The Continuum Regression Analysis with Preprocessing Variable Selection LASSO and SIR-LASSO

Adzkar Adlu Hasyr Suruddin[1*], Erfiani[1], and I Made Sumertajaya[1]
[1]Department of Statistic, IPB University, Bogor, Indonesia
*Corresponding author: adzkaradlu@apps.ipb.ac.id

Received: 5 September 2025
Revised: 14 January 2025
Accepted: 3 February 2025

ABSTRACT – **Analyzing high-dimensional data is a considerable challenge in statistics and data science. Issues like multicollinearity and outliers often arise, leading to unstable coefficients and diminished model effectiveness. Continuum regression is a useful method for calibration models because it effectively handles multicollinearity and reduces the number of dimensions in the data. This method condenses data into autonomous latent variables, resulting in a more stable, precise, and reliable model. It is possible to use the dimensionality reduction method without losing any important information from the original data. This makes it a useful tool for making calibration models work better. In the initial phase, minimizing dimensions via variable selection is crucial. The study aims to build and test the Continuum Regression calibration model using LASSO and SIR-LASSO variable selection preprocessing methods. SIR-LASSO is a method that integrates SIR with the variable selection capabilities of LASSO. This technique aims to handle high-dimensional data by identifying relevant low-dimensional structures. LASSO improves variable selection by applying a penalty to regression coefficients, reducing the impact of less significant or redundant variables. The integration improves SIR's efficacy in assessing high-dimensional data while also enhancing model stability and interpretability. This approach seeks to address the issues of multicollinearity and model instability. We conducted simulations using both low-dimensional and high-dimensional datasets to assess the efficacy of CR LASSO and CR SIR-LASSO. RStudio version 4.1.3 was used for the analysis. The "MASS" package was used to create data with a multivariate normal distribution. The "glmnet" package was used for LASSO variable selection, and the "LassoSIR" package was used for SIR-LASSO variable selection. In the simulation itself, LASSO surpasses SIR-LASSO in variable selection by yielding the lowest RMSEP value in every scenario. On the other hand, SIR-LASSO becomes less stable as the number of dimensions increases, which suggests that it is sensitive to large changes in variables. As shown by lower median RMSEP values across a range of sample sizes and situations, CR LASSO is usually better at making predictions than SIR-LASSO. The RMSEP distributions for LASSO are consistently tighter, which means that its performance is more stable and reliable compared to SIR-LASSO, whose data has more outliers and more variation. Even with a growing sample size, LASSO maintains its advantage, particularly when setting the value at 0.5. SIR-LASSO, although occasionally competitive, generally yields more variable results, particularly with larger sample sizes. Overall, LASSO appears to be a more reliable option for the CR model with pre-processed variable selection.**

Keywords– **continuum regression, High-dimensional, LASSO, SIR-LASSO, variable selection.**

I. INTRODUCTION

In regression analysis, the Ordinary Least Squares (OLS) method is a prevalent technique employed to describe the relationship between independent variables and dependent variables. Nonetheless, it is frequently necessary to address multicollinearity and outliers when implementing Ordinary Least Squares (OLS) [1]. Multicollinearity arises when two or more independent variables in a model exhibit a strong connection, resulting in unstable and difficult-to-interpret estimates of the regression coefficients [2]. These behaviours can induce biases in estimation and interpretation, hence diminishing the reliability of the regression model. Consequently, it is essential to consider the estimating method employed to enhance the reliability and accuracy of the analytical results.

Continuum Regression (CR) is an advancement of Least Squares Regression (LSR), Partial Least Squares Regression (PLSR), and Principal Component Regression (PCR) techniques, designed to address multicollinearity issues by diminishing the number of independent variables. This is achieved by condensing the data into new, independent latent variables with significantly reduced dimensions [3]. Introduce new variables in CR by optimizing the variance of independent variables and the covariance between the independent variable and the response variable. CR is implemented to finalize the calibration model in several case studies utilizing cross-validation index criteria, juxtaposed with varying adjustment parameters δ; the conclusion indicates that CR outperforms the findings of LSR, PLSR, and PCR [4]. Research by Setiawan investigated calibration models employing the CR approach, determining that CR offers advantages over PCR and PLSR in addressing multicollinearity issues across diverse independent variable matrix configurations [5]. Occasionally, when the dimensions of the data are very large, or when the number of observations is smaller than predictors induces singularities in the matrix structure of the independent variables, leading to computational issues [6]. Thus, dimensionality reduction entails decreasing the quantity of input variables to enhance efficiency. Initially, it is essential to diminish the dimensions from the original high-dimensional space (nxp) to a lower-

dimensional space (nxh), where $h<(n-1)<p$, while preserving the majority of the pertinent information from the original data; this procedure is referred to as preprocessing [7].

