arXiv:2301.00889v1 [math.ST] 2 Jan 2023

因果推断中协变量平衡的经验过程框架

埃弗伦·克鲁斯·科尔特斯
密歇根大学数据科学研究所 复杂系统研究中心 密歇根大学
encc@umich.edu

Kevin Josey
生物统计学系 哈佛大学陈曾熙公共卫生学院 kjosey@hsph.harvard.edu

Fan Yang
生物统计学与信息学系 科罗拉多公共卫生学院 fan.3.yang@cuanschutz.edu

Debashis Ghosh
生物统计学与信息学系 科罗拉多公共卫生学院 debashis.ghosh@cuanschutz.edu

摘要

我们提出一种新的视角，通过考察匹配程序所属函数类的复杂度来评估匹配程序。在这一视角下，我们利用有限样本集中不等式，为匹配后协变量平衡提供理论保证。我们将这一框架应用于粗化精确匹配以及使用倾向得分的匹配，并提出如何将其应用于其他算法。我们通过模拟研究评估这些程序。

关键词： 因果效应、经验分布函数、熵度量、超总体、尾部不等式、Vapnik-Chervonenkis 维数。

1 引言

因果推断是结果变量研究和政策研究的核心目标，在医学领域尤其如此。这一广阔研究领域涵盖诸多主题，其中包括利用非随机数据评估处理效应的方法。几乎每个科学学科都拥有大量观察性数据。然而，混杂造成的偏差是观察性研究固有的问题。在这种情况下，研究者必须以某种方式处理每一个潜在的混杂因素，才能确立因果关系。虽然随机化仍是推断的黄金标准，因为按定义它不存在混杂，但在某些研究设计中，将个体随机分配到处理组往往成本过高，有时甚至不符合伦理。

在潜在结果框架下（Neyman，1923年；Rubin，1974年），Rosenbaum和Rubin（1983年）阐明了倾向得分在利用观察性数据进行因果效应估计与推断时所发挥的关键作用。倾向得分定义为给定一组测量协变量时接受处理的概率。在强可忽略性假设下，倾向得分凭借其作为平衡评分的特性，能够消除由混杂造成的偏差（Rosenbaum和Rubin，1983年）。基于这一结果，研究者围绕倾向得分开发了许多因果效应估计方法，

其主要目标是实现协变量平衡（例如，Imai 和 Ratkovic（2014）；Zubizarreta（2015）；Chan 等人（2016））。然而，Rosenbaum和Rubin（1983年）关于倾向得分的结果是在渐近情形下推导得出的。这意味着，在有限样本情形下，倾向得分估计值可能无法充分平衡协变量分布。因此，许多方法会反复拟合倾向得分模型，并评估经倾向得分调整后的协变量的平衡诊断，然后再估计所关注的处理效应。Ho 等人（2007）和 Sekhon（2008）提出了一些评估平衡诊断的方法。倾向得分领域大致分为两个彼此重叠但又各有侧重的方向：一类利用倾向得分计算平衡权重（Hainmueller，2012；Imai 和 Ratkovic，2014；Chan 等人，2016），另一类则利用倾向得分等平衡评分构建匹配队列。

最近，Iacus 等人（2011）提出了一种多变量匹配方法，该方法使用观测协变量的粗化值。他们将其算法称为粗化精确匹配。该方法的主要目标之一，是免去反复重新匹配参与者、直到达到可接受的平衡程度这一迭代步骤。粗化精确匹配本质上相当简单，其基本启发式步骤如下：

1. 对每个混杂变量进行粗化，将其划分为一定数量的类别；
2. 根据粗化值的各种可能组合创建层；
3. 比较各层内处理组的结果变量，并适当调整层效应，以计算因果效应。

Iacus 等人（2011）为粗化精确匹配提供的理论依据，是他们称为“单调不平衡”的概念。他们指出，将混杂因素之间的距离限制在较小范围内，可以得到比基于 Rubin 及其合作者提出的等百分比偏差减少理论（Rubin，1976；Rubin和Thomas，1992；Rubin等人，2006）的程序更灵活的匹配程序。粗化精确匹配的一项主要优势是，它适用于利用大规模数据库查询方法进行因果推断：参见 Salimi 和 Suciu（2016）以及 Wang 等人（2017）。

然而，与逆概率加权估计量等其他方法相比，针对匹配估计量的技术性结果较少。Abadie 和 Imbens（2006）研究了匹配估计量的大样本渐近理论，并发现一般而言，基于匹配的平均因果效应估计量不具有通常的 $n^{1/2}$ 收敛性。其直观解释是，匹配算法会在因果效应估计中引入偏差，而且这种偏差不会随着样本量渐近增大而消失。该偏差项还会随着混杂因素数量的增加而增大。Abadie 和 Imbens（2011）提出了偏差校正估计量。Abadie 和 Imbens（2016）对使用估计倾向得分进行匹配的平均因果效应估计量的渐近行为开展了理论研究。

从概念上讲，实现协变量平衡是一个多元概念。若用 $\mathcal{L}(Z \mid T=0)$ 和$\mathcal{L}(Z \mid T=1)$ 表示给定处理状态时混杂因素的概率律，那么理想情况下，如同完全随机化的情形，这些分布在某种意义上是相等的。我们将这种相等称为协变量平衡。

大多数协变量平衡方法并未将混杂因素的联合分布纳入考虑，而是试图匹配混杂因素边际分布的矩。例如，Imai 和 Ratkovic（2014）提出在其算法中匹配协变量的一阶矩和二阶矩。在实际应用中，人们会使用一维诊断指标（例如比较处理组之间混杂因素的均值，或使用Kolmogorov-Smirnov统计量）来评估平衡。Wang 和 Zubizarreta（2019）指出，鉴于实现多元平衡本身存在复杂性，不如致力于实现混杂因素之间的近似平衡。

在此 pap中，我们提出一种新的理论方法，用于评估和理解协变量平衡。 我们在引入一种距离度量，用于评估两个多元分布彼此之间的接近程度

，并将协变量平衡定义为距离为零。该度量由匹配程序所属的函数族定义。后续的平衡评估取决于对相关函数类性质的理解。本文展示了以下内容：

1. 函数类的使用自然契合于使用概率度量（Zolotarev，1984）来比较概率律；在此情形下，比较的是给定处理条件下混杂因素的多元分布。2. 随后可以利用经验过程理论的结果（Van Der Vaart 和 Wellner，1996；Kosorok，2007）研究函数类的性质，并对理想平衡下匹配程序的收敛速率作出概率性陈述。3. 理想平衡为 Iacus 等人（2011）的方法提供了新的理论性样本外依据，也可用于评估其他算法策略。

基于这一框架，可以将本文中的技术视为一种可扩展策略的开发，旨在实现协变量平衡；从第3节所述的角度来看，该策略的复杂度相对较低。

2 背景与预备知识

2.1 数据结构与因果估计量

将数据表示为 (Y_i, T_i, Z_i)，$i = 1, \ldots, n$，它们是三元组 (Y, T, Z) 的随机样本，其中 Y 表示关注的响应变量，T 表示处理组，Z 是一个 p 维协变量向量。我们假设 T 取值于 $\{0, 1\}$。

我们现在简要回顾潜在结果框架（Rubin，1974年；Holland，1986）。令$\{Y(0)\ Y(1)\}$ 表示所有 n 个研究对象的潜在结果，观测到的响应变量与潜在结果之间的关系为

$$Y = (1 - T)Y(0) + TY(1).$$

在潜在结果框架中，因果效应被定义为基于潜在结果的个体内对比。一种常用的估计目标是平均因果效应，其定义为

$$\text{ACE} = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i(1) - Y_i(0) \right).$$

进行有效因果推断需要许多假设，包括一致性假设、处理正值性假设和强可忽略处理假设（Rosenbaum和Rubin，1983年），定义如下

$$T \perp \{Y(0), Y(1)\} \mid Z. \tag{2.1}$$

假设（2.1）意味着，在给定协变量的条件下，处理分配与潜在结果集合相互独立。处理正值性是指，对于 Z的所有取值，$1 > P(T = 1 \mid Z) > 0$。因此，直观而言，任何个体都有可能接受任一种处理。最后，一致性假设确保观测结果与观测处理下的潜在结果一致。

正如 Imbens 和 Rubin（2015）最近所述，因果推断通过利用观测协变量对分配机制进行建模。该建模自然产生的一个量是倾向得分（Rosenbaum 和 Rubin，1983年），即给定混杂因素时接受处理的概率。倾向得分定义为

$$e(Z) = P(T = 1 \mid Z).$$

根据式（2.1）中的处理可忽略性假设，还可由 Rosenbaum 和 Rubin（1983年）的定理3推出：给定倾向得分时，处理是强可忽略的，即

$$T \perp \{Y(0), Y(1)\} \mid e(Z).$$

基于这些假设和定义，我们可以采用以下方法进行因果推断：(a) 定义合适的因果估计目标；(b) 建立倾向评分模型；(c) 检查协变量平衡；(d) 如果满足(c)，则通过对倾向评分进行条件化来估计因果估计目标。我们注意到，在实践中，步骤(b)和(c)往往需要反复迭代。尽管本文的结果适用于倾向评分匹配分析，但也适用于更一般的匹配策略。

2.2 协变量平衡的既有结果

就协变量平衡而言，一大类理论结果源自对等百分比偏差缩减程序的研究（Rubin和Thomas，1992，1996）。等百分比偏差缩减是指，某种协变量匹配会使 Z 所有维度上的偏差都以相同幅度缩减。

