Notes

KCY

2024 年 11 月 18 日

1 记号

$$\operatorname{Cov}\left(Z_i, Z_j\right)=\begin{cases}\frac{n_1\left(n_1-1\right)}{n(n-1)}-\left(\frac{n_1}{n}\right)^2 & \text{if } i \neq j \\ \frac{n_1}{n}\left(1-\frac{n_1}{n}\right) & \text{if } i=j\end{cases}$$

$$\hat{\boldsymbol{\tau}}_X \equiv \overline{\boldsymbol{X}}_1-\overline{\boldsymbol{X}}_0=\frac{1}{n_1} \sum_{i=1}^n Z_i \boldsymbol{X}_i-\frac{1}{n_0} \sum_{i=1}^n\left(1-Z_i\right) \boldsymbol{X}_i$$

$$\begin{aligned} M=\hat{\boldsymbol{\tau}}_{\boldsymbol{X}}^{\top} \operatorname{Cov}^{-1}\left(\hat{\boldsymbol{\tau}}_{\boldsymbol{X}}\right) \hat{\boldsymbol{\tau}}_{\boldsymbol{X}} & =\hat{\boldsymbol{\tau}}_{\boldsymbol{X}}^{\top}\left(\frac{n}{n_1 n_0} \boldsymbol{S}_{\boldsymbol{X}}^2\right)^{-1} \hat{\boldsymbol{\tau}}_{\boldsymbol{X}} \\ & =\frac{n_1 n_0}{n}\left(\overline{\boldsymbol{X}}_1-\overline{\boldsymbol{X}}_0\right)^{\top}\left(\boldsymbol{S}_{\boldsymbol{X}}^2\right)^{-1}\left(\overline{\boldsymbol{X}}_1-\overline{\boldsymbol{X}}_0\right) . \end{aligned}$$

$$\boldsymbol{S}_{\boldsymbol{X}}^2=(n-1)^{-1} \sum_{i=1}^n\left(\boldsymbol{X}_i-\overline{\boldsymbol{X}}\right)\left(\boldsymbol{X}_i-\overline{\boldsymbol{X}}\right)^{\top}$$

$$
\begin{aligned}
\operatorname{Cov}\left(\hat{\boldsymbol{\tau}}_{\boldsymbol{X}}\right) &= \operatorname{Cov}\left(\frac{1}{n_1}\sum_{i=1}^{n} Z_i \boldsymbol{X}_i - \frac{1}{n_0}\sum_{i=1}^{n}\left(1-Z_i\right)\boldsymbol{X}_i\right) \\
&= \operatorname{Cov}\left(\frac{n}{n_0 n_1}\sum_{i=1}^{n} Z_i \boldsymbol{X}_i\right) \\
&= \left(\frac{n}{n_0 n_1}\right)^2 \sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{X_i}\boldsymbol{X_j}^T \operatorname{Cov}\left(Z_i, Z_j\right) \\
&= \frac{n}{n_0 n_1}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X_i}\boldsymbol{X_i}^T - \frac{1}{n(n-1)}\sum_{i\neq j} \boldsymbol{X_i}\boldsymbol{X_j}^T\right) \\
&\overset{*}{=} \frac{n}{n_0 n_1}\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)^\top\right) \\
&= \frac{n}{n_0 n_1}\boldsymbol{S}_{\boldsymbol{X}}^2
\end{aligned}
$$

$$\tau_i = Y_i(1) - Y_i(0)$$

$$\tau = n^{-1}\sum_{i=1}^{n}\tau_i$$

$$Y_i = Z_i Y_i(1) + \left(1 - Z_i\right) Y_i(0)$$

$$S_z^2 = (n-1)^{-1}\sum_{i=1}^{n}\left\{Y_i(z) - \bar{Y}(z)\right\}^2$$

$$S_{\boldsymbol{X}}^2 = (n-1)^{-1}\sum_{i=1}^{n}\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)\left(\boldsymbol{X}_i - \bar{\boldsymbol{X}}\right)^\top$$

$$\boldsymbol{S}_{z,\boldsymbol{X}} = \boldsymbol{S}_{\boldsymbol{X},z}^\top = (n-1)^{-1}\sum_{i=1}^{n}\left\{Y_i(z) - \bar{Y}(z)\right\}\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)^\top$$

