An Improved Generalized Ridge Estimator for High-dimensional Linear regression

Szu-Peng Yang

Advisor: Takeshi Emura

Graduate Institute of Statistics, National Central University

Abstract: In multiple linear regression, the least square estimator (LSE) is inappropriate for high-dimensional regressors. Hoerl and Kennard (1970) established that there exist some shrinkage parameter greater than zero such that the mean square error (MSE) of ridge estimator is less than that of the LSE. Here, we propose a generalized ridge estimator, which gives unequal shrinkage parameters using some thresholding technique. We choose the shrinkage and thresholding parameters through the generalized cross-validation (GCV)[2]. We also consider significance testing based on the proposed estimator. Simulations show that the new estimator performs better than the ridge in terms of MSE criterion, even when p is greater than n. We demonstrate the method using the non-small cell lung cancer data.

1. Introduction

Ridge regression is an effective method when the number of regressors is large than the sample size ($p > n$)[3]. Ridge regression is originally derive by Hoerl and Kennard (1970) to reduce the collinearity of LSE. They established that ridge estimator has less MSE than that of the LSE. In addition, in $p > n$ case, the ridge estimator is workable, but the LSE is not. Cule et al. (2011) implemented the ridge estimator on the data obtained by using microarray and further tested the significance of each component of it.

Generalized ridge regression[1] (GRR) is more flexible than the ordinary one. In the GRR, the identical matrix multiplied by λ is replaced by a diagonal matrix K with nonnegative elements. However, the generalized ridge regression has not been applied to the case of $p > n$. This is because the GRR involves a large number of shrinkage parameters, making it difficult to handle high-dimensional regressors. Hence, we propose a special class of generalized ridge estimator that adapts to high-dimenionality.

2. Proposed method

We consider the linear model $\mathbf{y}_{n\times 1} = X_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$ where $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, I)$. Then the ridge estimator is

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

2.1 Proposed estimator

We propose an estimator which reduces the MSE. Let the $\hat{\boldsymbol{\beta}}^0 = (\hat{\beta}_1^0, ..., \hat{\beta}_p^0)^{\mathrm{T}}$ denotes the initial estimate, where

$$\hat{\beta}_j^0 = \mathrm{x}_j^{\mathrm{T}}\mathbf{y}/\mathrm{x}_j^{\mathrm{T}}\mathrm{x}_j, \quad j = 1, ..., p,$$

and x_j denotes the jth column of design matrix X. Then the proposed estimator is

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

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

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

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

2.2 Generalized cross-validation

Golub et al. (1979) proposed the minimum of generalized cross-validation (GCV) function to estimate the λ for ridge estimator. The GCV function is defined as

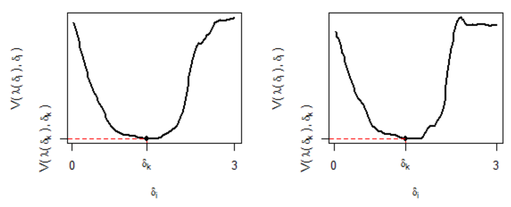
$$V(\lambda) = \frac{1}{n}\|\{I - A(\lambda)\}\mathbf{y}\|^2 \Big/ \left[\frac{1}{n}\mathrm{Tr}\{I - A(\lambda)\}\right]^2$$

where $A(\lambda) = X(X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}$. As the formula for ridge, we defined the GCV function for the proposed estimator as follow

$$V(\lambda, \Delta) = \frac{1}{n}\|\{I - A(\lambda, \Delta)\}\mathbf{y}\|^2 \Big/ \left[\frac{1}{n}\mathrm{Tr}\{I - A(\lambda, \Delta)\}\right]^2$$

where $A(\lambda, \Delta) = X\{X^{\mathrm{T}}X + \lambda\hat{W}(\Delta)\}^{-1}X^{\mathrm{T}}$.

The following figure shows that the minimum of GCV function is available for several cases.



2.3 Significance test

As in Cule et al. (2011), we implement the significance test for the proposed method. Consider the null hypothesis $H_{0j}: \beta_j = 0, \ j = 1, ..., p$. Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_p)^{\mathrm{T}}$ be either ridge or proposed estimator. Then the Wald test statistic is

$$Z_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \overset{H_{0j}}{\sim} N(0, 1).$$

In addition, we calculate the P-value for Z_j. Let Z be a random variable from the standard normal distribution. Then, the two-sided P-value is

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

3. Simulation

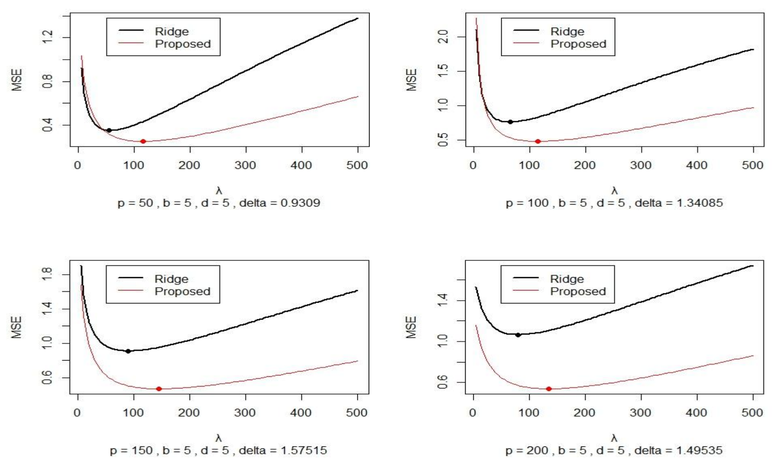
Throughout the simulation, we consider the sparse model with some groups of correlated regressors[4]. And we set $n = 100$ and $p \in \{50, 100, 150, 200\}$.

