Efficient Estimation and Inference in the Proportional Odds Model for Survival Data

Huang, Xifen; Xiong, Chaosong; Jiang, Tao; Lu, Junfeng; Xu, Jinfeng

Published in:
Mathematics

Published: 01/09/2022

Document Version:
Final Published version, also known as Publisher’s PDF, Publisher’s Final version or Version of Record

License:
CC BY

Publication record in CityU Scholars:
Go to record

Published version (DOI):
10.3390/math10183362

Publication details:

Citing this paper
Please note that where the full-text provided on CityU Scholars is the Post-print version (also known as Accepted Author Manuscript, Peer-reviewed or Author Final version), it may differ from the Final Published version. When citing, ensure that you check and use the publisher's definitive version for pagination and other details.

General rights
Copyright for the publications made accessible via the CityU Scholars portal is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights. Users may not further distribute the material or use it for any profit-making activity or commercial gain.

Publisher permission
Permission for previously published items are in accordance with publisher's copyright policies sourced from the SHERPA RoMEO database. Links to full text versions (either Published or Post-print) are only available if corresponding publishers allow open access.

Take down policy
Contact lbscholars@cityu.edu.hk if you believe that this document breaches copyright and provide us with details. We will remove access to the work immediately and investigate your claim.

Download date: 15/04/2024
Abstract: In modeling time-to-event data with long-term survivors, the proportional hazards model is widely used for its easy and direct interpretation as well as the flexibility to accommodate the past information and allow time-varying predictors. This becomes most relevant when the mortality of individuals converges with time, and the estimation and inference based upon the proportional odds model can often yield more accurate and reasonable results than the classical Cox’s proportional hazards model. Along with the fast development of the data science technologies, computational challenges for survival data with increasing sample size and diverging parameter dimension exist. Currently, existing methods for analyzing such data are computationally inconvenient. In this paper, we propose efficient computational methods for analyzing survival data in the proportional odds model, where the nonparametric maximum likelihood approach is combined with the minorization-maximization (MM) algorithm and the regularization scheme to yield fast and accurate estimation and inferential procedures. The illustration of the methodology using extensive simulation studies and then the application to the Veterans’ Administration lung cancer data is also given.

Keywords: long-term survivor; model selection; regularization method; time-varying covariates

MSC: 46N30; 62N01; 62N02

1. Introduction

The proportional odds model was initially proposed by McCullagh [1,2] with the purpose of analyzing ordinal data instead of censored survival data. Due to its easy and direct interpretation, it has been widely used in practice. Ref. [3,4] extended the model to fit survival data using the Newton-Raphson method. Many researchers started to use this model in survival analysis also because of its good prediction performance. Ref. [5] considered the rank-based estimation method. Ref. [6] proposed the semi-parametric proportional odds model, which is one important case of the general linear transformation model. In addition, ref. [7] employed the profited likelihood method and developed the model diagnostic procedures. For interval censored data, ref. [8,9] further introduced the sieve maximum likelihood estimation and obtained the consistency and asymptotic normality for the estimated parameters. Ref. [10] proposed an easy implementation of this proportional odds model based on the conditional logistic regression. Ref. [11] considered the semi-parametric proportional odds model, where the baseline function can be any non-decreasing function. As for the improvement of the parameter estimation efficiency, ref. [12] proposed the majorization-maximization (MM) algorithm for the proportional odds model and this algorithm performs well given high-dimensional data. Ref. [13] adopted the cubic spline for baseline estimation and [14] proposed to capture the spatial heterogeneity using a Bayesian hierarchical model for the analysis of spatially related data. Similarly, ref. [15] also proposed the semiparametric Bayesian proportional odds model, where the baseline is estimated through a monotone increasing spline. The Bayesian model framework was further extended to the application of clustered and multi-event data by [16]. Ref. [17] proposed an
extension of the original model in order to fit the data with a multivariate response, random intercept and non-linear effect of covariates. Furthermore, based on the interval censored data and Bayesian estimation method, ref. [18] introduced the cure rate proportional odds models for the corresponding data analysis. Different from the traditional proportional odds model, ref. [19] discussed the quantile-based definition for this model and presented the illustration using real life datasets. Ref. [20] added a log-concave constraint on the baseline distribution and developed the corresponding parameter and density estimation. Besides the MM algorithm, ref. [21] proposed an expectation-maximization (EM) scheme for the parameter estimation with a good computational efficiency for both parameter and baseline estimation. In addition, ref. [22] discussed the efficient estimation of the odds ratio for the proportional odds model with censored time-lagged outcome. [23] has pointed out the high computational complexity and inefficiency for the proportional odds model with right censored data. For example, ref. [7]’s method cannot perform estimation appropriately given the consecutively observed failure time. In addition, ref. [13] pointed out the method proposed by [9] is very complex and computationally unachievable. Moreover, the estimated baseline function from [13]’s method may not preserve the monotone property constructed by the natural cubic spline. Therefore, to tackle these problems, ref. [23] developed the EM algorithm for the proportional odds model with right censored data, where the baseline function is estimated using the spline. The method can be applied to a large dataset, while AIC or BIC criterion is applied to determine the optimal number of knots for the natural cubic spline estimation.

The covariates are important for the regression analysis, where some covariates with a tiny impact to the response might be captured during the estimation procedure. Therefore, when constructing the regression model, it is essential to conduct a variable selection procedure. The most popular way of variable selection is conducted by adding penalty functions to the original objective. The parameters are then estimated by maximizing the new objective function. In the survival analysis, the LASSO [24], SCAD [25], the adaptive LASSO [26] and the elastic SCAD are commonly applied for the proportional hazard models. As for the proportional odds model, only limited studies are conducted based on the penalized regression. Ref. [27] studied the application of LASSO and the adaptive LASSO in the proportional odds model, where the simulation results of the adaptive LASSO demonstrate better performance in both variable selection and parameter estimation.

As for optimizing the likelihood function of the proportional odds model, the Newton-Raphson method and EM algorithm perform well in finding the root without an explicit form. Ref. [28] first developed the MM algorithm to optimize the convex function iteratively. EM algorithm is a special case of the MM algorithm. The MM algorithm can help avoid high-dimensional matrix inversion, separate the parameters, linearize the optimization problem and transform the non-differentiable problem to smooth optimization. In real applications, the MM algorithm has been applied to optimize the objective function from statistical models such as quantile regression [28], Bradley-Terry model [29], zero-inflated and zero-truncated models [30,31], high-dimensional covariates selection problem [32], mixed Gaussian model [33] and finite mixture models. In addition, the MM algorithm has been also applied to survival analysis. Ref. [34] has developed a MM algorithm to solve the optimization problem based on the shared gamma frailty model. The estimation results perform well in their simulation and real data analysis.

In the following article, we first review the MM algorithm in Section 2. Then, we apply the usage of the MM algorithm in the proportional odds model in Section 3. In Section 4, we further conduct a variable selection procedure based on the proportional odds model using the MM algorithm. Simulation studies and real data analysis are conducted in Sections 5 and 6. Finally, we conclude the performance of our proposed methods and discuss the future work in Section 7.
2. MM Algorithm

2.1. Basic Principle

The MM algorithm performs optimization by constructing a simple surrogate function for the original objective function. Then, instead of optimizing the original objective function, we optimize the surrogate function with simplified expression. The process of the MM algorithm involves two steps; the first M step is the minorization step, which constructs a proper minorization function under the following conditions (1) and (2) by some commonly used inequations. The second M step is the maximization step, where a minorization function is maximized via the Newton-Raphson method or a quasi-Newton method, since the constructed minorization function is usually parameter-separable.