$$S_\tau^2 = (n-1)^{-1}\sum_{i=1}^{n}\left(\tau_i - \tau\right)^2$$

$$\boldsymbol{S}_{\tau,X} = \boldsymbol{S}_{\boldsymbol{X},\tau}^\top = (n-1)^{-1}\sum_{i=1}^{n}\left(\tau_i - \tau\right)\left(\boldsymbol{X}_i - \overline{\boldsymbol{X}}\right)^\top$$

$$\hat{\tau} \equiv \bar{Y}_1 - \bar{Y}_0 = \frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-Z_i)\,Y_i$$

$$S^2_{z|\boldsymbol{X}} = \boldsymbol{S}_{z,\boldsymbol{X}}\boldsymbol{S}_{\boldsymbol{X}}^{-2}\boldsymbol{S}_{\boldsymbol{X},z}$$

$$S^2_{\tau|\boldsymbol{X}} = \boldsymbol{S}_{\tau,\boldsymbol{X}}\boldsymbol{S}_{\boldsymbol{X}}^{-2}\boldsymbol{S}_{\boldsymbol{X},\tau}$$

$$\boldsymbol{u}_i = \left(r_0 Y_i(1) + r_1 Y_i(0), \boldsymbol{X}_i^\top\right)^\top \in \mathbb{R}^{K+1} \quad (r_0 = n_0/n, r_1 = n_1/n)$$

$$\begin{pmatrix}\hat{\tau}\\ \hat{\boldsymbol{\tau}}_{\boldsymbol{X}}\end{pmatrix} = \frac{n}{n_1 n_0}\sum_{i=1}^{n} Z_i \boldsymbol{u}_i - \frac{n}{n_0}\begin{pmatrix}\bar{Y}(0)\\ \overline{\boldsymbol{X}}\end{pmatrix}$$

$$\overline{\boldsymbol{u}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{u}_i, \boldsymbol{S}^2_{\boldsymbol{u}} = (n-1)^{-1}\sum_{i=1}^{n}\left(\boldsymbol{u}_i - \overline{\boldsymbol{u}}\right)\left(\boldsymbol{u}_i - \overline{\boldsymbol{u}}\right)^\top$$

$$\gamma_n \equiv \frac{(K+1)^{1/4}}{\sqrt{n r_1 r_0}}\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{S}_{\boldsymbol{u}}^{-1}\left(\boldsymbol{u}_i - \overline{\boldsymbol{u}}\right)\right\|_2^3$$

$$\Delta_n \equiv \sup_{\mathcal{Q}\in\mathcal{C}_{K+1}}\left|\mathbb{P}\left\{V^{-1/2}\begin{pmatrix}\hat{\tau}-\tau\\ \hat{\boldsymbol{\tau}}_X\end{pmatrix}\in\mathcal{Q}\right\} - \mathbb{P}(\boldsymbol{\varepsilon}\in\mathcal{Q})\right|$$

$$R^2 = \mathrm{Corr}^2\left(\hat{\tau}, \hat{\boldsymbol{\tau}}_X\right) = \frac{\boldsymbol{V}_{\tau\boldsymbol{x}}\boldsymbol{V}_{\boldsymbol{x}\boldsymbol{x}}^{-1}\boldsymbol{V}_{\boldsymbol{x}\tau}}{V_{\tau\tau}} = \frac{n_1^{-1}S^2_{1|\boldsymbol{X}} + n_0^{-1}S^2_{0|\boldsymbol{X}} - n^{-1}S^2_{\tau|\boldsymbol{X}}}{n_1^{-1}S_1^2 + n_0^{-1}S_0^2 - n^{-1}S_\tau^2}$$

$$\begin{aligned}\mathrm{Cov}\begin{pmatrix}\hat{\tau}\\ \hat{\boldsymbol{\tau}}_{\boldsymbol{X}}\end{pmatrix} &= \begin{pmatrix} n_1^{-1}S_1^2 + n_0^{-1}S_0^2 - n^{-1}S_\tau^2 & n_1^{-1}\boldsymbol{S}_{1,\boldsymbol{X}} + n_0^{-1}\boldsymbol{S}_{0,\boldsymbol{X}} \\ n_1^{-1}\boldsymbol{S}_{\boldsymbol{X},1} + n_0^{-1}\boldsymbol{S}_{\boldsymbol{X},0} & n/\left(n_1 n_0\right)\cdot\boldsymbol{S}^2_{\boldsymbol{X}}\end{pmatrix}\\ &\equiv \boldsymbol{V} \equiv \begin{pmatrix} V_{\tau\tau} & \boldsymbol{V}_{\tau\boldsymbol{x}} \\ \boldsymbol{V}_{\boldsymbol{x}\tau} & \boldsymbol{V}_{\boldsymbol{x}\boldsymbol{x}}\end{pmatrix}\end{aligned}$$