3.1 Mean square error

We evaluate the performance of estimators through MSE which is evaluated by

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

First, with fixed λ and Δ, the graphs below show that if the $(\hat{\lambda}, \hat{\Delta})$ is chosen properly, the MSE of the proposed estimate may be less than that of the ridge estimate. Second, directly estimating $(\hat{\lambda}, \hat{\Delta})$ by GCV[2], we have the results in table below. The MSE of the proposed method is always less than that of the ridge.



Comparison based on $n = 100$ sample and 200 replicates with different p.

setting	estimate	$E(\hat{\lambda})$	$E(\hat{\Delta})$	$MSE(\hat{\beta}_1)$	$MSE(\hat{\beta}_p)$	$MSE(\hat{\boldsymbol{\beta}})$
$p = 50$	Ridge	23.1953		0.0126	0.0086	0.4699
	Proposed	47.4188	0.9309	0.0121	0.0052	0.3818
$p = 100$	Ridge	24.5518		0.0087	0.0086	1.0246
	Proposed	48.0316	1.3409	0.0094	0.0055	0.7188
$p = 150$	Ridge	21.7419		0.0124	0.0069	1.2722
	Proposed	45.6736	1.5752	0.0192	0.0040	0.7616
$p = 200$	Ridge	11.4202		0.0038	0.0045	1.4059
	Proposed	32.0577	1.4954	0.0046	0.0027	0.8372

3.2 Significance testing

In order to know the precision of the test, the type I error,

$$\sum_{s=1}^{500} I(P_{50,s} < 0.05)/500,$$

is displayed in the table below, along with the averaged of $\hat{\beta}_{50}$ and Z_{50}.

Results of significance testing for the proposed estimator based on 500 replicates.

	$E(\hat{\beta}_{50})$	$sd(\hat{\beta}_{50})$	$E(Z_{50})$	$sd(Z_{50})$	Type I error at $\alpha = 0.05$
$p = 50$	-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

4. Data analysis

We use the non-small cell lung cancer (NSCLC) data which is available at http://www.ncbi.nlm.nih.gov/geo/ with accession number GSE33072. There are 131 patients with 33297 gene signatures in raw data. But we used only 124 patients along with 394 gene signatures. We view the EGFR index as the response and the gene signatures as the regressors.

4.1 4-fold cross-validation

1	2	3	4
Train $\Im_1$ (31 patients)	Train $\Im_2$ (31 patients)	Test $\Im_3$ (31 patients)	Train $\Im_4$ (31 patients)

In stead of the MSE, here, we use the prediction error (PE) to evaluate the performance of the estimator. The estimated PE is evaluated by

$$PE = \sum_{k=1}^{4}\sum_{i\in\Im_k}(y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{(-k)})^2/124.$$

The following table displays the results.

Comparison of the ridge and proposed method over 100 random cross validation.

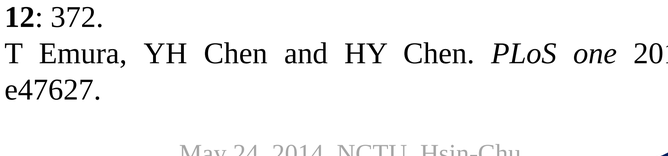
No. of rep.	$\hat{\lambda}^{ridge}$	$\hat{\lambda}^{proposed}$	$\hat{\Delta}^{proposed}$	PE^{ridge}	$PE^{proposed}$
1	294.401	410.763	1.448	0.502	0.454
2	258.612	349.376	1.418	0.703	0.753
3	315.201	431.229	1.598	0.481	0.441
4	310.829	419.522	1.598	0.495	0.452
≈	≈	≈	≈	≈	≈
100	285.245	393.704	1.583	0.505	0.461
Average	307.035	422.718	1.482	0.494	0.456

4.2 Regressor selection

By applying the significance testing, we are able to select few gene signatures which are most strongly associated with the EGFR index. The following table display the information of top 20 selected gene signatures. And the following plot demonstrates that the predictors from both ridge and proposed method are positively associated with the EGFR index.

The 20 most strongly associated genes based on ridge and proposed methods in order of.P-value

	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
≈	≈	≈	≈	≈	≈	≈
20	SLC6A14	-0.0273	0.0080	7895136 (ID_REF)	0.0242	0.0121
PE			0.7069			0.6648



5. Conclusion

1. We propose an estimation method of regression coefficient.
2. The method reduce the MSE much in all case and give a fine prediction (simulations).
3. The PE of the proposed estimate is almost always less than that of ridge, that is, give a better prediction (data analysis).

6. Reference

1. AE Hoerl and RW Kennard. Technometrics 1970, **12**: 55-67.
2. GH Golub, M Heath and G Wahba. *Technometrics* 1979, **21***:* 215-223.
3. E Cule, P Vineis and MD Lorio. *MBC Bioinformatrics* 2011, **12**: 372.
4. T Emura, YH Chen and HY Chen. *PLoS one* 2012, **7**: e47627.