Check for updates

*Original Research Article*

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2023, Vol. 32(1) 118–132
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09622802221133554
journals.sagepub.com/home/smm
SAGE

# Efficient algorithms for survival data with multiple outcomes using the frailty model

**Xifen Huang[1], Jinfeng Xu[2] and Yunpeng Zhou[3]**

**Abstract**

Survival data with multiple outcomes are frequently encountered in biomedical investigations. An illustrative example comes from Alzheimer's Disease Neuroimaging Initiative study where the cognitively normal subjects may clinically progress to mild cognitive impairment and/or Alzheimer's disease dementia. Transition time from normal cognition to mild cognitive impairment and that from mild cognitive impairment to Alzheimer's disease are expected to be correlated within subjects and the dependence is often accommodated by the frailty (random effects). Estimation in the frailty model unavoidably involves multiple integrations which may be intractable and hence leads to severe computational challenges, especially in the presence of high-dimensional covariates. In this paper, we propose efficient minorization–maximization algorithms in the frailty model for survival data with multiple outcomes. The alternating direction method of multipliers is further incorporated for simultaneous variable selection and homogeneity pursuit via regularization and fusion. Extensive simulation studies are conducted to assess the performance of the proposed algorithms. An application to the Alzheimer's Disease Neuroimaging Initiative data is also provided to illustrate their practical utilities.

**Keywords**

Alternating direction method of multipliers, Alzheimer's Disease Neuroimaging Initiative, homogeneity pursuit, minorization–maximization algorithm, sparsity, the frailty model

## Introduction

Survival data with multiple outcomes arise from time-to-occurrence studies when either of the two or more events (failures) occur for the same subject, or from identical events occurring to related subjects such as family members or classmates. In these studies, failure times are correlated within the cluster (subject or group), violating the independence of failure times assumption required in traditional survival analysis. For example, in the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, two important outcomes of interest are transition time from normal cognition to mild cognitive impairment (MCI) and that from MCI to Alzheimer's disease (AD) dementia, respectively. A primary objective of the ADNI study is to identify relevant clinical variables and biomarkers that are related to and predictive of these two outcomes. It is important to note that AD is the most common type of dementia, accounting for 60% to 80% of age-related dementia cases.[1] Roughly more than 5 million Americans are suffering from memory loss and dementia caused by AD, with a significant increase predicted in the near future if no disease-altering therapeutics are developed. As the therapeutic intervention is most likely to be beneficial in the early stage of the disease, identification of a biosignature that enables an earlier and more accurate diagnosis of AD is an important goal for researchers. This led to the development of MCI, which is a transitional stage between normal aging and the development of AD.[2,3] The shared frailty or random effects model has been widely used to model correlated failure time data[4–7] where the within-subject dependence is accommodated by the frailty. There are several computational challenges for the estimation of the frailty model for survival data. First, the

[1]School of Mathematics, Yunnan Normal University, Kunming, China
[2]Department of Biostatistics, City University of Hong Kong, Hong Kong, Hong Kong
[3]Department of Statistics & Actuarial Science, University of Hong Kong, Hong Kong

**Corresponding author:**
Jinfeng Xu, Department of Biostatistics, City University of Hong Kong.
Email: jinfenxu@cityu.edu.hk

observed likelihood involves multiple integrals which may be intractable and the computationally intensive Laplace approximation or Monte Carlo simulation method is usually employed. In addition, the most popular method in recent years is the h-likelihood approach which is capable of handling more complex random effects structures.[9,11,10,8] Note that the power variance family of frailty distributions may have a closed-form Laplace transform which leads to a tractable likelihood when using a parametric model for the survival time. The EM algorithm can be used to obtain the maximum likelihood estimates due to the closed-form Laplace transform in the Cox survival model.[14,12,13] However, the model parameters in frailty models consist of high-dimensional nonparametric components so that their computations are usually intensive. The existing approaches rely on the EM algorithms which use Newton's method and involve matrix inversion and may not perform well in these high-dimensional situations. So the second computational challenge is that the model parameters include both the parametric component (regression coefficients and the variance of frailty) and the nonparametric component (the baseline hazard function) and the estimation is often conducted by the profile likelihood method which is reliable but computationally intensive.[17,16,15] Last but not least, the computational complexity is even more severe when the number of covariates is large and it is important to conduct both variable selection and homogeneity pursuit[19,20,18] in addition to parameter estimation.