$$p_n \equiv \mathbb{P}\left(\chi^2_{K_n} \le a_n\right)$$

2 文章脉络

这篇文章主要介绍重随机化试验设计，证明了重随机化可以渐进达到理论最优的试验设计，同时保持协变量均衡。

历史：在随机化试验设计领域，一直有两派争论，一种是完全随机化试验派，可以在期望平均意义下排除因果推断中的共因 cofounders 对于试验结果的影响，但被 Rubin 等人证明这种完全随机化试验的方法容易产生协变量不均衡的情况：

> With k independent covariates, the chance of *at least one* covariate showing a "significant difference" between treatment and control groups, at significance level α, is $1-(1-\alpha)^k$. For a modest 10 covariates and a 5% significance level, this probability is 40%. "Most experimenters on carrying out a random assignment

另一种是最优试验设计派，即采用协变量协方差最小的试验方案，这种方案理论上是协变量最均衡的方案，但仍有问题，那就是这种条件太强，往往最后筛选出的方案只有一个或者两个，从而无法对该试验设计做鲁棒性分析或者统计推断，更无法进行渐进估计。

本文则是在渐进角度提出了二者之间的折中方案，那就是重随机化。

1. Collect the covariate data for the experimental units, and specify a covariate balance criterion.
2. Randomly assign n_1 units to treatment group and the remaining n_0 units to control group.
3. Check the covariate balance for the treatment assignment from Step 2. If the balance criterion is satisfied, proceed to Step 4; otherwise, return to Step 2.
4. Conduct the experiment using the acceptable treatment assignment from Step 3.

需要注意的是，从上述步骤可以看出，重随机化其实就是多次完全随机化试验，只不过给每次随机化试验加了一个限定条件，满足相应条件的随机化试验才能被接受，但每次随机化都是完全随机化试验，因此当考虑每次随机化试验的过程中，Z_i 永远都满足第一部分的概率分布。此外，CRE 并不考虑协变量的影响，这也是 CRE 与重随机化的区别之一。本文考虑的限

定条件是利用 Mahalanobis 距离来反映试验组和控制组协变量之间的“距离”，让 Mahalanobis 距离小于一个阈值来实现均衡协变量的效果。

本文并不是第一个提出的，Rubin 等人已提出应该使用重随机化的方法，并得到了个体平均因果效应的概率分布

$$\sqrt{n}(\hat{\tau}-\tau) \mid M \leq a \dot{\sim} \sqrt{nV_{\tau\tau}}\left(\sqrt{1-R^2}\cdot\varepsilon + R\cdot L_{K,a}\right)$$

bution. In (3), $\varepsilon \sim \mathcal{N}(0,1)$ follows a standard Gaussian distribution, $L_{K,a} \sim D_1 \mid \boldsymbol{D}^\top \boldsymbol{D} \leq a$ follows a constrained Gaussian distribution with $\boldsymbol{D} = (D_1, D_2, \ldots, D_K)^\top \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_K)$, and ε and $L_{K,a}$ are mutually independent. Besides, $V_{\tau\tau}$ is the variance of $\hat{\tau}$ under the CRE, R^2 is the squared multiple correlation between $\hat{\tau}$ and $\hat{\boldsymbol{\tau}}_{\boldsymbol{X}}$ under the CRE, and we defer the explicit expressions for $V_{\tau\tau}$ and R^2 to Section 3.1. From (3), the difference-in-means estimator

这个结果考虑的是固定的协变量个数 K，固定的 Mahalanobis 距离 a. 而本文则考虑的是随 n 变化的 K_n，a_n 等，也即，当协变量个数随样本量变化的情况下，协变量个数在怎么样的一个数量级下，重随机化仍能获得比较好的效果，本文最终证明了在特定的数量级下，重随机化能够渐进逼近理论最优试验设计，同时结果还有一定的鲁棒性。

