Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2015, Article ID 432986, 8 pages
http://dx.doi.org/10.1155/2015/432986

Research Article

Robust Stability Best Subset Selection for Autocorrelated Data Based on Robust Location and Dispersion Estimator

Hassan S. Uraibi,[1,2] Habshah Midi,[3] and Sohel Rana[3]

[1]*Laboratory of Computational Statistics and Operations Research, INSPEM, University Putra Malaysia, 43400 Serdang, Malaysia*
[2]*Department of Statistics, College of Administration and Economics, University of Al-Qadisiyah, Diwaniyah, Iraq*
[3]*Faculty of Science and Institute for Mathematical Research, University Putra Malaysia, 43400 Serdang, Malaysia*

Correspondence should be addressed to Hassan S. Uraibi; hssn.samil@gmail.com

Received 23 September 2015; Revised 7 December 2015; Accepted 8 December 2015

Academic Editor: Ramón M. Rodríguez-Dagnino

Stability selection (multisplit) approach is a variable selection procedure which relies on multisplit data to overcome the shortcomings that may occur to single-split data. Unfortunately, this procedure yields very poor results in the presence of outliers and other contamination in the original data. The problem becomes more complicated when the regression residuals are serially correlated. This paper presents a new robust stability selection procedure to remedy the combined problem of autocorrelation and outliers. We demonstrate the good performance of our proposed robust selection method using real air quality data and simulation study.

1. Introduction

The approach of splitting data into two parts is not new in the statistical inference and data analysis. Wasserman and Roeder [1] suggested combining the single-split approach with variable selection procedure. The variable selection algorithm is carried out in the first part (random half of data), followed by testing the significance of each selected variable based on p value of regression coefficient in the second part of data (the remaining half of data). However, this procedure does not guarantee reproducible results due to choosing arbitrary split [2].

A stability selection or multisplit approach is put forward to enhance and improve the performance of single-split variable selection method. The modern approaches of stability selection which rely on subsampling technique are proposed by [2, 3] for high dimensional data. The data is repeatedly split randomly into two parts with equal size of $n/2$. Unlike bootstrap, the stability selection approach repeatedly selects (without replacement) two subsamples with equal size $[n/2]$ from the original data. There is a possibility that any part of the split data may contain more outliers than the other parts of the split data. As a consequence, the existing classical linear regression stability selection procedure is easily affected by outliers, hence resulting in unreliable variables that are selected to the final model. This problem can be rectified by incorporating robust estimator in the selection procedure. However, this approach may not be adequate since robust estimation is expected to perform well only up to a certain percentage of outliers (Imon and Ali [4], Norazan et al. [5]). Since the selection procedure of the stability selection method is fairly closed to bootstrap [6], the idea of robust bootstrap may be used in stability selection procedure.

Following the idea of [4], in this paper, we propose diagnostic method before subsampling. The proposed diagnostic method is based on the Reweighted Fast Consistent and High (RFCH) breakdown estimator which is developed by [7] (cited by Alkenani and Yu [8], Özdemir and Wilcox [9], and Zhang et al. [10]). The suspected outliers are identified and deleted and random subsampling is performed from the remaining (clean) set of observations.

The proposed variable selection procedure also takes into consideration the autocorrelation problem. This problem, if not remedied, may provide misleading conclusions about the statistical significance of the regression coefficients [11]. Hence, the existing variable selection procedure may select

the wrong model. Appropriate remedial measures must be taken after detecting the presence of autocorrelation problems. One often used the Cochrane-Orcutt or Prais-Winsten methods (Greene [12], Gujarati and Porter [11]) to rectify autocorrelation problem. Nonetheless, these procedures are based on the OLS estimates, which are not robust and are therefore easily affected by outliers. Ann and Midi [13] proposed the Robust Cochrane-Orcutt Prais-Winsten (RCOPW) iterative method, based on high breakdown point and high efficiency MM-estimator [14], to overcome the combined problem of outliers and autocorrelated errors.

Hence, the main objective of this paper is to develop reliable, robust stability all-subset selection procedure in the presence of outliers and autocorrelation problem. The proposed method is formulated by rectifying the autocorrelation problem at the outset and subsequently the Reweighted Fast Consistent High (RFCH) breakdown estimator is incorporated in the algorithm. Upon convergence, the concentrated (clean) dataset is identified and all possible subsets procedures, namely, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) methods, were applied to the concentrated dataset in the last steps of the RFCH method. This approach is called concentrating all-subset selection and can be considered as a trade-off between the quality of data and the interpretability of a model.

2. The Consistency of Robust Stability Selection

Olive and Hawkins [7] showed that the RFCH estimator is Fast Consistent and High breakdown. The RFCH estimator is constructed using concentration algorithm in which the convergence is achieved after ten steps. At convergence, outliers are identified and deleted from the dataset. The remaining data will be used in the robust stability selection method whereby the former can be considered a source of consistency having the following properties:

(1) The all-subset selection of single-split data is consistent based on [7, Theorem 1].

