经济学中的因果推断问题

最近十多年，经济学实证研究的发展越来越注重变量之间因果关系的识别，如何通过清晰的实证设计识别变量间的因果关系是所有经济学实证研究者所要面对的头等问题（Angrist 和 Pischke，2010）。自然科学通过受控实验来识别因果关系，其基本思想是通过随机选择实验组和控制组，且在保持其他条件不变的情况下给予实验组某一处置，然后比较两组对象在结果上的差异。若结果有显著差异，那么这种差异唯一的来源就是实验所施加的处置。作为对自然科学实验方法的回应，经济学中的实验方法随Daniel Kahneman 和Vornon Lomax Smith 获得诺贝尔经济学奖而得到主流经济学界的认同。但是在经济学实证研究中通过实验的方法推断因果关系也存在诸多缺陷，比如能通过实验研究的经济关系有限、实验的道德考量、等待结果的时间漫长以及成本因素等。于是，自然实验 (Natural Experiment) 在实证研究中逐渐受到推崇 (Meyer, 1995; Angrist 和Krueger, 2001)，通过利用某一特殊事件造成的对所要研究解释变量的外生冲击来识别其对被解释变量的因果影响。但是在大多数实证研究还是要依靠观测数据的情况下，如果在观测数据中找不出对所要研究解释变量的外生冲击，此时，因果推断就成了一个难题。如何利用非随机的观测数据进行统计推断，以达到和利用随机实验数据相近的效果是统计学家和经济学家共同面对的问题，Rubin (1974) 首次对此在理论上给出了正式的阐述。

对观测数据进行因果推断最大的挑战在于“因果推断的基础性问题”，以下通过Rubin (1974) 的分析框架加以说明。假设研究者关注处置w 对结果 y 是否有因果影响，若某一个体 i 接受了处置， 则记$w_i = 1$; 若没有接受处置，则记 $w_i = 0$。在观测数据中，研究者只知道个体i 接受了处置的结果$y_i(1)$或没有接受处置的结果$y_i(0)$，而不能同时观察到个体 i 在两种状态下的情况——即同时观察到$y_i(1)$ 和$y_i(0)$。但因果推断所要估计的处置效应（记为τ_i）为同一个体在有和没有受到处置两种状态下结果的差异，即$\tau_i = y_i(1) - y_i(0)$，所以用观测数据推断因果关系就存在缺失数据问题——缺失观测单位处于反事实处置状态下的数据，这就是所谓的 “因果推断的基础性问题”。匹配法的产生就是为了解决基于观测数据推断因果关系所带来的缺失数据问题。

匹配的基本思想

假设从总体中获取的样本为(y_i, w_i, x_i)，其中i = 1, …, N; y_i为结果变量; w_i为处置指示变量; x_i为协变量。把样本个体分为实验组和受控组，分别代表接受处置的情形 ($w_i = 1$) 和未接受处置的情形 ($w_i = 0$)。对于样本个体i，其结果为$y_i(0)$或 $y_i(1)$。因此，样本中有N_i个体来自实验组，有$N - N_i$个个体来自控制组。在关于处置效应的理论和应用文献中，最广泛受到关注的是两个估计量:

①平均处置效应 (ATE)，定义为$\mathrm{T}_{\text{ate}} = E\left[y_i(1) - y_i(0) \mid x_i\right]$;

②受处置平均效应 (ATT)，定义为$\mathrm{T}_{\text{att}} = E\left[y_i(1) \mid w_i = 1\right] - E\left[y_i(0) \mid w_i = 1\right]$。

不同于实验产生的带有随机性质的数据，在使用观测数据的情况下必须施加一定的假设条件才能得到满足因果联系的估计值。最重要的两个假设条件是:

假设**1**: 不可知性(Ignorability): 以x为条件，w 与(y_0, y_1)是独立的，$w \perp (\mathrm{y}_0, \mathrm{y}_l) \mid x_0$。

