COMMUNICATIONS IN STATISTICS—SIMULATION AND COMPUTATION®
2017, VOL. 46, NO. 8, 6083–6105
http://dx.doi.org/10.1080/03610918.2016.1193195

Taylor & Francis
Taylor & Francis Group

A Bayesian approach with generalized ridge estimation for high-dimensional regression and testing

Szu-Peng Yang and Takeshi Emura

Graduate Institute of Statistics, National Central University, Taiwan

ABSTRACT

This paper adopts a Bayesian strategy for generalized ridge estimation for high-dimensional regression. We also consider significance testing based on the proposed estimator, which is useful for selecting regressors. Both theoretical and simulation studies show that the proposed estimator can simultaneously outperform the ordinary ridge estimator and the LSE in terms of the mean square error (MSE) criterion. The simulation study also demonstrates the competitive MSE performance of our proposal with the Lasso under sparse models. We demonstrate the method using the lung cancer data involving high-dimensional microarrays.

ARTICLE HISTORY
Received 28 September 2015
Accepted 19 May 2016

KEYWORDS
Bayes estimator; Compound covariate estimator; Linear model; Mean square error; Shrinkage estimator; Statistical decision theory

1. Introduction

When the number of regressors, p exceeds the sample size n (i.e., $p > n$), the least squares estimator (LSE) is not suitable to estimate regression coefficients in a linear model. There exists a large number of variable reduction methods to deal with the "high-dimensional" $p > n$ setting, including forward selection, Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), Dantzig Selector (Candes and Tao, 2007), SIS (Fan and Lv, 2008), just to name a few. See also the book of Hastie *et al.* (2009). As pointed out by Bühlmann (2013), the variable selection methods rarely address the uncertainty of the regressors (i.e., P-value). Concretely, if a regressor variable is selected, it is considered statistically significant without being quantified by P-values. Alternatively, one can perform ridge regression for significance testing for each regression coefficient. This method allows one to access P-values of all the p regressors (Bühlmann 2013; Cule et al. 2011; Cule and De Lorio 2013). See also the method of Zhang and Zhang (2014).

Ridge regression is an effective method when the number of regressors is larger than the sample size ($p > n$). Ridge regression was due to Hoerl and Kennard (1970) and was developed to reduce the multicollinearity problem for the linear regression model. Later, the ridge estimator is theoretically shown to work even under the $p > n$ case (Golub *et al.*, 1979). Ridge regression is a shrinkage type estimator that shrinks all regression coefficients toward zero (Hastie *et al.*, 2009), which is particularly suitable for modeling high-dimensional microarrays or single nucleotide polymorphism (SNP) data. For instance, Cule *et al.* (2011) applied the ridge estimator on the high-dimensional SNP data and performed significance testing for

CONTACT Takeshi Emura takeshiemura@gmail.com Graduate Institute of Statistics, National Central University, Taiwan.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.

selecting a subset of SNPs useful for prediction. There have been considerable recent applications of ridge regression to high-dimensional settings, including Whittaker *et al.* (2000), Zhao *et al.* (2011), and Cule and De Lorio (2013). In spite of these previous applications, theoretically solid understanding of significance testing under the ridge estimator is only given by a recent work of Bühlmann (2013).

Generalization of the ridge regression has been considered by many authors. The so-called generalized ridge regression is derived by Hoerl and Kennard (1970). Unlike the ordinary ridge regression that shrinks all regression coefficients uniformly, the generalized ridge regression allows different degrees of shrinkage under multiple shrinkage parameters. Interestingly, this generalization actually simplifies the optimal choice of the multiple shrinkage parameters and allows exact evaluation of mean square error (MSE) under estimated shrinkage parameters (Hoerl and Kennard, 1970; Jimichi, 2008). As detailed in Section 4, the generalized ridge estimator is a Bayes estimator that minimizes the posterior risk. From a frequentist viewpoint, the generalized ridge estimator performs better than the LSE in terms of the MSE under estimated optimal choices of the multiple shrinkage parameters (Jimichi 2008). Loesgen (1990) demonstrated that the multiple shrinkage parameters in the generalized ridge estimator arise naturally by utilizing prior information about regression coefficients. All these statistical properties of the generalized ridge estimator are derived under the traditional $p < n$ setting.

To the best of our knowledge, the generalized ridge regression has not been applied to the case of $p > n$. If it would be directly applied to the $p > n$ setting, the generalized ridge regression would involve a large number of shrinkage parameters, which are considerably difficult to be estimated.

In this paper, we propose a class of generalized ridge estimators that reduces the number of shrinkage parameters under a sparsity assumption. The proposed estimator is naturally interpreted from a Bayesian point of view and has a desired performance in terms of the MSE criterion. In addition, the proposed method provides a tool for significance testing and regressor selection (gene selection), which is useful for high-dimensional data analysis. We conduct simulations to study the performance of the proposed method under both $p < n$ and $p \geq n$ cases. Here, we compare our method with three existing methods (the LSE, the ordinary ridge regression, and the Lasso). Finally, we analyze the lung cancer data involving high-dimensional microarrays.

Section 2 provides the background. Section 3 introduces the proposed method, and Section 4 examines its theoretical properties. Sections 5 and 6 describe simulations and real data analysis, respectively. Section 7 concludes.

2. Background

2.1. Linear regression model

Consider the linear regression model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{bmatrix} = [x_1, \ldots, x_p] = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix};$$

X is a fixed (non-random) design matrix, $\boldsymbol{\beta} \in R^p$ is an unknown vector of regression coefficients and $\boldsymbol{\varepsilon}$ follows $N_n(\mathbf{0},\ \sigma^2 I)$, where $\sigma^2 > 0$ is unknown and I is the $n \times n$ identity matrix. Here, $\mathbf{x}_i^{\mathrm{T}}$ denotes the transpose of the $p \times 1$ vector $\mathbf{x}_i$. We assume that the design matrix is

standardized such that $\sum_{i=1}^{n} x_{ij} = 0$ and $\sum_{i=1}^{n} x_{ij}^2 = c$ for $j = 1, \ldots, p$, where c is a constant, usually n or $n-1$.

Provided $X^{\mathrm{T}}X$ is invertible (non-singular), the least squares estimator (LSE) is

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^{\mathrm{T}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathbf{y}.$$

The normality assumption on $\boldsymbol{\varepsilon}$ is not essential for statistical properties of the LSE and the proposed estimator in Section 3. The essential assumptions are $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $Cov(\boldsymbol{\varepsilon}) = \sigma^2 I$. However, we will use the normality assumption to make connections to some Bayesian interpretation in Section 4. Clearly, the LSE is not a suitable estimator under which $X^{\mathrm{T}}X$ is singular, especially when $p > n$.

2.2. Ridge regression and lasso

It is well-known that the LSE has the minimum mean square error (MSE) among all linear unbiased estimators. However, by allowing biased estimators, there exists an even better estimator which reduces the variance much and cost less bias.

Hoerl and Kennard (1970) defined the ridge regression estimator

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

where $\lambda > 0$ is a shrinkage parameter that gives the degree of shrinking $\hat{\boldsymbol{\beta}}(\lambda)$ toward the zero vector. By introducing bias, the ridge regression reduces the variance part of the MSE. An elegant result of Hoerl and Kennard (1970) is that there always exist some $\lambda > 0$ such that the ridge estimator has strictly smaller MSE than that of the LSE. If the eigenvalues of $X^{\mathrm{T}}X$ is $\lambda_1 \geq \cdots \geq \lambda_p > 0$ (i.e, the case of $p < n$), there exists a value $\lambda > 0$ such that

$$MSE(\hat{\boldsymbol{\beta}}(\lambda)) = E\{||\ \hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}\ ||^2\} < MSE(\hat{\boldsymbol{\beta}}) = \sigma^2 \sum_{j=1}^{p} (1/\lambda_j),$$

where $||\ \mathbf{a}\ ||^2 = \mathbf{a}^{\mathrm{T}}\mathbf{a}$ is the L_2-norm for a vector $\mathbf{a}$. The details of the above results are referred to the existence theorem (Theorem 4.3 of Hoerl and Kennard, 1970).

The ridge estimator is regarded as the minimizer of the L_2-penalized residual sum square:

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \{||\ \mathbf{y} - X\boldsymbol{\beta}||^2 + \lambda||\ \boldsymbol{\beta}\ ||^2\}.$$

If the L_2-norm in the penalty term is replaced by the L_1-norm $||\ \boldsymbol{\beta}\ ||_1 = |\beta_1| + \cdots + |\beta_p|$, the resultant estimator is the Lasso (Tibshirani, 1996)

$$\hat{\boldsymbol{\beta}}^{\mathrm{Lasso}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \{||\ \mathbf{y} - X\boldsymbol{\beta}||^2 + \lambda||\ \boldsymbol{\beta}\ ||_1\}.$$

Some of the Lasso coefficients $\hat{\beta}_j^{\mathrm{Lasso}}$ s are exactly zero, making the Lasso different from the ridge that yields all small but nonzero coefficients. This means that the Lasso induces a variable selection tool that selects regressors with non-zero coefficients.

2.3. Estimation of optimal λ

In practice, the value of λ in the ridge estimator is chosen based on criteria, such as the Allen's PRESS (1974), the generalized cross-validation criterion (GCV) (Golub *et al.*,1979), the effective degree of freedom (Hastie *et al.*, 2009), Mallows C_p (Mallows, 1973), and many others as comprehensively listed in Wong and Chiu (2015) and Kibria and Banik (2016).

We particularly introduce the GCV criterion whose asymptotic efficiency is theoretically justified under the $p \geq n$ setup (the CGV theorem of Golub *et al.*, 1979). Let $\hat{\boldsymbol{\beta}}^{(k)}(\lambda)$ be the ridge estimate of $\boldsymbol{\beta}$ without the k th data point $(y_k, \mathbf{x}_k^{\mathrm{T}})$. If λ is chosen properly, then the k-th component $[X\hat{\boldsymbol{\beta}}^{(k)}(\lambda)]_k$ of $X\hat{\boldsymbol{\beta}}^{(k)}(\lambda)$ predicts y_k well. The GCV is defined to be a weighted average of predicted square errors

$$V(\lambda) = \frac{1}{n}\sum_{k=1}^{n}\left([X\hat{\boldsymbol{\beta}}^{(k)}(\lambda)]_k - y_k\right)^2 w_k(\lambda)$$

where $w_k(\lambda) = \{1 - a_{kk}(\lambda)\}/\{1 - \mathrm{Tr}\, A(\lambda)/n\}$, and $a_{kk}(\lambda)$ is the kth diagonal of $A(\lambda) = X(X^{\mathrm{T}}X + \lambda I)X^{\mathrm{T}}$. Golub *et al.* (1979) give a computationally efficient version

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

The function $V(\lambda)$ is called the GCV function of the ridge. The GCV estimator of λ is defined as

$$\hat{\lambda} = \underset{\lambda \geq 0}{\arg\min}\, V(\lambda).$$

The GCV theorem (Golub *et al.* 1979) guarantees the asymptotic efficiency of the GCV estimator under both $p < n$ and $p \geq n$ setups.

