选的备择假设 $\boldsymbol{\mu}$。为比较这些方案，对于任意非负随机变量 S_1^{μ} 和 S_2^{μ}，我们定义 $S_1^{\mu} \succeq S_2^{\mu}$ 的含义为：对所有 $\boldsymbol{\mu} \in \mathrm{M}^k$，均有

满足 $\mathbb{E}_{P_{\boldsymbol{\mu}}}[\log\, S_1^{\boldsymbol{\mu}}] \geq \mathbb{E}_{P_{\boldsymbol{\mu}}}[\log S_2^{\boldsymbol{\mu}}]$。若 $S_1^{\boldsymbol{\mu}} \succ S_2^{\boldsymbol{\mu}}$，且存在 $\boldsymbol{\mu} \in \mathrm{M}^k$ 使得等式不成立，我们记作 $S_1^{\boldsymbol{\mu}} \succeq S_2$。从现在起，我们再次省略对 $\boldsymbol{\mu}$ 的依赖，即以 S 代替 $S^{\boldsymbol{\mu}}$。对于每个底层指数族 $\mathcal{M}$，我们显然有：

$$S_{\mathrm{PSEUDO}(\mathcal{M})} \succeq S_{\mathrm{GRO}(\mathcal{M})} \succeq S_{\mathrm{GRO(IID)}} \text{ and } S_{\mathrm{GRO}(\mathcal{M})} \succeq S_{\mathrm{COND}}. \tag{3.1}$$

下面我们给出定理 2 和定理 3（证明见附录）。这些结果超越了上述定性评估，通过数值方式界定了 $S_{\mathrm{PSEUDO}(\mathcal{M})}$ and $S_{\mathrm{GRO(IID)}}$ 与 $S_{\mathrm{PSEUDO}(\mathcal{M})}$ and S_{COND} 之间增长率差异的大小；由于 $S_{\mathrm{GRO}(\mathcal{M})}$ 必须位于二者之间，因此也界定了这两者与 $S_{\mathrm{GRO}(\mathcal{M})}$ 之间增长率差异的大小。定理 2 和定理 3 本质上是渐近的（以均值参数之间的差异来衡量）。若要给出更精确而非渐近的表述，我们需要区分不同的指数族；下一节将对此进行讨论。

为了陈述这些定理，我们需要定义效应量，即零假设与备择假设之间的差异。迄今为止，我们一直将备择假设视为固定的，并由 $\boldsymbol{\mu}$ 给出，但效应量通常以零假设为出发点来定义。为此，请注意，每个 $P_{\langle\mu_0\rangle} \in \mathcal{H}_0(\mathcal{M})$ 都对应一整组备择假设，其中 $P_{\langle\mu_0\rangle}$ 是零假设内 KL 散度最小的点。该组备择假设由 $\mathrm{M}^{(k)}(\mu_0) = \{\mu'_1, \ldots, \mu'_k \in \mathrm{M} : \frac{1}{k}\sum_{i=1}^k \mu'_i = \mu_0\}$ 参数化，如式 (2.3) 所示。我们可以使用式 (2.1) 中给出的特殊记号 $\langle\mu_0\rangle$，对该集合重新参数化如下。令 A 为 $\mathbb{R}^k$ 中各分量之和为 0 的单位向量集合，即 $\boldsymbol{\alpha} \in \mathrm{A}$ 当且仅当 $\sqrt{\sum_{j=1}^k \alpha_j^2} = 1$ 且 $\sum_{j=1}^k \alpha_j = 0$。显然，$\boldsymbol{\mu} \in \mathrm{M}^{(k)}(\mu_0)$ 当且仅当 μ_1、……、$\mu_k \in \mathrm{M}$ 且对于某个标量 $\delta \geq 0$ 和某个 $\boldsymbol{\alpha} \in \mathrm{A}$，$\boldsymbol{\mu} = \langle\mu_0\rangle + \delta\boldsymbol{\alpha}$。我们可以将 δ 视为效应的大小，将 $\boldsymbol{\alpha}$ 视为其方向。请注意，若 $k = 2$，则只有两个方向，即 $\mathrm{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_{-1}\}$，其中 $\boldsymbol{a}_1 = (1/\sqrt{2},\ -1/\sqrt{2})$ 和 $\boldsymbol{a}_{-1} = -\boldsymbol{a}_1$，分别对应正向和负向效应：当 $\boldsymbol{\alpha} = \boldsymbol{a}_1$ 时，$\mu_1 - \mu_2 = \sqrt{2}\cdot\delta$；当 $\boldsymbol{\alpha} = \boldsymbol{a}_{-1}$ 时，$\mu_1 - \mu_2 = -\sqrt{2}\cdot\delta$，如后文图 1 所示。还请注意，对于一般的 k，在下述定理中，我们可以直接将 δ 理解为 $\boldsymbol{\mu}$ 与 $\langle\mu_0\rangle$ 之间的欧几里得距离。

定理 2。固定某个 $\mu_0 \in \mathrm{M}$、某个 $\boldsymbol{\alpha} \in \mathrm{A}$，并令 $\boldsymbol{\mu} = \langle\mu_0\rangle + \delta\boldsymbol{\alpha}$ 对满足$\boldsymbol{\mu} \in \mathrm{M}^{(k)}(\mu_0)$的 $\delta \geq 0$ 成立。$S_{\mathrm{PSEUDO}(\mathcal{M})}$ 与 $S_{\mathrm{GRO(IID)}}$ 之间的增长率差为

$$\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log S_{\mathrm{PSEUDO}(\mathcal{M})} - \log S_{\mathrm{GRO(IID)}}\right] = \frac{1}{8}\int_x \frac{\left(f_x''(0)\right)^2}{f_x(0)} d\rho(x)\cdot\delta^4 + o\left(\delta^4\right) = O\left(\delta^4\right), \tag{3.2}$$

其中 $f_x(\delta) = \sum_{i=1}^k p_{\mu_0+\delta\alpha_i}(x) = \sum^k p_{\mu_i}(x)$，$f_x''$是 f_x的二阶导数，因此$_{i=1} f_x(0) = k p_{\mu_0}(x)$，并且（经过一些计算） $f_x''(0) = \frac{d^2}{d\mu^2} p_\mu(x)\,|_{\mu=\mu_0}$。

正如 $O(\cdot)$记号所隐含的那样，左侧的期望定义良好且有限，中间等式中的积分也有限。该定理表明，对于一般的指数族，只要距离 δ 在 $\mathcal{H}_1$ 与 $\mathcal{H}_0(\mathcal{M})$之间较小，$S_{\mathrm{GRO(IID)}}$ 在 GRO 意义下就会出乎意料地接近（$O(\delta^4)$）最优的 $S_{\mathrm{GRO}(\mathcal{M})}$。这意味着，只要 $S_{\mathrm{GRO}(\mathcal{M})} \neq S_{\mathrm{PSEUDO}(\mathcal{M})}$（因此 $S_{\mathrm{GRO}(\mathcal{M})}$ 难以计算，且 $S_{\mathrm{PSEUDO}(\mathcal{M})}$ 不是一个e变量），我们就可以考虑改用 $S_{\mathrm{GRO(IID)}}$：它会更加稳健（因为它对于大得多的假设 $\mathcal{H}_0$(IID)而言是一个e变量），而且在增长率方面只会略逊一筹。

定理 2 与下一个定理十分相似，后者涉及的是 S_{COND} 而非$S_{\mathrm{GRO(IID)}}$。为陈述该定理，我们先令 $X_k(x^{k-1}, z) := z - \sum_{i=1}^{k-1} x_i$，并将 $Z = \sum_{i=1}^{k} X_i$ 在 $P_{\boldsymbol{\mu}}$ 下的边缘分布记为 $P_{\boldsymbol{\mu};[Z]}$，注意其密度$p_{\boldsymbol{\mu};[Z]}$ 由下式给出

$$p_{\boldsymbol{\mu};[Z]}(z) = \int_{\mathcal{C}(z)} p_{\boldsymbol{\mu}}\left(x^{k-1}, x_k\right) d\rho(x^{k-1}), \tag{3.3}$$

其中，ρ 被扩展为 ρ 在 $\mathcal{X}^{k-1}$ 上的乘积测度，且

$$\mathcal{C}(z) := \left\{x^{k-1} \in \mathcal{X}^{k-1} : X_i(x^{k-1}, z) \in \mathcal{X}\right\}. \tag{3.4}$$

定理 3。固定某个 $\mu_0 \in \mathtt{M}$、$\boldsymbol{\alpha} \in \mathtt{A}$，并令 $\boldsymbol{\mu} = \langle\mu_0\rangle + \delta\boldsymbol{\alpha}$ 对于满足 $\boldsymbol{\mu} \in \mathtt{M}^{(k)}(\mu_0)$ 的 $\delta \geq 0$ 成立。$S_{\mathrm{PSEUDO}(\mathcal{M})}$ 与 S_{COND} 之间的增长率差为

$$\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log S_{\mathrm{PSEUDO}(\mathcal{M})} - \log S_{\mathrm{COND}}\right] = \frac{1}{8}\int_z \frac{\left(g_z''(0)\right)^2}{g_z(0)} d\rho_{[Z]}(z) \cdot \delta^4 + o\left(\delta^4\right) = O(\delta^4), \tag{3.5}$$

其中，$g_z(\delta) := p_{\langle\mu_0\rangle + \boldsymbol{\alpha}\delta;\,[Z]}(z)$，而 $\rho_{[Z]}$ 表示由 ρ 在 $\mathcal{X}^k$上的乘积测度诱导出的 Z 上的测度；$g_z''(0)$ 的显式表达式为

$$\int_{\mathcal{C}(z)} p_{\langle\mu_0\rangle}\left(x^k\right) \sum_{j=1}^{k} \left[I'(\mu_0)(x_j - \mu_0) - I(\mu_0)\right] d\rho(x^{k-1}),$$

其中，$I(\mu)$ 表示 μ 的费 *Fisher* 信息，而 $I'(\mu)$ 是其一阶导数。

同样，左侧的期望是定义良好且有限的，右侧的积分也是有限的。将定理 3 与定理 2 进行比较可知，$f_x(0)$（在 x处计算的 k 个相同密度之和）被 $g_z(0)$（在 z处计算的 k 个独立同分布随机变量之和的密度）所替代。

推论 1. 采用上述两个定理中的定义，增长率之差$\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\text{对数 } S_{\mathrm{COND}} - \text{对数}S_{\mathrm{GRO(IID)}}\right]$ 可写为

$$\frac{1}{8}\left(\int_x \frac{\left(f_x''(0)\right)^2}{f_x(0)} d\rho(x) - \int_z \frac{\left(g_z''(0)\right)^2}{g_z(0)} d\rho_{[Z]}(z)\right) \cdot \delta^4 + o\left(\delta^4\right) = O\left(\delta^4\right). \tag{3.6}$$

4 特定指数分布族的增长率比较

现在，我们将针对若干标准指数族，更精确地建立 k-样本检验中四种（伪）e变量之间的关系。这些指数族包括表1中列出的族，以及本节末尾列出的几个相关族。对于所考虑的每个族 $\mathcal{M}$，我们给出证明，说明不同的 e变量何时相同，即 $S = S'$，其中 S，$S' \in \{S_{\mathrm{GRO}(\mathcal{M})}, S_{\mathrm{COND}}, S_{\mathrm{GRO(IID)}}, S_{\mathrm{PSEUDO}(\mathcal{M})}\}$。只要我们能够证明，对于另一个$e$变量 $S \in \{S_{\mathrm{COND}}, S_{\mathrm{GRO(IID)}}\}$，有 $S_{\mathrm{GRO}(\mathcal{M})} \neq S$，便可推断出 $S_{\mathrm{GRO}(\mathcal{M})} \succ S$，因为 $S_{\mathrm{GRO}(\mathcal{M})}$ 是 $\mathcal{H}_0(\mathcal{M})$ 的 GROe变量。如果 S_{COND} 和 $S_{\mathrm{GRO(IID)}}$ 都不等于 $S_{\mathrm{GRO}(\mathcal{M})}$，我们将通过模拟考察 $S_{\mathrm{GRO(IID)}} \succ S_{\mathrm{COND}}$ 是否成立，或反之亦然——我们的理论结果并不适用于这种情况。本文中的所有模拟均针对 $k = 2$ 这一情形进行。

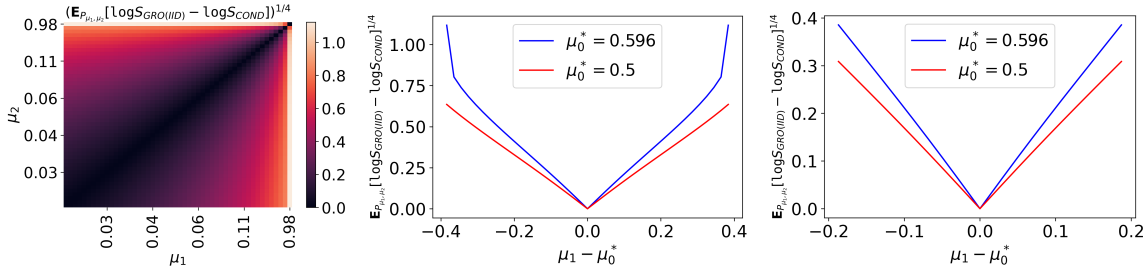
定理 2 和定理 3 表明，在 $\delta = 0$（$\mu_1, \ldots, \mu_k$ 都彼此接近）的邻域内，差值 $\mathbb{E}_{P_\mu}[\log S - \log S']$ 的阶为 δ^4，当 $S, S' \in \{S_{\mathrm{GRO}(\mathcal{M})}, S_{\mathrm{PSEUDO}(\mathcal{M})}, S_{\mathrm{GRO(IID)}}, S_{\mathrm{COND}}\}$ 时。因此，图中将展示 $(\mathbb{E}_{P_\mu}[\log S - \log S'])^{1/4}$，因为这样一来，我们预计随着远离对角线，距离将线性增大，从而使图示更具信息量。