假设**2**: 重叠(Overlap) : 对于所有x∈X，其中X是协变量的集合，$0 < P(w = 1|x) < 1$。

这两个假设被称为强不可知性条件。由假设1可推出:

①$E\left[y_i(1) \mid x_i, w_i\right] = E(y_i(1)|x_i)$;

②$E\left[y_i(0) \mid x_i, w_i\right] = E(y_i(0)|x_i)$。

下面，以平均处置效应τ_{ate}为例说明匹配的主要思想。注意到，结果变量 y_i 可表示为 $y_i = y_i(0) + w_i(y_i(1) - y_i(0))$，所以若 w=1, $E(y_i|x_i, w_i) = E(y_i(1)|x_i)$; 若 w=0, 则 $E(y_i|x_i, w_i) = E(y_i(0)|x_i)$. 所以平 均处置效应估计量 ({ }^{T}) ate 又可表示为:

$$\mathrm{T}_{\text{ate}} = \mathrm{E}\left[y_i(1) - y_i(0) \mid x_i\right] = E\left[y_i \mid x_i, w_i = 1\right] - E\left[y_i \mid x_i, w_i = 0\right] \tag{1}$$

由于$E[y_i|x_i, w_i = 1]$和$E[y_i|x_i, w_i = 0]$可以分别由试验组样本和控制组样本估计得到，所以平均处置效应为两组样本估计值的差值。记$y_i\hat{(0)}$为$E[y_i|x_i, w_i = 0]$的样本估计值，$y_i\hat{(0)}$为$E[y_i|x_i, w_i =]$的样本估计值，所以平均处置效应的样本估计值表示为$\hat{\tau} = N^{-1}\sum(y_i(1) - \hat{y_i(0)}$。由于存在数据缺失的问题——观测数据中只有个体在某一种状态下的结果，匹配法的基本思想就是通过在相反状态的样本中寻找和自身最“接近”的样本来近似替代自身反事实状态下的结果。例如，若某一接受了处置的个体其结果为$y_i(1)$，我们通过在未受处置的样本中寻找和 i 最“接近”的个体 $I(i)$ （记$I(i)$ 为和i “最接近”的受控组样本中个体的索引），用其结果来近似替代i 在未受处置状态下的结果; 同理，对于受控组中的样本个体在实验组的样本中找到“最接近”的样本来进行匹配。最终，匹配出的样本为

$$\hat{y}_i(0) = \begin{cases} y_i & w_i = 0 \\ y_{l(i)} & w_i = 1 \end{cases} \quad \hat{y}_i(1) = \begin{cases} y_{l(i)} w_i = 0 \\ y_i w_i = 1 \end{cases} \tag{2}$$

利用这些匹配出的样本数据，就可以计算平均处置效应。

匹配法的发展过程

应用匹配法的关键在于如何寻找反事实状态下的结果，以构建协变量平衡的匹配样本。协变量平衡指的是实验组样本协变量的共同分布和控制组样本协变量的共同分布相同或相似。按照匹配方式的不同，可以将匹配法分为协变量匹配和倾向得分匹配两类。从最初的协变量匹配到用单一维度的倾向得分进行匹配，关于用何种方式进行匹配经历了若干的发展过程。协变量匹配的想法较为直观，在最初的发展阶段得到了应用，但在数据量较少或协变量 较多的情况下，就存在所谓“维度诅咒”的问题。倾向得分匹配利用倾向得分这一包含了协变量所有信息的单一统计量进行匹配，克服了“维度的诅咒”而且满足平衡性要求。

