STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **6**, September 2018, pp 415–426.
Published online in International Academic Press (www.IAPress.org)

Generalized Ridge Regression Estimator in High Dimensional Sparse Regression Models

Mahdi Roozbeh

Department of Statistics, Semnan University, Iran

Abstract Modern statistical analysis often encounters linear models with the number of explanatory variables much larger than the sample size. Estimation in these high-dimensional problems needs some regularization methods to be employed due to rank deficiency of the design matrix. In this paper, the ridge estimators are considered and their restricted regression counterparts are proposed when the errors are dependent under a multicollinearity and high-dimensionality setting. The asymptotic distributions of the proposed estimators are exactly derived. Incorporating the information contained in the restricted estimator, a shrinkage type ridge estimator is also exhibited and its asymptotic risk is analyzed under some special cases. To evaluate the efficiency of the proposed estimators, a Monté-Carlo simulation along with a real example are considered.

Keywords Asymptotic distribution; High-dimension; Ridge regression; Shrinkage estimator; Tikhonov regularization

AMS 2010 subject classifications Primary: 62G08, Secondary: 62J07

DOI: 10.19139/soic.v6i3.581

1. Introduction

Consider a linear regression model:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i,\ i = 1, 2, \ldots, n, \tag{1.1}$$

where y_is are responses, $x_i = (x_{i1}, \ldots, x_{ip})^\top$ is design points, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is vector denoting unknown coefficients, ε_is are unobservable random errors and the superscript $(^\top)$ denotes the transpose of a vector or matrix. Further, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ has a cumulative distribution function $F(\varepsilon)$; $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{V}_n$, where σ^2 is finite and $\boldsymbol{V}_n$ is a known matrix belonging to the space of all positive definite matrices of dimension $n \times n$, denoted by $\mathcal{S}(n)$.

Now-a-day, many data problems nowadays carry the structure that the number of covariates p may exceed sample size n, known as small n, large p problems. Such cases, that can be seen in the studies of genomics, financial markets, mobile phone communication, bioinformatics and risk management, regularization methods should be considered for inferring (see [9, 5]) about parameters of interest.

One of the mostly used regularization methods is the Tikhonov [21] regularization which was brought into statistical contexts by Hoerl and Kennard [12] as ridge regression. To mention a few recent researches, see e.g., [1, 2, 3, 4, 14, 15, 16].

In this paper, when a subset of coefficients is zero, the underlying model is called sparse. Under the sparsity assumption, the vector of coefficients $\boldsymbol{\beta}$ can be partitioned as $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ where $\boldsymbol{\beta}_1$ is the coefficient vector for main effects and $\boldsymbol{\beta}_2$ is the vector for nuisance effects or insignificant coefficients. We are essentially interested in the estimation of $\boldsymbol{\beta}_1$ when it is reasonable that $\boldsymbol{\beta}_2$ is close to zero.

The paper is organized as follows: The problem of interest is stated in section 2. In section 3, sparse ridge model is considered by proposing the new estimators, while their asymptotic properties are derived in section 4. Section 5 is devoted to exhibiting a shrinkage type ridge estimator and analyzing its asymptotic properties. Efficiencies of the proposed shrinkage estimators relative to the ordinary estimator are evaluated through a

*Correspondence to: (Email: mahdi.roozbeh@semnan.ac.ir). Department of Statistics, Faculty of Mathematics, Statistics and Computer, Semnan University, Semnan, Iran , P.O. Box 35195-363.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

Monté-Carlo simulation as well as a real data example in section 6. We conclude our study in section 7 by giving a summary.

2. Generalized Ridge Estimator

blueIn this section, we discuss a biased estimation technique under multicollinearity for the regression models. So, the following preliminaries are needed. In a full matrix notation, the model (1.1) can be represented as

$$\boldsymbol{y}_n = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.2}$$

where $\boldsymbol{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ and $\boldsymbol{y}_n = (y_1, \ldots, y_n)^\top$.

As known, both of the ordinary least squares estimator (OLSE) and its covariance matrix are heavily dependent to the characteristics of the matrix $\boldsymbol{S}_n = \boldsymbol{X}_n^\top \boldsymbol{X}_n$. If $\boldsymbol{S}_n$ is ill-conditioned, then the OLSE may affect by various numerical errors. The problem of multicollinearity can be solved by collecting additional data, reparameterizing the model and reselecting the variables. There are two well-known mathematical methods to overcome multicollinearity: the 'principal components regression method' and the 'ridge regression method'. Here, we discuss the ridge regression method.

It is known that spectral decomposition of the (symmetric) positive definite matrix $\boldsymbol{S}_n$ can be given by

$$\boldsymbol{S}_n = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\top, \ \boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p), \tag{2.3}$$

where the columns of $\boldsymbol{\Gamma}$ are eigenvectors of the matrix $\boldsymbol{S}_n$ as well as the scalers λ_1,..., λ_p are its eigenvalues, satisfying

$$\lambda_1 \geq \ldots \geq \lambda_p > 0,$$

without loss of generality. Therefore, the orthogonal (canonical) version of the model (2.2) is given by

$$\boldsymbol{y}_n = \boldsymbol{X}_n^*\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \tag{2.4}$$

where $\boldsymbol{X}_n^* = \boldsymbol{X}_n\boldsymbol{\Gamma}$ and $\boldsymbol{\alpha} = \boldsymbol{\Gamma}^\top \boldsymbol{e}$. However, when $\boldsymbol{S}_n$ is ill-conditioned, there exists an approximate linear dependency among its columns. So, the OLSE of $\boldsymbol{\beta}$ has a large variance and multicollinearity is said to be appeared. In such situation, there exists a small positive constant ε such that

$$\lambda_1 \geq \ldots \geq \lambda_r \gg \varepsilon \geq \lambda_{r+1} \geq \ldots \geq \lambda_p > 0, \tag{2.5}$$

where r is called the numerical rank of $\boldsymbol{S}_n$ (Watkins, 2002). More closeness of the small eigenvalues to the origin leads to more strength of the multicollinearity. To overcome damaging effect of the small eigenvalues, we need to increase them.

In another point of view based on the spectral decomposition, in the ridge regression the matrix $\boldsymbol{S}_n$ is replaced by the matrix $\boldsymbol{S}_n(k)$ defined by

$$\boldsymbol{S}_n(k) = \boldsymbol{\Gamma}\boldsymbol{\Lambda}(k)\boldsymbol{\Gamma}^\top, \ \boldsymbol{\Lambda}(k) = \text{diag}(\lambda_1 + k, \ldots, \lambda_p + k). \tag{2.6}$$

Hence, we have

$$\left\|\boldsymbol{S}_n(k) - \boldsymbol{S}_n\right\|_2 = k,$$

and,

$$\kappa_2(\boldsymbol{S}_n(k)) = \frac{\lambda_1 + k}{\lambda_p + k} \leq \kappa_2(\boldsymbol{S}_n) = \frac{\lambda_1}{\lambda_p},$$

where $\kappa_2(.)$ stands for the spectral condition number and $\|\boldsymbol{x}\|_2$ denotes the Euclidean norm of the vector $\boldsymbol{x}$. Thus, computations of the ridge regression are numerically more stable than the OLSE.

In an extension scheme, the canonical generalized ridge estimator has been proposed by Hoerl and Kennard [12] as follows:

$$\hat{\boldsymbol{\alpha}}\left(\boldsymbol{K}_n^{(p)}\right) = \left(\boldsymbol{S}_n^* + \boldsymbol{K}_n^{(p)}\right)^{-1} \boldsymbol{X}_n^{*\top}\boldsymbol{y}_n = \boldsymbol{T}_n^*\left(\boldsymbol{K}_n^{(p)}\right)\hat{\boldsymbol{\alpha}}, \ \boldsymbol{T}_n^*\left(\boldsymbol{K}_n^{(p)}\right) = \left(\boldsymbol{K}_n^{(p)}\boldsymbol{S}_n^{*-1} + \boldsymbol{I}_p\right)^{-1}, \tag{2.7}$$

for some suitably chosen diagonal matrix $\boldsymbol{K}_n^{(p)} = \text{diag}\left(k_n^{(1)}, \ldots, k_n^{(p)}\right)$ of tuning parameters, with $k_n^{(i)} > 0$, i=1,...,p, $\boldsymbol{S}_n^* = \boldsymbol{X}_n^{*\top}\boldsymbol{X}_n^*$, and $\hat{\boldsymbol{\alpha}} = \boldsymbol{\Lambda}^{-1}\boldsymbol{X}_n^{*\top}\boldsymbol{y}_n$ is the canonical ordinary least-squares estimate of $\boldsymbol{\alpha}$. Hoerl