Common data compression techniques include Principal Component Analysis (PCA), Fourier transformation, wavelet transformation, pursuit projection, and the Least Absolute Shrinkage and Selection Operator (LASSO) [8]. LASSO is a pre-processing method used to select variables by shrinking the regression coefficients [9]. The LASSO process provides an alternative for variable selection and for reducing linear regression model coefficients to zero [10]. Another method for data compression is Sliced Inverse Regression (SIR), which replaces the original variables with a low-dimensional linear combination of predictors without losing information and without requiring model assumptions [11].

The study by Arwini on rainfall estimation using Continuum Regression (CR) with LASSO variable selection preprocessing showed that this approach enhances precision and yields fairly accurate predictions compared to using PCA preprocessing [8]. However, LASSO has limitations when the number of observations is smaller than predictors, as it only selects p variables to include in the model [12]. The LASSO model is also prone to bias and can be difficult to interpret due to information loss, particularly when an irrelevant predictor variable has a weak correlation with the response variable [10]. In contrast, the Sliced Inverse Regression (SIR) method has advantages over LASSO in identifying general patterns in high-dimensional data. SIR can reduce data dimensionality, capture nonlinear structures, address multicollinearity issues, enhance computational efficiency in high-dimensional data, and produce models that are easier to interpret [13]. Further developed the SIR method by incorporating a LASSO penalty into the SIR least squares formulation, resulting in the SIR-LASSO method [14]. SIR approach can be associated with many other dimension reduction methods, such as PCA, principal component regression, and PLS, as well as their derivatives [15]. The usual SIR, however, cannot work with problems where the number of predictors, p, exceeds the sample size, n, and can suffer when there is high collinearity among the predictors [16].

This research aims to enhance the Continuum Regression models by incorporating the LASSO variable selection preprocessing technique and the SIR-LASSO method. By adopting this approach, the study seeks to address the issue of multicollinearity, which is common in high-dimensional data, thereby improving the stability and accuracy of the regression model. The implementation of the SIR method is anticipated to boost computational efficiency and improve model interpretation by reducing data dimensions and capturing non-linear structures. Existing approaches to regression modeling in high-dimensional contexts, such as Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR), have been thoroughly investigated but often struggle with addressing multicollinearity and fall short in simultaneously managing variable selection and capturing non-linear relationships. Although the LASSO technique is well-known for its effectiveness in selecting relevant variables, its application alongside dimension reduction methods like SIR has been insufficiently explored, particularly within continuum regression models. This lack of integrated studies combining LASSO and SIR-LASSO underscores a critical gap in optimizing model accuracy and interpretability for high-dimensional data challenges. Recent developments in high-dimensional regression modeling focus on hybrid techniques that merge variable selection with dimension reduction. Approaches such as Sparse Partial Least Squares (SPLS) and Elastic Net have been introduced to address multicollinearity and enhance interpretability. However, these methods often require balancing trade-offs between predictive performance and computational efficiency. The integration of LASSO and SIR-LASSO offers a novel solution, combining the strengths of both methods to effectively manage multicollinearity while capturing key data structures. This research aims to push the boundaries of current methodologies by applying this combination to continuum regression, establishing an efficient framework for high-dimensional regression modeling.

II. LITERATURE REVIEW

A. Continuum Regression

Continuum regression is one method used to overcome multicollinearity [17]. Let $\mathbf{X}$ be a matrix of data of size $(\boldsymbol{n} \times \boldsymbol{p})$, and $\mathbf{y}$ is a vector of response variables of size $(\boldsymbol{n} \times \mathbf{1})$. Continuum regression was developed based on the classical linear regression model with a β parameter of size $(\boldsymbol{p} \times \mathbf{1})$. Mathematically, it can be expressed in the following equation:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon \tag{1}$$

where ε is an error vector of size (n × 1). The new (latent) variable in continuum regression is formulated in the model in the following equation and is formulated in the model as in the following equation:

$$\boldsymbol{y} = \boldsymbol{T}_h\xi + \varepsilon \tag{2}$$

with $\boldsymbol{T_h} = \boldsymbol{XW_h}$ and $\boldsymbol{W_h} = (\boldsymbol{w_1}, \boldsymbol{w_2}, \dots, \boldsymbol{w_h})$, the matrix contains h columns of variables with $\boldsymbol{h} < \boldsymbol{p}$ and is called the weighting matrix [18]. Where p is the number of independent variables, and h is the number of new variables from LASSO and SIR-LASSO selection. Stone and Brooks [4] formulate the vector $(\boldsymbol{w_i} = \boldsymbol{i} = \mathbf{1}, \mathbf{2}, \dots, \boldsymbol{h})$ as in the following equation:

$$\boldsymbol{w_i} = \arg max \left\{ Cov(\boldsymbol{x}_w, y)^2 (Var(\boldsymbol{x}_w))^{[\delta/(1-\delta)]-1} \right\} \tag{3}$$

with the constraints $\|\boldsymbol{w}_i\| = 1$ and $Cov(\boldsymbol{x}_{wi}, \boldsymbol{x}_{wj}) = 0$ for $i < j = 1,2, \dots, h$, the adjustment parameter δ is real number $0 \leq 0.5 < 1$. Equation (2) can be obtained with the values $\delta = (0, 0.5, 1)$, each of which is a generalization of the least squares, partial least squares, and PCR methods.

The estimation of the parameter ξ in equation (2) is carried out using the least squares method, which is formulated as follows:

$$\hat{\xi} = (\boldsymbol{T}_h^T \boldsymbol{T}_h)^{-1} \boldsymbol{T}_h^T y \quad (4)$$

$$\hat{\beta} = \boldsymbol{W}(\boldsymbol{T}_h^T \boldsymbol{T}_h)^{-1} \boldsymbol{T}_h^T y \quad (5)$$

$$\hat{y} = \boldsymbol{X}\boldsymbol{W}_h \hat{\xi} \quad (6)$$

B. Least Absolute and Shrinkage Selection Operator

The Least Absolute Shrinkage and Selection Operator (LASSO), introduced by [9], is widely used for constructing models that yield accurate results. LASSO serves as an alternative to the least squares method, offering a penalty approach that aids in variable selection and helps mitigate overfitting issues [19]. The Lagrangian equation also expresses the coefficient estimator, LASSO, as follows:

$$\hat{\beta}_{LASSO} = \arg min \left\{ \sum_{i=1}^{n} \left(\boldsymbol{Y}_i - \sum_{j=1}^{p} \beta_j \boldsymbol{X}_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} ; \ \lambda \geq 0 \quad (7)$$

with $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$ and λ is a penalty parameter (regularizer) that controls the amount of shrinkage. If $\lambda = 0$ then the LASSO estimator gives the same results as the least squares estimator. If $\lambda \to \infty$ then forces all the coefficients to be zero. If λ is large enough then the estimated coefficient will be exactly zero, so it can function as a selection variable. One way to find the optimal λ value is to use the Cross Validation (CV) method with a minimum CV value.

C. Sliced Inverse Regression (SIR)

Li introduced the Sliced Inverse Regression (SIR) model, which is one of the most general models for estimating adequate dimension reduction [11]. For a regression problem, a general nonlinear model can summarize the relationship between observations $y_{n\times 1}$ and predictor $x_{n\times p}$ as:

$$y = f(\boldsymbol{x}\beta_1, \boldsymbol{x}\beta_2, \dots, \boldsymbol{x}\beta_k, \varepsilon) \quad (8)$$

Where $\beta_s, \ s = 1, \dots, k$, are unknown vector denoting the contribution of each predictor in $\boldsymbol{x}$, $k(k < p)$ is the dimension we aim to reduce $\boldsymbol{x}$ to, $f(.)$ is a an unknown nonlinear function of k inputs, and ε denotes zero-mean random noise independent to $\boldsymbol{x}$ via k linear combinations of predictors:

$$y \equiv (\boldsymbol{x}|P_s\boldsymbol{x}), S = span(\beta_1, \dots, \beta_k) \quad (9)$$

Where P_s denotes the projection operator onto the k-dimensional subspace S. Therefore, we only need to estimate S generated by β_s to effectively reduce dimensionality. The efficient dimension reduction (e.d.r) for (8) can be estimated by the SIR method by finding an inverse regression curve $E(\boldsymbol{x}|y)$. It was proved in Li that if $\boldsymbol{x}$ has been standardized to have zero mean and identity covariance, the inverse regression curve will fall into the e.d.r space. SIR can estimate the inverse regression curve $E(\boldsymbol{x}|y)$ as the sliced mean value of $\boldsymbol{x}$, which is obtained by slicing y into several groups and partitioning $\boldsymbol{x}$ into several slices according to the values of y. Equation (8) shows the model when the response variable Y depends on the p-dimension. The SIR method divides the model into several slices based on the Y value, then combines information from all slices. SIR can be calculated through several conversion and arithmetic methods [20]. The following are the calculation stages of the SIR method using conversion and arithmetic:

Stage 1: Standardize $\boldsymbol{x}$ and estimate the sample mean $\overline{\boldsymbol{x}} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{i=1}^{n} \boldsymbol{x}_i$ and the sample covariance matrix $\hat{\Sigma}_{\boldsymbol{x}} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T (\boldsymbol{x}_i - \overline{\boldsymbol{x}})$.