2.1 CRE 下的 Gaussian 逼近

本文首先利用 Berry-Esseen 界和配对 coupling 方法给出了 CRE 条件下 Δ_n 与 γ_n 的控制关系

THEOREM 1. *For any $n \geq 2$, $K \geq 0$, $r_1, r_0 \in (0,1)$, and any finite population $\Pi_n \equiv \{(Y_i(1), Y_i(0), \boldsymbol{X}_i) : i = 1, 2, \ldots, n\}$ with nonsingular $\boldsymbol{V}$ defined as in (4), define γ_n and Δ_n as in (7) and (8). Then*:

(i) *there exists an absolute constant C_K that depends only on K such that $\Delta_n \leq C_K \gamma_n$;*
(ii) *if the conjecture in Raič (2015) hold, then there exists a universal constant C such that $\Delta_n \leq C\gamma_n$;*
(iii) $\Delta_n \leq 174\gamma_n + 7\gamma_n^{1/3}$.

THEOREM 2. *Under the same setting as Theorem 1,*

$$\Delta_n \leq 180\gamma_n + \frac{3(\log n)^{3/4}(K+1)^{3/4}}{n^{1/4}\sqrt{r_1 r_0}} \cdot \max_{1\leq i\leq n} \left\| \boldsymbol{S}_{\boldsymbol{u}}^{-1}(\boldsymbol{u}_i - \bar{\boldsymbol{u}}) \right\|_\infty,$$

and, for any $\iota \geq 2$, there exists a universal constant C_ι depending only on ι such that

$$\Delta_n \leq 174\gamma_n + \frac{C_\iota (K+1)^{3\iota/\{4(\iota+1)\}}}{n^{\iota/\{4(\iota+1)\}}\{r_1 r_0\}^{\iota/2}} \cdot \frac{1}{n}\sum_{i=1}^{n} \left\| \boldsymbol{S}_{\boldsymbol{u}}^{-1}(\boldsymbol{u}_i - \bar{\boldsymbol{u}}) \right\|_\iota^\iota.$$

这里 Raic 猜想右边关于 n 的系数可以收紧为 $n^{-1/2}$，但本文的上述定理最多收紧到 $n^{-1/4}$，因此这里是未来有待完成的一项工作。

2.2 接受概率

考虑到 Mahalanobis 距离 M 在 CRE 下服从 $\chi^2_{K_n}$ 分布，因此自然地有如下控制：

$$p_n - \Delta_n \leq \mathbb{P}(M \leq a_n) \leq p_n + \Delta_n$$

于是在 $p_n \gg \Delta_n$ 的条件下我们便可以用 p_n 近似接受概率，进而 a_n 可看成由 K_n, p_n 决定的函数（即为 χ^2 分布的分位数）

CONDITION 1. As $n \to \infty$, the sequence of finite populations satisfies that $\gamma_n \to 0$.

CONDITION 2. As $n \to \infty$, $p_n/\Delta_n \to \infty$.

CONDITION 3. As $n \to \infty$, $\log(p_n^{-1})/K_n \to \infty$.

CONDITION 4. As $n \to \infty$, $\limsup_{n\to\infty} R_n^2 < 1$.

CONDITION 5. As $n \to \infty$,

$$\frac{\max_{z\in\{0,1\}} \max_{1\le i\le n}\{Y_i(z) - \bar{Y}(z)\}^2}{r_0 S^2_{1\backslash \boldsymbol{X}} + r_1 S^2_{0\backslash \boldsymbol{X}}} \cdot \frac{\max\{K_n, 1\}}{r_1 r_0} \cdot \sqrt{\frac{\max\{1, \log K_n, -\log p_n\}}{n}} \to 0. \tag{16}$$

CONDITION 6. For each sample size n, $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n$ are i.i.d. random vectors in $\mathbb{R}^{K_n+1}$ with finite and nonsingular covariance matrix. The standardized random vector $\boldsymbol{\xi}_i = \mathrm{Cov}(\boldsymbol{u}_i)^{-1/2}(\boldsymbol{u}_i - \mathbb{E}\boldsymbol{u}_i) \in \mathbb{R}^{K_n+1}$ satisfies that $\sup_{\boldsymbol{v}\in\mathbb{R}^{K_n+1}:\boldsymbol{v}^\top\boldsymbol{v}=1} \mathbb{E}|\boldsymbol{v}^\top \boldsymbol{\xi}_i|^\delta = O(1)$ for some $\delta > 2$, that is, there exists an absolute constant $C < \infty$ such that $\mathbb{E}|\boldsymbol{v}^\top \boldsymbol{\xi}_i|^\delta \le C$ for all n and all unit vector $\boldsymbol{v}$ in $\mathbb{R}^{K_n+1}$.

