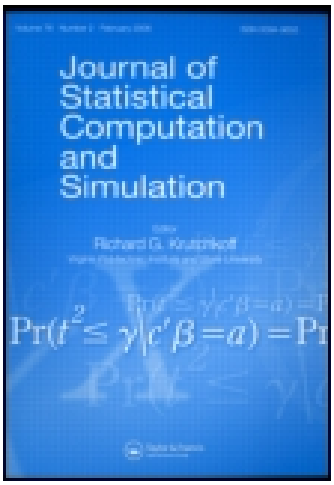

Journal of Statistical Computation and Simulation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gscs20

Model selection and parameter estimation of a multinomial logistic regression model

Shakhawat Hossain[a], S. Ejaz Ahmed[b] & Hatem A. Howlader[a]

[a] Department of Mathematics and Statistics, University of Winnipeg, Winnipeg, MB, Canada R3B 2E9

[b] Department of Mathematics, Brock University, St. Catharines, Ontario, Canada L2S 3A1

Published online: 26 Nov 2012.

To cite this article: Shakhawat Hossain, S. Ejaz Ahmed & Hatem A. Howlader (2014) Model selection and parameter estimation of a multinomial logistic regression model, Journal of Statistical Computation and Simulation, 84:7, 1412-1426, DOI: 10.1080/00949655.2012.746347

To link to this article: http://dx.doi.org/10.1080/00949655.2012.746347

Journal of Statistical Computation and Simulation, 2014
Vol. 84, No. 7, 1412–1426, http://dx.doi.org/10.1080/00949655.2012.746347

Model selection and parameter estimation of a multinomial logistic regression model

Shakhawat Hossain[a]*, S. Ejaz Ahmed[b] and Hatem A. Howlader[a]

[a]*Department of Mathematics and Statistics, University of Winnipeg, Winnipeg, MB, Canada R3B 2E9;*
[b]*Department of Mathematics, Brock University, St. Catharines, Ontario, Canada L2S 3A1*

(*Received 2 June 2012; final version received 31 October 2012*)

In the multinomial regression model, we consider the methodology for simultaneous model selection and parameter estimation by using the shrinkage and LASSO (least absolute shrinkage and selection operation) [R. Tibshirani, *Regression shrinkage and selection via the LASSO*, J. R. Statist. Soc. Ser. B 58 (1996), pp. 267–288] strategies. The shrinkage estimators (SEs) provide significant improvement over their classical counterparts in the case where some of the predictors may or may not be active for the response of interest. The asymptotic properties of the SEs are developed using the notion of asymptotic distributional risk. We then compare the relative performance of the LASSO estimator with two SEs in terms of simulated relative efficiency. A simulation study shows that the shrinkage and LASSO estimators dominate the full model estimator. Further, both SEs perform better than the LASSO estimators when there are many inactive predictors in the model. A real-life data set is used to illustrate the suggested shrinkage and LASSO estimators.

Keywords: shrinkage estimators; LASSO; asymptotic distributional bias and risk; Monte Carlo simulation; multinomial logistic regression; likelihood ratio test

1. Introduction

The multinomial logistic regression (MLR) is a generalization of the logistic regression for dichotomous response variables. It can be extended to any number of categories of the response variable [1–4]. At each combination of the levels of the predictor variables, the MLR model assumes that the outcomes of the categories of the response variable have a multinomial distribution. In addition to the health and life sciences, the MLR is used in econometrics, sociometrics, and other fields of application for the prediction of probabilities of polytomous response variables as a function of a set of predictor variables [5]. For example, in a social science research, the child maltreatment [6] with three categories (sexual abuse, other types of abuse, and neglect) may be related to the predictors, such as maternal age, community child poverty rate, birth year, region, sex of the child, birth order, and race or ethnicity.

In this paper, we consider the variable selection and parameter estimation for regression coefficients of an MLR model when some of the parameters are in a linear subspace, that is, when some of the predictors may be inactive for the response of interest. In a situation like this, the practice has been to use some prior information about the inactive predictors in the full model

*Corresponding author. Email: sh.hossain@uwinnipeg.ca, mhossain2005@gmail.com

to produce a candidate reduced model. The shrinkage estimation method, inspired by Stein's results, combines the full model estimator (FE) and reduced model estimator (RE) in an optimal way. This method has been found to be more accurate on average (in terms of mean-squared error) than any other classical method. A large number of studies have been conducted in the area of shrinkage estimation method (see [7–11] and others).

The motivation of the present work is to extend the shrinkage estimation method for variable selection and the parameter estimation for the MLR model by combining the information from recent literature on sparsity patterns and then compare the resulting estimator to a version of the penalty estimator, the so-called LASSO (least absolute shrinkage and selection operation). A major challenge in model selection is to decide which predictors, among many active ones, are to be included in the model. It is customary to use the stepwise and subset selection methods for the purpose. These selection processes are discrete in nature and they only consider whether a predictor is in or out of the model. Another method incorporates the effects of shrinkage by adding some bias to the estimator effectively reducing its variance. This method, the so-called shrinkage method, uses a two-step approach if the prior information on the predictors is not available. In the first step, the active predictors are selected based on one of the above selection methods and the traditional model selection criteria, such as Akaike information criterion and Bayesian information criterion. Next, a constraint on the full parameter space is placed by using the remaining inactive predictors. In the MLR model, let $\boldsymbol{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J)'$ be an unknown $pJ \times 1$ vector of parameters, p is the number of predictor variables, J (> 2) is the number of categories of the response variable, and $\boldsymbol{\beta}_j = (\beta_{1j}, \beta_{2j}, \ldots, \beta_{pj})', j = 1, 2, \ldots, J$. We would like to find a subspace that satisfies a set of $p_2 J$ linear restrictions

$$\boldsymbol{LB} = \boldsymbol{h}, \tag{1}$$

where $\boldsymbol{L}$ is $p_2 J \times pJ$ matrix of rank $p_2 J \leq pJ$, and $\boldsymbol{h}$ is a $p_2 J$ vector of constants. Since the rank of $\boldsymbol{L}$ is p_2, the p_2 equations in Equation (1) do not contain any redundant information about $\boldsymbol{B}$ for each category j. In the second step, we combine the FE and RE in an optimal way in order to achieve an improved estimator for the remaining active predictors. This approach can be implemented for moderate values of p_2. For large p_2, one can use penalty methods, such as LASSO and its variants, to obtain sparse models and then apply the shrinkage estimation strategy to obtain efficient estimators.

The LASSO for linear regression models has become a useful and well-studied approach for simultaneous execution of both parameter estimation and variable selection. It can be easily applied to other settings, for example, logistic regression [12,13]. Remarkable progress has been made in recent years in developing efficient and consistent algorithms for the LASSO with good features even in high-dimensional settings. Meier and Bühlmann [13] implemented the group LASSO for logistic regression and established the convergence rate of the estimator; however, they did not demonstrate the benefit of the group LASSO over the L_1-penalty. Recently, the coordinate descent algorithm was developed by Friedman *et al.* [14] for penalized MLR model and was shown that this algorithm efficiently calculates the regularization parameters. For further aspects of LASSO in regard to its variable selection, estimation and prediction properties, see, for example [15–19].

The rest of the paper is organized as follows. The model selection and estimation strategies are developed in Section 2. The asymptotic results of shrinkage estimators (SEs) for the MLR model are described in Section 3. In Section 4, we use a simulation study to evaluate the relative performance of the estimators with respect to FE. Section 5 illustrates the proposed estimation strategies with a real data application. Concluding remarks are given in Section 6.

