mathematics

MDPI

Article

A High-Dimensional Counterpart for the Ridge Estimator in Multicollinear Situations

Mohammad Arashi [1,2,*], Mina Norouzirad [3], Mahdi Roozbeh [4] and Naushad Mamode Khan [5]

[1] Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad P.O. Box 9177948974, Iran
[2] Department of Statistics, University of Pretoria, Pretoria 0002, South Africa
[3] Department of Statistics, Faculty of Mathematical Sciences, Shahrood University of Technology, Shahrood P.O. Box 3619995181, Iran; mina.norouzirad@gmail.com
[4] Department of Statistics, Faculty of Mathematics, Statistics and Computer Sciences, Semnan University, Semnan P.O. Box 3514799422, Iran; mahdi.roozbeh@semnan.ac.ir
[5] Department of Economics and Statistics, University of Mauritius, Réduit 80837, Mauritius; n.mamodekhan@uom.ac.mu
* Correspondence: arashi@um.ac.ir

Abstract: The ridge regression estimator is a commonly used procedure to deal with multicollinear data. This paper proposes an estimation procedure for high-dimensional multicollinear data that can be alternatively used. This usage gives a continuous estimate, including the ridge estimator as a particular case. We study its asymptotic performance for the growing dimension, i.e., $p \to \infty$ when n is fixed. Under some mild regularity conditions, we prove the proposed estimator's consistency and derive its asymptotic properties. Some Monte Carlo simulation experiments are executed in their performance, and the implementation is considered to analyze a high-dimensional genetic dataset.

Keywords: asymptotic; high–dimension; Liu estimator; multicollinear; ridge estimator

Citation: Arashi, M.; Norouzirad, M.; Roozbeh, M.; Khan, N.M. A High-Dimensional Counterpart for the Ridge Estimator in Multicollinear Situations. *Mathematics* **2021**, *9*, 3057. https://doi.org/10.3390/math9233057

Academic Editor: Jin-Ting Zhang

Received: 10 November 2021
Accepted: 24 November 2021
Published: 28 November 2021

1. Introduction

Consider the multiple regression model given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{Y} = (y_1, \ldots, y_n)^\top$ is a vector of n responses, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ is an $n \times p$ design matrix, with the ith predictor $\mathbf{x}_i \in \mathbb{R}^p$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is the coefficients vector, and $\boldsymbol{\epsilon}$ is an n-vector of unobserved errors. Further, we shall assume $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$, $\mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top) = \sigma^2\mathbf{I}_n$, $\sigma^2 > 0$.

When $p < n$, the ordinary least squares (LS) estimator of $\boldsymbol{\beta}$ is given by

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \mathcal{S}(\boldsymbol{\beta}), \quad \mathcal{S}(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}. \end{aligned} \tag{2}$$

However, for the high dimensional (HD) case, $p > n$ the LS estimator cannot be obtained, because $\mathbf{X}^\top\mathbf{X}$ is rank deficient. It is well known that the ridge regression (RR) estimator of [1], followed by [2] regularization, however, exists. The rationale is to add a positive value $k > 0$ to the eigenvalues of $\mathbf{X}^\top\mathbf{X}$ to efficiently estimate the parameters via $\hat{\boldsymbol{\beta}}^{\text{Ridge}} = (\mathbf{X}^\top\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{Y}$. Refer to Saleh et al. [3] for theory and application of the RR approach. Using the projection of $\boldsymbol{\beta}$ onto the row space of $\mathbf{X}$ is a well-described remedy. Wang et al. [4] used this technique and proposed a high dimensional LS estimator as a limiting case of the RR, while Buhlmann [5] also used the projection method and developed a bias correction in the RR estimator to propose a bias-corrected RR estimator

for the high dimensional setting. Shao and Deng [6] used the method and proposed to threshold the RR estimator when the projection vector is sparse, in the sense that many of its components are small and demonstrated consistency. Dicker [7] studied the minimum property of the RR estimator and derived its asymptotic risk for the growing dimension, i.e., $p \to \infty$. Although the RR estimator involves high dimensional problems, there exits a counterpart that has not been considered in high dimension.

An Existing Two-Parameter Biased Estimator

It is well known that the RR estimator is an efficient approach for multicollinear situations. Since then, many authors have developed ridge-type estimators to overcome the issue of multicollinearity. One drawback of the RR estimator is that it is a non-linear function of the tuning parameter. Hence, Liu [8] developed a similar estimator; however, it is linear for the tuning parameter via the following optimization problem, for the case $p < n$:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathcal{S}(\boldsymbol{\beta}) + (d\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (d\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{3}$$

The solution to the optimization problem (3) has the form

$$\hat{\boldsymbol{\beta}}^{\text{Liu}} = (\mathbf{X}^\top \mathbf{X} + \mathbf{I}_p)^{-1} (\mathbf{X}^\top \mathbf{Y} + d\hat{\boldsymbol{\beta}}), \tag{4}$$

where $d \in (0, 1)$ is termed as the biasing parameter.

Combining the advantages of the RR and Liu estimators, Ozkale and Kaciranlar [9] proposed a two-parameter estimator by solving the following optimization problem:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathcal{S}(\boldsymbol{\beta}) + k\Big[(d\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top (d\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - c\Big], \tag{5}$$

where c is a constant, and k is the Lagrangian multiplier. The resulting two-parameter ridge estimator has the form

$$\hat{\boldsymbol{\beta}}(k, d) = (\mathbf{X}^\top \mathbf{X} + k\mathbf{I}_p)^{-1} (\mathbf{X}^\top \mathbf{Y} + kd\hat{\boldsymbol{\beta}}) \tag{6}$$

The above estimator has several advantages and can be simplified to LS, RR, and Liu estimators as limiting cases (see Figure 1). It can be argued that this estimator can also be interpreted as a restricted estimator under stochastic prior information about $\boldsymbol{\beta}$.

