附录 1：数理证明

对任意的矩阵 M，记其矩阵范数为 $\| M \| = \sqrt{tr(M'M)}$，其中 $tr(\cdot)$ 表示矩阵的迹，而 M' 表示矩阵 M 的转置。对于方阵 A，记 $\lambda_{\min}(A)$ 表示 A 的最小特征值。令 C 表示一般的正常数，每次出现时可能代表不同的值。参考 Hirano 等（2003）中的方法，为逼近倾向得分函数 $\pi(\boldsymbol{x})$，本文采用幂级数逼近基序列 $\boldsymbol{R}^K(\boldsymbol{x})$ 并进一步将其进行正交化，于是有 $\mathrm{E}[\boldsymbol{R}^K(\boldsymbol{X})\boldsymbol{R}^K(\boldsymbol{X})'] = \boldsymbol{I}_K$ 且 $\zeta(K) \le CK$，其中 $\zeta(K) = \sup_{\boldsymbol{x}\in\mathcal{X}} \| \boldsymbol{R}^K(\boldsymbol{x}) \|$。

为了证明定理 1，需要先证明下面的一些引理。回忆正文中给出的记号，$L(u) = \exp(u)/(1+\exp(u))$ 为 logistic CDF，倾向得分函数 $\pi(\boldsymbol{x})$ 的估计量为 $\hat{\pi}_K(\boldsymbol{x}) = L(\boldsymbol{R}^K(\boldsymbol{x})'\hat{\boldsymbol{\beta}}_K)$，其中

$$\hat{\boldsymbol{\beta}}_K = \arg\min_{\boldsymbol{\beta}} M_n(\boldsymbol{\beta})$$

这里

$$M_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left[(1-D_i)e^{\boldsymbol{R}^K(\boldsymbol{X}_i)'\boldsymbol{\beta}} - D_i\left(\boldsymbol{R}^K(\boldsymbol{X}_i)'\boldsymbol{\beta}\right)\right]$$

记

$$S(D_i, \boldsymbol{X}_i, \boldsymbol{\beta}) = (D_i - 1)\exp\{\boldsymbol{R}^K(\boldsymbol{X}_i)'\boldsymbol{\beta}\} + D_i(\boldsymbol{R}^K(\boldsymbol{X}_i)'\boldsymbol{\beta})$$

$$S_n(D_i, \boldsymbol{X}_i, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}[(D_i - 1)\exp\{\boldsymbol{R}^K(\boldsymbol{X}_i)'\boldsymbol{\beta}\} + D_i(\boldsymbol{R}^K(\boldsymbol{X}_i)'\boldsymbol{\beta})]$$

则

$$\hat{\boldsymbol{\beta}}_K = \arg\max_{\boldsymbol{\beta}} S_n(D_i, \boldsymbol{X}_i, \boldsymbol{\beta})$$

直接计算可得到

$$\frac{\partial^2 S_n(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = \frac{1}{n}\sum_{i=1}^{n}(D_i - 1)e^{\boldsymbol{R}^K(\boldsymbol{X}_i)'\boldsymbol{\beta}}\boldsymbol{R}^K(\boldsymbol{X}_i)\boldsymbol{R}^K(\boldsymbol{X}_i)' < 0$$

因此 $S_n(D_i, \boldsymbol{X}_i, \boldsymbol{\beta})$ 在其定义域上是全局凹函数。对于任意固定的 K，定义

$$Q^*(\boldsymbol{\beta}) = \mathrm{E}[(\pi(\boldsymbol{X}) - 1)\exp\{\boldsymbol{R}^K(\boldsymbol{X})'\boldsymbol{\beta}\} + \pi(\boldsymbol{X})(\boldsymbol{R}^K(\boldsymbol{X})'\boldsymbol{\beta})]$$

$$Q_K(\boldsymbol{\beta}) = \mathrm{E}[(L(\boldsymbol{R}^K(\boldsymbol{X})'\boldsymbol{\beta}_K) - 1)\exp\{\boldsymbol{R}^K(\boldsymbol{X})'\boldsymbol{\beta}\} + L(\boldsymbol{R}^K(\boldsymbol{X})'\boldsymbol{\beta}_K)(\boldsymbol{R}^K(\boldsymbol{X})'\boldsymbol{\beta})]$$

它们的极大值点分别记为

$$\boldsymbol{\beta}_K^* = \arg\max_{\boldsymbol{\beta}} Q^*(\boldsymbol{\beta})$$

$$\boldsymbol{\beta}_K = \arg\max_{\boldsymbol{\beta}} Q_K(\boldsymbol{\beta})$$

那么当 $n \to \infty$ 时，有 $\hat{\boldsymbol{\beta}}_K$ 依概率收敛于 $\boldsymbol{\beta}_K^*$。最后定义 $\pi_K^*(\boldsymbol{x}) = L(\boldsymbol{R}^K(\boldsymbol{x})'\boldsymbol{\beta}_K^*)$。我们有下面的引理 1。

引理 1（倾向得分函数的近似）：假设

（1） $\boldsymbol{X}$ 的支撑集 $\mathcal{X}$ 是 $\mathbb{R}^m$ 上的一个紧子集；

（2）倾向得分函数 $\pi(\boldsymbol{x})$ 是 s 阶连续可微函数，其中 $s/m \geqslant 4$；

（3）倾向得分函数 $\pi(\boldsymbol{x})$ 在其支撑集上有界且远离 0 和 1；

（4） $\boldsymbol{X}$ 的密度函数在其支撑集上有界且远离 0。

那么我们有

$$\|\boldsymbol{\beta}_K - \boldsymbol{\beta}_K^*\| = O(K^{-s/(2m)})$$

和

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |\pi(\boldsymbol{x}) - \pi_K^*(\boldsymbol{x})| = O(K^{-s/(2m)}\zeta(K))$$

