Mixture Modeling of Time-to-Event Data in the Proportional Odds Model

Huang, Xifen; Xiong, Chaosong; Xu, Jinfeng; Shi, Jianhua; Huang, Jinhong

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/math10183375

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: 15/05/2024
Article

Mixture Modeling of Time-to-Event Data in the Proportional Odds Model

Xifen Huang 1, Chaosong Xiong 1, Jinfeng Xu 2,*, Jianhua Shi 2 and Jinhong Huang 2

1 School of Mathematics, Yunnan Normal University, Kunming 650092, China
2 School of Mathematics, Minnan Normal University, Zhangzhou 363000, China
* Correspondence: sjf@mnnu.edu.cn

Abstract: Subgroup analysis with survival data are most essential for detailed assessment of the risks of medical products in heterogeneous population subgroups. In this paper, we developed a semiparametric mixture modeling strategy in the proportional odds model for simultaneous subgroup identification and regression analysis of survival data that flexibly allows the covariate effects to differ among several subgroups. Neither the membership or the subgroup-specific covariate effects are known a priori. The nonparametric maximum likelihood method together with a pair of MM algorithms with monotone ascent property are proposed to carry out the estimation procedures. Then, we conducted two series of simulation studies to examine the finite sample performance of the proposed estimation procedure. An empirical analysis of German breast cancer data is further provided for illustrating the proposed methodology.

Keywords: heterogeneous covariate effects; mixture of proportional odds model; MM algorithm; nonparametric maximum likelihood

MSC: 62N01; 62N02

1. Introduction

In some clinical trials, a substantial proportion of patients respond favorably to a new treatment while the others may eventually relapse. Subgroup analyses aim to classify the patients into a few homogeneous groups and tailor a disease treatment specifically for each subgroup to optimize the treatment effect. In recent years, subgroup identification has received increasing attention in a wide range of fields such as clinical trials, public management, econometrics, and social science. For example, Refs. [1,2] conducted subgroup analysis in econometrics and marketing, while Refs. [3,4] implemented the subgroup analysis in epidemiology and biology, respectively.

Statistical methods for subgroup analysis have also been greatly developed recently. Among them, a finite mixture model has been recognized as an important tool and has been widely used for analyzing data from a heterogeneous population [5]. For example, there are many studies on the Gaussian mixture model for data clustering and classification [6–8]. Ref. [9] introduced a structured logistic-normal mixture model to identify subgroups in randomized clinical trials with differential treatment effects. Refs. [10,11] extended the mixture model-based approach to generalized linear models. Bayesian approaches for mixture regression models are studied by [12]. Moreover, nonparametric mixture models have also been under study in recent years. Ref. [13] proposed a nonparametric mixture model for cure rate estimation. Ref. [14] studied a semiparametric accelerated failure time mixture model for estimation of a biological treatment effect on a latent subgroup of interest in randomized clinical trials. Ref. [15] proposed a semiparametric Logistic–Cox mixture model for subgroup analysis when the interested outcome is event time with right censoring.
Mixture models are deeply connected to the expectation–maximization (EM) algorithm. The EM algorithm is a popular approach for maximum likelihood estimation in incomplete data problems, of which finite mixtures are canonical examples because the unobserved labels of the individuals (as in unsupervised clustering) give a direct interpretation of missing data [16]. Actually, the EM algorithm is a special member of the general family of MM algorithms [17]. The MM algorithm possesses great flexibility in solving optimization problems because the basic idea of MM algorithm is to convert a difficult optimization problem into a series of simpler ones. The MM algorithm has been a powerful tool for optimization problems and enjoys its greatest vogue in computational statistics. Thus far, the MM algorithm has been widely used in many statistical optimization problems. We can find applications of MM principle in a broad range of statistical contexts, including the Bradley–Terry model [18], quantile regression [19], variable selection [20,21], the proportional odds model [22], the shared frailty model [23], distance majorization [24] and so on. The key property of MM principle is that it can decompose a high-dimensional objective function into separable low-dimensional functions by the construction of surrogate function. In this paper, we introduce the general MM principle to the semiparametric mixture of proportional odds model [22], the shared frailty model [23], distance majorization [24] and so on. MM algorithms [17]. The MM algorithm possesses great flexibility in solving optimization problems and enjoys its greatest vogue in computational statistics. Thus far, the MM algorithm has been widely used in many statistical optimization problems. We can find applications of MM principle in a broad range of statistical contexts, including the Bradley–Terry model [18], quantile regression [19], variable selection [20,21], the proportional odds model [22], the shared frailty model [23], distance majorization [24] and so on. The key property of MM principle is that it can decompose a high-dimensional objective function into separable low-dimensional functions by the construction of surrogate function. In this paper, we introduce the general MM principle to the semiparametric mixture of proportional odds model for simultaneous subgroup identification and regression analysis. The rest of the paper is organized as follows. We first review the MM algorithm in Section 2. In Section 3, we present the latent proportional odds model and develop a pair of estimation procedures for the proposed model using the MM algorithm. In Section 4, we provide two parts of simulation studies to select the number of subgroups and assess the finite-sample performances of the proposed methods. We further provide an application of the German breast cancer study data to illustrate the practical utilities of the proposed methods in Section 5.

