Maximum Likelihood Estimation of Cumulative Link Models via the MM algorithm

Feiran Jiao[1], Kung-Sik Chan[1,*]

Abstract

The cumulative link model provides a flexible framework for regression analysis with ordinal response data. While the likelihood function of the cumulative link model generally admits a closed-form expression, maximum likelihood estimation as commonly implemented via iteratively re-weighted least squares and Fisher scoring often fails to converge due to the strong nonlinearity of the likelihood function. Empirical study shows that the failure rate of convergence can be substantial and generally increases with the extent of multicollinearity of the covariates. We develop minorization-maximization (MM) algorithms for maximum likelihood estimation of the cumulative link model, with link-function specific minorization function. The proposed approach inherits the strong advantage of the MM algorithm in ensuring that the likelihood always increases along the parametric iterates, thereby guaranteeing that the iterative MM scheme converges to the global maximizer, thanks to the concavity of the likelihood function for the cumulative link model, with commonly used links. Simulation studies show that the proposed approach generally yields more accurate estimates as compared to maximum likelihood estimates based on iteratively re-weighted least squares and Fisher scoring. The proposed MM algorithms can be sped up with monotonic over-relaxation. An empirical study shows that suitable monotone over-relaxation can significantly cut computation time.

Keywords: EM algorithm, monotone over-relaxation, ordinal response, proportional odds model

*Corresponding author

Email addresses: feiran-jiao@uiowa.edu (Feiran Jiao), kung-sik-chan@uiowa.edu (Kung-Sik Chan)

1. Introduction

Ordinal response variables abound in scientific and quantitative analyses, whose outcomes comprise a few categorical values that admit a natural ordering, so that their values are often represented by non-negative integers, for instance, pain score (0-10) or disease severity (0-3) in medical research. Ordinal variables play an important role where precise measurement is not always available. A popular model for regression analysis with ordinal response is the cumulative link model. The cumulative link model (McCullagh, 1980; Anderson and Philips, 1981; Agresti, 2002) assumes that the cumulative probability of the ordinal response is linked to some linear predictor with constant regression coefficients but a variable intercept term that preserves the ordering of the ordinal category; the link function is generally taken as the inverse function of some fixed cumulative distribution function, e.g., that of the standard normal distribution. The cumulative link model is known as the proportional odds model or the ordered logit model in the case of a logistic link (McCullagh, 1980; Greene and Hensher, 2010), the ordered probit model in the case of a probit link (Aitchison and Silvey, 1957) and the proportional hazards model in the case of a complementary log-log link.

A useful alternative representation of the cumulative link model stipulates that the ordinal response results from quantization of a latent continuous response variable driven by a linear regression model. This alternative representation renders it natural to consider link functions other than the logit and probit link, and opens up alternative approaches to drawing inference, e.g., model diagnostics with a cumulative link model. For instance, the error distribution in the latent regression model may be the standard Cauchy distribution or the standard Gumbel distribution. The standard Cauchy distribution is symmetric but, compared with the normal distribution, it has heavier tails. The standard Gumbel distribution (for modeling minima, which is a sub-model of the generalized extreme value distribution, with zero shape parameter, location at zero and unit scale) is an asymmetric distribution, which corresponds to using the complementary log-log link (Simonoff, 2013). As alluded to before, the cumulative link model with a complementary log-log link is also known as the *proportional hazards model*, which is commonly used in survival analysis. Interestingly, if the error distribution is the standard Gumbel distribution for modeling maxima, it corresponds to the log-log link function. The log-log link is then equivalent to the complementary log-log link with the the order of the categories reversed; thus, the

log-log link is generally omitted in the theoretical development below.

Other approaches for analyzing ordinal responses include the adjacent-categories logits and continuation-ratio logits (Agresti, 2002). We focus on the cumulative link model for it provides a familiar regression framework that is easy to interpret.

While the likelihood function of the cumulative link model has a closed-form solution for the aforementioned error distributions, its strong nonlinearity renders direct optimization of the likelihood to sometimes fail with substantial failure rate which tends to increase with the extent of collinearity in the covariates; see Section 3. To mitigate this problem, we propose specific minorization-maximization (MM) algorithms for maximum likelihood estimation of a cumulative link model for each of the preceding four error distributions. Our approach leverages on the latent variable representation of the cumulative link model, in which case, in principle, the expectation-maximization (EM) algorithm may be applied to derive an MM algorithm for maximum likelihood estimation. Specifically, a minorization function is naturally constructed in the E-step. Unfortunately, except for the probit link, the minorization function constructed via the EM algorithm does not admit a tractable solution. We solve the problem by deriving a link-specific minorization function to the minorization function constructed in the E-step, for Gumbel link and the Cauchy link. In the case of the complementary log-log link, we develop a local minorization function with a tractable maximizer within the domain of minorization; see Section 2 for details. An important property of the proposed MM algorithms is that the iterative estimation algorithms always increase the likelihood. We report some simulation results in Section 3, illustrating that the proposed algorithms generally result in more accurate estimates than the commonly used approach of maximum likelihood estimation via Fisher scoring and iteratively re-weighted least squares. In Section 4 we discuss a simple approach to accelerate the proposed MM methods to fit cumulative link models. All theoretical derivations are collected in Section 5. We conclude in Section 6.

2. Cumulative Link Model

The cumulative link model can be formulated by considering a continuous latent variable y^*. The observed ordinal response variable y, which takes values in $\{0, 1, 2, \cdots, q\}$, is observed in category p if and only if $\tau_{p-1} < y_i^* \leq \tau_p$, where $p = 0, \ldots, q$ and the fixed but unknown threshold parameters are

such that $-\infty = \tau_{-1} < \tau_0 < \tau_1 < \ldots < \tau_{q-1} < \tau_q = \infty$. Consider a linear model for the underlying latent variable y^*, for $i = 1, \ldots, n$,

$$y_i^* = \sum_{j=1}^{d} \beta_j X_{ij} + \varepsilon_i, \quad \varepsilon_i \sim F, \tag{1}$$

where the Xs are covariates, $\{\beta_j, j = 1, \ldots, d\}$ is a vector of unknown parameter to be estimated, the εs are independent and identically distributed, continuous random variables with known cumulative distribution F and f the corresponding probability density function (pdf); n is the sample size. Without loss of generality, all covariates are standardized. Consequently,

$$P(y_i = p) = P(\tau_{p-1} < y_i^* \leq \tau_p) = F\left(\tau_p - \sum_{j=1}^{d} \beta_j X_{ij}\right) - F\left(\tau_{p-1} - \sum_{j=1}^{d} \beta_j X_{ij}\right). \tag{2}$$

Naturally, $F(\cdot)$ and $f(\cdot)$ vary with the specific error distribution assumption. Let F^{-1} denote the inverse of F. Then,

$$F^{-1}\left\{P(y \leq p \mid \mathbf{X})\right\} = \tau_p - \sum_{j=1}^{d} \beta_j X_{ij}, \tag{3}$$

which links the cumulative probabilities to the linear predictors, thence the model is known as the cumulative link model.

Denote the vector of the observed responses by $\boldsymbol{y} = (y_1, \cdots, y_n)^\top$, and the vector of corresponding latent variables $\boldsymbol{y}^* = (y_1^*, \ldots, y_n^*)^\top$. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\tau}^\top)^\top$. Conditional on the covaraites, the log-likelihood for the cumulative link model is given by

$$\ell(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} \sum_{p=0}^{q} \mathbf{1}(y_i = p) \log\left\{F(\tau_p - \mathbf{x}_i^\top \boldsymbol{\beta}) - F(\tau_{p-1} - \mathbf{x}_i^\top \boldsymbol{\beta})\right\}, \tag{4}$$

where $\mathbf{1}(\cdot)$ is an indicator function of the enclosed event and we have suppressed X's from the notation $\ell(\boldsymbol{\theta}|\mathbf{y})$ and the like, for simplicity. Below, we sometimes write $\ell(\boldsymbol{\theta})$ for $\ell(\boldsymbol{\theta}|\mathbf{y})$. Burridge (1981) and Pratt (1981) showed that $\ell(\boldsymbol{\theta}|\mathbf{y})$ is a concave function of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau})$ for the cumulative link model with the logit, probit and complementary log-log links, hence the optimization problem can be solved by searching a local maximizer. The concavity

result does not extend to the cumulative link model with a Cauchy link. Maximum likelihood estimation with the cumulative link model is generally implemented by iteratively re-weighted least squares and Fisher scoring, as implemented in the `polr` function in the statistical computing software R (R Core Team, 2022; Venables and Ripley, 2002) As mentioned earlier, ML estimation via iteratively re-weighted least squares and Fisher scoring may fail, since the initial values may not be sufficiently close to the true value, see Section 3.

