arXiv:2304.01316v1 [stat.ME] 3 Apr 2023

匹配机器学习：利用学习得到的度量进行处理效应推断的广义框架

Marco Morucci[1],Cynthia Rudin[2], 和 Alexander Volfovsky[3]

[1]纽约大学数据科学中心[2]杜克大学计算机科学系[3]杜克大学统计科学系

摘要

我们提出了匹配机器学习，这一框架将黑箱机器学习的灵活性与匹配的可解释性结合起来；匹配是观察性因果推断中的一项长期工具。在因果推断的许多高风险应用中，可解释性至关重要。目前，用于非参数估计平均处理效应和个体化处理效应的方法都是黑箱，无法对估计量进行人工审计。我们的框架利用机器学习学习用于匹配个体和估计结果变量的最优度量，从而在实现黑箱机器学习性能的同时保持可解释性。我们的广义框架将若干已发表的研究作为特例纳入其中。我们为所提出的框架提供渐近推断理论，使用户能够为个体化处理效应和平均处理效应的估计量构建近似置信区间。实证结果表明，匹配机器学习的具体实例表现与黑箱机器学习方法不相上下，且优于针对类似问题的现有匹配方法。最后，在应用中，我们展示了即使协变量数据高度复杂，匹配机器学习仍可用于因果推断：我们研究了一个图像数据集，并获得了高质量的匹配结果和处理效应估计量。

*Key*词语：匹配、因果推断、非参数、降维 on

1 引言

匹配方法在观察性因果推断中有着悠久的历史。由于方法简单，即使非技术背景的受众也能理解；此外，它还能快速执行因果分析，且具有已知的统计性质，同时对结果变量或处理分布均不作参数假设（Rosenbaum and Rubin，1983年）。近年来，用于非参数估计的黑箱机器学习（机器学习）工具已逐渐取代匹配，成为利用上下文协变量进行处理效应估计的默认方法（例如，Hill，2011年；Wager和Athey，2018年；Chernozhukov 等，2018；Hahn 等，2020）。这其中有充分的理由：机器学习黑箱的复杂性使其能够以前所未有的精度预测处理效应，而匹配方法有时无法达到这种精度。然而，使用黑箱机器学习也要付出结果可解释性方面的代价。我们主要担心的是，无法解释的模型难以审计，也就是说，难以利用有关数据、问题或研究人群的上下文信息，评估不可解释模型结果的稳健性和可信度。这一点尤其值得关注，因为大多数因果推断数据集都可能存在各种形式的隐藏噪声以及未测量混杂。本文提出利用机器学习的精度来提升匹配的可审计性，从而弥合两者之间的差距。我们提出了一个通用框架：首先利用灵活的机器学习学习用于匹配个体的距离度量，然后采用基于所学度量构建高质量匹配结果的匹配算法。这些匹配结果旨在近似黑箱机器学习的预测，同时仍保持可审计性；完成匹配后，便弃用黑箱，仅依据匹配结果开展分析。分析人员只需检查匹配结果本身，就能审计匹配估计量的质量。该框架中的方法可估计条件平均处理效应（CATE），即对任一个体的预期效应（也称为个体化处理效应）、平均处理效应（ATE），以及处理组平均处理效应（ATT）；这些分别是对所有

或仅针对处理个体。　　本文对匹配与处理效应估计领域的文献作出了若干重要贡献：

总体而言：

（i）我们提出了一种生成可解释的处理效应估计值的框架，

同时在 CATE 方面仍保持黑箱模型的准确性。

（ii）我们推导了利用以下方法得到的 CATE 估计量的渐近分布和误差界

我们框架涵盖的任何匹配方法。

(iii) 我们将框架扩展为一个通用的双重稳健匹配框架，以便对平均处理效应/对处理组的平均处理效应进行有效的$\sqrt{n}$-渐近推断。

我们还通过实证表明，我们的方法在有限样本下的性能确实与黑箱机器学习相当；从理论上看，对于条件处理效应，我们的匹配估计量渐近正态，且其方差已知并可估计。此外，我们借助双重机器学习（DML）（Chernozhukov 等，2018）方法，将匹配得到的条件平均处理效应结合起来，构造一个双重稳健估计量，从而以 $\sqrt{n}$ 的速率获得渐近正态的平均处理效应估计量。重要的是，后一种方法使我们能够绕开 Abadie 和 Imbens（2006）提出的问题，即许多常见匹配方法的渐近偏倚消失得很慢。我们的框架推广了许多已知且广泛使用的匹配方法，其中一些汇总于表 1。重要的是，我们为该框架建立的理论保证适用于它所推广的任何匹配方法。我们将在第 6 节进一步阐述该框架如何推广表 1 中的匹配方法。我们将通过对图像进行匹配来展示该方法的强大功能与灵活性：我们利用卷积神经网络，学习图像协变量的低维表示。我们将表明，即使输入协变量复杂且维度很高（例如图像），该方法仍能生成可解释的匹配组。虽然用于匹配的表示难以解释，但其结果可进行可视化审计：人们能够直观检查匹配后的图像组，并轻松评估其质量与可信度。这将使我们能够研究

社交媒体上的品牌响应性是否与增加相关

表 1：M-M 的一些特殊情形匹配方法 L.

方法	ϕ 的选择	q 的选择
最近邻	$\phi(\mathbf{x}) = \mathbf{x}$,	2
倾向得分匹配 (Rosenbaum and Rubin，1983年)	$\phi(\mathbf{x}) = h(\mathbf{x}) \quad h \in$ 取最小值$_{h \in \mathcal{H}} \mathbb{E}_{\mathbf{X}}[(h(\mathbf{X}) - \Pr(T_i = t \mid \mathbf{X}))^2]$	1
预后评分匹配 (Hansen，2008)	$\phi(\mathbf{x}) = h(\mathbf{x}) \quad h \in$ 取最小值$_{h \in \mathcal{H}} \mathbb{E}_{\mathbf{X},\ Y_i(t)}[(h(\mathbf{X}) - Y_i(t))^2]$	1
自适应超盒* (Morucci 等人，2020)	$\phi(\mathbf{x}) = [h_t(\mathbf{x})\ h_{t'}(\mathbf{x})]$ $h_t \in$ 使其取最小值的参数$_{h \in \mathcal{H}} \mathbb{E}_{\mathbf{X}, Y_i(t)}[(h(\mathbf{X}) - Y_i(t))^2]$ $h_{t'} \in$ 使$_{h \in \mathcal{H}} \mathbb{E}_{\mathbf{X}, Y_i\ (t')}[(h(\mathbf{X}) - Y_i(t'))^2]$	1
粗化精确匹配 （Iacus 等人，2012年）	$\phi(\mathbf{x}) = \tilde{\Gamma}\mathbf{x}$ $\tilde{\Gamma} \in \mathbb{R}^{p \times p}_{diag}$ 是逐维卡尺矩阵	∞
遗传算法 (Diamond 和 Sekhon，2013年)	$\phi(\mathbf{x}) = \mathbf{M}\mathbf{x}$ $\mathbf{M} \in \arg\min_{\mathbf{M} \in \mathbb{R}^{p \times p}_{diag}} \sum_{i,j \in \mathrm{Tr}}$ 马哈$(\mathbf{x}_i, \mathbf{x}_j, \mathbf{M})$	2
MALTS (Parikh 等人，2022年)	$\phi(\mathbf{x}) = \mathbf{M}\mathbf{x}$, $\mathbf{M} \in \arg\min_{\mathbf{M} \in \mathbb{R}^{p \times p}_{diag}} \left\| \sum_{i,\ j \in \mathrm{Tr}} \frac{(y_i - y_j) \exp(-\mathrm{Maha}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{M}) \mathbb{I}(T_i = T_j))}{\sum_{k \in Tr} \exp(-\mathrm{Maha}(\mathbf{x}_i,\ \mathbf{x}_k,\ \mathbf{M}) \mathbb{I}(t_i = t_k))} \right\|$	2
精细平衡* （Rosenbaum 等人，2007）	$\phi(\mathbf{x}) = h(\mathbf{x}) \quad h \in \arg\min_{h \in \mathcal{H}} \mathbb{E}_{\mathbf{X}}[(h(\mathbf{X}) - \Pr(T_i = t \mid \mathbf{X}))^2]$	1

注： * 需要对匹配问题施加额外约束。参数 ϕ 和 q 是 M-ML超参数，稍后将予以定义；将它们设为表中所列的值，即可得到左栏中的方法。 ϕ 是从原始空间到有用特征空间的映射。该映射通过机器学习模型估计，该模型最小化表中所示的损失函数。 q定义了学习到的特征空间中的距离度量。有关此表的详细说明，请参见第 6 节。处理由随机变量 T 表示，特定的处理水平为 t，协变量为$\mathbf{X}$ （若非随机变量，则为$\mathbf{x}$），潜在结果为 $Y(t)$，Tr 是训练集，Maha($\mathbf{x}_i$,$\mathbf{x}_j$,$\mathbf{M}$) 表示个体 i 和 j 的协变量向量在正定实值对角矩阵$\mathbf{M}$加权下的马氏距离。

与品牌进行消费者互动。这是营销和消费者行为中的一个重要问题（例如，Laroche等，2013年），但迄今为止，尚未在我们的方法所能实现的粒度层面进行研究。本文将按如下结构展开：第2节总体介绍匹配方法。第3节概述匹配机器学习框架，该框架通过在学习得到的距离度量上匹配个体来估计条件平均处理效应。第4节介绍我们方法的大样本理论性质。第5节提出一种算法，用于渐近有效地估计平均处理效应。第7节展示我们的方法在模拟数据集上的性能的实证证据。最后，第8节将我们的方法应用于一个问题：研究社交媒体上的品牌响应性对消费者与品牌互动量的影响。

2 匹配与观察性因果推断

我们首先介绍符号。设有一个包含 n 个个体的样本： $i=1$、.....、 n；对于一种可取 M 种可能取值的处理， $Y_i(1)$.....、 $Y_i(M)$ $\in \mathbb{R}$表示潜在结果，其中 $t \in \{1,\ldots,M\}$。分配的处理用随机变量 $T_i \in \{1,\ldots,M\}$表示。我们从不观察每个个体的完整潜在结果向量，而是观察结果变量 $Y_i = \sum_{i=1}^{n} Y_i(t)\mathbb{I}[T_i = t]$，其中 $\mathbb{I}[E]$ 是事件 E的指示函数。每个个体都被分配一个 p维协变量随机向量$\mathbf{X}_i$，其取值属于 $\mathbb{X}$，其中 $\mathbb{X} \subset \mathbb{R}^p$是紧集；观测到的协变量向量记为$\mathbf{x}_i \in \mathbb{X}$。对于任意随机变量 A，我们用 f_A 表示 A的概率质量函数或概率分布函数，并用 F_A 表示 A的累积分布函数，还用 f_A 表示 A的分布。我们用符号$\mathbb{E}_A[A]$ 和 $\mathbb{V}_A[A]$ 分别表示相对于 A分布的 A 的期望和方差。不带下标使用这些符号时，表示期望算子是相对于方括号内所有随机量而言的。对于所有 i，我们作出以下经典假设：

A1（数据分布）：

（a）数据 $\mathcal{O}_n = \{\mathbf{O}_i\}_{i=1}^n = \{Y_i,\mathbf{X}_i,\ T_i\}_{i=1}^n$是一组$\mathbf{O}$的独立同分布副本。
（b）协变量分布的定义域 $\mathbb{X}$ 是 $\mathbb{R}^p$的一个紧子集。

（c）协变量的边缘分布具有可微累积分布函数（相对于勒贝格测度） $F_{\mathbf{X}}$（$\mathbf{x}$），且存在常数 $c_{f_{\mathbf{X}}}$，$C_{f_{\mathbf{X}}}$，使得 $0 < c_{f_{\mathbf{X}}} < f_{\mathbf{X}}$（$\mathbf{x}$） $< C_{f_{\mathbf{X}}} < \infty$ 在 $\mathbb{X}$上处处成立。

A2（重叠）：对于所有 $\mathbf{x} \in \mathbb{X}$和 $t = 1$，.....，M，都有 $0 < \Pr(T = t|\mathbf{X} = \mathbf{x}) < 1$。

A3（条件可忽略性）： $T \perp\!\!\!\perp$（Y（1） .....，Y（M）$|\mathbf{X}$。

A4（有界高阶矩）：对于所有 t，$t' \in \{1,\ldots,M\}$、所有$\mathbf{x} \in \mathbb{X}$，存在某个 $\delta > 0$和常数 C_δ，使得： $\mathbb{E}[|Y(t)|^{2+\delta}|\mathbf{X} = \mathbf{x}\ T = t'] \leq C_\delta$。

本文主要关注针对给定协变量向量 $\mathbf{x}$ 估计条件响应函数（CRF）和条件平均处理效应。它们分别记为：

$\mu(\mathbf{x}\ t) = \mathbb{E}[Y(t)|\mathbf{X} = \mathbf{x}]$, 和 $\tau(\mathbf{x}) = \mu(\mathbf{x}\ t) - \mu(\mathbf{x}\ t')$，对应两种处理$t$、$t'$。请注意，假设 3 允许我们写作：$\mathbb{E}[Y|\mathbf{X} = \mathbf{x}\ T = t] = \mathbb{E}[Y(t)|\mathbf{X} = \mathbf{x}]$，因此意味着我们可以利用观测数据对感兴趣的量进行一致估计。我们还将使用记号 $\sigma^2(\mathbf{x}\ t) = \mathbb{V}[Y(t)|\mathbf{X} = \mathbf{x}]$ 表示潜在结果的条件方差。本文后续将关注 CRF 和 CATE 的平均形式的估计，其定义如下：平均响应函数（ARF） $\mu(t) = \mathbb{E}[Y(t)]$, 平均处理效应（ATE）： $\tau(t, t') = \mathbb{E}[Y(t)] - \mathbb{E}[Y(t')]$, 以及处理组平均处理效应（ATT）：$\delta(t, t') = \mathbb{E}[Y(t)|T = t] - \mathbb{E}[Y(t')|T = t]$。我们将看到，对 CRF 进行一致估计如何能够实现对所有其他量的一致估计。匹配的核心思想是创建一个匹配组，即个体的一个子集：$\mathtt{MG}(\phi, \mathbf{x}\ t) \subset \{1, \ldots, n\}$，其中包含观测结果变量将用于估计上述给定目标值 $\mathbf{x}$ 下感兴趣量的个体。因此，匹配程序的主要问题是选择哪些个体应纳入 $\mathtt{MG}(\phi, \mathbf{x}\ t)$。下一节将概述我们为此采取的策略。

3 匹配机器学习：通用流程

本文的核心思想是利用强大的黑箱机器学习模型来指导如何构建匹配组，进而估计条件响应函数和条件平均处理效应。为此，我们提出使用机器学习构建协变量的 d 维表示，随后在这些表示上进行匹配，而非直接使用原始协变量值：机器学习将灵活地学习能够揭示 $\mathbf{X}$、 T 和 Y 之间关系的表示，从而获得更高质量的匹配结果。为实现上述目标，我们引入函数 ϕ： $\mathbb{X} \mapsto \mathbb{R}^d$，它是观测协变量的表示函数。我们假设对于各自定义域中的所有 $\mathbf{x}$， $\phi(\mathbf{x})$ 均存在且连续。 ϕ 也可以依赖于处理水平，

这意味着每个处理组都会分别估计一个 ϕ 。为便于阅读，我们在记号中省略了这一点。接着，我们提出在一个单独的训练集上使用机器学习方法学习 ϕ 。这种表示的理念是将协变量映射到一个空间中，使潜在结果相近的个体彼此接近，而这正是匹配的最终目标。例如，如果仅根据 X 来判断三个个体 i、j、k 之间的相似性，我们可能会将 i 与 k 匹配，尽管从（未观测到的）潜在结果来看，i 与 j 更相似。相反，如果我们考虑某种变换 ϕ(X)，且该变换能够揭示 X 与 $Y(t)$ 之间的关系，那么我们或许就能正确判断应将 i 与 j 而不是 k 匹配。指定一个有用的映射 ϕ，可以显著提升匹配质量，优于直接基于原始协变量值进行的简单匹配。现有匹配文献已探讨过这一思想，以往使用的 ϕ 的常见示例包括：针对协变量 x 和期望处理水平 t 的倾向得分，ϕ(x) $=$ Pr($T = t$|X $=$ x)（Rosenbaum and Rubin，1983）；预后得分，即对照条件下的期望潜在结果：ϕ(x) $= \mathbb{E}[Y(0)|\mathrm{X} = \mathrm{x}]$（Hansen，2008）（假设 $t = 0$ 表示对照条件）；以及马氏距离：ϕ(x) $=$ Mx，其中 M 是由正权重构成的对角 $p \times p$ 矩阵（Rubin，1980；Diamond 和 Sekhon，2013；Parikh 等人，2022）。确定用于匹配的协变量表示后，我们还需要一个距离度量，用来表征个体在变换后协变量上的相似性。为此，我们采用一般的 L_q 范数。对于任意 (u, v) $\in \mathbb{X} \times \mathbb{X}$，令：$D_\phi^q(\mathrm{u}, \mathrm{v}) = \left(\sum_{j=1}^{d} |\phi(\mathrm{u})_j - \phi(\mathrm{v})_j|^q\right)^{\frac{1}{q}}$，表示 u 和 v 之间的 q 范数距离。q 可以是任意整数或无穷大，我们按通常方式将其定义为 D_ϕ^∞(u, v) $= \max_{j=1,\ldots,d} |\phi(\mathrm{u})_j - \phi(\mathrm{v})_j|$。实践中，最常用的范数选择包括 $q = 2$ （例如 Rubin，1980；Abadie 和 Imbens，2011；Diamond 和 Sekhon，2013），用于在协变量本身上进行匹配，即采用 L_2 距离；绝对值距离（在一维情况下对应任意 L_q ）（例如 Rosenbaum and Rubin，1983），用于在协变量的一维表示（如倾向得分）上进行匹配；以及 $q = \infty$ （例如 Rosenbaum and Rubin，1984；Iacus 等人，2012），用于粗化匹配方法。我们的主要

算法如下:

算法：匹配机器学习（M-ML）

输入： 一个包含 n 个观测值 $\mathcal{D} = \{\mathbf{x}_i,\ y_i,\ t_i\}_{i=1}^n$ 的数据集，划分为训练集和匹配集。一个期望的协变量值$\mathbf{x}$、期望的处理水平 t、一个由个体组成的匹配集，以及一个独立的训练集；一个正实数卡尺 $\Gamma_n > 0$。

输出： 协变量值$\mathbf{x}$ 和处理水平 t对应的条件响应函数估计量。

阶段 1： 使用独立的训练集，构建表示$\frown$ $\frown\phi(\cdot)$的估计量，记为 $\phi(\cdot)$。计算匹配集中所有个体 i 的 $\phi(\mathbf{x}_i)$，以及输入$\widehat{\ }$值 $\phi(\mathbf{x})$。

阶段 2： 从匹配集中选择处理水平符合期望值且与$\mathbf{x}$的距离小于 Γ_n 的个体，构成匹配组:

$$\texttt{MG}(\widehat{\phi}, \mathbf{x}, t) = \left\{ i = 1, \ldots, n :\ D^q_{\widehat{\phi}}(\mathbf{x}, \mathbf{x}_i) \leq \Gamma_n,\ T_i = t \right\}. \tag{1}$$

第3阶段： 构造估计量:

$$\hat{\mu}(\mathbf{x}, t) = \frac{1}{|\texttt{MG}(\widehat{\phi}, \mathbf{x}, t)|} \sum_{i \in \texttt{MG}(\widehat{\phi}, \mathbf{x}, t)} y_i. \tag{2}$$

在第1阶段，我们将数据拆分为训练集和匹配集，并在训练集上学习函数 ϕ 的估计量，然后用它来估计匹配集中每个个体的 $\phi(\mathbf{x}_i)$。在第2阶段，我们通过最大化距离小于某个广义卡尺 Γ_n 的个体加权总和来构造匹配结果。在第3阶段，我们使用第2阶段构造的匹配组来估计$\mathbf{x}$的条件响应函数（CRF）。我们通过对式（3）中定义的表示距离施加约束，来控制匹配组的规模以及纳入哪些个体：与目标值的距离小于 Γ_n 的个体将纳入匹配组。Γ_n 的定义方式产生了 M-ML 算法的两个变体:

卡尺 M-ML: 在这种情况下， Γ_n 被定义为分析师选定的某个正实数 e

。这种技术称为卡尺匹配（例如，Rosenbaum and Rubin，1985b）。这种技术保证匹配组内个体之间的距离不会小于一个已知值。请注意，在卡尺 M-ML 中，可能会出现$\widehat{}$在 $\texttt{MG}(\phi,\ \mathbf{x}\ t)$ 中没有个体的情况：此时无法构造估计量；我们应使用更大的 Γ_n 值，以便能够构造匹配结果。

KNN M-ML：在这种情况下，$\Gamma_n = D^q_{\widehat{\phi}}(\mathbf{x},\mathbf{x}_{(k_n)})$，其中（$k_n$）是向量（$D^q_{\widehat{\phi}}(\mathbf{x},\mathbf{x}_1)\ldots,D^q_{\widehat{\phi}}(\mathbf{x},\mathbf{x}_n)$）的第 k_n^{th} 个顺序统计量。这种技术称为 K近邻匹配（例如，Rubin，1976）。该方法保证每个匹配组恰好包含k_n 个体（假设不存在并列值）；不过在这种情况下，卡尺 Γ_n 将是x和数据的函数，这意味着用户无法直接控制其取值。

我们将在下一节说明，在对 Γ_n 和 k_n作出适当选择的情况下，这两个 M-ML算法变体在渐近意义上基本等价。

在第3阶段，利用前一步得到的匹配结果，为 $\mu(\mathbf{x}\ t)$构造一个典型匹配估计量。直观而言，可以通过运行两次 M-ML算法来构造 CATE 估计量 $\tau(\mathbf{x}\ t,t')$：第一次以x 和 t 为输入，得到$\hat{}\mu(\mathbf{x}\ t)$；第二次以x和 t' 为输入，得到$\hat{}\mu(\mathbf{x}\ t')$；最后计算两者之差：$\hat{}\tau(\mathbf{x}\ t,t')\ =\hat{}\mu(\mathbf{x}\ t)\ -\hat{}\mu(\mathbf{x}\ t')$. 我们将在后续章节看到，这个估计量具有与使用 M-ML 构造的 $\hat{}\mu(\mathbf{x}\ t)$ 相同的理想渐近性质。

4 匹配机器学习：渐近性质

现在我们研究 M-ML 估计量的统计行为。这一点很重要，因为这样不仅能让我们给出随着 n 增大而变化的估计误差界，还能让我们为 CATE 和 ATE/ATT 估计值构造近似置信区间。以这种方式量化不确定性，对于几乎所有匹配的科学应用都至关重要，也是提高匹配估计结果可信度的必要条件。由于 M-ML 是非参数的，我们着重确立其

渐近性质，并通过模拟从经验角度研究其有限样本行为。本节将证明，对于数据分布定义域中的任意 x，CRF 和 CATE 估计量都能以 Stone（1982）意义下的非参数最优速率实现一致且有效的估计。我们将在第 5 节看到，达到这一速率后，可以将 M-ML 估计量作为输入应用双重机器学习方法，从而得到根号 n 渐近正态的 ARF、ATE 和 ATT 估计值。由这些$\widehat{\ }$结果可知，对 $\widehat{\mathrm{MG}}(\phi, \mathrm{x}\ t)$ 应用常规样本方差估计量，可以一致地估计$\mathbb{V}[Y(t)|\mathrm{X} = \mathrm{x}]$, ，因此能够为 M-ML 的 CRF 和 CATE 估计量构造近似置信区间。我们将看到，随着 n 增大而缩小卡尺 Γ_n 或增大匹配数量 k_n，是实现 M-ML 估计量一致性的关键。另一个重要的结果背后的直觉是，第一阶段估计量不会影响匹配估计量的渐近行为，只会影响其收敛速率；从渐近角度看，匹配估计量可以理解为基于 ϕ 的真实值进行匹配，这一点将在下文证明。下述结果的所有证明均见补充材料。关键假设是相对于 D_ϕ^q 满足以下条件：A5（Lipschitz条件）：对于所有 x、z $\in \mathbb{X}$ 和 $t \in \{1, \ldots, M\}$，存在常数C_L，使得：(a) $|\mu(\mathrm{x}\ t) - \mu(\mathrm{z}\ t)| \leq C_L D_\phi^q(\mathrm{x}, \mathrm{z})$，且 (b) $|\sigma^2(\mathrm{x}\ t) - \sigma^2(\mathrm{z}\ t)| \leq C_L D_\phi^q(\mathrm{x}, \mathrm{z})$。对结果函数作出这一（或类似的）光滑性假设，是几乎所有与匹配类似的非参数估计框架中的常见做法（例如，Kallus，2020；Farrell等人，2021；Wager和Athey，2018）。但这里的关键区别在于，我们希望该假设适用于变换后的协变量，而不一定适用于原始协变量。对原始协变量作光滑性假设（这是匹配和非参数方法中的常见做法）要强得多：只要 ϕ 关于协变量值是 Lipschitz连续的（对于许多 ϕ 的选择而言，这是一个简单且普遍满足的要求），就能直接推出变换后的协变量也满足该条件。我们框架的最后一个组成部分是采用灵活的机器学习方法，根据数据估计 ϕ 。为此，我们引入一个单独的训练集，其样本量为 ρn，占样本的比例为 $\rho \in (0, 1)$。

该训练集可以通过将整个数据随机划分为两组来获得：一组训练集和一组匹配集。为简化记号，我们假设个体总数为 $n+\rho n$。接下来假设将一种机器学习方法应用于训练$\widehat{}$数据，以构造 ϕ 的估计量，记为 ϕ。实际上， ϕ 可以建模为函数空间上某个总体损失函数的$\widehat{}$极小化者，而 ϕ 则是其经验$\widehat{}$对应形式，但情况不一定总是如此。为了使我们的理论结果成立，我们对 $\phi\widehat{}$ 所需的条件是： ϕ 是 ϕ 的一致估计量，并且二者都是可微函数。为了确立 M-ML估计量的渐近性质，我们还需要对第一阶段距离度量估计量作出一项额外但合理的假设:

