第二十二屆南區統計研討會

國立高雄大學統計學研究所

June 28-29, 2013

Gene selection for survival data under dependent censoring

-- a copula-based approach --

Takeshi Emura (中央大學統計所)

Joint work with Dr. Yi-Hau Chen (中研院)

Outline:

1) High-dimensional survival data

   -- Lung cancer data --

2) Univariate selection under independent censoring (popular method)

3) Proposed method under dependent censoring

4) Lung cancer data analysis

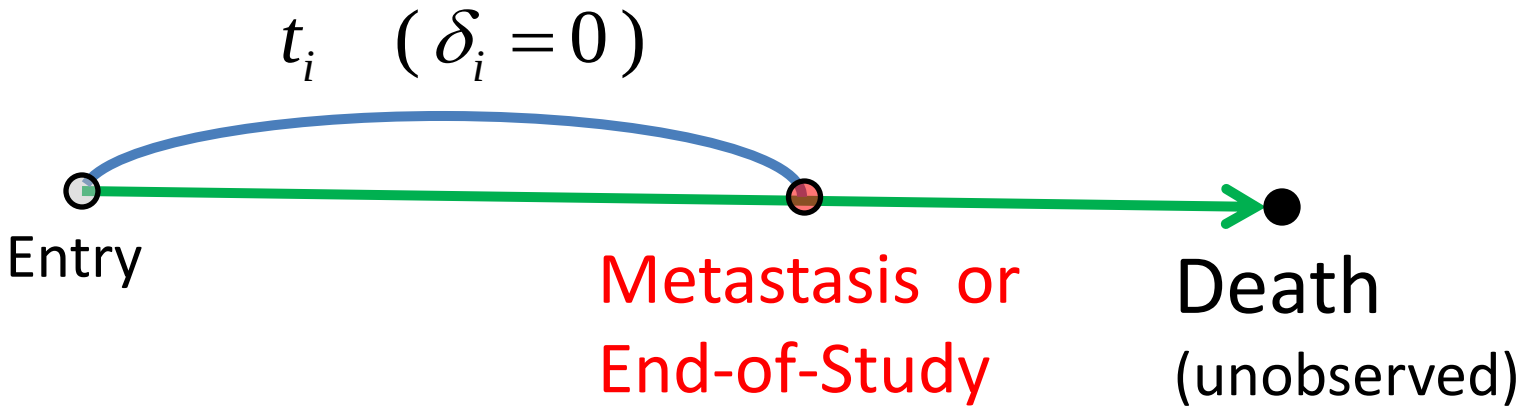
   Univariate selection vs. proposed method

High-dimensional Survival Data

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

t_i : either time to death or censoring

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



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

(Gene $\Leftrightarrow$ Covariate)

High-dimensional Survival Data

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

p=672 >> n = 125

(high-dimensionality)

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

ID_REF	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

High-dimensional Survival Data

- Genetic information is useful in survival prediction:

Breast cancer:

(Jenssen et al., 2002; van de Vijver et al., 2002;
van't Veer et al., 2002; Zhao et al., 2011)

Lung cancer:

(Beer et al., 2002; Chen et al., 2007; Shedden et al., 2008)

- Primary task is selecting a small fraction of genes that are relevant to survival

- Most common method in medical research:

  Gene selection via univariate Cox-regression

  Jenssen et al. (2002 Hum Genet), Matsui (2006 BMC Bioinformatics), Chen et al. (2007 NEJM) , name but a few

Univariate Selection

Step1: Univariate Cox model for a single gene j

$$\Pr(t \le t_i \le t + dt \mid t_i \ge t, x_{ij}) / dt = h_{0j}(t) \exp(\beta_j x_{ij}), \quad j = 1, \ldots, p$$