(2) The multisplit procedure in which single-split data is repeated B times is also consistent based on [2, Corollary 3.1].

3. Robust Stability All-Subset Selection Method

Let a multivariate location and scatter model be a joint distribution of the ith case of a $(p \times 1)$ random vector that is completely specified by a $p \times 1$ population location vector μ and a $(p \times p)$ symmetric positive definite population scatter matrix Σ. Assume that n cases are collected in an $n \times p$ matrix X, such that $X_1^T, X_2^T, \ldots, X_p^T$ are independent. Consider a linear regression model $Y = X\beta + \varepsilon$, where Y is an $(n \times 1)$ vector of response variables, β is an $(n \times 1)$ vector of regression parameters, X is an $(n \times p)$ matrix of independent variables, and ε is an $(n \times 1)$ vector of random errors, where $\varepsilon \sim N(0, \sigma^2 I_n)$. The algorithm of our proposed robust and fast consistent variable selection consists of three main stages that are summarized as follows.

Stage 1 (rectifying the autocorrelation problem). We follow a simple procedure of Robust Cochrane-Orcutt method which is proposed by Ann and Midi [13] to rectify the problem of autocorrelation in the presence of both types of outlying observations, vertical outliers, and leverage points. The procedure can be summarized as follows:

(1) Estimate the robust regression coefficients using the MM-estimator to get the residuals $\widehat{U}_t$.

(2) Regress $\widehat{U}_t$ with $\widehat{U}_{t-1}$ using the MM-estimator, to find the robust parameter $\widehat{\rho}$.

(3) Use $\widehat{\rho}$ in the equations below to remedy the autocorrelation problem, and obtain a new design matrix X^* and response variable Y^*:

$$\begin{aligned} Y^* &= Y_{2:n} - \widehat{\rho} Y_{1:(n-1)} \\ X^* &= X_{[2:n,j]} - \widehat{\rho} X_{(1:[n-1],j)}, \end{aligned} \tag{1}$$

where $j = 1, \ldots, p$.

Stage 2 (concentrating the data). The concentrating algorithm assumes that the normality assumption for a linear regression is violated due to outliers or other contamination. The RFCH algorithm is employed to clean the data. This procedure uses the Devlin, Gnanadesikan and Kettenring (DGK) [15], and Median Ball (MB) [16]. These algorithms are summarized as follows.

Suppose the matrix X is a combination of the response vector Y^* and the covariates matrix X^*.

(i) The DGK Algorithm

Step 1. Begin by computing the classical estimator $(\overline{X}, \mathrm{cov})$ of the original dataset to give the initial or starting point $(T_{0,\mathrm{Start}}, C_{0,\mathrm{Start}})$, and find the initial Mahalanobis distance:

$$D_{0,\mathrm{DGK}} = \sqrt{(X - T_{0,\mathrm{Start}})^t (C_{0,\mathrm{Start}})^{-1} (X - T_{0,\mathrm{Start}})}. \tag{2}$$

Step 2. Arrange the initial Mahalanobis distances in increasing order to compute their median. Those observations in the original dataset whose Mahalanobis distances are less than the median of all the Mahalanobis distances will be in the remaining set (half dataset) and will be denoted by $\widetilde{X}_{1,\mathrm{DGK}}$:

$$\begin{aligned} \mathrm{Med}_{0,\mathrm{DGK}} &= \mathrm{Median}\left(D_{0,\mathrm{DGK}}\right) \\ \widetilde{X}_{1,\mathrm{DGK}} &= \left\{X_{ij} : D_{0,\mathrm{DGK}} \le \mathrm{Med}_{0,\mathrm{DGK}}\right\}, \\ & \qquad i = 1, \ldots, p, \;\; j = 1, 2, \ldots, m. \end{aligned} \tag{3}$$

Step 3. Let $C_{0,\mathrm{DGK}}$ be equal to $C_{0,\mathrm{Start}}$, where $C_{0,\mathrm{Start}}$ is the variance-covariance matrix of the original data. Calculate the average and the variance-covariance estimators of $\widetilde{X}_{1,\mathrm{DGK}}$ to get the first attractor $(T_{1,\mathrm{DGK}}, C_{1,\mathrm{DGK}})$.

Step 4. If the diagonal elements of $C_{1,\mathrm{DGK}}$ are equal to $C_{0,\mathrm{Start}}$, then stop the algorithm. Otherwise, repeat Steps 1–3 until convergence, to get the final attractor $(T_{K,\mathrm{DGK}}, C_{K,\mathrm{DGK}})$ and $\widetilde{X}_{K,\mathrm{DGK}}$, where K is the convergence step.

