LMU
LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Faisal Maqbool Zahid & Gerhard Tutz

Ridge Estimation for Multinomial Logit Models with Symmetric Side Constraints

Technical Report Number 067, 2009
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Ridge Estimation for Multinomial Logit Models with Symmetric Side Constraints

Faisal Maqbool Zahid[a,*], Gerhard Tutz[b]

[a]Ludwig-Maximilians-University Munich, Ludwigstrasse 33, D-80539 Munich, Germany

[b]Ludwig-Maximilians-University Munich, Akademiestrasse 1, 80799 Munich, Germany

Abstract

In multinomial logit models, the identifiability of parameter estimates is typically obtained by side constraints that specify one of the response categories as reference category. When parameters are penalized, shrinkage of estimates should not depend on the reference category. In this paper we investigate ridge regression for the multinomial logit model with symmetric side constraints, which yields parameter estimates that are independent of the reference category. In simulation studies the results are compared with the usual maximum likelihood estimates and an application to real data is given.

Key words: logistic regression, penalization, side constraints, ridge regression, cross-validation, multinomial logit

1. Introduction

The multinomial logit model is the most widely used model in multi-categorical regression. It specifies the conditional probabilities of response categories through linear functions of covariate vector **x**. When the number of predictors is large as compared to the number of observations, the logit model suffers from problems such as complete separation, the estimates of parameters are not uniquely defined (some are infinite) and/or the maximum of log-likelihood is achieved at 0. The use of regularization methods can help

[*]Corresponding author. Tel.: ++49 89 2180 6408; fax.: ++49 89 2180 5040.

Email addresses: `faisal-maqbool.zahid@stat.uni-muenchen.de` (Faisal Maqbool Zahid), `tutz@stat.uni-muenchen.de` (Gerhard Tutz)

to overcome such problems.
Regularization methods based on penalization typically maximize a penalized log-likelihood. Ridge regression, one of the oldest penalization methods for linear models, was extended to GLM type models by Nyquist (1991), although a definition of a ridge estimator for the logistic regression model, which is a particular case of generalized linear models was suggested by Schaefer et al. (1984) and Schaefer (1986). Segerstedt (1992) discussed a generalization of ridge regression for ML estimation in GLM. Many alternative penalization/shrinkage methods were proposed for univariate GLMs, among them the Lasso (Tibshirani (1996)), which was adapted to GLMs by Park and Hastie (2007), the Dantzig selector (James and Radchenko (2009)), SCAD (Fan and Li (2001)) and boosting approaches (Bühlmann and Hothorn (2007), Tutz and Binder (2006)). However, few approaches have been proposed for multicategory responses. Krishnapuram et al. (2005) consider multinomial logistic regression with lasso type estimates, Zhu and Hastie (2004) use ridge type penalization and Friedman et al. (2008) use the penalties L1 (the lasso), L2(ridge regression) and mixture of the two (the elastic net).

In this paper we are defining the ridge regression (L2 penalty) for multicategory logit models with symmetric constraints. Zhu and Hastie (2004) used this symmetric constraint while using penalized logistic regression as an alternative to the SVM (support vector machine) for microarray cancer diagnostic problems. Friedman et al. (2008) also used the symmetric multinomial logit model for defining paths for generalized linear models using cyclical coordinate descent algorithm. In contrast to Zhu and Hastie (2004) and Friedman et al. (2008), our approach is based on Fisher scoring that uses a transformed version of the design matrix and a matrix other than the identity matrix in the ridge penalty.
In section 2 side constraints, interpretation of the parameters with symmetric side constraint, and the penalized model with L2-penalty is described. Section 3 compares the ridge estimates based on symmetric side constraint with the usual MLE in terms of MSE of $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\beta}}$ in a simulation study. Multinomial logit model with symmetric constraint is implemented on the real data in section 4. Section 5 concludes with some concluding remarks.

2. Side Constraints and Regularization

The multinomial logit model is one of most oftenly used regression models when a categorical response variable has more than two (unordered) categories. Let the response variable $Y \in \{1, \ldots, k\}$ have k possible values (categories). A generic form of the multinomial logit model is given by

$$P(Y=r|\mathbf{x}) = \frac{exp(\mathbf{x}^T\boldsymbol{\beta}_r)}{\sum_{s=1}^{k} exp(\mathbf{x}^T\boldsymbol{\beta}_s)} = \frac{exp(\eta_r)}{\sum_{s=1}^{k} exp(\eta_s)}, \tag{1}$$

where $\boldsymbol{\beta}_r^T = (\beta_{r0}, \ldots, \beta_{rp})$. It is obvious that one has to specify some additional constraints since the parameters $\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_k^T$ are not identifiable. An often used side constraint is based on choosing a reference category (RSC). When category k is chosen, one sets

$$\boldsymbol{\beta}_k^T = (0, \ldots, 0) \qquad \text{yielding} \qquad \eta_k = 0.$$

Of course any of the response categories can be chosen as reference. When category s is chosen one sets $\boldsymbol{\beta}_s^T = (0, \ldots, 0)$ yielding $\eta_s = 0$. Throughout the paper we will use reference category k when a model with a reference category is fitted. The corresponding model is

$$P(Y=r|\mathbf{x}) = \frac{exp(\mathbf{x}^T\boldsymbol{\beta}_r)}{1+\sum_{s=1}^{q} exp(\mathbf{x}^T\boldsymbol{\beta}_s)} \qquad \text{for } r = 1, \ldots, q. \tag{2}$$

An alternative side constraint that is more appropriate when defining regularization terms is the symmetric side constraint (SSC) given by

$$\sum_{s=1}^{k} \boldsymbol{\beta}_s^* = \mathbf{0}. \tag{3}$$

With $\boldsymbol{\beta}_r^*$ denoting the corresponding parameters, the multinomial logit model is

$$P(Y=r|\mathbf{x}) = \frac{exp(\mathbf{x}^T\boldsymbol{\beta}_r^*)}{\sum_{s=1}^{k} exp(\mathbf{x}^T\boldsymbol{\beta}_s^*)} = \frac{exp(\eta_r^*)}{\sum_{s=1}^{k} exp(\eta_s^*)} \qquad \text{for } r = 1, \ldots, q \tag{4}$$

Although the models are equivalent parameters for symmetric side constraint are different from parameters with a reference category and consequently have different interpretation. In the case of SSC, i.e., $\sum_{s=1}^{k}\boldsymbol{\beta}_s^* = \mathbf{0}$, the ”median” response can be viewed as the reference category, and is defined by the geometric mean. Then one obtains from (4)

$$\frac{P(Y=r|\mathbf{x})}{GM(\mathbf{x})} = \frac{exp(\boldsymbol{\eta}_r^*)}{\sqrt[k]{\prod_{s=1}^{k} P(Y=s|\mathbf{x})}}$$

and

$$\log\left(\frac{P(Y=r|\mathbf{x})}{GM(\mathbf{x})}\right) = \mathbf{x}^T\boldsymbol{\beta}_r^*.$$

Therefore $\boldsymbol{\beta}_r^*$ reflects the effects of $\mathbf{x}$ on the logits when $P(Y = r|\mathbf{x})$ is compared to the median response $GM(\mathbf{x})$.

