


Stochastic Restricted Modified Mixed Logistic Estimator

T. Kayathiri 

Postgraduate Institute of Science,
University of Peradeniya,
Peradeniya, Sri Lanka

M. Kayanan 

Department of Physical Science,
University of Vavuniya,
Vavuniya, Sri Lanka.

P. Wijekoon 

Department of Statistics and Computer Science,
University of Peradeniya,
Peradeniya, Sri Lanka

Abstract

In this study, we introduce a new estimator named the Stochastic Restricted Modified Mixed Logistic Estimator (SRMMLE), which is specifically designed to handle multicollinearity within the framework of stochastic linear restrictions. Further, we enhance the SRMMLE by modifying its coefficients, resulting in four distinct variants: Stochastic Restricted Modified Mixed Logistic Estimator 1 (SRMMLE1), Stochastic Restricted Modified Mixed Logistic Estimator 2 (SRMMLE2), Stochastic Restricted Modified Mixed Logistic Estimator 3 (SRMMLE3), and Stochastic Restricted Modified Mixed Logistic Estimator 4 (SRMMLE4). Based on the mean square error matrix criterion, we establish conditions for the superiority of SRMMLE over existing estimators, such as the Stochastic Restricted Maximum Likelihood Estimator (SRMLE), Stochastic Restricted Ridge Maximum Likelihood Estimator (SRRMLE), Stochastic Restricted Logistic Liu Estimator (SRLLE), and Stochastic Restricted Mixed Liu-Type Estimator (SRMLTE). In the simulation study, we determined the scalar mean square error and the K-fold cross-validated balanced accuracy of the estimators. Further, we present an empirical study and a real data application illustrating the superior performance of the proposed estimator. In particular, the SRMMLE4 outperforms others in terms of scalar mean square error and balanced accuracy.

Keywords: balanced accuracy, logistic regression, multicollinearity, predictive performance, R.

1. Introduction

A statistical modeling technique called logistic regression is used to predict the probability of a binary outcome based on one or more predictor variables. It is commonly used in various fields, including economics, healthcare, and social sciences. Let us consider a general form of

the logistic regression model

$$y_i = \pi_i + \epsilon_i, \quad i = 1, 2, \dots, n. \quad (1)$$

which follows the Bernoulli distribution with parameter π_i and is given by,

$$\pi_i = \frac{\exp(x_i' \beta)}{1 + \exp(x_i' \beta)} \quad (2)$$

where, x_i is the i^{th} row of X , which is an $n \times (p + 1)$ data matrix with p predictor variables and β is a $(p + 1) \times 1$ vector of coefficients, ϵ_i is independent with mean zero and variance $\pi_i(1 - \pi_i)$ of the response y_i .

The maximum likelihood estimation technique is a commonly used method to estimate the parameter (β), and the Maximum Likelihood Estimator (MLE) of β , which is given by:

$$\hat{\beta}_{MLE} = C^{-1} X' \hat{W} Z. \quad (3)$$

where, $C = X' \hat{W} X$; Z is the column vector with i^{th} element equals $\text{logit}(\hat{\pi}_i) + \frac{y_i - \hat{\pi}_i}{\hat{\pi}_i(1 - \hat{\pi}_i)}$

and $\hat{W} = \text{diag}[\hat{\pi}_i(1 - \hat{\pi}_i)]$, which is an unbiased estimate of β .

The covariance matrix of $\hat{\beta}_{MLE}$ is

$$\text{Cov}(\hat{\beta}_{MLE}) = (X' \hat{W} X)^{-1} = C^{-1} \quad (4)$$

When multicollinearity occurs, it can lead to unstable and unreliable estimates of the regression coefficients, making it difficult to interpret the relationship between the predictor variables and the outcome variable accurately. To reduce this issue several alternative estimators have been proposed in literature. In this study, we considered estimators based on the sample information and stochastic linear restrictions. The stochastic restricted methods are a flexible and effective approach to handle multicollinearity in logistic regression by incorporating prior information in a probabilistic manner. This results in more stable and reliable parameter estimates, which improves the overall performance of the model (Nagarajah and Wijekoon (2015), Li, Asar, and Wu (2020), Varathan and Wijekoon (2021)). Arashi, Kibria, and Valizadeh (2017) considered several stochastic restricted estimators in linear regression, while Nagarajah and Wijekoon (2015) introduced the Stochastic Restricted Maximum Likelihood Estimator (SRMLE) for logistic regression. Later, the same authors proposed, the Stochastic Restricted Ridge Maximum Likelihood Estimator (SRRMLE) Varathan and Wijekoon (2016), Stochastic Restricted Ridge Likelihood Estimator (SRRLE) Varathan and Wijekoon (2019a), Stochastic Restricted Liu Maximum Likelihood Estimator (SRLMLE) Varathan and Wijekoon (2019b), Stochastic Restricted Liu-Type Logistic Estimator (SRLTLE) Varathan and Wijekoon (2018), Stochastic Restricted Almost Unbiased Ridge Logistic Estimator (SRAURLE) Varathan and Wijekoon (2017b), Stochastic Restricted Almost Unbiased Liu Logistic Estimator (SRAULLE) Varathan and Wijekoon (2017a). Moreover, the Stochastic Restricted Logistic Liu Estimator (SRLLE), and the Stochastic Restricted Mixed Liu-Type estimator (SRMLTE) have been proposed by Li *et al.* (2020), and Yehia (2020), respectively, in the literature.

Further, Varathan and Wijekoon (2021) have introduced a stochastic restricted optimal logistic estimator (SROLE) for the logistic regression based on sample information and the prior information in the form of stochastic linear restrictions. Their study, the performance of the SROLE was compared with some existing logistic estimators such as SRMLE, SRRLE, SRLMLE, SRAULLE, SRAURLE, and SRLTLE in the scalar mean square error (SMSE) criterion.

The aforementioned literature primarily compares the performance of estimators in terms of SMSE. However, Logistic regression outputs probabilities, which are subsequently used

to classify instances into categories. To evaluate how well these probabilities translate into correct classifications, it is essential to use classification-specific metrics such as balanced accuracy, F1 score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). High values across these metrics collectively suggest robust model performance, with balanced accuracy emphasizing correct classifications, the F1 score focusing on precision and recall, and the AUC-ROC curve reflecting overall discriminative power. These metrics provide a meaningful assessment of the classification performance of the model. In contrast, the SMSE evaluates the average squared difference between predicted and actual values, making it appropriate for continuous outputs. Applying SMSE to logistic regression can be misleading because it does not account for the categorical nature of the predictions. Consequently, a model might exhibit a low SMSE while performing poorly in terms of classification accuracy or other relevant classification metrics. Thus, it is crucial to use appropriate classification metrics to accurately assess the performance of logistic regression models.

In this study, we propose a novel estimator, the Stochastic Restricted Modified Mixed Logistic Estimator (SRMMLE) for logistic regression. We evaluate the performance of the proposed estimator with existing estimators: SRMLE, SRRMLE, SRLLE, SRMLTE, and SROLE in terms of SMSE and classification metrics.

The paper is organized as follows: Section 2 presents the existing estimators, the construction of the new estimators, and their properties. Section 3 provides a theoretical and numerical discussion of the conditions under which the proposed estimators are superior to the existing estimators. Section 4 discusses the real data application and validates the theoretical conditions. Some concluding remarks are given in Section 5. Finally, the references are presented at the end of the paper.

2. The existing estimators and asymptotic properties

Suppose that the linear stochastic restriction is available in addition to the logistic regression model (1) of the form

$$h = H\beta + v; \quad E(v) = 0, \quad Cov(v) = \psi \quad (5)$$

where h is an $(q \times 1)$ stochastic known vector, H is a $(q \times (p+1))$ of full rank $q \leq (p+1)$ known elements and v is an $(q \times 1)$ random vector of errors with mean 0 and dispersion matrix ψ , and ψ is assumed to be known $(q \times q)$ positive definite matrix. Further, it is assumed that v is stochastically independent of ϵ , i.e. $E(\epsilon v') = 0$.

2.1. The existing estimators based on sample information

In this section, we considered some biased estimators based on sample information, including the Logistic Ridge Estimator (LRE) (Schaefer, Roi, and Wolfe (1984)), Logistic Liu Estimator (LLE) (Månsson, Kibria, and Shukur (2012)), and Liu-type logistic Estimator (LTLE) (Inan and Erdogan (2013)). The definitions of these estimators are shown respectively as follows:

$$\begin{aligned} \hat{\beta}_{LRE} &= Z_k \hat{\beta}_{MLE}, & \text{where } Z_k &= (C + kI)^{-1}C, \quad k > 0. \\ \hat{\beta}_{LLE} &= Z_d \hat{\beta}_{MLE}, & \text{where } Z_d &= (C + I)^{-1}(C + dI), \quad 0 < d < 1. \\ \hat{\beta}_{LTLE} &= Z_{k,d} \hat{\beta}_{MLE}, & \text{where } Z_{k,d} &= (C + kI)^{-1}(C - dI), \quad 0 < d < 1, \quad k > 0. \end{aligned} \quad (6)$$

Note that the three alternative estimators that we have given above are a function of $\hat{\beta}_{MLE}$,

and we can present them in general form as,

$$\hat{\beta}_{GLE} = L_{(i)} \hat{\beta}_{MLE}, \quad (7)$$

where $L_{(i)}$ is a positive definite matrix.

$$\hat{\beta}_{GLE} = \begin{cases} \hat{\beta}_{MLE} & \text{if } L_{(i)} = I; \\ \hat{\beta}_{LRE} & \text{if } L_{(i)} = Z_k; \\ \hat{\beta}_{LLE} & \text{if } L_{(i)} = Z_d; \\ \hat{\beta}_{LTLE} & \text{if } L_{(i)} = Z_{k,d}. \end{cases} \quad (8)$$

2.2. The existing stochastic restricted mixed estimators

Nagarajah and Wijekoon (2015), introduced the Stochastic Restricted Maximum Likelihood Estimator (SRMLE)

$$\hat{\beta}_{SRMLE} = \hat{\beta}_{MLE} + C^{-1}H'(\psi + HC^{-1}H')^{-1}(h - H\hat{\beta}_{MLE}) \quad (9)$$

By substituting $\hat{\beta}_{MLE}$, we obtain

$$\begin{aligned} \hat{\beta}_{SRMLE} &= C^{-1}X'\hat{W}z + C^{-1}H'(\psi + HC^{-1}H')^{-1}(h - HC^{-1}X'\hat{W}z) \\ &= (C^{-1} - C^{-1}H'(\psi + HC^{-1}H')^{-1}HC^{-1})(X'\hat{W}z + H'\psi^{-1}h) \end{aligned} \quad (10)$$

Lemma 1 (Rao and Statistiker (1973)). *Let A be an $n \times n$ matrix, B an $n \times m$ matrix, C an $m \times m$ matrix, and D an $m \times n$ matrix, where the necessary inverse exists. Then,*

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$$

By lemma 1, we can write the $\hat{\beta}_{SRMLE}$ as,

$$\begin{aligned} \hat{\beta}_{SRMLE} &= (C + H'\psi^{-1}H)^{-1}(X'\hat{W}z + H'\psi^{-1}h) \\ &= A(X'\hat{W}z + H'\psi^{-1}h). \end{aligned} \quad (11)$$

where $A = (C + H'\psi^{-1}H)^{-1}$.

