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

Szu-Peng Yang & Takeshi Emura

To cite this article: Szu-Peng Yang & Takeshi Emura (2016): A Bayesian approach with generalized ridge estimation for high-dimensional regression and testing, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2016.1193195

To link to this article: http://dx.doi.org/10.1080/03610918.2016.1193195

Accepted author version posted online: 28 Jul 2016.
Published online: 28 Jul 2016.

Submit your article to this journal

Article views: 37

View related articles

View Crossmark data
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.

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...
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 $y = X\beta + \varepsilon$, where

\[
y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.
\]

$X$ is a fixed (non-random) design matrix, $\beta \in \mathbb{R}^p$ is an unknown vector of regression coefficients and $\varepsilon$ follows $N_n(0, \sigma^2 I)$, where $\sigma^2 > 0$ is unknown and $I$ is the $n \times n$ identity matrix. Here, $x_i^T$ denotes the transpose of the $p \times 1$ vector $x_i$. 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 \( X^T X \) is invertible (non-singular), the least squares estimator (LSE) is

\[
\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T = (X^T X)^{-1} X^T y.
\]

The normality assumption on \( \varepsilon \) is not essential for statistical properties of the LSE and the proposed estimator in Section 3. The essential assumptions are \( E[\varepsilon] = 0 \) and \( \text{Cov}(\varepsilon) = \sigma^2 I \). 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 \( X^T X \) 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

\[
\hat{\beta}(\lambda) = (X^T X + \lambda I)^{-1} X^T y,
\]

where \( \lambda > 0 \) is a shrinkage parameter that gives the degree of shrinking \( \hat{\beta}(\lambda) \) 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 \( X^T X \) is \( \lambda_1 \geq \cdots \geq \lambda_p > 0 \) (i.e., the case of \( p < n \)), there exists a value \( \lambda > 0 \) such that

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

where \( \| a \|^2 = a^T a \) is the L2-norm for a vector \( a \). 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 L2-penalized residual sum square:

\[
\hat{\beta}(\lambda) = \arg \min_{\beta} \| y - X\beta \|^2 + \lambda \| \beta \|^2.
\]

If the L2-norm in the penalty term is replaced by the L1-norm \( \| \beta \|_1 = |\beta_1| + \cdots + |\beta_p| \), the resultant estimator is the Lasso (Tibshirani, 1996)

\[
\hat{\beta}^{\text{Lasso}}(\lambda) = \arg \min_{\beta} \| y - X\beta \|^2 + \lambda \| \beta \|_1.
\]

Some of the Lasso coefficients \( \hat{\beta}_j^{\text{Lasso}} \) 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 \( \lambda \)

In practice, the value of \( \lambda \) 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 \( \hat{\beta}^{(k)}(\lambda) \) be the ridge estimate of \( \beta \) without the \( k \) th data point \((y_k, x_k^T)\). If \( \lambda \) is chosen properly, then the \( k \)-th component \([X^{(k)} \hat{\beta}(\lambda)]_k\) of \( X^{(k)} \hat{\beta}(\lambda) \) predicts \( y_k \) well. The GCV is defined to be a weighted average of predicted square errors

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

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

\[
V(\lambda) = \frac{1}{n} \| \{I - A(\lambda)\} y \| ^2 \left[ \frac{1}{n} \text{Tr} \{I - A(\lambda)\} \right] ^2
\]

(1)

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

\[
\hat{\lambda} = \text{arg min}_{\lambda \geq 0} V(\lambda).
\]

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 \texttt{glmnet} package (Friedman et al. 2015).

3. Proposed method

3.1. Proposed idea

In the ridge estimator, the matrix \( X^T X \) is replaced by \( X^T X + \lambda I \) with a shrinkage parameter \( \lambda > 0 \). Alternatively, the generalized ridge estimator considers \( X^T X + W \) for a general diagonal matrix \( W = \text{diag}(w_1, \ldots, w_p) \), 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 \( W = \text{diag}(w_1, \ldots, w_p) \), where

\[
w_j = \begin{cases} 
\lambda \gamma & \text{if } \beta_j \neq 0, \\
\lambda & \text{if } \beta_j = 0,
\end{cases}
\]

(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 $\beta$ 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 $\hat{\beta}^0 = (\hat{\beta}_1^0, \ldots, \hat{\beta}_p^0)^T$, defined as

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

where $x_j$, for $j = 1, \ldots, p$, are the columns of $X$. Note that $\hat{\beta}^0$ 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 $\beta_j$ is more likely to be nonzero. Hence, we propose a special class of generalized ridge estimators

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

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

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

for $j = 1, \ldots, p$, $\text{SD}(\hat{\beta}^0) = \sqrt{\sum_{j=1}^p (\hat{\beta}_j^0 - \bar{\beta}_j^0)^2/(p-1)}$, and $\bar{\beta}_j^0 = \sum_{j=1}^p \hat{\beta}_j^0/p$. We call $\Delta$ “thresholding parameter.” Under the sparse model ($\beta \approx 0$), the histogram of $|\hat{\beta}_j^0|/\text{SD}(\hat{\beta}^0)$, $j = 1, \ldots, p$, is well-approximated by $N(0, 1)$. This implies that $|\hat{\beta}_j^0|/\text{SD}(\hat{\beta}^0)$ 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 $(\lambda, \Delta)$ by GCV

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

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

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

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

Given $\Delta$, the CGV function is continuous in $\lambda$, 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 $\Delta$, 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

$$
\hat{\beta}(\lambda, \Delta) = [X^T X + \lambda \hat{W}(\Delta)]^{-1} X^T y.
$$
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 $j$th component of $\hat{\beta}(\hat{\lambda}, \hat{\Delta})$. Define the Wald statistics