and Kennard [12] showed that for the known optimal values

$$k_n^{(i)} = \frac{\sigma^2}{\alpha_i^2},\ i = 1, \ldots, p, \tag{2.8}$$

the generalized ridge regression estimator is superior to all of the other ones within the class of biased estimators, where σ^2 is the error variance of the model (2.2) and α_i^2 is the i^{th} entry of $\boldsymbol{\alpha}$. Moreover, Hocking et al. [13] showed that with these optimal $k_n^{(i)}$, values the proposed estimator (2.7) is superior to all estimators within the class of biased estimators they considered. However, the optimal value of $k_n^{(i)}$ fully depends on the unknown values σ^2 and α_i which must be estimated from the observed data. Hoerl and Kennard [12] suggested to replace σ^2 and α_i^2 by their corresponding unbiased estimators. That is,

$$\hat{k}_n^{(i)} = \frac{\hat{\sigma}^2}{\hat{\alpha_i}^2},\ i = 1, \ldots, p, \tag{2.9}$$

where $\hat{\sigma}^2$ is an unbiased and effective estimator of σ^2 and $\hat{\alpha}_i$ is the i^{th} entry of $\hat{\boldsymbol{\alpha}}$ which is an unbiased estimator of $\boldsymbol{\alpha}$. Hoerl and Kennard [12] also suggested an iterative procedure to estimate $k_n^{(i)}$. Two other methods to select $k_n^{(i)}$ have been proposed by Hemmerle and Brantle [10]. Some exact finite sample properties for (2.7) can be found in Hemmerle and Carey [11]. Thus, the matrix $\boldsymbol{S}_n$ is replaced by the matrix $\boldsymbol{S}_n\left(\hat{\boldsymbol{K}}_n^{(p)}\right)$ defined by

$$\boldsymbol{S}_n\left(\hat{\boldsymbol{K}}_n^{(p)}\right) = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\left(\hat{\boldsymbol{K}}_n^{(p)}\right)\boldsymbol{\Gamma}^\top,\ \boldsymbol{\Lambda}\left(\hat{\boldsymbol{K}}_n^{(p)}\right) = \text{diag}(\lambda_1 + \hat{k}_n^{(1)}, \ldots, \lambda_p + \hat{k}_n^{(p)}). \tag{2.10}$$

Hence, we have

$$\left\|\boldsymbol{S}_n\left(\hat{\boldsymbol{K}}_n^{(p)}\right) - \boldsymbol{S}_n\right\|_2 = \max_{i=1,\ldots,p}\{\hat{k}_n^{(i)}\}, \tag{2.11}$$

and,

$$\kappa_2\left(\boldsymbol{S}\left(\hat{\boldsymbol{K}}_n^{(p)}\right)\right) = \frac{\max\limits_{i=1,\ldots,p}\{\lambda_i + \hat{k}_n^{(i)}\}}{\min\limits_{i=1,\ldots,p}\{\lambda_i + \hat{k}_n^{(i)}\}}.$$

Now, considering the high-dimensional case $p > n$, we are primary interested in estimating the regression vector-parameter $\boldsymbol{\beta}$ in the model (2.2). Since $\boldsymbol{S}_n$ is rank deficient, a regularization method is needed to combat this ill-conditioning, such as Tikhonov [21] regularization. In our setup, estimating $\boldsymbol{\beta}$ is equal to minimizing the following general criterion, under the L_2 norm

$$\left\|\boldsymbol{V}_n^{-\frac{1}{2}}(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta})\right\|_2^2 + \left\|\boldsymbol{K}_n^{(p)\frac{1}{2}}\boldsymbol{\beta}\right\|_2^2$$

Let $F(\boldsymbol{\beta}) = \left\{\left\|\boldsymbol{V}_n^{-\frac{1}{2}}(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta})\right\|_2^2 + \left\|\boldsymbol{K}_n^{(p)\frac{1}{2}}\boldsymbol{\beta}\right\|_2^2\right\}$. Then

$$\hat{\boldsymbol{\beta}}\left(\boldsymbol{K}_n^{(p)}\right) \quad = \quad \text{argmin}_{\boldsymbol{\beta}}F(\boldsymbol{\beta}) = \left(\boldsymbol{X}_n^\top\boldsymbol{V}_n^{-1}\boldsymbol{X}_n + \boldsymbol{K}_n^{(p)}\right)^{-1}\boldsymbol{X}_n^\top\boldsymbol{V}_n^{-1}\boldsymbol{y}_n, \tag{2.12}$$

where $\hat{\boldsymbol{\beta}}\left(\boldsymbol{K}_n^{(p)}\right)$ is called generalized ridge estimator (GRE).

For the case $\boldsymbol{K}_n^{(p)} = k\boldsymbol{I}_p$, $k > 0$, the GRE reduces to the ordinary ridge regression estimator introduced by Hoerl and Kennard [12]. If $k_n^{(i)} \to 0$, $i = 1, \ldots, p$, then $\hat{\boldsymbol{\beta}}(\boldsymbol{K}_n^{(p)})$ reduces to the well-known ordinary least square (GLS) estimator $\hat{\boldsymbol{\beta}}_n = \left(\boldsymbol{X}_n^\top\boldsymbol{V}_n^{-1}\boldsymbol{X}_n\right)^{-1}\boldsymbol{X}_n^\top\boldsymbol{V}_n^{-1}\boldsymbol{y}_n$.

In what follows, we discuss about the properties of $\hat{\boldsymbol{\beta}}(\boldsymbol{K}_n^{(p)})$.

Since $\hat{\boldsymbol{\beta}}(\boldsymbol{K}_n^{(p)}) = \text{argmin}F(\boldsymbol{\beta})$, $F(\hat{\boldsymbol{\beta}}(\boldsymbol{K}_n^{(p)})) \leq F(\boldsymbol{0})$. Therefore,

$$\begin{aligned}\left\|\boldsymbol{K}_n^{(p)\frac{1}{2}}\hat{\boldsymbol{\beta}}\right\|_2^2 &\leq \left\|\boldsymbol{V}_n^{-\frac{1}{2}}(\boldsymbol{y}_n - \boldsymbol{X}_n\hat{\boldsymbol{\beta}})\right\|_2^2 + \left\|\boldsymbol{K}_n^{(p)\frac{1}{2}}\hat{\boldsymbol{\beta}}\right\|_2^2 \\ &= F(\hat{\boldsymbol{\beta}}(\boldsymbol{K}_n^{(p)}))\end{aligned}$$

$$\leq \quad F(\mathbf{0}) = \left\| \boldsymbol{V}_n^{-\frac{1}{2}} \boldsymbol{y}_n \right\|_2^2$$

Thus, for $k_1 = \min(k_n^{(i)})$, we get

$$\left\| \hat{\boldsymbol{\beta}} \right\|_2^2 \leq k_1^{-1} \left\| \boldsymbol{V}_n^{-\frac{1}{2}} \boldsymbol{y}_n \right\|_2^2, \quad \text{and} \quad E \left\| \hat{\boldsymbol{\beta}} \right\|_2^2 \leq k_1^{-1} E \left\| \boldsymbol{V}_n^{-\frac{1}{2}} \boldsymbol{y}_n \right\|_2^2 \tag{2.13}$$

Now, we focus on the covariance. For the high-dimensional case $p > n$, using spectral decomposition

$$\boldsymbol{X}_n^\top \boldsymbol{V}_n^{-1} \boldsymbol{X}_n = \boldsymbol{\Gamma} \begin{bmatrix} \boldsymbol{\Lambda}_{n\times n} & \boldsymbol{O}_{n\times(p-n)} \\ \boldsymbol{O}_{n\times(p-n)}^\top & \boldsymbol{O}_{(p-n)\times(p-n)} \end{bmatrix} \boldsymbol{\Gamma}^\top, \ \boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \tag{2.14}$$

where $\lambda_1 \geq \ldots \geq \lambda_n > 0$.

