arXiv:2302.02859v1 [stat.ME] 6 Feb 2023

一种用于大数据因果推断的快速自助法算法

Matthew Kosko，乔治·华盛顿大学统计学系，华盛顿特区 20052；Lin Wang，普渡大学统计学系，印第安纳州西拉斐特 47907；以及 Michele Santacatterina，纽约大学人口健康系，纽约州纽约市 10016

2023年2月7日

摘要

从大型实验数据和观察性数据中估计因果效应，在工业界和研究领域都日益普遍。自助法是一种直观且强大的技术，用于构建估计量的标准误和置信区间。然而，在涉及大数据的情境下，应用自助法可能需要高昂的计算成本。此外，基于机器学习和优化技术的现代因果推断估计量进一步加重了自助法的计算负担。小样本自助法已被提出用于非因果的大数据情境，但尚未应用于评估因果效应估计量的性质。本文针对大数据因果推断，介绍一种名为因果小样本自助法的新型自助法算法。该新型

算法显著提高了传统自助法的计算效率，同时提供一致估计和理想的置信区间覆盖率。我们描述了其性质，提出实践考量，并在一项模拟研究中，从偏倚、真实95%置信区间的覆盖率以及计算时间等方面评估了所提出算法的性能。我们利用妇女健康倡议的大型观察性数据集，评估激素治疗对冠心病平均发病时间的影响。

关键词：因果自助法；真实世界数据；倾向得分；协变量平衡；机器学习

1 引言

近年来，用于推断因果关系的大规模实验数据集和观察性数据集在工业界（Kohavi 等，2009年；Tang 等，2010年；Peysakhovich 和 Lada，2016年；Kohavi 等，2013年）和医学研究领域（Dagan 等（2021）；Mohammadi 等，2021年；Sud 等，2020年）都日益普遍。例如，Tang 等（2010年）利用大规模 A/B 测试评估了用户可见的变化和机器学习算法；Dagan 等（2021）则在一个超过一百万名参与者的人群中评估了新冠 mRNA 疫苗的有效性。

估计因果效应的方法多种多样。基于倾向得分的技术被广泛采用，例如逆概率加权（IPW）估计量（Lunceford 与 Davidian（2004））以及协变量平衡倾向得分（CBPS）（Imai 与 Ratkovic 2014）。IPW 权重由倾向得分的倒数构成（Rosenbaum 与 Rubin 1983），倾向得分可通过参数和非参数机器学习技术进行估计，例如支持向量机（SVM）（Westreich 等，2010）。另一种方法是，CBPS 通过广义矩估计法找到能够平衡协变量的逻辑模型（Imai 与 Ratkovic 2014）。

自助法（Efron 1979年，Efron与Tibshirani 1994）已用于获取统计估计量精度的估计值，包括标准误和置信区间。随着人们能够更多地使用现代计算能力，自助法日益普及；然而，在涉及大型数据集的情形下，应用自助法仍可能需要高得难以承受的计算开销。

已有研究提出了一些方法来提高自助法的计算效率。Politis 等人（1999年）和 Bickel 等人（2012年）分别提出了子抽样和与之密切相关的m out of n 自助法。这些方法在较小的

子样本上获得自助法估计，从而改善了标准自助法在计算方面的局限性。不过，这些方法本身也有缺点（Samworth 2003年，Kleiner 等人（2014年））。具体而言，与自助法相比，它们的有限样本表现较差，而且对所选子样本大小十分敏感。为克服这些挑战，Kleiner 等人（2014年）提出了小样本自助法（BLB）。该方法不再将估计量应用于较小的子样本，而是在数据的多个子集（即“样本袋”）上实施自助法。最后，BLB 抽取的自助法样本与完整数据集大小相同。尽管 BLB 提高了自助法的可扩展性，但它尚未扩展到为因果 效应估计量提供精度估计。

本文介绍了一种新的自助法算法，称为小规模自助法袋（因果 BLB），用于大数据因果推断。我们的研究受到两方面的推动：越来越多的大数据集被用于推断因果关系，以及机器学习和协变量平衡技术被用于估计因果效应。在接下来的章节中，我们概述现有关于将自助法应用于因果推断的研究，并介绍本文对这一领域的贡献。我们将在第 2 节介绍所提出的方法，并讨论其性质。第 3 节将探讨实践考量；第 4 节将根据计算时间、偏倚以及 95%置信区间的覆盖率评估该方法的性能。我们将利用妇女健康倡议研究中的一个大型观察性数据集，应用所提出的方法评估激素治疗对冠心病发生时间的影响。第 6 节将给出结论。

1.1 相关工作

尽管已有大量文献探讨将自助法应用于各种统计问题（Wu等人 2022年、Zhu等人 2020年等），但关于

因果推断中自助法的文献相对较少，尤其是针对大数据因果推断的自助法研究。因果问题中采用自助法的常见方法是从数据中有放回地抽取样本，并对每个样本进行“重新设计”，以确保协变量平衡（Zhang & Sabbaghi 2021年，Dagan等人 2021年）。

关于将自助法用于匹配估计量的文献更为丰富。Abadie & Imbens（2008）以及 Abadie & Spiess（2022）都研究了匹配估计量背景下的自助法。Abadie & Imbens（2008）表明，标准自助法无法为匹配估计量提供有效推断，因此需要进行修改。在此基础上，Abadie & Spiess（2022）考察了匹配后如何为回归系数（包括处理效应系数）构造有效的标准误。作者提出了非参数自助法的分块版本，该方法对匹配组而非单个观测进行重新抽样；他们证明，这一程序能够产生有效推断。Otsu & Rai（2017）同样解决了 Abadie & Imbens（2008）发现的标准自助法用于匹配估计量时所存在的问题；作者构造了一种加权自助法程序，在计算自助法估计量时不重新计算某个观测被使用的次数，而是将其作为观测的一部分进行重新抽样。Adusumilli（2018）提出了一种依赖潜在误差概念的修正自助法。Zhao 等人（2019）将自助法应用于敏感性分析，并推导出了敏感性分析的置信区间。

已有一些研究评估自助法在因果推断中的表现。Austin（2016年）通过一项模拟研究，使用加权 Cox 模型评估了自助法在估计边际风险比方差方面的表现。在另一项模拟研究中，Austin 和 Small（2014年）展示了在使用无放回倾向得分匹配估计平均处理效应时自助法的表现。

Imbens 近期在因果推断自助法方面开展了严谨的研究

& Menzel（2021年）。他们提出了一种因果自助法算法，用于处理观察性数据和随机化数据中的因果推断，并证明了其具有良好的大样本性质。尤其是，他们表明自助法置信区间在渐近意义上是保守的。Imbens & Menzel（2021年）提出的程序，首先使用联结函数为处理组和对照组插补潜在结果，该函数将边际密度函数映射为联合密度。随后，程序通过从原始数据中抽样并模拟新的处理向量，模拟抽样分布和随机化分布。作者使用一种特定的联结函数，证明该程序能够产生保守的推断。对于观察性数据，插补步骤使用纳入平衡分数的加权经验累积分布函数，以确保处理与潜在结果相互独立。尽管该方法具有大样本方面的优势，但所需的插补步骤在处理大型数据集时计算开销很大。

1.2 我们的贡献

我们对这一领域文献的贡献在于提出了一种新颖的自助法算法，它在利用大数据一致地估计因果效应并获得理想的置信区间覆盖率的同时，缩短了传统自助法的计算时间。与Imbens & Menzel（2021年）不同，我们提出的方法无需插补，而是通过抽取加权自助抽样样本来实现。这样做显著提高了计算效率，同时保留了该方法直观的解释：反复抽取加权估计量。除此之外，我们还提供了该算法的 `R` 代码，网址为`https://github.com/mdk31/causalbootstrap`。

2 因果小袋自助法

我们考虑一项由 n 个单位组成的实验性或观察性研究，这些单位是独立同分布（iid）抽取的。根据潜在结果框架（Imbens & Rubin 2015年），对于每个单位 $i = 1, \ldots, n$，我们令 $Y_i(w) \in \mathbf{R}$ 表示处理 $w \in \{0, 1\}$的潜在结果。我们令 $X_i \in \mathcal{X}$ 表示观测到的混杂因素。我们将 W_i 设为接受处理 w的指示变量。本文重点在于估计平均处理效应（平均处理效应），其定义为

$$\tau \equiv E(Y(1) - Y(0)),$$

在假设一致性、无干扰和可忽略的处理分配的条件下，该量可以识别（Imbens & Rubin 2015, Hernan & Robins 2020, Rosenbaum & Rubin 1983）。

本文关注如何使用自助法计算标准误和置信区间。如前所述，随着 n 增加，自助法的计算负担会变得难以承受。现在我们介绍因果小样本自助法（因果 BLB），这是一种快速自助法算法，它将 Kleiner 等人（2014）提出的原始算法扩展到了因果推断。因果BLB算法首先从原始数据中抽取 s 个大小为b 的子集。随后，它针对每个子集，使用参数方法、机器学习或协变量平衡技术获得一组权重。然后，算法在每个子集内使用其权重进行 r次大小为 n 的自助法重抽样；对于从第 k个子集中抽取的第 j次自助法重抽样，使用以下加权估计量估计 τ：

$$\hat{\tau}_{j,k} = \sum_{i=1}^{b} \left(\frac{-1}{n_0}\right)^{W_{i,k}-1} \left(\frac{1}{n_1}\right)^{W_{i,k}} \tilde{M}_{i,j,k} Y_{i,k},$$

其中，n_0 和 n_1 分别是完整数据集中对照组和处理组个体的数量，$\tilde{M}_{j,\ k}$是由两个向量 M^1 和 M^0连接而成的向量，分别对应处理组和对照组

第 j 次自助法重复抽样中，分别按照归一化的基于倾向得分的权重进行多项分布抽样：

$$(\hat{w}^0(X_i), \hat{w}^1(X_i)) = \left(\frac{1/(1-\hat{\pi}(X_i))}{\sum_{i=1}^{n}(1-W_i)/(1-\hat{\pi}(X_i))}, \frac{1/\hat{\pi}(X_i)}{\sum_{i=1}^{n} W_i/\hat{\pi}(X_i)} \right),$$

其中，$\hat{\pi}(X_i)$ 是估计的倾向得分。这里，自助法对处理组和对照组中的数据都进行过度抽样，因此 M^1 和 M^0 中元素的总和分别为 n_1 和 n_0。由于任一自助法重抽样中都可能多次出现同一观测值，我们将自助法重新表述为对该子集中每个观测值进行整数倍复制，表示该观测值在自助法重抽样中出现的次数；这些整数服从多项分布，与 Præstgaard 与 Wellner（1993）中的做法相同。这确保了自助法重抽样的总样本量为 $n = n_0 + n_1$。

最后，因果 BLB 算法对各次重抽样以及各个子集中的 $\hat{\tau}_{j,\ k}$ 取平均，得到总体自助法估计，如下所示：

$$\begin{aligned}\hat{\tau} &= \frac{1}{s}\sum_{k=1}^{s}\frac{1}{r}\sum_{j=1}^{r}\hat{\tau}_{j,k} \\ &= \frac{1}{s}\sum_{k=1}^{s}\frac{1}{r}\sum_{j=1}^{r}\sum_{i=1}^{b}\left(\frac{-1}{n_0}\right)^{W_{i,k}-1}\left(\frac{1}{n_1}\right)^{W_{i,k}}\tilde{M}_{i,j,k}Y_{i,k}.\end{aligned}$$

通过这种做法，因果 BLB 算法沿用了在因果问题中应用自助法的传统方式，即从数据中抽取加权样本，从而确保协变量平衡，与 Zhang 与 Sabbaghi（2021）以及 Dagan 等（2021）的做法一致。因果 BLB 算法总结于算法1和图1中。

2.1 性质

在本章节中，我们讨论因果 BLB 的一些性质。具体而言，我们首先证明因果 BLB 能够提供平均处理效应的一致估计值，τ。随后，我们讨论其与.....的联系

算法1 因果 BLB

要求： 子集数量 s、子集大小 b，以及自助法重复次数， r

计算 $n_1 = \sum_{i=1} W_i$ 和 $n_0 = n - n_1$

对于 $k \leftarrow 1$ 至 s，**执行**

从 $\mathcal{I} = \{1, 2, \ldots, n\}$ 中无放回抽样，抽取一组索引 $I_k = \{i_1, \ldots, i_b\}$

对于子集数据 $\mathbf{O}_k = (W_{i_1}, \ldots, W_{i_b}, X_{i_1}, \ldots, X_{i_b})$，构建倾向得分模型 $\hat{}\pi_{i,k}(X_i)$，并平衡数据

