mathematics

MDPI

Article

Subgroup Identification and Regression Analysis of Clustered and Heterogeneous Interval-Censored Data

Xifen Huang and Jinfeng Xu *

School of Mathematics, Yunnan Normal University, Kunming 650500, China; 190004@ynnu.edu.cn
* Correspondence: 204203@ynnu.edu.cn

Abstract: Clustered and heterogeneous interval-censored data occur in many fields such as medical studies. For example, in a migraine study with the Netherlands Twin Registry, the information including time to diagnosis of migraine and gender was collected for 3975 monozygotic and dizygotic twins. Since each study subject is observed only at discrete and periodic follow-up time points, the failure times of interest (i.e., the time when the individual first had a migraine) are known only to belong to certain intervals and hence are interval-censored. Furthermore, these twins come from different genetic backgrounds and may be associated with differential risks for developing migraines. For simultaneous subgroup identification and regression analysis of such data, we propose a latent Cox model where the number of subgroups is not assumed a priori but rather data-driven estimated. The nonparametric maximum likelihood method and an EM algorithm with monotone ascent property are also developed for estimating the model parameters. Simulation studies are conducted to assess the finite sample performance of the proposed estimation procedure. We further illustrate the proposed methodologies by an empirical analysis of migraine data.

Keywords: clustered interval-censored data; EM algorithm; heterogeneous covariate effects; latent Cox model; migraine data; nonparametric maximum likelihood

Citation: Huang, X.; Xu, J. Subgroup Identification and Regression Analysis of Clustered and Heterogeneous Interval-Censored Data. *Mathematics* **2022**, *10*, 862. https://doi.org/10.3390/math10060862

Academic Editor: Jose Antonio Roldan-Nofuentes

Received: 19 January 2022
Accepted: 4 March 2022
Published: 8 March 2022

1. Introduction

Subgroup identification for heterogeneous data has become a ubiquitous problem in a broad range of applications including social science, marketing, and clinical trials. For instance, in clinical trials, heterogeneity may arise due to underlying differences among groups of patients. For patients with similar attributes, disease progression or treatment effects often exhibit close patterns. Therefore, it is valuable to classify the patients into a few homogeneous groups and tailor a disease treatment specifically for each subgroup to optimize the treatment effect. Conceptually, analyzing data from a heterogeneous population consisting of a few homogeneous subgroups is to view data as generated from a mixture of subgroups and leads to a finite mixture model. In unsupervised learning, parametric mixture models have been widely used in many fields. The books [1–4] and the review paper [5] provide a thorough introduction and applications on finite mixture models. In addition, a mixture model can be applied in reliability analysis ([6–8]).

Interval-censored data is a common type of data in real applications. In many clinical applications, the observations are recorded periodically, and the failure times of interest are known between each period, which causes the difficulties on analyzing on this type of data. Ref. [9–11] reviewed existing methods that applied the parametric models and nonparametric estimations for survival curves based on interval-censored data. Particularly, Ref. [12] proposed the nonparametric way for survival distribution estimation and [13] provided the score statistics for parameter estimation for interval-censored data. However, as for heterogeneous interval-censored data, mixture models should be considered for subgroup classification. Only limited research studies are targeting this area. Ref. [14] proposed the estimation methods for Gaussian mixtures using MCMC methodology and [15]

proposed a semi-parametric mixture model in the field of antimicrobial resistance with interval-censored observations. However, the methods mentioned above are density estimation using a mixture model without conducting regression analysis on other observed covariates. There exist computational difficulties on conducting group identification and regression analysis based on mixture models in survival analysis for interval-censored data.

In this paper, motivated by the Netherlands twin study on migraines, we propose a new latent Cox model for analyzing clustered and heterogeneous interval-censored data. The population is separated into a few subgroups according to the covariate effects. The baseline hazard functions for the subgroups as well as the number of subgroups are left unspecified to avoid restrictive distributional assumptions and allow for flexibility.

Compared with existing mixture survival models ([16–18]) in the literature for right-censored data, the proposed model aims to accomplish simultaneous subgroup identification and regression analysis. It is important to note that compared with right-censored data, for interval-censored survival data, the incomplete data information and computational complexity bring greater challenges for the aforementioned tasks. Moreover, we investigate the heterogeneity driven by unknown covariate effects without specifying the baseline hazard functions and the number of subgroups, which make the model estimation more challenging and the computation more intensive. Our new proposed nonparametric maximum likelihood estimation approach separates the parameters during estimation, which greatly reduces the computational complexity. In addition, the proposed EM algorithm has the monotone ascent property for estimating the model parameters. Numerical studies demonstrate its good performance. A modified Bayesian information criterion is also proposed to select the number of mixing components [19].

The rest of the paper is organized as follows. In Section 2, we present the latent Cox model for clustered interval-censored data. In Section 3, we develop an estimation procedure for the proposed model using the EM algorithm. Selecting the number of subgroups and assessing the finite-sample performance of the proposed methods are presented in Section 4. We further provide an application to migraine data to illustrate the practical utilities of the proposed methods in Section 5.

2. Data and Model

Let T_{ij} denote the response of interest (i.e., the failure time) for the jth subject in the ith cluster, where $j = 1, \ldots, n_i$, $i = 1, \ldots, n$, n_i is the number of subjects in the ith cluster and n is the number of clusters in the dataset. Furthermore, T_{ij} is interval-censored and only known to belong to the interval $(L_{ij}, R_{ij}]$. The q-dimensional vector of covariates is denoted by $\mathbf{X}_{ij} = (X_{ij1}, \ldots, X_{ijq})^\top$. The observations are summarized as $\mathrm{Y}_{obs} = \{(L_{ij}, R_{ij}], \mathbf{X}_{ij}; i = 1, \ldots, n, j = 1, \ldots, n_i\}$. For accommodating heterogeneous covariate effects that may exist among subgroups, we propose a latent Cox model for simultaneous subgroup identification and regression analysis. Specifically, the instantaneous hazard function for the jth subject in the ith cluster

$$\lambda_{ij}(t) = \lambda_{0i}(t) \exp\left(\mathbf{X}_{ij}^\top \boldsymbol{\beta}_i\right), i = 1, \ldots, n. \tag{1}$$

