Al-Qadisiyah Journal for Administrative and Economic Sciences ISSN$_{Online}$: 2312-9883
QJAE, Volume 26, Issue 4 (2024) ISSN$_{Print}$: 1816-9171

Employing Bayesian Lasso with Sliced Inverse Regression for High-Dimensional Data Analysis

Saif Hosam Raheem saif.hosam@qu.edu.iq
Raghad Hamza Mahdi Statistics.stp.24.8@qu.edu.iq
University of AL-Qadisiyah

Received: 13/10/2024
Accepted: 8/12/2024
Available online: 25 /12 /2024

Corresponding Author : Raghad Hamza Mahdi

Abstract : This research aims to develop an efficient model for high-dimensional data analysis by integrating the Bayesian Lasso method with Sliced Inverse Regression (BLSIR). The main problem addressed is that traditional methods, such as Ordinary Least Squares (OLS), struggle to handle data with many explanatory variables, especially in the presence of outliers or multicollinearity. By employing the BLSIR model, the results show that this method achieves higher accuracy in estimating significant variables and reducing the negative impact of non-influential variables compared to LASSO, SIR, and BLASSO methods. The model was tested on real data related to non-alcoholic fatty liver disease, where BLSIR outperformed other methods in minimizing the Mean Squared Error (MSE) and Mean Absolute Error (MAE). This model shows significant improvements in handling high-dimensional data, offering better accuracy and variable selection. Its application in non-alcoholic fatty liver disease analysis proves its potential for broader use in various scientific and medical fields. Based on these findings, we recommend applying this method in other fields to analyze multidimensional data. The implementation was carried out using the R language program.

Introduction:

Statistics is an important and basic science. It is concerned with studying the methods of collecting, analyzing, interpreting, and predicting data. The purpose is to obtain accurate and useful information in order to make the right decisions in various fields. Due to the quick development witnessed by technology and information sciences in our modern era, which has greatly impacted the advancement of medical, natural, and human sciences. This technological and informational progress has clearly influenced the field of statistics, owing to its close connection with these sciences. This development was accompanied by the emergence of the high-dimensional data problem, where the number of explanatory variables (P) is greater than the sample size (n), i.e., (P > n).This type of data often includes several standardization issues, such as the presence of multicollinearity, extreme (abnormal) values, and other related problems.

This type of data is also characterized by a number of properties, the most common of which is the property of high correlation between variables and the property of variables with a clustered structural structure and others. In this case, analyzing this data becomes difficult and complex, and traditional statistical methods cannot be applied, as they will yield incorrect results that could impact the decisions made. This is called the problem of dimensionality, and this problem has become the focus of attention of many researchers. The solution to this problem is to reduce the dimensions while preserving the characteristics and information of the regression in addition to improving the model. This is done by choosing a group of these variables called the subset (Subset). There are two methods for choosing the subset, which are the variable selection method (V.S) and the feature extraction method (F.E). In order to obtain sufficient dimensions, the method of reducing the sufficient dimension will be used, which is one of the feature extraction methods. This is done by reducing the number of variables (p) to a smaller number of dimensions (d), i.e. (p<d) while preserving the information and characteristics of the regression. There are two methods for extracting the feature, the first aims to find the (Central Suβspace) (finding an important subset of the original data) and an example of this is the Sliced Inverse Regression. The second method aims to find the (partial central distance) and the most famous of these is the Minimum Average Variance Estimator. Despite the advantages of these methods represented by reducing the dimensions, they suffer from reducing a number of variables (p) to a number of dimensions (d), but each dimension contains within it all the unimportant variables, the presence of which causes problems, including difficulty in interpretation, and sometimes leads to unreliable results.

Research Problem

The main problem is in applying the Ordinary Least Squares (OLS) method to high-dimensional data where efficient estimates cannot be obtained. OLS suffers from two main problems: low prediction accuracy due to high variance, and difficulty interpreting the model in the presence of a large number of predictive variables. In addition, OLS is very sensitive to outliers, which may affect the distribution of the model residuals and make it non-normal, which violates the conditions for applying OLS.

Research Objective

The main objective of the study is to develop an efficient regression model to deal with high-dimensional (HD) data and address the problems caused by outliers and multicollinearity among predictive variables. The research seeks to improve prediction accuracy and reduce high variance in estimates by using regularization methods such as different regularization methods that impose penalties on the size of the parameters to facilitate the process of estimating the parameters in the case of a large number of variables and a small sample size. The thesis aims to employ the Bayesian Lasso method with the piecewise inverse regression method.