协变量匹配

匹配法的应用最早出现在20 世纪40 年代 (Greenwood, 1945; Chapin, 1947)，但其在理论上的发展却始于20 世纪70 年代。其中，Rubin (1973a, 1973b) 以及Cochran 和Rubin (1973) 对此做出了开创性的贡献。他们关注了单个协变量的情形下受处置平均效应的估计问题。早期的匹配法通过单个关键协变量或通过对多个协变量加权来进行匹配 (Dehejia 和 Wahba, 2002) ，但这种协变量匹配在实践应用中却存在所谓的“维度诅咒”问题。如在 Chapin(1947) 的研究中，实验组样本数为671，控制组样本数为523，当通过6 个分类变量进行匹配时，最终只有23 对样本能匹配上。小样本情况下协变量匹配造成的样本数骤减无疑给统计推断的准确性带来了极大的不利影响; 同时，在大样本的情形下协变量匹配则存在计算上的困难。下面具体说明协变量匹配。

假设$f(w|X_k)$为协变量X_k的共同分布，其中 w = 0，1 表示是否受到处置; k = i, j 分别表示实验组样本和控制组样本。

$$\begin{aligned} &X_i = X_j \Rightarrow f_w(X_i) = f_w(X_j), w = 0,1 \\ &d(X_i X_j) < \varepsilon \Rightarrow d(f_w(X_i), f_w(X_j)) < \delta, w = 0,1 \end{aligned} \tag{3}$$

其中，和为数学意义上的距离。常见的距离定义为马氏距离① (Cochran和Rubin, 1973; Rosenbaum和Rubin, 1985) ，表示为$d_m = (X_i - X_j)D^{-l}(X_i - X_j)'$，其中D为协变量x 的方差-协方差矩阵。式 (3) 为精准匹配的情形: 实验组样本i 和与其相匹配的控制组样本j具有相同的协变量值，所以经过匹配后的两组样本其协变量的共同分布相同，满足了协变量平衡的条件。式 (4) 为更常见的情形: 实验组样本i 和与其相匹配的控制组样本j， 它们的距离小于某一设定的值ε，因此经过匹配后的两组样本其协变量的共同分布的距离小于某一设定的值δ，部分满足了协变量平衡的条件。

倾向得分匹配

由于协变量匹配在数据量较小以及协变量较多 的时候会出现难以匹配的情形，所以其在实际应用中受到诸多限制。Rosenbaum 和 Rubin (1983) 开创性地提出了倾向得分匹配的方法，通过倾向得分这一单一维度变量进行匹配以减轻协变量匹配对数据以及计算上的要求。定义倾向得分（记为e(x)）为在给定可观测协变量的情况下个体接受处置的条件概率 （表示为e(x)=prob(w=1 | x)）。他们证明了在满足不可知性条件（假设1）的情况下，以倾向得分$e(x)$为条件，w 与 (y_0, y_1) 是独立的，即$w \perp (y_0, y_1) \mid e(x)$。对通过倾向得分进行匹配构建的平衡样本进行均值差分能得到无偏的平均处置效应估计值。倾向得分匹配的基本想法可以表述为

$$\begin{aligned} &\text{prob}(X_i \mid T_i = 1, p(X_i) = p) = \text{prob}((X_i \mid T_i = 0, p(X_i) = p) = \text{prob}(X_i \mid p) \\ &d(p_k, p_l) < \varepsilon \Rightarrow d(\text{prob}(X_i \mid p_k), \text{prob}(X_i \mid p_j))s < \delta \end{aligned} \tag{4}$$

式 (5) 和式 (6) 分别对应协变量匹配中的式 (3) 和式 (4) 。式 (5) 表示当根据相同的倾向分值进行匹配时，实验组样本和控制组样本协变量的共同分布在倾向分值处相同。式 (5) 表示当根据倾向得分的差距在之内进行匹配时，实验组样本和控制组样本协变量的共同分布的距离小于δ。

