Thailand Statistician
April 2020; 18(2): 176-195
http://statassoc.or.th
Contributed paper

Post Selection Estimation and Prediction in Poisson Regression Model

Orawan Reangsephet*[a], Supranee Lisawadi [a] and Syed Ejaz Ahmed [b]
[a] Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathum Thani, Thailand.
[b] Department of Mathematics and Statistics, Brock University, St. Catharines, Ontario, Canada.
*Corresponding author; e-mail: reangsephet.o@gmail.com

Received: 29 August 2018
Revised: 2 March 2019
Accepted: 17 April 2019

Abstract

The use of subspace information for estimating parameters of the model has gained increasing attention in recent years. However, the quality of the subspace information is usually unknown, and in consequence the classical maximum likelihood estimation strategies, which rely on this information, become biased and inefficient. Our goal was to improve the performance of estimation strategies for a Poisson regression model for which subspace information is available. We proposed estimators based on the linear shrinkage, preliminary test, and Stein-type strategies and investigated their asymptotic properties using the notation of asymptotic distributional bias and risk. Comprehensive Monte Carlo simulations were conducted to assess the simulated relative efficiency of the proposed estimators. Further, comparisons were made with the two penalized likelihood estimators: least absolute shrinkage and selection operator (LASSO) and ridge. Finally, the proposed estimators were applied to a real data set, to confirm their usefulness. Based on our findings, the proposed estimators were more efficient than the classical estimator when the accuracy of the subspace information was unknown.

Keywords: Linear shrinkage, preliminary test, Stein-type, penalized likelihood, Monte Carlo simulation.

1. Introduction

In many fields, such as physical biology, social sciences, and epidemiology, the response of interest is represented by count data in which large count numbers are rare. A widely-used statistic tool in the analysis of count data is the Poisson regression model, given as follows:

$$f(y_i|\boldsymbol{x}_i)=\frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!},\quad y_i=0,1,2,...;\ i=1,2,...,n, \tag{1}$$

where y_i is the independent Poisson response variable for the i^{th} subject, $\boldsymbol{x}_i=(x_{i1},x_{i2},...,x_{ip})^{\mathrm{T}}$ is a $p\times 1$ predictor vector for the i^{th} subject, $\boldsymbol{\beta}=(\beta_1,\beta_2,...,\beta_p)^{\mathrm{T}}$ is a $p\times 1$ vector of regression

coefficients, and $\mu_i = \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$ is the mean parameter for the i^{th} subject. For a detailed discussion see, for example, Cameron and Trivedi (2013) or Myers et al. (2012).

Our primary focus was on parameter estimation for the Poisson regression model in cases when many predictors are available, but these may or may not be significant for the response of interest. A range of variable selection procedures can be used to produce the insignificant predictors, including the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). This information, which is commonly called subspace information, gives two choices of model. The first is a full model that takes all predictors into account. The second is a submodel, based on the subspace information that retains only the significant predictors. We can therefore split the parameter vector of the full model $\boldsymbol{\beta}$ into two subvectors as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ represent a $p_1 \times 1$ significant parameter subvector and a $p_2 \times 1$ insignificant parameter subvector, respectively, such that $p_1 + p_2 = p$. Under available subspace information, we are interested in the estimation of the significant parameter subvector $\boldsymbol{\beta}_1$ when $\boldsymbol{\beta}_2$ is a known vector $\boldsymbol{\beta}_2^0$, so that $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^0$. Without loss of generality, $\boldsymbol{\beta}_2^0$ may be set to zero vector. It is important to keep in mind that the efficiency of both full model and submodel estimators are directly impacted by the uncertain of subspace information.

As previously stated, Hossain and Ahmed (2012) studied Stein-type shrinkage estimators and applied three penalty procedures, including least absolute shrinkage and selection operator (LASSO), adaptive LASSO, and smoothly clipped absolute deviation (SCAD), to estimate the parameters in a Poisson regression model when the subspace information was available. They reported that the shrinkage estimators dominated classical maximum likelihood estimator across a wide class of models. However, these estimators outperformed the penalty estimators only when the number of insignificant predictors was moderate to large.

Another way of dealing with the problem from the uncertain of subspace information is to use the preliminary test strategy that checks the validity of subspace information. In this study, therefore, we extend the work of Hossain and Ahmed (2012) by applying a preliminary test to remove the uncertainty concerning the available subspace information before the information $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^0$ is incorporated into the estimation process. We further also applied the ridge estimator, which improves the precision of parameter estimation when many predictors are present in the model and/or the multicollinearity problem exists, see Ahmed (2014) and Mansson and Shukur (2011). Previous studies that have applied these strategies to estimation of a parameter of interest include Ahmed et al. (2015), Al-Kandari et al. (2007), Al-Momani et al. (2017), Gao et al. (2017), Reangsephet et al. (2018), and Yüzbasi and Ahmed (2015). These have reported that estimators based on preliminary test and Stein-type strategies performed better than the classical estimators.

The rest of this paper is organized as follows. In section 2, some common estimation strategies are discussed. We present and compare their asymptotic properties in Section 3. We conducted Monte Carlo simulations to compare the performance of the proposed estimators. The results are reported in Section 4, and a real data example is given in Section 5. Finally, in Section 6, we present our conclusions and make recommendations.

2. Estimation Strategies

Consider the log-likelihood function of the Poisson regression model defined in (1) is given by

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) - \ln(y_i!) \right\}. \tag{2}$$

The derivatives of the log-likelihood function with respect to $\boldsymbol{\beta}$ are obtained by solving the following score equation

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left\{ y_i - \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) \right\} \boldsymbol{x}_i = \mathbf{0}_p. \tag{3}$$

2.1. Full model and submodel estimators

The full model (FM) estimator, denoted by of $\hat{\boldsymbol{\beta}}^{\mathrm{FM}}$, is the maximum likelihood estimator (MLE), which is obtained by solving the score equation in (3). Since, this score equation is the nonlinear function in parameter $\boldsymbol{\beta}$, we need to solve (3) by using the Newton-Raphson iterative method to obtain the value of $\hat{\boldsymbol{\beta}}^{\mathrm{FM}}$. As the results of Santos and Neves (2008), we can state the following theorem.

Theorem 1 *Under the usual regularity conditions of MLE, as* $n \to \infty$, $\hat{\boldsymbol{\beta}}^{\mathrm{FM}} \xrightarrow{D} N_p(\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\beta})^{-1})$, *where* $\mathbf{V}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}) \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}$ *is the Fisher information matrix.*

The FM estimator $\hat{\boldsymbol{\beta}}^{\mathrm{FM}}$ can be partitioned as $\hat{\boldsymbol{\beta}}^{\mathrm{FM}} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} \\ \hat{\boldsymbol{\beta}}_2^{\mathrm{FM}} \end{bmatrix}$. Consequently, the Fisher information matrix $\mathbf{V}(\boldsymbol{\beta})$ can be written as $\mathbf{V}(\boldsymbol{\beta}) = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix}$, and then $\mathbf{V}(\boldsymbol{\beta})^{-1} = \begin{bmatrix} \mathbf{V}_{11.2}^{-1} & -\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1} \\ -\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\mathbf{V}_{11.2}^{-1} & \mathbf{V}_{22.1}^{-1} \end{bmatrix}$, where $\mathbf{V}_{11.2} = \mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{22}^{-1}\mathbf{V}_{21}$ and $\mathbf{V}_{22.1} = \mathbf{V}_{22} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}$.

Now, we consider the information $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^0$ and then add this information on model in (1). Hence, we have the candidate submodel (SM), so that only $\boldsymbol{\beta}_1$ is unknown vector. The SM estimator of $\boldsymbol{\beta}_1$, denoted by $\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}$, can be obtained by solving (3), subject to $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_2^0 = \mathbf{0}_{p_2}$.

2.2. Linear shrinkage estimator

The linear shrinkage (LS) estimator of the parameter vector $\boldsymbol{\beta}_1$, denoted by $\hat{\boldsymbol{\beta}}_1^{\mathrm{LS}}$, is a linear combination of the full model and the submodel estimator

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{LS}} = \hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \lambda(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}), \tag{4}$$

where $\lambda \in [0,1]$ represents the degree of confidence in the given subspace information. Its value may be set by using the researcher's belief in the accuracy of the available subspace information or by minimizing the mean squared error of this estimator.

2.3. Shrinkage pretest estimator

The shrinkage pretest (SP) estimator of the parameter vector $\boldsymbol{\beta}_1$ is defined as

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{SP}} = \hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - (\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{LS}}) I(D_n \le d_{n,\alpha}). \tag{5}$$

Alternatively,

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{SP}} = \hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \lambda(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{SM}})I(D_n \le d_{n,\alpha}), \tag{6}$$

where $I(.)$ is an indicator function, and $d_{n,\alpha}$ is the α-level critical value of the exact distribution of a suitable test statistic D_n for testing $H_0 : \boldsymbol{\beta}_2 - \boldsymbol{\beta}_2^0 = \mathbf{0}_{p_2}$. This estimator determines the choice of the full model or submodel. We note that this estimator is called as the preliminary test (PT) estimator when $\lambda = 1$. In this study, we suggest the likelihood ratio statistic D_n in (7) for testing H_0

$$D_n = n(\hat{\boldsymbol{\beta}}_2^{\mathrm{FM}} - \boldsymbol{\beta}_2^0)^{\mathrm{T}} \mathbf{V}_{22.1}(\hat{\boldsymbol{\beta}}_2^{\mathrm{FM}} - \boldsymbol{\beta}_2^0) + o_{p_2}(1). \tag{7}$$

Under H_0, as $n \to \infty$, the distribution of D_n converges to a χ^2 distribution with p_2 degrees of freedom.

