Al-Qadisiyah Journal for Administrative and Economic Sciences ISSN$_{Online}$: 2312-9883
QJAE, Volume, Issue ISSN$_{Print}$: 1816-9171
Manuscript ID:

Comparison of the Performance of Bayesian Elastic Net Regression and Elastic Net Regression a Simulation Study

[1]Ali Ahmad Hussein [2]Dr. Bahr Kadhim Mohammed

[1, 2] *(Dept. of Statistics, Faculty of Administration and Economics/*

University of Al-Qadisyiah, IRAQ,) ***stat.post13@qu.edu.iq , bahr.mahemmed@qu.edu.iq***

Corresponding Author : Ali Ahmad Hussein Affiliation : Dept. of Statistics, Faculty of Administration and Economics
Email: stat.post13@qu.edu.iq

Abstract

Bayesian elastic net and classical elastic net are regularization methods that provide variable selection procedure. We discuss the Bayesian elastic net by setting the scale mixture of normal distribution mixing with Rayleigh distribution as double exponential (Laplace) prior distribution of regression coefficient. The new proposed scale mixture produced normal distribution mixing with truncated gamma distribution. The hierarchical prior distributions and new Gibbs sample algorithm have developed. Therefore, variable selection have discussed through some simulation examples. The simulation results show the outperformance of the proposed model.

Keywords*: Bayesian elastic net, Hierarchical model, Gibbs sampler, Simulation.*

1-Introduction

Regression methodology founded to estimate the expected of the response variable based on the information that provided by the predictor variables. The parameter estimates of regression model are reliable estimates if it offers balance between the variance and bias, in addition to the model explain ability. It is well known that the OLS estimates are biased and inconsistent (inflated variance) when the multicollinearity problem appear in the design matrix X, or when the number of predictor variable p exceed or near the number of observations n. Therefore, in these circumantance the OLS estimates are usually not unique and instable with high variances. The high variance in the OLS estimates motivated the authors to explore the regularization methods that used to overcomes the limitations of least squares estimates quality, James et al. (2013). The ridge regression method adding a penalty function to residuals sum of squares (RSS) to address the problem multicollinearly, where the penalty function contains the L2-norm. The ridge parameter estimates cannot set to zero, Hoerl and Kennard (1970). Tibshirani,(1996) produced Lasso method which is essentially regards as penalized method that provide variable selection procedure. Consequently, many authors developed other shrinkage methods to provide variable selection procedure; such as, relaxed lasso, fused lasso, adaptive lasso, elastic net, etc. Model selection procedure in regression analysis aims to select the best fit estimated regression model through selecting the relevant predictor variables that affects the response variable and remove the irrelevant variables. In paper I consider the linear regression model where the ordinary least squares (OLS) estimates are no longer achieved by minimizing the residual sum squares (RSS). Instead of the OLS the elastic net have discussed in this paper, elastic net is the flexible regularization and variable selection method that combined two of penalties function. Moreover, the Elastic Net (**EN**) is another penalized method that proposed by Zou and Hastie (2005) to address the limitations of lasso method. EN method combined the ridge and lasso to the RSS term, **EN** method deal with many relevant predictors that have highly pairwise correlation and **EN** usually works better than lasso. Obviously the regression analysis methods are very widely popular tools that investigated the relationship between the response variable and the independent variable(s). This motivated many authors and researchers to develop various regression analysis tools that cope with the practical underlying situation. The ordinary least squared (OLS) method is very common tool to find the regression coefficient estimates. Moreover, violated the assumptions of (OLS) was the key idea behind searching for

substitution methods for regression coefficient estimates. In addition for that the investigation about the more explanation model developed along with the model selection and variable selection procedures. The Ordinary Least Squares provided unbiased and smallest variance parameters estimates through minimizing of the Residual Sum of Squares (RSS).

In regression analysis, the set of the independent variable that should be included in regression equation bring the attention of the researcher, because it is the first part of the regression analysis and then examine to see whether regression equation was correct. So, the variable selection problem related with the regression form specification. The residual mean squares (RMS) is a criterion for model selection, the smallest the RMS between two regression equation is preferred.Mallows (1973) developed the Mallows C_k criterion to judge the performance of the regression function by using the following form.

Akaike (1973) introduced the Akaike information criterion (AIC) as model selection criterion that combined the most fit equation and the smaller number of independent variables, the smallest AIC value the better model. Schwarz (1978) proposed a modification of the AIC is called Bayes Information criterion (BIC) which is defines as follows, the smallest BIC value the better model. Zou et al. (2007) discussed the using of of BIC criterion to choose the shrinkage parameter in lasso method. Hocking (1976) list the evaluating regression method that is called all possible equations which is gives 2^k equations (k is the number of independent variables), where we can use the (RME, C_k, R^2) to select the best model. The limitation of all possible equations is the larger number of equations when k getting larger. Efroymson (1960) introduced the stepwise method as variable selection procedure combined the mechanism of both Forward Selection (FS) Procedure and Backward Elimination (BS) procedure. The calculation of the stepwise method depends on the inclusion and deletion of independent variables, it is essentially a modification method for (FS and BE) methods. The AIC and BIC are used for select the best fitted model in the stepwise method. It is recommended obtaining the variance inflation factors (VIF) test or the eigenvalues of the correlation matrix of the independent variables as a first step to variable selection procedure. George and McCulloch (1993) proposed another method for utilizing an information criterion for model selection; this method is called stochastic search variable selection ($SSVS$). This method can be used in the well know Bayesian algorithm ($MCMC$), so it is depends on the probabilistic considerations in selecting of the subsets of independent variables. Hoerl and Kennard (1970) introduced a theory about ridge regression with penalized function to estimate the parameters of multiple regression model by adding a small positive quantity (λ) to the inverse of (X^tX) matrix to address the problem of linearly dependent (correlation) of the independent variable. The ridge estimator is biased but with the smallest variance. Also, ridge methods can be applied in the case of ($n \geq k$) and regards as regularization method. But ridge regression is not a variable selection method. Ridge uses the L2-norm as penalty function.The response variable in ridge regression is centered (Draper and Smith, (1998). Tibshirani (1996) proposed the new variable selection method that is called Lasso. Lasso method can be regards as regularization method that adds the L1-norm penalty function to the RSS. Due to the L1-norm, lasso provides variable selection procedure by setting the parameter estimates to zero. Also, in this paper there is a remarkable note about Bayes estimation for the linear regression model based on assuming that the parameter β is follows the double exponential distribution as prior density.