The GCV criterion is not applicable to the Lasso. Alternatively, one can apply the 10-fold cross-validation, which is implemented in R `glmnet` package (Friedman *et al.* 2015).

3. Proposed method

3.1. Proposed idea

In the ridge estimator, the matrix $X^{\mathrm{T}}X$ is replaced by $X^{\mathrm{T}}X + \lambda I$ with a shrinkage parameter $\lambda > 0$. Alternatively, the generalized ridge estimator considers $X^{\mathrm{T}}X + W$ for a general diagonal matrix $W = diag(w_1, \ldots, w_p)$, where $w_j \geq 0$, $j = 1, \ldots, p$, are shrinkage parameters. Under the usual $p < n$ setup, the weight matrix W is chosen so that the estimator optimizes some criteria, such as the MSE (Hoerl and Kennard, 1970) and PRESS (Allen, 1974). However, in the high-dimensional case of $p > n$, such optimization schemes yield over-fitting. This motivates us to propose a restricted class of W that reduces the number of shrinkage parameters.

We consider a special class $W = diag(w_1, \ldots, w_p)$, where

$$w_j = \begin{cases} \lambda\gamma & \text{if } \beta_j \neq 0, \\ \lambda & \text{if } \beta_j = 0, \end{cases} \tag{2}$$

for $j = 1, \ldots, p$ and for some $\gamma \in [0, 1]$. When $\beta_j \neq 0$, it is reasonable to choose the smaller weight w_j since it results in the greater value of $|\hat{\beta}_j(W)|$. The parameter $\lambda > 0$ represents the global amount of shrinkage, and the parameter $\gamma \in [0, 1]$ represents the ratio of shrinkage between zero and non-zero coefficients. The weight in Equation (2) has an adequate Bayesian interpretation and is justified from the MSE calculations as discussed in Section 4. If $\gamma = 1$, then Equation (2) results in the ordinary ridge estimator. If $\gamma = 0$, part of regressors do not have any shrinkage, leading to a worse performance in terms of the MSE under high-dimensionality. Hence, we suggest choosing an intermediate value $\gamma = 1/2$ that can also be

suggested by theoretical analysis of Section 4.2. Note that the weight matrix W is unknown since we do not know which components of $\boldsymbol{\beta}$ are nonzero. We even do not know how many components are nonzero. In practice, W must be estimated from data.

3.2. Proposed estimator and computation

We estimate W using the initial estimate $\hat{\boldsymbol{\beta}}^0 = (\hat{\beta}_1^0, \ldots, \hat{\beta}_p^0)^{\mathrm{T}}$, defined as

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

where x_j, for $j = 1, \ldots, p$, are the columns of X. Note that $\hat{\boldsymbol{\beta}}^0$ is a compound of the univariate LSEs, sometimes called "the compound covariate estimator" (Chen and Emura, 2016; Emura et al., 2012). If $|\hat{\beta}_j^0|$ is greater than some threshold, then the true value of β_j is more likely to be nonzero. Hence, we propose a special class of generalized ridge estimators

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

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

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

for $j = 1, \ldots, p$, $SD(\hat{\boldsymbol{\beta}}^0) = \sqrt{\sum_{j=1}^p (\hat{\beta}_j^0 - \bar{\beta}^0)^2/(p-1)}$, and $\bar{\beta}^0 = \sum_{j=1}^p \hat{\beta}_j^0/p$. We call Δ "thresholding parameter." Under the sparse model ($\boldsymbol{\beta} \approx \mathbf{0}$), the histogram of $\hat{\beta}_j^0/SD(\hat{\boldsymbol{\beta}}^0)$, $j = 1, \ldots, p$, is well-approximated by $N(0, 1)$. This implies that $|\hat{\beta}_j^0|/SD(\hat{\boldsymbol{\beta}}^0)$ falls in the range $[0, 3]$ with nearly 99.73%. Hence, we suggest a search range $\Delta \in [0, 3]$ which is free from model parameters such as n and p.

3.3. Computation of (λ, Δ) by GCV

The optimal value of (λ, Δ) in the proposed estimator is estimated in a similar fashion as the ordinary ridge estimator. We modify the GCV function in Equation (1) to

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

where $A(\lambda, \Delta) = X\{X^{\mathrm{T}}X + \lambda\hat{W}(\Delta)\}^{-1}X^{\mathrm{T}}$. Then the estimators $(\hat{\lambda}, \hat{\Delta})$ for the proposed method is defined as the global minimizer of $V(\lambda, \Delta)$,

$$(\hat{\lambda}, \hat{\Delta}) = \underset{\lambda \geq 0,\ \Delta \geq 0}{\arg\min}\, V(\lambda, \Delta).$$

Given Δ, the CGV function is continuous in λ, and hence it is easily minimized using any optimization scheme, such as R `optim` routine, to get $\hat{\lambda}(\Delta)$. Since $V(\hat{\lambda}(\Delta), \Delta)$ is discontinuous in Δ, we propose a grid search. It suffices to search on the grid $D = \{0, 3/100, \ldots, 300/100\}$, thought some efficient algorithms might also be applicable (e.g., Araki and Hattori, 2013). Hence, the "feasible" version of the proposed estimator is

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

The estimator can be interpreted as the empirical Bayes estimator in which hyper parameters W are estimated by $\hat{\lambda}\hat{W}(\hat{\Delta})$ (Section 4.3).

3.4. Significance testing

One can test the significance of each regressor using the proposed method. Consider a null hypothesis

$$H_{0j}: \beta_j = 0 \text{ vs. } H_{1j}: \beta_j \neq 0,$$

for $j = 1, \ldots, p$. Let $\hat{\beta}_j(\hat{\lambda}, \hat{\Delta})$ be jth component of $\hat{\boldsymbol{\beta}}(\hat{\lambda}, \hat{\Delta})$. Define the Wald statistics

$$Z_j = \hat{\beta}_j(\hat{\lambda}, \hat{\Delta})/se\{\hat{\beta}_j(\hat{\lambda}, \hat{\Delta})\},$$

where $se\{\hat{\beta}_j(\hat{\lambda}, \hat{\Delta})\}$ is the standard error. Similar to Cule *et al.* (2011), we define $se\{\hat{\beta}_j(\hat{\lambda}, \hat{\Delta})\}$ by the square root of the jth diagonal of the estimated covariance matrix,

$$Cov\{\hat{\boldsymbol{\beta}}(\hat{\lambda}, \hat{\Delta})\} = \hat{\sigma}^2\{X^{\mathrm{T}}X + \hat{\lambda}\hat{W}(\hat{\Delta})\}^{-1}X^{\mathrm{T}}X\{X^{\mathrm{T}}X + \hat{\lambda}\hat{W}(\hat{\Delta})\}^{-1},$$

where

$$\hat{\sigma}^2 \equiv \{\mathbf{y} - X\hat{\boldsymbol{\beta}}(\hat{\lambda}, \hat{\Delta})\}^{\mathrm{T}}\{\mathbf{y} - X\hat{\boldsymbol{\beta}}(\hat{\lambda}, \hat{\Delta})\}/\nu,$$
$$\nu \equiv \mathrm{Tr}\,\{I - A(\hat{\lambda}, \hat{\Delta})\}^2 = n - \mathrm{Tr}\,\{2A(\hat{\lambda}, \hat{\Delta}) - A(\hat{\lambda}, \hat{\Delta})^2\},$$

where $A(\hat{\lambda}, \hat{\Delta}) = X\{X^{\mathrm{T}}X + \hat{\lambda}\hat{W}(\hat{\Delta})\}^{-1}X^{\mathrm{T}}$. Note that ν is the effective residual degree of freedom, which reduces to $n - p$ if $p < n$ and $\hat{\lambda} = 0$. The P-value calculated from the usual Wald test is useful for regressor selection. For instance, one can choose a subset of the regressors whose P-values are less than some threshold.

4. Theoretical properties

This section gives some theoretical properties that support the proposed estimator under the sparse model ($\boldsymbol{\beta} \approx \mathbf{0}$). Such properties give us systematic reasons why the proposed method can outperform the existing methods. If readers are only interested in applying the statistical methods to read data, it is possible to skip this section.

4.1. Bayesian interpretation

We give a Bayesian interpretation for the proposed class in Equation (2). While a few different types of noninformative prior are commonly used, including the constant prior, the Jeffreys prior and reference prior (see Fan, 2001), we focus on the zero-mean multivariate normal prior. The zero-mean assumption in the prior implies the sparsity in the model, i.e., majority of the regression coefficients are nearly equal to zero.

We first review a Bayesian derivation and interpretation of the generalized ridge estimator along the line with Loesgen (1990). Consider the prior $\boldsymbol{\beta} \sim N_p(\mathbf{0}, \sigma^2 W^{-1})$, where W^{-1} is a $p \times p$ covariance matrix (hyperparameters). Also, $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I)$, as in Section 2. Assume that σ^2 is known. After some calculations, the posterior density of $\boldsymbol{\beta}$ becomes

$$f(\boldsymbol{\beta}|\mathbf{y}, X, W) \propto \exp\left[-\frac{1}{2\sigma^2}\{\boldsymbol{\beta} - (X^{\mathrm{T}}X + W)^{-1}X^{\mathrm{T}}\mathbf{y}\}^{\mathrm{T}}(X^{\mathrm{T}}X + W)\{\boldsymbol{\beta} - (X^{\mathrm{T}}X + W)^{-1}X^{\mathrm{T}}\mathbf{y}\}\right].$$

We see that the posterior mean of $\boldsymbol{\beta}$ is exactly the generalized ridge estimator,

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

Although the paper of Loesgen (1990) did not consider the setting of $p > n$ or the sparse model, the same line of thought can be applied here to choose a good matrix W. Note that, if $\beta_1, \ldots,\ \beta_p$ are independent in the prior, we have $W = diag(w_1, \ldots,\ w_p)$, and $w_1, \ldots,\ w_p \geq 0$. It follows that $E[\beta_j] = 0$ and $Var[\beta_j] = \sigma^2 w_j^{-1}$ which expresses the uncertainty of prior belief that β_j is exactly zero. It means that w_j^{-1} should be small if we believe strongly that $\beta_j = 0$, or w_j^{-1} should be large if β_j is considered far from zero. This gives a rule that, if the truth is $\beta_j = 0$, then w_j should be large; if the truth is $\beta_j \neq 0$, then w_j should be small. Hence, we proposed the weight in Equation (2), where γ is the variance ratio.

Throughout these Bayesian arguments, we attempt to reduce the p-dimensional hyper parameters $(w_1, \ldots,\ w_p)$ to the two-dimensional hyperparameters $(\lambda,\ \gamma)$. In Section 3.3, we used the data to estimate the hyperparameters $W = diag(w_1, \ldots,\ w_p)$ by $\hat{\lambda}\hat{W}(\hat{\Delta})$. This argument follows the empirical Bayes approach.

