Adaptive Elastic Net with Distance Correlation on the Grouping Effect and Robust of High Dimensional Stock Market Price

(Jaring Elastik Mudah Suai dengan Korelasi Jarak ke atas Kesan Pengelompokan dan Keteguhan Dimensi Tinggi Harga Pasaran Saham)

YUSRINA ANDU, MUHAMMAD HISYAM LEE* & ZAKARIYA YAHYA ALGAMAL

ABSTRACT

Stock market is found in many financial studies. Nonetheless, many of these literatures do not consider on the highly correlated stock market price. In particular, the studies on variable selection, grouping effects and robust dedicated to high dimension stock market price can be considered as scarce. Penalized linear regression using elastic net is one of the recognized methods to perform variable selection. However, the lack of consistency in variable selection may reduce the model performance. Hence, adaptive elastic net with distance correlation (AEDC) is proposed in this study and compared against elastic net, adaptive elastic net with elastic weight and adaptive elastic net with ridge weight. AEDC had lower mean squared error when the alpha increases from 0.05 to 0.95. Thus, the proposed method has successfully contributed to encouraging grouping effects between the highly correlated variables and also has an improved model performance in the presence of robustness.

Keywords: Adaptive elastic net; high dimensional data; penalized linear regression; robust; stock market price

ABSTRAK

Pasaran saham sering ditemui dalam banyak kajian kewangan. Namun begitu, kebanyakan literatur tidak mengambil kira mengenai harga pasaran saham yang berkorelasi tinggi. Secara terperincinya, kajian mengenai pemilihan pemboleh ubah, penggalakan kesan pengelompokan dan keteguhan yang didedikasikan terhadap harga pasaran saham berdimensi tinggi adalah kurang. Kaedah regresi linear terhukum merupakan salah satu kaedah yang diperakui untuk melakukan pemilihan pemboleh ubah. Namun demikian, pemilihan pemboleh ubah yang kurang tekal boleh menjejaskan keberhasilan model. Maka, jaring elastik mudah suai dengan korelasi jarak (EJMSKJ) diusulkan dalam kajian ini dan dibandingkan dengan jaring elastik, jaring elastik mudah suai dengan pemberat elastik dan jaring elastik mudah suai dengan pemberat batas. EJMSKJ mempunyai min ralat kuasa dua yang rendah apabila nilai alfa meningkat daripada 0.05 ke 0.95. Maka, kaedah yang diusulkan telah menyumbang kepada penggalakan kesan pengelompokan antara pemboleh ubah berkorelasi tinggi dan juga keberhasilan model yang lebih baik apabila keteguhan wujud.

Kata kunci: Data dimensi tinggi; harga pasaran saham; jaring elastik mudah suai; regresi linear terhukum; teguh

INTRODUCTION

In many high dimension studies, the explanatory variables are known to be highly correlated, of which this property is also applicable to the high dimension stock market price. When the number of stock market price is greater than the number of observations, applying of statistical approach may be a challenging task. This is due to the presence of overfitting and multicollinearity, which is a well-known problem in many high dimensional studies (Arashi & Roozbeh 2019; Dong et al. 2018).

The idea of adapting ordinary least squares (OLS) as initial weights may be compelling. However, it is not applicable in high dimension data. Therefore, penalized method as initial weight is more suitable in order to

encourage grouping effects between the stock market price variables. Ridge regression is among the widely used penalized method to overcome the multicollinearity problem that is usually presence in the highly correlated data between the variables (Arashi & Roozbeh 2019; Dong et al. 2018). Hoerl and Kennard (1970) whom introduced this method showed that the regression coefficients can never equal to zero although it converges towards zero. This result is obtained by adding an L_2-penalty in the sum of squares residual, of which also causes bias in the estimated parameters as well as increase the regression coefficients variances. Although ridge can be applied to highly correlated stock market price data, however, it might be restricted in performing variable selection. Hence, the interpretation of high dimensional model is not easily obtained (Tibshirani 1996).