(ii) The Median Ball (MB) Algorithm

Step 1. Suppose the initial variance-covariance matrix $C_{0,\mathrm{Start}} = \mathrm{diag}(p)$ of the identity matrix and suppose that Med is the median vector of the matrix X. Then, the Mahalanobis distance based on the median is defined as follows:

$$D_{0,\mathrm{MB}} = \sqrt{(X - \mathrm{Med})^t \left(C_{0,\mathrm{Start}}\right)^{-1} (X - \mathrm{Med})}. \quad (4)$$

Step 2. The location criterion cut-off point is the median of $D_{0,\mathrm{MB}}$ and is denoted by luct:

$$\mathrm{luct} = \mathrm{Med}_{0,\mathrm{MB}} = \mathrm{Median}\left(D_{0,\mathrm{MB}}\right), \quad (5)$$

where luct $\neq 0.5$. The cut-off point should be the quantile of $D_{0,\mathrm{MB}}$ whose probability equals 0.5. For the concentration of X, find the half dataset with only nonoutlying observations whose Mahalanobis distances are less than or equal to the median:

$$\widetilde{X}_{1,\mathrm{MB}} = \left\{X_{ij} : D_{0,\mathrm{MB}} \leq \mathrm{Med}_{0,\mathrm{MB}}\right\}, \quad i = 1, \ldots, p, \ j = 1, 2, \ldots, m. \quad (6)$$

Step 3. Compute the average and the variance-covariance matrix of $\widetilde{X}_{1,\mathrm{MB}}$.

Step 4. For more concentrations, compute the Mahalanobis distances again, and repeat Steps 1–3 until convergence at the final attractor $(T_{K,\mathrm{MB}}, C_{5,\mathrm{MB}})$ and $\widetilde{X}_{K,\mathrm{MB}}$, where K is the convergence step.

(iii) The Reweighted Fast and Consistent High (RFCH) Breakdown Algorithm. Olive and Hawkins [7] developed the MB estimator by adding the location criterion or cut-off point to select the attractor and proposed the so-called Fast Consistent and High (FCH) breakdown estimator. Olive and Hawkins [7] noted that the FCH estimator uses the attractors with the smallest determinant.

Step 1. Following the same approach as Olive and Hawkins [7], define the final attractors as follows:

$$T_{\mathrm{FCH}} = \begin{cases} T_{K,\mathrm{DGK}} & \text{if } \sqrt{|C_{K,\mathrm{DGK}}|} < \sqrt{|C_{K,\mathrm{MB}}|} \\ T_{K,\mathrm{MB}} & \text{Otherwise,} \end{cases}$$

$$C_{\mathrm{FCH}} = \begin{cases} \dfrac{\mathrm{MED}\left(D_i^2\left((T_{K,\mathrm{DGK}}, C_{K,\mathrm{DGK}})\right)\right)}{\chi^2_{(p,0.5)}} C_{K,\mathrm{DGK}}, & \text{if } \sqrt{|C_{K,\mathrm{DGK}}|} < \sqrt{|C_{K,\mathrm{DGK}}|} \\ \dfrac{\mathrm{MED}\left(D_i^2\left((T_{K,\mathrm{MB}}, C_{K,\mathrm{MB}})\right)\right)}{\chi^2_{(p,0.5)}} C_{K,\mathrm{MB}}, & \text{Otherwise,} \end{cases} \quad (7)$$

where $\chi^2_{(p,0.5)}$ is the 50th percentile of a Chi-square distribution with p degrees of freedom.

According to [7, Theorem 1], as long as the start (T_K, C_K) is a consistent estimator of either $(T_{K,\mathrm{DGK}}, s_0 C_{K,\mathrm{DGK}})$ or $(T_{K,\mathrm{MB}}, s_1 C_{K,\mathrm{MB}})$, the FCH attractor is a consistent estimator of $(T_{K,\mathrm{FCH}}, aC_{K,\mathrm{FCH}})$, where $s_0 = \mathrm{MED}(D_i^2(T_{K,\mathrm{DGK}}, C_{K,\mathrm{DGK}}))/\chi^2_{(p,0.5)}$ and $s_1 = \mathrm{MED}(D_i^2(T_{K,\mathrm{MB}}, C_{K,\mathrm{MB}}))/\chi^2_{(p,0.5)}$ are positive constants and $a = s_0$ or $a = s_1$ based on the criterion cut-off point.