2. Model selection and estimation strategies

Let the response variable $Y \in \{0, 1, 2, \ldots, J\}$ have $J + 1$ possible categories. In this paper, we treat the category with label 0 as the reference category. However, any category can be used as a reference category. For each category j $(1, 2, \ldots, J)$, there is a regression function in which the log odds of response in category j, relative to category 0, is a linear function of regression parameters $\boldsymbol{\beta_j} = (\beta_{1j}, \beta_{2j}, \ldots, \beta_{pj})'$ and a predictor vector $\boldsymbol{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$. The MLR model can then be written as

$$\log\left(\frac{\pi_{ij}}{\pi_{i0}}\right) = \log\frac{P(Y = j|\boldsymbol{x}_i)}{P(Y = 0|\boldsymbol{x}_i)} = \boldsymbol{x}_i'\boldsymbol{\beta}_j, \tag{2}$$

where $i = 1, 2, \ldots, n$. The J logits $\log(P(Y = 1|\boldsymbol{x}_i)/P(Y = 0|\boldsymbol{x}_i))$, $\log(P(Y = 2|\boldsymbol{x}_i)/P(Y = 0|\boldsymbol{x}_i)), \ldots, \log(P(Y = J|\boldsymbol{x}_i)/P(Y = 0|\boldsymbol{x}_i))$ given in Equation (2) determine the response probabilities $P(Y = 1|\boldsymbol{x}_i)$, $P(Y = 2|\boldsymbol{x}_i), \ldots, P(Y = J|\boldsymbol{x}_i)$ uniquely since $\sum_{j=1}^{J} P(Y = j|\boldsymbol{x}_i) = 1$. Therefore, only J response categories and parameter vectors will have to be specified. The inverse formulae are

$$\pi_{ij} = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta_j})}{1 + \sum_{k=1}^{J}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_k)}$$

and

$$\pi_{i0} = \frac{1}{1 + \sum_{k=1}^{J}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_k)}.$$

Let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iJ})'$ be a $J \times 1$ column vector of responses for the observation i, with the corresponding $J \times 1$ column vector of probabilities $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iJ})'$. Then, the likelihood function is given by

$$\begin{aligned} L(\boldsymbol{B}) &= \prod_{i=1}^{n} \pi_{i1}^{y_{i1}}\pi_{i2}^{y_{i2}}\cdots\pi_{iJ}^{y_{iJ}}(1 - \pi_{i1} - \cdots - \pi_{iJ})^{(1-y_{i1}-\cdots-y_{iJ})} \\ &= \prod_{i=1}^{n}\left(\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1)}{1 + \sum_{k=1}^{J}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_k)}\right)^{y_{i1}} \times \left(\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_2)}{1 + \sum_{k=1}^{J}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_k)}\right)^{y_{i2}} \\ &\quad \times \cdots \left(\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_J)}{1 + \sum_{k=1}^{J}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_k)}\right)^{y_{iJ}}. \end{aligned}$$

The corresponding log-likelihood is then given by

$$\begin{aligned} l(\boldsymbol{B}) &= l(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J) \\ &= \sum_{i=1}^{n}[(\boldsymbol{x}_i'\boldsymbol{\beta}_1)y_{i1} + (\boldsymbol{x}_i'\boldsymbol{\beta}_2)y_{i2} + \cdots + (\boldsymbol{x}_i'\boldsymbol{\beta}_J)y_{iJ}] - \sum_{i=1}^{n}\log\left[1 + \sum_{k=1}^{J}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_k)\right]. \end{aligned} \tag{3}$$

The FE of the parameter vector $\boldsymbol{\beta}_j$ for the category j is obtained from observations $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ by solving the following score equations of the log-likelihood:

$$S(\boldsymbol{B}) = \frac{\partial l}{\partial \boldsymbol{\beta}_j} = \boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{\Pi}) = \boldsymbol{0}, \tag{4}$$

where $\partial l/\partial\boldsymbol{\beta}_j$ is a $p \times 1$ vector, $\boldsymbol{X}' = \boldsymbol{X}_D \otimes \boldsymbol{I}_J$, $\boldsymbol{X}_D' = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$, $\boldsymbol{X}$ is a $nJ \times pJ$ matrix, $\otimes$ is the Kronecker product matrix operator, $\boldsymbol{I}_J$ is a $J \times J$ identity matrix, $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)'$, and $\boldsymbol{\Pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_n)'$.

2.1. *The FE and candidate RE*

The FE $\hat{\boldsymbol{B}}^{\mathrm{FE}}$ of $\boldsymbol{B}$ can be obtained by maximizing the log-likelihood function (3) with respect to $\boldsymbol{\beta}_j$, which is typically done by solving Equation (4). We use the Newton–Raphson iterative method to solve these nonlinear score equations in $\boldsymbol{\beta}_j$. Under usual regularity conditions [20], it can be shown that $\hat{\boldsymbol{B}}^{\mathrm{FE}}$ is consistent and asymptotically normal with a $pJ \times pJ$ variance–covariance matrix $(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}$, where $\boldsymbol{V}$ is an $nJ \times nJ$ block diagonal matrix with n times $J \times J$ blocks with the ijth element of $\boldsymbol{V}$, $\pi_{ij}(\delta_{jl} - \pi_{il}), l = 1, 2, \ldots, p$, with indicator $\delta_{jl} = 1$ if $j = l$ and $\delta_{jl} = 0$ if $j \neq l$.

The candidate RE, $\hat{\boldsymbol{B}}^{\mathrm{RE}}$ of $\boldsymbol{B}$ can be obtained by maximizing the log-likelihood function (3) under the linear restriction $\boldsymbol{L}\boldsymbol{B} - \boldsymbol{h} = \boldsymbol{0}$. The RE behaves better than the FE when the null hypothesis (1) is true but a different picture may emerge for possible departure from the null hypothesis; the RE may be biased, inefficient, and even inconsistent [21].

2.2. *Shrinkage and positive SEs*

The SE is a weighted average of the FE and RE and is defined as

$$\hat{\boldsymbol{B}}^{\mathrm{SE}} = \hat{\boldsymbol{B}}^{\mathrm{RE}} + (1 - (p_2 J - 2)\Psi^{-1})(\hat{\boldsymbol{B}}^{\mathrm{FE}} - \hat{\boldsymbol{B}}^{\mathrm{RE}}), \quad p_2 J \geq 3,$$

where the weight Ψ is defined as

$$\begin{aligned}\Psi &= 2[l(\hat{\boldsymbol{B}}^{\mathrm{FE}}) - l(\hat{\boldsymbol{B}}^{\mathrm{RE}})],\\ &= (\boldsymbol{L}\hat{\boldsymbol{B}}^{\mathrm{FE}} - \boldsymbol{h})'[\boldsymbol{L}(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{L}']^{-1}(\boldsymbol{L}\hat{\boldsymbol{B}}^{\mathrm{FE}} - \boldsymbol{h}) + o_p(1).\end{aligned}$$

The weight Ψ is used as a test statistic to test the null hypothesis $H_0 : \boldsymbol{L}\boldsymbol{B} = \boldsymbol{h}$. When H_0 is true, the distribution of Ψ converges to χ^2 with $p_2 J$ degrees of freedom as $n \to \infty$.

The major problem with the estimator $\hat{\boldsymbol{B}}^{\mathrm{SE}}$ is that it may have a different sign than the FE, $\hat{\boldsymbol{B}}^{\mathrm{FE}}$, perhaps due to over-shrinking. The change of sign certainly would make the researchers rather uncomfortable. To avoid the over-shrinking inherent in $\hat{\boldsymbol{B}}^{\mathrm{SE}}$, we define a positive shrinkage estimator (PSE) which will control the possible over-shrinking problem. It can be defined as

$$\hat{\boldsymbol{B}}^{\mathrm{PSE}} = \hat{\boldsymbol{B}}^{\mathrm{RE}} + (1 - (p_2 J - 2)\Psi^{-1})^{+}(\hat{\boldsymbol{B}}^{\mathrm{FE}} - \hat{\boldsymbol{B}}^{\mathrm{RE}}),$$

where $z^{+} = \max(0, z)$.

2.3. *LASSO estimator*

Tibshirani [22] originally proposed LASSO for linear regression models, which was subsequently adapted to work with categorical variables. Krishnapuram *et al.* [19] applied the L_1 penalty of the LASSO version in the MLR model by replacing the residual sum of squares by the corresponding negative log-likelihood function. The penalized likelihood function is given by

$$\hat{\boldsymbol{B}}_{\lambda}^{\mathrm{LASSO}} = \underset{(\boldsymbol{B})}{\operatorname{argmin}} \frac{1}{n}\left[-l(\boldsymbol{B}) + \lambda \sum_{j=1}^{J} \|\boldsymbol{B}\|_1\right],$$

where $\lambda \geq 0$ controls the amount of penalty applied to the estimate. Setting λ to zero produces the penalized to unpenalized likelihood function, which we minimize to obtain the maximum-likelihood estimates. Conversely, for moderate values of λ, LASSO shrinks the coefficients

towards zero, and some of the coefficients may even end up being exactly zero. In addition to shrinking coefficients and deleting the inactive predictors, LASSO gives good prediction accuracy by effectively balancing the bias and variance. The entire technique is computed by quadratic programming and λ is selected using 10-fold cross-validation. Note that the output of the LASSO resembles the shrinkage method by both shrinking and deleting coefficients. However, it is different from the shrinkage method in that it treats all the regression coefficients equally. The LASSO does not contemplate a linear subspace during optimization.