Below, we propose minorization-maximization (MM) algorithms (Hunter and Lange, 2004) to do maximum likelihood estimation of the cumulative link model, which is link specific for four commonly used links. An MM algorithm is an iterative method for obtaining maximum likelihood estimates, when the objective function (here the observed-data log-likelihood) is difficult to optimize directly. Each iteration of the MM algorithm comprises two steps, namely, the minorization step and the maximization step. The idea of an MM algorithm is to first create a surrogate function that minorizes the objective function and then maximize the surrogate function, which increases the objective function; the iterations are repeated until convergence. Following Hunter and Lange (2004), a function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)})$ is said to minorize $\ell(\boldsymbol{\theta})$ at the point $\boldsymbol{\theta}^{(m)}$, if

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)}) &\leq \ell(\boldsymbol{\theta}) \text{ for all } \boldsymbol{\theta}, \\ Q(\boldsymbol{\theta}^{(m)}|\boldsymbol{\theta}^{(m)}) &= \ell(\boldsymbol{\theta}^{(m)}). \end{aligned}$$

That is, the function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(m)})$ always lies below $\ell(\boldsymbol{\theta})$ while it coincides with $\ell(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^{(m)}$. The MM algorithm enjoys the property of increasing the objective function value along the iterates because

$$\ell(\boldsymbol{\theta}^{(m+1)}) \geq Q(\boldsymbol{\theta}^{(m+1)} \mid \boldsymbol{\theta}^{(m)}) \geq Q(\boldsymbol{\theta}^{(m)} \mid \boldsymbol{\theta}^{(m)}) = \ell(\boldsymbol{\theta}^{(m)}).$$

2.1. Cumulative logit model

The most commonly used approach for handling ordered categorical responses is the cumulative logit model, often known as the proportional odds model. This model specifies a logit link:

$$\text{logit}\left\{P(Y \leq p \mid \mathbf{X})\right\} = \log\left\{\frac{P(y \leq p \mid \mathbf{X})}{1 - P(y \leq p \mid \mathbf{X})}\right\} = \tau_p - \sum_{i=1}^{d} \beta_j X_j, \quad p = 0, 1, \ldots q. \tag{5}$$

It assumes that the cumulative logit lines are parallel as they share the same $\boldsymbol{\beta}$ but different intercepts τ_p per response value.

Anderson and Philips (1981) proposed to view this model using an underlying (latent) continuous response variable y^*. For the cumulative logit model, the error ε in (1) follows the standard logistic distribution. The shape of the logistic distribution is similar to that of the normal distribution but with heavier tails.

The celebrated expectation-maximization (EM) algorithm is a special case of the more general MM algorithm (Dempster et al., 1977; Zhou and Zhang, 2012). The E-step of the EM algorithm provides a general recipe for constructing a minorization function, which starts with the complete-data log-likelihood for the latent response y^*:

$$\begin{aligned} \ell^*(\boldsymbol{\beta}) = \ell^*(\boldsymbol{\beta}|\mathbf{y}^*) &= \sum_{i=1}^{n} \log f(y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} \left[-y_i^* + \mathbf{x}_i^\top \boldsymbol{\beta} - 2 \log \left\{ 1 + \exp(-y_i^* + \mathbf{x}_i^\top \boldsymbol{\beta}) \right\} \right], \quad (6) \end{aligned}$$

Then, a minorization function is obtained by taking the conditional expectation of the complete-data log-likelihood given the observed data and the current parametric iterate. Since the errors follow the logistic distribution, taking expectation of the complete-data log-likelihood does not yield a closed-form solution, unlike the case of normal error distribution. Hence, we shall develop a minoziation-maximization (MM) algorithm by further minorizing the function $E\{\ell^*(\boldsymbol{\beta})|y, \boldsymbol{\theta}^{(k)}\}$.

Let $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\tau}^{(0)}$ be the initial values for the unknown parameters $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$. For $k = 0, 1, 2, \ldots$, the MM algorithm proceeds as follows:

- Minorization step: Compute a minorization function $Q(\boldsymbol{\beta}|\boldsymbol{\theta}^{(k)})$ of the observed-data log-likelihood $\ell(\boldsymbol{\beta}, \boldsymbol{\tau}^{(k)}|\mathbf{y})$.
- Maximization step: Set $\boldsymbol{\beta}^{(k+1)}$ to be the argument maximizing $Q(\cdot|\boldsymbol{\theta}^{(k)})$ and set $\boldsymbol{\tau}^{(k+1)}$ to be the argument maximizing $\ell(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{\tau}|\mathbf{y})$.

The iteration alternates between the two M-steps, until some convergence criterion is met.

For the minorization step, the minorization function is obtained by applying a second-order Taylor expansion to the complete-data log-likelihood (6),

and taking conditional expectation given the observed data and the current parameter estimates. By the mean value theorem, there exists a $\widetilde{\boldsymbol{\beta}}$ between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^{(k)}$ such that

$$\begin{aligned} E\left\{\ell^*(\boldsymbol{\beta})|\mathbf{y},\boldsymbol{\theta}^{(k)}\right\} &= E\left\{\ell^*(\boldsymbol{\beta}^{(k)})|\mathbf{y},\boldsymbol{\theta}^{(k)})\right\} + (\boldsymbol{\beta}-\boldsymbol{\beta}^{(k)})^\top E\left\{\left.\frac{\partial\ell^*(\boldsymbol{\beta}^{(k)})}{\partial\boldsymbol{\beta}}\right|\mathbf{y}^*,\boldsymbol{\theta}^{(k)}\right\} + \\ &\quad \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^{(k)})^\top E\left\{\left.\frac{\partial^2\ell^*(\widetilde{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\right|\mathbf{y}^*,\boldsymbol{\theta}^{(k)}\right\}(\boldsymbol{\beta}-\boldsymbol{\beta}^{(k)}). \end{aligned}$$

Writing p_i for $1/\{1+\exp(-y_i^*+\mathbf{x}_i^\top\boldsymbol{\beta})\}$, the first derivative and Hessian matrix of $\ell^*(\boldsymbol{\beta})$ are given as belows:

$$\frac{\partial\ell^*(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \sum_{i=1}^n (2p_i-1)\,\mathbf{x}_i$$

$$\begin{aligned} \frac{\partial^2\ell^*(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} &= -\sum_{i=1}^n 2p_i(1-p_1)\mathbf{x}_i\mathbf{x}_i^\top \\ &\geq -\frac{1}{2}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top, \end{aligned}$$

because $p_i(1-p_i)$ is bounded above by $\frac{1}{4}$. Thus,

$$\begin{aligned} E\left\{\ell^*(\boldsymbol{\beta})|\mathbf{y},\boldsymbol{\theta}^{(k)}\right\} &\geq E\left\{\ell^*(\boldsymbol{\beta}^{(k)})|\mathbf{y},\boldsymbol{\theta}^{(k)}\right\} + (\boldsymbol{\beta}-\boldsymbol{\beta}^{(k)})^\top E\left\{\left.\frac{\partial\ell^*(\boldsymbol{\beta}^{(k)})}{\partial\boldsymbol{\beta}}\right|\mathbf{y},\boldsymbol{\theta}^{(k)}\right\} - \\ &\quad \frac{1}{4}(\boldsymbol{\beta}-\boldsymbol{\beta}^{(k)})^\top\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^{(k)}). \end{aligned} \tag{7}$$

The right hand side of the preceding inequality is a minorization function of $\ell(\boldsymbol{\beta},\boldsymbol{\tau}^{(k)}|\mathbf{y})$ around $\boldsymbol{\beta}^{(k)}$, which is denoted as $Q(\boldsymbol{\beta}|\boldsymbol{\theta}^{(k)})$.

