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

Hassan S. Uraibi, Habshah Midi, and Sohel Rana

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

Correspondence should be addressed to Hassan S. Uraibi; hssn.sami1@gmail.com

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

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 $\mu$ and a $(p \times p)$ symmetric positive definite population scatter matrix $\Sigma$. Assume that $n$ cases are collected in an $n \times p$ matrix $X$, such that $X_1^T, X_2^T, \ldots, X_n^T$ are independent. Consider a linear regression model $Y = X\beta + \epsilon$, where $Y$ is an $(n \times 1)$ vector of response variables, $\beta$ is an $(n \times 1)$ vector of regression parameters, $X$ is an $(n \times p)$ matrix of independent variables, and $\epsilon$ is an $(n \times 1)$ vector of random errors, where $\epsilon \sim N(0, \sigma^2I_n)$. 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 $\tilde{U}_t$.
2. Regress $\tilde{U}_t$ with $U_{t-1}$ using the MM-estimator, to find the robust parameter $\tilde{\rho}$.
3. Use $\tilde{\rho}$ in the equations below to remedy the autocorrelation problem, and obtain a new design matrix $X^*$ and response variable $Y^*$:

$$Y^* = Y_{2:n} - \tilde{\rho}Y_{1:n-1}$$

$$X^* = X_{[2:n,j]} - \tilde{\rho}X_{[1:n-1,j]}$$

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 $(\tilde{X}, \text{cov})$ of the original dataset to give the initial or starting point $(T_{0,\text{Start}}, C_{0,\text{Start}})$, and find the initial Mahalanobis distance:

$$D_{0,DGK} = \sqrt{(X - T_{0,\text{Start}})^T (C_{0,\text{Start}})^{-1} (X - T_{0,\text{Start}})}$$

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 $\tilde{X}_{1,DGK}$:

$$\tilde{X}_{1,DGK} = \{X_{ij} : D_{0,DGK} \leq \text{Med}_{0,DGK}\}.$$

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

---

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

(ii) The Median Ball (MB) Algorithm

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

\[
D_{0, MB} = \sqrt{(X - \text{Med})^T (C_{0, Start})^{-1} (X - \text{Med})}.
\] (4)

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

\[
\text{luct} = \text{Med}_{0, MB} = \text{Median}(D_{0, MB}),
\] (5)

where \( \text{luct} \neq 0.5 \). The cut-off point should be the quantile of \( D_{0, MB} \) 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:

\[
\hat{X}_{1, MB} = \left\{ X_{ij} : D_{0, MB} \leq \text{Med}_{0, MB} \right\},
\] (6)

\( i = 1, \ldots, p, \ j = 1, 2, \ldots, m. \)

Step 3. Compute the average and the variance-covariance matrix of \( \hat{X}_{1, MB} \).

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

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

\[
C_{FCH} = \begin{cases} \text{MED} \left( D^2_i \left( T_{K, DGK}, C_{K, DGK} \right) \right) / \chi_{(p, 0.5)}^2 & \text{if } \sqrt{|C_{K, DGK}|} < \sqrt{|C_{K, MB}|} \\ \text{MED} \left( D^2_i \left( T_{K, MB}, C_{K, MB} \right) \right) / \chi_{(p, 0.5)}^2 & \text{Otherwise,} \end{cases}
\] (7)

where \( \chi_{(p, 0.5)}^2 \) 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 \((T_{K, DGK}, s_2 C_{K, DGK})\) or \((T_{K, MB}, s_1 C_{K, MB})\), the FCH attractor is a consistent estimator of \((T_{K, FCH}, a C_{K, FCH})\), where \( s_2 = \text{MED}(D^2_i(T_{K, DGK}, C_{K, DGK})/\chi_{(p, 0.5)}^2) \) and \( s_1 = \text{MED}(D^2_i(T_{K, MB}, C_{K, MB})/\chi_{(p, 0.5)}^2) \) 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 \( D^2_i(T_{FCH}, C_{FCH}) \leq \chi_{(p, 0.975)}^2 \), and using the classical estimator to obtain \((T_{1, FCH}, C_{1, FCH})\) from

\[
\tilde{X}_{1, FCH} = \left\{ X_{ij} : D^2_i \left( T_{FCH}, C_{MFCH} \right) \leq \chi_{(p, 0.975)}^2 \right\},
\] (8)

\( i = 1, \ldots, p, \ j = 1, 2, \ldots, m. \)

Compute the new cut-off point as \( \text{MED}(D^2_i(T_{1, RFCH}, C_{1, FCH})/\chi_{(p, 0.5)}^2) \). The new variance-covariance matrix is

\[
C_{2, FCH} = \text{MED} \left( D^2_i \left( T_{1, FCH}, C_{1, FCH} \right) \right) / \chi_{(p, 0.5)}^2 \] (9)