2. MM Principle

The MM algorithm is an important and powerful tool for optimization problems and enjoys its greatest vogue in computational statistics. For example, \( \ell(\alpha|Y_{\text{obs}}) \) is the objective log-likelihood function, \( \alpha = (a_1, \ldots, a_q)^T \in \Theta \) are the vector of parameters to be estimated, and \( \Theta \) is the parameter space. The maximum likelihood estimate of \( \alpha \) is \( \hat{\alpha} = \arg \max_{\alpha \in \Theta} \ell(\alpha|Y_{\text{obs}}) \). The MM principle provides a general frame for constructing iterative algorithms with monotone convergence, which involves double duty. In maximization problems, the first M stands for minorize and the second M for maximize. The minorization step first constructs a surrogate function \( Q(\alpha|\alpha^{(k)}) \) such that

\[
Q(\alpha|\alpha^{(k)}) \leq \ell(\alpha|Y_{\text{obs}}), \forall \alpha, \alpha^{(k)} \in \Theta, Q(\alpha^{(k)}|\alpha^{(k)}) = \ell(\alpha^{(k)}|Y_{\text{obs}}),
\]

where \( \alpha^{(k)} \) denotes the current estimate of \( \alpha \) in the \( k \)-th iteration. The maximization step then updates \( \alpha^{(k)} \) by \( \alpha^{(k+1)} \), which maximizes the surrogate function \( Q(\cdot|\alpha^{(k)}) \) instead of \( \ell(\alpha|Y_{\text{obs}}) \), that is,

\[
\alpha^{(k+1)} = \arg \max_{\alpha \in \Theta} Q(\alpha|\alpha^{(k)}).
\]

Since

\[
\ell(\alpha^{(k+1)}|Y_{\text{obs}}) \geq Q(\alpha^{(k+1)}|\alpha^{(k)}) \geq Q(\alpha^{(k)}|\alpha^{(k)}) = \ell(\alpha^{(k)}|Y_{\text{obs}}),
\]

the constructed MM algorithm can increase the objective function at each iteration and possess the ascent property driving the objective optimization function \( \ell(\alpha|Y_{\text{obs}}) \) uphill.

3. Proportional Odds Model with Individual-Specific Covariate Effects

Let \( T \) be time to event. The proportional odds model postulates that

\[
\lambda_i(t | X) = \frac{\lambda_0(t) \exp(X_i^\top \beta)}{1 + \Lambda_0(t) \exp(X_i^\top \beta)},
\]
where $\lambda_i(t)$ is the hazard function of $T_i$ given the covariates $X_i$. Let the conditional survival function of $T$ be $S(t|X) = P(T > t|X)$. We know that $\lambda(t|X) = -\frac{d}{dt} \log S(t|X)$. In the proportional odds model, $\beta$ is the regression coefficients, quantifying the effect of the covariates $X$ on the time to event $T$ through the conditional hazard function. It is assumed to be the same for all subjects in the population. In practice, however, subjects may come from different subgroups, the covariate effects may differ and therefore it is more appropriate to assume the following proportional odds model with individual-specific covariate effects:

$$\lambda_i(t | X) = \frac{\lambda_0(t) \exp(X_i^\top \beta_i)}{1 + \Lambda_0(t) \exp(X_i^\top \beta_i)}.$$

In this model, we assume that the covariate effects $\beta_i$ for the subject $i$ may differ. For both parsimony and better interpretation, it is reasonable to assume that $\beta_i = \beta_{0,m}$ with probability $\pi_m$, $m = 1, ..., M$. In other words, there are only $M$ different subgroups for the covariate effects $\beta_i$, where $\beta_{0,m}$, $m = 1, ..., M$ are $M$ different regression coefficients. It is of our interest to estimate the number of groups $M$, $\beta_{0,m}$, $m = 1, ..., M$ and $\pi_m$, $m = 1, ..., M$. Note that $\sum_{m=1}^M \pi_m = 1$.