A6（表示函数）： 存在一个函数 $\widehat{\phi}(\mathbf{x}\ \mathcal{O}_n)$: $\mathbb{X}\times\Omega\ \mapsto\ \mathbb{R}^d$ 使得$\widehat{}$对于实值 $r_{ML}>0$:
（a）函数 $\phi(\mathbf{x}\ \mathcal{O}_n)$ 和 $\phi(\mathbf{x})$ 关于$\mathbf{x}$在所有$\mathbf{x}\in\mathbb{X}$处均以概率 1 连续（ $f_{\mathbf{O}}$-几乎处处），（b） $\|\phi(\mathbf{x}\ \mathcal{O}_n)-\phi(\mathbf{x})\|_{\mathcal{P},\ q}=o(n^{-r_{ML}})$ 在 $f_{\mathbf{X}}$ 上几乎处处成立$\widehat{}$，（c） $\|\phi(\mathbf{X}\ \mathcal{O}_n)\ -\phi(\mathbf{X})\|_{\mathcal{P},\ q}=o(n^{-r_{ML}})$。

该假设的 a) 部分将可能的表示限制为关于$\mathbf{x}$的连续可微函数，这是大多数机器学习算法的常见性质。该假设的 b) 部分$\widehat{}$指出，用于学习 ϕ 的过程$\widehat{}$必须使所得量对于所有可能的输入都具有适当的分布。只要 ϕ 是通过独立训练样本学习得到的，c) 部分就成立。最后，该假设的 d) 部分指出， ϕ的第一阶段机器学习估计必须在点态和均方意义上都收敛于 ϕ 的真实值。这一假设在非参数两阶段估计设定中相当标准（Chernozhukov 等，2018），并已针对多种不同的机器学习方法得到验证，例如 LASSO（Belloni等人，2014）、随机森林（Wager和Athey，2018）、支持向量机（Devroye等人，2013）和深度神经网络（Farrell等人，2021）；研究还表明，在相对温和的假设下，这些方法都以速率 $r\geq 1/4$ 收敛。

考虑一个数据之外的个体，其观测到的协变量为 $\mathbf{x}$。我们的数据中任一个体与 $\mathbf{x}$ 匹配的概率由下列引理给出。

引理 1. 设 *A1-A6* 成立。对于 $\mathbf{x} \in \mathbb{X}$，使用卡尺 *M-ML* 进行匹配，且固定 $\Gamma_n \geq 0$，即：$\mathtt{MG}(\phi, \mathbf{x}\ t) = \{i = 1, \ldots, n\colon\ D^q_{\widehat{\phi}}(\mathbf{X}_i, \mathbf{x}) \leq \Gamma_n, T_i = t\}$。那么，如果 $\Gamma_n \to 0$ 当 $n \to \infty$ 时，对于任意 $i \in 1$，...，n，有：

1. $\Gamma_n^{-d}\mathrm{Pr}_{\mathcal{O}_n,\ \mathbf{x}_i,\ T_i}(i \in \mathtt{MG}(\widehat{\phi},\ \mathbf{x}\ t)) \to V_d e(t) f_{\phi(\mathbf{X})|T=t}(\phi(\mathbf{x}))$ 对于所有 $\mathbf{x} \in \mathbb{X}$，

2. $\Gamma_n^{-d}\mathrm{Pr}_{\mathbf{x},\ \mathcal{O}_n,\ \mathbf{x}_i,\ T_i}(i \in \mathtt{MG}(\widehat{\phi},\ \mathbf{X}\ t)) \to V_d e(t) \mathbb{E}_{\mathbf{X}}[f_{\phi(\mathbf{X})|T=t}(\phi(\mathbf{X}))]$，

其中 $e(t) = \Pr(T = t)$，$V_d = \frac{2Ga}{-Ga}\left(\frac{2}{q}+1\ d\right)\left(\frac{d}{q}+1\right)$，其中 Ga 是伽马函数，$f_{\phi(\mathbf{X})|T=t}(\phi(\mathbf{x}))$ 是 $\phi(\mathbf{X})$ 在给定 $T = t$ 条件下的概率密度函数。

请注意，第一个陈述涉及关于 $f_{\mathbf{X}}$ 的几乎处处收敛，而第二个陈述涉及关于相同分布的期望收敛。上述结果既直观又有趣：其证明完全不依赖机器学习收敛速率 r，而是 $\widehat{\ }$ 利用了 ϕ 的连续性，并结合广义变量变换 $\widehat{\ }$ 来证明该结果。该定理表明，$\mathrm{Pr}(i \in \mathtt{MG}(\phi,\ \mathbf{x}\ t))$ 渐近正比于 Γ_n^d：这有一个重要推论，即渐近地，我们的卡尺 Γ_n 收缩速率是决定与任意单位 i 匹配概率的唯一相关因素，而 ϕ，d 的维度将影响这一速率，无论原始协变量数量多少。

建立这一结果后，可以利用它以及假设 A4 和 A5 中定义的常数，证明采用 M-ML 获得的 CRF 估计量的主要理论结果：

定理 *1*。（卡尺 *M-ML CRF* 估计量的渐近行为）设 *A1-A6* 均成立，且 *A5* 对某个 $\delta > 0$ 成立。设 $e(t)$、V_d 和 $f_{\phi(\mathbf{X})|T=t}(\phi(x))$ 按照引理 *1* 中的定义给出。设 $r = min\left(\frac{1}{2+d}, r_{ML}\right)$。对于采用卡尺 *M-ML* 进行的匹配结果，有：（*i*）对于卡尺 $\Gamma_n = Kn^{\frac{2r-1}{d}}$：$n^r(\tilde{\mu}(x\ t) - \mu(x\ t)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2(x\ t)}{K^d V_d e\ t}\left(\overline{f_{\phi(\mathbf{X})|T=t}(\phi(x))}\right)\right)$·（*ii*）对于卡尺 $\Gamma_n = n^{-\frac{1}{2+d}}$ 和 $s = 2 + \delta$：$\|\tilde{\mu}(X\ t) - \mu(X\ t)\|_{\mathcal{P},s} = \overline{O}(n^{-\frac{1}{2+d}}) + o(n^{-r_{ML}})$，且该界在所有可能的 Γ_n 取值中最小。

KNN M-ML 也有相同的结果：以下定理证明，在满足 k_n 适当条件时，采用该方法得到的 M-ML CRF 估计渐近地

等价于直接控制 Γ_n 得到的估计量。

定理 2. （*KNN M-ML CRF* 估计量的渐近行为）

假设 *A1-A6* 成立。令 $\mathtt{MG}(\phi, \mathbf{x}\ t) = \{i = 1, \ldots, n : \quad D^q_{\widehat{\phi}} \leq \Gamma_n, T_i = t\}$，其中 Γ_n 等于向量 $(D^q_{\widehat{\phi}}(\mathbf{x}, \mathbf{X}_i) \ldots, D^q_{\widehat{\phi}}(\mathbf{x}, \mathbf{X}_n))$ 的k_n^{th}阶顺序统计量（注意，此时 Γ_n 是 $\mathbf{x}$ 和 $\mathbf{X}_1, \cdots \mathbf{X}_n$的函数），其中 k_n 为正整数。令 $r = \min\left(\frac{1}{2+d}, r_{ML}\right)$。我们有：

(i) 对于 $k_n = \lfloor Kn^{2r} \rfloor$，存在正常数 $K > 0$: $n^r(\hat{\mu}(x\ t) - \mu(x\ t)) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2(x\ t)}{K}\right)$.

(ii) 对于 $k_n = \lfloor n^{\frac{2}{2+d}} \rfloor$ 和整数 $s > 2$: $\|\hat{\mu}(x\ t) - \mu(x\ t)\|_{\mathcal{P},s} = O(n^{-\frac{1}{2+d}}) + o(n^{-r_{ML}})$，且此界在 k_n的所有可能取值中为最小值。

引理 1 为我们提供了一种方法，用于确立定理 1 中匹配组大小 $|\mathtt{MG}(\phi, \mathbf{x}\ t)|$ 的渐近阶，以及定理 2 中 Γ_n 的渐近阶，因此在这两个定理的证明中都至关重要。这两个定理的一个重要推论是，Γ_n 和 k_n 的选择会直接影响 M-ML估计量的收敛速率和渐近方差。对于卡尺 M-ML，采用与定理 1 证明中相同的推理，通过设定 $\Gamma_n = \left(n^{r-1} K^d V_d e_{(t)} f_{\phi(\mathbf{x})|T=t}(\phi(\mathbf{x})) \sigma^2(\mathbf{x}\ t)\right)^{\frac{1}{p}}$，可以使渐近方差等于 1；将上述值乘以任意正数，则可以使渐近方差任意小。显然，这种方差的降低是以偏倚增大为代价的，因为匹配组更大，因此在大多数情况下，将渐近方差降至 1 以下并不切实际。KNN M-ML 也存在类似关系：此时，选择 k_n 的一种方法是将其设为最接近 $n^r K$ 的整数。在这种情况下，选择 $K = 1$ 时，渐近方差恰好为 $\sigma^2(\mathbf{x}\ t)$；然而，具体取决于数据，针对 $\mathbf{x}$ 可能没有 n^r 个高质量匹配结果，因此可能必须选择 $K < 1$，从而导致渐近方差更大。我们指出，关于 K近邻 回归渐近正态性的结果，通过纳入距离度量学习步骤，推广了 Stute 等人（1984）的结果。

定理1和定理2给出的界有两个重要推论。首先，KNN匹配和卡尺匹配的最优收敛速率相同，这确定了这两种长期沿用的匹配程序渐近等价。其次，我们的

界与多项研究中关于在原始协变量的 L_q 距离上进行匹配时 KNN 分类和回归收敛速率的结果直接相关，这些结果由 Gy¨orfi (1981)；Gy¨orfi 等人 (2002) 总结。值得注意的是，我们的界与其他关于非变换匹配的界之间的主要差异是相差一个 $o(n^{-r_{ML}})$ 因子，这是因为在我们的情形中必须学习协变量变换。这一事实对我们的设定有重要影响：与基于协变量原始值进行匹配相比，通过学习距离度量所带来的性能提升只能在有限样本中实现，而非渐近地实现。Rimanic 等人 (2020) 的近期研究也支持这一结论；他们在对表示函数 ϕ 采用不同假设的情况下，为 KNN 回归推导出了类似的界。我们将在模拟部分展示，这些有限样本下的收益相当显著。更重要的是，该界意味着通过机器学习增加距离学习步骤来大幅提高匹配的预测准确性，几乎不会损害收敛速率，因为误差界中匹配部分以其非参数速率递减。

接下来讨论潜在结果 $\sigma^2(\mathbf{x}\ t)$ 的渐近方差，这对于渐近推断十分重要。也可以通过 M-ML 估计该量：对采用 M-ML 构造的匹配组应用传统方差估计量即可。以下定理对此作出正式表述：

定理 3。（样本方差估计量的一致性）设 $\sigma^2(x\ t)$ 的样本方差估计量定义为：$\hat{\sigma}^2(x\ t) = \frac{1}{N\hat{\phi}}(\ ,x\ t)\sum_{i\in \mathtt{MG}(\hat{\phi},x\ t)}(Y_i - \hat{\mu}(x\ t))^2$.。设 *A1-A6* 成立，$\hat{}$ 且 $\mathtt{MG}(\phi,x\ t)$ 采用卡尺 *M-ML* 或 *KNN M-ML* 构造。则对所有 t，当 $n \to \infty$ 时，有：$\hat{\sigma}^2(x\ t) \xrightarrow{p} \sigma^2(x\ t)$。

请注意，无论是选择匹配到$\mathbf{x}\hat{}$的个体数量，还是选择 Γ_n 来控制 $\mathtt{MG}$的半径（ϕ，$\mathbf{x}\ t$），这一结果都成立。

最后，刚才引入的定理可直接推出以下推论，该推论允许我们构造 CATE 的渐近置信区间。

推论 *1*。（*CATE* 估计量的渐近正态性）设 *A1-A6* 成立，且令 $r = min\left(\frac{1}{2+d}, r_{ML}\right)$。对于两个处理水平 t, $t' \in \{1, \ldots, M\}$ 以及一个实数 $K > 0$：当 $n \to \infty$ 时：（*i*）若使用卡尺 *M-ML* 进行匹配，且 $\Gamma_n = Kn^{\frac{2r-1}{p}}$：$n^r(\hat{\tau}(x\ t, t') - \tau(x\ t, t')) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2(x\ t)}{K^d V_d e(t) f_{\phi(\mathbf{X})|T=t}(\phi(x))} + \frac{\sigma^2(x\ t')}{K^d V_d e(t') f_{\phi(\mathbf{X})|T=t'}(\phi(x))}\right)$。（*ii*）若使用 *KNN M-ML* 进行匹配，且 $k_n = \lfloor n^{2r} K \rfloor$：$n^r(\hat{\tau}(x\ t, t') - \tau(x\ t, t')) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2(x\ t)}{K} + \sigma^2(x\ t')\right)$。

该推论之所以成立，是因为只有观测到处理 $T_i = t$ 的个体被用于构建$\hat{\mu}$(x t)，而只有具有 $T_i = t'$ 的个体被用于构建$\hat{\mu}$(x t')：这使得两个估计量彼此独立。由于这两个估计量相互独立，对向量 ($\hat{\mu}$(x t), $\hat{\mu}$(x t')) 应用连续映射定理，便可得出该推论中的结论。该推论的结果意味着，对于固定的 x，可以构建近似的 $1-\alpha$ 置信区间：CI(x t, t') $= \left[\hat{\tau}(\right.$ x t, t') $\pm \Phi^{-1}(1-\alpha/2)$ $\sqrt{\hat{\sigma}^2$(x t) c(x t) $+ \hat{\sigma}^2$(x t') c(x t') $\left.\right]$，其中，对于卡尺匹配（定理 1），c(x t) $= n^r K^d V_d e(t) f_{\phi(\mathbf{X})|T=t}(\phi$(x)$)$，或者对于K近邻匹配（定理 2），$c$(x t) $= k_n$。这将使分析人员能够以一种被广泛接受且具有已知保证的方式，量化其 CATE 估计量周围的不确定性。

5 匹配双重机器学习：针对平均效应估计量

M-ML 框架可以自然地扩展到对平均处理效应和对处理组的平均处理效应进行非参数估计，同时保留匹配的可审计性，以及构造渐近有效置信区间的能力。这是可行的，因为 M-ML估计量可以作为增广逆倾向加权（AIPW）估计量第一阶段估计量的输入（Robins 等，1994）。研究发现，只要条件响应函数和倾向得分的第一阶段估计量收敛得足够快，这些估计量在渐近意义下服从正态分布，且方差已知（Chernozhukov 等，2018；Van Der Laan 和 Rubin,

2006年），正如上述分析所示，我们的方法具备这一性质。

得益于这一性质，我们可以仿照 Chernozhukov 等人（2018）的 DML 算法，构建一个用于估计平均处理效应和对处理组的平均处理效应的 M-ML 算法版本。我们提出的算法称为匹配双重机器学习（M-DML），定义如下：

平均处理效应或对处理组的平均处理效应估计：匹配双重机器学习（M-DML）

阶段 1： 将数据 $\{\mathbf{x}_i,\ y_i,\ t_i\}_{i=1}^n$ 随机划分为 L 折，每折大小为 n/L。令 S_ℓ表示第 ℓ折中所有个体的索引，$S_{\setminus\ell}$ 表示不在该折中的所有个体的索引。第 ℓ 折中的个体用于估计平均处理效应和对处理组的平均处理效应。

对 $\ell = 1$、.....、 L重复以下步骤。

阶段 2： 将 $S_{\setminus\ell}$ 中的个体进一步划分为训练集（记为 $\mathcal{TS}_\ell$ ）和匹配集（记为 $\mathcal{MS}_\ell$）。仅使用 $\mathcal{TS}_\ell$中的个体学习表示函数 $\phi(\mathbf{x})$。使用 $S_{\setminus\ell}$中的所有个体，为每个处理水平 t 构建倾向得分的一致估计量，记为 $\hat{e}(\mathbf{x}\ t)$；并构建接受处理的边际倾向概率的一致估计量 t: $\Pr(T = t)$，后者记为 $\hat{e}(t)$。

第3阶段： 对于每个单位 $i \in S_\ell$ 和处理水平 $t,\ t'$，针对每个处理水平运行 M-ML算法，并以 $\mathcal{MS}_\ell$ 中的所有个体作为匹配候选，以获得 $\hat{\mu}(\mathbf{x}_i,\ t)$、$\hat{\mu}(\mathbf{x}_i,\ t')$。此外，使用第2阶段学习得到的倾向模型，预测 i 在每个处理水平下的倾向得分 $\hat{e}(\mathbf{x}_i,\ t)$。

第3阶段：对于每个单位 $i \in S_\ell$，利用上一阶段的输出构造双重稳健得分函数：$\widehat{\psi}(\mathbf{x}_i,\ t) = \hat{\mu}(\mathbf{x}_i,\ t) + \frac{\mathbb{I}[t_i=t]}{\hat{e}(\mathbf{x}_i,\ t)}(y_i - \hat{\mu}(\mathbf{x}_i,\ t))$，并以类似方式计算 $\widehat{\psi}(\mathbf{x}_i,\ t')$，然后计算：$\widehat{\psi}(\mathbf{x}_i,\ t,\ t') = \frac{\mathbb{I}[t_i=t]}{\hat{e}(t)\hat{e}(\mathbf{x}_i,\ t)}(y_i - \hat{\mu}(\mathbf{x}_i,\ t')) - \frac{\mathbb{I}[t_i=t']}{\hat{e}(t')\hat{e}(\mathbf{x}_i,\ t')}(y_i - \hat{\mu}(\mathbf{x}_i,\ t'))$。

第4阶段： 构造估计量：对于 t,t'，分别构造 $\hat{\mu}_\ell^{DR}(t) = \frac{L}{n}\sum_{i\in S_\ell}\widehat{\psi}(\mathbf{x}_i, t)$、$\hat{\tau}_{ATE}^{DR}(t,t')_\ell = \hat{\mu}_\ell^{DR}(t) - \hat{\mu}_\ell^{DR}(t')$、$\hat{\tau}_{ATT}^{DR}(t,t')_\ell = \frac{L}{n}\sum_{i\in S_\ell}\widehat{\psi}(\mathbf{x}_i, t, t')$。

第5阶段： 对各折取平均：$\hat{\mu}^{DR}(t) = \frac{1}{L}\sum_\ell \hat{\mu}_\ell^{DR}(t)$、$\hat{\tau}_{ATE}^{DR}(t,t') = \frac{1}{L}\sum_\ell \hat{\tau}_{ATE}^{DR}(t,t')_\ell$，以及 $\hat{\tau}_{ATT}^{DR}(t,t') = \frac{1}{L}\sum_\ell \hat{\tau}_{ATT}^{DR}(t,t')_\ell$。这三个量就是 M-DML 的输出。

该算法 w　　通过将数据划分为 L 折（第1阶段），然后再进一步　　r

划分为训练集和匹配集。请注意，这是三路数据划分，这直观上是匹配步骤在机器学习预测步骤之上增加后所必需的；Wang 等人（2021）也采用了类似的划分流程。随后，对每一折分别拟合 M-ML 算法（第2阶段），并对结果取平均，以获得目标参数的最终估计值（第5阶段）。请注意，如果数据量足够，就可以只使用一个留出的训练集，无需再将折 ℓ 之外的个体进一步划分为训练集和匹配集。这样可以避免增加交叉拟合的复杂性，因为对于大型数据集而言，交叉拟合的计算成本可能很高。

给定用户感兴趣的两个处理水平，t 和 t'，M-ML算法在每一折上运行两次，分别针对处理水平 t 和 t'。使用 M-ML 输出的 $\hat{\mu}(\mathbf{x}\ t)$ 构造的估计量见公式（5）和（5），它们是双重稳健得分函数，用于 Robins 等人（1994）的增广逆概率加权估计量。这些估计量的思想是利用观测数据校正第一阶段预测，从而确保估计量具有渐近正态性。如算法所示，需要为所有个体 i 以及感兴趣的处理水平，获得倾向得分 $e(\mathbf{x}_i, t)$ 的一致估计量。对于这一量，存在许多良好的估计量；使用任何基于满足定理 4 中给定收敛速率要求的机器学习方法的估计量，都将使定理中关于 M-DML 的渐近保证成立。

M-DML算法是 Chernozhukov 等人（2018）定义 3.2 中 DML2 算法的一个特例。为使 M-DML 与该定义相对应，我们将 M-ML 作为定义中的第一阶段估计量。结合增强逆概率加权估计量的性质以及 M-ML 的收敛速率，我们可以保证 M-DML 具有渐近正态性和相应的收敛速率。以下定理对此作出说明。

定理 *4*。设观测到的结果变量和处理定义如下：$Y = \mu(X\ D) + U,\quad \mathbb{E}[U|X\ D] = 0$，以及 $D = e(X\ D) + V,\quad \mathbb{E}[V|X] = 0$。设 *A1-A6* 成立，并进一步假设用户选择的倾向得分估计量满足：对于某个正实数 r_e，*(i)* $\|\hat{e}(X\ t) - e(X\ t)\|_{\mathcal{P},2} = O(n^{-r_e})$，*(ii)* $\|\hat{e}(X\ t) - 1/2\|_{\mathcal{P},\infty} \leq 1/2 - \epsilon$，其中某个

$\epsilon > 0$, ***(iii)*** $0 < \hat{e}(\mathbf{x}\ t) < 1$ 对所有 $\mathbf{x}$ 以及 t 均成立。设 $r =$最小值 $\left(\frac{1}{2+d}, r_{ML}\right)$，并假设 $r + r_e \geq 1/2$。令 $\mathtt{MG}(\phi, \mathbf{X}\ t)$ 通过卡尺 ***M-ML*** 或 ***KNN M-ML*** 构造，其中采用 $\Gamma_n = Kn^{\frac{2r-1}{d}}$ 或 $k_n = Kn^{2r}$，固定整数 $K > 0$。则 ***M-DML*** 估计量满足以下结论：

1) 渐近正态性：$\sqrt{n}(\hat{\mu}^{DR}(t) - \mu(t)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(t))$，其中 $\sigma^2(t) = \mathbb{E}_{\mathbf{X}}[\psi(\mathbf{X}\ t)^2]$，$\sqrt{n}(\hat{\tau}^{DR}_{ATE}(t,t') - \tau(t,t')) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{ATE}(t,t'))$，其中 $\sigma^2_{ATE}(t,t') = \mathbb{E}_{\mathbf{X}}[(\psi(\mathbf{X}\ t) - \psi(\mathbf{X}\ t'))^2]$，且 $\sqrt{n}(\hat{\tau}^{DR}_{ATT}(t,t') - \tau(t,t')) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{ATT}(t,t'))$，其中 $\sigma^2_{ATT}(t,t') = \mathbb{E}_{\mathbf{X}}[\psi(\mathbf{X}\ t,t')^2]$。

2) 应用于第二阶段估计量的样本方差估计量的一致性，即：$\hat{\sigma}^2(t) = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\psi}(X_i, t) - \hat{\mu}^{DR}(t))^2 \xrightarrow{p} \sigma^2(t)$，$\hat{\sigma}^2_{ATE}(t,t') = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\psi}(X_i, t) - \widehat{\psi}(X_i, t') - \hat{\tau}^{DR}_{ATE}(t,t'))^2 \xrightarrow{p} \sigma^2_{ATE}(t,t')$，以及 $\hat{\sigma}^2_{ATT}(t,t') = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\psi}(X_i, t, t') - \hat{\tau}^{DR}_{ATT}(t,t'))^2 \xrightarrow{p} \sigma^2_{ATT}(t,t')$。*3)* 近似置信区间：

令 Δ 为 ***M-DML*** 算法第*5*阶段的 ***M-DML*** 估计器，Δ 为其对应的估计目标，$\hat{\sigma}^2$ 为其相应的渐近方差。目标参数的近似 $1-\alpha$ 置信区间为：$CI(\delta) = \left[\widehat{\Delta} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{\hat{\sigma}^2}{n}}\right]$，其中 $\Phi^{-1}(a)$ 为标准正态分布的 a^{th}分位数。

该结论几乎可直接由我们的定理 1、2 以及 Chernozhukov 等人（2018）的定理 5.1 得出。该结果证明，使用 M-DML 得到的平均处理效应和对处理组的平均处理效应估计量以 $\sqrt{n}$ 速率满足渐近正态性。这一点至关重要，因为我们可以利用估计量的渐近分布，近似计算感兴趣的平均参数的置信区间，从而提供因果推断所需的不确定性量化。请注意，如果对 ϕ 和 $\hat{e}$使用相同的机器学习估计量，那么对其速率的要求变为：$r_{ML} \geq \frac{1}{4}$，这与 Chernozhukov 等人（2018）给出的粗略要求相同。请注意，在 $\beta \geq \frac{p}{2}$ 的条件下，M-ML 和第一阶段方法可以达到这一速率；也就是说，如果结果变量是协变量的充分光滑函数，这也是非参数估计框架中的常见要求。为了达到所需速率，表示的维度 ϕ 也必须选择得当，使条件 $\frac{1}{2+d} + r_e \geq \frac{1}{2}$ 成立。这

可通过选择 $d \leq \frac{2r_e}{\frac{1}{2}-r_e}$ 来实现。如果 r_e 是具有 β-光滑倾向得分和 p 个协变量时的最优非参数收敛速率（Stone，1982），那么该条件可简化为 $d \leq \frac{4\beta}{p}$。

6 M-ML 和 M-DML：示例与扩展

本节介绍 M-ML 和 M-DML 的一些实际应用示例，并说明这些算法在特定情境下需要考虑的一些因素。此外，我们还介绍了 M-ML 算法的一项扩展，用户可借此在 M-ML 算法第 2 阶段的匹配优化问题中添加任意约束。