Let $\ell(\theta | Y_{\text{obs}})$ is the log-likelihood of observed data $Y_{\text{obs}}$, $\hat{\theta}$ is the parameter estimation through the maximization of log-likelihood. Assume $\theta(t)$ is the $t$-th iteration of $\hat{\theta}$, $Q(\theta | \theta(t))$ is the surrogate function determined by $\theta(t)$. If $\theta(t)$ satisfy,

\begin{align}
Q(\theta | \theta(t)) & \leq \ell(\theta | Y_{\text{obs}}) \quad \forall \theta, \theta(t) \in \Theta, \quad \text{and} \\
Q(\theta(t) | \theta(t)) & = \ell(\theta(t) | Y_{\text{obs}}).
\end{align}

Then, we call $Q(\theta | \theta(t))$ to be the minorization function of $\ell(\theta | Y_{\text{obs}})$. The maximizing of $Q(\theta | \theta(t))$ can substitute the maximization of the original objective. If $Q(\theta | \theta(t))$ obtains the optimal at $\theta(t+1)$,

$$\theta(t+1) = \arg \max_{\theta \in \Theta} Q(\theta | \theta(t)).$$

From (1) and (2), we have

$$\ell(\theta(t+1) | Y_{\text{obs}}) \geq Q(\theta(t+1) | \theta(t)) \geq Q(\theta(t) | \theta(t)) = \ell(\theta(t) | Y_{\text{obs}}).$$

The MM algorithm is a stable optimization procedure and has a convergence property due to the increase in the objective in every iteration. EM algorithm is a special case of MM algorithm and its first M step (also called E step) is to calculate the expectation of the complete-data log-likelihood function.

2.2. Commonly Used Inequalities

The key step of the MM algorithm is to find the appropriate surrogate function for the objective with less computational cost. In real applications, in order to construct the surrogate function, we suggest the following ways based on the commonly used algebraic inequalities.

The first method is Jensen’s inequality. Assume that $X$ is a random variable, and $\varphi$ is a concave function, then

$$\varphi(\mathbb{E}[X]) \geq \mathbb{E}[\varphi(X)].$$

In contrast, if $\varphi$ is convex, then,

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)].$$

In real applications, we always use the continuous and discrete version of Jensen’s Inequality. The continuous version is

$$\varphi\left(\int_{\Omega} f(x) \cdot g(x) dx\right) \geq \int_{\Omega} \varphi(f(x)) \cdot g(x) dx,$$
where $\Omega$ is the subset of the real line $\mathbb{R}$. $f(\cdot)$ is defined real function on $\Omega$, $g(\cdot)$ is the density function on $\Omega$. Discrete version,

$$
\varphi \left( \sum_{i=1}^{n} a_i x_i \right) \geq \sum_{i=1}^{n} a_i \varphi(x_i),
$$

where $\varphi(\cdot)$ is a concave function, $a_i \geq 0$ and $\sum_{i=1}^{n} a_i = 1$.

The second method is the inequality of arithmetic and geometric means,

$$
- \prod_{i=1}^{n} x_i^{a_i} \geq - \sum_{i=1}^{n} \frac{a_i}{\|a_i\|_1} x_i^{\|a_i\|_1},
$$

where $x_i$ and $a_i$ are nonnegative. The left-hand side of the inequality is a product of $x_i^{a_i}$ and the other side of this inequality is the sum of $x_i^{\|a_i\|_1}$ for $i = 1, \ldots, n$. The structure of this inequality says that it can be used to minorize product terms into the summation of other terms.

The third method is the supporting hyperplane inequality,

$$
- \log(x) \geq - \log(x_0) - \frac{x - x_0}{x_0}.
$$

3. Proportional Odds Model

Let $T$ denotes the failure time, $X$ is the covariate of $p \times 1$ dimension, $\beta$ is the corresponding coefficient. Suppose that given $X$, $T$ follow the proportional odds model

$$
\text{logit} \{ F(t \mid X) \} = \log \Lambda_0(t) + X^\top \beta,
$$

where $F(t \mid X)$ denotes the cumulative distribution function of $T$ given $X$, $\Lambda_0(t)$ denotes a cumulative baseline function, and $\text{logit} \ (x) = \log \frac{1}{1-x}$.

Based on the assumption above, we can easily obtain the survival and density function of the proportional odds model

$$
S(t \mid X) = \frac{1}{1 + \Lambda_0(t) \exp(X^\top \beta)},
$$

and

$$
f(t \mid X) = \frac{\lambda_0(t) \exp(X^\top \beta)}{[1 + \Lambda_0(t) \exp(X^\top \beta)]^2},
$$

respectively, with the hazard rate

$$
\lambda(t \mid X) = \frac{\lambda_0(t) \exp(X^\top \beta)}{1 + \Lambda_0(t) \exp(X^\top \beta)},
$$

where $\lambda_0(t) = d\Lambda_0(t)/dt$.

Here, we consider the failure time $T$ with a right censoring structure, and a failure time study that consists of $n$ independent individuals. Let $T_i$, $C_i$ and $X_i$ denote the failure time, the censoring time and the covariate of the $i^{th}$ individual, respectively. Moreover, we assume the $T_i$ and $C_i$ are independent. In addition, let $t_i = \min(T_i, C_i)$ denote the observation time of event, and $\delta_i = I(T_i \leq C_i)$ is the censoring indicator based on the observed data $\{t_i, \delta_i, X_i\}, i = 1, \ldots, n$. Then, the likelihood function is

$$
L_{\text{obs}} = \prod_{i=1}^{n} \left\{ \frac{\lambda_0(t_i) \exp(X_i^\top \beta)}{1 + \Lambda_0(t_i) \exp(X_i^\top \beta)} \right\}^{\delta_i} \frac{1}{1 + \Lambda_0(t_i) \exp(X_i^\top \beta)}.
$$
with the corresponding log-likelihood function

\[ \ell_{\text{obs}} = \sum_{i=1}^{n} \left\{ \delta_i \log \lambda_0(t_i) + \delta_i X_j^T \beta - (\delta_i + 1) \log \left[ 1 + \Lambda_0(t_i) \exp \left( X_j^T \beta \right) \right] \right\}. \quad (3) \]

In Equation (3), using the supporting hyperplane inequality,

\[- \log(x) \geq - \log(x_0) - \frac{x - x_0}{x_0},\]

we have

\[-(\delta_i + 1) \log \left[ 1 + \Lambda_0(t_i) \exp \left( X_j^T \beta \right) \right] \geq - (\delta_i + 1) \log \left( A_i^{(k)} \right) - (\delta_i + 1) \frac{1 + \Lambda_0(t_i) \exp \left( X_j^T \beta \right) - A_i^{(k)}}{A_i^{(k)}},\]

where \( A_i^{(k)} = 1 + \Lambda_0^{(k)}(t_i) \exp \left( X_j^T \beta^{(k)} \right) \). Then, we can obtain the surrogate function of \( \ell_{\text{obs}} \), as follows

\[ Q_{11}(\Lambda_0, \beta \mid \Lambda_0^{(k)}, \beta^{(k)}) = \sum_{i=1}^{n} \left\{ \delta_i \log \lambda_0(t_i) + \delta_i X_j^T \beta - \frac{(\delta_i + 1) \Lambda_0(t_i) \exp \left( X_j^T \beta \right)}{A_i^{(k)}} \right\}. \]

After the minimization step, we then apply two methods to estimate the surrogate function derived above.

