doi.org/10.36253/ijas-16752

A MODIFIED CORRELATION BASED REGULARIZATION TECHNIQUE FOR REGRESSION ESTIMATION AND FEATURE SELECTION

Isaac Adeola Adeniyi, Dolapo Abidemi Shobanke

Department of Statistics, Federal University Lokoja, P.M.B. 1154, Lokoja, Nigeria

Abstract. *Variable selection is important for making sense with (ultra) high-dimensional data. Penalized least squares such as the LASSO, elastic-net and the correlation based elastic-net (L1CP) are popular methods for carrying out variable selection and estimation simultaneously. This study proposes a modified version of the L1CP motivated by reasons similar to that given by Zou and Hastie (2005) where the naïve elastic net was rescaled to give the elastic net. The scaling transformation is derived such that the double shrinkage caused by applying two penalties is undone thereby reducing bias. The derived scaling transformations are found to depend on the correlations among the predictors. A robust worst-case quadratic solver is used to obtain estimates. An evaluation of the proposed method which is referred to as CL1CP alongside the L1CP, LASSO and elastic-net through simulation studies illustrate the advantages of the CL1CP compared to the other alternatives considered especially in correct selection of sparse models. In terms of variable selection, estimation and prediction accuracy the proposed CL1CP performs favourably compared to the L1CP, LASSO and elastic-net especially for "grouped-variables" selection. Results from applications to two real life datasets corroborate the findings from simulation studies.*

Keywords*: Dimension reduction; Penalization; Sparsity; Data mining; Machine learning.*

1. INTRODUCTION

The multiple linear regression model is frequently used in studies for carrying out predictions and studying the effects of some predictor variables on a continuous response variable. Let the response vector be denoted by $Y = (y_1, \dots, y_n)^T$ and $X = [X_1, \dots, X_p]^T$ denote the predictor matrix, where p is the number of predictor variables. The regression model that represents the relationship between Y and $X_1, X_2, \cdots, X_p$ is

$$Y = \beta_0 + X^T\beta + \epsilon, \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$ is a vector of unknown regression coefficients to be estimated with β_j representing the regression coefficient corresponding to predictor variable X_j, $j = 1, \ldots, p$; β_0 is the intercept and ϵ is a vector of error terms. Throughout this paper, we assume the errors ϵ are identically and independently distributed with zero mean and finite variance σ^2. The ordinary least squares (OLS) technique which is the classical approach for estimating the coefficients can be inefficient and not applicable when the predictors are highly correlated and when p is more than n respectively (Hoerl and Kennard, 1970; Ryan, 2009; Wang et al., 2016). With the advancement in technology, high and ultra-high dimensional data where number of variables p exceeds the sample size n are emerging from medical research and many other areas (Yahya et al., 2011; Hapfelmeier et al., 2012; Yahya et al., 2014; Yahya et al., 2015).

Variable selection is basically important for making sense with (ultra) high-dimensional data (Fan and Li, 2006) and it could lead to increased prediction performance of the fitted model. Classical model selection methods are the best-subset selection method and its step-wise variants. Nevertheless, best-subset selection is computationally impracticable when the number of predictors is large. Furthermore, as reported by Breiman (1996), subset selection is unstable while step-wise methods can perform very poorly resulting in a model with poor prediction accuracy. Also, when predictors are highly correlated, estimates of β by the OLS are not unique (Tan, 2012). The ridge regression (Hoerl and Kennard, 1970) uses an L_2-norm penalty to improve OLS when the predictors are correlated. The ridge estimator is obtained by solving the L_2-norm penalized least squares problem

$$\hat{\beta}_{ridge} = arg\min_{\beta} \|Y - \beta_0 - X^T\beta\|_2^2 + \lambda\|\beta\|_2^2$$

where $\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$ is the L_2-norm of β. The ridge regression tries to take care of the problems caused by presence of correlated predictors in the model but does not carry out variable selection.

One of the most popular methods that have been proposed to overcome the underlying drawbacks of subset selection, stepwise selection and the OLS is the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996). The LASSO estimator is obtained by solving the L_1 penalized least squares problem

$$\hat{\beta}_{lasso} = arg\min_{\beta} \|Y - \beta_0 - X^T\beta\|_2^2 + \lambda\|\beta\|_1,$$

where $\|\beta\|_1^1 = \sum_{j=1}^{p} |\beta_j|$ is the L_1-norm of β. Although the LASSO has shown success in many situations, some of its limitations include its inability to select more than n predictors in the $p > n$ situation; tendency to select only one

variable from a group of predictors with high pairwise correlations and poor performance compared to ridge regression in the $n > p$ situation. To overcome the problems of LASSO highlighted above, Zou and Hastie (2005) proposed the elastic-net which is a combination of the L_1 penalty of the LASSO and the L_2 penalty of the ridge. The naïve elastic net estimator $\beta_{naive-enet}$ is the solution to

$$\hat{\beta}_{naive-enet} = \arg\min_{\beta} \|Y - \beta_0 - X^T\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2. \quad (2)$$

Zou and Hastie (2005) called the estimator obtained from (2) the naïve elastic-net estimator and showed that it does not perform satisfactorily nor possess the minimax optimal property because of the double shrinkage due to the L_1 and L_2 penalties which introduces unnecessary extra bias. Zou and Hastie (2005) obtained the elastic-net (ENET) estimator as $\left(1 + \frac{\lambda_2}{n}\right)\hat{\beta}_{naive-enet}$ and $(1 + \lambda_2)\hat{\beta}_{naive-enet}$ if the predictors are standardized (each variable has mean zero and l_2-norm one). However, Anbari and Mkhadri (2014) remarked that the ENET seems to be slightly less reliable if the correlation between variables is not so extreme. Besides, ENET ignores the information concerning the correlation of the data in the l_2 norm penalty.