To tackle these challenges, in this paper, we leverage the minorization–maximization (MM) method to develop efficient estimation algorithms in the frailty model for survival data with multiple outcomes. The proposed algorithms exhibit the following advantages. First, the algorithms increase the likelihood at each iteration and reliably converge to the maximum from well-chosen initial values.[21,23,22] Second, unlike existing approaches where parametric and nonparametric components are treated differently and the nonparametric component is profiled out in the estimation,[15,7] our approach adopts a non-profile method which treats the parametric and nonparametric components in the same way, and directly decomposes a high-dimensional objective function into separable low-dimensional functions. This leads to its easy and fast numerical implementation. Third, the algorithms mesh well with regularized estimation in sparse high-dimensional regression models. The concave penalties with good unbiasedness properties such as the smoothly clipped absolute deviations penalty (SCAD, Fan and Li[24]) and the minimax concave penalty (MCP, Zhang et al.[25]) can be effectively and conveniently incorporated into the algorithms for regularization and fusion. Besides, we can also consider the LASSO proposed by Tibshirani[26] and the group LASSO proposed by Utazirubanda et al.[27] for penalized frailty model. For the Cox frailty models with penalized regression, Groll et al.[28] also proposed a penalization approach to distinguish different types of effects. In addition, Rondeau et al.[11] constructs the package for penalized regression using the general frailty model. However, the existing method mentioned above is regularization methods without fused penalty. However, our methods based on the MM principle can incorporate both the regularization and fusion parts in an effective and convenient way. Finally, as the constructed minorizing functions after the decomposition are usually concave and the alternating direction method of multipliers (ADMM, Boyd et al.[29]) has good convergence properties for such loss functions,[29,30] we further pair the ADMM with our algorithms to facilitate simultaneous variable selection and homogeneity pursuit via regularization and fusion.

The rest of the paper is organized as follows. In the "The frailty model" section, we present the frailty model for survival data with multiple outcomes. A non-profile MM algorithm that efficiently decomposes the objective function into separable low-dimensional functions is developed in the "A non-profile MM method" section. In the "Variable selection and homogeneity pursuit via regularization and fusion" section, the ADMM is further incorporated into the algorithm for regularized and fused estimation. The "Simulations" section presents simulation studies to assess the finite sample performance of the proposed methods. The "An application to the ADNI data" section illustrates the proposed methods using the ADNI data. Some concluding remarks and discussions are given in the "Discussion" section.

## The frailty model

Assume that there are $n$ subjects with $J$ outcomes of interest in the study. Let $T_{ij}$ be the survival time, $C_{ij}$ the censoring time, and $X_{ij}$ a $q$-dimensional vector of covariates for the $j$th outcome of the $i$th subject, $i = 1, \ldots, n; j = 1, \ldots, J$. The censoring time $C$ is assumed to be independent of $T$, given the covariates $X$. Define $t_{ij} = \min(T_{ij}, C_{ij})$ and $I_{ij} = I(T_{ij} \leq C_{ij})$, where $I(\cdot)$ denotes the indicator function. The observed data are hence $Y_{obs} = \{t_{ij}, I_{ij}, \mathbf{X}_{ij};\ i = 1, \ldots, n, j = 1, \ldots, J\}$. Let $\omega_i, i = 1, \ldots, n,$ be the subject-specific frailty terms which are independent and identically distributed with the density function $f(\omega_i|\boldsymbol{\theta})$ and the domain $\mathbb{W}$. The (shared) frailty model postulates that given the frailty $\omega_i$, $T_{ij}, j = 1, \ldots, J$ are independent. Furthermore, conditional on $\omega_i$, the hazard function of the event time $T_{ij}$ at time $t$ is

$$\lambda_{ij}(t|\mathbf{X}_{ij}, \omega_i) = \omega_i \lambda_{0j}(t) \exp\{\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}\}, \tag{1}$$

where $\boldsymbol{\beta}$ is an unknown vector of regression coefficients and $\lambda_{0j}(\cdot)$ is the baseline hazard function of the $j$th outcome. Let $\Lambda_{0j}(t) = \int_0^t \lambda_{0j}(u)du, j = 1, \ldots, J$ denote the baseline cumulative hazard functions and define $\Lambda_0 = (\Lambda_{01}, \ldots, \Lambda_{0J})$.

