arXiv:2302.02923v2 [stat.ML] 6 Jun 2023

探寻洞见，而非灵丹妙药：揭秘异质性处理效应估计中的模型选择困境

Alicia Curth [1] **Mihaela van der Schaar** [1 2]

摘要

在许多高风险应用中，人们往往关注个性化处理效应估计值。因此，在实际部署用于估计此类效应的模型之前，必须确保从不断扩充的机器学习工具箱中为这一任务选出了最佳候选模型。遗憾的是，由于实践中缺乏反事实信息，通常无法依靠标准验证指标来完成这一选择，这也导致处理效应估计文献中一个众所周知的模型选择困境。尽管近期已有研究探讨了一些解决方案，但人们仍缺乏对不同模型选择准则的优势与不足的系统性认识。因此，本文不试图宣布一个普遍适用的“赢家”，而是通过实证研究探讨不同选择准则的成功与和失败模式。我们强调，选择策略、候选估计器以及用于比较它们的数据之间存在复杂的相互作用，并在阐明不同准则相对优势与劣势的同时，提出了在这一背景下开展更具启发性的实证研究所应遵循的设计原则。

1. 引言

在实证科学中，关注处理（或行动、干预和政策）因果效应的应用无处不在；个性化效应估计最终可用于改善医疗保健、经济学和营销等多个领域的决策。机器学习（ML）在提供此类个性化效应估计方面展现出巨大潜力（Bica 等人，2021），相关机器学习文献在过去五年间日趋成熟：近期提出了大量用于估计条件平均处理效应（CATE）的新方法，包括与方法无关的所谓元学习策略，这些策略可使用任意机器学习预测方法实现（K¨unzel 等人，2019年；Nie & Wager，2021；Kennedy，2020；Curth & van derSchaar，2021b），以及针对特定机器学习方法的改造（Shalit 等人，2017；Alaa 和 van der Schaar，2018；Wager 与 Athey，2018），以适用于处理效应估计。

个性化处理效应估计在安全关键应用中往往至关重要，尤其是在医学和政策制定领域。因此，在实际应用之前，我们希望确保从种类繁多的工具中选出的机器学习估计量值得信赖，并能给出尽可能准确的估计。尽管这听起来似乎是一个直截了当的要求，但在实践中仍是一大难题：由于因果推断的基本问题（Holland，1986），真实处理效应通常无法获得，因而无法进行标准的模型验证；要克服这一模型选择困境，就需要考虑其他解决方案。正如下文所述，尽管近期已有新的模型选择策略提案，也有比较不同策略的实证研究，但我们认为，人们仍普遍缺乏对各种策略优缺点的理解，也不清楚它们与底层数据生成过程（DGPs）之间如何相互关联——这正是本文希望阐明的问题。

相关工作。 尽管模型选择问题对异质性处理效应估计具有实际意义，但迄今为止受到的关注仍非常有限（这也与近年来提出的大量新估计量形成鲜明对比，概述见附录 B）。一种直观且常用的解决方案是依赖简单的预测型验证，评估模型对与事实处理相关的观测结果（事实结果）进行 预测的表现，而非 评估效应估计的质量。近期已提出一些更有针对性的替代方案： Rolling &Yang(2014) 提议通过匹配最近邻的处理组单位和对照组单位，并比较其结果，来构造近似的效应验证目标；Nie &Wager (2021) 指出，他们的 R-学习器目标函数也可用于

[1]剑桥大学应用数学与理论物理系，英国 [2]艾伦·图灵研究所。通信请联系：Alicia Curth <amc253@cam.ac.uk>。

《 40th 国际机器学习会议，美国夏威夷州檀香山。PMLR 202，2023年。版权所有 © 2023 作者。

模型选择，Saito & Yasui(2020) 同样提议使用一种与 Kennedy (2020) 的元学习器目标函数相对应的准则，而 Alaa 和 Van Der Schaar(2019年) 则提议依靠影响函数对插件估计进行去偏。我们了解到有两项独立的基准研究比较了这些准则（或其中的部分准则）：Schuler 等人 (2018) 发现 R-学习器目标函数总体表现最佳，而 Mahajan 等人 (2022年) 发现，在所有考察的数据集上，没有任何准则始终优于其他准则（尤其是，他们并未发现 R-学习器目标函数表现突出）。这表明，我们仍有很大的空间去理解不同选择准则的相对优势及其相对劣势。

贡献。 本文着重于系统地理解不同模型选择准则的优缺点。需要指出的是，我们的目标并非提出新方法，而是建立对文献中已有工具的理解与洞见。我们认为，这是该领域迈向实践中真正采用个性化处理效应估计器的最关键、最必要的下一步之一，也可能启发进一步的方法学研究，以弥补这种理解所揭示的不足。为此，我们作出三项贡献：

1. 我们为一个高度复杂的选择问题建立直觉：揭示其内在挑战，通过对现有准则进行分类，为已有研究梳理出结构，并利用这些见解推导出关于其相对表现的假设。
2. 我们提出实验设计的要求，以便厘清这一问题中各种复杂因素的作用：我们倡导开展更完善的实验，通过采用具有有趣变化维度的数据生成过程，并以更透明的方式报告结果，从而系统地考察数据生成过程、候选估计器与选择准则之间的相互作用。

3. 我们通过实证考察现有准则的成功与失效模式，为条件平均处理效应模型选择问题提供了新的见解，并得出结论：在我们考察的所有实验条件下，没有任何一种现有选择准则始终表现最佳。除了重点介绍不同类型选择准则的若干性能趋势外，我们主要考察：i）候选估计器与选择准则在构造时具有相似归纳偏差所导致的相容性偏差；以及 ii）事实选择准则能够（或不能）实现什么。我们发现：i）依赖处理效应插件估计值的选择准则往往偏向于与其插件式估计相似的估计量，而依赖伪结果的选择准则中，这种相容性偏差较不明显；ii）事实选择有时表现不佳，不仅是因为它无法评估所有类型的条件平均处理效应估计器，bu也因为它并未充分针对效应估计进行设计 等。

2. 问题设定

我们考虑潜在结果框架（Rubin，2005）下如今已成为标准的条件平均处理效应估计设定。也就是说，我们假设可以获得一个由 n 独立同分布元组（X、A、Y）构成的数据集，其中 Y是我们关注的结果，X由处理前协变量组成，$A \in \{0,1\}$ 是根据某种（通常未知的）倾向概率$\pi(x) = \mathbb{P}(A = 1|X = x)$分配的二元处理（行动、干预或政策）。我们假设，从概念上讲，每个个体先验地对应两个潜在结果（POs）$Y(a)$ $a \in \{0,1\}$，分别表示接受处理 a时的结果；然而，我们只能观察到个体实际接受处理 A 所对应的结果，即 $Y = Y(A)$。因此，我们可以自然地通过（不可观测的）潜在结果对比$Y(1) - Y(0)$定义个体化处理效应。我们着重估计条件平均处理效应（CATE）$\tau(x)$，即协变量为 $X = x$的个体的预期潜在结果差：

$$\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x] = \mu_1(x) - \mu_0(x) \quad (1)$$

其中 $\mu_a(x) = \mathbb{E}[Y(a)|X = x]$。为确保能够从观测数据中识别并以非参数方式估计处理效应，我们依赖标准的可忽略性假设（Rosenbaum与Rubin，1983年）：

假设2.1（可忽略性。）。（i）一致性。对于处理分配为 A的个体，我们观测到相应的潜在结果，即 $Y = Y(A)$。（ii）无混杂性。不存在未观测混杂因素，因此$Y(0)$ $Y(1) \perp\!\!\!\perp A|X$。（iii）重叠。处理分配是非确定性的，即 $\pi(x) \in (0, 1)$。

因此，$\mathbb{E}[Y(a)|X = x] = \mathbb{E}[Y|X = x, A = a]$，观测到的统计关联具有因果解释。

2.1. CATE估计策略

近期文献提出了大量用于估计条件平均处理效应的策略。其中一类研究最近采用了元学习器框架Künzel 等人（2019年），其中元学习器提供了一套使用任意[1] 机器学习方法 $\mathcal{M}$估计条件平均处理效应的“配方”。由于这类方法易于结合不同的底层机器学习方法实现，且已有相应的理论认识，并与下一节讨论的模型选择策略相对应，本文将重点研究如何在此类元学习器之间进行选择。

按照 Curth 和 vander Schaar（2021b），我们区分了（i）间接估计策略，这类策略通过输出 PO 回归的估计值

[1] 而非对特定另一研究方向提出的机器学习方法进行调整，详见附录B。

以间接估计条件平均处理效应，然后设定处理效应估计值；以及（ii）直接估计策略，这类策略直接输出处理效应的估计值，而不将 PO 估计值作为副产品输出。Künzel 等人（2019年）讨论了两种间接学习器：T-学习器策略，即按处理组拆分训练数据，并在每个样本上分别训练两次，以输出两个回归器；以及 S-学习器策略，即将处理指示变量直接附加到协变量上，只训练一次，从而使用协变量和处理指示变量输出一个估计函数，该函数可用于插补两个潜在结果。由于后一种形式可能导致条件平均处理效应的隐式正则化（Schuler 等人，2018）（因为任何异质性效应都必须由处理指示变量与协变量的学习交互项表示），我们还纳入了一个扩展版本（ES-学习器），该版本显式地使用协变量和处理指示变量进行训练。

或者，也存在一些直接估计器，它们只需依赖于在预处理步骤中使用机器学习方法 $\mathcal{M}$ 得到的部分干扰参数估计 $\eta = (\mu_0(x)\ \mu_1(x)\ \pi(x))$，即可输出条件平均处理效应的估计值。大多数此类策略，尤其是 X-学习器（K¨unzel 等，2019年）、DR-学习器（Kennedy，2020）以及 RA-和 PW-学习器（Curth & van der Schaar，2021b），都采用伪结果方法：利用插件式干扰参数估计 $\hat{}\eta$ 构造伪结果，使得对于真实干扰参数 η，有 $\mathbb{E}[Y_\eta|X=x]=\tau(x)$；随后使用 $\mathcal{M}$ 将 $Y_{\hat{\eta}}$ 对 X 进行回归，从而得到 $\hat{}\tau(x)$ 的估计值。Nie & Wager（2021）的 R- 学习器思路相似，但改用修正损失函数。这些多阶段估计程序近年来在文献中日益流行，因为它们具有良好的理论性质（Curth & van der Schaar，2021b），更为稳健（Kennedy，2020），并且研究发现，在各种情境下，它们的表现均优于单独使用基础版 S-学习器和 T-学习器（Nie & Wager，2021）。

3. CATE 模型选择：理解挑战与现有策略

我们研究的问题是：如何从一组条件平均处理效应估计量中选择一个估计器 $\mathcal{T}=\{\hat{}\tau_1(\cdot)\ldots,\hat{}\tau_K(\cdot)\}$，其中包含不同的元学习器+与机器学习方法组合，使估计异质性效应的精确个体处理效应误差（PEHE）最小化（Hill，2011），该误差是在大小为 n的测试集上估计底层效应函数的均方根误差：

$$\arg\min_{\hat{\tau}_k\in\mathcal{T}} \mathcal{E}_\tau^{oracle}(\hat{\tau}_k) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tau(X_i)-\hat{\tau}_k(X_i))^2}$$

该指标是一个预言机指标，因为它在实践中无法评估——这使得条件平均处理效应模型选择问题极具挑战性。下文将对此问题固有的挑战进行深入讨论，并对为克服这些挑战而提出的模型选择准则进行分类。

3.1. CATE 模型选择为何具有挑战性?

挑战 1：缺乏针对个体处理效应的监督信号。由于因果推断中的根本问题（Holland，1986），也就是说，对任何一个个体，我们只能观测到 $Y(0)$ 或 $Y(1)$，因此用于估计 $\mathbb{E}[Y(1)-Y(0)|X=x]=\tau(x)$ 的真实监督目标标签 $Y(1)-Y(0)$ 无法通过标准的留出验证方法用于模型选择。监督标签的缺失也促使人们构建直接估计效应的元学习器，并使用伪结果进行估计（Kennedy，2020；Curth & van der Schaar，2021b），但这也带来了模型选择的另一项挑战：尽管可以直接估计效应，却无法仅凭观测数据，根据多个直接估计器的输出 $\hat{}\tau_k(x)$ 对其进行比较，因为没有可用于直接验证 $\hat{}\tau_k(x)$ 的自然结果。