匹配方法的基本步骤

匹配方法有以下四个主要步骤：

（1）定义“相似性(closeness)”：用于度量一个个体是否能很好地匹配另一个个体好的距离。

（2）给定“相似性”后，实施匹配方法。

（3）评估生成的匹配样本的匹配结果，步骤1和2可能需要重复多次，直到得到良好的匹配结果。

（4）给定步骤3中已完成的匹配，分析结果以及估计因果效应。

定义相似性

定义相似性包含两个层面：选择协变量以及将这些协变量综合成一个距离测度。

选择协变量

匹配方法依赖于条件独立性假设(ignorability)，它假设在控制了协变量之后，控制组和对照组之间没有未观察到的差异。引入实际上与结果无关的变量可能会导致方差略有增加。然而，就增加的偏误而言，遗漏潜在的重要混淆变量的代价可能非常高。因此，研究人员在引入可能与处理分配和/或结果相关的变量方面不应该过于严格。需要注意的是，可能已受相关处理影响的任何变量都不应包含在匹配过程中。

定义距离测度

下一步是定义“距离“D_{ij},具体来说就是两个个体(i和j)之间相似性的度量。有以下四种比较主流的方法：

(1)Exact:

$$D_{ij} = \begin{cases} 0, if X_i = X_j \\ \infty, if X_i \neq X_j \end{cases} \tag{5}$$

(2)Mahalanobis:

$$D_{ij} = (X_i - X_j)' \Sigma^{-1} (X_i - X_j) \tag{6}$$

其中，$Sigma$是协变量X的方差协方差矩阵。如果估计的因果效应参数是ATT，$Sigma$是控制组协变量的方差协方差矩阵；如果估计的是ATE，$Sigma$是所有协变量的方差协方差矩阵。

(3)Propensity score:

$$D_{ij} = |e_i - e_j| \tag{7}$$

其中e_k是个体k的倾向得分。

(4)Linear prospensity score:

$$D_{ij} = |logite_i - logite_j| \tag{8}$$

Exact distance和Mahalanobis distance测度的主要困难在于，当*X*是高维时，两者都不能很好地工作。要求精确匹配通常会导致大量体无法匹配，这会导致比匹配不精确但仍有更多个体留在分析中的情况更大的偏差(Rosenbaum and Rubin, 1985)。而且Mahalanobis distance在协变量不是正态分布时也表现不好(Gu and Rosenbaum, 1993)，这可能是因为Mahalanobis度量匹配本质上将*X*元素之间的所有交互关系视为同等重要。倾向得分将所有协变量汇总为一个标量：被处理的概率。首先倾向得分是一个平衡分数：在倾向分数的每个值处，定义倾向分数的协变量*X*的分布在治疗组和对照组中是相同的。 因此，至少在观察到的协变量方面，将具有相似倾向得分的个体分组重复了一个小型随机实验。 其次，如果给定协变量可以忽略治疗分配，那么给定倾向评分也可以忽略治疗分配。上述距离度量也可以组合，例如，对关键协变量进行精确匹配，然后在这些组内进行倾向得分匹配。如果感兴趣的关键协变量是连续的， Mahalanobis matching within propensity score calipers (Rubin and Thomas, 2000)将个体 *i* 和 *j* 之间的距离定义为

$$D_{ij} = \begin{cases} (Z_i - Z_j)' \Sigma^{-1} (Z_i - Z_j), \text{ if } |\text{logit}(e_i) - \text{logit}(e_j)| \leq c \\ \infty, \text{ if } |\text{logit}(e_i) - \text{logit}(e_j)| > c \end{cases} \tag{9}$$

其中c是卡尺，Z是一系列关键协变量，Σ是Z的方差协方差。

匹配实施方法

近邻匹配(Nearest Neighbor Matching)

在其最简单的形式里，1:1近邻匹配为每个处理组个体*i* 在控制组中寻找一个距离最近的控制组个体与之匹配。关于 1:1 匹配的一个常见问题是它会丢弃大量观测值，导致解释力的降低。但这种解释力的降低往往非常小，首先，在均值的两个样本组的比较中，精度主要由样本大小较小的组决定（Cohen, 1988)。所以处理组的大小保持不变的情况下，只有对照组样本大小的减少，那么整体解释里可能不会降低很多(Ho et al., 2007)。其次，当两个样本组更相似时，解释力会增加，因为此时外推减少了且精度变得更高。

