International Journal of Statistics and Probability; Vol. 12, No. 2; March 2023
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Olsavs: A New Algorithm For Model Selection

Nicklaus T. Hicks[1] & Hasthika S. Rupasinghe Arachchige Don[1]

[1] Department of Mathematical Sciences, Appalachian State University, USA

Correspondence: Hasthika S. Rupasinghe Arachchige Don, Department of Mathematical Sciences, Appalachian State University, Boone, NC, USA

Received: January 1, 2023 Accepted: March 11, 2023 Online Published: March 20, 2023

doi:10.5539/ijsp.v12n2p28 URL: https://doi.org/10.5539/ijsp.v12n2p28

Abstract

The shrinkage methods such as Lasso and Relaxed Lasso introduce some bias in order to reduce the variance of the regression coefficients in multiple linear regression models. One way to reduce bias after shrinkage of the coefficients would be to apply ordinary least squares to the subset of predictors selected by the shrinkage method used. This work extensively investigated this idea and developed a new variable selection algorithm. The authors named this technique OLSAVS (**O**rdinary **L**east **S**quares **A**fter **V**ariable **S**election). The OLSAVS algorithm was implemented in *R*. Simulations were used to illustrate that the new method is able to produce better predictions with less bias for various error distributions. The OLSAVS method was compared with a few widely used shrinkage methods in terms of their achieved test root mean square error and bias.

Keywords: Multiple Linear Regression, OLS, Lasso, Relax Lasso, Elastic Net, Bias, Variance

1. Introduction

Following (Pelawa Watagoda et al., 2021) and (Pelawa Watagoda, 2018), suppose that the response variable Y_i and at least one predictor variable $x_{i,j}$ are quantitative with $x_{i,1} \equiv 1$. Let $\boldsymbol{x}_i^T = (x_{i,1}, ..., x_{i,p}) = (1 \ \ \boldsymbol{u}_i^T)$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ where β_1 corresponds to the intercept. Then the multiple linear regression (MLR) model is

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{1}$$

for $i = 1, ..., n$. This model is also called the full model. Here n is the sample size, and assume that the random variables e_i are independent and identically distributed (iid) with variance $V(e_i) = \sigma^2$.

In matrix notation, these n equations become

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{2}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of response variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & x_{13} & \dots & x_{1p} \\ 1 & x_{22} & x_{23} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & x_{n3} & \dots & x_{np} \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \tag{3}$$

The ith fitted value $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ and the ith residual $r_i = Y_i - \hat{Y}_i$ where $\hat{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}$.

2. Variable Selection

Variable selection is the search for a subset of predictor variables that can be deleted with little loss of information if n/p is large, and so the model with the remaining predictors is useful for prediction. Following (Olive and Hawkins, 2005) and (Pelawa Watagoda and Olive, 2021), a *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \tag{4}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Suppose that S is a subset of I and that model (5) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I \tag{5}$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \boldsymbol{0}$ if $S \subseteq I$.

3. Estimating Model Coefficients

The most common method of obtaining model coefficients ($\boldsymbol{\beta}$) is the ordinary least squares. There are many methods for estimating $\boldsymbol{\beta}$, including, Lasso by (Tibshirani, 1996), Elastic Net by (Zou and Hastie, 2005), Relaxed Lasso by (Meinshausen, 2007), and ridge regression by (Hoerl and Kennard, 1970).

One can obtain the the least squares estimates for $\beta_1, \beta_1, \ldots, \beta_p$ by minimizing (6)

$$Q = \sum_{i=1}^{n}(Y_i - \beta_1 - \beta_2 X_{i,2} - \ldots - \beta_p X_{i,p})^2 \tag{6}$$

Ridge Regression coefficient estimates $\hat{\beta}^R$, are values that minimizes,

$$\sum_{i=1}^{n}(y_i - \beta_1 - \sum_{j=2}^{p}\beta_j X_{ij})^2 + \lambda\sum_{j=2}^{p}\beta_j^2 = RSS + \lambda\sum_{j=2}^{p}\beta_j^2 \tag{7}$$

Where $\lambda \geq 0$ is a tuning parameter. The term, $\lambda\sum_{j=2}^{p}\beta_j^2$ is known as the shrinkage penalty. This penalty value is small when $\beta_1, ..., \beta_p$ are close to zero, sending the β_j values to zero but never reaching zero. For this reason, ridge regression includes all predictors p in the model. The Lasso regression minimizes a similar quantity as in (7), except the shrinkage penalty changed to $\lambda\sum_{j=1}^{p}|\beta_j|$. Unlike ridge regression, often, some of the Lasso coefficients $\hat{\beta}_j$ are exactly equal to zero.

Following (Meinshausen, 2007), Relaxed Lasso controls model selection and shrinkage estimation by two separate parameters λ and ϕ. The Relaxed Lasso estimator is defined for $\lambda \in [0, \infty)$ and $\phi \in (0, 1]$ as

$$\hat{\boldsymbol{\beta}}^{\lambda,\phi} = \underset{\beta}{\operatorname{argmin}}\, n^{-1}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \boldsymbol{X}_i^T\{\boldsymbol{\beta}\cdot\boldsymbol{1}_{\mathcal{M}_\lambda}\})^2 + \phi\lambda|\boldsymbol{\beta}|_1 \tag{8}$$

Where $\boldsymbol{1}_{\mathcal{M}_\lambda}$ is the indicator function on the set of variables $\mathcal{M}_\lambda \subseteq \{1, ..., p\}$ so that for all $k \in \{1, \ldots, p\}$

$$\boldsymbol{\beta}\cdot\boldsymbol{1}_{\mathcal{M}_\lambda} = \begin{cases} 0 & k \notin \mathcal{M}_\lambda \\ \boldsymbol{\beta}_k & k \in \mathcal{M}_\lambda \end{cases}$$

