Article

The One Standard Error Rule for Model Selection: Does It Work?

Yuchen Chen ¹ and Yuhong Yang ^{2,*} ¹ Carlson School of Management, University of Minnesota, Minneapolis, MN 55455, USA; chen3912@umn.edu² School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA

* Correspondence: yangx374@umn.edu; Tel.: +1-6126268337

Abstract: Previous research provided a lot of discussion on the selection of regularization parameters when it comes to the application of regularization methods for high-dimensional regression. The popular “One Standard Error Rule” (1se rule) used with cross validation (CV) is to select the most parsimonious model whose prediction error is not much worse than the minimum CV error. This paper examines the validity of the 1se rule from a theoretical angle and also studies its estimation accuracy and performances in applications of regression estimation and variable selection, particularly for Lasso in a regression framework. Our theoretical result shows that when a regression procedure produces the regression estimator converging relatively fast to the true regression function, the standard error estimation formula in the 1se rule is justified asymptotically. The numerical results show the following: 1. the 1se rule in general does not necessarily provide a good estimation for the intended standard deviation of the cross validation error. The estimation bias can be 50–100% upwards or downwards in various situations; 2. the results tend to support that 1se rule usually outperforms the regular CV in sparse variable selection and alleviates the over-selection tendency of Lasso; 3. in regression estimation or prediction, the 1se rule often performs worse. In addition, comparisons are made over two real data sets: Boston Housing Prices (large sample size n , small/moderate number of variables p) and Bardet–Biedl data (large p , small n). Data guided simulations are done to provide insight on the relative performances of the 1se rule and the regular CV.

Citation: Chen, Y.; Yang, Y. The One Standard Error Rule for Model

Selection: Does It Work? *Stats* **2021**, *4*, 868–892. <https://doi.org/10.3390/stats4040051>

Keywords: subsampling; tuning parameter selection; estimation accuracy; regression estimation; variable selection

Academic Editor: Fulvia Mecatti

Received: 25 September 2021

Accepted: 1 November 2021

Published: 5 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Background

Resampling and subsampling methods were widely utilized in many statistical and machine learning applications. In high-dimensional regression learning, for instance, bootstrap and subsampling play important roles in model selection uncertainty quantification and other model selection diagnostics (see, e.g., [1] for references). Cross-validation (CV) (as a subsampling tool) and closely related methods were proposed to assess quality of variable selection performance in terms of F- and G-measures ([2]) and to determine variable importances ([3]).

In this paper, we focus on the examination of cross-validation as used for model selection. A core issue is how to properly quantify variabilities in subsamplings and associated evaluations, which is a really challenging issue that is yet to be solved, to the best of our knowledge. In application of variable and model selection, a popular practice is to use the “one standard error rule”. However, its validity hinges crucially on the goodness of the standard error formula used in the approach. Our aim in this work is to study the standard error estimation issue and its consequences on variable selection and regression estimation.

In this section, we provide a background on tuning parameter selection for high-dimensional regression.

1.1. Regularization Methods

Regularization methods are now widely used to tackle the problem of curse of dimensionality in high-dimensional regression analysis. By imposing a penalty on the complexity of the solution, it can solve an ill-posed problem or prevent overfitting. Examples include Lasso ([4]) and Ridge regression ([5,6]), which add an l_1 and l_2 penalty on the coefficient estimates, respectively, to the usual least-squares fit objective function.

Specifically, given a response vector $y \in \mathbb{R}^n$, a matrix $X \in \mathbb{R}^{n \times p}$ of predictor variables and a vector of intercept α , the Lasso estimate is defined as

$$\hat{\beta}^{lasso} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|y - \alpha - X\beta\|_2^2 + \lambda \sum_{j=1}^p |\beta_j|, \quad (1)$$

and the ridge estimate is defined as

$$\hat{\beta}^{ridge} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|y - \alpha - X\beta\|_2^2 + \lambda \sum_{j=1}^p \beta_j^2. \quad (2)$$

The tuning parameter λ controls the strength of the penalty. As is well known, the nature of the ℓ_1 penalty causes some coefficients to be shrunk to zero exactly while the ℓ_2 penalty can shrink all of the coefficients towards zero, but it typically does not set any of them exactly to zero. Thus, an advantage of Lasso is that it makes the results simpler and more interpretable. In this work, ridge regression will be considered when regression estimation (i.e., the focus is on the accurate estimation of the regression function) or prediction is the goal.

1.2. Tuning Parameter Selection

The regularization parameter, or the tuning parameter, determines the extent of penalization. A traditional way to select a model is to use an information criterion such as Akaike information criterion (AIC) ([7,8]), Bayesian information criterion (BIC) and its extension ([9–11]). A more commonly used way for tuning parameter selection is the cross-validation ([12–14]). For recent work on cross-validation for consistent model selection for high-dimensional regression, see [15].

The idea of K -fold cross validation is to split the data into K roughly equal-sized parts: $F_1, \dots, F_K$. For the k -th part where $k = 1, 2, \dots, K$, we consider training on $(x_i, y_i), i \notin F_k$ and evaluation on $(x_i, y_i), i \in F_k$. For each value of the tuning parameter $\lambda \in \{\lambda_1, \dots, \lambda_m\}$, a set of candidate values, we compute the estimate $\hat{f}_\lambda^{(-k)}$ on the training set, and record the total error on the validation set:

$$CV_k(\lambda) = \sum_{i \in F_k} (y_i - \hat{f}_\lambda^{(-k)}(x_i))^2. \quad (3)$$

For each tuning parameter value λ , we compute the average error over all folds:

$$CV(\lambda) = \frac{1}{K} \sum_{k=1}^K CV_k(\lambda). \quad (4)$$

We then choose the value of tuning parameter that minimizes the above average error,

$$\hat{\lambda}_{min} = \underset{\lambda \in \{\lambda_1, \dots, \lambda_m\}}{\operatorname{argmin}} CV(\lambda). \quad (5)$$

1.3. Goal for the One Standard Error Rule

The “one standard error rule” (1se rule) for model selection was first proposed for selecting the right sized tree ([16]). It is also suggested to pick the most parsimonious model within one standard error of the minimum cross validation error ([17]). Such a

rule acknowledges the fact that the bias-variance trade-off curve is estimated with error, and hence takes a conservative approach: a smaller model is preferred when the prediction errors are more or less indistinguishable. To estimate the standard deviation of $CV(\lambda)$ at each $\lambda \in \{\lambda_1, \dots, \lambda_m\}$, one computes the sample standard deviation of validation errors in $CV_1(\lambda), \dots, CV_K(\lambda)$:

$$SD(\lambda) = \sqrt{\text{Var}(CV_1(\lambda), \dots, CV_K(\lambda))}. \quad (6)$$

Then, we estimate the standard deviation of $CV(\lambda)$, which is declared the standard error of $CV(\lambda)$, by

$$SE(\lambda) = SD(\lambda) / \sqrt{K}. \quad (7)$$

The organization of the rest of the paper is as follows. Section 2 examines validity of the 1se rule from a theoretical perspective. Section 3 describes the objectives and approaches of our numerical investigations. Section 4 presents the experimental results of the 1se rule in three different aspects. Section 5 applies the 1se rule on two real data sets and examine its performances via data guided simulations. Section 6 concludes the paper. Appendix A gives additional numerical results and Appendix B provides the proof of the main theorem in the paper.

2. Theoretical Result on the One-Standard-Error Rule

2.1. When Is the Standard Error Formula Valid?

The 1se rule makes intuitive sense, but its validity is not clear at all. The issue is that the CV errors completed at different folds are not independent, which invalidates the standard error formula for the sample mean of i.i.d. observations. Note that the validity of the standard error is equivalent to the validity of the sample standard deviation expressed in Equation (6). Thus, it suffices to study if and to what extent the usual sample variance properly estimates the targeted theoretical variance.

In this section, we investigate the legitimacy of the standard error formula for a general learning procedure. We focus on the regression case, but similar results hold for classification as well.

Let δ be a learning procedure that produces the regression estimator $\hat{f}_\delta(x; \mathbb{D}^n)$ based on the data $\mathbb{D}^n = (x_i, y_i)_{i=1}^n$. Throughout the section, we assume K is given and for simplicity, n is a multiple of K . Define $CV_k(\delta) = \sum_{i \in F_k} (y_i - \hat{f}_\delta(x_i; \mathbb{D}_{(-k)}^n))^2$, where $\mathbb{D}_{(-k)}^n$ denote the observations in $\mathbb{D}^n$ that do not belong to F_k and $k = 1, 2, \dots, K$.

The SE of the CV errors of the learning procedure δ is defined as

$$SE(\delta) = \sqrt{\frac{\frac{1}{K-1} \sum_{k=1}^K (CV_k(\delta) - \frac{1}{K} \sum_{k=1}^K CV_k(\delta))^2}{K}}.$$

To study the SE formula, given that K is fixed, it is basically equivalent to studying the sample variance formula, i.e.,

$$\mathbb{S}_n(\delta) = \frac{1}{K-1} \sum_{k=1}^K (CV_k(\delta) - \frac{1}{K} \sum_{k=1}^K CV_k(\delta))^2.$$

The validity of these formulas (in an asymptotic sense) hinges on that the correlation between $CV_{k_1}(\delta)$ and $CV_{k_2}(\delta)$ for $k_1 \neq k_2$ approaches zero as $n \rightarrow \infty$.

Definition 1. The sample variance $\mathbb{S}_n(\delta)$ is said to be asymptotically relatively unbiased (ARU) if

$$\frac{\mathbb{E} \mathbb{S}_n(\delta) - \text{Var}(CV_k(\delta))}{\text{Var}(CV_k(\delta))} \rightarrow 0$$

as $n \rightarrow \infty$.

Clearly, if the property above holds, the SE formula is properly justified in terms of the relative estimation bias being negligible. Then, the 1se rule is sensible.

For $p > 0$, define the L_p -norm

$$\|f\| = \left(\int f^p(x) P_X(dx) \right)^{1/p},$$

where P_X denotes the probability distribution of X from which $X_1, X_2, \dots, X_n$ are drawn. Let $\hat{f}_{\delta,n}$ denote the estimator $\hat{f}_\delta(x; \mathbb{D}^n)$.

Theorem 1. If $\mathbb{E}(\|f - \hat{f}_{\delta,n}\|_4^4) = o(\frac{1}{n})$, then $\mathbb{S}_n(\delta)$ is asymptotically relatively unbiased.