挑战 2：混杂会导致处理组之间出现协变量偏移。仅使用事实性（观测到的）结果来验证估计量，一个直接的选择是根据其结果预测能力来评估它们；也就是说，使用 $Y_i-\hat{}\mu_{A_i}(X_i)$ 进行验证。然而，这种方法仅适用于间接学习器，因为它需要输出$\hat{}\mu_a(x)$。即使愿意将注意力限制在间接学习器上，以便采用事实评估策略，仍然存在一个固有的挑战：只在$\hat{}\mu_a(x)$ 的 $A=a$ 观测值个体上进行评估，本质上会受到协变量偏移的影响，只要 $\pi(x)$ 不是常数，因为处理并非完全随机分配。在效应估计中，这同样是一项挑战（Shalit 等人，2017）。

挑战 3：用于良好潜在结果估计和条件平均处理效应估计的选择可能并不相同。最后，即使只在间接学习器中进行选择，且不存在协变量偏移，在有限样本下，潜在结果估计表现最佳的估计量，在条件平均处理效应估计方面也未必表现最佳。显然，当模型设定正确且数据量无限时，完美估计潜在结果便会立即得到完美的条件平均处理效应估计值。然而，当数据有限和／或模型设定错误时，这可能会导致在良好估计潜在结果与良好估计其差值之间产生权衡[2]。

[2]为说明这一点，请注意，当将一个间接估计量 $\hat{}\mu_a^1(x) = \{\mu_1(x) + \tilde{\epsilon}_1(x)\ \mu_1(x) - \tilde{\epsilon}_1(x)\}$ 与另一个估计量 $\hat{}\mu_a^2(x) = \{\mu_1(x) + \tilde{\epsilon}_2(x)\ \mu_1(x) + \tilde{\epsilon}_2(x)\}$ 进行比较时，若对所有 x 都有估计误差$\tilde{\epsilon}_1(x) < \tilde{\epsilon}_2(x)$，那么估计潜在结果的均方误差会在估计量 $\hat{}\mu_a^1(x)$ 下更低——但它对条件平均处理效应的估计误差在每个 x 上都是 $2\tilde{\epsilon}_1$；而对潜在结果估计较差的估计量 $\hat{}\mu_a^2(x)$，其条件平均处理效应估计误差却为 0。

3.2. 模型选择准则的分类

为克服上述挑战，相关研究中已采用或提出了许多替代模型选择准则。下文根据现有模型选择准则最显著的特征，将其分为三类：事实（预测）准则、插件代理准则和伪结果代理准则。图中对这三类策略作了概念性概述。1总体而言，我们考虑的模型选择准则集合与以下基准测试研究所采用的准则集合（的并集）相似：Schuler 等人(2018); Mahajan 等人(2022年)[3]。

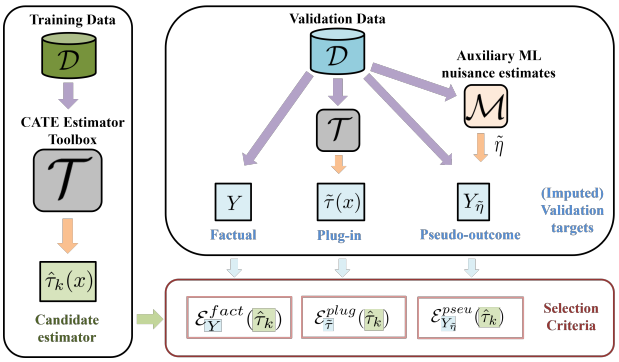


图 1。所考察选择准则的概念性概览。

1. 事实（预测）准则。 首先，如上一节所述，评估模型 $\hat{\tau}_k(x)$ 的一种可行方法是：若模型还输出一对回归器 $\{\hat{\mu}_0^k(x), \hat{\mu}_1^k(x)\}$，则采用一种简单的预测损失，仅考虑观测到的潜在结果

$$\mathcal{E}_Y^{fact}(\hat{\tau}_k) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}_{A_i}^k(X_i))^2}$$

为校正可能存在的协变量偏移，还可以利用 $w(X_i, A_i)$ 中的倾向得分估计，将其转换为重要性加权准则 $\mathcal{E}_Y^{fact,w} = A_i(\hat{\pi}(X_i))^{-1} + (1-A_i)(1-\hat{\pi}(X_i))^{-1}$。

2. 插件代理准则。 因此，要直接评估估计值 $\hat{\tau}_k(x)$，就必须为条件平均处理效应构造替代准则。一种方法是在留出数据上拟合一个新的条件平均处理效应估计器 $\tilde{\tau}(x)$，并用它与这些估计值进行比较：

$$\mathcal{E}_{\tilde{\tau}}^{plug}(\hat{\tau}_k) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tilde{\tau}(X_i) - \hat{\tau}_k(X_i))^2}$$

为了实现这一准则，可以使用任意条件平均处理效应估计器——因此，试图选择最佳插件式估计器，可能会让我们兜回原点，重新陷入最初试图解决的问题。相关研究（Alaa 和 Van Der Schaar，2019年；Mahajan 等人，2022年）仅将间接估计器（S 学习器和 T 学习器）视为插件式代理准则——可能是因为可以根据事实性结果在这些估计器之间进行选择——但我们指出，从技术上说，任何估计器，包括直接估计器，都可以用作 $\tilde{\tau}(x)$。

[3]Schuler 等人(2018)，这些研究早于大多数模型选择论文的发表，因此遗漏了一些插件式和伪结果准则。 Mahajan 等人(2022年)仅未考虑事实性（预测）准则。正如下一节所述，我们认为这类准则非常重要。我们仅将策略价值准则排除在考虑范围之外，即那些用于优化所推导出的治疗策略 $\mathbf{1}\{\hat{\tau}_k(x) > 0\}$的准则；这样做既是因为我们关注精确个体处理效应误差，也因为研究表明，即便以策略价值为评估标准，这些准则的表现也不佳(Schuler 等人，2018)。

3. 伪结果代理准则。 最后，可以采用与直接学习器所依据的相同伪结果策略：给定使用机器学习方法 $\mathcal{M}$ 从验证数据中获得的辅助干扰参数估计值 $\tilde{\eta} = (\tilde{\mu}_0(x)\ \tilde{\mu}_1(x(\ \tilde{\pi}(x)))$，可以构造伪结果 $Y_{\tilde{\eta}}$，使其满足：对于真实干扰参数 η，$\mathbb{E}[Y_\eta|X = x] = \tau(x)$；并且，与在学习器本身中将这些伪结果用作回归结果不同，可以将其用作验证目标

$$\mathcal{E}_{Y_{\tilde{\eta}}}^{pseu}(\hat{\tau}_k) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_{\tilde{\eta}} - \hat{\tau}_k(X_i))^2}$$

这是合理的，因为条件均值$\mathbb{E}[Y_{\tilde{\eta}}|X = x]$ 在期望意义下使均方误差 $\mathcal{E}_{Y_{\tilde{\eta}}}^{pseu}(\cdot)^2$ 最小。双重稳健伪结果的使用，

$$Y_{DR,\tilde{\eta}} = \left(\frac{A}{\hat{\pi}(X)} - \frac{(1-A)}{1-\hat{\pi}(X)}\right)Y + \left[\left(1 - \frac{A}{\hat{\pi}(X)}\right)\hat{\mu}_1(x) - \left(1 - \frac{1-A}{1-\hat{\pi}(X)}\right)\hat{\mu}_0(X)\right]$$

这正是直接元学习器 DR-学习器（Kennedy，2020）所依据的方法，并由此产生了 Saito & Yasui（2020）提出的选择标准。我们在此指出，也可以使用 Curth & van der Schaar（2021b）讨论的其他元学习器伪结果来实现这一目的，例如单重稳健倾向得分加权 $Y_{PW,\tilde{\eta}} = \left(\frac{A}{\pi\ X}\ \hat{(\ )} - (1-A)\ 1-\hat{\pi}(X)\right)Y$。我们也将 Rolling & Yang（2014）的基于匹配的模型选择策略归入这一类别。在~这种情况下，$Y_{\text{match},\tilde{\eta}} = (2A_i - 1)(Y_i - NN_{1-A_i}(X_i))$，其中~$NN_a(X)$是处理组 a中 X的最近邻；这本质上对应于 Curth & van der Schaar（2021b）的 RA-学习器所关联的伪结果，并使用 1近邻回归来估计干扰参数 $\hat{\mu}_a(x)$。我们还将 Alaa 和 Van Der Schaar（2019）基于影响函数的标准归入这一类别，并在附录C中讨论。最后，Nie & Wager（2021）的 R学习器目标函数要求对……进行估计

无条件处理均值 $\mu(x) = \mathbb{E}[Y|X=x]$，依赖于类似的思路[4]，也可用于选择任务（Nie & Wager，2020），由此得到如下标准

$$\mathcal{E}_R^{pseu}(\hat{\tau}_k) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \tilde{\mu}(X_i) - \hat{\tau}_k(X_i)(A_i - \tilde{\pi}(X_i)))^2}$$

4. 模型选择困境揭秘

在建立了模型选择策略的高层次分类之后，我们现在可以基于这些策略的内在特征（第4.1节），形成直觉并对其预期的优势与劣势进行假设。这些预期引出了许多有趣的研究问题，而我们认为，只有通过精心设计的受控实验，才能将这些问题区分开来——我们认为，相关研究迄今为止忽视了这一点。因此，接下来我们将讨论构建条件平均处理效应模型选择问题实证研究的设计原则（第4.2节），并在实验部分加以应用。

4.1. 比较模型选择准则：对优势与劣势的预期

结合我们对不同条件平均处理效应模型选择策略的高层次概览与前文讨论的挑战，我们认为，每一类准则都有其固有的优势与劣势。

事实（预测）准则。我们认为，事实性准则在实践中可能颇具吸引力，因为至少未加权的 $\mathcal{E}_Y^{fact}$ 无需估计任何干扰参数，只依赖观测数据；这意味着它不会带来额外开销，而且其结果也值得信赖，因为它不依赖于可能存在模型设定错误或偏差的干扰参数估计。不过，这类准则主要评估对潜在结果的估计表现；如上文所述，这可能会错误地优先考虑对潜在结果的良好拟合，而非良好的条件平均处理效应拟合——后者根本无法衡量。更关键的是，这也意味着事实性准则无法评估所有类型的方法，尤其无法评估所有直接估计器，因此只能从所有可能的估计量中选择较小的一个集合 $\mathcal{T}_{indirect} \subset \mathcal{T}$，也就可能因其自身构造而错过表现最佳的估计量。

插件代理准则。与事实值准则相比，插件代理准则的明显优势在于，它们可以评估所有类型的估计量，且目标正是我们感兴趣的结果（即条件平均处理效应）。然而，由于需要一个插件估计值 $\tilde{\tau}(x)$，这会带来额外的估计误差或方差风险。此外，$\tilde{\tau}(x)$ 可以是任何条件平均处理效应估计器，因此，选择一个好的插件 $\tilde{\tau}(x)$会让我们回到最初试图解决的困境。最后，我们认为，这类替代准则也可能受到一种我们称为相容性偏差的现象影响：它们可能偏好在结构上与其插件估计器 $\tilde{\tau}(x)$相似的条件平均处理效应估计量 $\hat{\tau}_k(x)$。尽管 $\hat{\tau}_k(x)$ 和 $\tilde{\tau}(x)$ 应在不同的数据折上拟合，我们仍预计，插件准则 $\tilde{\tau}(x)$可能会偏好带有相似归纳偏差的估计量。也就是说，我们预计，例如，使用线性回归实现的插件替代准则 $\tilde{\tau}(x)$，可能会偏好同样依赖线性回归的条件平均处理效应估计量 $\hat{\tau}_k(x)$（而非采用其他方法实现的估计量 $\mathcal{M}$）；而依赖 S-学习器替代准则 $\tilde{\tau}(x)$的准则，则可能更倾向于选择 S学习器 $\hat{\tau}_k(x)$。这里，我们借用了心理学文献中的“相容性偏差”一词，用来指个体可能会系统性地偏好与当前信念一致的信息（Hart et al., 2009）。