3. Asymptotic results and comparison

In this section, we present the asymptotic distributions of the estimators and the necessary test statistics. This will facilitate the derivation of the asymptotic distributional bias (ADB) and asymptotic distributional risk (ADR) of the estimators of $\boldsymbol{B}$. Under a fixed alternative and for any estimator $\hat{\boldsymbol{B}}^*$, the asymptotic distribution of $(\hat{\boldsymbol{B}}^* - \boldsymbol{B})/\text{se}(\hat{\boldsymbol{B}}^*)$ is equivalent to that of $(\hat{\boldsymbol{B}}^{\text{FE}} - \boldsymbol{B})/\text{se}(\hat{\boldsymbol{B}}^{\text{FE}})$. This suggests that in an asymptotic setup, there is not much to investigate under a fixed alternative such as $\boldsymbol{LB} \neq \boldsymbol{h}$. Therefore, to obtain meaningful asymptotics, a class of local alternatives, $K_{(n)}$, is considered:

$$K_{(n)} : \boldsymbol{LB} = \boldsymbol{h} + \frac{\boldsymbol{\delta}}{\sqrt{n}}, \tag{5}$$

where $\boldsymbol{\delta}$ is a $p_2 J$ vector of constants. Each element of $\boldsymbol{\delta}/\sqrt{n}$ is a measure of how much the local alternatives $K_{(n)}$ differ from the subspace (1). For such local alternatives, the SE, PSE, and RE may not be asymptotically unbiased. In order to investigate the asymptotic risk properties of the estimators, we consider the following quadratic loss function:

$$\mathcal{L}(\hat{\boldsymbol{B}}^*, \boldsymbol{B}; \boldsymbol{Q}) = [\sqrt{n}(\hat{\boldsymbol{B}}^* - \boldsymbol{B})]' \boldsymbol{Q} [\sqrt{n}(\hat{\boldsymbol{B}}^* - \boldsymbol{B})], \tag{6}$$

where $\boldsymbol{Q}$ is a $p \times p$ positive semi-definite weight matrix. Without loss of generality, a common choice of $\boldsymbol{Q}$ is the identity matrix, which we will use in Section 4.

Using the distribution of $\sqrt{n}(\hat{\boldsymbol{B}}^* - \boldsymbol{B})$ and taking expectations on both sides of Equation (6), we obtain the expected loss called the ADR, which can be shown to be

$$\mathbf{ADR}_n^0(\hat{\boldsymbol{B}}^*, \boldsymbol{B}; \boldsymbol{Q}) = \text{tr}(\boldsymbol{Q}\hat{\boldsymbol{\Sigma}}_n), \tag{7}$$

where $\hat{\boldsymbol{\Sigma}}_n$ is the asymptotic covariance matrix of $\sqrt{n}(\hat{\boldsymbol{B}}^* - \boldsymbol{B})$. Under the limit,

$$\lim_{n \to \infty} \hat{\boldsymbol{\Sigma}}_n = \boldsymbol{\Sigma},$$

Equation (7) takes the form

$$\mathbf{ADR}_n^0(\hat{\boldsymbol{B}}^*, \boldsymbol{B}; \boldsymbol{Q}) = \mathbf{ADR}^0(\hat{\boldsymbol{B}}^*, \boldsymbol{B}; \boldsymbol{Q}) = \text{tr}(\boldsymbol{Q}\boldsymbol{\Sigma}),$$

which is termed as the asymptotic risk. In our setup, we denote the distribution of $\sqrt{n}(\hat{\boldsymbol{B}}^* - \boldsymbol{B})$ by $G_n(\boldsymbol{y}), \boldsymbol{y} \in \boldsymbol{R}^p$. Suppose that $G_n \to G$ (at all points of continuity), as $n \to \infty$, and let Σ_G be the dispersion matrix of G. Then, the ADR of $\boldsymbol{B}^*$ becomes

$$\mathbf{ADR}^0(\hat{\boldsymbol{B}}^*, \boldsymbol{B}; \boldsymbol{Q}) = \text{tr}(\boldsymbol{Q}\boldsymbol{\Sigma}_G).$$

Note that if $\sqrt{n}(\hat{\boldsymbol{B}}^* - \boldsymbol{B})$ converges to the second moment, then the ADR is indeed the asymptotic risk. However, this stronger mode of convergence may be difficult to show analytically, especially

for SE and PSE as the explicit form of the variance–covariance matrix cannot be found, and thus, we will work with the ADR in our ongoing discussions.

The SEs are, in general, biased. However, the bias is accompanied by reduction in risk, and hence, the SEs do not have any serious impact on risk assessment. In this vein, we define the ADB as

$$\mathbf{ADB}_n^0(\hat{\boldsymbol{B}}^*, \boldsymbol{B}) = E[\sqrt{n}(\hat{\boldsymbol{B}}^* - \boldsymbol{B})],$$

which for any estimator is

$$\lim_{n\to\infty} \int \cdots \int \boldsymbol{y}\, \mathrm{d}G_n(\boldsymbol{y}) = \int \cdots \int \boldsymbol{y}\, \mathrm{d}G(\boldsymbol{y}).$$

Two central key results to the study of ADR and ADB of the SE and PSE are given in the following theorem.

THEOREM 3.1 *Under the local alternatives $K_{(n)}$ in Equation* (5) *and the usual regularity conditions*

(1) $\sqrt{n}(\boldsymbol{L}\hat{\boldsymbol{B}}^{\mathrm{FE}} - \boldsymbol{h}) \xrightarrow{\mathcal{L}} N(\boldsymbol{\delta}, \boldsymbol{L}(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{L}')$ *as* $n \to \infty$.
(2) *The test statistic* Ψ *converges to a non-central chi-squared distribution* $\chi^2_{Jp_2}(\boldsymbol{\Delta})$ *with* Jp_2 *degrees of freedom and non-centrality parameter*

$$\boldsymbol{\Delta} = \boldsymbol{\delta}'(\boldsymbol{L}(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{L}')^{-1}\boldsymbol{\delta},$$

where $\boldsymbol{\Delta}$ *is a* $J \times J$ *matrix of constants.*

With this theorem in place, we are in a position to use the results for the MLR model and give the main results of this subsection. We present (without proof) the results on SE and PSE. The proofs are similar to those given in [9].

THEOREM 3.2 *Under the class of local alternatives $K_{(n)}$ and the results of Theorem 3.1, the ADB of the positive SE is*

$$\begin{aligned}\mathbf{ADB}(\hat{\boldsymbol{B}}^{\mathrm{PSE}}) = \mathbf{ADB}(\hat{\boldsymbol{B}}^{\mathrm{SE}}) &- (Jp_2 - 2)\mathbf{ADB}(\hat{\boldsymbol{B}}^{\mathrm{RE}}) \\ &\times E[\chi^{-2}_{Jp_2+2}(\boldsymbol{\Delta})I(\chi^2_{Jp_2+2}(\boldsymbol{\Delta}) < (Jp_2 - 2))] \\ &+ \mathbf{ADB}(\hat{\boldsymbol{B}}^{\mathrm{RE}})H_{Jp_2+2}(Jp_2 - 2, \boldsymbol{\Delta}),\end{aligned}$$

where $H_{Jp_2+2}(Jp_2 - 2, \boldsymbol{\Delta})$ *is the distribution function of the* χ^2 *distribution with non-centrality parameter* $\boldsymbol{\Delta}$, $\mathbf{ADB}(\hat{\boldsymbol{B}}^{\mathrm{SE}}) = -(Jp_2 - 2)\boldsymbol{M}\boldsymbol{\delta}E(\chi^{-2}_{Jp_2+2}(\boldsymbol{\Delta}))$, $\mathbf{ADB}(\hat{\boldsymbol{B}}^{\mathrm{RE}}) = -\boldsymbol{M}\boldsymbol{\delta}$, *and* $\boldsymbol{M} = (\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{L}'[\boldsymbol{L}(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{L}']^{-1}$.

The constant matrix $\boldsymbol{\delta}$ is common to the ADBs of $\hat{\boldsymbol{B}}^{\mathrm{RE}}$, $\hat{\boldsymbol{B}}^{\mathrm{SE}}$, and $\hat{\boldsymbol{B}}^{\mathrm{PSE}}$; the ADBs thus differ only by a constant factor $\boldsymbol{\Delta}$. It then suffices to compare $\boldsymbol{\Delta}$ only. It is clear that the ADB of the $\hat{\boldsymbol{B}}^{\mathrm{RE}}$ is an unbounded function of $\boldsymbol{\Delta}$. On the other hand, the ADBs of both $\hat{\boldsymbol{B}}^{\mathrm{SE}}$ and $\hat{\boldsymbol{B}}^{\mathrm{PSE}}$ are bounded in $\boldsymbol{\Delta}$. Since $E(\chi^{-2}_{Jp_2+2}(\boldsymbol{\Delta}))$ is a decreasing function of $\boldsymbol{\Delta}$, the ADB of $\hat{\boldsymbol{B}}^{\mathrm{SE}}$ starts from the origin, increases to a maximum, and then decreases towards $\mathbf{0}$ as $\boldsymbol{\Delta} > \mathbf{0}$. The characteristics of $\hat{\boldsymbol{B}}^{\mathrm{PSE}}$ are similar to those of $\hat{\boldsymbol{B}}^{\mathrm{SE}}$.

