Biometrika (2020), **107**, 2, *pp.* 481–488
Printed in Great Britain

doi: 10.1093/biomet/asz074
Advance Access publication 27 January 2020

Miscellanea

Estimation of error variance via ridge regression

BY X. LIU

School of Statistics and Management, Shanghai University of Finance and Economics, 777 Guoding Road, Shanghai 200433, China
liu.xu@sufe.edu.cn

S. ZHENG

School of Mathematics & Statistics, Northeast Normal University, 5268 Renmin Street, Changchun 130024, China
zhengsr@nenu.edu.cn

AND X. FENG

School of Statistics and Management, Shanghai University of Finance and Economics, 777 Guoding Road, Shanghai 200433, China
feng.xingdong@sufe.edu.cn

SUMMARY

We propose a novel estimator of error variance and establish its asymptotic properties based on ridge regression and random matrix theory. The proposed estimator is valid under both low- and high-dimensional models, and performs well not only in nonsparse cases, but also in sparse ones. The finite-sample performance of the proposed method is assessed through an intensive numerical study, which indicates that the method is promising compared with its competitors in many interesting scenarios.

Some key words: High dimension; Random matrix theory; Ridge regression; Variance estimation.

1. INTRODUCTION

High-dimensional linear regression is an important tool in data analysis. In the classical cases where the covariate dimension is fixed and the sample size tends to infinity, the estimation of error variance in different regression models has been well studied. However, when the covariate dimension is large relative to the sample size, many classical methods become invalid in both theory and implementation.

Consider the high-dimensional linear regression model

$$y_i = x_i^{\mathrm{T}}\beta + \varepsilon_i \quad (i = 1, \ldots, n), \tag{1}$$

where y_i is the response of the ith individual, $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ is the p-dimensional covariate, $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the corresponding coefficient vector, and ε_i is the error, which is independent of x_i, with $\sigma^2 = \mathrm{var}(\varepsilon_i) < \infty$. The samples $\{(y_1, x_1^{\mathrm{T}})^{\mathrm{T}}, \ldots, (y_n, x_n^{\mathrm{T}})^{\mathrm{T}}\}$ are independent and identically distributed.

High-dimensional inferences and predictions have attracted increasing attention in recent years. For example, Zhong & Chen (2011) and Cui et al. (2017) proposed high-dimensional tests for regression coefficients, van de Geer et al. (2014) and Zhang & Zhang (2014) considered the confidence interval of each coefficient under high-dimensional regression models, and Dobriban & Wager (2018) studied the prediction of high-dimensional ridge regression with unit error variance. However, it is essential to have a satisfactory estimation of error variance for either inference or prediction.

Sun & Zhang (2012) proposed a scaled lasso for simultaneous estimation of regression coefficients and error variance. Fan et al. (2012) proposed a refitted cross-validation method to estimate error variance in high-dimensional regression models by splitting the data. Reid et al. (2016) studied the estimation of error variance based on the lasso method. However, in all of these methods sparsity is assumed for the coefficient vectors. Fan et al. (2012) claim that their method still works without the sparsity assumption, but they provided no theoretical evidence for this. Dicker (2014) presented an estimator based on the method of moments. A maximum likelihood method was introduced by Dicker & Erdogdu (2016) for cases with normally distributed noise. Janson et al. (2017) estimated the error variance by solving a convex optimization problem, and Verzelen & Gassiat (2018) considered an adaptive method, with both methods assuming known covariance structures for covariates and normally distributed noise. Yu & Bien (2019) used a lasso-based maximum likelihood method to estimate the error variance under the assumption of sparse coefficients and Gaussian noise.

In this note, we consider estimation of the error variance σ^2, which is fundamental, but still challenging in high-dimensional regression models. The main contribution of this work can be summarized as follows. We propose an estimator of the error variance under the model (1) based on ridge regression and random matrix theory, and we discuss its connection with the classical estimator. The proposed estimator outperforms its competitors in cases without sparsity, as demonstrated in our simulation study. The theoretical properties of the proposed estimator are established, with detailed proofs and further numerical results given in the Supplementary Material.

2. ESTIMATION

Under model (1), the ridge regression estimator (Hoerl & Kennard, 1970) is obtained by considering the optimization problem

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^{\mathrm{T}}\beta)^2 + \eta\|\beta\|^2,$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^p$ and η is a positive tuning parameter. The estimate can be expressed in the explicit form

$$\hat{\beta} = n^{-1}(S_n + \eta I_p)^{-1}X^{\mathrm{T}}Y,$$

where $X = (x_1, \ldots, x_n)^{\mathrm{T}}$, $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$, $S_n = n^{-1}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}$ and I_p is the $p \times p$ identity matrix. The average of squares of residues is then

$$\begin{aligned} n^{-1}(Y - X\hat{\beta})^{\mathrm{T}}(Y - X\hat{\beta}) &= n^{-1}Y^{\mathrm{T}}(I_p - A_{1n})^2 Y \\ &= \check{\sigma}^2 - \eta n^{-2}Y^{\mathrm{T}}X(S_n + \eta I_p)^{-2}XY, \end{aligned}$$

