計算機統計学
第22巻・第1号：2009
3 - 21

# データ適応型分布に基づく多分岐樹木構造接近法

下川敏雄* ・後藤昌司**

## 要 旨

生存時間研究において，樹木構造接近法は患者の生存時間の予後因子を評価するための強力な手段である．既存の生存時間研究における樹木構造接近法のほとんどが2分岐で構成されているが，応答(生存時間)に対する予後因子を2分岐で単純に扱える場合は少ない．衛藤 他(2007)は，予後因子の交互作用効果を(比較的)簡単なグラフィクスで表示できる多分岐型樹木構造接近法を提案している．この方法は，分岐基準にノンパラメトリック(検定)統計量を用いて構成されているが，データ省察から解釈までの手順に論理的な一貫性を保持するには，応答に何らかの潜在基礎分布を想定したほうが好ましい．本論文では，その潜在基礎分布にデータ適応型分布，それもベキ正規分布を想定したもとで多分岐樹木構造接近法を構成する．

その性能を，数個の文献事例およびシミュレーションにより評価した．その結果，既存のノンパラメトリック検定統計量に基づく多分岐樹木構造接近法と比肩し得る性能をもち，結果の解釈が容易であった．データの省察から予後因子の探索までを一貫して整合的に行うことができる点で，ここで提示したデータ適応型多分岐樹木構造接近法が推奨できる．

## 1. 序に代えて

生存時間研究では，患者の予後因子の探索が重要な課題の一つである．このとき，それらの要因が生存時間に対して個別に影響を与えることは少なく，多くの場合に，それらの要因の複合関係，あるいは交互作用を伴っている．従来，このような場面では比例ハザード・モデルが代表として用いられているが，このモデルでは要因間の関係を線形結合で表現しているため，輻湊した要因の構造を適切に捉えることが困難である．近年，予後因子の交互作用効果を捉えるための有用な手段として樹木構造接近法が脚光を浴び，諸種の方法が提案されている．ただし，既存の樹木構造接近法の多くが2分岐によって構成されており，複雑な要因構造を適切に捉えるには，過度に単純化され，しかも階層の多い深い樹木が構成され，その解釈に難をきたしている．この点に着目し，Shimokawa & Goto(2006)および衛藤 他(2007)は，より複雑な交互作用要因を比較的に簡単なグラフィクスで提示できる多分岐樹木構造接近法を提案し，その性能を評価している．多分岐樹木構造接近法は，2分岐型樹木構造接近法に比して解釈が平易な深さの浅い樹木を提示し，しかも2分岐樹木構造接近法の適切性を診断することができる．ただし，上記の多分岐樹木構造接近法は，分

* 山梨大学 大学院医学工学総合研究部 〒400-8511 甲府市武田4-3-11 (Tel. 055-220-8395),
E-mail: shimokawa@yamanashi.ac.jp
** 特定非営利活動法人 医学統計研究会 〒560-0085 豊中市上新田2丁目22-10-A411 (Tel. 06-6835-8790),
E-mail: gotoo@bra.or.jp

分岐基準にノンパラメトリック(検定)統計量を用い，得られた樹木の解釈には，Kaplan-Meier プロットを用いている．樹木構造接近法の一つの目的が，生存時間分布に対する予後因子の影響を定量的に解釈し，その結果に基づく知識を，その後の臨床研究につなげることであるが，データの分解から生存時間分布の比較，予後因子の探索までを一貫した形式の上位上で実行できることが望ましい(松原・後藤, 1989)．

2分岐樹木構造接近法には，パラメトリック接近法として，RECPAM(RECursive PArtition and Amagramation：Ciampi&Thiffault, 1988; Ciampi *et al.*, 1989) 法が提案されているが，その多分岐への拡張は試みられていない．さらに，生存時間分布を任意の理論分布に規定することは，柔軟な樹木構造接近法に「かたい」制約を与える惧れがある．そのため，データに適応した形式の柔軟な理論分布の利用が希求される．

本稿では，上記の背景と動機のもとに 生存時間分布にデータ適応型分布，とくにベキ正規分布 (Goto *et al.*, 1979；Goto *et al.*, 1983) を想定して，RECPAM 法を多分岐樹木構造接近法に拡張する．これにより，上述の一貫した解釈が可能になるだけでなく，得られた樹木の結果に基づく新たな臨床研究への動機を提供したり，その計画に依拠する統計的シミュレーションなどへの多くの応用が期待できる

2節では，既存の生存時間研究における樹木構造接近法の現状について略説し，検定統計量に基づく多分岐樹木構造接近法 (Shimokawa & Goto, 2006；衛藤 他, 2007) の要約を与える．3節では，データ適応型多分岐樹木構造接近法 (DAMUST: Data-Adaptive MUlti-Split Tree structured method) を提示する．4節では，その方法の性能を文献例およびシミュレーションにより評価する．文献例では，とくに生存時間分布にベキ正規分布を想定することの効用を評価する．シミュレーションでは，検定統計量に基づく多分岐樹木構造接近法との対比を通して，データ適応型樹木構造接近法の性能を吟味し，その有用性について触れる．5節では得られた結果の解釈を述べ 提案したデータ適応型多分岐樹木構造接近法の特徴を要約する

## 2. 生存時間研究における樹木構造接近法

生存時間研究では患者の予後因子を探索するために比例ハザード・モデルが広範に利用されている．比例ハザード・モデルでは，対数ハザードに対する共変量の回帰関係が加法項で表現されるが，実地でこのような仮定を満たすことは少ない．このような場合の対処として，一般化加法モデル (Hastie & Tibshirani, 1990) の枠組みで比例ハザード・モデルを拡張する方法 (菅田 他, 1992) が提案されている．ただし，この方法には結果の解釈に難がある．その対処のひとつとして，近年，樹木構造接近法が注目され，諸種の方法が提案されている (衛藤 他, 2007)．

Crowley & Ankerst(2005) および Shimokawa & Goto(2006) は，樹木の分岐過程に基づいて，樹木構造接近法を 2 種類に大別している．ひとつは，CART 法 (Breiman *et al.*, 1984) と同様に，ふし内の不均一性の測度を定義し，それを最小にするように樹木を構成する接近法である．たとえば，Therneau & Atkinson(1997) は，生存時間分布に指数分布を想定することで，(各ふしにおける比例ハザード・モデルに対する) 全尤度の偏分を定義し それをふし内の不均一性測度に用いている．LeBlanc & Crowley(1992) は，基礎ハザードに Breslow(1972) の累積ハザード推定量を用いている．同様に，Keles & Segal(2002) は，マルチンゲール残差に基づいて 2 分岐型樹木構造接近法を構成している．

もうひとつは，分岐によって構成される子ふ

し間の分離度を最大にする接近法である．たとえば，ふし間分離度測度に，一般化順位統計量(あるいはログランク統計量)を用いて樹木を構成する方法がある[Segal, 1988；松原 他, 1990]．

CART 法では，分岐手順と刈り込み手順を導入することで，樹木の複雑さと過剰適合のトレード・オフを考慮している．そのため，分岐手順において，過大な樹木を構成しなければならない．これに対して，検定統計量に基づく方法では，予め規定した有意水準によって樹木を規定する．そのため，刈り込み手順が存在しない．

### 2.1 検定統計量に基づく多分岐型樹木構造接近法

既存の樹木構造接近法の殆どが 2 分岐によって構成されている．これは，その過程が CART 型接近法あるいはその変法によって構成されているためである．したがって，分岐手順によって過大な樹木が構成される．他方，多分岐では早期に分岐がすることが指摘されており(Hastie *et al.*,2001)，分岐過程の要件を充たさない憾れがある．また，多分岐ふしは，多段階による分岐によって 2 分岐でも表現でき，さらに分岐候補点探索のアルゴリズムの複雑さと計算負荷の大きさも多分岐型樹木構造接近法がほとんど提案されない理由の一つである．ただし，真の分岐点が多分岐である場合に，上記の多段階による 2 分岐が適切に処理できるか否かは疑問であり (Kim & Loh, 2001)，また，検定統計量に基づく方法では，刈り込み手順が存在しないことから，分岐によって多くの候補 (2 分岐あるいは多分岐) を探索することは，既存の 2 分岐型樹木構造接近法の適切性の評価にも繋がる．Shimokawa & Goto(2006) および衛藤他 (2007) は，多標本一般化順位統計量 (Letón & Zuluaga, 2002) を適用することで，Segal(1988) の方法を多分岐樹木に拡張している．

いま，終結ふしでない任意のふし $\tau_c$ に属する個体 $\{t_i, \delta_i\}_{i=1}^{n^{(c)}}, i \in \tau^c (i = 1, 2, \ldots, n^{(c)})$ の共変量 $x_j$ に基づく $R$ 分岐を考える．ここに，$t_i$ は生存時間，$\delta_i$ は中途打ち切り指標 (1：生存，0：死亡)，そして $n^{(c)}$ はふし $\tau_c$ に属する個体数である．このとき，$R$ 群に対する一般化順位統計量 $MTW^{(c)}$(分岐基準) は

$$MTW^{(c)} = \boldsymbol{U}^{(c)^{\mathrm{T}}} \mathrm{Var}(\boldsymbol{U}^{(c)})^{-1} \boldsymbol{U}^{(c)}$$