Remark 1. The condition $\mathbb{E}(\|f - \hat{f}_{\delta,n}\|_4^4) = o(\frac{1}{n})$ holds if the regression procedure is based on a parametric model under regularity conditions. For instance, if we consider a linear model with p_0 terms, then under sensible conditions (e.g., [15], p. 111), $\mathbb{E}(\|f - \hat{f}_{\delta,n}\|_4^4) = \mathcal{O}(\frac{p_0^2}{n^2})$ and consequently as long as $p_0 = o(\sqrt{n})$, the variance estimator $\mathbb{S}_n(\delta)$ is asymptotically relatively unbiased.

Remark 2. The above result suggests that when relatively simple models are considered, the 1se rule may be reasonably applicable. In the same spirit, when sparsity oriented variable selection methods are considered as candidate learning methods, if the most relevant selected models are quite sparse, the 1se rule may be proper.

2.2. An Illustrative Example

Our theorem shows that the standard error formula is asymptotically valid when the correlation between the CV errors at different folds is relatively small, which is related to the rate of convergence of the regression procedure. In this section, we illustrate how the dependence affects the validity of the standard error formula in a concrete example. For ease in highlighting the dependence among the CV errors at different folds, we consider a really simple setting, where the true regression function is a constant. The simplicity allows us to calculate the variance and covariance of the CV errors explicitly and then see the essence. The insight gained, however, is more broadly applicable.

The true data generating model is $Y = \theta + \epsilon$, where $\theta \in \mathbb{R}$ is unknown. Given data $\mathbb{D}^n$, for an integer r as a factor of n , the regression procedure considers only the first, r -th, $(2r)$ -th, $\dots$, (n/r) -th observations and take the corresponding sample mean of Y to estimate $f(x) = \theta$.

Suppose $l = \frac{n}{Kr}$ is an integer. Then, when applying the K -fold CV, the training sample size is $(K-1)\frac{n}{K}$ but only $(K-1)l$ observations are actually used to estimate f . Let W_i, W_j denote the CV errors on the i -th and j -th fold, $1 \leq i, j \leq K$. Then, it can be shown that

$$\text{Var}(W_i) \asymp n(1 + \frac{n}{l^2}) \quad (8)$$

and

$$\text{Cov}(W_i, W_j) \asymp \frac{n^2}{l^2} \quad (9)$$

for $i \neq j$, where $\asymp$ denotes the same order.

Therefore, the standard error formula is ARU if and only if $\frac{l}{\sqrt{n}} \rightarrow \infty$.

In this illustrating example, when l is small (recall that only l observations are used in each fold for training), the regression estimator obtained from excluding one fold respectively are highly dependent, which makes the sample variance formula problematic. In contrast, when l is large, e.g., $l = \frac{n}{K}$ (i.e., all the observations are used), the dependence of the CV errors at different folds becomes negligible compared to that of the randomness in the response. In the extreme case of $l = \frac{n}{K}$, we have

$$\rho(W_i, W_j) \asymp \frac{1}{n} \quad (10)$$

for $i \neq j$.

Note also that the choice of l determines the convergence rate of the regression estimator, i.e., $E(\|f - \hat{f}_{\delta,n}\|_2^2)$ is of order $1/l$. Obviously, with large l , the estimator converges fast, which dilutes the dependence of the CV errors and helps the sample variance formula to be more reliable.

3. Numerical Research Objectives and Setup

One standard error rule is intended to improve the usual cross-validation method stated in Section 1.2, where λ is chosen to minimize the CV error. However, to date, no empirical or theoretical studies were conducted to justify the application of the one standard error rule in favor of parsimony, to the best of our knowledge. There are some interesting open questions that deserve investigations.

1. Does the 1se itself provide a good estimation of the standard deviation of the cross validation error $CV(\lambda)$ as intended?
2. Does the model selected by the 1se rule (model with λ_{1se}) typically outperform the model selected by minimizing the CV error (model with λ_{min}) in variable selection?
3. What if estimating the regression function or prediction is the goal?

In this paper, we study the estimation accuracy of the one standard error rule and its application in regression estimation and variable selection under various regression scenarios, including both simulations and real data examples.

3.1. Simulation Settings

Our numerical investigations are done in a regression framework where our simulated response is generated by linear models. Both dependent and independent predictors are considered. The data generating process is shown as follows.

Denote the sample size as n and the total number of predictors as p . Firstly, the predictors are generated by normal distributions. In the independent case, the vector of the explanatory variables $(X_1, X_2, \dots, X_p)$ has i.i.d. $N(0, 1)$ components and in the dependent case, the vector follows $N(0, \Sigma)$ distribution, where Σ takes either Σ_1 or Σ_2 as follows

$$\Sigma_1 = \begin{pmatrix} 1 & \rho & \dots & \rho^{n-1} \\ \rho & 1 & \dots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \dots & 1 \end{pmatrix}$$

or

$$\Sigma_2 = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix}.$$

where $0 < \rho < 1$. Then, the response vector $y \in \mathbb{R}^n$ is generated by

$$y = X\beta + \epsilon, \quad (11)$$

where β is the chosen coefficient vector (to be specified later) and the random error vector ϵ has mean 0 and covariance matrix $\sigma^2 I_{n \times n}$ for some $\sigma^2 > 0$.

3.2. Regression Estimation

For estimation of the regression function, we intend to minimize the estimation risk. Given the data, we consider a regression method and its loss. For Lasso, we define

$$\hat{\beta}_{min} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \alpha_i - x_i' \beta)^2 + \lambda \sum_{j=1}^p |\beta_j|, \quad (12)$$

where λ is chosen by the regular CV, i.e., $\lambda = \lambda_{min}$. We let $\hat{\beta}_{1se}$ be the parameter estimate obtained from the simplest model whose CV error is within the one standard error of the minimum CV error, i.e., $\lambda = \lambda_{1se}$. Given an estimate $\hat{\beta}$, the loss for estimating the regression function is

$$L(\hat{\beta}, \beta^*) = E(\mathbf{x}'\hat{\beta} - \mathbf{x}'\beta^*)^2, \quad (13)$$

where β^* is the true parameter vector and the expectation is taken with respect to the randomness of $\mathbf{x}$ that follows the same distribution as in the data generation. The loss can be approximated by

$$L(\hat{\beta}, \beta^*) \approx \frac{1}{J} \sum_{j=1}^J (z_j'\hat{\beta} - z_j'\beta^*)^2, \quad (14)$$

where $z_j, j = 1, \dots, J$ are i.i.d. with the same distribution as $\mathbf{x}$ in the data generation and J is large (In the following numerical exercise, we set $J = 500$). In the case of using real data to compare λ_{1se} and λ_{min} , clearly, it is not possible to compute the estimation loss as described above. We will use cross-validation and data guided simulations (DGS) instead for comparison of λ_{1se} and λ_{min} . See Section 4 for details.

3.3. Variable Selection

Parameter value specification can substantially affect the accuracy of the variable selection by Lasso. In our simulation study, we investigate the performances of the model with λ_{min} and that with λ_{1se} by comparing their probabilities of accurate variable selection. When considering real data, since we do not know the true data generation process, we employ data guided simulations to provide more insight. Let S^* denote the set of variables in the true model and $\hat{S}$ be the set of selected variables by a method. In the DGS over real data sets, we define the symmetric difference $S^* \nabla \hat{S}$ to be the variables that are either in S^* but not in $\hat{S}$ or in $\hat{S}$ but not in S^* . We use the size of the symmetric difference to measure the performance of the variable selection method. Performance of the model with λ_{min} and the model with λ_{1se} in terms of variable selection will be compared this way for the data examples, where some plausible models based on the data are used to generate data.

Organization of the numerical results is as follows. Section 4 presents the experimental results of the 1se rule in its estimation accuracy as well as its performances in regression estimation and variable selection, respectively. Section 5 applies the 1se rule over two real data sets and investigates its performance relative to the regular CV via DGS. Some additional figures and tables are in Appendix A. All of R codes used in this paper are available upon request.

4. Simulation Results

4.1. Is 1se on Target of Estimating the Standard Deviation?

This subsection presents a simulation study of the estimation accuracy of the 1se with respect to different parameter settings in three sampling settings, by varying value of the correlation (ρ) in the design matrix, error variance σ^2 , sample size n , and the number of cross validation fold K .

4.1.1. Data Generating Processes (DGP)

We consider the following three sampling settings through which we first simulate the dependent (Σ_1 with $\rho = 0.9$) or independent design matrix $X \in \mathbb{R}^{n \times p}$, and then generate the response $y = X\beta + \epsilon$ with either decaying or constant coefficient vector β . Recall that p is the total number of predictors and let q denote the number of nonzero coefficients.

DGP 1: Y is generated by 5 predictors (out of 30, $q = 5, p = 30$).

DGP 2: Y is generated by 20 predictors (out of 30, $q = 20, p = 30$).

DGP 3: Y is generated by 20 predictors (out of 1000, $q = 20, p = 1000$).

For the coefficient $\beta = (\beta_1, \dots, \beta_q, \beta_{q+1}, \dots, \beta_p)$, it equals $(1, \frac{1}{2}, \dots, \frac{1}{q}, 0, \dots, 0)$ for decaying coefficient case, and it equals $(\underbrace{1, 1, \dots, 1}_q, \underbrace{0, \dots, 0}_{p-q})$ for constant coefficient case.

In these three sampling settings, ϵ are i.i.d. $N(0, \sigma^2)$. Therefore, the first and third configuration yield sparse models and the third configuration also represents the case of high-dimensional data.

4.1.2. Procedure

We apply the Lasso and Ridge procedures on the simulated data sets with tuning parameter selection by the two cross validation versions. The simulation results obtained with various parameter choices are summarized in Tables A1–A4 in the Appendix A. The simulation choices are sample size $n \in \{100, 1,000\}$, the error variance $\sigma^2 \in \{0.01, 1, 4\}$ and K -fold cross validation: $K \in \{2, 5, 10, 20\}$.

