Simplex-based multinomial logistic regression with diverging numbers of categories and covariate

Fu, Sheng; Chen, Piao; Liu, Yufeng; Ye, Zhisheng

DOI
10.5705/ss.202021.0082

Publication date
2023

Document Version
Final published version

Published in
Statistica Sinica

Simplex-based Multinomial Logistic Regression with Diverging Numbers of Categories and Covariates

Sheng Fu[1], Piao Chen[*2] , Yufeng Liu[3] and Zhisheng Ye[1]

[1]*National University of Singapore,* [2]*Delft University of Technology*

[3]*University of North Carolina at Chapel Hill*

Abstract: Multinomial logistic regression models are popular in multicategory classification analysis, but existing models suffer several intrinsic drawbacks. In particular, the parameters cannot be determined uniquely because of the over-specification. Although additional constraints have been imposed to refine the model, such modifications can be inefficient and complicated. In this paper, we propose a novel and efficient simplex-based multinomial logistic regression technique, seamlessly connecting binomial and multinomial cases under a unified framework. Compared with existing models, our model has fewer parameters, is free of any constraints, and can be solved efficiently using the Fisher scoring algorithm. In addition, the proposed model enjoys several theoretical advantages, including Fisher consistency and sharp comparison inequality. Under mild conditions, we establish the asymptotical normality and convergence for the new model, even when the numbers of categories and covariates increase with the sample size. The proposed framework is illustrated by means of extensive simulations and real applications.

Key words and phrases: Asymptotics; Classification; Fisher consistency; Kernel learning; MLR; Simplex coding scheme

*Corresponding author: p.chen-6@tudelft.nl

1. Introduction

Logistic regression (LR) is the most frequently used regression model for analysis of categorical outcomes (Cramer, 2003; Yee, 2015; Fang and Yi, 2021; Mo and Liu, 2021). LR has been widely applied in epidemiology, biology, economics, and the social sciences, among others (Hosmer et al., 2013; Lemeshow and Hosmer, 2014). LR models can be divided into two types based on the number of categories, binomial LR and multinomial LR (MLR). The MLR model is also known as the conditional maximum entropy model in natural language processing (Berger et al., 1996; Malouf, 2002), and as the softmax regression in neural networks (Ng et al., 2013; Goodfellow et al., 2016).

The statistical theory of binomial LR is well established, whereas modeling and inference for MLR is more complicated. There have been numerous attempts to generalize the original binomial LR to the multinomial case; see McCullagh and Nelder (1989), Hastie et al. (2009), and Tutz (2011). For a k-categorical regression problem, a natural approach is to estimate k regression functions, one for each category. The unknown parameters in these functions are typically jointly estimated using a maximum likelihood or Bayesian updating. As shown in Bühlmann and Van De Geer (2011, §3.3.3), such an extension is over-specified, with unidentifiable parameters. In general, $k-1$ functions are sufficient to determine a k-categorical LR model. For instance, the binomial LR ($k=2$) is defined by a single regression function. Therefore, additional restrictions are needed on the regression functions to make the model identifiable. Roughly speaking, there are two common schemes. The first prespecifies a reference category and sets its regression function to zero, and the second uses a sum-to-zero constraint on the k functions (Hastie et al., 2015, §3.3). Both extensions sub-

sume the binomial LR as a special example, and are introduced in Section 3. However, the extra constraints complicate the parameter estimation and theoretical analysis. Specifically, the reference-based MLR does not treat all categories equally, and the explicit sum-to-zero constraint increases the computational cost. In addition, the relationship between these two constrained MLR (CMLR) models is not entirely clear.

A powerful method for circumventing the explicit constraints is to use simplex coding. The basic idea is to construct a k-vertex simplex structure in a $(k-1)$-dimensional Euclidean space, where each vertex represents one category. The covariate vector of each instance is then mapped to a point in this $(k-1)$-dimensional space. In other words, only $k-1$ regression functions need to be estimated under the simplex coding scheme, and the non-identifiability issue is resolved automatically, without further constraints on these $k-1$ functions. Therefore, simplex coding for multiclass learning is expected to have lower computational complexity in model training (Hill and Doucet, 2007; Lange and Wu, 2008; Mroueh et al., 2012). Such a scheme is also called angle-based classification by Zhang and Liu (2014), Zhang et al. (2018), and Fu et al. (2018), because the predicted label of a new observation corresponds to the vertex that has the smallest angle with the mapped $(k-1)$-dimensional point of the covariates. This geometric interpretation makes the coding scheme easy to understand.

The first objective of our study is to remove the cumbersome constraints in existing MLR models, and propose a novel MLR model using the delicate simplex structure, called simplex-based MLR (SMLR). Inheriting from simplex coding, the redundancy of the categorical space is removed and the representation of the regression functions is identifiable. Hence, the resulting SMLR model has a clear geometric explanation, and is computationally efficient in

terms of parameter estimation. Compared with regular CMLRs, the SMLR model enjoys more parsimonious parameter specification. Specifically, the SMLR model avoids the subjective selection of a reference category, gets rid of the sum-to-zero constraint, and provides a symmetric insight on all categories by treating them equally. With fewer parameters, the likelihood estimation of the SMLR solves an unconstrained optimization problem, which can be implemented efficiently using the Fisher scoring algorithm. The proposed SMLR can be treated as a unified framework recovering CMLRs, but the parameters involved have different interpretations.

The second objective of this study is to establish the asymptotic properties of the MLEs of an SMLR with a diverging number of categories, which is peculiar to multicategory classification applications. In practical problems, the granularity of the classification, in terms of the number of categories, is usually determined based on the size of the available training data (Dekel and Shamir, 2010). For example, photo sharing websites allow users to annotate their photos with keywords. The key task is to recommend keywords whenever new photos are uploaded. Assuming there are no restrictions on the keywords that may be used, the set of distinct keywords is likely to grow as additional photos are uploaded to the site. Similarly, web directory classification with Yahoo! taxonomies yields some rare categories that are ignored under a small sample size, but are considered for larger samples (Liu et al., 2005). Other examples include textual document categorization (Dekel and Shamir, 2010) and the identification of flowers, plants and products using images (Nilsback and Zisserman, 2008; Deng et al., 2010). The phenomenon of a diverging number of categories has attracted some attention in the literature, with most existing studies focusing on algorithmic development,

for example, distributed computing, hierarchical classification, and penalization techniques. For instance, Deng et al. (2009, 2010) exploited the semantic hierarchy of categories to obtain more informative image classifiers. Based on the hierarchical structure of categories, Price et al. (2019) proposed a group-fused MLR model that automatically combines the categories. However, there are a few asymptotic studies of MLR models when the class size increases with the sample size, probably because the constraint on existing CMLRs makes the asymptotic properties difficult to establish. In contrast, when the dimensions of the covariates and the categories are fixed, the asymptotic properties of MLRs are well established (Fahrmeir and Kaufmann, 1985; Van der Vaart, 1998; Tutz, 2011). In this study, we focus on the asymptotics of an SMLR model with varying category sizes. An important byproduct is that we also establish the asymptotics under a diverging number of covariates, which has received scant attention in the literature on MLR models.

The third objective of this study is to show the theoretical advantages of the SMLR model under certain settings. In particular, we explore kernel learning for SMLR, which enjoys a faster convergence rate than those of existing MLR models. Few studies have conducted convergence analysis for kernel MLR under a diverging number of categories. This study fills this gap by establishing the consistency of kernel SMLR, while letting the number of categories k go to infinity at the order of $o(n)$. In addition, we show that the proposed SMLR enjoys some desirable statistical properties, including Fisher consistency and comparison inequality, which are fundamental in understanding the nature of the SMLR model.

The rest of the paper is organized as follows. In Section 2, we introduce the notation

and briefly review regular MLR models. In Section 3, we propose the SMLR method, and explore its connections with regular MLR models. We establish the asymptotical results for an SMLR model with a diverging number of parameters in Section 4. Simulation studies and real applications demonstrate the performance of the proposed approach in Section 5. Section 6 concludes the paper. The main proofs are given in the Supplementary Material.

Throughout this paper, $\mathbf{0}$ and $\mathbf{1}$ represent vectors of zeros and ones, respectively, $\mathbf{e}_j$ is the jth column vector of an identity matrix $\mathbf{I}$, the dimensions of which can be inferred contextually, and $\mathbf{diag}(\mathbf{u})$ is a diagonal matrix with entries determined by a vector $\mathbf{u}$. The vectorization of a matrix $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{m\times n}$ is defined as

$$\overrightarrow{\mathbf{A}} := \mathrm{vec}(\mathbf{A}) = (a_{11}, \ldots, a_{m1}, a_{12}, \ldots, a_{m2}, \ldots, a_{1n}, \ldots, a_{mn})^\top \in \mathbb{R}^{mn}.$$

Let $\|\mathbf{A}\|_2$ be the spectral norm, defined as the largest singular value of $\mathbf{A}$. In general, $\|\mathbf{A}\| := \sqrt{\sum_{i,j} a_{ij}^2}$ is the Frobenius norm, including the Euclidean norm of a vector. For a square matrix $\mathbf{A}$, $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ denote the maximum and minimum eigenvalues, respectively. For any matrices $\mathbf{A}$ and $\mathbf{B}$ of the same dimensions, $\mathbf{A} \succeq \mathbf{B}$ or $\mathbf{B} \preceq \mathbf{A}$ denotes $\mathbf{A} - \mathbf{B}$ is positive semi-definite. $\mathbf{A} \succ \mathbf{0}$ implies that $\mathbf{A}$ is positive definite, where $\mathbf{0}$ is a matrix of zeros. For $\mathbf{A} \succ \mathbf{0}$, $\mathbf{A}^{\frac{1}{2}}$ denotes its symmetric square root, with $\mathbf{A} = (\mathbf{A}^{\frac{1}{2}})^2$ and $\mathbf{A}^{-\frac{1}{2}} = (\mathbf{A}^{\frac{1}{2}})^{-1}$. The Kronecker product of two matrices $\mathbf{A}$ and $\mathbf{B}$ is denoted by $\mathbf{A} \otimes \mathbf{B}$.

2. Review of Regular MLR Models

Consider a multicategory classification problem with k possible categories $\mathcal{Y} \triangleq \{1, 2, \ldots, k\}$. Suppose we are given a set of observations $\mathcal{D} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})^\top \in \mathcal{X} \subseteq \mathbb{R}^d$ is the covariate vector and $y_i \in \mathcal{Y}$ is the corresponding response category. We define

$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top \in \mathbb{R}^{n\times d}$ as the design matrix.

Let $\pi_y(\mathbf{x})$ be the conditional probability that the response category is y for the given covariate $\mathbf{x}$. Note that y must take one and only one class label from $\mathcal{Y}$. We need the sum-to-one condition $\sum_{y=1}^k \pi_y(\mathbf{x}) = 1$ to reflect this implicit nature of a multinomial regression. Under this condition, only $k-1$ free probabilities are informative, and the rest are redundant. For the observations $\mathcal{D}$, the joint distribution is $\prod_{i=1}^n \pi_{y_i}(\mathbf{x}_i)$. To link the probability $\{\pi_1, \ldots, \pi_k\}$ to the covariate $\mathbf{x}$, a generic MLR model considers the multinomial-Poisson transformation (Baker, 1994; Lang, 1996)

$$\pi_y(\mathbf{x}) = \frac{e^{\mathbf{x}^\top \boldsymbol{\theta}_y}}{\sum_{j=1}^k e^{\mathbf{x}^\top \boldsymbol{\theta}_j}}, \quad \mathbf{x} \in \mathcal{X},\ y = 1, \ldots, k, \tag{2.1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k) \in \mathbb{R}^{d\times k}$ is the regression coefficient matrix and $\boldsymbol{\theta}_j \in \mathbb{R}^d$ is the jth column vector. Model (2.1) can be interpreted as a neural network (Ripley, 1996). Based on the observations $\mathcal{D}$, the log-likelihood function for (2.1) is

$$\ell_n(\boldsymbol{\theta}) = \log\Big\{\prod_{i=1}^n \pi_{y_i}(\mathbf{x}_i)\Big\} = \sum_{i=1}^n \mathbf{x}_i^\top \boldsymbol{\theta}_{y_i} - \sum_{i=1}^n \log\Big(\sum_{j=1}^k e^{\mathbf{x}_i^\top \boldsymbol{\theta}_j}\Big). \tag{2.2}$$

Note that if we add a common vector $\mathbf{b} \in \mathbb{R}^d$ to each $\boldsymbol{\theta}_j$, the probabilities in (2.1) remain unchanged and $\ell_n(\boldsymbol{\theta} + \mathbf{b}\mathbf{1}_k^\top) = \ell_n(\boldsymbol{\theta})$. Therefore, the MLR model (2.1) is not identifiable with over-specified $\boldsymbol{\theta}$. In fact, $\boldsymbol{\theta}$ has $k-1$ free columns, because there are $k-1$ free informative conditional category probabilities. This problem may be readily resolved by adopting some restrictions on $\boldsymbol{\theta}_j$, leading to the constrained MLR (CMLR).