As in [10,20], we make the following two assumptions. (A1) L_{ij} and R_{ij} are random and (A2) T_{ij} are independent of $(L_{ij}, R_{ij}]$. It is important to note that the baseline hazard functions and the covariate effects are allowed to vary across the clusters and accordingly accommodate the heterogeneity. In the same spirit of mixture modeling and for extrapolation and interpretation purposes, further assume that n clusters are from M subgroups with $M \geqslant 1$ and the clusters in the same subgroup have the same baseline hazard and covariate effects. In other words, let $\mathcal{G} = (\mathcal{G}_1, \ldots, \mathcal{G}_M)$ be a partition of $\{1, \ldots, n\}$. Let the mixing probabilities be $\pi_m, m = 1, \ldots, M$ and $\pi_1 + \ldots + \pi_M = 1$. For each cluster $i = 1, \ldots, n$, with probability π_m, we have $i \in \mathcal{G}_m$ and $\lambda_{0i}(\cdot) = \lambda_{0m}(\cdot)$ and $\boldsymbol{\beta}_i = \boldsymbol{\beta}_m$. In practice, the number of subgroups M is unknown and will be estimated in a data-driven way. However, in practice, it is usually reasonable to assume that M is much smaller than n. Our goal is to estimate M and the model parameters $\boldsymbol{\Lambda}_0 = (\Lambda_{01}, \ldots, \Lambda_{0M})$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M)$ and

$\mathbf{\Pi} = (\pi_1, \ldots, \pi_M)$. The observed likelihood function of the i-th cluster $\{L_{ij}, R_{ij}, \mathbf{X}_{ij}\}_{j=1}^{n_i}$ can be written as

$$f_i(\mathbf{\Lambda}_0, \boldsymbol{\beta}, \mathbf{\Pi}|Y_{obs}) = \sum_{m=1}^{M} \pi_m \cdot f_{i(m)}(\Lambda_{0m}, \boldsymbol{\beta}_m), \tag{2}$$

where $f_{i(m)}(\Lambda_{0m}, \boldsymbol{\beta}_m)$ denotes the likelihood function of the i-th cluster when it belongs to the m-th subgroup. When the i-th cluster comes from the m-th subgroup, its hazard function $\lambda_{ij(m)}(t) = \lambda_{0m}(t)\exp(\mathbf{X}_{ij}^\top \boldsymbol{\beta}_m)$ for $j = 1, \ldots, n_i$ where $\lambda_{0m}(\cdot)$ is the unspecified baseline hazard function and $\Lambda_{0m}(\cdot)$ is the corresponding cumulative baseline risk of the m-th subgroup. $\boldsymbol{\beta}_m$ is the corresponding effect of $\mathbf{X}_{ij}$ in the m-th subgroup. Furthermore, we suppose that T_{ij} is monitored at a sequence of positive time points $U_{ij1} < \cdots < U_{ijK_{ij}}$ and $\{U_{ijk} : k = 1, \ldots, K_{ij}, i = 1, \ldots, n, j = 1, \ldots, n_i\}$ are independent of $\{T_{ij} : i = 1, \ldots, n, j = 1, \ldots, n_i\}$ as a conventional assumption for interval-censored data. Let $(L_{ij}, R_{ij}]$ be the shortest time interval that brackets T_{ij}, i.e., $L_{ij} = \max\{U_{ijk} : U_{ijk} < T_{ij}, k = 0, \ldots, K_{ij}\}$ and $R_{ij} = \min\{U_{ijk} : U_{ijk} \geqslant T_{ij}, k = 1, \ldots, K_{ij} + 1\}$, where $U_{ij0} = 0$ and $U_{ij,K_{ij}+1} = \infty$. Then, we have

$$f_{i(m)}(\Lambda_{0m}, \boldsymbol{\beta}_m) = \prod_{j=1}^{n_i} \left\{ \exp\left[-\int_0^{L_{ij}} \lambda_{0m}(t_k)\exp(\mathbf{X}_{ij}^\top \boldsymbol{\beta}_m)\right] - \exp\left[-\int_0^{R_{ij}} \lambda_{0m}(t_k)\exp(\mathbf{X}_{ij}^\top \boldsymbol{\beta}_m)\right]\right\},$$

and the log-likelihood $\ell(\mathbf{\Lambda}_0, \boldsymbol{\beta}, \mathbf{\Pi}|Y_{obs})$ based on the observed data $\{(L_{ij}, R_{ij}, \mathbf{X}_{ij}), i = 1, \ldots, n, j = 1, \ldots, n_i\}$ is

$$\ell(\mathbf{\Lambda}_0, \boldsymbol{\beta}, \mathbf{\Pi}|Y_{obs}) = \sum_{i=1}^{n} \log\left[\sum_{m=1}^{M} \pi_m \cdot f_{i(m)}(\Lambda_{0m}, \boldsymbol{\beta}_m)\right]. \tag{3}$$

To estimate $\mathbf{\Lambda}_0$, $\boldsymbol{\beta}$ and $\mathbf{\Pi}$, we adopt the nonparametric maximum likelihood estimation approach. Let $0 = t_0 < t_1 < \cdots < t_K < \infty$ be the ordered sequence of all L_{ij} and R_{ij} with $R_{ij} < \infty$. The estimator for Λ_{0m} is a step function that jumps only at those time points with respective jump sizes of $0, \lambda_{0m}(t_1), \ldots, \lambda_{0m}(t_K)$. It follows that (3) can be rewritten as

$$\sum_{i=1}^{n} \log\left(\sum_{m=1}^{M} \pi_m \prod_{j=1}^{n_i} \left\{ \exp\left[-\sum_{t_k \leqslant L_{ij}} \lambda_{0m}(t_k)\exp(\mathbf{X}_{ij}^\top \boldsymbol{\beta}_m)\right] - \exp\left[-\sum_{t_k \leqslant R_{ij}} \lambda_{0m}(t_k)\exp(\mathbf{X}_{ij}^\top \boldsymbol{\beta}_m)\right]\right\}\right). \tag{4}$$

