FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 9 No. 1, January, 2025, pp 29 - 40
DOI: https://doi.org/10.33003/fjs-2025-0901-2774

REGRESSION ESTIMATION AND FEATURE SELECTION USING MODIFIED CORRELATION-ADJUSTED ELASTIC NET PENALTIES

***[1]Olayiwola Babarinsa, [1]Helen Edogbanya, [2]Ovye Abari and [3]Isaac Adeniyi**

[1]Department of Mathematics, Federal University Lokoja, P.M.B. 1154, Lokoja, Nigeria
[2]Department of Computer Science, Federal University Lokoja, P.M.B. 1154, Lokoja, Nigeria
[3]Department of Statistics, Federal University Lokoja, P.M.B. 1154, Lokoja, Nigeria

*Corresponding authors' email: modibbojrm@gmail.com; modibbomahmud@nsuk.edu.ng

ABSTRACT

Regularized regression techniques such as the least absolute shrinkage and selection operator (LASSO), elastic-net, and the type 1 and type 2 correlation adjusted elastic-net (CAEN1 and CAEN2 respectively) are used for simultaneously carrying out variable selection and estimation of coefficients in machine learning. Modified estimators based on the CAEN1 and CAEN2 are proposed in this study by rescaling the estimates to undo the double shrinkage incurred due to the application of two penalties. The scale factors are derived by decomposing the correlation matrix of the predictors. The derived scale factors, which depend on the magnitude of correlations among the predictors, ensure that the elastic-net is included as a special case. Estimation is carried out using a robust worst-case quadratic solver algorithm. Simulations show that the proposed estimators referred to as corrected correlation adjusted elastic-net (CCAEN1 and CCAEN2) perform competitively with the CAEN1, CAEN2, LASSO, and elastic-net in terms of variable selection, estimation and prediction accuracy with CCAEN1 yielding the best results when the number of predictors is more than the number of observations and CCAEN2 producing the best performance when there is grouping effect, where highly correlated predictors tend to be included in or excluded from the model together. Applications to two real-life datasets further demonstrate the advantage of the proposed methods for machine learning.

Keywords: Variable selection, Regularization, High-dimensional data, Grouping effect, LASSO, Machine Learning

INTRODUCTION

Regression analysis is a popular tool for building predictive models and studying the influence of some predictor variables on a continuous response variable (Biecek & Burzykowski, 2021). Let $Y = (y_1, \dots, y_n)^T$ and $X = [X_1, \dots, X_p]^T$ denote the response vector and matrix of predictors (or features) respectively, where p is the number of predictor variables. In a regression analysis, the relationship between Y and $X_1, X_2, \cdots, X_p$ is represented by

$$Y = X^T\beta + \epsilon, \qquad (1)$$

where $\beta = (\beta_1, \dots, \beta_p)^T$ with β_j, $j = 1, \dots, p$ is the regression coefficient representing the effect of a predictor variable X_j on Y.

The ordinary least squares (OLS) technique, a classical technique for estimating the regression coefficients, can perform poorly when the predictors are highly correlated and is not applicable when the number of predictors p exceeds the sample size n, see (Hoerl & Kennard, 1970; Ryan, 2008; Wang, Dunson, & Leng, 2016). However, technological breakthroughs in medicine, artificial intelligence, machine learning and other areas have given rise to large data situations where p is greater than n, see (Garba, Yahya, & Aremu, 2016; Hapfelmeier, Babatunde, Yahya, & Ulm, 2012) and the references therein.

One important aspect of regression modelling is variable selection especially when dealing with high-dimensional data (Fan & Li, 2006). Correctly selecting important predictors determines the performance of the fitted model. The best-subset and stepwise model selection methods are popular for variable selection. However, it is computationally impractical to use the best-subset selection method when the number of predictors is large and when $p > n$ the best-subset method is limited to only models having the number of predictors less than n (Hanke, Dijkstra, Foraita, & Didelez, 2024). Moreover, Breiman (1996) reported that the subset selection can be unstable while the performance of the step-wise methods can result to a model with poor predictive performance. Also, when predictors are highly correlated, estimates of β by the OLS are unstable. The ridge regression was introduced to circumvent the problem encountered by the OLS when dealing with collinear predictors (Hoerl & Kennard, 1970). The ridge regression involves imposing an l_2-norm penalty on the OLS objective function. The ridge estimator is obtained by solving the $l_2$regularized least squares problem. Thus

$$\beta_{ridge} = arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \qquad (2)$$

where $\|\beta\|_2^2 = \sum_{j=1}^{p}\beta_j^2$ is the l_2-norm of β. The ridge regression tries to handle the problems caused by collinear predictors in the model but does not perform variable selection.

A popular method proposed to perform variable selection and estimation of regression coefficients simultaneously is the least absolute shrinkage and selection operator (LASSO) technique (Tibshirani, 1996). This technique handles some of the problems associated with subset selection, stepwise selection and the OLS especially when $p \gg n$ and when the predictors are collinear. The LASSO estimator is obtained by solving the l_1 regularized least squares problem:

$$\hat{\beta}_{lasso} = arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1^1, \qquad (3)$$

where $\|\beta\|_1^1 = \sum_{j=1}^{p}|\beta_j|$ is the l_1-norm of β. Although the LASSO possesses some good statistical properties and has been found to provide good results in many situations, it has some drawbacks which include its inability to select more than n predictors in the $p > n$ situation; propensity to select only one variable from a group of highly correlated predictors and poor performance compared to ridge regression in the $n > p$ situation. To overcome these problems of LASSO, Zou and Hastie (2005) proposed a technique which combines

the l_1 penalty of the LASSO and the l_2 penalty of the ridge termed the elastic-net (ENET). The elastic-net estimator $\hat{\beta}_{naive-enet}$ is obtained by solving

$$\hat{\beta}_{naive-enet} = arg\min_{\beta}\|Y - X\beta\|_2^2 + \lambda_1\|\beta\|_1^1 + \lambda_2\|\beta\|_2^2. \quad (4)$$

Zou and Hastie (2005) referred to the estimator obtained from equation (4) as the naive elastic-net estimator and showed that its performance is not always satisfactory, nor does it possess the minimax optimal property because of the double shrinkage due to the simultaneous use of the l_1 and l_2 penalties which introduce unnecessary extra bias. The authors derived the elastic-net (ENET) estimator as $\left(1 + \frac{\lambda_2}{n}\right)\hat{\beta}_{naive-enet}$ or $(1 + \lambda_2)\hat{\beta}_{naive-enet}$ if the predictors are standardized (each variable has a mean zero and l_2-norm one). Nonetheless, Anbari and Mkhadri (2014) observed that the ENET seems to be slightly less reliable if the correlation between variables is not so extreme. Besides, the ENET like the LASSO does not factor in the information concerning the correlation among the predictors during shrinkage.

