Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 6607330, 10 pages
https://doi.org/10.1155/2022/6607330

Research Article

Evaluating the Performance of Feature Selection Methods Using Huge Big Data: A Monte Carlo Simulation Approach

Faridoon Khan,[1] Amena Urooj,[1] Saud Ahmed Khan,[1] Saima K. Khosa,[2] Sara Muhammadullah,[1] and Zahra Almaspoor[3]

[1]*PIDE School of Economics, Pakistan Institute of Development Economics, Islamabad, Pakistan*
[2]*Department of Statistics, Bahauddin Zakariya University, Multan, Pakistan*
[3]*Department of Statistics, Yazd University, Yazd, 89175-741, Iran*

Correspondence should be addressed to Zahra Almaspoor; z.almaspoor@stu.yazd.ac.ir

Received 7 September 2021; Revised 2 December 2021; Accepted 23 December 2021; Published 19 January 2022

Academic Editor: Caroline Mota

In this article, we compare autometrics and machine learning techniques including Minimax Concave Penalty (MCP), Elastic Smoothly Clipped Absolute Deviation (E-SCAD), and Adaptive Elastic Net (AEnet). For simulation experiments, three kinds of scenarios are considered by allowing the multicollinearity, heteroscedasticity, and autocorrelation conditions with varying sample sizes and the varied number of covariates. We found that all methods show improved their performance for a large sample size. In the presence of low and moderate multicollinearity and low and moderate autocorrelation, the considered methods retain all relevant variables. However, for low and moderate multicollinearity, excluding AEnet, all methods keep many irrelevant predictors as well. In contrast, under low and moderate autocorrelation, along with AEnet, the Autometrics retain less irrelevant predictors. Considering the case of extreme multicollinearity, AEnet retains more than 93 percent correct variables with an outstanding gauge (zero percent). However, the potency of remaining techniques, specifically MCP and E-SCAD, tends towards unity with augmenting sample size but capturing massive irrelevant predictors. Similarly, in case of high autocorrelation, E-SCAD has shown good performance in the selection of relevant variables for a small sample, while in gauge, Autometrics and AEnet are performed better and often retained less than 5 percent irrelevant variables. In the presence of heteroscedasticity, all techniques often hold all relevant variables but also suffer from overspecification problems except AEnet and Autometrics which circumvent the irrelevant predictors and establish the true model precisely. For an empirical application, we take into account the workers' remittance data for Pakistan along its twenty-seven determinants spanning from 1972 to 2020 for Pakistan. The AEnet selected thirteen relevant covariates of workers' remittance while E-SCAD and MCP suffered from an overspecification problem. Hence, the policymakers and practitioners should focus on the relevant variables selected by AEnet to improve workers' remittance in the case of Pakistan. In this regard, the Pakistan government has devised policies that make it easy to transfer remittances legally and mitigate the cost of transferring remittances from abroad. The AEnet approach can help policymakers arrive at relevant variables in the presence of a huge set of covariates, which in turn produce accurate predictions.

1. Introduction

"Big Data" has arrived, but big insights have not [1]. In regression analysis, researchers are often interested in discovering the important features while predicting the response variable. Therefore, it is important to identify the potential features for knowledge discovery and the predictive ability of the model [2]. However, variable selection is one of the crucial steps while constructing a linear regression model. Picking too many covariates is more likely to enhance the variance of the estimated or trained model. Stated differently, including more variables in the model leads to high variability in the least squares fit, resulting in overfitting and thus providing poor prediction in the future [3]. In contrast, selecting a few covariates may result in unpredictable output or biased results [3, 4]. As [5] stated that for valid results, all relevant predictors should be incorporated in the regression model. Missing a single predictor might

lead to a misspecified model and the conclusion we draw can be fallacious. According to [6, 7], if the covariates are highly interrelated to each other, then the confidence interval associated with each estimated coefficient becomes wider and leads to wrong inferences.

In the recent era, a substantial mass of research has concentrated on the analysis of "Big Data" in the field of economics. As a result, a substantial focus is being paid to the wide variety of techniques that are available in the areas of data mining, machine learning, dimension reduction, and penalized least squares [8, 9]. Recently, in the regression context, [1] categorized Big Data into three classes: Tall Big Data, Huge Big Data, and Fat Big Data. Each type can be defined as follows:

(i) Tall Big Data: more observations and several covariates ($N >> P$)

(ii) Huge Big Data: more observations and more covariates ($N > P$)

(iii) Fat Big Data: fewer observations and more covariates ($N < P$)

Here, N and P represent the number of observations and covariates, respectively. We graphically represent the types of Big Data in Figure 1.

It is quite obvious that Big Data's handling is not an easy task and to date in literature, there exist just a couple of methods, which can be utilized for improving the least squares estimates under a data-rich environment (Big Data). In Figure 2, we identify all common methods and their modification.

Now, we briefly discuss these methods. Penalized least square methods are an integral component of machine learning (ML). It has already been shown in the literature that ML methods are efficient approaches for using Big Data [10]. Penalized regression methods are the modified form of ordinary least squares (OLS). Mathematically we can write the modified form:

$$\sum_{c=1}^{n}\left(y_c - \alpha_0 - \sum_{d=1}^{m}\alpha_d x_{cd}\right)^2 + k * \vartheta \sum_{d=1}^{m}\left|\alpha_d\right| + k * (1-\vartheta)\sum_{d=1}^{m}\alpha_d^2. \tag{1}$$

Like in classical regression, the first component is the sum of squared residuals and the remaining part represents the shrinkage penalty. Here "k" refers to the tuning parameter and is often selected by cross-validation. The other parameter is ϑ; hence by altering its value, we get different models. More specifically, equating $\vartheta = 0$, results in ridge regression model form and if $\vartheta = 1$ is taken as there in Lasso regression. While for the value of ϑ between zero and one, we get the model for elastic net [6]. As its name reflects penalized least square methods are based on some constraints. A good penalty consists of the following three oracle properties: unbiasedness, continuity, and sparsity [11]. Methods belonging to the family of penalized regression like ridge, Lasso and Elastic net do not satisfy all the aforementioned oracle properties [12, 13]. Although in the literature, some modified methods satisfy

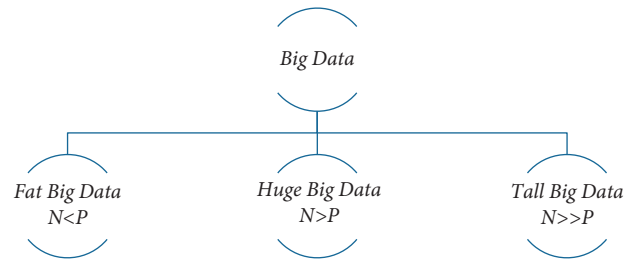

Figure 1: Types of Big Data.