在一对一最近邻匹配中，往往采用贪婪匹配(greedy matching)，对每一个干预组个体都在控制组中寻找一个距离最近的。但是保证每一对距离最近，对全部干预组个体而言，匹配上的控制组样本并不一定是总体上最近的。另一种匹配方法，称为最优匹配(optimal matching)，不是一个一个个体地进行匹配，而是总体上对所有处理组个体同时进行匹配，寻找对所有处理组个体而言匹配上的总距离最小。如果我们想要得到每个个体的匹配效果，最优匹配方法会得到更平衡的成对匹配。当控制组个体很多的时候，有时可以为每个处理组的给找到多个良好匹配，称为"ratio matching"(Smith, 1997; Rubin and Thomas, 2000)。这里选择匹配数量涉及偏差：方差权衡。为每个处理过的个体选择多个对照通常会增加偏差，因为根据定义，第 2、第 3 和第 4 最接近的匹配比第 1 最接近的匹配离被处理个体更远。另一方面，由于更大的匹配样本大小，使用多个匹配可以减少方差。一种对"ratio matching"进行改进的形式叫做“variable ratio matching"，其允许比率变化，也就是不同的处理个体可以接受不同数量的匹配。另一个关键问题是一个控制组个体是否可以用作多个干预组个体的匹配：匹配应该“with replacement”还是“without replacement”。Matching with replacement通常可以减少偏差，因为可以多次使用与许多干预个体相似的控制个体。然而，当使用matcing with replacement时，推断变得更加复杂，因为被匹配的控制个体不再独立——有些在匹配的样本中不止一次，这需要在结果分析中加以考虑。

Subclassification，Full Matching and Weighting

与最近邻匹配个体被赋予 0 或 1 的权重（取决于他们是否被选中） 作为匹配相反，这些方法可以被认为是赋予所有个体（隐式或显式）介于 0 和 1 之间的权重。

Subclassification

分层匹配是根据协变量或倾向指数进行分层(stratification)，使层内两组个体特征比较相似，从而降低估计偏差。

Full Matching

全匹配是一种更复杂的subclassification形式，这种方法可以自动选择子类的数量(Rosenbaum, 1991; Hansen, 2004; Stuart and Green, 2008)。Full matching构建了一系列匹配集，其中每个匹配集包含至少一个处理个体和至少一个控制个体（每个匹配集可能有许多来自其他组）。就最小化每个匹配组内每个处理个体与每个对照个体之间的平均距离而言，完全匹配是最优的。因此，full matching可能对不愿丢弃某些对照个体但希望在倾向得分上获得最佳平衡的研究人员具有吸引力。

Weighting Adjustments

倾向得分可以也可以直接用作估计ATE中的逆权重，这种方法被称为处理权重的逆概率(inverse probability of treatment wrighting, IPTW)(Czajka et al., 1992; Robins, Hernan and Brumback, 2000; Lunceford and Davidian, 2004)。权重可以通过下式计算：

$$w_i = \frac{T_i}{\hat{e}_i} + \frac{1 - T_i}{1 - \hat{e}_i} \tag{10}$$

其中，$\hat{e}_k$是个体k的倾向得分估计值。这种权重可以用于对处理组和对照组进行加权至整个样本，就像调查抽样权重将样本加权至总体一样。另一种加权方式的权重可以表示为$w_i = T_i + (1 - T_i)\frac{\hat{e}_i}{1-\hat{e}_i}$。有了这个权重，接受干预的个体就会得到一个权重为1的加权。控制组个体通过$\frac{1}{1-\hat{e}_i}$项加权至全样本，处理组使用$\hat{e}_i$项。主要用于经济学文献的第三种加权方法是核加权，它对每个处理组个体的对照组中的多个个体进行平均，权重由个体之间的距离定义。加权方法的一个潜在缺点是，如果权重非常大，方差可能非常大。