With similar intent, the octagonal shrinkage and clustering algorithm for regression (OSCAR) which is based on the penalized least squares with a penalty function combining the L_1 and the pairwise L_∞ norms was proposed by Bondell and Reich (2008). The OSCAR constrains some coefficients to be identically equal, allowing correlated predictors that have similar effect on the response to form clusters represented by the same coefficients. Nevertheless, Anbari and Mkhadri (2014) observed that obtaining the OSCAR estimates can be slow for large p. Alternatively, Tutz and Ulbricht (2009) proposed a combination of a correlation based penalty (CP) and a blockwise boosting procedure (BB) for carrying out shrinkage and variable selection. But, in practice determining the step length factor and the stopping number of iterations for the blockwise boosting procedure can sometimes be tough and affects the sparsity of the solution as well as the speed of the procedure. As a result, Anbari and Mkhadri (2014) introduced the correlation based elastic-net estimator $\left(\hat{\beta}_{cp}\right)$ which is the minimizer of

$$\|Y - \beta_0 - X^T\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2 \sum_{j=1}^{p-1}\sum_{k>j}\left[\frac{\left(\beta_j - \beta_k\right)^2}{1 - r_{jk}} + \frac{\left(\beta_j + \beta_k\right)^2}{1 + r_{jk}}\right], \quad (3)$$

where r_{jk} is the empirical correlation between the predictors X_j and X_k. The L1CP is the combination of the LASSO and the correlation based penalty (CP) of Tutz and Ulbricht (2009). Similarly, Tan (2012) also proposed two correlation adjusted

elastic net (namely CAEN1 and CAEN2) penalties for linear regression and were further extended for the Poisson regression by Algamal and Lee (2015). The L1CP was found to have good empirical performance compared to the LASSO and ENET. Nonetheless, the double shrinkage incurred by using both the L_1 and correlation based penalties was not accounted for in the L1CP which can lead to unnecessary bias. Thus, the performance of the L1CP can be improved if the double shrinkage is undone.

Our goal in this study is to introduce scaled version of the L1CP. The main idea is to improve the performance of the L1CP by scaling the estimates in order to undo the double shrinkage obtained by combining both the L_1 and correlation based penalties. In this paper, we call the estimator proposed by Anbari and Mkhadri (2014) naïve L1CP while we call our proposed estimators CL1CP. We obtain the scales by carrying out a decomposition of the CP estimator (Tutz and Ulbricht, 2009) using ideas similar to those used by Zou and Hastie (2005) to correct the naïve ENET to give the ENET estimator.

The rest of the paper is organized as follows; In Section 2, we introduce the CL1CP and the estimators with corresponding proposed methods of rescaling. We carry out simulation studies to evaluate the finite sample performance of our proposals in comparison to the naïve L1CP as well as other competitors such as the LASSO and ENET in Section 3. In Section 4, we apply all the methods considered in Section 3 on real life data. Concluding remarks are given in Section 5.

2. THE METHODOLOGY- RESCALED L1CP (CL1CP)

In this section, we present the elements of the correction for the L1CP procedure to give the CL1CP. The objective function to be minimized to yield the L1CP regression estimates is

$$\|Y - \beta_0 - X^T\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2 P(\beta), \tag{4}$$

where

$$P(\beta) = \sum_{j=1}^{p-1}\sum_{k>j}\left[\frac{(\beta_j - \beta_k)^2}{1 - r_{jk}} + \frac{(\beta_j + \beta_k)^2}{1 + r_{jk}}\right].$$

Our proposed corrected estimator is of the form

$$\hat{\beta}_{CL1CP} = S \times \left[arg\min_{\beta}\|Y - \beta_0 - X^T\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2 P(\beta)\right], \tag{5}$$

where S is the scaling factor and r_{jk} is the empirical Pearson's product moment correlation between X_j and X_k. Details on the scaling factor are given in the

following. Tutz and Ulbricht (2009) showed that $P(\beta)$ can be written in a simple quadratic form:

$$P(\beta) = \beta^T W_{cp} \beta,$$

where $W_{cp} = \left(w_{ij}\right)_{1 \le i,j \le p}$ is a matrix such that

$$w_{ij} = \begin{cases} 2\sum_{s \neq i} \frac{1}{1 - r_{is}^2}, & i = j \\ \frac{-2r_{ij}}{1 - r_{ij}^2}, & i \neq j \end{cases},$$

and $r_{ij}^2 \neq 1, i, j = 1, \dots, p$. The estimator proposed by Anbari and Mkhadri (2014) which we call naïve L1CP estimator is a two-stage procedure: the correlation based penalty (CP) (Tutz and Ulbricht, 2009) regression coefficients are first obtained for each fixed λ_2, and then the lasso type shrinkage is carried out along the lasso coefficient solution paths. This leads to a double amount of shrinkage which introduces unnecessary extra bias, compared with pure CP or LASSO. We, therefore, introduce a corrected L1CP (CL1CP) estimator $\hat{\beta}_{CL1CP}$ which is given as

$$diag\left(\lambda_2 W_{cp} + I\right)\left[arg\min_{\beta} \|Y - \beta_0 - X^T\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2 \beta^T W_{cp} \beta\right], \qquad (6)$$

where I is a $p \times p$ identity matrix and $diag(A)$ is a diagonal matrix such that its diagonal elements are the same as the diagonal elements of A. It is easy to see that the CL1CP estimator is given by (5) when

$$S = diag\left(\lambda_2 W_{cp} + I\right). \qquad (7)$$

The scale factor for a single β_j can be simplified to

$$S_{jj} = 1 + 2\lambda_2 \sum_{s \neq j}^{p} \frac{1}{1 - r_{js}^2}.$$

The motivation for adopting the $diag\left(\lambda_2 W_{cp} + I\right)$ is similar to the argument used by Zou and Hastie (2005) to improve the performance of the ENET by rescaling the naïve ENET. First, it should be recalled that the correlation based penalty (CP) (Tutz and Ulbricht, 2009) estimator is given as

$$\hat{\beta}_{cp} = \left(X^T X + \lambda_2 W_{cp}\right)^{-1} X^T Y.$$

Now, let us consider a decomposition of the CP estimator with $\boldsymbol{X}$ standardized, we have

$$X^T X = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{12} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \dots & 1 \end{bmatrix}.$$

Thus, the CP estimator can be rewritten as

$$\hat{\beta}_{cp} = R^{-1} \begin{bmatrix} 1 & r_{12}\left(1 - \frac{2\lambda_2}{1 - r_{12}^2}\right) S_1^{-1} & \dots & r_{1p}\left(1 - \frac{2\lambda_2}{1 - r_{1p}^2}\right) S_1^{-1} \\ r_{12}\left(1 - \frac{2\lambda_2}{1 - r_{12}^2}\right) S_2^{-1} & 1 & \dots & r_{2p}\left(1 - \frac{2\lambda_2}{1 - r_{2p}^2}\right) S_2^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p}\left(1 - \frac{2\lambda_2}{1 - r_{1p}^2}\right) S_p^{-1} & r_{2p}\left(1 - \frac{2\lambda_2}{1 - r_{2p}^2}\right) S_p^{-1} & \dots & 1 \end{bmatrix}^{-1} X^T Y, \quad (8)$$

where $R = \begin{bmatrix} S_1 & 0 & \dots & 0 \\ & S_2 & \dots & 0 \\ & & \ddots & \vdots \\ & & & S_p \end{bmatrix}$.