The octagonal shrinkage and clustering algorithm for regression (OSCAR) is a regularized least squares with a penalty function that combines the l_1 and the pairwise l_∞ norms (Bondell & Reich, 2006). The OSCAR operates by constraining some coefficients to be identically equal, causing correlated predictors that have identical effects on the response to form groups represented with the same coefficients. However, obtaining the Oscar estimates for large p can be computationally burdensome (Anbari & Mkhadri, 2014). Other proposed penalties for simultaneous estimation and variable selection in regression problems include the Smoothly Clipped Absolute Deviation (SCAD) penalty and the Minimax Concave Penalty (MCP) (Fan & Li, 2001; Zhang, 2010).

In addition, Tutz and Ulbricht (2009) introduced a technique that combines a correlation-based penalty and a blockwise boosting (BB) procedure for performing shrinkage and variable selection. Determination of the appropriate step length factor and number of iterations for the BB technique can sometimes be problematic in practice and this affects the sparsity of the solution and the speed of the technique. Consequently, Anbari and Mkhadri (2014) introduced the L1CP estimator which combines the LASSO and the correlation-based penalty of Tutz and Ulbricht (2009).

Similarly, Tan (2012) proposed two correlation-adjusted elastic-net (namely CAEN1 and CAEN2) penalties for linear regression which were further extended to Poisson regression by Algamal (2015). The CAEN1 and CAEN2 estimators are the minimizers of

$$CAEN1 = \|Y - X\beta\|_2^2 + \lambda_1\|\beta\|_1^1 + \lambda_2\left[\sum_{j=1}^{p-1}\left[\left(\beta_j - r_{j,j+1}\beta_{j+1}\right)^2\right] + \beta_p^2\right], \quad (5)$$

and

$$CAEN2 = \|Y - X\beta\|_2^2 + \lambda_1\|\beta\|_1^1 + \lambda_2\left[\sum_{j=1}^{p-1}\sum_{k>j}\left[\left(\beta_j - r_{j,k}\beta_k\right)^2\right] + \beta_p^2\right], \quad (6)$$

respectively. The CAEN1 and CAEN2 performed better than the ENET for count data using the Poisson regression. However, the unnecessary bias that may be introduced due to double shrinkage by both the l_1 and correlation-adjusted penalties were not accounted for in the CAEN1 and CAEN2. Also, empirical studies on the finite sample performance of the CAEN1 and CAEN2 for linear regression are not available in the literature. More details can be found on how the CAEN1 and CAEN2 methods can be augmented to become LASSO-type problems (Tan, 2012). The LASSO-type problems can be solved by a quadratic solver algorithm. The quadratic solver showed that the algorithm is robust and highly computationally efficient compared to other algorithms for sparse regression such as proximal (Beck & Teboulle, 2009) and the coordinate descent methods (Fu, 1998).

Our goal in this study is to introduce scaled versions of the CAEN1 and CAEN2 methods. The main idea is to improve the performance of the CAEN1 and CAEN2 by scaling the estimates to undo the double shrinkage thereby reducing the bias of the estimates. In this paper, we call the estimators proposed by Tan (2012) the naïve CAEN1 and CAEN2 while we call our proposed estimators CCAEN1 and CCAEN2. We obtain the scaling transformation using ideas similar to those used by Zou and Hastie (2005) to correct the naïve ENET to give the ENET estimator.

All through this paper, the errors ϵ are assumed to be identically and independently distributed (IID) with zero mean and finite variance σ^2. The rest of the paper is organized as follows; In Section 2, the corrected correlation adjusted elastic net (CCAEN1 and CCAEN2) estimators are introduced alongside corresponding proposed scaling transformations. In Section 3, simulation studies are carried out to evaluate the finite sample performance of the proposed methods (CCAEN1 and CCAEN2) alongside the naive CAEN1 and CAEN2 as well as other competitors such as the LASSO and ENET. The applications of all the methods considered for real-life datasets are discussed in Section 4. Section 5 contains the concluding remarks.

MATERIALS AND METHODS

This section introduces scaled versions of the naive CAEN1 and CAEN2 methods - termed CCAEN1 and CCAEN2 respectively - and improve their performance by scaling the estimates to undo the double shrinkage thereby reducing the bias of the estimates. The scaling transformation is obtain using ideas similar to methods in Zou and Hastie (2005) to correct the naive ENET to give the ENET estimator. Firstly, the general form of the objective function to be minimized to yield the regression estimates is

$$\|Y - X\beta\|_2^2 + \lambda_1\|\beta\|_1^1 + \lambda_2 P(\beta) \quad (7)$$

Our proposed corrected estimators are of the form

$$S\left[arg\min_{\beta}\|Y - X\beta\|_2^2 + \lambda_1\|\beta\|_1^1 + \lambda_2 P(\beta)\right], \quad (8)$$

where S is the scaling factor. In the following section, we show the derivation of the scaling factors used in CCAEN1 and CCAEN2.

The Rescaled CAEN1 and CAEN2

Tan (2012) proposed two correlated adjusted elastic-net (CAEN) estimators. The first type CAEN which we call the naïve CAEN1 is obtained when $P(\beta)$ in (6) is set to

$$P(\beta) = \sum_{j=1}^{p-1}\left[\left(\beta_j - r_{j,j+1}\beta_{j+1}\right)^2\right] + \beta_p^2. \quad (9)$$

The second type (naïve CAEN2) is obtained when

$$P(\beta) = \sum_{j=1}^{p-1}\sum_{k>j}\left[\left(\beta_j - r_{j,k}\beta_k\right)^2\right] + \beta_p^2. \quad (10)$$

Tan (2012) showed that equation (9) and (10) can be written in the following simple quadratic form:

$P(\beta) = \beta^T W_1 \beta,$

and

$P(\beta) = \beta^T W_2 \beta,$

respectively, where, $W_1 = D_1^T D_1$, $W_2 = D_2^T D_2$ and

$$D_1 = \begin{pmatrix} 1 & -r_{1,2} & 0 & \dots & 0 & 0 \\ 0 & 1 & -r_{2,3} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & -r_{P-1,P} \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix},$$

and

$$D_2 = \begin{pmatrix} 1 & -r_{1,2} & 0 & 0 & \dots & 0 & 0 \\ 1 & 1 & -r_{1,3} & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & \dots & 0 & -r_{1,p} \\ 0 & 1 & -r_{2,3} & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & -r_{2,4} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 & \dots & 0 & -r_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & -r_{p-1,p} \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}.$$