6.1 用匹配解释机器学习预测

M-ML 最直接的应用，是根据黑箱机器学习模型估计的潜在结果变量进行匹配，从而通过基于案例的推理审计模型的预测。在因果推断文献中，根据 $\mu(\mathbf{x}\ t)$ 的估计值进行匹配称为预后评分匹配（Hansen，2008），它具有许多倾向得分匹配的性质。在这种情况下，我们将 ϕ 定义为 $\min_h \mathbb{E}[\ell(Y(t), \mathbf{X}\ h)],$，其中 ℓ 是损失函数。例如，对于回归问题，可以使用损失 $\ell(Y(t), \mathbf{X}\ h) = (Y(t) - h(\mathbf{X}))^2$。不难证明，在这种情况下，$\phi(\mathbf{x}) = \mathbb{E}[Y(t)|\mathbf{X} = \mathbf{x}] = \mu(\mathbf{x}\ t)$。传统机器学习方法会通过在假设函数空间 $\mathcal{H}$ 上最小化经验风险 $\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i(t), \mathbf{X}_i, h)$ 来估计 ϕ。这个空间可能非常庞大、复杂且难以解释，因此几乎不可能审计。正如前文所述，匹配可以通过将黑箱 $h(\mathbf{x}\ t)$ 的输出替换为邻近个体观测结果变量的平均值，来弥补可解释性的不足；如果我们构造的距离度量得当，这些个体的预测 $Y(t)$ 值就会相近。重要的是，在这种情况下，$d = 1$，这意味着 M-ML 的收敛速率为 $\min(n^{\frac{1}{3}}, n^{r_{ML}})$，而这几乎总是等于 r_{ML}，因为

任何机器学习方法都极不可能达到高于 $1/3$ 的收敛速率，尤其是在 $p > 1$ 的情况下。因此，在机器学习之上增加匹配以实现可审计性，几乎不会影响机器学习预测的收敛速率。值得注意的是，在这种情况下， ϕ 的维度恰好为 $d = 1$。这意味着，我们的估计量中匹配部分的收敛速率恰好等于非参数最优速率（参见 Stone（1982）以及 Gy¨orfi 等人（2002）第 6.3 节）。这意味着，在这种情况下，CATE M-ML 估计量的收敛速率为 $r = r_{ML}$；也就是说，M-ML 估计量的收敛速度与任何后端机器学习估计量对 ϕ 的估计一样快。这一点有着重要意义：在机器学习预测之上加入基于匹配的可解释性层，不会影响收敛速率。

6.2 使用学习得到的权重进行马氏距离匹配

M-ML 的一种潜在用途是构造协变量权重，用以描述每个特征对于结果生成的重要性，然后使用这些权重计算加权 L_2 距离并据此进行匹配。已有多种匹配方法实现了这一做法（Wang 等人，2021；Parikh 等人，2022；Diamond 和 Sekhon，2013），它们分别针对不同的权重集合，以确保匹配结果具备不同的理想属性。在 M-ML 框架中，可以通过设定 $\phi(\mathbf{x}) = \mathbf{M}\mathbf{x}$ 来表示这种情形，其中 $\mathbf{M}$ 是一个对角权重矩阵，其权重可以是已知的，也可以从数据中学习得到。在这种情况下， $\phi(\mathbf{x})$ 可逆，而在我们的框架中用于估计卡尺匹配渐近方差的匹配渐近概率 $n^r K^d V_d e(t) f_{\phi(\mathbf{x})|T=t}(\phi(\mathbf{x}))$ 变为： $n^r K^d V_d e(t) f_{\mathbf{x}|T=t}(\mathbf{x}) \prod_{j=1}^{p} m_j$，其中 m_j是协变量 j对应权重的真实值。

6.3 M-ML 作为一种通用匹配框架

如前面的示例所示，进行放回匹配时，无论是对待匹配个体的两两距离设置卡尺，还是设定固定的匹配结果数量，M-ML 算法都可以视为若干其他流行匹配方法的推广。

本文介绍了 M-ML 的特例方法，并在表 1 中进行了总结。具体而言，以下方法都是 M-ML 框架的特例：最近邻、倾向得分匹配、预后得分匹配、预后得分匹配、自适应超盒、粗化精确匹配、遗传匹配、遗传匹配、学习拉伸后的匹配（MALTS）和精细平衡。

下面更详细地说明这些方法如何成为 M-ML 的特例，先从马氏距离匹配谈起；它包括 MALTS 和遗传匹配。马氏距离匹配（Rubin，1980年）根据马氏距离匹配彼此接近的个体。对于两个向量 u、v $\in \mathbb{R}^p$ 和一个方阵 M $\in \mathbb{R}^{p\times p}$，马氏距离定义为：Maha(u, v, M) $= \sqrt{(u - v)^T M(u - v)}$。M 通常取对角矩阵，本文也将使用对角矩阵 M。将 ϕ(x) =Mx 作为变换函数，然后依据 L_2 距离进行匹配，即可将马氏匹配实现为 M-ML。在相同情形下，令 M ~~= I，其中 I 是单位~~矩阵，就会得到简单的最近邻匹配，例如 Abadie 和 Imbens（2006年）使用的方法。遗传匹配（Diamond 和 Sekhon，2013年）和 MALTS（Parikh 等人，2022年）等方法按上述方式进行马氏匹配，但会增加一个首先学习最优 M 的步骤，因此也利用了 M-ML 的学习组件。对于在单独的训练集上学习距离的遗传匹配版本，这一步可以表示为求解以下问题，以找到 M：M $\in$ arg min $\sum_{i,j\in}$ Maha(x_i, x_j, M), $M\in\mathbb{R}^{p\times p}_{diag}$ Tr，其中 Tr 是一组索引，用于指示哪些个体属于训练集。对于 MALTS，M 通过以下优化问题求得：M $\in$ arg min $M\in\mathbb{R}^{p\times p}_{diag}$ $\left|\sum_{i\ j\in}\right.$, Tr $(y_i-y_j)\mathbb{I}(t_i=t_j)$ exp(–Maha(x_i,x_j,M)) $\sum_{k\in}$Tr $\mathbb{I}(t_i=t_k)$ exp(–Maha(x_i,x_k,M)) $\left|\right.$。

使用处理预测变量和结果变量并据此进行匹配的方法，也可以作为 M-ML 的特例来实现。对于倾向得分匹配（Rosenbaum and Rubin，1985a），ϕ(x) $= h$(x) 和 h 被选为函数类 $\mathcal{H}$中处理分配的最佳预测变量，即：$h \in$ argmin$_{h\in\mathcal{H}}$ $\mathbb{E}_{\mathbf{X}}[(h(\mathbf{X})-\Pr(T_i = t|\mathbf{X}))^2]$。对于预后评分匹配（Hansen，2008），采用相同的方法，但此时 h 是对照结果的最佳预测变量，记为 t'：$h \in$ arg min$_{h\in\mathcal{H}}$ $\mathbb{E}_{\mathbf{X},Y(t')}[(h(\mathbf{X}) - Y(t'))^2]$。自适应超盒（AHB）匹配算法（Morucci 等人，2020）的预训练版本

也可以构造为 M-ML 的一种形式：对感兴趣的两个处理水平 t, t'，采用与预后评分匹配相同的 h 形式，并构造向量：$\phi(\mathbf{x}) = (h_t(\mathbf{x}) \; h_{t'}(\mathbf{x}))$，其中 $h_t \in \arg\min_{h \in \mathcal{H}} \mathbb{E}_{\mathbf{x}, Y(t)}[(h(\mathbf{X}) - Y(t))^2],$，而 $h_{t'}$的定义方式与之类似。此外，AHB 会在从数据中学习得到的协变量空间超矩形区域内匹配个体：虽然这无法在 M-ML 中直接实现，但可以预先指定协变量空间的矩形区域 $\mathbf{H}(\mathbf{x})$，并在 M-ML 匹配组中添加约束，要求匹配单位位于该区域内：$\texttt{MG}(\phi, \mathbf{x}\; t) = \{i = 1, \ldots, n\colon \; t_i = t, D_{\hat{\phi}}^q(\mathbf{x}_i, \mathbf{x}) \leq \Gamma_n, \mathbf{x}_i \in \mathbf{H}(\mathbf{x})\}$。在估计出 ϕ之后，所有这些匹配方法都使用 L_1 距离进行匹配；M-ML 也可以采用该距离。粗化精确匹配（Iacus 等人，2012）可以通过创建一个逐维卡尺的对角矩阵 Γ，将其实现为 M-ML；矩阵的对角线由向量 $\left[\frac{1}{\gamma_1}, \ldots, \frac{1}{\gamma_p}\right]$给出，其中 γ_j的每个值都是两个单位在第 j^{th} 维上允许的最大距离。若要完全模拟 CEM，则应使用上确界范数进行匹配，即令 $q = \infty$。类似地，可以通过向匹配组添加形式为 $|x_{ij} - x_j| \leq \gamma_j$的约束，在 j^{th}协变量上实现近似精细平衡匹配（Rosenbaum 等人，2007）；其中 γ_j 是第j^{th} 维上的一个较小卡尺。将 $\gamma_j = 0$设为相应值即可实现精细平衡。精细平衡匹配可以采用任意选择的 ϕ 和 q；但若要模拟 Rosenbaum 等人（2007）所采用的实现方式，则应将 ϕ 设为倾向得分，并令 $q = 1$。

— —

最后，分析人员可能会感兴趣于通过求解某个优化问题的加权版本来构建最优匹配组；在该问题中，纳入$\texttt{MG}(\phi, \mathbf{x}\; t)$的个体数量被选为累积距离与个体数量加权组合的最优值。这是一种偏差-方差权衡，因为较大的组方差较低，但偏倚较高，这是因为它们纳入了距离更远的点。以下引理表明，式（1）中定义的匹配组构造也是此类加权问题的一个最优解：

引理 2。 *(M-ML* 是加权匹配问题的一个解。*)* 令 $W_i(\phi, \mathbf{x}\; t) =$

$\widehat{\mathbb{I}}[i \in \mathtt{MG}(\phi, x\ t)]$ 表示单位 i 是否包含在 $\mathtt{MG}(\phi, x\ t)$中，那么在 *(1)* 中定义的匹配组$\widehat{\mathtt{MG}}(\phi, x\ t)$也是该问题的一个解：$\min_{W_1(x\ t)\ \ldots, W_n(x\ t):\in\{0,1\}^n} \sum_{i=1}^{n} D_{\widehat{\phi}}^{q}(x, X_i) - \widehat{\Gamma}_n \sum_{i=1}^{n} W_i(\phi, x\ t)$。

上述表述例如见于 Morucci 等人（2020），也见于 Zubizarreta（2012）的一些优化问题中，并结合了针对匹配问题的额外约束。该引理重要地表明，M-ML 框架涵盖了以组合优化问题为目标的匹配算法，这意味着我们为 M-ML 得出的渐近结果和经验结果也可以推广到这类方法。

总之，我们已经证明，我们的方法和理论结果普遍适用于许多现有的匹配算法，这也使得能够轻松计算这些算法的渐近置信区间。

6.4 相关工作

在文献中，研究者已在各种特定且受限的情境下，探讨了依据协变量的学习函数进行匹配这一思路。在这些情境中，最早出现且得到广泛研究的是倾向得分匹配（PSM）（Rosenbaum 和 Rubin，1983年）。在关于倾向得分匹配的各类分析中，与我们的研究设定最接近的是 Rubin 和 Thomas（1992年）的研究，以及 Abadie 和 Imbens（2016年）的研究。前者表明，当协变量变换为仿射变换且结果变量服从椭球分布时，可以推导出匹配方法有限样本偏差和方差的理论保证。后者表明，仅针对平均效应，当倾向得分为线性时，倾向得分匹配确实具有有效的渐近行为。其他近期研究也考虑了基于协变量变换进行匹配，包括 Luo 和 Zhu（2020）对协变量线性变换的匹配研究，以及 Kallus（2020）提出的一种 ATT 估计方法：该方法无需额外的非参数调整，依赖协变量的核映射，并具有 $\sqrt{n}$-一致性。我们的论文更具一般性，因为我们的匹配框架

适用于协变量的任意 变换，对结果数据的分布假设要求极少，并且涵盖 ATE/ATT 和 CATE 的估计。除统计推断工具外，处理效应估计中的匹配文献还涌现出大量直接基于协变量未经变换的原始取值进行匹配的方法（例如，Iacus 等人，2012年；Diamond 和 Sekhon，2013年；Zubizarreta，2012年）。不过，这些方法大多侧重于优化个体间某种汇总质量指标，因此在估计 CATE 时表现不佳，不同于我们提出的方法；而在估计 ATE/ATT 时，它们容易受到 Abadie 和 Imbens（2006年）所述收敛问题的影响。该研究表明，最近邻匹配方法无法使 ATE/ATT 达到名义上的 $\sqrt{n}$ 收敛速率，这一问题与我们的研究设定密切相关。S¨avje（2022年）的近期研究也表明，不放回匹配方法同样无法达到这一速率。我们提出了一种基于近期有效两阶段估计结果的方法（Chernozhukov 等，2018；Van Der Laan 和 Rubin，2006年），使我们的平均估计量能够以名义速率达到一致。解决这一问题的其他方法包括 Abadie 和 Imbens（2011年）以及 Otsu 和 Rai（2017年）的研究，但这两种方法都要求将匹配与独立的结果函数和倾向得分函数非参数估计相结合，而这些估计并不涉及匹配，因此最终估计结果难以审计。另一种解决这一问题的现有方法由 Wang 和 Zubizarreta（即将发表）提出；他们表明，针对平均平衡而非每个个体的最近邻平衡进行匹配，可以达到名义速率。我们的结果和所提出的框架与他们的方法不同：本框架使用最近邻匹配结果，计算速度快于同时优化全部匹配结果（这需要通过混合整数规划实现），并且能够直接估计个体层面的处理效应，而他们的方法无法做到这一点。此外，还有一类关于个体化处理效应估计的匹配方法，将距离函数的机器学习与匹配相结合（Dieng 等人，2019年；Parikh 等人，2022年；Morucci 等人，2020；Wang 等人，2021年）：我们的论文旨在将这些方法统一到一个框架中加以推广，

并为这些方法的使用者提供一种对其输出估计进行推断的方法。最后，非参数条件平均处理效应估计的相关文献也与我们的工作有关。该领域的研究主要关注强大的黑箱方法（Chipman等人，2010年；Wager和Athey，2018年；Farrell等人，2021年），分析人员和决策者无法审计这些方法的预测，因此在许多情境下，结果的可信度显著降低，而且可能是错误的。我们的方法通过匹配明确地解决了这一问题。

7 模拟

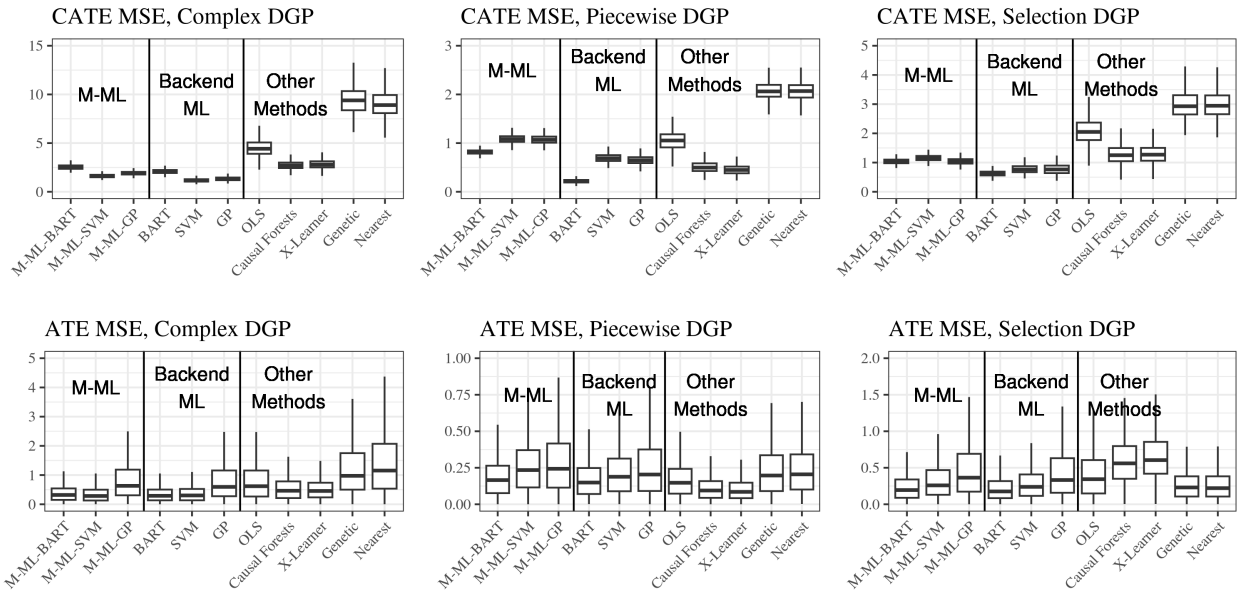
我们展示了对 M-ML 在若干已知真实因果效应的模拟数据集上估计 CATE 和平均处理效应的性能进行实证评估所得的结果。作为应用场景，我们聚焦于将 M-ML 用作机器学习的审计工具：通过黑箱算法生成的结果变量预测进行匹配，因为这是 M-ML 最自然的用途之一。我们表明，平均而言，M-ML 的性能与其所基于的黑箱方法相当，在某些设定下甚至优于这些方法。所有模拟均采用相同的总体设定：我们为 $n = 20000$个体和 $p = 20$ 个协变量生成数据，其中 5000 个体用于训练，其余个体用于匹配/估计。对于$i = 1, \ldots, n$, 我们生成：$\mathbf{X}_i \sim \mathcal{N}_p(1, 1)$ $\sigma_i \sim$均匀分布$(1, 2)$ $\boldsymbol{\lambda} \sim$均匀分布$_p(-4, 4)$ $\boldsymbol{\beta}^{lin} =\boldsymbol{\lambda}$, $\boldsymbol{\beta}^{qua} \sim$均匀分布$_p(0, 1) + \boldsymbol{\lambda}$, $\boldsymbol{\beta}^{cos} \sim$均匀分布$_p(0, 1) + \boldsymbol{\lambda}$, $\boldsymbol{\delta} \sim$均匀分布$_p(-1, 1)$ $\boldsymbol{\delta}^{int} \sim$均匀分布$_{p\times p}(-0.5, 0.5)$ $\tau = 5$, $\alpha = 5$, $\epsilon_i \sim \mathcal{N}(0, \sigma_i$其中，对于某个向量$\mathbf{u} \in \mathbb{R}^p$，$\mathbf{u} \sim f_p$表示由来自同一分布 f的 p 个抽样构成的向量。随后，我们根据以下数据生成过程生成结果变量，处理水平为 $t = 0$和 1：非线性：$Y_i(t) = \alpha+t\tau+\mathbf{X}_i\boldsymbol{\beta}^{lin}+\mathbf{X}_i^2\boldsymbol{\beta}^{qua}+\cos(\mathbf{X}_i)\boldsymbol{\beta}^{cos}+t\mathbf{X}_i\boldsymbol{\delta}+t\sum_{j=1}^{p}\sum_{k=1}^{p}X_{ij}X_{ik}\delta_{jk}^{int}+\epsilon_i$分段：$Y_i(t) = \alpha + t\tau + \sum_{j=1}^{p}\mathbb{I}(X_{ij} > 0)\beta_j^{lin} + t\sum_{j=1}^{p}\mathbb{I}(X_{ij} > 0)\delta_j + \epsilon_i$选择：$Y_i(t) = \alpha + t\tau + \sum_{j=1}^{10}X_j\beta_j^{lin} + t\sum_{j=1}^{10}X_j\delta_j + \epsilon_i$.我们选择上述数据生成过程，是因为它们模拟了三种在非参数处理中较为复杂、但可能出现在应用场景中的设定。非线性设定是

一种结果变量是协变量复杂函数的设定，其变化方式可能出乎意料；分段设定则较为简单，但将每个协变量视为一个简单的阈值，因此函数中包含阶梯式成分。最后，选择设定也是线性的，但包含变量选择成分：生成结果变量时仅使用 20 个模拟协变量中的 10 个，因此估计方法需要足够灵活，能够排除或降低未使用协变量的权重，以获得最优结果。最后，我们还生成倾向得分和处理指标，具体如下：

$u_i \sim$均匀分布$(0.1, 1)$ $e(\mathbf{X}_i, 1) = \exp(u_i(Y_i(1) + Y_i(0))/2 - \epsilon_i)\,1 + \exp(u_i(Y_i(1) + Y_i(0))/2 - \epsilon_i)$ $T_i \sim$伯努利分布$(e(\mathbf{X}_i, 1))$,请注意，潜在结果与处理分配之间的相关性由随机变量 u_i控制。我们引入该变量，以避免处理与结果完全相关，并防止可能出现的重叠性违背。我们的模拟将采用贝叶斯加性回归树、高斯过程和支持向量机作为 M-ML 的基线预测器，同时还会纳入包括因果森林在内的其他几种比较方法。所用方法的完整列表见附录中的表 2。对于模拟中采用的所有 M-ML 方法，我们均基于未加权 L2 距离进行匹配，其中 ϕ为倾向得分，或采用三种机器学习方法之一估计的潜在结果之差。我们采用 KNN M-ML 匹配，固定k_n ，并将 k_n 设为 $\lfloor n^{1/2} \rfloor$，因此，在对数据作出充分正则性假设的情况下， $\sqrt{k_n}$ 是大多数机器学习方法第一阶段收敛速率的下界（Chernozhukov 等，2018）。首先，我们展示图1顶行中的条件平均处理效应估计结果。在这一设定下，各方法的条件平均处理效应均通过其拟合的潜在结果预测值之差直接估计。该设定下的结果表明，尽管 M-ML 通常并不优于我们与之比较的非参数黑箱方法，但通常也不会逊于这些方法。这为以下观点提供了证据：在机器学习方法之上增加匹配，可以提升可解释性和可审计性，并支持不确定性量化，同时几乎不损失或完全不损失性能。结果还

表明，数据生成过程确实会影响性能：具体而言，M-ML 在非线性和选择数据生成过程中似乎表现更好，并且在这些设定下优于因果森林和 X-learner 等非参数因果方法。底行的

图1：估计误差

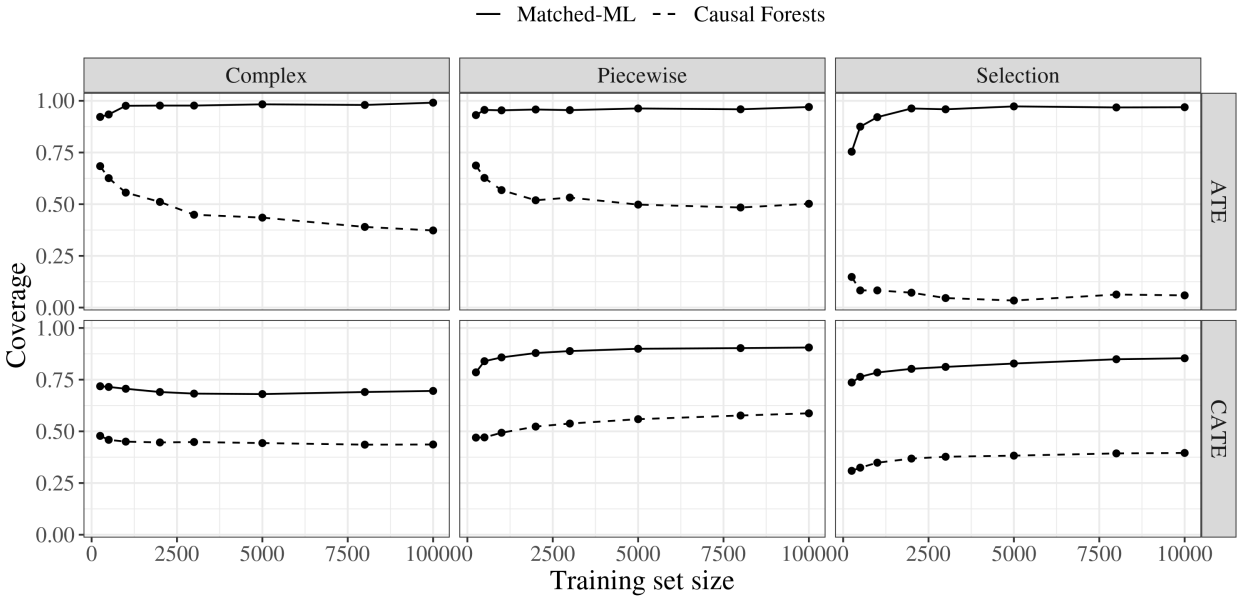


注：上排为 CATE，下排为 ATE。横轴表示所比较的不同方法，纵轴表示每次迭代中的平均绝对估计误差。各缩写的含义见表2。

图1展示了一组类似模拟的结果，但针对的是平均处理效应估计。我们使用 M-DML算法和 KNN M-ML算法来构建条件响应函数的第一阶段预测。因果森林和偏差校正匹配方法则采用原始论文中提供的估计量，以及实现这些方法的软件包。我们将结果与旨在估计平均处理效应的匹配方法（如 GenMatch）以及其他非参数机器学习方法进行比较。结果基于500轮模拟；每轮均按照之前相同的数据生成过程生成一个包含 $n = 5000$ 个体的数据集，并将其中3000个个体划分为训练集。除 ϵ，$\mathbf{X}$，和 Y 以外的参数均先行生成，并在全部500轮模拟中保持不变；其他变量则在每轮模拟中生成。我们再次发现，M-ML的表现与其他机器学习方法相当，并且在某些情形下优于其中一些方法。再次表明，基线的选择

算法和数据生成过程似乎会影响性能。最后，我们进行了一组模拟，以研究使用 M-ML和贝叶斯加性回归树机器学习后端得到的95%渐近置信区间的覆盖率。我们选择将 M-ML的覆盖率与因果森林（Wager和Athey，2018年）进行比较，因为据我们所知，这是唯一另一种能够为条件平均处理效应估计提供渐近有效置信区间的方法。我们运行与之前相同的一组模拟，但每次改变训练集的大小。对于每种训练集大小，我们从$\mathbf{X}$的分布中随机抽取250个条件平均处理效应进行估计，并计算各自的估计置信区间所覆盖的比例。每种设置下均重复这一过程，并对1000次模拟的结果取平均。结果报告在图2的底行。我们看到，在图2中，M-ML的表现远优于因果森林，在我们所有的模拟设置中覆盖率都更高。此外，我们还可以看到，M-ML仍未达到名义覆盖率水平：这是意料之中的，因为鉴于问题本身的难度，现有的条件平均处理效应估计方法也很少能达到这一水平（K¨unzel等，2019年）。接下来转向

