STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–11.
Published online in International Academic Press (www.IAPress.org)

linearized shrinkage estimator for Semiparametric Regression Model

Nadwa Khazaal Rashad [1], Nawal Mahmood Hammood [1], Zakariya Yahya Algamal [2,*]

[1]*Department of Management Information Systems, College of Administration and Economics, University of Mosul, Mosul, Iraq*
[2]*Department of Statistics and Informatics, University of Mosul, Mosul, Iraq*

Abstract Semiparametric regression models are an appropriate tool to study complex data sets since they make use of elements of parametric and nonparametric regression. The nonparametric methods are effective when there are variables that operate by a particular rule, and others that have their effects that are more intricate or hard to predict. With the help of semiparametric regression model, researchers run the risk of revealing the problem of multicollinearity. This is the reason why it is important to consider the shrinkage parameters when analyzing the Liu estimator data. Within the Liu estimator, the parameter of shrinkage is set according to several steps which are discussed and described. After this, the slime mould algorithm (SMA) has been applied in selecting the appropriate parameter of shrinkage in Liu regression when dealing with issues of multicollinearity. In our study, it was found that there are estimators that would significantly improve on mean squared error when compared to their counterparts.

Keywords Semiparametric model, slime mould algorithm, shrinkage parameter, Liu estimator

DOI: 10.19139/soic-2310-5070-3363

1. Introduction

Regression models are fundamental tools for studying the relationship between a response variable and a set of explanatory variables in many scientific fields, including economics, epidemiology, engineering, and the social sciences. Classical parametric regression models, such as the linear regression model, specify the functional form of the mean response using a finite-dimensional parameter vector. This parametric specification is attractive because it is simple, interpretable, and computationally convenient, and it allows the use of powerful likelihood-based and least-squares methods for estimation and inference.

However, parametric models are often too restrictive to capture the complex, nonlinear, or heterogeneous relationships that arise in real data. When the presumed functional form is misspecified, parameter estimates can be biased and inconsistent, leading to misleading conclusions. Nonparametric regression models relax the functional form assumptions by allowing the regression function to be essentially unrestricted within a large function space. Methods such as kernel regression, local polynomial smoothing, and splines can flexibly approximate a wide range of relationships. This flexibility comes at a cost: nonparametric estimators typically converge more slowly, suffer from the curse of dimensionality in multivariate settings, and may be less interpretable, especially when the primary scientific interest lies in the effect of specific covariates.

Semiparametric regression models have emerged as a compromise between these two extremes. They combine a finite-dimensional parametric component with an infinite-dimensional nonparametric component. In doing so, semiparametric models aim to retain the interpretability and efficiency of parametric models for key covariate

[*] Correspondence to: Zakariya Yahya Algamal (Email: zakariya.algamal@uomosul.edu.iq). Department of Statistics and Informatics, University of Mosul, Mosul, Iraq.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

effects, while still allowing substantial flexibility to model complex and unknown relationships in other parts of the model. This balance makes semiparametric regression particularly attractive in modern applications where both interpretability and flexibility are required.

Semiparametric regression models have received considerable attention in statics and econometrics, because of their flexibility in modeling events [1, 2, 21, 22, 23]. Consider a semiparametric regression model given by

$$y_i \, = \, x_i\beta \, + \, f(t_i) \, + \, \varepsilon_{i\,}, \quad i = 1, 2, ..., n, \tag{1}$$

where $x_i \, = (x_{i1}, x_{i2}, . . . , x_{ip})$ is a vector of explanatory variable, $\beta \, = \, (\beta_{1\,,} \beta_{2\,,} . . . \, , \, \beta_p)'$ is an unknown p-dimensional parameter vector, the t_i's are known and non-random in some bounded domain $D \subset \mathbb{R}, \;\; f(t_i)$ is an unknown smooth function and ε_i's are independent and identically distributed random error with mean 0, variance σ^2, which are independent of $(x_i \, , \, t_i)$ [3, 24, 25, 26, 27, 28].

Most of the approaches for the semiparametric regression model are based on different nonparametric regression procedures. The have been several approaches to estimating β and $f(.)$. An alternative approach to nonparametric procedure is differencing methodology. This incoming ,used differences to remove the trend in the data that arises from the function $f(.)$and does not require an estimator of the function $f(.)$and often called difference-based procedure. Provided that $f(.)$is differentiable and the t ordinates are closely spaced , it is possible to remove the effect of the function $f(.)$ by differencing the data appropriately [29, 30, 31, 32, 33, 34, 35]. In model Eq. (1), [5, 36, 37] concentrated on estimation of the linear component and used difference-based estimation procedure is optimal in the sense that the estimator of the linear component is asymptotically efficient and the estimator of the nonparametric component is asymptotically minimax rate optimal for the semiparametric model used higher order differences for optimal efficiency in estimating the linear party by using special class of difference sequences [38, 39, 40, 41, 42].

Now consider a semiparametric regression model in the presence of multicollinearitry. The existence of multicollinearity may lead to wide confidence intervals for the individual parameters or linear combination of the parameters and signs. For our purpose we only employ the ridge regression concept due to [15, 43, 44, 45], to combat multicollinearity. There are a lot of work adopting ridge regression methodology to overcome the multicollinearity problem.

2. Differencing Approach

In this section, we use a difference-based technique to estimate the linear regresstion coefficient vector β. This technique has been used to remove the nonparametric component in the semiparametric regression model by varios authors [16, 17, 46, 47, 48].Consider the following semiparametric regression model

$$y \, = \, x\beta \, + f(t) \, + \, \varepsilon, \tag{2}$$

Where $y \, = \, (y_1 \, , \, y_2 \, , \, . . . \, , \, y_n)' \, , \, x \, = \, (x_1 \, , x_2 \, , . . . , \, x_n)'$ is then $n \times p$matrix, $f(t) = (f(t_1) \, , . . . \, , \, f(t_n))'$ and $\varepsilon = \, (\varepsilon_1, . . . \, , \, \varepsilon_n)'$. We assume that in general, ε is a vector of disturbances distributed with $\mathrm{E}(\varepsilon) \, = \, 0$and $\mathrm{E}(\varepsilon\varepsilon') = \sigma^2 V$, where $V$is symmetric, positive definite known matrix and $\sigma^2$is an unknown parameter.

[4, 49, 50, 51, 52] suggested estimating β on the basis of the m^{th}order differencing equation when $V \, = \, I_n$ as

$$\sum_{j=0}^{m} d_j y_{i-j} = \left(\sum_{j=0}^{m} d_j x_{i-j} \right) \beta \, + \sum_{j=0}^{m} d_j f(t_{i-1}) \, + \sum_{j=0}^{m} d_i \varepsilon_{i-j}, \tag{3}$$

where $d_0, \, d_1 \, , \, . . . \, , d_m$ are differencing weights.

