修士論文 要旨

検出力評価に基づく干渉を考慮したランダム化検定の改良

数理情報学専攻 48206238 柳 泉穂

指導教員 清 智也 教授

1 はじめに

統計的因果推論とは, 処置の与える効果について推論を行う枠組みである. 因果推論では通常, 各個体の結果変数は他の個体への処置に依存しないことを仮定するが, 現実的にはこの仮定が成り立たない, すなわち個体間に干渉が存在する状況が考えられる. 干渉の下で古典的な検定法である randomization test を実行するためには, 個体と割り当ての集合を適切に制限する必要がある [1, 2]. [3] はグラフからのアプローチにより適切な部分集合の選択法を与え, 汎用的な検定法である biclique test を提案した. しかし, その検出力は選択する部分集合の特徴に依存するため , その選択規則を洗練させることで検出力を向上させる余地が残されていた. 本研究では randomization test の検出力評価を行い, それに基づき bicliqe test を改良する手法を提案した.

2 既存研究

個体の集合を $\mathbb{U}=\{1,\dots,N\}$, 各個体への処置の割り当てを $z=(z_1,\dots,z_N)^\intercal\in\{-1,1\}^N(1:\text{処置},\ -1:\text{無処置})$, 割り当て z の下での潜在結果変数を $Y(z)=(Y_1(z),\dots,Y_N(z))^\intercal$ で表す. 割り当ては既知の確率分布 $P(z)$ から無作為に選ばれ, 実現し得る割り当ての集合を $\mathbb{Z}=\{z\in\{-1,1\}^N: P(z)>0\}$ と書く. 干渉のある設定では各個体に対し, 写像 $f_i(z)$ により割り当てが "曝露" に対応付けられる. これを用いて, 興味のある帰無仮説は次のように表現できる:

$$H_0^{\mathcal{F}}: Y_i(z)=Y_i(z') \text{ for any } (i,z,z')\in\mathbb{U}\times\mathbb{Z}^2 \\ \text{such that } f_i(z),f_i(z')\in\mathcal{F}.$$

2.1 randomization test

検定統計量を $T(z,Y)$ とする. 古典的な randomization test では, 帰無仮説 H_0 に対し

1. $Z^{\text{obs}}\sim P(z),\ Y^{\text{obs}}=Y(Z^{\text{obs}})$
2. $T^{\text{obs}}=T(Z^{\text{obs}},Y^{\text{obs}})$
3. $\text{pval}(Z^{\text{obs}})=\mathrm{E}_Z\left[\mathbf{1}\{T(Z,Y(Z))\geq T^{\text{obs}}\}\middle| H_0\right]$

の手続きにより有効な p 値を得る. ここで p 値の算出には, H_0 の下での観測されなかった任意の割り当て z に対する結果変数 $Y(z)$ の値が必要となる. いま H_0 の下で, 任意の z に対し $Y(z)=Y^{\text{obs}}$ が成り立つならば, $T(Z,Y(Z))=T(Z,Y^{\text{obs}})$ より観測された値のみから p 値を計算できる. このような条件を満たす帰無仮説 H_0 を sharp であると言い, randomization test の実行には H_0 が sharp であることが必要となる.

2.2 biclique test

通常, 干渉が存在する下での帰無仮説 $H_0^{\mathcal{F}}$ は sharp でないため, $H_0^{\mathcal{F}}$ が sharp となるような個体, 割り当ての部分集合に制限することで検定の実行を可能にする. [3] は, このような部分集合が, $H_0^{\mathcal{F}}$ から構成される二部グラフである null exposure graph $G_f^{\mathcal{F}}=(V,E)$,

$$V=\mathbb{U}\cup\mathbb{Z},\ E=\{(i,z)\in\mathbb{U}\times\mathbb{Z}\mid f_i(z)\in\mathcal{F}\}$$

の中の完全部分二部グラフ (biclique) $C=(U,\mathcal{Z})$ $(U\subseteq\mathbb{U},\mathcal{Z}\subseteq\mathbb{Z})$ に対応するという考察から, 適切な部分集合の選択を二部グラフ中の biclique の探索に帰着させ, 汎用的な検定法の biclique test を提案した. biclique test では予め null exposure graph を分割するような biclique の集合 $\mathcal{C}$ を構成し, それに応じて実現した割り当て Z^{obs} に対応する biclique を一意的に選択する. このような $\mathcal{C}$ を biclique decomposition と呼ぶ. 以上を踏まえ biclique test の手続きを以下に示す.

1. $Z^{\text{obs}}\sim P(z),\ Y^{\text{obs}}=Y(Z^{\text{obs}})$
2. $C=(U,\mathcal{Z})\in\mathcal{C}$ s.t. $Z^{\text{obs}}\in\mathcal{Z}$
3. $T^{\text{obs}}=T(Z^{\text{obs}},Y^{\text{obs}};C)$
4. $\text{pval}(Z^{\text{obs}};C)=\mathrm{E}_{Z\sim r}\left[\mathbf{1}\{T(Z,Y^{\text{obs}};C)\geq T^{\text{obs}}\}\right]$
   $(r(z)\propto P(z)\mathbf{1}\{z\in\mathcal{Z}(C)\})$

biclique decomposition $\mathcal{C}$ は次の貪欲なアルゴリズムにより得ることができる. ここで, $C\in G$ は C が二部グラフ G の biclique であることを表す.