Step2: P-value of gene j for testing $H_{oj} : \beta_j = 0$

via Wald statistics $\hat{\beta}_j / sd\{\hat{\beta}_j\}$

Step3 : Gene selection with smaller P-values

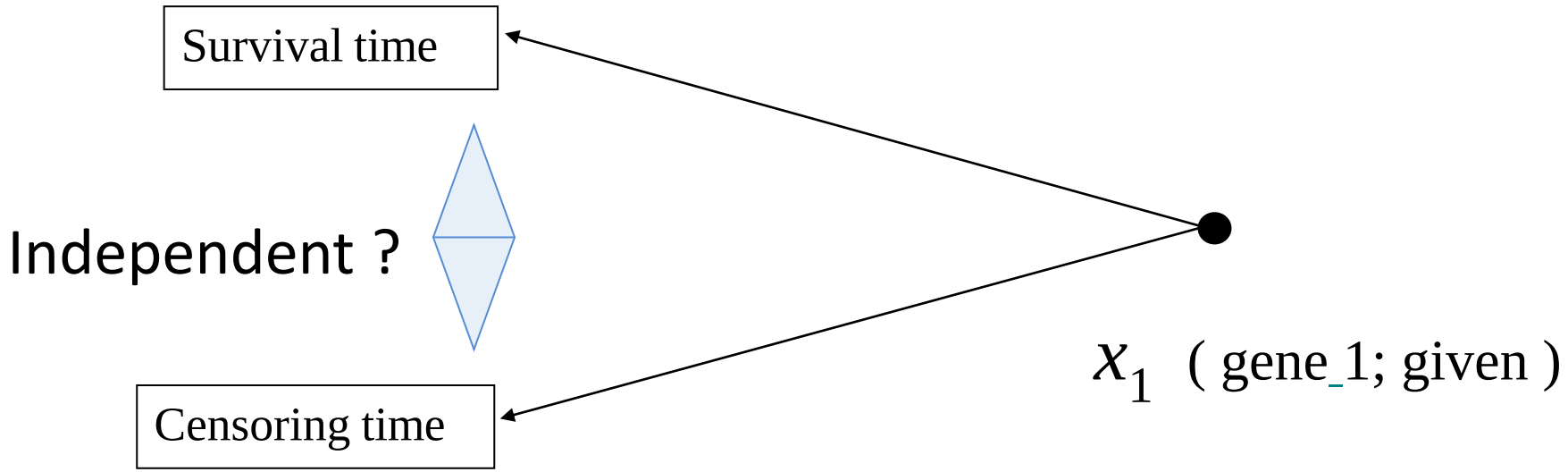
(e.g., P-value < 0.05)

Threshold can be determined by various different criteria

CV (Masui 2006), FDR (Witten & Tibs 2010), etc.

Independent censoring assumption

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



- Under the independent censoring assumption

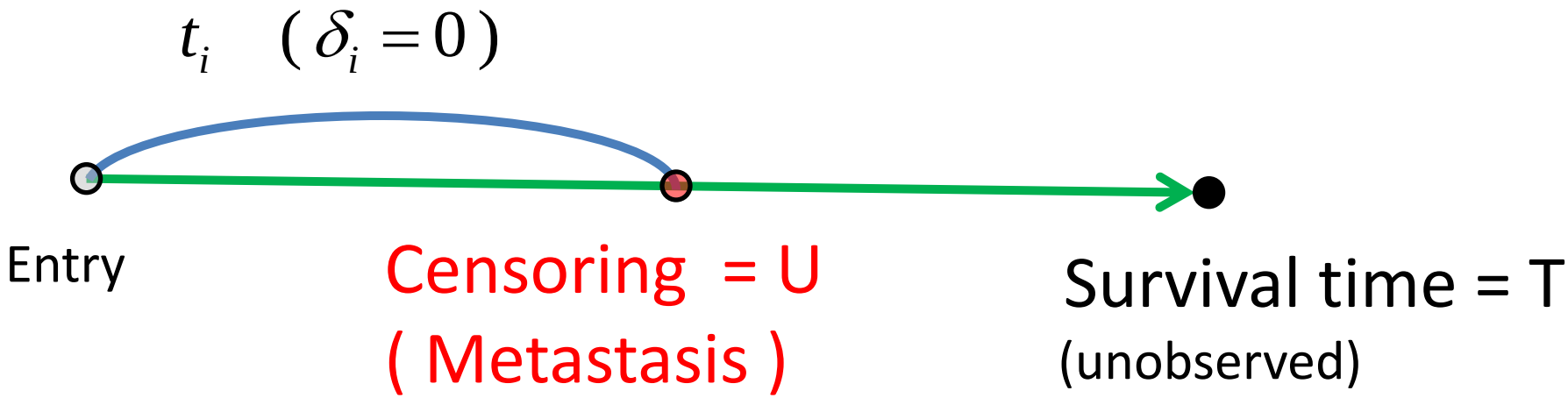
$$\hat{\beta}_j \xrightarrow{P} \beta_j, \quad j = 1, \quad \ldots, \quad p$$