The maximization step requires finding $\boldsymbol{\beta}^{(k+1)}$ and $\boldsymbol{\tau}^{(k+1)}$ by maximizing $Q(\boldsymbol{\beta}|\boldsymbol{\theta}^{(k)})$ and the observed-data log-likelihood $\ell(\boldsymbol{\beta}^{(k+1)},\boldsymbol{\tau}|\mathbf{y})$ respectively. Taking partial derivative of $Q(\boldsymbol{\beta}|\boldsymbol{\theta}^{(k)})$, the quadratic function is maximized

at

$$\begin{aligned}\boldsymbol{\beta}^{(k+1)} &= \boldsymbol{\beta}^{(k)} + 2\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1} E\left\{\frac{\partial \ell^*(\boldsymbol{\beta}^{(k)})}{\partial \boldsymbol{\beta}}\middle|\mathbf{y},\boldsymbol{\theta}^{(k)}\right\} \\ &= \boldsymbol{\beta}^{(k)} + 2\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1} \sum_{i=1}^{n}\left[E\left\{\frac{2}{1+\exp(-y_i^* + \mathbf{x}_i^\top\boldsymbol{\beta})}\middle|\mathbf{y},\boldsymbol{\theta}^{(k)}\right\} - 1\right]\mathbf{x}_i,\end{aligned} \tag{8}$$

where

$$\begin{aligned}& E\left[\left\{1+\exp(-y_i^* + \mathbf{x}_i^\top\boldsymbol{\beta})\right\}^{-1}\ |\mathbf{y},\boldsymbol{\theta}^{(k)}\right] \\ = \ & \frac{1}{2}\left\{\frac{2\exp(\tau_{y_i-1}^{(k)} - \mathbf{x}_i^\top\boldsymbol{\beta}^{(k)})+1}{(\exp(\tau_{y_i-1}^{(k)} - \mathbf{x}_i^\top\boldsymbol{\beta}^{(k)})+1)^2} - \frac{2\exp(\tau_{y_i}^{(k)} - \mathbf{x}_i^\top\boldsymbol{\beta}^{(k)})+1}{(\exp(\tau_{y_i}^{(k)} - \mathbf{x}_i^\top\boldsymbol{\beta}^{(k)})+1)^2}\right\} \times \\ & \left\{F(\tau_{y_i}^{(k)} - \mathbf{x}_i^\top\boldsymbol{\beta}^{(k)}) - F(\tau_{y_i-1}^{(k)} - \mathbf{x}_i^\top\boldsymbol{\beta}^{(k)})\right\}^{-1};\end{aligned}$$

here F is the standard logistic cdf and we use the convention that $\exp(-\infty) = F(-\infty) = 0$ and $\exp(\infty) = F(\infty) = 1$.

Because the threshold parameters $\boldsymbol{\tau}$s do not play a role in the complete-data log-likelihood (6), they are, instead, updated by maximizing the observed-data log-likelihood (4) with $\boldsymbol{\beta}$ fixed using the Fisher scoring algorithm. Pratt (1981) showed that the observed-data log-likelihood of the cumulative logit model is concave. Therefore, a local maximizer is necessarily the global maximizer. We apply the Fisher scoring algorithm to obtain the maximizer of $\boldsymbol{\tau}$. The Fisher information matrix of the cumulative logit model is derived in Appendix 5, which is shown to be positive definite, under the mild assumption that the covariates are not collinear. Since $\boldsymbol{\tau}$ follows the constraint that $\tau_{-1} = -\infty < \tau_0 \le \tau_1 \le \ldots < \tau_{q-1} < \tau_q = \infty$, we reparametrizate $\boldsymbol{\tau}$ using the unconstrained $\boldsymbol{\delta}$ such that

$$\begin{pmatrix} \delta_0 = \tau_0 \\ \delta_1 = \log(\tau_1 - \tau_0) \\ \vdots \\ \delta_{q-1} = \log(\tau_{q-1} - \tau_{q-2}) \end{pmatrix}.$$

Applying the Fisher scoring algorithm, the maximizer of $\boldsymbol{\delta}$ is given by

$$\boldsymbol{\delta}^{(k+1)} = \boldsymbol{\delta}^{(k)} + \mathcal{I}^{-1}(\boldsymbol{\delta}^{(k)})S(\boldsymbol{\delta}^{(k)}),$$

where $S(\boldsymbol{\delta}^{(k)})$ is the score function for $\boldsymbol{\delta}^{(k)}$ and $\mathcal{I}(\delta^{(k)})$ is the Fisher information matrix evaluated at $\boldsymbol{\delta}^{(k)}$; see Section 5 for the expression of the Fisher information matrix and the proof of its positive definiteness. Note that $\mathcal{I}(\boldsymbol{\delta}) = \mathbf{J}^\top \mathcal{I}(\boldsymbol{\tau})\mathbf{J}$, where the (i,j)th element of the $q \times q$ Jacobian matrix $\mathbf{J}$ is defined by $J_{i,j} = \frac{\partial \tau_i}{\partial \delta_j}$ and $S(\boldsymbol{\delta}) = \mathbf{J}^\top S(\boldsymbol{\tau})$. Hence, the maximizer of $\boldsymbol{\tau}$ can be obtained by iteratively applying the following equation

$$\begin{aligned} \boldsymbol{\delta}^{(k+1)} &= \boldsymbol{\delta}^{(k)} + \{\mathbf{J}^\top \mathcal{I}(\boldsymbol{\tau})\mathbf{J}\}^{-1}\mathbf{J}^\top S(\boldsymbol{\tau}) \\ &= \boldsymbol{\delta}^{(k)} + \mathbf{J}^{-1}\mathcal{I}^{-1}(\boldsymbol{\tau})S(\boldsymbol{\tau}). \end{aligned} \tag{9}$$

2.2. Cumulative probit model

The cumulative probit model employs the probit link, which assumes a conditional normal distribution for the latent response $\mathbf{y}^*$. The observed-data log-likelihood of the cumulative probit model is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\tau}|\mathbf{y}) = \sum_{i=1}^{n}\sum_{p=0}^{q} \mathbf{1}(y_i = p) \log\left\{\Phi\left(\tau_p - \mathbf{x}_i^\top \boldsymbol{\beta}\right) - \Phi\left(\tau_{p-1} - \mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\}, \tag{10}$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution and $\mathbf{1}(\cdot)$ is the indicator function of the enclosed event. The complete-data log-likelihood has the following form:

$$\ell^*(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta}) = -\frac{1}{2}\sum_{i=1}^{n}\left(y_i^* - x_i^\top \boldsymbol{\beta}\right)^2. \tag{11}$$

Here, the E-step results in a tractable minorization function:

$$\begin{aligned} Q(\boldsymbol{\beta}|\boldsymbol{\theta}^{(k)}) &= -\frac{1}{2}\sum_{i=1}^{n} E[(y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta})^2 | y_i, \boldsymbol{\theta}^{(k)}] \\ &= -\frac{1}{2}\sum_{i=1}^{n}\left\{E(y_i^*|y_i, \boldsymbol{\theta}^{(k)}) - \mathbf{x}_i^\top \boldsymbol{\beta}\right\}^2 - \frac{1}{2}\sum_{i=1}^{n} \mathrm{var}(y_i^*|y_i, \boldsymbol{\theta}^{(k)}) \end{aligned}$$

Since $\mathrm{var}(y_i^*|y_i, \boldsymbol{\theta}^{(k)})$ is a constant, we shall drop it so that

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = -\frac{1}{2}\sum_{i=1}^{n}\left\{E(y_i^*|y_i, \boldsymbol{\theta}^{(k)}) - \mathbf{x}_i^\top \boldsymbol{\beta}\right\}^2.$$

It is readily checked in Barr and Sherrill (1999), upon letting $y_i = p \in \{0, 1, \ldots, q\}$ and $\phi(\cdot)$ be the standard normal pdf, then

$$E(y_i^*|\boldsymbol{\theta}^{(k)}, y_i) = \mathbf{x}_i^\top \boldsymbol{\beta}^{(k)} + \frac{\phi\left(\tau_{p-1}^{(k)} - x_i^\top \boldsymbol{\beta}^{(k)}\right) - \phi\left(\tau_p^{(k)} - x_i^\top \boldsymbol{\beta}^{(k)}\right)}{\Phi\left(\tau_p^{(k)} - x_i^\top \boldsymbol{\beta}^{(k)}\right) - \Phi\left(\tau_{p-1}^{(k)} - x_i^\top \boldsymbol{\beta}^{(k)}\right)}, \quad (12)$$

with the convention that $\phi(-\infty) = \phi(\infty) = 0$, $\Phi(-\infty) = 0$ and $\Phi(\infty) = 1$.

The M-step requires finding $\boldsymbol{\beta}^{(k+1)} = \arg\max_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}|\boldsymbol{\theta}^{(k)})$ which equals

$$\boldsymbol{\beta}^{(k+1)} = (X^\top X)^{-1} X^\top E(y^*|y, \boldsymbol{\theta}^{(k)}). \quad (13)$$

The maximizer of the threshold parameter $\boldsymbol{\tau}$ is similarly obtained using Eqn. (9), but under the normal error distribution assumption.