To overcome the ridge limitation, another known penalized method which is least absolute shrinkage and selection operator (LASSO) was introduced by Tibshirani (1996). LASSO utilised L_1-penalty instead of L_2-penalty, in which the variable selection can be performed by assigning zero values to some stock market price coefficients. Indeed, LASSO has garnered the attention of many researches particularly in high dimensional studies. On the contrary, it also has some flaws, for instance, the selection of stock market price is usually less than the number of observations. In addition, this method attempts to choose only one stock market price among the highly correlated stock market price. Besides that, the method also has no oracle properties where the probability of choosing the right set of stock market price that have nonzero coefficients converged to one. Similarly, if the zero coefficients were known antecedent, it will have a similar means and covariances as the asymptotically normal nonzero coefficients estimators. Notwithstanding, LASSO has become the baseline of many penalized methods and has elucidate several extensions in diverse practical applications.

Alternatively, Zou and Hastie (2005) has then proposed elastic net method to overcome limitation issue in LASSO by combining both L_1-penalty and L_2-penalty. It is regarded that elastic net often performs much better than LASSO in selection of correlated explanatory variables and prediction accuracy (Zhou 2013). Nevertheless, elastic net is also lack of oracle properties, particularly in the consistency of variable selection. Hence, the adaptive elastic net was introduced

to overcome these drawbacks. The adaptive elastic net developed by Zou and Zhang (2009) proposed the use of elastic net estimates as initial weights. However, when applying adaptive elastic net with elastic net as initial weight, it may cause lower precision and some significant stock market price variables may be incorrectly assigned with smaller weight values in the initial estimator. As a result, the important variables would be falsely removed from the model by the penalized method, which subsequently lower the accuracy of the prediction in the selection of informative stock market price variables. Furthermore, if the pairwise correlation between the stock market price is low, the adaptive elastic net may underperform. Therefore, to encourage grouping effect and reduce bias in variable selection, it is pertinent to propose an alternative initial weight which is more suitable for highly correlated stock market price data.

Most of the methods available in the literature explicitly relies on the assumption of normality. Nonetheless, much of the real applications presents departure from normality. This is often the case of high dimensional data as the existence of heavy tailed distribution or presence of outliers in the response variable. Therefore, this may affect the consistency of conducting variable selection, encouraging grouping effect and robustness. Due to the presence of outliers in the response variable, it is known that conventional method such as OLS might not fully provide the desired estimation (Alhamzawi 2015). Henceforth, robust penalized regression methods have attracted interest of some researchers due to its capability to perform robust variable selection and robust estimation. Nonetheless, majority of the previous studies do not consider the high correlation between the stock market price, as they only focused on solving the outlier problems. Indeed, elastic net is better in encouraging grouping effect as compared to LASSO. However, the limitation of elastic net should also be considered. This can be observed in their poor model performance and lower prediction error, which is contributed further by the inconsistency of variable selection. Furthermore, the lack of oracle properties in elastic net also cannot be neglected. To deal with this shortcomings, adaptive elastic net was proposed (Zou & Zhang 2009). Thus, to overcome the shortcomings mentioned in the existing penalized methods, this study proposed a distance correlation weight as an alternative weight in adaptive elastic net (AEDC).

As of today, many stock market price researches are concentrated only on the financial aspects of the stock market price. The four main financial observations are the trading volume of the stock market prices, the behavior of the stock market (Jafry et al. 2020), the linkage between exchange rates and the stock market (Gharleghi et al. 2014), and the volatility of the stock market prices after a phenomenon such as a recession or news of a company's takeover. However, many of these studies do not discussed the stock market price when they are highly correlated. Therefore, this study will apply the proposed method on the high dimensional stock market price to further improve the grouping effects between the stock market price and enhanced the oracle property of the adaptive elastic net by introducing initial weight.

MATERIALS AND METHODS

PENALIZED LINEAR REGRESSION

The relationship between the explanatory variables and response variable can be describe through linear regression model. The linear regression model between the response vector, $y \in \mathbb{R}^{n \times 1}$ and explanatory variable, $X \in \mathbb{R}^{n \times p}$ (Kurnaz et al. 2018) can be obtained as

$$y = X\beta + \varepsilon \tag{1}$$

where $y = (y_1, ..., y_n)^T$, $X = (x_0, x_1 ..., x_p)^T$, $\beta = (\beta_0, \beta_1..., \beta_p)^T \in \mathbb{R}^{p+1}$ are the unknown regression coefficients and $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T \in \mathbb{R}^n$ is the error vector. The first column of the X is the ones considered for the β_0, of which all of the components in (1) are independently and identically distributed (iid) with mean and variance of $(0, \sigma_\varepsilon^2)$.