以上为本文中使用的条件，后续作者证明了这几种条件都会以高概率成立，而这几个条件是保证论文结论成立的几条关键性条件，进而结论最终将会以高概率成立，保证了结论的有效性。条件 1 保证了渐进分布逼近 Gaussian 分布，而条件 2 保证了 Mahalanobis 界 a_n 不能过小，过小会导致不精确的渐进估计和无效的随机试验。

THEOREM 3. *Under ReM and Conditions 1 and 2, as* $n \to \infty$,

$$\sup_{c\in\mathbb{R}}\left|\mathbb{P}\{V_{\tau\tau}^{-1/2}(\hat{\tau}-\tau)\le c \mid M \le a_n\} - \mathbb{P}\left(\sqrt{1-R_n^2}\varepsilon_0 + \sqrt{R_n^2}L_{K_n,a_n} \le c\right)\right| \to 0. \tag{10}$$

在条件 1、2 成立的情况下，ReM 下的平均因果效应的渐进分布为正态分布与条件正态分布的卷积和。上述定理是对现有结论的推广，因为式中引入了随样本量 n 变化的 Mahalanobis 界 a_n 和 R_n, K_n.

2.3 ReM 相较 CRE 的改进程度

对于 CRE，可以看成不考虑协变量以及 Mahalanobis 界为正无穷的 ReM，进而可以得到

$$\sup_{c\in\mathbb{R}}\left|\mathbb{P}\left\{V_{\tau\tau}^{-1/2}(\hat{\tau}-\tau)\le c\right\} - \mathbb{P}\left(\varepsilon_0 \le c\right)\right| \to 0$$

由此可以计算 ReM 相较于 CRE 的改进程度

COROLLARY 1. *Under ReM and Conditions 1 and 2, the asymptotic distribution of* $V_{\tau\tau}^{-1/2}(\hat{\tau}-\tau)$, $\sqrt{1-R_n^2}\varepsilon_0+\sqrt{R_n^2}L_{K_n,a_n}$, *as shown in* (10) *is symmetric and unimodal around zero. Compared to the asymptotic distribution in* (11) *under the CRE, the percentage reductions in asymptotic variance and length of asymptotic* $1-\alpha$ *symmetric quantile range for* $\alpha \in (0,1)$ *are, respectively,*

$$(1-v_{K_n,a_n})R_n^2 \quad \text{and} \quad 1-\frac{\nu_{1-\alpha/2,K_n,a_n}(R_n^2)}{z_{1-\alpha/2}}. \tag{12}$$

Both percentage reductions in (12) *are nonnegative and are uniquely determined by* (R_n^2, p_n, K_n), *and they are nondecreasing in* R_n^2 *and nonincreasing in* p_n *and* K_n.

其中 $v_{K,a} = \mathrm{Var}\left(L_{K,a}\right) = \mathbb{P}(\chi^2_{K+2} \le a)/\mathbb{P}(\chi^2_K \le a), \nu_{\alpha,K,a}(R^2)$ 为 $\sqrt{1-R^2}\varepsilon_0+$

$\sqrt{R^2}L_{K,a}$ 的分位数，z_α 为正态分布的分位数. 为了让 ReM 的优势更明显，自然希望渐进方差下降得更多，例如让 R_n^2 更大，但这不意味着我们可以使用尽可能多的协变量，因为随着 K_n 的增大，对于渐进方差的下降会产生副作用，因此我们的目标就是能否在让 ReM 相较于 CRE 的优势尽可能大的情况下，渐进条件下 $V_{\tau\tau}^{-1/2}(\hat{\tau}-\tau)$ 的分布能够逼近 Gaussian 分布，即条件 Gaussian 分布 L_{K_n,a_n} 是渐进可忽略的。因此本文继续研究 L_{K_n,a_n} 的渐进形态

2.4 条件限制的 Gaussian 分布

研究其方差 v_{K_n,a_n} 即可，该量刻画了 ReM 相较 CRE 的优势。

THEOREM 4. *As* $n \to \infty$:

(i) *if* $\log(p_n^{-1})/K_n \to \infty$, *then* $v_{K_n,a_n} \to 0$;
(ii) *if* $\limsup_{n\to\infty} \log(p_n^{-1})/K_n < \infty$, *then* $\liminf_{n\to\infty} v_{K_n,a_n} > 0$;
(iii) *if* $\liminf_{n\to\infty} \log(p_n^{-1})/K_n > 0$, *then* $\limsup_{n\to\infty} v_{K_n,a_n} < 1$;
(iv) *if* $\log(p_n^{-1})/K_n \to 0$, *then* $v_{K_n,a_n} \to 1$.