### 3.1. Heterogeneity Regression Pursuit via MM Algorithm

The joint density function of $(T, \delta)$ can be written as

$$f(t, \delta|X) = \sum_{m=1}^M \pi_m f_m(t, \delta|X),$$

where

$$f_m(t, \delta|X) = \left\{ \frac{\lambda_0(t) \exp(X^\top \beta_m)}{1 + \Lambda_0(t) \exp(X^\top \beta_m)} \right\}^\delta \frac{1}{1 + \Lambda_0(t) \exp(X^\top \beta_m)}$$

denotes the density function of the $m$-th subgroup, $m = 1, 2, ..., M$, $\beta_m$ is the corresponding effect parameter of $X$ in the $m$-th subgroup. Given the observed data $Y_{obs} = \{ (t_i)_{i=1}^n, \{d_i\}_{i=1}^n, \{X_i\}_{i=1}^n \}$, we have the observed log-likelihood function as

$$\ell(\Lambda_0, \beta, \pi|Y_{obs}) = \sum_{i=1}^n \log \left\{ \sum_{m=1}^M \pi_m f_m(t_i, \delta_i|X_i) \right\},$$

where $\Lambda_0(t) = \sum_{i=1}^n I(t_i \leq t) \lambda_0(t_i)$, $\beta = (\beta_1^\top, ..., \beta_M^\top)^\top$, $\pi = (\pi_1, ..., \pi_M)$. Given the parameters in the $k$-th iteration and denoting

$$\nu_{mi}^{(k)} = \frac{\pi_m^{(k)} \cdot f_m^{(k)}(t_i, \delta_i|X_i)}{\sum_{k=1}^K \pi_m^{(k)} \cdot f_m^{(k)}(t_i, \delta_i|X_i)},$$

then we can rewrite $\ell(\Lambda_0, \beta, \pi|Y_{obs})$ as

$$\ell(\Lambda_0, \beta, \pi|Y_{obs}) = \sum_{i=1}^n \log \left\{ \sum_{m=1}^M \nu_{mi}^{(k)} \cdot \frac{\pi_m \cdot f_m(t_i, \delta_i|X_i)}{\nu_{mi}^{(k)}} \right\}. \tag{2}$$

By the continuous version of Jensen’s inequality as $\varphi(\int f(x) \cdot g(x) dx) \geq \int f(x) \cdot \varphi(g(x)) dx$, we can transfer the function $\varphi(\cdot)$ outside the integral to the inside of the integral, where $g(x)$ is a density function. Inspired by this feature, we construct a density
The surrogate function $v_{mi}^{(k)}$ in Equation (2) which plays the role of function $g(x)$, the rest of the part

$$
\pi_m \cdot f_m(t_i, \delta_i|X_i) / v_{mi}^{(k)}
$$

plays the role of function $f(x)$. By the following calculation,

$$
\sum_{i=1}^{n} \log \left\{ \frac{\sum_{m=1}^{M} v_{mi}^{(k)} \cdot \frac{\pi_m \cdot f_m(t_i, \delta_i|X_i)}{v_{mi}^{(k)}}}{\sum_{i=1}^{n} \sum_{m=1}^{M} v_{mi}^{(k)} \cdot \{ \log \pi_m + \log f_m(t_i, \delta_i|X_i) \}} \right\} 
$$

the logarithmic function on the outside is transferred to the inside of the integral, which breaks down the product terms into a summation. Hence, we construct the surrogate function for $\ell(\Lambda_0, \beta, \pi|Y_{obs})$ as

$$
Q(\Lambda_0, \beta, \pi|\Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)}) = \sum_{i=1}^{n} \sum_{m=1}^{M} v_{im}^{(k)} \cdot \{ \log \pi_m + \log f_m(t_i, \delta_i|X_i) \},
$$

where

$$
Q(\pi|\Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)}) = \sum_{i=1}^{n} \sum_{m=1}^{M} v_{im}^{(k)} \cdot \log \pi_m,
$$

(3)

and

$$
Q(\Lambda_0, \beta|\Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)})
$$

$$
= \sum_{i=1}^{n} \sum_{m=1}^{M} v_{im}^{(k)} \cdot \log f_m(t_i, \delta_i|X_i),
$$

(4)

$$
= \sum_{i=1}^{n} \delta_i \log \lambda_0(t_i) + \sum_{i=1}^{n} \sum_{m=1}^{M} v_{im}^{(k)} X_i^\top \beta_m - \sum_{i=1}^{n} \sum_{m=1}^{M} v_{im}^{(k)} (\delta_i + 1) \log \left[ 1 + \Lambda_0(t_i) \exp(X_i^\top \beta_m) \right].
$$

The surrogate function $Q(\Lambda_0, \beta, \pi|\Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)})$ separates the parameters $\pi$ and $(\Lambda_0, \beta)$ into (3) and (4), respectively. All the parameters $\{\pi_m\}_{m=1}^{M}$ in (3) are separated from each other so that updating $\pi_m$ is as straightforward as

$$
\hat{\pi}_m = \frac{\sum_{i=1}^{n} v_{im}^{(k)}}{n}, \quad m = 1, \ldots, M.
$$

(5)

To update $(\Lambda_0, \beta)$, we apply the supporting hyperplane inequality to Equation (4) to release the object $x$ from the logarithmic function,