证明：由序列逼近以及 logistic CDF 的单调性，类似于 Hirano 等（2003）中引理 1 和 Heiler （2021）中定理 4.1 的证明，我们可以得到

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |\pi(\boldsymbol{x}) - L(\boldsymbol{R}^K(\boldsymbol{x})'\boldsymbol{\beta}_K)| < CK^{-s/m} \tag{A.1}$$

令 $\varepsilon = \inf\{\inf_{\boldsymbol{x}\in\mathcal{X}} \pi(\boldsymbol{x}), \inf_{\boldsymbol{x}\in\mathcal{X}}(1-\pi(\boldsymbol{x}))\}$，由假设 4（ii）知 $\varepsilon > 0$。定义集合

$$\Phi_K = \{\boldsymbol{\beta} \in \mathbb{R}^K : \inf\{\inf_{\boldsymbol{x}\in\mathcal{X}} L(\boldsymbol{R}^K(\boldsymbol{x})'\boldsymbol{\beta}), \inf_{\boldsymbol{x}\in\mathcal{X}}(1 - L(\boldsymbol{R}^K(\boldsymbol{x})'\boldsymbol{\beta}))\} \geqslant \varepsilon/2\}$$

由于(A.1)式，因此对足够大的 K，我们有 $\boldsymbol{\beta}_K \in \Phi_K$。对任意的 $\boldsymbol{\beta} \in \Phi_K$，有 $e^{\varepsilon/2} \leqslant \boldsymbol{R}^K(\boldsymbol{x})'\boldsymbol{\beta} \leqslant e^{1-\varepsilon/2}$，于是 $e^{\boldsymbol{R}^K(\boldsymbol{x})'\boldsymbol{\beta}}$ 和 $\boldsymbol{R}^K(\boldsymbol{x})'\boldsymbol{\beta}$ 在 Φ_K 上有界，从而根据 （A.1）式知存在常数 C_1 使得

$$\sup_{\boldsymbol{\beta}\in\Phi_K} |Q^*(\boldsymbol{\beta}) - Q_K(\boldsymbol{\beta})| \leqslant C_1 K^{-s/m}$$

对固定的常数 C_2，定义集合

$$\tilde{\Phi}_K = \{\boldsymbol{\beta} \in \mathbb{R}^K : \|\boldsymbol{\beta} - \boldsymbol{\beta}_K\| \leqslant C_2 K^{-s/(2m)}\} \tag{A.2}$$

根据 Hirano 等（2003）中引理 1 和 Heiler（2021）中定理 4.1 的证明，可以得到当 K 足够大时，有 $\tilde{\Phi}_K \subset \Phi_K$。注意到

$$-\frac{\partial^2 Q_K(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = \mathrm{E}\left\{(1 - L(\boldsymbol{R}^K(\boldsymbol{X})'\boldsymbol{\beta}_K))e^{\boldsymbol{R}^K(\boldsymbol{X})'\boldsymbol{\beta}}\boldsymbol{R}^K(\boldsymbol{X})\boldsymbol{R}^K(\boldsymbol{X})'\right\}$$

因此，根据 $\mathrm{E}[\boldsymbol{R}^K(\boldsymbol{X})\boldsymbol{R}^K(\boldsymbol{X})'] = \boldsymbol{I}_K$，对于 $\boldsymbol{\beta} \in \tilde{\Phi}_K$ 以及足够大的 K 使得 $\tilde{\Phi}_K \subset \Phi_K$，我们有

$$\lambda_{\min}\left(-\frac{\partial^2 Q_K(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right) = \lambda_{\min}\left(\mathrm{E}[(1 - L(\boldsymbol{R}^K(\boldsymbol{X})'\boldsymbol{\beta}_K))e^{\boldsymbol{R}^K(\boldsymbol{X})'\boldsymbol{\beta}}\boldsymbol{R}^K(\boldsymbol{X})\boldsymbol{R}^K(\boldsymbol{X})']\right) \geqslant \varepsilon/2$$

在（A.2）中选择 C_2 使其满足 $C_2 > \sqrt{4C_1/\varepsilon}$。取足够大的 K 使得 $\tilde{\Phi}_K \subset \Phi_K$，于是，对任意的 $\boldsymbol{\beta} \in \tilde{\Phi}_K$， $Q^*(\boldsymbol{\beta}_K)$ 与 $Q^*(\boldsymbol{\beta})$ 的差满足

$$
\begin{aligned}
Q^*(\boldsymbol{\beta}_K)-Q^*(\boldsymbol{\beta}) &\geqslant Q_K(\boldsymbol{\beta}_K)-C_1K^{-s/m}-Q_K(\boldsymbol{\beta})-C_1K^{-s/m} \\
&\geqslant -\frac{\partial Q_K(\boldsymbol{\beta}_K)}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}-\boldsymbol{\beta}_K)-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_K)'\frac{\partial^2 Q_K(\tilde{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}(\boldsymbol{\beta}-\boldsymbol{\beta}_K)-2C_1K^{-s/m} \\
&\geqslant \frac{C_2^2\varepsilon}{2}K^{-s/m}-2C_1K^{-s/m}>0
\end{aligned}
$$