2.4. Stein-type and positive-part Stein-type shrinkage estimators

The Stein-type shrinkage estimator (SE) and positive-part Stein-type shrinkage estimator (PSE) for Poisson regression model are proposed and discussed by Hossain and Ahmed (2012). We now briefly introduce these estimators. The SE and PSE of the parameter vector $\boldsymbol{\beta}_1$ are respectively given by

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{SE}} = \hat{\boldsymbol{\beta}}_1^{\mathrm{SM}} + (1 - (p_2 - 2)D_n^{-1})(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}), \quad p_2 \ge 3 \tag{8}$$

and

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{PSE}} = \hat{\boldsymbol{\beta}}_1^{\mathrm{SM}} + (1 - (p_2 - 2)D_n^{-1})^{+}(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}), \quad p_2 \ge 3, \tag{9}$$

where $(1 - (p_2 - 2)D_n^{-1})^{+} = \max\{0, 1 - (p_2 - 2)D_n^{-1}\}$. Alternatively, the PSE can be written in the canonical form as

$$\hat{\boldsymbol{\beta}}_1^{\mathrm{PSE}} = \hat{\boldsymbol{\beta}}_1^{\mathrm{SE}} - (1 - (p_2 - 2)D_n^{-1})(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{SM}})I(D_n \le (p_2 - 2)), \quad p_2 \ge 3. \tag{10}$$

2.5. Penalized likelihood estimator

The penalized likelihood estimation strategies are different from the previously mentioned strategies in that they shrink all the coefficients toward zero equally. In this study, we consider the widely recognized penalized likelihood procedures, which commonly produce more precise and accurate estimates, including the least absolute shrinkage and selection operator estimator (LASSO) and ridge estimator. Suppose that $\pi \ge 0$ represents the tuning parameter and it controls the amount of shrinkage. The LASSO estimator performs simultaneous variable selection and parameter estimation. It uses an L_1 penalty and is therefore given by

$$\hat{\boldsymbol{\beta}}^{\mathrm{LASSO}} = \arg\min_{\boldsymbol{\beta}} \left\{ -\sum_{i=1}^{n} \left\{ y_i x_i^T \beta - \exp(x_i^T \beta) - \ln(y_i !) \right\} + \pi \sum_{j=1}^{p} |\beta_j| \right\}. \tag{11}$$

The ridge estimator of Hoerl and Kennard (1970) uses an L_2 penalty and is defined by

$$\hat{\beta}^{Ridge} = \arg\min_{\beta} \left\{ -\sum_{i=1}^{n} \left\{ y_i x_i^T \beta - \exp(x_i^T \beta) - \ln(y_i !) \right\} + \pi \sum_{j=1}^{p} \beta_j^2 \right\}. \tag{12}$$

Practically, the cross-validation method is a way of selecting an optimal tuning parameter π for the penalized likelihood estimators.

3. Asymptotic Properties

To study the asymptotic properties in terms of bias and risk of the proposed estimators, we now define the sequence of local alternatives $H_{(n)}$ as follows

$$H_{(n)}: \boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^{(n)}, \quad \boldsymbol{\beta}_2^{(n)} = \boldsymbol{\beta}_2^0 + \frac{\boldsymbol{\delta}}{\sqrt{n}}. \tag{13}$$

Here, $\boldsymbol{\delta} = (\delta_1, \delta_2, ..., \delta_{p_2})^{\mathrm{T}} \in \mathbb{R}^{p_2}$ is a $p_2 \times 1$ fixed vector. The quantity of $\frac{\boldsymbol{\delta}}{\sqrt{n}}$ measures the extent to which the local alternatives $H_{(n)}$ differ from the subspace information $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^0$.

For simplicity, we use the notation $\psi_v(x;\Delta)$ as to represent the cumulative distribution function of a noncentral χ^2 distribution with non-centrality parameter Δ and v degrees of freedom. Further, $E\left[\chi_v^{-2j}(\Delta)\right] = \int_0^\infty x^{-2j} d\psi_v(x;\Delta)$. In order to prove the asymptotic properties of the proposed estimators, we first present important lemmas:

Lemma 1 *Following Judge and Bock (1978), let* $\mathbf{z}$ *be a* k*-dimensional random vector that follows multivariate normal distribution with mean* $\boldsymbol{\mu}_{\mathbf{z}}$ *and covariance matrix* $\boldsymbol{\Sigma}_{\mathbf{z}}$*. Then, for any measurable function* φ*, we have*

$$E[\mathbf{z}\varphi(\mathbf{z}^{\mathrm{T}}\mathbf{z})] = \boldsymbol{\mu}_{\mathbf{z}} E[\varphi(\chi^2_{k+2}(\Delta))], \tag{14}$$

$$E[\mathbf{z}\mathbf{z}^{\mathrm{T}}\varphi(\mathbf{z}^{\mathrm{T}}\mathbf{z})] = \boldsymbol{\Sigma}_{\mathbf{z}} E[\varphi(\chi^2_{k+2}(\Delta))] + \boldsymbol{\mu}_{\mathbf{z}}\boldsymbol{\mu}_{\mathbf{z}}^{\mathrm{T}} E[\varphi(\chi^2_{k+4}(\Delta))], \tag{15}$$

where Δ *is the non-centrality parameter.*

Lemma 2 *Under the sequence of local alternative* $H_{(n)}$ *and the usual regularity condition of MLE, as* $n \to \infty$*, the test statistic* D_n *converges to a non-central* χ^2 *distribution with non-centrality parameter* $\Delta = \boldsymbol{\delta}^{\mathrm{T}}\mathbf{V}_{22.1}\boldsymbol{\delta}$ *and* p_2 *degrees of freedom.*

Lemma 3 *Under the sequence of local alternative* $H_{(n)}$ *and the usual regularity condition of MLE, as* $n \to \infty$*,*

$$\begin{pmatrix} \boldsymbol{W}_n \\ \boldsymbol{Z}_n \end{pmatrix} \xrightarrow{D} \begin{pmatrix} \boldsymbol{W} \\ \boldsymbol{Z} \end{pmatrix} \sim N_{p_1+p_2}\left(\begin{bmatrix} \mathbf{0}_{p_1} \\ \boldsymbol{\delta} \end{bmatrix}, \begin{bmatrix} \mathbf{V}_{11.2}^{-1} & -\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1} \\ -\mathbf{V}_{22}^{-1}\mathbf{V}_{21}\mathbf{V}_{11.2}^{-1} & \mathbf{V}_{22.1}^{-1} \end{bmatrix} \right), \tag{16}$$

$$\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \sim N_{p_1}(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta}, \mathbf{V}_{11.2}^{-1} - \mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}), \tag{17}$$

$$\boldsymbol{Y}_n \xrightarrow{D} \boldsymbol{Y} \sim N_{p_1}(-\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta}, \mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}), \tag{18}$$

where $\boldsymbol{W}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \boldsymbol{\beta}_1)$, $\boldsymbol{Z}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_2^{\mathrm{FM}} - \boldsymbol{\beta}_2)$, $\boldsymbol{X}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}} - \boldsymbol{\beta}_1)$, *and* $\boldsymbol{Y}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{SM}})$.

Proof: See Appendix A.

Throughout this work, we assumed the normalized parameter estimators to be uniformly integrable. We present the asymptotic distributional bias and the asymptotic distributional risk results of the estimators in the following section.

3.1. Asymptotic distributional bias

The asymptotic distributional bias (ADB) for any estimator $\hat{\boldsymbol{\beta}}_1^*$ can be defined as

$$B(\hat{\boldsymbol{\beta}}_1^*) = \lim_{n\to\infty} E\left[\sqrt{n}(\hat{\boldsymbol{\beta}}_1^* - \boldsymbol{\beta}_1)\right]. \tag{19}$$

By virtue of Lemmas 1 to 3 and by using the definition of ADB in (19), we give the ADBs of the proposed estimators in the following theorem.

Theorem 2 *Under the sequence of local alternatives* $H_{(n)}$ *and the usual regularity conditions of MLE, as* $n \to \infty$,

$$B(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}}) = \mathbf{0}, \tag{20}$$

$$B(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}) = \mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta}, \tag{21}$$

$$B(\hat{\boldsymbol{\beta}}_1^{\mathrm{LS}}) = \lambda\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta}, \tag{22}$$

$$B(\hat{\boldsymbol{\beta}}_1^{\mathrm{SP}}) = \lambda\psi_{p_2+2}(\chi^2_{p_2,\alpha};\Delta)\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta}. \tag{23}$$

Here, $\Delta = \boldsymbol{\delta}^{\mathrm{T}}\mathbf{V}_{22.1}\boldsymbol{\delta}$.

Proof: See Appendix B.