如果匹配程序在协变量的仿射变换下保持不变，则称该匹配方法为仿射不变。若假定在给定 T 的条件下， Z 服从所谓的椭圆对称分布，则Rubin和Thomas（1992）的定理3.1以及推论3.1和3.2适用，因此任何仿射不变的匹配方法都具有等百分比偏差缩减性质。椭圆对称分布的例子包括多元正态分布和t分布。尽管给定处理组时混杂因素服从椭圆对称分布这一假设限制较强，但Rubin等人（2006）的近期研究放宽了这一假设。他们假设在给定 T 的条件下， Z 的条件分布是椭圆对称分布的判别混合。Rubin等人（2006）证明，在这种设定下，等百分比偏差缩减性质的推广形式同样成立。

因此，对于等百分比偏差缩减方法，我们可以保证，尝试提高某个变量的平衡性不会导致其他变量的平衡性失真。然而，在实践中，等百分比偏差缩减成立所需的假设似乎过于严格。Iacus 等人（2011）采取了另一种方法，着重关注样本内协变量差异，并要求处理组和对照组受试者的样本均值之间的最大差异不超过某个常数。他们将这一概念推广到数据的任意函数，并将其称为不平衡界定；他们把这样的匹配方法定义为单调不平衡界定匹配方法：对变量应用单调函数后所产生的差异，由一个特定于混杂因素的项设定上界。因此，可以更严格地要求某个变量达到平衡，而不影响所有混杂因素中的最大不平衡程度。

要求满足单调不平衡界定性质会带来许多重要影响。首先，许多混杂因素调整方法，例如Cochran和Rubin（1973年）所定义的最近邻匹配或卡尺匹配，并不属于单调不平衡界定方法，因为它们会固定各层内处理组和对照组观测的数量，而单调不平衡界定方法则允许观测数量变化。相比之下，如果卡尺匹配程序允许针对每个混杂因素使用不同的卡尺，那么它就属于单调不平衡界定方法。

Iacus 等人（2011）还指出，因果效应估计的一个关键目标是降低模型依赖性（Ho 等人，2007），也就是说，不应将潜在结果外推到协变量空间中没有观测值的区域。在对潜在结果模型作出一些假设的条件下，他们证明，对于单调失衡界定方法，模型依赖性有一个上界，该上界由涉及失衡参数的项构成。此外，使用单调不平衡界定匹配方法估计平均因果效应时，估计误差也可以由涉及该参数的项给出上界。

作为一个具体的新型单调不平衡界定方法示例，Iacus 等人（2011）提出了一种更粗用于创建层的精确匹配算法。其步骤如下：

1. 对于每个变量 $Z_j (j = 1, \ldots, p)$，将其粗化为函数 $C_j(Z_j)$，使其取值少于 Z_j 的唯一值；

2. 使用向量在处理组与对照组观测之间进行精确匹配

$$(C_1(Z_1), C_2(Z_2), \ldots, C_p(Z_p)) .$$

这实际上会基于以下各项的唯一组合创建层 $\mathcal{S}_1, \ldots, \mathcal{S}_J$

$$(C_1(Z_1), C_2(Z_2), \ldots, C_p(Z_p)) .$$

3. 丢弃仅包含 $T = 0$ 观测的层。对于仅包含来自 $T = 1$ 总体的观测的层，使用可用的对照组对潜在结果 $Y(0)$ 进行外推，或者通过将感兴趣的因果效应限定为无需进一步基于模型的假设即可识别因果效应的处理组单位来舍弃这些层。对于同时包含处理组和对照组观测的层，则比较两个总体的结果变量。

Iacus 等人（2011）开发了非常易于使用的软件包，可在 R 和 Stata 中实施粗化精确匹配。他们证明，粗化精确匹配方法对于各种感兴趣的泛函都满足单调不平衡界定性质。此外，他们还对粗化精确匹配试图模拟的对象作出了非常直观的解释。传统的倾向得分方法试图模拟随机化研究，而采用粗化精确匹配的分析则模拟随机区组设计，其中区组按定义能够预测潜在结果。众所周知，在这种情况下，随机区组设计会产生更有效率的估计量（例如，Box、Hunter 和 Hunter，1978）。

另一种近年来受到关注的方法，是将协变量平衡纳入因果效应估计过程。例如，Imai 和 Ratkovic（2014）提出使用广义矩估计法进行因果效应估计，并在该程序中将协变量平衡视为一项约束。Chan 等人（2016）提出使用校准估计量进行因果效应估计，其中协变量平衡约束会导出一个受约束的拉格朗日对偶优化问题。对于这些方法，作者能够推导出因果效应估计量的一致性和渐近正态性结果。

如第 3.1 节将更详细说明的那样，我们将使用积分概率度量来评估两个总体之间的协变量平衡。Kallus（2020）也使用了类似的度量。他们将这种度量定义为目标误差，并以此为目标，在估计处理组平均处理效应时求得最优加权系数。虽然我们的方法与之互为补充，但两者仍有几个显著差异。首先，在 Kallus（2020）中，他们使用该度量来寻找与已知匹配方法相对应的权重。其度量中涉及的函数表示潜在结果与协变量之间的期望关系。而在我们的方法中，我们采用任意匹配程序，并根据匹配的测度，利用涉及表示匹配程序本身的函数的概率度量对其进行界定，同时给出匹配效果的概率界限。此外，在 Kallus（2020）中，他们假设总体固定，因此协变量取值不存在随机性；而我们关注的则是这些协变量的样本分布。这两种方法之间的差异将在第 2.3 节进一步解释。

2.3 推断模式与协变量平衡

在考察各种实现协变量平衡的方案时，重新思考进行因果推断的方式会很有帮助。Imbens 和 Rubin（2015）对因果推断中有限总体模式与超总体模式的区别作了很好的概述。有限总体模式下的因果推断将抽样单位视为研究总体。实验的随机性完全源自处理机制，因此随机性仅涉及处理分配。如果采用有限样本视角

进行因果推断，就可以使用基于随机化的方法来推断因果效应。

相比之下，超总体推断模式考虑两种变异来源。第一种源于处理分配的随机性，第二种源于抽样单元是从超总体中随机抽取的样本。因此，这种方法假定存在一个超总体，抽样单元来自其中。

回顾 2.2 节的相关工作，等百分比偏差减少理论以及 Iacus 等人（2011）的研究，在有限总体因果推断模式的假设下，提出了关于协变量平衡的结果。因此，这些方法的协变量平衡结果涉及对抽样单元进行子抽样和匹配，且平衡是针对匹配样本而言的。我们将在下一节引入的平衡概念可以适用于这两种推断模式。

3 主要结果

3.1 理想平衡

本节中，我们希望从比较分布$\mathcal{L}(Z \mid T=0)$和 $\mathcal{L}(Z \mid T=1)$的角度研究协变量平衡。为此，我们必须确定如何进行这种比较。我们首先定义概率伪度量。

定义 3.1（伪度量）。设 $\mathcal{A}$ 为定义在同一个可测空间上的概率测度集合。函数 $m: \mathcal{A} \times \mathcal{A} \to [0, \infty)$ 若对所有 μ、 ν、 $\lambda \in \mathcal{A}$ 均满足以下条件，则称其为 $\mathcal{A}$ 上的**伪度量**:

1. $m(\mu, \mu) = 0$。2. $m(\mu, \nu) = m(\nu, \mu)$。3. $m(\mu, \nu) \leq m(\mu, \lambda) + m(\lambda, \nu)$。

请注意，这些性质几乎使 m 成为 $\mathcal{A}$上的度量，但值得注意的是，我们并未假设：若两个元素之间的距离为零，则这两个元素相同。本文中我们将宽泛地使用术语，把伪度量称为度量。

本文将使用的度量类定义如下:

$$\gamma_{\mathcal{F}}(\mu, \nu) = \sup_{f \in \mathcal{F}} \left| \int f d\mu - \int f d\nu \right|, \tag{3.1}$$

其中 $\mathcal{F}$是一个函数类。在（3.1）中，Zolotarev（1984）将 $\gamma_{\mathcal{F}}(\mu, \nu)$ 称为概率度量的一个例子。在我们的记法中，我们省略 $\gamma_{\mathcal{F}}$ 对 $\mathcal{F}$ 的依赖，并将其写作 γ。现在，我们将基于（3.1）定义理想平衡。

定义 3.2（理想平衡）：设 μ 和 ν 是定义在同一概率空间上的分布， m 是一个伪度量，则称 μ 和 ν 相对于 m 满足**理想平衡**，如果 $m(\mu, \nu) = 0$。

当 μ 和 ν 是第 2 节所述的给定处理组的协变量条件分布时，理想平衡是对总体的约束。如果它们是数据的经验分布，那么理想平衡就是样本约束。从某种意义上说，匹配方法旨在使匹配数据在某个 m下达到理想平衡。

请注意，在这一阶段，我们只讨论了总体分布定律，并未说明如何利用真实数据估计或计算这些量。在实践中，我们不会期望观察性研究能够达到理想平衡。不过，理想平衡确实可作为一个有用的基准，让我们据此研究各种函数约束的性质。这里，(3.1) 中的函数空间 $\mathcal{F}$ 起到了约束的作用；更复杂的函数空间则意味着对 $Z|T=1$ 和 $Z|T=0$的联合分布施加了更多约束。

