傾向スコアを用いた因果推論のためのモデル選択基準

九州大学 大学院数理学府 馬場 崇充
九州大学 マス・フォア・インダストリ研究所 二宮 嘉行

1 はじめに

因果推論で重用される周辺構造モデルは潜在的な結果変数を用いた反実仮想モデルであり，推定は傾向スコアに基づくことが一般的である．そして，IPW（inverse probability weighted）推定量（Rubin 1985 BayesStat）やその改良版である DR（doubly robust）推定量（例えば Bang & Robins 2005 Biometrics）が脚光を浴びてきた．一方，この設定においてもモデル選択は不可欠な作業であるにもかかわらず，そしてこの設定では AIC や C_p の形式的な利用は妥当でないにもかかわらず，情報量規準の開発は十分でない．例えば Platt et al.（2013 StatMed）ではある周辺構造モデルに対して QIC という情報量規準を提案しているが，やはり理論的妥当性がない．そこで，一般的な周辺構造モデルに対し，QIC の基となっているリスクに対応する平均二乗誤差を考え，その漸近不偏推定量として C_p 型の情報量規準を導出する．当日は，導出した漸近 C_p 基準を Platt et al.（2013 StatMed）のモデルに対して書き下すと QIC とは大きく異なってくることを確認するとともに，異なる種の平均二乗誤差を考えてそれに対応する漸近 C_p 基準も導く．

2 モデル

Brookhart & van der Laan（2006 CSDA）や Platt et al.（2013 StatMed）の周辺構造モデルを含む形の

$$\boldsymbol{y} = \sum_{h=1}^{H} t^{(h)} \boldsymbol{y}^{(h)} = \sum_{h=1}^{H} t^{(h)} \left(\boldsymbol{X}^{(h)} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \right)$$

というモデルを考える．ここで，$\boldsymbol{y}^{(h)}$ $(\in \mathbb{R}^r)$ は従属変数ベクトル，$\boldsymbol{X}^{(h)}$ $(\in \mathbb{R}^{r \times p})$ はランダムな共変量ベクトル $\boldsymbol{z}$ $(\in \mathbb{R}^q)$ の一部を含んでもよい独立変数行列とする．また，式における表現の煩雑さを軽減するため，本質的ではないがこの独立変数は $\mathrm{E}[\sum_{h=1}^{H} \boldsymbol{X}^{(h)t} \boldsymbol{X}^{(h)}]$ が p 次元単位行列 $\boldsymbol{I}_p$ となるように基準化されているものとする．そして，$\boldsymbol{\varepsilon}$ は期待値 $\boldsymbol{0}$ の誤差変数ベクトルであり，通常の回帰と同様に $\boldsymbol{X}^{(h)}$ と $\boldsymbol{\varepsilon}$ は独立であると仮定する．また，H 個の群を考え，$t^{(h)}$ は $\boldsymbol{y}^{(h)}$ が観測されると 1，観測されないと 0 となる指示変数とする．$\boldsymbol{z}$ を条件付けたもとでこの $t^{(h)}$ が 1 となる確率, $\mathrm{P}(t^{(h)} = 1 \mid \boldsymbol{z})$，がいわゆる一般化傾向スコア $e^{(h)}(\boldsymbol{z})$ である（Imbens 2000 Biometrika）．このモデルに対して N 個の独立なサンプルがあるとし，第 i サンプルの変数には添え字 i を付けることにする．

3 主結果

傾向スコア解析において通常仮定される（弱く）無視できる割り当て条件 $y^{(h)} \perp\!\!\!\perp t^{(h)} \mid \boldsymbol{z}$ があれば，漸近 C_p 基準を導出することができ，それを基に周辺構造モデルのモデル選択を行うことができる．

定理 1 傾向スコアが既知のとき，IPW 推定量 $\hat{\boldsymbol{\beta}}^{\mathrm{IPW}}$ を用いたときの漸近 C_p 基準は次で与えられる:

$$\sum_{h=1}^{H} \left\{ \sum_{i=1}^{N} \frac{t_i^{(h)}}{e_i^{(h)}(\boldsymbol{z})} \left(\boldsymbol{y}_i - \boldsymbol{X}_i^{(h)} \hat{\boldsymbol{\beta}}^{\mathrm{IPW}} \right)^t \left(\boldsymbol{y}_i - \boldsymbol{X}_i^{(h)} \hat{\boldsymbol{\beta}}^{\mathrm{IPW}} \right) \right\} + 2 \sum_{h=1}^{H} \mathrm{E} \left[\frac{1}{e^{(h)}(\boldsymbol{z})} \boldsymbol{\varepsilon}^t \boldsymbol{X}^{(h)} \boldsymbol{X}^{(h)t} \boldsymbol{\varepsilon} \right].$$

傾向スコアが未知のときも IPW 推定量を用いたときの漸近 C_p 基準を導くことができるが，そのときは DR 推定量を用いる傾向が強い．

定理 2 傾向スコアが未知のとき，DR 推定量 $\hat{\boldsymbol{\beta}}^{\mathrm{DR}}$ を用いたときの漸近 C_p 基準は，上の $\hat{\boldsymbol{\beta}}^{\mathrm{IPW}}$ を $\hat{\boldsymbol{\beta}}^{\mathrm{DR}}$ で，$e_i^{(h)}(\boldsymbol{z})$ を最尤推定量 $\hat{e}_i^{(h)}(\boldsymbol{z})$ で置き換えたものに次の項を加えたもので与えられる:

$$2 \sum_{h=1}^{H} \mathrm{E} \left[\left(1 - \frac{1}{e^{(h)}(\boldsymbol{z})} \right) \mathrm{E}[\boldsymbol{\varepsilon}^t \mid \boldsymbol{z}] \boldsymbol{X}^{(h)} \boldsymbol{X}^{(h)t} \mathrm{E}[\boldsymbol{\varepsilon} \mid \boldsymbol{z}] \right] + 2 \sum_{h' \neq h} \mathrm{E} \left[\mathrm{E}[\boldsymbol{\varepsilon}^t \mid \boldsymbol{z}] \boldsymbol{X}^{(h')} \boldsymbol{X}^{(h)t} \mathrm{E}[\boldsymbol{\varepsilon} \mid \boldsymbol{z}] \right].$$

これらに現れる期待値はその経験版に，$e^{(h)}(\boldsymbol{z})$ は $\hat{e}^{(h)}(\boldsymbol{z})$ に置き換えることで簡単に評価できる．