依存打ち切りを考慮した生存木構築について

東京理科大学　下川 朝有
東京理科大学　宮岡 悦良

はじめに

本研究では生存時間と打ち切り時間の間の依存性を考慮した，木構造モデルの構築を扱う．観測時間及び打ち切り指標という一般的な生存データが与えられた下，生存時間と打ち切り時間の依存は，追加情報無しに同定不可能であることは広く知られている．そこで本研究では，共変量が与えられた下での，生存時間及び打ち切り時間の結合分布に対してコピュラを用いてモデル化を行う．これらの仮定の下，生存木構築のための手法を提案し，シミュレーション及び実データへの適用を通してその性能を比較する．

記述・モデル

U は死亡時間，C は打ち切り時間，そして $X = \min(U, C)$ は観測時間を表す確率変数とする．また $\delta = I(X = U)$ はイベント指標を表し，$\boldsymbol{Z} = (Z_1, \cdots, Z_q)$ は q 次元共変量ベクトル，また対応する共変量空間を $\mathcal{Z}$ を用いて表すとする．このとき観測データは $\mathscr{L} = \{(x_i, \delta_i, \boldsymbol{z}_i); i = 1, \cdots, n\}$ で与えられる．

$S_i(u) = \Pr(U > u | \boldsymbol{Z} = \boldsymbol{z}_i)$ を被験者 i の条件付き生存確率としたとき，以下の形で与えられるモデルの構築を考える：

$$S_i(u) = S(u; \boldsymbol{\mu}_k, \boldsymbol{\eta}_k), \quad \boldsymbol{z}_i \in t_k,$$

ただし $t_k \subseteq \mathcal{Z}$ 共変量空間の分割による部分空間を表す（$k = 1, 2, \cdots, K$）．$\boldsymbol{\mu}_k$ は関心のある p 次元パラメータベクトルであり，$\boldsymbol{\eta}_k$ はその他のパラメータベクトルを表すと仮定する．

ここで U と C の依存をモデル化するため，コピュラを用いたモデルを仮定する．すなわち，$F_i(u)$ 及び $G_i(c)$ をそれぞれ U と C に関する $\boldsymbol{z}_i$ を持つ被験者の累積分布関数を表すとし，その結合累積分布関数は以下で与えられるとする：

$$\Pr(U \le u, C \le c | \boldsymbol{Z} = \boldsymbol{z}_i) = H_i\{F_i(u), G_i(c); \alpha_i\},$$

ただし $H_i\{\cdot, \cdot; \alpha_i\}$ は依存度を表すパラメータ α_i を持つ，2 次元コピュラ関数を表す．この仮定の下，K を固定したとき，U と C に関する尤度は以下で与えられる：

$$L = \prod_k \prod_{i \in t_k} \left\{ f_i(x_i) - \frac{\partial}{\partial u} H_i\{F_i(u), G_i(x_i); \alpha_i\}\Big|_{u=x_i} \right\}^{\delta_i} \left\{ g_i(x_i) - \frac{\partial}{\partial c} H_i\{F_i(x_i), G_i(c); \alpha_i\}\Big|_{c=x_i} \right\}^{1-\delta_i},$$

ここで $i \in t_k$ はその共変量の値が t_k に含まれる被験者の指標の集合を表し，また $f_i(u) = \frac{d}{du} F_i(u)$，$g_i(c) = \frac{d}{dc} G_i(c)$ である．

木構造モデル

木構造モデルは共変量空間の分割ルールと，その結果得られる空間の部分集合（ノード）により成る．木構造を T，ノードを t で表すとする．ノード t に対する分割ルールは "$\boldsymbol{Z} \in t_L$?" の形で与えられる．ここで $t_L \subset \mathcal{Z}$ 及び，$t_R = t - t_L$ は t の子ノードと呼ばれる．ノード t がある分割ルールにより t_L 及び t_R に分割されるとき，ノード t に含まれる標本 $\mathscr{L}_t$ による尤度は，以下の形で与えられる：

$$L_t = L_{t_L}(\boldsymbol{\mu}_{t_L}, \boldsymbol{\eta}_{t_L}, \boldsymbol{\theta}_{t_L}, \alpha_{t_L}) \times L_{t_R}(\boldsymbol{\mu}_{t_R}, \boldsymbol{\eta}_{t_R}, \boldsymbol{\theta}_{t_R}, \alpha_{t_R}),$$

ただし $\boldsymbol{\theta}$ は C に関するモデル内のパラメータを表し，またコピュラに関するパラメータ α_i は各ノード内の被験者間で共通と仮定している．

$\mathscr{L}$ を用いて，$\mathcal{Z}$ を再帰的に分割することにより T は構築されるが，その際各ノードにおいて最適な分割ルールを決定するため，分割基準を定める必要がある．そこで本研究では複合仮説 $H_0 : \boldsymbol{\xi}_{t_L} = \boldsymbol{\xi}_{t_R}$ に対する検定統計量を用いた場合について調べる．加えて，t_L 及び t_R 内の原因別ハザードに関わるパラメータ間の同等性に対する検定統計量についても検討する．