2018年度、統計関連学会連合大会
９月９日～１３日

単変量Cox回帰にもとづく遺伝子選択と複合共変量による生存期間の予測

発表者；江村剛志（國立中央大學，統計研究所）
松井茂之（名古屋大学大学院、医学部）
Hsuan-yu Chen (中央研究院、統計科学研究所)

本発表の詳細な解説は大会ウェブページに掲載される「詳細論文」
または次の原論文

Emura T, Matsui S, Chen HY (2019), compound.Cox: univariate feature selection and compound covariate for predicting survival, in revision *Computer Methods and Programs in Biomedicine*.

研究の設定＆動機

生存期間 = 応答変数
遺伝子発現量 = 共変量（予後因子）

大量の遺伝子の中から、**生存期間に関連する遺伝子を選択**したい

- 肺がん患者

  ERBB3, LCK, DUSP6, STAT2 (Chen et al., 2006 NEJM)
- 乳がん患者

  ECRG4 (Sabatier et al., 2011, PLoS ONE)
- 卵巣がん患者

  CXCL12 (Popple et al., 2012, British J. of Cancer)

データの解析例

- 肺がん患者

ERBB3,　LCK,　DUSP6,　STAT2

など16の遺伝子が生存期間に関連

(**Chen** et al., 2007 NEJM)

☝データから**単変量Cox回帰**に基づいて、16の関連遺伝子を選択。

使用したデータ;

n＝125人の肺がん患者（台湾の大学病院）

（死亡38人　＋　打ち切りによる生存87人）

p＝485個の遺伝子

（選択された関連遺伝子は16個）

肺がんデータ in 『*compound.Cox*』 **R**パッケージ (Emura et al. 2018)

訓練標本（n＝63）または
テスト標本（n＝62）の指標

```
> library(compound.Cox)
> data("Lung")
> Lung
```

打ち切り ↓

↓*p*=97個の遺伝子（離散値）

生存期間(月)

→→t .vec	d.vec	train	VHL	IHPK1	…	RPL5
1 47.06271	0	FALSE	2	2		4
2 49.27393	0	TRUE	3	4		4
3 20.06601	1	TRUE	2	3		1
4 26.99670	1	TRUE	2	4		2
5 39.90099	0	FALSE	3	4		4
⋮ ⋮	⋮	⋮	⋮	⋮		⋮
125 56.84141	0	FALSE	3	2	…	3

Chen et al.（2007）では
訓練標本（*n*＝63）に単変量Cox回帰を用いて関連遺伝子を選択、
テスト標本（*n*＝62）で関連遺伝子の予測能力を評価

T = 死亡時間

x_j = j番目の遺伝子発現量

T と x_j *関連強➡単変量Cox回帰の推定値が大*

$$h(t \mid x_j) = \frac{\Pr(t \le T < t + dt \mid T \ge t, x_j)}{dt} = h_0(t)\exp(\beta_j x_j)$$

データの形式

$\{ (t_i, \delta_i, \mathbf{x}_i); i = 1, \ldots, n \}$,

- t_i : survival time or censoring time,
- δ_i : censoring indicator ($\delta_i = 1$ if t_i is survival time, or $\delta_i = 0$ if t_i is censoring time),
- $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$: p -dimensional features (genes).

単変量Cox回帰にもとづく遺伝子選択法

Step1: 単変量Coxモデルで全遺伝子を個々に検定（多重検定）

$$H_{oj}: \beta_j = 0 \quad \text{vs.} \quad H_{1j}: \beta_j \neq 0$$

$$h_{0j}(t)\exp(\beta_j x_{ij}), \quad j = 1,\ldots,p$$

（1）ワルド検定；Z値＝ $\hat{\beta}_j / SE(\hat{\beta}_j)$

$\hat{\beta}_j$ = 単変量部分尤度推定値

（2）スコア検定； Z値＝ $S_j / V_j^{1/2}$ ＝単変量スコア統計量÷SD

Step2：棄却された遺伝子を選択

例； P値が0.05以下である遺伝子を選択

Step3：選択された遺伝子の

（1）FDR値を計算（簡便法、Permutation法）

（2）CVL値を計算（部分尤度の*K*分割クロスバリデーション）

Step4：選択された複数の遺伝子で生存期間の予後予測（後述）

本研究では**Steps1-4**を実行するRパッケージ、*compound.Cox* を提案

P値0.05でRパッケージを訓練標本に適用、遺伝子選択

```
> res=uni.Wald(t.vec,d.vec,X.mat)
> res$beta[res$P<0.05]
```

