

Compound Cox: univariate feature selection and compound covariates for predicting survival

Takeshi Emura

Graduate Institute of Statistics, NCU

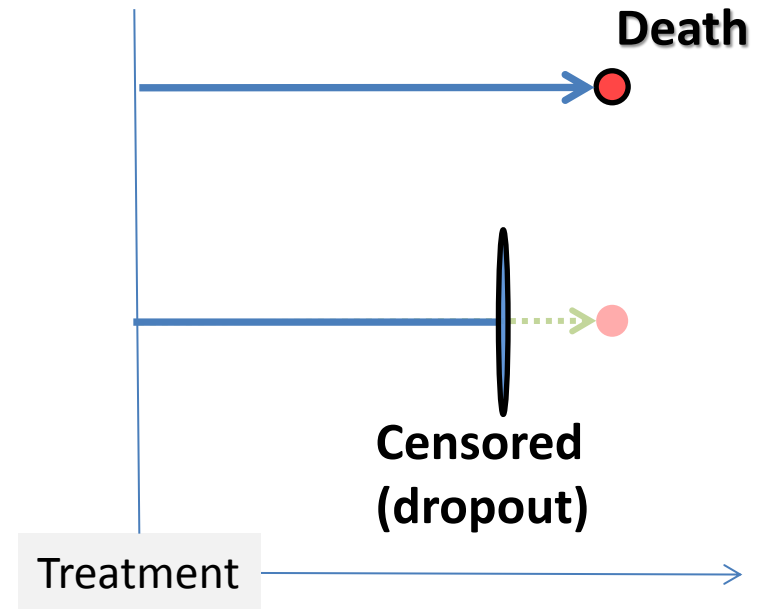
**3rd Pacific Rim Cancer Biostatistics
Conference
6/27-28, 2019**

- Survival Data (right-censored)

t_i : time - to - death or censoring

$$\delta_i = \begin{cases} 1 & \text{if death} \\ 0 & \text{if censoring} \end{cases}$$

$\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$, possibly $p > n$



t_i Time-to-event	δ_i Censor	x_{i1} AP3S1	x_{i2} APMAP	x_{i3} ARHGAP28	x_{i4} CXCL12		x_{i127} ASB7	$x_{i,128}$ B4GALT5
1650	0	-0.52	1.12	-0.37	1.30		0.354	-1.015
30	1	-0.18	-0.69	-0.93	1.28		0.026	0.38
⋮	⋮							
1800	1	-1.08	0.70	-0.29	-0.529		-0.50	-1.09

↑
Short Survival

↑
High Expression

P-value < 0.05, but....

Errors in multiple tests

- $\alpha=0.05$ is the error rate of **ONLY** one test
100 tests $\rightarrow 0.05 \times 100 = \underline{5 \text{ rejections}}$
(False Discoveries)

In statistical process control (Montgomery 2009)

- **3-sigma** rule, $\alpha=0.0023$
- set α to be **ARL=370**

In design of microarray experiments (Simon 2001)

- $\alpha=0.001$, one error in 1000 tests
- Set α to be **FDR=0.20**

Today's points

- “death” may be unobservable
→ Deal with ***censoring (or dropout)***
Tool: Cox's partial likelihood (Cox 1972)

- A ***large*** number of features
→ Select a subset of ***features***
Tool: Multiple significance tests
Tool: False discovery rate (Witten & Tibshirani 2010)
Tool: Cross-validated likelihood (CVL) (Matsui 2006)

- A multigene predictor
Tool: Tukey's compound covariate
(Tukey 1993; Shyr and Kim 2003; Matsui 2006; Emura et al. 2012)

We developed “*compound.Cox*” R package

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>

Tools

- Lung cancer data
- Feature selection
- Multiple tests
- Predictor calculation
- Prediction error (CVL)
- False discovery rate (FDR)
- Copula methods



Lung cancer data in *Lung* object

Training samples (n=63)

Test samples (n=62)

```
> library(compound.Cox)
```

```
> data("Lung")
```

```
> Lung
```

	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

Censor

$p=97$ features

Survival time (months)

Univariate Cox regression

T = Survival time

x_j = j -th feature

Proportional hazards model (Cox 1972)

$$\Pr(t \leq T < t + dt \mid T \geq t, x_j) = h_0(t) \exp(\beta_j x_j) dt$$

Partial likelihood estimator (Cox regression)

$$\hat{\beta}_j = \arg \max \ell_j(\beta_j)$$

$$\ell_j(\beta_j) = \sum_{i=1}^n \delta_i \left[\beta_j x_{ij} - \log \left(\sum_{\ell \in R_i} \exp(\beta_j x_{\ell j}) \right) \right]$$

$$SE(\hat{\beta}_j) = [-\partial^2 \ell_j(\beta_j) / \partial \beta_j^2]^{-1/2}$$