Zou and Hastie (2005) introduced the so called elastic net, which is regards as regularization method that combined the ridge and lasso methods. It can be considered as variable selection method that works simultaneously as variable selection and shrinkage method. Furthermore, the elastic net dealing well with a grouping effect of correlated independent variables as contrast of lasso.

Li and Lin (2010), introduced the parameter estimation of the elastic net model from the Bayesian perception. By using the Gibbs sampler algorithm based on considering that the prior density is a scale mixture of normal mixing with truncated Gamma. The linear regression model studied for variable selection and prediction accuracy, the proposed model outperform in variable selection procedure and is a comparable model in the terms of prediction accuracy. Park and Casella (2008) developed Gibbs sample algorithm based on new Bayesian hierarchal prior model. The scale mixture of normal mixing with exponential density have used as representation form for the double exponential prior distribution through the lasso linear. Mallick and Yi (2014) introduced new Bayesian lasso method that depends on new representation of the double exponential prior density as scale mixture of uniform mixing with special case of gamma distribution. Variable selection procedure has performed and parameter estimation explained based on the new lasso method.

Flaih et al. (2020a) introduced new scale mixture of normal mixing Rayleigh density to represent the double exponential prior density. New hierarchal prior model have developed and therefore new Gibbs sample algorithm have implement to calculate the mode of the posterior density of lasso regression model parameter. The proposed model is comparable in terms of variable selection and estimation accuracy. Fadel et al. (2020b) developed an extension for lasso Tobit and adaptive lasso Tobit regression models based on the proposed scale mixture in Flaih et al. (2020a).

2-Bayesian Hierarchical Prior models

Elastic net penalized method is very common used in regression model as a regularization method which combines the ridge and lasso penalty functions the elastic net method. Zou and Hastie (2005) introduced the elastic net method as sparsity procedure that can deal with the effect of correlated variables of covariates , the elastic net estimator is defined as follows ,

$$\hat{\beta} = argmin \, \|y - X\beta\|^2_2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|^2 \ldots (1)$$

Where the elastic net penalized function is

$$h(\beta) = \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|^2$$

here $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ the penalties parameters.

Flaih et al. (2020) introduced the Bayesian lasso regression model based on scale mixture representation of normal mixing with Rayleigh density. In this paper I assumed the above scale mixture by considering the linear regression model:

$$E\,(y \,/\, X, \beta) = X\beta$$

Suppose that the scale mixture of Laplace distribution that mixing normal with Rayleigh distribution defined as follows,

If $x/y \sim N\,(\mu\,, y^2)$ with y~Ray (b), then x~Laplace (μ, b), that is:

$$\frac{1}{2b} e^{-\frac{|x-\mu|}{b}} = \int_0^\infty \frac{1}{\sqrt{2\pi y^2}} e^{-\frac{(x-\mu)^2}{2y^2}} \frac{y}{b} e^{-\frac{y^2}{2b}} dy \quad \ldots (2)$$

by letting μ=0, X =β, and b = $\frac{\sigma^2}{\lambda_1}$, then (2) become as follows :

$$\frac{\lambda_1}{2\sigma^2} e^{-\frac{\lambda_1|\beta|}{2\sigma^2}} = \int_0^{\infty} \frac{1}{\sqrt{2\pi y^2}} e^{-\frac{\beta_j^2}{2y^2}} \frac{\lambda y}{\sigma^2} e^{-\frac{\lambda_1 y^2}{2\sigma^2}} dy \quad \dots (3)$$

Zou and Hastie (2005) introduced the prior distribution of elastic net method $\pi(\beta)$ as:

$$\pi(\beta) \propto e^{-\lambda_1\|\beta\|^1 - \lambda_2\|\beta\|_2^2}, \qquad \dots (4)$$

From (4), we can deal with β_j/σ^2 as Scale mixture of normal distributions N $(0,\frac{\sigma^2(t-1)}{\lambda_2 t})$ mixing truncated gamma with shape parameter (1/2) and Scale parameter ($\frac{2\lambda_2}{\lambda_1}$), see Almusaedi and Flaih (2021a, 2021b), Alsafi and Flaih (2021) for more information. By formula (4), we have the following elastic net linear regression (ENLR) hierarchical model,

$$\left.\begin{array}{c} y = X\beta + e\,, \\ y|X\,,\beta\,,\sigma^2 \sim N\,(X\beta\,,\sigma^2\,In\,), \\ \beta\,\Big|\lambda_2,\sigma^2,t \sim \prod_{j=1}^{p} N\,(0,\left(\frac{\lambda_2}{\sigma^2}\,\frac{t_j}{t_j-1}\right)^{-1}) \\ t|\lambda_1,\lambda_2 \sim \prod_{j=1}^{p} truncated\ gamma\left(\frac{1}{2}\,,\frac{2\lambda_2}{\lambda_1}\right); t \in (1,\infty), \\ \sigma^2 \sim Inverse\ Gamma(\alpha,\tau) \end{array}\right\} \dots (5)$$

3- Full Conditional Posterior Distributions of ENLR

By using the hierarchical model (5), the full joint distribution is well defined as follows:

$\pi\left(\beta|y,X\,,\sigma^2\,,t\right) \propto \pi(y/X\,,\beta\,,\sigma^2)$

$f\left(y|\beta,\sigma^2\right)\pi\left(\sigma^2\right)\prod_{j=1}^{p}\pi\left(\beta_j|\ t_j\,,\sigma^2\right)\pi(t_j) \ =$

$$\left(\frac{1}{\sqrt{2\,\pi\,\sigma^2}}\right)^n e^{-\frac{1}{2\,\sigma^2}(y-X\beta)'(y-X\beta)} \cdot \frac{\tau^{\alpha}}{\sqrt{\alpha}}\ (\sigma^2)^{-\alpha-1}\ e^{-\frac{\tau}{\sigma^2}} \quad \prod_{j=1}^{p}\sqrt{\frac{\lambda_2\,t}{\sigma^2\,(t-1)}} \quad e^{-\frac{\beta_j^2}{2}\left(\frac{\lambda_2}{\sigma^2}\cdot\frac{t}{(t_j-1)}\right)}\ t^{-\frac{1}{2}}\ e^{\frac{-\lambda_1}{2\lambda_2}t} \quad \dots (6)$$

Where α is the shape parameter and τ is the rate parameter of inverse gamma distribution. Now the full conditional posterior distributions are as follows:

The parts that includes β, $\pi(\beta)$ in the joint distribution (6) is

$$e^{-\frac{1}{2\,\sigma^2}(y-X\beta)'(y-X\beta)-\frac{1}{2\sigma^2}\lambda_2\beta'\,A\,\beta} \quad \text{, where } A = \left(\frac{t}{t-1}\right)$$

$\pi(\beta) = \exp\left[-\frac{1}{2\,\sigma^2}\left\{\left(\beta'\,(X'X)\,\beta - 2yx\beta + y'y\right) + \lambda_2\,\beta'\,A\,\beta\right\}\right]$

$= \exp\left[-\frac{1}{2\sigma^2}\left\{\beta'\left(X'X + \lambda_2 A\right)\beta - 2yX\beta + y'y\right\}\right]$

$$= \exp\left[-\frac{1}{2\,\sigma^2}\left\{\left(\beta^{'} C\beta - 2yx\beta + y^{'}y\right)\right\}\right]$$

Where C= $X^{'}X + \lambda_2 A$

$$\pi(\beta) = \exp\left\{-\frac{1}{2\sigma^2}\left(\beta^{'}C\beta - 2yX\beta + y^{'}y\right)\right\} \ldots \quad (7)$$

Let $\left(\beta - C^{-1}X^{'}y\right)^{'}C\left(\beta - C^{-1}X^{'}y\right) = \beta C^{'}\beta - 2yX\beta + y^{'}(XC^{-1}X)^{'}y$

then (7) Can rewrite as follows:

$$\exp\left[-\frac{1}{2\,\sigma^2}\left\{\left(\beta - C^{-1}X^{'}y\right)^{'} C\left(\beta - C^{-1}X^{'}y\right) + y^{'}(In - XC^{-1}X^{'})y\right]\right. \quad \ldots (8)$$

The second part of (8) does not involve β, so we can reduce (8) as follows

$$\pi(\beta) = \exp\left[-\frac{1}{2\,\sigma^2}\left\{\left(\beta - C^{-1}X^{'}y\right)^{'} C\left(\beta - C^{-1}X^{'}y\right)\right]\right. \ldots \quad (9)$$

We can say that (9) is the multivariable normal distribution with mean $C^{-1}\,X^{'}y$ and variance $\sigma^2 C^{-1}$.

The second Conditional posterior distribution is for σ^2, $\pi(\sigma^2)$. The terms that involve σ^2 in the full joint distribution (6) are as follows

$$\pi(\sigma^2) = (\sigma^2)^{-\frac{n}{2}}\ (\sigma^2)^{-\alpha-1}\ \ (\sigma^2)^{-\frac{p}{2}} - \frac{1}{e^{2\sigma^2}}(y - X\beta\acute{)}(y - X\beta) - \frac{\tau}{\sigma^2} - \frac{\beta^{'}\lambda_2 A\beta}{2\sigma^2}$$

$$(\sigma^2)^{-\frac{n}{2}-\frac{p}{2}-\alpha-1}\ -\frac{1}{e^{2\sigma^2}}\{(y-X\beta)^{'}(y-X\beta)+\tau+\beta^{'}\lambda 2A\beta\} \quad \ldots \quad (10)$$

The formula (10) is the inverse gamma distribution with shape parameter

$$\left(\frac{n}{2} + \frac{p}{2} + \alpha\right) \text{ and Scale parameter } \frac{(y-X\beta)^{'}(y-X\beta)}{2} + \frac{\beta^{'}\lambda 2A\beta}{2} + \tau\ .$$

The third part in the conditional posterior distribution of (t_j). The parts of (6) that involves (t_j) are

$$\sqrt{\frac{\lambda_2}{\sigma^2}\frac{tj}{t_j - 1}}\ e^{-\frac{\beta_j^2}{2}\left(\frac{\lambda_2}{\sigma^2}\frac{t_j}{t_j - 1}\right)tj - \frac{1}{2}}\ e^{-\frac{\lambda 1}{2\lambda 2}t_j}$$

Then based on the (Chhicara and Folks 1988) works, the distribution of $(t-1)$ is the generalized inverse Gaussian distribution and defined as follows,

$$(t-1) \sim GIG\left(\lambda = \frac{1}{2}, a = \frac{\lambda 1}{4\lambda_2\sigma^2}, \chi = \frac{\lambda_2\beta_j^2}{\sigma^2}\right), \ldots \quad (11)$$

Then, $(t-1)^{-1}$ variable follows the full conditional inverse Gaussian distribution with $\mu = \frac{\sqrt{\lambda_1}}{(2\lambda_2|\beta_j|)}$ and $\lambda = \frac{\lambda_1}{4\lambda_2\sigma^2}$.

See (Chhikara and Folks 1988) for more details. The choosing of the Shrinkage parameters λ_1 andλ_2 conducted by Li and Lin (2010) and Park and Casella (2008), they used the empirical Bayes procedure.