The asymptotic properties of SRMLE are

$$E(\hat{\beta}_{SRMLE}) = \beta, \quad (12)$$

and

$$Cov(\hat{\beta}_{SRMLE}) = (C + H'\psi^{-1}H)^{-1} = A. \quad (13)$$

As a result, the mean square error matrix is produced as

$$MSEM(\hat{\beta}_{SRMLE}) = (C + H'\psi^{-1}H)^{-1} = A. \quad (14)$$

Varathan and Wijekoon (2016), introduced the Stochastic Restricted Ridge Logistic Estimator (SRRMLE) by replacing $\hat{\beta}_{MLE}$ by $\hat{\beta}_{LRE}$ in equation (9), which is defined as

$$\begin{aligned} \hat{\beta}_{SRRMLE} &= \hat{\beta}_{LRE} + C^{-1}H'(\psi + HC^{-1}H')^{-1}(h - H\hat{\beta}_{LRE}) \\ &= Z_k\hat{\beta}_{MLE} + C^{-1}H'(\psi + HC^{-1}H')^{-1}(h - HZ_k\hat{\beta}_{MLE}) \\ &= Z_kC^{-1}X'\hat{W}z + C^{-1}H'(\psi + HC^{-1}H')^{-1}(h - HZ_kC^{-1}X'\hat{W}z) \end{aligned}$$

Since C is non singular matrix,

$$\begin{aligned} Z_k C^{-1} &= (C + kI)^{-1} C C^{-1} \\ &= (C + kI)^{-1} I \\ &= C^{-1} (I + kC^{-1})^{-1} \\ &= C^{-1} (C + kI)^{-1} C \\ &= C^{-1} Z_k \end{aligned}$$

$$\begin{aligned} \hat{\beta}_{SRRMLE} &= C^{-1} Z_k X' \hat{W} z + C^{-1} H' (\psi + HC^{-1} H')^{-1} (h - HC^{-1} Z_k X' \hat{W} z) \\ &= (C^{-1} - C^{-1} H' (\psi + HC^{-1} H')^{-1} HC^{-1}) (Z_k X' \hat{W} z + H' \psi^{-1} h) \\ &= (C + H' \psi^{-1} H)^{-1} (Z_k X' \hat{W} z + H' \psi^{-1} h) \\ &= A(Z_k X' \hat{W} z + H' \psi^{-1} h). \end{aligned} \quad (15)$$

Li *et al.* (2020) introduced the Stochastic Restricted Logistic Liu Estimator (SRLLE) by replacing $\hat{\beta}_{MLE}$ by $\hat{\beta}_{LLE}$ in equation (9), which is defined as

$$\begin{aligned} \hat{\beta}_{SRLLE} &= \hat{\beta}_{LLE} + C^{-1} H' (\psi + HC^{-1} H')^{-1} (h - H \hat{\beta}_{LLE}) \\ &= Z_d \hat{\beta}_{MLE} + C^{-1} H' (\psi + HC^{-1} H')^{-1} (h - H Z_d \hat{\beta}_{MLE}) \\ &= A(Z_d X' \hat{W} z + H' \psi^{-1} h). \end{aligned} \quad (16)$$

By replacing $\hat{\beta}_{MLE}$ by $\hat{\beta}_{LTLE}$ in equation (9), Yehia (2020), developed the Stochastic Restricted Mixed Liu-Type estimator (SRMLTE).

$$\begin{aligned} \hat{\beta}_{SRMLTE} &= \hat{\beta}_{LTLE} + C^{-1} H' (\psi + HC^{-1} H')^{-1} (h - H \hat{\beta}_{LTLE}) \\ &= Z_{k,d} \hat{\beta}_{MLE} + C^{-1} H' (\psi + HC^{-1} H')^{-1} (h - H Z_{k,d} \hat{\beta}_{MLE}) \\ &= A(Z_{k,d} X' \hat{W} z + H' \psi^{-1} h). \end{aligned} \quad (17)$$

2.3. The general form of existing stochastic restricted mixed logistic estimators

The general form of the SRMLE, SRRMLE, SRLLE, and SRMLTE can be written as

$$\begin{aligned} \hat{\beta}_{SRGMLE} &= \hat{\beta}_{GLE} + C^{-1} H' (\psi + HC^{-1} H')^{-1} (h - H \hat{\beta}_{GLE}) \\ &= L_{(i)} \hat{\beta}_{MLE} + C^{-1} H' (\psi + HC^{-1} H')^{-1} (h - H L_{(i)} \hat{\beta}_{MLE}) \\ &= (C + H' \psi^{-1} H)^{-1} (L_{(i)} X' \hat{W} z + H' \psi^{-1} h) \\ &= A(L_{(i)} X' \hat{W} z + H' \psi^{-1} h). \end{aligned} \quad (18)$$

where $L_{(i)}$ is a non negative definite matrix,

$$\hat{\beta}_{SRGMLE} = \begin{cases} \hat{\beta}_{SRMLE} & \text{if } L_{(i)} = I; \\ \hat{\beta}_{SRRMLE} & \text{if } L_{(i)} = Z_k; \\ \hat{\beta}_{SRLLE} & \text{if } L_{(i)} = Z_d; \\ \hat{\beta}_{SRMLTE} & \text{if } L_{(i)} = Z_{k,d}. \end{cases} \quad (19)$$

The asymptotic properties of the general form of stochastic restricted mixed logistic estimators are

$$\begin{aligned} E[\hat{\beta}_{SRGMLE}] &= E[(C + H' \psi^{-1} H)^{-1} (L_{(i)} X' \hat{W} z + H' \psi^{-1} h)] \\ &= (C + H' \psi^{-1} H)^{-1} (L_{(i)} C + H' \psi^{-1} H) \beta \\ &= A(L_{(i)} C + H' \psi^{-1} H) \beta. \end{aligned} \quad (20)$$

and the dispersion matrix;

$$\begin{aligned} D[\hat{\beta}_{SRGMLE}] &= Cov[\hat{\beta}_{SRGMLE}] \\ &= (C + H' \psi^{-1} H)^{-1} (L_{(i)} C L_{(i)}' + H' \psi^{-1} H) (C + H' \psi^{-1} H)^{-1} \\ &= A(L_{(i)} C L_{(i)}' + H' \psi^{-1} H) A. \end{aligned} \quad (21)$$

The bias vector and Mean square error matrix (MSE) are

$$\begin{aligned} B[\hat{\beta}_{SRGMLE}] &= E[\hat{\beta}_{SRGMLE}] - \beta \\ &= (C + H' \psi^{-1} H)^{-1} (L_{(i)} - I) C \beta \\ &= A(L_{(i)} - I) C \beta. \end{aligned} \quad (22)$$

and

$$\begin{aligned} MSE[\hat{\beta}_{SRGMLE}] &= D[\hat{\beta}_{SRGMLE}] + B[\hat{\beta}_{SRGMLE}] B' [\hat{\beta}_{SRGMLE}] \\ &= A(L_{(i)} C L_{(i)}' + H' \psi^{-1} H) A + A(L_{(i)} - I) C \beta \beta' C (L_{(i)} - I)' A. \end{aligned} \quad (23)$$

The scalar mean square error (SMSE) can be obtained as,

$$\begin{aligned} SMSE[\hat{\beta}_{SRGMLE}] &= tr[MSE(\hat{\beta}_{SRGMLE})] \\ &= tr(A(L_{(i)} C L_{(i)}' + H' \psi^{-1} H) A) + \beta' C (L_{(i)} - I)' A A (L_{(i)} - I) C \beta. \end{aligned} \quad (24)$$

2.4. The stochastic restricted optimal logistic estimator (SROLE) and its properties

Varathan and Wijekoon (2021) introduced the SROLE, as the optimal estimator of SRMLE, SRRLE, SRLMLE, SRAULLE, SRAURLE, and SRLTLE. It is defined as,

$$\hat{\beta}_{SROLE} = \hat{F}_{opt} \hat{\beta}_{SRMLE}$$

where, $\hat{F}_{opt} = \beta \beta' (A + \beta \beta')^{-1}$.

The expectation, bias, covariance, MSE, and SMSE of SROLE are defined as:

$$E[\hat{\beta}_{SROLE}] = \hat{F}_{opt} \beta. \quad (25)$$

$$B[\hat{\beta}_{SROLE}] = (\hat{F}_{opt} - I) \beta \quad (26)$$

$$D[\hat{\beta}_{SROLE}] = Cov[\hat{\beta}_{SROLE}] = \hat{F}_{opt} A \hat{F}_{opt}'. \quad (27)$$

$$\begin{aligned} MSE[\hat{\beta}_{SROLE}] &= D[\hat{\beta}_{SROLE}] + B[\hat{\beta}_{SROLE}] B' [\hat{\beta}_{SROLE}] \\ &= \hat{F}_{opt} A \hat{F}_{opt}' + (\hat{F}_{opt} - I) \beta \beta' (\hat{F}_{opt} - I)'. \end{aligned} \quad (28)$$

$$\begin{aligned}
& SMSE[\hat{\beta}_{SROLE}] \\
& = tr(\hat{F}_{opt} A \hat{F}_{opt}') + \beta' (\hat{F}_{opt} - I)' (\hat{F}_{opt} - I) \beta.
\end{aligned} \tag{29}$$

2.5. The proposed estimators

The general form of stochastic restricted mixed Logistic estimator (equation (18)) consists of the matrix $L_{(i)}$, which takes different choices depending on different estimators. Now, we minimize the SMSE of SRGMLE with respect to $L_{(i)}$ to identify the most suitable.

$$\begin{aligned}
& \frac{\partial SMSE(\hat{\beta}_{SRGMLE})}{\partial L_{(i)}} \\
& = \frac{\partial tr[A(L_{(i)} C L_{(i)} + H' \psi^{-1} H) A]}{\partial L_{(i)}} + \frac{\partial \beta' C (L_{(i)} - I)' A A (L_{(i)} - I) C \beta}{\partial L_{(i)}} \\
& = \frac{\partial tr(A L_{(i)} C L_{(i)} A)}{\partial L_{(i)}} + \frac{\partial \beta' C L_{(i)} A A L_{(i)} C \beta - 2 \beta' C A A L_{(i)} C \beta + \beta' C A A C \beta}{\partial L_{(i)}} \\
& = \frac{\partial tr(L_{(i)} A A L_{(i)} C)}{\partial L_{(i)}} + \frac{\partial \beta' C L_{(i)} A A L_{(i)} C \beta}{\partial L_{(i)}} - 2 \frac{\partial \beta' C A A L_{(i)} C \beta}{\partial L_{(i)}}
\end{aligned} \tag{30}$$

We consider the following lemmas to further simplify the above equation.