重新排列观测值，使前 $b_k^0 = \sum_{l=1}^{b}(1 - W_{i_l})$ 个为对照组，最后 $b - b_k^0$ 个为处理组

为子集数据构造归一化倾向得分逆权重

$(\hat{w}_{i,k}^0(X_i), \hat{w}_{i,k}^1(X_i)) = \left(\frac{1/(1-\hat{\pi}_{i,k}(X_i))}{\sum_{i=1}^{b}(1-W_i)/(1-\hat{\pi}_{i,k}(X_i))}, \frac{1/\hat{\pi}_{i,k}(X_i)}{\sum_{i=1}^{b} W_i/\hat{\pi}_{i,k}(X_i)}\right)$

对于 $j \leftarrow 1$ 至 r，**执行**

Sample $\left(M_{1,j,k}^1, \ldots, M_{b_k^1,j,k}^1\right) \sim \text{Multinomial}(n_1, \hat{\mathbf{w}}_k^1)$

Sample $\left(M_{1,j,k}^0, \ldots, M_{b_k^0,j,k}^0\right) \sim \text{Multinomial}(n_0, \hat{\mathbf{w}}_k^0)$

Compute $\tilde{M}_{j,k} = (M_{1,j,k}^0, \ldots, M_{b_k^0,j,k}^0, M_{1,j,k}^1, \ldots, M_{b_k^1,j,k}^1)$

Calculate $\hat{\tau}_{j,k} = \sum_{i=1}^{b} \left(\frac{-1}{n_0}\right)^{W_{i,k}-1} \left(\frac{1}{n_1}\right)^{W_{i,k}} \tilde{M}_{i,j,k} Y_{i,k}$

否则f

$\hat{\tau}_k \leftarrow \frac{1}{r}\sum_{j=1}^{r} \hat{\tau}_{j,k}$

$\text{se}(\hat{\tau}_k) \leftarrow \sqrt{\frac{1}{r-1}\sum_{j=1}^{r}(\hat{\tau}_{j,k} - \hat{\tau}_k)^2}$

$\text{CI}(\tau)_k \leftarrow (\hat{\tau}_k^{0.025}, \hat{\tau}_k^{0.975})$ (2.5% and 97.5% percentiles of the bootstrap draws)

结束循环

$\hat{}\mathcal{T} \leftarrow \frac{1}{s}\sum_{k=1}^{s} \hat{}\tau_k$标准误($\hat{}\tau$)

$\leftarrow \frac{1}{s}\sum_{k=1}^{s}$ 标准误($\hat{}\tau_k$) 置信区间(τ)

$\leftarrow \left(\frac{1}{s}\sum_{k=1}^{s} \hat{}\tau_k^{0.025}, \quad \frac{1}{s}\sum_{k=1}^{s} \hat{}\tau_k^{0.975}\right)$

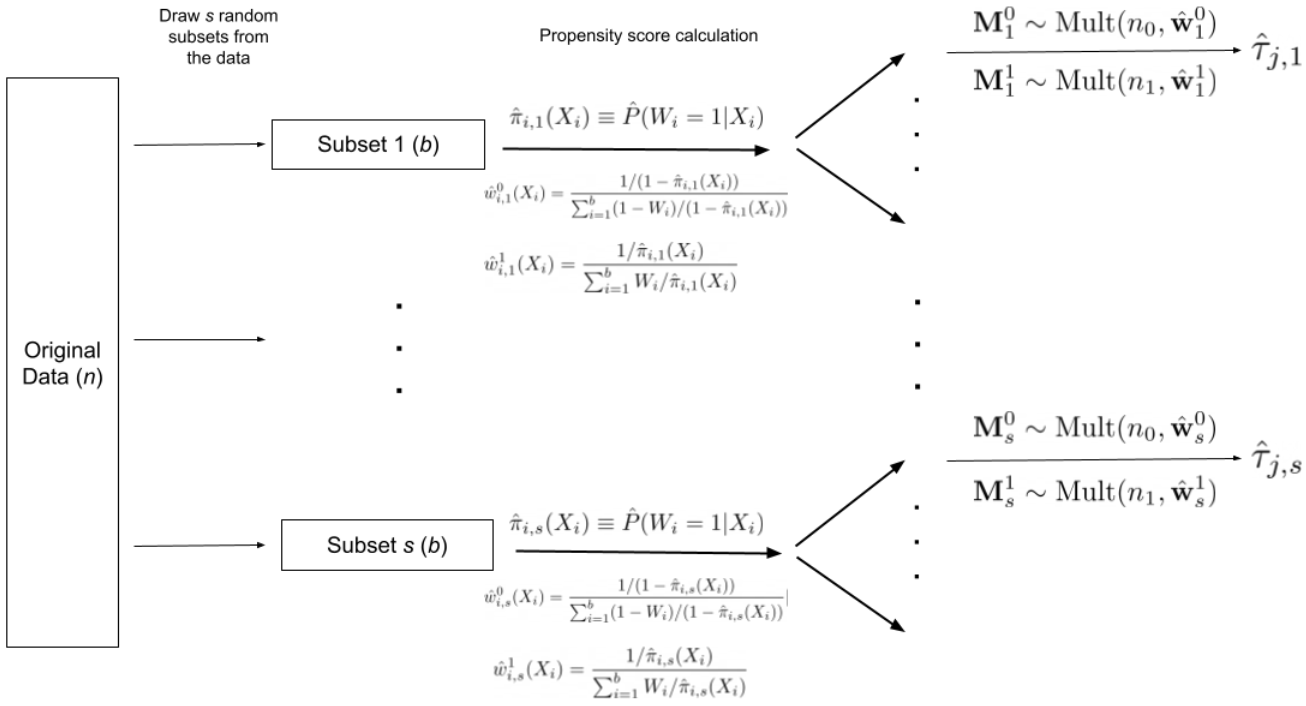

图1：因果BLB算法的图形表示。

我们将所提出的因果 BLB 与 Imbens & Menzel（2021）提出的因果自助法算法进行比较。最后，我们通过将因果 BLB 与传统自助法进行比较，讨论时间复杂度的改进。

定理 1（一致性）。 在满足（因果）一致性、无干扰、可忽略的处理分配以及倾向得分模型设定正确的条件下，$\hat{}\tau_{\mathrm{p}} \to \tau$。

证明见附录的 A 章节。总之，我们证明了因果 BLB 估计量的表现与方差稳定化 IPW 估计量类似；特别地，我们证明，只要倾向得分模型正确，它就依概率收敛于 IPW 估计量。

与 Imbens & Menzel（2021）的因果自助法的联系 Imbens & Menzel（2021）将不确定性视为源于分配的随机性，而非抽样不确定性，并据此提出了因果自助法。他们使用保序 copula $\hat{} F_{01}^{iso} = C^{iso}(\hat{} F_0, \hat{} F_1)$，对照组和处理组潜在结果的联合分布进行插补，其中 $\hat{} F_0$ 和 $\hat{} F_1$ 分别是对照组和处理组的边际经验累积分布函数（CDFs）。具体而言，对于随机化试验，他们仅需根据以下方式对缺失的反事实结果进行插补：

$$\tilde{Y}_i(0) := \begin{cases} Y_i & \text{if } W_i = 0, \\ \hat{F}_0^{-1}(\hat{F}_1(Y_i)) & \text{otherwise,} \end{cases}$$

$$\tilde{Y}_i(1) := \begin{cases} Y_i & \text{if } W_i = 1, \\ \hat{F}_1^{-1}(\hat{F}_0(Y_i)) & \text{otherwise.} \end{cases}$$

然后，他们通过反复从插补后的经验总体中抽取 n 个单位来模拟随机化分布。他们还证明，分布的一个估计量

$\hat{\tau}_{ATE}$（假设采用保序耦合）在任意阶近似下渐近保守。

这种因果自助法框架对小型或中等规模的样本可能效果良好，但由于涉及插补，在大数据场景下计算成本高得难以承受；对于观察性设计，尤其是在需要进行倾向得分估计和核插补时更是如此。事实上，已知当 $Y_i(1) = T(Y_i(0))$ 时，可以达到保序 copula，其中 T 是严格递增的变换（例如，参见 Schmidt（2007年））。现在我们证明，在假设 $Y_i(1) = T(Y_i(0))$ 的条件下，抽样不确定性等同于分配不确定性，因此我们的基于抽样的自助法框架等同于 Imbens & Menzel（2021）的因果自助法。我们仅针对实验试验进行说明；观察性设计的情形也可以用类似的方法证明，只是符号表达较为繁琐。注意，$\hat{F}_1(Y_i(1)) = \hat{F}_0(T^{-1}(Y_i(1))) = \hat{F}_0(Y_i(0))$，而 $\tilde{Y}_i(0)$ 的插补累积分布函数为

$$
\begin{aligned}
Pr(\tilde{Y}_i(0) \leq y) &= \frac{n_0}{n} Pr(Y_i \leq y | W_i = 0) + \frac{n_1}{n} Pr(\hat{F}_0^{-1}(\hat{F}_1(Y_i)) \leq y | W_i = 1) \\
&= \frac{n_0}{n} Pr(Y_i \leq y | W_i = 0) + \frac{n_1}{n} Pr(\hat{F}_1(Y_i) \leq \hat{F}_0(y) | W_i = 1) \\
&= \frac{n_0}{n} Pr(Y_i \leq y | W_i = 0) + \frac{n_1}{n} Pr(\hat{F}_0(Y_i) \leq \hat{F}_0(y) | W_i = 0) \\
&= \hat{F}_0(y),
\end{aligned}
$$

这就是观测到的 $Y_i(0)$ 的经验累积分布函数。类似地，我们可以证明，插补总体 $\tilde{}$ 的 $Y_i(1)$ 的累积分布函数等于观测到的 $Y_i(1)$ 的经验累积分布函数。因此，从插补总体中抽样等价于从样本中抽样。

改进的时间复杂度表1展示了使用因果 BLB 进行因果推断时，常用于获取加权估计量权重的方法的时间复杂度。与

方法	求解器	传统 TC	因果 BLB TC
逻辑回归	IRLS	$O(n)$	$O(sn^{\gamma})$
CBPS	BFGS	$O(n^2)$	$O(sn^{2\gamma})$
SVM	SMO	$O(n^3)$	$O(sn^{3\gamma})$

表 1：因果推断中所用算法的时间复杂度（TC）。

Kleiner 等人（2014）一样，我们用参数 γ 表示子集大小 b，因此 $b = n^{\gamma}$（经过适当取整，使 b 为整数）。例如，当 $\gamma = 0.7$ 且 $n = 10000$ 时，$b = 631$。

我们可以看到，对于协变量平衡倾向评分和支持向量机而言，时间复杂度表明，使用因果 BLB 方法具有明显的时间优势，而使用标准逻辑回归几乎没有什么收益。当我们考察图 2 中模拟所耗用的时间时，确实观察到了这一点。需要注意的是，当子集大小相对于数据规模过小时，迭代算法的运行时间可能会更长，而数据拆分可能会加剧计时问题。关于这一点以及协变量平衡倾向评分，请参见第 5.3 节中的讨论。

数据结构无关 因果 BLB 在数据结构方面具有无关性，即既可应用于随机化数据，也可应用于观察性数据。例如，可以使用因果 BLB，通过将倾向得分估计为处理分配的边际概率或条件概率，来从大型实验中估计平均处理效应（Lunceford 与 Davidian（2004））。

3 实践考量

因果 BLB 算法取决于获取倾向得分的方法及其超参数、子集的数量 s 和子集的大小 b、

自助法样本的数量以及置信区间的类型。本章节将提供一些关于如何选择这些参数的实用指南。对于超参数的选择，我们的建议基于补充材料 B.1 章节中描述的相对误差模拟。

估计倾向得分的方法。尽管逻辑回归是逆概率加权估计量中最常用的倾向得分估计方法，但也有多种算法可用于获取倾向得分；其中包括非参数方法和机器学习方法（Lee 等，2010）。不过，因果 BLB 的优势在于不要求使用任何特定模型；然而，如第2.1节所述，由于复杂性不同，某些算法在时间效率方面具有明显优势。

在第4节介绍的模拟中，我们使用了三种不同的方法来估计倾向得分权重，即逻辑回归、使用线性核且代价参数等于 0.01 的支持向量机（SVM），以及协变量平衡倾向评分（CBPS）。

当子集数量相对于数据较少时，这三种方法的表现都相当不错。然而，图 2 显示，每个子集的大小会极大地影响计算时间。特别是，在使用相对较小的 b 和较大的 s 时，采用 CBPS 和 SVM 的因果 BLB 能够显著减少计算时间。此外，对于完整数据集而言可能在计算上难以承受的超参数调优，应改为在各个子集上执行。