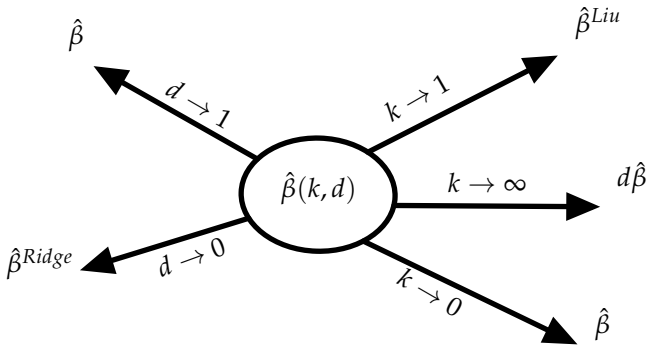

Figure 1. Special limiting cases.

With growing dimensions p, $p > n$, the LS estimator (2) cannot be obtained, so it is not possible to use the two-parameter ridge estimator in Equation (6). Hence, developing a high-dimensional two-parameter version of this estimator and studying its asymptotic performance is interesting and worthwhile. Therefore, in this paper, we propose a high-dimensional version of Ozkale and Kaciranlar's estimator and give the asymptotic properties. The paper's organization is as follows: In Section 2, a high-dimensional two-parameter estimator is proposed, and its asymptotic characteristics are discussed. Section 3

indicates the generalized cross validation for choosing the parameters. In Section 4, some simulation experiments are presented to assess the novel estimator's statistical and computational performance, and an application to the AML data is illustrated in this section The conclusion is presented in the last section.

2. The Proposed Estimator

In this section, we develop an HD estimator and establish its asymptotic properties. To show a component is dependent to p, we shall use the subscript p and particularly consider the scenarios in which $p \to \infty$ and n is fixed. This is termed *large p, fixed n*, which is more general than scenarios with $p/n \to \rho \in (0, \infty)$, a common assumption in high-dimensional settings.

Consider a diverging number of variables case, in which p is allowed to tend to infinity. This case fulfills the high-dimensional case $p > n$. Under this setting, the inverse of $\mathbf{X}^\top\mathbf{X}$ does not exist; however, the RR estimator is still valid and applicable. Further, the Liu estimator cannot be obtained. As a remedy, one can use the Moore–Penrose inverse of $\mathbf{X}^\top\mathbf{X}$, a particular case of the generalized inverse. Wang and Leng [10] showed that $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ can be seen as the Moore-Penrose inverse of $\mathbf{X}$ for $p < n$, and that $\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}$ is the Moore–Penrose inverse of $\mathbf{X}$ when $p > n$. This gives, for any $p, n > 0$,

$$(\mathbf{X}^\top\mathbf{X} + s\mathbf{I}_p)^{-1}\mathbf{X}^\top = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + s\mathbf{I}_n)^{-1}, \tag{7}$$

where s is an arbitrary nonegative constant.

Multiplying both sides of (7) by $\mathbf{Y}$ reveals that the LS estimator can be represented as

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \lim_{s\to\infty}(\mathbf{X}^\top\mathbf{X} + s\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{Y} \\ &= \lim_{s\to\infty}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + s\mathbf{I}_n)^{-1}\mathbf{Y} \\ &= \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{Y}. \end{aligned} \tag{8}$$

Now, for the HD case, substitute (8) in (6) to obtain

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{\mathrm{HD}} &= (\mathbf{X}^\top\mathbf{X} + k_p\mathbf{I}_p)^{-1}(\mathbf{X}^\top\mathbf{Y} + k_p d_p\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{Y}) \\ &= (\mathbf{X}^\top\mathbf{X} + k_p\mathbf{I}_p)^{-1}(\mathbf{X}^\top + k_p d_p\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1})\mathbf{Y} \\ &= (\mathbf{X}^\top\mathbf{X} + k_p\mathbf{I}_p)^{-1}(\mathbf{X}^\top + k_p d_p\mathbf{X}^+)\mathbf{Y}, \end{aligned} \tag{9}$$

where $\mathbf{X}^+ = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}$ is the Moore–Penrose inverse of $\mathbf{X}$.

We impose the following regularity conditions for studying the asymptotic performance of the estimator. $\hat{\boldsymbol{\beta}}^{\mathrm{HD}}$ given by (9).

(A1) $1/k_p = o(1)$. There exists a constant $0 \le \delta < 0.5$, such that a component of $\mathbf{X}$ is $O\left(k_p^{\delta}\right)$.

(A2) $d_p = o(1)$. There exists a constant $0 \le \eta < 0.5$, such that a component of $\mathbf{X}^+$ is $O\left(d_p^{-\eta}\right)$.

(A3) For sufficiently large p, there is a vector $\mathbf{b}_{p\times 1}$, such that $\boldsymbol{\beta} = \mathbf{X}^\top\mathbf{X}\mathbf{b}$, and there exists a constant $\varepsilon > 0$, such that each component of the vector $\mathbf{b}_{p\times 1}$ is $O(1/p^{\varepsilon+1.5})$, and $k_p = o(p^{\varepsilon}a_p)$, with $a_p = o(1)$. (An example of such choice is $k_p = \sqrt{p}$ and $\varepsilon = 0.5 + \delta$).

(A4) For sufficiently large p, there exists a constant $\delta > 0$, such that each component of $\boldsymbol{\beta}$ is $O(p^{-2-\delta})$ and $1/d_p = o(p^{\delta})$. Further, $k_p^{\delta-1} = o(d_p)$.

Assumption **(A3)** is adopted from Luo [11]. Let $\hat{\boldsymbol{\beta}}^{\mathrm{HD}} = \left(\hat{\beta}_1^{\mathrm{HD}}, \ldots, \hat{\beta}_p^{\mathrm{HD}}\right)^\top$.