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

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 $j$th diagonal of the estimated covariance matrix,

$$\text{Cov}(\hat{\beta}(\hat{\lambda}, \hat{\Delta})) = \hat{\sigma}^2[X^TX + \hat{\lambda}\hat{W}(\hat{\Delta})]^{-1}X^TX[\hat{X}^TX + \hat{\lambda}\hat{W}(\hat{\Delta})]^{-1},$$

where

$$\hat{\sigma}^2 \equiv (y - X\hat{\beta}(\hat{\lambda}, \hat{\Delta}))(y - X\hat{\beta}(\hat{\lambda}, \hat{\Delta}))/v,$$

$$v \equiv \text{Tr}\{I - A(\hat{\lambda}, \hat{\Delta})\}^2 = n - \text{Tr}\{2A(\hat{\lambda}, \hat{\Delta}) - A(\hat{\lambda}, \hat{\Delta})^2\},$$

where $A(\hat{\lambda}, \hat{\Delta}) = X(X^TX + \hat{\lambda}\hat{W}(\hat{\Delta}))^{-1}X^T$. Note that $v$ 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 ($\beta \approx 0$). 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 $\beta \sim N_p(0, \sigma^2W^{-1})$, where $W^{-1}$ is a $p \times p$ covariance matrix (hyperparameters). Also, $y = X\beta + \varepsilon$, where $\varepsilon \sim N_n(0, \sigma^2I)$, as in Section 2. Assume that $\sigma^2$ is known. After some calculations, the posterior density of $\beta$ becomes

$$f(\beta|y, X, W) \propto \exp\left[\frac{-1}{2\sigma^2}(\beta - (X^TX + W)^{-1}X^T\hat{y})^T(X^TX + W)(\beta - (X^TX + W)^{-1}X^T\hat{y})\right].$$
We see that the posterior mean of $\beta$ is exactly the generalized ridge estimator,
\[
E(\beta | y, X, W) = (X^T X + W)^{-1} X^T y.
\]

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 $W = \text{diag}(w_1, \ldots, w_p)$, and $w_1, \ldots, w_p \geq 0$. It follows that $E[\beta_j] = 0$ and $\text{Var}[\beta_j] = \sigma^2 w_j^{-1}$ which expresses the uncertainty of prior belief that $\beta_j$ 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 $\beta_j$ 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 $\gamma$ 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 $W = \text{diag}(w_1, \ldots, w_p)$ by $\hat{\lambda} \hat{W} (\hat{\lambda})$. 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 $\hat{\beta}$ be any estimator of $\beta$. The MSE matrix is defined as a $p \times p$ matrix
\[
M(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T] = C + dd^T,
\]
where $C = \text{Cov}(\hat{\beta})$ is the covariance matrix of $\hat{\beta}$ and $d = \text{Bias}(\hat{\beta}) = E(\hat{\beta}) - \beta$ is the bias of $\hat{\beta}$. The MSE of $\hat{\beta}$ is the trace of $M(\hat{\beta})$. Note that the diagonals of a nonnegative definite (n.n.d.) matrix are nonnegative. This implies that, if $M(\hat{\beta}_1) - M(\hat{\beta}_2)$ is n.n.d., then $\text{MSE}(\hat{\beta}_1) \geq \text{MSE}(\hat{\beta}_2)$ for two estimators $\hat{\beta}_1$ and $\hat{\beta}_2$.

In our analysis of the generalized ridge estimator, we compare the performance between the ordinary ridge estimator $\hat{\beta}(\lambda)$ with $\lambda = 1$ and the proposed class
\[
\hat{\beta}(\lambda, \gamma) = [X^T X + W(\lambda, \gamma)]^{-1} X^T y,
\]
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
\[
M(\hat{\beta}(1)) - M(\hat{\beta}(\lambda, \gamma)), \text{ and } M(\hat{\beta}(0)) - M(\hat{\beta}(\lambda, \gamma))
\]
are n.n.d. Trenkler (1985) established a useful lemma as follows:

**Lemma 1:** Suppose $A$ is a symmetric $p \times p$ matrix, $a$ is an $p \times 1$ vector and $\eta$ is a positive real number. Then $\eta A - aa^T$ is n.n.d. if and only if
\begin{enumerate}
  \item $A$ is n.n.d.,
  \item $a = Av$ for some $v \in R^p$
\end{enumerate}
and
iii) \( a^T A^{-1} a \leq \eta. \)
where \( A^{-1} \) 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 \( M(\hat{\beta}_1) - M(\hat{\beta}_2) = (C_1 + \mathbf{d}_1 \mathbf{d}_1^T) - (C_2 + \mathbf{d}_2 \mathbf{d}_2^T) \) is n.n.d., where \( C_i \) is the covariance and \( \mathbf{d}_i \) is the bias for the ridge-type estimator \( \hat{\beta}_i = (X^T X + W_i)^{-1} X^T y, i = 1, 2. \) As a version of Lemma 1, it is convenient to establish the following theorem.

**Theorem 1:** \( M(\hat{\beta}_1) - M(\hat{\beta}_2) \) is n.n.d. if all the three conditions hold:

i) \( \frac{C_1 - C_2}{\sigma^2} \) is n.n.d.,

ii) \( \mathbf{d}_2 = (C_1 - C_2 + \mathbf{d}_1 \mathbf{d}_1^T) \mathbf{v} \) for some \( \mathbf{v} \in \mathbb{R}^p, \)

iii) \( \mathbf{d}_1^T (C_1 - C_2 + \mathbf{d}_1 \mathbf{d}_1^T)^{-1} \mathbf{d}_2 \leq 1. \)

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

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

i.e., \( C_1 - C_2 + \mathbf{d}_1 \mathbf{d}_1^T \) 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

\[
(C_1 - C_2 + \mathbf{d}_1 \mathbf{d}_1^T) - \mathbf{d}_2 \mathbf{d}_2^T = M(\hat{\beta}_1) - M(\hat{\beta}_2)
\]

is n.n.d. \( \square \)

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

**Example**

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

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

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

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

\[
\frac{1}{\sigma^2} (C_1 - C_2) = \text{diag} \left( \frac{1}{4} - \frac{1}{(1 + \gamma \lambda)^2}, \frac{1}{4} - \frac{1}{(1 + \lambda)^2}, \ldots, \frac{1}{4} - \frac{1}{(1 + \lambda)^2} \right),
\]

\[
C_1 - C_2 + \mathbf{d}_1 \mathbf{d}_1^T = \text{diag} \left( \frac{\sigma^2(1 + \gamma \lambda)^2 - 4}{4(1 + \gamma \lambda)^2}, \frac{\sigma^2(\lambda + 3)(\lambda - 1)}{4(1 + \lambda)^2}, \ldots, \frac{\sigma^2(\lambda + 3)(\lambda - 1)}{4(1 + \lambda)^2} \right).
\]
mean that \( \rho \) right panel shows the non-orthonormal case zero as diag then \( \lambda \) are satisfied when \( 2 \) MSE Therefore, the proposed class. Under \( \sigma \) H 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 \( d_2^\text{T}(C_1 - C_2 + d_1 d_1^\text{T})^{-1} d_2 \leq 1 \) if \( \beta_1 \) is near 0. For the orthonormal case, Equation (3) allows the expression.

\[
d_2^\text{T}(C_1 - C_2 + d_1 d_1^\text{T})^{-1} d_2 = \frac{6.25\beta_1^2}{\sigma^2\{(1 + 1.25\gamma)^2 - 4\}} + (1 + 1.25\gamma)^2\beta_1^2 = \frac{6.25\beta_1^2}{1.0625\sigma^2 + 5.0625\beta_1^2}.
\]

Hence, the range of \( \beta_1 \) is \([-0.8947, 0.8947]\). In this range, Theorem 1 verifies the relation \( MSE(\hat{\beta}_1) \geq MSE(\hat{\beta}_2) \) [see also Yang (2014) who numerically verified the relation \( MSE(\hat{\beta}_1) \geq MSE(\hat{\beta}_2) \)].

