LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
INSTITUT FÜR STATISTIK

Faisal Maqbool Zahid & Christian Heumann

Regularized Proportional Odds Models

Technical Report Number 103, 2011
Department of Statistics
University of Munich

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

Regularized Proportional Odds Models

Faisal Maqbool Zahid* and Christian Heumann†

March 18, 2011

Abstract

The proportional odds model is commonly used in regression analysis to predict the outcome for an ordinal response variable. The maximum likelihood approach becomes unstable or even fails in small samples with relatively large number of predictors.The ML estimates also do not exist with complete separation in the data. An estimation method is developed to address these problems with MLE. The proposed method uses pseudo observations to regularize the observed responses by sharpening them so that they become close to the underlying probabilities. The estimates can be computed easily with all commonly used statistical packages supporting the fitting of proportional odds models with weights. Estimates are compared with MLE in a simulation study and two real life data sets.

KEY WORDS: Data sharpening, Logistic regression, Proportional odds model, Pseudo data, Regularization, Shrinkage estimation.

*Faisal Maqbool Zahid. Department of Statistics, Ludwig Strasse 33, 80539. Ludwig-Maximilians University Munich Germany (E-mail: faisalmz99@yahoo.com).

†Christian Heumann is Assistant Professor in Department of Statistics, Ludwig-Maximilians University Munich Germany (E-mail: christian.heumann@stat.uni-muenchen.de)

1 Introduction

The maximum likelihood approach with favourable asymptotic properties, plays a key role to fit the proportional odds models. The maximum likelihood estimation is sensitive to large number of predictors with small samples. The MLE does not respond with $p > n$ and/or complete separation in proportional odds models (POM). Penalized likelihood estimation is used to obtain the estimates in high-dimensional settings and/or ill-conditioned design space. Regularization techniques based on penalization typically maximize a penalized log-likelihood. Ridge regression, one of the oldest penalization methods for linear models was defined for logistic regression by Schaefer et al. (1984) and Schaefer (1986). Later on Nyquist (1991) and Segerstedt (1992) extended the concepts of ridge estimation for GLM type models. Different alternatives of MLE have been proposed in the literature for univariate GLMs e.g., 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 the boosting approach (Bühlmann and Hothorn (2007), Tutz and Binder (2006)). However, few approaches have been proposed for multi-categories response models. For example, Zhu and Hastie (2004) used ridge type penalization, Krishnapuram et al. (2005) considered multinomial logistic regression with lasso type estimates and Friedman et al. (2010) provided an efficient algorithm for the complete regularization path using L_1 penalty. Rousseeuw and Christmann (2003) proposed robust estimates for binary regression which always exist and are based on the responses which are closely related to the unobservable true responses. Tutz and Leitenstorfer (2006) considered a shrinkage type estimator for binary regression that has an improved existence than the usual MLE in different situations.

In this paper we are shrinking the parameter estimates without using any penalty term. We are exploiting the discreteness of the responses for regularization rather than restricting the range of parameters. The proposed technique is an extended version of the shrinkage estimates by Tutz and Leitenstorfer (2006), for proportional odds models. The shrinkage estimates are simple and easy to compute and interpret. We are downgrading the observed multinomial response $y_{ij} = 1$ and shrinking it towards the underlying probability by introducing $q = k - 1$ shadow (pseudo) data sets for a $k-$categories POM. For example, if the

observed response results in the jth category among category labels $1, \ldots, j, \ldots, k$, then the corresponding $k-1$ shadow responses have the category labels $1, \ldots, j-1, j+1, \ldots, k$. So each of $k-1$ categories other than the jth category in original data gets the representation against the ith response with same predictor vector but with different weights. As a result we are working with weighted log-likelihood for a data with $k \cdot n$ observations instead of usual log-likelihood (unweighted) for n observations of the original data. The estimates obtained with the use of shadow responses are more stable and have improved existence than the usual MLE. The estimates can be computed with any statistical software/package that fits the POM with weights. The layout of the paper is as follows:
In Section 2, the basic idea of regularizing the observed responses is described. In Section 3, three different approaches for regularization are discussed. The first simplest approach is based on some optimal values of the tuning parameters that are used for deciding the weights and are linked with the outcome category of each response. The second approach takes the fit into account and uses a weighting scheme associated with leverage measures where the weights are linked with the individual responses rather than the observed categories. The third approach is different from the second in the sense that it is based on the the diagnostic measures for logistic regression proposed by Pregibon (1981). The performance of all of these approaches is investigated and compared with the usual MLE in Section 4. An expression for the standard errors of shrinkage estimates is derived in Section 5. In Section 6, the estimates are computed for two real data sets. Section 7 completes the discussion with some concluding remarks.

2 Regularization of observed responses

For the ordinal responses, there are several models discussed in the literature. Ananth and Kleinbaum (1997) described these models with interpretation of the models' parameters (also see Agresti (1999)). Proportional odds model (also called cumulative logit model by McCullagh (1980)) is commonly used to model the ordinal responses. In this text, for regularization we are considering POM, however the proposed technique(s) can be applied in the same way to the other models available for ordinal responses. Let for a

given predictor vector $\mathbf{x}$, there is an observable variable $Y \in \{1, \ldots, k\}$ that is connected with an unobservable latent variable Z as $Y = r \;\Leftrightarrow\; \gamma_{0,r-1} < Z \leq \gamma_{0r},\ r = 1, \ldots, k$ where $-\infty = \gamma_{00} < \gamma_{01} < \ldots < \gamma_{0k} = \infty$. This indicates that Y is a categorized version of Z determined by $\gamma_{01}, \ldots, \gamma_{0q}$ (for $q = k - 1$). The cumulative logistic model has the form

$$P(Y \leq r|\mathbf{x}_i) = P(Z \leq \gamma_{0r}|\mathbf{x}_i) = \frac{\exp(\gamma_{0r} - \mathbf{x}_i^T \boldsymbol{\gamma}_r)}{1 + \exp(\gamma_{0r} - \mathbf{x}_i^T \boldsymbol{\gamma}_r)} \qquad r = 1, \ldots, q = k - 1, \quad (1)$$

or alternatively

$$\log\left[\frac{\phi_{ir}(\mathbf{x}_i)}{1 - \phi_{ir}(\mathbf{x}_i)}\right] = \gamma_{0r} - \mathbf{x}_i^T \boldsymbol{\gamma} \qquad r = 1, \ldots, k - 1, \quad (2)$$

where $\phi_{ir}(\mathbf{x}_i) = P(Y \leq r|\mathbf{x}_i)$ is the cumulative probability up to and including the category r for the covariate vector $\mathbf{x}_i$. In (2), each cumulative logit is increasing in r with its own intercept γ_{0r} and a global parameter vector $\boldsymbol{\gamma}$. The parameters $\{\gamma_{0r}\}$ and $\boldsymbol{\gamma}$ are unknown and $\{\gamma_{0r}\}$ must satisfy $\gamma_{01} < \ldots < \gamma_{0q}$, to ensure that the fitted probabilities are positive. For the estimation of parameters in POM with k response categories and p covariates, let the cumulative logit model has the form

$$\text{logit}(\phi_{ir}) = \mathbf{X}_i \boldsymbol{\beta}.$$

Here $(q \times p^*)$-matrix $\mathbf{X}_i$ (with $p^* = p + q$) given by

$$\mathbf{X}_i = \begin{bmatrix} 1 & & & \mathbf{x}_i^T \\ & 1 & & \mathbf{x}_i^T \\ & & \ddots & \vdots \\ & & 1 & \mathbf{x}_i^T \end{bmatrix},$$

is a component of the complete design matrix $\mathbf{X}$ of order $nq \times p^*$ which is given by