3. Estimation and Algorithm

The observed data with unknown subgroup memberships can be formulated as an incomplete-data problem in the EM framework. We view the observed data $\{(L_{ij}, R_{ij}], \mathbf{X}_{ij}; i = 1, \ldots, n, j = 1, \ldots, n_i\}$ as being incomplete and introduce the unobserved Bernoulli random variables $Z_{im} \sim \text{Bernoulli}(\pi_m)$ for $m = 1, \ldots, M$,

$$Z_{im} = \begin{cases} 1, & \text{if the } i\text{-th cluster } \{(L_{ij}, R_{ij}], \mathbf{X}_{ij}\}_{j=1}^{n_i} \text{ belongs to the } m\text{-th subgroup,} \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

and Poisson random variables W_{mijk} $(k = 1, \ldots, K)$ with means $\lambda_{0m}(t_k)\exp(\mathbf{X}_{ij}^\top \boldsymbol{\beta}_m)$. Define $A_{mij} = \sum_{t_k \leqslant L_{ij}} W_{mijk}$ and $B_{mij} = I(R_{ij} < \infty)\sum_{L_{ij} \leqslant t_k \leqslant R_{ij}} W_{mijk}$. Since the probability of observing $A_{mij} = 0$ and $B_{mij} > 0$ is $\exp[-\sum_{t_k \leqslant L_{ij}} \lambda_{0m}(t_k)\exp(\mathbf{X}_{ij}^\top \boldsymbol{\beta}_m)] - I(R_{ij} < \infty)\exp[-\sum_{t_k \leqslant R_{ij}} \lambda_{0m}(t_k)\exp(\mathbf{X}_{ij}^\top \boldsymbol{\beta}_m)]$, the likelihood from the observations $\{(L_{ij}, R_{ij}], \mathbf{X}_{ij}, A_{mij} = 0, B_{mij} > 0 : i = 1 \ldots, n; j = 1, \ldots, n_i, m = 1, \ldots, M\}$ is the same as (4). Therefore, we develop an EM algorithm to maximize (4) by treating $W_{mijk}(t_k \leqslant R_{ij}^*)$, Z_{im} as missing

data, where $R_{ij}^* = L_{ij} I(R_{ij} = \infty) + R_{ij} I(R_{ij} < \infty)$. Then, the complete-data log-likelihood is proportional to

$$\ell_{com}(\boldsymbol{\Lambda}_0, \boldsymbol{\beta}, \boldsymbol{\Pi}) \propto \sum_{i=1}^{n} \sum_{m=1}^{M} Z_{im} \Big\{ \log(\pi_m) + I(t_k \leqslant R_{ij}^*) \sum_{k=1}^{K} \sum_{j=1}^{n_i} \Big[W_{mijk} \log(\lambda_{0m}(t_k)) + W_{mijk} \mathbf{X}_{ij}^\top \boldsymbol{\beta}_m - \lambda_{0m}(t_k) \exp(\mathbf{X}_{ij}^\top \boldsymbol{\beta}_m) \Big] \Big\}. \tag{6}$$

In the M-step, we maximize (6) for any given $\boldsymbol{\beta}_m$, then we have

$$\hat{\pi}_m = \sum_{i=1}^{n} Z_{im} / n, \tag{7}$$

$$\hat{\lambda}_{0m}(t_k) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n_i} I(t_k \leqslant R_{ij}^*) Z_{im} W_{mijk}}{\sum_{i=1}^{n} \sum_{j=1}^{n_i} I(t_k \leqslant R_{ij}^*) Z_{im} \exp(\mathbf{X}_{ij}^\top \boldsymbol{\beta}_m)}, \tag{8}$$

where $m = 1, \ldots, M;\ k = 1, \ldots, K$. After incorporating (8) into (6), we obtain

$$\ell_{com}(\boldsymbol{\beta}) \propto \sum_{m=1}^{M} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \sum_{k=1}^{K} I(t_k \leqslant R_{ij}^*) Z_{im} W_{mijk} \left[\mathbf{X}_{ij}^\top \boldsymbol{\beta}_m - \log \left\{ \sum_{i'=1}^{n} \sum_{j'=1}^{n_{i'}} I(t_k \leqslant R_{i'j'}^*) Z_{i'm} \exp(\mathbf{X}_{i'j'}^\top \boldsymbol{\beta}_m) \right\} \right].$$

To update $\boldsymbol{\beta}_m$, we employ the following Newton–Raphson algorithm

$$\boldsymbol{\beta}_m^{(t+1)} = \boldsymbol{\beta}_m^{(t)} + I^{-1}(\boldsymbol{\beta}_m^{(t)}) \nabla \ell_{com}(\boldsymbol{\beta}_m^{(t)}) \tag{9}$$

where

$$\nabla \ell_{com}(\boldsymbol{\beta}^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \sum_{k=1}^{K} I(t_k \leqslant R_{ij}^*) Z_{im} W_{mijk} \times \left[\mathbf{X}_{ij}^\top - \frac{\sum_{i'=1}^{n} \sum_{j'=1}^{n_{i'}} I(t_k \leqslant R_{i'j'}^*) Z_{i'm} \exp(\mathbf{X}_{i'j'}^\top \boldsymbol{\beta}_m^{(t)}) \mathbf{X}_{i'j'}}{\sum_{i'=1}^{n} \sum_{j'=1}^{n_{i'}} I(t_k \leqslant R_{i'j'}^*) Z_{i'm} \exp(\mathbf{X}_{i'j'}^\top \boldsymbol{\beta}_m^{(t)})} \right],$$

$$I^{-1}(\boldsymbol{\beta}^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \sum_{k=1}^{K} I(t_k \leqslant R_{ij}^*) Z_{im} W_{mijk} \left[\frac{\sum_{i'=1}^{n} \sum_{j'=1}^{n_{i'}} I(t_k \leqslant R_{i'j'}^*) Z_{i'm} \exp(\mathbf{X}_{i'j'}^\top \boldsymbol{\beta}_m^{(t)}) \mathbf{X}_{i'j'} \mathbf{X}_{i'j'}^\top}{\sum_{i'=1}^{n} \sum_{j'=1}^{n_{i'}} I(t_k \leqslant R_{i'j'}^*) Z_{i'm} \exp(\mathbf{X}_{i'j'}^\top \boldsymbol{\beta}_m^{(t)})} - \frac{\{ \sum_{i'=1}^{n} \sum_{j'=1}^{n_{i'}} I(t_k \leqslant R_{i'j'}^*) Z_{i'm} \exp(\mathbf{X}_{i'j'}^\top \boldsymbol{\beta}_m^{(t)}) \mathbf{X}_{i'j'} \} \{ \sum_{i'=1}^{n} \sum_{j'=1}^{n_{i'}} I(t_k \leqslant R_{i'j'}^*) Z_{i'm} \exp(\mathbf{X}_{i'j'}^\top \boldsymbol{\beta}_m^{(t)}) \mathbf{X}_{i'j'}^\top \}}{\{ \sum_{i'=1}^{n} \sum_{j'=1}^{n_{i'}} I(t_k \leqslant R_{i'j'}^*) Z_{i'm} \exp(\mathbf{X}_{i'j'}^\top \boldsymbol{\beta}_m^{(t)}) \}^2} \right].$$