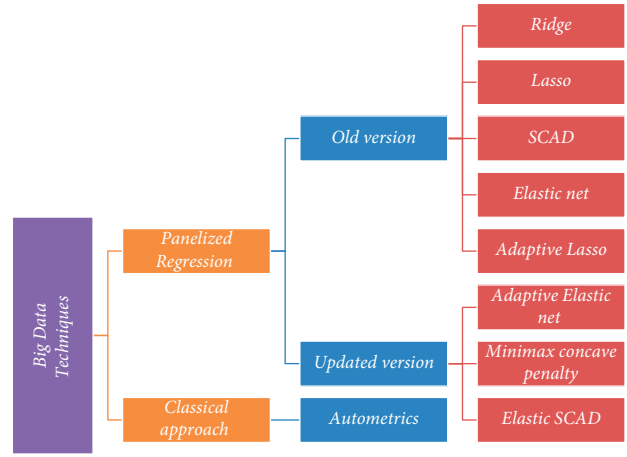

Figure 2: Classical and penalized regression methods.

the required oracle properties including smoothly clipped absolute deviation (SCAD) and adaptive lasso, but the drawback associated with these two methods is as follows: they only select one variable from a group of correlated covariates and ignore other variables. The selected variable may or may not be theoretically important. [14] modified SCAD by adding another property to its penalty, which spurs a set of highly correlated covariates to be in or out of the model at the same time. In other words, the new version of SCAD is able to select a group of correlated variables instead of a single one. Similarly, [2] modified the elastic net in the form of an adaptive elastic net, which achieved an oracle property. The method is capable of including and excluding features simultaneously. Minimax concave penalty (MCP) is another extended method, which is developed by [6] and is based on the concave penalty. The method also enjoys an oracle property. To summarize, Adaptive Elastic net, MCP and Elastic SCAD are the updated forms of penalization techniques, primarily used for variable selection and will be explored in the next sections.

Another approach for automatic model selection was proposed by [15, 16], known as PcGets. This method is based on the idea of general to specific (gets) modeling. It starts from a general unrestricted model which captures the key attributes of the underlying dataset. Their standard testing approaches are utilized to decrease its complexity by removing statistically insignificant variables, inspecting the validity of the reductions at every stage to ensure the congruence of the selected model. They studied PcGets

probabilities recovering the data generating process (DGP) through Monte Carlo experiments and got reliable results. The consistency of the PcGets procedure was established by [17].

The new version of the PcGets algorithm was proposed by [18] as Autometrics. This version is based on the same principles as PcGets. Autometrics utilizes a tree-path search to identify and knock out statistically insignificant covariates. If the relevant covariate is eliminated by chance, the algorithm works and does not get stuck even in a single route, containing other covariates as proxies (like in stepwise regression). The beauty of this algorithm is that it works well even if the number of covariates exceeds the number of observations [10].

Our study contributes theoretically as well as empirically to literature. There exists immense literature on using conventional approaches like vector autoregressions, vector error correction models, etc. Such approaches adjust not more than 10 covariates, as more covariates create serious issues, due to which the results are invalid. More precisely, increasing the number of predictors (Big Data) leads to a few major problems in the models, such as degrees of freedom, high variability, and multicollinearity. For fixing these problems and achieving valid results, this study adopts several updated classical and machine learning techniques. The techniques will be compared under simulated scenarios for multicollinearity, heteroscedasticity, and autocorrelation, and will there be applied to macroeconomic data to provide conclusive solutions to the predictability and validity of distinct theoretical scenarios simultaneously. Our study aims to provide an improved technique to help policymakers; the improved tool is not restricted to worker's remittances (in our case) but is valid for any macroeconomic data set under Huge Big Data ($P < N$).

The goal of this study is to compare the performance of the classical approach (Autometrics) with improved shrinkage methods including Adaptive Elastic net; Elastic Smoothly Clipped Absolute Deviation; Minimax Concave Penalty under different scenarios like multicollinearity, heteroscedasticity and autocorrelation in terms of variable selection. In this study, we focus solely on exploring these techniques for the case of Huge Big Data.

The rest of the article is arranged as Section 2 gives an overview of methods. Section 3 discusses the simulation exercise. Section 4 carries out the real data analysis. Section 5 comprises conclusion.

2. Methods

In statistics and econometrics, it is imperative to investigate the performance of statistical models theoretically and empirically. This work attempts to describe both aspects of the included methods. Our study considers a variety of modified forms of penalization techniques and classical approaches. The methods considered here are Adaptive Elastic net, Elastic Smooth Clipped Absolute Deviation, Minimax Concave Penalty, and Autometrics. Here, we provide a detailed description of each method.

2.1. Adaptive Elastic Net (AEnet). The lasso estimator has been designed to improve the performance of the ridge estimator. It is certainly useful, particularly when most coefficients of the true model are zero. Albeit, ridge regression performs better than lasso when a correlation between predictors is high [19].