where $\check{\sigma}^2 = n^{-1}Y^{\mathrm{T}}(I_n - A_{1n})Y$ and $A_{1n} = n^{-1}X(S_n + \eta I_p)^{-1}X^{\mathrm{T}}$. As indicated in the proof of Theorem 1, we have

$$\{1 - n^{-1}\mathrm{tr}(A_{1n})\}^{-1}\check{\sigma}^2 = \sigma^2 + o_{\mathrm{p}}(1), \quad \eta n^{-2}Y^{\mathrm{T}}X(S_n + \eta I_p)^{-2}XY = o_{\mathrm{p}}(1)$$

under certain conditions. Hence,

$$n^{-1}(Y - X\hat{\beta})^{\mathrm{T}}(Y - X\hat{\beta}) = \{1 - n^{-1}\mathrm{tr}(A_{1n})\}\sigma^2 + o_{\mathrm{p}}(1).$$

Therefore, the error variance σ^2 can be estimated by

$$\hat{\sigma}^2 = \{1 - n^{-1}\mathrm{tr}(A_{1n})\}^{-1}\breve{\sigma}^2. \tag{2}$$

Before summarizing the asymptotic properties of the estimate $\hat{\sigma}^2$ in Theorem 1, we introduce some assumptions.

Assumption 1. We have that $\tau_n = p/n \to \tau \in [0, +\infty)$ as $n \to \infty$.

Assumption 2. The fourth moment of the model error, $E(\varepsilon_i^4)$, is finite.

Assumption 3. The covariate x_i follows the component structure $x_i = \mu + \Sigma^{1/2} z_i$, where μ and Σ are an unknown vector and matrix, respectively, and $z_i = (z_{i1}, \ldots, z_{ip})^{\mathrm{T}}$ is such that $E(z_i) = 0_p$ and $\mathrm{var}(z_i) = I_p$. Furthermore, $E(z_{\ell i}^4) = 3 + \Delta$ for some constant Δ, and $E(z_{1j_1}^{\ell_1} \cdots z_{1j_d}^{\ell_d}) = E(z_{1j_1}^{\ell_1}) \cdots E(z_{1j_d}^{\ell_d})$ for unequal positive integers $j_1, \ldots, j_d$ and nonnegative integers $\ell_1, \ldots, \ell_d$ with $\sum_{v=1}^{d} \ell_v \leqslant 4$.

Assumption 4. There exist two positive bounds C_1 and c_1 such that $C_1 \geqslant \limsup_{p\to\infty} \lambda_{\max}(\Sigma) \geqslant \liminf_{p\to\infty} \lambda_{\min}(\Sigma) \geqslant c_1$, where $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)$ denote the maximum and minimum eigenvalues of Σ, respectively.

Assumption 5. The limit of $n^{-1}\mathrm{tr}(A_{1n} \circ A_{1n})$ exists, and $\mathrm{tr}(A_{1n}^2 \circ A_{1n}^2) = o_{\mathrm{p}}[\{\mathrm{tr}(A_{1n}^2) - \mathrm{tr}(A_{1n} \circ A_{1n})\}^2]$ where $\mathrm{tr}(A \circ B) = \sum_{j=1}^{p} e_j^{\mathrm{T}} A e_j e_j^{\mathrm{T}} B e_j$ with A and B being any two $p \times p$ matrices and e_j the jth column of I_p.

Remark 1. Assumption 1 specifies the convergence regime of the covariate dimension p and the sample size n. Assumption 2 imposes a moment constraint on the model errors. Assumptions 3 and 4 were considered for the factor structure by Bai & Saranadasa (1996), Chen & Qin (2010), Zhong & Chen (2011), Wang et al. (2015) and Cui et al. (2017). In Assumption 5, assuming the existence of the limit of $n^{-1}\mathrm{tr}(A_{1n} \circ A_{1n})$ is common in random matrix theory, and the assumption that $\mathrm{tr}(A_{1n}^2 \circ A_{1n}^2) = o_{\mathrm{p}}[\{\mathrm{tr}(A_{1n}^2) - \mathrm{tr}(A_{1n} \circ A_{1n})\}^2]$ is similar to the condition $\mathrm{tr}(\Sigma^4) = o\{\mathrm{tr}^2(\Sigma^2)\}$ considered by Bai & Saranadasa (1996), Chen & Qin (2010) and Zhong & Chen (2011).

