2018 年 2 月 8 日

多次元セミパラメトリック階層混合モデルと optimal discovery 法を用いた関連遺伝子のスクリーニング

オミクスデータ（遺伝子多型や遺伝子発現のデータなど）と興味のある臨床変数の関連解析では，サンプルのサブグループを規定する重要な変数が別に存在することが多い．これらの変数は共変量として回帰モデルに取り込まれることが多いが，サブグループの数が比較的少ないときにはサブグループ別のオミクスデータと臨床変数の関連統計量の同時分布をモデリングするアプローチが考えられる[1].

例として，多発性骨髄腫に対して標準的な化学療法のレジメン（コントロール群）とサリドマイドを併用するレジメン（サリドマイド群）を比較するランダム化臨床試験を考える [2]．この試験では，治療前に採取されたがん形質細胞に対してマイクロアレー遺伝子発現解析が行われ，約 5 万 5 千個のプローブセットの発現量が測定されている．遺伝子発現データと主要エンドポイントである生存期間 (overall survival) との関連解析として標準的なものは，遺伝子，治療，これらの交互作用項を共変量とした Cox 回帰モデル解析であろう．遺伝子別にこのモデルをあてはめて遺伝子の主効果，治療との交互作用項の検定を行ってみると，FDR = 5%に対して，前者では 793 個の遺伝子が有意となったが，後者では有意な遺伝子は一つもなかった．

一方，提案するアプローチでは治療群別に遺伝子のみを共変量とした Cox モデルを当てはめる．具体的には，遺伝子 j に対して発現量が標準偏差分だけ増加したときの対数ハザード比の推定値 $\mathbf{b}_j = (b_j{}^{(0)}, b_j{}^{(1)})$ を用いる（それぞれコントロール群，サリドマイド群での推定値）．これを全プローブセットでプロットすると図 1 が得られる．

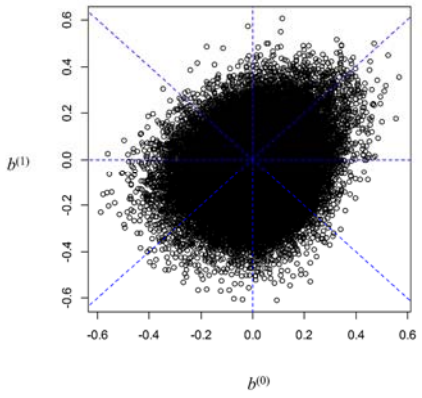


図 1．全遺伝子の $\mathbf{b}_j = (b_j{}^{(0)}, b_j{}^{(1)})$のプロット

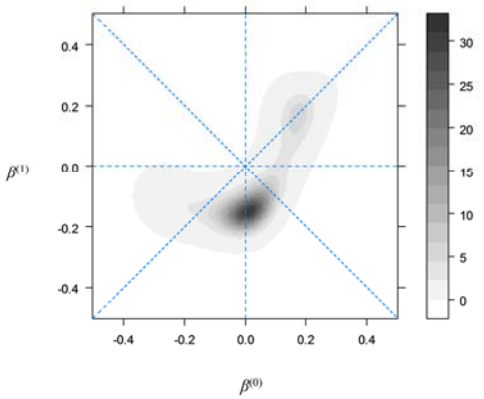
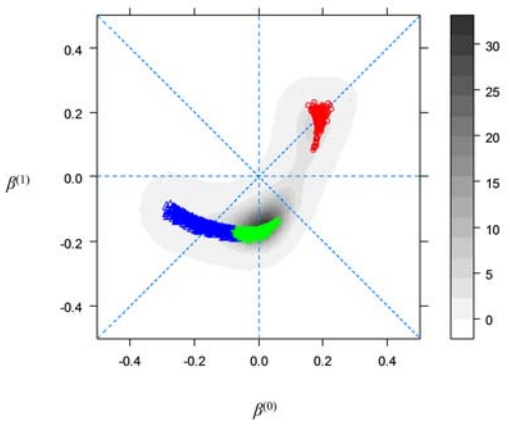


図 2．階層混合モデルにより推定された効果サイズ分布

このプロットには，ノイズ成分として生存時間と関連のない遺伝子と（母集団からの）サンプリング誤差が含まれる．これらをセミパラメトリック階層混合モデル[1]を用いて取り除くと，シグナル成分としての関連遺伝子の割合，関連ありの遺伝子の効果サイズ分布が得られる（図 2）．大変興味深いことに，真の対数ハザード比 $\mathbf{b}_j = (b_j^{(0)}, b_j^{(1)}) = (0, -0.2)$ 周辺に大きなピークがみられる．これはサリドマイドを受けることで死亡リスクが下がることを意味し，このような関連プロファイルをもつ遺伝子（治療効果予測遺伝子）が多く存在することを示唆している．

一旦背後にある分布をうまく推定できれば，最強力検定としての optimal discovery 検定統計量 [3-4] をモデルベースに構成でき，遺伝子ランキングや FDR の推定などを有効に行える．今の例では，FDR = 5%に対して 1,615 個の遺伝子が有意となった．有意となった遺伝子の効果サイズ推定（縮小推定）もモデルベースに可能であり，図 3 にプロットした．解釈が容易となるよう三つのクラスターに分けているが，効果サイズ分布のピークに近いクラスター（緑色）は治療効果関連遺伝子のクラスターと考えられる．これには器官形成をはじめとする発生関連の遺伝子が多く含まれていた．



従来の回帰モデルによる治療と遺伝子の交互作用検定では一つの遺伝子も検出できなかったが，図 2 の効果サイズ分布をみると，交互作用検定が主に想定している関連プロファイル $\beta^{(1)} = -\beta^{(0)}$ の方向にはあまり関連遺伝子が存在していないことがわかる．これに対して，提案法は，（事前に想定することが難しい）多次元空間のもとでの効果サイズ分布をノンパラメトリックに捉え，その効果サイズ分布に基づいて最強力検定を行うことで効率的なスクリーニング解析を可能にするものと考えることができる．

文献：

1. Matsui S, Noma H, Qu P, Yoshio Sakai, Matsui K, Heuck C, Crowley J. (2018). Multi-subgroup gene screening using semi-parametric hierarchical mixture models and the optimal discovery procedure: application to a randomized clinical trial in multiple myeloma. Biometrics 74,

313-320.

2. Barlogie B, Anaissie E, van Rhee F, Shaughnessy J, Szymonifka J, Hoering A, Petty N, Crowley J. (2010). Reiterative survival analyses of total therapy 2 for multiple myeloma elucidate follow-up time dependency of prognostic variables and treatment arms. Journal of Clinical Oncology 28, 3023-3027.
3. Storey JD. (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. Journal of the Royal Statistical Society, Series B 69, 347-368.
4. Noma H, Matsui S. (2012). The optimal discovery procedure in multiple significance testing: an empirical Bayes approach. Statistics in Medicine, 31, 165-176.