Scholars Journal of Physics, Mathematics and Statistics

OPEN ACCESS

Abbreviated Key Title: Sch J Phys Math Stat
ISSN 2393-8056 (Print) | ISSN 2393-8064 (Online)
Journal homepage: https://saspublishers.com

On Using the Ridge Regression to solve the Multi-collinearity Problem

Ahmed M. Mami[1*], Abdelbaset Abdalla[1], Eisay H. Bin Ismaeil[1]

[1]Department of Statistics, Faculty of Science, University of Benghazi, Benghazi, Libya

DOI: 10.36347/sjpms.2021.v08i08.003 | **Received:** 14.09.2021 | **Accepted:** 23.10.2021 | **Published:** 27.10.2021

***Corresponding author:** Ahmed M. Mami

Abstract | **Original Research Article**

Multicollinearity is a phenomenon when two or more predictors are highly correlated that leading the matrix $\boldsymbol{X^T X}$ to be singular, and hence identifying the least squares estimates will encounter numerical problems. In this work, we proposed two remedial measures for handling severe multicollinearity in the least-squares estimation, namely, the Ridge regression, and Least Absolute Shrinkage and Selection Operator (Lasso). A simulation study was conducted to compare the two proposed methods under different settings. These settings include different sample sizes, a variety of several explanatory variables used in the model along with the difference in the degree of correlation that exists among the explanatory variables, and finally the dependency of the error terms on the normal or non-normal distributions. This simulation study is novel in the field of Shrinkage Estimators, also may increase the effective capabilities of Ridge Regression, and several interesting results have been achieved.

Keywords: Multi-collinearity, Shrinkage estimators, Ridge Regression, and LASSO.

1. INTRODUCTION

Regression analysis is a statistical method for studying such a relationship that exists between one dependent variable, y, and the explanatory variables $(x_1, x_1, \dots, x_p)$. It is probably one of the oldest topics in the area of mathematical statistics dating back to about two centuries ago. The traditional method of parameters estimation for the linear regression models is the Ordinary Least Squares Estimation. Four of the major problem areas in the least Squares Analysis relate to failures of the basic assumptions, they are namely: Non-normality Problem, Heterogeneous Variances Problem, Correlated Errors Problem, Collinearity Problem (Baltagi 2001). Although, an alternative to Least Squares regression when the assumptions are not satisfied is known as the Robust Regression. Robust Regression refers to a general class of statistical techniques designed to reduce the sensitivity of the estimates to failures in the assumptions of the parametric model. A Robust Regression procedure would decrease the impact of such errors by reducing the weight given to large residuals. This can be done by minimizing the sum of absolute residuals, instead of the sum of squared residuals (Huber (1981) and Hampel *et al*., (1986).

1.1 The Collinearity Problem

The $\boldsymbol{X}$ matrix contains the explanatory variables and may cause singularity when some linear combinations of the columns of X are exactly equal to zero. It comes more obvious when the least-squares analysis is computed because the unique solution of $(X^tX)^{-1}$does not exist. The difficulties that arise from X being nearly singular are known as the collinearity problem. The impact of collinearity on least squares is very serious if the purpose is to estimate the regression coefficients or if the purpose is to identify the important variables involved in the process. The estimates of the regression coefficients can differ greatly from the parameters they are estimating, even to the extent of having an opposite sign. Moreover, the collinearity allows important variables to be replaced in the model with related variables that are involved in the near singularity. Therefore, the regression analysis provides small suggestions of the relative importance of the explanatory variables.

In this paper, our prime interest is to handle the collinearity problem when estimating the coefficients for linear regression models.

There are several ways for near-singularity to emerge:

1. A bad mathematical model that puts restrictions on explanatory variables that forces them to add to a constant will generate collinearity.
2. Explanatory variables of the application may show near-linear dependencies because of the biological, physical, or chemical restrictions.
3. Insufficient sample size may create data in which the near-linear dependencies are an artifact of the data collection process (Yan and Su 2005).

It is not easy to identify the origin of the collinearity problem, but it is extremely important to understand its nature as much as possible. Having known the nature of the collinearity problem will always help to determine its origin and, in turn, find suitable ways of handling the problem and of interpreting the regression results.

1.2 Introduction to Biased Regression

Biased regression refers to these methods of regression in which unbiasedness is no longer an essential condition. Thus, the Biased regression methods have been recommended as a possible solution to the collinearity problem. The motivation behind the biased regression methods is based on the possibility for obtaining estimators that are very close on average, to the parameter being estimated other than those obtained using the least-squares estimators. The MSE is considered to be the best measure of averaging the "nearness" of an estimator to the parameter being estimated.

If $\tilde{\theta}$ is a biased estimator having a smaller mean squared error than an unbiased estimator $\hat{\theta}$, the MSE of $\tilde{\theta}$ can be defined as

$$\text{MSE}(\tilde{\theta}) = E(\tilde{\theta} - \theta)^2 \quad (1)$$

Recall that the variance of an estimator $\tilde{\theta}$ can also be defined as

$$\text{Var}(\tilde{\theta}) = E\left[\tilde{\theta} - E(\tilde{\theta})\right]^2 \quad (2)$$

That means, the MSE is computing the average squared deviation of the estimator from the parameter being estimated, whereas the variance is computing the average squared deviation of the estimator from its expectation. If the estimator is unbiased, then $E(\tilde{\theta}) = \theta$ and $\text{MSE}(\tilde{\theta}) = \sigma^2(\tilde{\theta})$. If the estimator is biased, then the MSE is equal to the variance of the estimator plus the square of its bias, where $\text{Bias}(\tilde{\theta}) = E(\tilde{\theta}) - \theta$. The biased estimator can obtain a variance that is sufficiently smaller than the variance of an unbiased estimator to compensate for the bias introduced (Hoerl et al 1975). Therefore, it may be possible to find an estimator for which the sum of its squared bias and its variance (*i.e.* the MSE) is smaller than the variance of the unbiased estimator. Many biased regression methods have been proposed as solutions to the collinearity problem. Stein shrinkage (Stein(1960)), Ridge Shrinkage Regression (Hoerl and Kennard (1970), the LASSO was proposed by Tibshirani (1996).