From (8), one can observe that L1CP estimator involves scaling the correlations r_{ij} by a factor $\left(1 - \frac{2\lambda_2}{1 - r_{ij}^2}\right) S_{jj}^{-1}$ followed by direct shrinkage with R^{-1} leading to double amount of shrinkage. When we combine the CP with the LASSO, the direct R^{-1} shrinkage step is not needed and removed by our scaling factor. In our proposed method, the shrinkage by LASSO is sufficient for controlling the variance and obtaining sparsity and we, therefore undo the R^{-1} shrinkage step by multiplying the naïve L1CP estimates by $diag(\lambda_2 W_{cp} + I)$. It can be easily shown that

$$R = \begin{bmatrix} S_{11} & 0 & \dots & 0 \\ & S_{22} & \dots & 0 \\ & & \ddots & \vdots \\ & & & S_{pp} \end{bmatrix} = diag(\lambda_2 W_{cp} + I).$$

2.1 Estimation and Selection of Tuning Parameters λ_1 and λ_2

The properties of the L1CP highlighted in Anbari and Mkhadri (2014) are preserved in our proposed scaling transformations. Anbari and Mkhadri (2014) showed that the L1CP can be augmented to become a LASSO type problem. Therefore, existing computational techniques for penalized regression methods can easily be used for obtaining the CL1CP. Following Anbari and Mkhadri (2014), the L1CP optimization problem can be written as

$$arg\min_{\beta^*}\left\|Y^* - \beta_0^* - X^{*T}\beta^*\right\|_2^2 + \lambda_1\|\beta^*\|_1, \tag{9}$$

where,

$X^*_{(n+p)\times p} = \begin{pmatrix} X \\ \sqrt{\lambda_2}L^T \end{pmatrix}, Y^*_{(n+p)} = \begin{pmatrix} Y \\ 0 \end{pmatrix}$ and $L^T = W_{cp}^{\frac{1}{2}}$.

L^T is obtained from the Cholesky decomposition of W_{cp}, that is, $LL^T = W_{cp}$. Note that W_{cp} is a real symmetric positive-definite square matrix, and that the L1CP estimator becomes the naïve Elastic-Net estimator when $W_{cp} = I$. See Anbari and Mkhadri (2014) for more details on the proof of how the L1CP method can be augmented to become LASSO type problems. The CL1CP estimator $\hat{\beta}_{CL1CP}$ can, therefore, be expressed as

$$diag(\lambda_2 W_{cp} + I) \times \left[arg\min_{\beta^*}\left\|Y^* - \beta_0^* - X^{*T}\beta^*\right\|_2^2 + \lambda_1\|\beta^*\|_1\right]. \tag{10}$$

It can be clearly seen that the CL1CP estimator becomes the Elastic-Net estimator when $W_{cp} = I$ whereas the L1CP can only reduce to the naïve Elastic-Net estimator. This is another justification for the need to rescale the L1CP as Zou and Hastie (2005) showed that the naïve Elastic-Net estimator does not perform adequately.

We adopt the quadratic solver method proposed by Grandvalet et al. (2017) to obtain the correlation based regression coefficients as well as the LASSO and elastic-net regression coefficients. In this approach, the inner minimization problem is seen as a simple unconstrained quadratic problem, and an optimization strategy based on the iterative resolution of plain quadratic problems is used to obtain minimizers of the objective function. The algorithm is highly computationally efficient and numerically robust compared to other optimization strategies for sparse regression such as coordinate descent (Fu, 1998) or proximal methods (Beck and Teboulle, 2009). The algorithm relies on second order techniques to solve a series of small-size quadratic approximations to the objective function defined from the currently estimated worst-case perturbation leading to stable convergence and reduced sensitivity to numerical errors, unlike first-order methods such as coordinate descent or proximal methods. The general

scheme of the algorithm starts from a vector of zeros and continues to keep an active set of variables. Only the active variables are involved when solving the underlying optimization problem, hence the size of the underlying problems solved is related to the number of nonzero coefficients in the vector of parameters. Therefore, by considering a decreasing grid of values for λ_1 and fixing λ_2, the entire path of solutions can be explored at a reasonable computational cost. The R (R Core Team, 2018) package *quadrupen* (Grandvalet et al., 2017) contains implementation of this algorithm. For more details on the algorithm, see Grandvalet et al. (2017).

Selecting the tuning parameters λ_1 and λ_2 is not a trivial task in practice and is very important in order to achieve a good prediction and estimation accuracy. The tuning parameters can be chosen by minimizing an estimate of the out-of-sample prediction error. A validation set can be used to estimate this directly if available, otherwise, one can use five-fold or ten-fold cross validation (Hastie *et al.*, 2001; Efron and Tibshirani, 1997; Tibshirani, 1996; Kohavi, 1995; Efron and Tibshirani, 1993). In this paper, we use ten-fold cross validation (10-fold CV) to select tuning parameters in applications on real life datasets while validation sets are generated to select tuning parameters in the simulation studies. The cross-validation (or validation) is done on a two dimensional surface since there are two tuning parameters in the ENET, L1CP and CL1CP. Usually a grid of λ_2 is chosen first, then for each λ_2 the entire solution path of the ENET, L1CP or CL1CP is produced by the quadratic solver algorithm. The second tuning parameter λ_1 is chosen by the 10-fold CV error. The chosen pair of λ_1 and λ_2 is the one with the smallest 10-fold CV error. See Hastie *et al.* (2001) for more information on selection of tuning parameters.

3. A SIMULATION STUDY

In this section, we present a simulation study to examine the performance of the CL1CP under various conditions with the OLS, stepwise, L1CP, LASSO and ENET. The methods are examined under different cases of medium, high and ultrahigh ($p > n$) dimensional settings. The true underlying regression model from which we simulate data is given by

$$Y = X^T\beta + \sigma\epsilon, \qquad \epsilon \sim N(0,1). \tag{11}$$

Each simulated data consists of a training set for fitting the model, a validation set for selecting the tuning parameters, and a test set on which the test errors are computed for evaluation of performance. The notation ·/·/· represents the number of observations in the training, validation, and test datasets, respectively. For

example, 100/200/300 indicates that the training, validation, and test datasets contain 100, 200, and 300 observations, respectively. In these simulation experiments, the smoothing parameters λ_1 and λ_2 were selected over a grid of 100 equally spaced values, ranging from 10^{-2} to $10^{2.2}$ for each parameter using the validation datasets.

3.1 Simulation Setting

The methods under consideration are examined under five simulation settings. The simulation settings represent different data scenarios and are similar to those used by Tutz and Ulbricht (2006) and Zou and Hastie (2005).

Setting 1: We simulated 100 data sets consisting of $n/10n/200$ observations and 8 predictors. We set $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$, $n \in \{50, 100\}$ and $\sigma = 2$. The pairwise correlation between X_i and X_j was set to be $\rho(i,j) = \theta^{|i-j|}$ for all i, j, where $\theta \in \{0.5, 0.99\}$. This setting represents a sparse situation.