The model parameters consist of three parts, i.e. $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\Lambda_0$. For ease of expression, write $\boldsymbol{\alpha} = (\boldsymbol{\theta},\ \boldsymbol{\beta},\ \Lambda_0)$.

The observed likelihood function takes the form

$$L(\boldsymbol{\alpha}|Y_{obs}) = \prod_{i=1}^{n} \int_{\mathbb{W}} f(\omega_i|\boldsymbol{\theta}) \prod_{j=1}^{J} \left[\lambda_{0j}(t_{ij})\omega_i \exp\left(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}\right)\right]^{I_{ij}} \exp\left[-\Lambda_{0j}(t_{ij})\omega_i \exp\left(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}\right)\right] \mathrm{d}\omega_i. \tag{2}$$

In general, the Laplace transform of the frailty's distribution is intractable and hence (2) does not exhibit an explicit form. To avoid the computationally intensive Laplace approximation or Monte Carlo simulation methods for directly optimizing the objective function (2), we use the MM principle to construct a sequence of minorization functions and optimize them and this strategy results in an algorithm that is numerically much more convenient to implement. Unlike existing methods which are developed for the gamma frailty model where the observed likelihood takes an explicit form,[32,16,31] our method allows (2) to be intractable and also uses a non-profile method which does not require the step that profiles out the non-parametric component and treats the estimation of parametric and nonparametric components in the same fashion.

## Methods

## A description of MM principle

We first briefly describe the MM principle. Let $\ell(\boldsymbol{\alpha}|Y_{obs})$ be the objective function to be maximized, where $\boldsymbol{\alpha}$ denotes the unknown vector of parameters, $\boldsymbol{\alpha} \in \boldsymbol{\Theta}$, and $\boldsymbol{\Theta}$ the parameter space. The MM method iterates between the minorization step and the maximization step until convergence. The minorization step first constructs a surrogate function $Q(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(k)})$ such that

$$Q(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(k)}) \leq \ell(\boldsymbol{\alpha}|Y_{obs}), \quad \forall\, \boldsymbol{\alpha}, \boldsymbol{\alpha}^{(k)} \in \boldsymbol{\Theta} \quad \text{and} \quad Q(\boldsymbol{\alpha}^{(k)}|\boldsymbol{\alpha}^{(k)}) = \ell(\boldsymbol{\alpha}^{(k)}|Y_{obs}), \tag{3}$$

where $\boldsymbol{\alpha}^{(k)}$ denotes the current estimate of $\hat{\boldsymbol{\alpha}}$ in the $k$th iteration. Note that the function $Q(\cdot\,|\boldsymbol{\alpha}^{(k)})$ always lies under $\ell(\cdot\,|Y_{obs})$ and is tangent to it at the point $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(k)}$. 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 $\ell(\boldsymbol{\alpha}|Y_{obs})$. Note that

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

The MM algorithm increases the objective function at each iteration and possesses the ascent property driving the target function $\ell(\boldsymbol{\alpha}|Y_{obs})$ uphill. The key step in the construction of MM algorithms is how to establish the minorizing function $Q(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(k)})$. In the following subsection, the proposed minorizing function $Q(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(k)})$ possesses the good property of parameter separation completely. Hence, under some general and verifiable conditions about the target function $\ell(\boldsymbol{\alpha}|Y_{obs})$ given in Vaida,[33] the proposed MM method may lead to the satisfactory estimates and the related convergence properties of the proposed MM algorithm can be established in a similar way as Huang et al.[31]