HMMR	LCK	ANXA5	IRF4	STAT2	ERBB3	NF1
0.5156711	-0.8447389	-1.0876762	0.5176704	0.5849869	0.5509026	0.4715235
DLG2	HGF	CPEB4	ZNF264	MMD	RNF4	FRAP1
1.3215044	0.5086750	0.5891676	0.5473276	0.9151541	0.6463635	-0.7696768
STAT1	DUSP6					
-0.5844262	0.7524497					

↑ Chen et al. (2007 NEJM)と同じ結果

- この16の遺伝子の中に、偽陽性はいつくあるか？
  ⇒FDR（False discovery rate）を計算

False Discovery Rate (FDR)とは

FDR=選択された遺伝子中の無関係な遺伝子の割合
(通常、20%以下であることが望ましい(？))

	選択された遺伝子	選択されなかった遺伝子	全遺伝子
関連する遺伝子			
無関係な遺伝子	*f*		
	q=16		***p=97***

FDR = *f*/16

FDR = 0.05 *×97/16* = 0.30　(30%とやや大きい)

FDRに関する一般的注釈

- **Permutation法でも計算可**（Witten & Tibshirani 2010 *SMMR*）
  - *compound.Cox* パッケージで計算を実装

- **FDRは期待値**；
  実際の偽陽性の数や、どれが偽陽性なの不明

- **FDRは選択された遺伝子の予測能力ではない**
  （後述のCVL値を予測能力として使用）

CVL（Cross-validated likelihood）とは

P値の閾値を与えたもとで選択された複数遺伝子の予測能力の指標で、部分尤度の*K*分割クロスバリデーションを用いて、次の式で定義

$$CVL = \sum_{k=1}^{K} \{ \ell(\hat{\gamma}_{-k}) - \ell_{-k}(\hat{\gamma}_{-k}) \},$$

$$\text{where} \quad \hat{\gamma}_{-k} = \arg\max_{\gamma} \ell_{-k}(\gamma),$$

$$\ell(\gamma) = \sum_{i} \delta_i \left[\gamma \mathrm{CC}_{i,-k} - \log \left\{ \sum_{\ell \in R_i} \exp\left(\gamma \mathrm{CC}_{\ell,-k} \right) \right\} \right],$$

