arXiv:2601.05420v1 [cs.LG] 8 Jan 2026

E噪声 LLM 作为评审评估的高效推断

Yiqun T. Chen* Sizhu Lu† Sijia Li‡ Moran Guo§ Shengyi Li¶2026年1月12日

摘要

大语言模型（LLM）越来越多地被用作生成式人工智能输出的自动评估器，这种范式通常称为“LLM 作为评审”。在实践中，LLM 评判器对潜在真实值的预测并不完美，还可能存在系统性、非随机误差。近期提出了两种主要方法来应对这一问题：(i) 基于错分模型（例如 Rogan–Gladen 型估计量）的直接测量误差校正，以及 (ii) 替代结果方法，例如预测增强推断（PPI），后者通过在一小组金标准人工标注上校准预测残差来校正偏差。本文系统研究了这两种方法在估计均值参数（例如平均基准分数或成对胜率）时的表现。借助半参数效率理论的工具，我们通过推导有效影响函数（EIF）型高效估计量的显式形式，统一了这两类估计量，并刻画了 PPI 型估计量的渐近方差严格小于测量误差校正的条件。我们通过模拟实验验证了理论结果，并用一个真实数据示例展示了这些方法。我们在 https://github.com/yiqunchen/debias-llm-as-a-judge 提供了基准测试方法及比较工具的实现。

1 引言

自动化模型评估是机器学习（ML）中一项至关重要但颇具挑战的任务，尤其是在生成式人工智能（AI）时代。可扩展性带来了海量待标注输出，输出形式也多种多样，从数值预测和分类标签，到推理轨迹、数学证明等开放式段落不一而足；传统的均方误差或误分类率等指标都无法充分衡量这些输出。此外，越来越多的研究开始利用大语言模型（LLMs）生成大量新问题及其对应答案，但逐一对每组问答进行人工核验并不可行。

受大语言模型近期进展的推动，越来越多的研究提出使用大语言模型作为自动“评判者”来评估其他模型的输出，例如判断模型的回答是否正确、分配质量分数，或从两个

*约翰斯·霍普金斯大学生物统计学系和计算机科学系，美国马里兰州巴尔的摩，邮编 21205；通讯邮箱：yiqun.t.chen@gmail.com

†加州大学伯克利分校统计学系，美国加利福尼亚州，邮编 94720

‡加州大学洛杉矶分校生物统计学系，美国加利福尼亚州，邮编 90095§约翰斯·霍普金斯大学生物统计学系，美国马里兰州巴尔的摩，邮编 21205

¶约翰斯·霍普金斯大学生物统计学系，美国马里兰州巴尔的摩 21205

候选回答中选出更优者，这种做法称为“LLM 作为评审”。这一方法有望实现可扩展且低成本的评估，但也带来了新的误差来源：大语言模型评审器并不能完美代表人类判断。总体而言，LLM 作为评审的工作流程通常依赖于：

- 一个大型评估集，其中仅有大语言模型判断可用，而
- 一个较小的校准集，其中既有人工标注，也有大语言模型判断。

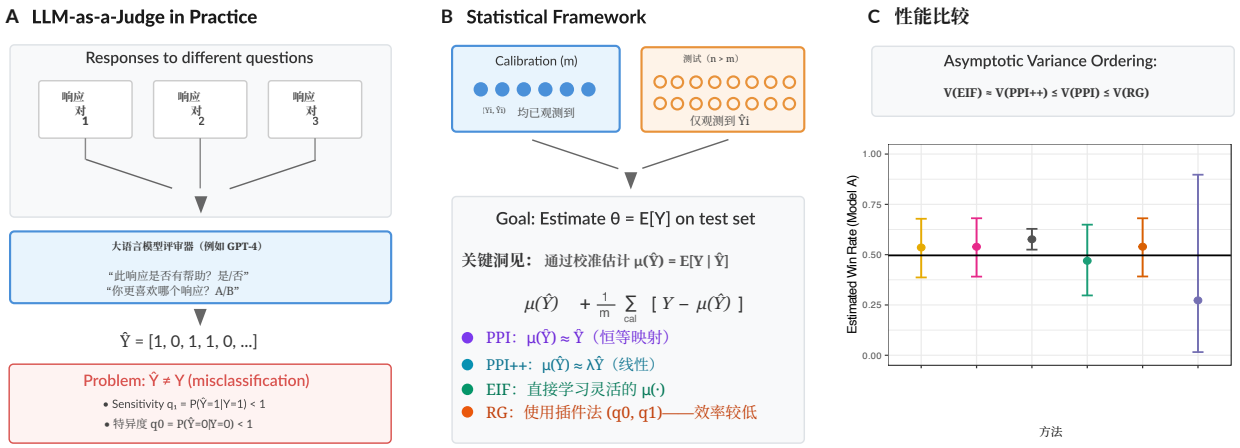
这种设置带来了一个挑战：仅依赖校准集会导致评估方差较大，而依赖评估集标签进行评估则会产生偏差，因为众所周知，大语言模型会生成同质化判断，并倾向于避开极端类别。因此，近期研究提出了一个关键问题：如何最优地结合这两个数据集，为评估指标构建有效且高效的估计，并提供经过校准的不确定性（例如置信区间）？

从统计学角度看，LLM 作为评审的评估可以表述为一个经典的测量误差问题。考虑二元结果情形，其中人工标签 Y 代表潜在真值，而 LLM 判断 Y 是一个含噪替代变量。对于一个较小的校准子集，我们观测到成对数据（Y_i, Y_i）；而对于大部分评估集，只有 Y_i 可用。该领域一种由来已久的方法是对错分机制 Pr($Y \mid Y$)建模，并采用直接测量误差校正来估计真实准确率，例如使用 Rogan–Gladen 估计量或混淆矩阵求逆 [Rogan 和 Gladen，1978年，Fuller， 2009]。另一类方法，包括预测增强推断（PPI） [Angelopoulos 等人，2023a] 及相关研究 [Wang et al., 2020, Miao et al., 2025, Salerno et al., 2025, Ji et al., 2025, Egami et al., 2023, Rister Portinari Maranca et al.,2025], , 将 LLM 判断视为 Y 的一般代理结果，并利用校准集 Y来校正其平均偏差。这种方法无需显式估计错分模型的参数，即可实现有效的统计推断。

尽管这两种方法传统均取得了进展，但针对其统计有效性和效率的直接比较仍然很少。这一缺口可能反映了这些方法通常应用于不同的文化和社群：近年来，测量误差模型主要用于医疗场景，例如在测量不完美的情况下估计 COVID-19 和死因患病率 [Rosin 等人，2023年，Gonz´alez 等人，2017年，Fiksel 等人， 2022], 而 PPI 风格的估计量则更常由现代机器学习/人工智能应用所推动。一个近期的例外是Lee 等人 [2025], , 他们为如何将测量误差校正应用于 LLM 作为评审场景提供了一套方案（详见第2.2节）。

虽然本文主要关注均值参数推断有效性的保证，但也有越来越多优秀的互补研究致力于改进 LLM 作为评审的ˆ ˆ评分（Y），以便在 $Y \approx Y$ 各种任务和大语言模型上 [Tian 等人，2025年，Liu 等人，2024年，Shi 等人，2024年，Sahoo 等人，2025年，Schroeder 和 Wood-Doughty，2024年，Li 等人， 2025]。我们认为这类研究与本文互为补充：我们的目标是无论 Y质量如何，都提供相应保证；而改进 Y 的校准，则可以进一步降低我们所考察的各估计量的方差（详见第2节）。

本文研究了这两种方法在估计均值参数方面的表现，并从半参数效率理论的视角将它们统一起来（概览见图1 ）：根据替代变量均值估计的基础有效影响函数（EIF）理论 $\widehat{\quad}$ $\widehat{\quad}$标签 Y，最优估计量取决于回归函数 $\mathbb{E}[Y \mid Y]$。这一视角$\widehat{\quad}$揭示了流行的基于校准的估计量可以理解为采用不同函数形式对 $\mathbb{E}[Y \mid Y]$ 的近似：PPI 隐式地使用恒等映射 $Y \approx Y$；PPI++ 使用一个 $\widehat{\quad}$ 带调节参数 $\widehat{\quad}$的单参数线性映射 $Y \approx \lambda Y \quad \lambda$；而 RePPI 则以灵活的参数化方法取代这些



ˆ 图 1：**校准 LLM 作为评审的评估**。（**A**）大语言模型生成的标签 Y 存在噪声，其灵敏度为 $q_1 < 1$，特异度为 $q_0 < 1$。（**B**）人工标注校准集可利用 PPI、Rogan–Gladen（RG）或基于 EIF 的估计量进行偏差校正。（**C**）比较朴素方法、RG、PPI 和基于 EIF 的估计量在评估 LLM 作为评审的表现时的效果。

ˆ形式，采用灵活的机器学习模型直接估计 $\mathbb{E}[Y \mid Y]$ [Ji 等人， 2025]。特别是在二元结果情形下，PPI++ 与最优 λ 相一致，基于 EIF 的估计量亦然；但对于更一般的结果，这种一致性未必成立。我们得出的这些见解与缺失数据高效推断的丰富文献密切相关，尤其是完全随机缺失的设定 [Chen等人，2008，Robins 和 Rotnitzky， 1995]。最近，Xu 等人 [2025]提出了一种通用的“安全PPI”框架，在感兴趣参数的维度为 PPI++时优于 PPI/ $p > 1$，并得出了类似结论：除非 ˆ PPI 的方差 ++ 具有严格的线性形式，否则可得到严格改进。我们的工作可以视为针对标量结果情形的平行探索与实现，该情形涵盖了最常见的应用，例如 LLM 作为评审。

从经验上看，在广泛的参数设定下，我们提出的基于 EIF 的估计量优于基础PPI；并且就有限样本偏差和置信区间的效率而言，这两种方法都优于 Rogan–Gladen 估计量。我们进一步从理论上解释了这些经验发现，证明了：（1）对于采用最优调节参数的二元结果，EIF 估计量与 PPI++ 渐近等价；（2）PPI 估计量的渐近方差严格小于 Rogan–Gladen 估计量的渐近方差。

我们的工作有助于连接经典统计学中关于估计效率理论日益丰富的研究成果，与应用人工智能领域中快速扩展的 LLM 作为评审实践及校准工作。随着大语言模型日益成为数据分析工作流程中不可或缺的一部分，我们的工作为大语言模型评判系统的可靠不确定性量化提供了实用且亟需的指导。

2 二元设置下的校准方法综述

2.1 问题设置

我们将其形式化为LLM 作为评审的评估如下：对于 $i = 1, \ldots\ldots, N$，我们观察到独立同分布的配对数据

$$Y_i \sim \text{伯努利分布}(\theta), \quad \text{denoting the human label}, \qquad \hat{Y}_i \in \{0,1\} \quad \text{denoting the LLM-as-a-judge labe} \tag{1}$$

目标参数是人工结果的均值

$$\theta = \mathbb{E}[Y] = \Pr(Y = 1). \tag{2}$$

作为一个具体示例，如果目标是比较不同大语言模型生成的两个回答 R_1 和 R_2，令 $Y_i = 1$ 表示人工标注员更偏好 R_1 而非 R_2 对提示 i 的回答（否则为 0）。那么估计目标 θ 表示 R_1 相对于 R_2的人工胜率。

大小为 N 的完整样本被划分为两个相互独立且不重叠的子集：大小为 $\hat{n}$ 的测试集，其中仅观测到代理结果 Y_i；以及大小为 $m = N - n$ 的校准集，其中同时观测到真结果 Y_j 和代理变量 Y_j。这一设置完全对应于完全随机缺失（MCAR）机制：对于测试集 [中的观测值，真结果 Y_i 缺失（Robins 和 Rotnitzky，1995；Little 和 Rubin， 2002], ），因为校准集与测试集的划分是通过独立的伯努利分配生成的，且该分配不依赖于 Y_i 或 Y_i。在渐近分析中，我们考虑 m 和 n 均随 $\frac{n}{m} \to \gamma_1 \in (0, \infty)$ 趋于无穷的情形。也就是说，校准集中的观测值比例收敛到一个固定常数 $\frac{1}{1+\gamma_1} \in (0, 1)$。 —

用以下符号表示测试集中经大语言模型评判器判断为阳性的样本比例：

$$\hat{\theta}_{\text{naive}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{Y}_i. \tag{3}$$

$\hat{}$ 当评判者并非完美（即 $q_0 + q_1 < 2$）时，有 $\mathbb{E}[\theta_{\text{naive}}] = \theta + (q_0 + q_1 - 2)\theta + (1 - q_0)$，这会导致点估计有偏、方差估计不正确，推断也无效。因此，我们回顾几种基于校准的去偏估计量：Rogan–Gladen（RG）估计量 [Rogan 和 Gladen，1978年，Lee 等人， 2025], 预测增强推断（PPI） [Angelopoulos 等人，2023a], 以及其功效调优变体（PPI++） [Angelopoulos 等人，2023b]。

2.2 Rogan–Gladen 估计量

Rogan–Gladen 估计量会考虑大语言模型评审器的灵敏度和特异度，

$$q_1 = \Pr(\hat{Y} = 1 \mid Y = 1), \qquad q_0 = \Pr(\hat{Y} = 0 \mid Y = 0). \tag{4}$$

$\hat{}$令 $p = \Pr(Y = 1)$ 表示大语言模型将一个实例判为阳性的边际概率。经典的错误分类关系为

$$p = q_1\theta + (1 - q_0)(1 - \theta) = (q_0 + q_1 - 1)\theta + (1 - q_0). \tag{5}$$

求解（5）得 θ，得到 $\theta = \frac{p+q_0-1}{q_0+q_1-1}$. 在实际应用中， p、 q_0和 q_1 均未知，需根据数据进行估计：$\hat{}p = \frac{1}{n}\sum_{i=1}^{n} \hat{}Y_i$。 在校准集上，记 $m_1 = \sum_{j=1}^{m} \mathbf{1}\{Y_j = 1\}$ 和 $m_0 = m - m_1$ 并定义

$$\hat{q}_1 = \frac{\sum_{j=1}^{m} \mathbf{1}\{\hat{Y}_j = 1, Y_j = 1\}}{m_1}, \qquad \hat{q}_0 = \frac{\sum_{j=1}^{m} \mathbf{1}\{\hat{Y}_j = 0, Y_j = 0\}}{m_0}.$$

针对 θ 的 Rogan–Gladen 估计量为

$$\hat{\theta}_{\mathrm{RG}} = \frac{\hat{p} + \hat{q}_0 - 1}{\hat{q}_0 + \hat{q}_1 - 1}. \tag{6}$$

这是存在测量误差时经典的患病率估计量 [Rogan 和 Gladen，1978年，Lang 和 Reiczigel，2014] 以及 Lee 等人在中提出的相同估计量 [2025]。

命题 1 （（ θ_{RG} 在（6）中的渐近分布）。 假设测试样本和校准样本相互独立，且 $q_0 + q_1 - 1 \neq 0$ 满足 q_0 和 q_1 定义于（4）。令 $N = n + m \to \infty$ 和$n/m \to \gamma_1$，则有

$$\sqrt{N}\left(\hat{\theta}_{\mathrm{RG}} - \theta\right) \xrightarrow{d} \mathcal{N}\big(0, V_{\mathrm{RG}}\big),$$

其中

$$V_{\mathrm{RG}} = \frac{1+\gamma_1}{\gamma_1} \frac{1}{(q_0+q_1-1)^2} \left[p(1-p) \; + \; \gamma_1 \left\{(1-\theta)\, q_0(1-q_0) + \theta\, q_1(1-q_1)\right\}\right]. \tag{7}$$

2.3 预测增强推断（PPI）

预测增强推断将大语言模型标签视为 Y 的替代变量，并利用校准集校正其偏差，而不对灵敏度和特异度进行显式建模。按照第 2.2 节中的设定，我们有 $\mathbb{E}[Y] = p$， $\mathbb{E}[Y - Y] = \Delta = p - \theta$，从而得到以下估计量：

$$\hat{\theta}_{\mathrm{PPI}} = \hat{p} - \frac{1}{m}\sum_{j=1}^{m}(\hat{Y}_j - Y_j). \tag{8}$$

命题 2. 联合模型 (1) 下，测试样本和校准样本相互独立 ：

(i) 对所有 n, m， *PPI* 都是无偏的：

$$\mathbb{E}[\hat{\theta}_{\mathrm{PPI}}] = \theta.$$

(ii) 有限样本方差为

$$\mathrm{Var}(\hat{\theta}_{\mathrm{PPI}}) = \frac{p(1-p)}{n} + \frac{(1-\theta)(1-q_0) + \theta(1-q_1) - (\theta-p)^2}{m}, \tag{9}$$

其中 $p = \Pr(Y=1) \; = (1-\theta)(1-q_0) \; + \theta q_1$。若 $N = n + m \to \infty$ 且 $n/m \to \gamma_1 \in (0, \infty)$，

$$\sqrt{N}\,(\hat{\theta}_{\mathrm{PPI}} - \theta) \xrightarrow{d} \mathcal{N}(0, V_{\mathrm{PPI}}),\; V_{\mathrm{PPI}} = \left(\frac{1+\gamma_1}{\gamma_1}\right)\left\{p(1-p) + \gamma_1\left[(1-\theta)(1-q_0) + \theta(1-q_1) - (\theta-p)^2\right]\right\}.$$

2.4 检验效能调节型 PPI (PPI++)

PPI++ 通过引入调节参数 $\lambda \in \mathbb{R}$进行推广：

$$\hat{\theta}_{\mathtt{PPI++}}(\lambda) := \frac{1}{m}\sum_{j=1}^{m} Y_j \; + \; \lambda\left(\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i \; - \; \frac{1}{m}\sum_{j=1}^{m}\hat{Y}_j\right). \tag{10}$$