## A non-profile MM method

By (2), the observed log-likelihood function

$$\ell(\boldsymbol{\alpha}|Y_{obs}) = \sum_{i=1}^{n} \log \int_{\mathbb{W}} \tau_i(\omega_i|\boldsymbol{\alpha})\, \mathrm{d}\omega_i, \tag{4}$$

where

$$\tau_i(\omega_i|\boldsymbol{\alpha}) = f(\omega_i|\boldsymbol{\theta}) \prod_{j=1}^{J} \left\{ \left[\lambda_{0j}(t_{ij})\omega_i \exp\left(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}\right)\right]^{I_{ij}} \exp\left[-\Lambda_{0j}(t_{ij})\omega_i \exp\left(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}\right)\right] \right\}.$$

Define

$$v_i(\omega_i|\boldsymbol{\alpha}^{(k)}) = \frac{\tau_i(\omega_i|\boldsymbol{\alpha}^{(k)})}{\int_{\mathbb{W}} \tau_i(\omega_i|\boldsymbol{\alpha}^{(k)})\, \mathrm{d}\omega_i},$$

where $\boldsymbol{\alpha}^{(k)}$ denotes the estimates of the parameters in the $k$th iteration. Then (4) can be rewritten as

$$\ell(\boldsymbol{\alpha}|Y_{obs}) = \sum_{i=1}^{n} \log \left[ \int_{\mathbb{W}} \frac{\tau_i(\omega_i|\boldsymbol{\alpha})}{v_i(\omega_i|\boldsymbol{\alpha}^{(k)})} \cdot v_i(\omega_i|\boldsymbol{\alpha}^{(k)})\, \mathrm{d}\omega_i \right]. \tag{5}$$

Recall that Jensen's inequality states

$$\varphi\left[\int_{\mathbb{X}} h(x)\cdot g(x)\,\mathrm{d}x\right] \geq \int_{\mathbb{X}} \varphi\big[h(x)\big]\cdot g(x)\,\mathrm{d}x,$$

where $\mathbb{X}$ is a subset of the real line $\mathbb{R}$, $\varphi(\cdot)$ is a concave function, $h(\cdot)$ is an arbitrary real-valued function defined on $\mathbb{X}$ and $g(\cdot)$ is a density function defined on $\mathbb{X}$. In (5), $v_i(\omega_i|\boldsymbol{\alpha}^{(k)})$ is a density function, choosing $h(x)$ as $\tau_i(\omega_i|\boldsymbol{\alpha})/v_i(\omega_i|\boldsymbol{\alpha}^{(k)})$, by Jensen's inequality, we construct the following surrogate function for $\ell(\boldsymbol{\alpha}|Y_{obs})$,

$$Q_1(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(k)}) = Q_{11}(\boldsymbol{\theta}|\boldsymbol{\alpha}^{(k)}) + Q_{12}(\boldsymbol{\beta},\,\Lambda_0|\boldsymbol{\alpha}^{(k)}) + c_1. \tag{6}$$

In equation (6), the last term $c_1 = -\sum_{i=1}^{n}\int_{\mathbb{W}} v_i(\omega_i|\boldsymbol{\alpha}^{(k)})\log\left[v_i(\omega_i|\boldsymbol{\alpha}^{(k)})\right]\mathrm{d}\omega_i$ does not depend on parameters to be estimated which can be treated as a constant in the following minorizing process. And

$$Q_{11}(\boldsymbol{\theta}|\boldsymbol{\alpha}^{(k)}) = \sum_{i=1}^{n}\int_{\mathbb{W}} \log\left[f(\omega_i|\boldsymbol{\theta})\right]\cdot v_i(\omega_i|\boldsymbol{\alpha}^{(k)})\,\mathrm{d}\omega_i, \tag{7}$$

$$Q_{12}(\boldsymbol{\beta},\,\Lambda_0|\boldsymbol{\alpha}^{(k)}) = \sum_{i=1}^{n}\sum_{j=1}^{J}\left[I_{ij}\log\left(\lambda_{0j}(t_{ij})\right)+I_{ij}\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}-A_i^{(k)}\Lambda_{0j}(t_{ij})\exp\left(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}\right)\right], \tag{8}$$