3.1. Profile MM Method

Given the value of \( \beta \), we can obtain the estimate of \( \Lambda_0 \)

\[ d\hat{\Lambda}_0(t_i) = \frac{\delta_i}{\sum_{j=1}^{n} I(t_j \geq t_i) (\delta_j + 1) \exp \left( X_j^T \beta \right) / A_j^{(k)}}. \quad (4) \]

Substitute (4) into \( Q_{11}(\Lambda_0, \beta \mid \Lambda_0^{(k)}, \beta^{(k)}) \), we have

\[ Q_{12}(\beta \mid \Lambda_0^{(k)}, \beta^{(k)}) = \sum_{i=1}^{n} \delta_i X_j^T \beta - \sum_{i=1}^{n} \left\{ \delta_i \log \sum_{j=1}^{n} I(t_j \geq t_i) (\delta_j + 1) \exp \left( X_j^T \beta \right) / A_j^{(k)} \right\} + c_1, \]

where \( c_1 \) is a constant. Using the supporting hyperplane inequality again to minimize \( Q_{12}(\beta \mid \Lambda_0^{(k)}, \beta^{(k)}) \), we have

\[ Q_{13}(\beta \mid \Lambda_0^{(k)}, \beta^{(k)}) = \sum_{i=1}^{n} \delta_i X_j^T \beta - \sum_{i=1}^{n} \left\{ \delta_i \sum_{j=1}^{n} I(t_j \geq t_i) (\delta_j + 1) \exp \left( X_j^T \beta / A_j^{(k)} \right) / B_i^{(k)} \right\} + c_2, \quad (5) \]

where \( c_2 \) is a constant and \( B_i^{(k)} = \sum_{j=1}^{n} I(t_j \geq t_i) (\delta_j + 1) \exp \left( X_j^T \beta \right) / A_j^{(k)}. \)

Taking first and second derivatives for \( Q_{13}(\beta \mid \Lambda_0^{(k)}, \beta^{(k)}) \) to \( \beta \), then

\[ Q'_{13}(\beta \mid \Lambda_0^{(k)}, \beta^{(k)}) = \sum_{i=1}^{n} \delta_i X_j^T - \sum_{i=1}^{n} \delta_i \sum_{j=1}^{n} I(t_j \geq t_i) (\delta_j + 1) \exp \left( X_j^T \beta / A_j^{(k)} \right) X_j / B_i^{(k)}, \]

\[ Q''_{13}(\beta \mid \Lambda_0^{(k)}, \beta^{(k)}) = -\sum_{i=1}^{n} \delta_i \sum_{j=1}^{n} I(t_j \geq t_i) (\delta_j + 1) \exp \left( X_j^T \beta \right) X_j X_j / B_i^{(k)}. \]
The estimation function for $\beta$ is

$$
\hat{\beta}^{(k+1)} = \beta^{(k)} - Q_{13}^{x_{i}} \left( \beta^{(k)} | A^{(k)}_{0}, \beta^{(k)} \right)^{-1} Q_{13}^{x_{i}} \left( \beta^{(k)} | A^{(k)}_{0}, \beta^{(k)} \right).
$$

The algorithm is given below,

1. Let the initial value for $(\beta, \Lambda)$ be $(\beta^{(0)}, \Lambda^{(0)})$.
2. Update the estimate of $\beta$ by Equation (5).
3. Using the updated value of $\beta$, calculate the estimate of $\Lambda_{0}$ by Equation (4).
4. Iterate step 2 and 3 until it converges.

### 3.2. Non-Profile MM Method

Use the inequality of arithmetic and geometric means

$$
- \prod_{i=1}^{n} x_{i}^{|a|_{i}} \geq - \sum_{i=1}^{n} \frac{d_{i} ||a||_{i}}{d_{i} ||a||_{i}}
$$

to minimize $Q_{11}(\Lambda_{0}, \beta | A_{0}^{(k)}, \beta^{(k)})$, let $x_{1} = \Lambda_{0}(t_{i}) / A_{0}^{(k)}(t_{i})$ and $x_{2} = \exp(X_{i}^{T} \beta) / \exp(X_{i}^{T} \beta^{(k)})$, we have

$$
- \frac{\Lambda_{0}(t_{i}) \exp(X_{i}^{T} \beta)}{A_{0}^{(k)}(t_{i}) \exp(X_{i}^{T} \beta^{(k)})} \geq - \frac{\Lambda_{0}^{2}(t_{i})}{2A_{0}^{(k)}(t_{i})} - \frac{\Lambda_{0}(t_{i})}{2 \exp(X_{i}^{T} \beta^{(k)})} \exp(2X_{i}^{T} \beta).
$$

By computation, we obtain

$$
- \Lambda_{0}(t_{i}) \exp(X_{i}^{T} \beta) \geq - \frac{\exp(X_{i}^{T} \beta^{(k)})}{2A_{0}^{(k)}(t_{i})} \Lambda_{0}^{2}(t_{i}) - \frac{\Lambda_{0}(t_{i})}{2 \exp(X_{i}^{T} \beta^{(k)})} \exp(2X_{i}^{T} \beta).
$$

Therefore, the surrogate function is

$$
Q_{14}(\Lambda_{0}, \beta | A_{0}^{(k)}, \beta^{(k)}) = Q_{14}(\Lambda_{0} | A_{0}^{(k)}, \beta^{(k)}) + Q_{14}(\beta | A_{0}^{(k)}, \beta^{(k)}),
$$

where

$$
Q_{14}(\Lambda_{0} | A_{0}^{(k)}, \beta^{(k)}) = n \sum_{i=1}^{n} \delta_{i} \log \Lambda_{0}(t_{i}) - \sum_{i=1}^{n} \frac{(\delta_{i} + 1) \exp(X_{i}^{T} \beta^{(k)})}{2A_{0}^{(k)}(t_{i})} \Lambda_{0}^{2},
$$

$$
Q_{14}(\beta | A_{0}^{(k)}, \beta^{(k)}) = \sum_{i=1}^{n} \delta_{i} X_{i}^{T} \beta - \sum_{i=1}^{n} \frac{(\delta_{i} + 1) \Lambda_{0}^{(k)}}{2A_{0}^{(k)}(t_{i})} \exp(X_{i}^{T} \beta^{(k)}) \exp(2X_{i}^{T} \beta).
$$

Let $\frac{\partial Q_{14}(\Lambda_{0}, \beta | A_{0}^{(k)}, \beta^{(k)})}{\partial \Lambda_{0}} = 0$, the estimation equation of $\Lambda_{0}$ is

$$
d\Lambda_{0}(t_{i}) = \frac{\delta_{i}}{\sum_{j=1}^{n} I(t_{j} \geq t_{i})(\delta_{j} + 1) \exp(X_{j}^{T} \beta) / A_{0}^{(k)}(t_{j})}.
$$