伪结果代理准则。我们预计，这些准则将兼具插件准则的优点和部分缺点（即需要估计额外参数，以及由此可能导致误差或方差增加）。由于它们不使用最终的 $\tau(x)$ 估计，而仅使用伪结果 $Y_{\tilde{\eta}}$，我们预计它们不太容易受到相容性偏差的影响，但仍可能偏向具有相似归纳偏差的估计量，例如基于相同伪结果训练的直接估计器。

4.1.1. 由此提出的研究问题。

本文的研究动机来自下文列出的许多有趣研究问题，这些问题自然地源于上述讨论；我们认为，相关研究尚未解决其中任何一个问题——我们认为，这是因为迄今为止，研究重点一直在于确立某种准则的全局最佳性能。相反，我们希望了解不同准则之间可能出现性能差异的情形，并希望这能帮助从业者在具体应用中选择合适的准则。具体而言，本文将深入探讨三个问题：

- **问题1：** 不同的选择准则分别在什么情况下效果更好或更差？如果存在系统性规律，这些规律是否与以往研究中针对不同估计策略 观察到的规律相似？
- **问题2：** 替代指标选择准则是否确实会像预期的那样受到共轭偏差的影响？不同类型之间是否存在差异？
- **Q3.** 依赖事实验证 $\mathcal{E}_{\mu_a}^{fact}$会让我们在什么方面、何时遭受损失？限制估计器池是否有影响？优化了错误的目标是否有影响？

[4] $\mathcal{E}_R^{pseu}$ 可以改写为伪结果形式：$Y_{R,\,\tilde{\eta}} = \frac{Y_i - \tilde{\mu}}{(X_i)\,A_i - \tilde{\pi}(X_i)}$ 若结合权重 $(A_i - \tilde{\pi}(X_i))^2$ 用于均方根误差求和项中（Knaus 等人，2021）。

4.2. 确立实验设计的理想要求 ign

如上所述，我们认为，在这个条件平均处理效应模型选择困境中，还有许多值得探讨的有趣问题。我们所知的现有实证研究——无论是提出新模型选择准则的研究（Alaa 和 Van Der Schaar，2019年；Saito & Yasui，2020），还是对现有准则进行基准测试的研究（Schuler 等人，2018；Mahajan 等人，2022年）——主要关注的问题都是“哪种策略在总体上效果最好？”，并且采用了黑箱方法：使用不透明的数据集进行基准测试，考虑大量难以拆分的候选估计量，并且主要报告多个不同数据生成过程下的平均结果。

近期研究指出，机器学习条件平均处理效应估计文献中的基准测试实践普遍存在诸多不足（Curth 等人，2021年），尤其是过度依赖单一的半合成数据集[5]，这些数据集的数据生成过程中编码了非常具体的问题特征，却没有讨论这些选择所产生的影响。因此，我们认为，精心设计受控实验环境至关重要，这样才能区分数据生成过程的不同组成部分对选择准则表现的影响。下文将讨论设计实证研究时应满足的理想要求，以便深入了解我们所关注的模型选择表现；下一节中，我们将据此设计一项实证研究。

1. 使用具有有意义“实验旋钮”的数据生成过程。为了系统地了解不同选择准则在何种情况下表现良好，我们需要能够系统地改变底层实验特征——正如Dorie 等人（2019年）所建议的那样，我们设计模拟实验，使其能够调节重要的“实验旋钮”。为了深入了解选择准则的相对表现，我们将选择考察那些已被证明会影响估计量表现的维度，以及我们预期也会产生影响的其他维度。

2. 考察候选估计器的表现$\mathcal{T}$。为了了解不同选择器在何种情况下表现良好，我们认为，首先弄清楚 $\mathcal{T}$ 中底层候选估计器的表现十分重要。据我们所知，所有相关研究都跳过了这一步，但我们认为这至关重要，因为不同选择策略的表现可能与底层估计器的表现密切交织：例如，如果选择器与估计器之间具有相合性，和／或某类估计器在特定数据集上占据优势，那么相应的选择器可能从一开始就更容易表现出色。

[5]请注意，由于真实数据中缺乏 $Y(1) - Y(0)$，通常需要在实验中模拟结果，以便获得已知的真实值；大多数现有基准数据集都是半合成的，即使用真实协变量，但模拟结果。

3. 分析选择准则的表现如何以及何时发生变化。最后，在构建数据生成过程并考察估计器之后，我们旨在详细分析选择准则的表现。也就是说，我们希望明确了解，随着实验因素的改变，选择器的相对表现如何变化。在许多相关研究中，这一点无法做到，因为结果是以平均值的形式汇总多个不同设置（Alaa 和 Van Der Schaar，2019年；Mahajan 等人，2022年），从而掩盖了系统性表现差异中可能存在的有趣见解。

5. 实证研究

设置。本节按照上述三个步骤开展实证研究，比较条件平均处理效应选择准则。整个研究中，我们采用两种机器学习方法$\mathcal{M}$ 来实现所有元学习器和选择准则：极端梯度提升树（Chen 和 Guestrin，2016）（梯度提升）以及带岭惩罚的线性回归（逻辑回归）。选择这两种方法，是因为它们体现了截然不同的归纳偏差，使我们能够深入了解灵活模型与刚性模型之间的性能差异。在需要倾向得分估计时，我们使用逻辑回归进行估计。作为待选元学习器，我们考虑（间接的）S、T 和 ES 学习器，以及（直接的）DR 和 R 学习器。作为选择准则，正文中考虑 $\mathcal{E}_Y^{fact}$（事实性），$\mathcal{E}^{plug,\ ES}_{\tilde{\tau}}$，$\mathcal{E}^{plug,\ T}_{\tilde{\tau}}$，$\mathcal{E}^{plug,\ DR}_{\tilde{\tau}}$ & $\mathcal{E}^{plug,\ R}_{\tilde{\tau}}$（替插件式）以及 $\mathcal{E}^{pseudo\ DR}_{Y_{\tilde{\eta}}}$ & $\mathcal{E}^{pseudo\ R}_{Y_{\tilde{\eta}}}$（替代伪结果）。使用通常表现较差的准则所得到的更多结果见附录D.1。实现细节[6] 见附录C。

5.1. 第一步：设计一个有洞察力的数据生成过程。

我们构建了一个数据生成过程，其设计大致借鉴了Curth& van der Schaar （2021a）所采用的设置：我们同样使用 X ACIC2016 数据集中的协变量（Dorie 等人，2019年），并自行模拟结果和处理分配，以提高透明度和可控性。我们首先在随机抽取的截断点处，将所有连续变量二值化，得到处理后的协变量 X^*，然后在 $\mu_0(\cdot)$的线性模型中使用这些协变量，其中包含 X^*的最高三阶交互项；而 $\tau(\cdot)$ 则仅是 X^*的线性函数。该设置包含三个主要的实验调节因素：

1. **条件平均处理效应复杂度**。我们的第一个实验调节因素是 $\rho \in \{0, .1, .3\}$，即 $\tau(\cdot)$中输入项非零系数所占的比例，它控制了 $\tau(\cdot)$ 的复杂度（稀疏性）。这一因素曾被用于Curth& van derSchaar (2021a)，研究显示它会影响不同估计量的相对表现。

2. **模型设定错误**。其次，我们特意引入了协变量的变换，因为这使我们能够

[6]复现所有实验的代码可在以下网址获取：
https://github.com/AliciaCurth/CATESelection

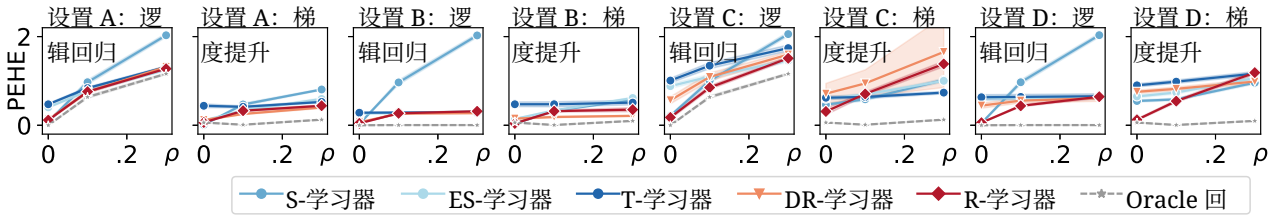


图 2. **不同候选估计器的条件平均处理效应估计误差（PEHE），见** $\mathcal{T}$。所有学习器均使用线性回归（LR）和极端梯度提升（GB）实现，并在 4 种不同设置下进行评估，其中 $\tau(x)$ 的复杂度随 ρ 增加。阴影区域表示一个标准误（SE）。

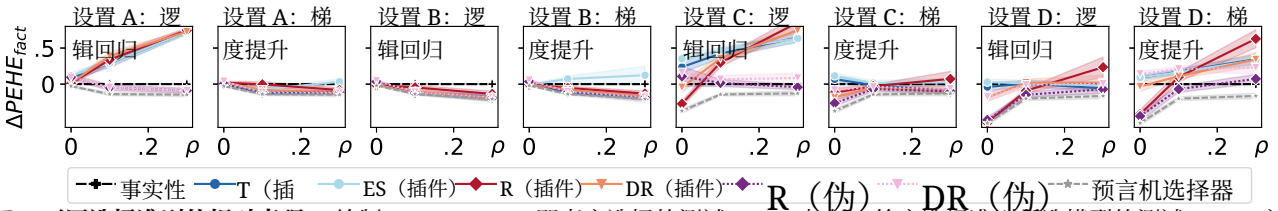


图 3.**不同选择准则的相对表现**。绘制 ΔPEHE_{fact}，即事实选择的测试 PEHE 与任一给定选择准则所选模型的测试 PEHE 之差（越低越好，负值表示优于事实选择）。这些模型使用线性回归（LR）和极端梯度提升（GB）实现。每种准则均可使用 $\mathcal{T}$，即图 2 上方所示性能对应的 10 个候选估计器的完整集合。阴影区域表示一个标准误（SE）。

探究模型（误）设定的影响。也就是说，算法（估计量和选择准则）的输入是 X 还是 X^* X^{input}，是我们考虑的第二个实验调节因素：当输入为原始数据 X时，这会隐式偏向基于树的模型，如梯度提升（GB）（该数据生成过程模拟了决策树中的分裂），因为逻辑回归（LR）无法完全恢复 X^*，因而无法学习 $\tau(\cdot)$ 或 $\mu_0(\cdot)$ 中的模式。即使输入变换后的数据 X^* 作为 X^{input}，简单的逻辑回归也无法拟合潜在结果（POs）$\mu_a(\cdot)$，这是由于存在高阶交互项；然而，处理效应本身关于 X^* 是线性的，因此，如果我们观察到 $Y(1) - Y(0)$，就可以用逻辑回归进行拟合。

3. **混杂**。第三，我们比较混杂设置与随机设置；只有在前一种设置中，处理才会依据进入 $\mu_0(\cdot)$的变量进行分配。正文中，倾向得分对数几率关于 X^{input}呈线性，因此在所有设置下均可通过逻辑回归一致估计。我们在附录 D.4中还考察了其他倾向概率的设置，并探讨处理组规模不平衡的影响（而正文中各组规模相同）。

我们在图的 x 轴上改变 ρ，并根据特征 2 和 3 将其分为四种设置：设置 A 中不存在混杂，估计量和选择器以 X 为输入；设置 B 中不存在混杂，估计量和选择器以 X^* 为输入；设置 C 中存在混杂，估计量和选择器以 X 为输入；设置 D 中存在混杂，估计量和选择器以 X^* 为输入。数据生成过程的更正式描述见附录C。此外，在全文中，我们将大小为 n 的训练数据划分为 $n_{train} = \frac{2n}{3}$，用于训练所有估计量；以及 $n_{val} = \frac{n}{3}$，供选择准则使用，并使用大小为 $n_{test} = 500$ 的独立测试集来根据其选出的最佳模型在测试集上的 PEHE 来评估准则。我们默认使用 $n = 1000 + 500$，但也会在附录 $n = 2000 + 1000$ 中考察数据量增加（$n = 2000 + 1000$）或减少（$n = 500 + 250$）的影响 D.2。每项实验均使用 20 个随机种子重复进行，并报告其均值和标准误（SE）。最后，我们在附录E中补充了使用标准 ACIC2016 和 IHDP 基准数据集的结果，并说明了我们为何认为这些数据集不太适合开展有趣的分析——这进一步表明，精心设计数据生成过程对于获得有意义的洞见至关重要。

