A class of generalized ridge estimator for high-dimensional linear regression

Advisor: Takeshi Emura
Presenter: Szu-Peng Yang

June 23, 2014

Graduate Institute of Statistics, NCU

Outline

- Introduction
- Methodology
- Numerical analysis
- Conclusion

Introduction

- Introduction
- Methodology
- Numerical analysis
- Conclusion

Background

- Model

$$\mathbf{y}_{n\times 1} = X_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}, \quad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I)$$

- Least square estimator (LSE)

$$\hat{\boldsymbol{\beta}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathbf{y}$$

  - Advantage: unbiasedness, minimum variance
  - Disadvantage: high-dimensionality ($p > n$)

Performance in terms of MSE?

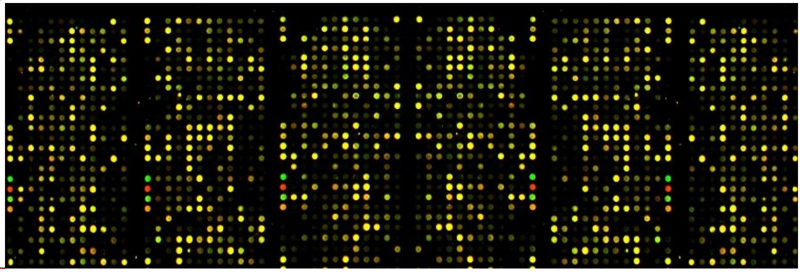
Background – High-dimensionality

- Microarray data

$$p = 394 >> n = 124$$

394 gene signatures

EGFR index
from 124 patients

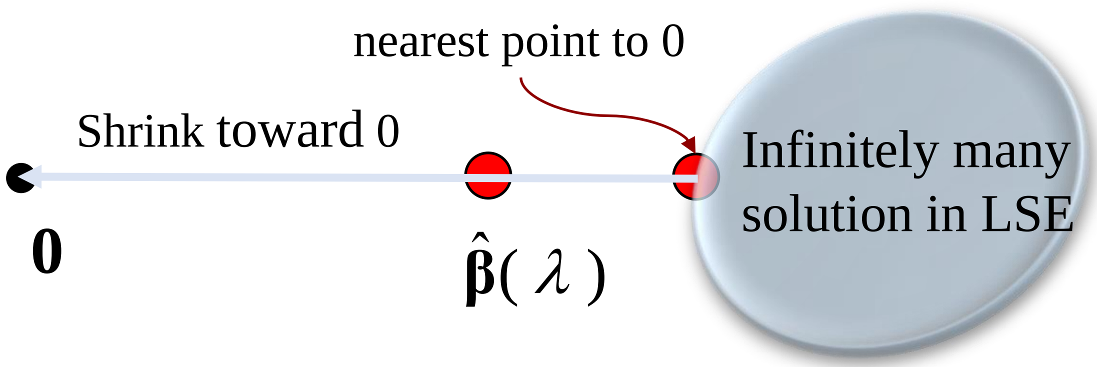


Ridge regression

- Ridge estimator – Hoerl and Kennard (1970)

$$\hat{\boldsymbol{\beta}}(\lambda) = (X^{\mathrm{T}}X + \lambda I)^{-1} X^{\mathrm{T}}\mathbf{y}, \quad \lambda > 0.$$

Shrinkage parameter



Ridge regression

Theorem 1 (Existence Theorem, Hoerl and Kennard, 1970)

There always exists a $\lambda > 0$ such that $MSE\{\hat{\boldsymbol{\beta}}(\lambda)\} < MSE(\hat{\boldsymbol{\beta}})$.

Note: $MSE(\hat{\boldsymbol{\beta}}) \equiv E\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}$

Let P be orthogonal matrix such that

$$X^{\mathrm{T}}X = P^{\mathrm{T}}\Lambda P$$

where Λ is a diagonal matrix. Then

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^{\mathrm{T}} = P\boldsymbol{\beta},$$

and $\alpha_{\max}^2 \equiv \max\{\alpha_1^2, \ldots, \alpha_p^2\}$.