The estimators proposed are two-stage procedures: the correlation-adjusted penalty regression coefficients based on CAEN1 or CAEN2 are first obtained for each fixed λ_2, and the lasso-type shrinkage is carried out along the lasso coefficient solution paths. To reduce a double amount of shrinkage, CCAEN1 and CCAEN2 estimators are introduced as follows

$$diag(\lambda_2 W_1 + I)\left[arg\min_{\beta}\|Y - \beta_0 - X\beta\|_2^2 + \lambda_1\|\beta\|_1^1 + \lambda_2\beta^T W_1\beta\right], \quad (11)$$

and

$$diag(\lambda_2 W_2 + I)\left[arg\min_{\beta}\|Y - \beta_0 - X\beta\|_2^2 + \lambda_1\|\beta\|_1^1 + \lambda_2\beta^T W_2\beta\right], \quad (12)$$

respectively, where I is a $p \times p$ identity matrix and $diag(A)$ is a diagonal matrix such that the diagonal elements are the same as the diagonal elements of A (Babarinsa et al., 2022). It is easy to see, for example, that the CCAEN1 estimator is given by equation (8) when S is set as

$$S = diag(\lambda_2 W_1 + I) \quad (13)$$

The scale factor for the CCAEN1 (S^{c1}) of a single β_j can be simplified to

$$S_j^{C1} = \begin{cases} 1 + \lambda_2, & j = 1 \\ 1 + \lambda_2\left(1 + r_{j,j+1}^2\right), & j = 2,3,\dots p \end{cases}. \quad (14)$$

Similarly, the scale factor for the CCAEN2 (S^{c2}) of a single β_j in the case of CCAEN2 can be simplified to

$$S_j^{C2} = \begin{cases} 1 + \lambda_2(p-1), & j = 1 \\ 1 + \lambda_2\left(p - j + \sum_{i=1}^{j-1} r_{i,j}^2\right), & j = 2,3,\dots p \end{cases}. \quad (15)$$

The motivation for adopting $diag(\lambda_2 W_s + I), s = 1,2$ as scale factors are similar to the argument used by (Zou & Hastie, 2005) to improve the performance of the ENET by rescaling the naïve ENET. Without the LASSO penalty, the correlation-adjusted penalty (CAP) regression estimator (Tan 2012) is given as

$$\hat{\beta}_{CAP} = (X^T X + \lambda_2 W_1)^{-1} X^T Y. \quad (16)$$

Now, consider a decomposition of the first type CAP estimator with X standardized, to obtain

$$X^T X = \begin{bmatrix} 1 & r_{1,2} & \dots & r_{1,p} \\ r_{1,2} & 1 & \dots & r_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1,p} & r_{2,p} & \dots & 1 \end{bmatrix}, \quad (17)$$

where $X = \begin{bmatrix} 1 & r_{1,2} & \dots & r_{1,p} \\ 0 & 1 & \dots & r_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$

and $W_1 = \left(w_{ij}\right)_{1 \le i,j \le p}$ is a real symmetric $p \times p$ matrix such that

$$w_{ij} = \begin{cases} 1, & if\ i = j = 1 \\ 1 + r_{i,i+1}^2, & if\ i = j = 2,3,\dots p \\ -r_{i,j} & if\ j = i + 1 \\ 0, & if\ otherwise \end{cases}. \quad (18)$$

Hence, the CAP estimator can be rewritten as

$$\hat{\beta}_{CAP} = R^{-1} Z^{-1} X^T Y, \quad (19)$$

where $R = \begin{bmatrix} S_1^{C1} & 0 & \dots & 0 \\ & S_2^{C1} & \dots & 0 \\ & & \ddots & \vdots \\ & & & S_p^{C1} \end{bmatrix}$. $Y = \left(y_{ij}\right)_{1 \le i,j \le p}$ and $Z = \left(z_{ij}\right)_{1 \le i,j \le p}$ is a $p \times p$ matrix such that

$$z_{ij} = \begin{cases} 1, & if\ i = j \\ \frac{r_{i,j}(1-\lambda_2)}{S_j^{C1}}, & if\ j = i + 1\ or\ i = j + 1 \\ \frac{r_{1,j}}{S_j^{C1}} & if\ otherwise \end{cases} \quad (20)$$

The matrix Z can further be expressed as

$$Z = \begin{bmatrix} 1 & \frac{r_{1,2}(1-\lambda_2)}{S_1^{C1}} & \frac{r_{1,3}}{S_1^{C1}} & \frac{r_{1,4}}{S_1^{C1}} & \dots & \frac{r_{1,p}}{S_1^{C1}} \\ \frac{r_{1,2}(1-\lambda_2)}{S_2^{C1}} & 1 & \frac{r_{2,3}(1-\lambda_2)}{S_2^{C1}} & \frac{r_{2,4}}{S_2^{C1}} & \dots & \frac{r_{2,p}}{S_2^{C1}} \\ \frac{r_{1,3}}{S_3^{C1}} & \frac{r_{2,3}(1-\lambda_2)}{S_3^{C1}} & 1 & \frac{r_{3,4}(1-\lambda_2)}{S_3^{C1}} & \dots & \frac{r_{3,p}}{S_3^{C1}} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{r_{1,p-1}}{S_{p-1}^{C1}} & \frac{r_{2,p-1}}{S_{p-1}^{C1}} & \frac{r_{3,p-1}}{S_{p-1}^{C1}} & \frac{r_{4,p-1}}{S_{p-1}^{C1}} & \dots & \frac{r_{p-1,p}(1-\lambda_2)}{S_{p-1}^{C1}} \\ \frac{r_{1,p}}{S_p^{C1}} & \frac{r_{2,p}}{S_p^{C1}} & \frac{r_{3,p}}{S_p^{C1}} & \frac{r_{4,p}}{S_p^{C1}} & \dots & 1 \end{bmatrix} \quad (21)$$

From (15), the CAP estimator involves scaling the correlations by a factor of $(1-\lambda_2)(S_1^{C1})^{-1}$ or $(S_1^{C1})^{-1}$ followed by direct shrinkage with R^{-1} leading to double the amount of shrinkage. When combining the CAP with the LASSO, the direct R^{-1} shrinkage step is not needed and is removed by the proposed scaling factor. In the proposed method, the shrinkage by LASSO is sufficient for controlling the variance, and obtaining sparsity and therefore undoing the $R^{-1} = [diag(\lambda_2 W_1 + I)]^{-1}$ shrinkage step by multiplying the naïve CAEN1 estimates by $diag(\lambda_2 W_1 + I)$.

Similar to the case of the naïve CAEN1, the naïve CAEN2 estimates incur double shrinkage by scaling the correlation and shrinking the estimates' direct scaling. We undo the direct scaling by rescaling the estimates using the proposed scaling factors given in equation (15) for the CAEN2. The scaling factors vary for each coefficient according to the correlation values and position of the corresponding covariate during estimation. It can be easily observed that CAEN1 and CAEN2 estimates change when the arrangement of the predictors is changed. This is also the case with CCAEN1 and CCAEN2 estimators.

It can be observed that the CCAEN1 and CCAEN2 estimators become the Elastic-Net estimator when W_1 and W_2 are equal to a $p \times p$ identity matrix respectively, whereas the CAEN1 and CAEN2 can only reduce to the naïve Elastic-Net estimator. This is another justification for the need to rescale the CAEN1 and CAEN2. Henceforth, we refer to the original formulations of the correlation-based methods as the "naïve correlation-based methods".

