國家衛生研究院
群體健康科學研究所演講
2015/1/29

Gene selection for survival data under dependent censoring -- a copula-based approach –

published in *Statistical Methods in Medical Research* (2014),
DOI: 10.1177/0962280214533378

Takeshi Emura
Graduate Institute of Statistics National Central University
Joint work with Dr. Yi-Hau Chen (Academia Sinica)

Outline:

1) Survival analysis -- overview --
2) Dependent censoring
3) Proposed method

   -- Copula approach –

4) Lung cancer data analysis

Survival analysis -- overview --

Common endpoint (time-to-event):

1) Overall survival (OS)

Focus of this talk

= time-to-death due to any cause

2) Progression-free survival (PFS)

= time to the first evidence of disease progression (e.g., death, relapse, or metastasis)

Use of these endpoints are discussed by many authors

Sherrill et al. (2008), Rondeau, Pignon and Michiels (2011), Cheema and Burkes (2013) and Singh, Wang and Law (2014)

Survival analysis -- overview --

OS (Overall survival) = Response
Gene expression = Covariate

Cox regression is used to find genes associated with OS

- Lung cancer patients

  ERBB3, LCK, DUSP6, STAT1 (Chen et al., 2006 NEJM)

- Breast cancer patients

  ECRG4 (Sabatier et al., 2011, PLoS ONE)

- Ovarian cancer patients

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

Survival analysis -- overview --

OS (Overall survival)

= time-to-death due to any cause

Censoring = drop out (not death)

Mutually exclusive (competing) events

Example: Lung cancer data (Chen et al 2007, NEJM)

- 38 patients (died)
- 87 patients (censored)

n = 125 patients

Cox regression (Cox 1972) analysis is valid under:

Independent censoring assumption:

'death' and 'dropout' are independent

Survival analysis -- overview --

- Survival data :

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

$$t_i := \min\{\ \text{time to death}\ ,\ \text{censoring}\ \}$$

$$\delta_i = \begin{cases} 1 & \text{if time-to-death} \\ 0 & \text{if censoring time (drop out)} \end{cases}$$

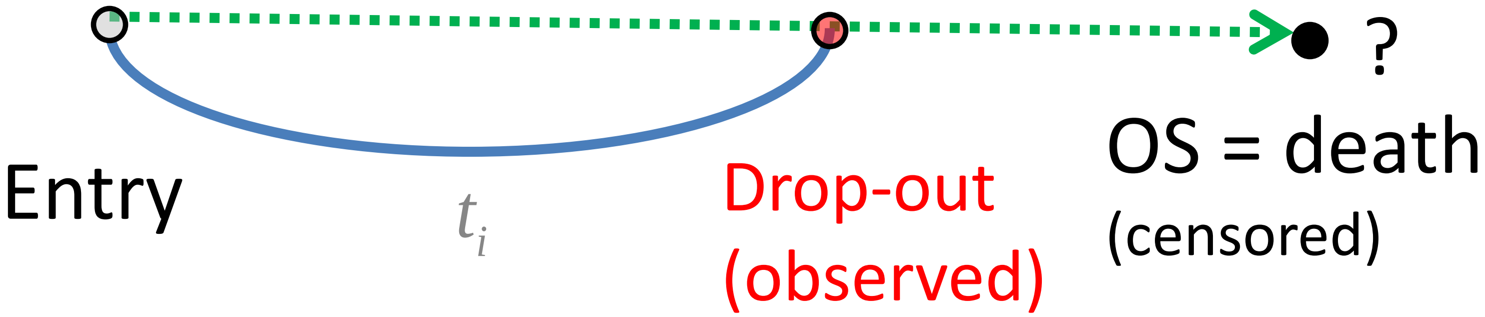


Fig. Case of censoring ($\delta_i = 0$)

Non-small-cell lung cancer data: Chen et al. (2007, NEJM)

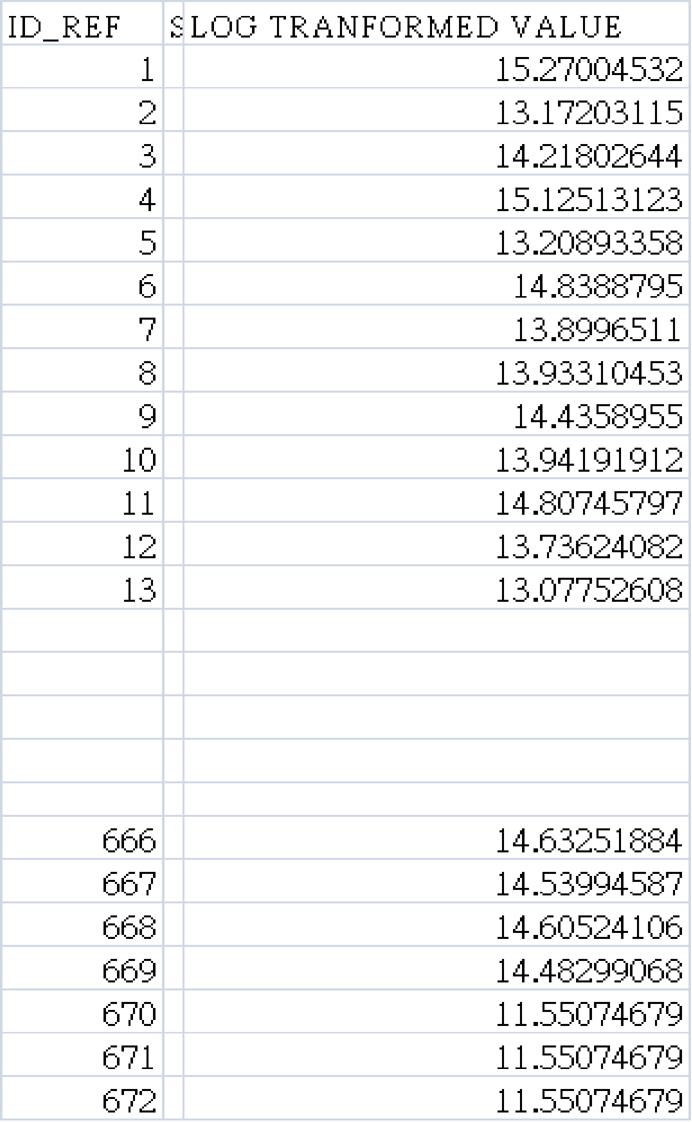
- Gene vector : $\mathbf{x}_i = (x_{i1}, \ldots, x_{i672})'$