不难看出，（10）对于任意独立于 λ 的 Y_i均保持无偏，但不同的选择会导致不同的渐近方差（因而检验效能也不同），详见命题 3。

命题 3。 假设 λ是固定的（不依赖于数据），在具有独立测试样本和校准样本的联合模型（1）下，则：

(i) (无偏性) 对所有 $n,\ m,$

$$\mathbb{E}\Big[\hat{\theta}_{PPI++}(\lambda)\Big] = \theta.$$

(ii) (有限样本方差)

$$\mathrm{Var}\Big(\hat{\theta}_{PPI++}(\lambda)\Big) = \frac{\theta(1-\theta)}{m} + \lambda^2\left(\frac{p(1-p)}{n} + \frac{p(1-p)}{m}\right) - \frac{2\lambda}{m}\,\theta\big(q_1 - p\big). \tag{11}$$

(iii) (渐近分布) 若 $N = n + m \to \infty$ 满足 $n/m \to \gamma_1 \in (0,\ \infty)$，则 n

$$\sqrt{N}\left(\hat{\theta}_{PPI++}(\lambda) - \theta\right) \xrightarrow{d} \mathcal{N}\big(0,\, V_{PPI++}(\lambda)\big)\,,$$

其中

$$V_{PPI++}(\lambda) = \frac{1+\gamma_1}{\gamma_1}\left[\gamma_1\,\theta(1-\theta) + \lambda^2(1+\gamma_1)\,p(1-p) - 2\lambda\,\gamma_1\,\theta\big(q_1 - p\big)\right]. \tag{12}$$

(iv) (方差最小化调优) $V_{PPI++}(\lambda)$ 在 $\lambda \in \mathbb{R}$ 处取得最小值

$$\lambda^\star = \frac{\gamma_1\,\theta\big(q_1 - p\big)}{(1+\gamma_1)\,p(1-p)}.$$

2.5 最大似然估计量

除 PPI/PPI++ 等（基于矩的）去偏估计量外，另一个自然的候选方案是在完整的三参数误分类模型下针对（Y, Y）使用联合最大似然估计量 ˆ。与 RG（在校准样本上使用灵敏度和特异度的插件估计量）不同，MLE 通过最大化模型在测试样本和校准样本上的联合似然，显式拟合（θ, q_0, q_1）。

在模型（1）下，测试观测值满足 ˆ

$$\Pr(Y=1,\hat{Y}=1) = \theta q_1, \qquad \Pr(Y=1,\hat{Y}=0) = \theta(1-q_1),$$
$$\Pr(Y=0,\hat{Y}=1) = (1-\theta)(1-q_0), \qquad \Pr(Y=0,\hat{Y}=0) = (1-\theta)q_0.$$

ˆ似然与估计方程。 令 $O_i = (R_i,\ Y_i,\ Y_i)$ 表示 $i = 1,\ \ldots\ldots,\ n+m$，其中 $R_i \in \{0,1\}$ 表示 Y_i 是否被观测到（$R_i = 1$ 表示校准，$R_i = 0$ 表示测试）。在此设定中，我们假设各个 R_i 均为独立同分布抽样，且 $\Pr(R = 1) = m/(n+m)$，观测数据似然可分解为校准部分（对应 $R = 1$）和仅测试部分（对应 $R = 0$）。记作 $p = (1-\theta)(1-q_0) + \theta q_1$，则每个观测的贡献为

$$f(O_i;\theta,q_0,q_1) = \Big[\Pr(Y_i,\hat{Y}_i)\Big]^{R_i}\Big[\Pr(\hat{Y}_i)\Big]^{1-R_i},$$

因此，完整的对数似然（忽略常数项）为

$$\ell(\theta, q_0, q_1) = \sum_{i=1}^{n+m} (1 - R_i)\Big\{\hat{Y}_i \log p + (1 - \hat{Y}_i) \log(1 - p)\Big\} + \sum_{i=1}^{n+m} R_i\Big\{\mathbf{1}\{Y_i = 1, \hat{Y}_i = 1\} \log(\theta q_1) + \mathbf{1}\{Y_i = 1, \hat{Y}_i = 0\} \log(\theta(1 - q_1)) + \mathbf{1}\{Y_i = 0, \hat{Y}_i = 1\} \log((1 - \theta)(1 - q_0)) + \mathbf{1}\{Y_i = 0, \hat{Y}_i = 0\} \log((1 - \theta)q_0)\Big\}. \quad (13)$$

最大似然估计定义为

$$(\hat{\theta}_{\mathrm{MLE}}, \hat{q}_{0,\mathrm{MLE}}, \hat{q}_{1,\mathrm{MLE}}) = \arg \max_{\theta, q_0, q_1 \in (0,1)} \ell(\theta, q_0, q_1),$$

等价地，它是得分方程的一个解

$$\frac{\partial}{\partial \theta}\ell(\theta, q_0, q_1) = 0, \qquad \frac{\partial}{\partial q_0}\ell(\theta, q_0, q_1) = 0, \qquad \frac{\partial}{\partial q_1}\ell(\theta, q_0, q_1) = 0.$$

在实践中，这些方程通过数值方法求解（例如使用牛顿–拉夫森法），并受约束 $q_0 + q_1 \neq 1$（可识别性）。

命题 4（最大似然估计的渐近正态性与效率）。 假设针对ˆ（Y, Y, R）的三参数错分模型设定正确，且满足标准正则条件。设 $N = n + m \to \infty$ *with* $n/m \to \gamma_1 \in$（0，∞），并定义

$$p = (1 - \theta)(1 - q_0) + \theta q_1, \qquad \eta = (\theta, q_0, q_1)^\top.$$

将基于合并后的测试（n）和校准（m）样本的最大似然估计记为

$$\hat{\eta}_{\mathrm{MLE}} := (\hat{\theta}_{\mathrm{MLE}}, \hat{q}_{0,\mathrm{MLE}}, \hat{q}_{1,\mathrm{MLE}})^\top.$$

Then

$$\sqrt{N}\,(\hat{\eta}_{\mathrm{MLE}} - \eta) \xrightarrow{d} \mathcal{N}\big(0,\ I_{\gamma_1}(\eta)^{-1}\big),$$

其中极限费舍尔信息为

$$I_{\gamma_1}(\theta, q_0, q_1) = \frac{\gamma_1}{1 + \gamma_1} I_{\mathrm{test}}(\theta, q_0, q_1) + \frac{1}{1 + \gamma_1} I_{\mathrm{cal}}(\theta, q_0, q_1),$$

with

$$I_{\mathrm{test}}(\theta, q_0, q_1) = \frac{1}{p(1 - p)} \begin{pmatrix} (q_0 + q_1 - 1)^2 & -(q_0 + q_1 - 1)(1 - \theta) & \theta(q_0 + q_1 - 1) \\ -(q_0 + q_1 - 1)(1 - \theta) & (1 - \theta)^2 & -\theta(1 - \theta) \\ \theta(q_0 + q_1 - 1) & -\theta(1 - \theta) & \theta^2 \end{pmatrix},$$

$$I_{\mathrm{cal}}(\theta, q_0, q_1) = \mathrm{diag}\left(\frac{1}{\theta(1 - \theta)},\ \frac{1 - \theta}{q_0(1 - q_0)},\ \frac{\theta}{q_1(1 - q_1)}\right).$$

因此 q y,

$$\sqrt{N}\,(\hat{\theta}_{\mathrm{MLE}} - \theta) \xrightarrow{d} \mathcal{N}\big(0,\ \big[I_{\gamma_1}(\eta)^{-1}\big]_{11}\big),\ \textit{where}$$

$$\big[I_{\gamma_1}(\eta)^{-1}\big]_{11} = (1+\gamma_1)\,\theta(1-\theta)\,\frac{p(1 - p) + \gamma_1\Big[(1 - \theta)q_0(1 - q_0) + \theta q_1(1 - q_1)\Big]}{p(1 - p) + \gamma_1\Big[(q_0 + q_1 - 1)^2\theta(1 - \theta) + (1 - \theta)q_0(1 - q_0) + \theta q_1(1 - q_1)\Big]}.$$

3 通过有效影响函数审视原则性效率

现在，我们有多个渐近无偏的估计量来估计 $\theta = \mathbb{E}[Y]$（RG、PPI、PPI++ 和最大似然估计）。这引出了一个自然的问题：

> 在给定统计模型下，可达到的最小渐近方差是多少，哪些估计量能够达到这一方差？

我们借助统计学中的半参数效率理论来回答这个问题（详细论述请参阅 Bickel 等人[1993], Van der Vaart [2000] 及其所引文献）。令 $Z_1, \ldots, Z_N$ 表示来自分布 P 的独立同分布观测值，并令 $\theta = \theta(P)$ 为目标参数。若存在均值为零的函数 $\phi_P(Z)$，使得 $\hat{\theta} - \theta(P) = \frac{1}{N}\sum_{i=1}^{N} \phi_P(Z_i) + o_p(N^{-1/2})$，则估计量 θ 是正则且渐近线性的。任何此类 ϕ_P 都称为 $\hat{\theta}$ 的影响函数（IF），直接应用中心极限定理可得

$$\sqrt{N}\big(\hat{\theta} - \theta(P)\big) \xrightarrow{d} \mathcal{N}\big(0,\ \mathrm{Var}_P(\phi_P(Z))\big).$$

在所有 IF 中，θ 在 P 下的有效影响函数（EIF）——若其存在——是渐近方差最小的唯一影响函数。这个最小方差称为半参数效率界。影响函数与 EIF 一致的估计量达到该界，因此称为半参数有效。

具体来说，P 指的是 $(X, Y, \hat{Y}, R)$ 的联合分布的一般模型，其中 $\hat{Y} \in \{0,1\}$ 是人工标签；$Y \in \{0,1\}$ 是大语言模型判定标签；$R \in \{0,1\}$ 表示某个观测值是否属于校准集（观测到 Y）或测试集（缺失 Y）。除观测值之间满足独立同分布假设外，不作其他约束，并且 R 与 $\hat{Y}$ 和 Y 独立。回顾一下，我们观测到 n 个测试点，仅有 Y；以及 m 个校准点，同时拥有 $(Y, \hat{Y})$。令 $R \in \{0,1\}$ 表示是否观测到 Y（校准集对应 $R = 1$，测试集对应 $R = 0$，且 $\Pr(R = 1) = \frac{m}{m+n}$），并定义

$$\mu(\hat{y}) := \mathbb{E}[Y \mid \hat{Y} = \hat{y}], \qquad \hat{y} \in \{0,1\}.$$

一种有助于理解 EIF 的方法是注意到，在此设计中，关于 $\theta = \mathbb{E}[Y]$ 有两个自然的信息来源：(i) 大型测试样本能够精确估计 $\hat{Y}$ 的分布；(ii) 校准样本告诉我们 Y 与 $\hat{Y}$ 的关系。函数 $\mu(\hat{Y})$ 正是将 $\hat{Y}$“转换”为 Y 估计值的最佳方法：根据迭代期望定律，$\theta = \mathbb{E}[Y] = \mathbb{E}\{\mathbb{E}[Y \mid \hat{Y}]\} = \mathbb{E}[\mu(\hat{Y})]$。因此，一种自然的“先插补再求平均”估计量，是在（大型）测试集上对 $\mu(\hat{Y})$ 求平均。不过，由于 $\mu(\cdot)$ 未知，我们需要用校准样本来学习它。

命题5（有效影响函数）。给定 $(Y, \hat{Y}, R) \sim P$，其中 R 与 Y 独立，且 $\hat{Y}$，则 $\theta = \mathbb{E}[Y]$ 在 P 下的有效影响函数为

$$\phi^\star(R, Y, \hat{Y}; P) = \mu(\hat{Y}) - \theta + \frac{R}{\Pr(R = 1)}\big\{Y - \mu(\hat{Y})\big\}. \tag{14}$$

命题 6。定义

$$\hat{\mu}(1) := \frac{\sum_{j=1}^{m} \mathbf{1}\{\hat{Y}_j = 1\}\, Y_j}{\sum_{j=1}^{m} \mathbf{1}\{\hat{Y}_j = 1\}},\ \hat{\mu}(0) := \frac{\sum_{j=1}^{m} \mathbf{1}\{\hat{Y}_j = 0\}\, Y_j}{\sum_{j=1}^{m} \mathbf{1}\{\hat{Y}_j = 0\}},\ \hat{\mu}(\hat{Y}) = \hat{\mu}(0)\mathbf{1}\{\hat{Y} = 0\} + \hat{\mu}(1)\mathbf{1}\{\hat{Y} = 1\}.$$

那么，在 θ 下估计(1) 的渐近有效估计量可写为

$$\hat{\theta}_{\mathrm{EIF}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\mu}(\hat{Y}_i) \;+\; \frac{1}{m}\sum_{j=1}^{m}\Big(Y_j - \hat{\mu}(\hat{Y}_j)\Big), \tag{15}$$

其渐近分布为

$$\sqrt{N}\big(\hat{\theta}_{\mathrm{EIF}} - \theta\big) \xrightarrow{d} \mathcal{N}\big(0, V_{\mathrm{EIF}}\big), \tag{16}$$

其中，二元情形下的方差可简化为

$$V_{\mathrm{EIF}} = \frac{\theta(1-\theta)}{p(1-p)}\left[\theta(1-\theta)(q_0+q_1-1)^2 + (\gamma_1+1)\big(q_1(1-q_1)\theta + q_0(1-q_0)(1-\theta)\big)\right]. \tag{17}$$

PPI 通常并非有效。 一般效率理论表明，EIF 估计量 θ_{EIF} 在(15) 中，在 (Y, Y, R) 的一般独立同分布模型下渐近有效 $\sim P$。因此，在该模型中，式 (17) 中的渐近方差是所有正则渐近线性估计量中最小的。此外，经过一些代数运算可以验证，第2节介绍的所有估计量都是正则且渐近线性的。特别地，PPI 对应于式 ($\mu(Y) = Y$) 中的特殊选择 14)。这产生了一个无偏估计量，但除非 $\mathbb{E}[Y \mid Y] = Y$，否则通常并非有效；在这种情况下， Y 是 Y的完美预测。

经过最优调优的PPI++ 在二元情形下与 EIF 相匹配。 与 PPI 不同，PPI++在一个由无偏估计量构成的单参数族中进行搜索，并选择 λ 以最小化渐近方差。一般而言，PPI++ 与 EIF 估计量并不一致，因为 $\mathbb{E}[Y \mid Y]$ 不一定等于 λY对于某个单一标量而言的取值 λ。然而，当 Y， $Y \in \{0, 1\}$时，回归函数 $\mu(Y) = \mathbb{E}[Y \mid Y]$总是关于 Y的仿射函数，因此，PPI++ 族足够丰富，可以与 EIF 估计量相匹配。事实上，令$\gamma_1 = n/m$，命题3 中的方差最小化选择可写为

$$\lambda^\star = \frac{\gamma_1 \operatorname{Cov}(Y, \hat{Y})}{(1+\gamma_1)\operatorname{Var}(\hat{Y})} = \frac{\gamma_1\,\theta(1-\theta)\,(q_0+q_1-1)}{(1+\gamma_1)\,p(1-p)}, \qquad p = (1-\theta)(1-q_0) + \theta q_1.$$

另一方面，根据贝叶斯法则，

$$\mu(1) = \Pr(Y=1 \mid \hat{Y}=1) = \frac{\theta q_1}{p}, \qquad \mu(0) = \Pr(Y=1 \mid \hat{Y}=0) = \frac{\theta(1-q_1)}{1-p}. \tag{18}$$

因此，

$$\begin{aligned}\mu(1) - \mu(0) &= \frac{\theta q_1}{p} - \frac{\theta(1-q_1)}{1-p} = \theta\,\frac{q_1 - p}{p(1-p)} \\ &= \theta\,\frac{q_1 - \big[(1-\theta)(1-q_0) + \theta q_1\big]}{p(1-p)} = \frac{\theta(1-\theta)(q_0+q_1-1)}{p(1-p)}.\end{aligned}$$

因此，$\lambda^\star = \frac{\gamma_1}{1+\gamma_1}\big\{\mu(1) - \mu(0)\big\}$，这正是 $\mu(Y)$仿射形式的斜率，表明经过最优调节的 PPI++ 与二元设置下的 EIF 估计量渐近等价。

$\hat{}$最大似然估计等价于 EIF 在模型设定正确的情况下（Y、Y、R）$\sim P$，经典似然理论表明，最大似然估计对于 θ（命题4）是渐近有效的，因此达到了可能的最小渐近方差 [Bickel 等人，1993]。在二元结果情形下，模型由三个参数（θ、q_0、q_1）饱和，因此我们预期 EIF 估计量的渐近方差将与最大似然估计的渐近方差一致。证明方差表达式等价性的详细代数推导留待附录。

RG 的效率低于 PPI。 最后，比较（6）和（8），可以得到命题7中的以下结果。

命题 7。 固定 $\theta \in (0、1)$ 以及 q_0、$q_1 \in (0、1)$，其中 $q_0 + q_1 - 1 > 0$；并假设 n、$m \to \infty$ 满足 $n/m \to \gamma_1 \in (0、\infty)$。则

$$V_{\mathrm{PPI}} \le V_{\mathrm{RG}},$$

其中

$$\frac{\gamma_1}{1+\gamma_1} V_{\mathrm{PPI}} = p(1-p) + \gamma_1 \Big[(1-\theta)(1-q_0) + \theta(1-q_1) - (\theta - p)^2\Big]$$

and

$$\frac{\gamma_1}{1+\gamma_1} V_{\mathrm{RG}} = \frac{1}{(q_0+q_1-1)^2}\left[p(1-p) \;+\; \gamma_1 \left\{(1-\theta)\, q_0(1-q_0) + \theta\, q_1(1-q_1)\right\}\right]$$