2.3. Cumulative link model with Cauchy latent variable

In the case of the Cauchy link, i.e., the link function is the inverse of the standard Cauchy cdf, the complete-data log-likelihood has the following form:

$$\ell^*(\boldsymbol{\beta}) \;=\; -\sum_{i=1}^{n} \left[\log \pi + \log\{1 + (y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta})^2\}\right]. \quad (14)$$

We derive below an MM algorithm using an approach similar to the case of the cumulative logit model by first applying a second-order Taylor expansion to the complete-data log-likelihood, then taking conditional expectation given the observed data and a current estimate, and finally bounding the second partial derivative by a known matrix. The first derivative of $\ell^*(\boldsymbol{\beta})$ takes the form

$$\frac{\partial \ell^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \;=\; \sum_{i=1}^{n} \frac{2(y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta})}{1 + (y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta})^2} \mathbf{x}_i.$$

The second derivative of $\ell^*(\boldsymbol{\beta})$ can be lower-bounded as follows:

$$\begin{aligned} \frac{\partial^2 \ell^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} &= \sum_{i=1}^{n} \frac{-2 + 2(y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{\{1 + (y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta})^2\}^2} \mathbf{x}_i \mathbf{x}_i^\top \\ &\geq -2 \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top. \end{aligned}$$

Altogether, we have the following minorizing function:

$$\begin{aligned} Q(\boldsymbol{\beta}|\boldsymbol{\theta}^{(k)}) &= E\left\{\ell^*(\boldsymbol{\beta}^{(k)})|\mathbf{y},\boldsymbol{\theta}^{(k)}\right\} + (\boldsymbol{\beta}-\boldsymbol{\beta}^{(k)})^\top E\left\{\left.\frac{\partial\ell^*(\boldsymbol{\beta}^{(k)})}{\partial\boldsymbol{\beta}}\right|\mathbf{y},\boldsymbol{\theta}^{(k)}\right\} - \\ &\quad (\boldsymbol{\beta}-\boldsymbol{\beta}^{(k)})^\top \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^{(k)}) \end{aligned} \tag{15}$$

The maximization step proceeds to maximizing this quadratic Q function, which is maximized at

$$\begin{aligned} \boldsymbol{\beta}^{(k+1)} &= \boldsymbol{\beta}^{(k)} + \frac{1}{2}\left(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\right)^{-1} E\left\{\left.\frac{\partial\ell^*(\boldsymbol{\beta}^{(k)})}{\partial\boldsymbol{\beta}}\right|\mathbf{y},\boldsymbol{\theta}^{(k)}\right\} \\ &= \boldsymbol{\beta}^{(k)} + \frac{1}{2}\left(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\right)^{-1} \sum_{i=1}^n \left\{E\left(\left.\frac{2(y_i^*-\mathbf{x}_i^\top\boldsymbol{\beta})}{1+(y_i^*-\mathbf{x}_i^\top\boldsymbol{\beta})^2}\right|\mathbf{y},\boldsymbol{\theta}^{(k)}\right)\mathbf{x}_i\right\}, \end{aligned}$$

where

$$\begin{aligned} & E\left(\left.\frac{2(y_i^*-\mathbf{x}_i^\top\boldsymbol{\beta})}{1+(y_i^*-\mathbf{x}_i^\top\boldsymbol{\beta})^2}\right|\mathbf{y},\boldsymbol{\theta}^{(k)}\right) \\ = \; & \frac{1}{\pi}\left(\frac{1}{(\tau_{y_i-1}^{(k)}-\mathbf{x}_i^\top\boldsymbol{\beta}^{(k)})^2+1} - \frac{1}{(\tau_{y_i}^{(k)}-\mathbf{x}_i^\top\boldsymbol{\beta}^{(k)})^2+1}\right)\times \\ & \left\{F(\tau_{y_i}^{(k)}-\mathbf{x}_i^\top\boldsymbol{\beta}^{(k)}) - F(\tau_{y_i-1}^{(k)}-\mathbf{x}_i^\top\boldsymbol{\beta}^{(k)})\right\}^{-1}; \end{aligned}$$

here F is the standard Cauchy cdf and we use the convention that $1/\infty = 0$, $F(-\infty) = 0$ and $F(\infty) = 0$.

The maximizer of the threshold parameter $\boldsymbol{\tau}$ is similarly obtained using Eqn. (9), but under the Cauchy error distribution assumption.

2.4. Cumulative link model with complementary log-log link

Another choice of link in Equation (3) is the complementary log-log transformation

$$\log[-\log\{1-P(y\le p\,|\,\mathbf{X})\}] = \tau_p - \sum_{j=1}^d \beta_j X_{ij},$$

in which case the latent variable $\mathbf{y}^*$ follows a standard Gumbel distribution (minimum), which belongs to the family of the generalized extreme value distribution (Embrechts et al., 2013). Normal, logistic and Cauchy distributions

are all unimodal symmetric distributions, and their cumulative distribution functions approach zero at the same rate they approach one. But the Gumbel distribution for minima has an asymmetric density which approaches zero slowly and one quickly. The probability density function of the Gumbel distribution with zero mean and unit scale is

$$f(x) = \exp\{x - \exp(x)\}, \text{ for } x \in (-\infty, \infty).$$

The corresponding cumulative distribution function is

$$F(x) = 1 - \exp\{-\exp(x)\}, \text{ for } x \in (-\infty, \infty). \tag{16}$$

In the case of a Gumbel link, the complete-data log-likelihood has the following form:

$$\ell^*(\boldsymbol{\beta}) \;=\; \sum_{i=1}^{n} y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta} - \exp\left(y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta}\right). \tag{17}$$

The first derivative and the Hessian matrix of $\ell^*(\boldsymbol{\beta})$ are given by

$$\frac{\partial \ell^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left\{\exp(y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta}) - 1\right\} \mathbf{x}_i$$

$$\frac{\partial^2 \ell^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = -\sum_{i=1}^{n} \exp(y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta}) \mathbf{x}_i \mathbf{x}_i^\top.$$

Unfortunately, there is no global lower bound for $\frac{\partial^2 \ell^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}$. So we develop a local lower bound to enable a local minorization function, with the local minorization function suitably chosen such that it is maximized within the minorization domain. Let $K = \max\{\|\mathbf{x}_j\|_\infty, 1 \le j \le n\}$ where $\|\cdot\|_\infty$ is the max norm. It can be shown that for $\boldsymbol{\beta} \in \mathcal{N}(\Delta) = \{|\boldsymbol{\beta} - \boldsymbol{\beta}^{(k)}| \le \Delta\}$,

$$\frac{\partial^2 \ell^*(\widetilde{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \ge -\sum_{i=1}^{n} \exp(y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta}^{(k)}) \exp(\mathbf{K}\Delta) \mathbf{x}_i \mathbf{x}_i^\top. \tag{18}$$

Then the local minorizing function is

$$\begin{aligned} Q(\boldsymbol{\beta}|\boldsymbol{\theta}^{(k)}) \;=\; & E\{\ell^*(\boldsymbol{\beta}^{(k)} \,|\mathbf{y}, \boldsymbol{\theta}^{(k)}\} + (\boldsymbol{\beta} - \boldsymbol{\beta}^{(k)})^\top E\left\{ \frac{\partial \ell^*(\boldsymbol{\beta}^{(k)})}{\partial \boldsymbol{\beta}} \middle| \mathbf{y}, \boldsymbol{\theta}^{(k)} \right\} - \\ & \frac{1}{2} \exp(\mathbf{K}\Delta)(\boldsymbol{\beta} - \boldsymbol{\beta}^{(k)})^\top \sum_{i=1}^{n} E\{\exp(y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta}^{(k)}) \,|\mathbf{y}, \boldsymbol{\theta}^{(k)}\} \mathbf{x}_i \mathbf{x}_i^\top (\boldsymbol{\beta} - \boldsymbol{\beta}^{(k)}). \end{aligned}$$

Write $\left[\sum_{i=1}^{n} E\{\exp(y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta}^{(k)}) \,|\mathbf{y}, \boldsymbol{\theta}^{(k)}\} \mathbf{x}_i \mathbf{x}_i^\top\right]^{-1} E\left\{\frac{\partial \ell^*(\boldsymbol{\beta}^{(k)})}{\partial \boldsymbol{\beta}} \,|\mathbf{y}, \boldsymbol{\theta}^{(k)}\right\}$ as $H^{(k)}$. Set

$$\Delta = \frac{\sqrt{1 + 4\mathbf{K}|H^{(k)}|} - 1}{2\mathbf{K}}.$$

Then it can be checked that the Q function is maximized at

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \exp(-\mathbf{K}\Delta) H^{(k)} \in \mathcal{N}(\Delta).$$

Note that

$$\begin{aligned} & E\{\exp(y_i^* - \mathbf{x}_i^\top \boldsymbol{\beta}^{(k)}) \,|\mathbf{y}, \boldsymbol{\theta}^{(k)}\} \\ = \; & \Big[\exp\{-\exp(\tau_{y_i-1}^{(k)} - \mathbf{x}_i^\top \boldsymbol{\beta})\} \times \{\exp(\tau_{y_i-1}^{(k)} - \mathbf{x}_i^\top \boldsymbol{\beta}) + 1\} - \\ & \quad \exp\{-\exp(\tau_{y_i}^{(k)} - \mathbf{x}_i^\top \boldsymbol{\beta})\} \times \{\exp(\tau_{y_i}^{(k)} - \mathbf{x}_i^\top \boldsymbol{\beta}) + 1\}\Big] \{F(\tau_{y_i}^{(k)} - \mathbf{x}_i^\top \boldsymbol{\beta}) - F(\tau_{y_i-1}^{(k)} - \mathbf{x}_i^\top \boldsymbol{\beta})\}^{-1}, \end{aligned}$$

where $\exp(-\infty) = 0$ and $\exp(\infty) = 1$.