Suppose $t_i$are equally spaced on the unit interval and $f'(.) \leq L$. By the mean value theorem, for some $t_i^* \, \in \, [t_{i-1}, t_i]$we have

$$f(ti) - \, f(ti - 1) \, = f'(t_i^*)(t_i - \, t_{i-1}) \leq \, \frac{L}{n}.$$

Note that with $m = p = 1$ from Eq. (3) we have

$$\begin{aligned} y_i - y_{i-1} &= (x_i - x_{i-1})\beta + f(t_i) - f(t_{i-1}) + \varepsilon_i - \varepsilon_{i-1} \\ &= (x_i - x_{i-1})\beta + O(\tfrac{1}{n}) + \varepsilon_i - \varepsilon_{i-1} \cdot \\ &\cong (x_i - x_{i-1})\beta + \varepsilon_i - \varepsilon_{i-1} \cdot \end{aligned}$$

We then estimate the linear regression coefficient β by the ordinary least-square estimators based on the differences [53, 54, 55]. Then we obtainthe least-squares estimate $\overset{\wedge}{\beta}_{diff} = \frac{\sum_{i=2}^{n}(x_i - x_{i-1})(y_i - y_{i-1})}{\sum_{i=2}^{n}(x_i - x_{i-1})^2}$.
Now let $d = (d_0, \ . \ . \ . \ , d_m)$ be a $(m+1)$-vector, where $m$is the order of differencing and $d_0 \, , \, d_1 \, , \ . \ . \ . \ , d_m$ are differencing weights minimizing the variance of the estimators i.e.,

$$, \min_{d_0 \, , \, d_1 \, , \ . \ . \ . \ , d_m} \sum_{i=1}^{m} \left(\sum_{j=0}^{m} d_j \, d_{i+j} \right)^2$$

Satisfying the conditions

$$\sum_{j=0}^{m} dj = 0 \, , \, \sum_{j=0}^{m} d_j^2 = 1. \tag{4}$$

The role of constraints (Eq. (3)) is now evident. The first condition ensures that, as the t's become close, the nonparametric effect is removed and the second one ensures that the variance of the sum of weighted residuals remains equals to σ^2in Eq. (2).
Now, we define the $(n-m) \times \, n$ differencing matrix Dwhose element satisfy Eq. (3) as

$$D = \begin{pmatrix} d_0 & d_1 & \dots & d_m & 0 & 0 & \dots & 0 \\ 0 & d_0 & d_1 & \dots & d_m & 0 & \dots & 0 \\ \vdots & \ddots & & & & & & \vdots \\ 0 & 0 & \dots & 0 & d_0 & d_1 & \dots & d_m \end{pmatrix}$$

This and related matrices are given, for example, in [4, 6, 7, 8, 9, 56, 57, 58, 59, 60].

Applying the differencing matrix to model (Eq. (2)) permits direct estimation of the parametric effect. As a result of development in [18, 61, 62, 63] it is know that the parameter vector β in (Eq. (1)) can be estimated with parametric efficiency. We now show the difference-based estimators that can be used for this purpose. Since the data have been ordered so that the values of the nonparametric variable(s) are close, the application of the differencing matrix Din model (Eq. (2)) removes the nonparamertric effect in large samples. If $f(.)$is an unknown function , i.e ., the inferential object and has a bound first derivative, then $D \, f(t)$ is close to 0 , so that by applying the differencing matrix , we may rewrite (Eq. (1)) as

$$Dy \, = \, DX\beta \, + \, D\varepsilon,$$

Or

$$y_D \, = \, X_D\beta + \varepsilon D, \tag{5}$$

where $y_D = D_y$, $X_D = D_X, \varepsilon_D = D_\varepsilon$. So, $\varepsilon_D$is a(n-m)-vector of disturbances distributed with $\mathrm{E}(\varepsilon_D) = 0$ and $\mathrm{E}(\varepsilon_D {\varepsilon_D}^{'}) = \sigma^2 V_D$ where $V_D = DVD' \neq I_{n-m}$. For arbitrary differencing coefficients satisfying Eq. (2) and (3), [19] defined a simple differencing estimator of the parameter β in a semiparametric regression model (SPRM) when $V = I_n$ as

$$\hat{\beta} D \, = (X_D' X_D)^{-1} X_D' y_D \, . \tag{6}$$

Thus, differencing allows one to perform inferences on β as if there were no nonparametric component $f(.)$in the model (Eq. (1)), [20, 64, 65, 66].Once β is estimated, a variety of nonparametric techniques could be applied to estimate $f(.)$as if β were known .

To account the parameter β in Eq. (4), we introduce the modified estimator of σ^2, defined as

$$\sigma^2 D = \frac{y'_D V_D^{-\frac{1}{2}}(I-P)V_D^{-\frac{1}{2}} yD}{tr(D'(I-P)D)}\,, \tag{7}$$

where $tr(.)$is the trace function for a squared matrix and $p$is the projection matrix defined as

$$p\;=\;V_D^{-\frac{1}{2}} X_D (X'_D V_D^{-1} X_D)^{-1} X'_D V_D^{-\frac{1}{2}}.$$

3. Liu Estimator

To overcome the effect of multicollinearity, Liu estimator is usually utilized. The Liu estimator for the semiparametric regression model (Liu) is defined as

$$\hat{\beta}_{Liu}\;=(X'_D X_D + I)^{-1}(X'_D X_D + dI)X'_D y_D\,, \tag{8}$$

where $0 < d < 1$is the shrinkage parameter. Several methods were proposed to estimate the value of d. The mean square error (MSE) of $\hat{\beta}_{Liu}$is defined as:

$$\text{MSE}\;(\hat{\beta}_{Liu}) = \sum_{j=1}^{p} \frac{(\lambda_j + d)^2}{\lambda_j(\lambda_j+1)^2} + (d-1)^2 \sum_{j=1}^{p} \frac{\alpha_j^2}{(\lambda_j+1)^2}\;, \tag{9}$$

where α_j is defined as the jth element of $\gamma\,\hat{\beta}_{Liu}$and γ is the eigenvector of the$X'_D X_D$.