with $A_i^{(k)} = \int_{\mathbb{W}}\omega_i\cdot v_i(\omega_i|\boldsymbol{\alpha}^{(k)})\,\mathrm{d}\omega_i$, $i = 1,\,\ldots,\,n$. The minorizing function $Q_1(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(k)})$ separates the parameters $\boldsymbol{\theta}$ and $(\boldsymbol{\beta},\,\Lambda_0)$ into (7) and (8), respectively.

Next, we further separate $\boldsymbol{\beta}$ and $\Lambda_0$ in (8). As in Lange and Zhou,[34] we use the arithmetic-geometric mean inequality

$$-\prod_{i=1}^{n} x_i^{a_i} \geq -\sum_{i=1}^{n}\frac{a_i}{\|\mathbf{a}\|_1}x_i^{\|\mathbf{a}\|_1}. \tag{9}$$

Here $x_i$, $a_i$ are nonnegative. Choosing $x_1 = \Lambda_{0j}(t_{ij})/\Lambda_{0j}^{(k)}(t_{ij})$ and $x_2 = \exp\left(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}\right)/\exp\left(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}^{(k)}\right)$ in inequality (9), we obtain the following surrogate function for (8)

$$\begin{aligned} &Q_2(\boldsymbol{\beta},\,\Lambda_0|\boldsymbol{\alpha}^{(k)}) \\ &\quad= \sum_{i=1}^{n}\sum_{j=1}^{J}\left[I_{ij}\log\left(\lambda_{0j}(t_{ij})\right) + I_{ij}\mathbf{X}_{ij}^{\top}\boldsymbol{\beta} - \frac{A_i^{(k)}\exp\left(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}^{(k)}\right)}{2\Lambda_{0j}^{(k)}(t_{ij})}\Lambda_{0j}(t_{ij})^2 - \frac{A_i^{(k)}\Lambda_{0j}^{(k)}(t_{ij})}{2\exp\left(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}^{(k)}\right)}\exp\left(2\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}\right)\right], \\ &\quad\hat{=} Q_{21}(\Lambda_0|\boldsymbol{\alpha}^{(k)}) + Q_{22}(\boldsymbol{\beta}|\boldsymbol{\alpha}^{(k)}) \end{aligned} \tag{10}$$

where

$$Q_{21}(\Lambda_0|\boldsymbol{\alpha}^{(k)}) = \sum_{i=1}^{n}\sum_{j=1}^{J}\left[I_{ij}\log\left(\lambda_{0j}(t_{ij})\right) - \frac{A_i^{(k)}\exp\left(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}^{(k)}\right)}{2\Lambda_{0j}^{(k)}(t_{ij})}\Lambda_{0j}(t_{ij})^2\right], \tag{11}$$

and

$$Q_{22}(\boldsymbol{\beta}|\boldsymbol{\alpha}^{(k)}) = \sum_{i=1}^{n}\sum_{j=1}^{J}\left[I_{ij}\mathbf{X}_{ij}^{\top}\boldsymbol{\beta} - \frac{A_i^{(k)}\Lambda_{0j}^{(k)}(t_{ij})}{2\exp\left(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}^{(k)}\right)}\exp\left(2\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}\right)\right]. \tag{12}$$

Now $\boldsymbol{\beta}$ and $\Lambda_0$ are separated in the maximization. To update $\Lambda_0$, we maximize (11). For ease of computation, we set the first order derivative of (11) equal to 0, then we have

$$\mathrm{d}\hat{\Lambda}_{0j}(t_{ij}) = \frac{I_{ij}}{\sum_{r=1}^{n} I(t_{rj}\geq t_{ij})A_r^{(k)}\exp\left(\mathbf{X}_{rj}^{\top}\boldsymbol{\beta}^{(k)}\right)}, \tag{13}$$