To overcome the shortcomings of lasso and ridge regression, the elastic net method was proposed by [19] and used both lasso and ridge penalty simultaneously. The penalty function of the elastic net (EN) is given by the following:

$$\widehat{\alpha}^{\mathrm{EN}} = \left(1 + \frac{k_2}{n}\right) \mathrm{Arg\,min} \sum_{c=1}^{n} \left(y_c - \alpha_o - \sum_{d=1}^{m} \alpha_d x_{cd} \right) + k_2 \sum_{d=1}^{m} \alpha_d^2 + k_1 \sum_{d=1}^{m} |\alpha_d|. \tag{2}$$

Using a cross-validation approach, the tuning parameters k_1 and k_2 control the relative significance of L_1 norm and L_2 norm penalty. Both Lasso and Ridge regression are the special form of the elastic net, which have already been discussed in Section 1. In this sense, the elastic net contains dual features that are shrinkage and variable selection.

To estimate $\widehat{\alpha}^{\mathrm{EN}}$, [19] proposed an algorithm called least angle regression (LAR). This is the fact that EN does not satisfy an oracle property like Adaptive Lasso, albeit it performs better than Adaptive Lasso [11]. Later on, the idea of the Adaptive Lasso and the Elastic net regularization was combined to achieve further improvement known as Adaptive Elastic net (AEnet) and is defined as follows:

$$\widehat{\alpha}^{\mathrm{AEnet}} = \left(1 + \frac{k_2}{n}\right) \mathrm{Arg\,min} \sum_{c=1}^{n} \left(y_c - \alpha_o - \sum_{d=1}^{m} \alpha_d x_{cd} \right) + k_2 \sum_{d=1}^{m} \alpha_d^2 + k_1 \sum_{d=1}^{m} |\alpha_d|. \tag{3}$$

$\widehat{\omega}_d$ ($d = 1, 2, \ldots, m$) are adaptive data-driven weights. According to [2], initially, we estimate $\widehat{\alpha}^{\mathrm{EN}}$ by using an EN method as given in (2) and then utilize it while computing the weights as $\widehat{\omega}_d = |\widehat{\alpha}_d^{\mathrm{EN}}|^{-\tau}$; here τ is constant and should be positive. Thus, AEnet, the modified form of elastic net, attains an oracle property.

2.2. Elastic Smoothly Clipped Absolute Deviation (E-SCAD). Reference [12] developed a new regularization method known as SCAD. This method is nonconvex and fulfills the properties of a good penalty. This method not only selects the important features consistently and yields the estimates of unknown coefficients more efficiently given that the true model is known. Therefore, the SCAD function covers all the limitations related to the existing methods like Ridge and Lasso.

The penalty function of SCAD is defined as follows:

$$p_k(|\tau|) = k\left\{I(\tau \le k) + \frac{(\gamma k - \tau)}{(\gamma - 1)k} + I(\tau > k)\right\}. \quad (4)$$

They considered the value of γ equal to 3.7, and the unknown tuning parameter k was computed by generalized cross-validation. As foregoing, the penalty function is continuous, and the resulting solution is given by the following:

$$p_k(|\tau|) = \begin{cases} k|\tau|, & |\tau| < k, \\ -\dfrac{\left(\tau^2 - 2\gamma k|\tau| + k^2\right)}{2(\gamma - 1)}, & k < |\tau| \le \gamma k, \\ \dfrac{(\gamma + 1)k^2}{2}, & |\tau| > \gamma k. \end{cases} \quad (5)$$

The tuning parameters can be induced from the data-driven techniques. The idea of a combination of SCAD and L_2 penalty was proposed by [14] and called it Elastic SCAD. Mathematically, E-SCAD can be written as follows:

$$\text{pen}_k(|\tau|) = \sum_{d=1}^{D} p_k(|\tau|) + \lambda_{2p} \sum_{d=1}^{m} \alpha_d^2. \quad (6)$$

2.3. Minimax Concave Penalty. The idea of minimax concave penalty (MCP) was initially proposed by [20]. This method provides the convexity of the penalized loss in sparse regions to the greatest extent, given certain thresholds for variable selection and unbiasedness. The MCP is described as follows:

$$S_{\text{MCP}}(t; k) = \left\{ \begin{array}{ll} kt - \dfrac{t^2}{2\gamma} & \text{if } |t| \le \gamma k \\ \dfrac{1}{2}\gamma k^2 & \text{if } |t| > \gamma k \end{array} \right\}. \quad (7)$$

The tuning parameter ($\gamma > 0$) reduces the maximal concavity subject to the following constraints, i.e., unbiasedness and features selection:

$$\begin{aligned} \rho(t; k) &= 0, \quad \forall t \ge \gamma k, \\ \rho(0+; k) &= k, \\ &\cdot \sum_{d=1}^{m} p_d\left(|\alpha_d|; k; \gamma\right). \end{aligned} \quad (8)$$

The role of dual tuning parameters in concave penalty regression is to control the amount of regularization. Besides, the concavity of the MCP penalty substantially prevents the sparse convexity on account of reducing the maximal concavity. As the value of the regularization parameter rises, a result bears more convexity and attain near an unbiased penalty [20]. The penalty function is a part of the quadratic spline function and dual tuning parameters.

2.4. Classical Approach. Autometrics comprises five fundamental phases. The initial phase concerns the construction of a linear model known as General Unrestricted Model (GUM); the second step yields the estimates of unknown parameters and statistical testing of the GUM; the third step consists of the presearch process; the fourth step provides the tree-path search; the last step involves a selection of the final model.

The complete algorithm is precisely delineated in [18]. The key notion is to commence modeling with a linear model incorporating each essential feature. Estimate the GUM by the least square method and then execute the statistical tests to ensure the congruency of a model. If the estimated GUM contains statistically insignificant coefficients at prespecified criteria, then again estimate the simpler model by utilizing different path searches and ratified by statistical or diagnostic tests. As some terminal models are detected, Autometrics undertakes their union testing. Rejected models are eliminated, and the union of those terminal models who survived induces new GUM for another tree-path search iteration. This whole inspection process remains, and the terminal models are statistically examined against their union. If two or more terminal models assure the encompassing tests, then the prechosen information criterion is a gateway to a final decision.