で与えられる．ここに，$\boldsymbol{U}^{(c)} = (U_1^{(c)}, U_2^{(c)}, \ldots, U_{R-1}^{(c)})^{\mathrm{T}}$，および $\mathrm{Var}(\boldsymbol{U}^{(c)}) = \{\mathrm{Cov}(U_r^{(c)}, U_{r'}^{(c)})\}_{rr'}$ は，それぞれ

$$U_r^{(c)} = \sum_{k=1}^{K} w_{rk}^{(c)} \left( d_{rk}^{(c)} - d_k^{(c)} \frac{n_{rk}^{(c)}}{n_k^{(c)}} \right),$$

および，

$$\mathrm{Cov}(U_r^{(c)}, U_{r'}^{(c)}) = \sum_{k=1}^{K} w_{rk}^{(c)^2} \frac{n_{rk}^{(c)} d_k^{(c)} (n_k^{(c)} - d_k^{(c)})}{n_k^{(c)} (n_k^{(c)} - 1)} \left[ \kappa_{rr'}^{(c)} - \frac{n_{r'k}^{(c)}}{n_k^{(c)}} \right]$$

であり，$n_{rk}^{(c)} \left( n_k^{(c)} = \sum_{r=1}^{R} n_{rk}^{(c)} \right), k = 1, 2, \ldots, K$ は，それぞれ，$R$ 個の子ふしに属する個体の時点 $t_k$ でリスクに曝された個体数，$d_{rk}^{(c)} \left( d_k^{(c)} = \sum_{r=1}^{R} d_{rk}^{(c)} \right)$ は時点 $t_k$ における死亡数，$\kappa_{rr'}^{(c)}$ はもし $r = r'$ ならば 1 であり，$r \neq r'$ ならば 0 である．また，重み $w_i^{(r)}$ は，$w_{rk}^{(c)} = 1$ のときに $R$ 標本のログランク統計量，$w_{rk}^{(c)} = n_r^{(c)}$ のときに $R$ 標本の Gehan 統計量，$w_{rk}^{(c)} = n_r^{(c)^{1/2}}$ のときに $R$ 標本の Tarone & Ware 統計量，$w_{rk}^{(c)} = \hat{S}_r^{(c)}$ のときに $R$ 標本の Peto-Peto 統計量に対応している．

以下に検定統計量に基づく多分岐型樹木構造接近法 (MUSTGRAS:MUlti-Split Tree structured method based on Generalized RAnk Statistics) のアルゴリズムを略説する：

**M0** 初期設定として，重み $w_k^{(c)}$，停止基準 (有

意水準, 終結ふし, 最小個体数), および最大分割数$R_{\max}$に任意の数値を代入する.

**M1** 末端のふしの個数$C$を1とする.

**M2** 停止条件 ((a) すべてのふしが予め規定した有意水準未満になる, (b) ふし内標本サイズが最小ふしサイズよりも小さくなる) のいずれかを満たすまで以下の処理を実行する.

**M2-1** 末端のふし$\tau(c=1,2,\ldots,C)$について以下の手順を行う.

**M2-1-1** 分岐数$R=2,\ldots,R_{\max}$について, 以下の手順を行う (ただし, $x_j$の最大可能分岐数が$R_{\max}$以下の場合には, 分岐可能な最大個数まで以下の手順を行う).

**M2-1-1a** 共変量$x_j$のカテゴリを$R$分割する全組み合わせ対で$R$標本に対する一般化順位統計量$MTW_{Rjl}^{(c)}(j=1,\ldots,p;l=1,\ldots)$を求め, そのp値$p_{Rjl}^{(c)}=F_{\chi^2,R-1}(MTW_{Rjl}^{(c)})$を算出する. ここに$l$は探索した分岐候補を表す添字であり, $F_{\chi^2,R-1}$は, 自由度$R-1$のカイ2乗分布の上側確率である.

**M2-1-1b** $p_{Rj^*l^*}^{(c)}=\min_{j,l}\left\{p_{Rjl}^{(c)}\right\}$をもつ変数$x_{j^*}$と$l^*$番目の分岐点が, ふし$\tau^{(c)}$において, 生存時間の最良の$R$分岐を与えるとして, 末端のふし$\tau^{(c)}$の$R$分岐の候補とする.

**M2-1-2** $p_{R^*j^*l^*}^{(c)}=\min_{R}\left\{p_{Rjl^*}^{(c)}\right\}$をもつ$R^*$分岐を最良の分岐数を与えるとする.

**M2-2** $p_{R^*j^*l^*}^{(c^*)}=\min_{c}\left\{p_{R^*j^*l^*}^{(c)}\right\}$をもつ末端のふし$\tau^{(c^*)}$が最良の分岐ふしとして選定される. そして, 末端のふし$\tau^{c^*}$に属する個体が分岐変数$x_{j^*}$によって, $l^*$番目の分岐候補で$R^*$分岐される.

**M2-3** $C\rightarrow C+(R^*-1)$とする.

多分岐樹木構造接近法では, 1個のふしにおける最大分割数を制限している. これは, 多分岐樹木構造接近法では, 分岐点の選定に多くの計算負荷を要するためである.

## 3. データ適応型分布に基づく樹木構造接近法

2.1節で略説した, 既存の多分岐型樹木構造接近法では, 多標本一般化順位統計量 (ノンパラメトリック検定統計量) が分岐基準として利用されている. ただし, 樹木構造接近法の目標は, 類似した予後をもつ患者群に分類する (たとえば, ステイジ分類) だけでなく, 何らかの治療法に対する適応患者像を抽出する場合にも用いられる.

このとき, 上述の分岐基準を用いることは, 生存時間分布の同定から治療法の比較および予後の予測といった, データ解析の多段階の手順を実行していくときに解析結果の接合で解釈が難しくなる (松原・後藤, 1989). たとえば, 任意の治療法に対する適応患者像を樹木構造接近法によって探索することで, 新たな臨床研究の動機が得られることがある. このとき, ノンパラメトリック接近法では, その研究計画につなぐ一貫した形式での手順を与えることは困難である. そのため, 本稿では, データに適応的で柔軟な生存時間分布にデータ適応型分布 (族) を各ふしの生存時間分布に想定する. データ適応型分布 (族) には, ベキ正規分布, 対数ガンマ分布 (Prentice, 1974), あるいは一般化ロジスティック分布 (Prentice, 1976) が代表的である. 本稿

では,(a) 結果の解釈が平易であること, (b) パラメータの推定が安定していること (局所最適値あるいは初期値によるパラメータの推定で発散が起こりにくい), (c) 後続の解析に繋げることや解釈が比較的容易であること, からベキ正規分布を用いる.

これにより, 樹木の構成から結果の解釈および, 後続の研究に至るまでの過程をデータ適応型分布の範疇で一貫してとり扱うことができ, しかも影響要因を (多分岐樹木により) 比較的簡素な形式で提示することができる.

### 3.1 ベキ正規分布

生存時間$t$にベキ変換 (Box & Cox, 1964)

$$t^{(\lambda)} = \begin{cases} (t^{\lambda}-1)/\lambda & \lambda, \neq 0, \\ \log t, & \lambda = 0 \end{cases} \qquad (3.1)$$

を施したとき, 変換後の分布は正規分布に従うことが意図されている. ベキ正規分布は, このベキ変換の目標を正規化においたときに変換前に規定される分布である (Goto *et al.*, 1979 : Goto *et al.*, 1983 : Johnson *et al.*, 1994). ここに, $\lambda$はベキ変換パラメータである. $\lambda \neq 0$のとき, ベキ正規分布$\mathrm{PND}(\lambda, \mu, \sigma^2)$の確率密度関数$f_{\mathrm{PND}}$は

$$f_{\mathrm{PND}}(t;\lambda,\mu,\sigma) = \frac{t^{\lambda-1}}{A(\lambda)\sqrt{2\pi}\sigma} \exp\left\{-\frac{(t^{(\lambda)}-\mu)^2}{2\sigma^2}\right\}$$

で定義される (Goto *et al.*, 1979 : Goto *et al.*, 1983 : Goto & Inoue, 1980 : Johnson, *et al.*, 1994). ここに, $\mu$と$\sigma^2$はベキ変換後の観測値$t^{(\lambda)}$が近似的に正規分布に従うときの平均と分散を意味する. $A(\lambda)$はベキ正規分布の確率比例定数項

$$A(\lambda) = \begin{cases} \Phi\left(\mathrm{sign}(\lambda)\left|(\lambda\mu+1)/(\lambda\sigma)\right|\right), & \lambda \neq 0, \\ 1, & \lambda = 0 \end{cases}$$

である. ここに, $\Phi(\cdot)$は標準正規分布の累積分布関数である (Goto *et al.*, 1979 : Goto *et al.*, 1983 : Goto & Inoue, 1980).

ベキ正規分布の推測において, $A(\lambda)$を考慮に入れたパラメータの精確な推定を行うことは, 一般的に困難である (Goto & Inoue, 1980). そのため, 通常では, $A(\lambda) \approx 1$と想定したもとでパラメータの推定を行う. 因に, これらの推測の一致性などについて後藤 他 (1991), 濱崎・後藤 (1996), Hamasaki & Goto(1998) に与えられている.

### 3.2 RECPAM 法

Ciampi & Thiffault(1988) および Ciampi *et al.*(1989) は, 生存時間分布を予め規定することで樹木を構成する, RECPAM(RECurcive Partition and AMalgamation) 法を提案している. この方法では, 各ふしに何らかのパラメトリックな生存時間分布を想定することで, 分割 (および合併) 基準には尤度比統計量あるいは情報量規準が用いられている.