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

we have

$$
- \log \left[ 1 + \Lambda_0(t_i) \exp(X_i^\top \beta_m) \right] \geq - \log(A_{im}^{(k)}) - \frac{1 + \Lambda_0(t_i) \exp(X_i^\top \beta_m) - A_{im}^{(k)}}{A_{im}^{(k)}},
$$

where $A_{im}^{(k)} = 1 + \Lambda_0^{(k)}(t_i) \exp(X_i^\top \beta_m^{(k)})$. Then, we obtain the following surrogate function for $Q(\Lambda_0, \beta|\Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)})$,

$$
Q_1(\Lambda_0, \beta|\Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)})
$$

$$
= \sum_{i=1}^{n} \delta_i \log \lambda_0(t_i) + \sum_{i=1}^{n} \sum_{m=1}^{M} v_{im}^{(k)} X_i^\top \beta_m - \sum_{i=1}^{n} \sum_{m=1}^{M} v_{im}^{(k)} (\delta_i + 1) \frac{\Lambda_0(t_i) \exp(X_i^\top \beta_m)}{A_{im}^{(k)}}.
$$
3.2. Profile MM Method

Following [25,26], we consider the profile estimation approach and first profile out $\Lambda_0$ in $Q_1(\Lambda_0, \beta | \Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)})$ for any given $\beta$. This leads to the estimate of $\Lambda_0$ given $\beta$ as

$$\hat{\Lambda}_0(t_i) = \frac{\delta_i}{\sum_{j=1}^n I(t_j \geq t_i) \sum_{m=1}^M v_{jm}^{(k)} (\delta_j + 1) \exp(X_j^\top \beta_m) / A_{jm}^{(k)}}. \quad (6)$$

Substituting (6) into $Q_1(\Lambda_0, \beta | \Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)})$ yields the function

$$Q_2(\beta | \Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)})$$

$$= \sum_{i=1}^n \sum_{m=1}^M v_{im}^{(k)} (\delta_i X_i^\top \beta_m) - \sum_{i=1}^n \delta_i \log \left[ \sum_{j=1}^n I(t_j \geq t_i) \sum_{m=1}^M v_{jm}^{(k)} (\delta_j + 1) \exp(X_j^\top \beta_m) / A_{jm}^{(k)} \right].$$

We use the supporting hyperplane inequality again to deal with $Q_2(\beta | \Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)})$, then we obtain the following $Q_3(\beta | \Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)})$ where all $\beta_m (m = 1, \ldots, M)$ are separated from each other,

$$Q_3(\beta | \Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)})$$

$$= \sum_{m=1}^M \left\{ \sum_{i=1}^n v_{im}^{(k)} (\delta_i X_i^\top \beta_m) - \sum_{i=1}^n \delta_i \log \left[ \sum_{j=1}^n I(t_j \geq t_i) v_{jm}^{(k)} (\delta_j + 1) \exp(X_j^\top \beta_m) / A_{jm}^{(k)} \right] \right\}$$

$$= \sum_{m=1}^M Q_3(\beta_m | \Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)}),$$

where $B_i^{(k)} = \sum_{j=1}^n I(t_j \geq t_i) \sum_{m=1}^M v_{jm}^{(k)} (\delta_j + 1) \exp(X_j^\top \beta_m) / A_{jm}^{(k)}$. Finally, the estimate of each $\beta_m$ can be obtained by one step Newton iteration.

3.3. Non-Profile MM Method

For the above profile MM method, the estimate of $\Lambda_0$ is highly related to the estimate of $\beta$ because we treat nonparametric component $\Lambda_0$ as a function of $\beta$ in the profile step. Inspired by the parameter-separable property of the MM principle, we further separate the nonparametric part $\Lambda_0$ with $\beta$ according to the decomposition rules. That is, we use the following inequality of arithmetic and geometric means to the function $Q_1(\Lambda_0, \beta | \Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)})$ as

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

Here, we let $x_1 = \Lambda_0(t_i) / \Lambda_0^{(k)}(t_i)$ and $x_2 = \exp(X_i^\top \beta_m) / \exp(X_i^\top \beta_m^{(k)})$, then we have

$$-\frac{\Lambda_0(t_i) \exp(X_i^\top \beta_m)}{\Lambda_0^{(k)}(t_i) \exp(X_i^\top \beta_m^{(k)})} \geq -\frac{\Lambda_0^{(k)}(t_i)}{2\Lambda_0^{(k)}(t_i)} - \frac{\exp(2X_i^\top \beta_m^{(k)})}{2\exp(X_i^\top \beta_m^{(k)})}.$$