Theorem 1. *Assume* **(A1)** *and* **(A2)**. *Then,* $\mathrm{var}\left(\hat{\boldsymbol{\beta}}_i^{\mathrm{HD}}\right) = o(1)$ *for all* $i = 1, \ldots, p$.

Proof. For the proof, refer to Appendix A. □

Theorem 2. *Assume* **(A1)–(A3)**. *Further, suppose* $\lambda_{ip} = O(k_p)$, *where* $\lambda_{ip} > 0$ *is the ith eigenvalue of* $\mathbf{X}^\top\mathbf{X}$. *Then,* $\text{bias}(\hat{\beta}_i^{\text{HD}}) = o(1)$ *for all* $i = 1, 2, \ldots, p$.

Proof. For the proof, refer to Appendix A. □

Using Theorems 1 and 2, it can be verified that the HD estimator $\hat{\beta}^{\text{HD}}$ is a consistent estimator for $\boldsymbol{\beta}$ as $p \to \infty$.

The following result reveals the asymptotic distribution of this estimator as $p \to \infty$.

Theorem 3. *Assume* $1/k_p = o(1)$, *and for sufficiently large* p, *there exists a constant* $\delta > 0$, *such that each component of* $\boldsymbol{\beta}$ *is* $O(1/p^{2+\delta})$. *Let* $k_p = o(p^\delta)$, $\lambda_{ip} = o(k_p)$. *Furthermore, suppose that* $\epsilon \sim \mathcal{N}_n(0, \sigma^2\mathbf{I}_n)$, $\sigma^2 > 0$. *Then,*

$$\frac{1}{d_p}\left(\hat{\boldsymbol{\beta}}^{\text{HD}} - \boldsymbol{\beta}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{X}^+\mathbf{X}^{+\top}\right) \qquad as \quad p \to \infty. \tag{10}$$

Proof. For the proof, refer to Appendix A. □

3. Generalized cross Validation

As noted, the estimator $\hat{\beta}^{\text{HD}}$ depends on both the ridge parameter k_p and Liu parameter d_p that must be optimized in practice. To do this, we use the generalized cross-validation (GCV) criterion. The GCV uses to choose the ridge and Liu parameters by minimizing an estimate of the unobservable risk function

$$\begin{aligned} \text{R}(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}}^{\text{HD}}) &= \frac{1}{n}\left(\text{E}(\mathbf{Y}) - \hat{\mathbf{Y}}^{\text{HD}}(k_p, d_p)\right)^\top\left(\text{E}(\mathbf{Y}) - \hat{\mathbf{Y}}^{\text{HD}}(k_p, d_p)\right) \\ &= \frac{1}{n}\left\|\text{E}(\mathbf{Y}) - \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{HD}}(k_p, d_p)\right\|^2, \end{aligned}$$

where

$$\begin{aligned} \hat{\mathbf{Y}}^{\text{HD}}(k_p, d_p) &= \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{HD}} \\ &= (\mathbf{X}^\top\mathbf{X} + k_p\mathbf{I}_p)^{-1}(\mathbf{X}^\top + k_p d_p\mathbf{X}^+)\mathbf{Y} \\ &= \mathbf{H}(k_p, d_p)\mathbf{Y}, \end{aligned} \tag{11}$$

with $\mathbf{H}(k_p, d_p) = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + k_p\mathbf{I}_p)^{-1}(\mathbf{X}^\top + k_p d_p\mathbf{X}^+)$, termed as the hat matrix of $\mathbf{Y}$.

This is straightforward to demonstrate, as in [12].

$$\begin{aligned} \text{E}\left(\text{R}(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}}^{\text{HD}})\right) &= \frac{1}{n}\left\|\left(\mathbf{I}_n - \mathbf{H}(k_p, d_p)\right)\mathbf{X}\boldsymbol{\beta}\right\|^2 + \frac{\sigma^2}{n}\,\text{tr}\left(\mathbf{H}(k_p, d_p)^\top\mathbf{H}(k_p, d_p)\right) \\ &= \nu_1^2(k_p, d_p) + \sigma^2\nu_2(k_p, d_p), \end{aligned}$$

where $\nu_1^2(k_p, d_p) = \frac{1}{n}\left\|\left(\mathbf{I}_n - \mathbf{H}(k, d)\right)\mathbf{X}\boldsymbol{\beta}\right\|^2$ and $\nu_2(k_p, d_p) = \frac{1}{n}\,\text{tr}\left(\mathbf{H}(k_p, d_p)^\top\mathbf{H}(k_p, d_p)\right)$.

The GCV function is then defined as

$$\begin{aligned} \text{GCV}(\hat{\boldsymbol{\beta}}^{\text{HD}}) &= \frac{\frac{1}{n}\left\|\left(\mathbf{I}_n - \mathbf{H}(k_p, d_p)\right)\boldsymbol{y}\right\|^2}{\left(1 - \frac{1}{n}\,\text{tr}\left(\mathbf{H}(k_p, d_p)\right)\right)^2} \\ &= \frac{\frac{1}{n}\left\|\left(\mathbf{I}_n - \mathbf{H}(k_p, d_p)\right)\boldsymbol{y}\right\|^2}{\left(1 - \mu_1(k_p, d_p)\right)^2}, \end{aligned} \tag{12}$$

where $\mu_1(k_p, d_p) = \frac{1}{n}\text{tr}\left(\mathbf{H}(k_p, d_p)\right)$.

The following theorem extends the GCV theorem proposed by Akdeniz and Roozbeh [13].