いま, 生存時間$t_i (i = 1, \ldots, n)$, 中途打ち切り指標$\delta_i (\delta_i = 1$は生存, $\delta_i = 0$は中途打ち切りを表す) が$p$個の共変量$\boldsymbol{X}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})$とともにとられたとする. このとき, RECPAM 法では, 任意のふし$\tau$に含まれる生存時間$t_i (i \in \tau)$がパラメータ$\boldsymbol{\theta}$で規定される分布に従うと仮定したもとで, 評価基準に尤度比統計量 (帰無仮説 : $\mathrm{H}_0 : \boldsymbol{\theta}_l = \boldsymbol{\theta}_r = \boldsymbol{\theta}$, 対立仮説 : $\mathrm{H}_1 : \boldsymbol{\theta}_l \neq \boldsymbol{\theta}_r$) を用いて分岐点を探索する. ここに, $\boldsymbol{\theta}_l$および$\boldsymbol{\theta}_r$は, それぞれ分岐された左右の子ふしにおける生存時間分布のパラメータである.

したがって, 末端のふし$\tau_c (c = 1, \ldots, C)$での, 変数$x_j$の分岐候補点$l$における分岐基準は

$$\rho_{jl}^{(c)} = -2\{\log L(\hat{\boldsymbol{\theta}}^{(c)};t) - \log L(\hat{\boldsymbol{\theta}}_l^{(c)};t) L(\hat{\boldsymbol{\theta}}_r^{(c)};t)\}$$

である. ここに, $L$は生存時間分布の尤度関数,

$\hat{\boldsymbol{\theta}}^{(c)}$ は仮説$H_0$のもとでの$\boldsymbol{\theta}^{(c)}$の最尤推定量，そして$\hat{\boldsymbol{\theta}}_l^{(c)}, \hat{\boldsymbol{\theta}}_r^{(c)}$ は仮説$H_1$のもとでの$\boldsymbol{\theta}_l^{(c)}, \boldsymbol{\theta}_r^{(c)}$の最尤推定量である．$\rho_{jl}^{(c)}$は正則条件のもとで漸近的に自由度$q$(生存時間分布のパラメータ数)のカイ2乗分布に従う．最良分岐点$s^*(\tau)$は，$\rho_{jl}^{(c)}$のp値が最も小さなものを選択する．この分岐は，予め規定した有意水準$\alpha$まで行われる．

樹木の成長は，いいかえれば尤度比検定統計量の増加系列によって構成される．ただし，任意の有意水準にのみ依拠した樹木の構成は，過大あてはめを惹き起こす恐れがある．また，検定統計量がデータに依存するという事実を考慮せずに無視していることから，この適切性基準は適当でない．そのため，RECPAM法では，分岐後に，すべての部分樹木系列のなかから最適な部分樹木を選択するための手順を設けている．そこでは，(a) 1例削除交差確認法 (Stone, 1974), (b) ジャックナイフ法 (Efron, 1982), (c) 赤池の情報量規準 (AIC；Akaike, 1974) のいずれかが用いられる．これら3種類の方法のうち，前者の2手法はリサンプリングを必要とすることから，多大な計算負荷を要する．そのため，AICを分岐基準に用いるのが一般的である．

### 3.3 データ適応型多分岐樹木構造接近法

RECPAM法では，既存の樹木構造接近法と同様に，2分岐によって樹木が構成される．他方，多分岐への拡張は試みられていない．本稿では，RECPAM法の多分岐への拡張を試みる．これにより，通常のRECPAM法の適切性をデータ適応型多分岐樹木構造接近法 (DAMUST: Data-Adaptive MUlti-Split Tree structred method) によって評価することができる．

本稿では，生存時間分布にベキ正規分布を想定する．このとき，確率比例定数項$A(\lambda)$を含めたもとでの推測は，計算が困難であるため，Box & Cox(1964) に倣い，$A(\lambda) = 1$を仮定する．このことは，ベキ変換後の観測値が正規分布に従うことを意味する (Goto & Hamasaki, 2002；濱崎・後藤,2002)．

すなわち，末端のふし$\tau_c (c = 1, \ldots, C)$での，変数$x_j$の分岐候補点$l$における$R$分岐に対する尤度比統計量に基づく分岐基準は

$$\rho_{Rjl}^{(c)} = -2\sum_{r=1}^{R}\left\{\log L_{\mathrm{PND}}(\hat{\boldsymbol{\theta}}^{(c)}; t) - \log L_{\mathrm{PND}}(\hat{\boldsymbol{\theta}}_r^{(c)}; t)\right\}$$

である．ここに，$\hat{\boldsymbol{\theta}}^{(c)} = (\hat{\lambda}_c, \hat{\mu}_c, \hat{\sigma}_c)$は，ベキ正規分布のパラメータの最尤推定値，$\log L_{\mathrm{PND}}(\hat{\boldsymbol{\theta}}^{(c)}; t)$は，ベキ正規分布の対数尤度関数

$$\begin{aligned}
&\log L_{\mathrm{PND}}(\hat{\boldsymbol{\theta}}^{c}; t)\\
&= \sum_{i\in\tilde{\tau}_c}\delta_i \log f(t_i) + \sum_{i\in\tilde{\tau}_c}(1-\delta_i)\log S(t_i)\\
&= \sum_{i\in\tilde{\tau}_c}\left\{-\frac{1}{2}\log(2\pi\hat{\sigma}_c^2) - \frac{1}{2\hat{\sigma}_c^2}\left(\frac{t_i^{\hat{\lambda}_c}-1-\hat{\lambda}_c\hat{\mu}_c}{\hat{\lambda}_c}\right)^2\right.\\
&\left.\times(\hat{\lambda}_c-1)\log t_i\right\} + \sum_{i\in\tilde{\tau}_c}\left\{1-\Phi\left(\frac{t_i^{\hat{\lambda}_c}-1-\hat{\lambda}_c\hat{\mu}_c}{\hat{\lambda}_c\hat{\sigma}_c}\right)\right\}
\end{aligned}$$

であり，$S(t_i)$はベキ正規分布における生存時間関数である．式 (3.2) は，帰無仮説$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \cdots = \boldsymbol{\theta}_R = \boldsymbol{\theta}$のもとで，漸近的に自由度$3(R-3)$のカイ2乗分布に従う．

DAMUSTでは，式(3.2)により得られた検定統計量のp値を分岐基準に用いる．データ適応型多分岐樹木構造接近法の実行手順は，3段階，すなわち，分岐，最適部分樹木の選定，および併合によって構成される．以下にその流れを詳説する．

**分岐：** 分岐は，2.1節の検定統計量に基づく方法と同様の流れで実行される．すなわち，多標本一般化順位統計量$MTH^{(c)}$の代わりに，任意の分岐候補点$c$によって分割された個体の生存時間にあてはめたベキ正規分布のパラメータの

推定値$\boldsymbol{\theta}^{(c)}$に基づく尤度比統計量$\rho^{(c)}$を用い，そのp値が最も小さくなる分岐候補で再帰的に分岐させる．以下にそのアルゴリズムを提示する：

**D0** 初期設定として，停止基準（最小有意水準，終結ふしの最小個体数），および最大分割数$R_{\max}$に任意の数値を代入する．

**D1** 末端のふしτの個数$C$を1とする．

**D2** 停止条件（(a) すべてのふしが予め規定した有意水準未満になる，(b) ふし内標本サイズが最小ふしサイズよりも小さくなる）のいずれかを満たすまで以下の処理を実行する．

**D2–1** $C$個の末端のふし$\tau^{(c)}(c=1,\ldots,C)$についてベキ正規分布$\mathrm{PND}(\lambda_c,\mu_c,\sigma_c)$のパラメータを推定する$(\hat{\lambda}_c,\hat{\mu}_c,\hat{\sigma}_c)$．

**D2–2** それぞれの末端のふしに対して，以下の手順を行う．

**D2–2–1** 分岐数$R=2,\ldots,R_{\max}$について手順を行う（ただし，$x_j$の最大可能分岐数が$R_{\max}$以下の場合には，分岐可能な最大個数まで以下の手順を行う）．

**D2–2–1a** 共変量$x_j$のカテゴリーを$R$分割したもとで，それぞれの部分集合の生存時間に対して，ベキ正規分布$\mathrm{PND}(\lambda_{rjl}^{(c)},\mu_{rjl}^{(c)},\sigma_{rjl}^{(c)}),r=1,2,\ldots,R$をあてはめる$(\hat{\lambda}_{rjl}^{(c)},\hat{\mu}_{rjl}^{(c)},\hat{\sigma}_{rjl}^{(c)})$．

**D2–2–1b** あてはめたベキ正規分布に基づき，尤度比統計量$\rho_{Rjl}^{(c)}(j=1,\ldots,p;\ l=1,\ldots)$を求め，そのp値$p_{Rjl}^{(c)}=F_{\chi^2,3(R-3)}(\rho_{Rjl}^{(c)})$を算出する．ここに，$l$は探索した分岐候補を表す添え字であり，$F_{\chi^2,3(R-3)}$は，自由度$3(R-3)$のカイ2乗分布の上側確率である．

**D2 2 1c** $p_{Rj^*l^*}^{(c)}=\min_{j,l}\left\{p_{Rjl}^{(c)}\right\}$をもつ変数$x_{j^*}$と$l^*$番目の分岐点が，ふし$\tau^{(c)}$において，生存時間の最良の$R$分岐を与えるとして，末端のふし$\tau^{(c)}$の$R$分岐の候補とする．