$$\mathrm{CC}_{i,-k} = \sum\nolimits_{j \in \Omega_{-k}} w_{j,-k} x_{ij}$$

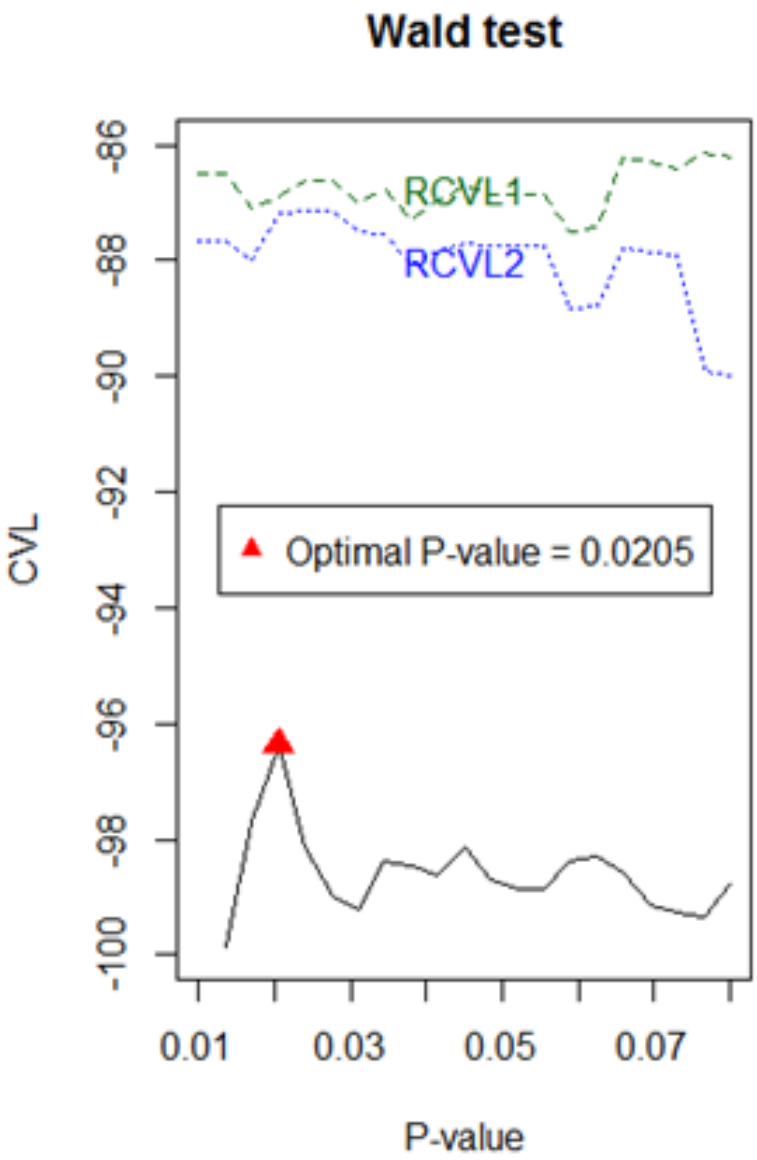
$$\ell_{-k}(\gamma) = \sum_{i \in \Im_{-k}} \delta_i \left[\gamma \mathrm{CC}_{i,-k} - \log \left\{ \sum_{\ell \in R_i \cap \Im_{-k}} \exp\left(\gamma \mathrm{CC}_{\ell,-k} \right) \right\} \right],$$