Some simplification is possible as

\[
d_1^\text{T}(C_1 - C_2 + d_1 d_1^\text{T})^{-1} d_2 = \frac{4\lambda^2\gamma^2\beta_1^2}{\sigma^2\{(1 + \lambda\gamma)^2 - 4\}} + (1 + \lambda\gamma)^2\beta_1^2.
\] (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 \( d_2 = (C_1 - C_2 + d_1 d_1^\text{T})v \) for \( v^\text{T} = (v_1, 0^\text{T}). \) That is, Conditions (i) and (ii) of Theorem 1 are satisfied when \( \lambda = 2.5 \) and \( \gamma \geq 2/5. \) From Equation (3), \( d_1^\text{T}(C_1 - C_2 + d_1 d_1^\text{T})^{-1} d_2 \) goes to zero as \( \beta_1 \to 0 \) (see also Fig. 1). Thus, if \( \beta_1 \) is near 0, Condition iii) of Theorem 1 is verified. Therefore, \( M(\hat{\beta}_1) - M(\hat{\beta}_2) \) is n.n.d., implying \( MSE(\hat{\beta}_1) \geq MSE(\hat{\beta}_2) \). 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 \( d_1^\text{T}((C_1 - C_2 + d_1 d_1^\text{T})^{-1} d_2 \) can be strictly less than one.

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

\[
d_1 = 0, d_2 = (X^\text{T}X + W_2)^{-1}X^\text{T}X\beta - \beta = \frac{-\lambda\gamma\beta_1}{1 + \lambda\gamma} \begin{pmatrix} 1 \\ 0 \end{pmatrix},
\]

\[
\frac{1}{\sigma^2}(C_1 - C_2) = \text{diag}(1 - \frac{1}{(1 + \lambda\gamma)^2}, 1 - \frac{1}{(1 + \lambda)^2}, \ldots, 1 - \frac{1}{(1 + \lambda)^2}).
\]
\[
d_2^T(C_1 - C_2 + d_1d_1^T)^{-1}d_2 = \frac{\lambda^2\gamma^2\beta^2_1}{\sigma^2\{(1 + \lambda\gamma)^2 - 1\}},
\]

Similar to the case for the ridge, Conditions i)-iii) hold when \(\beta_1\) 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 \(\beta_1\) is near 0. This conclusion would continue to hold even for non-orthonormal cases of \(\rho \neq 0\), where the tractable formula of \(d_2^T(C_1 - C_2 + d_1d_1^T)^{-1}d_2\) is no longer available. We compute \(d_2^T(C_1 - C_2 + d_1d_1^T)^{-1}d_2\) numerically with \(\rho = 0.9\), and verify \(d_2^T(C_1 - C_2 + d_1d_1^T)^{-1}d_2 \leq 1\) if \(\beta_1\) 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 \(\beta\), 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 \(\text{rank}(C_1 - C_2 + d_1d_1^T) = \text{rank}([C_1 - C_2 + d_1d_1^T|d_2])\), then \(d_2\) belongs to the column space of \(C_1 - C_2 + d_1d_1^T\). This implies \(d_2 = (C_1 - C_2 + d_1d_1^T)v\) for some \(v \in R^p\). If so, Condition (ii) in Theorem 1 is satisfied. One can use \(\text{qr}()\) \$rank\ in R to obtain the rank of a matrix. For instance, one can check Condition (ii) under \(p = 100, n = 50, \lambda = 2.5, \gamma = 1/2, \sigma = 1, W_1 = I_{100}\).

\[
W_2 = \text{diag}(\lambda\gamma, \ldots, \lambda\gamma, \lambda, \ldots, \lambda), \quad \beta = (0.5, \ldots, 0.5, 0, \ldots, 0)^T
\]

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

\[
C_1 - C_2 + d_1d_1^T =
\begin{bmatrix}
-0.0046 & 0.0059 & \cdots & 0.0026 \\
0.0059 & 0.0070 & \cdots & 0.0086 \\
\vdots & \vdots & \ddots & \vdots \\
0.0026 & 0.0086 & \cdots & 0.0065
\end{bmatrix}
\]

\[
[C_1 - C_2 + d_1d_1^T|d_2] =
\begin{bmatrix}
-0.0046 & 0.0059 & \cdots & 0.0026 & -0.0148 \\
0.0059 & 0.0070 & \cdots & 0.0086 & -0.0912 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0.0026 & 0.0086 & \cdots & 0.0065 & -0.0401
\end{bmatrix}
\]

Condition (ii) is verified as \(\text{rank}(C_1 - C_2 + d_1d_1^T) = \text{rank}([C_1 - C_2 + d_1d_1^T|d_2]) = 70\). 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 \(\hat{\beta} = Ly\), 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 $\beta$ has $(q + r)$ nonzero terms and $p - (q + r)$ zero terms such that

$$
\beta = \left( \frac{b}{q}, \ldots, \frac{b}{q}, \frac{d}{r}, \ldots, \frac{d}{r}, 0, \ldots, 0 \right)^T,
$$

for $b, d \in \mathbb{R}$. 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 $X = (x_1, \ldots, x_n)^T$ by letting

$$
x_i^T = \left( \frac{z_{i1} + u_i}{\sqrt{2}}, \ldots, \frac{z_{q} + u_i}{\sqrt{2}}, \frac{z_{(q+1)} + v_i}{\sqrt{2}}, \ldots, \frac{z_{(q+r)} + v_i}{\sqrt{2}}, \frac{z_{(q+r+1)}}{\sqrt{2}}, \ldots, \frac{z_{ip}}{\sqrt{2}} \right),
$$

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

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

The design matrix that has the same correlation structure is generated by $X \cdot \text{pathway routine in R compound.Cox package}$ (Emura et al., 2017a).