The maximizer of the threshold parameter $\boldsymbol{\tau}$ is similarly obtained using Eqn (9), but under the standard Gumbel error distribution assumption.

3. Simulations

We report some simulation results showing that fitting a cumulative link model by the proposed MM algorithm generally outperforms the standard, iterative re-weighted least squares as implemented by the R function `polr`, in terms of estimation accuracy and empirical coverage of confidence intervals, especially in the presence of strong multicollinearity among the explanatory variables. Realizations from the cumulative link model can be simulated in two stages. First, the latent responses are simulated according to the linear model, $\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon}$ is a random sample from one of the four error distributions, namely, normal, logistic, Cauchy or Gumbel distribution. Second, the thresholds, $\tau_{-1} = -\infty < \tau_0 < \tau_1 < \ldots < \tau_{q-1} < \tau_q = \infty$, are set as certain quantiles of the unconditional marginal distribution of y^*, so as to make a pre-determined unconditional probability distribution for the response variable. Finally, y_i is set to be j if and only if $\tau_{j-1} < y_i^* \leq \tau_j$. In the experiment, the ordinal response variable takes value from five categories: 0, 1, 2, 3, and 4, and the thresholds are chosen to make the five categories equally frequent. The number of covariates $d = 5$, with the true regression coefficients being $\boldsymbol{\beta}_0 = (1, 3, -2, 5, 0.5)^\top$. Multicollinearity of

various degrees is introduced by drawing the covariates from a d-dimensional Gaussian distribution $N(\mathbf{0}, \mathbf{\Sigma})$, where the (h, l)th entry of $\mathbf{\Sigma} = \rho^{|h-l|}$. The marginal distribution of y^* is determined by which distribution we simulate data from. For normal distribution, y^* is normally distributed with zero mean and variance $1 + \boldsymbol{\beta}_0^\top \Sigma \boldsymbol{\beta}_0$. For other error distributions, it is difficult to obtain the marginal distribution of y^*. However, the mean and variance can be calculated for the logistic and extreme value errors. For the logistic error distribution, the marginal distribution of y^* has zero mean and variance $\frac{\pi^2}{3} + \boldsymbol{\beta}_0^\top \Sigma \boldsymbol{\beta}_0$. For the Gumbel distribution, y^* has mean γ and variance $\frac{\pi^2}{6} + \boldsymbol{\beta}_0^\top \Sigma \boldsymbol{\beta}_0$, where γ is the Euler's constant. For Cauchy errors, the mean and variance of y^* are undefined. The thresholds are determined so that the five ordinal outcomes are equally likely.

We set the initial values for the estimation schemes as follows. We can initialize $\boldsymbol{\beta}^{(0)}$ as the ordinary least square estimate obtained by regressing the ordinal response on the covariates, i.e., treating the ordinal response as continuous response. Initial values of $\boldsymbol{\tau}^{(0)}$ can be set as the maximum likelihood estimates for the case that none of the covariates are relevant, i.e., $\boldsymbol{\beta} = 0$, in which case the maximum likelihood estimator of $F(\hat{\tau}_j)$ is the fraction of observed response values not greater than j and F depends on assumptions of the error distribution. We use the same sets of initial values for $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ for the proposed MM algorithm and the `polr` function to fit the cumulative link models under different error assumptions.

As alluded to earlier, maximum likelihood estimation of the cumulative link models via direct optimization as implemented by the commonly used R function `polr` may fail to converge, due to the strong nonlinearity of the objective function. Table 1 lists the frequency when `polr` fails to converge until the accumulation of 400 successful fits, with data simulated from the cumulative link model with the probit link and ρ ranging from 0 to 0.8 with increment of 0.2. With increasing correlation among the covariates, `polr` tends to fail more frequently. On the other hand, maximum likelihood estimation using the proposed MM algorithm, however, is superior to the `polr` in that it always converges, based on our stopping criterion.

[Table 1 about here.]

Below, we report further simulation results for the case of $\rho = 0$ and 0.8, with the four commonly used link functions. Each simulation is replicated 1000 times.

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

[Table 5 about here.]

[Table 6 about here.]

[Table 7 about here.]

[Table 8 about here.]

Define the model error as $\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2/n$, where n is the number of observations (100 here). Table 2 shows that the model error increases with multicollinearity, i.e, larger ρ. Model fits using the proposed MM algorithm yielded smaller model errors than those using the `polr` function, under all four error distributions. Tables 3 and 4 list the average estimators of $\boldsymbol{\beta}$ using the proposed MM algorithm and the R function `polr`. As ρ increases, the estimation bias increases. For probit, logit and Cauchy distribution, biases of the estimates based on the MM algorithm are relatively smaller than those based on the `polr` function. However, for the complementary log-log link, the `polr` estimates have smaller biases when $\rho = 0$. The threshold estimators for these two methods tended to perform similarly as shown in Tables 5 and 6. However, for the complementary log-log link, the proposed MM algorithm yielded much closer threshold estimates to their true values. Note that the true thresholds change somewhat with ρ. We also observe that variability in the estimates for the proposed MM algorithm tended to be smaller, as compared with the `polr` estimates. Empirical coverage rates for both methods (Tables 7 and 8) are similar and both close to the nominal 95%, except that for the complementary log-log link, the `polr` estimates have much lower coverage rates than 95%. Overall, the simulation results confirm the robust performance of the proposed MM estimation method.

4. Acceleration of the MM algorithm

One disadvantage of the proposed MM algorithm is that it is computationally intensive. Yu (2012) proposed the *monotonic over-relaxation* for accelerating an EM algorithm. We adopt it to speed up the convergence of the MM algorithm for fitting a cumulative link model.

The idea of monotonic over-relaxation applies to the maximization step. At the maximization-step, update $\boldsymbol{\beta}^{(k+1)}$ as

$$\boldsymbol{\beta}^{(k+1)} = (1+\omega)\boldsymbol{\beta}_{MM}^{(k+1)} - \omega\boldsymbol{\beta}^{(k)}, \tag{19}$$

where $\omega \geq 0$ is the over-relaxation parameter. Because the minorizing function is a quadratic function of $\boldsymbol{\beta}$, hence for any $\omega \in [0,1]$,

$$Q(\boldsymbol{\beta}^{(k+1)}|\boldsymbol{\beta}^{(k)}) \geq Q(\boldsymbol{\beta}^{(k)}|\boldsymbol{\beta}^{(k)}).$$

For the MM algorithm, any parameter update that increases the minorizing function does the same for the observed data log-likelihood. Thus monotonicity of the observed data log-likelihood is maintained by updating $\boldsymbol{\beta}^{(k)}$ to $\boldsymbol{\beta}^{(k+1)}$ with a suitable $\omega \in [0,1]$, but could still be true for larger ω that further speeds up the convergence. Although the preceding equation is guaranteed to be valid for $\omega \in [0,1]$, this inequality may still hold with larger ω. We have performed an experiment to assess potential improvements of the over-relaxation method in reducing the computation time and yet obtaining correct estimation results. We illustrate the gains of the over-relaxation method in Table 9. Computation was done using a MaxOS computer with a 1.8 Ghz i5 Core with 4GB memory. We observe significant improvements due to monotonic over-relaxation, in terms of reduced computation time for these four cumulative link models.

[Table 9 about here.]

