線形重回帰で効果がマスクされている変数の効率的探索

川崎　能典　モデリング研究系　教授

1. シミュレーションデザイン

想定している問題を例示する設定から説明する．確率変数ベクトル $z_i = (z_i^{(1)}, z_i^{(2)}, z_i^{(3)})$, $i = 1, \ldots, n$ を生成する．ここで各 $z_i^{(k)}$ $(k = 1, 2, 3)$ は各々 p_k 個の要素 $(z_{i,1}^{(k)}, \ldots, z_{i,p_k}^{(k)})$ から成るものとする．

まず $(z_{i,1}^{(1)}, \ldots, z_{i,p_1}^{(1)})$ は互いに独立な p_1 個の $[0, 1]$ 上の一様分布に従うものとする．次に $(z_{i,1}^{(2)}, \ldots, z_{i,p_2}^{(2)})$ は互いに独立な p_2 次元多変量正規分布であり，その平均はゼロ，共分散 Σ_2 は対角成分が 1 で非対角成分は 0.5 とする．最後に $(z_{i,1}^{(3)}, \ldots, z_{i,p_3}^{(3)})$ は互いに独立な p_3 次元多変量正規分布であり，その平均はゼロ，共分散 Σ_3 は対角成分が 1 で非対角成分は $0.8^{|j-k|}$ $(j, k = 1, \ldots, p_3)$ とする．

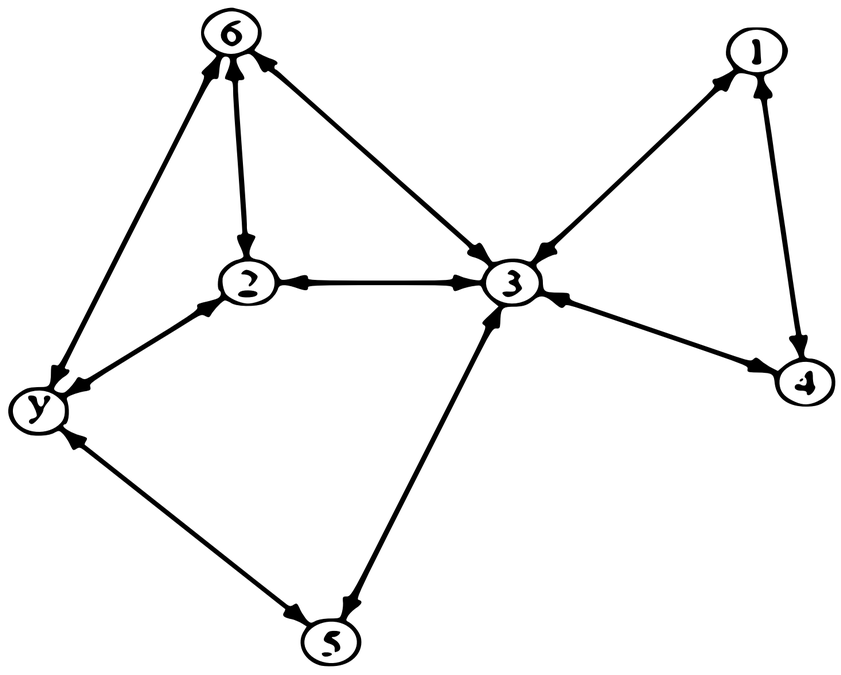
応答変数 y_i は，ϵ_i を独立な標準正規確率変数として，

$$y_i = z_{i,1}^{(1)} + z_{i,1}^{(2)} + z_{i,1}^{(3)} + \epsilon_i$$

として生成される．ここで想定する重要な設定は，我々は $z_{i,1}^{(1)}, z_{i,1}^{(2)}, z_{i,1}^{(3)}$ は観測できないということである．その代わりに，z_i の要素から生成される，以下の $x_{i,j}$ $(j = 1, \ldots, p_1 + p_2 + p_3)$ が観測できるとする．ただし以下の定義で ϵ_i' は ϵ_i とは独立な標準正規確率変数であり，B_i は独立なベルヌーイ確率変数で 1 を取る確率は 0.2 とする．

$$\begin{aligned}
x_{i,1} &= z_{i,1}^{(1)} - z_{i,2}^{(1)} \\
x_{i,2} &= z_{i,3}^{(1)} \\
x_{i,3} &= z_{i,3}^{(1)} - z_{i,1}^{(1)} \\
x_{i,4} &= z_{i,4}^{(1)} + \epsilon_i' + x_{i,1} \\
x_{i,5} &= z_{i,5}^{(1)} + z_{i,1}^{(1)} + z_{i,2}^{(1)} \\
x_{i,6} &= z_{i,6}^{(1)} + 0.3B_i + x_{i,2} \\
x_{i,j} &= z_{i,j}^{(1)} \quad \text{for} \quad j = 7, \ldots, p_1 \\
x_{i,j+p_1} &= z_{i,j}^{(2)} \quad \text{for} \quad j = 1, \ldots, p_2 \\
x_{i,j+p_1+p_2} &= z_{i,j}^{(3)} \quad \text{for} \quad j = 1, \ldots, p_3
\end{aligned}$$