Setting 2: This setting is similar to that of Setting 1 except that $\beta_j = 0.85$, for all j.

Setting 3: In this instance, the simulated data sets consist of $n/10n/200$ observations and 40 predictors and we set $\beta = (\underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10})$, $n \in \{50, 100\}$, $\sigma = 8$ and $\rho(i,j) = 0.5^{|i-j|}$ for all i, j. In this setting there are 40 sparse grouped predictors with only 20 being relevant.

Setting 4: In this setting, the simulated data sets consisting of $n/10n/200$ observations and 40 predictors and we set $\beta = (\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{25})$, $n \in \{50, 100\}$ and $\sigma = 15$. The predictors X are generated as follows:

$X_i = Z_1 + w_i^x$, $Z_1 \sim N(0,1)$, $i = 1, \ldots, 5$,
$X_i = Z_2 + w_i^x$, $Z_2 \sim N(0,1)$, $i = 6, \ldots, 10$,
$X_i = Z_3 + w_i^x$, $Z_3 \sim N(0,1)$, $i = 11, \ldots, 15$.

X_i are independent identically distributed (iid) $N(0,1)$, for $i = 16, \ldots, 40$ and w_i^x are iid $N(0, 0.01)$. This setting implies there are three equally important groups with each containing 5 members.

Setting 5: In this situation, the simulated data sets consist of $n/10n/100$ observations and 200 predictors and we set $\beta = (\underbrace{5, \ldots, 5}_{20}, \underbrace{0, \ldots, 0}_{180})$, $n = 100$, $\sigma = 12$ and $\rho(i,j) = 0.5^{|i-j|}$ for all i, j.

3.2 Simulation Results

The performance of the methods are evaluated over 100 replications of each setting discussed above. The evaluation criteria are: prediction mean-squared

errors (MSE_Y) defined as $\frac{1}{n_{test}}\left\|Y_{test} - X_{test}^T\hat{\beta}\right\|^2$; mean-squared errors of estimates (MSE_β) defined as $\left\|\hat{\beta} - \beta\right\|^2$, size ($S$) which is the number of non-zero estimated regression coefficients; hits (C) which is the number of truly non-zero coefficients correctly estimated to be non-zero false positive (IC) which is the number of truly zero coefficients incorrectly estimated to be non-zero. For each method and simulation setting, each of the evaluation criteria was computed over 100 replications. Tables 1-5 summarizes the medians of MSE_Y, MSE_β, S, C and IC, while Figures 1-5 gives graphical representation of the results. The acronym TS stands for the true size of the model, that is, the number of important variables in the model.

The simulation results for Setting 1 (sparse situation) at $\theta = 0.5$ and $\theta = 0.99$ are presented in Table 1 and summarized in Figure 1 ($\theta = 0.5$, $n = 100$). The results show that OLS has the highest estimation and prediction error followed by Stepwise under all scenarios in Setting 1. When the correlations among the predictors are relatively low ($\theta = 0.5$), the CL1CP has the best performance in terms of prediction at both sample sizes considered. The performance of the ENET in terms of estimation at $n = 50$ is the best while that of CL1CP is the best at $n = 100$. The L1CP outperforms the OLS, Stepwise and LASSO in terms of prediction and estimation. In terms of variable selection, the CL1CP, ENET and Stepwise have the best performance while the L1CP and LASSO have similar performance. At high level of multicollinearity ($\theta = 0.99$) and $n = 50$, the L1CP and CL1CP have similar performance with respect to prediction but the L1CP seem to be slightly better than CL1CP in terms of estimation, while the LASSO have the worst performance in terms of prediction and estimation out of the penalized methods. Also, the CL1CP seem to be better than L1CP for variable selection at $\theta = 0.99$ and $n = 50$. Furthermore, at $\theta = 0.99$ and $n = 100$, the CL1CP is slightly better than L1CP in terms of prediction while for estimation L1CP seem to be slightly better. The LASSO has the best variable selection performance at $\theta = 0.99$ and $n = 50$ wth the L1CP and ENET having similar performances.

Table 1: Medians of mean squared errors of estimation and prediction (MSE_β and MSE_Y); median absolute average errors of estimation (AE); median estimated model sizes (S), median Hits (C) and median FP (IC) for Setting 1 when $\theta = 0.5, 0.99$ based on 100 replications. TS stands for the true size of the model.

θ	n	*Method*	MSE_β	MSE_Y	S $(TS = 3)$	C	IC
0.5	50	***OLS***	0.88	4.77	-	-	-
		Stepwise	0.62	4.54	4	3	1
		LASSO	0.45	4.47	6	3	3
		ENET	0.28	4.26	4	3	1
		L1CP	0.44	4.43	6	3	3
		CL1CP	0.29	4.25	4	3	1
	100	***OLS***	0.43	4.33	-	-	-
		Stepwise	0.27	4.26	4	3	1
		LASSO	0.23	4.22	6	3	3
		ENET	0.14	4.14	4	3	1
		L1CP	0.22	4.22	6	3	3
		CL1CP	0.13	4.14	4	3	1
0.99	50	***OLS***	48.91	4.75	-	-	-
		Stepwise	35.44	4.59	3	1	1
		LASSO	9.65	4.19	4	2	2
		ENET	5.92	4.09	7	3	4
		L1CP	5.5	4.09	7	3	4
		CL1CP	5.6	4.09	6	3	3
	100	***OLS***	23.44	4.33	-	-	-
		Stepwise	19.89	4.31	3	2	1
		LASSO	6.38	4.16	4	3	2
		ENET	4.64	4.13	6	3	3.5
		L1CP	4.65	4.15	6	3	4
		CL1CP	4.71	4.11	7	3	4

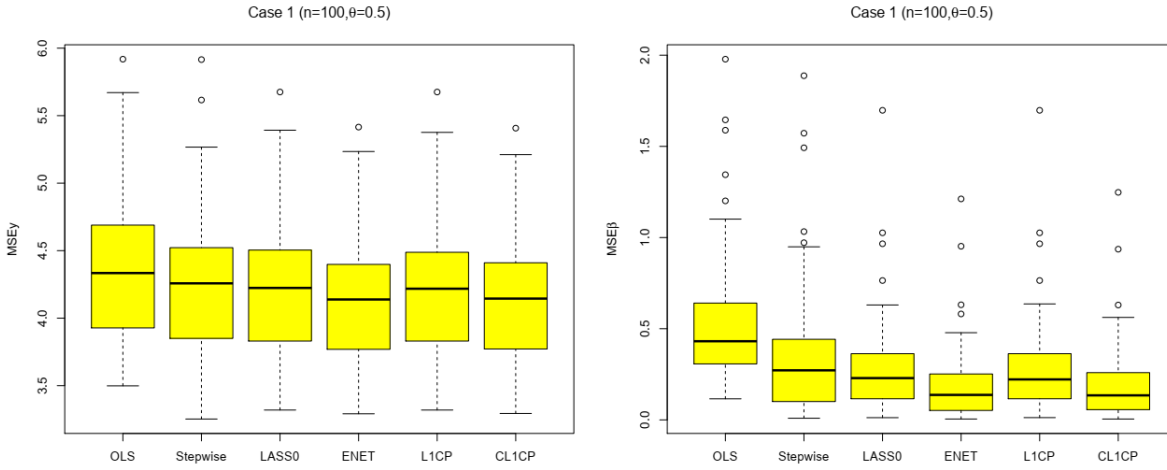


Figure 1: Boxplots of MSE_{β} and MSE_{Y} over 100 replications for Setting 1 at $n = 100$ and $\theta = 0.5$.