Theorem 1
Let $k_n = \max_{1\leq i\leq n} \left(k_n^{(i)}\right)$ and $k_o = \sum_{j=1}^n \left(1/k_n^{(j)}\right)^2$. Assume $\lambda_j = o(k_n)$, for $j = 1, \ldots, n$. Then, we have

$$\lim_{p\to\infty} \operatorname{tr} \left(\operatorname{Cov}(\hat{\boldsymbol{\beta}}(\boldsymbol{K}_n^{(p)}) - \boldsymbol{\beta}) \right) = \sigma^2 k_o \operatorname{tr} \left(\boldsymbol{X}_n^\top \boldsymbol{V}_n^{-1} \boldsymbol{X}_n \right).$$

Proof: By definition, after some algebra,

$$\begin{aligned} \operatorname{Cov}(\hat{\boldsymbol{\beta}}(\boldsymbol{K}_n^{(p)}) - \boldsymbol{\beta}) &= \sigma^2 \left(\boldsymbol{X}_n^\top \boldsymbol{V}_n^{-1} \boldsymbol{X}_n + \boldsymbol{K}_n^{(p)} \right)^{-1} \boldsymbol{X}_n^\top \boldsymbol{V}_n^{-1} \boldsymbol{X}_n^\top \left(\boldsymbol{X}_n^\top \boldsymbol{V}_n^{-1} \boldsymbol{X}_n + \boldsymbol{K}_n^{(p)} \right)^{-1} \\ &= \sigma^2 \left(\boldsymbol{X}_n^\top \boldsymbol{V}_n^{-1} \boldsymbol{X}_n + \boldsymbol{K}_n^{(p)} \right)^{-1} - \sigma^2 \left(\boldsymbol{X}_n^\top \boldsymbol{V}_n^{-1} \boldsymbol{X}_n + \boldsymbol{K}_n^{(p)} \right)^{-1} \\ &\quad \times \boldsymbol{K}_n^{(p)} \left(\boldsymbol{X}_n^\top \boldsymbol{V}_n^{-1} \boldsymbol{X}_n + \boldsymbol{K}_n^{(p)} \right)^{-1} \\ &= \sigma^2 \begin{bmatrix} \frac{\lambda_1}{\left(\lambda_1 + k_n^{(1)}\right)^2} & & & \\ & \ddots & & \boldsymbol{O}_{n\times(p-n)} \\ & & \frac{\lambda_n}{\left(\lambda_n + k_n^{(n)}\right)^2} & \\ & \boldsymbol{O}_{n\times(p-n)}^\top & & \boldsymbol{O}_{(p-n)\times(p-n)} \end{bmatrix} \end{aligned}$$

since

$$\begin{aligned} \left(\boldsymbol{X}_n^\top \boldsymbol{V}_n^{-1} \boldsymbol{X}_n + \boldsymbol{K}_n^{(p)} \right)^{-1} &= \left\{ \boldsymbol{\Gamma} \begin{bmatrix} \boldsymbol{\Lambda}_{n\times n} & \boldsymbol{O}_{n\times(p-n)} \\ \boldsymbol{O}_{n\times(p-n)}^\top & \boldsymbol{O}_{(p-n)\times(p-n)} \end{bmatrix} \boldsymbol{\Gamma}^\top + \boldsymbol{K}_n^{(p)} \right\}^{-1} \\ &= \begin{bmatrix} \lambda_1 + k_n^{(1)} & & & & \\ & \ddots & & \boldsymbol{O}_{n\times(p-n)} & \\ & & \lambda_n + k_n^{(n)} & & \\ & & & k_n^{(n+1)} & \\ & \boldsymbol{O}_{n\times(p-n)}^\top & & & \ddots \\ & & & & & k_n^{(p)} \end{bmatrix}^{-1} \end{aligned} \tag{2.15}$$

Hence, $\operatorname{tr} \operatorname{Cov}(\hat{\boldsymbol{\beta}}(\boldsymbol{K}_n^{(p)}) - \boldsymbol{\beta}) = \sigma^2 \sum_{j=1}^n \frac{\lambda_j}{\left(\lambda_j + k_n^{(j)}\right)^2}$. Under the assumption $\lambda_j = o(k_n)$, for $j = 1, \ldots, n$ we have

$$\lim_{p\to\infty} \sum_{j=1}^n \frac{1}{\left(\lambda_j + k_n^{(j)}\right)^2} = \sum_{j=1}^n \left(\frac{1}{k_n^{(j)}} \right)^2 = k_o. \tag{2.16}$$

So, the proof is complete. □

Under some mild conditions, it can be shown that $\frac{1}{\sigma^2 k_o} (\hat{\boldsymbol{\beta}}(\boldsymbol{K}_n^{(p)}) - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \operatorname{diag} \boldsymbol{X}_n^\top \boldsymbol{V}_n^{-1} \boldsymbol{X}_n)$. For more detail on high-dimensional properties, we refer to [7].

3. Sub-model Approach

Since $p > n$, one methodology to infer about regression parameters is to reduce the number of features from p to $p_1 < n$. Hence, the regression parameter $\boldsymbol{\beta}$ is partitioned as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$, where the sub-vector $\boldsymbol{\beta}_i$ has dimension p_i, $i = 1, 2$ and $p_1 + p_2 = p$. Thus, the underlying model has form

$$\boldsymbol{y}_n = \boldsymbol{X}_n^{(1)}\boldsymbol{\beta}_1 + \boldsymbol{X}_n^{(2)}\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\boldsymbol{X}_n$ is partitioned to $(\boldsymbol{X}_n^{(1)}, \boldsymbol{X}_n^{(2)})$ in such a way that $\boldsymbol{X}_n^{(i)}$ is a $n \times p_i$ sub-matrix, $i = 1, 2$. With respect to this partitioning, the generalized least square estimators (GOLSEs) of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are respectively given by

$$\hat{\boldsymbol{\beta}}_n^{(1)} = \boldsymbol{C}_n^{(1)^{-1}} \boldsymbol{X}_n^{(1)^\top} \boldsymbol{\Sigma}_n^{(2)^{-1}} \boldsymbol{y}_n, \quad \boldsymbol{C}_n^{(1)} = \boldsymbol{X}_n^{(1)^\top} \boldsymbol{\Sigma}_n^{(2)^{-1}} \boldsymbol{X}_n^{(1)} \tag{3.2}$$

$$\hat{\boldsymbol{\beta}}_n^{(2)} = \boldsymbol{C}_n^{(2)^{-1}} \boldsymbol{X}_n^{(2)^\top} \boldsymbol{\Sigma}_n^{(1)^{-1}} \boldsymbol{y}_n, \quad \boldsymbol{C}_n^{(2)} = \boldsymbol{X}_n^{(1)^\top} \boldsymbol{\Sigma}_n^{(1)^{-1}} \boldsymbol{X}_n^{(1)} \tag{3.3}$$

where

$$\boldsymbol{\Sigma}_n^{(i)^{-1}} = \boldsymbol{V}_n^{-1} - \boldsymbol{V}_n^{-1}\boldsymbol{X}_n^{(i)} \left(\boldsymbol{X}_n^{(i)^\top} \boldsymbol{V}_n^{-1} \boldsymbol{X}_n^{(i)}\right)^{-1} \boldsymbol{X}_n^{(i)^\top} \boldsymbol{V}_n^{-1}, \quad i = 1, 2.$$

The sparse model is defined when $\mathcal{H}_o : \boldsymbol{\beta}_2 = 0$ is true. In this paper, we refer restricted regression model (RRM) to the sparse model.

For the RRM, the generalized restricted estimator based on OLS estimator (GROLSE) has form

$$\hat{\boldsymbol{\beta}}_n^{R1} = \left(\boldsymbol{X}_n^{(1)^\top} \boldsymbol{V}_n^{-1} \boldsymbol{X}_n^{(1)}\right)^{-1} \boldsymbol{X}_n^{(1)^\top} \boldsymbol{V}_n^{-1} \boldsymbol{y}_n. \tag{3.4}$$

According to [20], the GROLSE performs better than GOLSE when model is sparse. However, the former estimator performs poorly as $\boldsymbol{\beta}_2$ is different from zero. In the following result, in a similar fashion as in [22], the relation between the sub-model and full-model estimators of $\boldsymbol{\beta}_1$ is obtained. It can be easily shown that

$$\hat{\boldsymbol{\beta}}_n^{(1)} = \hat{\boldsymbol{\beta}}_n^{R1} - \left(\boldsymbol{X}_n^{(1)^\top} \boldsymbol{V}_n^{-1} \boldsymbol{X}_n^{(1)}\right)^{-1} \boldsymbol{X}_n^{(1)^\top} \boldsymbol{V}_n^{-1} \boldsymbol{X}_n^{(2)} \hat{\boldsymbol{\beta}}_n^{(2)}. \tag{3.5}$$

3.1. Sparse ridge model

Under the sparsity assumption, i.e., $\boldsymbol{\beta}_2 = 0$ in the high-dimensional problem, following [19], the generalized restricted ridge estimator (GRRE), is given by

$$\begin{aligned} \hat{\boldsymbol{\beta}}_n^{R1}\left(\boldsymbol{K}_n^{(p_1)}\right) &= \left(\boldsymbol{X}_n^{(1)^\top} \boldsymbol{V}_n^{-1} \boldsymbol{X}_n^{(1)} + \boldsymbol{K}_n^{(p_1)}\right)^{-1} \boldsymbol{X}_n^{(1)^\top} \boldsymbol{V}_n^{-1} \boldsymbol{y}_n \\ &= \left(\boldsymbol{I}_{p_1} + \left(\boldsymbol{X}_n^{(1)^\top} \boldsymbol{V}_n^{-1} \boldsymbol{X}_n^{(1)}\right)^{-1} \boldsymbol{K}_n^{(p_1)}\right)^{-1} \hat{\boldsymbol{\beta}}_n^{R1} \\ &= \boldsymbol{T}_1\left(\boldsymbol{K}_n^{(p_1)}\right) \hat{\boldsymbol{\beta}}_n^{R1}, \end{aligned} \tag{3.6}$$

where $\boldsymbol{K}_n^{(p_1)} = \operatorname{diag}\left(k_n^{(1)}, \ldots, k_n^{(p_1)}\right)$ is the ridge parameter matrix as a function of sample size n and $\boldsymbol{T}_1\left(\boldsymbol{K}_n^{(p_1)}\right) = \left(\boldsymbol{I}_{p_1} + \left(\boldsymbol{X}_n^{(1)^\top} \boldsymbol{V}_n^{-1} \boldsymbol{X}_n^{(1)}\right)^{-1} \boldsymbol{K}_n^{(p_1)}\right)^{-1}$.

