107年CSA國際統計學術研討會 **11月９日～10日**

Univariate feature selection and compound covariate for predicting survival

Takeshi Emura

Joint work with

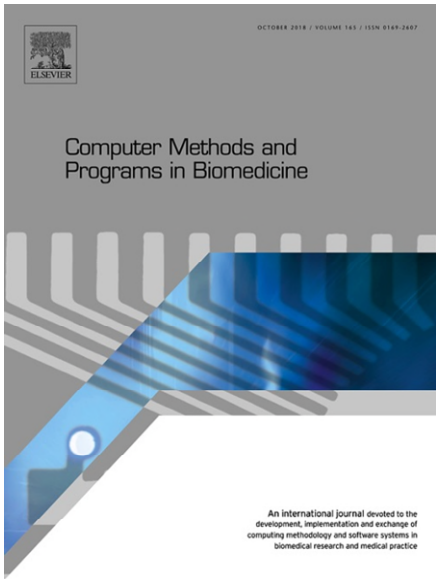
Sigeyuki Matsui & Hsuan-Yu Chen

Emura T, Matsui S, Chen HY (2019),

Computer Methods and Programs in Biomedicine

Volume 168: 21-37

https://doi.org/10.1016/j.cmpb.2018.10.020



Motivation & Setting

Survival = Response
Gene expressions = Features

Objective: Select features associated with survival

- Lung cancer

  ERBB3, LCK, DUSP6, STAT2 (Chen et al., 2006 NEJM)
- Breast cancer

  ECRG4 (Sabatier et al., 2011, PLoS ONE)
- Ovarian cancer

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

Data example

- **Lung cancer patients**

ERBB3, LCK, DUSP6, STAT2, …., etc.

-- 16 genes -- predictive of survival

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

☝ Used univariate feature selection

Dataset

***n*＝125** patients（from Taiwan）

（Death = 38 ＋ Censoring = 87）

***p*＝485** genes

（Only 16 genes selected）

Lung cancer data in 『*compound.Cox*』 package (Emura et al. 2019)

Training sample（n＝63）
Test sample（n＝62）

Survival time (Month) ↓ t .vec

Censor status ↓ d.vec

↓p=97 genes

	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
˜H	˜H	˜H	˜H	˜H	˜H		˜H
125	56.84141	0	FALSE	3	2	…	3

Chen et al.（2007）

Training samples（n＝63）→ Select genes

Test samples（n＝62）→ Evaluate predictive capability of selected genes

T = Survival time

x_j = expression of j-th gene

Association between T & x_j ➡ *Cox regression*

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

Data inputs

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

- 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).

Feature selection with univariate Cox regression

Step1: **Test individual features（multiple tests）**

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

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

（1）Wald test； $Z = \hat{\beta}_j / SE(\hat{\beta}_j)$ $\hat{\beta}_j$ = PL estimate

（2）Score test； $Z = S_j / V_j^{1/2}$ = (Score statistic) ÷ SD

Step2： **Select features**

Example； P-value<0.05

Step3： **Evaluate features**

（1）FDR（False discovery rate）

（2）CVL（Cross-validated likelihood）

Step4： **Survival prediction by selected features**（later）

To do **Steps1-4 automatically, we develop** R package *compound.Cox*

Efficient computation of score tests

- Individual z-statistic for the *j*-th gene

$$S_j = \sum_{i=1}^{n} \delta_i \left(x_{ij} - S_{ij}^{(1)} / S_{ij}^{(0)} \right) \qquad V_j = \sum_{i=1}^{n} \delta_i \left(S_{ij}^{(2)} / S_{ij}^{(0)} - (S_{ij}^{(1)} / S_{ij}^{(0)})^2 \right)$$

where $S_{ij}^{(k)} = \sum_{\ell \in R_i} x_{\ell j}^k$

- Vector z-statistics: $\mathbf{Z} = \mathbf{S} / \mathbf{V}^{1/2}$

$$\mathbf{S} = \boldsymbol{\delta}'(\mathbf{X} - \mathbf{S}^{(1)} / \mathbf{S}^{(0)}) \qquad \mathbf{V} = \boldsymbol{\delta}'(\mathbf{S}^{(2)} / \mathbf{S}^{(0)} - (\mathbf{S}^{(1)} / \mathbf{S}^{(0)})^2)$$