5.2. 步骤 2：考察候选估计器的性能

这里，我们在图中简要考察基础学习器本身的表现。2。我们纳入一个预言机，直接基于（通常未知的）$\tau(X_i)$ 进行训练，以此提供由模型设定错误导致的误差下界（特别是，这表明在设置 A 和 C 中，逻辑回归模型无法很好地捕捉条件平均处理效应，而梯度提升模型可以）。我们发现，不同学习器和设置之间确实存在有趣的性能差异，这意味着没有任何单一学习器、估计策略或机器学习方法能够始终优于其他所有方法。正如预期，R-学习器和 DR-学习器表现良好，尤其是在条件平均处理效应相对简单时。对于 $\rho = 0$，T-学习器和 ES-学习器表现最差，但在 ρ 较大时表现相对更好；S-学习器则恰恰相反。请注意，逻辑回归 S-学习器在 $\rho > 0$ 上的表现尤其差，因为它只能学习恒定的处理效应（即，对于 $\rho > 0$，其模型设定错误十分严重）。在附录 D.4中，我们还考察了处理组规模不平衡的情况，发现这主要会降低 R-学习器的表现。

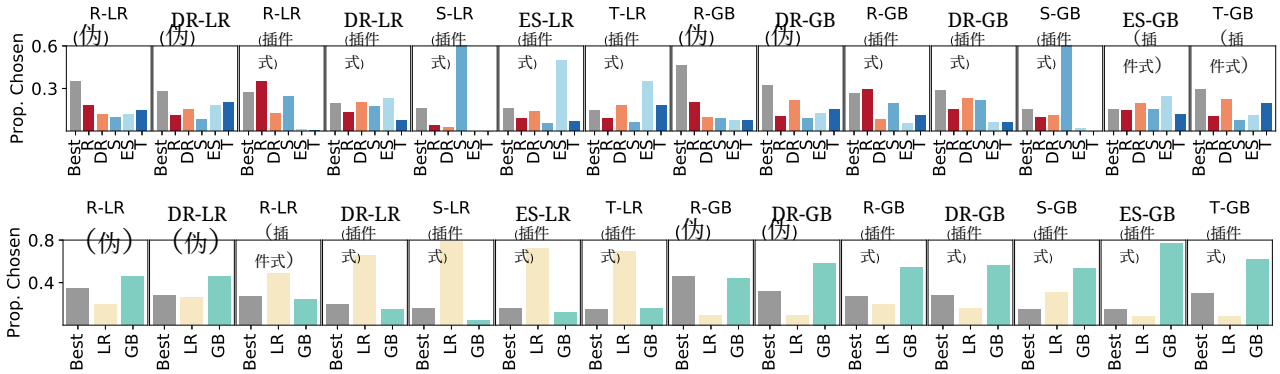


图*4*。 **考察选择准则与估计量之间是否存在相合性偏差：估计策略（上）和机器学习方法（下）**。 衡量在所有所考察的设置中，真实最优估计器被选中的比例（每个图中最左侧的柱），以及在真实最优估计器未 被选中时所选择的估计器类型。

5.3. 步骤3：了解不同选择准则的性能

现在，我们可以考察不同模型选择准则的表现了：在图3中，我们展示了不同准则在全部 10 种学习器-方法组合中进行选择的结果。为便于阅读，我们以 ΔPEHE_{fact}来报告表现，即任一给定选择器的测试 PEHE 与事实选择之间的差值（越低越好，负值表示优于事实选择）。

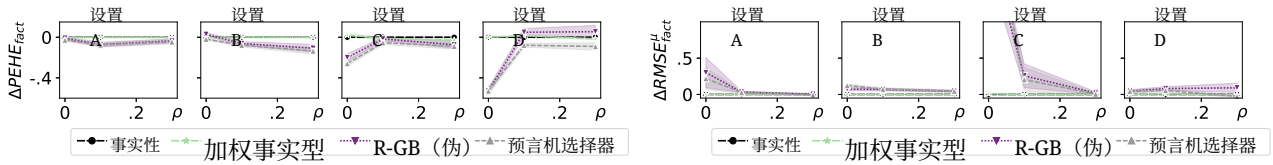
问题1：总体性能趋势。 在图 3中，我们观察到，在所有选择器中，当以下情况出现时，相对于事实性性能，大多数性能差异和提升才会出现：i）处理是随机分配的且 处理效应复杂（ρ 在A和B中较大）；ii）存在混杂 且 处理效应简单（ρ 在C和D中较小）；以及iii）对于潜在结果而言，逻辑回归模型的设定错误，但对于条件平均处理效应而言没有设定错误（B和D）。我们观察到，插件准则的性能通常遵循其底层方法的性能（即比较图 2 和3中的趋势）。此外，基于ES学习器和T-学习器的插件准则总体表现最差，尤其是在使用逻辑回归模型实现时。在设置C和D中，基于R-学习器的插件替代准则在条件平均处理效应复杂度较低时表现良好，但在复杂度较高时则表现不佳——这与上一节中观察到的底层估计量模式相似。总体而言，伪结果准则的表现似乎最佳，其中R伪结果通常与预言机选择器最为相似。还需注意的是，尽管基于间接学习器的插件准则在模型设定错误时（设置A和C中的逻辑回归模型）性能显著下降，但即使使用设定错误的模型（逻辑回归模型）实现，伪结果准则仍能表现良好。

Q2. 相容性偏差。 接下来，我们考察是否存在相容性偏差的证据——也就是说，插件式或伪结果代理准则是否似乎会内在地偏向于选择归纳偏差与用于提供验证目标的策略相似的估计量。我们拟通过计算以下情况发生的比例来衡量这一点：(i) 当某种估计策略或 (ii) 底层机器学习方法被验证准则选中 但未选出最佳估计量时（直观地说，我们作此区分是因为，只要某个估计量是预言机选择，选中它就不应被视为存在偏差）。[7] 在附录 D.3中，我们给出了一个类似的图，但未作此区分。

在图 4（上）中，我们考察选择准则与估计量策略（即使用任一 机器学习方法实现的 R-、DR-、S-、ES- 或 T-学习器）之间的相合性，并汇总了图 3中所有设置下的结果。我们发现，有明确证据表明，某些插件式准则与其对应的学习策略之间存在相合性偏差；这种现象在依赖间接学习器的准则中最为显著，尤其是插件式 S-学习器和 ES-学习器。基于 R-学习器的插件式准则也明显受到这种偏差的影响，而 DR 插件式准则表现出的这种现象较弱。总体而言，伪结果准则对自身策略的偏好不那么明显；其中，伪结果 R 准则和 DR 准则的逻辑回归实现，整体上最少表现出此类相合性偏差。

在图 4 （下图）中，我们进一步考察选择准则与估计量方法（即LR或GB，用于任何估计策略）之间的相容性偏差。在这里，我们同样观察到几乎所有准则都存在明显相容性偏差的证据：基于LR（GB）的准则似乎更倾向于选择使用LR（GB）实现的学习器。唯一的例外似乎是

[7]更形式化地说，我们考察选择器 $\mathcal{E}_j$ 选择使真实（不可观测）精确个体处理效应误差 $\mathcal{E}_\tau^{oracle}$最小的最佳估计量$\hat{\tau}^*$ 的频率；如果没有选择这一首选项$\hat{\tau}^*$，则考察它选择了哪种其他类型的估计量。也就是说，在图 4 中，我们通过标为“最佳”的最左侧柱形来衡量正确选出最佳估计量的频率——即 $\hat{P}(\arg\min_{\hat{\tau}_k} \mathcal{E}_j(\hat{\tau}_k)=\hat{\tau}^*) = 1-\alpha_{\mathcal{E}_j}$，其中 $\alpha_{\mathcal{E}_j}$是选择器的错误率——其他柱形则表示在 $\mathcal{E}_j$ 未能作出正确选择时，选中某一特定类型估计量 $\hat{\tau}_l \neq \hat{\tau}^*$ 的比例。如果某类估计量被选中的频率明显偏高——即 $\hat{P}(\arg\min_{\hat{\tau}_k} \mathcal{E}_j(\hat{\tau}_k) = \hat{\tau}_l | \hat{\tau}_l \neq \hat{\tau}^*) = 1 - \alpha_{\mathcal{E}_j} >> \frac{\alpha_{\mathcal{E}_j}}{|\mathcal{T}|}$——我们便认为这表明选择器 $\mathcal{E}_j$ 可能偏向于选择$\hat{\tau}_l$这类估计量。



（a）条件平均处理效应估计性能　　（b）潜在结果估计性能

图 *5*。 **仅在间接学习器之间进行选择时，不同选择准则的相对性能**。 左图：绘制 ΔPEHE_{fact}，即事实选择与任一给定选择准则选出的模型之间测试 PEHE 的差值。右图：绘制 $\Delta\text{RMSE}^{\mu}_{fact}$，即使用事实选择选出的模型与任一给定选择准则选出的模型估计潜在结果时，其平均 RMSE 之间的差值。（在两图中，数值越低越好，负值表示优于事实选择。）

LR伪结果选择器实际上更常选择GB学习器；这或许可以部分解释它们相较于其他基于LR的选择器表现较好的原因。

问题 3：事实评估会让我们失去什么？ 最后，我们来考察这样一个问题： 为什么 我们在图 3 中看到，在某些情景下，如此多的使用处理效应代理指标的选择器能够优于事实选择。一种可能的解释是，将直接学习器排除在可用的候选池之外 $\mathcal{E}_Y^{fact}$ ——然而，正如图 2所示，通常存在某个间接学习器能够达到与直接学习器相当的性能。我们希望检验另外两种解释：*i*）混杂导致的协变量偏移；以及 *ii*）事实选择错误地将重点放在潜在结果估计性能上。我们可以通过纳入一个重要性加权事实选择器来检验 *i*），并通过将所有选择器可用的候选估计器池限制为所有选择器 均可用的估计器 $\mathcal{T}_{indirect}$ （即将 R-学习器和 DR-学习器排除在估计器池之外）来检验 ii）。

在图 5（左）中，我们观察到，i）似乎并非如此，因为加权和未加权事实选择的表现完全相同（对此的一种可能解释是，所考虑的间接估计器都没有自行进行内部协变量偏移校正）。然而，针对 ii），我们确实观察到，预言机和伪结果选择器似乎都选择了与 $\mathcal{E}_Y^{fact}$ 不同（且更优）的间接估计器，即使不存在协变量偏移也是如此（在设置 B 中尤为明显）。图 5（右）则表明，这些估计器在估计潜在结果本身时表现确实较差，这种权衡正是我们在第 4节中所预期的。

6. 结论

我们研究了条件平均处理效应模型选择问题，着重理解近期研究中采用或提出的不同模型选择策略（使用事实值、插件式替代指标或伪结果代理指标）的优势与劣势。我们没有试图评定一个普遍适用的“赢家”，而是通过实证研究考察了不同策略的成功与失效模式，并由此发现，在某些情境下，事实选择可能是合适的；而在另一些情境下，伪结果代理指标方法可能表现更好（只有插件式替代指标方法似乎始终容易表现不佳）。我们希望，本文提出的一些见解能为从业者提供一个起点，帮助他们判断自身应用可能具备的特征对应于我们考察的哪些情境。为此，我们还在附录中以与一位虚构读者进行问答的形式，提供了对研究发现的补充概述。A。我们还指出，选择策略、候选估计器与用于测试的数据生成过程之间存在复杂的相互作用；当估计器和选择策略的归纳偏差相契合时，很可能会产生相容性偏差。通过这一研究，我们也希望向学界表明，有必要开展更多模拟研究，并依托精心构建的数据生成过程，以便厘清这一问题中各种因素的作用，从而进行更细致入微的分析。

局限性。 最后请注意，我们并不声称研究结果已面面俱到：为了获得有趣且细致的见解，我们需要将研究范围限定在特定问题和候选估计器上。我们认为，该领域有大量值得探索的问题，而我们仅初步选择了一些问题，作为讨论的起点。例如，下一步可以考察不同标准在其他类别估计器之间进行选择时的表现，例如扩展 Shalit 等人（2017）研究的、特定于方法的神经网络估计器；或者在更微观的层面上，研究如何为任何特定方法调整超参数。我们的实验结果仅限于回答一些我们认为最引人关注的问题，但我们希望，本文讨论的实验设计要求以及所提供的实现，能够使未来研究轻松拓展至更多问题及相关数据生成过程。

致谢

我们谨向 Toon Vanderschueren、vanderschaar-lab 成员以及匿名审稿人致谢，感谢他们对本文早期稿件提出的富有洞见的意见并参与讨论。AC 衷心感谢 AstraZeneca 提供的资助。

参考文献

Alaa, A. 和 van der Schaar, M. 异质性处理效应估计的局限：实用算法设计指南。载于国际机器学习会议，第 129–138 页。PMLR，2018 年。

Alaa, A. 和 Van Der Schaar, M. 使用影响函数验证因果推断模型。载于国际机器学习会议，第 191–201 页。PMLR，2019年。

Assaad, S., Zeng, S., Tao, C., Datta, S., Mehta, N., Henao, R., Li, F. 和 Duke, L. C. 使用平衡权重进行反事实表征学习。载于国际人工智能与统计学会议，第 1972–1980 页。PMLR，2021。

Athey, S. 和 Imbens, G. 用于异质性因果效应的递归划分。《美国国家科学院院刊》，113(27):7353–7360，2016。

Athey, S., Tibshirani, J., and Wager, S. 广义随机森林. 统计年刊, 47(2):1148–1178, 2019年.