The Elastic Net estimator is defined as follows:

Given dataset $(\boldsymbol{y}, \boldsymbol{X})$, penalty parameter (λ_1, λ_2) and augmented data $(\boldsymbol{y}^*, \boldsymbol{X}^*)$

$$\boldsymbol{X}^*_{(n+p)\times p} = (1+\lambda_2)^{(-1/2)}\begin{pmatrix}\boldsymbol{X}\\ \sqrt{\lambda_2}\boldsymbol{I}\end{pmatrix}, \qquad \boldsymbol{y}^*_{(n+p)} = \begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{0}\end{pmatrix}$$

Let $\gamma = \dfrac{\lambda_1}{\sqrt{1+\lambda_2}}$ and $\boldsymbol{\beta}^* = \sqrt{1+\lambda_2}\boldsymbol{\beta}$. Then the naïve Elastic Net solves a Lasso-type problem

$$\hat{\boldsymbol{\beta}}^* = \underset{\beta^*}{\operatorname{argmin}}\, n^{-1}\sum_{i=1}^{n}(\boldsymbol{Y}_i^* - \boldsymbol{X}_i^{*T}\boldsymbol{\beta}^*)^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}}|\boldsymbol{\beta}^*|_1 \tag{9}$$

Naïve Elastic Net estimator is a two steps procedure: for each fixed λ_2 it first finds the ridge regression coefficients, and then it does the Lasso type shrinkage along the Lasso type solution path. As a result, the predictors will shrink unnecessarily (double shrinkage). This would not help to reduce the variances much and also it will introduce unnecessary extra bias, compared to the original Lasso or ridge. As a solution, it uses a correction factor $\sqrt{1+\lambda_2}$ to get the Elastic Net solutions.

Finally, the Elastic Net solutions can be written as $\hat{\boldsymbol{\beta}} = \sqrt{1+\lambda_2}\hat{\boldsymbol{\beta}}^*$.

4. A New Method for Model Selection: OLSAVS

The shrinkage methods such as Lasso and Relaxed Lasso introduce some bias in order to reduce the variance of the regression coefficients. As briefly mentioned in (Hastie et al., 2015), one way to reduce bias after shrinkage of the coefficients would be to apply ordinary least squares to the subset of predictors selected by the shrinkage method used. This work extensively explores this idea to develop a new variable selection algorithm. The authors named this technique Ordinary Least Squares After Variable Selection (OLSAVS). OLSAVS method was implemented in *R*. The set of functions can be found at `https://hasthika.github.io/olsvspack.txt`. The algorithm of the OLSAVS method is as follows:

Algorithm: Ordinary Least Squares After Variable Selection (OLSAVS)

Repeat: following steps with a different shrinkage method

1) Apply the first shrinkage method to $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$.
2) Obtain the k non-zero predictors selected by the shrinkage method in 1)
3) Apply Ordinary Least Squares on the subset of k predictors obtained in 2)

Stop

4) Select a single best model using cross-validated prediction error, C_p , (AIC), BIC, or adjusted R^2

5. Simulation

This section contains the simulation setup and the results.

5.1 Simulation Setup

The statistical software, R (see (R Core Team, 2020)) was used to generate $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. The regression parameters $\boldsymbol{\beta}$ were set to $(1, 1, \ldots, 1, 0, \ldots, 0)$ with $k+1$ ones, $p-k-1$ zeroes where p is the total number of predictors, and k is the number of non-trivial predictors. Then for a given regression method, the regression coefficients, $\hat{\boldsymbol{\beta}}$ were obtained using the proposed method. This process was repeated 5000 times (runs). For each run, the difference between the regression parameters $\boldsymbol{\beta}$ and the regression coefficients, $\hat{\boldsymbol{\beta}}$ were obtained using the Minkowski distance. The average difference (Diff) was calculated by averaging all 5000 runs. The test root mean square error was also obtained using a set of test observations and averaged over the 5000 runs (TRMSE). $p = n/5, n/2$, or $n-1$ were used as the total number of predictors and $k = 1, 19$, or $p-1$ as the number of non-trivial predictors in the model. As per the easiness of coding the relation $cor(x_i, x_j) = \rho = (2\psi + (p-3)\psi^2)/(1 + (p-2)\psi^2)$ was used, for $i \neq j$, where, x_i, x_j are non-trivial predictors. As ψ increases the correlation between preceptors, ρ grows. $\psi = 0, 0.3$ or 0.9 were used with five error distributions with zero mean.

1. $N(0, 1)$, the normal distribution with mean 0 and variance 1 which is commonly used in simulation studies.
2. t_3, a t distribution with degrees of freedom 3, one of the heavy-tailed distributions.
3. $EXP(1) - 1$, an exponential distribution with mean 0. This distribution is not very commonly used in simulations but found in many real-life situations, a non-symmetric error distribution
4. $uniform(-1, 1)$, a uniform distribution in the rage of -1 and 1.
5. $0.9N(0, 1) + 0.1N(0, 100)$, a mixture of normal distributions.

The simulation study was conducted in *R*.

