國立中興大學統計所

Survival prediction using the Cox proportional hazard models with a high-dimensional covariates

Emura , Chen & Chen H-Y (2012), “Survival prediction based on compound covariate method under Cox proportional hazard models" PLoS ONE 7(10). doi:10.1371/journal.pone.0047627

Takeshi Emura (NCU)

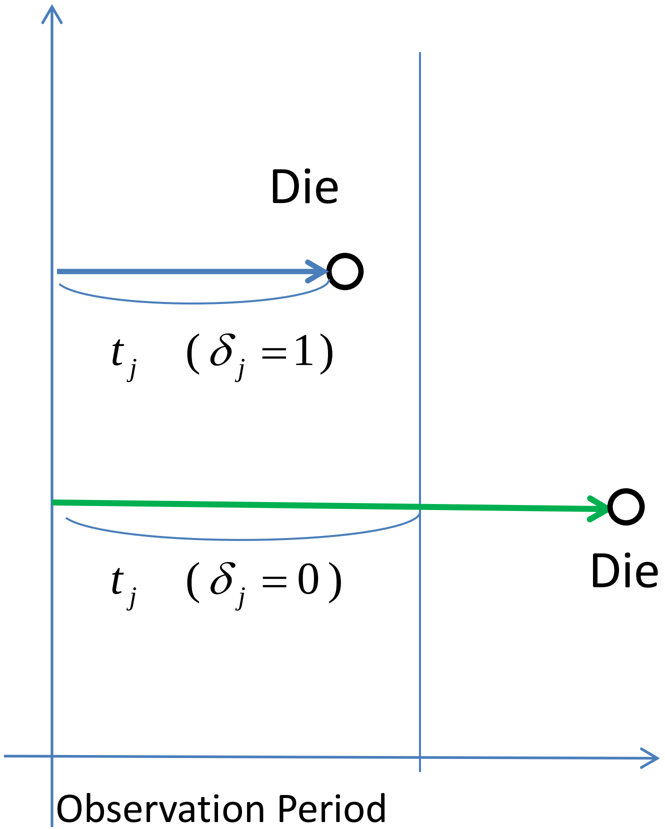
Joint work with Dr. Yi-Hau Chen and Dr. Hsuan-Yu Chen (Sinica)

- Survival data :

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

t_i : either time to death or censoring

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



- High-dimensional covariates :

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

(Gene $\Leftrightarrow$ Covariate)

Lung cancer data from Chen et al. 2007 NEJM

Patient ID	Survival_Status
100	alive , 47 months (censored)
101	alive, 49 months (censored)
102	death, 20 months
109	death, 26
110	alive, 39 (censored)
113	alive, 35 (censored)
115	alive, 45 (censored)
116	death, 9
128	death, 21
.	
.	
.	
.	
.	
365	alive, 5 months (censored)

$$t_i = 47, \ \delta_i = 0$$

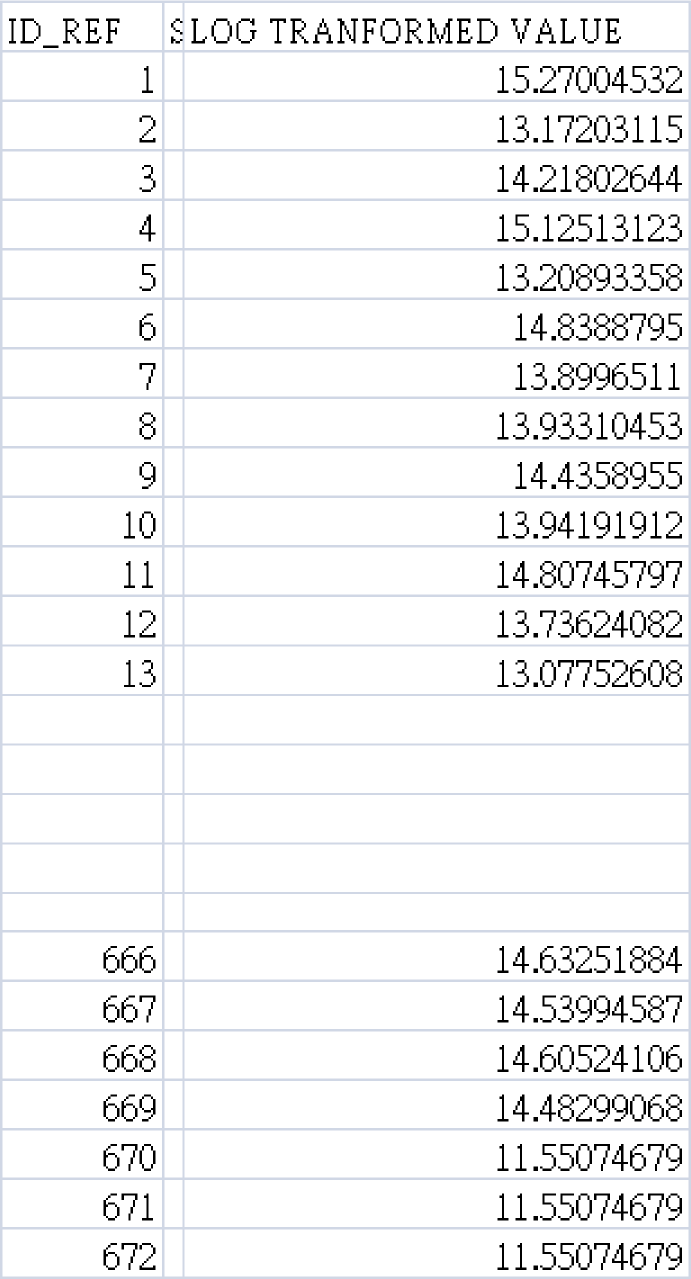
$$t_i = 20, \ \delta_i = 1$$

n=125 samples

1^{st} Patient (ID = 100)