其中 $\tilde{\boldsymbol{\beta}}$ 是位于 $\boldsymbol{\beta}$ 与 $\boldsymbol{\beta}_K$ 之间的点。因此，$Q^*(\boldsymbol{\beta})$ 的局部极大值点在 $\tilde{\Phi}_K$ 的内部取得。此外，由于 $Q^*(\boldsymbol{\beta})$ 是全局凹函数，故 $\boldsymbol{\beta}_K^*=\arg\max_{\boldsymbol{\beta}} Q^*(\boldsymbol{\beta})$ 必然满足 $\boldsymbol{\beta}_K^*\in\tilde{\Phi}_K$，此即证明了 $\|\boldsymbol{\beta}_K-\boldsymbol{\beta}_K^*\|=O(K^{-s/(2m)})$。注意到

$$
\left|L(\boldsymbol{R}^K(\boldsymbol{X})'\boldsymbol{\beta}_K)-L(\boldsymbol{R}^K(\boldsymbol{X})'\boldsymbol{\beta}_K^*)\right|\leqslant\|\boldsymbol{R}^K(\boldsymbol{X})\|\cdot\|\boldsymbol{\beta}_K^*-\boldsymbol{\beta}_K\|=O(\zeta(K)K^{-s/(2m)})
$$

因此，根据（A.1）式和三角不等式我们可得到

$$
\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\pi(\boldsymbol{x})-L(\boldsymbol{R}^K(\boldsymbol{x})'\boldsymbol{\beta}_K^*)\right|=O(\zeta(K)K^{-s/(2m)})
$$

至此，我们证明了引理 1。

引理 2. 假设引理 1 中的四个条件成立。此外，假设 $K(n)$ 是 K 的一序列取值，其满足 $K(n)\to\infty$ 和 $\zeta(K(n))^4/n\to 0$。那么我们有

$$
\left\|\hat{\boldsymbol{\beta}}_{K(n)}-\boldsymbol{\beta}_{K(n)}^*\right\|=O_p\left(\sqrt{\frac{K(n)}{n}}\right)
$$

证明：首先证明

$$
\frac{\partial S_n(\boldsymbol{\beta}_K^*)}{\partial \boldsymbol{\beta}}=O_p\left(\sqrt{\frac{K}{n}}\right) \tag{A.3}
$$

注意到

$$
\frac{\partial S_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}=\frac{1}{n}\sum_{i=1}^{n}\left\{(D_i-1)e^{\boldsymbol{R}^K(\boldsymbol{X}_i)'\boldsymbol{\beta}}\boldsymbol{R}^K(\boldsymbol{X}_i)+D_i\boldsymbol{R}^K(\boldsymbol{X}_i)\right\}
$$

于是，

$$
\begin{aligned}
\mathrm{E}\left[\left\|\frac{\partial S_n(\boldsymbol{\beta}_K^*)}{\partial \boldsymbol{\beta}}\right\|^2\right] &= \frac{1}{n}\mathrm{tr}\left(\mathrm{E}\left[\frac{\partial S_n(\boldsymbol{\beta}_K^*)}{\partial \boldsymbol{\beta}}\frac{\partial S_n(\boldsymbol{\beta}_K^*)}{\partial \boldsymbol{\beta}'}\right]\right) \\
&= \frac{1}{n}\mathrm{tr}\left(\mathrm{E}\left[\left((1-\pi(\boldsymbol{X}_i))e^{2\boldsymbol{R}^K(\boldsymbol{X}_i)'\boldsymbol{\beta}_K^*}+\pi(\boldsymbol{X}_i)\right)\boldsymbol{R}^K(\boldsymbol{X}_i)\boldsymbol{R}^K(\boldsymbol{X}_i)'\right]\right) \\
&\leqslant \frac{1}{n}\mathrm{tr}\left(\mathrm{E}\left[\sup_{\boldsymbol{x}\in\mathcal{X}}\left\{(1-\pi(\boldsymbol{x}))e^{2\boldsymbol{R}^K(\boldsymbol{x})'\boldsymbol{\beta}_K^*}+\pi(\boldsymbol{x})\right\}\boldsymbol{R}^K(\boldsymbol{X}_i)\boldsymbol{R}^K(\boldsymbol{X}_i)'\right]\right) \\
&\leqslant \frac{C}{n}\mathrm{tr}\left(\mathrm{E}\left[\boldsymbol{R}^K(\boldsymbol{X}_i)\boldsymbol{R}^K(\boldsymbol{X}_i)'\right]\right) \\
&= \frac{CK}{n}
\end{aligned}
$$

由 Markov 不等式知 $\frac{\partial S_n(\boldsymbol{\beta}_K^*)}{\partial \boldsymbol{\beta}} = O_p\left(\sqrt{\frac{K}{n}}\right)$。接下来我们需要证明对任意满足条件 $\left\|\boldsymbol{\beta}-\boldsymbol{\beta}_K^*\right\| = C\sqrt{K/n}$ 的 $\boldsymbol{\beta}$，$S_n(\boldsymbol{\beta}) < S_n(\boldsymbol{\beta}_K^*)$ 以任意大的概率成立。由中值定理知

$$S_n(\boldsymbol{\beta}) - S_n(\boldsymbol{\beta}_K^*) = \frac{\partial S_n(\boldsymbol{\beta}_K^*)}{\partial \boldsymbol{\beta}'}(\boldsymbol{\beta}-\boldsymbol{\beta}_K^*) + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_K^*)'\frac{\partial^2 S_n(\tilde{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}(\boldsymbol{\beta}-\boldsymbol{\beta}_K^*)$$