Two customary constraints are used in the literature to refine the parameters. The first is to choose a reference category, denoted by $r \in \mathcal{Y}$, and set $\boldsymbol{\theta}_r$ as the zero vector $\mathbf{0}$, leading to the reference-based MLR (CMLR1); see Anderson (1972), Anderson and Blair (1982), Albert

and Anderson (1984), Böhning (1992), Krishnapuram et al. (2005), and Hastie et al. (2009), among others. As a result, the CMLR1 estimator is given by

$$\widehat{\boldsymbol{\theta}}^{\mathrm{rb}} = \arg\max_{\boldsymbol{\theta}} \ \ell_n(\boldsymbol{\theta}), \quad \text{s.t. } \boldsymbol{\theta}_r = \mathbf{0}. \tag{2.3}$$

The other is a sum-to-zero constraint $\sum_{j=1}^{k} \boldsymbol{\theta}_j = \mathbf{0}$, leading to the symmetric constrained MLR (CMLR2), which has been studied by Friedman et al. (2000), Zhu and Hastie (2004, 2005), Friedman et al. (2010), Zahid and Tutz (2013b), and Hastie et al. (2015), among others. Consequently, the estimator for the CMLR2 model is defined as

$$\widehat{\boldsymbol{\theta}}^{\mathrm{sc}} = \arg\max_{\boldsymbol{\theta}} \ \ell_n(\boldsymbol{\theta}), \quad \text{s.t. } \sum_{j=1}^{k} \boldsymbol{\theta}_j = \mathbf{0}. \tag{2.4}$$

Note that the total number of parameters in (2.3) or (2.4) is dk, and the exact number of free parameters is $d(k-1)$, owing to the extra constraints. Comprehensive discussions on MLRs can be found in Tutz (2011, §8), Hosmer et al. (2013, §8), Yee (2015), and the references therein.

The above-mentioned CMLRs have some inherent deficiencies. On the one hand, when the reference category for CMLR1 alters, the estimated parameters and the corresponding interpretation change as well. As a result, CMLR1 lacks systematic insights on the categories, and the choice of reference category is subjective and confusing in practice, especially with an increasing number of categories. On the other hand, as shown in Zhang and Liu (2014), the explicit sum-to-zero constraint in CMLR2 can be theoretically inefficient and computationally expensive. We next propose a competent MLR formulation with appealing properties that is free of constraints on the parameters.

3. Simplex-based MLR

In this section, we formulate a novel and efficient simplex-based MLR (SMLR) in Section 3.1, develop the estimation procedure for the SMLR model in Section 3.2, and compare the proposed SMLR with existing MLRs in Section 3.3.

3.1 Methodology

To address the limitations in existing CMLRs, we borrow from recent multicategory classification studies (Zhang and Liu, 2014; Zhang et al., 2018; Fu et al., 2018), and propose an attractive simplex-based MLR model. A well-designed simplex in a $(k-1)$-dimensional Euclidean space plays a central role in reducing the parameter redundancy. To begin with, consider k vertices $\{\mathbf{w}_j \in \mathbb{R}^{k-1},\ j = 1, \ldots, k\}$

$$\mathbf{w}_j = \begin{cases} (k-1)^{-1/2}\mathbf{1}_{k-1}, & \text{if } \ j = 1 \\ -(1+k^{1/2})/\{(k-1)^{3/2}\}\mathbf{1}_{k-1} + \{k/(k-1)\}^{1/2}\mathbf{e}_{j-1}, & \text{if } \ 2 \le j \le k \end{cases}, \tag{3.1}$$

where $\mathbf{1}$ and $\mathbf{e}_j$ are vectors in $\mathbb{R}^{k-1}$. One can verify that each $\mathbf{w}_j$ has Euclidean norm 1 and $\mathbf{W}\mathbf{1}_k = \sum_{j=1}^{k} \mathbf{w}_j = \mathbf{0}$. Denote the matrix of vertex vectors as $\mathbf{W} \triangleq (\mathbf{w}_1, \ldots, \mathbf{w}_k) \in \mathbb{R}^{(k-1)\times k}$. Clearly, $\mathbf{W}$ is of full row rank $k-1$. Note that alternative constructions of the simplex exist, such as those of Hill and Doucet (2007) and Mroueh et al. (2012). However, we can always connect $\mathbf{W}$ with those simplices using proper linear transformations in $\mathbb{R}^{k-1}$, and then show that they are equivalent.

The simplex $\mathbf{W}$ can be used to reduce the dimension of the categorical space to $k-1$. Consider a k-categorical distribution with $P(Y = y) = p_y > 0$ $(y \in \mathcal{Y})$, with the sum-to-one condition $\sum_{j=1}^{k} p_j = 1$. The conventional multinomial distribution encodes Y as a one-hot

vector in $\mathbb{R}^k$, that is, $\mathbf{Y} = \mathbf{e}_Y$. Here, $\mathbf{Y}$ is redundant, because $\mathbf{1}_k^\top \mathbf{Y} \equiv 1$. Let $\mathbf{p} = (p_1, \ldots, p_k)^\top$ be the probability vector. Then, the covariance matrix of $\mathbf{Y}$ is $\mathbf{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top$ (Forbes et al., 2011). Without loss of generality, we can encode category j to the jth vertex $\mathbf{w}_j$ to obtain another multinomial random vector $\mathbf{Z} = \mathbf{W}\mathbf{Y} \in \mathbb{R}^{k-1}$, with $P(\mathbf{Z} = \mathbf{w}_j) = P(\mathbf{Y} = \mathbf{e}_j) = p_j$. The following proposition states a useful result on the covariance matrix of $\mathbf{Z}$.

Proposition 1. *The covariance matrix of* $\mathbf{Z} = \mathbf{W}\mathbf{Y}$ *is positive definite.*

Note that the original $\mathbf{Y}$ is redundant with the covariance matrix $\mathbf{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top \succeq \mathbf{0}$. On the other hand, by Proposition 1, the refined category vector $\mathbf{Z} = \mathbf{W}\mathbf{Y}$ has the covariance matrix $\mathbf{W}(\mathbf{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top)\mathbf{W}^\top \succ \mathbf{0}$. Therefore, the simplex $\mathbf{W}$ in $\mathbb{R}^{k-1}$ leads to a refined categorical space without redundancy. We can use the vertex $\mathbf{w}_j$ to represent the jth category, a strategy we call simplex coding.

Next, we use $\mathbf{W}$ to remove the explicit sum-to-zero constraint in CMLR2. For the coefficients $\boldsymbol{\theta} \in \mathbb{R}^{d\times k}$ with $\sum_{j=1}^k \boldsymbol{\theta}_j = \mathbf{0}$, we can find a matrix $\boldsymbol{\beta} \in \mathbb{R}^{d\times(k-1)}$ such that $\boldsymbol{\theta}_j = \boldsymbol{\beta}\mathbf{w}_j$, for $j = 1, \ldots, k$. It is obvious that $\sum_{j=1}^k \boldsymbol{\theta}_j = \sum_{j=1}^k \boldsymbol{\beta}\mathbf{w}_j = \boldsymbol{\beta}(\sum_{j=1}^k \mathbf{w}_j) = \mathbf{0}$, using the fact that $\sum_{j=1}^k \mathbf{w}_j = \mathbf{0}$. For example, a possible choice of $\boldsymbol{\beta}$ is $(1 - \frac{1}{k})\boldsymbol{\theta}\mathbf{W}^\top$. The equivalence between $\boldsymbol{\theta}$ and $\boldsymbol{\beta}\mathbf{W}$ is shown in the following proposition.

Proposition 2. $\{\boldsymbol{\theta} \in \mathbb{R}^{d\times k} \mid \sum_{j=1}^k \boldsymbol{\theta}_j = \mathbf{0}\}$ *is equivalent to* $\{\boldsymbol{\beta}\mathbf{W} \mid \boldsymbol{\beta} \in \mathbb{R}^{d\times(k-1)}\}$.

By Proposition 2, the parameters $\boldsymbol{\theta}$ with a sum-to-zero constraint can be reformulated as $\boldsymbol{\beta}\mathbf{W}$, without loss of information. By replacing $\boldsymbol{\theta}_j$ in the classical MLR (2.1) with $\boldsymbol{\beta}\mathbf{w}_j$, the simplex-based MLR model takes the form

$$\widetilde{\pi}_j(\mathbf{x}) = \frac{e^{\mathbf{x}^\top \boldsymbol{\beta}\mathbf{w}_j}}{\sum_{s=1}^k e^{\mathbf{x}^\top \boldsymbol{\beta}\mathbf{w}_s}}, \quad \mathbf{x} \in \mathcal{X},\ j = 1, \ldots, k. \tag{3.2}$$

We can interpret the quantity $\mathbf{x}^\top \boldsymbol{\beta} \mathbf{w}_j$ as the inner product $\langle \boldsymbol{\beta}^\top \mathbf{x}, \mathbf{w}_j \rangle$ in $\mathbb{R}^{k-1}$. In other words, we first map a covariate $\mathbf{x}$ to a point $\boldsymbol{\beta}^\top \mathbf{x}$ in $\mathbb{R}^{k-1}$ using the coefficient matrix $\boldsymbol{\beta}$, and then take the inner product with the encoded vertex $\mathbf{w}_j$. Because $\widetilde{\pi}_j(\mathbf{x})$ is increasing in $\mathbf{x}^\top \boldsymbol{\beta} \mathbf{w}_j$, the predicted rule is $\widehat{y}(\mathbf{x}) = \arg\max_j \widetilde{\pi}_j(\mathbf{x}) = \arg\max_j \{\mathbf{x}^\top \boldsymbol{\beta} \mathbf{w}_j\}$, which is computationally more efficient than that of the reference-based MLR. In particular, SMLR simply requires computing k individual inner products, whereas the reference-based MLR involves calculating k probabilities using (2.1). As shown in Zhang and Liu (2014), the largest inner product rule is equivalent to the least angle rule. Hence, the SMLR can be treated as an alternative multicategory angle-based classifier.

The proposed SMLR model enjoys a parsimonious model specification without constraints. Each of the CMLRs involves dk parameters under a linear constraint, whereas SMLR requires only $d(k-1)$ parameters. Note that $\boldsymbol{\beta}\mathbf{W}$ plays the same role as $\boldsymbol{\theta}$ in existing MLRs. Under the factorization $\boldsymbol{\beta}\mathbf{W}$, the $k-1$ rows of $\mathbf{W}$ can be viewed as latent outcome variables, and each row has a loading on each of the k categories. The $k-1$ columns of $\boldsymbol{\beta}$ specify parameter vectors for these latent outcome variables.

To further interpret the parameters of SMLR and investigate their relationships with the CMLRs, we examine the log-odds forms. For a CMLR1 model with a reference category r (that is, $\boldsymbol{\theta}_r \equiv \mathbf{0}$), we have

$$\log\left(\frac{\pi_j(\mathbf{x})}{\pi_r(\mathbf{x})}\right) = \mathbf{x}^\top \boldsymbol{\theta}_j, \quad j = 1, \ldots, k.$$

Clearly, $\boldsymbol{\theta}_j$ depends on the selection of the reference category, and different selections yield different interpretations of the parameters of CMLR1. On the other hand, the parameter interpretation for CMLR2 is related to the median response (Tutz, 2011), which is defined

as the geometric mean

$$\mathrm{GM}(\mathbf{x}) = \prod_{j=1}^{k} \{\pi_j(\mathbf{x})\}^{1/k}.$$

Then, for CMLR2 with the constraint $\sum_{j=1}^{k} \boldsymbol{\theta}_j = \mathbf{0}$, we have

$$\log\left(\frac{\pi_j(\mathbf{x})}{\mathrm{GM}(\mathbf{x})}\right) = \mathbf{x}^\top \boldsymbol{\theta}_j, \quad j = 1, \ldots, k.$$

Therefore, $\boldsymbol{\theta}_j$ shows the effects of $\mathbf{x}$ on the comparison of $Y = j$ with the geometric mean response $\mathrm{GM}(\mathbf{x})$. Similarly, we can define the simplex-based geometric mean

$$\mathrm{SGM}(\mathbf{x}) = \prod_{j=1}^{k} \{\widetilde{\pi}_j(\mathbf{x})\}^{1/k}.$$

After some algebra, we can show that

$$\log\left(\frac{\widetilde{\pi}_j(\mathbf{x})}{\mathrm{SGM}(\mathbf{x})}\right) = \mathbf{x}^\top \boldsymbol{\beta} \mathbf{w}_j, \quad j = 1, \ldots, k.$$

Thus, SMLR is a generalization of CMLR2 under the simplex-based coding scheme. The modeling of CMLR1 is based on asymmetric comparisons of the categories, while CMLR2 and SMLR use symmetric comparisons. However, because SMLR removes the sum-to-zero constraint in CMLR2 and involves fewer parameters, SMLR is more desirable and computationally efficient.