The forecasting model is obtained by using Autometrics approach on the GUM:

$$y_t = \gamma_0 + \sum_{i=1}^{n} \sum_{k=0}^{K} \delta_{i,k} x_{i,t-k} + e_t. \quad (9)$$

Here, two strategies are widely used for variable selection, a conservative and a superconservative (Liberal) strategy. This study adopts the super conservative strategy based on a one percent level of significance instead of five percent.

3. Simulation Study

Our simulation experiment involves three main scenarios, namely simulations on a data generating process (DGP) with (i) multicollinearity, (ii) heteroscedasticity, and (iii) autocorrelation. In each case, we vary the DGP characteristics as the correlation structure among predictors, the level of variance of the error term, and the level of correlation between the current and lagged value of the error term.

3.1. Data Generating Process. We generate data from the following equation:

$$Y_t = X_t^T \gamma + \varepsilon_t, \quad (10)$$

where Y_t is an outcome variable. The features set, $X_t = x_1, x_2, \ldots, x_P$, is generated from multivariate normal distribution as $X_t \sim \text{MVN}(0, \sum)$ where the mean of covariates is zero and $\sum$ is the variance-covariance matrix. The same data generating process (DGP) was used by [1, 21] as mentioned in equation (9) for artificial data generation. Three sorts of sample sizes are to be used in the simulation exercise. Moreover, we

assume two sets of candidate variables with varying the number of relevant (p) and irrelevant variables (q) respectively, presented in Figure 3.

In the first scenario, we generate the pairwise correlation between the predictors, i.e., x_m and x_n as $\text{cov}(x_m, x_n) = \sum^{|m-n|}$. The population covariance matrix is generated as follows:

$$\sum_p = \begin{bmatrix} 1 & \cdots & \sum^{|n-m|} \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \sum^{|m-n|} & \cdots & 1 \end{bmatrix}. \tag{11}$$

With varying the parameter, $\sum$, we get the different pairwise correlation; here, we assume the values for $\sum$ as {0.25, 0.5, 0.9} followed by [22]. In the second scenario, we generate a correlation between the current and lagged residuals (autocorrelation), denoted by ρ. The autocorrelation is generated by the following equation:

$$\varepsilon_t = \rho\varepsilon_{t-1} + \mu_t. \tag{12}$$

We assign the following values to the coefficient of lagged residuals: $\rho \in \{0.25, 0.5, 0.9\}$. Third scenario: in the case of heteroscedasticity, the variance of the error term is not constant and varies across observations by σ_k.

$$E\left(\varepsilon_t^2\right) = \sigma_k. \tag{13}$$

Thus, we divide the variance σ_k into two parts, i.e., σ_1 and σ_2. For half of the observations (n/2), we set the variance by σ_1 and σ_2 for the remaining (n/2) data points. Our experiment assumes three cases of heteroscedasticity and set the values of $\pi_i = (\sigma_1/\sigma_2)$, where $i = 1, 2, 3$ as $\pi_i \in \{0.1/0.3, 0.2/0.6, 0.3/0.9\}$. This study attempts to evaluate the performance of Autometrics, AEnet, E-SCAD, and MCP using Huge Big Data under all preceding scenarios. Tenfold cross-validation is executed to determine the optimal value of the tuning parameter.

To evaluate the performance, the authors [1] have used potency and gauge to assess the best model in features selection relatively. Therefore, we follow the same criteria for model selection as well. The entire process is replicated 1,000 times. The comparison of regularization techniques and Autometrics is assessed in the form of incorrect zero identification, namely gauge and correct zero identification, namely potency [1]. For simulation as well as empirical analysis, we use R software.

3.2. Simulation Results. The Monte Carlo simulation results are described in Tables 1–3.

Table 1 depicts the findings of simulation in the case of low, moderate, and high multicollinearity for different combinations of observations (n) and covariates. The performance of all methods is improving with increasing sample size:

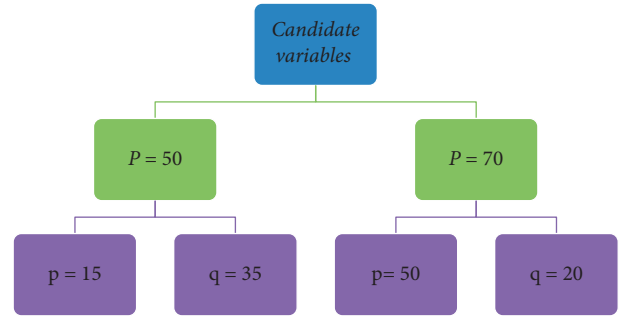

Figure 3: Distribution of candidate variables into relevant (p) and irrelevant (q).

(1) In the case of low and moderate multicollinearity, the potency associated with all methods is one under most simulated scenarios, clearly revealing that they retain all the relevant variables under low multicollinearity. Increasing the level of multicollinearity tends to improve the performance of AEnet and Autometrics in such a way that holds less irrelevant variables but adversely affects the MCP performance, particularly in small and moderate samples. Across low and moderate multicollinearity, the gauge associated with AEnet is lower than the gauge of other methods, which exhibits that it retains less irrelevant covariates. Comparatively, the E-SCAD retains more irrelevant variables and thus overspecify the true model.

(2) In the case of high collinearity: high collinearity among variables substantially distorts the performance of MCP and Autometrics in terms of potency and gauge specification. The AEnet retained more than 93 percent correct variables with an outstanding gauge (zero percent). However, the potency and gauge of other methods tend to increase with increasing sample size, particularly MCP and E-SCAD significantly overspecifying the true model (retain more irrelevant variables). AEnet showed an outstanding performance in terms of gauge. Under the large sample, improvement in the E-SCAD gauge was achieved in contrast to the case of low and moderate levels of multicollinearity.

Table 2 presents the simulation results by varying heteroscedasticity along with sample size and many covariates (both relevant and irrelevant).