4-Simulation Experimental

In this section , simulation study will be conducted to show the behavior of our proposed model , Bayesian elastic net regression using R package (BANTER) and compared with different exists models , the elastic net regression model (AN) by implementing the (rn) R package , and the lasso elastic net regression model (lnr) by implementing the R package . Our comparison is based on the parameters estimates of the different models different elastic net . Also , we used the median mean absolute deviation (MMAD) criterion .

$$mmad = median\left[mean\left|x^T\hat{\beta} - x^T\beta^{true}\right|\right]\ldots.(12)$$

The MMAD and the standard deviation (SD) are used to measure the performance of prediction accuracy for different model . The Gibbs sample algorithm have been used with 10000 iterations to generate the stability of the posterior distribution of the interested parameter assuming the number of observation is n = 400 , the first 1000 iterations have burned in . We generated the observations of predictor variables from $X \sim N(0,\Sigma)$, where the matrix $\sum_{ij} = \rho^{|i-j|}$, with three distributions of (**i.i.d.**) error terms .

1-First example

In this example , we assumed that the true vector of the true parameter (very sparse), $\boldsymbol{\beta} = (\mathbf{1},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0})$ with error distributed according to standard normal $\boldsymbol{e_i \sim N}\ (\mathbf{0},\mathbf{1})$,
$\boldsymbol{e_i \sim N}\ (\mathbf{1},\mathbf{1}), \boldsymbol{e_i \sim N}\ (\mathbf{2},\mathbf{2}) + \boldsymbol{N}\ (\mathbf{2},\mathbf{2}), \boldsymbol{e_i \sim lap}\ (\mathbf{1},\mathbf{0}), \boldsymbol{and}\ \boldsymbol{e_i \sim X^2_{(4)}}$. I generated the observations of the predictors $\mathbf{X_1}, \ldots\ldots, \mathbf{X_9}$ from the multivariate normal $\mathbf{N_{n=9}}\ (\mathbf{0}, \sum)$, here $\sum$ is the **var.cov** matrix defined as $\sum_{\mathbf{ij}} = \mathbf{0.7^{|i-j|}}$. The true relationship between the predictor variables and response variable based on the above true vector is $f(X) = \sum_{j=1}^{9} X_j\beta_j$, So the correct model is defined by$f(X) = X_1\beta_1$,

Table (1) values of MMAD and SD in example one

Sample Size	Comparison Methods		$e_i \sim N(0,1)$	$e_i \sim N(1,1)$	$e_i \sim N(2,2) + N(2,2)$	$e_i \sim Lap(1,0)$	$e_i \sim x^2_{(4)}$
	n=15	BANETR	1.235(0.672)	1.130(0.845)	1.521 (0.832)	1.303 (0.672)	1.612(0.792)
		ANETR	1.543(0.830)	1.317(0.992)	1.834 (0.970)	1.452 (0.757)	1.704 (0.822)
	n=25	BANETR	1.373(0.238)	1.240(0.152)	1.234 (0.632)	1.546(0.499)	1.703(0.643)
		ANETR	1.645(0.536)	1.325(0.273)	1.564 (0.874)	1.769 (0.682)	1.892(0.782)
	n=35	BANETR	1.245(0.563)	1.547(0.482)	1.529 (0.353)	1.346 (0.482)	1.446(0.583)
Small Sample		ANETR	1.482(0.834)	1.618(0.834)	1.865 (0.932)	1.634 (0.542)	1.782 (0.671)

Meddle Sample	n=45	BANETR	1.282(0.451)	1.417(0.683)	1.220 (0.493)	1.030 (0.534)	1.106 (0.585)
		ANETR	1.597(0.780)	1.632(0.745)	1.836 (0.698)	1.573 (0.840)	1.839 (0.732)
	n=55	BANETR	1.256(0.245)	1.361(0.391)	1.435 (0.634)	1.310 (0.427)	1.420 (0.370)
		ANETR	1.562(0.792)	1.620(0.407)	1.834 (0.803)	1.478 (0.896)	1.838 (0.550)
	n=65	BANETR	1.069(0.327)	1.452(0.075)	1.520 (0.311)	1.564 (0.183)	1.305 (0.183)
		ANETR	1.623(0.832)	1.971(0.621)	1.733 (0.504)	1.826 (0.202)	1.352(0.420)
Large Sample	n=100	BANETR	1.855(0.358)	1.746(0.352)	1.107 (0.432)	1.523(0.453)	1.352(0. 534)
		ANETR	(1.964)(0.563)	(1.832)(0.832)	(1.543)(0.678)	(1.854) (0.704)	(1.676)(0.828)
	n=200	BANETR	1.241(0.332)	1.230(0.282)	1.781 (0.405)	1.530 (0.204)	1.682 (0.387)
		ANETR	(1.537)(0.564)	(1.676)(0.564)	(1970)(0.653)	(1.675)(0.734)	(1.754)(0.673)
	n=300	BANETR	1.604(0.450)	1.638(0.564)	1.754 (0.563)	1.039 (0.356)	1.651(0.432)
		ANETR	(1.722)(0.792)	(1.927)(0.671)	(1.934)(0.892)	(1.527)(0.643)	1.643)(0.854)

Table (**1**) provided the values of the MMAD and SD quality measures of the estimated regression models for the proposed method (BENLR) and the (ENLR) based on three types of sample sizes , small samples **(n=15 , n=25 , n=35)** , middle samples **(n=45 , n=55 , n=65)** , and large samples **(n=100 , n=200 , n=300)** . Clearly the values of MMAD criterion are the smallest in the proposed methods compared with the other methods under all different type of error distributions. Also, the SD criterion shows the preference of the proposed methods under different types of sample sizes and under different error distributions .Consequently, the proposed method is a promising regularization method.