Variables selection procedure[3] [4] [7]

The regression model includes a large number of explanatory variables and it is not known which explanatory variables will affect the dependent variable. Therefore, a subset of the original variables is selected to obtain the smallest subset of important variables that have a significant effect on the dependent variable and achieve the best estimation model with high explanatory and predictive power. Variable selection plays an important role in the analysis of high-dimensional data because it reduces unimportant variables, reduces bias, provides faster and less expensive models, improves model prediction, and gives a good understanding of the data set. To achieve these goals, researchers have proposed many methods for variable selection V.S. They are divided into two types: traditional methods (Backward elimination procedure, Forward selection procedure, Stepwise selection procedure, Akaike Information Criteria). And regularization methods (Lasso, Group Lasso, Adaptive Lasso Method., Elastic Net Method, Adaptive Elastic Net Method, Reciprocal Lasso method).

Lasso penalty function[13] [14]

Lasso method was proposed by (Tibshirani , 1996) and Lasso means Least absolute shrinkage and selection operator. It is an effective and powerful method for processing high-dimensional data (HD). It is also part of the family of penalty least squares, as it works on selecting variables and estimating parameters at the same time. It works on reducing some parameters and zeroing others to zero completely, and thus it can automatically achieve variable selection. Therefore, this method is very similar to the Ridge regression method from a theoretical point of view. Ridge regression adds the sum of squared coefficients (penalty I_2) ($\sum_{k=1}^{p} \hat{\beta}_k^2$).

But Lasso adds the absolute value of the sum of $\sum_{k=1}^{p} |\hat{\beta}|$. The researchers (Fan, Li, 2001) showed that this method produces biased estimates of large coefficients. Therefore, it does not have the oracle property. The properties of OP include (consistency, sparsity, homoscedasticity, and the ability to choose the true model with probability of one). Although these attractive features of Lasso have proven successful in a variety of situations, Lasso faces some problems. The first problem is that when n>p, the Lasso method cannot handle sets of highly correlated independent variables. The second problem is that the Lasso method cannot handle the information provided by explanatory variables that form sums. The third problem is if we have P explanatory variables and n observations. And if p>n then we choose n explanatory variables at most and in this way we will neglect some variables that have an effect on the model. This estimator is obtained by adding the penalty function to the least squares loss as in the following equation:

$$\hat{\beta} = \mathbf{argmin} \sum_{i=1}^{n} (y_i - \sum_j X_{ij} \beta_j)^2 \quad \text{subject to} \sum_j |\beta_j| \leq \lambda \qquad (1)$$

$$\hat{\beta} = \mathbf{argmin} \sum_{i=1}^{n} \left(y_i - \sum_j X_{ij} \beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \qquad (2)$$

λ: Represents the penalty parameter or tuning parameter.
$\hat{\beta}$: Represents the estimators of the OLS method.
P: Represents the number of variables since (J=1,……,P).
n: Sample size since (i=1,….,n)
The first part of the above equation represents the least squares loss, and the second part represents the Lasso penalty function. It controls the degree of shrinkage (contraction) of the estimator, so it plays a fundamental role in the process of selecting the significant variable and works to control the severity of the penalty. When it is 0= λ, no parameter is deleted and thus the number of variance increases, but when the value of λ>0 increases, it gives a greater reduction in

the complex model and provides criteria for selecting the variable, i.e. it gives models that are subject to change (Alkenani, Yn, 2013). The value of λ is determined through GCV (Generalize Cross Validation) as in the following equation:

$$GCV = \frac{SSR}{n(1 - p(\lambda)\backslash nX)^2} \qquad (3)$$

$$RSS = \sum_{i=1}^{n} (y_i - X_i^T)^2 \qquad (4)$$

The estimator of the Lasso method can be represented as follows:

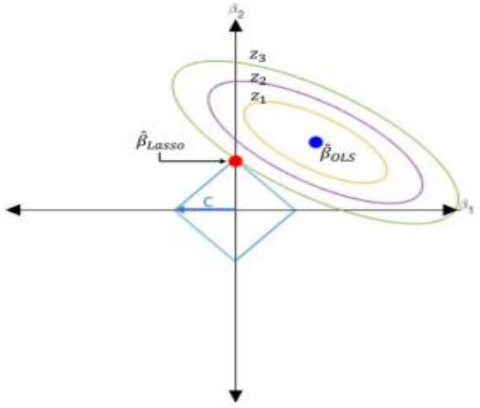

Figure (2-1) An illustration of the Lasso method.

The colored circles explain the sum of squares of the error with the estimator value of the least squares method that touches the penalty function represented by the diamond shape.

Bayesian Lasso[(1) (13)]

It is a statistical model used in regression analysis, and combines two main concepts: Lasso and Bayesian Models. The Bayesian Lasso was proposed by the researcher (Robert Tibshirani) in 1996. The Bayesian regularization method was developed due to the difficulties in statistical inference of regression coefficients. On the other hand, the Bayesian method provides accurate inference even with small sample sizes, in addition to accurate estimation when p is greater than n (n<p).