In this paper, we will address one of the problems which is the Multicollinearity that indicates strong correlations among some of the explanatory variables. We suggest in this paper two methods in what so-called Shrinkage estimators of Ridge regression and LASSO. The Ridge Shrinkage Regression is trying to solve the Multicollinearity problem by reducing the severity of the phenomenon, and this is at the expense of the bias of feature estimates. While the LASSO method works to end or cancel "delete" variables most affected by the linear correlation problem and at the expense of shrinking explanatory variables.

The statistical procedures to be presented in this paper can be formulated as usual regression analysis. These procedures differ in the **X**-matrix working and how that matrix is determined. Whatever the **X**-matrix used, there will be a set of regression coefficients. Two suggestions have been offered for how to control the scale of regression coefficients, which are:

1. The L_1–penalty which means constraining the sum of the absolute values of the regression coefficients to be less than some constant C.
2. The L_2–penalty which means constraining the sum of the squared regression coefficients to be less than some constant *C*.

Both suggestions lead to "shrinkage methods. When shrinkage is applied to usual regression estimates there can be, as noted above, two goals. First, one might be interested in model selection. The *lasso* can provide useful alternatives to usual model selection procedures. Second, one might be interested in striking a good balance between the bias and the variance, the Ridge Regression is then used.

A Simulation study is conducted to make comparisons between the two suggested methods under different settings.

2. METHODOLOGY

In this section, some necessary groundwork will be laid concerning proposed remedial measures for handling severe multicollinearity that will be used in this paper.

2.1 Shrinkage Estimators

Let the explanatory variables used in this study be arranged in matrix form which we call the X-matrix. The procedures to be presented in this section can be formulated as a straightforward extension to regression analysis. These procedures differ in X-matrix working and also how that X-matrix is determined. Whatever the X-matrix used, there will be equal to the number of regression coefficients. The larger the absolute value of these coefficients the more the fitted values can vary. If

the regression coefficient is equal to zero, the fitted values will be a straight line (*i.e*. parallel to the *x*-axis, positioned at the unconditional mean of the response). As the regression coefficient gets increasing, the resulting step function will have a step of increasing size. The fitted line becomes more irregular. Generally, the potential for irregularity fitting is greater as the regression coefficients increase (Hoerl *et al*., 1975). Two popular suggestions have been offered for how to control the magnitude of regression coefficients (as stated in section 1.2), there are:

However, the smaller the value of C is, the smaller the sum. As well as the smaller the sum is, the smaller the regression coefficients in magnitude. These two constraints lead to so-called shrinkage methods. Simply, the regression coefficients have been shrunk toward zero, making the fitted values more homogeneous. Therefore, the main goal is to introduce a small amount of bias into the computed regression coefficients in trade for a considerable amount of reduction in their variance (Baltagi 2001).

2.2 Introduction to Ridge Shrinkage Regression

In order to motivate the theoretical development of the Ridge Shrinkage Regression estimator, take a closer look at the mean squared error of the least-squares estimator of $\boldsymbol{\beta}$

$$\text{MSE}(\widehat{\boldsymbol{\beta}}) = \text{E}||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2 \quad (3)$$

Remember that the MSE is usually used as a measure for assessing the quality of estimation, which consists of two parts: the squared bias and the variance, and can be written in the following form:

$$E\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = \sum_j E(b_j - \beta_j)^2 = \sum_j \{E(b_j) - \beta_j\}^2 + \sum_j Var(b_j) \quad (4)$$

The Gauss-Markov theorem states that the least-squares approach achieves the smallest variance among all unbiased linear estimates. Although, the minimum MSE is not necessarily guaranteed. To make a better understanding of different types of shrinkage estimators, let $\widehat{\boldsymbol{\beta}}^{LS}$ denote the ordinary least squares estimator.

The multiple linear regression model is $y = \text{X}\beta + \varepsilon$

The estimator $\hat{\beta}^{LS} = (\text{X}^t\text{X})^{-1}\text{X}^t y$ is an unbiased estimator of β in addition,

$\text{E}(\hat{\beta}^{LS}) = \beta$ and $Cov(\hat{\beta}^{LS}) = \sigma^2.(\text{X}^t\text{X})^{-1}$

We have

$\text{MSE}(\hat{\beta}^{LS}) = \text{E}\|\hat{\beta}^{LS}\|^2 - \|\beta\|^2.$

$= \text{tr}\{\sigma^2(\text{X}^t\text{X})^{-1}\} = \sigma^2.tr\{(\text{X}^t\text{X})^{-1}\}$ (5)

Therefore, by rearranging (5), we get

$E\left(\|\hat{\beta}^{LS}\|^2\right) = \|\beta\|^2 + \sigma^2.tr\{(\text{X}^t\text{X})^{-1}\}$ (6)

Because of the ill-conditioned in X^tX, the resultant least-square estimate of $\hat{\beta}^{LS}$would be large in length$\|\hat{\beta}^{LS}\|$ and related to large standard errors. As well, this large variation would lead to the poor model prediction.

The Ridge Shrinkage Regression is a constrained type of least squares. It solves the estimation problem by producing a biased estimator, however, with small variances (Weisberg 2005).

2.3 Theoretical Development of Ridge Shrinkage Estimator

For any least squares estimator $\hat{\beta}$, the least-squares criterion can be rewritten as its minimum, reached at$\|\hat{\beta}^{LS}\|$. The quadratic form in b:

$\text{Q(b)} = \|y - X\hat{\beta}^{LS} + X\hat{\beta}^{LS} - Xb\|^2$

$= (y - X\hat{\beta}^{LS})^t(y - X\hat{\beta}^{LS}) + (\text{b} - \hat{\beta}^{LS})^t X^t X\,(\text{b} - \hat{\beta}^{LS})$

$= Q_{min} + \phi(b)$ (7)

Contours for each constant of the quadratic form $\phi(b)$ are hyperellipsoids centered at the ordinary LSE $\hat{\beta}^{LS}$.It is reasonable to expect from (7) that, if one moves away from Q_{min}, the movement is in a direction that shortens the length of $\hat{\beta}$.