Theorem 4. *According to the definition of GCV, we have*

$$\frac{\mathrm{E}\Big(\mathrm{R}(\boldsymbol{\beta};\hat{\boldsymbol{\beta}}^{\mathrm{HD}})\Big)-\mathrm{E}\Big(\mathrm{GCV}(\hat{\boldsymbol{\beta}}^{\mathrm{HD}})\Big)+\sigma^2}{\mathrm{E}\Big(\mathrm{R}(\boldsymbol{\beta};\hat{\boldsymbol{\beta}}^{\mathrm{HD}})\Big)} = \left(1-\frac{\sigma^2}{\big(1-\mu_1(k_p,d_p)\big)^2}\right) + \frac{1}{D(k,d)}\times\frac{\sigma^2\mu_1(k_p,d_p)^2}{\big(1-\mu_1(k_p,d_p)\big)^2},$$

where $D(k,d)=\nu_1^2(k_p,d_p)+\sigma^2\nu_2(k_p,d_p)$, *and consequently,*

$$\frac{\Big|\mathrm{E}\Big(\mathrm{R}(\boldsymbol{\beta};\hat{\boldsymbol{\beta}}^{\mathrm{HD}})\Big)-\mathrm{E}\Big(\mathrm{GCV}(\hat{\boldsymbol{\beta}}^{\mathrm{HD}})\Big)+\sigma^2\Big|}{\mathrm{E}\Big(\mathrm{R}(\boldsymbol{\beta};\hat{\boldsymbol{\beta}}^{\mathrm{HD}})\Big)} < \frac{\sigma^2}{\big(1-\mu_1(k_p,d_p)\big)^2} \times\left(2\mu_1(k_p,d_p)+\frac{\mu_1(k_p,d_p)^2}{\nu_2(k_p,d_p)}\right),$$

whenever $0<\mu_1(k_p,d_p)<1$.

Proof. r For the proof, refer to Appendix A. □

4. Numerical Investigations

In this section, for performance assessment of the proposed HD estimator $\hat{\boldsymbol{\beta}}^{\mathrm{HD}}$, we conduct a simulation study along with the analysis of real data.

4.1. Simulation

Here, we consider the multiple regression model with varying squared multiple correlation coefficient R^2 and error distribution, given by the following relation:

$$\mathbf{Y}=c\mathbf{X}\boldsymbol{\beta}+\sigma\boldsymbol{\epsilon},$$

where $\boldsymbol{\beta}=(\boldsymbol{\beta}_1,\mathbf{0})^\top$, $\boldsymbol{\beta}_1$ is the active set, and its dimension is $p_1=0.4p$. The absolute values of a normal distribution with mean 0 and standard deviation 5 is considered $\boldsymbol{\beta}_1$. The remaining $p-p_1$ components are zero.

In this example, motivated by McDonald and Galarneau [14], the explanatory variables are computed by

$$\mathbf{x}_j=\sqrt{1-\rho^2}\mathbf{z}_j+\rho\mathbf{z}_{p_1},\quad j=1,\ldots,p,$$

where the $\mathbf{z}_j$s are independent standard normal pseudo-random vectors, and ρ is specified such that the correlation between any two explanatory variables is given by ρ^2. Similarly to Zhu et al. [15], the variance is set to $\sigma^2=6.83$, and two different kinds of error distribution are taken for $\boldsymbol{\epsilon}$: (1) the standard normal is $\mathcal{N}_n(\mathbf{0},\mathbf{I}_n)$, and (2) standard t with 5 degrees of freedom $t_n(\mathbf{0},\mathbf{I}_n,5)$. The constant c is also varied to control the signal-to-noise ratio, and it is set to 0.5, 1, and 2 with the corresponding $R^2=20\%,50\%$ and 80%. R^2 represents the proportion of the variable for a dependent variable that is explained by an independent variable or variables in a regression model.

We consider $\rho\in\{0.8,0.95\}$; the sample size and the number of covariates are set to $n\in\{30,50,100\}$, $p\in\{256,512,1024\}$, respectively. Following regularity conditions (**A1**)–(**A4**), we set $k_p=\sqrt{p}$. For $\delta=0.25=1/4$, we take $d_p=p^{-1/5}$, which guarantees (**A4**). We then simulate $\hat{\boldsymbol{\beta}}^{\mathrm{HD}}$ and $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}$ 100 times using Equation (9) and $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}=(\mathbf{X}^\top\mathbf{X}+k_p\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{Y}$.

For comparison purposes, the quadratic bias (QB) and mean squared error (MSE) are computed according to

$$\mathrm{QB}(\hat{\boldsymbol{\beta}}^*) = \frac{1}{100}\sum_{j=1}^{100}(\hat{\boldsymbol{\beta}}_j^* - \beta)^\top(\hat{\boldsymbol{\beta}}_j^* - \beta), \quad \text{and} \quad \mathrm{MSE}(\hat{\boldsymbol{\beta}}^*) = \frac{1}{100}\sum_{j=1}^{100}(\hat{\boldsymbol{\beta}}_j^* - \boldsymbol{\beta})^\top(\hat{\boldsymbol{\beta}}_j^* - \boldsymbol{\beta}),$$

respectively, where $\hat{\boldsymbol{\beta}}^*$ is one of $\hat{\boldsymbol{\beta}}^{\mathrm{HD}}$ or $\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}$.

4.2. Review of Results

In Theorem 2, the condition for which the proposed $\hat{\boldsymbol{\beta}}^{\mathrm{HD}}$ is unbiased is investigated based on the eigenvalues of $\mathbf{X}^\top\mathbf{X}$. Here, we numerically analyze the biasedness of this estimator by comparing the ridge estimator concerning the parameters of the model. For this purpose, the difference in QB is reported in Table 1 by evaluating

$$\mathrm{diff} = \mathrm{QB}(\hat{\boldsymbol{\beta}}^{\mathrm{HD}}) - \mathrm{QB}(\hat{\boldsymbol{\beta}}^{\mathrm{Ridge}}).$$

If diff is positive, then the quadratic bias of the proposed estimator is larger than that of the ridge estimator.

Table 1. The difference between quadratic biases of the high dimensional and ridge estimators.