The Bayesian regularization model includes two steps:

First: Determine the prior distribution of regression coefficients, which is the most important step in the Bayesian method for selecting variables and estimating coefficients together. The main idea of Bayesian analysis is to minimize the variance of the estimator while increasing the bias. Therefore, the choice of the prior distribution must be accurate because choosing an inaccurate or incorrect prior distribution can lead to a number of problems including Giββs sample convergence problems, posterior estimation problems, and instability. Second: Calculating the posterior distribution, Tibshirani (1996) proposed that if the regression coefficients are identical and independent (i.e. double exponential), the Lasso estimate can be interpreted as a posterior estimate, and as a result, many Lasso techniques have been proposed over the years by other researchers to use the Laplace analogy, for example, (Figueiredo2003;Bae and Mallick 2004;Yuan and Lin 2005) and in 2008 Park, Casella developed a complete Bayesian analysis based on Lasso analysis based on the description of the Laplace conditional model.

$$\pi(\beta/\sigma^2) = \prod_{k=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_k|/\sqrt{\sigma^2}} \qquad (5)$$

Sufficient dimension reduction (SDR)[(6) (10)]

Cook (1998) proposed a theory of sufficient dimension reduction to reduce a number of explanatory variables without losing regression information. Dimension reduction methods are one of the main methods that researchers need to analyze high-dimensional (HD) data, especially in recent years. This is after the development of data collection methods, data storage, and storage capacity. These developments contributed to the emergence of the "curse of dimensionality" problem. This problem has become the focus of attention of many researchers. The main task of the concept of dimensionality (SDR) is to reduce the number of explanatory variables and simplify the study of relationships by selecting a subset of variables that actually affect the model. That is, transforming data from a high-

dimensional space to a low-dimensional space while preserving the characteristics and regression information. That is, it replaces the original predictive variables with linear structures while preserving regression information. There are two types of dimensionality reduction methods. The first type is subset selection. The second type represents feature extraction methods.

Sliced Inverse Regression (SIR)

The Sliced Inverse Regression (SIR) method is considered one of the most important methods for reducing the dimensions of the variables included in the analysis, where the data has high dimensions (i.e. it suffers from the problem of dimensionality CD). This method was proposed by the researcher (1991, Li). It is known that regression analysis studies the relationship between the dependent variable y and the independent variables (X'S) represented by (E(y /x)). As for the inverse regression of the SIR slices, it studies the relationship in an inverse manner between the dependent variable (y) (Dependent Variaβle) and the independent variables (X'S) (Independent Variaβle), i.e. it makes the variable y represent the independent variable, and X represents the dependent variable, represented by (E(x ⁄ y)) . This method has been applied and used widely in various fields, including finance, economics, and medical fields. This method divides the model into multiple slices according to the values of the dependent variable y, and then different statistical operations are performed for each slice. It also works to integrate the information of all slices and obtain the latent root information. The largest of them is chosen to represent the effective trends (e. d .r) of the SIR, and here we mean (e. d .r) The vector resulting from the reduction process that represents the new shape of the data, and its dispersion is proportional to the dispersion of the original variables (X). The inverse regression curve cannot be straight, and this curvature plays an important role in finding trends (e. d .r). If the inverse regression curve is straight, we may not be able to find more than one trend. It is difficult to estimate the normal regression parameters for high-dimensional data if there is a dimensionality problem CD, inverse regression can address this problem by dividing the regression into P slices (E(X_i/y)) since (i=1,2,…,p). In this case, the dimensionality problem will be neglected if we assume that $\bar{X}$represents the arithmetic mean of the variable X and assume that $(\hat{z} = \hat{\Sigma}^{-1|2}(X - \bar{X}))$ is a simplified version of Z, where $\hat{\Sigma}$ represents the covariance matrix of the variable X. Let h represent the number of slices and n_y represent the number of observations of the y^{th}slice. So m is a simplified version of $\widehat{M} = \sum_{y=1}^{h} \hat{f}_y \hat{z}_y \hat{z}_y^T$ where $\hat{f_y}$and $\hat{Z}_y$is the Z-rate of slice y. Let us assume that $(\hat{\lambda}_1 > \hat{\lambda}_2 > \cdots \hat{\lambda}_p \geq 0)$ which represents the eigenvalues corresponding to the eigenvectors $(\hat{v}_1, \hat{v}_2, \hat{v}_3, \ldots\ldots, \hat{v}_p)$ of $\widehat{M}$. If the area d of $(S_{y/z})$ is known, then span($\hat{\beta}$)= span$(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \ldots\ldots, \hat{\beta}_P)$ is a consistent estimator of $(S_{y/z})$, when $\hat{\beta}_i = \hat{\Sigma}^{-1/2}\hat{v}_i$. The SIR method relies on finding effective trend estimates that serve as parameters (β'S) through which the data is transformed into the reduced form and the original data is replaced for ease of handling, and in turn it is considered to address the problem of dimensions or (the curse of dimensions), that the model on which the SIR method relies is similar to the nonparametric and semiparametric regression model and is represented by the following formula:

$$y = f\left(B_1^T X, B_2^T X, B_3^T X, \ldots\ldots, B_k^T X, \varepsilon\right) \qquad (6)$$

Where:

B_k: is a vector of unknown parameters, and k=1,2,3,…,m.