匹配效果诊断

在匹配之前就需要检验协变量的平衡性，如果协变量比较平衡，也就没有必要实施匹配了，可以直接利用回归等方法进行因果效应的估计。匹配方法相当于从观测数据中将隐藏的随机化实验样本寻找出来。因而，对匹配完成后形成的匹配样本，需要检验是否近似于随机化实验。常用的检验指标包括标准化平均值差异(standardized difference in average)和对数标准差比(log ratio of standard deviations)。

标准化平均值差异定义为：

$$\hat{\Delta}_\alpha \equiv \frac{\bar{X}_t - \bar{X}_c}{\sqrt{\left(s_t^2 + s_c^2\right)/2}} \tag{11}$$

其中，$\bar{X}_d$,$d = t$,c表示干预组或控制组某协变量的平均值，s_d^2表示干预组或控制组协变量的方差，分别定义为：

$$s_t^2 = 1/\left(N_t - 1\right) \sum_{i_i D_i = 1} \left(X_i - \bar{X}_t\right)^2 \quad s_c^2 = 1/\left(N_c - 1\right) \sum_{i, D_i = 0} \left(X_i - \bar{X}_c\right)^2 \tag{12}$$

如果两组个体协变量完全平衡，标准化平均值差异将接近于0，因而，$\hat{\Delta}_\alpha$的值越接近于0，说明样本越有可能平衡，而且其与样本容量无关。

对数标准差比考察的是二阶矩的差异，定义为：

$$\hat{\Gamma}_a = \ln(s_t) - \ln(s_c) \tag{13}$$

如果两组协变量分布平衡，那么两组协变量标准差将相同，从而两组协变量标准差的对数比将接近于0.

但在非正态分布中，前两阶矩不一定决定整个分布，因此可以检验倾向指数的平衡性，如果两组倾向指数的期望值相同，那么两组个体的协变量分布将相同。

$$\hat{\Delta}_\alpha^l \equiv \frac{\bar{l}_t - \bar{l}_c}{\sqrt{\left(s_{l,t}^2 + s_{l,c}^2\right)/2}} \tag{14}$$

其中，$\bar{l}_d$是两组个体线性化倾向指数的平均值，$s_{l,t}^2$是两组倾向指数估计值的样本方差。

因果效应估计

After k:1 Matching

所有匹配估计量可以写成下列形式：

$$\hat{\tau}_{\mathrm{ATT}} = \frac{1}{N_t} \sum_{i, D_i=1} \left[Y_i - \sum_{j \in M_{j(i)}} w(i,j) Y_j \right] \tag{15}$$

其中，$0 < w(i,j) \leq 1$，$M_{j(i)}$是上文定义的与干预组个体i相匹配的控制组个体的集合。不同匹配方法的主要差别在于权重的差异。

匹配偏差可以写为：

$$\begin{aligned} \hat{B} &= \frac{1}{N_i} \sum_{i, D_i=1} \left[Y_i - \sum_{j \in M_{ji}} w(i,j) Y_j \right] - \frac{1}{N_t} \sum_{i_i, D_i - 1} (Y_i - Y_{0i}) \\ &= \frac{1}{N_i} \sum_{i: D_i=1} \left(Y_{0i} - \sum_{j \in M_{j(i)}} Y_j \right) \end{aligned} \tag{16}$$

其中，第一行的第一项是干预组平均因果效应的匹配估计量，第二项是干预组的平均因果效应，两者之差是估计偏差。干预组个体i的匹配误差可以表示为：

$$\begin{aligned} B_i(X_i) =& E\left[Y_i - Y_{0m_{j(i)}} \mid X_i, X_{m_{j(i)}}, D_i = 1\right] \\ & - E\left[Y_{1i} - Y_{0i} \mid X_i, X_{m_{j(i)}}, D_i = 1\right] \\ =& E\left[Y_{0i} \mid X_i, X_{m_{j(i)}}, D_i = 1\right] - E\left[Y_{0i} \mid X_i, X_{m_{j(i)}}, D_i = 0\right] \\ =& E\left[Y_{0i} \mid X_i\right] - E\left[Y_{0i} \mid X_{m_{j(i)}}\right] \\ =& \mu_c(X_i) - \mu_c\left(X_{m_{j(i)}}\right) \end{aligned} \tag{17}$$