			$\rho = 0.8$		$\rho = 0.95$	
			$\mathcal{N}(0, \mathbf{I}_n)$	$t_n(0, \mathbf{I}_n, 5)$	$\mathcal{N}(0, \mathbf{I}_n)$	$t_n(0, \mathbf{I}_n, 5)$
p	c	n	**diff**	**diff**	**diff**	**diff**
256	0.5	30	5.7657	5.7643	10.0535	10.1134
		50	6.4911	6.4941	11.4722	11.5088
		100	17.8314	17.8493	30.1008	30.4137
	1	30	22.9671	487.4169	39.6556	459.3473
		50	25.8621	522.6298	45.1138	480.5501
		100	70.8693	798.6551	118.1326	676.1919
	2	30	91.7526	2413.9746	158.0026	2256.1664
		50	103.3996	2587.7382	179.8509	2357.4111
		100	283.0549	3922.0057	470.4114	3259.2211
512	0.5	30	3.1943	3.2012	6.5528	6.6001
		50	4.4800	4.4781	9.5861	9.6151
		100	10.2121	10.2489	20.1828	20.3366
	1	30	12.7657	926.7540	26.0911	916.2663
		50	17.8861	1009.3595	38.0549	969.4353
		100	40.7254	1192.4455	79.9094	1095.9628
	2	30	51.0605	4621.0862	104.2892	4555.0569
		50	71.5157	5029.3107	151.9461	4809.2595
		100	162.7616	5920.6337	318.7343	5397.7878
1024	0.5	30	1.7594	1.7584	3.7384	3.7410
		50	3.9188	3.9345	9.2523	9.3437
		100	5.1236	5.1189	12.6469	12.6455
	1	30	7.0318	1637.6798	14.8960	1636.5664
		50	15.6758	1804.8548	36.9649	1763.7468
		100	20.4564	1940.6091	50.2993	1856.0197
	2	30	28.1221	8181.4255	59.5312	8167.9835
		50	62.7157	9008.4246	147.8715	8781.1968
		100	81.7756	9682.7404	147.8715	9229.7803

To comprise the MSEs, we use the relative mean square error (RMSE) given by

$$\text{RMSE} = \frac{\text{MSE}(\hat{\boldsymbol{\beta}}^{\text{Ridge}})}{\text{MSE}(\hat{\boldsymbol{\beta}}^{\text{HD}})}.$$

The results are reported in Table 2. If RMSE > 1, then the proposed estimator has a smaller MSE compared to the ridge.

Table 2. The relative MSE of the high dimensional and ridge estimators.

			$\rho = 0.8$		$\rho = 0.95$	
			$\mathcal{N}(0, \mathbf{I}_n)$	$t_n(0, \mathbf{I}_n, 5)$	$\mathcal{N}(0, \mathbf{I}_n)$	$t_n(0, \mathbf{I}_n, 5)$
p	c	n	**RMSE**	**RMSE**	**RMSE**	**RMSE**
256	0.5	30	1.0050	1.0050	1.0140	1.0139
		50	1.0058	1.0058	1.0161	1.0160
		100	1.0222	1.0222	1.0543	1.0539
	1	30	1.0032	1.0032	1.0179	1.0178
		50	1.0039	1.0039	1.0209	1.0209
		100	1.0221	1.0220	1.0883	1.0876
	2	30	0.9816	0.9816	0.9852	0.9851
		50	0.9793	0.9793	0.9829	0.9829
		100	0.9434	0.9435	0.9587	0.9584
512	0.5	30	1.0011	1.0011	1.0031	1.0031
		50	1.0016	1.0016	1.0048	1.0048
		100	1.0041	1.0041	1.0119	1.0119
	1	30	1.0004	1.0004	1.0029	1.0029
		50	1.0007	1.0007	1.0048	1.0048
		100	1.0023	1.0023	1.0139	1.0139
	2	30	0.9948	0.9948	0.9924	0.9924
		50	0.9929	0.9929	0.9895	0.9895
		100	0.9843	0.9843	0.9810	0.9809
1024	0.5	30	1.0003	1.0003	1.0009	1.0009
		50	1.0007	1.0007	1.0022	1.0022
		100	1.0009	1.0009	1.0031	1.0031
	1	30	1.0001	1.0001	1.0006	1.0006
		50	1.0002	1.0002	1.0017	1.0017
		100	1.0003	1.0003	1.0025	1.0025
	2	30	0.9984	0.9984	0.9973	0.9973
		50	0.9964	0.9964	0.9933	0.9933
		100	0.9954	0.9954	0.9910	0.9911

Based on the results of Tables 1 and 2, the following conclusions are made:

(1) The performance of the estimators is affected by the number of observations (n), the number of variables (p), the signal to noise ratio (c), and the degree of multicollinearity (ρ).

(2) By increasing the degree of multicollinearity, ρ, although for both cases of error distributions, the QB of the proposed estimator increases for $c = 0.5$ and 1, its MSE decreases dramatically since the RMSE increases.

(3) The signal-to-noise shows the effect of $\boldsymbol{\beta}$ in the model. Lower values (less than 1) are a sign of model sparsity, since, when c is small, the proposed estimator performs better than the ridge. This is evidence that our estimator is a better candidate as an

alternative in sparse models in the MSE sense. However, the QB increases for large c values, which forces the model to overestimate the parameters.

(4) As p increases, although the proposed estimator is superior to the ridge in sparse models (small c values), the efficiency decreases. This is more evident when the ratio p/n becomes larger. This fact may come as poor performance of the proposed estimators, but our estimator is still preferred in high dimensions for sparse models.

(5) Obviously, as n increases, so does the RMSE; however, the QB becomes very large, and it is due to the nature of the proposed estimators because of its complicated form. It must be noted that this does not contradict the results of Theorem 2, since the simulation scheme does not obey the regularity condition.