我们的发现，包括证明和模拟结果，总结于表1。对于每个指数族，我们列出了伪e变量与阶'$\succ$'相比时的秩。以黑色标注的秩已在附录 D 中得到证明，而以蓝色标注的秩只是基于上述模拟结果提出的猜想。支持这些猜想的模拟结果见图 1。此外，当 $S_{\mathrm{PSEUDO}(\mathcal{M})}$不是该模型的$e$变量时，其秩以红色标注，详情见附录。请注意，只要任何两个e变量的秩相同，它们就必定在 ρ-几乎处处相等；这是由对数函数的严格凹性以及指数族中分布的全支撑性质所致。例如，表中的结果反映出，对于伯努利族，我们已证明 $S_{\mathrm{PSEUDO}(\mathcal{M})} = S_{\mathrm{GRO}(\mathcal{M})} = S_{\mathrm{GRO(IID)}}$，且 $S_{\mathrm{PSEUDO}(\mathcal{M})} \succ S_{\mathrm{COND}}$。此外，对于几何族，以及方差自由且均值固定的贝塔分布，我们已证明$S_{\mathrm{PSEUDO}(\mathcal{M})}$ 不是e变量，且 $S_{\mathrm{GRO}(\mathcal{M})} \neq S_{\mathrm{GRO(IID)}}$，并且 $S_{\mathrm{GRO}(\mathcal{M})} \neq S_{\mathrm{COND}}$，因此由（3.1）可得 $S_{\mathrm{PSEUDO}(\mathcal{M})} \succ S_{\mathrm{GRO}(\mathcal{M})}$、$S_{\mathrm{GRO}(\mathcal{M})} \succ S_{\mathrm{GRO(IID)}}$ 和 $S_{\mathrm{GRO}(\mathcal{M})} \succ S_{\mathrm{COND}}$。接着，图 1a 所示的模拟结果表明，对于方差自由且均值固定的贝塔分布，$S_{\mathrm{GRO(IID)}} \succ S_{\mathrm{COND}}$；图 1b 的结果则表明，对于几何族，$S_{\mathrm{COND}} \succ S_{\mathrm{GRO(IID)}}$，但这些结论尚未得到证明。图 1c 表明，对于方差自由且均值固定的高斯分布，$S_{\mathrm{GRO(IID)}} \succ S_{\mathrm{COND}}$。最后，图 1d 表明，对于指数分布，$S_{\mathrm{GRO(IID)}}$ 与 S_{COND}之间没有明确的关系。也就是说，对于某些$\mu_1, \ldots, \mu_k \in \mathrm{M}$，$S_{\mathrm{GRO(IID)}}$ 的增长速度快于 S_{COND}，而对于另一些则慢于 S_{COND}，表中用秩 (3) – (4) 表示了这一点。

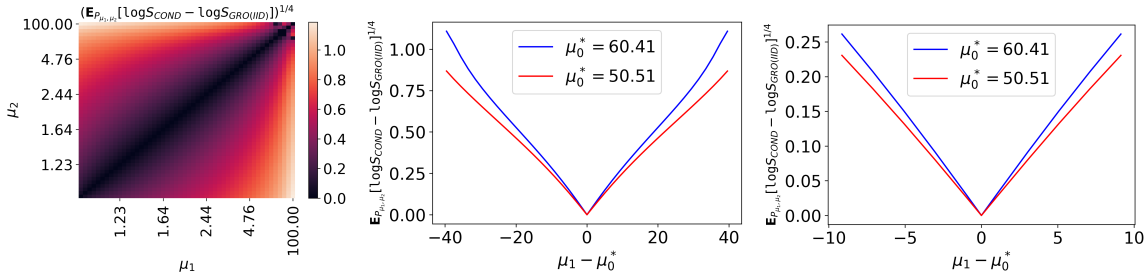
指数族	$S_{\mathrm{PSEUDO}(\mathcal{M})}$	$S_{\mathrm{GRO}(\mathcal{M})}$	$S_{\mathrm{GRO(IID)}}$	S_{COND}
伯努利分布	(1)	(1)	(1)	(2)
均值自由的高斯分布，方差固定	(1)	(1)	(2)	(1)
泊松分布	(1)	(1)	(2)	(1)
自由的 β，固定的 α贝塔分布	(1)	(2)	(3)	(4)
几何分布	(1)	(2)	(4)	(3)
均值固定、方差自由的高斯分布	(1)	(2)	(3)	(4)
指数分布	(1)	(2)	(3)-(4)	(3)-(4)

表1：与关系'$\succ$'相比时，四种不同e变量的秩。
黑色标注的秩已在附录 D 中证明，而蓝色标注的秩则是基于
关于图 1 中的模拟。当 $S_{\mathrm{PSEUDO}(\mathcal{M})}$ 不属于相应秩时，其秩以红色标出
一个 e 变量，如附录 D 所示

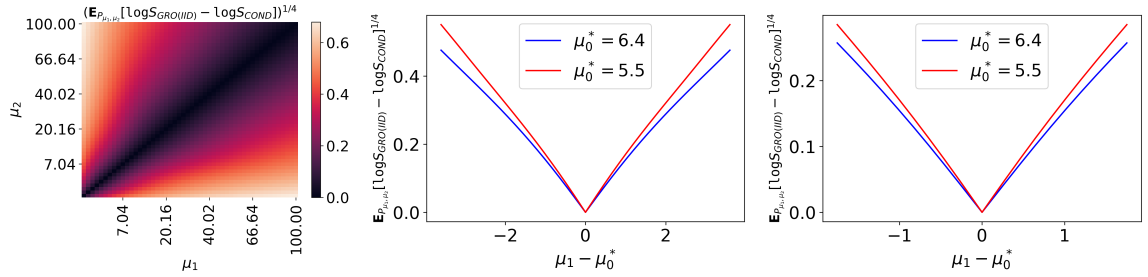
最后，我们注意到，对于表中列出的每个族，所得结果必然可以推广到其他任何族，只要将其化为自然形式 (1.2) 后，该族与表中族相同。例如，具有固定最小参数 v 的帕累托分布族可以化为指数分布族：如果 $U \sim \mathrm{Pareto}(v, \alpha)$，那么我们可以进行变换 $X = t(U)$，其中 $t(U)$ 为 $\log(U/v)$，随后得到 $X \sim \mathrm{Exp}(\alpha)$。因此，对于具有帕累托$(v, \alpha)$ 分布且以 α 为自由参数的 k 样本问题，其等价于具有指数分布的 k 样本问题；所得到的 e 值 $S_{\mathrm{GRO}(\mathcal{M})}$



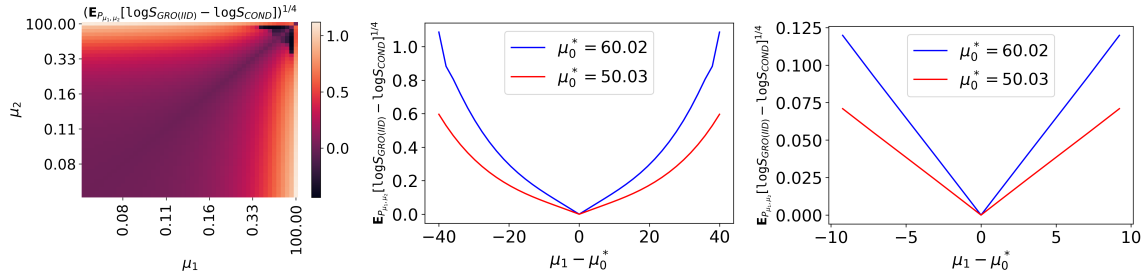
(a) 方差自由的贝塔分布，且 β 固定 α



(b) 几何分布



(c) 均值固定、方差自由的高斯分布



(d) 指数分布

图 1：四种指数族的 $S_{\mathrm{GRO(IID)}}$ 与 S_{COND} 比较。我们在一个由 50×50 个备择假设构成的网格上评估了预期增长差异（μ_1，μ_2），这些备择假设在标准参数化下等距排列（这解释了图示的均值参数化中的非线性缩放）。左侧为对应的热图。右侧为这些热图沿对角线的“切片”：红色曲线对应主对角线（左上至右下），蓝色曲线对应从左上方第二个刻度（第 10 个离散化点）到右下方第二个刻度的对角线。这些切片关于 0 对称，其值仅取决于 $\delta = |\mu_1 - \mu_2| / \sqrt{2} = |\mu_1 - \mu_0^*| \cdot \sqrt{2}$，其中 $\mu_0^* = (\mu_1 + \mu_2)/2$，而 δ 如定理 2 所述

在帕累托情形下，对于观测 U，给定某个特定备择假设时，与指数情形下对应备择假设对于观测 $X = t\,(U)$ 时的 $S_{\mathrm{GRO}(\mathcal{M})}$ 相同； $S_{\mathrm{GRO(IID)}}$ 和 S_{COND} 也同理。因此，帕累托分布的排序必须与表1中的指数分布排序相同。类似地，对数正态分布（均值或方差自由）的e变量可以简化为对应的两个正态分布e变量。

5 用模拟近似 RIPr

由于具有增长最优性，有时我们可能仍希望使用 GROe变量 $S_{\mathrm{GRO}(\mathcal{M})}$，即使它不等于容易计算的 $S_{\mathrm{PSEUDO}(\mathcal{M})}$。为此，我们需要对其进行数值近似。本节有两个目标：首先，我们想说明原则上这是可行的；其次，我们将展示这会引出一些关于未来工作的有趣问题。因此，下文将更详细地探讨如何模拟近似 $S_{\mathrm{GRO}(\mathcal{M})}$，针对的是我们此前考虑过的、具有 $S_{\mathrm{GRO}(\mathcal{M})} \neq S_{\mathrm{PSEUDO}(\mathcal{M})}$ 的指数族，即贝塔分布、几何分布、指数分布和方差自由的高斯分布；为简单起见，我们只考虑 $k = 2$ 的情形。附录E中提供了一些图表，展示 RIPr 概率密度在 μ_1, μ_2 的特定取值下的情况；这里我们重点讨论如何近似这些密度，并将针对 $k = 2$ 得出的发现作为推测更大的 k 时情况的依据。

5.1 通过Li算法近似RIPr

Li [1999] 提供了一种算法，可以根据 KL 散度，将具有密度q 的分布 Q 在分布集合 $\mathcal{P}$ 的凸包 $\mathrm{CONV}(\mathcal{P})$上的 RIPr 任意精确地近似，其中每个 $P \in \mathcal{P}$ 都具有密度p。在第 m 步，该算法输出一个有限混合 $P_{(m)} \in \mathrm{CONV}(\mathcal{P})$，其中至多包含 m 个 $\mathcal{P}$的元素。对于 $m > 1$，这些混合通过迭代地设定 $P_{(m)} := \alpha P_{(m-1)} + (1-\alpha)P'$来确定，其中 $\alpha \in [0,1]$ 和$P' \in \mathcal{P}$ 的选择旨在使 KL 散度 $D(Q\|\alpha P_{(m-1)} + (1-\alpha)P')$最小。在此， $P_{(1)}$ 被定义为 $\mathcal{P}$ 中使 $D(Q\|P_{(1)})$最小的唯一元素。因此，这是一种贪心算法，但 Li 表明，在关于 $\mathcal{P}$的某些正则条件下，有 $D(Q\|P_{(m)}) \to \inf_{P \in \mathrm{CONV}(\mathcal{P})} D(Q\|P)$。也就是说， $P_{(m)}$ 在 KL 散度意义下近似 RIPr。这表明（但本身不足以证明）$\sup_{P \in \mathcal{P}} \mathbb{E}_P[q(X)/p_{(m)}(X)] \to 1$，即似然比确实趋于一个 e变量。

我们通过数值计算，考察这一结论是否适用于我们熟悉的情形： $k = 2$, Q等于 $P_{\boldsymbol{\mu}}$，其中某个 $\boldsymbol{\mu} = (\mu_1, \mu_2) \in \mathrm{M}^2$, 和 $\mathcal{P} = \mathcal{H}_0(\mathcal{M})$。为此，我们对贝塔分布、指数分布、几何分布和方差自由的高斯分布，针对多种取值（μ_1, μ_2）应用了Li算法。在所有这些情形下，经过至多 $m = 15$ 次迭代后，我们发现$\sup_{\mu_0 \in \mathrm{M}} \mathbb{E}_{P_{\mu_0,\ \mu_0}}[p_{\mu_1,\mu_2}(X_1,\ X_2)/q_{(m)}(X_1,\ X_2)]$ 不超过 1.005：Li算法收敛得相当快；收敛情况及模拟中的设计选择见附录E中的图示。