Step 2. Obtain the Reweighted FCH attractors by isolating the observation with $D_i^2(T_{\mathrm{FCH}}, C_{\mathrm{FCH}}) \leq \chi^2_{(p,0.975)}$, and using the classical estimator to obtain $(T_{1,\mathrm{FCH}}, C_{1,\mathrm{FCH}})$ from

$$\widetilde{X}_{1,\mathrm{FCH}} = \left\{X_{ij} : D_i^2\left(T_{\mathrm{FCH}}, C_{\mathrm{MFCH}}\right) \leq \chi^2_{(p,0.975)}\right\}, \quad i = 1, \ldots, p, \ j = 1, 2, \ldots, m. \quad (8)$$

Compute the new cut-off point as $\mathrm{MED}(D_i^2(T_{1,\mathrm{RFCH}}, C_{1,\mathrm{FCH}}))/\chi^2_{(p,0.5)}$. The new variance-covariance matrix is

$$C_{2,\mathrm{FCH}} = \frac{\mathrm{MED}\left(D_i^2\left(T_{1,\mathrm{FCH}}, C_{1,\mathrm{FCH}}\right)\right)}{\chi^2_{(p,0.5)}} C_{1,\mathrm{FCH}}. \quad (9)$$

Step 3. Repeat Steps 1-2 with the new cut-off point until convergence, to get the final attractors $(T_{\mathrm{RFCH}}, C_{\mathrm{RFCH}})$ and $\widetilde{X}_{\mathrm{RFCH}}$.

Stage 3 (robust stability selection based on all-subset selection). The concentrated data $\widetilde{X}_{\mathrm{RFCH}}$ involves the concentrated response vector Y^*_{RFCH} and the concentrated design matrix X^*_{RFCH}. Assume that $\widetilde{X}^{n_1}_{\mathrm{RFCH}}$ is a single random subsample that is drawn from $\widetilde{X}_{\mathrm{RFCH}}$, and $\widetilde{X}^{n_2}_{\mathrm{RFCH}}$ is the remaining subsample, where $n_1 = n_2 = [n/2]$ such that n is the number of rows in the concentrated design matrix X^*_{MRFCH}.

All-subset regression method guarantees that all possible potential covariates will be included in the submodels. The classical BIC criteria have the ability to determine the best model. We propose that all-subset procedure be applied to the first part of data $\widetilde{X}^{n_1}_{\text{RFCH}}$. The best model is the one that has coefficients with p values less than $\alpha^* = 0.05/d$, where d is the number of all candidate covariates. Repeat this procedure B times until convergence to get B best subsets such that $\widehat{S}_{kj} = \{kj; \widehat{\beta}_j \neq 0\}$, where $j = 1, 2, \ldots, m$; m is number of parameters estimation in subset k, where $k = 1, 2, \ldots, B$.

Following Meinshausen and Bühlmann [2], the threshold is defined as

$$\pi_{\text{th}} = \sqrt{\text{EV} \times p \times (2 \times \lambda - 1)}, \tag{10}$$

where EV is the expected number of variables falsely selected, p is the number of covariates in the specific subset, and λ is the highest chosen selection probability with the most selected covariates in the hole path of solution. In this study, we used $\lambda = 0.95$. Let δ be the number of $\widehat{\beta}_j$'s repeated in $\widehat{S}_{kj}$; then, the selected variables are those that belong to $\widehat{S}^*_j$ such that $\widehat{S}^*(\widehat{\beta}_j) = \{j; \delta(\widehat{\beta}_j)/B > \pi^*_{\text{th}}\}$. We multiply π_{th} by p to create the threshold measured by percentage; that is, $\pi^*_{\text{th}} = \pi_{\text{th}} \times p$, where p is the number of covariates in certain subset.

4. Simulation Study

Here, we report a simulation study that was designed to assess the performance of our proposed robust variable selection technique under two different outlier scenarios. In this experiment, we consider multiple linear regression model with the following relation:

$$Y = 7X_1 + 6X_3 + 5X_4 + 7X_6 + 7X_9 + 0\left[X_D\right] + \varepsilon, \tag{11}$$

where $D = 2, 5, 7, 8, 10$.

A design matrix was generated from a multivariate normal distribution with covariance structure $\text{cov}(X_j; X_k) = \rho^{|j-k|}$, where $\rho = 0.5$, $j, k = 1, 2, \ldots, 10$, and $n = 500$.

The random errors ε were drawn from a standard normal distribution. To create the autocorrelation problem, we considered the following setting:

$$\begin{aligned} Y^* &= Y_{[2:n]} + \rho Y_{[1:(n-1)]}, \\ X^* &= X_{[2:n]} + \rho X_{[1:(n-1)]}, \end{aligned} \tag{12}$$

where $\rho = 0.9$.

As in [17], two outlier scenarios were added to the data. The first scenario contaminated the residuals by ε symmetric outliers with the slash distribution, where $\varepsilon = 0.10$, and the random errors were generated as $\varepsilon \sim (1 - \varepsilon)N(0, 1) + \varepsilon N(0, 1)/u(0, 1)$. The second outlier scenario was generated by replacing 10% of the original values with high leverage points and vertical outliers. The vertical outliers were generated as asymmetric outliers, where $\varepsilon = 0.10$, and the errors were generated as $\varepsilon \sim (1 - \varepsilon)N(0, 1) + \varepsilon N(20, 1)$. To create the leverage points, each covariate was contaminated with 10% outlying observations generated from $N(50, 1)$.