Univariate significance tests

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

(1) Wald test:

$$\text{Z-value: } z_j = \hat{\beta}_j / SE(\hat{\beta}_j) \sim N(0,1)$$

$$\text{P-value: } P_j = \Pr(|Z| > |z_j|)$$

(2) Score test:

$$S_j = \sum_{i=1}^n \delta_i \left(x_{ij} - \bar{x}_j(t_i) \right) \quad V_j = \text{Var}(S_j) = \sum_{i=1}^n \delta_i \left(\bar{x}_j^2(t_i) - (\bar{x}_{ij}(t_i))^2 \right)$$

$$\text{Z-value: } z_j = S_j / \sqrt{V_j} \sim N(0,1)$$

$$\text{P-value: } P_j = \Pr(|Z| > |z_j|)$$

-Simple algebraic computation

Score tests via matrix (Matsui's idea at 2nd PRCBC)

- Z-values: $\mathbf{Z} = \mathbf{S} / \mathbf{V}^{1/2}$

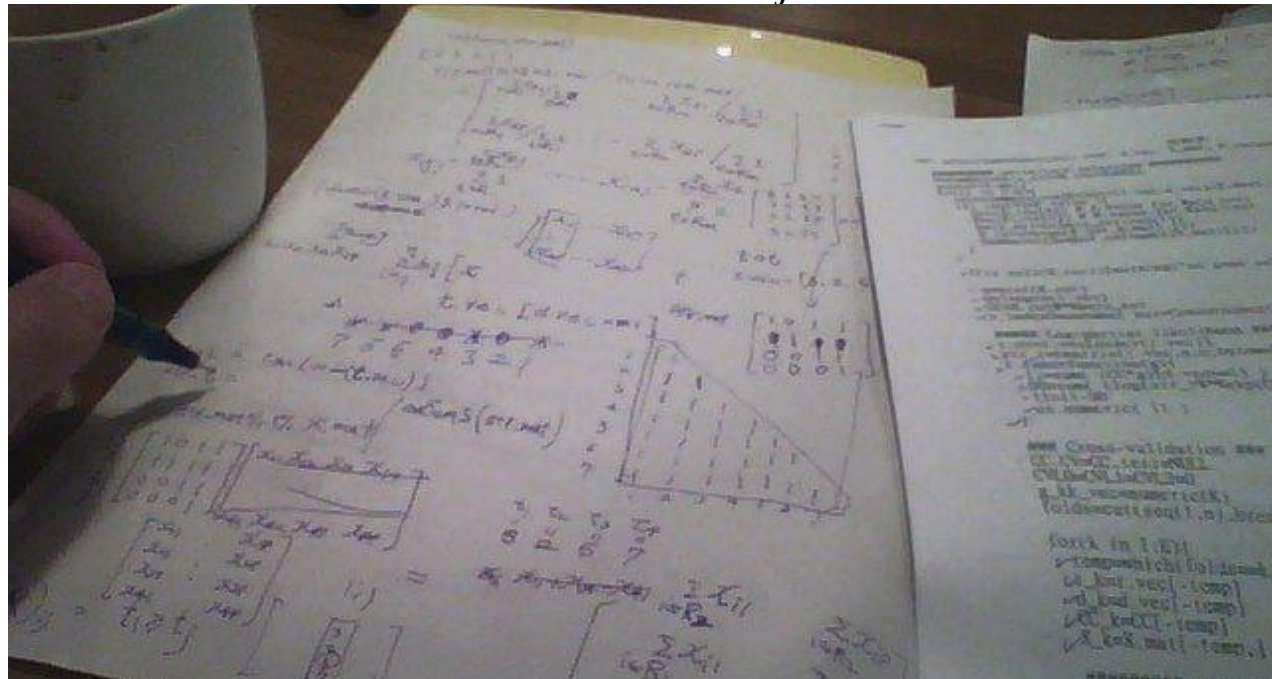
$$\mathbf{S} = \delta'(\mathbf{X} - \mathbf{S}^{(1)} / \mathbf{S}^{(0)})$$

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

$$\delta = (\delta_1, \dots, \delta_n)'$$

$\mathbf{S}^{(k)}$ is a $n \times p$ matrix with elements $S_{ij}^{(k)}$

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



Whole algorithms

Step 1: Test $H_{0j} : \beta_j = 0$ vs. $H_{1j} : \beta_j \neq 0$

Select a feature (via P-value < 0.05)

Step 2 : Critical evaluation of selected features

(1) False negatives ?