（请注意，由于我们已经证明了 $S_{\mathrm{GRO}(\mathcal{M})} = S_{\mathrm{PSEUDO}(\mathcal{M})}$ 对于伯努利分布、泊松分布和均值自由的高斯分布成立，因此无需对这些分布族近似 $S_{\mathrm{GRO}(\mathcal{M})}$。）

5.2 通过暴力法近似RIPr

虽然Li算法的收敛速度相当快，但由于其采用贪心策略，在迭代 $m = 2$时仍远非最优。这促使我们研究：通过仅使用两个分量的混合，我们能够在多大程度上接近e变量。因此，我们设 p_A（x^k）：$= \alpha p_{\langle\mu_{01}\rangle}$（$x^k$）+（$1-\alpha$）$p_{\langle\mu_{02}\rangle}$（$x^k$），并针对 $\boldsymbol{\mu} =$（μ_1，μ_2）的各种取值，考察了

$$S_{\mathrm{APPR}} := \frac{p_{\boldsymbol{\mu}}(X^k)}{p_A(X^k)} \tag{5.1}$$

作为近似 e 变量，对于使.....最小化的特定 $\alpha \in [0,1]$ 和 μ_{01}，μ_{02}

$$\sup_{\mu_0 \in \mathtt{M}} \mathbb{E}_{P_{\langle\mu_0\rangle}}[S_{\mathrm{APPR}}].$$

（实际操作中，我们在 M 上取 1000 个等间距网格点，对 μ_0 求最大值；并在包含 100 个等距网格点的网格上，对 α，μ_{01}，μ_{02} 求最小值。网格在 M 上的左右端点通过反复试验确定。）

模拟结果列于表2，其中 $k = 2$ 取特定值，μ_1，μ_2 取特定值，指数族则选取适合进行近似的族（即 $S_{\mathrm{GRO}(\mathcal{M})} \neq S_{\mathrm{PSEUDO}(\mathcal{M})}$）。我们还尝试了更多参数值，并得到了类似结果；附录E中的表3为每个族另列了一组参数值。对于所有这些族，$\sup_{\mu_0 \in \mathtt{M}} \mathbb{E}_{P_{\langle\mu_0\rangle}}[S_{\mathrm{APPR}}]$ 都非常接近 1。Gr¨unwald 等人 [2023] 的推论2表明，若上确界恰好为 1，即 S_{APPR} 是一个 e变量，那么 S_{APPR} 也必定是相对于 $P_{\boldsymbol{\mu}}$的 GRO e 变量。这使我们推测，所有超过 1 的部分或许都源于离散化和数值误差；以下情况可能成立（也可能不成立——我们未能找到证明或证伪该说法的方法）：

猜想 对于 $k = 2$，RIPr，即所实现的分布

$$\min_{Q \in \mathrm{CONV}(\mathcal{H}_0(\mathcal{M}))} D(P_{\mu_1,\mu_2} \| Q)$$

可以写成 $\mathcal{H}_0(\mathcal{M})$ 中两个元素的混合。

6 结论与未来工作

在本文中，我们介绍并分析了四类e变量，用于检验 k组数据是否按照指数族中的同一元素分布。这四类e变量包括：GRO e变量（$S_{\mathrm{GRO}(\mathcal{M})}$）、条件$e$变量（$S_{\mathrm{COND}}$）、混合$e$变量（$S_{\mathrm{GRO}(\mathrm{IID})}$）以及伪$e$变量（$S_{\mathrm{PSEUDO}(\mathcal{M})}$）。我们比较了简单备择假设下$e$变量的增长率；在该备择假设下，$k$ 个组中的每一组都具有指数族中不同但固定的分布。我们证明，对于任意两个e变量

$$S\ S' \in \{S_{\mathrm{GRO}(\mathcal{M})}, S_{\mathrm{COND}}, S_{\mathrm{GRO}(\mathrm{IID})}, S_{\mathrm{PSEUDO}(\mathcal{M})}\} \qquad \mathbb{E}[\quad S - \quad S'] =, ,$$

$O\ \delta^4 \qquad \ell_2$

如果该备择分布的参数与零假设参数空间之间的距离为 δ，则有 log log ( )。这表明，当效应量较小时，所有e变量的表现都出人意料地相似。对于更一般的效应量，根据定义，我们知道$S_{\mathrm{GRO}(\mathcal{M})}$ 具有最高的增长率。计算 $S_{\mathrm{GRO}(\mathcal{M})}$ 需要计算备择分布在零假设上的反向信息投影，而这通常是一个困难的

分布	$(\mu_1,\ \mu_2)$	α	$(\mu_{01},\ \mu_{02})$	$\sup_{\mu_0 \in \mathtt{M}} \mathbb{E}_{X_1,\ X_2 \sim P_{\mu_0,\ \mu_0}}[S_{\mathrm{APPR}}]$
beta	(0.5, 0.25)	0.22	(0.24, 0.81)	1.0052
指数分布 方差自由的高斯分布 且固定均值	(0.5, 0.25)	0.56	(0.35, 0.51)	1.0000
	(0.5, 0.25)	0.37	$(0.5,\ \ 0.5)$	1.0000
指数分布 几何分布	$(\frac{10}{3}, \frac{5}{4})$	0.51	(0.62, 0.31)	1.0047
	$(\frac{10}{3}, \frac{5}{4})$	0.47	(1.84, 2.97)	1.0008
方差自由的高斯分布 且均值固定	$(\frac{10}{3}, \frac{5}{4})$	0.08	(3.64, 2.73)	1.0002

表2：对于给定的 $\boldsymbol{\mu} = (\mu_1,\ \mu_2)$，我们展示了相应双分量混合$\alpha p_{\mu_{01}}(X_1)\, p_{\mu_{01}}(X_2) + (1-\alpha)\, p_{\mu_{02}}(X_1)\, p_{\mu_{02}}(X_2)$ 的 α，μ_{01}和 μ_{02}；该混合分布通过按照第 5.2 节中的方法对 KL 散度进行穷举最小化得到。我们还展示了相应似然比 S_{APPR} 作为 e变量的接近程度

问题。不过，我们证明了存在一些指数族，使得以下条件之一成立： $S_{\mathrm{PSEUDO}(\mathcal{M})} = S_{\mathrm{GRO}(\mathcal{M})}$、$S_{\mathrm{COND}} = S_{\mathrm{GRO}(\mathcal{M})}$ 或 $S_{\mathrm{GRO(IID)}} = S_{\mathrm{GRO}(\mathcal{M})}$，从而大大简化了问题。如果希望检验一个不满足该条件的指数族，则可以使用算法来估计反向信息投影。我们已通过数值计算验证，反向信息投影的近似也会得到 $S_{\mathrm{GRO}(\mathcal{M})}$的近似。不过，由于计算上的优势，相较于 $S_{\mathrm{GRO}(\mathcal{M})}$，使用 S_{COND} 或 $S_{\mathrm{GRO(IID)}}$ 可能仍然更为可取。我们的模拟表明，具体应优先选择哪一个取决于特定的指数族；有时甚至不存在明确的优先顺序。

致谢 感谢 Rosanne Turner 和 Wouter Koolen 与我们进行了多次极有帮助的讨论。

声明：财务支持、无利益冲突 我们获得了国家留学基金管理委员会国家公派留学基金项目编号202006280045的部分财务支持。作者没有需要就与本文内容相关的任何利益冲突作出声明。

参考文献

RJ Adams。《 $2\times\ 2$ 列联表的安全假设检验》。硕士论文，代尔夫特理工大学，2020年。Akshay Balsubramani 和 Aaditya Ramdas。利用迭代对数律进行序贯非参数检验。人工智能中的不确定性，2016年。O.E. Barndorff-Nielsen。《统计理论中的信息与指数族》。威利，英国奇切斯特，1978年。Lawrence D. Brown。《统计指数族基础及其在统计决策理论中的应用》，*IMS* 讲义专著系列第9卷。IMS，1986年。

D.A. Darling 和 H. Robbins。均值、方差和中位数的置信序列。美国国家科学院院刊，58(1):66–68，1967年。Boyan Duan、Aaditya Ramdas 和 Larry Wasserman。通过下注进行交互式秩检验。载于第一届因果学习与推理会议，机器学习研究论文集第177卷，第201–235页，2022年4月11–13日。P. Gr¨unwald。E后验。伦敦皇家学会哲学汇刊 *A*辑，2023年。Peter Gr¨unwald。《最小描述长度原理》。麻省理工学院出版社，2007年。Peter Gr¨unwald、Alexander Henzi 和 Tyron Lardy。model-x 下条件独立性的任意时刻有效检验。*arXiv* 预印本 *arXiv:2209.12637*，2022年。Peter Gr¨unwald、Rianne de Heide 和 Wouter Koolen。安全检验。*arXiv* 预印本 *arXiv:1906.07801*，2023年。已被《皇家统计学会会刊 B辑》接收。Alexander Henzi 和 Johanna F. Ziegel。对概率预测表现进行有效的序贯推断。*Biometrika*，2022年。John L. Kelly。对信息率的一种新诠释。贝尔系统技术期刊，35：第917–26页，1956年。Alix Lh´eritier 和 Fr´ed´eric Cazals。序贯非参数多变量双样本检验。*IEEE* 信息论汇刊，64(5):3361–3370，2018年。Qiang Jonathan Li。《混合模型估计》。耶鲁大学，1999年。Teodora Pandeva、Tim Bakker、Christian A Naesseth 和 Patrick Forr´e。评估分类器双样本检验。*arXiv* 预印本 *arXiv:2210.13027*，2022年。Aaditya Ramdas、Peter Gr¨unwald、Vladimir Vovk 和 Glenn Shafer。博弈论统计学与安全的任意时刻有效推断。*arXiv* 预印本 *arXiv:2210.01948*，2022年。Shalev Shaer、Gal Maman 和 Yaniv Romano。通过下注检验实现条件独立性的无模型序贯检验。*arXiv* 预印本 *arXiv:2210.00354*，2022年。Glenn Shafer。下注检验：一种统计与科学交流策略（附讨论与回应）。皇家统计学会会刊 *A* 辑，184(2):407–478，2021。Rosanne Turner 和 Peter Gr¨unwald。列联表及其他情形的任意时刻有效置信区间。*arXiv* 预印本 *arXiv:2203.09785*，2022a。Rosanne Turner 和 Peter Gr¨unwald。分层计数数据的安全序贯检验与效应估计。载于*2023*年第二十六届国际人工智能与统计学会议（*AISTATS*），机器学习研究论文集第206卷，2022b。Rosanne Turner、Alexander Ly 和 Peter Gr¨unwald。列联表及其他情形的安全检验与始终有效置信区间。*arXiv* 预印本 *arXiv:2106.02693*，2021。

Vladimir Vovk 和 Ruodu Wang. E值：校准、组合与应用。《统计年鉴》，49:1736–1754，2021。

Abraham Wald.《序贯分析》。John Wiley & Sons, Inc.，纽约；Chapman & Hall, Ltd.，伦敦，1947年。

Ulla-Britt Wennerholm、Sissel Saltvedt、Anna Wessberg、M ̊arten Alkmark、Christina Bergh、Sophia Brismar Wendel、Helena Fadl、Maria Jonsson、Lars Ladfors、Verena Sengpiel 等。41周时引产与期待管理，以及42周时引产的比较（瑞典过期妊娠引产研究，swepis）：一项多中心、开放标签、随机优效性试验。《英国医学杂志》，367，2019年。

David Williams.《鞅与概率》。剑桥大学出版社，1991年。

William HenryYoung. 关于可求和函数类及其傅里叶级数。《伦敦皇家学会会刊》。*A*辑，数学与物理学论文，87(594):225–229，1912年。

A 实际应用： k 个独立同分布数据流

在最简单的实际应用中，我们一次观察一个数据块，即在时间 n，我们已经观察到 $\boldsymbol{X}_{(1)}$、.....、$\boldsymbol{X}_{(n)}$，其中每个 $\boldsymbol{X}_{(i)} = (X_i, 1, \ldots..., X_i, k)$ 都是一个数据块，也就是一个向量，其中每个 k 组各有一个结果。这是一个相当受限的设置，但我们可以轻松将其扩展到这样的数据块：每个组包含数量不同的结果。例如，如果数据以数据块的形式到达，组 j 中有 m_j 个结果，对于 $j = 1 \ldots k$、$X_{(i)} = (X_i, 1, 1, \ldots..., X_i, 1, m_1, X_i, 2, 1, \ldots..., X_i, 2, m_2, \ldots..., X_i, k, 1, \ldots..., X_i, k, m_k)$，我们可以将其重新组织为具有$k' = \sum_{j=1}^{k} m_j$个组、每个组包含1个结果的形式，并考虑这样一个备择假设：结果向量的前 m_1 个条目具有相同的均值$\mu'_1 = \ldots = \mu'_{m_1} = \mu_1$；接下来的 m_2 个条目具有相同的均值 $\mu'_{m_1+1} = \ldots = \mu'_{m_1+m_2} = \mu_2$；依此类推。