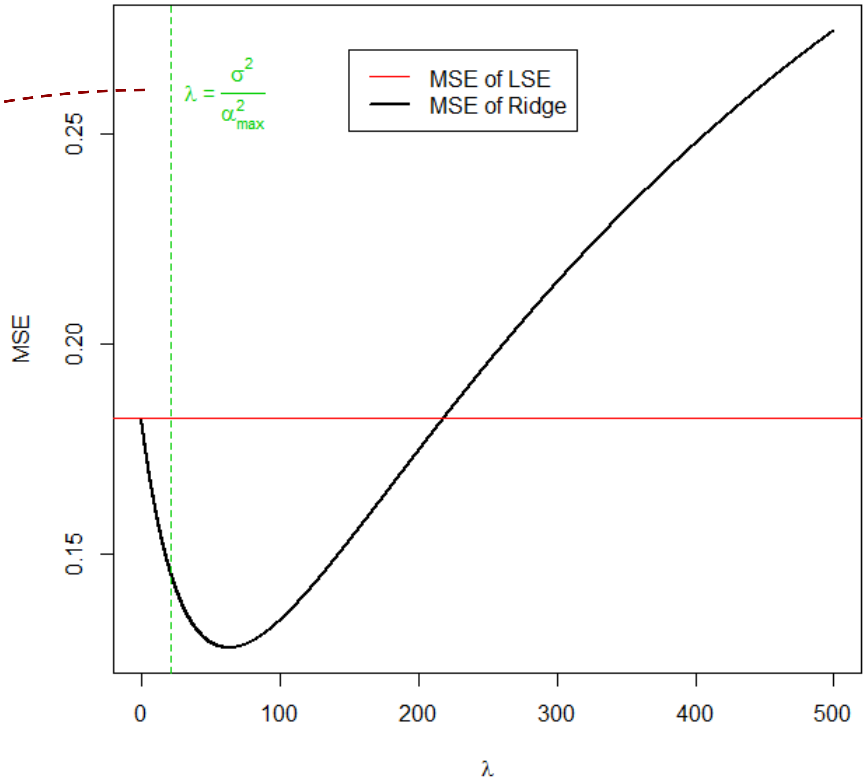


Figure 1 The plot of the MSE of the ridge estimator.

Ridge regression

- Is the ridge estimator adapt to $p > n$ case ?
  - Whittaker, Thompson and Denham (2000)
  - Zhao, Rødland and Sørlie, et al. (2011)
  - Cule, Vieis and De Iorio (2011)

    SNP data, Significance testing

Generalized ridge regression

- Generalized ridge estimator – Hoerl and Kennard (1970)

$$\hat{\boldsymbol{\beta}}(W) = (X^{\mathrm{T}}X + W)^{-1}X^{\mathrm{T}}\mathbf{y},$$

- W is a diagonal matrix.
  - Allen (1974)
  - McLachlan (1980)
  - Loesgen (1990)

Generalized ridge regression

- Benefit

1. different weights for different regressors
2. Performance in terms of MSE (McLachlan, 1980)

- Problem:

No application in $p > n$ case.

Methodology

- Introduction
- Methodology
- Numerical analysis
- Conclusion

Baysian interpretation

Good Bayesian interpretation (Loesgen, 1990):

- Prior $\boldsymbol{\beta} \sim N_p(\mathbf{0}, \sigma^2 W^{-1})$
- $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \ \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I)$

➔ Bayes estimator (posterior mean)

$$E[\boldsymbol{\beta} \mid X, \mathbf{y}] = (X^{\mathrm{T}} X + W)^{-1} X^{\mathrm{T}} \mathbf{y}$$

Baysian interpretation

- $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^{\mathrm{T}}$, $\beta_1, ..., \beta_p$ are independent

➔ $\sigma^2 W^{-1}$ is a diagonal matrix and $W = diag(w_1, ..., w_p)$.

- $W^{-1} \propto \sigma^2 W^{-1}$ (precision of prior)

➔ Precision of prior knowledge $(\beta_j = 0) \propto w_j^{-1}$

Baysian interpretation

- How to choose shrinkage parameters ?

1. More likely $\beta_j = 0$ ➔ w_j^{-1} small ➔ w_j large
2. More likely $\beta_j \neq 0$ ➔ w_j^{-1} large ➔ w_j small

Proposed method – Idea

- Define $W = diag(w_1, \ldots, w_p)$,

$$w_j = \begin{cases} \lambda\gamma & \text{if} \quad \beta_j \neq 0 \\ \lambda & \text{if} \quad \beta_j = 0 \end{cases}, \ \gamma \in [0, 1], \ \lambda > 0, \quad j = 1, \ldots, p.$$

- Problem: $\beta_1, \ldots, \beta_p$ unknown

➔ estimate W from data.

Proposed method – Estimator

- Initial estimate $\hat{\boldsymbol{\beta}}^0 = (\hat{\beta}_1^0, \ldots, \hat{\beta}_p^0)^{\mathrm{T}}$,

$$\hat{\beta}_j^0 = \frac{\mathrm{x}_j^{\mathrm{T}}\mathbf{y}}{\mathrm{x}_j^{\mathrm{T}}\mathrm{x}_j} \quad \text{for } j = 1, \ldots, p$$

where x_j, for $j = 1, \ldots, p$, are the columns of X.