Estimation and Selection of Tuning Parameters (λ_1 And λ_2)

The proposed estimators preserve the properties of the CAEN1 and CAEN2 estimators. Since the CAEN can be

reduced to a LASSO-type problem, existing computational techniques for regularised regression methods can easily be adapted to obtain the CAEN1, CAEN2, CCAEN1, and CCAEN2 estimates.

To demonstrate the implementation of the proposed techniques and carry out the evaluation, the quadratic solver method proposed by Grandvalet, Chiquet, and Ambroise (2012) is adopted to obtain the regression coefficients for CAEN1, CAEN2, CCAEN1, CCAEN2, LASSO and elastic-net regression coefficients. This technique allows the inner minimization problem to be viewed as a simple unconstrained quadratic problem, and an optimization strategy based on the iterative resolution of small-size quadratic problems is used to obtain minimizers of the objective function. The R (R Core Team, 2021) package *quadrupen* implements this algorithm.

Selecting the tuning parameters λ_1 and λ_2 appropriately is very important in practice to achieve good prediction and estimation accuracy. Minimizing an out-of-sample prediction error can be used to determine the tuning parameters. Out-of-sample prediction error can be estimated using a validation dataset if available, otherwise, a k-fold cross-validation can be used, see (Efron & Tibshirani, 1997; Efron & Tibshirani, 1994; Kohavi, 1995). In this study, validation datasets are generated for selecting tuning parameters for simulation studies while ten-fold cross-validation (10-fold CV) is used to choose tuning parameters in applications on real datasets. The cross-validations (or validations) for determining CAEN1, CAEN2, CCAEN1, CCAEN2 and ENET are done on a two-dimensional surface because the methods involve two tuning parameters. We adopt the typical approach for carrying out validation or cross-validation by first choosing a grid of λ_2, and then for each λ_2 the entire solution path of the CAEN1, CAEN2, CCAEN1, CCAEN2 or ENET is produced by the quadratic solver algorithm. The chosen pair of λ_1 and λ_2 is the one that produces the least out-of-sample prediction error.

RESULTS AND DISCUSSION

Simulation Study

In this section, we discuss a simulation study carried out to examine the performance of the CCAEN1, and CCAEN2 under various conditions alongside CAEN1, CAEN2, 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

$$Y = X^T\beta + \epsilon,\ \epsilon \sim N(0, \sigma^2), \qquad (22)$$

where σ^2 is the error variance. 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 $\cdot/\cdot/\cdot$ is used to represent the number of observations in the training, validation and test set, respectively, for example, $100/200/300$ implies that there are 100, 200 and 300 observations in the training, validation and test datasets respectively.

Simulation Setting

The simulation settings for the six cases considered here are similar to those used in (Tutz & Ulbricht, 2009; Zou & Hastie, 2005).

Case 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 $\theta \in \{0.5, 0.99\}$.

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

Case 3: In this case, the generated data sets consist of $n/10n/200$ observations and 40 predictors and $\beta = (\underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10})$, $n \in \{100, 200\}$, $\sigma = 5$ and $\rho(i,j) = 0.5$ for all i, j.

Case 4: Each simulated dataset contains $n/10n/200$ observations and 40 predictors and we set $\beta = (\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{25})$, $n \in \{100, 200\}$ 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 and identically distributed (IID) $N(0,1)$, for $i = 16, \ldots, 40$ and w_i^x are iid $N(0, 0.01)$. This setting implies three equally important groups with each containing 5 members.

Case 5: In this case, the simulated datasets consist of $100/1000/200$ observations and 200 predictors and we set $\beta = (\underbrace{5, \ldots, 5}_{20}, \underbrace{0, \ldots, 0}_{180})$, $\sigma = 3$ and $\rho(i,j) = 0.5^{|i-j|}$ for all i, j. This represents a $p \gg n$ situation with 20 of the 200 predictors being relevant.

Case 6: This setting is similar to that of Case 1 except that $= (3, 1.5, 0, 0, 2, \underbrace{0, \ldots, 0}_{195})$, $\sigma = 3$ and only cases of $n = 100$ and $\theta = 0.5$ are considered. Here, $p \gg n$ but only 3 of the 200 predictors are relevant.

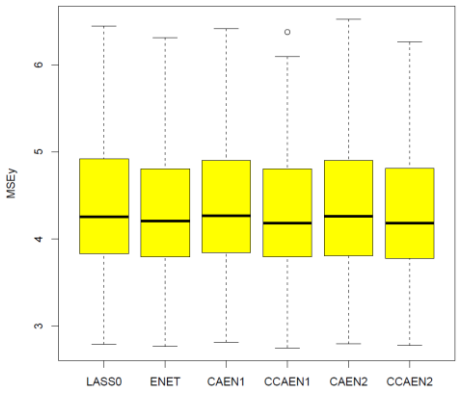
Simulation Results

The performance of the methods is evaluated over 100 replications of each setting discussed above. The evaluation criteria are: prediction mean-squared errors on the test data (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 $\left\| \hat{\beta} - \beta \right\|^2$ for assessing coefficients' estimation accuracy, size (S) which is the number of non-zero estimated regression coefficients; hits (TP) which is the number of truly non-zero coefficients correctly estimated to be non-zero , false positive (FP) which is the number of truly zero coefficients incorrectly estimated to be non-zero. For each method and simulation case, each of the evaluation criteria was computed over 100 replications. Tables 1-6 summarizes the medians of MSE_Y, MSE_β, S, TP and FP, while Figure 1-6 gives graphical representations of MSE_Y.

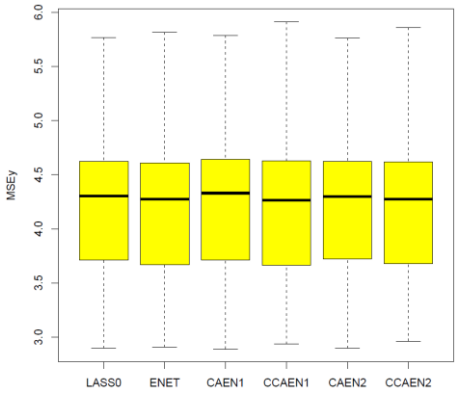
Table 1: Medians of mean squared errors of estimation and prediction (MSE_β and MSE_Y); median estimated model sizes (S), median Hits (TP) and median false positives (FP) for CASE 1 when $\theta = 0.5, 0.99$ based on 100 replications. TS stands for the true size of the model