不过，更一般地说，我们将面对 k 个彼此独立同分布的数据流，而且每个数据流中的数据到达速率可能不同。我们仍然可以处理这种情况，方法是预先为每个数据流确定一个重数 $m_1, \ldots, m_k$。随着数据不断到达，我们用每组 j 的 m_j 个结果填充虚拟「数据块」 $j = 1 \ldots k$。一旦一个或多个虚拟数据块被完全填满，就可以像往常一样进行分析，但仅限于已填满的数据块。也就是说，如果对于某个整数 B，我们已经观察到数据流 j 中有 Bm_j 个结果（对所有 $j = 1 \ldots k$ 都成立），但对于某个j，我们尚未观察到 $(B + 1)m_j$ 个结果，而我们决定停止分析并计算反对零假设的证据，那么我们就输出前 B 个数据块对应的e变量乘积，并暂时忽略任何额外数据。重要的是，如果我们在分析数据流时发现某些数据流提供数据的速度远快于其他数据流，就可以动态调整 $m_1, \ldots, m_k$：每当一个虚拟数据块完成时，我们都可以为下一个数据块确定备择假设重数；关于 $k = 2$ 的情形，详见 Turner等人 [2021]的详细说明。

B 第2节的证明

在证明中，我们会直接使用指数族的（对数）密度导数的基本性质，而不再特别说明。这些性质都可以在例如 Barndorff-Nielsen 的著作中找到

[1978]。

B.1 命题 1 的证明

证明。由于引理 2 中已经证明 $S_{\mathrm{GRO}(\mathcal{M})}$ 是一个 E 变量，因此该命题的“若”部分成立。“仅若”部分直接由[Gr¨unwald 等人提出的定理 1 的推论2得出， 2023], 该推论指出，形如 $p_{\mu}(X^k)/r(X^k)$的 E 变量至多只有一个，其中 r 是 X^k的概率密度。 □

B.2 命题 2 的证明

证明。定义 $g(\mu_0) := \mathbb{E}_{p_{\langle\mu_0\rangle}}\left[S_{\mathrm{PSEUDO}(\mathcal{M})}\right]$ 和 $B(\mu_i) := A(\lambda(\mu_i) + \lambda(\mu_0) - \lambda(\mu_0^*))$ 。

$$
\begin{aligned}
g(\mu_0) &= \mathbb{E}_{p_{\langle\mu_0\rangle}}\left[\prod_{i=1}^{k}\frac{p_{\mu_i}(X_i)}{p_{\mu_0^*}(X_i)}\right] = \prod_{i=1}^{k}\mathbb{E}_{Y\sim p_{\mu_0}}\left[\frac{p_{\mu_i}(Y)}{p_{\mu_0^*}(Y)}\right] \\
&= \prod_{i=1}^{k}\int \exp\left(\lambda(\mu_0)y - A\left(\lambda(\mu_0)\right)\right)\cdot\frac{\exp\left(\lambda(\mu_i)y - A\left(\lambda(\mu_i)\right)\right)}{\exp\left(\lambda(\mu_0^*)y - A\left(\lambda(\mu_0^*)\right)\right)}d\rho(y) \\
&= \prod_{i=1}^{k}\int \exp\left(\left(\lambda(\mu_i)+\lambda(\mu_0)-\lambda(\mu_0^*)\right)y - A\left(\lambda(\mu_i)\right) - A\left(\lambda(\mu_0)\right) + A\left(\lambda(\mu_0^*)\right)\right)d\rho(y) \\
&= \prod_{i=1}^{k}\exp\left(A\left(\lambda(\mu_0^*)\right) - A\left(\lambda(\mu_i)\right) - A\left(\lambda(\mu_0)\right)\right)\exp\left(B(\mu_i)\right) \\
&\qquad\cdot\int \exp\left(\left(\lambda(\mu_i)+\lambda(\mu_0)-\lambda(\mu_0^*)\right)y - B(\mu_i)\right)d\rho(y) \\
&= \prod_{i=1}^{k}\exp\left(A\left(\lambda(\mu_0^*)\right) - A\left(\lambda(\mu_i)\right) - A\left(\lambda(\mu_0)\right)\right)\exp\left(B(\mu_i)\right)\cdot 1 \\
&= \exp\left(kA\left(\lambda(\mu_0^*)\right) - \sum_{i=1}^{k}A\left(\lambda(\mu_i)\right) - kA\left(\lambda(\mu_0)\right) + \sum_{i=1}^{k}B(\mu_i)\right). \qquad \text{(B.1)}
\end{aligned}
$$

对 μ_0 求一阶和二阶导数，可得

$$
\frac{d}{d\mu_0}g(\mu_0) = g(\mu_0)\cdot\frac{d}{d\mu_0}\left(\sum_{i=1}^{k}B(\mu_i) - kA\left(\lambda(\mu_0)\right)\right) \qquad \text{(B.2)}
$$

and

$$\begin{aligned}\frac{d^2}{d\mu_0^2}g(\mu_0) =& \left(\frac{d}{d\mu_0}g(\mu_0)\right)\cdot\frac{d}{d\mu_0}\left(\sum_{i=1}^{k}B(\mu_i)-kA\left(\lambda(\mu_0)\right)\right)\\ &+g(\mu_0)\cdot\frac{d^2}{d\mu_0^2}\left(\sum_{i=1}^{k}B(\mu_i)-kA\left(\lambda(\mu_0)\right)\right)\\ =&g(\mu_0)\left(\sum_{i=1}^{k}(\mu_i+\mu_0-\mu_0^*)-k\mu_0\right)^2\\ &+g(\mu_0)\left(\sum_{i=1}^{k}\mathrm{VAR}_{P_{\mu_i+\mu_0-\mu_0^*}}[X]-k\mathrm{VAR}_{P_{\mu_0}}[X]\right)\\ =&g(\mu_0)\left(\sum_{i=1}^{k}\mathrm{VAR}_{P_{\mu_i+\mu_0-\mu_0^*}}[X]-k\mathrm{VAR}_{P_{\mu_0}}[X]\right)=g(\mu_0)\cdot f(\mu_0).\end{aligned} \tag{B.3}$$

其中，第二个等式由（B.2）、$(d/d\lambda(\mu))A(\lambda(\mu)) = \mathbb{E}_{P_\mu}[X]$ 和 $(d^2/d\lambda(\mu)^2)A(\lambda(\mu)) = \mathrm{VAR}_{P_\mu}[X]$ 得出。（B.3）关于 μ_0 连续。因此，如果 $f(\mu_0^*) > 0$ 成立，就意味着存在一个区间 $\mathbf{M}^* \subset \mathbf{M}$，其中 μ_0^* 位于 $\mathbf{M}^*$ 的内部，并且（B.1）在该区间上严格凸。那么必定存在一点 $\mu_0' \in \mathbf{M}^*$，满足 $\mathbb{E}_{P_{\langle\mu'\rangle}}\left[S_{\mathrm{PSEUDO}(\mathcal{M})}\right] > \mathbb{E}_{P_{\langle\mu^*\rangle}}\left[S_{\mathrm{PSEUDO}(\mathcal{M})}\right] =$

1，即 $S_{\mathrm{PSEUDO}(\mathcal{M})}$ 不是一个 E 变量。反之，$f(\mu_0^*) < 0$ 意味着存在一个区间 $\mathbf{M}^* \subset \mathbf{M}$，使得 μ_0^* 位于 $\mathbf{M}^*$的内部，并且（B.1）在该区间上严格凹。结论得证。 □

B.3 定理1的证明

为了准备定理1的证明，我们先回顾杨氏不等式：

引理3。 **[杨氏不等式]** 设 p、q 为满足 $\frac{1}{p}+\frac{1}{q}=1$的正实数。那么，如果 a、b 为非负实数，则有 $ab \le \frac{a^p}{p}+\frac{b^q}{q}$.

定理1的证明完全遵循 Turner等人 [2021] 在 $\mathcal{M}$ 为伯努利模型这一特殊情形下证明该陈述时所采用的相同论证。

证明。 我们首先证明，定理陈述中定义的 $S_{\mathrm{GRO(IID)}}$ 是一个 E 变量。为此，令 $p_0^*(X) = \frac{1}{k}\sum_{i=1} p_{\mu_i}(X)$。我们有：

$$\mathbb{E}_{X^k\sim P_{\langle\mu_0\rangle}}\left[S_{\mathrm{GRO(IID)}}\right] = \mathbb{E}_{X_1\sim P_{\mu_0}}\left[\frac{p_{\mu_1}(X_1)}{p_0^*(X_1)}\right]\cdot\ldots\cdot\mathbb{E}_{X_k\sim P_{\mu_0}}\left[\frac{p_{\mu_k}(X_k)}{p_0^*(X_k)}\right]. \tag{B.4}$$

我们还有

$$\begin{aligned}&\frac{1}{k}\mathbb{E}_{X_1\sim P_{\mu_0}}\left[\frac{p_{\mu_1}(X_1)}{p_0^*(X_1)}\right]+\cdots+\frac{1}{k}\mathbb{E}_{X_k\sim P_{\mu_0}}\left[\frac{p_{\mu_k}(X_k)}{p_0^*(X_k)}\right]\\ =&\frac{1}{k}\mathbb{E}_{X\sim P_{\mu_0}}\left[\frac{p_{\mu_1}(X)}{\frac{1}{k}\sum_{i=1}^{k}p_{\mu_i}(X)}+\cdots+\frac{p_{\mu_k}(X)}{\frac{1}{k}\sum_{i=1}^{k}p_{\mu_i}(X)}\right]=1.\end{aligned} \tag{B.5}$$

我们需要证明 (B.4) ≤ 1，为此可以使用 (B.5)。更简单地说，只需证明 $\prod r_i \leq 1$ with $\frac{1}{k}\sum r_i \leq 1$, $r_i \in \mathbb{R}^+$。但这很容易得出：

$$\begin{aligned}
\frac{1}{k}\sum_{i=1}^{k} r_i = \frac{k-1}{k}\cdot\frac{\sum_{i=1}^{k-1} r_i}{k-1} + \frac{r_k}{k} &\geq \left(\frac{\sum_{i=1}^{k-1} r_i}{k-1}\right)^{\frac{k-1}{k}} r_k^{\frac{1}{k}} \\
&= \left(\frac{k-2}{k-1}\cdot\frac{\sum_{i=1}^{k-2} r_i}{k-2} + \frac{r_{k-1}}{k-1}\right)^{\frac{k-1}{k}} r_k^{\frac{1}{k}} \\
&\geq \left(\frac{\sum_{i=1}^{k-2} r_i}{k-2}\right)^{\frac{k-2}{k}} r_{k-1}^{\frac{1}{k}} r_k^{\frac{1}{k}} \\
&\vdots \\
&\geq \left(\frac{r_1+r_2}{2}\right)^{\frac{2}{k}} \prod_{i=3}^{k} r_i^{\frac{1}{k}} \geq \prod_{i=1}^{k} r_i^{\frac{1}{k}} \qquad \text{(B.6)}
\end{aligned}$$

其中，第一个不等式成立是因为在引理3中令 $\frac{1}{p} := \frac{k-1}{k}$，$\frac{1}{q} := \frac{1}{k}$，$a^p := \sum_{i=1}^{k-1} r_i$，~~$b^q := r_k$~~ 后应用了杨氏不等式。其余不等式可用同样的方法证明。由此可得 $\prod r_i^k \leq 1$，进而得到 $\prod r_i \leq 1$。

这表明 $S_{\mathrm{GRO(IID)}}$ 是一个 e变量。接下来需要证明 $S_{\mathrm{GRO(IID)}}$ 确实是相对于 $\mathcal{H}_0$(IID)的 GRO e 变量；证明这一点后，由引理2可知它是唯一的此类 e变量，因此由引理1可知，P_0^* 在引理1中达到最小值。由于我们已经知道 $S_{\mathrm{GRO(IID)}}$ 是一个 e变量，因此根据 Gr¨unwald 等人 [2023], 的定理1推论2，$S_{\mathrm{GRO(IID)}}$ 相对于 $\mathcal{H}_0$(IID)是 GRO e 变量；该推论指出，形如 $p_{\boldsymbol{\mu}}(X^k)/r(X^k)$ 的 e 变量至多有一个，其中 r 是概率密度。由于 $S_{\mathrm{GRO(IID)}}$ 是这样的 e 变量，根据引理1，它必定是 GRO e 变量。

□

B.4 命题3的证明

证明。 X_1，X_2、.....、X_k的观测值记为 x^k（:= x_1、.....、x_k）。令 X_k（x^{k-1}，z）：= $z - \sum_{i=1}^{k-1} x_i$，并按照（3.4）定义 $\mathcal{C}$（z），按照（3.3）定义 $p_{\boldsymbol{\mu};[Z]}$（z）和 ρ（x^{k-1}），

可得:

$$
\begin{aligned}
p_{\boldsymbol{\mu}}\left(x^{k-1}\middle|Z=z\right) &= \frac{p_{\boldsymbol{\mu}}\left(x^k\right)}{p_{\boldsymbol{\mu};[Z]}\left(z\right)} \\
&= \frac{\exp\left(\sum_{i=1}^{k}\left(\lambda(\mu_i)x_i - A(\lambda(\mu_i))\right)\right)}{\int_{y^{k-1}\in\mathcal{C}(z)}\exp\left(\sum_{i=1}^{k-1}\left(\lambda(\mu_i)y_i - A(\lambda(\mu_i)) + \lambda(\mu_k)X_k(y^{k-1},z)\right) - A(\lambda(\mu_k))\right)d\rho(y^{k-1})} \\
&= \frac{\exp\left(\lambda(\mu_k)z + \sum_{i=1}^{k-1}(\lambda(\mu_i)-\lambda(\mu_k))x_i\right)}{\int_{y^{k-1}\in\mathcal{C}(z)}\exp\left(\lambda(\mu_k)z + \sum_{i=1}^{k-1}(\lambda(\mu_i)-\lambda(\mu_k))y_i\right)d\rho(y^{k-1})} \\
&= \frac{\exp\left(\sum_{i=1}^{k-1}(\lambda(\mu_i)-\lambda(\mu_k))x_i\right)}{\int_{y^{k-1}\in\mathcal{C}(z)}\exp\left(\sum_{i=1}^{k-1}(\lambda(\mu_i)-\lambda(\mu_k))y_i\right)d\rho(y^{k-1})}.
\end{aligned}
$$