(2) Predictive capability ?

Step 3 : Compound covariate predictor

Z-value: $z_1x_1 + z_2x_2 + \cdots + z_px_q$

β -value: $\hat{\beta}_1x_1 + \hat{\beta}_2x_2 + \cdots + \hat{\beta}_qx_q$

Step 4 : Test the predictor by independent test samples

“compound.Cox” performs the whole algorithms

uni.selection()

features ($n \times p$ matrix)

Survival Time

Censor

P-value

```
> uni.selection(t.vec,d.vec,X.mat,K=20,P.value=0.05,score=FALSE,permutation=TRUE) ## Wald test
```

\$beta

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

\$Z ← Z -value

.....

\$P ← P-value

.....

\$CVL

-98.66365 CVL: Cross-validation

\$FDR

P.value * (No. of genes)

0.3031250

Permutation

0.3128125

← FDR: False Discovery Rate

Regression coefficients

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

We selected **16 features** ($P < 0.05$), but...

```
> uni.selection(t.vec,d.vec,X.mat,K=20,P.value=0.05,score=FALSE,permutation=TRUE)
```

\$beta

ANXA5	DLG2	ZNF264	DUSP6	CPEB4	LCK	STAT1
-1.087	1.321	0.547	0.752	0.589	-0.844	-0.584
RNF4	IRF4	STAT2	HGF	ERBB3	NF1	FRAP1
0.646	0.517	0.584	0.5086	0.550	0.471	-0.769
MMD	HMMR					
0.915	0.515					

- **Q1: Is there any FALSE feature?**
⇒ Compute FDR (False discovery rate)
- **Q2: Is “P-value < 0.05” optimal ?**
⇒ Compute CVL (Cross-validated likelihood)
- **Q3: Any biased due to dropout ?**
⇒ Adjust the bias by a copula method

False Discovery Rate (FDR)

FDR= Proportion of false rejection
= $E[\mathbf{f}/16]$, where $\mathbf{f}$ is unknown

	Rejected	Accepted
$\beta \neq 0$	16- $\mathbf{f}$	
$\beta = 0$	$\mathbf{f}$	
	$q=16$	$p=97$

→ **FDR** and $E[\mathbf{f}]$ can be estimated by
a permutation method (Witten & Tibshirani 2010)

Random M permutations (Witten & Tibshirani 2010)

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

$P_j^{(m)}$ is the P-value for testing $H_{0j} : \beta_j = 0$ vs. $H_{1j} : \beta_j \neq 0$

- **FDR does not guarantee predictive ability**
(small FDR $\rightarrow$  $\rightarrow$ high prediction ability)

CVL (Cross-validated likelihood)

: predictive capability of selected features

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

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

CVL = - (Allen's PRESS)
for Gaussian likelihood &
uncensored data (Allen 1974)

Test-data
Likelihood $\rightarrow$

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



Training-data
likelihood $\rightarrow$

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

High CVL $\Rightarrow$ High prediction (classification) capability

Matsui 2006; Emura, Matsui and Chen 2019

$$\begin{bmatrix} t_1, \delta_1, x_{11}, x_{12}, \dots, x_{1p} \\ \vdots \\ t_n, \delta_n, x_{n1}, x_{n2}, \dots, x_{np} \end{bmatrix} : \text{survival data with } n \text{ samples}$$

⇓ **Step 1:** Divide between "testing" vs. "training" samples

$$\left. \begin{bmatrix} t_1, \delta_1, x_{11}, x_{12}, \dots, x_{1p} \\ \vdots \\ \text{-----} \\ \vdots \\ t_n, \delta_n, x_{n1}, x_{n2}, \dots, x_{np} \end{bmatrix} \right\} \begin{array}{l} \frac{n}{K} \text{ testing samples, } i \in \mathfrak{T}_k \\ \text{-----} \\ n - \frac{n}{K} \text{ training samples, } i \in \mathfrak{T}_{-k} \end{array}$$