Note that Zhang and Liu (2014) proposed a flexible multicategory classification framework under the simplex structure that considers a general large-margin loss function $\ell(\cdot)$. According to Theorem 3 in Zhang and Liu (2014), if we consider an exponential loss of the form $\ell(z) = e^{-z}$ and $\ell'(z) = -e^{-z}$, the relationship between the conditional class probability and the theoretical minimizer $\boldsymbol{f}^*$ can be expressed as

$$P_j(\mathbf{x}) = \frac{\ell'(\langle \boldsymbol{f}^*(\mathbf{x}), \mathbf{w}_j \rangle)^{-1}}{\sum_{i=1}^{k} \ell'(\langle \boldsymbol{f}^*(\mathbf{x}), \mathbf{w}_i \rangle)^{-1}} = \frac{\exp(\langle \boldsymbol{f}^*(\mathbf{x}), \mathbf{w}_j \rangle)}{\sum_{s=1}^{k} \exp(\langle \boldsymbol{f}^*(\mathbf{x}), \mathbf{w}_s \rangle)}, \quad \mathbf{x} \in \mathcal{X},\ j = 1, \ldots, k,$$

which recovers the probabilistic assumption (3.2) of the proposed SMLR. This interesting connection sheds some light on SMLR from the perspective of large-margin classification.

3.2 Maximum Likelihood Estimation

The log-likelihood function for the SMLR model (3.2) based on the data set $\mathcal{D}$ is

$$\mathcal{L}_n(\boldsymbol{\beta}) = \ell_n(\boldsymbol{\beta}\mathbf{W}) = \sum_{i=1}^{n} \mathbf{x}_i^\top \boldsymbol{\beta}\mathbf{w}_{y_i} - \sum_{i=1}^{n} \log\Big(\sum_{j=1}^{k} e^{\mathbf{x}_i^\top \boldsymbol{\beta}\mathbf{w}_j}\Big), \tag{3.3}$$

and the corresponding estimator by maximizing the likelihood is defined as

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \ \mathcal{L}_n(\boldsymbol{\beta}). \tag{3.4}$$

Compared with CMLRs, our SMLR solves an unconstrained problem with fewer parameters.

The Fisher scoring algorithm, which is equivalent to Newton's method for the SMLR model, can be used to solve (3.4). For notational simplicity, let $\widetilde{\boldsymbol{\pi}}_i \triangleq \widetilde{\boldsymbol{\pi}}(\mathbf{x}_i)$ be the probability vector for the ith observation, where $\widetilde{\boldsymbol{\pi}}_i$ depends on $\boldsymbol{\beta}$, as determined by (3.2). Let $\boldsymbol{\Lambda}(\mathbf{u}) = \mathbf{diag}(\mathbf{u}) - \mathbf{u}\mathbf{u}^\top$ be a matrix associated with $\mathbf{u}$. After some tedious computation, the score vector is

$$U_n(\boldsymbol{\beta}) = \frac{\partial \mathcal{L}_n(\boldsymbol{\beta})}{\partial \overrightarrow{\boldsymbol{\beta}}} = \sum_{i=1}^{n} (\mathbf{W} \otimes \mathbf{x}_i)(\mathbf{e}_{y_i} - \widetilde{\boldsymbol{\pi}}_i), \tag{3.5}$$

and the negative Hessian matrix is

$$\mathbf{Q}_n(\boldsymbol{\beta}) = -\frac{\partial^2 \mathcal{L}_n(\boldsymbol{\beta})}{\partial \overrightarrow{\boldsymbol{\beta}} \partial \overrightarrow{\boldsymbol{\beta}}^\top} = \sum_{i=1}^{n} [\mathbf{W}\boldsymbol{\Lambda}(\widetilde{\boldsymbol{\pi}}_i)\mathbf{W}^\top] \otimes (\mathbf{x}_i \mathbf{x}_i^\top). \tag{3.6}$$

Observe that $\mathbf{Q}_n(\boldsymbol{\beta})$ involves only the design matrix $\mathbf{X}$, and does not depend on the response Y. In what follows, we consider the setting of a fixed design. Then, $\mathbf{Q}_n(\boldsymbol{\beta})$ is also known as the Fisher information. The following proposition presents some basic results for $\mathbf{Q}_n(\boldsymbol{\beta})$.

Proposition 3. *For any design* $\mathbf{X}$, $\mathbf{Q}_n(\boldsymbol{\beta}) \succeq \mathbf{0}$ *and* $\mathcal{L}_n(\boldsymbol{\beta})$ *is concave in* $\boldsymbol{\beta}$. *In addition,* $\mathbf{Q}_n(\boldsymbol{\beta})$ *is positive definite if and only if the design matrix* $\mathbf{X}$ *is of full column rank.*

Owing to the concavity of $\mathcal{L}_n(\boldsymbol{\beta})$, the MLE $\widehat{\boldsymbol{\beta}}$ is a solution to the likelihood equation $U_n(\boldsymbol{\beta}) = \mathbf{0}$, which can be solved using standard convex programming (Boyd and Vandenberghe, 2004). In addition, the full column rank of the design matrix $\mathbf{X}$ implies that the number of observations n could be as low as d, that is, $n \geq d$. If $\mathbf{Q}_n(\boldsymbol{\beta}) \succ \mathbf{0}$, then $\mathcal{L}_n(\boldsymbol{\beta})$ is strictly concave. In this case, as long as the estimate $\widehat{\boldsymbol{\beta}}$ exists, it must be unique.

The Fisher scoring procedure can be reformulated further as an iteratively reweighted least squares algorithm, and the corresponding updating scheme is

$$\begin{aligned}\overrightarrow{\boldsymbol{\beta}}^{\text{new}} &= \overrightarrow{\boldsymbol{\beta}}^{\text{old}} + \mathbf{Q}_n^{-1}(\boldsymbol{\beta}^{\text{old}}) U_n(\boldsymbol{\beta}^{\text{old}}) \\ &= \overrightarrow{\boldsymbol{\beta}}^{\text{old}} + \Big(\sum_{i=1}^{n} [\mathbf{W}\boldsymbol{\Lambda}(\widetilde{\boldsymbol{\pi}}_i^{\text{old}})\mathbf{W}^\top] \otimes (\mathbf{x}_i \mathbf{x}_i^\top) \Big)^{-1} \Big(\sum_{i=1}^{n} (\mathbf{W} \otimes \mathbf{x}_i)(\mathbf{e}_{y_i} - \widetilde{\boldsymbol{\pi}}_i^{\text{old}}) \Big),\end{aligned}$$

where $\widetilde{\boldsymbol{\pi}}_i^{\text{old}}$ has the jth entry $\widetilde{\pi}_{ij}^{\text{old}} = \frac{e^{\mathbf{x}_i^\top \boldsymbol{\beta}^{\text{old}} \mathbf{w}_j}}{\sum_{s=1}^{k} e^{\mathbf{x}_i^\top \boldsymbol{\beta}^{\text{old}} \mathbf{w}_s}}$, for $j = 1, \ldots, k$.

3.3 Comparison with Regular CMLRs

As a new member of the MLR family, the SMLR is closely related to existing MLRs. The main results are stated in the following theorem.

Theorem 1. *The three estimators* (2.3), (2.4), *and* (3.4) *defined for the same observations* $\mathcal{D}$ *achieve the same log-likelihood values, that is,* $\ell_n(\widehat{\boldsymbol{\theta}}^{\text{rb}}) = \ell_n(\widehat{\boldsymbol{\theta}}^{\text{sc}}) = \mathcal{L}_n(\widehat{\boldsymbol{\beta}})$. *Moreover, if* $\widehat{\boldsymbol{\beta}}$ *is unique, then the regular estimators can be recovered by* $\widehat{\boldsymbol{\theta}}^{\text{rb}} = \widehat{\boldsymbol{\beta}}(\mathbf{W} - \mathbf{w}_r \mathbf{1}_k^\top)$ *and* $\widehat{\boldsymbol{\theta}}^{\text{sc}} = \widehat{\boldsymbol{\beta}}\mathbf{W}$.

Theorem 1 states that the two CMLR estimators $\widehat{\boldsymbol{\theta}}^{\text{rb}}$ and $\widehat{\boldsymbol{\theta}}^{\text{sc}}$ can be uniquely determined by the SMLR estimator $\widehat{\boldsymbol{\beta}}$, under some linear transformations. For any observation $\mathbf{x}$, we

have the following prediction results:

$$\widehat{\pi}_j^{\mathrm{rb}}(\mathbf{x}) = \widehat{\pi}_j^{\mathrm{sc}}(\mathbf{x}) = \widehat{\widetilde{\pi}}_j(\mathbf{x}) = \frac{e^{\mathbf{x}^\top \widehat{\boldsymbol{\beta}}\mathbf{w}_j}}{\sum_{s=1}^k e^{\mathbf{x}^\top \widehat{\boldsymbol{\beta}}\mathbf{w}_s}}.$$

Therefore, these three MLR models are equivalent in terms of probability estimation and label prediction. If an alternative simplex is applied, the estimated coefficients of SMLR may be different, but these three MLR models still share the same prediction outputs.

Theorem 1 also implies connections between the two CMLRs, that is, $\widehat{\boldsymbol{\theta}}^{\mathrm{rb}} = \widehat{\boldsymbol{\theta}}^{\mathrm{sc}}(\mathbf{I}_k - \mathbf{e}_r\mathbf{1}_k^\top)$ and $\widehat{\boldsymbol{\theta}}^{\mathrm{sc}} = \widehat{\boldsymbol{\theta}}^{\mathrm{rb}}(\mathbf{I}_k - \frac{\mathbf{1}_k\mathbf{1}_k^\top}{k})$, where $\mathbf{I}_k$ is a $k \times k$ identity matrix. To compare the three MLRs, we consider a corner case in which there are two distinct categories, that is, $k = 2$. Let $P(\mathbf{x}) = P(Y = 1|\mathbf{x})$ be the probability that the first category happens, conditioning on the given covariates $\mathbf{x}$. Clearly, $P(Y = 2|\mathbf{x}) = 1 - P(\mathbf{x})$. Then, the CMLR1 model with parameter $(\boldsymbol{\theta}, \mathbf{0})$ is

$$P(\mathbf{x}) = \frac{e^{\mathbf{x}^\top\boldsymbol{\theta}}}{e^{\mathbf{x}^\top\boldsymbol{\theta}} + 1} = \frac{1}{1 + e^{-\mathbf{x}^\top\boldsymbol{\theta}}},$$

where the second category is viewed as a reference. For CMLR2, owing to the sum-to-zero constraint $\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2 = \mathbf{0}$, we can simplify $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ as $(\boldsymbol{\theta}_1, -\boldsymbol{\theta}_1)$, yielding the model

$$P(\mathbf{x}) = \frac{e^{\mathbf{x}^\top\boldsymbol{\theta}_1}}{e^{\mathbf{x}^\top\boldsymbol{\theta}_1} + e^{\mathbf{x}^\top\boldsymbol{\theta}_2}} = \frac{1}{1 + e^{\mathbf{x}^\top(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)}} = \frac{1}{1 + e^{-2\mathbf{x}^\top\boldsymbol{\theta}_1}}.$$

The SMLR model (3.2) with parameter $\boldsymbol{\beta}$ becomes

$$P(\mathbf{x}) = \frac{e^{\mathbf{x}^\top\boldsymbol{\beta}}}{e^{\mathbf{x}^\top\boldsymbol{\beta}} + e^{-\mathbf{x}^\top\boldsymbol{\beta}}} = \frac{1}{1 + e^{-2\mathbf{x}^\top\boldsymbol{\beta}}}.$$

Denote the three corresponding estimates as $(\widehat{\boldsymbol{\theta}}, \mathbf{0})$, $(\widehat{\boldsymbol{\theta}}_1, -\widehat{\boldsymbol{\theta}}_1)$, and $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^d$, respectively. Based on Theorem 1, we have $\widehat{\boldsymbol{\theta}} = 2\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}_1 = \widehat{\boldsymbol{\beta}}$. This leads to the following corollary.