Lemma 2 (Rao and Toutenburg (1995)). *Let a be a $n \times 1$ vector, N a symmetric $t \times t$ matrix, and M a $t \times n$ matrix, then*

$$\frac{\partial (a' M' N M a)}{\partial M} = 2 N M a a'$$

Lemma 3 (Rao and Toutenburg (1995)). *Let a is a vector of order $n \times 1$, b is another vector of order $m \times 1$, and M is an $n \times m$ matrix, then*

$$\frac{\partial (a' M b)}{\partial M} = a b'$$

Lemma 4 (Rao and Toutenburg (1995)). *For the differentials of the trace we have*

$$\frac{\partial tr(X' A X B)}{\partial X} = A X B + A' X B'$$

by applying Lemmas 2, 3, and 4 in equation (30), we obtain

$$\begin{aligned}
\frac{\partial SMSE(\hat{\beta}_{SRGMLE})}{\partial L_{(i)}} & = A A L_{(i)} C + A A L_{(i)} C + 2 A A L_{(i)} C \beta \beta' C - 2 A A C \beta \beta' C \\
& = 2 A A L_{(i)} C + 2 A A L_{(i)} C \beta \beta' C - 2 A A C \beta \beta' C
\end{aligned} \tag{31}$$

$$\frac{\partial SMSE(\hat{\beta}_{SRGMLE})}{\partial L_{(i)}} \Big|_{L_{(i)} = \hat{L}} = 0 \tag{32}$$

Lemma 5 (Rao and Toutenburg (1995)). *Let $A : n \times n$ and $B : n \times n$ such that $A > 0$ and $B : n \times n \geq 0$. Then $C = A + B > 0$.*

According to lemma 5, $(I + C\beta\beta')$ is a nonsingular matrix. By equating (32) to a null matrix, we obtain the \hat{L} ,

$$\begin{aligned} 2AA\hat{L}C + 2AA\hat{L}C\beta\beta'C - 2AAC\beta\beta'C &= 0 \\ 2AA(\hat{L} + \hat{L}C\beta\beta')C &= 2AAC\beta\beta'C \\ \hat{L} + \hat{L}C\beta\beta' &= C\beta\beta' \\ \hat{L}(I + C\beta\beta') &= C\beta\beta' \\ \hat{L} &= C\beta\beta'(I + C\beta\beta')^{-1}. \end{aligned} \quad (33)$$

By replacing $\hat{\beta}_{GLE}$ with $\hat{L}\hat{\beta}_{GLE}$ in equation (18), now we propose a new estimator called a Stochastic Restricted Modified Mixed Logistic Estimator (SRMMLE),

$$\begin{aligned} \hat{\beta}_{SRMMLE} &= \hat{L}\hat{\beta}_{GLE} + C^{-1}H'(\psi + HC^{-1}H')^{-1}(h - H\hat{L}\hat{\beta}_{GLE}) \\ &= \hat{L}L_{(i)}\hat{\beta}_{MLE} + C^{-1}H'(\psi + HC^{-1}H')^{-1}(h - H\hat{L}L_{(i)}\hat{\beta}_{MLE}) \\ &= (C + H'\psi^{-1}H)^{-1}(\hat{L}L_{(i)}X'\hat{W}z + H'\psi^{-1}h) \\ &= (C + H'\psi^{-1}H)^{-1}(J_{(i)}X'\hat{W}z + H'\psi^{-1}h). \end{aligned} \quad (34)$$

where, $J_{(i)} = \hat{L}L_{(i)}$.

The proposed estimator, SRMMLE involves an unknown parameter β within its term \hat{L} , so it becomes essential to look for a vector with known β values. As such, following [Varathan and Wijekoon \(2021\)](#), we replace β with the normalized eigenvector corresponding to the largest eigenvalue of the matrix C , subject to the constraint $\beta'\beta = 1$.

By adopting $\hat{\beta}_{MLE}$, $\hat{\beta}_{LRE}$, $\hat{\beta}_{LLE}$, and $\hat{\beta}_{LTLE}$ in place of $\hat{\beta}_{GLE}$ in equation (34), we propose four new estimators namely, Stochastic Restricted Modified Mixed Logistic Estimator 1 (SRMMLE 1), Stochastic Restricted Modified Mixed Logistic Estimator 2 (SRMMLE 2), Stochastic Restricted Modified Mixed Logistic Estimator 3 (SRMMLE 3), and Stochastic Restricted Modified Mixed Logistic Estimator 4 (SRMMLE 4), respectively, and defined as,

$$\hat{\beta}_{SRMMLE} = \begin{cases} \hat{\beta}_{SRMMLE1} & \text{if } J_{(i)} = \hat{L}I; \\ \hat{\beta}_{SRMMLE2} & \text{if } J_{(i)} = \hat{L}Z_k; \\ \hat{\beta}_{SRMMLE3} & \text{if } J_{(i)} = \hat{L}Z_d; \\ \hat{\beta}_{SRMMLE4} & \text{if } J_{(i)} = \hat{L}Z_{k,d}. \end{cases} \quad (35)$$

2.6. Asymptotic properties of the proposed estimators

In this section, we defined the asymptotic properties of the proposed estimator, SRMMLE. Since the present estimator (SRMMLE) is in a similar form of the estimator SRGMLE, it can be obtained by replacing $L_{(i)}$ with $J_{(i)}$ in equations (20) - (24).

The expected value of SRMMLE can be written as,

$$E[\hat{\beta}_{SRMMLE}] = A(J_{(i)}C + H'\psi^{-1}H)\beta. \quad (36)$$

and the dispersion matrix;

$$D[\hat{\beta}_{SRMMLE}] = Cov[\hat{\beta}_{SRMMLE}] = A(J_{(i)}CJ_{(i)} + H'\psi^{-1}H)A. \quad (37)$$

The bias vector and MSE can be

$$B[\hat{\beta}_{SRMMLE}] = A(J_{(i)} - I)C\beta \quad (38)$$

and

$$\begin{aligned} MSE[\hat{\beta}_{SRMMLE}] &= D[\hat{\beta}_{SRMMLE}] + B[\hat{\beta}_{SRMMLE}]B'[\hat{\beta}_{SRMMLE}] \\ &= A(J_{(i)}CJ_{(i)} + H'\psi^{-1}H)A + A(J_{(i)} - I)C\beta\beta'C(J_{(i)} - I)'A. \end{aligned} \quad (39)$$

The SMSE can be obtained as,

$$\begin{aligned} SMSE[\hat{\beta}_{SRMMLE}] \\ = tr(A(J_{(i)}CJ_{(i)} + H'\psi^{-1}H)A) + \beta'C(J_{(i)} - I)'AA(J_{(i)} - I)C\beta. \end{aligned} \quad (40)$$

3. Comparison among the estimators

3.1. Theoretical comparison

In this section, we compare the performance of the proposed estimator SRMMLE with the existing estimators SRGMLE in terms of the mean square error matrix criterion by following lemmas.

Lemma 6 (Rao and Toutenburg (1995)). *Let A be positive definite and B be a regular matrix, then $B'AB > 0$.*

Lemma 7 (Rao, Shalabh, Toutenburg, and Heumann (2008)). *Let the two $n \times n$ matrices $M > 0$, $N \geq 0$, then $M > N$ if and only if $\lambda_{max}(NM^{-1}) < 1$.*

Lemma 8 (Trenkler and Toutenburg (1990)). *Let $\hat{\beta}_j = A_j y, j = 1, 2$ be two competing homogenous linear estimators of β . Suppose that $D = Cov(\hat{\beta}_1) - Cov(\hat{\beta}_2) > 0$; where $Cov(\hat{\beta}_j), j = 1, 2$ denotes the covaraince matrix of $\hat{\beta}_j$. Then $\Delta(\hat{\beta}_1, \hat{\beta}_2) = MSEM(\hat{\beta}_1) - MSEM(\hat{\beta}_2) \geq 0$ if and only if $d'_2(D + d'_1d_1)d_2 \leq 1$, where $MSEM(\hat{\beta}_j), d_j; j = 1, 2$ denote the Mean Square Error Matrix and bias vector of $\hat{\beta}_j$, respectively.*

Theorem 3.1. *When $\lambda_{max}[AJ_{(i)}CJ_{(i)}'A(AL_{(i)}CL_{(i)}'A)^{-1}] < 1$, the estimator SRMMLE is superior to SRGMLE if and only if $\delta'_{new}(D_1 + \delta'_G\delta_G)^{-1}\delta_{new} \leq 1$.*

Proof. Consider,

$$\begin{aligned} MSEM(\hat{\beta}_{SRGMLE}) - MSEM(\hat{\beta}_{SRMMLE}) \\ = A(L_{(i)}CL_{(i)} + H'\psi^{-1}H)A + A(L_{(i)} - I)C\beta\beta'C(L_{(i)} - I)'A \\ - A(J_{(i)}CJ_{(i)} + H'\psi^{-1}H)A + A(J_{(i)} - I)C\beta\beta'C(J_{(i)} - I)'A \\ = A(L_{(i)}CL_{(i)} - J_{(i)}CJ_{(i)})A + A[(L_{(i)} - I)C\beta\beta'C(L_{(i)} - I)' \\ - (J_{(i)} - I)C\beta\beta'C(J_{(i)} - I)']A. \end{aligned} \quad (41)$$