It should be noted that whatever side constraint is used, the log-odds between two response probabilities and the corresponding weights are easily computed by

$$\log\left[\frac{P(Y = r|\mathbf{x})}{P(Y = s|\mathbf{x})}\right] = \mathbf{x}^T(\boldsymbol{\beta}_r^* - \boldsymbol{\beta}_s^*),$$

which follows from (2) and (4) for any choice of response categories $r, s \in \{1, \dots, k\}$. Let in the following $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_q^T)$ and $\boldsymbol{\beta}^{*T} = (\boldsymbol{\beta}_1^{*T}, \dots, \boldsymbol{\beta}_q^{*T})$ denote the parameter vectors for the multinomial logit model under the two situations i.e., reference category side constraint ($\boldsymbol{\beta}_k = \mathbf{0}$) and symmetric side constraint ($\sum_{s=1}^k \boldsymbol{\beta}_s^* = \mathbf{0}$). For illustration we consider the case of a response variable with three categories. With a model which contains only the intercept, logits are given as

$$\log\left(\frac{\pi_1}{\pi_3}\right) = \beta_{10}, \quad \log\left(\frac{\pi_2}{\pi_3}\right) = \beta_{20}$$

with side constraint $\beta_{30} = 0$, and

$$\log\left(\frac{\pi_1^*}{\pi_3^*}\right) = \beta_{10}^* - \beta_{30}^* = 2\beta_{10}^* + \beta_{20}^*, \qquad \log\left(\frac{\pi_2^*}{\pi_3^*}\right) = \beta_{20}^* - \beta_{30}^* = \beta_{10}^* + 2\beta_{20}^*$$

with symmetric side constraint $\sum_{s=1}^3 \beta_{s0}^* = \mathbf{0}$. Equating the corresponding logits in both situations, one obtains

$$\boldsymbol{\beta}^* = \mathbf{T}\boldsymbol{\beta}\ , \qquad \boldsymbol{\beta} = \mathbf{T}^{-1}\boldsymbol{\beta}^*, \tag{5}$$

where $\boldsymbol{\beta}^{*T} = (\beta_{10}^* \quad \beta_{20}^*)$, $\boldsymbol{\beta}^T = (\beta_{10} \quad \beta_{20})$, and

$$\mathbf{T} = \begin{bmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{bmatrix}, \qquad \mathbf{T}^{-1} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

For a model with an intercept and p covariates, logits are given by

$$\begin{aligned} \log\left(\tfrac{\pi_r}{\pi_3}\right) &= \mathbf{x}^T\boldsymbol{\beta}_r \qquad r = 1, 2, \\ \log\left(\tfrac{\pi_r^*}{\pi_3^*}\right) &= \mathbf{x}^T\boldsymbol{\beta}_r^* \qquad r = 1, 2. \end{aligned}$$

Equating the logits for these two cases, we get $2(p + 1)$ equations which can easily be solved to get the result

$$\mathbf{B}^* = (\mathbf{T}\mathbf{B}^T)^T \quad \text{or} \quad \mathbf{B} = (\mathbf{T}^{-1}\mathbf{B}^{*T})^T,$$

where $\mathbf{T}$ and $\mathbf{T}^{-1}$ are the 2×2 matrices from above and $\mathbf{B} = (\boldsymbol{\beta}_1 \quad \boldsymbol{\beta}_2)$, and $\mathbf{B}^* = (\boldsymbol{\beta}_1^* \quad \boldsymbol{\beta}_2^*)$ are $(p+1) \times 2$ matrices composed of parameter vectors with RSC and SSC respectively.

In the general case let $\boldsymbol{\beta}_{.j}^T = (\beta_{1j}, \ldots, \beta_{k-1,j})$, $\boldsymbol{\beta}_{.j}^{*T} = (\beta_{1j}^*, \ldots, \beta_{k-1,j}^*)$, $j = 0, \ldots, p$, collect parameter vectors for single variables with reference category k or symmetric side constraints respectively. Then one obtains the transformation

$$\boldsymbol{\beta}_{.j}^* = \mathbf{T}\boldsymbol{\beta}_{.j} \qquad \text{for } j = 0, 1, \ldots, p \tag{6}$$

given as

$$\begin{bmatrix} \beta_{1j}^* \\ \beta_{2j}^* \\ \vdots \\ \beta_{k-2,j}^* \\ \beta_{k-1,j}^* \end{bmatrix} = \begin{bmatrix} \frac{k-1}{k} & -\frac{1}{k} & \cdots & -\frac{1}{k} & -\frac{1}{k} \\ -\frac{1}{k} & \frac{k-1}{k} & \cdots & -\frac{1}{k} & -\frac{1}{k} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{1}{k} & -\frac{1}{k} & \cdots & \frac{k-1}{k} & -\frac{1}{k} \\ -\frac{1}{k} & -\frac{1}{k} & \cdots & -\frac{1}{k} & \frac{k-1}{k} \end{bmatrix} \begin{bmatrix} \beta_{1j} \\ \beta_{2j} \\ \vdots \\ \beta_{k-2,j} \\ \beta_{k-1,j} \end{bmatrix}$$

with the inverse transformation

$$\begin{bmatrix} \beta_{1j} \\ \beta_{2j} \\ \vdots \\ \beta_{k-2,j} \\ \beta_{k-1,j} \end{bmatrix} = \begin{bmatrix} 2 & 1 & \cdots & 1 & 1 \\ 1 & 2 & \cdots & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \cdots & 2 & 1 \\ 1 & 1 & \cdots & 1 & 2 \end{bmatrix} \begin{bmatrix} \beta_{1j}^* \\ \beta_{2j}^* \\ \vdots \\ \beta_{k-2,j}^* \\ \beta_{k-1,j}^* \end{bmatrix} \tag{7}$$

i.e., $\boldsymbol{\beta}_{.j} = \mathbf{T}^{-1}\boldsymbol{\beta}_{.j}^*$ (for $j = 0, 1, \ldots, p$), where $\mathbf{T}^{-1}$ is a $(q \times q)$-matrix with diagonal entries 2 and off-diagonal elements 1. The same transformation holds for ML estimates. Estimates of the parameters with symmetric side constraint can be computed by transforming (reparameterizing) estimates with reference category side contraint and vice versa.

