ORMANCE OF SHRINKAGE METHODS – A MONTE CARLO STUDY

GAFAR MATANMI OYEYEMI[1], EYITAYO OLUWOLE OGUNJOBI[2],
& ADEYINKA IDOWU FOLORUNSHO[3]
[1]Department of Statistics, University of Ilorin.
[2]Department of Mathematics and Statistics,
The Polytechnic Ibadan, Adeseun Ogundoyin Campus, Eruwa.
[3]Department of Mathematics and Statistics, Osun State Polytechnic Iree, Nigeria
E-mail: gmoyeyemi@gmail.com

## Abstract
*Multicollinearity has been a serious problem in regression analysis, Ordinary Least Squares (OLS) regression may result in high variability in the estimates of the regression coefficients in the presence of multicollinearity. Least Absolute Shrinkage and Selection Operator (LASSO) methods is a well established method that reduces the variability of the estimates by shrinking the coefficients and at the same time produces interpretable models by shrinking some coefficients to exactly zero. We present the performance of LASSO -type estimators in the presence of multicollinearity using Monte Carlo approach. The performance of LASSO, Adaptive LASSO, Elastic Net, Fused LASSO and Ridge Regression (RR) in the presence of multicollinearity in simulated data sets are compared using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) criteria. A Monte Carlo experiment of 1000 trials was carried out at different sample sizes n (50, 100 and 150) with different levels of multicollinearity among the exogenous variables (ρ = 0.3, 0.6, and 0.9). The overall performance of Lasso appears to be the best but Elastic net tends to be more accurate when the sample size is large.*

Keywords: Multicollinearity, Elastic net, Ridge, Adaptive Lasso, Fused Lasso.

## Introduction
Multicollinearity can cause serious problem in estimation and prediction when present in a set of predictors. Traditional statistical estimation procedures such as Ordinary Least Squares (OLS) tend to perform poorly, have high prediction variance, and may be difficult to interpret (Brown, 1993) i.e. because of its large variance's and covariance's which means the estimates of the parameters tend to be less precise and lead to wrong inferences (Muhammad, Maria & Muhammad, 2013). In such situations it is often beneficial to use shrinkage i.e. shrink the estimator towards zero vector, which in effect involves introducing some bias so as to decrease the prediction variance, with the net result of reducing the mean squared error of prediction, they are nothing more than penalized estimators, due to estimation there is objective functions with the addition of a penalty which is based on the parameter. Various assumptions have been made in the literature where penalty of $\ell_1$ - norm, $\ell_2$- norm or both $\ell_1$ and $\ell_2$ which stand as tuning parameters ($\lambda$ ) were used to influence the parameter estimates in order to minimize the effect of the collinearity. Shrinkage methods are popular among the researchers for their theoretical properties e.g. parameter estimation.

Over the years, the LASSO - type methods have become popular methods for parameter estimation and variable selection due to their property of shrinking some of the model coefficients to exactly zero see (Tibshirani, 1996), (Xun & Liangjun, 2013). Tibshirani, (1996) proposed a new shrinkage method Least Absolute Shrinkage and Selection Operator (LASSO) with tuning parameter $\lambda \geq 0$ which is a penalized method, (Knight, & Fu, 2000) for the first systematic study of the asymptotic properties of Lasso – type estimators (Xun & Liangjun, 2013). The LASSO shrinks some coefficients while setting others to exactly zero,

and thus theoretical properties suggest that the LASSO potentially enjoys the good features of both subset selection and ridge regression. Frank and Friedman (1993) had earlier proposed Ridge regression which minimizes the Residual Sum of Squares subject to constraint with $\gamma \geq 0$.