Similarly, the generalized ridge estimators (GREs) of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ respectively have forms

$$\begin{aligned} \hat{\boldsymbol{\beta}}_n^{(1)}\left(\boldsymbol{K}_n^{(p_1)}\right) &= \left(\boldsymbol{X}_n^{(1)^\top} \boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{K}_n^{(p_2)}\right) \boldsymbol{X}_n^{(1)} + \boldsymbol{K}_n^{(p_1)}\right)^{-1} \boldsymbol{X}_n^{(1)^\top} \boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{K}_n^{(p_2)}\right) \boldsymbol{y}_n \\ &= \left(\boldsymbol{I}_{p_1} + \left(\boldsymbol{X}_n^{(1)^\top} \boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{K}_n^{(p_2)}\right) \boldsymbol{X}_n^{(1)}\right)^{-1} \boldsymbol{K}_n^{(p_1)}\right)^{-1} \hat{\boldsymbol{\beta}}_n^{(1)} \\ &= \boldsymbol{R}_1\left(\boldsymbol{K}_n^{(p_1)}\right) \hat{\boldsymbol{\beta}}_n^{(1)}, \end{aligned} \tag{3.7}$$

$$\begin{aligned} \hat{\boldsymbol{\beta}}_n^{(2)}\left(\boldsymbol{K}_n^{(p_2)}\right) &= \left(\boldsymbol{X}_n^{(2)^\top} \boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{K}_n^{(p_1)}\right) \boldsymbol{X}_n^{(2)} + \boldsymbol{K}_n^{(p_2)}\right)^{-1} \boldsymbol{X}_n^{(2)^\top} \boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{K}_n^{(p_1)}\right) \boldsymbol{y}_n \\ &= \left(\boldsymbol{I}_{p_2} + \left(\boldsymbol{X}_n^{(2)^\top} \boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{K}_n^{(p_1)}\right) \boldsymbol{X}_n^{(2)}\right)^{-1} \boldsymbol{K}_n^{(p_2)}\right)^{-1} \hat{\boldsymbol{\beta}}_n^{(2)} \end{aligned}$$

$$= \quad \boldsymbol{R}_2\left(\boldsymbol{K}_n^{(p_2)}\right)\hat{\boldsymbol{\beta}}_n^{(2)}, \tag{3.8}$$

where

$$\boldsymbol{R}_1\left(\boldsymbol{K}_n^{(p_1)}\right) = \left(\boldsymbol{I}_{p_1} + \left(\boldsymbol{X}_n^{(1)\top}\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{K}_n^{(p_2)}\right)\boldsymbol{X}_n^{(1)}\right)^{-1}\boldsymbol{K}_n^{(p_1)}\right)^{-1},$$

$$\boldsymbol{R}_2\left(\boldsymbol{K}_n^{(p_2)}\right) = \left(\boldsymbol{I}_{p_2} + \left(\boldsymbol{X}_n^{(2)\top}\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{K}_n^{(p_1)}\right)\boldsymbol{X}_n^{(2)}\right)^{-1}\boldsymbol{K}_n^{(p_2)}\right)^{-1},$$

$$\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{K}_n^{(p_i)}\right) = \boldsymbol{V}_n^{-1} - \boldsymbol{V}_n^{-1}\boldsymbol{X}_n^{(i)}\left(\boldsymbol{X}_n^{(i)\top}\boldsymbol{V}_n^{-1}\boldsymbol{X}_n^{(i)} + \boldsymbol{K}_n^{(p_i)}\right)^{-1}\boldsymbol{X}_n^{(i)\top}\boldsymbol{V}_n^{-1}, \quad i = 1, 2. \tag{3.9}$$

It is easy to show that

$$\hat{\boldsymbol{\beta}}_n^{(1)}\left(\boldsymbol{K}_n^{(p_1)}\right) = \hat{\boldsymbol{\beta}}_n^{R1}\left(\boldsymbol{K}_n^{(p_1)}\right) - \left(\boldsymbol{X}_n^{(1)T}\boldsymbol{V}_n^{-1}\boldsymbol{X}_n^{(1)}\right)^{-1}\boldsymbol{X}_n^{(1)T}\boldsymbol{V}_n^{-1}\boldsymbol{X}_n^{(2)}\hat{\boldsymbol{\beta}}_n^{(2)}\left(\boldsymbol{K}_n^{(p_2)}\right). \tag{3.10}$$

4. Asymptotics

In this section we study the asymptotic performance of the sub-model estimator when $p_1 < n$. For the purpose of this paper, we need to take the following assumptions as regularity conditions:

(A1) $\max_{1\le i\le n}\boldsymbol{x}_i^\top\left(\boldsymbol{X}_n^\top\boldsymbol{V}_n^{-1}\boldsymbol{X}_n + \boldsymbol{K}_n^{(p)}\right)^{-1}\boldsymbol{x}_i = o(n)$, where $\boldsymbol{x}_i^\top$ is the i^{th} row of $\boldsymbol{X}_n$.
(A2) $\frac{k_n^{(i)}}{n} \to k_o$ as $n \to \infty$, $i = 1, \ldots, p$.
(A3) Let

$$\begin{aligned}\boldsymbol{A}_n &= \boldsymbol{X}_n{}^\top\boldsymbol{V}_n^{-1}\boldsymbol{X}_n + \boldsymbol{K}_n^{(p)} = \begin{pmatrix}\boldsymbol{X}_n^{(1)\top}\\ \boldsymbol{X}_n^{(2)\top}\end{pmatrix}\boldsymbol{V}_n^{-1}\left(\boldsymbol{X}_n^{(1)}\ \boldsymbol{X}_n^{(2)}\right) + \boldsymbol{K}_n^{(p)}\\ &= \begin{pmatrix}\boldsymbol{X}_n^{(1)\top}\boldsymbol{V}_n^{-1}\boldsymbol{X}_n^{(1)} & \boldsymbol{X}_n^{(1)\top}\boldsymbol{V}_n^{-1}\boldsymbol{X}_n^{(2)}\\ \boldsymbol{X}_n^{(2)\top}\boldsymbol{V}_n^{-1}\boldsymbol{X}_n^{(1)} & \boldsymbol{X}_n^{(2)\top}\boldsymbol{V}_n^{-1}\boldsymbol{X}_n^{(2)}\end{pmatrix} + \begin{pmatrix}\boldsymbol{K}_n^{(p_1)} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{K}_n^{(p_2)}\end{pmatrix}\\ &= \begin{pmatrix}\boldsymbol{A}_{n11} & \boldsymbol{A}_{n12}\\ \boldsymbol{A}_{n21} & \boldsymbol{A}_{n22}\end{pmatrix}.\end{aligned}$$

Then, there exists a positive definite matrix $\boldsymbol{A}$ such that $\frac{1}{n}\boldsymbol{A}_n \to \boldsymbol{A} = \begin{pmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22}\end{pmatrix}$, as $n \to \infty$.

(A4) $\boldsymbol{X}_n^{(2)\top}\boldsymbol{V}_n^{-1}\boldsymbol{X}_n^{(2)} + \boldsymbol{K}_n^{(p_2)} = o(\sqrt{n})$.
(A5) $\boldsymbol{x}_i^\top\boldsymbol{x}_j = o(\sqrt{n})$, $i, j = 1, \ldots, n$.

By (A2), (A4) and (A5), one can directly conclude that $\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{K}_n^{(p_2)}\right) \to \boldsymbol{V}_n^{-1}$ as $n \to \infty$.

According to [20], the test statistic for testing $\mathcal{H}_o : \boldsymbol{\beta}_2 = 0$ diverges as $n \to \infty$, under any fixed alternatives $A_{\boldsymbol{\xi}} : \boldsymbol{\beta}_2 = \boldsymbol{\xi}$. To overcome this difficulty, in sequel, the following local alternatives are considered

$$K_{(n)} : \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{2(n)} = n^{-\frac{1}{2}}\boldsymbol{\xi}, \tag{4.1}$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{p_2})^\top \in \mathbb{R}^{p_2}$ is a fixed vector.