(1) In the case of heteroscedasticity: the potency of all included methods is one in almost all scenarios, certainly manifesting that they hold all the active covariates. In contrast, the gauge of AEnet and Autometrics exhibit that it avoids the irrelevant variables and very precisely identifies the true model. Higher level of Autocorrelation adversely affects the potency of Autometrics in contrast to rival methods. The results suggest that MCP drops the inactive variables, particularly when the sample size is increased. E-SCAD has considerably overspecified the model. Increasing the number of covariates tends to

Table 1: Variable selection under multicollinearity from Monte Carlo Simulation.

Models	$\sum = 0.25,\ P = 50$		$\sum = 0.25,\ P = 70$	
$n = 80/160/320$	Potency	Gauge	Potency	Gauge
MCP	1/1/1	0.04/0.02/0.02	0.99/1/1	0.05/0.02/0.01
E-SCAD	1/1/1	0.12/0.10/0.10	1/1/1	0.11/0.10/0.09
AEnet	0.99/1/1	0.01/0/0	0.99/1/1	0.02/0/0
Autometrics	0.99/1/1	0.04/0.01/0.01	0.99/1/1	0.04/0.01/0.01
n = 80/160/320	$\sum = 0.50,\ P = 50$		$\sum = 0.50,\ P = 70$	
MCP	0.99/1/1	0.06/0.02/0.01	0.99/1/1	0.09/0.01/0.01
E-SCAD	1/1/1	0.10/0.07/0.06	0.99/1/1	0.09/0.06/0.06
AEnet	0.99/1/1	0/0/0	0.99/1/1	0/0/0
Autometrics	0.99/1/1	0.02/0.01/0.01	0.98/1/1	0.06/0.01/0.01
n = 80/160/320	$\sum = 0.90,\ P = 50$		$\sum = 0.90,\ P = 70$	
MCP	0.68/0.94/0.99	0.19/0.22/0.09	0.59/0.92/0.99	0.16/0.23/0.09
E-SCAD	0.91/0.98/0.99	0.13/0.09/0.03	0.89/0.98/0.99	0.12/0.09/0.03
AEnet	0.93/0.98/0.99	0/0/0	0.91/0.98/0.99	0/0/0
Autometrics	0.63/0.89/0.99	0.06/0.02/0.02	0.61/0.87/0.99	0.17/0.03/0.01

Table 2: Variable selection under heteroscedasticity from Monte Carlo Simulation.

Models	$\pi_1 = 0.1/0.3,\ P = 50$		$\pi_1 = 0.1/0.3,\ P = 70$	
n = 80/160/320	Potency	Gauge	Potency	Gauge
MCP	1/1/1	0.08/0.02/0.01	1/1/1	0.01/0.01/0.01
E-SCAD	1/1/1	0.10/0.11/0.11	1/1/1	0.09/0.10/0.10
AEnet	1/1/1	0/0/0	1/1/1	0/0/0
Autometrics	1/1/1	0.01/0.01/0.01	1/1/1	0.04/0.01/0.01
n = 80/160/320	$\pi_2 = 0.2/0.6,\ P = 50$		$\pi_2 = 0.2/0.6,\ P = 70$	
MCP	1/1/1	0.02/0.01/0.02	1/1/1	0.01/0.01/0.01
E-SCAD	1/1/1	0.10/0.10/0.12	1/1/1	0.09/0.10/0.10
AEnet	1/1/1	0/0/0	1/1/1	0/0/0
Autometrics	1/1/1	0.01/0.01/0.01	1/1/1	0.04/0.01/0.01
n = 80/160/320	$\pi_3 = 0.3/0.9,\ P = 50$		$\pi_3 = 0.3/0.9,\ P = 70$	
MCP	1/1/1	0.02/0.01/0.02	1/1/1	0.01/0.01/0.01
E-SCAD	1/1/1	0.10/0.10/0.10	1/1/1	0.09/0.10/0.10
AEnet	1/1/1	0/0/0	1/1/1	0/0/0
Autometrics	1/1/1	0.01/0.01/0.01	0.99/1/1	0.04/0.01/0.01

affect the gauge associated with Autometrics and AEnet.

Table 3 portrays the simulation's output by varying autocorrelation, sample size, and several covariates (both active and inactive). Low (0.25), moderate (0.5), and high autocorrelation (0.9) are considered here:

(1) In the case of low and moderate autocorrelation: under mostly simulated schemes, the methods have found all the right variables, but E-SCAD and MCP retain a huge set of irrelevant variables that overspecify the model. In contrast, the AEnet and Autometrics provide the best results under almost all combinations of n and p. In other words, AEnet and Autometrics avoid the irrelevant variables and correctly specify the true model very well. Increasing the length of covariates, the E-SCAD gauge is improved but negatively affects the gauge of Autometrics and AEnet.

(2) In the case of high autocorrelation: comparatively rival methods, E-SCAD has shown good performance in selecting relevant variables considering a small sample. However, the same method collapsed under gauge. Similarly, Autometrics and AEnet performed better in gauge and often held less than 5 percent inactive variables. Expanding the covariates' window adversely affects the AEnet and Autometrics performance in terms of gauge.

4. Real Data Implications

After Monte Carlo experiments, this study performs real data analysis using Huge Big Data. We consider worker's remittances inflow and all its possible determinants data for real data analysis. There are so many factors that affect the worker's remittances inflow. Some covariates are recommended by economic theory to be included in the model. Apart from this, a long list of variables has been recommended by past studies. This study considers all the possible determinants based on economic theories and literature to make a general model. In econometrics literature, such a model is known as the general unrestricted model (GUM).

4.1. Data Source. This study collects the yearly data for Pakistan from 1972 to 2020 using different sources such as world development indicators (WDI), international financial statistics (IFS), international country risk guide, and state bank of Pakistan. The few missing observations in the data set are replaced by averaging the neighbor observations. Most variables are transformed into logarithm form to ensure normality. Detail regarding the variables has been given in Table 4. Table 4 describes the variables, symbols, definition of each variable, and data source.