Stage 2: Bin y into m slices, $G_1, \dots, G_m$, and calculate the proportion of $\boldsymbol{y}_i$ that falls into the slice G_j , $j = 1, \dots, m$, as $\hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} \xi_j(\boldsymbol{y}_i)$, where $\xi_j(\boldsymbol{y}_i)$ equals 1 or 0 depending on whether y_i falls into the jth slice or not.

Stage 3: For each slice, calculate the sliced mean $\overline{\boldsymbol{x}}_j = \frac{1}{\hat{p}_j} \sum \boldsymbol{x}_i$ and weighted covariance $\hat{\Sigma}_W = \frac{1}{(n-1)\,\hat{p}_j} \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T (\boldsymbol{x}_i - \overline{\boldsymbol{x}})$ for the sliced means.

Stage 4: Conduct a weighted PCA for $\hat{\Sigma}_{\boldsymbol{x}}$ and $\hat{\Sigma}_W$ in the following way: $\hat{\Sigma}_W \hat{\beta}_s = \hat{\lambda}_s \hat{\Sigma}_{\boldsymbol{x}} \hat{\beta}_s$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$. By solving this generalized eigen-decomposition problem. SIR directions β_s can be estimated.

In Stages 2 and 3, we can obtain the estimates of the standardized inverse regression curve $E(\boldsymbol{x}|y)$. It should be noticed that we only need to transform the sliced G_j to $\overline{\boldsymbol{x}}_j$ become PCA, rather than to transform all $\boldsymbol{x}_i$.

D. *SIR-LASSO*

The SIR approach provides a method for estimating low-dimensional subspaces that contain the most important information about the relationship between predictor and response variables. This subspace is represented by Span$(\hat{\beta})$, where the elements of $\hat{\beta} \in R^{p\times d}$ are often not equal to zero. When we have many predictors (especially highly correlated ones, we do not need to use them all to predict the response variable accurately. What needs to be done is to select a subset of predictors that is informative and not redundant. This is done to overcome problems such as multicollinearity and overfitting, which occur when using to many predictors [21].

Lin introduced LASSO with SIR with the aim of replacing the Ordinary Least Squares (OLS) estimator in the SIR

algorithm with the LASSO estimator, and $\sum_{l=1}^{p}|\beta_{jl}| \leq \tau$ is the Lasso restriction, β_{jl} is the lth coordinate of β_j, and τ is the shrinkage factor [14]. SIR-LASSO can be generalized to the multiple index model (8). Let $\hat{\lambda}_i, 1 \leq i \leq d$, be the d-top eigenvalues $\hat{\Lambda}_H$ and $\hat{\eta} = \hat{\eta}_1, \dots, \hat{\eta}_d$ be the corresponding eigen-vectors. We define a multivariate pseudo response

$$\tilde{Y} = \frac{1}{c}\boldsymbol{M}\boldsymbol{M}^T\boldsymbol{X}^T\hat{\eta}\ diag\left(\frac{1}{\hat{\lambda}_1}, \dots, \frac{1}{\hat{\lambda}_s}\right) \tag{10}$$

Then apply the LASSO on each column of the pseudo response matrix to produce the corresponding estimate. For each $1 \leq i \leq d$, solve the LASSO optimization problem

$$\hat{\beta}_{SIR-LASSO} = \arg min\, \mathcal{L}_{\beta,i} \tag{11}$$

Where $\mathcal{L}_{\beta,i} = \frac{1}{2n}\left\|\tilde{Y}_{*,i} - \right\|_2^2 + \mu_i\|\beta\|_1$ and $\mu_i = C\sqrt{\frac{\log(p)}{n\hat{\lambda}_i}}$ for sufficiently large constant C. The following are the stages of SIR via LASSO:

Stage 1: Determine the values of τ, then initial $\beta_j = \hat{V}_j$.

Stage 2: Find $t(y) = \left(\hat{E}(X|y_1), \hat{E}(X|y_2), \dots, \hat{E}(X|y_n)\right)^T \beta_j$, $\hat{E}(X|y_i)$ refers to the estimated sample evaluated at y_i, $i = 1,2, \dots, n$.