For each case, we generated 500 independent simulated datasets. The problem of autocorrelated errors first is rectified and then randomly split each of the dataset into training n^{tr} (70%) and test n^{ts} (30%) sets. The proposed robust stability selections (R. multisplit-AIC and R. multisplit-BIC), the existing stability selections (multisplit-AIC and multisplit-BIC), and the single-split all-subsets-AIC and the single-split all-subsets-BIC methods were then applied to the training datasets. This process was repeated 500 times. The average Root Mean Squares Errors (RMSE) of the test sets over 500 simulation runs and the percentage chances for each variable of the training sets being selected in the final model over 500 simulation runs are presented in Tables 1–3. The potential variables being selected are also exhibited in the tables. The best method is the one that has the lowest RMSE and selects the correct variables (variables X_1, X_3, X_4, X_6, X_9) in the final model with no noise variable. The results in Table 1 show that when there is no outlier in the data, all the six methods are reasonably closed to each other. The results indicate that our proposed method is comparable with other existing methods.

Nevertheless, the results change dramatically in the presence of both outliers scenarios. It can be observed from Table 2 that the classical multisplit-AIC and multisplit-BIC methods are much affected in the presence of high leverage and vertical outliers. Both methods have the highest RMSEs and tend to be underfitting. In this situation, both the single-split-AIC and single-split-BIC variable selection techniques also fail to select the correct variables. Both methods tend to be overfitting because they also select noise variables in the final model. The presence of symmetric outliers as can be seen from Table 3 changes things amazingly. The RMSEs of the single-split-AIC and single-split-BIC are relatively larger than those of the other methods and both tend to be underfitting. Surprisingly, the multisplit-AIC and multisplit-BIC methods select the correct variables in this situation. Nonetheless, their RMSEs are slightly larger than the R. multisplit-AIC and the R. multisplit-BIC. On the other hand, the RMSEs of the R. multisplit-AIC and the R. multisplit-BIC are consistently the smallest among the six methods. Both methods select the correct variables with no noise variables, when no contamination occurs in the model and also in both outliers scenarios. Hence, it can be concluded that our proposed R. multisplit-AIC and the R. multisplit-BIC methods are the best variable selection methods in the linear regression model with autocorrelated errors because they are stable and consistently select the correct variables without choosing any noise variable.

5. Air Quality Data

In this study, hourly air pollution data which are taken from the Department of Environment (DoE), Malaysia, is used to further assess the performance of our method. The data consists of the PM10 concentration and ten independent variables, of which six are pollutant variables (sulphur dioxide (SO_2), nitrogen dioxide (NO_2), nitrogen monoxide (NO), nitrogen oxide (NO_x), carbon monoxide (CO), and ozone (O_3)) and four are meteorological variables (wind speed (WS), wind direction (WD), temperature (Temp), and

Table 1: Selected variables, average RMSE, and percentage for each variable being selected for clean data (threshold = 71.41).

	Single-split-AIC	Single-split-BIC	Multisplit-AIC	Multisplit-BIC	R. multisplit-AIC	R. multisplit-BIC
RMSE	0.67	0.67	0.65	0.65	0.65	0.65
1	100	100	100	100	99.9	99.9
2	17.9	1.79	17.5	2.60	0.84	0.84
3	100	99.9	100	100	99.7	99.7
4	100	99.9	100	100	99.6	99.6
5	13	1.65	19.9	3.90	0.84	0.84
6	100	99.9	100	100	99.7	99.7
7	14	1.51	17.00	1.70	0.88	0.88
8	12	1.47	20.3	3.40	0.75	0.75
9	99.9	99.9	100	100	99.8	99.8
10	16.6	1.71	16.7	3.2	0.81	0.81
Selected variables	1, 3, 4, 6, 9	1, 3, 4, 6, 9	1, 3, 4, 6, 9	1, 3, 4, 6, 9	1, 3, 4, 6, 9	1, 3, 4, 6, 9

Table 2: Selected variables, average RMSE, and percentage for each variable being selected for high leverage and vertical outliers (threshold = 71.41).

	Single-split-AIC	Single-split-BIC	Multisplit-AIC	Multisplit-BIC	R. multisplit-AIC	R. multisplit-BIC
RMSE	0.039	0.039	21.93	22.29	0.036	0.036
1	100	100	43.6	16.5	100	100
2	99.89	97.72	28.7	5.4	2.16	2.16
3	100	100	66.7	45.4	100	100
4	100	100	49.6	25.2	100	100
5	17.22	1.72	97.5	78.1	1.04	1.04
6	100	100	100	99.9	100	100
7	15.8	2.08	16.8	2.8	0.49	0.49
8	19.38	2.79	16.3	3.4	1.31	1.31
9	100	100	97.1	92.3	99.9	99.9
10	16.65	2.66	16.2	2.1 0	1.27	1.27
Selected variables	1, 2, 3, 4, 6, 9	1, 2, 3, 4, 6, 9	5, 6, 9	5, 6, 9	1, 3, 4, 6, 9	1, 3, 4, 6, 9

Figure 1: Location site of Seberang Perai, Penang [18].

relative humidity (Hum)). PM10 is a particulate matter 10 micrometers or less in diameter of solid or semisolid material found in the air. The value of each variable was recorded from the monitoring station at Seberang Perai, Penang (Figure 1), on an hourly basis every day from January 2005 to December 2013.