To more conveniently update $\boldsymbol{\beta}$, we apply Jensen's inequality to the concave function $-\exp(\cdot)$ in $Q_{22}(\boldsymbol{\beta}|\boldsymbol{\alpha}^{(k)})$ by rewriting

$$2\mathbf{X}_{ij}^{\top}\boldsymbol{\beta} = \sum_{p=1}^{q}\delta_{pij}\left[2\delta_{pij}^{-1}X_{pij}(\beta_p - \beta_p^{(k)}) + 2\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}^{(k)}\right],$$

where $\delta_{pij} = |X_{pij}|/\sum_{p=1}^{q}|X_{pij}|$. The minorizing function for $Q_{22}(\boldsymbol{\beta}|\boldsymbol{\alpha}^{(k)})$ is then

$$Q_{23}(\beta_1, \ldots, \beta_q|\boldsymbol{\alpha}^{(k)}) \doteq \sum_{p=1}^{q} Q_{23p}(\beta_p|\boldsymbol{\alpha}^{(k)}), \tag{14}$$

where

$$Q_{23p}(\beta_p|\boldsymbol{\alpha}^{(k)}) = \sum_{i=1}^{n}\sum_{j=1}^{J}\left\{I_{ij}X_{pij}\beta_p - \frac{\delta_{pij}A_i^{(k)}\Lambda_{0j}^{(k)}(t_{ij})\exp\left[2\delta_{pij}^{-1}X_{pij}(\beta_p - \beta_p^{(k)}) + 2\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}^{(k)}\right]}{2\exp(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}^{(k)})}\right\}, \tag{15}$$

for $p = 1, \ldots, q$. To summarize, we construct a non-profile MM (NPMM) algorithm where the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ are updated by maximizing the the surrogate function

$$Q_{npmm}(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{\alpha}^{(k)}) = Q_{11}(\boldsymbol{\theta}|\boldsymbol{\alpha}^{(k)}) + \sum_{p=1}^{q} Q_{23p}(\beta_p|\boldsymbol{\alpha}^{(k)}), \tag{16}$$

and $\Lambda_0$ is updated with the explicit formula (13). From (16), we can see that $Q_{npmm}(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{\alpha}^{(k)})$ is a sum of $q + 1$ univariate functions. Hence, the updates in the maximization step only involves $q + 1$ simple univariate optimizations. It is worth noting that this parameter-separable feature of the algorithm also facilitates the incorporation of simple off-the-shelf accelerators for boosting computational effectiveness. The proposed algorithm is summarized as follows.

**Algorithm 1.** The non-profile MM (NPMM) algorithm.

**Step 1.** Give initial values of $\theta$, $\beta$ and $\Lambda_0$.
**Step 2.** Update the estimate of $\theta$ via (3).
**Step 3.** Update the estimate of $\beta_p$ via (25).
**Step 4.** Update the estimate of $\Lambda_0$ via (24).
**Step 5.** Iterate steps 2 to 4 until convergence.

## Variable selection and homogeneity pursuit via regularization and fusion

In high-dimensional regression, the regularization often yields effective variable selection.[24] In the meantime, pairwise fusion also helps to cluster regression coefficients into different homogeneity groups. In the regularized and fused estimation, the objective function is

$$\ell(\boldsymbol{\theta}, \boldsymbol{\beta}, \Lambda_0|Y_{obs}) - n\sum_{p=1}^{q} P_1(|\beta_p|, a\lambda) - \sum_{1\leqslant p<l\leqslant q} P_2(|\beta_p - \beta_l|, \lambda), \tag{17}$$

where $\ell(\boldsymbol{\theta}, \boldsymbol{\beta}, \Lambda_0|Y_{obs})$ is the log-likelihood function, $q$ is the dimension of $\boldsymbol{\beta}$, $P_1(\cdot, a\lambda)$ and $P_2(\cdot, \lambda)$ are nonnegative concave penalty functions, $\lambda \geqslant 0$ is a tuning parameter and $a \geqslant 0$ is another tuning parameter that controls the relative ratio between sparsity and fusion penalties. Define

$$\ell^{P}(\boldsymbol{\alpha}|Y_{obs}) = \ell(\boldsymbol{\theta}, \boldsymbol{\beta}, \Lambda_0|Y_{obs}) - n\sum_{p=1}^{q} P(|\beta_p|, \lambda). \tag{18}$$