Stage 3: β_j updated as LASSO solution with $t(y)$ as a response, X predictor, and shrinkage factor τ.

Stage 4: β_j normalized as $\beta_j = \beta_j/|\beta_j|$. If $j > 1$, β_j orthonomalized in such a way $\beta_j^T\beta_j = 1$ and $\eta_K^T\hat{\Sigma}_x\beta_j = 0$, with $\hat{\eta}_k = 1,2, \dots, j$ is the first estimated SIR $j-1$.

Stage 5: Stage 2-4 are repeated until β_j convergen.

III. METHODOLOGY

This research uses simulation data analyzed by RStudio 4.1.3 software. The simulation data generate several levels correlation (r). This research refers to [14], with some adjustments of the correlation level ($r = 0.5; 0.9$). This study's optimal δ^* value in Continuum Regression between ($0 < \delta^* < 1$) is $\delta^* = 0.5$. A small δ^*, such as 0.5, reaches the optimal value faster than a larger δ^* value can extract more relevant information from the independent variables (the results of the experiment δ^* between the values 0 to 1) [22]. When $\delta^* = 0$ Continuum Regression includes Least Square Regression, $\delta^* = 0.5$ Continuum Regression includes Partial Least Squares Regression, and $\delta^* = 1$ Continuum Regression includes Principal Component Regression. This simulation was conducted on low low dimensional data ($n = 75, p = 50$) and high dimensional data ($n = 75, p = 100; 150$). The simulation study consists of the following stages:

1. Determine the number of the observations $n = 75$ and the number of explanatory variables $p = 50; 100; 150$
2. Generating a p-dimensional vector of explanatory variables for the *ith* observation $\boldsymbol{x}_i = \left(x_{1i}, x_{2i}, \dots, x_{pi}\right)^T$through a multivariate normal distribution $\boldsymbol{x}_i \sim N_p(0, \Sigma)$ where is the covariance matrix $\Sigma = r_{p\times p}; r_{jk} = \rho^{|j-k|}; r = \{0.5; 0.9\}$ and $i = 1,2,3, \dots, n; j, k = 1,2,3, \dots, p$
3. Generate regression model: $y = X\beta + \varepsilon$, where $\varepsilon \sim N(0,1)$ for $p = 5$ (ordered significant variables).
4. Perform preprocessing with LASSO and SIR-LASSO variable selection

$$\hat{\beta}_{Lasso} = \arg min\left\{\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}\beta_j\boldsymbol{X}_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

$$\hat{\beta}_{LASSO-SIR} = argmin\left\{\frac{1}{2n}\left\|\tilde{Y}_{*,i} - \boldsymbol{X}^T\beta\right\|_2^2 + \mu_i\|\beta\|_1\right\}$$

Where Y_i is the observed value of the response variable at the i-th observation $\boldsymbol{X}_i = (\boldsymbol{X}_{1i}, \dots, \boldsymbol{X}_{ij})$, $\beta_j\boldsymbol{X}_{ij}$ is the predicted value for the i-th observation, obtained by multiplying the vectors feature $\boldsymbol{X}_i$ with regression coefficient vector β [9]. $\tilde{Y}_i$ is multivariate pseudo response at the i-th observation $\boldsymbol{X}$, $\boldsymbol{X}^T\beta$ is the predicted value for the i-th observation, obtained by multiplying the vectors feature $\boldsymbol{X}$ with the regression coefficient vector β. The selection of independent variables LASSO and SIR-LASSO selection based on the optimum lambda (λ) with a minimum value of Mean Squared Error Cross-Validation (MSECV) [23]. The procedure for k-fold cross-validation is as follows:

   a. Randomly divides data into k parts into k subsamples.
   b. For each k subsamples, one subsample will be used as *testing data* and (k-1) subsamples as *training data*.
   c. The optimum λ is obtained based on the minimum MSECV value [24]

$$MSECV = \frac{1}{k}\sum_{k=1}^{N}(y_i - \hat{y}_{-k}(x_i))^2 \tag{12}$$

   $\hat{y}_{-k}(x_i$ is the predicted response value for x_i when the model is obtained from data without involving the k-th subsample, and y_i is the i-th response variable in the testing data.

   d. Repeat steps (b) to (d) k times to obtain the minimum CV. Selection of predictor variables is based on selecting the optimum lambda value with the smallest cross-validation value.
5. Use the CR approach to model using selection variables LASSO and SIR-LASSO

$$\boldsymbol{y} = \boldsymbol{T}_h\xi + \varepsilon$$

6. Repeat steps 2 through 7 to 1000 times.
7. Calculate the Root Mean Square Error of Prediction (RMSEP) and R^2 value:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (13)$$

$$R^2 = 1 - \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}} \quad (14)$$

y_i is the observed value for the i-th data, and $\hat{y}_i$ is the predicted value for the i-th data obtained from the prediction model. $\bar{y}_i$ is the average of the actual observation values, $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the sum of squares of residuals, and $\sum_{i=1}^{n}(y_i - \bar{y}_i)^2$ is the total sum of squares.