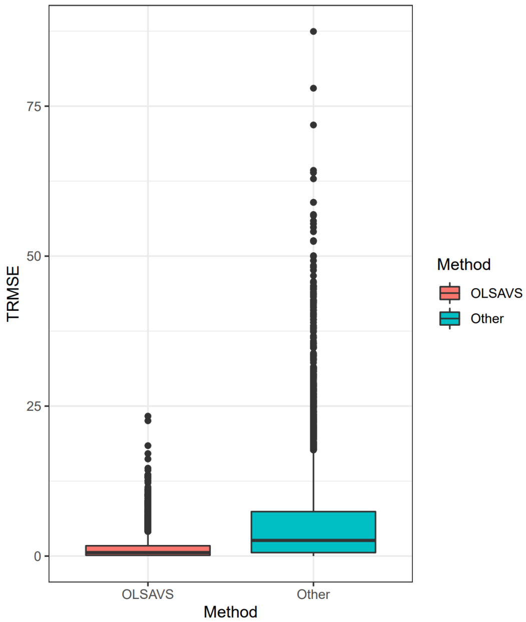
5.2 Simulation Results

Table 1 compares the OLSAVS method to Lasso with normal errors with mean 0 and variance 1. Notice in the TRMSE column, when the number of non-trivial predictors (k) is low, Lasso and the OLSAVS method perform equally well. However, as k and the correlation between the predictors increase, the OLSAVS method outperformed Lasso with noticeably larger distances between the OLSAVS and Lasso TRMSE values. This trend continues throughout the table. The OLSAVS stays consistent throughout for the TRMSE values, whereas the Lasso shows a lot of variability. Except for the case of $k = 1$ and $\psi = 0.9$, the Lasso either came close or betters OLSAVS in the difference value (Diff column). Other than this certain case, the OLSAVS method significantly bettered the Lasso.

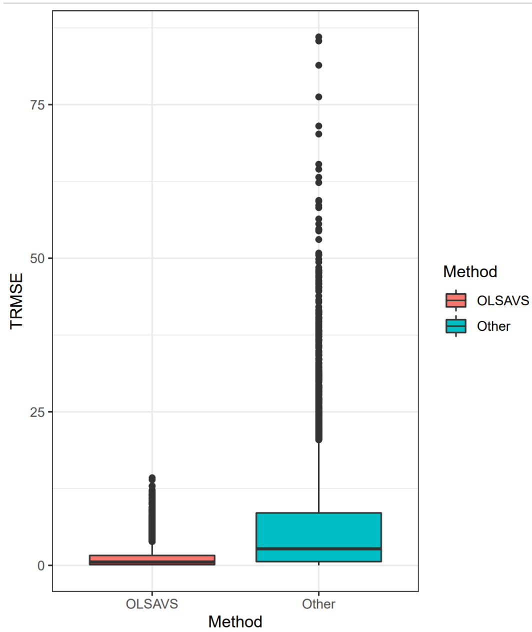
The side-by-side boxplots in Figure 1, compare the TRMSE results for two randomly selected rows in table 1 for $n = 100$ and $n = 200$. OLSAVS has a smaller median TRMSE and a lower variation in the results than the Lasso. It also appears

Table 1. TRMSE and difference values for OLSAVS vs. Lasso for $e_i \sim N(0, 1)$

				TRMSE		Diff	
n	p	k	ψ	OLSAVS	Lasso	OLSAVS	Lasso
100	20	1	0	1.0478	1.021	0.01657	0.1449
100	20	1	0.3	1.0417	1.0199	0.1124	0.1549
100	20	1	0.9	1.0083	1.0057	0.6455	0.6635
100	20	19	0	1.1112	1.1107	0.0037	0.0122
100	20	19	0.3	1.1112	1.1377	0.0055	0.0158
100	20	19	0.9	1.1543	2.3451	1.5667	7.5674
100	50	1	0	1.0916	1.0369	0.0268	0.181
100	50	1	0.3	1.0708	1.0303	0.1654	0.2028
100	50	1	0.9	1.0142	1.0059	0.9511	0.5929
100	50	49	0	1.3973	1.4014	0.0043	0.0149
100	50	49	0.3	1.5522	4.353	0.0163	0.0557
100	50	49	0.9	2.1118	9.4775	8.8016	43.9716
200	40	1	0	1.0386	1.0175	0.0225	0.1208
200	40	1	0.3	1.0316	1.0157	0.1044	0.1327
200	40	1	0.9	1.0108	1.005	0.7825	0.4165
200	40	39	0	1.117	1.119	0.002	0.0095
200	40	39	0.3	1.117	2.5606	0.0028	0.0369
200	40	39	0.9	1.6985	6.8533	5.8545	32.8388
200	100	1	0	1.0691	1.0261	0.0354	0.1454
200	100	1	0.3	1.0605	1.0251	0.1392	0.1653
200	100	1	0.9	1.0184	1.0045	0.9712	0.0855
200	100	99	0	1.4495	1.4561	0.0034	0.0132
200	100	99	0.3	3.166	11.4553	0.1379	0.488
200	100	99	0.9	3.7274	27.9646	17.4699	93.9076



(a) Table 1, Plot 1 (n = 100)



(b) Table 1, Plot 2 (n = 200)

Figure 1. Box plots for table 1 showing simulation results for OLSAVS vs. Lasso regression with error type 1

for this simulation, for Lasso with normal errors, when the sample size increases, so does the variation in the regression coefficients while OLSAVS maintains the same low variability.

Table 2 has the same structure as in Table 1 but the simulation was done with error type 2. The Lasso seems to narrowly outperform OLSAVS when $k = 1$. However, as the k increases, the TRMSE favored the OLSAVS method significantly. Additionally, as the correlation between the non-trivial predictors increases, notice the large distance increase between the two methods in the TRMSE. In the Diff column in Table 2, the OLSAVS method outperformed Lasso by a large majority.

Figure 2 shows two simulation plots pulled from table 2. Looking at two cases, each varying in sample size, the OLSAVS does edge-out Lasso regression with the error type being from a t-distribution. However, unlike in Figure 1, the variation in the plot now decreases as expected when the sample size increases.