For notational convenience, let $-k_o\boldsymbol{A}^{-1}\boldsymbol{\beta} = \boldsymbol{\mu} = \left(\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top\right)^\top$, $\boldsymbol{\delta} = \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{\xi}$, $\boldsymbol{\mu}_{11.2} = \boldsymbol{\mu}_1 - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\Big((\boldsymbol{\beta}_2 - \boldsymbol{\xi}) - \boldsymbol{\mu}_2\Big)$, $\boldsymbol{\gamma} = \boldsymbol{\mu}_{11.2} + \boldsymbol{\delta}$, $\boldsymbol{A}_{22.1} = \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}$, $\boldsymbol{B} = \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-2}\boldsymbol{A}_{12}\boldsymbol{A}_{22.1}^{-1}$. Then, we have

Theorem 2
Under the regularity conditions (A1)-(A3) and local alternatives $\{K_{(n)}\}$

(i) $\boldsymbol{V}_{(n)}^{(1)} = \sqrt{n}\left(\hat{\boldsymbol{\beta}}_n^{(1)}\left(\boldsymbol{K}_n^{(p_1)}\right) - \boldsymbol{\beta}_1\right) \xrightarrow{\mathcal{D}} N_{p_1}\left(-\boldsymbol{\mu}_{11.2}, \sigma^2\boldsymbol{A}_{11.2}^{-1}\right)$
(ii) $\boldsymbol{V}_{(n)}^{(2)} = \sqrt{n}\left(\hat{\boldsymbol{\beta}}_n^{R1}(\boldsymbol{K}_n^{(p_1)}) - \boldsymbol{\beta}_1\right) \xrightarrow{\mathcal{D}} N_{p_1}(-\boldsymbol{\gamma}, \sigma^2\boldsymbol{A}_{11}^{-1})$

(iii) $\boldsymbol{V}_{(n)}^{(3)} = \sqrt{n}\left(\hat{\boldsymbol{\beta}}_n^{R1}\left(\boldsymbol{K}_n^{(p_1)}\right) - \hat{\boldsymbol{\beta}}_n^{(1)}\left(\boldsymbol{K}_n^{(p_1)}\right)\right) \xrightarrow{\mathcal{D}} N_{p_1}\left(-\boldsymbol{\gamma} + \boldsymbol{\mu}_{11.2}, \sigma^2\left(\boldsymbol{A}_{11}^{-1} - \boldsymbol{A}_{11.2}^{-1}\right)\right)$

where $\boldsymbol{A}_{11.2} = \boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}$.

If $\boldsymbol{A}_{12} = \boldsymbol{0}$, then $\boldsymbol{\delta} = \boldsymbol{0}$, $\boldsymbol{\gamma} = \boldsymbol{\mu}_{11.2}$ and $\boldsymbol{A}_{11.2} = \boldsymbol{A}_{11}$, and all the asymptotic risk functions reduce to common value $\sigma^2\,\mathrm{tr}(\boldsymbol{A}_{11}^{-1}) + \boldsymbol{\mu}_{11.2}^\top\boldsymbol{\mu}_{11.2}$ for all $\boldsymbol{\xi}$. In a similar fashion as in [20], if $\boldsymbol{A}_{12} \neq \boldsymbol{0}$, then when $\Delta^2 = \left(\boldsymbol{\xi}^\top\boldsymbol{A}_{22.1}^{-1}\boldsymbol{\xi}\right)\sigma^{-2}$, $\boldsymbol{A}_{22.1} = \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}$ moves away from 0, the asymptotic risk of $\hat{\boldsymbol{\beta}}_n^{R1}\left(\boldsymbol{K}_n^{(p_1)}\right)$ becomes unbounded. Further, if $\boldsymbol{\xi} = \boldsymbol{0}$, $\hat{\boldsymbol{\beta}}_n^{R1}\left(\boldsymbol{K}_n^{(p_1)}\right)$ is superior to $\hat{\boldsymbol{\beta}}_n^{(1)}\left(\boldsymbol{K}_n^{(p_1)}\right)$ (denoted by $\hat{\boldsymbol{\beta}}_n^{R1}\left(\boldsymbol{K}_n^{(p_1)}\right) > \hat{\boldsymbol{\beta}}_n^{(1)}\left(\boldsymbol{K}_n^{(p_1)}\right)$) in the sense of having smaller asymptotic risk function.

5. Shrinkage Estimator

Under the aforementioned analysis in previous section, one may consider improving the GRE of $\boldsymbol{\beta}_1$ by shrinking toward $\boldsymbol{0}$ (see [20] for more details). In this respect, one plausible choice is to combine both GRE and GRRE of $\boldsymbol{\beta}_1$ to obtain

$$\hat{\boldsymbol{\beta}}_n^{S}\left(\boldsymbol{K}_n^{(p_1)}\right) = \alpha_1\hat{\boldsymbol{\beta}}_n^{(1)}\left(\boldsymbol{K}_n^{(p_1)}\right) + \alpha_2\hat{\boldsymbol{\beta}}_n^{R1}\left(\boldsymbol{K}_n^{(p_1)}\right). \tag{5.1}$$

Now, let $\boldsymbol{\beta}^*$ be any estimator of $\boldsymbol{\beta}_1$. Then the asymptotic distributional risk (ADR) of $\boldsymbol{\beta}^*$ is evaluated by

$$ADR(\boldsymbol{\beta}_1; \boldsymbol{\beta}^*) = \lim_{n\to\infty} E\left[n(\boldsymbol{\beta}^* - \boldsymbol{\beta}_1)^\top(\boldsymbol{\beta}^* - \boldsymbol{\beta}_1)\right]$$

Under the pronounced regularity conditions and local alternatives $\{K_{(n)}\}$, the asymptotic risk of shrinkage ridge estimator (SRE) $\hat{\boldsymbol{\beta}}_n^{S}\left(\boldsymbol{K}_n^{(p_1)}\right)$, after some algebra, is given by

$$\begin{aligned} ADR\left(\boldsymbol{\beta}_1; \hat{\boldsymbol{\beta}}_n^{S}\left(\boldsymbol{K}_n^{(p_1)}\right)\right) &= \alpha_1^2\left[\sigma^2\,\mathrm{tr}(\boldsymbol{A}_{11.2}^{-1}) + \boldsymbol{\mu}_{11.2}^\top\boldsymbol{\mu}_{11.2}\right] + \alpha_2^2\left[\sigma^2\,\mathrm{tr}(\boldsymbol{A}_{11}^{-1}) + \boldsymbol{\gamma}^\top\boldsymbol{\gamma}\right] \\ &\quad +(\alpha_1+\alpha_2-1)^2\boldsymbol{\xi}^\top\boldsymbol{\xi} - 2\alpha_1(\alpha_1+\alpha_2-1)\boldsymbol{\mu}_{11.2}^\top\boldsymbol{\xi} - 2\alpha_2(\alpha_1+\alpha_2-1)\boldsymbol{\gamma}^\top\boldsymbol{\xi} \\ &\quad +\alpha_1\alpha_2\left(2\boldsymbol{\mu}_{11.2}^\top\boldsymbol{\xi} + 2\boldsymbol{\gamma}^\top\boldsymbol{\xi} - 2\boldsymbol{\mu}_{11.2}^\top\boldsymbol{\gamma} - \boldsymbol{\xi}^\top\boldsymbol{\xi}\right). \end{aligned} \tag{5.2}$$

Obviously, under $\boldsymbol{\xi} = \boldsymbol{0}$, $\hat{\boldsymbol{\beta}}_n^{S}\left(\boldsymbol{K}_n^{(p_1)}\right) > \hat{\boldsymbol{\beta}}_n^{(1)}\left(\boldsymbol{K}_n^{(p_1)}\right)$ whenever

$$(1-\alpha_1^2)\sigma^2\,\mathrm{tr}(\boldsymbol{A}_{11.2}^{-1}) - \alpha_2^2\sigma^2\,\mathrm{tr}(\boldsymbol{A}_{11}^{-1}) + \left[1-(\alpha_1-\alpha_2)^2\right]\boldsymbol{\mu}_{11.2}^\top\boldsymbol{\mu}_{11.2} > 0;$$

and for the special case $\alpha_1 = 1$, it can be revealed that $\hat{\boldsymbol{\beta}}_n^{S}\left(\boldsymbol{K}_n^{(p_1)}\right) > \hat{\boldsymbol{\beta}}_n^{(1)}\left(\boldsymbol{K}_n^{(p_1)}\right)$ whenever

$$0 < \alpha_2 < \frac{2\boldsymbol{\mu}_{11.2}^\top\boldsymbol{\mu}_{11.2}}{\sigma^2\,\mathrm{tr}(\boldsymbol{A}_{11}^{-1}) + \boldsymbol{\mu}_{11.2}^\top\boldsymbol{\mu}_{11.2}}. \tag{5.3}$$