For the purpose of the statistical analysis, the hourly data were converted to a daily average, giving 3,287 readings. Missing values and calibration hours of certain variables are replaced by the coordinate medians for these variables. Let us first observe the plots in Figure 2. Both the histogram (b) and the quantile-quantile (Q-Q) plot (c) of Figure 2 show that the residuals are contaminated with a heavy-tailed mixture distribution. Since some points in the Q-Q plot do not fall on the straight line and the histogram is skewed to the right, this indicates that this data is not normal. Thus, we suspect that there are outliers in this dataset. Figure 2(d) also shows that there are some leverage points in each covariate.

Figure 2(a) indicates the existence of autocorrelation or serial correlation between the residuals, and it seems that there is high order autoregression AR(p).

Our proposed robust stability all-subset selection procedures and the existing methods were then applied to the data (3287 observations) to investigate which important variables influenced PM10. The dataset consists of 3287 observations, which include the PM10 as the response variable and the ten independent variables already mentioned. Since the air quality data are taken in time sequence, the Durbin Watson (DW) test is applied to the data to check the existence of

Table 3: Selected variables, average RMSE, and percentage for each variable being selected for symmetric outlier (threshold = 71.41).

	Single-split-AIC	Single-split-BIC	Multisplit-AIC	Multisplit-BIC	R. multisplit-AIC	R. multisplit-BIC
RMSE	0.663	0.663	0.23	0.23	0.212	0.212
1	91.18	87.54	92.1	96.7	100	100
2	16.97	3.69	15.8	6.0	1.136	1.136
3	76.36	63.75	91.0	81.9	100	100
4	88.45	84.26	89.8	75.1	100	100
5	18.07	4.21	18.3	6.4	1.22	1.22
6	85.88	78.71	93.6	96.8	100	100
7	17.88	3.28	15.4	3.5	1.21	1.21
8	158	3.02	17.8	3.3	0.78	0.78
9	68.45	51.92	91.5	96.5	99.5	99.5
10	18.33	3.696	17.9	4.6	0.95	0.95
Selected variables	1, 3, 4, 6	1, 4, 6	1, 3, 4, 6, 9	1, 3, 4, 6, 9	1, 3, 4, 6, 9	1, 3, 4, 6, 9