- $\bar{\beta}^0 = \sum_{j=1}^{p} \hat{\beta}_j^0 \Big/ p$
- $se(\hat{\boldsymbol{\beta}}^0) = \sqrt{\sum_{j=1}^{p} (\hat{\beta}_j^0 - \bar{\beta}^0)^2 \Big/ (p-1)}$
- Under the sparse model ($\boldsymbol{\beta} \approx \mathbf{0}$),

$$\hat{\beta}_j^0 / se(\hat{\boldsymbol{\beta}}^0) \dot{\sim} N(0,1)$$

Proposed method – Estimator

- Proposed estimator

$$\hat{\boldsymbol{\beta}}(\lambda, \Delta) = \{ X^{\mathrm{T}} X + \lambda \hat{W}(\Delta) \}^{-1} X^{\mathrm{T}} \mathbf{y}, \quad \Delta \geq 0,$$

where $\hat{W}(\Delta) = diag\{ \hat{w}_1(\Delta), \ldots, \hat{w}_p(\Delta) \}$ and

$$\hat{w}_j(\Delta) = \begin{cases} 1/2 & \text{if} \quad |\hat{\beta}_j^0| / se(\hat{\boldsymbol{\beta}}^0) \geq \Delta, \\ 1 & \text{otherwise}, \end{cases}$$

Thresholding parameter

for $j = 1, \ldots, p.$

Proposed method – Parameters estimation

- Allen's PRESS (1974)

  generalized cross-validation (GCV) (Golub, Heath and Wahba, 1979)

- Minimizer of GCV

- $\hat{\beta}_j^0 / se(\hat{\boldsymbol{\beta}}^0) \sim N(0,1)$ ➔ Choose $\hat{\Delta}$ among $D = \{0, 3/100, \ldots, 300/100\}$

Table 1 Formula of GCV function

Ridge estimator	$V(\lambda) = \frac{1}{n} \| \{ I - A(\lambda) \} \mathbf{y} \|^2 \Big/ \left[\frac{1}{n} \mathrm{Tr}\{ I - A(\lambda) \} \right]^2$ $A(\lambda) = X(X^{\mathrm{T}} X + \lambda I)^{-1} X^{\mathrm{T}}$
Proposed estimator	$V(\lambda, \Delta) = \frac{1}{n} \| \{ I - A(\lambda, \Delta) \} \mathbf{y} \|^2 \Big/ \left[\frac{1}{n} \mathrm{Tr}\{ I - A(\lambda, \Delta) \} \right]^2$ $A(\lambda, \Delta) = X\{ X^{\mathrm{T}} X + \lambda \hat{W}(\Delta) \}^{-1} X^{\mathrm{T}}$

MSE comparison

- $MSE(\hat{\boldsymbol{\beta}}) = Tr\{\, MSEM(\hat{\boldsymbol{\beta}})\,\}$

Mean square error matrix

- Trenkler (1985)
- Loesgen (1990)

$$MSEM(\hat{\boldsymbol{\beta}}) = E\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}\}$$
$$= C + \mathbf{d}\mathbf{d}^{\mathrm{T}}$$

$$C = Cov(\hat{\boldsymbol{\beta}})$$
$$= E\{\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\}\{\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\}^{\mathrm{T}}$$

$$\mathbf{d} = Bias(\hat{\boldsymbol{\beta}}) = E(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}$$

Mean square error

$$MSE(\hat{\boldsymbol{\beta}}) = E\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\}$$

MSE comparison

Lemma 1 (Trenkler, 1985)

Suppose A is a symmetric $n \times n$ matrix, $\mathbf{a}$ is an $n \times 1$ vector and γ is a positive real number. Then $\gamma A - \mathbf{a}\mathbf{a}^{\mathrm{T}}$ is n.n.d. if and only if

1) A is n.n.d. 2) $\mathbf{a} = A\mathbf{v}$ for some $\mathbf{v} \in R^p$ 3) $\mathbf{a}^{\mathrm{T}} A^{-} \mathbf{a} \le \gamma$

where A^{-} is the generalized inverse of A.

Note: the diagonals of a nonnegative definite (n.n.d.) matrix are nonnegative.

MSE comparison

- $\hat{\boldsymbol{\beta}}_i = (X^{\mathrm{T}}X + W_i)^{-1}X^{\mathrm{T}}\mathbf{y}$ for $i = 1, 2$.
- $MSEM(\hat{\boldsymbol{\beta}}_1) - MSEM(\hat{\boldsymbol{\beta}}_2) = (C_1 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}}) - (C_2 + \mathbf{d}_2\mathbf{d}_2^{\mathrm{T}})$

Theorem 2

$MSEM(\hat{\boldsymbol{\beta}}_1) - MSEM(\hat{\boldsymbol{\beta}}_2)$ is n.n.d. if all the three conditions hold:

1) $(C_1 - C_2)/\sigma^2$ is .n.n.d.
2) $C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}}$ is full rank
3) $\mathbf{d}_2^{\mathrm{T}}(C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}})^{-1}\mathbf{d}_2 \leq 1$

MSE comparison – Example

- $\boldsymbol{\beta} = (\beta_1, 0)^{\mathrm{T}}, \quad \beta_1 \neq 0$
- $W_1 = I = diag(1, 1)$