$\hat{}$对所有此类参数均成立，当且仅当 $q_0 = q_1 = 1$ 时取等号（即 $\Pr(Y = Y) = 1$）。

渐近效率总结。 上述推导的关系表明，性能的排序十分明确。在 (Y, Y) 的一般独立同分布模型下，EIF 估计量 θ_{EIF} 达到半参数效率界。在二元情形下，经过最优调优的 PPI++ 渐近等价于 θ_{EIF}，因此也等价于最大似然估计。形式上，对于 $N = n + m \to \infty$ 和 $\frac{n}{m} \to \gamma_1$：

$$\hat{\theta}_{\mathrm{EIF}} \;=_{\mathrm{asymp}}\; \hat{\theta}_{\mathrm{MLE}} \;=_{\mathrm{asymp}}\; \hat{\theta}_{\mathtt{PPI++}}(\lambda^\star) \;\prec_{\mathrm{asymp}}\; \hat{\theta}_{\mathrm{PPI}} \;\prec_{\mathrm{asymp}}\; \hat{\theta}_{\mathrm{RG}},$$

其中，$=_{\mathrm{asymp}}$ 表示相差至多 o_p（$n^{-1/2}$），而 $\prec_{\mathrm{asymp}}$ 表示渐近方差更小。

4 $\hat{}$超越二元情形：$\mathbb{E}[Y \mid Y]$的非参数估计

$\hat{}\,\hat{}$ 在二元情形下，$\mu(Y) = \mathbb{E}[Y \mid Y]$ 必然是仿射函数，因此经过最优调优的 PPI++ 能够达到效率界。此外，(Y, Y)上的非参数分布等价于二元情形下用于最大似然估计的三参数分布；这种“饱和”性质使最大似然估计能够达到效率界，而无需作出任何限制性参数假设。

对于更加高维的替代输出 Y（例如高维替代输出序数评分），EIF 的函数形式，因此在（15）中的形式，保持不变。然而，回归函数 $\mu(\cdot)$ 可能是非线性的，而效率取决于估计量 $\widehat{\mu}(\cdot)$ 与真条件均值函数的接近程度。

$\hat{}\,\hat{}$ 当 Y 和 Y 均为一维时，估计 $\mu(Y)$ 可归结为一个单变量回归问题，可以使用从简单线性回归到灵活的非参数平滑器等模型来解决。给定一致估计量 $\hat{}\mu(\cdot)$ 和独立同分布模型 $(Y, \hat{}Y, R)$，我们便可以构造代入式 EIF 估计量：

$$\hat{\theta}_{\mathrm{EIF}} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\mu}(\hat{Y}_i) + \frac{1}{\hat{P}(R=1)} R_i\{Y_i - \hat{\mu}(\hat{Y}_i)\}\right). \tag{19}$$

如果我们设计标注过程 R，例如，以概率 $P(R=1)=\frac{\gamma_1}{1+\gamma_1}$ 进行标注，那么只要 $\hat{}\mu(\cdot)$ 是 $\mu(\cdot)$ 的一致估计量，(19) 就是渐近有效的。

最后需要指出的是，与二元情形（或具有有限类别的分类情形）不同，在连续情形下，如果不对 Y 和 Y 之间的关系作出额外的参数假设，我们无法找到与最大似然估计或 Rogan–Gladen 估计量的直接类比。

5 模拟研究

5.1 二元 Y

我们在模拟中比较以下估计量的表现：式（3）中的朴素估计量、式（6）中的 Rogan–Gladen 估计量、式（8）中的 PPI 估计量、式（PPI++ ）中的估计量、式（10）中的最大似然估计量，以及式（13）中的EIF估计量（15）。

5.2 模拟设置

数据生成过程。 每次重复试验由 $N=2{,}000$ 个独立同分布样本组成。真二元标签生成为 $Y_i \overset{\text{独立同分布}}{\sim}$ 伯努利分布(θ)，其中 $\theta \in \{0.1, 0.2, \ldots, 0.9\}$ 为目标患病率。

作为评判者的[illegible]的错分模型生成，分别为 $\Pr(Y_i=1 \mid Y_i=1)=q_1$ 和 $\Pr(Y_i=0 \mid Y_i=0)=q_0$。

参数网格。我们通过完整的析因设计改变以下参数：

- 真实患病率： $\theta \in \{0.1, 0.2, \ldots, 0.9\}$。
- 评判者准确率： q_0， $q_1 \in \{0.6, 0.7, 0.8\}$。
- 标注预算（校准规模）： $\{1\%, 5\%, 10\%\}$ 的 N（即期望获得 $m \in \{20, 100, 200\}$ 个人工标签）。

校准集通过随机抽样 从完整数据集中获得；因此，校准集中的类别计数 m_0和 m_1 是随机的，并取决于基础患病率 θ。对于每种配置，我们进行 $B=1{,}000$次蒙特卡洛重复试验，以计算偏差，以及90%置信区间的覆盖率和平均宽度。

置信区间构造。 我们对所有置信区间应用逻辑斯蒂变换，以改善 θ 接近边界时的有限样本覆盖率 [Brown 等人，2001年，Stone， 1996]。 $\hat{}\hat{}$给定点估计 θ 和方差估计 V，我们构造

$$\mathrm{CI}_{1-\alpha}=\operatorname{expit}\left(\operatorname{logit}(\hat{\theta}) \pm z_{\alpha/2}\cdot\frac{\sqrt{\widehat{V}}}{\hat{\theta}(1-\hat{\theta})}\right),$$

其中，logit 尺度上的标准误通过德尔塔法获得。

评估指标。对于每个估计量 $\hat{\theta}$，令 $\hat{\theta}^{(b)}$ 表示其在重复试验 $b = 1, \ldots, B$中的取值，并令 CI $^{(b)} = [L^{(b)}, U^{(b)}]$ 表示其对应的 90% 置信区间。我们计算：

- **偏差：** $\text{Bias}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^{(b)} - \theta\right)$;
- **覆盖概率：** $\text{Cov}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\mathbf{1}\left\{ L^{(b)} \leq \theta \leq U^{(b)} \right\}$;
- **平均置信区间宽度：** $\text{Width}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\left(U^{(b)} - L^{(b)}\right)$.

5.3 对称评判准确率下的结果（$q_0 = q_1$）

图 2–4 按行对应不同准确率水平 $q_0 = q_1 \in \{0.6, 0.7, 0.8\}$、按列对应不同标注预算进行排列。

偏差。 图 2显示，除朴素估计量外，所有估计量的平均偏差均小于朴素样本均值估计量。回顾一下，朴素估计量的偏差为 $(1 - q_0) + \theta(q_0 + q_1 - 2)$，当 $\theta > 0.5$ 较大时为负，当 $\theta < 0.5$较小时为正，这与我们的模拟估计结果一致。在其余估计量中， θ_{RG} 表现出相对较大的有限样本偏差。

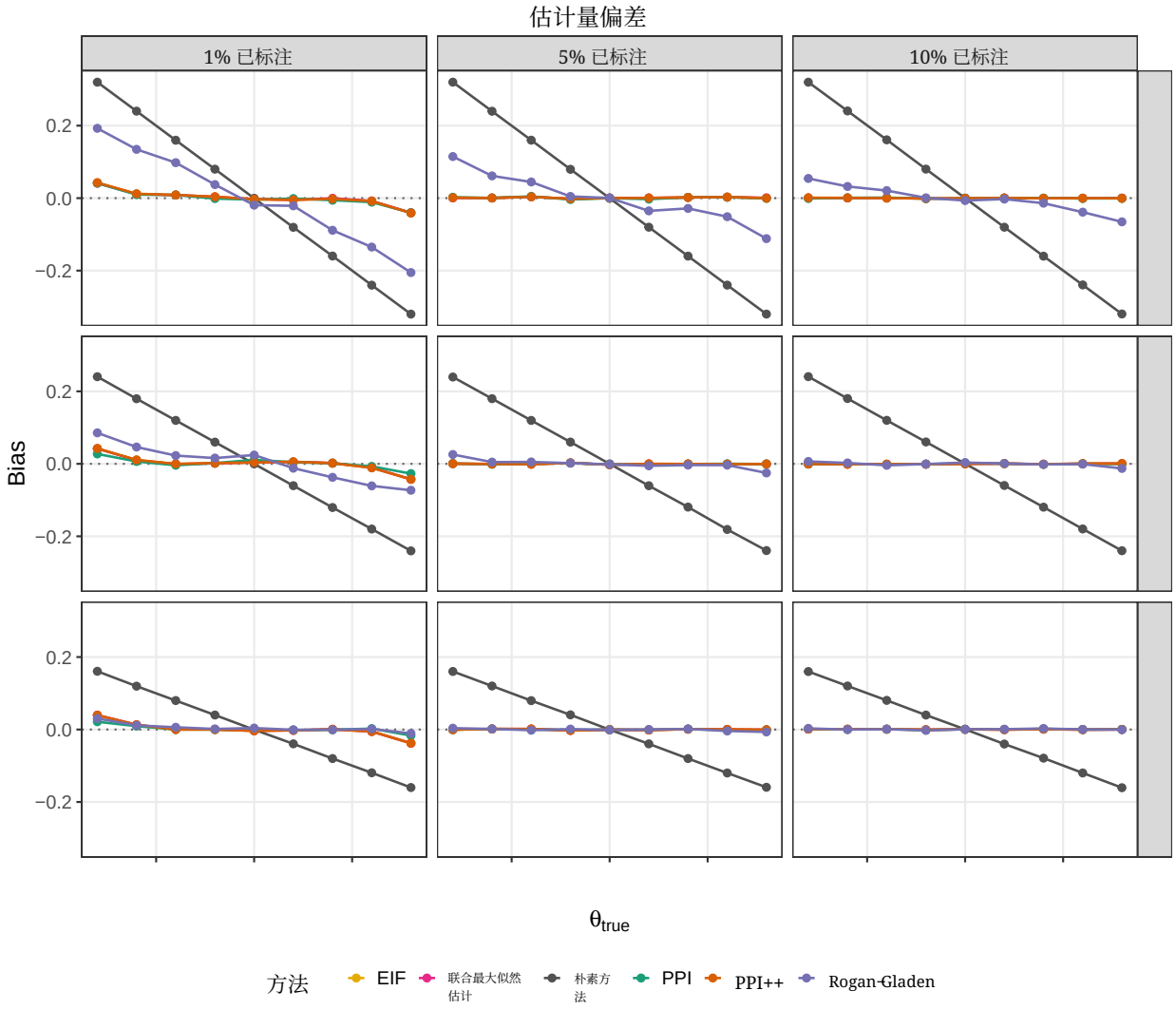
覆盖率。图3报告了经验覆盖率。朴素估计量出现了严重的覆盖不足，这与其渐近无效性一致（在 $\theta = 0.5$ 朴素估计量无偏，因此覆盖率更接近名义水平）。所有其他估计量在各种设置下都保持了接近名义水平的覆盖率。RG 估计量及其相应的置信区间往往覆盖过度，尤其是在已标注观测数量较少时（即校准集有限时）。

置信区间宽度。图4 比较了有效估计量的区间宽度（即排除 $\hat{\theta}_{\text{naive}}$ ）。正如我们在第3节的理论分析所预测，`EIF` 和`PPI++` 产生的区间几乎相同，且始终最窄，通常比标准`PPI`窄 35–55%。`MLE` 区间稍宽，但总体上与`EIF` 和`PPI++`相当。低准确率评判者会导致 Rogan–Gladen 区间大幅变宽，这是由于其渐近方差中的 $1/(q_0+q_1-1)^2$ 放大因子所致——在 $q_0 = q_1 = 0.6$时，RG 区间大约比 $10\times$ 宽`EIF`/`PPI++`。

5.4 连续 Y

此处，我们考虑 Ji 等人 [2025] 其中 Y 取离散值（例如，1–5 量表上的大语言模型评分），而真实结果 Y 是连续的。当大语言模型被用作评判者并给出序数评分时，这种情形很常见，但最终目标是推断模型的连续质量尺度 Y。

数据生成过程。 令 $Z \sim \text{Unif}(\{1, 2, 3\})$ 表示一个潜在混合成分，其条件分布为 $Y \mid Z \sim \mathcal{N}(\mu_Z, \sigma^2)$，其中 $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)^\top$。替代预测在成分 Z 上为 $\widehat{\ }$，很容易看出，我们需要谨慎地对 $\mathbb{E}[Y|Y]$ 建模。即使 $\widehat{\ }$预测 Y 3 $\mathbb{E}[Y]$ 能完美识别潜在类别，式（ ）中的朴素估计量对 的估计仍有偏差，因为 Y 和 Y 的尺度不同。具体而言，偏差为 $2 - (\mu_1 + \mu_2 + \mu_3)/3$，当 μ_k 值偏离其类别索引时，该偏差会增大。



ˆ 图 2： θ的估计量偏差。所有去偏估计量都实现了接近零的偏差；朴素估计量 ˆ（红色）在许多设置下表现出较大的偏差。当仅有 1% 的数据已标注时， θ_{RG} 也表现出相当大的偏差。

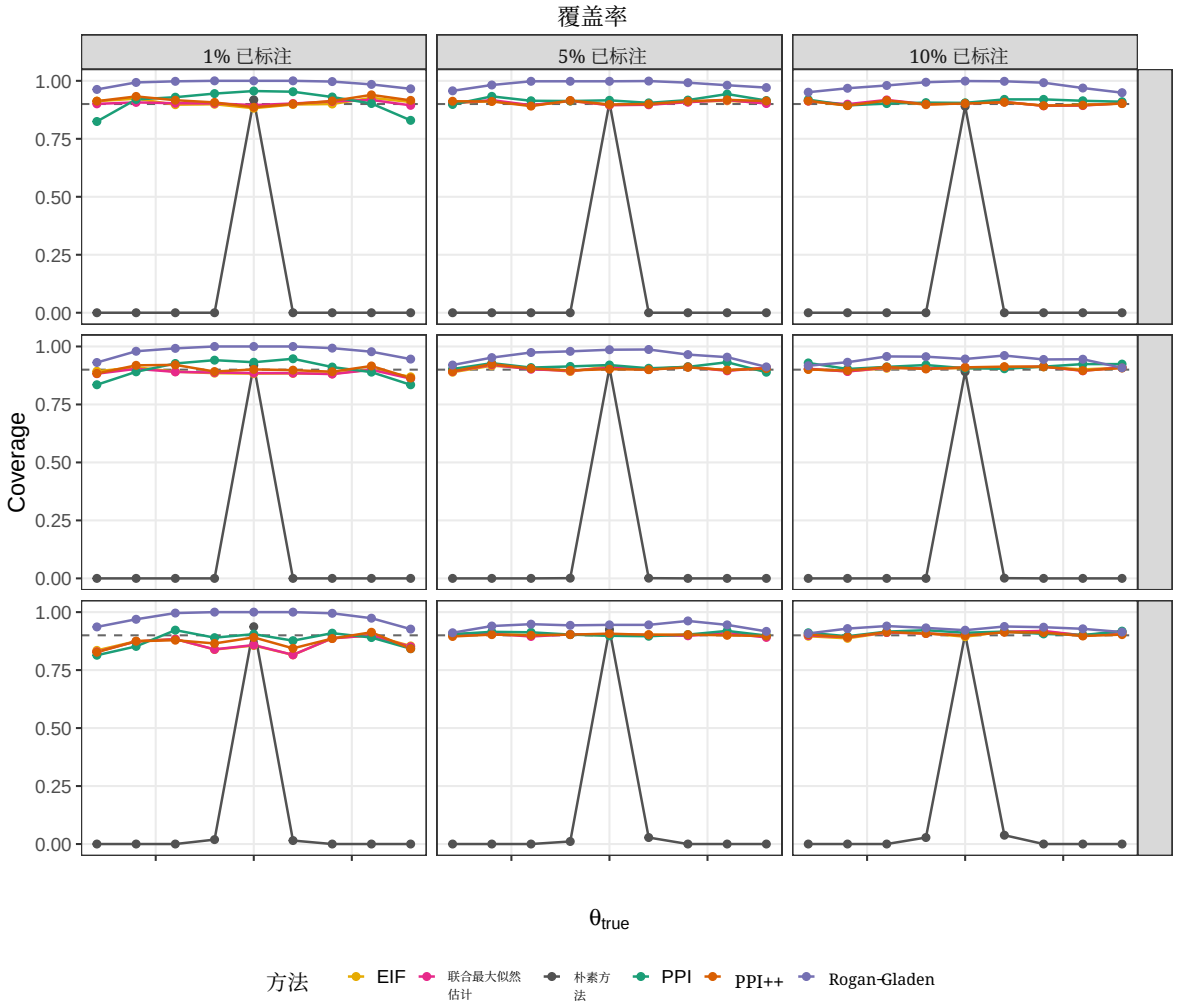


图 3：对称情形 $q_0 = q_1$ 下的经验覆盖率。朴素估计量表现出严重的覆盖不足。RG、PPI 和 PPI++通常能达到名义覆盖率；不过，当标注比例 m/n 较小且评判者质量较低（即 $q_0 + q_1$较小）时，RG 往往会出现覆盖过高的情况。

模拟设计。 我们固定 $N = 2000$, $\mu_1 = 1$, $\mu_2 = 2$, $\sigma = 1$，并改变 $\mu_3 \in \{3, 4, \ldots, 9\}$ 以控制偏差幅度。标注比例依次为 {5%、10%、20%}。我们报告 90%置信区间的覆盖率，以及 $B = 500$ 次重复试验中的均方根误差。