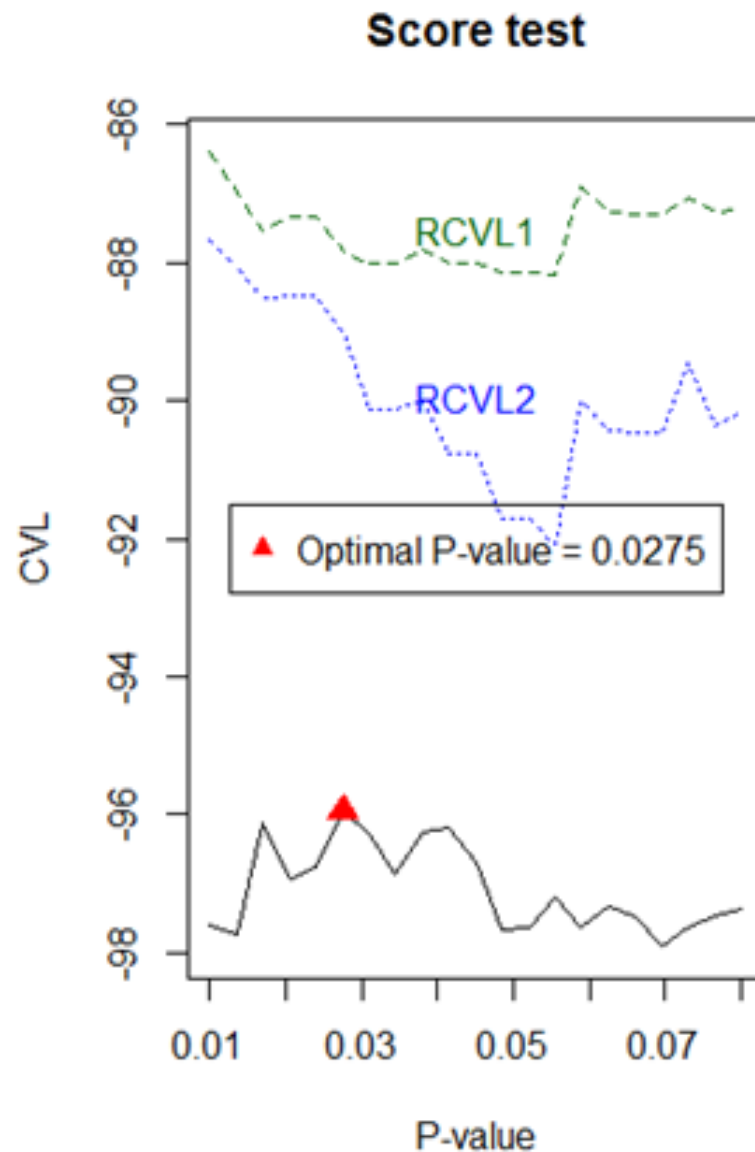
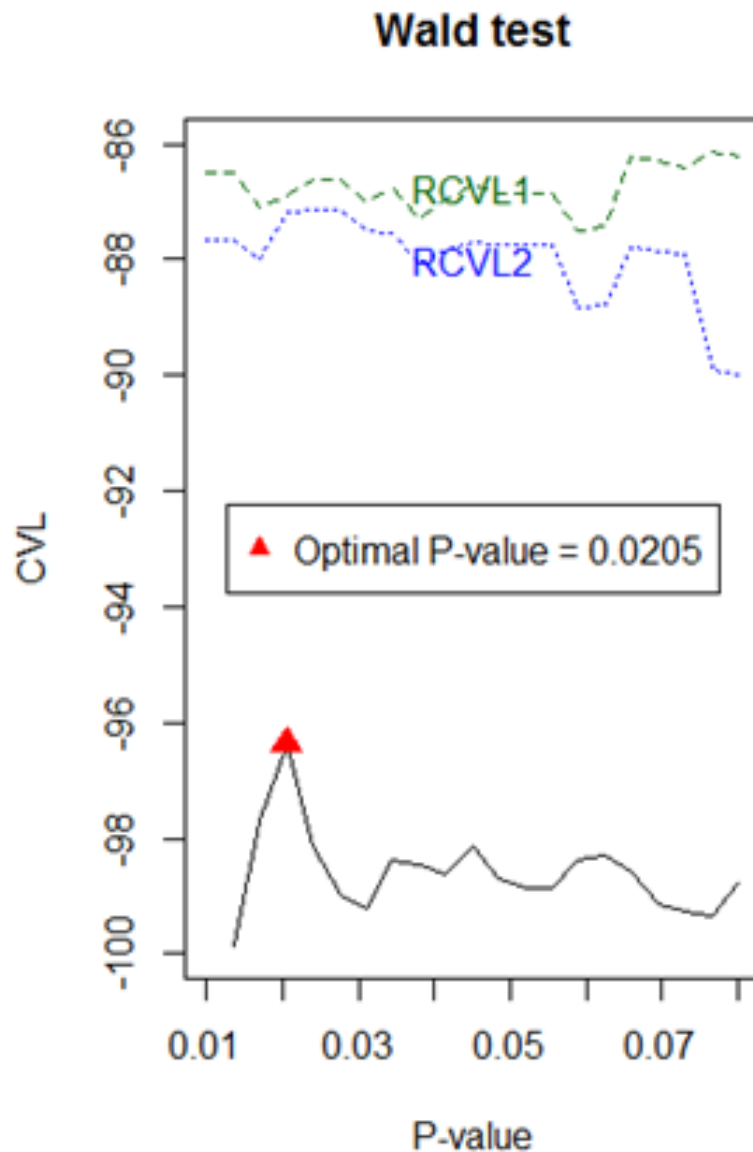
⇓ **Step 2:** Feature Selection & Prediction (P-values < P)