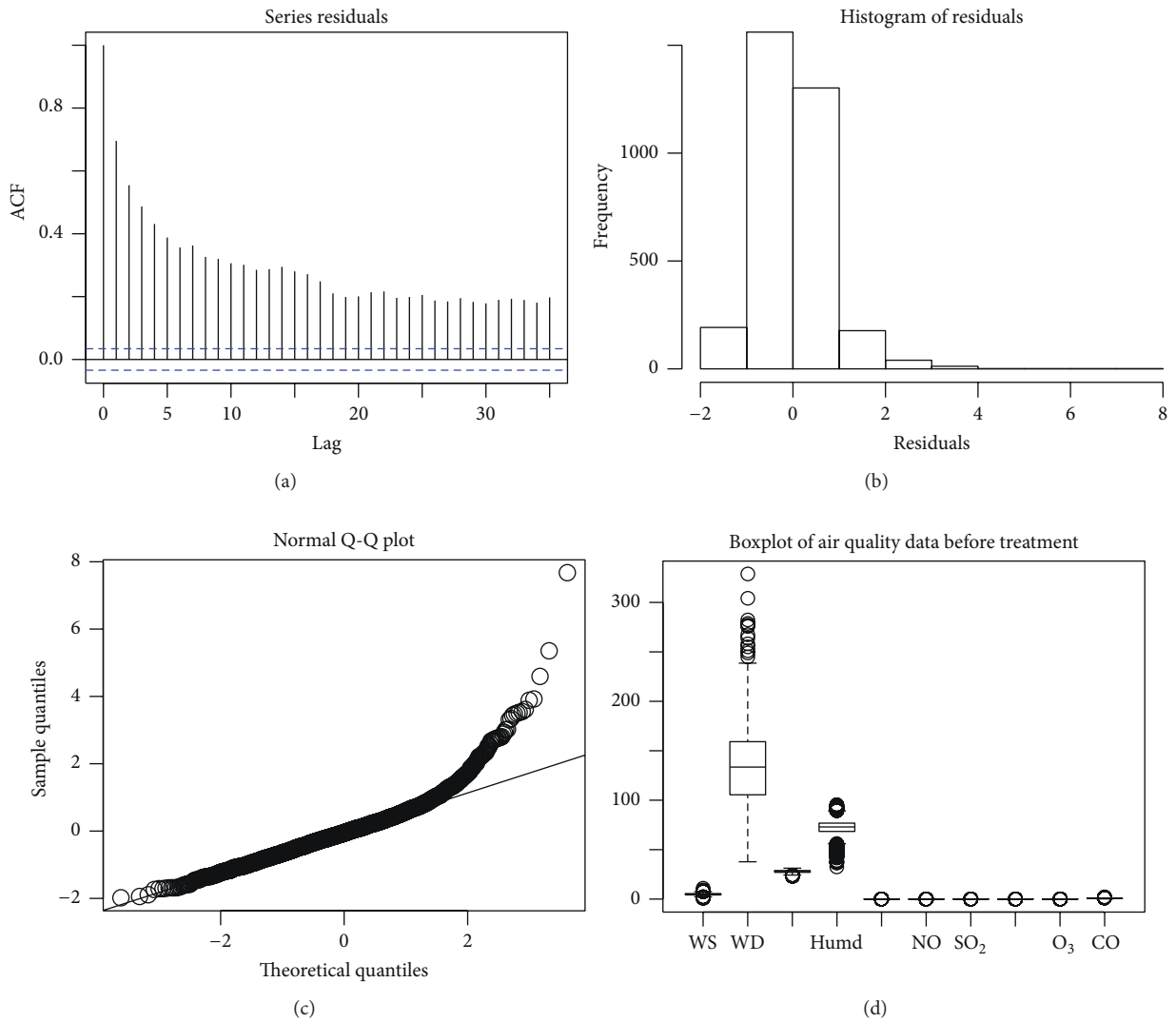

Figure 2: Q-Q plot, histogram of residuals, and plot of PM10 versus each component of air quality data, Seberang Perai, Pinang.

Table 4: Selected variables, average RMSE, and percentage chance for each variable being selected, for air quality data (threshold = 67.08).

	Single-split-AIC	Single-split-BIC	Multisplit-AIC	Multisplit-BIC	R. multisplit-AIC	R. multisplit-BIC
RMSE	0.51	0.51	0.51	0.51	0.4	0.4
WS	8.53	0.4	24.24	0.64	23.07	6.6
WD	100	76.86	73.34	14.24	**100**	**99.77**
Temp	100	100	100	100	**100**	**100**
Hum	100	100	100	100	**100**	**100**
NO_x	91	45.93	79.30	61.22	54.43	26.3
NO	96.5	47.8	87.66	67.54	49.6	24.63
SO_2	89.23	13.63	7.12	0.06	**99.93**	**91.33**
NO_2	10.46	54.36	32.42	47.06	**71**	**84.77**
O_3	100	100	100	100	**100**	**100**
CO	100	100	100	100	**100**	**100**
Selected variables	2, 3, 4, 5, 6, 7, 9, 10	2, 3, 4, 9, 10	2, 3, 4, 5, 6, 9, 10	3, 4, 5, 8, 9	2, 3, 4, 7, 8, 9, 10	2, 3, 4, 7, 8, 9, 10

autocorrelation problem. The results of Durbin Watson statistics for the original air quality data ($p \ll 0.01$) confirmed the existence of autocorrelation and no autocorrelation ($p > 0.05$) after treating the autocorrelation problem.

After correcting the autocorrelation problem, the data is then randomly divided into training (70%) and test (30%) sets.

This process is repeated 3,000 times. The RFCH is used to concentrate the training and test set data. Following Meinshausen and Bühlmann [2], each training set and each test set are split randomly into two sets of equal size and this process is repeated 50 times. The six variable selection methods were then applied to the first part of the training dataset. The variables that are selected in the final model are determined. For cross validation, the coefficients of each training model are used to predict the response (PM10) using test set data. The model residuals and the RMSEs are then computed. Table 4 exhibits the selected variables and the percentage of each variable being selected for training set data and the average RMSE for test set data over 3,000 runs. The threshold value in Table 4 is calculated as follows:

$$\pi_{\text{th}} = \sqrt{5 \times 10 \times (2 \times (0.95) - 1)} \times 10 = 67.08. \quad (13)$$

A candidate variable is the one whose percentage of being selected in a model exceeds the threshold value. The best method is the one that has the lowest average of RMSE.

The results in Table 4 show that the RMSE of our proposed method, based on both AIC and BIC, is the smallest compared to the existing methods. This suggests that our proposed method correctly identified the potential variables, namely, WD, Temp, Hum, SO_2, NO_2, O_3, and CO, to be included in the final model. The single-split-AIC method selects eight covariates, while the single-split-BIC method selects only six covariates. The classical multisplit-AIC selects seven covariates and multisplit-BIC selects five covariates.

It is interesting to observe that our proposed methods select all the pollutant variables except NO_x and NO and all the meteorological variables except WS. From the results in Table 4, we can clearly infer that the R. multisplit-AIC and R. multisplit-BIC methods are more efficient than the classical methods, because the final model that is selected by these methods is sufficient to include all the nonzero covariates and has the smallest RMSE. The results of the model validation suggest that WD, Temp, Hum, SO_2, NO_2, O_3, and CO should be included in the final model.