Corollary 1. *If $k = 2$, the three MLR models are equivalent up to a scaling factor.*

Theorem 1 can also be used to reveal some interesting properties of the regular CMLRs. For example, consider the relation between the CMLR1 estimators under two different reference categories, r and s. By Theorem 1, we have $\widehat{\boldsymbol{\theta}}^{(\mathrm{r})} = \widehat{\boldsymbol{\beta}}(\mathbf{W} - \mathbf{w}_r\mathbf{1}_k^\top)$ and $\widehat{\boldsymbol{\theta}}^{(\mathrm{s})} = \widehat{\boldsymbol{\beta}}(\mathbf{W} - \mathbf{w}_s\mathbf{1}_k^\top)$. The relation between $\widehat{\boldsymbol{\theta}}^{(\mathrm{r})}$ and $\widehat{\boldsymbol{\theta}}^{(\mathrm{s})}$ is shown in the following corollary.

Corollary 2. *For the CMLR1 model with two different reference categories r and s, $\widehat{\boldsymbol{\theta}}^{(\mathrm{r})}$ and $\widehat{\boldsymbol{\theta}}^{(\mathrm{s})}$ have the following relation:*

$$\widehat{\boldsymbol{\theta}}^{(\mathrm{r})} = \widehat{\boldsymbol{\theta}}^{(\mathrm{s})} - \widehat{\boldsymbol{\theta}}_r^{(\mathrm{s})}\mathbf{1}_k^\top.$$

For illustration, consider $k = 2$ and an available estimate $(\widehat{\boldsymbol{\theta}}, \mathbf{0})$ for the CMLR1 model with the second reference. By Corollary 2, the estimate for CMLR1 with the first reference is $(\mathbf{0}, -\widehat{\boldsymbol{\theta}})$.

According to Theorem 1, it is possible to transfer the properties of regular CMLRs to the SMLR framework. For instance, Albert and Anderson (1984) showed that the MLE for the CMLR1 model exists only when the data sets overlap, where existence means the finiteness of an estimate. We can extend the concept of overlapping to the SMLR model, as follows.

Definition 1 (Overlapping)**.** We say that the observations $\mathcal{D}$ are overlapping if for every nonzero matrix $\boldsymbol{\beta} \in \mathbb{R}^{d\times(k-1)}$, there exists a duplet (i, t), with $i \in \{1, \ldots, n\}$ and $t \in \mathcal{Y}\backslash y_i$, such that $\mathbf{x}_i^\top\boldsymbol{\beta}(\mathbf{w}_{y_i} - \mathbf{w}_t) < 0$.

Applying Theorem 1, we have the following corollary for the existence of the SMLR estimator.

Corollary 3. *The MLE for the SMLR model exists if and only if the observations $\mathcal{D}$ overlap.*

Although the three MLRs are closely connected, the SMLR model treats all categories equally, and provides systematic insights in a concise framework without further constraints. Hence, the SMLR model serves as a unified framework that includes the binomial LR and two regular CMLRs. In what follows, we focus on the SMLR model for succinct technical presentation. All conclusions for the SMLR model can be adapted to regular CMLRs using the link functions in Theorem 1.

4. Theoretical Properties

In this section, we establish the asymptotical results and statistical learning theory for the proposed SMLR model, including the existence and uniqueness of the MLE, Fisher consistency, the comparison inequalities as a classifier, and the convergence results for the kernel SMLR.

4.1 Asymptotical Results

We are interested in the asymptotic behavior of the SMLR model under a complicated diverging setting, where both the number of covariates and the number of categories can increase with the sample size. In the literature, the diverging number of covariates is well studied, but few studies examine the diverging number of categories, even though such a setting is not uncommon in practice. In the general "large n, diverging d and diverging k" setup, we denote $d = d_n$ and $k = k_n$ to emphasize the effect of the sample size n. Then, the number of parameters for the SMLR is $(k_n - 1)d_n$. Hereafter, we replace $(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}})$ by $(\boldsymbol{\beta}_n, \widehat{\boldsymbol{\beta}}_n)$ to emphasize their dependencies on the sample size n.

For the matrix $\mathbf{\Lambda}(\widetilde{\boldsymbol{\pi}}_i)$ in $\mathbf{Q}_n(\boldsymbol{\beta}_n)$, because $\widetilde{\boldsymbol{\pi}}_i$ depends on $\boldsymbol{\beta}_n$, we rewrite $\mathbf{\Lambda}_i(\boldsymbol{\beta}_n) = \mathbf{\Lambda}(\widetilde{\boldsymbol{\pi}}_i)$. Hence, the Fisher information matrix $\mathbf{Q}_n(\boldsymbol{\beta}_n) = \sum_{i=1}^n [\mathbf{W}\mathbf{\Lambda}_i(\boldsymbol{\beta}_n)\mathbf{W}^\top] \otimes (\mathbf{x}_i\mathbf{x}_i^\top)$ is a function of $\boldsymbol{\beta}_n$. Assume that the true coefficient matrix is $\boldsymbol{\beta}_{n0}$. Denote $\mathbf{G}_{n0} = \mathbf{Q}_n(\boldsymbol{\beta}_{n0})$. When the underlying model is correctly specified,

$$P(Y = j|X = \mathbf{x}) = \frac{e^{\mathbf{x}^\top \boldsymbol{\beta}_{n0}\mathbf{w}_j}}{\sum_{s=1}^k e^{\mathbf{x}^\top \boldsymbol{\beta}_{n0}\mathbf{w}_s}}.$$

For the score vector $U_n(\cdot)$ defined in (3.5), it follows that $E[U_n(\boldsymbol{\beta}_{n0})] = \mathbf{0}$. Define $\mathbf{S}_n = \mathbf{X}^\top\mathbf{X} = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$. Because the covariates $\mathbf{x}_i$ are assumed to be deterministic, $\mathbf{Q}_n(\boldsymbol{\beta}_n)$ and $\mathbf{S}_n$ are not random. Given a fixed $\delta > 0$, we consider a region of interest for $\boldsymbol{\beta}_n$, that is, $N_n(\delta) = \{\boldsymbol{\beta} : \|\mathbf{G}_{n0}^{1/2}(\overrightarrow{\boldsymbol{\beta}} - \overrightarrow{\boldsymbol{\beta}_{n0}})\| \leq \delta\{d_n(k_n - 1)\}^{1/2}\}$.

In order to obtain the asymptotic results, we need the following assumptions.

Assumption 1. The true parameter $\boldsymbol{\beta}_{n0}$ is contained in the interior of a compact subset $\mathcal{B}_n$ in $\mathbb{R}^{d_n \times (k_n - 1)}$.

Assumption 2. Each covariate of X is uniformly bounded by a constant $C > 0$.

Assumption 3. There exist two positive constants c_1 and c_2 such that $c_1 \leq \lambda_{\min}(\mathbf{S}_n/n) \leq \lambda_{\max}(\mathbf{S}_n/n) \leq c_2$.

Assumption 4. There exists a constant $L_0 > 0$ such that $\|\mathbf{\Lambda}(\boldsymbol{\beta}) - \mathbf{\Lambda}(\boldsymbol{\beta}')\|_2 \leq L_0 \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|$, for any $\boldsymbol{\beta}, \boldsymbol{\beta}' \in B_n$.

The above assumptions are motivated by the literature on M-estimation with a diverging dimension of covariates; see Portnoy (1985), Portnoy (1988), He and Shao (2000), Wang (2011), Liang and Du (2012), and Gao et al. (2018). Assumption 1 restricts the true parameter $\boldsymbol{\beta}_{n0}$ under the doubly diverging setting. Assumption 2 ensures that each observed

covariate vector is bounded, that is, $\|\mathbf{x}_i\| \leq C\sqrt{d_n}$, for any $i = 1, \ldots, n$. Assumption 3 gives some regular conditions on the design matrix $\mathbf{X}$. Assumption 4 is specially tailored for the SMLR model, and can be viewed as a generalized Lipschitz condition.

Remark 1. Assumption 3 is also used by Wang (2011) and Liang and Du (2012) to establish the asymptotics for a binomial LR with diverging covariates. Based on Assumption 3, we can find a positive constant c_0 such that $c_0 \leq \max_{i=1,\ldots,n} \min_{j=1\ldots,k} \widetilde{\pi}_{ij}(\boldsymbol{\beta}_{n0})$. Furthermore, using Lemma 4 in the Supplementary Material, for the Fisher information matrix $\mathbf{G}_{n0}$, we have

$$\frac{\mathbf{G}_{n0}}{n} = \frac{1}{n}\sum_{i=1}^{n}[\mathbf{W}\boldsymbol{\Lambda}_i(\boldsymbol{\beta}_n)\mathbf{W}^\top] \otimes (\mathbf{x}_i\mathbf{x}_i^\top) \succeq \frac{k_n c_0}{k_n - 1}\,\frac{\mathbf{S}_n}{n} \succeq c\mathbf{I}_{(k_n-1)d_n}, \tag{4.1}$$

with a constant $c \in (0, c_0 c_1]$. If $k = 2$, the result (4.1) is reduced to that of Liang and Du (2012). In summary, there exists a positive constant c such that $\lambda_{\min}(\mathbf{G}_{n0}/n) \geq c$, and (4.1) is required to establish the convergence and asymptotic normality of the estimated coefficients under certain diverging settings.

Remark 2. Under Assumption 2, we have $\|\mathbf{S}_n/n\|_2 \leq n^{-1}\sum_{i=1}^{n}\mathbf{x}_i^\top\mathbf{x}_i \leq d_n C^2$. Therefore, Assumption 3 holds naturally in the "fixed d" case.

Remark 3. Owing to the complicated nature of an MLR, Assumption 4 is necessary to control the behavior of different coefficients on the component of the Fisher information matrix, which depicts the continuousness of a matrix function in another manner. Technical proofs of the consistency and asymptotic normality can be simplified using Assumption 4.

Next, when the number of parameters $d_n(k_n - 1)$ increases, the existence and consistency of the MLE is guaranteed by the following theorem.

Theorem 2. *Suppose Assumptions 1–4 hold. If* $\sqrt{d_n k_n/n} \to 0$*, then there exists a sequence* $\{\widehat{\boldsymbol{\beta}}_n\}$ *such that as* $n \to \infty$*,*

$$P\Big(\widehat{\boldsymbol{\beta}}_n \in N_n(\delta) \text{ and } U_n(\widehat{\boldsymbol{\beta}}_n) = \mathbf{0}\Big) \to 1 \text{ and } \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\| = O_p\big(\sqrt{d_n k_n/n}\big).$$

For the classical setting with fixed d and k, the conditions become $1/n \to 0$. For a fixed dimension d and varying categories k_n, we require $k_n = o(n)$. When the number of categories k is fixed, d_n should satisfy $d_n = o(n)$.

The following theorem ensures the asymptotic normality of $\widehat{\boldsymbol{\beta}}_n$.

Theorem 3. *Suppose Assumptions 1–4 hold. If* $d_n k_n/\sqrt{n} \to 0$*, then for any* $d_n(k_n - 1) \times l$ *matrix* $\mathbf{V}_n$ *with* l *fixed and such that* $\mathbf{V}_n^\top \mathbf{V}_n = \mathbf{I}_l$*,*

$$\mathbf{V}_n^\top \mathbf{G}_{n0}^{1/2}\Big(\overrightarrow{\widehat{\boldsymbol{\beta}}_n} - \overrightarrow{\boldsymbol{\beta}_{n0}}\Big) \longrightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_l), \text{ in distribution,}$$

where $\mathcal{N}(\cdot, \cdot)$ *is a multivariate normal distribution.*

In particular, let $\mathbf{U}_n = \mathbf{G}_{n0}^{-1/2}\mathbf{V}_n\big(\mathbf{V}_n^\top \mathbf{G}_{n0}^{-1}\mathbf{V}_n\big)^{-1/2}$. Then $\mathbf{U}_n^\top \mathbf{U}_n = \mathbf{I}_l$. Based on Theorem 3, we have the following corollary which gives the asymptotic distribution of $\mathbf{V}_n^\top\Big(\overrightarrow{\widehat{\boldsymbol{\beta}}_n} - \overrightarrow{\boldsymbol{\beta}_{n0}}\Big)$.

Corollary 4. *Under the same conditions as in Theorem 3, as* $n \to \infty$*, we have*

$$\big(\mathbf{V}_n^\top \mathbf{G}_{n0}^{-1}\mathbf{V}_n\big)^{-1/2}\mathbf{V}_n^\top\Big(\overrightarrow{\widehat{\boldsymbol{\beta}}_n} - \overrightarrow{\boldsymbol{\beta}_{n0}}\Big) \longrightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_l), \text{ in distribution.}$$

If d and k are fixed, we only need $n^{-1/2} \to 0$, which is trivially true. When the number of categories k_n diverges and d is fixed, the conditions in Theorem 3 reduce to $k_n/\sqrt{n} \to 0$, which implies a sufficient condition $k_n = o\big(\sqrt{n}\big)$. For a diverging number of covariates d_n and fixed k, we require $d_n/\sqrt{n} \to 0$, that is, $d_n = o(\sqrt{n})$. In particular, for the binomial LR

model, we recover the result in Portnoy (1985, 1988), and our result is stronger than those of He and Shao (2000) and Wang (2011). Hence, our conditions in Theorem 2 and 3 are general for MLR models with a diverging number of parameters, even with a diverging number of categories.