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

P=672 covariates >> n = 125



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

- 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)

Hazard function:

$$h(t \mid \mathbf{x}_i) = \Pr(t \le t_i \le t + dt \mid t_i \ge t, \mathbf{x}_i) / dt$$

- Cox proportional hazard model (Cox 1972 JRSSB)

$$h(t \mid \mathbf{x}_i) = h_0(t) \exp(\boldsymbol{\beta}' \mathbf{x}_i)$$

•Partial likelihood estimator:

$$\hat{\boldsymbol{\beta}} : \quad L_n^1(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}' \mathbf{x}_i)}{\sum_{t_l \le t_i} \exp(\boldsymbol{\beta}' \mathbf{x}_l)} \right)^{\delta_i}$$

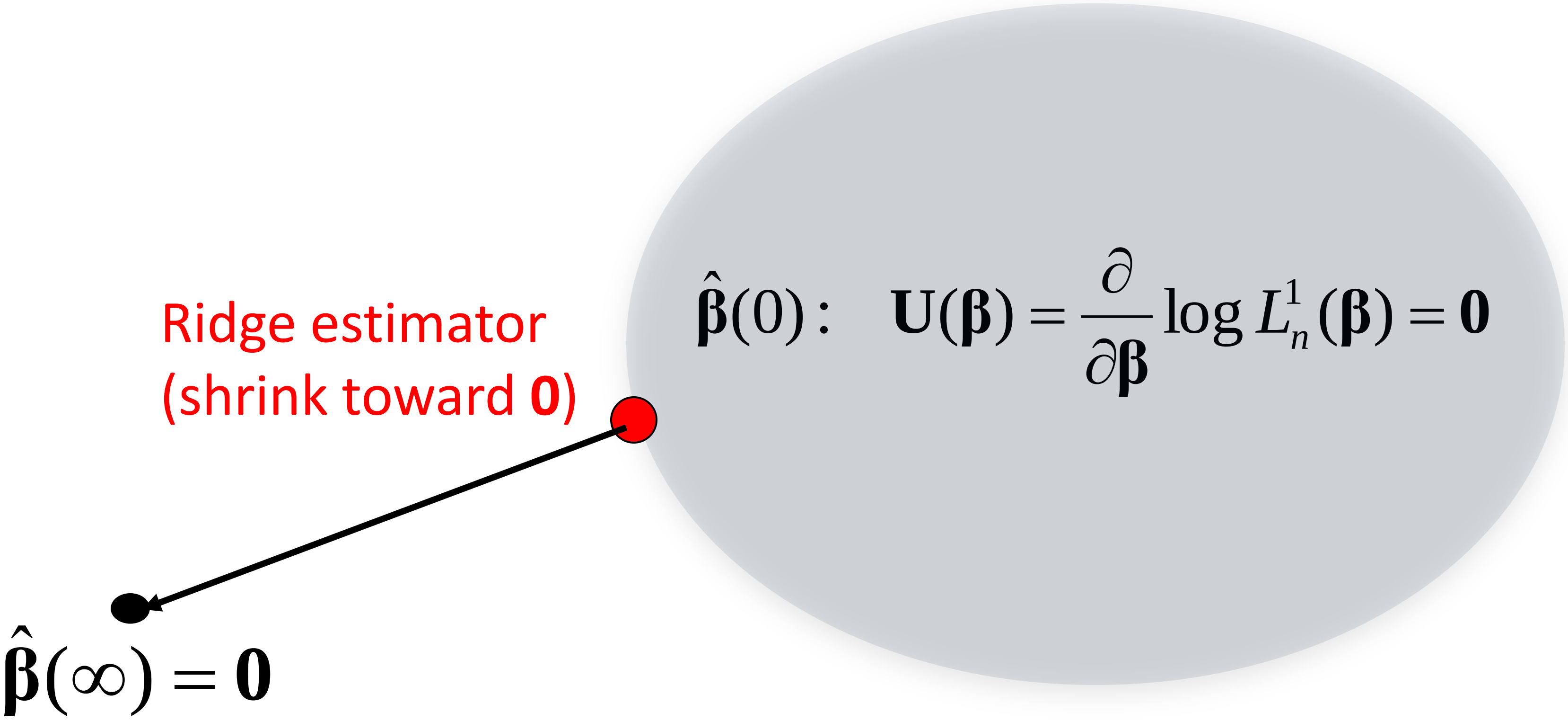
If *n* go to infinity, $\hat{\boldsymbol{\beta}} \rightarrow_P \boldsymbol{\beta}_0$