Now consider,

$$\begin{aligned} D(\hat{\beta}_{SRGMLE}) - D(\hat{\beta}_{SRMMLE}) &= A(L_{(i)}CL_{(i)}' - J_{(i)}CJ_{(i)}')A \\ &= D_1. \end{aligned} \quad (42)$$

Note that, $AL_{(i)}CL_{(i)}'A$ and $AJ_{(i)}CJ_{(i)}'A$ are positive definite matrices (by lemma 6) since C is a positive definite matrix. Consequently, by lemma 7, if $\lambda_{max}[AJ_{(i)}CJ_{(i)}'A(AL_{(i)}CL_{(i)}'A)^{-1}] < 1$ then D_1 is a positive definite matrix, where $\lambda_{max}[AJ_{(i)}CJ_{(i)}'A(AL_{(i)}CL_{(i)}'A)^{-1}]$ is the largest eigenvalue of $[AJ_{(i)}CJ_{(i)}'A(AL_{(i)}CL_{(i)}'A)^{-1}]$. Further by lemma 8, $MSEM(\hat{\beta}_{SRGMLE}), MSEM(\hat{\beta}_{SRMMLE})$ is non negative definite if $\delta'_{new}(D_1 + \delta'_G\delta_G)^{-1}\delta_{new} \leq 1$, where $\delta_{new} = A(J_{(i)} - I)C\beta$ and $\delta_G = A(L_{(i)} - I)C\beta$. \square

Note that the above theorem outlines the necessary and sufficient conditions for the superiority of the proposed estimator (SRMMLE) over the general existing estimator (SRGMLE). By substituting $L_{(i)}$ with an appropriate matrix, we can derive the following conditions for the superiority of SRMMLE over the existing estimators SRMLE, SRRMLE, SRLLE, and SRMLTE with respect to mean square error matrix (MSEM).

- If $L_{(i)} = I$; SRMMLE is superior than SRMLE when $\lambda_{max}[AJ_{(i)}CJ_{(i)}'A(ACA)^{-1}] < 1$.
- If $L_{(i)} = Z_k$; SRMMLE is superior than SRRMLE when $\lambda_{max}[AJ_{(i)}CJ_{(i)}'A(AZ_kCZ_k'A)^{-1}] < 1$.
- If $L_{(i)} = Z_d$; SRMMLE is superior than SRLLE when $\lambda_{max}[AJ_{(i)}CJ_{(i)}'A(AZ_dCZ_d'A)^{-1}] < 1$.
- If $L_{(i)} = Z_{k,d}$; SRMMLE is superior than SRMLTE when $\lambda_{max}[J_{(i)}CJ_{(i)}'A(AZ_{k,d}CZ_{k,d}'A)^{-1}] < 1$.

3.2. Numerical illustration

The proposed estimators, SRMMLE1, SRMMLE2, SRMMLE3, and SRMMLE4, are compared with the existing estimators SRMLE, SRRMLE, SRLLE, SRMLTE, and SROLE using the following classification metrics.

The confusion matrix offers a clear understanding of how effectively the model distinguishes between positive and negative instances. Table 1 provides the components of the confusion matrix.

The evaluation metrics are,

Table 1: Confusion Matrix
Predicted Outcome

		Predicted Outcome	
		P	N
Actual	P	True Positive (TP)	False Negative (FN)
	N	False Positive (FP)	True Negative (TN)

Sensitivity / Recall: This measures how effectively the model identifies true positive cases.

$$Sensitivity/Recall = \frac{TP}{(TP + FN)} \quad (43)$$

Specificity: Specificity measures the true negative rate, indicating how accurately the model identifies negative cases.

$$Specificity = \frac{TN}{(TN + FP)} \quad (44)$$

Precision: Precision measures the proportion of true positive predictions among all positive predictions.

$$Precision = \frac{TP}{(TP + FP)} \quad (45)$$

Balanced accuracy: This metric averages sensitivity and specificity to give a more balanced view of performance of the model.

$$\text{Balanced accuracy} = \frac{(\text{Sensitivity} + \text{Specificity})}{2} \quad (46)$$

F1 Score: The F1 Score is the harmonic mean of precision and recall, balancing the two.

$$F1score = 2 * \frac{(\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}} \quad (47)$$

AUC Value (Area Under the Curve): AUC measures the overall ability of the model to distinguish between positive and negative cases.

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC): This is a widely used metric for evaluating the performance of binary classifiers. It measures the likelihood that a model will correctly rank a randomly chosen positive instance higher than a randomly chosen negative one.

Simulation study

Following McDonald and Galarneau (1975), and Alheety, Månsson, and Kibria (2021), we generate the predictor variables using following equation (48).

$$x_{i,j} = \sqrt{(1 - \rho^2)}z_{i,j} + \rho z_{i,p+1} \quad ; i = 1, 2, \dots, n. \quad j = 1, 2, \dots, p. \quad (48)$$

where z_{ij} are pseudo-random numbers from a standard normal distribution and ρ represent the correlation between any two predictor variables. Four predictor variables are generated, and we choose $\rho = (0.7, 0.9, 0.99, 0.995, 0.999)$ from moderate to severe multicollinearity. Further, in this study, we considered five different sample sizes: 30, 50, 100, 200, and 1000. The output was unreliable when we chose sample sizes of 30 and 50 for cross-validated balanced accuracy. Therefore, we chose 100 as the small sample for balanced accuracy calculation. The dependent variable y_i is obtained from the Bernoulli distribution with $\pi_i = \frac{\exp(x'_i\beta)}{1 + \exp(x'_i\beta)}$. The

parameter values of $\beta_1, \beta_2, \dots, \beta_p$ are chosen so that $\beta' \beta = 1$ and $\beta_1 = \beta_2 = \dots, \beta_p$.

Following Varathan and Wijekoon (2021), we take the restriction as follows:

$$H = \begin{pmatrix} 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \quad h = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \text{ and } \psi = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (49)$$

The simulation is repeated 2000 times by generating new pseudo-random numbers, and we estimate the SMSE of each estimator for different (k, d) . The initial values of k and d are set at 0.01, and are incrementally increased by 0.1. For $k > 0, 0 < d < 1$.

$$SMSE(\hat{\beta}^*) = \frac{1}{2000} \sum_{r=1}^{2000} (\hat{\beta}_r - \beta)' (\hat{\beta}_r - \beta) \quad (50)$$

where $\hat{\beta}_r$ denotes any estimator considered in the r^{th} simulation.

Further, a 5-fold cross-validation was implemented to compute the average cross-validated balanced accuracy for each estimator. The results of the simulations are presented in Tables (2) - (5), showing the minimum SMSE values and the highest balanced accuracy along with their respective shrinkage parameters. Additionally, these findings are depicted in Figure (1) - (2).

According to the tables 2 - 5, our proposed estimators exhibit enhanced performance relative to the existing methods. In particular, SRMMLE1 outperforms SRMLE, whereas SRMMLE2 demonstrates superiority over SRMLE, SRLLE, SROLE, and SRRMLE. Similarly, SRMMLE3 achieves better results than SRLLE and SROLE. Notably, SRMMLE4 surpasses all the considered estimators across various sample sizes and ρ values.

From Tables 3 - 5, SRMLTE exhibits comparatively better balanced accuracy for $\rho = 0.7$ and 0.9 when $n = 100$, for $\rho = 0.9, 0.995$, and 0.999 when $n = 200$, and $\rho = 0.7, 0.99$ and 0.999 when $n = 1000$. Meanwhile, SRMMLE4 achieves better balanced accuracy compared to other estimators for $\rho = 0.99, 0.995$, and 0.999 when $n = 100$, for $\rho = 0.99$ when $n = 200$, and $\rho = 0.9, 0.99$ and 0.995 when $n = 1000$.

Further, we identified the parameter combination that minimizes SMSE and the combination that maximizes average cross-validated balanced accuracy. Based on Table 3 - 5, the minimum SMSE and maximum balanced accuracy for each estimator are shown for different shrinkage values.

From Figure (1), the performance of the estimators in terms of SMSE for each sample size and correlation can be ranked as follows: SRMMLE4, SRMLTE, SRMMLE2, SRRMLE, SRMMLE3, SROLE, SRLLE, SRMMLE1, and SRMLE. Moreover, an increase in both the sample size and the correlation between the two predictor variables (ρ) resulted in a decrease in the SMSE of SRMLE, demonstrating enhanced performance. Conversely, the performance of other estimators declined as the sample size and correlation increased.

Table 2: The SMSE of the simulation study when $n=30$ and 50

ρ	Estimators	$n = 30$	$n = 50$
$\rho = 0.7$	SRMLE	5.44984	5.48675
	SRLLE	1.60335 ($d = 0.01$)	1.74686 ($d = 0.01$)
	SROLE	1.59354	1.73397
	SRRMLE	1.11852 ($k = 4.91$)	1.14744 ($k = 4.91$)
	SRMLTE	0.85906 ($k = 2.91, d = 0.99$)	0.86632 ($k = 2.51, d = 0.99$)
	SRMMLE1	1.84187	1.87487
	SRMMLE2	1.03409 ($k = 4.91$)	1.04527 ($k = 4.91$)
	SRMMLE3	1.14574 ($d = 0.01$)	1.19038 ($d = 0.01$)
	SRMMLE4	0.82459 ($k = 0.81, d = 0.99$)	0.82457 ($k = 0.51, d = 0.99$)
$\rho = 0.9$	SRMLE	4.56791	4.73790
	SRLLE	1.69309 ($d = 0.01$)	1.87863 ($d = 0.01$)
	SROLE	1.68225	1.86559
	SRRMLE	1.14488 ($k = 4.91$)	1.18847 ($k = 4.91$)
	SRMLTE	0.89289 ($k = 2.31, d = 0.99$)	0.89706 ($k = 1.91, d = 0.99$)
	SRMMLE1	1.92399	2.02651
	SRMMLE2	1.04908 ($k = 4.91$)	1.06897 ($k = 4.91$)
	SRMMLE3	1.19796 ($d = 0.01$)	1.27216 ($d = 0.01$)
	SRMMLE4	0.85085 ($k = 0.71, d = 0.99$)	0.84083 ($k = 0.51, d = 0.99$)
$\rho = 0.99$	SRMLE	4.35108	4.40716
	SRLLE	1.76694 ($d = 0.01$)	1.93341 ($d = 0.01$)
	SROLE	1.75057	1.91655
	SRRMLE	1.16777 ($k = 4.91$)	1.20752 ($k = 4.91$)
	SRMLTE	0.88659 ($k = 1.81, d = 0.99$)	0.91407 ($k = 1.21, d = 0.99$)
	SRMMLE1	2.02324	2.08420
	SRMMLE2	1.06082 ($k = 4.91$)	1.08168 ($k = 4.91$)
	SRMMLE3	1.23639 ($d = 0.01$)	1.31219 ($d = 0.01$)
	SRMMLE4	0.84491 ($k = 0.71, d = 0.99$)	0.86988 ($k = 0.51, d = 0.99$)
$\rho = 0.995$	SRMLE	4.21095	4.39081
	SRLLE	1.73206 ($d = 0.01$)	1.93321 ($d = 0.01$)
	SROLE	1.71624	1.91604
	SRRMLE	1.15866 ($k = 4.91$)	1.20736 ($k = 4.91$)
	SRMLTE	0.91913 ($k = 2.11, d = 0.99$)	0.90347 ($k = 1.01, d = 0.99$)
	SRMMLE1	1.99014	2.09026
	SRMMLE2	1.05818 ($k = 4.91$)	1.08160 ($k = 4.91$)
	SRMMLE3	1.22620 ($d = 0.01$)	1.31206 ($d = 0.01$)
	SRMMLE4	0.85898 ($k = 0.71, d = 0.99$)	0.85498 ($k = 0.41, d = 0.99$)
$\rho = 0.999$	SRMLE	4.21423	4.38388
	SRLLE	1.74376 ($d = 0.01$)	1.93586 ($d = 0.01$)
	SROLE	1.72744	1.91839
	SRRMLE	1.16212 ($k = 4.91$)	1.20760 ($k = 4.91$)
	SRMLTE	0.90851 ($k = 1.91, d = 0.99$)	0.89331 ($k = 0.91, d = 0.99$)
	SRMMLE1	1.99998	2.08873
	SRMMLE2	1.05947 ($k = 4.91$)	1.08137 ($k = 4.91$)
	SRMMLE3	1.23081 ($d = 0.01$)	1.31222 ($d = 0.01$)
	SRMMLE4	0.85896 ($k = 0.71, d = 0.99$)	0.85651 ($k = 0.41, d = 0.99$)