THEOREM 3.3 *Under the class local alternatives $K_{(n)}$ and the results of Theorem* 3.1, *the ADR functions of the positive SE are*

$$\begin{aligned}\mathbf{ADR}(\hat{\boldsymbol{B}}^{\mathrm{PSE}};\boldsymbol{Q}) &= \mathbf{ADR}(\hat{\boldsymbol{B}}^{\mathrm{SE}};\boldsymbol{Q})\\ &\quad - \boldsymbol{\delta}'(\boldsymbol{M}'\boldsymbol{Q}\boldsymbol{M})\boldsymbol{\delta}E[(1-(Jp_2-2)\chi^{-2}_{Jp_2+4}(\boldsymbol{\Delta}))^2 I(\chi^2_{Jp_2+4}(\boldsymbol{\Delta})<Jp_2-2)]\\ &\quad - \mathrm{trace}[\boldsymbol{Q}\boldsymbol{M}\boldsymbol{L}\boldsymbol{B}^{-1}]E[(1-(Jp_2-2)\chi^{-2}_{Jp_2+2}(\boldsymbol{\Delta}))^2 I(\chi^2_{Jp_2+4}(\boldsymbol{\Delta})<Jp_2-2)]\\ &\quad + 2\boldsymbol{\delta}'(\boldsymbol{M}'\boldsymbol{Q}\boldsymbol{M})\boldsymbol{\delta}E[(1-(Jp_2-2)\chi^{-2}_{Jp_2+4}(\boldsymbol{\Delta}))I(\chi^2_{Jp_2+4}(\boldsymbol{\Delta})<Jp_2-2)],\end{aligned}$$

where

$$\begin{aligned}\mathbf{ADR}(\hat{\boldsymbol{B}}^{\mathrm{SE}};\boldsymbol{Q}) &= \mathbf{ADR}(\hat{\boldsymbol{B}}^{\mathrm{FE}};\boldsymbol{Q}) - 2(Jp_2-2)\mathrm{trace}[\boldsymbol{Q}\boldsymbol{M}\boldsymbol{L}\boldsymbol{B}^{-1}]\{2E(\chi^{-2}_{Jp_2+2}(\boldsymbol{\Delta}))\\ &\quad - (Jp_2-2)E(\chi^{-4}_{Jp_2+2}(\boldsymbol{\Delta}))\} + (Jp_2-2)\boldsymbol{\delta}'(\boldsymbol{M}'\boldsymbol{Q}\boldsymbol{M})\boldsymbol{\delta}\{2E(\chi^{-2}_{Jp_2+2}(\boldsymbol{\Delta}))\\ &\quad - 2E(\chi^{-4}_{Jp_2+2}(\boldsymbol{\Delta})) + (Jp_2-2)E(\chi^{-4}_{Jp_2+4}(\boldsymbol{\Delta}))\}\end{aligned}$$

and $\mathbf{ADR}(\hat{\boldsymbol{B}}^{\mathrm{RE}};\boldsymbol{Q}) = \mathbf{ADR}(\hat{\boldsymbol{B}}^{\mathrm{FE}};\boldsymbol{Q}) - \mathrm{trace}[\boldsymbol{Q}\boldsymbol{M}\boldsymbol{L}(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}] + \boldsymbol{\delta}'(\boldsymbol{M}'\boldsymbol{Q}\boldsymbol{M})\boldsymbol{\delta}$, *where* $\mathbf{ADR}(\hat{\boldsymbol{B}}^{\mathrm{FE}};\boldsymbol{Q}) = \mathrm{trace}[\boldsymbol{Q}(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}]$. *The outline of proof is provided in the appendix.*

For a suitable of choice of $\boldsymbol{Q}$, when $\boldsymbol{\Delta}$ moves away from $\mathbf{0}$ vector, the risk of $\hat{\boldsymbol{B}}^{\mathrm{RE}}$ increases and becomes unbounded. This clearly indicates that the performance of $\hat{\boldsymbol{B}}^{\mathrm{RE}}$ is strongly dependent upon the validity of the restriction. The ADR of $\hat{\boldsymbol{B}}^{\mathrm{SE}}$ is smaller than or equal to the ADR of $\hat{\boldsymbol{B}}^{\mathrm{FE}}$ in the entire parameter space and the upper limit is attained when $\boldsymbol{\Delta}\to\infty$. The risk of $\hat{\boldsymbol{B}}^{\mathrm{PSE}}$ is asymptotically superior to that of $\hat{\boldsymbol{B}}^{\mathrm{SE}}$ for all components of $\boldsymbol{\Delta}>\mathbf{0}$. To compare the risks of $\hat{\boldsymbol{B}}^{\mathrm{SE}}$ and $\hat{\boldsymbol{B}}^{\mathrm{FE}}$, it can be easily shown that, under certain conditions $\mathbf{ADR}(\hat{\boldsymbol{B}}^{\mathrm{SE}};\boldsymbol{Q}) \le \mathbf{ADR}(\hat{\boldsymbol{B}}^{\mathrm{FE}};\boldsymbol{Q})$ for all $\boldsymbol{\Delta}\ge\mathbf{0}$. Hence, $\hat{\boldsymbol{B}}^{\mathrm{PSE}}$ dominates $\hat{\boldsymbol{B}}^{\mathrm{FE}}$ and we have $\mathbf{ADR}(\hat{\boldsymbol{B}}^{\mathrm{PSE}};\boldsymbol{Q}) \le \mathbf{ADR}(\hat{\boldsymbol{B}}^{\mathrm{SE}};\boldsymbol{Q}) \le \mathbf{ADR}(\hat{\boldsymbol{B}}^{\mathrm{FE}};\boldsymbol{Q})$.

4. A simulation study

In this section, we conduct a simulation study to evaluate the proposed methods and compare the risk performance (namely mean squared error (MSE)) of the estimators $\hat{\boldsymbol{B}}^{\mathrm{RE}}$, $\hat{\boldsymbol{B}}^{\mathrm{SE}}$, $\hat{\boldsymbol{B}}^{\mathrm{PSE}}$, and $\hat{\boldsymbol{B}}^{\mathrm{LASSO}}$ with respect to the FE $\hat{\boldsymbol{B}}^{\mathrm{FE}}$. Our simulation study is based on sample sizes $n=250$ and 300. Models with different numbers of covariates and a response variable with three categories are considered, $j=0$ being the reference category . The responses are generated from the following model:

$$\log\left(\frac{\pi_{ij}}{\pi_{i0}}\right) = \boldsymbol{x}_i'\boldsymbol{\beta}_j, \quad j=1,2,$$

with $i=1,2,\ldots,250$, or 300, $\pi_{ij}=P(Y_i=j)$, $\pi_{i0}=P(Y_i=0)=1-P(Y_i=1)-P(Y_i=2)$, and the covariate values $\boldsymbol{x}_i=(x_{i1},x_{i2},\ldots,x_{ip})'$ are drawn from a multivariate standard normal distribution. All computations were performed using the function *multinom* of the library *nnet* in *R*.

First, we consider the hypothesis $H_0:\boldsymbol{L}\boldsymbol{B}=\mathbf{0}$, where the first p_1 columns of the matrix $\boldsymbol{L}$ are zeros for each category j. We then partition the matrix $\boldsymbol{B}=(\boldsymbol{B_1},\boldsymbol{B_2})'$, where $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are $p_1\times 2$ and $p_2\times 2$ matrices, respectively, and $p=p_1+p_2$. The true parameter values used for different settings considered in the study are $\boldsymbol{B}=(\boldsymbol{B_1},\boldsymbol{B_2})'=(\boldsymbol{B_1},\mathbf{0})'$ with $\boldsymbol{B_1}=$

$((-1.3, 1.2, -1.1), (2.0, -1.5, -0.4))'$. We assume the weight matrix $\boldsymbol{Q} = \boldsymbol{I}_1$, where $\boldsymbol{I}_1$ is an identity matrix. We report the simulation results for the cases $\boldsymbol{B}_2 = \boldsymbol{0}$ and $\boldsymbol{B}_2 \neq \boldsymbol{0}$.