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

The parameter vector $\boldsymbol{\beta}$ has length p^* with components $\boldsymbol{\beta}^T = (\gamma_{01}, \ldots, \gamma_{0q}, \gamma_1, \ldots, \gamma_p)$. For obtaining a regularized version of the responses, we generate q shadow/pseudo data sets of the original data. Each response y_{ij} in the original data with design point $\mathbf{x}_i$ has q shadow responses with identical value of the design point. In a simple case, let the category 1 is observed in a three category POM (with category labels $1, 2$ and 3) with predictor vector $\mathbf{x}$. For this response, we have two shadow responses with category labels 2 and 3 with the same predictor vector $\mathbf{x}$. In other words, using the vector of dummies, the response $(0, 0)$ has two shadow responses $(1, 0)$ and $(0, 1)$. In general, with $k-$categories response model we are generating $q = k - 1$ shadow responses corresponding to each response in the original data and the original sample of size n is increased to $k \cdot n$. Different weights are assigned to the ith original response and the corresponding shadow responses in such a way that sum of these weights equals one. If α_{is} $(s = 1, \ldots, q)$ for $\alpha_{is} \in [0, \frac{1}{k}]$ is the weight assigned to the ith observation of the sth shadow data then $\alpha_{i0} = 1 - \sum_{s=1}^{q} \alpha_{is}$ is the corresponding weight for the ith observation in the original data. The weighted log likelihood function with the original and shadow data is given by

$$l_w(\boldsymbol{\beta}) = \sum_{s=0}^{q} \sum_{i=1}^{n} w_{is} l_{is}(\boldsymbol{\beta}), \tag{3}$$

where

$$l_{is}(\boldsymbol{\beta}) = \sum_{r=1}^{k} \log(\pi_{irs})^{y_{irs}} = \sum_{r=1}^{k} \log\big[\phi_{irs} - \phi_{i,r-1,s}\big]^{y_{irs}},$$

with y_{irs} as the ith response with category r in the sth data set ($s = 0$ represents the original data and $s = 1, \ldots, q$ represent each of q shadow data sets). The weights w_{is} can

be given as

$$w_{is} = \begin{cases} \alpha_{i0} & \text{for } s = 0 \\ \alpha_{is} & \forall\ s = 1, \ldots, q. \end{cases}$$

For the shadow data sets if we have $\alpha_{is} = 0, \forall i$, we are left with the usual (unweighted) log-likelihood function for the original observations. With increasing values of α_{is} $(s = 1, \ldots, q)$, shadow data will get more weights and small weights are assigned to the original responses.

We define the weighted log-likelihood function with original and shadow responses in the simplest case of a response variable with three (ordered) categories labeled $1, 2$ and 3. The weights for the ith observation of two shadow data sets are α_{i1} and α_{i2} respectively and the corresponding weight for the original response is $\alpha_{i0} = 1 - \sum_{s=1}^{2} \alpha_{is}$. The weighted log-likelihood function in this case is given by

$$\begin{aligned} l_w(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \Bigg[\alpha_{i0} \Big\{ y_{i1}\log(\pi_{i1}) + y_{i2}\log(\pi_{i2}) + y_{i3}\log(\pi_{i3}) \Big\} \\ &\qquad + \alpha_{i1} \Big\{ y_{i3}\log(\pi_{i1}) + y_{i1}\log(\pi_{i2}) + y_{i2}\log(\pi_{i3}) \Big\} \\ &\qquad + \alpha_{i2} \Big\{ y_{i2}\log(\pi_{i1}) + y_{i3}\log(\pi_{i2}) + y_{i1}\log(\pi_{i3}) \Big\} \Bigg] \\ &= \sum_{i=1}^{n} \Bigg[\Big\{ y_{i1} + (y_{i3} - y_{i1})\alpha_{i1} + (y_{i2} - y_{i1})\alpha_{i2} \Big\} \log(\pi_{i1}) \\ &\qquad + \Big\{ y_{i2} + (y_{i1} - y_{i2})\alpha_{i1} + (y_{i3} - y_{i2})\alpha_{i2} \Big\} \log(\pi_{i2}) \\ &\qquad + \Big\{ y_{i3} + (y_{i2} - y_{i3})\alpha_{i1} + (y_{i1} - y_{i3})\alpha_{i2} \Big\} \log(\pi_{i3}) \Bigg] \\ &= \sum_{i=1}^{n} \tilde{y}_{i1}\log(\pi_{i1}) + \tilde{y}_{i2}\log(\pi_{i2}) + \tilde{y}_{i3}\log(\pi_{i3}) = \sum_{i=1}^{n}\sum_{j=1}^{3} \tilde{y}_{ij}\log(\pi_{ij}). \end{aligned}$$

The final expression for $l_w(\boldsymbol{\beta})$ indicates that the use of shadow responses with different weights transform the original response y_{ij} into $\tilde{y}_{ij}$. As a result the weighted log-likelihood with shadow data simplifies to an un-weighted log-likelihood of transformed responses $\tilde{y}_{ij}$

with probabilities π_{ij}. We can say that the shadow responses with different weights are used to process the exaggerated value of the original response $y_{ij} = 1$ to transform it into a more regularized version with a value smaller than 1. The log-likelihood function with the transformed responses (or weighted log-likelihood with shadow data sets) for a $k-$categories response model can be given as

$$l_w(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{k} \tilde{y}_{ij}\log(\pi_{ij}) \tag{4}$$

where the ith transformed observation for the jth category is given by

$$\tilde{y}_{ij} = \begin{cases} y_{ij} + \sum_{r=2}^{k}(y_{ir} - y_{ij})\alpha_{i,k-(r-1)} & \text{if } j = 1, \\ y_{ij} + \sum_{r=1}^{k-1}(y_{ir} - y_{ij})\alpha_{i,j-r} & \text{if } j = k, \\ y_{ij} + \sum_{r=1}^{j-1}(y_{ir} - y_{ij})\alpha_{i,j-r} + \sum_{r=j+1}^{k}(y_{ir} - y_{ij})\alpha_{i,(k+j)-r} & \text{otherwise.} \end{cases}$$

We can proceed with the log-likelihood function given in (4) instead of working with weighted log-likelihood given in (3), but working with the weighted version is easy and simple using any statistical software that fits the proportional odds models using weights. Also in the weighted version the sample size is artificially increased with shadow data which not only improves the existence of estimates but also the estimates become more stable than the usual MLE. The weighted score function for $k \cdot n$ observations comprised of original and shadow responses is given by

$$s_w(\boldsymbol{\beta}) = \frac{\partial l_w(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{k\cdot n} s_{wi}(\boldsymbol{\beta}),$$

with the ith component as

$$s_{wi}(\boldsymbol{\beta}) = \mathbf{X}_i^T \text{diag}(\mathbf{w}_i)\mathbf{D}_i(\boldsymbol{\beta})\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta})[\mathbf{y}_i - h(\boldsymbol{\eta}_i)],$$

where $\mathbf{w}_i$ are the weights associated with the ith observation, $\mathbf{D}_i(\boldsymbol{\beta}) = \frac{\partial h(\boldsymbol{\eta}_i)}{\partial \boldsymbol{\eta}}$ is the derivative of $h(\boldsymbol{\eta})$ evaluated at $\boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_i(\boldsymbol{\beta}) = \text{cov}(\mathbf{y}_i)$ is the covariance matrix of $\mathbf{y}_i$ given the parameter vector $\boldsymbol{\beta}$. Alternatively the score function can be written as $s_{wi}(\boldsymbol{\beta}) = \mathbf{X}_i^T \text{diag}(\mathbf{w}_i)\mathbf{W}_i(\boldsymbol{\beta})\frac{\partial g(\boldsymbol{\mu}_i)}{\partial \boldsymbol{\mu}^T}[\mathbf{y}_i - h(\boldsymbol{\eta}_i)]$ with $\mathbf{W}_i(\boldsymbol{\beta}) = \mathbf{D}_i(\boldsymbol{\beta})\boldsymbol{\Sigma}_i^{-1}\mathbf{D}_i^T(\boldsymbol{\beta}) = \left\{\frac{\partial g(\boldsymbol{\mu}_i)}{\partial \boldsymbol{\mu}^T}\boldsymbol{\Sigma}_i(\boldsymbol{\beta})\frac{\partial g(\boldsymbol{\mu}_i)}{\partial \boldsymbol{\mu}}\right\}^{-1}$. In matrix notation