If *p* > *n*, $\hat{\boldsymbol{\beta}}$ is not unique (infinitely many maxima)

- Penalized Cox-regression (Verweij & Houwelingen 1994)

$$\hat{\boldsymbol{\beta}}(\lambda): \quad L_{\lambda}^{Ridge}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{\sum_{l \in R_i} \exp(\boldsymbol{\beta}'\mathbf{x}_l)} \right)^{\delta_i} - \lambda \| \boldsymbol{\beta} \|^2$$

(Ridge estimator)



$$\hat{\boldsymbol{\beta}}(0): \quad \mathbf{U}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log L_n^1(\boldsymbol{\beta}) = \mathbf{0}$$

Ridge estimator
(shrink toward **0**)

$$\hat{\boldsymbol{\beta}}(\infty) = \mathbf{0}$$

Existing methods for high-dimensional survival data

- Lasso (Cox-regression with L_1 penalty)

  Gui & Li (2005 Bioinformatics), Segal (2006 Biostatistics)

- Ridge regression (Cox-regression with L_2 penalty)

  Verveij & van Howelingen(1994 Stat. Med.), Zhao et al. (2011 PLoS ONE)

- Gene selection via univariate Cox-regression

  Jenssen et al. (2002 Nature Med.), Chen et al. (2007 NEJM), name but a few

- Others (PC, supervised PC, partial lease square, etc.)

Among above methods, ridge regression has the best performance in terms of survival prediction

(Bovelstad et al., 2007; van Weieringen e al., 2009; Bovelstad and Borgan, 2011)

Two objectives of our study:

1. Theoretical study for
   compound covariate prediction method

   *Previously used in microarrays datasets

   Tukey (1993 Controlled Clinical Trial), Beer et al. (2002 Nature Med.)
   Chen et al. (2007 NEJM), Radamacher et al (2002 J. of Theoretical Bio.)
   Matsui (2006 BMC Bioinformatics)

   *No theoretical analysis in the literature

2. Propose to refine the compound covariate
   prediction via *shrinkage* technique

Compound covariate prediction

Step1: For each gene $j(=1,\ldots,p)$, fit a univariate Cox model

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

Step2: A set of *p* regression coefficients

$$\hat{\boldsymbol{\beta}}(0) = (\hat{\beta}_1,\ldots,\hat{\beta}_p)', \quad \text{where } \hat{\beta}_j = \arg\max \prod_{i=1}^{n} \left(\frac{\exp(\beta_j x_{ij})}{\sum_{t_l \ge t_i} \exp(\beta_j x_{lj})} \right)^{\delta_i}$$

Remark: This is possible even when *p* > *n*

Step 3: *Compound covariate prediction*

For a future patient with genes $\mathbf{x} = (x_1, \ldots, x_p)'$,

$$\hat{\boldsymbol{\beta}}'(0)\mathbf{x} < c \ \text{(Good prognosis)} \ ; \ \hat{\boldsymbol{\beta}}'(0)\mathbf{x} > c \ \text{(Poor prognosis)}$$

Compound covariate method:

- Univariate method to resolve the high dimensionality

- Empirically perform well in microarray studies

- Its theoretical studies have not yet done

- **Assumption**: The Cox model holds with

$$h(t \mid \mathbf{x}_i) = h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}_i) = h_0(t)\exp(\beta_1 x_{i1} + \cdots + \beta_p x_{ip})$$

at the true parameter $\boldsymbol{\beta}' = \boldsymbol{\beta}_0' = (\beta_{0,1}, \ldots, \beta_{0,p}) \neq \mathbf{0}$

- **Remark:** Under the multivariate Cox model assumption, the *univariate Cox model does not hold*, i.e,

$$h(t \mid x_{ij}) = -\frac{\partial}{\partial t}\log E[\exp\{-H_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}_i)\} \mid x_{i1}]$$

$$\not\propto \exp(\beta_j x_{ij}).$$

- Univariate Cox model for each gene $j(=1,\ldots,p)$

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

is a misspecified model (a working model)

Ref:

Struthers & Kalbfleisch (1986) Misspecified proportional hazard models, Biometrika 73 pp.363-9.

- Univariate partial likelihood equation

$$\hat{\beta}_j : \quad \text{Solution to} \quad 0 = U_j(\beta_j) = \frac{1}{n}\sum_{i=1}^{n}\delta_i\left\{ x_{ij} - \frac{\sum_{\ell=1}^{n} I(t_\ell \ge t_i)x_{\ell j}\exp(\beta_j x_{\ell j})}{\sum_{\ell=1}^{n} I(t_\ell \ge t_i)\exp(\beta_j x_{\ell j})} \right\}$$

$$\beta_j^* \quad \text{Solution to} \quad 0 = u_j(\beta_j) \xleftarrow{P} U_j(\beta_j)$$

$$\hat{\beta}_j \xrightarrow{P} \beta_j^* \neq \beta_{0j} \quad \text{(true value in the Assumption)}$$

Remark I: If all genes $\mathbf{x} = (x_1, ..., x_p)'$ are independent

$$\text{sign}(\beta_j^*) = \text{sign}(\beta_{0j}), \qquad |\beta_j^*| \leq |\beta_{0j}|$$

Remark II:

Let $\boldsymbol{\beta}^*(0) = (\beta_1^*, \ldots, \beta_p^*)'$ and $\mathbf{0} = (0, \ldots, 0)'$.