Athey, S., Imbens, G. W., Metzger, J., and Munro, E. 使用Wasserstein生成对抗网络设计蒙特卡罗模拟. 计量经济学杂志, 2021.

Bica, I., Alaa, A. M., Lambert, C., and Van Der Schaar, M. 利用机器学习从真实世界患者数据中估计个体化治疗效果：应对潜在挑战的当前与未来方法. 临床药理学与治疗学, 109(1):87–100, 2021.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. 机器学习软件的API设计：scikit-learn项目的经验. In*ECML PKDD*研讨会：数据挖掘与机器学习语言, pp. 108–122, 2013.

Chen, T. 和 Guestrin, C. XGBoost：一种可扩展的梯度提升树系统。载于 第 *22* 届 *ACM SIGKDD* 知识发现与数据挖掘国际会议论文集，第 785–794 页，2016。

Cui, Y. 和 Tchetgen, E. T. 双重稳健泛函的选择性机器学习。*arXiv* 预印本 *arXiv:1911.02029*，2019年。

Curth, A. 和 van der Schaar, M. 关于异质性处理效应估计中的归纳偏差。《神经信息处理系统进展》，34:15883– 15894，2021a。

Curth, A. 和 van der Schaar, M. 异质性处理效应的非参数估计：从理论到学习算法。载于国际人工智能与统计学会议，第 1810–1818 页。PMLR，2021b。

Curth, A.、Svensson, D.、Weatherall, J. 和 van der Schaar, M. 条件平均处理效应的估计真的做得很出色吗？对处理效应估计中机器学习基准测试实践的批判性审视。载于第三十五届神经信息处理系统会议数据集与基准测试专题（第二轮），2021。

Dorie, V.、Hill, J.、Shalit, U.、Scott, M. 和 Cervone, D. 自动化方法与自主方法在因果推断中的比较：从数据分析竞赛中汲取的经验。统计科学，34(1)：43–68，2019年。

Hahn, P. R.、Murray, J. S.、Carvalho, C. M. 等。用于因果推断的贝叶斯回归树模型：正则化、混杂与异质性效应（附讨论）。贝叶斯分析，15(3)：965–1056，2020。

Hart, W., Albarrac´ın, D., Eagly, A. H., Brechan, I., Lindberg, M. J., 和 Merrill, L. 感觉自己得到认可，而非事实正确：关于选择性接触信息的荟萃分析。心理学通报, 135(4):555, 2009.

Hassanpour, N. 和 Greiner, R. 使用重要性抽样权重进行反事实回归。载于国际人工智能联合会议, 第 5880– 5887 页, 2019a.

Hassanpour, N. 和 Greiner, R. 学习用于反事实回归的解耦表征。载于国际学习表征会议, 2019b.

Hill, J. L. 用于因果推断的贝叶斯非参数建模。计算与图形统计学杂志, 20(1):217–240, 2011.

Holland, P. W. 统计学与因果推断。《美国统计学会会刊》, 81(396):945–960, 1986.

Johansson, F.、Shalit、U. 和 Sontag、D. 学习用于反事实推断的表示。在《国际机器学习会议》上，第 3020–3029 页。PMLR，2016。

Johansson, F. D.、Kallus、N.、Shalit、U. 和 Sontag、D. 学习用于跨设计泛化的加权表示。《*arXiv* 预印本 *arXiv:1802.08598*》，2018。

Kennedy, E. H. 异质性因果效应的最优双重稳健估计。*arXiv* 预印本 *arXiv:2004.14497*，2020。

Knaus, M. C., Lechner, M. 和 Strittmatter, A. 异质性因果效应的机器学习估计：蒙特卡洛实证证据。《计量经济学期刊》，24(1):134–161，2021。

K¨unzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. 使用机器学习估计异质性处理效应的元学习器。《美国国家科学院院刊》，116(10):4156–4165，2019年。

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. 使用深度潜变量模型进行因果效应推断。载于《神经信息处理系统进展》，第6446–6456页，2017年。

Mahajan, D., Mitliagkas, I., Neal, B., and Syrgkanis, V. 异质性因果效应估计中模型选择的实证分析。*arXiv*预印本 *arXiv:2211.01939*，2022年。Nie, X. and Wager, S. 异质性处理效应的准预言机估计。《生物统计学》，108(2):299–319，2021年。Parikh, H., Varjao, C., Xu, L., and Tchetgen, E. T. 验证因果推断方法。载于国际机器学习会议，第17346–17358页。PMLR，2022年。Rolling, C. A. and Yang, Y. 处理效应估计的模型选择。《英国皇家统计学会会刊：*B*辑（统计方法学）》，76(4):749–769，2014年。Rosenbaum, P. R. and Rubin, D. B. 倾向得分在观察性研究因果效应分析中的核心作用。《生物统计学》，70(1):41–55，1983年。Rubin, D. B. 使用潜在结果进行因果推断：设计、建模与决策。《美国统计学会会刊》，100(469):322–331，2005年。Saito, Y. and Yasui, S. 反事实交叉验证：用于因果推断模型的稳定模型选择程序。载于国际机器学习会议，第8398–8407页。PMLR，2020年。Schuler, A., Jung, K., Tibshirani, R., Hastie, T., and Shah, N. 合成验证：为给定数据集选择最佳因果推断方法。*arXiv*预印本 *arXiv:1711.00083*，2017年。Sc huler, A., Baiocchi, M., Tibshirani, R., and Shah, N. 个体处理效应估计中模型选择方法的比较。*arXiv* 预印本 *arXiv:1804.05146*，2018年。Shalit, U., Johansson, F. D., and Sontag, D. 个体处理效应估计：泛化界与算法。载于国际机器学习会议，第3076–3085页。PMLR，2017年。

Wager, S. 和 Athey, S. 使用随机森林估计异质处理效应并进行推断。《美国统计学会会刊》，113(523):1228–1242，2018。Wu, P. 和 Fukumizu, K. \beta-intact-vae：在重叠有限的情况下识别和估计因果效应。*arXiv* 预印本 *arXiv:2110.05225*，2021。Yoon, J.、Jordon, J. 和 Van Der Schaar, M. Ganite：使用生成对抗网络估计个体化处理效应。载于国际学习表征会议，2018。Zhang, Y.、Bellot, A. 和 van der Schaar, M. 学习重叠表示以估计个体化处理效应。*arXiv* 预印本 *arXiv:2001.04754*，2020。

附录

本附录的结构如下：在附录A 中，我们以与一位虚构读者进行问答的形式，补充概述关键要点。在附录B中，我们补充回顾了相关文献，讨论近期机器学习文献中提出的条件平均处理效应估计量。附录C 讨论实验细节，附录D则补充呈现正文中实证研究的结果。附录E呈现其他数据集上的结果。

A. 要点：与一位虚构读者的问答，探讨本文的关键见解

我们研究了 CATE 模型选择问题，着重于理解近期研究中已采用或提出的不同模型选择策略——使用事实值、插件代理指标或伪结果代理指标——各自的优势与不足。我们没有试图断定哪种策略是普遍的“赢家”，而是通过实证研究不同策略的成功与失败情形，希望本文提供的一些见解能为从业者提供起点，帮助他们判断自身应用可能具有的特征如何对应到我们考察的情形。鉴于本文结果可能较难消化，我们在下文以与一位虚构读者进行问答的形式，简要讨论本文的要点。

免责声明：正如我们上文所强调的，我们的结果并未涵盖、也不可能涵盖所有可能的情形。我们根据自己的实证研究以及在整篇论文中建立的直觉来回答以下问题，但请注意，这些结论仅基于我们能够考察的有限（但细致入微的）情形。

• 问：那么，最佳的模型选择准则是什么？ 答：目前还没有什么灵丹妙药（真的？！）。

• 问：好吧。那么有哪些不错的候选准则？ 答：这似乎取决于你的数据，但总体而言，我们发现伪结果代理指标和事实选择在不同情境下表现良好。

• 问：那我来介绍一下我的数据。我有混杂数据，该怎么办？ 答：我们在实验中观察到，尤其是在处理效应简单且数据存在混杂的情况下，使用 R目标函数或双重稳健目标函数的伪结果准则，表现远优于其他准则。当处理效应是更复杂的函数时，事实值准则似乎表现更好。

• 问：如果我有来自临床试验的无混杂数据呢？ 答：当数据无混杂且模型设定正确时，我们发现选择准则的影响略小。只有在情境（B）中，即处理效应是比潜在结果简单得多的函数时，我们才观察到使用其他准则带来了一些改进——在这种情况下，随着 ρ 增加，插件式和伪结果代理指标的表现都更好。

• 问：我预计处理效应异质性不如处理与否情况下的结果异质性显著（预后信息远多于预测信息）。我该怎么办？ 答：伪结果代理指标，尤其是R学习器目标函数，在这种情形下似乎效果很好。

• 问：我预计情况正好相反——处理效应很可能是特征的一个非常复杂的函数——这对我意味着什么？ 在这种情况下，使用任何替代指标都可能弊大于利，带来更多噪声（尤其是数据集较小且存在混杂时）；此时或许更适合依靠事实验证。

• 问：我更愿意依靠事实值进行验证，因为我最信任这种方法，而且它不需要我估计任何额外参数。这样做会有什么损失？ 答：我们发现，在某些情形下，采用事实评估的效果可能更差，原因有二：它意味着你只能评估较少的一组估计量，而且它针对的是错误的目标——当潜在结果相对于条件平均处理效应非常复杂时，这一点似乎尤为重要。

• 问：你提到的这种“相容性偏差”是什么？答：我们认为，相容性偏差是指替代指标验证目标可能会偏向条件平均处理效应估计量$\hat{}\tau_k(x)$，这些估计量在结构上与所用替代指标相似，因为它们具有相似的归纳偏差。这个术语本身通常用于心理学文献，表示个体可能会系统性地偏好与当前信念一致的信息（Hart 等人，2009）。

• 问：我明白了！哪类选择准则受这种偏差影响最大？ 答：在我们的实验中，我们发现插件式替代指标选择标准似乎比伪结果选择标准更容易受到这种偏差的影响。

• 问：这些内容确实很有启发性，但你们没有涉及我感兴趣的模型选择的一个方面。对于如何设计自己的实证研究来探究这个问题，你们有什么建议吗？我们同意还有许多值得探讨的有趣问题，也希望未来能进一步研究！在第 4.2 节中，我们介绍了设计自己的实证研究时，如何进行有启发性的实验设计；此外，我们未来也会发布代码，便于轻松接入新的选择准则和数据集！

B. 其他文献综述

B.1. 进一步的CATE估计策略。

在正文中，我们聚焦于条件平均处理效应估计文献中两个主要研究方向之一，即依赖所谓元学习器的方向。使用任何机器学习预测方法即可轻松实现这类方法。该方向最初由K¨unzel 等人(2019年提出，之后由Nie&Wager(2021)；Kennedy (2020)；以及Curth & van der Schaar (2021b)进一步拓展。