4.2. Correlation Matrix. In Figure 4, blue and red colors exhibit Positive and negative correlations between the variables. The colors severity and area of the circles indicate a high pairwise correlation. Besides the right side of the correlogram, the legend color shows the pairwise correlation. We can observe numerous severe color circles in blue and red, evidence of high pairwise correlation.

Table 3: Variable selection under Autocorrelation from Monte Carlo Simulation.

Models	$\rho = 0.25,\ P = 50$		$\rho = 0.25,\ P = 70$	
$n = 80/160/320$	Potency	Gauge	Potency	Gauge
MCP	1/1/1	0.04/0.02/0.02	1/1/1	0.04/0.02/0.02
E-SCAD	1/1/1	0.13/0.10/0.10	1/1/1	0.12/0.09/0.09
AEnet	0.99/1/1	0.01/0/0	0.99/1/1	0.02/0/0
Autometrics	0.99/1/1	0.01/0.01/0.01	0.99/1/1	0.05/0.01/0
n = 80/160/320	$\rho = 0.50,\ P = 50$		$\rho = 0.50,\ P = 70$	
MCP	0.99/1/1	0.06/0.02/0.02	0.99/1/1	0.08/0.02/0.01
E-SCAD	1/1/1	0.15/0.10/0.10	0.99/1/1	0.14/0.09/0.09
AEnet	0.99/1/1	0.02/0/0	0.99/1/1	0.03/0/0
Autometrics	0.99/1/1	0.01/0.01/0.01	0.99/1/1	0.05/0.01/0.01
n = 80/160/320	$\rho = 0.90,\ P = 50$		$\rho = 0.90,\ P = 70$	
MCP	0.91/0.99/1	0.16/0.12/0.05	0.82/0.99/1	0.14/0.11/0.05
E-SCAD	0.98/0.99/1	0.28/0.23/0.15	0.96/0.99/1	0.26/0.22/0.13
AEnet	0.94/0.99/0.99	0.04/0.01/0	0.92/0.99/0.99	0.06/0.01/0
Autometrics	0.82/0.98/0.99	0.03/0.01/0.01	0.76/0.97/0.99	0.10/0.01/0.01

Table 4: Variables description.

Sr. no.	Variables	Symbol	Definition/construction	Source of data
1	Workers' remittances	WR	The transfer of foreign money by migrated workers to Pakistan.	SBP
2	Interest rate	INT	Call money rate	SBP
3	Gold prices	GOLD	Gold prices is defining the price of gold in which the gold is traded on gold market.	SBP
4	Development expenditure	DEX	It is the type of expenditure that helps the economic and social development of the country—for example, the expenditure on education, health, etc.	SBP
5	Major agriculture crops	AGC	Major agriculture crops are wheat, rice, cotton, sugarcane, maize etc.	SBP
6	Inflation	INF	Inflation is the increase in the price of goods and services over time at a general level. Inflation rate is measured by CPI_t - $CPI_{t-1}/CPI_{t-1} * 100$	SBP
7	Foreign direct investment	FDI	FDI is the type of investment in which the people or organization of one country invested in company of property of other countries.	SBP
8	Trade openness	TO	Trade openness is defined as the ratio of trade to GDP	SBP
9	Exchange rate/Nominal exchange rate	EXR	Value of the rupees per unit of US dollar	IFS
10	Stock market performance	SP	Share prices	IFS
11	Investment return of pak	IRPak	$0.8INT_{Pk} + 0.2dLn\ (SP_{Pk})$ where INT_{Pk} is interest rate and SP_{Pk} is share prices of Pakistan.	IFS
12	Investment return of US	IRUS	$0.8INT_{US} + 0.2dLn\ (SP_{US})$ Where INT_{US} is interest rate and SP_{US} is share prices of US.	IFS
13	Real domestic product	GDP	It is defined as the total value of final goods and services which are produced inside the boundary of the country in a given period.	WDI
14	Unemployment	UEMP	Unemployment is defined as the people who want to work but do not have a job.	WDI
15	Foreign debts	DEBT	Foreign debt is money that one country borrowed from an outside country or organization. It is also known as external debt.	WDI
16	Real effective exchange rate	REER	It is defined as the nominal effective exchange rate which is divided by a price deflator.	WDI
17	Secondary school enrolment	SSEN	Secondary school enrolment is defined as the number of students who are enrolled in secondary school.	WDI
18	Financial liberalization	FINL	The data on financial liberalization is taken from Shabbir (2013). He used the following formula for the construction of financial liberalization.	Shabbir (2013)
19	Job skill index		The job skill index is constructed with the help of weighted index of the different skill categories.	Bureau of emigration and overseas employment
20	Wage rate	WAGE	The amount of wage that is paid to the worker per unit of time.	Bhatti(2018)

Table 4: Continued.

Sr. no.	Variables	Symbol	Definition/construction	Source of data
21	Democracy	DMOC	Democracy is the type of government in which people elect their representatives.	ICRG
22	Internal conflict	ICNF	Internal conflict is defined as the political violence inside the country and its actual influence on the governance.	ICRG
23	External conflict	XCNF	External conflict is defined as the problem such as diplomatic pressures, trade restrictions, etc., to the mandatory government from the foreign action to violent external pressure.	ICRG
24	Law and order situation	LAOR	Law and order situation is defined as the condition when people follow the rule and regulations. There is no violence or threats, and the police control all the crimes, etc.	ICRG
25	Corruption	CRRP	The illegal actions by powerful people such as bureaucrats, government, police, etc.	ICRG
26	Terrorism index (no' of attacks)	TIND	It is the use of violence and threats for the purpose of achieving political and ideological objectives.	ICRG
27	Government stability	GS	Whenever the representative of the govt. change without any threats of violence, it is known as political stability.	ICRG
28	Black market premium	BMP	Black market premium is defined as the percentage difference between the black market exchange rate and official exchange rate.	ICRG