With $\boldsymbol{\pi}_i^T = (\pi_{i1}, \ldots, \pi_{iq})$ $(q = k-1)$ denoting the $(q \times 1)$-vector of probabilities with $\pi_{ir} = P(Y = r|\mathbf{x}_i)$, the multinomial logit model has the form

$$\boldsymbol{\pi}_i = h(\mathbf{X}_i\boldsymbol{\beta}) = h(\boldsymbol{\eta}_i), \tag{8}$$

where h is a vector-valued response function, $\mathbf{X}_i$ is a $(q \times (p+1))$-design matrix composed of $\mathbf{x}_i$ (with first term 1 for the intercept) and given as

$$\mathbf{X}_i = \begin{bmatrix} \mathbf{x}_i^T & & & \\ & \mathbf{x}_i^T & & \\ & & \ddots & \\ & & & \mathbf{x}_i^T \end{bmatrix}$$

and $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_q^T)$ is the vector of unknown parameters of length $(q \times (p+1))$. The multinomial logit model is given by

$$\pi_{ir} = \frac{exp(\mathbf{x}_i^T \boldsymbol{\beta}_r)}{1 + \sum_{s=1}^{q} exp(\mathbf{x}_i^T \boldsymbol{\beta}_s)} \qquad r = 1, \dots, q$$

which for side constraint with reference category k yields

$$\log\left[\frac{P(Y = r|\mathbf{x})}{P(Y = k|\mathbf{x})}\right] = \mathbf{x}^T \boldsymbol{\beta}_r, \qquad r = 1, \dots, q \tag{9}$$

The log-odds compare $\pi_r = P(Y = r|\mathbf{x})$ to the probability $\pi_k = P(Y = k|\mathbf{x})$ of the reference category k. The q logits $\log(P(Y = 1|\mathbf{x})/P(Y = k|\mathbf{x})), \dots, \log(P(Y = q|\mathbf{x})/P(Y = k|\mathbf{x}))$ given by (9) determine the response probabilities $P(Y = 1|\mathbf{x}), \dots, P(Y = k|\mathbf{x})$ uniquely since the constraint $\sum_{r=1}^{k} P(Y = r|\mathbf{x}) = 1$. holds. Therefore only $q = k - 1$ response categories and parameter vectors have to be specified. The representation of the multinomial logit model in (8) and the corresponding response function h depend distinctly on the choice of the reference category. Since the parameters $\boldsymbol{\beta}^*$ with SSC may be obtained by reparameterization of the parameters $\boldsymbol{\beta}$ with RSC, the numerical computation of maximum likelihood estimates of $\boldsymbol{\beta}^*$ makes use of a transformation of the design matrix $\mathbf{X}$. The transformed design matrix for SSC has the form

$$\mathbf{X}^* = \mathbf{X}\mathbf{T}^*,$$

where $\mathbf{X}$ is the total design matrix of order $q(n \times (p+1))$ given as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix}$$

with $\mathbf{X}_i$, a $q \times q(p+1)$ matrix (composed of $\mathbf{x}_i$) as defined earlier. $\mathbf{T}^*$ is a $q((p+1) \times (p+1))$ matrix composed of the elements of $\mathbf{T}^{-1}$ in order to satisfy $\boldsymbol{\beta}_{.j} = \mathbf{T}^{-1}\boldsymbol{\beta}^*_{.j}$ (for $j = 0, 1, \dots, p$). For example, with $k = 3$ and $p = 2$, $\mathbf{T}^*$, one obtains

$$\mathbf{T}^* = \mathbf{T}^{-1}_{q \times q} \otimes \mathbf{I}_{(p+1)\times(p+1)} = \begin{bmatrix} 2 & 0 & 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & 0 & 0 & 1 \\ 1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 2 & 0 \\ 0 & 0 & 1 & 0 & 0 & 2 \end{bmatrix}$$

where $\otimes$ is the Kronecker matrix product. The corresponding score function

$$s(\boldsymbol{\beta}^*) = \frac{\partial l(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^*} = \sum_{i=1}^{n} s_i(\boldsymbol{\beta}^*),$$

has components

$$s_i(\boldsymbol{\beta}^*) = \mathbf{X}_i^{*T}\mathbf{D}_i(\boldsymbol{\beta}^*)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta}^*)[\mathbf{y}_i - h(\boldsymbol{\eta}_i^*)],$$

where $\mathbf{D}_i(\boldsymbol{\beta}^*) = \frac{\partial h(\boldsymbol{\eta}_i^*)}{\partial \boldsymbol{\eta}^*}$ is derivative of $h(\boldsymbol{\eta}^*)$ evaluated at $\boldsymbol{\eta}_i^* = \mathbf{X}_i^*\boldsymbol{\beta}^*$ and $\boldsymbol{\Sigma}(\boldsymbol{\beta}^*) = cov(\mathbf{y}_i)$ is the covariance matrix of ith observation of $\mathbf{y}$ given parameter vector $\boldsymbol{\beta}^*$. In matrix notation one has

$$s(\boldsymbol{\beta}^*) = \mathbf{X}^{*T}\mathbf{D}(\boldsymbol{\beta}^*)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)[\mathbf{y} - h(\boldsymbol{\eta}^*)],$$

where $\mathbf{y}$ and $h(\boldsymbol{\eta}^*)$ are given by

$$\mathbf{y}^T = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T), \quad h(\boldsymbol{\eta}^*)^T = (h(\boldsymbol{\eta}_1^*)^T, \ldots, h(\boldsymbol{\eta}_n^*)^T).$$