□

C 第3节的证明

C.1 定理 2 的证明

证明。 我们通过对下文定义的 $F(\delta)$ 在 $\delta = 0$ 处进行复杂的泰勒展开来证明该定理。我们首先计算 $F(\delta)$ 的前四阶导数。因此，我们定义并推导出，$_k$其中 $\mu_i = \mu_0 + \alpha_i\delta$ ，且 $f_y(\delta)$ $= \sum p_{\mu_i}(y)$ 的定义与定理陈述中的相同，$_{i=1}$

$$
\begin{aligned}
F(\delta) :=& \mathbb{E}_{P_{\langle\mu_0\rangle+\boldsymbol{\alpha}\delta}}\left[\log S_{\mathrm{PSEUDO}(\mathcal{M})} - \log S_{\mathrm{GRO(IID)}}\right] \\
=& \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log\prod_{j=1}^{k}\left(\frac{1}{k}\sum_{i=1}^{k}p_{\mu_i}(X_j)\right) - \log p_{\langle\mu_0\rangle}(X^k)\right] \\
=& \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\sum_{j=1}^{k}\log f_{X_j}(\delta) - \sum_{j=1}^{k}\log p_{\mu_0}(X_j)\right] - k\log k \\
\overset{(a)}{=}& \sum_{j=1}^{k}\mathbb{E}_{X\sim P_{\mu_j}}\left[\log f_X(\delta) - \log p_{\mu_0}(X)\right] - k\log k \\
\overset{(b)}{=}& \overbrace{\int_{y\in\mathcal{X}} f_y(\delta)\log f_y(\delta)d\rho(y)}^{F_1(\delta)} + \overbrace{\left(-\int_{y\in\mathcal{X}} f_y(\delta)\log p_{\mu_0}(y)d\rho(y)\right)}^{F_2(\delta)} - k\log k, \qquad \text{(C.1)}
\end{aligned}
$$

其中，我们定义 $F_1(\delta)$ 等于 (C.1) 中最左边的一项， $F_2(\delta)$ 等于第二项；只要满足以下条件，(a) 和 (b) 就都成立:

$$
\text{for all } j\in\{1,\ldots,k\}\text{: } \mathbb{E}_{X_j\sim P_{\mu_j}}\left[\left|\log f_{X_j}(\delta) - \log p_{\mu_0}(X_j)\right|\right] < \infty \qquad \text{(C.2)}
$$

有限。我们在在线补充材料中验证了，对于所有足够接近 0 的 δ，该条件以及许多相关的绝对值期望有限条件都成立。综上，这些条件不仅意味着 (a) 和 (b) 成立，还意味着 (c)：在计算 $F_1(\delta)$ 和 $F_2(\delta)$ 的一阶 k 导数时，对于所有此类 δ，我们可以自由交换对 y 的积分与对 δ 的微分，其中 k 为任意有限值；此外 (d)，当 δ 取包含 0 的某个紧区间内的值时，所有这些导数都有限（由于细节直接明了，但相当繁琐冗长，我们将其推迟到补充材料中）。因此，利用 (c)，在下文证明的其余部分中，我们将自由地在积分号下进行微分；利用 (d)，我们可以得出最终结果有限的结论。

对于每个导数，我们首先计算 $F_1(\delta)$ 的导数，然后计算 $F_2(\delta)$ 的导数。

$$
\begin{aligned}
F_1'(\delta) &= \int f_y'(\delta) d\rho(y) + \int f_y'(\delta) \log f_y(\delta) d\rho(y) = 0, \\
F_2'(\delta) &= -\int f_y'(\delta) \log p_{\mu_0}(y) d\rho(y) = 0, \text{ so } \quad F'(0) = F_1'(0) + F_2'(0) = 0,
\end{aligned} \tag{C.3}
$$

其中，上述公式成立是因为对于所有 $x \in \mathcal{X}$，$f_x'(0) = 0$，由此可以得到 by

$$
\begin{aligned}
f_x'(\delta^\circ) =& \sum_{j=1}^{k} \frac{dp_{\mu_j}(x)}{d\mu_j} \frac{d\mu_j}{d\delta}(\delta^\circ), \\
f_x'(0) =& \frac{dp_{\mu_0}(x)}{d\mu_0} \sum_{j=1}^{k} \frac{d\mu_j}{d\delta}(0) = \frac{dp_{\mu_0}(x)}{d\mu_0} \sum_{j=1}^{k} \alpha_j = 0,
\end{aligned} \tag{C.4}
$$

其中我们利用了在 $\delta = 0$ 处所有 μ_j 都等于 μ_0 这一事实。下面讨论二阶导数 s:

$$
\begin{aligned}
F_1''(\delta) &= \int f_y''(\delta) d\rho(y) + \int \left(f_y''(\delta) \log f_y(\delta) + \frac{\left(f_y'(\delta)\right)^2}{f_y(\delta)} \right) d\rho(y) \\
&= \int \left(f_y''(\delta) \log f_y(\delta) + \frac{\left(f_y'(\delta)\right)^2}{f_y(\delta)} \right) d\rho(y) \\
F_1''(0) &= \int \left(f_y''(0) \log f_y(0) + \frac{\left(f_y'(0)\right)^2}{f_y(0)} \right) d\rho(y); \\
&= \int f_y''(0) \log p_{\mu_0}(y) d\rho(y) + \int_{y \in \mathcal{X}} \left(f_y''(0) \log k \right) d\rho(y) \\
&= \int \left(f_y''(0) \log p_{\mu_0}(y) \right) d\rho(y),
\end{aligned} \tag{C.5}
$$

其中 $\int f_y''(\delta) d\rho(y) = 0$，因为 $\int f_y(\delta) d\rho(y) = k$，其中 k 是一个不依赖于 δ 的常数。于是 $F_2''(\delta)$ 可表示为

$$
\begin{aligned}
F_2''(\delta) = -\int f_y''(\delta) \log p_{\mu_0}(y) d\rho(y) \; ; \; F_2''(0) = -\int f_y''(0) \log p_{\mu_0}(y) d\rho(y), \text{ so} \\
F''(0) = F_1''(0) + F_2''(0) = 0.
\end{aligned} \tag{C.6}
$$

现在我们计算 $F(\delta)$ 的三阶导数，记作 $F_{(3)}(\delta)$。

$$
\begin{aligned}
F_1^{(3)}(\delta) &= \int \left(f_y^{(3)}(\delta) \log f_y(\delta) + \frac{f_y''(\delta) f_y'(\delta)}{f_y(\delta)} + \frac{2 f_y''(\delta) f_y'(\delta) f_y(\delta) - (f_y'(\delta))^3}{(f_y(\delta))^2} \right) d\rho(y) \\
F_1^{(3)}(0) &= \int f_y^{(3)}(0) \log f_y(0) d\rho(y) \\
&= \int f_y^{(3)}(0) \log p_{\mu_0}(y) d\rho(y) + \int f_y^{(3)}(0) \log k d\rho(y) \\
&= \int f_y^{(3)}(0) \log p_{\mu_0}(y) d\rho(y) \\
F_2^{(3)}(\delta) &= - \int f_y^{(3)}(\delta) \log p_{\mu_0}(y) d\rho(y) \\
F_2^{(3)}(0) &= - \int f_y^{(3)}(0) \log p_{\mu_0}(y) d\rho(y), \text{ so } F^{(3)}(0) = F_1^{(3)}(0) + F_2^{(3)}(0) = 0,
\end{aligned}
\tag{C.7}
$$

这成立是因为 $f_y'(0) = 0$ 以及 $\int f_y(0) d\rho(y) = k$。

$F(\delta)$ 的四阶导数可按如下方式计算：

$$
\begin{aligned}
F_1^{(4)}(\delta) = & \int \left(f_y^{(4)}(\delta) \log f_y(\delta) + \frac{f_y^{(3)}(\delta) f_y'(\delta)}{f_y(\delta)} \right) d\rho(y) \\
& + \int 3 \cdot \frac{\left(f_y^{(3)}(\delta) f_y'(\delta) + (f_y''(\delta))^2 \right) f_y(\delta) - f_y''(\delta) \left(f_y'(\delta) \right)^2}{\left(f_y(\delta) \right)^2} d\rho(y) \\
& - \int \frac{3 \left(f_y(\delta) f_y'(\delta) \right)^2 \cdot f_y''(\delta) - 2 \left(f_y'(\delta) \right)^4 \cdot f_y(\delta)}{\left(f_y(\delta) \right)^4} d\rho(y) \ ;
\end{aligned}
\tag{C.8}
$$

$$
\begin{aligned}
F_1^{(4)}(0) &= \int \left(f_y^{(4)}(0) \log f_y(0) + \frac{3 \left(f_y''(0) \right)^2}{f_y(0)} \right) d\rho(y) \\
&= \int f_y^{(4)}(0) \log p_{\mu_0}(y) d\rho(y) + \log k \int_{y \in \mathcal{X}} f_y^{(4)}(0) d\rho(y) + \int_{y \in \mathcal{X}} \frac{3 \left(f_y''(0) \right)^2}{f_y(0)} d\rho(y) \\
&= \int f_y^{(4)}(0) \log p_{\mu_0}(y) d\rho(y) + \int_{y \in \mathcal{X}} \frac{3 \left(f_y''(0) \right)^2}{f_y(0)} d\rho(y),
\end{aligned}
$$

而 $F_2^{(4)}(\delta)$ 可通过以下方式计算

$$
\begin{aligned}
& F_2^{(4)}(\delta) = - \int f_y^{(4)}(\delta) \log p_{\mu_0}(y) d\rho(y), \ F_2^{(4)}(0) = - \int f_y^{(4)}(0) \log p_{\mu_0}(y) d\rho(y), \text{ so} \\
& F^{(4)}(0) = F_1^{(4)}(0) + F_2^{(4)}(0) = \int \frac{3 \left(f_y''(0) \right)^2}{f_y(0)} d\rho(y) > 0.
\end{aligned}
$$

基于 基于上述导数，我们现在可以对 $F(\delta)$ 进行四阶泰勒展开

在 $\delta = 0$ 附近展开，得到：

$$\begin{aligned}\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log S_{\mathrm{PSEUDO}(\mathcal{M})} - \log S_{\mathrm{GRO(IID)}}\right] =& \frac{1}{4!}F^{(4)}(0)\delta^4 + o(\delta^4) \\ =& \frac{1}{8}\int_{y\in\mathcal{X}_k}\frac{\left(f_y''(0)\right)^2}{f_y(0)}d\rho(y)\cdot\delta^4 + o\left(\delta^4\right),\end{aligned}$$

其中 $f_y(0) = \sum_{i=1}^{k} p_{\mu_0}(y) = kp_{\mu_0}(y)$，且 $f_y''(0) = \left(\sum_{i=1}^{k}\alpha_i^2\right)\cdot\frac{d^2}{d\mu^2}p_{\mu}(y)\ |_{\mu=\mu_0} = \frac{d^2}{d\mu^2}p_{\mu}(y)\ |_{\mu=\mu_0}$。

□

C.2 定理 3 的证明

证明。我们通过比前一定理中更加复杂的泰勒展开得到结果。与该定理一样，我们可以自由地在积分号下对 δ 求导——在线补充材料中也再次验证了这样做是允许的。

令 $\boldsymbol{\mu}, \boldsymbol{\alpha}, \mathcal{C}(z)$ $\rho(x^{k-1})$ $P_{\boldsymbol{\mu}}$ 等均如定理陈述所述。我们有：

$$\begin{aligned}&f(\delta) := \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log S_{\mathrm{PSEUDO}(\mathcal{M})} - \log S_{\mathrm{COND}}\right] \\ =&\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log\frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle\mu_0\rangle}\left(X^k\right)} - \log\frac{p_{\boldsymbol{\mu}}\left(X^{k-1}\mid Z\right)}{p_{\langle\mu_0\rangle}\left(X^{k-1}\mid Z\right)}\right] \\ =&\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log\frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle\mu_0\rangle}\left(X^k\right)} - \log\frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle\mu_0\rangle}\left(X^k\right)} + \log\frac{\int_{\mathcal{C}(z)}p_{\boldsymbol{\mu}}\left(x^k\right)d\rho(x^{k-1})}{\int_{\mathcal{C}(z)}p_{\langle\mu_0\rangle}\left(x^k\right)d\rho(x^{k-1})}\right] \\ =&D\left(P_{\langle\mu_0\rangle+\boldsymbol{\alpha}\delta;[Z]}\|P_{\langle\mu_0\rangle;[Z]}\right).\end{aligned}$$

