USC RANKING

A New Estimator to Combat Multicollinearity in Logistic Regression Model

Prof. Dr. Monira Ahmed Hussein
Professor of Statistics at Department
of Insurance, Statistics and Mathematics
Faculty of Commerce,
University of Sadat City

Mostafa Kamal Abd El-Rahman
Lecturer Assistant at Department
of Insurance, Statistics & Mathematics
Faculty of Commerce,
University of Sadat City

المجلة العلمية للدراسات والبحوث المالية والإدارية
كلية التجارة - جامعة مدينة السادات
المجلد السادس عشر - العدد الأول – مارس ٢٠٢٤

التوثيق المقترح وفقًا لنظام APA:

Hussein, M. A. & Abd El-Rahman, M. K. (2024). A New Estimator to Combat Multicollinearity in Logistic Regression Model, المجلة العلمية للدراسات والبحوث المالية والإدارية، كلية التجارة، جامعة مدينة السادات، المجلد السادس عشر، العدد الأول، ٥٢٠-٥٥١.

رابط المجلة : https://masf.journals.ekb.eg

Abstract

This paper proposes a new estimator based on the singular value decomposition technique of the design matrix to remedy multicollinearity in the binary logistic model. The proposed estimator is called the SVD-based maximum likelihood logistic estimator. The theoretical properties of this estimator and its superiority over some existing estimators is derived in the sense of the matrix mean squared error criterion. The choice of scalar parameter for this estimator is discussed. A Monte Carlo simulation study has been conducted to compare the performance of the proposed estimator with the existing maximum likelihood estimator and ridge logistic estimator in terms of the mean squared error criterion. Moreover, a real data application is presented to illustrate the potential benefits of the proposed estimator and satisfy the theoretical findings. The results from the simulation study and the empirical application reveal that the proposed estimator works well and outperforms existing estimators in scalar mean squared error sense.

Keywords: Logistic regression, Maximum Likelihood, Multicollinearity, Ridge estimator, Singular value decomposition.

1. Introduction

Logistic regression is an appropriate statistical method to model binary or dichotomous data when the response variable has two categories either success or failure. The explanatory variables in logistic regression may take any type of variable whether continuous, discrete, ordinal, nominal or any mixture of these variables. The binary logistic regression model has wide applications in biostatistics, economics, finance, social sciences, medical sciences, machine learning, classification problem, and many other binary data fields.

The most common and frequently used method to estimate the parameters in logistic regression is the maximum likelihood estimation method (MLE). The ML estimator (β^_ML) can be obtained by using numerical iterative algorithms such as iteratively re-weighted least squares (IRLS) by Newton–Raphson algorithm, which is an asymptotically unbiased estimate of β.

One of the major assumptions for binary logistic regression is there should be no high correlations or multicollinearity among the explanatory variables of the regression model. However, in applied research, there is often the problem of multicollinearity that is due to the existence of nearly linear dependency among the explanatory variables.

The existence of multicollinearity leads to unstable parameters of the ML estimator for the logistic regression model. Moreover, the variance and asymptotic mean squared error (MSE) of the regression coefficients may be inflated. Consequently, the inference and conclusions about the model parameters based on the ML estimator may not be responsible.

To overcome the multicollinearity problem in the logistic regression model, many estimators have been introduced in the literature as alternatives to the ML estimator. Firstly, Schaefer et al. (1984) proposed the ridge logistic estimator (RLE) to handle the multicollinearity problem in the logistic regression model. They suggested a biasing parameter (k) added to the diagonal elements of the information matrix in the ML estimator to reduce the effect of multicollinearity.

In addition, there are many studies that focused on the ridge logistic estimator (RLE), such as; Kibria et al. (2012) evaluated some biasing ridge parameters (k), Nja et al. (2013) introduced the modified logistic ridge regression estimator (MLRE), Wu and Asar (2016) suggested the almost unbiased ridge logistic estimator (AURLE), Asar and Genc (2017) proposed the two-parameter ridge estimator in logistic regression. Varathan and Wijekoon (2017) introduced an optimal generalized logistic estimator based on quasi-likelihood (QL) estimation, Jadhav (2020) proposed the linearized ridge logistic estimator (LRLE), Lukman et al. (2020) introduced the modified ridge type logistic estimator, Varathan (2022) proposed a modified almost unbiased ridge logistic estimator. Abonazel et al. (2023) proposed the probit modified ridge and probit Dawoud −Kibria estimators for the probit regression model.

Also, other studies interested in Liu estimator in the logistic regression model, for instance; Mansson et al. (2012) proposed the Liu-Estimator in logistic

regression, Inan and Erdogan (2013) suggested Liu-type estimator, Xinfeng (2015) proposed the almost unbiased Liu logistic estimator (AULLE), Varathan and Wijekoon (2019) introduced the modified almost unbiased Liu logistic estimator (MAULLE).

Recently, Roozbeh et al. (2016) introduced a biased estimator based on the decomposition technique to solve the problem of multicollinearity in linear regression models. This technique depends on decomposing the design matrix into two factors; the isometric matrix with orthonormal columns and the upper triangular matrix, such that. They suggested positive value added to small diagonal elements of the R matrix. Consequently, the estimator of based on the decomposition depends on a modified such as, which is called the-based least-squares estimator (QRLSE).

This paper aims to propose an estimator to combat multicollinearity in binary logistic regression model based on the singular value decomposition (SVD) technique. In addition, the theoretical properties of this new estimator were derived. Moreover, A Monte Carlo simulation study and an empirical application are conducted to evaluate the performance of this estimator and illustrate its benefits in a real data application.

The rest of the paper is organized as follows; In Section 2, we describe the logistic regression model and maximum likelihood estimator (MLE) with their asymptotic mean squared error (MSE). Also, the multicollinearity problem and some biased estimators dealt with this problem in the logistic regression model are illustrated. The proposed estimator is constructed in Section 3. In Section 4, we drive the asymptotic properties of the proposed estimators in the context of the bias, variance-covariance matrix, and mean squared error. In Section 5, we evaluate the superiority of the proposed estimator over the MLE and the logistic ridge estimator (LRE) in the sense of the matrix mean squared error criteria. The choice of scalar parameter is discussed in Section 6. A simulation study is carried out in order to evaluate the performance of the proposed estimator with the existing ones in Section 7. In addition, Section 8 provides a real data application to support the theoretical results. Finally, in Section 9, we present a summary and conclusions.

Logistic Regression Model and Multicollinearity

Consider the binary logistic regression model such that the response variable (y_i) is assumed to be independent of each other and coded as zero or one with $y_i \sim Bernoulli\,(\pi_i)$ such as

$$\pi_i = \frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)} = \frac{1}{1+\exp\left(-x_i'\beta\right)}, \qquad i = 1, 2, ..., n \tag{1}$$

where $x_i' = [1 \;\; x_{i1} \;\; \ldots \;\; x_{i(p+1)}]$ is the ith row of an $n \times (p+1)$ design matrix X with p explanatory variables, and $\beta = (\beta_0, \beta_1, ..., \beta_p)'$ is a vector of regression parameters.

The maximum likelihood estimation method is the most common technique to estimate the logistic regression parameters (β). Therefore, the corresponding log-likelihood function of the model (1) is given by

$$L(\beta) = \sum_{i=1}^{n} y_i \log(\pi_i) + \sum_{i=1}^{n} (1 - y_i) \log(1 - \pi_i)$$

$$= \sum_{i=1}^{n} y_i x_i' \beta - \sum_{i=1}^{n} \ln[1 + \exp(x_i' \beta)]. \qquad (2)$$

ML estimator can be obtained by maximizing the above log-likelihood function by differentiating Equation (2) with respect to the parameter vector (β) and equating the first derivative to zero, which yields

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n} (y_i - \pi_i) x_i = 0. \qquad (3)$$

Since Equation (3) is nonlinear in β, the numerical iterative algorithms techniques must be used such as iteratively re-weighted least squares algorithm (IRLS), and a numerical solution is obtained as follows

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + \left(X' W^{(t)} X\right)^{-1} X' \left(y - \hat{\pi}^{(t)}\right), \qquad (4)$$

where $\hat{\pi}^{(t)}$ is the estimated vector of π using $\hat{\beta}^{(t)}$ and $W^{(t)} = \text{diag}[\hat{\pi}^{(t)}{}_i (1 - \hat{\pi}_i^{(t)})]$.

With many iterative of Equation (4) the convergence is obtained and the maximum likelihood estimator (MLE) for the logistic regression model can be written as

$$\hat{\beta}_{ML} = \left(X' \hat{W} X\right)^{-1} X' \hat{W} \hat{z}, \qquad (5)$$

where $\hat{W} = \text{diag}[\hat{\pi}_i(1-\hat{\pi}_i)]$ and $\hat{z}_i = \log(\hat{\pi}_i) + \frac{y_i - \hat{\pi}_i}{\hat{\pi}_i(1-\hat{\pi}_i)}$. The asymptotic variance-covariance matrix of the ML estimator $(\hat{\beta}_{ML})$ is

$$Cov\left(\hat{\beta}_{ML}\right) = \left(X'\hat{W}X\right)^{-1}. \qquad (6)$$

Since ML estimator is an asymptotically unbiased estimate of β, the asymptotic matrix mean squared error (MMSE) of $\hat{\beta}_{ML}$ is

$$MMSE\left(\hat{\beta}_{ML}\right) = \left(X'\hat{W}X\right)^{-1}. \qquad (7)$$

And scaler mean squared error (SMSE) of $\hat{\beta}_{ML}$ is

$$SMSE\left(\hat{\beta}_{ML}\right) = trace\left(\left(X'\hat{W}X\right)^{-1}\right) = \sum_{j=1}^{p+1}\frac{1}{\lambda_j}, \qquad (8)$$

where λ_j is the *j*th eigenvalues of the information matrix $X'\hat{W}X$.

When the explanatory variables are highly correlated, the columns of the information matrix $(X'\hat{W}X)$ are close to being dependent and this matrix becomes a near-singular ill-conditioned matrix. It implies that some of the eigenvalues $(\lambda_j's)$ of $X'\hat{W}X$ become too small and close to zero. Thus, the mean squared error value of the regression estimate produced by the ML estimator is inflated. So, the estimates have large variances and large confidence intervals, which leads to inefficient estimates [see e.g. Mansson and Shukur (2011) & Kibria *et al.* (2012)]. As a result, in the presence of a multicollinearity problem, the logistic regression model becomes unstable and the estimation of the model parameters becomes inaccurate.