子集大小 b。 </style}每个子集的大小与子集数量密切相关，二者相互依赖。一般而言，每个子集使用的观测值越少，就需要更多子集才能获得足够好的结果。在最初的 BLB 论文中，Kleiner 等（2014，第 20 页）建议根据参数 γ 选择子集大小 b，其中

$b = n^{\gamma}$。根据他们进行的大量模拟，作者发现，在许多情况下， $\gamma = 0.7$ 是“合理且有效的选择”（Kleiner 等（2014，第 20 页））。在我们自己的模拟中（见补充材料图 8 和图 7 的相对误差轨迹），对于因果 BLB，我们发现适当的 γ 会因估计方法而异。对于支持向量机这类数据密集型机器学习方法，增加子集数量不足以使因果 BLB 算法收敛到较低的相对误差。这些方法要求每个子集具有较大的样本量， γ 约为 0。8。相比之下，只要使用足够多的子集，逻辑回归即使在 γ 低至 0.5 时也能收敛（见下一段讨论）。

子集数量 s。 同样，从误差轨迹中可以看出，对于较大的 b ，只需要少量子集。总体而言，与 Kleiner 等人（2014）的建议类似，我们建议$s \geq 2$ 用于 $\gamma = 0.9$ ， $s > 10$ 用于 $\gamma = 0.6$。

自助抽样样本数量 r。 相对误差轨迹显示，即使 $r = 100$ 相当小，算法也能收敛。与 Kleiner 等人（2014）的建议类似，补充材料中的图 9 展示了逻辑回归中相对误差随 r 和 s 变化的情况。我们发现，随着子集数量增加，每个子集的自助抽样样本数量几乎没有影响。

置信区间。 与传统自助法类似，可以利用渐近正态性或百分位数（Efron 和 Tibshirani（1994））获得置信区间。在模拟中，我们展示了通过百分位数（图4、图5和图6）以及渐近法（补充材料中的图10、图11和图12）获得的置信区间。结果表明，这两种技术都能提供适当的名义覆盖率。

缺乏重叠。 尽管逆概率加权估计量得到了广泛应用，但当倾向得分接近0时，它们可能会产生极端权重并导致错误推断（Kang & Schafer 2007）。我们建议按照标准做法截断倾向得分。或者，我们建议剔除导致极端权重的子集，并通过有放回抽样抽取一个新样本。

协变量平衡。 可在每个自助抽样样本和子集中，使用标准化均值差（Stuart 2010）计算协变量平衡。实际应用中，如果某个子集的协变量平衡未达到理想水平，可以将其舍弃并重新抽取一个子集。

并行计算。 因果 BLB 支持在现代计算平台上进行并行和分布式实现。其实现细节遵循 Kleiner 等人（2014）提出的原始 BLB 算法。

4 模拟

我们使用具有两个协变量和恒定处理效应的数据生成过程来评估因果 BLB 的性能，具体如下所述。我们参照 Morris 等人（2019）的展示指南来描述模拟设置。

$$
\begin{aligned}
X_1, X_2 &\overset{\text{i.i.d.}}{\sim} \text{Normal}(0, 1) \\
\Pr(W = 1 | X_1, X_2) &= \frac{1}{1 + \exp(-0.5X_1 - 0.5X_2)} \\
\epsilon &\sim \text{Normal}(0, 1) \\
Y(0) &= X_1 + X_2 + \epsilon \\
Y(1) &= Y(0) + 2W
\end{aligned}
\tag{1}
$$

数据生成机制：数据生成机制（DGM）如式（1）所示。我们为每位受试者 i 抽取两个相互独立的协变量（X_{1i}， X_{2i}），并将协变量的线性组合取逆对数几率，构造倾向评分。随后，我们抽取一个服从标准正态分布的独立同分布误差项 ϵ ，并构造对照组和处理组的结果。所有受试者的处理效应均为 2。协变量重叠程度由逆对数几率方程中的参数控制； $-\frac{1}{2}$ 确保处理组和对照组在部分协变量上存在一定的不平衡。

估计目标：我们感兴趣的估计目标 τ 是平均处理效应（ATE）： $\tau = E(Y(1) - Y(0))$。

方法：我们构建了 500 次数据生成机制（DGM）重复试验，样本量从 5,000 变化到 20,000，以考察结果的直方图以及结果如何随样本量增加而变化。为了说明因果 BLB 使用几种不同估计量时的性能，我们分别使用逻辑回归、支持向量机（SVM）和协变量平衡倾向评分（CBPS）计算出的倾向得分来估计权重。

性能指标：我们评估（1）不同方法的运行时间，（2）偏倚，以及（3）自助法置信区间的覆盖率。具体而言，我们希望确保自助法置信区间至少达到名义覆盖率，即

$$\Pr(\tau \in \mathrm{CI}^*) \geq 1 - \alpha.$$

4.1 结果

计时。 图 2 展示了不同子集数量下，我们采用的各种倾向得分估计方法所耗费的时间。在比较耗时的时候，我们希望确保各项计算使用相同的样本量；因此， b 被设为 n/s。我们发现，对于协变量平衡倾向评分和支持向量机这类计算密集型方法，增加子集数量有明显的时间优势

此外，我们发现，对于逻辑回归这类方法，增加子集数量并没有什么实际好处。我们在第2.1节讨论因果BLB性质时提到过这一点，而模拟结果也证实了这一点。

偏倚。 图3展示了在三种不同样本量、子集数量和倾向评分加权估计方法下，对数据生成机制进行500次重复试验所得的平均处理效应估计值。在此模拟中，每次计算均使用了 $\gamma = 0.8$。可以看到，所有方法都得到了无偏估计。

覆盖率 图4、图5和图6展示了不同子集规模下百分位数置信区间的“拉链图”。这些图呈现了通过500次重复试验得到的自助法置信区间。纵轴表示 $\frac{|\hat{\tau}-0.5|}{\mathrm{se}(\hat{\tau})}$ 的分位百分数。这些图表明，对于 $\gamma = 0.8$， $s = 2$ 个子集不足以达到至少名义覆盖率；这一问题在支持向量机模型中更为明显。随着子集数量增加，可以看到覆盖率逐渐接近 $s > 4$的名义覆盖率。

类似地，补充材料中的图10、图11和图12展示了针对每个子集 k 推导的渐近正态（Efron 和 Tibshirani（1994））置信区间，形式如下：

$$[\hat{\tau}_k^{\mathrm{IPW}} - z^{1-\alpha} \cdot \mathrm{se}(\hat{\tau}_k), \hat{\tau}_k^{\mathrm{IPW}} - z^{\alpha} \cdot \mathrm{se}(\hat{\tau}_k)]$$

其中，$\hat{\tau}_k^{\mathrm{IPW}}$是在整个子集上估计得到的平均处理效应估计量（它不是算法1中的自助法$\hat{\tau}_k$），标准误通过自助法估计。这是计算渐近置信区间的常用方法，参见 Efron与Tibshirani（1994年，第168页）。此处的标准误通过计算自助重抽样样本中平均处理效应估计值的标准差得到；随后对各子集的置信区间界限取平均。我们看到，渐近结果给出的结果相似，但略微更保守