(6) There is evidence of robustness for the distribution tail for sparse models, i.e., the QB and RMSE are the same for both normal and t distributions. However, as c increases, the QB of the proposed estimator explodes for the heavier tail distribution. This may be seen as a disadvantage of the proposed estimators, but even for large values of c, the RMSE stays the same, evidence of relatively small variance for the heavier tail distribution.

4.3. AML Data Analysis

This section assesses the performance of the proposed estimators using the mean prediction error (MPE) and MSE criteria of a data set adopted from Metzeler et al. [16], in which the information for 79 patients was collected. The data can be accessed from the Gene Expression Omnibus (GEO) data repository (http://www.ncbi.nlm.nih.gov/geo/ (accessed on 1 January 2021)) by the National Center for Biotechnology Information (NCBI), where the data is available under GEO accession number GSE12417. We only use the data set that was used as a test set. This contains gene expression data for 79 adult patients with cytogenetically normal acute myeloid leukemia (CN-AML), showing heterogeneous treatment outcomes. According to Sill et al. [17], we reduce the total number of 54,675 gene expression features that have been measured with the Affymetrix HG-U133 Plus 2.0 microarray technology to the top $p \in \{1000, 2000\}$ features with the largest variance across all 79 samples. We considered overall survival time based on month as the response variable. The condition number of the design matrix for the AML data set is approximately 1095.80, evident of severe multicollinearity among columns of the design matrix ([18], see p. 298). To find the optimum values of k and d, denoted by k_{opt} and d_{opt} for practical purposes, we use the GCV given by Equation (12). Hence, we use the following formulas:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{\text{HD}*} &= (\mathbf{X}^\top\mathbf{X} + k_{\text{opt}}\mathbf{I}_p)^{-1}(\mathbf{X}^\top\mathbf{Y} + k_{\text{opt}}d_{\text{opt}}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{Y})\\ \hat{\boldsymbol{\beta}}^{\text{Ridge}*} &= (\mathbf{X}^\top\mathbf{X} + k_{\text{opt}}\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{Y}.\end{aligned}$$

To compute the MPE and MSE, we divide the whole data set into two train ($\mathcal{T} = (\mathbf{X}^{\text{train}}, \mathbf{Y}^{\text{trian}})$) and validation ($\mathcal{V} = (\mathbf{X}^{\text{valid}}, \mathbf{Y}^{\text{valid}})$) sets, comprising 70% and 30%, respectively. Then, the measures are evaluated using

$$\begin{aligned}\text{MPE}_{\text{boot}}(\hat{\boldsymbol{\beta}}^*) &= \frac{1}{\text{N.boot}}\sum_{j=1}^{\text{N.boot}}(\mathbf{X}^{\text{valid}}\hat{\boldsymbol{\beta}}_j^{\text{train}*} - \mathbf{Y}^{\text{valid}})^\top(\mathbf{X}^{\text{valid}}\hat{\boldsymbol{\beta}}_j^{\text{train}*} - \mathbf{Y}^{\text{valid}}),\\ \text{MSE}_{\text{boot}}(\hat{\boldsymbol{\beta}}^*) &= \frac{1}{\text{N.boot}}\sum_{j=1}^{\text{N.boot}}(\hat{\boldsymbol{\beta}}_j^{\text{train}*} - \boldsymbol{\beta}^{\text{HD}*})^\top(\hat{\boldsymbol{\beta}}_j^{\text{train}*} - \boldsymbol{\beta}^{\text{HD}*}),\end{aligned}$$

where N.boot stands for the number of bootstrapped sample, $\hat{\boldsymbol{\beta}}^*$ is one of the proposed and ridge estimators, and $\hat{\boldsymbol{\beta}}^{\text{HD}*}$ is the assumed true parameter obtained by Equation (9) from the whole data set.

$$\text{RMPE}_{\text{boot}} = \frac{\text{MPE}_{\text{boot}}(\hat{\boldsymbol{\beta}}^{\text{Ridge}*})}{\text{MPE}_{\text{boot}}(\hat{\boldsymbol{\beta}}^{\text{HD}*})} \qquad \text{RMSE}_{\text{boot}} = \frac{\text{MSE}_{\text{boot}}(\hat{\boldsymbol{\beta}}^{\text{Ridge}*})}{\text{MSE}_{\text{boot}}(\hat{\boldsymbol{\beta}}^{\text{HD}*})}$$

The results are tabulated in Table 3 for the number of bootstrap N.boot = 200. The following conclusions are obtained from Table 3:

(1) Using the GCV, the proposed estimator is shown to be consistently superior to the ridge estimator, relative to RMSE and RMPE criteria.
(2) Similarly to the results of simulations, with growing p, the MSE of the proposed estimator increases compared to the ridge estimator. However, as p gets larger the mean prediction error becomes smaller, which shows the superiority for prediction purposes.

Further, Figure 2 depicts the MSE and MPE values for both HD and ridge estimators, for the case $p = 1000$. It is obvious that the high-dimensional estimator performs better compared to the ridge. For the case $p = 2000$, we obtained similar results.

Table 3. RMPE and RMSE values for 200 bootstrapped samples in the analysis of AML data

Criterion	$p = 1000$	$p = 2000$
$\text{RMPE}_{\text{boot}}$	1.001981	1.002278
$\text{RMSE}_{\text{boot}}$	1.046073	1.039997

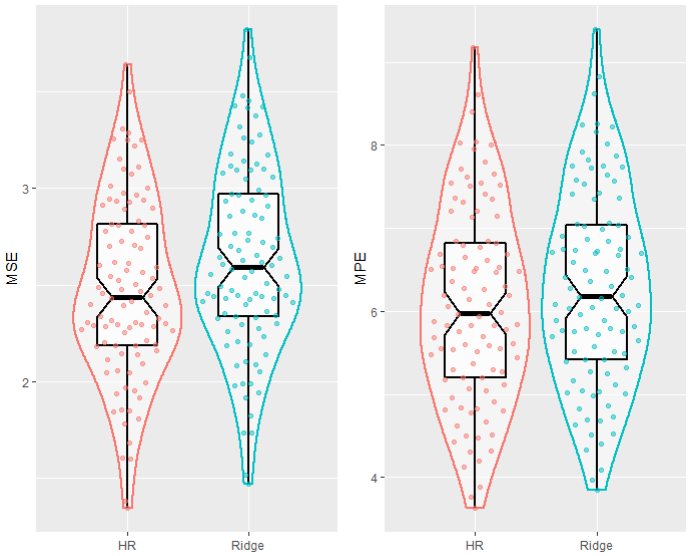

Figure 2. Box-plot of the MSE and MPE values for $p = 1000$ in the AML data.