θ	n	Method	MSE_β	MSE_Y	S ($TS = 3$)	TP	FP
0.5	50	*LASSO*	0.50	4.38	6	3	3
		ENET	0.36	4.27	4	3	1
		CAEN1	0.49	4.38	6	3	3
		CCAEN1	0.35	4.28	4	3	1
		CAEN2	0.50	4.39	6	3	3
		CCAEN2	0.39	4.24	4	3	1
	100	*LASSO*	0.21	4.17	6	3	3
		ENET	0.14	4.11	4	3	1
		CAEN1	0.21	4.16	6	3	3
		CCAEN1	0.13	4.12	4	3	1
		CAEN2	0.2	4.16	6	3	3
		CCAEN2	0.17	4.16	4	3	1
0.99	50	*LASSO*	10.57	4.09	4	2	2
		ENET	6.50	4.04	7	3	4
		CAEN1	6.93	4.06	8	3	5
		CCAEN1	7.49	4.05	6	3	3
		CAEN2	6.84	4.07	7	3	4
		CCAEN2	6.67	4.07	7	3	4
	100	*LASSO*	7.13	4.16	4	2	2
		ENET	5.41	4.16	7	3	4
		CAEN1	5.92	4.18	8	3	5
		CCAEN1	6.05	4.16	6	3	3
		CAEN2	5.60	4.19	6	3	3
		CCAEN2	5.62	4.18	6	3	3

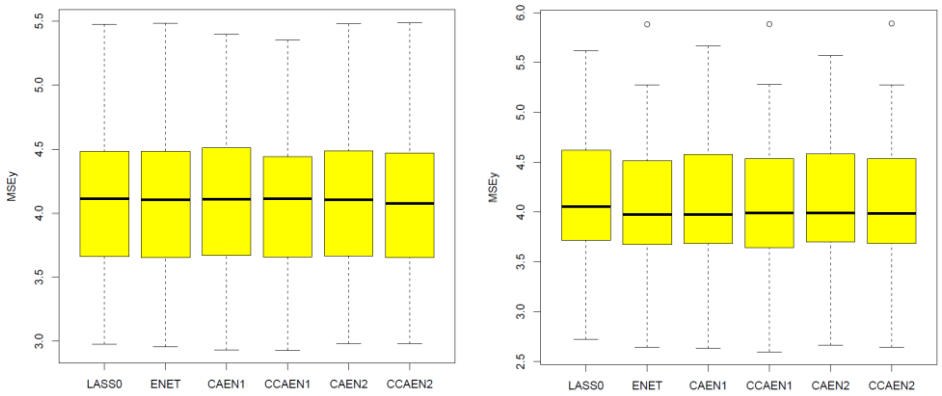
The simulation results for case 1 at training sample sizes (n) of 50 and 100 for $\theta = 0.5$ and $\theta = 0.99$ are presented in Table 1 while the boxplots in Figure 1 provide a graphical view of predictive performance (MSE_Y). The results generally show that CCAEN1 and CCAEN2 outperform CAEN1 and CAEN2 respectively in respect to prediction accuracy and variable selection. The results also show that CCAEN1 has the least estimation error at both sample sizes considered when the correlation among the predictors is moderate to low ($\theta = 0.5$), while ENET has the best estimation performance at both sample sizes when $\theta = 0.99$. The LASSO expectedly performs poorly when there is a high correlation among some of the predictors ($\theta = 0.99$), especially with respect to estimation.



(a) $n = 50$, $\theta = 0.5$



(b) $n = 100$, $\theta = 0.5$



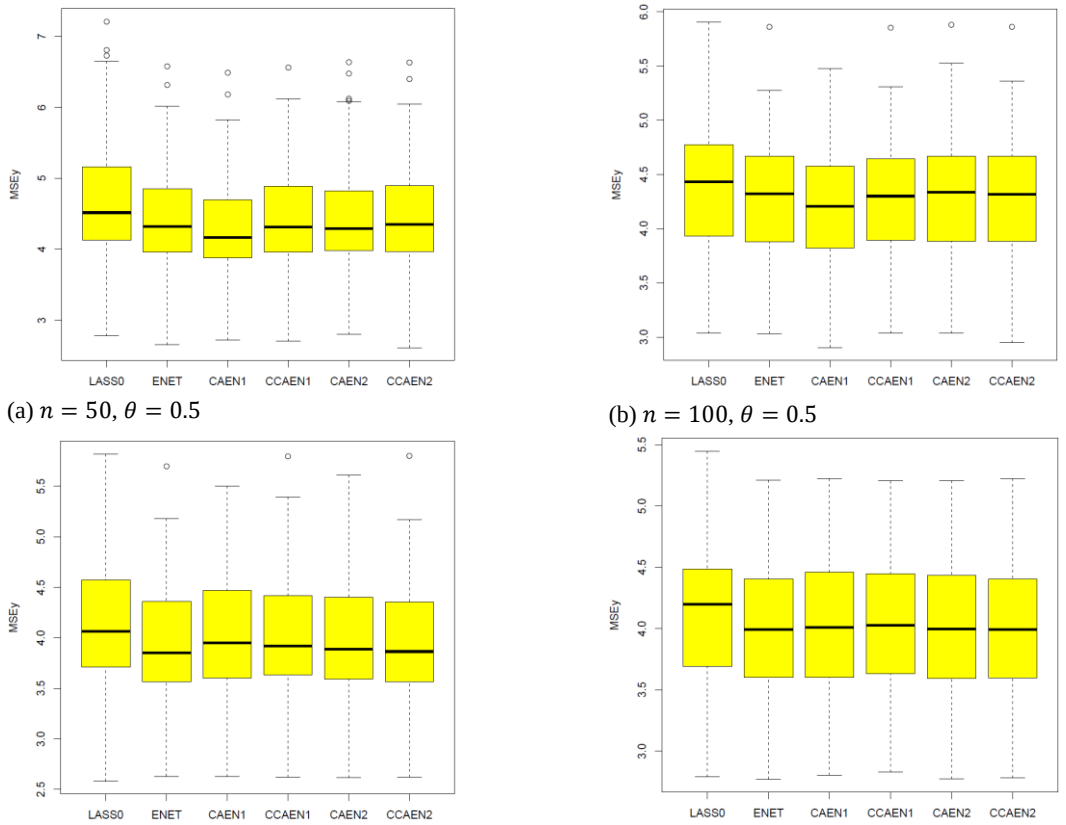
(c) $n = 50, \theta = 0.99$ (d) $n = 100, \theta = 0.99$

Figure 1: Boxplots of Empirical MSE_Y comparing LASSO, ENET, CAEN1, CCAEN1, CAEN2 and CCAEN2 for case 1 over 100 replications at (a) $n = 50, \theta = 0.5$; (b) $n = 100, \theta = 0.5$; (c) $n = 50, \theta = 0.99$ and (d) $n = 100, \theta = 0.99$

Table 2: Medians of mean squared errors of estimation and prediction (MSE_β and MSE_Y); median estimated model sizes (S), median Hits (TP) and median false positives (FP) for CASE 2 when $\theta = 0.5, 0.99$ based on 100 replications