4.2. MSE comparison

We show that, with an appropriate choice of tuning parameters $(\lambda,\ \gamma)$ in Equation (2), the proposed class improves upon the ordinary ridge estimator and the LSE simultaneously. Our mathematical arguments follow the MSE matrix comparison originated from Theobald (1974). His approach has been particularly useful for comparison of ridge-type estimators; see the overview of the general theory for the MSE matrix comparison given by Trenkler and Toutenburg (1990), and some practical assessment of the MSE matrix in Jang and Anderson-Cook (2015). Note that the theory developed here is applicable even when $p > n$.

Let $\hat{\boldsymbol{\beta}}$ be any estimator of $\boldsymbol{\beta}$. The MSE matrix is defined as a $p \times p$ matrix

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

where $C = Cov(\hat{\boldsymbol{\beta}})$ is the covariance matrix of $\hat{\boldsymbol{\beta}}$ and $\mathbf{d} = Bias(\hat{\boldsymbol{\beta}}) = E(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}$ is the bias of $\hat{\boldsymbol{\beta}}$. The MSE of $\hat{\boldsymbol{\beta}}$ is the trace of $\mathbf{M}(\hat{\boldsymbol{\beta}})$. Note that the diagonals of a nonnegative definite (n.n.d.) matrix are nonnegative. This implies that, if $\mathbf{M}(\hat{\boldsymbol{\beta}}_1) - \mathbf{M}(\hat{\boldsymbol{\beta}}_2)$ is n.n.d., then $MSE(\hat{\boldsymbol{\beta}}_1) \geq MSE(\hat{\boldsymbol{\beta}}_2)$ for two estimators $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$.

In our analysis of the generalized ridge estimator, we compare the performance between the ordinary ridge estimator $\hat{\boldsymbol{\beta}}(\lambda)$ with $\lambda = 1$ and the proposed class

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

where $W(\lambda,\ \gamma)$ is defined in Equation (2). We also compare the performance between the LSE and the proposed class. Accordingly, we seek conditions under which both

$$\mathbf{M}(\hat{\boldsymbol{\beta}}(1)) - \mathbf{M}(\hat{\boldsymbol{\beta}}(\lambda,\ \gamma)),\ and\ \mathbf{M}(\hat{\boldsymbol{\beta}}(0)) - \mathbf{M}(\hat{\boldsymbol{\beta}}(\lambda,\ \gamma))$$

are n.n.d. Trenkler (1985) established a useful lemma as follows:

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

i) A is n.n.d.,
ii) $\mathbf{a} = A\mathbf{v}$ for some $\mathbf{v} \in R^p$

and
iii) $\mathbf{a}^{\mathrm{T}}A^{-}\mathbf{a} \leq \eta$.
where A^{-} is the generalized inverse of A.

Lemma 1 with $\eta = 1$ is discussed in Trenkler and Toutenburg (1990). Similar to this paper, we give simple sufficient conditions that $\mathbf{M}(\hat{\boldsymbol{\beta}}_1) - \mathbf{M}(\hat{\boldsymbol{\beta}}_2) = (C_1 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}}) - (C_2 + \mathbf{d}_2\mathbf{d}_2^{\mathrm{T}})$ is n.n.d., where C_i is the covariance and $\mathbf{d}_i$ is the bias for the ridge-type estimator $\hat{\boldsymbol{\beta}}_i = (X^{\mathrm{T}}X + W_i)^{-1}X^{\mathrm{T}}\mathbf{y}$, $i = 1,\ 2$. As a version of Lemma 1, it is convenient to establish the following theorem.

Theorem 1: $\mathbf{M}(\hat{\boldsymbol{\beta}}_1) - \mathbf{M}(\hat{\boldsymbol{\beta}}_2)$ is n.n.d. if all the three conditions hold:
i) $(C_1 - C_2)/\sigma^2$ is n.n.d.,
ii) $\mathbf{d}_2 = (C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}})\mathbf{v}$ *for some* $\mathbf{v} \in R^p$,
iii) $\mathbf{d}_2^{\mathrm{T}}(C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}})^{-1}\mathbf{d}_2 \leq 1$.

Proof: We apply Lemma 1 with $\eta = 1$, $A = C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}}$ and $\mathbf{a} = \mathbf{d}_2$. If $(C_1 - C_2)/\sigma^2$ is n.n.d., then $\mathbf{x}^{\mathrm{T}}(C_1 - C_2)\mathbf{x} \geq 0$ for any $\mathbf{x} \neq \mathbf{0}$. Then, for $\mathbf{x} \neq \mathbf{0}$,

$$\mathbf{x}^{\mathrm{T}}(C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}})\mathbf{x} = \mathbf{x}^{\mathrm{T}}(C_1 - C_2)\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{d}_1\mathbf{d}_1^{\mathrm{T}}\mathbf{x} \geq (\mathbf{d}_1^{\mathrm{T}}\mathbf{x})^2 \geq 0,$$

i.e., $C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}}$ is also n.n.d.. Hence, Condition i) of Lemma 1 is satisfied. In addition, Conditions ii) and iii) in Lemma 1 are satisfied with $\eta = 1$. By Lemma 1, we have verified

$$(C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}}) - \mathbf{d}_2\mathbf{d}_2^{\mathrm{T}} = \mathbf{M}(\hat{\boldsymbol{\beta}}_1) - \mathbf{M}(\hat{\boldsymbol{\beta}}_2)$$

is n.n.d. □

Roughly speaking, only Condition iii) of Theorem 1 is essential for the MSE improvement of $\hat{\boldsymbol{\beta}}_2$ over $\hat{\boldsymbol{\beta}}_1$. If the quantity in Condition iii) is strictly less than one, the MSE of $\hat{\boldsymbol{\beta}}_2$ can be less than the MSE of $\hat{\boldsymbol{\beta}}_1$.

Example

Here is a simple example for illustrating Theorem 1. Let $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_1,\ \mathbf{0}^{\mathrm{T}}) \in R^p$, where $\beta_1 \neq 0$. This is a simplified case of more general sparse models that will be discussed in Section 5. Let $X^{\mathrm{T}}X = (1-\rho)I + \rho\mathbf{1}\mathbf{1}^{\mathrm{T}}$, where ρ is the correlation between columns of X and $\mathbf{1} = (1, \ldots,\ 1)^{\mathrm{T}}$. We start with the orthonormal case ρ=0, namely $X^{\mathrm{T}}X = I$, as in p.152 of Loesgen (1990). This means $p < n$.

First, we compare between the ordinary ridge and the proposed class by letting $W_1 = I = diag(1,\ \ldots\ ,\ 1)$ and $W_2 = diag(\lambda\gamma,\ \lambda,\ \ldots,\ \lambda)$ for $0 < \gamma < 1$ and $\lambda > 0$. Then,

$$\mathbf{d}_1 = (X^{\mathrm{T}}X + W_1)^{-1}X^{\mathrm{T}}X\boldsymbol{\beta} - \boldsymbol{\beta} = \frac{-\beta_1}{2}\begin{pmatrix}1\\ \mathbf{0}\end{pmatrix},$$

$$\mathbf{d}_2 = (X^{\mathrm{T}}X + W_2)^{-1}X^{\mathrm{T}}X\boldsymbol{\beta} - \boldsymbol{\beta} = \frac{-\lambda\gamma\beta_1}{1+\lambda\gamma}\begin{pmatrix}1\\ \mathbf{0}\end{pmatrix},$$

$$\frac{1}{\sigma^2}(C_1 - C_2) = diag\left(\frac{1}{4} - \frac{1}{(1+\lambda\gamma)^2},\ \frac{1}{4} - \frac{1}{(1+\lambda)^2},\ \cdots,\ \frac{1}{4} - \frac{1}{(1+\lambda)^2}\right),$$

$$C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}} = diag\left(\frac{\sigma^2\{(1+\lambda\gamma)^2-4\}+(1+\lambda\gamma)^2\beta_1^2}{4(1+\lambda\gamma)^2},\ \frac{\sigma^2(\lambda+3)(\lambda-1)}{4(1+\lambda)^2},\ \cdots,\ \frac{\sigma^2(\lambda+3)(\lambda-1)}{4(1+\lambda)^2}\right).$$

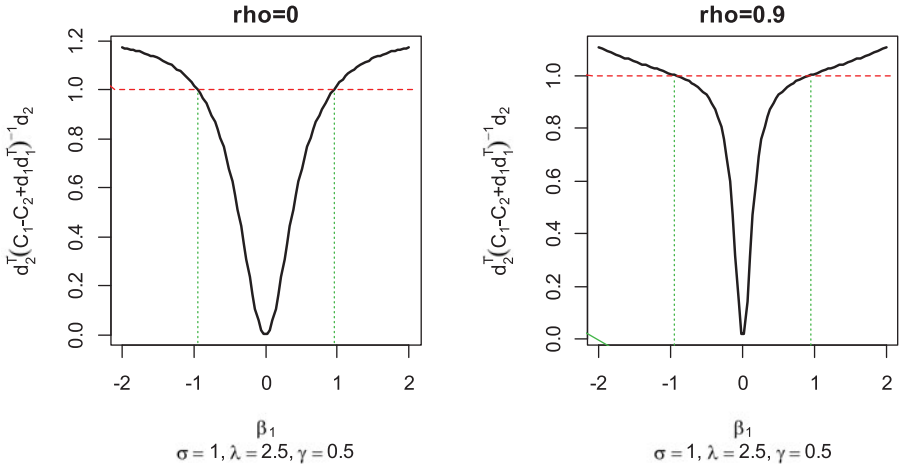

Figure 1. The plots of $\mathbf{d}_2^{\mathrm{T}}(C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}})^{-1}\mathbf{d}_2$ against β_1 for comparing between the ordinary ridge and the proposed class. Under $\sigma^2 = 1$, $\lambda = 2.5$, and $\gamma = 0.5$, the left panel shows the orthonormal case $\rho = 0$; the right panel shows the non-orthonormal case $\rho = 0.9$. The figures show that $\mathbf{d}_2^{\mathrm{T}}(C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}})^{-1}\mathbf{d}_2 \leq 1$ if β_1 is near 0. For the orthonormal case, Equation (3) allows the expression.

$$\mathbf{d}_2^{\mathrm{T}}(C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}})^{-1}\mathbf{d}_2 = \frac{6.25\beta_1^2}{\sigma^2\{\,(1+1.25)^2 - 4\,\} + (1+1.25)^2\beta_1^2} = \frac{6.25\beta_1^2}{1.0625\sigma^2 + 5.0625\beta_1^2}.$$

Hence, the range of β_1 is $[-0.8947,\ 0.8947]$. In this range, Theorem 1 verifies the relation $MSE(\hat{\boldsymbol{\beta}}_1) \geq MSE(\hat{\boldsymbol{\beta}}_2)$ [see also Yang (2014) who numerically verified the relation $MSE(\hat{\boldsymbol{\beta}}_1) \geq MSE(\hat{\boldsymbol{\beta}}_2)$].