Table 2 and Figure 2 (for $\theta = 0.5$, $n = 100$) summarize the simulation results for Setting 2 (dense scenario). When the correlation among the predictors is relatively low ($\theta = 0.5$) and $n = 50$, LASSO has the worst performance in terms of estimation followed by OLS while the L1CP has the best performance in terms of estimation and prediction. The Stepwise has the worst performance at $n = 50$ in terms of prediction while CL1CP and the ENET have similar performances with respect to all the criteria in Setting 2. At $\theta = 0.5$ and $n = 100$, LASSO has the worst performance while in terms of variable selection, all the penalized methods perform equally. Furthermore, under Setting 2 when there is high correlation among the predictors ($\theta = 0.99$), the L1CP and CL1CP have similar performances in terms of variable selection at both sample sizes considered while Stepwise has poorest performance followed by the LASSO.

Table 2: Medians of mean squared errors of estimation and prediction (MSE_β and MSE_Y); median absolute average errors of estimation (AE); median estimated model sizes (S), median Hits (C) and median FP (IC) for SETTING 2 when $\theta = 0.5, 0.99$ based on 100 replications.

θ	n	*Method*	MSE_β	MSE_Y	S $(TS = 3)$	C	IC
0.5	50	*OLS*	0.86	4.49	-	-	-
		Stepwise	1.39	4.77	7	7	0
		LASSO	0.88	4.46	8	8	0
		ENET	0.69	4.36	8	8	0
		L1CP	0.45	4.24	8	8	0
		CL1CP	0.69	4.34	8	8	0
	100	*OLS*	0.44	4.21	-	-	-
		Stepwise	0.45	4.26	8	8	0
		LASSO	0.43	4.30	8	8	0
		ENET	0.33	4.26	8	8	0
		L1CP	0.25	4.25	8	8	0
		CL1CP	0.34	4.26	8	8	0
0.99	50	*OLS*	44.89	4.49	-	-	-
		Stepwise	21.10	4.36	2	2	0
		LASSO	10.70	4.18	4	4	0
		ENET	0.09	3.93	5	5	0
		L1CP	0.07	3.96	8	8	0
		CL1CP	0.30	3.94	8	8	0
	100	*OLS*	23.58	4.21	-	-	-
		Stepwise	18.02	4.16	3	3	0
		LASSO	7.82	4.16	5	5	0
		ENET	0.05	4.07	8	8	0
		L1CP	0.03	4.03	8	8	0
		CL1CP	0.22	4.07	8	8	0

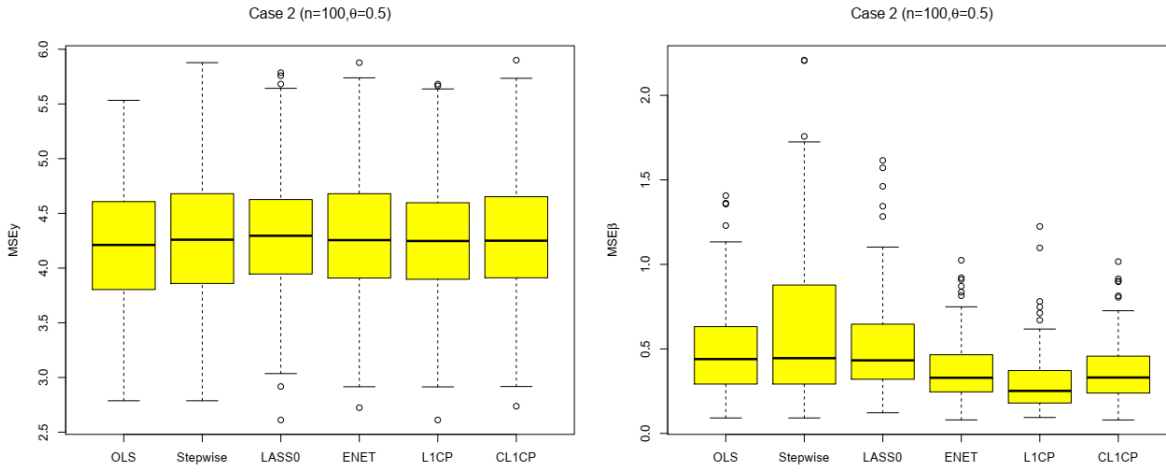


Figure 2: Boxplots of MSE_{β} and MSE_{Y} over 100 replications for Setting 2 at $n = 100$ and $\theta = 0.5$.

In terms of estimation and prediction, OLS and Stepwise perform very poorly while L1CP has the best performance in terms of estimation. At $n = 50$, the CL1CP is slightly better than L1CP in terms of prediction while at $n = 100$ the L1CP is the best. One can also observe that the performances of CL1CP and ENET are similar in terms of prediction but the LASSO is the poorest based on all criteria. It can also be noted that both the L1CP and CL1CP outperform the ENET in terms of variable selection at $n = 50$, an indication that having small sample size may affect the performance of the ENET and LASSO when the predictors are highly collinear. Under Setting 2, the L1CP mostly outperforms the CL1CP in terms of prediction and estimation, this is in line with the results of Zou and Hastie (2005) where the Naïve ENET outperforms the ENET under a similar setting like this.

From results for Settings 1 and 2, it appears that when there is high correlations among the predictors, the extra shrinkage incurred by the correlation based penalties is needed and should not be undone. Therefore, the naïve ENET and L1CP methods are better in terms of estimation and prediction when the correlations among the predictors are high and when all the predictors included in the model are relevant. However, the naïve methods are conservative and usually do not select sparse models compared to our proposed CL1CP as will see in the results from the remaining examples.