In Ridge Shrinkage Regression, the optimization problem can be defined as:

minimizing $\|\beta\|^2$subject to $(\beta - \hat{\beta}^{LS})^t X^t X\,(\beta - \hat{\beta}^{LS}) = \phi_0$ (8)

For some constant ϕ_0. The imposed constrain guarantees a reasonably small residual sum of squares $\text{Q}(\beta)$ when compared to its minimum Q_{min}. Figure (1) displays the contours of the residual sum of squares together with theL_2 ridge shrinkage constraint in the two-dimensional case (Kotz and Nadarajah 2004).

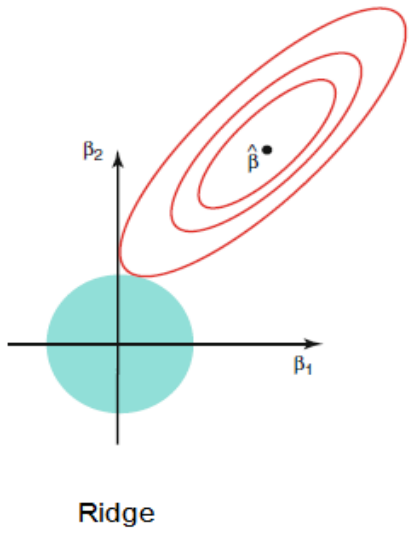

Figure 1: Contours of the Sum of Squares of the Residual and the Constraint Functions in Ridge Shrinkage Regression

In the view of the Lagrangian problem, it is equivalent to minimizing

$Q^*(\beta) = \|\beta\|^2 + (1/k)\{(\beta - \hat{\beta}^{LS})^t X^t X (\beta - \hat{\beta}^{LS}) - \phi_0\}$ (9)

Where k is the deflection factor chosen to satisfy the constraint.

Therefore, differentiate $Q^*(\beta)$with respect to β

$\frac{\partial Q^*(\beta)}{\partial \beta} = 2\beta + (1/k)\{2(X^tX)\beta - 2(X^tX)\hat{\beta}^{LS}\} = 0$ (10)

That yields the Ridge Shrinkage estimator as follows

$\hat{\beta}^R = \{X^tX + k\mathrm{I}\}^{-1}X^ty$ (11)

An alternative way is to state the Ridge Shrinkage problem in the constrained least-squares form by minimizing$\|y - X\beta\|^2$, subject to$\|\beta\|^2 \leq s$,

For some constant value of s.

Hence, the Lagrangian problem becomes simply minimizing that

$$\|y - X\beta\|^2 + \lambda . \|\beta\|^2$$

which produces the same estimator given in (11). The penalty parameter$\lambda \geq 0$ controls the amount of shrinkage in $||\beta||^2$. As the value of λ gets larger, the greater amount of shrinkage. For this reason, the Ridge Shrinkage estimator is often called the shrinkage estimator. There is a one-to-one correspondence among four values λ, s, k and ϕ_0(Bates and Watts 1988). It is extremely important to note that the formal Ridge Shrinkage solution is not invariant under the scaling of the explanatory variables. Therefore, standardization of both the explanatory variables and the response is essential, that is:

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_{xj}} \; and \; y'_i = \frac{y_i - \bar{y}}{s_y}$$

Before using the Ridge Shrinkage estimator in (11). It is helpful to adopt the following standardized variables notation, the matrices $\mathrm{X^tX}$ and $\mathrm{X^ty}$ becomes as follows:

$\mathrm{X^tX} = \mathrm{R_{XX}}$ and $\mathrm{X^ty} = \mathrm{r}_{XY}$

Note that $\mathrm{R_{XX}}$ denotes the correlation matrix among X_j's, and $\mathrm{r_{XY}}$denotes correlation vector between Y and all X_j's. Now, the Ridge Shrinkage estimator can be written as:

$\hat{\beta}^R = \{\mathrm{R}_{XX} + k\mathrm{I}\}^{-1}\mathrm{r}_{XY}$ (12)

If the explanatory variables are orthogonal ($\mathrm{X^tX} = \mathrm{I}$), then the Ridge Shrinkage estimates are just a scaled version of least squares estimates (it is equivalent to, $\hat{\beta}^R = \frac{1}{1+k}.\hat{\beta}^{LS}$for some shrinkage constant ($0 \leq \frac{1}{1+k} \leq 1$).

In addition, the intercept value β_0 goes to0 when working with standardized data. Having obtained a Ridge Shrinkage estimator $\hat{\beta}^R$, transformation step of its components is necessary in order to get the fitted linear regression equation between the original Y and $X_j$values. It is suitable to express in matrix form the normalization and its inverse transformation involved. Let $\mathbf{X}_0$ be the original design matrix. Its centered version is given by:

$$\mathrm{X}_C = (\mathrm{I} - \mathrm{j}_n\mathrm{j}_n^t/n)\mathrm{X}_0$$

And its normalized version is

$\mathrm{X} = \mathrm{X}_C\mathrm{L}^{-1/2}$

Where $\mathrm{j}_n$be the *n*-dimensional vector with all elements are ones and L be a diagonal matrix with diagonal elements from the matrix $\mathrm{X}_c^t\mathrm{X}_C$, i.e.,

$\mathrm{L} = \mathrm{diag}\,(\mathrm{X}_c^t\mathrm{X}_C)$.

Likewise, the original response vector y_0 can be normalized as

$\mathrm{y} = \frac{(\mathrm{I} - \mathrm{j}_n\mathrm{j}_n^t/n)\mathrm{y}_0}{s_y}$,

Where s_y is the sample standard deviation of y_0 .