In the E-step, we evaluate the conditional expectations of Z_{im} and W_{mijk} involved in the M-step. The posterior mean of Z_{im} is

$$\hat{E}(Z_{im}) = \frac{\pi_m \cdot f_{i(m)}(\Lambda_{0m}, \boldsymbol{\beta}_m)}{\sum_{i=1}^{n} \pi_m \cdot f_{i(m)}(\Lambda_{0m}, \boldsymbol{\beta}_m)}, \tag{10}$$

where

$$f_{i(m)}(\Lambda_{0m}, \boldsymbol{\beta}_m) = \prod_{j=1}^{n_i} \left\{ \exp\left[-\sum_{t_k \leqslant L_{ij}} \lambda_{0m}(t_k) \exp(\mathbf{X}_{ij}^\top \boldsymbol{\beta}_m) \right] - \exp\left[-\sum_{t_k \leqslant R_{ij}} \lambda_{0m}(t_k) \exp(\mathbf{X}_{ij}^\top \boldsymbol{\beta}_m) \right] \right\}.$$

In addition, the conditional expectation of W_{mijk} for $t_k \leqslant R_{ij}^*$ is

$$\hat{E}(W_{mijk}) = I(L_{ij} < t_k \leqslant R_{ij} < \infty) \frac{\lambda_{0m}(t_k) \exp(\mathbf{X}_{ij}^\top \boldsymbol{\beta}_m)}{1 - \exp\left\{ -\sum_{L_{ij} < t_{k'} \leqslant R_{ij}} \lambda_{0m}(t_{k'}) \exp(\mathbf{X}_{ij}^\top \boldsymbol{\beta}_m) \right\}}. \quad (11)$$

Now, we summarize iteration processes between the E-step and M-step for the proposed algorithm as follows.

Step 1. Give initial values of $\boldsymbol{\beta}$, $\boldsymbol{\Pi}$, and $\boldsymbol{\Lambda}_0$.

Step 2. Calculate the conditional expectations of Z_{im} and W_{mijk} via (10) and (11).

Step 3. Replace Z_{im} in (7) by $\hat{E}(Z_{im})$ and update the estimate of $\boldsymbol{\Pi}$ via (7).

Step 4. Replace Z_{im} and W_{mijk} in (8) by $\hat{E}(Z_{im})$ and $\hat{E}(W_{mijk})$, then update the estimate of $\boldsymbol{\Lambda}_0$ via (8).

Step 5. Replace Z_{im} and W_{mijk} in (9) by $\hat{E}(Z_{im})$ and $\hat{E}(W_{mijk})$, then update the estimate of $\boldsymbol{\beta}$ via (9).

Step 6. Iterate steps 2 to 5 until convergence.

We iterate between the E-step and M-step until the sum of the absolute differences of the estimates at two iterations is less than ϵ, i.e., the stopping criterion is set to be

$$||\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}||_1 + ||\boldsymbol{\Pi}^{(t+1)} - \boldsymbol{\Pi}^{(t)}||_1 + ||\boldsymbol{\Lambda}_0^{(t+1)} - \boldsymbol{\Lambda}_0^{(t)}||_1 < \epsilon,$$

where $||\boldsymbol{\alpha}||_1$ indicates the L_1 norm for $\boldsymbol{\alpha}$, i.e., $||\boldsymbol{\alpha}||_1 = \sum_{i=1}^q |\alpha_i|$ with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)$.

In the following section, we let $\epsilon = 10^{-3}$, and simulation studies are conducted to assess the finite sample performance of the proposed method and in particular, we propose a modified BIC criterion to select the number of subgroups M.

4. Simulation Study

As in the mixture model [21], the number of subgroups M in the proposed model is unknown and will be estimated in 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

$$\text{BIC}(M) = -2\ell(\hat{\boldsymbol{\Lambda}}_0, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Pi}}) + M * q * \log(N), \quad (12)$$

where $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_M)$, $N = \sum_{i=1}^n n_i$ is the sample size, and q is the dimension of $\boldsymbol{\beta}_i$.

In the following, we conduct a set of simulation studies to assess the finite sample performance of our proposed method.

Example 1. *We generate clustered interval-censored data from a latent Cox model with two covariates and three subgroups*

$$\lambda_{ij}(t) = \lambda_{0i}(t) \exp\left(\mathbf{X}_{ij}^\top \boldsymbol{\beta}_i\right), \quad i = 1, \ldots, n. \quad (13)$$

where the covariates X_{ij1} and X_{ij2} are independent and both follow the standard normal distribution. The n clusters are randomly assigned into three subgroups with equal probabilities, i.e., we let $P(i \in \mathcal{G}_1) = P(i \in \mathcal{G}_2) = P(i \in \mathcal{G}_3) = 1/3$, so that $\boldsymbol{\beta}_i = (0.5, 3)$, $\Lambda_{0i}(t) = (t/4)^2$ for $i \in \mathcal{G}_1$, $\boldsymbol{\beta}_i = (-2, -1)$, $\Lambda_{0i}(t) = \log(1 + t/8)$ for $i \in \mathcal{G}_2$ and $\boldsymbol{\beta}_i = (2, -3)$, $\Lambda_{0i}(t) = 2t$ for $i \in \mathcal{G}_3$. The cluster size is set to be m for each cluster. We consider different combinations of the number of clusters (n) and the cluster size (m) to assess the performance of the proposed estimation procedure.

We identify the number of subgroups M by minimizing the modified BIC given in (12). Table 1 presents the mean, median, and standard error (s.d.) of the estimated number of subgroups, denoted

by $\hat{M}$, and the empirical percentage of $\hat{M}$ equal to the true number of subgroups based on 100 *replications. It can be seen from Table 1 that for $(n, m) = (400, 4)$ and $(n, m) = (800, 2)$, the BIC identifies the true number of subgroups among all* 100 *replications, indicating its favorable performance. Table 2 reports the estimation results for the regression coefficients $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$, and $\hat{\boldsymbol{\beta}}_3$, and the mixing probabilities $\hat{\pi}_1$ and $\hat{\pi}_2$, based on* 100 *replications. The proposed method preforms well and yields the estimators with small biases for the case of $(n, m) = (400, 4)$. However, as we increase the number of clusters to 800 and decrease the number of subjects within each cluster to 2, the biases of some estimators become larger even though the total sample sizes are the same for the two cases.*

Table 1. The sample mean, median, and standard error (s.d.) of $\hat{M}$ and the empirical percentage (per) of $\hat{M}$ equal to the true number of subgroups based on 100 replications in Example 1.