After generating the designed matrix, we set $\mathbf{y} = X \hat{\beta} + \varepsilon$, where $\varepsilon \sim N_n(0, I)$. Based on 500 replications (on $\varepsilon$), 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 $\lambda$ and $\Delta$
We compare the performances of the proposed method and ridge regression in terms of the MSE curve, which is the plot of the MSE

$$
MSE(\hat{\beta}(\lambda)) = \mathbb{E}[\left(\hat{\beta}(\lambda) - \beta\right)^T(\hat{\beta}(\lambda) - \beta)],
$$

$$
MSE(\hat{\beta}(\lambda, \Delta^*)) = \mathbb{E}[\left(\hat{\beta}(\lambda, \Delta^*) - \beta\right)^T(\hat{\beta}(\lambda, \Delta^*) - \beta)],
$$

against $\lambda$. 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 $\lambda$ are determined. Nevertheless, the curve gives us some insight about potential gain of the MSE with varying
values of $\lambda$. The MSE of the LSE is calculated as $MSE(\hat{\beta}) = MSE(\hat{\beta}(0))$ only for the case of $p = 50$.

Figures 2–5 depict the two MSE curves, $MSE(\hat{\beta}(\lambda))$ and $MSE(\hat{\beta}(\lambda, \Delta^*))$. 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., $MSE(\hat{\beta}(\lambda)) < MSE(\hat{\beta}(0))$. 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 $\hat{\beta}(\hat{\lambda}, \hat{\Delta})$ can be less than that of the ordinary ridge estimator $\hat{\beta}(\lambda)$. In addition, the superiority of

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

**Figure 3.** The MSE curves of the ridge and proposed estimators against $\lambda$ with $b = d = 10$. 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 $\lambda$ 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 $\lambda$ is properly estimated by data, the proposed method should give stronger shrinkage toward the zero vector.

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

**Figure 5.** The MSE curves of the ridge and proposed estimators against $\lambda$ with $b = 10$ and $d = -10$. The points denote the minimum of each curve.
5.3. MSE comparison for estimated $\lambda$ and $\Delta$

Rather than fixing $\lambda$ and $\Delta$ in Section 5.2, we consider the variability of estimating $\lambda$ and $\Delta$ for the ridge and proposed estimators. The performances are then evaluated by the “feasible” MSE for the ridge and proposed estimator, respectively defined as

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

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

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

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

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

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

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

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 $MSE(\hat{\beta})$. 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 $MSE(\hat{\beta})$ by half. The first column shows that the shrinkage parameter

<table>
<thead>
<tr>
<th>$b = d = 5$</th>
<th>$p = 50$</th>
<th>Ridge</th>
<th>Lasso</th>
<th>Proposed</th>
<th>$E(\hat{\lambda})$</th>
<th>$E(\hat{\Delta})$</th>
<th>$MSE(\hat{\beta}_1)$</th>
<th>$MSE(\hat{\beta}_p)$</th>
<th>$MSE(\hat{\beta})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = 100$</td>
<td>Ridge</td>
<td>23.30</td>
<td>-</td>
<td>0.0112</td>
<td>0.0077</td>
<td>0.4663</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lasso</td>
<td>20.96</td>
<td>-</td>
<td>0.0274</td>
<td>0.0023</td>
<td>0.5978</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed</td>
<td>47.34</td>
<td>0.92</td>
<td>0.0107</td>
<td>0.0048</td>
<td>0.3763</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 150$</td>
<td>Ridge</td>
<td>23.86</td>
<td>-</td>
<td>0.0080</td>
<td>0.0086</td>
<td>1.0228</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lasso</td>
<td>58.27</td>
<td>-</td>
<td>0.0244</td>
<td>0.0001</td>
<td>1.0655</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed</td>
<td>46.96</td>
<td>1.31</td>
<td>0.0086</td>
<td>0.0058</td>
<td>0.7191</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 200$</td>
<td>Ridge</td>
<td>20.45</td>
<td>-</td>
<td>0.0118</td>
<td>0.0065</td>
<td>1.2909</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lasso</td>
<td>28.77</td>
<td>-</td>
<td>0.0350</td>
<td>0.0003</td>
<td>0.6628</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed</td>
<td>43.83</td>
<td>1.53</td>
<td>0.0171</td>
<td>0.0037</td>
<td>0.7822</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$b = d = 10$</td>
<td>$p = 50$</td>
<td>Ridge</td>
<td>9.51</td>
<td>-</td>
<td>0.0202</td>
<td>0.0119</td>
<td>0.7755</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lasso</td>
<td>44.13</td>
<td>-</td>
<td>0.0262</td>
<td>0.0004</td>
<td>0.6301</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed</td>
<td>21.55</td>
<td>0.81</td>
<td>0.0174</td>
<td>0.0083</td>
<td>0.6177</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 100$</td>
<td>Ridge</td>
<td>9.32</td>
<td>-</td>
<td>0.0314</td>
<td>0.0178</td>
<td>2.4243</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lasso</td>
<td>174.06</td>
<td>-</td>
<td>0.0317</td>
<td>0.0000</td>
<td>5.1991</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed</td>
<td>20.28</td>
<td>1.28</td>
<td>0.0266</td>
<td>0.0116</td>
<td>1.7087</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 150$</td>
<td>Ridge</td>
<td>5.18</td>
<td>-</td>
<td>0.0320</td>
<td>0.0124</td>
<td>3.3783</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lasso</td>
<td>30.09</td>
<td>-</td>
<td>0.0356</td>
<td>0.0002</td>
<td>0.6453</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed</td>
<td>14.98</td>
<td>1.55</td>
<td>0.0437</td>
<td>0.0086</td>
<td>1.8486</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 200$</td>
<td>Ridge</td>
<td>0.76</td>
<td>-</td>
<td>0.0053</td>
<td>0.0065</td>
<td>3.3301</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lasso</td>
<td>41.14</td>
<td>-</td>
<td>0.0348</td>
<td>0.0002</td>
<td>0.6711</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed</td>
<td>5.58</td>
<td>1.50</td>
<td>0.0079</td>
<td>0.0046</td>
<td>1.8204</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>$E(\hat{\lambda})$</th>
<th>$E(\hat{\lambda}/\Delta_1)$</th>
<th>$MSE(\hat{\beta}_1)$</th>
<th>$MSE(\hat{\beta}_p)$</th>
<th>$MSE(\hat{\beta})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b = 5$, $d = -5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 50$</td>
<td>Ridge</td>
<td>21.36</td>
<td>0.0121</td>
<td>0.0081</td>
<td>0.4872</td>
</tr>
<tr>
<td></td>
<td>Lasso</td>
<td>19.65</td>
<td>0.0275</td>
<td>0.0026</td>
<td>0.6106</td>
</tr>
<tr>
<td></td>
<td>Proposed</td>
<td>43.49</td>
<td>0.91</td>
<td>0.0108</td>
<td>0.3981</td>
</tr>
<tr>
<td>$p = 100$</td>
<td>Ridge</td>
<td>26.69</td>
<td>0.0075</td>
<td>0.0098</td>
<td>0.8622</td>
</tr>
<tr>
<td></td>
<td>Lasso</td>
<td>103.43</td>
<td>0.0248</td>
<td>0.0000</td>
<td>0.8379</td>
</tr>
<tr>
<td></td>
<td>Proposed</td>
<td>52.78</td>
<td>1.03</td>
<td>0.0073</td>
<td>0.5693</td>
</tr>
<tr>
<td>$p = 150$</td>
<td>Ridge</td>
<td>17.84</td>
<td>0.0053</td>
<td>0.0169</td>
<td>1.5185</td>
</tr>
<tr>
<td></td>
<td>Lasso</td>
<td>51.01</td>
<td>0.0268</td>
<td>0.0003</td>
<td>0.7663</td>
</tr>
<tr>
<td></td>
<td>Proposed</td>
<td>38.25</td>
<td>1.34</td>
<td>0.0061</td>
<td>0.9057</td>
</tr>
<tr>
<td>$p = 200$</td>
<td>Ridge</td>
<td>11.76</td>
<td>0.0173</td>
<td>0.0057</td>
<td>1.3875</td>
</tr>
<tr>
<td></td>
<td>Lasso</td>
<td>29.63</td>
<td>0.0333</td>
<td>0.0004</td>
<td>0.6588</td>
</tr>
<tr>
<td></td>
<td>Proposed</td>
<td>32.55</td>
<td>1.52</td>
<td>0.0192</td>
<td>0.8059</td>
</tr>
<tr>
<td>$b = 10$, $d = -10$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 50$</td>
<td>Ridge</td>
<td>8.79</td>
<td>0.0220</td>
<td>0.0123</td>
<td>0.7997</td>
</tr>
<tr>
<td></td>
<td>Lasso</td>
<td>32.04</td>
<td>0.0266</td>
<td>0.0010</td>
<td>0.6056</td>
</tr>
<tr>
<td></td>
<td>Proposed</td>
<td>19.76</td>
<td>0.0191</td>
<td>0.0087</td>
<td>0.6527</td>
</tr>
<tr>
<td>$p = 100$</td>
<td>Ridge</td>
<td>10.71</td>
<td>0.0194</td>
<td>0.0241</td>
<td>1.9339</td>
</tr>
<tr>
<td></td>
<td>Lasso</td>
<td>263.10</td>
<td>0.0256</td>
<td>0.0000</td>
<td>2.8712</td>
</tr>
<tr>
<td></td>
<td>Proposed</td>
<td>23.13</td>
<td>0.96</td>
<td>0.0180</td>
<td>1.3168</td>
</tr>
<tr>
<td>$p = 150$</td>
<td>Ridge</td>
<td>3.61</td>
<td>0.0092</td>
<td>0.0472</td>
<td>3.9864</td>
</tr>
<tr>
<td></td>
<td>Lasso</td>
<td>161.97</td>
<td>0.0360</td>
<td>0.0000</td>
<td>3.0474</td>
</tr>
<tr>
<td></td>
<td>Proposed</td>
<td>12.07</td>
<td>1.37</td>
<td>0.0111</td>
<td>2.0756</td>
</tr>
<tr>
<td>$p = 200$</td>
<td>Ridge</td>
<td>1.32</td>
<td>0.0495</td>
<td>0.0131</td>
<td>3.5478</td>
</tr>
<tr>
<td></td>
<td>Lasso</td>
<td>32.65</td>
<td>0.0329</td>
<td>0.0002</td>
<td>0.6272</td>
</tr>
<tr>
<td></td>
<td>Proposed</td>
<td>6.83</td>
<td>1.56</td>
<td>0.0479</td>
<td>1.7845</td>
</tr>
</tbody>
</table>