其中 $\tilde{\boldsymbol{\beta}}$ 位于 $\boldsymbol{\beta}$ 与 $\boldsymbol{\beta}_K^*$ 之间。既然 $\tilde{\boldsymbol{\beta}}$ 位于 $\boldsymbol{\beta}$ 与 $\boldsymbol{\beta}_K^*$ 之间，于是 $\tilde{\boldsymbol{\beta}} \in \tilde{\Phi}_K$。对任意固定的 K，

$$\begin{aligned}
-\mathrm{E}\left[\frac{\partial^2 S_n(\tilde{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(1-D_i)e^{\boldsymbol{R}^K(\boldsymbol{X}_i)'\tilde{\boldsymbol{\beta}}}\boldsymbol{R}^K(\boldsymbol{X}_i)\boldsymbol{R}^K(\boldsymbol{X}_i)'\right] \\
&= \mathrm{E}\left[(1-\pi(\boldsymbol{X}_i))e^{\boldsymbol{R}^K(\boldsymbol{X}_i)'\tilde{\boldsymbol{\beta}}}\boldsymbol{R}^K(\boldsymbol{X}_i)\boldsymbol{R}^K(\boldsymbol{X}_i)'\right] \\
&\geqslant \inf_{\boldsymbol{x}\in\mathcal{X}}\left\{(1-\pi(\boldsymbol{x}))e^{\boldsymbol{R}^K(\boldsymbol{x})'\tilde{\boldsymbol{\beta}}}\right\}\mathrm{E}\left[\boldsymbol{R}^K(\boldsymbol{X}_i)\boldsymbol{R}^K(\boldsymbol{X}_i)'\right] \\
&\geqslant A\boldsymbol{I}_K
\end{aligned}$$

这就意味着

$$-\frac{\partial^2 S_n(\tilde{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} \geqslant C'\boldsymbol{I}_K + o_p(1)$$

因此，我们可以找到一个样本大小 n_0，使得当 $n \geqslant n_0$ 时，以至少 $1-\varepsilon/2$ 的概率有 $\lambda_{\min}\left(\partial^2 S_n(\tilde{\boldsymbol{\beta}})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'\right) \geqslant A/2$。由（A.3）知当选择足够大的 $n \geqslant n_0$ 以及常数 C，有

$$\mathrm{P}\left(\left\|\frac{\partial S_n(\boldsymbol{\beta}_K^*)}{\partial\boldsymbol{\beta}}\right\| < \frac{A}{4}C\sqrt{\frac{K}{n}}\right) \geqslant 1-\frac{\varepsilon}{2}$$

因此，对 K 以及足够大的 n，当 $\left\|\boldsymbol{\beta}-\boldsymbol{\beta}_K^*\right\| = C\sqrt{K/n}$ 时，以至少 $1-\varepsilon$ 的概率有

$$\begin{aligned}
S_n(\boldsymbol{\beta}) - S_n(\boldsymbol{\beta}_K^*) &\leqslant \left\|\frac{\partial S_n(\boldsymbol{\beta}_K^*)}{\partial\boldsymbol{\beta}}\right\|\left\|\boldsymbol{\beta}-\boldsymbol{\beta}_K^*\right\| - \frac{A}{4}\left\|\boldsymbol{\beta}-\boldsymbol{\beta}_K^*\right\|^2 \\
&\leqslant \left(\left\|\frac{\partial S_n(\boldsymbol{\beta}_K^*)}{\partial\boldsymbol{\beta}}\right\| - \frac{A}{4}C\sqrt{\frac{K}{n}}\right)\left\|\boldsymbol{\beta}-\boldsymbol{\beta}_K^*\right\| \\
&< 0
\end{aligned}$$

注意到 $S_n(\boldsymbol{\beta})$ 是连续函数，因此它在紧集 $\left\{\boldsymbol{\beta}: \left\|\boldsymbol{\beta}-\boldsymbol{\beta}_K^*\right\| \leqslant C\sqrt{K/n}\right\}$ 上必有极大值。此外，由上面的不等式知 $S_n(\boldsymbol{\beta})$ 在紧集 $\left\{\boldsymbol{\beta}: \left\|\boldsymbol{\beta}-\boldsymbol{\beta}_K^*\right\| \leqslant C\sqrt{K/n}\right\}$ 上的极大值点 $\hat{\boldsymbol{\beta}}_K$ 必然满足 $\left\|\hat{\boldsymbol{\beta}}_K-\boldsymbol{\beta}_K^*\right\| < C\sqrt{K/n}$，从而在点 $\hat{\boldsymbol{\beta}}_K$ 上满足一阶条件且由 $S_n(\boldsymbol{\beta})$ 的全局凹性知 $\hat{\boldsymbol{\beta}}_K$ 也是 $S_n(\boldsymbol{\beta})$ 在 $\mathbb{R}^K$ 上的极大值点。根据以上的讨论，$\hat{\boldsymbol{\beta}}_K$ 是 $S_n(\boldsymbol{\beta})$ 在 $\mathbb{R}^K$ 上的极大值点这一事件至少以概率 $1-\varepsilon$ 发生，于是 $\hat{\boldsymbol{\beta}}_K$ 存在且以趋于 1 的概率满足一阶条件，因此 $\left\|\hat{\boldsymbol{\beta}}_K-\boldsymbol{\beta}_K^*\right\| = O_p(\sqrt{K/n})$。