That is,

$$-\Lambda_0(t_i) \exp(X_i^\top \beta_m) \geq -\frac{\exp(X_i^\top \beta_m^{(k)})}{2\Lambda_0^{(k)}(t_i)} \Lambda_0^{(k)}(t_i) - \frac{\Lambda_0^{(k)}(t_i)}{2\exp(X_i^\top \beta_m^{(k)})} \exp(2X_i^\top \beta_m).$$
Substituting the above inequality back to $Q_4(\Lambda_0, \beta | \Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)})$, we may obtain

$$Q_4(\Lambda_0, \beta | \Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)})$$

$$= \sum_{i=1}^{n} \delta_i \log \Lambda_0(t_i) + \sum_{i=1}^{n} \sum_{m=1}^{M} v_{im}^{(k)} \delta_i X_i^T \beta_m$$

$$- \sum_{i=1}^{n} \sum_{m=1}^{M} v_{im}^{(k)} (\delta_i + 1) \left[ \exp(X_i^T \beta_m^{(k)}) / 2 \Lambda_0^{(k)}(t_i) \right] \sum_{i=1}^{n} \sum_{m=1}^{M} v_{im}^{(k)} (\delta_i + 1) \exp(2X_i^T \beta_m) / A_{im}^{(k)}$$

$$= Q_4(\Lambda_0 | \Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)}) + Q_4(\beta | \Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)})$$

where

$$Q_4(\Lambda_0 | \Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)}) = \sum_{i=1}^{n} \delta_i \log \Lambda_0(t_i) - \sum_{i=1}^{n} \sum_{m=1}^{M} v_{im}^{(k)} (\delta_i + 1) \left[ \exp(X_i^T \beta_m^{(k)}) / 2 \Lambda_0^{(k)}(t_i) \right]$$

$$\left( \sum_{i=1}^{n} \sum_{m=1}^{M} v_{im}^{(k)} (\delta_i + 1) \exp(2X_i^T \beta_m) / A_{im}^{(k)} \right)$$

and

$$Q_4(\beta | \Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)})$$

$$= \sum_{i=1}^{n} \sum_{m=1}^{M} v_{im}^{(k)} \delta_i X_i^T \beta_m - \sum_{i=1}^{n} \sum_{m=1}^{M} v_{im}^{(k)} (\delta_i + 1) \Lambda_0^{(k)}(t_i) \exp(2X_i^T \beta_m) / A_{im}^{(k)}.$$

It is observed that the parameters $\Lambda_0$ and $\beta_m$ are completely separated, then the corresponding parameter estimators can be obtained by differentiating them separately. Letting $\partial Q_4(\Lambda_0 | \Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)}) / \partial \Lambda_0 = 0$, we obtain the estimate of $\Lambda_0$ by

$$\hat{\lambda}_0(t_i) = \frac{\delta_i}{\sum_{j=1}^{n} I(t_j \geq t_i) \sum_{m=1}^{M} v_{im}^{(k)} (\delta_i + 1) \exp(X_i^T \beta_m) / A_{im}^{(k)}}.$$

To update $\beta_m$, we calculate the first and second derivatives of $Q_4(\beta | \Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)})$ as follows:

$$Q_{4\beta_m}(\beta | \Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)})$$

$$= \sum_{i=1}^{n} \sum_{m=1}^{M} v_{im}^{(k)} \delta_i X_i^T \beta_m - \sum_{i=1}^{n} \sum_{m=1}^{M} v_{im}^{(k)} (\delta_i + 1) \Lambda_0^{(k)}(t_i) \exp(2X_i^T \beta_m) / A_{im}^{(k)}.$$

and

$$Q''_{4\beta_m}(\beta | \Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)}) = - \sum_{i=1}^{n} \sum_{m=1}^{M} v_{im}^{(k)} (\delta_i + 1) \Lambda_0^{(k)}(t_i) \exp(2X_i^T \beta_m) / A_{im}^{(k)}.$$

Then, $\beta_m$ can be estimated by

$$\beta_m^{(k+1)} = \beta_m^{(k)} - Q''_{4\beta_m}(\beta_m^{(k)} | \Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)})^{-1} Q_{4\beta_m}(\beta_m^{(k)} | \Lambda_0^{(k)}, \beta^{(k)}, \pi^{(k)}).$$

4. Simulation Study

According to the estimation equation derived in previous sections, we simulate the data to analyze the estimation result at finite sample size. As the number of groups $M$ in the mixture of proportional odds model is unknown and will be estimated by a data-driven manner. Here, we use the modified Bayesian information criterion (BIC [19]) to choose the number of components $M$ by minimizing the criterion function:

$$BIC_M = -2 \cdot \ell(\hat{\Lambda}_0, \hat{\beta}, \hat{\pi} | Y_{obs}) + M \cdot q \cdot log(n).$$
where \( n \) is the sample size and \( q \) is the dimension of \( \beta_m \). Note that this is strictly related to the marginal likelihood computation as can be seen in [27–29].

**Scenario 1.** We generate clustered right-censored data from a mixture of proportional odds model with two subgroups and two covariates

\[
\lambda_i(t \mid X) = \frac{\lambda_0(t) \exp(X_i^\top \beta_i)}{1 + \Lambda_0(t) \exp(X_i^\top \beta_i)},
\]

where the two covariates \( X_{i1} \) and \( X_{i2} \) are independent and follow the standard normal distribution, \( \Lambda_0(t) = \left(t/2\right)^2 \). We randomly assign the sample size \( n \) into two subgroups with equal probabilities, i.e., we let \( P(i \in G_1) = P(i \in G_2) = 0.5 \) so that \( \beta_i = (3, -1) \top \) for \( i \in G_1 \), \( \beta_i = (-3, 2) \top \) for \( i \in G_2 \). We choose different sample sizes \( n = 150, 250, 500 \) and set the censoring proportion at 30% to assess their performance of the proposed estimation procedures.