Table 3 compares OLSAVS with Lasso with the error being from the exponential distribution. A similar trend occurs in the TRMSE column as in Table 2. As the non-trivial predictors are increased to $p - 1$ or the correlation between the predictors increased, OLSAVS produced lower TRMSE than Lasso. The OLSAVS seems to dominate the majority of the difference values.

Figure 3 shows boxplots gathered from two simulations in table 3. In the two plots shown, yet again the OLSAVS edges out the Lasso and has a much shorter variation in the box plot. Lasso with an exponential error does compete in the $n = 100$ plot but then has a much larger gap when the sample size is doubled.

Table 4 uses uniformly distributed errors with zero mean. Once again, the same trend appears in the TRMSE values between the OLSAVS and Lasso estimates as before. The difference for the simulations remains mostly the same as for previous simulations for Lasso with an exponential distribution.

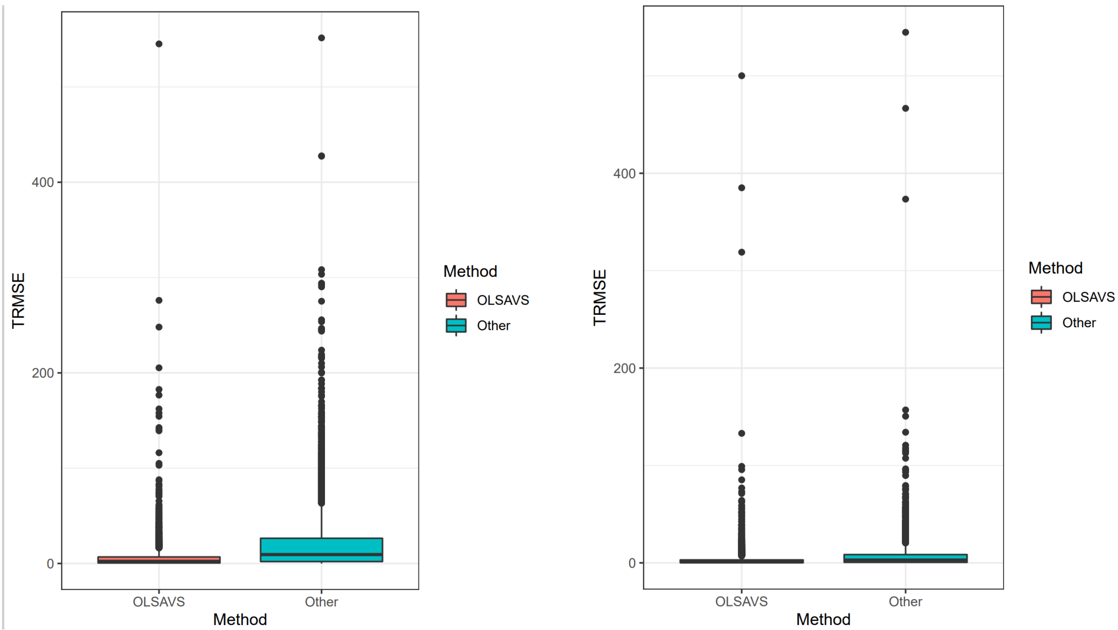
In figure 4, the variation for either method is large compared to figure 1, however, OLSAVS still has the smaller average TRMSE. Once the sample size increases, both the variation and the average TRMSE shrinks for each method, but the OLSAVS still maintains the advantage in each.

Tables 5 and 6, compare the OLSAVS method with Relaxed Lasso regression with normal errors and with exponential errors respectively. Much like the results for Lasso regression, The Relaxed Lasso appears to have a slight advantage when $k = 1$. However, once k increases, OLSAVS begins to have much smaller TRMSE values. However, the Relaxed Lasso is more competitive for normally distributed errors than the Lasso. The differences in table 6 show the OLSAVS method performing well in terms of bias of the regression coefficients. Additionally, the OLSAVS provides more consistent differences overall.

Figure 5 shows simulation results between OLSAVS and Relaxed Lasso with error type 1 while figure 6 shows results

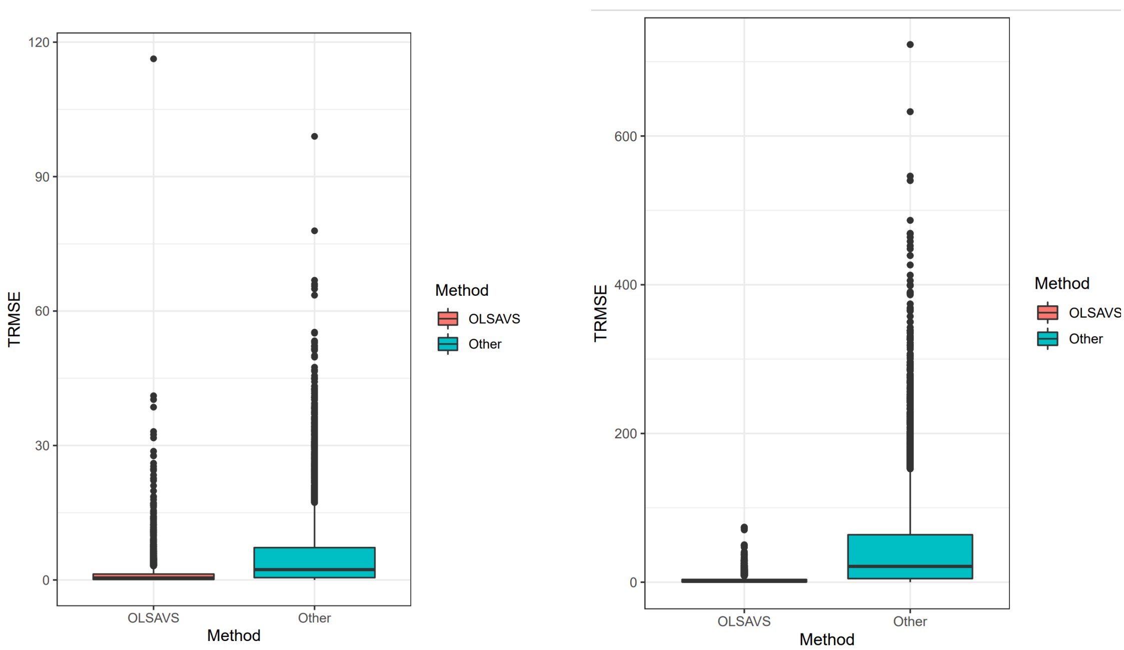
Table 2. TRMSE and difference values for OLSAVS vs. Lasso for $e_i \sim t_3$