6. Conclusions and Recommendation

The main aim of this study was to develop a reliable alternative approach that is capable of selecting the correct variables in the final model for data having the combined problem of outliers and autocorrelated errors. We have considered the well known all-subsets-AIC and all-subsets-BIC, multisplit-AIC and multisplit-BIC variables selection methods in this regard. All the existing methods are not effective in choosing the correct variables in the final model. In this study, we proposed a robust stability selection method by incorporating a high efficient and high breakdown MM-estimator, the RFCH estimator, and applied the all-subset-BIC and the all-subset-AIC to the concentrated data. The real air quality data and simulation experiments show that our proposed methods successfully and consistently select the correct variables in the final model with the smallest RMSE. The commonly used methods failed to correctly select the correct variables in the final model. Hence, we can consider our proposed methods as better variable selection methods and strongly recommend using them especially when outliers and autocorrelated errors occur in the data.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to thank the Ministry of Science, Technology & Innovation, Malaysia, that has supported this work under eScienceFund Research Grant no. 06-01-04-SF1764. Special thanks also go to Department of Environment, Ministry of Natural Resources & Environment, Malaysia, that has provided the air pollution data to be used in this research.

References

[1] L. Wasserman and K. Roeder, "High-dimensional variable selection," *Annals of Statistics*, vol. 37, no. 5, pp. 2178–2201, 2009.

[2] N. Meinshausen and P. Bühlmann, "Stability selection," *Journal of the Royal Statistical Society B: Statistical Methodology*, vol. 72, no. 4, pp. 417–473, 2010.

[3] R. D. Shah and R. J. Samworth, "Variable selection with error control: another look at stability selection," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 75, no. 1, pp. 55–80, 2013.

[4] A. H. M. Imon and M. M. Ali, "Bootstrapping regression residuals," *Journal of the Korean Data and Information Science Society*, vol. 16, no. 3, pp. 665–682, 2005.

[5] M. R. Norazan, H. Midi, A. H. M. R. Imon, and S. Chen, "Weighted bootstrap with probability in regression," in *Proceedings of the 8th WSEAS International Conference on Applied Computer and Applied Computational Science*, Hangzhou, China, 2009.

[6] P. Bühlmann and B. Yu, "Analyzing bagging," *Annals of Statistics*, vol. 30, no. 4, pp. 927–961, 2002.

[7] D. J. Olive and D. M. Hawkins, "Robust multivariate location and dispersion," 2010, http://lagrange.math.siu.edu/Olive/pphbmld.pdf.

[8] A. Alkenani and K. Yu, "A comparative study for robust canonical correlation methods," *Journal of Statistical Computation and Simulation*, vol. 83, no. 4, pp. 690–720, 2013.

[9] A. F. Özdemir and R. Wilcox, "New results on the small-sample properties of some robust univariate estimators of location," *Communications in Statistics: Simulation and Computation*, vol. 41, no. 9, pp. 1544–1556, 2012.

[10] J. Zhang, D. J. Olive, and P. Ye, "Robust covariance matrix estimation with canonical correlation analysis," *International Journal of Statistics and Probability*, vol. 1, no. 2, p. 119, 2012.

[11] D. N. Gujarati and D. Porter, *Basic Econometrics*, McGraw-Hill, New York, NY, USA, 2009.

[12] W. H. Greene, *Econometric Analysis*, Pearson Education, New Delhi, India, 2003.

[13] L. H. Ann and H. Midi, "The effect of high leverage points on the robust autocorrelation test in multiple linear regression," in *Proceedings of the 11th WSEAS International Conference on Applied Computer Science (ACS '11)*, Penang, Malaysia, October 2011.

[14] V. J. Yohai, "High breakdown-point and high efficiency robust estimates for regression," *The Annals of Statistics*, vol. 15, no. 2, pp. 642–656, 1987.

[15] S. J. Devlin, R. Gnanadesikan, and J. R. Kettenring, "Robust estimation of dispersion matrices and principal components," *Journal of the American Statistical Association*, vol. 76, no. 374, pp. 354–362, 1981.

[16] D. J. Olive and D. M. Hawkins, "High breakdown multivariate estimators," http://lagrange.math.siu.edu/Olive/pphbrs.pdf.

[17] C. Agostinelli and M. Salibian-Barrera, "Robust model selection with lars based on S-estimators," in *Proceedings of COMPSTAT'2010*, pp. 69–78, Springer, Berlin, Germany, 2010.

[18] A. Z. Ul-Saufie, A. S. Yahaya, N. A. Ramli, and H. A. Hamid, "Robust regression models for predicting PM10 concentration in an industrial area," *International Journal of Engineering and Technology*, vol. 2, no. 3, pp. 364–370, 2012.