THEOREM 1. *Suppose that Assumptions 1–5 hold and $\eta\|\beta\|^2 = o(1)$. Then $\hat{\sigma}^2 \to \sigma^2$ in probability and*

$$n^{1/2}\sigma_{\hat{\sigma}}^{-1}\left\{\hat{\sigma}^2 - \sigma^2 - \frac{\eta\beta^{\mathrm{T}} S_n (S_n + \eta I_p)^{-1}\beta}{1 - n^{-1}\mathrm{tr}(A_{1n})}\right\} \to N(0, 1)$$

in law, where

$$\sigma_{\hat{\sigma}}^2 = \frac{\tau\{E(\varepsilon_1^4) - 3\sigma^4\}}{\{1 - \tau + \tau\eta m(\eta)\}^2} \lim_{n\to\infty} p^{-1}\mathrm{tr}(A_{1n} \circ A_{1n}) + \frac{E(\varepsilon_1^4) - \sigma^4}{\{1 - \tau + \tau\eta m(\eta)\}^2}[1 - 2\tau\{1 - \eta m(\eta)\}]$$
$$+ \frac{2\tau\sigma^4}{\{1 - \tau + \tau\eta m(\eta)\}^2}\left\{1 - 2\eta m(\eta) - \eta^2\frac{\mathrm{d}m(\eta)}{\mathrm{d}\eta}\right\},$$

with $m(\eta)$ being the Stieltjes transform of the limiting spectral distribution of S_n given in Lemma 3 *of the Supplementary Material. Furthermore, if $n^{1/2}\eta\|\beta\|^2 = o(1)$, then*

$$n^{1/2}\sigma_{\hat{\sigma}}^{-1}(\hat{\sigma}^2 - \sigma^2) \to N(0, 1)$$

in law.

Remark 2. If the error ε_i is normally distributed, then we have $E(\varepsilon_i^4) = 3\sigma^2$, which implies that the asymptotic variance of $\hat{\sigma}^2$ can be simplified to

$$\sigma_{\hat{\sigma}}^2 = \frac{2\sigma^4}{\{1 - \tau + \tau\eta m(\eta)\}^2}\left\{1 - \tau - \eta^2\tau\frac{\mathrm{d}m(\eta)}{\mathrm{d}\eta}\right\};$$

it then follows from the Supplementary Material that the variance $\sigma^2_{\hat\sigma}$ can be consistently estimated by

$$\hat\sigma^2_{\hat\sigma} = \frac{2\hat\sigma^4}{\{1 - n^{-1}\mathrm{tr}(A_{1n})\}^2}\left\{1 - 2n^{-1}\mathrm{tr}(A_{1n}) + n^{-1}\mathrm{tr}(A^2_{1n})\right\}.$$

Remark 3. Consider the case of $\tau \in [0, 1)$ where $p < n$. When the matrix S_n is invertible, the proposed estimator is the same as the classical estimator of the error variance σ^2, denoted by $\hat\sigma^2_0$, upon setting the tuning parameter η to zero in (2). Furthermore, if the error ε_i is normally distributed and $\Sigma = \sigma^2_x I_p$ where σ_x is some positive constant, then it follows from Lemma 4 of the Supplementary Material that

$$\lim_{\eta\to 0}\sigma^2_{\hat\sigma} = \frac{2\sigma^4}{1-\tau},$$

which happens to be the variance of $n^{1/2}\hat\sigma^2_0$ if $\tau = p/n$. The mean squared error of the proposed estimate $\hat\sigma^2$ is smaller than that of the classical estimate $\hat\sigma^2_0$ if $n\eta\sigma^{-4}\|\beta\|^4 + 4\sigma^{-2}\|\beta\|^2 < 4\sigma_x^{-2}(1-\tau)^{-1}$ for sufficiently large n and small η.

However, if $\tau = 1$ where $p = n$, we have

$$\sigma^2_{\hat\sigma} = \frac{\sigma^4 p^{-1}\mathrm{tr}\{(S_n + \eta I_p)^{-2}\}}{[p^{-1}\mathrm{tr}\{(S_n + \eta I_p)^{-1}\}]^2}.$$

Fan et al. (2012) obtained similar asymptotic properties based on sure screening methods under a sparsity assumption, which is not necessary in our method. Numerical comparisons are conducted in the next section, and they indicate that our proposed method is much more efficient than its competitors when the sparsity assumption is strongly violated.

3. SIMULATION STUDY

We carry out a simulation study to assess the finite-sample performance of the proposed method. To save space, only some of the results are reported here; the remaining results are summarized in the Supplementary Material.

As in Zhong & Chen (2011), the covariates x_{ij} are generated from the model $x_{ij} = \sum_{t=1}^{T}\rho_t z_{i(j+t-1)}$ $(j = 1, \ldots, p)$ for some $T < p$, where $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ and $z_i = (z_{i1}, \ldots, z_{i(p+T-1)})^{\mathrm{T}} \sim N(0, I_{p+T-1})$. The parameters $\{\rho_t\}_{t=1}^{\mathrm{T}}$ are independently generated from the uniform distribution $\mathrm{Un}(0, 1)$. The covariate vector x_i follows the p-dimensional normal distribution with covariance matrix

$$\Sigma_1 = \left\{\sum_{k=1}^{T-|j-l|}\rho_k\rho_{k+|j-l|}I(|j-l| < T)\right\}_{j,l=1}^{p}.$$

The responses $\{y_i\}_{i=1}^{n}$ are generated from the linear regression model

$$y_i = x_i^{\mathrm{T}}\beta + \varepsilon_i \quad (i = 1, \ldots, n),$$

where the errors $\varepsilon_1, \ldots, \varepsilon_n$ are independently drawn from two distributions: the normal distribution $N(0, \sigma^2)$ and the distribution $(0.5\sigma^2)^{1/2}t_4$, where t_4 refers to the t distribution with four degrees of freedom. We repeatedly generate 1000 datasets in each experiment of this simulation study.