Following Hossain and Ahmed (2012), we found that the ADBs of the SE and PSE are respectively, as follows:

$$B(\hat{\boldsymbol{\beta}}_1^{\mathrm{SE}}) = (p_2-2)\,\mathrm{E}[\chi^{-2}_{p_2+2}(\Delta)]\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta}, \tag{24}$$

and

$$B(\hat{\boldsymbol{\beta}}_1^{\mathrm{PSE}}) = \left\{(p_2-2)\,\mathrm{E}[\chi^{-2}_{p_2+2}(\Delta)]I(\chi^2_{p_2+2}(\Delta) > (p_2-2)) + \psi_{p_2+2}(p_2-2;\Delta)\right\}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta}. \tag{25}$$

Further, when $\lambda = 1$, $B(\hat{\boldsymbol{\beta}}_1^{\mathrm{SP}})$ becomes the ADB of the PT estimator, denoted by $\hat{\boldsymbol{\beta}}_1^{\mathrm{PT}}$, which is given by $B(\hat{\boldsymbol{\beta}}_1^{\mathrm{PT}}) = \psi_{p_2+2}(\chi^2_{p_2,\alpha};\Delta)\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta}$. The ADB expressions for the estimators are not in scalar form. To obtain a scalar quantity of $B(\hat{\boldsymbol{\beta}}_1^*)$, we take the recourse by converting them to the quadratic form. This is called the asymptotic quadratic distributional bias (ADQB), and is defined as

$$QB(\hat{\boldsymbol{\beta}}_1^*) = [B(\hat{\boldsymbol{\beta}}_1^*)]^{\mathrm{T}}\mathbf{V}_{11.2}[B(\hat{\boldsymbol{\beta}}_1^*)]. \tag{26}$$

Using the definition in (26), we present the ADQBs of the estimators as follows:

$$QB(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}}) = 0,$$

$$QB(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}) = \Delta^*,$$

$$QB(\hat{\boldsymbol{\beta}}_1^{\mathrm{LS}}) = \lambda^2\Delta^*,$$

$$QB(\hat{\boldsymbol{\beta}}_1^{\mathrm{SP}}) = \left\{\lambda\psi_{p_2+2}(\chi^2_{p_2,\alpha};\Delta)\right\}^2\Delta^*,$$

$$QB(\hat{\boldsymbol{\beta}}_1^{\mathrm{SE}}) = \left\{(p_2-2)\,\mathrm{E}[\chi^{-2}_{p_2+2}(\Delta)]\right\}^2\Delta^*,$$

$$QB(\hat{\boldsymbol{\beta}}_1^{\mathrm{PSE}}) = \left\{(p_2-2)\mathrm{E}[\chi_{p_2+2}^{-2}(\Delta)]I(\chi_{p_2+2}^2(\Delta) > (p_2-2)) + \psi_{p_2+2}(p_2-2;\Delta)\right\}^2 \Delta^*.$$

Here, $\Delta^* = \boldsymbol{\delta}^{\mathrm{T}}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{11.2}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta}$.

The proofs of the above ADQB expressions can be easily obtained by using results of ADB in Theorem 2. From the bias results, the $\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}}$ is an unbiased estimator of $\boldsymbol{\beta}_1$, while other estimators are biased and their biases depend on the quantity of Δ^*. The AQDBs of $\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}$ and $\hat{\boldsymbol{\beta}}_1^{\mathrm{LS}}$ are the unbounded functions of Δ^*. The AQDB of $\hat{\boldsymbol{\beta}}_1^{\mathrm{SP}}$ depends on both the degree of confidence in the subspace information λ and size of the test α. That of $\hat{\boldsymbol{\beta}}_1^{\mathrm{SP}}$ increases to a maximum and then slowly decreases to zero as Δ increases. Similarly, at $\Delta = 0$, the AQDBs of $\hat{\boldsymbol{\beta}}_1^{\mathrm{SE}}$ and $\hat{\boldsymbol{\beta}}_1^{\mathrm{PSE}}$ start from zero, increase to a maximum point, and then decrease again towards zero, for the reason that $E[\chi_{p_2+2}^{-2}(\Delta)]$ is the decreasing log convex function of Δ. Moreover, the AQDB of $\hat{\boldsymbol{\beta}}_1^{\mathrm{PSE}}$ is always smaller than or equal to $\hat{\boldsymbol{\beta}}_1^{\mathrm{SE}}$ for all values of Δ.

3.2. Asymptotic distributional risk

Let $\boldsymbol{Q}$ be a known positive semi-definite matrix, and $\hat{\boldsymbol{\beta}}_1^*$ be any estimator of $\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}$, $\hat{\boldsymbol{\beta}}_1^{\mathrm{LS}}$, $\hat{\boldsymbol{\beta}}_1^{\mathrm{SP}}$, $\hat{\boldsymbol{\beta}}_1^{\mathrm{SE}}$ or $\hat{\boldsymbol{\beta}}_1^{\mathrm{PSE}}$. We consider the quadratic loss function $L(\hat{\boldsymbol{\beta}}_1^*;\boldsymbol{Q}) = n(\hat{\boldsymbol{\beta}}_1^* - \boldsymbol{\beta}_1)^{\mathrm{T}}\boldsymbol{Q}(\hat{\boldsymbol{\beta}}_1^* - \boldsymbol{\beta}_1)$. Then, the asymptotic distributional risk (ADR) of $\hat{\boldsymbol{\beta}}_1^*$ is defined as

$$R(\hat{\boldsymbol{\beta}}_1^*;\boldsymbol{Q}) = trace[\boldsymbol{Q}MSE(\hat{\boldsymbol{\beta}}_1^*)]. \tag{27}$$

Here, $MSE(\hat{\boldsymbol{\beta}}_1^*)$ is the asymptotic mean squared error matrix (MSE) of the estimator $\hat{\boldsymbol{\beta}}_1^*$, which can be defined as

$$MSE(\hat{\boldsymbol{\beta}}_1^*) = \lim_{n\to\infty} E\left[\sqrt{n}(\hat{\boldsymbol{\beta}}_1^* - \boldsymbol{\beta}_1)(\hat{\boldsymbol{\beta}}_1^* - \boldsymbol{\beta}_1)^{\mathrm{T}}\right]. \tag{28}$$

Using the definitions in (27) and (28) and Lemmas 1 to 3, the ADRs of the proposed estimators are contained in the following theorem.

Theorem 3 *Expressions for the ADRs of the proposed estimators under the sequence of local alternatives* $H_{(n)}$ *and the usual regularity conditions of MLE, as* $n \to \infty$,

$$R(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}};\boldsymbol{Q}) = trace[\boldsymbol{Q}\mathbf{V}_{11.2}^{-1}], \tag{29}$$

$$R(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}};\boldsymbol{Q}) = trace[\boldsymbol{Q}\mathbf{V}_{11.2}^{-1}] - trace[\boldsymbol{Q}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}] + \Delta_R, \tag{30}$$

$$R(\hat{\boldsymbol{\beta}}_1^{\mathrm{LS}};\boldsymbol{Q}) = trace[\boldsymbol{Q}\mathbf{V}_{11.2}^{-1}] - \lambda(2-\lambda)trace[\boldsymbol{Q}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}] + \lambda^2\Delta_R, \tag{31}$$

$$\begin{aligned} R(\hat{\boldsymbol{\beta}}_1^{\mathrm{SP}};\boldsymbol{Q}) = {} & trace[\boldsymbol{Q}\mathbf{V}_{11.2}^{-1}] - \lambda(2-\lambda)\psi_{p_2+2}(\chi_{p_2,\alpha}^2;\Delta)trace[\boldsymbol{Q}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}] \\ & + \lambda\left\{2\psi_{p_2+2}(\chi_{p_2,\alpha}^2;\Delta) - (2-\lambda)\psi_{p_2+4}(\chi_{p_2,\alpha}^2;\Delta)\right\}\Delta_R. \end{aligned} \tag{32}$$

Here, $\Delta_R = \boldsymbol{\delta}^{\mathrm{T}}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}\boldsymbol{Q}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta}$.

Proof: See Appendix C.

Following Hossain and Ahmed (2012), we get

$$R(\hat{\boldsymbol{\beta}}_1^{\mathrm{SE}};\boldsymbol{Q}) = trace[\boldsymbol{Q}\mathbf{V}_{11.2}^{-1}] - (p_2-2)\begin{Bmatrix} 2E[\chi_{p_2+2}^{-2}(\Delta)] \\ -(p_2-2)E[\chi_{p_2+2}^{-4}(\Delta)] \end{Bmatrix} trace[\boldsymbol{Q}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}]$$
$$+(p_2-2)\left\{(p_2-2)E[\chi_{p_2+4}^{-4}(\Delta)] - 2\left(E[\chi_{p_2+4}^{-2}(\Delta)] - E[\chi_{p_2+2}^{-2}(\Delta)]\right)\right\}\Delta_R,$$

and

$$R(\hat{\boldsymbol{\beta}}_1^{\mathrm{PSE}};\boldsymbol{Q}) = R(\hat{\boldsymbol{\beta}}_1^{\mathrm{SE}};\boldsymbol{Q})$$
$$-E\left[\left\{1-(p_2-2)\chi_{p_2+2}^{-2}(\Delta)\right\}^2 I(\chi_{p_2+2}^2(\Delta)\le p_2-2)\right] trace[\boldsymbol{Q}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}]$$
$$+\begin{Bmatrix} 2E\left[\left\{1-(p_2-2)\chi_{p_2+2}^{-2}(\Delta)\right\} I(\chi_{p_2+2}^2(\Delta)\le p_2-2)\right] \\ -E\left[\left\{1-(p_2-2)\chi_{p_2+4}^{-2}(\Delta)\right\}^2 I(\chi_{p_2+4}^2(\Delta)\le p_2-2)\right] \end{Bmatrix}\Delta_R.$$

We see that if $\mathbf{V}_{12}=\mathbf{0}$, all ADR results reduce to a common value $trace[\boldsymbol{Q}\mathbf{V}_{11.2}^{-1}]$ for all $\boldsymbol{Q}$. Assuming that $\mathbf{V}_{12}\neq\mathbf{0}$, then

(i) $R(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}};\boldsymbol{Q})$ remains a constant with $trace[\boldsymbol{Q}\mathbf{V}_{11.2}^{-1}]$.