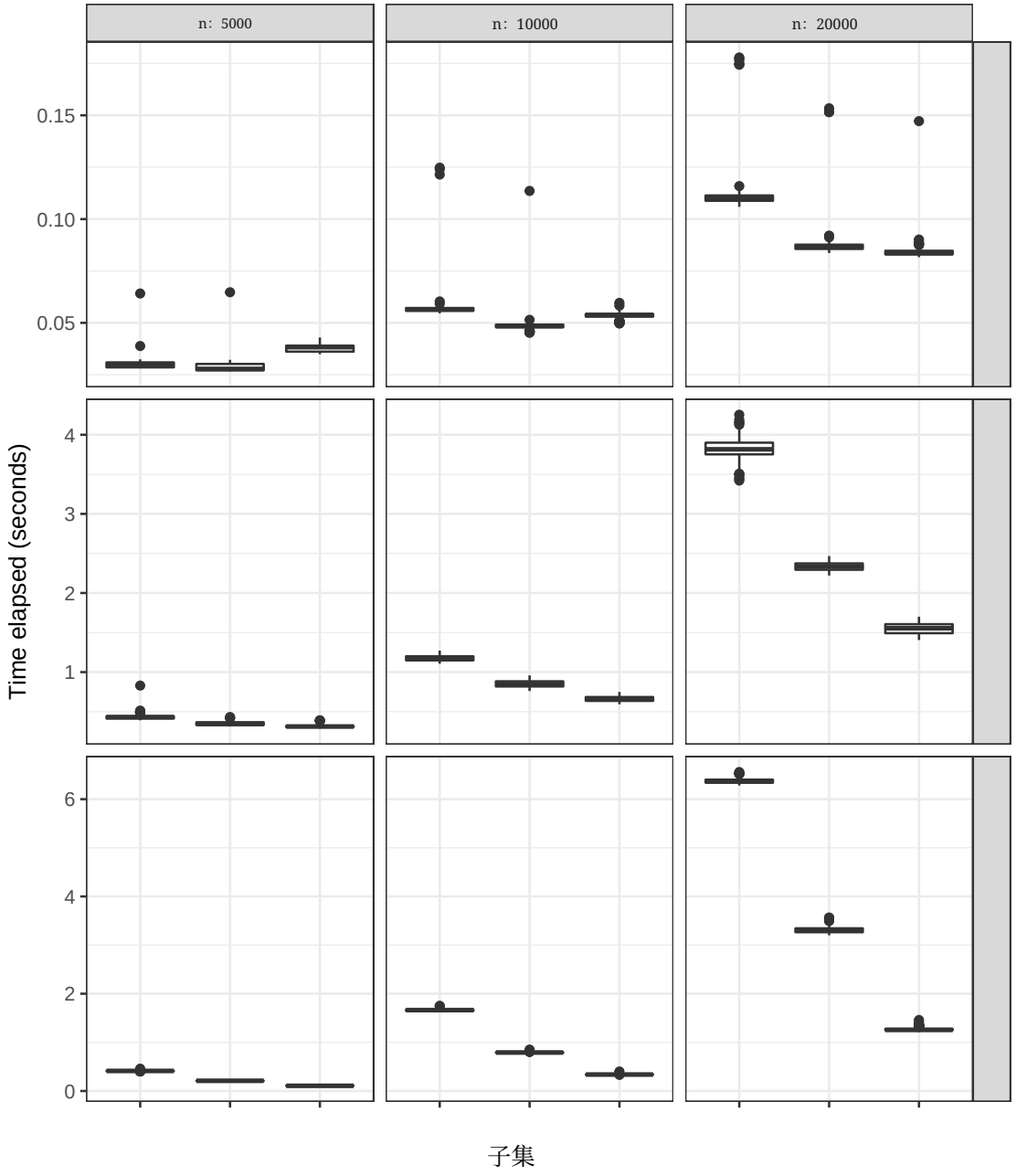

图 2：100 次重复试验中因果 BLB 运行时间的箱线图（$B = 100$）。

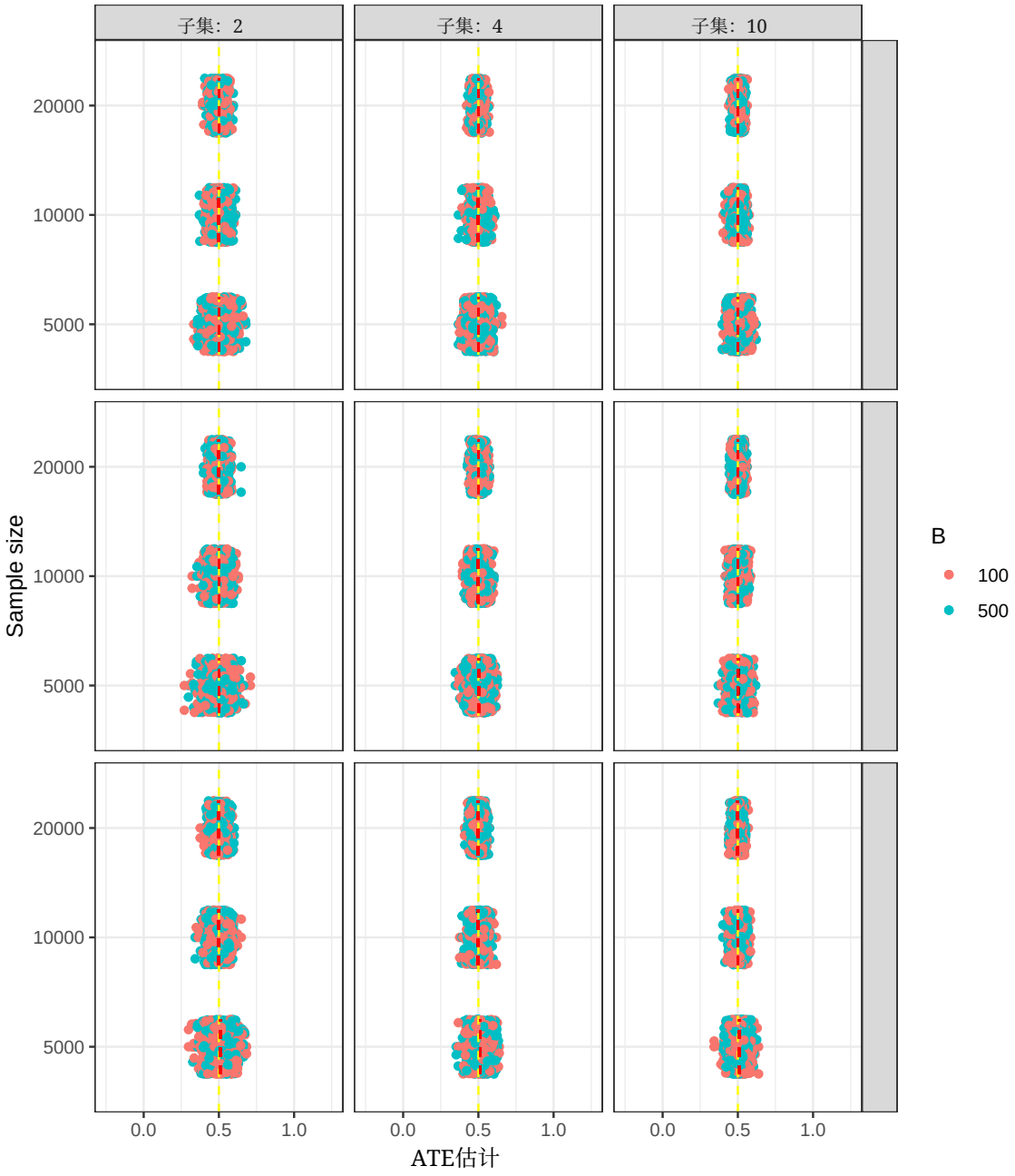

图3：500次重复实验得到的平均处理效应估计值（虚线表示真实值，红线表示均值）。

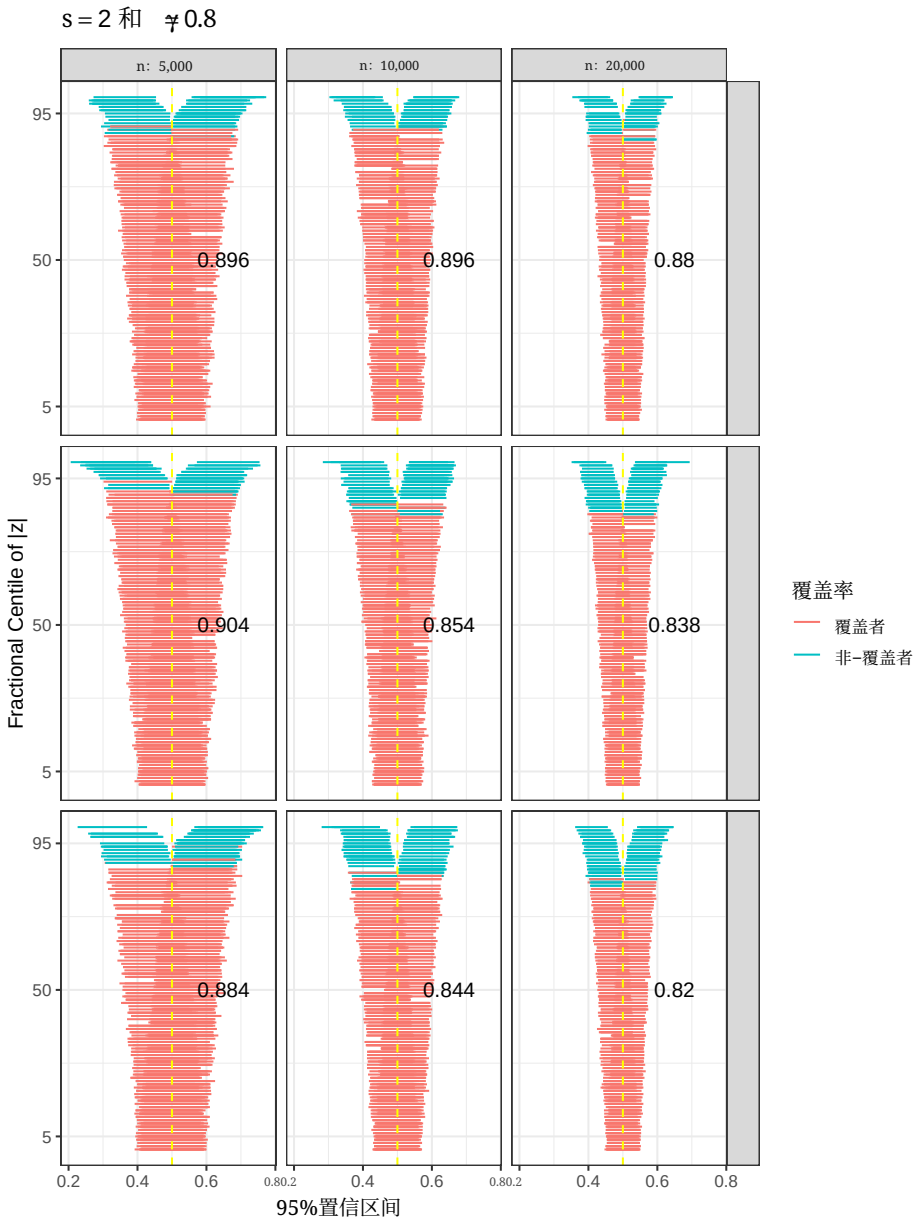

图 4：因果BLB算法进行 500 次重复试验得到的置信区间，2 个子集（黄色线表示真实值，$B = 500$）。

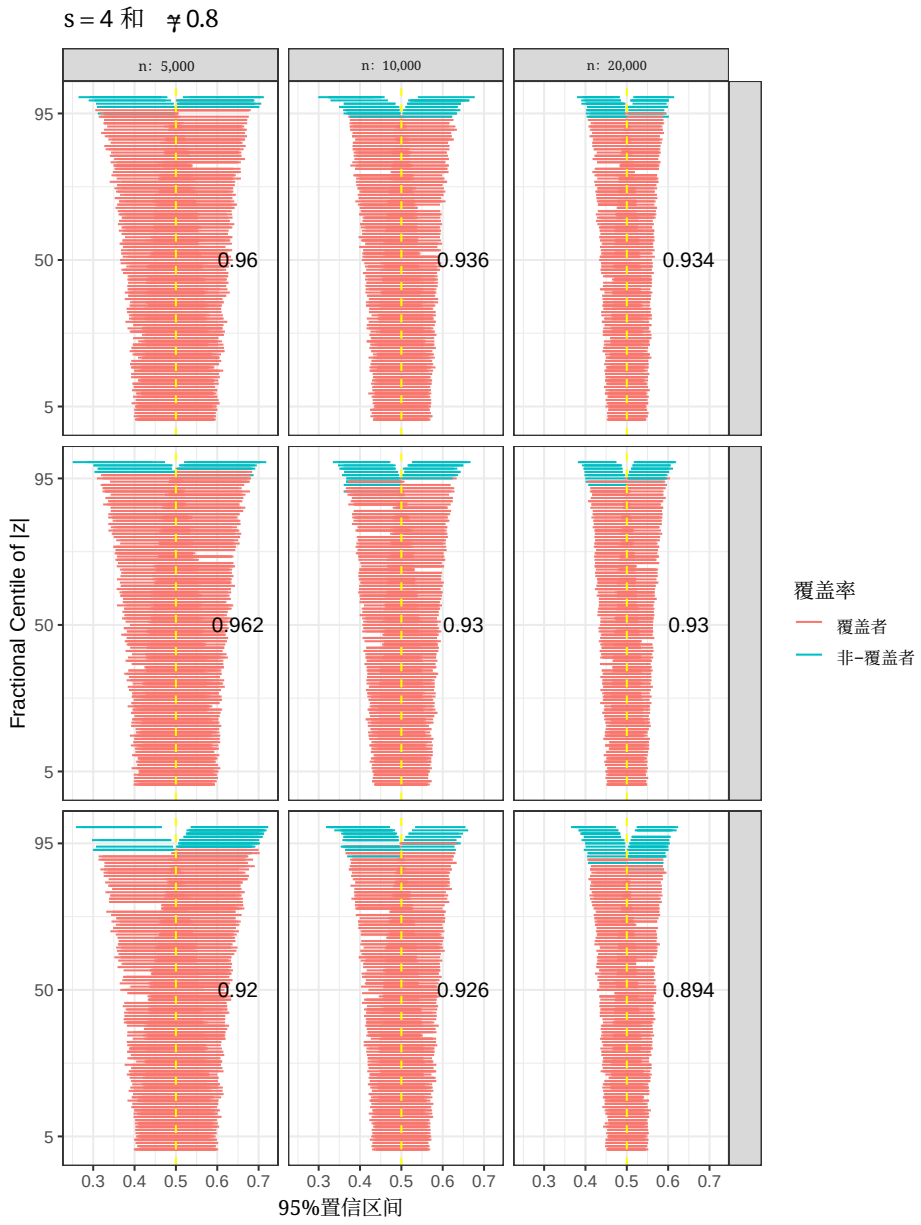

图5：因果BLB算法进行500次重复试验得到的置信区间，4个子集（真实值由黄色线表示，$B = 500$）。

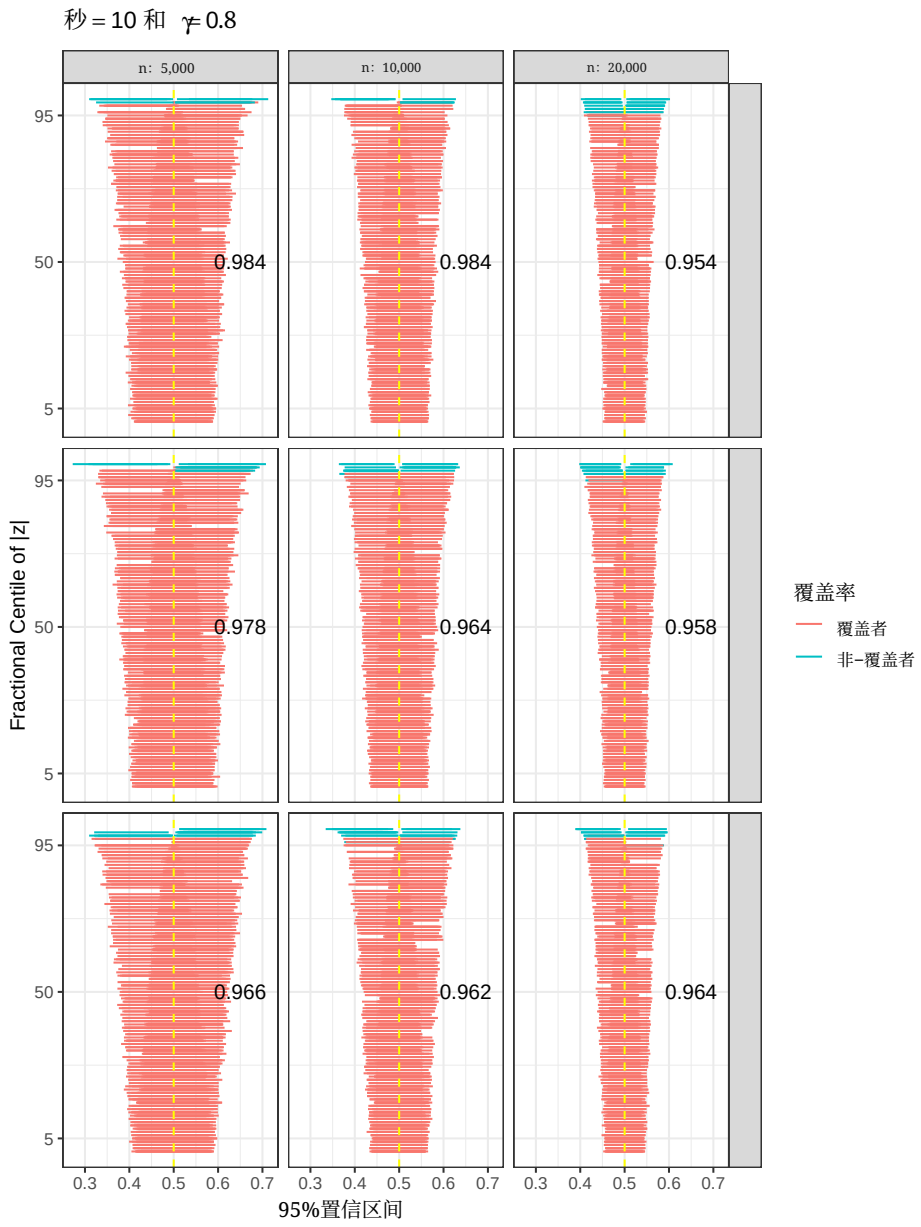

图 6：因果BLB算法进行 500 次重复试验所得的置信区间，共 10 个子集（黄色线表示真实值，$B = 500$）

推断。

5 在妇女健康倡议中的应用

为在真实数据集上评估因果 BLB，我们将其应用于妇女健康倡议的数据。妇女健康倡议是一项于1992年启动的临床研究，旨在了解绝经后女性最常见的死亡和患病原因 Group 等（1998）。该倡议包括一项随机试验和一项观察性研究。在随机试验中，16,608名女性被随机分配每日服用0.625毫克结合雌激素（CEE）和2.5毫克醋酸甲羟孕酮（MPA），或服用安慰剂。在观察性研究中，共有53,054名女性入组，其中约三分之一（17,503名）正在使用雌激素加孕激素 Prentice 等（2005）。