p=672 >> n = 125
(high-dimensionality)

$$(\quad \text{Covariate} \quad \Rightarrow \quad \text{Gene} \quad)$$

- Select small subset of genes via **univariate** Cox regression

(e.g., Jenssen et al. 2002
Matsui 2006)



ID_REF	S	LOG TRANFORMED VALUE
1		15.27004532
2		13.17203115
3		14.21802644
4		15.12513123
5		13.20893358
6		14.8388795
7		13.8996511
8		13.93310453
9		14.4358955
10		13.94191912
11		14.80745797
12		13.73624082
13		13.07752608
666		14.63251884
667		14.53994587
668		14.60524106
669		14.48299068
670		11.55074679
671		11.55074679
672		11.55074679

Univariate Cox model

- $T = \text{OS}$ (Overall survival)
- $x_j = j - \text{th}$ gene expression

Hazard rate function assumed to be this form :

$$h(t \mid x_j) = \Pr(t \le T < t + dt \mid T \ge t, x_j) / dt$$

$$= h_0(t) \exp(\beta_j x_j)$$

Interpretation

$$\beta_j = \log \frac{h(t \mid x_j = 1)}{h(t \mid x_j = 0)}$$

for a unit change in j - th expression :

$x_j = 0 \quad \rightarrow \quad x_j = 1$

Univariate Selection

Step1: Univariate Cox model for a single gene j

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

Step2: Wald test for $H_{oj}: \beta_j = 0 \quad \text{vs.} \quad H_{1j}: \beta_j \neq 0$

using $\hat{\beta}_j / sd\{\hat{\beta}_j\}$

Step3 : Gene selection with smaller P-values than some **threshold**

1) P-value < 0.05
2) Cross-validated partial-likelihood (Masui 2006),
3) FDR (Witten & Tibshirani 2010), etc.

Univariate selection

- Gene selection via univariate Cox-regression is a simple strategy to overcome high-dimensionality

  Jenssen et al. (2002 Hum Genet)

  Matsui (2006 BMC Bioinformatics),

  Chen et al. (2007 NEJM)

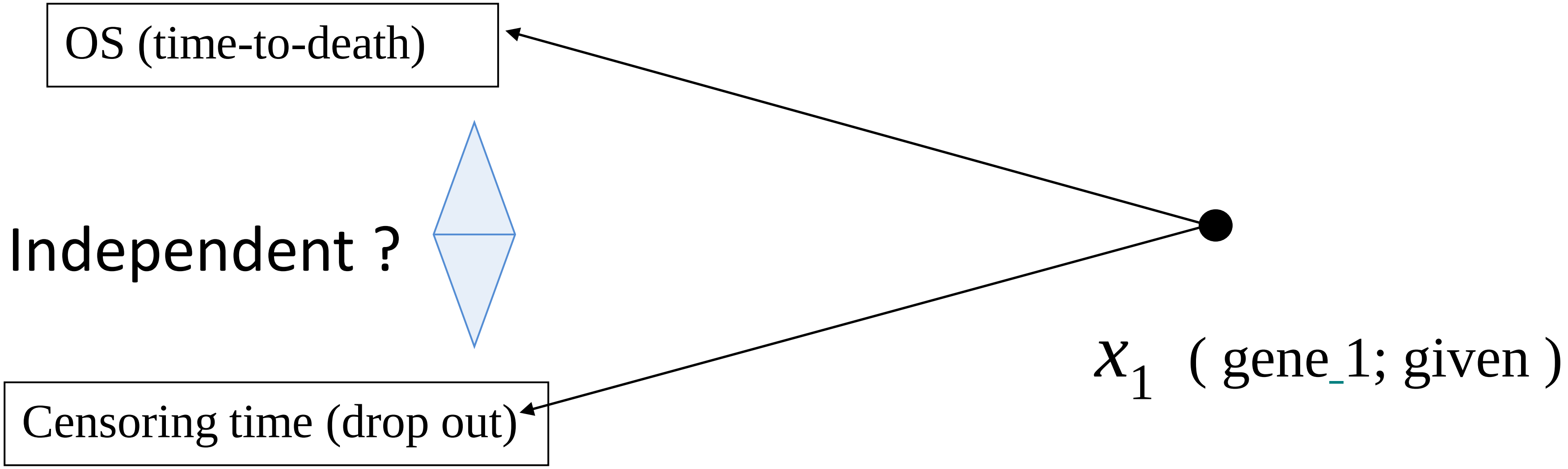
  Matsui et al. (2012 Clinical Cancer Res)

  just name a few

- Univariate selection is valid under independent censoring assumption

Independent censoring assumption

- ***Assumption:*** *The survival time* T *and censoring time* U *are conditionally independent given a gene* x_j *for all* $j = 1, \dots, p$.