1. Do the j -th simulation ($j = 1, \dots, N$). Simulate a data set of sample size n given the parameter values.
2. K -fold cross validation. Randomly split the data set into K equal-sized parts: $F_1, F_2, \dots, F_K$ and record the cross validation error for each part $CV_k(\lambda)$ where $k = 1, 2, \dots, K$.
3. Calculate the total cross validation error, $CV.err_j(\lambda)$, by taking the mean of these K numbers. Find λ_{min} that minimizes the total CV error. Now, we take $\lambda = \lambda_{min}$ throughout the remaining steps. Calculate the standard error of cross validation error $SE_j(\lambda)$ by one standard error rule:

$$SE_j(CV.error) = \sqrt{Var(CV_{1,j}, CV_{2,j}, \dots, CV_{K,j})/K}.$$

4. Repeat step 1 to step 3 for N times ($N = 100$). Calculate the standard deviation of the cross validation error $CV.err_j(\lambda)$, and call it $SD(CV.error)$, and the mean standard error of cross validation error (as used for 1se rule), $SE(CV.error) = \frac{1}{N} \sum SE_j(CV.error)$.
5. Performance assessment. Calculate the ratio of (the claimed) standard error over (the simulated true) standard deviation:

$$\frac{SE(CV.error)}{SD(CV.error)}.$$

If the 1se works well, the ratio should be close to 1. However, the results in the Tables A1–A4 show that, in some cases, the standard deviation is even more than 100% overestimated or more than 50% underestimated. The histograms of the ratios (Figures A1 and A2) also show that the standard error estimates are not close to the targets. Not surprisingly, 2-fold is particularly untrustworthy. But for the common choice of 10-fold, even for independent predictors with constant coefficient, the standard error estimate by 1se can be 40% over or under the target. In addition, there is no obvious pattern on when the 1se is performing well or not. Overall, this simulation study provides evidence that the 1se fails to provide good estimation as intended (see [18] for more discussions on the standard error issues).

4.2. Is 1se Rule Better for Regression Estimation?

Since the standard error estimation in 1se is not the end goal, the poor performance of 1se in the previous subsection does not necessary mean that the 1se rule is poor for regression estimation and variable selection. In this part, we study the performance of model λ_{1se} in regression estimation and compare it with that of model λ_{min} . A large number of samples are drawn so that we know in which cases the 1se rule can consistently outperforms the regular CV in regression estimation.

By simulations, we compare λ_{1se} and λ_{min} in term of accuracy in estimating the regression function. We consider several factors: nonzero coefficients $\beta_{(1)}$, number of zero

coefficients $p - q$, sample size n , standard deviation of the noise σ and the dependency of the predictors.

4.2.1. Model Specification

We consider the data generation model: $y = X\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$ and $\beta = (\beta_{(1)}, \beta_{(2)})$, with $\beta_{(1)} = (\beta_1, \beta_2, \dots, \beta_q)$ being the nonzero coefficients and $\beta_{(2)} = (\beta_{q+1}, \beta_{q+2}, \dots, \beta_p)$ being the zero coefficients. The vector of explanatory variables is generated by the multivariate normal distribution,

$$X \sim MV(0, \Sigma_i),$$

where $i = 1, 2$.

4.2.2. Procedure

1. Simulate a fixed validation set of 500 observations of the predictors for estimation of the loss.
2. Each time randomly simulate a training set of n observations from the same data generating process.
3. Apply the 1se rule (10-fold cross validation) over the training set and record the selected model: model with λ_{1se} and the model with λ_{min} . Calculate the estimation losses of these two models: $\frac{1}{500} \sum_{i=1}^{500} (z_i' \hat{\beta} - z_i' \beta^*)^2$, where $\hat{\beta}$ is based on λ_{1se} or λ_{min} respectively, and z_i are independently generated from the same distribution used to generate X .
4. Repeat this process for M times ($M = 500$) and calculate the fraction that model with λ_{min} has a smaller loss.

In our simulations, we make various comparisons: small or large nonzero coefficients; small or large error variance; small or large sample size; low- or high-dimensional data; independent or dependent design matrix. The parameter values used in our study are displayed in Table 1 below. The results are reported in Tables A5 and A6. Note that when λ_{min} and λ_{1se} give identical selection over the $M = 500$ runs, the entry in the table is NaN.

Table 1. Simulation settings for regression estimation and variable selection.

Parameter	Values
$\beta_{(1)}$	$\{0.1, 1, 3\}$ for constant coefficients $\frac{1}{i}$ where $i = 1, \dots, q$ for decaying coefficients
σ^2	$\{0.01, 4\}$
ρ	$\{0, 0.5, 0.9\}$
n	$\{100, 1, 000\}$
$p - q$	$\{20, 990\}$ where $q = 10$
Σ_i	$i = 1$ for AR(1) covariance; $i = 2$ for compound symmetry covariance

The simulation results show that, in general, for the design matrix with either AR(1) correlation or constant correlation, the model with λ_{min} tends to provide better regression estimation (with smaller estimation errors), especially when the data is generated with relatively large coefficient. In addition, we consider two more settings below.

4.2.3. Case 1: Large p/q

In this case, we consider high-dimensional data with AR(1) covariance design matrix, Σ_1 , where $q = 10$ and $p = 1000$. We set $\sigma^2 = 1$, $n \in \{100, 1000\}$, $\rho \in \{0.01, 0.9\}$, constant nonzero coefficients with value $\beta \in \{1, 6, 11, 16, 21\}$. We report the probability that model λ_{min} estimates better. Clearly, in this case, the model λ_{min} significantly outperforms the model λ_{1se} for a range of β values at both sample size (See Figure 1).

Figure 1. Probability that model with λ_{min} outperforms model with λ_{1se} for estimation.

4.2.4. Case 2: Small p/q

In this case, we consider low-dimensional data with AR(1) covariance design matrix, Σ_1 , where $q = 10$ and $p - q \in \{2, 5, 10\}$. We set $\sigma^2 = 0.01$, $n \in \{100, 1000\}$, $\rho = 0.1$, the common nonzero coefficient value $\beta \in \{0.2, 0.4, 0.6, 0.8, 1, 1.2\}$. We report the probability that model λ_{min} estimates the regression function better. In this case, model λ_{min} significantly outperforms model λ_{1se} as the common coefficient value increases over 1, getting close to or even over 0.8 (See Figure 2).

Figure 2. Probability that model with λ_{min} outperforms model with λ_{1se} for estimation.

4.3. Is 1se Rule Better for Variable Selection?

There are two main issues of the application of the 1se rule in variable selection. One is that the accuracy of variable selection by Lasso itself is very sensitive to the parameter values in the data generating process. In various situations (especially when the covariance are complicatedly correlated such as in gene expression type of data), Lasso may perform very poorly in accurate variable selection. The second issue is that λ_{min} and λ_{1se} returned by the glmnet function are the same in many cases and then there is no difference in the result of variable selection. The cases where $\lambda_{min} = \lambda_{1se}$ are eliminated in the following

general study. To examine the real difference between λ_{min} and λ_{1se} , we consider the difference of their probabilities of selecting the true model given that they have distinct select outcome.

4.3.1. Procedure

1. Randomly simulate a data set of sample size n given the parameter values.
2. Perform variable selection with λ_{min} and λ_{1se} over the simulated data set, respectively. If the two returned models are the same, then discard the result and go back to step 1. Otherwise, check their variable selection results.
3. Repeat the above process for M times ($M = 500$) and calculate the fraction of correct variable selection, denoted by $P_c(\lambda_{1se})$ and $P_c(\lambda_{min})$, respectively. We report the proportion difference $P_c(\lambda_{1se}) - P_c(\lambda_{min})$. Positive means 1se rule is better.

We adopt the same framework and parameter settings (Table 1) as that used for regression estimation. The results show that λ_{1se} tends to outperform λ_{min} in most of the cases when the data are generated with AR(1) correlation design matrix (see Tables A7 and A8). But with constant correlation design matrix, these two methods have no practically difference in variable selection.

Below we consider several special cases where model λ_{min} still needs to be considered, and although it can only do slightly better when the error variance is large, nonzero coefficients are small and no extra variables exist to interrupt the selection process.

4.3.2. Case 1: Constant Coefficients

In this case, let $q = 5$, $p - q \in \{0, 1, 2, 3, 4\}$, $\sigma^2 = 4$, $n \in \{100, 1000\}$, $\rho = 0.01$ and the common nonzero $\beta \in [0.5, 1.3]$ with Σ_1 . The result is in Figure 3. In the case of small sample size, say $n = 100$, model λ_{min} has higher probability of correct variable selection when the constant coefficient lies between (0.5, 1). The largest difference appears when $p - q = 0$, not surprisingly. As the common coefficient value becomes larger than 1, the probability increases from negative to positive, which indicates the reverse to be true: model λ_{1se} now works better. However, in the case of large sample size, say $n = 1000$, model λ_{1se} significantly outperforms model λ_{min} unless when $p - q = 0$.

Figure 3. Difference in proportions of selecting true model: $P_c(\lambda_{1se}) - P_c(\lambda_{min})$ (positive means 1se is better).

4.3.3. Case 2: Decaying Coefficients

In this case, let $q = 5, \sigma^2 = 4, n = 1000, \rho = 0.5, p - q \in \{0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$, and $\beta_i = \frac{1}{i}, i = 1, 2, \dots, q$ with Σ_1 . Figure 4 indicates that, with decaying coefficients lie between (0,1) and large sample size $n = 1000$, model λ_{min} has higher probability of correct variable selection when the number of zero-coefficient variables $p - q$ is small. The largest difference appears when $p - q = 0$, i.e., there is no extra variables provided to interrupt the variable selection process.

Figure 4. Difference in proportions of selecting true model: $P_c(\lambda_{1se}) - P_c(\lambda_{min})$ (positive means 1se is better).

4.3.4. Case 3: Hybrid Case

Let $n = 100, \sigma^2 = 4$ and $\rho \in \{0, 0.5\}$ with Σ_1 . In this data generating process, we set the first ten coefficients to be positive, i.e., $q = 10$, among which the first nine nonzero coefficients are constant 1, while the last one β_{10} increases from 0 to 10, i.e., $\beta_{10} \in \{0, 0.05, 0.1, 0.15, 0.2, 0.4, 0.6, 0.8, 1, 3, 5, 10\}$. The number of zero coefficients $p - q \in \{0, 1, 2, 3, 5, 20\}$. From Figure 5, when β_{10} is very small, i.e., $\beta_{10} \in \{0.05, 0.2\}$ and $p - q = 0$, λ_{min} performs better. For other situations, λ_{1se} is significantly better.