To solve the problem of multicollinearity in logistic regression, various alternative biased estimators are proposed in the literature instead of the ML estimator. Schaefer *et al.* (1984) introduced the ridge logistic estimator (RLE) as an alternative to MLE when there exists strong dependence among explanatory variables. The form of ridge logistic estimator (RLE) is defined as

$$\hat{\beta}_{RLE} = \left(X'\hat{W}X + kI\right)^{-1} X'\hat{W}X\hat{\beta}_{ML}, \qquad k > 0 \qquad (9)$$

where I is an $n \times n$ identity matrix and k is the shrinkage or biasing parameter which defined as [Schaefer *et al.* (1984) & Smith *et al.* (1991)]:

$$k_1 = \frac{1}{\hat{\beta}'_{ML}\,\hat{\beta}_{ML}},\quad k_2 = \frac{p}{\hat{\beta}'_{ML}\,\hat{\beta}_{ML}},\quad k_3 = \frac{p+1}{\hat{\beta}'_{ML}\,\hat{\beta}_{ML}}. \qquad (10)$$

The asymptotic variance-covariance matrix of $\hat{\beta}_{RLE}$ is defined as follows

$$Cov\left(\hat{\beta}_{RLE}\right) = \left(X'\hat{W}X + kI\right)^{-1}\left(X'\hat{W}X\right)\left(X'\hat{W}X + kI\right)^{-1}. \qquad (11)$$

Also, the asymptotic MMSE and SMSE of $\hat{\beta}_{RLE}$ are defined as

$$MMSE\left(\hat{\beta}_{RLE}\right) = \left(X'\hat{W}X + kI\right)^{-1}\left(X'\hat{W}X\right)\left(X'\hat{W}X + kI\right)^{-1} + \eta\eta' \qquad (12)$$

where $\eta = \left(\left(X'\hat{W}X + kI\right)^{-1}X'\hat{W}X - I\right)\beta$, and

$$SMSE\left(\hat{\beta}_{RLE}\right) = trace\left(MMSE\left(\hat{\beta}_{RLE}\right)\right) = \sum_{j=1}^{p+1}\frac{\lambda_j}{(\lambda_j + k)^2} + \sum_{j=1}^{p+1}\frac{k^2\hat{\alpha}_j^2}{(\lambda_j + k)^2}$$

, (13)

where the first term in (13) is the asymptotic variance of $\hat{\beta}_{RLE}$ and the second term is its squared bias, $\hat{\alpha} = \gamma'\hat{\beta}_{ML}$ and γ is an orthogonal matrix whose columns are the eigenvectors corresponding to the ordered eigenvalues of $X'\hat{W}X$ matrix.

Mansson *et al.* (2012) generalized a Liu estimator for the logistic regression model and called it the logistic Liu estimator (LLE). This estimator is given by

$$\hat{\beta}_{LLE} = \left(X'\hat{W}X + I\right)^{-1}\left(X'\hat{W}X + dI\right)\hat{\beta}_{ML}, \qquad (14)$$

where d is the shrinkage parameter, $0 < d < 1$.

Then, Inan and Erdogan (2013) suggested the Liu-type logistic regression estimator, which is given as

$$\hat{\beta}_{LLTE} = \left(X'\hat{W}X + kI\right)^{-1}\left(X'\hat{W}X - dI\right)\hat{\beta}_{ML}. \qquad (15)$$

Lukman *et al.* (2020) developed the logistic version of the modified ridge-type estimator in the linear regression model which is proposed by Lukman *et al.* (2019). The logistic modified ridge-type estimator is defined as

$$\hat{\beta}_{LMRT} = \left(X'\hat{W}X + k\left(1+d\right) I \right)^{-1} X'\hat{W}X\hat{\beta}_{ML}, \qquad (16)$$

where $k > 0$ and $0 < d < 1$.

Roozbeh *et al.* (2016) proposed a new biased estimator based on the QR decomposition to overcome multicollinearity in linear regression models. They used the QR decomposition technique to factorize the ill-conditional design matrix (X) into the isometric matrix Q with orthonormal columns and the upper triangular matrix R. They mentioned that, when multicollinearity occurs for matrix (X), some diagonal entries of matrix R become too small, and more closeness of the small entries values of the R matrix leads to more strength of the multicollinearity.

To overcome the multicollinearity problem, they added a positive scalars (μ) to the small diagonal entries of the upper triangular matrix (R), and the modified version of R matrix becomes $R_{(\mu)} = R + \text{diag}\,(0, ..., 0, \mu_{r+1}, ..., \mu_p)$. Consequently, the new biased estimator based on QR decomposition, which is called the QR-based least squares estimator $(QRLSE)$ for linear regression model, is defined as

$$\hat{\beta}_{(\mu)} = \left(X'_{\mu} X_{\mu} \right)^{-1} X'Y \qquad (17)$$

where $X_{\mu} = QR_{(\mu)}$ is a modified design matrix obtained form $R_{(\mu)}$.

In the following section, we developed an estimator to overcome the multicollinearity problem in the binary logistic regression model as an extension to the QR-based least-squares estimator (QRLSE) for the linear regression model, which was introduced by Roozbeh *et al.* (2016). While our proposed estimator is based on the singular value decomposition technique (SVD) to overcome multicollinearity for the binary logistic regression model.

1. Construction of the Proposed Estimator

Watkins (2002) mentioned that the singular value decomposition (*SVD*) may be the most important matrix decomposition technique of all, and a powerful tool

for both theoretical and computational purposes. First, a needful definitions and theory will be briefly presented.

Theorem 3.1. (Watkins, 2002) *Given a matrix* $A \in \mathbb{R}^{n\times m}$ *be a nonzero matrix with rank* r. *Then* A *can be factorized as a product*

$$A = U \sum V^{T},$$

where $U \in \mathbb{R}^{n\times n}$ *and* $V \in \mathbb{R}^{m\times m}$ *are orthogonal matrices, and* $\sum \in \mathbb{R}^{n\times m}$ *is a rectangular diagonal matrix.*

The diagonal elements of $\sum$ are unique non-negative values called the singular values, which are listed in descending order such as $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r \geq 0$. The matrices U and V are unitary matrices such that $U'U = V'V = I$, with orthonormal columns which are called left and right singular vectors, respectively.

In this context, the design matrix X can be factorized into the product of three matrices; left singular vectors matrix $\mathcal{U}$, right singular vectors matrix $\mathcal{V}$ which are orthonormal matrices, and rectangular diagonal matrix of singular values $\mathcal{D}$. Therefore, the singular value decomposition (SVD) of the $n \times p$ design matrix X can be expressed as follows

$$X_{n\times p} = \mathcal{U}_{n\times n}\, \mathcal{D}_{n\times p}\, \mathcal{V}'_{p\times p}, \qquad (18)$$

where $\mathcal{D}$ is diagonal matrix with entries called the uniquely singular values (δ's), which are ordered as $\delta_{11} = \delta_{\max} \geq \delta_{22} \geq ... \geq \delta_{pp} = \delta_{\min} \geq 0$, p is the number of explanatory variables that refers to the exact rank of the full column rank matrix X.

In the presence of a multicollinearity problem, the X becomes an ill-conditioned matrix, and the diagonal matrix $\mathcal{D}$ becomes having r large singular values, while the others are relatively small which perhaps close to zero [see, e.g. Kibria *et al.* (2012)]. To determine which singular values of $\mathcal{D}$ are small, we need to define a threshold or a positive constant (ω) that separates the large and small singular values [see, Watkins (2002)], such as

$$\delta_{11} \geq \delta_{22} \geq ... \geq \delta_{rr} >> \omega \geq \delta_{r+1,r+1} \geq ... \geq \delta_{pp}, \qquad (19)$$

where r is the numerical rank of ill-conditioned matrix X, which is defined as the number of singular values of X that are substantially larger than ω. Cattell (1966) introduced the scree plot that draws the singular values in a coordinate system and then r is chosen as the "large gap" or "elbow" of the graph. Furthermore, MATLAB has a "*rank"* command to compute the numerical rank of the matrix, which uses a default threshold and can also be overridden by the user.

To overcome the multicollinearity problem, we keep the large singular values $(\delta_{11} \geq \delta_{22} \geq ... \geq \delta_{rr})$ as they are because they are large enough. On the other hand, it is reasonable to increase the small singular values $(\delta_{r+1,r+1} \geq ... \geq \delta_{pp})$ of the diagonal matrix $\mathcal{D}$ by some positive quantities (τ_i) as follows:

$$\delta_{ii} \leftarrow \delta_{ii} + \tau_i, \qquad i = r+1, ..., p, \tag{20}$$

where τ_i are positive scalar parameters will be derived in Section 6. Therefore, we get a modified version of the ill-conditioned design matrix X such as

$$X_\tau = \mathcal{U}\mathcal{D}_\tau \mathcal{V}', \tag{21}$$

where $\mathcal{D}_\tau = \mathcal{D} + \text{diag}(0, ..., 0, \tau_{r+1}, ..., \tau_p)$.

According to the previous discussion, now we introduce our proposed estimator based on *SVD* for the logistic regression model. The proposed estimator is called the *SVD*-based Maximum Likelihood Logistic Estimator $\left(SVD - MLLE(\tau)\right)$ which is denoted $\left(\hat{\beta}_{SVD}^{ML}(\tau)\right)$ and can be obtained as follows

$$\hat{\beta}_{SVD}^{ML}(\tau) = \left(X_\tau' \hat{W} X_\tau\right)^{-1} X' \hat{W} X \hat{\beta}_{ML} \tag{22}$$

where τ_i are positive scalar parameters, $\hat{W} = \text{diag}[\hat{\pi}_i(1-\hat{\pi}_i)]$ which is a weight matrix estimated based on MLE. One can note that, if the positive scalar (τ_i) values equal zero, we obtain the MLE estimator.

With some algebraic calculations, the SVD-based Maximum Likelihood Logistic Estimator $\left(SVD - MLLE(\tau)\right)$ is defined as

$$\hat{\beta}_{SVD}^{ML}(\tau) = \mathcal{V}\left(\mathcal{D}_\tau' \mathcal{U}' \hat{W} \mathcal{U} \mathcal{D}_\tau\right)^{-1} \mathcal{D}' \mathcal{U}' \hat{W} \mathcal{U} \mathcal{D} \mathcal{V}' \hat{\beta}_{ML}, \tag{23}$$

As we have mentioned in Section 2, the ridge logistic estimator (RLE) suggested by Schaefer *et al.* (1984) is summarized by adding a small positive quantity (ridge parameter k) to all diagonal entries of the information matrix $(X'\hat{W}X)$. So, this approach changes all columns of the design matrix (X). Hence, this may lead to an increase in the bias of ML estimator.

On the other hand, our proposed estimator based on the SVD technique for logistic regression is summarized by adding the positive scalars (τ_i) to the last $(p-r)$ singular values only, which are too small of the diagonal matrix $\mathcal{D}$. So, this proposed estimator changes only the last columns of the design matrix (X), which are considered a noisy data. From this point of view, we hope that the $\left(\hat{\beta}_{SVD}^{ML}(\tau)\right)$ estimator may reduce the bias of the logistic regression estimation and make our estimator more efficient than other biased estimators.

2. Asymptotic properties of the proposed estimator

This section considers the statistical properties of the proposed estimator for the binary logistic regression model. We drive the bias, variance-covariance matrix, and matrix mean squares error (MMSE).