**D2–2–2** $p_{R^*j^*l^*}^{(c)}=\min_{R}\left\{p_{Rj^*l^*}^{(c)}\right\}$をもつ$R^*$分岐が末端のふし$\tau^{(c)}$の最良の分岐数を与えるとする．

**D2–3** $p_{R^*j^*l^*}^{(c^*)}=\min_{c}\left\{p_{Rj^*l^*}^{(c)}\right\}$をもつ末端のふし$\tau^{(c^*)}$が最良の分岐ふしとして選定される．そして，末端のふし$\tau^{c^*}$に属する個体が分岐変数$x_{j^*}$によって，$l^*$番目の分岐候補で$R^*$分岐される．

**D2–4** $C\to C+(R^*-1)$とする．

本アルゴリズムは，検定統計量に基づく多分岐型樹木構造接近法に倣って構成されている．そのため，本手法のプログラムは，衛藤 他(2007)で提示されたMATLAB toolboxを用いて，評価基準およびベキ正規分布のパラメータ推定を導入することで構成された．

**最適部分樹木の探索：** CART接近法では，刈り込み手順により，根幹ふしまで逐次に樹木を剪定することで，部分樹木系列を得る．そして，そのなかで最適な部分樹木を交差確認法によるリスク推定値から選定する．他方，DAMUSTでは，Ciampi & Thiffault(1988) およびCiampi *et al.*(1989) に倣い，赤池の情報量規準 (AIC) を用いる．すなわち，AICは，末端のふし（終結ふしおよび剪定候補となるふし）にあてはめたベキ正規分布に基づいて計算される．

いま，分岐過程によって得られた最大樹木を$T$とし，その$C$個の終結ふしを$\tau_c$，および$C^\dagger$個の非終結ふしを$\tau_{c^\dagger}$とする．したがって，ふし$\tau$によって剪定された部分樹木は，$T-T_{\tau_{c^\dagger}}\cup\tau_{c^\dagger}$により得られる．このとき，個々の終結ふしにおけるベキ正規分布のパラメータ数は3であることから，最適部分樹木の選定過程では，すべての非終結ふし$\tau_{c^\dagger}(c^\dagger=1,2,\ldots,C^\dagger)$に対して

$$AIC(\tau_{c^\dagger}) = -2\left\{\sum_{c\notin\tau_{c^\dagger}}\log L_c(\hat{\boldsymbol{\theta}}_c)+\log L_{c^\dagger}(\hat{\boldsymbol{\theta}}_{c^\dagger})\right\} +6(C_{-c^\dagger}+1)$$

を計算し，$\min_{c^\dagger} AIC(\tau_{c^\dagger})$となる部分樹木を最適部分樹木として選定する．ここに，$C_{-c^\dagger}$は，ふし$\tau_{c^\dagger}$に含まれない終結ふしの数である．よって，$\tau_{c^\dagger}$を末端のふしとしたときの(いいかえれば，$\tau_{c^\dagger}$で剪定したときの)，末端のふし数は$C_{-c^\dagger}+1$である．

AICによる最適部分樹木の選定の利点は，樹木の適合度を定量的に示すことができる点にある．本稿では，生存時間分布にベキ正規分布を想定しているが，指数分布あるいはWeibull分布といった，他の理論分布でも構成可能である．このとき，生存時間分布の適切性を，AICによって評価することができる．

**併合：** 分岐によって構成される樹木は，すべての終結ふし対で，明確に異なる生存時間曲線をもつことは保障されていない．したがって，最終的に異なる生存時間曲線をもつ群分けが望まれるステイジ分類に樹木構造接近法を，そのままの形式で利用することは困難である．RECPAM法では，終結ふしを任意の検定統計量に基づいてクラスタリング(併合)することで，この問題を回避している．併合アルゴリズムは，最も類似する終結ふしの対を併合することからはじまり，最も類似なクラスターの対での任意の検定統計量が有意水準$\alpha$で有意にならなくなるまで連続的に併合する．検定統計量には，Ciampi & Thiffault(1988)およびCiampi *et al.*(1989)は尤度比統計量を用い，松原 他(1990)は，ログランク統計量を用いている．

データ適応型樹木構造接近法では，生存時間分布にベキ正規分布を想定しており，一貫して尤度に基いて樹木を構成する．そのため，本稿では，すべての終結ふしの組み合わせによって構成されるクラスターに対して尤度比統計量(3.2)を算出することで，最良なクラスターを探索する．

## 4. 事例検討および数値検証

### 4.1 事例検討

ここでは，事例検討を通して，DAMUSTの特性，および生存時間分布にデータ適応型分布を想定することで有用な特徴および示唆される数値検証の仮説を探求する．

#### 4.1.1 乳癌データ

乳癌に罹患している患者に対して，ホルモン療法の効果を検討するために，ドイツ乳癌研究グループが7個の予後因子(年齢，閉経の有無，腫瘍径，ステイジ，リンパ節転移個数，プロゲステロン・レセプタ個数，エストロゲン・レセプタ個数，およびホルモン療法が行われたか否か)とともに生存時間を評価している．ここでの目標は，患者の生存時間に及ぼす共変量の影響(とくに，ホルモン療法との交互作用効果)を評価することである．

図1は，本データにMUSTGRASを適用した結果である．先ず，すべての個体は，リンパ節転移個数で3分岐した．次いで，3分岐した部分

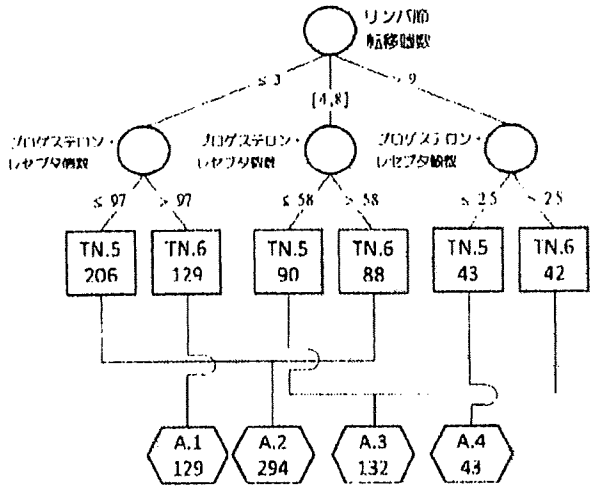
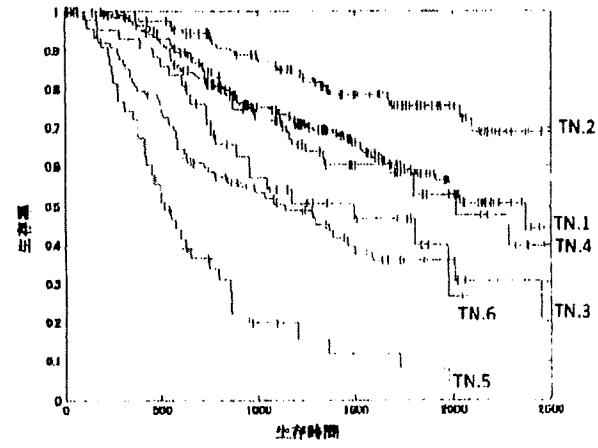



図 1: 乳癌データに対する MUSTGRAS の結果

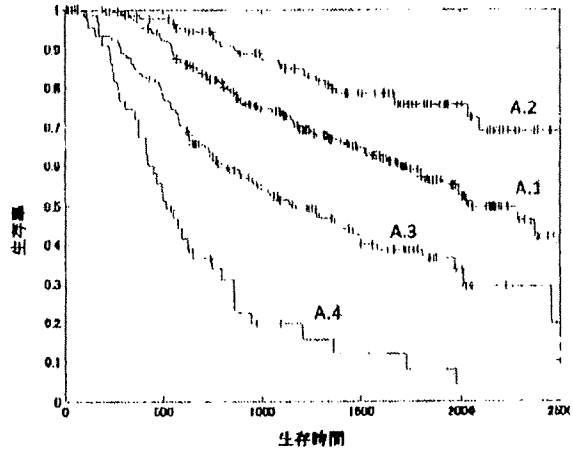
集合は，すべてプロゲステロン・レセプタ個数で分岐した．このとき，分岐点は，リンパ節転移個数が多くなるにつれて，小さくなる傾向にあった．ただし，ホルモン療法の有無は影響因子と現れていない．

図 2 (a) は，得られた樹木に対する Kaplan-Meier プロットである．リンパ節転移個数が多くなるにつれて，生存時間が短くなる傾向にあった．このとき，プロゲステロン・レセプタ個数が多い患者は，少ない患者に比して生存時間が長かった．ログランク検定の p 値に基づいてすべての終結ふし対が有意になるように併合を行った．その結果，TN.1と TN.4および TN.3と TN.6が併合した．したがって，併合手順では，プロゲステロン・レセプタ個数が多い終結ふしは，リンパ節転移個数が少なく，かつプロゲステロン・レセプタ個数が少ない終結ふしと結合した．図 4 (b) は，得られた合併ふしに対する Kaplan-Meier プロットである．4 個の併合ふしの生存時間分布が明瞭に分岐した．

図 3 は，DAMUST を適用した結果である．すべての個体は，MUSTGRAS の場合と同様にリンパ節転移個数で分岐した．リンパ節転移個数が 3 個以下の患者は，ホルモン療法の有無で分岐した．リンパ節転移個数が 4 個以上の 2 個の