Then, $\boldsymbol{\beta}^*(0)$ is between $\boldsymbol{\beta}_0$ and $\mathbf{0}$.

Above results deduced from :

Struthers & Kalbfleisch (1986 Biometrika) ; Bretagnolle & Huber-Carol(1988 Scand. JS)

Proposed estimator

- Univariate *compound* likelihood (unique maxima)

$$L_n^0(\boldsymbol{\beta}) = \prod_{j=1}^{p}\prod_{i=1}^{n}\left(\frac{\exp(\beta_j x_{ij})}{\sum_{t_l \geq t_i}\exp(\beta_j x_{lj})}\right)^{\delta_i}$$

- Multivariate likelihood (infinitely many maxima when *p* > *n*)

$$L_n^1(\boldsymbol{\beta}) = \prod_{i=1}^{n}\left(\frac{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}{\sum_{l \in R_i}\exp(\boldsymbol{\beta}'\mathbf{x}_l)}\right)^{\delta_i}$$

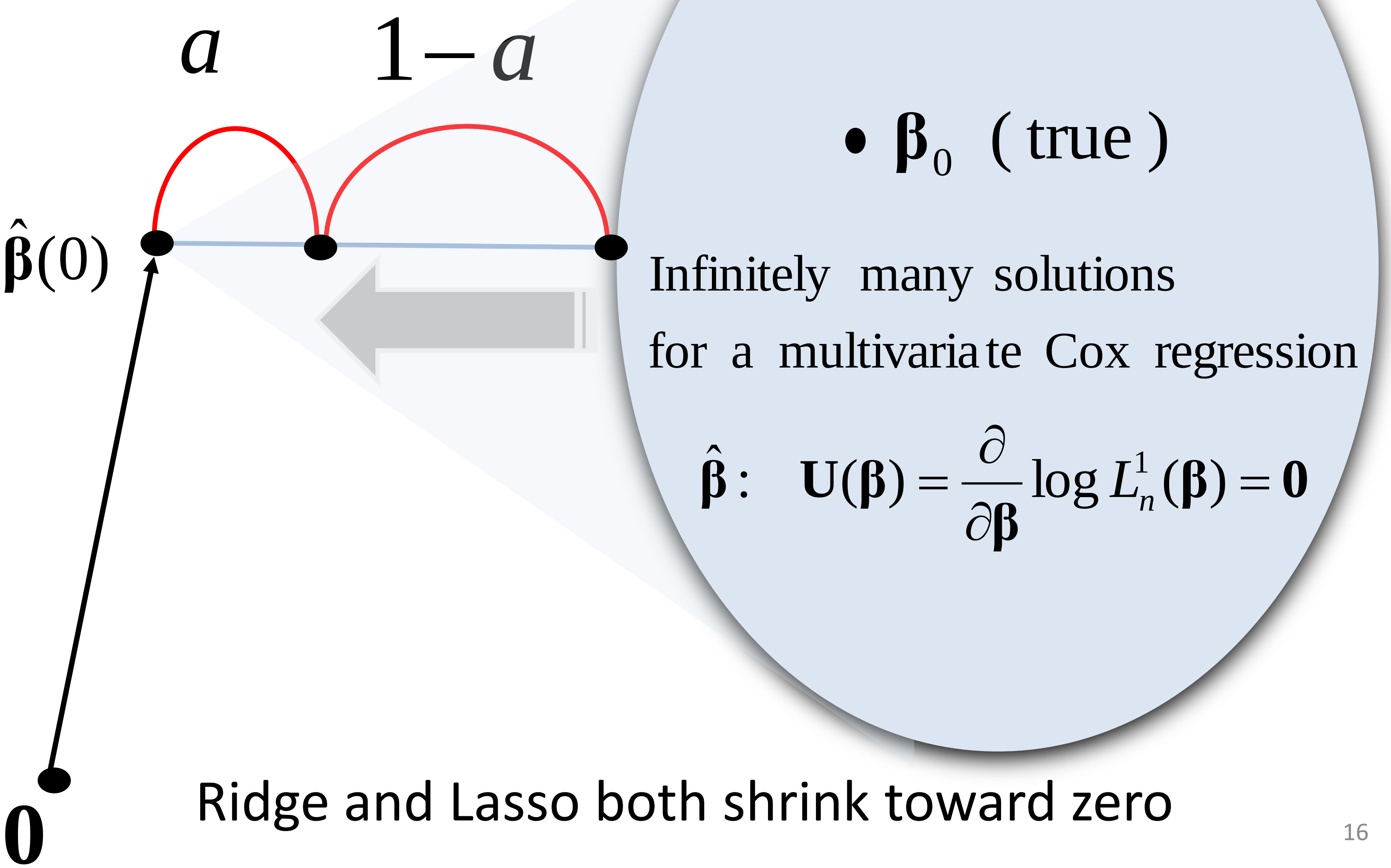
- Idea: Mixture of univariate and multivariate likelihood

$$\hat{\boldsymbol{\beta}}(a) = \operatorname{argmax}\left\{ a \log L_n^1(\boldsymbol{\beta}) + (1-a)\log L_n^0(\boldsymbol{\beta}) \right\}, \quad a \in [0,1]$$