According to the Table 3 which presents the results for Setting 3 (equally grouped sparse variables) and Figure 3 (for $n = 100$), the CL1CP outperforms all the methods based on all the criteria with the ENET following closely and OLS and Stepwise having the worst performance. Of the penalized methods, the LASSO has the worst performance in terms of estimation and prediction when $n = 50$ while L1CP is the worst when $n = 100$. In terms of variable selection, the LASSO is better than the L1CP. The results here show that scaling the L1CP increases its performance in terms of estimation, prediction and ability to do "grouped variables" selection.

In Settings 4 and 5, the OLS and Stepwise methods are not included since they always behave poorly as in the previous settings and are not applicable in Setting 5 where $p > n$. The results for Setting 4 presented in Table 4 and summarized in Figure 4 ($n = 100$ only) reveal that L1CP outperforms all the other methods in terms of all the evaluation criteria at $n = 50$ followed by the ENET while the ENET has the best performance in terms of estimation and prediction at $n = 100$ followed closely by CL1CP. The LASSO performs poorly in terms of prediction, estimation and variable selection omitting relevant variables while the L1CP selects the highest number of irrelevant variables. The performance of the CL1CP here is significantly better than that of L1CP, this further shows that scaling the L1CP estimator improves its ability to do group selection.

The results for Setting 5 ($p > n$ situation) are summarized in Table 5 and Figure 5. According to the results, the ENET has the best performance in terms of estimation, prediction and variable selection followed by the LASSO and then CL1CP but in terms of variable selection. The results also show that the performance of the CL1CP is significantly better than that of the L1CP in terms of the criteria. The L1CP has poor performance in terms of estimation and prediction and selects a ridiculously large number of irrelevant variables compared to the CL1CP.

In general, the proposed CL1CP outperform the L1CP in almost all situations considered. The performance of the CL1CP does not strongly differ where the L1CP is better. The L1CP always select sparser models compared to the L1CP without omitting the relevant variables. Also, we find that the CL1CP outperforms the ENET and LASSO at times especially under situations where there are grouped variables and competes favourably in other situations.

Table 3: Medians of mean squared errors of estimation and prediction (MSE_β and MSE_Y); median absolute average errors of estimation (AE); median estimated model sizes (S), median Hits (C) and median FP (IC) for SETTING 3 based on 100 replications.

n	*Method*	MSE_β	MSE_Y	S $(TS = 20)$	C	IC
50	***OLS***	5.29	7.36	-	-	-
	Stepwise	5.11	7.18	31	20	11
	LASSO	3.33	6.15	29	20	9
	ENET	2.41	5.47	23	20	3
	L1CP	3.09	6.11	31	20	11
	CL1CP	2.42	5.50	23	20	3
100	***OLS***	2.67	5.58	-	-	-
	Stepwise	2.26	5.34	25	20	5
	LASSO	1.58	5.17	28	20	8
	ENET	1.22	4.90	23	20	3
	L1CP	1.59	5.23	30	20	10
	CL1CP	1.21	4.90	23	20	3

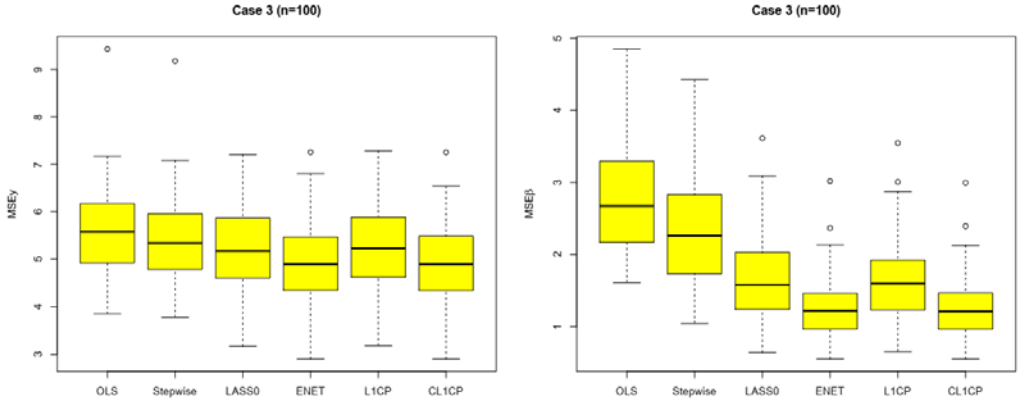


Figure 3: Boxplots of MSE_β and MSE_Y over 100 replications for Setting 3 at $n = 100$.

Table 4: Medians of mean squared errors of estimation and prediction (MSE_β and MSE_Y); median absolute average errors of estimation (AE); median estimated model sizes (S), median Hits (C) and median FP (IC) for SETTING 4 based on 100 replications.

n	*Method*	MSE_β	MSE_Y	S $(TS = 15)$	C	IC
50	***LASSO***	385.67	288.97	10	3	7
	ENET	13.70	264.02	18	15	3
	L1CP	28.48	289.77	25	15	10
	CL1CP	11.95	258.58	18	15	3
100	***LASSO***	420.18	250.61	10	3	7
	ENET	2.74	235.40	17	15	2
	L1CP	9.95	249.53	23	15	8
	CL1CP	3.28	236.12	17	15	2

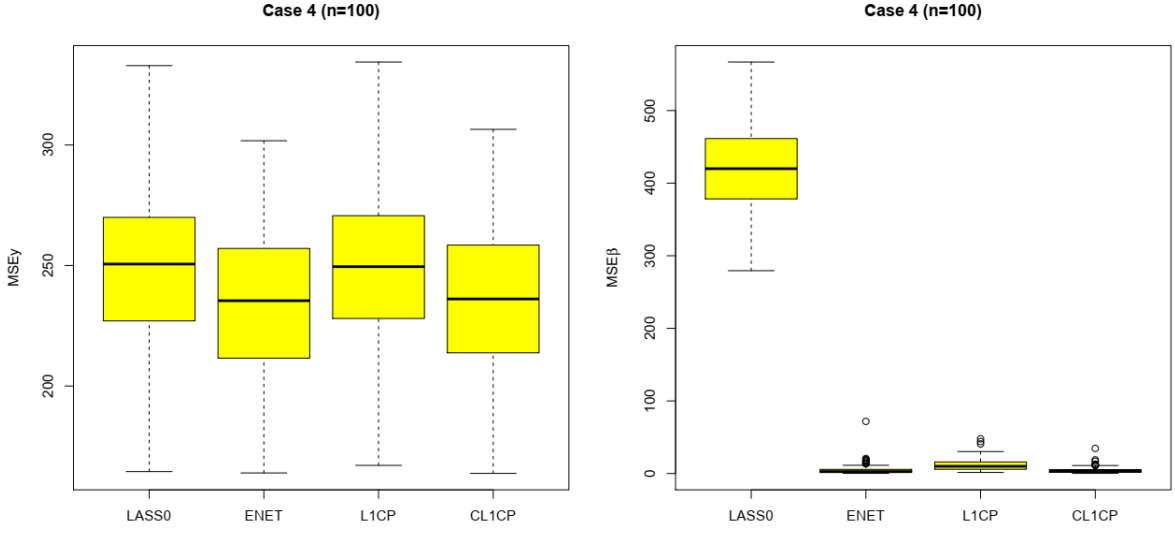


Figure 4: Boxplots of MSE_β and MSE_Y over 100 replications for Setting 4 at $n = 100$.