The following theorem and its corollary suggest that one can approximate $\mathbf{G}_{n0}$ using $\mathbf{Q}_n(\widehat{\boldsymbol{\beta}}_n)$ when applying Theorem 3 and Corollary 4 for interval estimation.

Theorem 4. *Under the same conditions as in Theorem 3, as $n \to \infty$, we have*

$$\left\|\mathbf{V}_n^\top \mathbf{Q}_n^{-1}(\widehat{\boldsymbol{\beta}}_n)\mathbf{V}_n - \mathbf{V}_n^\top \mathbf{G}_{n0}^{-1}\mathbf{V}_n\right\|_2 \longrightarrow 0, \text{ in probability.}$$

Corollary 5. *Under the same conditions as in Theorem 3, as $n \to \infty$, we have*

$$\left[\mathbf{V}_n^\top \mathbf{Q}_n^{-1}(\widehat{\boldsymbol{\beta}}_n)\mathbf{V}_n\right]^{-\frac{1}{2}} \mathbf{V}_n^\top \left(\overrightarrow{\widehat{\boldsymbol{\beta}}_n} - \overrightarrow{\boldsymbol{\beta}_{n0}}\right) \longrightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_l), \text{ in distribution,}$$

and

$$\left(\overrightarrow{\widehat{\boldsymbol{\beta}}_n} - \overrightarrow{\boldsymbol{\beta}_{n0}}\right)^\top \mathbf{V}_n \left[\mathbf{V}_n^\top \mathbf{Q}_n^{-1}(\widehat{\boldsymbol{\beta}}_n)\mathbf{V}_n\right]^{-1} \mathbf{V}_n^\top \left(\overrightarrow{\widehat{\boldsymbol{\beta}}_n} - \overrightarrow{\boldsymbol{\beta}_{n0}}\right) \longrightarrow \chi_l^2, \text{ in distribution,}$$

where χ_l^2 is the χ^2 distribution with l degrees of freedom.

According to Corollary 5 and the MLE, an asymptotic $1-\alpha$ confidence interval ($0 < \alpha < 1$) for β_{ij} is

$$\widehat{\beta}_{ij} \pm z_{\alpha/2} \mathbf{v}_{ij}^\top \top \mathbf{Q}_n^{-1} \mathbf{v}_{ij},$$

where $z_{\alpha/2}$ denotes the upper $(\alpha/2)$-quantile of the standard normal distribution, and $\mathbf{v}_{ij}$ is the unit vector of length $d_n(k_n - 1)$, with the $[(j-1)d_n + i]$th element equal to one and all other elements equal to zero. We can also apply these results when testing the following linear hypothesis:

$$H_0 : \mathbf{V}_n^\top \overrightarrow{\boldsymbol{\beta}_{n0}} = \boldsymbol{a} \iff H_1 : \mathbf{V}_n^\top \overrightarrow{\boldsymbol{\beta}_{n0}} \neq \boldsymbol{a},$$

where the vector $\boldsymbol{a} \in \mathbb{R}^l$ is known and $\mathbf{V}_n$ is a $\{(k_n - 1)d_n\} \times l$ matrix such that $\mathbf{V}_n^\top \mathbf{V}_n = \mathbf{I}_l$. The large-sample Wald test statistic is defined as

$$T_n = \Big(\overrightarrow{\widehat{\boldsymbol{\beta}}_n}^\top \mathbf{V}_n - \boldsymbol{a}^\top\Big)\big[\mathbf{V}_n^\top \mathbf{Q}_n^{-1}(\widehat{\boldsymbol{\beta}}_n)\mathbf{V}_n\big]^{-1}\Big(\mathbf{V}_n^\top \overrightarrow{\widehat{\boldsymbol{\beta}}_n} - \boldsymbol{a}\Big).$$

Corollary 5 shows that the Wald test remains valid, that is, $T_n \longrightarrow \chi_l^2$ in distribution under the null hypothesis H_0, even when the numbers of covariates and categories diverge with the sample size.

4.2 Fisher Consistency and Error Analysis

Fisher consistency is a fundamental property for classifiers, and is also called infinite-sample consistency by Zhang (2004) and classification calibration by Bartlett et al. (2006) and Tewari and Bartlett (2007). In this section, we show these desired properties for the SMLR model.

First, let $P_j(\mathbf{x}) = P(Y = j|X = \mathbf{x})$ be the underlying class conditional probability for any $\mathbf{x} \in \mathcal{X}$, and define a vector $\mathbf{p}(\mathbf{x}) = (P_1(\mathbf{x}), \ldots, P_k(\mathbf{x}))^\top$. Consider a classifier $\mathcal{C} : \mathcal{X} \mapsto \mathcal{Y}$. The expected misclassification error is given by

$$\mathcal{R}(\mathcal{C}) = E[\mathbb{1}(Y \neq \mathcal{C}(X))] = 1 - E_X[P(Y = \mathcal{C}(\mathbf{x})|X = \mathbf{x})].$$

One can verify that the optimal classifier minimizing $\mathcal{R}(\mathcal{C})$, often called the Bayes rule, is denoted as $\mathcal{C}_{\mathrm{B}}(\mathbf{x}) = \arg\max_j P_j(\mathbf{x})$. Denote $\mathcal{R}^* = \mathcal{R}(\mathcal{C}_{\mathrm{B}}) = 1 - E_X[\max_j P_j(X)]$.

Under the simplex coding scheme, it is sufficient to directly use $k - 1$ functions for multicategory classification. Let $\boldsymbol{f} = (f_1, \ldots, f_{k-1})^\top : X \mapsto \mathbb{R}^{k-1}$ be a generic classification function. Then the prediction rule induced by $\boldsymbol{f}$ is $\mathcal{C}_{\boldsymbol{f}}(\mathbf{x}) = \arg\max_j \langle \boldsymbol{f}(\mathbf{x}), \mathbf{w}_j \rangle$. For the SMLR log-likelihood (3.3), we can define the SMLR loss function for an observation $(\mathbf{x}, y)$

as

$$V(\boldsymbol{f}(\mathbf{x}), y) = \log\left(\sum_{j=1}^{k} e^{\langle \boldsymbol{f}(\mathbf{x}), \mathbf{w}_j \rangle}\right) - \langle \boldsymbol{f}(\mathbf{x}), \mathbf{w}_y \rangle. \tag{4.2}$$

We are interested in the expected V-risk

$$\mathcal{E}(\boldsymbol{f}) = E[V(\boldsymbol{f}(X), Y)] = E_X\{E[V(\boldsymbol{f}(X), Y)|X]\}.$$

Consider the hypothesis space

$$\mathfrak{F} = \{\boldsymbol{f} : \mathcal{X} \mapsto \mathbb{R}^{k-1} \mid E_X[\|\boldsymbol{f}(X)\|] < \infty\},$$

where $\|\cdot\|$ is the standard Euclidean norm in $\mathbb{R}^{k-1}$. Note that $\mathcal{E}(\boldsymbol{f})$ is a functional of $\boldsymbol{f}$, and we define its minimizer over $\mathfrak{F}$ as $\boldsymbol{f}^* = \arg\min_{\boldsymbol{f}\in\mathfrak{F}} \mathcal{E}(\boldsymbol{f})$. Fisher consistency requires that $\mathcal{C}_{\mathrm{B}}(\mathbf{x}) = \mathcal{C}_{\boldsymbol{f}^*}(\mathbf{x})$, for any $\mathbf{x} \in \mathcal{X}$.

Theorem 5. *Assume that $P_j(\mathbf{x}) > 0$, for $j = 1, \ldots, k$. The expected V-risk $\mathcal{E} : \mathfrak{F} \mapsto \mathbb{R}_+$ is a convex and continuous functional, with the minimizer $\boldsymbol{f}^*(\mathbf{x}) = (1 - 1/k)\sum_{i=1}^{k}[\log P_i(\mathbf{x})]\mathbf{w}_i$. Moreover, the SMLR loss function* (4.2) *is Fisher consistent.*

Theorem 5 states a one-to-one correspondence between $\boldsymbol{f}^*$ and $\mathbf{p}$, where the argument $\mathbf{x}$ is suppressed for brevity. Specifically, the explicit form of $\boldsymbol{f}^*$ is a linear expression of $\mathbf{w}_i$, with coefficients determined uniquely by P_j. Because $\langle \boldsymbol{f}^*(\mathbf{x}), \mathbf{w}_j \rangle = \log P_j(\mathbf{x}) - \frac{1}{k}\sum_{i=1}^{k} \log P_i(\mathbf{x})$, for $j = 1, \ldots, k$, we conclude that if $P_i > P_j$, then $\langle \boldsymbol{f}^*, \mathbf{w}_i \rangle > \langle \boldsymbol{f}^*, \mathbf{w}_j \rangle$, and if $P_i = P_j$, then $\langle \boldsymbol{f}^*, \mathbf{w}_i \rangle = \langle \boldsymbol{f}^*, \mathbf{w}_j \rangle$. However, if $P_j \to 0$, $\boldsymbol{f}^*$ becomes unbounded and meaningless.

For a classification function $\boldsymbol{f}$, $\mathcal{R}(\mathcal{C}_{\boldsymbol{f}}) - \mathcal{R}^*$ and $\mathcal{E}(\boldsymbol{f}) - \mathcal{E}(\boldsymbol{f}^*)$ are called the excess misclassification risk and the excess V-risk in SMLR, respectively. Then, the following theorem provides an essential comparison inequality.

Theorem 6. *For any* $\boldsymbol{f} \in \mathfrak{F}$*, we have* $\mathcal{R}(\mathcal{C}_{\boldsymbol{f}}) - \mathcal{R}^* \leq \sqrt{2}\{\mathcal{E}(\boldsymbol{f}) - \mathcal{E}(\boldsymbol{f}^*)\}^{1/2}$.

Theorem 6 covers the results for the binary LR in Bartlett et al. (2006). The upper bound can be improved under some regularity conditions. To this end, we introduce the following generalized Tsybakov's low-noise assumption (Tsybakov, 2004).

Assumption 5. Let $P_{(1)}(\mathbf{x})$ and $P_{(2)}(\mathbf{x})$ be the largest conditional probability and the second largest conditional probability, respectively. Assume that there exist $C > 0$ and $\alpha \geq 0$ such that for all $0 < h \leq 1$,

$$P_X(\{\mathbf{x} \in \mathcal{X} | P_{(1)}(\mathbf{x}) - P_{(2)}(\mathbf{x}) \leq h\}) \leq Ch^{\alpha}. \tag{4.3}$$

Intuitively, it is clear that the misclassification error is particularly large when it is difficult to separate the class with the highest probability from the others. In many multicategory classification problems, it is reasonable to assume that the $P_{(1)}(\mathbf{x})$ is unlikely to be very close to $P_{(2)}(\mathbf{x})$, for $\mathbf{x} \in \mathcal{X}$. Hence, Assumption 5 is a meaningful low-noise condition that depends on the parameter α. We consider two extreme values of α. If $\alpha = 0$, this imposes the case without any assumption on the noise, as discussed in Theorem 6. If $\alpha = \infty$, it reduces to the noiseless case.

We can improve the bound for the excess misclassification risk under Assumption 5. Note that similar results are established in Theorem 2 of Mroueh et al. (2012) for multicategory support vector machines equipped with a hinge loss or a quadratic loss. Theorem 7 establishes the results for the MLR models.

Theorem 7. *For each* $\boldsymbol{f} \in \mathfrak{F}$*, if Assumption 5 holds, then we have*

$$\mathcal{R}(\mathcal{C}_{\boldsymbol{f}}) - \mathcal{R}^* \leq (8C_{\alpha})^{\frac{\alpha+1}{\alpha+2}}\{\mathcal{E}(\boldsymbol{f}) - \mathcal{E}(\boldsymbol{f}^*)\}^{\frac{\alpha+1}{\alpha+2}},$$

where $C_\alpha = (\alpha+1)C^{\frac{1}{\alpha+1}}\alpha^{-\frac{\alpha}{\alpha+1}} > 0$ *is a constant.*

Remarkably, Theorem 6 is a particular case of Theorem 7 with $\alpha = 0$. Furthermore, Theorem 7 is a refined version of Theorem 6, because $\frac{\alpha+1}{\alpha+2} > 1/2$ when $\alpha > 0$.

4.3 Convergence Analysis of Kernel SMLR

Motivated by the kernel MLR of Zhu and Hastie (2005), we investigate the consistency of the kernel SMLR and conduct a convergence analysis under a diverging number of categories, which few works have done.