记 $\delta_0 = \mathrm{E}\left(\frac{1-D}{1-\pi(\boldsymbol{X})}\Delta Y\right)$，$\hat{\delta}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{1-D_i}{1-\hat{\pi}(\boldsymbol{X}_i)}\Delta Y_i$ 和 $\eta_0(\boldsymbol{x}) = \mathrm{E}(\Delta Y \mid D=0, \boldsymbol{X}=\boldsymbol{x})$，那么有如下的引理 3。

引理 3. 在定理 1 的假设下，有

$$\sqrt{n}(\hat{\delta}_n-\delta_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{1-D_i}{1-\pi(\boldsymbol{X}_i)}\Delta Y_i-\delta_0-\frac{\eta_0(\boldsymbol{X}_i)}{1-\pi(\boldsymbol{X}_i)}\left(\pi(\boldsymbol{X}_i)-D_i\right)\right\}+o_p(1)$$

证明：本引理的证明可根据类似于 Hirano 等（2003）中定理 1 的证明过程得到，为节省篇幅，这里省略其详细证明过程。

定理 1 的证明：由前面的定义知

$$\alpha_0=\frac{1}{\mathrm{E}(D)}\left\{\mathrm{E}(\Delta Y)-\mathrm{E}\left(\frac{1-D}{1-\pi(\boldsymbol{X})}\Delta Y\right)\right\}=\frac{1}{\mathrm{E}(D)}\left(\mathrm{E}(\Delta Y)-\delta_0\right)$$

定义

$$\tilde{\alpha}_n=\frac{1}{\mathrm{E}(D)}\left(\frac{1}{n}\sum_{i=1}^{n}\Delta Y_i-\hat{\delta}_n\right)$$

于是有

$$\begin{aligned}
&\sqrt{n}(\tilde{\alpha}_n-\alpha_0)\\
&=\frac{1}{E(D)}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\Delta Y_i-\mathrm{E}(\Delta Y)\right)-\sqrt{n}(\hat{\delta}_n-\delta_0)\right\}\\
&=\frac{1}{\mathrm{E}(D)}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\Delta Y_i-\mathrm{E}(\Delta Y)\right)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{1-D_i}{1-\pi(\boldsymbol{X}_i)}\Delta Y_i-\delta_0-\frac{\eta_0(\boldsymbol{X}_i)}{1-\pi(\boldsymbol{X}_i)}\left(\pi(\boldsymbol{X}_i)-D_i\right)\right)+o_p(1)\right\}\\
&=\frac{1}{\mathrm{E}(D)}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{D_i-\pi(\boldsymbol{X}_i)}{1-\pi(\boldsymbol{X}_i)}\left(\Delta Y_i-\eta_0(\boldsymbol{X}_i)\right)-\left(\mathrm{E}(\Delta Y)-\delta_0\right)\right\}+o_p(1)\\
&=\frac{1}{\mathrm{E}(D)}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{D_i-\pi(\boldsymbol{X}_i)}{1-\pi(\boldsymbol{X}_i)}\left(\Delta Y_i-\eta_0(\boldsymbol{X}_i)\right)-\mathrm{E}(D)\alpha_0\right\}+o_p(1)
\end{aligned}$$

注意到

$$\begin{aligned}
\sqrt{n}(\hat{\alpha}_n-\tilde{\alpha}_n)&=\sqrt{n}\left(\frac{1}{\hat{\mu}_n}-\frac{1}{\mu_0}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\Delta Y_i-\hat{\delta}_n\right)\\
&=-\sqrt{n}\left(\hat{\mu}_n-\mu_0\right)\frac{\hat{\alpha}_n}{\hat{\mu}_n}\\
&=-\frac{\hat{\alpha}_n}{\sqrt{n}}\sum_{i=1}^{n}\frac{D_i-\mu_0}{\hat{\mu}_n}
\end{aligned}$$

因此，我们有

$$\begin{aligned}
\sqrt{n}(\hat{\alpha}_n-\alpha_0)&=\sqrt{n}(\tilde{\alpha}_n-\alpha_0)+\sqrt{n}(\hat{\alpha}_n-\tilde{\alpha}_n)\\
&=\frac{1}{\mathrm{E(D)}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{D_i-\pi(\boldsymbol{X}_i)}{1-\pi(\boldsymbol{X}_i)}\left(\Delta Y_i-\eta_0(\boldsymbol{X}_i)\right)-\mathrm{E}(D)\alpha_0\right\}-\frac{\hat{\alpha}_n}{\sqrt{n}}\sum_{i=1}^{n}\frac{D_i-\mathrm{E}(D)}{\hat{\mu}_n}+o_p(1)\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\lambda(\Delta Y_i,\boldsymbol{X}_i,D_i;\alpha_0)+\left(\frac{\alpha_0}{\mu_0}-\frac{\tilde{\alpha}_n}{\hat{\mu}_n}\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(D_i-\mu_0\right)+o_p(1)
\end{aligned}$$

其中

$$\lambda(\Delta Y_i, \boldsymbol{X}_i, D_i; \alpha_0) = \frac{1}{\mathrm{E}(D)}\left[\frac{D_i - \pi(\boldsymbol{X}_i)}{1-\pi(\boldsymbol{X}_i)}\left(\Delta Y_i - \eta_0(\boldsymbol{X}_i)\right) - D_i\alpha_0\right]$$

由中心极限定理知 $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(D_i - \mu_0\right) = O_p(1)$，其加上 $\tilde{\alpha}_n$ 依概率收敛于 α_0 且 $\hat{\mu}_n$ 依概率收敛于 μ_0 意味着

$$\sqrt{n}(\hat{\alpha}_n - \alpha_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\lambda(\Delta Y_i, \boldsymbol{X}_i, D_i; \alpha_0) + o_p(1) \to_d N(0, V)$$