  $W_2 = diag(2.5\gamma, 2.5), \; 0 < \gamma < 1$
- $X^{\mathrm{T}}X = I$ (For simplicity)

$\Rightarrow$ $(C_1 - C_2)/\sigma^2 = diag\left(\frac{1}{4} - \frac{1}{(1+2.5\gamma)^2}, \frac{33}{196}\right)$

$$C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}} = diag\left(\frac{\sigma^2\{(1+2.5\gamma)^2 - 4\} + (1+2.5\gamma)^2\beta_1^2}{4(1+2.5\gamma)^2}, \frac{33}{196}\right)$$

$$\mathbf{d}_2^{\mathrm{T}}(C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}})^{-1}\mathbf{d}_2 = \frac{25\gamma^2\beta_1^2}{\sigma^2\{(1+2.5\gamma)^2 - 4\} + (1+2.5\gamma)^2\beta_1^2}$$

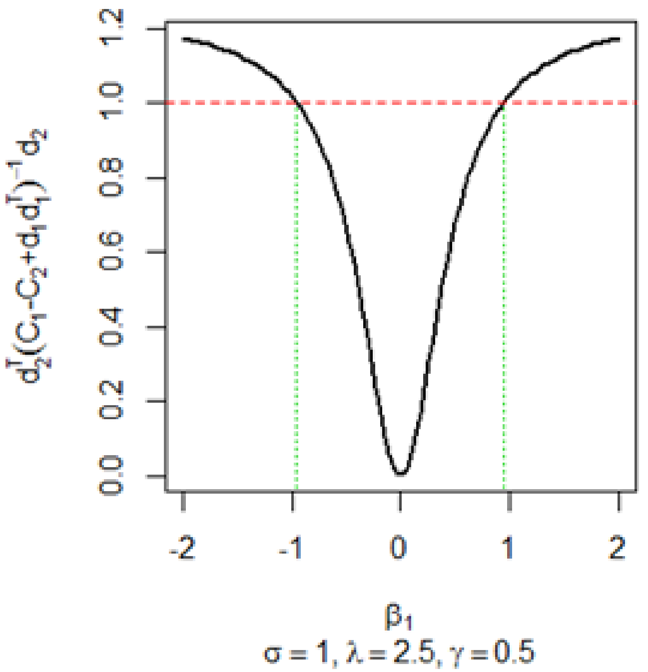


Figure 2 The plot of $\mathbf{d}_2^{\mathrm{T}}(C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}})^{-1}\mathbf{d}_2$ against β_1.

Result:

$|\beta_1|$ not too large, $\gamma \geq 2/5$ $\Rightarrow$ $MSEM(\hat{\boldsymbol{\beta}}_1) - MSEM(\hat{\boldsymbol{\beta}}_2)$ n.n.d.

$\Rightarrow$ $MSE(\hat{\boldsymbol{\beta}}_1) \geq MSE(\hat{\boldsymbol{\beta}}_2)$

Significance testing (Cule, Vineis and De Iorio, 2011)

- Hypothesis

$$H_{0j}: \beta_j = 0 \quad \text{v.s.} \quad H_{1j}: \beta_j \neq 0, \quad j = 1, \ldots, p$$

- Wald test statistics

$$Z_j = \frac{\hat{\beta}_j(\hat{\lambda}, \hat{\Delta})}{se\{\hat{\beta}_j(\hat{\lambda}, \hat{\Delta})\}} \overset{H_{0j}}{\dot{\sim}} N(0,1), \quad j = 1, \ldots, p$$

- Two-sided P-value

$$P_j = \Pr(Z > |Z_j| \; or \; Z < -|Z_j|) = 2 \times \Pr(Z > |Z_j|)$$

Numerical analysis

- Introduction
- Methodology
- **Numerical analysis**
- Conclusion

Simulation – Model setting

- Sparse model
  - Emura, Chen and Chen (2012)
  - Ing and Lai (2011)
  - Binder, et al. (2009)

Correlated regressors

- $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \ \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, I)$

- $\boldsymbol{\beta} = (\underbrace{\overbrace{b/q, \ldots, b/q}^{q}}_{Corr = 0.5}, \underbrace{\overbrace{d/r, \ldots, d/r}^{r}}_{Corr = 0.5}, \overbrace{0, \ldots, 0}^{p-(q+r)})^{\mathrm{T}}$

Simulation – Model setting

- $n = 100, \quad p \in \{50, 100, 150, 200\}, \quad q = r = 10$
- $\boxed{\text{Case I} \qquad b = d = 5}$