$$\begin{bmatrix} t_1, \delta_1, CC_{1,-k} \\ \vdots \\ t_m, \delta_m, CC_{m,-k} \end{bmatrix}, \quad CC_{i,-k} = w_{1,-k}x_{i1} + w_{2,-k}x_{i2} + \dots + w_{q,-k}x_{iq}, \quad i \in \mathfrak{T}_{-k}$$

⇓ **Step 3:** Average (Predicted likelihood - Baseline)

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

The optimal **P-value** threshold by the CVL plot



Optimal Wald tests ($P < 0.0205$)

→ Select 7 features (genes)

```
> 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

$FDR = 0.0205 \times 97/7 = 0.29$ (29%)

Optimal score tests $P < 0.0275$)

→ Select 10 features (genes)

```
> 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

$FDR = 0.0275 \times 97/10 = 0.30$ (30%)

Prediction & Classification

- **Selected Features;** $(x_1, \dots, x_q)$

e.g. $q = 10$ in the score tests

- **Compound Covariate:**

$$CC = w_1 x_1 + \dots + w_p x_q$$

Wald test: $(w_1, \dots, w_q) = (\hat{\beta}_1, \dots, \hat{\beta}_q)$

Score test: $(w_1, \dots, w_q) = (z_1, \dots, z_q)$

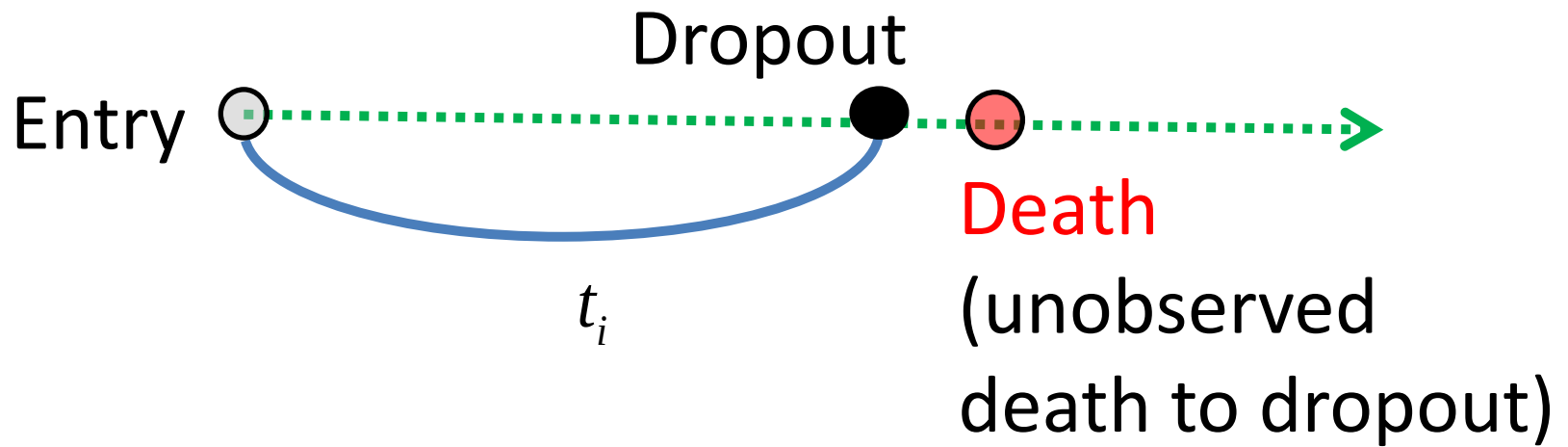
- **Classification** $CC < c \Rightarrow$ Good prognosis (Long survival)
 $CC > c \Rightarrow$ Poor prognosis (Short survival)

 cut-off value

Are weights (beta) unbiased?