图2：95%渐近置信区间覆盖率



平均处理效应方面，我们使用与前文相同的模拟设置，将 M-DML 的覆盖率与因果森林的覆盖率进行比较。结果见图2上排。显然，M-DML

在所有设置下均达到标称覆盖率（95%），而因果森林的表现则远逊于此。附录中的图5和图6进一步比较了 CF 与 M-DML 的区间大小。总体而言，这些结果表明，M-ML 和 M-DML 输出的区间略大于因果森林，但我们的方法所实现的覆盖率远高于因果森林，这在很大程度上足以证明区间增大的合理性。因此，我们的模拟结果表明，M-ML 能够为高精度黑箱机器学习方法增加审计能力，同时不降低性能。这适用于条件平均处理效应和平均处理效应估计。

8 应用：基于图像数据的匹配

我们将该方法应用于研究品牌对社交媒体粉丝的响应性所带来的回报。这是网络营销和消费者行为领域一个经过充分研究的问题：品牌关系与社交媒体参与之间存在周期性关系。在社交媒体上与品牌互动可以加强消费者与品牌关系（Laroche等，2013年；Labrecque，2014年）。当消费者感到品牌对其作出响应时，社交媒体参与会加强消费者与品牌关系（Labrecque，2014年）。然而，大量证据也表明，已经与品牌建立牢固关系的消费者更有可能在社交媒体上与该品牌互动（John等，2017年；Simon 和 Tossan，2018年）。因此，理解品牌对社交媒体消费者的响应性所产生的影响，显然构成了一个因果推断挑战，这正是我们试图在此解决的问题。具体而言，为了控制相关关系中潜在的混杂因素，我们根据日期/时间和评论数量等元数据匹配帖子，同时也根据所发布的图像进行匹配。根据图像进行匹配十分重要，因为帖子中图像的内容很可能会影响用户与该帖子互动的可能性。这一应用还展示了 M-M 如何实现基于图像等复杂协变量匹配个体，以及如何通过简单检查匹配结果来审计结果。

8.1 数据与方法论

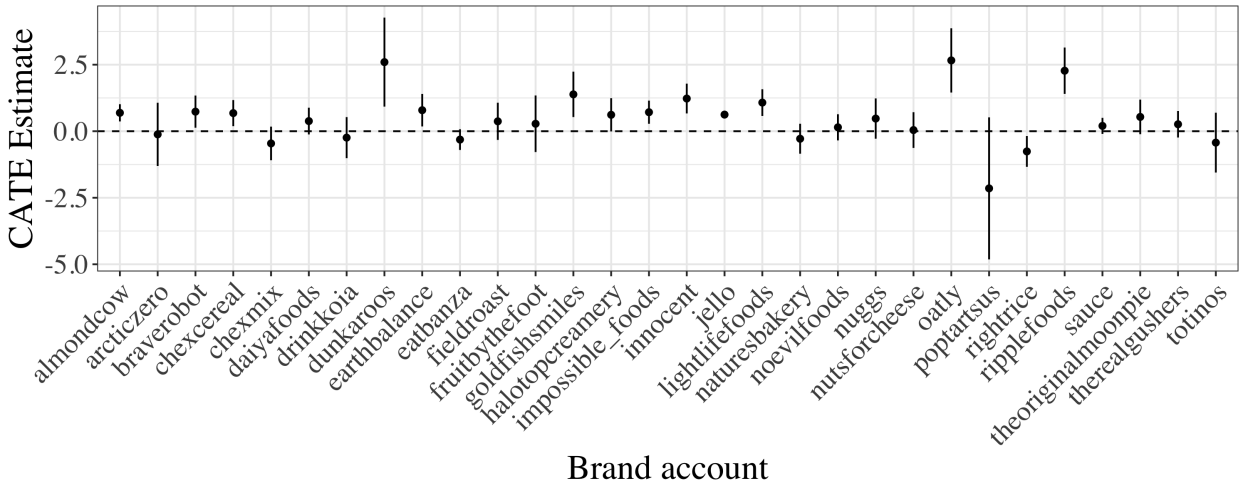
我们使用了一个数据集，其中包含31个食品品牌在2020年5月15日至2021年5月15日期间发布的 Instagram 图片帖子。如果帖子发布当天，品牌本身曾对观看者留下的评论作出任何回应，则处理变量取值为1；结果变量是帖子发布一天或更长时间后收到的评论数量。控制协变量包括原始帖子发布的小时、月份和星期，以及用户在帖子发布当天留下的评论总数，还包括帖子本身的图像。后者使我们能够直接控制帖子内容。对于该应用，我们首先在训练集上训练变分自编码器（Kingma 和 Welling，2013年），从而构建机器学习表示函数。我们使用表示函数，为匹配集中的每幅图像构建100维表示。我们的编码器-解码器架构由两个完全相同的、包含4层的卷积神经网络组成。我们使用70%的帖子数据训练模型，并使用剩余30%的数据进行推断。最终，我们得到 $n = 1677$ 个帖子，并对其进行匹配和条件平均处理效应估计。我们按如下方式选择匹配组的大小：对于 $t = 0$，1，我们将 k_{nt} set设为最接近 $n_t^{1/2}$ 的整数，其中 n_t 是匹配集中处理为 t 的组的大小；对处理组和对照组均采用这一过程后，我们获得 $k_1 = 7$ 来估计 $\mu(\mathbf{x}, 1)$，以及 $k_0 = 6$个体来估计 $\mu(\mathbf{x}, 0)$。

8.2 结果

在帖子发布当天至少进行过一次品牌互动，对该帖子随后几天收到的评论数量的平均处理效应为 0.36，95% 渐近置信区间为 0.27 至 0.63。由于结果变量是计数的自然对数，因此可以将该结果解释为：在发帖当天增加一次品牌互动，会使该帖子随后几天收到的评论数量增加约 *44%*。图 3 展示了更多结果，其中显示了每个品牌的单独平均处理效应估计值。这些

估计值是使用 M-DML 估计器，通过汇总每个品牌各个帖子的条件平均处理效应得到的。值得注意的是，大多数品牌似乎都表现出相似的正向处理效应，而只有 RightRice 的处理效应为负且具有统计显著性。这表明处理效应存在异质性，值得进一步研究。最后，我们展示了一些由所创建的匹配组中的示例图像

图 3：按品牌划分的条件平均处理效应



通过采用变分自编码器后端的 M-ML 实现。这一点很重要，因为我们可以检查用于生成估计量的案例（即匹配组），从而更好地审计这些估计量：如果匹配组不符合直觉，那么就有理由怀疑它们在处理效应估计中的实用性和整体可信度。图 4 所示的样本组突显了具有相似元素的图像如何被匹配在一起：大多数组别都包含来自同一品牌的帖子，且展示了相似食物的视觉表示。补充材料中还提供了展示类似模式的其他样本匹配组。这表明，我们的结果在人类层面上具有可解释性，并且依据的是匹配协变量中清晰可见的视觉线索。

图 4：样本匹配组



9 结论

在使用因果推断的高风险决策场景中，可解释性至关重要，因为它使分析人员能够依据自身的背景知识审计估计结果。本文介绍了匹配机器学习，这是一种旨在结合黑箱机器学习方法的预测能力与匹配的可审计性和易用性的方法。我们在一个通用框架下提出了用于条件平均处理效应和平均处理效应估计的 M-ML算法。匹配度量、表示函数和附加约束的许多不同选择，都可以纳入 M-ML 框架。我们从理论上证明，在合理条件下，用于条件平均处理效应估计的 M-ML 能以接近非参数最优的速率实现渐近正态性和一致性。我们还证明，将 M-ML 估计量作为增广逆概率加权估计量的输入，可以以 $\sqrt{n}$ 速率一致地估计平均处理效应和对处理组的平均处理效应。在实证方面，我们表明 M-ML 并未以牺牲准确性为代价来换取可审计性。在我们的应用中，我们展示了如何使用 M-ML 分析图像等非标准因果数据。总体而言，M-ML 拓展了可解释且准确的因果推断研究边界。M-ML 的潜在扩展包括：通过机器学习方法转换连续处理后进行匹配，以及在本文提出的条件基础上，进一步发展对第一阶段机器学习估计要求更宽松的理论条件。

–

参考文献

Alberto Abadie 和 Guido W Imbens。平均处理效应匹配估计量的大样本性质。《计量经济学》，74(1)：235–267，2006年。

Alberto Abadie 和 Guido W Imbens。平均处理效应的偏差校正匹配估计量。《商业与经济统计学杂志》，29(1)：1–11，2011年。

Alberto Abadie 和 Guido W Imbens。匹配估计量的鞅表示形式。Journal of the American Statistical Association，107(498)：833–843，2012年。

Alberto Abadie 和 Guido W Imbens。基于估计倾向得分的匹配。《计量经济学》，84(2)：781–807，2016年。

Alexandre Belloni、Victor Chernozhukov、Lie Wang 等。非参数回归中的平方根 LASSO 枢轴估计。《统计年鉴》，42(2)：757–788，2014年。

Victor Chernozhukov、Denis Chetverikov、Mert Demirer、Esther Duflo、Christian Hansen、Whitney Newey 和 James Robins。用于估计处理效应和结构参数的双重/去偏机器学习，2018年。

Hugh A Chipman、Edward I George、Robert E McCulloch 等。贝叶斯加性回归树。《应用统计年鉴》，4(1):266–298，2010年。

Luc Devroye、L´aszl´o Gy¨orfi 和 G´abor Lugosi。《模式识别概率理论》，第31卷。施普林格科学与商业媒体，2013年。

Alexis Diamond 和 Jasjeet S Sekhon。用于估计因果效应的遗传匹配：一种实现观察性研究平衡的通用多变量匹配方法。《经济学与统计学评论》，95(3):932–945，2013年。

Awa Dieng、Yameng Liu、Sudeepa Roy、Cynthia Rudin 和 Alexander Volfovsky。《用于因果推断的可解释近乎精确匹配》。载于第22届国际人工智能与统计学会议论文集，第2445–2453页。PMLR，2019年。

Harris Drucker、Chris JC Burges、Linda Kaufman、Alex Smola、Vladimir Vapnik 等。《支持向量回归机》。神经信息处理系统进展，9：155–161，1997年。

Max H Farrell、Tengyuan Liang 和 Sanjog Misra。用于估计和推断的深度神经网络。《计量经济学》，89(1)：181–213，2021年。

L Gy¨orfi。k n-近邻回归估计量和分类规则的收敛速率（通信）。IEEE信息论汇刊，27(3)：362–364，1981。

L´aszl´o Gy¨orfi、Michael Kohler、Adam Krzy˙zak 和 Harro Walk。非参数回归的分布无关理论，第1卷。Springer，2002。

P Richard Hahn、Jared S Murray 和 Carlos M Carvalho。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（含讨论）。《贝叶斯分析》，15(3)：965–1056，2020。

Ben B Hansen。倾向得分的预后类比物。Biometrika，95(2)：481–488，2008。

Jennifer L Hill。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年。

Stefano M Iacus、Gary King 和 Giuseppe Porro。无需平衡性检验的因果推断：粗化精确匹配。《政治分析》，第 1–24 页，2012年。

Leslie K John、Oliver Emrich、Sunil Gupta 和 Michael I Norton。“点赞”会带来喜爱吗？加入品牌社交网络对营销结果的影响。《营销研究杂志》，54(1)：144–155，2017年。

Nathan Kallus。《因果推断的广义最优匹配方法》。《机器学习研究杂志》，21(62):1–54，2020。

Diederik P Kingma 和 Max Welling。《自编码变分贝叶斯》。arXiv 预印本 arXiv:1312.6114，2013年。

S¨oren R K¨unzel、Jasjeet S Sekhon、Peter J Bickel 和 Bin Yu。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。

Lauren I Labrecque。在社交媒体环境中培育消费者-品牌关系：拟社会互动的作用。《互动营销杂志》，28(2)：134–148，2014年。

Michel Laroche、Mohammad Reza Habibi 和 Marie-Odile Richard。在社交媒体中存在还是缺席：社交媒体如何影响品牌忠诚度？《国际信息管理杂志》，33(1)：76–82，2013年。

Qi Li 和 Jeffrey Scott Racine。非参数计量经济学：理论与实践。普林斯顿大学出版社，2007。

Wei Luo 和YeyingZhu。使用充分维度约简进行因果推断的匹配。《商业与经济统计学杂志》，38(4)：888–900，2020。

Marco Morucci、Vittorio Orlandi、Sudeepa Roy、Cynthia Rudin 和 Alexander Volfovsky。用于可解释的个体化处理效应估计的自适应超盒匹配。载于《人工智能不确定性会议》，第 1089–1098 页。PMLR，2020。

Taisuke Otsu 和 Yoshiyasu Rai。平均处理效应的匹配估计量的自助法推断。Journal of the American Statistical Association，112(520):1720–1732，2017年。

Harsh Parikh、Cynthia Rudin 和 Alexander Volfovsky。MALTS：学习拉伸后的匹配。《机器学习研究杂志》，2022年。

Carl Edward Rasmussen。《机器学习中的高斯过程》。载于《机器学习暑期学校》，第 63–71 页。Springer，2003 年。

Luka Rimanic、Cedric Renggli、Bo Li 和 Ce Zhang。《特征变换下最近邻分类器的收敛性》。arXiv 预印本 arXiv:2010.07765，2020 年。

James M Robins、Andrea Rotnitzky 和 Lue Ping Zhao。某些回归变量并非始终可观测时回归系数的估计。Journal of the American Statistical Association，89(427):846–866，1994年。

Paul R Rosenbaum 和 Donald B Rubin。倾向得分在因果效应观察性研究中的核心作用。Biometrika，70(1):41–55，1983年。

Paul R Rosenbaum 和 Donald B Rubin。利用倾向得分亚分类减少观察性研究中的偏倚。Journal of the American Statistical Association，79(387):516–524，1984年。

Paul R Rosenbaum 和 Donald B Rubin。不完全匹配导致的偏倚。Biometrics，第103–116 页，1985a。

Paul R Rosenbaum 和 Donald B Rubin。使用纳入倾向得分的多变量匹配抽样方法构建对照组。美国统计学家，39(1):33–38，1985b。

Paul R Rosenbaum、Richard N Ross 和 Jeffrey H Silber。观察性研究中采用精细平衡的最小距离匹配抽样：卵巢癌处理研究。《Journal of the American Statistical Association》，102(477)：75–83，2007。

Donald B Rubin。具有相同偏倚降低比例的多变量匹配方法 I：若干示例。《Biometrics》，109–120页，1976。

Donald B Rubin。使用马氏距离匹配降低偏倚。《Biometrics》，293–298页，1980年。

Donald B Rubin 和 Neal Thomas。《椭球分布下的仿射不变匹配方法》。《统计年鉴》，第 1079–1093 页，1992年。

Fredrik S¨avje。《无放回匹配的不一致性》。Biometrika，109(2)：551–558，2022年。

Fran¸coise Simon 和 Vesselina Tossan。品牌与消费者的社交分享是否重要？品牌与消费者参与品牌运营的社交媒体的关系框架。《Journal of Business Research》，85：175–184，2018年。

Charles J Stone。非参数回归的最优全局收敛速率。《统计年鉴》，页码 1040–1053，1982年。

Winfried Stute 等。最近邻回归函数估计量的渐近正态性。《The Annals of Statistics》，12(3)：917–926，1984年。

Mark J Van Der Laan 和 Daniel Rubin。目标最大似然学习。《国际生物统计学杂志》，2(1)，2006年。

Stefan Wager 和 Susan Athey。使用随机森林估计和推断异质性处理效应。《Journal of the American Statistical Association》，113(523)：1228–1242，2018年。

Tianyu Wang、Marco Morucci、M Usaid Awan、Yameng Liu、Sudeepa Roy、Cynthia Rudin 和 Alexander Volfovsky。Flame：一种用于因果推断的快速、大规模、近似精确匹配方法。《机器学习研究杂志》，22(31)：1–41，2021年。

Xianfu Wang. 广义单位球的体积。Mathematics Magazine, 78(5):390–395, 2005年。

Yixin Wang 和 Jos´e R Zubizarreta. 平衡匹配的大样本性质。Statistica Sinica，即将出版。

Jos´e R Zubizarreta。《使用混合整数规划进行匹配：一项关于手术后肾衰竭的观察性研究》。Journal of the American Statistical Association，107（500）：1360–1371，2012年。

补充材料

A 预备知识

A.1 假设与主要符号

这里重述本文的主要符号和假设。设 $(\Omega\ \mathcal{F}, \mathcal{P})$ 为一个概率空间，其中 Ω $= \mathbb{R} \times \mathbb{X} \times \{1, \ldots, M\}$，并设 $\mathbf{O}\ = (Y, \mathbf{X}\ T)$ 为该空间上的一组随机变量，且 $Y = \sum_{j=1}^{M} Y(j)\ \mathbb{I}[T = j]$。注意，$Y(t)$ 的定义域属于 $\mathbb{R}$，$\mathbf{X}$ 属于 $\mathbb{X}$，而 T 属于 $\{1, \ldots, M\}$。用 $\mathcal{P}$ 表示 $\mathbf{O}$ 的联合分布。对于随机变量A，我们用 F_A 表示其累积分布函数，用 f_A 表示其概率密度函数，并用 $\mathbb{E}_A$ 和 $\mathbb{V}_A$ 分别表示关于 A 的期望和方差。当符号 $\mathbb{E}[\cdot]$ 或 $\mathbb{V}[\cdot]$ 未带任何下标使用时，均指相对于括号内的所有随机变量而言。对于函数g：$\mathbb{R}^p \mapsto \mathbb{R}^d$，定义距离函数：$D_g^q(\mathrm{u}, \mathrm{v}) := \|g(\mathrm{u}) - g(\mathrm{v})\|_q$，其中 $\|\cdot\|_q$ 为标准 q-范数。设 A 为 $\mathbb{R}^p$ 上的随机变量，并设 $f(A) : \mathbb{R}^p \mapsto \mathbb{R}^d$。我们将使用符号 $\|f(A)\|_{\mathcal{P},q} = \sqrt[q]{\int \max_{j=1,\ldots,d} |f(a)_j|^q d\mathcal{P}(a)}$ 表示关于测度$\mathcal{P}$的 L_q范数。我们使用以下符号：$\mu(\mathbf{x}\ t) := \mathbb{E}[Y|\mathbf{X}\ = \mathbf{x}\ T = t]$, $\sigma^2(\mathbf{x}\ t)\ = \mathbb{V}[Y|\mathbf{X}\ = \mathbf{x}\ T = t]$，以及$\tau(\mathbf{x}\ t, t')\ = \mu(\mathbf{x}\ t)\ - \mu(\mathbf{x}\ t')$，来表示感兴趣的量。

我们作如下假设：

A1（数据分布）：

(a) 数据 $\mathcal{O}_n = \{\mathbf{O}_i\}_{i=1}^n = \{Y_i, \mathbf{X}_i,\ T_i\}_{i=1}^n$ 是$\mathbf{O}$的 n 独立同分布副本组成的集合 .

(b) 协变量分布的定义域 $\mathbb{X}$ 是 $\mathbb{R}^p$的紧子集 .

(c) 协变量具有边缘分布，其累积分布函数相对于勒贝格测度可微 $F_{\mathbf{X}}$（$\mathbf{x}$），并且存在常数 $c_{f_{\mathbf{X}}}$，$C_{f_{\mathbf{X}}}$，使得 $0 < c_{f_{\mathbf{X}}} < f_{\mathbf{X}}$（$\mathbf{x}$）$< C_{f_{\mathbf{X}}} < \infty$ 在 $\mathbb{X}$ 上处处成立。

A2（重叠性）：对于所有 $\mathbf{x} \in \mathbb{X}$和 $t = 1$，.....，M都有 $0 < \Pr(T = t|\mathbf{X}\ = \mathbf{x}) < 1$。

A3（条件可忽略性）：$T \perp\!\!\!\perp$（Y（1）.....，Y（M））$|\mathbf{X}$。

A4（有界高阶矩）：对于所有 t，$t' \in \{1, \ldots, M\}$、所有 $\mathbf{x} \in \mathbb{X}$以及某个 $\delta > 0$和常数 C_δ，都有：$\mathbb{E}[|Y\ (t)\ |^{2+\delta}|\mathbf{X}\ = \mathbf{x}\ T = t'] \leq C_\delta$。

A5（Lipschitz条件）：对于所有$\mathbf{x}$、$\mathbf{z}\ \in \mathbb{X}$和 $t \in \{1, \ldots, M\}$，存在一个常数C_L，使得：(a) $|\mu(\mathbf{x}\ t)\ - \mu(\mathbf{z}\ t)| \leq C_L D_\phi^q(\mathbf{x},\mathbf{z})$ (b) $|\sigma^2(\mathbf{x}\ t)\ - \sigma^2(\mathbf{z}\ t)| \leq C_L D_\phi^q(\mathbf{x},\mathbf{z})$

A6（表示函数）：存在一个函数 $\widehat{\phi}(\mathbf{x}\ \mathcal{O}_n) : \mathbb{X} \times \Omega \mapsto \mathbb{R}^d$，使得对于实值 $r_{ML} > 0$ 和 $q > 0$:

(a) 函数 $\widehat{\phi}(\mathbf{x}\ \mathcal{O}_n)$ 和 $\widehat{\phi}(\mathbf{x})$ 关于$\mathbf{x}$ 在所有$\mathbf{x}\ \in \mathbb{X}$处均几乎必然连续。

(b) $\|\widehat{\phi}(\mathbf{x}\ \mathcal{O}_n)\ - \phi(\mathbf{x})\|_{\mathcal{P}\,q} = o(n^{-r_{ML}})$ 几乎必然地关于 $f_{\mathbf{X}}$。

(c) $\|\widehat{\phi}(\mathbf{X}\ \mathcal{O}_n)\ - \phi(\mathbf{X})\|_{\mathcal{P}\,q} = o(n^{-r_{ML}})$。

在本附录中，我们还将使用一些专用符号来表示$\widehat{\ }$匹配操作。我们用 $\mathtt{MG}(\phi,\ \mathbf{x}\ t) \subset \{1, \ldots, n\}$ 表示所形成的匹配组

围绕协变量值 $\mathbf{x}$、处理值 t以及表示函数 ϕ。请注意，$\widehat{\texttt{MG}}(\phi, \mathbf{x}\ t)$ 的定义会因使用卡尺 M-ML 还是 KNN-M-ML 而异，具体采用哪种定义将在各个定理中说明。我们还将使用 $\widehat{W}_i(\phi, \mathbf{x}\ t) = \mathbb{I}[i \in \widehat{\texttt{MG}}(\phi, \mathbf{x}\ t)]$ 表示个体 i属于 $\widehat{\texttt{MG}}(\phi, \mathbf{x}\ t)$，并用 $\widehat{N}(\phi, \mathbf{x}\ t) = \sum_{i=1}^{n} W_i(\phi, \mathbf{x}\ t)$ 计算 $\widehat{\texttt{MG}}(\phi, \mathbf{x}\ t)$ 中个体的数量。

B 证明

B.1 引理 1 的证明

证明。 在证明该结果之前，我们先确立一些重要事实。考虑量 $\widehat{\phi}(\mathbf{x}\ \mathcal{O}_n) =: \phi(\mathbf{x})$，为简化记号，我们省略其对 $\mathcal{O}_n$ 的显式依赖 ϕ 。根据 A6 (b) 和 (c)，我们知道它必定是在某个子集 $A \subset \mathbb{R}^d$ 上的随机变量（即可测函数）。记该随机变量的累积分布函数为 $F_{\widehat{\phi}(\mathbf{x})}$。根据 A6 (b)，对于任意 $\mathbf{u} \in A$，有：

$$\lim_{n\to\infty} F_{\widehat{\phi}(\mathbf{x})}(\mathbf{u}) = F_{\phi(\mathbf{x})}(\mathbf{u}) = \begin{cases} 1 & \text{if } u_j \geq \phi(\mathbf{x})_j,\ j = 1, \ldots, d \\ 0 & \text{otherwise.} \end{cases},$$

这是因为 $\phi(\mathbf{x})$ 相对于 $\mathbf{u}$ 是常数。这也意味着 $dF_{\phi(\mathbf{x})}(\mathbf{u}) = \delta(\phi(\mathbf{x}) - \mathbf{u})d\mathbf{u}$，其中 $\delta(\mathbf{u})$ 是狄拉克 δ 函数，它在 0 处的密度为 1，在其他所有位置的密度均为 0。由此，我们还可以得出，变量对 $(\phi(\mathbf{x})\ \phi(\mathbf{z}))$ 的联合累积分布函数（记为 $F_{(\widehat{\phi}(\mathbf{x}),\widehat{\phi}(\mathbf{z}))}(\mathbf{u}, \mathbf{v})$）将收敛于 $F_{\phi(\mathbf{x})}(\mathbf{u})F_{\phi(\mathbf{z})}(\mathbf{v})$。

对于任意$\mathbf{x}$、$\mathbf{z} \in \mathbb{X}$和 $\Gamma_n \geq 0$，我们可以将以下量写成刚刚研究过的各个量的函数：$\Pr_{\mathcal{O}_n}(\|\phi(\mathbf{x}) - \phi(\mathbf{z})\|_q \leq \Gamma_n)$（其中随机性来自训练数据）：

$$\Pr_{\mathcal{O}_n}(\|\widehat{\phi}(\mathbf{x}) - \widehat{\phi}(\mathbf{z})\|_q \leq \Gamma_n) = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathbb{I}[\|\mathbf{u} - \mathbf{v}\|_q \leq \Gamma_n] dF_{(\widehat{\phi}(\mathbf{x}),\widehat{\phi}(\mathbf{z}))}(\mathbf{u}, \mathbf{v}).$$

现在将变量从 $\mathbf{u}$，$\mathbf{v}$ 变换为 $\mathbf{u}$，$\mathbf{r}$，其中 $\mathbf{r} = \frac{\mathbf{u}-\mathbf{v}}{\Gamma_n}$，雅可比行列式等于 Γ_n^d。我们有：

$$\Pr_{\mathcal{O}_n}(\|\widehat{\phi}(\mathbf{x}) - \widehat{\phi}(\mathbf{z})\|_q \leq \Gamma_n) = \Gamma_n^d \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathbb{I}[\|\mathbf{r}\|_q \leq 1] dF_{(\widehat{\phi}(\mathbf{x}),\widehat{\phi}(\mathbf{z}))}(\mathbf{u}, \mathbf{u} - \Gamma_n\mathbf{r}).$$