θ	n	Method	MSE_β	MSE_Y	S ($TS = 8$)	TP	FP
0.5	50	*LASSO*	1.07	4.52	8	8	0
		ENET	0.72	4.32	8	8	0
		CAEN1	0.32	4.16	8	8	0
		CCAEN1	0.69	4.31	8	8	0
		CAEN2	0.54	4.29	8	8	0
		CCAEN2	0.74	4.35	8	8	0
	100	*LASSO*	0.48	4.43	8	8	0
		ENET	0.33	4.32	8	8	0
		CAEN1	0.18	4.21	8	8	0
		CCAEN1	0.33	4.30	8	8	0
		CAEN2	0.32	4.34	8	8	0
		CCAEN2	0.34	4.32	8	8	0
0.99	50	*LASSO*	10.13	4.07	4	4	0
		ENET	0.05	3.85	8	8	0
		CAEN1	0.49	3.95	8	8	0
		CCAEN1	1.08	3.92	8	8	0
		CAEN2	0.04	3.89	8	8	0
		CCAEN2	0.15	3.86	8	8	0
	100	*LASSO*	7.05	4.20	5	5	0
		ENET	0.04	3.99	8	8	0
		CAEN1	0.31	4.01	8	8	0
		CCAEN1	0.78	4.03	8	8	0
		CAEN2	0.02	4.00	8	8	0
		CCAEN2	0.07	3.99	8	8	0

Table 2 and the boxplots in Figure 2 summarize the simulation results for case 2 which represents a situation where the predictors have a non-sparse structure. When the correlation among the predictors is moderate to low ($\theta = 0.5$), all the regularized methods have similar performances with regards to prediction and variable selection while CAEN1 has the best performance as regards estimation. With high correlation among some of the predictors ($\theta = 0.99$), LASSO produced the worst performance among the regularized methods and CAEN2 has the best performance regarding estimation. The performances of all the methods improve as sample size increases. These findings are similar to the observations made by Zou and Hastie (2005) regarding the naive elastic-net and elastic-net under a similar simulation setting. Summarily, when the model is dense as in case 2 and there is moderate to low correlation among the predictors, CAEN1 is the best but if the correlation among some of the predictors is high when the model is dense, then CAEN2 produces the best performance.



(a) $n = 50, \theta = 0.5$ (b) $n = 100, \theta = 0.5$

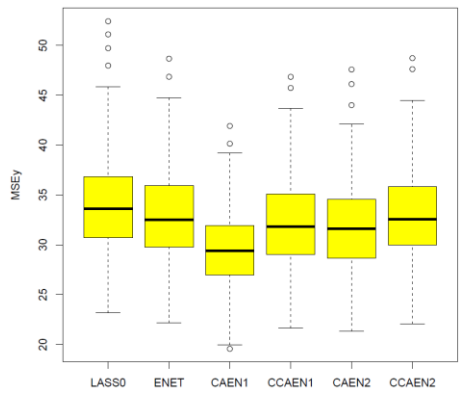
(c) $n = 50, \theta = 0.99$ (d) $n = 100, \theta = 0.99$

Figure 2: Boxplots of Empirical MSE_Y comparing LASSO, ENET, CAEN1, CCAEN1, CAEN2 and CCAEN2 for case 2 over 100 replications at (a) $n = 50, \theta = 0.5$; (b) $n = 100, \theta = 0.5$; (c) $n = 50, \theta = 0.99$ and (d) $n = 100, \theta = 0.99$

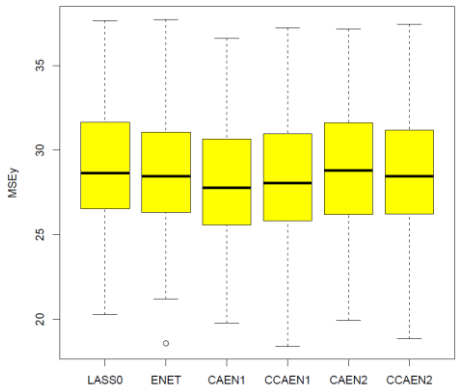
Table 3: Medians of mean squared errors of estimation and prediction (MSE_β and MSE_Y); median estimated model sizes (S), median Hits (TP) and median false positives (FP) for CASE 3 based on 100 replications

n	*Method*	MSE_β	MSE_Y	S $(TS = 20)$	TP	FP
100	***LASSO***	17.9	33.61	29	20	9
	ENET	16.13	32.50	29	20	9.5
	CAEN1	9.35	29.40	36.5	20	16.5
	CCAEN1	15.12	31.83	28	20	8
	CAEN2	13.87	31.63	35	20	15
	CCAEN2	16.16	32.57	29	20	9
200	***LASSO***	7.87	28.65	29	20	9
	ENET	7.50	28.44	30	20	10
	CAEN1	5.73	27.78	36	20	16
	CCAEN1	6.58	28.06	28	20	8
	CAEN2	7.51	28.79	32	20	12
	CCAEN2	7.55	28.45	30	20	10

Table 3 and Figure 3 summarize the results for case 3 which represents a situation where there is a moderate number of relevant and irrelevant slightly correlated predictors. The generated dataset for case 3 has 20 relevant and 20 noise predictors indicating a situation where the predictors have a moderately dense structure. Also, the pairwise correlation among the 40 generated predictors is set at 0.5. The results indicate that CAEN1 is the best in terms of prediction and estimation. However, our proposed methods here have the best performance in reference to variable selection with CCAEN1 producing the best performance. Zou and Hastie (2005) reported similar findings in that the naive elastic-net outperformed the elastic-net regarding prediction under a similar simulation setting. It can also be observed that CCAEN1 and CCAEN2 produce results that are identical to the ENET under this setting.



(a) $n = 100$



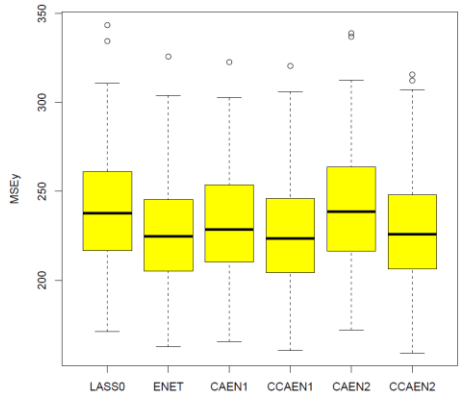
(b) $n = 200$

Figure 3: Boxplots of Empirical MSE_Y comparing LASSO, ENET, CAEN1, CCAEN1, CAEN2 and CCAEN2 for case 3 over 100 replications at (a) $n = 100$ and (b) $n = 200$