Without loss of generality, the assumptions follow that the explanatory variables are standardized, $\sum_{i=1}^n x_{ij} = 0$ and $(n^{-1}) \sum_{i=1}^n = 1$ where $\forall j \in \{1, 2, ..., p\}$ The response variable is also assumed centred as $\sum_{i=1}^n y_i = 0$ without β_0 consideration. Penalized linear regression (PLR) model is defined as the following:

$$\text{PLR} = (y - \beta X)^T (y - \beta X) + \lambda \sum_{j=1}^p P(|\beta_j|) \tag{2}$$

where $\lambda \in [0, \infty]$ which is the tuning parameter and $P(\cdot)$ is the penalty term that represents the coefficients function.

The penalty term has the capability to conduct variable selection which is through shrinking the parameters to zero controlled by the λ values. When the $\lambda = 0$, this is the OLS estimation. Conversely, the coefficient estimates which are influenced by the shrinkage amount increase when the λ has large values. The coefficient estimates can be obtained by the minimization of (2) as:

$$\hat{\beta}_{PLR} = \underset{\beta}{\text{argmin}} \, \text{PLR}. \tag{3}$$

Likewise, Equations (1) and (3) can be written as a constrained minimization problem

$$\hat{\beta}_{PLR} = \underset{\beta}{\text{argmin}} (y - \beta X)^T (y - \beta X) \tag{4}$$

with subject to

$$\sum_{j=1}^p P_\lambda (|\beta_j|) \leq H \tag{5}$$

where the tuning parameter is denoted by $H \geq 0$.

It is worthy of note the advantages of penalized methods especially in their selection stability and computing efficiency as compared to the conventional variable selection methods (Hastie et al. 2015; Rish & Grabarnik 2014). Therefore, various penalty terms have been developed and their properties are methodically discussed.

ELASTIC NET

Elastic net was developed by combining the L_1-norm penalty and the L_2-norm penalty which are from ridge and LASSO, respectively (Zou & Hastie 2005). The ridge penalty is to overcome the highly correlated problem, whereas the LASSO penalty is for the variable selection problem. Elastic net estimator with penalized linear regression is defined as follows,

$$\text{PLR}_{\text{Elastic}}(\beta; \lambda_1, \lambda_2) = (y - \beta X)^T (y - \beta X) + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2 \tag{6}$$

where the two tuning parameters are $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$. The sparsity in the regression coefficients is encourage by the first tuning parameter of λ_1 whereas the grouping effect is encouraged from the second tuning parameter

λ_2 in (6). Meanwhile, increasing λ_1 shrinkage value will subsequently decrease the number of stock market price selected. On the other hand, there is another tuning parameter which is α, a mixing proportion of the LASSO (L_1) and ridge (L_2) penalties in the value of (0, 1). Under the assumption of $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$, a linear combination of both penalties can be obtained by rewriting (6) as:

$$\text{PLR}_{\text{Elastic}}(\boldsymbol{\beta}; \alpha) = (y - \boldsymbol{\beta X})^T (y - \boldsymbol{\beta X}) + (1-\alpha)\sum_{j=1}^{p}|\beta_j| + \alpha\sum_{j=1}^{p}\beta_j^2. \quad (7)$$

Equation (7) is a ridge regression if $\alpha = 1$. On the contrary, it becomes the LASSO regression if $\alpha = 1$. in (7). Noteworthy, when is between 0 and 1, (7) can shrink the coefficients as ridge regression and select variables as LASSO. The penalized linear regression using elastic net regression coefficient estimates can be obtained as:

$$\hat{\beta}_{\text{PLR}}^{\text{Elastic}} = \underset{\beta}{\text{argmin}}\,\text{PLR}_{\text{Elastic}}(\boldsymbol{\beta}; \alpha). \quad (8)$$