这里，我们比较式（3）中的朴素估计量、式（8）中的预测增强估计量、经检验效能调优的 PPI++ 估计量（式10）、EIF 估计量（式19）的性能。对于 EIF 估计量，$\hat{}$ $\hat{}$ 我们使用以下三种变体来近似 $\mathbb{E}[Y|Y]$: （ 1 $\hat{}$ $g\ z\ = \mathbb{E}[Y \mid Y = z]$ ）EIF：针对每个 $z \in \{1, 2, 3\}$分别学习（）；（2）线性 EIF：通过线性回归拟合 g（Y） $= a + b \cdot Y$（请注意，这与 PPI++渐近等价）；（3）EIF-GAM/样条：采用 $\hat{}$进行非参数校准，将 Y 视为平滑函数。由于我们没有对 Y施加额外的分布假设，因此在连续情形下无法写出参数似然函数，也就无法轻易将 Rogan–Gladen 和最大似然估计量扩展到这一设定。

$\hat{}$结果。图5展示了推断性能如何取决于 $\mathbb{E}[Y \mid Y]$ 的估计方式。$\hat{}$ $\hat{}$ $\hat{}$当 $\boldsymbol{\mu} = (1, 2, 3)$ 时，条件均值关于 Y 呈线性，因此使用 $\mathbb{E}[Y \mid Y] = Y$ 的模型设定正确，除朴素估计量外，所有方法都达到了名义覆盖率。此外，各方法的置信区间宽度几乎没有差异，因为 EIF、PPI 和 PPI++ 渐近等价。然而，随着 μ_3 增大，条件均值不再是线性的。虽然所有去偏估计量都保持了正确的覆盖率，但效率梯度十分明显：PPI 的效率最低，$\hat{}$其次是 PPI++ 和使用线性模型估计 $\mathbb{E}[Y \mid Y]$ 的 EIF（在对 PPI++ 进行最优调节时，这两者渐近$\hat{}$等价）。最后，使用灵活模型估计 $\mathbb{E}[Y \mid Y]$，例如 GAM（ [Wood 2012] ）或样条，或者拟合一个三分类模型，针对每个 $z \in \{1, 2, 3\}$ 分别学习 g （ z ） $= \mathbb{E}[Y \mid Y = z]$，都能达到效率最高的水平。

6 真实数据应用

6.1 数据与实验设置

我们在 arena-human-preference-140k数据集中的真实人类偏好数据上评估这些估计量。该数据集采集自 Chatbot Arena [Chiang 等人， 2024], 这是一个在线平台，用户可以比较两个人工智能模型的响应并选出胜者。我们聚焦于三个出现频率较高的模型对，它们均以 Claude Opus 4 作为模型 A：(1) Claude Opus 4 与 Gemini 2.5 Flash（$n = 493$）；(2) Claude Opus 4 与 Gemini 2.5 Pro（$n = 414$）；(3) Claude Opus 4 与 Qwen3-235B（$n = 494$）。

对于每场对决，真实标签 $Y_i = 1$ 表示人工评审员选择了模型 A（Claude）， $Y_i = 0$则表示并非如此。我们使用两个大语言模型评审器 GPT-4o-mini 和 GPT-5.2，根据向人工评审员展示的相同提示-响应对生成代理标签$\tilde{}$ $\tilde{Y}_i \in \{0, 1\}$ 。

为模拟这样一种现实场景：我们只能获取部分偏好数据，使用数据的 10% 随机样本作为带标签校准集，以估计 $\mathbb{E}[Y \mid Y]$ 或误分类概率（q_0， q_1），其余 90% 作为无标签测试集。所有置信区间均在 90% 的名义水平下构建。

6.2 结果

评判者质量。 表 1 报告了每种评判者–数据集组合的估计特异度（$\hat{}q_0$）和灵敏度（$\hat{}q_1$），这些数值基于随机抽取的 10% 校准集计算得出。条件$q_0 + q_1 > 1$ 表示其判别能力超出随机猜测。

表1：根据校准数据估算的评判者质量指标（$\hat{q}_0$：特异度；$\hat{q}_1$：灵敏度）。 $q_0 + q_1$ 的值接近 1 和 2 分别表示表现接近随机水平和接近完美。

模型对	评判者	$\hat{q}_0$	$\hat{q}_1$	$\hat{q}_0 + \hat{q}_1$
Claude 4 与 Gemini 2.5 Flash	GPT-4o-mini	0.74	0.69	1.43
Claude 4 对比 Gemini 2.5 Flash	GPT-5.2	0.47	0.67	1.14
Claude 4 对比 Gemini 2.5 Pro	GPT-4o-mini	0.50	0.81	1.31
Claude 4 对比 Gemini 2.5 Pro	GPT-5.2	0.64	0.82	1.46
Claude 4 与 Qwen3-235B	GPT-4o-mini	0.44	0.70	1.14
Claude 4 与 Qwen3-235B	GPT-5.2	0.50	0.64	1.14

评判者质量在不同设置下存在显著差异。GPT-4o-mini 在评估 Claude 4 与 Gemini 2.5 Flash 的比较时表现最佳，但在评估 Claude 4 与 Gemini 2.5 Pro 的比较时表现较差。GPT-5.2 在 Claude 4 与 Gemini 2.5 Pro 的比较中取得了最高的总体判别能力。两种评判者在 Claude 4 与 Qwen3-235B 的比较中表现不佳，尽管灵敏度中等，特异度却很低。这些具有异质性的错误率进一步凸显了偏差校正的必要性：朴素估计会继承评判者的系统性误差。

单个数据划分上的示例。 图 6 展示了三个模型对在一次随机校准/测试划分中的点估计和 90% 置信区间。黑色水平线表示根据所有人工标注计算得出的真实胜率。我们观察到：(i) 朴素估计量始终存在偏差；当使用 GPT-4o-mini 作为评判者评估 Claude 4 Opus 时，它高估了真实胜率；(ii) Rogan–Gladen 估计量给出的区间极宽，在实际应用中基本没有参考价值；(iii) `PPI++`、`EIF`和最大似然估计量得出的点估计相近，区间宽度适中，且覆盖真实值。

覆盖率与效率。 为评估覆盖率特性，图 7绘制了 1,000 次随机校准/测试划分中，各估计量的经验覆盖率与平均置信区间宽度。朴素估计量并不可靠，其覆盖率会随评判者偏差的大小和方向而剧烈变化：当系统性误差恰好与真实患病率一致时（例如，GPT-5.2 评判 Gemini 2.5 Pro），它几乎达到完美覆盖；但当偏差较大时，覆盖率会降至 0%，例如 GPT-4o-mini 评判 Gemini 2.5 Pro，以及两位评判者评判 Qwen3-235B 的情况。

相比之下，PPI++、EIF 和 MLE 在所有评判者—数据集组合中均实现了接近名义覆盖率的表现，平均区间宽度为 0.27–0.30。PPI 的覆盖率略高于名义值，区间也更宽（0.35–0.44）；而 Rogan–Gladen 则非常保守，以区间极宽（0.76–0.96）为代价实现了 100% 的覆盖率。在覆盖率有效的方法中，PPI++、EIF 和 MLE 的覆盖率—宽度权衡最佳，这与我们的理论预测一致：它们在达到名义覆盖率的同时，区间宽度约为 Rogan–Gladen 的三分之一。

7 讨论

我们通过有效影响函数，为理解 LLM 作为评审的评估中的校准提供了一个统一视角。我们通过理论分析和模拟实验比较了两种互补策略：通过错分模型直接进行测量误差校正，以及基于代理变量的方法（如 PPI），后者将大语言模型判断视为代理结果。当目标是均值结果（这在 LLM 作为评审的场景中很常见）时，RG 和 PPI 变体都能提供简单可靠的解决方案；鉴于其方差优势，优先推荐 PPI++/基于 EIF 的方法。具体而言，对于二元结果，经过最优调优的 PPI++ 等同于基于 EIF 的策略。

自然可以进行若干扩展。一个重要方向是研究实例依赖型误分类，即 LLM 的错误率会随输入特征或示例难度而变化。将我们的框架扩展到这一情形，与半参数高效推断文献中的做法类似：在 $m(X,\ \hat{Y}) = \mathbb{E}[Y \mid X, \hat{Y}]$ 下，我们的最终估计量涉及如下形式的影响函数

$$\phi^{\star}(O) = m(X, \hat{Y}) - \theta + \frac{R}{\Pr(R = 1 \mid X, \hat{Y})}\{Y - m(X, \hat{Y})\}, \tag{20}$$

另一个方向是开发多评审集成方法，将不同类型的 LLM 评估器组合起来，以提高稳健性。这需要仔细处理错误率各异的多个评判者（$q_0,\ q_1$）在（4）中的情形，并以有原则的方式聚合其代理标签 $\hat{Y}$（例如通过最优加权或学习得到的组合规则）。

最后，对分布偏移的稳健性——包括协变量偏移（例如校准集与测试集之间响应风格的系统性差异）和标签偏移（例如校准与部署期间真实标签 Y 患病率的变化，如将模型推广到新的文化或地理语境时）——值得开展更深入的方法学和实证研究；这些理论与计算方面的进展留待未来工作。

随着 LLM 作为评审的评估在模型开发和科学评估中日益重要，我们希望这些结果能够为规范的校准、有效的不确定性量化以及有限人工标注的高效利用提供实用基础。

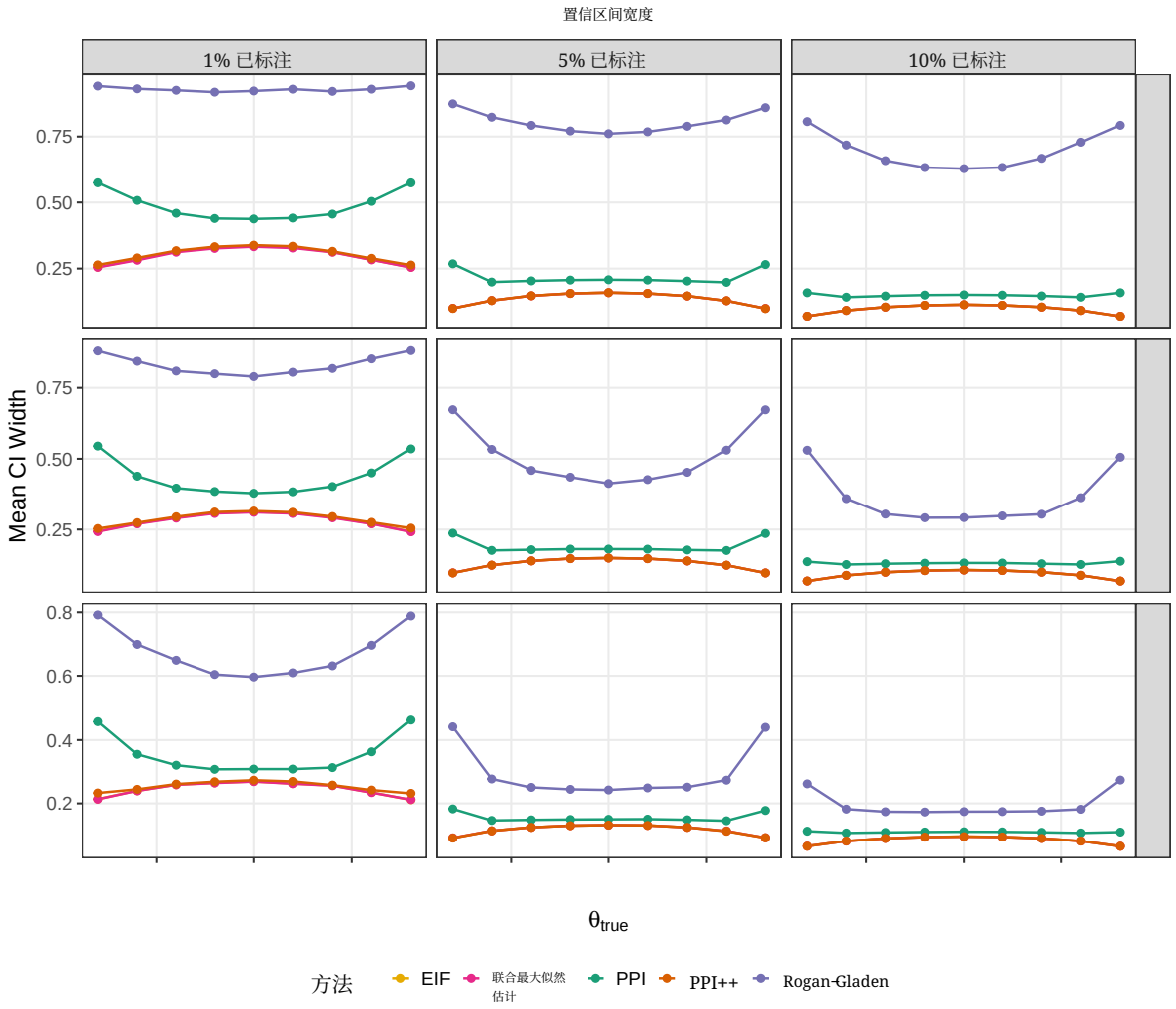


图4：对称情形 $q_0 = q_1$下的平均置信区间宽度。EIF 和 PPI++ 产生几乎相同且最窄的置信区间，相较于 Rogan–Gladen，其表现优越约 3–15$\times$ 倍，具体取决于标注比例。当 $q_0 + q_1 - 1$ 较小时（即大语言模型评判器更接近随机猜测时），这一优势最为明显。最大似然估计产生的置信区间略宽，但覆盖率更接近标称值。由于朴素置信区间过窄，无法达到预期覆盖率，因此不予列出。

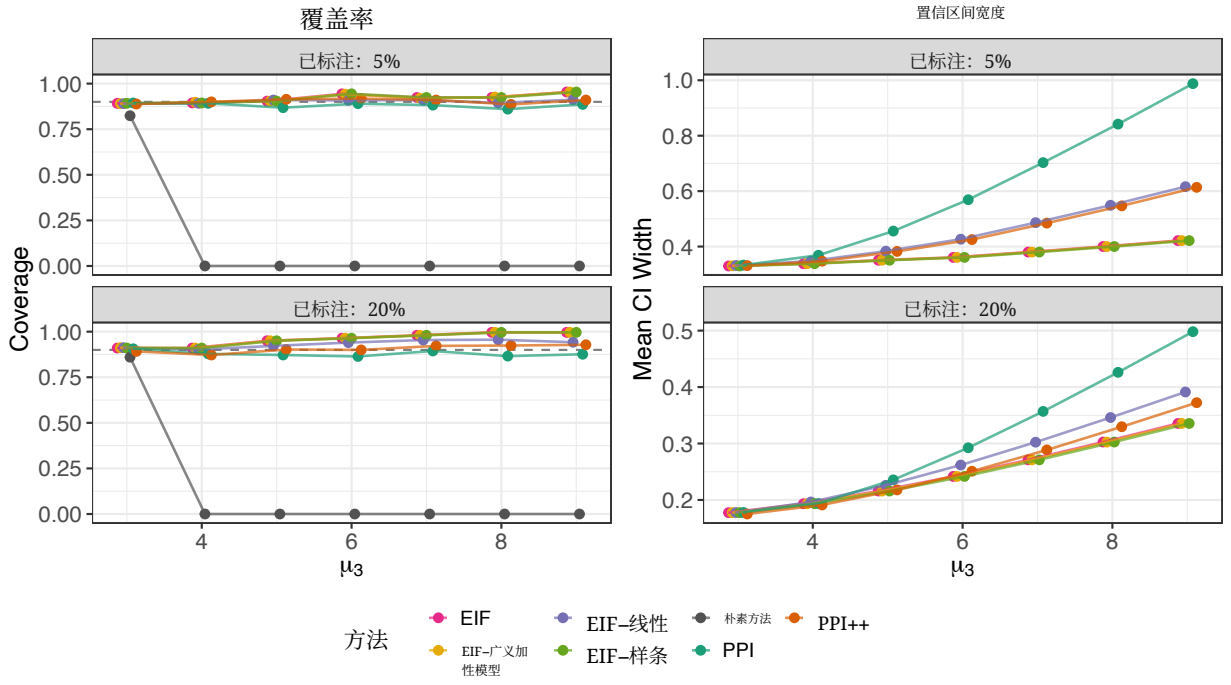
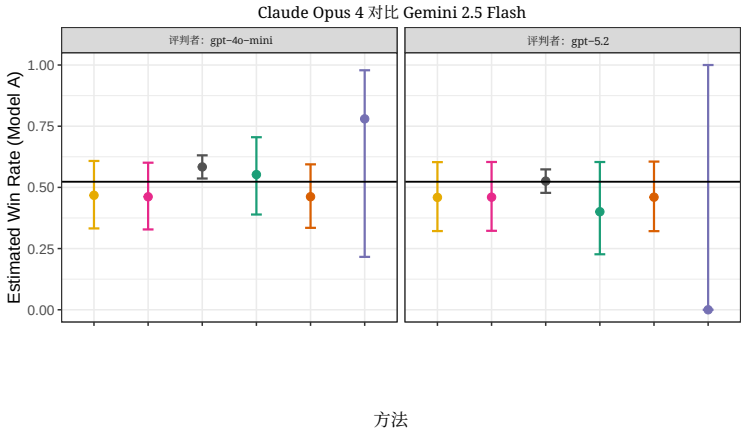
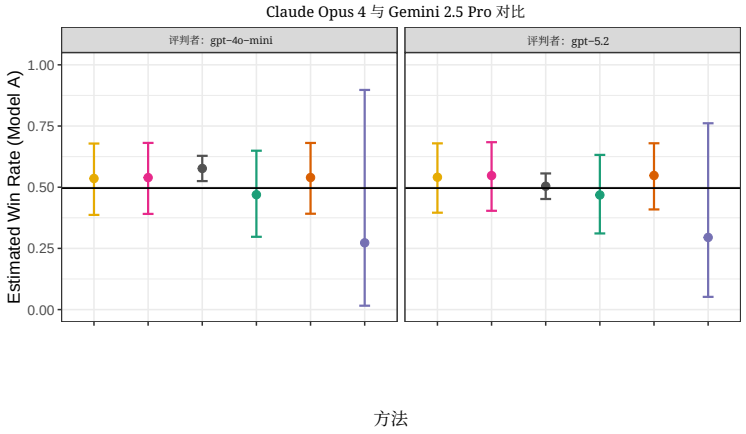