Table 5: Medians of mean squared errors of estimation and prediction (MSE_β and MSE_Y); median absolute average errors of estimation (AE); median estimated model sizes (S), median Hits (C) and median FP (IC) for SETTING 5 based on 100 replications at $n = 100$.

Method	MSE_β	MSE_Y	S $(TS = 15)$	C	IC
LASSO	105.04	246.57	43	20	23
ENET	68.94	216.65	42	20	22
L1CP	260.57	824.63	161	20	141
CL1CP	148.98	327.36	24.5	20	5

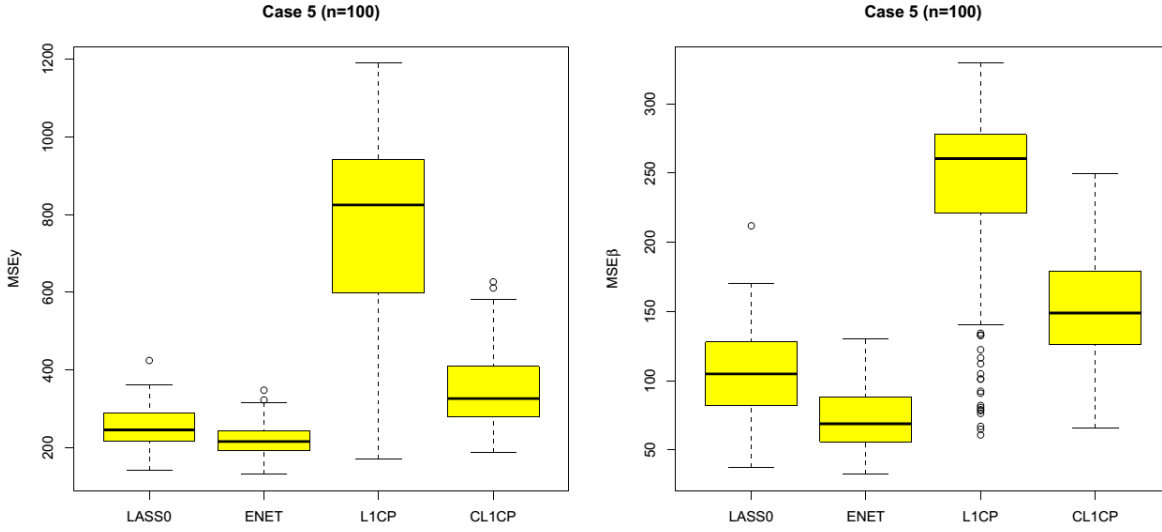


Figure 4: Boxplots of MSE_β and MSE_Y over 100 replications for Setting 5.

4. APPLICATION TO REAL LIFE DATASETS

In this section, applications of the methods on two real life datasets are considered. The first dataset is the prostate dataset used in Zou and Hastie (2005) which comes from a study of prostate cancer by Stamey et al. (1989) involving

97 men. The data consist of a response variable which is the log of prostate specific antigen (lpsa) and eight predictors: log cancer volume (lcavol), log prostate weight (lweight), age, log of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), log capsular penetration (lcp), Gleason score (gleason) and percentage Gleason score 4 or 5 (pgg45).

The second dataset is the gene expression data from the microarray experiments on 120 mammalian eye tissue samples by Scheetz et al. (2006). The dataset consist of 200 predictors which represents 200 gene probes of 120 rats. The response variable is the expression level of TRIM32 gene. In contrast to the first dataset, the dimension of the gene expression data is very high with the sample size ($n = 120$) less than the number of predictors ($p = 200$). The OLS, Stepwise selection method, LASSO, ENET, L1CP and the CL1CP methods were all applied to the prostate dataset while only the penalized methods were applied to the gene expression dataset.

In the two applications presented in this section, the selection of the smoothing parameters λ_1 and λ_2 was performed using a grid search approach. Specifically, we considered a grid of 100 equally spaced values for each parameter, ranging from 10^{-2} to $10^{2.2}$. On a Windows 64-bit system with 8 GB RAM and an Intel Core i5-8365U CPU (1.60 GHz, 1.90 GHz turbo), the typical runtime for 10-fold cross-validation was 6 seconds for the prostate cancer dataset and 82 seconds for the eye tissue dataset. The Eye tissue dataset required more computational time than the prostate cancer dataset due to its higher dimensionality and larger sample size, which increased the complexity of the quadratic optimization and expanded the active set of variables during cross-validation, resulting in longer processing times. These results indicate that the validation process remains computationally efficient, making the proposed method practical for real-world applications, even in high-dimensional settings.

Firstly, the prostate cancer data were randomly split into a training set with 50 observations, and a test set with 47 observations. The training dataset were used for model fitting and selection of tuning parameters by10-foldcross validation. The performance of the methods were then compared based on their prediction mean squared error (MSEy) on the test dataset and the number of non-zero coefficients. For the gene expression dataset, the training set consists of 60 observations while the test set consists of 60 observations likewise. The process of data splitting, model fitting and computation of MSEy were repeated 100 times. The results for both datasets are summarized in Table 6.

Table 6 obviously shows that the non-corrected versions of the correlation based method (L1CP) selects denser models compared to the corresponding

corrected version (CL1CP) with no substantial gain in prediction accuracy in both datasets. The LASSO has the best performance in both datasets. For the prostate cancer data, the CL1CP selected the fewest number of predictors, which is same as the number of predictors selected by the stepwise method, and has prediction accuracy not significantly different from that of the L1CP which selects 3 variables more. Out of the penalized methods, the L1CP selects the highest number of predictors (176) with no significant gain in prediction accuracy.

In the case of the gene expression dataset, which is of ultrahigh dimension, the CL1CP outperforms all the methods in terms of sparsity with no significant loss in prediction accuracy. The L1CP produced a prediction error of about 0.007, albeit with median number of predictors of 145.5 which is almost six times that of the CL1CP (23.5 predictors) with prediction error of 0.008. So, the L1CP performs poorly in variable selection in cases of $p > n$ compared to CL1CP. Also, the CL1CP outperforms the ENET in terms of sparsity though they have the same prediction accuracy. The results from this section further show that the scaling the correlation based method (L1CP) improve its performances in terms of variable selection and can make it outperform the ENET.