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

Stage 3 (robust stability selection based on all-subset selection). The concentrated data \( \tilde{X}_{RFCH} \) involves the concentrated response vector \( Y_{RFCH}^* \) and the concentrated design matrix \( X_{RFCH}^* \). Assume that \( \tilde{X}_{RFCH}^* \) is a single random subsample that is drawn from \( \tilde{X}_{RFCH} \), and \( \tilde{X}_{RFCH}^* \) is the remaining subsample, where \( n_1 = n_2 = [n/2] \) such that \( n \) is the number of rows in the concentrated design matrix \( X_{RFCH}^* \).
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 $X_{BIC}^n$. 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 $\hat{S}_{kj} = \{kj; \hat{\beta}_j \neq 0\}$, 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

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

where EV is the expected number of variables falsely selected, $p$ is the number of covariates in the specific subset, and $\lambda$ 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 $\delta$ be the number of $\hat{\beta}_j$’s repeated in $\hat{S}_{kj}$; then, the selected variables are those that belong to $\hat{S}_j^*$ such that $\hat{S}_j^* = \{j; \delta(\hat{\beta}_j)/B > \pi_{th}^j\}$. We multiply $\pi_{th}^j$ by $p$ to create the threshold measured by percentage; that is, $\pi_{th}^j = \pi_{th} \times p$, 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:

$$Y = 7X_1 + 6X_3 + 5X_4 + 7X_6 + 7X_9 + 0 \left[ X_D \right] + \epsilon, \quad (11)$$

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 $\epsilon$ were drawn from a standard normal distribution. To create the autocorrelation problem, we considered the following setting:

$$Y^* = Y_{[2:n]} + \rho Y_{[1:(n-1)]}, \quad (12)$$

$$X^* = X_{[2:n]} + \rho X_{[1:(n-1)]},$$

where $\rho = 0.9$.

As in [17], two outlier scenarios were added to the data. The first scenario contaminated the residuals by $\epsilon$ symmetric outliers with the slash distribution, where $\epsilon = 0.10$, and the random errors were generated as $\epsilon \sim (1 - \epsilon)N(0, 1) + \epsilon N(0, 1)/\mu(0, 1)$. 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 $\epsilon = 0.10$, and the errors were generated as $\epsilon \sim (1 - \epsilon)N(0, 1) + \epsilon N(20, 1)$. 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 $n^t$ (70%) and test $n^t$ (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 oxide ($NO_x$), 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).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>0.67</td>
<td>0.67</td>
<td>0.65</td>
<td>0.65</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99.9</td>
<td>99.9</td>
</tr>
<tr>
<td>2</td>
<td>17.9</td>
<td>1.79</td>
<td>17.5</td>
<td>2.60</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>99.9</td>
<td>100</td>
<td>100</td>
<td>99.7</td>
<td>99.7</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>99.9</td>
<td>100</td>
<td>100</td>
<td>99.6</td>
<td>99.6</td>
</tr>
<tr>
<td>5</td>
<td>13</td>
<td>1.65</td>
<td>19.9</td>
<td>3.90</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>99.9</td>
<td>100</td>
<td>100</td>
<td>99.7</td>
<td>99.7</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>1.51</td>
<td>17.00</td>
<td>1.70</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>1.47</td>
<td>20.3</td>
<td>3.40</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>9</td>
<td>99.9</td>
<td>99.9</td>
<td>100</td>
<td>100</td>
<td>99.8</td>
<td>99.8</td>
</tr>
<tr>
<td>10</td>
<td>16.6</td>
<td>1.71</td>
<td>16.7</td>
<td>3.2</td>
<td>0.81</td>
<td>0.81</td>
</tr>
</tbody>
</table>

Selected variables: 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).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>0.039</td>
<td>0.039</td>
<td>21.93</td>
<td>22.29</td>
<td>0.036</td>
<td>0.036</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>100</td>
<td>43.6</td>
<td>16.5</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>99.89</td>
<td>97.72</td>
<td>28.7</td>
<td>5.4</td>
<td>2.16</td>
<td>2.16</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>100</td>
<td>66.7</td>
<td>45.4</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>100</td>
<td>49.6</td>
<td>25.2</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>1722</td>
<td>1.72</td>
<td>97.5</td>
<td>78.1</td>
<td>1.04</td>
<td>1.04</td>
</tr>
<tr>
<td>6</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99.9</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>7</td>
<td>15.8</td>
<td>2.08</td>
<td>16.8</td>
<td>2.8</td>
<td>0.49</td>
<td>0.49</td>
</tr>
<tr>
<td>8</td>
<td>19.38</td>
<td>2.79</td>
<td>16.3</td>
<td>3.4</td>
<td>1.31</td>
<td>1.31</td>
</tr>
<tr>
<td>9</td>
<td>100</td>
<td>100</td>
<td>97.1</td>
<td>92.3</td>
<td>99.9</td>
<td>99.9</td>
</tr>
<tr>
<td>10</td>
<td>16.65</td>
<td>2.66</td>
<td>16.2</td>
<td>2.10</td>
<td>1.27</td>
<td>1.27</td>
</tr>
</tbody>
</table>

Selected variables: 1, 2, 3, 4, 6, 9

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.