It is straightforward to use the Ridge Shrinkage estimator $\hat{\beta}^R$ in (11) to predict with a new data matrix X_{new} (which is m×p on the original data scale).

The predicted vector $\hat{\boldsymbol{y}}_{new}$is then given as :

$\hat{y}_{new} = s_y\{(\mathrm{X}_{new} - \mathrm{j}_m\mathrm{j}_n^t\mathrm{X}/n)\mathrm{L}^{-1/2}\hat{\beta}^R + \mathrm{j}_m\mathrm{j}_n^t y/n\}$ (13)

Thus, the computation of the expectation and variance of $\hat{\beta}^R$ can be obtained using the following relation

$\hat{\beta}^R = \mathrm{Z}\,\hat{\beta}^{LS}$ (14)

Were

$\mathrm{Z} = \{\mathrm{I} + k(\mathrm{X^tX})^{-1}\}^{-1}$

It follows that

$\mathrm{E}(\hat{\beta}^R) = \mathrm{Z}\,\beta$ (15)

$\mathrm{Cov}(\hat{\beta}^R) = \sigma^2.\mathrm{Z}(\mathrm{X^tX})^{-1}Z^t$ (16)

Finally, comparison can be achieved between $\hat{\beta}^R$ with $\hat{\beta}^{LS}$ to see which estimator has a smaller MSE for certain values of *k*.

Let the ascending order sequence of the eigenvalues of the Z matrix as follows:

$$\lambda_{max} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m = \lambda_{min} > 0$$

From standard least square estimation, it is well known that

$MSE(\hat{\beta}^{LS}) = \sigma^2.\sum_j 1/\lambda_j$

For the Ridge Shrinkage estimator, the components of the Mean squared errors can be found from (15) and (16). The first component is the sum of their squared biases is

$$\sum_j \{E(\hat{\beta}_j^R) - \beta_j\}^2 = \{E(\hat{\beta}^R) - \beta\}^t \{E(\hat{\beta}^R) - \beta\}$$

$$\sum_j \{E(\hat{\beta}_j^R) - \beta_j\}^2 = \beta^t (I - Z)^t (I - Z)\beta$$

$$= k^2 \beta^t (X^t X + kI)^{-2} \quad (17)$$

And the second component is the sum of their variances is

$$\text{tr}\{\text{Cov}(\hat{\beta}^R)\} = \sigma^2 . tr\{(X^t X)^{-1} z^t z\}$$

$$= \sigma^2 \sum_j \left\{\frac{1}{\lambda_j} . \frac{\lambda_j^2}{(\lambda_j + k)^2}\right\} = \sigma^2 \sum_j \left\{\frac{\lambda_j}{(\lambda_j + k)^2}\right\} \quad (18)$$

Therefore, the MSE for the Ridge Shrinkage estimator is as follows

$$\text{MSE}(\hat{\beta}^R, k) = \sigma^2 \sum_j \left\{\frac{\lambda_j}{(\lambda_j + k)^2}\right\} + k^2 \beta^t (X^t X) + kI)^{-2} \beta$$

$$= \gamma_1(k) + \gamma_2(k) \quad (19)$$

It is worth noting that the first quantity $\gamma_1(k)$ is a monotonic decreasing function of k while the second quantity $\gamma_2(k)$ is monotonically increasing. The constant k reflects the amount of bias increased and the variance decreased. Whereas, when k = 0, it turns into the usual Least Squares Estimates (Hoerl *et al*., and Kennard 1970).

Had shown that there always exists a k >0 such that

$$\text{MSE}(\hat{\beta}^R, k) < \text{MSE}(\hat{\beta}^R, 0) = \text{MSE}(\hat{\beta}^{LS})$$

Finally, the Ridge Shrinkage estimator can be superior in comparison with the Least Squares Estimator in terms of providing a smaller MSE. However, in practice, the right choice of k is yet to be determined and hence there is no guarantee that a smaller MSE always is achieved by the Ridge Shrinkage Regression. The statistical properties of the Ridge Shrinkage estimator are tabulated in Table 1.

Table 1: Some of the most important statistical properties of the Ridge Shrinkage estimator

Sr	Property	Formula
1	Mean	$E(\hat{\beta}_R) = (X^t X + kI_p)^{-1} X^t X \beta$
2	Variance	$\text{Var}(\hat{\beta}_R) = \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2}$
3	Var-Cov matrix	$\text{Cov}(\hat{\beta}_R) = \text{Cov}(Z\hat{\beta})$ $= \sigma^2 (X^t X + kI_p)^{-1} X^t X (X^t X + kI_p)^{-1}$ $= \sigma^2$[VIF]
4	Bias	$\text{Bias}(\hat{\beta}_R) = -k(X^t X + kI_p)^{-1} \beta$ $= -k\, P\left(\frac{1}{\lambda_j + k}\right) P^t \beta$
5	MSE	$\text{MSE}(\hat{\beta}_R) = \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} + \sum_{j=1}^{p} \frac{k^2 \alpha_j^2}{(\lambda_j + k)^2}$

2.4 The LASSO and Other Extensions of Ridge Shrinkage Regression

The LASO is another shrinkage method like Ridge Shrinkage Regression, however with an important and attractive feature in variable selection. Rather, the Ridge Shrinkage Regression makes the selection process continuous by varying shrinkage parameter k and hence becomes more stable. Through this process, the Ridge Shrinkage Regression does not set any coefficients to zero, since it does not give an easily interpretable model as in subset selection (Greene2000). The LASSO technique is proposed to maintain the advantages of both subset selection and ridge regression by shrinking some coefficients and setting other coefficients to 0. The LASSO estimator of β is obtained by

minimizing $\|y - X\beta\|^2$, subject to $\sum_{j=1}^{p} |\beta_j| \leq s$.