Figure 5. Difference in proportions of selecting true model: $P_c(\lambda_{1se}) - P_c(\lambda_{min})$ (positive means 1se is better).

5. Data Examples

In this section, the performances of the 1se rule and the regular CV in regression estimation and variable selection are examined over real data sets: Boston Housing Price (an example of $n > p$) and Bardet–Biedl data (an example of $p > n$).

Boston Housing Price. The data were collected by [19] for the purpose of discovering whether or not clean air influenced the value of houses in Boston. It consists of 506 observations and 14 non constant independent variables. Of these, *medv* is the response variable while the other 13 variables are possible predictors.

Bardet–Biedl. The gene expression data are from the microarray experiments of mammalian eye tissue samples of [20]. The data consists of 120 rats with 200 gene probes and the expression level of TRIM32 gene. It is interesting to know which gene probes are more able to explain the expression level of TRIM32 gene by regression analysis.

5.1. Regression Estimation

To test which method does better in regression estimation, we perform both cross validation and data guided simulation (DGS) over these two real data sets as follows.

5.1.1. Procedure for Cross Validation

1. We randomly select n_1 observations from the data sets as training set and the rest as validation set, $n_1 \in \{100, 200, 400\}$ for Boston Housing Price and $n_1 \in \{40, 80\}$ for Bardet–Biedl data.
2. Apply K-fold cross validation $K \in \{5, 10, 20\}$, for Boston Housing Price and $K \in \{5, 10\}$ for Bardet–Biedl data over the training set and compute the mean square prediction error of model λ_{1se} and model λ_{min} over the validation set.
3. Repeat the above process for 500 times and compute the proportion of better estimation for each method.

The results (Tables 2 and 3) show that, for all the training size n_1 considered, the proportion of model λ_{min} doing better is about 52~57% for Boston Housing Price; 56~59% for Bardet–Biedl data. The consistent slight advantage of λ_{min} agrees with our earlier simulation results in supporting λ_{min} as the better method for regression estimation.

Table 2. Regression estimation (cross validation): Boston Housing Price.

	Proportion of λ_{min} Being Better in Prediction		
	5-fold	10-fold	20-fold
train set $n_1 = 100$	0.554	0.526	0.538
train set $n_1 = 200$	0.560	0.556	0.564
train set $n_1 = 400$	0.552	0.564	0.558

Table 3. Regression estimation (cross validation): Bardet–Biedl.

	Proportion of λ_{min} Being Better in Prediction	
	5-fold	10-fold
train set $n_1 = 40$	0.582	0.566
train set $n_1 = 80$	0.580	0.594

5.1.2. Procedure for DGS

1. Obtain model λ_{min} and model λ_{1se} with K-fold cross validation, $K \in \{5, 10, 20\}$, over the data and also the coefficient estimates $\hat{\beta}_{min}$, $\hat{\beta}_{1se}$, the standard deviation estimates $\hat{\sigma}_{min}$, $\hat{\sigma}_{1se}$ and the estimated responses $\hat{y}_{min,1}$, $\hat{y}_{1se,1}$ (fitted values).
2. Simulate the new responses in two scenarios:
 Scenario 1: $y = \hat{y}_{min,1} + \epsilon$ with $\epsilon \sim N(0, \hat{\sigma}_{min}^2 I)$,
 Scenario 2: $y = \hat{y}_{1se,1} + \epsilon$ with $\epsilon \sim N(0, \hat{\sigma}_{1se}^2 I)$.

- Here, apply Lasso with λ_{min} and λ_{1se} (using the same K-fold CV) on the new data set (i.e., new response and the original design matrix) and get the new estimated responses $\hat{y}_{min,2}$ and $\hat{y}_{1se,2}$ for each of the two scenarios.
- Calculate the mean square estimation error $\frac{1}{n} \sum_{i=1}^n (\hat{y}_{min,2} - \hat{y}_{min,1})^2$ and $\frac{1}{n} \sum_{i=1}^n (\hat{y}_{1se,2} - \hat{y}_{min,1})^2$ for scenario 1; $\frac{1}{n} \sum_{i=1}^n (\hat{y}_{min,2} - \hat{y}_{1se,1})^2$ and $\frac{1}{n} \sum_{i=1}^n (\hat{y}_{1se,2} - \hat{y}_{1se,1})^2$ for scenario 2.
- Repeat the above resampling process for 500 times and compute the proportion of better estimation for each method.

The estimation results (Table 4) show that, overall, the proportion of better estimation for the model λ_{min} is between 46.8~52.8% and for the model λ_{1se} , it is between 47.2~53.2%. Therefore, based on the DGS, in term of regression estimation accuracy, there is not much difference before λ_{min} and λ_{1se} : the observed properties are not significantly different from 50% at 0.05 level.

Table 4. Regression estimation (DGS).

	Proportion of λ_{min} Being Better in Estimation				
	Boston Housing Price			Bardet–Biedl	
	5-fold	10-fold	20-fold	5-fold	10-fold
Scenario 1	0.478	0.486	0.468	0.528	0.523
Scenario 2	0.515	0.489	0.487	0.494	0.480

5.2. Variable Selection: DGS

To test which model does better in variable selection, we perform the DGS over the above real data sets. Since the exact variable selection might not be reached, especially when applied to the high-dimensional data, we use symmetric difference $|S^* \nabla \hat{S}|$ to measure their performance, which is the number of variables either in S^* but not in $\hat{S}$ or in $\hat{S}$ but not in S^* . Below is the algorithm for the DGS.

Procedure

- Apply the 10-fold (default) cross validation over real data set and select the true set of variables S^* : S_{min}^* or S_{1se}^* , either by the model with λ_{min} or by the model with λ_{1se} .
- Do least square estimation by regressing the response on the true set of variables and obtain the estimated response $\hat{y}$ and the residual standard error $\hat{\sigma}$.
- Simulate the new response y_{new} by adding the error term randomly generated from $N(0, \hat{\sigma}^2)$ to $\hat{y}$. Apply K-fold cross validation, $K \in \{5, 10, 20\}$, over the simulated data set (i.e., y_{new} and the original design matrix) and select set of variables $\hat{S}$: $\hat{S}_{min}$ or $\hat{S}_{1se}$, either by model λ_{min} or by model λ_{1se} . Repeat this process for 500 times.
- Calculate the symmetric difference $|S^* \nabla \hat{S}|$.

The distributions of the symmetric difference are shown in the appendix. (Figures A3 and A4 for Boston Housing Price; Figures A5 and A6 for Bardet–Biedl data) The mean of $| \hat{S} |$ is reported in Tables 5 and 6. For Boston Housing Price, the size of the candidate variables is 13 while model λ_{min} selects 11 variables and model λ_{1se} selects 8 variables on average, i.e., the mean of $| S_{min}^* |$ is 11 and the mean of $| S_{1se}^* |$ is 8. For Bardet–Biedl data, the size of the candidate variables is 200, while model λ_{min} selects 21 variables and model λ_{1se} selects 19 variables on average, i.e., the mean of $| S_{min}^* |$ is 21 and the mean of $| S_{1se}^* |$ is 19.

On average, the 1se rule does better in variable selection since the mean of $| \hat{S}_{1se} |$ is more closed to $| S^* |$. In the case of $n > p$, if we want to have a perfect or near perfect variable selection result (i.e., $| S^* \nabla \hat{S} | \leq 1$), then the 1se rule is a good choice, despite the fact that its selection result is more unstable: the range of $| S^* \nabla \hat{S}_{1se} |$ is larger than that of $| S^* \nabla \hat{S}_{min} |$. In the case of $n < p$, both models fail to provide perfect selection result. But overall the 1se rule is more accurate and stable given that $| S^* \nabla \hat{S}_{1se} |$ has smaller mean and range.

Table 5. Mean of selected model size $|\hat{S}|$: Boston Housing Price.

		5-fold	10-fold	20-fold
$ S_{min}^* = 11$	$ \hat{S}_{min} $	12.744	12.672	12.636
	$ \hat{S}_{1se} $	10.564	10.608	10.560
$ S_{1se}^* = 8$	$ \hat{S}_{min} $	10.792	10.824	10.772
	$ \hat{S}_{1se} $	8.250	8.254	8.246

Table 6. Mean of selected model size $|\hat{S}|$: Bardet–Biedl.

		5-fold	10-fold	20-fold
$ S_{min}^* = 21$	$ \hat{S}_{min} $	32.272	33.604	35.500
	$ \hat{S}_{1se} $	21.064	21.864	22.404
$ S_{1se}^* = 19$	$ \hat{S}_{min} $	24.208	25.426	26.350
	$ \hat{S}_{1se} $	19.852	19.992	20.274

6. Conclusions

The one standard error rule is proposed to pick the most parsimonious model within one standard error of the minimum. Despite the fact that it was widely used as a conservative and alternative approach for cross validation, there is no evidence confirming that the 1se rule can consistently outperform the regular CV.

Our theoretical result shows that the standard error formula is asymptotically valid when the regression procedure converges relatively fast to the true regression function. Our illustrative example also shows that when the regression estimator converges slowly, the CV errors on the different evaluation folds may be highly dependent, which unfortunately makes the usual sample variance (or sample standard deviation) formula problematic. In such a case, the use of the 1se rule may easily lead to a choice of inferior candidates.

Our numerical results offer finite-sample understandings. First of all, the simulation study also casts doubt on the estimation accuracy of the 1se rule in terms of estimating the standard error of the minimum cross validation error. In some cases, the 1se is likely to have 100% overestimation or underestimation, depending on the number of cross validation fold and the dependency among the predictors. More specifically, for example, the 1se tends to have overestimation when used with 2-fold cross validation in case of dependent predictors; but it tends to have underestimation when used with 20-fold cross validation in case of independent predictors.

Secondly, the performances of the 1se rule and the regular CV in regression estimation and variable selection are compared via both simulated and real data sets. In general, on the one hand, the 1se rule often performs better in variable selection for sparse modeling; on the other hand, it often does worse in regression estimation.

While our theoretical and numerical results clearly challenge indiscriminate uses of the 1se rule in general model selection, its parsimony nature may still be appealing when a sparse model is highly desirable. For the real data sets, for regression estimation, the regular CV does better than the 1se rule but the difference is small. There is almost no difference in the estimation accuracy between these two methods when regression estimation is assessed by the DGS. Overall, considering the model simplicity and interpretability, the 1se rule would be a better choice here.