The number of replications was initially varied. Finally, each realization was repeated 2000 times to obtain stable results. For each realization, we calculated the bias and mean-square error of the estimators. We define the constant

$$\boldsymbol{\Delta}^* = (\boldsymbol{B} - \boldsymbol{B}^0)'(\boldsymbol{B} - \boldsymbol{B}^0),$$

where $\boldsymbol{\Delta}^*$ is a 2×2 diagonal matrix of constants, $\boldsymbol{B}^0 = (\boldsymbol{B}_1, \boldsymbol{0})'$ corresponding to the reduced model, and $\boldsymbol{B}$ is the parameter in model (A1). Samples were generated using diagonal values of $\boldsymbol{\Delta}^*$ lying between 0 and 4.

The risk performance of any estimator, say $\hat{\boldsymbol{B}}^*$, is measured by comparing its MSE with that of the FE $\hat{\boldsymbol{B}}^{\mathrm{FE}}$ defined as

$$\mathrm{REF}(\hat{\boldsymbol{B}}^{\mathrm{FE}} : \hat{\boldsymbol{B}}^*) = \frac{\text{Simulated MSE}(\hat{\boldsymbol{B}}^{\mathrm{FE}})}{\text{Simulated MSE}(\hat{\boldsymbol{B}}^*)}.$$

The amount of relative efficiency (REF) larger than 1 indicates by how much the estimator $\hat{\boldsymbol{B}}^*$ is efficient than the estimator $\hat{\boldsymbol{B}}^{\mathrm{FE}}$. In terms of ADR, this means that $\hat{\boldsymbol{B}}^{\mathrm{FE}}$ has a higher risk than $\hat{\boldsymbol{B}}^*$. We used this notation in Theorems 3.2 and 3.3.

4.1. *Comparison of* RE, SE, PSE, *and* LASSO *estimators with respect to* FE *when* $LB = h$

When $\boldsymbol{B}_1 \neq \boldsymbol{0}, \boldsymbol{B}_2 = \boldsymbol{0}$, the LASSO can be expected to estimate $\boldsymbol{B}_1$ consistently well by selecting λ so that many of the elements of matrix $\boldsymbol{B}_2$ will be set to 0. On the other hand, the SEs can be expected to perform well by placing almost all weights to the RE $\hat{\boldsymbol{B}}^{\mathrm{RE}}$. We now compare these two methods.

The relative efficiencies of RE, SE, PSE, and LASSO with respect to FE are shown in Tables 1 and 2 for $n = 250$ and 300 when 3 out of 21 regression coefficients are not zero and $\boldsymbol{\Delta}^* = \boldsymbol{0}$. We used the 10-fold cross-validation to choose the best value of the tuning parameter λ for computing LASSO estimator. Detailed results are provided for $(p_1, p_2) = \{(3,3), (3,6), (3,9), (3,12), (3,18)\}$. We also report the simulated values of the estimates and their corresponding standard errors in Table 3 when $k_1 = 3$ and $k_2 = 9$. To save space, we do not report these results for other scenarios.

From Tables 1 and 2 and Figures 1 and 2, we can see that the REF of the SEs is higher than $\hat{\boldsymbol{\beta}}^{\mathrm{FE}}$ at and near $\boldsymbol{\Delta}^* = \boldsymbol{0}$, indicating (as expected) that the greatest ADR reductions occur for parameter

Table 1. Relative efficiencies of RE, SE, PSE, and LASSO with respect to FE when the categories of response = 2 and 3, reference category = 1, and $\boldsymbol{\Delta}^* = \boldsymbol{0}$.

Method	$p_2 = 3$	$p_2 = 6$	$p_2 = 9$	$p_2 = 12$	$p_2 = 18$
			$n = 250$, Category = 2		
RE	1.45	1.78	3.35	4.22	7.37
SE	1.25	1.54	2.65	3.16	4.75
PSE	1.29	1.57	2.76	3.27	4.92
LASSO	1.35	1.42	1.89	2.05	2.64
			$n = 250$, Category = 3		
RE	1.42	1.76	3.31	4.37	6.53
SE	1.23	1.52	2.58	3.19	4.43
PSE	1.26	1.54	2.67	3.28	4.53
LASSO	1.79	1.93	2.43	2.79	3.36

Table 2. Relative efficiencies of RE, SE, PSE, and LASSO with respect to FE when the categories of response = 2 and 3, reference category = 1, and $\boldsymbol{\Delta}^* = \mathbf{0}$.

Method	$p_2 = 3$	$p_2 = 6$	$p_2 = 9$	$p_2 = 12$	$p_2 = 18$
		$n = 300$, Category = 2			
RE	1.33	1.69	3.31	4.16	6.97
SE	1.20	1.43	2.53	3.12	4.72
PSE	1.22	1.44	2.62	3.23	4.88
LASSO	1.31	1.46	1.56	1.79	2.56
		$n = 300$, Category = 3			
RE	1.33	1.57	3.16	4.12	6.41
SE	1.19	1.41	2.51	3.08	4.31
PSE	1.21	1.43	2.59	3.18	4.42
LASSO	1.62	1.71	1.91	2.53	3.24

Table 3. Simulated estimates and standard error (in parenthesis) of RE, SE, PSE, and LASSO with respect to FE when the categories of response = 2 and 3, reference category = 1, $\boldsymbol{\Delta}^* = \mathbf{0}$, $k_2 = 9$, and $n = 250$.

Estimates	Category = 2			Category = 3		
True value	−1.30	1.20	−1.10	2.00	−1.50	−0.40
FE	−1.47 (0.35)	1.35 (0.33)	−1.24 (0.27)	2.29 (0.47)	−1.71 (0.37)	−0.44 (0.25)
RE	−1.36 (0.30)	1.25 (0.28)	−1.15 (0.24)	2.10 (0.39)	−1.57 (0.32)	−0.42 (0.23)
SE	−1.37 (0.31)	1.27 (0.29)	−1.16 (0.24)	2.13 (0.41)	−1.60 (0.33)	−0.42 (0.23)
PSE	−1.38 (0.31)	1.25 (0.29)	−1.17 (0.24)	2.15 (0.41)	−1.58 (0.33)	−0.42 (0.23)
LASSO	−0.98 (0.25)	0.92 (0.22)	−0.52 (0.22)	1.65 (0.31)	−1.20 (0.28)	−0.12 (0.08)

values near the restricted parameter space. We also observe that the REF is an increasing function of the number of inactive variables in the model. Moreover, at $\boldsymbol{\Delta}^* = \mathbf{0}$, as we would expect, $\hat{\boldsymbol{B}}^{\mathrm{RE}}$ is best because of its unbiasedness, and the shrinkage and positive SEs perform better than $\hat{\boldsymbol{B}}^{\mathrm{FE}}$. Tables 1 and 2 also reveal that the LASSO performs better than the shrinkage strategy when the number of inactive predictors p_2 in the model is small. On the other hand, the SEs outshine the LASSO estimator for larger values of p_2.

Generally speaking, in the presence of a relatively large number of inactive predictors in the model, the shrinkage strategy does well relative to the LASSO estimator [23]. However, adaptive SEs are preferable when the number of inactive predictors is relatively large.

4.2. *Comparison of* RE*,* SE*, and* PSE *with respect to* FE *when* $\boldsymbol{LB \neq h}$

The LASSO estimator is omitted from this section since it does not take into consideration the fact that the regression parameters lie in a subspace $\boldsymbol{LB} = \boldsymbol{h}$. The SEs are expected to do better by adapting to the case $\boldsymbol{B}_2 \neq \mathbf{0}$.

The simulation results for the case when the active values of the parameter matrix $\boldsymbol{B}_1 = ((-1.3, 1.2, -1.1), (2.0, -1.5, -0.4))'$ and the inactive values of the parameter matrix $\boldsymbol{B}_2 = ((b_3, \boldsymbol{a}), (b_3, \boldsymbol{a}))'$, where $b_3 = 0$, 0.45, 0.63, 0.77, 0.89, 1, 1.1, and 2, and $\boldsymbol{a}$ is a zero vector with different dimensions, are presented in Figures 1 and 2. Note that there are $k = p_2 - 1$ inactive variables and that the value of $\boldsymbol{\Delta}^*$ is b_3^2 in this setting. The findings are summarized as follows:

(i) The RE $\hat{\boldsymbol{B}}^{\mathrm{RE}}$ is better than all the estimators at and near $\boldsymbol{\Delta}^* = \mathbf{0}$. On the contrary, when $\boldsymbol{\Delta}^* > \mathbf{0}$, the REF of $\hat{\boldsymbol{B}}^{\mathrm{RE}}$ converges to zero or ADR of $\hat{\boldsymbol{B}}^{\mathrm{RE}}$ converges to ∞; whereas the relative efficiencies (or ADR) of all the other estimators remain bounded and approach 1. This is in agreement with the asymptotic results of Section 4.