5. The Information Matrix for a Cumulative Link Model

Consider the observed log-likelihood

$$\ell(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n}\sum_{p=0}^{q} d_{i,p}\log[F(\tau_p - \mathbf{x}_i^\top\boldsymbol{\beta}) - F(\tau_{p-1} - \mathbf{x}_i^\top\boldsymbol{\beta})],$$

where $d_{i,p} = 1$ if $y_i = p$ and 0 otherwise. We denote $F(\tau_p - \mathbf{x}_i^\top \boldsymbol{\beta})$ as $F_{i,p}$, $f(\tau_p - \mathbf{x}_i^\top \boldsymbol{\beta})$ as $f_{i,p}$, and $\tau_p - \mathbf{x}_i^\top \boldsymbol{\beta}$ as $\alpha_{i,p}$. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \tau_0, \ldots, \tau_{q-1})^\top$. The observed score function of $\boldsymbol{\theta}$ is

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} \begin{pmatrix} \sum_{p=0}^{q} d_{i,p} \left[\frac{f_{i,p-1} - f_{i,p}}{F_{i,p} - F_{i,p-1}} \right] \mathbf{x}_i \\ d_{i,0} \frac{f_{i,0}}{F_{i,0}} - d_{i,1} \frac{f_{i,0}}{F_{i,1} - F_{i,0}} \\ d_{i,1} \frac{f_{i,1}}{F_{i,1} - F_{i,0}} - d_{i,2} \frac{f_{i,1}}{F_{i,2} - F_{i,1}} \\ \vdots \\ d_{i,q-1} \frac{f_{i,q-1}}{F_{i,q-1} - F_{i,q-2}} - d_{i,q} \frac{f_{i,q-1}}{1 - F_{i,q-1}} \end{pmatrix}.$$

Next we take the second derivative of the observed-data log-likelihood, it can be written as

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \sum_{i=1}^{n} \begin{pmatrix} D_i & C_i \\ B_i & A_i \end{pmatrix}, \tag{20}$$

where A_i is a $q \times q$ matrix, B_i is $q \times d$ matrix, C_i is a $d \times q$ matrix and D_i is a $d \times d$ matrix. We denote for $k = 0, \cdots, q-1$,

$$\begin{aligned} \gamma_{i,k} &= d_{i,k} \frac{\frac{\partial f_{i,k}}{\partial \tau_k}(F_{i,k} - F_{i,k-1}) - f_{i,k}^2}{(F_{i,k} - F_{i,k-1})^2} - d_{i,k+1} \frac{\frac{\partial f_{i,k}}{\partial \tau_k}(F_{i,k+1} - F_{i,k}) + f_{i,k}^2}{(F_{i,k+1} - F_{i,k})^2}, \\ \eta_{i,k} &= -d_{i,k+1} \frac{-f_{i,k} f_{i,k+1}}{(F_{i,k+1} - F_{i,k})^2}. \end{aligned}$$

Moreover,

$$\begin{aligned} & \frac{\partial \left(d_{i,k} \frac{f_{i,k}}{F_{i,k} - F_{i,k-1}} - d_{i,k+1} \frac{f_{i,k}}{F_{i,k+1} - F_{i,k}} \right)}{\partial \boldsymbol{\beta}} \\ = \; & d_{i,k} \frac{\frac{\partial f_{i,k}}{\partial \boldsymbol{\beta}}(F_{i,k} - F_{i,k-1}) - f_{i,k}^2 \frac{\partial f_{i,k}}{\partial \boldsymbol{\beta}}}{(F_{i,k} - F_{i,k-1})^2} - d_{i,k+1} \frac{\frac{\partial f_{i,k}}{\partial \boldsymbol{\beta}}(F_{i,k+1} - F_{i,k}) - f_{i,k}(f_{i,k+1} \frac{\partial f_{i,k+1}}{\partial \boldsymbol{\beta}} - f_{i,k} \frac{\partial f_{i,k}}{\partial \boldsymbol{\beta}})}{(F_{i,k+1} - F_{i,k})^2}. \end{aligned}$$

We note that $\frac{\partial f_{i,k}}{\partial \boldsymbol{\beta}} = \frac{\partial f_{i,k}}{\partial \tau_k}(-\mathbf{x}_i)$, then

$$A_i = \begin{pmatrix} \gamma_{i,0} & \eta_{i,0} & 0 & 0 & \cdots & 0 \\ \eta_{i,0} & \gamma_{i,1} & \eta_{i,1} & 0 & \cdots & 0 \\ 0 & \eta_{i,1} & \gamma_{i,2} & \eta_{i,2} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \eta_{i,q-3} & \gamma_{i,q-2} & \eta_{i,q-2} \\ 0 & 0 & \cdots & 0 & \eta_{i,q-2} & \gamma_{i,q-1} \end{pmatrix},$$

$$B_i = \begin{pmatrix} -(\gamma_{i,0} + \eta_{i,0})\mathbf{x}_i^\top \\ -(\gamma_{i,1} + \eta_{i,0} + \eta_{i,1})\mathbf{x}_i^\top \\ \vdots \\ -(\gamma_{i,q-1} + \eta_{i,q-2})\mathbf{x}_i^\top \end{pmatrix},$$

$$C_i = B_i^\top$$

$$D_i = \sum_{p=0}^{q} d_{i,p}\left[\frac{(f_{i,p-1} - f_{i,p})^2 - (\alpha_{i,p-1}f_{i,p-1} - \alpha_{i,p}f_{i,p})(F_{i,p} - F_{i,p-1})}{\big(F_{i,p} - F_{i,p-1}\big)^2}\right](\mathbf{x}_i\mathbf{x}_i^\top).$$

For normal errors,

$$\frac{\partial f_{i,k}}{\partial \tau_k} = \phi(\tau_k - \mathbf{x}_i^\top\boldsymbol{\beta})\{-(\tau_k - \mathbf{x}_i\boldsymbol{\beta})\}.$$

For logistic errors,

$$\frac{\partial f_{i,k}}{\partial \tau_k} = \frac{\exp(\tau_k - \mathbf{x}_i^\top\boldsymbol{\beta})\{1 - \exp(\tau_k - \mathbf{x}_i^\top\boldsymbol{\beta})\}}{\{1 + \exp(\tau_k - \mathbf{x}_i^\top\boldsymbol{\beta})\}^3}.$$

For Cauchy errors,

$$\frac{\partial f_{i,k}}{\partial \tau_k} = \frac{-2(\tau_k - \mathbf{x}_i^\top\boldsymbol{\beta})}{\{(\tau_k - \mathbf{x}_i^\top\boldsymbol{\beta})^2 + 1\}^2}.$$

For Gumbel errors,

$$\frac{\partial f_{i,k}}{\partial \tau_k} = -\exp\{\tau_k - \mathbf{x}_i^\top\boldsymbol{\beta} - \exp(\tau_k - \mathbf{x}_i^\top\boldsymbol{\beta})\}\{\exp(\tau_k - \mathbf{x}_i^\top\boldsymbol{\beta}) - 1\}.$$

Next, we show that expected Fisher information matrix is positive definite, under the mild assumption that the covariates are not collinear, i.e., there exists no non-zero linear combination of X that is a constant almost everywhere. Let $\boldsymbol{\omega} = (\boldsymbol{\omega}_1^\top, \boldsymbol{\omega}_2^\top)^\top = (\omega_1, \ldots, \omega_q, \omega_1^*, \ldots, \omega_d^*)^\top$ be a non-zero vector. Our aim is to verify that

$$\begin{aligned}
\boldsymbol{\omega}^\top E\Big(-\frac{\partial^2 \ell}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\Big)\boldsymbol{\omega} &= \boldsymbol{\omega}^\top \sum_{i=1}^{n} E\left\{E\begin{pmatrix} A_i & B_i \\ C_i & D_i \end{pmatrix}\Big|\,\mathbf{x}_i\right\}\boldsymbol{\omega} \\
&= \sum_{i=1}^{n} E\left\{\boldsymbol{\omega}_1^\top E(A_i|\mathbf{x}_i)\boldsymbol{\omega}_1 + 2\boldsymbol{\omega}_1^\top E(B_i|\mathbf{x}_i)\boldsymbol{\omega}_2 + \boldsymbol{\omega}_2^\top E(D_i|\mathbf{x}_i)\boldsymbol{\omega}_2\right\} \\
&= T_1 + T_2 + T_3 > 0.
\end{aligned}$$