Similarly, $\hat{\boldsymbol{\beta}}_n^{S}\left(\boldsymbol{K}_n^{(p_1)}\right) > \hat{\boldsymbol{\beta}}_n^{R1}\left(\boldsymbol{K}_n^{(p_1)}\right)$ whenever

$$(1-\alpha_2^2)\sigma^2\,\mathrm{tr}(\boldsymbol{A}_{11}^{-1}) - \alpha_1^2\sigma^2\,\mathrm{tr}(\boldsymbol{A}_{11.2}^{-1}) + \left[1-(\alpha_1-\alpha_2)^2\right]\boldsymbol{\mu}_{11.2}^\top\boldsymbol{\mu}_{11.2} > 0;$$

and for the special case $\alpha_2 = 1$, it can be revealed that $\hat{\boldsymbol{\beta}}_n^{S}\left(\boldsymbol{K}_n^{(p_1)}\right) > \hat{\boldsymbol{\beta}}_n^{R1}\left(\boldsymbol{K}_n^{(p_1)}\right)$ whenever

$$0 < \alpha_1 < \frac{2\boldsymbol{\mu}_{11.2}^\top\boldsymbol{\mu}_{11.2}}{\sigma^2\,\mathrm{tr}(\boldsymbol{A}_{11.2}^{-1}) + \boldsymbol{\mu}_{11.2}^\top\boldsymbol{\mu}_{11.2}}. \tag{5.4}$$

6. Numerical Results

In this section, the Monté-Carlo studies and a real data example about riboflavin production data set (see [6]) are consider to justify the assertions.

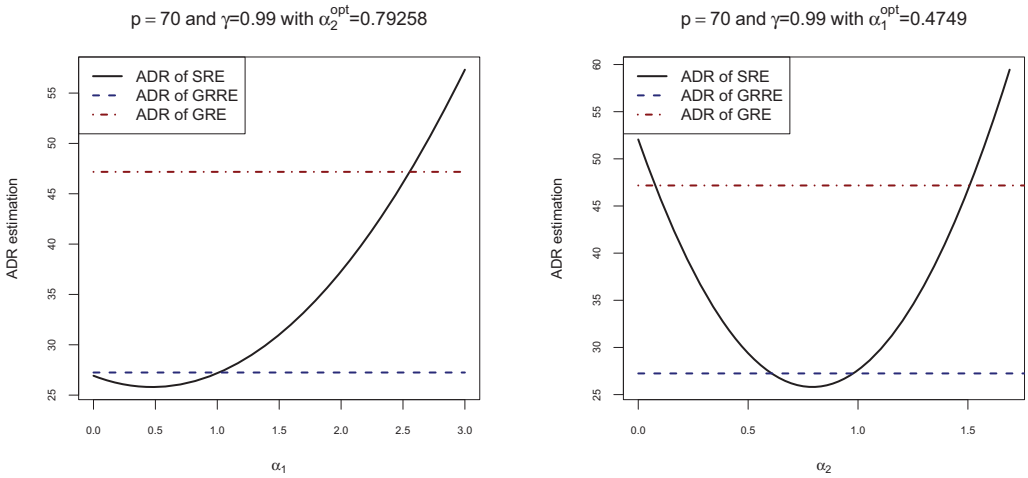

Figure 1. The diagram of ADRs versus α_1 and α_2 for simulated data set.

6.1. The Monté-Carlo simulation studies

To examine the performance of the proposed estimators, a Monté-Carlo simulation is performed. To achieve different degrees of collinearity, following [18, 8], the explanatory variables were generated using the following device for $n = 25$ and $p = 30,\ 50$ and 70 from the following model:

$$x_{ij} = (1-\gamma^2)^{\frac{1}{2}} z_{ij} + \gamma z_{ip}, \qquad i = 1, 2, ..., n, \quad j = 1, 2, ..., p \tag{6.1}$$

where z_{ij} are independent standard normal pseudo-random numbers and γ is specified so that the correlation between any two explanatory variables is given by γ^2. These variables are then standardized so that $\boldsymbol{X}_n^\top \boldsymbol{X}_n$ and $\boldsymbol{X}_n^\top \boldsymbol{y}_n$ are in correlation forms. Three different values $\gamma = 0.75, 0.90$ and 0.99 are considered for the correlation. Then the observations for the dependent variable are determined by

$$\boldsymbol{y}_n = \boldsymbol{X}_n^{(1)} \boldsymbol{\beta}_1 + \boldsymbol{X}_n^{(2)} \boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \tag{6.2}$$

with $\boldsymbol{\beta}_1 = (-1.25, 1, 2.5, 4, -3, -5)^\top$, $p_1 = 6$ and $n = 25$. To achieve the sparse model, $\boldsymbol{\beta}_2$ is generated as $N_{p_2}(\boldsymbol{0}, 0.01 \times \boldsymbol{I}_{p_2})$, $p_2 = p - p_1$. The error term of the model (6.2) is generated as

$$\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{V}_n),\ \sigma^2 = 1.44,\ \boldsymbol{V}_n[i,j] = \exp\big(-7|i-j|\big),\ i, j = 1, \ldots, n.$$

The Monté-Carlo simulation is performed with $M = 10^3$ replications, obtaining the ridge estimators $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_n^{(1)}\left(\boldsymbol{K}_n^{(p_1)}\right)$, $\hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_n^{R1}\left(\boldsymbol{K}_n^{(p_1)}\right)$, $\hat{\boldsymbol{\beta}}_3 = \hat{\boldsymbol{\beta}}_n^{S}\left(\boldsymbol{K}_n^{(p_1)}\right)$, in the sparse restricted regression model.

The relative efficiencies of the above methods with respect to the first method are estimated as

$$Eff(\hat{\boldsymbol{\beta}}_i) = \frac{\frac{1}{M}\sum_{m=1}^{M} \left\| \hat{\boldsymbol{\beta}}_1^{(m)} - \boldsymbol{\beta}_1 \right\|_2^2}{\frac{1}{M}\sum_{m=1}^{M} \left\| \hat{\boldsymbol{\beta}}_i^{(m)} - \boldsymbol{\beta}_1 \right\|_2^2},\ i = 1, 2, 3,$$

where $\hat{\boldsymbol{\beta}}_i^{(m)}$ is the estimators obtained in the mth iteration.

All numerical computations were conducted using the statistical package R. In Tables 1 to 4, the proposed estimators along with $ADR(.)$, efficiencies of proposed estimators relative to GRE and optimal values of α_1 and α_2 were computed. The asymptotic risk of SRE was employed to select the optimal parameters α_1 and α_2, numerically, by minimizing the ADR function of SRE. The minimum of ADR approximately occurred at $\alpha_1^{opt} = 0.4749$ and $\alpha_2^{opt} = 0.7449$ for $p = 70$ and $\gamma = 0.99$. Figure 1 shows the ADR functions versus α_1 and α_2. Since the results were similar across cases, only the results for $p = 70$ and $\gamma = 0.99$ were reported to save space.

As it can be found from Tables 1-3, the SRE performs better than the other proposed estimators in the sense of having smaller ADR. Also, because of the sparsity assumption, both GRRE and SRE are more efficient than GRE. Moreover, since the simulated model is sparse, as it is evident from Table 4, the allocated weight to the GRRE (α_2) is greater than that of the GRE (α_1). From Figure 1, one can distinguish the optimal regions of superiority of SRE over GRE and GRRE.

Table 1. Evaluation of the proposed estimators for $\gamma = 0.75$

Method / Coefficients	$p = 30$			$p = 50$			$p = 70$		
	GRE	GRRE	SRE	GRE	GRRE	SRE	GRE	GRRE	SRE
$\hat{\beta}_1$	-1.3423	-1.2296	-1.2932	-1.0754	-1.2549	-1.3500	-0.9888	-1.4900	-1.5778
$\hat{\beta}_2$	0.4666	1.0251	1.0085	0.0492	0.9809	0.9399	0.2205	0.5115	0.5116
$\hat{\beta}_3$	1.7686	2.5435	2.5604	0.7694	2.4766	2.4758	0.2324	1.7544	1.8014
$\hat{\beta}_4$	2.9308	4.0521	4.0972	1.5382	3.9842	4.0167	0.7041	3.0047	3.1030
$\hat{\beta}_5$	-2.6616	-2.5916	-2.7564	-1.9584	-2.6366	-2.8188	-1.8737	-4.1290	-4.2485
$\hat{\beta}_6$	-3.2513	-4.8248	-4.8671	-1.6499	-4.9238	-4.9359	-1.5163	-3.9774	-4.1091
k^{opt}	1e-04	0.01	—	0.01	0.13	—	0.14	0.36	—
$\hat{ADR}(\hat{\boldsymbol{\beta}}_i)$	9.6646	1.0270	0.7930	24.091	0.9813	0.7926	32.490	6.5667	6.1225
$Eff(\hat{\boldsymbol{\beta}}_i)$	1.0000	9.4102	12.187	1.0000	24.550	30.394	1.0000	4.9476	5.3066

