mathematics

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: xjf@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

Citation: Huang, X.; Xiong, C.; Xu, J.; Shi, J.; Huang, J. Mixture Modeling of Time-to-Event Data in the Proportional Odds Model. *Mathematics* **2022**, *10*, 3375. https://doi.org/10.3390/math10183375

Academic Editors: Min Wang, Haijun Gong, Liucang Wu and Songfeng Zheng

Received: 18 August 2022
Accepted: 11 September 2022
Published: 16 September 2022

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] studied 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 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, $\ell(\boldsymbol{\alpha}|Y_{obs})$ is the objective log-likelihood function, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_q)^T \in \boldsymbol{\Theta}$ are the vector of parameters to be estimated, and $\boldsymbol{\Theta}$ is the parameter space. The maximum likelihood estimate of $\boldsymbol{\alpha}$ is $\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}\in\boldsymbol{\Theta}} \ell(\boldsymbol{\alpha}|Y_{obs})$. 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$Q(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(k)})$ such that

$$Q(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(k)}) \le \ell(\boldsymbol{\alpha}|Y_{obs}), \forall \boldsymbol{\alpha}, \boldsymbol{\alpha}^{(k)} \in \boldsymbol{\Theta}, Q(\boldsymbol{\alpha}^{(k)}|\boldsymbol{\alpha}^{(k)}) = \ell(\boldsymbol{\alpha}^{(k)}|Y_{obs}), \tag{1}$$

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

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

Since

$$\ell(\boldsymbol{\alpha}^{(k+1)}|Y_{obs}) \ge Q(\boldsymbol{\alpha}^{(k+1)}|\boldsymbol{\alpha}^{(k)}) \ge Q(\boldsymbol{\alpha}^{(k)}|\boldsymbol{\alpha}^{(k)}) = \ell(\boldsymbol{\alpha}^{(k)}|Y_{obs}),$$

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

3. Proportional Odds Model with Individual-Specific Covariate Effects

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

$$\lambda_i(t \mid X) = \frac{\lambda_0(t)\exp\left(X_i^\top \boldsymbol{\beta}\right)}{1 + \Lambda_0(t)\exp\left(X_i^\top \boldsymbol{\beta}\right)},$$

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 $\lambda(t|X) = -\frac{d(-\log S(t|X))}{dt}$. In the proportional odds model, $\boldsymbol{\beta}$ 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:

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

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

3.1. Heterogeneity Regression Pursuit via MM Algorithm

The joint density function of (T, δ) can be written as

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

where

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

denotes the density function of the m-th subgroup, $m = 1, 2, ..., M$, $\boldsymbol{\beta}_m$ 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

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

where $\Lambda_0(t) = \sum_i^n I(t_i \leqslant t)\lambda_0(t_i)$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_M^T)^T$, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_M)$. Given the parameters in the k-th iteration and denoting

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

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

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

By the continuous version of Jensen's inequality as $\varphi\left(\int_\Omega f(x) \cdot g(x)\mathrm{d}x\right) \geq \int_\Omega \varphi(f(x)) \cdot g(x)\mathrm{d}x$, 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 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,

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

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 $\ell(\Lambda_0, \boldsymbol{\beta}, \boldsymbol{\pi}|Y_{obs})$ as

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

where

$$Q(\boldsymbol{\pi}|\Lambda_0^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\pi}^{(k)}) = \sum_{i=1}^{n}\sum_{m=1}^{M} v_{im}^{(k)} \cdot \log \pi_m, \tag{3}$$

and

$$\begin{aligned} &Q(\Lambda_0, \boldsymbol{\beta}|\Lambda_0^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\pi}^{(k)}) \\ &= \sum_{i=1}^{n}\sum_{m=1}^{M} v_{im}^{(k)} \log f_m(t_i, \delta_i|X_i), \\ &= \sum_{i=1}^{n} \delta_i \log \lambda_0(t_i) + \sum_{i=1}^{n}\sum_{m=1}^{M} v_{im}^{(k)} \delta_i X_i^\top \boldsymbol{\beta}_m - \sum_{i=1}^{n}\sum_{m=1}^{M} v_{im}^{(k)} (\delta_i + 1) \log\Big[1 + \Lambda_0(t_i) \exp(X_i^\top \boldsymbol{\beta}_m)\Big]. \end{aligned} \tag{4}$$

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

$$\hat{\pi}_m = \frac{\sum_{i=1}^{n} v_{im}^{(k)}}{n}, \quad m = 1, \ldots, M. \tag{5}$$