$$\begin{aligned} s_w(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \mathbf{X}_i^T \text{diag}(\mathbf{w}_i)\mathbf{D}_i(\boldsymbol{\beta})\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta})[\mathbf{y}_i - h(\boldsymbol{\eta}_i)] \\ &= \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{D}_i(\boldsymbol{\beta})\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta})[\tilde{\mathbf{y}}_i - h(\boldsymbol{\eta}_i)] \qquad (5) \\ &= \mathbf{X}^T\mathbf{D}(\boldsymbol{\beta})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta})[\tilde{\mathbf{y}} - h(\boldsymbol{\eta})] \end{aligned}$$

where $\mathbf{y}$ and $h(\boldsymbol{\eta})$ are given by $\mathbf{y}^T = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)$, and $h(\boldsymbol{\eta}) = (h(\boldsymbol{\eta}_1), \ldots, h(\boldsymbol{\eta}_n))^T$ respectively. The matrices have block diagonal form as $\boldsymbol{\Sigma}(\boldsymbol{\beta}) = \text{diag}(\boldsymbol{\Sigma}_i(\boldsymbol{\beta}))$, $\mathbf{W}(\boldsymbol{\beta}) = \text{diag}(\mathbf{W}_i(\boldsymbol{\beta}))$, $\mathbf{D}(\boldsymbol{\beta}) = \text{diag}(\mathbf{D}_i(\boldsymbol{\beta}))$. The general form of the score equations with transformed (regularized) responses can be written as

$$\frac{\partial l_w(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{x}_i(\tilde{\mathbf{y}}_i - h(\boldsymbol{\eta}_i)) = 0 \qquad (6)$$

Score equations in (6) use a regularized version of the original responses y_{ij} such that if $y_{ij} = 1$, the regularized version $\tilde{y}_{ij}$ assumes the value $1 - \sum_{s=1}^{q} \alpha_s$, which is less than 1. If we use the equal weights i.e., $\alpha_{is} = \frac{1}{k}$ $(\forall s = 1, \ldots, q)$, the score equations in (6) lead to a solution with $\boldsymbol{\beta} = \mathbf{0}$.

3 Regularization techniques

3.1 Category specific regularization (CSR)

The basic idea behind the regularization of the observed responses is that instead of using the exaggerated values of $y_{ij} = 1$, a smoothed version (where y_{ij} assumes a value less than 1) of these responses should be used. The use of shadow data helps to downgrade $y_{ij} = 1$ using different weights for shadow responses. In the simplest situation, the weights for each shadow response can be chosen in the interval $[0, \frac{1}{k}]$. In this section we are introducing a weighting scheme that assigns the same weight to each response in the shadow data with same category. We are exploiting the property of MLE for the proportional odds models (with intercept) given by

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_{ij} = \bar{y}_{.j}, \qquad j = 1, \ldots, k \tag{7}$$

Here $\bar{y}_{.j}$ is the mean of responses corresponding to the jth category. Different weights should be assigned to each of y_{ij} $(j = 1, \ldots, k)$ to hold the property (7). From (6) we have $k - 1$ score equations corresponding to the intercept terms which are of the form

$$\sum_{i=1}^{n}\tilde{\mathbf{y}}_{ij} = \sum_{i=1}^{n}\hat{\boldsymbol{\pi}}_{ij}. \tag{8}$$

For category specific weights (i.e., the same weight to all responses resulting in the jth category in the sahdow data), let $\tilde{\alpha}_j$ be the weight associated with category j. From (8), while holding (7) we obtain $(k - 1)$ equations as

$$\sum_{\substack{j=1 \\ j\neq r}}^{k}\bar{y}_{.j}\tilde{\alpha}_j - (k-1)\bar{y}_{.r}\tilde{\alpha}_r = 0 \qquad r = 1, \ldots, k-1. \tag{9}$$

solving this system of $k-1$ equations for $\tilde{\alpha}$'s, we get

$$\tilde{\alpha}_j = \frac{\bar{y}_{.k}}{\bar{y}_{.j}}\,\tilde{\alpha}_k, \quad j = 1, \ldots, k-1. \tag{10}$$

If $\tilde{\alpha}_k = \bar{y}_{.j}$, we have $\tilde{\alpha}_j = \bar{y}_{.k} \;\; \forall j$, i.e., each response $y_{ij} = 1$ is shrunk towards $\bar{y}_{.k}$. The optimal value of the tuning parameter $\tilde{\alpha}_j$ that minimizes the prediction error can be searched in the interval $[0, \bar{y}_{.j}]$. But if all the tuning parameters assumes the same value i.e., $\frac{1}{k}$, the solution for the estimates will be $\hat{\boldsymbol{\beta}} = \mathbf{0}$. If $\bar{y}_{.j} > \frac{1}{k}$, it is intuitive to search for optimal value of the weight $\tilde{\alpha}_j$ in the interval $[0, \frac{1}{k}]$. Since for $\tilde{\alpha}_j = \bar{y}_{.j}$, each response with $y_{ij} = 1 \quad (j = 1, \ldots, q)$ shrinks towards $\bar{y}_{.k}$, it is sensible to shrink the response $y_{ik} = 1$ towards the mean of the rest of $k-1$ response categories i.e., shrinking $y_{ik} = 1$ towards $\frac{1}{k-1}\sum_{j=1}^{k-1} \bar{y}_{.j}$. The resulting weighting scheme for the original data ($s = 0$) and the sth ($s = 1, \ldots, q$) shadow data set is given by

$$w_i = \begin{cases} 1 - \sum_{s=1}^{k-1} w_{is} & \text{for } s = 0 \\ w_{is} & \forall s = 1, \ldots, q, \end{cases} \tag{11}$$

where

$$w_{is} = \begin{cases} \frac{\bar{y}_{.k}}{\bar{y}_{.j}} \cdot \alpha_j & \text{for } \; y_{ij} = 1 \; (j \in \{1, \ldots, q\}) \\ \frac{1}{k-1}\sum_{j=1}^{k-1} \alpha_j & \text{for } \; y_{ik} = 1. \end{cases}$$

For the optimal values of q tuning parameters α_j $(j = 1, \ldots, q)$, we are using the leave-one-out cross-validation with the following distance measures:
Kullback-Leibler discrepancy given by

$$\mathrm{L_{KL}} = \sum_{j=1}^{k}\sum_{i=1}^{n} \pi_{ij} \log\left(\frac{\pi_{ij}}{\hat{\pi}_{ij}}\right),$$

with a convention that $0.\log(0) = 0$. The averaged squared error computed as

$$\mathrm{ASE} = \frac{1}{k \cdot n}\sum_{j=1}^{k}\sum_{i=1}^{n} (\pi_{ij} - \hat{\pi}_{ij})^2,$$

and the averaged L1-distances given by

$$\text{AL1} = \frac{1}{k \cdot n} \sum_{j=1}^{k} \sum_{i=1}^{n} |\pi_{ij} - \hat{\pi}_{ij}|.$$

For the leave-one-out cross-validation, the fit is computed using all the data except the ith observation. However to save the time and reduce the computational burden one can use k-fold cross-validation for searching optimal values of the tuning parameters.