本章节中，我们使用妇女健康倡议观察性研究（1993年9月至2010年9月）的数据，评估雌激素加孕激素疗法对50至79岁绝经后女性冠心病（CHD）平均发生时间的影响。

5.1 研究人群

我们将目标研究人群界定为绝经后女性：她们在 WHI 观察性研究中报告称，在入组前 2 年内未使用雌激素疗法、孕酮疗法或两者的组合。基线定义为首次随访访视，并从基线开始随访女性，直至确诊冠心病、失访、死亡或 2010年9月30日，以最先发生者为准。在 WHI 原始观察性研究的 93,676 名女性中，37,080 名在入组前 2 年内使用过任何激素治疗，另有 30,960 名缺少入组以来天数的信息，1,567 名缺少绝经后时间的信息。最终

研究人群由 24,069 名女性组成。我们考察了 34 个混杂因素，详见补充材料 D 章节。绝经后时间已被认为是影响激素治疗对冠心病风险和获益的重要因素（Carrasquilla 等人，2015年，2017年）。因此，我们通过按绝经后时间分为三组进行分层分析，评估雌激素加孕激素疗法对冠心病平均发生时间的影响：0-10 年、11-20 年和 $20+$年。

5.2 估计

我们的目标是使用因果 BLB 估计量估计激素治疗对冠心病发生时间的平均处理效应。为获得倾向得分，我们使用了逻辑回归、支持向量机和协变量平衡倾向评分。对于支持向量机估计，由于倾向得分很可能是混杂因素的复杂函数，我们使用了径向基核和 0.01 的代价参数。根据模拟结果，对于每种估计方法和每个分层数据集（按绝经后时间分层，详见第 5.1 节），我们使用 $s = 2$和 $s = 4$运行因果BLB算法。我们没有根据 γ选择子集大小 b，而是选择了 $b = {}^{n}_{s}$，以便准确比较使用不同子集时因果 BLB 的计算时间。由于每个分层数据集都相对较小（每个数据集均少于 10,000 个观测值），我们省略了 $s = 10$ 情形。我们将自助法重复抽样次数设为 $r = {}^{500}_{s}$。最后，我们计算了因果 BLB 估计值、标准误和百分位数置信区间。自助法标准误的计算方法是：先计算每个子集的估计值的标准差，再对各子集的结果取平均。为进行经验比较，我们还使用估计的倾向得分计算了标准逆概率加权估计量。

5.3 结果

结果见表2。在该表中，“方法”表示用于计算倾向得分的方法。 s、 b 和 r 分别表示子集数量、子集大小和自助法重复抽样次数。“完整”表示在每个按“绝经后时间”分层的数据集上估计的标准逆概率加权估计量。$\hat{\tau}$、se($\hat{\tau}$)和 (CI(τ))分别表示按照算法1所述方法计算的点估计、标准误和百分位数置信区间。为计算计算时间，我们运行算法 100 次，然后计算以秒为单位的中位时间，结果显示在 median(T)列中。[1]我们看到，对于每个年份区间，结果都相似，表现为较小的负向且在统计学上不显著的结果。特别是，支持向量机得到的负向效应略大。需要注意的是，与模拟结果相反，随着 s 增加且 b 下降，协变量平衡倾向评分所需的时间会增加。我们在图 2 中已经看到，对于规模与分层妇女健康倡议数据集相当的数据集，使用协变量平衡倾向评分时，因果 BLB 的时间加速效果相当有限。然而，时间的绝对增加取决于倾向得分估计中所使用协变量数量的增加。使用默认最大迭代次数的协变量平衡倾向评分算法，在低 b、高 s 的情形下会运行更多次迭代。关于这一点的进一步讨论，见附录 E。

6 结论

本文介绍了因果 BLB，这是一种新颖的自助法算法，可在从大数据中一致估计因果效应的同时，缩短传统自助法的计算时间。我们证明，所获得的估计量与一个

[1]点估计、标准误和置信区间结果来自算法的一次运行；该算法仅运行 100 次，以估计所耗用的时间。

绝经后时间							
绝经	方法	s	b	Full	$\hat{\tau}$	se(ˆτ) (CI(τ))	median(T)
0-10	CBPS	2	3708	0.00	0.00	0.05（-0.1，0.09）	6.34
0-10	CBPS	4	1854	0.00	0.00	0.05（-0.09，0.09）	7.77
0-10	PS	2	3708	-0.01	-0.03	0.04（-0.11，0.06）	0.61
0-10	PS	4	1854	-0.01	-0.05	0.04（-0.13，0.04）	0.59
0-10	SVM	2	3708	-0.08	-0.13	0.04（-0.21，-0.04）	5.15
0-10	SVM	4	1854	-0.08	-0.08	0.04（-0.17，0）	2.80
<10-20	CBPS	2	4768	0.00	0.00	0.06（-0.11，0.1）	8.73
<10-20	CBPS	4	2384	0.00	0.00	0.05（-0.11，0.09）	11.17
<10-20	PS	2	4768	-0.05	-0.07	0.05（-0.17，0.04）	0.76
<10-20	PS	4	2384	-0.05	-0.08	0.05 (-0.16, 0.02)	0.75
<10-20	SVM	2	4768	-0.13	-0.19	0.05 (-0.28, -0.09)	6.42
<10-20	SVM	4	2384	-0.13	-0.19	0.05（-0.28，-0.1）	3.52
20+	CBPS	2	3434	0.00	0.00	0.08（-0.16，0.14）	8.44
20+	CBPS	4	1717	0.00	-0.01	0.09 (-0.19, 0.15)	14.46
20+	PS	2	3434	-0.06	-0.03	0.1（-0.22，0.16）	0.61
20+	PS	4	1717	-0.06	-0.05	0.09（-0.21，0.12）	0.62
20+	SVM	2	3434	-0.15	-0.05	0.09（-0.23，0.12）	2.37
20+	SVM	4	1717	-0.15	-0.27	0.1（-0.45, -0.08）	1.34

表2：使用妇女健康倡议观察性研究的数据，应用因果 BLB 估计激素治疗对冠心病发病时间的平均处理效应。“方法”表示用于计算倾向得分的方法。 s、 b 和 r 分别表示子集数量、子集大小和自助法重复次数（$r = \frac{500}{s}$ ）。“Full”表示按“绝经后时间”分层的逆概率加权估计。ˆτ、se(ˆτ)和（CI(τ)）分别是按照算法1所述方法计算的点估计、标准误和百分位数置信区间。中位时间（秒）。

方差稳定化逆概率加权估计量。我们还表明，该技术在支持向量机等复杂机器学习方法中具有计算优势。我们方法的主要局限在于必须正确估计倾向得分模型。因此，后续工作将包括把因果 BLB 扩展到双重稳健方法。此外，根据有限样本模拟取得的良好结果，未来的研究方向包括证明因果 BLB 能够得出渐近正确的置信区间。

参考文献

Abadie, A. & Imbens, G.W.（2008），《匹配估计量中自助法失效的问题》，《计量经济学》76(6), 1537–1557。

Abadie, A. & Spiess, J.（2022年），《稳健的匹配后推断》，《美国统计学会会刊》117(538), 983–995。

Adusumilli, K.（2018），倾向得分匹配的自助法推断，技术报告，工作论文。

Austin, P. C.（2016年），《使用治疗逆概率加权（IPTW）时的方差估计：生存分析》，《医学统计学》35(30), 5642–5655。

Austin, P. C. & Small, D. S.（2014年），《无放回倾向得分匹配中自助法的使用：一项模拟研究》，《医学统计学》33(24)，4306–4319。

Bickel, P. J., G¨otze, F. & vanZwet, W.R. (2012)，《重抽样少于 n 个观测值：收益、损失及弥补损失的方法》，载于《Willem van Zwet 选集》，施普林格，第 267–297 页。

Carrasquilla, G. D., Berglund, A., Gigante, B., Landgren, B.-M., de Faire, U., Hallqvist, J. & Leander, K. (2015), 《绝经后早期开始绝经期激素治疗能否降低心肌梗死风险？一项基于人群的病例对照研究》, 《绝经》**22**(6), 598–606。Carrasquilla, G. D., Frumento, P., Berglund, A., Borgfeldt, C., Bottai, M., Chiavenna, C., Eliasson, M., Engstr¨om, G., Hallmans, G., Jansson, J.-H. 等 (2017), 《绝经后激素治疗与卒中风险：基于人群的队列研究数据的合并分析》, 《*PLoS*医学》**14**(11), e1002445。Dagan, N., Barda, N., Kepten, E., Miron, O., Perchik, S., Katz, M. A., Hern´an, M. A., Lipsitch, M., Reis, B. & Balicer, R. D. (2021), 《全国大规模疫苗接种背景下的 BNT162b2 mRNA 新冠疫苗》, 《新英格兰医学杂志》。Efron, B. (1979), 《自助法：重新审视刀切法》, 《统计年刊》**7**, 1–26。Efron, B. & Tibshirani, R. J. (1994), 《自助法导论》, CRC出版社。Group, W. H. I. S. 等 (1998), 《妇女健康倡议临床试验和观察性研究的设计》, 《临床试验控制》**19**, 61–109。Hernan, M. & Robins, J. (2020), 《因果推断：假如……。博卡拉顿：Chapman & Hill/CRC》。Imai, K. & Ratkovic, M. (2014), 《协变量平衡倾向评分》, 《皇家统计学会会刊：*B* 辑（统计方法学）》**76**(1), 243–263。Imbens, G. & Menzel, K. (2021), 《因果自助法》, 《统计年刊》**49**(3), 1460–1488。

Imbens, G. W. & Rubin, D. B.（2015年），统计学、社会科学与生物医学科学中的因果推断，剑桥大学出版社。

Kang, J. D. & Schafer, J. L.（2007年），“揭开双重稳健性的神秘面纱：从不完整数据中估计总体均值的替代策略比较”，《统计科学》 **22**(4)，523–539。

Kleiner, A.、Talwalkar, A.、Sarkar, P. & Jordan, M. I.（2014年），“适用于海量数据的可扩展自助法”，《皇家统计学会会刊：*B*辑（统计方法学）》**76**(4)，795–816。

Kohavi, R.、Crook, T.、Longbotham, R.、Frasca, B.、Henne, R.、Ferres, J.L. & Melamed, T.（2009年），“微软的在线实验”，《数据挖掘案例研究》 **11**（2009年），39。

Kohavi, R.、Deng, A.、Frasca, B.、Walker, T.、Xu, Y. 和 Pohlmann, N.（2013），大规模在线受控实验，载于 《第19届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集》，第1168–1176页。

Lee, B. K.、Lessler, J. 和 Stuart, E. A.（2010），《使用机器学习改进倾向性评分加权》，《医学统计学》**29**(3)，337–346。

Lunceford, J. K. 和 Davidian, M.（2004），《通过倾向性评分进行分层和加权以估计因果治疗效应：一项比较研究》，《医学统计学》**23**(19)，2937–2960。

Mohammadi, N., Alimohammadian, M., Feizesani, A., Poustchi, H., Alizadeh, A., Yaseri, M., Mansournia, M. A. & Sadjadi, A. (2021年)，‘使用参数化 g 公式研究鸦片消费对上消化道癌症死亡的边际因果效应：对伊朗戈勒斯坦队列研究中 49,946 例病例的分析’，《公共科学图书馆·综合》 **16**(1), e0246004.