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 $MSE(\hat{\beta})$.

The MSE of the Lasso produces quite different pattern from the two ridge estimators (Table 1). In general, the Lasso gives the smallest $MSE(\hat{\beta})$ 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

$$H_0 : \beta_1 = 0 \ v.s. \ H_1 : \beta_1 \neq 0,$$

$$H_0 : \beta_{50} = 0 \ v.s. \ H_1 : \beta_{50} \neq 0.$$
Table 2. Simulation results for testing $H_0 : \beta_{50} = 0$ using the proposed estimator (the LSE in parenthesis) based on 500 replicates.

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

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

$$Rejection\ rate = \frac{1}{500} \sum_{i=1}^{500} I(\ Z^{(i)} > Z_{a/2}),$$

where $I(\cdot)$ is the indicator function, and $Z^{(i)}$ 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.

<table>
<thead>
<tr>
<th>$\beta_1 = 5/10 = 0.5, b = d = 5$</th>
<th>$E(\hat{\beta}_1)$</th>
<th>$sd(\hat{\beta}_1)$</th>
<th>$E(Z_1)$</th>
<th>$sd(Z_1)$</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p = 50$</td>
<td>0.465(0.502)</td>
<td>0.097(0.202)</td>
<td>4.999(2.562)</td>
<td>1.009(1.065)</td>
<td>0.998(0.706)</td>
</tr>
<tr>
<td>$p = 100$</td>
<td>0.514</td>
<td>0.091</td>
<td>5.968</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>$p = 150$</td>
<td>0.411</td>
<td>0.095</td>
<td>5.637</td>
<td>0.940</td>
<td>1</td>
</tr>
<tr>
<td>$p = 200$</td>
<td>0.499</td>
<td>0.065</td>
<td>6.530</td>
<td>0.928</td>
<td>1</td>
</tr>
<tr>
<td>$\beta_1 = 10/10 = 1, b = d = 10$</td>
<td>$p = 50$</td>
<td>0.956(1.002)</td>
<td>0.124(0.202)</td>
<td>7.573(5.177)</td>
<td>1.034(1.143)</td>
</tr>
<tr>
<td>$p = 100$</td>
<td>1.067</td>
<td>0.148</td>
<td>8.338</td>
<td>1.137</td>
<td>0.998</td>
</tr>
<tr>
<td>$p = 150$</td>
<td>0.851</td>
<td>0.147</td>
<td>7.847</td>
<td>0.967</td>
<td>1</td>
</tr>
<tr>
<td>$p = 200$</td>
<td>1.020</td>
<td>0.084</td>
<td>8.494</td>
<td>0.919</td>
<td>1</td>
</tr>
<tr>
<td>$\beta_1 = 5/10 = 0.5, b = 5, d = -5$</td>
<td>$p = 50$</td>
<td>0.465(0.502)</td>
<td>0.096(0.201)</td>
<td>4.766(2.562)</td>
<td>0.990(1.065)</td>
</tr>
<tr>
<td>$p = 100$</td>
<td>0.479</td>
<td>0.083</td>
<td>5.882</td>
<td>1.032</td>
<td>0.996</td>
</tr>
<tr>
<td>$p = 150$</td>
<td>0.484</td>
<td>0.076</td>
<td>5.830</td>
<td>0.926</td>
<td>1</td>
</tr>
<tr>
<td>$p = 200$</td>
<td>0.398</td>
<td>0.085</td>
<td>5.450</td>
<td>0.882</td>
<td>1</td>
</tr>
<tr>
<td>$\beta_1 = 10/10 = 1, b = 10, d = -10$</td>
<td>$p = 50$</td>
<td>0.947(1.002)</td>
<td>0.128(0.201)</td>
<td>7.308(5.177)</td>
<td>1.020(1.143)</td>
</tr>
<tr>
<td>$p = 100$</td>
<td>0.982</td>
<td>0.133</td>
<td>8.126</td>
<td>1.110</td>
<td>0.998</td>
</tr>
<tr>
<td>$p = 150$</td>
<td>1.009</td>
<td>0.105</td>
<td>7.991</td>
<td>0.921</td>
<td>1</td>
</tr>
<tr>
<td>$p = 200$</td>
<td>0.830</td>
<td>0.133</td>
<td>7.339</td>
<td>0.794</td>
<td>1</td>
</tr>
</tbody>
</table>