OS (time-to-death)

Independent ?

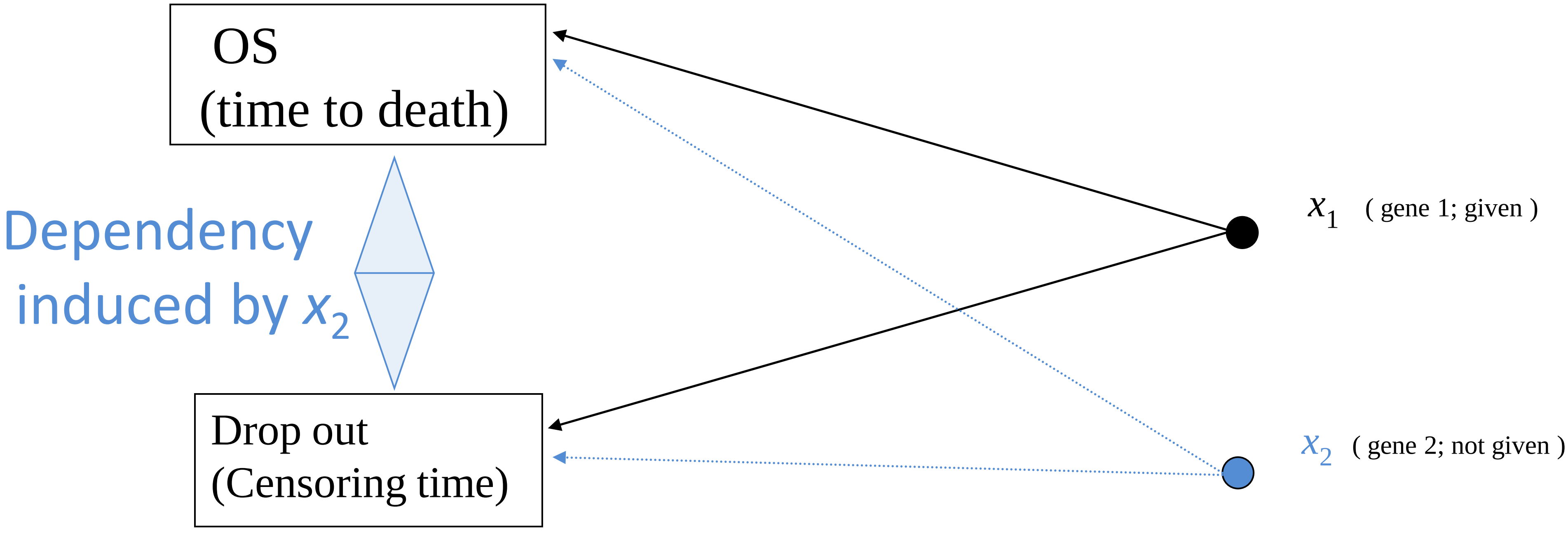
Censoring time (drop out)

x_1 (gene_1; given)

- Under the independent censoring assumption

  $\hat{\beta}_1 \xrightarrow{P} \beta_1$ (true effect of gene 1)

How independent censoring violate?



- OS (T) and censoring (U) usually cannot be conditionally independent given only x_1
  - ∨ unobserved x_2 affect both OS and censoring.
    - ➔ Frailty model (Oakes 1989) for bivariate lifetimes

How independent censoring violate?

- Given only *j*-th gene x_j

Dependency between Survival (T) and censoring (U) times is induced by $\boldsymbol{x}_{(-j)}$

$$\Pr(T > t, U > u \mid x_j)$$

$$= \varphi_{\boldsymbol{\beta}(-j),\boldsymbol{\gamma}(-j)}[\,\varphi^{-1}_{\boldsymbol{\beta}(-j)}\{\Pr(T > t \mid x_j)\}, \varphi^{-1}_{\boldsymbol{\gamma}(-j)}\{\Pr(U > u \mid x_j)\}\,]$$

where $\varphi_{\boldsymbol{\beta}(-j),\boldsymbol{\gamma}(-j)}$, $\varphi_{\boldsymbol{\beta}(-j)}$ and $\varphi_{\boldsymbol{\gamma}(-j)}$ are Laplace transforms

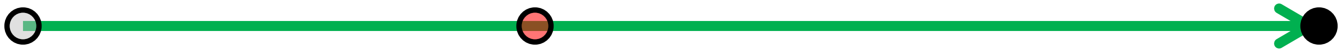
Details : Emura T & Chen YH (2014)

Univariate selection:

- Popular gene selection method in medical research
- Rely on the independence censoring
- If dependent censoring occurs, univariate selection may not correctly identify truly effective genes

- In this talk, we propose a gene selection that adjusts for dependent censoring using a copula

Copula: review



Entry Censoring = U Overall survival= T

$$\Pr(T \le t, U \le u) = C[\Pr(T \le t), \Pr(U \le u)]$$

- A copula function $C: [0,1] \times [0,1] \mapsto [0,1]$ characterize the dependence structures (Nelsen, 2006):