Table 3: The results of simulation study when $n=100$

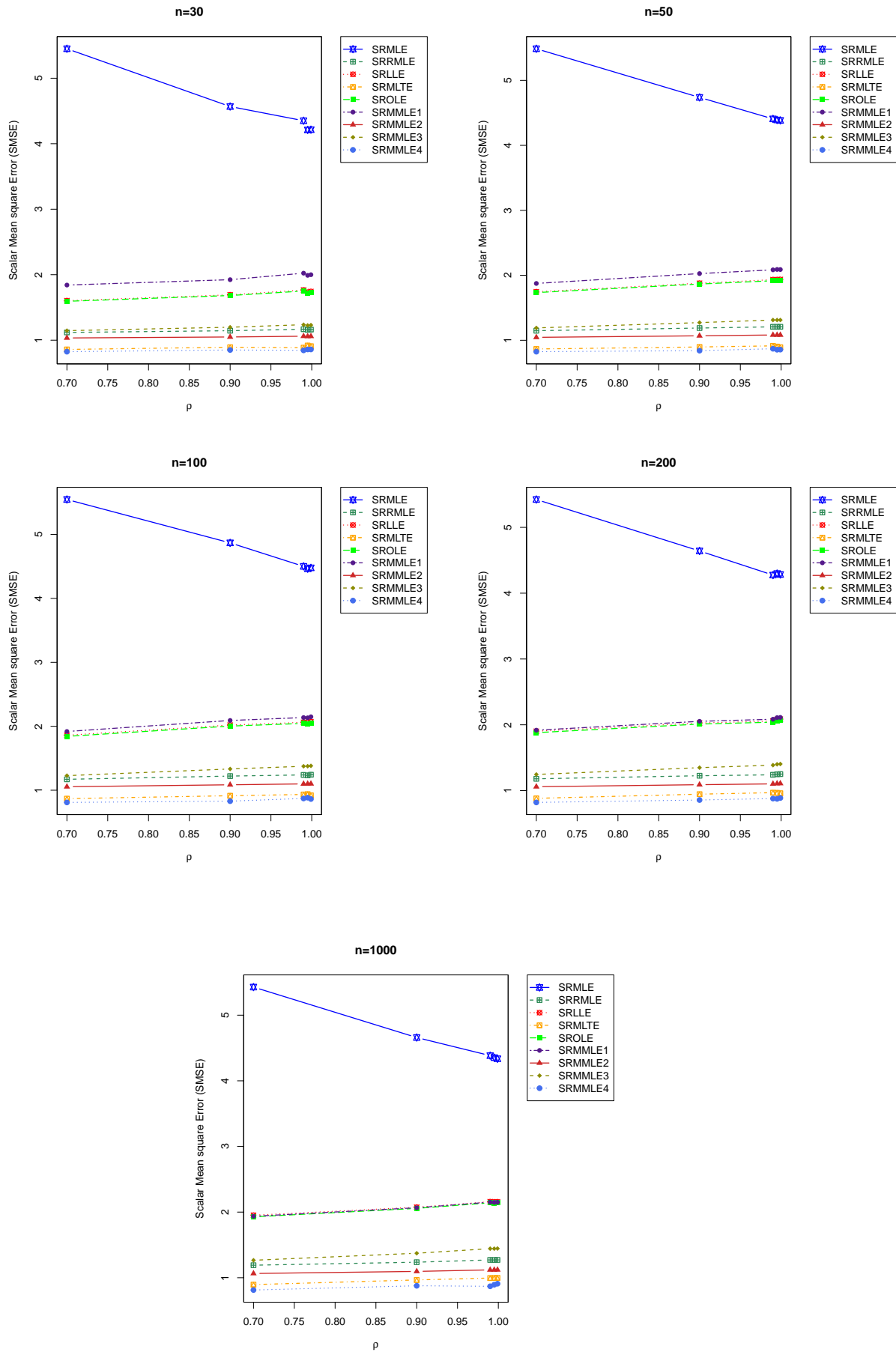
ρ	Estimators	SMSE	Balanced Accuracy
$\rho = 0.7$	SRMLE	5.54738	0.51734
	SRLLE	1.86034 ($d = 0.01$)	0.51775 ($d = 0.31$)
	SROLE	1.84213	0.51701
	SRRMLE	1.17075 ($k = 4.91$)	0.51785 ($k = 0.31$)
	SRMLTE	0.86858 ($k = 2.31, d = 0.99$)	0.51789 ($k = 0.21, d = 0.01$)
	SRMMLE1	1.91839	0.51723
	SRMMLE2	1.05467 ($k = 4.91$)	0.51735 ($k = 0.01$)
	SRMMLE3	1.22791 ($d = 0.01$)	0.51733 ($d = 0.61$)
	SRMMLE4	0.80917 ($k = 0.31, d = 0.99$)	0.51736 ($k = 0.11, d = 0.01$)
$\rho = 0.9$	SRMLE	4.86889	0.52242
	SRLLE	2.01943 ($d = 0.01$)	0.52298 ($d = 0.11$)
	SROLE	2.00329	0.52274
	SRRMLE	1.22126 ($k = 4.91$)	0.52307 ($k = 0.11$)
	SRMLTE	0.91406 ($k = 1.81, d = 0.99$)	0.52389 ($k = 0.51, d = 0.21$)
	SRMMLE1	2.09087	0.52269
	SRMMLE2	1.08500 ($k = 4.91$)	0.52271 ($k = 1.61$)
	SRMMLE3	1.33229 ($d = 0.01$)	0.52283 ($d = 0.41$)
	SRMMLE4	0.82909 ($k = 0.21, d = 0.99$)	0.52292 ($k = 0.51, d = 0.11$)
$\rho = 0.99$	SRMLE	4.50125	0.52541
	SRLLE	2.06244 ($d = 0.01$)	0.52542 ($d = 0.71$)
	SROLE	2.04493	0.52546
	SRRMLE	1.23941 ($k = 4.91$)	0.52546 ($k = 0.41$)
	SRMLTE	0.93352 ($k = 0.99, d = 0.99$)	0.52564 ($k = 0.01, d = 0.11$)
	SRMMLE1	2.13653	0.52553
	SRMMLE2	1.09883 ($k = 4.91$)	0.52548 ($k = 0.01$)
	SRMMLE3	1.37484 ($d = 0.01$)	0.52560 ($d = 0.51$)
	SRMMLE4	0.87144 ($k = 0.21, d = 0.99$)	0.52576 ($k = 4.51, d = 0.21$)
$\rho = 0.995$	SRMLE	4.46385	0.52584
	SRLLE	2.05931 ($d = 0.01$)	0.52598 ($d = 0.01$)
	SROLE	2.04181	0.52598
	SRRMLE	1.23908 ($k = 4.91$)	0.52605 ($k = 0.01$)
	SRMLTE	0.93906 ($k = 0.81, d = 0.99$)	0.52607 ($k = 0.11, d = 0.31$)
	SRMMLE1	2.13242	0.52566
	SRMMLE2	1.09939 ($k = 4.91$)	0.52598 ($k = 0.21$)
	SRMMLE3	1.37589 ($d = 0.01$)	0.52612 ($d = 0.31$)
	SRMMLE4	0.88362 ($k = 0.21, d = 0.99$)	0.52623 ($k = 2.71, d = 0.11$)
$\rho = 0.999$	SRMLE	4.47801	0.52630
	SRLLE	2.07160 ($d = 0.01$)	0.52635 ($d = 0.21$)
	SROLE	2.05381	0.52630
	SRRMLE	1.24261 ($k = 4.91$)	0.52631 ($k = 0.01$)
	SRMLTE	0.92097 ($k = 0.51, d = 0.99$)	0.52635 ($k = 0.11, d = 0.31$)
	SRMMLE1	2.14853	0.52578
	SRMMLE2	1.10089 ($k = 4.91$)	0.52630 ($k = 0.11$)
	SRMMLE3	1.38124 ($d = 0.01$)	0.52626 ($d = 0.01$)
	SRMMLE4	0.86249 ($k = 0.21, d = 0.99$)	0.52636 ($k = 1.11, d = 0.21$)