3.2 Response specific regularization (RSR1)

The weighting scheme discussed in Section 3.1 is category dependent because it assigns the same weight to each response resulting in the jth category. These weights are based on some optimal values of q tuning parameters. The category specific weights have some drawbacks as: (i) We have to perform a grid search for searching the optimal values of the tuning parameters. The grid search increases the computational burden with the increase of number of response categories and making the approach less efficient regarding the time required for computing the estimates. (ii) Each response resulting in the jth category gets the same weight in the data irrespective of the residual value corresponding to a particular observation. We can overcome these problems by using a weighting scheme that does not depend on the tuning parameters and assigns the weights to the responses which are response dependant rather than the category dependant. To have such weights we exploit the information available in the hat matrix. The hat matrix provides a measure of leverage of the data and is a building block of the regression diagnostics. In case of multinomial response, iteratively reweighted least squares estimation is linked with the iterative fitting of pseudo observations $\mathbf{z} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{D}^{-\mathbf{1}}(\mathbf{y} - \hat{\boldsymbol{\mu}})$ (e.g., McCullagh and Nelder 1989; Fahrmeir and Tutz 2001). At convergence the maximum likelihood estimate is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{z}.$$

The estimate $\hat{\boldsymbol{\beta}}$ is a weighted least square solution for the linear problem $\tilde{\mathbf{z}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, or alternatively we can say that $\hat{\boldsymbol{\beta}}$ is an unweighted least squares solution of the linear problem $\tilde{\mathbf{z}}_\mathbf{0} = \mathbf{X}_\mathbf{0}\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}$ with $\tilde{\mathbf{z}}_\mathbf{0} = \mathbf{W}^{\mathbf{T/2}}\tilde{\mathbf{z}}$ and $\mathbf{X}_\mathbf{0} = \mathbf{W}^{\mathbf{T/2}}\mathbf{X}$. For this model the hat matrix is given by

$$\begin{aligned} \mathbf{H} &= \mathbf{X}_\mathbf{0}(\mathbf{X}_\mathbf{0}^\mathbf{T}\mathbf{X}_\mathbf{0})^{-\mathbf{1}}\mathbf{X}_\mathbf{0}^\mathbf{T} \\ &= \mathbf{W}^{\mathbf{T/2}}\mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{W}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\mathbf{T}\mathbf{W}^{\mathbf{1/2}} \end{aligned} \tag{12}$$

The same form of the hat matrix for multinomial response models was used by Lesaffre and Albert (1989). Here $\mathbf{W} = \mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}^T$ with $\mathbf{D} = \text{diag}(\mathbf{D}_i) = \text{diag}(\frac{\partial h(\boldsymbol{\eta}_i)}{\partial \boldsymbol{\eta}})$. In the context of linear models diagonal elements $h_{ii}(0 \leq h_{ii} \leq 1)$ of the hat matrix $\mathbf{H}$ provide the information about the extreme design points corresponding to the high values (close to one) of h_{ii}. In contrast to the classical linear models, the hat matrix here not only depends on the design matrix $\mathbf{X}$ but also on the fit and we may have extreme points in the design space even for a smaller value of h_{ii}. In case of multi-category response the $(nq \times nq)$-hat matrix $\mathbf{H}$ has n diagonal matrices $\mathbf{H}_{ii}$ of order $(q \times q)$ on its diagonal. The measures tr$(\mathbf{H}_{ii})$ or det$(\mathbf{H}_{ii})$ of the block diagonal matrix $\mathbf{H}_{ii}$ can be used as an indicator of the leverage of $\mathbf{y}_i$. In this section we use these measures to construct our weighting scheme. Like the category specific regularization one can shrink all $y_{ij} = 1$ (in the shadow data) towards $\frac{1}{k}$ by selecting $w_i = \frac{1}{k}$tr$(\mathbf{H}_{ii})$ or $w_i = \frac{1}{k}$det$(\mathbf{H}_{ii})$, which will give maximum shrinkage when the leverage measure i.e., tr$(\mathbf{H}_{ii})$ or det$(\mathbf{H}_{ii})$ assumes the value 1. But here we opt a little different approach and shrinking each $y_{ij} = 1 \quad (j = 1, \ldots, k)$ towards the average of means of the rest of $(k-1)$ categories. The weights for original data $(s = 0)$ and sth shadow data $(s = 1, \ldots, q)$ with this approach are given as:

$$w_i = \begin{cases} 1 - \sum_{s=1}^{k-1} w_{is} & \text{for } s = 0, \\ w_{is} & \forall s = 1, \ldots, q, \end{cases} \tag{13}$$

with

$$w_{is} = \frac{1}{k-1}(1 - \bar{y}_{.j})\ (\text{det/tr}(\mathbf{H}_{ii})) \quad \text{for} \quad y_{ij} = 1.$$

3.3 Regularization with Pregibon's Hat Matrix (RSR2)

Pregibon (1981) developed some diagnostic measures for the logistic models to measure the effect of outlying responses and extreme design points on the maximum likelihood fit. The $(nq \times nq)$-matrix $\mathbf{H}$ given in (12) is a symmetric and idempotent matrix. Since $\mathbf{X_0}\hat{\boldsymbol{\beta}} = \mathbf{H}\tilde{\mathbf{z}}_\mathbf{0}$, the matrix $\mathbf{H}$ is a projection matrix mapping the observations $\tilde{\mathbf{z}}_\mathbf{0}$ into the fitted values $\mathbf{W}^{T/2}\mathbf{X}\hat{\boldsymbol{\beta}}$. Pregibon (1981) considered another symmetric and idempotent matrix $\mathbf{H}^*$ given by

$$\mathbf{H}^* = \mathbf{W}^{T/2}\mathbf{X}^*(\mathbf{X}^{*T}\mathbf{W}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}\mathbf{W}^{1/2}, \quad \mathbf{X}^* = (\mathbf{X}, \mathbf{z}) \tag{14}$$

In Section 3.2, for regularization we used the information contained in the block diagonal matrices $\mathbf{H}_{ii}$ of the hat matrix $\mathbf{H}$ given by (12). We used there the trace and determinant of matrices $\mathbf{H}_{ii}$. If we consider the trace of sub-diagonal matrices $\mathbf{H}_{ii}$ and $\mathbf{H}^*_{ii}$ as the diagnostic measure then after some algebraic derivation, it can be shown that the usual hat matrix $\mathbf{H}$ and that one given by Pregibon are connected with each other by the relation

$$\text{tr}(\mathbf{H}^*_{ii}) = \text{tr}(\mathbf{H}_{ii}) + \frac{\chi_i^2}{\chi^2}, \tag{15}$$

where $\chi^2 = \sum_{i=1}^n \chi_i^2$, with the ith component $\chi_i^2 = (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)$. The extreme points in the design space are reflected by the large values of $\text{tr}(\mathbf{H}_{ii})$ and χ_i^2/χ^2 shows the relative poor fit. So the large values of $\text{tr}(\mathbf{H}^*_{ii})$ indicate extreme points in the design space, poorly fitted observations or both. The value of $\text{tr}(\mathbf{H}^*_{ii})$ provides a measure of influence of the observation $\mathbf{y}_i$ like $\text{tr}(\mathbf{H}_{ii})$ of the usual hat matrix, so we can shrink the responses using $\text{tr}(\mathbf{H}^*_{ii})$. The weighting scheme for original data ($s = 0$) and sth shadow data ($s = 1, \ldots, q$) is similar to (13) and is given by

$$w_i^* = \begin{cases} 1 - \displaystyle\sum_{s=1}^{k-1} w_{is}^* & \text{for } s = 0, \\ w_{is}^* & \forall s = 1, \ldots, q, \end{cases} \tag{16}$$

with

$$w_{is}^{*} = \frac{1}{k-1}(1-\bar{y}_{.j})\ \text{tr}(\mathbf{H}_{ii}^{*}) \quad \text{for} \quad y_{ij}=1,$$

Although $\mathbf{H}$ and $\mathbf{H}^{*}$ have connection given in (15) only in terms of trace of sub-diagonal matrices, but the determinant may also be used as the leverage measure as in the case of usual hat matrix. In the simulation study we also use $\det(\mathbf{H}_{ii}^{*})$ along with $\text{tr}(\mathbf{H}_{ii}^{*})$ as a leverage measure.