To find the optimal value of d, the first derivative of Eq. (8) with respect to d is defined as

$$\frac{d\text{MSE}\;(\hat{\beta}_{Liu})}{dd} = 2\sum_{j=1}^{p} \frac{\lambda_j + d}{\lambda_j(\lambda_j+1)^2} + 2(d-1)\sum_{j=1}^{p} \frac{\alpha_j^2}{(\lambda_j+1)^2}. \tag{10}$$

By setting the resulting $d\text{MSE}\;(\hat{\beta}_{Liu})/dd$ to zero and solving for d, the optimal value is obtained as [10]:

$$d_{optimal} = \frac{\tau(\alpha^2-1)}{(1/\lambda)+\alpha^2}. \tag{11}$$

According to Eq. (10), when $\alpha_j^2 < 1$ the $d_{optimal}$ becomes negative and becomes positive when $\alpha_j^2 > 1$. To guarantee that $d_{optimal}$be between 0 and 1, the following methods have been proposed to estimate the $d_{optimal}$ [10, 11]:
$d_1 = \max\left(0, \frac{\hat{\tau}(\hat{\alpha}^2_{\max}-1)}{(1/\hat{\lambda}_{\max})+\hat{\alpha}^2_{\max}}\right)$, $d_2 = \max\left(0, median\left(\frac{\hat{\tau}(\hat{\alpha}_j^2-1)}{(1/\hat{\lambda}_j)+\hat{\alpha}_j^2}\right)\right)$, $d_3 = \max\left(0, \frac{\hat{\tau}}{p}\sum_{j=1}^{p} \frac{(\hat{\alpha}_j^2-1)}{(1/\hat{\lambda}_j)+\hat{\alpha}_j^2}\right)$, $d_4 = \max\left(0, \max\left(\frac{\hat{\tau}(\hat{\alpha}_j^2-1)}{(1/\hat{\lambda}_j)+\hat{\alpha}_j^2}\right)\right)$, and $d_5 = \max\left(0, \min\left(\frac{\hat{\tau}(\hat{\alpha}_j^2-1)}{(1/\hat{\lambda}_j)+\hat{\alpha}_j^2}\right)\right)$, where $\hat{\alpha}$ is defined as the jth element of $\gamma\,\hat{\beta}_{Liu}$and γ is the eigenvector of the $X'_D X_D$ matrix, $\hat{\alpha}_{\max}$is the maximum value of $\hat{\alpha}$.

4. The proposed approach

People have applied stochastic optimization techniques to solve various optimization problems. The reason these algorithms are trending these days is that they have shown impressive results when dealing with different optimization problems [12]. Lots of these algorithms take their inspiration from mechanisms in the natural world, including particle swarm optimization, genetic algorithm, and firefly algorithm. To optimize a study is to identify the best value for the study's parameters among the all possible values to help achieve the best results [13].

SMA, or the Slime Mould Algorithm [14], mimics the effective and irregular moves made by slime moulds in search of food and insight found in the way these organisms form networks.

For Liu estimator, we have one biasing parameter, d. This parameter is treated as a solute in the initial population. Consequently, our proposed algorithm is as:

Step 1: The number of the population size, NP, is set to 40 and the maximum number of iterations is $\mathrm{T}_{\max}= 500$.

Step 2: For d, the initial values are randomly generated from uniform distribution with 0 and 1.

Step 3: The fitness function is defined as

$$\text{fitness} = \min \left[\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_{i(k)})^2 \right] . \tag{12}$$

According to Eq. (11), each solute has its fitness value, and, therefore, the best solution is calculated.

Step 5: The movement of the solutions are updated according to SMA algorithm.

Step 6: Steps 4 and 5 are repeated until a $\mathrm{T}_{\max}$ is reached.

5. Simulation Results

A Mont Carlo simulation scheme to evaluate the performance of the estimating methods for the ridge estimator shrinkage parameter. The explanatory variables $\mathbf{x}_i^T = (x_{i1}, x_{i2}, ..., x_{in})$ have been generated from the following formula

$$x_{ij} = (1-\rho^2)^{1l2} w_{ij} + \rho w_{ip}, \quad i = 1, 2, ..., n, \quad j = 1, 2, ..., p, \tag{13}$$

where ρ represents the correlation between the explanatory variables and w_{ij}'s are independent standard normal pseudo-random numbers. Because the sample size has direct impact on the prediction accuracy, three representative values of the sample size are considered: 30, 50, 100, and 250. Further, because we are interested in the effect of multicollinearity, in which the degrees of correlation considered more important, three values of the pairwise correlation are considered with $\rho = \{0.90, 0.95, 0.99\}$. Then n observations for the dependent variable are determined by

$$y_i = \sum_{j=1}^{6} x_{ij}\beta_j + f(t_i) + \varepsilon_i \ , i = 1, \ . \ . \ . \ , n, \tag{14}$$

where $\beta = (3,\ 1,\ -3\ ,\ 2\ ,\ -5\ ,\ 4\ )'$ and $f(t) = \frac{1}{3}\ [\Phi(t;-3, 0.81)\ + \Phi(t; 0, 0.36)\ + \Phi(t; 3, 0.81)\ ]$, Which is mixture of normal densities for $t \in [-5\ , 5]$ $and\, \Phi(x; \mu, \sigma^2)$is a normal density function with mean μ and variance σ^2. The main reason of selecting such structure for nonlinear part is to check the efficiency of nonparametric estimation for wavy function. Moreover, $\varepsilon \ \sim \ N(0, \sigma^2 V)$for which the element of $V$are $v_{ij} = (\frac{1}{n})^{|i-j|}$.Four values of $\sigma^2 = 0.01$is investigated. We use a fourth–order differencing coefficients, $dd_o = 0.8873,\ dd_1 = 0.3099,\ dd_2 = -0.2426,\ dd_3 = -0.1910, and\, dd_4 = -0.1409$ in which $m = 4$.

The estimated MSE for all the different selection methods of d and the combination of n and ρ, are respectively summarized in Tables 1 – 4. Several observations can be obtained as follows:

1. MSE values increases for most methods as multicollinearity rises ($\rho = 0.90 \rightarrow \rho = 0.95 \rightarrow \rho = 0.99$), reflecting the growing challenge of parameter estimation in highly correlated settings.
2. SMA consistently outperforms all other methods, maintaining the lowest MSE across all ρ values. It dominates with MSEs of 1.211 ($\rho = 0.90$), 1.245 ($\rho = 0.95$), and 1.264 ($\rho = 0.99$), when n=30, highlighting its robustness to multicollinearity.
3. d1 and d4 methods show competitive MSEs at lower $\rho = 0.90$, but degrade at $\rho = 0.99$.
4. SPRM, d3, and d5 exhibit high MSEs, especially at $\rho = 0.99$.
5. In terms of ρ values, there is increasing in the MSE values when the correlation degree increases regardless the value of n.
6. With respect to the value of n, The MSE values decrease when n increases, regardless the value of ρ.