ADAPTIVE ELASTIC NET WITH DISTANCE CORRELATION

Zou and Zhang (2009) proposed an adaptive version of the elastic net. This is through merging the ridge penalty with the adaptive weight of L_1-norm penalty. The adaptive elastic net is capable of encouraging the grouping effects between the variables. Meanwhile, distance correlation (DC) follows that two random vectors is equivalent to zero if and only if they are both independent to each other. In addition to this, the two univariate random normal variables DC function is strictly increasing of the absolute Pearson correlation value of these two random normal variables (Székely et al. 2007). The dependence between two random vectors of DC can be measured by a non-negative weight function of $w(t, s)$ where t and s are the dimensions of $p \in \mathbb{R}^m$ and $q \in \mathbb{R}^n$ as:

$$w(t, s) = (c_m c_n |t|^{1+m}\,|s|^{1+n})^{-1},$$

with $c_m = \frac{\pi^{1+m/2}}{\Gamma((1+m)/2)}$, a non-negative constant associated to the dimensionality of c_m and the Gamma complete function is denoted by $\Gamma(\cdot)$ (Shen et al. 2020; Székely et al. 2007). The Euclidean norm of is defined as $\|a\| = a^T a$ if is real vector and $\|\psi\|^2 = \psi\bar{\psi}$ for a complex-valued ψ function, with $\bar{\psi}$ is the conjugate of ψ. Thus, AEDC is

obtained as follows:

$$\text{PLR}_{\text{AEDC}}(\boldsymbol{\beta}; \lambda_1, \lambda_2) = (y - \boldsymbol{\beta X})^T(y - \boldsymbol{\beta X}) + \lambda_1\sum_{j=1}^{p}w_j|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2,$$

where the estimation of adaptive elastic net is denoted by

$$\tilde{\beta}_{\text{AEDC}} = \underset{\beta}{\text{argmin}}\left\{(y - \boldsymbol{\beta}\tilde{X})^T(y - \boldsymbol{\beta}\tilde{X}) + \lambda_1\sum_{j=1}^{p}w_j|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2\right\} \quad (9)$$

and w_j in (9) is the variables screen using DC approach which is

$$w_j = dCorr(x, y) = \frac{dCov(x,y)}{\sqrt{(dVar(x) \cdot dVar(y))}}, \quad j = 1, \ldots, p. \quad (10)$$

RESULTS AND DISCUSSION

SIMULATION STUDY

The simulation matrix is generated according to the combination of linear, nonlinearity and seasonality pattern. Apart from resembling the stock market price, it is also pertinent to include the error that has a heavy-tailed distribution or outliers in the response variable. Thus, the response variable is generated to have a seasonality pattern. The following shows the generated matrix simulation. In the linear model, the matrix simulation is built up as the following. A random uniform distribution matrix of $n = 30$ and $p = 200$ in the interval of 0 to 1 is generated through (11)

$$Y_t = 0.4Y_{t-1} - 0.3Y_{t-2} + u_r \quad (11)$$

where u_r and Y_t with $1 \le r \le p$ and $1 \le t \le n$.
Meanwhile, in the nonlinear matrix design, it was also generated randomly using uniform distribution of $\mathbf{U}[0,1]$ where $n = 30$ and $p = 400$ as in shown in (12).

$$Y_r = u_r - 0.4u_{r-1} + 0.3u_{r-2} + 0.5u_r u_{r-2} \quad (12)$$

The seasonality pattern was generated following the uniform distribution of [0,1] whereby $u_{i,t}$ is $1 \le t \le T$ and $1 \le i \le m$ and f_t, $-2 \le t \le T$ respectively, with $n = 30$ and $p = 401$ as in (13). In addition, this also includes the response variable matrix value.

$$z_{i,t} = 10\sin\left(2\pi\left(\frac{i}{m}\right)\right)f_t + 10\cos\left(2\pi\left(\frac{i}{m}\right)\right)f_{t-1} + u_{i,t} \quad (13)$$

The computing steps for the simulation is presented in Table 1.