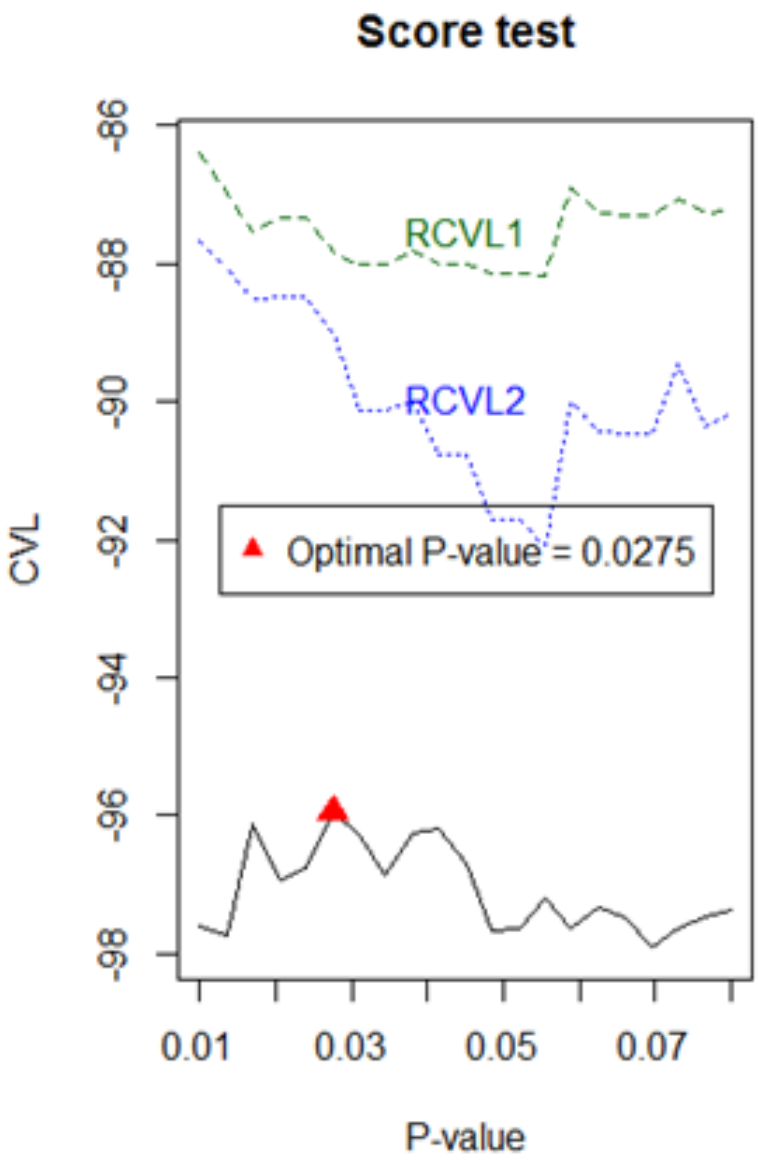
高CVL値➡高予測能力;　詳細は下記文献

Matsui 2006 ; *BMC Bioinformatics*

Emura, Matsui and Chen 2019 *Computer Methods and Programs in Biomed*

CVL値を最適化するP値の閾値を求める





ワルド検定の最適閾値（P＝0.0205）で7個の遺伝子を選択

```
> uni.selection(t.vec,d.vec,X.mat,K=20, P.value=0.0205,score=FALSE)
$beta
ANXA5       DLG2       ZNF264      DUSP6      CPEB4       LCK         STAT1
-1.0876762  1.3215044  0.5473276   0.7524497  0.5891676   -0.8447389  -0.5844262

$CVL   -96.37303
```

↑CVL値

FDR値=0.0205×97/7=0.29（29%）

スコア検定の最適閾値（P＝0.0275）で10個の遺伝子を選択

```
> uni.selection(t.vec,d.vec,X.mat,K=20, P.value=0.0275,score=TRUE)
$Z
ANXA5       DLG2       ZNF264     DUSP6      CPEB4       LCK         STAT1       STAT2
-3.363578   3.111772   2.814363   2.710854   2.538888    -2.511423   -2.445038   2.369334
RNF4        IRF4
2.345912    2.231286
⋮
$CVL     -95.95690
```

↑CVL値

FDR値=0.0275×97/10=0.30（30%）

生存期間の予後予測

- **選択された複数遺伝子の発現量**; $(x_1, \ldots, x_q)$

  スコア検定の最適閾値では　q＝10

- **複合共変量（Compound Covariate）**:

$$\mathbf{CC} = w_1 x_1 + \cdots + w_p x_q$$

β値; $(w_1, \ldots, w_q) = (\hat{\beta}_1, \ldots, \hat{\beta}_q)$

z値; $(w_1, \ldots, w_q) = (z_1, \ldots, z_q)$

- 予後分類:

CC $< c$ ➡ 予後は良い

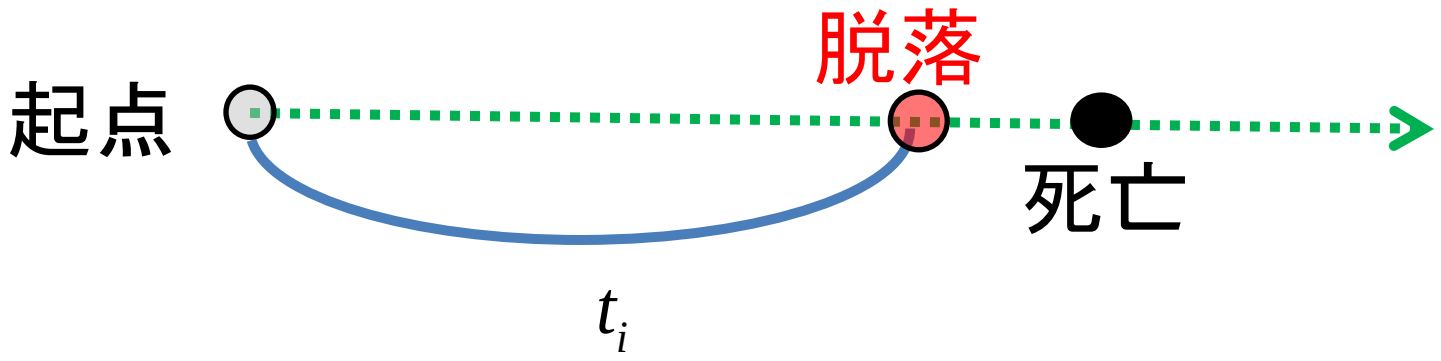
CC $> c$ ➡ 予後は悪い

ここで、c ＝**CC**の閾値

従属センサー（Dependent censoring）問題

死亡の直前に脱落する患者

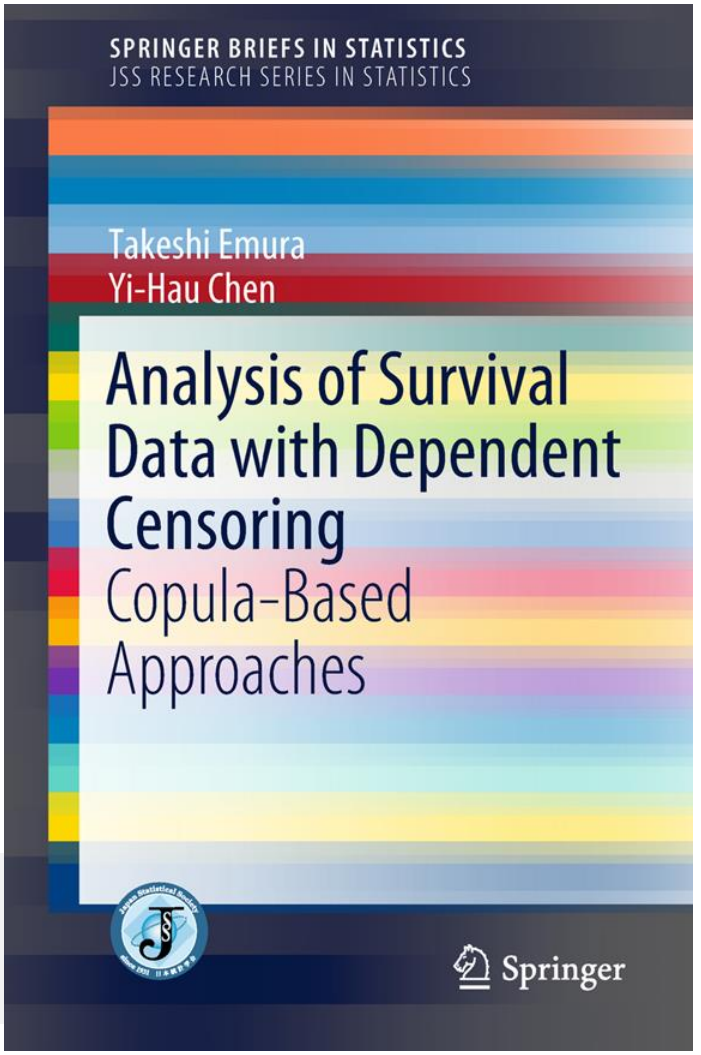
➡センサリングと死亡に相関あり



独立センサーの仮定が崩れ、
Cox回帰の推定値にバイアスが生じる
（Emura & Chen 2016 *SMMR*）

従属センサーのコピュラ・モデル

T = 生存期間（死亡時間）
U = センサー（脱落）時間
x_j = j番目の遺伝子発現量
C_α = コピュラ関数
（αは相関パラメータ）



↓ 2次元生存関数

$$\Pr(T_i > t, U_i > u \mid x_{ij}) = C_\alpha\{\Pr(T_i > t \mid x_{ij}), \Pr(U_i > u \mid x_{ij})\}$$