3.2 一个集中不等式结果

设 $\mathcal{F}$为一个函数空间，$\|\cdot\|$为一个范数。覆盖数 $N(\epsilon, \mathcal{F}, \|\cdot\|)$是覆盖 $\mathcal{F}$所需的半径为 ϵ 的 $\|\cdot\|$-球的最小数量，其中以 $f \in \mathcal{F}$为中心的球是集合 $\{g \mid \|f-g\| \leq \epsilon\}$。直观地说，可以将覆盖数视为函数类 $\mathcal{F}$复杂度的一种度量。对于测度 μ，范数 $L_r(\mu)$-范数（针对 $r \geq 1$）定义为$\|f\|^r_{L_r(\mu)} = \int |f|^r d\mu$。本文始终假设 $\mathcal{F}$一致有界。请注意，若 μ是任意概率测度，且满足一致有界性，则可在 $\mathcal{F}$上赋予范数$L_r(\mu)$，而不必删去其中任何元素。除非另有说明，我们假设 $\mathcal{F}$ 中函数的值域为 $[0,1]$。最后，对于函数类 $\mathcal{F}$，其包络函数定义为任意满足以下条件的函数 h：对 $\mathcal{F}$ 中的所有 f，不等式

$$|f(x)| \leq |h(x)|$$

对任意 x均成立。

设 $\{Z_i\}_{i=1}^n$ 是一个样本，其中每个 Z_i 都服从分布 Q。我们用 $\mathbb{Q}_n$表示经验分布。以 $\mathcal{F}$为索引的经验过程 $\mathbb{G}_n^Q$ 定义为将任意 $f \in \mathcal{F}$ 映射到

$$\mathbb{G}_n^Q(f) = \sqrt{n}\left(\int f d\mathbb{Q}_n - \int f dQ\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(f(Z_i) - \int f dQ\right).$$

定理 3.3. 设 $\mathbb{Q}_{n_0}^0$ 和 $\mathbb{Q}_{n_1}^1$ 分别是从 Q^0 和 Q^1中抽取的观测值所构成的两个经验分布，并假设相对于 γ，Q^0 和 Q^1 满足理想平衡。设 M 为概率测度集合。若存在常数 C 和 K，使得 $\mathcal{F}$ 满足

$$\sup_{\mu \in M} N(\epsilon, \mathcal{F}, \|\cdot\|_{L_r(\mu)}) \leq \left(\frac{K}{\epsilon}\right)^C,$$

对每个 $0 < \epsilon < C$，则

$$Pr\{\gamma(\mathbb{Q}_{n_0}^0, \mathbb{Q}_{n_1}^1) > \delta\} \leq \left(\frac{D\delta}{2\sqrt{C}}\right)^C \left(n_0^{C/2} \exp(-n_0\delta^2/2) + n_1^{C/2} \exp(-n_1\delta^2/2)\right), \tag{3.2}$$

其中 D 是一个仅取决于 K 的常数。

定理 3.3 的证明及后续结果见补充材料。全文中，我们将用 $B_n(\delta, D, C)$ 表示定理 3.3 中的界，其中下标n 提醒我们该界取决于样本量。

注释 3。 4。我们注意到，(3.2) 中的界是非渐近的，对任意样本量均成立。 .

注释 3.5。在此框架中，函数类发挥着重要作用。定理 3.3 给出了一个关于相关函数类熵数的界。特别是，该方法倾向于选择低复杂度函数。一个关键的技术问题是确保满足定理中的覆盖数条件。为此，我们将主要运用 Vapnik-Chervonenkis 理论（Chervonenkis 和 Vapnik，1971）的结果来确定适当的覆盖数。

在大多数情况下，我们感兴趣的函数类并非实值函数类，而是向量值函数类。以下简单的结果可用于处理这些情况。

引理 3.6。设 $\{\mathcal{F}_i\}_{i=1}^d$ 为实值函数空间的一个集合，且对每个 $1 \leq i \leq d$，(P^i, Q^i) 在 $\gamma_{\mathcal{F}_i}$ 下满足理想平衡。令$(\mathbb{P}^i, \mathbb{Q}^i)$ 表示它们各自的经验分布，并隐含地依赖于样本量。则

$$Pr\left(\sum_{i=1}^d \gamma_{\mathcal{F}_i}(\mathbb{P}^i, \mathbb{Q}^i) > \delta\right) \leq \sum_{i=1}^d B(\delta/d, D_i, C_i).$$

现在，考虑集合 $\{\mathcal{F}_i\}_{i=1}^d$，其中每个 $\mathcal{F}_i$ 都是实值函数空间。对所有 $i\}$，定义 $\mathcal{F}=\{f=(f_1,$ ……， $f_d)^T \mid f_i \in \mathcal{F}_i$。令 π_ℓ 为坐标投影，即对于有限维向量 $x=(x_1,$ ……， $x_d)$， $\pi_\ell(x)=x_\ell$。最后，定义 $\mathcal{F}_\pi=\{\pi_\ell \circ f \mid f \in \mathcal{F}, 1 \leq \ell \leq d\}$。注意， $\mathcal{F}_\pi$ 中的元素均为实值。下面的引理说明，我们可以假设 μ 和ν 相对于每个 $\gamma_{\mathcal{F}_i}$满足理想平衡，也可以假设它们相对于$\gamma_{\mathcal{F}_\pi}$满足理想平衡。

引理 3.7。 设 $\mathcal{F}$、 $\{\mathcal{F}_i\}_{i=1}^d$和 $\mathcal{F}_\pi$ 如上所述，并设 μ 和 ν 表示两个概率测度。则以下条件等价：

1. μ 和 ν 相对于 $\gamma_{\mathcal{F}_\pi}$满足理想平衡；
2. μ 和 ν 分别相对于 $\gamma_{\mathcal{F}_i}$、 $1 \leq i \leq d$满足理想平衡。
3. $\max_i \gamma_{\mathcal{F}_i}(\nu, \mu)=0$。

以下推论将非常有用：

推论 3.8。 设 $\mathcal{F}$ 和 $\mathcal{F}_\pi$ 如上所述，且对所有 i， $\mathcal{F}_i=\mathcal{F}^*$ 成立。假设 $\mathcal{F}^*$ 具有多项式覆盖数。令 $\{X_j^0\}_{j=1}^{n_0} \sim Q^0$ 和 $\{X_j^1\}_{j=1}^{n_1} \sim Q^1$，其中 Q^0 和 Q^1 相对于 $\gamma_{\mathcal{F}_\pi}$满足理想平衡。固定 $f^* \in \mathcal{F}$，则

$$Pr\left(\left\|\frac{1}{n_0}\sum_{j=1}^{n_0} f^*(X_j^0)-\frac{1}{n_1}\sum_{j=1}^{n_1} f^*(X_j^1)\right\|_{\ell_p} > \delta\right) \leq dB(\delta/d^{1/p}, D^*, C^*),$$

对于有限的 $p \geq 1$，且

$$Pr\left(\left\|\frac{1}{n_0}\sum_{j=1}^{n_0} f^*(X_j^0)-\frac{1}{n_1}\sum_{j=1}^{n_1} f^*(X_j^1)\right\|_{\ell_\infty} > \delta\right) \leq dB(\delta, D^*, C^*),$$

其中 D^*、 C^* 仅取决于 $\mathcal{F}^*$。

定义 3.9（Vapnik-Chervonenkis 维数）。 函数类 $\mathcal{F}$ 在环境集合 $\mathcal{X}$ 上的**Vapnik-Chervonenkis** 维数，是被 $\mathcal{F}$ 打散的最大子集的基数。如果对于 $S \in \mathcal{X}$ 的每一种可能的 $0-1$ 标记方式，都至少存在一个函数 $f \in \mathcal{F}$ 能够实现这种标记，则称函数类 $\mathcal{F}$ 打散集合 $S \in \mathcal{X}$。

我们将使用的一个关键结果是 Van Der Vaart 和 Wellner（1996）定理 2.6.7 的一个推论，该结果表明，如果函数类 $\mathcal{G}$ 的 Vapnik-Chervonenkis 维数 v 有限，那么

$$\sup_\mu N(\epsilon, \mathcal{G}, L_2(\mu)) \leq \left(\frac{K}{\epsilon}\right)^{C^*},$$

其中 $C^*=2v-2$。

4 个示例

4.1 粗化函数类上的平衡

考虑 Iacus 等人（2011）所述的粗化精确匹配。设 $\mathcal{Z}_0=\{Z_i^0\}_{i=1}^{n_0}$ 和$\mathcal{Z}_1=\{Z_j^1\}_{j=1}^{n_1}$ 分别为对照样本和处理样本。在粗化精确匹配中，我们将样本空间划分为若干分区，并匹配位于同一分区中的样本

，同时舍弃那些不含对侧组样本的子集中的样本。我们关注的量为

$$\Delta = \frac{1}{m_0}\sum_{i\in M_0} w_i^0 Z_i^0 - \frac{1}{m_1}\sum_{j\in M_1} w_j^1 Z_j^1,$$