				TRMSE		Diff	
n	p	k	ψ	OLSAVS	Lasso	OLSAVS	Lasso
100	20	1	0	1.3677	1.3362	0.0260	0.1882
100	20	1	0.3	1.3551	1.3317	0.1497	0.2031
100	20	1	0.9	1.3178	1.3169	0.6539	0.7482
100	20	19	0	1.4441	1.4442	0.0047	0.0130
100	20	19	0.3	1.4441	1.4590	0.0061	0.0142
100	20	19	0.9	1.4487	2.4245	1.8395	7.2843
100	50	1	0.3	1.3704	1.3311	0.2075	0.2549
100	50	1	0.9	1.3060	1.3015	0.9523	0.6991
100	50	49	0	1.8360	1.8418	0.0076	0.0076
100	50	49	0.3	1.9877	4.4636	0.0245	0.0245
100	50	49	0.9	2.2778	9.6242	8.8102	43.9438
200	40	1	0	1.3453	1.3165	0.0351	0.1550
200	40	1	0.3	1.3324	1.3163	0.1351	0.1707
200	40	1	0.9	1.3075	1.3033	0.8243	0.5671
200	40	39	0	1.4351	1.4355	0.0036	0.0107
200	40	39	0.3	1.4351	2.5826	0.0048	0.0376
200	40	39	0.9	1.8981	6.7108	5.7510	32.8242
200	100	1	0	1.3837	1.3189	0.0460	0.1847
200	100	1	0.3	1.3559	1.3153	0.1774	0.2115
200	100	1	0.9	1.3096	1.2998	1.1920	0.2210
200	100	99	0	1.8397	1.8487	0.0058	0.0147
200	100	99	0.3	3.3891	11.4710	0.1501	0.4976
200	100	99	0.9	3.8424	27.4098	17.2990	93.9071



(a) Table 2, Plot 1 ($n = 100$) (b) Table 2, Plot 2 ($n = 200$)

Figure 2. Box plots for table 2 showing simulation results for OLSAVS vs. Lasso regression with error type 2



(a) Table 3, Plot 1 ($n = 100$) (b) Table 3, Plot 2 ($n = 200$)

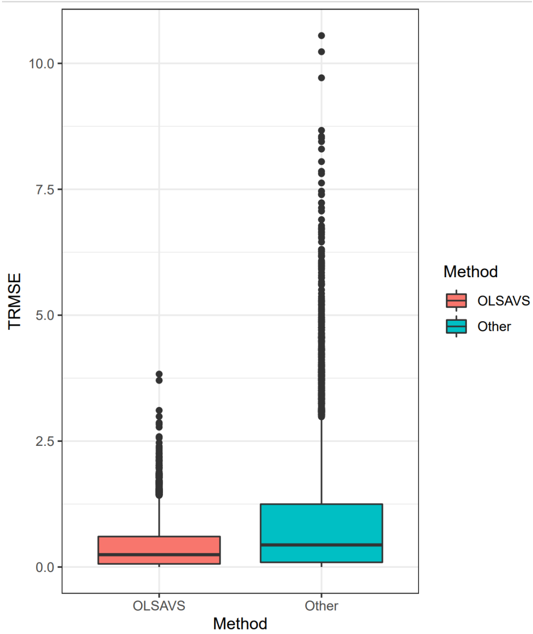
Figure 3. Box plots for table 3 showing simulation results for OLSAVS vs. Lasso regression with error type 3

Table 3. TRMSE and difference values for OLSAVS vs. Lasso for $e_i \sim exp(1) - 1$