The first and second derivatives of $\beta$ are

$$
Q_{14}^{x_{i}}(\beta | A_{0}^{(k)}, \beta^{(k)}) = \sum_{i=1}^{n} \delta_{i} X_{i}^{T} - \sum_{i=1}^{n} \frac{(\delta_{i} + 1) \Lambda_{0}^{(k)}}{A_{0}^{(k)}(t_{i}) \exp(X_{i}^{T} \beta^{(k)})} \exp(2X_{i}^{T} \beta) X_{i}^{T},
$$

$$
Q_{14}^{x_{i}x_{i}}(\beta | A_{0}^{(k)}, \beta^{(k)}) = - \sum_{i=1}^{n} \frac{2(\delta_{i} + 1) \Lambda_{0}^{(k)}}{A_{0}^{(k)}(t_{i}) \exp(X_{i}^{T} \beta^{(k)})} \exp(2X_{i}^{T} \beta) X_{i}^{T} X_{i}.
$$
Thus, we have the estimation function of $\beta$ which is
\[
\beta^{(k+1)} = \beta^{(k)} - Q''_{14}(\beta^{(k)} | \Lambda_0^{(k)}, \beta^{(k)})^{-1} Q'_{14}(\beta^{(k)} | \Lambda_0^{(k)}, \beta^{(k)}).
\]

The algorithm is as follows,
1. Let the initial value of $({\beta, \Lambda_0})$ to be $({\beta^{(0)}, \Lambda_0^{(0)})}$.
2. Update $\beta$ using Equation (6).
3. Use the updated value of $\beta$, calculate the estimate of $\Lambda_0$ using Equation (7).
4. Iterate step 2 and 3 until convergence.

To sum up, both profile MM and non-profile MM methods have decomposed the objective function into two separate parts. That is, the non-parametric component $\Lambda_0$ is separated from the regression vector $\beta$, which makes the next maximization step more simple than directly optimizing the objective log-likelihood function. It is worth noting that the parameter-separable feature is one of the advantages of the MM algorithm, which can easily incorporate the quasi-Newton acceleration and other simple off-the-shelf accelerators for boosting computational effectiveness.

4. Variable Selection in the Proportional Odds Model

4.1. Parameter Separated Estimation Method

Notice that the estimate of $\beta$ relies on the Newton-Raphson algorithm, which is sensitive to the initial value, and may lead to computational inefficiency due to the inappropriate choice of initial value. Particularly, the higher the dimension of $\beta$, the higher order the matrix of Henssian from the Newton-Raphson method with a high computational cost in doing matrix inverse. Under this circumstance, the proposed MM method can avoid such a matrix inversion problem with much lower computational cost. In the following section, we describe in detail on the parameter-separated estimation method following the previous two methods of Section 3.

First, let
\[
X_i^\top \beta = \sum_{q=1}^{p} \omega_{iq} \left[ \omega_{iq}^{-1} x_{iq}(\beta_q - \beta_q^{(k)}) + X_i^\top \beta^{(k)} \right],
\]
when $x_{iq} = 0$, we let $1/\omega_{iq} = 0$. Then, using the discrete form of Jensen’s inequality,
\[
\varphi \left( \sum_{i=1}^{n} a_i x_i \right) \geq \sum_{i=1}^{n} a_i \varphi(x_i),
\]
where, $\varphi(\cdot)$ is a concave function, $a_i \geq 0$ and $\sum_{i=1}^{n} a_i = 1$.

Let $\varphi(\cdot) = -\exp(\cdot)$ in Equation (9) and use the expression form of Equation (8), we can minimize Equation (5) by
\[
Q_{15}(\beta | \Lambda_0^{(k)}, \beta^{(k)}) = \sum_{q=1}^{p} Q_{15q}(\beta_q | \Lambda_0^{(k)}, \beta^{(k)}),
\]
where
\[
Q_{15q}(\beta_q | \Lambda_0^{(k)}, \beta^{(k)})
= \sum_{j=1}^{n} \left\{ \delta_j x_{ij} \beta_q - \delta_j \sum_{j=1}^{n} I(t_j > t_i) (\delta_j + 1) \omega_{ij} \exp(\omega_{ij}^{-1} x_{ij}(\beta_q - \beta_q^{(k)}) + X_i^\top \beta^{(k)})/A_j^{(k)} \right\}.
\]
Similarly, let \( \varphi(\cdot) = -\exp(\cdot) \) in Equation (9) and use the expression form of Equation (8), we can also minimize Equation (6) by

\[
Q_{16}(\beta | \Lambda_{0}^{(k)}, \beta^{(k)}) = \sum_{q=1}^{p} Q_{16q}(\beta_q | \Lambda_{0}^{(k)}, \beta^{(k)}),
\]

where

\[
Q_{16q}(\beta_q | \Lambda_{0}^{(k)}, \beta^{(k)}) = \sum_{i=1}^{n} \left\{ \delta_i x_i q \beta_q - \frac{(\delta_i + 1) \Lambda_{0}^{(k)}}{2 \Lambda_{i}^{(k)}} \omega_{i q} \exp \left( 2 \omega_{i q} \left( \beta_q - \beta_q^{(k)} \right) + 2 X_{i}^{\top} \beta^{(k)} \right) \right\}.
\]

From Equations (10) and (11), it can be observed that the two resulting MM methods only involves \( p + 1 \) separate univariate optimizations in its maximization step and matrix inversion is not needed. Thus, the proposed methods can highly reduce the computational cost.

### 4.2. The Variable Selection Based on SCAD and MCP Penalties

The variable selection is an important field in high-dimensional data analysis. Using variable selection methods to select significant explanatory variables and remove the insignificant ones can improve the prediction accuracy of the statistical model. In this article, we applied the SCAD and MCP penalties with oracle properties to conduct variable selection in the proportional odds model. The penalized log-likelihood function becomes the new objective to be minimized, which can be written as,

\[
\ell_{\text{pen}}(\beta) = \ell_{\text{obs}}(\beta) - n \sum_{q=1}^{p} \rho(\beta_q | \epsilon, \gamma),
\]

where \( \rho(\cdot | \epsilon, \gamma) \) is the penalized term. Where the SCAD proposed by [35] is

\[
\rho(t | \epsilon) = \epsilon \int_{0}^{t} \min \left\{ 1, \frac{(\gamma - x/\epsilon)}{\gamma - 1} \right\} \mathrm{d}x, t \geq 0, \epsilon \geq 0, \gamma \geq 2,
\]

\( \gamma \) is suggested to take value 3.7 and the MCP proposed by [36] is

\[
\rho(t | \epsilon) = \epsilon \int_{0}^{t} \left( 1 - \frac{x}{\epsilon \gamma} \right) \mathrm{d}x, t \geq 0, \epsilon \geq 0, \gamma > 1,
\]

where \( \gamma \) here takes 3.