We call it "compound shrinkage estimator"

Compound shrinkage estimator :

$$\hat{\boldsymbol{\beta}}(a) = \operatorname{argmax}\left\{ a \log L_n^1(\boldsymbol{\beta}) + (1-a) \log L_n^0(\boldsymbol{\beta}) \right\}$$



- $\boldsymbol{\beta}_0$ (true)

Infinitely many solutions for a multivariate Cox regression

$$\hat{\boldsymbol{\beta}}: \quad \mathbf{U}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log L_n^1(\boldsymbol{\beta}) = \mathbf{0}$$

Ridge and Lasso both shrink toward zero

- Proposition 2: (in our paper)

$$\sqrt{n}(\hat{\boldsymbol{\beta}}(\hat{a})-\boldsymbol{\beta}_0) \to N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)) \text{ with } \quad \hat{a} = \text{argmax}\, CV(a).$$

(CV = Cross-Validated likelihood of Verveij & Houwelingen 1993)

- Plug-in variance estimator $\boldsymbol{\Sigma}_n^{\hat{a}}(\hat{\boldsymbol{\beta}}(\hat{a}))$

$$\boldsymbol{\Sigma}_n^a(\boldsymbol{\beta}) = \mathbf{A}_n^a(\boldsymbol{\beta})\{\mathbf{V}_n^a(\boldsymbol{\beta})/n\}^{-1}\mathbf{A}_n^a(\boldsymbol{\beta})'$$

$$\mathbf{A}_n^a(\boldsymbol{\beta}) = \mathbf{V}_n^a(\boldsymbol{\beta})^{-1}\dot{\mathbf{h}}_n(\boldsymbol{\beta})\{-d^2CV(a)/da^2\}^{-1}\dot{\mathbf{h}}_n(\boldsymbol{\beta})' + \mathbf{I}_p$$

$$\dot{h}_n(\boldsymbol{\beta}) = \partial\mathbf{U}_n^a(\boldsymbol{\beta})/\partial a, \text{ where } \mathbf{U}_n^a(\boldsymbol{\beta}) = \text{Score function}$$

$$\frac{d}{da}CV(a) = \text{Estimating function of } a,$$

$$\mathbf{V}_n^a(\boldsymbol{\beta}) = \text{observed Fisher information}$$

*Reasonable performance even when *p* > *n*.

Numerical comparison

$\hat{\boldsymbol{\beta}}$ is obtained by 4 methods

1. Compound covariate (CC) estimator

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)', \quad \text{where } \hat{\beta}_j = \text{univariate Cox regression estimators}$$

2. Compound shrinkage (CS) estimator

$$a \log L_n^1(\boldsymbol{\beta}) + (1-a) \log L_n^0(\boldsymbol{\beta})$$

3. Ridge estimator

$$\log L_n^1(\boldsymbol{\beta}) - (\lambda/2) \sum_{j=1}^{p} \beta_j^2$$

4. Lasso estimator

$$\log L_n^1(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} |\beta_j|$$

* $\hat{a}$ or $\hat{\lambda}$ is obtained by cross-validation (Verveij & Houwelingen 1993 Stat.Med.)

Simulation set up

- Cox model: $h(t \mid \mathbf{x}_i) = \exp(\beta_1 x_{i,1} + \cdots + \beta_{100} x_{i,100})$

  $\Rightarrow T_i \sim Exp(\lambda_i),$

  where $\lambda_i = \exp(\beta_1 x_{i,1} + \cdots + \beta_{100} x_{i,100}).$

- Random censoring: $C_i \sim U(0,1)$

  (Censoring 54~63%)

- Data: $\{ (t_i , \delta_i , \mathbf{x}_i); i = 1, \ldots, 100 \}$

  where $t_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$

Simulation set up

1) Training set $\{ (t_i , \delta_i , \mathbf{x}_i); i = 1, \ldots, 100 \}$

$$\hat{\boldsymbol{\beta}}' = \begin{cases} \text{compound covariate} \\ \text{compound shrinkage} \\ \text{Ridge regression} \\ \text{Lasso} \end{cases}$$

R “compound.Cox” package
Emura & Chen (2012)

R “penalized” package
Goeman (2010)

2) Testing set $\{ (t_i^*, \delta_i^*, \mathbf{x}_i^*); i = 1, \ldots, 100 \}$

$\hat{\boldsymbol{\beta}}'\mathbf{x}_i^* < c$ (Good prognosis) ; $\hat{\boldsymbol{\beta}}'\mathbf{x}_i^* > c$ (Poor prognosis)

P-value from a two-sample Log-rank test

(Smaller P-value corresponds to better discrimination power)

*Repeat 50 times

Table 1. Simulation results under **sparse cases**.