Frank and Friedman (1993) argued that the optimal choice of parameter $\lambda$ yields reasonable predictors because it controls the degree of precision for true coefficient of $\beta$ to aligned with original variable axis direction in the predictor space. Fan and Li (2001) Introduced the Smoothing Clipped Absolute Deviation (SCAD) which penalized Least Square estimate to reduce bias and satisfy certain conditions to yield continuous solutions. Hoerl and Kennard (1970a) was first to propose Ridge Regression which minimizes the Residual Sum of Squares subject to constraint with $\gamma = 2$, thus regarded as $\ell_2$ - norm. Efron, Hastie, Johnstone and Tibshirani (2004) developed Least Angle Regression Selection (LARS) for a model selection algorithm (Wang & Leng, 2008), Wei and Huang (2010) study the properties of adaptive group Lasso. In 2006, Yuan and Lin, (2006) proposed a Generalization of LASSO and other shrinkage methods include Dantzig Selector with Sequential Optimization, (DASSO) (James, Radchenko, & Lv, 2009), Elastic Net (Zou & Hastie, 2005), Variable Inclusion and Selection Algorithm, (VISA) (Radchenko & James, 2008), Adaptive LASSO (Zou, 2006) among others.

LASSO-type estimators are the techniques that are often suggested to handle the problem of multicollinearity in regression model. More often than none, Bayesian simulation with secondary data has been used. When the ordinary least squares are adopted there is tendency to have poor inferences, but with LASSO-type estimators which have recently been adopted may still come with its shortcoming by shrinking important parameters, we intend to examine how these shrink parameters may be affected asymptotically. However, the performances of other estimators have not been exhaustively compared in the presence of all these problems. Moreover, the question of which estimator is robust in the presence of a LASSO-type estimators of these problems have not been fully addressed. This is the focus of this research work.

## Material and method

Consider a simple least squares regression model.

$$y_i = x_i' \beta + e_i, \qquad (1)$$

where $x_i$ are exogenous, $e_i$ are $i.i.d.$ $i = 1, \cdots n$, random variable with mean zero and finite variance $\sigma^2$. $\beta$ is $p \times 1$ vector. Suppose $\beta$ takes the largest possible dimension , in other words the number of regressors may be at most $p$, but the true p is somewhere between 1 and $p$. The issue here is to come up with the true model and estimate it at the same time.

The least squares estimate without model selection is

$$\hat{\beta}_{LS} = (\sum_{i=1}^{n} x\, x_i')^{-1} \ (\sum_{i=1}^{n} x_i y_i) \ .$$

with $p \times 1$ estimates.

Shrinkage estimators are not that easy to calculate as ordinary least squares. Thus the objective functions for the shrinkage estimators:

$$\hat{\beta} = \operatorname{argmin}_{\beta} \left[\sum_{i=1}^{n} (y_i - x_i' \beta)^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|^{\gamma}\right] \qquad (2)$$

Where $\lambda_n$ is a tuning parameter (for penalization), it is a positive sequence, and $\lambda > 0$ .$\lambda_n$ will not be estimated, and $\gamma$ will be specified by us. The objective function consists of 2 parts, the first one is the LS objective function part, and then the penalty factor.

Thus, taking the penalty part only

$$\lambda_n \sum_{j=1}^{p} |\beta_j|^{\gamma}$$

If $\lambda_n$ is going to infinity or to a constant, the values of $\beta$ that minimizes that part should be the case that $\beta = 0_p$ . We get all zeros if we minimize only the penalty part. So the penalty part will shrink the coefficients to zero. This is the function of the penalty.

## Ridge Regression (RR)

Ridge Regression (RR) by (Hoerl & Kennard, 1970b) is ideal if there are many predictors, all with non-zero coefficients and drawn from a normal distribution (Friedman, Hastie & Tibshirani, 2010). In particular, it performs well with many predictors each having small effect and prevents coefficients of linear regression models with many correlated variables from being poorly determined and exhibiting high variance. RR shrinks the coefficients of correlated predictors equally towards zero. For example, given k identical predictors, each would get identical coefficients equal to $(\frac{1}{k})th$ the size that any one predictor would get if fit singly (Friedman, Hastie & Tibshirani, 2010). Ridge regression does not force coefficients to vanish and hence cannot select a model with only the most relevant and predictive subset of predictors. The ridge regression estimator solves the regression problem in [17] using $\ell_2$ penalized least squares:

$$\hat{\beta}\ (ridge) = \underset{\beta}{argmin} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \qquad (3)$$