  Case II $\qquad b = d = 10$

  Case III $\qquad b = 5, \; d = -5$

  Case IV $\qquad b = 10, \; d = -10$

Simulation – MSE comparison

- $MSE(\hat{\boldsymbol{\beta}}(\lambda)) = E\{(\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta})^{\mathrm{T}}(\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta})\}$
- $MSE(\hat{\boldsymbol{\beta}}(\lambda, \Delta^{*})) = E\{(\hat{\boldsymbol{\beta}}(\lambda, \Delta^{*}) - \boldsymbol{\beta})^{\mathrm{T}}(\hat{\boldsymbol{\beta}}(\lambda, \Delta^{*}) - \boldsymbol{\beta})\}$

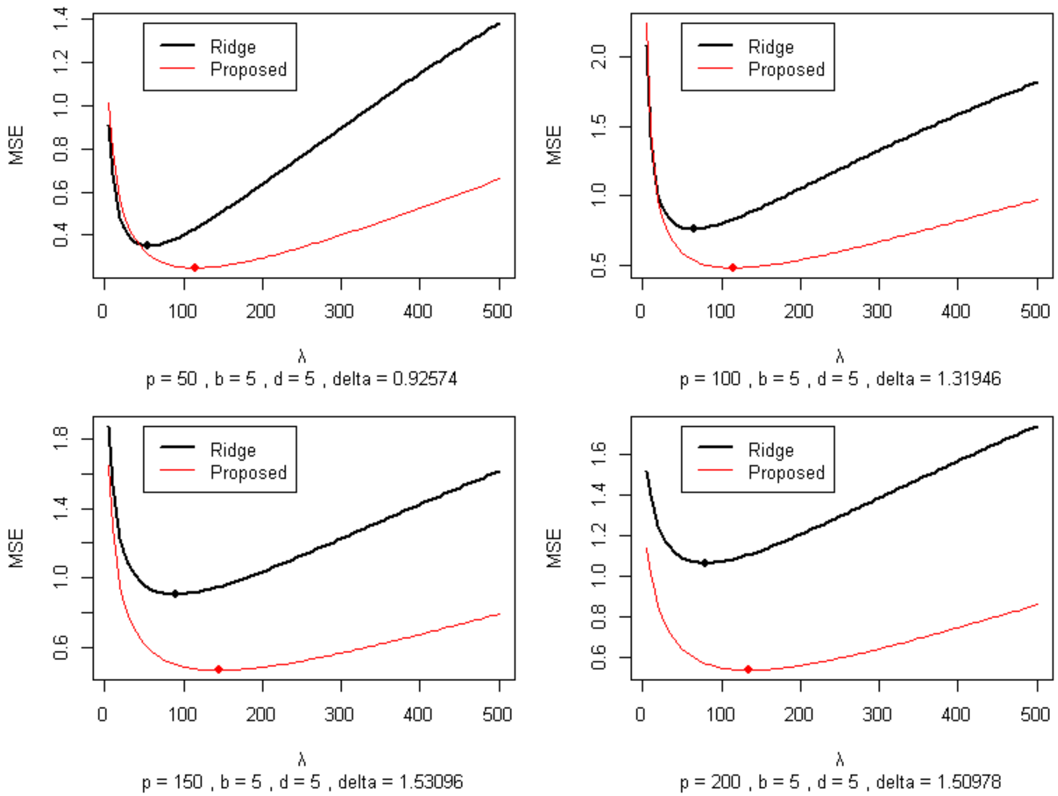


Figure 2 The plot of the MSE curve with $b=d=5$.

Simulation – MSE comparison

- $MSE(\hat{\boldsymbol{\beta}}(\hat{\lambda})) = E\{(\hat{\boldsymbol{\beta}}(\hat{\lambda}) - \boldsymbol{\beta})^{\mathrm{T}}(\hat{\boldsymbol{\beta}}(\hat{\lambda}) - \boldsymbol{\beta})\}$
- $MSE(\hat{\boldsymbol{\beta}}(\hat{\lambda}, \hat{\Delta})) = E\{(\hat{\boldsymbol{\beta}}(\hat{\lambda}, \hat{\Delta}) - \boldsymbol{\beta})^{\mathrm{T}}(\hat{\boldsymbol{\beta}}(\hat{\lambda}, \hat{\Delta}) - \boldsymbol{\beta})\}$
- $MSE(\hat{\beta}_j) = E(\hat{\beta}_j - \beta_j)^2, \quad j = 1, \ldots, p$

Table 2 Simulation results based on 500 replicates with *b*=*d*=5.

setting	estimate	$E(\hat{\lambda})$	$E(\hat{\Delta})$	$MSE(\hat{\beta}_1)$	$MSE(\hat{\beta}_p)$	$MSE(\hat{\boldsymbol{\beta}})$
$p = 50$	Ridge	23.2051		0.0112	0.0077	0.4663
	Proposed	47.3432	0.9257	0.0107	0.0048	0.3763
$p = 100$	Ridge	23.8657		0.0080	0.0086	1.0228
	Proposed	46.9601	1.3195	0.0086	0.0058	0.7191
$p = 150$	Ridge	20.4585		0.0118	0.0065	1.2909
	Proposed	43.8348	1.5310	0.0171	0.0037	0.7822
$p = 200$	Ridge	10.6090		0.0036	0.0046	1.4137
	Proposed	30.8835	1.5098	0.0043	0.0027	0.8364