Future studies are encouraged to provide more theoretical understandings on the 1se rule. Also, it may be studied numerically in more complex settings (e.g., when used to compare several nonlinear/nonparametric classification methods) than those in this work. In a larger context, since it is well known that the penalty methods for model selection are often unstable (e.g., [1,21,22]), alternative methods to strengthen reliability and reproductivity of variable selection may be applied (see reference in the above mentioned papers and [23]).

Author Contributions: Conceptualization, Y.Y.; methodology, Y.C. and Y.Y.; numerical results, Y.C.; writing—original draft preparation, Y.C.; writing—review and editing, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

Appendix A. Additional Numerical Results

Figure A1. Histogram of Tables A1 and A2.

Figure A2. Histogram of Tables A3 and A4.

Table A1. Mean ratio of claimed SE and true SD: dependent predictors (Σ_1) and decaying coefficients.

			DGP1				DGP2				DGP3			
			2-fold	5-fold	10-fold	20-fold	2-fold	5-fold	10-fold	20-fold	2-fold	5-fold	10-fold	20-fold
$n = 100$	$\sigma^2 = 0.01$	Lasso	1.489	1.305	1.275	1.271	1.068	1.185	1.183	1.127	1.015	1.217	1.266	1.222
		Ridge	1.187	1.104	0.975	0.924	1.370	1.171	1.082	1.031	0.918	1.190	1.252	1.194
	$\sigma^2 = 1$	Lasso	1.759	1.376	1.145	1.088	1.247	1.181	1.148	1.134	0.886	1.094	0.906	0.728
		Ridge	1.499	1.350	1.231	1.129	1.512	1.310	1.202	1.102	1.003	1.092	1.093	1.049
	$\sigma^2 = 4$	Lasso	1.260	1.162	1.067	1.071	1.333	1.202	1.164	1.119	1.121	0.580	0.319	0.229
		Ridge	1.540	1.332	1.217	1.108	1.553	1.360	1.252	1.142	1.080	0.969	0.968	0.949
$n = 1000$	$\sigma^2 = 0.01$	Lasso	1.221	1.226	1.123	1.134	1.876	1.100	1.028	0.997	1.519	1.166	1.101	1.111
		Ridge	1.725	1.011	0.880	0.853	1.490	0.882	0.780	0.777	1.385	1.299	1.254	1.251
	$\sigma^2 = 1$	Lasso	1.960	1.233	1.068	1.031	1.725	1.219	1.073	0.959	1.853	0.889	0.711	0.637
		Ridge	2.173	1.243	1.060	1.029	2.069	1.188	1.010	0.978	1.346	1.224	1.183	1.177
	$\sigma^2 = 4$	Lasso	1.650	1.031	0.955	0.873	2.122	1.184	1.003	0.880	1.868	0.831	0.668	0.598
		Ridge	2.139	1.218	1.032	0.996	2.154	1.244	1.057	1.020	1.180	1.104	1.085	1.079

Table A2. Mean ratio of claimed SE and true SD: independent predictors and decaying coefficients.

			DGP1				DGP2				DGP3			
			2-fold	5-fold	10-fold	20-fold	2-fold	5-fold	10-fold	20-fold	2-fold	5-fold	10-fold	20-fold
$n = 100$	$\sigma^2 = 0.01$	Lasso	1.275	1.225	1.130	1.077	1.029	1.087	1.110	1.128	1.221	0.952	0.581	0.379
		Ridge	0.973	0.987	0.918	0.909	1.035	1.012	0.964	0.905	1.152	1.130	1.143	1.108
	$\sigma^2 = 1$	Lasso	1.532	1.197	1.055	0.999	1.121	1.054	1.132	1.030	1.243	0.978	0.651	0.459
		Ridge	1.039	0.989	0.936	0.924	1.083	1.043	0.977	0.946	1.049	0.986	1.004	0.967
	$\sigma^2 = 4$	Lasso	1.569	1.241	1.074	1.026	1.109	1.125	1.061	1.041	1.236	1.033	0.767	0.584
		Ridge	1.027	0.990	0.938	0.933	1.050	1.030	0.963	0.946	1.006	0.951	0.963	0.930
$n = 1000$	$\sigma^2 = 0.01$	Lasso	1.126	0.914	0.748	0.668	1.499	1.046	0.864	0.820	1.855	0.961	0.789	0.741
		Ridge	1.509	1.108	0.964	0.909	1.677	1.007	0.900	0.854	1.203	1.127	1.100	1.091
	$\sigma^2 = 1$	Lasso	1.100	0.749	0.641	0.597	1.624	1.181	0.984	0.888	2.065	1.026	0.823	0.747
		Ridge	1.264	1.037	0.975	0.959	1.327	0.994	0.944	0.926	1.003	0.962	0.962	0.964
	$\sigma^2 = 4$	Lasso	1.864	1.193	0.964	0.879	1.866	0.965	0.797	0.710	2.102	1.042	0.836	0.756
		Ridge	1.145	1.004	0.963	0.960	1.131	0.975	0.944	0.938	0.955	0.935	0.937	0.941

Table A3. Mean ratio of claimed SE and true SD: dependent predictors (Σ_1) and constant coefficients.

			DGP1				DGP2				DGP3			
			2-fold	5-fold	10-fold	20-fold	2-fold	5-fold	10-fold	20-fold	2-fold	5-fold	10-fold	20-fold
$n = 100$	$\sigma^2 = 0.01$	Lasso	1.452	1.346	1.322	1.307	0.958	1.132	1.140	1.103	0.953	1.155	1.214	1.161
		Ridge	1.211	1.279	1.068	1.047	1.573	1.391	1.246	1.184	1.046	1.370	1.399	1.356
	$\sigma^2 = 1$	Lasso	1.595	1.284	1.170	1.134	0.955	1.186	1.195	1.136	0.947	1.178	1.231	1.203
		Ridge	1.307	1.280	1.144	1.074	1.783	1.428	1.384	1.290	1.068	1.397	1.440	1.380
	$\sigma^2 = 4$	Lasso	1.529	1.415	1.176	1.070	0.998	1.280	1.260	1.254	1.039	1.312	1.290	1.270
		Ridge	1.432	1.340	1.221	1.130	1.566	1.312	1.208	1.111	1.092	1.422	1.471	1.389
$n = 1000$	$\sigma^2 = 0.01$	Lasso	1.193	1.187	1.110	1.119	1.084	0.804	0.773	0.744	1.492	1.128	1.122	1.079
		Ridge	2.087	1.123	0.970	0.912	1.742	1.448	1.423	1.301	1.726	1.596	1.509	1.482
	$\sigma^2 = 1$	Lasso	1.995	1.469	1.330	1.292	1.389	1.035	0.950	0.948	1.573	1.225	1.154	1.159
		Ridge	2.017	1.196	1.029	1.011	2.315	1.448	1.200	1.088	1.744	1.599	1.505	1.486
	$\sigma^2 = 4$	Lasso	2.200	1.023	0.883	0.806	2.128	1.375	1.162	1.127	2.144	1.385	1.235	1.204
		Ridge	2.163	1.260	1.078	1.054	1.991	1.261	1.050	1.008	1.783	1.612	1.500	1.490

Table A4. Mean ratio of claimed SE and true SD: independent predictors and constant coefficients.

			DGP1				DGP2				DGP3			
			2-fold	5-fold	10-fold	20-fold	2-fold	5-fold	10-fold	20-fold	2-fold	5-fold	10-fold	20-fold
$n = 100$	$\sigma^2 = 0.01$	Lasso	1.180	1.248	1.178	1.158	0.921	1.082	1.160	1.109	1.197	1.172	0.599	0.386
		Ridge	0.918	1.149	1.099	1.104	1.121	1.089	1.009	0.851	1.103	1.112	1.096	1.050
	$\sigma^2 = 1$	Lasso	1.311	1.241	1.133	1.075	0.930	1.110	1.195	1.129	1.371	1.166	0.594	0.380
		Ridge	0.953	0.980	0.898	0.886	1.116	0.997	0.858	0.740	1.122	1.116	1.096	1.054
	$\sigma^2 = 4$	Lasso	1.366	1.202	1.087	1.057	0.966	1.104	1.151	1.123	1.387	1.155	0.597	0.382
		Ridge	1.018	0.985	0.917	0.906	1.135	1.016	0.896	0.799	1.159	1.107	1.092	1.052
$n = 1000$	$\sigma^2 = 0.01$	Lasso	0.939	0.978	0.842	0.763	1.149	1.030	1.093	1.003	1.068	1.050	1.057	1.067
		Ridge	1.475	0.969	0.768	0.719	1.101	1.045	0.942	0.896	1.543	1.446	1.412	1.396
	$\sigma^2 = 1$	Lasso	1.311	1.003	0.833	0.788	0.991	1.001	1.065	0.954	1.181	1.027	1.063	1.017
		Ridge	1.511	1.086	0.924	0.874	1.362	0.572	0.460	0.433	1.370	1.185	1.154	1.142
	$\sigma^2 = 4$	Lasso	0.996	0.824	0.650	0.631	1.317	0.971	0.909	0.798	1.293	1.000	1.014	0.957
		Ridge	1.405	1.099	0.974	0.934	1.552	0.793	0.679	0.652	1.303	1.157	1.136	1.131

Table A5. Regression estimation: probability of λ_{min} doing better, AR(1) correlation.

			Low-Dimension				High-Dimension			
			$\beta = 0.1$	$\beta = 1$	$\beta = 3$	Decay β	$\beta = 0.1$	$\beta = 1$	$\beta = 3$	Decay β
$n = 100$	$\rho = 0$	$\sigma^2 = 0.01$	0.52	0.62	1.00	0.52	0.46	0.64	0.84	0.54
		$\sigma^2 = 4$	0.42	0.52	0.52	0.54	0.58	0.50	0.50	0.52
	$\rho = 0.5$	$\sigma^2 = 0.01$	0.54	0.83	0.92	0.52	0.58	0.70	0.90	0.58
		$\sigma^2 = 4$	0.44	0.56	0.54	0.58	0.38	0.60	0.60	0.60
	$\rho = 0.9$	$\sigma^2 = 0.01$	0.56	0.88	1.00	0.60	0.54	0.84	0.96	0.54
		$\sigma^2 = 4$	0.54	0.56	0.56	0.54	0.57	0.52	0.52	0.46
$n = 1000$	$\rho = 0$	$\sigma^2 = 0.01$	0.58	0.57	NaN	0.58	0.64	0.71	NaN	0.74
		$\sigma^2 = 4$	0.45	0.58	0.56	0.52	0.57	0.66	0.66	0.72
	$\rho = 0.5$	$\sigma^2 = 0.01$	0.50	NaN	NaN	0.54	0.68	NaN	NaN	0.68
		$\sigma^2 = 4$	0.52	0.50	0.52	0.54	0.64	0.68	0.68	0.70
	$\rho = 0.9$	$\sigma^2 = 0.01$	0.56	NaN	NaN	0.60	0.66	NaN	NaN	0.70
		$\sigma^2 = 4$	0.56	0.54	0.58	0.54	0.68	0.66	0.64	0.66

Table A6. Regression estimation: probability of λ_{min} doing better, constant correlation.