To update $(\Lambda_0, \boldsymbol{\beta})$, 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

$$\begin{aligned} -\log\Big[1 + \Lambda_0(t_i) \exp\Big(X_i^\top \boldsymbol{\beta}_m\Big)\Big] &\geq -\log(A_{im}^{(k)}) \\ &\quad - \frac{1 + \Lambda_0(t_i) \exp(X_i^\top \boldsymbol{\beta}_m) - A_{im}^{(k)}}{A_{im}^{(k)}}, \end{aligned}$$

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

$$\begin{aligned} &Q_1(\Lambda_0, \boldsymbol{\beta}|\Lambda_0^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\pi}^{(k)}) \\ &= \sum_{i=1}^{n} \delta_i \log \lambda_0(t_i) + \sum_{i=1}^{n}\sum_{m=1}^{M} v_{im}^{(k)} \delta_i X_i^\top \boldsymbol{\beta}_m - \sum_{i=1}^{n}\sum_{m=1}^{M} v_{im}^{(k)} (\delta_i + 1) \frac{\Lambda_0(t_i) \exp(X_i^\top \boldsymbol{\beta}_m)}{A_{im}^{(k)}}. \end{aligned}$$

3.2. Profile MM Method

Following [25,26], we consider the profile estimation approach and first profile out Λ_0 in $Q_1(\Lambda_0, \boldsymbol{\beta}|\Lambda_0^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\pi}^{(k)})$ for any given $\boldsymbol{\beta}$. This leads to the estimate of Λ_0 given $\boldsymbol{\beta}$ as

$$\hat{\lambda}_0(t_i) = \frac{\delta_i}{\sum_{j=1}^{n} I(t_j \geqslant t_i) \sum_{m=1}^{M} v_{jm}^{(k)} (\delta_j + 1) \exp\left(X_j^\top \boldsymbol{\beta}_m\right) / A_{jm}^{(k)}}. \tag{6}$$

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

$$\begin{aligned} & Q_2(\boldsymbol{\beta}|\Lambda_0^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\pi}^{(k)}) \\ & = \sum_{i=1}^{n} \sum_{m=1}^{M} v_{im}^{(k)} \delta_i X_i^\top \boldsymbol{\beta}_m - \sum_{i=1}^{n} \delta_i \log\left[\sum_{j=1}^{n} I(t_j \geqslant t_i) \sum_{m=1}^{M} v_{jm}^{(k)} (\delta_j + 1) \exp\left(X_j^\top \boldsymbol{\beta}_m\right) / A_{jm}^{(k)}\right]. \end{aligned}$$

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

$$\begin{aligned} & Q_3(\boldsymbol{\beta}|\Lambda_0^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\pi}^{(k)}) \\ & = \sum_{i=1}^{n} \sum_{m=1}^{M} v_{im}^{(k)} \delta_i X_i^\top \boldsymbol{\beta}_m - \sum_{i=1}^{n} \delta_i \frac{\sum_{j=1}^{n} I(t_j \geqslant t_i) \sum_{m=1}^{M} v_{jm}^{(k)} (\delta_j + 1) \exp\left(X_j^\top \boldsymbol{\beta}_m\right) / A_{jm}^{(k)}}{B_i^{(k)}} \\ & = \sum_{m=1}^{M} \left\{ \sum_{i=1}^{n} v_{im}^{(k)} \delta_i X_i^\top \boldsymbol{\beta}_m - \sum_{i=1}^{n} \delta_i \frac{\sum_{j=1}^{n} I(t_j \geqslant t_i) v_{jm}^{(k)} (\delta_j + 1) \exp\left(X_j^\top \boldsymbol{\beta}_m\right) / A_{jm}^{(k)}}{B_i^{(k)}} \right\} \\ & = \sum_{m=1}^{M} Q_3(\boldsymbol{\beta}_m|\Lambda_0^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\pi}^{(k)}), \end{aligned}$$

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

3.3. Non-Profile MM Method

For the above profile MM method, the estimate of Λ_0 is highly related to the estimate of $\boldsymbol{\beta}$ because we treat nonparametric component Λ_0 as a function of $\boldsymbol{\beta}$ in the profile step. Inspired by the parameter-separable property of the MM principle, we further separate the nonparametric part Λ_0 with the $\boldsymbol{\beta}$ according to the decomposition rules. That is, we use the following inequality of arithmetic and geometric means to the function $Q_1(\Lambda_0, \boldsymbol{\beta}|\Lambda_0^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\pi}^{(k)})$ as