Many dropout censors before death

➡ Dependence between censoring and death



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

Model of Dependent Censoring

T = Survival (death) time

U = Censoring (dropout) time

x_j = j -th feature

C_α = Copula; α = copula parameter

⇓ Joint 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 feature on T

↑ Marginal survival function

Estimation under dependent censoring

Semi-parametric MLE (Emura and Chen 2016)

$$\begin{aligned} & \ell(\beta_j, \gamma_j, \Lambda_{0j}, \Gamma_{0j} | \alpha) \\ &= \sum_i \delta_i [\beta_j x_{ij} + \log \eta_{1ij}(t_i; \beta_j, \gamma_j, \Lambda_{0j}, \Gamma_{0j} | \alpha) + \log d\Lambda_{0j}(t_i)] \\ &+ \sum_i (1 - \delta_i) [\gamma_j x_{ij} + \log \eta_{2ij}(t_i; \beta_j, \gamma_j, \Lambda_{0j}, \Gamma_{0j} | \alpha) + \log d\Gamma_{0j}(t_i)] \\ &- \sum_i \Phi_\alpha [\exp \{ -\Lambda_{0j}(t_i) e^{\beta_j x_{ij}} \}, \exp \{ -\Gamma_{0j}(t_i) e^{\gamma_j x_{ij}} \}], \end{aligned}$$

Computed by “*compound.Cox*” R package


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



Compute weight w_j the CC

Survival prediction

1. Optimal Wald (7 features):

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

2. Optimal score (10 features) :

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

3. Optimal Wald + copula (7 features) :

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

4. Optimal score + copula (10 features) :

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

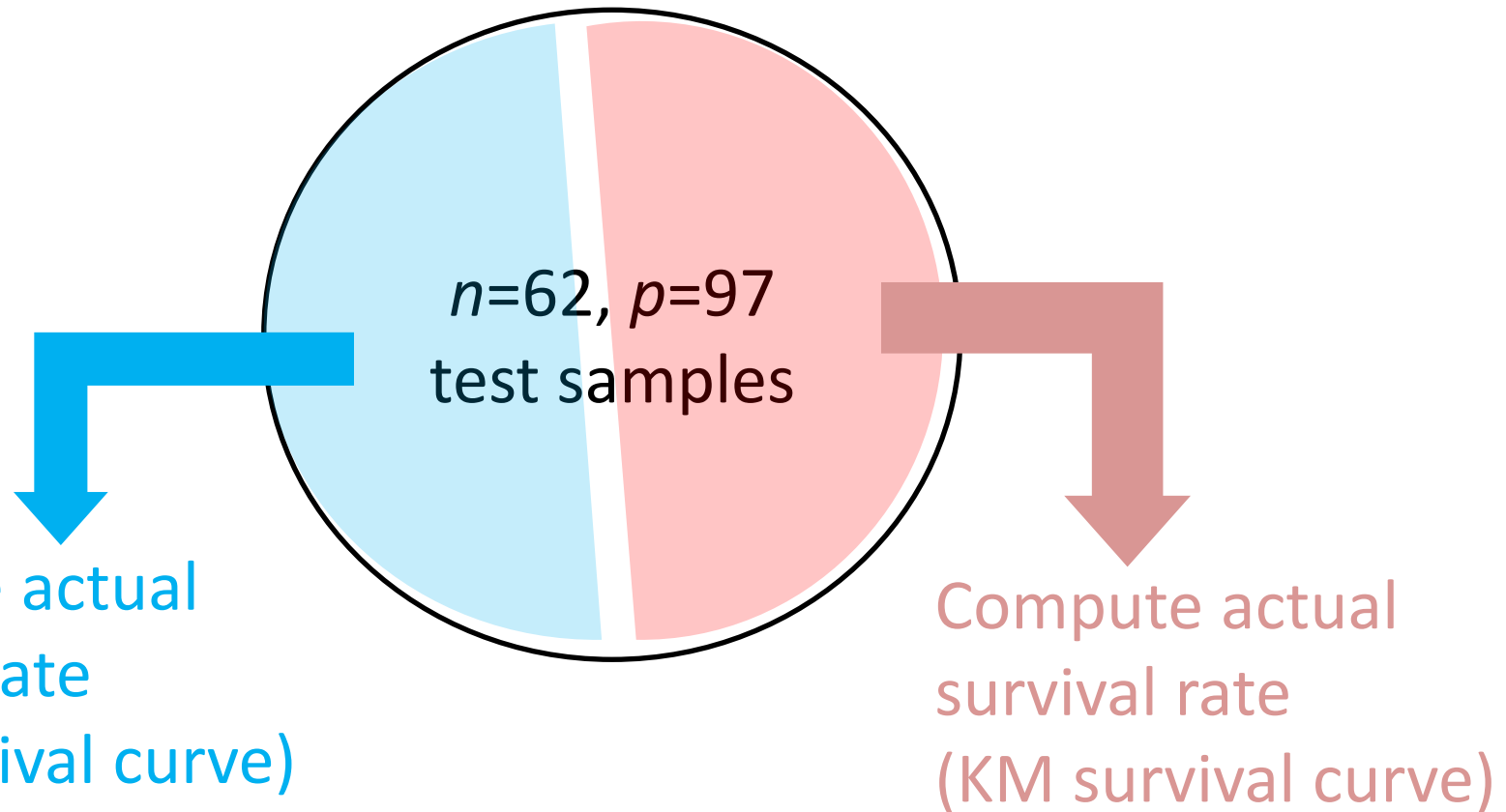
$CC < c \Rightarrow$ Good prognosis (High survival rate)