Table 1. MSE values when $n = 30$

Method			
	$\rho = 0.90$	$\rho = 0.95$	$\rho = 0.99$
SPRM	*1.792*	*3.153*	*3.802*
d1	*1.66*	*3.22*	*3.393*
d2	*1.711*	*3.313*	*3.46*
d3	*1.726*	*1.741*	*3.263*
d4	*1.438*	*2.638*	*3.317*
d5	*1.77*	*2.97*	*3.03*
SMA	*1.336*	*1.37*	*1.389*

Table 2. MSE values when $n = 50$

Method			
	$\rho = 0.90$	$\rho = 0.95$	$\rho = 0.99$
SPRM	*1.803*	*2.998*	*3.591*
d1	*1.601*	*3.153*	*3.14*
d2	*1.652*	*3.148*	*3.197*
d3	*1.667*	*1.574*	*3.021*
d4	*1.379*	*2.471*	*3.054*
d5	*1.711*	*2.803*	*2.767*
SMA	*1.324*	*1.336*	*1.351*

Table 3. MSE values when $n = 100$

Method			
	$\rho = 0.90$	$\rho = 0.95$	$\rho = 0.99$
SPRM	*3.774*	*4.861*	*6.378*
d1	*1.552*	*2.857*	*3.067*
d2	*1.593*	*1.981*	*2.934*
d3	*1.608*	*1.407*	*2.737*
d4	*1.32*	*2.304*	*2.791*
d5	*1.652*	*2.636*	*2.504*
SMA	*1.307*	*1.331*	*1.348*

Table 4. MSE values when $n = 250$

Method			
	$\rho = 0.90$	$\rho = 0.95$	$\rho = 0.99$
SPRM	*3.895*	*4.982*	*6.499*
d1	*1.673*	*2.978*	*3.188*
d2	*1.714*	*2.102*	*3.056*
d3	*1.729*	*1.528*	*2.858*
d4	*1.441*	*2.429*	*2.912*
d5	*1.773*	*2.757*	*2.627*
SMA	*1.428*	*1.452*	*1.469*

6. Real data application

Investments in companies and sectors are typically expressed in relation to assets, which may be classified as either current (short-term) or fixed (long-term). This categorization also applies to total capital formation, which is the sum of all investment capital that flows into an economy over a specific period. As each investment expenditure contributes towards accumulating a balance of flows over the years, the status of these investments is subject to change due to various internal and external variables.

One such variable is the periodic and continuous increase in the prices of basic commodities, which can weaken the purchasing power of investors and erode the value of their accumulated balances over time. Additionally, fluctuations in local currency exchange rates, denominated in a specific currency or group of currencies, can impact the value of investments held by foreign investors. Another variable that affects capital formation is the balance of payments, which reflects the economic and financial activities of residents and non-residents within a country. It is primarily comprised of two components: the current account and the capital account. The first measures the value of imported, exported, and consumed goods and services, while the second tracks the flow of financial (capital) assets, including both portfolio and direct investment, such as foreign direct investment. These figures can be absolute numbers or attributed to the gross domestic product (GDP), which is the sum of the monetary values of goods and services produced during a specific year in a particular country by residents and foreign residents within its geographical boundaries.

Numerous variables can influence the increase or decrease in GDP, and consequently, total capital formation. Hence, this thesis aims to study the effect of two crucial economic variables affecting capital formation: the current account balance and foreign direct investment, as explanatory variables that follow a parameter function with unknown parameters. In addition, the thesis will examine the impact of inflation as a variable that follows a non-parametric function on total capital formation.

The data for this study was obtained from the World Bank's official website. Specifically, from the Global Development Indicators dataset, which had been updated as of July 1st, 2021. The dataset included information related to several countries around the world, with a total of 78 countries being considered for the year 2019. The data focused on identifying the most significant and prominent factors that affect the ratio of total capital formation to gross domestic product. In defining the terms used in this study's variables, the following descriptions are based on information sourced from the official website of the World Bank (2021):

1. Gross capital formation: This represents the total expenditure on increasing the fixed assets of an economy and includes net changes in the level of stocks. It encompasses fixed assets such as machinery, equipment purchases, construction of roads, railways, offices, hospitals, private residences, commercial and industrial buildings, inventories, and other similar investments.
2. Current account balance: This refers to the sum of net exports of goods and services, net income, and net current transfers of a country's economy.
3. Foreign direct investment: This refers to the net inflow of investments aimed at obtaining a permanent share in the management of an enterprise operating in an economy other than that of the investor.
4. Inflation: This reflects the annual percentage change in the cost to consumers for acquiring a basket of goods and services measured by the consumer price index. This basket of goods can remain constant or change over specific periods of time, such as yearly.

Before estimating the model, the first step involves analyzing the relationship between the independent explanatory variables and the dependent response variable graphically. This analysis is necessary to determine whether the relationship between the variables is linear or non-linear. Based on this assessment, the variables for each of the two components of the model are identified. In this study, a statistical program R, was used to create a scatter plot, which is shown in Figure 1. Upon examining the scatter plot, it was determined that the relationship between each of the two explanatory variables (current account balance X1, foreign direct investment X2) and the dependent response variable (total capital formation Y) is linear. On the other hand, the relationship between the dependent response variable (Y) and the third explanatory variable (inflation x3) is non-linear.

To detect linear multicollinearity among the explanatory variables affecting total capital accumulation, a simple correlation matrix (c) was calculated. The correlation coefficients are: rx1x2=0.903, rx1x3=-1.779, and rx2x3=-1.473. By observing the correlation, we note that there are varying correlations between all explanatory variables, and that the highest level of these correlations was between variables (X2, X1), which means that there is a possibility of linear multicollinearity in the data.

In Table 5, the MSE values of the proposed difference-based estimators are reported for m=7,m=5, and 3. For all considered values of mm, the difference-based estimators d1–d5 exhibit smaller MSEs than the baseline SPRM estimator, indicating improved estimation accuracy. Moreover, there is a clear decreasing pattern in MSE as we move from d1 to d5, with d5 consistently outperforming the other difference-based estimators. The SMA estimator achieves the lowest MSE in all settings, suggesting that it provides the most efficient estimates among the competing methods under the studied conditions.