Morris, T. P., White, I. R. & Crowther, M. J. (2019), ‘使用模拟研究评估统计方法’, 《医学统计学》 38(11), 2074–2102. Otsu, T. & Rai, Y. (2017年), ‘平均处理效应匹配估计量的自助法推断’, 《美国统计学会会刊》 112(520), 1720–1732. Peysakhovich, A. & Lada, A. (2016年), ‘结合观察数据和实验数据以发现异质性处理效应’, *arXiv* 预印本 *arXiv:1611.02385* . Politis, D. N., Romano, J.P.&Wolf, M.(1999 年), 非平稳时间序列的子抽样, 见 ‘子抽样’, 施普林格, 第 101–119 页。Præstgaard, J. & Wellner, J. A. (1993年), ‘一般经验过程的可交换加权自助法’, 《概率年刊》 第 2053–2086 页。

Prentice, R. L., Langer, R., Stefanick, M. L., Howard, B. V., Pettinger, M., Anderson, G., Barad, D., Curb, J. D., Kotchen, J., Kuller, L. 等（2005年）, 《绝经后激素治疗与心血管疾病：弥合观察性研究与妇女健康倡议临床试验之间的差异》, 《美国流行病学杂志》162(5), 404–414。

Rosenbaum, P. R. & Rubin, D. B.（1983年）, 《倾向得分在因果效应观察性研究中的核心作用》, 《生物计量学》70(1), 41–55。Samworth, R.（2003年）, 《关于恢复自助法一致性的方法的一则说明》, 《生物计量学》90(4), 985–990。Schmidt, T.（2007年）, 《应对 Copula》, *Copula*：从理论到金融应用3, 34。

Stuart, E. A.（2010年），《因果推断的匹配方法：综述与展望》，《统计科学：数理统计学会评论期刊》**25(1)**, 1。Sud, M., Han, L., Koh, M., Austin, P. C., Farkouh, M. E., Ly, H. Q., Madan, M., Natarajan, M. K., So, D. Y., Wijeysundera, H. C. 等（2020年），《冠状动脉疾病患者中遵循血流储备分数处理阈值与主要不良心血管事件之间的关联》，*JAMA* **324(23)**, 2406–2414。

Tang, D., Agarwal, A., O'Brien, D. & Meyer, M.（2010年），重叠实验基础设施：开展更多、更好、更快的实验，载于《第16届ACM SIGKDD知识发现与数据挖掘国际会议论文集》，第17–26页。

Westreich, D., Lessler, J. & Funk, M. J.（2010年），《倾向得分估计：作为逻辑回归替代方案的机器学习和分类方法》，《临床流行病学杂志》**63(8)**, 826。

Wu, R. L., Idris, A. H., Berkowitz, N. M., Happe, M., Gaudinski, M. R., Buettner, C., Strom, L., Awan, S. F., Holman, L. A., Mendoza, F. 等（2022年），《低剂量皮下注射或静脉注射单克隆抗体预防疟疾》，《新英格兰医学杂志》**387(5)**, 397–407。

Zhang, Y. & Sabbaghi, A. (2021年), '大型观察性数据中的因果推断设计型自助法',《统计理论与实践杂志》**15(4)**, 1–26.

Zhao, Q., Small, D. S. & Bhattacharya, B. B.（2019），《通过百分位自助法对逆概率加权估计量进行敏感性分析》，《皇家统计学会会刊：*B*辑（统计方法学）》**81(4)**, 735–761。

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R. 等（2020），《2019年中国肺炎患者中发现的新型冠状病毒》，《新英格兰医学杂志》。

自助法估计量的一致性

对于每个子集 k，将数据排序，使所有对照组个体排在前面，所有处理组个体排在后面。子集 k 中用于估计平均处理效应的标准化逆概率加权估计量为：

$$\hat{\tau}_k = \sum_{i=1}^{b} \left(\frac{W_i Y_i w_{i,k}^1(X_i)}{\sum_{i=1}^{b} W_i w_{i,k}^1(X_i)} - \frac{(1-W_i) Y_i w_{i,k}^0(X_i)}{\sum_{i=1}^{b} (1-W_i) w_{i,k}^0(X_i)} \right) \tag{2}$$

$$= \sum_{i=1}^{b} \left(W_i Y_i w_{i,k}^{1'}(X_i) - (1-W_i) Y_i w_{i,k}^{0'}(X_i) \right) \tag{3}$$

假设子集 k 中有 b_k^1 个处理组个体和 $b_k^0 \equiv b - b_k^1$ 个对照组个体。现在，对于每次重复抽样 j，考虑以下多项分布抽样：

$$\left(M_{1,j,k}^1, \ldots, M_{b_k^1,j,k}^1 \right) \sim \text{Multinomial}(n_1; w_{1,k}^{1'}(X_1), \ldots, w_{b_k^1,k}^{1'}(X_{b_k^1}))$$
$$\left(M_{1,j,k}^0, \ldots, M_{b_k^0,j,k}^0 \right) \sim \text{Multinomial}(n_0; w_{1,k}^{0'}(X_1), \ldots, w_{b_k^0,k}^{0'}(X_{b_k^0}))$$

现在构造 $\hat{\tau}_{j,\ k}$

$$\hat{\tau}_{j,k} = \sum_{i=1}^{b} \left(\frac{-1}{n_0} \right)^{W_{i,k}-1} \left(\frac{1}{n_1} \right)^{W_{i,k}} \tilde{M}_{i,j,k} Y_{i,k}$$

给定子集中观测到的数据，多项分布抽样向量的期望为：

$$E(\tilde{M}_{i,j,k}) = \begin{cases} n_1 w_{i,k}^{1'} & \text{if } W_i = 1 \\ n_0 w_{i,k}^{0'} & \text{if } W_i = 0 \end{cases}$$

因此，对于上述求和中的一项，

$$\mathrm{E}_{\mathbf{M}}\left(\left(\frac{-1}{n_0}\right)^{W_{i,k}-1}\left(\frac{1}{n_1}\right)^{W_{i,k}}\tilde{M}_{i,j,k}Y_{i,k}\right)=\frac{W_iY_in_1w_{i,k}^{1'}}{n_1}-\frac{(1-W_i)Y_in_0w_{i,k}^{0'}}{n_0}$$
$$=W_iY_iw_{i,k}^{1'}-(1-W_i)Y_iw_{i,k}^{0'}$$

因此，根据弱大数定律，在 r 次加权自助法重抽样中（对应于 r 次处理组和对照组的多项分布抽样）

$$\frac{1}{r}\sum_{i=1}^{r}\hat{\tau}_{i,k}\xrightarrow{p}\hat{\tau}_k$$

由于每个子集都是从完整数据集中随机选取的子集，在通常的因果推断假设下，并假设倾向得分设定正确，$\hat{}_{\tau_k\,\mathrm{p}}\!\!\rightarrow\tau$。

B 附加参数考量

B.1 相对误差模拟

为确定实用指南，我们进行了与 Kleiner 等人（2014年）类似的模拟。我们使用第 4 章节中描述的相同数据生成机制。首先，我们生成了 2000 个相互独立的模拟数据集副本，每个副本包含 $n = 20000$。对于每个数据集，我们估计逆概率加权方差稳定化平均处理效应估计量，然后对结果计算分位数，以获得方差稳定化 IPW 估计量的“真实”置信区间。接着，在一个独立数据集上，我们迭代运行因果BLB算法，将每个子集的自助法重复次数设为 $r = 100$，更新平均处理效应估计，并记录处理每个子集所需的时间。每次更新估计值后，我们按如下方式计算相对误差：设 ξ_{lo} 和 ξ_{up} 分别为真实置信区间的下界和上界

，设 c_{lo}^{s} 和 c_{up}^{s} 分别为处理完第 s^{th} 个子集后 BLB置信区间的下界和上界。那么处理完子集 s 后的相对误差指标为：

$$\mathrm{Err}^{s}=\frac{\frac{\left|c_{\mathrm{lo}}^{s}-\xi_{\mathrm{lo}}\right|}{\xi_{\mathrm{lo}}}+\frac{\left|c_{\mathrm{up}}^{s}-\xi_{\mathrm{up}}\right|}{\xi_{\mathrm{up}}}}{2}$$

我们在独立数据实现上重复了 5 次，并对每次迭代的估计值和时间取平均。这些模拟得出了相对误差随时间变化的轨迹。绘制这些结果后，我们可以直观地查看不同 γ 取值下达到“收敛”所需的时间。

图 7 展示了 $\gamma \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$ 下的相对误差轨迹，其中倾向得分权重使用逻辑回归计算。类似地，图 8 展示了相同的轨迹，但倾向得分是使用带线性核的支持向量机估计的，代价参数为 0.01。图 9 展示了因果 BLB 在不同 r 和 s 取值下达到的相对误差。