The RSR1 and RSR2 have certain advantages over CSR as: (i) each response observation gets different weight on the basis of some leverage measure whereas CSR assigns weights to the observations subject to the outcome category without considering the corresponding fit. (ii) no grid search is to be performed as in CSR for the selection of optimal values of tuning parameters and processing time for grid search and cross-validation is saved. As a result it makes the fitting procedure more efficient with respect to processing time especially with increasing number of response categories. (iii) simulation study shows that RSR1 and RSR2 perform excellently in terms of MSE($\hat{\boldsymbol{\beta}}$) and MSE($\hat{\boldsymbol{\pi}}$). In response specific regularization (RSR) we need the hat matrix based on maximum likelihood fit, which will be missing in case of non-existence of usual MLE. This problem can be overcome with CSR approach. We can compute the response shrinkage estimates with smallest possible values of the tuning parameters that the numerical procedures allow and from this fit we can get the hat matrix to proceed to the response specific regularization.

4 Simulation Study

In a simulation study, we generated Gaussian data with n observations and p covariates. We used different number of combinations of n ($n = 30, 50$ and 100) and p ($p = 2, 5, 10, 15$ and 20). The values of the global parameters $\gamma_j = (-1)^j \exp(-2(j-1)/20)$ for $j = 1, \ldots, p$ and the intercept values $\gamma_{01} = -0.3$ and $\gamma_{02} = 0.8$ are used. The covariates are drawn from $N(0, 1)$. In each combination of n and p, $S = 200$ data sets are generated. The function `polr` of the package `MASS` in statistical environment/ language R is used to compute the usual MLE and the estimates with CSR, RSR1 and RSR2. To

compare our estimates with usual MLE, only those samples are considered in the study for which the usual MLE exists. The likelihood estimates for the setting $n = 30$ & $p = 20$ are not given in Table 1 because we could not get $S = 200$ samples in this setting for which MLE exist. For this setting we get the estimates for RSR1 and RSR2 on the basis of hat matrix obtained from CSR with smallest possible values of the tuning parameters allowed by the numerical procedure.

Table 1: Results of simulation study: Comparison of MLE and response shrinkage methods in terms of MSE($\hat{\boldsymbol{\pi}}$) and MSE($\hat{\boldsymbol{\beta}}$)

		MLE		CSR, CV(KL)				CSR, CV(SE)			
p	n	MSE($\hat{\pi}$)	MSE($\hat{\boldsymbol{\beta}}$)	MSE($\hat{\pi}$)	$lR_{\mathrm{ML}}(\hat{\pi})$	MSE($\hat{\boldsymbol{\beta}}$)	$lR_{\mathrm{ML}}(\hat{\boldsymbol{\beta}})$	MSE($\hat{\pi}$)	$lR_{\mathrm{ML}}(\hat{\pi})$	MSE($\hat{\boldsymbol{\beta}}$)	$lR_{\mathrm{ML}}(\hat{\boldsymbol{\beta}})$
2	30	0.0412	1.1648	0.0404	−0.0063	0.6726	−0.3618	0.0389	−0.0641	0.6820	−0.3904
	50	0.0245	0.6195	0.0240	−0.0565	0.4564	−0.2018	0.0246	−0.0424	0.4829	−0.1827
	100	0.0102	0.2435	0.0100	−0.0322	0.1920	−0.1546	0.0103	−0.0069	0.1964	−0.1180
5	30	0.0672	3.1974	0.0652	−0.0382	1.2136	−0.6687	0.0632	−0.0779	1.2662	−0.6597
	50	0.0419	1.9956	0.0369	−0.1102	0.9683	−0.3375	0.0367	−0.1019	0.9994	−0.3083
	100	0.0210	0.5667	0.0194	−0.0751	0.4292	−0.1992	0.0199	−0.0490	0.4551	−0.1595
10	30	0.1058	5.3151	0.1044	0.0094	2.6814	−0.4098	0.1034	−0.0015	2.6602	−0.4194
	50	0.0749	4.4559	0.0575	−0.2535	1.8036	−0.6856	0.0589	−0.2293	1.9302	−0.6376
	100	0.0371	1.4670	0.0302	−0.2179	0.7558	−0.5471	0.0305	−0.2055	0.7828	−0.5083
15	30	0.1509	13.5638	0.2181	0.3867	4.4315	−0.7432	0.2196	0.3954	4.4388	−0.7414
	50	0.0985	10.2853	0.0847	−0.1293	2.7326	−1.0134	0.0832	−0.1470	2.7152	−1.0223
	100	0.0563	3.0929	0.0415	−0.2977	1.1419	−0.8075	0.0422	−0.2815	1.1897	−0.7801
20	30	–	–	0.1542	–	11.2196	–	0.1664	–	15.4287	–
	50	0.1205	7.3818	0.1298	0.1019	3.7875	−0.4733	0.1309	0.1065	3.7890	−0.4740
	100	0.0731	5.6302	0.0518	−0.3340	1.5527	−1.0575	0.0518	−0.3367	1.5495	−1.0597

Table 1: Continued

		CSR, CV(L1)				RSR1, (using trace)				RSR2, (using trace)			
p	n	MSE($\hat{\pi}$)	$lR_{\text{ML}}(\hat{\pi})$	MSE($\hat{\boldsymbol{\beta}}$)	$lR_{\text{ML}}(\hat{\boldsymbol{\beta}})$	MSE($\hat{\pi}$)	$lR_{\text{ML}}(\hat{\pi})$	MSE($\hat{\boldsymbol{\beta}}$)	$lR_{\text{ML}}(\hat{\boldsymbol{\beta}})$	MSE($\hat{\pi}$)	$lR_{\text{ML}}(\hat{\pi})$	MSE($\hat{\boldsymbol{\beta}}$)	$lR_{\text{ML}}(\hat{\boldsymbol{\beta}})$
2	30	0.0434	0.0130	1.0546	−0.0473	0.0339	−0.2000	0.6958	−0.3887	0.0317	−0.2830	0.5274	−0.6278
	50	0.0250	−0.0106	0.5828	−0.0531	0.0215	−0.1537	0.4643	−0.2368	0.0224	−0.1196	0.4125	−0.2983
	100	0.0092	−0.0958	0.1984	−0.1552	0.0095	−0.0649	0.2049	−0.1235	0.0093	−0.0918	0.1838	−0.1984
5	30	0.0643	−0.0558	2.3479	−0.1901	0.0543	−0.2086	1.3368	−0.6809	0.0526	−0.2271	1.0213	−0.8352
	50	0.0392	−0.0654	1.5575	−0.1502	0.0345	−0.1937	1.0441	−0.4121	0.0317	−0.2728	0.7248	−0.6120
	100	0.0195	−0.0764	0.4600	−0.1608	0.0189	−0.1046	0.4258	−0.2283	0.0185	−0.1284	0.3798	−0.3144
10	30	0.0950	−0.1003	3.3757	−0.3014	0.1008	−0.0351	2.4709	−0.5743	0.1048	0.0165	2.2349	−0.6124
	50	0.0701	−0.0648	3.5005	−0.1777	0.0578	−0.2470	1.7323	−0.7799	0.0596	−0.2096	1.5587	−0.8126
	100	0.0336	−0.1045	1.0893	−0.2596	0.0326	−0.1321	0.8584	−0.4458	0.0309	−0.1828	0.7033	−0.6056
15	30	0.1306	−0.1414	4.6206	−0.7870	0.2012	0.3252	5.0608	−0.6819	0.2124	0.3848	4.5655	−0.7385
	50	0.0907	−0.0835	6.8066	−0.3189	0.0863	−0.1146	3.4459	−0.9103	0.0831	−0.1499	2.4549	−1.1624
	100	0.0508	−0.1035	2.1482	−0.3046	0.0458	−0.2033	1.4261	−0.6565	0.0437	−0.2447	1.1083	−0.8477
20	30	0.2715	–	61.2610	–	0.1748	–	6.9364	–	0.1808	–	6.2498	–
	50	0.1095	−0.0936	4.5475	−0.4179	0.1085	−0.0866	3.2278	−0.6961	0.1062	−0.1069	2.8398	−0.7925
	100	0.0641	−0.1326	3.2490	−0.4310	0.0612	−0.1698	2.2120	−0.8000	0.0588	−0.2060	1.6879	−0.9977