ε: represents the term of the random error independent of (X).

And (f) represents an unknown function.

In this case, the problem of dimensionality was addressed using the inverse regression method. The inverse regression model can be represented as follows:

$$F(A, C) = \sum_{y=1}^{h} \| f_y^{-1/2} \hat{Z}_y A C_y \|^2$$

Basic algorithm for (SIR):

1-Unify the values of the variable (X) by asymptotic transformation to obtain

$$\breve{X} = \hat{\Sigma}_{xx}^{-\frac{1}{2}}(X_i - \bar{X}), \qquad i = 1,2, \ldots, n \qquad (7)$$

Where $\hat{\Sigma}$represents the sample covariance matrix, and $\bar{x}$ represents the sample mean respectively.

2-Dividing the range from y_i to H into slices $(I_1, I_2, \ldots\ldots, I_H)$, assuming that the proportion of $y_i$that falls in slices H is$\hat{p}_h$. The experimental proportion is calculated using the following formula:

$$\hat{P}_h = (1/n) \sum_{i=1}^{n} \delta_y(y_i) \qquad (8)$$

Where $\delta_h y_i$can take two values either 0 or 1 depending on whether $y_i$is in slice H or not.

3-In each slice the sample mean of $\breve{X}_I$is calculated and is referred to as with $(\hat{m}_h(h = 1,2,3, \ldots, H))$ where $(\hat{m}_h = (1/n\hat{p}_h) \sum_{y_i \in I_h} \breve{X}_i)$.

4-Perform a weighted PCA for the data at ((h=1,2,3,…,H)mˆ_h) as follows: Find the weighted covariance matrix $\hat{\mathrm{V}} = \sum_{\mathrm{h}}^{\mathrm{H}} \hat{\mathrm{P}}_{\mathrm{h}} \hat{\mathrm{m}}_{\mathrm{h}} \hat{\mathrm{m}}_{\mathrm{h}}^{\mathrm{T}}$ and then search for the characteristic values and vectors.

5-Let k be the largest characteristic vector which is (the row of vectors) (k=1,2,3,….,k) To get:

$$\hat{\beta}_{\mathrm{K}} = \hat{\mathrm{n}}_{\mathrm{K}} \hat{\Sigma}_{\mathrm{xx}}^{-1/2} \qquad (\mathbf{9})$$

$\mathrm{n}_{\mathrm{k}}\beta_{\mathrm{k}} \sum_{\mathrm{XX}}^{1/2} (k = 1,2,3, \ldots., \mathrm{K})$ ، Since k=1,2,3,…,K

$\boldsymbol{\Sigma}_{\mathbf{xx}}$ denotes the covariance matrix of the matrix X. Steps 2 and 3 produce an initial estimate of the standard inverse regression curve E(Z⁄y). The adjustment of the weights in the principal components analysis in step 4 is to take into account the presence of different sample sizes in different sectors. The first k components determine the most important subarea for tracing the path of the inverse regression curve E(Z ⁄y). Finally, step 5 rescales the original scale. Thus, $\hat{\beta}_{\mathrm{K}}\mathrm{S}$ can be used to bias as an estimate of the direction of the standard inverse regression curve (EDR) and the area of β(EDR) estimated by$\hat{\beta}$, the area generated by $\hat{\beta}_{\mathrm{k}}\mathrm{S}$.

Shrinkage Sliced Inverse Regression Method[(9)(10)]

In (2005), the researcher (NI et al.) proposed the SIR shrinkage estimator (SSIR-L) method that combines the Lasso penalty and the SIR method. SIR provides an estimated Span of($\hat{\beta}$) the central subspace $\mathrm{S}_{\mathrm{y/x}}$ and the elements of $\hat{\beta} \in \mathrm{R}^{\mathrm{P}\times\mathrm{d}}$ are usually non-zero. If there are many independent variables or those variables are highly correlated, in this case we will need a subset of these variables to obtain sufficient predictors. Assuming that some rows of the coefficient matrix β are all zero, the Lasso method was used to develop the inverse regression by compressing some rows of the matrix β to zero. In order to improve interpretability, Cook (2004) formulated the inverse spline regression to improve some regression problems by reducing:

$$\mathrm{F}(\mathrm{A},\mathrm{C}) = \sum_{\mathrm{y}=1}^{\mathrm{h}} \|\hat{\mathrm{f}}_{\mathrm{y}}^{-1/2} \hat{\mathrm{Z}}_{\mathrm{y}} - \mathrm{AC}_{\mathrm{y}}\|^2 \qquad (10)$$