8. Compare the model CR LASSO and CR SIR-LASSO obtained by the RMSEP and R^2 values.

V. RESULTS AND DISCUSSIONS

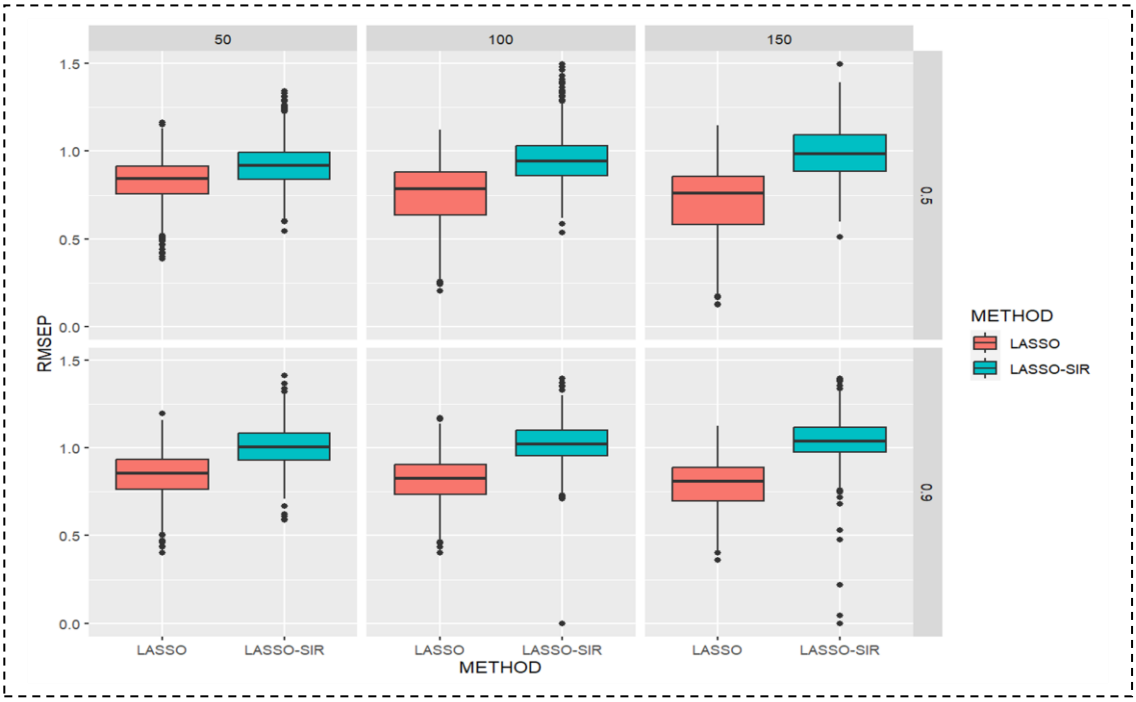

Figure 1 Boxplot RMSEP value on CR

The boxplot comparison reveals that the LASSO method generally outperforms SIR-LASSO in terms of predictive accuracy, as indicated by lower median RMSEP values across varying sample sizes and conditions. LASSO consistently tighter distributions of RMSEP, suggesting more stable and reliable performance compared to SIR-LASSO, which exhibits greater variability and more outliers in its results. Despite the increasing sample size, LASSO maintains its advantage, particularly under the condition marked the value 0.5. SIR-LASSO, while sometimes competitive, tends to produce more variable outcomes, especially at larger sample size. Overall, LASSO seems to more dependable choice for predictive modelling in the scenarios depicted in **Figure 1**.

We assessed the goodness-of-fit of the CR LASSO and CR SIR-LASSO models using their RMSEP and R^2 values. RMSEP is a metric that assesses the accuracy of a prediction model. The lower the RMSEP number, the lower the average model prediction error. R^2 or coefficient of determination, is a statistical measure that indicates how well the observed data matches the model expectations. The greater the R^2 score, the better the model produces [25]. **Table 1** compares the RMSEP and R^2 values for LASSO and SIR-LASSO preprocessing in different settings.