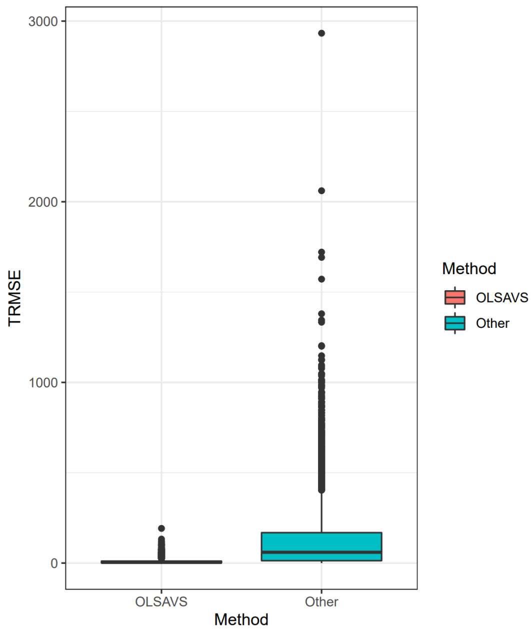
				TRMSE		Diff	
n	p	k	ψ	OLSAVS	Lasso	OLSAVS	Lasso
100	20	1	0	1.0589	1.0366	0.0208	0.1416
100	20	1	0.3	1.0287	1.0099	0.1123	0.1533
100	20	1	0.9	0.9913	0.9900	0.6264	0.6519
100	20	19	0	1.1055	1.1063	0.0033	0.0118
100	20	19	0.3	1.1055	1.1583	0.0046	0.0148
100	20	19	0.9	1.1335	2.3835	1.5415	7.5833
100	50	1	0	1.0948	1.0456	0.0232	0.1804
100	50	1	0.3	1.0767	1.0414	0.1689	0.2039
100	50	1	0.9	1.0263	1.0188	0.9343	0.5814
100	50	49	0	1.4078	1.4105	0.0045	0.0147
100	50	49	0.3	1.5744	4.3752	0.0153	0.0520
100	50	49	0.9	2.1137	9.5345	8.8138	44.0251
200	40	1	0	1.0298	1.0071	0.0202	0.1201
200	40	1	0.3	1.0202	1.0037	0.1063	0.1331
200	40	1	0.9	0.9969	0.9925	0.7664	0.4067
200	40	39	0	1.1034	1.1048	0.0026	0.0097
200	40	39	0.3	1.1034	2.4771	0.0035	0.0380
200	40	39	0.9	1.7076	6.7040	5.8761	32.8648
200	100	1	0	1.0522	1.0122	0.0371	0.0371
200	100	1	0.3	1.0356	1.0101	0.1402	0.1657
200	100	1	0.9	1.0067	0.9953	0.9636	0.0896
200	100	99	0	1.4138	1.4192	0.0037	0.0131
200	100	99	0.3	3.1658	11.2754	0.1451	0.4836
200	100	99	0.9	3.6705	27.4919	17.5606	93.9063

Table 4. TRMSE and difference values for OLSAVS vs. Lasso for $e_i \sim uniform(-1, 1)$

				TRMSE		Diff	
n	p	k	ψ	OLSAVS	Lasso	OLSAVS	Lasso
100	20	1	0	0.6087	0.5952	0.0112	0.0843
100	20	1	0.3	0.6024	0.5945	0.0662	0.0900
100	20	1	0.9	0.5905	0.5879	0.5773	0.4067
100	20	19	0	0.6360	0.6369	0.0023	0.0103
100	20	19	0.3	0.6360	0.9626	0.0031	0.0308
100	20	19	0.9	0.7599	2.3075	1.3762	7.8005
100	50	1	0	0.6265	0.5999	0.0177	0.1043
100	50	1	0.3	0.6169	0.5955	0.0943	0.1160
100	50	1	0.9	0.5917	0.5826	0.7318	0.2120
100	50	49	0	0.8221	0.8299	0.0028	0.0127
100	50	49	0.3	1.0467	4.3592	0.0121	0.0573
100	50	49	0.9	1.9089	9.4600	9.0251	44.0993
200	40	1	0	0.6078	0.5907	0.0127	0.0697
200	40	1	0.3	0.6016	0.5904	0.0604	0.0760
200	40	1	0.9	0.5886	0.5836	0.4935	0.1028
200	40	39	0	0.6506	0.6506	0.0013	0.0088
200	40	39	0.3	0.6506	2.5639	0.0018	0.0388
200	40	39	0.9	1.4729	6.8174	5.9630	33.0055
200	100	1	0	0.6130	0.5903	0.0123	0.0831
200	100	1	0.3	0.5967	0.5967	0.0898	0.0973
200	100	1	0.9	0.5902	0.5830	0.5646	0.0216
200	100	99	0	0.8125	0.8243	0.0021	0.0119
200	100	99	0.3	2.9615	11.3001	0.1460	0.4744
200	100	99	0.9	3.5458	27.3187	17.4181	93.9087

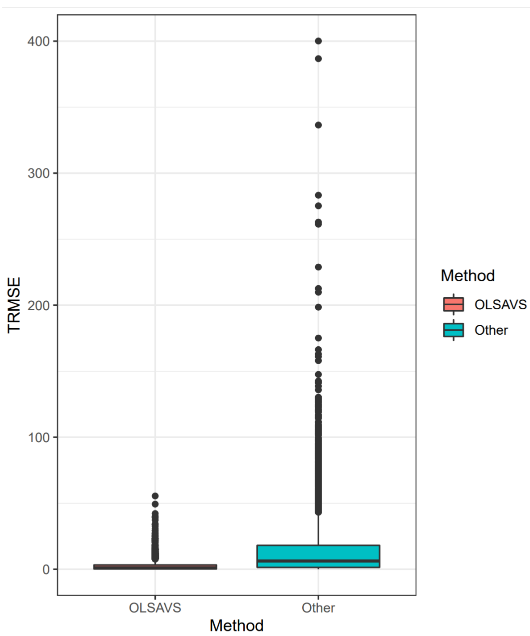


(a) Table 4, Plot 1 (n = 100)

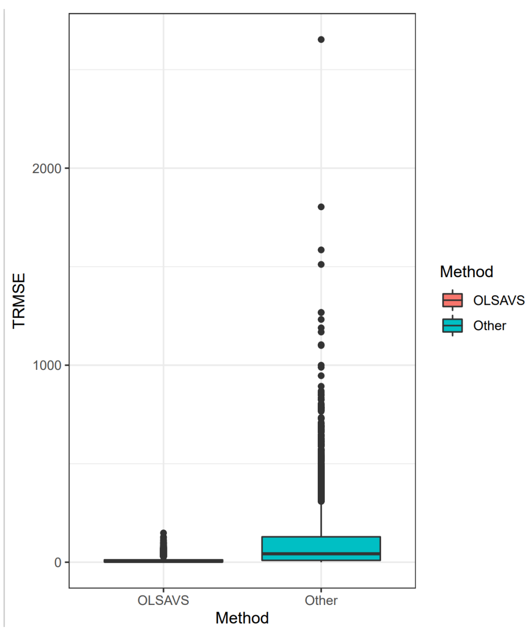


(b) Table 4, Plot 2 (n = 200)