Figures 2-4 report the percentage reduction in MSE achieved by the competing methods relative to the SMA method when m=7, m=5, and m=3, where larger values indicate greater improvement. As shown in Figure 2, for

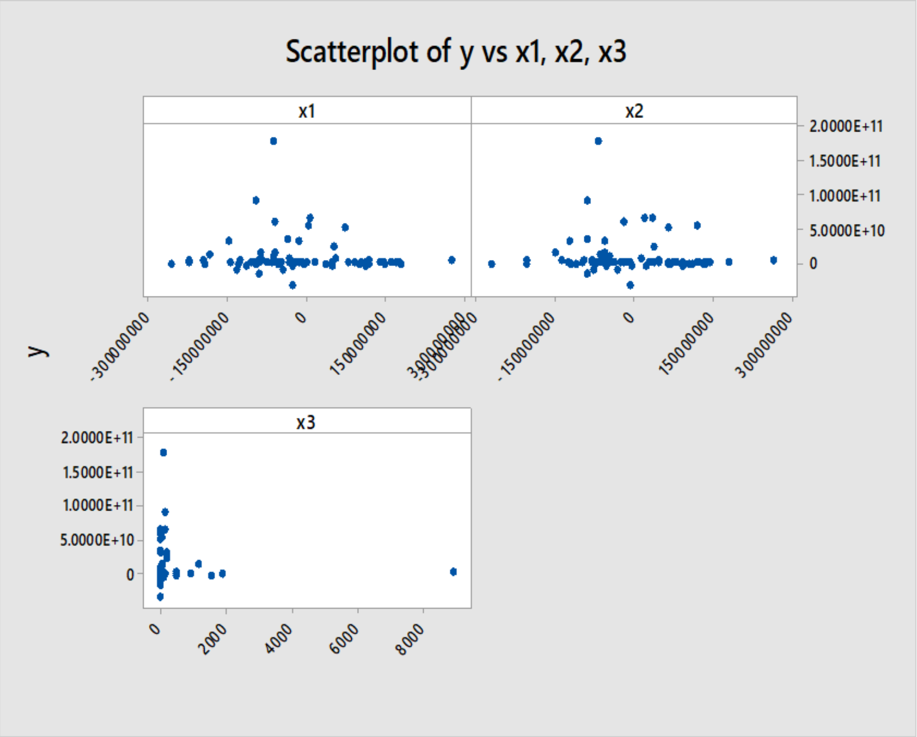

Figure 1. the nature of the relationship between the response variable and the explanatory variables of the study data.

example, SPRM attains the largest reduction (53.70%), indicating the strongest overall error decrease among all approaches considered in this experiment. The remaining approaches (d1–d5) also yield positive MSE reductions, but with a clear monotone decline as we move from d1 to d5. Specifically, the reductions are 47.54% (d1), 43.73% (d2), 37.04% (d3), 29.97% (d4), and 26.53% (d5), demonstrating progressively smaller gains over the baseline.

Overall, These figures highlights that SPRM provides the most substantial improvement, while the d-method family exhibits diminishing returns across d1–d5 in terms of MSE reduction.

Table 5. MSE valuesfor the used methods

m=7	m=5	m=3	Methods of The Difference based Estimators
8.7014	8.0134	7.7548	SPRM
7.6794	7.851	6.4589	d1
7.1593	7.4098	6.0975	d2
6.3986	6.057	5.0631	d3
5.7524	5.611	4.7349	d4
5.4831	5.405	4.6124	d5
4.0284	3.886	3.5621	SMA