Table 1 Average RMSEP and R^2

Method	r	p	Evaluation	
			RMSEP	R^2
LASSO	0.5	50	0.830	0.953
SIR-LASSO			0.924	0.943
LASSO	0.9		0.840	0.968
SIR-LASSO			1.007	0.955
LASSO	0.5	100	0.750	0.961
SIR-LASSO			0.953	0.940
LASSO	0.9		0.815	0.970
SIR-LASSO			1.025	0.952
LASSO	0.5	150	0.705	0.965
SIR-LASSO			0.993	0.935
LASSO	0.9		0.907	0.972
SIR-LASSO			1.041	0.950

Average RMSEP (Root Mean Square Error of Prediction) and R^2 values for two methods, LASSO and SIR-LASSO, across different scenarios defined by the parameters r (with values of 0.5 and 0.9) and p representing sample sizes of 50, 100, and 150. Generally, LASSO outperforms SIR-LASSO in terms of RMSEP, consistently achieving lower error values across all scenarios. Specifically, LASSO shows a noticeable advantage in both the (r = 0.5) and (r = 0.9) conditions, with RMSEP values remaining below 1 in every case, whereas SIR-LASSO tends to have higher RMSEP values, especially as the sample size increases.

In terms of R^2, which measures the proportion of variance in the dependent variable that is predictable from the independent variables, both methods perform well, with R^2 values generally above 0.94. LASSO tends to achieve slightly higher R^2 values across most conditions, indicating a marginally better fit and predictive accuracy. The consistency of R^2 values across different conditions suggests that both methods can explain a significant portion of the variability in the data, but LASSO offers a more robust and reliable model with better generalization across different sample sizes and conditions.

Nonetheless, practical implementations may provide more difficulties, including nonlinear connections and heteroscedasticity, which could affect model efficacy. Future research may investigate alternate hybrid methodologies, the adaptive calibration of SIR-LASSO parameters, or the evaluation of CR with other variable selection techniques. Implementing these strategies on actual datasets, whether genetic, financial, or environmental data, would further substantiate their efficacy in practical applications.

V. CONCLUSIONS AND SUGGESTIONS

Continuum Regression (CR) is a good way to solve these problems, especially when used with variable selection methods like LASSO and SIR-LASSO. LASSO improves model stability by penalizing less significant variables, while SIR-LASSO integrates dimension reduction with feature selection. Simulations indicated that CR LASSO consistently outperformed CR SIR-LASSO by yielding lower RMSEP values and displaying more stable predictions across various conditions. SIR-LASSO, though sometimes competitive, exhibited heightened unpredictability and sensitivity to changes in dimensionality. LASSO is the more reliable option for preprocessing in CR calibration models.

To make Continuum Regression (CR) work better with high-dimensional data and outliers, future research could look into hybrid methods that combine the best parts of both LASSO and SIR-LASSO. Although LASSO demonstrates more stability and reliability, incorporating adaptive approaches or refining parameter selection in SIR-LASSO should alleviate its susceptibility to dimensional variations and the impact of outliers. Furthermore, integrating CR with other sophisticated variable selection techniques, such as elastic net or robust principal component analysis (PCA), may augment resilience to outliers and elevate model efficacy. Evaluating these methodologies across many datasets and practical applications would enhance the generalizability of the results.

REFERENCES

[1] K. Lakshmi, B. Mahaboob, M. Rajaiah, and C. Narayana, "Ordinary least squares estimation of parameters of linear model," *Journal of Mathematical and Computational Science*, vol. 11, no. 2, pp. 2015–2030, 2021, doi: 10.28919/jmcs/5454.

[2] M. Tsagris and N. Pandis, "Multicollinearity," May 01, 2021, *NLM (Medline)*. doi: 10.1016/j.ajodo.2021.02.005.

[3] X. Chen and L. P. Zhu, "Connecting continuum regression with sufficient dimension reduction," *Stat Probab Lett*, vol. 98, pp. 44–49, Mar. 2015, doi: 10.1016/j.spl.2014.12.007.

[4] M. Stone and R. J. Brooks, "Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 52, no. 2, pp. 237–258, 1990.

[5] K. Setiawan Notodiputro, "Regresi Kontinum dengan Prapemrosesan Transformasi Wavelet Diskret (Continum Regression with Discrete Wavelet Transformation Preprocessing)," *Jurnal ILMU DASAR*, vol. 8, no. 2, pp. 103-109, 2007.