(ii) The ADRs of $\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}$ and $\hat{\boldsymbol{\beta}}_1^{\mathrm{LS}}$ are unbound functions of $\Delta_R\in[0,\infty)$. Since $\lambda\in[0,1]$, $\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}$ outperforms $\hat{\boldsymbol{\beta}}_1^{\mathrm{LS}}$ when Δ_R is equal to or close to zero. However, it has a higher risk than $\hat{\boldsymbol{\beta}}_1^{\mathrm{LS}}$ as $\Delta_R\to\infty$. In summary, $R(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}};\boldsymbol{Q})\le R(\hat{\boldsymbol{\beta}}_1^{\mathrm{LS}};\boldsymbol{Q})\le R(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}};\boldsymbol{Q})$ for $\Delta_R>0$.

(iii) $R(\hat{\boldsymbol{\beta}}_1^{\mathrm{SP}};\boldsymbol{Q})$ is bounded in Δ. $\hat{\boldsymbol{\beta}}_1^{\mathrm{SP}}$ achieves its smallest risk and outperforms $\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}}$ at $\Delta=0$. However, as Δ increases, its risk increases to a maximum value which is higher than $R(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}};\boldsymbol{Q})$. After passing through the maximum point, $R(\hat{\boldsymbol{\beta}}_1^{\mathrm{SP}};\boldsymbol{Q})$ monotonically approaches $R(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}};\boldsymbol{Q})$.

(iv) For all values of Δ with $p_2\ge 3$, we have $R(\hat{\boldsymbol{\beta}}_1^{\mathrm{PSE}};\boldsymbol{Q})\le R(\hat{\boldsymbol{\beta}}_1^{\mathrm{SE}};\boldsymbol{Q})\le R(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}};\boldsymbol{Q})$. However, when Δ is small, $\hat{\boldsymbol{\beta}}_1^{\mathrm{SE}}$ and $\hat{\boldsymbol{\beta}}_1^{\mathrm{PSE}}$ are less beneficial than the other estimators, except $\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}}$.

4. Simulation Results

In this section, we report the results of Monte Carlo simulations conducted to investigate the performance of the proposed estimators. The risk was estimated in term of the simulated mean squared error (SMSE) in estimation, and the performance of the listed estimators was compared using the simulated relative efficiency (SRE). The SRE of the estimator $\hat{\boldsymbol{\beta}}_1^{*}$ with respect to $\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}}$ was defined as

$$SRE(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}};\hat{\boldsymbol{\beta}}_1^{*}) = \frac{SMSE(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}})}{SMSE(\hat{\boldsymbol{\beta}}_1^{*})}. \tag{33}$$

Here, $\hat{\boldsymbol{\beta}}_1^*$ is any estimator of $\hat{\boldsymbol{\beta}}_1^{\text{SM}}$, $\hat{\boldsymbol{\beta}}_1^{\text{LS}}$, $\hat{\boldsymbol{\beta}}_1^{\text{SP}}$, $\hat{\boldsymbol{\beta}}_1^{\text{SE}}$, $\hat{\boldsymbol{\beta}}_1^{\text{PSE}}$, $\hat{\boldsymbol{\beta}}_1^{\text{LASSO}}$ or $\hat{\boldsymbol{\beta}}_1^{\text{Ridge}}$. An SRE is greater than one indicates that $\hat{\boldsymbol{\beta}}_1^*$ is superior to $\hat{\boldsymbol{\beta}}_1^{\text{FM}}$.

Our simulations were based on a Poisson regression model with sample size $n = 60$. The simulated Poisson response was generated from the following model: $y_i = \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, where $\boldsymbol{x}_i \sim N_p(\mathbf{0}_p, \boldsymbol{\Sigma})$ for $i = 1,2,...,n$. We considered two correlation structures of $\boldsymbol{\Sigma}$, AR(1) and constant. They were commonly used in many studies including Arashi et al. (2018), Yüzbasi et al. (2017b), and and Yüzbasi et al. (2017a). For AR(1) structure, the $(j,k)^{\text{th}}$ elements of $\boldsymbol{\Sigma}$ were defined to be equal to $\Sigma_{jk} = r^{|j-k|}$, $j = 1,2,...,p$; $k = 1,2,...,p$. For constant structure, all off-diagonal elements of $\boldsymbol{\Sigma}$ were defined to be equal to r. We set $r = 0.00$ and 0.75 in order to distinguish between uncorrelated and correlated predictors. The condition number (CN) was used to detect the existence of multicollinearity. A rule of thumb is that if CN is greater than 30, there is a reason to be concerned about multicollinearity problem.

Without loss of generality, we considered the hypothesis $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^0 = \mathbf{0}_{p_2}$ in the simulation studies. We partitioned $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are the $p_1 \times 1$ and $p_2 \times 1$ subvectors, respectively, so that $p_1 + p_2 = p$. We defined a parameter Δ^{sim} as $\Delta^{\text{sim}} = (\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)})^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)})$, where $\boldsymbol{\beta}^{(0)} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \mathbf{0}_{p_2}^{\mathrm{T}})^{\mathrm{T}}$, to represent how far the true value of parameter vector $\boldsymbol{\beta}$ deviated from the parameter vector of the model under subspace information $\boldsymbol{\beta}^{(0)}$. In this study, we considered the two following cases for the true value of $\boldsymbol{\beta}$:

1. At $\Delta^{\text{sim}} = 0$, $\boldsymbol{\beta}_1 = (1.20, -0.80, 0.15)^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = \mathbf{0}_{p_2}$.
2. At $\Delta^{\text{sim}} \geq 0$, $\boldsymbol{\beta}_1 = (-1.20, 0.80, 0.25)^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (\beta_4, \mathbf{0}_{p_2-1}^{\mathrm{T}})^{\mathrm{T}}$, so that $\Delta^{\text{sim}} = \beta_4^2$.

We set $p_1 = 3$, $p_2 = 3,5,7,10,15$, and $\Delta^{\text{sim}} \in [0,2]$. We also used the three common significance levels $\alpha =$ 0.01, 0.05 and 0.10 and the different values $\lambda = 0.25$, 0.50 and 0.75 to study their impact on the proposed estimators. The other values of r and other correlation structures, including compound symmetric and unstructured, were studied. However, their results were similar. Hence, we did not report them for space saving.

In this study, the tuning parameters π of two penalized likelihood estimators were estimated using 10-fold cross-validation. $N = 1,000$ iterations were run to obtain stable results for each configuration. All computations and graphics were conducted using the R programming (R Core Team 2015).

4.1. Case 1 subspace information is correct

The SREs of listed estimators with respect to the FM estimator are reported in Tables 1 and 3. According to these results, the SREs of all estimators increased as p_2 increased, and, with the exception of $\hat{\boldsymbol{\beta}}_1^{\text{Ridge}}$, were superior to $\hat{\boldsymbol{\beta}}_1^{\text{FM}}$ in the case of uncorrelated predictors. At $\Delta^{\text{sim}} = 0$, as we would anticipate, $\hat{\boldsymbol{\beta}}_1^{\text{SM}}$ produced the maximum risk reduction, especially when p_2 and r were large. For fixed p_2, the SRE of $\hat{\boldsymbol{\beta}}_1^{\text{LS}}$ decreased sharply to 1 as $\lambda \to 0$, while the performance of $\hat{\boldsymbol{\beta}}_1^{\text{SP}}$

became poorer as α increased and λ decreased. $\hat{\boldsymbol{\beta}}_1^{\text{PSE}}$ dominated $\hat{\boldsymbol{\beta}}_1^{\text{SE}}$ in every case, and its SRE increased rapidly as p_2 increased.

$\hat{\boldsymbol{\beta}}_1^{\text{LASSO}}$ outperformed $\hat{\boldsymbol{\beta}}_1^{\text{SE}}$ only when p_2 was small and $r = 0.00$. For $r = 0.00$, $\hat{\boldsymbol{\beta}}_1^{\text{Ridge}}$ performed poorly, but improved when multicollinearity was a serious concern. Moreover, as p_2 increased, $\hat{\boldsymbol{\beta}}_1^{\text{Ridge}}$ became dominated by LASSO, as it does not perform variable selection.

Table 1 Simulated relative efficiency of SM, LS, SP, SE, PSE, LASSO, and Ridge estimators with respect to FM at $\Delta^{\text{sim}} = 0$ when the predictors are perfectly uncorrelated $(r = 0.00)$

Estimator	Number of insignificant predictors				
	3	5	7	10	15
RE	2.460	4.573	6.557	12.882	20.871
LS					
$\lambda = 0.25$	1.354	1.500	1.542	1.657	1.714
$\lambda = 0.50$	1.810	2.335	2.512	3.119	3.497
$\lambda = 0.75$	2.264	3.731	4.852	7.337	9.310
SP					
$\lambda = 0.25, \alpha = 0.01$	1.339	1.481	1.504	1.634	1.692
$\lambda = 0.25, \alpha = 0.05$	1.278	1.388	1.446	1.535	1.621
$\lambda = 0.25, \alpha = 0.10$	1.238	1.335	1.403	1.467	1.534
$\lambda = 0.50, \alpha = 0.01$	1.764	2.258	2.347	2.986	3.345
$\lambda = 0.50, \alpha = 0.05$	1.593	1.923	2.119	2.483	2.914
$\lambda = 0.50, \alpha = 0.10$	1.491	1.755	1.969	2.203	2.478
$\lambda = 0.75, \alpha = 0.01$	2.175	3.316	4.601	6.467	8.085
$\lambda = 0.75, \alpha = 0.05$	1.867	2.718	3.560	4.280	5.588
$\lambda = 0.75, \alpha = 0.10$	1.697	2.325	2.806	3.232	3.934
SE	1.208	1.947	2.574	3.937	5.275
PSE	1.402	2.316	3.265	5.096	8.250
LASSO	1.209	1.511	2.203	3.255	4.636
Ridge	0.829	0.898	0.920	1.045	1.563
CN	3.133	3.920	4.505	5.137	13.965

4.2. Case 2 subspace information may be correct or incorrect

In this case, the penalized likelihood estimators were not included in the $\Delta^{\text{sim}} \geq 0$ case because these estimators do not take advantage of the subspace information $\boldsymbol{\beta}_2 = \mathbf{0}_{p_2}$. For the sake of brevity, we report here only the results for $p_2 = 5$, 10 and 15 with $\lambda = 0.75$ and $r = 0.00$. The SREs of the proposed estimators are reported in Table 4, and to ease comparison, shown as graphs in Figures 1 to 3.