The matrices have block diagonal form

$$\boldsymbol{\Sigma}(\boldsymbol{\beta}^*) = \text{diag}(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta}^*)), \quad \mathbf{W}(\boldsymbol{\beta}^*) = \text{diag}(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta}^*)), \quad \mathbf{D}(\boldsymbol{\beta}^*) = \text{diag}(\mathbf{D}_i(\boldsymbol{\beta}^*)).$$

Then Fisher scoring iteration, which can also be viewed as an iteratively reweighted least square procedure, has the form

$$\hat{\boldsymbol{\beta}}^{*(k+1)} = \hat{\boldsymbol{\beta}}^{*(k)} + \left(\mathbf{X}^{*T}\ \mathbf{W}(\hat{\boldsymbol{\beta}}^{*(k)})\ \mathbf{X}^*\right)^{-1} s(\hat{\boldsymbol{\beta}}^{*(k)}).$$

2.1. Regularization

Regularization methods using penalization are based on penalized log-likelihood

$$l_p(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}) - \frac{\lambda}{2}J(\boldsymbol{\beta}),$$

where $l_i(\boldsymbol{\beta})$ is the usual log-likelihood contribution of the ith observation, λ is a tuning parameter and $J(\boldsymbol{\beta})$ is a functional which penalizes the size of parameter. In high dimensional problems, which may also cause the non-existence of maximum-likelihood estimators, the use of regularization methods is advantageous because penalized estimators will exist and have better prediction error than the usual ML estimator. Ridge penalty, introduced by Hoerl & Kennard (1970) for linear models and then extended to generalized linear models by Nyquist (1991), is one of the oldest penalization methods. It uses the penalty $J(\boldsymbol{\beta}) = \sum_{i=1}^{p} \beta_i^2$, yielding for binary responses the penalized log-likelihood

$$l_p(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\beta}) - \frac{\lambda}{2} \sum_{i=1}^{p} \beta_i^2.$$

For multi-categorical response model, instead of one parameter vector one has the collection of parameter vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$, which are identifiable only under some side constraint. A straightforward extension of the binary case is the penalty

$$J(\boldsymbol{\beta}) = \sum_{r=1}^{q} \sum_{j=1}^{p} \beta_{rj}^2 = \sum_{j=1}^{p} \boldsymbol{\beta}_{.j}^T \boldsymbol{\beta}_{.j},$$

where $\boldsymbol{\beta}_{.j}^T = (\beta_{1j}, \ldots, \beta_{k-1,j})$ and $\beta_{kj} = 0$, which specifies k as reference category. However, if a different reference category is chosen the corresponding ridge estimator would yield different estimates, even after transformation.

A more natural choice for defining the multi-category ridge estimator is the use of symmetrically constrained parameters. Therefore we will use the definition

$$J(\boldsymbol{\beta}^*) = \sum_{r=1}^{k} \sum_{j=1}^{p} \beta_{rj}^{*2} \tag{10}$$

with $\sum_{r=1}^{k} \beta_{rj}^* = 0$. It can also be written as

$$J(\boldsymbol{\beta}^*) = \sum_{j=1}^{p} \boldsymbol{\beta}_{.j}^{*T} \mathbf{P} \boldsymbol{\beta}_{.j}^* \tag{11}$$

where $\boldsymbol{\beta}_{.j}^{*T} = (\beta_{1j}^*, \ldots, \beta_{k-1,j}^*)$ and $\mathbf{P} = \mathbf{T}^{-1}$. Transformation to parameters with side constraint $\boldsymbol{\beta}_k = \mathbf{0}$ yields

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \boldsymbol{\beta}_{.j}^T \mathbf{T}^T \mathbf{P} \mathbf{T} \boldsymbol{\beta}_{.j}. \tag{12}$$

The use of matrix $\mathbf{T}^T\mathbf{PT}$ instead of the identity matrix $\mathbf{I}$, will cause $J(\boldsymbol{\beta})$ to penalize the size of parameters for all k categories while working with the q logits under the constraint given in (3). For the complete design one obtains

$$J(\boldsymbol{\beta}^*) = \boldsymbol{\beta}^{*T} \mathbf{P}^* \boldsymbol{\beta}^*,$$

where $\boldsymbol{\beta}^*$ has length $q(p+1)$ and matrix $\mathbf{P}^*$ differs from matrix $\mathbf{T}^*$ only by having the zero rows corresponding to the intercepts $\beta_{.0}$ (i.e., each of $[r(p+1)+1]$th row is zero for $r = 0, 1, \ldots, k-2$), since intercept terms are not penalized.

A general form of the penalty term for multi-categorical responses has the additive form

$$\lambda J(\boldsymbol{\beta}) = \lambda \sum_{r=1}^{q} \sum_{j=1}^{p} \left|\boldsymbol{\beta}_{rj}\right|^{\gamma}, \qquad \gamma > 0$$

Multi-categorical ridge and lasso are special cases with $\gamma = 2$ and $\gamma = 1$ respectively. Since shrinkage should not depend on the reference category, the penalties should use the symmetric constraints which transform to different functions when reference categories are used.

If we consider multinomial logit model with SSC described in (4) and the penalty term given in (11), then the penalized log-likelihood is given by

$$\begin{aligned} l_p(\boldsymbol{\beta}^*) &= \sum_{i=1}^{n} l_i(\boldsymbol{\beta}^*) - \frac{\lambda}{2} J(\boldsymbol{\beta}^*) \\ &= \sum_{i=1}^{n} l_i(\boldsymbol{\beta}^*) - \frac{\lambda}{2} \sum_{j=1}^{p} \boldsymbol{\beta}_{.j}^{*T} \mathbf{P} \, \boldsymbol{\beta}_{.j}^{*} \end{aligned}$$