In order to handle the singularity around the origin of \( \rho(\beta_q | \epsilon, \gamma) \), [35] suggested to use the quadratic local approximation of the penalized term, which can be written as

\[
-\rho(\beta_q | \epsilon, \gamma) \geq -\rho(\beta_q^{(k)} | \epsilon, \gamma) - \frac{\rho'(\beta_q^{(k)} | \epsilon, \gamma)}{2 \beta_q^{(k)}} \left( \beta_q^{2} - \beta_q^{(k)2} \right).
\]

Thus, the penalized surrogate function can be written as

\[
Q_{\text{pen}}(\beta) = \ell_{\text{obs}}(\beta) - n \sum_{q=1}^{p} \frac{\rho'(\beta_q^{(k)} | \epsilon, \gamma)}{2 \beta_q^{(k)}} \beta_q^{2}.
\]

where \( \ell_{\text{obs}}(\beta) \) can be minorized by Equation (10) or Equation (11).
In order to obtain a good variable selection result, we applied the BIC criterion to select the tuning parameter \( \epsilon \). The likelihood function with BIC criterion suggested by [37] is

\[
BIC = -2\ell(\hat{\Lambda}_0, \hat{\beta}) + q \log(n),
\]

where \( n \) is the sample size and \( q \) denotes the dimension of \( \hat{\beta} \), which is the number of selected non-zero parameters. For the exact process of tuning parameters’ selection, we first find the proper range of \( \lambda \) values by searching from \((0, +\infty)\) using this BIC criteria with an initial sequence \( s_1, s_2, ..., s_k \). The solution path can be plotted using this sequence and \( s_i \) is selected where the minimum BIC is obtained. Then, some grid points are constructed in a range \((s_{i-1}, s_{i+1})\) for a more accurate search where the optimal \( \lambda \) is selected from this sequence with the minimum BIC score.

5. Simulation Study

According to the estimation equation derived in previous sections, we simulate the data to analyze the estimation result at a finite sample size.

5.1. Parameter Estimation of the Proportional Odds Model

**Case 1:** The data is simulated to verify the method given from Section 3. Let \((X_1, X_2, X_3)\) be the covariates, which follow the standard normal distribution and the true parameter \( \beta \) is set to be \( (2, 1, -3)^\top \), \( \Lambda_0(t) = (t/2)^2 \). The censoring times are generated from the uniform distribution \( U(0, b) \) to yield two censoring proportions of about 30% or 50% separately. We take sample size to be 250 and 500, the corresponding censoring rate are 30% and 50%. The simulation is conducted 500 times repeatedly. The BIAS, MSE (mean square error), SD (standard deviation) and median number of iterations (K) are reported in the following Tables 1 and 2.

**Case 2:** Let \( X_1, X_2 \) follows standard normal distribution, \( X_3 \) follows bernoulli distribution with rate 0.5, the true parameters \( \beta \) are set to be \( (2, 1, -1)^\top \), \( \Lambda_0(t) = \log(t + 1) \). The censoring times are also generated from uniform distribution \( U(0, b) \). The simulation result under 50% censoring rate is given by Table 3.

From Tables 1–3, we can observe that the two proposed MM algorithms in Section 3 perform similarly well with small MSEs and SDs at different sample sizes and censoring proportions. The estimation for both the parametric part and nonparametric part are accurate with small estimation bias. Moreover, with the increasing of sample size, the estimation result becomes more stable for both the Profile MM method and Non-profile MM method. In addition, compared with the Non-profile MM methods, the Profile MM algorithm performs much more efficiently with less iteration numbers. From Figure 1, we set sample size as 250; the dotted line can fit the solid line well when the censoring rate is 30% and 50%, respectively. Similarly, from Figure 2, when the right censoring rate is 50%, the curves of the true baseline cumulative hazard function and estimated baseline cumulative hazard function are coincident in both cases, where the sample size is 250 and 500, which indicates the consistency of our estimator. The results demonstrated in the Figures 1 and 2 are consistent with those shown in Tables 1–3, and we can conclude that our proposed methods have an excellent performance in estimating the baseline cumulative hazard function and other parameters, given the high right censoring rate and small sample size.
Table 1. The estimation result of Case 1 with 30% censoring rate.

<table>
<thead>
<tr>
<th>n</th>
<th>Parameter</th>
<th>Profile MM</th>
<th>Non-Profile MM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BIAS</td>
<td>MSE</td>
</tr>
<tr>
<td>250</td>
<td>$\beta_1$</td>
<td>-0.0066</td>
<td>0.0316</td>
</tr>
<tr>
<td></td>
<td>$\beta_2$</td>
<td>-0.0077</td>
<td>0.0204</td>
</tr>
<tr>
<td></td>
<td>$\beta_3$</td>
<td>-0.0001</td>
<td>0.0513</td>
</tr>
<tr>
<td></td>
<td>$\Lambda_0(0.5)$</td>
<td>0.0085</td>
<td>0.0003</td>
</tr>
<tr>
<td></td>
<td>$\Lambda_0(1)$</td>
<td>0.0052</td>
<td>0.0035</td>
</tr>
<tr>
<td>500</td>
<td>$\beta_1$</td>
<td>-0.0068</td>
<td>0.0143</td>
</tr>
<tr>
<td></td>
<td>$\beta_2$</td>
<td>0.0007</td>
<td>0.0089</td>
</tr>
<tr>
<td></td>
<td>$\beta_3$</td>
<td>-0.0154</td>
<td>0.0241</td>
</tr>
<tr>
<td></td>
<td>$\Lambda_0(0.5)$</td>
<td>-0.0002</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>$\Lambda_0(1)$</td>
<td>-0.0028</td>
<td>0.0013</td>
</tr>
</tbody>
</table>

Table 2. The estimation result of Case 1 with 50% censoring rate.

<table>
<thead>
<tr>
<th>n</th>
<th>Parameter</th>
<th>Profile MM</th>
<th>Non-Profile MM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BIAS</td>
<td>MSE</td>
</tr>
<tr>
<td>250</td>
<td>$\beta_1$</td>
<td>-0.0029</td>
<td>0.0462</td>
</tr>
<tr>
<td></td>
<td>$\beta_2$</td>
<td>-0.0021</td>
<td>0.0235</td>
</tr>
<tr>
<td></td>
<td>$\beta_3$</td>
<td>0.0001</td>
<td>0.0682</td>
</tr>
<tr>
<td></td>
<td>$\Lambda_0(0.5)$</td>
<td>-0.0003</td>
<td>0.0003</td>
</tr>
<tr>
<td></td>
<td>$\Lambda_0(1)$</td>
<td>-0.0001</td>
<td>0.0032</td>
</tr>
<tr>
<td>500</td>
<td>$\beta_1$</td>
<td>0.0012</td>
<td>0.0196</td>
</tr>
<tr>
<td></td>
<td>$\beta_2$</td>
<td>-0.0105</td>
<td>0.0116</td>
</tr>
<tr>
<td></td>
<td>$\beta_3$</td>
<td>0.0063</td>
<td>0.0348</td>
</tr>
<tr>
<td></td>
<td>$\Lambda_0(0.5)$</td>
<td>-0.0002</td>
<td>0.0002</td>
</tr>
<tr>
<td></td>
<td>$\Lambda_0(1)$</td>
<td>-0.0017</td>
<td>0.0018</td>
</tr>
</tbody>
</table>

Table 3. The estimation result of Case 2 with 50% censoring rate.