然而，机器学习文献中近期提出的许多方法属于另一类文献流派，专注于将特定的机器学习方法应用于条件平均处理效应估计。这类文献几乎全部依赖间接估计策略，提出的模型旨在学习针对每种处理做出无偏结果预测（从而可以将条件平均处理效应估计为两者之差）。这一领域中非常流行的一种做法，最初由 Johansson 等人（2016）；Shalit 等人（2017）提出，之后又由 Johansson 等人（2018）；Hassanpour & Greiner（2019a;b）；Assaad 等人（2021）；Curth & van der Schaar（2021a;b）等加以扩展，即使用神经网络学习共享表示 $\Phi(x)$，随后通过两个处理特异性输出头 h_a 来估计 $\hat{\mu}_a(x) = h_a(\Phi(x))$。在这一框架中，由混杂引起的协变量偏移通常通过添加平衡正则化项来处理，该项惩罚处理组之间 $\Phi(x)$ 分布的差异（Shalit 等人，2017），并结合重要性加权（Johansson 等人，2018；Hassanpour & Greiner，2019a；Assaad 等人，2021）。其他研究则探讨了使用高斯过程（Alaa & van der Schaar，2018）、生成对抗网络（Yoon 等人，2018）、变分自编码器（Louizos 等人，2017；Wu & Fukumizu，2021）和深度核学习（Zhang 等人，2020）进行潜在结果估计。

另一方面，统计学和计量经济学文献除了采用正文讨论的元学习器策略外，主要依赖基于树的方法，其中最突出的是采用 S-学习器风格的贝叶斯加性回归树（Hill，2011）以及以因果树（Athey & Imbens，2016）和因果森林（Wager & Athey，2018；Athey et al.，2019；Hahn et al.，2020）形式出现的直接估计器。

B.2. 更多模型选择策略。

正文讨论的模型选择策略，即 Rolling & Yang（2014）、Nie & Wager（2021）、Saito & Yasui（2020）、Alaa 和 Van Der Schaar（2019）、Schuler 等人（2018）以及 Mahajan 等人（2022）提出和/或研究的策略，均使用给定的观察性数据直接估计模型性能，只插补部分干扰参数或替代目标。另有一支文献（Schuler 等人，2017；Athey 等人，2021；Parikh 等人，2022）提出，通过从手头的观察性数据集中学习生成模型，并用它模拟多个与目标数据集具有某些共同特征、但处理效应已知的测试数据集，来验证因果推断模型；这些已知处理效应可用于比较候选估计器的处理效应估计值。随后，应在真实数据上使用在这类生成数据集上表现最佳的估计量（Schuler 等人，2017；Athey 等人，2021；Parikh 等人，2022）。最后，在处理效应估计领域还存在其他有趣的模型选择问题，例如 Cui & Tchetgen（2019）所探讨的，如何为多阶段处理效应估计器选择最佳的第一阶段干扰估计器。

C. 实验详情

C.1. 实现细节

所有代码均采用 Python 编写，并遵循 scikit-learn 风格，以提高模块化程度并便于重复使用。所有代码均可在 `https://github.com/AliciaCurth/CATESelection` 以及 vanderschaar-lab 仓库中获取 `https://github.com/vanderschaarlab`。

底层机器学习方法

- 对于带有岭（L2）惩罚的线性回归（逻辑回归），我们使用 `RidgeCV`（其实现位于 scikit-learn(Buitinck 等人，2013）；这样，每当逻辑回归被用作任何元学习器的子程序（用于估计 $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4\}$ 时，我们都可以自动遍历岭惩罚系数$\mu(x)$，$\mu_0(x)$，$\mu_1(x)$，或 $\tau(x)$）。
- 对于极端梯度提升树（梯度提升），我们使用 `XGBRegressor`，其实现位于`xgboost` Python 包中（Chen 和 Guestrin，2016）。每当梯度提升被用作任何元学习器的子程序时，我们都会使用 scikit-learn的 5 折 `GridSearchCV`，对以下参数的所有组合进行搜索：学习率 $\in \{.1, .3\}$、最大深度 - -

$\in \{1,3,6\}$ 和估计量个数 $\in \{20,100\}$，这些超参数经观察会影响所有学习器。

• 最后，对于曾用于估计倾向得分的逻辑回归 $\pi(x)$，我们使用 scikit-learn 的实现 LogisticRegressionCV，以便在每次使用时遍历正则化参数 $\lambda \in \{10^{-5}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$。

请注意，我们会在所有子程序 $\mathcal{M}$ 中重新优化超参数，这意味着对于每个实验的每个种子中的每个元学习器，都会重新选择这些超参数。

元学习器 给定一个 按照上述讨论实现机器学习方法 $\mathcal{M}$后，我们按如下方式训练不同的元学习器 s:

• S-学习器：将 A 附加到 X，得到 $X' = (X,\ A)$。调用 $\mathcal{M}$.fit($X',\ Y$）。
• 扩展 S-学习器（ES-学习器）：将 A 和 $A * X$ 附加到 X，得到 $X' = (X,\ A * X,\ A)$。调用 $\mathcal{M}$.fit($X',\ Y$）。
• T-学习器：根据处理指示变量拆分数据。对每个处理组调用 $\mathcal{M}$.fit($X[A == a],\ Y[A == a]$)$a \in \{0,1\}$

• DR-学习器：调用 T-学习器以获得 $\mu_a(x)$的估计值。调用 $\mathcal{M}$.`fit`($X,\ A$)来拟合倾向得分估计。使用这些估计值计算伪结果$Y_{DR,\ \tilde{\eta}}$，具体方法见正文。调用 $\mathcal{M}$.`fit`($X,\ Y_{DR,\ \tilde{\eta}}$)。（注：我们还按照Kennedy (2020的建议测试了使用 5 折交叉拟合以确保一致性，但发现这并未提升性能。）

•R-学习器：按照 DR-学习器的方法拟合倾向得分估计。调用$\mathcal{M}$.fit(X, A)拟合无条件均值估计 $\mu(x)$。计算伪结果 $Y_{R,\tilde{\eta}} = {}^{Y_i-\tilde{\mu}}(X_i)$ ($_{A_i-\tilde{\pi}}(X_i)$ 和权重 $\beta_i = (A_i - \tilde{\pi}(X_i))^2$。~~调用$\mathcal{M}$.fit~~($X, Y_{R,\tilde{\eta}}$, sample weight $= \beta$)。（注意：我们还按照 Nie & Wager (2021) 的建议测试了 5 折交叉拟合，以确保一致性，但发现这并未提升性能。）

选择准则 所有选择准则均仅通过考察验证数据来计算。对于插件代理准则，采用上文讨论的策略，在验证数据上计算处理效应的插件估计值。

对于伪结果替代准则，我们进行 5 折交叉拟合，以避免干扰参数估计与结果之间的相关性；也就是说，我们将验证数据分成 5 折，仅使用某个数据点未被分入的其余 4 折来插补其伪结果。

虽然正文中未作介绍，但在附录 D.1 中，我们还使用 PW 伪结果计算了伪结果代理指标 $Y_{PW,\ \tilde{\eta}} = \left({}_{\pi}{}^{A}{}_{\hat{}}(X) \ - {}_{\hat{\pi}(X)}{}^{(1-A)1-} \right) Y$，以及 Rolling & Yang (2014) 提出的匹配伪结果，该结果通过寻找欧几里得距离最近的邻居计算得出。我们还计算了 Alaa 和 Van Der Schaar (2019) 的影响函数验证标准，该标准通过选择使以下表达式最小化的 $\hat{}\tau_k(x)$来实现：

$$Y_{IF} = (1-B)\tilde{\tau}(x)^2 + BY(\tilde{\tau}(x) - \hat{\tau}(x)) - D * (\tilde{\tau}(x) - \hat{\tau}_k(x))^2 + \hat{\tau}_k(x)^2 \tag{2}$$

其中 $D = A - \tilde{\pi}(x)$、 $B = 2AC^{-1}$ 和 $C = \tilde{\pi}(1-\tilde{\pi})$ 以及 $\tilde{\tau}(x)$ 是 T-学习器估计值。所有干扰参数均通过 5 折交叉估计进行估计。

C.2. 数据生成过程（DGP）

我们的实验基于（Curth& van derSchaar，2021a）中使用的数据生成过程。主要区别在于：a）我们对数据进行随机二值化，以考察模型设定错误的影响；b）考虑高阶交互作用，使条件平均处理效应与潜在结果之间的差异更加显著；c）同时引入混杂。

我们还使用了协作围产期项目提供的协变量数据[8]，这些数据用于首届大西洋因果推断竞赛（ACIC2016）（Dorie 等人，2019年，并按照竞赛所用的转换方法处理所有协变量[9]。原始数据集包含 $d = 58$ 个协变量，其中我们排除了3 个分类变量。在剩余的55个协变量中，5 个为二元变量，27 个为计数数据，23 个为连续变量。由于我们发现现有的二元变量和计数数据非常稀疏，因此决定改为对变量进行随机二值化：在每列中随机选取一个观测值，并仅保留23 个连续列，从而得到一个新的输入数据集 X^*，用于创建一个模拟决策树的数据生成过程。（请注意，虽然其他列未用于结果模拟，但它们仍然是 X^{input} 提供给估计量和选择准则的数据的一部分，因此所有估计量还必须学会区分有信息列和无信息列。）

[8]可从https://jenniferhill7.wixsite.com/acic-2016/competition[9]我们使用以下代码：
https://github.com/vdorie/aciccomp/blob/master/2016/R/transformInput.R

与.....类似当前 th & van der Schaar (2021a)，随后我们在带有交互项的线性模型中使用输入数据 s:

$$Y_i = c + \sum_{j=1}^{d} \beta_j X_j^* + \sum_{j,l}^{d} \beta_{j,l} X_j^* X_l^* + \sum_{j,l,k}^{d} \beta_{j,l,k} X_j^* X_l^* X_k^* + \sum_{j,l,k}^{d} \beta_{j,l,k,m} X_j^* X_l^* X_k^* X_m^* + A_i \sum_{j=1}^{d} \gamma_j X_j + \epsilon_i \quad (3)$$

其中 $\epsilon_i \sim N(0,\ 1)$、$\beta_j \sim \mathcal{B}(.3)$ 和 $\gamma_j \sim \mathcal{B}(\rho)$。我们将每个变量随机纳入一个一阶、二阶和三阶交互项中，然后模拟系数 $\beta. \sim \mathcal{B}(.2)$。我们将每个系数设为二元变量，以避免不同模拟运行中潜在结果和条件平均处理效应的尺度出现较大方差，从而使均方根误差在各次运行间保持可比。

接下来，我们通过依据倾向得分 $\pi(x) = expit(\xi \mathcal{Z}(X^{input}\beta))$ 来分配处理，从而模拟混杂；其中$\mathcal{Z}(\cdot)$ 表示在模拟数据中进行标准化，而 β 是公式 3中的线性系数，确保所有变量都是真实混杂变量。还需注意，X^{input} ——即估计量和选择器所观察到的数据——会进入倾向得分，因此逻辑回归在估计倾向得分时始终是正确设定的。我们在附录D.4中进一步探讨了其他设置。

因此，我们考察的三个主要实验因素是条件平均处理效应复杂度 ρ、混杂强度 ξ，以及估计量和选择器获取输入数据的方式 X 与 X^*。在实验中，我们始终改变 $\rho \in \{0, .1, .3\}$ 并定义如下设置:

- A：连续输入数据 $X^{input} = X$，无混杂 $\xi = 0$
- B：二值化输入数据 $X^{input} = X^*$，无混杂 $\xi = 0$
- C：连续输入数据 $X^{input} = X$，无混杂 $\xi = 3$
- D：二值化输入数据 $X^{input} = X^*$，无混杂 $\xi = 3$

D. 使用主要数据生成过程的附加结果

D.1. 其他选择准则

在图 6 中，我们还展示了匹配性能、另外两种伪结果（Curth& van derSchaar (2021b）中讨论的 PW 伪结果和 RA 伪结果）、影响函数（IF）以及加权事实验证（wFactual）的表现，这些内容未在正文中呈现。为便于阅读，且因为它们并未在事实选择方面带来任何改进（事实上，它们的表现通常更差），我们没有在正文中讨论这些内容。