(a) 終結ふし




(b) 合併ふし

図 2: 乳癌データに対する MUSTGRAS の結果

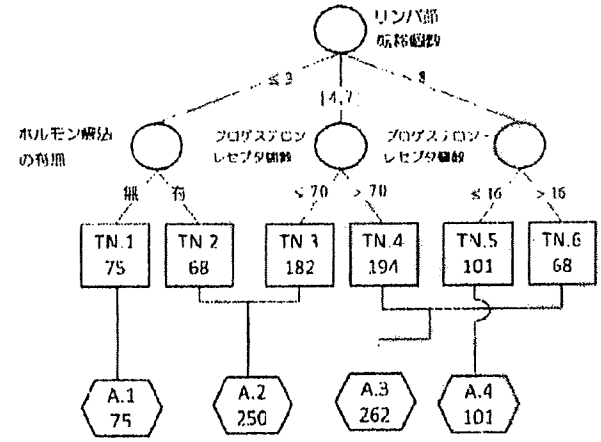
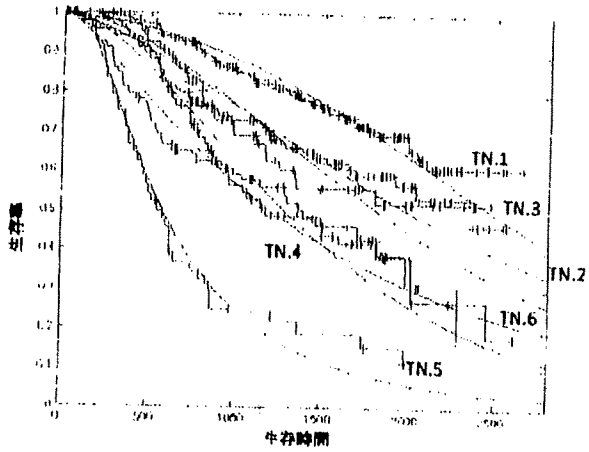
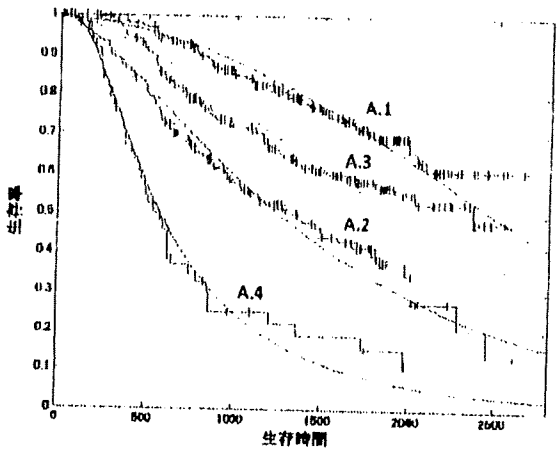



図 3: 乳癌データに対する DAMUST の結果

ふしは，いずれもプロゲステロン・レセプタ個数により分岐した．このとき，分岐点の個数は，MUSTGRAS と同様にリンパ節転移個数が多




(a) 終結ふし




(b) 合併ふし

図 4: 乳癌データに対する DAMUST の Kaplan-Meier プロット

くなるにつれて，少なくなる傾向にあった．

図4(a) は，得られた樹木に対する Kaplan-Meier プロットである．リンパ節転移個数が 3 個以下で，かつホルモン療法を受けた患者の予後が最良であり，リンパ節転移個数が 8 個以上で，かつもプロゲステロン・レセプタ個数が16個超の患者の予後が極端に悪かった．したがって，ホルモン療法の有効性は，リンパ節転移の少ない，進行程度の浅い患者に影響を受けるようである．

表 1 は，生存時間分布に他の理論分布を想定した場合の AIC および終結ふしの数である．終結ふしの数には，大差がないものの，指数分布の AIC 値が4573.8であり極端に高かった．また，

表 1: 乳癌データに対するそれぞれの生存時間分布における AIC 値

| 生存時間分布 | AIC 値 | 終結ふし数 |
|---|---|---|
| ベキ正規分布 | 4292.7 | 6 |
| Weibull 分布 | 4316.6 | 6 |
| 指数分布 | 4573.8 | 6 |
| ガンマ分布 | 4558.6 | 5 |

ベキ正規分布が (それぞれの終結ふしにおけるベキ変換パラメータ $\lambda$ はそれぞれ，0.12, 0.43, 0.02, 0.27, 0.14, -0.02であった) 4292.7であり最良の適合を示した．したがって，生存時間分布にベキ正規分布を想定することの適切性が示唆される．

ただし，TN.2および TN.3の生存率に明瞭な差は認められなかった．これは，DAMUST では，すべての終結ふし対で，明確に異なる生存曲線をもつことは保障されていないためである．そのため，併合手順を用いて，すべてのふし対における尤度比統計量の p 値が有意 (有意水準0.05) になるまで併合を行った (図 3 の 8 角形が併合ふしである)．その結果，TN.2と TN.3 が併合され，TN.4と TN.6が併合された．これに対して，TN.1は併合されなかった．すなわち，リンパ節転移個数が少ない症例に対してホルモン療法が有効であることが改めて示唆された．図 2 (b) は，併合ふしに対する Kaplan-Meier プロットである．その結果，リンパ節転移個数が多く，かつプロゲステロン・レセプタ個数が少ない患者の予後の悪いことが示唆された．

### 4.1.2 心臓移植データ

本データは，スタンフォード大学で検討された，心臓移植患者64名 (死亡40例，中途打ち切り24名) に対して，生存時間 (日数)，年齢，事前治療の有無，ミスマッチの個数，ミスマッチ評点，ADL 後退の有無がとられている (Clowley & Hu, 1977)．

図5は，本データに対するMUSTGRASの結果である(ここで，一般化順位統計量の重みを1とし($R$標本のログランク統計量)，終結ふしの最小個体数を5個，年齢およびミスマッチ評点の最大分岐数を5分岐とした．さらに，有意水準は併合手順を考慮に入れ0.10とした)．全データは，先ず年齢で分岐し，次いで，52歳以下の患者は，年齢で分岐し，53歳以上の患者はミスマッチ評点で分岐した．最終的には，6個の終結ふしが得られた．このとき，終結ふしの説明変数には，年齢およびミスマッチ評点のみが選定された．さらに，年齢での分岐が多かった．図6(a)は，その結果に対応するKaplan-Meierプロットである．すなわち，53歳以上の患者の予後が最悪であった．これに対して，TN.1((年齢$\leq$ 38) $\bigcap$ (ミスマッチ評点$\leq$ 1.46))の患者の予後は最良であった．ログランク検定を用い，有意水準0.05のもとで併合を行った結果，TN.1とTN.2が新たな併合ふし(A.1)を構成し，TN.3からTN.6までがもう一つの併合ふし(A.2)を構成した．このとき，併合手順後のKaplan-Meierプロットを図6(b)に示す．その結果，A1((年齢$\leq$ 43) $\bigcap$ (ミスマッチ評点$\leq$ 1.46))の患者が他の患者に比して良好な予後をもつことが示唆された．

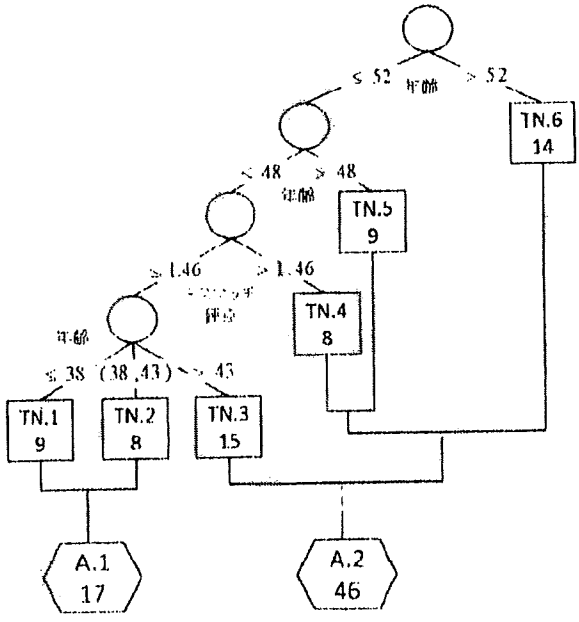
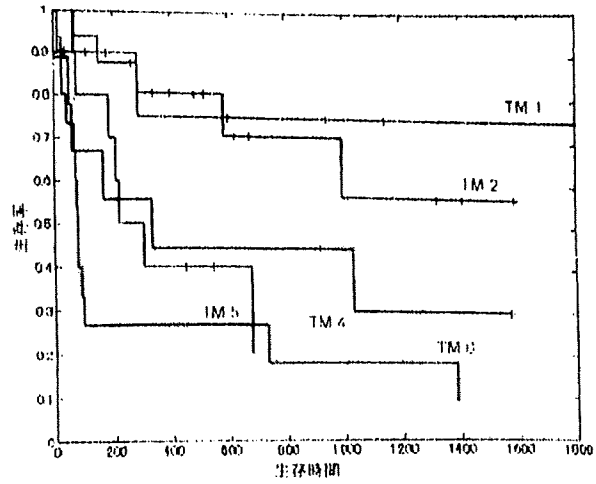
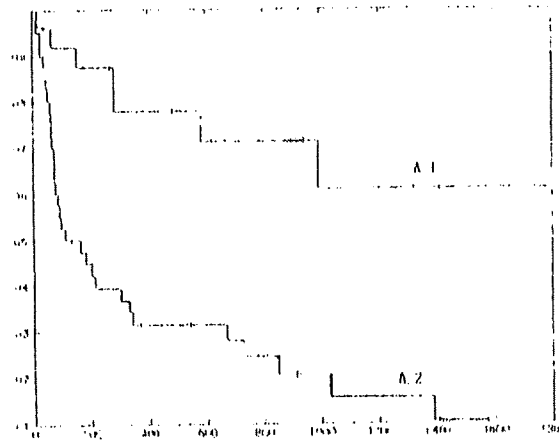