<table>
<thead>
<tr>
<th>n</th>
<th>Parameter</th>
<th>Profile MM</th>
<th>Non-Profile MM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BIAS</td>
<td>MSE</td>
</tr>
<tr>
<td>250</td>
<td>$\beta_1$</td>
<td>0.0147</td>
<td>0.0336</td>
</tr>
<tr>
<td></td>
<td>$\beta_2$</td>
<td>0.0075</td>
<td>0.0207</td>
</tr>
<tr>
<td></td>
<td>$\beta_3$</td>
<td>-0.0328</td>
<td>0.0777</td>
</tr>
<tr>
<td></td>
<td>$\Lambda_0(0.5)$</td>
<td>0.0074</td>
<td>0.0089</td>
</tr>
<tr>
<td></td>
<td>$\Lambda_0(1)$</td>
<td>0.0138</td>
<td>0.0242</td>
</tr>
<tr>
<td>500</td>
<td>$\beta_1$</td>
<td>0.0041</td>
<td>0.0186</td>
</tr>
<tr>
<td></td>
<td>$\beta_2$</td>
<td>-0.0042</td>
<td>0.0120</td>
</tr>
<tr>
<td></td>
<td>$\beta_3$</td>
<td>-0.0015</td>
<td>0.0398</td>
</tr>
<tr>
<td></td>
<td>$\Lambda_0(0.5)$</td>
<td>-0.0015</td>
<td>0.0039</td>
</tr>
<tr>
<td></td>
<td>$\Lambda_0(1)$</td>
<td>-0.0005</td>
<td>0.0115</td>
</tr>
</tbody>
</table>
Figure 1. True and estimated baseline cumulative functions with different censoring rate when sample size is 250. The solid and dotted lines plot the true and estimated baseline cumulative hazard functions, respectively. The estimated baseline cumulative hazard function is the empirical average of the estimated baseline cumulative hazard functions based on 500 replications.

Figure 2. True and estimated baseline cumulative functions with different sample size when censoring rate is 50%. The solid and dotted lines plot the true and estimated baseline cumulative hazard functions, respectively. The estimated baseline cumulative hazard function is the empirical average of the estimated baseline cumulative hazard functions based on 500 replications.
5.2. Simulation on Variable Selection

**Case 3:** We simulate the data to verify the method discussed in Section 4.1. The dimension of the covariates is set to be 10 and they follow standard normal distribution. Let \( \Lambda_0(t) = (t/2)^2 \), the censoring times are generated from uniform distribution \( U(0, b) \) to yield a censoring proportion of about 50%, assume \( \beta = (-4, -4, -2, -2, 1, 1, 3, 3, 5, 5)^\top \). Based on 500 replications, the BIAS, MSE (mean square error), SD (standard deviation) and median number of iterations (K) of the estimated parameters are reported in the Table 4.

<table>
<thead>
<tr>
<th>( n )</th>
<th>Parameter</th>
<th>Profile MM</th>
<th>Non-Profile MM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BIAS</td>
<td>MSE</td>
<td>SD</td>
</tr>
<tr>
<td>200</td>
<td>( \beta_1 )</td>
<td>-0.0711</td>
<td>0.1664</td>
</tr>
<tr>
<td></td>
<td>( \beta_3 )</td>
<td>-0.0208</td>
<td>0.0640</td>
</tr>
<tr>
<td></td>
<td>( \beta_5 )</td>
<td>0.0020</td>
<td>0.0433</td>
</tr>
<tr>
<td></td>
<td>( \beta_7 )</td>
<td>0.0496</td>
<td>0.1134</td>
</tr>
<tr>
<td></td>
<td>( \beta_9 )</td>
<td>0.1008</td>
<td>0.2523</td>
</tr>
<tr>
<td>400</td>
<td>( \beta_1 )</td>
<td>-0.0230</td>
<td>0.0663</td>
</tr>
<tr>
<td></td>
<td>( \beta_3 )</td>
<td>-0.0149</td>
<td>0.0317</td>
</tr>
<tr>
<td></td>
<td>( \beta_5 )</td>
<td>0.0093</td>
<td>0.0177</td>
</tr>
<tr>
<td></td>
<td>( \beta_7 )</td>
<td>0.0139</td>
<td>0.0451</td>
</tr>
<tr>
<td></td>
<td>( \beta_9 )</td>
<td>0.0372</td>
<td>0.0982</td>
</tr>
</tbody>
</table>

**Case 4:** Let \( X_1, X_2, \cdots, X_7, X_8 \) follows standard normal distribution, \( X_9, X_{10} \) follows bernoulli distribution with rate 0.5, \( \Lambda_0(t) = 2t \), parameters \( \beta = (1, 1, 1, -2, -2, -2, -2, -1)^\top \), the simulation result under 30% censoring rate is presented by Table 5.

<table>
<thead>
<tr>
<th>( n )</th>
<th>Parameter</th>
<th>Profile MM</th>
<th>Non-Profile MM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BIAS</td>
<td>MSE</td>
<td>SD</td>
</tr>
<tr>
<td>200</td>
<td>( \beta_5 )</td>
<td>-0.0474</td>
<td>0.0438</td>
</tr>
<tr>
<td></td>
<td>( \beta_9 )</td>
<td>0.0216</td>
<td>0.0977</td>
</tr>
<tr>
<td></td>
<td>( \beta_{10} )</td>
<td>0.0060</td>
<td>0.0899</td>
</tr>
<tr>
<td>400</td>
<td>( \beta_5 )</td>
<td>-0.0178</td>
<td>0.0193</td>
</tr>
<tr>
<td></td>
<td>( \beta_{10} )</td>
<td>-0.0084</td>
<td>0.0411</td>
</tr>
</tbody>
</table>

From the simulation results demonstrated in Tables 4 and 5, the proposed parameter separated MM method in Section 4.1 can estimate the parameters accurately with small estimation bias and we also observe that the MSE, SD and K decrease with the sample size increases. Different from the results in Tables 1–3, the non-profile MM method has fewer iterations than profile MM method in Tables 4 and 5.

**Case 5:** In this part, we illustrate the utility of the proposed MM method for the regularized estimation in the sparse high-dimensional proportional odds regression model with 10 covariates \((X_1, \cdots, X_{10})^\top\). Let \( \Lambda_0(t) = (t/2)^2 \), the censoring times are generated from uniform distribution \( U(0, b) \) to yield a censoring proportion of about 30%, where the marginal distribution of \((X_1, \cdots, X_{10})^\top\) is the standard normal with correlation \( \rho = 0.2 \). Assume the true value of \( \beta \) is \((-2, 1, 3, 0, 0, 0, 0, 0, 0, 0)^\top\), where \( \beta_1, \beta_2, \beta_3 \) are non-zero parameters. The simulation is repeated 500 times.
In order to assess the effectiveness of our methods, we calculate the RMSE to test the average difference between the true and estimated parameters.

\[
RMSE = \sqrt{\frac{1}{p} \sum_{j=1}^{p} (\beta_j - \hat{\beta}_j)^2},
\]

where \( p \) is the number of explanatory variables. Moreover, we calculate the FDR (false discovery rate) and PSR (positive select rate) proposed by [38]

\[
FDR = \begin{cases} 
\frac{FP}{TP + FP}, & TP + FP > 0, \\
0, & TP + FP = 0,
\end{cases}
\]

\[
PSR = \frac{TP}{m},
\]

where FP (false positive) is the number of parameters, which are estimated to be non-zero with true value equals to zero, TP (true positive) denotes the correctly excluded insignificant parameters, and \( m \) denotes the number of non-zero parameters. Thus, the low FDR or high PSR indicates the good parameter selection result.