我们将通过在 $\delta = 0$ 处对 $f(\delta)$ 作泰勒展开来证明该结果。显然，$f(0) = 0$，且一阶导数 $f'(0) = 0$，因为 $f(0)$ 是开集上 $f(\delta)$ 的最小值，并且 $f(\delta)$ 可微。接下来计算 $f(\delta)$ 的二阶导数，使用定理陈述中的记号 $g_z(\delta) = p_{\langle\mu_0\rangle+\boldsymbol{\alpha}\delta;[Z]}(z)$，其中 g_z' 和 g_z'' 分别表示一阶导数和二阶导数。

$$\begin{aligned}f'(\delta) &= \int g_z'(\delta)\log\frac{g_z(\delta)}{g_z(0)}d\rho_{[Z]}(z) + \int g_z'(\delta)d\rho_{[Z]}(z) = \int g_z'(\delta)\log\frac{g_z(\delta)}{g_z(0)}d\rho_{[Z]}(z). \\ f''(\delta) &= \int g_z''(\delta)\log\frac{g_z(\delta)}{g_z(0)}d\rho_{[Z]}(z) + \int\frac{\left(g_z'(\delta)\right)^2}{g_z(\delta)}d\rho_{[Z]}(z),\end{aligned}$$

其中，第一行中的第二个等号成立，是因为交换微分与积分的顺序不会改变第二项，并且 $\int g_z(\delta)dz = 1$ 在 δ 中为常数。由此得到

$$f''(0) = \int\frac{\left(g_z'(0)\right)^2}{g_z(0)}d\rho_{[Z]}(z), \tag{C.9}$$

and，其中 x_k 设为 $X_k(x^{k-1},\ z)$，并回想 $\boldsymbol{\mu} = \langle\mu_0\rangle + \boldsymbol{\alpha}\delta$ 和 $\mu_j = \mu_0 + \alpha_j\delta$ ，

$$
\begin{aligned}
g_z'(\delta) &= \int_{\mathcal{C}(z)} \frac{d}{d\delta} p_{\langle\mu_0\rangle+\boldsymbol{\alpha}\delta}(x^k) d\rho(x^{k-1}) \\
&= \int_{\mathcal{C}(z)} \sum_{j=1}^{k} \prod_{i\in\{1,\ldots,k\}\setminus j} p_{\mu_i}(x_i) \frac{dp_{\mu_j}(x_j)}{d\delta} d\rho(x^{k-1}) \\
&= \int_{\mathcal{C}(z)} \sum_{j=1}^{k} p_{\mu_1,\ldots,\mu_{j-1},\mu_{j+1},\ldots,\mu_k}(x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_k) \frac{dp_{\mu_j}(x_j)}{d\mu_j} \frac{d\mu_j}{d\delta} d\rho(x^{k-1}) \\
&= \int_{\mathcal{C}(z)} \sum_{j=1}^{k} p_{\boldsymbol{\mu}}(x^k) \frac{d\log p_{\mu_j}(x_j)}{d\mu_j} \alpha_j d\rho(x^{k-1}) \\
&= \int_{\mathcal{C}(z)} \sum_{j=1}^{k} p_{\boldsymbol{\mu}}(x^k) \left(I(\mu_j)x_j - \mu_j I(\mu_j)\right) \alpha_j d\rho(x^{k-1})
\end{aligned}
$$

其中 $I(\mu_j)$ 是费 Fisher 信息。最后一个等号成立，是因为对于与 μ_j 对应的规范参数 $\lambda(\mu_j)$，我们有 $d\lambda(\mu_j)/d\mu_j = I(\mu_j)$ 且 $dA(\beta)/d\beta|_{\beta=\lambda(\mu_j)} = \mu_j$；例如参见 [Gr¨unwald，2007 年，第 18] 章。现在

$$
\begin{aligned}
g_z'(0) &= \int_{\mathcal{C}(z)} \sum_{j=1}^{k} p_{\langle\mu_0\rangle}(x^k) \left(I(\mu_0)x_j - \mu_0 I(\mu_0)\right) \alpha_j d\rho(x^{k-1}) \\
&= \int_{\mathcal{C}(z)} p_{\langle\mu_0\rangle}(x^k) I(\mu_0) \sum_{j=1}^{k} x_j\alpha_j d\rho(x^{k-1}) \qquad \text{(C.10)} \\
&= I(\mu_0) \cdot \int_{\mathcal{C}(z)} p_{\langle\mu_0\rangle}(x^k) \sum_{j=1}^{k} x_j\alpha_j d\rho(x^{k-1}) \qquad \text{(C.11)}
\end{aligned}
$$

k其中第二个等号由 $\sum_{j=1} \alpha_j = 0$ 得出。由于 X^k 独立同分布 $\sim P_{\mu_0}$ 于 $P_{\langle\mu_0\rangle}$，且 (C.10) 中的积分是在可交换序列构成的集合上进行的，（为了理解该陈述，我们可以考虑简单情形：$k = 2$、X_1 和 X_2 对于给定的 $\mathcal{C}(z)$ 是“对称的”，因此可以交换。）我们必然有：即使以轮询方式重新排列 α_j，(C.10) 仍然成立；也就是说，对所有 $i = 1..k$，取 $\alpha_{j.}$ $_i = \alpha_{(j+i-1) \text{ 模 } k}$，

$$
g_z'(0) = I(\mu_0) \cdot \int_{\mathcal{C}(z)} p_{\langle\mu_0\rangle}(x^k) \sum_{j=1}^{k} x_j\alpha_{j,i} d\rho(x^{k-1}).
$$

k将这些 k 方程相加，并利用 $\sum \alpha_i = 0$，可得 $kg_z'(0) = 0$，因此 $g_z'(0) = 0$。$_{i=1}$由 (C.9) 可知

$$
f''(0) = 0.
$$

现在我们计算 $f(\delta)$的三阶导数，记作 $f^{(3)}(\delta)$:

$$f^{(3)}(\delta) = \int \left(g_z^{(3)}(\delta) \log \frac{g_z(\delta)}{g_z(0)} + \frac{g_z''(\delta) g_z'(\delta)}{g_z(\delta)} \right) d\rho_{[Z]}(z) + \int \left(\frac{2 g_z''(\delta) g_z'(\delta) g_z(\delta) - (g_z'(\delta))^3}{(g_z(\delta))^2} \right) d\rho_{[Z]}(z)$$

因此，由于 $g_z'(0) = 0$，我们也必须有

$$f^{(3)}(0) = 0.$$

$f(\delta)$ 的四阶导数现在计算如下:

$$f^{(4)}(\delta) = \int \left(g_z^{(4)}(\delta) \log \frac{g_z(\delta)}{g_z(0)} + \frac{g_z^{(3)}(\delta) \cdot g_z'(\delta)}{g_z(\delta)} \right) d\rho_{[Z]}(z) + \int 3 \cdot \frac{\left(g_z^{(3)}(\delta) \cdot g_z'(\delta) + (g_z''(\delta))^2 \right) g_z(\delta) - g_z''(\delta) \cdot (g_z'(\delta))^2}{(g_z(\delta))^2} d\rho_{[Z]}(z).$$

Then

$$f^{(4)}(0) = \int \frac{3 \left(g_z''(0) \right)^2}{g_z(0)} d\rho_{[Z]}(z) > 0.$$

现在，我们已具备对 $f(\delta)$ 在 $\delta = 0$ 附近进行四阶泰勒展开所需的所有要素，得到:

$$\mathbb{E}_{P_\mu} \left[\log S_{\text{PSEUDO}(\mathcal{M})} - \log S_{\text{COND}} \right] = \frac{1}{8} \int \frac{\left(g_z''(0) \right)^2}{g_z(0)} d\rho_{[Z]}(z) \cdot \delta^4 + o\left(\delta^4\right)$$

这正是我们要证明的结论。 □

D 第4节的证明

本节中，我们证明表1中的所有结论。

D.1 伯努利族

我们证明，当 $\mathcal{M}$ 等于伯努利族时，有 $S_{\text{PSEUDO}(\mathcal{M})} = S_{\text{GRO}(\mathcal{M})} = S_{\text{GRO(IID)}} \succ S_{\text{COND}}$。

k证明。 我们设定 $\bar{\mu}_0^* = \frac{1}{k}\sum_{i=1}^{k} \mu_i$。

$$S_{\mathrm{GRO(IID)}} := \frac{p_{\boldsymbol{\mu}}(X^k)}{\prod_{j=1}^{k}\left(\frac{1}{k}\sum_{i=1}^{k} p_{\mu_i}(X_j)\right)} = \frac{p_{\boldsymbol{\mu}}(X^k)}{\prod_{j=1}^{k}\left(\frac{1}{k}\sum_{i=1}^{k}\left(\mu_i^{X_j}(1-\mu_i)^{1-X_j}\right)\right)} \tag{D.1}$$

$$= \frac{p_{\boldsymbol{\mu}}(X^k)}{\prod_{j=1}^{k}\left((\mu_0^*)^{X_j}(1-\mu_0^*)^{1-X_j}\right)}$$

$$= \frac{p_{\boldsymbol{\mu}}(X^k)}{\prod_{j=1}^{k} p_{\mu_0^*}(X_j)} = S_{\mathrm{PSEUDO}(\mathcal{M})} \tag{D.2}$$

其中第三个等号成立是因为 $X_i \in \{0,1\}$。因此，根据定理1， $S_{\mathrm{PSEUDO}(\mathcal{M})}$ 是一个 e 变量，且$S_{\mathrm{PSEUDO}(\mathcal{M})} = S_{\mathrm{GRO}(\mathcal{M})}$。结合（3.1）以及以下事实即可得出结论：当 $Z=0$ 或 $Z=2$ 时，我们有 $S_{\mathrm{COND}} = 1$，而其他e变量则不满足这一点，因此 $S_{\mathrm{COND}} \neq S_{\mathrm{GRO}(\mathcal{M})} = S_{\mathrm{PSEUDO}(\mathcal{M})} = S_{\mathrm{GRO(IID)}}$。由此，结合（3.1）即可得到结果。 □

D.2 均值自由、方差固定的泊松分布与高斯分布族

我们证明，当 $\mathcal{M}$ 等于均值自由、方差固定的高斯分布族 σ^2时，有 $S_{\mathrm{PSEUDO}(\mathcal{M})} = S_{\mathrm{GRO}(\mathcal{M})} = S_{\mathrm{COND}} \succ S_{\mathrm{GRO(IID)}}$。对于 $\mathcal{M}$ 等于泊松分布族的情形，证明过程完全类似，故略去。

证明。 注意，若令 $Z := \sum_{i=1}^{k} X_i$，则当 $X^k \sim P_{\boldsymbol{\mu}}$ 时有 $Z \sim \mathcal{N}(\sum_{i=1}^{k}\mu_i, k\sigma^2)$。令 μ_0^* 由（2.3）给出，并相对于固定的备择假设 $P_{\boldsymbol{\mu}}$ ，如（2.3）下方对 $S_{\mathrm{PSEUDO}(\mathcal{M})}$的定义所示。由于 $k\mu_0^* = \sum_{i=1}^{k}\mu_i$，对于 $X^k \sim P_{\langle\mu_0^*\rangle}$， Z 具有相同分布。由此可写为

$$S_{\mathrm{COND}} = \frac{p_{\boldsymbol{\mu}}\left(X^k \mid Z\right)}{p_{\langle\mu_0^*\rangle}\left(X^k \mid Z\right)} = \frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle\mu_0^*\rangle}\left(X^k\right)}\frac{p_{\langle\mu_0^*\rangle}(Z)}{p_{\boldsymbol{\mu}}(Z)} = \frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle\mu_0^*\rangle}\left(X^k\right)} = S_{\mathrm{PSEUDO}(\mathcal{M})}.$$

因此， $S_{\mathrm{PSEUDO}(\mathcal{M})}$ 也是一个e变量，所以根据定理1可得 $S_{\mathrm{PSEUDO}(\mathcal{M})} = S_{\mathrm{GRO}(\mathcal{M})}$。此外， $S_{\mathrm{GRO(IID)}}$ 的分母由不同于 $p_{\langle\mu_0^*\rangle}$的分布给出，因此 $S_{\mathrm{GRO(IID)}} \neq S_{\mathrm{GRO}(\mathcal{M})} = S_{\mathrm{PSEUDO}(\mathcal{M})} = S_{\mathrm{COND}}$。结论随即由（3.1）得出。

□

D.3 对于哪些族， $S_{\mathbf{pseudo}(\mathcal{M})}$ 不是 e变量

这里，我们证明，当 $\mathcal{M}$ 等于方差自由且固定均值的贝塔分布族时， $S_{\mathrm{PSEUDO}(\mathcal{M})}$ 不是一个e变量。由此可得 (3.1) 中的结论 $S_{\mathrm{PSEUDO}(\mathcal{M})} \succ S_{\mathrm{GRO}(\mathcal{M})}$。(3.1) 还给出 $S_{\mathrm{GRO}(\mathcal{M})} \succeq S_{\mathrm{GRO(IID)}}$ 和 $S_{\mathrm{GRO}(\mathcal{M})} \succeq S_{\mathrm{COND}}$。当 $\mathcal{M}$ 等于几何分布族，或等于方差自由且固定均值的高斯分布族时，结论也相同，因为证明 $S_{\mathrm{PSEUDO}(\mathcal{M})}$ 不是一个e变量，与下文给出的贝塔分布证明完全类似