CC = compound covariate, CS = compound shrinkage.

LR-test = Log_{10} P-value for discriminating poor / good patients.

Scenario 1: Tag gene / Scenario 2: Gene pathway

		$\boldsymbol{\beta} = (1.5, 1.5, \underbrace{0, \ldots, 0}_{\times 98})$			
		CC	CS	Ridge	Lasso
Scenario1	LR-test	-5.89	-5.88	-4.99	-10.59
Scenario2	LR-test	-8.88	-9.35	-7.01	-12.39
		$\boldsymbol{\beta} = (\underbrace{0.8, \ldots, 0.8}_{\times 5}, \underbrace{0, \ldots, 0}_{\times 95})$			
		CC	CS	Ridge	Lasso
Scenario1	LR-test	-3.88	-4.31	-4.21	-6.64
Scenario2	LR-test	-13.71	-13.69	-11.38	-14.52

Table 2. Simulation results under **Non-sparse cases**.

CC = compound covariate, CS = compound shrinkage.

LR-test = Log_{10} P-value for discriminating poor / good patients.

Scenario 1: Tag gene / Scenario 2: Gene pathway

		$\boldsymbol{\beta} = (\underbrace{0.2,\ldots,0.2}_{\times 10}, \underbrace{-0.2,\ldots,-0.2}_{\times 10}, \underbrace{0,\ldots,0}_{\times 80})$			
		CC	CS	Ridge	Lasso
Scenario1	LR-test	-1.22	-1.28	-1.29	-0.39
Scenario2	LR-test	-10.35	-9.49	-9.33	-9.11
		$\boldsymbol{\beta} = (\underbrace{0.1,\ldots,0.1}_{\times 15}, \underbrace{-0.1,\ldots,-0.1}_{\times 15}, \underbrace{0,\ldots,0}_{\times 70})$			
		CC	CS	Ridge	Lasso
Scenario1	LR-test	-0.55	-0.61	-0.61	-0.40
Scenario2	LR-test	-7.93	-6.80	-6.67	-6.05

Mostly, $\hat{\boldsymbol{\beta}}' = \mathbf{0}$ for Lasso

Simulation results: Summary

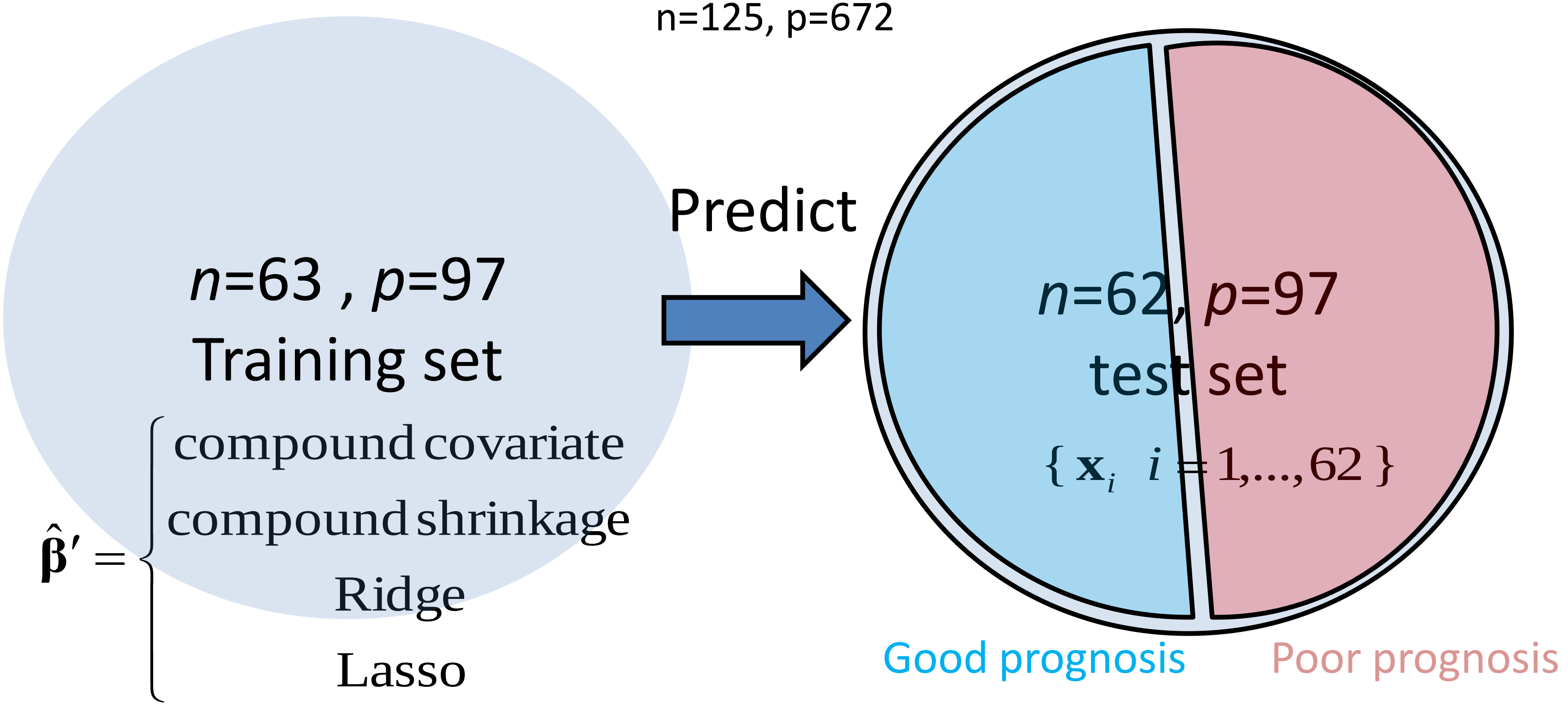
- Lasso is best in sparse cases
- Ridge and compound shrinkage are better than Lasso in non-sparse cases
- Compound shrinkage is slightly better than Ridge (sparse cases)