The corresponding penalized score function $s_p(\boldsymbol{\beta}^*)$ is given by

$$\begin{aligned} s_p(\boldsymbol{\beta}^*) &= \sum_{i=1}^{n} \mathbf{X}_i^{*T} \mathbf{D}_i(\boldsymbol{\beta}^*) \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta}^*)[\mathbf{y}_i - h(\boldsymbol{\eta}_i^*)] - \lambda \mathbf{P}^* \boldsymbol{\beta}^* \\ &= \mathbf{X}^{*T} \mathbf{D}(\boldsymbol{\beta}^*) \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)[\mathbf{y} - h(\boldsymbol{\eta}^*)] - \lambda \mathbf{P}^* \boldsymbol{\beta}^* \end{aligned}$$

yielding the estimation equations

$$\mathbf{X}^{*T} \mathbf{D}(\boldsymbol{\beta}^*) \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}^*)[\mathbf{y} - h(\boldsymbol{\eta}^*)] - \lambda \mathbf{P}^* \boldsymbol{\beta}^* = \mathbf{0}$$

where $\boldsymbol{\beta}^*$ is a vector of parameters of length $q \times (p+1)$, and $\mathbf{P}^*$ is a $q \times ((p+1) \times (p+1))$ diagonal matrix whose elements are the q times repetition of diagonal of $\mathbf{P}$. Fisher scoring iteration provides

$$\hat{\boldsymbol{\beta}}^{*(k+1)} = \hat{\boldsymbol{\beta}}^{*(k)} + \left(\mathbf{X}^{*T} \; \mathbf{W}(\hat{\boldsymbol{\beta}}^{*(k)}) \; \mathbf{X}^* + \lambda \mathbf{P}^*\right)^{-1} s_p(\hat{\boldsymbol{\beta}}^{*(k)}).$$

At convergence, if $\boldsymbol{\beta}^*$ are the estimates (penalized) of true parameter $\boldsymbol{\beta}$, then for the covariance matrix one obtains

$$cov(\hat{\boldsymbol{\beta}}^*) \;=\; \left(\mathbf{X}^{*T}\,\mathbf{W}(\hat{\boldsymbol{\beta}}^*)\,\mathbf{X}^* + \lambda\mathbf{P}^*\right)^{-1}\left(\mathbf{X}^{*T}\,\mathbf{W}(\hat{\boldsymbol{\beta}}^*)\,\mathbf{X}^*\right)\left(\mathbf{X}^{*T}\,\mathbf{W}(\hat{\boldsymbol{\beta}}^*)\,\mathbf{X}^* + \lambda\mathbf{P}^*\right)^{-1}$$

and the hat matrix

$$\mathbf{H}^* = \mathbf{W}^{*T/2}\mathbf{X}^*\left(\mathbf{X}^{*T}\,\mathbf{W}(\hat{\boldsymbol{\beta}}^*)\,\mathbf{X}^* + \lambda\mathbf{P}^*\right)^{-1}\mathbf{X}^{*T}\mathbf{W}^{*1/2}$$

which we need in section 3 while deciding about the optimum value of the tuning parameter λ on the basis of generalized cross-validation.

3. Simulation Study

In a simulation study the results of penalization using the ridge penalty with symmetric constraint were compared with its counterpart i.e., penalization with a reference category and usual MLE. In this study, for multinomial logit models with three response categories different number of continuous (independent and correlated) and categorical covariates were considered for different sample sizes ($n = 30, 50, 70$ and 100). The situations with different number and type of covariates used for multinomial logit models in the simulation study were:
Ik3: Independent covariates drawn from standard normal distribution,
Mk3: Covariates with moderate correlation of magnitude 0.3 between covariates,
Hk3: Covariates with high correlation of magnitude 0.9 between covariates.
The parameter values used for the vector $\boldsymbol{\beta}$ of length $q(p+1)$ for situations Ik3, Mk3 and Hk3 were:

$$\begin{aligned} p = 5: &\quad \boldsymbol{\beta}^T = (1, 5/6, \ldots, 1/6, 1/6, 2/6, \ldots, 1),\\ p = 10: &\quad \boldsymbol{\beta}^T = (1, 10/11, \ldots, 1/11, 1/11, 2/11, \ldots, 1),\\ p = 20: &\quad \boldsymbol{\beta}^T = (1, 19/20, \ldots, 1/20, 1/20, 2/20, \ldots, 1), \end{aligned}$$

In addition, simulations with categorical covariates were performed:
ICk3: 10 independent standard normal covariates, one categorical covariate with three categories, two binay and one covariate with four categories,
MCk3: 10 correlated covariates with correlation 0.3, one categorical covariate with three categories, two binary and one covariate with four categories,
HCk3: 10 correlated covariates with correlation 0.9, one categorical covariate with three categories, two binary and one covariate with four categories.

TABLE 1: Simulation results for comparison of ridge and MLE with SSC in terms of MSE($\hat{\boldsymbol{\pi}}$) and MSE($\hat{\boldsymbol{\beta}}$)

			MLE		**SSC Ridge**			
situation	p	n	MSE($\hat{\pi}$)	MSE($\hat{\boldsymbol{\beta}}$)	MSE($\hat{\pi}$)	$lR_{\mathrm{ML}}(\hat{\pi})$	MSE($\hat{\boldsymbol{\beta}}$)	$lR_{\mathrm{ML}}(\hat{\boldsymbol{\beta}})$
Ik3	5	30	0.1359	25.8665	0.0975	−0.3500	4.2130	−1.2507
		50	0.0824	5.9079	0.0658	−0.2420	2.4656	−0.7061
		70	0.0549	1.5856	0.0474	−0.1510	0.9373	−0.4328
		100	0.0385	0.7731	0.0354	−0.0913	0.5924	−0.2516
	10	30	0.2026	216.7891	0.1504	−0.3437	152.3266	−2.1192
		50	0.1641	60.6962	0.1292	−0.2756	44.7314	−1.0421
		70	0.1118	9.7789	0.0925	−0.2084	6.1794	−0.7286
		100	0.0740	3.5824	0.0617	−0.1957	2.1672	−0.5316
	20	30	–	–	0.3377	–	26.6171	–
		50	–	–	0.2825	–	51.4760	–
		70	0.2011	485.6023	0.1794	−0.1511	426.7621	−0.9383
		100	0.1529	104.4179	0.1415	−0.0950	94.7832	−0.4985
Mk3	10	30	0.2140	408.9403	0.1464	−0.4080	116.3031	−2.6988
		50	0.1475	168.1165	0.1076	−0.3844	121.2961	−1.6518
		70	0.1086	31.1482	0.0846	−0.2880	17.2707	−1.1140
		100	0.0777	6.8345	0.0598	−0.2903	3.7639	−0.7489
	20	30	–	–	0.2741	–	22.8738	–
		50	–	–	0.2065	–	41.5213	–
		70	–	–	0.1753	–	44.9323	–
		100	0.1318	222.6823	0.1130	−0.1846	163.3361	−1.1780
Hk3	10	30	0.2068	827.0594	0.1076	−0.8392	1008.4812	−3.5708
		50	0.1521	401.0319	0.0929	−0.6888	232.4513	−2.3256
		70	0.1098	288.5410	0.0546	−0.8251	13.7831	−2.7377
		100	0.0791	133.4725	0.0391	−0.7781	7.0180	−2.3192
	20	30	–	–	0.1872	–	37.9161	–
		50	–	–	0.1555	–	55.5950	–
		70	–	–	0.1425	–	65.5167	–
		100	0.1466	1402.6828	0.1131	−0.3869	1393.6544	−1.6206
ICk3	17	30	–	–	0.2759	–	49.5630	–
		50	–	–	0.1854	–	52.4761	–
		70	0.1633	388.9480	0.1049	−0.5511	289.5700	−2.1113
		100	0.1195	174.8915	0.0805	−0.4609	71.6874	−1.9694
MCk3	17	30	–	–	0.2852	–	49.7037	–
		50	–	–	0.1769	–	46.1630	–
		70	0.1650	526.4872	0.1093	−0.4984	171.6959	−2.5559
		100	0.1308	220.6468	0.0940	−0.3964	199.1444	−1.7828
HCk3	17	30	–	–	0.2154	–	53.2974	–
		50	–	–	0.1796	–	70.5122	–
		70	0.1602	907.3471	0.0894	−0.7270	679.7028	−2.7218
		100	0.1228	411.4471	0.0646	−0.7030	16.8915	−3.0567