。对于所有这些情况，通过模拟很容易发现，通常 $S_{\mathrm{GRO}(\mathcal{M})} \neq S_{\mathrm{GRO(IID)}}$ 和 $S_{\mathrm{GRO}(\mathcal{M})} \neq S_{\mathrm{COND}}$，因此可得 $S_{\mathrm{GRO}(\mathcal{M})} \succ S_{\mathrm{GRO(IID)}}$ 和 $S_{\mathrm{GRO}(\mathcal{M})} \succ S_{\mathrm{COND}}$。

证明。首先，令 $Q_{\alpha,\beta}$ 表示采用标准参数化的贝塔分布，因此其密度为

$$q_{\alpha,\beta}(u) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-1}(1-u)^{\beta-1}, \qquad \alpha,\beta > 0; u \in [0,1].$$

为简化证明，我们在此假设 $\alpha = 1$。于是

$$q_{1,\beta}(u) = \frac{\Gamma(1+\beta)}{\Gamma(\beta)}(1-u)^{\beta-1} = \frac{1}{1-u}\exp\left(\beta\log(1-u) - \log\frac{1}{\beta}\right)$$

其中，第一个等号成立是因为 $\Gamma(1+\beta) = \beta\Gamma(\beta)$。将其与 (1.1) 对比可知，$\beta$ 是族 $\{Q_{1,\beta}: \beta > 0\}$ 对应的规范参数，并且有

$$\lambda(\mu) = \beta, \quad t(u) = \log(1-u), \quad A(\beta) = \log\frac{1}{\beta}.$$

根据命题 2，要证明该结论，只需证明：对于任意并非全都相等的 $\mu_1, \ldots, \mu_k$，令 $X = t(U) = \log(1-U)$，并令 $\mu_0^* = \frac{1}{k}\sum \mu_i$ $_{i=1}$ 按 (2.3) 定义，则有

$$\sum_{i=1}^{k} \mathrm{VAR}_{P_{\mu_i}}[X] - k\mathrm{VAR}_{P_{\mu_0^*}}[X] > 0. \tag{D.3}$$

直接计算可得

$$\mathrm{VAR}_{P_{\mu_i}}[X] = \mathrm{VAR}_{Q_{1,\beta_i}}[X] = \frac{d^2}{d^2\beta_i}(\log\frac{1}{\beta_i}) = \frac{1}{\beta_i^2} \text{ in particular } \mathrm{VAR}_{P_{\mu_0^*}}[X] = \frac{1}{(\beta_0^*)^2} \tag{D.4}$$

其中，β_i 对应于 μ_i，即 $\mathbb{E}_{Q_{1,\beta_i}}[(X)] = \mu_i$。我们还有：

$$\mathbb{E}_{P_{\beta_0^*}}[(X)] = \mu_0^* = \frac{1}{k}\sum_{i=1}^{k}\mu_i = \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_{P_{\beta_i}}[(X)]. \tag{D.5}$$

$_k$ 由于 $\mathbb{E}_{P_{\beta_i}}[(X)] = \frac{d}{d\beta_i}(\log\frac{1}{\beta_i}) = -\frac{1}{\beta_i}$，因此 $\frac{1}{\beta_0^*} = \frac{1}{k}\sum_{i=1}\frac{1}{\beta_i}$。结合 (D.4) 和 (D.5)，我们得到

$$\sum_{i=1}^{k}\mathrm{VAR}_{P_{\mu_i}}[(X)] - k\mathrm{VAR}_{P_{\mu_0^*}}[(X)] = \sum_{i=1}^{k}\frac{1}{(\beta_i)^2} - k\left(\frac{1}{k}\sum_{i=1}^{k}\frac{1}{\beta_i}\right)^2. \tag{D.6}$$

根据詹森不等式，只要至少有一个 μ_i 不等于 μ_0^*，(D.6) 就严格为正，这正是我们要证明的结论。 □

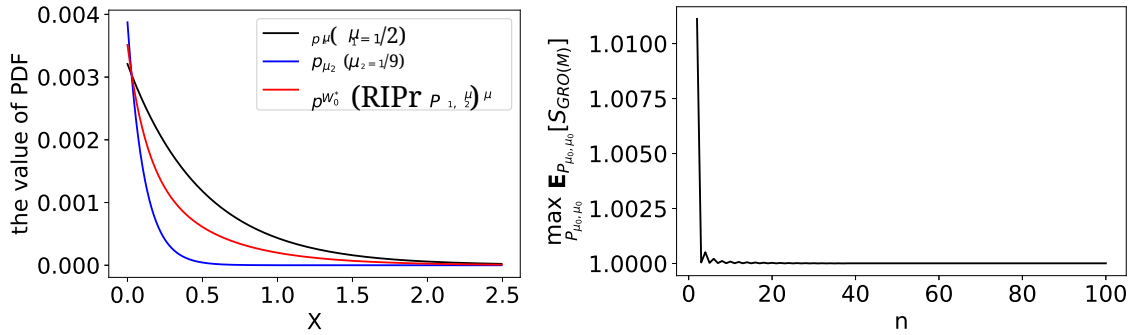


图 2：指数分布。右侧的 n 表示使用 Li算法的迭代次数，从第 2 次迭代开始

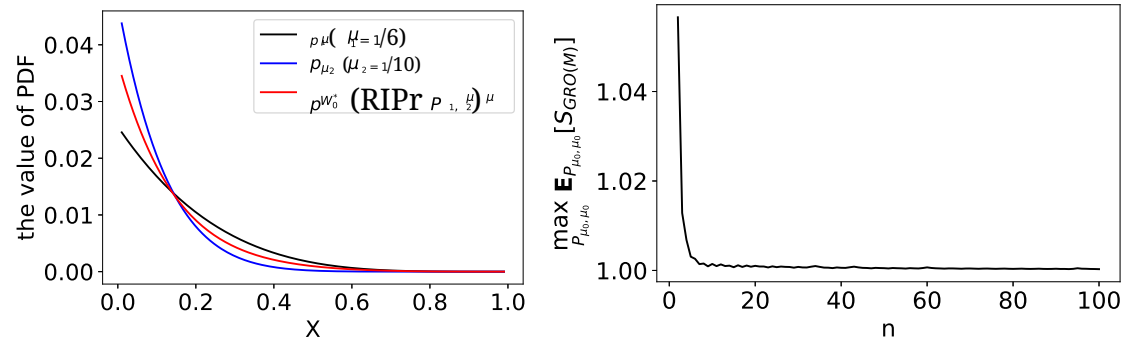


图 3：贝塔分布， β 为自由参数， α 为固定参数。右侧的 n 表示使用 Li算法的迭代次数，从第 2 次迭代开始

E RIPr 近似及Li算法收敛的图示

我们使用四种分布来说明 RIPr 近似以及 Li算法的收敛性：指数分布、具有自由 β 和固定 α 的贝塔分布、几何分布，以及方差自由且均值固定的高斯分布；每种分布均采用一组特定的（随机选取的）参数设置。图 2–5 左侧的图展示了 Li算法经过 $n = 100$ 次迭代后的概率密度函数（对于几何分布，则为离散概率质量函数）。右侧的图展示了 Li算法的收敛速度。右侧的图未显示第一次迭代（对于几何分布和方差自由的高斯分布，则未显示前两次迭代），因为在这些初始步骤中，最坏情况期望 $\sup_{\mu_0 \in \mathtt{M}}[S_{\mathrm{GRO}(\mathcal{M})}]$ 通常大得无法比较。通过实验我们发现，Li算法在计算真实 $S_{\mathrm{GRO}(\mathcal{M})}$ 方面收敛得相当快。在 Li算法的每一步中，我们都在 $[0, 1]$, 内均匀间隔的 100 个点组成的网格上搜索 $P_{(m)}$ 中的最佳混合权重 α，并通过在参数空间 M 内由 100 个等距点组成的网格上搜索 μ' 来确定新分量 $P' = P_{\mu',\ \mu'}$；其中，网格左右端点通过反复试验确定。显然，采用这种临时的离散化策略，我们无法保证任何形式上的近似结果，但在实践中它始终表现良好：在所有情况下，经过 15 次迭代后，我们都发现 $\max_{\mu_0 \in \mathtt{M}} \mathbb{E}_{P_{\mu_0,\ \mu_0}}[S_{\mathrm{GRO}(\mathcal{M})}] \lesssim 1.005$。作为比较，我们展示了能够达到的最佳近似

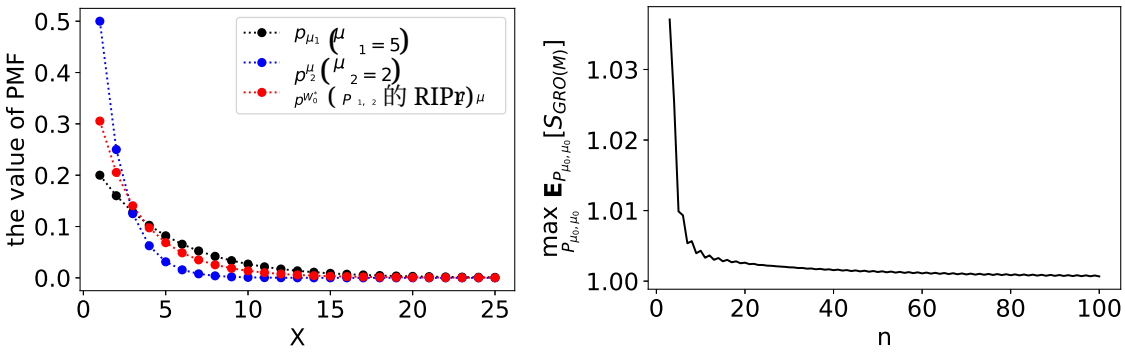


图 4：几何分布。右侧的 n 表示 Li算法从第 3 次迭代开始的迭代次数

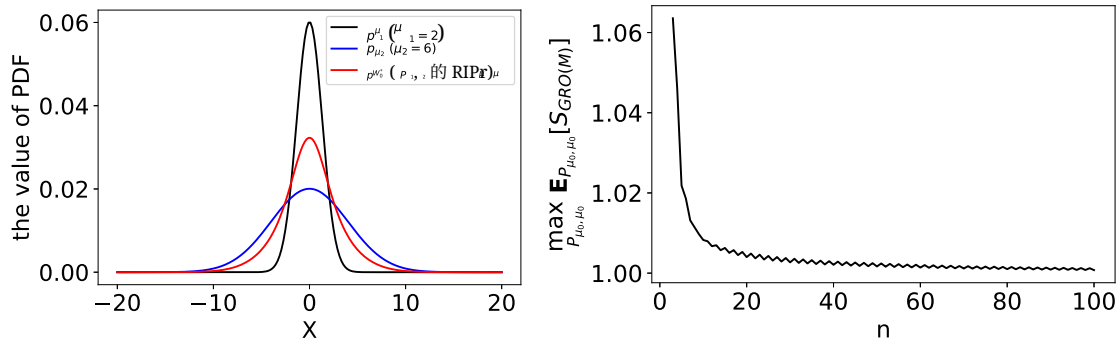


图 5：均值固定、方差自由的高斯分布。右侧的 n 表示 Li算法从第 3 次迭代开始的迭代次数

通过对表 3 中相同参数值下仅两个分量进行暴力组合得到。

分布	$(\mu_1,\ \mu_2)$	α	$(\mu_{01},\ \mu_{02})$	$\sup_{\mu_0\in\mathtt{M}} \mathbb{E}_{X_1,\ X_2\sim P_{\mu_0,\ \mu_0}}[S]$
beta	$(\frac{1}{6}-, \frac{1}{10})$	0.57	(0.12, 0.16)	1.00071
几何分布	(5, 2)	0.39	(2.52, 4.21)	1.00035
指数分布	$(\frac{1}{9}2, 1-)$	0.53	(0.13, 0.51)	1.00083
方差自由的高斯分布且均值固定	(2, 6)	0.41	(5.82, 3.36)	1.00035

表格 3：与表2类似，μ_1，μ_2 对应于图2中使用的参数——

5

补充材料

在本补充材料中，我们验证了定理 2 和定理 3 的证明中隐式使用富比尼定理和在积分号下进行微分所需的所有条件均已满足，并且所有相关导数均有界。

定理 2

文中采用的记号如下:

$$\begin{aligned}
\mu_j &= \mu_0 + \delta\alpha_j \\
\lambda(\mu_j) &= \text{nat. param. } \lambda \text{ corresponding to mean } \mu = \mu_j \\
p_\mu(y) &= e^{\lambda(\mu)y - A(\lambda(\mu))} \\
f_y(\delta) &= \sum_{j=1}^{k} p_{\mu_i}(y).
\end{aligned}$$

As t为简化导数的记号，我们记 $g_y(\lambda) = e^{\lambda y - A(\lambda)}$，于是

$$f_y(\delta) = \sum_{j=1}^{k} g_y(\lambda(\mu_j)) \text{ and } p_{\mu_0}(y) = g_y(\lambda(\mu_0)). \tag{E.1}$$