2-Second example

In this example, we supposed that the true vector of parameters (sparse) $\boldsymbol{\beta} = (\mathbf{1},\mathbf{0},\mathbf{0},\mathbf{1},\mathbf{0},\mathbf{1},\mathbf{0},\mathbf{1},\mathbf{0})$ with error distributed according to standard normal $\boldsymbol{e_i}\sim\boldsymbol{N}\ (\mathbf{0},\mathbf{1})$, $\boldsymbol{e_i}\sim\boldsymbol{N}\ (\mathbf{1},\mathbf{1}), \boldsymbol{e_i}\sim\boldsymbol{N}\ (\mathbf{2},\mathbf{2}) + \boldsymbol{N}\ (\mathbf{2},\mathbf{2}), \boldsymbol{e_i}\sim\boldsymbol{lap}\ (\mathbf{1},\mathbf{0}), \boldsymbol{and}\ \boldsymbol{e_i}\sim\boldsymbol{X^2_{(4)}}$ Igenerated the observations of the covariates $\mathbf{X_1}$, …. . , $\mathbf{X_9}$ from the multivariate normal $\mathbf{N_{n=9}}$ ($\mathbf{0}$, $\sum$), here $\sum$ is the **var-cov** matrix defined as $\sum_{\mathbf{ij}} = \mathbf{0}.\mathbf{7}^{|\mathbf{i}-\mathbf{j}|}$. The true relationship between the predictor variables and response variable base on the above true vector is

$$f(X) = \sum_{j=1}^{9} X_j \beta_j,$$

So the correct model is defined by

$$f(X) = X_1\beta_1 + X_4\beta_4 + X_6\beta_6 + X_8\beta_8,$$

Table (2). values of MMAD and SD of example Two

Sample Size	Comparison Methods		$e_i \sim N(0,1)$	$e_i \sim N(1,1)$	$e_i \sim N(2,2) + N(2,2)$	$e_i \sim Lap(1,0)$	$e_i \sim x^2_{(4)}$
	n=15	BANETR	1.364(0.453)	1.223 (0.573)	1.332 (0.353)	1.165 (0.758)	1.232 (0.563)
		ANETR	1.573 (0.748)	1.473 (0.736)	1.637 (0.572)	1.342 (0.394)	1.640 (0.662)
	n=25	BANETR	1.443 (0.283)	1.234 (0.263)	1.323 (0.157)	1.439(0.231)	1.006(0.346)
		ANETR	1.647 (0.834)	1.634 (0.463)	1.839 (0.453)	1.854 (0.537)	1.538(0.782)
Small Sample	n=35	BANETR	1.362 (0.334)	1.433 (0.273)	1.245 (0.434)	1.245 (0.245)	1.234(0.456)
		ANETR	1.563 (0.673)	1.823 (0.439)	1.734 (0.664)	1.547 (0.465)	1.482(0.706)
Meddle Sample	n=45	BANETR	1.282 (0.436)	1.275 (0.055)	1.493 (0.565)	1.265 (0.346)	1.464(0.161)
		ANETR	1.453 (0.764)	1.453 (0.673)	1.745 (0.856)	1.733 (0.546)	1.845(0.456)
	n=55	BANETR	1.645 (0.453)	1.238 (0.459)	1.334 (0.264)	1.464 (0.354)	1.365(0.579)
		ANETR	1.934 (0.854)	1.852 (0.673)	1.652 (0.345)	1.655 (0.566)	1.934(0.935)
	n=65	BANETR	1.178 (0.327)	1.005 (0.075)	1.563 (0.352)	1.045 (0.164)	1.156(0.254)
		ANETR	1.465 (0.845)	1.454 (0.756)	1.745 (0.564)	1.456 (0.303)	1.458(0.846)
Large Sample	n=100	BANETR	1.045 (0.045)	1.273 (0.245)	1.354 (0.322)	1. 007(0.322)	1.256(0.233)
		ANETR	(1.212)(0.845)	(1.565)(0.372)	(1.783)(0.173)	(1.222)(0.433)	(1.475)(0.452)
	n=200	BANETR	1.435 (0.332)	1.543 (0.346)	1.697 (0.435)	1.157 (0.253)	1.235(0.633)
		ANETR	1.635 (0.732)	(1.812)(0.845)	(1.845)(0.674)	(1.676)(0.445)	(1.875)(0.343)
	n=300	BANETR	1.665 (0.553)	1.453 (0.334)	1.534 (0. 389)	(1.274)(0.341)	(1.452)(0.323)

		ANETR	(1.912)(0.675)	(1.756)(0.311)	(1.781)(0.564)	(1.771)(0.422)	(1.881)(0.706)

Table (**2**) displays MMAD and SD values as measurement for testing the quality of the estimated regression models based on the proposed methods (BENLR) and the (ENLR) under three types of sample sizes , small samples **(n=15 , n=25 , n=35)** , middle samples **(n=45 , n=55 , n=65)** , and large samples **(n=100 , n=200 , n=300)** . Obviously, the values of MMAD criterion are the smallest in the proposed methods compared with the other methods under all different type of error distributions .In addition , the SD criterion show the preference of the proposed methods under different type of sample sizes and error terms distributions .Hence , the proposed method is a promising regularization method.

3-Third example

In this example, we assumed that the true vector of the true parameter (dense)
$\boldsymbol{\beta} = (\mathbf{0.85}, \mathbf{0.85}, \mathbf{0.85}, \mathbf{0.85}, \mathbf{0.85}, \mathbf{0.85}, \mathbf{0.85}, \mathbf{0.85}, \mathbf{0.85})$ with error distributed according to standard normal
$\boldsymbol{e_i \sim N}\ (\mathbf{0}, \mathbf{1})\ , \boldsymbol{e_i \sim N}\ (\mathbf{2}, \mathbf{2})\ + \boldsymbol{N}\ (\mathbf{2}, \mathbf{2})\ , \boldsymbol{e_i \sim lap}\ (\mathbf{1}, \mathbf{0})\ , \boldsymbol{and}\ \boldsymbol{e_i \sim X^2_{(4)}}$