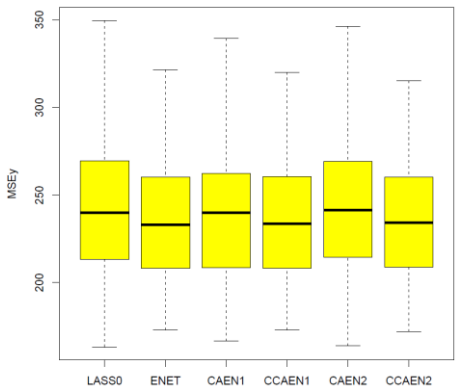
Table 4: Medians of mean squared errors of estimation and prediction (MSE_β and MSE_Y); median estimated model sizes (S), median Hits (TP) and median false positives (FP) for CASE 4 based on 100 replications

n	*Method*	MSE_β	MSE_Y	S $(TS = 15)$	TP	FP
100	***LASSO***	420.38	237.60	10	3	7
	ENET	1.62	225.80	17	15	2
	CAEN1	16.99	228.53	40	15	25
	CCAEN1	14.60	224.41	16	15	1
	CAEN2	99.12	238.58	18	11	7
	CCAEN2	1.96	227.32	16	15	1
200	***LASSO***	450.72	239.86	10	3	6
	ENET	0.72	233.07	16	15	1
	CAEN1	16.01	239.74	40	15	25
	CCAEN1	13.89	233.68	16	15	1
	CAEN2	69.21	241.24	19	13	6
	CCAEN2	0.72	234.23	16	15	1

The results for case 4, which represents situations where there are grouped predictors are presented in Table 4 and Figure 4. The results reveal that CCAEN1 and CCAEN2 perform better than CAEN1 and CAEN2 in all criteria. CAEN2 seems not to be able to do group variable selection adequately. The LASSO as expected has the worst performance in this case while the ENET and CCAEN2 have the best performance with respect to estimation error and variable selection. However, the CAEN1 while selecting all the relevant predictors always selects larger models containing many irrelevant predictors compared to others.



(a) $n = 100$



(b) $n = 200$

Figure 4: Boxplots of Empirical MSE_Y comparing LASSO, ENET, CAEN1, CCAEN1, CAEN2 and CCAEN2 for case 4 over 100 replications at (a) $n = 100$ and (b) $n = 200$

Table 5: Medians of mean squared errors of estimation and prediction (MSE_β and MSE_Y); median estimated model sizes (S), median Hits (TP) and median false positives (FP) for CASE 5 based on 100 replications at $n = 100$

Method	MSE_β	MSE_Y	$S\ (TS = 20)$	TP	FP
LASSO	7.01	15.69	42	20	22
ENET	6.88	15.28	39.5	20	19.5
CAEN1	6.05	15.45	46.5	20	26.5
CCAEN1	4.64	12.47	32	20	12
CAEN2	7.74	17.83	43	20	23
CCAEN2	6.92	15.39	40	20	20

The simulation setting for case 5 represents the scenario where $p \gg n$ and the results are presented in Table 5. The results show that all the correlation based methods can also produce sparse solutions when $p \gg n$. The proposed CCAEN1 yields the best results in all criteria considered. It can also be noted that CCAEN1 and CCAEN2 produce better predictions and select more parsimonious models compared to CAEN1 and CAEN2 respectively.

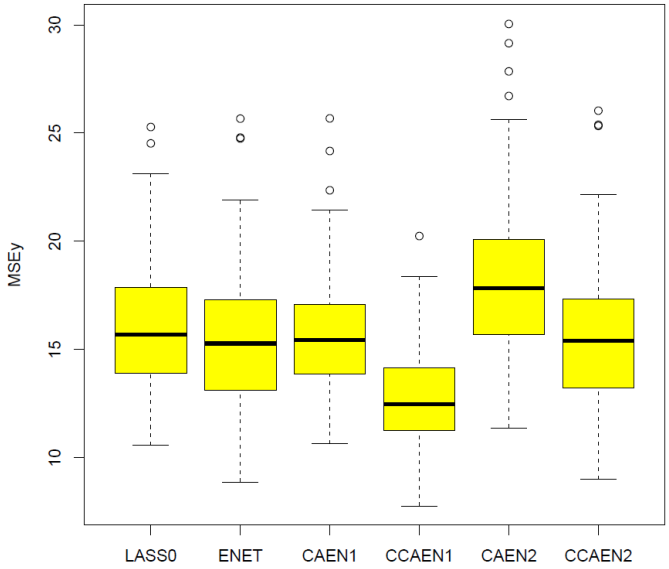


Figure 5: Boxplots of empirical MSE_Y comparing LASSO, ENET, CAEN1, CCAEN1, CAEN2 and CCAEN2 for case 5 over 100 replications at $n = 100$

Table 6: Medians of mean squared errors of estimation and prediction (MSE_β and MSE_Y); median estimated model sizes (S), median Hits (TP) and median false positives (FP) for CASE 6 based on 100 replications at $n = 100$

Method	MSE_β	MSE_Y	$S\ (TS = 20)$	TP	FP
LASSO	1.29	10.79	14	3	11
ENET	1.17	10.46	13	3	10
CAEN1	1.3	10.82	15	3	12
CCAEN1	1.17	10.12	12	3	9
CAEN2	1.42	10.94	15	3	12
CCAEN2	1.33	10.57	13	3	10

Table 6 and Figure 6 provides the results for case 6 in which the predictor matrix X has a sparse structure with $p \gg n$. The results are similar to those of case 5 with CCAEN1 and CCAEN2 outperforming CAEN1 and CAEN2 respectively. Again, CCAEN1 seems to produce the best performance while CAEN2 seems to be outperformed by all other methods.

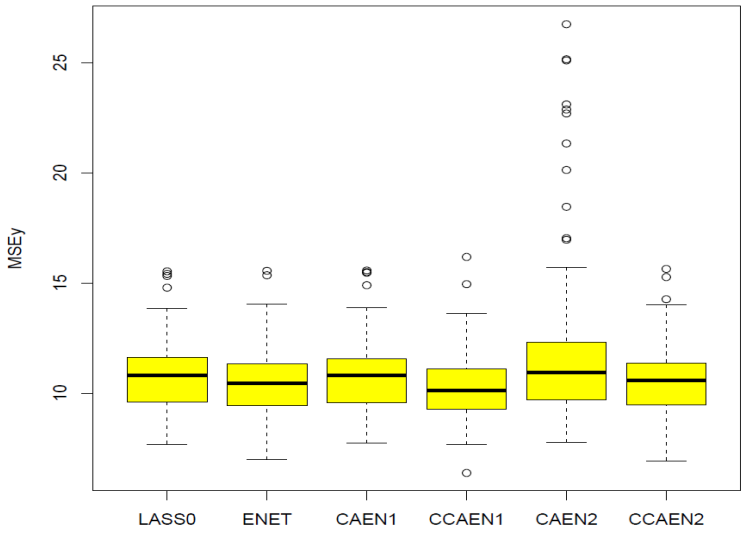


Figure 6: Boxplots of empirical MSE_Y comparing LASSO, ENET, CAEN1, CCAEN1, CAEN2 and CCAEN2 for case 6 over 100 replications at $n = 100$

Generally, from the simulation study, it appears that when the model is dense and there is moderate to high correlations among the predictors, the extra shrinkage incurred by the correlation-based penalties is needed and should not be undone. Therefore, the naïve correlation-based methods are better in terms of estimation and prediction when the correlations among the predictors when the model is dense. However, the naïve methods (CAEN1 and CAEN2) always select larger models including irrelevant predictors compared to our proposed methods implying that our proposed methods are better at producing sparse solutions while still preserving the ability to do 'grouped variables' selection.