More Explicitly, the L_2penalty $\sum_j \beta_j^2$ in Ridge, Shrinkage Regression is replaced by the L_1penalty $\sum_j |\beta_j|$ in LASSO. If s is chosen to be greater than or equal to $\sum_j |\beta_j^{LS}|$, then the LASSO estimates are the same as the least-squares estimation. If s is chosen to be smaller than $\sum_j |\beta_j^{LS}|$, then it will cause shrinkage of the solutions towards zero.

Figure 2 displays the contours of the residual sum of squares together with the L_1LASSO constraint in the two-dimensional case (Greene 2000).

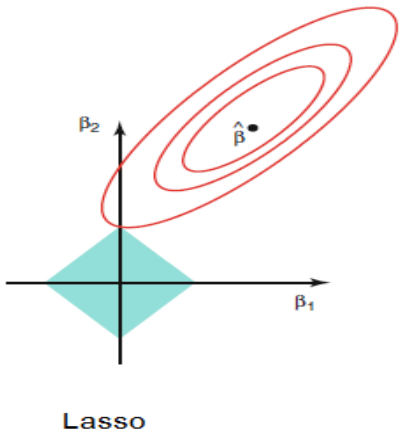

Figure 2: Contours of the Sum of Squares of the Residual and the Constraint Functions in LASSO

It is worth noting that, in Figure 2.1, the constraint region in Ridge Shrinkage Regression has a disk shape. While, in Figure 2.2, the constraint region in LASSO has a diamond shape. It can be seen that both the Ridge Shrinkage Regression and the LASSO methods start by finding the first point at which the elliptical contours hit the constraint region. But, unlike the disk of the Ridge Shrinkage Regression case, the diamond in the LASSO has corners. If the solution occurs at a corner, then it has one coefficient $\hat{\beta}_j$ equal to zero.

It can also be viewed that, the LASSO solution is competitive with the Ridge Shrinkage Regression solution but with many zero coefficient estimates. In the case of or the normal designs where $\mathbf{X^tX = I}$, the LASSO estimator can be written as**:**

$$\hat{\beta}_j^{lasso} = \text{sign}\,(\hat{\beta}_j^{LS})\{|\hat{\beta}_j^{LS}| - \gamma\}_+ \quad (20)$$

Where γ is determined by the condition $\sum_j |\hat{\beta}_j^{lasso}| = s$

In Conclusion, the coefficients whose values are less than the threshold γ would be automatically forced to go to 0 while the coefficients whose values are larger than γwould be shrunk by a unit of γ. Therefore, the LASSO technique performs as a variable selection operator (Baltagi 2001).

In general, the non-smooth behavior of the LASSO constraint makes the solutions nonlinear in the response variable y. Efron et al (2004) in their initial proposal of LASSO had used quadratic programming to solve the optimization problem. It is based on using the fact that the condition$\sum_j |\beta_j| \leq s$ is equivalent to $\delta_i^t$for all i= 1, 2, . . . , 2^p, where $\delta_i$is the p-tuples of form (±1,±2, . . . ,±p).

Others such as Efron *et al*., (2004) have developed a compact descent method for solving the constrained LASSO problem for any fixed *s*.

Lately, Efron *et al*., (2004) have derived a different approach, called the Least Angle Regressions (LARS). The LARS enables a variable selection method in a specific way. Since the entire path of LASSO solutions as *s* varies from 0 to +∞ can be extracted with a small modification on LARS. The LARS method works only with normalized data and employed iteratively technique to predict the response $\hat{\mu}$ with updating steps (Freund and Littell, 2000).

We have compared the performance of the Ridge Shrinkage Regression and the LASSO in a larger context. A major interest in this paper is the patterns of shrinkage as the λ changes. Ridge Shrinkage Regression tends to shrink the coefficients so that they all reach zero together as λ gets large. The LASSO shrinks the coefficients so that some reach zero well before others as λ gets large (Hastie *et al*., 2002).

3. THE SIMULATION STUDY

This section is devoted to comparing the performance of the two proposed estimators: namely: The Ridge and the Lasso estimators that used to treat the problem of Collinearity via simulation. In order to better identify the properties of the ridge estimator and some of its alternatives, we have computed two sets of simulations. The exact procedures are described in the following section (3.1). The simulations include several ridge and Lasso methods. We did include the related lasso approach despite the fact it is primarily used for model selection, not estimation.

All computations and graphics in this thesis were carried out using the software package R, which is based on the statistical language S (Statistical Science, Inc. 2015). However, we believe that our results are very useful for assessing the practical performances of the two proposed estimators that are used to solve the Collinearity problem.

3.1 Description of The Experiment

In this simulation study, the ridge estimator is compared to both OLS and Lasso using various choices of correlation (ρ = 0.10, 0.25, 0.50, 0.75, and 0.90) between predictors. This was done in order to better quantify some of the commonly cited advantages of the ridge estimator, such as it performs best when the predictors are strongly correlated. To cover the effects of various situations of Collinearity on the regression model, the study is classified into two different patterns, as follows: Pattern 1: Error terms are distributed as normal, Pattern 2: Error terms are distributed as non-normal. In each Pattern, several choices of the number of independent variables (p=2, 5, 10, and 20). The sample sizes considered were n = 25, 50, and 100, and the model was of the form Y= $X\beta + \varepsilon$. Lastly, two different marginal distributional errors were used: namely the normal distribution, and the heavy-tailed t distribution.

Experimenting with different choices of correlations (ρ = 0.10, 0.25, 0.50, 0.75, and 0.90) between the choice number predictors (p), and three different sample sizes (*n*) 25, 50, and 100, also the two different marginal errors' distributions for 1000 independently different runs, each gives the means and standard deviations of the mean square error (MSE) obtained by utilizing the three proposed estimators, namely the OLS estimator, the Ridge estimators, and Lasso estimator.

Next, we present the graphical summary obtained from the simulation study in each pattern of errors distribution.