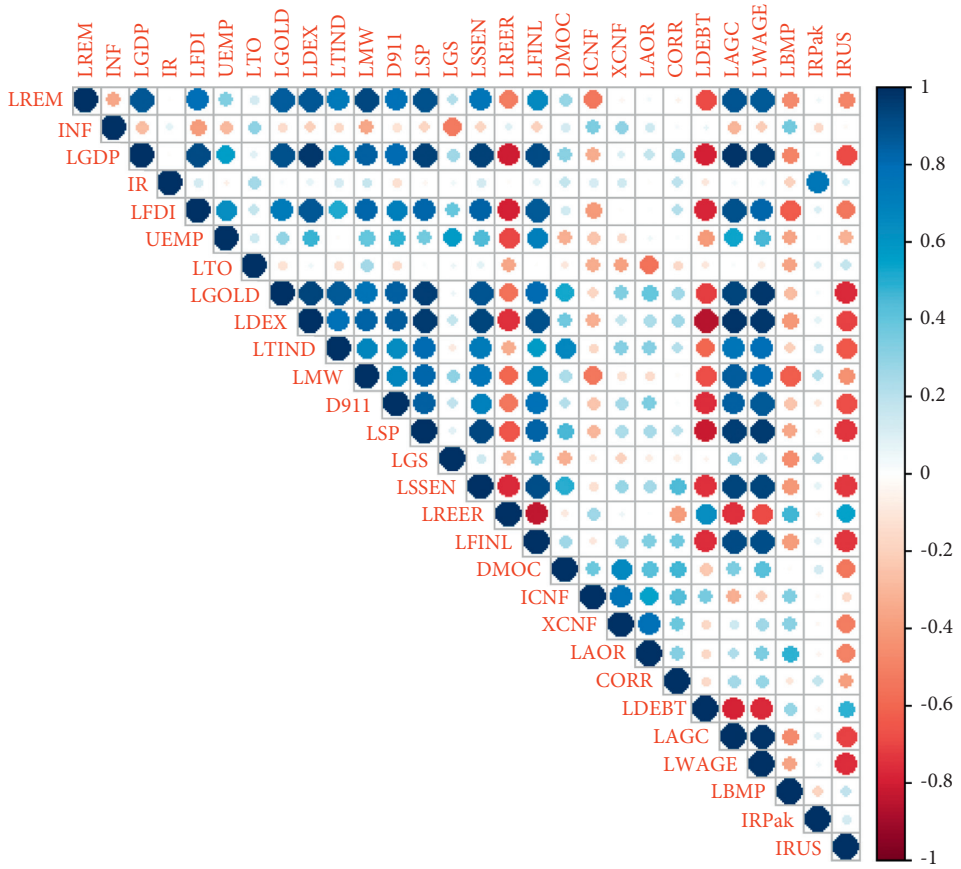

Figure 4: Correlation structure among covariates.

Figure 4 shows that there exists high multicollinearity among the predictors using the data period spanning from 1972 to 2020. We noted that in Monte Carlo simulations in the case of high multicollinearity, the AEnet outperformed the rival counterparts in terms of potency and gauge, mainly when the sample size is small. It reveals that AEnet is more robust in such circumstances, and thus we should proceed with AEnet output.

We performed diagnostic tests and found that the residuals of an estimated model are homoscedastic and uncorrelated.

Table 5: Features selection based on real data.

Variables	MCP	E-SCAD	AEnet	Autometrics
✓LGDP	✓	✓	✓	✓
✗INF	✗	✗	✗	✗
✗IR	✗	✗	✗	✗
✗LFDI	✗	✗	✗	✗
✗UEMP	✗	✗	✗	✗
✓LTO	✓	✓	✓	✓
✓LGOLD	✓	✓	✓	✓
✗LDEX	✗	✗	✗	✗
✓D911	✓	✓	✓	✓
✗LTIND	✗	✗	✗	✗
✓LMW	✓	✓	✓	✓
✗LSP	✗	✗	✗	✗
✓LSSEN	✓	✓	✓	✓
✓LREER	✓	✓	✓	✓
✓LFINL	✓	✓	✓	✓
✓DMOC	✓	✓	✓	✓
✗ICNF	✗	✗	✗	✗
✗XCNF	✗	✗	✗	✗
✗LAOR	✗	✗	✗	✗
✗CORR	✗	✗	✗	✗
✓LDEPT	✓	✓	✓	✓
✗LGS	✗	✗	✗	✗
✗IRUS	✗	✗	✗	✗
✓IRPAK	✓	✓	✓	✓
✗LAGC	✗	✗	✗	✗
✗LWAGE	✗	✗	✗	✗
✓LBMP	✓	✓	✓	✓

Tick marks show the selected variable, and cross marks show the nonselected variable.

Table 5 depicts the features selection based on real data using classical and shrinkage methods. In Table 5, the AEnet suggests almost 13 important determinants of workers' remittance among 27 determinants. In contrast, MCP and E-SCAD recommend many unrelated determinants of workers' remittance. In other words, we can conclude that they have overspecified the model. Apart from this, Autometrics keep the least number of irrelevant variables. The selection of an irrelevant set of covariates leads to poor forecasting.

In contrast, the right set of covariates can improve forecasting, leading to low forecast error. Consequently, an accurate forecast can help the government and other sectors in decision-making. To summarize the results, the empirical application strongly supports the findings of the simulation exercise.

5. Conclusion and Recommendations

This study compares Autometrics and three machine learning techniques, namely, Minimax Concave Penalty (MCP), Elastic Smoothly Clipped Absolute Deviation (E-SCAD), and Adaptive Elastic net (AEnet), under different scenarios: multicollinearity, heteroscedasticity, and autocorrelation with varying sample size and several covariates. We conducted Monte Carlo experiments to compare all methods in terms of variable selection using potency and gauge. All methods are improving their performance with expanding sample size. Considering the cases of low and moderate multicollinearity as well as low and moderate autocorrelation, the techniques retain all relevant predictor variables. However, for low and moderate multicollinearity, except AEnet, all methods keep many irrelevant predictors as well, whereas under low and moderate autocorrelation, including AEnet, the Autometrics also retain less irrelevant predictor variables. In presence of extreme multicollinearity, AEnet retains more than 93 percent of correct variables. Albeit, the potency of remaining techniques, specifically MCP and E-SCAD tends towards unity with increasing sample size but capturing massive irrelevant predictors as well. Considering the higher level of autocorrelation, E-SCAD has shown good performance in the selection of relevant variables under small sample. However, the same method collapsed under gauge. Similarly, Autometrics and AEnet performed better in gauge and often held less than 5 percent irrelevant variables. In the presence of heteroscedasticity, all techniques often hold all relevant variables but also suffer from overspecification problems except AEnet and Autometrics, which avoid the irrelevant predictors and identify the true model precisely.