The asymptotic bias of $\hat{\beta}_{SVD}^{ML}(\tau)$ can be obtained as follows

$$Bias\left(\hat{\beta}_{SVD}^{ML}(\tau)\right)=E\left(\hat{\beta}_{SVD}^{ML}(\tau)\right)-\beta$$

$$=E\left(\mathcal{V}\left(\mathcal{D}_{\tau}'\mathcal{U}'\hat{W}\mathcal{U}\mathcal{D}_{\tau}\right)^{-1}\mathcal{D}'\mathcal{U}'\hat{W}\mathcal{U}\mathcal{D}\mathcal{V}'\hat{\beta}_{ML}\right)-\beta$$

$$=\mathcal{V}\left(\mathcal{D}_{\tau}'\mathcal{U}'\hat{W}\mathcal{U}\mathcal{D}_{\tau}\right)^{-1}\mathcal{D}'\mathcal{U}'\hat{W}\mathcal{U}\mathcal{D}\mathcal{V}'E\left(\hat{\beta}_{ML}\right)-\beta.$$

Since $\hat{\beta}_{ML}$ estimator is an asymptotically unbiased estimate of β. Therefore, the asymptotic bias of the proposed $\hat{\beta}_{SVD}^{ML}(\tau)$ estimator can be obtained as

$$Bias\left(\hat{\beta}_{SVD}^{ML}(\tau)\right)=\left(\mathcal{V}\left(\mathcal{D}_{\tau}'\mathcal{U}'\hat{W}\mathcal{U}\mathcal{D}_{\tau}\right)^{-1}\mathcal{D}'\mathcal{U}'\hat{W}\mathcal{U}\mathcal{D}\mathcal{V}'-I\right)\beta$$

$$= \left(\mathcal{V}\Omega_{\tau}^{-1}\Psi\,\mathcal{V}' - I \right)\beta = \mathcal{V}\left(\Omega_{\tau}^{-1}\Psi - I \right)\alpha . \qquad (24)$$

where $\Omega_{\tau} = \mathcal{D}_{\tau}'\mathcal{U}'\hat{W}\,\mathcal{U}\mathcal{D}_{\tau}$, $\Psi = \mathcal{D}'\mathcal{U}'\hat{W}\,\mathcal{U}\mathcal{D}$ and $\alpha = \mathcal{V}'\beta$.

In addition, the asymptotic variance-covariance matrix of $\hat{\beta}_{SVD}^{ML}(\tau)$ can be derived as

$$Cov\left(\hat{\beta}_{SVD}^{ML}(\tau) \right) = Cov\left(\mathcal{V}\left(\mathcal{D}_{\tau}'\mathcal{U}'\hat{W}\,\mathcal{U}\mathcal{D}_{\tau} \right)^{-1} \mathcal{D}'\mathcal{U}'\hat{W}\,\mathcal{U}\mathcal{D}\mathcal{V}'\hat{\beta}_{ML} \right)$$

$$= \mathcal{V}\left(\mathcal{D}_{\tau}'\mathcal{U}'\hat{W}\,\mathcal{U}\mathcal{D}_{\tau} \right)^{-1} \mathcal{D}'\mathcal{U}'\hat{W}\,\mathcal{U}\mathcal{D}\mathcal{V}' \left[Cov\left(\hat{\beta}_{ML} \right) \right] \mathcal{V}\,\mathcal{D}'\mathcal{U}'\hat{W}\mathcal{U}\mathcal{D}$$
$$\times \left(\mathcal{D}_{\tau}'\mathcal{U}'\hat{W}\,\mathcal{U}\mathcal{D}_{\tau} \right)^{-1} \mathcal{V}' .$$

Since, $Cov(\hat{\beta}_{ML}) = (X'\hat{W}X)^{-1} = (\mathcal{V}\mathcal{D}'\mathcal{U}'\hat{W}\mathcal{U}\mathcal{D}\mathcal{V}')^{-1}$, then the asymptotic variance-covariance matrix of $\hat{\beta}_{SVD}^{ML}(\tau)$ can be defined as

$$Cov\left(\hat{\beta}_{SVD}^{ML}(\tau) \right) = \mathcal{V}\left(\mathcal{D}_{\tau}'\mathcal{U}'\hat{W}\,\mathcal{U}\mathcal{D}_{\tau} \right)^{-1} \mathcal{D}'\mathcal{U}'\hat{W}\mathcal{U}\mathcal{D}\left(\mathcal{D}_{\tau}'\mathcal{U}'\hat{W}\,\mathcal{U}\mathcal{D}_{\tau} \right)^{-1} \mathcal{V}'$$

$$= \Omega_{\tau}^{-1}\Psi\,\Omega_{\tau}^{-1} . \qquad (25)$$

Consequently, the asymptotic matrix mean squared error (MMSE) can be obtained as

$$MMSE\left(\hat{\beta}_{SVD}^{ML}(\tau) \right) = Cov\left(\hat{\beta}_{SVD}^{ML}(\tau) \right) + Bias\left(\hat{\beta}_{SVD}^{ML}(\tau) \right)\left(Bias\left(\hat{\beta}_{SVD}^{ML}(\tau) \right) \right)'$$

$$= \Omega_{\tau}^{-1}\Psi\Omega_{\tau}^{-1} + \mathcal{V}\left(\Omega_{\tau}^{-1}\Psi - I \right)\alpha\alpha'\left(\Psi\Omega_{\tau}^{-1} - I \right)\mathcal{V}' . \qquad (26)$$

3. The Superiority of the Proposed Estimator.

In this section, we check the performance of the proposed estimator with existing logistic maximum likelihood estimator and ridge logistic estimator in terms of the asymptotic matrix mean square error criterion. The following lemmas are needful to perform the theoretical comparisons.

Lemma 5.1. (Rao and Toutenburg, 1996) *Let* $A: n \times n$ *and* $B: n \times n$ *be any two matrices such that* A *is a positive definite and* B *is a non- negative definite matrix. Then* $A + B$ *is nonnegative definite matrix.*

Lemma 5.2. (Wu, 2016) *Suppose that* $\mathcal{M}$ *be a positive definite matrix and* $\mathcal{K}$ *be a non-negative definite matrix, then*:

$$\mathcal{M} - \mathcal{K} \geq 0 \Leftrightarrow \lambda_{\max}\left(\mathcal{K}\mathcal{M}^{-1}\right) \leq 1.$$

3.1 The Comparison between $\widehat{\boldsymbol{\beta}}_{\boldsymbol{MLE}}$ and $\widehat{\boldsymbol{\beta}}_{\boldsymbol{SVD}}^{\boldsymbol{ML}}(\boldsymbol{\tau})$.

The asymptotic matrix means squared error (MMSE) of $\hat{\beta}_{ML}$ in Eq. (7) can also be written in the form of singular value decomposition as follows

$$MMSE\left(\hat{\beta}_{ML}\right) = \left(\mathcal{V}\mathcal{D}'\mathcal{U}'\hat{W}\mathcal{U}\mathcal{D}\mathcal{V}'\right)^{-1}$$
$$= \mathcal{V}\left(\mathcal{D}'\mathcal{U}'\hat{W}\mathcal{U}\mathcal{D}\right)^{-1}\mathcal{V}' = \Psi^{-1}. \qquad (27)$$

Theorem 5.1. *The proposed estimator* $\hat{\beta}_{SVD}^{ML}(\tau)$ *is superior to the maximum likelihood estimator,* $\hat{\beta}_{ML}$, *in the MMSE sense if and only if* $\lambda_{\max}\left(\mathcal{K}_1\mathcal{M}_1^{-1}\right) \leq 1$.

Proof.

Let $\Delta_1 = MMSE\left(\hat{\beta}_{ML}\right) - MMSE\left(\hat{\beta}_{SVD}^{ML}(\tau)\right)$

$$= \Psi^{-1} - \left[\Omega_\tau^{-1}\Psi\Omega_\tau^{-1} + \mathcal{V}\left(\Omega_\tau^{-1}\Psi - I\right)\alpha\alpha'\left(\Psi\Omega_\tau^{-1} - I\right)\mathcal{V}'\right]$$
$$= \mathcal{M}_1 - \mathcal{K}_1,$$

where $\mathcal{M}_1 = \Psi^{-1}$ and $\mathcal{K}_1 = \left[\Omega_\tau^{-1}\Psi\Omega_\tau^{-1} + \mathcal{V}\left(\Omega_\tau^{-1}\Psi - I\right)\alpha\alpha'\left(\Psi\Omega_\tau^{-1} - I\right)\mathcal{V}'\right]$. Since $\mathcal{M}_1$ and $\Omega_\tau^{-1}\Psi\Omega_\tau^{-1}$ are a positive definite matrices (*p.d.*) and $\mathcal{V}\left(\Omega_\tau^{-1}\Psi - I\right)\alpha\alpha'$ is a non-negative definite matrix, then according to Lemma 5.1., $\mathcal{K}_1$ is a positive definite matrix (*p.d.*). Also, by Lemma 5.2., if $\lambda_{\max}\left(\mathcal{K}_1\mathcal{M}_1^{-1}\right) \leq 1$, then $\mathcal{M}_1 - \mathcal{K}_1$ is a positive definite matrix, where

$\lambda_{\max}$ is the largest eigen value of $\mathcal{K}_1 \mathcal{M}_1^{-1}$, and the comparison can be concluded.

3.2 The Comparison between $\widehat{\boldsymbol{\beta}}_{\boldsymbol{RLE}}$ and $\widehat{\boldsymbol{\beta}}_{\boldsymbol{SVD}}^{\boldsymbol{ML}}(\boldsymbol{\tau})$.

The bias, asymptotic variance-covariance matrix, and asymptotic MMSE of $\hat{\beta}_{RLE}$ can also be written in the form of singular value decomposition as follows

$$Bias\left(\hat{\beta}_{RLE}\right)=\left[\left(\mathcal{V}\mathcal{D}'\mathcal{U}'\hat{W}\mathcal{U}\mathcal{D}\mathcal{V}'+kI\right)^{-1}\left(\mathcal{V}\mathcal{D}'\mathcal{U}'\hat{W}\mathcal{U}\mathcal{D}\mathcal{V}'\right)-I\right]\beta$$

$$=-k\mathcal{V}\left(\mathcal{V}\mathcal{D}'\mathcal{U}'\hat{W}\mathcal{U}\mathcal{D}+kI\right)^{-1}\alpha$$

$$=-k\mathcal{V}\left(\Psi+kI\right)^{-1}\alpha,$$

$$Cov\left(\hat{\beta}_{RLE}\right)=\left(\mathcal{V}\mathcal{D}'\mathcal{U}'\hat{W}\mathcal{U}\mathcal{D}\mathcal{V}'+kI\right)^{-1}\left(\mathcal{V}\mathcal{D}'\mathcal{U}'\hat{W}\mathcal{U}\mathcal{D}\mathcal{V}'\right)\left(\mathcal{V}\mathcal{D}'\mathcal{U}'\hat{W}\mathcal{U}\mathcal{D}\mathcal{V}'+kI\right)^{-1}$$

$$=\left[\mathcal{V}\left(\Psi+kI\right)^{-1}\Psi\left(\Psi+kI\right)^{-1}\right]\mathcal{V}', \quad \text{and}$$

$$MMSE\left(\hat{\beta}_{RLE}\right)=\mathcal{V}\left[\left(\Psi+kI\right)^{-1}\Psi\left(\Psi+kI\right)^{-1}\right]\mathcal{V}'+k^2\mathcal{V}\left(\Psi+kI\right)^{-1}\alpha\alpha'\left(\Psi'+kI\right)^{-1}\mathcal{V}'$$
$$=\mathcal{V}\left[\left(\Psi+kI\right)^{-1}\Psi\left(\Psi+kI\right)^{-1}+k^2\left(\Psi+kI\right)^{-1}\alpha\alpha'\left(\Psi'+kI\right)^{-1}\right]\mathcal{V}'. \quad (28)$$

Theorem 5.2. *The proposed estimator* $\hat{\beta}_{SVD}^{ML}(\tau)$ *is superior to the ridge logistic estimator,* $\hat{\beta}_{RLE}$, *in the MMSE sense if and only if* $\lambda_{\max}\left(\mathcal{K}_2\mathcal{M}_2^{-1}\right)\leq 1$.