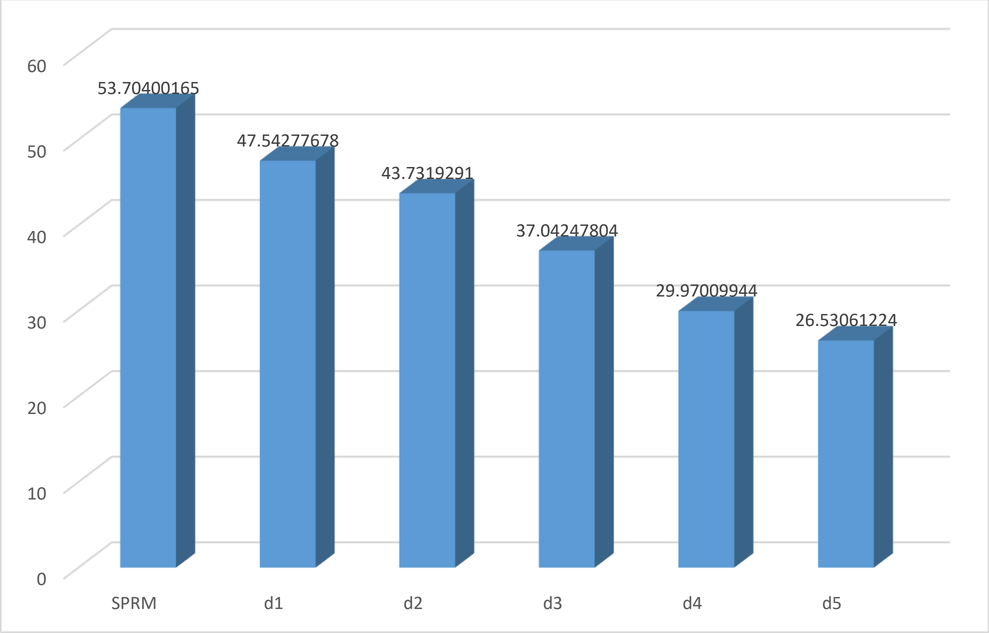

Figure 2. Reduction in MSE using SMA when m=7

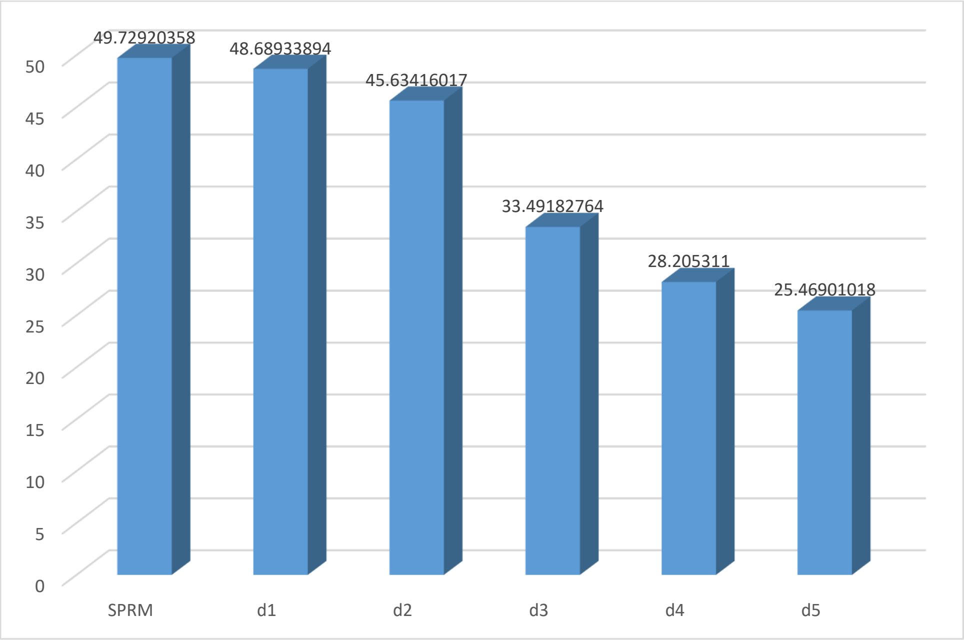

Figure 3. Reduction in MSE using SMA when m=5

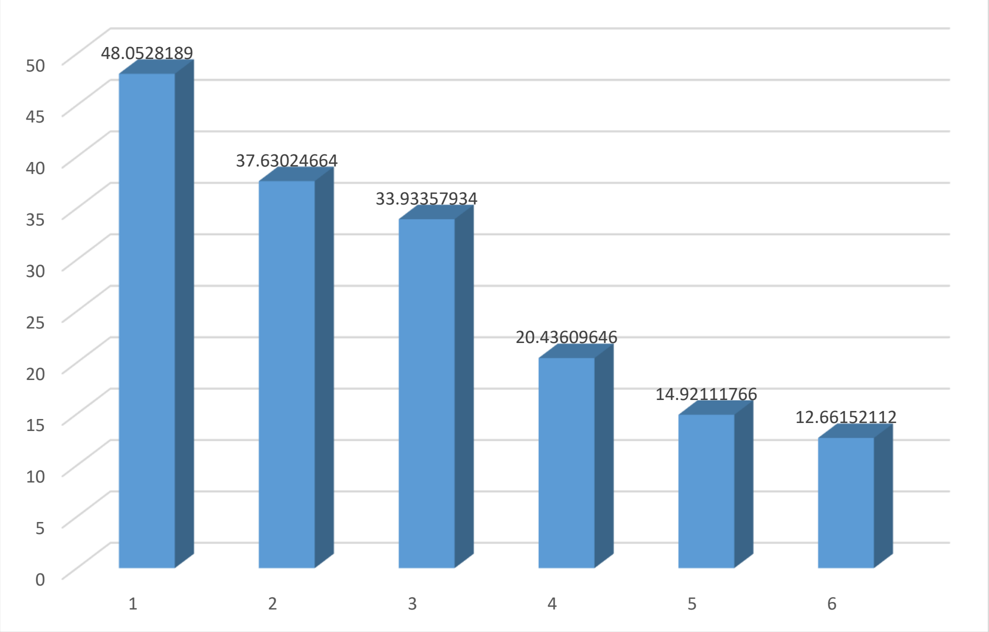

Figure 4. Reduction in MSE using SMA when m=3

7. Conclusion

In this paper, methods of selecting the Liu estimator shrinkage are discussed and outlined. The SMA is a viable way of estimating the Liu regression shrinkage. Due to its bio-inspired searching, SMA can process a significant amount of data points quickly, as well as, it solves the uncertainty and inconsistency posed by multicollinearity. Overall, this method is believed to be a methodological approach towards making decisions more efficiently in terms of the needed shrinkage. Under Monte Carlo simulations, we are able to know that some estimator is much better than the other based on MSE. The results in all the cases evidenced that the SMA performed better than the SPRM approach in Liu regression. SMA has its sensitivity to algorithmic settings like population size, maximum iterations, and the initialization range of the shrinkage parameter

REFERENCES

1. F. Akdeniz, E. Akdeniz, G. Kaçiranlar, and H. Hu, *Efficiency of the generalized difference-based Liu estimators in semiparametric regression models with correlated errors*, Journal of Statistical Computation and Simulation, vol. 85, no. 1, pp. 147–165, 2013.
2. F. Akdeniz and E. A. Duran, *Liu-type estimator in semiparametric regression models*, Journal of Statistical Computation and Simulation, vol. 80, no. 8, pp. 853–871, 2010.
3. H. Emami, *Local influence for Liu estimators in semiparametric linear models*, Statistical Papers, vol. 59, no. 2, pp. 529–544, 2016.
4. A. Yatchew, *An elementary estimator of the partial linear model*, Economics Letters, vol. 57, no. 2, pp. 135–143, 1997.
5. C. Wei and X. Wang, *Liu-type estimator in semiparametric partially linear additive models*, Journal of Nonparametric Statistics, vol. 28, no. 3, pp. 459–468, 2016.
6. M. Roozbeh, *Shrinkage ridge estimators in semiparametric regression models*, Journal of Multivariate Analysis, vol. 136, pp. 56–74, 2015.
7. M. Roozbeh and M. Arashi, *New Ridge Regression Estimator in Semiparametric Regression Models*, Communications in Statistics - Simulation and Computation, vol. 45, no. 10, pp. 3683–3715, 2015.
8. M. Roozbeh, M. Arashi, and H. A. Niroumand, *Semiparametric Ridge Regression Approach in Partially Linear Models*, Communications in Statistics - Simulation and Computation, vol. 39, no. 3, pp. 449–460, 2010.
9. A. Ullah, X. Wang, H. H. Zhang, and M. Roy, *A semiparametric generalized ridge estimator and link with model averaging*, Econometric Reviews, vol. 36, no. 1-3, pp. 370–384, 2015.
10. K. Månsson, *Developing a Liu estimator for the negative binomial regression model: method and application*, Journal of Statistical Computation and Simulation, vol. 83, no. 9, pp. 1773–1780, 2013.