On the application side, we take the workers' remittance data along its twenty-seven determinants spanning from 1972 to 2020. AEnet keeps thirteen predictors of workers' remittance. MCP and E-SCAD have selected many irrelevant determinants and consequently overspecified the model. This study has several recommendations:

(i) When there is a low/moderate multicollinearity case, and the sample size is small, practitioners and policymakers can use E-SCAD provided if there are less number of irrelevant covariates. Except for this case, AEnet is recommended in the presence of multicollinearity, particularly if the covariates are highly correlated with each other.

(ii) The study recommends AEnet when the residuals are heteroscedastic.

(iii) In the presence of autocorrelation, if there are more active variables and fewer inactive variables, then researchers should adopt E-SCAD if the scenario is converse, then use AEnet or Autometrics.

(iv) In the case of Pakistan, the AEnet showed remarkable performance in relevant variables. Hence, the policymakers and practitioners should focus on the relevant variables selected by AEnet to improve workers' remittance in the case of Pakistan. In this regard, the Pakistan government has devised policies that make it easy to transfer remittances legally and mitigate the cost of transferring remittances from abroad. The AEnet approach can help policymakers arrive at relevant variables in the presence of a huge set of covariates, which in turn produce accurate predictions.

Appendix

Table 4 describes the variables, symbols, definition of each variable, and source of data.

Data Availability

Data can be shared upon request to the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Conceptualization was proposed by Faridoon Khan, Amena Urooj, and Saud Ahmed Khan. Methodology and formal analysis were performed by Faridoon Khan. The original draft was written by Faridoon Khan. Supervision and cosupervision were done by Amena Urooj and Saud Ahmed Khan. Software was provided by Faridoon Khan, Sara Muhammadullah, and Zahra Almaspoor.Investigations were conducted by Amena Urooj and Saud Ahmed Khan. Review and editing were done by Faridoon Khan, Zahra Almaspoor, and Saima K. Khosa.

References

[1] J. A. Doornik and D. F. Hendry, "Statistical model selection with big data," *CogentEconomics and Finance*, vol. 3, no. 1, 2015.

[2] H. Zou and H. H. Zhang, "On the adaptive elastic-net with a diverging number of parameters," *Annals of Statistics*, vol. 37, no. 4, pp. 1733–1751, 2009.

[3] L. Breiman, "Better subset regression using the nonnegative garrote," *Technometrics*, vol. 37, no. 4, pp. 373–384, 1995.

[4] I. Savin, "A comparative study of the lasso-type and heuristic model selection methods," *Jahrbucher für Nationalokonomie und Statistik*, vol. 233, no. 4, pp. 526–549, 2013.

[5] E. E. Leamer and E. E. Leamer, *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, Vol. 53, John Wiley & Sons Incorporated, , Hoboken, NJ, USA, 1978.

[6] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, Springer, Berlin, Germany, 2013.

[7] S. Ali, H. Khan, I. Shah, M. M. Butt, and M. Suhail, "A comparison of some new and old robust ridge regression estimators," *Communications in Statistics - Simulation and Computation*, vol. 50, no. 8, pp. 1–19, 2019.

[8] J. L. Castle, J. A. Doornik, and D. F. Hendry, "Modelling non-stationary 'big data," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1556–1575, 2021.

[9] H. R. Varian, "Big data: new tricks for econometrics," *The Journal of Economic Perspectives*, vol. 28, no. 2, pp. 3–28, 2014.

[10] N. R. Swanson and W. Xiong, "Big data analytics in economics: what have we learned so far, and where should we go from here," *Canadian Journal of Economics/Revue canadienne d'économique*, vol. 51, no. 3, pp. 695–746, 2018a.

[11] Z. Y. Algamal and M. H. Lee, "Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification," *Computers in Biology and Medicine*, vol. 67, pp. 136–145, 2015.

[12] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[13] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[14] L. Zeng and J. Xie, "Group variable selection via SCAD-L2," *Statistics*, vol. 48, no. 1, pp. 49–66, 2014.

[15] H. M. Krolzig and D. F. Hendry, "Computer automation of general-to-specific model selection procedures," *Journal of Economic Dynamics and Control*, vol. 25, no. 6-7, pp. 831–866, 2001.

[16] K. D. Hoover and S. J. Perez, "Data mining reconsidered: encompassing and the general-to-specific approach to specification search," *The Econometrics Journal*, vol. 2, no. 2, pp. 167–191, 1999.

[17] J. Campos, D. F. Hendry, and H.-M. Krolzig, "Consistent model selection by an automatic gets approach," *Oxford Bulletin of Economics & Statistics*, vol. 65, no. s1, pp. 803–819, 2003.

[18] J. A. Doornik, *Econometric Model Selection With More Variables Than Observations*, Economics Department, University of Oxford, England, UK, 2009.

[19] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B*, vol. 67, no. 2, pp. 301–320, 2005.

[20] C. H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[21] F. Khan, A. Urooj, K. Ullah, B. Alnssyan, and Z. Almaspoor, "A comparison of Autometrics and penalization techniques under various error distributions: evidence from Monte CarloSimulation," *Complexity*, vol. 2021, Article ID 9223763, 8 pages, 2021.

[22] N. Xiao and Q.-S. Xu, "Multi-step adaptive elastic-net: reducing false positives in high-dimensional variable selection," *Journal of Statistical Computation and Simulation*, vol. 85, no. 18, pp. 3755–3765, 2015.