where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)'$

➔ Efficiently programmed in R

Faster computation than Wald tests

Apply *compound.Cox* to the 63 training samples

```
> 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					

↓ Predictor proposed by Chen et al. (2007 NEJM)

The 16-gene predictor = (-1.09×ANXA5) + (1.32×DLG2) + (0.55×ZNF264) + (0.75×DUSP6)

+ (0.59×CPEB4) + (-0.84×LCK) + (-0.58×STAT1) + (0.65×RNF4) + (0.52×IRF4)

+ (0.58×STAT2) + (0.51×HGF) + (0.55×ERBB3) + (0.47×NF1) + (-0.77×FRAP1)

+ (0.92×MMD)+ (0.52×HMMR).

- How many **false positives** among the 16 genes?
  ⇒FDR（False discovery rate）

False Discovery Rate (FDR)

FDR=proportion of false positives (<0.20 is recommended)

	No. of selected features		No. of features
No. of informative features			
No. of non-informative features	*f*		
	q=16		***p=97***

FDR = $f/16$, where f is unknown

FDR = $(0.05 \times 97) / 16 = 0.30$ (simple method)

Remarks on FDR

- **FDR can be computed by**

(1) A simple method (FDR = P-value * No. of tests)

(2) A permutation method (implemented in *compound.Cox*)

$$\mathrm{FDR} = \frac{\text{The expected number of false discoveries}}{\text{The number of rejections}} = \frac{\frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{p} I(P_j^{(m)} < P)}{\sum_{j=1}^{p} I(P_j < P)}$$

$$\frac{\frac{1}{M}\sum_{m=1}^{M}\sum_{j=1}^{p} I(P_j^{(m)} < P)}{\sum_{j=1}^{p} I(P_j < P)} \approx \frac{E\left[\sum_{j=1}^{p} I(P_j^{(m)} < P)\right]}{q} \approx \frac{p \times E[I(P_j^{(m)} < P)]}{q} = \frac{p \times P}{q}$$

- **FDR is just an expectation** ;

  Actual number of false positives is unknown

- **FDR is not a capability of selected features**

CVL（Cross-validated likelihood）

CVL = Index of predictive capability of selected gene under a given P-value.

Defined by a K-fold cross-validation on a partial likelihood

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

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

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

$$CC_{i,-k} = \sum_{j \in \Omega_{-k}} w_{j,-k} x_{ij}$$

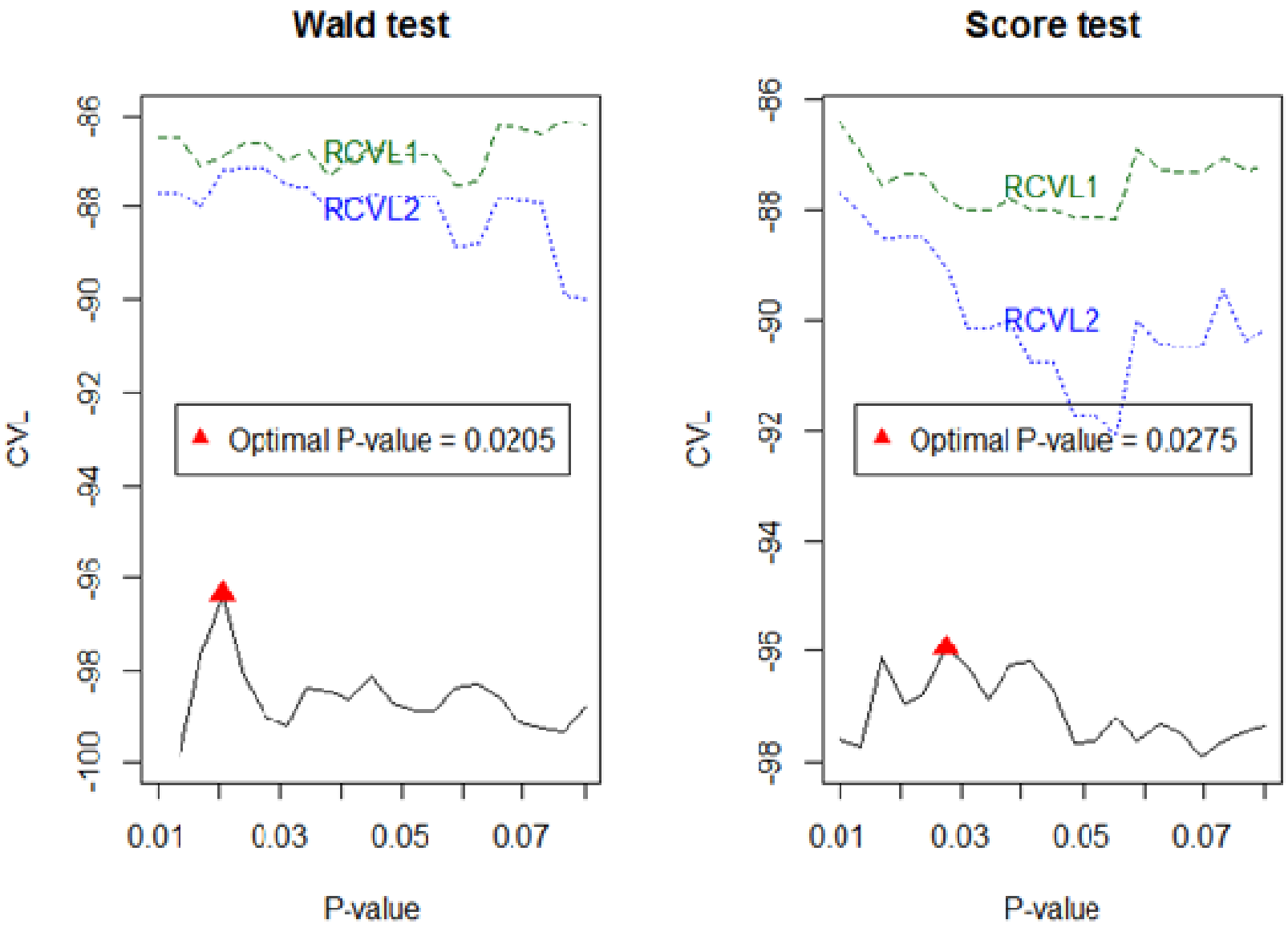
$$\ell_{-k}(\gamma) = \sum_{i \in \Im_{-k}} \delta_i \left[\gamma CC_{i,-k} - \log \left\{ \sum_{\ell \in R_i \cap \Im_{-k}} \exp\left(\gamma CC_{\ell,-k}\right) \right\} \right],$$

High CVL ➡ High predictive capability;

Matsui 2006 ; *BMC Bioinformatics*

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

Optimal P-value cut-off by maximizing CVL



Wald test: Optimal cut-off (P = 0.0205)
➔ 7 Featurs selected