Proof.

Let $\Delta_2 = MMSE\left(\hat{\beta}_{RLE}\right)-MMSE\left(\hat{\beta}_{SVD}^{ML}(\tau)\right)$

$$=\left[\mathcal{V}\left(\Psi+kI\right)^{-1}\Psi\left(\Psi+kI\right)^{-1}\mathcal{V}'+k^2\mathcal{V}\left(\Psi+kI\right)^{-1}\alpha\alpha'\left(\Psi'+kI\right)^{-1}\mathcal{V}'\right]$$
$$-\left[\Omega_\tau^{-1}\Psi\Omega_\tau^{-1}+\mathcal{V}\left(\Omega_\tau^{-1}\Psi-I\right)\alpha\alpha'\left(\Psi\Omega_\tau^{-1}-I\right)\mathcal{V}'\right]$$

$$=\left[C_k+B_kB_k'\right]-\left[C_\tau+B_\tau B_\tau'\right]\ C_\tau$$

$$=\mathcal{M}_2-\mathcal{K}_2,$$

where $\mathcal{M}_2 = C_k + B_k B_k'$, $\mathcal{K}_2 = C_\tau + B_\tau B_\tau'$, $C_k = \mathcal{V}\left[(\Psi + kI)^{-1}\Psi(\Psi + kI)^{-1}\right]\mathcal{V}'$, $B_k = -k\mathcal{V}(\Psi + kI)^{-1}\alpha$, $C_\tau = \Omega_\tau^{-1}\Psi\Omega_\tau^{-1}$ and $B_\tau = \mathcal{V}\left(\Omega_\tau^{-1}\Psi - I\right)\alpha$. Since C_k and C_τ are a positive definite matrices (*p.d.*) and B_k and B_τ are a non-negative definite matrices, then according to Lemma 5.1., $\mathcal{M}_2$ and $\mathcal{K}_2$ are a positive definite matrices (*p.d.*). Also, by Lemma 5.2., if $\lambda_{\max}\left(\mathcal{K}_2\mathcal{M}_2^{-1}\right) \leq 1$, then $\mathcal{M}_2 - \mathcal{K}_2$ is a positive definite matrix, where $\lambda_{\max}$ is the largest eigen value of $\mathcal{K}_2\mathcal{M}_2^{-1}$, and the theorem can be stated.

4. The Choice of the Scalar Parameter (τ)

Roozbeh *et al.* (2016) mentioned that there is no closed-form expression for the scalar parameter in their estimator which is based on the *QR* decomposition technique and called the *QR*-based least squares estimator for the linear regression model. Therefore, they conducted some numerical comparisons to derive the best values (regions) for the scalar parameter in the sense of minimum mean square error (MSE) by some simulation and graphical results to evaluate the performance of their proposed estimator with existing ones.

On the other hand, some authors mentioned that there is no definite rule on how to choose the ridge parameter [see, e.g. Månsson and Shukur (2011), & Kibria *et al.* (2012)]. Also, they suggested some formulas for ridge parameters and evaluated them by means of Monte Carlo simulations. Other authors generalized the ridge parameters k, which were developed for linear ridge regression, to be applicable for logistic ridge regression (LRR).

In this context, this paper considers two approaches to determine the scalar parameter (τ_i). The first one follows Roozbeh *et al.* (2016) to find the best values of the scalar parameter (τ_{opt}) numerically by computing the proposed estimator with many scalar parameter (τ_i) values (say, τ is from 1 to 10000). Then, we plot the MSEs versus τ values and consider the best value (τ_{opt}) which produces the minimum MSE value. In the same way, the optimal ridge parameter (k_{opt}) can be considered.

The second approach is to suggest some scalar parameter formulas which are evaluated by conducting a lot of simulation comparisons in order to define the best formula of scaler parameters (τ_i) in the sense of minimum MSE.

It is reasonable to believe that the construction of the best formula for the scalar parameters perhaps depends on the following factors

1. The distance between any small singular value and the next one of matrix $\mathcal{D}$

2. The ridge parameter (k), which is somewhat like the scalar parameters (τ_i). Since ridge parameter (k) is added to all diagonal elements of the matrix $X'\hat{W}X$ while the scalar parameters (τ_i) are added to the small singular values only of the matrix $\mathcal{D}$.
3. The eigenvalues of $X'\hat{W}X$ matrix.
4. The number of explanatory variables (p) in the logistic model.

Consequently, the suggested scalar parameters may be a function of the previous factors and defined as follows

i. $$\tau_i(1) = \left(\frac{2(p+1)}{\lambda_{\min}}\right) \times \max\left(k, \frac{\delta_{r+i-1,r+i-1} - \delta_{r+i,r+i}}{\delta_{r+i,r+i}}\right),$$
$$i = 1, \ldots, (p-r) \quad (29)$$

ii. $$\tau_i(2) = \max\left(k, \frac{\delta_1 - \delta_{r+i,r+i}}{\delta_{r+i,r+i}}\right), \qquad i = 1, \ldots, (p-r)$$
(30)

iii. $$\tau_i(3) = \left(\frac{p^2}{\delta_{r+i,r+i}}\right) \times \max\left(k, \frac{\delta_{r+i-1,r+i-1} - \delta_{r+i,r+i}}{\delta_{r+i,r+i}}\right),$$
$$i = 1, \ldots, (p-r) \quad (31)$$

where p is the number of explanatory variables, $\lambda_{\min}$ is the minimum eigenvalue of $X'\hat{W}X$ matrix, k is ridge parameter and δ,s are singular values of the matrix $\mathcal{D}$.

Inan and Erdogan (2013) mentioned that if the condition index is 15, multicollinearity is a concern; if it is greater than 30, multicollinearity is a very serious concern. In this context, we suggest using the condition index of the $X'\hat{W}X$ matrix to determine the small singular values in the $\mathcal{D}$ matrix which are increased by τ_i,s values. Such that the small singular values are those values whose square condition index is greater than 15 which suffer from a concern multicollinearity problem.

We evaluate the performance of the proposed estimator at different values of these suggested scalar parameters $(\tau_i's)$ by a Monte Carlo simulation in the next section.

5. Simulation Study

In this section, a simulation study is conducted to compare the performance of the proposed SVD-MLLE estimator at different scalar parameter values with the existing MLE and RLE estimators in the sense of the MSE

criteria. The explanatory variables are generated using the simulation procedure presented by Kibria (2003) and Lukman *et al.* (2019) as follows

$$x_{ij} = \left(1-\rho^2\right)^{1/2} z_{ij} + \rho z_{ip}, \quad i = 1, 2, ..., n; \;\; j = 1, 2, ..., p,$$

where z_{ij} are independent pseudo-random numbers from standardized normal distribution and ρ is the correlation between the explanatory variables. The values of ρ are chosen to be 0.85, 0.90, 0.95, 0.99 and 0.999. The response variable is generated from a Bernoulli distribution with parameter π_i where

$$\pi_i = \frac{1}{1+\exp\left(-x_i'\beta\right)},$$

where β are the true parameter values chosen such that $\beta'\beta = 1$, which is a common restriction in this type of simulation study. The sample size n is taken to be 75, 100, 150, and 200. We consider different numbers of explanatory variables (p) equal to 2, 4, 6, and 8, respectively. In this paper, we choose the value of ridge parameter $k = 1/\hat{\beta}'_{ML}\hat{\beta}_{ML}$ due to Schaefer *et al* .(1984) for the ridge logistic estimator (RLE). For the proposed estimator, we consider the optimal value of the scalar parameter (τ_{opt}) numerically by plotting the mean square error of SVD-MLLE estimator with many values of τ_i (say from 1 to 3000) to obtain the best value (τ_{opt}) which has the minimum MSE value. In addition, the formulas for the scalar parameter [$\tau_i(1)$, $\tau_i(2)$, and $\tau_i(3)$] defined in equations [29-31] are considered.

Then the experiment is replicated 1000 times, and the estimated MSE values of the estimators are calculated using the following equation

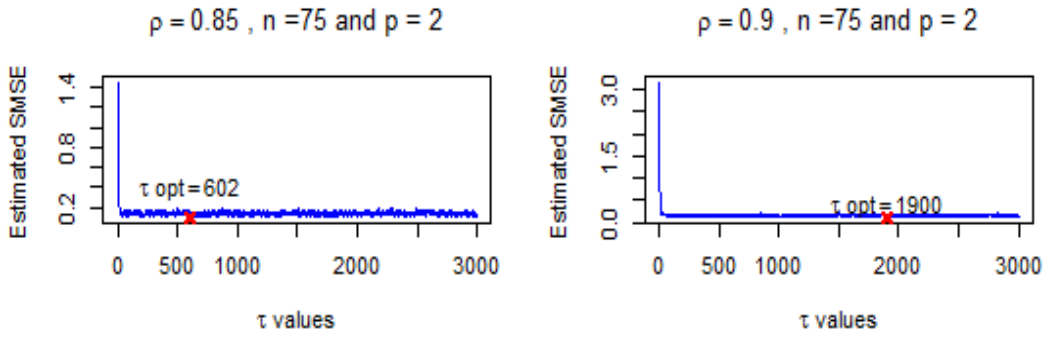
$$MSE\left(\hat{\beta}\right) = \frac{1}{1000}\sum_{j=1}^{1000}(\hat{\beta}_j - \beta)'(\hat{\beta}_j - \beta),$$

where $\hat{\beta}_j$ are the estimate parameters in the *j*th replication obtained from MLE, RLE or SVD-MLLE estimators. The MSEs of the estimators are presented for different values of ρ, n and p in Tables 7.1-7.4. Also, Figures 7.1-7.4 show the mean square error of SVD-MLLE estimator versus many values of τ_i to obtain the best scalar parameter (τ_{opt}) which has the minimum MSE value.

Table 7.1. The estimated MSE values of the estimators for different ρ when $p = 2$.

	Estimator	$\rho = 0.85$	$\rho = 0.90$	$\rho = 0.95$	$\rho = 0.99$	$\rho = 0.999$
$n = 75$	MLE	2.6625	5.6239	1.9283	7.0025	89.6949
	RLE *with* k	2.0183	4.4567	1.0730	3.4924	38.2704
	SVD-MLLE *with*	0.1308	0.0979	0.0206	0.0312	0.0403
	SVD-MLLE *with*	0.1566	0.1685	0.0428	0.0798	0.1025
	SVD-MLLE *with*	0.4657	0.6918	0.0798	0.0873	0.1040
	SVD-MLLE *with*	0.3639	0.5106	0.0513	0.0803	0.1029
$n = 100$	MLE	2.4386	1.1671	1.1286	5.2406	61.6192
	RLE *with* k	1.9670	0.9003	0.7098	2.4561	28.8561
	SVD-MLLE *with*	0.0968	0.1743	0.0077	0.0162	0.0219
	SVD-MLLE *with*	0.2285	0.2375	0.0292	0.0478	0.0536
	SVD-MLLE *with*	0.5683	0.3164	0.0831	0.0559	0.0602
	SVD-MLLE *with*	0.5913	0.3036	0.0516	0.0483	0.0591
$n = 150$	MLE	0.7500	0.4216	0.6029	2.8533	47.3767
	RLE *with* k	0.6165	0.3143	0.3884	1.5085	25.6532
	SVD-MLLE *with*	0.0699	0.0054	0.0105	0.0091	0.0136
	SVD-MLLE *with*	0.2515	0.0434	0.0339	0.0279	0.0400
	SVD-MLLE *with*	0.3047	0.0901	0.0684	0.0357	0.0413
	SVD-MLLE *with*	0.3638	0.0999	0.0557	0.0288	0.0410
$n = 200$	MLE	0.3146	0.2894	0.5007	3.2087	29.6207
	RLE *with* k	0.2583	0.2309	0.3413	1.6161	14.5658
	SVD-MLLE *with*	0.0225	0.0124	0.0122	0.0075	0.0078
	SVD-MLLE *with*	0.1027	0.0790	0.0365	0.0203	0.0264
	SVD-MLLE *with*	0.1144	0.0914	0.0734	0.0255	0.0278
	SVD-MLLE *with*	0.1378	0.1090	0.0677	0.0208	0.0274

Source: by the researcher from R outputs.