TABLE 1. Algorithm of initial weight in the simulation

Algorithm: The implementation of initial weight for simulation

Step 1:	**Data**: Generate the high dimension matrix data which are linear, nonlinear and have seasonality pattern using (11), (12) and (13). Set **runs** = 0.
	Initialize: runs ← runs + 1
	For each combination, generate the following (x, y) at random $\mathbf{U}[0,1]$ of which dimension of $n = 30$, $p = 1001$, as
	• Equation (11) with $n = 1, ..., 30$, $p = 1, ..., 200$ at
	• Equation (12) with $n = 1, ..., 30$, $p = 201, ..., 600$
	• Equation (13) with $n = 1, ..., 30$, $p = 601, ..., 1001$
	end at **runs** = 1000.
Step 2:	Find w_j, where $\tilde{x}_i = x_j/\lvert w_j \rvert$, where $j = 1, ..., p$ by screening the variables using distance correlation $dCorr(x, y)$.
Step 3:	Define
Step 4:	Calculate the estimation of adaptive elastic net with proposed initial weights as $$\tilde{\hat{\beta}}_{\text{AEDC}} = \underset{\beta}{\arg\min} \left\{ (y - \beta\tilde{X})^T (y - \beta\tilde{X}) + \lambda_1 \sum_{j=1}^{p} w_j \lvert \beta_j \rvert + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\}$$ at $\alpha = 0.05$ to $\alpha = 0.95$.
Step 5:	Repeat Step 1- 4 and displayed the results as an average of the simulation runs.

A 5-fold cross-validation was performed to assist λ in selecting the optimal tuning parameters in all the first part of all the three adaptive elastic net excluding the elastic net. The next step was introduced the initial weights in these adaptive elastic net methods namely with elastic net weight (AEN), adaptive elastic net with ridge weight (AER), and AEDC. On the other hand, there was no initial weight introduced in the elastic net (EN) method.

In AEDC, DC between each explanatory variables and response variable was calculated to screen the most relevant variables. The number of rows of both variables must be equivalent and without any missing values. After the initial weights were introduced, the 5-fold cross-validation were run once more to obtain the minimum λ value in AEN, AER, and AEDC. Since no additional weights were added in the elastic net, thus, the simulation step was initiate here. Next, the generalized linear model was fitted using linear regression, using the minimum λ obtained previously. The subsequent calculation steps are executed accordingly.

The simulation was repeated for 1000 times to ensure the stability of all methods in this section. Additionally, the performance of each method is assessed by different alpha (α) values ranging from 0.05 to 0.95. AEDC was compared with elastic net, AEN and AER based on the mean squared error (MSE) and number of selected stock market price. The result are summarized as in Tables 2 and 3, where the standard deviation is denoted in parentheses.

TABLE 2. Simulation result of mean squared error at different α

α	EN	AEN	AER	AEDC
0.05	0.0013 (0.0012)	0.0013 (0.0012)	0.0013 (0.0012)	0.0013 (0.0012)
0.10	0.0035 (0.0031)	0.0034 (0.0030)	0.0034 (0.0030)	0.0033 (0.0031)
0.15	0.0063 (0.0055)	0.0061 (0.0054)	0.0061 (0.0054)	0.0060 (0.0055)
0.20	0.0096 (0.0083)	0.0094 (0.0081)	0.0094 (0.0082)	0.0092 (0.0083)
0.25	0.0133 (0.0114)	0.0131 (0.0111)	0.0130 (0.0113)	0.0128 (0.0113)
0.30	0.0174 (0.0146)	0.0170 (0.0143)	0.0169 (0.0145)	0.0167 (0.0145)
0.35	0.0216 (0.0180)	0.0212 (0.0176)	0.0211 (0.0179)	0.0208 (0.0179)
0.40	0.0260 (0.0214)	0.0256 (0.0210)	0.0255 (0.0213)	0.0250 (0.0212)
0.45	0.0306 (0.0248)	0.0300 (0.0243)	0.0299 (0.0247)	0.0294 (0.0246)
0.50	0.0351 (0.0281)	0.0345 (0.0276)	0.0344 (0.0281)	0.0338 (0.0279)
0.55	0.0396 (0.0314)	0.0390 (0.0308)	0.0388 (0.0313)	0.0382 (0.0311)
0.60	0.0441 (0.0346)	0.0435 (0.0339)	0.0432 (0.0345)	0.0426 (0.0343)
0.65	0.0486 (0.0377)	0.0479 (0.0370)	0.0476 (0.0376)	0.0469 (0.0374)
0.70	0.0529 (0.0406)	0.0522 (0.0399)	0.0519 (0.0406)	0.0512 (0.0404)
0.75	0.0571 (0.0435)	0.0564 (0.0427)	0.0560 (0.0434)	0.0553 (0.0433)
0.80	0.0612 (0.0462)	0.0605 (0.0454)	0.0600 (0.0461)	0.0593 (0.0460)
0.85	0.0652 (0.0487)	0.0644 (0.0478)	0.0639 (0.0486)	0.0632 (0.0485)
0.90	0.0690 (0.0511)	0.0682 (0.0501)	0.0677 (0.0510)	0.0670 (0.0509)
0.95	0.0726 (0.0533)	0.0718 (0.0522)	0.0713 (0.0532)	0.0705 (0.0531)