Where $\|y - X\beta\|_2^2 = \sum_{i=1}^{n}(y_i - x_i^T\beta)^2$ is the $\ell_2$–norm (quadratic) loss function (i.e. residual sum of squares), $x_i^T$ is the $i-th\ row$ of X, $\|\beta\|_2^2 = \sum_{j=1}^{p}\beta_j^2$ is the $\ell_2$ –norm penalty on $\beta$, and $\lambda \geq 0$ is the tuning parameter (penalty, regularization, or complexity) which regulates the strength of the penalty (linear shrinkage) by determining the relative importance of the data-dependent empirical error and the penalty term. The larger the value of $\lambda$, the greater is the amount of shrinkage. As the value of $\lambda$ is dependent on the data it can be determined using data-driven methods, such as cross-validation. The intercept is assumed to be zero in equation (3) due to mean centering of the phenotypes.

## Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO regression methods are widely used in domains with massive datasets, such as genomics, where efficient and fast algorithms are essential (Friedman, Hastie & Tibshirani, 2010). The LASSO is, however, not robust to high correlations among predictors and will arbitrarily choose one and ignore the others and break down when all predictors are identical (Friedman, Hastie & Tibshirani, 2010). The LASSO penalty expects many coefficients to be close to zero, and only a small subset to be larger (and nonzero).

The LASSO estimator uses the $\ell_1$ penalized least squares criterion to obtain a sparse solution to the following optimization problem:

$$\hat{\beta}\ (lasso) = \underset{\beta}{argmin} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1, \qquad (4)$$

Where $\|\beta\|_1 = \sum_{i=}^{p}|\beta_j|$ is the $\ell_1$ -norm penalty on $\beta$, which induces sparsity in the solution, and $\lambda \geq 0$ is a tuning parameter.

The $\ell_1$ penalty enables the LASSO to simultaneously regularize the least squares fit and shrinks some components of $\hat{\beta}\ (lasso)$ to zero for some suitably chosen $\lambda$. The cyclical coordinate descent algorithm, (Friedman, Hastie & Tibshirani, 2010), efficiently computes the entire lasso solution paths for $\lambda$ for the lasso estimator and is faster than the well-known

LARS algorithm (Efron, Hastie, Johnstone & Tibshirani, 2004). These properties make the lasso an appealing and highly popular variable selection method.

### Fused LASSO

To compensate the ordering limitations of the LASSO, (Tibshirani, Saunders, Rosset, Zhu & Knight, 2005) introduced the fused LASSO. The fused LASSO penalizes the $\ell_1$ -norm of both the coefficients and their differences:

$$\hat{\beta}_F = \arg\min_{\beta}(\bar{y} - X\beta)'(\bar{y} - X\beta) + \lambda_1 \sum_{j=1}^{p} |\beta_J| + \lambda_2 \sum_{j=1}^{p} |\beta_J - \beta_{J-1}| \quad (5)$$

where $\lambda_1$ and $\lambda_2$ are tuning parameters. They provided the theoretical asymptotic limiting distribution and a degrees of freedom estimator.

### Elastic Net

Zou and Hastie (2005) proposed the elastic net, a new regularization of the LASSO, for the unknown group of variables and for the multicollinear predictors. The elastic net method overcomes the limitations of the LASSO method which uses a penalty function based on

$$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_J|$$

Use of this penalty function has several limitations. For instance, in the "large *p*, small *n*" case the LASSO selects at most n variables before it saturates. Also if there is a group of highly correlated variables, then the LASSO tends to select one variable from a group and ignore the others. To overcome these limitations, the elastic net adds a quadratic part to the penalty($\|\beta\|^2$), which when used alone is ridge regression (known also as Tikhonov regularization). The elastic net estimator can be expressed as

$$\hat{\beta}_{EN} = \arg\min_{\beta}(\bar{y} - X\beta)'(\bar{y} - X\beta) + \lambda_1 \sum_{j=1}^{p} |\beta_J| + \lambda_2 \sum_{j=1}^{p} |\beta_J|^2 \quad (7)$$

where $\lambda_1$ and $\lambda_2$ are tuning parameters. As a result, the elastic net method includes the LASSO and ridge regression: in other words, each of them is a special case where $\lambda_1 = \lambda, \lambda_2 = 0$ *or* $\lambda_1 = 0, \lambda_2 = \lambda$.