这里便需要做一个权衡，从上述定理可以看出我们最好用更小的 p_n 和更小的 K_n。然而 p_n 不能太小，因为会导致不精确的随机试验和渐进逼近；K_n 如果太小，还会导致 R_n 变小，从而降低相较于 CRE 的提升度。因此需要找到一个 “适当的” K_n 和 p_n。条件 3 则保证了条件限制 Gaussian 项是渐进可忽略的，而条件 4 则确保我们所需要的 Gaussian 项不会渐进可忽略，否则结论就不成立了。

THEOREM 5. *Under ReM and Conditions* 1–4,

$$\sup_{c\in\mathbb{R}} \left|\mathbb{P}\left\{V_{\tau\tau}^{-1/2}(\hat{\tau}-\tau) \le c \mid M \le a_n\right\} - \mathbb{P}\left(\sqrt{1-R_n^2}\varepsilon_0 \le c\right)\right| \to 0. \tag{13}$$

上述定理便是本文的核心观点，即在最优 ReM 下，是可以达到渐进最优试验的。接下来只需对所用到的条件进行逐一验证即可。

THEOREM 6. *Under ReM and Conditions 1 and 4:*

(i) *if and only if* $\log(\Delta_n^{-1})/K_n \to \infty$, *there exists a sequence* $\{p_n\}$ *such that both Conditions 2 and 3 hold, under which ReM achieves its ideally optimal precision and the asymptotic Gaussian approximation in (13) holds*;

(iii) *if* $\limsup_{n\to\infty} \log(\Delta_n^{-1})/K_n < \infty$, *then for any sequence* $\{p_n\}$ *satisfying Condition 2 such that the asymptotic approximation in (10) holds,* $\liminf_{n\to\infty} v_{K_n,a_n} > 0$;

(iii) *if* $\liminf_{n\to\infty} \log(\Delta_n^{-1})/K_n > 0$, *then there exists a sequence* $\{p_n\}$ *satisfying Condition 2 such that (10) holds and* $\limsup_{n\to\infty} v_{K_n,a_n} < 1$;

(iv) *if* $\log(\Delta_n^{-1})/K_n \to 0$, *then for any sequence* $\{p_n\}$ *satisfying Condition 2 such that (10) holds, the corresponding* $v_{K_n,a_n} \to 1$ *as* $n \to \infty$, *under which ReM asymptotically provides no gain on estimation precision compared to the CRE.*

2.5 大样本推断

$$s^2_{z\backslash \boldsymbol{X}} = s^2_z - s_{z,\boldsymbol{X}} \boldsymbol{S}^{-2}_{\boldsymbol{X}} \boldsymbol{s}_{\boldsymbol{X},z}$$

$$s^2_{\tau|\boldsymbol{X}} = (s_{1,\boldsymbol{X}} - s_{0,\boldsymbol{X}})\, S^{-2}_{\boldsymbol{X}}\, (s_{\boldsymbol{X},1} - s_{\boldsymbol{X},0})$$

$$\hat{V}_{\tau\tau} = n_1^{-1}s_1^2 + n_0^{-1}s_0^2 - n^{-1}s^2_{\tau|\boldsymbol{X}}, \quad \hat{R}^2_n = 1 - \hat{V}^{-1}_{\tau\tau}\left(n_1^{-1}s^2_{1\backslash\boldsymbol{X}} + n_0^{-1}s^2_{0\backslash\boldsymbol{X}}\right)$$

可以得到关于 τ 的置信区间

$$\hat{\mathcal{C}}_\alpha = \left[\hat{\tau} - \hat{V}^{1/2}_{\tau\tau} \cdot v_{1-\alpha/2,K_n,a_n}\left(\hat{R}^2_n\right), \quad \hat{\tau} + \hat{V}^{1/2}_{\tau\tau} \cdot v_{1-\alpha/2,K_n,a_n}\left(\hat{R}^2_n\right)\right]$$

THEOREM 7. *Under ReM and Conditions 1, 2 and 5, as* $n \to \infty$:

(i) *the estimators in (14) are asymptotically conservative in the sense that*

$$\max\{|\hat{V}_{\tau\tau}(1-\hat{R}^2_n) - V_{\tau\tau}(1-R^2_n) - S^2_{\tau\backslash\boldsymbol{X}}/n|, |\hat{V}_{\tau\tau}\hat{R}^2_n - V_{\tau\tau}R^2_n|\}$$
$$= o_{\mathbb{P}}(V_{\tau\tau}(1-R^2_n) + S^2_{\tau\backslash\boldsymbol{X}}/n);$$

(ii) *for any* $\alpha \in (0,1)$, *the resulting* $1-\alpha$ *confidence interval in (15) is asymptotically conservative, in the sense that* $\liminf_{n\to\infty} \mathbb{P}(\tau \in \hat{\mathcal{C}}_\alpha \mid M \le a_n) \ge 1-\alpha$;

(iii) *if further* $S^2_{\tau\backslash\boldsymbol{X}} = nV_{\tau\tau}(1-R^2_n)\cdot o(1)$, *the* $1-\alpha$ *confidence interval in (15) becomes asymptotically exact, in the sense that* $\lim_{n\to\infty} \mathbb{P}(\tau \in \hat{\mathcal{C}}_\alpha \mid M \le a_n) = 1-\alpha$.

如果条件 3 和条件 4 成立，那么置信区间可以得到简化，因为此时平均因果效应的边际分布即为 Gaussian 分布

$$\tilde{\mathcal{C}}_\alpha = \left[\hat{\tau} - \sqrt{\hat{V}_{\tau\tau}\left(1-\hat{R}^2_n\right)}\cdot z_{1-\alpha/2}, \quad \hat{\tau} + \sqrt{\hat{V}_{\tau\tau}\left(1-\hat{R}^2_n\right)}\cdot z_{1-\alpha/2}\right]$$

2.6 正则化条件

COROLLARY 2. *If Condition 6 holds, $r_z^{-1} = O(1)$ for $z = 0, 1$ and $K_n = o(n^{2/7-4/(7\delta)})$, then*:

(i) *$\gamma_n = o_{\mathbb{P}}(1)$, and thus Δ_n in (8) is of order $o_{\mathbb{P}}(1)$;*
(ii) *the finite population and superpopulation squared multiple correlations R_n^2 and $R_{\text{sup},n}^2$ are asymptotically equivalent, in the sense that $R_n^2 - R_{\text{sup},n}^2 = o_{\mathbb{P}}(1)$;*
(iii) *if further the standardized potential outcomes have bounded bth moments for some $b > 4$, both $\text{Var}(Y(1))$ and $\text{Var}(Y(0))$ are of the same order as $\text{Var}(r_0 Y(1) + r_1 Y(0))$, $\limsup_{n\to\infty} R_{\text{sup},n}^2 < 1$ and $K_n = O(n^c)$ and $-\log p_n = o(n^{1-4/b-2c})$ for some $c < 1/2 - 2/b$, then the quantity on the left-hand side of (16) is of order $o_{\mathbb{P}}(1)$.*

上述推论的（i）表明条件 1 以高概率成立；（ii）说明在一定条件下，条件 4 以高概率成立；（iii）则表明一定条件下，条件 5 以高概率成立。

COROLLARY 3. *If Condition 6 holds, $r_z^{-1} = O(1)$ for $z = 0, 1$, $R_{\text{sup},n}^2 \le 1 - c$ for some absolute constant $c > 0$, and $K_n = o(\log n)$, then there exists a sequence of acceptance probabilities p_n (or equivalently a sequence of thresholds a_n) such that, with probability converging to 1, the distribution of the difference-in-means estimator under ReM can be asymptotically approximated by a Gaussian distribution with mean zero and variance $V_{\tau\tau}(1 - R_n^2)$, that is,*

$$\sup_{c\in\mathbb{R}} \left| \mathbb{P}\{V_{\tau\tau}^{-1/2}(\hat{\tau} - \tau) \le c \mid M \le a_n, \mathcal{U}_n\} - \mathbb{P}(\sqrt{1 - R_n^2}\varepsilon_0 \le c \mid \mathcal{U}_n) \right| = o_{\mathbb{P}}(1).$$