其中， m_ℓ 是 ℓ^{th} 组的匹配样本数， M_ℓ 是其索引集，而$\{w_i^0, w_j^1\}_{i\in M_0, j\in M_1}$是权重。

在补充材料中，我们说明如何将此匹配程序表示为关于变量 Z_i^0和 Z_j^1的函数f。这使我们能够用 f表示 Δ。我们进一步指定函数空间 $\mathcal{F}$，使得

$$\|\Delta\| \le \gamma_{\mathcal{F}}(\mathbb{Q}_{n_0}^0, \mathbb{Q}_{n_1}^1)$$

在适当的范数下成立。利用 $\mathcal{F}$ 的性质并结合上述界限，我们可以推导出感兴趣的结果：

$$Pr(|\Delta_k| \ge \delta) \le B(\delta, D, C^*),$$

其中，常数为 C^*，且 Δ_k 是 Δ的 k^{th} 分量。类似地，

$$Pr(\|\Delta\|_{\ell_p} \ge \delta) \le dB(\delta/d^{1/p}, D, C^*)$$

and

$$Pr(\|\Delta\|_{\ell_\infty} \ge \delta) \le dB(\delta, D, C^*).$$

4.2 线性倾向得分上的协变量平衡

如第3节所述，已有大量研究基于线性判别分析推导匹配结果。也就是说，我们假设 $P(Z \mid T = \ell)$ 服从 $N(\mu_\ell, \Sigma)$。在此模型下，所考虑的度量是倾向得分的 $logit$（见 Stuart（2010））。在补充材料中，我们证明距离 $|logit(e(Z)) - logit(e(Z')|$ 可以用线性判别分析超平面向量表示。事实上，若协变量的维数为 p，我们可以构造一个由超平面导出的函数空间 $\mathcal{F}$，其 Vapnik-Chervonenkis 维数为$p+1$，使得

$$\begin{aligned}\Delta &= \left|\frac{1}{m_0}\sum_{i\in M_0} logit(e(Z_i)) - \frac{1}{m_1}\sum_{j\in M_1} logit(e(Z_j))\right| \\ &\le \gamma_{\mathcal{F}}(\mathbb{Q}_{n_0}^0, \mathbb{Q}_{n_1}^1),\end{aligned}$$

从而使我们能够利用定理 3.3 确定所关注的界：

$$Pr\{\Delta > \delta\} \le B(\delta, D, 2p).$$

4.3 使用核方法实现协变量平衡

许多作者（Hazlett，2016；Wong和Chan，2018；Zhu 等人，2018）主张使用核方法进行匹配并评估协变量平衡。这相当于假设 (3.1) 中的 $\mathcal{F}$ 表示一个再生核希尔伯特空间。关于这些函数空间的更多细节，请参见补充材料。

为了将定理 3.3 应用于核设定，我们注意到，第4.2节中的线性判别分析有一种形式可以扩展到再生核希尔伯特空间设定（Baudat和Anouar，2000）。设 $\mathcal{H}$ 为一个再生核希尔伯特空间，$\|\cdot\|_{\mathcal{H}}$ 为与之关联的范数，那么核化匹配程序中一个自然的度量是

$$\Delta_{\mathcal{H}} = \left\|\frac{1}{m_0}\sum_{i\in M_0} f(Z_i) - \frac{1}{m_1}\sum_{j\in M_1} f(Z_j)\right\|_{\mathcal{H}},$$

它是 Δ在第4.2节中的泛函推广，其中 $f \in \mathcal{H}$ 是用户选择的一个适当函数。于是 $\Delta_{\mathcal{H}} \leq \gamma_{\mathcal{F}}(\mathbb{Q}_{n_0}^0, \mathbb{Q}_{n_1}^1)$，我们只需稍作调整即可使用前面的结果。我们在补充材料中证明了

$$P(\Delta_{\mathcal{H}} > \delta) \leq B(\delta, D, C^*),$$

其中， C^* 取决于 $\mathcal{H}$的光滑性。

5 实际应用

到目前为止，我们已经给出了理论结果，说明在理想平衡假设下，各种函数类对应的算法会如何表现。如前所述，理想平衡定义虽然严格，但能够对各种算法进行理论刻画。于是自然会产生一个问题：如何在实践中运用前几节的理论结果?

请注意，可以将公式 (3.1) 中的度量视为一种多元平衡度量，这使它有别于文献中的许多其他平衡度量。Zhu 等人（2018）使用了公式 (3.1)，其中$\mathcal{F}$ 是一个再生核希尔伯特空间，并将其作为协变量平衡诊断。他们发现，在某些情况下，与单变量诊断以及基于预后评分的诊断相比，该诊断对发现协变量不平衡更为敏感（Hansen，2008）。

考虑估计处理组平均因果效应的问题。在实践中，处理组和对照组总体不太可能达到理想平衡。也就是说，$\gamma_{\mathcal{F}}\left(Q^0,\ Q^1\right) \neq 0$，除非处理是随机分配的。因此，在观察性研究中，我们无法使用定理 3.3。不过，可以稍作修改，使分析过程基本保持不变。

设 $w \in \mathcal{W} \subset \mathbb{R}^{n_0}$ 为一个权重向量，并定义

$$\mathbb{Q}_w^0 = \frac{1}{\sum_{i:T_i=0} w_i} \sum_{i:T_i=0} w_i \delta_{X_i}.$$

因果推断中的大多数方法都旨在找到合适的权重 w，使得$\mathbb{Q}_w^0$ 收敛到某个满足与 Q^1的理想平衡的分布 Q^*，即满足 $\gamma_{\mathcal{F}}\left(Q^*,\ Q^1\right) = 0$。为了使这一修改可行，我们只需修改定理 3.3 的证明，并纳入 $\mathbb{Q}_w^0$ 收敛到 Q^*的收敛速率；该速率可能因问题而异。完成后，我们便可沿用类似的论证方式。

令 $f^* \in \mathcal{F}$ 表示一种具有平衡诊断的匹配程序

$$\Delta = \left| \int f d\mathbb{Q}_w^0 - \int f d\mathbb{Q}_{n_1}^1 \right|,$$

那么，根据 $\gamma_{\mathcal{F}}$的定义，

$$\Delta \leq \gamma_{\mathcal{F}}\left(\mathbb{Q}_w^0, \mathbb{Q}_{n_1}^1\right).$$

因此，如果我们能找到满足以下条件的权重： $\mathbb{Q}_w^0$ 收敛到 Q^*，且 $\gamma_{\mathcal{F}}(Q^*,\ Q^1) = 0$，那么我们就能界定 Δ 超过某个阈值 δ的概率。

有许多方法可以求得 $w \in \mathcal{W}$，其中最直接的方法是使用逆概率 a-的 tr 能力疗权重，

$$w_i = T_i + \frac{e(Z_i)(1 - T_i)}{1 - e(Z_i)}.$$

即使是因果推断文献中常见的、规定严格的匹配算法，也会得到一些权重 $w \in \mathcal{W}$，如 Abadie 和 Imbens (2006) 所述。在有放回的一对一匹配中，令 $\mathcal{J}(i) = \{j_1(i)\ j_2(i)\ \ldots\}$ 表示与单元

单元 $i = 1, 2, \ldots\ldots, n$相匹配的索引集合。若不存在并列情况，则 $\mathcal{J}(i) = j(i)$。若存在并列情况（这种情况经常出现，尤其是在精确匹配中（参见粗化精确匹配）），$\mathcal{J}(i)$可能包含多个匹配索引。通过求解，我们可以在匹配过程中为每个单元生成权重

$$w_i = \sum_{\{l:T_l=1\}} \frac{I[i \in \mathcal{J}(l)]}{\#\mathcal{J}(l)} \text{ for all } i \in \{i : T_i = 0\}$$

其中 $\#\mathcal{J}(i)$ 表示 $\mathcal{J}(i)$ 的基数。

6 模拟研究

我们进行了一项模拟研究，以评估第4节中报告的距离的分布。我们还考察这些距离对估计处理组平均处理效应所产生的下游影响。我们考虑了两种数据生成机制。此外，我们改变样本量和响应变量的方差，共设置八种情景。对于下文所述的每种情景，我们均重复模拟1000次。我们报告了第4节（表1）中考察的三个距离（Δ）的均值和蒙特卡罗标准误，并给出了一个具有代表性的情景的核密度估计（图1）。我们还使用 Matching 软件包（Sekhon，2008）实现 Abadie 和 Imbens（2006）所述的一对一匹配方法，评估这些 Δ 统计量对平均处理效应的下游影响（表2和表6）。

对于 $i = 1, 2, \ldots\ldots, n$，令 $Z_{i1} \sim \mathcal{N}(1, 4)$、$Z_{i2} \sim \text{Bin}(1, 0.3)$、$Z_{i3} \sim \mathcal{N}(0, 1)$ 和 $Z_{i4} \sim \text{Bin}(1, 0.5)$，其中 T_i 表示二元处理分配。处理组的结果变量条件均值 $\mu_1(Z_i)$ 和对照组的结果变量条件均值 $\mu_0(Z_i)$ 构造如下

$$\begin{aligned} \mu_0(Z_i) &= 10 - 3Z_{i1} - Z_{i2} + Z_{i3} + 3Z_{i4} \text{ and} \\ \mu_1(Z_i) &= \mu_0(Z_i) + 5 + 3Z_{i1} - Z_{i2} + Z_{i3} - 3Z_{i4}. \end{aligned} \tag{6.1}$$

我们从 $T_i \sim \text{Bin}(1, 0.5)$ 分布中抽样。对于 $i = 1, 2, \ldots, n$，我们抽取反事实响应$Y_i(1) \sim \mathcal{N}[\mu_1(Z_i), \sigma^2]$和 $Y_i(0) \sim \mathcal{N}[\mu_0(Z_i), \sigma^2]$。观测结果为 $Y_i = T_iY_i(1) + (1 - T_i)Y_i(0)$。我们将这些条件称为“基线”。对于误差方差，我们设定$\sigma^2 \in \{5, 10\}$。

对于标记为“稀疏”的情景，我们加入一组额外的协变量，但这些协变量最终不会影响结果变量。结果变量由式 (6.1) 中的潜在结果模型决定，但我们考虑的方法也会纳入噪声协变量 $Z_{i5} \sim \mathcal{N}(-1, 4)$、$Z_{i6} \sim \text{Bin}(1, 0.7)$、$Z_{i7} \sim \mathcal{N}(0, 1)$ 和 $Z_{i8} \sim \text{Bin}(1, 0.5)$。