Therefore, $\mathrm{C} = (\mathrm{C}_1, \mathrm{C}_2, \ldots., \mathrm{C}_{\mathrm{h}})$ and $\mathrm{A} \in \mathrm{R}^{\mathrm{p}\times\mathrm{d}} and\ C_y \in R^P$. Assuming that ($\hat{A}$ & $\hat{C}$) represent the values of (A&C) that minimize F . Span($\hat{\mathrm{A}}$)is equal to the space spanned by the largest (d) among the eigenvectors of ($\hat{\mathrm{M}}_{\mathrm{SIR}}$), it is not necessary that the value of $\hat{\mathrm{A}}$ is unique. By focusing on the coefficients of the independent variables, the researcher (Ni et al.) reformulated F(A&C) as follows:

$$\mathrm{G}(\mathrm{B},\mathrm{C}) = \sum_{\mathrm{y}=1}^{\mathrm{h}} (\hat{\mathrm{f}}_{\mathrm{y}}^{-\frac{1}{2}} \hat{\Sigma}^{-\frac{1}{2}} \hat{\mathrm{Z}}_{\mathrm{y}} - \mathrm{BC}_{\mathrm{y}})^{\mathrm{T}}\, \hat{\Sigma}\, (\hat{\mathrm{f}}_{\mathrm{y}}^{-\frac{1}{2}} \hat{\Sigma}^{-\frac{1}{2}} \hat{\mathrm{Z}}_{\mathrm{y}} - \beta \mathrm{C}_{\mathrm{y}} \qquad (11)$$

Since:

β: represents the value that zeroes the previously mentioned equation, which is the value of β.

span $\left(\widehat{\beta}\right)$=span($\hat{\Sigma}^{-1/2}\hat{\mathrm{A}}$)represents the estimator of $\mathrm{S}_{\mathrm{y/x}}$ and represents

Then the researchers explained the shrinkage estimator of the inverse regression method for ($\mathrm{S}_{\mathrm{y/x}}$)

by minimizing the following equation:

$$\sum_{\mathrm{y}=1}^{\mathrm{h}} \| \hat{\mathrm{f}}_{\mathrm{y}}^{1/2}\, \hat{\mathrm{Z}}_{\mathrm{y}} - \hat{\Sigma}^{\frac{1}{2}} \mathrm{diag}\left(\widehat{\beta}\hat{\mathrm{C}}_{\mathrm{y}}\right) \propto \|^2 + \lambda \sum_{\mathrm{i}=1}^{\mathrm{p}} |\beta_{\mathrm{i}}| \qquad (12)$$

Where ($\hat{\beta}$&$\hat{\mathrm{C}}$))reduces the value of G(β,C).

Inverse Lasso Regression:

Researcher Li et al. in 2019 defined the traditional inverse Lasso regression and indicated that this regression is estimated by studying the following optimization problem:

$$\arg min_{\beta} = \| \frac{1}{n} x'x\beta - \mathrm{n} \|^2 + \lambda \ \| \beta \| \qquad (13)$$

where n is the eigenvector associated with the largest eigenvalue. Li et al. also pointed out that as long as $\beta \propto \Sigma^{-1}\mathrm{n}$ Formula (13) can be followed to find the space-spanned solutions β. Formula (31) can be considered as a penalized least squares problem under the condition ‖β‖<λ. The matrix x in (13) is p×n. The estimated latent vector $\hat{\eta}$ extracted from the estimated variance-covariance matrix $\hat{V}[Ex|y]$according to the SIR model should be a linear combination of the vertical vectors of the matrix x. Hence, it can be assumed that $\tilde{y} \in R^n$ is a vector of the artificial response variable where:

$\hat{\eta} = \frac{1}{n} x\tilde{y}$

Thus, we can find an estimate of β by finding the solution to the following optimization problem:

$argmin\ \frac{1}{2n} \| \tilde{y} - X\beta \|_2^2 + \lambda \| \beta \|$ (14)

In 2019, researchers Li et al. provided an algorithm called Lasso-SIR, which is an effective algorithm for estimating the vector β. From formula (14), we find that:

$\tilde{y} = \frac{1}{C\hat{M}} M \acute{M} \acute{X} \hat{\eta}$ (15)

Where M is a matrix n×H, C is a constant, M is the largest latent value of

$\boldsymbol{\acute{\Lambda}_H = \frac{1}{H} X_H \acute{X}_H}$

$\eta\hat{}$ is the latent vector $X_H = \frac{XM}{C}$. Therefore, when multiplying the latent vector $\hat{\eta}$ by the latent value $\widehat{M}$, we find that formula (15) is satisfied and we have

$$\hat{\eta} = \frac{1}{n} x\ \tilde{y} \qquad (16)$$

Below is the Lasso-SIR algorithm as indicated by researchers (Li et al.) in 2019.