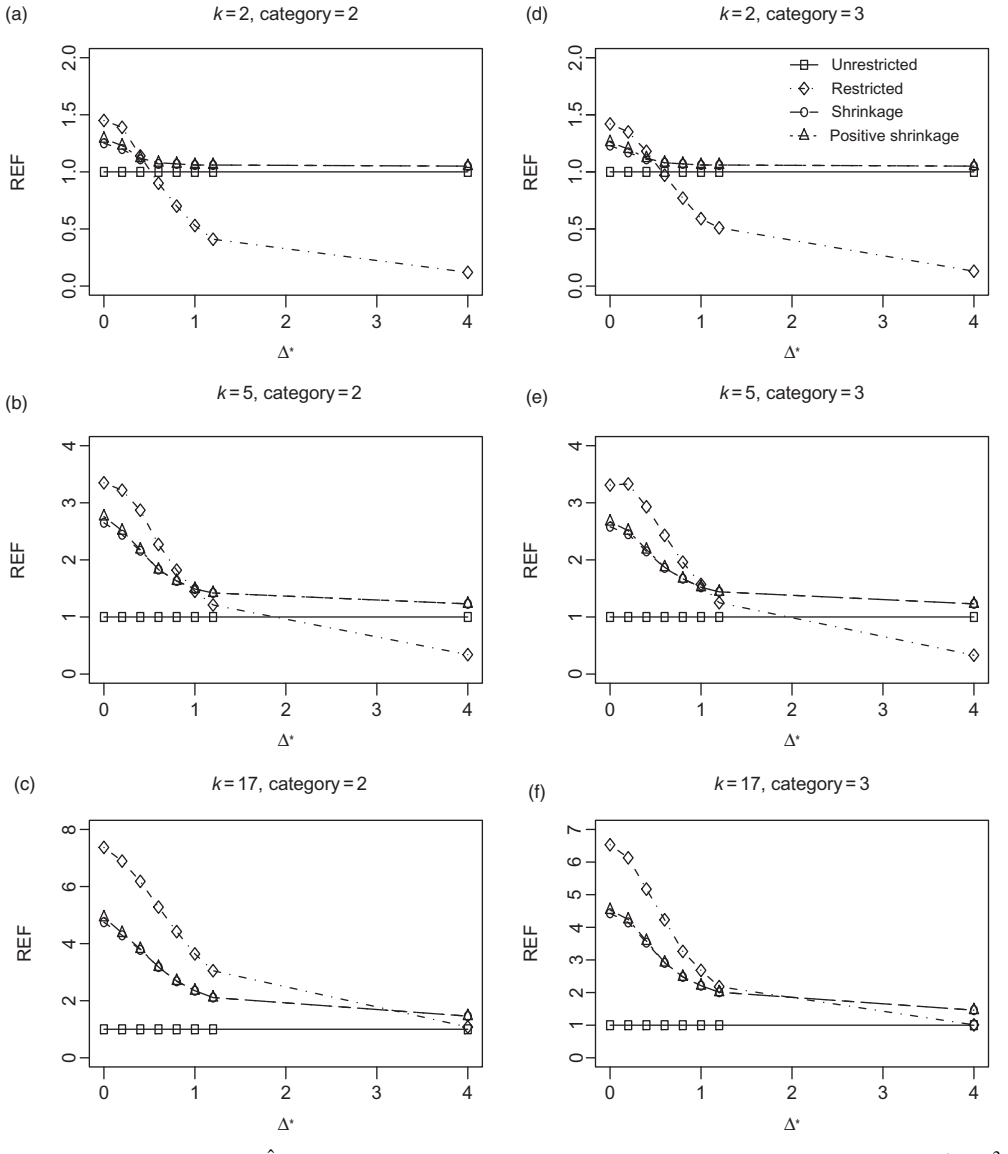

Figure 1. REF with respect to $\hat{\boldsymbol{B}}$ of the estimators when the subspace misspecifies b_3 as zero as a function of $\boldsymbol{\Delta}^* = b_3^2$. Here k is equal to the number of inactive predictors, reference category $= 1$, and $n = 250$.

(ii) Comparing side by side Figure 1(a)–(c) with Figure 1(e)–(g), one can see that the higher the parameter dimensions the better enunciated is the risk dominance of the SEs over the FE.

5. Real data example

We extracted the data from the Demographic and Health Survey study conducted by the National Institute for Population Research and Training in Bangladesh during 2004 under the supervision of the Demographic and Health Surveys programme, Calverton, MD, USA . These data are publicly available in http://measuredhs.com/data. The main goal of this survey was to investigate the important risk factors associated with women's unintended pregnancy. It covered a nationally

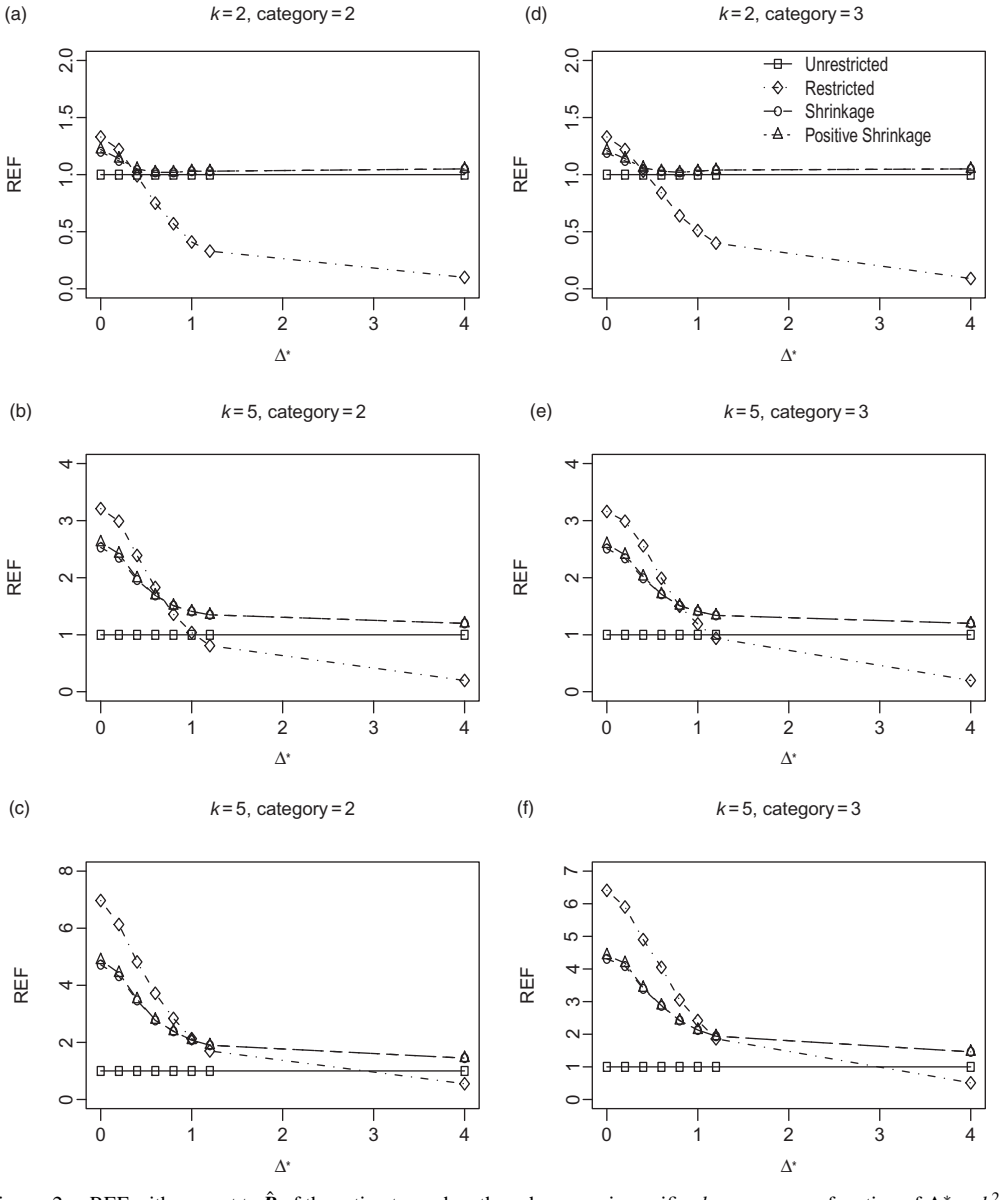

Figure 2. REF with respect to $\hat{\boldsymbol{B}}$ of the estimators when the subspace misspecifies b_3 as zero as a function of $\boldsymbol{\Delta}^* = b_3^2$. Here k is equal to the number of inactive predictors, reference category $= 1$, and $n = 300$.

representative sample of 11, 440 ever-married women from the ages of 10 to 49 years. The analysis is based on 1069 women who were pregnant during the survey. We considered the pregnancy intention status (wanted, mistimed, and unwanted; *wanted* is a reference category) as a response variable and a set of predictor variables: current age (x_1), education level (x_2: 0, no education; 1, education), religion (x_3: 1, Muslim; 0, non-Muslim), number of living children (x_4: 0, no children; 1, children), age at first marriage (x_5), wealth index (two dummy variables: poor versus rich (x_6) and middle versus rich (x_7)), used modern method of family planning prior to pregnancy (x_8: 0, no; 1, yes), current working status (x_9: 0, no; 1, yes), and access to media (x_{10}: 0, no; 1, yes).