Table 1 reports the mean and median of the estimator \( \hat{M} \) and the proportion of \( \hat{M} \) equal to the true number of subgroups based on 500 replications. Table 2 reports the empirical bias, mean square error (MSE), and standard error (s.d.) of the estimators \( \hat{\pi}, \hat{\beta}_1, \) and \( \hat{\beta}_2 \) based on 500 replications. We found that the mean of \( \hat{M} \) gradually approaches the true number of subgroups 2, and the median of \( \hat{M} \) remains at 2, and the proportion of correctly identifying the true number of subgroups is close to 1 with the increase of sample size. Moreover, our methods can estimate the parameters well with small empirical bias, small MSE, and small standard error, even at small sample sizes.

### Table 1. The mean, median, standard error (s.d.), and the proportion (Pro) of \( \hat{M} \) in Scenario 1.

<table>
<thead>
<tr>
<th>Method</th>
<th>n</th>
<th>Mean</th>
<th>Median</th>
<th>Pro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profile MM</td>
<td>150</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>2.03</td>
<td>2</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Non-profile MM</td>
<td>150</td>
<td>2</td>
<td>2</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>2.03</td>
<td>2</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>2.005</td>
<td>2</td>
<td>0.995</td>
</tr>
</tbody>
</table>

### Table 2. Parameter estimation results in Scenario 1.

<table>
<thead>
<tr>
<th>n</th>
<th>Parameter</th>
<th>True</th>
<th>Profile MM</th>
<th>Non-Profile MM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bias</td>
<td>MSE</td>
</tr>
<tr>
<td>150</td>
<td>( \pi_1 )</td>
<td>0.5</td>
<td>-0.0014</td>
<td>0.0029</td>
</tr>
<tr>
<td></td>
<td>( \beta_{11} )</td>
<td>3</td>
<td>0.0162</td>
<td>0.1731</td>
</tr>
<tr>
<td></td>
<td>( \beta_{12} )</td>
<td>-1</td>
<td>-0.0216</td>
<td>0.0977</td>
</tr>
<tr>
<td></td>
<td>( \beta_{21} )</td>
<td>-3</td>
<td>-0.0085</td>
<td>0.1785</td>
</tr>
<tr>
<td></td>
<td>( \beta_{22} )</td>
<td>2</td>
<td>-0.0011</td>
<td>0.1234</td>
</tr>
<tr>
<td>250</td>
<td>( \pi_1 )</td>
<td>0.5</td>
<td>0.0013</td>
<td>0.0019</td>
</tr>
<tr>
<td></td>
<td>( \beta_{11} )</td>
<td>3</td>
<td>-0.0106</td>
<td>0.0911</td>
</tr>
<tr>
<td></td>
<td>( \beta_{12} )</td>
<td>-1</td>
<td>-0.0119</td>
<td>0.0551</td>
</tr>
<tr>
<td></td>
<td>( \beta_{21} )</td>
<td>-3</td>
<td>-0.0100</td>
<td>0.1068</td>
</tr>
<tr>
<td></td>
<td>( \beta_{22} )</td>
<td>2</td>
<td>0.0032</td>
<td>0.0744</td>
</tr>
<tr>
<td>500</td>
<td>( \pi_1 )</td>
<td>0.5</td>
<td>0.0002</td>
<td>0.0008</td>
</tr>
<tr>
<td></td>
<td>( \beta_{11} )</td>
<td>3</td>
<td>0.0089</td>
<td>0.0431</td>
</tr>
<tr>
<td></td>
<td>( \beta_{12} )</td>
<td>-1</td>
<td>-0.0082</td>
<td>0.0240</td>
</tr>
<tr>
<td></td>
<td>( \beta_{21} )</td>
<td>-3</td>
<td>-0.0096</td>
<td>0.0485</td>
</tr>
<tr>
<td></td>
<td>( \beta_{22} )</td>
<td>2</td>
<td>0.0076</td>
<td>0.0318</td>
</tr>
</tbody>
</table>
Scenario 2. We generate right-censored data from a proportional odds model with three covariates

\[
\lambda_i(t \mid X) = \frac{\lambda_0(t) \exp(X_i^\top \beta_i)}{1 + \Lambda_0(t) \exp(X_i^\top \beta_i)},
\]

where the three covariates \(X_{i1}, X_{i2}, \text{and } X_{i3}\) are independent and follow the standard normal distribution. We set \(\beta = (1, -3, 2)^\top\) and \(\Lambda_0(t) = (t/2)^2\) for all subjects. Note that the model corresponds to the latent proportional odds model with the true number of subgroups \(M\) being 1. We set the censoring proportion at 30% and choose different sample sizes \(n = 250, 500\) to assess their performance of the proposed estimation procedures.