Table 2. Evaluation of the proposed estimators for $\gamma = 0.90$

Method / Coefficients	$p = 30$			$p = 50$			$p = 70$		
	GRE	GRRE	SRE	GRE	GRRE	SRE	GRE	GRRE	SRE
$\hat{\beta}_1$	-1.3071	-1.2120	-1.2890	-0.9863	-1.2109	-1.6814	-0.7243	-1.2703	-1.7947
$\hat{\beta}_2$	0.4390	0.9715	0.9271	0.1208	1.0719	0.8296	0.0910	0.9842	0.7958
$\hat{\beta}_3$	1.7053	2.5750	2.5639	0.9122	2.7129	2.5980	0.5531	2.6233	2.5118
$\hat{\beta}_4$	2.8883	4.1709	4.1653	1.6350	4.3926	4.3216	1.0719	4.2234	4.2779
$\hat{\beta}_5$	-2.6237	-2.0233	-2.3803	-1.7832	-0.1338	-1.9457	-1.2949	-0.0663	-1.8843
$\hat{\beta}_6$	-2.8491	-4.8424	-4.7643	-1.0767	-4.5903	-3.9601	-0.6508	-4.6289	-3.9573
k^{opt}	0.10	0.30	—	0.17	0.43	—	0.43	1.07	—
$\hat{ADR}(\hat{\boldsymbol{\beta}}_i)$	13.391	3.0308	2.2409	28.250	13.213	6.4622	36.533	14.320	6.9432
$Eff(\hat{\boldsymbol{\beta}}_i)$	1.0000	4.4181	5.9755	1.0000	2.1380	4.3716	1.0000	2.5512	5.2617

Table 3. Evaluation of the proposed estimators for $\gamma = 0.99$

Method / Coefficients	$p = 30$			$p = 50$			$p = 70$		
	GRE	GRRE	SRE	GRE	GRRE	SRE	GRE	GRRE	SRE
$\hat{\beta}_1$	-1.0841	-1.3174	-1.3561	-0.9212	-1.4586	-1.5295	-0.9274	-1.2769	-1.4437
$\hat{\beta}_2$	0.2518	0.4380	0.5003	0.1785	0.6240	0.6248	0.1715	0.4523	0.3484
$\hat{\beta}_3$	1.1976	1.8524	1.9290	0.8593	2.1653	2.1419	0.3954	1.5600	1.5314
$\hat{\beta}_4$	2.0972	3.1175	3.2193	1.5161	3.6718	3.6278	0.8316	2.6495	2.6456
$\hat{\beta}_5$	-2.0060	-2.3669	-2.4796	-1.7413	-0.3643	-1.3213	-1.6408	-3.0978	-3.2068
$\hat{\beta}_6$	-1.7052	-3.4797	-3.5661	-0.8283	-4.3776	-3.8814	-0.8940	-2.6663	-2.8464
k^{opt}	0.23	0.29	—	0.21	0.49	—	1.01	1.64	—
$\hat{ADR}(\hat{\boldsymbol{\beta}}_i)$	30.241	15.841	13.648	35.054	16.917	13.038	37.868	20.420	18.396
$Eff(\hat{\boldsymbol{\beta}}_i)$	1.0000	1.9090	2.2157	1.0000	2.0722	2.6887	1.0000	1.8545	2.0585

Table 4. Optimal values of shrinkage parameters for different p and γ values

	$\gamma = 0.75$			$\gamma = 0.90$			$\gamma = 0.99$		
	$p = 30$	$p = 50$	$p = 70$	$p = 30$	$p = 50$	$p = 70$	$p = 30$	$p = 50$	$p = 70$
α_1^{opt}	0.1505	0.1660	0.1019	0.2571	1.0863	1.5641	0.2242	0.5922	0.4461
α_2^{opt}	0.9015	0.9464	1.0027	0.8182	0.6054	0.6386	0.8660	0.7700	0.8169

6.2. Application to Riboflavin Production Data

To illustrate the usefulness of the suggested strategies for high-dimensional data in the semiparametric regression model, the data set about riboflavin (vitamin B2) production is considered in Bacillus subtilis, which can be found in R package "hdi". There is a single real valued response variable which is the logarithm of the

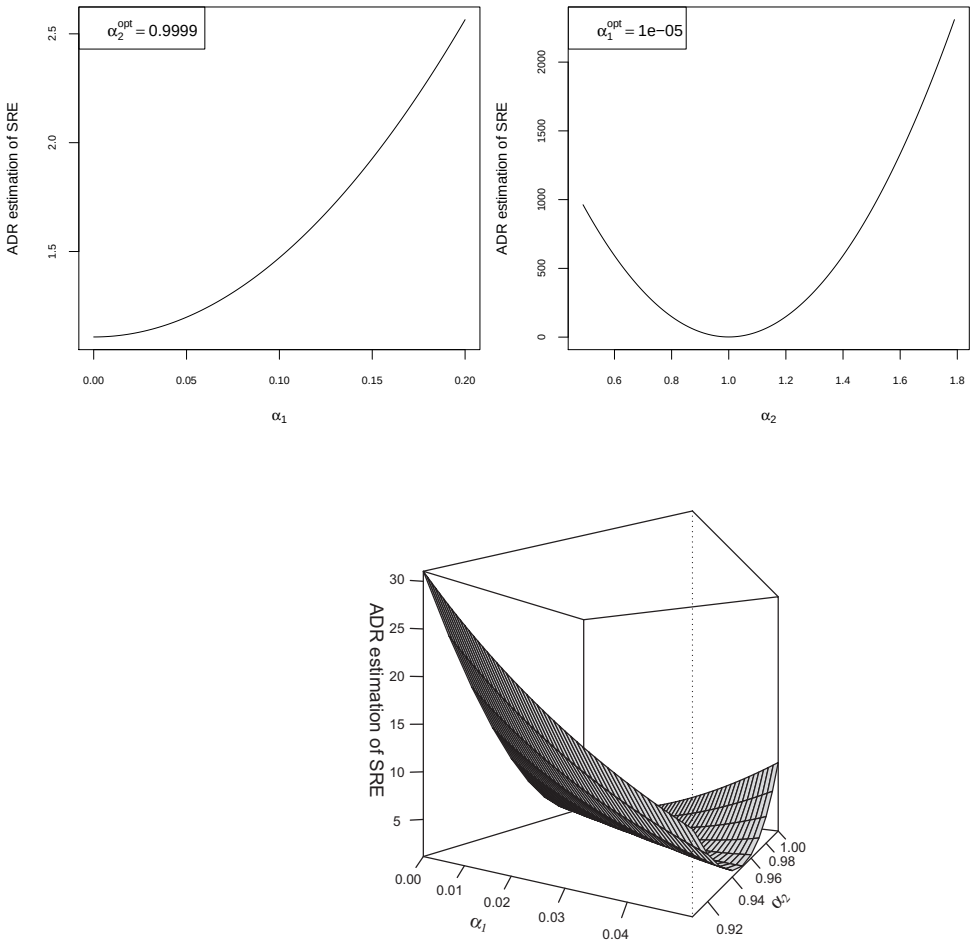

Figure 2. The diagram of ADR of SRE versus α_1 and α_2 for real data set.

riboflavin production rate. Furthermore, there are $p = 4088$ explanatory variables measuring the logarithm of the expression level of 4088 genes. There is one rather homogeneous data set from $n = 71$ samples that were hybridized repeatedly during a fed batch fermentation process where different engineered strains and strains grown under different fermentation conditions were analyzed. Based on 100-fold cross validation, the Lasso shrinks 4047 parameters to zero and remains $p_1 = 41$ significant explanatory variables. So, the specification of the sparse regression model is

$$\boldsymbol{y}_n = \boldsymbol{X}_n^{(1)}\boldsymbol{\beta}_1 + \boldsymbol{X}_n^{(2)}\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \tag{6.3}$$

where $p_1 = 41$ and $p_2 = 4047$.

According to [23], the unknown $\boldsymbol{V}_n$ can be estimated by a consistent estimator

$$\hat{\boldsymbol{V}}_n = \frac{1}{n-p_1}\left(\boldsymbol{y}_n - \boldsymbol{X}_n^{(1)}\boldsymbol{b}_n^{R1}\right)\left(\boldsymbol{y}_n - \boldsymbol{X}_n^{(1)}\boldsymbol{b}_n^{R1}\right)^{\top}, \tag{6.4}$$

where $\boldsymbol{b}_n^{R1} = \left(\boldsymbol{X}_n^{(1)\top}\boldsymbol{X}_n^{(1)}\right)^{-1}\boldsymbol{X}_n^{(1)\top}\boldsymbol{y}_n$ is ordinary least square estimator for the sparse regression model.