(n, m)	Mean	Median	s.d.	Per
$(400, 4)$	3	3	0	1
$(800, 2)$	3	3	0	1

Table 2. The empirical bias and sample standard error (s.d.) of the estimators $\hat{\pi}_1$, $\hat{\pi}_2$, $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$, and $\hat{\boldsymbol{\beta}}_3$ based on 100 replications in Example 1.

(n, m)		π_1	π_2	$\boldsymbol{\beta}_1$		$\boldsymbol{\beta}_2$		$\boldsymbol{\beta}_3$	
				β_{11}	β_{12}	β_{21}	β_{22}	β_{31}	β_{32}
$(400, 4)$	True	$1/3$	$1/3$	0.5	3	-2	-1	2	-3
	Bias	-0.0210	0.0123	0.0194	0.1204	0.2007	0.1023	0.1131	-0.0868
	s.d.	0.0425	0.0278	0.1823	0.3687	0.2379	0.1926	0.2407	0.2273
$(800, 2)$	True	$1/3$	$1/3$	0.5	3	-2	-1	2	-3
	Bias	-0.0192	0.0154	0.1039	0.1790	0.3054	0.1789	0.2167	-0.1255
	s.d.	0.0487	0.0302	0.2097	0.4058	0.3331	0.2723	0.2745	0.2339

Example 2. *We simulate data from a latent Cox model with three covariates and two subgroups*

$$\lambda_{ij}(t) = \lambda_{0i}(t) \exp\left(\mathbf{X}_{ij}^{\top} \boldsymbol{\beta}_i\right), i = 1, \ldots, n. \tag{14}$$

where the covariates $\mathbf{X}_{ij} = (X_{ij1}, X_{ij2}, X_{ij3})^{\top}$ are generated from a multivariate normal distribution with mean zero and a first-order autoregressive covariance structure $\Sigma = (\sigma_{st})$ with $\sigma_{st} = 0.5^{|s-t|}$ for $s, t = 1, 2, 3$. The clusters are randomly assigned into two subgroups with equal probabilities, i.e., we let $P(i \in \mathcal{G}_1) = P(i \in \mathcal{G}_2) = 1/2$, and $\boldsymbol{\beta}_i = (-0.5, -1, -2)$, $\Lambda_{0i}(t) = 4t^2$ for $i \in \mathcal{G}_1$, $\boldsymbol{\beta}_i = (0.5, 1, 2)$, $\Lambda_{0i}(t) = \log(1 + t/8)$ for $i \in \mathcal{G}_2$. The cluster size is set to be m for all n clusters. We consider the cases of $(n, m) = (400, 3)$ and $(600, 2)$. As in Example 1, we estimate the number of subgroups M by minimizing the modified BIC given in (12). Table 3 reports the mean, median, and standard error (s.d.) of the estimator $\hat{M}$ and the empirical percentage of $\hat{M}$ equal to the true number of subgroups based on 100 *replications. We observe that the median of $\hat{M}$ is equal to the true number of subgroups* 2*, and the mean also gets closer to* 2 *as the number of clusters increases. Moreover, the empirical percentage of correctly identifying the true number of subgroups is close to* 1 *as the cluster number becomes moderately large. The estimation results for the regression coefficients $\boldsymbol{\beta}$ and mixing probabilities Π are summarized in Table 4. It can be seen that in terms of the estimation accuracy, the proposed estimation procedure performs quite well and yields the estimators with small biases for $(n, m) = (400, 3)$. Similar with Example 1, as we increase the number of clusters to 600 and decrease the number of subjects within each cluster to 2, the biases of some estimators become larger even though the total sample sizes are the same for the two cases.*

Table 3. The sample mean, median, and standard error (s.d.) of $\hat{M}$ and the empirical percentage (per) of $\hat{M}$ equal to the true number of subgroups based on 100 replications in Example 2.

(n, m)	Mean	Median	s.d.	Per
$n = (400, 3)$	2.02	2	0.14	0.98
$n = (600, 2)$	2	2	0	1

Table 4. The empirical bias and standard error (s.d.) of the estimators $\hat{\pi}_1$, $\hat{\beta}_1$, and $\hat{\beta}_2$ based on 100 replications in Example 2.

(n, m)		π_1	β_1			β_2		
			β_{11}	β_{12}	β_{13}	β_{21}	β_{22}	β_{23}
$(400, 3)$	True	0.5	-0.5	-1	-2	0.5	1	2
	Bias	0.0068	-0.0173	-0.0395	-0.0663	-0.0124	-0.0431	-0.1016
	s.d.	0.0313	0.2585	0.2483	0.2980	0.1174	0.1603	0.2289
$(600, 2)$	True	0.5	-0.5	-1	-2	0.5	1	2
	Bias	-0.0048	-0.0253	-0.0467	-0.0983	-0.0299	-0.0528	-0.1121
	s.d.	0.0311	0.2415	0.2617	0.2921	0.1525	0.1840	0.2219

Example 3. *We next generate data from the Cox model with two covariates*

$$\lambda_{ij}(t) = \lambda_0(t) \exp(\mathbf{X}_{ij}^\top \boldsymbol{\beta}), i = 1, \ldots, n. \tag{15}$$

where the covariates $\mathbf{X}_{ij} = (X_{ij1}, X_{ij2})^\top$ *are generated from a multivariate normal distribution with mean zero and a first-order autoregressive covariance structure* $\Sigma = (\sigma)_{st}$ *with* $\sigma_{st} = 0.5^{|s-t|}$ *for* $s, t = 1, 2$. *We set* $\boldsymbol{\beta} = (1, 3)$, $\Lambda_0(t) = t^2/16$ *and consider* $(n, m) = (200, 4)$ *or* $(400, 2)$. *Note that the model corresponds to the latent Cox model with the true number of subgroups M being* 1.