K_n	γ_n	p_n	v_{K_n,a_n}
$K_n = o(\log n)$	$o_{\mathbb{P}}(\frac{(\log n)^{7/4}}{n^{1/2-1/\delta}})$	$p_n \asymp n^{-\beta}, \beta \in (0, \frac{\kappa}{2} - \frac{\kappa}{\delta})$	$o(1)$
$K_n \asymp \log n$	$O_{\mathbb{P}}(\frac{(\log n)^{7/4}}{n^{1/2-1/\delta}})$	$p_n \asymp n^{-\beta}, \beta \in (0, \frac{\kappa}{2} - \frac{\kappa}{\delta})$	$(0, 1)$
$K_n \asymp n^{\zeta}, \zeta \in (0, \frac{2}{7} - \frac{4}{7\delta})$	$O_{\mathbb{P}}(n^{-(\frac{1}{2}-\frac{1}{\delta}-\frac{7\zeta}{4})})$	$p_n \asymp n^{-\beta}, \beta \in (0, \frac{\kappa}{2} - \frac{\kappa}{\delta} - \frac{7\zeta\kappa}{4})$	$1 - o(1)$

2.7 数量级选取

关于 K_n：最多 $O(\log n)$，尽管比最优的试验方案 $o(\log n)$ 多一些，但已能够比 CRE 更优化。倘若更大则无法体现 ReM 的优势。

关于 γ_n：令

$$\tilde{\boldsymbol{X}} = (\boldsymbol{X}_1, \boldsymbol{X}_2 \ldots, \boldsymbol{X}_n)^\top \in \mathbb{R}^{n \times K_n}$$

$$\boldsymbol{H} = \tilde{\boldsymbol{X}} \left(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} \right)^{-1} \tilde{\boldsymbol{X}}^\top \in \mathbb{R}^{n \times n}$$

从而有

$$\gamma_n = \frac{(K_n+1)^{1/4}}{\sqrt{nr_1r_0}} \frac{1}{n} \sum_{i=1}^{n} \left(e_i^2 + (n-1)H_{ii} \right)^{3/2} \in \left[\frac{1}{4\sqrt{2}} \tilde{\gamma}_n, \sqrt{2}\tilde{\gamma}_n \right]$$

其中

$$\tilde{\gamma}_n = \frac{(K_n+1)^{1/4}}{\sqrt{r_1r_0n}} \frac{1}{n} \sum_{i=1}^{n} |e_i|^3 + \frac{(K_n+1)^{1/4}}{\sqrt{r_1r_0}} \sum_{i=1}^{n} H_{ii}^{3/2}$$

可以证明，$\tilde{\gamma}_n$ 的两项都能被很好地控制。

关于 p_n：选择使得 v_{K_n,a_n} 非常小（例如 0.01）的 p_n，检验其渐进逼近，实际运用中还需考虑计算复杂度，因为得到一个可接受的重随机化方案平均需要 $1/p_n$ 次随机化试验。

3 个人评价

本文主要干的事情实质上是改进了 Rubin 的研究成果，通过研究渐进形态并探寻渐进分布能否逼近 Gaussian 分布，进而通过渐进的方法寻找到了完全随机化和理论最优试验之间的折中方法：最优重随机化。

这也为今后研究提供了一个很好的思路，以最新研究成果为出发点，在基础上进行优化和改良，进而获得更好的结果。通过广泛阅读某一问题相关的文献，确认这一问题目前发展的最新进度，然后在最新进度的基础上在各方面尝试进行推广，例如固定的推广为渐进的，Mahalanobis 距离更换为其他度量方式等等。作者在最后一部分也提出了一些其他的思考：

In this paper, we mainly focused on the asymptotic properties of rerandomized treatment-control experiments using the Mahalanobis distance criterion; see the Supplementary Material (Wang and Li (2022)) for extension to regression adjustment under rerandomization. Beyond that, the derived theory, including both the finite population central limit theorem and the asymptotic behavior of the constrained Gaussian random variable, can also be useful for analyzing other covariate balance criteria, such as the Mahalanobis distance criterion with tiers of covariates (Li, Ding and Rubin (2018), Morgan and Rubin (2015)). It will also be interesting to extend the theory to rerandomization in more complex experiments, such as blocked experiments (Johansson and Schultzberg (2022), Wang, Wang and Liu (2021)), factorial experiments (Branson, Dasgupta and Rubin (2016), Li, Ding and Rubin (2020)), survey experiments (Yang, Qu and Li (2021)) and sequential experiments (Zhou et al. (2018)). Besides, we mainly considered finite population inference focusing on the average treatment effect of the experimental units in hand. It will be interesting to also consider superpopulation inference of some population average treatment effect when the units are randomly sampled from some superpopulation (Schultzberg and Johansson (2020)).