其中

$$V = \mathrm{E}[\phi^2(\Delta Y_i, \boldsymbol{X}_i, D_i; \alpha_0)]$$

定理 2 的证明：本定理的证明可根据类似于 Hirano 等（2003）中定理 2 的证明得到，为节省篇幅，这里省略其详细证明过程。

附录 2：方法在其他模型中的应用

为讨论本文提出方法在不同模型下的适用性，下面考虑本文提出方法在以下两种不同模型下的应用：一是重复截面数据（Repeated Cross-section Data）模型，二是交错双重差分（Staggered Difference-in-Differences）模型。

一、重复截面数据模型

标准的双重差分法一般用于面板数据类型，但有时研究者并不能获得对每个个体结果的重复观测，而只能获得重复截面数据。本文提出方法仍然适用于这种数据类型。为了在这种情形下讨论本文提出的方法，我们需要重新定义模型框架。

具体来说，假设研究者可得到的数据类型是重复截面数据，即研究者可观测到接受处理之前与接受处理之后的混合样本（Pooled Sample）$(Y_i, \boldsymbol{X}_i, D_i, T_i)_{i=1}^{n}$，其中 $Y_i = Y_{i0} + T_i(Y_{i1} - Y_{i0})$，$T_i$ 是时间指示变量，它表示如果观测值属于接受处理之后样本则取值为 1，否则取值为 0。令 n_1 与 n_0 分别表示接受处理之后与接受处理之前的样本量，于是有 $n = n_1 + n_0$。更具体地，我们假定混合重复截面数据（Pooled Repeated Cross-section Data）$\{Y_i, \boldsymbol{X}_i, D_i, T_i\}_{i=1}^{n}$ 是来自以下混合分布中的独立同分布样本

$$\begin{aligned}\mathrm{P}(Y \leqslant y, D = d, \boldsymbol{X} \leqslant \boldsymbol{x}, T = t) &= t \cdot \lambda \cdot \mathrm{P}(Y_1 \leqslant y, D = d, \boldsymbol{X} \leqslant \boldsymbol{x} \mid T = 1) \\ &\quad + (1-t)\cdot(1-\lambda)\cdot \mathrm{P}(Y_0 \leqslant y, D = d, \boldsymbol{X} \leqslant \boldsymbol{x} \mid T = 0)\end{aligned}$$

其中 $(y, \boldsymbol{x}, d, t) \in \mathcal{R} \times \mathcal{R}^m \times \{0,1\} \times \{0,1\}$，$\lambda = \mathrm{P}(T = 1) \in (0,1)$，且 $(D, \boldsymbol{X})$ 的分布不随时间 T 变化。

在以上设定下，研究者感兴趣的因果效应参数仍然为接受处理组的平均处理效应，即

$$\alpha_0 = \mathrm{E}[Y_1(1) - Y_1(0) \mid D = 1] = \mathrm{E}[Y_1(1) \mid D = 1] - \mathrm{E}[Y_1(0) \mid D = 1]$$

Abadie（2005）证明了因果效应参数 α_0 具有如下识别结果：

$$\alpha_0 = \frac{1}{\mathrm{E}(D)}\mathrm{E}\left[\frac{D-\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}\frac{T-\lambda}{\lambda(1-\lambda)}Y\right]$$

根据以上识别结果，我们可将其重写为

$$\alpha_0 = \frac{1}{\mathrm{E}(D)}\left[\mathrm{E}\left\{\frac{T-\lambda}{\lambda(1-\lambda)}Y\right\} - \mathrm{E}\left\{\frac{1-D}{1-\pi(\boldsymbol{X})}\cdot\frac{T-\lambda}{\lambda(1-\lambda)}Y\right\}\right]$$

与前面类似，倾向得分函数 $\pi(\boldsymbol{x})$ 满足对任意的可积函数 $\xi(\boldsymbol{x})$ 有

$$\mathrm{E}\left[\frac{1-D}{1-\pi(\boldsymbol{X})}\xi(\boldsymbol{X})\right] = \mathrm{E}[\xi(\boldsymbol{X})]$$

于是，可以使用正文中提出的协变量平衡法估计倾向得分函数 $\pi(\boldsymbol{x})$，进而得到重复截面数据模型下感兴趣因果效应参数 α_0 的估计量为

$$\hat{\alpha}_n = \frac{1}{\hat{\mu}_n}\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{T_i-\hat{\lambda}_n}{\hat{\lambda}_n(1-\hat{\lambda}_n)}Y_i - \frac{1-D_i}{1-\hat{\pi}(\boldsymbol{X}_i)}\cdot\frac{T_i-\hat{\lambda}_n}{\hat{\lambda}_n(1-\hat{\lambda}_n)}Y_i\right\}$$

其中 $\hat{\lambda}_n = \frac{1}{n}\sum_{i=1}^{n}T_i$， $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n}D_i$， $\hat{\pi}(\boldsymbol{X}_i)$ 是本文提出的 $\pi(\boldsymbol{X}_i)$ 的协变量平衡估计量。

二、交错双重差分模型

前面考虑的标准双重差分模型需要假设处理组在同一个时间点接受处理且只有两期数据，然而现实中存在相当多的政策并非是一次性全面实施，而是先通过小面积试点再逐步推广，在渐进的过程中推而行之，这方面的政策如增值税转型、土地确权、新农保实施、高铁修建等。在这种情形下，标准的双重差分法并不能很好地对这些政策进行评估与分析。为了在这种情形下识别与估计因果效应，文献中一个常用的方法是交错双重差分法（Staggered Difference-in-Differences）。