Table 4: The results of the simulation study when $n=200$

ρ	Estimators	SMSE	Balanced Accuracy
$\rho = 0.7$	SRMLE	5.42375	0.51196
	SRLLE	1.90213 ($d = 0.01$)	0.51205 ($d = 0.11$)
	SROLE	1.88081	0.51236
	SRRMLE	1.17966 ($k = 4.91$)	0.51201 ($k = 4.91$)
	SRMLTE	0.88296 ($k = 2.21, d = 0.99$)	0.51232 ($k = 4.81, d = 0.21$)
	SRMMLE1	1.91855	0.51221
	SRMMLE2	1.05930 ($k = 4.91$)	0.51217 ($k = 0.01$)
	SRMMLE3	1.24572 ($d = 0.01$)	0.51221 ($d = 0.99$)
	SRMMLE4	0.82075 ($k = 0.21, d = 0.99$)	0.51229 ($k = 0.31, d = 0.41$)
$\rho = 0.9$	SRMLE	4.64176	0.51604
	SRLLE	2.02783 ($d = 0.01$)	0.51647 ($d = 0.11$)
	SROLE	2.01111	0.51643
	SRRMLE	1.22611 ($k = 4.91$)	0.51649 ($k = 0.31$)
	SRMLTE	0.94615 ($k = 2.01, d = 0.99$)	0.51693 ($k = 0.41, d = 0.41$)
	SRMMLE1	2.05288	0.51667
	SRMMLE2	1.09014 ($k = 4.91$)	0.51663 ($k = 0.01$)
	SRMMLE3	1.34892 ($d = 0.01$)	0.51675 ($d = 0.81$)
	SRMMLE4	0.85735 ($k = 0.11, d = 0.99$)	0.51656 ($k = 0.01, d = 0.01$)
$\rho = 0.99$	SRMLE	4.27511	0.51830
	SRLLE	2.05689 ($d = 0.01$)	0.51838 ($d = 0.21$)
	SROLE	2.04049	0.51833
	SRRMLE	1.24056 ($k = 4.91$)	0.51835 ($k = 0.01$)
	SRMLTE	0.96910 ($k = 1.21, d = 0.99$)	0.51833 ($k = 0.11, d = 0.01$)
	SRMMLE1	2.08495	0.51801
	SRMMLE2	1.10251 ($k = 4.91$)	0.51834 ($k = 0.11$)
	SRMMLE3	1.38696 ($d = 0.01$)	0.51834 ($d = 0.01$)
	SRMMLE4	0.87942 ($k = 0.11, d = 0.99$)	0.51841 ($k = 1.21, d = 0.21$)
$\rho = 0.995$	SRMLE	4.29871	0.51818
	SRLLE	2.08073 ($d = 0.01$)	0.51833 ($d = 0.01$)
	SROLE	2.06410	0.51833
	SRRMLE	1.24878 ($k = 4.91$)	0.51833 ($k = 0.41$)
	SRMLTE	0.96669 ($k = 0.91, d = 0.99$)	0.51878 ($k = 0.01, d = 0.11$)
	SRMMLE1	2.10871	0.51784
	SRMMLE2	1.10691 ($k = 4.91$)	0.51833 ($k = 3.21$)
	SRMMLE3	1.40017 ($d = 0.01$)	0.51832 ($d = 0.01$)
	SRMMLE4	0.87508 ($k = 0.11, d = 0.99$)	0.51854 ($k = 1.91, d = 0.31$)
$\rho = 0.999$	SRMLE	4.28388	0.51870
	SRLLE	2.08583 ($d = 0.01$)	0.51873 ($d = 0.51$)
	SROLE	2.06918	0.51870
	SRRMLE	1.25122 ($k = 4.91$)	0.51871 ($k = 0.11$)
	SRMLTE	0.96073 ($k = 0.51, d = 0.99$)	0.51887 ($k = 0.01, d = 0.21$)
	SRMMLE1	2.11258	0.51808
	SRMMLE2	1.10853 ($k = 4.91$)	0.51870 ($k = 0.01$)
	SRMMLE3	1.40460 ($d = 0.01$)	0.51868 ($d = 0.01$)
	SRMMLE4	0.88825 ($k = 0.11, d = 0.99$)	0.51882 ($k = 2.61, d = 0.31$)

Table 5: The results of the simulation study when n=1000

ρ	Estimators	SMSE	Balanced Accuracy
$\rho = 0.7$	SRMLE	5.42790	0.50561
	SRLLE	1.95269 ($d = 0.21$)	0.50572 ($d = 0.01$)
	SROLE	1.92775	0.50573
	SRRMLE	1.19100 ($k = 4.91$)	0.50576 ($k = 3.31$)
	SRMLTE	0.89547 ($k = 2.21, d = 0.99$)	0.50582 ($k = 1.01, d = 0.31$)
	SRMMLE1	1.93633	0.50573
	SRMMLE2	1.06466 ($k = 4.91$)	0.50575 ($k = 2.11$)
	SRMMLE3	1.26606 ($d = 0.01$)	0.50574 ($d = 0.01$)
	SRMMLE4	0.81290 ($k = 0.01, d = 0.99$)	0.50578 ($k = 0.11, d = 0.21$)
$\rho = 0.9$	SRMLE	4.65941	0.50754
	SRLLE	2.07462 ($d = 0.01$)	0.50774 ($d = 0.01$)
	SROLE	2.05651	0.50773
	SRRMLE	1.23671 ($k = 4.91$)	0.50778 ($k = 3.01$)
	SRMLTE	0.96645 ($k = 2.51, d = 0.99$)	0.50778 ($k = 0.71, d = 0.01$)
	SRMMLE1	2.06604	0.50769
	SRMMLE2	1.09649 ($k = 4.91$)	0.50773 ($k = 2.01$)
	SRMMLE3	1.37312 ($d = 0.01$)	0.50772 ($d = 0.01$)
	SRMMLE4	0.87739 ($k = 0.01, d = 0.99$)	0.50787 ($k = 1.51, d = 0.51$)
$\rho = 0.99$	SRMLE	4.38372	0.50860
	SRLLE	2.15837 ($d = 0.01$)	0.50866 ($d = 0.31$)
	SROLE	2.14144	0.50865
	SRRMLE	1.27132 ($k = 4.91$)	0.50866 ($k = 0.11$)
	SRMLTE	0.99530 ($k = 2.41, d = 0.99$)	0.50872 ($k = 0.11, d = 0.41$)
	SRMMLE1	2.15153	0.50862
	SRMMLE2	1.11965 ($k = 4.91$)	0.50865 ($k = 1.71$)
	SRMMLE3	1.44307 ($d = 0.01$)	0.50865 ($d = 0.31$)
	SRMMLE4	0.86959 ($k = 0.01, d = 0.99$)	0.50872 ($k = 1.41, d = 0.41$)
$\rho = 0.995$	SRMLE	4.35272	0.50865
	SRLLE	2.15492 ($d = 0.01$)	0.50868 ($d = 0.71$)
	SROLE	2.13809	0.50866
	SRRMLE	1.27055 ($k = 4.91$)	0.50865 ($k = 4.31$)
	SRMLTE	0.99663 ($k = 2.11, d = 0.99$)	0.50875 ($k = 0.01, d = 0.21$)
	SRMMLE1	2.14810	0.50864
	SRMMLE2	1.11962 ($k = 4.91$)	0.50867 ($k = 0.01$)
	SRMMLE3	1.44296 ($d = 0.01$)	0.50866 ($d = 0.21$)
	SRMMLE4	0.89384 ($k = 0.01, d = 0.99$)	0.50878 ($k = 0.11, d = 0.41$)
$\rho = 0.999$	SRMLE	4.33609	0.50876
	SRLLE	2.15626 ($d = 0.01$)	0.50877 ($d = 0.71$)
	SROLE	2.13949	0.50876
	SRRMLE	1.27131 ($k = 4.91$)	0.50876 ($k = 0.01$)
	SRMLTE	0.99767 ($k = 1.31, d = 0.99$)	0.50897 ($k = 0.01, d = 0.21$)
	SRMMLE1	2.14969	0.50866
	SRMMLE2	1.12032 ($k = 4.91$)	0.50876 ($k = 0.11$)
	SRMMLE3	1.44498 ($d = 0.01$)	0.50876 ($d = 0.01$)
	SRMMLE4	0.90888 ($k = 0.01, d = 0.91$)	0.50882 ($k = 1.61, d = 0.51$)

Figure 1: The minimum SMSE of the estimators with correlation (ρ)

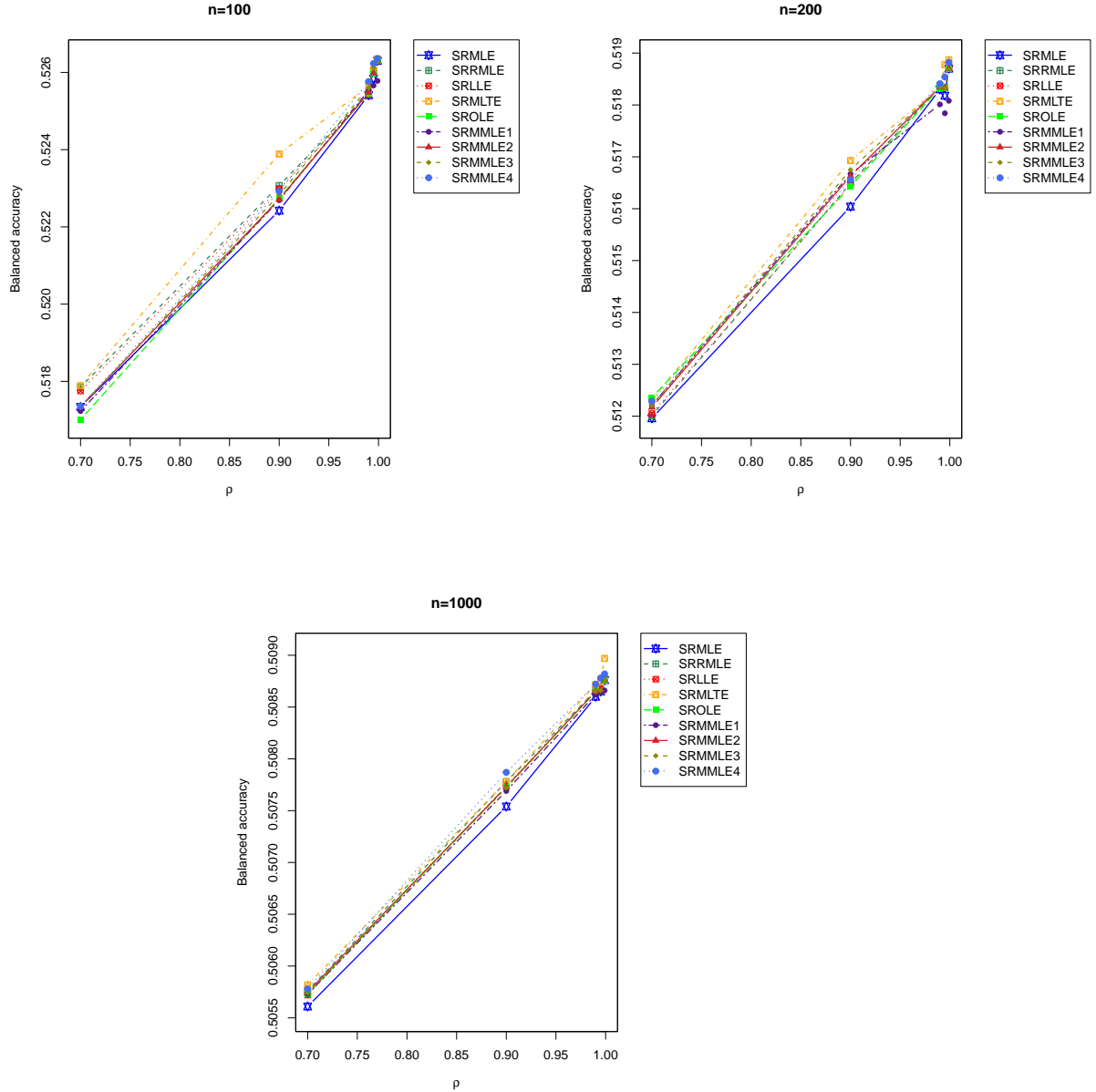


Figure 2: The maximum balanced accuracy of the estimators when $n=100$, 200 and 1000