Table 1: Continued

		RSR1, (using determinant)				RSR2, (using using determinant)			
p	n	MSE($\hat{\pi}$)	$lR_{\mathrm{ML}}(\hat{\pi})$	MSE($\hat{\boldsymbol{\beta}}$)	$lR_{\mathrm{ML}}(\hat{\boldsymbol{\beta}})$	MSE($\hat{\pi}$)	$lR_{\mathrm{ML}}(\hat{\pi})$	MSE($\hat{\boldsymbol{\beta}}$)	$lR_{\mathrm{ML}}(\hat{\boldsymbol{\beta}})$
2	30	0.0409	−0.0049	1.5310	−0.0074	0.0407	−0.0109	1.1346	−0.0213
	50	0.0244	−0.0022	0.6175	−0.0029	0.0244	−0.0023	0.6151	−0.0062
	100	0.0102	−0.0004	0.2432	−0.0007	0.0102	−0.0014	0.2426	−0.0028
5	30	0.0666	−0.0095	3.1155	−0.0198	0.0661	−0.0176	3.0584	−0.0448
	50	0.0417	−0.0051	1.9727	−0.0076	0.0416	−0.0084	1.9616	−0.0150
	100	0.0210	−0.0010	0.5656	−0.0016	0.0210	−0.0020	0.5642	−0.0039
10	30	0.1044	−0.0142	5.0886	−0.0320	0.1034	−0.0243	4.9594	−0.0643
	50	0.0744	−0.0070	4.3635	−0.0166	0.0743	−0.0075	4.3572	−0.0188
	100	0.0371	−0.0013	1.4609	−0.0035	0.0371	−0.0022	1.4574	−0.0063
15	30	0.1491	−0.0111	12.6369	−0.0434	0.1479	−0.0194	12.2154	−0.0961
	50	0.0975	−0.0098	9.9831	−0.0245	0.0974	−0.0108	9.9573	−0.0296
	100	0.0561	−0.0028	3.0692	−0.0064	0.0561	−0.0031	3.0674	−0.0072
20	30	0.3481	−	212.6165	−	0.3442	−	232.4355	−
	50	0.1195	−0.0089	7.2101	−0.0208	0.1190	−0.0132	7.1309	−0.0362
	100	0.0729	−0.0030	5.5685	−0.0085	0.0728	−0.0033	5.5650	−0.0096

In Table 1, regularization techniques CSR (with weights given in (11)) and RSR (with weights given in (13) and (16)) are compared with the usual MLE in terms of MSE($\hat{\boldsymbol{\beta}}$) and MSE($\hat{\boldsymbol{\pi}}$). For the CSR, optimal values of tuning parameters are chosen by cross validation based on error measures described in section (3.1) and the corresponding results are denoted by CV(KL), CV(SE) and CV(L1) for Kullback-Leibler, squared error loss and the L1 distance respectively. MSE($\hat{\boldsymbol{\beta}}$) and MSE($\hat{\boldsymbol{\pi}}$) are computed as:

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

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 $k\cdot n$ and $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$ are of length $p+k-1$. Let MSE_s is the MSE of $\hat{\boldsymbol{\pi}}$ (or $\hat{\boldsymbol{\beta}}$) for a particular regularization approach and MSE_s^{ML} is the corresponding MSE for the maximum likelihood estimate. The ratio $\text{MSE}_s/\text{MSE}_s^{ML}$ for the sth simulation will provide a measure of improvement of a particular regularization method over MLE. The distribution of the ratios $\text{MSE}_s/\text{MSE}_s^{ML}$ is skewed and therefore the logarithms of these ratios are considered. In Table 1 along with the MSE, the means of $\log(\text{MSE}_s/\text{MSE}_s^{ML})$ denoted by $lR_{ML}(\hat{\pi})$ and $lR_{ML}(\hat{\boldsymbol{\beta}})$ are considered for comparing the regularization techniques with the usual MLE. The negative values of these log-ratios refer to an improvement of regularized estimates over the usual ML estiamtes. Our main focus is to develop an estimation technique that assures the existence of parameter estimates especially in case of large number of covariates with small samples or with no overlapping observations in the data. However we also consider a simple case of only two covariates even with a large sample size to observe the behaviour of our regularized estimates. As the asymptotic theory behaves well with increasing sample size and one can expect better performance of usual ML estimates relative to our estimates in a weighted fashion. An overall view of Table 1 reflects a better performance of our estimates in every situation not only with respect to the parameter estimates but also the fit. The use of L1 distance although provides better results than MLE but they are close to the usual MLE. According to our experience when L1 distances are used as a distance measure, some of the optimal values of α's are

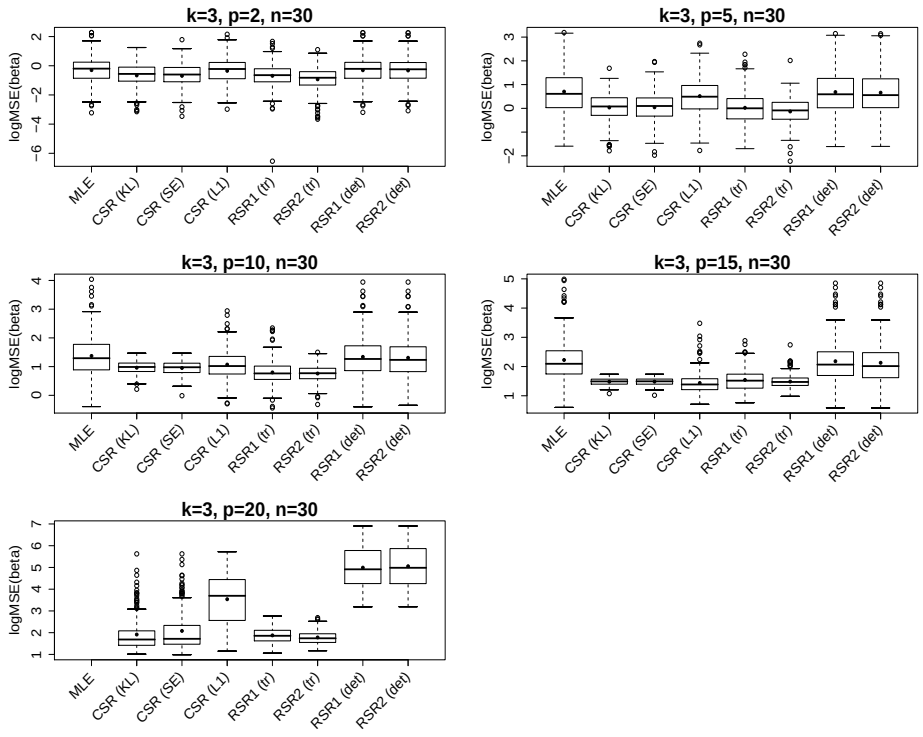

Figure 1: Illustration of the simulation study: Box plots for comparing different methods with different number of predictors for $n = 30$ in terms of MSE($\hat{\boldsymbol{\beta}}$).