図 5: 心臓移植データに対するMUSTGRASの結果




(a) 終結ふし




(b) 合併ふし

図 6: 乳癌データに対するMUSTGRASのKaplan-Meierプロット

図7は，本データに対するDAMUSTの結果である．初期の分岐には，MUSTGRASと同様に年齢が選択され，分岐点は48歳で，僅かに低かった．次いでいずれのふしも，事前治療の有無により分岐した．そして最終的には，6個の終結ふしが得られた．このとき，分岐に関わる変数として，年齢，事前治療の有無，ミスマッチ評点の3個の変数が選択され，MUSTGRASの結果に比して多かった．

表2は，生存時間分布に他の理論分布を想定

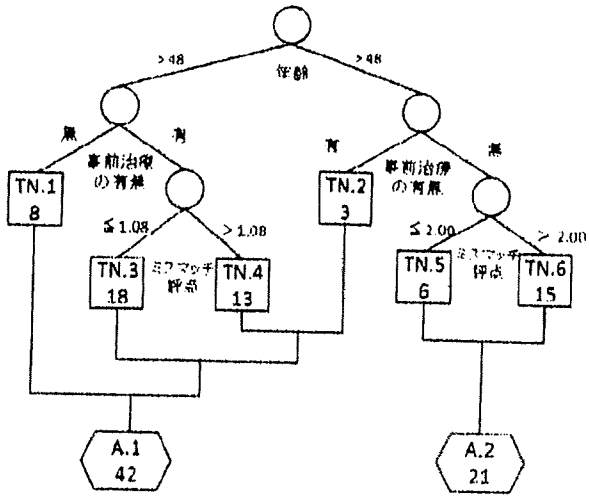


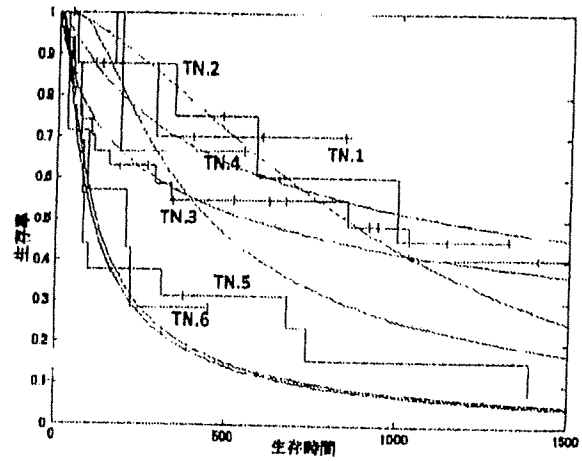

図 7: 心臓移植データに対する DAMUST の結果

表 2: 心臓移植データに対するそれぞれの生存時間分布における AIC 値

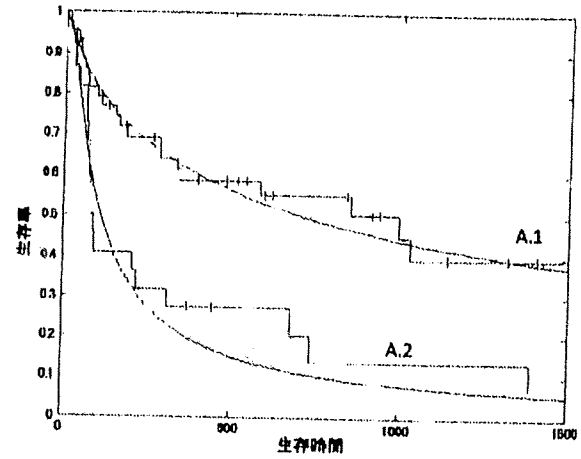
| 生存時間分布 | AIC 値 | 終結ふし数 |
|---|---|---|
| ベキ正規分布 | 594.1 | 6 |
| Weibull 分布 | 596.7 | 6 |
| 指数分布 | 597.3 | 6 |
| ガンマ分布 | 599.0 | 6 |

した場合の AIC および終結ふしの数である. ベキ正規分布の AIC 値が最良の適合を示した. したがって, 生存時間分布にベキ正規分布 (それぞれの終結ふしのベキ変換パラメータ $\lambda$ は-0.02, 0.27, 0.12, 0.03, 0.41であった) を想定することの適切性が示唆された.

図 8 (a) は, DAMUST に対する Kaplan-Meier プロットである. TN.1((年齢≤ 48)∩(ミスマッチ評点≤ 1.08)∩(事前治療=有)) の生存率が最も高く, TN.6((年齢> 48)∩(ミスマッチ評点≤ 2.00)∩(事前治療=無)) の生存率が最も低かった. さらに, TN.3と TN.4に想定したベキ正規分布の生存時間曲線が交錯しており, 解釈が困難であった. これは, MUSTGRAS では, 分岐基準に多標本ログランク統計量が用いられており, 各ふしの分岐において, 比例ハザード性が仮定されている. そのため, 生存時間分布が交差することは少ない. 他方, DAMUST では, 潜在基礎分布を予め想定しており, 交差ハザードに対する適応対処の手段がない. そのため, 既存の多分岐樹木構造接近法に比して, 併合手順が重要な役割をもつと考えられる.




(a) 終結ふし




(b) 合併ふし

図 8: 心臓移植データに対する MUSTGRAS の Kaplan-Meier プロット

有意水準0.05のもとで併合を行った結果, TN.1から TN.4までが併合ふし A.1を構成し, TN.5と TN.6が併合ふし A.2を構成した. すなわち, 年齢が48歳超でかつ事前治療がない患者の予後がその他の患者と異なることが示唆された. 図 8 (b) は, 併合後の MUSTGRAS の Kaplan-Meier プロットである. A.1のほうが A.2の予後よりも高かった. MUSTGRAS では, 年齢による影響が顕著だったのに対して, DAMUST では事前治療の有無が予後を左右し

ていた．本データにおいて，48歳以上に対する，事前治療の有用性が示唆されたことは，今後の治療計画に対する一つの知見となるであろう．

### 4.2 数値検証

文献例では，多分岐型樹木構造接近法の生存時間分布にベキ正規分布を想定することの有用性が示唆されたが，MUSTGRASの性能の評価は行われていない．本節では，生存時間研究の特殊性に配慮し，若干の数値検証によってMUSTGRASとDAMUSTの性能の評価を行う．

**目標とデザイン：** 文献例では，生存時間分布にデータ適応型分布を想定することの有用性を提示することができた．ただし，文献例では，既存の真の樹木構造が不明なため，DAMUSTとMUSTGRASとの性能の比較は行っていない．ここでは，(a)DAMUSTが有用な状況，(b)真の構造がデータ適応型分布に従うとき，MUSTGRASおよび他の生存時間分布を想定した多分岐型樹木構造接近法の挙動，に注目して数値検証を行う．ここでは，DAMUST, MUSTGRASおよび指数分布を想定した場合の多分岐樹木構造接近法をとり上げて評価する．

衛藤 他 (2007) では，真のモデル構造が樹木によって平易に表すことができる場合のみに焦点をあててシミュレーションを行っている．本稿では，衛藤 他 (2007) の示唆に倣い，数値検証を実行する．ここでは，多分岐樹木構造接近法の有用性が示唆されたモデル

$$
\text{//} \sim \begin{cases} \text{ふし 1：PND}(\lambda, \mu_1, \sigma) \\ \qquad \text{if} \quad (x_1 \le 2) \cap (x_2 \le 3) \\ \text{ふし 2：PND}(\lambda, \mu_2, \sigma) \\ \qquad \text{if} \quad (x_1 \le 2) \cap (x_2 > 3) \\ \text{ふし 3：PND}(\lambda, \mu_3, \sigma) \\ \qquad \text{if} \quad (x_1 \in (2,4]) \cap (x_2 \le 3) \\ \text{ふし 4：PND}(\lambda, \mu_4, \sigma) \\ \qquad \text{if} \quad (x_1 \in (2,4]) \cap (x_2 > 3) \\ \text{ふし 5：PND}(\lambda, \mu_5, \sigma) \\ \qquad \text{if} \quad (x_1 > 4) \cap (x_2 \le 3) \\ \text{ふし 6：PND}(\lambda, \mu_6, \sigma) \\ \qquad \text{if} \quad (x_1 > 4) \cap (x_2 > 3) \end{cases} \tag{4.1}
$$