[6] S. M. Ajeel and H. A. Hashem, "Comparison Some Robust Regularization Methods in Linear Regression via Simulation Study," *Academic Journal of Nawroz University*, vol. 9, no. 2, p. 244, Aug. 2020, doi: 10.25007/ajnu.v9n2a818.

[7] S. Sivaranjani, S. Ananya, J. Aravinth, and R. Karthika, "Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction," in *2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021*, Institute of Electrical and Electronics Engineers Inc., Mar. 2021, pp. 141–146. doi: 10.1109/ICACCS51430.2021.9441935.

[8] A. Arwini, A. H. Wigena, and A. Mohamad Soleh, "Continuum Regression Modeling with LASSO to Estimate Rainfall," 2020. doi: 10.29322/ijsrp.10.10.2020.p10651.

[9] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J R Stat Soc Series B Stat Methodol*, vol. 58, no. 1, pp. 267–288, 1996.

[10] S. Agus Mohammad and Aunuddin, "LASSO: SOLUSI ALTERNATIF SELEKSI PEUBAH DAN PENYUSUTAN KOEFISIEN MODEL REGRESI LINIER," *Forum Statistika Dan Komputasi*, vol. 18, no. 1, 2013.

[11] K.-C. Li, "Sliced inverse regression for dimension reduction," *J Am Stat Assoc*, vol. 86, no. 414, pp. 316–327, 1991.

[12] A. F. Fikri, W. Agwil, and D. Agustina, "PERFORMA TEKNIK REGULARISASI DALAM PENANGANAN MASALAH MULTIKOLINIERITAS," Journal UNIB, vol.2, no. 1, pp. 45-51, 2022. [Online]. Available: https://ejournal.unib.ac.id/diophantine, 2022. [Online]. Available: https://ejournal.unib.ac.id/diophantine,

[13] Y. Tu, Y. S. Hung, L. Hu, and Z. Zhang, "PCA-SIR: a new nonlinear supervised dimension reduction method with application to pain prediction from EEG," in *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, IEEE, 2015, pp. 1004–1007.

[14] Q. Lin, Z. Zhao, and J. S. Liu, "Sparse Sliced Inverse Regression via Lasso," *J Am Stat Assoc*, vol. 114, no. 528, pp. 1726–1739, Oct. 2019, doi: 10.1080/01621459.2018.1520115.

[15] S. Girard, H. Lorenzo, and J. Saracco, "Advanced topics in Sliced Inverse Regression," *J Multivar Anal*, vol. 188, 2022, doi: 10.1016/j.jmva.2021.104852ï.

[16] L. Li and X. Yin, "Sliced inverse regression with regularizations," *Biometrics*, vol. 64, no. 1, pp. 124–131, 2008, doi: 10.1111/j.1541-0420.2007.00836.x.

[17] S. Sutikno, S. Setiawan, and H. Purnomoadi, "Statistical downscaling output GCM modeling with continuum regression and pre-processing PCA approach," *IPTEK The Journal for Technology and Science*, vol. 21, no. 3, 2010.

[18] I. Ismah, E. Erfiani, A. H. Wigena, and B. Sartono, "Performance Analysis of Robust Functional Continuum Regression to Handle Outliers," *InPrime: Indonesian Journal of Pure and Applied Mathematics*, vol. 6, no. 1, pp. 52–62, 2024.

[19] S. K. Safi, M. Alsheryani, M. Alrashdi, R. Suleiman, D. Awwad, and Z. N. Abdalla, "Migration Letters Optimizing Linear Regression Models with Lasso and Ridge Regression: A Study on UAE Financial Behavior during COVID-19," vol. 20, no. 6, pp. 139–153, 2023, [Online]. Available: www.migrationletters.com

[20] W. K. Härdle and L. Simar, *Applied multivariate statistical analysis*. Springer Nature, 2019.

[21] L. Li, "Sparse sufficient dimension reduction," *Biometrika*, vol. 94, no. 3, pp. 603–613, 2007.

[22] Z. Xie, X. Feng, X. Chen, and G. Huang, "Optimizing a vector of shrinkage factors for continuum regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 206, p. 104141, Nov. 2020, doi: 10.1016/J.CHEMOLAB.2020.104141.

[23] S. Lee, M. H. Seo, and Y. Shin, "The lasso for high dimensional regression with a possible change point," 2015. [Online]. Available: https://academic.oup.com/jrsssb/article/78/1/193/7040660

[24] F. Emmert-Streib and M. Dehmer, "High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection," Dec. 01, 2019, *MDPI*. doi: 10.3390/make1010021.

[25] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput Sci*, vol. 7, p. e623, 2021.