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

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

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

That is,

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

Substituting the above inequality back to $Q_1(\Lambda_0, \boldsymbol{\beta}|\Lambda_0^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\pi}^{(k)})$, we may obtain

$$\begin{aligned}
&Q_4(\Lambda_0, \boldsymbol{\beta}|\Lambda_0^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\pi}^{(k)}) \\
&= \sum_{i=1}^{n} \delta_i \log \lambda_0(t_i) + \sum_{i=1}^{n}\sum_{m=1}^{M} v_{im}^{(k)} \delta_i X_i^\top \boldsymbol{\beta}_m \\
&- \sum_{i=1}^{n}\sum_{m=1}^{M} v_{im}^{(k)}(\delta_i + 1)\left[\frac{\exp(X_i^\top \boldsymbol{\beta}_m^{(k)})}{2\Lambda_0^{(k)}(t_i)}\Lambda_0^2(t_i) + \frac{\Lambda_0^{(k)}(t_i)}{2\exp(X_i^\top \boldsymbol{\beta}_m^{(k)})}\exp(2X_i^\top \boldsymbol{\beta}_m)\right] / A_{im}^{(k)} \\
&\hat{=} Q_4(\Lambda_0|\Lambda_0^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\pi}^{(k)}) + Q_4(\boldsymbol{\beta}|\Lambda_0^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\pi}^{(k)}),
\end{aligned}$$

where

$$Q_4(\Lambda_0|\Lambda_0^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\pi}^{(k)}) = \sum_{i=1}^{n} \delta_i \log \lambda_0(t_i) - \sum_{i=1}^{n}\sum_{m=1}^{M} v_{im}^{(k)}(\delta_i + 1)\frac{\exp(X_i^\top \boldsymbol{\beta}_m^{(k)})}{2\Lambda_0^{(k)}(t_i)}\Lambda_0^2(t_i) / A_{im}^{(k)}$$

and

$$\begin{aligned}
&Q_4(\boldsymbol{\beta}|\Lambda_0^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\pi}^{(k)}) \\
&= \sum_{i=1}^{n}\sum_{m=1}^{M} v_{im}^{(k)} \delta_i X_i^\top \boldsymbol{\beta}_m - \sum_{i=1}^{n}\sum_{m=1}^{M} \frac{v_{im}^{(k)}(\delta_i + 1)\Lambda_0^{(k)}(t_i)}{2\exp(X_i^\top \boldsymbol{\beta}_m^{(k)})}\exp(2X_i^\top \boldsymbol{\beta}_m) / A_{im}^{(k)}.
\end{aligned}$$

It is observed that the parameters Λ_0 and $\boldsymbol{\beta}_m$ are completely separated, then the corresponding parameter estimators can be obtained by differentiating them separately. Letting $\partial Q_4(\Lambda_0|\Lambda_0^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\pi}^{(k)})/\partial\Lambda_0 = 0$, we obtain the estimate of Λ_0 by

$$\hat{\lambda}_0(t_i) = \frac{\delta_i}{\sum_{j=1}^{n} I(t_j \geqslant t_i) \sum_{m=1}^{M} v_{jm}^{(k)}(\delta_j + 1)\exp\left(X_j^\top \boldsymbol{\beta}_m\right) / A_{jm}^{(k)}}.$$

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

$$\begin{aligned}
&Q'_{4\boldsymbol{\beta}_m}(\boldsymbol{\beta}|\Lambda_0^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\pi}^{(k)}) \\
&= \sum_{i=1}^{n}\sum_{m=1}^{M} v_{im}^{(k)} \delta_i X_i^\top - \sum_{i=1}^{n}\sum_{m=1}^{M} \frac{v_{im}^{(k)}(\delta_i + 1)\Lambda_0^{(k)}(t_i)}{\exp(X_i^\top \boldsymbol{\beta}_m^{(k)})}\exp(2X_i^\top \boldsymbol{\beta}_m) X_i^\top / A_{im}^{(k)}
\end{aligned}$$

and

$$Q''_{4\boldsymbol{\beta}_m}(\boldsymbol{\beta}|\Lambda_0^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\pi}^{(k)}) = -\sum_{i=1}^{n}\sum_{m=1}^{M} \frac{v_{im}^{(k)}(\delta_i + 1)\Lambda_0^{(k)}(t_i)}{\exp(X_i^\top \boldsymbol{\beta}_m^{(k)})}\exp(2X_i^\top \boldsymbol{\beta}_m) X_i^\top X_i / A_{im}^{(k)}.$$