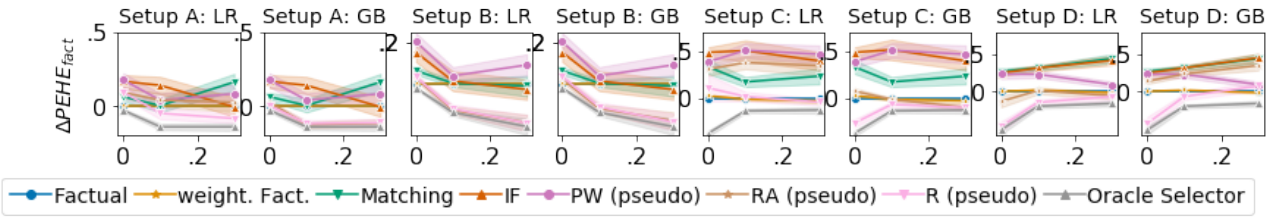
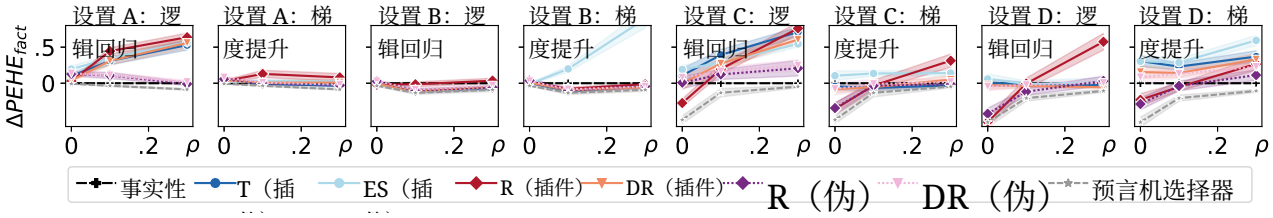
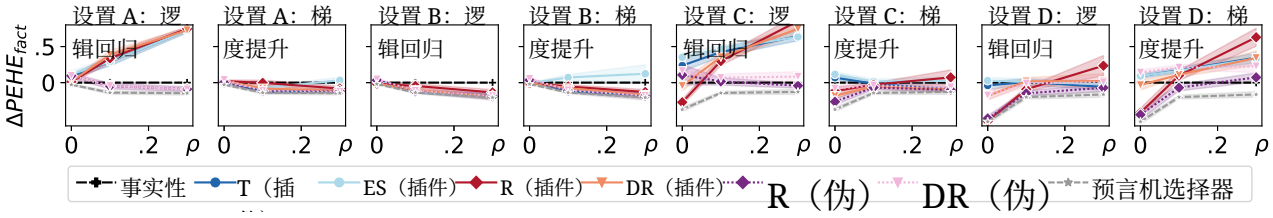


图 6。 ΔPEHE$_{fact}$：对于不同的选择准则，在 4 种不同设置下分别使用线性回归（LR）和极端梯度提升（GB）实现时，任一给定选择准则的测试 PEHE 与事实选择之间的差异（数值越低越好，负值表示优于事实选择）；其中包括正文未展示的其他准则：RA- 和 PW 伪结果、匹配、影响函数（IF）以及加权事实验证（wFactual）。此处，$\tau(x)$ 的复杂度在 ρ 中增加。阴影区域表示一个标准误（SE）。

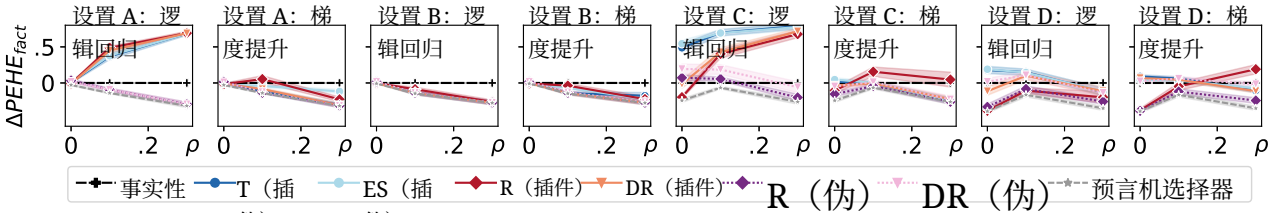
D.2. 额外样本量



(a) $n = 500 + 250$



(b) $n = 1000 + 5000$ （改编自正文）



(c) $n = 2000 + 1000$

图7。**不同样本量下各选择准则的相对表现**。绘制 ΔPEHE_{fact}，即事实选择与任一给定选择准则所选模型的测试 PEHE 之差（越低越好，负值表示优于事实选择）；所用方法为线性回归（LR）和极端梯度提升（梯度提升）。

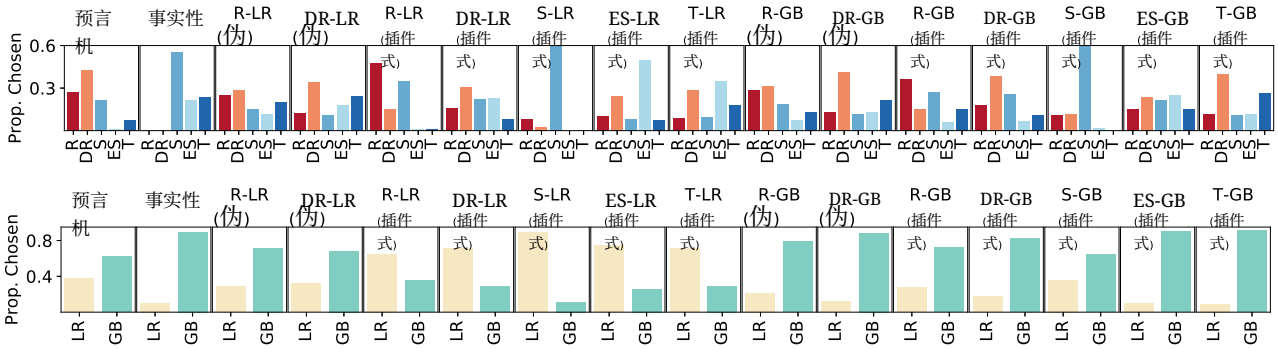


图8。**考察选择准则与估计量之间是否存在相容性偏差：估计策略（上）和机器学习方法（下）**。衡量在正文所考虑的所有设置中，选择某一特定估计量的比例。请注意，与图 4 不同，本图衡量的是任意估计量被选中的绝对比例（而非像正文那样，衡量出现错误时选中某种估计器类型的频率）。

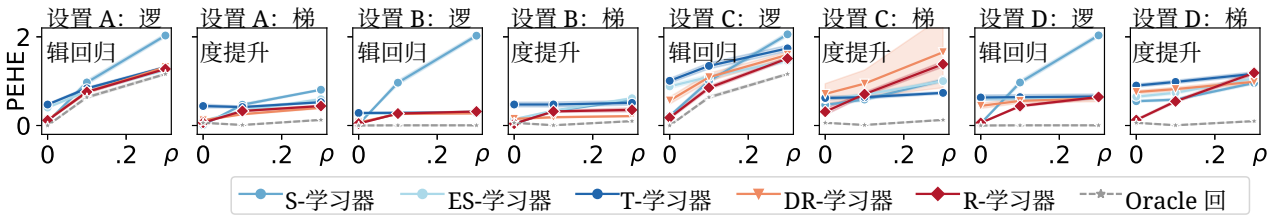
D.3. 更多相容性图

在图 8中，我们展示了任一选择准则选择任一估计量类型的绝对 次数——这与正文中的图 4 不同，后者关注的是选择器出错时所选估计量的类型。只考察选择器所犯的错误时，相合性会明显得多，但其中一些相合性

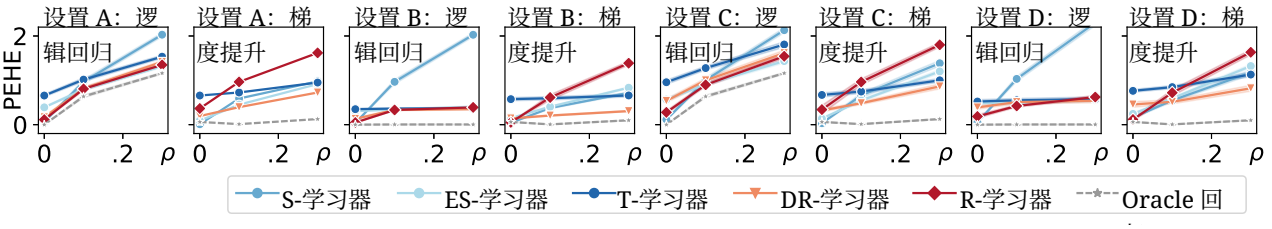
正文中讨论的模式也清晰地反映在图8中。

D.4. 采用其他倾向得分的其他设置

不平衡。 在图9和10 中，我们通过重新缩放处理分配倾向得分，研究加入不平衡处理组规模的影响。不同于正文中所考虑的平衡边际处理倾向得分 $\pi = 0.5$，现在我们以边际倾向概率 $\pi = .2$ 分配处理，使得对照组单位显著多于处理组单位。在图9中，就基础学习器而言，我们观察到最显著的差异是，在效应异质性较大时，R-学习器现在表现相对较差。在图10中，我们发现，在不平衡设置下，在混杂设置中，当 ρ较小时，相对于事实选择的改进幅度要小得多；而在平衡设置下，这正是性能提升最大的情形。



(a) $\pi = 0.5$（正文中复现）

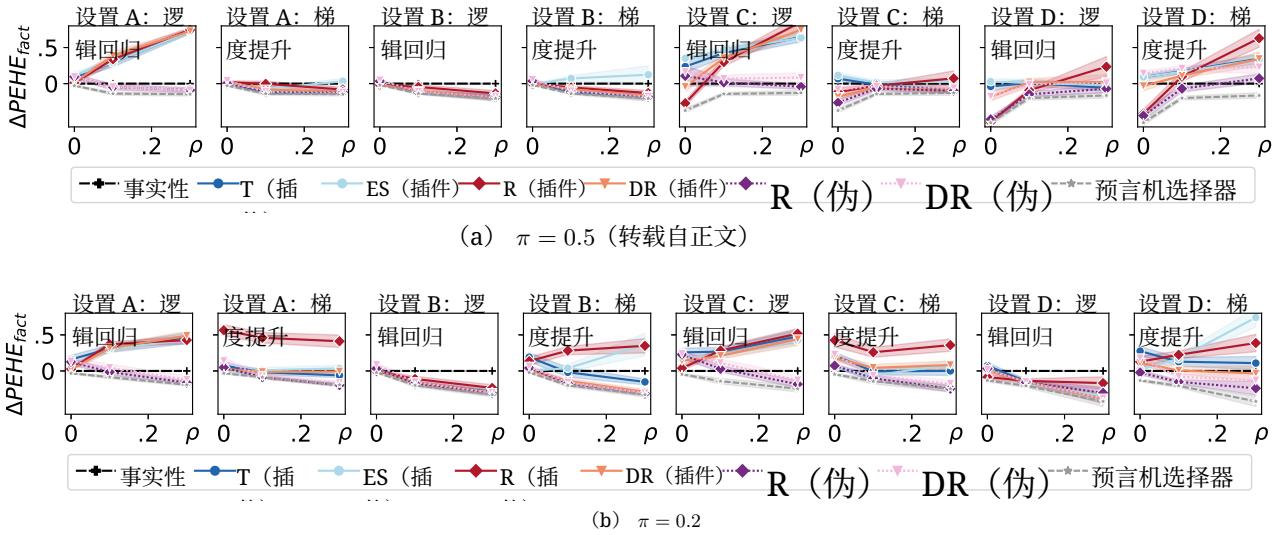


(b) $\pi = 0.2$

图 9。 **不同候选估计器的条件平均处理效应估计误差（精确个体处理效应误差）$\mathcal{T}$，其中不考虑处理组不平衡（上图）和考虑处理组不平衡（下图）**。所有学习器均使用线性回归（逻辑回归）和极端梯度提升（梯度提升）实现，并在 4 种不同设置下进行考察，其中 $\tau(x)$ 的复杂度随 ρ 增加。阴影区域表示一个标准误。