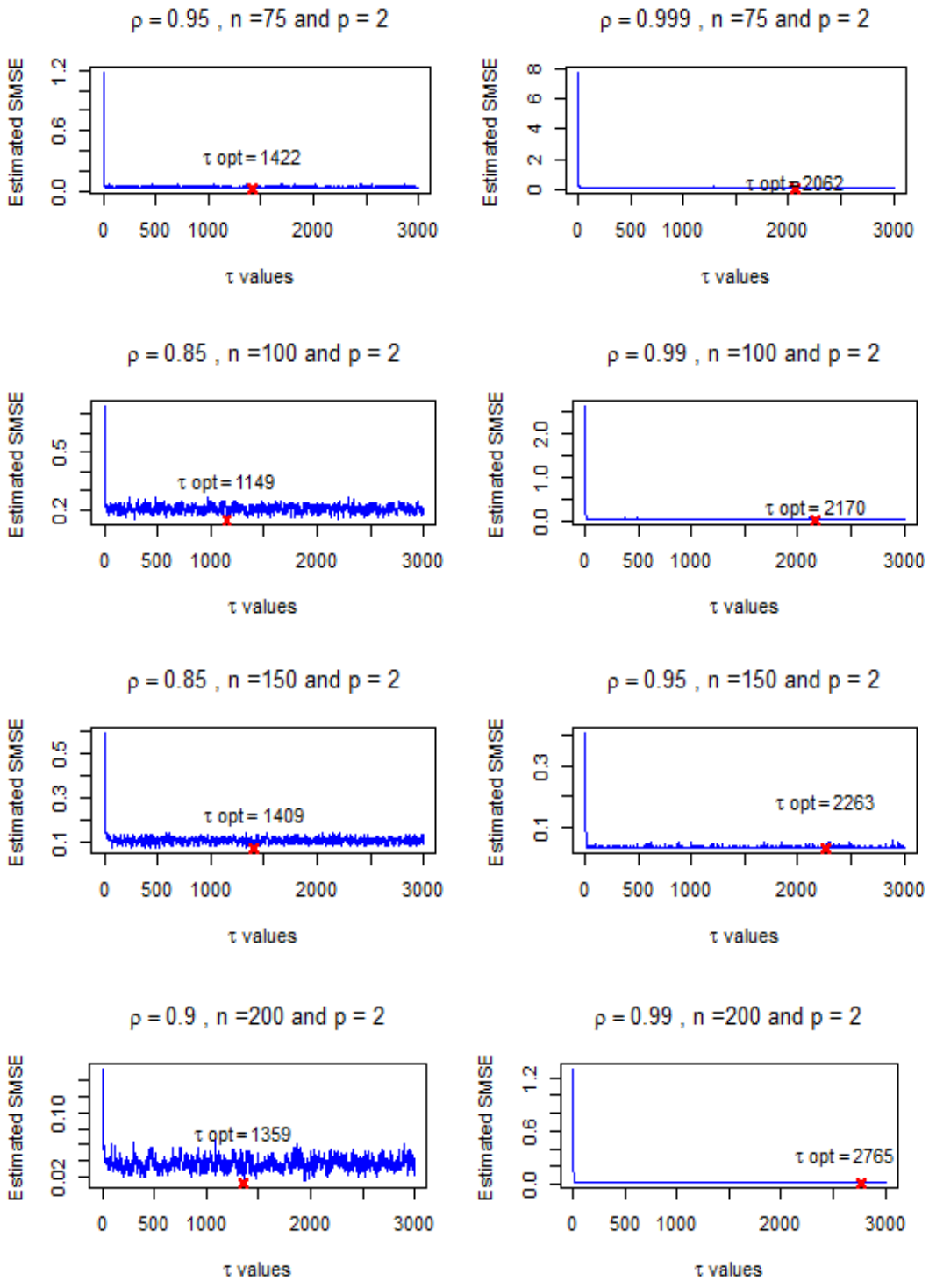


Fig. 7.1. The best value (τ_{opt}) for SVD-MLLE when $p = 2$ at some different ρ and n.

Table 7.2. The estimated MSE values of the estimators for different ρ when $p = 4$.

	Estimator	$\rho = 0.85$	$\rho = 0.90$	$\rho = 0.95$	$\rho = 0.99$	$\rho = 0.999$

$n = 75$	MLE	6.3005	4.1124	6.8434	31.5243	403.9054
	RLE *with* k	4.9356	2.8028	4.3185	19.2620	245.0803
	SVD-MLLE *with* τ_{opt}	0.3269	0.1496	0.0574	0.0511	0.0458
	SVD-MLLE *with* $\tau_i(1)$	1.2400	1.2667	1.0582	0.1485	0.1977
	SVD-MLLE *with* $\tau_i(2)$	1.3346	1.3427	1.1222	0.1654	0.2140
	SVD-MLLE *with* $\tau_i(3)$	1.5823	1.4079	1.1563	0.2818	0.2184
$n = 100$	MLE	5.0630	2.7939	3.9864	21.3965	265.1667
	RLE *with* k	4.1334	2.0643	2.6561	13.3465	153.5516
	SVD-MLLE *with* τ_{opt}	0.2633	0.0920	0.0331	0.0378	0.0288
	SVD-MLLE *with* $\tau_i(1)$	1.2666	0.5362	0.4039	0.0992	0.0708
	SVD-MLLE *with* $\tau_i(2)$	1.1650	0.4381	0.3619	0.1189	0.0743
	SVD-MLLE *with* $\tau_i(3)$	1.6551	0.7247	0.5678	0.1927	0.0730
$n = 150$	MLE	1.0822	1.4437	2.7817	13.1915	169.5922
	RLE *with* k	0.8399	1.0702	1.9009	8.1725	98.6048
	SVD-MLLE *with* τ_{opt}	0.2248	0.1023	0.0301	0.0216	0.0203
	SVD-MLLE *with* $\tau_i(1)$	0.6396	0.6026	0.3513	0.1130	0.0401
	SVD-MLLE *with* $\tau_i(2)$	0.5191	0.5486	0.2899	0.0859	0.0442
	SVD-MLLE *with* $\tau_i(3)$	0.6539	0.6365	0.5266	0.1765	0.0549
$n = 200$	MLE	0.8575	1.3178	1.7364	11.0441	112.038
	RLE *with* k	0.6980	1.0312	1.1938	6.8050	65.9296
	SVD-MLLE *with* τ_{opt}	0.1117	0.0711	0.0123	0.0097	0.0127
	SVD-MLLE *with* $\tau_i(1)$	0.5934	0.6595	0.2710	0.0974	0.0278
	SVD-MLLE *with* $\tau_i(2)$	0.4251	0.4512	0.2408	0.0535	0.0317
	SVD-MLLE *with* $\tau_i(3)$	0.5700	0.6959	0.3918	0.2064	0.0504

Table 7.3. The estimated MSE values of the estimators for different ρ when p =6.

	Estimator	$\rho = 0.85$	$\rho = 0.90$	$\rho = 0.95$	$\rho = 0.99$	$\rho = 0.999$
$n = 75$	MLE	8.4319	9.4226	17.2377	66.8203	736.7599
	RLE *with* k	6.5200	6.3327	11.8453	44.0930	455.7728
	SVD-MLLE *with* τ_{opt}	0.5876	0.5715	0.1395	0.1285	0.1629
	SVD-MLLE *with* $\tau_i(1)$	2.2249	3.0206	2.0793	0.3534	0.1414
	SVD-MLLE *with* $\tau_i(2)$	2.4594	3.1235	2.2562	0.5168	0.1548

	SVD-MLLE *with* $\tau_i(3)$	2.3804	3.0551	2.1668	0.3977	0.1453
$n = 100$	MLE	3.4076	5.4817	8.0117	47.7288	527.7479
	RLE *with* k	2.5519	3.8209	5.5933	31.1958	344.1768
	SVD-MLLE *with* τ_{opt}	0.2847	0.2039	0.1503	0.0606	0.0911
	SVD-MLLE *with* $\tau_i(1)$	1.2417	1.6254	1.4610	0.0934	0.0860
	SVD-MLLE *with* $\tau_i(2)$	1.2464	1.6612	1.4509	0.1494	0.0963
	SVD-MLLE *with* $\tau_i(3)$	1.2500	1.6722	1.4271	0.1286	0.0948
$n = 150$	MLE	1.8340	2.7918	4.9056	29.9405	288.3237
	RLE *with* k	1.4343	2.0817	3.4983	19.9363	188.2218
	SVD-MLLE *with* τ_{opt}	0.2746	0.2001	0.0565	0.0257	0.0211
	SVD-MLLE *with* $\tau_i(1)$	0.9030	1.2194	0.7964	0.2183	0.0398
	SVD-MLLE *with* $\tau_i(2)$	0.8904	1.1967	0.7342	0.1379	0.0497
	SVD-MLLE *with* $\tau_i(3)$	0.8928	1.2342	0.7727	0.1385	0.0489
$n = 200$	MLE	1.7378	1.6244	3.8956	20.3783	215.9051
	RLE *with* k	1.4455	1.2503	2.7861	13.4498	142.5207
	SVD-MLLE *with* τ_{opt}	0.2017	0.1462	0.0124	0.0175	0.0172
	SVD-MLLE *with* $\tau_i(1)$	0.9392	0.8598	0.4756	0.1121	0.0309
	SVD-MLLE *with* $\tau_i(2)$	0.8003	0.8485	0.4216	0.0992	0.0367
	SVD-MLLE *with* $\tau_i(3)$	0.8624	0.8617	0.4898	0.1158	0.0416

Table 7.4. The estimated MSE values of the estimators for different ρ when $p = 8$.