Table 4. Estimates (first row) and standard errors (second row) of the coefficients for the effect of current age, number of living children, age at first marriage, and wealth index (poor versus rich) for the response category 2 with respect to 1.

Estimators	β_{21}	β_{24}	β_{25}	β_{28}	REF
FE	0.022	0.413	−0.071	−0.262	1.000
	0.024	0.307	0.063	0.324	
RE	0.010	0.411	−0.064	−0.172	2.447
	0.024	0.291	0.061	0.211	
SE	0.022	0.413	−0.070	−0.296	1.330
	0.024	0.302	0.062	0.278	
PSE	0.022	0.412	−0.071	−0.296	1.330
	0.024	0.302	0.062	0.278	
LASSO	0.001	0.230	0.043	−0.084	1.516
	0.004	0.260	0.009	0.186	

Note: The REF column gives the relative mean-square error of the estimators with respect to the FE.

Table 5. Estimates (first row) and standard errors (second row) of the coefficients for the effect of current age, number of living children, age at first marriage, and wealth index (poor versus rich) for the response category 3 with respect to 1.

Estimators	β_{31}	β_{34}	β_{35}	β_{38}	REF
FE	0.230	0.931	−0.269	−0.380	1.000
	0.023	0.276	0.068	0.336	
RE	0.211	0.613	−0.237	−0.328	1.591
	0.020	0.257	0.058	0.224	
SE	0.226	0.869	−0.263	−0.184	1.541
	0.022	0.276	0.065	0.289	
PSE	0.226	0.869	−0.263	−0.182	1.542
	0.022	0.267	0.064	0.288	
LASSO	0.189	0.352	−0.162	−0.106	1.601
	0.018	0.223	0.047	0.216	

Note: The REF column gives the relative mean-square error of the estimators with respect to the FE.

Since the prior information was not available in this study, we used the stepwise method in the first step of the shrinkage method to form a subset of the total number of predictors. It showed that the current age, number of living children, age at first marriage, and wealth index (poor versus rich) were the active predictors, and the effects of the remaining six predictors may be ignored. We then formed a constraint on the full model by using the inactive predictors. Here, the restricted subspace $\boldsymbol{B}_2 = ((\beta_{22}, \beta_{23}, \beta_{26}, \beta_{27}, \beta_{29}, \beta_{210}), (\beta_{32}, \beta_{33}, \beta_{36}, \beta_{37}, \beta_{39}, \beta_{310}))' = ((0,0,0,0,0,0), (0,0,0,0,0,0))'$, $p = 10$, $p_1 = 4$, and $p_2 = 6$.

The point estimates, the standard errors, and relative efficiencies based on bootstrap size 1000 are displayed in Tables 4 and 5. The findings in these tables are consistent with the simulation results (Tables 1 and 2). The LASSO estimator performs well compared to the SEs when the number of inactive variables is relatively small. On the other hand, the SEs perform well when there are moderate to large number of inactive predictors in the model.

However, the RE outperforms the LASSO since the deleted predictors in the sub-model are indeed irrelevant or nearly irrelevant for the response.

6. Concluding remarks

In this paper, we presented the shrinkage and LASSO methods for simultaneous variable selection and parameter estimation for the MLR model. We established the risk properties of the full model,

reduced model, shrinkage and positive SEs via ADRs and a Monte Carlo simulation study. We also compared the performance of the shrinkage and positive SEs with the LASSO estimator through the same Monte Carlo simulation study.

We compared the relative efficiencies for the RE as well as shrinkage, positive shrinkage, and LASSO estimators with respect to the FE. It is found that when the restriction on the hypothesis is true ($\boldsymbol{\Delta}^* = \mathbf{0}$), the RE is superior to all other estimators. However, as we deviate from the pivot ($\boldsymbol{\Delta}^* > \mathbf{0}$), the REF approaches 0, that is, the risk of the RE becomes unbounded. The relative efficiencies of the shrinkage and positive SEs decrease gradually with the increase in the value of the elements of the $\boldsymbol{\Delta}^*$ matrix, and they perform at the steady rate over a wide range of alternative parameter subspace. In particular, when the number of inactive predictors is large, the positive SE is more efficient than all other estimators.

Our findings in the simulation study show that LASSO is competitive to shrinkage and positive SEs when there is a moderate to large number of predictors in the model and only a few of them are inactive. On the other hand, shrinkage and positive SEs using the data-based weights perform well when p and the number of inactive predictors p_2 are moderate to large.

The results from a real-life example used to illustrate the proposed methods are consistent with our analytical and numerical findings.

Acknowledgements

The authors thank the editor and the three referees for their helpful comments which has improved the earlier version of the manuscript. The research of Shakhawat Hossain and S. Ejaz Ahmed was supported by the University of Winnipeg and the Natural Sciences and the Engineering Research Council of Canada, respectively.

References

[1] Y.H. Chan, *Multinomial logistic regression*, Singapore Med. J. 46 (2004), pp. 259–269.
[2] D.W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed., John Wiley & Sons, New York, 2000.
[3] A. Agresti, *Categorical Data Analysis*, 2nd ed., John Wiley & Sons, New York, 2002.
[4] A.W. van der Vaart, *Asymptotic Statistics*, Cambridge University Press, Cambridge, 1998.
[5] T. Briz and R.W. Ward, *Consumer awareness of organic products in Spain: An application of multinomial logit models*, Food Policy 34 (2009), pp. 295–304.
[6] B. Lee and R.M. Goerge, *Poverty, early childbearing and child maltreatment: A multinomial analysis*, Child. Youth Serv. Rev. 21 (1999), pp. 755–780.
[7] G. Judge and R. Mittelhammaer, *A semiparametric basis for combining estimation problem under quadratic loss*, J. Am. Statist. Assoc. 99 (2004), pp. 479–487.
[8] K. Ohtani and A. Wan, *Comparison of the stein and the usual estimators for the regression error variance under the Pitman nearness criterion when variables are omitted*, Statist. Pap. 50 (2008), pp. 151–160.
[9] S.E. Ahmed, S. Hossain, and K.A. Doksum, *LASSO and shrinkage estimation in Weibull censored regression models*, J. Statist. Plann. Inference 142 (2012), pp. 1273–1284.
[10] S. Hossain and S.E. Ahmed, *Shrinkage and penalty estimators of a Poisson regression model*, Aust. NZ J. Stat. (2012), in press.
[11] K. Golam and A.K.M.E. Saleh, *Improving the estimators of the parameters of a probit regression model: A ridge regression approach*, J. Statist. Plann. Inference 142 (2012), pp. 1421–1435.
[12] G.L. Tian, M.L. Tang, H.B. Fang, and M. Tan, *Efficient methods for estimating constrained parameters with applications to regularized (LASSO) logistic regression*, Comput. Stat. Data Anal. 52 (2008), pp. 3528–3542.
[13] S. van de Geer, L. Meier, and P. Bühlmann, *The group LASSO for logistic regression*, J. R. Statist. Soc. Ser. B 70 (2008), pp. 53–71.
[14] J. Friedman, T. Hastie, and R. Tibshirani, *Regularization paths for generalized linear models via coordinate descent*, J. Statist. Softw. 33 (2010), pp. 1–22.
[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed., Springer, New York, 2009.
[16] S. GeerVan de , *High-dimensional generalized linear models and the LASSO*, Ann. Stat. 36 (2008), pp. 614–645.
[17] M. Park and T. Hastie, *An L_1 regularization-path algorithm for generalized linear models*, J. R. Statist. Soc. Ser. B 69 (2007), pp. 659–677.
[18] P. Zhao and B. Yu, *On model selection consistency of LASSO*, J. Mach. Learn. Res. 7 (2006), pp. 2541–2563.
[19] L. Krishnapuram, B. Carin, M. Figueiredo, and A. Hartemink, *Sparse multinomial logistic regression: Fast algorithms and generalization bounds*, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005), pp. 957–968.