Some simplification is possible as

$$\mathbf{d}_2^{\mathrm{T}}(C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}})^{-1}\mathbf{d}_2 = \frac{4\lambda^2\gamma^2\beta_1^2}{\sigma^2\{\,(1+\lambda\gamma)^2 - 4\,\} + (1+\lambda\gamma)^2\beta_1^2}. \tag{3}$$

We know that a diagonal matrix is n.n.d. if and only if its diagonals are all nonnegative. It means that $(C_1 - C_2)/\sigma^2$ is n.n.d. if and only if:

1) $(1+\lambda\gamma)^2 \geq 4$ and **2)** $(1+\lambda)^2 \geq 4$.

Equivalently, $\lambda \geq 1$ and $\gamma \geq 1/\lambda$. For instance, if $\lambda = 2.5$, then $\gamma \geq 2/5$. If we let

$$v_1 = \frac{-4\lambda\gamma\beta_1(1+\lambda\gamma)}{\sigma^2\{(1+\lambda\gamma)^2 - 4\} + (1+\lambda\gamma)^2\beta_1^2},$$

then $\mathbf{d}_2 = (C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}})\mathbf{v}$ for $\mathbf{v}^{\mathrm{T}} = (v_1,\ \mathbf{0}^{\mathrm{T}})$. That is, Conditions (i) and (ii) of Theorem 1 are satisfied when $\lambda = 2.5$ and $\gamma \geq 2/5$. From Equation (3), $\mathbf{d}_2^{\mathrm{T}}(C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}})^{-1}\mathbf{d}_2$ goes to zero as $\beta_1 \to 0$ (see also Fig. 1). Thus, if β_1 is near 0, Condition iii) of Theorem 1 is verified. Therefore, $\mathbf{M}(\hat{\boldsymbol{\beta}}_1) - \mathbf{M}(\hat{\boldsymbol{\beta}}_2)$ is n.n.d., implying $MSE(\hat{\boldsymbol{\beta}}_1) \geq MSE(\hat{\boldsymbol{\beta}}_2)$. While the choice $\gamma = 2/5$ is allowed, this does not strictly improve the MSE as Condition iii) yields the equality. We suggest a slightly larger value, say $\gamma = 1/2 > 2/5$ so that $\mathbf{d}_2^{\mathrm{T}}(C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}})^{-1}\mathbf{d}_2$ can be strictly less than one.

Second, we compare between the LSE and the proposed class by letting $W_1 = diag(0, \ldots,\ 0)$ and $W_2 = diag(\lambda\gamma,\ \lambda, \ldots,\ \lambda)$ for $0 < \gamma < 1$ and $\lambda > 0$. Then,

$$\mathbf{d}_1 = \mathbf{0},\ \mathbf{d}_2 = (X^{\mathrm{T}}X + W_2)^{-1}X^{\mathrm{T}}X\boldsymbol{\beta} - \boldsymbol{\beta} = \frac{-\lambda\gamma\beta_1}{1+\lambda\gamma}\begin{pmatrix}1\\ \mathbf{0}\end{pmatrix},$$

$$\frac{1}{\sigma^2}(C_1 - C_2) = diag\left(1 - \frac{1}{(1+\lambda\gamma)^2},\ 1 - \frac{1}{(1+\lambda)^2},\ \cdots,\ 1 - \frac{1}{(1+\lambda)^2}\right),$$

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

Similar to the case for the ridge, Conditions i)-iii) hold when β_1 is near 0 under $\lambda = 2.5$ and $\gamma = 1/2 > 2/5$.

Therefore, we have theoretically verified our choice $\gamma = 1/2$ in Equation (2) such that the proposed estimator simultaneously improves upon both the ordinary ridge and the LSE when β_1 is near 0. This conclusion would continue to hold even for non-orthonormal cases of $\rho \neq 0$, where the tractable formula of $\mathbf{d}_2^{\mathrm{T}}(C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}})^{-1}\mathbf{d}_2$ is no longer available. We compute $\mathbf{d}_2^{\mathrm{T}}(C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}})^{-1}\mathbf{d}_2$ numerically with $\rho = 0.9$, and verify $\mathbf{d}_2^{\mathrm{T}}(C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}})^{-1}\mathbf{d}_2 \leq 1$ if β_1 is near 0 as shown in Fig. 1.

Remark I: An important implication from the above simple example is that the global shrinkage parameter ($\lambda = 2.5$) in the proposed method must be larger than that of the ridge ($\lambda = 1$). However, the amount of shrinkage corresponding to the nonzero coefficients ($\lambda\gamma = 2.5 \times 2/5 = 1$) remains the same as the ridge. Hence, the proposed method improves upon the ridge by imposing higher shrinking rate for the zero coefficients.

Although the above example considers the simple case of $p < n$, Conditions of Theorem 1 can be satisfied for more general X and $\boldsymbol{\beta}$, including the case of p > n. Under high-dimensionality, however, Conditions (i)–(iii) are difficult to be verified without relying on computer programs.

We propose a way to verify Conditions (i)–(iii) in Theorem 1, under the case of $p > n$. Note that if $rank(C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}}) = rank([C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}}|\mathbf{d}_2])$, then $\mathbf{d}_2$ belongs to the column space of $C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}}$. This implies $\mathbf{d}_2 = (C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}})\mathbf{v}$ for some $\mathbf{v} \in R^p$. If so, Condition (ii) in Theorem 1 is satisfied. One can use `qr()$rank` in R to obtain the rank of a matrix. For instance, one can check Condition (ii) under $p = 100$, $n = 50$, $\lambda = 2.5$, $\gamma = 1/2$,$\sigma = 1$, $W_1 = I_{100}$,

$$W_2 = diag(\overbrace{\lambda\gamma, \ldots, \lambda\gamma}^{20}, \overbrace{\lambda, \ldots, \lambda}^{80}), \quad \boldsymbol{\beta} = (\overbrace{0.5, \ldots, 0.5}^{20}, \overbrace{0, \ldots, 0}^{80})^{\mathrm{T}}$$

First, we generate the design matrix as in the simulations (Section 5.1). Then we obtain

$$C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}} = \begin{bmatrix} -0.0046 & 0.0059 & \cdots & 0.0026 \\ 0.0059 & 0.0070 & \cdots & 0.0086 \\ \vdots & \vdots & \ddots & \vdots \\ 0.0026 & 0.0086 & \cdots & 0.0065 \end{bmatrix},$$

$$[C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}}|\mathbf{d}_2] = \left[\begin{array}{cccc|c} -0.0046 & 0.0059 & \cdots & 0.0026 & -0.0148 \\ 0.0059 & 0.0070 & \cdots & 0.0086 & -0.0912 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0.0026 & 0.0086 & \cdots & 0.0065 & -0.0401 \end{array}\right].$$

Condition (ii) is verified as $rank(C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}}) = rank([C_1 - C_2 + \mathbf{d}_1\mathbf{d}_1^{\mathrm{T}}|\mathbf{d}_2]) = 70$. Conditions (i) and (iii) can be checked more easily.

Remark II: The developed framework of evaluating the MSE matrix is applicable for estimators having the expressions of both variance and bias. This is mainly the case of linear estimators $\hat{\boldsymbol{\beta}} = L\mathbf{y}$, where L is a deterministic matrix, including the LSE, ordinary ridge, and the proposed class. It is typically not possible to evaluate nonlinear estimators, in particular

the Lasso estimator. The Lasso estimator is often evaluated by the upper bound of the MSE (Hansen, 2016).

5. Simulations

We conducted Monte Carlo simulations to study the performance for the proposed method.

5.1. Model design

We consider a sparse high-dimensional model, where the true $\boldsymbol{\beta}$ has $(q+r)$ nonzero terms and $p-(q+r)$ zero terms such that

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

for $b,\ d \in \mathrm{R}$. We set $p \in \{50,\ 100,\ 150,\ 200\}$ and $q = r = 10$. We consider four cases: (**I**) $b = d = 5$ (**II**) $b = d = 10$ (**III**) $b = 5$ and $d = -5$ (**IV**) $b = 10$ and $d = -10$. This type of sparse high-dimensional models is adopted in many papers such as Emura *et al.* (2012) and Bühlmann (2013). The sample size is fixed at $n = 100$ throughout the simulations.

The marginal distributions of p regressors follow $N(0,\ 1)$. We introduce correlation among columns of the design matrix $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$ by letting

$$\mathbf{x}_i^{\mathrm{T}} = \left(\frac{z_{i1} + u_i}{\sqrt{2}}, \ldots, \frac{z_{iq} + u_i}{\sqrt{2}}, \frac{z_{i(q+1)} + v_i}{\sqrt{2}}, \ldots, \frac{z_{i(q+r)} + v_i}{\sqrt{2}}, z_{i(q+r+1)}, \ldots, z_{ip} \right),$$

where $z_{i1}, \ldots, z_{ip}$ and $u_i,\ v_i$ all independently follow $N(0,\ 1)$ for $i = 1, \ldots, n$. This yields the correlation

$$Corr(x_{ij},\ x_{ij'}) = \begin{cases} 0.5 & \text{if } j,\ j' \in \{1, \ldots, q\}, \\ 0.5 & \text{if } j,\ j' \in \{q+1, \ldots, q+r\}, \\ 0 & \text{otherwise.} \end{cases}$$

The design matrix that has the same correlation structure is generated by `X.pathway` routine in R `compound.Cox` package (Emura *et al.*, 2017a).

After generating the designed matrix, we set $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0},\ I)$. Based on 500 replications (on $\boldsymbol{\varepsilon}$), the performances of the proposed estimator will be examined by the MSE criterion (Section 5.2 and 5.3). We will also study the performance of the proposed significance test (Section 5.4).

5.2. MSE comparison for fixed λ and Δ

We compare the performances of the proposed method and ridge regression in terms of the MSE curve, which is the plot of the MSE

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

against λ. Here, $\Delta^* = E(\hat{\Delta})$ is given prior to the simulations. The MSE curve is often called "infeasible MSE" since the point estimates are not obtained unless the values λ are determined. Nevertheless, the curve gives us some insight about potential gain of the MSE with varying

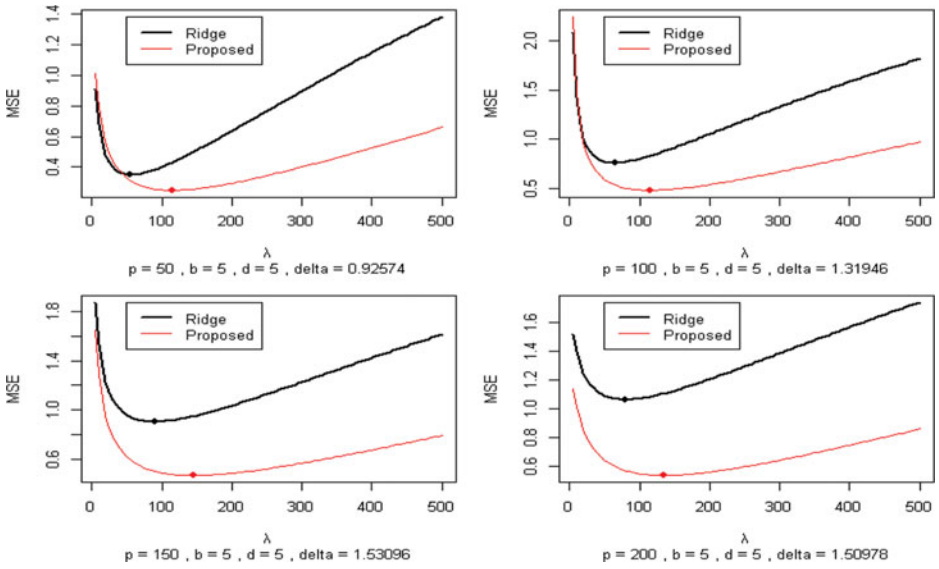

Figure 2. The MSE curves of the ridge and proposed estimators against λ with $b = d = 5$. The points denote the minimum of each curve.

values of λ. The MSE of the LSE is calculated as $MSE(\hat{\boldsymbol{\beta}}) = MSE(\hat{\boldsymbol{\beta}}(0))$ only for the case of p= 50.