In this study, we consider the sparse case with $\beta_1 = \beta_2 = \|\beta\|/\surd 2$ and $\beta_3 = \cdots = \beta_p = 0$, and the nonsparse cases with $\beta_1 = \cdots = \beta_{\alpha_k p} = (\alpha_k p)^{1/2}\|\beta\|$ and $\beta_{\alpha_k p+1} = \cdots = \beta_p = 0$ for $\alpha_k = 0.1, 0.3, 0.5, 0.8$ and 1.

To obtain the ridge regression estimator, we need to invert the $p \times p$ matrix $S_n + \eta I_p$, which is time-consuming when p is much greater than n. We use the Woodbury matrix identity to speed up its computation (Hager, 1989), that is,

$$(X^{\mathrm{T}}X/n + \eta I_p)^{-1} = \eta^{-1}\{I_p - X^{\mathrm{T}}(XX^{\mathrm{T}}/n + \eta I_n)^{-1}X/n\}.$$

Motivated by the method of Friedman et al. (2010), which was originally proposed for the lasso, the tuning parameter used in the proposed method is

$$\eta = \alpha \max_{1 \leqslant j \leqslant p} \left| x_{(j)}^{\mathrm{T}} Y \right| / (np),$$

where α is a preset constant, $x_{(j)} = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}$ is the jth column of X $(j = 1, \ldots, p)$ and $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$. From experience, we suggest taking $\alpha = 0.1$, although the performance of the proposed method is robust to a wide range of choices of this constant, as demonstrated in the Supplementary Material.

In this setting, the method of Verzelen & Gassiat (2018) frequently fails in its computation with their suggested estimates of precision matrices when the true covariances are unknown. Therefore, we compare our proposed method with five different methods: the refitted cross-validation method of Fan et al. (2012), the maximum likelihood method of Dicker & Erdogdu (2016), the moment-based method of Dicker (2014), the method of Janson et al. (2017) based on eigenvalues, and the scaled lasso proposed by Sun & Zhang (2012).

As Table 1 shows, the performance of our method is comparable to that of the refitted cross-validation method in those cases with relatively small dimensions p and weak signals, where the advantage of the lasso is not quite obvious yet. In extremely sparse scenarios, the refitted cross-validation method performs well. The proposed method performs quite robustly with respect to the sparsity rate $1 - \|\beta\|_0/p$, where $\|\cdot\|_0$ denotes the L_0-norm, but the performance of the refitted cross-validation method becomes increasingly poor as the sparsity rate decreases, because it is challenging for the lasso to capture all signals when the sparsity assumption is strongly violated. The method of Dicker & Erdogdu (2016) performs relatively poorly compared with the proposed method and the refitted cross-validation method in this numerical study. In many cases where $p \gg n$, the methods of Dicker (2014) and Janson et al. (2017) cannot guarantee positive estimates, owing to the absence of numerically satisfactory estimates of precision matrices since n is not large.

A similar conclusion can be drawn based on the results reported in Table 2, since the errors ε_i are generated from $(0.5\sigma^2)^{1/2}t_4$. Obviously, Assumption 2 is violated in this case, but the performance of the proposed method is still satisfactory.

When $\Sigma_1 = I_p$, the methods of Janson et al. (2017) and Sun & Zhang (2012) often underestimate the error variances in cases with weak signals, while the other methods perform well; see the Supplementary Material.

4. Discussion

The proposed estimator is useful for constructing confidence and prediction intervals that can be directly related to error variance. Moreover, it is likely to be helpful in testing the hypothesis $H_0 : C\beta = 0_r$ versus $H_1 : C\beta \neq 0_r$, where C is a known matrix of full row rank with dimension $r \times p$ and 0_r is an r-dimensional zero vector. Under high-dimensional models, the rank r may be allowed to go to infinity as the sample size increases. This inference problem for high-dimensional regression models has rarely been touched upon in the current literature. Motivated by the present work, one can adopt a similar idea to account for correlations of covariates, and develop reasonable tests for the hypothesis under more general model settings.

Table 1. *Averages and standard errors (in parentheses) in the cases where* $\varepsilon_i \sim N(0,\sigma^2)$, $x_i \sim N(0,\Sigma_1)$ *and* $T = 10$