Example 1: Independence copula: $C[v, w] = vw$

Example 2: Clayton copula: $C_\alpha(v, w) = (v^{-\alpha} + w^{-\alpha} - 1)^{-1/\alpha}$,
(Clayton, 1978)

$$\alpha \begin{cases} = 0 & \text{independence} \\ > 0 & \text{positively dependece} \end{cases}$$

Proposed method

Copula model + Proportional hazards model

(Heckman & Honore 1989; Escarela & Carriere 2003; Chen 2010)

- Survival copula for dependent censoring:

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

- T_i : Survival Time

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

True Effect of gene j on overall survival

- U_i : Censoring Time

$$\Pr(U_i > u \mid x_{ij}) = \exp\{-\Gamma_{0j}(u)e^{\gamma_j x_{ij}}\}$$

Proposed method

Semiparametric MLE (Chen 2010, JRSSB)

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

Maximize:

R compound.Cox package (Emura & Chen 2014)

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

Estimated effect of gene j on overall survival

Proposed method

- Estimation of α is difficult

  (Unidentifiablility Tsiatis 1975)

- ML estimator for α

$$\hat{\alpha} = \arg\max_{\alpha} \ell(\hat{\beta}_j(\alpha), \hat{\gamma}_j(\alpha), \hat{\Lambda}_{0j}(\alpha), \hat{\Gamma}_{0j}(\alpha) \,|\, \alpha)$$

do not work !

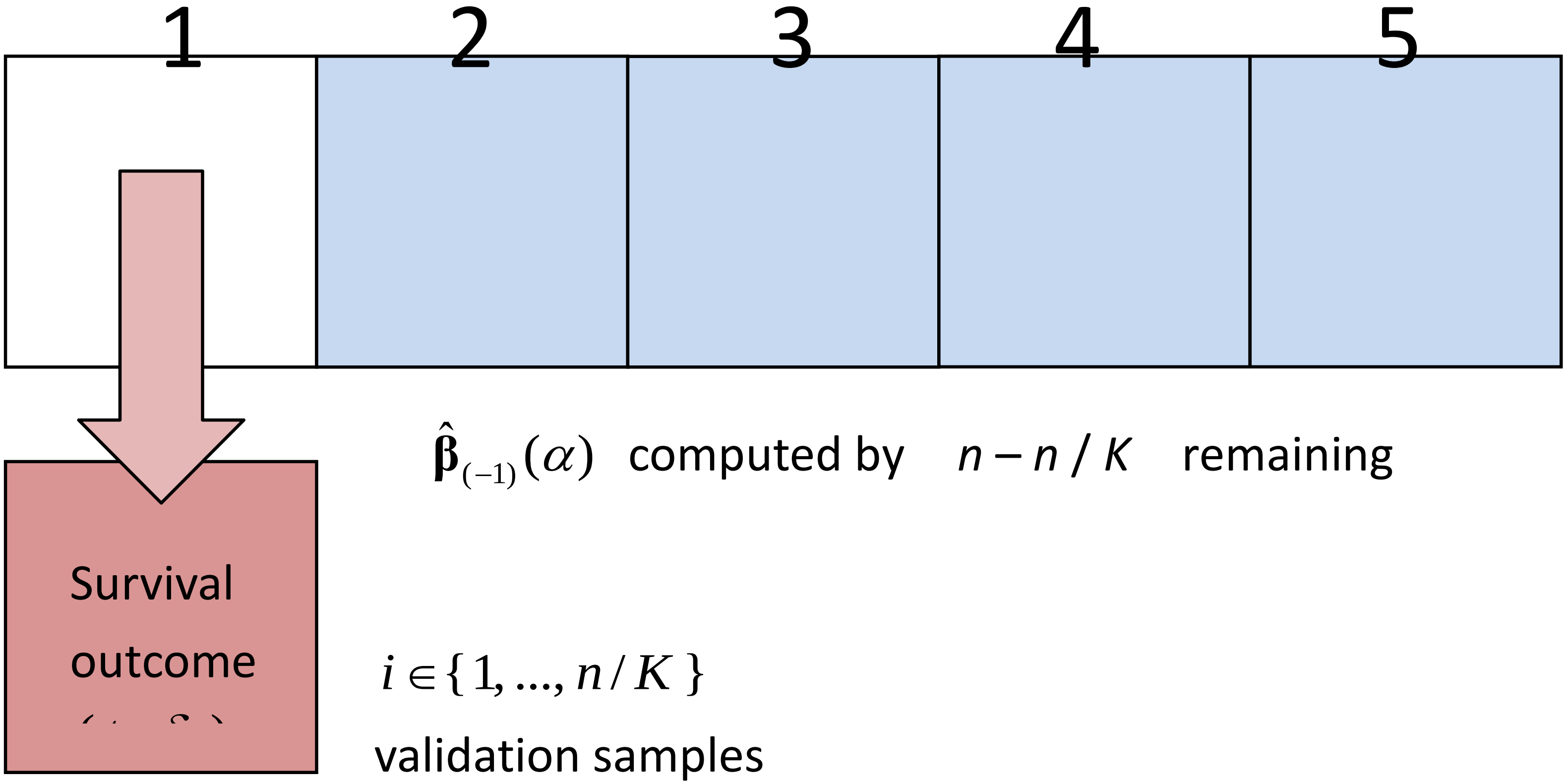
- Our strategy:

  Estimate α from prediction point of view

  ➔ Optimize a cross-validated prediction measure

Illustration of the *K* = 5 Cross validation:

- The individuals in the subset $k = 1$ are removed (Red color).
- $\hat{\boldsymbol{\beta}}_{(-1)}(\alpha)$ is computed by *n* – *n* / *K* remaining samples (Blue color)
- The outcome (t_i, δ_i) is validated by the $\mathrm{PI}_i(\alpha) = \hat{\boldsymbol{\beta}}'_{(-1)}(\alpha) x_i$,



$\hat{\boldsymbol{\beta}}_{(-1)}(\alpha)$ computed by $n - n/K$ remaining

$i \in \{1, \ldots, n/K\}$

validation samples

Proposed method

$$\left\{\begin{array}{l}\hat{\beta}_1(\alpha) <---- \text{Univariate estimate for Gene 1} \\ . \\ . \\ . \\ \hat{\beta}_p(\alpha) <---- \text{Univariate estimate for Gene } p\end{array}\right.$$

==> Compound univariate estimator

(Emura, et al., 2012 PLoS ONE):

$$\hat{\boldsymbol{\beta}}(\alpha) = (\ \hat{\beta}_1(\alpha), \ldots, \hat{\beta}_p(\alpha)\ )$$

$*$ Treat it as p-variate regressors

$*$ Still work even when $p > n$

Proposed method

- Prognostic index (PI)

$$\mathrm{PI_i}(\alpha) = \hat{\beta}_1(\alpha)x_{i1} + \cdots + \hat{\beta}_p(\alpha)x_{ip}$$

$$\Rightarrow \begin{cases} \text{High --> Poor prognosis} \\ \text{Low --> Good prognosis} \end{cases}$$

- Cross-validated *c*-index (Harrell’s *c*-index)

$$CV(\alpha) = \frac{\sum_{i<j}\{\,\mathbf{I}(\,t_i < t_j\,)\mathbf{I}(\,\mathrm{PI}_i(\alpha) > \mathrm{PI}_j(\alpha)\,)\delta_i + \mathbf{I}(\,t_j < t_i\,)\mathbf{I}(\,\mathrm{PI}_j(\alpha) > \mathrm{PI}_i(\alpha)\,)\delta_j\,\}}{\sum_{i<j}\{\,\mathbf{I}(\,t_i < t_j\,)\delta_i + \mathbf{I}(\,t_j < t_i\,)\delta_j\,\}}$$

- Proposedestimator for dependence parameter :