Independent censoring assumption

$t_i = \min\{ T_i, U_i \}$

- T_i : Survival Time
- U_i : Censoring Time

Patient ID = 365
Age: 68.4 years-old
Gender: Male
Survival time (month): 4.55
Metastasis time (month): 1.186



* T and U may positively be dependent

Univariate selection:

- Most 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

Propose a gene selection method:

Adjust for dependent censoring via copula

Copula: review

Entry Censoring = U Survival time = T

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

- The function: $C: [0,1]\times[0,1] \mapsto [0,1]$, called copula, characterize the dependence structures

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

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

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

Proposed method

Proportional hazards with dependent censoring

(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 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)

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

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

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

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

Maximize:

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

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

Estimated effect of gene j on survival

Proposed method

- Estimation of α is impossible

  Unidentifiablility (Tsiatis 1975; Chen 2010)

- Prognostic index (PI).

$$\mathrm{PI}_i(\alpha) = \hat{\beta}_1(\alpha)x_{i1} + \cdots + \hat{\beta}_p(\alpha)x_{ip} \Rightarrow \begin{cases} \text{High} \text{ - - } > \text{Poor prognosis} \\ \text{Low} \text{ - - } > \text{Good prognosis} \end{cases}$$

- Maximize concordance (Harrell's *c*-index)