を用いる．

**データの生成：** 本数値検証において，各終結ふし内の生存時間分布は，ベキ正規分布に従う乱数によって生成する．このとき，生存時間分布の差は，位置パラメータの差$\Delta\mu$によって規定する．中途打ち切りは，以下のように設定する．

いま，試験の打ち切り時点 (中途打ち切り時点) を$T^*$とし，試験開始から登録までの時間を$W \sim \mathrm{U}(0, T^*)$とする．ここで$\mathrm{U}(T_1^\dagger, T_2^\dagger)$は，区間$(T_1^\dagger, T_2^\dagger]$の一様分布を表す．$C$を真の終結ふしの数とするとき，$T_a^c, c = 1, \ldots, C$は，$c$番目の真の終結ふしでの生存時間であり，生存時間分布$f(t; \boldsymbol{\theta}^{(c)})$に従うとする．$W$と$T_a^c$が独立の分布に従うことを仮定すると，試験開始から死亡までの時間は，$Z^c = W + T_a^c$で与えられ，実際の生存時間は，死亡のときには$T = T_a^c$で与えられ，生存 (中途打ち切り) のときには$T_a^c = T^* - W$で与えられる．このとき，$Z^c$の密度関数$g(z^*; \boldsymbol{\theta}^{(c)})$と累積分布関数$G(z^*; \boldsymbol{\theta}^{(c)})$は，一様分布と生存時間分布$f(t; \boldsymbol{\theta}^{(c)})$のたたみ込みより

$$g(z^*;\boldsymbol{\theta}^{(c)}) = \frac{1}{T^*}\int_0^{T^*} f(z^{c*}-w;\boldsymbol{\theta}^{(c)})dw,$$

$$G(z^*;\boldsymbol{\theta}^{(c)}) = \frac{1}{T^*}\int_0^{t^{c*}} g(u;\boldsymbol{\theta}^{(c)})du$$

と表すことができる (後藤・松原, 1982；下川他,2002；衛藤 他,2007). 生存時間分布にベキ正規分布を想定するとき, $c$ 番目の終結ふしでの中途打ち切り比率$\eta_c(T^*)$ は

$$\eta_c(T^*) = \begin{cases} \frac{1}{T^* A(\lambda_c)}\int_0^{T^*}\Phi\left[\frac{y^{\lambda_c}-(\lambda_c\mu_c+1)}{\lambda_c\sigma_c}\right]dy \\ \qquad\qquad \times\frac{1}{A(\lambda_c)}\Phi\left[\frac{\lambda_c\mu_c+1}{\lambda_c\sigma_c}\right] & ,\lambda_c>0 \\ \frac{1}{T^* A(\lambda_c)}\int_0^{T^*}\Phi\left[\frac{\log y-\mu_c}{\sigma_c}\right]dy & ,\lambda_c=0 \\ \frac{1}{T^* A(\lambda_c)}\int_0^{T^*}\Phi\left[\frac{y^{\lambda_c}-(\lambda_c\mu_c+1)}{\lambda_c\sigma_c}\right]dy & ,\lambda_c<0 \end{cases}$$

で与えられる. ただし, 樹木構造接近法では, 各終結ふしで中途打ち切り比率が異なるため, 終結ふしの中途打ち切り比率の平均値

$$\eta_0(T^*) = \frac{1}{nT^*}\sum_{c=1}^{C} n_c\left\{1-G(T^*;\boldsymbol{\theta}^{(c)})\right\}$$

を用いる (衛藤 他, 2007). ここに, $n_c$ は, $c$ 番目の終結ふしでの個体数である.

樹木構造接近法の結果に影響を及ぼす要因として, 標本サイズ (S), 中途打ち切り比率 (D), 共変量の個数 (C), 位置ベクトルの差 ($\Delta\mu$), 手法 (M) を選定する. 本稿の数値検証では, 小規模な状況を想定する. そのため, 共変量$p$は離散一様分布 U(1,6) を想定し, 3 (2) 7 個の 3 水準とする. このとき, $R_{max}=6$ とし, 分岐有意水準を0.05とする. 変数間の相関構造を崩すことがないようにコンピュータ志向型実験デザイン (CADEX [Computer Aided Design of EXperimant] ; Kennard & Stone, 1969; Goto & Matsubara, 1979) を利用する. 中途打ち切り比率は 5 (5)15(%) の 3 水準とし, 標本サイズ$n$は, 実地での適用の範囲を考慮して100(50)250の 4 水準とした. 各終結ふしの位置パラメータは, 各ふしの 5 年生存率が等間隔になるように規定したもとで設定する. このとき, 終結ふし $c$ の 5 年生存率は, $p_c = 0.10\times c + 0.05$ および $p_c = 0.10\times c + 0.10$ の 2 水準とする.

**評価の方法：** 事例検討では, 樹木構造接近法によって推定されたモデルの適切性を情報量規準などで評価したが, 既存の多分岐樹木構造接近法では, 分岐過程に検定統計量を用いる. そのため, ここでは, 生成された200組のデータに 2 種類の多分岐樹木構造接近法を個別に適用し, その上で, 真の樹木モデルと推定樹木モデルのハザードの平均平方誤差 (HMSE : Hazard Mean Squared Error)

$$h_{dis} = \frac{1}{n}\sum_{i=1}^{n}\left\{h(t_i;\boldsymbol{\theta}_{i\in r_c}) - \hat{h}(t_i\in r_c^{\dagger})\right\}^2$$

によって評価する ($c=1,\ldots,C, c^*=1,\ldots,C^{\dagger}$). ここに, $h(t_i;\boldsymbol{\theta}_{i\in r_c})$ は, 終結ふし$r_c$ に含まれる$i$ 番目の個体のハザードであり, $\hat{h}(t_{i\in r_c^{\dagger}})$ は樹木構造接近法によって得られた終結ふしにおけるハザードの推定値 (Nelson-Arlen 推定量) である. このとき, 中途打ち切り比率, 共変量の数, 標本サイズ, 手法, 位置パラメータの差といった変動因子が樹木構造接近法の結果に及ぼす影響を ( 4 元分類) 分散分析の結果の p 値と寄与率により定量的に評価する. なお, この分散分析では, 2 因子交互作用までを解釈する.

**結果：** 4 元分類分散分析の結果を表 3 に示す. 標本サイズ (S) の寄与率 (63.43%) が最も高かった. 次いで, 共変量の数数 (C) の寄与率 (20.90%), が相対的に高かった. 手法 (M) の p 値が near0, および寄与率が8.98(%) であることから, 手法による差異が認められた.

図 9 (a) は, 手法に対するウィンドウ・プロッ

表 3: 数値検証

| 因子 | p 値 | 寄与率 |
|---|---|---|
| 標本サイズ (N) | 63.43 | near0 |
| 共変量の個数 (P) | 20.90 | near0 |
| $\mu$ の差 ($\Delta$) | 1.44 | near0 |
| 手法 (M) | 8.98 | near0 |
| N × P | 0.01 | 0.085 |
| N × $\Delta$ | 0.51 | 0.064 |
| N × M | 4.02 | near0 |
| P × $\Delta$ | 0.02 | 0.025 |
| P × M | 0.46 | 0.015 |
| $\Delta$ × M | 0.01 | 0.028 |

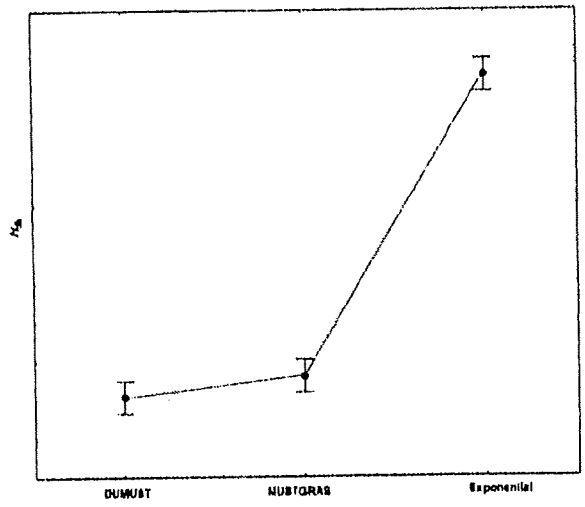
トである. 指数分布を想定した場合には, 極端にHMSEが高かった. DAMUSTとMUSTGRASでは, DAMUSTがMUSTGRASよりも僅かに良好な性能を示した.

さらに, 標本サイズ× 手法の2次交互作用のみが有意だった. また, 標本サイズの主効果は, いずれのモデルで大きく, 共変量の影響も比較的高かった. このことから, 以降では標本サイズおよび共変量の個数と手法の関係を省察する.

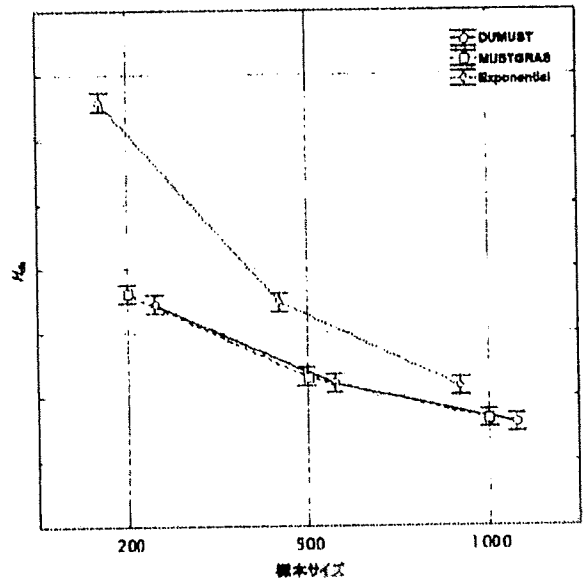
手法と標本サイズの組み合わせ水準におけるHMSEによるウィンドウ・プロットを図9 (b)に示す. 標本サイズが1000のとき, MUSTGRASとDAMUSTの性能には殆ど差異が認められなかったが, 標本サイズが小さくなるにつれて, DUMSTのHMSEはMUSTGRASのHMSEよりも小さくなる傾向にあった.