I generated the observations of the covariates $\mathbf{X_1}$, …. . , $\mathbf{X_9}$ from the multivariate normal $\mathbf{N_{n=9}}$ ($\mathbf{0}$, $\sum$), here $\sum$ is the **var-cov** matrix defined as $\sum_{\mathbf{ij}} = \mathbf{0.7^{|i-j|}}$. The true relationship between the predictor variables and response variable base on the above true vector is

$$f(X) = \sum_{j=1}^{9} X_j \beta_j,$$

So the correct model is defined by

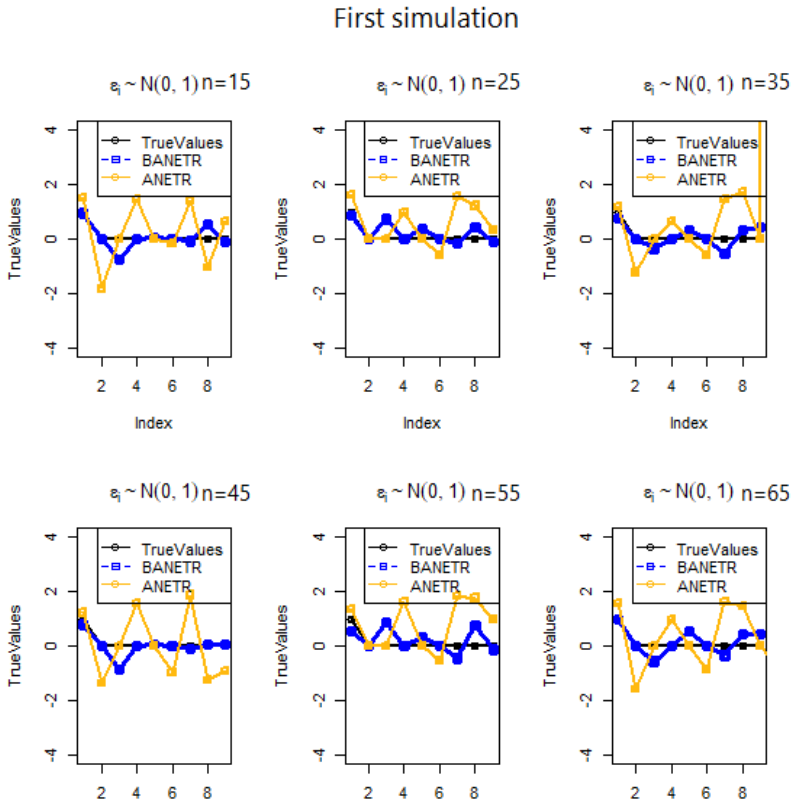
$$f(X) = \sum_{j=1}^{9} 0.85 X_j,$$

Table (3). values of MMAD and SD of example Three

Sample Size	Comparison Methods		$e_i \sim N(0,1)$	$e_i \sim N(1,1)$	$e_i \sim N(2,2) + N(2,2)$	$e_i \sim Lap(1,0)$	$e_i \sim x^2_{(4)}$
	n=15	BANETR	1.675 (0.344)	1.386 (0. 776)	1.285 (0.937)	1.930 (0.393)	1.383 (0.362)
		ANETR	1.283 (0.385)	1.495 (0.350)	1.364 (0.296)	1.696 (0.535)	1.672 (0.582)
	n=25	BANETR	1.898 (0.483)	1.292 (0.272)	1.352 (0.317)	1.782(0.562)	1.231 (0.452)
		ANETR	2.021 (0.536)	2.200 (0.652)	2.564 (0.674)	2.069 (0.682)	2.003 (0.563)
Small Sample	n=35	BANETR	1.565 (0.346)	1.654 (0.452)	1.845 (0.453)	1.456 (0.456)	1.674 (0.450)
		ANETR	2.464 (0.834)	2.065 (0.834)	2.078 (0.732)	2.004 (0.786)	2.454 (0.780)
Meddle Sample	n=45	BANETR	2.002 (0.051)	1.417 (0.683)	1.672 (0.493)	1.653 (0.385)	1.452 (0.230)
		ANETR	2.597 (0.433)	2.003 (0.792)	2.112 (0.562)	2.573 (0.840)	2.021(0.3657)

	n=55	BANETR	1.564 (0.674)	1.673 (0.564)	1.562 (0.634)	1.423 (0.427)	1.008 (0.543)
		ANETR	2.456 (0.857)	2.620 (0.407)	2.005 (0.460)	2.478 (0.096)	1.845 (0.670)
	n=65	BANETR	1.956 (0.463)	1.563 (0.194)	1.206 (0.435)	1.546 (0.354)	1.563 (0.322)
		ANETR	2.071 (0.544)	2.534 (0.342)	2.116 (0.537)	2.399 (0.653)	2.054 (0.745)
Large Sample	n=100	BANETR	1.782 (0.452)	1.765 (0.653)	1.435 (0.175)	1.534 (0.264)	1.845 (0.343)
		ANETR	2.071 (0.544)	2.564 (0.934)	2.005 (0.341)	2.071 (0.544)	2.563 (0.544)
	n=200	BANETR	1.673 (0.364)	1.830 (0.282)	1.781 (0.405)	1.530 (0.204)	1.807 (0.432)
		ANETR	2.342 (0.649)	2.023 (0.450)	2.217 (0.671)	2.115 (0.620)	2.316 (0.782)
	n=300	BANETR	1.759 (0.423)	1.673 (0.337)	1.673 (0.452)	1.867 (0.340)	1.684 (0.632)
		ANETR	2.342 (0.685)	2.125 (0.644)	2.233 (0.6754)	2.43 (0.564)	2.231 (0.875)

Table (**3**) illustrate the MMAD and SD which are the measures of quality for the estimated regression models of the proposed methods (BENLR) and the (ENLR) based on three types of sample sizes , small samples **(n=15 , n=25 , n=35)** , middle samples **(n=45 , n=55 , n=65)** , and large samples **(n=100 , n=200 , n=300)** . It is very clear that the values of MMAD criterion are the smallest values in the proposed methods compared with the other methods under all different type of error distributions. As well as, the SD criterion shows the preference of the proposed method under different type, of sample sizes and error terms distributions. Eventually, the proposed method is a promising regularization method. Figure (1) shows different plots for $e \sim N\ (0,1)$ error term distributions and different sample sizes, three lines of the parameter estimates based the proposed model (BENLR) , (ENLR) model , and the true vector of the coefficients.