To start with, we consider a Mercer kernel K defined over $\mathcal{X} \times \mathcal{X}$ and the reproducing kernel Hilbert space H_K induced by K with the inner product $\langle \cdot, \cdot \rangle$, stratifying $\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle = K(\mathbf{x}, \mathbf{x}')$ with a feature map $\phi : \mathcal{X} \mapsto H_K$. Then, we introduce the following notation:

$$\mathcal{F} = \Big\{ \boldsymbol{f} : \mathcal{X} \mapsto \mathbb{R}^{k-1},\ f_j(\mathbf{x}) = \langle \mathbf{u}_j, \phi(\mathbf{x}) \rangle \mid E_X[\|\boldsymbol{f}(X)\|] < \infty,\ \mathbf{u}_j \in H_K,\ \forall j = 1, \ldots, k-1 \Big\},$$

$$\mathcal{F}_A = \Big\{ \boldsymbol{f} \in \mathcal{F} \ \Big|\ \Big(\sum\nolimits_{j=1}^{k-1} \langle \mathbf{u}_j, \mathbf{u}_j \rangle \Big)^{1/2} \leq A \Big\} \subseteq \mathcal{F},$$

where $A > 0$ is a constant used to bound the hypothesis in $\mathcal{F}$. Note that if $\phi(\mathbf{x}) = \mathbf{x}$, the classical setting of linear learning is recovered. We also need some technical assumptions about the boundedness of the kernel function.

Assumption 6. There exists a constant $C > 0$ such that $\sqrt{K(\mathbf{x}, \mathbf{x})} \leq C$, for any $\mathbf{x} \in \mathcal{X}$.

Under Assumption 6, we know $E_X[\|\boldsymbol{f}(X)\|] \leq AC$ for any $\boldsymbol{f} \in \mathcal{F}_A$, which ensures $\mathcal{F}_A \subseteq \mathcal{F}$. Theoretically, the expected risk minimizer for the kernel SMLR model is denoted as

$$\boldsymbol{f}^* = \arg\min_{\boldsymbol{f} \in \mathcal{F}}\ \mathcal{E}(\boldsymbol{f}) = E[V(\boldsymbol{f}(X), Y)].$$

To ensure the uniqueness of $\boldsymbol{f}^*$, we further define $\boldsymbol{f}^* = \arg\min_{\boldsymbol{f}\in\mathcal{B}}(\sum_{j=1}^{k-1}\langle \mathbf{u}_j, \mathbf{u}_j\rangle)^{1/2}$, where $\mathcal{B} = \{\boldsymbol{f} \in \mathcal{F} | \boldsymbol{f} = \arg\min_{\boldsymbol{f}\in\mathcal{F}} \mathcal{E}(\boldsymbol{f})\}$. In fact, our aim is to learn the empirical risk minimizer from $\mathcal{F}_A$, defined by

$$\widehat{\boldsymbol{f}} = \arg\min_{\boldsymbol{f}\in\mathcal{F}_A} \widehat{\mathcal{E}}_{\mathcal{D}}(\boldsymbol{f}) = \frac{1}{n}\sum_{i=1}^{n} V(\boldsymbol{f}(\mathbf{x}_i), y_i).$$

Proposition 4. *Suppose Assumption 6 is met. For the function class $\mathcal{F}_A$, we have*

$$E_{\mathcal{D}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}_A}\left|\frac{1}{n}\sum_{i=1}^{n} V(\boldsymbol{f}(\mathbf{x}_i), y_i) - \mathcal{E}(\boldsymbol{f})\right|\right] \leq 4\sqrt{2}AC(k-1)^{1/2}n^{-1/2}.$$

The concentration inequality in Proposition 4 involves the expectation of the supremum of the empirical process, and it is a main tool to establish the consistency of $\widehat{\boldsymbol{f}}$ under the diverging number of categories, as shown in the following Theorem.

Theorem 8. *Assume Assumption 6 holds and there exists a proper A such that the theoretical minimizer $\boldsymbol{f}^* \in \mathcal{F}_A$. If $\sqrt{(k_n-1)/n} \to 0$, then we have*

$$\lim_{n\to\infty} E_{\mathcal{D}}[\mathcal{E}(\widehat{\boldsymbol{f}})] = \min_{\boldsymbol{f}\in\mathcal{F}_A} \mathcal{E}(\boldsymbol{f}) = \mathcal{E}(\boldsymbol{f}^*).$$

Theorem 8 implies that the consistency of $\widehat{\boldsymbol{f}}$ requires that k should increase at the order $o(n)$. In particular, this order is identical to that of Theorem 2 for the linear SMLR model. The main difference is that Theorem 2 is established for the MLE of the linear SMLR, whereas Theorem 8 is based on the negative log-likelihood loss function for the kernel SMLR. For the kernel CMLR2 with fixed k, Zhang (2004) showed a convergence result under the condition $\sqrt{k/n}\ln^{3/2} n \to 0$. On the other hand, our Theorem 8 requires $\sqrt{(k-1)/n} \to 0$, which is faster than that of Zhang (2004).

5. Numerical Studies

In this section, we conduct several experiments to demonstrate the numerical performance of the proposed SMLR model. In particular, we study three simulated examples in Section 5.1, and consider two real-world applications in Sections 5.2 and 5.3.

5.1 Simulated Examples

Consider the following SMLR model:

$$\widetilde{\pi}_y(\mathbf{x}_i) = \frac{e^{\mathbf{x}_i^\top \boldsymbol{\beta}_{n0} \mathbf{w}_y}}{\sum_{j=1}^{k} e^{\mathbf{x}_i^\top \boldsymbol{\beta}_{n0} \mathbf{w}_j}}, \quad i = 1, \ldots, n; \ y \in \mathcal{Y}, \tag{5.1}$$

where $\boldsymbol{\beta}_{n0}$ is a $d_n \times (k_n - 1)$ matrix of parameters. Specifically, $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id_n})^\top$, for $i = 1, \ldots, n$, are independently generated from a multivariate normal distribution with mean zero and marginal standard deviation 0.2. In the simulation, we concentrate on the model (5.1) with a diverging number of parameters, with the following explicit settings. Let $\lfloor a \rfloor$ be the largest integer not greater than the number a.

Example 1 (diverging k and fixed d)**.** Consider $k_n = \lfloor \sqrt{n} \rfloor$ and $d = 3$. Let $\mathbf{a}_n = \{(k_n - 1)d\}^{-1/2} \mathbf{1}_{(k_n-1)d}$. The true parameter matrix is

$$\boldsymbol{\beta}_{n0} = \left(\tfrac{k_n-1}{k_n} \cdot \mathbf{1}_d \quad -\tfrac{k_n-2}{k_n} \cdot \mathbf{1}_d \quad \cdots \quad (-1)^{k_n-1} \tfrac{2}{k_n} \cdot \mathbf{1}_d \quad (-1)^{k_n} \tfrac{1}{k_n} \cdot \mathbf{1}_d \right).$$

Example 2 (diverging d and fixed k)**.** Consider $d_n = \lfloor 2\sqrt{n} \rfloor$ and $k = 3$. Let $d_0 = \lfloor d_n/4 \rfloor$

and $\mathbf{V}_n = \begin{pmatrix} \mathbf{v}_{n1} & \mathbf{v}_{n2} \end{pmatrix} = d_n^{-1/2}\mathbf{I}_2 \otimes \mathbf{1}_{d_n}$. The true parameter matrix is

$$\boldsymbol{\beta}_{n0} = \begin{pmatrix} 0.4 \cdot \mathbf{1}_{d_0} & -0.1 \cdot \mathbf{1}_{d_0} \\ -0.3 \cdot \mathbf{1}_{d_0} & 0.2 \cdot \mathbf{1}_{d_0} \\ 0.2 \cdot \mathbf{1}_{d_0} & -0.3 \cdot \mathbf{1}_{d_0} \\ -0.1 \cdot \mathbf{1}_{d_n - 3d_0} & 0.4 \cdot \mathbf{1}_{d_n - 3d_0} \end{pmatrix}.$$

Example 3 (diverging k and d simultaneously)**.** Consider $k_n = \lfloor 3n^{1/4} \rfloor$and $d_n = \lfloor 2n^{1/4} \rfloor$. Let $d_0 = \lfloor d_n/2 \rfloor$ and $\mathbf{b}_n = \{(k_n - 1)d_n\}^{-1/2}\mathbf{1}_{(k_n-1)d_n}$. The true parameter matrix is

$$\boldsymbol{\beta}_{n0} = \begin{pmatrix} \frac{k_n-1}{k_n} \cdot \mathbf{1}_{d_0} & -\frac{k_n-2}{k_n} \cdot \mathbf{1}_{d_0} & \cdots & (-1)^{k_n-1}\frac{2}{k_n} \cdot \mathbf{1}_{d_0} & (-1)^{k_n}\frac{1}{k_n} \cdot \mathbf{1}_{d_0} \\ -\frac{k_n-1}{k_n} \cdot \mathbf{1}_{d_n-d_0} & \frac{k_n-2}{k_n} \cdot \mathbf{1}_{d_n-d_0} & \cdots & (-1)^{k_n}\frac{2}{k_n} \cdot \mathbf{1}_{d_n-d_0} & (-1)^{k_n+1}\frac{1}{k_n} \cdot \mathbf{1}_{d_n-d_0} \end{pmatrix}.$$

The sample size n varies from $\{100, 500, 1000, 5000, 10000\}$ for each example, and we conduct 10000 replications for each simulation. Because the dimension of the estimated coefficient matrix changes as n increases, we measure the accuracy of the estimation using the simulated average mean squared error (AMSE), which is obtained by averaging $\frac{\left\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\right\|^2}{d_n(k_n-1)}$ over all simulated samples. We also report the asymptotic behaviors of some linear combinations of $\widehat{\boldsymbol{\beta}}_n$, including the average estimation bias (Bias), standard error of the estimates (SE_emp), average estimated standard error (SE_est), and coverage probability (CP) of a 95% confidence interval over 10000 replications. All simulations are implemented in `R` (R Core Team, 2021).

From the theoretical results in Section 4.1, we can design certain diverging settings, as shown in Examples 1–3. Tables 1–3 demonstrate the simulation results for the three considered examples, where #para denotes the total number of unknown parameters. In general, the biases and standard errors decrease as the sample size n increases, and the coverage probabilities are close to the nominal value when n is large. These results suggest

that the performance of the MLE is satisfactory under the considered diverging settings, and that the asymptotic properties are valid.

Table 1: Results for simulated Example 1.

n	k	#para	AMSE	Bias	SE_emp	SE_est	CP
100	10	27	2.9282	0.0110	1.7587	1.6932	0.9278
500	22	63	1.0520	0.0114	1.1039	1.0905	0.9326
1000	31	90	0.7218	0.0023	0.8439	0.8458	0.9422
5000	70	207	0.3473	0.0080	0.6015	0.5997	0.9426
10000	100	297	0.2472	-0.0011	0.5077	0.5092	0.9493

Table 2: Results for simulated Example 2.

n	d	#para	AMSE	Estimate: $\mathbf{v}_{n1}^\top \overrightarrow{\widehat{\boldsymbol{\beta}}_n}$				Estimate: $\mathbf{v}_{n2}^\top \overrightarrow{\widehat{\boldsymbol{\beta}}_n}$			
				Bias	SE_emp	SE_est	CP	Bias	SE_emp	SE_est	CP
100	20	40	1.2580	0.1010	1.0868	0.9118	0.9137	0.0646	1.0786	0.9129	0.9173
500	44	88	0.1496	0.0323	0.4050	0.3813	0.9342	0.0305	0.4141	0.3906	0.9355
1000	63	126	0.0687	0.0252	0.2627	0.2517	0.9394	0.0445	0.2656	0.2538	0.9370
5000	141	282	0.0133	0.0223	0.1141	0.1133	0.9452	0.0198	0.1141	0.1134	0.9462
10000	200	400	0.0068	0.0169	0.0818	0.0817	0.9467	0.0171	0.0818	0.0821	0.9473

Lastly, we explore hypothesis testing based on the large-sample Wald test. Consider the model (5.1) with $n = 10000$; the other settings remain the same as before. For each scenario with a null hypothesis H_0, we are interested in comparing its estimated density curve and the density curve of its corresponding χ^2 distribution, and the Q-Q plots for the

Table 3: Results for simulated Example 3.

n	d	k	#para	AMSE	Bias	SE_emp	SE_est	CP
100	6	9	48	2.9184	0.0229	1.7543	1.6355	0.9233
500	9	14	117	0.6827	0.0070	0.8175	0.8085	0.9401
1000	11	16	165	0.3995	0.0020	0.6576	0.6482	0.9418
5000	16	25	384	0.1233	-0.0030	0.3675	0.3656	0.9466
10000	20	30	580	0.0744	0.0024	0.2841	0.2819	0.9479