为强调对 δ 的依赖，以下我们用 $\mu_j(\delta)$ 代替μ_j 。

步骤 1 我们首先证明有限性条件 (C.2)。注意到

$$\begin{aligned}
\log \sum_{j=1}^{k} g_y(\lambda(\mu_j(\delta))) &\le \log(\max_j g_y(\lambda(\mu_j(\delta)))k) \\
&= \max_j \log(g_y(\lambda(\mu_j(\delta)))) + \log k \\
&\le \max_j \log(\max\{g_y(\lambda(\mu_j(\delta))), 1\}) + \log k \\
&\le \sum_j \log(\max\{g_y(\lambda(\mu_j(\delta))), 1\}) + \log k \\
&\le \sum_j |\lambda(\mu_j(\delta))y - \log A(\lambda(\mu_j(\delta)))| + \log k.
\end{aligned}$$

and

$$\begin{aligned}\log \sum_{j=1}^{k} g_y(\lambda(\mu_j(\delta))) &= \log \frac{1}{k} \sum_{j=1}^{k} g_y(\lambda(\mu_j(\delta))) + \log k \\ &\geq \frac{1}{k} \sum_{j=1}^{k} \log g_y(\lambda(\mu_j(\delta))) + \log k \\ &= \frac{1}{k} \sum_{j=1}^{k} \lambda(\mu_j(\delta))y - A(\lambda(\mu_j(\delta))) + \log k.\end{aligned}$$

将这些结果结合起来，我们可以看到

$$\begin{aligned}&|\log f_y(\delta)| \leq \\ &\max\left\{\sum_j |\lambda(\mu_j(\delta))y - A(\lambda(\mu_j(\delta)))| + \log k, \left|\frac{1}{k}\sum_{j=1}^{k}(\lambda(\mu_j(\delta))y - A(\lambda(\mu_j(\delta)))) + \log k\right|\right\} \\ &\leq \sum_j |\lambda(\mu_j(\delta))y - A(\lambda(\mu_j(\delta)))| + \log k, \qquad \text{(E.2)}\end{aligned}$$

以及，更显而易见的是，

$$|\log g_y(\lambda(\mu_0))| \leq |\lambda(\mu_0)y - A(\lambda(\mu_0)|\,. \qquad \text{(E.3)}$$

我们知道， $\lambda(\mu_j(\delta))$ 和 $A(\lambda(\mu_j(\delta)))$ 都是光滑函数，因此当 $\mu_j(\delta)$ 位于均值参数空间 M 的内部时，它们是有限函数（参见 [Barndorff-Nielsen，1978年，第 9 章，定理 9.1 和式 (2)]）。由于 M 是开集，且对所有 $j = 1..k$, $\mu_j(0) = \mu_0 \in \mathsf{M}$，因而可知，对于所有 δ，只要它们位于包含 0 作为内点的 $\mathbb{R}$ 的紧子集中， $|\log f(y)(\delta) - \log g_y(\lambda(\mu_0))|$ 就可以写成关于 $|y|$ 的光滑函数，特别是有限函数。由于在所有满足 $\mu \in \mathsf{M}$ 的 P_μ 下， $|y| \leq 1 + y^2$ 的期望有限，因此根据 (E.1)，(C.2) 有限。

步骤 2 现在我们着手证明，对于位于包含 0 作为内点的 $\mathbb{R}$ 紧子集中的 δ，可以对 δ 求导。证明将用到 (E.2) 和 (E.3)。我们将函数 f_y 和 g_y 的导数记为

$$g_y^s(\lambda) = \frac{\mathrm{d}^s}{\mathrm{d}\lambda^s} g_y(\lambda) \quad \text{and} \quad f_y^s(\delta) = \frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f_y(\delta).$$

我们将论证，对于任意 $s \in \mathbb{N}$，族 $\{\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f_y(\delta) \log f_y(\delta) - f_y(\delta) \log g_y(\lambda(\mu_0)) : \delta \in \Delta\}$ 对于任意紧集 $\Delta \subset \mathbb{R}$ 都是一致可积的，因此我们可以交换微分与积分 [，例如参见 Williams，1991年，第 A 章16]。

利用指数族的标准结果，对于 λ 位于规范参数空间的内部，我们有：

$$\begin{aligned}g_y^{(1)}(\lambda) &= (y - \mu(\lambda))g_y(\lambda) \\ g_y^{(2)}(\lambda) &= -I(\lambda)g_y(\lambda) + (y - \mu(\lambda))^2 g_y(\lambda),\end{aligned}$$

其中， $\mu(\lambda)$ 表示与 λ 对应的均值参数，而 $I(\lambda)$ 表示相应的费 Fisher 信息。

利用 $(d^s/d\lambda^s)A(\lambda)$ 对所有 s 都连续这一事实继续推导，可得

$$g_y^{(s)}(\lambda) = g_y(\lambda) \cdot h_{y,s}(\lambda) \text{ with } h_{y,s}(\lambda) = \sum_{t=1}^{s} h_{[t,s]}(\lambda)(y - \mu(\lambda))^t \tag{E.4}$$

对于某些光滑函数 $h_{[1,s]}$, $h_{[2,s]}$, ..., $h_{[s,s]}$ 的 λ（我们不需要知道这些函数的精确定义）。类似地

$$f_y^{(1)}(\delta) = \sum_j g_y^{(1)}(\lambda_{\mu_j(\delta)}) \cdot (\lambda(\mu_j(\delta)))'$$

其中 $\lambda(\mu_j(\delta))' = \mathrm{d}\ \mathrm{d}\delta\lambda(\mu_j(\delta))$. 我们知道，对于均值参数空间 M 内部的 $\mu_j(\delta)$, $\lambda'(\mu_j(\delta))$ 及更高阶导数是光滑函数（参见 [Barndorff-Nielsen，1978年，第 9 章、定理 9.1 和公式 (2)]）。由于该空间是开集，并且对于所有 $j = 1..k$, $\mu_j(0) = \mu_0 \in \mathrm{M}$，因此可知，对于 $\mathbb{R}$ 的一个以 0 为内点的紧子集中的 δ, $\lambda'(\mu_j(\delta))$ 是 δ 的光滑函数。于是，类似于我们上面对 $g^{(s)}$ 所做的处理，可得

$$f_y^{(s)}(\delta) = \sum_j \sum_{t=1}^{s} g_y^{(t)}(\lambda(\mu_j(\delta))) \cdot r_{t,s}(\mu_j) \tag{E.5}$$

对于某些光滑函数 $r_{t,\ s}$，其具体细节我们无需了解。特别地，结合

$$b_y^{(s)} := \frac{f_y^{(s)}(\delta)}{f_y(\delta)}$$

that

$$\begin{aligned}\left|b_y^{(s)}\right| &\leq \frac{\sum_j g_y(\lambda(\mu_j(\delta))) \cdot (\sum_{t=1}^{s} |h_{y,t}(\lambda(\mu_j(\delta))) \cdot r_{t,s}(\mu_j(\delta))|)}{\sum_j g_y(\lambda(\mu_j(\delta)))} \\ &\leq \sum_j \sum_{t=1}^{s} |h_{y,t}(\lambda(\mu_j(\delta))) \cdot r_{t,s}(\mu_j(\delta))|.\end{aligned}$$

检查正文中的证明，我们可以非正式地指出， $F_0(\delta)$ 和 $F_1(\delta)$ 的前四阶导数中所有不含对数的项，都可以写成乘积 $f_y(\delta) \cdot b^{(s_1)y}(\delta) \cdot \underbrace{\ldots}_{\ldots} \cdot b^{(s_u)y}(\delta)$，其中的 $b_y^{(s)}$ 我们刚刚已用关于 $|y|$ 的多项式加以界定；同样，利用 (E.2) 和 (E.3)，含对数的项也可以用这类多项式来界定，这表明所有积分中的所有项都可以如此界定。事实确实如此：将上述推理形式化后，我们发现

$$\begin{aligned}&\int \left(\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f_y(\delta) \log f_y(\delta) - f_y(\delta) \log g_y(\lambda(\mu_0))\right)^2 d\rho(y) = \\ &\int \left(f_y^{(s)}(\log f_y(\delta) - \log g_y(\lambda(\mu_0))) + f_y(\delta) \sum_u c_u \cdot b_y^{(s_2)}(\delta) \cdot \ldots \cdot b_y^{(s_u)}(\delta)\right)^2 d\rho(y) \\ &= \int (f_y^{(s)}(\log f_y(\delta) - \log g_y(\lambda(\mu_0))))^2 + \left(f_y(\delta) \sum_u c_u \cdot b_y^{(s_1)}(\delta) \cdot \ldots \cdot b_y^{(s_u)}(\delta)\right)^2 \\ &\quad + f_y(\delta) f_y^{(s)}(\log f_y(\delta) - \log g_y(\lambda(\mu_0))) \sum_u c_u \cdot b_y^{(s_1)}(\delta) \cdot \ldots \cdot b_y^{(s_u)}(\delta) d\rho(y).\end{aligned}$$

根据 (E.2)、(E.3) 以及上文给出的 $|b_y^{(s)}|$ 的界，积分中的所有项都可以用关于 y （或 $|y|$）的多项式界定，因此该积分可表示为关于 ρ 和 P_μ 的矩的线性函数。因此，同时注意到 ρ 本身是一个概率测度，也是所考虑指数族中的一个成员（等于 P_μ，其中 $\lambda(\mu) = 0$），可知该积分在均值参数空间的一个紧子集内对 δ 一致有界。由此可知，族 $\{ \mathrm{d}^s \mathrm{d}\delta^s f_y(\delta) \log f_y(\delta) - f_y(\delta) \log g_y(\lambda(\mu_0)) : \delta \in \Delta\}$ 一致可积 [，参见例如 Williams，1991，第 13 章。3], 因此，积分与微分可以自由交换 [，参见例如 Williams，1991，第 A 章16]。此外，定理陈述右侧的量也是有界的。

定理 3

与定理 3 的证明中一样，令 $f(\delta) = \mathbb{E}_{P_\mu}\left[\log \frac{p_\mu(X^k)}{p_{\langle\mu_0\rangle}(X^k)} - \log \frac{p_\mu(X^k|Z)}{p_{\langle\mu_0\rangle}(X^k|Z)}\right]$。

为验证正文中的证明，我们只需证明 $f(\delta)$ 有限，并证明关于 δ，我们可以在一个包含 $\delta = 0$ 的紧区间上交换微分与期望。因此，我们希望证明，对于任意 $s \in \mathbb{N}$，都有

$$\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f(\delta) = \mathbb{E}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\left(\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle\mu_0\rangle}(X^k)} - \log \frac{p_{\boldsymbol{\mu}}(X^k \mid Z)}{p_{\langle\mu_0\rangle}(X^k \mid Z)}\right)\right].$$

为证明这一点，首先注意到，$\mathbb{E}_{P_\mu}\left[\log \frac{p_\mu(X^k)}{p_{\langle\mu_0\rangle}(X^k)}\right]$ 和 $\mathbb{E}_{P_\mu}\left[\log \frac{p_\mu(X^k|Z)}{p_{\langle\mu_0\rangle}(X^k|Z)} \mid Z\right]$ 都是指数族成员之间的 KL 散度（例如，Brown [1986] 证明了在给定充分统计量之和的条件下，会得到一个新的导出的完整指数族）；只要 δ 位于一个内部包含 0 的充分小区间内，这些散度就是有限的（因为此时 μ 位于均值参数空间的内部）。这已经表明 $f(\delta)$ 有限，也使我们能够将其改写为

$$f(\delta) = \mathbb{E}_{P_\mu}\left[\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle\mu_0\rangle}(X^k)}\right] - \mathbb{E}_{P_\mu}\left[\log \frac{p_{\boldsymbol{\mu}}(X^k \mid Z)}{p_{\langle\mu_0\rangle}(X^k \mid Z)}\right].$$

此外，[Brown, 1986, 定理 2.2] 结合 Barndorff-Nielsen [1978]的定理 9.1 和第 9 章公式 2 可知，对于任意完整指数族和任意有限的$k > 0$，KL 散度关于其第一个自变量的 k阶导数（以均值参数化表示）存在且有限，并且可以通过在积分号下求微分得到；对于均值参数空间内部的任意 μ，此结论均成立。因此，对于任何包含 0 的紧区间内的所有 δ，我们可以分别交换这类项中的期望与微分。因此，从上一个公式开始，我们可以

写出

$$\begin{aligned}
&\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f(\delta) = \frac{\mathrm{d}^s}{\mathrm{d}\delta^s} \mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0 \rangle}(X^k)} \right] - \frac{\mathrm{d}^s}{\mathrm{d}\delta^s} \mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\log \frac{p_{\boldsymbol{\mu}}(X^k \mid Z)}{p_{\langle \mu_0 \rangle}(X^k \mid Z)} \right] \\
&= \mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} \log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0 \rangle}(X^k)} \right] - \mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} \log \frac{p_{\boldsymbol{\mu}}(X^k \mid Z)}{p_{\langle \mu_0 \rangle}(X^k \mid Z)} \right] \\
&= \mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} \log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0 \rangle}(X^k)} \right] - \mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} \log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0 \rangle}(X^k)} + \log \frac{p_{\boldsymbol{\mu};[Z]}(Z)}{p_{\langle \mu_0 \rangle;[Z]}(Z)} \right] = \\
&\mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} \log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0 \rangle}(X^k)} \right] - \mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} \log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0 \rangle}(X^k)} \right] + \mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} \log \frac{p_{\boldsymbol{\mu};[Z]}(Z)}{p_{\langle \mu_0 \rangle;[Z]}(Z)} \right] \\
&= \mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} \log \frac{p_{\boldsymbol{\mu};[Z]}(Z)}{p_{\langle \mu_0 \rangle;[Z]}(Z)} \right],
\end{aligned}$$

其中，最后一行用到了所有涉及的项都是有限的这一事实。这正是我们需要证明的。