$$\Pr(T_i > t \mid x_{ij}) = \exp\{-\Lambda_{0j}(t) e^{\beta_j x_{ij}}\}$$

遺伝子 j が T に与える影響

コピュラによる従属センサーへの対応

セミパラメトリック最尤推定量 (Chen 2010, JRSSB)

$$\ell(\beta_j, \gamma_j, \Lambda_{0j}, \Gamma_{0j} \mid \alpha)$$

$$= \sum_i \delta_i [\beta_j x_{ij} + \log \eta_{1ij}(t_i; \beta_j, \gamma_j, \Lambda_{0j}, \Gamma_{0j} \mid \alpha) + \log d\Lambda_{0j}(t_i)]$$

$$+ \sum_i (1-\delta_i)[\gamma_j x_{ij} + \log \eta_{2ij}(t_i; \beta_j, \gamma_j, \Lambda_{0j}, \Gamma_{0j} \mid \alpha) + \log d\Gamma_{0j}(t_i)]$$

$$- \sum_i \Phi_\alpha[\exp\{-\Lambda_{0j}(t_i)e^{\beta_j x_{ij}}\}, \exp\{-\Gamma_{0j}(t_i)e^{\gamma_j x_{ij}}\}],$$

compound.Cox パッケージで計算

$$(\hat{\beta}_j(\alpha), \hat{\gamma}_j(\alpha), \hat{\Lambda}_{0j}(\alpha), \hat{\Gamma}_{0j}(\alpha))$$

予測に使う j 番目の遺伝子の重み w_j を計算

複数遺伝子を使った生存予後予測

1. **Optimal Wald（7個の遺伝子）:**

$$\text{CC} = \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_7 x_{i7}$$

2. **Optimal score（10個の遺伝子）:**

$$\text{CC} = z_1 x_{i1} + \cdots + z_{10} x_{i10}$$

3. **Optimal Wald + copula（7個の遺伝子）:**

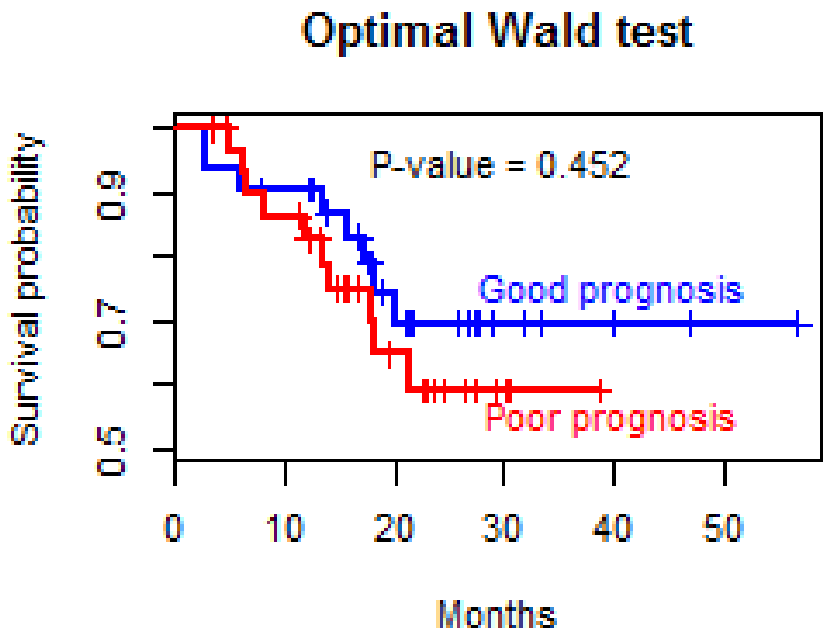
$$\text{CC} = \hat{\beta}_1(\hat{\alpha}) x_{i1} + \cdots + \hat{\beta}_7(\hat{\alpha}) x_{i7}$$