Figure 2 illustrates that the balanced accuracy of each estimator improves with increasing correlation (ρ) but declines as the sample size (n) increases.

Empirical study

To check the prediction performance of the existing stochastic restricted estimators, we generate one data set using equations (2) and (48) when $n = 600$, $p = 4$ and $\rho = 0.998$. We used `set.seed(67)` in R Program.

Since unknown real values of the parameter vector β , we used the maximum likelihood estimator of β when calculating matrix C . The eigenvalues of the matrix C are computed as $0.78290, 0.00082, 0.00078, 0.00073$. To test the multicollinearity among the explanatory variables, we use condition index (CI), computed as $CI = \sqrt{\frac{\max(\lambda_j)}{\min(\lambda_j)}} = 32.74852$. There was severe multicollinearity when the condition index exceeded 30. Thus, the result provides evidence of severe multicollinearity among the explanatory variables.

We applied the same restriction as in the simulation study, and the shrinkage parameters were

obtained to minimize the SMSE and maximize the average cross-validated balanced accuracy. The results of an empirical application are presented in Table 6.

Table 6: The results of an empirical study

Estimators	SMSE	Balanced Accuracy
SRMLE	7.03624	0.52897
SRLLE	3.33357 (d=0.01)	0.52897 (d=0.01)
SROLE	3.30318	0.52897
SRRMLE	1.62500 (k=4.91)	0.52897 (k=0.01)
SRMLTE	0.98502 (k=0.81, d=0.99)	0.52940 (k=0.11, d=0.11)
SRMMLE1	3.33030	0.52897
SRMMLE2	1.29040 (k=4.91)	0.52897 (k=0.01)
SRMMLE3	1.98447 (d=0.01)	0.52897 (d=0.01)
SRMMLE4	0.10068 (k=0.01, d=0.99)	0.53230 (k=0.31, d=0.71)

From Table 6, we notice that our proposed estimator, SRMMLE4, shows superior performance. Next, SRMLTE demonstrates better performance than the other estimators in both SMSE and balanced accuracy. Moreover, we can observe the consequences associated with the outcomes of the simulation study.

4. Real data application

Myopia data was used to check the performance of the proposed estimators SRMMLE1, SRMMLE2, SRMMLE3, and SRMMLE4 with existing estimators, such as SRMLE, SRRMLE, SRLLE, SRMLTE, and SRGLE. This dataset is about a study of myopia taken from [Hosmer Jr, Lemeshow, and Sturdivant \(2013\)](#) and also studied by [Asar, Arashi, and Wu \(2017\)](#). In this data, 618 subjects who were not myopic when they entered the study were followed up for at least five years, and observations were made on 17 parameters. However, following [Asar et al. \(2017\)](#), we focused our analysis on four explanatory variables: spherical equivalent refraction (SPHEQ), axial length (AL), anterior chamber depth (ACD), and vitreous chamber depth (VCD). These variables are continuous and measured on the same scale (mm). We limited our analysis to the first 300 observations. The dependent variable indicates whether a subject has myopia (coded as 1) or not (coded as 0). In this dataset, the dependent variable consists of 263 cases of '0' and 37 cases of '1', indicating an imbalanced dataset.

The imbalance in our dataset, with the dominance of class 0, poses challenges for classification models, leading to biased predictions and poor identification of the minority class. This imbalance can also result in misleading performance metrics, such as an artificially high accuracy that fails to reflect the model's true predictive ability. Furthermore, insufficient representation of the minority class may hinder the model's ability to generalize effectively. To mitigate these issues, we use performance measurement adjustments that emphasize precision, recall, and F1-score to improve the evaluation of minority class predictions. Additionally, we utilize ROC-AUC which provides more robust estimates in imbalanced data scenarios.

The condition number, a measure of multicollinearity, is obtained as 10.4571. This result provides evidence of moderate multicollinearity among the predictor variables in the dataset. We examined the predictor variables with positive and negative correlations to one another and discovered that the correlation between the variables AL and VCD is very strong (0.9402). To evaluate the classification metrics of the estimators, we split the dataset into two so that

seventy percent of the data belongs to the training set and thirty percent of the data belongs to the test set. We trained the model using the training set and then evaluated the classification metrics using the testing set. Additionally, we applied the same restriction as in the simulation study, as referenced in Equation (49). Furthermore, during the model fitting process, we used shrinkage parameters optimized to minimize the SMSE in the empirical study discussed above. The results of the real data application are summarized in Table (7).

Table 7: The results of real data application

Estimators	SMSE	Sensitivity/Recall	Specificity	Balanced Accuracy	Precision	F1 Score	AUC value
SRMLE	8.1319	0.9512	0.1429	0.5470	0.4815	0.6393	0.6351
SRMLE	1.7785	0.8605	0.0638	0.4622	0.4568	0.5968	0.6063
SROLE	2.0741	0.8837	0.0851	0.4844	0.4691	0.6129	0.4650
SRRMLE	1.2031	0.8667	0.0667	0.4667	0.4815	0.6191	0.6241
SRMLTE	1.4127	0.8043	0.0000	0.4022	0.4568	0.5827	0.8957
SRMMLE1	3.4610	0.8723	0.0698	0.4711	0.5062	0.6406	0.6680
SRMMLE2	1.1077	0.8723	0.0698	0.4711	0.5062	0.6406	0.6680
SRMMLE3	1.4661	0.8723	0.0698	0.4711	0.5062	0.6406	0.6680
SRMMLE4	0.8392	0.9302	0.1277	0.5289	0.4938	0.6452	0.6680

From Table 7, it is evident that SRMLTE exhibits the highest AUC value (0.8957), indicating superior overall discriminative ability. Additionally, SRMMLE4 demonstrates superior overall performance by achieving the lowest SMSE, the highest F1 Score, and a slightly better-balanced accuracy.

Next, we validate the theoretical condition under Theorem 3.2 using the Myopia dataset. The corresponding results are presented in Table 8.

Based on Table 8, our proposed estimators, SRMMLE1 and SRMMLE3, outperform both

Table 8: Validation of the theoretical conditions for the myopia data

Existing Estimators	Proposed Estimators	Condition $\lambda_{max}[AJ_{(i)}CJ_{(i)}'A(AL_{(i)}CL_{(i)}'A)^{-1}]$	Decision
SRMLE $L_{(i)} = I$	SRMMLE1 $J_{(i)} = \hat{L}$	0.10772 < 1	SRMMLE1, SRMMLE2, SRMMLE3, and SRMMLE4 are superior than SRMLE.
	SRMMLE2 $J_{(i)} = \hat{L}Z_k$	0.00089 < 1	
	SRMMLE3 $J_{(i)} = \hat{L}Z_d$	0.01215 < 1	
	SRMMLE4 $J_{(i)} = \hat{L}Z_{kd}$	0.00092 < 1	
SRRMLE $L_{(i)} = Z_k$	SRMMLE1 $J_{(i)} = \hat{L}$	13.31009 > 1	SRMMLE2 and SRMMLE4 are superior than SRRMLE.
	SRMMLE2 $J_{(i)} = \hat{L}Z_k$	0.11000 < 1	
	SRMMLE3 $J_{(i)} = \hat{L}Z_d$	1.50198 > 1	
	SRMMLE4 $J_{(i)} = \hat{L}Z_{kd}$	0.11400 < 1	
SRMLE $L_{(i)} = Z_d$	SRMMLE1 $J_{(i)} = \hat{L}$	0.96512 < 1	SRMMLE1, SRMMLE2, SRMMLE3, and SRMMLE4 are superior than SRMLE.
	SRMMLE2 $J_{(i)} = \hat{L}Z_k$	0.00798 < 1	
	SRMMLE3 $J_{(i)} = \hat{L}Z_d$	0.10891 < 1	
	SRMMLE4 $J_{(i)} = \hat{L}Z_{kd}$	0.00827 < 1	
SRMLTE $L_{(i)} = Z_{kd}$	SRMMLE1 $J_{(i)} = \hat{L}$	12.63360 > 1	SRMMLE2 and SRMMLE4 are superior than SRMLTE.
	SRMMLE2 $J_{(i)} = \hat{L}Z_k$	0.10441 < 1	
	SRMMLE3 $J_{(i)} = \hat{L}Z_d$	1.42564 > 1	
	SRMMLE4 $J_{(i)} = \hat{L}Z_{kd}$	0.10821 < 1	

SRMLE and SRLLE, whereas SRMMLE2 and SRMMLE4 demonstrate superiority over SRMLE, SRRMLE, SRLLE, and SRMLTE. These results align with the simulation study findings in Tables 2 - 5, except for SRMMLE1 versus SRLLE and SRMMLE2 versus SRMLTE. When compared to the real data application results in Table 7, the validation of theoretical conditions holds, except for SRMMLE1 versus SRLLE.

The AUC-ROC of the myopia application is presented in the Figure 3.