$$\hat{\alpha} = \arg\max CV(\alpha)$$

Proposed method

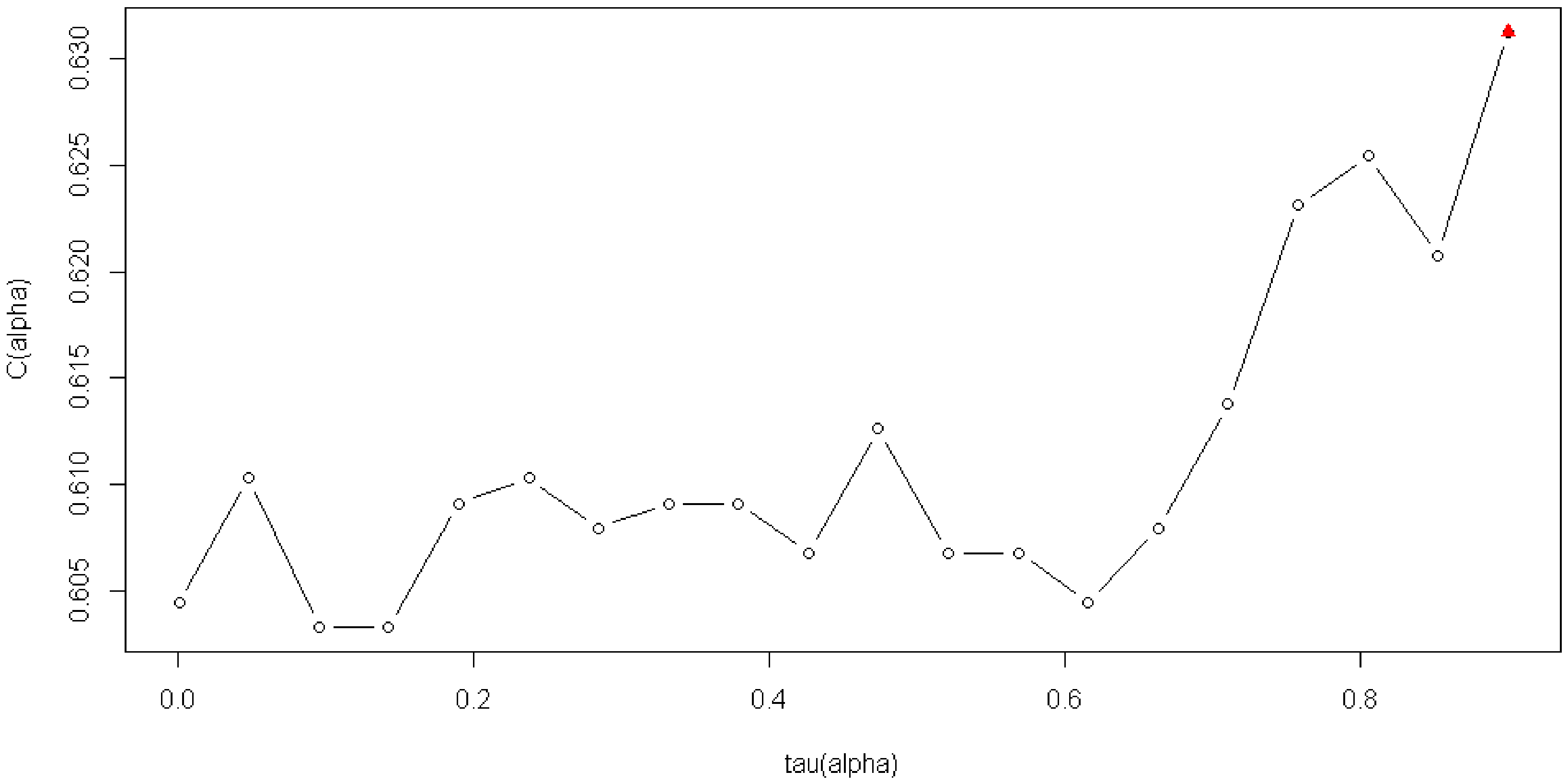


Fig. 6: The cross-validated *c*-index for the 63 training set from the lung cancer data. The cross-validated *c*-index is maximized at α =18, which corresponds to Kendall’s tau = 0.90.

Proposed method

Step1: Fit the copula-Cox model for a single gene j

$$\Pr(T_i > t, U_i > u \mid x_{ij}) = C_\alpha\{\exp\{-\Lambda_{0j}(t)e^{\beta_j x_{ij}}\}, \exp\{-\Gamma_{0j}(u)e^{\gamma_j x_{ij}}\}\}$$

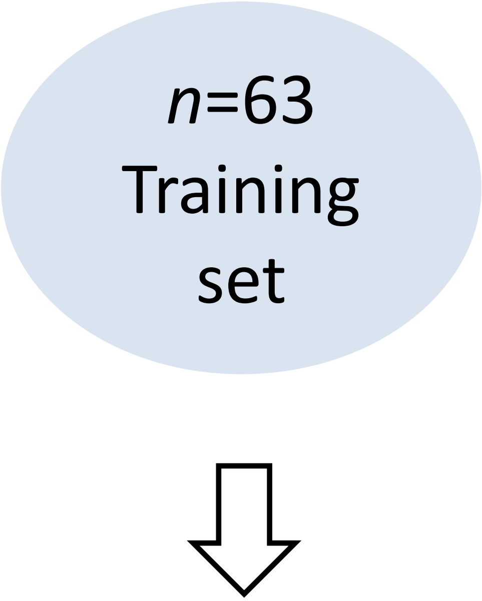
Step2: Wald test for $H_{oj}: \beta_j = 0$ via $\hat{\beta}_j(\hat{\alpha}) / sd\{\hat{\beta}_j(\hat{\alpha})\}$

(R compound.Cox package, Emura & Chen 2012)

Step3 : Gene selection with smaller P-values

NOTE: If $\alpha = 0$, then the proposed method is identical to univariate selection.

- Data: Lung cancer data (Chen et al., 2007 NEJM)



n=63
Training
set

Select 16 top genes (as in Chen et al. 2007)

1. Univariate selection
2. Proposed method

   (Claytoncopula with $\hat{\alpha} = 18$)

The 16 most strongly associated genes

	Univariate selection			Proposed method		
No.	Gene	Coefficient	P-value	Gene	Coefficient	P-value
1	ANXA5	-1.09	0.0039	ZNF264	0.51	0.0004
2	DLG2	1.32	0.0041	MMP16	0.50	0.0005
3	ZNF264	0.55	0.0079	HGF	0.50	0.0010
4	DUSP6	0.75	0.0086	HCK	-0.49	0.0012
5	CPEB4	0.59	0.0162	NF1	0.47	0.0016
~~						
14	FRAP1	-0.77	0.0408	DUSP6	0.40	0.0121
15	MMD	0.92	0.0419	ENG	-0.37	0.0139
16	HMMR	0.52	0.0481	CKMT1A	-0.41	0.0155

Gray shading signifies genes that appear in both univariate selection and the proposed

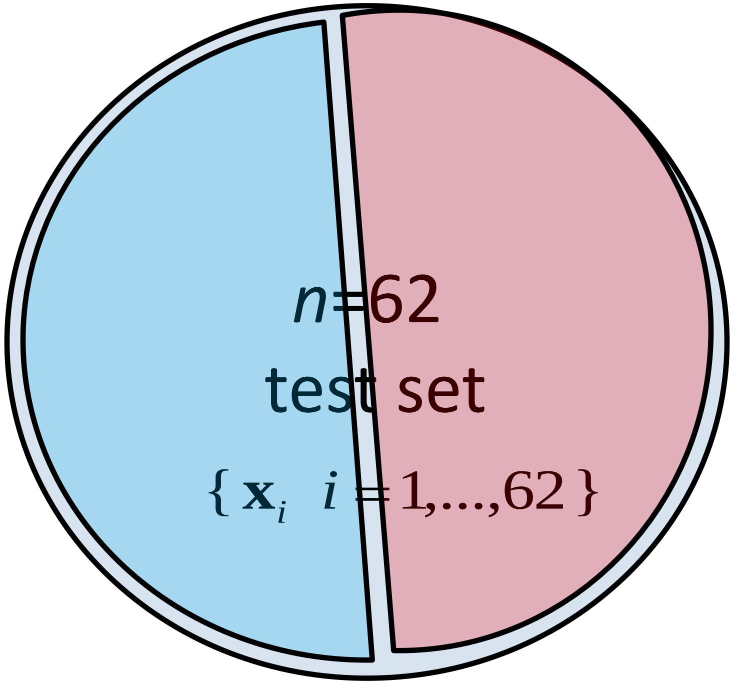
- Data: Lung cancer data (Chen et al., 2007 NEJM)