The parameter values used for situations ICk3, MCk3 and HCk3 were:

$$p = 17: \quad \boldsymbol{\beta}^T = \big((1, 10/11, \ldots, 1/11, 1, 6/7, \ldots, 1/7), (1/11, 2/11, \ldots, 1, 1/7, 2/7, \ldots, 1)\big).$$

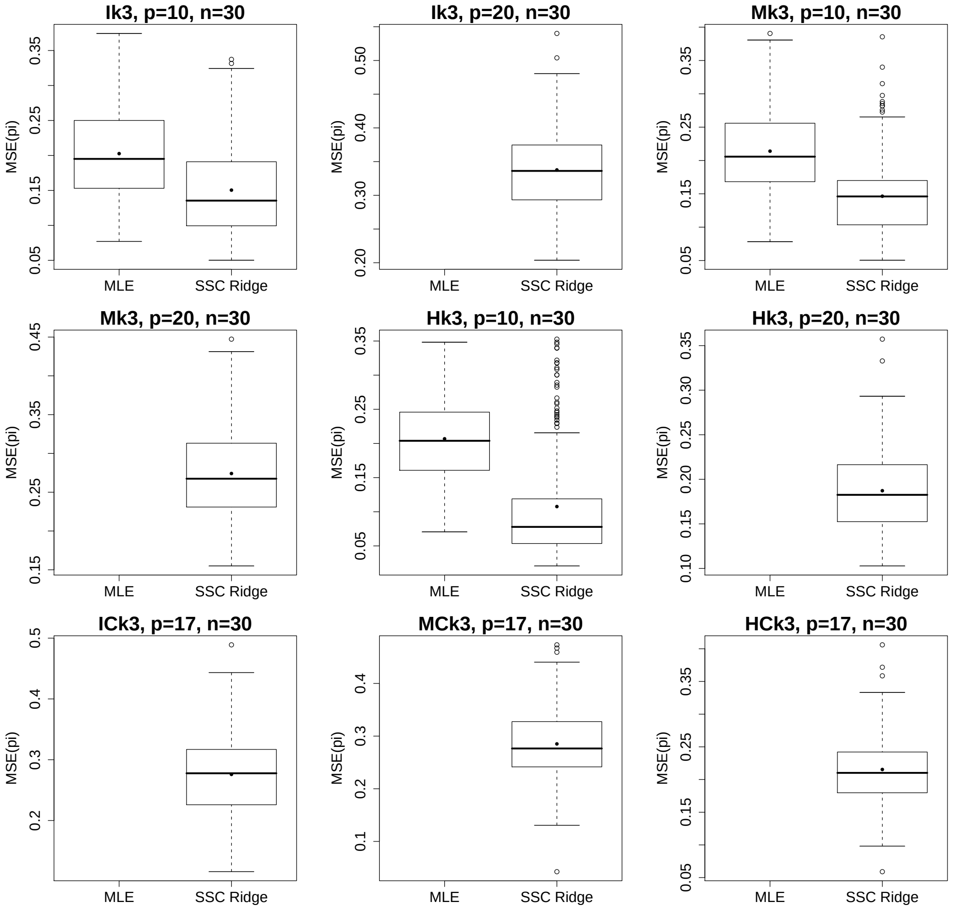

Figure 1: Illustration of the simulation study; Box plots for comparing ridge and MLE with SSC for $n = 30$ in terms of MSE($\hat{\boldsymbol{\pi}}$) .

In the study, independent continuous covariates were drawn from a standard normal distribution and for each setting $S = 200$ data sets were used. For computing the usual ML estimates, `multinom` function of library `nnet` in R was used. The results of usual MLE are not given in Table 1 if ML estimates were not converging and/or produced infinitely large standard errors. The values of tuning parameter λ for SSC-ridge were chosen by use of generalized cross-validation (GCV). The results of ridge estimates with symmetric side