### Adaptive LASSO

Fan and Li (2001) showed that the LASSO can perform automatic variable selection but it produces biased estimates for the large coefficients. Zou (2006) introduced the adaptive LASSO estimator as

$$\hat{\beta}_{AL} = \arg\min_{\beta}(\bar{y} - X\beta)'(\bar{y} - X\beta) + \lambda \sum_{j=1}^{p} \widehat{w_J} |\beta_J| \quad (8)$$

with the weight vector $\hat{w} = 1/|\hat{\beta}|$ where $\hat{\beta}$ is a $\sqrt{n}$ consistent estimator such as $\hat{\beta}$(OLS) and

$\gamma > 0$. where $\widehat{\omega_j}(j = 1, \ldots, p)$ are the adaptive data-driven weights, which can be estimated by,

$\widehat{\omega_j} = \left(|\hat{\beta}_j^{ini}|\right)^{-\gamma}$, where $\lambda$ is a positive constant and $\hat{\beta}^{ini}$ is an initial consistent estimator of $\beta$ obtained through least squares or ridge regression if multicollinearity is important (Zou, 2006). The optimal value of $\lambda > 0$ and $\lambda$ can be simultaneously selected from a grid of values, with values of $\lambda$ selected from {0.5, 1, 2}, using two-dimensional cross-validation (Zou, 2006). The weights allow the adaptive LASSO to apply different amounts of shrinkage to different coefficients and hence to more severely penalize coefficients with small values. The flexibility introduced by weighting each coefficient differently corrects for the undesirable tendency of the lasso to shrink large coefficients too much yet insufficiently

shrink small coefficients by applying the same penalty to every regression coefficient (Zou, 2006).

## Monte Carlo Study

In this section we carried out simulation to examine the finite sample performance for LASSO, Adaptive LASSO, Elastic LASSO, Fused LASSO and Ridge Regression using AIC and BIC.

We infected the data with multicollinearity by generating sets of variables of sample sizes n (n = 50, 100 and 150) using normal distribution respectively. The level of multicollinearity among the variables are small (r = 0.1 – 0.3), mild ( r = 0.4 – 0.6) and serious (r = 0.7 – 0.9). Each simulation was repeated 1000 times for consistency using R package.

Table 1: Mean AIC and BIC of the fitted model using the five methods

| | | Ridge Regression | | Adaptive | | Elastic Net | | Fused | | LASSO | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| N | R | AIC | BIC | AIC | BIC | AIC | BIC | AIC | BIC | AIC | BIC |
| 50 | 0.1 – 0.3 | -65.62 -65.68 | -65.39 -65.45 | -65.57 - | -65.35 - | -65.62 -65.63 | -65.39 -65.40 | - 65.64 | - 65.41 | -65.62 -65.56 | -65.39 -65.34 |
| | 0.4 – 0.6 | -65.60 | -65.37 | 65.79 -65.69 | 65.56 -65.36 | -65.64 | -65.41 | -65.63 - | -65.41 - | -65.56 | .65.33 |
| | 0.7 – 0.9 | | | | | | | 65.79 | 65.56 | | |
| 100 | 0.1 – 0.3 | - 67.48 | - 67.32 | -67.27 -64.64 | -67.12 -64.49 | -67.20 -64.65 | -67.04 -64.49 | -67.35 - | -67.20 - | -66.82 -64.69 | -66.67 -64.53 |
| | 0.4 – 0.6 | -64.58 -67.44 | -64.42 -67.29 | - 69.61 | - 67.46 | -67.39 | -67.23 | 64.87 -68.14 | 64.72 -67.99 | -68.03 | -67.87 |
| | 0.7 – 0.9 | | | | | | | | | | |
| 150 | 0.1 – 0.3 | -65.22 -65.79 | -65.09 -65.67 | - 65.30 | - 65.18 | -65.12 -65.77 | -64.99 -65.65 | -65.15 - | -65.03 - | -65.19 -65.75 | -65.07 -65.62 |
| | 0.4 – 0.6 | -65.50 | -65.38 | -65.64 - | -65.52 - | -65.48 | -65.36 | 65.94 -65.58 | 65.82 -65.46 | -65.66 | -65.54 |
| | 0.7 – 0.9 | | | 65.73 | 65.61 | | | | | | |