如前所述，我们测试第4节所述的三个示例，评估它们能否对处理组的平均处理效应进行有效且无偏的估计。线性判别分析将 f 设为每个单位接受处理的拟合后验概率的 Logit 变换。支持向量机示例采用在线性核假设下各点到所得分离超平面的距离。粗化精确匹配的实施方式与 Iacus 等人（2011）的描述类似，并使用 cem R 软件包实现。表1展示了模拟实验的结果。由于本次模拟中随机化已实现平衡，我们还报告了未匹配粗略的平均因果效应估计值，以供参考。此处，Δ 表示未加权协变量的样本均值差绝对值的最大值。

在此示例中，Δ 的值不一定可以直接比较。它们确实代表了我们在定理中界定其尾概率的分布。模拟旨在刻画这些统计量的一些密度，以便我们更好地了解，对于第4节中的不同平衡方法，δ 的哪些取值是可以接受的。我们看到，粗化精确匹配后 Δ的取值最为集中，其次是以下取值

n	σ^2	情景	θ	A	B	C	D
1000	5	基线	6.2	0.11 (0.07)	0.03 (0.02)	0.02 (0.01)	0.09 (0.04)
1000	5	稀疏	6.2	0.15 (0.07)	0.01 (0.01)	0.03 (0.02)	0.13 (0.05)
1000	10	基线	6.2	0.12 (0.07)	0.03 (0.02)	0.02 (0.01)	0.09 (0.05)
1000	10	稀疏	6.2	0.15 (0.07)	0.01 (0.01)	0.03 (0.02)	0.13 (0.05)
2000	5	基线	6.2	0.08 (0.05)	0.02 (0.01)	0.01 (0.01)	0.06 (0.03)
2000	5	稀疏	6.2	0.11 (0.05)	0.01 (0.01)	0.02 (0.01)	0.09 (0.04)
2000	10	基线	6.2	0.08 (0.05)	0.02 (0.01)	0.01 (0.01)	0.06 (0.03)
2000	10	稀疏	6.2	0.11 (0.05)	0.01 (0.01)	0.02 (0.01)	0.09 (0.04)

表1：实验中发现的 Δ 的平均值和蒙特卡罗标准误。在此表中，方法A为未加权估计，方法B指粗化精确匹配，方法C指线性判别分析，方法D指支持向量机。由于A和B都会生成向量值 Δ ，我们报告其最大值。

由线性判别分析生成的取值。支持向量机的平衡诊断以及未加权比较所得的取值则分散得多。

不同 Δ 估计值之间可以直接比较的一点，是各种平衡方法在估计平均处理效应时产生的下游影响。本部分模拟研究旨在说明， Δ 分布的集中程度可能与平均处理效应估计的实际质量——因果推断最终关注的结果——关系不大。尽管粗化精确匹配下 Δ分布的集中程度，比线性判别分析和支持向量机下 Δ的其他密度都更窄，但其估计的平均处理效应也偏差最大。其蒙特卡罗标准误似乎也高于另外两种平衡方法。线性判别分析同样使 Δ统计量的分布较为集中，但除蒙特卡罗标准误最小的未加权估计外，它产生的平均处理效应估计最有效。这一结果很有意思，因为未加权诊断中 Δ 的取值分散程度最大。因此，我们认为，在评估平衡并据此判断哪种方法最适合评估处理效应时，必须仔细考虑 Δ 统计量的尺度。

n	σ^2	情景	θ	A	B	C	D
1000	5	基线	6.2	6.20 (0.33)	6.24 (0.33)	6.20 (0.42)	6.20 (0.36)
1000	5	稀疏	6.2	6.20（0.34）	6.29（1.24）	6.21（0.45）	6.20（0.39）
1000	10	基线	6.2	6.20 (0.37)	6.22 (0.40)	6.20 (0.47)	6.20 (0.42)
1000	10	稀疏	6.2	6.19 (0.35)	6.31 (1.46)	6.20 (0.46)	6.22 (0.42)
2000	5	基线	6.2	6.19 (0.24)	6.21 (0.24)	6.20 (0.29)	6.20 (0.25)
2000	5	稀疏	6.2	6.20 (0.23)	6.34 (0.71)	6.21 (0.29)	6.21 (0.26)
2000	10	基线	6.2	6.21 (0.25)	6.21 (0.26)	6.19 (0.32)	6.21 (0.28)
2000	10	稀疏	6.2	6.21 (0.25)	6.38 (0.79)	6.21 (0.31)	6.21 (0.27)

表 2：模拟估计值和蒙特卡罗标准误汇总。与“基线”和“稀疏”对应的模拟场景将在第 6 节中进一步详细介绍。这里， θ指处理组中的总体平均处理效应。在此表中，方法A为未加权估计，方法B指粗化精确匹配，方法C为线性判别分析，方法D为支持向量机。

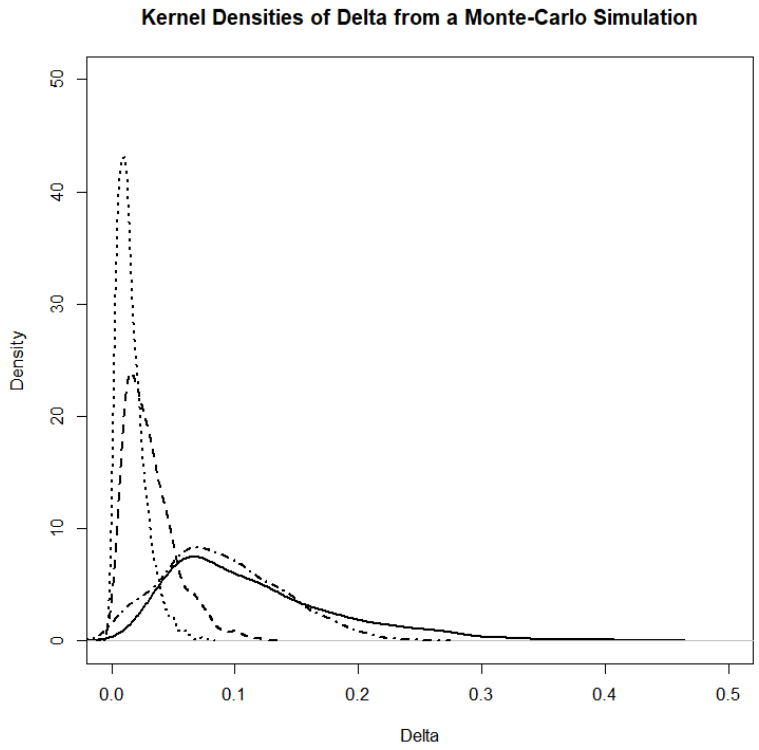


图 1：基准情景下 Δ 平衡统计量的核密度，其中 $n = 1000$ 和$\sigma^2 = 10$。实线表示未加权估计值的分布，虚线表示粗化精确匹配的分布，点线表示线性倾向得分的分布，点划线表示支持向量机示例的分布。

n	σ^2	情景	θ	A	B	C	D
1000	5	基线	6.2	0.952	0.937	0.941	0.929
1000	5	稀疏	6.2	0.944	0.955	0.934	0.917
1000	10	基线	6.2	0.941	0.918	0.935	0.912
1000	10	稀疏	6.2	0.955	0.950	0.951	0.931
2000	5	基线	6.2	0.931	0.945	0.937	0.923
2000	5	稀疏	6.2	0.956	0.945	0.939	0.918
2000	10	基线	6.2	0.959	0.936	0.926	0.928
2000	10	稀疏	6.2	0.953	0.946	0.948	0.935

表3：模拟实验中覆盖概率的汇总。与“基线”“交互作用”“正值性”和“稀疏”相对应的模拟情景将在第 6 节中进一步详细说明。此处， θ 指处理组中的总体平均处理效应。在本表中，方法 A 为未加权估计，方法B指粗化精确匹配，方法C指线性判别分析，方法D指支持向量机。

致谢

作者谨此感谢以下机构提供的资金支持：美国国立卫生研究院、美国国家科学基金会、美国退伍军人管理局以及 Grohne-

科罗拉多大学癌症中心的 Stepp捐赠基金。

附录

定理 3.3 的证明

为减轻读者的符号负担，我们将使用 P 和 Q，而非 Q^0 和 Q^1。

证明。 根据 γ的定义：

$$\begin{aligned}
\gamma(\mathbb{P}_{n_0}, \mathbb{Q}_{n_1}) &= \sup_{f\in\mathcal{F}} \left| \int f d\mathbb{P}_{n_0} - \int f d\mathbb{Q}_{n_1} \right| \\
&= \sup_{f\in\mathcal{F}} \left| \int f d\mathbb{P}_{n_0} \pm \int f dP \pm \int f dQ - \int f d\mathbb{Q}_{n_1} \right| \\
&\leq \sup_{f\in\mathcal{F}} \left| \int f d\mathbb{P}_{n_0} - \int f dP - \int f d\mathbb{Q}_{n_1} + \int f dQ \right| + \sup_{f\in\mathcal{F}} \left| \int f dP - \int f dQ \right| \\
&= \sup_{f\in\mathcal{F}} \left| \int f d\mathbb{P}_{n_0} - \int f dP - \int f d\mathbb{Q}_{n_1} + \int f dQ \right|,
\end{aligned}$$