	Estimator	$\rho = 0.85$	$\rho = 0.90$	$\rho = 0.95$	$\rho = 0.99$	$\rho = 0.999$
$n = 75$	MLE	8.6822	11.2042	21.5837	135.1982	1748.015
	RLE *with* k	6.3498	8.0906	15.1695	92.2424	1172.628
	SVD-MLLE *with* τ_{opt}	1.3830	0.8932	0.7361	0.2655	0.5503
	SVD-MLLE *with* $\tau_i(1)$	3.4043	4.1649	4.0173	2.1586	1.4905
	SVD-MLLE *with* $\tau_i(2)$	3.6018	4.1577	3.9267	2.2256	1.5184
	SVD-MLLE *with* $\tau_i(3)$	3.3721	3.9640	3.5401	2.1005	1.4946
$n = 100$	MLE	4.5937	9.2370	19.0242	75.8786	653.9457
	RLE *with* k	3.5461	6.6311	13.2548	53.9368	447.9042
	SVD-MLLE *with* τ_{opt}	0.8826	0.5113	0.1073	0.2366	0.1210

	SVD-MLLE *with* $\tau_i(1)$	2.4974	3.0187	2.4530	0.2858	0.0987
	SVD-MLLE *with* $\tau_i(2)$	2.5022	3.1350	2.6286	0.4387	0.1084
	SVD-MLLE *with* $\tau_i(3)$	2.4207	2.9653	2.3818	0.2552	0.0996
$n = 150$	MLE	2.6593	3.9894	7.7975	60.4453	482.243
	RLE *with* k	2.1193	3.0281	5.8529	41.0284	328.6728
	SVD-MLLE *with* τ_{opt}	0.5375	0.3776	0.0646	0.0435	0.0463
	SVD-MLLE *with* $\tau_i(1)$	1.3709	1.8631	1.2810	0.1577	0.0590
	SVD-MLLE *with* $\tau_i(2)$	1.3508	1.8787	1.1098	0.3439	0.0714
	SVD-MLLE *with* $\tau_i(3)$	1.3184	1.8313	1.0588	0.2090	0.0642
$n = 200$	MLE	1.7522	2.964	5.8831	32.1895	423.2236
	RLE *with* k	1.4253	2.3208	4.2680	22.9358	283.9812
	SVD-MLLE *with* τ_{opt}	0.3243	0.1404	0.0382	0.0191	0.0267
	SVD-MLLE *with* $\tau_i(1)$	1.0922	1.4016	0.6682	0.1979	0.0520
	SVD-MLLE *with* $\tau_i(2)$	1.0872	1.3808	0.6597	0.1615	0.0623
	SVD-MLLE *with* $\tau_i(3)$	1.0652	1.3687	0.5776	0.1131	0.0583

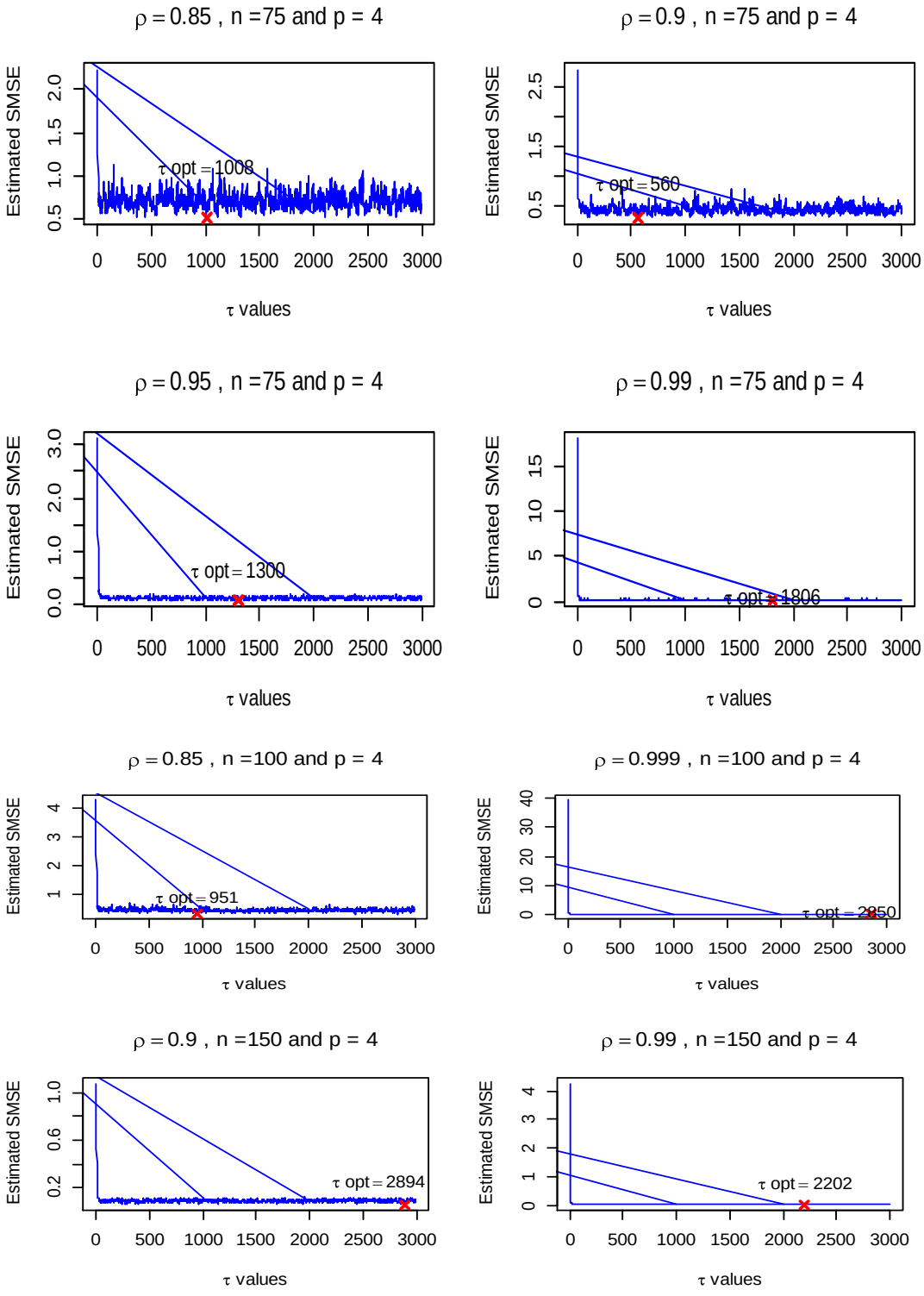
Source: by the researcher from R outputs.

From the results of Tables 7.1-7.4 it can be observed that, the proposed SVD-MLLE estimator outperforms the MLE and RLE in the mean squared error sense for all cases. Moreover, the advantage of using SVD-MLLE increases in the presence of high correlation degrees among the explanatory variables.

In general, increasing the degree of correlation among explanatory variables has a negative effect on the MLE and RLE estimators, while the proposed SVD-MLLE estimator performs very well. On the other hand, increasing the sample size has a positive impact on all estimators whether at any number of explanatory variables or correlation levels. Also, the MSEs for all estimators increase when more explanatory variables are included in the model.

To investigate the performance of the suggested scalar parameters $\tau_i's$, one can note than the best value of the scalar parameter (τ_{opt}) yields the minimum MSE in all cases as expected. In addition, with a small sample size or nearly perfect correlation between the explanatory variables ($\rho = 0.999$), the proposed estimator performs well with $\tau_i(1)$. While with a large sample size, $\tau_i(2)$ outperforms $\tau_i(1)$ and $\tau_i(3)$. Also, with a small number of explanatory variables, $\tau_i(1)$yields a minimum MSEs than $\tau_i(2)$ and $\tau_i(3)$. While, when more

explanatory variables are included in the model, $\tau_i(3)$ works well, except at $\rho = 0.999$, the $\tau_i(1)$ outperforms $\tau_i(2)$ and $\tau_i(3)$.



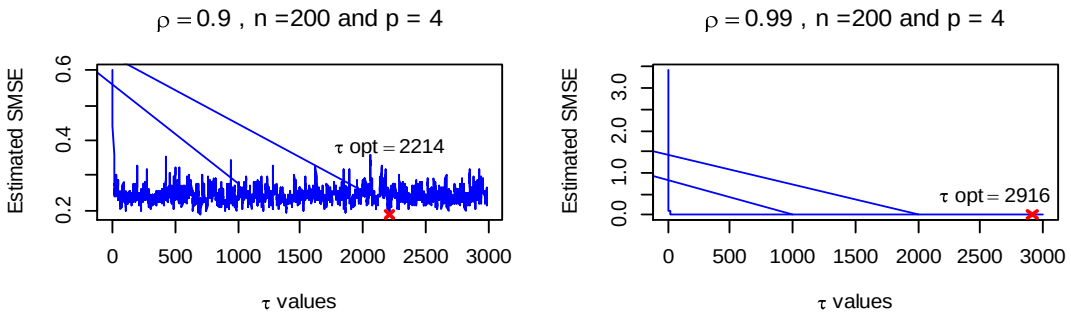
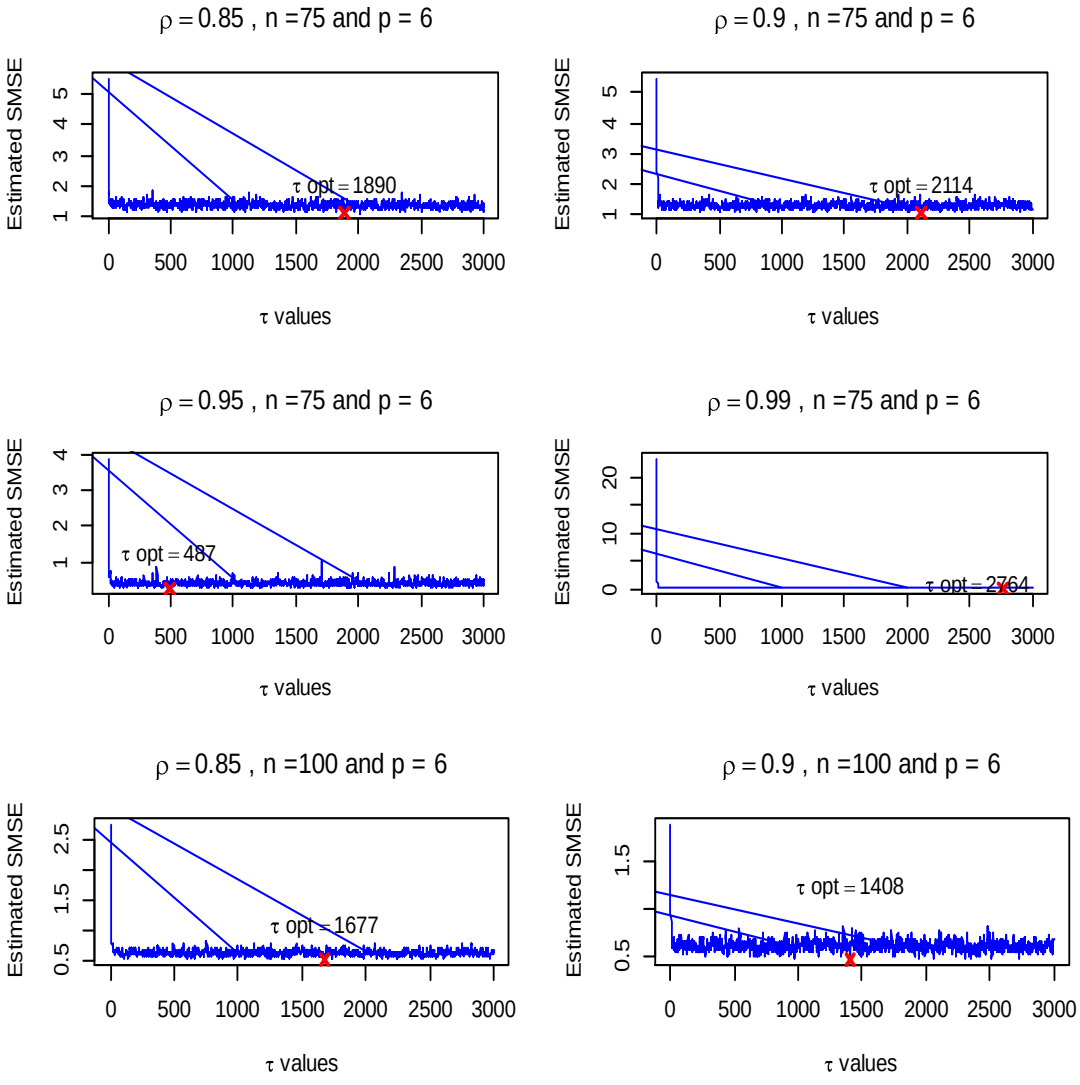


Fig. 7.2. The best value (τ_{opt}) for SVD-MLLE when $p = 4$ at some different ρ and n.