Figure 4. Box plots for table 4 showing simulation results for OLSAVS vs. Lasso regression with error type 4



(a) Table 5, Plot 1 (n = 100)



(b) Table 5, Plot 2 (n = 200)

Figure 5. Box plots for table 5 showing simulation results for OLSAVS vs. Relaxed Lasso regression with error type 1

Table 5. TRMSE and difference values for OLSAVS vs. Relax Lasso for $e_i \sim N(0, 1)$

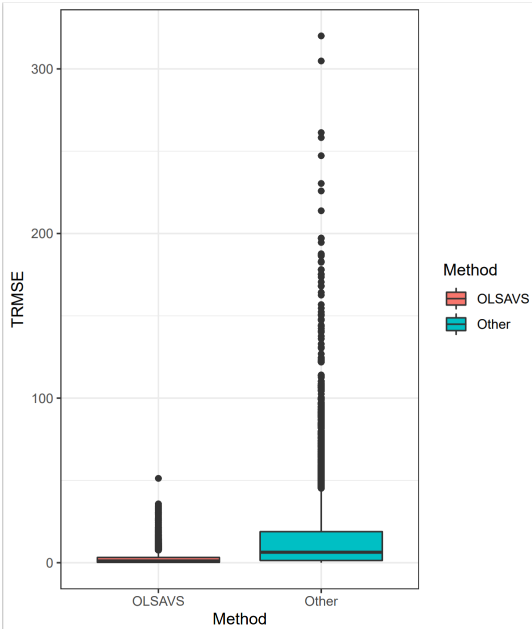
				TRMSE		Diff	
n	p	k	ψ	OLSAVS	R Lasso	OLSAVS	R Lasso
100	20	1	0	1.05004	1.03955	0.01180	0.04069
100	20	1	0.3	1.04307	1.03993	0.11183	0.11699
100	20	1	0.9	1.00809	1.00586	0.63942	0.67805
100	20	19	0	1.11518	1.11527	0.00315	0.00298
100	20	19	0.3	1.11518	1.12323	0.00384	0.00384
100	20	19	0.9	1.14518	1.44021	1.54564	7.55956
100	50	1	0	1.10563	1.08454	0.02597	0.05857
100	50	1	0.3	1.09096	1.08971	0.16601	0.16668
100	50	1	0.9	1.03093	1.02483	0.95682	0.60006
100	50	49	0	1.42151	1.42127	0.00415	0.00446
100	50	49	0.3	1.57851	3.89696	0.01567	0.01567
100	50	49	0.9	2.12296	4.80282	8.87674	45.66343
200	40	1	0	1.03855	1.03304	0.02126	0.02466
200	40	1	0.3	1.03295	1.03285	0.10517	0.10524
200	40	1	0.9	1.01005	1.00655	0.78436	0.42916
200	40	39	0	1.12531	1.12531	0.00267	0.00267
200	40	39	0.3	1.12531	1.80026	0.00386	0.12240
200	40	39	0.9	1.70105	3.57716	5.88467	33.41152
200	100	1	0	1.06191	1.05162	0.02599	0.03373
200	100	1	0.3	1.04883	1.04896	0.13610	0.13720
200	100	1	0.9	1.01643	1.00291	0.96788	0.96788
200	100	99	0	1.40641	1.40518	1.40518	0.00281
200	100	99	0.3	3.15561	10.08956	0.14074	0.92836
200	100	99	0.9	3.71833	9.86930	17.35998	97.95932

Note – R Lasso: Relaxed Lasso

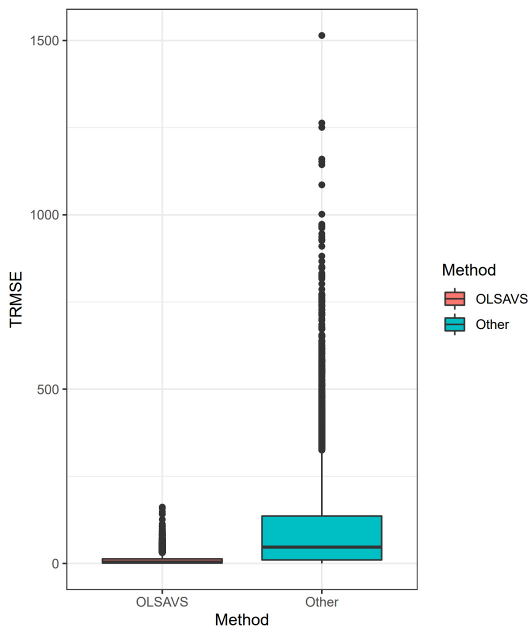
Table 6. TRMSE and difference values for OLSAVS vs. Relaxed Lasso for $e_i \sim EXP(1) - 1$