C 渐近置信区间

本章节纳入了渐近置信区间。

D 混杂因素

WHI观察性数据的混杂因素列表如下：

- 不含矿物质的复合维生素使用（是，否）
- 含矿物质的复合维生素使用（是，否），

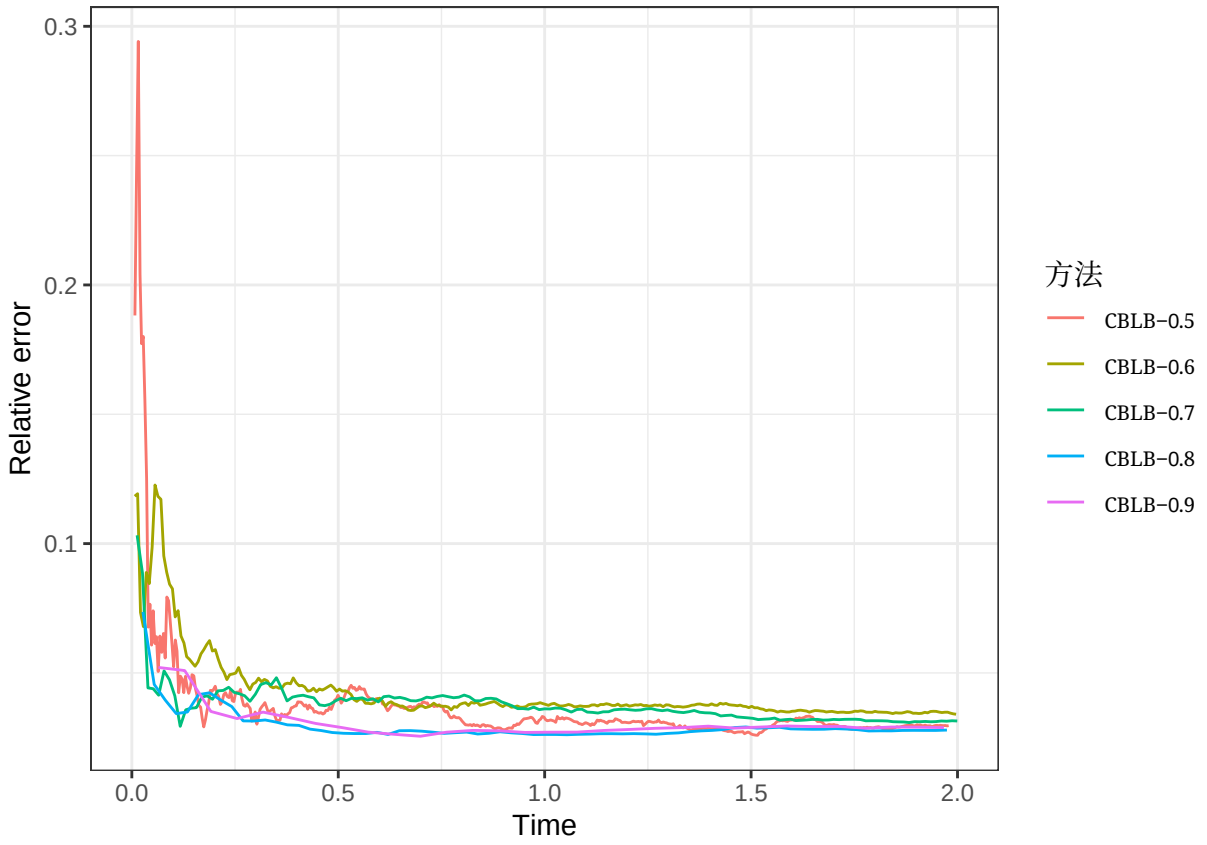

图 7: $n = 20000$ 的相对误差轨迹与处理时间的关系，其中倾向得分权重由逻辑回归估计。

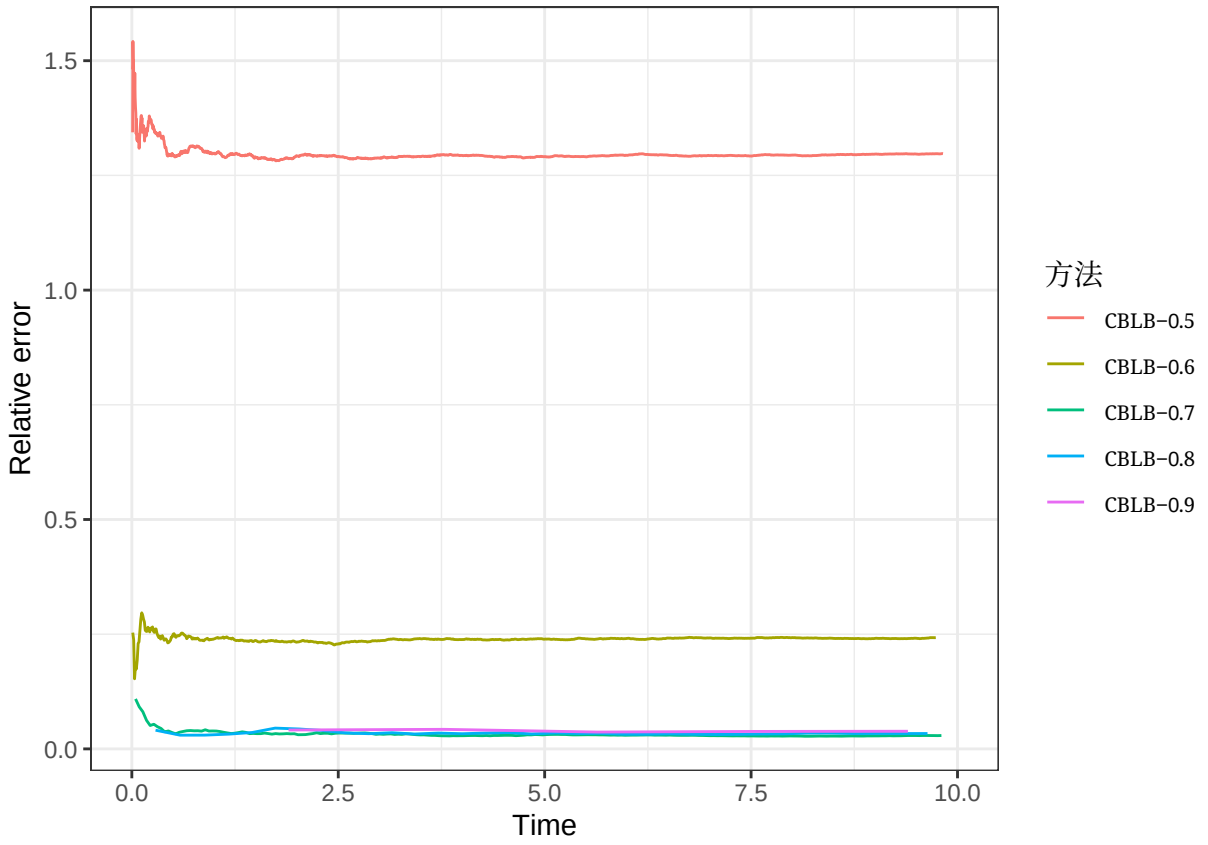

图 8: $n = 20000$ 的相对误差轨迹与处理时间的关系，倾向得分权重由支持向量机估计。

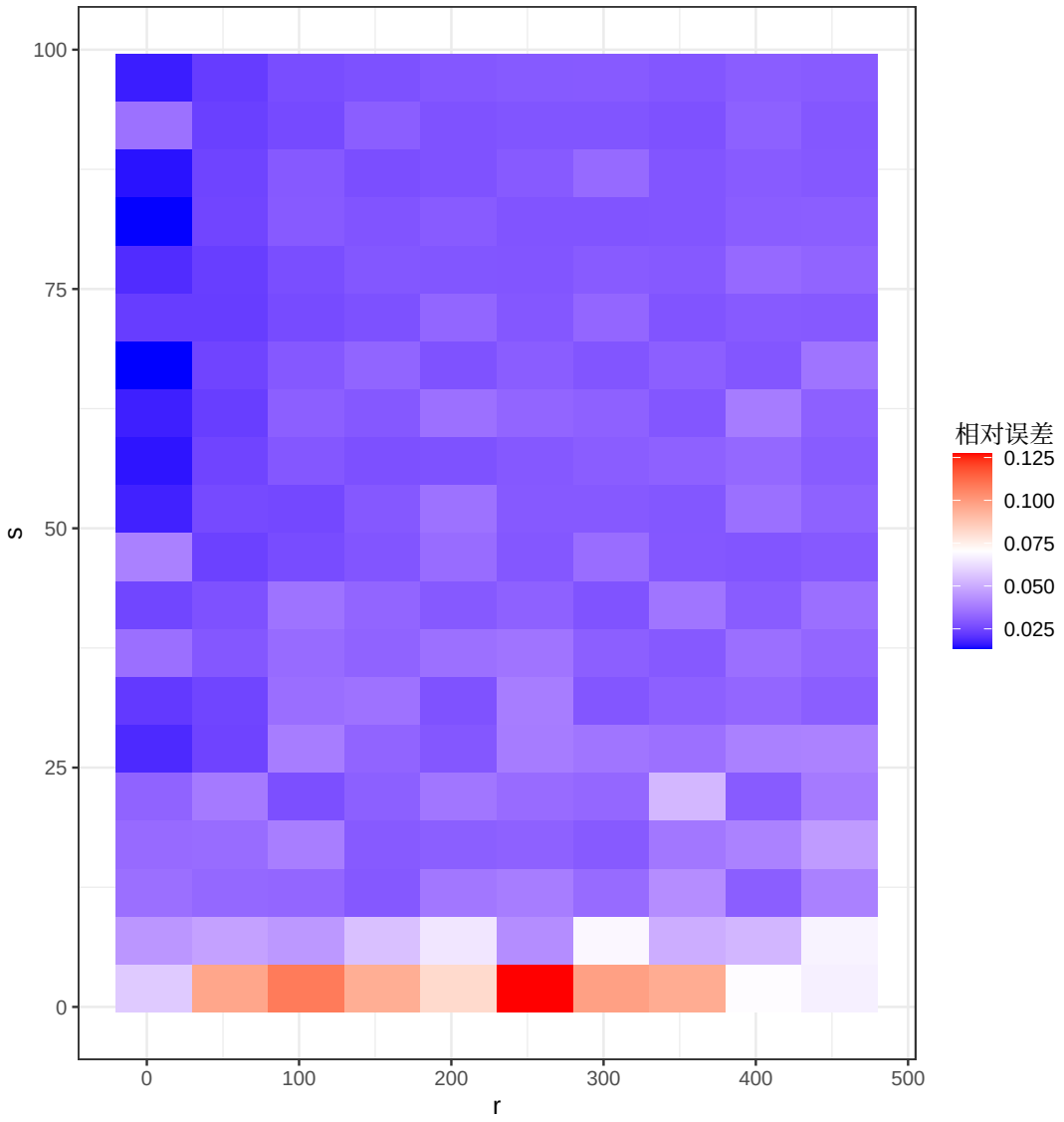

图 9：对于不同的 r 和 s，通过逻辑回归估计倾向得分权重所达到的相对误差，$n = 20000$。

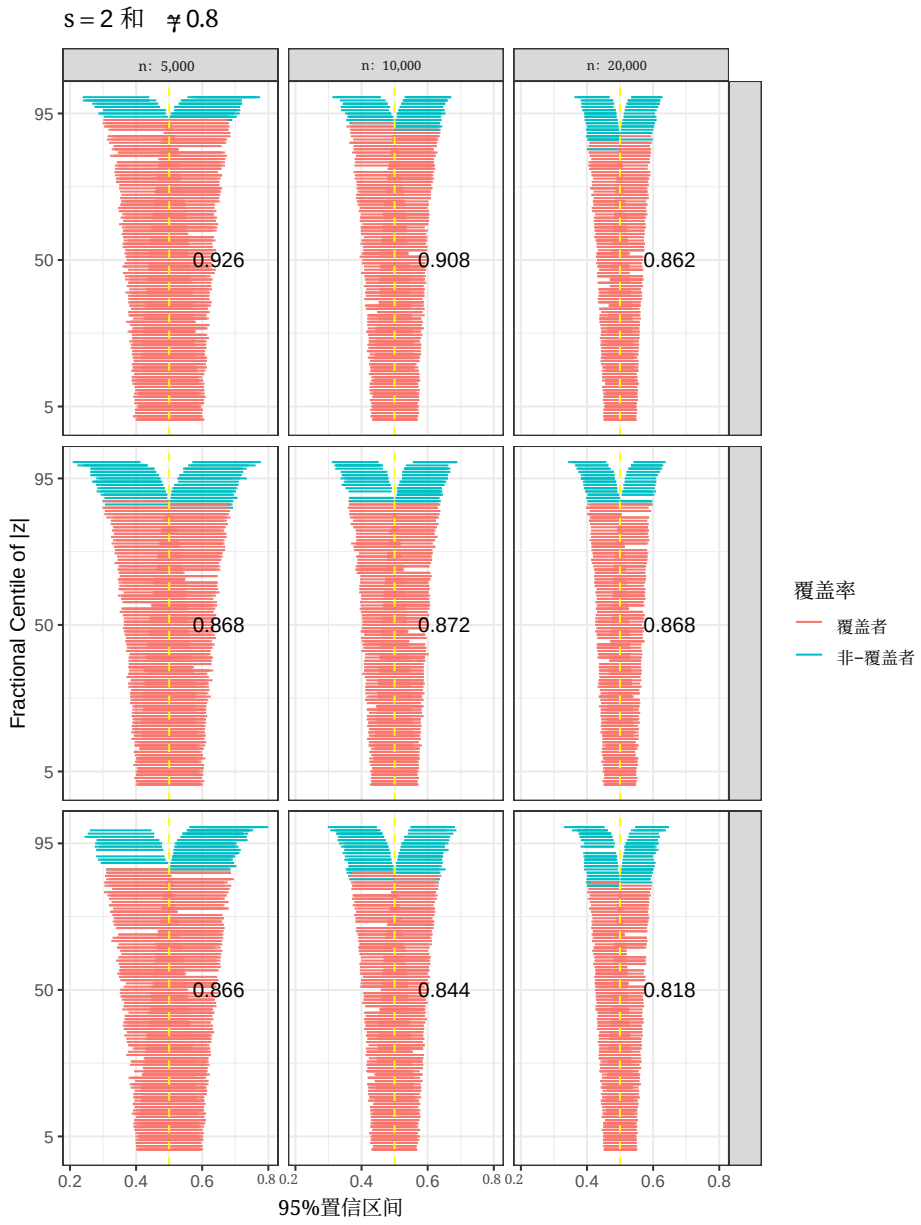

图10：因果BLB算法经过 500 次重复试验得到的渐近置信区间，2 个子集（黄色线表示真实值，$B = 500$）。

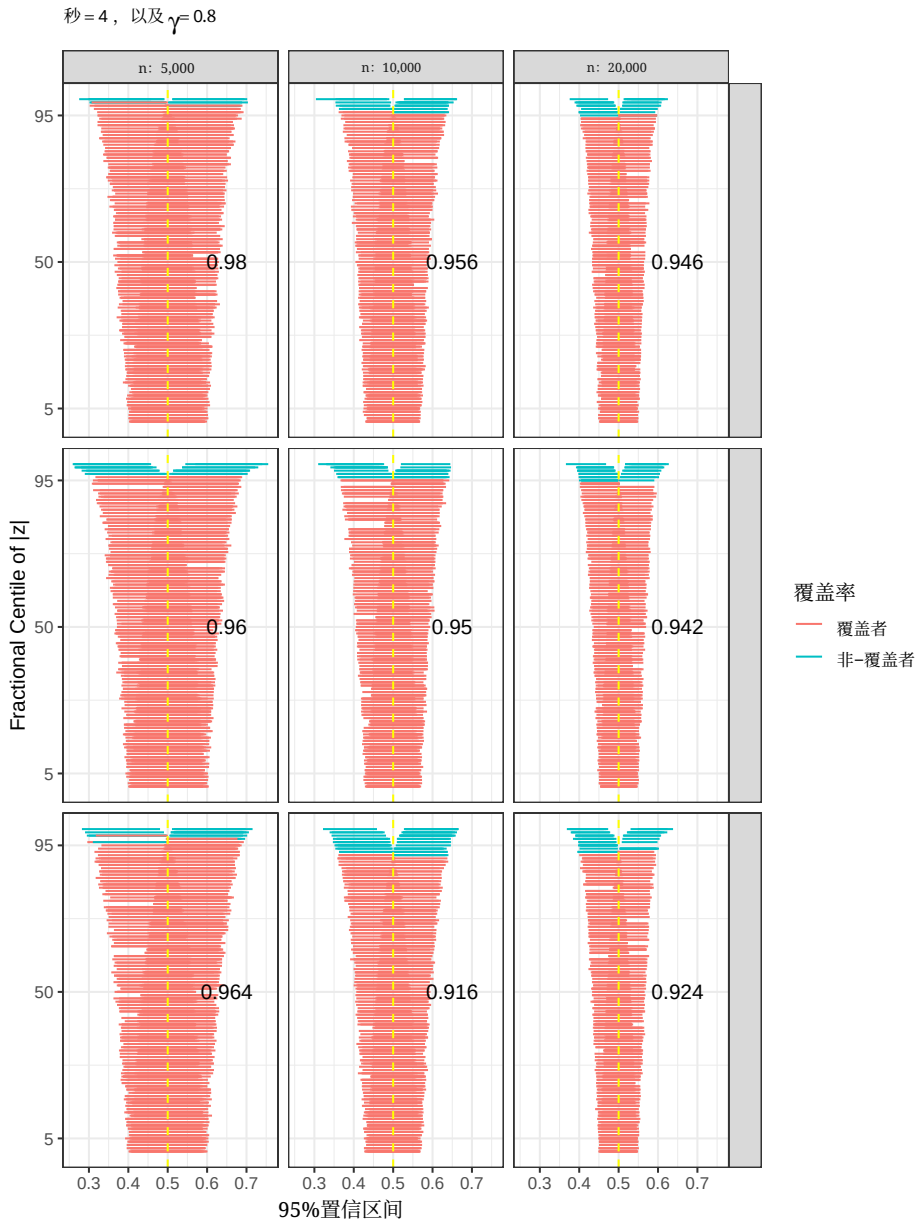

图11：因果BLB算法进行 500 次重复试验所得的渐近置信区间，包含 4 个子集（黄色线表示真实值， $B = 500$）。

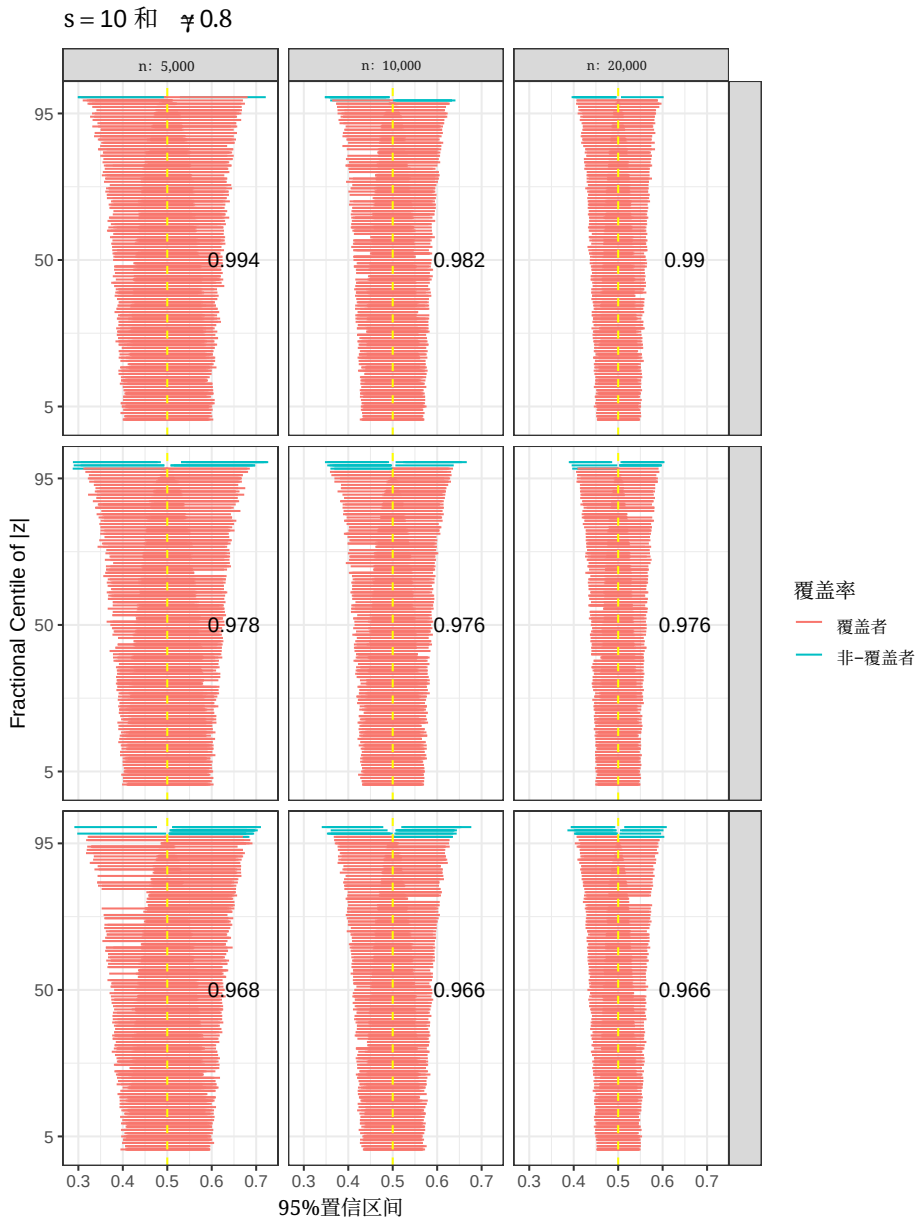

图12：因果BLB算法进行500次重复试验所得的渐近置信区间，包含10个子集（黄色线表示真实值，$B = 500$）。

- 族裔（白人、黑人、西班牙裔、美洲原住民、A 亚洲/太平洋岛民，未知），
- 妊娠次数（7 个类别）
- 双侧卵巢切除术（是、否），
- 绝经年龄（数值），
- 曾患乳腺癌（是、否）
- 曾患结肠癌（是、否）
- 曾患子宫内膜癌（是、否）
- 曾患皮肤癌（是、否）
- 曾患黑色素瘤（是、否）