Pattern 1: Errors are distributed as normal and p = 2,5,10, and 20:

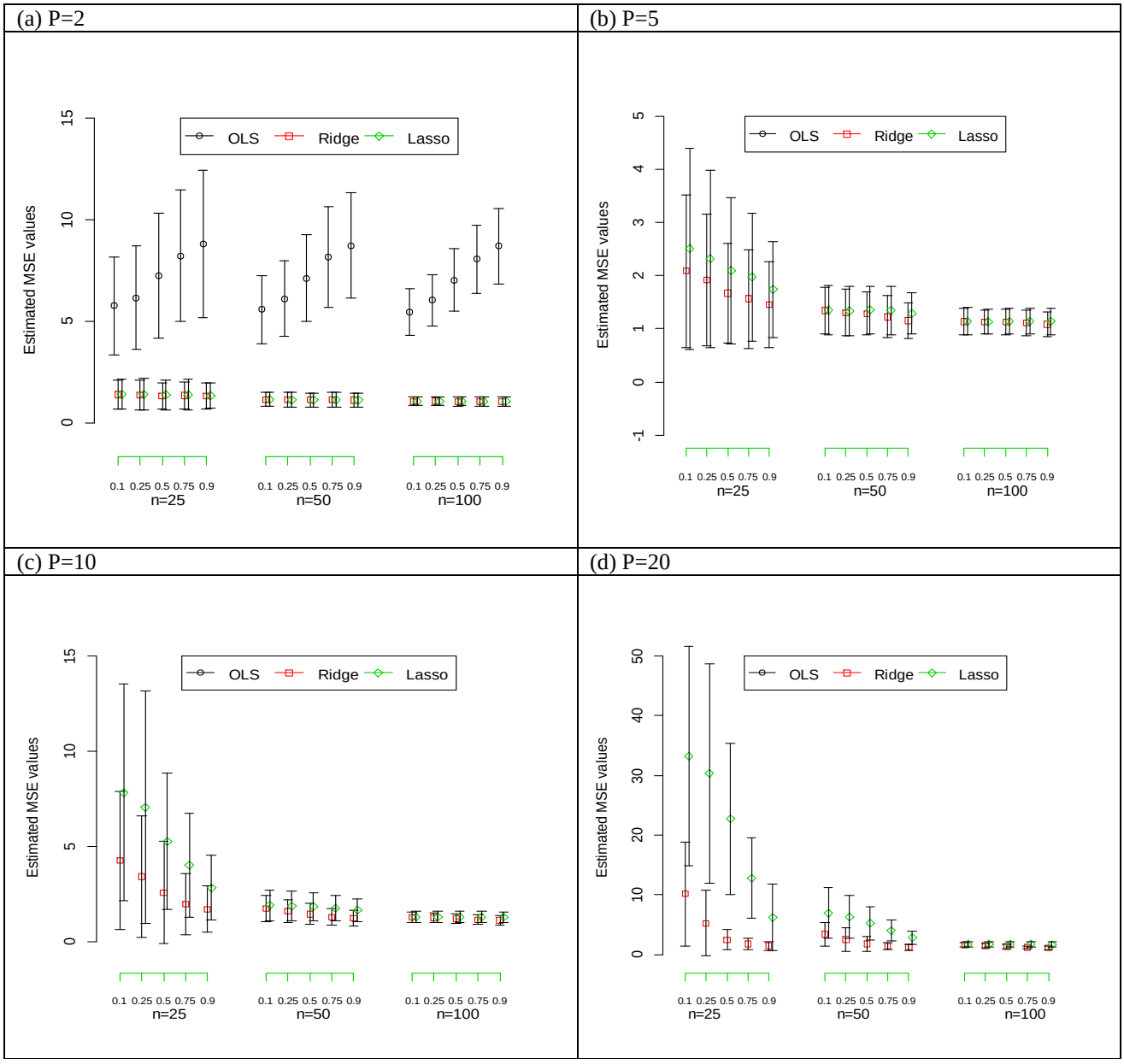

Figure 3: The Box-plots demonstrate how various choices of correlation affect the mean (standard deviation) of the MSE value for the three proposed estimators (OLS, Ridge, and Lasso) for (a) P=2, (b) P=5, (c) P=10, and (d) P=20 in pattern1

Having examined Figure 3 very carefully **in the case of errors belonging to Normal distribution and P= 2, 5, 10, and 20,** we have noted the following points:

1. The values of MSE_{OLS} increase very dramatically and they behave badly in comparing with the corresponding counterparts of MSE_{Lasso} and MSE_{Ridge} in the case of multi-collinearity presence.
2. The numerical results support the superiority of the Ridge estimator followed closely by the Lasso estimator as the sample sizes increase from 25, 50, and 100.
3. When increasing the sample size (n = 25, 50, and 100) and in the case of the explanatory variables (p = 5), the Ridge estimator performs best (e.g. MSERidge = 1.083815 when ρ = 0.25 and n = 100 where MSELasso = 1.133132 at the same values of n and ρ).
4. When increasing the sample size (n = 25, 50, and 100) and in the case of the explanatory variables (p = 10),the Ridge estimator performs best (e.g. MSERidge = 1.113801 when ρ = 0.9 and n = 100 where MSELasso = 1.282817 at the same values of n and ρ).
5. When increasing the sample size (n = 25, 50, and 100) and in the case of the explanatory variables (p = 20), the Ridge estimator performs best (e.g. MSERidge= 1.141078 when ρ = 0.9 and n = 100 where MSELasso = 1.689402 at the same values of n and ρ).

6. In general, the best performance (the smallest values of averaged MSE) has been achieved by the Ridge estimator for all possible choices of P = 2, 5, 10, and 20, all possible choices of correlations ρ (0.10, 0.25, 0.50, 0.75, and 0.90) as well as all possible sample sizes (n=25, n=40, and n=100) when the marginal errors' distributed as normal.