n=63
Training
set

Select 16 gene

1. Univariate selection
2. Proposed method

Predict



n=62
test set

$\{ \mathbf{x}_i \ \ i = 1,\ldots,62 \}$

Good prognosis

Poor prognosis

$$\mathrm{PI_i}(\alpha) = \hat{\beta}_1(\alpha) x_{i1} + \cdots + \hat{\beta}_{16}(\alpha) x_{i16}$$

$$\mathrm{PI_i}(\alpha) < c \ \ (\text{Good prognosis})$$

$$\mathrm{PI_i}(\alpha) > c \ \ (\text{Poor prognosis})$$

1. PI (univariate selection) =

(-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).

2. PI (proposed method) =

(0.51*ZNF264) + (0.50*MMP16) + (0.50*HGF) + (-0.49*HCK) + (0.47*NF1)

+ (0.46*ERBB3) + (0.57*NR2F6) + (0.77*AXL) + (0.51*CDC23) + (0.92*DLG2)

+ (-0.34*IGF2) + (0.54*RBBP6) + (0.51*COX11) + (0.40*DUSP6) + (-0.37*CKMT1A)

+ (-0.41*ENG).

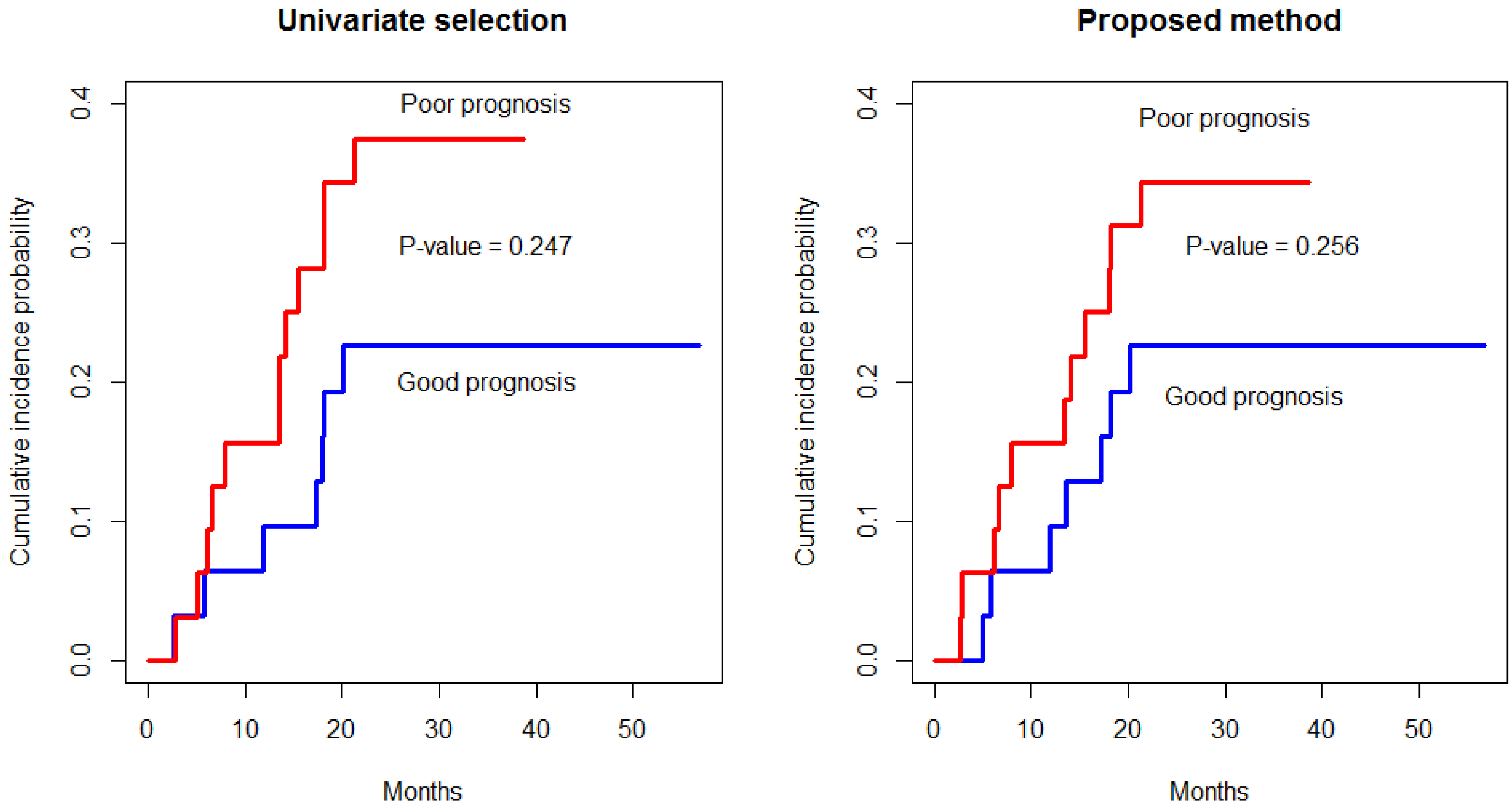


Figure 5 The cumulative incidence curves for the good (or poor) prognosis group separated by the top 16 genes. The good (or poor) group is determined by the low (or high values of the 16-gene prognostic index with equal sample sizes.