Based on Table 2, the MSE value was similar among the four methods at $\alpha = 0.05$. It can be clearly observed that when the α increases from 0.05 to 0.95, the MSE also increases. Nonetheless, the MSE for our proposed method were the lowest compared to the other methods as α increases up to 0.95 from 0.10. This was followed close by AEN and AER. Conversely, the elastic net had the highest MSE than of other three methods. Furthermore, the increment of the α values causes a wider MSE gap when it is compared to our proposed method.

In contrast, the number of selected variables in Table 3 decreases as the α increases to 0.95 in all elastic net, AEN, AER and AEDC, respectively. It can be observed clearly that AER had the highest number of selected variables when $\alpha = 0.05, 0.10$ and $\alpha = 0.20$, when compared to other three methods. Contrariwise, the remaining α showed that AEDC successfully outperformed elastic net, AEN and AER having the highest number of variables. The fact that AEN and elastic net did not outperform the other two methods were practically due to the inability of addressing the heavy-tailed distribution error or presence of outliers in the response variable.

TABLE 3. Simulation result of selected variables at different α

α	EN	AEN	AER	AEDC
0.05	163.76 (21.90)	163.97 (21.90)	164.38 (21.73)	164.26 (22.03)
0.10	99.71 (14.19)	99.68 (14.10)	100.03 (13.83)	100.01 (14.07)
0.15	72.87 (11.15)	72.92 (11.20)	73.21 (10.97)	73.23 (11.06)
0.20	57.78 (9.51)	57.95 (9.60)	57.99 (9.63)	57.94 (9.64)
0.25	47.68 (8.81)	47.85 (8.71)	48.07 (8.87)	48.08 (8.82)
0.30	40.66 (8.68)	40.68 (8.67)	41.02 (8.67)	41.05 (8.67)
0.35	35.24 (8.59)	35.33 (8.60)	35.51 (8.60)	35.79 (8.68)
0.40	30.90 (8.62)	30.96 (8.34)	31.30 (8.69)	31.41 (8.52)
0.45	27.45 (8.69)	27.48 (8.39)	27.76 (8.79)	27.99 (8.67)
0.50	24.56 (8.79)	24.70 (8.63)	24.88 (8.93)	25.04 (8.72)
0.55	22.07 (8.89)	22.19 (8.67)	22.44 (8.95)	22.59 (8.80)
0.60	19.93 (8.95)	20.12 (8.78)	20.31 (8.80)	22.59 (8.80)
0.65	18.11 (9.03)	18.27 (8.90)	18.41 (8.95)	18.69 (8.95)
0.70	16.55 (9.10)	16.66 (8.95)	16.80 (9.04)	17.07 (9.04)
0.75	15.07 (9.06)	15.29 (8.98)	15.41 (9.10)	15.61 (9.10)
0.80	13.88 (9.11)	14.03 (8.98)	14.16 (9.13)	14.38 (9.13)
0.85	12.78 (9.11)	12.83 (8.90)	13.10 (9.08)	13.23 (9.08)
0.90	11.80 (9.07)	11.87 (8.85)	12.17 (9.02)	12.19 (9.02)
0.95	10.98 (9.04)	10.98 (8.78)	11.30 (9.20)	11.33 (8.97)

In addition, from both of the MSE and number of selected variables, the simulation results showed that the DC weight has significantly improved the performance of AEDC. Modelling with AEDC have better result performance which was contributed by the lower values of MSE. Thus, based on simulation study, AEDC has proved the capability to performed variable selection efficiently, especially in the presence of a heavy-tailed distribution error or outliers in the response variables.