```
> 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

FDR=0.0205 × 97/7=0.29（29%）

Score test: Optimal cut-off (P = 0.0275)
➔ 10 featurs selected

```
> 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
˜H
$CVL    -95.95690
```

↑CVL

FDR=0.0275 × 97/10=0.30（30%）

Prediction of survival

- **Selected features**

$$(x_1, \ldots, x_q) \qquad \text{e.g. } q = 10$$

- **Compound Covariate:**

Ensemble of univariate tests

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

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

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

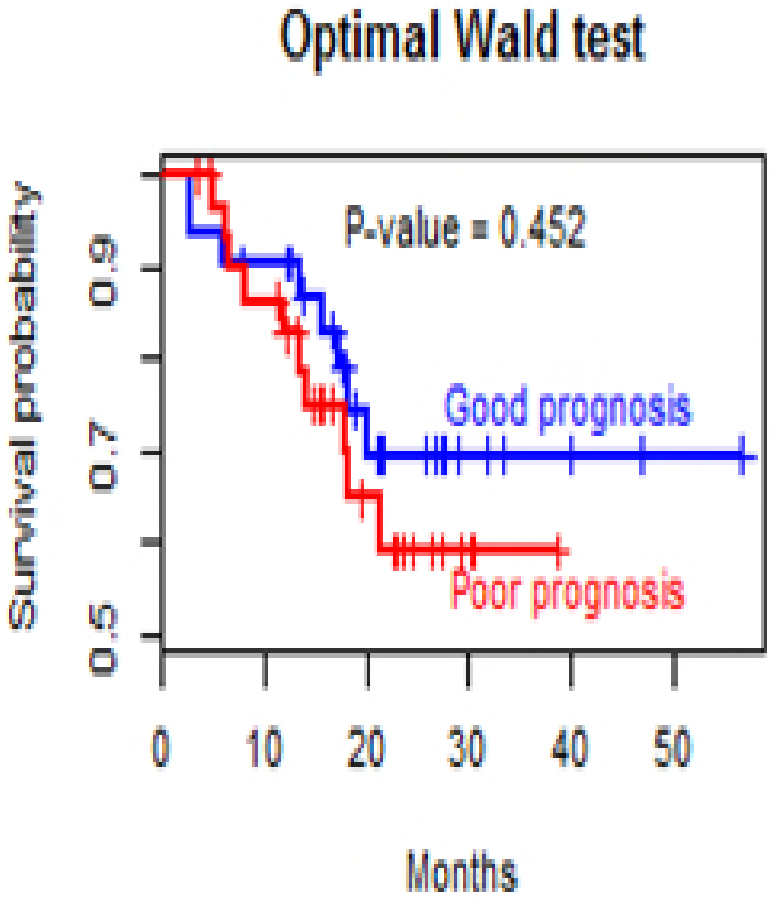
- Classification:

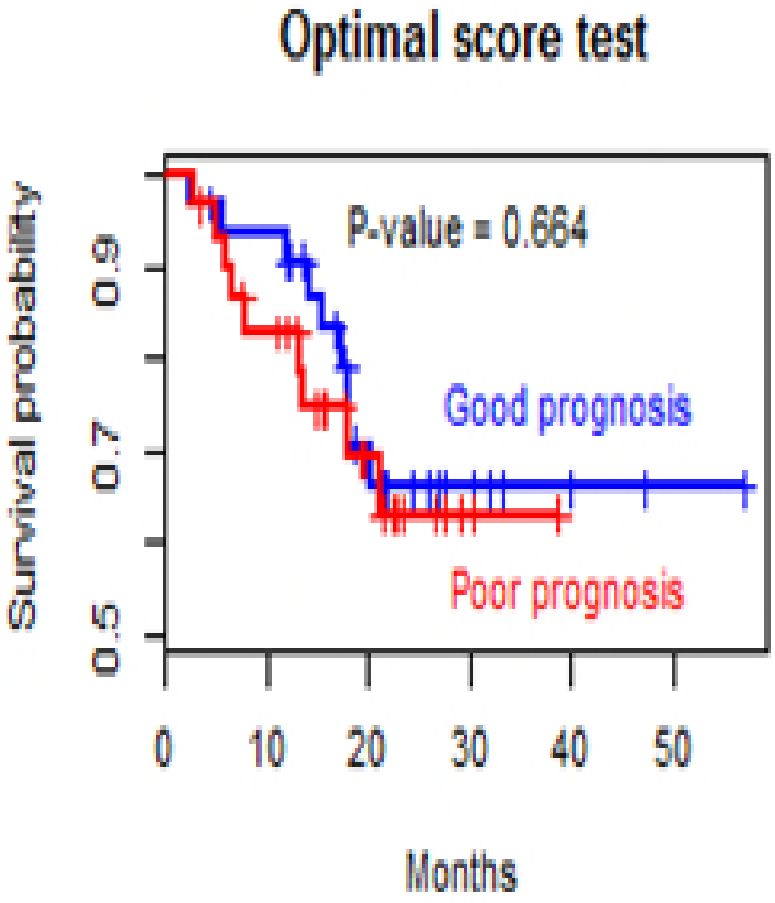
CC < c ➡ Good prognosis

CC > c ➡ Poor prognosis

c = cut-off value

Classification results on *n*=62 test samples

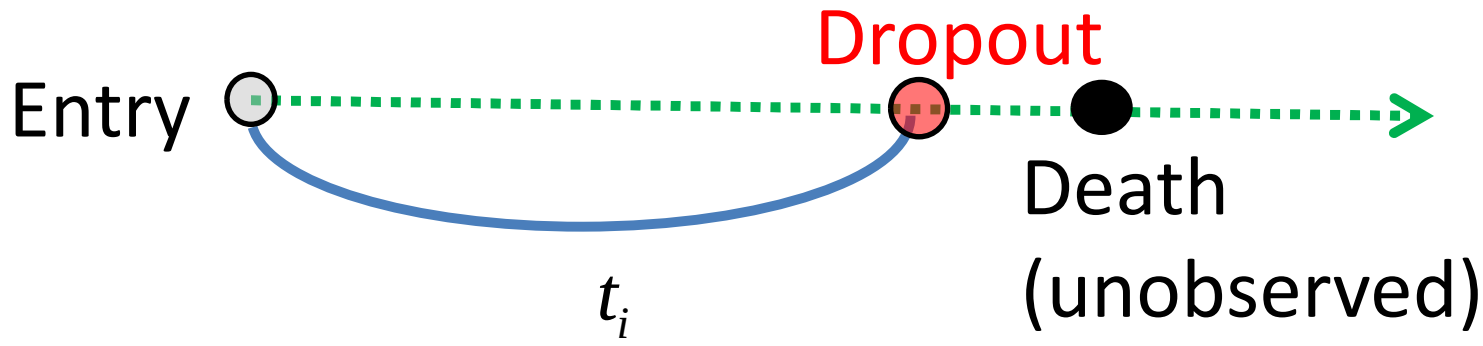




Dependent censoring is suspected

Dropout just before death

➡ Positive dependence between censoring and death



Estimate $\hat{\beta}_j = \arg\max \ell(\beta_j)$ is biased

(Emura & Chen 2016; 2018)

Copula model for dependent censoring