chosen as zero which provides the estimates close to the MLE. For the case $p = 20$ and $n = 30$, MLE is not existing but the estimates do exist with regularization approaches. The use of determinant as the influence measure in RSR approach does not provide as good results as can be obtained with tr($\mathbf{H}$) or tr($\mathbf{H}^*$) as a leverage measure. The reason is that when determinant is used in RSR approach, maximum possible weights (almost 1) are assigned to the original responses and minimum (almost 0) weights are assigned to the shadow responses which results in estimates very close to the usual ML estimates. But using the L1 distance in CSR and using determinant as leverage measure in RSR can still provide the results close to MLE (actually little improved results) in the case where MLE does not exist. These results are also included in Table 1 to reflect this aspect of regularization approach. Table 1 also shows that with RSR, the use of matrix $\mathbf{H}^*$ given by Pregibon i.e., RSR2 approach gives more better results as compared to RSR1 which

uses the usual hat matrix. Although results of regularization techniques are always better than usual MLE but in particular, in terms of $\hat{\boldsymbol{\pi}}$, CSR approach taking edge over RSR with increasing sample size for larger number of covariates. In terms of MSE($\hat{\boldsymbol{\beta}}$), RSR2 (using trace) almost knocked out the other approaches. All approaches shown in Table 1 are compared in terms of box plots with respect to MSE($\hat{\boldsymbol{\beta}}$) in Fig. 1 for the most interesting case of small samples, i.e., $n = 30$. The bullets (solid circles) within the boxes are the mean of 200 values for which the box plots are drawn. The results of some simulation studies (not shown here) with more than three response categories also showed better performance of regularization techniques than usual MLE.

5 Estimation of standard errors

The score function of weighted log-likelihood function given in (5) can be written as

$$\begin{aligned} s_{\text{weighted}}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{D}_i(\boldsymbol{\beta}) \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta}) [\tilde{\mathbf{y}}_i - h(\boldsymbol{\eta}_i)] \\ &= \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{D}_i(\boldsymbol{\beta}) \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta}) [\mathbf{y}_i - h(\boldsymbol{\eta}_i) + \mathbf{y}_i^*] \end{aligned} \tag{17}$$

with $\mathbf{y}_i^* = \tilde{\mathbf{y}}_i - \mathbf{y}_i$. The first order approximation yields

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \approx \left(\frac{-\partial s_{\text{weighted}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \right)^{-1} s_{\text{weighted}}(\boldsymbol{\beta}).$$

From (17) the weighted score function can be written as

$$s_{\text{weighted}}(\boldsymbol{\beta}) = s(\boldsymbol{\beta}) + s_w(\boldsymbol{\beta}),$$

where

$$
\begin{aligned}
s(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{D}_i(\boldsymbol{\beta}) \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta})[\mathbf{y}_i - h(\boldsymbol{\eta}_i)] \quad \text{and} \\
s_w(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{D}_i(\boldsymbol{\beta}) \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta}) \mathbf{y}_i^*
\end{aligned}
$$

The derivatives needed here are

$$
-\frac{\partial s}{\partial \boldsymbol{\beta}^T} = F + \sum_{i=1}^{n} X_i X_i^T \frac{\partial^2 h}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} \{\mathbf{y}_i - \boldsymbol{\pi}_i\},
$$

where F is the weighted Fisher matrix,

$$
F = \sum_{i=1}^{n} X_i X_i^T \boldsymbol{\Sigma}_i^{-1} \left(\frac{\partial h}{\partial \boldsymbol{\eta}}\right) \left(\frac{\partial h}{\partial \boldsymbol{\eta}}\right)^T, \tag{18}
$$

and

$$
-\frac{\partial s_w}{\partial \boldsymbol{\beta}^T} = \sum_{i=1}^{n} X_i X_i^T \boldsymbol{\Sigma}_i^{-1} \frac{\partial^2 h}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} \mathbf{y}_i^*.
$$

The property of score function that $E(s(\boldsymbol{\beta})) = 0$ is not fulfilled for our weighted score function because $E(s_w(\boldsymbol{\beta})) \neq 0$. Using the basic definition of covariance, after some laborious derivation for the covariance of weighted score function we get

$$
\text{cov}(s_{\text{weighted}}(\boldsymbol{\beta})) = \sum_{i=1}^{n} \boldsymbol{\Gamma}_i^T \boldsymbol{\Sigma}_i \boldsymbol{\Gamma}_i,
$$

where $\boldsymbol{\Gamma}_i = \mathbf{A}_i^T \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_i^T \mathbf{X}_i$. $\mathbf{A}_i$ is a $q \times q$ matrix given as

$$
\mathbf{A}_i = \begin{bmatrix}
w_{i(0)} - w_{i(1)} & w_{i(k-1)} - w_{i(1)} & w_{i(k-2)} - w_{i(1)} & \cdots & w_{i(2)} - w_{i(1)} \\
w_{i(1)} - w_{i(2)} & w_{i(0)} - w_{i(2)} & w_{i(k-1)} - w_{i(2)} & \cdots & w_{i(3)} - w_{i(2)} \\
w_{i(2)} - w_{i(3)} & w_{i(1)} - w_{i(3)} & w_{i(0)} - w_{i(3)} & \cdots & w_{i(4)} - w_{i(3)} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
w_{i(k-2)} - w_{i(k-1)} & w_{i(k-3)} - w_{i(k-1)} & w_{i(k-4)} - w_{i(k-1)} & \cdots & w_{i(0)} - w_{i(k-1)}
\end{bmatrix},
$$

where $w_{i(s)}$ is the weight corresponding to the ith observation in the original data ($s = 0$) or sth ($s = 1, \ldots, q$) shadow data sets. The term $\partial s_w / \partial \boldsymbol{\beta}^T$ in the expression for $\text{cov}(s_{\text{weighted}}(\boldsymbol{\beta}))$ may be neglected asymptotically as $w_{i(s)} \to 0, \forall i, s = 1, \ldots, q$ with increasing sample size and one obtains as approximation the sandwich matrix

$$\text{cov}(\hat{\boldsymbol{\beta}}) = F(\boldsymbol{\beta})^{-1} \text{cov}(s_{\text{weighted}}(\boldsymbol{\beta})) F(\boldsymbol{\beta})^{-1}. \tag{19}$$

With $w_{i(s)} = 0$ for $\forall i, s$, the expression for covariance coincides with that of usual MLE. The accuracy of approximation of standard errors for such type of estimates has been investigated by Tutz and Leitenstorfer (2006) for binary responses.

6 Application

In this section we are using two real data sets from the medical field to compare our regularized estimates with the usual MLE.

Knee Injuries Data:

The first data being considered in this section was used by Fahrmeir et al. (1999). The data is aobut a clinical study focusing on the healing of sports related knee injuries, where 140 patients took part in the study but 13 patients with missing values were eliminated and 127 patients were used in the analysis. The patients visited the physician before treatment (baseline) and $3, 7$ and 10 days after the treatment. By random design one of the two therapies were chosen. In the treatment group, an anti-inflammatory spray was used while in the placebo group a spray without active ingredients was used. In each visit of the patient, the severity of injuries and the healing process were assessed by different indicators. The variable of primary interest was "pain from pressure". The pain Y occurring during the movement was assessed on a five point scale ranging from 1 (no pain) to 5 (severe pain). In addition to the treatment (1 : treatment; 0 : placebo), the covariates sex (1 : male; 0 : female) and age are considered. As the covariates assume the same values for the responses at different time points, we are using the data of 127 patients for the last time point i.e., responses after 10 days treatment.