Figures 2–5 depict the two MSE curves, $MSE(\hat{\boldsymbol{\beta}}(\lambda))$ and $MSE(\hat{\boldsymbol{\beta}}(\lambda,\ \Delta^*))$. We see that there always exists some $\lambda > 0$ such that the ordinary ridge estimator has strictly smaller MSE than that of the LSE, i.e., $MSE(\hat{\boldsymbol{\beta}}(\lambda)) < MSE(\hat{\boldsymbol{\beta}}(0))$. This is the consequence of the existence theorem as mentioned in Section 2.2. The proposed method gives a quite similar pattern of the MSE curve to that of the ordinary ridge. However, the minimum of the MSE curves for the proposed method is smaller than that for the ordinary ridge in all cases. Hence, if estimates $(\hat{\lambda},\ \hat{\Delta})$ are chosen properly, the MSE of the proposed estimator $\hat{\boldsymbol{\beta}}(\hat{\lambda},\ \hat{\Delta})$ can be less than that of the ordinary ridge estimator $\hat{\boldsymbol{\beta}}(\hat{\lambda})$. In addition, the superiority of

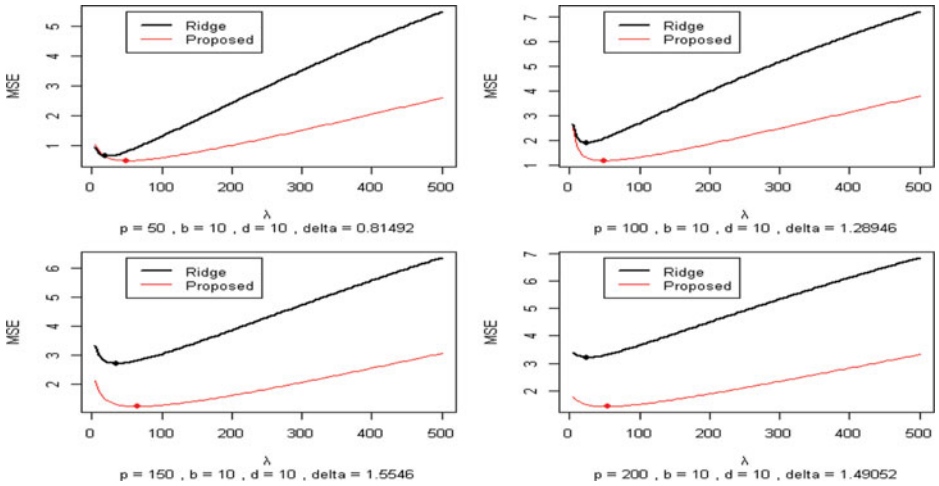

Figure 3. The MSE curves of the ridge and proposed estimators against λ with $b = d = 10$. The points denote the minimum of each curve.

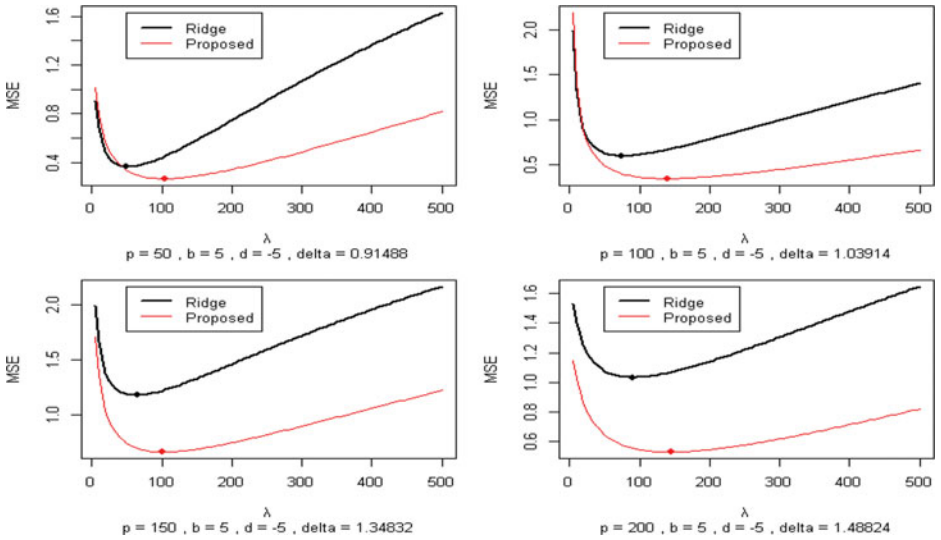

Figure 4. The MSE curves of the ridge and proposed estimators against λ with $b = 5$ and $d = -5$. The points denote the minimum of each curve.

the proposed method over the ordinary ridge tends to be greater when p is larger ($p = 150$ and 200). Hence, one can expect the greater benefit of the proposed estimator when p is larger.

An important finding from Figures 2–5 is that the optimal λ for the proposed method is always larger than that for the ordinary ridge. This result agrees with our theoretical analysis of Section 4.2; the proposed method improves the MSE by choosing larger shrinkage parameter than the ridge does (i.e., $\lambda \geq 1$). Hence, if λ is properly estimated by data, the proposed method should give stronger shrinkage toward the zero vector.

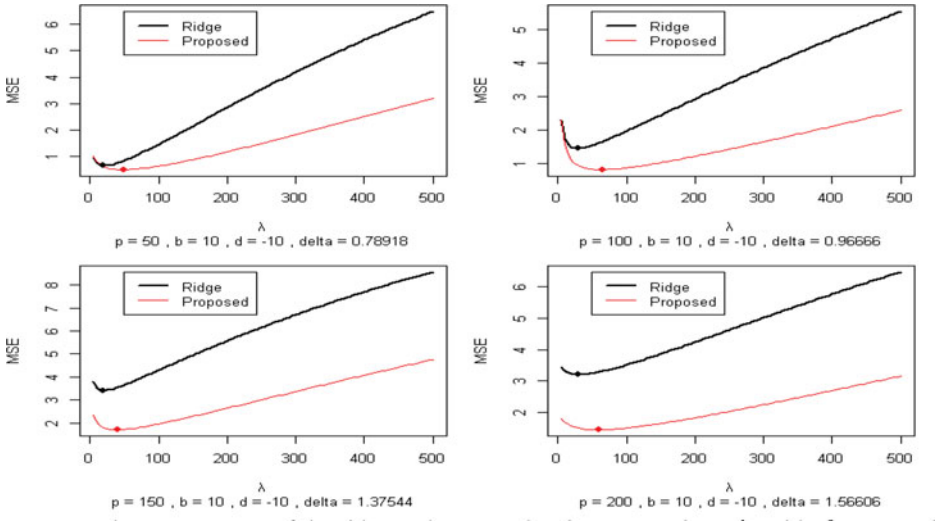

Figure 5. The MSE curves of the ridge and proposed estimators against λ with $b = 10$ and $d = -10$. The points denote the minimum of each curve.

5.3. MSE comparison for estimated λ and Δ

Rather than fixing λ and Δ in Section 5.2, we consider the variability of estimating λ and Δ for the ridge and proposed estimators. The performances are then evaluated by the "feasible" MSE for the ridge and proposed estimator, respectively defined as

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

where $\hat{\lambda} = \arg\min_{\lambda \geq 0} V(\lambda)$ and $(\hat{\lambda},\ \hat{\Delta}) = \arg\min_{\lambda \geq 0,\ \Delta \geq 0} V(\lambda,\ \Delta)$ (see Sections 2.3 and 3.3, respectively. We also evaluate the MSE for the Lasso

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

where $\hat{\lambda}^{\mathrm{Lasso}}$ is first obtained from R `cv.glmnet` routine (10-fold cross validation), and then $\hat{\boldsymbol{\beta}}^{\mathrm{Lasso}}(\hat{\lambda}^{\mathrm{Lasso}})$ is obtained by R `glmnet` routine (Friedman *et al.* 2015).

Let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots,\ \hat{\beta}_p)^{\mathrm{T}}$ be an estimator of $\boldsymbol{\beta} = (\beta_1, \ldots,\ \beta_p)^{\mathrm{T}}$, which is either $\hat{\boldsymbol{\beta}}(\hat{\lambda})$, $\hat{\boldsymbol{\beta}}(\hat{\lambda},\ \hat{\Delta})$, or $\hat{\boldsymbol{\beta}}^{\mathrm{Lasso}}(\hat{\lambda}^{\mathrm{Lasso}})$. We also examine two components of the MSE defined as

$$MSE(\hat{\beta}_j) = E(\hat{\beta}_j - \beta_j)^2, \quad for\ j = 1 \text{ and } p$$

Table 1 compares the MSE among the proposed method, the ordinary ridge estimator, and the Lasso. We see that the performance of proposed estimator is always better than that of the ridge in terms of $MSE(\hat{\boldsymbol{\beta}})$. The advantage of the proposed method is more remarkable when p is larger, as expected from the results in Section 5.2. For instance, when $p = 200$, the proposed method reduces $MSE(\hat{\boldsymbol{\beta}})$ by half. The first column shows that the shrinkage parameter

Table 1. **(a)** Simulation results comparing three estimators (Ridge, Lasso and proposed estimator) based on 500 replicates.

			$E(\hat{\lambda})$	$E(\hat{\Delta})$	$MSE(\hat{\beta}_1)$	$MSE(\hat{\beta}_p)$	$MSE(\hat{\boldsymbol{\beta}})$
$b = d = 5$	$p = 50$	Ridge	23.20	–	0.0112	0.0077	0.4663
		Lasso	20.96	–	0.0274	0.0023	0.5978
		Proposed	47.34	0.92	0.0107	0.0048	0.3763
	$p = 100$	Ridge	23.86	–	0.0080	0.0086	1.0228
		Lasso	58.27	–	0.0244	0.0001	1.0655
		Proposed	46.96	1.31	0.0086	0.0058	0.7191
	$p = 150$	Ridge	20.45	–	0.0118	0.0065	1.2909
		Lasso	28.77	–	0.0350	0.0003	0.6628
		Proposed	43.83	1.53	0.0171	0.0037	0.7822
	$p = 200$	Ridge	10.60	–	0.0036	0.0046	1.4137
		Lasso	29.10	–	0.0335	0.0006	0.6993
		Proposed	30.88	1.50	0.0043	0.0027	0.8364
$b = d = 10$	$p = 50$	Ridge	9.51	–	0.0202	0.0119	0.7755
		Lasso	44.13	–	0.0262	0.0004	0.6301
		Proposed	21.55	0.81	0.0174	0.0083	0.6177
	$p = 100$	Ridge	9.32	–	0.0314	0.0178	2.4243
		Lasso	174.06	–	0.0317	0.0000	5.1991
		Proposed	20.28	1.28	0.0266	0.0116	1.7087
	$p = 150$	Ridge	5.18	–	0.0320	0.0124	3.3783
		Lasso	30.09	–	0.0356	0.0002	0.6453
		Proposed	14.98	1.55	0.0437	0.0086	1.8486
	$p = 200$	Ridge	0.76	–	0.0053	0.0065	3.5301
		Lasso	41.14	–	0.0348	0.0002	0.6711
		Proposed	5.58	1.50	0.0079	0.0046	1.8204

NOTE: We set the sample size $n = 100$.