			Low-Dimension				High-Dimension			
			$\beta = 0.1$	$\beta = 1$	$\beta = 3$	Decay β	$\beta = 0.1$	$\beta = 1$	$\beta = 3$	Decay β
$n = 100$	$\rho = 0.5$	$\sigma^2 = 0.01$	0.56	0.88	0.89	0.53	0.54	0.84	1.00	0.64
		$\sigma^2 = 4$	0.53	0.56	0.54	0.52	0.61	0.52	0.56	0.60
	$\rho = 0.9$	$\sigma^2 = 0.01$	0.60	1.00	1.00	0.70	0.56	0.82	0.90	0.70
		$\sigma^2 = 4$	0.58	0.54	0.62	0.62	0.62	0.58	0.64	0.64
$n = 1000$	$\rho = 0.5$	$\sigma^2 = 0.01$	0.50	NaN	NaN	0.50	0.74	NaN	NaN	0.78
		$\sigma^2 = 4$	0.50	0.42	0.50	0.44	0.72	0.74	0.76	0.78
	$\rho = 0.9$	$\sigma^2 = 0.01$	0.48	NaN	NaN	0.54	0.74	NaN	NaN	0.78
		$\sigma^2 = 4$	0.52	0.50	0.52	0.46	0.72	0.74	0.76	0.76

Table A7. Variable selection: $P_c(\lambda_{1se}) - P_c(\lambda_{min})$, AR(1) correlation.

			Low-Dimension				High-Dimension			
			$\beta = 0.1$	$\beta = 1$	$\beta = 3$	Decay β	$\beta = 0.1$	$\beta = 1$	$\beta = 3$	Decay β
$n = 100$	$\rho = 0$	$\sigma^2 = 0.01$	0.13	0.13	0	0.06	0	0	0.02	NaN
		$\sigma^2 = 4$	0	0.12	0.11	0.01	0	0	0	0.90
	$\rho = 0.5$	$\sigma^2 = 0.01$	0.45	0	0	0	0.25	0.01	0	NaN
		$\sigma^2 = 4$	0	0.40	0.43	0.05	0	0.16	0.27	0.87
	$\rho = 0.9$	$\sigma^2 = 0.01$	0.40	0	0	0	0.61	0	0	NaN
		$\sigma^2 = 4$	0	0.14	0.39	0.01	0	0.25	0.62	0.7
$n = 1000$	$\rho = 0$	$\sigma^2 = 0.01$	0.72	0.10	NaN	0.70	0.27	0.16	NaN	NaN
		$\sigma^2 = 4$	0	0.71	0.70	0	0	0.25	0.25	0.52
	$\rho = 0.5$	$\sigma^2 = 0.01$	0.86	NaN	NaN	0.03	0.86	NaN	NaN	NaN
		$\sigma^2 = 4$	0	0.85	0.85	0	0	0.85	0.85	0.95
	$\rho = 0.9$	$\sigma^2 = 0.01$	0.56	NaN	NaN	0	0.76	NaN	NaN	NaN
		$\sigma^2 = 4$	0	0.67	0.36	0	0	0.77	0.17	0.65

Table A8. Variable selection: $P_c(\lambda_{1se}) - P_c(\lambda_{min})$: constant correlation.

			Low-Dimension				High-Dimension			
			$\beta = 0.1$	$\beta = 1$	$\beta = 3$	Decay β	$\beta = 0.1$	$\beta = 1$	$\beta = 3$	Decay β
$n = 100$	$\rho = 0.5$	$\sigma^2 = 0.01$	0	0	0.1	0	0	0	0	0
		$\sigma^2 = 4$	0	0	0	0	0	0	0	0
	$\rho = 0.9$	$\sigma^2 = 0.01$	0	0	0	0	0	0	0	0
		$\sigma^2 = 4$	0	0	0	0	0	0	0	0
$n = 1000$	$\rho = 0.5$	$\sigma^2 = 0.01$	0	NaN	NaN	0.01	0	NaN	NaN	0
		$\sigma^2 = 4$	0	0	0	0	0	0	0	0
	$\rho = 0.9$	$\sigma^2 = 0.01$	0	NaN	NaN	0.01	0	NaN	NaN	0
		$\sigma^2 = 4$	0	0	0	0	0	0	0	0

Figure A3. Boston Housing Price: $|S \nabla \hat{S}|; S = var_{min}$, (a) $\hat{S} = var_{min}$; (b) $\hat{S} = var_{1se}$.

Figure A4. Boston Housing Price: $|S \nabla \hat{S}|; S = var_{1se}$, (a) $\hat{S} = var_{min}$; (b) $\hat{S} = var_{1se}$.

Figure A5. Bardet-Biedl data: $|S \nabla \hat{S}|; S = var_{min}$, (a) $\hat{S} = var_{min}$; (b) $\hat{S} = var_{1se}$.

Figure A6. Bardet-Biedl data: $|S \nabla \hat{S}|$; $S = \text{var}_{1se}$, (a) $\hat{S} = \text{var}_{min}$; (b) $\hat{S} = \text{var}_{1se}$.

Appendix B. Proof of the Main Result

Appendix B.1. A Lemma

Lemma A1. Let $W_1, W_2, \dots, W_K$ have identical variance and common pairwise correlation ρ . Then

$$\mathbb{E} \frac{1}{K-1} \sum_{i=1}^K (W_i - \bar{W})^2 = (1 - \rho) \text{Var}(W_i). \quad (\text{A1})$$

Remark A1. Depending on $\rho = 0$, $\rho > 0$ or $\rho < 0$, the sample variance is unbiased, under-estimating, or over-estimating the variance of W_i . For a general learning procedure, the correlation between the test errors from different folds can be both positive and negative, which makes the estimated standard error based on the usual sample variance formula under-estimating or over-estimating the true standard error. This is seen in our numerical results.

Remark A2. Note that for the sample mean $\bar{W}$, we have $\text{Var}(\bar{W}) = \text{Var}(W_1)/K + \text{Var}(W_1)\rho(K-1)/K$. Thus for $\text{Var}(W_1)/K$ to be asymptotically valid (in approximation) for $\text{Var}(\bar{W})$ in the sense that their ratio approaches 1, we must have $\rho \rightarrow 0$, which is also the key requirement for the sample variance formula to be asymptotically valid (ARU).

Appendix B.2. Proof of Lemma A1

Since the sample variance stays the same with a common shift of the observations, without loss of generality, we assume $\mathbb{E} W_i = 0$. Then

$$\begin{aligned} S^2 &= \frac{1}{K-1} \sum_{i=1}^K (W_i - \bar{W})^2 \\ &= \frac{1}{K-1} \left(\sum_{i=1}^K W_i^2 - K\bar{W}^2 \right) \\ &= \frac{1}{K-1} \left(\sum_{i=1}^K W_i^2 - \frac{1}{K} \sum_{i=1}^K W_i^2 - \frac{2}{K} \sum_{1 \leq i < j \leq K} W_i W_j \right) \\ &= \frac{1}{K} \sum_{i=1}^K W_i^2 - \frac{2}{K(K-1)} \sum_{i < j} W_i W_j. \end{aligned} \quad (\text{A2})$$

Thus we conclude $\mathbb{E} S^2 = (1 - \rho) \text{Var}(W_i)$.

Appendix B.3. Proof of Theorem 1

From Lemma A1, we only need to show that the correlation between the test errors on any two folds approaches 0 as $n \rightarrow \infty$. In the following, we examine the covariance between two test errors and show it is negligible compared to the variance of the test error under the condition in the theorem.

Let $W_1, \dots, W_K$ denote the CV test errors for the k -th fold, $k = 1, 2, \dots, K$. Note $W_k = \sum_{j \in F_k} (Y_j - \hat{f}_{(-F_k)}(X_j))^2$, where $\hat{f}_{(-F_k)}$ denote the estimator of f by the learning procedure $\hat{\delta}$ based on $\mathbb{D}_{(-F_k)}^n = \{(x_i, y_i) : 1 \leq i \leq n, i \notin F_k\}$. Below, we first calculate the variance of W_1 and then bound $\rho(W_{k_1}, W_{k_2})$ with $k_1 \neq k_2$. Note that $\text{Var}(W_k)$ is equal to