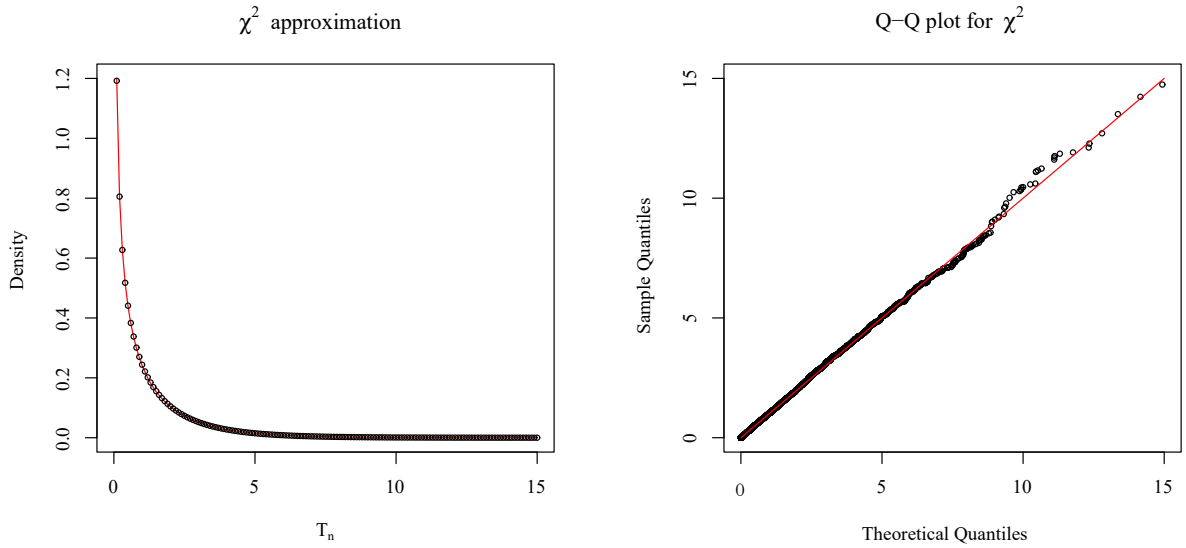

Figure 1: Asymptotic results for testing $H_0: \mathbf{a}_n^\top \overrightarrow{\boldsymbol{\beta}_n} = \mathbf{a}_n^\top \overrightarrow{\boldsymbol{\beta}_{n0}}$ in Example 1. The left panel gives the estimated null density of the large-sample Wald test (circle points) and the density of the chi-square distribution χ_1^2 under H_0 (solid line). The right panel gives the Q-Q plot for the Wald test statistic under H_0.

Wald test statistic under H_0. The related results are shown in Figures 1–3. As seen, the χ^2 approximation for the null distribution is reasonably accurate, and the theoretical quantiles are approximated very well by the sample quantiles.

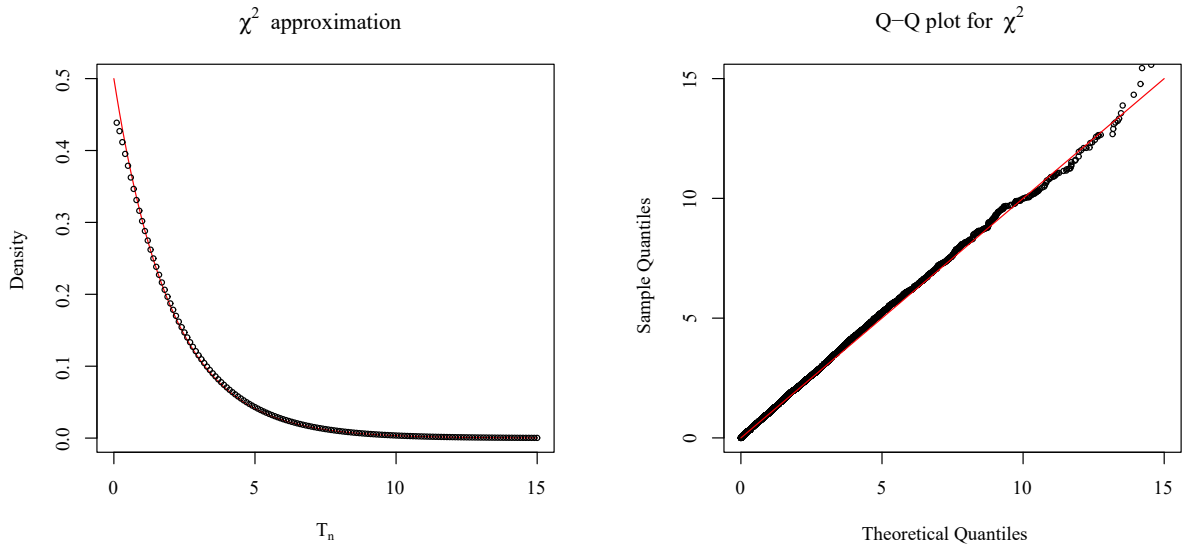

Figure 2: Asymptotic results for testing $H_0: \mathbf{V}_n^\top \overrightarrow{\boldsymbol{\beta}_n} = \mathbf{V}_n^\top \overrightarrow{\boldsymbol{\beta}_{n0}}$ in Example 2. The left panel gives the estimated null density of the large-sample Wald test (circle points) and the density of the chi-square distribution χ^2_2 under H_0 (solid line). The right panel gives the Q-Q plot for the Wald test statistic under H_0.

5.2 Application I

In this section, we apply the asymptotic results of the SMLR model to a real data set for statistical inference. We consider the 1996 American National Election Study (NES96) data set, which can be found in the CRAN package `faraway`, and contains data on 944 respondents, where each respondent consists of 10 related variables. Following Faraway (2016) and Price et al. (2019), we consider three explanatory variables, namely, education level (categorical with seven levels), income (categorical with 24 levels), and age (numerical). Each covariate is standardized to have mean zero and standard deviation one. The intercepts are considered as well. In addition, the response variable is the self-identified political affiliation of voters. As seen from Table 4, the response variable originally has seven categories (Original Categories),

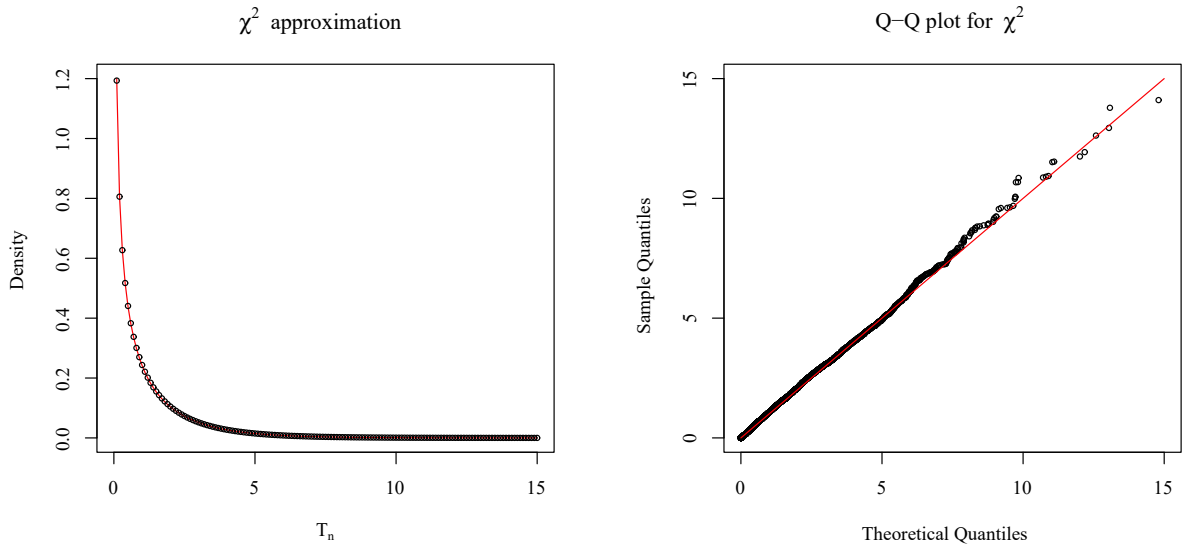

Figure 3: Asymptotic results for testing $H_0: \mathbf{b}_n^\top \overrightarrow{\boldsymbol{\beta}_n} = \mathbf{b}_n^\top \overrightarrow{\boldsymbol{\beta}_{n0}}$ in Example 3. The left panel gives the estimated null density of the large-sample Wald test (circle points) and the density of the chi-square distribution χ_1^2 under H_0 (solid line). The right panel gives the Q-Q plot for the Wald test statistic under H_0.

which is later reduced to three (Grouped Categories) by Faraway (2016). The objective here is to use the proposed SMLR model to check whether the subjective grouping of the response categories is recommended.

The fitted coefficient matrix $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_{ij}) \in \mathbb{R}^{4\times(k-1)}$ for the original and grouped categories is shown in Table 5, where the jth row vector of $\boldsymbol{\beta}$ is denoted by $\boldsymbol{\beta}_{j\bullet}$. Note that $\widehat{\boldsymbol{\beta}}_{1\bullet}$ is very close to zero under the grouped categories, which implies that the age factor may have a weak effect. However, the effect of age on the original categories seems strong. To verify this, we conduct the following hypothesis test to investigate the effects of the age factor:

$$H_0 : \boldsymbol{\beta}_{1\bullet} = \mathbf{0} \iff H_1 : \boldsymbol{\beta}_{1\bullet} \neq \mathbf{0}. \tag{5.2}$$

By Corollary 5, with $\mathbf{V} = \mathbf{I}_{k-1} \otimes (0,1,0,0)^\top$, the Wald test statistic $\widehat{\boldsymbol{\beta}}_{1\bullet}^\top [\mathbf{V}^\top \mathbf{Q}_n^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{V}]^{-1} \widehat{\boldsymbol{\beta}}_{1\bullet}$

Table 4: Summary of NES96 categories.

Original Categories (size)	Grouped Categories (size)
Strong Democrat (200)	Democrat (380)
Weak Democrat (180)	
Independent Democrat (108)	Independent (239)
Independent (37)	
Independent Republican (94)	
Weak Republican (150)	Republican (325)
Strong Republican (175)	

Table 5: Fitted coefficients of the SMLR model based on the NES96 data.

	Original Categories						Grouped Categories	
	$\widehat{\boldsymbol{\beta}}_{\bullet 1}$	$\widehat{\boldsymbol{\beta}}_{\bullet 2}$	$\widehat{\boldsymbol{\beta}}_{\bullet 3}$	$\widehat{\boldsymbol{\beta}}_{\bullet 4}$	$\widehat{\boldsymbol{\beta}}_{\bullet 5}$	$\widehat{\boldsymbol{\beta}}_{\bullet 6}$	$\widehat{\boldsymbol{\beta}}_{\bullet 1}$	$\widehat{\boldsymbol{\beta}}_{\bullet 2}$
Intercept ($\widehat{\boldsymbol{\beta}}_{0\bullet}$)	0.6304	0.1824	-0.8353	0.0667	0.5098	0.6198	0.0094	0.2519
Age ($\widehat{\boldsymbol{\beta}}_{1\bullet}$)	-0.1222	-0.0794	0.0815	0.2118	0.0728	0.1836	-0.0474	0.0103
Education Level ($\widehat{\boldsymbol{\beta}}_{2\bullet}$)	0.0391	0.1050	-0.2889	0.0010	0.0175	0.0925	-0.0365	0.0120
Income ($\widehat{\boldsymbol{\beta}}_{3\bullet}$)	-0.5525	-0.1564	0.0168	-0.1116	-0.1451	-0.0377	-0.2570	-0.2312

follows χ^2_{k-1} under H_0. For the grouped categories with $k = 3$, the Wald test statistic is 1.0572 and the p-value is 0.5894. As a result, the age factor plays no role in determining the political affiliation of voters under the grouped categories. On the other hand, the Wald test statistic is 18.3178 and the p-value is 0.0055 for the original categories, meaning that the effects of the age factor are non-negligible. These results imply that the grouping of the

response categories is not supported by the data, which is consistent with the conclusion in Price et al. (2019).

5.3 Application II

As discussed, the main advantages of the proposed SMLR over CMLR1 and CMLR2 are from computational and asymptotical perspectives, with these three MLRs proved to be equivalent in Section 3.3. Therefore, these advantages may not be evident in a finite-sample real application. On the other hand, the penalized counterparts of these three MLRs are no longer equivalent. The penalty terms of CMLR1 and CMLR2 depend on prespecified constraints to make the models identifiable, whereas the regularized SMLR solves an unconstrained optimization problem without the reference category. The simplex-based structure is expected to be more convenient and efficient in terms of statistical analysis and algorithmic design (Zhang and Liu, 2014).