现在考虑 $\Gamma_n^{-d}\Pr_{\mathcal{O}_n}(\|\widehat{\phi}(\mathbf{x}) - \widehat{\phi}(\mathbf{z})\|_q \le \Gamma_n)$ 的极限行为，我们有：

$$\begin{aligned}
\lim_{n\to\infty} \Gamma_n^{-d} \Pr_{\mathcal{O}_n}(\|\widehat{\phi}(\mathbf{x}) - \widehat{\phi}(\mathbf{z})\|_q \le \Gamma_n) &= \lim_{n\to\infty} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathbb{I}[\|\mathbf{r}\|_q \le 1] dF_{(\widehat{\phi}(\mathbf{x}),\widehat{\phi}(\mathbf{z}))}(\mathbf{u}, \mathbf{u} - \Gamma_n \mathbf{r}) \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathbb{I}[\|\mathbf{r}\|_q \le 1] \lim_{n\to\infty} dF_{(\widehat{\phi}(\mathbf{x}),\widehat{\phi}(\mathbf{z}))}(\mathbf{u}, \mathbf{u} - \Gamma_n \mathbf{r}) \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathbb{I}[\|\mathbf{r}\|_q \le 1] dF_{\phi(\mathbf{x})}(\mathbf{u}) F_{\phi(\mathbf{z})}(\mathbf{u}) \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \mathbb{I}[\|\mathbf{r}\|_q \le 1] \delta(\phi(\mathbf{x}) - \mathbf{u}) \delta(\phi(\mathbf{z}) - \mathbf{u}) d\mathbf{u} d\mathbf{r} \\
&= \int_{\mathbb{R}^d} \mathbb{I}[\|\mathbf{r}\|_q \le 1] d\mathbf{r} \int_{\mathbb{R}^d} \delta(\phi(\mathbf{x}) - \mathbf{u}) \delta(\phi(\mathbf{z}) - \mathbf{u}) d\mathbf{u} \\
&= \frac{2Ga(\frac{2}{q}+1)^d}{Ga(\frac{d}{q}+1)} \delta(\phi(\mathbf{x}) - \phi(\mathbf{z})) = V_d \delta(\phi(\mathbf{x}) - \phi(\mathbf{z})).
\end{aligned}$$

其中，第二个等号由控制收敛定理（DCT）推出，因为指示函数的上界为 1。第四个等号源于 $F_{\phi(\mathbf{x})}$ 相对于勒贝格测度绝对连续，因此其密度等于 $\delta(\phi(\mathbf{x}))$。最后一个等号则由以下事实推出：根据 q-范数定义的、以 0 为中心的 d 维单位球的体积为 $V_d = \int_{\mathbb{R}^d} \mathbb{I}[\|\mathbf{r}\|_q \le 1] d\mathbf{r} = \frac{2Ga(\frac{2}{q}+1)^d}{Ga(\frac{d}{q}+1)}$，其中 Ga 是伽马函数（例如参见 Wang, 2005），并且狄拉克 δ 函数具有已知性质。

现在我们可以研究 $\Pr_{\mathcal{O}_n}(\mathbf{Z} \in \mathrm{MG}(\widehat{\phi}, \mathbf{x}, t))$，对于 $(\mathbf{Z}, T) \sim f_{\mathbf{X},T}$，使用刚刚得到的结果并再次应用支配收敛定理。首先定义密度函数 $f_{\phi(\mathbf{X})}(\mathbf{u}) = \int_{\mathbb{X}} f_{\mathbf{X}}(\mathbf{z}) \delta(\phi(\mathbf{z}) - \mathbf{u}) d\mathbf{z}$。现在固定匹配数据中的任意一个单位 i，其

协变量 $\mathbf{X}_i = \mathbf{Z}$ 以及处理指示变量 $T_i = T$。我们有：

$$\begin{aligned}
\lim_{n\to\infty} \Gamma_n^{-d} \Pr_{\mathcal{O}_n,\mathbf{Z},T}(\mathbf{Z} \in \mathtt{MG}(\widehat{\phi},\mathbf{x},t)) &= \lim_{n\to\infty} \Gamma_n^{-d} \Pr_{\mathcal{O}_n,\mathbf{Z},T}(D^q_{\widehat{\phi}}(\mathbf{Z},\mathbf{x}) \le \Gamma_n, T = t) \\
&= \lim_{n\to\infty} \Gamma_n^{-d} \Pr(T=t) \mathbb{E}_{\mathbf{Z}}[\Pr_{\mathcal{O}_n}(D^q_{\widehat{\phi}}(\mathbf{Z},\mathbf{x}) \le \Gamma_n | T = t) | \mathbf{Z}] \\
&= \lim_{n\to\infty} \Gamma_n^{-d} e(t) \int_{\mathbb{X}} \Pr_{\mathcal{O}_n}(D^q_{\widehat{\phi}}(\mathbf{z},\mathbf{x}) \le \Gamma_n | \mathbf{Z} = \mathbf{z}, T = t) dF_{\mathbf{X}|T=t}(\mathbf{z}) \\
&= \lim_{n\to\infty} \Gamma_n^{-d} e(t) \int_{\mathbb{X}} \Pr_{\mathcal{O}_n}(D^q_{\widehat{\phi}}(\mathbf{z},\mathbf{x}) \le \Gamma_n) dF_{\mathbf{X}|T=t}(\mathbf{z}) \\
&= e(t) \int_{\mathbb{X}} \lim_{n\to\infty} \Gamma_n^{-d} \Pr_{\mathcal{O}_n}(\|\widehat{\phi}(\mathbf{x}) - \widehat{\phi}(\mathbf{z})\|_q \le \Gamma_n) dF_{\mathbf{X}|T=t}(\mathbf{z}) \\
&= e(t) \int_{\mathbb{X}} V_d \delta(\phi(\mathbf{x}) - \phi(\mathbf{z})) dF_{\mathbf{X}|T=t}(\mathbf{z}) \\
&= V_d e(t) \int_{\mathbb{X}} f_{\mathbf{X}}(\mathbf{z}) \int_{\mathbb{R}^d} \delta(\phi(\mathbf{x}) - \mathbf{u}) \delta(\phi(\mathbf{z}) - \mathbf{u}) d\mathbf{u} d\mathbf{z} \\
&= V_d e(t) \int_{\mathbb{R}^d} \delta(\phi(\mathbf{x}) - \mathbf{u}) \int_{\mathbb{X}} f_{\mathbf{X}}(\mathbf{z}) \delta(\phi(\mathbf{z}) - \mathbf{u}) d\mathbf{z} d\mathbf{u} \\
&= V_d e(t) \int_{\mathbb{R}^d} \delta(\phi(\mathbf{x}) - \mathbf{u}) f_{\phi(\mathbf{X})|T=t}(\mathbf{u}) d\mathbf{u} \\
&= V_d e(t) f_{\phi(\mathbf{X})|T=t}(\phi(\mathbf{x})).
\end{aligned}$$

注意，第二个和第四个等式中概率的独立性源于$\mathbf{Z}$ 和 T 独立于 $\mathcal{O}_n$ 抽取这一事实；第五个等式则是应用支配收敛定理（DCT）所得，这得益于积分中的概率被 1 所界定。至此，点 (i) 得证，下面转向点 (ii)。

由于 $\Pr_{\mathcal{O}_n,\mathbf{Z},T}(\mathbf{Z} \in \mathtt{MG}(\widehat{\phi},\mathbf{x},t))$ 被 1 所界定，因此对积分 $\mathbb{E}_{\mathbf{X}}[\Pr_{\mathcal{O}_n,\mathbf{Z},T}(\mathbf{Z} \in \mathtt{MG}(\widehat{\phi},\mathbf{X},t))] = \int_{\mathbb{X}} f_{\mathbf{X}}(\mathbf{x}) \Gamma_n^d \Pr_{\mathcal{O}_n,\mathbf{Z},T}(\mathbf{Z} \in \mathtt{MG}(\widehat{\phi},\mathbf{x},t)) d\mathbf{x}$ 再应用一次支配收敛定理即可验证第二个陈述。 □

B.2 定理 1 的证明

通过 本节余下部分将使用校正后的估计量 r:

$$\tilde{\mu}(\mathbf{x},t) = \frac{1}{N(\widehat{\phi},\mathbf{x},t)+1} \sum_{i\in\mathtt{MG}(\widehat{\phi},\mathbf{x},t)} Y_i.$$

该估计量的渐近行为与 $\hat{\mu}(\mathbf{x},t)$ 相同，但更易于处理，因为即使$\mathbf{x}$的匹配组 t 为空，它仍有定义。我们将证明 $\tilde{\mu}(\mathbf{x},t)$ 的结果，但这些结果同样适用于 $\hat{\mu}(\mathbf{x},t)$。

证明。 我们将先证明 (i) 中的结果，再证明 (ii) 中的结果。

命题（i） 回忆一下 $n^r := \min(n^{\frac{1}{2+d}}, n^{r_{\mathcal{MLE}}})$。根据引理 3（证明见下文），我们知道，为了证明 $\tilde{\mu}(\mathbf{x},t)$ 的渐近正态性，需要验证 3 个条件，

即：1） $R_\phi(\widehat{\phi},\ \mathbf{x}\ t)\ =\ o_p(n^{-r})$，2）$n^{2r}N(\widehat{\phi},\ \mathbf{x}\ t)\,(N(\widehat{\phi},\ \mathbf{x}\ t)+1)_2 \xrightarrow{p} \frac{1}{K}$对某个常数 K 成立，以及 3）$\mathbb{E}\left[\left(N\ \widehat{\phi}\ t^{n^{2r}}(,\ \mathbf{x})+1\right)\right]=O(1)$。

从第一个条件出发，令 $\mathbf{X}_{max}=X_{i^*}$，使得：$i^*\in\arg\max_{i\in \mathtt{MG}(\widehat{\phi},\mathbf{x},t)} D_\phi^q(\mathbf{X}_i,\mathbf{x})$，并回顾定义：$R_\phi(\phi,\mathbf{x}\ t)\ =D_\phi^q(\mathbf{X}max,\mathbf{x})$。根据匹配结果是使用卡尺 M-ML 得出的这一假设，我们有 $D_{\widehat{\phi}}^q(\mathbf{X}max,\mathbf{x})\ \leq\Gamma_n$，并根据 $\Gamma_n\asymp n^{\frac{2r-1}{d}}$ 这一假设可得：

$$\begin{aligned} n^r D_{\widehat{\phi}}^q(\mathbf{X}_{max},\mathbf{x}) &\leq n^r\Gamma_n = O(n^r n^{\frac{2r-1}{d}}) \\ &= O(n^{r(1+\frac{2}{d})-\frac{1}{d}})\to 0, \end{aligned} \tag{3}$$

对于任意 $r<\frac{1}{2+d}$。利用这一事实以及 A6c，我们可以证明 $n^r R_\phi(\widehat{\phi},\mathbf{x}\ t)\xrightarrow{p}0$，并如下验证条件 1：

$$\begin{aligned} n^r R_\phi(\widehat{\phi},\mathbf{x},t) &= n^r D_\phi^q(\mathbf{X}_{max},\mathbf{x}) \\ &= n^r\|\phi(\mathbf{X}_{max})-\phi(\mathbf{x})+\widehat{\phi}(\mathbf{X}_{max})-\widehat{\phi}(\mathbf{X}_{max})+\widehat{\phi}(\mathbf{x})-\widehat{\phi}(\mathbf{x})\|_q \\ &\leq n^r\|\phi(\mathbf{X}_{max})-\widehat{\phi}(\mathbf{X}_{max})\|_q + n^r\|\widehat{\phi}(\mathbf{x})-\phi(\mathbf{x})\|_q \\ &+ n^r\|\widehat{\phi}(\mathbf{X}_{max})-\widehat{\phi}(\mathbf{x})\|_q \\ &= o_p(1)+o_p(1)+n^r D_{\widehat{\phi}}^q(\mathbf{X}_{max},\mathbf{x})=o_p(1). \end{aligned}$$

该不等式由三角不等式得出，最后一行则由假设 A6 和式 (3) 得出。

其次，我们将通过证明 $\frac{n^{2r}N(}{\phi,\ \mathbf{x}\ t)\ (N(\widehat{\phi},\ \mathbf{x}\ t)+1)^2}\rightsquigarrow_{\nu_\phi(\mathbf{x}\ t)}$，来说明条件 2 成立，其中：$\nu_\phi(\mathbf{x}\ t)=V_d e(t) f_{\phi(\mathbf{X})|T=t}(\phi(\mathbf{x}))$ 的定义见引理 1。注意，$N(\phi,\ \mathbf{x}\ t)$ 是一个大小为 n、概率为 $\Pr(i\in \mathtt{MG}(\phi,\ \mathbf{x}\ t))$ 的二项分布随机变量。根据引理 1，我们知道：

$$E_{N(\widehat{\phi},\mathbf{x},t)}\left[\frac{N(\widehat{\phi},\mathbf{x},t)}{n^{2r}}\right]=n^{1-2r}\Pr(i\in\mathtt{MG}(\widehat{\phi},\mathbf{x},t))\xrightarrow{p}\nu_\phi(\mathbf{x},t) \tag{4}$$

只要 $\Gamma_n\asymp n^{\frac{2r-1}{d}}$，正如我们在该定理中所假设的。根据马尔可夫不等式，这意味着 $^{N}(\widehat{\phi},\ \mathbf{x}\ t\ \xrightarrow{p}\ \nu_\phi)$，现在考虑这个量：$(N(\widehat{\phi},\ \mathbf{x}\ t)\ +1)^2$：根据公式（4），我们知道：$N(\widehat{\phi},\ \mathbf{x}\ t)\ =O_p(n^{2r})$，这意味着 $N(\widehat{\phi},\ \mathbf{x}\ t)^2=O_p(n^{4r})$，因此：

$$n^{-2r}(N(\widehat{\phi},\mathbf{x},t)+1)^2\geq n^{-2r}N(\widehat{\phi},\mathbf{x},t)^2=n^{-2r}O_p(n^{4r})\xrightarrow{p}\infty.$$

然后，我们可以对 $g(N(\phi,\ \mathbf{x}\ t)\ +1)$ 应用连续映射定理，其中：$g(a)\ =\ \frac{1}{n^{-2r}a}-\frac{1}{n^{-2r}a^2}$，因为在我们的情境下，$a=N(\widehat{\phi},\ \mathbf{x}\ t)+1>0$，且所定义的 $g(a)$ 是连续的

在该定义域上。然后，条件 2 得到如下验证：

$$\begin{aligned}\frac{n^{2r}N(\widehat{\phi},\mathbf{x},t)}{(N(\widehat{\phi},\mathbf{x},t)+1)^2} &= \frac{N(\widehat{\phi},\mathbf{x},t)+1-1}{n^{-2r}(N(\widehat{\phi},\mathbf{x},t)+1)^2} \\ &= \frac{1}{n^{-2r}(N(\widehat{\phi},\mathbf{x},t)+1)} - \frac{1}{n^{-2r}(N(\widehat{\phi},\mathbf{x},t)+1)^2} \xrightarrow{p} \nu_\phi(\mathbf{x},t)^{-1}.\end{aligned}$$

为验证引理 3 所需的最后一个条件，我们先考察两个相关量。

首先，我们知道 $N(\phi, \mathbf{x}\ t)$ 服从样本量为 n、概率为 $\Pr(i \in \mathtt{MG}(\phi, \mathbf{x}\ t))$ 的二项分布，因此其二阶矩为 $\mathbb{E}[N(\widehat{\phi}, \mathbf{x}\ t)^2] = n(n-1)\gamma^2 + n\gamma$，其中 $\gamma = \Pr(i \in \mathtt{MG}(\widehat{\phi}, \mathbf{x}\ t)) = O(\Gamma_n^d) = O(n^{2r-1})$，并且根据引理 1 和连续映射定理，$(n^{1-2r}\gamma)^2 \to \nu_\phi(\mathbf{x}\ t)^2$,。因此，我们有：

$$\begin{aligned}\mathbb{E}\left[\left(\frac{N(\widehat{\phi},\mathbf{x},t)}{n^{2r}}\right)^2\right] &= \frac{n(n-1)}{n^{4r}}\gamma^2 + \frac{n}{n^{4r}}\gamma \\ &= n^{2-4r}\gamma^2 - n^{1-4r}\gamma^2 + n^{1-4r}\gamma \\ &= (n^{1-2r}\gamma)^2 - n^{1-4r}O(n^{4r-2}) + n^{1-4r}O(n^{2r-1}) \\ &= (n^{1-2r}\gamma)^2 - O(n^{-1}) + O(n^{-2r}) \to \nu_\phi(\mathbf{x},t)^2.\end{aligned}$$

对刚才所述结果的另一种应用可得：

$$\begin{aligned}\mathbb{E}\left[\left(\frac{N(\widehat{\phi},\mathbf{x},t)+1}{n^{2r}}\right)^2\right] &= \mathbb{E}\left[\frac{N(\widehat{\phi},\mathbf{x},t)^2 + 2N(\widehat{\phi},\mathbf{x},t)+1}{n^{4r}}\right] \\ &= \mathbb{E}\left[\left(\frac{N(\widehat{\phi},\mathbf{x},t)}{n^{2r}}\right)^2\right] + \frac{2}{n^{4r}}\mathbb{E}[N(\widehat{\phi},\mathbf{x},t)] + n^{-4r} \\ &= \mathbb{E}\left[\left(\frac{N(\widehat{\phi},\mathbf{x},t)}{n^{2r}}\right)^2\right] + o(1) + o(1) = O(1) \to \nu_\phi(\mathbf{x},t)^2.\end{aligned}$$

由上面的推导可知，根据马尔可夫不等式，$\left(\frac{N(\widehat{\phi},\mathbf{x},t)+1}{n^{2r}}\right)^2 \xrightarrow{p} \nu_\phi(\mathbf{x}\ t)^2$。此外，由于函数 $g(a) = 1/a$ 在正实数范围内连续，且 $N(\phi, \mathbf{x}\ t) + 1$属于该范围，我们可以应用连续映射定理，得出：$\left(\frac{n^{2r}}{N\widehat{\phi}\ t(,\mathbf{x})+1}\right) \xrightarrow{p} \nu_\phi(\mathbf{x}\ t)^{-2}$。由于 $g(a)$ 在同一范围内也以上界 1 为限，因此可以应用控制收敛定理，得出 $\mathbb{E}\left[\left(\frac{n^{2r}}{N\widehat{\phi}\ t(,\mathbf{x})+1}\right)^2\right] \xrightarrow{p} \nu_\phi(\mathbf{x}\ t)^{-2}$ 同样成立，这直接意味着我们需要验证的条件成立。由于三个条件均已验证，定理中的结果 (i) 可由引理 3 得出。

断言 (ii) 设 $\bar{\mu}(\mathbf{x}\ t) = {}_{N}\widehat{\phi}^{\,1}{}_{\overline{(,\ \mathbf{x}\ t)\ +1}} \sum_{i=1}^{n} W_i(\widehat{\phi},\ \mathbf{x}\ t)\mu(\mathbf{X}_i,\ t)$：我们应用三角不等式，将误差分解为两个包含该项的部分：

$$\begin{aligned}\|\tilde{\mu}(\mathbf{x},t)-\mu(\mathbf{x},t)\|_{\mathcal{P},s} &= (\mathbb{E}[|\tilde{\mu}(\mathbf{x},t)-\bar{\mu}(\mathbf{x},t)+\bar{\mu}(\mathbf{x},t)-\mu(\mathbf{x},t)|^s])^{1/s} \\ &\le 2^{1-1/s}\left((\underbrace{\mathbb{E}[|\tilde{\mu}(\mathbf{x},t)-\bar{\mu}(\mathbf{x},t)|^s]}_{I_1})^{1/s}+(\underbrace{\mathbb{E}[|\bar{\mu}(\mathbf{x},t)-\mu(\mathbf{x},t)|^s]}_{I_2})^{1/s}\right).\end{aligned} \tag{5}$$

我们分别对 I_1 和 I_2 给出上界。先考察分量 I_2，回忆一下，$\widehat{\ }\qquad\widehat{\ }\qquad\widehat{\ } N(\phi,\mathbf{x}\ t) = \sum_{i=1}^{n} W_i(\phi,\mathbf{x}\ t)$，且 $N(\phi,\mathbf{x}\ t)$ 服从二项分布。根据引理 1 和连续映射，我们有 $(n\Gamma_n^d)^{-s}N(\widehat{\phi},\mathbf{x}\ t)^s \xrightarrow{p} \nu_\phi(\mathbf{x}\ t)^s$，因此：$(n\Gamma_n^d)^s\ (\overline{N(\widehat{\phi},\mathbf{x}\ t)+1)_s} \xrightarrow{p} \nu_\phi(\mathbf{x}\ t)^{-s}$，这同样由连续映射定理得到。最后，由控制收敛定理可得：

$$\mathbb{E}\left[\frac{1}{(N(\widehat{\phi},\mathbf{x},t)+1)^s}\right] = O((n\Gamma_n^d)^{-s}). \tag{6}$$

$$\begin{aligned}
I_2 &= \mathbb{E}[|\bar{\mu}(\mathbf{x},t)-\mu(\mathbf{x},t)|^s] = \mathbb{E}\left[\left|\frac{1}{N(\widehat{\phi},\mathbf{x},t)+1}\sum_{i=1}^{n} W_i(\widehat{\phi},\mathbf{x},t)\mu(\mathbf{X}_i,t)-\mu(\mathbf{x},t)\right|^s\right] \\
&= \mathbb{E}\left[\left|\frac{1}{N(\widehat{\phi},\mathbf{x},t)+1}\sum_{i=1}^{n}\Big(W_i(\widehat{\phi},\mathbf{x},t)\mu(\mathbf{X}_i,t)-\mu(\mathbf{x},t)\Big)+\frac{\mu(\mathbf{x},t)}{N(\widehat{\phi},\mathbf{x},t)+1}\right|^s\right] \\
&\le 2^{s-1}\mathbb{E}\left[\left|\frac{1}{N(\widehat{\phi},\mathbf{x},t)+1}\sum_{i=1}^{n}\Big(W_i(\widehat{\phi},\mathbf{x},t)\mu(\mathbf{X}_i,t)-\mu(\mathbf{x},t)\Big)\right|^s+\left|\frac{\mu(\mathbf{x},t)}{N(\widehat{\phi},\mathbf{x},t)+1}\right|^s\right] \\
&\le 2^{s-1}\mathbb{E}\left[\left|\frac{1}{N(\widehat{\phi},\mathbf{x},t)+1}\sum_{i=1}^{n} W_i(\widehat{\phi},\mathbf{x},t)C_L D_\phi^q(\mathbf{X}_i,\mathbf{x})\right|^s+\left|\frac{\mu(\mathbf{x},t)}{N(\widehat{\phi},\mathbf{x},t)+1}\right|^s\right] \\
&\le 2^{s-1}\mathbb{E}\left[\left|\frac{C_L}{N(\widehat{\phi},\mathbf{x},t)+1}\sum_{i=1}^{n} W_i(\widehat{\phi},\mathbf{x},t)R_\phi(\widehat{\phi},\mathbf{x},t)\right|^s\right]+2^{s-1}\mathbb{E}\left[\left|\frac{\mu(\mathbf{x},t)}{N(\widehat{\phi},\mathbf{x},t)+1}\right|^s\right] \\
&= 2^{s-1}C_L^s\mathbb{E}\left[\left(\frac{N(\widehat{\phi},\mathbf{x},t)}{N(\widehat{\phi},\mathbf{x},t)+1}\right)^s R_\phi(\widehat{\phi},\mathbf{x},t)^s\right]+2^{s-1}|\mu(\mathbf{x},t)|^s\mathbb{E}\left[\frac{1}{(N(\widehat{\phi},\mathbf{x},t)+1)^s}\right] \\
&\le 2^{s-1}C_L^s\mathbb{E}\left[\|\phi(\mathbf{X}_{max})-\widehat{\phi}(\mathbf{X}_{max})+\widehat{\phi}(\mathbf{X}_{max})-\widehat{\phi}(\mathbf{x})+\widehat{\phi}(\mathbf{x})-\phi(\mathbf{x})\|_q^s\right]+O((n\Gamma_n^d)^{-s}) \\
&\le 2^{s-1}C_L^s\mathbb{E}[\|\widehat{\phi}(\mathbf{X}_{max})-\phi(\mathbf{X}_{max})\|_q^s]+2^{s-1}C_L^s\mathbb{E}[\|\widehat{\phi}(\mathbf{x})-\phi(\mathbf{x})\|_q^s] \\
&\quad+2^{s-1}C_L^s\mathbb{E}[\|\widehat{\phi}(\mathbf{X}_{max})-\widehat{\phi}(\mathbf{x})\|_q^s]+O((n\Gamma_n^d)^{-s}) \\
&= o(n^{-sr_{ML}})+o(n^{-sr_{ML}})+O(\Gamma_n^s)+O((n\Gamma_n^d)^{-s}).
\end{aligned} \tag{7}$$

$\widehat{\ }$第一个不等式由 A5 得出，第二个根据 $R_\phi(\phi,\mathbf{x}\ t)$ 的定义得出，第四个则由三角不等式得出。另请注意，第四行中的分数始终小于 1。$\widehat{\ }\qquad\widehat{\ }$最后一行的陈述由 A6 以及以下事实得出：$\|\phi(\mathbf{X}_{max})-\phi(\mathbf{x})\|_q^s \le \Gamma_n$

在卡尺 M-ML 条件下。

~ ~我们使用引理 4 来对 I_1 给出上界。令 $Y = (Y_1, \ldots Y_n)$, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$, ~和 $T = (T_1, \ldots, T_n)$, 则有:

$$
\begin{aligned}
I_1 &= \mathbb{E}\left[\left|\frac{1}{N(\widehat{\phi}, \mathbf{X}, t)+1}\sum_{i=1}^{n}(Y_i - \mu(\mathbf{X}, t))W_i(\widehat{\phi}, \mathbf{X}, t)\right|^s\right] \\
&= \mathbb{E}\left[\left|\frac{1}{N(\widehat{\phi}, \mathbf{X}, t)+1}\right|^s \mathbb{E}_{\tilde{Y}|\tilde{\mathbf{X}},\tilde{T}}\left[\left|\sum_{i=1}^{n}(Y_i - \mu(\mathbf{X}, t))W_i(\widehat{\phi}, \mathbf{X}, t)\right|^s\right]\right] \\
\text{(By Lemma 4)} &\leq \mathbb{E}\left[\frac{B_s C_s N(\widehat{\phi}, \mathbf{x}, t)^{s/2}}{(N(\widehat{\phi}, \mathbf{X}, t)+1)^s}\right] \\
&\leq B_s C_s \mathbb{E}\left[\frac{N(\widehat{\phi}, \mathbf{x}, t)^{s/2}}{(N(\widehat{\phi}, \mathbf{X}, t))^s}\right] = B_s C_s \mathbb{E}\left[\frac{1}{(N(\widehat{\phi}, \mathbf{X}, t))^{s/2}}\right]. \qquad (8)
\end{aligned}
$$