$$\begin{aligned}
 & \text{Var} \left(\sum_{j \in F_k} \left(f(X_j) + \epsilon_j - \hat{f}_{(-F_k)}(X_j) \right)^2 \right) \\
 &= \text{Var} \left(\sum_{j \in F_k} \left(f(X_j) - \hat{f}_{(-F_k)}(X_j) \right)^2 + \sum_{j \in F_k} \epsilon_j^2 + 2 \sum_{j \in F_k} \left(f(X_j) - \hat{f}_{(-F_k)}(X_j) \right) \epsilon_j \right) \\
 &= \text{Var} \left(\sum_{j \in F_k} \left(f(X_j) - \hat{f}_{(-F_k)}(X_j) \right)^2 \right) + \text{Var} \left(\sum_{j \in F_k} \epsilon_j^2 \right) + 4 \text{Var} \left(\sum_{j \in F_k} \left(f(X_j) - \hat{f}_{(-F_k)}(X_j) \right) \epsilon_j \right) \\
 &\quad + 2 \text{Cov} \left(\sum_{j \in F_k} \left(f(X_j) - \hat{f}_{(-F_k)}(X_j) \right)^2, \sum_{j \in F_k} \epsilon_j^2 \right) \\
 &\quad + 4 \text{Cov} \left(\sum_{j \in F_k} \left(f(X_j) - \hat{f}_{(-F_k)}(X_j) \right)^2, \sum_{j \in F_k} \left(f(X_j) - \hat{f}_{(-F_k)}(X_j) \right) \epsilon_j \right) \\
 &\quad + 4 \text{Cov} \left(\sum_{j \in F_k} \epsilon_j^2, \sum_{j \in F_k} \left(f(X_j) - \hat{f}_{(-F_k)}(X_j) \right) \epsilon_j \right) \\
 &= \text{Var} \left(\sum_{j \in F_k} \left(f(X_j) - \hat{f}_{(-F_k)}(X_j) \right)^2 \right) + \sum_{j \in F_k} \text{Var}(\epsilon_j^2) + 4 \sum_{j \in F_k} \text{Var}(\epsilon_j) \mathbb{E} \|f - \hat{f}_{(-F_k)}\|_2^2 \\
 &\quad + 0 + 0 + 4 \sum_{j \in F_k} \mathbb{E} \epsilon_j^3 \left(f(X_j) - \hat{f}_{(-F_k)}(X_j) \right) \\
 &= \text{Var} \left(\sum_{j \in F_k} \left(f(X_j) - \hat{f}_{(-F_k)}(X_j) \right)^2 \right) + n_0 \text{Var}(\epsilon_j^2) + 4n_0 \sigma^2 \mathbb{E} \|f - \hat{f}_{(-F_k)}\|_2^2,
 \end{aligned} \tag{A3}$$

where we have used the assumption that the error ϵ_j is independent of X_j and $\mathbb{E} \epsilon_j^3 = 0$. Now,

$$\begin{aligned}
 & \text{Var} \left(\sum_{j \in F_k} \left(f(X_j) - \hat{f}_{(-F_k)}(X_j) \right)^2 \right) \\
 &= \mathbb{E} \left[\text{Var} \left(\sum_{j \in F_k} \left(f(X_j) - \hat{f}_{(-F_k)}(X_j) \right)^2 \middle| \mathbb{D}_{(-F_k)}^n \right) \right] + \text{Var} \left[\mathbb{E} \left(\sum_{j \in F_k} \left(f(X_j) - \hat{f}_{(-F_k)}(X_j) \right)^2 \middle| \mathbb{D}_{(-F_k)}^n \right) \right] \\
 &= \mathbb{E} n_0 \left[\text{Var} \left(\left(f(X_j) - \hat{f}_{(-F_k)}(X_j) \right)^2 \middle| \mathbb{D}_{(-F_k)}^n \right) \right] + \text{Var} \left(n_0 \|f - \hat{f}_{(-F_k)}\|_2^2 \right) \\
 &= n_0 \mathbb{E} \left(\|f - \hat{f}_{(-F_k)}\|_4^4 - \|f - \hat{f}_{(-F_k)}\|_2^4 \right) + n_0^2 \text{Var} \left(\|f - \hat{f}_{(-F_k)}\|_2^2 \right).
 \end{aligned} \tag{A4}$$

Therefore,

$$\begin{aligned}
 & \text{Var}(W_k) \\
 &= n_0 \mathbb{E} \left(\|f - \hat{f}_{(-F_k)}\|_4^4 - \|f - \hat{f}_{(-F_k)}\|_2^4 \right) + n_0^2 \text{Var} \left(\|f - \hat{f}_{(-F_k)}\|_2^2 \right) \\
 &\quad + n_0 \text{Var}(\epsilon_j^2) + 4n_0 \sigma^2 \mathbb{E} \|f - \hat{f}_{(-F_k)}\|_2^2.
 \end{aligned} \tag{A5}$$

Now for W_{k_1} and W_{k_2} , $k_1 \neq k_2$, we calculate $Cov(W_{k_1}, W_{k_2})$. Clearly, WLOG, we can take $k_1 = 1, k_2 = 2$.

$$\begin{aligned}
 W_1 &= \sum_{j \in F_1} \left(f(X_j) - \hat{f}_{(-F_1)}(X_j) + \epsilon_j \right)^2 \\
 &= \sum_{j \in F_1} \left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j) + \epsilon_j + \hat{f}_{(-F_1, -F_2)}(X_j) - \hat{f}_{(-F_1)}(X_j) \right)^2 \\
 &= \sum_{j \in F_1} \left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j) + \epsilon_j \right)^2 + \sum_{j \in F_1} \left(\hat{f}_{(-F_1, -F_2)}(X_j) - \hat{f}_{(-F_1)}(X_j) \right)^2 \\
 &\quad + 2 \sum_{j \in F_1} \left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j) + \epsilon_j \right) \left(\hat{f}_{(-F_1, -F_2)}(X_j) - \hat{f}_{(-F_1)}(X_j) \right) \\
 &\triangleq C + R_{1,1} + R_{1,2}.
 \end{aligned} \tag{A6}$$

Similarly,

$$\begin{aligned}
 W_2 &= \sum_{j \in F_2} \left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j) + \epsilon_j \right)^2 + \sum_{j \in F_2} \left(\hat{f}_{(-F_1, -F_2)}(X_j) - \hat{f}_{(-F_2)}(X_j) \right)^2 \\
 &\quad + 2 \sum_{j \in F_2} \left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j) + \epsilon_j \right) \left(\hat{f}_{(-F_1, -F_2)}(X_j) - \hat{f}_{(-F_2)}(X_j) \right) \\
 &\triangleq D + R_{2,1} + R_{2,2}.
 \end{aligned} \tag{A7}$$

Then

$$\begin{aligned}
 Cov(W_1, W_2) &= Cov(C, D) + Cov(C, R_{2,1}) + Cov(C, R_{2,2}) \\
 &\quad + Cov(R_{1,1}, D) + Cov(R_{1,1}, R_{2,1}) + Cov(R_{1,1}, R_{2,2}) \\
 &\quad + Cov(R_{1,2}, D) + Cov(R_{1,2}, R_{2,1}) + Cov(R_{1,2}, R_{2,2}).
 \end{aligned} \tag{A8}$$

For $Cov(C, D)$, we have

$$\begin{aligned}
 Cov(C, D) &= \mathbb{E} CD - (\mathbb{E} C)(\mathbb{E} D) \\
 &= \mathbb{E} \left[\sum_{j \in F_1} \left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j) + \epsilon_j \right)^2 \right] \left[\sum_{j \in F_2} \left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j) + \epsilon_j \right)^2 \right] \\
 &\quad - \left\{ \mathbb{E} \sum_{j \in F_1} \left[\left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j) \right)^2 + \epsilon_j^2 + 2\epsilon_j \left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j) \right) \right] \right\}^2.
 \end{aligned}$$

For the first part, we have

$$\begin{aligned}
 \mathbb{E} CD &= \mathbb{E} \mathbb{E} \left[CD | \mathbb{D}_{(-F_1, -F_2)}^n \right] \\
 &= \mathbb{E} \mathbb{E} \left[\sum_{j \in F_1} \left[\left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j) \right)^2 + \epsilon_j^2 + 2\epsilon_j \left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j) \right) \right] \right. \\
 &\quad \times \left. \sum_{j \in F_2} \left[\left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j) \right)^2 + \epsilon_j^2 + 2\epsilon_j \left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j) \right) \right] | \mathbb{D}_{(-F_1, -F_2)}^n \right] \\
 &= n_0^2 \mathbb{E} \|f - \hat{f}_{(-F_1, -F_2)}\|_2^4 + n_0^2 \sigma^4 + 2n_0^2 \sigma^2 \mathbb{E} \|f - \hat{f}_{(-F_1, -F_2)}\|_2^2,
 \end{aligned}$$

and

$$\begin{aligned}
 & (\mathbb{E}C)(\mathbb{E}D) \\
 &= (\mathbb{E}C)^2 \\
 &= \left[n_0 \left(\mathbb{E} \|f - \hat{f}_{(-F_1, -F_2)}\|_2^2 + \sigma^2 \right) \right]^2 \\
 &= n_0^2 \left(\mathbb{E} \|f - \hat{f}_{(-F_1, -F_2)}\|_2^2 \right)^2 + 2n_0^2 \sigma^2 \mathbb{E} \|f - \hat{f}_{(-F_1, -F_2)}\|_2^2 + n_0^2 \sigma^4.
 \end{aligned} \tag{A9}$$

Thus, $Cov(C, D) = n_0^2 Var\left(\|f - \hat{f}_{(-F_1, -F_2)}\|_2^2\right)$. For the other terms in the earlier expression of $Cov(W_1, W_2)$, they can be handled similarly. Indeed, $|Cov(C, R_{2,1})| \leq \sqrt{Var(C)}\sqrt{Var(R_{2,1})}$ and the other terms are bounded similarly.

Next, we calculate $Var(C)$. Actually, the calculation is basically the same as for $Var(W_k)$ and we get

$$\begin{aligned}
 Var(C) = & n_0 \mathbb{E} \left(\|f - \hat{f}_{(-F_1, -F_2)}\|_4^4 - \|f - \hat{f}_{(-F_1, -F_2)}\|_2^4 \right) + n_0^2 Var\left(\|f - \hat{f}_{(-F_1, -F_2)}\|_2^2\right) \\
 & + n_0 Var(\epsilon_j^2) + 4n_0 \sigma^2 \mathbb{E} \|f - \hat{f}_{(-F_1, -F_2)}\|_2^2.
 \end{aligned}$$

Note that

$$\begin{aligned}
 Var(R_{1,1}) &= Var\left(\sum_{j \in F_1} \left(\hat{f}_{(-F_1, -F_2)}(X_j) - \hat{f}_{(-F_1)}(X_j)\right)^2\right) \\
 &= \mathbb{E}\left(Var\left(\sum_{j \in F_1} \left(\hat{f}_{(-F_1, -F_2)}(X_j) - \hat{f}_{(-F_1)}(X_j)\right)^2 \middle| \mathbb{D}_{(-F_1)}^n\right)\right) \\
 &\quad + Var\left(\mathbb{E}\left(\sum_{j \in F_1} \left(\hat{f}_{(-F_1, -F_2)}(X_j) - \hat{f}_{(-F_1)}(X_j)\right)^2 \middle| \mathbb{D}_{(-F_1)}^n\right)\right) \\
 &= \mathbb{E} n_0 \left(\|\hat{f}_{(-F_1, -F_2)} - \hat{f}_{(-F_1)}\|_4^4 - \|\hat{f}_{(-F_1, -F_2)} - \hat{f}_{(-F_1)}\|_2^4 \right) \\
 &\quad + n_0^2 Var\left(\|\hat{f}_{(-F_1, -F_2)} - \hat{f}_{(-F_1)}\|_2^2\right) \\
 &\leq n_0 \mathbb{E} \|\hat{f}_{(-F_1, -F_2)} - \hat{f}_{(-F_1)}\|_4^4 + n_0^2 \mathbb{E} \|\hat{f}_{(-F_1, -F_2)} - \hat{f}_{(-F_1)}\|_2^4.
 \end{aligned} \tag{A10}$$