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 $\mathcal{S}_k$, $k = 1, 2, 3, 4$ (Fig. 6).

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

$$PE = \frac{1}{124} \sum_{k=1}^{4} \sum_{i \in \mathcal{S}_k} (y_i - x_i^T \hat{\beta}^{(-k)})^2,$$

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

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train $\mathcal{S}_1$</td>
<td>Train $\mathcal{S}_2$</td>
<td>Test $\mathcal{S}_3$</td>
<td>Train $\mathcal{S}_4$</td>
</tr>
<tr>
<td></td>
<td>(31 patients)</td>
<td>(31 patients)</td>
<td>(31 patients)</td>
<td>(31 patients)</td>
</tr>
</tbody>
</table>

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 $\mathcal{S}_3$ are removed and the remaining patients in $\mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_4$ are used for calculating regression coefficients $\hat{\beta}^{(-3)}$. 
Table 4. Comparison between the ridge regression and the proposed method over 100 random cross-validations on the lung cancer data.

<table>
<thead>
<tr>
<th>No. of replicate</th>
<th>Ridge</th>
<th>Proposed</th>
<th>Threshold</th>
<th>Prediction Error (PE)</th>
<th>Prediction Error (Proposed)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\lambda}$</td>
<td>$\hat{\lambda}$</td>
<td>$\Delta$</td>
<td>(Ridge)</td>
<td>(Proposed)</td>
</tr>
<tr>
<td>1</td>
<td>294.4</td>
<td>410.7</td>
<td>1.448</td>
<td>0.502</td>
<td>&gt;</td>
</tr>
<tr>
<td>2</td>
<td>258.6</td>
<td>349.3</td>
<td>1.418</td>
<td>0.703</td>
<td>&lt;</td>
</tr>
<tr>
<td>3</td>
<td>315.2</td>
<td>431.2</td>
<td>1.598</td>
<td>0.481</td>
<td>&gt;</td>
</tr>
<tr>
<td>4</td>
<td>310.8</td>
<td>419.5</td>
<td>1.598</td>
<td>0.495</td>
<td>&gt;</td>
</tr>
<tr>
<td>5</td>
<td>306.7</td>
<td>418.2</td>
<td>1.545</td>
<td>0.471</td>
<td>&gt;</td>
</tr>
<tr>
<td>6</td>
<td>325.6</td>
<td>447.3</td>
<td>1.538</td>
<td>0.518</td>
<td>&gt;</td>
</tr>
<tr>
<td>7</td>
<td>312.5</td>
<td>442.8</td>
<td>1.500</td>
<td>0.474</td>
<td>&gt;</td>
</tr>
<tr>
<td>8</td>
<td>323.6</td>
<td>435.8</td>
<td>1.523</td>
<td>0.476</td>
<td>&gt;</td>
</tr>
<tr>
<td>9</td>
<td>307.6</td>
<td>423.9</td>
<td>1.470</td>
<td>0.507</td>
<td>&gt;</td>
</tr>
<tr>
<td>10</td>
<td>303.9</td>
<td>426.2</td>
<td>1.313</td>
<td>0.472</td>
<td>&gt;</td>
</tr>
<tr>
<td>≈</td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
</tr>
<tr>
<td>99</td>
<td>320.8</td>
<td>441.7</td>
<td>1.448</td>
<td>0.481</td>
<td>&gt;</td>
</tr>
<tr>
<td>100</td>
<td>285.2</td>
<td>393.7</td>
<td>1.583</td>
<td>0.505</td>
<td>&gt;</td>
</tr>
<tr>
<td>Average</td>
<td>307.0</td>
<td>422.7</td>
<td>1.482</td>
<td>0.494</td>
<td>&gt;</td>
</tr>
</tbody>
</table>

NOTE: PE (Prediction Error) is defined as $PE = \frac{1}{24} \sum_{k=1}^{24} \sum_{i \in \mathcal{I}_k} (y_i - x_i^T \hat{\beta}_k)^2$.

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 $\hat{\beta}(\hat{\lambda}$, $\hat{\Delta}$). 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](image-url) - The profile plots of the GCV function for estimating $\Delta$ (left graph) and $\lambda$ (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 $\lambda$, 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.