Then, $\boldsymbol{\beta}_m$ can be estimated by

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

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:

$$BIC_M = -2 * \ell(\hat{\Lambda}_0, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}|Y_{obs}) + M * q * \log(n).$$

where n is the sample size and q is the dimension of $\boldsymbol{\beta}_m$. 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

$$\lambda_i(t \mid X) = \frac{\lambda_0(t)\exp\left(X_i^\top \boldsymbol{\beta_i}\right)}{1 + \Lambda_0(t)\exp\left(X_i^\top \boldsymbol{\beta_i}\right)},$$

where the two covariates X_{i1} and X_{i2} are independent and follow the standard normal distribution, $\Lambda_0(t) = (t/2)^2$, 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 $\boldsymbol{\beta}_i = (3, -1)^\top$ for $i \in G_1$, $\boldsymbol{\beta}_i = (-3, 2)^\top$ 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 $\hat{\pi}$, $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$ 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.

Method	n	Mean	Median	Pro
Profile MM	150	2	2	1
	250	2.03	2	0.97
	500	2	2	1
Non-profile MM	150	2.03	2	0.97
	250	2.03	2	0.97
	500	2.005	2	0.995

Table 2. Parameter estimation results in Scenario 1.

n	Parameter	True	Profile MM			Non-Profile MM		
			BIAS	MSE	s.d.	BIAS	MSE	s.d.
150	π_1	0.5	−0.0014	0.0029	0.0543	−0.0035	0.0032	0.0565
	β_{11}	3	0.0162	0.1731	0.4162	0.0235	0.1812	0.4255
	β_{12}	−1	−0.0216	0.0977	0.3122	0.0047	0.0925	0.3045
	β_{21}	−3	−0.0085	0.1785	0.4228	−0.0233	0.1913	0.4372
	β_{22}	2	−0.0011	0.1234	0.3516	0.0332	0.1267	0.23548
250	π_1	0.5	0.0013	0.0019	0.0441	0.0018	0.0017	0.0417
	β_{11}	3	−0.0106	0.0911	0.3019	−0.0112	0.1038	0.3222
	β_{12}	−1	−0.0119	0.0551	0.2347	−0.0134	0.0487	0.2206
	β_{21}	−3	−0.0100	0.1068	0.3270	−0.0067	0.1030	0.3212
	β_{22}	2	0.0032	0.0744	0.2730	−0.0059	0.0724	0.2693
500	π_1	0.5	0.0002	0.0008	0.0287	−0.0014	0.0008	0.0277
	β_{11}	3	0.0089	0.0431	0.2076	0.0085	0.0462	0.2150
	β_{12}	−1	−0.0082	0.0240	0.1550	0.0043	0.0220	0.1483
	β_{21}	−3	−0.0096	0.0485	0.2202	−0.0158	0.0434	0.2079
	β_{22}	2	0.0076	0.0318	0.1784	−0.0053	0.0349	0.1870

Scenario 2. We generate right-censored data from a proportional odds model with three covariates

$$\lambda_i(t \mid X) = \frac{\lambda_0(t)\exp\left(X_i^\top \boldsymbol{\beta_i}\right)}{1 + \Lambda_0(t)\exp\left(X_i^\top \boldsymbol{\beta_i}\right)},$$

where the three covariates X_{i1}, X_{i2} and X_{i3} are independent and follow the standard normal distribution. We set $\boldsymbol{\beta} = (1, -3, 2)^\top$ 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 $\boldsymbol{\beta}$ 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 $\boldsymbol{\beta}$ under different sample sizes.

Table 3. The mean, median, and the proportion (Pro) of $\hat{M}$ in Scenario 2.

Method	n	Mean	Median	Pro
Profile MM	250	1.005	1	0.995
	500	1	1	1
Non-profile MM	250	1	1	1
	500	1	1	1

Table 4. Parameter estimation results in Scenario 2.

n	Parameter	True	Profile MM			Non-Profile MM		
			BIAS	MSE	s.d.	BIAS	MSE	s.d.
250	β_1	1	0.0021	0.0185	0.1361	−0.0049	0.0206	0.1436
	β_2	−3	0.0194	0.0517	0.2268	0.0078	0.0449	0.2121
	β_3	2	−0.0190	0.0319	0.1779	0.0060	0.0309	0.1760
500	β_1	1	−0.0012	0.0093	0.0966	0.0004	0.0097	0.0986
	β_2	−3	−0.0014	0.02439	0.1563	0.0001	0.0247	0.1574
	β_3	2	0.0114	0.0167	0.1288	−0.0013	0.0149	0.1221