For the other term,

$$\begin{aligned}
\text{Var}(R_{1,2}) &= 4\text{Var}\left(\sum_{j \in F_1} \left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j) + \epsilon_j\right) \left(\hat{f}_{(-F_1, -F_2)}(X_j) - \hat{f}_{(-F_1)}(X_j)\right)\right) \\
&= 4\mathbb{E} \text{Var}\left(\sum_{j \in F_1} \left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j) + \epsilon_j\right) \left(\hat{f}_{(-F_1, -F_2)}(X_j) - \hat{f}_{(-F_1)}(X_j)\right) | \mathbb{D}_{(-F_1)}^n\right) \\
&\quad + 4\text{Var}\left(\mathbb{E}\left(\sum_{j \in F_1} \left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j) + \epsilon_j\right) \left(\hat{f}_{(-F_1, -F_2)}(X_j) - \hat{f}_{(-F_1)}(X_j)\right) | \mathbb{D}_{(-F_1)}^n\right)\right) \\
&= 4\mathbb{E}\left(n_0 \text{Var}\left(\left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j) + \epsilon_j\right) \left(\hat{f}_{(-F_1, -F_2)}(X_j) - \hat{f}_{(-F_1)}(X_j)\right) | \mathbb{D}_{(-F_1)}^n\right)\right) \\
&\quad + 4\text{Var}\left(n_0 \mathbb{E}\left(\left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j)\right) \left(\hat{f}_{(-F_1, -F_2)}(X_j) - \hat{f}_{(-F_1)}(X_j)\right) | \mathbb{D}_{(-F_1)}^n\right)\right) \\
&= 4n_0 \mathbb{E} \text{Var}\left(\left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j)\right) \left(\hat{f}_{(-F_1, -F_2)}(X_j) - \hat{f}_{(-F_1)}(X_j)\right) | \mathbb{D}_{(-F_1)}^n\right) \\
&\quad + 4n_0 \mathbb{E} \sigma^2 \text{Var}\left(\|\hat{f}_{(-F_1, -F_2)} - \hat{f}_{(-F_1)}\|_2^2\right) \\
&\quad + 4n_0^2 \text{Var}\left(\mathbb{E}\left[\left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j)\right) \left(\hat{f}_{(-F_1, -F_2)}(X_j) - \hat{f}_{(-F_1)}(X_j)\right) | \mathbb{D}_{(-F_1)}^n\right]\right) \\
&\leq 4n_0 \mathbb{E}\left(\mathbb{E}\left(\left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j)\right)^2 \left(\hat{f}_{(-F_1, -F_2)}(X_j) - \hat{f}_{(-F_1)}(X_j)\right)^2 | \mathbb{D}_{(-F_1)}^n\right)\right) \\
&\quad + 4n_0 \sigma^2 \mathbb{E}\left(\|\hat{f}_{(-F_1, -F_2)} - \hat{f}_{(-F_1)}\|_2^2\right) \\
&\quad + 4n_0^2 \mathbb{E}\left(\mathbb{E}\left(\left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j)\right)^2 | \mathbb{D}_{(-F_1)}^n\right) \mathbb{E}\left(\left(\hat{f}_{(-F_1, -F_2)}(X_j) - \hat{f}_{(-F_1)}(X_j)\right)^2 | \mathbb{D}_{(-F_1)}^n\right)\right) \\
&\leq 4n_0 \mathbb{E}\left(\mathbb{E}\left(\left(f(X_j) - \hat{f}_{(-F_1, -F_2)}(X_j)\right)^4 | \mathbb{D}_{(-F_1)}^n\right)\right)^{\frac{1}{2}} \left(\mathbb{E}\left(\left(\hat{f}_{(-F_1, -F_2)}(X_j) - \hat{f}_{(-F_1)}(X_j)\right)^4 | \mathbb{D}_{(-F_1)}^n\right)\right)^{\frac{1}{2}} \\
&\quad + 4n_0 \sigma^2 \mathbb{E}\|\hat{f}_{(-F_1, -F_2)} - \hat{f}_{(-F_1)}\|_2^2 + 4n_0^2 \mathbb{E}\left(\|f - \hat{f}_{(-F_1, -F_2)}\|_2^2 \|\hat{f}_{(-F_1, -F_2)} - \hat{f}_{(-F_1)}\|_2^2\right) \\
&\leq 4n_0 \mathbb{E}\left(\|f - \hat{f}_{(-F_1, -F_2)}\|_4^2 \|\hat{f}_{(-F_1, -F_2)} - \hat{f}_{(-F_1)}\|_4^2\right) + 4n_0 \sigma^2 \mathbb{E}\|\hat{f}_{(-F_1, -F_2)} - \hat{f}_{(-F_1)}\|_2^2 \\
&\quad + 4n_0^2 \mathbb{E}\|f - \hat{f}_{(-F_1, -F_2)}\|_2^2 \|\hat{f}_{(-F_1, -F_2)} - \hat{f}_{(-F_1)}\|_2^2 \\
&\leq 4n_0 \left(\mathbb{E}\|f - \hat{f}_{(-F_1, -F_2)}\|_4^4\right)^{\frac{1}{2}} \left(\mathbb{E}\|\hat{f}_{(-F_1, -F_2)} - \hat{f}_{(-F_1)}\|_4^4\right)^{\frac{1}{2}} + 4n_0 \sigma^2 \mathbb{E}\|\hat{f}_{(-F_1, -F_2)} - \hat{f}_{(-F_1)}\|_2^2 \\
&\quad + 4n_0^2 \left(\mathbb{E}\|f - \hat{f}_{(-F_1, -F_2)}\|_2^4\right)^{\frac{1}{2}} \left(\mathbb{E}\|\hat{f}_{(-F_1, -F_2)} - \hat{f}_{(-F_1)}\|_2^4\right)^{\frac{1}{2}}.
\end{aligned} \tag{A11}$$

Under the condition $\mathbb{E}\|f - \hat{f}_{(\delta, n)}\|_4^4 = o(1/n)$, it is seen that $\text{Var}(W_1)$ and $\text{Var}(C)$ are both of order n , $\text{Var}(R_{1,1})$, $\text{Var}(R_{1,2})$, and $\text{Cov}(C, D)$ are all of order $o(n)$, which together imply $|\text{Cov}(W_1, W_2)| = o(\text{Var}(W_1))$ and consequently $\rho(W_1, W_2) \rightarrow 0$ as $n \rightarrow \infty$. This completes the proof of the theorem.

References

1. Nan, Y.; Yang, Y. Variable selection diagnostics measures for high-dimensional regression. *J. Comput. Graph. Stat.* **2014**, *23*, 636–656. [\[CrossRef\]](#)
2. Yu, Y.; Yang, Y.; Yang, Y. Performance assessment of high-dimensional variable identification. *arXiv* **2017**, arXiv:1704.08810.
3. Ye, C.; Yang, Y.; Yang, Y. Sparsity oriented importance learning for high-dimensional linear regression. *J. Am. Stat. Assoc.* **2018**, *113*, 1797–1812. [\[CrossRef\]](#)
4. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. (Methodol.)* **1996**, *58*, 267–288. [\[CrossRef\]](#)
5. Tikhonov, A.N. On the stability of inverse problems. *Dokl. Akad. Nauk SSSR* **1943**, *39*, 195–198.
6. Horel, A.E. Applications of ridge analysis to regression problems. *Chem. Eng. Progress* **1962**, *58*, 54–59.
7. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [\[CrossRef\]](#)
8. Hurvich, C.M.; Tsai, C.L. Regression and time series model selection in small samples. *Biometrika* **1989**, *76*, 297–307. [\[CrossRef\]](#)
9. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [\[CrossRef\]](#)

10. Chen, J.; Chen, Z. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **2008**, *95*, 759–771. [[CrossRef](#)]
11. Wang, H.; Li, R.; Tsai, C.L. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **2007**, *94*, 553–568. [[CrossRef](#)]
12. Allen, D.M. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **1974**, *16*, 125–127. [[CrossRef](#)]
13. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. (Methodol.)* **1974**, *36*, 111–133. [[CrossRef](#)]
14. Geisser, S. The predictive sample reuse method with applications. *J. Am. Stat. Assoc.* **1975**, *70*, 320–328. [[CrossRef](#)]
15. Zhang, Y.; Yang, Y. Cross-validation for selecting a model selection procedure. *J. Econom.* **2015**, *187*, 95–112. [[CrossRef](#)]
16. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Routledge: Oxfordshire, UK, 2017.
17. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer Series in Statistics; Springer: Berlin/Heidelberg, Germany, 2009.
18. Yang, Y. Comparing learning methods for classification. *Stat. Sin.* **2006**, *16*, 635–657.
19. Harrison, D., Jr.; Rubinfeld, D.L. Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Manag.* **1978**, *5*, 81–102. [[CrossRef](#)]
20. Scheetz, T.E.; Kim, K.Y.; Swiderski, R.E.; Philp, A.R.; Braun, T.A.; Knudtson, K.L.; Dorrance, A.M.; DiBona, G.F.; Huang, J.; Casavant, T.L.; et al. Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 14429–14434. [[CrossRef](#)] [[PubMed](#)]
21. Meinshausen, N.; Bühlmann, P. Stability selection. *J. R. Stat. Soc. Ser. (Stat. Methodol.)* **2010**, *72*, 417–473. [[CrossRef](#)]
22. Lim, C.; Yu, B. Estimation stability with cross-validation (ESCV). *J. Comput. Graph. Stat.* **2016**, *25*, 464–492. [[CrossRef](#)]
23. Yang, W.; Yang, Y. Toward an objective and reproducible model choice via variable selection deviation. *Biometrics* **2017**, *73*, 20–30. [[CrossRef](#)] [[PubMed](#)]