[20] Y. Ding, *On the asymptotic normality of multinomial population size estimators with application to back calculation of AIDS epidemic*, Biometrika 83 (1996), pp. 695–699.
[21] N.M. Kiefer and G.R. Skoog, *Local asymptotic specification error analysis*, Econometrica 52 (1984), pp. 873–885.
[22] R. Tibshirani, *Regression shrinkage and selection via the LASSO*, J. R. Statist. Soc. Ser. B 58 (1996), pp. 267–288.
[23] S.E. Ahmed, K. Doksum, S. Hossain, and J. You, *Shrinkage, pretest and LASSO estimators in partially linear models*, Aust. NZ J. Stat. 49 (2007), pp. 461–471.
[24] M.S. Hossain, *Shrinkage, pretest and LASSO estimators in parametric and semiparametric linear models*, Unpublished Ph.D. dissertation, University of Windsor, Windsor, 2008.

Appendix 1

Proof of Theorem 3.3 To prove this theorem, we first derive the asymptotic covariance matrix of the estimators. The asymptotic covariance of an estimator $\boldsymbol{B}^*$ is defined as follows:

$$\boldsymbol{\Sigma}(\hat{\boldsymbol{B}}^*) = \lim_{n\to\infty} nE\{(\hat{\boldsymbol{B}}^* - \boldsymbol{B})(\hat{\boldsymbol{B}}^* - \boldsymbol{B})'\}.$$

The ADR of any estimator is

$$\mathbf{ADR}(\hat{\boldsymbol{B}}^*; \boldsymbol{Q}) = \mathrm{tr}(\boldsymbol{Q}\boldsymbol{\Sigma}(\hat{\boldsymbol{B}}^*)).$$

Therefore, $\boldsymbol{\Sigma}(\hat{\boldsymbol{B}}^{\mathrm{FE}}) = (\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}$. Using the asymptotic distribution of $\hat{\boldsymbol{B}}^{\mathrm{FE}}$ in Section 2.1 and Theorem 3.1, we can write

$$\begin{aligned}\boldsymbol{\Sigma}(\hat{\boldsymbol{B}}^{\mathrm{RE}}) &= \lim_{n\to\infty} E\{n(\hat{\boldsymbol{B}}^{\mathrm{RE}} - \boldsymbol{B})(\hat{\boldsymbol{B}}^{\mathrm{RE}} - \boldsymbol{B})'\} \\ &= (\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1} - \boldsymbol{M}\boldsymbol{L}(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1} + \boldsymbol{\delta}'(\boldsymbol{M}'\boldsymbol{Q}\boldsymbol{M})\boldsymbol{\delta}.\end{aligned}$$

A similar proof of above is available in [24]. The covariance matrix of the SE

$$\begin{aligned}\boldsymbol{\Sigma}(\hat{\boldsymbol{B}}^{\mathrm{SE}}y) &= \lim_{n\to\infty} nE\{(\hat{\boldsymbol{B}}^{\mathrm{SE}} - \boldsymbol{B})(\hat{\boldsymbol{B}}^{\mathrm{SE}}y - \boldsymbol{B})'\} \\ &= \lim_{n\to\infty}\left\{nE\left[\left\{(\hat{\boldsymbol{B}}^{\mathrm{FE}} - \boldsymbol{B}) - \frac{Jp_2 - 2}{\Psi}(\hat{\boldsymbol{B}}^{\mathrm{FE}} - \hat{\boldsymbol{B}}^{\mathrm{RE}})\right\}\right]\right. \\ &\quad \left.\times\left\{(\hat{\boldsymbol{B}}^{\mathrm{FE}} - \boldsymbol{B})' - \frac{Jp_2 - 2}{\Psi}(\hat{\boldsymbol{B}}^{\mathrm{FE}} - \hat{\boldsymbol{B}}^{\mathrm{RE}})'\right\}\right. \\ &= \lim_{n\to\infty} nE[(\hat{\boldsymbol{B}}^{\mathrm{FE}} - \boldsymbol{B})(\hat{\boldsymbol{B}}^{\mathrm{FE}} - \boldsymbol{B})'] \\ &\quad + (Jp_2 - 2)^2 \lim_{n\to\infty} nE[(\hat{\boldsymbol{B}}^{\mathrm{FE}} - \hat{\boldsymbol{B}}^{\mathrm{RE}})(\hat{\boldsymbol{B}}^{\mathrm{FE}} - \hat{\boldsymbol{B}}^{\mathrm{RE}})'\Psi^{-2}] \\ &\quad - 2(Jp_2 - 2)\lim_{n\to\infty} nE[(\hat{\boldsymbol{B}}^{\mathrm{FE}} - \hat{\boldsymbol{B}}^{\mathrm{RE}})(\hat{\boldsymbol{B}}^{\mathrm{FE}} - \boldsymbol{B})'\Psi^{-1}] \\ &= \boldsymbol{\Lambda}^{-1} + (Jp_2 - 2)^2 \lim_{n\to\infty} nE[(\hat{\boldsymbol{B}}^{\mathrm{FE}} - \hat{\boldsymbol{B}}^{\mathrm{RE}})(\hat{\boldsymbol{B}}^{\mathrm{FE}} - \hat{\boldsymbol{B}}^{\mathrm{RE}})'\Psi^{-2}] \\ &\quad - 2(Jp_2 - 2)\lim_{n\to\infty} nE\{[(\hat{\boldsymbol{B}}^{\mathrm{FE}} - \hat{\boldsymbol{B}}^{\mathrm{RE}})[(\hat{\boldsymbol{B}}^{\mathrm{FE}} - \hat{\boldsymbol{B}}^{\mathrm{RE}}) \\ &\quad - \boldsymbol{\Lambda}^{-1}\boldsymbol{L}'\boldsymbol{\Lambda}^{-1}(\boldsymbol{L}\boldsymbol{B} - \boldsymbol{h})]'\Psi^{-1}]\} \\ &= \boldsymbol{\Lambda}^{-1} + (Jp_2 - 2)^2 \lim_{n\to\infty} \boldsymbol{M}E[n(\boldsymbol{L}\hat{\boldsymbol{B}}^{\mathrm{FE}} - \boldsymbol{h})(\boldsymbol{L}\hat{\boldsymbol{B}}^{\mathrm{FE}} - \boldsymbol{h})'\Psi^{-2}]\boldsymbol{M}' \\ &\quad - 2(Jp_2 - 2)\lim_{n\to\infty} \boldsymbol{M}E[n(\boldsymbol{L}\hat{\boldsymbol{B}}^{\mathrm{FE}} - \boldsymbol{h})(\boldsymbol{L}\hat{\boldsymbol{B}}^{\mathrm{FE}} - \boldsymbol{h})'\Psi^{-1}]\boldsymbol{M}' \\ &\quad + 2(Jp_2 - 2)\lim_{n\to\infty} n\{E[(\hat{\boldsymbol{B}}^{\mathrm{FE}} - \hat{\boldsymbol{B}}^{\mathrm{RE}}y)'\Psi^{-1}]\ (\boldsymbol{L}\boldsymbol{B} - \boldsymbol{h})'(\boldsymbol{L}\boldsymbol{\Lambda}^{-1}\boldsymbol{L}')^{-1}\boldsymbol{\Lambda}^{-1}\Big\},\end{aligned}$$

where $\boldsymbol{\Lambda} = \boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}$. Therefore, by some algebra and using the identity

$$E(\chi^{-2}_{Jp_2+4}) = E(\chi^{-2}_{Jp_2+2}) + E(\chi^{-4}_{Jp_2+4}), \tag{A1}$$

we get

$$\begin{aligned}\boldsymbol{\Sigma}(\hat{\boldsymbol{B}}^{\mathrm{SE}}) = {} & \boldsymbol{\Lambda}^{-1} - 2(Jp_2 - 2)\boldsymbol{MLB}^{-1}\{2E(\chi^{-2}_{Jp_2+2}(\boldsymbol{\Delta})) \\ & - (Jp_2 - 2)E(\chi^{-4}_{Jp_2+2}(\boldsymbol{\Delta}))\} + (Jp_2 - 2)\boldsymbol{\delta}'(\boldsymbol{M}'\boldsymbol{M})\boldsymbol{\delta}\{2E(\chi^{-2}_{Jp_2+2}(\boldsymbol{\Delta})) \\ & - 2E(\chi^{-4}_{Jp_2+2}(\boldsymbol{\Delta})) + (Jp_2 - 2)E(\chi^{-4}_{Jp_2+4}(\boldsymbol{\Delta}))\}.\end{aligned}$$

The derivation of the covariance matrix for $\hat{\boldsymbol{B}}^{\mathrm{PSE}}$ follows from some algebra, similar steps of the derivation of the covariance matrix of $\hat{\boldsymbol{B}}^{\mathrm{SE}}y$, and identity (A1). ■