Based on the BIC criterion given in (12), we estimate the number of subgroups M and report the sample mean, median, and standard error (s.d.) of the estimated number of subgroups $\hat{M}$ *and the empirical percentage of* $\hat{M}$ *equal to the true number of subgroups M based on* 100 *replications. We consider* $(n, m) = (200, 4)$ *and* $(400, 2)$. *The results are given in Table 5. We observe that for each replication, the number of subgroups is correctly identified to be* 1. *The estimation results are summarized in Table 6. We find that the regression coefficients are estimated accurately with small biases for* $(n, m) = (200, 4)$. *Similar with Examples 1 and 2, as we increase the number of clusters to 400 and decrease the number of subjects within each cluster to 2, the biases of some estimators become larger even though the total sample sizes are the same for the two cases. To assess the estimation accuracy of the cumulative baseline hazard rate function, by plotting them in Figure 1, we show the difference between the true cumulative hazard rate function* $\Lambda_0(t)$ *and the estimated baseline cumulative hazard curves* $\hat{\Lambda}_0(t)$. *From Figure 1, it can be seen that two curves are quite close to each other during the time periods of* $(0, 2)$ *and* $(6, 12)$. *However, because there are no sample points falling in the time period of* $(2, 6)$, *the two curves exhibit a significant difference during this time period.*

Table 5. The sample mean, median, and standard error (s.d.) of $\hat{M}$ and the empirical percentage (per) of $\hat{M}$ equal to the true number of subgroups M based on 100 replications in Example 3.

(n, m)	Mean	Median	s.d.	Per
$n = (200, 4)$	1	1	0	1
$n = (400, 2)$	1	1	0	1

Table 6. The empirical bias and sample standard error (s.d.) of the estimator $\hat{\beta}_{11}$, $\hat{\beta}_{12}$ and $\hat{\Lambda}_0(8)$ based on 100 replications in Example 3.

(n, m)		β_{11}	β_{12}	$\Lambda_0(8)$
$(200, 4)$	True	1	3	4
	Bias	0.0021	−0.0074	−0.0710
	s.d.	0.1180	0.2327	0.63830
$(400, 2)$	True	1	3	4
	Bias	0.0099	0.0166	−0.2143
	s.d.	0.1447	0.2348	0.6849

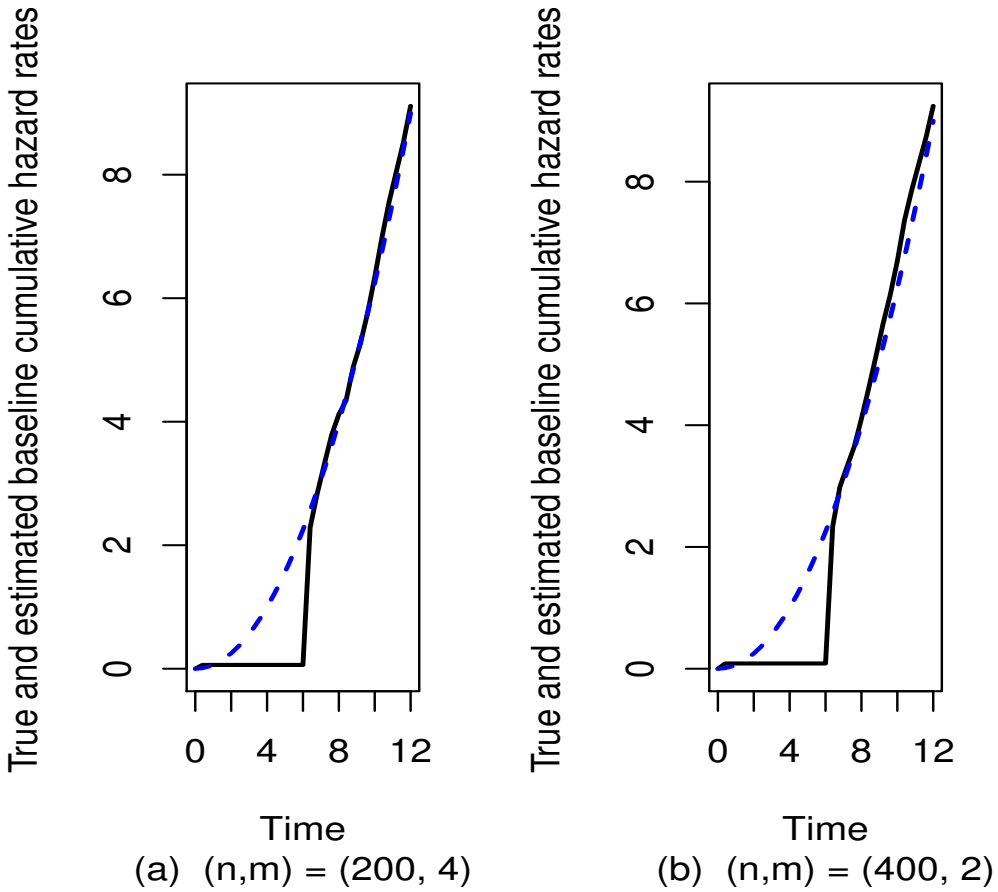

Figure 1. The dotted and solid lines plot the true and estimated baseline cumulative hazard functions, respectively. The estimated baseline cumulative hazard function is the empirical average of the estimated baseline cumulative hazard functions based on 50 replications with $(n, m) = (200, 4)$ or $(400, 2)$ in Example 3.

5. An Application to the Netherlands Twin Study on Migraine

We now apply the proposed model to analyze the Netherlands twin migraine data. The participants were volunteer members of the Netherlands Twin Registry, which is maintained by the Department of Biological Psychology at the Vrije Universiteit in Amsterdam [22]. The data were collected between 1991 and 2002 as part of an ongoing study of health, lifestyle, and genetics involving a large cohort of Dutch twins and their relatives. The primary response of interest in the migraine study is the time when the individual first had a migraine. Since the individuals were followed up on a periodic basis, the time to event may be known only to belong to intervals and hence be interval-censored. The twins form into the clusters with the cluster size 2 and come from different genetic backgrounds, which naturally can be classified into heterogeneous subgroups based on the genetic pro-

files of the twin families, which are not directly observed. Our analysis is based on 3975 monozygotic and dizygotic twin pairs. The left and right endpoints of the interval in which the individual had migraine (in years) are denoted by L_{ij} and R_{ij}, respectively, for the jth individual in the ith cluster. In this dataset, L_{ij} and R_{ij} are random and are independent of the event time. Furthermore, two covariates included in the model are gender (1 = male, 0 = female) and the type of twins (1 = monozygotic, 0 = dizygotic).