Remark: Ridge is reported as the best method in the literature

Bovelstad et al., 2007; van Weieringen e al., 2009; Bovelstad and Borgan, 2011

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

n=125, p=672



n=63 , *p*=97

Training set

Predict

n=62, *p*=97

test set

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

Good prognosis

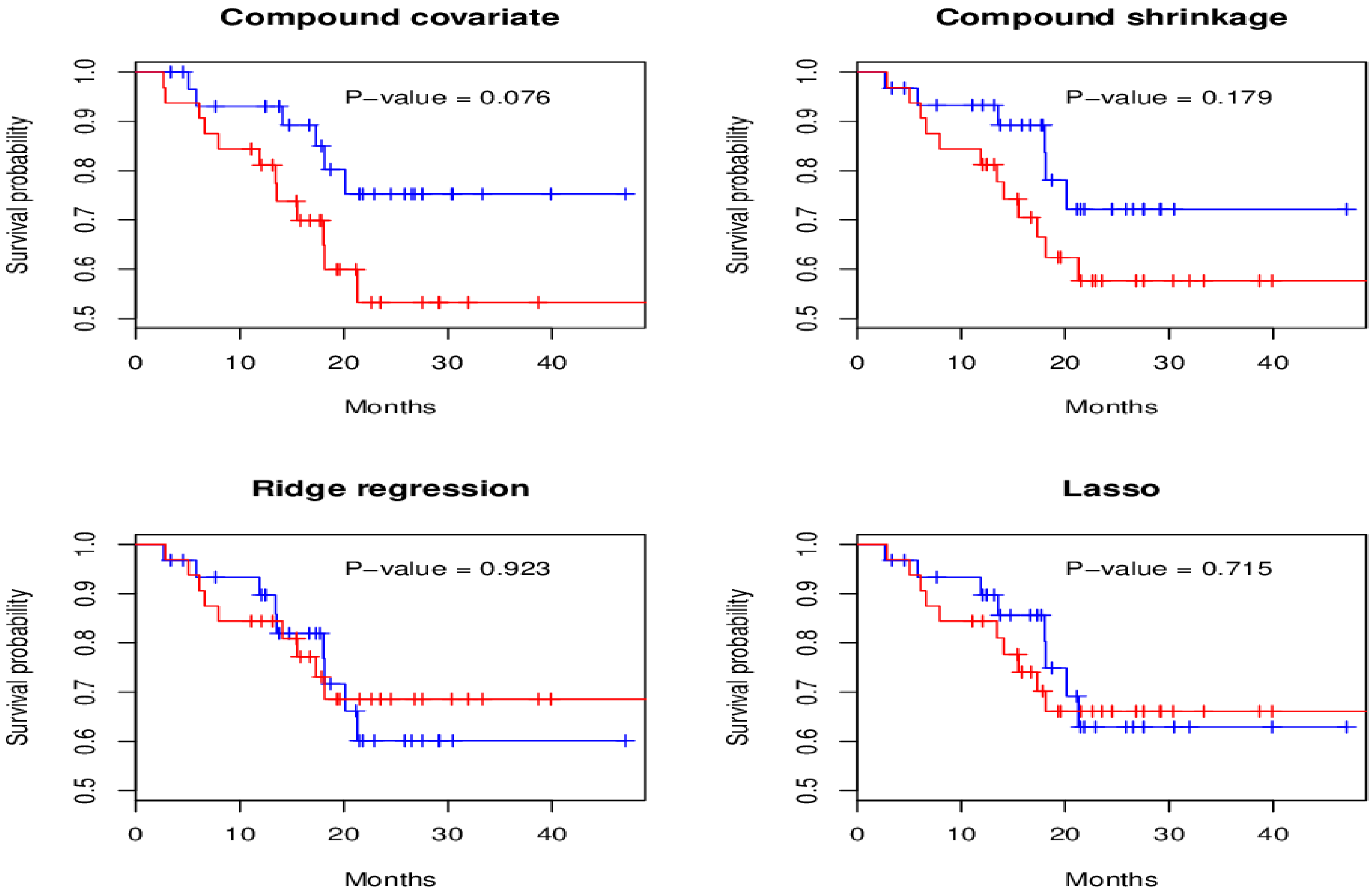
Poor prognosis

$$\hat{\boldsymbol{\beta}}' = \begin{cases} \text{compound covariate} \\ \text{compound shrinkage} \\ \text{Ridge} \\ \text{Lasso} \end{cases}$$