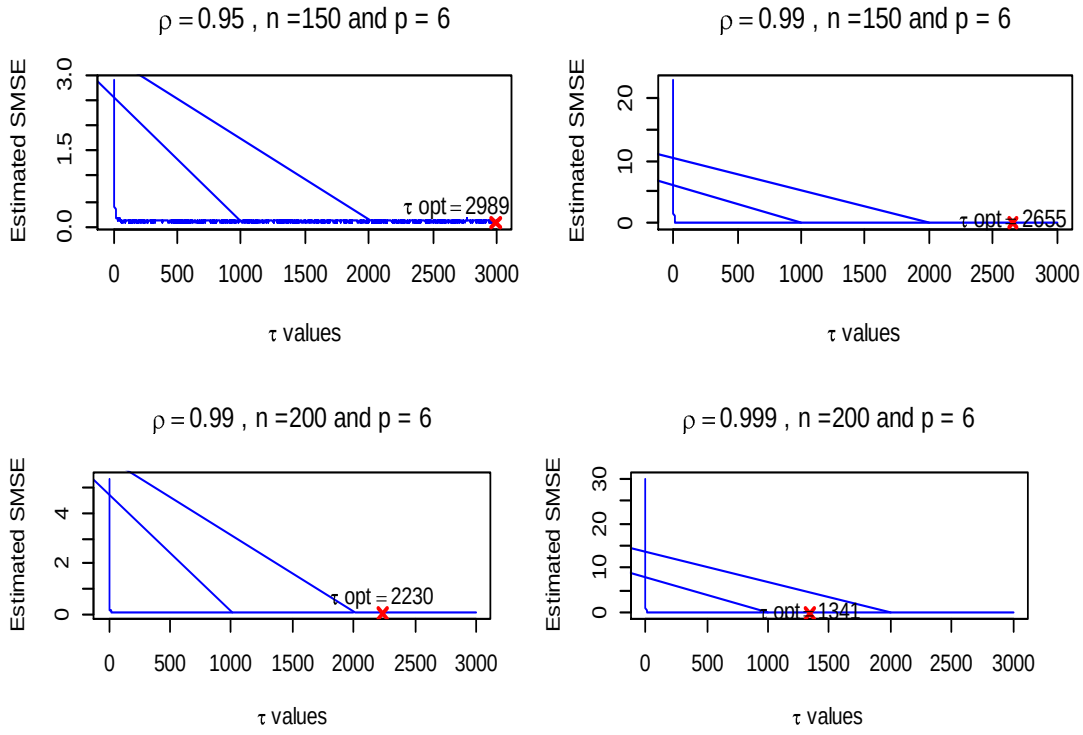
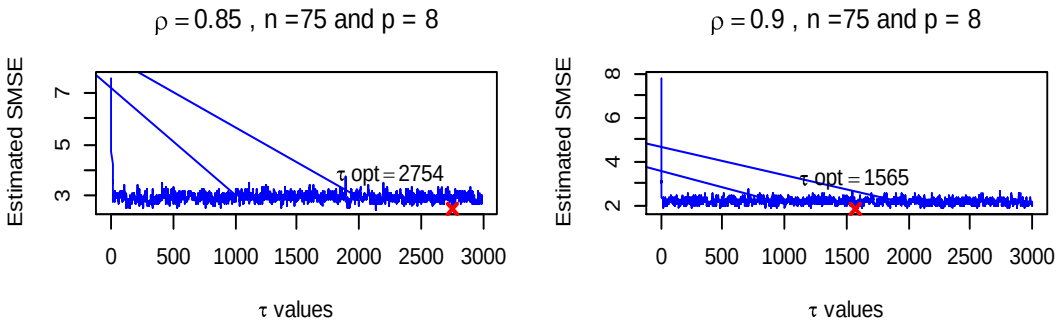


Fig. 7.3. The best value (τ_{opt}) for SVD-MLLE when $p = 6$ at some different ρ and n.



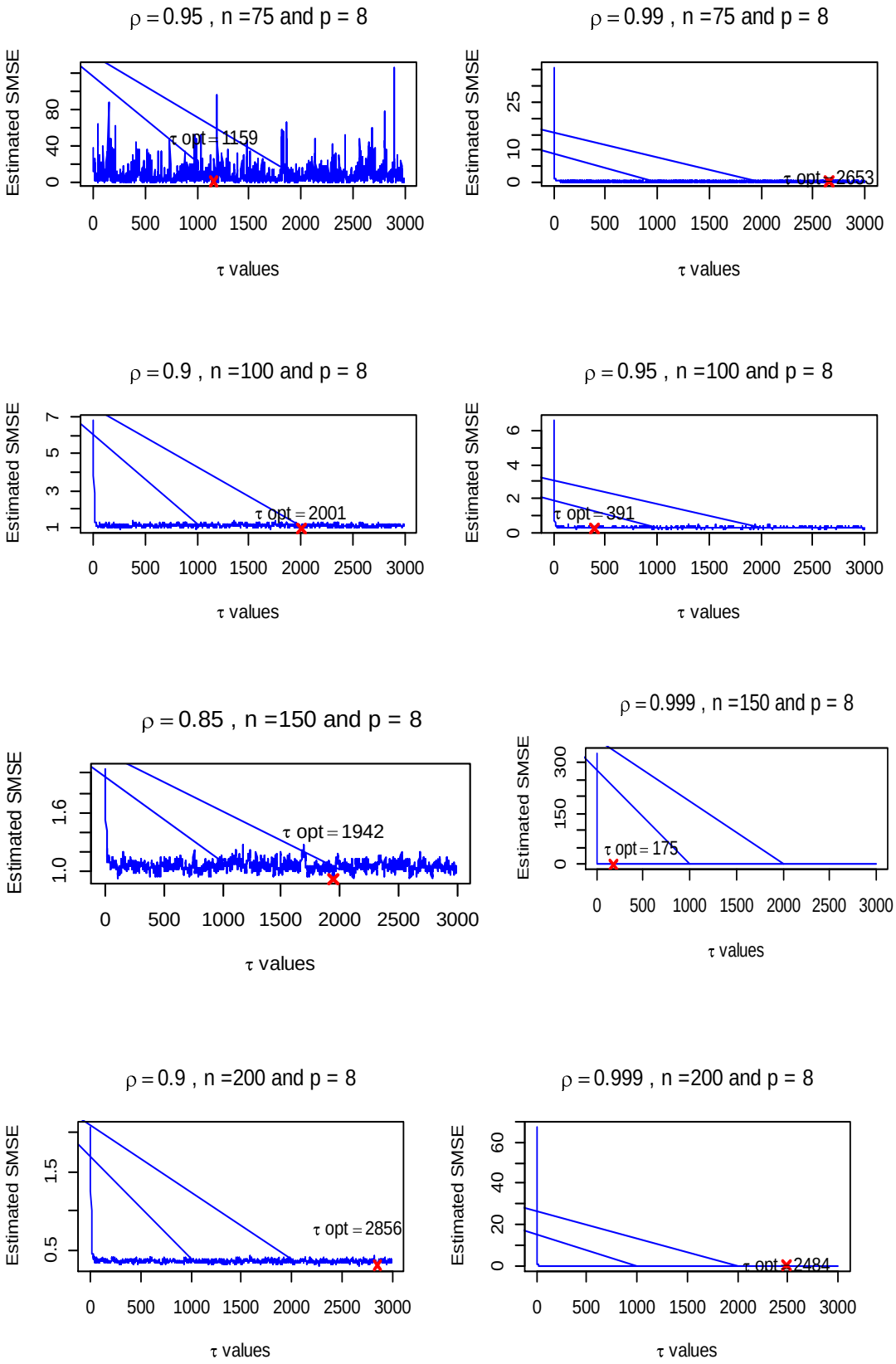


Fig. 7.4. The best value (τ_{opt}) for SVD-MLLE when $p = 8$ at some different ρ and n.

6. Empirical Application

In this section, we consider a real data application in order to evaluate the performance of our proposed SVD-MLLE estimator with the existing MLE and RLE estimators. In addition, the benefits of this new estimator in real-world fields are illustrated. Also, the results and conclusions are discussed.

The data set used in this paper is obtained from the official web page of Statistics Sweden (www.scb.se.). We will estimate a logistic regression model for the full sample consisting of 290 municipalities in Sweden for 2022. The response variable considered the net population change which is defined as follows

$$y_i = \begin{cases} 1 \text{ if there is an increase in the population in the municipali ty } i, \\ 0 \quad \text{otherwise.} \end{cases}$$

The response variable is explained by the following independent variables which are defined as

X_1: The population size,

X_2: Number of unemployed people,

X_3: Number of apartments built,

X_4: Number of bankrupt firms.

The correlation coefficients between the explanatory variables can be used as an initial step to identify the existence of a multicollinearity (Midi *et al.* (2013)). The bivariate correlations between the explanatory variables are obtained in Table 8.1.

Table 8.1. Bivariate correlation matrix for the explanatory variables.

	X_1	X_2	X_3	X_4
X_1	1.0000			
X_2	0.9562	1.0000		
X_3	0.9986	0.9521	1.0000	
X_4	0.9488	0.8899	0.9538	1.0000

Source: by the researcher through R outputs.

Table 7.1. shows that the correlation coefficients among the explanatory variables are very high, all are greater than 0.8899, and some of them are close to one.

Moreover, the condition index (CI) and condition number (κ) can be used as powerful measures to detect the degree of multicollinearity among explanatory variables which are computed as $CI = \sqrt{\lambda_{\max} / \lambda_k}$ .and $\kappa = \lambda_{\max} / \lambda_{\min}$, respectively, where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum and minimum eigenvalues of the $X'\hat{W}X$ matrix [see, e.g. Weissfeld and Sereika (1991), Lukman et al. (2020) & Awwad (2022)]. According to the literature, multicollinearity is a concern when the condition index or condition number is 15, while multicollinearity is a serious concern when they are greater than 30.

The eigenvalues, condition indices, condition number of the $X'\hat{W}X$ matrix, and eigenvalues and singular values of $X'X$ are presented in Table 8.2.

Table 8.2. Condition indices & number of $X'\widehat{W}X$, and singular values of $X'X$ matrix.

	$X'\widehat{W}X$		$X'X$	
λ	**Eigenvalue**	**Condition index**	**Eigenvalue**	**Singular values of $\mathcal{D}$**
1	75833070000	1.0000	2566675000000	1602085
٢	38171180	44.57194	1074156000	32774
٣	18657560	63.75319	421959000	20542
٤	1851.312	6400.141	278333	528
Condition Number		**40961799**		

Source: by the researcher through R outputs.