因此，我们有:

$$
I_1 \leq B_s C_s \mathbb{E}\left[\frac{1}{(N(\widehat{\phi}, \mathbf{X}, t))^{s/2}}\right] = O((n\Gamma_n^d)^{-s/2}). \qquad (9)
$$

合并 （5）、（9）和（7），并应用詹森不等式，可得 :

$$
\|\tilde{\mu}(\mathbf{x}, t) - \mu(\mathbf{x}, t)\|_{\mathcal{P},s} \leq 2^{1-1/s}(I_1 + I_2)^{1/s} = O((n\Gamma_n^d)^{-1/2}) + O(\Gamma_n) + o(n^{-r_{ML}}). \qquad (10)
$$

通过令 $\Gamma_n = n^{\frac{2r-1}{d}}$ 和 $r = \frac{1}{2+d}$ 并代入（10），即可得到定理中的界: —

$$
\|\tilde{\mu}(\mathbf{x}, t) - \mu(\mathbf{x}, t)\|_{\mathcal{P},s} = O(n^{-\frac{1}{2+d}}) + o(n^{-r_{ML}}). \qquad (11)
$$

由于 $\mathbf{X}$ 的定义域有界，我们可以对 $\|\tilde{\mu}(\mathbf{X}\ t) - \mu(\mathbf{X}\ t)\|_{\mathcal{P},\ s} = \mathbb{E}_{\mathbf{X}}[\|\tilde{\mu}(\mathbf{X}\ t) - \mu(\mathbf{X}\ t)\|_{\mathcal{P}}^s]^{1/s}$ 应用支配收敛定理（DCT），从而可知该界对 $\mathbf{X}$ 的期望也成立。证毕。

□

B.3 定理 2 的证明

证明。 **断言 (ii)** 我们首先对误差进行上界估计: $\|\hat{}\mu(\mathbf{x}\ t) - \mu(\mathbf{x}\ t)\|_{\mathcal{P},\ s}$ 对任意$\mathbf{x} \in \mathbb{X}$均成立，因为引入这一结果后，命题（i）的证明中的许多步骤都会变得简单。对这一初步结果应用支配收敛定理，即可得到断言 (ii) 中的结果。回忆一下，我们已定义 $\bar{\mu}(\mathbf{x}\ t) = \frac{1}{k_n}\sum_{i=1}^{n} W_i(\widehat{\phi},\ \mathbf{x}\ t)\mu(\mathbf{X}_i,\ t)$，我们应用三角不等式 —

将误差拆分为两个包含该项的分量：

$$\|\hat{\mu}(\mathbf{x},t)-\mu(\mathbf{x},t)\|_{\mathcal{P},s}=\mathbb{E}[\|\hat{\mu}(\mathbf{x},t)-\bar{\mu}(\mathbf{x},t)+\bar{\mu}(\mathbf{x},t)-\mu(\mathbf{x},t)\|^s]^{1/s}$$
$$\leq 2^{1-1/s}\left[\underbrace{(\mathbb{E}[\|\hat{\mu}(\mathbf{x},t)-\bar{\mu}(\mathbf{x},t)\|^s])}_{I_1}^{1/s}+\underbrace{(\mathbb{E}[\|\bar{\mu}(\mathbf{x},t)-\mu(\mathbf{x},t)\|^s])}_{I_2}^{1/s}\right].$$

我们使用引理 4 来给 I_1 求上界。令 $Y=(Y_1,\ldots Y_n)$、$\mathbf{X}=(\mathbf{X}_1,\ldots,\mathbf{X}_n)$，以及 $T=(T_1,\ldots,T_n)$，则有：

$$I_1=\mathbb{E}\left[\left|\frac{1}{k_n}\sum_{i=1}^n(Y_i-\mu(\mathbf{X},t))W_i(\widehat{\phi},\mathbf{X},t)\right|^s\right]=\mathbb{E}\left[\left|\frac{1}{k_n}\right|^s\mathbb{E}_{\tilde{Y}|\tilde{\mathbf{X}},\tilde{T}}\left[\left|\sum_{i=1}^n(Y_i-\mu(\mathbf{X},t))W_i(\widehat{\phi},\mathbf{X},t)\right|^s\right]\right]$$
$$(\text{By Lemma 4})\leq\mathbb{E}\left[\frac{B_sC_sk_n^{s/2}}{k_n^s}\right]=B_sC_s\frac{1}{k_n^{s/2}}.$$

在直接对偏差项 I_2给出上界之前，我们先为以下量建立一个上界：$\mathbb{E}[\|\phi(\mathbf{X}_{(k_n)})-\phi(\mathbf{x})\|_q^s]=D_{\widehat{\phi}}^q(\mathbf{X}_{(k_n)},\mathbf{x})^s$。注意，变换后的协变量 $\phi(\mathbf{X})$具有连续且有界的密度函数，这是由于原始协变量的密度函数 $f_{\mathbf{X}}(\cdot)$有界且连续，并且 $\phi(\mathbf{x})$在所有$\mathbf{x}\in\mathbb{X}$处连续。基于这些事实，我们可以将 $\phi(\mathbf{X}_1)\ldots,\ \phi(\mathbf{X}_n)$作为输入，并使用 $\lambda=-s$ 和 $\gamma=0$（λ 和 γ 在引理 5 中定义），应用引理 5。由此得到：

$$\mathbb{E}[D_{\widehat{\phi}}^q(\mathbf{X}_{(k_n)},\mathbf{x})^s]=O\left(\left(\frac{k_n}{n}\right)^{s/d}\right).\tag{12}$$

利用这一结果，我们现在可以转而对偏差项 I_2给出上界。回忆一下，$R_\phi(\phi,\ \mathbf{x}\ t)=\max_{i=1,\ldots,n}W_i(\phi,\ \mathbf{x}\ t)D_\phi^q(\mathbf{x},\ \mathbf{X}_i)$，并注意在此情形下，$R_\phi(\phi,\ \mathbf{x}\ t)$是$\mathbf{x}$ 与其在匹配样本中的 ϕ-距离最近邻之间的距离。我们有：

$$I_2=\mathbb{E}[|\bar{\mu}(\mathbf{x},t)-\mu(\mathbf{x},t)|^s]=\mathbb{E}\left[\left|\frac{1}{k_n}\sum_{i=1}^nW_i(\widehat{\phi},\mathbf{x},t)\mu(\mathbf{X}_i,t)-\mu(\mathbf{x},t)\right|^s\right]$$
$$=\mathbb{E}\left[\left|\frac{1}{k_n}\sum_{i=1}^nW_i(\widehat{\phi},\mathbf{x},t)(\mu(\mathbf{X}_i,t)-\mu(\mathbf{x},t))\right|^s\right]\leq\mathbb{E}\left[\left|\frac{1}{k_n}\sum_{i=1}^nW_i(\widehat{\phi},\mathbf{x},t)C_LD_\phi^q(\mathbf{X}_i,\mathbf{x})\right|^s\right]$$
$$\leq\mathbb{E}\left[\left|\frac{C_L}{k_n}\sum_{i=1}^nW_i(\widehat{\phi},\mathbf{x},t)R_\phi(\widehat{\phi},\mathbf{x},t)\right|^s\right]=C_L\mathbb{E}[R_\phi(\widehat{\phi},\mathbf{x},t)^s]$$

其中，第一个不等式由假设 5 得出，第二个不等式则由 $R_\phi(\widehat{\phi},\ \mathbf{x}\ t)$ 的定义得出。令 $\mathbf{X}_{(k_n)}$ 表示就 $D_{\widehat{\phi}}^q$ 而言，$\mathbf{x}$ 的 k_n^{th}最近邻的协变量，并注意

因此，利用（12），我们有：

$$\begin{aligned}\mathbb{E}[R_\phi(\widehat{\phi},\mathbf{x},t)^s] &= \mathbb{E}[\|\phi(\mathbf{X}_{(k_n)})-\phi(\mathbf{x})\|_q^s] \\ &\le 2^{s-1}(\mathbb{E}[\|\phi(\mathbf{X}_{(k_n)})-\widehat{\phi}(\mathbf{X}_{(k_n)})\|_q^s]+\mathbb{E}\|\widehat{\phi}(\mathbf{x})-\phi(\mathbf{x})\|_q^s]+\mathbb{E}[\|\widehat{\phi}(\mathbf{X}_{(k_n)})-\widehat{\phi}(\mathbf{x})\|_q^s]) \\ &= o(n^{-r_{ML}s})+o(n^{-r_{ML}s})+2^{s-1}\mathbb{E}[D_{\widehat{\phi}}^q(\mathbf{X}_{(k_n)},\mathbf{x})^s] \\ &= o(n^{-r_{ML}s})+O\left(\left(\frac{k_n}{n}\right)^{s/d}\right) \end{aligned} \tag{13}$$

其中事实 th $\widehat{\ }$诸如 $\mathbb{E}[\|\phi(\mathbf{x})-\phi(\mathbf{x})\|_q]$ 的项为 $o(n^{-r_{ML}})$，这一点由假设6得出。

最后，我们可以将迄今得到的界合并起来，从而证明定理中的结论：

$$\begin{aligned}\|\hat{\mu}(\mathbf{x},t)-\mu(\mathbf{x},t)\|_{\mathcal{P},s} &\le 2^{1-1/s}(I_1+I_2)^{1/s} \\ &\le 2^{1-1/s}\left[(B_sC_sk_n^{-s/2})^{\frac{1}{s}}+\left(O\left(\left(\frac{k_n}{n}\right)^{s/d}\right)+(o(n^{-sr_{ML}}))\right)^{\frac{1}{s}}\right] \\ &\le 2^{1-1/s}\left[C_sB_s^{1/s}k_n^{-1/2}+O\left(\left(\frac{k_n}{n}\right)^{1/d}\right)+o(n^{-r_{ML}})\right] \\ &= O(k_n^{-1/2})+O\left(\left(\frac{k_n}{n}\right)^{1/d}\right)+o(n^{-r_{ML}}).\end{aligned}$$

通过对$\|\widehat{\ }\mu(\mathbf{X}\ t)-\mu(\mathbf{X}\ t)\|_{\mathcal{P},\ s}=(\mathbb{E}_{\mathbf{x}\sim f_\mathbf{X}}[\|\widehat{\ }\mu(\mathbf{x}\ t)-\mu(\mathbf{x}\ t)\|_{\mathcal{P},\ s}^s])^{1/s}$应用支配收敛定理，可以轻易地将该结果推广到 $\|\widehat{\ }\mu(\mathbf{X}\ t)-\mu(\mathbf{X}\ t)\|_{\mathcal{P},\ s}$。将 $k_n=n^{\frac{2}{2+d}}$ 代入上述界并化简，即可得到定理中的最终下界。

命题（i）我们最后证明（i）中给出的渐近正态性结果。我们将借助引理 3 验证这一命题，因此需要确认本例满足引理要求的三个条件。由（13）和马尔可夫不等式可知，条件 1 成立：$R_\phi(\widehat{\phi},\ \mathbf{x}\ t)=o_p(n^{-r})$ 在 $f_\mathbf{X}$ 上几乎处处成立，只要我们令 $k_n=Kn^{2r}$，其中 $r=\min(\frac{1}{2+d},\ r_{ML})$。其次，对于条件 2 和条件 3，我们知道 $N(\widehat{\phi},\ \mathbf{x}\ t)=k_n$，因此：$N\widehat{\phi}(,\ \mathbf{x}\ t)=\frac{n^{2r}}{k_n}\to\frac{1}{K}$由假设成立。这也意味着 $\mathbb{E}\left[\left(N\widehat{\phi}(,\ \mathbf{x}\ t)\right)^2\right]\to\frac{1}{K^2}=O(1)$。因此，引理 3的条件 2 和条件 3 均已验证，引理直接推出所需结论。

□

B.4 定理 3 的证明

证明。我们基本上将沿用此前用于证明这一事实的所有论证。令$\eta(\mathbf{x}\ t)=\mathbb{E}[Y(t)^2|\mathbf{X}=\mathbf{x}],$，并回顾一下，我们先前已将 $\mu(\mathbf{x}\ t)=\mathbb{E}[Y(t)|\mathbf{X}=\mathbf{x}],\widehat{\ }$以及 $W_i(\phi,\mathbf{x}\ t)$ 定义为一个二元变量，用于表示属于 $\mathtt{MG}(\phi,\mathbf{x}\ t)$ 的成员资格。我们首先着重证明：$N\widehat{\phi}^1(,\mathbf{x}\ t)$ $\sum_{i=1}^n W_i(\widehat{\phi},\mathbf{x}\ t)Y_i^2\xrightarrow{p}\eta(\mathbf{x}\ t)$。要说明这一点确实成立

在这种情况下，注意：

$$\frac{1}{N(\widehat{\phi},\mathbf{x},t)}\sum_{i=1}^{n}W_i(\widehat{\phi},\mathbf{x},t)Y_i^2-\eta(\mathbf{x},t)$$
$$=\frac{1}{N(\widehat{\phi},\mathbf{x},t)}\sum_{i=1}^{n}W_i(\widehat{\phi},\mathbf{x},t)(Y_i^2-\eta(\mathbf{X}_i,t))+\frac{1}{N(\widehat{\phi},\mathbf{x},t)}\sum_{i=1}^{n}W_i(\widehat{\phi},\mathbf{x},t)(\eta(\mathbf{X}_i,t)-\eta(\mathbf{x},t)).$$

从第一项开始，对于任意 i，有：$\mathbb{E}[Y_i^2-\eta(\mathbf{X}_i,t)|T_i,\ \mathbf{X}_i]=0$，根据$\eta(\mathbf{X}_i,t)$的定义，这意味着 $\mathbb{E}[W_i(\widehat{\phi},\mathbf{x}\ t)(Y_i^2-\eta(\mathbf{X}_i,t))]=\mathbb{E}_{T_i,\ \mathbf{X}_i}[W_i(\widehat{\phi},\mathbf{x}\ t)\mathbb{E}_{Y_i|\mathbf{X}_i,T_i}[(Y_i^2-\eta(\mathbf{X}_i,t))]]=0$。因此，根据大数弱律：$\frac{1}{N(\widehat{\phi},\mathbf{x}\ t)}\sum_{i=1}^{n}W_i(\widehat{\phi},\mathbf{x}\ t)(Y_i^2-\eta(\mathbf{X}_i,t))\xrightarrow{p}0$。接着看第二项，首先注意，假设 A5 表明，对于所有 $\mathbf{u}$、$\mathbf{v}\in\mathcal{X}$：

$$\begin{aligned}|\mu(\mathbf{u},t)^2-\mu(\mathbf{v},t)^2|&=|(\mu(\mathbf{u},t)-\mu(\mathbf{v},t))(\mu(\mathbf{u},t)+\mu(\mathbf{v},t))|\\&\le|(\mu(\mathbf{u},t)-\mu(\mathbf{v},t))||(\mu(\mathbf{u},t)+\mu(\mathbf{v},t))|\\&\le C_LD_\phi^q(\mathbf{u},\mathbf{v})|(\mu(\mathbf{u},t)+\mu(\mathbf{v},t))|\\&\le C_LD_\phi^q(\mathbf{u},\mathbf{v})2C_\delta,\end{aligned}$$

其中，第一个不等式由柯西-施瓦茨不等式得出，第二个不等式由假设 A5 的 Lipschitz 条件得出，第三个不等式则由假设 A4 得出；假设 A4 意味着 $|\mu(\mathbf{u}\ t)|$ 由某个常数 C_δ 所界。将上述结果应用于前一个表达式的第二项，可得：

$$\begin{aligned}&\frac{1}{N(\widehat{\phi},\mathbf{x},t)}\sum_{i=1}^{n}W_i(\widehat{\phi},\mathbf{x},t)(\eta(\mathbf{X}_i,t)-\eta(\mathbf{x},t))\\&=\frac{1}{N(\widehat{\phi},\mathbf{x},t)}\sum_{i\in\mathtt{MG}(\widehat{\phi},\mathbf{x},t)}(\sigma^2(\mathbf{X}_i,t)+\mu(\mathbf{X}_i,t)^2-\sigma^2(\mathbf{x},t)-\mu(\mathbf{x},t)^2)\\&=\frac{1}{N(\widehat{\phi},\mathbf{x},t)}\sum_{i\in\mathtt{MG}(\widehat{\phi},\mathbf{x},t)}(\sigma^2(\mathbf{X}_i,t)-\sigma^2(\mathbf{x},t))+(\mu(\mathbf{X}_i,t)^2-\mu(\mathbf{x},t)^2)\\&\le\frac{1}{N(\widehat{\phi},\mathbf{x},t)}\sum_{i\in\mathtt{MG}(\widehat{\phi},\mathbf{x},t)}C_LD_\phi^q(\mathbf{X}_i,\mathbf{x})+2C_\delta C_LD_\phi^q(\mathbf{X}_i,\mathbf{x})\\&\le(C_L+2C_\delta C_L)\max_{i\in\mathtt{MG}(\widehat{\phi},\mathbf{x},t)}D_\phi^q(\mathbf{X}_i,\mathbf{x})=O_p(R_\phi(\widehat{\phi},\mathbf{x},t)),\end{aligned}$$

其中，第一个等式由方差的定义得出，第一个不等式由假设 A5 得出，第二个不等式则由 max 的定义得出。通过与定理 1 和定理 2 的证明相同的论证，我们知道 $R_\phi(\widehat{\phi},\mathbf{x}\ t)\xrightarrow{p}0$；无论是使用卡尺进行匹配（根据公式 (7)），还是使用固定的 k_n（根据公式 (12)），结论均成立。因此，$\frac{1}{N(\widehat{\phi},\mathbf{x}\ t)}\sum_{i=1}^{n}W_i(\widehat{\phi},\mathbf{x}\ t)(\eta(\mathbf{X}_i,t)-\eta(\mathbf{x}\ t))\xrightarrow{p}0$。这证明了$\frac{1}{N(\widehat{\phi},\mathbf{x}\ t)}\sum_{i\in\mathtt{MG}(\widehat{\phi},\mathbf{x}\ t)}Y_i^2$ 收敛于 $\eta(\mathbf{x}\ t)$。利用这一结果，我们可以证明简单方差估计量确实是 CATE 的一致估计量

方差：

$$\frac{1}{N(\widehat{\phi},\mathbf{x},t)}\sum_{i\in \mathtt{MG}(\widehat{\phi},\mathbf{x},t)}(Y_i-\hat{\mu}(\mathbf{x},t))^2=\frac{1}{N(\widehat{\phi},\mathbf{x},t)}\sum_{i\in \mathtt{MG}(\widehat{\phi},\mathbf{x},t)}(Y_i^2-2Y_i\hat{\mu}(\mathbf{x},t)+\hat{\mu}(\mathbf{x},t)^2)$$

$$=\left(\frac{1}{N(\widehat{\phi},\mathbf{x},t)}\sum_{i\in \mathtt{MG}(\widehat{\phi},\mathbf{x},t)}Y_i^2\right)-\left(\hat{\mu}(\mathbf{x},t)\frac{2}{N(\widehat{\phi},\mathbf{x},t)}\sum_{i\in \mathtt{MG}(\widehat{\phi},\mathbf{x},t)}Y_i\right)+\hat{\mu}(\mathbf{x},t)^2$$

$$=\frac{1}{N(\widehat{\phi},\mathbf{x},t)}\sum_{i\in \mathtt{MG}(\widehat{\phi},\mathbf{x},t)}Y_i^2-\hat{\mu}(\mathbf{x},t)^2$$

$$\xrightarrow{p}\eta(\mathbf{x},t)-\mu(\mathbf{x},t)^2=\sigma^2(\mathbf{x},t),$$

其中，$\hat{}\mu(\mathbf{x}\ t)^2$ 到 $\mu(\mathbf{x}\ t)^2$ 的收敛性，可由将连续映射定理应用于 $\hat{}\mu(\mathbf{x}\ t)\xrightarrow{p}\mu(\mathbf{x}\ t)$ 得出。

□

B.5 定理 4 的证明

证明。 该结论的证明可通过将同一论文中的定理 5.1（Chernozhukov⁀等人，2018年）应用于估计量 $\psi(\mathbf{X}_i,t)$ 得出。由于 M-DML 是该论文定义 3.2 中 DML2 算法的一种情形，且该情形由定理 5.1 覆盖，因此可以应用该定理。该定理要求我们检验估计量是否满足同一论文中的假设 5.1，该假设由若干条件组成。我们首先考察 Chernozhukov 等人（2018）的假设 5.1 中的原始条件。这些条件均由我们的主要假设满足，具体而言：条件 (i) 由定理中的假设满足，条件 (ii) 由 A4 满足，条件 (iii) 由 A2 满足，条件 (iv) 由 A1 满足，条件 (v) 由 A4 满足。

现在需要验证关于 $\hat{}\mu$ 和 $\hat{}e$ 的估计量的三个条件。首先，令 $\eta(\mathbf{x})=(\mu(\mathbf{x},1)\ \mu(\mathbf{x},0)\ e(\mathbf{x},1)\ e(\mathbf{x},0))$，其中我们用 $\eta(\mathbf{x})_j$ 表示该向量的第 j^{th} 个分量，$j=1,\ldots,4$。此外，令该量对应的估计量为 $\hat{}\eta_{\mathcal{O}_{\setminus\ell}}(\mathbf{x})=(\hat{}\mu(\mathbf{x},1)\ \mu(\mathbf{x},0),\hat{}e(\mathbf{x},1),\hat{}e(\mathbf{x},0))$，其中我们使用记号 $\mathcal{O}_{\setminus\ell}$ 来强调估计量 $\hat{}\eta_{\mathcal{O}_{\setminus\ell}}(\mathbf{x})$ 依赖于不属于第 ℓ 折的数据：$\mathcal{O}_{\setminus\ell}=\{\mathbf{X}_i,T_i,Y_i\}_{i\in S_{\setminus\ell}}$。对于 $s>2$，定义：

$$\|\hat{\eta}_{\mathcal{O}_{\setminus\ell}}-\eta\|_{\mathcal{P},s}=\max_{j\in 1,\ldots,4}\int_{\mathcal{O}_{\setminus\ell},\mathbf{X}}\|\hat{\eta}_D(\mathbf{x})_j-\eta(\mathbf{x})_j\|_q f_{\mathcal{O}_{\setminus\ell}}(D)f_{\mathbf{X}}(\mathbf{x})dD\mathbf{x}.$$

为了证明 Chernozhukov 等人（2018）的假设 5.1 在我们的设定下成立，我们需要证明，对于某个严格为正的 C，$\|\hat{}\eta_{\mathcal{O}_{\setminus\ell}}-\eta\|_{\mathcal{P},\ s}\leq C$。首先注意到，$\|\hat{}e(\mathbf{x}\ t)-e(\mathbf{x}\ t)\|_{\mathcal{P},\ s}<1$，因为 $0<\hat{}e(\mathbf{x}\ t)<1$，并且对于所有数据集 D、$\mathbf{x}$和t，$0<e(\mathbf{x}\ t)<1$ 根据构造成立。剩下需要证明 $\|\hat{}\mu-\mu\|_{\mathcal{P},\ s}\leq C$。这由我们的 A4 得出，方法是

在其中设定 $\delta = s-2$。那么，对于任意单位， i、 t 和 $\mathbf{x}$：

$$\begin{aligned}\mathbb{E}_{Y_i}[\|Y_i-\mu(\mathbf{x},t)\|_s|\mathbf{X}_i=\mathbf{x}_i,T_i=t] &\leq \mathbb{E}_Y[\|Y_i\|_s|\mathbf{X}_i=\mathbf{x}_i,T_i=t]+\|\mu(\mathbf{x},t)\|_s\\ &=\mathbb{E}_{Y_i}[\sqrt[s]{|Y_i|^s}|\mathbf{X}_i=\mathbf{x}_i,T_i=t]+\mathbb{E}_Y[\sqrt[s]{|Y|^s}|\mathbf{X}=\mathbf{x},T=t]\\ &\leq \mathbb{E}_{Y_i}[\|Y_i\|^s|\mathbf{X}_i=\mathbf{x}_i,T_i=t]+\mathbb{E}_Y[\|Y\|^s|\mathbf{X}=\mathbf{x},T=t]\\ &\leq C_s+C_s,\end{aligned}$$

因此：

$$\begin{aligned}\|\hat{\mu}-\mu\|_{\mathcal{P},s} &= \mathbb{E}[\mathbb{E}_{Y_{\setminus\ell}}[\|\hat{\mu}(\mathbf{x},t)-\mu(\mathbf{x},t)\|_q|\mathbf{X}=\mathbf{x},\mathbf{X}_{\setminus\ell},T_{\setminus\ell}]]\\ &=\mathbb{E}\left[\mathbb{E}_{Y_{\setminus\ell}}\left[\left\|\frac{1}{N(\widehat{\phi},\mathbf{x},t)}\sum_{i\in\texttt{MG}(\widehat{\phi},\mathbf{x},t)}Y_i-\mu(\mathbf{x},t)\right\|_q\middle|\mathbf{X}=\mathbf{x},\mathbf{X}_{\setminus\ell},T_{\setminus\ell}\right]\right]\\ &\leq\mathbb{E}\left[\mathbb{E}_{Y_{\setminus\ell}}\left[\max_{i\in\texttt{MG}(\widehat{\phi},\mathbf{x},t)}\left\|Y_i-\mu(\mathbf{x},t)\right\|_q\middle|\mathbf{X}=\mathbf{x},\mathbf{X}_{\setminus\ell},T_{\setminus\ell}\right]\right]\\ &\leq 2C_s.\end{aligned}$$

其次，我们必须证明 $\|\hat{\eta}_{\mathcal{O}_{\setminus\ell}}-\eta\|_{\mathcal{P},2}=o(1)$。根据假设， $\|\hat{e}-e\|_{\mathcal{P},2}$ 成立；根据定理 1，对于卡尺 M-ML，该结论成立；对于 KNN-MML，根据定理 2，该结论也成立。最后，Chernozhukov 等人（2018）的假设 5.1 的最后一个要求是：

$$\|\hat{\mu}-\mu\|_{\mathcal{P},2}\times\|\hat{e}-e\|_{\mathcal{P},2}=O(n^{-\frac{1}{2}}),$$