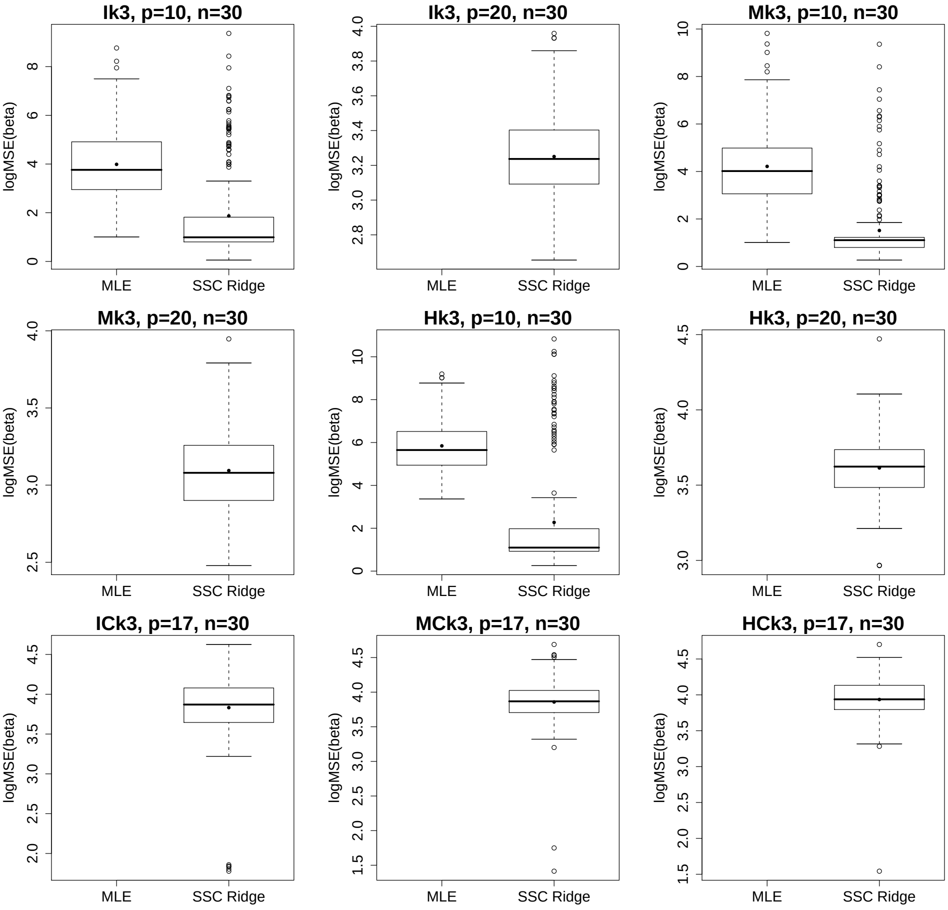

FIGURE 2: Illustration of the simulation study; Box plots for comparing ridge and MLE with SSC for $n = 30$ in terms of $\log(\text{MSE}(\hat{\boldsymbol{\beta}}))$.

constraint (SSC-ridge) and the ML estimates for SSC are compared on the basis of MSE (mean squared error) of $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\beta}}$. MSEs were computed using the estimates of all k logits as:

$$\text{MSE}(\hat{\boldsymbol{\pi}}) = \tfrac{1}{S}\textstyle\sum_s \text{MSE}_s(\hat{\boldsymbol{\pi}}) \qquad \text{with} \qquad \text{MSE}_s(\hat{\boldsymbol{\pi}}) = \frac{1}{kn}\sum_{i=1}^{n}\sum_{r=1}^{k}(\hat{\pi}_{ir} - \pi_{ir})^2 \text{ for the } s\text{th sample}$$

and

$$\text{MSE}(\hat{\boldsymbol{\beta}}) = \tfrac{1}{S}\textstyle\sum_s \|\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}\|^2$$

where $\hat{\boldsymbol{\pi}}$ is a vector of length kn and $\hat{\boldsymbol{\beta}}$ (vector of parameter estimates using SSC) and $\boldsymbol{\beta}$ are of length $k(p + 1)$.

Let MSE_{ssc} and MSE_{ML} represent the MSE's of $\hat{\boldsymbol{\pi}}$ (or $\hat{\boldsymbol{\beta}}$) for ridge and the usual MLE using the symmetric side constraint respectively. In Table 1 SSC-ridge estimates are compared with ML estimates. Improvement of estimates of SSC-ridge over MLE for simulation s can be measured by $\text{MSE}_{\text{ssc}}/\text{MSE}_{\text{ml}}$, but because the distribution of these ratios is skewed, we considered the mean across logarithms. In case of mean across logarithms we have $S^{-1}\sum_s \log(\text{MSE}_{\text{ssc}}/\text{MSE}_{\text{ML}}) = \log((\prod_s \text{MSE}_{\text{ssc}}/\text{MSE}_{\text{ML}})^{1/S})$ which refers to the logarithm of geometric mean.
In Table 1, $lR_{ML}(\hat{\boldsymbol{\pi}})$ and $lR_{ML}(\hat{\boldsymbol{\beta}})$ represent the means of $\log(\text{MSE}_{\text{ssc}}/\text{MSE}_{\text{ML}})$. The negative values of lR_{ML} indicate the improvement of the ridge method over usual MLE. Table 1 shows that usual MLEs do not exist for large number of covariates when samples size is small, but ridge estimates do. As the number of covariates increases and also in the case of collinearity ridge estimators definitely outperform MLEs in terms of $\text{MSE}(\hat{\boldsymbol{\pi}})$ and $\text{MSE}(\hat{\boldsymbol{\beta}})$. In Fig. 1 and Fig. 2, SSC-ridge is compared with MLE (if exists) in terms of box plots with respect to $\text{MSE}(\hat{\boldsymbol{\pi}})$ and $\text{MSE}(\hat{\boldsymbol{\beta}})$ respectively for the most interested case of small samples i.e., $n = 30$. The solid circles within the boxes of each box plot represent the mean of 200 values for which the box plots are drawn.

4. Application

In this section usual ML estimates (with reference category and symmetric side constraint) and the SSC-ridge estimates are computed for a data used by Agresti (2002) consisting of the factors influencing the primary food choice of 219 alligators captured in Florida lakes. Agresti (2002) fitted the baseline-category logit model using 'primary food choice' with five categories: Fish (F), Invertebrate (I), Reptile (R), Bird (B), and Others (O) as the response variable with 'Fish' as the reference category. The covariates used are L=Lake of capture (Hancock, Oklawaha, Trafford, George), G=gender (male, female) and S=size (<= 2.3 meters long, > 2.3 meters long. While comparing different models on the basis of G^2-values, the best fitted model is the (L+S) fitted on the data after grouping them over gender. We fit this model to get ML estimates (with RSC and SSC). The SSC-ridge estimates and their standard errors for this model are computed to compare them with the ML estimates. In Table 2 the estimates and their standard errors (within brackets) are shown for MLE with RSC (each of four logits is compared to the reference category "F") and SSC, and SSC-ridge (each logit is compared to the median response given by the geometric mean). The optimum value of the tuning parameter for SSC-ridge is $\lambda = 1.9$.