其他倾向得分设定。在图11和图12中，我们进一步考察了所使用的倾向得分设定的变化。具体而言，请注意，如附录C所述，我们通过对 X^{input} 进行逻辑回归，确保倾向得分始终被正确设定。作为副作用，这意味着在正文所考虑的设置 C 中，π 是 X 的函数，而 μ 是 X^* 的函数；在设置 D 中，π 和 μ 都是二值化（观测到的）X^* 的函数，因此二者更加一致。因此，我们在此进一步考虑两种设置，其中 π 是未观测版本的协变量的函数（即设置 C* 中的 X^*，其估计量表示为 X；以及设置 D* 中的 X，其估计量表示为X^*）。这可能会以两种方式影响结果：一方面，在新的设置 C* 和 D* 中，倾向得分估计量存在模型设定错误，这可能会对依赖这些估计量和选择器的估计量产生不利影响。另一方面，请注意，在旧设置 D 和新设置 C* 中，π 和 μ 依赖于协变量 $X*$ 的相同变换；而在旧设置 C 和新设置 D* 中，π 和 μ 并不依赖于协变量的相同变换 —— 在 D 和 C* 中，倾向得分与结果更加一致，这通常会使估计更加困难。

在图 11和 12中，我们观察到第二种效应似乎强于第一种：尤其是在使用错设模型（设置 C 和 C* 中的逻辑回归模型）时，对于候选估计器和选择准则，我们都观察到，使 π和 μ 更加对齐（即从设置 C 转向 C*）会降低它们的性能——无论是否确实需要估计倾向得分（例如，不包含倾向得分估计的 S 学习器和 T 学习器，性能也会下降）。



（a） $\pi = 0.5$（转载自正文）

（b） $\pi = 0.2$

图 *10*。**在没有处理组不平衡（上）和存在处理组不平衡（下）的情况下，不同选择准则的相对表现**。绘制 Δ精确个体处理效应误差$_{fact}$，即事实选择的测试 PEHE 与采用任一给定选择准则所选模型的测试 PEHE 之差（越低越好，负值表示优于事实选择）。模型分别使用线性回归（逻辑回归，LR）和极端梯度提升（梯度提升，GB）实现。每种准则都可以使用 $\mathcal{T}$，即由 10 个候选估计器构成的完整集合，其表现如上文图9 所示。我们考虑 4 种不同设置，其中 $\tau(x)$ 的复杂度随 ρ增加。阴影区域表示一个标准误。

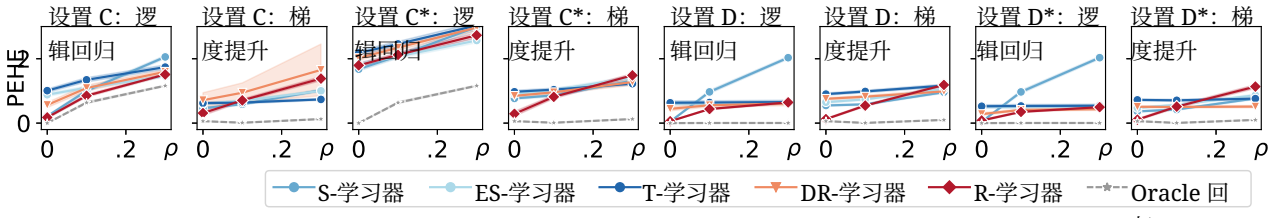


图 *11*。**在 $\mathcal{T}$ 中，不同候选估计器在采用不同倾向得分设定的混杂设置下进行条件平均处理效应估计时的误差（精确个体处理效应误差）**。

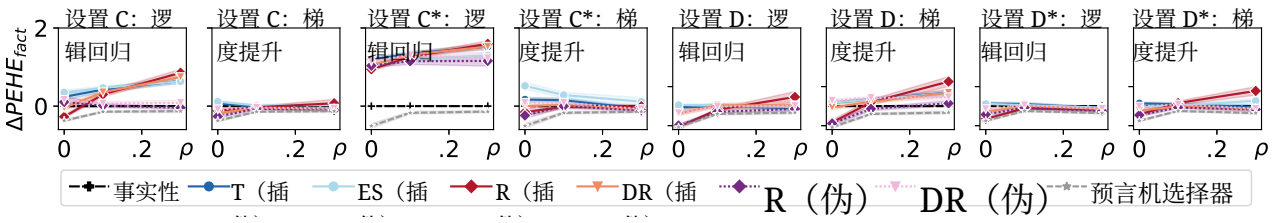


图 *12*。**采用不同倾向得分设定的混杂设置下，不同选择准则的相对表现**。

E. 其他数据集：IHDP 和 ACIC2016

为求完整，我们重复了 Curth 等人 (2021) 所进行的实验。他们指出，当前条件平均处理效应估计文献中使用 ACIC2016 和 IHDP 数据集存在一些问题，尤其是忽视了这些数据集潜在的结构特征（即它们所偏向的实验参数设置）。有鉴于此，我们使用 Curth 等人 (2021) 讨论的设置，重复了正文中的相同实验；详情请参阅他们的论文。

尤其需要注意，　　由 Hill （2011）构建的原始 IHDP 数据集具有非常复杂的处理效应函数　　n,

可能使间接学习器占优。 Curth 等人 （2021）构建了一个修改版本，其中 POs 仍然很复杂，但处理效应只是线性的，因此现在可能使直接学习器占优。此外，Curth 等人 （2021）研究了 ACIC2016 竞赛的 77 个设置中的设置 3（编号为 2、26 和 7）（Dorie 等人，2019年）：这些设置的唯一区别在于“条件平均处理效应复杂度”，其中设置 2 不存在异质性，设置 26 存在一定异质性，而设置 7 具有非常强的异质性——因此对相对性能有着相似的预期。IHDP 数据集本身已经非常小，总样本量为 747 ，因此我们保留原始的训练集—测试集划分，并像之前一样将训练数据划分为 $2/3$ 用于 $1/3$ 训练和验证。ACIC2016 数据集使用与我们的模拟相同的底层协变量，因此我们考虑相同的三种样本量设置。最后需要注意的是，Curth 等人 （2021）观察到，即使采用相同的估计策略，不同设置下的绝对均方根误差也存在非常高的方差。因此，为稳定结果，我们报告相对均方根误差（$RMSE$（$method$）$/RMSE$（$baseline$））， 其中所有选择准则的基准仍为事实选择，而 IHDP 上学习器的基准为 T-LR，ACIC2016 上的基准为 T-GB（分别是表现最佳的 T-学习器）。

比较基础学习器的性能。我们在表 E 中报告基础学习器的相对性能。我们观察到，在两个数据集上，基础学习器的相对性能均符合预期，并且与Curth 等人(2021) 的结果相似：在条件平均处理效应较简单的数据生成过程中，直接学习器具有优势；而在条件平均处理效应复杂的设置中，T-学习器则更具优势。我们还观察到，即使条件平均处理效应复杂度较低，R-GB 学习器的表现通常也比正文中的结果差，这可能是所有数据集中处理组规模不平衡所致。

表 *1*。IHDP 和 ACIC 设置下基础学习器的相对精确个体处理效应误差。结果分别对 IHDP 的全部 100 次模拟和 ACIC 的每种设置各 10 次模拟取平均。

设置	预言机-梯度提升	预言机-逻辑回归	S-GB	S-LR	ES-GB	ES-LR	T-GB	T-LR	DR-GB	DR-LR	R-GB	R-LR
原始 IHDP	0.58	0.78	1.58	2.58	1.54	1.58	1.14	1.00	1.25	1.11	2.60	1.65
修改版 IHDP	0.05	0.00	0.85	0.46	1.07	1.24	1.35	1.00	0.94	0.67	4.46	1.54
ACIC 2 , n=750	0.00	0.00	0.31	0.32	0.52	0.92	1.00	1.10	0.64	0.46	1.06	0.31
ACIC 2 , n=1500	0.00	0.00	0.37	0.15	0.53	0.84	1.00	1.08	0.67	0.40	0.91	0.29
ACIC 2 , n=3000	0.00	0.00	0.50	0.20	0.69	0.93	1.00	1.04	0.49	0.47	1.33	0.31
ACIC 26 , n=750	0.59	1.07	1.12	1.61	1.04	1.26	1.00	1.28	0.94	1.26	1.24	1.27
ACIC 26 , n=1500	0.58	1.32	1.20	2.03	1.08	1.57	1.00	1.47	0.91	1.46	1.22	1.47
ACIC 26 , n=3000	0.58	1.55	1.20	2.41	1.03	1.74	1.00	1.71	0.95	1.70	1.40	1.72
ACIC 7 , n=750	1.10	1.27	1.15	1.71	1.08	1.47	1.00	1.44	0.95	1.46	1.37	1.44
ACIC 7 , n=1500	0.68	1.56	1.23	2.20	1.10	1.75	1.00	1.71	1.01	1.77	1.43	1.74
ACIC 7 , n=3000	0.64	1.74	1.25	2.54	1.09	1.93	1.00	1.90	0.99	1.91	1.57	1.91

比较选择器性能。 我们在表 E 中报告选择器的相对性能。IHDP 上的相对选择器性能基本符合预期，但在修改后的设置中，伪 R 准则和 DR 准则的表现不如预期，而插件准则的表现更好。很难确定其原因，因为 IHDP 数据集也具有以下特点：i）对照组人群远大于处理组人群；ii）重叠有限；iii）样本量非常小。ACIC 数据集上的结果也不尽相同；在这里，我们观察到，事实选择的改进主要出现在规模最小的数据集上，以及条件平均处理效应较简单的情形（设置 2）。这可能部分是因为没有任何方法能够很好地拟合高异质性的数据生成过程，因为在设置 26 和 7 中，预言机选择器性能仅略好于事实选择。请注意，这些 ACIC 模拟也存在重叠有限以及处理组和对照组不平衡的问题。我们希望这番讨论能说明为什么我们认为有必要构建自己的数据生成过程：由于这些因素在现有数据集中相互交织，因此极难利用这些数据来区分不同因素对性能的影响。

表 2. 模型选择指标的精确个体处理效应误差 相对于 IHDP 和 ACIC 设置中事实选择的表现。对 IHDP 的 100 次模拟以及 ACIC 的每种设置各 10 次模拟取平均。

在 IHDP 的 100 次模拟中取平均，以及每种设置各 10 次模拟

设定	预言机	事实性	S-GB	S-LR	ES-GB	ES-LR	T-GB	T-LR	PlugDR GB	PlugDR LR	PlugR GB	PlugR LR	PseuDR GB	PseuDR LR	PseuR GB	PseuR LR
原始 IHDP	0.91	1.00	1.53	2.57	1.53	1.43	1.02	1.02	1.12	1.10	1.81	1.63	1.03	1.03	1.10	1.11
修改版 IHDP	0.58	1.00	0.91	0.71	0.97	1.36	0.85	0.79	0.77	0.76	3.32	1.91	1.07	1.01	1.36	1.38
ACIC 2, n=750	0.45	1.00	2.38	6.86	2.38	5.10	3.11	3.99	3.20	0.78	5.57	2.45	3.50	5.80	4.50	4.56
ACIC 2, n=1500	0.35	1.00	0.54	1.13	0.98	0.90	2.27	1.83	2.08	1.29	2.31	0.51	2.70	2.80	2.16	2.74
ACIC 2, n=3000	0.35	1.00	0.62	0.39	1.12	1.57	4.16	1.79	2.96	1.04	3.42	1.55	4.83	4.94	4.48	4.21
ACIC 26, n=750	0.91	1.00	1.20	1.65	1.12	1.27	0.95	1.23	0.97	1.26	1.11	1.31	1.00	0.96	0.95	0.96
ACIC 26, n=1500	0.88	1.00	1.25	2.00	1.16	1.48	0.90	1.41	0.93	1.46	1.30	1.46	0.98	1.00	0.94	1.00
ACIC 26, n=3000	0.86	1.00	1.25	2.46	1.08	1.79	1.00	1.78	1.01	1.78	1.27	1.77	1.06	1.08	1.12	1.09
ACIC 7, n=750	0.85	1.00	1.34	1.65	1.12	1.38	0.90	1.35	0.90	1.32	1.21	1.46	0.90	0.90	0.88	0.91
ACIC 7, n=1500	0.94	1.00	1.24	2.13	1.20	1.72	1.00	1.66	1.03	1.66	1.35	1.78	0.99	1.04	1.08	1.07
ACIC 7, n=3000	0.98	1.00	1.23	2.53	1.12	1.91	0.99	1.89	1.05	1.92	1.25	1.99	0.98	1.04	1.08	1.04