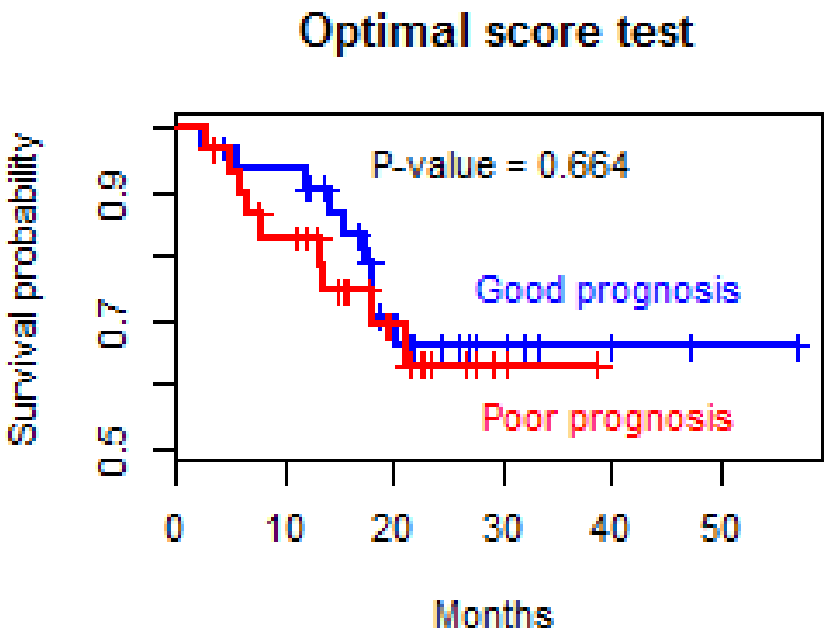
4. **Optimal score + copula（10個の遺伝子）**

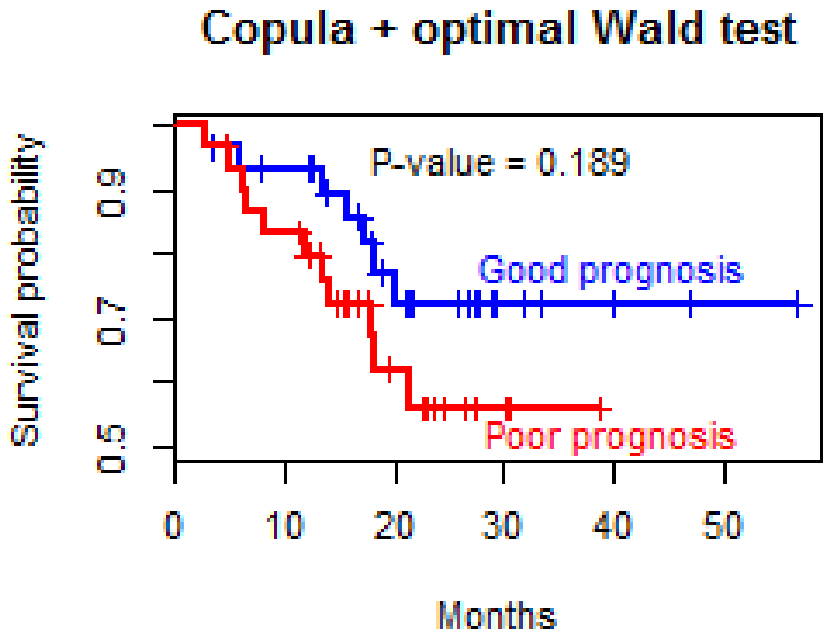
$$\text{CC} = \hat{\beta}_1(\hat{\alpha}) x_{i1} + \cdots + \hat{\beta}_{10}(\hat{\alpha}) x_{i10}$$