<table>
<thead>
<tr>
<th>No.</th>
<th>Ordinary Ridge</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Gene symbol</td>
<td>Coefficient</td>
</tr>
<tr>
<td>1</td>
<td>FGA</td>
<td>-0.0381</td>
</tr>
<tr>
<td>2</td>
<td>AKR1B10</td>
<td>-0.0462</td>
</tr>
<tr>
<td>3</td>
<td>CPS1</td>
<td>-0.0411</td>
</tr>
<tr>
<td>4</td>
<td>KRT6A</td>
<td>-0.0345</td>
</tr>
<tr>
<td>5</td>
<td>MSMB</td>
<td>-0.0446</td>
</tr>
<tr>
<td>6</td>
<td>FGG</td>
<td>-0.0337</td>
</tr>
<tr>
<td>7</td>
<td>CYP2B7P1</td>
<td>0.0302</td>
</tr>
<tr>
<td>8</td>
<td>SERPINB5</td>
<td>-0.0290</td>
</tr>
<tr>
<td>9</td>
<td>FGB</td>
<td>-0.0285</td>
</tr>
<tr>
<td>10</td>
<td>CYP2B6</td>
<td>0.0232</td>
</tr>
<tr>
<td>11</td>
<td>LOC344887</td>
<td>-0.0302</td>
</tr>
<tr>
<td>12</td>
<td>SERPINB3</td>
<td>-0.0263</td>
</tr>
<tr>
<td>13</td>
<td>GPR110</td>
<td>0.0310</td>
</tr>
<tr>
<td>14</td>
<td>HSD17B6</td>
<td>0.0260</td>
</tr>
<tr>
<td>15</td>
<td>CRP</td>
<td>-0.0138</td>
</tr>
<tr>
<td>16</td>
<td>DKK1</td>
<td>-0.0389</td>
</tr>
<tr>
<td>17</td>
<td>SLC3A4A2</td>
<td>0.0217</td>
</tr>
<tr>
<td>18</td>
<td>MUC13</td>
<td>-0.0318</td>
</tr>
<tr>
<td>19</td>
<td>CPN1</td>
<td>-0.0083</td>
</tr>
<tr>
<td>20</td>
<td>SLC6A14</td>
<td>-0.0273</td>
</tr>
</tbody>
</table>

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-ketoreductase 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 \( \hat{\beta}_{\text{Train}} \) be either the ordinary ridge

**Figure 8.** The plots of the EGFR index \( y_i \) against its predictor \( (x_i^{\text{Test}})^T \hat{\beta}_{\text{Train}} \). The dash lines (red color) denote the distance between the value \( y_i \) and the predictor \( (x_i^{\text{Test}})^T \hat{\beta}_{\text{Train}} \). The prediction error is \( PE = \sum_{i \in \text{Test}} (y_i - (x_i^{\text{Test}})^T \hat{\beta}_{\text{Train}})^2 / 62 \).
Table 6. The 45 most strongly associated genes from the lung cancer data based on the ordinary ridge and proposed methods.