Discussion

Application to Real Life Datasets

This section presents applications of the proposed methods on two real-life datasets. The data consists of a response variable which is the log of prostate specific antigen (lpsa) and eight predictors. The predictors are 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 prostate cancer dataset were analysed in (Zou & Hastie, 2005).

The second dataset, which is referred to as the gene expression dataset, comes from the microarray experiments on 120 mammalian eye tissue samples (Scheetz et al., 2006). The dataset consists of 200 predictors which represent 200 gene probes of 120 rats. The response 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 LASSO, ENET, CAEN1, CCAEN1, CAEN2 and CCAEN2 were all applied to the prostate dataset while only the regularized methods were applied to the gene expression dataset.

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 was used for model fitting and selection of tuning parameters by 10-fold cross-validation. The performance of the methods is 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 was repeated 100 times. The results for both datasets are summarized in Table 7.

The first dataset which we refer to as the prostate cancer dataset comes from a study of prostate cancer by Stamey et al. (2001)) involving 97 men. The data consist of a response variable which is the log of prostate specific antigen (lpsa) and eight predictors. The predictors are log cancer volume

Table 7: 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
LASSO	0.594	5	0.008	23
ENET	0.597	5	0.008	31
CAEN1	0.601	6	0.007	102
CCAEN1	0.595	5	0.008	22
CAEN2	0.605	6	0.007	145.5
CCAEN2	0.614	5	0.007	23

The results in Table 7 indicate that the naive versions of the correlation-based methods select larger models compared to the corresponding scaled versions with no substantial gain in prediction accuracy in both datasets. For the prostate cancer data, only the lasso performs better than CCAEN1 in terms of prediction and sparsity. The LASSO, ENET, CCAEN1 and CCAEN2 select 5 predictors while CAEN1 and CAEN2

select the higher number of predictors with no significant gain in prediction accuracy.

The second gene expression dataset is a real life situation when the number of predictors exceeds the sample size. The results from applying the considered methods on the gene expression dataset show that the proposed rescaled correlation-based methods outperform their corresponding nave versions in terms of sparsity with no significant loss in prediction accuracy terms. The CAEN2 produced a prediction error of about 0.007, albeit with the median number of predictors of 145.5 which is almost six times that of the CCAEN2 with the same prediction error. The CCAEN1 selected the fewest number of predictors (22 predictors) with no significant difference between its prediction error (0.008) and the lowest prediction error of 0.007 which was produced by the CAEN1, CAEN2 and CCAEN2 using an average of 102, 145.5, 23 and 176 predictors respectively. Out of all the nave correlation-based methods, the CAEN1 is the best in terms of sparsity and prediction accuracy. The results from this section further show that scaling the correlation-based methods improves their performances in terms of variable selection.

CONCLUSION

Decomposition of the ridge operator is important to rescale and undo the double shrinkage incurred by using both the ridge and LASSO penalties. The CAEN1 and CAEN2 assess the finite sample performance where an efficient and robust worst-case quadratic solver method was adopted for estimation. The CCAEN1 and CCAEN2 through simulations and applications to real-life datasets perform better in terms of variable selection and ability to handle grouping effects than the naive CAEN1, CAEN2, LASSO and ENET. The CCAEN1 produce the best results in high-dimensional situations ($p \gg n$) while CCAEN2 outperforms all other methods when there are grouped variables. By correctly identifying relevant variables in high-dimensional problems, CCAEN1 and CCAEN2 have made a significant contribution to big data analytics and mining.

ACKNOWLEDGMENT

This work was supported by the Tertiary Education Trust Fund (TETFund) Institutional Based Research (IBR) grant awarded to Federal University Lokoja in 2024. We acknowledge Mrs Chinenye Ezenweke Pauline of the Department of Statistics, Federal University Lokoja for her insightful discussions on the simulation study.

REFERENCES

Algamal, Z. Y. (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*, 82-102.

Babarinsa, O., Sofi, A. Z. M., Mohd, A. H., Eluwole, A., Sunday, I., Adamu, W., Daniel, L. (2022). Note on the history of (square) matrix and determinant. *FUDMA JOURNAL OF SCIENCES, 6*(3), 177-190.

Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences, 2*(1), 183-202.

Biecek, P., & Burzykowski, T. (2021). *Explanatory model analysis: explore, explain, and examine predictive models*: Chapman and Hall/CRC.

Bondell, H. D., & Reich, B. J. (2006). *Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR*. Retrieved from

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The annals of statistics, 24*(6), 2350-2383.

Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American statistical Association, 92*(438), 548-560.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*: Chapman and Hall/CRC.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association, 96*(456), 1348-1360.

Fan, J., & Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *arXiv preprint math/0602133*.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics, 7*(3), 397-416.

Garba, W., Yahya, G., & Aremu, M. (2016). Multiclass Sequential Feature Selection and Classification Method for Genomic Data. *Blood, 7*(10).

Grandvalet, Y., Chiquet, J., & Ambroise, C. (2012). Sparsity by Worst-Case Penalties. *arXiv preprint arXiv:1210.2077*.

Hanke, M., Dijkstra, L., Foraita, R., & Didelez, V. (2024). Variable selection in linear regression models: Choosing the best subset is not always the best choice. *Biometrical Journal, 66*(1), 2200209.

Hapfelmeier, A., Babatunde, W., Yahya, R. R., & Ulm, K. (2012). 26 Predictive modeling of gene expression data. *Handb Stat Clin Oncol, 4*, 71.

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

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection.* Paper presented at the International Joint Conference on Artificial Intelligence.

Ryan, T. (2008). *Modern regression methods* (Vol. 655): John Wiley & Sons.

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., . . . Casavant, T. L. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences, 103*(39), 14429-14434.

Stamey, T. A., Warrington, J. A., Caldwell, M. C., Chen, Z., Fan, Z., Mahadevappa, M., . . . Zhang, Z. (2001). Molecular genetic profiling of Gleason grade 4/5 prostate cancers

compared to benign prostatic hyperplasia. *The Journal of urology*, *166*(6), 2171-2177.

Tan, Q. E. A. (2012). *Correlation adjusted penalization in regression analysis.* (Ph.D.), University of Manitoba Canada.

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

Tutz, G., & Ulbricht, J. (2009). Penalized regression with correlation-based penalty. *Statistics and Computing*, *19*, 239-253.

Wang, X., Dunson, D., & Leng, C. (2016). *No penalty no tears: Least squares in high-dimensional linear models.* Paper presented at the International Conference on Machine Learning.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, *101* 1418-1429.

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.