The maximum SRE of all estimators was observed at $\Delta^{\text{sim}} = 0$. The submodel estimator $\hat{\boldsymbol{\beta}}_1^{\text{SM}}$ dominated all other estimators when the subspace information was either true or nearly true. As Δ^{sim} moved away from zero, its SRE decreased and converged to zero. The SRE of $\hat{\boldsymbol{\beta}}_1^{\text{LS}}$ approached zero as Δ^{sim} moved away from zero, though more slowly than that of $\hat{\boldsymbol{\beta}}_1^{\text{SM}}$. However, it outperformed other estimators in some space of Δ^{sim}.

Table 2 Simulated relative efficiency of SM, LS, SP, SE, PSE, LASSO, and Ridge estimators with respect to FM at $\Delta^{sim} = 0$ when the predictors are correlated $(r = 0.75)$ with AR(1)

Estimator	Number of insignificant predictors				
	3	5	7	10	15
RE	3.242	5.617	6.542	11.292	27.955
LS					
$\lambda = 0.25$	1.440	1.558	1.596	1.659	1.729
$\lambda = 0.50$	2.094	2.593	2.772	3.140	3.611
$\lambda = 0.75$	2.867	4.328	4.928	6.800	10.404
SP					
$\lambda = 0.25, \alpha = 0.01$	1.417	1.530	1.581	1.329	1.706
$\lambda = 0.25, \alpha = 0.05$	1.363	1.469	1.500	1.545	1.348
$\lambda = 0.25, \alpha = 0.10$	1.283	1.400	1.428	1.470	1.552
$\lambda = 0.50, \alpha = 0.01$	2.014	2.464	2.695	2.962	3.438
$\lambda = 0.50, \alpha = 0.05$	1.835	2.212	2.327	2.533	3.066
$\lambda = 0.50, \alpha = 0.10$	1.605	1.962	2.053	2.216	2.562
$\lambda = 0.75, \alpha = 0.01$	2.686	3.897	4.633	5.846	8.811
$\lambda = 0.75, \alpha = 0.05$	2.309	3.181	3.461	4.120	6.346
$\lambda = 0.75, \alpha = 0.10$	1.887	2.588	2.773	3.193	4.206
SE	1.329	1.983	2.644	3.540	6.538
PSE	1.509	2.525	3.278	4.918	9.333
LASSO	1.127	1.566	2.114	2.974	4.937
Ridge	1.161	1.306	1.329	1.794	2.584
CN	44.160	50.347	60.671	100.850	150.459

Table 3 Simulated relative efficiency of SM, LS, SP, SE, PSE, LASSO, and Ridge estimators with respect to FM at $\Delta^{sim} = 0$ when the predictors are correlated $(r = 0.75)$ with constant structure

Estimator	Number of insignificant predictors				
	3	5	7	10	15
RE	2.749	4.550	6.935	12.135	24.443
LS					
$\lambda = 0.25$	1.388	1.521	1.596	1.670	1.719
$\lambda = 0.50$	1.918	2.420	2.782	3.206	3.543
$\lambda = 0.75$	2.484	3.740	5.034	7.150	9.801
SP					
$\lambda = 0.25, \alpha = 0.01$	1.361	1.498	1.566	1.651	1.678
$\lambda = 0.25, \alpha = 0.05$	1.301	1.426	1.506	1.582	1.594
$\lambda = 0.25, \alpha = 0.10$	1.246	1.362	1.427	1.506	1.515
$\lambda = 0.50, \alpha = 0.01$	1.832	2.324	2.631	3.088	3.261
$\lambda = 0.50, \alpha = 0.05$	1.656	2.058	2.370	2.710	2.773
$\lambda = 0.50, \alpha = 0.10$	1.512	1.836	2.055	2.358	2.401
$\lambda = 0.75, \alpha = 0.01$	2.311	3.462	4.457	6.461	7.553
$\lambda = 0.75, \alpha = 0.05$	1.978	2.793	3.614	4.736	5.002
$\lambda = 0.75, \alpha = 0.10$	1.732	2.318	2.793	3.575	3.703
SE	1.256	1.890	2.565	3.792	5.733
PSE	1.425	2.328	3.290	5.444	8.041
LASSO	1.120	1.229	2.242	2.952	5.428
Ridge	0.844	1.395	1.530	1.837	3.780
CN	28.121	44.005	55.230	73.926	230.650

Table 4 Simulated relative efficiency of SM, LS, SP, SE, and PSE with respect to the FM for $\lambda = 0.75$ and $r = 0.00$

p_2	Δ^{sim}	Estimator						
		SM	LS	SP			SE	PSE
				$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$		
5	0.00	4.024	3.377	2.996	2.438	2.083	1.788	2.160
	0.05	1.023	1.395	1.170	1.093	1.058	1.263	1.306
	0.10	0.556	0.852	0.885	0.919	0.937	1.148	1.155
	0.15	0.416	0.662	0.870	0.926	0.950	1.093	1.095
	0.20	0.314	0.511	0.905	0.953	0.983	1.077	1.078
	0.25	0.253	0.419	0.910	0.968	0.984	1.055	1.055
	0.50	0.117	0.199	0.989	1.000	1.000	1.022	1.022
	0.75	0.075	0.130	1.000	1.000	1.000	1.015	1.015
	1.00	0.052	0.091	1.000	1.000	1.000	1.007	1.007
	2.00	0.020	0.035	1.000	1.000	1.000	1.003	1.003
10	0.00	8.786	5.904	5.215	3.920	2.962	3.464	4.364
	0.05	2.139	2.730	1.883	1.478	1.309	1.986	2.081
	0.10	1.251	1.791	1.244	1.128	1.083	1.620	1.636
	0.15	0.864	1.319	1.058	1.019	1.004	1.435	1.436
	0.20	0.704	1.104	0.989	0.993	0.998	1.348	1.349
	0.25	0.557	0.896	0.973	0.986	0.989	1.262	1.267
	0.50	0.270	0.459	0.996	0.998	0.998	1.136	1.136
	0.75	0.169	0.291	1.000	1.000	1.000	1.077	1.077
	1.00	0.117	0.203	1.000	1.000	1.000	1.057	1.057
	2.00	0.048	0.084	1.000	1.000	1.000	1.017	1.017
15	0.00	14.968	8.027	6.582	4.703	3.644	5.254	7.090
	0.05	4.049	4.401	2.941	2.058	1.661	2.987	3.224
	0.10	2.239	2.911	1.565	1.321	1.223	2.179	2.207
	0.15	1.655	2.323	1.279	1.108	1.065	1.898	1.904
	0.20	1.220	1.796	1.117	1.048	1.030	1.656	1.665
	0.25	0.969	1.503	1.029	1.006	1.003	1.543	1.543
	0.50	0.503	0.827	0.997	0.998	0.998	1.260	1.260
	0.75	0.333	0.562	1.000	1.000	1.000	1.172	1.172
	1.00	0.227	0.390	1.000	1.000	1.000	1.116	1.116
	2.00	0.099	0.174	1.000	1.000	1.000	1.043	1.043

The shrinkage pretest estimator $\hat{\boldsymbol{\beta}}_1^{SP}$ performed well at $\Delta^{sim} = 0$, then became inferior to $\hat{\boldsymbol{\beta}}_1^{FM}$ as Δ^{sim} increased, before equaling $\hat{\boldsymbol{\beta}}_1^{FM}$ as Δ^{sim} increased further. Both estimators based on the Stein-type strategy outperformed $\hat{\boldsymbol{\beta}}_1^{FM}$ across the entire space of Δ^{sim}, and also outperformed all other estimators in the wider space of Δ^{sim}. Further, their gain in risk reduction became more pronounced as p_2 increased. Clearly, the departure from the subspace information is central to $\hat{\boldsymbol{\beta}}_1^{SM}$ and $\hat{\boldsymbol{\beta}}_1^{LS}$, but had a smaller impact on the preliminary test and Stein-type strategies. These results were in agreement with the asymptotic results presented in Section 3. Though none of all estimators uniformly outperformed all others, the estimators based on the linear shrinkage, preliminary test, and Stein-type strategies were robust when there was the subspace information with uncertainty of correctness.