Retinopathy Data:

The second data set that we are using for the computation of regularized estimates, is taken from Bender and Grouven (1998). In a 6-year follow up study on diabetes and retinopathy, 613 diabetic patients were reported by Bender and Grouven (1998). The objective was to investigate how the retinopathy status is associated with the risk factors. The considered risk factor is a binary variable "smoking (SM)" 1 : if the patients smoked during the study period; 0 : otherwise) adjusted for the known risk factors "diabetes duration (DIAB)" (measured in years), "glycosylated hemoglobin (GH)" (measured in percent), and "diastolic blood pressure (BP)" (measured in mmHg). The "retinopathy status" is a response variable with three response categories (1 : no retinopathy; 2 : non-proliferative retinopathy; 3 : advanced retinopathy or blind).

Table 2: Estimates and standard errors for "Knee Injuries" data

Method of Estimation	Intercept 1	Intercept 2	Intercept 3	Intercept 4	Therapy	Age	Sex
MLE	-0.9861	0.1988	1.1538	3.1409	-0.9438	0.0159	-0.0499
	(0.6290)	(0.6276)	(0.6399)	(0.7307)	(0.3355)	(0.0170)	(0.3731)
CSR, CV(KL)[a]	-1.0147	0.1498	1.0773	2.7864	-0.8916	0.0150	-0.0388
	(0.5857)	(0.5794)	(0.5870)	(0.6553)	(0.3160)	(0.0160)	(0.3384)
CSR, CV(SE)[b]	-1.0470	0.1011	0.9967	2.6388	-0.8655	0.0147	-0.0276
	(0.5735)	(0.5666)	(0.5731)	(0.6316)	(0.3088)	(0.0156)	(0.3314)
RSR1, (using trace)	-1.0120	0.1584	1.0971	2.9214	-0.8888	0.0143	-0.0211
	(0.5631)	(0.5569)	(0.5647)	(0.6420)	(0.3068)	(0.0153)	(0.3256)
RSR2, (using trace)	-1.0257	0.1478	1.0848	2.8884	-0.8666	0.0137	-0.0110
	(0.5630)	(0.5565)	(0.5642)	(0.6392)	(0.3063)	(0.0153)	(0.3254)

[a] Results are based on optimal values of tuning parameters $\boldsymbol{\alpha}^T = (0.02222222, 0.02187227, 0.02222222, 0.01049869)$.
[b] Results are based on optimal values of tuning parameters $\boldsymbol{\alpha}^T = (0.02222222, 0.02187227, 0.02222222, 0.04724409)$.

We are using the proportional odds model in both examples under the assumption that proportional odds assumption is fulfilled. The results for the parameter estimates and their standard error (within brackets) for usual MLE and the regularization techniques are presented in Table 2 and 3 for the "knee injury" and "retinopathy" data sets respectively. For CSR approach, the optimal values of α's are decided on the basis of leave-one-out

Table 3: Estimates and standard errors for "Retinopathy" data

Method of Estimation	Intercept 1	Intercept 2	SM	DIAB	GH	BP
MLE	12.3025	13.6733	−0.2549	−0.1398	−0.4597	−0.0724
	(1.2923)	(1.3197)	(0.1931)	(0.0139)	(0.0758)	(0.0136)
CSR [a]	11.1014	12.4013	−0.2189	−0.1274	−0.4156	−0.0656
	(1.1852)	(1.2077)	(0.1790)	(0.0124)	(0.0693)	(0.0126)
RSR1, (using trace)	12.1241	13.4944	−0.2486	−0.1382	−0.4554	−0.0712
	(1.2602)	(1.2867)	(0.1883)	(0.0132)	(0.0727)	(0.0132)
RSR2, (using trace)	11.9793	13.3457	−0.2428	−0.1371	−0.4499	−0.0703
	(1.2540)	(1.2802)	(0.1878)	(0.0131)	(0.0725)	(0.0132)

[a] Results are based on the same optimal vector $\boldsymbol{\alpha}^T = (0.02749942, 0.01246796)$ of tuning parameters for Kullback-Leibler discrepancy and the squared error loss.

cross-validation and are given as the table footnotes. The results of Tables 2 and 3 show that regularization not only shrinks the parameter estimates but also provides improved estimates of standard errors than the corresponding maximum likelihood estimates. For the RSR approach, we do not consider the determinant as a measure of influence to formulate the weighting scheme because it gives the results which are almost similar to usual MLE.

7 Conclusion

The proposed regularization technique aims at securing the existence of estimates by sharpening the responses to shrink them towards the underlying probabilities. Shadow responses with different weights are used as a tool to compute the estimates. The weighting scheme is designed in such a way that the weights associated with an original response and its corresponding shadow responses add to one. The use of q shadow responses for each observed y_{ij} with the same predictor vector $\mathbf{x}_i$ also resolves the problem of complete separation in the data. In case of CSR approach, selection of positive values for the tuning parameters $0 < \alpha_j \leq \bar{y}_{\cdot j}$ (for $j = 1, \ldots, q$) assures the existence of unique estimates. The use of L1-distance as a distance measure for the selection of optimal tuning parameters is not a good choice because it selects zero as the optimal value for most of the tuning

parameters. The squared error loss and Kullback-Leibler distance provide good results especially in terms of fit. The regularization technique based on the unique weights for each response i.e., RSR is more efficient than CSR not only in terms of processing time but also in terms of MSE's for the parameter estimates and the fit. For using the diagnostic measures it is recommended to use Pregibon's hat matrix rather than the usual hat matrix because it provides much better results. The proposed technique shrinks the parameter estimates without using any penalty term and does not require any special treatment for the categorical predictors which is required in case of penalized likelihood (see Zahid and Tutz (2010)). In sum, our regularized estimates are easy to compute, have better performance than MLE and have improved existence i.e., they also exist with no overlapping observation in the data and in the situations where MLE does not exist with large number of predictors relative to the sample size.

References

Agresti, A., 1999. Modelling ordered categorical data: Recent advances and future challenges. Statistics in Medicine 18, 2191–2207.

Ananth, C. V., Kleinbaum, D. G., 1997. Regression models for ordinal responses: A review of methods and applications. International Journal of Epidemiology 26, Number 6, 1323–1333.

Bender, R., Grouven, U., 1998. Using binary logistic regression models for ordinal data with non-proportional odds. Journal of Clinical Epidemiology 51, Issue 10, 809–816.

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

Fahrmeir, L., Gieger, C., Heumann, C., 1999. An application of isotonic longitudinal marginal regression to monitoring the healing process. Biometrics 55, 951–956.

Fahrmeir, L., Tutz, G., 2001. Multivariate Statistical Modelling based on Generalized Linear Models. second ed. Springer-Verlag NewYork, Inc.

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., 2010. Regularization paths for generalized linear models via coordinate descent. Jounal of Statistical Software 33, Issue 1.

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.

Lesaffre, E., Albert, A., 1989. Multiple-group losgistic regression diagnostics. Applied Statistics 38, 425–440.

McCullagh, P., 1980. Regression models for ordinal data. Journal of the Royal Statistical Society B 42, 109–142.

McCullagh, P., Nelder, J., 1989. Generalized Linear Models. second ed. Chapman & Hall, NewYork.

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.

Pregibon, D., 1981. Logistic regression diagnostics. The Annals of Statistics 9, 705–724.

Rousseeuw, P., Christmann, A., 2003. Robustness against separation and outliers in logistic regression. Computational Statistics and Data Analysis 43, 315–332.

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 the 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.

Tutz, G., Leitenstorfer, F., 2006. Response shrinkage estimators in binary regression. Computational Statistics & Data Analysis 50, Issue 10, 2878–2901.

Zahid, F. M., Tutz, G., 2010. Multinomial logit models with implicit variable selection. Technical Report No. 89. Institute of Statistics, Ludwig-Maximilians-University Munich, Germany.

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