- 过去10年患过其他癌症（是、否）
- 是否曾患深静脉血栓形成（是、否）
- 是否曾患卒中（是、否）、是否曾患心肌梗死（是、否）
- 是否曾患糖尿病（是、否）
- 是否曾因高胆固醇服用药物（是、否）
- 是否曾患骨质疏松症（是、否）
- 是否曾患心血管疾病（是、否）
- 冠状动脉旁路移植术（是，否）

- 是否曾发生心房颤动（是，否）
- 曾患主动脉瘤（是、否）
- 心绞痛（是、否），55岁或以上时髋部骨折（是、否）
- 曾吸食至少100支香烟（是、否）
- 酒精摄入量（不饮酒、过去饮酒、每月少于1杯、每周少于1杯、每周1至7杯、每周 $7+$ 杯）
- 水果中位数份/天（数值）
- 蔬菜中位数份/天（数值）
- 膳食能量（千卡）
- 收缩压（数值型）
- 舒张压（数值型）
- 体重指数（数值型）
- 教育程度（11 个类别）
- 收入（10 个类别）。

E 协变量平衡倾向评分 时间

为考察存在多个混杂因素时 CBPS 的性能，我们构建了以下模拟。我们创建了一个简单模型，其中二元结局 y 的成功概率由 p 个预测变量的线性组合决定。我们将 y 视为

处理指示变量，并将预测变量 **X** 视为混杂因素。为了评估因果 BLB 在不同 n-p 组合下的表现，我们构建了两个数据集，一个大小为 $n/2$ ，另一个大小为 $n/4$。我们在第一个数据集上运行 CBPS 算法 2 次，在第二个数据集上运行 4 次，以模拟因果 BLB 在数据子集上的操作（因为有 s 个子集时，CBPS 会运行 s 次）。每种 n-p 组合运行 10 次，并计算中位时间。结果如图 13 所示。我们看到，在预测变量数量较少时，采用 $s = 4$ 的因果 BLB 可以优于 $s = 2$，但随着预测变量数量p 增加，需要更多数据。

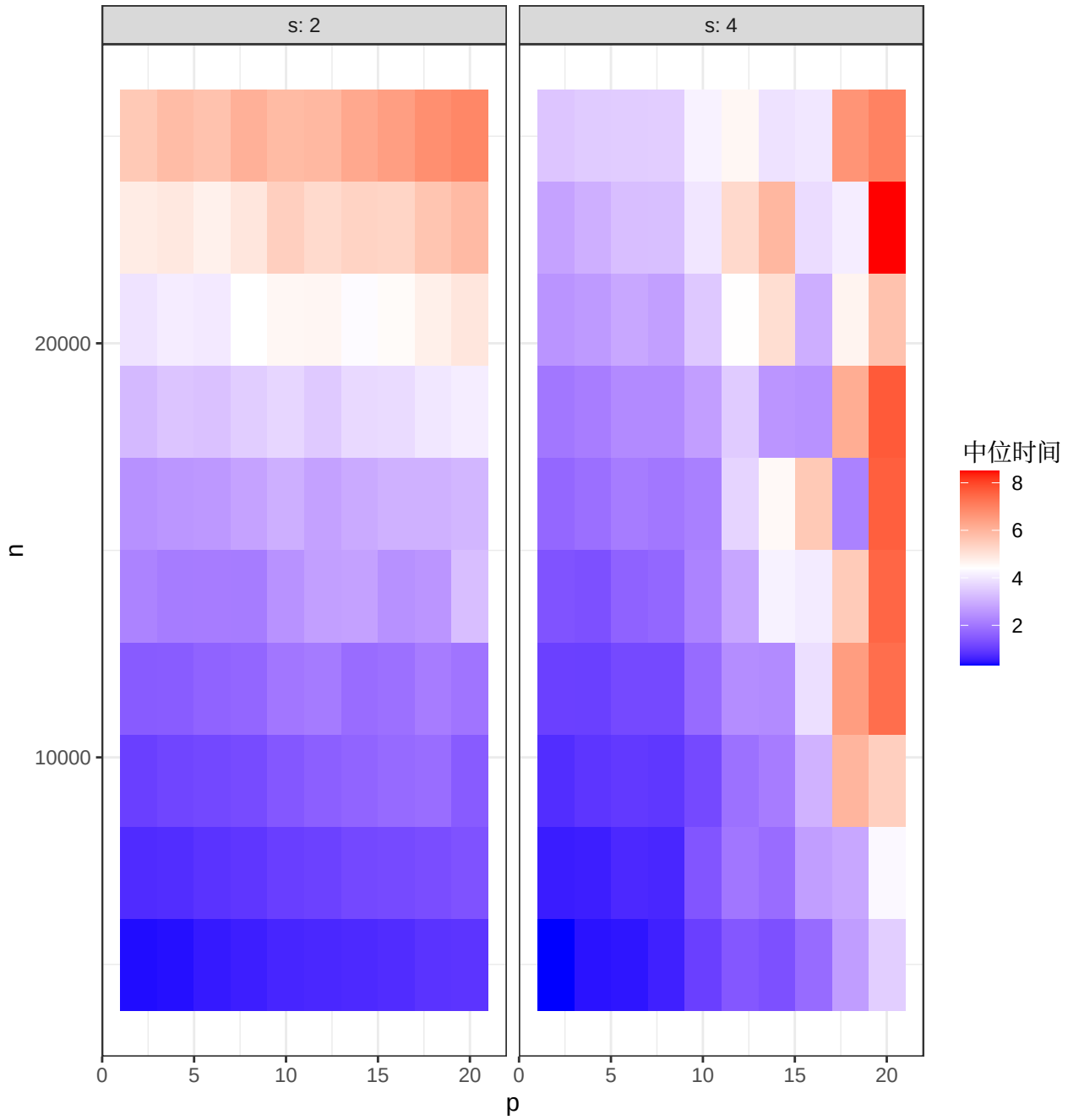

图 13：协变量平衡倾向评分的中位时间随 n 和 p的变化。