Figure 3 shows that the SROLE performs poorly, as indicated by an AUC of 0.465, which

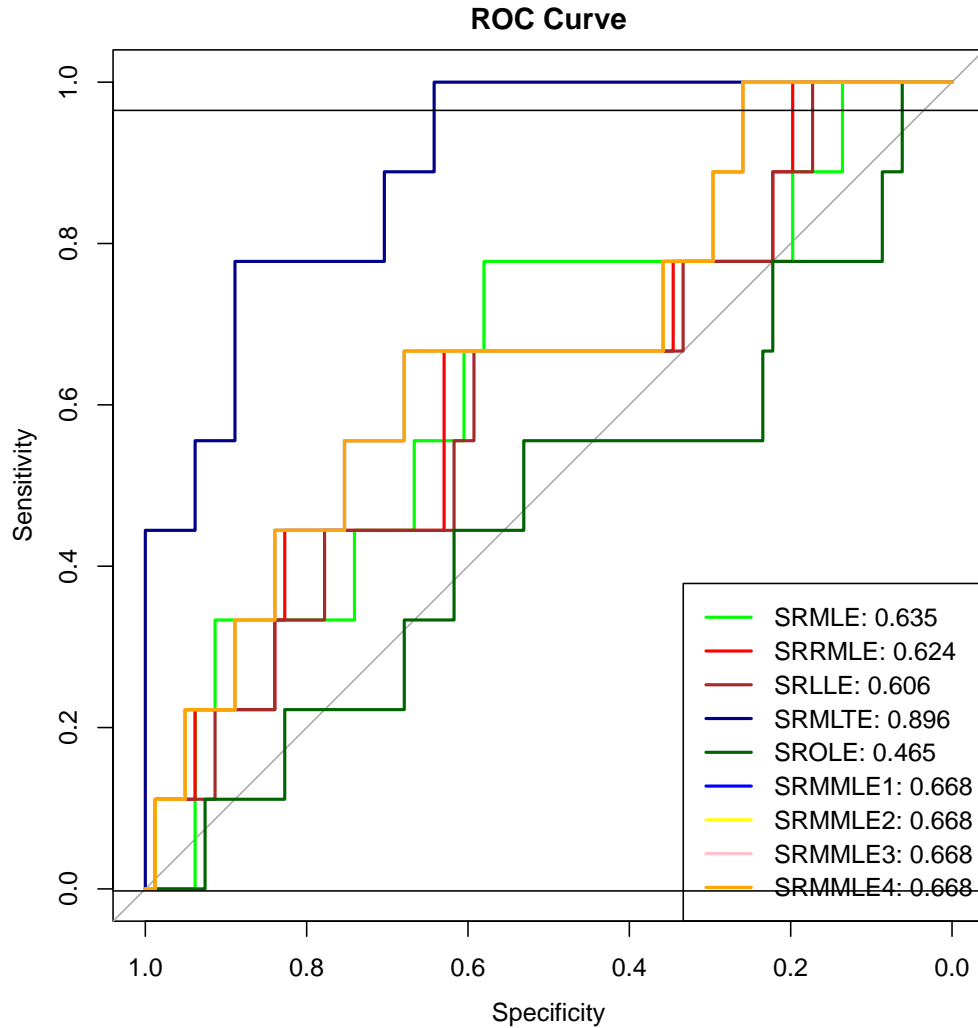


Figure 3: The ROC curve of real data application

is worse than random guessing ($AUC = 0.50$). This suggests that the model may be misclassifying a significant number of instances. On the other hand, the SRMLE, SRLLE, and SRRMLE models, with AUC values ranging from 0.60 to 0.64, demonstrate only moderate effectiveness. These models perform slightly better than random guessing but do not exhibit strong discriminative power. The SRMMLE1, SRMMLE2, SRMMLE3, and SRMMLE4 models each have same pattern with an AUC of 0.668, indicating they are better classifiers than those in the moderate range. Finally, the SRMLTE model stands out with an AUC of 0.896, reflecting a strong ability to distinguish between positive and negative classes.

5. Concluding remarks

In this study, we introduced the Stochastic Restricted Modified Mixed Logistic Estimator (SRMMLE) for the logistic regression models with stochastic linear restrictions. By modifying its coefficients, we derived four variants: SRMMLE1, SRMMLE2, SRMMLE3, and SRMMLE4. Our findings demonstrate that SRMMLE4 achieved superior performance in both SMSE and predictive accuracy compared to other estimators. Additionally, SRMLTE also exhibited better performance in both SMSE and prediction. Notably, SRMMLE4 and SRMLTE are both based on the $Z_{(k,d)}$ function, suggesting that the inclusion of the two parameters, k and d , contributes to improved results. Furthermore, SRMMLE2 and SRMMLE3 outperformed the existing estimators SRMLE, SRRMLE, SRLLE, and SROLE in terms of SMSE. However, they did not exhibit superior performance in prediction. Therefore, when comparing logistic regression estimators, it is essential to use classification metrics.

References

- Alheety M, Månsson K, Kibria BMG (2021). “A New Kind of Stochastic Restricted Biased Estimator for Logistic Regression Model.” *Journal of Applied Statistics*, **48**(9), 1559–1578. doi:[10.1080/02664763.2020.1769576](https://doi.org/10.1080/02664763.2020.1769576).
- Arashi M, Kibria BMG, Valizadeh T (2017). “On Ridge Parameter Estimators under Stochastic Subspace Hypothesis.” *Journal of Statistical Computation and Simulation*, **87**(5), 966–983. doi:<https://doi.org/10.1080/00949655.2016.1239104>.
- Asar Y, Arashi M, Wu J (2017). “Restricted Ridge Estimator in the Logistic Regression Model.” *Communications in Statistics-Simulation and Computation*, **46**(8), 6538–6544. doi:<https://doi.org/10.1080/03610918.2016.1206932>.
- Hosmer Jr DW, Lemeshow S, Sturdivant RX (2013). *Applied Logistic Regression*. John Wiley & Sons. doi:[10.1002/9781118548387](https://doi.org/10.1002/9781118548387).
- Inan D, Erdogan BE (2013). “Liu-type Logistic Estimator.” *Communications in Statistics-Simulation and Computation*, **42**(7), 1578–1586. doi:[10.1080/03610918.2012.667480](https://doi.org/10.1080/03610918.2012.667480).
- Li Y, Asar Y, Wu J (2020). “On the Stochastic Restricted Liu Estimator in Logistic Regression Model.” *Journal of Statistical Computation and Simulation*, **90**(15), 2766–2788. doi:<https://doi.org/10.1080/00949655.2020.1790561>.
- Månsson K, Kibria BMG, Shukur G (2012). “On Liu Estimators for the Logit Regression Model.” *Economic Modelling*, **29**(4), 1483–1488.
- McDonald GC, Galarneau DI (1975). “A Monte Carlo Evaluation of Some Ridge-type Estimators.” *Journal of the American Statistical Association*, **70**(350), 407–416. doi:[10.1080/01621459.1975.10479882](https://doi.org/10.1080/01621459.1975.10479882).
- Nagarajah V, Wijekoon P (2015). “Stochastic Restricted Maximum Likelihood Estimator in Logistic Regression Model.” *Open Journal of Statistics*, **5**(7), 837. doi:[10.4236/ojs.2015.57082](https://doi.org/10.4236/ojs.2015.57082).
- Rao CR, Shalabh, Toutenburg H, Heumann C (2008). *Linear Model and Generalizations*. Springer, Berlin.
- Rao CR, Statistiker M (1973). *Linear Statistical Inference and Its Applications*, volume 2. Wiley New York. doi:[10.1002/9780470316436](https://doi.org/10.1002/9780470316436).
- Rao CR, Toutenburg H (1995). *Linear Models: Least Squares and Alternatives, Second Edition*. Springer. doi:https://doi.org/10.1007/978-1-4899-0024-1_2.

- Schaefer RL, Roi LD, Wolfe RA (1984). “A Ridge Logistic Estimator.” *Communications in Statistics-Theory and Methods*, **13**(1), 99–113. doi:[10.1080/03610928408828664](https://doi.org/10.1080/03610928408828664).
- Trenkler G, Toutenburg H (1990). “Mean Squared Error Matrix Comparisons between Biased Estimators—An Overview of Recent Results.” *Statistical Papers*, **31**(1), 165–179. doi:[10.1007/BF02924687](https://doi.org/10.1007/BF02924687).
- Varathan N, Wijekoon P (2016). “Ridge Estimator in Logistic Regression under Stochastic Linear Restrictions.” *British Journal of Mathematics & Computer Science*, **15**(3), 1–14. doi:[10.9734/BJMCS/2016/24585](https://doi.org/10.9734/BJMCS/2016/24585).
- Varathan N, Wijekoon P (2017a). “More on the Restricted Almost Unbiased Liu-estimator in Logistic Regression.” *arXiv preprint arXiv:1711.10156*. doi:<https://doi.org/10.48550/arXiv.1711.10156>.
- Varathan N, Wijekoon P (2017b). “A Stochastic Restricted Almost Unbiased Ridge Estimator in Logistic Regression.” In *Proceedings of the Annual Conference on iPURSE*. University of Peradeniya, Sri Lanka.
- Varathan N, Wijekoon P (2018). “Liu-Type Logistic Estimator under Stochastic Linear Restrictions.” *Ceylon Journal of Science*, **47**(1), 21–34. doi:[10.4038/cjs.v47i1.7483](https://doi.org/10.4038/cjs.v47i1.7483).
- Varathan N, Wijekoon P (2019a). *Improvement of Ridge Type Estimators in Logistic Regression*. Ph.D. thesis, Postgraduate Institute of Science, University of Peradeniya, Sri Lanka.
- Varathan N, Wijekoon P (2019b). “Logistic Liu Estimator under Stochastic Linear Restrictions.” *Statistical Papers*, **60**, 945–962. doi:[10.1007/s00362-016-0856-6](https://doi.org/10.1007/s00362-016-0856-6).
- Varathan N, Wijekoon P (2021). “Optimal Stochastic Restricted Logistic Estimator.” *Statistical Papers*, **62**(2), 985–1002. doi:[10.1007/s00362-019-01121-y](https://doi.org/10.1007/s00362-019-01121-y).
- Yehia EG (2020). “A Stochastic Restricted Mixed Liu-Type Estimator in Logistic Regression Model.” *Applied Mathematical Sciences*, **14**(7), 311–322.

Affiliation:

Thayaparan Kayathiri
 Postgraduate Institute of Science,
 University of Peradeniya,
 Peradeniya, Sri Lanka
 E-mail: skayathiri1994@gmail.com