根据定理 1 和定理 2，我们知道 $\|\hat{\mu}-\mu\|_{\mathcal{P},2}=O(n^{-r})$；根据假设， $\|\hat{e}-e\|_{\mathcal{P},2}=O(n^{-r_e})$。由于我们假设 $r+r_e=1/2$，因此有：

$$\|\hat{\mu}-\mu\|_{\mathcal{P},2}\times\|\hat{e}-e\|_{\mathcal{P},2}=O(n^{-r})\times O(n^{-r_e})\leq O(n^{-1/2}).$$

完全相同的论证也可以用来证明，对处理组的平均处理效应的情形同样成立。

□

B.6 引理 2 的证明

证明。令 $\texttt{MG}^*=\{i:\ D^q_{\widehat{\phi}}(\mathbf{X}_i,\ \mathbf{x})\leq\Gamma_n\}$，并定义相应的成员指示变量 ：

$$W_i^*=\begin{cases}1 & \text{if } D^q_{\widehat{\phi}}(\mathbf{X}_i,\mathbf{x})\leq\Gamma_n\\ 0 & \text{otherwise.}\end{cases}$$

任取 $W_i, \ldots, W_n \in \{0,1\}^n$，并令 MG 为相应的匹配组。我们有：

$$\begin{aligned}
&\sum_{i \in \texttt{MG}}^{n} D^q_{\widehat{\phi}}(\mathbf{x}, \mathbf{X}_i) - \Gamma_n \sum_{i=1}^{n} W_i = \sum_{i=1}^{n} W_i (D^q_{\widehat{\phi}}(\mathbf{x}, \mathbf{X}_i) - \Gamma_n) \\
&= \underbrace{\sum_{i:\, D^q_{\widehat{\phi}}(\mathbf{X}_i, \mathbf{x}) > \Gamma_n} W_i (D^q_{\widehat{\phi}}(\mathbf{x}, \mathbf{X}_i) - \Gamma_n)}_{\geq 0} + \underbrace{\sum_{i:\, D^q_{\widehat{\phi}}(\mathbf{X}_i, \mathbf{x}) \leq \Gamma_n} W_i (D^q_{\widehat{\phi}}(\mathbf{x}, \mathbf{X}_i) - \Gamma_n)}_{\leq 0} \\
&\geq \sum_{i:\, D^q_{\widehat{\phi}}(\mathbf{X}_i, \mathbf{x}) \leq \Gamma_n} W_i (D^q_{\widehat{\phi}}(\mathbf{x}, \mathbf{X}_i) - \Gamma_n) \geq \sum_{i:\, D^q_{\widehat{\phi}}(\mathbf{X}_i, \mathbf{x}) \leq \Gamma_n} (D^q_{\widehat{\phi}}(\mathbf{x}, \mathbf{X}_i) - \Gamma_n) \\
&= \sum_{i=1}^{n} W_i^* (D^q_{\widehat{\phi}}(\mathbf{x}, \mathbf{X}_i) - \Gamma_n).
\end{aligned}$$

因此，$W_1^*, \ldots, W_n^*$ 和 MG* 是使目标函数最小化的匹配指标和匹配组。

□

B.7 引理 3

引理 *3.* 设 *A1-A6* 成立。对于任意的 $x \in \mathbb{X}$ 和 $t \in \{1, \ldots, M\}$，令匹配组 $\widehat{\texttt{MG}}(\phi, x\ t) \subset \{1, \ldots, n : T_i = t\}$。定义变量 $\widehat{W}_i(\phi, x\ t) = \mathbb{I}[i \in \widehat{\texttt{MG}}(\phi, x\ t)]$，作为指示单位 i 是否属于 $\widehat{\texttt{MG}}(\phi, x\ t)$ 的指标。此外，将匹配组中的个体数量定义为：$N(\phi, x\ t) = |\widehat{\texttt{MG}}(\phi, x\ t)|$，并令 $R_\phi(\phi, x\ t) = max_{i \in \widehat{\texttt{MG}}(\widehat{\phi}, \mathbf{x}, t)} D^q_\phi(X_i, x)$ 为匹配组的半径。若 $\widehat{\texttt{MG}}(\phi, x\ t)$ 满足以下要求，其中常数为 $c \geq 0$ 和 $r > 0$：

1. $R_\phi(\widehat{\phi}, \mathbf{x}, t) = o_p(n^{-r})$,
2. $\frac{n^{2r} N(\widehat{\phi}, \mathbf{x}, t)}{(N(\widehat{\phi}, \mathbf{x}, t) + c)^2} \xrightarrow{p} \frac{1}{K}$ *for some constant,* K,
3. $\mathbb{E}\left[\left(\frac{n^{2r}}{N(\widehat{\phi}, \mathbf{x}, t) + c}\right)^2\right] = O(1)$

那么，估计量：$\hat{\mu}(x\ t) = \frac{1}{N(\widehat{\phi}, x, t) + c} \sum_{i=1}^{n} W_i(\widehat{\phi}, x\ t) Y_i$ 满足：$n^r(\hat{\mu}(x\ t) - \mu(x\ t)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(x\ t)\ K)$。

证明。首先，我们将 Abadie 和 Imbens（2012）为处理组平均处理效应（ATT）的匹配估计量构造的鞅表示，调整为适用于任一匹配估计量 $\hat{\mu}(\mathbf{x}\ t)$ 的鞅表示。我们先将匹配条件平均处理效应估计量的估计误差写成两项之和：$\hat{\mu}(\mathbf{x}\ t) - \mu(\mathbf{x}\ t) =$

$\widehat{\ }$ $\widehat{\ }\mathcal{D}(\phi,\ \mathbf{x}\ t)+B(\phi,\ \mathbf{x}\ t)$，如下所示：

$$
\begin{aligned}
\hat{\mu}(\mathbf{x},t)-\mu(\mathbf{x},t) =& \frac{1}{N(\widehat{\phi},\mathbf{x},t)+c}\sum_{i\in\mathtt{MG}(\widehat{\phi},\mathbf{x},t)} Y_i-\mu(\mathbf{x},t) \\
&+\frac{1}{N(\widehat{\phi},\mathbf{x},t)+c}\sum_{i\in\mathtt{MG}(\widehat{\phi},\mathbf{x},t)}\mu(\mathbf{X}_i,t)-\frac{1}{N(\widehat{\phi},\mathbf{x},t)+c}\sum_{i\in\mathtt{MG}(\widehat{\phi},\mathbf{x},t)}\mu(\mathbf{X}_i,t) \\
=& \underbrace{\frac{1}{N(\widehat{\phi},\mathbf{x},t)+c}\sum_{i\in\mathtt{MG}(\widehat{\phi},\mathbf{x},t)} Y_i-\mu(\mathbf{X}_i,t)}_{\mathcal{D}(\widehat{\phi},\mathbf{x},t)}+\underbrace{\frac{1}{N(\widehat{\phi},\mathbf{x},t)+c}\sum_{i\in\mathtt{MG}(\widehat{\phi},\mathbf{x},t)}\mu(\mathbf{X}_i,t)-\mu(\mathbf{x},t)}_{B(\widehat{\phi},\mathbf{x},t)}.
\end{aligned}
\tag{14}
$$

$\widehat{\ }$我们暂且忽略 $B(\phi,\mathbf{x}\ t)$ 项，稍后再回过头来讨论它，因为我们将看到，$\hat{\ }\mu(\mathbf{x}\ t)$ 的渐近正态性所需的条件意味着该项趋于零。$\widehat{\ }$基于这一直观认识，只需证明 $n^r\mathcal{D}(\phi,\mathbf{x}\ t)$ 的渐近正态性即可，因为根据 Slutzky 定理，该项的渐近分布与 $\widehat{\ }$ $n^r(\hat{\ }\mu(\mathbf{x}\ t)-\mu(\mathbf{x}\ t))$ 的渐近分布相同。为了研究 $\mathcal{D}(\phi,\mathbf{x}\ t)$ 的渐近分布，我们希望将其写成鞅的形式，从而能够利用$\widehat{\ }$鞅数组的中心极限定理来确立其渐近正态性。通过将 $\mathcal{D}(\phi,\mathbf{x}\ t)$ 改写为 $\widehat{\ }$ $\widehat{\ }$鞅差之和，我们可以证明，它相对于某个滤过确实构成一个鞅。回顾一下，二元变量 $W_i(\phi,\mathbf{x}\ t)=\mathbb{I}[i\in\mathtt{MG}(\phi,\mathbf{x}\ t)]$ 表示$\widehat{\ }$ $\widehat{\ }$匹配集单位 i 在 $\mathtt{MG}(\phi,\mathbf{x}\ t)$ 中的成员资格。注意，仅当 $T_i=t$ 时，$W_i(\phi,\mathbf{x}\ t)=1$，并且对于任意 $i\in 1,\ldots,n$，定义：

$$
\xi_{n,i}(\mathbf{x})=\frac{n^r}{N(\widehat{\phi},\mathbf{x},t)+c}W_i(\widehat{\phi},\mathbf{x},t)(Y_i-\mu(\mathbf{X}_i,t))
$$

这些量将成为我们在表示中构造的鞅差，$\widehat{\ }$并且根据定义：$n^r\mathcal{D}(\phi,\ \mathbf{x}\ t)=\sum_{i=1}^n\xi_{n,\ i}(\mathbf{x})$。最后，定义以下 σ-域：

$$
\mathcal{F}_{n,i}(\mathbf{x})=\sigma\{\mathbf{X}_1,\ldots,\mathbf{X}_n,T_1,\ldots,T_n,Y_1,\ldots,Y_i,\widehat{\phi}\}. \tag{15}
$$

$\widehat{\ }$注意，我们将表示 ϕ 直接纳入所定义的滤过中；但如果该表示是从单独的训练集学习得到的，则可以将该数据集而非表示纳入滤过，从而得到相同的结果。由于$\xi_{n,\ i}(\mathbf{x})$的均值为零，且适应于滤过$\mathcal{F}_{n,\ i}(\mathbf{x})$，因此该数组：

$$
\left\{\sum_{j=1}^{i}\xi_{n,j}(\mathbf{x}),\ \mathcal{F}_{n,i}(\mathbf{x}),\ 1\le i\le n\right\} \tag{16}
$$

按照 Abadie 和 Imbens（2012年）中的相同论证，对每个 $n>2$ 而言都是一个鞅。

现在，我们可以将适用于三角鞅数组的 Lindeberg 中心极限定理应用于式（16）中定义的鞅数组。这里所讨论的中心极限定理指出，如果满足以下条件：

1. (条件 1) $\mathbb{E}[\xi_{n,i}(\mathbf{x})] = 0$

2.（条件 2） $\sum_{i=1}^{n} \mathbb{E}[\xi_{n,i}(\mathbf{x})^2 | \mathcal{F}_{n,\ i-1}(\mathbf{x})] \xrightarrow{p} \sigma^2$，对某个常数 σ^2，当 $n \to \infty$

3.（条件 3） $\forall \epsilon > 0$： $\sum_{i=1}^{N} \mathbb{E}[\xi_{n,i}(\mathbf{x})^2 \mathbb{I}_{[|\xi_{n,i}(\mathbf{x})| > \epsilon]}] \to 0$，当 $n \to \infty$,

则 $n^r \mathcal{D}(\widehat{\phi},\ \mathbf{x}\ t) \xrightarrow{d} \mathcal{N}(0,\ \sigma^2)$。下面将证明这三个条件均成立。

从条件 1 开始，该条件很容易验证，因为 $\mathbb{E}[\xi_{n,i}(\mathbf{x}) | \mathcal{F}_{n,\ i-1}(\mathbf{x})] = 0$，这是因为 $\xi_{n,\ i}(\mathbf{x})$ 是一个鞅差项，因此 $\mathbb{E}[\xi_{n,i}(\mathbf{x})] = \mathbb{E}[\mathbb{E}[\xi_{n,i}(\mathbf{x}) | \mathcal{F}_{n,\ i-1}(\mathbf{x})]] = \mathbb{E}[0] = 0$。

其次，条件 3（通常称为林德伯格条件）可由更容易检验的李雅普诺夫条件推出。李雅普诺夫条件指出，对于某个 $\delta > 0$:

$$\lim_{n \to \infty} \sum_{i=1}^{N} \mathbb{E}[|\xi_{n,i}(\mathbf{x})|^{2+\delta}] = 0.$$

由以下论证可知，在我们的情形下李雅普诺夫条件成立：固定一个 $s > 0$，则,

$$\begin{aligned}
\mathbb{E}[|\xi_{n,i}(\mathbf{x})|^{2+\delta}] &= \mathbb{E}\left[\left|\underbrace{\frac{n^r}{N(\widehat{\phi}, \mathbf{x}, t) + c}}_{\geq 0} \underbrace{W_i(\widehat{\phi}, \mathbf{x}, t)}_{\in \{0,1\}} (Y_i - \mu(\mathbf{X}_i, t))\right|^{2+\delta}\right] \\
&= \mathbb{E}\left[\frac{n^{2r+r\delta}}{(N(\widehat{\phi}, \mathbf{x}, t) + c)^{2+\delta}} W_i(\widehat{\phi}, \mathbf{x}, t) |Y_i - \mu(\mathbf{X}_i, t)|^{2+\delta}\right],
\end{aligned}$$

现在回顾一下，根据假设 A5，$\mathbb{E}[|Y_i|^{2+\delta} | \mathbf{X}_i,\ T_i] \leq C_\delta$，其中 $C_\delta < \infty$ 是某个常数。此外，由于 $Y_i \perp\!\!\!\perp \widehat{\phi}$（根据 A6.b），可得：$\mathbb{E}[|Y_i|^{2+\delta} | \widehat{\phi}, \mathbf{X}_i,\ T_i] \leq C_\delta$。由于在给定 ϕ 的条件下，$\widehat{\phi}$ 和 $\mathbf{X}_i$ 和 T_i，所有 $\mu(\mathbf{X}_i,\ t)$、$N(\phi,\ \mathbf{x}\ t)$、$W_i(\phi,\ \mathbf{x}\ t)$ 都是常数，因此可得:

$$\begin{aligned}
&\mathbb{E}\left[W_i(\widehat{\phi}, \mathbf{x}, t) \frac{n^{2r+r\delta}}{(N(\widehat{\phi}, \mathbf{x}, t) + c)^{2+\delta}} |Y_i - \mu(\mathbf{X}_i, t)|^{2+\delta}\right] \\
&= \mathbb{E}\left[W_i(\widehat{\phi}, \mathbf{x}, t) \frac{n^{2r+r\delta}}{(N(\widehat{\phi}, \mathbf{x}, t) + c)^{2+\delta}} \mathbb{E}[|Y_i - \mu(\mathbf{X}_i, t)|^{2+\delta} | \widehat{\phi}, \mathbf{X}_i, T_i]\right] \\
&\leq \mathbb{E}\left[W_i(\widehat{\phi}, \mathbf{x}, t) \frac{n^{2r+r\delta}}{(N(\widehat{\phi}, \mathbf{x}, t) + c)^{2+\delta}} C_\delta\right].
\end{aligned}$$

综合这些界，可得（当 $n \to \infty$ 时）：

$$\sum_{i=1}^{N} \mathbb{E}[|\xi_{n,i}(\mathbf{x})|^{2+\delta}] \leq C_\delta \sum_{i=1}^{n} \mathbb{E}\left[W_i(\widehat{\phi}, \mathbf{x}, t) \frac{n^{2r+r\delta}}{(N(\widehat{\phi}, \mathbf{x}, t) + c)^{2+\delta}}\right] = C_\delta \mathbb{E}\left[\frac{n^{2r+r\delta}}{(N(\widehat{\phi}, \mathbf{x}, t) + c)^{1+\delta}}\right]$$
$$= C_\delta \mathbb{E}\left[\frac{n^{2r(1+\delta)}}{n^{2r(1+\delta)}} \frac{n^{2r+r\delta}}{(N(\widehat{\phi}, \mathbf{x}, t) + c)^{1+\delta}}\right] = C_\delta \frac{1}{n^{r\delta}} \mathbb{E}\left[\left(\frac{n^{2r}}{(N(\widehat{\phi}, \mathbf{x}, t) + c)}\right)^{1+\delta}\right].$$

最后，根据引理的条件（3），对于 $\delta = 1$，我们有：

$$C_\delta \frac{1}{n^{r\delta}} \mathbb{E}\left[\left(\frac{n^{2\delta}}{N(\widehat{\phi}, \mathbf{x}, t) + c}\right)^{1+\delta}\right] = C_\delta \frac{1}{n^r} O(1) \to 0.$$

因此，条件 3 成立。

接下来讨论条件 2。首先回顾：$\mathbb{E}[(Y_i - \mu(\mathbf{X}_i, t))^2 | \mathcal{F}_{n,i-1}(\mathbf{x})] = \sigma^2(\mathbf{X}_i, t)$，其中 $\sigma^2(\mathbf{X}_i, t) = \mathbb{V}[Y(t) | \mathbf{X} = \mathbf{X}_i]$。据此，我们可以写成：

$$\sum_{i=1}^{n} \mathbb{E}[\xi_{n,i}(\mathbf{x})^2 | \mathcal{F}_{n,i-1}(\mathbf{x})] = \sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{n^r}{N(\widehat{\phi}, \mathbf{x}, t) + c}(W_i(\widehat{\phi}, \mathbf{x}, t)(Y_i - \mu(\mathbf{X}_i, t)))\right)^2 \middle| \mathcal{F}_{n,i-1}(\mathbf{x})\right]$$
$$= \sum_{i=1}^{n} \frac{n^{2r}}{(N(\widehat{\phi}, \mathbf{x}, t) + c)^2} (W_i(\widehat{\phi}, \mathbf{x}, t) \mathbb{E}[(Y_i - \mu(\mathbf{X}_i, t))^2 | \mathcal{F}_{n,i-1}(\mathbf{x})]$$
$$= \sum_{i=1}^{n} \frac{n^{2r}}{(N(\widehat{\phi}, \mathbf{x}, t) + c)^2} W_i(\widehat{\phi}, \mathbf{x}, t) \sigma^2(\mathbf{X}_i, t)$$
$$= \sum_{i=1}^{n} \frac{n^{2r}}{(N(\widehat{\phi}, \mathbf{x}, t) + c)^2} W_i(\widehat{\phi}, \mathbf{x}, t)(\sigma^2(\mathbf{X}_i, t) - \sigma^2(\mathbf{x}, t))$$
$$+ \sum_{i=1}^{n} \frac{n^{2r}}{(N(\widehat{\phi}, \mathbf{x}, t) + c)^2} W_i(\widehat{\phi}, \mathbf{x}, t) \sigma^2(\mathbf{x}, t)$$
$$= \sum_{i=1}^{n} \frac{n^{2r}}{(N(\widehat{\phi}, \mathbf{x}, t) + c)^2} W_i(\widehat{\phi}, \mathbf{x}, t)(\sigma^2(\mathbf{X}_i, t) - \sigma^2(\mathbf{x}, t)) \tag{17}$$
$$+ \sum_{i=1}^{n} \frac{n^{2r}}{(N(\widehat{\phi}, \mathbf{x}, t) + c)^2} \sigma^2(\mathbf{x}, t). \tag{18}$$

回想一下，根据假设 A5：$\sigma^2(\mathbf{X}_i, t) - \sigma^2(\mathbf{x}, t) \leq C_L D_\phi^q(\mathbf{x}, \mathbf{X}_i) \leq C_L R_\phi(\widehat{\phi}, \mathbf{x}, t)$ 对于任意

i，因此：

$$\begin{aligned}
&\sum_{i=1}^{n} \frac{n^{2r}}{(N(\widehat{\phi}, \mathbf{x}, t) + c)^2} W_i(\widehat{\phi}, \mathbf{x}, t)(\sigma^2(\mathbf{X}_i, t) - \sigma^2(\mathbf{x}, t)) \\
&\leq C_L \sum_{i=1}^{n} \frac{n^{2r}}{(N(\widehat{\phi}, \mathbf{x}, t) + c)^2} W_i(\widehat{\phi}, \mathbf{x}, t) R_\phi(\widehat{\phi}, \mathbf{x}, t) \\
&= \frac{n^{2r} N(\widehat{\phi}, \mathbf{x}, t)}{(N(\widehat{\phi}, \mathbf{x}, t) + c)^2} C_L R_\phi(\widehat{\phi}, \mathbf{x}, t) \\
&= O_p(1) o_p(1) C_L = o_p(1).
\end{aligned}$$

其中，最后一个等式由本引理的条件（1）和（2）得出。对于另一项（公式（18）），我们有：

$$\sum_{i=1}^{n} \frac{n^{2r}}{(N(\widehat{\phi}, \mathbf{x}, t) + c)^2} \sigma^2(\mathbf{x}, t) = \frac{n^{2r} N(\widehat{\phi}, \mathbf{x}, t)}{(N(\widehat{\phi}, \mathbf{x}, t) + c)^2} \sigma^2(\mathbf{x}, t) \xrightarrow{p} \frac{1}{K} \sigma^2(\mathbf{x}, t),$$

由本引理的条件（2）和 Slutzky 定理可得。因此：

$$\sum_{i=1}^{n} \mathbb{E}[\xi_{n,i}(\mathbf{x})^2 | \mathcal{F}_{n,i-1}(\mathbf{x})] = o_p(1) + \frac{1}{K} \sigma^2(\mathbf{x}, t),$$

这证明了林德伯格中心极限定理的条件 2。由于所有条件均已满足，该中心极限定理意味着：$n^r \mathcal{D}(\widehat{\phi}, \mathbf{x}\ t) \xrightarrow{d} \mathcal{N}(0, \frac{1}{K}\sigma^2(\mathbf{x}\ t))$。将式(14)中的分解代入，立即可得 $\hat{\mu}(\mathbf{x}\ t) - \mu(\mathbf{x}\ t)$ 的结果：$n^r(\hat{\mu}(\mathbf{x}\ t) - \mu(\mathbf{x}\ t)) = n^r \mathcal{D}(\phi, \mathbf{x}\ t) + n^r B(\phi, \mathbf{x}\ t)$。如上所示，根据林伯格中心极限定理，第一个量收敛于 $\mathcal{N}(0, \frac{1}{K}\sigma^2(\mathbf{x}\ t))$；根据引理的条件（1）和条件（3），第二个量依概率收敛于 0：

$$\begin{aligned}
n^r B(\widehat{\phi}, \mathbf{x}, t) &= \frac{n^r}{N(\widehat{\phi}, \mathbf{x}, t) + c} \sum_{i=1}^{n} W_i(\widehat{\phi}, \mathbf{x}, t) \mu(\mathbf{X}_i, t) - \mu(\mathbf{x}, t) \\
&= \frac{n^r}{N(\widehat{\phi}, \mathbf{x}, t) + c} \sum_{i=1}^{n} (\mu(\mathbf{X}_i, t) - \mu(\mathbf{x}, t)) W_i(\widehat{\phi}, \mathbf{x}, t) + \frac{n^r c \mu(\mathbf{x}, t)}{N(\widehat{\phi}, \mathbf{x}, t) + c} \\
&\leq \frac{n^r}{N(\widehat{\phi}, \mathbf{x}, t) + c} \sum_{i=1}^{n} C_\mu D_\phi^q(\mathbf{x}, \mathbf{X}_i) W_i(\widehat{\phi}, \mathbf{x}, t) + \frac{n^r c \mu(\mathbf{x}, t)}{N(\widehat{\phi}, \mathbf{x}, t) + c} \\
&\leq C_\mu \frac{n^r}{N(\widehat{\phi}, \mathbf{x}, t) + c} \sum_{i=1}^{n} R_\phi(\widehat{\phi}, \mathbf{x}, t) W_i(\widehat{\phi}, \mathbf{x}, t) + \frac{n^r c \mu(\mathbf{x}, t)}{N(\widehat{\phi}, \mathbf{x}, t) + c} \\
&\leq \underbrace{C_\mu n^r R_\phi(\widehat{\phi}, \mathbf{x}, t)}_{o_p(1) \text{ by Req. (1)}} + \underbrace{\frac{n^r c \mu(\mathbf{x}, t)}{N(\widehat{\phi}, \mathbf{x}, t) + c}}_{o_p(1) \text{ by Req. (3)}}.
\end{aligned} \tag{19}$$

注意，根据马尔可夫不等式，引理的条件（3）意味着 $\frac{n^r c\mu_{(\mathbf{x}\,t)}}{N(\widehat{\phi},\,\mathbf{x}\,t)+c} = o_p(1)$。至此，证明完成。 □

引理 4.（匹配结果的矩界）设 MG（ϕ、$\mathbf{x}$ t）为一组单位，其中处理 t 与 $\mathbf{x}$ 匹配，并设 N（ϕ、$\mathbf{x}$ t）表示该组的大小。设假设 *A1–A4* 成立，并令其他所有量均按本文其余部分中的定义给出。对于 $s \geq 2$，我们有：

$$\mathbb{E}\left[\left|\sum_{i=1}^{n} W_i(\widehat{\phi},\mathbf{x},t)(Y_i-\mu(\mathbf{X}_i,t))\right|^s\right] \leq B_s C_s N(\widehat{\phi},\mathbf{x},t)^{s/2}, \tag{20}$$

对于一个仅取决于 s 的常数 B_s，以及假设 *4* 中定义的 C_s .