According to Table 6, the regularized estimation methods proposed in this paper can correctly select the significant parameters. With the increase in sample size, we can observe that the parameter selection results perform better. Under the same framework, the results produced by Profile MM method and Non-profile MM method are similar as well. In both cases, where the sample size is 200 and 400, no parameter with true value equals to 0 is selected. For the selection of true parameters, both penalties generate good results with PSR greater than 0.9, while SCAD performs slightly better than MCP. As presented by Table 7, the estimation of BIAS, MSE and SD are small for non-zero parameters, which indicates a good estimation performance. That is, the MM method can effectively deal with the parameter selection for the proportional odds model under SCAD and MCP penalties.

<table>
<thead>
<tr>
<th>( n )</th>
<th>Index</th>
<th>Profile MM</th>
<th>Non-Profile MM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SCAD</td>
<td>MCP</td>
</tr>
<tr>
<td>200</td>
<td>FDR</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>PSR</td>
<td>0.9927</td>
<td>0.9387</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.1205</td>
<td>0.1480</td>
</tr>
<tr>
<td>400</td>
<td>FDR</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>PSR</td>
<td>0.9987</td>
<td>0.9740</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.0794</td>
<td>0.0927</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>( n )</th>
<th>( \beta )</th>
<th>BIAS</th>
<th>MSE</th>
<th>SD</th>
<th>BIAS</th>
<th>MSE</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCAD</td>
<td>200</td>
<td>( \beta_1 )</td>
<td>0.0242</td>
<td>0.0475</td>
<td>0.2167</td>
<td>0.0279</td>
<td>0.0495</td>
<td>0.2210</td>
</tr>
<tr>
<td></td>
<td>( \beta_2 )</td>
<td>-0.0639</td>
<td>0.1089</td>
<td>0.3242</td>
<td>-0.0510</td>
<td>0.0990</td>
<td>0.3108</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \beta_3 )</td>
<td>-0.0215</td>
<td>0.0628</td>
<td>0.2501</td>
<td>-0.0270</td>
<td>0.0632</td>
<td>0.2502</td>
<td></td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>( \beta_1 )</td>
<td>0.0143</td>
<td>0.0295</td>
<td>0.1714</td>
<td>0.0026</td>
<td>0.0207</td>
<td>0.1440</td>
</tr>
<tr>
<td></td>
<td>( \beta_2 )</td>
<td>-0.0228</td>
<td>0.0381</td>
<td>0.1941</td>
<td>-0.0127</td>
<td>0.0276</td>
<td>0.1658</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \beta_3 )</td>
<td>-0.0007</td>
<td>0.0357</td>
<td>0.1892</td>
<td>-0.0027</td>
<td>0.0285</td>
<td>0.1692</td>
<td></td>
</tr>
<tr>
<td>MCP</td>
<td>200</td>
<td>( \beta_1 )</td>
<td>0.0707</td>
<td>0.0692</td>
<td>0.2537</td>
<td>0.0971</td>
<td>0.0702</td>
<td>0.2468</td>
</tr>
<tr>
<td></td>
<td>( \beta_2 )</td>
<td>-0.1722</td>
<td>0.2078</td>
<td>0.4225</td>
<td>-0.2093</td>
<td>0.2544</td>
<td>0.4594</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \beta_3 )</td>
<td>-0.0619</td>
<td>0.0707</td>
<td>0.2588</td>
<td>0.0571</td>
<td>0.0637</td>
<td>0.2462</td>
<td></td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>( \beta_1 )</td>
<td>0.0376</td>
<td>0.0388</td>
<td>0.1935</td>
<td>0.0504</td>
<td>0.0358</td>
<td>0.1827</td>
</tr>
<tr>
<td></td>
<td>( \beta_2 )</td>
<td>-0.0735</td>
<td>0.0894</td>
<td>0.2902</td>
<td>-0.1179</td>
<td>0.1421</td>
<td>0.3584</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \beta_3 )</td>
<td>-0.0348</td>
<td>0.0334</td>
<td>0.1796</td>
<td>-0.0262</td>
<td>0.0329</td>
<td>0.1796</td>
<td></td>
</tr>
</tbody>
</table>
6. Real Data Analysis

We apply our methods to the Veterans’ administration lung cancer study data with sample size 137, and the number of covariate is 8. The data can be retrieved from the R package “survival” and the information for the covariates are given in Table 8.

Table 8. Covariates of lung cancer data.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Detailed Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trt</td>
<td>treatment, 1 = Standard 2 = test</td>
</tr>
<tr>
<td>Celltype</td>
<td>Squamous, smallcell, adeno, large</td>
</tr>
<tr>
<td>Time</td>
<td>Survival time</td>
</tr>
<tr>
<td>Status</td>
<td>Censoring status</td>
</tr>
<tr>
<td>Karno</td>
<td>Karnofsky performance score (100 = good)</td>
</tr>
<tr>
<td>Diagtime</td>
<td>Months from diagnosis to randomisation</td>
</tr>
<tr>
<td>Age</td>
<td>In years</td>
</tr>
<tr>
<td>Prior</td>
<td>Prior therapy (0 = no, 10 = yes)</td>
</tr>
</tbody>
</table>

This dataset is generally applied for the illustration of the proportional odds model. For the purpose of comparing against other studies, we use the data of patient with no prior therapy where the censoring rate is 6.19% and the covariates are “celltype” and “karno”. Two MM algorithms are applied to estimate the parameters. Moreover, the standard deviation is estimated by 1000 times of bootstraps, where $\hat{\beta}^*_g$, $g = 1, \ldots, G$ is the estimate of $g^{th}$ bootstrap. We construct the 100(1 – $\alpha$)% confidence interval of $\beta$ using normal approximation.

\[
\left( \hat{\beta}^* \pm z_{\alpha/2} \hat{se}^*(\hat{\beta}) \right),
\]

(12)

where $\hat{\beta}^* = (1/G) \sum_{g=1}^{G} \hat{\beta}^*_g$ and

\[
\hat{se}^*(\hat{\beta}) = \sqrt{\frac{1}{(G-1)} \sum_{g=1}^{G} (\hat{\beta}^*_g - \hat{\beta}^*)^2}
\]

(13)

bootstrap confidence interval of 100(1 – $\alpha$)% is

\[
\left( \hat{\beta}^L, \hat{\beta}^U \right),
\]

(14)

where $\hat{\beta}^L$ and $\hat{\beta}^U$ denotes the quantile of $\left\{ \hat{\beta}^*_1, \ldots, \hat{\beta}^*_G \right\}$’s $\alpha/2$ and $1 - \alpha/2$ separately.

Tables 9 and 10 present the estimation result of two MM algorithms, where the standard estimated error (SE) is defined by Equation (13) and the 95% confidence interval (CI1) is defined by Equation (12). Moreover, the 95% confidence interval (CI2) is defined by Equation (14). In addition, the estimated cumulative hazard function are plotted in Figure 3.

Table 9. The estimation result from profile MM method.