To explore the heterogeneity across the twins as indicated by their initial health status, household lifestyle, disease progression, and genetic profiles, we assume that the twins can be classified into a few homogeneous subgroups for each of which the conditional hazard function is postulated by a Cox model. We fit the migraine data by the proposed model with varying M, and the number of subgroups M is estimated by minimizing the BIC criterion function in (12). We found that by the BIC criterion, the optimal M is 3. In Table 7, for the number of subgroups $M = 1, 2, 3, 4$, we report the maximum log-likelihood values (LL), the BIC values (BIC), and the estimated parameters. We found that the model with three subgroups yields the best fit. The twins can be classified into three homogeneous subgroups with mixing probabilities of 77%, 19%, and 4%, respectively. The estimated regression coefficients for three subgroups are also detailed in Table 7. In addition, the baseline cumulative hazard functions for three subgroups are plotted in Figure 2.

Table 7. Estimation results for migraine data with $M = 1, 2, 3, 4$: the number of subgroups (M), the maximum log-likelihood values (LL), the BIC values (BIC), and the estimated parameters.

M	LL	BIC	Estimated Parameters
1	−6600.132	13,218.23	$\hat{\boldsymbol{\beta}} = (-0.5325, -0.1151)$
2	−6579.397	13,194.72	$\hat{\pi}_1 = 0.2003,\ \hat{\boldsymbol{\beta}}_1 = (-0.3968, 0.1302)$ $\hat{\pi}_2 = 0.7997,\ \hat{\boldsymbol{\beta}}_2 = (-0.6769, 0.2541)$
3	−6559.832	13,173.55	$\hat{\pi}_1 = 0.1881,\ \hat{\boldsymbol{\beta}}_1 = (-0.9552, 0.1364)$ $\hat{\pi}_2 = 0.0383,\ \hat{\boldsymbol{\beta}}_2 = (0.3508, 0.5495)$ $\hat{\pi}_3 = 0.7736,\ \hat{\boldsymbol{\beta}}_3 = (-0.4493, -0.2817)$
4	−6597.269	13,266.39	$\hat{\pi}_1 = 0.0211,\ \hat{\boldsymbol{\beta}}_1 = (-0.3509, 0.5877)$ $\hat{\pi}_2 = 0.0193,\ \hat{\boldsymbol{\beta}}_2 = (1.1539, 0.3783)$ $\hat{\pi}_3 = 0.7893,\ \hat{\boldsymbol{\beta}}_3 = (-0.2949, -0.2022)$ $\hat{\pi}_4 = 0.1703,\ \hat{\boldsymbol{\beta}}_4 = (-0.5758, 0.1578)$

For the optimal model selected by BIC criterion, we calculate the empirical standard error and 95%CI of the parameters by the bootstrap method. We repeatedly generated bootstrap samples for G times and obtained bootstrap estimates $(\hat{\boldsymbol{\Pi}}_g, \hat{\boldsymbol{\beta}}_g), g = 1, \ldots, G$ with $G = 500$. Then, the normal-based $100(1-\alpha)\%$ bootstrap interval for π_1 is

$$[\bar{\pi}_1 - z_{\alpha/2}\hat{se}(\pi_1),\ \ \bar{\pi}_1 + z_{\alpha/2}\hat{se}(\pi_1)] \tag{16}$$

where $\bar{\pi}_1 = (\sum_{g=1}^{G} \hat{\pi}_{1g})/G$, $\hat{se}(\pi_1) = \sqrt{[\sum_{g=1}^{G}(\hat{\pi}_{1g} - \bar{\pi}_1)^2]/(G-1)}$. The bootstrap $100(1-\alpha)\%$ percentile interval for π_1 is $[\hat{\pi}_{1L},\ \ \hat{\pi}_{1U}]$; here, $\hat{\pi}_{1L}$ and $\hat{\pi}_{1U}$ are the $(\alpha/2)G$-th and $(1-\alpha/2)G$-th order statistics of $\{\hat{\pi}_{1g}\}_{g=1}^{G}$. The confidence intervals and the empirical standard errors for other parameters can be calculated in a similar way, and the results are reported in Table 8.

Table 8. The estimated parameters for the optimal model.

Parameters	SE	95% Bootstrap CI †	95% Bootstrap CI ‡
π_1	0.0307	[0.1174, 0.2379]	[0.1116, 0.2374]
π_2	0.0044	[0.0231, 0.0404]	[0.0232, 0.0406]
π_3	0.0301	[0.7315, 0.8495]	[0.7058, 0.8308]
β_{11}	0.2428	[−1.3256, −0.3735]	[−1.3017, −0.5147]
β_{12}	0.1384	[−0.1210, 0.4218]	[−0.1386, −0.1386]
β_{21}	0.0613	[0.2426, 0.4829]	[0.2412, 0.4811]
β_{22}	0.0571	[0.4279, 0.6518]	[0.4266, 0.6503]
β_{31}	0.0584	[−0.5606, −0.3314]	[−0.5538, −0.3549]
β_{32}	0.0622	[−0.3997, −0.1557]	[−0.4247, −0.1747]

Notes: SE, the empirical standard error based on the bootstrap samples; CI †, normal-based bootstrap CI; CI ‡, percentile bootstrap CI.