				TRMSE		Diff	
n	p	k	ψ	OLSAVS	R Lasso	OLSAVS	R Lasso
100	20	1	0	1.36487	1.34437	0.02908	0.08346
100	20	1	0.3	1.34847	1.34209	1.34209	0.15530
100	20	1	0.9	1.30332	1.30211	0.62132	0.74412
100	20	19	0	1.44316	1.44326	0.00553	0.00605
100	20	19	0.3	1.44316	1.44720	0.00764	0.03972
100	20	19	0.9	1.43871	1.65548	1.83681	7.26831
100	50	1	0	1.44446	1.40636	0.03604	0.10807
100	50	1	0.3	1.41096	1.40477	0.20648	0.21225
100	50	1	0.9	1.34176	1.33864	0.95980	0.71164
100	50	49	0	1.85165	1.85164	0.00554	0.00698
100	50	49	0.3	2.00740	4.14320	4.14320	4.14320
100	50	49	0.9	2.31574	4.91246	8.71224	45.59930
200	40	1	0	1.33118	1.31585	0.03086	0.05388
200	40	1	0.3	1.31769	1.31472	0.13591	0.13865
200	40	1	0.9	1.29278	1.28860	0.84014	0.58889
200	40	39	0	1.44431	1.44436	0.00272	0.00272
200	40	39	0.3	1.44431	2.02409	0.00388	0.13289
200	40	39	0.9	1.91426	3.72937	5.77462	33.17816
200	100	1	0	1.35272	1.32707	0.04958	0.07102
200	100	1	0.3	1.33008	1.32910	0.18092	0.18116
200	100	1	0.9	1.28972	1.27810	1.19952	0.21588
200	100	99	0	1.80251	1.80167	0.00513	0.00464
200	100	99	0.3	3.31817	10.32727	0.14145	0.92488
200	100	99	0.9	3.76114	9.86319	17.36192	97.95763

Note – R Lasso: Relaxed Lasso



(a) Table 6, Plot 1 (n = 100)



(b) Table 6, Plot 2 (n = 200)

Figure 6. Box plots for table 6 showing simulation results for OLSAVS vs. Relaxed Lasso regression with error type 2.

with error type 2. Figure 6 follows closely to the results from figure 5. The Relaxed Lasso seems to have a much larger variation for error types 1 and 2 compared to OLASVS.

6. Real Data Example

Wisconsin nursing home data set provided by the Wisconsin Department of Health and Family Services (DHFS) was used as the real data example, see (Rosenberg et al., 2007). The goal of this data set is to utilize nursing home capacity. The years 2000 and 2001 were considered, with 362 and 355 facilities respectively. However, 10 observations were removed for containing missing values. The data set contains 12 variables, with *total patient-years* (TPY) being the response variable.

To determine how OLSAVS performs, the data set was split into testing and training sets. 60% of the data was allocated to be in the training set and 40% was in the testing set. The TRMSE was recorded and compared the values of the OLSAVS method to those of Lasso, Relaxed Lasso, and Elastic Net.

Table 7, summarizes the results. Each of these side-by-side results shows that the OLSAVS method edges out each of the other methods considered for this real-world example.

Table 7. TRMSE comparison for OLSAVS vs. common methods using Wisconsin nursing home data

Method	**OLSAVS**	Elastic Net	**OLSAVS**	Lasso	**OLSAVS**	Relaxed Lasso
TRMSE	7.956877	8.071371	8.007678	8.044224	8.007678	8.013652

7. The *R* Package

The *R* function used for the simulation, a function we used to produce graphs, and a function for performing the OLSAVS regression can be found under `https://hasthika.github.io/olsvspack.txt`. To load this package: use `source("https://hasthika.github.io/olsvspack.txt")`

8. Conclusions

The new method for variable selection, OLSAVS involves applying ordinary least squares to a subset of predictors selected from a specific variable selection method such as Lasso, Relax Lasso, or Elastic Net. We expected the OLSAVS to reduce the bias in the regression coefficients introduced by the shrinkage method and lead to the model being close-fitting while keeping the consistency of the reduced variance from the shrinkage method. Simulation results show that the OLSAVS method not only reduced the bias of the regression coefficient but also further reduced the variance of the estimates.

Furthermore, the OLSAVS method performs well in terms of prediction error as well. As discussed in section 5.2, the test root mean square errors when using OLSAVS for all error types studied are either significantly low or equal to the competing shrinkage method. It is interesting to notice that the prediction accuracy drastically decreases as the correlation between the predictors increases when using commonly used shrinkage methods. Prediction accuracy decreases further with the number of non-trivial predictors. OLSAVS method outperformed the other shrinkage method studied in both of the scenarios mentioned above and produced much lower test root mean square error values.

Acknowledgment

The authors thank Dr. Lasanthi Watagoda and the reviewer for the comments that improved this article.

References

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman amp; Hall/CRC.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics, 12*(1), 5567.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics Data Analysis, 52*, 374393.

Olive, D. J., & Hawkins, D. M. (2005). Variable Selection for 1D Regression Models. *Technometrics, 47*(1), 4350. Publisher: Taylor & Francis eprint: https://doi.org/10.1198/004017004000000590

Pelawa Watagoda, L. C. R. (2018). A Sub-Model Theorem for Ordinary Least Squares. *International Journal of Statistics and Probability, 8*(1), 40.

Pelawa Watagoda, L. C. R., Arnholt, A. T., & Arachchige Don, H. S. R. (2021). HRLR regression. *RMS: Research in Mathematics & Statistics, 8*(1), 1921904. Publisher: Taylor & Francis eprint. https://doi.org/10.1080/27658449.2021.1921904

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rosenberg, M. A., Frees, E. W., Sun, J., Johnson, P. H., & Robinson, J. (2007). Predictive Modeling with Longitudinal Data. *North American Actuarial Journal, 11*(3), 5469. Publisher: Routledge eprint: https://doi.org/10.1080/10920277.2007.10597466.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267288. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.25176161.1996.tb02080.x.

Pelawa Watagoda, L. C. R., & Olive, D. J. (2021). Comparing six shrinkage estimators with large sample theory and asymptotically optimal prediction intervals. *Statistical Papers, 62*(5).

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301320.

Copyrights