Table 3 reports the mean and median of the estimator \(\hat{M}\) and the proportion of \(\hat{M}\) equal to the true number of subgroups based on 200 replications. Table 4 reports the empirical bias, mean square error (MSE), and standard error (s.d.) of the estimators \(\beta\) based on 500 replications. Based on the profile MM method, we observed that the median of \(\hat{M}\) is equal to the true number 1, the mean also gets closer to 1, and the empirical percentage of \(\hat{M}\) is close to 1 as the sample size increases. Based on the non-profile MM method, we found that the mean and median of \(\hat{M}\) are both the true number 1, and the proportion of \(\hat{M}\) is 1 when the sample sizes are 250 and 500. Furthermore, our methods show excellent performance in parameter estimation. We obtain great estimates of \(\beta\) under different sample sizes.

Table 5 reports the mean and median of the estimator \(\hat{M}\) and the proportion of \(\hat{M}\) equal to the true number of subgroups based on 500 replications. Table 6 reports the empirical bias, mean square error (MSE), and standard error (s.d.) of the estimators \(\beta\) based on 500 replications. Based on the profile MM method, we observed that the median of \(\hat{M}\) is equal to the true number 1, the mean also gets closer to 1, and the empirical percentage of \(\hat{M}\) is close to 1 as the sample size increases. Based on the non-profile MM method, we found that the mean and median of \(\hat{M}\) are both the true number 1, and the proportion of \(\hat{M}\) is 1 when the sample sizes are 250 and 500. Furthermore, our methods show excellent performance in parameter estimation. We obtain great estimates of \(\beta\) under different sample sizes.

Scenario 3. We generate clustered right-censored data from a mixture of proportional odds model with two subgroups and two correlated covariates

\[
\lambda_i(t \mid X) = \frac{\lambda_0(t) \exp(X_i^\top \beta_i)}{1 + \Lambda_0(t) \exp(X_i^\top \beta_i)},
\]

where the two covariates are generated from a multivariate normal distribution with mean zero and a first-order autoregressive structure \(\rho^{r-s}\) for \(r, s = 1, 2\). Set \(\Lambda_0(t) = (t/2)^2\), sample size \(n = 200\). Then, we randomly assign the sample size \(n\) into two subgroups with equal probabilities, i.e., we let \(P(i \in G_1) = P(i \in G_2) = 0.5\) so that \(\beta_1 = (3, -1)^\top\) for \(i \in G_1, \beta_1 = (-3, 2)^\top\) for \(i \in G_2\). We choose different values of \(\rho\) with \(\rho = 0.2, 0.8\) and set the censoring proportion at 30% to assess their performance of the proposed estimation procedures.

Table 5 reports the mean and median of the estimator \(\hat{M}\) and the proportion of \(\hat{M}\) equal to the true number of subgroups based on 500 replications. Table 6 reports the empirical bias, mean square error (MSE), and standard error (s.d.) of the estimators \(\beta\) based on 500 replications. Based on the profile MM method, we observed that the median of \(\hat{M}\) is equal to the true number 1, the mean also gets closer to 1, and the empirical percentage of \(\hat{M}\) is close to 1 as the sample size increases. Based on the non-profile MM method, we found that the mean and median of \(\hat{M}\) are both the true number 1, and the proportion of \(\hat{M}\) is 1 when the sample sizes are 250 and 500. Furthermore, our methods show excellent performance in parameter estimation. We obtain great estimates of \(\beta\) under different sample sizes.
bias, mean square error (MSE), and standard error (s.d.) of the estimators \( \hat{\pi} \), \( \beta_1 \), and \( \beta_2 \) based on 500 replications. In Table 5, the results of the profile MM method and non-profile MM method are basically consistent, the proportions of \( \hat{M} \) are very close to 1 and the smaller the value of \( \rho \), the larger the value of Pro. It shows that our proposed methods can accurately identify the number of subgroups. In Table 6, the estimation results at a smaller value of \( \rho \) perform better and more stably than the results at a larger value of \( \rho \) for both the profile MM method and the non-profile MM method.

**Table 5.** The mean, median, and the proportion (Pro) of \( \hat{M} \) in Scenario 3.

<table>
<thead>
<tr>
<th>Method</th>
<th>( \rho )</th>
<th>Mean</th>
<th>Median</th>
<th>Pro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profile MM</td>
<td>0.2</td>
<td>2.005</td>
<td>2</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>2.015</td>
<td>2</td>
<td>0.985</td>
</tr>
<tr>
<td>Non-profile MM</td>
<td>0.2</td>
<td>2.005</td>
<td>2</td>
<td>0.995</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>2.015</td>
<td>2</td>
<td>0.985</td>
</tr>
</tbody>
</table>