(n,σ^2)	p	$\lVert\beta\rVert^2$ (RidgeVar)		$\lVert\beta\rVert^2$ (RCV)		$\lVert\beta\rVert^2$ (MLE)		$\lVert\beta\rVert^2$ (MM)		$\lVert\beta\rVert^2$ (EigenPrism)		$\lVert\beta\rVert^2$ (Scalreg)	
		0.025	0.1	0.025	0.1	0.025	0.1	0.025	0.1	0.025	0.1	0.025	0.1
							SR $= 1 - 2/p$						
(60, 1)	100	1.015	1.049	1.106	1.084	1.186	1.715	1.059	1.203	1.019	1.119	0.685	0.681
		(0.045)	(0.045)	(0.030)	(0.035)	(0.027)	(0.040)	(0.030)	(0.038)	(0.030)	(0.035)	(0.022)	(0.022)
	400	1.097	1.359	1.126	1.131	1.178	1.703	1.050	1.199	0.994	1.112	0.229	0.222
		(0.034)	(0.038)	(0.030)	(0.038)	(0.028)	(0.041)	(0.041)	(0.055)	(0.041)	(0.054)	(0.015)	(0.015)
(60, 3)	100	3.003	3.046	3.144	3.315	3.149	3.482	3.046	3.119	2.965	2.999	2.267	2.260
		(0.128)	(0.128)	(0.083)	(0.092)	(0.077)	(0.086)	(0.087)	(0.093)	(0.087)	(0.090)	(0.068)	(0.068)
	400	3.069	3.342	3.072	3.306	3.082	3.405	2.998	3.064	2.858	2.907	1.278	1.270
		(0.091)	(0.096)	(0.081)	(0.088)	(0.072)	(0.079)	(0.111)	(0.121)	(0.111)	(0.119)	(0.057)	(0.057)
							SR $= 90\%$						
(60, 1)	100	1.018	1.057	1.137	1.140	1.559	3.212	0.865	0.431	0.828	0.353	0.649	0.622
		(0.045)	(0.045)	(0.036)	(0.036)	(0.037)	(0.076)	(0.037)	(0.088)	(0.032)	(0.051)	(0.021)	(0.021)
	400	1.215	1.828	1.458	1.966	1.711	3.834	0.650	−0.406	0.603	−0.448	0.172	0.113
		(0.035)	(0.041)	(0.045)	(0.083)	(0.041)	(0.092)	(0.062)	(0.168)	(0.056)	(0.133)	(0.013)	(0.011)
(60, 3)	100	3.005	3.057	3.270	3.485	3.394	4.446	2.890	2.507	2.813	2.411	2.216	2.152
		(0.128)	(0.127)	(0.090)	(0.109)	(0.083)	(0.109)	(0.090)	(0.114)	(0.087)	(0.094)	(0.067)	(0.066)
	400	3.199	3.870	3.378	4.192	3.448	4.871	2.654	1.661	2.533	1.594	1.209	1.087
		(0.091)	(0.098)	(0.090)	(0.127)	(0.080)	(0.114)	(0.121)	(0.188)	(0.117)	(0.160)	(0.057)	(0.057)
							SR $= 70\%$						
(60, 1)	100	1.018	1.057	1.265	1.349	1.699	3.771	0.666	−0.358	0.681	−0.227	0.613	0.542
		(0.045)	(0.045)	(0.042)	(0.054)	(0.040)	(0.090)	(0.046)	(0.133)	(0.036)	(0.081)	(0.020)	(0.019)
	400	1.217	1.841	1.595	3.049	1.736	3.946	0.613	−0.561	0.568	−0.596	0.155	0.070
		(0.034)	(0.040)	(0.046)	(0.098)	(0.041)	(0.093)	(0.066)	(0.182)	(0.059)	(0.152)	(0.013)	(0.009)
(60, 3)	100	3.007	3.062	3.341	3.795	3.498	4.880	2.704	1.757	2.703	1.966	2.180	2.047
		(0.128)	(0.127)	(0.091)	(0.122)	(0.084)	(0.116)	(0.094)	(0.146)	(0.088)	(0.104)	(0.066)	(0.064)
	400	3.205	3.896	3.420	4.587	3.474	4.971	2.607	1.490	2.495	1.451	1.198	1.034
		(0.090)	(0.096)	(0.092)	(0.133)	(0.081)	(0.114)	(0.125)	(0.200)	(0.120)	(0.171)	(0.056)	(0.056)
							SR $= 50\%$						
(60, 1)	100	1.017	1.057	1.337	1.553	1.729	3.884	0.621	−0.530	0.653	−0.336	0.597	0.498
		(0.045)	(0.045)	(0.044)	(0.065)	(0.041)	(0.093)	(0.048)	(0.142)	(0.037)	(0.087)	(0.020)	(0.018)
	400	1.219	1.846	1.635	3.338	1.747	3.976	0.602	−0.601	0.567	−0.601	0.150	0.063
		(0.034)	(0.039)	(0.047)	(0.104)	(0.041)	(0.095)	(0.068)	(0.192)	(0.062)	(0.159)	(0.013)	(0.010)
(60, 3)	100	3.007	3.063	3.386	3.962	3.528	4.991	2.650	1.555	2.675	1.860	2.162	1.987
		(0.128)	(0.127)	(0.093)	(0.133)	(0.086)	(0.122)	(0.096)	(0.156)	(0.089)	(0.107)	(0.066)	(0.065)
	400	3.207	3.899	3.428	4.680	3.474	4.985	2.598	1.446	2.490	1.420	1.196	1.019
		(0.091)	(0.097)	(0.092)	(0.132)	(0.082)	(0.118)	(0.125)	(0.204)	(0.119)	(0.173)	(0.057)	(0.057)
							SR $= 20\%$						
(60, 1)	100	1.018	1.058	1.402	1.882	1.743	3.946	0.601	−0.620	0.640	−0.396	0.585	0.458
		(0.045)	(0.045)	(0.046)	(0.076)	(0.040)	(0.092)	(0.049)	(0.146)	(0.037)	(0.090)	(0.020)	(0.018)
	400	1.221	1.851	1.675	3.513	1.757	4.014	0.586	−0.667	0.552	−0.671	0.147	0.058
		(0.034)	(0.039)	(0.046)	(0.106)	(0.041)	(0.094)	(0.067)	(0.185)	(0.061)	(0.158)	(0.013)	(0.010)
(60, 3)	100	3.007	3.063	3.566	4.409	3.766	5.987	2.481	0.918	2.511	1.290	1.780	1.517
		(0.128)	(0.128)	(0.100)	(0.151)	(0.089)	(0.140)	(0.098)	(0.189)	(0.087)	(0.121)	(0.057)	(0.054)
	400	3.208	3.901	3.643	5.555	3.724	5.940	2.472	0.958	2.365	0.940	0.392	0.212
		(0.090)	(0.097)	(0.098)	(0.159)	(0.087)	(0.139)	(0.139)	(0.256)	(0.129)	(0.215)	(0.035)	(0.028)
							SR $= 0\%$						
(60, 1)	100	1.018	1.058	1.420	2.019	1.746	3.958	0.609	−0.594	0.641	−0.395	0.580	0.442
		(0.045)	(0.045)	(0.047)	(0.079)	(0.040)	(0.090)	(0.047)	(0.141)	(0.037)	(0.089)	(0.020)	(0.018)
	400	1.223	1.859	1.679	3.575	1.768	4.044	0.579	−0.691	0.544	−0.694	0.147	0.058
		(0.034)	(0.039)	(0.046)	(0.110)	(0.041)	(0.093)	(0.067)	(0.188)	(0.061)	(0.159)	(0.013)	(0.010)
(60, 3)	100	3.007	3.063	3.571	4.528	3.771	6.002	2.480	0.917	2.510	1.290	1.778	1.501
		(0.128)	(0.128)	(0.099)	(0.153)	(0.090)	(0.142)	(0.100)	(0.193)	(0.088)	(0.123)	(0.057)	(0.054)
	400	3.209	3.905	3.651	5.572	3.722	5.938	2.480	0.976	2.369	0.946	0.392	0.215
		(0.090)	(0.097)	(0.097)	(0.161)	(0.087)	(0.141)	(0.139)	(0.256)	(0.129)	(0.216)	(0.035)	(0.028)