Simulation – Significance testing

- Hypothesis

$$H_{0,50}: \beta_{50} = 0 \quad \text{v.s.} \quad H_{1,50}: \beta_{50} \neq 0$$

Note: $\beta_{50} = 0$,

$H_{0,50}$ is true

- Wald test statistic

$$Z_{50} = \frac{\hat{\beta}_{50}(\hat{\lambda}, \hat{\Delta})}{se\{\hat{\beta}_{50}(\hat{\lambda}, \hat{\Delta})\}} \overset{H_{0,50}}{\dot{\sim}} N(0,1)$$

- Type I error

$$\frac{\sum_{s=1}^{500} \mathbf{I}(|Z_{50}^{(s)}| > Z_{\alpha/2})}{500}$$

$Z_{\alpha/2} \equiv$ the upper $\alpha/2$ percent point of $N(0,1)$

Simulation – Significance testing

- $\alpha = 0.05$

Table 3 Simulation results for significance testing based on 500 replicates with *b*=*d*=*5*.

	$E(\hat{\beta}_{50})$	$sd(\hat{\beta}_{50})$	$E(Z_{50})$	$sd(Z_{50})$	Type I error
$p = 50$	-0.0027	0.0691	-0.0424	0.9683	0.042
$p = 100$	-0.0075	0.0660	-0.1137	0.9412	0.028
$p = 150$	-0.0222	0.0549	-0.3889	0.8860	0.046
$p = 200$	0.0332	0.0515	0.5549	0.8414	0.048

Note:

1. $E(\hat{\beta}_{50}) = \bar{\hat{\beta}}_{50} = \sum_{s=1}^{500} \hat{\beta}_{50}^{(s)} / 500$
2. $sd(\hat{\beta}_{50}) = \sqrt{\sum_{s=1}^{500} (\hat{\beta}_{50}^{(s)} - \bar{\hat{\beta}}_{50})^2 / 499}$
3. $E(Z_{50}) = \bar{\hat{Z}}_{50} = \sum_{s=1}^{500} Z_{50}^{(s)} / 500$
4. $sd(Z_{50}) = \sqrt{\sum_{s=1}^{500} (Z_{50}^{(s)} - \bar{\hat{Z}}_{50})^2 / 499}$

Simulation – Significance testing

- Hypothesis

$$H_{0,1} : \beta_1 = 0 \quad \text{v.s.} \quad H_{1,1} : \beta_1 \neq 0$$

- Wald test statistic

$$Z_1 = \frac{\hat{\beta}_1(\hat{\lambda}, \hat{\Delta})}{se\{\hat{\beta}_1(\hat{\lambda}, \hat{\Delta})\}} \overset{H_{0,1}}{\dot{\sim}} N(0,1)$$

- Power

$$\frac{\sum_{s=1}^{500} \mathbf{I}(|Z_1^{(s)}| > Z_{\alpha/2})}{500}$$

$Z_{\alpha/2} \equiv$ the upper $\alpha/2$ percent point of $N(0,1)$

Note: $\beta_1 \neq 0$,
$H_{1,1}$ is true

Simulation – Significance testing

- $\alpha = 0.05$

Table 4 Simulation results for significance testing based on 500 replicates with *b*=*d*=*5*.