**Table 6.** Parameter estimation results in Scenario 3.

<table>
<thead>
<tr>
<th>( \rho )</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>0.2</td>
<td>( \pi_1 )</td>
<td>0.0003</td>
<td>0.0023</td>
</tr>
<tr>
<td></td>
<td>( \beta_{11} )</td>
<td>0.0285</td>
<td>0.1228</td>
</tr>
<tr>
<td></td>
<td>( \beta_{12} )</td>
<td>0.0161</td>
<td>0.0625</td>
</tr>
<tr>
<td></td>
<td>( \beta_{21} )</td>
<td>-0.0356</td>
<td>0.1194</td>
</tr>
<tr>
<td></td>
<td>( \beta_{22} )</td>
<td>-0.0149</td>
<td>0.0866</td>
</tr>
<tr>
<td>0.8</td>
<td>( \pi_1 )</td>
<td>0.0023</td>
<td>0.0043</td>
</tr>
<tr>
<td></td>
<td>( \beta_{11} )</td>
<td>0.0251</td>
<td>0.2466</td>
</tr>
<tr>
<td></td>
<td>( \beta_{12} )</td>
<td>-0.0155</td>
<td>0.1753</td>
</tr>
<tr>
<td></td>
<td>( \beta_{21} )</td>
<td>-0.0990</td>
<td>0.3442</td>
</tr>
<tr>
<td></td>
<td>( \beta_{22} )</td>
<td>0.0965</td>
<td>0.2601</td>
</tr>
</tbody>
</table>

5. **Real Data Analysis**

Now, we apply the proposed method to analyze the German Breast Cancer Study data which can be available from R package “pec”. The data contain the observations of 686 women where the censoring rate is 56.41%. In order to analyze whether there is heterogeneity in the data, we consider “tgrade(1 vs. III, II vs. III)” and “pnodes” as explanatory variables of interest, where “tgrade” indicates tumor grade which is an ordered factor at levels I vs. III or II vs. III, “pnodes” indicates the number of positive lymph nodes. Then, we use the BIC criterion function to determine the number of subgroups \( M \). In Table 7, we report the maximum log-likelihood values (LL), the BIC values (BIC), and the estimated parameters under the number of subgroups \( M = 1, 2, 3 \). Based on the results in Table 7, we found that the optimal \( M \) is 1 by comparing the BIC values. The estimated regression coefficients are detailed in Table 7.
Table 7. Estimation results for breast cancer data.

<table>
<thead>
<tr>
<th>Method</th>
<th>M</th>
<th>LL</th>
<th>BIC</th>
<th>Estimated Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>( \hat{\beta} = (-1.3489, -0.3919, -0.0937) )</td>
</tr>
<tr>
<td>Profile MM</td>
<td>1</td>
<td>-2049.165</td>
<td>4117.923</td>
<td>( \hat{\pi}_1 = 0.7012, \hat{\beta}_1 = (-1.2093, -0.0392, 0.0630) )</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-2046.936</td>
<td>4133.057</td>
<td>( \hat{\pi}_1 = 0.2680, \hat{\beta}_1 = (-1.6517, -1.4040, 0.2596) )</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-2044.792</td>
<td>4148.361</td>
<td>( \hat{\pi}_1 = 0.2680, \hat{\beta}_1 = (-1.6516, -1.4038, 0.2596) )</td>
</tr>
<tr>
<td>Non-profile MM</td>
<td>1</td>
<td>-2049.165</td>
<td>4117.923</td>
<td>( \hat{\pi}_1 = 0.7012, \hat{\beta}_1 = (-1.2093, -0.0393, 0.0630) )</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-2046.936</td>
<td>4133.057</td>
<td>( \hat{\pi}_1 = 0.2680, \hat{\beta}_1 = (-1.6516, -1.4040, 0.2596) )</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-2044.792</td>
<td>4148.361</td>
<td>( \hat{\pi}_1 = 0.2680, \hat{\beta}_1 = (-1.6516, -1.4038, 0.2596) )</td>
</tr>
</tbody>
</table>

6. Conclusions

In this work, we introduce the MM algorithm into a semiparametric mixture modeling strategy in the proportional odds model for subgroup analysis of survival data that flexibly allows the covariate effects to differ among several subgroups. Both proposed MM methods to the semiparametric mixture of proportional odds model are able to conduct simultaneous subgroup identification and regression analysis, which provides a general frame for constructing iterative algorithms with monotone convergence. The main advantage of our MM algorithm is that it can separate the nonparametric baseline hazard rate with other regression parameters and can help to avoid matrix inversion in high-dimensional regression analysis, which makes the estimation process more efficient. Furthermore, our algorithm can mesh well with the existing quasi-Newton acceleration and other simple off-the-shelf accelerators to further boost the estimation process. Such estimation procedures derived for the semiparametric mixture proportional odds model can be easily extended to other semiparametric or nonparametric mixture models. Although our proposed MM algorithms are developed for the mixture of proportional odds models, a parallel approach can essentially be developed for the more general mixture of transformation models. We will investigate this in our future work.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References