因为 $\gamma(P,\ Q) = 0$。利用初等概率论论证，我们有

$$\begin{aligned}
Pr\{\gamma(\mathbb{P}_{n_0}, \mathbb{Q}_{n_1}) > \delta\} &= Pr\left(\sup_{f\in\mathcal{F}} \left| \int f d\mathbb{P}_{n_0} - \int f dP - \int f d\mathbb{Q}_{n_1} + \int f dQ \right| > \delta \right) \\
&= Pr\left(\sup_{f\in\mathcal{F}} \left| \frac{1}{\sqrt{n_0}} \mathbb{G}^P_{n_0}(f) - \frac{1}{\sqrt{n_1}} \mathbb{G}^Q_{n_1}(f) \right| > \delta \right) \\
&\leq Pr\left(\sup_{f\in\mathcal{F}} |\mathbb{G}^P_{n_0}(f)| > \sqrt{n_0}\delta/2 \right) + Pr\left(\sup_{f\in\mathcal{F}} |\mathbb{G}^Q_{n_1}(f)| > \sqrt{n_1}\delta/2 \right),
\end{aligned}$$

其中 $\mathbb{G}^P_{n_0}(f)$ 和 $\mathbb{G}^Q_{n_1}(f)$ 分别表示由 P 和 Q 索引的 $\mathcal{F}$ 经验过程。

应用 Van Der Vaart 和 Wellner（1996）中的定理2.14.9，我们可以对各项作如下界定：

$$\begin{aligned}
Pr\left(\sup_{f\in\mathcal{F}} |\mathbb{G}^P_{n_0}(f)| > \sqrt{n_0}\delta/2 \right) &< \left(\frac{D\sqrt{n_0}\delta}{2\sqrt{C}} \right)^C \exp(-n_0\delta^2/2) \\
Pr\left(\sup_{f\in\mathcal{F}} \left|\mathbb{G}^Q_{n_1}(f)\right| > \sqrt{n_1}\delta/2 \right) &< \left(\frac{D\sqrt{n_1}\delta}{2\sqrt{C}} \right)^C \exp(-n_1\delta^2/2),
\end{aligned}$$

其中， D 是一个仅取决于 K的常数。将这两个界代入（6.2），即可完成证明。

□

引理 3.6 的证明

证明。 定义 $\gamma_i = \gamma_{\mathcal{F}_i}(\mathbb{P}^i, \mathbb{Q}^i)$。则：

$$\begin{aligned} Pr\left(\sum_i \gamma_i > \delta\right) &= 1 - Pr\left(\sum_i \gamma_i < \delta\right) \\ &\leq 1 - Pr(\gamma_i < \delta/d \ \forall i) \\ &= Pr(\exists \ i \ni \gamma_i > \delta/d) \\ &\leq \sum_i Pr(\gamma_i > \delta/d) \\ &\leq \sum_i B(\delta/d, D_i, C_i), \end{aligned}$$

其中，第二个不等式使用了并集界。 □

引理 3.7 的证明

证明。 假设对所有 i 都有 $\gamma_{\mathcal{F}_i}(\mu, \nu) = 0$，则

$$\begin{aligned} \gamma_{\mathcal{F}_\pi}(\mu, \nu) &= \sup_{f^\pi \in \mathcal{F}_\pi} \left| \int f^\pi d\mu - \int f^\pi d\nu \right| \\ &= \max_\ell \sup_{f \in \mathcal{F}} \left| \int \pi_\ell \circ f d\mu - \int \pi_\ell \circ f d\nu \right| \\ &= \max_\ell \sup_{f \in \mathcal{F}} \left| \int f_\ell d\mu - \int f_\ell d\nu \right| \\ &= \max_\ell \sup_{f_\ell \in \mathcal{F}_\ell} \left| \int f_\ell d\mu - \int f_\ell d\nu \right| \\ &= \max_\ell \gamma_{\mathcal{F}_\ell}(\mu, \nu) = 0. \end{aligned}$$

反之，假设 $\gamma_{\mathcal{F}_\pi}(\mu, \nu) = 0$，则有

$$\begin{aligned} \gamma_{\mathcal{F}_i}(\mu, \nu) &= \sup_{f_\ell \in \mathcal{F}_\ell} \left| \int f_\ell d\mu - \int f_\ell d\nu \right| \\ &= \sup_{f \in \mathcal{F}} \left| \int \pi_\ell \circ f d\mu - \int \pi_\ell \circ f d\nu \right| \\ &\leq \max_\ell \sup_{f \in \mathcal{F}} \left| \int \pi_\ell \circ f d\mu - \int \pi_\ell \circ f d\nu \right| \\ &= \gamma_{\mathcal{F}_\pi}(\mu, \nu) = 0. \end{aligned}$$

这证明了前两个等价关系。第三个是证明过程中的一个副产品。 □

推论 3.8 的证明

证明。为避免繁琐的记号，令 $v = \frac{1}{n_0}\sum_{j=1}^{n_0} f^*(X_j^0) - \frac{1}{n_1}\sum_{j=1}^{n_1} f^*(X_j^1)$，并注意到 $v_\ell = \frac{1}{n_0}\sum_{j=1}^{n_0} f_\ell^*(X_j^0) - \frac{1}{n_1}\sum_{j=1}^{n_1} f_\ell^*(X_j^1)$，则：

$$\begin{aligned}
Pr\left(\|v\|_{\ell_p} > \delta\right) &= Pr\left(\|v\|_{\ell_p}^p > \delta^p\right) \\
&= Pr\left(\sum_\ell |v_\ell|^p > \delta^p\right) \\
&\le Pr\left(\sum_\ell \gamma_{\mathcal{F}_\ell}(\mathbb{Q}_{n_0}^0, \mathbb{Q}_{n_1}^1)^p > \delta^p\right) \\
&\le \sum_\ell Pr\left(\gamma_{\mathcal{F}_\ell}(\mathbb{Q}_{n_0}^0, \mathbb{Q}_{n_1}^1)^p > \delta^p/d\right) \\
&= \sum_\ell Pr\left(\gamma_{\mathcal{F}_\ell}(\mathbb{Q}_{n_0}^0, \mathbb{Q}_{n_1}^1) > \delta/d^{1/p}\right) \\
&\le \sum_\ell B(\delta/d^{1/p}, D^*, C^*) = dB(\delta/d^{1/p}, D^*, C^*),
\end{aligned}$$

其中第二个和第三个不等式分别由引理 3.6 的一个简单变体以及引理 3.7 的应用得出。对于 ℓ_∞ 情形，我们有：

$$\begin{aligned}
Pr\left(\|v\|_{\ell_\infty} > \delta\right) &\le Pr\left(\max_\ell |\gamma_\ell| > \delta\right) \\
&\le \sum_\ell B(\delta, D^*, C^*),
\end{aligned}$$

至此证明完毕。 □

粗化函数的平衡

我们将证明，粗化精确匹配程序属于一类 Vapnik-Chervonenkis 维数易于处理的函数。考虑由固定数量元素 R 的分区构成的集合 $\mathcal{S}$。对于给定的分区 $S \in \mathcal{S}$，使得 $S = \{s_1, \ldots, s_R\}$，定义 $f_S^{k\alpha}$ 为：

$$f_S^{k\alpha}(x) = \sum_{i=1}^{R} k_i \alpha_i \chi_{s_i}(x),$$

其中，对于常数 k，χ_{s_i} 是 s_i 的示性函数，而 $\alpha := (\alpha_1, \ldots\ldots, \alpha_R)$ 是一个二元向量，也就是说，对于每个 i，$\alpha_i \in \{0, 1\}$。换言之，如果在 s_i 中找到 x，那么当 α_i 为 1 时，f 将返回 x 的缩放版本，否则返回零。

现在令 $\mathcal{F} := \{f_S^{k\alpha}\}_{S\in\mathcal{S},\ \alpha\in A,\ k\le\kappa}$，其中 A 是所有大小为 R 的二元向量的集合，且 $\kappa \in \mathbb{R}$。因此，粗化精确匹配程序属于这一类函数，因为在这种情况下，α_i 表示层 s_i中是否至少有两个来自不同组的成员。对于任意样本点 x，权重通常按以下方式选取：如果 x 是处理组单位，则 $w_i^1 = 1$；否则，$w_i^0 = (m_1^s/m_1)/(m_0^s/m_0)$，其中 s 是 x 所属的层。适当地令 $k_i = w_i^\ell n_\ell/m_\ell$取值，就能为匹配样本赋予权重。我们只需增加一个温和的假设：每个层 s 中的样本量与匹配样本量之比的增长速度不超过 $\sqrt{\kappa}$，也就是说，对所有$s \in S$，都有 $n_\ell/m_\ell^s \le \sqrt{\kappa}$，因为在这种情况下 $w_i^0 \le m_0/m_0^s \le n_0/m_0^s \le \sqrt{\kappa}$ 且 $n_\ell/m_\ell \le \sqrt{\kappa} m_\ell^s/m_\ell \le \sqrt{\kappa}$，因此 $k_i \le \kappa$。最后请注意，任何分区大小更小的类似函数，都可以表示为 $\mathcal{F}$中的一个函数，因此，只要不超过合理的上限 R，我们就可以考虑可变的分区大小。