$$\hat{\boldsymbol{\beta}}'\mathbf{x}_i < c \ \ (\text{Good prognosis}) \ ; \ \ \hat{\boldsymbol{\beta}}'\mathbf{x}_i > c \ \ (\text{Poor prognosis}),$$

$$\text{where } c \ \text{ is the median of } \{ \hat{\boldsymbol{\beta}}'\mathbf{x}_i, \mathrm{i} = 1,\dots,\mathrm{n} \}$$

Survival curves for Poor vs. Good prognosis groups in n=62 testing data; p-value for Log-rank test



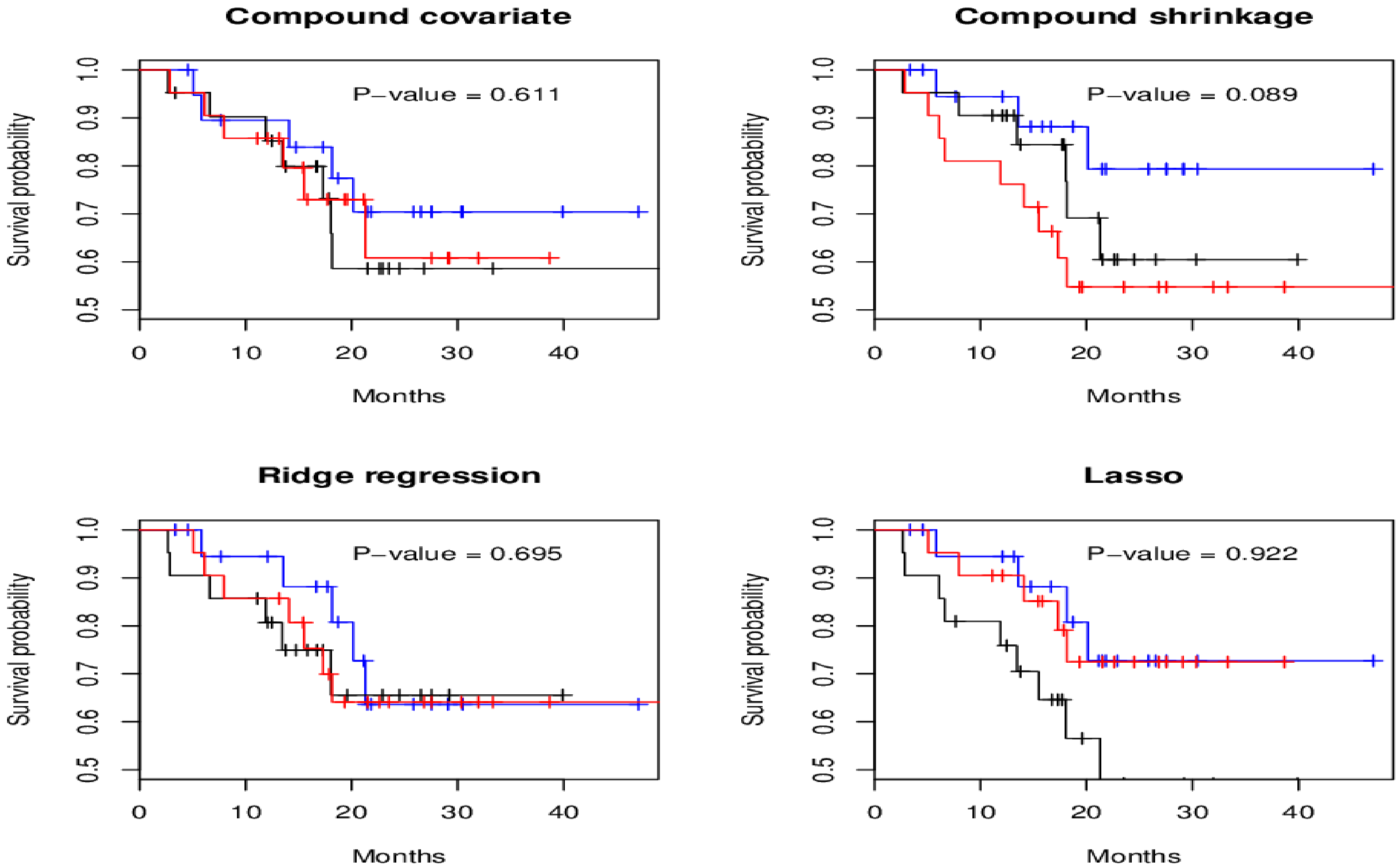
Compound covariate

Compound shrinkage

Ridge regression

Lasso

Survival curves for Poor, Medium, Good prognosis groups for n=62 testing data; p-value for Log-rank trend test



Compound covariate

Compound shrinkage

Ridge regression

Lasso

Thank you for your attention