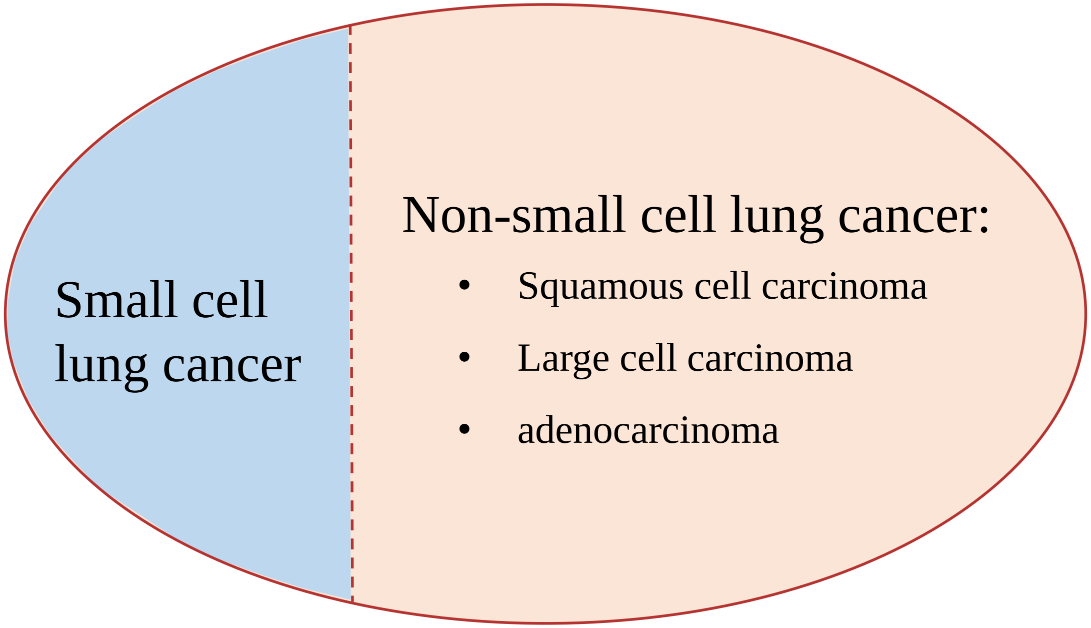
	$E(\hat{\beta}_1)$	$sd(\hat{\beta}_1)$	$E(Z_1)$	$sd(Z_1)$	Power
$p = 50$	0.4647	0.0973	4.9996	1.0090	0.998
$p = 100$	0.5135	0.0918	5.9685	0.9981	0.998
$p = 150$	0.4111	0.0958	5.6371	0.9409	1
$p = 200$	0.4998	0.0657	6.5303	0.9281	1

Note:

1. $E(\hat{\beta}_1) = \bar{\hat{\beta}}_1 = \sum_{s=1}^{500} \hat{\beta}_1^{(s)} / 500$
2. $sd(\hat{\beta}_1) = \sqrt{\sum_{s=1}^{500} (\hat{\beta}_1^{(s)} - \bar{\hat{\beta}}_1)^2 / 499}$
3. $E(Z_1) = \bar{\hat{Z}}_1 = \sum_{s=1}^{500} Z_1^{(s)} / 500$
4. $sd(Z_1) = \sqrt{\sum_{s=1}^{500} (Z_1^{(s)} - \bar{\hat{Z}}_1)^2 / 499}$

NSCLC data

Epithelial lung cancer



$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

394 selected gene signatures

EGFR index from 124 patients

$\Rightarrow$ $n = 124, \quad p = 394$

NSCLC data – 4-fold cross validation

1	2	3	4
Train $\mathfrak{I}_1$ (31 patients)	Train $\mathfrak{I}_2$ (31 patients)	Test $\mathfrak{I}_3$ (31 patients)	Train $\mathfrak{I}_4$ (31 patients)

- $\mathbf{x}_i \equiv i$th row of X
- $\hat{\boldsymbol{\beta}}^{(-k)} \equiv$ estimator based on all the data not in $\mathfrak{I}_k$
- Prediction error $PE = \dfrac{1}{124}\sum_{k=1}^{4}\sum_{i\in\mathfrak{I}_k}(y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{(-k)})^2$

NSCLC data – 4-fold cross validation

Table 5 *PE* comparison of the ridge regression and the proposed method over 100 random cross validation.

No. of replicate	$\hat{\lambda}^{ridge}$	$\hat{\lambda}^{proposed}$	$\hat{\Delta}^{proposed}$	PE^{ridge}	$PE^{proposed}$
1	294.401	410.763	1.448	0.502	0.454
2	258.612	349.376	1.418	0.703	0.753
3	315.201	431.229	1.598	0.481	0.441
4	310.829	419.522	1.598	0.495	0.452
5	306.745	418.203	1.545	0.471	0.441
6	325.630	447.352	1.538	0.518	0.472
7	312.526	442.847	1.500	0.474	0.433
8	323.645	435.895	1.523	0.476	0.438
9	307.684	423.964	1.470	0.507	0.464
10	303.928	426.297	1.313	0.472	0.436
≈	≈	≈	≈	≈	≈
99	320.884	441.779	1.448	0.481	0.442
100	285.245	393.704	1.583	0.505	0.461
Average	307.035	422.718	1.482	0.494	0.456

NSCLC data – Prediction

Regressors selection

Estimation → Prediction

62 patients (Training sample)	62 patients (Testing sample)