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

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

Proposed method

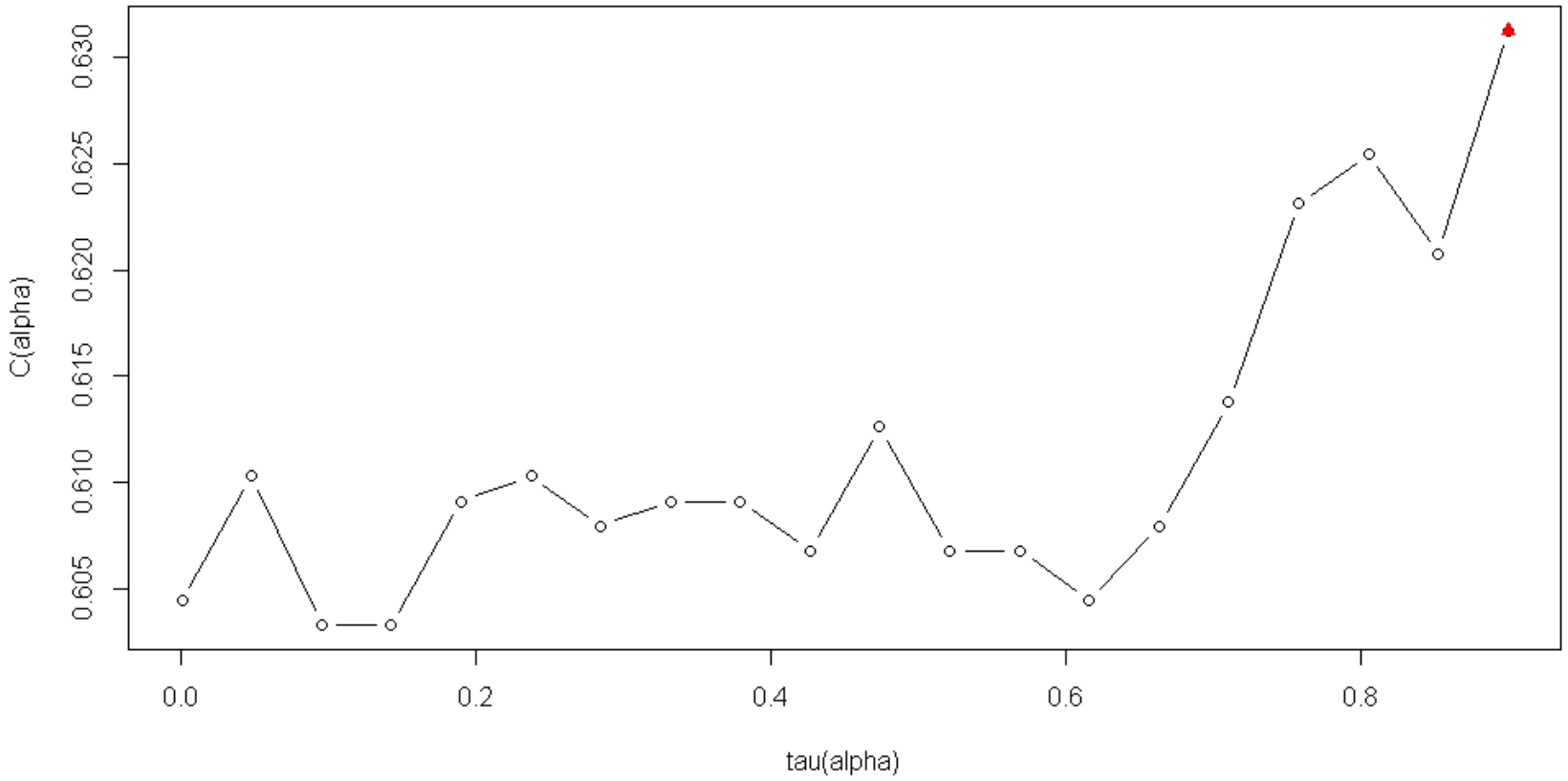


Fig. 6: The cross-validated *c*-index for the 63 training set from the lung cancer data. The cross-validated *c*-index is maximized at α =18, which corresponds to Kendall's tau = 0.90.

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: P-value of gene j for testing $H_{oj}: \beta_j = 0$

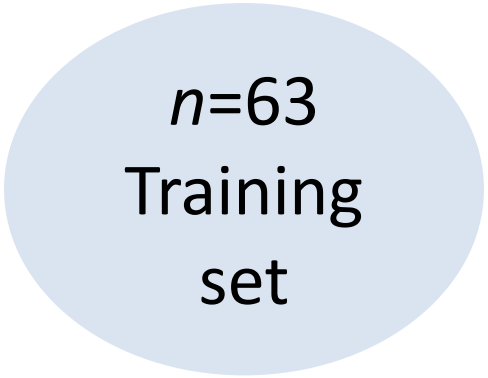
via Wald statistics $\hat{\beta}_j(\hat{\alpha}) / sd\{\hat{\beta}_j(\hat{\alpha})\}$

(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)



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

1. Univariate selection
2. Proposed method

   (Clayton copula 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

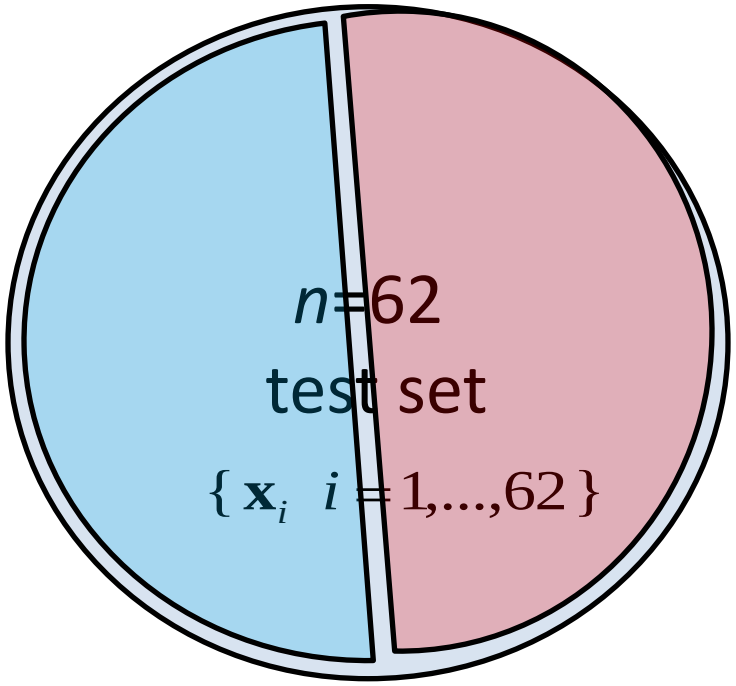
•Data: Lung cancer data (Chen et al., 2007 NEJM)

n=63
Training
set

Select 16 gene

1. Univariate selection
2. Proposed method

Predict



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

$$\mathrm{PI}_i(\alpha) < c \quad (\text{Good prognosis})$$

$$\mathrm{PI}_i(\alpha) > c \quad (\text{Poor prognosis})$$

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).