$CC > c \Rightarrow$ Poor prognosis (Low survival rate)

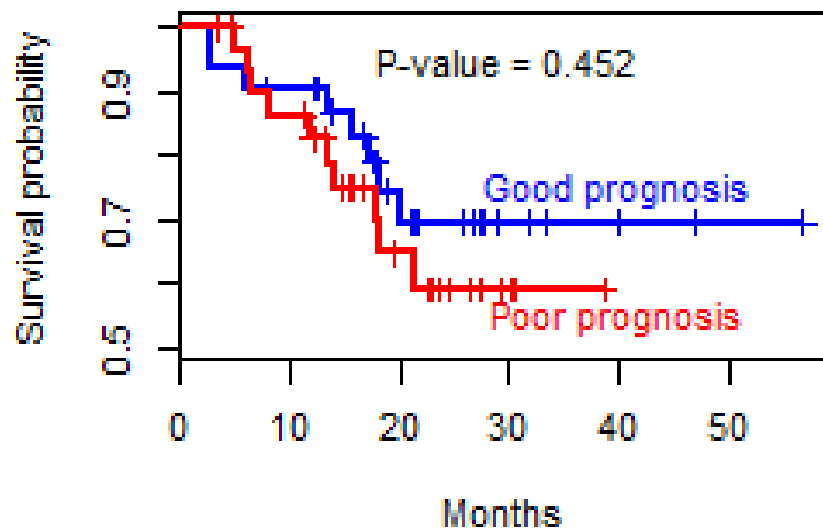
Classify the **test** samples ($n=62$)

Class 1 ; Good prognosis (Low CC)

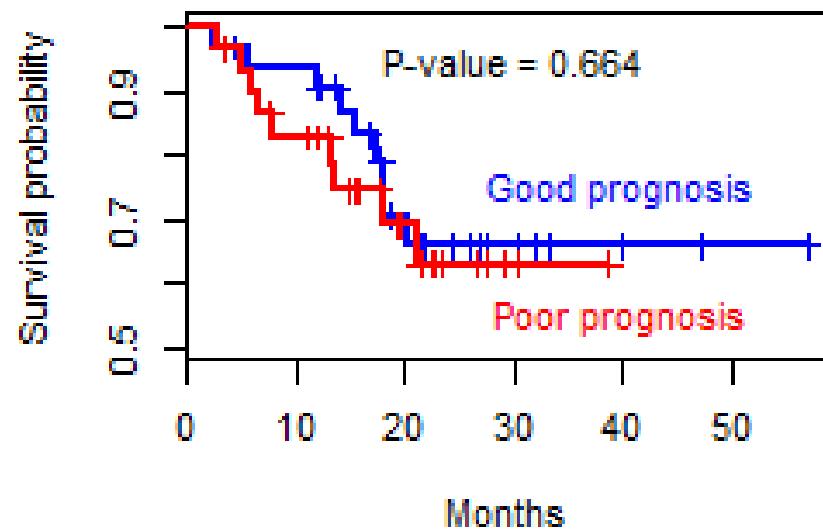
Class 2: Poor prognosis (High CC)