其中$m_{j(i)}$表示与个体i的匹配误差，匹配上的两组个体的协变量或倾向指数相似，当协变量差异比较小时，线性回归函数是对条件期望函数非常好的近似，因而可以用线性回归方法估计$\mu_c(X_i)$

After Subclassification or Full Matching

使用标准的分层匹配，通常是先在每个子类内再跨子类聚合估计影响(Rosenbaum and Rubin, 1984)。通过每个子类中受到处理的个体数量对子类估计进行加权估计 ATT；对每个子类中个体总数进行加权估计 ATE。每个子类中可能存在大量的不平衡个体，因此需要将干预指标和协变量作为预测变量在每个子类内进行回归调整(Lunceford and Davidian, 2004)。当子类的数量太大 ，也就是每个子类中的个体数太小，无法估计每个子类内的单独回归模型时，可以拟合联合模型。具体来说，可以拟合$Y_{ij} = \beta_{0j} + \beta_{1j}T_{ij} + \gamma X_{ij} + e_{ij}$ 这种模型，这里i代表不同的个体，j代表不同的组别，然后跨组别加权平均这些效应以获得整体干预效应：$\beta = \frac{N_j}{N} \sum Jj = 1\beta_{1j}$，$J$是分组数量，$N_j$是分组$j$的个体总数，$N$是个体总数。

结语

当使用观测值估计因果效应时，最好通过获得具有相似协变量分布的治疗组和对照组来尽可能接近地重复随机实验。 通常，研究者可以通过选择原始处理组和控制组的匹配良好的样本来实现这一目标，从而减少因协变量造成的偏差。本文通过合并现有研究对关于匹配方法的文献进行了一个总结。本研究详细阐述了匹配的实际操作方法，以及不同匹配方法的对比，进而为读者提供了合适的方法及建议。

ROSENBAUM, P. R. and RUBIN, D. B. (1985b). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* **39** 33–38.

GU, X. and ROSENBAUM, P. R. (1993). Comparison of multivari- ate matching methods: Structures, distances, and algorithms. *J. Comput. Graph. Statist.* **2** 405–420.

RUBIN, D. B. and THOMAS, N. (2000). Combining propensity score matching with additional adjustments for prognostic co- variates. *J. Amer. Statist. Assoc.* **95** 573–585.

COHEN, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Earlbaum, Hillsdale, NJ.

HO, D. E., IMAI, K., KING, G. and STUART, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* **15** 199–236.

SMITH, H. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Method- ology* **27** 325–353.

ROSENBAUM, P. R. (1991). A characterization of optimal designs for observational studies. *J. Roy. Statist. Soc. Ser. B* **53** 597–610. MR1125717

H A N S E N , B . B . (2004). Full matching in an observational study of coaching for the SAT. *J. Amer. Statist. Assoc.* **99** 609–618. MR2086387

STUART, E. A. and GREEN, K. M. (2008). Using full matching to estimate causal effects in non-experimental studies: Examin- ing the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology* **44** 395–406.

CZAJKA, J. C., HIRABAYASHI, S., LITTLE, R. and RUBIN, D. B.(1992). Projecting from advance data using propensity modeling. *J. Bus. Econom. Statist.* **10** 117–131.

ROBINS, J. M., HERNAN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiol- ogy. *Epidemiology* **11** 550–560.

LUNCEFORD, J. K. and DAVIDIAN, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Stat. Med.* **23** 2937–2960.

ROSENBAUM, P. R. and RUBIN, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc.* **79** 516–524.