11. K. Mansson, B. M. G. Kibria, and G. Shukur, *Improved Liu estimators for the Poisson regression model*, International Journal of Statistics and Probability, vol. 1, no. 1, pp. 1–6, 2012.
12. S. Mirjalili, *SCA: A Sine Cosine Algorithm for solving optimization problems*, Knowledge-Based Systems, vol. 96, pp. 120–133, 2016.
13. S. Li, H. Fang, and X. Liu, *Parameter optimization of support vector regression based on sine cosine algorithm*, Expert Systems with Applications, vol. 91, pp. 63–77, 2018.
14. A. A. Ewees, M. A. Elaziz, O. A. Alanezi, and E. H. Houssein, *Enhanced feature selection technique using slime mould algorithm: a case study on chemical data*, Neural Computing and Applications, vol. 35, no. 4, pp. 3307–3324, 2023.
15. Hoerl, Arthur E and Kennard, Robert W, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics, vol. 12, no. 1, pp. 55–67, 1970.
16. Levine, Michael, *Robust functional estimation in the multivariate partial linear model*, Annals of the Institute of Statistical Mathematics, vol. 71, no. 4, pp.743–770, 2017.
17. Roozbeh, Mahdi and Arashi, Mohammad, *New ridge regression estimator in semiparametric regression models*, Communications in Statistics-Simulation and Computation, vol. 45, no. 10, pp.3683–3715, 2016.
18. Speckman, Paul, *Kernel smoothing in partial linear models*, Journal of the Royal Statistical Society Series B: Statistical Methodology, vol. 50, no. 3, pp.413–436, 1988.
19. Yatchew, Adonis, *An elementary estimator of the partial linear model*, Economics letters, vol. 57, no. 2, pp.135–143, 1997.
20. Yatchew, Adonis and Sun, Yiguo and Deri, Catherine, *Efficient estimation of semiparametric equivalence scales with evidence from South Africa*, Journal of Business & Economic Statistics, vol. 21, no. 2, pp.247–257, 2003.
21. Algamal, Z. Y., & Asar, Y. *Liu-type estimator for the gamma regression model* , Communications in Statistics-Simulation and Computation, vol. 49, no . 8, p. 2035-2048, 2020.
22. Algamal, Z. Y., & Lee, M. H. *A new adaptive L1-norm for optimal descriptor selection of high-dimensional QSAR classification model for anti-hepatitis C virus activity of thiourea derivatives* , SAR and QSAR in Environmental Research, vol. 28, no. 1 , p. 75-90, 2017.
23. Kahya, M. A., Altamir, S. A., & Algamal, Z. Y. *Improving whale optimization algorithm for feature selection with a time-varying transfer function* , Numerical Algebra, Control and Optimization, vol. 11, no. 1 , p. 87-98 , 2020.
24. Algamal, Z. Y., Qasim, M. K., & Ali, H. T. M. *A QSAR classification model for neuraminidase inhibitors of influenza A viruses (H1N1) based on weighted penalized support vector machine* , SAR and QSAR in Environmental Research, vol. 28 , no. 5, p. 415-426 , 2017.
25. Algamal, Z. Y., Lee, M. H., & Al-Fakih, A. M. *High-dimensional quantitative structure–activity relationship modeling of influenza neuraminidase a/PR/8/34 (H1N1) inhibitors based on a two-stage adaptive penalized rank regression* , Journal of Chemometrics, vol. 30 , no.2 m p. 50-57 , 2016.
26. Algamal, Z. Y., Lee, M. H., Al-Fakih, A. M., & Aziz, M. *High-dimensional QSAR classification model for anti-hepatitis C virus activity of thiourea derivatives based on the sparse logistic regression model with a bridge penalty* , Journal of Chemometrics, vol. 31 , p.6, p. e2889 , 2017.
27. Algamal, Z. Y., Qasim, M. K., Lee, M. H., & Ali, H. T. M. *High-dimensional QSAR/QSPR classification modeling based on improving pigeon optimization algorithm* , Chemometrics and Intelligent Laboratory Systems, vol. 206, p. 104170 , 2020.
28. Ismael, O. M., Qasim, O. S., & Algamal, Z. Y. *Improving Harris hawks optimization algorithm for hyperparameters estimation and feature selection in v-support vector regression based on opposition-based learning* , Journal of Chemometrics, vol 34 , no. 11, e3311, 2020.
29. Abonazel, M. R., Algamal, Z. Y., Awwad, F. A., & Taha, I. M. *A new two-parameter estimator for beta regression model: method, simulation, and application* , Frontiers in Applied Mathematics and Statistics, vol. 7,p. 780322 ,2022.
30. Algamal, Z. Y., & Abonazel, M. R. *Developing a Liu-type estimator in beta regression model* , Concurrency and Computation: Practice and Experience, vol.34 ,no. 5 , p. e6685.
31. Algamal, Z., & Ali, H. M. *An efficient gene selection method for high-dimensional microarray data based on sparse logistic regression* , Electronic Journal of Applied Statistical Analysis, vol . 10 , no . 1 , 242-256 , 2017.
32. Salih, A. M., Algamal, Z., & Khaleel, M. A. *A new ridge-type estimator for the gamma regression model*, Iraqi Journal for Computer Science and Mathematics, vol . 5 , no .1 , p. 85-98, 2023.
33. Alharthi, A. M., Kadir, D. H., Al-Fakih, A. M., Algamal, Z. Y., Al-Thanoon, N. A., & Qasim, M. K. *Quantitative structure-property relationship modelling for predicting retention indices of essential oils based on an improved horse herd optimization algorithm*, SAR and QSAR in Environmental Research, vol . 34 , no .10 , p. 831-846, 2023.
34. Mahmood, S. W., Basheer, G. T., & Algamal, Z. Y. *Quantitative Structure–Activity Relationship Modeling Based on Improving Kernel Ridge Regression* , Journal of Chemometrics, vol. 39, no. 5, p. e70027, 2025.
35. Mahmood, S. W., Basheer, G. T., & Algamal, Z. Y. *Improving kernel ridge regression for medical data classification based on meta-heuristic algorithms* , Kuwait Journal of Science, vol. 52, no. 3 , p. 100408 , 2025
36. Algamal, Z. Y., Alhamzawi, R., & Ali, H. T. M. *Gene selection for microarray gene expression classification using Bayesian Lasso quantile regression*, Computers in biology and medicine, vol. 97, p. 145-152, 2018
37. Algamal, Z. Y., & Lee, M. H. *A novel molecular descriptor selection method in QSAR classification model based on weighted penalized logistic regression*, Journal of Chemometrics, vol. 31, no.(10), e2915, 2017
38. Qasim, M. K., Algamal, Z. Y., & Ali, H. M. *A binary QSAR model for classifying neuraminidase inhibitors of influenza A viruses (H1N1) using the combined minimum redundancy maximum relevancy criterion with the sparse support vector machine*, SAR and QSAR in Environmental Research, vol. 29, no.(7), p.517-527, 2018
39. Algamal, Z. Y., & Lee, M. H. *Applying penalized binary logistic regression with correlation based elastic net for variables selection*, Journal of Modern Applied Statistical Methods, vol. 14, no.(1), p.15, 2015
40. Algamal, Z. Y., Lee, M. H., Al-Fakih, A. M., & Aziz, M. *High-dimensional QSAR modelling using penalized linear regression model with L 1/2-norm*, SAR and QSAR in Environmental Research, vol. 27, no.(9), p.703-719, 2016