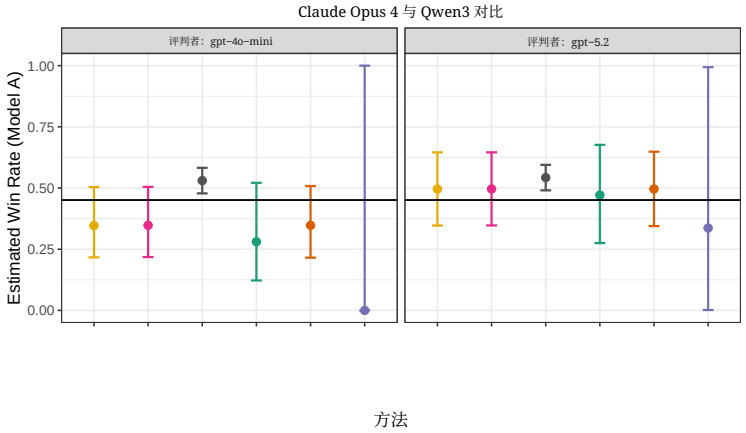
图5：离散预测模拟实验的覆盖率（左）和置信区间宽度（右）。虚线表示名义90%覆盖率。大多数方法在不同偏差幅度下都能实现近似有效的覆盖率（μ_3），而随着偏差增大，朴素区间的覆盖率会严重不足。采用样条和广义加性模型校准的 EIF 得到的区间最窄，体现了其校正离散预测的能力。正如预测的那样，线性校正和 PPI++ 表现相近。



(a) Claude Opus 4 对比 Gemini 2.5 Flash（真实胜率：0.523）



(b) Claude Opus 4 对比 Gemini 2.5 Pro（真实胜率：0.496）



(c) Claude Opus 4 对比 Qwen3-235B（真实胜率：0.451）

图6：使用两个 LLM 评判器（GPT-4o-mini 和 GPT-5.2）对三个模型对进行胜率估计的点估计和90%置信区间。黑色水平线表示真实胜率。不同的估计量和置信区间以不同颜色显示。

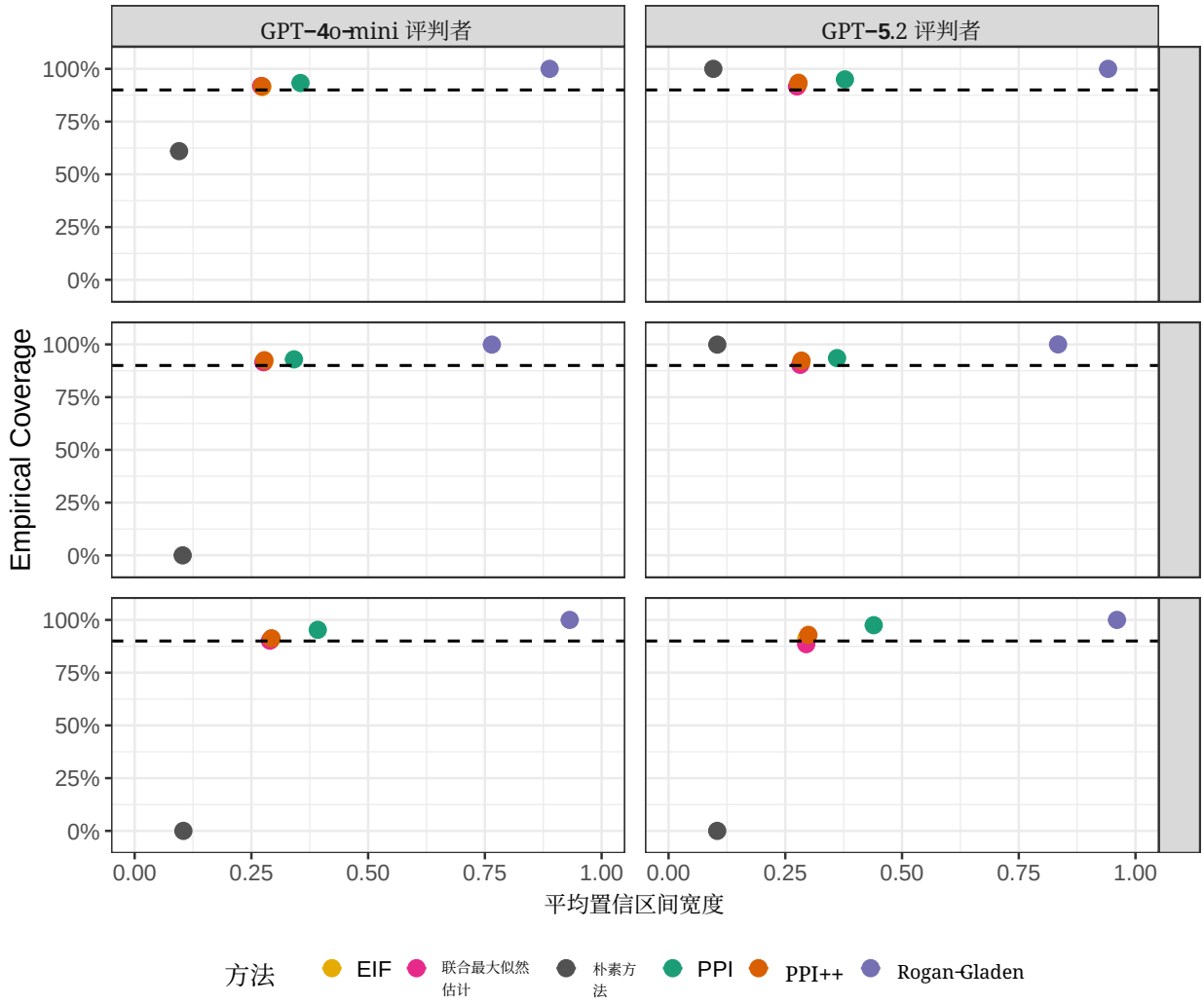


图7：在1,000次随机90/10测试/校准划分中，经验覆盖率与平均置信区间宽度的关系。虚线表示90%的名义覆盖率。误差条表示90%置信区间。当评判者偏差较大时，朴素估计（灰色）的覆盖率为0%；而PPI++、EIF和最大似然估计（MLE）则能维持名义覆盖率，同时保持区间较窄。

参考文献

Anastasios N Angelopoulos、Stephen Bates、Clara Fannjiang、Michael I Jordan 和 Tijana Zrnic。《预测增强推断》。《科学》，382(6671):669–674，2023a。Anastasios N Angelopoulos、John C Duchi 和 Tijana Zrnic。PPI++：高效的预测增强推断。arXiv 预印本 arXiv:2311.01453，2023b。Peter J Bickel、Chris AJ Klaassen、Peter J Bickel、Ya'acov Ritov、J Klaassen、Jon A Wellner 和 YA'Acov Ritov。《半参数模型的高效自适应估计》，第4卷。施普林格，1993年。Lawrence D Brown、T Tony Cai 和 Anirban DasGupta。《二项比例的区间估计》。《统计科学》，16(2):101–133，2001年。Xiaohong Chen、Han Hong 和 Alessandro Tarozzi。辅助数据下广义矩估计模型的半参数效率。《统计年鉴》，36(2)，2008年4月。ISSN 0090-5364。doi: 10.1214/ 009053607000000947。URL http://dx.doi.org/10.1214/009053607000000947。Wei-Lin Chiang、Lianmin Zheng、Ying Sheng、Anastasios Nikolas Angelopoulos、Tianle Li、Dacheng Li、Hao Zhang、Banghua Zhu、Michael Jordan、Joseph E. Gonzalez 和 Ion Stoica。Chatbot Arena：一个根据人类偏好评估大语言模型的开放平台。arXiv 预印本 arXiv:2403.04132，2024年。Naoki Egami、Musashi Hinck、Brandon Stewart 和 Hanying Wei。使用不完美代理变量进行下游推断：面向大语言模型社会科学应用的基于设计的监督学习。《神经信息处理系统进展》，36:68589–68601，2023年。Jacob Fiksel、Abhirup Datta、Agbessi Amouzou 和 Scott Zeger。数据集偏移下的广义贝叶斯量化学习。《美国统计学会会刊》，117(540):2163–2181，2022年。Wayne A Fuller。《测量误差模型》。约翰·威立父子出版公司，2009年。Pablo Gonz´alez、Alberto Casta˜no、Nitesh V Chawla 和 Juan Jos´e Del Coz。量化学习综述。《ACM计算调查》（CSUR），50(5):1–40，2017年。Wenlong Ji、Lihua Lei 和 Tijana Zrnic。预测作为代理变量：重新审视人工智能时代的代理结果。arXiv 预印本 arXiv:2501.09731，2025年。Zsolt Lang 和 Jen˝o Reiczigel。针对估计的灵敏度和特异度进行校正的疾病患病率置信限。《预防兽医学》，113(1):13–22，2014年。Chungpa Lee、Thomas Zeng、Jongwon Jeong、Jy-yong Sohn 和 Kangwook Lee。如何正确报告 LLM 作为裁判评估。arXiv 预印本 arXiv:2511.21140，2025年。Haitao Li、Junjie Chen、Qingyao Ai、Zhumin Chu、Yujia Zhou、Qian Dong 和 Yiqun Liu。Calibrae-val：校准预测分布以减轻大语言模型评审中的选择偏差。载于《第63届计算语言学协会年会论文集（第1卷：长篇论文）》，第16537–16552页，2025年。

R.J.A. Little 和 D.B. Rubin。缺失数据的统计分析。Wiley 概率与数理统计丛书。概率与数理统计。Wiley，2002年。ISBN 9780471183860。URLhttp://books.google.com/books?id=aYPwAAAAMAAJ。Yuxuan Liu、Tianchi Yang、Shaohan Huang、Zihan Zhang、Haizhen Huang、Furu Wei、Weiwei Deng、Feng Sun 和 Qi Zhang。校准基于大语言模型的评估器。载于*2024*年计算语言学、语言资源与评估联合国际会议论文集（*LREC-COLING 2024*），第2638–2656页，2024年。Jiacheng Miao、Xinran Miao、Yixuan Wu、Jiwei Zhao 和 Qiongshi Lu。弱假设且数据自适应的预测后推断。机器学习研究杂志，26(179):1–31，2025年。Alessandra Rister Portinari Maranca、Jihoon Chung、Musashi Hinck、Adam D Wolsky、Naoki Egami 和 Brandon M Stewart。利用基于设计的监督学习校正图像分析中人工智能辅助标注的测量误差。《社会学方法与研究》，第00491241251333372页，2025年。James M Robins 和 Andrea Rotnitzky。缺失数据下多元回归模型的半参数效率。《美国统计学会会刊》，90(429):122–129，1995年。Walter J Rogan 和 Beth Gladen。根据筛查试验结果估计患病率。《美国流行病学杂志》，107(1):71–76，1978年。Samuel P Rosin、Bonnie E Shook-Sa、Stephen R Cole 和 Michael G Hudgens。估计 SARS-CoV-2 血清阳性率。《皇家统计学会杂志 *A*辑：社会统计学》，186 (4):834–851，2023年。Aishwarya Sahoo、Jeevana Kruthi Karnuthala、Tushar Parmanand Budhwani、Pranchal Agarwal、Sankaran Vaidyanathan、Alexa Siu、Franck Dernoncourt、Jennifer Healey、Nedim Lipka、Ryan Rossi 等。定量大语言模型评审器。*arXiv* 预印本 *arXiv:2506.02945*，2025年。Stephen Salerno、Jiacheng Miao、Awan Afiaz、Kentaro Hoffman、Anna Neufeld、Qiongshi Lu、Tyler H McCormick 和 Jeffrey T Leek。ipd：用于对预测数据进行推断的 R 软件包。《生物信息学》，41(2):btaf055，2025年。Kayla Schroeder 和 Zach Wood-Doughty。你能相信大语言模型判断吗？LLM 作为评审的可靠性。*arXiv* 预印本 *arXiv:2412.12509*，2024年。Lin Shi、Chiyu Ma、Wenhua Liang、Xingjian Diao、Weicheng Ma 和 Soroush Vosoughi。评判评判者：对 LLM 作为评审中的位置偏差进行系统研究。*arXiv* 预印本 *arXiv:2406.07791*，2024年。Charles Joel Stone。概率与统计教程，第19卷。Duxbury Press Belmont，1996年。Zailong Tian、Zhuoheng Han、Yanzhe Chen、Haozhe Xu、Xi Yang、Richeng Xuan、Houfeng Wang 和 Lizi Liao。大语言模型评审器中的过度自信：诊断与置信度驱动的解决方案。*arXiv* 预印本 *arXiv:2508.06225*，2025年。

Aad W Van der Vaart。《渐近统计学》，第3卷。剑桥大学出版社，2000年。Siruo Wang、Tyler H McCormick 和 Jeffrey T Leek。基于机器学习预测结果校正推断的方法。《美国国家科学院院刊》，117（48）：30266–30275，2020年。Simon Wood。mgcv：采用广义交叉验证／赤池信息准则／限制最大似然进行平滑度估计的混合广义加性模型计算工具。2012年。Zichun Xu、Daniela Witten 和 Ali Shojaie。半参数高效半监督学习的统一框架。*arXiv* 预印本 *arXiv:2502.17741*，2025年。

统一用于 LLM 作为裁判评估的去偏方法的补充材料

A 附加结果

A.1 校准估计值的有限样本变异性

Rogan–Gladen 估计量依赖于代入估计值 $\hat{}q_0$ 和 $\hat{}q_1$。图 8 展示了它们在不同模拟设置下的有限样本变异性。尽管这些估计值近似无偏，但当标注预算较小时，其方差可能相当大。在1%的标注比例下（$m = 20$），正类可能仅包含 $m_1 \approx 3$ 个观测值，从而使 $\hat{}q_1$的均方根误差达到0.3，而 $\hat{}q_0$的均方根误差为0。

由于 Rogan–Gladen 估计量要除以 $\hat{}q_0 + \hat{}q_1 - 1$，当 m 较小时，或类别不平衡导致 m_1 （或 m_0）的观测值很少时，所得置信区间可能会很宽。

相比之下，PPI 类估计量和基于 EIF 的方法不需要显式估计 $\hat{}(q_0, q_1)$。相反，它们利用无偏校正来学习 $\mathbb{E}[Y \mid Y]$，因此在小规模校准情形下更加稳定。

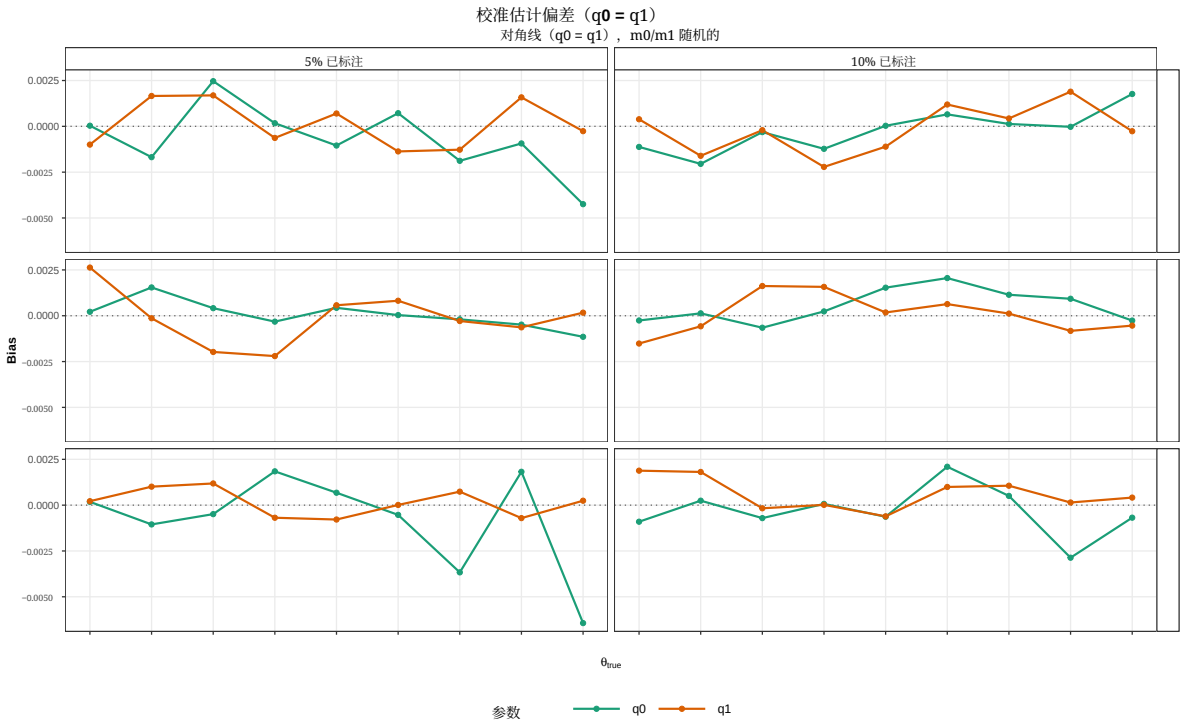


图 8：校准估计值 $\hat{}q_0$ 和 $\hat{}q_1$的有限样本均方根误差。

B 命题证明

B.1 命题1 的证明

ˆ ˆ证明。 由于测试样本由 $Y \in \{0,1\}$ 的独立同分布抽样组成，且具有 $p = \mathbf{Pr}(Y = 1)$，因此对于 $i = 1, \ldots, n$，有ˆ $Y_i \sim \mathbf{Bern}(p)$。对于样本均值ˆ$p = \frac{1}{n}\sum_{i=1}^{n}$ ˆY_i，当 $n \to \infty$ 时，根据中心极限定理，我们有

$$\sqrt{n}\,(\hat{p} - p) \xrightarrow{d} \mathcal{N}\big(0, p(1-p)\big).$$

在校准样本上，定义

$$X_{ab} = \sum_{j=1}^{m} \mathbf{1}\{Y_j = a,\ \hat{Y}_j = b\}, \qquad \pi_{ab} = \Pr(Y = a, \hat{Y} = b), \qquad a, b \in \{0,1\}.$$