Table 2: Estimates and standard errors for ”Primary Food Choice of Alligator” data

Logit	Method of Estimation	Intercept	size <= 2.3	Hancock	Oklawaha	Trafford
I vs F	MLE with RSC	−1.5490 (0.4249)	1.4581 (0.3959)	−1.6581 (0.6128)	0.9372 (0.4719)	1.1220 (0.4905)
I vs median	MLE with SSC	0.2232 (0.3906)	1.2966 (0.3159)	−1.8795 (0.5384)	0.3875 (0.4595)	−0.2103 (0.4107)
I vs median	SSC-Ridge	0.1170 (0.2478)	0.9982 (0.2391)	−1.1208 (0.2863)	0.4489 (0.2581)	0.0553 (0.2521)
R vs F	MLE with RSC	−3.3145 (1.0531)	−0.3513 (0.5800)	1.2428 (1.1854)	2.4589 (1.1181)	2.9353 (1.1164)
R vs median	MLE with SSC	−1.5423 (0.8427)	−0.5128 (0.4509)	1.0216 (0.9513)	1.9092 (0.9089)	1.6030 (0.8788)
R vs median	SSC-Ridge	−0.6189 (0.2774)	−0.4913 (0.2917)	0.0201 (0.3044)	0.6573 (0.2891)	0.5386 (0.2851)
B vs F	MLE with RSC	−2.0934 (0.6623)	−0.6306 (0.6425)	0.6954 (0.7813)	−0.6526 (1.2020)	1.0881 (0.8417)
B vs median	MLE with SSC	−0.3209 (0.5597)	−0.7922 (0.5088)	0.4740 (0.6555)	−1.2029 (0.9733)	−0.2445 (0.6731)
B vs median	SSC-Ridge	−0.7121 (0.3042)	−0.4426 (0.3066)	0.4290 (0.3021)	−0.4223 (0.3086)	0.0414 (0.3023)
O vs F	MLE with RSC	−1.9043 (0.5258)	0.3316 (0.4483)	0.8263 (0.5575)	0.0058 (0.7766)	1.5165 (0.6214)
O vs median	MLE with SSC	−0.1321 (0.4595)	0.1700 (0.3551)	0.6050 (0.4975)	−0.5441 (0.6604)	0.1841 (0.5039)
O vs median	SSC-Ridge	−0.2307 (0.2655)	0.1157 (0.2624)	0.6056 (0.2718)	−0.3643 (0.3022)	0.2464 (0.2744)

Moreover, ridge estimates are compared with ML estimates in terms of MSPE (mean squared prediction error). For this purpose 50 random permutations of the 219 observations were taken and each was divided into two parts: the training data set with 169 observations and the parameter estimates obtained from these observations are used to get the squared prediction error from the test data set of other 50 observations using the formula

$$\mathrm{SPE}_s = \frac{1}{kn}\sum_{i=1}^{n}\sum_{r=1}^{k}(\hat{\pi}_{ir}^{test} - \pi_{ir}^{test})^2,$$

where π’s are the observed responses in the form of dummy variables 0 or 1. The MSPE for 50 random permutations computed as

$$\mathrm{MSPE} = \frac{1}{50}\sum_{s=1}^{50}\mathrm{SPE}_s.$$

The mean squared prediction error for MLE=33.9568, and for SSC-ridge=33.89551. Because the sample size is sufficiently large, the asymptotic theory supports the results of usual MLE and we do not see a significant improvement of ridge estimates over ML estimates. The results however show a little improvement of SSC-ridge over the MLE.
To compare the MLE and ridge estimates with respect to their existence and performance in small samples, we drew 50 random samples for each of size 30 and 50 from the original data of 219 observations and computed MLE as well as ridge estimates for each sample. The results (not shown here) indicated that MLE fails to exist in all samples of size $n = 30$ and $n = 50$ but ridge estimates do exist in every case.

5. Concluding Remarks

In multinomial logit models, the identifiability of parameter estimates calls for some side constraint, which typically means that some response category is chosen as the reference category, so that the parameter estimates can describe the effect of $\mathbf{x}$ on the logits when $P(Y = r|\mathbf{x}), \quad (r = 1, \ldots, k-1)$ is compared to the pre-defined reference category. The penalized estimates should be independent of the choice of the reference category. The use of symmetric side constraint given in (3) leads us to the use of ”median” response given by the geometric mean of all responses as the reference category rather than using a particular category as reference. The use of ”median” response as reference makes the penalization independent of reference category choice. This objective can be achieved for L2-penalty using the Fisher scoring in a very simple way, just by making a transformation of the actual design matrix and then using a matrix other than the identity matrix in the ridge penalty (as defined in (11)). In case of multicategory response, using symmetric side constraint is appropriate than to work with a reference category side constraint but one should be careful while interpreting the parameter estimates for each logit as these estimates are now subject to the ”median” response category as the reference rather than a particular response category of the data. However once these estimates with SSC are computed, one can transform these estimates back to the reference category scale by using the inverse transformation given in (7).

References

Agresti, A., 2002. Categorical Data Analysis. second ed. Chichester: Wiley, NewYork.

Bühlmann, P., Hothorn, T., 2007. Boosting algorithms: regularization, prediction and model fitting (with discussion). Statistical Science 22, 477–505.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the Americal Statistical Association 96, 1348–1360.

Friedman, J., Hastie, T., Tibshirani, R., 2008. Regularization paths for generalized linear models via coordinate descent.

James, G., Radchenko, P., 2009. A generalized dantzig selector with shrinkage tuning. Biometrika 96, 323–337.

Krishnapuram, B., Carin, L., Figueiredo, M. A., Hartemink, A. J., 2005. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 957–968.

Nyquist, H., 1991. Restricted estimation of generalized linear models. Journal of Applied Statistics 40, 133–141.

Park, M. Y., Hastie, T., 2007. L1-regularization path algorithm for generalized linear models. Journal of the Royal Statistical Society B 69, 659–677.

Schaefer, R., 1986. Alternative estimators in logistic regression when the data are collinear. Journal of Statistical Computation and Simulation 25, 75–91.

Schaefer, R., Roi, L., Wolfe, R., 1984. A ridge logistic estimator. Communications in Statistics: Theory and Methods 13, 99–113.

Segerstedt, B., 1992. On ordinary ridge regression in generalized linear models. Communications in Statistics: Theory and Methods 21, 2227–2246.

Tibshirani, R., 1996. Regression shrinkage and selection via lasso. Journal of the Royal Statistical Society B 58, 267–288.

Tutz, G., Binder, H., 2006. Generalized additive modelling with implicit variable selection by likelihood based boosting. Biometrcs 62, 961–971.

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