5. Conclusions

In this note, we propose a high-dimensional two-parameter ridge estimator to the conventional ridge and Liu estimators. Its asymptotic properties have also been discussed. This estimator, via simulation and real-life experiments, is efficient in high dimensional problems and can potentially overcome multicollinearity. Additionally, the proposed high-dimensional ridge estimator yields superior performance in the small mean squared error sense.

Author Contributions: Conceptualization, M.A. and N.M.K.; methodology, M.A. and N.M.K.; validation, M.A., M.N., M.R. and N.M.K.; formal analysis, M.A., M.N., M.R. and N.M.K.; investigation, M.A., M.N., M.R. and N.M.K.; resources, M.N.; writing—original draft preparation, M.A.; writing—review and editing, M.A., M.N., M.R. and N.M.K.; visualization, M.A., M.N. and M.R.; supervision, M.A. and N.M.K.; project administration, M.A., M.N., M.R. and N.M.K.; funding acquisition, M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was based upon research supported, in part, by the visiting professor program, University of Pretoria, and the National Research Foundation (NRF) of South Africa, SARChI Research Chair UID: 71199; Reference: IFR170227223754 grant No. 109214. The work of M. Norouzirad

and M. Roozbeh is based on the research supported in part by the Iran National Science Foundation (INSF) (grant number 97018318). The opinions expressed and conclusions arrived at are those of the authors and are not necessarily attributed to the NRF.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this article may be simulated in R, using the stated seed value and parameter values. The real data set is available at http://www.ncbi.nlm.nih.gov/geo/ (accessed on 1 January 2021).

Acknowledgments: We would like to sincerely thank two anonymous reviewers for their constructive comments, which led us to put many details in the paper and improved the presentation.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proof of the Main Results

Proof of Theorem 1. By definition, we have

$$\begin{aligned}\operatorname{var}\left(\hat{\boldsymbol{\beta}}^{\mathrm{HD}}\right) &= \sigma^2\left(\mathbf{X}^\top\mathbf{X}+k_p\mathbf{I}_p\right)^{-1}\left(\mathbf{X}^\top+k_pd_p\mathbf{X}^+\right)\left(\mathbf{X}^\top+k_pd_p\mathbf{X}^+\right)^\top\left(\mathbf{X}^\top\mathbf{X}+k_p\mathbf{I}_p\right)^{-1}\\ &= \sigma^2\left(\frac{\mathbf{X}^\top\mathbf{X}}{k_p}+\mathbf{I}_p\right)^{-1}\left(\frac{\mathbf{X}^\top}{k_p}+d_p\mathbf{X}^+\right)\left(\frac{\mathbf{X}}{k_p}+d_p\mathbf{X}^{\top+}\right)\left(\frac{\mathbf{X}^\top\mathbf{X}}{k_p}+\mathbf{I}_p\right)^{-1} \quad \text{(A1)}\end{aligned}$$

By **(A1)**, $\mathbf{X}/k_p = O(1)k_p^{\delta-1} = o(1)$ and $\mathbf{X}^\top\mathbf{X}/k_p+\mathbf{I}_p \to \mathbf{I}_p$. By **(A2)**, $d_p\mathbf{X}^+ = O(1)d_p^{1-\eta} = o(1)$. Hence, $\operatorname{var}\left(\hat{\beta}_i^{\mathrm{HD}}\right) \to 0$ as $p \to \infty$, and the proof is complete. □

Proof of Theorem 2. By definition

$$\begin{aligned}\mathbb{E}\left(\hat{\boldsymbol{\beta}}^{\mathrm{HD}}\right) &= (\mathbf{X}^\top\mathbf{X}+k_p\mathbf{I}_p)^{-1}(\mathbf{X}^\top+k_pd_p\mathbf{X}^+)\mathbf{X}\boldsymbol{\beta}\\ &= \left(\frac{\mathbf{X}^\top\mathbf{X}}{k_p}+\mathbf{I}_p\right)^{-1}\left(\frac{\mathbf{X}^\top\mathbf{X}}{k_p}+d_p\mathbf{X}^+\mathbf{X}\right)\boldsymbol{\beta}\\ &= \left(\frac{\mathbf{X}^\top\mathbf{X}}{k_p}+\mathbf{I}_p\right)^{-1}\left(\frac{\mathbf{X}^\top\mathbf{X}}{k_p}\right)\boldsymbol{\beta}+d_p\mathbf{X}^+\mathbf{X}\boldsymbol{\beta}. \quad \text{(A2)}\end{aligned}$$

Under **(A2)**, $d_p\mathbf{X}^+\mathbf{X} = o(1)$. The proof is complete using Theorem 2 of Luo [11]. □

Proof of Theorem 3. We have

$$\begin{aligned}\frac{1}{d_p}(\hat{\boldsymbol{\beta}}^{\mathrm{HD}}-\boldsymbol{\beta}) &= \frac{1}{d_p}\left\{(\mathbf{X}^\top\mathbf{X}+k_p\mathbf{I}_p)^{-1}(\mathbf{X}^\top+k_pd_p\mathbf{X}^+)(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})-\boldsymbol{\beta}\right\}\\ &= \left(\frac{\mathbf{X}^\top\mathbf{X}}{k_p}+\mathbf{I}_p\right)^{-1}\left(\frac{\mathbf{X}^\top}{k_pd_p}+\mathbf{X}^+\right)\boldsymbol{\epsilon}\\ &\quad+\frac{1}{d_p}\left(\frac{\mathbf{X}^\top\mathbf{X}}{k_p}+\mathbf{I}_p\right)^{-1}(d_p\mathbf{X}^+\mathbf{X}-\mathbf{I}_p)\boldsymbol{\beta}.\end{aligned}$$