Table 6: Median mean squared errors of prediction (MSE_Y) and median estimated model sizes (S), based on 100 replications.

Method	Prostate data		Eye tissue data	
	MSE_Y	S	MSE_Y	S
OLS	0.613	8	-	-
Stepwise	0.623	4	-	-
LASSO	0.594	5	0.008	23
ENET	0.597	5	0.008	31
L1CP	0.603	7	0.007	176
CL1CP	0.661	4	0.008	24

5. CONCLUDING REMARKS

In this paper, we have proposed a scaled version of the correlation based penalized elastic net penalty for carrying out variable selection and regression simultaneously. The motivation behind our idea is from Zou and Hastie (2005) where decomposition of the ridge operator was used to justify the need for

rescaling in other to ameliorate the challenges that are associated with double shrinkage through the use of the ridge and LASSO penalties. Following the ideas of Zou and Hastie (2005), we derive the scales by a decomposition of the correlation based estimators. We call the proposed scaled or corrected method CL1CP while the corresponding original version is referred to as the naïve L1CP. The efficient and robust worst-case quadratic solver method (Grandvalet et al., 2017) was adopted for estimation.

We have demonstrated via simulations and real life dataset that the proposed method compared to the naïve versions is better in terms of variable selection while retaining the ability to handle grouping effects and produce good prediction accuracy. The results also reveal that the scaled version of the correlation based method outperforms the LASSO and elastic-net in some situations especially in terms of variable selection in the presence of group effect. Generally, the new CL1CP proposed method possesses better performance in terms of correct identification of relevant variables which is vital in high dimensional regression problems. Finally, the new rescaled correlation based penalized regression method proposed in this work may attract wider applications in data mining and big data analytics.

A key advantage of the proposed methodology is its favourable computational efficiency. By leveraging an iterative quadratic solver that focuses only on the active set of variables, the algorithm significantly reduces the dimensionality of the optimization problem at each step. This selective approach results in a smaller computational cost compared to traditional sparse regression techniques such as coordinate descent or proximal methods.

The method proposed here may be extended to generalized linear models such as the binary logistic regression, Poisson regression, COM-Poisson regression (Shmueli et al., 2005; Adeniyi et al., 2019), and structured modelling (Ezenweke et al., 2023) settings by adopting a penalized likelihood approach.

Acknowledgments

We are grateful to two anonymous reviewers whose helpful reviews and suggestions improved the quality of this paper.

Source of Finance

During this study, no financial support was received neither from any individual or organization.

Conflict of Interest

There is no conflicts of interest declared by the authors.

REFERENCES

Adeniyi, I.A., Shobanke, D. A.,Edogbanya H. O. (2019). Re-parameterization of the COM-Poisson distribution using spectral algorithms. *Pakistan Journal of Statistics and Operation Research*, 15(3), 701-712.

Algamal, Z. Y., Lee, M. H. (2015). Penalized Poisson regression model using adaptive modified elastic net penalty. *Electronic Journal of Applied Statistical Analysis*, 8(2), 236-245.

Anbari, M.E., Mkhadri, A. (2014). Penalized regression combining the L_1 norm and a correlation based penalty**.** *Sankhya* B, 76 (1), 82–102.

Beck, A., Teboulle, M. (2009). Fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2,183–202.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statist.* 24, 2350–2383.

Bondell, H. D., Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR. *Biometrics* 64, 115 − 123.

Efron, B.,Tibshirani, R. (1993). *An Introduction to Bootstrap*. Chapman and Hall, London.

Efron, B., Tibshirani, R., (1997). Improvements on cross-validation. The 0.632+ bootstrap method. *Journal of the American Statistical Association*, 2, 548–560.

Ezenweke, C. P., Adeniyi, I. A., Yahya, W. B., Onoja, R. E. (2023). Determinants and spatial patterns of anaemia and haemoglobin concentration among pregnant women in Nigeria using structured additive regression models. *Spatial and Spatio-Temporal Epidemiology*, *45*, 100578. https://doi.org/10.1016/j.sste.2023.100578

Fan, J. Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In *International Congress of Mathematicians*, **3,** 595–622.

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3), 397–416.

Grandvalet, Y., Chiquet, J., Ambroise, C. (2017). Sparsity by worst-case quadratic penalties. *arXiv preprint arXiv:1210.2077v2 [stat.ML]*.

Hapfelmeier, A., Yahya, W.B., Rosenberg, R., Ulm, K. (2012). Predictive modelling of gene dxpression Data. In J. Crowley and A. Hoering (Eds), *Handbook of Statistics in Clinical Oncology*, 3rd ed. (pp. 463-475). Chapman and Hall/CRC, New York.

Hastie, T., Tibshirani, R., Friedman, J. (2001). *The Elements of Statistical Learning*, Springer.

Hoerl, A.,Kennard, R. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12, 55 − 67.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection, *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1137–1145.

Ryan, T. (2009). *Modern Regression Methods (Second edition)*. John Wiley & Sons, Hoboken, NJ.

Scheetz, T. E., Kim, K. Y., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Sheffield, V. C., Stone, E. M. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(39), 14429–14434.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., Boatwright, P. (2005). A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. *Applied Statistics*, 54, 127–142.

Tan, Q. E. A. (2012). *Correlation Adjusted Penalization in Regression Analysis.* PhD. Thesis, The University of Manitoba, Canada. Retrieved from http://mspace.lib.umanitoba.ca/handle/1993/9147.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of*
the Royal statistical Society, Series B. 58, 267 − 288.

Tutz, g., Ulbricht, j. (2009). Penalized regression with correlation based penatly. *Statistics and. Computing.*, **19**, 239–253.

Wang, X., Dunson, D.B., Leng, C. (2016). No penalty no tears: Least squares in high-dimensional linear models. *Proceedings of Machine Learning Research,* 48, 1814-1822.

Yahya, W.B., Aremu, G.T. Garba, M.K. (2015). Multiclass sequential feature selection and classification method for genomic data. *Journal of Applied Science and Technology, 20* (1&2): 50-60.

Yahya, W.B., Rosenberg, R., Ulm, K. (2014). Microarray-based classification of histopathologic responses of locally advanced rectal carcinomas to neoadjuvant radiochemiotherapy treatment. *Türkiye Klinikleri Journal of Biostatistics, 6*(1), 8-23.

Yahya, W.B., Ulm, K., Fahrmeir, L., Hapfelmeier, A. (2011). k-SS: a Sequential feature selection and prediction method in microarray study. *International Journal of Artificial Intelligence*, 6(S11), 19-47.

Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal statistical Society, Series B*. 67, 301 − 320.