对于任意一个大小为 R 的点集，都存在一个分区 S，其中不同元素各包含一个点，因此存在一个 α，可以任意地将每个点分配为 0 或 1。所以 $\mathcal{F}$ 能打散这样的点集。然而，如果再增加一个点，由于分区数量受到限制，该点必然与先前的某个点共享一个分区元素，因此在 $f_s^{k\alpha}$下无法进行这样的分配。所以，$\mathcal{F}$ 的 Vapnik- Chervonenkis 维数为 R。最后，令 $g(\mathcal{Z}_\ell) = \mathbb{Q}_{n_\ell}^\ell$，其中 $\mathbb{Q}_{n_\ell}^\ell$ 是组 ℓ的样本 $\mathcal{Z}_\ell$ 的经验分布。令 k^* 按上述方式选取，并令 (S^*, α^*) 为粗化精确匹配所用的特定分区和二元向量。那么，对于 ℓ^{th} 分量，我们得到：

$$\begin{aligned}\left|\frac{1}{m_0}\sum_{i\in M_0} w_i^0 Z_{i,\ell}^0 - \frac{1}{m_1}\sum_{j\in M_1} w_j^1 Z_{j,\ell}^1\right| &= \left|\frac{1}{n_0}\sum_{i=1}^{n_0} f_{S^*,\ell}^{k^*\alpha^*}(Z_i^0) - \frac{1}{n_1}\sum_{j=1}^{n_1} f_{S^*,\ell}^{k^*\alpha^*}(Z_j^1)\right| \\ &\le \sup_{f_\ell\in\mathcal{F}^*}\left|\frac{1}{n_0}\sum_{i=1}^{n_0} f_\ell(Z_i^0) - \frac{1}{n_1}\sum_{j=1}^{n_1} f_\ell(Z_j^1)\right| \\ &= \gamma_{\mathcal{F}^*}(\mathbb{Q}_{n_0}^0, \mathbb{Q}_{n_1}^1) = \gamma_{\mathcal{F}^*}(g(\mathcal{Z}_0), g(\mathcal{Z}_1)).\end{aligned}$$

因此，每个维度上匹配样本之间的差异都受未匹配样本的 $\gamma_{\mathcal{F}^*}$ 距离所限。最后，只要假设域是紧致域（对于大多数粗化精确匹配情形，这样的假设是合理的），函数 $h(x) := \kappa x$ 就是 $\mathcal{F}$ 的包络函数，且其范数为 $\|h\|_{L_2(\mu)} < \infty$。于是，根据 Van Der Vaart 和 Wellner（1996）的定理 2.6.7：

$$\sup_\mu N(\epsilon, \mathcal{F}, L_2(\mu)) \le \left(\frac{K}{\epsilon}\right)^{C^*},$$

其中某个常数为 K，且 $C^* = 2(R-1)$。

由此得到最终结果：假设总体概率对于$\gamma_{\mathcal{F}_\pi}$满足理想平衡，那么对于 ℓ^{th} 分量，我们有：

$$Pr\left(\left|\frac{1}{m_0}\sum_{i\in M_0} w_i^0 Z_{i,\ell}^0 - \frac{1}{m_1}\sum_{j\in M_1} w_j^1 Z_{j,\ell}^1\right| > \delta\right) \le B(\delta, D, C^*).$$

如果我们关注的是完整向量的 ℓ_p 范数，那么根据推论 3.8：

$$Pr\left\{\left\|\frac{1}{m_0}\sum_{i\in M_0} w_i^0 Z_i^0 - \frac{1}{m_1}\sum_{j\in M_1} w_j^1 Z_j^1\right\|_{\ell_p} > \delta\right\} \le dB(\delta/d^{1/p}, D, C^*),$$

对于有限的 $p \ge 1$。而

$$Pr\left\{\left\|\frac{1}{m_0}\sum_{i\in M_0} w_i^0 Z_i^0 - \frac{1}{m_1}\sum_{j\in M_1} w_j^1 Z_j^1\right\|_{\ell_\infty} > \delta\right\} \le dB(\delta, D, C^*).$$

使用倾向评分进行平衡

回忆一下 $e(Z) = P(T = 1 \mid Z)$，并且我们假设 $Z \mid T = \ell \sim N(\mu_\ell, \Sigma)$。设 p_ℓ为 $N(\mu_\ell, \Sigma)$ 的概率密度函数，即高斯密度，那么根据贝叶斯定理的密度形式，我们有

$$p(T = 1 \mid Z = z) = \frac{p_1 P(T = 1)}{p_1 P(T = 1) + p_0 P(T = 0)}.$$

因此，我们可以将 $e(Z)$ 的对数几率 表示为

$$logit(e(Z)) = \log\left(\frac{e(Z)}{1-e(Z)}\right) = \log\left(\frac{p_1 P(T=1)}{p_0 P(T=0)}\right).$$

现在定义 $L_k := logit(e(Z_k))$，那么匹配程序基于差值 $|L_i - L_j|$。根据上述计算，经过几个简单步骤，我们得到

$$\begin{aligned} |L_i - L_j| &= \left|(\mu_1 - \mu_0)^T \Sigma^{-1}(Z_i - Z_j)\right| \\ &= \left|f^*(Z_i) - f^*(Z_j)\right|, \end{aligned}$$

其中，对于 $w \in \mathbb{R}^p$， $f^*(x) = w^T x$。注意，向量 w与线性判别分析中使用的向量相同，因此，加入偏移参数后，将 f^* 视为一个超平面会很有帮助。

设 M_0^j 为分配给处理单元 j的对照单元。我们假设每个处理单元分配的对照组数量固定，因此 $m_0 = |M_0^j| m_1$。那么

$$\begin{aligned} \Delta &:= \left|\frac{1}{m_1}\sum_{j\in M_1} logit(e_j) - \frac{1}{m_0}\sum_{i\in M_0} logit(e_i)\right| \\ &= \left|\frac{1}{m_1}\sum_{j\in M_1} L_j - \sum_{j\in M_1}\frac{1}{m_0}\sum_{i\in M_0^j} L_i\right| \\ &= \left|\sum_{j\in M_1}\left(\frac{1}{m_1}L_j - \frac{1}{m_0}\sum_{i\in M_0^j} L_i\right)\right| \\ &= \left|\sum_{j\in M_1}\left(\frac{1}{m_1}\sum_{i\in M_0^j}\frac{L_j}{|M_0^j|} - \frac{1}{m_0}\sum_{i\in M_0^j} L_i\right)\right| \\ &= \left|\sum_{j\in M_1}\sum_{i\in M_0^j}\left(\frac{L_j}{m_1|M_0^j|} - \frac{L_i}{m_0}\right)\right| \\ &= \left|\sum_{j\in M_1}\sum_{i\in M_0^j}\frac{1}{m_0}(L_j - L_i)\right| \\ &= \left|\sum_{j\in M_1}\sum_{i\in M_0^j}\frac{1}{m_0}(f^*(Z_j) - f^*(Z_i))\right| \\ &= \left|\frac{1}{m_1}\sum_{j\in M_1} f^*(Z_j) - \frac{1}{m_0}\sum_{i\in M_0} f^*(Z_i)\right|. \end{aligned}$$

也就是说，我们可以用判别函数的均值差来表示 $logit$s 的均值差。设 p 为协变量的维数，设 $\mathcal{F}$ 为由p维超平面组成的集合，注意 $f^* \in \mathcal{F}$。已知 $\mathcal{F}$ 的 Vapnik-Chervonenkis 维数为 $p+1$（Mohri 等人，2018）。我们希望用 γ 来界定 Δ，但首先需要对 f^*进行一些调整。

匹配程序确定了一组 $\mathcal{Z}_M = \{Z_k \mid k \in M\}$ 匹配样本，其中$M = M_0 \cup M_1$。根据高斯假设，Z服从高斯混合分布，因此两个样本点相同的概率为零。因此，存在一个 $\epsilon > 0$，使得对于所有

$k \in M$， $\mathcal{Z} \cap B_\epsilon(Z_k) = \{Z_k\}$，也就是说，以匹配样本为中心的每个 ϵ 球都不包含任何其他样本点（此处 $\mathcal{Z}$ 为样本集）。令 $S_\epsilon = \cup_k B_\epsilon(Z_k)$。注意， S_ϵ 是一个可测集。令 $\beta_{S_\epsilon}(x) := x\chi^{S_\epsilon}(x)$，该函数将未匹配的点映射为零，将匹配的点映射为其自身。此外，令 $\beta_\ell(x) := \frac{m_\ell}{n_\ell}\chi^{M_\ell}(x) + \chi^{M_\ell^C}(x)$，其中 $\ell \in \{0,1\}$。每个 β_ℓ 都将 M_ℓ 中的元素按因子 $\frac{m_\ell}{n_\ell}$ 缩放，其余元素保持不变。

注意， $f_M^* := f^* \circ \beta_1 \circ \beta_0 \circ \beta^{S_\epsilon}$ 在 $k \in M_k$ 时将 Z_k 映射到 $\frac{m_\ell}{n_\ell} w^T Z_k$，否则映射到 0。于是我们可以将 Δ 表示为

$$\begin{aligned}\Delta &= \left| \frac{1}{m_1} \sum_{j \in M_1} f^*(Z_j) - \frac{1}{m_0} \sum_{i \in M_0} f^*(Z_i) \right| \\ &= \left| \frac{1}{n_1} \sum_{j=1}^{n_1} f_M^*(Z_j) - \frac{1}{n_0} \sum_{i=1}^{n_0} f_M^*(Z_i) \right|.\end{aligned}$$