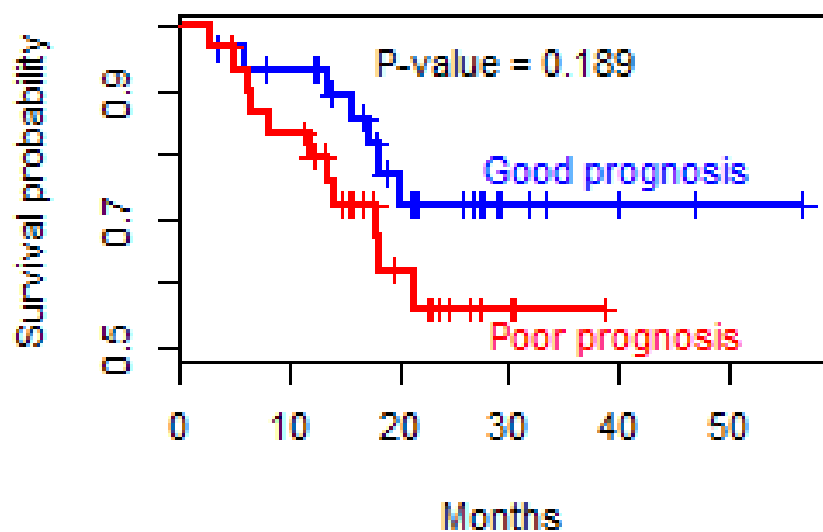
Optimal Wald test



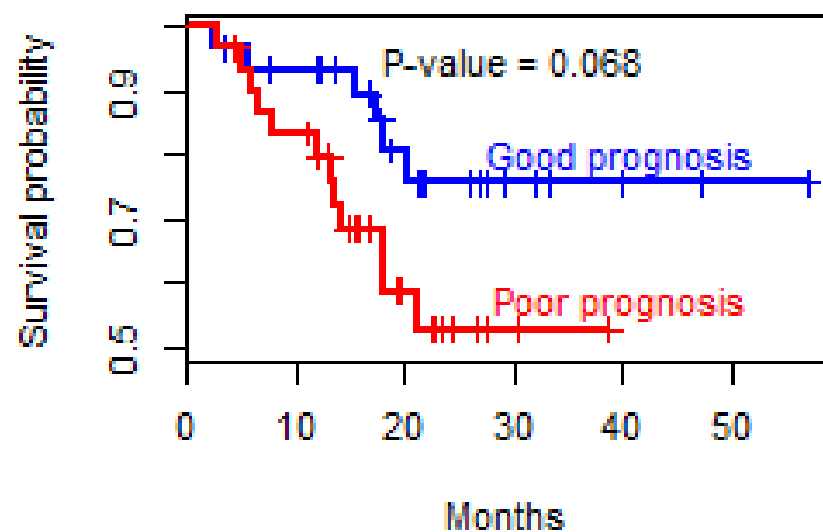
Optimal score test



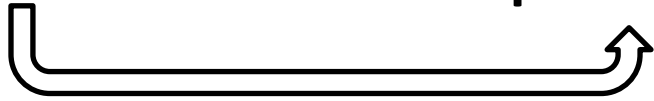
Copula + optimal Wald test



Copula + optimal score test



Summary

- An R package “*compound.Cox*”
 - Multiple tests for feature selection
≠ likelihood-based method (e.g., stepwise, Lasso)
 - Compound covariate for prediction
(Test → selection → prediction) in a coherent way

 - Matrix computation of score tests
(a number of tests done in a single matrix $\mathbf{Z} = \mathbf{S} / \mathbf{V}^{1/2}$)
- Implemented the evaluation measures:
 - Predictive capability (CVL) Matsui (2006)
 - False discovery rate (FDR) Witten and Tibshirani (2010)
- Copula model to deal with dependent censoring
 - improve prediction accuracy

References

- [1] **Allen DM** (1974), The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction, *Technometrics* 16, 125–27
- [2] **Witten M, Tibshirani R**. Survival analysis with high-dimensional covariates. *Statist Method Med Res* 2010; 19: 29-51.
- [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**, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. *N Engl J Med* 2007; 356: 11-20.
- [5] **Emura T, Chen YH, Chen HY**. Survival prediction based on compound covariate under Cox proportional hazard models. *PLoS One* 2012; 7(10): e47627.
- [6] **Emura T, Chen YH**, Gene selection for survival data under dependent censoring, a copula-based approach, *Statist Method Med Res* 2016; 25(6): 2840-57.
- [7] **Montgomery DC** (2009). *Statistical quality control* (Vol. 7), Wiley.
- [8] **Simon RM** (2003). *Design and analysis of DNA microarray investigations*, Springer
- [9] **Shyr Y, Kim K** (2003). Weighted flexible compound covariate method for classifying microarray data, *Springer*
- [10] **Tukey JW** (1993) Tightening the clinical trial *Controlled clinical trials* 14(4) 266-85