则 $X = (X_{00},\ X_{01},\ X_{10},\ X_{11})^\top$ 满足

$$X \sim \text{Multinomial}\big(m;\, \pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}\big),$$

其中

$$\begin{aligned}
\pi_{00} &= \Pr(Y = 0, \hat{Y} = 0) = q_0(1-\theta), \\
\pi_{01} &= \Pr(Y = 0, \hat{Y} = 1) = (1-q_0)(1-\theta), \\
\pi_{10} &= \Pr(Y = 1, \hat{Y} = 0) = (1-q_1)\theta, \\
\pi_{11} &= \Pr(Y = 1, \hat{Y} = 1) = q_1\theta.
\end{aligned}$$

设 $\pi = (\pi_{00},\ \pi_{01},\ \pi_{10},\ \pi_{11})^\top$ 和ˆ$P(R = 1)^{\text{cal}} = X/m$。当 $m \to \infty$ 时，根据多元中心极限定理，

$$\sqrt{m}\,(\hat{P}(R=1)^{\text{cal}} - \pi) \xrightarrow{d} \mathcal{N}\Big(0,\, \text{diag}(\pi) - \pi\pi^\top\Big).$$

由于

$$\hat{q}_1 = \frac{\sum_{j=1}^{m} \mathbf{1}\{\hat{Y}_j = 1,\ Y_j = 1\}}{m_1}, \qquad \hat{q}_0 = \frac{\sum_{j=1}^{m} \mathbf{1}\{\hat{Y}_j = 0,\ Y_j = 0\}}{m_0},$$

由此可得

$$\hat{q}_1 = \frac{X_{11}}{X_{11} + X_{10}} = \frac{\hat{\pi}_{11}^{\text{cal}}}{\hat{P}(R=1)_{11}^{\text{cal}} + \hat{P}(R=1)_{10}^{\text{cal}}}, \qquad \hat{q}_0 = \frac{X_{00}}{X_{00} + X_{01}} = \frac{\hat{P}(R=1)_{00}^{\text{cal}}}{\hat{P}(R=1)_{00}^{\text{cal}} + \hat{P}(R=1)_{01}^{\text{cal}}}.$$

定义 h：$(0,\ 1)^4 \to (0,\ 1)^2$ 为

$$h(x_{00}, x_{01}, x_{10}, x_{11}) = \left(\frac{x_{00}}{x_{00} + x_{01}},\ \frac{x_{11}}{x_{10} + x_{11}}\right)^\top,$$

使得 $h(\pi) = (q_0,\ q_1)^\top$ 且 $h(\hat{P}(R = 1)^{\text{cal}}) = (\hat{q}_0,\ \hat{q}_1)^\top$。由于 $\pi_{00} + \pi_{01} = 1 - \theta > 0$和 $\pi_{11} + \pi_{10} = \theta > 0$，映射 h 在 π处连续可微。其在 π 处的雅可比矩阵为

$$\nabla h(\pi) = \begin{pmatrix} \dfrac{\pi_{01}}{(\pi_{00}+\pi_{01})^2} & -\dfrac{\pi_{00}}{(\pi_{00}+\pi_{01})^2} & 0 & 0 \\ 0 & 0 & -\dfrac{\pi_{11}}{(\pi_{10}+\pi_{11})^2} & \dfrac{\pi_{10}}{(\pi_{10}+\pi_{11})^2} \end{pmatrix}.$$

当 $m \to \infty$ 时，根据多元 delta 方法，

$$\sqrt{m}\begin{pmatrix}\hat{q}_0 - q_0\\ \hat{q}_1 - q_1\end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(0,\, \nabla h(\pi)\big(\mathrm{diag}(\pi) - \pi\pi^\top\big)\nabla h(\pi)^\top\Big).$$

注意到 $\theta = \pi_{11} + \pi_{10}$；因此，直接计算可得

$$\nabla h(\pi)\big(\mathrm{diag}(\pi) - \pi\pi^\top\big)\nabla h(\pi)^\top = \begin{pmatrix}\dfrac{q_0(1-q_0)}{1-\theta} & 0\\ 0 & \dfrac{q_1(1-q_1)}{\theta}\end{pmatrix},$$

and

$$\sqrt{m}\begin{pmatrix}\hat{q}_0 - q_0\\ \hat{q}_1 - q_1\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0,\, \begin{pmatrix}\dfrac{q_0(1-q_0)}{1-\theta} & 0\\ 0 & \dfrac{q_1(1-q_1)}{\theta}\end{pmatrix}\right).$$

由于 $n/m \to \gamma_1 \in (0, \infty)$，我们有 $\sqrt{n} = \sqrt{n/m}\,\sqrt{m}$ 以及 $\sqrt{n/m} \to \sqrt{\gamma_1}$。根据斯卢茨基定理，

$$\sqrt{n}\begin{pmatrix}\hat{q}_0 - q_0\\ \hat{q}_1 - q_1\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0,\, \gamma_1\begin{pmatrix}\dfrac{q_0(1-q_0)}{1-\theta} & 0\\ 0 & \dfrac{q_1(1-q_1)}{\theta}\end{pmatrix}\right).$$

此外，由于测试样本与校准样本相互独立，$\hat{}p$ 与 $(\hat{}q_0, \hat{}q_1)$ 相互独立。因此，当 $n \to \infty$ 时，

$$\sqrt{n}\begin{pmatrix}\hat{p} - p\\ \hat{q}_0 - q_0\\ \hat{q}_1 - q_1\end{pmatrix} \xrightarrow{d} \mathcal{N}\Big(0,\, \mathrm{diag}\Big(p(1-p),\ \gamma_1\frac{q_0(1-q_0)}{1-\theta},\ \gamma_1\frac{q_1(1-q_1)}{\theta}\Big)\Big).$$

定义 $g(p, q_0, q_1) = \frac{p+q_0-1}{q_0+q_1-1}$，使得 $\theta = g(p, q_0, q_1)$。代数运算可得

$$\frac{\partial g}{\partial p} = \frac{1}{q_0+q_1-1}, \qquad \frac{\partial g}{\partial q_0} = \frac{1-\theta}{q_0+q_1-1}, \qquad \frac{\partial g}{\partial q_1} = -\frac{\theta}{q_0+q_1-1},$$

因此

$$\nabla g(p, q_0, q_1) = \frac{1}{q_0+q_1-1}\big(1,\ 1-\theta,\ -\theta\big).$$

根据多元 delta 方法，

$$\sqrt{n}\,(\hat{\theta}_{\mathrm{RG}} - \theta) \xrightarrow{d} \mathcal{N}\Big(0,\, \nabla g(p, q_0, q_1)\ \mathrm{diag}\Big(\pi(1-\pi),\ \gamma_1\frac{q_0(1-q_0)}{1-\theta},\ \gamma_1\frac{q_1(1-q_1)}{\theta}\Big)\ \nabla g(p, q_0, q_1)^\top\Big).$$

通过代数运算并假设 $\frac{N}{n} = \frac{m+n}{n} \to \frac{1+\gamma_1}{\gamma_1}$，我们得到

$$V_{\mathrm{RG}} = \left(\frac{1+\gamma_1}{\gamma_1}\right)\frac{1}{(q_0+q_1-1)^2}\left\{p(1-p) + \gamma_1\big[(1-\theta)q_0(1-q_0) + \theta q_1(1-q_1)\big]\right\}.$$

□

命题 2 的 B.2 证明

证明。 我们首先使用以下分解

$$\hat{\mu} = \hat{p} = \frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i, \qquad \hat{\Delta} = \frac{1}{m}\sum_{j=1}^{m}(\hat{Y}_j - Y_j), \qquad \hat{\theta}_{\text{PPI}} = \hat{\mu} - \hat{\Delta}.$$

步骤 1：的矩ˆμ。 由于测试样本中ˆ Y_i 独立同分布∼伯努利分布(p),

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}[\hat{Y}] = p, \qquad \text{Var}(\hat{\mu}) = \frac{1}{n}\text{Var}(\hat{Y}) = \frac{p(1-p)}{n}.$$

ˆˆ**步骤 2：的矩** Δ。 在校准样本中，被加数 $Y_j - Y_j$ 独立同分布，因此

$$\mathbb{E}[\hat{\Delta}] = \mathbb{E}[\hat{Y} - Y], \qquad \text{Var}(\hat{\Delta}) = \frac{1}{m}\text{Var}(\hat{Y} - Y).$$

ˆˆ因为 Y， $Y \in \{0, 1\}$，我们有 $Y - Y \in \{-1, 0, 1\}$ 且

$$(\hat{Y} - Y)^2 = \mathbf{1}\{\hat{Y} \neq Y\}.$$

因此,

$$\mathbb{E}[(\hat{Y} - Y)^2] = \Pr(\hat{Y} \neq Y) = \Pr(\hat{Y} = 1, Y = 0) + \Pr(\hat{Y} = 0, Y = 1). \tag{21}$$

根据 q_0， q_1 的定义,

$$\Pr(\hat{Y} = 1, Y = 0) = (1 - q_0)(1 - \theta), \qquad \Pr(\hat{Y} = 0, Y = 1) = (1 - q_1)\theta,$$

因此（21）变为

$$\mathbb{E}[(\hat{Y} - Y)^2] = (1 - \theta)(1 - q_0) + \theta(1 - q_1). \tag{22}$$

ˆ接下来，由于 $\mathbb{E}[Y] = p$ 和 $\mathbb{E}[Y] = \theta$,

$$\mathbb{E}[\hat{Y} - Y] = \mathbb{E}[\hat{Y}] - \mathbb{E}[Y] = p - \theta,$$

因此

$$\text{Var}(\hat{Y} - Y) = \mathbb{E}[(\hat{Y} - Y)^2] - \{\mathbb{E}[\hat{Y} - Y]\}^2 = (1 - \theta)(1 - q_0) + \theta(1 - q_1) - (p - \theta)^2. \tag{23}$$

结合 Var(ˆΔ) $= m^{-1}$Var(ˆ$Y - Y$), 得到

$$\text{Var}(\hat{\Delta}) = \frac{(1 - \theta)(1 - q_0) + \theta(1 - q_1) - (\theta - p)^2}{m}.$$

ˆ**步骤 3： 的无偏性**。 根据期望的线性性,

$$\mathbb{E}[\hat{\theta}_{\text{PPI}}] = \mathbb{E}[\hat{\mu}] - \mathbb{E}[\hat{\Delta}] = p - (p - \theta) = \theta.$$

ˆ**步骤4：的有限样本方差** θ_{PPI}。 由于测试样本和校准样本相互独立，ˆμ 和 Δ 也相互独立。因此,

$$\text{Var}(\hat{\theta}_{\text{PPI}}) = \text{Var}(\hat{\mu} - \hat{\Delta}) = \text{Var}(\hat{\mu}) + \text{Var}(\hat{\Delta}) = \frac{p(1-p)}{n} + \frac{(1 - \theta)(1 - q_0) + \theta(1 - q_1) - (\theta - p)^2}{m}.$$

步骤5：渐近正态性。根据中心极限定理，

$$\sqrt{n}\,(\hat{\mu}-p)\xrightarrow{d}\mathcal{N}\big(0,p(1-p)\big),\qquad \sqrt{m}\,(\hat{\Delta}-(p-\theta))\xrightarrow{d}\mathcal{N}\big(0,\mathrm{Var}(\hat{Y}-Y)\big).$$

若 $n,\ m\to\infty$ 且 $n/m\to\gamma_1\in(0,\ \infty)$，则

$$\sqrt{n}\,(\hat{\Delta}-(p-\theta))=\sqrt{\frac{n}{m}}\,\sqrt{m}\,(\hat{\Delta}-(p-\theta))\xrightarrow{d}\mathcal{N}\big(0,\gamma_1\,\mathrm{Var}(\hat{Y}-Y)\big),$$

根据斯卢茨基定理。两个样本的独立性意味着它们联合收敛至相互独立的极限，因此

$$\sqrt{n}\,(\hat{\theta}_{\mathrm{PPI}}-\theta)=\sqrt{n}\,(\hat{\mu}-p)\ -\ \sqrt{n}\,(\hat{\Delta}-(p-\theta))\xrightarrow{d}\mathcal{N}\Big(0,\,p(1-p)+\gamma_1\,\mathrm{Var}(\hat{Y}-Y)\Big).$$

通过代数运算，并根据假设 $\frac{N}{n}=\frac{m+n}{n}\Rightarrow\frac{1+\gamma_1}{\gamma_1}$，我们得到

$$V_{\mathrm{PPI}}=\frac{1+\gamma_1}{\gamma_1}\Big\{p(1-p)+\gamma_1\Big[(1-\theta)(1-q_0)+\theta(1-q_1)-(\theta-p)^2\Big]\Big\}.$$

□

B.3 命题 3 的证明

证明。我们可以将具有固定调节参数 $\lambda\in\mathbb{R}$ 的 PPI++ 估计量重写为：

$$\hat{\theta}_{\texttt{PPI++}}(\lambda)=\frac{1}{m}\sum_{j=1}^{m}Y_j\ +\ \lambda\left(\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i-\frac{1}{m}\sum_{j=1}^{m}\hat{Y}_j\right)=\lambda\,\hat{p}+\Big(\hat{\theta}_{\mathrm{class}}-\lambda\,\bar{\hat{Y}}_{\mathrm{cal}}\Big),$$

其中

$$\hat{\theta}_{\mathrm{class}}=\frac{1}{m}\sum_{j=1}^{m}Y_j,\qquad \bar{\hat{Y}}_{\mathrm{cal}}=\frac{1}{m}\sum_{j=1}^{m}\hat{Y}_j,\qquad \hat{p}=\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i.$$

无偏性。在联合模型 (1) 下，$\mathbb{E}[\hat{\theta}_{\mathrm{class}}]=\mathbb{E}[Y]=\theta$ 且 $\mathbb{E}[\hat{p}]=\mathbb{E}[Y_{\mathrm{cal}}]=\mathbb{E}[Y]=p$。因此，

$$\mathbb{E}\Big[\hat{\theta}_{\texttt{PPI++}}(\lambda)\Big]=\mathbb{E}[\hat{\theta}_{\mathrm{class}}]+\lambda\Big(\mathbb{E}[\hat{p}]-\mathbb{E}[\bar{\hat{Y}}_{\mathrm{cal}}]\Big)=\theta+\lambda(p-p)=\theta.$$

有限样本方差。由于测试样本和校准样本相互独立，$\hat{p}$ 与 $(\hat{\theta}_{\mathrm{class}},\bar{\hat{Y}}_{\mathrm{cal}})$ 独立。因此，我们有

$$\mathrm{Var}\Big(\hat{\theta}_{\texttt{PPI++}}(\lambda)\Big)=\mathrm{Var}(\lambda\hat{p})+\mathrm{Var}\Big(\hat{\theta}_{\mathrm{class}}-\lambda\bar{\hat{Y}}_{\mathrm{cal}}\Big).$$

此外，

$$\mathrm{Var}(\hat{p})=\frac{1}{n}\mathrm{Var}(\hat{Y})=\frac{p(1-p)}{n},\ \mathrm{Var}(\hat{\theta}_{\mathrm{class}})=\frac{1}{m}\mathrm{Var}(Y)=\frac{\theta(1-\theta)}{m},\ \mathrm{Var}(\bar{\hat{Y}}_{\mathrm{cal}})=\frac{1}{m}\mathrm{Var}(\hat{Y})=\frac{p(1-p)}{m}.$$

对于协方差项，我们将其简化为

$$\mathrm{Cov}(\hat{\theta}_{\mathrm{class}}, \bar{\hat{Y}}_{\mathrm{cal}}) = \mathrm{Cov}\left(\frac{1}{m}\sum_{j=1}^{m} Y_j,\ \frac{1}{m}\sum_{j=1}^{m}\hat{Y}_j\right) = \frac{1}{m}\mathrm{Cov}(Y, \hat{Y}),$$

由于跨样本项因 j之间的独立性而消失。此外，

$$\mathrm{Cov}(Y, \hat{Y}) = \mathbb{E}[Y\hat{Y}] - \mathbb{E}[Y]\mathbb{E}[\hat{Y}] = \Pr(Y=1, \hat{Y}=1) - \theta p = \theta q_1 - \theta p = \theta(q_1 - p).$$

将这些结果合并可得

$$\mathrm{Var}\Big(\hat{\theta}_{\texttt{PPI++}}(\lambda)\Big) = \frac{\theta(1-\theta)}{m} + \lambda^2\left(\frac{p(1-p)}{n} + \frac{p(1-p)}{m}\right) - \frac{2\lambda}{m}\,\theta(q_1 - p).$$