$ISSN_{Online}$: 2312-9883
$ISSN_{Print}$: 1816-9171

Figure (1) parameter estimates fitted lines of example one

Clearly, the proposed model (BANETR) is a comparable and gives best fit. Where the first simulation assumed the very sparse vector $\beta = (\ 1,0,0,0,0,0,0,0,0\ )$ with black lined, the proposed model parameter estimates with blue line, and (ENLR) model parameter estimates with orange line. Hence, the blue line fits the true vector in all different plots.Also, fig (2) shows different plots for $e \sim N\ (0,1)$ error term distributions and different sample sizes for the second simulation example (sparse model) $\boldsymbol{\beta} = (\ \mathbf{1},\mathbf{0},\mathbf{0},\mathbf{1},\mathbf{0},\mathbf{1},\mathbf{0},\mathbf{1},\mathbf{0}\ )$, three lines of the parameter estimates based the proposed model (BENLR) , (ENLR) model , and the true vector of the coefficients.

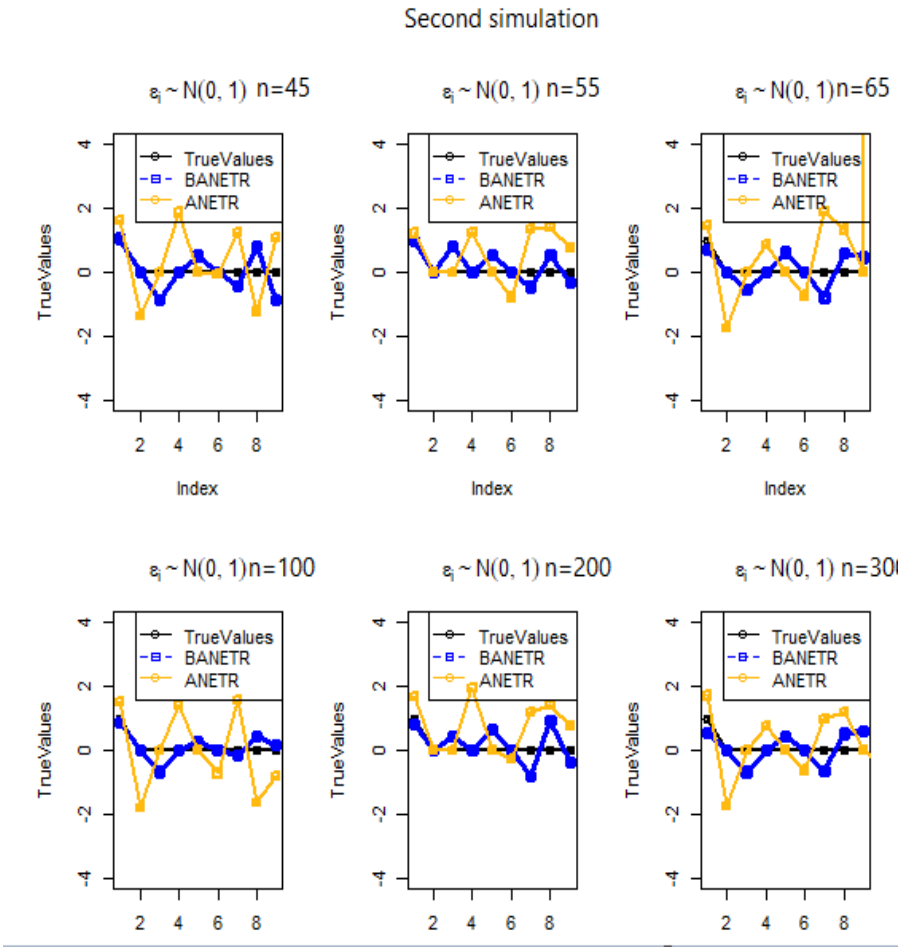


Fig (2) parameter estimates fitted lines of example two

Obviously, the proposed model (BANETR) is a comparable and gives best fit. Where the second simulation assumed the sparse vector with black lined, the proposed model parameter estimates with blue line, and (ENLR) model parameter estimates with orange line. Hence, the blue line fits the true vector in all different plots. Fig (3) shows different plots for $e \sim N\ (0,1)$ error term distributions and different sample sizes for the second simulation example (dense model) $\boldsymbol{\beta} = (\mathbf{0.85},\mathbf{0.85},\mathbf{0.85},\mathbf{0.85},\mathbf{0.85},\mathbf{0.85},\mathbf{0.85},\mathbf{0.85},\mathbf{0.85}\ )$ three lines of the parameter estimates based the proposed model (BENLR) , (ENLR) model , and the true vector of the coefficients