Algorithm 1 biclique decomposition algorithm

```
C ← ∅, G ← G_f^F
while |Z(G)| > 0 do
    C* = argmax_{C∈G} |E(C)|
    E(G) ← E(G) \ E(C*)
    Z(G) ← Z(G) \ Z(C*)
    C ← C ∪ {C*}
end while
return C
```

3 提案手法

以下では $\mathcal{F}=\{\mathbf{a},\mathbf{b}\}$ と表される帰無仮説を考える．biclique test の検出力は個々の biclique における randomization test の検出力の平均となるため, 高い検出力を得るためには "望ましい"biclique に分解することが重要となる．そこで, randomization test の検出力評価を通して検出力を特徴づける要素を明らかにし, それに基づいて bicliqe decomposition を構成することで biclique test を改善することを考える．

[4] は処置群と対照群のサイズが等しい場合において randomization test の検出力を評価したが, その結果を拡張したより一般的な設定の下で検出力の評価を行う．全個体が共通の処置効果 τ を持つモデル

$$Y(z)=\frac{\tau}{2}z+Y_0 \ \left(\tau\in\mathbb{R},\ Y_0\in\mathbb{R}^N\right)$$

を仮定し, 帰無仮説を $\tau=0$ とする．割り当ては $\mathbb{Z}=\{z^{(1)},\ldots,z^{(m)}\}$ から一様ランダムに選ばれるとする．加えて, 次の 3 つの仮定を置く．

仮定 1. (a) $Y_{0,i}\sim N(\mu,\sigma^2)$ *i.i.d.* $(i=1,\ldots,N)$
(b) $\frac{|\{i|z_i^{(k)}=1\}|}{N}=p\in(0,1) \quad (k=1,\ldots,m)$
(c) $\frac{\widetilde{z^{(k)}}\cdot z^{(l)}}{2}=\rho\in[0,1) \quad (k\neq l)$

ここで, $\tilde{z}$ は z を次のように変換したものを指す:

$$\tilde{z}_i=\begin{cases}|\{j|z_j=1\}|^{-1} & (z_i=1)\\ -|\{j|z_j=-1\}|^{-1} & (z_i=-1).\end{cases}$$

以上の設定の下, 次の評価を得る．

定理 2. 仮定 *1* の下で, 検定統計量として両群の平均の差を用いたときの有意水準 α の *randomization test* の平均的な検出力は次の表式で表される*:*

$$\int F_{\mathrm{bin}}\left(\lfloor m\alpha\rfloor-1;\ m-1,\ \Phi(z-\Theta)\right)\phi(z)dz.$$

ただし, $\Theta=\frac{\tau}{\sigma}\sqrt{N\cdot p(1-p)\cdot(1-\rho)}$, $F_{\mathrm{bin}}(k;n,p)$ は二項分布の累積分布関数, $\Phi(z),\phi(z)$ はそれぞれ標準正規分布の累積分布関数, 確率密度関数である．

得られた表式を $\mathcal{P}(\Theta,m;\alpha)$ $(\Theta\in\mathbb{R},\ m\in\mathbb{N},\ \alpha\in(0,1))$ と表すと, これは以下の性質を満たす．

命題 3. (a) $\Theta=0$ のとき, $\mathcal{P}(\Theta,m;\alpha)\leq\alpha$．特に, $\lfloor m\alpha\rfloor=m\alpha$ のとき等号が成立．
(b) $\mathcal{P}(\Theta,m;\alpha)$ は Θ についての狭義単調増加関数．
(c) $\Theta>0$ とする．m が $\lfloor m\alpha\rfloor=m\alpha$ を満たす整数であるとき, $\mathcal{P}(\Theta,m;\alpha)$ はこのような m についての狭義単調増加関数.

(a) は検出力の表式が定めた有意水準を満たすこと, (b), (c) はそれぞれ, Θ,m が大きいほど検出力が高くなることを示唆する．実際には所与の割り当て集合 $\mathbb{Z}$ が仮定 1 を満たすとは限らないが, その場合でも $\mathbb{Z}$ から計算される Θ の自然な代替量 $\hat{\Theta}_0$ が検出力をよく表現することを数値的に確認できる．これより, より直接的に望ましい biclique を選択するよう, Algorithm 1 の 3 行目において $|E(C)|$ を $\hat{\Theta}_0(C)$ に修正したものを提案手法とする．

4 数値実験

空間的な干渉が存在する設定において, 異なるパラメータの設定の下で既存手法と提案手法の検出力を比較した結果を図 1 に示す．いずれの設定においても提案手法は既存手法の検出力を上回ることが確認できる．ただし, 検出力の差異はパラメータの設定によって異なっており, これは分解を行う null exposure graph の構造の違いに由来すると考えられる．提案手法はグラフが疎で曝露の偏りが大きい, 既存手法が検出力を大きく落とす設定においても比較的高い検出力を保つ．

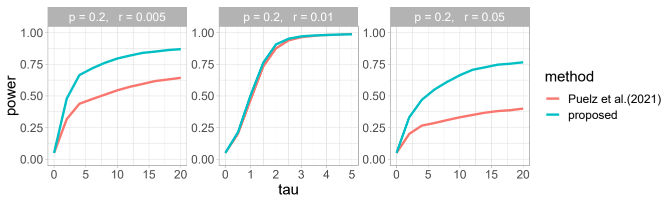

図 1. 既存手法と提案手法の検出力の比較

参考文献

[1] G. W. Basse, A. Feller, and P. Toulis. Randomization tests of causal effects under interference. *Biometrika*, Vol. 106, No. 2, pp. 487–494, 2019.

[2] S. Athey, D. Eckles, and G. W. Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, Vol. 113, No. 521, pp. 230–240, 2018.

[3] D. Puelz, G. Basse, A. Feller, and P. Toulis. A graph-theoretic approach to randomization tests of causal effects under general interference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pp. 1–32, 2021. Advance online publication. https://doi.org/10.1111/rssb.12478.

[4] A. M. Krieger, D. Azriel, M. Sklar, and A. Kapelner. Improving the power of the randomization test. *arXiv: 2008.05980*, 2020.