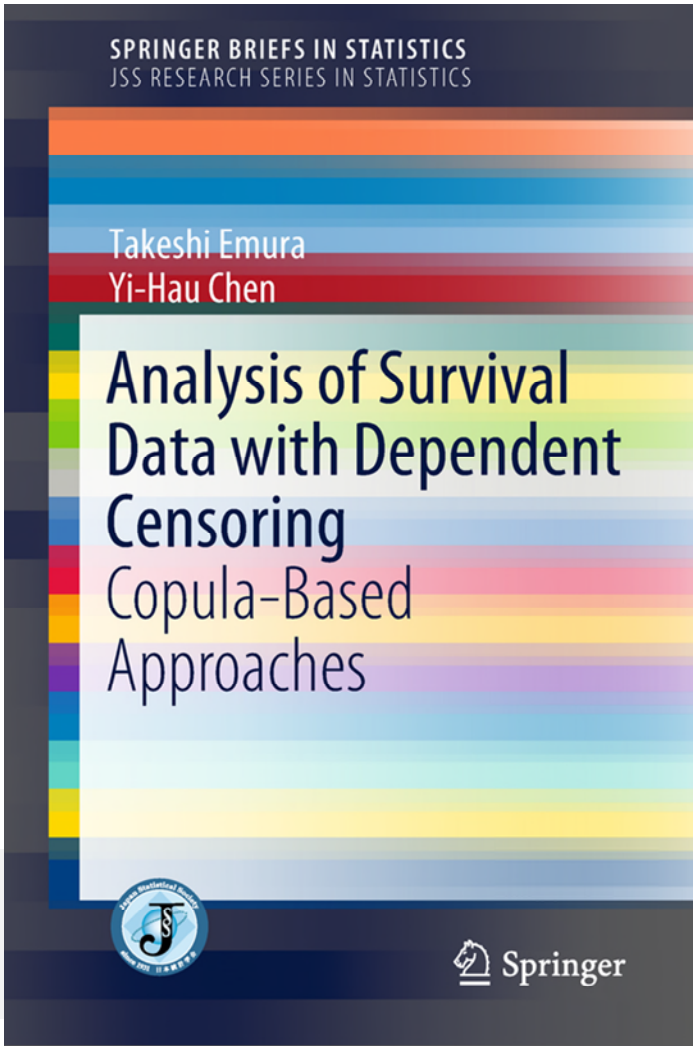
T = Survival time

U = Censoring time

x_j = the j-th gene

C_α = Copula function

（α = copula parameter）



↓ Bivariate survival function

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

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

Effect of j–th gene on T

Estimation under dependent censoring

Semi-parametric MLE (Chen 2010; Emura and Chen 2016)