Table 8.2. clearly shows a high value of the condition number ($\kappa > 30$). Hence, this revealed that a severe multicollinearity problem exists in this data set. We consider the small singular values that are whose square of the condition index of the $X'\hat{W}X$ matrix is greater than 15. Hence, there are three of squared condition indices greater than 15, we can conclude that there are three small singular values in the $\mathcal{D}$ matrix which are increased by the positive scalars (τ_i) values to obtain the adjusted $\mathcal{D}_\tau$ and then a modified X_τ matrix.

Since there are many rules in the literature to choose the ridge parameter (k), we consider $k = 1/\hat{\beta}'_{ML}\hat{\beta}_{ML}$ for the ridge estimator and the corresponding scalar parameters $\tau_i(2)$ for the proposed SVD-MLLE estimator defined in Eq. [30].

Table 8.3 gives the regression coefficients, standard errors, and the SMSE values of MLE, RLE, and the proposed SVD-MLLE for the considered value of ridge parameter (k = 1472.964) and corresponding scalar parameters τ_{12}= 1472.964, τ_{22}= 1472.964 and τ_{32}= 3035.709 for the proposed SVD-MLLE estimator. Figure 8.1 presents the mean square errors versus different values of k & τ_i for RLE and SVDMLLE to obtain the best k_{opt} & τ_{opt}, respectively, which have the minimum MSE value.

Table 8.3. The coefficients, standard errors and SMSE values (in 10−4) of estimators.

		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**SMSE**
MLE	*Estimate*	4.6718	- 7.2661	- 8.4514	260.2773	5.4024
	Std. Error	0.8961	2.5188	1.6839	232.4086	
RLE	*Estimate*	4.6671	- 6.5666	- 8.3829	144.9327	3.0066
	Std. Error	0.8961	2.2305	1.6800	129.4299	
SVD-MLLE *with* τ_{opt}	*Estimate*	1.4323	-13.4924	-0.1245	209.6456	1.3220
	Std. Error	1.5389	5.0432	3.9643	102.4408	
SVD-MLLE *with* $\tau_i(2)$	*Estimate*	1.6186	- 14.4568	- 0.3716	196.5363	1.3954
	Std. Error	1.6709	5.4836	4.2779	98.559	

Source: by the researcher from R outputs.

According to Table 8.3, it is observed that the SVD-MLLE estimator has less SMSE than the MLE and RLE estimators. In addition, the parameters of the new estimator have fewer standard errors than all parameters of MLE and some parameters of RLE. Therefore, the results reveal that the proposed estimator works well and outperforms the MLE and LRE in the SMSE sense.

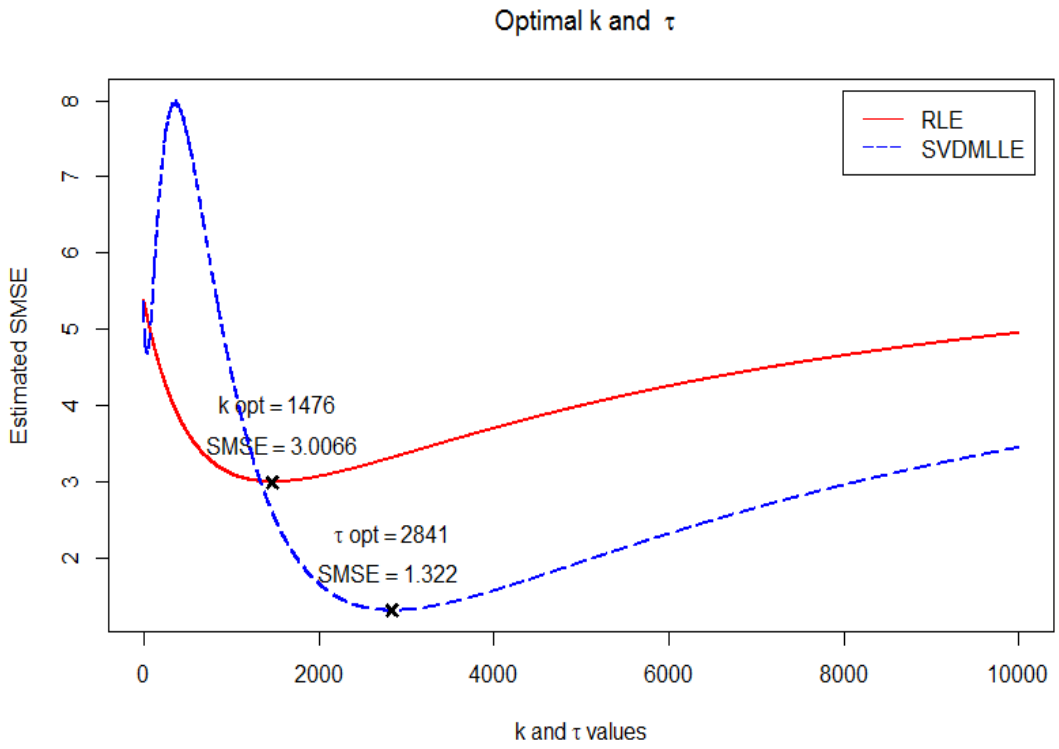


Fig. 8.1. The best values (k_{opt}) and (τ_{opt}) for LRE and SVD-MLLE with minimum MSE.

7. Conclusion

This paper introduces a new estimator to solve the multicollinearity problem in the binary logistic model. This estimator is based on the SVD technique for the design matrix (X) and is called the SVD-based maximum likelihood logistic estimator (SVD-MLLE). Also, we derived some statistical properties of this estimator such as bias, variance-covariance matrix, and scalar mean squared error. The results of the simulation study and the real data application reveal that the proposed estimator outperforms the existing MLE and RLE estimators in the SMSE criterion under different situations. Further, the proposed SVD-MLLE estimator performs well compared to MLE and RLE when the multicollinearity among the explanatory variables is high.

In addition, we can conclude that determining the optimal value of the scalar parameter (τ_{opt}) numerically by plotting the mean square error of the SVD-MLLE estimator versus many values of τ_i to obtain the best value (τ_{opt}) which has the minimum MSE value may gain benefits.

Since the choice of scalar parameters $(\tau_i's)$ affects the performance of the SVD-MLLE estimator, we recommend using $\tau_i(1)$ for small sample size or nearly perfect correlation between the explanatory variables. While $\tau_i(2)$ works well when the model includes a fewer number of explanatory variables. In contrast, $\tau_i(3)$ performs well with increasing the number of explanatory variables in the model.

Therefore, the results show that the ML estimator provides the least performance as expected when multicollinearity exists. So, the ML estimator should not be used in the presence of a multicollinearity problem since the parameter becomes unstable and it has a large SMSE. This problem is especially severe when the correlation between explanatory variables is high and the data size is small.

References:

Abonazel, M., Dawoud, I., Awwad, F. and Tag-Eldin, E. (2023). New estimators for the probit regression model with multicollinearity. Scientific African, Vol, 19.

Asar, Y. and Genç, A. (2016). New Shrinkage Parameters for the Liu-type Logistic Estimators. J. Communications in Statistics, Vol. 45, pp. 1094–1103.

Asar, Y. and Genç, A. (2017). Two-parameter ridge estimator in the binary logistic regression. Communications in Statistics-Simulation and Computation, Vol. 46(6), pp. 7088-7099.

Awwad, F., Odeniyi, K., Dawoud, I., Algamal, Z., Abonazel, M., Kibria, B. and Eldin, E. (2022). New two-parameter estimators for the logistic regression model with multicollinearity. WSEAS Trans. Math, Vol. 21, pp. 403-414.

Cattell, R. (1966). The scree test for the number of factors. Multivar Behav Res Vol. 1(2), pp. 245–276.

Inan, D. and Erdogan, B. (2013). Liu-type logistic estimator. Comm. Stat. and Simulation and Computation, Vol 42, pp. 1578-1586.

Jadhav, N. (2020). On linearized ridge logistic estimator in the presence of multicollinearity. Computational Statistics, Vol. 35(2), pp. 667-687.

Kibria, B. (2003). Performance of some new ridge regression estimators. Commun. Statist. Theor. Meth., Vol. 32, pp. 419-435.

Kibria, B., Månsson, K. and Shukur, G. (2012). Performance of some logistic ridge regression estimators. Comput Econ, Vol. 40, pp. 401-414.

Liu, X. (2010). A Class of generalized shrunken least squares estimators in linear model. MSc. thesis. University of Manitoba, Canada.

Lukman, A., Emmanuel, A., Clement, O. and Ayinde, K. (2020). A modified ridge-type logistic estimator. Iranian Journal of Science and Technology, Transactions A: Science, Vol. 44, pp. 437-443.

Lukman, A., Ayinde, K., Binuomote, S. and Onate, A. (2019) Modified ridge-type estimator to combat multicollinearity: application to chemical data. Journal of Chemometrics. Vol. 33.

Mansson, K., Kibria, B. and Shukur, G. (2012). On Liu estimators for the logit regression model. Economic Model, Vol. 29, pp. 1483–1488.

Mansson, K. and Shukur, G. (2011). On ridge parameters in logistic regression. Comm Statist Theory Methods, Vol. 40, pp. 3366–3381.

Midi, H., Sarkar, S. and Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. Journal of interdisciplinary mathematics, Vol. 13(3), pp.253-267.

Nja, M., Ogoke, U. and Nduka, E. (2013). The logistic regression model with a modified weight function. Journal of Statistical and Econometric Method, Vol. 2, pp.161−171.

Roozbeh, M., Kafakia, S. and Arashi, M. (2016). A class of biased estimators based on QR decomposition. Linear Algebra and its Applications, ELSEVIER, Vol. 508, pp. 190–205.

Schaefer, R., Roi, L. and Wolfe, R. (1984). A ridge logistic estimator. Comm. Stat. Theory Methods, Vol. 13, pp. 99–113.

Smith, K., Slattery, M. and French, T. (1991). Collinear nutrients and the risk of colon cancer. Journal of Clinical Epidemiolgy, Vol.44, pp.715-723.

Varathan, N. (2022). An improved ridge type estimator for logistic regression. Statistics in Transition New Series, Vol.23, pp.113-126.

Varathan, N. and Wijekoon, P. (2017) Optimal generalized logistic estimator. Commun Stat Theory Methods, Vol. 47(2): pp . 463–474.

Varathan, N. and Wijekoon, P. (2022). Modified almost unbiased Liu estimator in logistic regression. Communications in Statistics-Simulation and Computation, Vol. 50(11), pp. 3530-3546.

Watkins, D. (2002). Fundamentals of matrix computations. 2nd ed., John Wiley, New York.

Weissfeld, L. and Sereika, S. (1991). A multicollinearity diagnostic for generalized linear models. Commun Stat Theory Methods, Vol 20(4), pp.1183–1198.

Wu, J. and Asar, Y. (2016). On almost unbiased ridge logistic estimator for the logistic regression model. Hacettepe Journal of Mathematics and Statistics, Vol. 45(3), pp. 989–998.

Xinfeng, C. (2015). On the almost unbiased ridge and Liu estimator in the logistic regression model. International Conference on Social Science, Education Management and Sports Education. Atlantis Press, pp. 1663−1665.