RidgeVar, the proposed method; RCV, method of Fan et al. (2012); MLE, method of Dicker & Erdogdu (2016); MM, method of Dicker (2014); EigenPrism, method of Janson et al. (2017); Scalreg, method of Sun & Zhang (2012); SR, sparsity rate given by SR $= 1 - \lVert\beta\rVert_0/p$ where $\lVert\cdot\rVert_0$ is the L_0-norm.

Table 2. *Averages and standard errors (in parentheses) in the cases where $\varepsilon_i \sim (0.5\sigma^2)^{1/2} t_4$, $x_i \sim N(0, \Sigma_1)$ and $T = 10$*

(n, σ^2)	p	$\|\beta\|^2$ (RidgeVar)		$\|\beta\|^2$ (RCV)		$\|\beta\|^2$ (MLE)		$\|\beta\|^2$ (MM)		$\|\beta\|^2$ (EigenPrism)		$\|\beta\|^2$ (Scalreg)	
		0.025	0.1	0.025	0.1	0.025	0.1	0.025	0.1	0.025	0.1	0.025	0.1
							SR $= 1 - 2/p$						
(60, 1)	100	1.035	1.067	1.134	1.113	1.219	1.758	1.083	1.224	1.048	1.144	0.703	0.699
		(0.116)	(0.116)	(0.140)	(0.139)	(0.136)	(0.141)	(0.140)	(0.144)	(0.136)	(0.138)	(0.090)	(0.090)
	400	1.081	1.346	1.113	1.116	1.168	1.697	1.058	1.212	1.004	1.128	0.217	0.212
		(0.072)	(0.075)	(0.074)	(0.083)	(0.072)	(0.078)	(0.079)	(0.087)	(0.077)	(0.084)	(0.019)	(0.019)
(60, 3)	100	2.971	3.013	3.083	3.256	3.108	3.599	2.990	3.097	2.906	2.958	1.888	1.872
		(0.214)	(0.215)	(0.165)	(0.177)	(0.163)	(0.166)	(0.171)	(0.176)	(0.169)	(0.172)	(0.113)	(0.113)
	400	3.032	3.293	3.101	3.359	3.119	3.607	2.958	3.063	2.826	2.903	0.482	0.465
		(0.159)	(0.161)	(0.162)	(0.172)	(0.156)	(0.161)	(0.170)	(0.179)	(0.167)	(0.174)	(0.044)	(0.043)
							SR = 90%						
(60, 1)	100	1.037	1.076	1.162	1.166	1.601	3.273	0.889	0.450	0.853	0.370	0.671	0.643
		(0.116)	(0.116)	(0.151)	(0.136)	(0.136)	(0.150)	(0.140)	(0.162)	(0.134)	(0.139)	(0.088)	(0.086)
	400	1.196	1.811	1.431	1.884	1.700	3.826	0.659	−0.390	0.620	−0.418	0.167	0.112
		(0.072)	(0.076)	(0.078)	(0.112)	(0.078)	(0.113)	(0.094)	(0.185)	(0.087)	(0.153)	(0.019)	(0.016)
(60, 3)	100	2.976	3.029	3.246	3.369	3.479	5.068	2.762	2.199	2.687	2.093	1.830	1.752
		(0.214)	(0.214)	(0.169)	(0.191)	(0.164)	(0.182)	(0.173)	(0.201)	(0.169)	(0.178)	(0.112)	(0.109)
	400	3.167	3.832	3.528	4.570	3.657	5.766	2.461	1.071	2.360	1.035	0.411	0.295
		(0.160)	(0.163)	(0.172)	(0.218)	(0.162)	(0.194)	(0.184)	(0.270)	(0.177)	(0.235)	(0.043)	(0.038)
							SR = 70%						
(60, 1)	100	1.038	1.077	1.282	1.355	1.732	3.813	0.693	−0.331	0.713	−0.197	0.635	0.566
		(0.116)	(0.116)	(0.122)	(0.129)	(0.137)	(0.156)	(0.143)	(0.192)	(0.135)	(0.155)	(0.087)	(0.084)
	400	1.200	1.827	1.590	3.043	1.730	3.948	0.611	−0.568	0.571	−0.593	0.149	0.070
		(0.072)	(0.076)	(0.079)	(0.125)	(0.078)	(0.111)	(0.093)	(0.189)	(0.086)	(0.160)	(0.018)	(0.013)