In this section, we apply the three ridge-penalized MLR models to three real datasets from the UCI machine learning repository to further illustrate the usefulness of the SMLR-based models. A summary of the data sets is provided in Table 6, where $n/d/k$ denote the sample size of the training set, the dimension of the covariates and the number of categories, respectively. In addition, $n_{\min}$ and $n_{\max}$ represent the sizes of the minority and majority categories, respectively. Note that the ridge-penalized CMLR1 may have several versions, depending on the choice of the reference category, whereas the regularized CMLR2 and SMLR are unique. Given a fixed penalty factor λ, we train all three ridge-penalized MLR models for each data set, and compare their performance in terms of their label prediction accuracy.

As seen from Table 6, the selection of the reference category has a significant effect on

Table 6: Summary of real data sets and results of training ridge penalized MLRs.

Dataset	Information					Accuracy			
	n	d	k	$n_{\min}$	$n_{\max}$	λ	CMLR1	CMLR2	SMLR
Breast	106	9	6	14	22	1.5	0.7830; 0.8019; 0.7925; 0.7547; 0.7830; 0.8113	0.8019	0.8113
Segmentation	210	19	7	30	30	1.1	0.9524; 0.9524; 0.9429; 0.9571 0.9619; 0.9571; 0.9524	0.9571	0.9619
Glass	214	9	6	9	76	0.5	0.7196; 0.7196; 0.7150; 0.7243; 0.7243; 0.7196	0.7336	0.7336

the prediction accuracy, which is expected and is widely accepted in the literature (e.g., Tutz et al., 2015). This finding suggests that we need to be cautious when using the penalized CMLR1 and need to select an appropriate reference category beforehand. In practice, an extra tuning process on all possible reference categories may be needed to guarantee stable performance of the penalized CMLR1. On the other hand, the penalized CMLR2 and SMLR do not depend on the reference categories. Such MLR-based models offer symmetric and systematic insight into all categories, and become more appealing; see Friedman et al. (2010), Zahid and Tutz (2013a), Zahid and Tutz (2013b), Hastie et al. (2015), Powers et al. (2018), and de Jong et al. (2019). Between these two MLRs, the penalized SMLR is preferred for its computational convenience and the establishment of its statistical properties. Similar arguments appear in Zhang and Liu (2014). Nevertheless, substantial efforts are needed to investigate the theoretical properties of the penalized SMLR, which is left to future work. In summary, the advantages of the SMLR can be better exploited in a regularization setting,

and this study serves as a building block for further exploration of SMLR-based models.

6. Conclusion

We have proposed a novel SMLR model that has several distinct features. Compared with regular MLRs, the proposed SMLR circumvents restrictions such as reference category selection and the sum-to-zero constraint, and hence is computationally efficient. In addition, the SMLR can be treated as a unified framework that connects binomial and multinomial logistic regressions. Moreover, the asymptotic results, statistical learning theory, and properties under a general kernel of the SMLR are well established, even when the number of covariates and the number of categories increase with the sample size. Note that few studies examine these statistical properties for the MLR under a diverging number of categories. This study fills this gap because of the close relationship between the SMLR and regular MLRs. Lastly, numerical simulations and real examples show the excellent performance of the SMLR under a variety of scenarios.

One possible future research direction is to extend the SMLR to more complicated data sets such as clustered multinomial data and count data, which are ubiquitous in areas such as genomics, sports, imaging analysis, and text mining. For example, Wang (2011) studies a clustered binary LR model with a diverging number of covariates, although its extension to clustered multinomial data remains challenging. In terms of categorical count data, Zhang et al. (2017) proposed the reference-based MLR model. However, the selection of the reference category is subjective, and is not computationally efficient, in general. The proposed SMLR model could be a promising tool to deal with such multicategory data. Another potential

research topic is to adapt the SMLR to high-dimensional settings, where, as discussed in Section 5.3, the penalized SMLR is useful in such cases.

Supplementary Material

The online Supplementary Material provides the following: (i) preliminary lemmas used to establish the theoretical results in the manuscript, and (ii) technical proofs for all propositions and theorems presented in the manuscript.

Acknowledgements

We are grateful to the editors and three referees for their helpful comments and suggestions. Ye was supported by the National Science Foundation of China (72071138) and Singapore MOE AcRF Tier 2 grant (R-266-000-143-112).

References

Albert, A. and J. A. Anderson (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika.* **71**, 1–10.

Anderson, J. and V. Blair (1982). Penalized maximum likelihood estimation in logistic regression and discrimination. *Biometrika.* **69**, 123–136.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika.* **59**, 19–35.

Baker, S. G. (1994). The multinomial-poisson transformation. *J. Roy. Stat. Soc. Series D.* **43**, 495–504.

Bartlett, P. L., M. I. Jordan, and J. D. McAuliffe (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.* **101**, 138–156.

Berger, A. L., V. J. D. Pietra, and S. A. D. Pietra (1996). A maximum entropy approach to natural language processing. *Comput. Linguist.* **22**, 39–71.

Böhning, D. (1992). Multinomial logistic regression algorithm. *Ann. Inst. Stat. Math.* **44**, 197–200.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. Cambridge University Press.

Bühlmann, P. and S. Van De Geer (2011). *Statistics for High-dimensional Data: methods, theory and applications*. Springer.

Cramer, J. S. (2003). *Logit Models from Economics and Other Fields*. Cambridge University Press.

de Jong, V. M., M. J. Eijkemans, B. van Calster, D. Timmerman, K. G. Moons, E. W. Steyerberg, and M. van Smeden (2019). Sample size considerations and predictive performance of multinomial logistic prediction models. *Stat. Med.* **38**, 1601–1619.

Dekel, O. and O. Shamir (2010). Multiclass-multilabel classification with more classes than examples. In *International Conference on Artificial Intelligence and Statistics*, pp. 137–144.

Deng, J., A. C. Berg, K. Li, and F.-F. Li (2010). What does classifying more than 10,000 image categories tell us? In *European Conference on Computer Vision*, pp. 71–84. Springer.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li (2009). Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255.

Fahrmeir, L. and H. Kaufmann (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Stat.* **13**, 342–368.

Fang, J. and G. Y. Yi (2021). Matrix-variate logistic regression with measurement error. *Biometrika.* **108**, 83–97.

Faraway, J. J. (2016). *Extending the Linear Model with R: generalized linear, mixed effects and nonparametric regression models* (2nd ed.). Chapman and Hall/CRC.

Forbes, C., M. Evans, N. Hastings, and B. Peacock (2011). *Statistical Distributions* (4th ed.). John Wiley & Sons.

Friedman, J., T. Hastie, and R. Tibshirani (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Stat.* **28**, 337–407.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22.

Fu, S., S. Zhang, and Y. Liu (2018). Adaptively weighted large-margin angle-based classifiers. *J. Multivar. Anal.* **166**, 282–299.

Gao, Q., X. Du, X. Zhou, and F. Xie (2018). Asymptotic properties of maximum quasi-likelihood estimators in generalized linear models with diverging number of covariates. *J. Syst. Sci. Complex.* **31**, 1362–1376.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. MIT Press.

Hastie, T., R. Tibshirani, and J. H. Friedman (2009). *The Elements of Statistical Learning: data mining, inference and prediction* (2nd ed.). Springer.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical Learning with Sparsity: the lasso and generalizations*. CRC Press.

He, X. and Q.-M. Shao (2000). On parameters of increasing dimensions. *J. Multivar. Anal.* **73**, 120–135.

Hill, S. I. and A. Doucet (2007). A framework for kernel-based multi-category classification. *J. Artif. Intell. Res.* **30**, 525–564.

Hosmer, D. W., S. Lemeshow, and R. X. Sturdivant (2013). *Applied Logistic Regression* (3rd ed.). John Wiley & Sons.

Krishnapuram, B., L. Carin, M. A. Figueiredo, and A. J. Hartemink (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Trans. Pattern. Anal. Mach. Intell.* **27**, 957–968.

Lang, J. B. (1996). On the comparison of multinomial and poisson log-linear models. *J. Roy. Stat. Soc. Series B.* **58**, 253–266.

Lange, K. and T. T. Wu (2008). An MM algorithm for multicategory vertex discriminant analysis. *J. Comput. Graph. Stat.* **17**, 527–544.

Lemeshow, S. and D. W. Hosmer (2014). Logistic regression in practice. In *Wiley StatsRef: Statistics Reference Online*, pp. 1–15. Wiley Online Library.

Liang, H. and P. Du (2012). Maximum likelihood estimation in logistic regression models with a diverging number of covariates. *Electronic J. of Statistics.* **6**, 1838–1846.

Liu, T.-Y., Y. Yang, H. Wan, Q. Zhou, B. Gao, H.-J. Zeng, Z. Chen, and W.-Y. Ma (2005). An experimental study on large-scale web categorization. In *International Conference on World Wide Web*, pp. 1106–1107. Association for Computing Machinery.

Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation. In *Conference on Natural Language Learning*, volume 20, pp. 1–7. Association for Computational Linguistics.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models.* CRC Press.

Mo, W. and Y. Liu (2021). Supervised learning. In *Wiley StatsRef: Statistics Reference Online*, pp. 1–20. Wiley Online Library.

Mroueh, Y., T. Poggio, L. Rosasco, and J.-J. Slotine (2012). Multiclass learning with simplex coding. In *Adv. Neural Inf. Process. Syst.*, pp. 2789–2797.

Ng, A., J. Ngiam, C. Y. Foo, Y. Mai, and C. Suen (2013). *Softmax Regression.* `http://ufldl.stanford.edu/wiki/index.php/Softmax_Regression`, UFLDL Tutorial.

Nilsback, M.-E. and A. Zisserman (2008). Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729.

Portnoy, S. (1985). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large; ii. normal approximation. *Ann. Stat.* **13**, 1403–1417.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Stat.* **16**, 356–366.

Powers, S., T. Hastie, and R. Tibshirani (2018). Nuclear penalized multinomial regression with an application to predicting at bat outcomes in baseball. *Stat. Model.* **18**, 388–410.

Price, B. S., C. J. Geyer, and A. J. Rothman (2019). Automatic response category combination in multinomial logistic regression. *J. Comput. Graph. Stat.* **28**, 758–766.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Tewari, A. and P. L. Bartlett (2007). On the consistency of multiclass classification methods. *J. Mach. Learn. Res.* **8**, 1007–1025.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Stat.* **32**, 135–166.

Tutz, G. (2011). *Regression for Categorical Data*. Cambridge University Press.

Tutz, G., W. Pößnecker, and L. Uhlmann (2015). Variable selection in general multinomial logit models. *Comput. Stat. Data Anal.* **82**, 207–222.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Wang, L. (2011). GEE analysis of clustered binary data with diverging number of covariates. *Ann. Stat.* **39**, 389–417.

Yee, T. W. (2015). *Vector Generalized Linear and Additive Models: with an implementation in R*. Springer.

Zahid, F. M. and G. Tutz (2013a). Multinomial logit models with implicit variable selection. *Adv. Data Anal. Classif.* **7**, 393–416.

Zahid, F. M. and G. Tutz (2013b). Ridge estimation for multinomial logit models with symmetric side constraints.

Comput. Stat. **28**, 1017–1034.

Zhang, C. and Y. Liu (2014). Multicategory angle-based large-margin classification. *Biometrika.* **101**, 625–640.

Zhang, C., M. Pham, S. Fu, and Y. Liu (2018). Robust multicategory support vector machines using difference convex algorithm. *Math. Program.* **169**, 277–305.

Zhang, T. (2004). Statistical analysis of some multi-category large margin classification methods. *J. Mach. Learn. Res.* **5**, 1225–1251.

Zhang, Y., H. Zhou, J. Zhou, and W. Sun (2017). Regression models for multivariate count data. *J. Comput. Graph. Stat.* **26**, 1–13.

Zhu, J. and T. Hastie (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics.* **5**, 427–443.

Zhu, J. and T. Hastie (2005). Kernel logistic regression and the import vector machine. *J. Comput. Graph. Stat.* **14**, 185–205.

Sheng Fu

Department of Industrial Systems Engineering & Management, National University of Singapore

E-mail: fusheng1007@gmail.com

Piao Chen

Delft Institute of Applied Mathematics, Delft University of Technology, The Netherlands

E-mail: p.chen-6@tudelft.nl

Yufeng Liu

Department of Statistics and Operations Research, University of North Carolina at Chapel Hill

E-mail: yfliu@email.unc.edu

Zhisheng Ye

Department of Industrial Systems Engineering & Management, National University of Singapore

E-mail: yez@nus.edu.sg