为讨论本文提出方法在交错双重差分模型下的应用，考虑 Callaway 和 Sant'Anna（2021）中的模型框架。从直观上说，交错双重差分模型下的处理效应可能在第一次接受处理的时点不同以及接受处理时长不同之间存在异质性，为了刻画这种异质性，文献中考虑了“组别-时期平均处理效应”（Group-time Average Treatment Effect），而研究者关心的动态处理效应和平均处理效应则可以在组别-时期平均处理效应的基础上通过加权求和得到。

具体而言，我们令 T 表示观测时间长度，而 t 表示某一特定的观测时间，其中 $t=1,\ldots,T$ 。在标准的双重差分模型下，$T=2$ 且在 $t=1$ 时没有个体接受处理。令 D_{it} 表示个体 i 在时间 t 是否接受处理的二元变量，如果个体接受了处理则取值为 1 否则取值为 0。令 G 表示个体第一次接受处理的时间。定义指示变量 G_{ig}，它表示当个体 i 第一次接受处理的时间是 g 时取值为 1 否则取值为 0，即 $G_{ig} = I\{G_i = g\}$，其中 $I\{\cdot\}$ 为示性函数。此外，定义虚拟变量 C_i，它表示如果个体 i 从未接受处理则取值为 1 否则取值为 0。于是，对于样本中的每个个体，在二元变量集合 $\{G_1,\ldots,G_T,C\}$ 中必有一个二元变量取值为 1。

定义广义倾向得分函数为 $\pi_g(\boldsymbol{X})=\mathrm{P}(G_g=1\mid \boldsymbol{X},G_g+C=1)$，它表示在给定协变量 $\boldsymbol{X}$ 和处理状态下，个体在时间 g 第一次接受处理的概率。令 $Y_{it}(0)$ 表示个体在时间 t 未接受处理时得到的潜在结果。对于 $g=2,\ldots,T$，我们用 $Y_{it}(g)$ 表示在时点 g 首次接受处理的个体在时间 t 时的潜在结果，而 Y_{it} 则表示个体在时间 t 的观测结果。于是，观测结果与潜在结果具有如下关系

$$Y_{it}=Y_{it}(0)+\sum_{g=2}^{T}(Y_{it}(g)-Y_{it}(0))\cdot G_{ig}$$

最后，定义如下的组别-时期平均处理效应（Group-time Average Treatment Effect）：

$$ATT(g,t)=\mathrm{E}[Y_t(g)-Y_t(0)\mid G_g=1]$$

在以上设定和存在从未接受处理组[①]（Never-Treated Groups）以及其它一些假设下，Callaway 和 Sant'Anna（2021）证明了 $ATT(g,t)$ 具有如下的逆概率加权识别结果：

$$ATT(g,t)=\mathrm{E}\left[\left(\frac{G_g}{\mathrm{E}[G_g]}-\frac{\pi_g(\boldsymbol{X})C}{1-\pi_g(\boldsymbol{X})}\bigg/\mathrm{E}\left[\frac{\pi_g(\boldsymbol{X})C}{1-\pi_g(\boldsymbol{X})}\right]\right)(Y_t-Y_{g-1})\right]$$

该参数背后的直觉在于计算 $ATT(g,t)$ 时仅需考虑组别为第一次接受处理时间为 g 的样本和从未接受处理为控制组的样本（由 C 表示），而忽略所有其他样本。与此同时，给予控制组中的那些与处理组内出现更频繁的个体特征更相似的样本以更大的权重，反之则给予较小的权重。这种做法确保了处理组和控制组的个体在特征方面的平衡性。

在获得 $ATT(g,t)$ 的基础上，将其进行加权平均即可获得平均处理效应 θ：

$$\theta=\sum_{g}\sum_{t=2}^{T}w(g,t)\cdot ATT(g,t)$$

其中 $w(g,t)$ 是研究者仔细选择的已知或者可估计的权重函数，且其不同选择使得研究者可考虑不同维度的处理效应异质性。特别地，Callaway 和 Sant'Anna（2021）指出可以使用接受处理时间为 g 的组别的占比 $\mathrm{P}(G=g\mid G\leqslant T)$ 为权重，从而定义出平均处理效应对应的参数：

$$\theta_w^0=\frac{1}{\kappa}\sum_{g}\sum_{t=2}^{T}I\{t\geqslant g\}ATT(g,t)\mathrm{P}(G=t\mid G\leqslant T)$$

其中 $\kappa=\sum_{g}\sum_{t=2}^{T}I\{t\geqslant g\}\mathrm{P}(G=g\mid G\leqslant T)$。有关权重函数 $w(g,t)$ 选择的更多讨论，可参见 Callaway 和 Sant'Anna（2021）中的表 1。

[①]Callaway 和 Sant'Anna（2021）指出，研究者在某些情形下可能怀疑“从未接受处理组”在协变量特征上与接受处理组的样本存在较大差异，出于这种顾虑可删除数据中所有从未接受政策处理的样本，而在估计时选择使用“尚未接受处理组”（Not-Yet-Treated Groups）作为控制组，但需要注意的是两者在计算 $ATT(g,t)$ 时的公式上有所不同。下文考虑的协变量平衡法同样适用于这种情形下对广义倾向得分函数的估计，为了节约篇幅，本文仅给出了以从未接受处理组为控制组时的估计方法。