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

$$\lambda_i(t \mid X) = \frac{\lambda_0(t)\exp\left(X_i^\top \boldsymbol{\beta_i}\right)}{1 + \Lambda_0(t)\exp\left(X_i^\top \boldsymbol{\beta_i}\right)},$$

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 $\boldsymbol{\beta}_i = (3, -1)^\top$ for $i \in G_1$, $\boldsymbol{\beta}_i = (-3, 2)^\top$ for $i \in G_2$. We choose different values of ρ 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 $\hat{\pi}$, $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$ 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 ρ, 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 ρ perform better and more stably than the results at a larger value of ρ 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.

Method	ρ	Mean	Median	Pro
Profile MM	0.2	2.005	2	0.995
	0.8	2.015	2	0.985
Non-profile MM	0.2	2.005	2	0.995
	0.8	2.015	2	0.985

Table 6. Parameter estimation results in Scenario 3.

ρ	Parameter	Profile MM			Non-Profile MM		
		BIAS	MSE	s.d.	BIAS	MSE	s.d.
0.2	π_1	0.0003	0.0023	0.0487	−0.0036	0.0023	0.0484
	β_{11}	0.0285	0.1228	0.3502	−0.0146	0.1356	0.3689
	β_{12}	0.0161	0.0625	0.2502	0.0095	0.0822	0.2873
	β_{21}	−0.0356	0.1194	0.3446	−0.0021	0.1351	0.3684
	β_{22}	−0.0149	0.0866	0.2945	−0.0059	0.0918	0.3037
0.8	π_1	0.0023	0.0043	0.0661	−0.0022	0.0039	0.0630
	β_{11}	0.0251	0.2466	0.4972	−0.0131	0.2413	0.4923
	β_{12}	−0.0155	0.1753	0.4195	−0.0005	0.1648	0.4070
	β_{21}	−0.0990	0.3442	0.5797	−0.0209	0.2590	0.5098
	β_{22}	0.0965	0.2601	0.5020	0.0108	0.2128	0.4624

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(I 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.

Method	M	LL	BIC	Estimated Parameters
Profile MM	1	−2049.165	4117.923	$\hat{\boldsymbol{\beta}} = (-1.3489, -0.3919, -0.0937)$
	2	−2046.936	4133.057	$\hat{\pi}_1 = 0.7012, \hat{\boldsymbol{\beta}}_1 = (-1.2093, -0.0392, 0.0630)$ $\hat{\pi}_2 = 0.2988, \hat{\boldsymbol{\beta}}_2 = (-1.6508, -1.2284, 0.2415)$
	3	−2044.792	4148.361	$\hat{\pi}_1 = 0.2680, \hat{\boldsymbol{\beta}}_1 = (-1.6517, -1.4040, 0.2596)$ $\hat{\pi}_2 = 0.0757, \hat{\boldsymbol{\beta}}_2 = (-2.1672, 0.5394, 0.5748)$ $\hat{\pi}_3 = 0.6563, \hat{\boldsymbol{\beta}}_3 = (-1.1245, -0.0477, 0.0658)$
Non-profile MM	1	−2049.165	4117.923	$\hat{\boldsymbol{\beta}} = (-1.3489, -0.3918, -0.0937)$
	2	−2046.936	4133.057	$\hat{\pi}_1 = 0.7012, \hat{\boldsymbol{\beta}}_1 = (-1.2093, -0.0393, 0.0630)$ $\hat{\pi}_2 = 0.2988, \hat{\boldsymbol{\beta}}_2 = (-1.6508, -1.2282, 0.2415)$
	3	−2044.792	4148.361	$\hat{\pi}_1 = 0.2680, \hat{\boldsymbol{\beta}}_1 = (-1.6516, -1.4038, 0.2596)$ $\hat{\pi}_2 = 0.0757, \hat{\boldsymbol{\beta}}_2 = (-2.1672, 0.5392, 0.5748)$ $\hat{\pi}_3 = 0.6563, \hat{\boldsymbol{\beta}}_3 = (-1.1245, -0.0477, 0.0658)$

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

1. Frühwirth-Schnatter, S. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *J. Am. Stat. Assoc.* **2001**, *96*, 194–209.
2. Rossi, P.E.; Allenby, G.M. Bayesian statistics and marketing. *Mark. Sci.* **2003**, *22*, 304–328.
3. Green, P.J.; Richardson, S. Hidden Markov models and disease mapping. *J. Am. Stat. Assoc.* **2002**, *97*, 1055–1070.
4. Wang, P.; Puterman, M.L.; Cockburn, I.; Le, N. Mixed Poisson regression models with covariate dependent rates. *Biometrics* **1996**, *52*, 381–400.
5. Everitt, B. *Finite Mixture Distributions*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
6. Banfield, J.D.; Raftery, A.E. Model-based Gaussian and non-Gaussian clustering. *Biometrics* **1993**, *49*, 803–821.