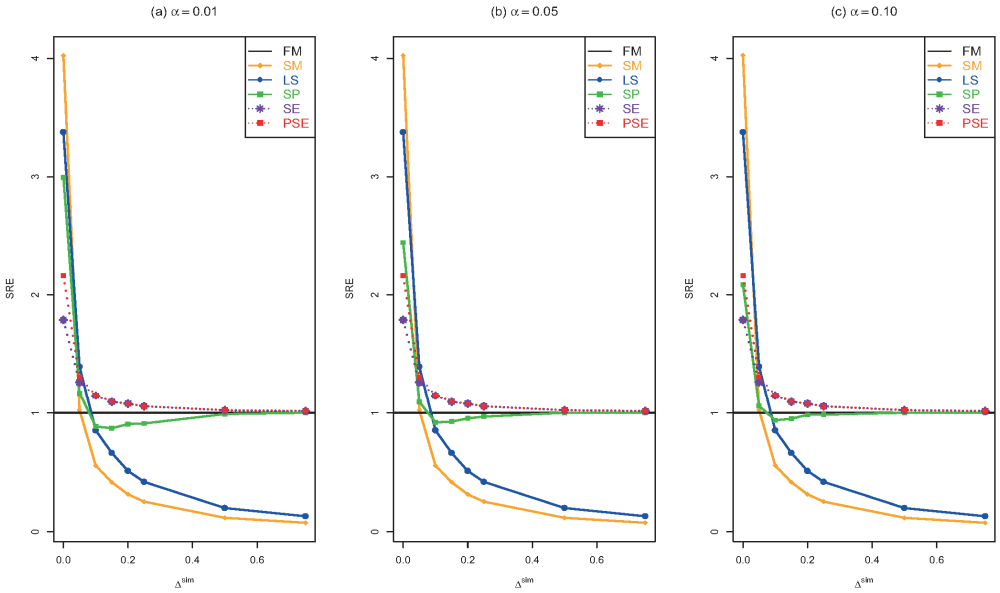

Figure 1 Simulated relative efficiency of FM, SM, LS, SP, SE, and PSE as a function of Δ^{sim} when $p_2 = 5$ for $\alpha = 0.01, 0.05$ and 0.10

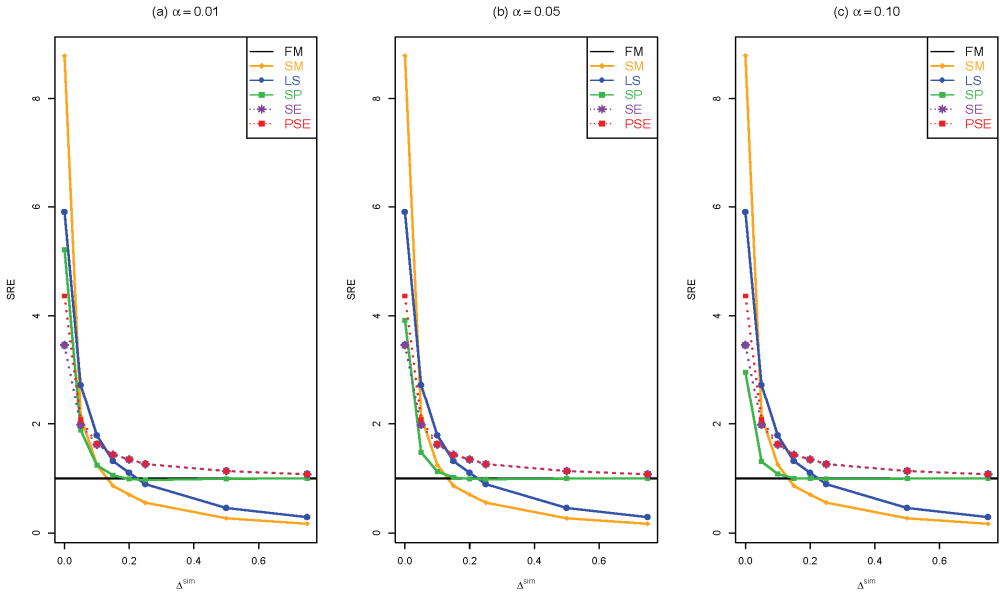

Figure 2 Simulated relative efficiency of FM, SM, LS, SP, SE, and PSE as a function of Δ^{sim} when $p_2 = 10$ for $\alpha = 0.01, 0.05$ and 0.10

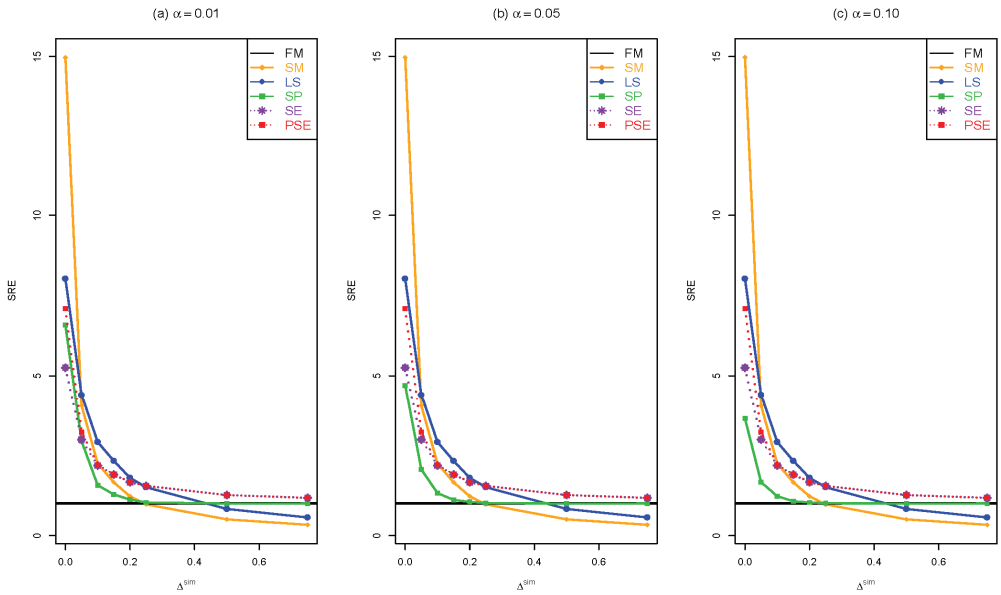

Figure 3 Simulated relative efficiency of FM, SM, LS, SP, SE, and PSE as a function of Δ^{sim} when $p_2 = 15$ for $\alpha = 0.01, 0.05$ and 0.10

5. Real Data Example

A data set of 189 pregnant women (reported in Hosmer and Lemeshow, 2000) was used to test the proposed estimators. The task was to make predictions about the number of physician visits during the first trimester. Multiple predictors were used: age of mother (AGE), history of hypertension (HT), where 1 = yes and 0 = no, weight of mother at last menstrual period (WEIGHT), smoking status during pregnancy (SMOKE), where 1 = yes and 0 = no, history of premature labor (PL), presence of uterine irritability (UI), where 1 = yes and 0 = no, birth weight (BW), and race (RACE), where 1 = white, 2 = black, and 3 = other. We found that a multicollinearity problem existed in the data ($\text{CN} > 30$).

We first applied variable selection methods based on AIC and BIC to generate the subspace information. AIC classified AGE, HT, and WEIGHT as significant predictors, whereas BIC assigned only AGE as significant. Two possible candidate submodels were therefore considered. For assessing the efficiency of the proposed estimators, $m = 100$ bootstrap rows were drawn with replacement $N = 1,000$ times from the complete dataset. The estimation of regression coefficients for only significant predictors selected by AIC and BIC are reported in Table 5.

Table 5 Point estimates and standard errors (in the parenthesis) for only significant coefficients selected by AIC and BIC

Criteria	Estimator							
	FM	SM	LS	SP	SE	PSE	LASSO	Ridge
AIC								
Intercept	−1.559	−1.822	−1.691	−1.634	−1.668	−1.167	−1.017	−0.933
	(1.125)	(0.766)	(0.904)	(1.008)	(0.957)	(0.954)	(0.985)	(0.963)
AGE	0.045	0.046	0.045	0.045	0.046	0.045	0.023	0.020
	(0.026)	(0.024)	(0.024)	(0.025)	(0.024)	(0.024)	(0.028)	(0.027)
HT	−0.563	−0.599	−0.581	−0.573	−0.579	−0.578	−0.127	−0.157
	(0.726)	(0.672)	(0.691)	(0.709)	(0.696)	(0.695)	(0.578)	(0.534)
WEIGHT	0.002	0.004	0.003	0.002	0.003	0.001	0.001	0.001
	(0.007)	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)	(0.004)	(0.004)
BIC								
Intercept	−1.514	−1.404	−1.459	−1.533	−1.505	−1.504	−1.008	−0.945
	(1.132)	(0.698)	(0.807)	(1.029)	(0.884)	(0.883)	(1.057)	(1.017)
AGE	0.043	0.048	0.046	0.044	0.045	0.045	0.022	0.020
	(0.027)	(0.026)	(0.025)	(0.027)	(0.026)	(0.026)	(0.030)	(0.029)

In fact, the true parameter values in the real data are unknown, making the exact value of Δ^{sim} to be unknown. Note that if a candidate submodel yields a most accurate prediction of response, it is indicated that $\Delta^{\text{sim}} = 0$. In contrast, when such submodel performs poorly in predicting, $\Delta^{\text{sim}} > 0$. Hence, the performance of each of the proposed estimators was evaluated using the simulated relative prediction error (SRPE), derived as

$$SRPE(\hat{\boldsymbol{\beta}}_1^{\text{FM}}; \hat{\boldsymbol{\beta}}_1^{*}) = \frac{\text{Simulated}\sum_{i=1}^{m}\left\{y_i - \exp(\boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_1^{\text{FM}})\right\}^2}{\text{Simulated}\sum_{i=1}^{m}\left\{y_i - \exp(\boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}_1^{*})\right\}^2}, \quad i = 1, 2, ..., m, \tag{34}$$

where $\hat{\boldsymbol{\beta}}_1^{*}$ is any proposed estimator. Here, the degree to which SRPE exceeds one reflects the degree of superiority of $\hat{\boldsymbol{\beta}}_1^{*}$ over $\hat{\boldsymbol{\beta}}_1^{\text{FM}}$. Since the accuracy of the subspace information were unknown, we conservatively selected $\lambda = 0.50$ and $\alpha = 0.05$, while the tuning parameters of the LASSO and ridge estimators were computed using 10-fold cross-validation. Table 6 shows the results.