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

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>0.663</td>
<td>0.663</td>
<td>0.23</td>
<td>0.23</td>
<td>0.212</td>
<td>0.212</td>
</tr>
<tr>
<td>1</td>
<td>91.18</td>
<td>87.54</td>
<td>92.1</td>
<td>96.7</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>16.97</td>
<td>3.69</td>
<td>15.8</td>
<td>6.0</td>
<td>1.136</td>
<td>1.136</td>
</tr>
<tr>
<td>3</td>
<td>76.36</td>
<td>63.75</td>
<td>91.0</td>
<td>81.9</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>88.45</td>
<td>84.26</td>
<td>89.8</td>
<td>75.1</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>18.07</td>
<td>4.21</td>
<td>18.3</td>
<td>6.4</td>
<td>1.22</td>
<td>1.22</td>
</tr>
<tr>
<td>6</td>
<td>85.88</td>
<td>78.71</td>
<td>93.6</td>
<td>96.8</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>7</td>
<td>17.88</td>
<td>3.28</td>
<td>15.4</td>
<td>3.5</td>
<td>1.21</td>
<td>1.21</td>
</tr>
<tr>
<td>8</td>
<td>158</td>
<td>3.02</td>
<td>178</td>
<td>3.3</td>
<td>0.78</td>
<td>0.78</td>
</tr>
<tr>
<td>9</td>
<td>68.45</td>
<td>51.92</td>
<td>91.5</td>
<td>96.5</td>
<td>99.5</td>
<td>99.5</td>
</tr>
<tr>
<td>10</td>
<td>18.33</td>
<td>3.696</td>
<td>17.9</td>
<td>4.6</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>Selected variables</td>
<td>1, 3, 4, 6</td>
<td>1, 4, 6</td>
<td>1, 3, 4, 6, 9</td>
<td>1, 3, 4, 6, 9</td>
<td>1, 3, 4, 6, 9</td>
<td>1, 3, 4, 6, 9</td>
</tr>
</tbody>
</table>

Series residuals

Histogram of residuals

Normal Q-Q plot

Boxplot of air quality data before treatment

Figure 2: Q-Q plot, histogram of residuals, and plot of PM10 versus each component of air quality data, Seberang Perai, Pinang.
autocorrelation problem. The results of Durbin Watson statistics for the original air quality data \( p < 0.01 \) 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 dataset. The variables that are selected 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.

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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>0.51</td>
<td>0.51</td>
<td>0.51</td>
<td>0.51</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>WS</td>
<td>8.53</td>
<td>0.4</td>
<td>24.24</td>
<td>0.64</td>
<td>23.07</td>
<td>6.6</td>
</tr>
<tr>
<td>WD</td>
<td>100</td>
<td>76.86</td>
<td>73.34</td>
<td>14.24</td>
<td>100</td>
<td>99.77</td>
</tr>
<tr>
<td>Temp</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Hum</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>NO(_x)</td>
<td>91</td>
<td>45.93</td>
<td>79.30</td>
<td>61.22</td>
<td>54.43</td>
<td>26.3</td>
</tr>
<tr>
<td>NO</td>
<td>96.5</td>
<td>47.8</td>
<td>87.66</td>
<td>67.54</td>
<td>49.6</td>
<td>24.63</td>
</tr>
<tr>
<td>SO(_2)</td>
<td>89.23</td>
<td>13.63</td>
<td>7.12</td>
<td>0.06</td>
<td>99.93</td>
<td>91.33</td>
</tr>
<tr>
<td>NO(_2)</td>
<td>10.46</td>
<td>54.36</td>
<td>32.42</td>
<td>47.06</td>
<td>71</td>
<td>84.77</td>
</tr>
<tr>
<td>O(_3)</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>CO</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Selected variables: 2, 3, 4, 5, 6, 7, 9, 10

\[ \pi_{th} = \sqrt{5 \times 10 \times (2 \times (0.95) - 1) \times 10} = 67.08. \] (13)
References


Submit your manuscripts at http://www.hindawi.com