证明。这是一个对均值为零的随机变量普遍成立的著名结果。这里我们只需将其证明调整到当前设定。令 $Y = (Y_1, \ldots\ldots, Y_n)$、$\mathbf{X} = (\mathbf{X}_1, \ldots\ldots, \mathbf{X}_n)$，以及 $T = (T_1, \ldots\ldots, T_n)$。由于随机变量 $Y_i - \mu(\mathbf{X}_i,\ t)$ 在给定 $\mathbf{X}_i$ 的条件下均值为 0，根据马尔琴凯维奇–兹格蒙德不等式，存在一个常数 B_s 使得：

$$\mathbb{E}_{\tilde{Y}|\tilde{\mathbf{X}},\tilde{T}}\left[\left|\sum_{i=1}^{n} W_i(\widehat{\phi},\mathbf{x},t)(Y_i-\mu(\mathbf{X}_i,t))\right|^s\right] \leq B_s \mathbb{E}_{\tilde{Y}|\tilde{\mathbf{X}},\tilde{T}}\left[\left(\sum_{i=1}^{n}|W_i(\widehat{\phi},\mathbf{x},t)(Y_i-\mu(\mathbf{X}_i,t))|^2\right)^{s/2}\right]$$

$$= B_s N(\widehat{\phi},\mathbf{x},t)^{s/2}\mathbb{E}_{\tilde{Y}|\tilde{\mathbf{X}},\tilde{T}}\left[\left(\frac{1}{N(\widehat{\phi},\mathbf{x},t)}\sum_{i=1}^{n}|W_i(\widehat{\phi},\mathbf{x},t)(Y_i-\mu(\mathbf{X}_i,t))|^2\right)^{s/2}\right]$$

然后，根据詹森不等式，我们有：

$$\leq B_s N(\widehat{\phi},\mathbf{x},t)^{s/2}\mathbb{E}_{\tilde{Y}|\tilde{\mathbf{X}},\tilde{T}}\left[\left(\frac{1}{N(\widehat{\phi},\mathbf{x},t)}\sum_{i=1}^{n}|W_i(\widehat{\phi},\mathbf{x},t)(Y_i-\mu(\mathbf{X}_i,t))|^s\right)\right]$$

$$= B_s N(\widehat{\phi},\mathbf{x},t)^{s/2}\left(\frac{1}{N(\widehat{\phi},\mathbf{x},t)}\sum_{i=1}^{n} W_i(\widehat{\phi},\mathbf{x},t)\mathbb{E}_{Y_i|\mathbf{X}_i,T_i}[|(Y_i-\mu(\mathbf{X}_i,t))|^s]\right)$$

并且由于 $\mathbb{E}_{Y_i|\mathbf{X}_i,T_i}[|(Y_i-\mu(\mathbf{X}_i,t))|^s] \leq C_s$，根据假设 4，且 $\delta = s-2$ ：

$$\leq B_s N(\widehat{\phi},\mathbf{x},t)^{s/2}\left(\frac{1}{N(\widehat{\phi},\mathbf{x},t)}\sum_{i=1}^{n} W_i(\widehat{\phi},\mathbf{x},t)C_s\right)$$

$$= B_s C_s N(\widehat{\phi},\mathbf{x},t)^{s/2}.$$

□

引理 5.（KNN 距离的界，*Li and Racine 2007* 年的引理 *14.1*）设 $\mathbf{Z}_1$、……、$\mathbf{Z}_n$ 为独立同分布观测值，其连续密度有界，且 $f_{\mathbf{Z}}$ 支撑集包含于 $\mathbb{R}^d$ 的某个子集。对于支撑集 $f_{\mathbf{Z}}$ 中的点 $\mathbf{z}$，令 $B(\mathbf{z}\ r)$ 表示以 $\mathbf{z}$ 为中心、半径为 r 的球，并定义 $G(r) = \Pr_{\mathbf{Z}}(\mathbf{Z} \in B(\mathbf{z}\ r))$。此外，令 $R_k(\mathbf{z})$ 表示 $\mathbf{z}$ 与

其在$\mathbf{Z}_1$、……、$\mathbf{Z}_n$中的 k^{th} 最近邻之间的欧几里得距离。最后，令 λ 和 γ 为整数，使得函数 $\Phi(R_k(\mathbf{z})) := \frac{1}{R_k \lambda G(\mathbf{z})(R_k(\mathbf{z}))^\gamma}$ 存在。则：

$$\mathbb{E}_{\mathbf{Z}_1,\ldots,\mathbf{Z}_n}[\Phi(R_k(\mathbf{z}))] = O\left(\left(\frac{k}{n}\right)^{-\frac{\lambda}{d}}\right). \tag{21}$$

该引理是 Li and Racine（2007）引理 14.1 的重述，且已在其中得到证明 .

C 模拟中使用的方法

表2：模拟实验中使用的方法

首字母缩略词	方法	引用
线性回归	线性回归	
BART	贝叶斯加性回归树	Chipman 等人（2010）
CF	因果森林	Wager 和 Athey（2018）
GP	高斯过程回归	（例如，Rasmussen，2003）
SVM	支持向量机	Drucker 等人（1997）
随机森林 X-Learner	结合随机森林的 X-Learner	K¨unzel 等人（2019年）
偏差校正 GenMatch	带偏差校正的遗传匹配	Diamond 和 Sekhon（2013年）
偏差校正 l2Match	带偏差校正的L2距离匹配	Abadie 和 Imbens（2011年）
MML-BART-Y	基于BART估计的结果变量的M-ML	
MML-SVM-Y	基于SVM估计的结果变量的M-ML	
MML-GP-Y	基于GP估计的结果变量的M-ML	

D 额外模拟结果

E 品牌数据上的更多样本匹配组

图 5：CATE 的 95% 渐近置信区间大小

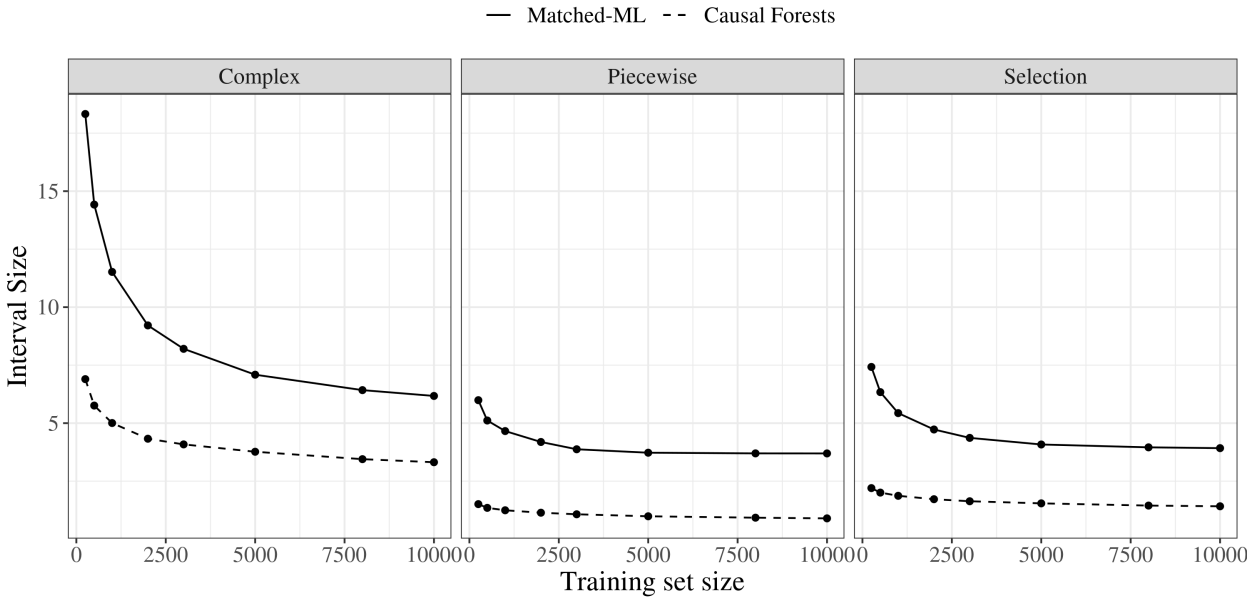


图 6：ATE 的 95% 渐近置信区间大小

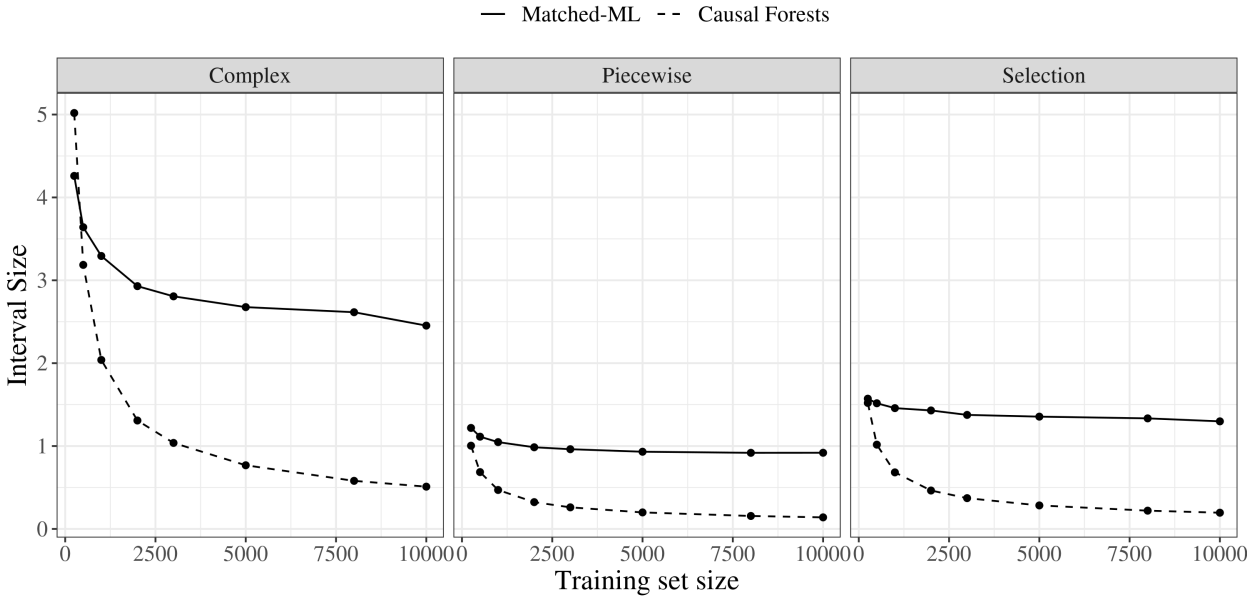


图 7：不包含零的置信区间比例，条件平均处理效应

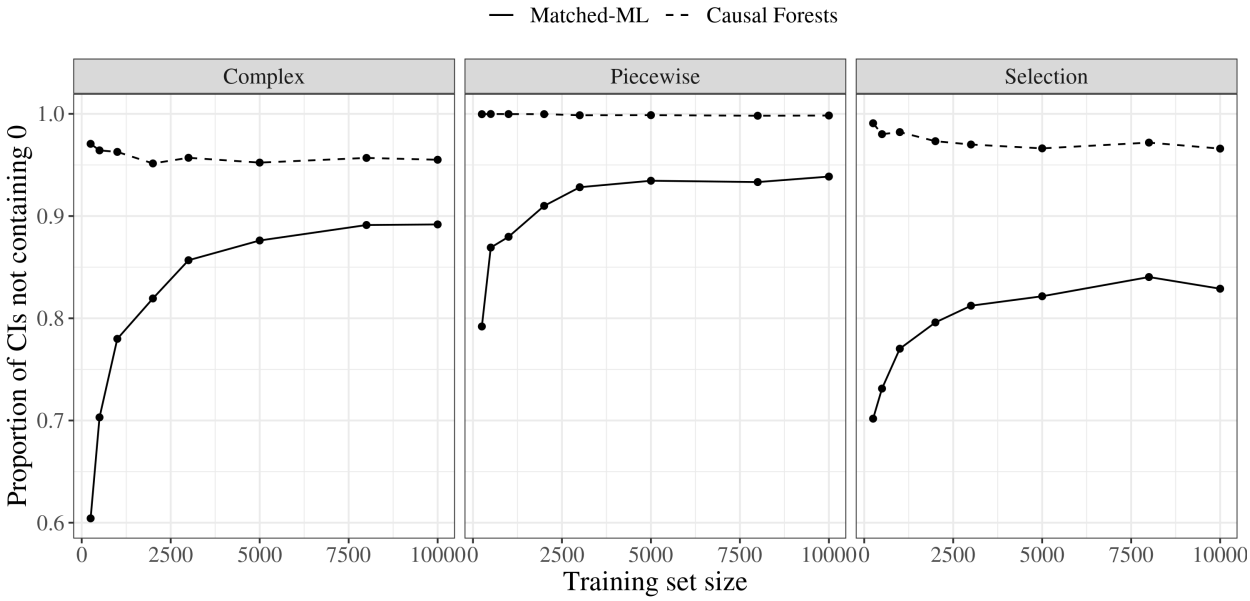
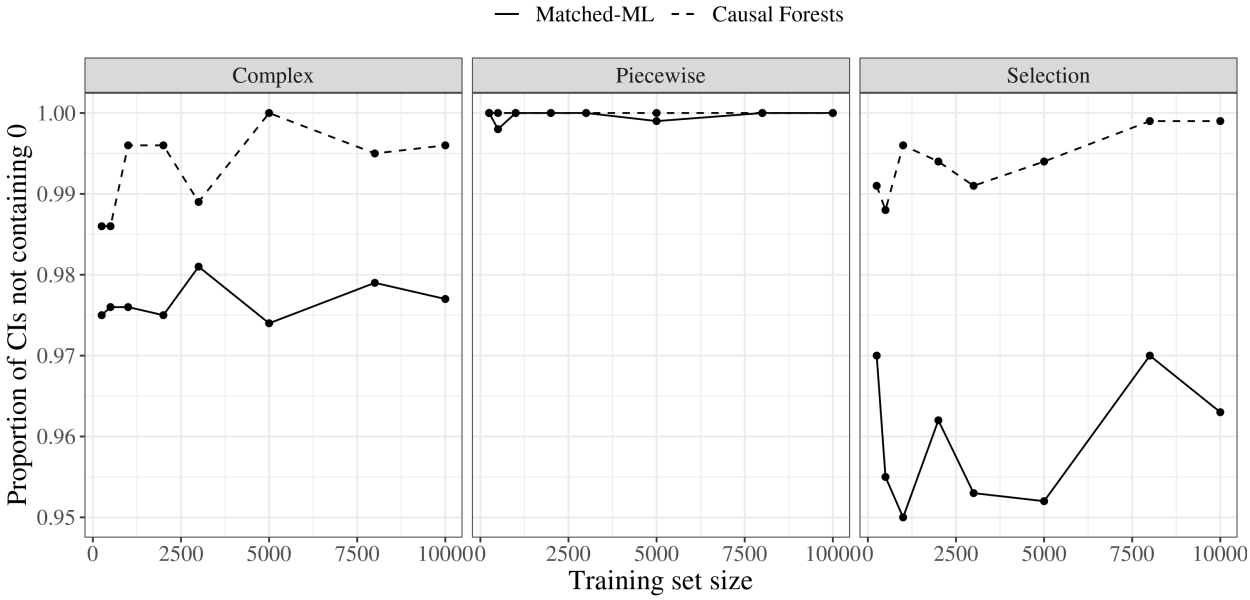
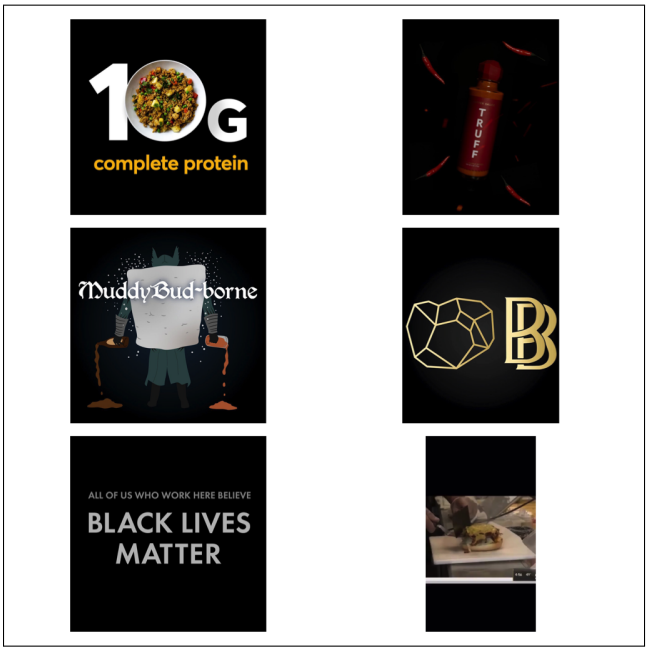
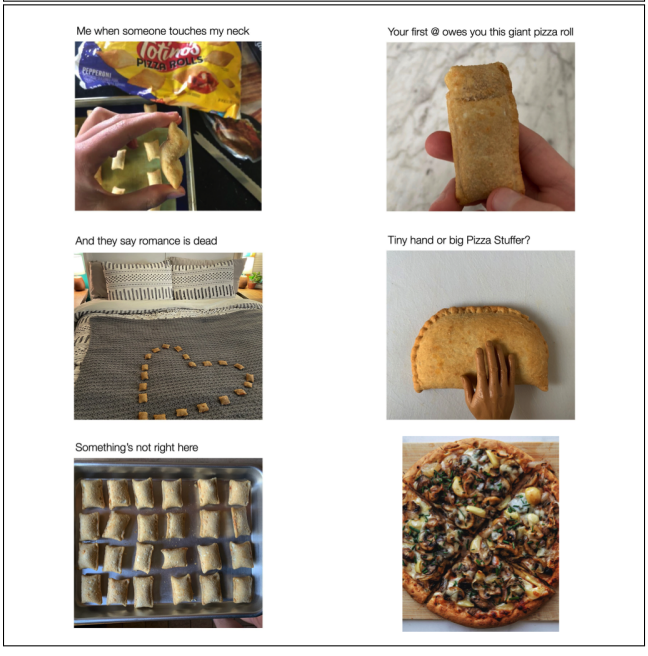


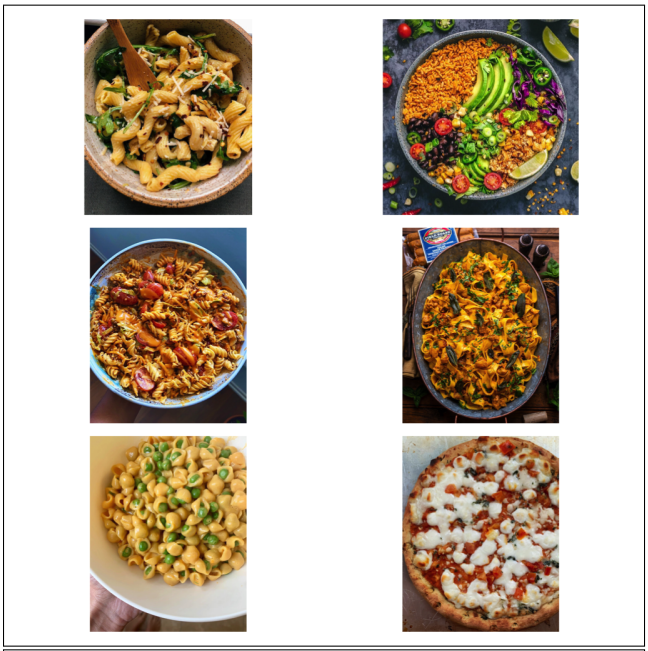
图 8：不包含零的置信区间比例，平均处理效应

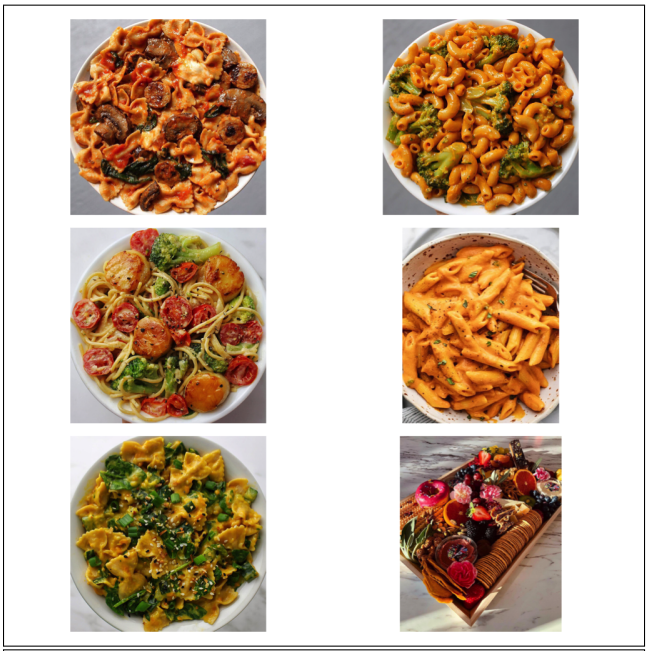



"SHARKS ARE THE TEACHERS OF THE OCEAN."
"SOME PEOPLE THINK PLANT-BASED OPTIONS ARE 'WEIRD,' BUT I GIVE THEM FLAVORS THEY RECOGNIZE, MADE WITH PLANT-BASED INGREDIENTS."
"I WANT TO BREAK STEREOTYPES AND TAKE CARE OF THE ENVIRONMENT WHILE DOING IT."
"THE BEAUTIFUL THING ABOUT COLLECTING PLASTIC IN THE OCEAN, IT MAKES IT GRASPABLE. GO OUT THERE AND PICK SOMETHING UP, THAT'S THE GATEWAY, AND NEXT THING YOU KNOW, YOU'RE DOING SOMETHING BIGGER."
"WE GIVE A PAIR OF WOOL SOCKS TO THE PARTNERS WE WORK WITH. THAT GRANNY SPIRIT IS ALWAYS WITH US."
– Leena
SINCE INCEPTION, WE'VE TAKEN INSPIRATION FROM STREETWEAR AND URBAN CULTURE. OUR FOUNDERS ARE MINORITIES AND OUR CUSTOMERS COME FROM ALL OVER. OUR BACKGROUNDS MAKE US UNIQUE. WE ARE MAKING A $10,000 DONATION TO THE NAACP IN THE FIGHT AGAINST RACIAL DISCRIMINATION AND POLICE BRUTALITY.
#BLACKLIVESMATTER
- TRUFF FAMILY

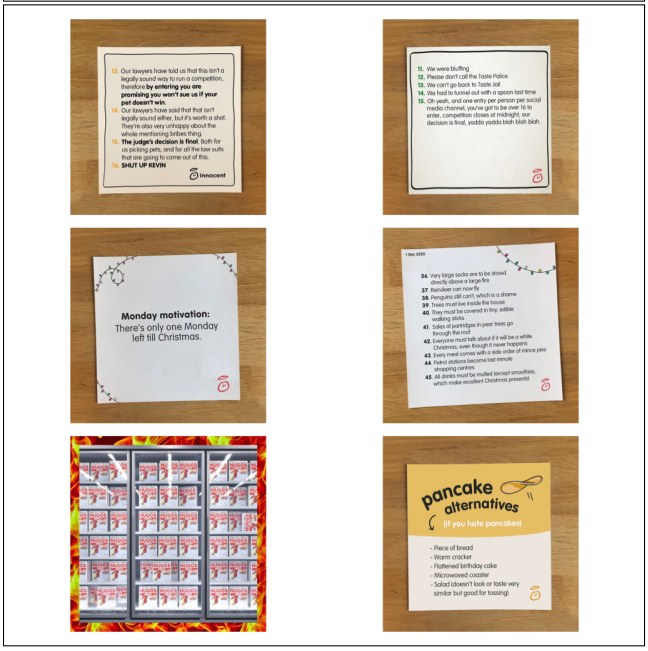

10G
complete protein
TRUFF
Muddy Bud-borne
ALL OF US WHO WORK HERE BELIEVE
BLACK LIVES MATTER


Me when someone touches my neck
Your first @ owes you this giant pizza roll
And they say romance is dead
Tiny hand or big Pizza Stuffer?
Something's not right here






IT'S A MATCH!
SEND MESSAGE
IG: @jailynisfeliz
TikTok: @jailynisfeliz
Twitter: @jailynisfeliz
"Helping others has always been a passion of mine. Giving voice to the voiceless and the unheard. Telling the stories that need to be told and more importantly, encouraging and giving people room to tell their own. Black artists, writers, musicians, and creators of all kinds should be allowed to freely express their talents AND their grievances to the world. No one draws the line for us. I have so many interests I want to explore and create upon, but no matter what I make, I want it to be bigger than me. Impact the life, or even just the day, of one person."
-Jailyn
#BlackVoicesCreate
"WE HAVE TO ACT NOW AND FIND WAYS TO USE OUR SPACE MORE EFFICIENTLY AND ECOLOGICALLY."
IG: @jailynisfeliz
TikTok: @jailynisfeliz
Twitter: @jailynisfeliz
"Helping others has always been a passion of mine. Giving voice to the voiceless and the unheard. Telling the stories that need to be told and more importantly, encouraging and giving people room to tell their own. Black artists, writers, musicians, and creators of all kinds should be allowed to freely express their talents AND their grievances to the world. No one draws the line for us. I have so many interests I want to explore and create upon, but no matter what I make, I want it to be bigger than me. Impact the life, or even just the day, of one person."
-Jailyn
#BlackVoicesCreate
Y'all are so vulnerable asking for dating advice, a friend in the DMs, and more that I think it's time that I share how I really feel. This year, I'm going to write each of you a personalized Valentine's Day card* in an effort to spread the love y'all show this salty snack mix on the daily. Head to the link in my bio to request one.
Love, Chex Mix
* while supplies last


11. Our lawyers have told us that this isn't a legally sound way to run a competition, therefore by entering you are promising you won't sue us if your pet doesn't win.
12. Our lawyers have said that that isn't legally sound either, but it's worth a shot. They're also very unhappy about the whole mentioning bribes thing.
13. The judge's decision is final. Both for us picking pets, and for all the law suits that are going to come out of this.
14. SHUT UP KEVIN
innocent
11. We were bluffing
12. Please don't call the Taste Police
13. We can't go back to Taste Jail
14. We had to tunnel out with a spoon last time
15. Oh yeah, and one entry per person per social media channel, you've got to be over 16 to enter, competition closes at midnight, our decision is final, yadda yadda blah blah blah.
Monday motivation:
There's only one Monday left till Christmas.
1 Dec 2020
34. Very large socks are to be stored directly above a large fire
37. Reindeer can now fly
38. Penguins still can't, which is a shame
39. Trees must live inside the house
40. They must be covered in tiny, edible bolling sticks
41. Sales of partridges in pear trees go through the roof
42. Everyone must talk about if it will be a white Christmas, even though it never happens
43. Every meal comes with a side order of mince pies
44. Petrol stations become last minute shopping centres
45. All drinks must be mulled (except smoothies, which make excellent Christmas presents)
pancake alternatives
(if you hate pancakes)
- Piece of bread
- Warm cracker
- Flattened birthday cake
- Microwaved coaster
- Salad (doesn't look or taste very similar but good for tossing)