<table>
<thead>
<tr>
<th>No.</th>
<th>Gene symbol</th>
<th>Coefficient</th>
<th>P-value</th>
<th>Gene symbol</th>
<th>Coefficient</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FGA</td>
<td>-0.0381</td>
<td>2.6×10^{-7}</td>
<td>FGA</td>
<td>-0.0507</td>
<td>3.7×10^{-7}</td>
</tr>
<tr>
<td>2</td>
<td>AKR1B10</td>
<td>-0.0462</td>
<td>7.9×10^{-7}</td>
<td>AKR1B10</td>
<td>-0.0590</td>
<td>1.7×10^{-6}</td>
</tr>
<tr>
<td>3</td>
<td>CPS1</td>
<td>-0.0411</td>
<td>2.6×10^{-5}</td>
<td>CPS1</td>
<td>-0.0562</td>
<td>2.7×10^{-5}</td>
</tr>
<tr>
<td>4</td>
<td>KRT6A</td>
<td>-0.0345</td>
<td>4.5×10^{-5}</td>
<td>FGG</td>
<td>-0.0465</td>
<td>8.1×10^{-5}</td>
</tr>
<tr>
<td>5</td>
<td>MSMB</td>
<td>-0.0446</td>
<td>0.0001</td>
<td>MSMB</td>
<td>-0.0643</td>
<td>0.0001</td>
</tr>
<tr>
<td>6</td>
<td>FGG</td>
<td>-0.0337</td>
<td>0.0001</td>
<td>KRT6A</td>
<td>-0.0445</td>
<td>0.0001</td>
</tr>
<tr>
<td>7</td>
<td>CYP2B7P1</td>
<td>0.0302</td>
<td>0.0002</td>
<td>CYP2B7P1</td>
<td>0.0413</td>
<td>0.0004</td>
</tr>
<tr>
<td>8</td>
<td>SERPINB5</td>
<td>-0.0290</td>
<td>0.0002</td>
<td>FGB</td>
<td>-0.0372</td>
<td>0.0007</td>
</tr>
<tr>
<td>9</td>
<td>FGB</td>
<td>-0.0285</td>
<td>0.0003</td>
<td>CYP2B6</td>
<td>0.0163</td>
<td>0.0009</td>
</tr>
<tr>
<td>10</td>
<td>CYP2B6</td>
<td>0.0232</td>
<td>0.0005</td>
<td>SERPINB5</td>
<td>-0.0339</td>
<td>0.0018</td>
</tr>
<tr>
<td>11</td>
<td>LOC344887</td>
<td>-0.0302</td>
<td>0.0011</td>
<td>GRP110</td>
<td>0.0427</td>
<td>0.0022</td>
</tr>
<tr>
<td>12</td>
<td>SERPINB3</td>
<td>-0.0263</td>
<td>0.0014</td>
<td>LOC344887</td>
<td>-0.0366</td>
<td>0.0023</td>
</tr>
<tr>
<td>13</td>
<td>GPRIR10</td>
<td>0.0310</td>
<td>0.0022</td>
<td>CRP</td>
<td>-0.0206</td>
<td>0.0034</td>
</tr>
<tr>
<td>14</td>
<td>HSD17B6</td>
<td>0.0260</td>
<td>0.0035</td>
<td>SERPINB3</td>
<td>-0.0318</td>
<td>0.0044</td>
</tr>
<tr>
<td>15</td>
<td>CRP</td>
<td>-0.0138</td>
<td>0.0037</td>
<td>SLC6A14</td>
<td>-0.0210</td>
<td>0.0051</td>
</tr>
<tr>
<td>16</td>
<td>DKK1</td>
<td>-0.0389</td>
<td>0.0041</td>
<td>HSD17B6</td>
<td>0.0349</td>
<td>0.0058</td>
</tr>
<tr>
<td>17</td>
<td>SLC3A2</td>
<td>0.0217</td>
<td>0.0043</td>
<td>DKK1</td>
<td>-0.0492</td>
<td>0.0066</td>
</tr>
<tr>
<td>18</td>
<td>MUC13</td>
<td>-0.0318</td>
<td>0.0045</td>
<td>MUC13</td>
<td>-0.0408</td>
<td>0.0077</td>
</tr>
<tr>
<td>19</td>
<td>CPN1</td>
<td>-0.0083</td>
<td>0.0054</td>
<td>CPN1</td>
<td>-0.0115</td>
<td>0.0018</td>
</tr>
<tr>
<td>20</td>
<td>SLC6A14</td>
<td>-0.0273</td>
<td>0.0080</td>
<td>789536a</td>
<td>0.0242</td>
<td>0.0121</td>
</tr>
<tr>
<td>21</td>
<td>7978626a</td>
<td>0.0345</td>
<td>0.0088</td>
<td>CYP2A1</td>
<td>-0.0418</td>
<td>0.0149</td>
</tr>
<tr>
<td>22</td>
<td>GPX2</td>
<td>-0.0229</td>
<td>0.0099</td>
<td>SLC22A10</td>
<td>-0.0157</td>
<td>0.0167</td>
</tr>
<tr>
<td>23</td>
<td>CYP2A1</td>
<td>-0.0322</td>
<td>0.0107</td>
<td>7978626a</td>
<td>0.0423</td>
<td>0.0178</td>
</tr>
<tr>
<td>24</td>
<td>8070755a</td>
<td>0.0258</td>
<td>0.0121</td>
<td>8070755a</td>
<td>0.0178</td>
<td>0.0227</td>
</tr>
<tr>
<td>25</td>
<td>789536a</td>
<td>0.0315</td>
<td>0.0127</td>
<td>SLC3A2</td>
<td>0.0240</td>
<td>0.0231</td>
</tr>
<tr>
<td>26</td>
<td>S100A7</td>
<td>-0.0211</td>
<td>0.0174</td>
<td>AZGP1</td>
<td>0.0153</td>
<td>0.0293</td>
</tr>
<tr>
<td>27</td>
<td>AKR1C2</td>
<td>-0.0207</td>
<td>0.0187</td>
<td>C9</td>
<td>-0.0046</td>
<td>0.0295</td>
</tr>
<tr>
<td>28</td>
<td>C9</td>
<td>-0.0066</td>
<td>0.0206</td>
<td>GPX2</td>
<td>-0.0253</td>
<td>0.0368</td>
</tr>
<tr>
<td>29</td>
<td>SLC22A10</td>
<td>-0.0103</td>
<td>0.0212</td>
<td>8151583a</td>
<td>-0.0132</td>
<td>0.0372</td>
</tr>
<tr>
<td>30</td>
<td>SERPINB4</td>
<td>-0.0227</td>
<td>0.0221</td>
<td>BPIFA1</td>
<td>-0.0158</td>
<td>0.0392</td>
</tr>
<tr>
<td>31</td>
<td>VSIG1</td>
<td>0.0252</td>
<td>0.0234</td>
<td>MMP7</td>
<td>0.0800</td>
<td>0.0403</td>
</tr>
<tr>
<td>32</td>
<td>IGJ</td>
<td>0.0191</td>
<td>0.0234</td>
<td>SERPINB4</td>
<td>-0.0275</td>
<td>0.0442</td>
</tr>
<tr>
<td>33</td>
<td>AKR1C1</td>
<td>-0.0182</td>
<td>0.0247</td>
<td>GYS2</td>
<td>0.0046</td>
<td>0.0447</td>
</tr>
<tr>
<td>34</td>
<td>FGL1</td>
<td>-0.0183</td>
<td>0.0303</td>
<td>FGL1</td>
<td>-0.0124</td>
<td>0.0452</td>
</tr>
<tr>
<td>35</td>
<td>AZGP1</td>
<td>0.0200</td>
<td>0.0333</td>
<td>MMP13</td>
<td>0.0188</td>
<td>0.0478</td>
</tr>
<tr>
<td>36</td>
<td>MMP10</td>
<td>-0.0189</td>
<td>0.0365</td>
<td>VSIG1</td>
<td>0.0302</td>
<td>0.0480</td>
</tr>
<tr>
<td>37</td>
<td>C4BPA</td>
<td>0.0163</td>
<td>0.0391</td>
<td>GLYATL1</td>
<td>0.0068</td>
<td>0.0498</td>
</tr>
<tr>
<td>38</td>
<td>BPIFA1</td>
<td>-0.0206</td>
<td>0.0412</td>
<td>AKR1C2</td>
<td>-0.0226</td>
<td>0.0501</td>
</tr>
<tr>
<td>39</td>
<td>APCS</td>
<td>-0.0050</td>
<td>0.0446</td>
<td>S100A7</td>
<td>-0.0246</td>
<td>0.0509</td>
</tr>
<tr>
<td>40</td>
<td>8128714a</td>
<td>-0.0195</td>
<td>0.0464</td>
<td>MMP10</td>
<td>-0.0246</td>
<td>0.0557</td>
</tr>
<tr>
<td>41</td>
<td>PLG</td>
<td>-0.0045</td>
<td>0.0453</td>
<td>PLG</td>
<td>-0.0033</td>
<td>0.0571</td>
</tr>
<tr>
<td>42</td>
<td>MMP7</td>
<td>0.0228</td>
<td>0.0465</td>
<td>ZBBX</td>
<td>-0.0121</td>
<td>0.0571</td>
</tr>
<tr>
<td>43</td>
<td>C4BPB</td>
<td>0.0153</td>
<td>0.0481</td>
<td>7892787a</td>
<td>-0.0185</td>
<td>0.0575</td>
</tr>
<tr>
<td>44</td>
<td>MFAP5</td>
<td>-0.0193</td>
<td>0.0482</td>
<td>C4BPB</td>
<td>0.0019</td>
<td>0.0600</td>
</tr>
<tr>
<td>45</td>
<td>GYS2</td>
<td>0.0060</td>
<td>0.0492</td>
<td>SPANXB2</td>
<td>-0.0208</td>
<td>0.0608</td>
</tr>
</tbody>
</table>

\[ PE_{i} = 0.5846 \quad PE = 0.5420 \]

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 \( \mathbf{x}_{i}^{\text{Test}} = (x_{i(1)}, \ldots, x_{i(20)})^{T} \) be the corresponding 20 gene signatures in the testing patients \( (i \in \text{Test}) \). Then we plot the EGFR index \( y_{i} \) against its predictor \( (\mathbf{x}_{i}^{\text{Test}})^{T} \beta^{\text{Train}} \) for all \( i \in \text{Test} \) (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

\[ PE = \frac{1}{62} \sum_{i \in \text{Test}} \left( y_i - (x_{i \text{Test}}^T \hat{\beta}^{\text{Train}}) \right)^2. \]

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 \( j \)th 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 par-
ticular, 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 frame-
work 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
\( \beta_j \), 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 sim-
ilar 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