Table 1. (b) Simulation results comparing three estimators (Ridge, Lasso and proposed estimator) based on 500 replicates.

			$E(\hat{\lambda})$	$E(\hat{\Delta})$	$MSE(\hat{\beta}_1)$	$MSE(\hat{\beta}_p)$	$MSE(\hat{\boldsymbol{\beta}})$
$b = 5,\ d = -5$	$p = 50$	Ridge	21.36	–	0.0121	0.0081	0.4872
		Lasso	19.65	–	0.0275	0.0026	0.6106
		Proposed	43.49	0.91	0.0108	0.0051	0.3981
	$p = 100$	Ridge	26.69	–	0.0075	0.0098	0.8622
		Lasso	103.43	–	0.0248	0.0000	0.8379
		Proposed	52.78	1.03	0.0073	0.0047	0.5693
	$p = 150$	Ridge	17.84	–	0.0053	0.0169	1.5185
		Lasso	51.01	–	0.0268	0.0003	0.7663
		Proposed	38.25	1.34	0.0061	0.0086	0.9057
	$p = 200$	Ridge	11.76	–	0.0173	0.0057	1.3875
		Lasso	29.63	–	0.0333	0.0004	0.6588
		Proposed	32.55	1.52	0.0192	0.0028	0.8059
$b = 10,\ d = -10$	$p = 50$	Ridge	8.79	–	0.0220	0.0123	0.7997
		Lasso	32.04	–	0.0266	0.0010	0.6056
		Proposed	19.76	0.78	0.0191	0.0087	0.6527
	$p = 100$	Ridge	10.71	–	0.0194	0.0241	1.9339
		Lasso	263.10	–	0.0256	0.0000	2.8712
		Proposed	23.13	0.96	0.0180	0.0118	1.3168
	$p = 150$	Ridge	3.61	–	0.0092	0.0472	3.9864
		Lasso	161.97	–	0.0360	0.0000	3.0474
		Proposed	12.07	1.37	0.0111	0.0217	2.0756
	$p = 200$	Ridge	1.32	–	0.0495	0.0131	3.5478
		Lasso	32.65	–	0.0329	0.0002	0.6272
		Proposed	6.83	1.56	0.0479	0.0058	1.7845

NOTE: We set the sample size $n = 100$.

estimates $\hat{\lambda}$ are larger in the proposed estimator than that in the ordinary ridge estimator. This implies that the proposed estimator reduces the MSE by shrinking more toward zero than the ridge estimator does. This finding is consistent with our theoretical calculations of the MSE in Section 4.2.

It is interesting to point out that the proposed method sometimes produces larger $MSE(\hat{\beta}_1)$ than the ridge does (Table 1) for $\beta_1 \neq 0$. This implies that the proposed method does not necessarily produce better estimates for non-zero regression coefficients. On the other hand, the proposed method always produces smaller $MSE(\hat{\beta}_p)$ than the ridge does for $\beta_p = 0$. Since the majority of regression coefficients are zero, the proposed methods has overall better performance in terms of $MSE(\hat{\boldsymbol{\beta}})$.

The MSE of the Lasso produces quite different pattern from the two ridge estimators (Table 1). In general, the Lasso gives the smallest $MSE(\hat{\boldsymbol{\beta}})$ for the case of $p = 200$. On the other hand, for the cases of $p = 50$ and $p = 100$, the proposed estimator performs better than the Lasso. Some unstability in the performance of the Lasso is found, especially when $p = 100$.

5.4. Performance of significance testing

We assess the performance of the proposed significance testing procedure defined in Section 3.4. We set the problem of testing hypotheses

$$H_0 : \beta_1 = 0 \quad v.s. \quad H_1 : \beta_1 \neq 0,$$
$$H_0 : \beta_{50} = 0 \quad v.s. \quad H_1 : \beta_{50} \neq 0.$$

Table 2. Simulation results for testing $H_0 : \beta_{50} = 0$ using the proposed estimator (the LSE in parenthesis) based on 500 replicates.

		$E(\hat{\beta}_{50})$	$sd(\hat{\beta}_{50})$	$E(Z_{50})$	$sd(Z_{50})$	Type I error
$\beta_{50} = 0, b = d = 5$	$p = 50$	−0.007(0.001)	0.069(0.141)	−0.042(0.000)	0.968(1.025)	0.042(0.054)
	$p = 100$	−0.005	0.066	−0.113	0.941	0.028
	$p = 150$	−0.022	0.054	−0.388	0.886	0.046
	$p = 200$	0.033	0.051	0.554	0.841	0.048
$\beta_{50} = 0, b = d = 10$	$p = 50$	−0.001(0.001)	0.091(0.142)	−0.011(0.000)	0.972(1.025)	0.048(0.054)
	$p = 100$	−0.003	0.096	−0.049	0.971	0.036
	$p = 150$	−0.034	0.075	−0.393	0.811	0.038
	$p = 200$	0.049	0.068	0.539	0.736	0.018
$\beta_{50} = 0, b = 5,\ d = -5$	$p = 50$	−0.005(0.001)	0.071(0.142)	−0.071(0.000)	0.969(1.025)	0.044(0.054)
	$p = 100$	−0.009	0.062	−0.143	0.941	0.028
	$p = 150$	−0.006	0.056	−0.115	0.853	0.018
	$p = 200$	−0.029	0.051	−0.504	0.857	0.054
$\beta_{50} = 0, b = 10,\ d = -10$	$p = 50$	−0.003(0.001)	0.093(0.142)	−0.036(0.000)	0.972(1.025)	0.044(0.054)
	$p = 100$	−0.008	0.088	−0.105	0.943	0.028
	$p = 150$	−0.008	0.078	−0.090	0.759	0.016
	$p = 200$	−0.056	0.066	−0.628	0.733	0.040

NOTE: We set the sample size $n = 100$; thus the LSE is applicable only for $p = 50$.

Since $\beta_1 \neq 0$ and $\beta_{50} = 0$ by the simulation setting, $H_0 : \beta_1 = 0$ is false and $H_0 : \beta_{50} = 0$ is true. Based on 500 replicates, we evaluate the rejection rates

$$Rejection\ rate = \frac{1}{500}\sum_{s=1}^{500} \mathbf{I}(\mid Z^{(s)} \mid > Z_{\alpha/2}),$$

where $\mathbf{I}(\cdot)$ is the indicator function, and $Z^{(s)}$ is the Wald statistic at the s-th replication. The rejection rate is "Type I error" under $H_0 : \beta_{50} = 0$ or "power" under $H_0 : \beta_1 = 0$.

Tables 2 and 3 display the simulation results for the proposed testing procedure. The Type I error rates for all cases, except only one case ($p = 200,\ b = 5$ and $d = -5$), are less than the nominal level $\alpha = 0.05$. Hence, the Type I error rates are generally kept below the nominal level. Powers for most cases are exactly equal to or quite close to one. In summary, the proposed test has conservative Type I error rate and high statistical power. This implies the test

Table 3. Simulation results for testing $H_0 : \beta_1 = 0$ using the proposed estimator (the LSE in parenthesis) based on 500 replicates.

		$E(\hat{\beta}_1)$	$sd(\hat{\beta}_1)$	$E(Z_1)$	$sd(Z_1)$	Power
$\beta_1 = 5/10 = 0.5, b = d = 5$	$p = 50$	0.465(0.502)	0.097(0.202)	4.999(2.562)	1.009(1.065)	0.998(0.706)
	$p = 100$	0.514	0.091	5.968	0.998	0.998
	$p = 150$	0.411	0.095	5.637	0.940	1
	$p = 200$	0.499	0.065	6.530	0.928	1
$\beta_1 = 10/10 = 1, b = d = 10$	$p = 50$	0.956(1.002)	0.124(0.202)	7.573(5.117)	1.034(1.143)	1(1)
	$p = 100$	1.067	0.148	8.338	1.137	0.998
	$p = 150$	0.851	0.147	7.847	0.967	1
	$p = 200$	1.020	0.084	8.494	0.919	1
$\beta_1 = 5/10 = 0.5, b = 5,\ d = -5$	$p = 50$	0.461(0.502)	0.096(0.201)	4.766(2.562)	0.990(1.065)	0.996(0.706)
	$p = 100$	0.479	0.083	5.882	1.032	0.996
	$p = 150$	0.484	0.076	5.830	0.926	1
	$p = 200$	0.398	0.085	5.450	0.882	1
$\beta_1 = 10/10 = 1, b = 10,\ d = -10$	$p = 50$	0.947(1.002)	0.128(0.201)	7.308(5.117)	1.020(1.143)	1(1)
	$p = 100$	0.982	0.133	8.126	1.110	0.998
	$p = 150$	1.009	0.105	7.991	0.921	1
	$p = 200$	0.830	0.133	7.339	0.794	1

NOTE: We set the sample size $n = 100$; thus the LSE is applicable only for $p = 50$.

has a good ability to select regressors with nonzero coefficients with a small rate to select null regressors.

Tables 2 and 3 also compare the proposed test with the test based on the LSE. Since we have set $n = 100$ and $p \in \{50,\ 100,\ 150,\ 200\}$ (Section 5.1), the LSE is applicable only for the case of $p = 50$. In this case, we see that the LSE has unbiased estimates for regression coefficients and Type I error rates close to the nominal level $\alpha = 0.05$. However, in terms of power, the proposed method is superior to the LSE.

6. Data analysis

6.1. Non-small cell lung cancer data

We investigated the lung cancer data, containing 131 patients with refractory non-small cell lung cancer. There are 33297 gene signatures per patient. The data is available at a genomics data repository http://www.ncbi.nlm.nih.gov/geo/ with accession number GSE33072.

The data record epidermal growth factor receptor (EGFR) that is the cell-surface receptor for members of the epidermal growth factor family of extracellular protein ligands. A patient with high EGFR index tends to have a cancer relapse or less probability of recovery (Dicker and Rodeck, 2005). We treat the EGFR index as a response variable (y_i) in our analyses.

Since the EGFR index is missing for 7 patients, we removed them and kept the remaining 124 patients for our analysis ($n = 124$). As commonly done, (e.g., Kim and Lee, 2007), we pre-filtered the top 394 of gene signatures with a high coefficient of variation (CV > 0.2) to insure the quality of regressors themselves, independently of the responses. After the pre-filtering, we standardized the design matrix.