渐近正态性。根据中心极限定理，

$$\sqrt{n}\,(\hat{p} - p) \xrightarrow{d} \mathcal{N}\big(0, p(1-p)\big).$$

对于校准部分，定义独立同分布变量

$$W_j = Y_j - \lambda\hat{Y}_j, \qquad \mathbb{E}[W_j] = \theta - \lambda p, \qquad \mathrm{Var}(W_j) = \mathrm{Var}(Y - \lambda\hat{Y}) = \theta(1-\theta) + \lambda^2 p(1-p) - 2\lambda\,\theta(q_1 - p).$$

Then

$$\hat{\theta}_{\mathrm{class}} - \lambda\bar{\hat{Y}}_{\mathrm{cal}} = \frac{1}{m}\sum_{j=1}^{m} W_j,$$

因此，再次根据中心极限定理，

$$\sqrt{m}\left\{\big(\hat{\theta}_{\mathrm{class}} - \lambda\bar{\hat{Y}}_{\mathrm{cal}}\big) - (\theta - \lambda p)\right\} \xrightarrow{d} \mathcal{N}\Big(0,\ \mathrm{Var}(Y - \lambda\hat{Y})\Big).$$

如果 $n,\ m \to \infty$ 且 $n/m \to \gamma_1 \in (0,\ \infty)$，那么 $\sqrt{n} = \sqrt{n/m}\sqrt{m}$ 且 $\sqrt{n/m} \to \sqrt{\gamma_1}$，因此根据斯卢茨基定理，

$$\sqrt{n}\left\{\big(\hat{\theta}_{\mathrm{class}} - \lambda\bar{\hat{Y}}_{\mathrm{cal}}\big) - (\theta - \lambda p)\right\} \xrightarrow{d} \mathcal{N}\Big(0,\ \gamma_1\,\mathrm{Var}(Y - \lambda\hat{Y})\Big).$$

最后，利用测试样本和校准样本之间的独立性，

$$\begin{aligned}\sqrt{n}\big(\hat{\theta}_{\texttt{PPI++}}(\lambda) - \theta\big) &= \lambda\sqrt{n}(\hat{p} - p) + \sqrt{n}\left\{\big(\hat{\theta}_{\mathrm{class}} - \lambda\bar{\hat{Y}}_{\mathrm{cal}}\big) - (\theta - \lambda p)\right\} \\ &\xrightarrow{d} \mathcal{N}\Big(0,\ \lambda^2 p(1-p) + \gamma_1\,\mathrm{Var}(Y - \lambda\hat{Y})\Big),\end{aligned}$$

其中

$$\begin{aligned}\mathrm{Var}(Y - \lambda\hat{Y}) &= \mathrm{Var}(Y) + \lambda^2\mathrm{Var}(\hat{Y}) - 2\lambda\,\mathrm{Cov}(Y, \hat{Y}) \\ &= \theta(1-\theta) + \lambda^2 p(1-p) - 2\lambda\,\theta(q_1 - p).\end{aligned}$$

将其代入渐近方差和 $\frac{N}{n}$ 缩放后，我们得到

$$\begin{aligned}V_{\texttt{PPI++}}(\lambda) &= \frac{1+\gamma_1}{\gamma_1}\left\{\lambda^2 p(1-p) + \gamma_1\Big[\theta(1-\theta) + \lambda^2 p(1-p) - 2\lambda\,\theta(q_1 - p)\Big]\right\} \\ &= \frac{1+\gamma_1}{\gamma_1}\left\{\gamma_1\,\theta(1-\theta) + \lambda^2(1+\gamma_1)\,p(1-p) - 2\lambda\,\gamma_1\,\theta(q_1 - p)\right\}.\end{aligned}$$

□

B.4 命题 4 的证明

证明。将组合对数似然写作

$$\ell_{n,m}(\eta) = \sum_{i=1}^{n} \ell_{\text{test}}(\hat{Y}_i;\eta) + \sum_{j=1}^{m} \ell_{\text{cal}}(Y_j,\hat{Y}_j;\eta), \qquad \eta = (\theta, q_0, q_1)^\top,$$

其中，测试样本仅贡献 Y，而校准样本贡献 $(Y,\ Y)$。

步骤 1：评分和费舍尔信息。对于测试样本观测值 $Y \in \{0,1\}$，模型表示为

$$P(\hat{Y}=1) = p(\eta) = (1-\theta)(1-q_0) + \theta q_1,$$

因此

$$\ell_{\text{test}}(\hat{Y};\eta) = \hat{Y}\log p + (1-\hat{Y})\log(1-p).$$

Let

$$g(\eta) = \nabla_\eta p(\eta) = \big(q_0 + q_1 - 1,\ -(1-\theta),\ \theta\big)^\top.$$

对 η 求导可得得分

$$\dot{\ell}_{\text{test}}(\hat{Y};\eta) = \frac{\hat{Y}-p}{p(1-p)}\, g(\eta).$$

由于 $\mathrm{Var}(Y) = p(1-p)$，因此测试样本提供的（每次观测的）费舍尔信息为

$$I_{\text{test}}(\eta) = \mathbb{E}\Big[\dot{\ell}_{\text{test}}(\hat{Y};\eta)\dot{\ell}_{\text{test}}(\hat{Y};\eta)^\top\Big] = \frac{1}{p(1-p)}\, g(\eta)g(\eta)^\top,$$

这与给定的矩阵一致。

对于一个 c 校准样本观测值 $(Y,\ Y)$，完整数据似然可分解为

$$P(Y=y,\hat{Y}=\hat{y}) = P(Y=y)\,P(\hat{Y}=\hat{y}\mid Y=y),$$

其中 $P(Y=1) = \theta$, $P(Y=1\mid Y=0) = 1-q_0$，且 $P(Y=1\mid Y=1) = q_1$。因此

$$\begin{aligned}\ell_{\text{cal}}(Y,\hat{Y};\eta) &= Y\log\theta + (1-Y)\log(1-\theta)\\ &\quad + (1-Y)\Big[\hat{Y}\log(1-q_0) + (1-\hat{Y})\log q_0\Big] + Y\Big[\hat{Y}\log q_1 + (1-\hat{Y})\log(1-q_1)\Big].\end{aligned}$$

得分分量为

$$\dot{\ell}_\theta = \frac{Y-\theta}{\theta(1-\theta)}, \qquad \dot{\ell}_{q_0} = (1-Y)\Big(\frac{1-\hat{Y}}{q_0} - \frac{\hat{Y}}{1-q_0}\Big), \qquad \dot{\ell}_{q_1} = Y\Big(\frac{\hat{Y}}{q_1} - \frac{1-\hat{Y}}{1-q_1}\Big).$$

我们首先计算信息矩阵中的非对角元素。

交叉矩 $\mathbb{E}[\ell_\theta\ell_{q_0}]$： ℓ_{q_0} 当 $Y=1$ 时等于 0。利用迭代期望，我们有

$$\mathbb{E}[\dot{\ell}_\theta\dot{\ell}_{q_0}] = \mathbb{E}[\mathbb{E}[\dot{\ell}_\theta\dot{\ell}_{q_0}\mid Y]] = (1-\theta)\mathbb{E}[\dot{\ell}_\theta\dot{\ell}_{q_0}\mid Y=0].$$

给定 $Y=0$、$\dot{\ell}_\theta = \frac{-\theta}{1-\theta}(\,) = \frac{-1}{1-\theta}$，因此：

$$\mathbb{E}[\dot{\ell}_\theta\dot{\ell}_{q_0}\mid Y=0] = \frac{-1}{1-\theta}\mathbb{E}[\dot{\ell}_{q_0}\mid Y=0] = \frac{-1}{1-\theta}\times 0 = 0.$$

交叉矩 $\mathbb{E}[\ell_\theta \ell_{q_1}]$**：** 首先注意到，$\ell_{q_1} = 0$ 当 $Y = 0$ 时，因此我们有

$$\mathbb{E}[\dot{\ell}_\theta \dot{\ell}_{q_1}] = \theta \mathbb{E}[\dot{\ell}_\theta \dot{\ell}_{q_1} \mid Y = 1].$$

给定 $Y = 1$: $\dot{\ell}_\theta = \frac{1}{\theta}$，我们得出

$$\mathbb{E}[\dot{\ell}_\theta \dot{\ell}_{q_1} \mid Y = 1] = \frac{1}{\theta} \mathbb{E}[\dot{\ell}_{q_1} \mid Y = 1] = \frac{1}{\theta} \times 0 = 0.$$

交叉矩 $\mathbb{E}[\ell_{q_0} \ell_{q_1}]$**：** $\ell_{q_0} \ell_{q_1} = (1-Y) \cdot Y \cdot (\cdots)$；由于 $Y \in 0,\ 1$ 总是蕴含 $(1-Y) \cdot Y = 0$。因此，校准费歇尔信息矩阵为对角矩阵，其对角元素为

$$I_{\text{cal}}(\eta) = \operatorname{diag}\Big(\mathbb{E}[\dot{\ell}_\theta^2],\ \mathbb{E}[\dot{\ell}_{q_0}^2],\ \mathbb{E}[\dot{\ell}_{q_1}^2]\Big) = \operatorname{diag}\left(\frac{1}{\theta(1-\theta)},\ \frac{1-\theta}{q_0(1-q_0)},\ \frac{\theta}{q_1(1-q_1)}\right).$$

步骤 2：联合最大似然估计的渐近正态性。令 $N = n + m$ 并记完整得分函数为 as

$$\dot{\ell}_{n,m}(\eta) = \sum_{i=1}^{n} \dot{\ell}_{\text{test}}(\hat{Y}_i; \eta) + \sum_{j=1}^{m} \dot{\ell}_{\text{cal}}(Y_j, \hat{Y}_j; \eta).$$

根据构造，$\mathbb{E}[\ell_{n,m}(\eta)] = 0$，并且根据独立性，

$$\operatorname{Var}\Big(\frac{1}{\sqrt{N}} \dot{\ell}_{n,m}(\eta)\Big) = \frac{n}{N} I_{\text{test}}(\eta) + \frac{m}{N} I_{\text{cal}}(\eta) \ \to\ \frac{\gamma_1}{1+\gamma_1} I_{\text{test}}(\eta) + \frac{1}{1+\gamma_1} I_{\text{cal}}(\eta) =: I_{\gamma_1}(\eta),$$

因为 $n/m \to \gamma_1 \in (0,\ \infty)$ 蕴含 $n/N \to \gamma_1/(1+\gamma_1)$ 和 $m/N \to 1/(1+\gamma_1)$。因此，多元中心极限定理给出

$$\frac{1}{\sqrt{N}} \dot{\ell}_{n,m}(\eta) \ \xrightarrow{d}\ \mathcal{N}\big(0,\ I_{\gamma_1}(\eta)\big).$$

在标准正则条件下，我们有

$$\sqrt{N}\,(\hat{\eta}_{\text{MLE}} - \eta) = \Big(-\frac{1}{N} \ddot{\ell}_{n,m}(\tilde{\eta})\Big)^{-1} \Big(\frac{1}{\sqrt{N}} \dot{\ell}_{n,m}(\eta)\Big) \ \xrightarrow{d}\ \mathcal{N}\big(0,\ I_{\gamma_1}(\eta)^{-1}\big)\,.$$

步骤 3：求 $(1,1)$ **元素的闭式**。令 $w = \gamma_1/(1+\gamma_1)$，并注意到

$$I_{\gamma_1}(\eta) = w\, I_{\text{test}}(\eta) + \frac{1}{1+\gamma_1} I_{\text{cal}}(\eta) = D + uu^\top,$$

其中

$$D = \frac{1}{1+\gamma_1} I_{\text{cal}}(\eta) \quad \text{(diagonal)}, \qquad u = \sqrt{\frac{w}{p(1-p)}}\, g(\eta).$$

由于 D 可逆（参数位于内部），由 Sherman–Morrison公式可得

$$(D + uu^\top)^{-1} = D^{-1} - \frac{D^{-1} uu^\top D^{-1}}{1 + u^\top D^{-1} u}.$$

在此，

$$D^{-1} = (1+\gamma_1) \operatorname{diag}\left(\theta(1-\theta),\ \frac{q_0(1-q_0)}{1-\theta},\ \frac{q_1(1-q_1)}{\theta}\right),$$

直接计算可得

$$u^\top D^{-1}u = \frac{\gamma_1}{p(1-p)}\Big[(q_0+q_1-1)^2\theta(1-\theta) + (1-\theta)q_0(1-q_0) + \theta q_1(1-q_1)\Big].$$

此外，第一个分量满足

$$\big(D^{-1}u\big)_1^2 = \frac{\gamma_1(1+\gamma_1)\,\theta^2(1-\theta)^2\,(q_0+q_1-1)^2}{p(1-p)}.$$

代入 Sherman–Morrison表达式并化简可得

$$\big[I_{\gamma_1}(\eta)^{-1}\big]_{11} = (1+\gamma_1)\,\theta(1-\theta)\,\frac{p(1-p)+\gamma_1\Big[(1-\theta)q_0(1-q_0)+\theta q_1(1-q_1)\Big]}{p(1-p)+\gamma_1\Big[(q_0+q_1-1)^2\theta(1-\theta)+(1-\theta)q_0(1-q_0)+\theta q_1(1-q_1)\Big]},$$

ˆ 这就是所述的闭式。取多元极限的第一个坐标，即可得到 θ_{MLE} 的边际渐近正态性。

□

B.5 命题 5 的证明

ˆ证明。 回顾一下，观测数据为 $O=(R,\ RY,\ Y)$，其中 $R\in\{0,1\}$ 表示是否观测到 Yˆ，且标记机制满足 $P(R=1\mid Y,\ Y)=P(R=1)\in(0,\ 1)$。

定义

$$\mu(\hat{Y}) = \mathbb{E}[Y\mid\hat{Y}],\qquad \theta = \mathbb{E}[Y] = \mathbb{E}\big[\mu(\hat{Y})\big].$$

我们通过在观测数据模型中识别 θ 的典范梯度，推导有效影响函数。

步骤 1：切空间表示。 令 $\{P_\varepsilon:\ \varepsilon\in(-\delta,\delta)\}$ 为一个经过真实分布 P_0 的任意正则参数ˆ子模型（其中 $P_{\varepsilon=0}=P_0$）。在 $P(R=1\mid Y,Y)=\pi$下，观测数据密度可分解为

$$p_\varepsilon(o) = p_{\varepsilon,R}(r)\,p_{\varepsilon,\hat{Y}}(\hat{y})\,\big\{p_{\varepsilon,Y\mid\hat{Y}}(y\mid\hat{y})\big\}^r,$$

因此，观测数据得分可分解为

$$s_\varepsilon(O) = s_R(R) + s_{\hat{Y}}(\hat{Y}) + R\,s_{Y\mid\hat{Y}}(Y,\hat{Y}), \tag{24}$$

其中各分量满足通常的零均值约束

$$\mathbb{E}[s_R(R)] = 0,\qquad \mathbb{E}\Big[s_{\hat{Y}}(\hat{Y})\Big] = 0,\qquad \mathbb{E}\Big[s_{Y\mid\hat{Y}}(Y,\hat{Y})\mid\hat{Y}\Big] = 0.$$

因此，观测数据切空间（的闭包）由所有形如（24）的函数组成。

ˆ**步骤 2：的路径导数** θ。 沿着该子模型，$\theta(\varepsilon)=\mathbb{E}_\varepsilon[Y]=\mathbb{E}_\varepsilon[\mu_\varepsilon(Y)]$,，其中 ˆ$\mu_\varepsilon$(ˆ$y$) $=\mathbb{E}_\varepsilon[Y\mid Y=$ ˆ$y]$。在 $\varepsilon=0$ 处求导并运用标准得分演算可得

$$\frac{d}{d\varepsilon}\theta(\varepsilon)\Big|_{\varepsilon=0} = \mathbb{E}\Big[\{\mu(\hat{Y})-\theta\}\,s_{\hat{Y}}(\hat{Y})\Big] + \mathbb{E}\Big[(Y-\mu(\hat{Y}))\,s_{Y\mid\hat{Y}}(Y,\hat{Y})\Big]. \tag{25}$$

ˆ 由于 θ 仅取决于 (Y,Y) 的边际分布，因此 $s_R(R)$ 不产生贡献。

步骤 3：候选影响函数与验证。考虑

$$\phi^\star(O) = \mu(\hat{Y}) - \theta + \frac{R}{\pi}\{Y - \mu(\hat{Y})\}. \tag{26}$$

ˆ 首先， $\mathbb{E}[\phi^\star(O)] = 0$ ，因为 $\mathbb{E}[\mu(Y)] = \theta$ ，并且

$$\mathbb{E}\left[\frac{R}{\pi}\{Y - \mu(\hat{Y})\} \,\middle|\, \hat{Y}\right] = \frac{\mathbb{E}[R \mid \hat{Y}]}{\pi}\,\mathbb{E}[Y - \mu(\hat{Y}) \mid \hat{Y}] = 0.$$

接下来，我们证明 $\phi^\star$ 表示路径导数：对于任意形式为 ($s(O)$ 的得分 24,

$$\mathbb{E}[\phi^\star(O)\, s(O)] = \mathbb{E}[\phi^\star(O)\, s_R(R)] + \mathbb{E}\Big[\phi^\star(O)\, s_{\hat{Y}}(\hat{Y})\Big] + \mathbb{E}\Big[\phi^\star(O)\, R s_{Y|\hat{Y}}(Y, \hat{Y})\Big].$$