NSCLC data – Prediction

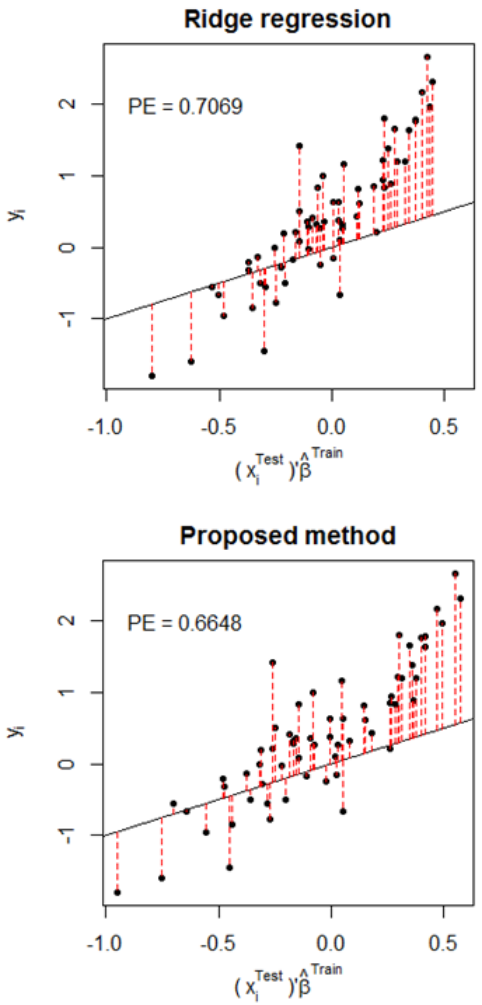


Figure 3 The plots of y_i against $(\mathbf{x}_i^{Test})^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{Train}$

Table 6 The 20 most strongly associated genes

	Ridge			Proposed method		
No.	Gene symbol	Coefficient	P-value	Gene symbol	Coefficient	P-value
1	FGA	-0.0381	2.6050E-07	FGA	-0.0507	3.7370E-07
2	AKR1B10	-0.0462	7.9121E-07	AKR1B10	-0.0590	1.7486E-06
3	CPS1	-0.0411	2.5736E-05	CPS1	-0.0562	2.6618E-05
4	KRT6A	-0.0345	4.5128E-05	FGG	-0.0465	8.0691E-05
5	MSMB	-0.0446	0.0001	MSMB	-0.0603	0.0001
6	FGG	-0.0337	0.0001	KRT6A	-0.0445	0.0001
7	CYP2B7P1	0.0302	0.0002	CYP2B7P1	0.0413	0.0004
8	SERPINB5	-0.0290	0.0002	FGB	-0.0372	0.0007
9	FGB	-0.0285	0.0003	CYP2B6	0.0163	0.0009
10	CYP2B6	0.0232	0.0005	SERPINB5	-0.0339	0.0018
11	LOC344887	-0.0302	0.0011	GPR110	0.0427	0.0022
12	SERPINB3	-0.0263	0.0014	LOC344887	-0.0366	0.0023
13	GPR110	0.0310	0.0022	CRP	-0.0206	0.0034
14	HSD17B6	0.0260	0.0035	SERPINB3	-0.0318	0.0044
15	CRP	-0.0138	0.0037	SLC6A14	-0.0210	0.0051
16	DKK1	-0.0389	0.0041	HSD17B6	0.0349	0.0058
17	**SLC34A2**	0.0217	0.0043	DKK1	-0.0492	0.0066
18	MUC13	-0.0318	0.0045	MUC13	-0.0408	0.0077
19	CPN1	-0.0083	0.0054	CPN1	-0.0115	0.0108
20	SLC6A14	-0.0273	0.0080	**(7895136)**	0.0242	0.0121
PE			0.7069			0.6648

Conclusion

- Introduction
- Methodology
- Numerical analysis
- Conclusion

Conclusion

- Proposed estimator under high-dimensionality
- Able to utilize prior knowledge
- Good Baysian interpretation and theoretical calculation
- Reduction of the MSE
- Significance testing
- Apply to microarray data

Reference

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16**, 125-127.

Binder, H., Allignol, A., Schumacher, M., and Beyersmann, J. (2009). Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics* **25**, 890-896.

Cule, E., Vineis, P. and De Iorio, M. (2011). Significance testing in ridge regression for genetic data. *BMC Bioinformatics* **12**, 372.

Emura, T., Chen, Y. H., and Chen, H. Y. (2012). Survival prediction based on compound covariate under cox proportional hazard models. *PLoS ONE* **7**, e47627.

Golub, G. H., Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215-223.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67.

Ing, C.-K. and Lai, T.-L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica* **21**, 1473-1513.

Loesgen, K.-H. (1990). A generalization and Bayesian interpretation of ridge-type estimators with good prior means. *Statistical Papers* **31**, 147-154.

McLachlan, G. J. (1980). On the mean square error associated with adaptive generalized ridge regression. *Biometrical Journal* **22**, 125-129.

Trekler, G. (1985). Mean square error matrix comparisons of estimators in linear regression. *Communications in Statistics A***14**, 2495-2509.

Whittaker, J. C., Thompson, R., and Denham, M. C., (2000). Marker-assisted selection using ridge regression. *Genetical Research* **75**, 249-252.

Zhao, X., Rødland, E. A., and Sørlie, T., et al. (2011). Combining gene signatures improves prediction of breast cancer survival. *PLoS ONE* **6**, e17845.

Thank you for your listening.