(60, 3)	100	2.975	3.029	3.377	3.748	3.629	5.676	2.506	1.189	2.530	1.464	1.781	1.618
		(0.214)	(0.213)	(0.175)	(0.206)	(0.166)	(0.192)	(0.179)	(0.238)	(0.172)	(0.191)	(0.112)	(0.107)
	400	3.174	3.858	3.615	5.286	3.696	5.913	2.403	0.853	2.304	0.828	0.395	0.239
		(0.162)	(0.166)	(0.176)	(0.215)	(0.166)	(0.201)	(0.188)	(0.283)	(0.179)	(0.245)	(0.043)	(0.037)
							SR = 50%						
(60, 1)	100	1.038	1.077	1.344	1.593	1.758	3.918	0.648	−0.505	0.684	−0.308	0.619	0.522
		(0.116)	(0.116)	(0.122)	(0.141)	(0.138)	(0.160)	(0.144)	(0.197)	(0.135)	(0.156)	(0.086)	(0.081)
	400	1.202	1.832	1.641	3.340	1.739	3.975	0.601	−0.605	0.567	−0.604	0.144	0.063
		(0.071)	(0.074)	(0.082)	(0.124)	(0.077)	(0.112)	(0.093)	(0.198)	(0.086)	(0.166)	(0.018)	(0.013)
(60, 3)	100	2.976	3.030	3.413	4.070	3.658	5.798	2.457	0.985	2.501	1.341	1.764	1.552
		(0.214)	(0.214)	(0.171)	(0.202)	(0.166)	(0.197)	(0.182)	(0.251)	(0.173)	(0.196)	(0.112)	(0.108)
	400	3.177	3.863	3.640	5.423	3.700	5.919	2.406	0.869	2.313	0.859	0.392	0.223
		(0.161)	(0.165)	(0.179)	(0.216)	(0.163)	(0.195)	(0.189)	(0.281)	(0.181)	(0.249)	(0.043)	(0.036)
							SR = 20%						
(60, 1)	100	1.038	1.077	1.421	1.902	1.776	3.989	0.623	−0.604	0.668	−0.375	0.606	0.481
		(0.116)	(0.116)	(0.143)	(0.178)	(0.136)	(0.158)	(0.145)	(0.199)	(0.136)	(0.158)	(0.087)	(0.083)
	400	1.201	1.832	1.654	3.489	1.746	4.005	0.582	−0.678	0.550	−0.679	0.139	0.056
		(0.072)	(0.074)	(0.079)	(0.125)	(0.078)	(0.111)	(0.091)	(0.192)	(0.085)	(0.167)	(0.018)	(0.013)
(60, 3)	100	2.976	3.031	3.479	4.329	3.679	5.876	2.432	0.876	2.484	1.274	1.756	1.502
		(0.214)	(0.214)	(0.175)	(0.208)	(0.167)	(0.198)	(0.183)	(0.253)	(0.173)	(0.197)	(0.112)	(0.108)
	400	3.182	3.875	3.665	5.547	3.710	5.939	2.400	0.854	2.307	0.850	0.385	0.209
		(0.161)	(0.164)	(0.177)	(0.209)	(0.160)	(0.188)	(0.190)	(0.288)	(0.182)	(0.253)	(0.042)	(0.035)
							SR = 0%						
(60, 1)	100	1.038	1.078	1.434	2.055	1.773	3.989	0.631	−0.577	0.669	−0.372	0.601	0.464
		(0.116)	(0.116)	(0.146)	(0.141)	(0.137)	(0.159)	(0.144)	(0.194)	(0.136)	(0.158)	(0.087)	(0.083)
	400	1.201	1.835	1.669	3.569	1.749	4.021	0.578	−0.696	0.547	−0.692	0.139	0.054
		(0.071)	(0.074)	(0.079)	(0.119)	(0.078)	(0.112)	(0.091)	(0.199)	(0.086)	(0.172)	(0.018)	(0.012)
(60, 3)	100	2.976	3.030	3.490	4.439	3.674	5.872	2.430	0.874	2.484	1.274	1.754	1.484
		(0.214)	(0.214)	(0.177)	(0.217)	(0.168)	(0.200)	(0.181)	(0.251)	(0.173)	(0.197)	(0.113)	(0.108)
	400	3.184	3.880	3.650	5.519	3.715	5.950	2.399	0.855	2.303	0.840	0.386	0.208
		(0.161)	(0.164)	(0.167)	(0.203)	(0.159)	(0.187)	(0.192)	(0.290)	(0.184)	(0.258)	(0.043)	(0.035)