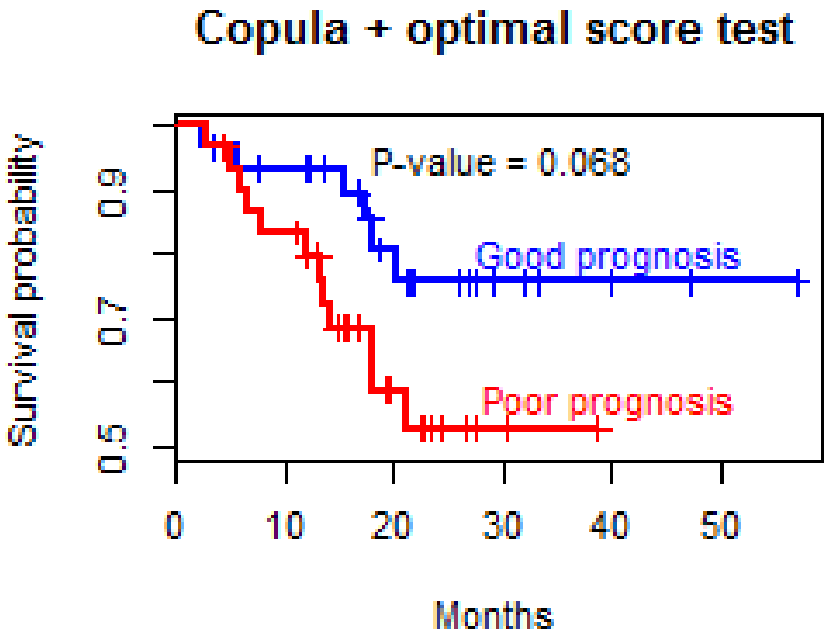
$CC < c$ ➡ 予後は良い（高生存率）

$CC > c$ ➡ 予後は悪い（低生存率）


Optimal Wald test
P-value = 0.452
Good prognosis
Poor prognosis
Survival probability
0.5
0.7
0.9
0
10
20
30
40
50
Months


Optimal score test
P-value = 0.664
Good prognosis
Poor prognosis
Survival probability
0.5
0.7
0.9
0
10
20
30
40
50
Months


Copula + optimal Wald test
P-value = 0.189
Good prognosis
Poor prognosis
Survival probability
0.5
0.7
0.9
0
10
20
30
40
50
Months


Copula + optimal score test
P-value = 0.068
Good prognosis
Poor prognosis
Survival probability
0.5
0.7
0.9
0
10
20
30
40
50
Months

本研究のまとめ

- **Rパッケージ；compound.Cox　を提案**
- 単変量Cox回帰法にもとづく遺伝子選択を多重検定で実行
  （Lassoなど罰則付尤度を使う流行りの手法と異なる）

- **選択された遺伝子の評価法を実装**
  - 予測能力評価（CVL値）　Matsui (2006 *BMC Bioinformatics*)
  - 無関係な遺伝子の割合（FDR値）
    Witten and Tibshirani（2010 *SMMR*）

- **従属センサーの問題への対処法を実装**
- 回帰推定値のバイアス補正　Emura and Chen（2016 *SMMR*）
- コピュラ･グラフィック推定量による、生存関数のバイアス補正
  Rivest and Wells (2001 *JMVA*)

今後の課題

- **単変量Cox回帰の理論的裏付け**

  なぜ単変量回帰が、多変量回帰よりも良いのか？

  シミュレーションや実データで実証されてきたが。。。

- **死亡時間＋中間イベント時間が観測されるケース**

  compound.Cox は利用可能であるが、中間イベントが死亡によって**従属センサー**されるなど問題点もある。

- **動的予測への応用**；（遺伝子）＋（中間イベント情報）の同時利用で生存期間の予測精度を上げる

  例；増悪あり･なしの影響を予測モデルに組み入れる。

  Emura et al.（2018 *SMMR*） 医者や臨床試験に組み込むには。。。

- **平均余命の予測**；

  パラメトリックモデルで患者の生存関数形を特定しているときの遺伝子選択は。。。

  Wu, Michimae and Emura (2019- submitted)

  Shinohara, Emura, Michimae, Takeuchi (2019 in preparation）