1-Let $\hat{\lambda}$ and $\hat{\eta}$ be the first eigenvalue and eigenvector of $\hat{\Lambda}_H$, respectively ;

2-Let $\tilde{y} = \frac{1}{C\hat{\lambda}} MM^{\tau} X^{\tau} \hat{\eta}$ and solve the Lasso optimization problem

$$\hat{\beta}(\mu) = argmin\ \mathcal{L}_{\beta}.\ where\ \mathcal{L}_{\beta,i} = \frac{1}{2n} \| \tilde{y} - X^{\tau}\beta \|_2^2 + \mu \| \beta \|_1,$$

Where $\mu = C\sqrt{\frac{\log(p)}{n\hat{\lambda}}}$ for sufficiently large constant C;

3-Estimate p_{β} by $p_{\hat{\beta}(\mu)}$

3-Bayesian Lasso Sliced Inverse Regression (BLSIR)[9] [10]

Based on model (6) and formula (14), we can study the inverse Lasso model according to Bayes' method and estimate the values of (β)Col that represent all linear combinations of the β space (space spanned). Because there is no previous study to employ the Bayes method for the Lasso-SIR model, we will rely on the method proposed by researchers (Casella, Park) in 2008 to estimate the parameters of the SIR regression model as a non-linear model. Here, the hierarchical model will be assumed for the studied model in addition to the subsequent distributions and the Giββs algorithm to generate samples.

Hierarchical model:

As we mentioned previously and based on model (6) and (14) and assuming that β is a variable that follows the Laplace distribution and after reformulating formula (14), we have the following:

$argmin_{\beta}(\tilde{y} - X\beta\acute{)}\ (\tilde{y} - X\beta) + \lambda \sum_{j=1}^{k} |\beta_j|$

The pyramid model can be written as follows:

$$\begin{aligned}
&\tilde{y}|X,\beta,\sigma^2 \backsim N_n(X\beta,\sigma^2 I_n),\\
&\qquad \beta|\sigma^2,\tau_1^2,\dots,\tau_k^2 \backsim N_k(o_p,\sigma^2,w_{\tau}),\\
&w_{\tau} = diag(\tau_1^2,\dots,\tau_k^2),\\
&\qquad \sigma^2,\tau_1^2,\dots,\tau_k^2 \backsim \pi(\sigma^2)d\sigma^2\pi_{j=1}^{k}\frac{\lambda^2}{2}e^{\frac{\lambda^2\tau_j^2}{2}}d\tau_j^2 \qquad (17)\\
&\sigma^2,\tau_1^2,..,\tau_k^2 > 0,\\
&\sigma^2 \backsim Inverse\ Gemma(a,\gamma),\\
&\lambda^2 \backsim Gamma(\theta_1,\theta_2)
\end{aligned}$$

Where the normal mixed-exponential distribution is adopted

to represent the prior distribution (Laplace distribution) of the parameter β.

Complete conditional posterior distributions[9][10]

Based on the hierarchical model (17), we can assume that the complete conditional posterior distributions of the assumed parameters of the studied model are as follows:

The complete joint density function for all prior distributions
$f(y \mid \beta, \sigma^2, X)\pi(\sigma^2)\pi_{j=1}^{k}\pi(\beta_j|\tau_j^2, \sigma^2).\pi(\tau_j^2)$

Now we list the complete posterior distributions.

1- The complete posterior distribution of β is the multivariate normal distribution with mean $A^{-1}\acute{x}\tilde{y}$ and variance $\sigma^2 A^{-1}$ where $A = \acute{X}X + w_\tau^{-1}$.

2- The complete posterior distribution of σ^2 is the inverse gamma distribution with shape parameter $\frac{n-1}{2} + \frac{k}{2} + a$ and scale parameter $(\tilde{y} - X\beta\acute{)}\,(y - X\beta) / 2 + \acute{\beta}\ w_\tau^{-1}\beta/2 + \gamma$.

3- The complete posterior distribution of $\frac{1}{\tau_j^2}$ is the inverse Gaussian distribution with shape parameter λ^2 and mean parameter $\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}$.

4- The complete posterior distribution of λ^2 is a complete distribution of shape parameter $k + \theta_1$ and rate parameter $(\text{rate})\ \sum_{j=1}^{k}\frac{\tau_j^2}{2}$.

RLSIR Model Algorithm:

In this section, the computational steps of the Giββs algorithm will be included in generating samples from posterior distributions as follows, noting that all values of x will be converted to standard values:

1- Sampling ỹ: where samples of the variable ỹ will be generated, which is an artificial variable after assuming that x∽N(0,Σ) from the multivariate normal distribution with mean xβ and variance $\sigma^2 I_n$. Here, σ^2 is considered as **Σ** the variance and covariance matrix of the multivariate normal distribution.

2- Sampling σ^2 : where samples of the variable σ^2will be generated from the inverse gamma model with a shape parameter $\frac{n-1}{2} + \frac{k}{2} + a$ And a measurement parameter
$(\tilde{y} - x\beta\acute{)}(\tilde{y} - x\beta) / 2 + \acute{\beta}\, w_c^{-1}\beta / 2 + \gamma\ )$

3- β Preview: where the samples of the variable β will be generated from the multivariate normal distribution predictor. $(A^{-1}x\tilde{y}, \Sigma A^{-1})$.