7. Hastie, T.; Tibshirani, R. Discriminant analysis by Gaussian mixtures. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 155–176.
8. McNicholas, P.D. Model-based classification using latent Gaussian mixture models. *J. Stat. Plan. Inference* **2010**, *140*, 1175–1181.
9. Shen, J.; He, X. Inference for subgroup analysis with a structured logistic-normal mixture model. *J. Am. Stat. Assoc.* **2015**, *110*, 303–312.
10. Chaganty, A.T.; Liang, P. Spectral experts for estimating mixtures of linear regressions. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; pp. 1040–1048.
11. Hurn, M.; Justel, A.; Robert, C.P. Estimating mixtures of regressions. *J. Comput. Graph. Stat.* **2003**, *12*, 55–79.
12. Frühwirth-Schnatter, S.; Frèuhwirth-Schnatter, S. *Finite Mixture and Markov Switching Models*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 425.
13. Peng, Y.; Dear, K.B. A nonparametric mixture model for cure rate estimation. *Biometrics* **2000**, *56*, 237–243.
14. Altstein, L.; Li, G. Latent subgroup analysis of a randomized clinical trial through a semiparametric accelerated failure time mixture model. *Biometrics* **2013**, *69*, 52–61.
15. Wu, R.f.; Zheng, M.; Yu, W. Subgroup analysis with time-to-event data under a logistic-Cox mixture model. *Scand. J. Stat.* **2016**, *43*, 863–878.
16. Becker, M.P.; Yang, I.; Lange, K. EM algorithms without missing data. *Stat. Methods Med. Res.* **1997**, *6*, 38–54.
17. Lange, K.; Hunter, D.R.; Yang, I. Optimization transfer using surrogate objective functions. *J. Comput. Graph. Stat.* **2000**, *9*, 1–20.
18. Hunter, D.R.; et al. MM algorithms for generalized Bradley-Terry models. *Ann. Stat.* **2004**, *32*, 384–406.
19. Hunter, D.R.; Lange, K. Quantile regression via an MM algorithm. *J. Comput. Graph. Stat.* **2000**, *9*, 60–77.
20. Hunter, D.R.; Li, R. Variable selection using MM algorithms. *Ann. Stat.* **2005**, *33*, 1617–1642.
21. Yen, T.J. A majorization–minimization approach to variable selection using spike and slab priors. *Ann. Stat.* **2011**, *39*, 1748–1775.
22. Hunter, D.R.; Lange, K. Computing estimates in the proportional odds model. *Ann. Inst. Stat. Math.* **2002**, *54*, 155–168.
23. Huang, X.; Xu, J.; Tian, G. On profile MM algorithms for gamma frailty survival models. *Stat. Sin.* **2019**, *29*, 895–916.
24. Chi, E.C.; Zhou, H.; Lange, K. Distance majorization and its applications. *Math. Program.* **2014**, *146*, 409–436.
25. Johansen, S. An extension of Cox's regression model. *Int. Stat. Rev. Int. Stat.* **1983**, *51*,165–174.
26. Klein, J.P. Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* **1992**, *48*, 795–806.
27. Knuth, K.H.; Habeck, M.; Malakar, N.K.; Mubeen, A.M.; Placek, B. Bayesian evidence and model selection. *Digit. Signal Process.* **2015**, *47*, 50–67.
28. Llorente, F.; Martino, L.; Curbelo, E.; López-Santiago, J.; Delgado, D. On the safe use of prior densities for Bayesian model selection. *Wiley Interdiscip. Rev. Comput. Stat.* **2022**, e1595. https://doi.org/10.1002/wics.1595.
29. DiCiccio, T.J.; Kass, R.E.; Raftery, A.; Wasserman, L. Computing Bayes factors by combining simulation and asymptotic approximations. *J. Am. Stat. Assoc.* **1997**, *92*, 903–915.