41. Al-Taweel, Y., & Algamal, Z. *Some almost unbiased ridge regression estimators for the zero-inflated negative binomial regression model*, Periodicals of Engineering and Natural Sciences, vol. 8, no.(1), p.248-255, 2020
42. Ewees, A. A., Algamal, Z. Y., Abualigah, L., Al-Qaness, M. A., Yousri, D., Ghoniem, R. M., & Abd Elaziz, M. *A cox proportional-hazards model based on an improved aquila optimizer with whale optimization algorithm operators*, Mathematics, vol. 10, no.(8), p.1273, 2022
43. Shamany, R., Alobaidi, N. N., & Algamal, Z. Y. *A new two-parameter estimator for the inverse Gaussian regression model with application in chemometrics*, Electronic Journal of Applied Statistical Analysis, vol. 12, no.(2), p.453-464, 2019
44. Awwad, F. A., Odeniyi, K. A., Dawoud, I., Algamal, Z. Y., Abonazel, M. R., Kibria, B. G., & Eldin, E. T. *New two-parameter estimators for the logistic regression model with multicollinearity*, WSEAS Trans. Math, vol. 21, p.403-414, 2022
45. Almishlih1, Zaynab Ayham, Qasim, Omar Saber, Algamal, Zakariya Yahya *Binary Arithmetic Optimization Algorithm Using a New Transfer Function for Fusion Modeling*, Fusion: Practice and Applications, vol. 18, no.2, p. 157-168, 2025
46. Kahya, M. A., Altamir, S. A., & Algamal, Z. Y. *Improving whale optimization algorithm for feature selection with a time-varying transfer function* , Numerical Algebra, Control and Optimization, vol. 11, no. 1 , p. 87-98 , 2020.
47. Algamal, Z. Y., Qasim, M. K., Lee, M. H., & Ali, H. T. M. *QSAR model for predicting neuraminidase inhibitors of influenza A viruses (H1N1) based on adaptive grasshopper optimization algorithm*, SAR and QSAR in Environmental Research, vol.31, no.11, p.803-814, 2020.
48. Al-Fakih, A. M., Algamal, Z. Y., Lee, M. H., Aziz, M.,& Ali, H. T. M. *QSAR classification model for diverse series of antifungal agents based on improved binary differential search algorithm*, SAR and QSAR in Environmental Research, vol.30, no.2, p.131-143, 2019.
49. Alharthi, A. M., Kadir, D. H., Al-Fakih, A. M., Algamal, Z. Y., Al-Thanoon, N. A., Qasim, M. K. *Improving golden jackel optimization algorithm: An application of chemical data classification*, Chemometrics and Intelligent Laboratory Systems, vol.250, 105149, 2024.
50. Ewees, A. A., Al-Qaness, M. A., Abualigah, L., Algamal, Z. Y., Oliva, D., Yousri, D., & Elaziz, M. A. *Enhanced feature selection technique using slime mould algorithm: A case study on chemical data* , Neural Computing and Applications, vol. 35, 3307-3324, 2023.
51. Qasim, O. S., & Algamal, Z. Y. *A gray wolf algorithm for feature and parameter selection of support vector classification* , International Journal of Computing Science and Mathematics, vol. 13, 93-102, 2021.
52. Algamal, Z. Y. *Variable selection in count data regression model based on firefly algorithm* , Statistics, Optimization & Information Computing, vol. 7, 520-529, 2019.
53. Kahya, M. A., Altamir, S. A., & Algamal, Z. Y. *Improving firefly algorithm-based logistic regression for feature selection*, Journal of Interdisciplinary Mathematics, vol. 22, 1577-1581, 2019.
54. Yousif, H. M., & Algamal, Z. Y. *Bandwidth Selection in Geographically Weighted Poisson Regression Model Using Firefly Optimization Algorithm with Application to Cancer Rate Data*, European Journal of Statistics, vol. 5, 10, 2025.
55. AL-Taie, F. A. Y., Qasim, O. S., & Algamal, Z. Y. *Improving kernel semi-parametric regression model based on a bat optimization algorithm*, AIP Conference Proceedings, vol. 3036, p. 040003, 2024.
56. Andu, Y., Lee, M. H., & Algamal, Z. Y. *Generalized dynamic principal component for monthly nonstationary stock market price in technology sector*, Journal of Physics: Conference Series, vol. 1132, p. 012076, 2018.
57. Andu, Y., Lee, M. H., & Algamal, Z. Y. *Generalized dynamic principal component for monthly nonstationary stock market price in technology sector*, Journal of Physics: Conference Series, vol. 1132, p. 012076, 2018.
58. Almishlih, Z. A., Qasim, O. S., & Algamal, Z. Y. *Design and evaluation of a new tent-shaped transfer function using the Polar Lights Optimizer algorithm for feature selection*, Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska, vol. 15, p. 27-31, 2025.
59. Al-Thanoon, N. A., Qasim, O. S., & Algamal, Z. Y. *Improving the binary tree growth algorithm with fuzzy mutual information for feature selection*, AIP Conference Proceedings, vol. 3318, p. 030008, 2025.
60. Al-Fakih, A. M., Qasim, M. K., Algamal, Z. Y., Alharthi, A. M., & Zainal-Abidin, M. H. *QSAR classification model for diverse series of antifungal agents based on binary coyote optimization algorithm* SAR and QSAR in Environmental Research, vol.34, no.4, p. 285-298, 2023.
61. Al-Fakih, A. M., Algamal, Z. Y., & Qasim, M. K. *An improved opposition-based crow search algorithm for biodegradable material classification* SAR and QSAR in Environmental Research, vol.33, no.5, p. 403-415, 2022.
62. Hawa, N. S., Mustafa, M. Y., Kibria, B. G., & Algamal, Z. Y. *Bootstrap Liu-type estimator for Conway-Maxwell-Poisson regression model* Communications in Statistics-Simulation and Computation, p. 1-12, 2025.
63. Naziyah, A. A., & Algamal, Z. Y. *Jackknifed Liu-type estimator in Poisson regression model* Journal of the Iranian Statistical Society, vol.19, no.1, p. 21-37, 2020.
64. Al-Taweel, Y., & Algamal, Z. Y. *Some almost unbiased ridge regression estimators for the zero-inflated negative binomial regression model* Periodicals of Engineering and Natural Sciences, vol.8, no.1, p. 248-255, 2020.
65. Mahmood, S., & Algamal, Z. Y. *Kernel ridge regression improving based on golden eagle optimization algorithm for multi-class classification* Statistics, Optimization & Information Computing, vol.15, no.1, p. 354–371, 2026.
66. Al-Fakih, A. M., Qasim, M. K., Algamal, Z. Y., Alharthi, A. M., & Zainal-Abidin, M. H. *QSAR classification model for diverse series of antifungal agents based on binary coyote optimization algorithm* Statistics, SAR and QSAR in Environmental Research, vol.34, no.4, p. 285-298, 2023.