YEARLY STOCK MARKET PRICE

The stock market price obtained from Yahoo Finance consists of 4362 stock market prices over the period of 30 years from 1987 to 2017. By removing the missing observations, the finalized data consists of 901 stock market prices. Majority of S&P 500 index values were derived from the 901 stock market prices, therefore it was chosen as the response variable (Andu et al. 2020). The performance results were summarised as in Tables 4 and 5.

TABLE 4. Result of mean squared error at different of stock market price data

α	EN	AEN	AER	AEDC
0.05	5.49	5.98	5.22	6.09
0.10	11.95	13.05	10.88	9.89
0.15	19.84	21.89	18.17	16.68
0.20	30.16	32.58	29.48	24.48
0.25	39.87	42.38	36.96	35.38
0.30	53.78	57.96	48.87	45.14
0.35	70.07	77.63	67.60	60.30
0.40	91.49	98.73	85.18	80.14
0.45	114.28	122.86	103.70	96.50
0.50	133.88	145.55	126.12	113.61
0.55	157.87	180.03	154.35	132.51
0.60	191.88	204.74	169.64	162.05
0.65	212.59	229.63	195.69	181.32
0.70	240.79	267.42	222.83	204.29
0.75	267.98	291.11	247.95	229.50
0.80	313.46	337.76	282.94	261.81
0.85	346.86	369.51	322.12	293.23
0.90	375.25	405.57	357.75	325.34
0.95	407.74	448.77	377.37	349.50

TABLE 5. Number of selected stock market price at different of stock market price data

α	EN	AEN	AER	AEDC
0.05	194	201	192	146
0.10	165	172	164	165
0.15	121	126	123	124
0.20	104	99	105	104
0.25	94	94	90	90
0.30	83	79	81	84
0.35	71	77	75	75
0.40	69	71	67	69
0.45	64	63	65	64
0.50	56	56	54	59
0.55	55	49	56	56
0.60	42	48	54	49
0.65	40	43	40	39
0.70	37	39	40	40
0.75	36	33	34	39
0.80	30	30	31	33
0.85	25	28	26	25
0.90	24	24	26	25
0.95	22	23	24	24

Table 4 shows that the value of MSE increase as the α increase. AEDC had the highest MSE at $\alpha = 0.05$ compared to the other three penalized methods. Nonetheless, it outperformed the rest of the methods in terms of model performance as it had the lowest MSE with the increment of α. Furthermore, AEDC also showed higher prediction power. Meanwhile, AER had better model performance than AEN based on the lower MSE. Although elastic net had smaller MSE compared to AER at $\alpha = 0.05$, however, as an overall, it did not perform well in both model performance and prediction.

On the contrary, Table 5 shows that the number of selected variables decreases with the increment of α value and it was similar to the simulation result pattern. There were three α that have similar fair share in the number of selected variables, for example, 94 ($\alpha = 0.25$) in AEN and elastic net method, and 40 ($\alpha = 0.70$) and 24 ($\alpha = 0.95$) in AER and AEDC method. It can also be observed that AEN had eight variables selected, followed by six equally shared variables selected in AER and AEDC, within the range of α provided. Although AEN shows higher number of variables selected, nevertheless, it may not be efficient in explaining the presence of outliers or heavy tailed distribution in the variables. This reason also applies similarly to AER. Notwithstanding, the variation in the selected number of stock market price showed the efficiency of AEDC in dealing with the presence of outliers or heavy-tailed distribution in the response variables.

Conclusion

Our study had successfully proposed initial weights using distance correlation in the adaptive elastic net. Remarkably, the comparison in both simulation and real application showed that AEDC yield better prediction power and achieved a satisfactory model performance when compared to elastic net, AEN, and AER. Based on the performance results, the proposed initial weight was able to further enhance AEN explicitly in the consistency of the number of stock market price selected. In addition, AEDC can be efficiently applied in different range of α values which has been evidently presented through simulation study and stock market price application.