応答変数 y_i に対して，$p_1 + p_2 + p_3 \equiv p$ 個の説明変数 $x_i = (x_{i,1}, \ldots, x_{i,p})$ が用意されている．構成上，最初の 6 個の変数 $x_{i,1}, \ldots, x_{i,6}$ だけが真の回帰関数に含まれる変数 $z_{i,1}^{(1)}, z_{i,2}^{(1)}, z_{i,3}^{(1)}$ を含む．変数間のグラフ関係は以下の図のようになっている．



以降，$p_1 = 400$, $p_2 = p_3 = 100$ とし，n に関しては $n = 300$ と $n = 500$ の 2 つの設定を用意する．我々の目的は，観測データ $x_i = (x_{i,1}, \ldots, x_{i,p})$ の中から，真の因果変数 $z_{i,1}^{(1)}, z_{i,2}^{(1)}, z_{i,3}^{(1)}$ と関わりのある $x_{i,1}, \ldots, x_{i,6}$ を漏らさず，しかし過不足なく拾い上げることにある．上記の設定から，デザイン行列は既に横長で，普通に回帰分析を行おうと思うと正則化が必要な状況であることに注意．

2. シミュレーション結果

手法の説明は後回しにして，以下の表に我々の提案手法(表中UKT列)とElastic Net(表中Enet)が，繰り返し数1,000回としてどれだけの頻度で $x_{i,1}, \ldots, x_{i,6}$ を正しく拾えたかをまとめてある．

変数	n	UKT(%)	Enet(%)
1	300	55	3
	500	99	2
2	300	55	97
	500	99	100
3	300	55	1
	500	99	1
4	300	49	4
	500	99	3
5	300	55	100
	500	99	100
6	300	55	17
	500	99	17

Elastic net は，応答変数 y_i と周辺相関を有する，グラフ上近い変数しか拾えていないことがわかる．これに対して提案手法では，$p = 600$ に対して $n = 500$ もあれば，真の因果変数を含む観測値を確実に拾い上げていることが見て取れる．

提案手法は偽陽性率の観点からも優れている．$x_{i,1}, \ldots, x_{i,6}$ 以外に，一つでも無関係な変数を拾ってきたケースを数え上げると，提案手法では $n = 300$ でも 0% である! これに対し Elastic net が無関係な変数を一つ以上拾ってきたケースは実験回数の 95% で，平均して17個の無関係な変数を選択した．

3. 提案手法の概要

実は我々の提案手法では一切回帰分析をしない．背後の理論はともかく，手順だけを記せば以下の通りである．

1. 説明変数 $x_j = (x_{1,j}, \ldots, x_{n,j})^T$ $(j = 1, \ldots p)$ どうしの，また各 x_j と $y = (y_1, \ldots, y_n)^T$ の相関係数の検定を行う．
2. 検定結果に基づいて，先の例に描いたグラフのように，各変数間の「接続性」がわかる．ここである説明変数 x_j から応答変数 y までの最短経路をダイクストラのアルゴリズムで求める．
3. ダイクストラアルゴリズムで得られるパスは冗長かもしれないが，x_j と y の経路上の変数 $(\hat{A})$ は十分絞り込めているはずであるので，$\mathrm{cor}(y, Q_C x_j)$ が最大となるような，$\hat{A}$ の部分集合 C を探索する．(Q_C は変数群 C を説明変数としたときの射影行列．)
4. こうして，y への経路が存在する x_j を拾い上げていく．

左段に例示したグラフから類推できる通り，y から遠い変数は x_j は $\mathrm{cor}(y, x_j) \approx 0$，すなわち y との周辺相関を検出できないので，実は y と関連しているにも関わらず通常の手法では捨てられてしまう．つまり x_j の効果はマスクされているわけだが，間に適切な変数集合 A を挟むことで応答変数 y に対する x_j の効果が浮かび上がってくる．すなわち，$\mathrm{cor}(y, Q_A x_j) \neq 0$ である．

$\mathrm{cor}(y, Q_A x_j)$ が最大となるような変数セット A を力業で探索するのは大変である．そこで，ペアワイズに調べた変数の接続性を元に，ダイクストラアルゴリズムの助けを借りて候補変数を十分に絞り込んだところで，仕上げの変数探索を行うのである．ゲノム解析への応用が，Ueki, Kawasaki and Tamiya (2017) に見られる．本研究は植木優夫博士(理化学研究所)，田宮元教授(東北大学)との共同研究であり，一部は統計数理研究所共同研究(2015-共研-1013)に基づくものである．

参考文献 Ueki, M., Kawasaki, Y. and Tamiya, G. (2017). Detecting genetic association through shortest paths in a bidirected graph, *Genetic Epidemiology*, Vol. 41, 481–497. DOI: 10.1002/gepi.22051