By **(A1)**, $\mathbf{X}^\top\mathbf{X}/k_p+\mathbf{I}_p \to \mathbf{I}_p$, by **(A2)**, $d_p\mathbf{X}^+\mathbf{X} = o(1)$, and by **(A4)**, $\mathbf{X}/k_pd_p = o(1)$. Hence,

$$\frac{1}{d_p}(\hat{\boldsymbol{\beta}}^{\mathrm{HD}}-\boldsymbol{\beta}) \to \mathbf{X}^+\boldsymbol{\epsilon}$$

The proof is complete. □

Proof of Theorem 4. It is straightforward to verify that

$$\mathrm{E}\Big(\mathrm{GCV}(\hat{\boldsymbol{\beta}}^{\mathrm{HD}})\Big) = \frac{v_1^2(k_p,d_p) + \sigma^2\big(1 - 2\mu_1(k_p,d_p) + v_2(k_p,d_p)\big)}{\big(1-\mu_1(k_p,d_p)\big)^2}.$$

Hence

$$\begin{aligned}\mathrm{E}\Big(\mathrm{R}(\boldsymbol{\beta};\hat{\boldsymbol{\beta}}^{\mathrm{HD}})\Big) - \mathrm{E}\Big(\mathrm{GCV}(\hat{\boldsymbol{\beta}}^{\mathrm{HD}})\Big) &= \mathrm{E}\Big(R(\hat{\boldsymbol{\beta}}^{\mathrm{HD}}(k_p,d_p);\boldsymbol{\beta})\Big)\left(1 - \frac{1}{\big(1-\mu_1(k_p,d_p)\big)^2}\right) \\ &\quad -\sigma^2\frac{1-2\mu_1(k_p,d_p)}{\big(1-\mu_1(k_p,d_p)\big)^2},\end{aligned}$$

which leads to the required result. □

References

1. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
2. Tikhonov, A.N. Solution of incorrectly formulated problems and the regularization method. *Sov. Math. Dokl.* **1963**, *4*, 1035–1038.
3. Saleh, A.K.M.E.; Arashi, M.; Kibria, B.M.G. *Theory of Ridge Regression Estimation with Applications*; John Wiley: Hoboken, NJ, USA, 2019.
4. Wang, X.; Dunson, D.; Leng, C. No penalty no tears: Least squares in high-dimensional models. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; pp. 1814–1822.
5. Bühlmann, P. Statistical significance in high-dimensional linear models. *Bernoulli* **2013**, *19*, 1212–1242. [CrossRef]
6. Shao, J.; Deng, X. Estimation in high-dimensional linear models with deterministic design matrices. *Ann. Stat.* **2012**, *40*, 812–831. [CrossRef]
7. Dicker, L.H. Ridge regression and asymptotic minimum estimation over spheres of growing dimension. *Bernoulli* **2016**, *22*, 1–37. [CrossRef]
8. Liu, K. A new class of biased estimate in linear regression. *Commun. Stat. Theory Methods* **1993**, *22*, 393–402.
9. Ozkale, M.R.; Kaciranlar, S. The restricted and unrestricted two-parameter estimators. *Commun. Stat. Theory Methods* **2007**, *36*, 2707–2725. [CrossRef]
10. Wang, X.; Leng, C. High dimensional ordinary least squares projection for screening variables. *J. R. Stat. Soc. Ser. B* **2015**. [CrossRef]
11. Luo, J. The discovery of mean square error consistency of ridge estimator. *Stat. Probab. Lett.* **2010**, *80*, 343–347. [CrossRef]
12. Amini, M.; Roozbeh, M. Optimal partial ridge estimation in restricted semiparametric regression models. *J. Multivar. Anal.* **2015**, *136*, 26–40. [CrossRef]
13. Akdeniz, F.; Roozbeh, M. Generalized difference-based weighted mixed almost unbiased ridge estimator in partially linear models. *Stat. Pap.* **2019**, *60*, 1717–1739. [CrossRef]
14. McDonald, G.C.; Galarneau, D.I. A Monte Carlo of Some Ridge-Type Estimators. *J. Am. Stat. Assoc.* **1975**, *70*, 407–416. [CrossRef]
15. Zhu, L.P.; Li, L.; Li, R.; Zhu, L.X. Model-free feature screening for ultrahigh dimensional data. *J. Am. Stat. Assoc.* **2011**, *106*, 1464–1475. [CrossRef] [PubMed]
16. Metzeler, K.H.; Hummel, M.; Bloomfield, C.D.; Spiekermann, K.; Braess, J.; Sauerl, M.C.; Heinecke, A.; Radmacher, M.; Marcucci, G.; Whitman, S.P.; et al. An 86 Probe Set Gene Expression Signature Predicts Survival in Cytogenetically Normal Acute Myeloid Leukemia. *Blood* **2008**, *112*, 4193–4201. [CrossRef] [PubMed]
17. Sill, M.; Hielscher, T.; Becker, N.; Zucknick, M. c060: Extended Inference for Lasso and Elastic-Net Regularized Cox and Generalized Linear Models; R Package Version 0.2-4; 2014. Available online: http://CRAN.R-project.org/package=c060 (accessed on 1 January 2021).
18. Montgomery, D.C.; Peck, E.A.; Vining, G.G. *Introduction to Linear Regression Analysis*, 5th ed.; Wiley: Hoboken, NJ, USA, 2012.