The efficiency of AEDC in handling the robustness particularly in the existence of a heavy tailed distribution error or outliers in the response variable of the stock market price has been notably presented. Notwithstanding, the number of stock market price selected, correspondingly,

the distance correlation had successfully encouraged grouping effects between the selected stock market price. Furthermore, the proposed weight also has effectively shown the improvement in the prediction performance, model interpretation and the oracle property in the stock market price application.

The assessment of high dimension stock market price using proposed initial weight in adaptive elastic net has been effectively performed in this study. In addition, our proposed method of AEDC would be an added value to the existing literature of stock market price studies, where previous studies were mostly focused on the financial aspects and machine learning approaches. Nevertheless, the result performances of high dimension stock market price using AEDC obtained in this study may also be useful to the financial experts to identify the significant stock market price on the stock index as well as for prediction purpose.

Since the presented work in the high dimension focused on highly correlated data, nevertheless, the proposed methods here could be extended to data that do not necessarily own this property. Further study can be conducted in the case where the high dimension data has lower or moderate correlation between the explanatory variables and the response variable.

REFERENCES

Alhamzawi, R. 2015. Model selection in quantile regression models. *Journal of Applied Statistics* 42(2): 445-458.

Andu, Y., Lee, M.H. & Algamal, Z.Y. 2020. Variable selection of yearly high dimension stock market price using ordered homogenous pursuit lasso. *AIP Conference Proceedings* 2266(1): 090012.

Arashi, M. & Roozbeh, M. 2019. Some improved estimation strategies in high-dimensional semiparametric regression models with application to riboflavin production data. *Statistical Papers* 60(3): 317-336.

Dong, Y., Song, L. & Amin, M. 2018. SCAD-ridge penalized likelihood estimators for ultra-high dimensional models. *Hacettepe Journal of Mathematics and Statistics* 47(2): 423-436.

Gharleghi, B., Shaari, A.H. & Sarmidi, T. 2014. Application of the threshold model for modelling and forecasting of exchange rate in selected ASEAN countries. *Sains Malaysiana* 43(10): 1609-1622.

2764

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

Hoerl, E. & Kennard, R.W. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12(1): 55-67.

Jafry, N.H.A., Ab Razak, R. & Ismail, N. 2020. Ukuran kebersandaran bagi pulangan lima-minit berbanding pulangan harian menggunakan kopula statik dan dinamik. *Sains Malaysiana* 49(8): 2023-2034.

Kurnaz, F.S., Hoffmann, I. & Filzmoser, P. 2018. Robust and sparse estimation methods for high-dimensional linear and logistic regression. *Chemometrics and Intelligent Laboratory Systems* 172: 211-222.

Rish, I. & Grabarnik, G. 2014. *Sparse Modeling: Theory, Algorithms, and Applications*. Boca Raton: CRC Press.

Shen, C., Priebe, C.E. & Vogelstein, J.T. 2020. From distance correlation to multiscale graph correlation. *Journal of the American Statistical Association* 115(529): 280-291.

Székely, G.J., Rizzo, M.L. & Bakirov, N.K. 2007. Measuring and testing dependence by correlation of distances. *Annals of Statistics* 35(6): 2769-2794.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1): 267-288.

Zhou, D.X. 2013. On grouping effect of elastic net. *Statistics and Probability Letters* 83(9): 2108-2112.

Zou, H. & Zhang, H.H. 2009. On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics* 37(4): 1733-1751.

Zou, H. & Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2): 301-320.

Yusrina Andu
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA, Kuala Pilah Campus
72000 Kuala Pilah, Negeri Sembilan Darul Khusus
Malaysia

Yusrina Andu & Muhammad Hisyam Lee*
Department of Mathematical Sciences
Faculty of Science
Universiti Teknologi Malaysia
81310 Johor Bahru, Johor Darul Takzim
Malaysia

Zakariya Yahya Algamal
Department of Statistics and Informatics
University of Mosul
Mosul, Iraq

*Corresponding author; email: mhl@utm.my