6.2. Numerical results

The performance of the ridge and proposed methods are compared by prediction error. First, we divide 124 patients into 4 groups of equal size, denoted by $\Im_k$, $k = 1,\ 2,\ 3,\ 4$ (Fig. 6).

Second, the estimator based on all the data not in $\Im_k$ is calculated and denoted by $\hat{\boldsymbol{\beta}}^{(-k)}$. Then the prediction error (PE) is defined as

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

where $\hat{\boldsymbol{\beta}}^{(-k)}$ denotes either the ridge or proposed estimate with $p = 394$ regressors. The values of PE are evaluated over the 100 randomly chosen folds of $\Im_k$, $k = 1,\ 2,\ 3,\ 4$.

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

Figure 6. The 4-fold cross-validation. The n =124 patients are randomly divided into 4 groups each containing $124/4 = 31$ patients. For instance, patients in $\Im_3$ are removed and the remaining patients in $\Im_1 \cup \Im_2 \cup \Im_4$ are used for calculating regression coefficients $\hat{\boldsymbol{\beta}}^{(-3)}$.

Table 4. Comparison between the ridge regression and the proposed method over 100 random cross-validations on the lung cancer data.

	Shrinkage parameters		Threshold	Prediction Error (PE)		
No. of replicate	$\hat{\lambda}^{\text{Ridge}}$	$\hat{\lambda}^{\text{Proposed}}$	$\hat{\Delta}^{\text{Proposed}}$	*PE* (Ridge)		*PE* (Proposed)
1	294.4	410.7	1.448	0.502	>	0.454
2	258.6	349.3	1.418	0.703	<	0.753
3	315.2	431.2	1.598	0.481	>	0.441
4	310.8	419.5	1.598	0.495	>	0.452
5	306.7	418.2	1.545	0.471	>	0.441
6	325.6	447.3	1.538	0.518	>	0.472
7	312.5	442.8	1.500	0.474	>	0.433
8	323.6	435.8	1.523	0.476	>	0.438
9	307.6	423.9	1.470	0.507	>	0.464
10	303.9	426.2	1.313	0.472	>	0.436
≈	≈	≈	≈	≈		≈
99	320.8	441.7	1.448	0.481	>	0.442
100	285.2	393.7	1.583	0.505	>	0.461
Average	307.0	422.7	1.482	0.494	>	0.456

NOTE: PE (Prediction Error) is defined as $PE = \frac{1}{124}\sum_{k=1}^{4}\sum_{i\in\Im_k}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}^{(-k)})^2$.

Table 4 compares the *PE* between the ordinary ridge and proposed estimators. First, we see that the shrinkage parameter $\hat{\lambda}$ of proposed method is always greater than that of that of the ordinary ridge. This result is consistent with both the theoretical and simulation results. The *PE* of the proposed method is almost always less than that of the ordinary ridge over the 100 random cross-validations. Hence, the proposed method performs better than the ridge in terms of predicting the EGFR index.

Next, the performance on regressor selection is compared between the ordinary ridge and the proposed method. We first separate the $n = 124$ samples into two parts; the 62 training samples and the remaining 62 testing samples. The 62 training samples is used to estimate the shrinkage parameters $(\hat{\lambda},\ \hat{\Delta})$ and regression coefficients $\hat{\boldsymbol{\beta}}(\hat{\lambda},\ \hat{\Delta})$. Figure 7 demonstrates that the minimizers of $V(\lambda,\ \Delta)$ are computed as $\hat{\lambda}= 220$ and $\hat{\Delta}= 1.53$.

Table 5 compares the ordinary ridge and the proposed method in terms of the 20 selected genes based on P-values (Section 3.4). We see that the selected genes are very similar between

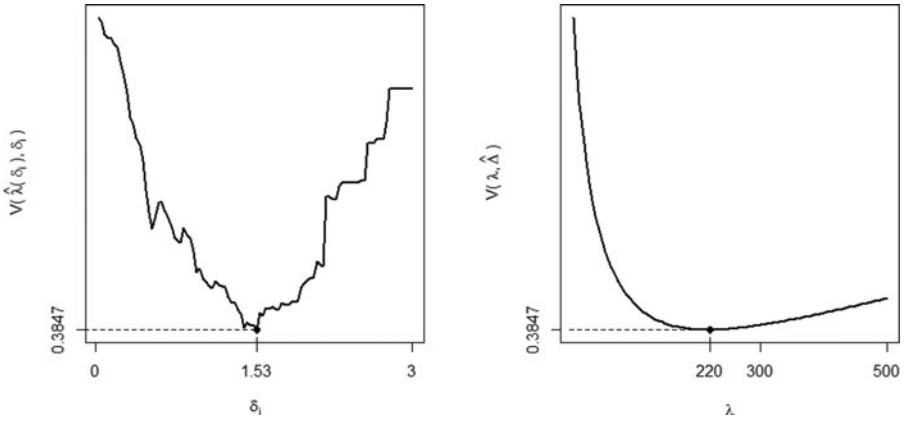

Figure 7. The profile plots of the GCV function for estimating Δ(left graph) and λ(right graph). All plots are based on 62 training samples. The left graph plots $V\{\hat{\lambda}(\delta_i),\ \delta_i\}$ against $\delta_i \in [0,\ 3]$. The right graph plots $V(\lambda,\ \hat{\Delta})$ against λ, where $\hat{\Delta} = 1.53$ is given.

Table 5. The 20 most strongly associated genes from the lung cancer data based on the ordinary ridge and proposed methods.

	Ordinary Ridge			Proposed method		
No.	Gene symbol	Coefficient	P-value	Gene symbol	Coefficient	P-value
1	FGA	−0.0381	2.6×10^{-7}	FGA	−0.0507	3.7×10^{-7}
2	AKR1B10	−0.0462	7.9×10^{-7}	AKR1B10	−0.0590	1.7×10^{-6}
3	CPS1	−0.0411	2.6×10^{-5}	CPS1	−0.0562	2.7×10^{-5}
4	KRT6A	−0.0345	4.5×10^{-5}	FGG	−0.0465	8.1×10^{-5}
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			PE=0.6648	

NOTE: The genes are ordered according to their P-values. The bottom row shows the prediction error (PE) of the linear predictors based on the 20 genes. The gene symbol is a specified abbreviation of the information of the gene (Wain, *et al.*, 2002). For instance, AKR1B10 is the abbreviation of "aldo-keto reductase family 1, member B10"; MSMB is the abbreviation of "microseminoprotein, beta-." The missing gene symbol is indicated by "∗" where the ID_REF of the original data is used.

the two methods, but they have different ordering (Table 5). For instance, the gene FGG is more strongly significant than the gene KRT6A for the proposed method, but their orders are reversed for the ordinary ridge.

The 20 selected regressors (genes) by the training samples are used to predict the response (the EGFR index) of the remaining 62 testing samples. Let $\hat{\boldsymbol{\beta}}^{Train}$ be either the ordinary ridge

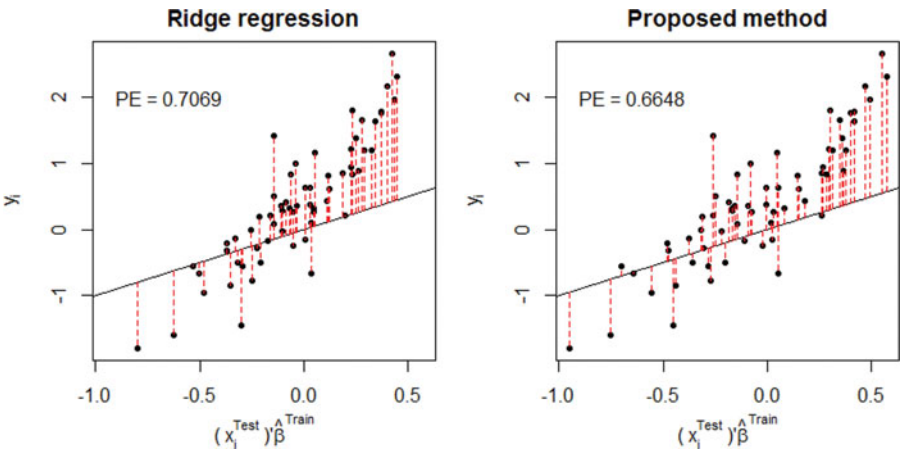

Figure 8. The plots of the EGFR index y_i against its predictor $(\mathbf{x}_i^{Test})^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{Train}$. The dash lines (red color) denote the distance between the value y_i and the predictor $(\mathbf{x}_i^{Test})^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{Train}$. The prediction error is $PE = \sum_{i\in Test}\{y_i - (\mathbf{x}_i^{Test})^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{Train}\}^2/62$.

Table 6. The 45 most strongly associated genes from the lung cancer data based on the ordinary ridge and proposed methods.

	Ordinary Ridge			Proposed method		
No.	Gene symbol	Coefficient	P-value	Gene symbol	Coefficient	P-value
1	FGA	−0.0381	2.6×10^{-7}	FGA	−0.0507	3.7×10^{-7}
2	AKR1B10	−0.0462	7.9×10^{-7}	AKR1B10	−0.0590	1.7×10^{-6}
3	CPS1	−0.0411	2.6×10^{-5}	CPS1	−0.0562	2.7×10^{-5}
4	KRT6A	−0.0345	4.5×10^{-5}	FGG	−0.0465	8.1×10^{-5}
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
21	7978626∗	0.0345	0.0088	CYP24A1	−0.0418	0.0149
22	GPX2	−0.0229	0.0099	SLC22A10	−0.0157	0.0167
23	CYP24A1	−0.0322	0.0107	7978626∗	0.0423	0.0178
24	8070755∗	0.0258	0.0121	8070755∗	0.0178	0.0227
25	7895136∗	0.0315	0.0127	SLC34A2	0.0240	0.0231
26	S100A7	−0.0211	0.0174	AZGP1	0.0155	0.0293
27	AKR1C2	−0.0207	0.0187	C9	−0.0046	0.0295
28	C9	−0.0066	0.0206	GPX2	−0.0253	0.0368
29	SLC22A10	−0.0103	0.0212	**8151583∗**	−0.0132	0.0372
30	SERPINB4	−0.0227	0.0221	BPIFA1	−0.0158	0.0392
31	VSIG1	0.0252	0.0234	MMP7	0.0180	0.0403
32	**IGJ**	0.0191	0.0234	SERPINB4	−0.0275	0.0442
33	**AKR1C1**	−0.0182	0.0247	GYS2	0.0046	0.0447
34	FGL1	−0.0183	0.0303	FGL1	−0.0124	0.0452
35	AZGP1	0.0200	0.0333	**MMP13**	0.0188	0.0478
36	MMP10	−0.0189	0.0365	VSIG1	0.0302	0.0480
37	**C4BPA**	0.0163	0.0391	**GLYATL1**	0.0068	0.0498
38	BPIFA1	−0.0206	0.0412	AKR1C2	−0.0226	0.0501
39	**APCS**	−0.0050	0.0446	S100A7	−0.0246	0.0509
40	**8128714∗**	−0.0195	0.0446	MMP10	−0.0246	0.0557
41	PLG	−0.0045	0.0453	PLG	−0.0033	0.0571
42	MMP7	0.0228	0.0465	**ZBBX**	−0.0121	0.0571
43	C4BPB	0.0153	0.0481	**7892787∗**	−0.0185	0.0575
44	**MFAP5**	−0.0193	0.0482	C4BPB	0.0109	0.0600
45	GYS2	0.0060	0.0492	**SPANXB2**	−0.0208	0.0608
		PE=0.5846				$PE = 0.5420$