第一项为零，因为 $\mathbb{E}[\phi^\star(O) \mid R] = 0$ （由与上文相同的计算可知），且 $\mathbb{E}[s_R(R)] = 0$。对于第二项，残差部分消失:

$$\mathbb{E}\left[\frac{R}{\pi}\{Y - \mu(\hat{Y})\}\, s_{\hat{Y}}(\hat{Y})\right] = \mathbb{E}\left[s_{\hat{Y}}(\hat{Y})\,\mathbb{E}\left[\frac{R}{\pi}\{Y - \mu(\hat{Y})\} \,\middle|\, \hat{Y}\right]\right] = 0,$$

so

$$\mathbb{E}\Big[\phi^\star(O)\, s_{\hat{Y}}(\hat{Y})\Big] = \mathbb{E}\Big[\{\mu(\hat{Y}) - \theta\}\, s_{\hat{Y}}(\hat{Y})\Big].$$

对于第三项 ˆˆˆˆ，注意到 $\mathbb{E}[s_{Y|\hat{}Y}(Y,\ Y) \mid Y] = 0$蕴含 $\mathbb{E}[R\, s_{Y|\hat{}Y}(Y,\ Y) \mid Y] = \pi \cdot 0 = 0$，因此

$$\mathbb{E}\Big[\{\mu(\hat{Y}) - \theta\}\, R s_{Y|\hat{Y}}(Y, \hat{Y})\Big] = 0,$$

因此

$$\mathbb{E}\Big[\phi^\star(O)\, R s_{Y|\hat{Y}}(Y, \hat{Y})\Big] = \mathbb{E}\left[\frac{R}{\pi}\{Y - \mu(\hat{Y})\}\, R s_{Y|\hat{Y}}(Y, \hat{Y})\right] = \mathbb{E}\Big[(Y - \mu(\hat{Y}))\, s_{Y|\hat{Y}}(Y, \hat{Y})\Big],$$

使用 $R^2 = R$和 $\mathbb{E}[Rf(Y,\ \hat{}Y)] = \pi\mathbb{E}[f(Y,\ \hat{}Y)]$在 $P(R = 1 \mid Y,\ \hat{}Y) = \pi$下。

组合各部分可得

$$\mathbb{E}[\phi^\star(O)\, s(O)] = \mathbb{E}\Big[\{\mu(\hat{Y}) - \theta\}\, s_{\hat{Y}}(\hat{Y})\Big] + \mathbb{E}\Big[(Y - \mu(\hat{Y}))\, s_{Y|\hat{Y}}(Y, \hat{Y})\Big],$$

这与（25）中的路径导数相符。因此， $\phi^\star$ 是 θ 在观测数据模型中的典范梯度，因此也是有效影响函数。这正是（14）。 □

B.6 命题6的证明

证明。**步骤 1：估计量的推导。**根据命题5， θ 的一个有效影响函数为

$$\phi^\star(R, Y, \hat{Y}) = \mu(\hat{Y}) - \theta + \frac{R}{P(R=1)}\big\{Y - \mu(\hat{Y})\big\}.$$

标准 EIF（单步）估计量可通过求解经验估计方程 ˆ 得到，其中 μ 替换为 $\hat{}\mu$， $P(R = 1)$ 替换为 $P(R = 1)$:

$$0 = \frac{1}{N}\sum_{i=1}^{N} \phi^\star\big(R_i, Y_i, \hat{Y}_i; \theta, \hat{\mu}, \hat{P}(R=1)\big).$$

代入 $\phi^\star$ 并展开可得

$$0 = \frac{1}{n+m}\sum_{i=1}^{n+m}\left[\hat{\mu}(\hat{Y}_i) - \theta + \frac{R_i}{\hat{P}(R=1)}\{Y_i - \hat{\mu}(\hat{Y}_i)\}\right].$$

整理后可得 θ：

$$\theta = \frac{1}{N}\sum_{i=1}^{N}\hat{\mu}(\hat{Y}_i) + \frac{1}{N}\sum_{i=1}^{N}\frac{R_i}{\hat{P}(R=1)}\{Y_i - \hat{\mu}(\hat{Y}_i)\}.$$

最后，在我们的设定下 $P(R=1) = m/(n+m)$，因此将其代入作为估计值，可得（15）。

步骤 2：渐近正态性。根据单步估计量的标准 M估计理论 [Bickel 等人， 1993]，

$$\sqrt{N}(\hat{\theta}_{\mathrm{EIF}} - \theta) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\phi^\star(R_i, Y_i, \hat{Y}_i) + o_p(1).$$

由中心极限定理可得 $\sqrt{N}(\hat{\theta}_{\mathrm{EIF}} - \theta) \xrightarrow{d} \mathcal{N}(0,\ V_{\mathrm{EIF}})$，其中 $V_{\mathrm{EIF}} = \mathrm{Var}(\phi^\star)$。

步骤 3：方差计算。记 $A = \mu(\hat{Y}) - \theta$，$B = \frac{R}{P(R=1)}\{Y - \mu(\hat{Y})\}$。由于 $\mathbb{E}[Y - \mu(Y) \mid Y] = 0$，我们有 $\mathrm{Cov}(A,\ B) = 0$，因此

$$V_{\mathrm{EIF}} = \mathrm{Var}(A) + \mathrm{Var}(B).$$

对于第一项：

$$\mathrm{Var}(A) = \mathrm{Var}(\mu(\hat{Y})) = p(1-p)(\mu_1 - \mu_0)^2.$$

对于第二项，利用 $R^2 = R$ 以及 R 与 $(Y,\ \hat{Y})$ 的独立性：

$$\mathrm{Var}(B) = \frac{1}{P(R=1)^2}\mathbb{E}[R(Y - \mu(\hat{Y}))^2] = \frac{1}{P(R=1)}\mathbb{E}[\mathrm{Var}(Y \mid \hat{Y})].$$

由于 $P(R=1) = m/N = 1/(1+\gamma_1)$，其中 $\gamma_1 = n/m$，我们有

$$\mathrm{Var}(B) = (1+\gamma_1)\big[(1-p)\mu_0(1-\mu_0) + p\,\mu_1(1-\mu_1)\big].$$

步骤4：二元情形的简化。代入 $\mu_1 = q_1\theta/p$、$\mu_0 = (1-q_1)\theta/(1-p)$ 和 $p = q_1\theta + (1-q_0)(1-\theta)$，经过代数运算可得：

$$\mu_1 - \mu_0 = \frac{\theta(1-\theta)(q_0 + q_1 - 1)}{p(1-p)},$$
$$(1-p)\mu_0(1-\mu_0) + p\,\mu_1(1-\mu_1) = \frac{\theta(1-\theta)[q_0(1-q_0)(1-\theta) + q_1(1-q_1)\theta]}{p(1-p)}.$$

合并后得到（17）。 □

V_{MLE} 与 V_{EIF}的渐近等价性

命题 8。 在模型 （1）且满足 $\gamma_1 = n/m$的条件下，最大似然估计的渐近方差等于 *EIF* 估计量的渐近方差：

$$V_{\mathrm{MLE}} := \big[I_{\gamma_1}(\eta)^{-1}\big]_{11} = V_{\mathrm{EIF}}.$$

证明。定义

$$C := (q_0 + q_1 - 1)^2\theta(1-\theta),$$
$$B := (1-\theta)q_0(1-q_0) + \theta q_1(1-q_1).$$

ˆ关键恒等式。 对 Y应用全方差定律：

$$p(1-p) = \mathrm{Var}(\hat{Y}) = \underbrace{\mathrm{Var}(\mathbb{E}[\hat{Y} \mid Y])}_{=C} + \underbrace{\mathbb{E}[\mathrm{Var}(\hat{Y} \mid Y)]}_{=B} = C + B.$$

化简 得 V_{MLE}。 根据命题 4，逆费舍尔信息矩阵的（1，1）项 is

$$V_{\mathrm{MLE}} = (1+\gamma_1)\theta(1-\theta)\,\frac{p(1-p) + \gamma_1 B}{p(1-p) + \gamma_1(C+B)}.$$

将分母中的 $p(1-p) = C + B$ 代入：

$$\begin{aligned} V_{\mathrm{MLE}} &= (1+\gamma_1)\theta(1-\theta)\,\frac{(C+B) + \gamma_1 B}{(C+B) + \gamma_1(C+B)} \\ &= (1+\gamma_1)\theta(1-\theta)\,\frac{C + (1+\gamma_1)B}{(1+\gamma_1)(C+B)} \\ &= \frac{\theta(1-\theta)}{p(1-p)}\big[C + (1+\gamma_1)B\big]. \end{aligned}$$

与 V_{EIF}比较。 根据命题6：

$$V_{\mathrm{EIF}} = \frac{\theta(1-\theta)}{p(1-p)}\big[\theta(1-\theta)(q_0+q_1-1)^2 + (1+\gamma_1)\big(q_0(1-q_0)(1-\theta) + q_1(1-q_1)\theta\big)\big].$$

这正是 $^{\theta}{}_{(1-\theta)\ p(1-p)}[C + (1+\gamma_1)B]$，由此完成证明 □

B.8 命题7的证明

证明。令 $\kappa = q_0 + q_1 - 1 \in (0, 1]$, , 其中 $\kappa = 1$ 当且仅当 $q_0 = q_1 = 1$ （完美分类）。我们引入以下符号来表示关键量：

$$V_1 = p(1-p) = \mathrm{Var}(\hat{Y}),$$
$$V_2 = (1-\theta)(1-q_0) + \theta(1-q_1) - (\theta - p)^2 = \mathrm{Var}(Y - \hat{Y}),$$
$$V_3 = (1-\theta)q_0(1-q_0) + \theta q_1(1-q_1) = \mathbb{E}[\mathrm{Var}(\hat{Y}|Y)].$$

这里，V_2 是 $\delta = Y - \hat{Y}$ 的方差。注意，$\mathbb{E}[\delta] = \theta - p$ 和 $\mathbb{E}[\delta^2] = \mathbb{P}(Y \neq \hat{Y}) = (1-\theta)(1-q_0) + \theta(1-q_1)$。

我们进一步将两边都除以因子 $\frac{1+\gamma}{\gamma}$，得到

$$V_{\mathrm{PPI}} = V_1 + \gamma_1 V_2, \qquad V_{\mathrm{RG}} = \frac{1}{\kappa^2}(V_1 + \gamma_1 V_3).$$

步骤 1：分解差值。我们有

$$\begin{aligned} V_{\mathrm{RG}} - V_{\mathrm{PPI}} &= \frac{V_1 + \gamma_1 V_3}{\kappa^2} - V_1 - \gamma_1 V_2 \\ &= V_1\left(\frac{1}{\kappa^2} - 1\right) + \gamma_1\left(\frac{V_3}{\kappa^2} - V_2\right) \\ &= \frac{(1-\kappa^2)V_1}{\kappa^2} + \frac{\gamma_1(V_3 - \kappa^2 V_2)}{\kappa^2} \\ &= \frac{(1-\kappa)(1+\kappa)V_1}{\kappa^2} + \frac{\gamma_1(V_3 - \kappa^2 V_2)}{\kappa^2}. \end{aligned}$$

步骤 2：第一项非负。由于 $0 < \kappa \leq 1$，我们有 $(1-\kappa) \geq 0$，$(1+\kappa) > 0$，且 $V_1 = p(1-p) \geq 0$。因此

$$\frac{(1-\kappa)(1+\kappa)V_1}{\kappa^2} \geq 0,$$

当且仅当 $\kappa = 1$ 或 $p \in \{0, 1\}$ 时，等号成立。

步骤 3：第二项非负。接下来只需证明 $V_3 - \kappa^2 V_2 \geq 0$。我们借助全方差定律来证明这一点。

引理 1. $V_3 = V_1 - \theta(1-\theta)\kappa^2$。

证明。对 Y 应用全方差定律：

$$\mathrm{Var}(\hat{Y}) = \mathbb{E}[\mathrm{Var}(\hat{Y}|Y)] + \mathrm{Var}(\mathbb{E}[\hat{Y}|Y]).$$

我们有 $\mathbb{E}[Y|Y=0] = \mathbb{P}(Y=1|Y=0) = 1 - q_0$ 和 $\mathbb{E}[Y|Y=1] = q_1$。因此：

$$\mathrm{Var}(\mathbb{E}[\hat{Y}|Y]) = (1-\theta)(1-q_0-p)^2 + \theta(q_1-p)^2.$$

使用 $p = (1-\theta)(1-q_0) + \theta q_1$，我们计算得：

$$\begin{aligned} 1 - q_0 - p &= (1-q_0) - (1-\theta)(1-q_0) - \theta q_1 \\ &= \theta(1-q_0) - \theta q_1 = -\theta(q_0 + q_1 - 1) = -\theta\kappa, \\ q_1 - p &= q_1 - (1-\theta)(1-q_0) - \theta q_1 \\ &= (1-\theta)q_1 - (1-\theta)(1-q_0) = (1-\theta)(q_0 + q_1 - 1) = (1-\theta)\kappa. \end{aligned}$$

因此：

$$\mathrm{Var}(\mathbb{E}[\hat{Y}|Y]) = (1-\theta)\theta^2\kappa^2 + \theta(1-\theta)^2\kappa^2 = \theta(1-\theta)\kappa^2.$$

因此 $V_3 = V_1 - \theta(1-\theta)\kappa^2$。 □

根据引理 1：

$$V_3 - \kappa^2 V_2 = V_1 - \theta(1-\theta)\kappa^2 - \kappa^2 V_2 = V_1 - \kappa^2\big(V_2 + \theta(1-\theta)\big).$$

我们声称，对于某个 $Q \geq 0$，该表达式等于 $(1-\kappa)Q$ 。通过代数运算（已进行符号验证）可得：

$$V_3 - \kappa^2 V_2 = (1-\kappa) \cdot Q(\theta, q_0, q_1),$$

其中，对于所有 $\theta \in (0,1)$ 和 $q_0, q_1 \in (0,1)$，有 $Q(\theta, q_0, q_1) \geq 0$，且 $q_0 + q_1 > 1$。

由于 $(1-\kappa) = 2 - q_0 - q_1 \geq 0$ （因为 $q_0, q_1 < 1$）且 $Q \geq 0$，因此有：

$$V_3 - \kappa^2 V_2 \geq 0,$$

当且仅当 $\kappa = 1$ 时等号成立（即 $q_0 = q_1 = 1$）。

步骤4：结论。结合步骤2和步骤3：

$$V_{\mathrm{RG}} - V_{\mathrm{PPI}} = \frac{(1-\kappa)(1+\kappa)V_1}{\kappa^2} + \frac{\gamma_1(V_3 - \kappa^2 V_2)}{\kappa^2} \geq 0.$$

因此 $V_{\mathrm{PPI}} \leq V_{\mathrm{RG}}$。

等号成立条件。当且仅当两项都为零时，等号成立。第一项在 $\kappa = 1$ 或 $p \in \{0,1\}$时为零。第二项在 $\kappa = 1$时为零。由于 $p = (1-\theta)(1-q_0) + \theta q_1$和 θ， q_0， $q_1 \in (0,1)$，因此有 $p \in (0,1)$。所以，等号成立要求 $\kappa = 1$，这意味着 $q_0 + q_1 = 2$，即 $q_0 = q_1 = 1$ （分类器是完美的）。□

B.9 第 5.4 节结果的校准拟合5.4

$\hat{}$ 我们在图 $\mu(\hat{}y) = \mathbb{E}[Y \mid Y = \hat{}y]$ 中展示了各方法估计得到的校准函数 9。真条件期望（黑色虚线）由模拟数据生成过程确定：由于$Y \mid Z \sim \mathcal{N}(\mu_Z, \sigma^2)$ 和 $\hat{}Y = Z$，校准函数是分段常数函数，在 $\hat{}Y \in \{1,2,3\}$处存在跳跃。样条/GAM方法能很好地近似这一阶梯函数，而线性校准采用了限制性较强的函数形式，当真实关系为非线性时会引入偏差。这解释了为何线性EIF估计量的置信区间更宽。

Y 与 $\hat{Y}$ 的校准拟合（离散预测）DGP：Z ~ Unif{1,2,3}，Y|Z ~ N(Z, 1)，$\hat{Y}$ = Z。虚线 = 表示完美校准。

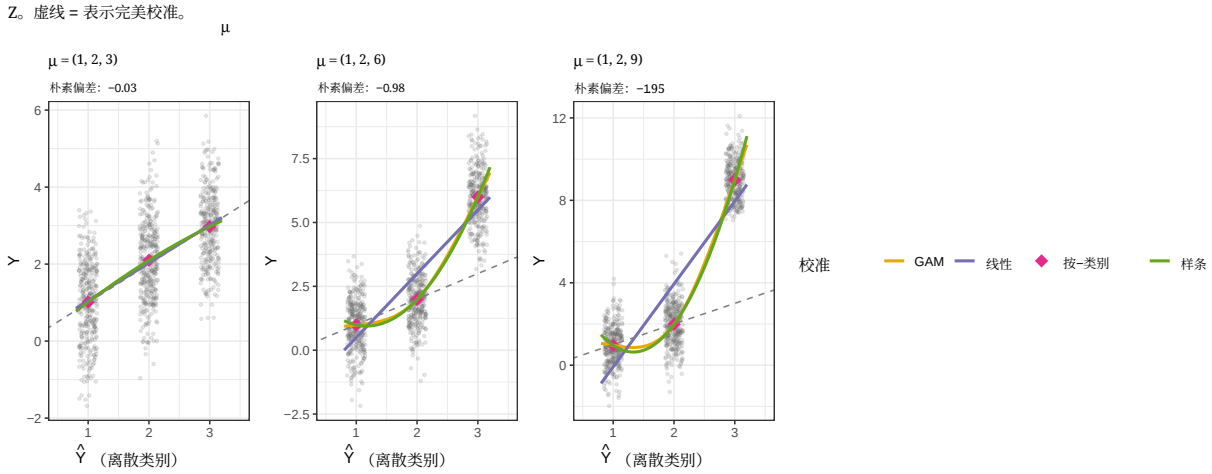


图9：连续 Y 模拟的校准拟合。