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

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

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

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

Computed by "*compound.Cox*" R package

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

Compute the weight w_j

Survival prediction

1. **Optimal Wald（7 genes）:**

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

2. **Optimal score（10 genes）:**

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

3. **Optimal Wald + copula（7 genes）:**

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

4. **Optimal score + copula（10 genes ）:**

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

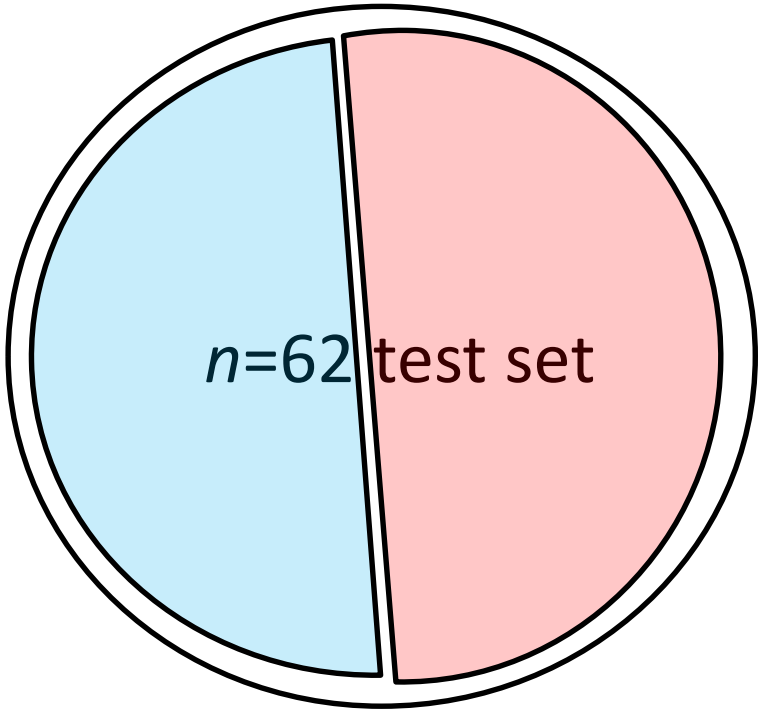
$CC < c$ ➡ Good prognosis（High survival rate）

$CC > c$ ➡ Poor prognosis（Low survival rate）

Test the 4 classifiers by a validation set（n＝62）

Classification rule;

Good prognosis（Low CC） vs. Poor prognosis（High CC）



Diagnosed as Good prognosis

⇓

⇓

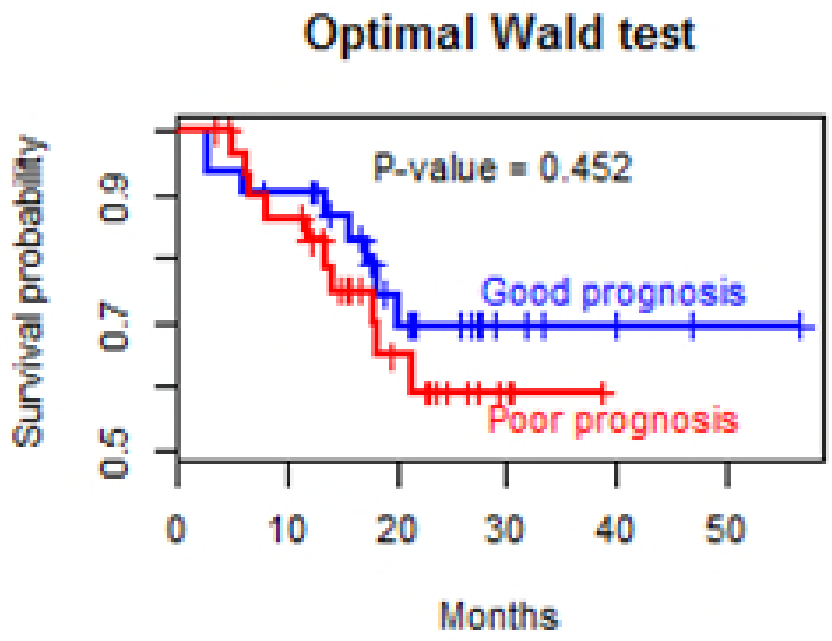
Compute actual survival rate (K-M estimator)