Final remark, for a fixed choice of correlation ρ, the change in the estimated values of MSE is insignificant as the value of sample sizes increases from n=25, 50, and 100.

In summary, the estimated values of MSE decrease as the sample size increases from n = 25, 50, and 100. The Ridge estimator has provided better-estimated MSE values regardless of the choice of the sample size as well as the choice of correlation ρ exists among the two explanatory variables.

The best performance (the smallest values of averaged MSE and stander deviation and its corresponding standard deviations) has been achieved by the Ridge estimator for all possible choices of correlations (0.10, 0.25, 0.50, 0.75, and 0.90) as well as all possible of sample sizes (n=25, n=50, and n=100) when the marginal errors' distributed as normal

Pattern 2: Errors are distributed as Non-normal and p =2, 5, 10, and 20

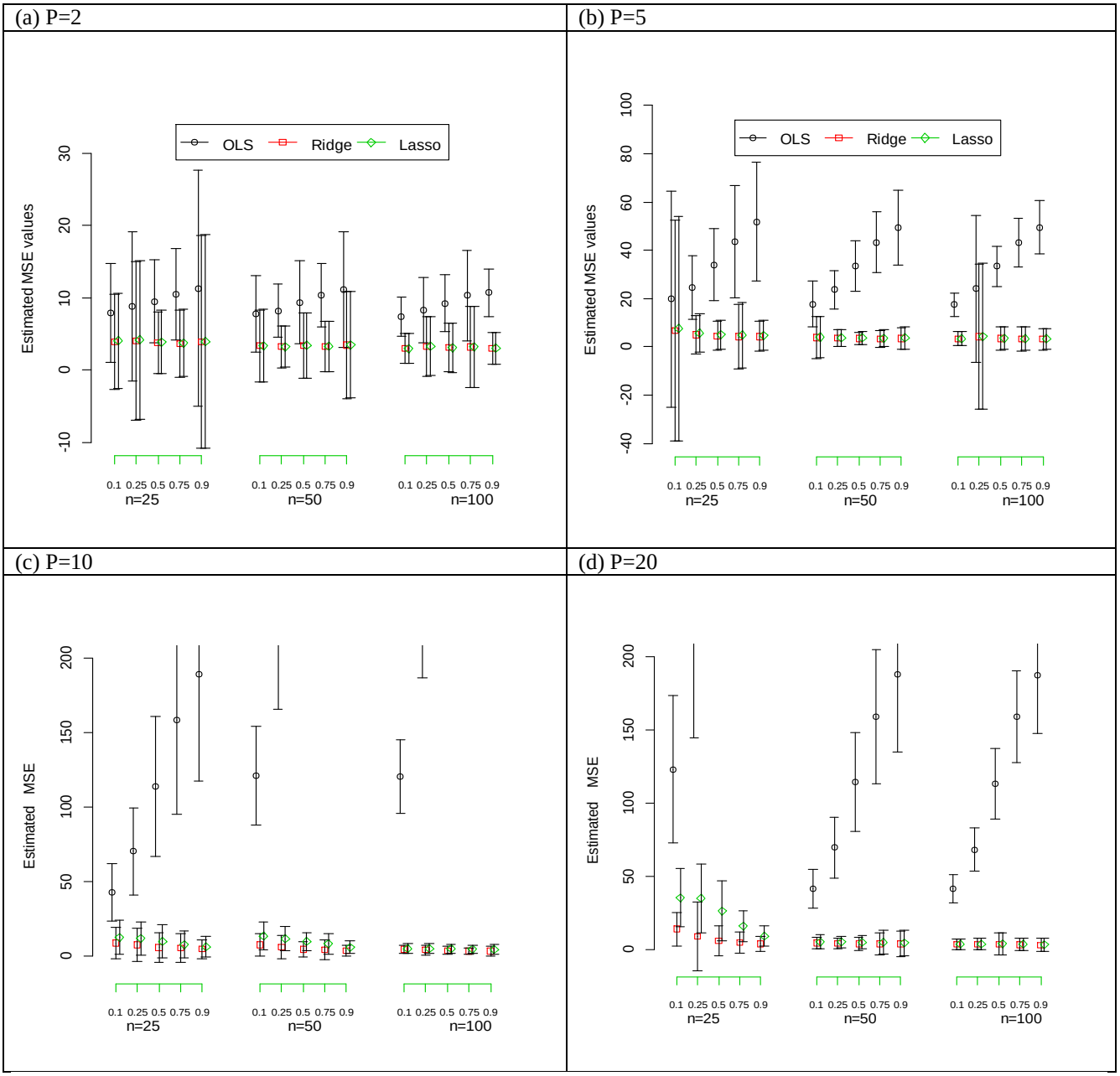

Figure 4: The Box plots demonstrate how various choices of correlation affect the mean (standard deviation) of the MSE value for the three proposed estimators (OLS, Ridge, and Lasso) for (a) P=2, (b) P=5, (c) P=10, and (d) P=20 in pattern 2

Having examined Figure 4 very carefully in the case of errors belonging to Non-Normal distribution and P= 2, 5, 10, and 20, we have noted the following points:

1. The values of MSE_{OLS} increase very dramatically and they behave badly in comparing with the corresponding counterparts of MSE_{Lasso} and MSE_{Ridge} in the case of multi-collinearity presence.
2. The numerical results support the superiority of the Ridge estimator followed closely by the Lasso estimator as the sample sizes increase from 25, 50, and 100.
3. When increasing the sample size (n = 25, 50, and 100) and in the case of the explanatory variables (p = 5), the Ridge estimator performs best (e.g. MSE_{Ridge} = 3.129667 when ρ = 0.9 and n = 100 where MSE_{Lasso} = 3.259719at the same values of n and ρ).
4. When increasing the sample size (n = 25, 50, and 100) and in the case of the explanatory variables (p = 10), the Ridge estimator performs best (e.g. MSE_{Ridge} = 3.164759whenρ = 0.9 and n = 100 where MSE_{Lasso} = 3.54715at the same values of n and ρ).
5. When increasing the sample size (n = 25, 50, and 100) and in the case of the explanatory variables (p = 20), the Ridge estimator performs best (e.g. MSE_{Ridge}= 3.231649 when ρ = 0.9 and n = 100 where MSE_{Lasso} = 4.326057 at the same values of n and ρ).
6. Once more, the best performance (the smallest values of averaged MSE) has been achieved by the Ridge estimator for all possible choices of correlations (0.10, 0.25, 0.50, 0.75, and 0.90) as well as all possible sample sizes (n=25, n=40, and n=100) when the marginal errors' distributed as non-normal.