Table 6 Simulated relative prediction error of SM, LS, SP, SE, PSE, LASSO, and Ridge estimators with respect to the FM estimator

Model	Estimator						
	SM	LS	SP	SE	PSE	LASSO	Ridge
AIC	2.008	1.774	1.315	1.569	1.578	1.746	1.724
BIC	3.934	3.189	2.712	2.490	2.496	2.142	2.131

As can be seen, all estimators outperformed the full model estimator. The prediction accuracy of both candidate submodel estimators was superior to that of all other estimators, suggesting that the

sparse models based on the AIC-based and BIC-based subspace information were suitable for this data, especially BIC-based. Based on the more suitable BIC-based model, it can be concluded that only the age of mother was a highly significant effect on the number of physician visits during the first trimester.

The performance of the linear shrinkage and shrinkage pretest estimators depended on the validity of the subspace information, whereas that of Stein-type estimators improved as the number of insignificant predictors increased. The LASSO and ridge estimator were dominated by the Stein-type estimators when p_2 was large. The ridge estimator had a high degree of precision, but was less efficient than the LASSO estimator because the sparsity pattern was satisfied. These results were consistent with the simulation results, in which case the obtained subspace information was correct $(\Delta^{\text{sim}} = 0)$ and p_2 increased.

6. Conclusions

In this study, we considered estimators based on the preliminary test, Stein-type, and penalized likelihood strategies for the parameter estimation of a Poisson regression model under restricted subspace information. Their asymptotic distributional bias and risk were derived and discussed. A Monte Carlo simulation was implemented to support the theoretical analysis. Further, the predictive performance of the proposed estimators was studied using a real application.

From these theoretical and numerical findings, the full model estimation suffered from an overfitting, as too many confounding predictors are retained in the model. However, the unreliable subspace information consequently resulted in submodel estimator becoming inefficient. The proposed estimators showed higher performance than the submodel estimators when the information was unreliable. The performance of the linear shrinkage and shrinkage pretest estimators depended on either the degree of confidence in the subspace information or the significance level, outperforming the submodel estimator when the information was unreliable. They were also shown to work better than the estimators based on Stein-type shrinkage strategy when the information was either true or nearly true.

Regardless of whether the subspace information was trustworthy or not, the Stein-type shrinkage-based estimator proved superior to the full model estimator, especially in its truncated version. They outperformed the other estimators in some part of the parameter space. The LASSO estimator was superior to the Stein-type shrinkage-based estimators only when the number of insignificant predictors was small. The ridge estimator performed poorly with sparse models and uncorrelated predictors, however, it performed well when the predictors were highly correlated.

When the accuracy of the subspace information is unknown, it is safe to use the preliminary test and Stein-type estimation strategies, given their superior performance. It would be interesting to extend our approach to high-dimensional data $(n \ll p)$, though we leave this for further study.

Acknowledgments

The research work of Professor Syed Ejaz Ahmed was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada. The first author gratefully acknowledges the financial support provided by Thammasat University (TU) under the TU Research Scholar.

References

Ahmed SE. Penalty, shrinkage and pretest strategies: variable selection and estimation. New York: Springer; 2014.

Ahmed SE, Hussein A, Al-Momani M. Efficient estimation for the conditional autoregressive model. J Stat Comput Simul. 2015; 85(13): 2569-2581.

Al-Kandari NM, Buhamra SS, Ahmed SE. Testing and merging information for effect size estimation. J Appl Stat. 2007; 34(1): 47-60.

Al-Momani M, Hussein AA, Ahmed SE. Penalty and related estimation strategies in the spatial error model. Stat Neerl, 2017; 71(1): 4-30.

Arashi M, Norouzirad M, Ahmed SE, Yüzbasi B. Rank-based Liu regression. Comput Stat, 2018; 33(3), 1525-1561.

Cameron AC, Trivedi PK. Regression analysis of count data, New York: Cambridge university press; 2013.

Gao X, Ahmed SE, Feng Y. Post selection shrinkage estimation for high-dimensional data analysis. Appl Stoch Models Bus Ind. 2017; 33(2): 97-120.

Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970; 12(1): 55-67.

Hosmer DW, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons; 2000.

Hossain S, Ahmed SE. Shrinkage and penalty estimators of a Poisson regression model. Aust NZ J Stat. 2012; 54(3):359-373.

Judge GG, Bock M. The statistical implications of pre-test and stein-rule estimators in econometrics. Amsterdam: North-Holland; 1978.

Mansson K, Shukur G. A Poisson ridge regression estimator. Econ Model. 2011; 28(4): 1475-1481.

Myers RH, Montgomery DC, Vining GG, Robinson TJ. Generalized linear models with applications in engineering and the sciences, New York: John Wiley & Sons; 2012.

R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2015.

Reangsephet O, Lisawadi S, Ahmed SE. Improving Estimation of Regression Parameters in Negative Binomial Regression Model. In: Xu J, Cooke FL, Gen M, Ahmed SE, editors. ICMSEM2018: Proceedings of the Twelfth International Conference on Management Science and Engineering Management; 2018 Aug 1-4; Australia. Springer; 2018. pp.265-275.

Santos JA, Neves MM. A local maximum likelihood estimator for Poisson regression. Metrika. 2008; 68(3): 257-270.

Yüzbasi B, Ahmed SE. Shrinkage ridge regression estimators in high-dimensional linear models. In: Xu J, Nickel S, Machado VC, Hajiyev A, editors. ICMSEM2015: Proceedings of the Ninth International Conference on Management Science and Engineering Management; 2015 Jul 21-23; Germany. Springer; 2015. pp.793-807.

Yüzbasi B, Ahmed SE, Aydin D. Ridge-type pretest and shrinkage estimations in partially linear models. Stat Pap. 2017a; 1-30.

Yüzbasi B, Arashi M, Ahmed SE. Shrinkage estimation strategies in generalized ridge regression models under low/high-dimension regime. arXiv preprint arXiv:1707.02331. 2017b.

Appendices

Appendix A. Proof of Lemma 3

Under the sequence of local alternative $H_{(n)}$, the asymptotic distribution of $\boldsymbol{W}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \boldsymbol{\beta}_1)$ and $\boldsymbol{Z}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_2^{\mathrm{FM}} - \boldsymbol{\beta}_2)$ are obtained by using Theorem 1 and we have the covariance between $\boldsymbol{W}$ and $\boldsymbol{Z}$ is given by $Cov(\boldsymbol{W}, \boldsymbol{Z}^{\mathrm{T}}) = -\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1}$.

Consider $\boldsymbol{X}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}} - \boldsymbol{\beta}_1) = \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} + \mathbf{V}_{11}^{-1}\mathbf{V}_{12}(\hat{\boldsymbol{\beta}}_2^{\mathrm{FM}} - \boldsymbol{\beta}_2) - \boldsymbol{\beta}_1) = \boldsymbol{W}_n + \mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{Z}_n$, which is a linear combination of $\boldsymbol{W}_n$ and $\boldsymbol{Z}_n$. Therefore, by Slutsky's theorem, as $n \to \infty$,

$$\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \sim N_{p_1}(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta}, \mathbf{V}_{11.2}^{-1} - \mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}).$$

Similarly, $\boldsymbol{Y}_n$ can be written as follows:

$$\boldsymbol{Y}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \boldsymbol{\beta}_1 + \boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}) = -\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{Z}_n,$$

which is a linear function of $\boldsymbol{Z}_n$. Again, by Slutsky's theorem, as $n \to \infty$, we get $\boldsymbol{Y}_n \xrightarrow{D} \boldsymbol{Y} \sim N_{p_1}(-\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta}, \mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1})$. The covariance between $\boldsymbol{W}$ and $\boldsymbol{Y}$ is derived by $Cov(\boldsymbol{W}, \boldsymbol{Y}^{\mathrm{T}}) = Cov(\boldsymbol{W}, \boldsymbol{Z}^{\mathrm{T}}(-\mathbf{V}_{11}^{-1}\mathbf{V}_{12})^{\mathrm{T}}) = Cov(\boldsymbol{W}, \boldsymbol{Z}^{\mathrm{T}})(-\mathbf{V}_{11}^{-1}\mathbf{V}_{12}) = \mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}$.

Appendix B. Proof of Theorem 2

Using Lemmas 1 to 3, under the sequence of local alternative $H_{(n)}$, we get

$$B(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}}) = \lim_{n\to\infty} E\left[\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \boldsymbol{\beta}_1)\right] = \lim_{n\to\infty} E[\boldsymbol{W}_n] = E[\boldsymbol{W}] = \mathbf{0},$$

$$B(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}) = \lim_{n\to\infty} E\left[\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}} - \boldsymbol{\beta}_1)\right] = \lim_{n\to\infty} E[\boldsymbol{X}_n] = E[\boldsymbol{X}] = \mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta},$$

$$\begin{aligned} B(\hat{\boldsymbol{\beta}}_1^{\mathrm{LS}}) &= \lim_{n\to\infty} E\left[\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{LS}} - \boldsymbol{\beta}_1)\right] = \lim_{n\to\infty} E\left[\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \lambda(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}) - \boldsymbol{\beta}_1)\right] \\ &= \lim_{n\to\infty} E[\boldsymbol{W}_n - \lambda\boldsymbol{Y}_n] = E[\boldsymbol{W}] - \lambda E[\boldsymbol{Y}] \\ &= \lambda\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta}, \end{aligned},$$

$$\begin{aligned} B(\hat{\boldsymbol{\beta}}_1^{\mathrm{SP}}) &= \lim_{n\to\infty} E\left[\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SP}} - \boldsymbol{\beta}_1)\right] = \lim_{n\to\infty} E\left[\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \lambda(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \hat{\boldsymbol{\beta}}_1^{\mathrm{SM}})I(D_n \le d_{n,\alpha}) - \boldsymbol{\beta}_1)\right] \\ &= \lim_{n\to\infty} E[\boldsymbol{W}_n - \lambda\boldsymbol{Y}_n I(D_n \le d_{n,\alpha})] = E[\boldsymbol{W}] - \lambda E[\boldsymbol{Y} I(\chi^2_{p_2}(\Delta) \le \chi^2_{p_2,\alpha})] \\ &= \lambda E[I(\chi^2_{p_2}(\Delta) \le \chi^2_{p_2,\alpha})]\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta} \\ &= \lambda\psi_{p_2+2}(\chi^2_{p_2,\alpha}; \Delta)\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta}. \end{aligned}$$