现在，考虑集合 $\mathcal{F}_M := \{f \circ \beta_1 \circ \beta_0 \circ \beta_S | f \in \mathcal{F}, S \in \Sigma\}$，其中 Σ 是根据 Z 的分布定义的可测集集合。 $\mathcal{F}_M$ 的 Vapnik-Chervonenkis 维数与 $\mathcal{F}$ 的相同，即 $p+1$。为此，我们注意到，超平面情形的标准推导涉及在 $\mathbb{R}^p$ 中打散标准基 $\mathcal{B}$。以概率 1，不会有样本点等于标准基向量，因此存在一个 $\epsilon' > 0$，使得我们可以构造集合 $s = \cup_{x \in \mathcal{B}} B_{\epsilon'}(x)$，满足 $s \in \Sigma$，且没有样本点属于 s。考虑用于打散 $\mathcal{B}$ 的 $\mathcal{F}$ 中的函数 $\{f_\nu\}$，并利用 s，我们也可以用 $\mathcal{F}_M$ 中的函数 $\{f_\nu \circ \beta_1 \circ \beta_0 \circ \beta_s\}$ 打散 $\mathcal{B}$。因此，Vapnik-Chervonenkis 维数至少为 $p+1$。由于函数 β_1、 β_0 和 β^S 要么为零，要么为缩放后的恒等映射，因此不会增加复杂度，维数不超过 $p+1$，所以确实为 $p+1$。对于包络函数，我们可以选择 $h(x) =< w_e, x >$。 w_e 的范数必须足够大，才能保持 $p+1$ 的 Vapnik-Chervonenkis 维数。由于用于确保这一维数的向量的范数为 $p+1$， w_e 的范数至少必须为 $p+1$。因此，我们可以选择任意大的常数 $C > p+1$。由于我们关注的是形如 $w = \Sigma^{-1}\Delta\mu$ 的向量，因此有 $\|w\| \leq \|S^{-1}\|_F \|\Delta\mu\|_2$，用户必须选择能够界定这些范数的常数。此外，我们还必须假设协变量本身有界，这能确保 h 的范数有限。

最后，我们有

$$\begin{aligned}\Delta &= \left| \frac{1}{n_1} \sum_{j=1}^{n_1} f_M^*(Z_j) - \frac{1}{n_0} \sum_{i=1}^{n_0} f_M^*(Z_i) \right| \\ &\leq \sup_{f \in \mathcal{F}_M} \left| \frac{1}{n_1} \sum_{j=1}^{n_1} f(Z_j) - \frac{1}{n_0} \sum_{i=0}^{n_0} f(Z_i) \right| \\ &= \gamma_{\mathcal{F}_M}(\mathbb{Q}_{n_0}^0, \mathbb{Q}_{n_1}^1).\end{aligned}$$

假设总体概率满足理想平衡，并结合定理 3.3 应用 Van Der Vaart 和 Wellner（1996）的定理 2.6.7，可得

$$Pr\{\Delta > \delta\} \leq B(\delta, D, 2p).$$

再生核希尔伯特空间的覆盖数上界

关于再生核希尔伯特空间的精彩综述，请参阅 Wahba (1990)；Berlinet 和 Thomas-Agnan (2011)；Steinwart 和 Christmann (2008)。粗略地说，映射 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 若满足以下性质，就称为与再生核希尔伯特空间 $\mathcal{H}$ 相关联的再生核：(a) $k(\cdot, x) \in \mathcal{H}$ 对于任意 $x \in \mathcal{X}$；(b) $f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$ 对于所有 $f \in \mathcal{H}$ 和 $x \in \mathcal{X}$。性质 (b) 通常称为再生性质。

要将定理 3.3 应用于再生核情形，我们需要基于不同于 Vapnik-Chervonenkis 理论的论证，直接对覆盖数进行界定。定义空间

$$\mathcal{H}_q^m(\mathbb{R}^p) = \{f \in L_q(\mathbb{R}^p) \mid D^j f \in L_q(\mathbb{R}^p) \ \forall j \in \{1, \ldots, m\};\ \|f\|_q < \infty\},$$

其中

$$\|f\|_q = \sum_{0 \le |\alpha| \le s} \|D^\alpha f\|_{L_q}$$

且 D^α 表示分布意义下的偏导数。于是，根据 Nickl 和 P¨otscher（2007）的定理 1，若 $m - q/p > 0$，则

$$N(\epsilon, \mathcal{H}, \|\cdot\|_q) \le b_1 \epsilon^{-q},$$

而若 $m - q/p < 0$,

$$N(\epsilon, \mathcal{H}, \|\cdot\|_q) \le b_2 \epsilon^{-p/m},$$

基于这一结果，便可应用定理 3.3 证明理想平衡下的收敛速率。请注意，这并不涵盖高斯核的情形，因为高斯核是无穷次可微的，因此空间 $\mathcal{H}_q^m(\mathbb{R}^p)$ 并不适用。对于关注高斯情形的读者，我们推荐阅读 Steinwart 和 Fischer（2020）近期发表的论文。

参考文献

Abadie, A. 和 G. W. Imbens（2006）。用于估计平均处理效应的匹配估计量的大样本性质。《计量经济学》*74*(1), 235–267。Abadie, A. 和 G. W. Imbens（2011）。用于估计平均处理效应的偏差校正匹配估计量。《商业与经济统计杂志》*29*(1), 1–11。Abadie, A. 和 G. W. Imbens（2016）。基于估计倾向得分进行匹配。《计量经济学》*84*(2), 781–807。Baudat, G. 和 F. Anouar（2000）。使用核方法的广义判别分析。《神经计算》*12*(10), 2385–2404。Berlinet, A. 和 C. Thomas-Agnan（2011）。概率论与统计学中的再生核希尔伯特空间。施普林格科学与商业媒体。Chan, K. C. G.、S. C. P. Yam 和 Z. Zhang（2016）。通过经验平衡校准加权对平均处理效应进行全局有效的非参数推断。《英国皇家统计学会会刊：*B*辑（统计方法学）》*78*(3), 673–700。Chervonenkis, A. 和 V. Vapnik（1971）。事件发生频率一致收敛至其概率（独立试验序列中事件频率一致收敛至其发生概率）。《概率论及其应用》*16*, 264–279。Hainmueller, J.（2012）。因果效应的熵平衡：一种用于在观察性研究中生成平衡样本的多变量重新加权方法。《政治分析》*20*(1), 25–46。Hansen, B. B.（2008）。倾向得分的预后类似物。*Biometrika 95*(2), 481–488。Hazlett, C.（2016）。核平衡：一种用于估计因果效应的灵活非参数加权程序。Ho, D. E.、K. Imai、G. King 和 E. A. Stuart（2007）。将匹配作为非参数预处理，以降低参数因果推断中的模型依赖性。《政治分析》*15*(3), 199–236。

Holland, P. W.（1986）。统计学与因果推断。《美国统计学会杂志》*81*(396), 945–960。Iacus, S. M.、G. King 和 G. Porro（2011）。具有单调不平衡界定性质的多变量匹配方法。《美国统计学会杂志》*106*(493), 345–361。Imai, K. 和 M. Ratkovic（2014）。协变量平衡倾向得分。《英国皇家统计学会会刊：*B*辑（统计方法学）》*76*(1), 243–263。Imbens, G. W. 和 D. B. Rubin（2015）。统计学、社会科学与生物医学科学中的因果推断。剑桥大学出版社。Kallus, N.（2020）。用于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。Kosorok, M. R.（2007）。经验过程与半参数推断导论。施普林格科学与商业媒体。Mohri, M.、A. Rostamizadeh 和 A. Talwalkar（2018）。机器学习基础。麻省理工学院出版社。Neyman, J.（1923）。概率论在农业试验中的应用：原则探讨。*Roczniki Nauk Rolniczych 10*, 1–51。Nickl, R. 和 B. M. P¨otscher（2007）。Besov 型和 Sobolev 型函数类的括号度量熵率与经验中心极限定理。理论概率杂志 *20*(2), 177–199。Rosenbaum, P. R. 和 D. B. Rubin（1983）。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。Rubin, D. B.（1974）。估计随机研究与非随机研究中的处理效应。教育心理学杂志 *66*(5), 688。Rubin, D. B.（1976）。具有等百分比偏差缩减性质的多变量匹配方法，I：若干示例。*Biometrics*, 109–120。Rubin, D. B.、E. A. Stuart 等（2006）。基于比例椭球对称分布的判别混合分布的仿射不变匹配方法。统计年刊 *34*(4), 1814–1826。Rubin, D. B. 和 N. Thomas（1992）。基于椭球分布的仿射不变匹配方法。统计年刊, 1079–1093。Salimi, B. 和 D. Suciu（2016）。Zaliql：一种基于SQL的框架，用于从大数据中进行因果推断。*arXiv*预印本 *arXiv:1609.03540*。Sekhon, J. S.（2008）。结合自动平衡优化的多变量匹配与倾向得分匹配软件：R匹配软件包。统计软件杂志，即将发表。Steinwart, I. 和 A. Christmann（2008）。支持向量机。施普林格科学与商业媒体。Steinwart, I. 和 S. Fischer（2020）。深入考察高斯核的覆盖数界。复杂性杂志, 101513。Stuart, E. A.（2010）。因果推断的匹配方法：综述与展望。统计科学：数理统计学会综述期刊 *25*(1), 1。

Van Der Vaart, A. W. 和 J. A. Wellner（1996）。弱收敛。载于弱收敛与经验过程，第 16–28 页。施普林格。Wahba, G.（1990）。观测数据的样条模型。工业与应用数学学会。Wang, T., M. Morucci, M. U. Awan, Y. Liu, S. Roy, C. Rudin 和 A. Volfovsky（2017）。Flame：一种用于因果推断的快速、大规模、近似匹配且精确的方法。Wang, Y. 和 J. R. Zubizarreta（2019）。最小离散度近似平衡权重：渐近性质与实践考量。*Biometrika*。Wong, R. K. 和 K. C. G. Chan（2018）。用于观察性研究的基于核的协变量函数平衡。*Biometrika 105*（1），199–213。Zhu, Y., J. S. Savage 和 D. Ghosh（2018）。一种用于平衡评估的基于核的度量。《因果推断杂志》*6*（2）。Zolotarev, V. M.（1984）。概率度量。《概率论及其应用》*28*（2），278–302。Zubizarreta, J. R.（2015）。用于不完整结局数据估计的稳定协变量平衡权重。《美国统计学会杂志》*110*（511），910–922。