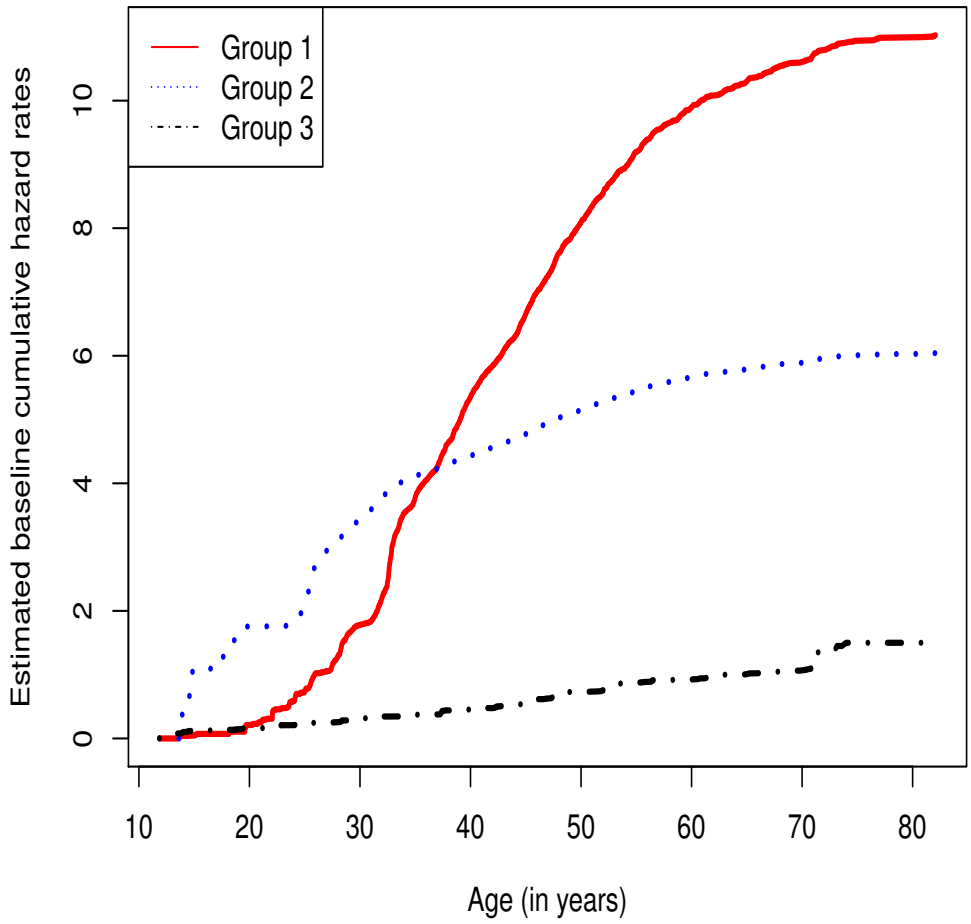

Figure 2. The estimated baseline cumulative hazard functions for migraine data in the optimal model with three subgroups.

Author Contributions: Data curation, X.H. and J.X.; Formal analysis, X.H.; Funding acquisition, J.X.; Investigation, J.X.; Methodology, J.X. 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.

Acknowledgments: The authors are very grateful to Marianne Jonker, D. I. Boomsma, and Aad van der vaart for sharing the migraine data from the Netherlands Twin Registry.

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

References

1. McLachlan, G.J.; Lee, S.X.; Rathnayake, S.I. Finite mixture models. *Annu. Rev. Stat. Its Appl.* **2019**, *6*, 355–378. [CrossRef]
2. Everitt, B. *Finite Mixture Distributions*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
3. Lindsay, B.G. *Mixture Models: Theory, Geometry and Applications*; NSF-CBMS Regional Conference Series in Probability and Statistics; JSTOR: New York, NY, USA, 1995; pp. 1–163.
4. Titterington, D.M.; Afm, S.; Smith, A.F.; Makov, U. *Statistical Analysis of Finite Mixture Distributions*; John Wiley & Sons Incorporated: Chichester, UK; New York, NY, USA, 1985; Volume 198.
5. McLachlan, G.; Chang, S. Mixture modelling for cluster analysis. *Stat. Methods Med. Res.* **2004**, *13*, 347–361. [CrossRef] [PubMed]
6. Aslam, M.; Yousaf, R.; Ali, S. Two-Component Mixture of Transmuted Fréchet Distribution: Bayesian Estimation and Application in Reliability. *Proc. Natl. Acad. Sci. India Sect. A Phys. Sci.* **2021**, *91*, 309–336. [CrossRef]
7. Rachid, A.; Naima, B. The Weibull log-logistic mixture distributions: Model, theory and application to lifetime data. *Qual. Reliab. Eng. Int.* **2021**, *37*, 1599–1627. [CrossRef]
8. Sindhu, T.N.; Hussain, Z.; Aslam, M. Parameter and reliability estimation of inverted Maxwell mixture model. *J. Stat. Manag. Syst.* **2019**, *22*, 459–493. [CrossRef]
9. Lindsey, J.C.; Ryan, L.M. Methods for interval-censored data. *Stat. Med.* **1998**, *17*, 219–238. [CrossRef]
10. Zhang, Z.; Sun, J. Interval censoring. *Stat. Methods Med. Res.* **2010**, *19*, 53–70. [CrossRef]
11. Sun, J. *The Statistical Analysis of Interval-Censored Failure Time Data*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 3.
12. Turnbull, B.W. Nonparametric estimation of a survivorship function with doubly censored data. *J. Am. Stat. Assoc.* **1974**, *69*, 169–173. [CrossRef]
13. Rabinowitz, D.; Tsiatis, A.; Aragon, J. Regression with interval-censored data. *Biometrika* **1995**, *82*, 501–513. [CrossRef]
14. Komárek, A. A new R package for Bayesian estimation of multivariate normal mixtures allowing for selection of the number of components and interval-censored data. *Comput. Stat. Data Anal.* **2009**, *53*, 3932–3947. [CrossRef]
15. Jaspers, S.; Aerts, M.; Verbeke, G.; Beloeil, P.A. A new semi-parametric mixture model for interval censored data, with applications in the field of antimicrobial resistance. *Comput. Stat. Data Anal.* **2014**, *71*, 30–42. [CrossRef]
16. Peng, Y.; Dear, K.B. A nonparametric mixture model for cure rate estimation. *Biometrics* **2000**, *56*, 237–243. [CrossRef] [PubMed]
17. 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. [CrossRef] [PubMed]
18. 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. [CrossRef]
19. McLachlan, G.J.; Peel, D. *Finite Mixture Model*; Wiley: New York, NY, USA, 2000.
20. Ma, L.; Hu, T.; Sun, J. Cox regression analysis of dependent interval-censored failure time data. *Comput. Stat. Data Anal.* **2016**, *103*, 79–90. [CrossRef]
21. McLachlan, G.J.; Krishnan, T. *The EM Algorithm and Extensions*; Wiley: New York, NY, USA, 1997.
22. Boomsma, D.I.; De Geus, E.J.; Vink, J.M.; Stubbe, J.H.; Distel, M.A.; Hottenga, J.J.; Posthuma, D.; Van Beijsterveldt, T.C.; Hudziak, J.J.; Bartels, M.; et al. Netherlands Twin Register: from twins to twin families. *Twin Res. Hum. Genet.* **2006**, *9*, 849–857. [CrossRef] [PubMed]