Appendix C. Proof of Theorem 3

To verify ADR expressions, we first derive the asymptotic mean squared error matrix of the proposed estimators. Using Lemmas 1 to 3, under the sequence of local alternative $H_{(n)}$, we have

$$MSE(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}}) = \lim_{n\to\infty} E\left[\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \boldsymbol{\beta}_1)(\hat{\boldsymbol{\beta}}_1^{\mathrm{FM}} - \boldsymbol{\beta}_1)^{\mathrm{T}}\right] = \lim_{n\to\infty} E[\boldsymbol{W}_n\boldsymbol{W}_n^{\mathrm{T}}] = E[\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}] = Var[\boldsymbol{W}] = \mathbf{V}_{11.2}^{-1},$$

$$\begin{aligned} MSE(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}}) &= \lim_{n\to\infty} E\left[\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}} - \boldsymbol{\beta}_1)(\hat{\boldsymbol{\beta}}_1^{\mathrm{SM}} - \boldsymbol{\beta}_1)^{\mathrm{T}}\right] = \lim_{n\to\infty} E[\boldsymbol{X}_n\boldsymbol{X}_n^{\mathrm{T}}] = E[\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}] \\ &= Var[\boldsymbol{X}] + E[\boldsymbol{X}]E[\boldsymbol{X}^{\mathrm{T}}] \\ &= \mathbf{V}_{11.2}^{-1} - \mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1} + (\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta})(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta})^{\mathrm{T}}, \end{aligned}$$

$$
\begin{aligned}
MSE(\hat{\boldsymbol{\beta}}_1^{\mathrm{LS}}) &= \lim_{n\to\infty} E\left[\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{LS}}-\boldsymbol{\beta}_1)(\hat{\boldsymbol{\beta}}_1^{\mathrm{LS}}-\boldsymbol{\beta}_1)^{\mathrm{T}}\right] \\
&= \lim_{n\to\infty} E[(\boldsymbol{W}_n-\lambda\boldsymbol{Y}_n)(\boldsymbol{W}_n-\lambda\boldsymbol{Y}_n)^{\mathrm{T}}] \\
&= E[(\boldsymbol{W}-\lambda\boldsymbol{Y})(\boldsymbol{W}-\lambda\boldsymbol{Y})^{\mathrm{T}}] \\
&= E[\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}]-2\lambda E[\boldsymbol{W}\boldsymbol{Y}^{\mathrm{T}}]+\lambda^2 E[\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}] \\
&= \mathbf{V}_{11.2}^{-1}-2\lambda\left\{Cov(\boldsymbol{W},\boldsymbol{Y}^{\mathrm{T}})+E[\boldsymbol{W}]E[\boldsymbol{Y}^{\mathrm{T}}]\right\}+\lambda^2\left\{Var[\boldsymbol{Y}]+E[\boldsymbol{Y}]E[\boldsymbol{Y}^{\mathrm{T}}]\right\} \\
&= \mathbf{V}_{11.2}^{-1}-2\lambda\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}+\lambda^2\left\{\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}+(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta})(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta})^{\mathrm{T}}\right\} \\
&= \mathbf{V}_{11.2}^{-1}-\lambda(2-\lambda)\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}+\lambda^2(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta})(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta})^{\mathrm{T}}, \\
MSE(\hat{\boldsymbol{\beta}}_1^{\mathrm{SP}}) &= \lim_{n\to\infty} E\left[\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{\mathrm{SP}}-\boldsymbol{\beta}_1)(\hat{\boldsymbol{\beta}}_1^{\mathrm{SP}}-\boldsymbol{\beta}_1)^{\mathrm{T}}\right] \\
&= \lim_{n\to\infty} E\left[(\boldsymbol{W}_n-\lambda\boldsymbol{Y}_n I(D_n\le d_{n,\alpha}))(\boldsymbol{W}_n-\lambda\boldsymbol{Y}_n I(D_n\le d_{n,\alpha}))^{\mathrm{T}}\right] \\
&= E\left[(\boldsymbol{W}-\lambda\boldsymbol{Y}I(\chi^2_{p_2}(\Delta)\le\chi^2_{p_2,\alpha}))(\boldsymbol{W}-\lambda\boldsymbol{Y}I(\chi^2_{p_2}(\Delta)\le\chi^2_{p_2,\alpha}))^{\mathrm{T}}\right] \\
&= E[\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}]-2\lambda\underbrace{E[\boldsymbol{W}\boldsymbol{Y}^{\mathrm{T}}I(\chi^2_{p_2}(\Delta)\le\chi^2_{p_2,\alpha})]}_{(A)}+\lambda^2\underbrace{E[\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}I(\chi^2_{p_2}(\Delta)\le\chi^2_{p_2,\alpha})]}_{(B)}.
\end{aligned}
$$

By using the rule of conditional expectation and Lemma 1, we have

$$
\begin{aligned}
(\mathrm{A}) &= E[\boldsymbol{W}\boldsymbol{Y}^{\mathrm{T}}I(\chi^2_{p_2}(\Delta)\le\chi^2_{p_2,\alpha})] \\
&= E\left[E[\boldsymbol{W}\boldsymbol{Y}^{\mathrm{T}}I(\chi^2_{p_2}(\Delta)\le\chi^2_{p_2,\alpha})]\,|\,\boldsymbol{Y}\right] \\
&= E\left[E[\boldsymbol{W}\,|\,\boldsymbol{Y}]\boldsymbol{Y}^{\mathrm{T}}I(\chi^2_{p_2}(\Delta)\le\chi^2_{p_2,\alpha})\right] \\
&= E\left[\{\boldsymbol{Y}+(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta})\}\boldsymbol{Y}^{\mathrm{T}}I(\chi^2_{p_2}(\Delta)\le\chi^2_{p_2,\alpha})\right] \\
&= E\left[\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}I(\chi^2_{p_2}(\Delta)\le\chi^2_{p_2,\alpha})\right]-(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta})E\left[\boldsymbol{Y}^{\mathrm{T}}I(\chi^2_{p_2}(\Delta)\le\chi^2_{p_2,\alpha})\right] \\
&= (\mathrm{B})-\psi_{p_2+2}(\chi^2_{p_2,\alpha};\Delta)(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta})(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta})^{\mathrm{T}},
\end{aligned}
$$

and

$$
\begin{aligned}
(\mathrm{B}) &= E[\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}I(\chi^2_{p_2}(\Delta)\le\chi^2_{p_2,\alpha})] \\
&= Var[\boldsymbol{Y}]E[I(\chi^2_{p_2+2}(\Delta)\le\chi^2_{p_2,\alpha})]+E[\boldsymbol{Y}]E[\boldsymbol{Y}^{\mathrm{T}}]E[I(\chi^2_{p_2+4}(\Delta)\le\chi^2_{p_2,\alpha})] \\
&= \psi_{p_2+2}(\chi^2_{p_2,\alpha};\Delta)\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}+\psi_{p_2+4}(\chi^2_{p_2,\alpha};\Delta)(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta})(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta})^{\mathrm{T}}.
\end{aligned}
$$

Substituting (B) into (A), we obtain

$$
(\mathrm{A}) = \psi_{p_2+2}(\chi^2_{p_2,\alpha};\Delta)\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}+\left\{\psi_{p_2+4}(\chi^2_{p_2,\alpha};\Delta)-\psi_{p_2+2}(\chi^2_{p_2,\alpha};\Delta)\right\}(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta})(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta})^{\mathrm{T}}.
$$

Thus,

$$
\begin{aligned}
MSE(\hat{\boldsymbol{\beta}}_1^{\mathrm{SP}}) &= \mathbf{V}_{11.2}^{-1} \\
&\quad -2\lambda\left\{\psi_{p_2+2}(\chi^2_{p_2,\alpha};\Delta)\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}+\begin{Bmatrix}\psi_{p_2+4}(\chi^2_{p_2,\alpha};\Delta)\\-\psi_{p_2+2}(\chi^2_{p_2,\alpha};\Delta)\end{Bmatrix}(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta})(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta})^{\mathrm{T}}\right\} \\
&\quad +\lambda^2\left\{\psi_{p_2+2}(\chi^2_{p_2,\alpha};\Delta)\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}+\psi_{p_2+4}(\chi^2_{p_2,\alpha};\Delta)(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta})(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta})^{\mathrm{T}}\right\} \\
&= \mathbf{V}_{11.2}^{-1}-\lambda(2-\lambda)\psi_{p_2+2}(\chi^2_{p_2,\alpha};\Delta)\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22.1}^{-1}\mathbf{V}_{21}\mathbf{V}_{11}^{-1} \\
&\quad +\lambda\left\{2\psi_{p_2+2}(\chi^2_{p_2,\alpha};\Delta)\text{-}(2-\lambda)\psi_{p_2+4}(\chi^2_{p_2,\alpha};\Delta)\right\}(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta})(\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\boldsymbol{\delta})^{\mathrm{T}}.
\end{aligned}
$$

Consequently, it is easy to verify Theorem 3 by using (27) and the above asymptotic mean squared error matrix expressions.