下面考虑使用本文提出的协变量平衡法估计广义倾向得分函数 $\pi_g(\boldsymbol{x})$。首先注意到

$$\pi_g(\boldsymbol{X})=\frac{\mathrm{E}[G_g\mid \boldsymbol{X}]}{\mathrm{E}[G_g+C\mid \boldsymbol{X}]},\ 1-\pi_g(\boldsymbol{X})=\frac{\mathrm{E}[C\mid \boldsymbol{X}]}{\mathrm{E}[G_g+C\mid \boldsymbol{X}]}$$

通过简单计算可得到

$$\mathrm{E}\left[\frac{\pi_g(\boldsymbol{X})C}{1-\pi_g(\boldsymbol{X})}\right]=\mathrm{E}[G_g]$$

因此，上面关于 $ATT(g,t)$ 的逆概率加权识别结果可被重新写为

$$ATT(g,t)=\frac{1}{\mathrm{E}[G_g]}\mathrm{E}\left[(G_g+C)(Y_t-Y_{g-1})-\frac{C}{1-\pi_g(\boldsymbol{X})}(Y_t-Y_{g-1})\right]$$

此外，对任意的可积函数 $\zeta(\boldsymbol{X})$，有

$$\mathrm{E}\left[\frac{C}{1-\pi_g(\boldsymbol{X})}\zeta(\boldsymbol{X})\right]=\mathrm{E}[(G_g+C)\zeta(\boldsymbol{X})]$$

即广义倾向得分函数 $\pi_g(\boldsymbol{x})$ 可平衡控制组（由 C 表示）与“研究总体”（由 G_g+C 表示）协变量的所有阶矩。

与前面一样，我们考虑使用协变量平衡法估计广义倾向得分函数 $\pi_g(\boldsymbol{x})$。令 $n_{g,c}$ 表示属于组别 G_g 或者组别 C 的所有样本数。注意到 $\pi_g(\boldsymbol{X})=\mathrm{P}(G_g=1\mid \boldsymbol{X},G_g+C=1)$，因此我们提出的 $\pi_g(\boldsymbol{x})$ 的估计量为

$$\hat{\pi}_g(\boldsymbol{x})=L\left(\boldsymbol{R}^K(\boldsymbol{x})'\hat{\boldsymbol{\gamma}}_K\right)$$

其中

$$\hat{\boldsymbol{\gamma}}_K=\arg\min_{\boldsymbol{\gamma}} Q_n(\boldsymbol{\gamma})$$

这里

$$Q_n(\boldsymbol{\gamma})=\frac{1}{n_{g,c}}\sum_{i:G_{ig}+C_i=1}\left[(1-G_{ig})e^{\boldsymbol{R}^K(\boldsymbol{X}_i)'\boldsymbol{\gamma}}-G_{ig}(\boldsymbol{R}^K(\boldsymbol{X}_i)'\boldsymbol{\gamma})\right]$$

将 $Q_n(\boldsymbol{\gamma})$ 对 $\boldsymbol{\gamma}$ 求二阶导数，可得到

$$\frac{\partial^2 Q_n(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'}>0$$

即 $Q_n(\boldsymbol{\gamma})$ 关于 $\boldsymbol{\gamma}$ 是一个全局严格凸函数。此外，将 $\hat{\boldsymbol{\gamma}}_K$ 代入到 $Q_n(\boldsymbol{\gamma})$ 的一阶导数并令其等于 0，然后经过一些简单代数运算可得到

$$\frac{1}{n_{g,c}}\sum_{i:G_{ig}+C_i=1}\left\{\left(\frac{1-G_{ig}}{1-\hat{\pi}_g(\boldsymbol{X}_i)}-1\right)\boldsymbol{R}^K(\boldsymbol{X}_i)\right\}=\frac{1}{n_{g,c}}\sum_{i:G_{ig}+C_i=1}\left\{\left(\frac{C_i}{1-\hat{\pi}_g(\boldsymbol{X}_i)}-1\right)\boldsymbol{R}^K(\boldsymbol{X}_i)\right\}=\boldsymbol{0}$$

因此，只要研究者选择合适的逼近基向量 $\boldsymbol{R}^K(\boldsymbol{X})$，估计量 $\hat{\pi}_g(\boldsymbol{x})$ 能够平衡控制组与“研究

总体”协变量的所有阶样本矩。此外，如果我们在逼近基序列 $\boldsymbol{R}^K(\boldsymbol{X})$ 中包含截距项，则有

$$\frac{1}{n_{g,c}}\sum_{i:G_{ig}+C_i=1}\frac{C_i}{1-\hat{\pi}_g(\boldsymbol{X}_i)}=1$$

即估计权重 $\{C_i/(1-\hat{\pi}_g(\boldsymbol{X}_i)):i\in\{G_{ig}+C_i=1\}\}$ 自动满足归一化条件。

最后，一旦研究者获得 $\pi_g(\boldsymbol{x})$ 的估计量 $\hat{\pi}_g(\boldsymbol{x})$ 之后，通过样本矩估计总体矩的方法，我们可得到 $ATT(g,t)$ 的估计量为

$$\hat{\alpha}_n(g,t)=\frac{1}{\frac{1}{n}\sum_{i=1}^{n}G_{ig}}\frac{1}{n}\sum_{i=1}^{n}\left[(G_{ig}+C_i)(Y_{it}-Y_{ig-1})-\frac{C_i}{1-\hat{\pi}_g(\boldsymbol{X}_i)}(Y_{it}-Y_{ig-1})\right]$$

基于此，研究者可进一步得到各种加权处理效应的估计量。