数値検証の結果, 生存時間分布に指数分布を想定することは, HMSEでの拙い結果を提示した. 生存時間分布がベキ正規分布に従う状況では, DAMUSTはMUSTGRASに比して僅かに良好な性能を示した. この傾向変化は, 標本サイズが小さくなるにつれて顕著であった.

DAMUSTの性能は, MUSTGRASに比して劇的な性能の向上を示さなかった. ただし, 1節で述べたが, DAMUSTの目標はこのような数値検証にみる性能の向上ではない. DAMUST




(a) 手法のみ




(b) 手法と標本サイズの組み合わせ水準

図 9: 数値検証に対するウィンドウ・プロット

では, 終結ふしの生存時間分布にベキ正規分布を想定することで, 得られた結果 (終結ふしにおける生存時間分布) に対する省察、比較、予後要因の探索, さらには臨床試験の動機の提供 (Green, Benedetti & Crowley,2002), あるいはその計画に依存する統計的シミュレーションのデザインといった議論をベキ正規分布の枠組みで行うことができる. したがって, 2手法に大差がないことは, DAMUSTを上述の目標を満たすための手法として適用できる.

## 5. 結びに代えて

本稿では，生存時間分布にデータ適応型分布を想定したもとでRECPAM法を多分岐に拡張した，データ適応型多分岐樹木構造接近法を提示した．これにより，樹木の分岐，最適樹木サイズの決定，および合併過程までをベキ正規分布のもとで一貫して行うことができ，その結果は，得られた結果に基づく臨床研究や統計的シミュレーションに繋げることができる．

データ適応型多分岐樹木構造接近法の性能を，文献事例および数値検証により評価した．事例検討の結果，生存時間分布にデータ適応型分布を想定することが，他の分布 (Weibull 分布および指数分布など) を想定するよりも，適合性能に優れた樹木を構成することができることを提示した．数値検証の結果，既存の検定統計量に基づく多分岐樹木構造接近法に比肩し得る性能が得られた．


### 謝辞

本論文の構成に際し，ご教授いただいた特定非営利活動法人医学統計研究会の先生方，ならびに丁寧な査読を通して貴重なご意見とご指摘を頂戴した本誌2名の審査員と編集理事の先生方に心から深謝の謝意を表します．


### 参考文献


Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19(6)**, 716–723.

Box, G.E.P. & Cox, D.R. (1964). An analysis of transformations (with discussions). *J. Roy. Statist. Soc.*, B**26**, 211–252.

Breiman, L., Friedman, J.H. Olshen, R.A. & Stone. C.J.(1984).*Classification and Regression Trees.* Wadsworth & Brooks.

Breslow, N.(1972). Disussion of regression models and life-tables by Cox, D.R. *J. Roy. Statist. Assoc.*, B**34**, 216–217.

Ciampi, A. & Thiffault, J. (1988). Recursive partition and amalgamation (RECPAM) for censored survival data: Criteria for tree selection. *Statistical software Newsletter,* **14(2)**, 78–81.

Ciampi, A., Hogg, S.A., Mckinney, S. & Thiffault, J. (1989). RECPAM : a computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics. *Computer Methods and Programs in Biomedicine*, **26**, 239–256.

Crowley, J. & Ankerst, D.P.(2005). *Handbook Of Statistics In Clinical Oncology*, Marcel Dekker.

Crowley, J. & Hu, M. (1977). Covariance analysis of heart transplant survival data, *J. Amer. Statist. Assoc.*, **72**, 27–36.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans, *Society of Industrial and Applied Mathematics CBMS-NSF Monographs*, 38.

Goto,M & Matsubara, Y. (1979). Evaluation of ordinary ridge regression. *Bull. Math. Statist.*, **19**, 1–35.

Goto, M., Matsubara, Y. & Tsuchiya, Y. (1983). Power-normal distribution and its applications.*Rep. Statist. Appl. Res., JUSE.*, **30(3)**, 8–28.

Goto, M., Uesaka, H. & Inoue, T. (1979). Some linear models for power transformed data.*Invited paper at The 10th International Biometric Conference*, August, 6–10 (*Res. Rep. No.93, Res. Instit. Fund.*

*Infor. Sc., Kyushu University*).

Goto, M. & Inoue, T. (1980). Some properties of power-normal distribution. *Bulletin of the Biometric Soc.*, 1, 28–54.

Goto, M. & Hamasaki, T. (2002). The bivariate power-normal distribution. *Bulletin of Informatics and Cybernetics*, **34(1–2)**, 29–49.

Green, S., Benedetti, J. & Crowley, J.(2002). *Clinical Trials in Oncology*, CRC Press [福田治彦・新美三由紀・石塚直樹 訳 (2004). 米国 SWOG に学ぶがん臨床試験の実践 — 臨床医と統計家の協調をめざして. 医学書院].

Johnson, N.L., Kots, S. & Balakrishanan, N.(1994). *Continuous Univariate Distributions*, Vol.1, (2nd ed.). John Wiley & Suns.

Hamasaki, T. & Goto, M. (1998). Inferences based on grouped observations from the bivariate power-normal distribution. *Journal of the Japanese Society of Computational Statistics*, **11(1)**, 95–119.

Hastie, T. & Tibshirani, R.(1990). *Generalized Additive Models*, Chapman & Hall.

Hastie, T., Tibshirani, R. & Friedman. J.H.(2001). *The Elements of Statistical Learning: ata mining, inference and prediction.* Springer.

Keles, S.D. & Segal, M.R. (2002). Residual-based tree-structured survival analysis **21**, 313–326.

Kennard, R.W. & Stone, L.A.(1969). Computer aided design of experiments. *Technometrics*, **11**, 137–148.

Kim,H & Loh,W.Y.(2001). Classification trees with unbiased multiway splits. *J. Amer. Statist. Assoc.*, **96**, 589–604.

LeBlanc, M. & Crowley, J.(1992). Relative risk trees for censored survival data. *Biometrics*, **48**, 411–425.

Letón, E. & Zuluaga, P. (2002). Survival tests for $r$ groups, *Biometrical Journal*,**44(1)**, 15–27.

Prentice, R.L. (1974). A log gamma model and its maximum likelihood estimation. *Biometrika*, **61**, 539–544.

Prentice, R.L. (1976). Use of logistic models in retrospective studies.*Biometrics*, **33**,599–606.

Segal. M.R.(1988). Regression trees for censored data. *Biometrics*, **44**,35–47.

Shimokawa, T. & Goto, M. (2006). Multi-split tree strutuerd method in survival analysis. it Contributed CD-ROM of the 23rd of The International Biometric Conference.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc.*, **B36**, 111–147.

Therneau, T. & Atkinson, E.(1997). An introduction to recursive partitioning using the rpart routine. *Technical Report, Section of Biostatistics, Mayo Clinic*, 61.

後藤昌司・山本成志・井上俊昭 (1991). ベキ正規分布のパラメータの推定：推定量の漸近挙動について. 計算機統計学, **4(1)**, 45–60.

衛藤俊寿・下川敏雄・後藤昌司 (2007). 生存時間研究における多分岐型樹木構造接近法. 行動計量学, **34(1)**, 1–20.

後藤昌司・松原義弘 (1982). 比例ハザードモデルとその周辺. 応用統計学, **11(1)**, 1–26.

下川敏雄・杉本知之・後藤昌司 (2002). 生存時間に絡む影響要因の樹木構造とその安定性の評価, 日本行動計量学会第30回大会論文

集, 252–255.

惣田隆生・田崎武僖・後藤昌司 (1992). データ適応型ハザード・モデルのあてはめと診断. 計算機統計学, **5(2)**, 117–126.

濱崎俊光・後藤昌司 (1996). ベキ変換の変換尺度の不変化調整. 計算機統計学, **9(1)**, 37–53.

濱崎俊光・後藤昌司 (2002). 2 変量ベキ正規分布の推測とその評価, 行動計量学, **29(2)**, 199–222.

松原義弘・後藤昌司 (1989). 生存時間解析におけるグラフィカル表現. 応用統計学, **18**, 85–97.

松原義弘・渡辺秀章・後藤昌司 (1990).データ解析における樹木構造表現法の諸法. 日本分類学会シンポジウム 発表抄録, 28 35.

(2008年 9 月13日受付, 2009年 7 月31日採択)

# MULTI-SPLIT TREE STRUCTURED METHOD BASED ON DATA-ADAPTIVE DISTRIBUTION

Toshio Shimokawa* and Masashi Goto**

*Division of Graduate School of Medicine and Engineering, University of Yamanashi,
4-3-11 Takeda, Kofu City 400-8511 Japan
**Biostatistical Research Association, NPO,
4-3-11 Kamishinden, Toyonaka City 560-0085, Japan

In survival analysis, the useful tool for exploration of the factors is tree structured method. Eto *et al.*(2007) have propose the multi-split tree structured method based on $k$-samples generalized rank test statistics (MUSTGRAS). However, these nonparametric approaches do not have consistency in data analysis process. Then, we proposed the methodology of data-adaptive multi-split tree structured method (DAMUST), assuming the power-normal distribution as the survival distribution of each terminal node, where power-normal distribution is defined as the distribution specified before the power-normal transformation.

We evaluated the performance of the DAMUST by some practical examples with survival data and small scale simulation. As a result, DAMUST has better performance than MUSTGRAS. On the other hand, we can evaluate the survival distribution for each terminal nodes based on the power normal distribution using DAMUST by way of simulation, clinical study, etc. Consequently, DAMUST is better useful method than MUSTGRAS.

**Key words:** Multiple interactions, Graphical representation, Power-normal distribution, Akaike's Information Criteria