To sum up, when the sample size was fixed at n = 100 and choices of correlation ρ varies from 0.10 to 0.90, it was concluded that the Ridge estimator once more gives smaller values of MSE along with their corresponding counterparts standard errors followed closely by the values obtained by Lasso estimator.

Final comment, for a fixed choice of correlation ρ, the change in the estimated values of MSE is insignificant as the value of sample sizes increase from n = 25, 50, and 100.

In summary, the estimated values of MSE decrease as the sample size increases from n = 25, 50, and 100 for all proposed estimators used. The Ridge estimator has provided better-estimated MSE values regardless of the choice of the sample size as well as the choice of correlation ρ exists among the two explanatory variables.

The best performance (the smallest values of averaged MSE and stander deviation and its corresponding standard deviations) has been achieved by the Ridge estimator for all possible choices of correlations (0.10, 0.25, 0.50, 0.75, and 0.90) as well as all possible of sample sizes (n=25, n=50, and n=100) when the marginal errors' distributed as non-normal. This means that the Ridge estimator works well regardless of the choice of the error term distribution of either normally or non-normally distributed.

4. DISCUSSION OF THE RESULTS AND CONCLUSION

The purpose of this section is to summarize the similarities and the differences between the Ridge regression estimator and the Lasso estimator which are used to handle and solve the "multi-collinearity" problem when the errors are either normally distributed or non-normally distributed.

4.1 The Discussion of the 1st Simulation Study

The comparisons among (OLS, Ridge Regression, Lasso) were made using the mean values of (MSE) as well as their corresponding values of standard deviations assuming that the error terms are distributed as normal. The following two interesting points have been concluded:

1. Ridge regression is one of the more common, although debated, estimation procedures for solving the multi-collinearity problem. The procedures discussed in this simulation fall into the category of biased estimation techniques. They are based on this idea: despite that OLS gives the best linear unbiased estimators (BLUE), there is no upper bound on the variance of the estimators and the presence of multicollinearity may produce large variances. Consequently, one can visualize that, under the condition of multi-collinearity, a huge price must be paid for the unbiasedness property that one achieves by using OLS. Biased estimation is used to accomplish a substantial reduction in variance with an accompanying increase instability of the regression coefficients. The coefficients become biased and, if one is successful, the reduction in variance is of greater magnitude than the bias made in the estimators.
2. The Ridge estimator provides the smallest mean and standard deviation values of MSE followed by the Lasso estimator. Whereas, the OLS estimator provides the largest mean and standard deviation values of MSE, with a normal distribution of errors and different selected values of the correlation coefficient. We noted that the values of MSE and their corresponding standard deviations of the Ridge estimator maintained the smallest for all selected correlations regardless of the number of different explanatory variables or the sample size.

4.2 The Discussion of the 2nd Simulation Study

In the second simulation study, it was assumed that the errors were distributed according to the heavy-tailed t-distribution (Non-normal distribution). All other

settings were made as in the first simulation study. The following four interesting points have been reached:

1. We have noted that the three estimators (OLS, Ridge Regression, Lasso) behaved in the same manner as when the errors were distributed normally.
2. The OLS estimator continued to give very large MSE values due to the problem of multi-collinearity.
3. The Ridge estimator maintains its superiority over the Lasso estimator, albeit with a slight difference such that it gives the smallest MSE value as well as the smallest standard deviation value.
4. Shrinkage estimators of Ridge regression and Lasso have proven to be robust due to non-normal errors' distribution, since no significant effect on the simulation results.

REFERENCES

1. Baltagi, B. H. (2001). Econometric Analysis of Panel Data. Wiley, John & Sons: New York.
2. Bates, D. M., & Watts, D. G. (1988). Nonlinear Regression Analysis and Its Applications. John Wiley and Sons.
3. Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least Angle Regression (with discussion). Annals of Statistics, 32, pp. 407-499.
4. Freund, R. J., & Littell, R. C. (2000). SAS System for Regression, 3rd ed. Cary, NC: SAS Institute.
5. Greene, W. H. (2000). Econometric Analysis, 4th Edition. Prentice.
6. Hampel, E. M. Ronchetti, P. J., Rousseeuw., & Stahel, W. A. (1986). Robust Statistics, The Approach Based on Influence Functions. Wiley, New York.
7. Hastie, T., Tibshirani, R., & Friedman, J. (2002). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY: Springer-Verlag.
8. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. Technometrics, 12, pp. 55-67.
9. Hoerl, A. E., Kannard, R. W., & Baldwin, K. F. (1975). Ridge regression: some simulations. *Communications in Statistics-Theory and Methods*, *4*(2), 105-123.
10. Kotz, S., & Nadarajah, S. (2004). Multivariate t Distributions and Their Applications. Cambridge University Press.
11. Stein. (1960). Multiple regression. In Contributions to Probability and Statistics, Essays in Honor of Harold Hotelling. Stanford University Press, Stanford, California.
12. Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B, 58, pp. 267-288.
13. Weisberg, S. (2005). Applied Linear Regression. 3rd edition. Wiley and Sons, Inc.
14. Yan, X., & Su, X. G. (2005). Testing for Qualitative Interaction. Encyclopedia of Biopharmaceutical Statistics: Second Edition, Revised and Expanded, Ed.