First we notice that

$$E(A_i|\mathbf{x}_i) = \begin{pmatrix} \frac{f_{i,0}^2}{F_{i,0}} + \frac{f_{i,0}^2}{F_{i,1}-F_{i,0}} & \frac{-f_{i,0}f_{i,1}}{F_{i,1}-F_{i,0}} & 0 & 0 & \cdots & 0 \\ \frac{-f_{i,0}f_{i,1}}{F_{i,1}-F_{i,0}} & \frac{f_{i,1}^2}{F_{i,1}-F_{i,0}} + \frac{f_{i,1}^2}{F_{i,2}-F_{i,1}} & \frac{-f_{i,1}f_{i,2}}{F_{i,2}-F_{i,1}} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \frac{-f_{i,q-2}f_{i,q-1}}{F_{i,q-1}-F_{i,q-2}} & \frac{f_{i,q-1}^2}{F_{i,q-1}-F_{i,q-2}} + \frac{f_{i,q-1}^2}{1-F_{i,q-1}} \end{pmatrix}$$

and

$$\begin{aligned} \boldsymbol{\omega}_1^\top E(A_i|\mathbf{x}_i)\boldsymbol{\omega}_1 &= \omega_1^2 \frac{f_{i,0}^2}{F_{i,0}} + \frac{1}{F_{i,1}-F_{i,0}}(\omega_1 f_{i,0} - \omega_2 f_{i,1})^2 + \frac{1}{F_{i,2}-F_{i,1}}(\omega_2 f_{i,1} - \omega_3 f_{i,2})^2 \\ &\quad + \cdots + \frac{1}{F_{i,q-1}-F_{i,q-2}}(\omega_{q-1} f_{i,q-2} - \omega_q f_{i,q-1})^2 + \omega_q^2 \frac{f_{i,q-1}^2}{1-F_{i,q-1}}. \end{aligned} \tag{21}$$

This must be greater than 0, which can be proved by contradiction. Suppose $\boldsymbol{\omega}_1^\top E(A_i|\mathbf{x}_i)\boldsymbol{\omega}_1 = 0$. Every term on the right side of (21) must be equal to 0. Therefore $\omega_1 = \omega_2 = \ldots = \omega_q = 0$. However this contradicts the assumption that $\boldsymbol{\omega}$ is a non-zero vector. Hence $\boldsymbol{\omega}_1^\top E(A_i|\mathbf{x}_i)\boldsymbol{\omega}_1$ is strictly greater than 0 and $E(A_i)$ is positive definite. It can be readily verified that $E(B_i|\mathbf{x}_i) = -E(A_i|\mathbf{x}_i)\mathbf{1}\mathbf{x}_i^\top, E(D_i|\mathbf{x}_i) = E(\mathbf{1}^\top A_i \mathbf{1}|\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\top$. By Cauchy-Schwarz inequality,

$$\begin{aligned} T_1+T_2+T_3 &= \sum_{i=1}^n E\Big(\boldsymbol{\omega}_1^\top E(A_i|\mathbf{x}_i)\boldsymbol{\omega}_1 + 2\boldsymbol{\omega}_1^\top E(B_i|\mathbf{x}_i)\boldsymbol{\omega}_2 + \boldsymbol{\omega}_2^\top E(D_i|\mathbf{x}_i)\boldsymbol{\omega}_2\Big) \\ &= \sum_{i=1}^n E\Big(\boldsymbol{\omega}_1^\top E(A_i|\mathbf{x}_i)\boldsymbol{\omega}_1 - 2\boldsymbol{\omega}_1^\top E(A_i|\mathbf{x}_i)\mathbf{1}\mathbf{x}_i^\top\boldsymbol{\omega}_2 + \boldsymbol{\omega}_2^\top \mathbf{1}^\top E(A_i|\mathbf{x}_i)\mathbf{1}\mathbf{x}_i\mathbf{x}_i^\top\boldsymbol{\omega}_2\Big) \\ &= \sum_{i=1}^n E\Big(\boldsymbol{\omega}_1^\top E(A_i|\mathbf{x}_i)\boldsymbol{\omega}_1 - 2\frac{\boldsymbol{\omega}_1^\top E(A_i|\mathbf{x}_i)\mathbf{1}}{\sqrt{\mathbf{1}^\top E(A_i|\mathbf{x}_i)\mathbf{1}}}\sqrt{\mathbf{1}^\top E(A_i|\mathbf{x}_i)\mathbf{1}}\mathbf{x}_i^\top\boldsymbol{\omega}_2 + \boldsymbol{\omega}_2^\top \mathbf{1}^\top E(A_i|\mathbf{x}_i)\mathbf{1}\mathbf{x}_i\mathbf{x}_i^\top\boldsymbol{\omega}_2\Big] \\ &\geq \sum_{i=1}^n E\Big(\boldsymbol{\omega}_1^\top E(A_i|\mathbf{x}_i)\boldsymbol{\omega}_1 - 2\sqrt{\boldsymbol{\omega}_1^\top E(A_i|\mathbf{x}_i)\boldsymbol{\omega}_1}\sqrt{\mathbf{1}^\top E(A_i|\mathbf{x}_i)\mathbf{1}}\mathbf{x}_i^\top\boldsymbol{\omega}_2 + \boldsymbol{\omega}_2^\top \mathbf{1}^\top E(A_i|\mathbf{x}_i)\mathbf{1}\mathbf{x}_i\mathbf{x}_i^\top\boldsymbol{\omega}_2\Big) \\ &= \sum_{i=1}^n E\Big(\sqrt{\boldsymbol{\omega}_1^\top E(A_i|\mathbf{x}_i)\boldsymbol{\omega}_1} - \sqrt{\mathbf{1}^\top E(A_i|\mathbf{x}_i)\mathbf{1}}\mathbf{x}_i^\top\boldsymbol{\omega}_2\Big)^2. \end{aligned}$$

Cauchy-Schwarz inequality indicates that the inequality is an equality if and only if $\boldsymbol{\omega}_1 = \kappa\mathbf{1}$ for some constant κ. Hence, $T_1+T_2+T_3 = 0$ if and only if there exists κ such that

$$E\{\mathbf{1}^\top E(A_i|\mathbf{x}_i)\mathbf{1}(\kappa - \mathbf{x}_i^\top\omega_2)^2\} = 0.$$

Since $\mathbf{1}^\top E(A_i|\mathbf{x}_i)\mathbf{1} > 0$ for all $\mathbf{x}_i$, $T_1+T_2+T_3 = 0$ implies that $E(\kappa-\mathbf{x}_i^\top\omega_2)^2 = 0$, which entails that $\kappa - \mathbf{x}_i^\top\omega_2 = 0$ almost everywhere. Hence $\mathbf{x}_i^\top\omega_2$ is a constant. By the non-collinearity assumption for the covariates, $\omega_2 = 0$, but then $\kappa = 0$ and $\omega_1 = 0$, which is a contradiction. Therefore, we conclude that $T_1 + T_2 + T_3 > 0$ of $\boldsymbol{\omega} \neq 0$.

6. Conclusion

We have developed MM algorithms for doing maximum likelihood estimation of the cumulative link model, for four commonly used links. The maximum likelihood estimates so obtained are generally more accurate than those based on iteratively re-weighted least squares and Fisher scoring. The latter approach suffers from the problem of high failure rate of convergence, especially with highly correlated covariates. The MM algorithms can be readily extended to penalized likelihood estimation with high-dimensional covariate, which will be reported elsewhere.

References

Agresti, A., 2002. Categorical data analysis. 2 ed., New York: John Wiley & Sons.

Aitchison, J., Silvey, S.D., 1957. The generalization of probit analysis to the case of multiple responses. Biometrika , 131–140.

Anderson, J., Philips, P., 1981. Regression, discrimination and measurement models for ordered categorical variables. Applied Statistics , 22–31.

Barr, D.R., Sherrill, E.T., 1999. Mean and variance of truncated normal distributions. The American Statistician 53, 357–361.

Burridge, J., 1981. A note on maximum likelihood estimation for regression models using grouped data. Journal of the Royal Statistical Society. Series B (Methodological) , 41–45.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) , 1–38.

Embrechts, P., Klüppelberg, C., Mikosch, T., 2013. Modelling extremal events: for insurance and finance. volume 33. Springer Science & Business Media.

Greene, W.H., Hensher, D.A., 2010. Modeling ordered choices: A primer. Cambridge University Press.

Hunter, D.R., Lange, K., 2004. A tutorial on MM algorithms. The American Statistician 58, 30–37.

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

Pratt, J.W., 1981. Concavity of the log likelihood. Journal of the American Statistical Association 76, 103–106.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Simonoff, J.S., 2013. Analyzing categorical data. Springer Science & Business Media.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. Fourth ed., Springer, New York. URL: http://www.stats.ox.ac.uk/pub/MASS4. iSBN 0-387-95457-0.

Yu, Y., 2012. Monotonically overrelaxed EM algorithms. Journal of Computational and Graphical Statistics 21, 518–537.

Zhou, H., Zhang, Y., 2012. EM vs MM: A case study. Computational Statistics & Data Analysis 56, 3909–3920.

List of Tables

Table 1: Failure frequencies for the R function `polr` until 400 successes were obtained.