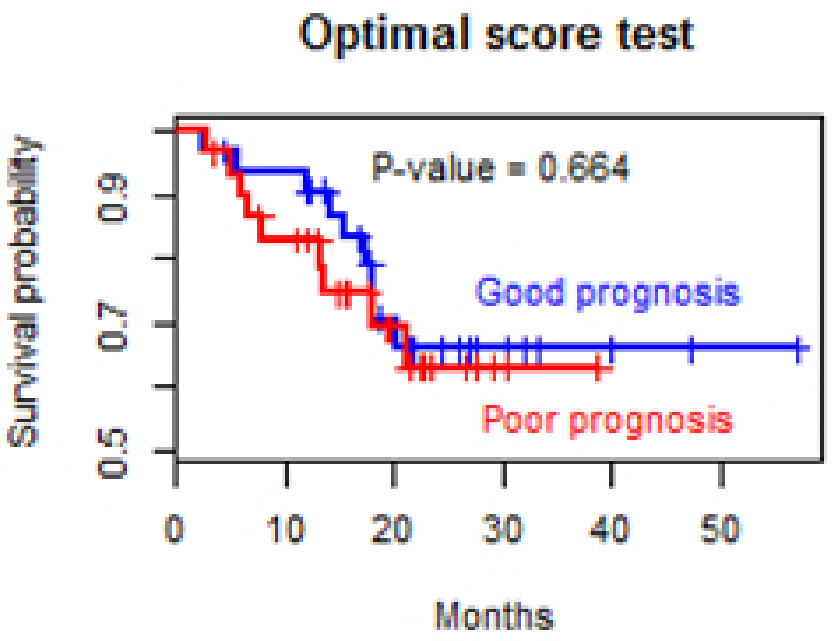
Diagnosed as Poor prognosis

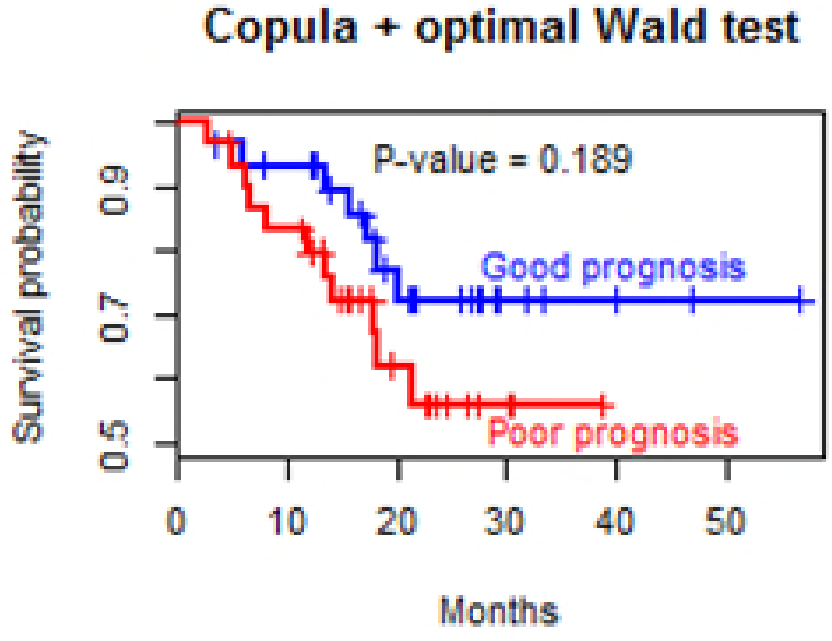
⇓

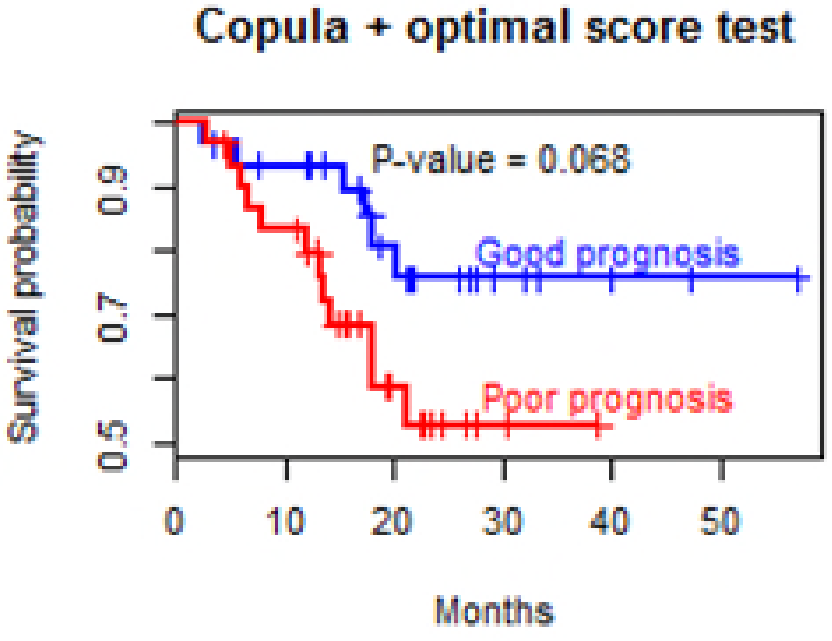
⇓

Compute actual survival rate (K-M estimator)


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

Summary

- **Developed an R package "*compound.Cox*"**
  - Use **multiple tests** for feature selection (frequently used in medical research)
  - Use **compound covariate** for prediction (an ensemble of multiple tests)
  - Very different from Lasso type method
  - Use a **vector computation** of score tests (new method)
- **Implemented the evaluation measures:**
  - Predictive capability (CVL) Matsui (2006)
  - False discovery rate (FDR) Witten and Tibshirani (2010)
- **Used copula for deal with dependent censoring**
  - More accurate predictor if censoring is informative

References

[1] Witten M, Tibshirani R. Survival analysis with high-dimensional covariates. Statist Method Med Res 2010; 19: 29-51.

[2] Wang Y, Klijn JG, Zhang Y, Sieuwerts, AM, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet 2005; 365(9460): 671-79.

[3] Matsui S. Predicting survival outcomes using subsets of significant genes in prognostic marker studies with microarrays. BMC Bioinformatics 2006; 7:156.

[4] Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. N Engl J Med 2007; 356: 11-20.

[5] Yoshihara K, Tajima A, Yahata T, Kodama S, Fujiwara H et al (2010). Gene expression profile for predicting survival in advanced-stage serous ovarian cancer across two independent datasets. PloS One 5(3), e9615

[6] Emura T, Nakatochi M, Matsui S, Michimae H, Rondeau V, Personalized dynamic prediction of death according to tumour progression and high-dimensional genetic factors; meta-analysis with a joint model, Statist Methods Med Res 2018; 27:2842-2858.

[7] Waldron L, Haibe-Kains B, Culhane AC, Riester M, Ding J, et al. Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. J Natl Cancer Inst 2014 106(5): dju049

[8] Emura T, Chen YH, Chen HY. Survival prediction based on compound covariate under Cox proportional hazard models. PLoS One 2012; 7(10): e47627. DOI:10.1371/journal.pone.0047627.

[9] Emura T, Chen YH, Gene selection for survival data under dependent censoring, a copula-based approach, Statist Method Med Res 2016; 25(6): 2840-2857.

[10] Emura T, Chen YH, Analysis of Survival Data with Dependent Censoring, Copula-Based Approaches, JSS Research Series in Statistics, Springer, Singapore; 2018.