NOTE: The genes are ordered according to their P-values. The bottom row shows the prediction error (PE) of the linear predictors based on the 45 genes. The gene symbol is a specified abbreviation of the information of the gene (Wain, *et al.*, 2002). For instance, AKR1B10 is the abbreviation of "aldo-keto reductase family 1, member B10"; MSMB is the abbreviation of "microseminoprotein, beta-." The missing gene symbol is indicated by "∗" where the ID_REF of the original data is used.

or proposed estimator of the 20 regression coefficients from the training sample (Table 5). Also, let $\mathbf{x}_i^{Test} = (x_{i(1)}, \ldots, x_{i(20)})^{\mathrm{T}}$ be the corresponding 20 gene signatures in the testing patients ($i \in Test$). Then we plot the EGFR index y_i against its predictor $(\mathbf{x}_i^{Test})^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{Train}$ for all $i \in Test$ (Fig. 8). Fig. 8 demonstrates that the predictors from both the ordinary ridge and proposed methods are highly predictive of the EGFR index. We compare the performance of

the two methods by the prediction error (PE) defined as

$$PE = \frac{1}{62}\sum_{i\in Test}\{y_i - (\mathbf{x}_i^{Test})^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{Train}\}^2.$$

The predictive performance of the proposed estimator ($PE = 0.6648$) is better than that of the ordinary ridge estimator ($PE = 0.7069$).

We also compare PE using the top 45 genes. The results are summarized in Table 6. The prediction error of the proposed method ($PE = 0.5420$) is still less than that of the ridge ($PE = 0.5846$). Thus, the proposed method gives a consistently better prediction for the EGFR index than the ordinary ridge without influenced by the threshold number.

7. Conclusion

This paper proposed a special class of the generalized ridge estimators under high-dimensionality. Unlike the ordinary ridge regression, the proposed method can utilize some prior knowledge on regression coefficient; if one is sure that jth regression coefficient is nonzero, one should give it less shrinkage. We showed that the proposed idea can be justified from the Bayesian view and the theoretical MSE calculations. In particular, the theoretical MSE calculations allow one to understand the mechanisms of the proposed class to improve upon both the LSE and the ordinary ridge by shrinking strongly on the null coefficients while shrinking weakly on the nonzero coefficients. Such theoretical results give us plausible reasons why the proposed method can outperform the existing procedures.

In simulations, we demonstrated the advantage of the proposed method under the sparse models, especially for the reduction of the MSE over the ordinary ridge regression. In particular, when the number of regressors is larger, the advantage of proposed method is more remarkable. Compared to the Lasso, the proposed method is superior for the moderately large dimensions ($p = 50$ and 100 with $n = 100$) while it is inferior for large dimension ($p = 200$ with $n = 100$). However, we believe that the disadvantage does not exclude the usefulness of the proposed method as it offers P-values for all the coefficients which are not directly possible by the Lasso. In addition, in many medical applications, the dimension p is reasonably reduced by initial quality controls for the regressors themselves (see Section 6.1).

In addition to proposing a new regression estimator, we developed significance testing for each regressor, namely $H_{0j}: \beta_j = 0$, which is extremely useful for regressor selection. This is important in applications to genetic (SNP or microarrays) data, in which biomedical researchers typically evaluate the significance of each gene in terms of P-values and select small fraction of significant genes. Applying the proposed method to the lung cancer microarray data, we successfully chose a small subset of regressors (genes) that are highly predictive to the response (the EGFR index). While our significance testing immediately follows the framework of Cule et al. (2011), we fail to give a theoretical support for it. Instead, we have justified its satisfactory control of the Type I error rate (a bit conservative) and power by simulations. As pointed out by Bühlmann (2013), the ridge estimator only produces a biased estimator for β_j that is not identifiable under high-dimensionality. An alternative proposal of Bühlmann (2013) is the test based on the bias-corrected ridge estimator, which gives mathematically more rigorous control for the Type I error, or even control for multiple testing. See also a similar bias-correction method of Zhang and Zhang (2014) to construct confidence intervals. So far, these bias-correction approaches seem to be restricted to a linear estimator, including the ordinary ridge and LSE. It is an interesting but challenging topic to follow this approach under the proposed non-linear estimator.

An important prerequisite to apply the proposed approach is a good initial estimate based on the univariate LSE (Section 3.2). While many time-to-event data analyses tend to use univariate estimates or univariate selection (Beer et al., 2002; Chen et al., 2007; Emura and Chen, 2014; Emura and Chen, 2016; Emura et al., 2017a,b; Jenssen, et al., 2002; Matsui, 2006) as a successful initial screening process for microarrays, there seems little theoretical support for it. In our simulations (not shown), we also observe that the univariate LSE has high sensitivity to separate the nonzero coefficients from the null coefficients, and hence it successfully serves as an initial screening process. We are currently trying to find more theoretical and numerical justifications for the univariate LSE under high-dimensionality.

Acknowledgements

We thank the anonymous reviewers for their helpful comments that improve the manuscript. We are also thankful to Prof. Tsai-Hung Fan, Dr. Chen Yi-Hau and Prof. Sheng-Mao Chang for their comments on an earlier version of our paper. This work was financially supported by the National Science Council of Taiwan (NSC101-2118-M008-002-MY2).

References

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

Araki, Y., Hattori, S. (2013). Efficient regularization parameter selection via information criteria. *Communications in Statistics—Simulation and Computation* 42(2):280–293.

Beer, D.G., Kardia, S.L.R., Huang, C.C., Giordano, T.J., Levin, A.M., et al. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine* 8:816–824.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* 19:1212–1242.

Candes, E., Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics* 35:2313–2351.

Chen, A. C., Emura, T. (2016). A modified Liu-type estimator with an intercept term under mixture experiments. *Communications in Statistics-Theory and Method*. DOI:10.1080/03610926.2015.1132327.

Chen, H.Y., Yu, S.L., Chen, C.H., Chang, G.C., Chen, C.Y., et al. (2007). A five-gene signature and clinical outcome in non-small-cell lung cancer. *New England Journal of Medicine* 356:11–20.

Cule, E., De Iorio, M. (2013). Ridge regression in prediction problems: Automatic choice of the ridge parameter. *Genetic Epidemiology* 37:704–714.

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

Dicker, A. P., Rodeck, U. (2005). Predicting the future from trials of the past: Epidermal growth factor receptor expression and outcome of fractionated radiation therapy trials. *Journal of Clinical Oncology* 23:5437–5439.

Emura, T., Chen, Y. H. (2016). Gene selection for survival data under dependent censoring: a copula based approach. *Statistical Methods in Medical Research* 25(6):2840–2857.

Emura, T., Chen, H. Y., Chen, Y. H. (2017a). R compound.Cox: Estimation, Gene Selection, and Survival Prediction Based on the Compound Covariate Method Under the Cox Proportional Hazard Model, version 3.1, CRAN.

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

Emura, T., Nakatochi, M., Matsui, S., Michimae, H., Rondeau, V. (2017b). Personalized dynamic prediction of death according to tumour progression and high-dimensional genetic factors: Meta-analysis with a joint model, *Statistical Methods in Medical Research*. DOI:10.1177/0962280216688032.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96:1348–1360.

Fan, J., Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. (with discussion). *Journal of Royal Statistical Society B* 70:849–911.
Fan, T. H. (2001). Noninformative Bayesian estimation for the optimum in a single factor quadratic response model. *Test* 10(2):225–240.
Friedman, J., Hastie, T., Simon, N., Tibshirani, R. (2015). Glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models, version 2.0–2. CRAN.
Golub, G. H., Heath, M., Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21:215–223.
Hansen, B. E. (2016). The risk of James–Stein and Lasso shrinkage. *Econometric Reviews* 35(8-10):1456–1470.
Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer-Verlag.
Hoerl, A. E., Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12:55–67.
Jang, D.-H., Anderson-Cook, C. M. (2015). Visualization approaches for evaluating ridge regression estimators in mixture and mixture-process experiments. *Quality and Reliability Engineering International* 31(8):1483–1494.
Jenssen, T.K., Kuo, W.P., Stokke, T., Hovig, E. (2002). Association between gene expressions in breast cancer and patient survival. *Human Genetics* 111:411–420.
Jimichi, M. (2008). Exact moments of feasible generalized ridge regression estimator and numerical evaluations. *Journal of the Japanese Society of Computationl Statistics* 21:1–20.
Kibria, B. M. G., Banik, S. (2016). Some ridge regression estimators and their performances. *Journal of Modern Applied Statistical Methods* 15(1):206–238.
Kim, S.-Y., Lee, J.-W. (2007). Ensemble clustering method based on the resampling similarity measure for gene expression data. *Statistical Methods in Medical Research* 16:539–564.
Loesgen, K.-H. (1990). A generalization and Bayesian interpretation of ridge-type estimators with good prior means. *Statistical Papers* 31:147–154.
Mallows, C. L. (1973). Some comments on Cp. *Technometrics* 15:661–675.
Matsui, S. (2006). Predicting survival outcomes using subsets of significant genes in prognostic marker studies with microarrays. *BMC Bioinformatics* 7:156.
Theobald, C. M. (1974). Generalizations of mean square error applied to ridge regression. *Journal of the Royal Statistical Society B* 36:103–106.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B* 58:267–288.
Trenkler, G. (1985). Mean square error matrix comparisons of estimators in linear regression. *Communications in Statistics A* 14:2495–2509.
Trenkler, G., Tourenburg, H. (1990). Mean squared error matrix comparisons between biased estimators – an overview of recent results. *Statistical Papers* 31:165–179.
Wain, J. M., Bruford, E. A., Lovering, R. C., et al. (2002). Guidelines for human gene nomenclature. *Genomics* 79:464–470.
Whittaker, J. C., Thompson, R., Denham, M. C. (2000). Marker-assisted selection using ridge regression. *Genetical Research* 75:249–252.
Wong, K. Y., Chiu, S. N. (2015). An iterative approach to minimize the mean squared error in ridge regression. *Computational Statistics* 30(2):625–639.
Yang, S. P. (2014). A class of generalized ridge estimator for high-dimensional linear regression. *National Central University Electronic Theses & Dissertations*. Taiwan, 1–41.
Zhang, C. H., Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society B* 76(1):217–242.
Zhao, X., Rødland, E. A., Sørlie, T., et al. (2011). Combining gene signatures improves prediction of breast cancer survival. *PLoS ONE* 6:e17845.