Table 5 shows a summary of the results. In this Table, the RSS, MSE and R^2 respectively are the residual sum of squares, mean square error and coefficient of determination of the model, i.e., $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, $1/(n-p_1)\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and $R^2 = 1 - RSS/S_{yy}$, where $\hat{y}_i = \boldsymbol{x}_i^{\top}\hat{\boldsymbol{\beta}}$.

Table 5. Evaluation of proposed estimators for real data set

Method	GRE	GRRE	SRE
k^{opt}	3.4241	7.6132	—
RSS	39.39465	1.108829	1.107424
MSE	1.313155	0.036961	0.036914
R^2	0.335704	0.981302	0.981326

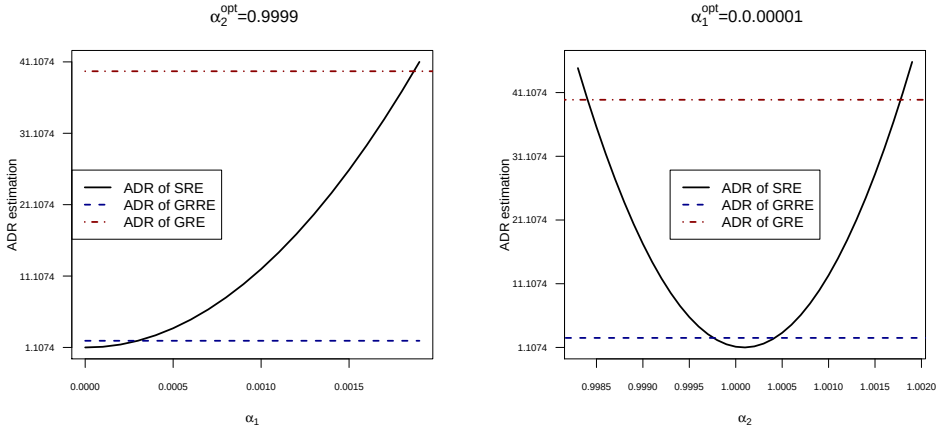

Figure 3. The diagram of ADRs versus α_1 and α_2 for real data set.

According to Figure 2, the minimum of ADR approximately occurred at $\alpha_1^{opt} = 0.00001$ and $\alpha_2^{opt} = 0.9999$. Also, the ADR functions of proposed estimators versus α_1 and α_2 are plotted in Figure 3, which can be used in a similar fashion as in the simulation.

7. Summary

In this paper, the generalized ridge estimators were proposed for the sparse multiple regression model. In this context, a generalized restricted ridge estimator exhibited for the sparsity test $\boldsymbol{\beta}_2 = \mathbf{0}$, where the regression vector-parameter β partitioned as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$. Information contained in the generalized restricted ridge estimator was employed to construct a shrinkage ridge estimator. Some asymptotic distributional results were given for the proposed estimators and superiority conditions discussed under some regularity conditions. All the analysis were conducted under a classical view point, using the Tikhonov regularization.

It is well-known that as the sample size tends to infinity, both Bayesian and classical analysis give almost same results. However, under a finite sample size, one may consider a generalized Tikhonov regularization and derive ridge regression estimators in Bayesian context. To see this, take $\boldsymbol{\Sigma}$, $\boldsymbol{\beta}_o$ and $\boldsymbol{\Omega}$ as the covariance of $\boldsymbol{y}_n$, expected value of $\boldsymbol{\beta}$ and covariance of $\boldsymbol{\beta}$, respectively. Therefore, the generalized ridge estimator has form

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_n^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_n + \boldsymbol{\Omega}^{-1})^{-1} (\boldsymbol{X}_n^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_n + \boldsymbol{\Omega}^{-1} \boldsymbol{\beta}_o). \tag{7.5}$$

Then, generalized restricted ridge estimator can be verified for the sparse (restricted) model.

Results of the Monté-Carlo simulation for different γ, $n = 25$, $p = 30, 50$ and 70, were presented in Tables 1 to 4 and Figure 1. From these Tables, it is realized that $\hat{\boldsymbol{\beta}}_n^S(\boldsymbol{K}_n^{(p_1)})$ is leading to be the best estimator among others, since it offers smaller risk and bigger efficiency value, in all cases. According to Figure 1, results show that the global minimum occurs under the ADR of GRE and GRRE. Also, there exists an optimal region for performance of SRE with respect to GRE and GRRE. For the real example, from Table 5 as well as Figures 2 and 3, it can be deduced that $\hat{\boldsymbol{\beta}}_n^S(\boldsymbol{K}_n^{(p_1)})$ is quite efficient in the sense that it has significant goodness of fit value.

Acknowledgments

The author's research is supported in part by Semnan University.

REFERENCES

1. F. Akdenïz, and G. Tabakan, *Restricted ridge estimators of the parameters in semiparametric regression model*, Communication in Statistics-Theory and Methods, vol. 38, no. 11, pp. 1852–1869, 2009.
2. E. Akdenïz Duran, W.K. Härdle, M. Osipenko, *Difference based ridge and Liu type estimators in semiparametric regression models*, Journal of Multivariate Analysis, vol. 105, no. 1, pp. 164–175, 2012.
3. F. Akdenïz, and M. Roozbeh, *Generalized difference-based weighted mixed almost unbiased ridge estimator in partially linear models*, Statistical Papers, DOI: 10.1007/s00362-017-0893-9, 2017.
4. M. Amini, and M. Roozbeh, *Optimal partial ridge estimation in restricted semiparametric regression models*, Journal of Multivariate Analysis, vol. 136, no. C, pp. 26–40, 2015.
5. P. Bühlmann, and S. van de Geer, *Statistics for High-dimensional Data: Methods, Theory and Applications*, Springer, Heidelberg, 2011.
6. P. Bühlmann, M. Kalisch, and L. Meier, *High-dimensional statistics with a view towards applications in biology*, Annual Review of Statistics and Its Application, vol. 1, pp. 255–278, 2014.
7. N. El Karoui, *Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results*, Arxiv: 1311.2445, 2016.
8. D.G. Gibbons, *A simulation study of some ridge estimators*, Journal of the American Statistical Association, vol. 76, no. 373, pp. 131–139, 1981.
9. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, second edition, Springer, New York, 2009.
10. W.J. Hemmerle, and T.F. Brantle, *Explicit and Constrained Generalized Ridge Estimation*, Technometrics, vol. 20, pp, 109–119, 1978.
11. W.J. Hemmerle, and M.B. Carey, *Some Properties of Generalized Ridge Estimators*, Communications in Statistics-Simulation and Computation, vol. 12, pp, 239–253, 1983.
12. A.E. Hoerl, and R.W. Kennard, *Ridge regression: biased estimation for non-orthogonal problems*, Technometrics, vol. 12, no. 1, pp. 55–67, 1970.
13. R.R. Hocking, F.M., Speed, and M.J. Lynn, *A Class of Biased Estimators in Linear Regression*, Technometrics, vol. 18, pp. 425–437, 1976.
14. B.M.G. Kibria, (2003). *Performance of some new ridge regression estimators*, Communications in Statistics-Simulation and Computation, vol. 32, pp.419–435.
15. B.M.G. Kibria, and S. Banik, *Some Ridge Regression Estimators and Their Performances*, Journal of Modern Applied Statistical Methods, vol. 15, no. 1, pp. 206–238, 2016.
16. B.M.G. Kibria, and A.K.Md.E. Saleh, *Improving the Estimators of the Parameters of a Probit Regression Model: A Ridge Regression Approach*, Journal of Statistical Planning and Inference, vol. 142, no. 6, pp. 1421–1435, 2011.
17. K. Knight, and W. Fu, *Asymptotic for lasso-type estimators*, Annals of Statistics, vol. 28, no. 5, pp. 1356–1378, 2000.
18. G.C. McDonald, and D.I. Galarneau, *A monte carlo evaluation of some ridge-type estimators*, Journal of the American Statistical Association, vol.70, no. 350, pp. 407–416, 1975.
19. M. Roozbeh, (2015). *Shrinkage ridge estimators in semiparametric regression models*, Journal of Multivariate Analysis, vol. 136, no. C, pp. 56–74, 2015.
20. A.K.Md.E. Saleh, *Theory of Preliminary Test and Stein-type Estimation with Applications*, John Wiley, New York, 2006.
21. A.N. Tikhonov, A. N. (1963). *Solution of incorrectly formulated problems and the regularization method*, Soviet Mathematics vol. 4, no. 4, pp. 1035–1038, 1963.
22. B. Yuzbashi, and S.E. Ahmed, *Shrinkage ridge regression estimators in high-dimensional linear models*, Proceedings of the Ninth International Conference on Management Science and Engineering Management, pp. 793-807, 2015.
23. A. Zellner, *An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias*, Journal of the American Statistical Association, vol. 57, no. 298, pp. 348–368, 1962.