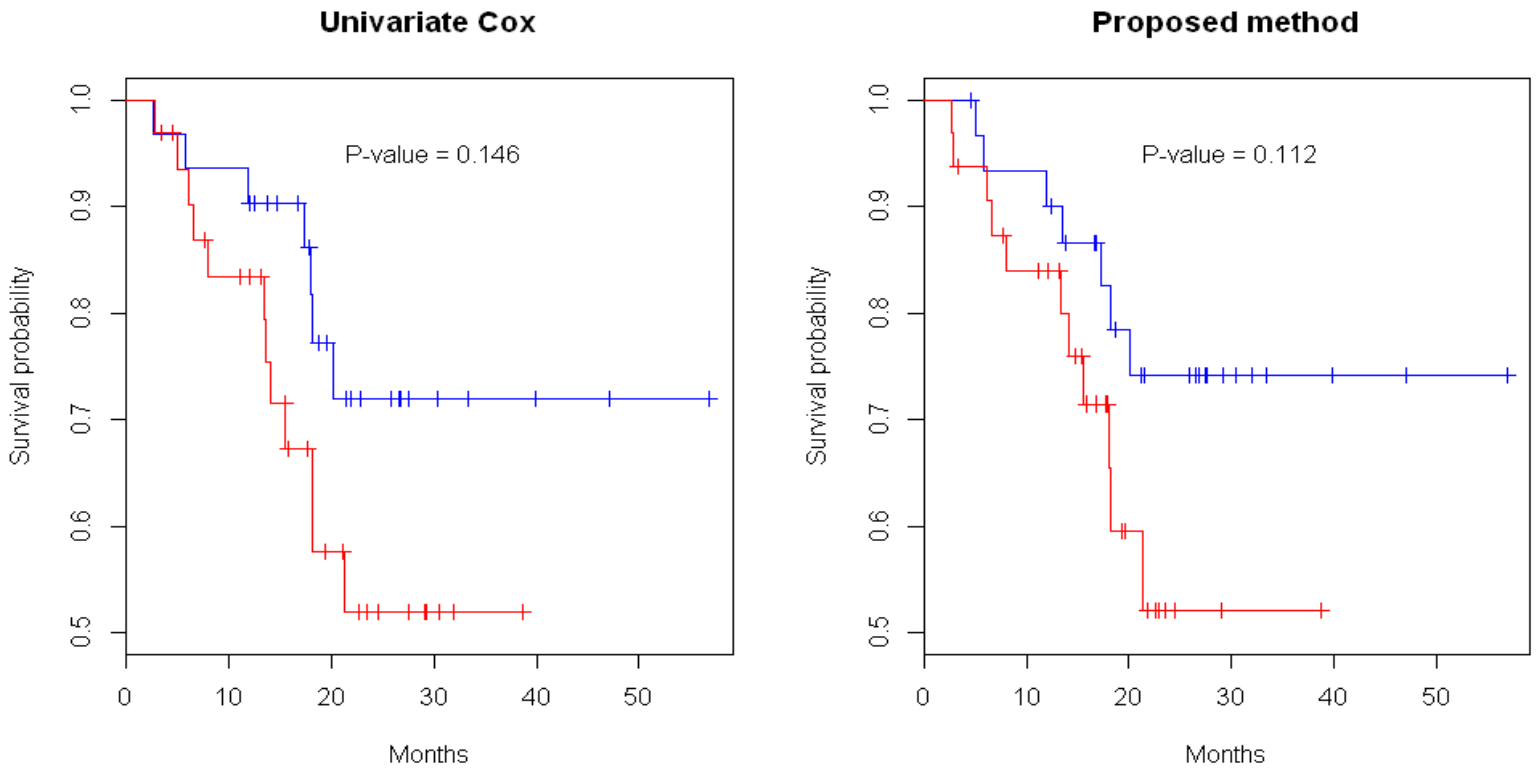


Fig. 4: The Kaplan-Meier plots for the good (or poor) prognosis group separated by the top 16 genes. The good (or poor) group is determined by the low (or high) values of the 16-gene prognostic index with equal sample sizes.