ρ	0	0.2	0.4	0.6	0.8
Number of Failures (Non-convergence)	249	220	216	252	330

Table 2: Average model error comparison

Model Error	probit		logit		cauchy		cloglog	
	MM	polr	MM	polr	MM	polr	MM	polr
$\rho = 0$	3.96	6.05	1.96	2.15	2.36	5.73	0.80	3.32
$\rho = 0.8$	5.94	15.37	2.06	2.27	2.25	6.96	1.13	5.04

Table 3: $\rho = 0$: Comparison of average regression estimates (standard deviation of estimators enclosed in parentheses)

	$\boldsymbol{\beta}$	1	3	-2	5	0.5
Probit	MM	1.20 (0.32)	3.56 (0.77)	-2.38 (0.54)	5.94 (1.25)	0.58 (0.24)
	`polr`	1.23 (0.39)	3.66 (0.96)	-2.44 (0.66)	6.10 (1.56)	0.60 (0.28)
Logit	MM	1.10 (0.33)	3.31 (0.59)	-2.22 (0.43)	5.52 (0.90)	0.55 (0.28)
	`polr`	1.11 (0.34)	3.34 (0.61)	-2.24 (0.44)	5.57 (0.93)	0.56 (0.28)
Cauchy	MM	1.10 (0.36)	3.30 (0.67)	-2.19 (0.50)	5.49 (1.05)	0.56 (0.33)
	`polr`	1.18 (0.46)	3.57 (0.99)	-2.37 (0.71)	5.94 (1.59)	0.60 (0.38)
Cloglog	MM	1.06 (0.24)	3.15 (0.40)	-2.11 (0.30)	5.27 (0.62)	0.53 (0.21)
	`polr`	1.02 (0.38)	3.01 (0.87)	-2.01 (0.60)	5.04 (1.40)	0.50 (0.28)

Table 4: $\rho = 0.8$: Comparison of average regression estimates (standard deviation of estimators enclosed in parentheses)

	$\boldsymbol{\beta}$	1	3	-2	5	0.5
Probit	MM	1.18 (0.46)	3.58 (1.02)	-2.37 (0.74)	5.95 (1.52)	0.62 (0.42)
	`polr`	1.24 (0.61)	3.75 (1.59)	-2.48 (1.06)	6.25 (2.59)	0.65 (0.49)
Logit	MM	1.10 (0.50)	3.32 (0.78)	-2.23 (0.70)	5.51 (1.02)	0.55 (0.49)
	`polr`	1.11 (0.51)	3.36 (0.80)	-2.25 (0.71)	5.57 (1.05)	0.55 (0.49)
Cauchy	MM	1.07 (0.55)	3.23 (0.87)	-2.17 (0.79)	5.40 (1.16)	0.54 (0.52)
	`polr`	1.19 (0.69)	3.56 (1.21)	-2.39 (1.01)	5.96 (1.79)	0.60 (0.61)
Cloglog	MM	0.98 (0.37)	2.96 (0.69)	-1.95 (0.55)	4.91 (0.99)	0.50 (0.34)
	`polr`	1.02 (0.52)	3.09 (1.12)	-2.04 (0.85)	5.11 (1.67)	0.53 (0.47)

Table 5: $\rho = 0$: Comparison of average threshold estimates (standard deviation of estimators enclosed in parentheses)

	$\boldsymbol{\tau}$	0|1	1|2	2|3	3|4
	True $\boldsymbol{\tau}$	-5.33	-1.60	1.62	5.36
Probit	MM	-6.34 (1.40)	-1.89 (0.53)	1.91 (0.57)	6.35 (1.30)
	`polr`	-6.51 (1.72)	-1.93 (0.58)	1.96 (0.61)	6.53 (1.75)
	True $\boldsymbol{\tau}$	-5.46	-1.60	1.69	5.49
Logit	MM	-5.99 (1.03)	-1.79 (0.52)	1.86 (0.53)	6.06 (1.02)
	`polr`	-6.05 (1.06)	-1.80 (0.53)	1.87 (0.54)	6.11 (1.05)
	True $\boldsymbol{\tau}$	-6.06	-1.76	1.84	6.10
Cauchy	MM	-6.63 (1.30)	-1.92 (0.61)	2.05 (0.62)	6.70 (1.37)
	`polr`	-7.12 (1.89)	-2.06 (0.77)	2.20 (0.79)	7.19 (1.95)
	True $\boldsymbol{\tau}$	-5.90	-2.14	1.12	4.82
Cloglog	MM	-6.55 (1.03)	-2.34 (0.51)	1.34 (0.43)	5.51 (0.92)
	`polr`	-4.78 (1.58)	-1.02 (0.63)	2.26 (0.57)	6.02 (1.48)

Table 6: $\rho = 0.8$: Comparison of average threshold estimates (standard deviation of estimators enclosed in parentheses)

	$\boldsymbol{\tau}$	0|1	1|2	2|3	3|4
	True $\boldsymbol{\tau}$	-5.73	-1.75	1.68	5.68
Probit	MM	-6.86 (1.79)	-2.08 (0.63)	2.01 (0.67)	6.77 (1.76)
	`polr`	-7.19 (2.92)	-2.19 (1.01)	2.10 (0.87)	7.10 (2.85)
	True $\boldsymbol{\tau}$	-5.86	-1.76	1.76	5.82
Logit	MM	-6.47 (1.10)	-1.97 (1.56)	1.94 (0.55)	6.42 (1.06)
	`polr`	-6.53 (1.13)	-1.99 (0.57)	1.96 (0.56)	6.48 (1.09)
	True $\boldsymbol{\tau}$	-6.41	-1.88	1.93	6.42
Cauchy	MM	-6.88 (1.33)	-2.00 (0.60)	2.05 (0.59)	6.90 (1.34)
	`polr`	-7.54 (2.10)	-2.19 (0.79)	2.25 (0.79)	7.56 (2.08)
	True $\boldsymbol{\tau}$	-6.37	-2.33	1.18	5.18
Cloglog	MM	-6.25 (1.13)	-2.26 (0.51)	1.22 (0.41)	5.19 (1.00)
	`polr`	-5.36 (1.94)	-1.25 (0.72)	2.37 (0.65)	6.51 (1.83)

Table 7: $\rho = 0$: Average empirical coverage rates for the estimates

		β_1	β_2	β_3	β_4	β_5	τ_0	τ_1	τ_2	τ_3
Probit	MM	0.94	0.95	0.94	0.94	0.95	0.94	0.94	0.95	0.95
	`polr`	0.94	0.94	0.93	0.93	0.95	0.93	0.94	0.94	0.95
Logit	MM	0.94	0.94	0.94	0.94	0.95	0.96	0.95	0.95	0.95
	`polr`	0.94	0.94	0.93	0.93	0.95	0.95	0.95	0.95	0.94
Cauchy	MM	0.97	0.98	0.97	0.98	0.95	0.98	0.96	0.97	0.98
	`polr`	0.97	0.98	0.96	0.98	0.96	0.98	0.96	0.96	0.98
Cloglog	MM	0.95	0.98	0.97	0.98	0.95	0.97	0.96	0.94	0.98
	`polr`	0.84	0.75	0.79	0.75	0.86	0.49	0.24	0.28	0.78

Table 8: $\rho = 0.8$: Average empirical coverage rates for the estimates

		β_1	β_2	β_3	β_4	β_5	τ_0	τ_1	τ_2	τ_3
Probit	MM	0.94	0.94	0.94	0.95	0.94	0.94	0.95	0.93	0.94
	`polr`	0.94	0.93	0.92	0.93	0.94	0.92	0.94	0.92	0.93
Logit	MM	0.94	0.95	0.94	0.94	0.93	0.95	0.95	0.95	0.95
	`polr`	0.94	0.94	0.94	0.94	0.93	0.94	0.95	0.95	0.95
Cauchy	MM	0.97	0.96	0.97	0.97	0.96	0.98	0.97	0.96	0.98
	`polr`	0.97	0.97	0.96	0.97	0.96	0.97	0.97	0.96	0.98
Cloglog	MM	0.94	0.93	0.94	0.93	0.95	0.92	0.92	0.93	0.92
	`polr`	0.86	0.80	0.82	0.79	0.87	0.53	0.30	0.30	0.79

Table 9: Average computation time (in seconds) per fitting a cumulative link model, using the monotonic over-relaxation method, and based on 100 replications

ω	0	0.25	0.5	0.75	1	3
Probit	3.4	2.8	2.3	2.0	2.0	0.9
Logit	1.02	0.84	0.71	0.62	0.55	0.28
Cauchy	7.6	7.5	6.9	6.2	5.5	3.3
Cloglog	8.0	6.8	5.9	5.2	4.6	2.6