$ISSN_{Online}$: 2312-9883
$ISSN_{Print}$: 1816-9171

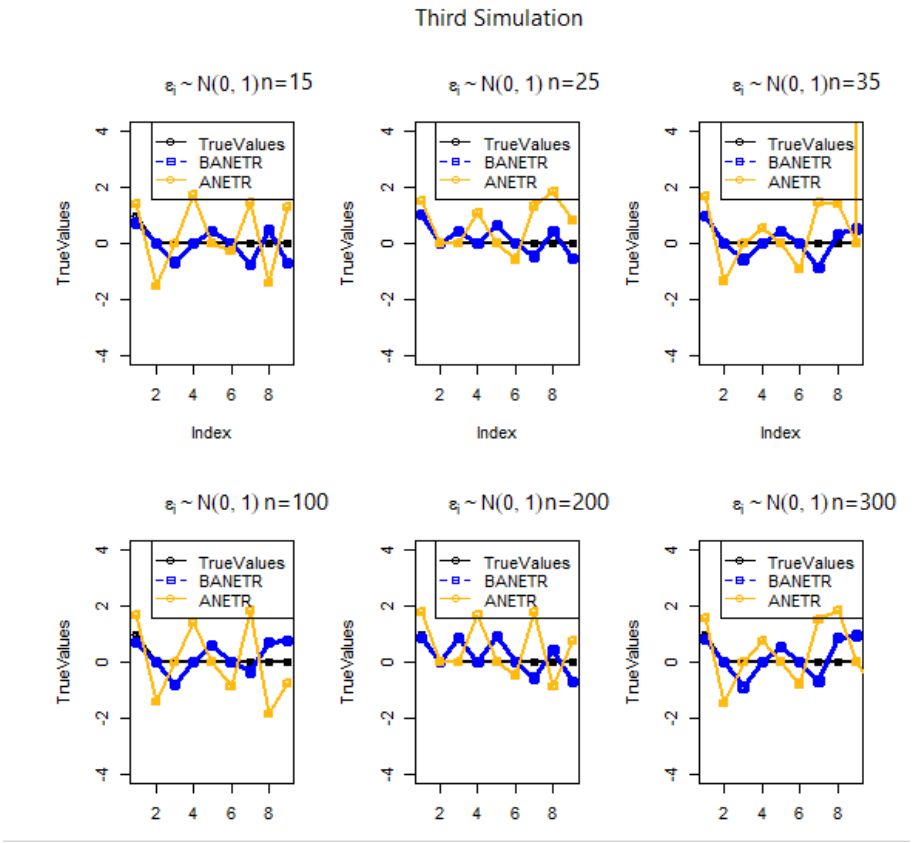


Figure (3) parameter estimates fitted lines of example three

Obviously, the proposed model (BANETR) is a comparable and gives best fit. Where the second simulation assumed the sparse vector $\boldsymbol{\beta} = (\mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{1}, \mathbf{0}, \mathbf{1}, \mathbf{0}, \mathbf{1}, \mathbf{0})$, with black lined, the proposed model parameter estimates with blue line, and (ENLR) model parameter estimates with orange line. Hence, the blue line fits the true vector in all different plots.

5-Conclusions

New scale mixture of Rayleigh distribution mixing with normal distribution have developed as the prior distribution of the Laplace distribution. Consequently, we produced new Bayesian hierarchical model for elastic net in linear regression. Gibbs sampler algorithm has developed to examine the convergence of the proposed posterior distributions. Some simulation scenarios have implemented based on the proposed method. The result of simulation shows that the proposed method clearly outperforms the other method from the variable selection procedure view.

References

[1] James, G.,Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer Publishing Company, Incorporated, New York.

[2] Hoerl, A. E., and Kennard, R. W. (1970a). Ridge regression: Applications to nonorthogonal problems. *Technometrics, 12*(1), 69-82.

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

[4] 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.

[5] Mallows, C. L., & Sloane, N. J. (1973). An upper bound for self-dual codes. Information and Control, 22(2), 188-200.

[6] Akaike, H. (1973), "Information Theory and an Extension of Maximum Likelihood Principle," in Second International Symposium on Information Theory (B. N. Petrov and F. Caski, Eds.) Akademia Kiado, Budapest, 267-28 1.

[7] Schwarz, G. (1978), "Estimating the Dimensions of a Model," Annals of Statistics, 121,461-464.

[8] Zou, P. X., Zhang, G., & Wang, J. (2007). Understanding the key risks in construction projects in China. International journal of project management, 25(6), 601-614.

[9] Hocking, R.R. (1976) The Analysis and Selection of Variables in Linear Regression. Biometrics, 32, 1-50.

[10] Efroymson, M.A. (1960). "Multiple Regression Analysis," In: A. Ralston and H. S. Wilf, Eds., Mathematical Methods for Digital Computers, John Wiley, New York.

[11] George, E.I. and McCulloch, R.E..(1993).Variable Selection Via Gibbs Sampling. Journal of the American Statistical Association, Vol. 88, No. 423. (Sep., 1993), pp. 881-889.

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

[13] Draper, N. R., & Smith, H. (1998). Applied regression analysis (3rd ed.). New York: John Wiley & Sons.

[14] Li, Q., & Lin, N. (2010). The Bayesian elastic net. Bayesian analysis, 5(1), 151-170.

[15] Park, T., & Casella, G. (2008). The bayesian lasso. Journal of the American Statistical Association, 103(482), 681-686. Choi, W. W., Weisenburger, D. D., Greiner, T. C., Piris, M.

[16] Mallick, H., & Yi, N. (2014). A new Bayesian lasso. Statistics and its interface, 7(4), 571-582.

[17] Flaih, A.N, Alshaybawee, and Al husseini F.H, (2020a). sparsity via new Bayesian Lasso . Periodicals of Engineering and Natural Sciences , vol.8 , is suc 1, 345-359.

[18] Flaih, A. N., Al-Saadony, M. F., & Elsalloukh, H. (2020b). New Scale Mixture for Bayesian Adaptive Lasso Tobit Regression. Al-Rafidain University College For Sciences, (46).

[19] Almusaedi, M.H.M. and Flaih, A.N. (2021a). Simulation Study for Penalized Bayesian Elastic Net Quantile Regression. Al-Qadisiyah journal of pure science. P- ISSN 1997-2490 , E- ISSN 2411-3514.

[20] Almusaedi, M.H.M. and Flaih, A.N.(2021b). Employing the Bayesian Elastic Net in Quantile Regression with an Application. AL-Qadisiyah Journal For Administrative and Economic sciences. (Accepted for publishing).

[21] ALsafi, M.R. and Flaih, A.N. (2021). Sparsity in Bayesian Elastic Net in Tobit Regression with an Application. AL-Qadisiyah Journal For Administrative and Economic sciences. (Accepted for publishing).

[22] Chhikara, R. (1988). The Inverse Gaussian Distribution: Theory: Methodology, and Applications (Vol. 95). CRC Press.

[23] Park, T., & Casella, G. (2008). The bayesian lasso. Journal of the American Statistical Association, 103(482), 681-686.

.