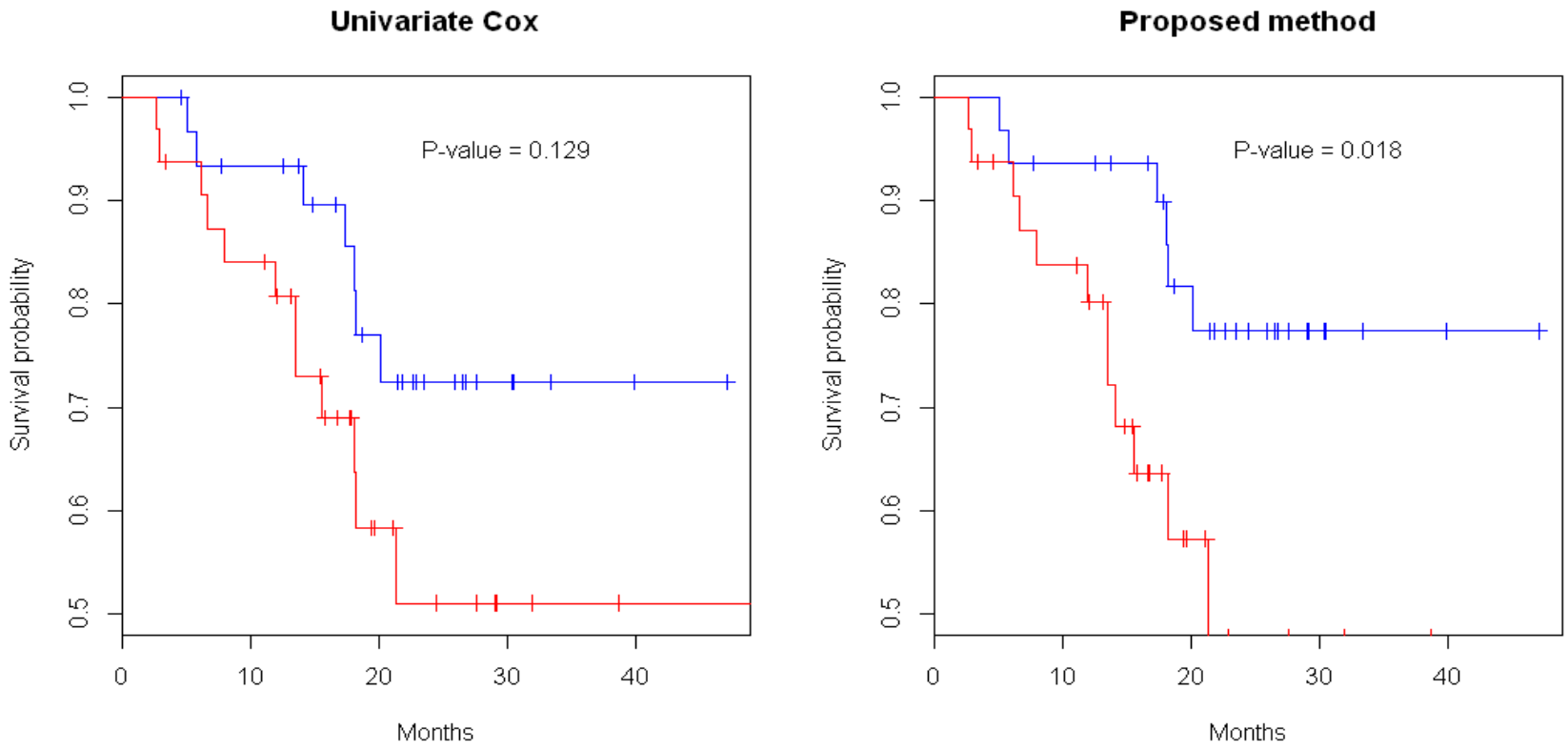


Fig. 5: The Kaplan-Meier plots for the good (or poor) prognosis group separated by the top **80 genes**. The good (or poor) group is determined by the low (or high) values of the **80-gene** prognostic index with equal sample sizes.

Thank you for your attention