Main focus:

Predictive value on overall survival

- Kaplan-Meier survival curves are not consistent under dependent censoring
- Copula-graphic survival curves under dependent censoring

  Zheng & Klein 1995 Biometrika,

  Rivest & Wells 2001 JMVA

  (algorithm easy to compute)

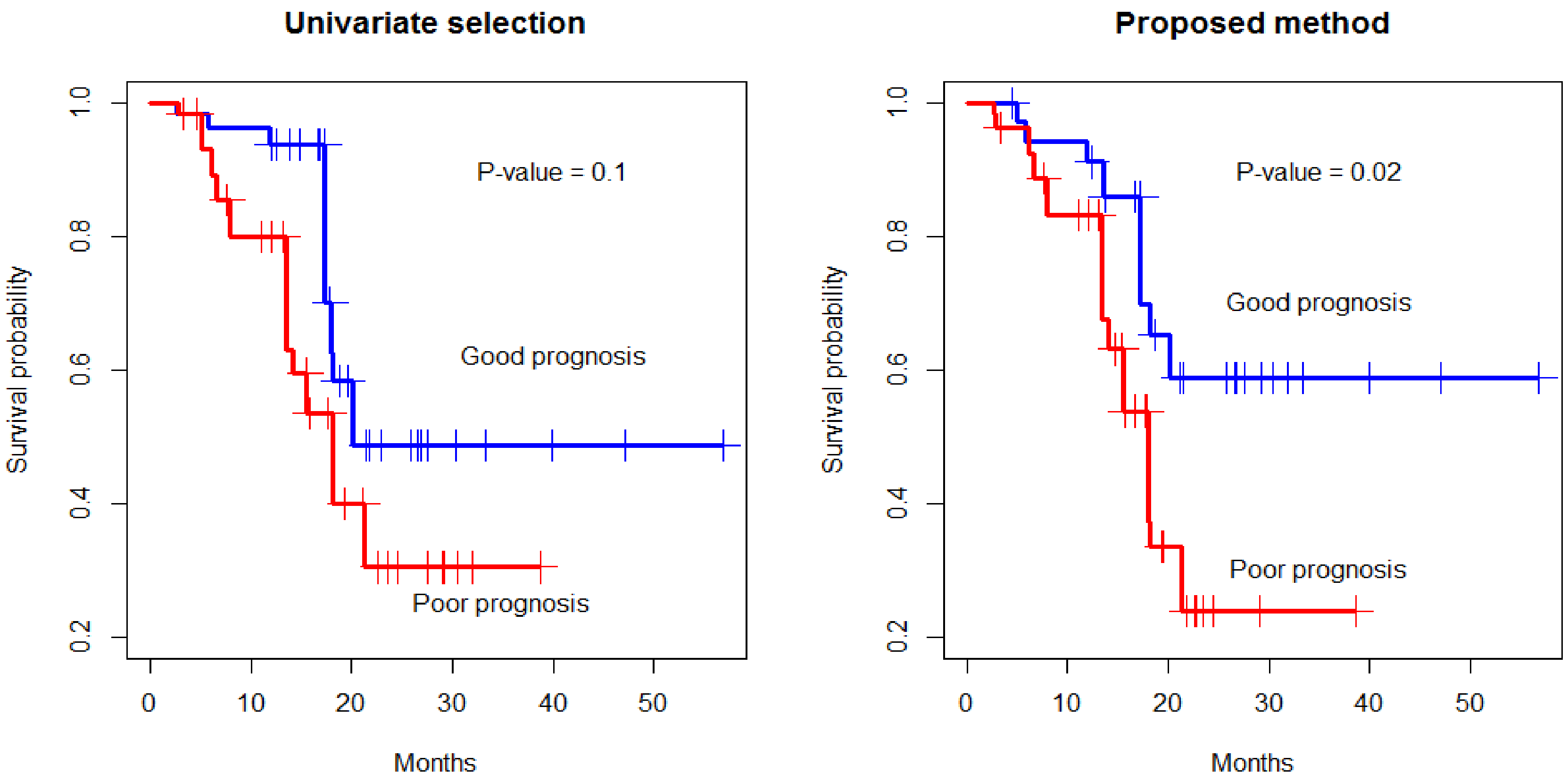


Figure 6 The marginal survival curves for the good (or poor) prognosis group separated by the top 16 genes. The good (or poor) group is determined by the low (or high) values of the 16-gene prognostic index with equal sample sizes.

Summary: Propose a gene selection method under **dependent censoring**

i) Copula approach for dependence model

➔Semi-parametric MLE (Chen 2010 JRSSB)

ii) New idea of estimating dependence parameter

➔ Cross-validated c-index

iii) Evaluation predictive power of selected gene:

➔ Copula-graphic estimator for survival curve

(Rivest & Wells 2001 JMVA)

iv) Software: **R** compound.Cox package

(Emura & Chen, version 1.4. 2014)