<table>
<thead>
<tr>
<th>Variable</th>
<th>MLE</th>
<th>SE</th>
<th>CI1</th>
<th>CI2</th>
</tr>
</thead>
<tbody>
<tr>
<td>karno</td>
<td>-0.0532</td>
<td>0.0105</td>
<td>[-0.0741, -0.0329]</td>
<td>[-0.0714, -0.0363]</td>
</tr>
<tr>
<td>squamous vs large</td>
<td>-0.1814</td>
<td>0.6382</td>
<td>[-1.4255, 1.0761]</td>
<td>[-1.2761, 0.8135]</td>
</tr>
<tr>
<td>small vs large</td>
<td>1.3827</td>
<td>0.4816</td>
<td>[0.4820, 2.3699]</td>
<td>[0.6734, 2.2438]</td>
</tr>
<tr>
<td>adenovs large</td>
<td>1.3138</td>
<td>0.4691</td>
<td>[0.4521, 2.2910]</td>
<td>[0.6341, 2.1755]</td>
</tr>
</tbody>
</table>
Table 10. The estimation result from non-profile MM method.

<table>
<thead>
<tr>
<th>Variable</th>
<th>MLE</th>
<th>SE</th>
<th>CI1</th>
<th>CI2</th>
</tr>
</thead>
<tbody>
<tr>
<td>karno</td>
<td>−0.0532</td>
<td>0.0106</td>
<td>[−0.0746, −0.0329]</td>
<td>[−0.0722, −0.0377]</td>
</tr>
<tr>
<td>squamous vs large</td>
<td>−0.1814</td>
<td>0.6589</td>
<td>[−1.4696, 1.1134]</td>
<td>[−1.2175, 0.8777]</td>
</tr>
<tr>
<td>small vs large</td>
<td>1.3827</td>
<td>0.5105</td>
<td>[0.4431, 2.4442]</td>
<td>[0.6459, 2.2644]</td>
</tr>
<tr>
<td>adeno vs large</td>
<td>1.3138</td>
<td>0.4659</td>
<td>[0.4654, 2.2919]</td>
<td>[0.6267, 2.1613]</td>
</tr>
</tbody>
</table>

Figure 3. Estimated baseline cumulative functions for lung cancer data.

From Tables 9 and 10, for the same patient, every increase of 1 degree of “karno”, the decrease in possibility of death is $\exp(-0.0532) = 0.9482$. From Figure 3, the estimation results of the baseline cumulative hazard function of two MM methods are almost the same. We can find that our method produces a similar result, as concluded by [3], and is a little bit different from the result presented by [5,7,11], which is shown by Table 11 from [11].

Table 11. Estimation result from other studies.

<table>
<thead>
<tr>
<th>Variable</th>
<th>(Bennett, [3])</th>
<th>(Pettitt, [5])</th>
<th>(Murphy et al., [7])</th>
<th>(Lam &amp; Leung, [11])</th>
</tr>
</thead>
<tbody>
<tr>
<td>karno</td>
<td>−0.053</td>
<td>−0.055</td>
<td>−0.055</td>
<td>−0.053</td>
</tr>
<tr>
<td>squamous vs large</td>
<td>−0.181</td>
<td>−0.177</td>
<td>−0.217</td>
<td>−0.247</td>
</tr>
<tr>
<td>small vs large</td>
<td>1.383</td>
<td>1.438</td>
<td>1.440</td>
<td>1.367</td>
</tr>
<tr>
<td>adeno vs large</td>
<td>1.314</td>
<td>1.302</td>
<td>1.339</td>
<td>1.316</td>
</tr>
</tbody>
</table>

Then, we consider all eight covariates for model fitting where the censoring rate is 6.57%. We first use the method from Section 4.1 to estimate the parameters. After obtaining the estimation results, we apply the SCAD and MCP penalties discussed in Section 4.2 for parameter selection, and the result is presented by Table 12.

From Table 12, the Profile MM method and Non-profile MM method produce similar results. Without considering the penalties, the results of parameter separated MM algorithms demonstrate that four parameters including “small vs large”, “adenov vs large”, “karno” and “prior” significantly affect the death rate. After introducing the penalties, we conduct the variable selection and 5 covariates are shrunk to 0. Only three significant parameters are preserved, which are “small vs large”, “adenov vs large” and “karno”. “small vs large” and “adenov vs large” lead to a positive effect to the death rate while “karno” is negatively related to the death rate, which is in line with the common sense in reality.
Table 12. The result of variable selection for lung cancer data.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Profile MM</th>
<th>Non-Profile MM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLE</td>
<td>SCAD</td>
</tr>
<tr>
<td>trt</td>
<td>−0.0141</td>
<td>0</td>
</tr>
<tr>
<td>squamous vs large</td>
<td>−0.0348</td>
<td>0</td>
</tr>
<tr>
<td>small vs large</td>
<td>1.2412</td>
<td>1.1960</td>
</tr>
<tr>
<td>adeno vs large</td>
<td>1.3251</td>
<td>1.3653</td>
</tr>
<tr>
<td>karno</td>
<td>−0.0597</td>
<td>−0.0582</td>
</tr>
<tr>
<td>diagtime</td>
<td>−0.0025</td>
<td>0</td>
</tr>
<tr>
<td>age</td>
<td>−0.0141</td>
<td>0</td>
</tr>
<tr>
<td>prior</td>
<td>−0.1663</td>
<td>0</td>
</tr>
</tbody>
</table>

7. Conclusions and Future Work

The proportional odds models are more competitive than proportional hazards models in dealing with right-censored survival data, where mortality tends to be uniform over time. The MM algorithm has the advantages of simple structure, strong interpretability and easy implementation. Hence, it is a useful tool for optimization problems and has a broad range of applications in statistics. In this work, we introduce the MM algorithm into the estimation of the proportional odds model. We first develop two MM algorithms for the estimation of proportional odds models, which greatly simplify the estimation process by constructing two simple surrogate functions for the log-likelihood function. The proposed MM algorithms successfully separate the parameters and decompose the high-dimensional maximization into separated low-dimensional ones, which may avoid the matrix inversion and can be used to more general scenarios. Moreover, we apply the MM methods to the regularized estimation in sparse high-dimensional proportional odds regression models with SCAD and MCP penalties. We find that the proposed MM algorithms with the property of separating parameters can mesh well with the SCAD and MCP penalties, which yield good results in simultaneous parameter estimation and variable selection.

The advantage of our algorithm is that we separate the estimation of the baseline hazard and other parameters, which makes the estimation process more efficient. In future studies, such technique derived from the semi-parametric model can be further extended to the application of parameter estimation for fully non-parametric models. In addition, as we mentioned in previous sections, the existing methods for the proportional odds model with right censored data involve high computation complexity when dealing with high-dimensional data. However, our proposed algorithm can help to avoid matrix inversion, which is capable of high-dimensional regression analysis. Furthermore, the advantage of our algorithm is that it can mesh well with the existing quasi-Newton acceleration and other simple off-the-shelf accelerators to further boost the estimation process. Although our proposed MM algorithms are developed for the proportional odds models, a parallel approach can essentially be developed for the more general transformation models. We will investigate this in our future work.

Author Contributions: Data curation, C.X., T.J. and J.L.; Formal analysis, X.H., C.X. and J.X.; Investigation, T.J. and J.L.; Methodology, X.H. and J.X.; Project administration, X.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. McCullagh, P. A logistic model for paired comparisons with ordered categorical data. Biometrika 1977, 64, 449–453. [CrossRef]