Table 2: Summary of the result

| Sample size (n) | r | Best |
|---|---|---|
| 50 | Low | Lasso |
| 100 | | Adaptive Lasso |
| 150 | | Elastic Net |
| 50 | Medium | Elastic Net |
| 100 | | Lasso |
| 150 | | Lasso |
| 50 | High | Lasso |
| 100 | | Lasso |
| 150 | | Elastic Net |

Table 1 shows both the AIC and BIC of the fitted model using the five methods while Table 2 presents the summary of Table 1. It is of interest to note that both criteria agreed in selecting the best method in all the cases considered. It can be observed that LASSO performed better at all the three levels of multicollinearity (small sample with low multicollinearity, medium sample size with medium multicollinearity and at small and medium sample sizes with high multicollinearity).

Elastic Net competed favourably with LASSO because it was also better at all levels of multicollinearity (high sample size with low multicollinearity, small sample size with medium multicollinearity and large sample size with high multicollinearity). Adaptive LASSO performed best only with medium sample size at low multicollinearity. Generally, it can be seen that LASSO performs best when the correlation is high but Elastic net tend to be more accurate when the sample size *n* is large. LASSO appears to have best overall performance among all the five methods. Therefore, the LASSO method is more suitable due to its significant advantage over others.

## Conclusion

We have considered Lasso type estimators in the presence of multicollinearity in linear regression model. The Ordinary Least Squares (OLS) method brings about poor parameters estimate and produce wrong inferences. Lasso type estimators are more stable and provide performances better than OLS approach of parameters estimation in the case of correlated predictors and produce consistent solution. Elastic net performed better for large sample size especially for high value of multicollinearity. While the LASSO is better for medium and high level of multicollinearity. Performances of both Fused and Ridge regression were poor compared with the other methods considerd.

## References


Brown, J. M. (1993). *Measurement, regression and calibration.* UK: Oxford University Press.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression (with discussion). *Annals of Statistics*, 32, 407 - 451.

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

Frank, I. E. & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics,* 35, 109-148.

Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization paths for generalized linear model via coodinate descent. *Journal of Statistical Software,* 33, 1 – 22

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

Hoerl, A. E. & Kennard, R. W. (1970b). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12, 69– 82.

James, G. M., Radchenko, P. & Lv, J. (2009). Connections between the dantzing selector and Lasso. *Journal of the Royal Statistical Society, Series B*, 71,121- 142.

Knight, K. & Fu, W. (2000). Asymptotic for Lasso-type estimators. *Annals of Statistics*, 28, 1356 - 1378.

Muhammad, I., Maria, J. & Muhammad, A. R. (2013). Comparison of shrinkable regression for remedy of multicollinearity problem. *Middle – East Journal of Scientific Research*, 14(4), 570 – 579.

Radchenko, P. & James, G. (2008). Variable inclusion and shrinkage algorithms. *Journal of American Statistical Association,* 103, 1304 – 1315.

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005). Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society. Series B* 67, 91-108

Wang, H. & Leng, C. (2008). A note of adaptive group Lasso. *Computational Statistics and Analysis*, 52, 5277 - 5286.

Wei, F. & Huang, J. (2010). Consistent group selection in high dimensional linear regression. *Bernoulli,* 16,1369 – 1384.

Xun, L. & Liangjun, S. (2013). *Shrinkage estimation of dynamic panel data models with interactive fixed effects.* Singapore Management University.

Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B.,* 68, 49-67.

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

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* 101, 1418 - 1429.