4- Sample $\frac{1}{t_j^2}$: where the samples of the variable $\frac{1}{t_j^2}$ will be generated from an inverse Gaussian distribution with shape parameter λ^2 and mean parameter $\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}$.

5- Preview λ^2 : where λ^2 values will be generated from the gamma distribution.

$\lambda^2 \backsim Gamma(k + \theta_1, \sum_{j=1}^{k}\frac{\tau_j^2}{2})$

The Real Data

In this part , the researcher dealt with the analysis of real data related to the level of fat in the liver (response variable) and its relationship with a set of explanatory variables, as the data was collected through the records registered at the Specialized Center for Digestive Medicine and Surgery in Diwaniyah for the period from 1/1/2023 to 31/12/2023. This data represents patients with non-alcoholic fatty liver disease, of both sexes, as the researcher will apply the proposed method and the comparison methods referred to in the experimental aspect to this data with the aim of identifying the most important factors affecting the level of fat accumulation in the liver and contributing to the development of this disease in patients.

Regression Model Variables

1- The dependent variable (Y) is Liver Fat Percentage and represents the percentage of fat accumulation in the liver.

The independent variables can be detailed as follows:

2. X1: Age
3. X2: Obesity (BMI)
4. X3: Hypertension

5. X4: Unhealthy diet
6. X5: Insulin resistance
7. X6: Sedentary lifestyle (Physical Inactivity)
8. X7: Blood cholesterol level
9. X8: Type 2 diabetes
10. X9: Blood triglycerides level
11- X10: Liver enzymes level (ALT) and (AST)

The proposed method (BLSSIR) was applied in addition to the comparison methods, which are the traditional Lasso method (LASSO), the Bayesian Lasso method (BLASSO), and the segmented inverse regression method (SIR). For the purpose of comparing the proposed method with the comparison methods, as well as for the purpose of finding the parameter estimates for the study variables and identifying the most important variables with a significant effect on the response variable. The table below shows the estimated parameters according to the mentioned methods.

Table No. (1) Estimated values of regression model parameters for real data

Methods	BLSIR	LASSO	BLASSO	SIR
$\boldsymbol{\beta_1}$	0.001	0.056	0.014	2.015
$\boldsymbol{\beta_2}$	6.238	3.164	3.118	4.254
$\boldsymbol{\beta_3}$	0.009	1.954	1.221	2.021
$\boldsymbol{\beta_4}$	3.214	4.985	5.635	5.324
$\boldsymbol{\beta_5}$	7.120	4.958	2.001	2.685
$\boldsymbol{\beta_6}$	0.007	1.025	1.002	1.952
$\boldsymbol{\beta_7}$	0.002	3.256	0.031	0.072
$\boldsymbol{\beta_8}$	10.233	7.120	6.023	6.365
$\boldsymbol{\beta_9}$	6.125	1.854	2.232	2.791
$\boldsymbol{\beta_{10}}$	1.035	2.954	2.965	2.635

From the table above, and after determining the factors as follows:

1. Age variable (X1): We notice that the estimated values of parameter (β1) for this variable are very low in all methods, indicating a small effect of age on the response variable, as we find that the BLSSIR method excluded this variable.
2. Obesity variable (BMI) (X2): All estimated values of parameter (β2) indicate a significant positive effect of obesity on the response variable, with the highest estimate in the BLSIR method (6.254), which means that an increase in BMI is associated with an increase in the response variable.
3. Hypertension variable (X3): The estimated values of the parameter (β3) indicate a positive effect across all comparison methods, but we find that the BLSIR method excluded the impact of this variable on the dependent variable, indicating that hypertension does not have a significant effect on the response variable, and this was confirmed by the doctors at the center after presenting the results to them.
4. Unhealthy diet variable (X4): The estimated values of the parameter (β4) show a significant positive effect on the response variable, with the highest estimate in the BLASSO method, where the parameter value was 5.635.

5. Insulin resistance variable (X5): The estimated values of the parameter (β5) vary between methods, with the highest estimate in the BLSSIR method, where the parameter value was 7.120, which is the highest value of the parameter among the other methods, which indicates a direct effect on the dependent variable.
6. Physical Inactivity (X6): The estimated values of the parameter (β6) show a small positive effect on the response variable, with the lowest estimate in the BLSSIR method, as we find that the proposed method has excluded the effect of this variable and we did not notice this in the other methods.
7. Blood cholesterol level variable (X7): The estimated values of the parameter (β7) show a very small or no effect in all methods, indicating a non-significant effect of this variable on the response variable.
8. Type 2 diabetes variable (X8): The estimated values of the parameter (β8) show a large positive effect on the response variable, as we notice the highest estimate in the BLSSIR method, where the estimated parameter value was 10.233, which shows the superiority of the proposed method over other methods through the importance of this variable as it directly affects non-alcoholic fatty liver disease.
9. Triglycerides (X9) variable: The estimated values of the parameter (β9) show a significant positive effect on the response variable, with the highest estimate in the BLSSIR method.
10. Liver enzymes (X10) variable ((ALT and AST)): The estimated values (β10) show an almost equal positive effect across all methods, with slight differences.
11. We conclude from this that the proposed method is superior in estimating and identifying the influential variables and excluding variables that have no effect on the model, as we note that the BLSSIR method has excluded four variables (age, high blood pressure, sedentary lifestyle, blood cholesterol level), where the result was presented to specialist doctors who confirmed the validity of the results to a very large extent.
12. For the purpose of further confirmation, the MSE and MAE criteria were calculated for all methods and their values were as in the following table:

Table (2) MAE and MSE values

Methods	MSE	MAE
BLSIR	1.142	0.685
LASSO	2.965	0.856
BLASSO	2.265	0.693
SIR	2.395	0.725

From the above table, we notice that the BLSIR method obtained the lowest value for the MSE (1.142) and MAE (0.685), making it the best among the four methods in estimating the parameters of the study variables. It is followed in second place by the BLASSO method, which obtained an MSE of (2.265) and an MAE of (0.693), indicating its good performance but slightly lower than BLSIR. In third place came the SIR method, which obtained an MSE of (2.395) and an MAE of (0.725), reflecting an average performance. Finally, the LASSO method was the least effective based on the high values of the MSE (2.965) and MAE (0.856), making it the least accurate in estimation among the mentioned methods.

Conclusions

In this section, the most important conclusions reached by the researcher through his study of this thesis will be discussed in its theoretical, experimental and applied aspects. The most important conclusions were as follows:
1- The simulation results showed in all the tested examples that the proposed method (BLSIR) was superior after comparing it to the methods (LASSO), (BLASSO), and (SIR), and based on the two criteria (MSE) and (MAE), as it achieved the lowest value for the two criteria.
2-It has been observed that the proposed method is more accurate as the sample size increases, as well as when the correlation increases.
3- Four variables were excluded from the model using the proposed BLSIR method.
4-The most important variables that have an impact on non-alcoholic fatty liver disease are (obesity, unhealthy diet, insulin resistance, type 2 diabetes, level of triglycerides in the blood, level of liver enzymes).
5- Variables that have no effect on non-alcoholic fatty liver disease are (age, high blood pressure, sedentary lifestyle, blood cholesterol level).

Recommendations

Based on the conclusions reached from the experimental side and real data, a number of recommendations were reached, which are summarized as follows:

1- The study recommends using the BLSIR method in the case of high-dimensional data because of its accuracy in excluding variables that have no effect on the model.

2- Use the proposed method (BLSIR) in fields other than the medical field because of its accuracy in estimating and selecting variables.

3- Adopting the BLSIR method in the Specialized Center for Digestive Medicine and Surgery in Diwaniyah for the purpose of contributing to knowledge of the factors that affect non-alcoholic fatty liver disease for the purpose of avoiding or preventing it.

4- Use the Bayesian inverse Lasso method (BRLASSO) with the segmented inverse regression (SIR) method, as it is one of the good methods for estimating parameters as well as selecting variables.

References

1- Agresti, A. (2010). Analysis of ordinal categorical data (Vol. 656). John Wiley & Sons.

2- Alhamzawi, R., Yu, K., & Benoit, D. F. (2012). Bayesian adaptive Lasso quantile regression. *Statistical Modelling*, *12*(3), 279-297.

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

4- Fu, W., & Knight, K. (2000). Asymptotics for lasso-type estimators. The Annals of statistics, 28(5), 1356-1378.

5- Griffin, J. E., & Brown, P. J. (2007). Bayesian adaptive lassos with non-convex penalization.

6- Hasan Zeinah & Ali Omer . (2017). using sliced inverse regression with other methods in dimension reduction. thesis Submitted to The Council of the College of Administration and Economics at University of Baghdad.

7- Hesterberg, T., Choi, N. H., Meier, L., & Fraley, C. (2008). Least angle and ℓ 1 penalized regression: A review.

8- Kaur, S., & Ghosh, S. M. (2016). A survey on dimension reduction techniques for classification of multidimensional data. Int. J. Sci. Technol. Eng, 2(12), 31-37.

9- Li, K. C. (1991). Sliced inverse regression for dimension reduction. Journal of the American Statistical Association, 86(414), 316-327.

10-Lin, Q., Zhao, Z., & Liu, J. S. (2019). Sparse sliced inverse regression via lasso. *Journal of the American Statistical Association*.

11-Park, T., & Casella, G. (2008). The bayesian lasso. Journal of the american statistical association, 103(482), 681-686.

12-Shen, Y, Traganitis, P. A., & Giannakis, G. B. (2017, December). Nonlinear dimensionality reduction on graphs. In 2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP) (pp. 1-5). IEEE.

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

14-Wang, T., Xu, P., & Zhu, L. (2013). Penalized minimum average variance estimation. Statistica Sinica, 543-569.