RidgeVar, the proposed method; RCV, method of Fan et al. (2012); MLE, method of Dicker & Erdogdu (2016); MM, method of Dicker (2014); EigenPrism, method of Janson et al. (2017); Scalreg, method of Sun & Zhang (2012); SR, sparsity rate given by SR $= 1 - \|\beta\|_0/p$ where $\|\cdot\|_0$ is the L_0-norm.

Acknowledgement

We thank the editor, associate editor and two referees for many helpful comments. This work was partially supported by the National Natural Science Foundation of China, the Program for Innovative Research Team of Shanghai University of Finance and Economics, and the Key Laboratory for Applied Statistics of Ministry of Education, Northeast Normal University.

Supplementary material

Supplementary Material available at *Biometrika* online includes additional results of the simulation study and technical details. An R (R Development Core Team, 2020) package implementing the proposed method is available at `https://github.com/xliusufe/RidgeVar`.

References

BAI, Z. & SARANADASA, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statist. Sinica* **6**, 311–29.

CHEN, S. & QIN, Y. (2010). A two-sample test for high dimensional data with application to gene-set testing. *Ann. Statist.* **38**, 808–35.

CUI, H., GUO, W. & ZHONG, W. (2017). U-type test for high-dimensional regression coefficients using refitted cross-validation variance estimation. *Ann. Statist.* **46**, 958–88.

DICKER, L. H. (2014). Variance estimation in high-dimensional linear models. *Biometrika* **101**, 269–84.

DICKER, L. H. & ERDOGDU, M. A. (2016). Maximum likelihood for variance estimation in high-dimensional linear models. *Proc. Mach. Learn. Res.* **51**, 159–67. Proc. 19th Int. Conf. Artificial Intelligence and Statistics (AISTATS 2016).

DOBRIBAN, E. & WAGER, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *Ann. Statist.* **46**, 247–79.

FAN, J., GUO, S. & HAO, N. (2012). Variance estimation using refitted cross-validation in ultrahigh-dimensional regression. *J. R. Statist. Soc.* B **74**, 37–65.

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Software* **33**, 1–22.

HAGER, W. W. (1989). Updating the inverse of a matrix. *SIAM Rev.* **31**, 221–39.

HOERL, A. E. & KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

JANSON, L., BARBER, R. F. & CANDES, E. (2017). EigenPrism: Inference for high dimensional signal-to-noise ratios. *J. R. Statist. Soc.* B **79**, 1037–65.

R DEVELOPMENT CORE TEAM (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

REID, S., TIBSHIRANI, R. & FRIEDMAN, J. (2016). A study of error variance estimation in lasso regression. *Statist. Sinica* **26**, 35–67.

SUN, T. & ZHANG, C. (2012) Scaled sparse linear regression. *Biometrika* **99**, 879–98.

VAN DE GEER, S., BÜHLMANN, P., RITOV,Y. & DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42**, 1166–202.

VERZELEN, N. & GASSIAT, E. (2018). Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli* **24**, 3683–710.

WANG, L., PENG, B. & LI, R. (2015). A high-dimensional nonparametric multivariate test for mean vector. *J. Am. Statist. Assoc.* **110**, 1658–69.

YU, G. & BIEN, J. (2019). Estimating the error variance in a high-dimensional linear model. *Biometrika* **106**, 533–46.

ZHANG, C. H. & ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc.* B **76**, 217–42.

ZHONG, P. S. & CHEN, S. (2011). Tests for high-dimensional regression coefficients with factorial designs. *J. Am. Statist. Assoc.* **106**, 260–74.

[*Received on* 6 *October* 2018. *Editorial decision on* 21 *August* 2019]