Using the non-profile MM technique in (16), the minorization function for (18) is

$$Q_{npmm}(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{\alpha}^{(k)}) - n\sum_{p=1}^{q} P(|\beta_p|, \lambda). \tag{19}$$

When $P(\cdot, \lambda)$ is piecewise differentiable, nondecreasing and concave on $(0, \infty)$ such as MCP and SCAD penalties,[35] the penalty term $-P(\cdot, \lambda)$ can be minorized by a local quadratic approximation form as

$$-P(|\beta_p|, \lambda) \geqslant -P(|\beta_p^{(k)}|, \lambda) - \frac{[\beta_p^2 - (\beta_p^{(k)})^2]P'(|\beta_p^{(k)}|_+, \lambda)}{2|\beta_p^{(k)}|}. \tag{20}$$

Combining (20) with (16), we obtain the surrogate function for (18) as follows

$$\begin{aligned} Q^{P}_{npmm}(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{\alpha}^{(k)}, \lambda) &= Q_{npmm}(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{\alpha}^{(k)}) - n\sum_{p=1}^{q}\frac{\beta_p^2 \cdot P'(|\beta_p^{(k)}|_+, \lambda)}{2|\beta_p^{(k)}|} + c_2 \\ &= Q_{11}(\boldsymbol{\theta}|\boldsymbol{\alpha}^{(k)}) + \sum_{p=1}^{q}\left[Q_{23p}(\beta_p|\boldsymbol{\alpha}^{(k)}) - \frac{n\beta_p^2 \cdot P'(|\beta_p^{(k)}|_+, \lambda)}{2|\beta_p^{(k)}|}\right] + c_2, \end{aligned} \tag{21}$$

where $c_2 = n\sum_{p=1}^{q}\{|\beta_p^{(k)}| \cdot P'(|\beta_p^{(k)}|_+, \lambda)/2 - P(|\beta_p^{(k)}|, \lambda)\}$ which does not contain any parameters to be estimated and can be omitted. It follows that the surrogate function for (17) is

$$Q^{P_1}_{npmm}(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{\alpha}^{(k)}, a\lambda) - \sum_{1\leqslant p<l\leqslant q} P_2(|\beta_p - \beta_l|, \lambda). \tag{22}$$

Next we incorporate the ADMM to deal with the fusion term. We first introduce a new set of parameters $\eta_{pl} = \beta_p - \beta_l$. The maximization of (22) is reformulated as the constraint optimization problem (23),

$$\begin{gathered} Q^{P_1}_{npmm}(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{\alpha}^{(k)}, a\lambda) - \sum_{1\leqslant p<l\leqslant q} P_2(|\eta_{pl}|, \lambda), \\ \text{subject to } \beta_p - \beta_l - \eta_{pl} = 0, \end{gathered} \tag{23}$$

where $\boldsymbol{\eta} = \{\eta_{pl},\ p < l\}^{\top}$. By the augmented Lagrangian method, the estimates can be obtained by maximizing

$$\begin{aligned} &G_{npmm}(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\nu}) \\ &= Q^{P_1}_{npmm}(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{\alpha}^{(k)}, a\lambda) - \sum_{p<l} P_2(|\eta_{pl}|, \lambda) - \sum_{p<l}\nu_{pl}(\beta_p - \beta_l - \eta_{pl}) - \frac{\rho}{2}\sum_{p<l}(\beta_p - \beta_l - \eta_{pl})^2, \end{aligned} \tag{24}$$

where the dual variables $\boldsymbol{\nu} = \{\nu_{pl},\ p < l\}^{\top}$ are Lagrange multipliers and $\rho$ is the penalty parameter. Specifically, for given $(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\nu})$, the maximization of (24) is equivalent to maximizing

$$-\frac{\rho}{2}(\sigma_{pl} - \eta_{pl})^2 - P_2(|\eta_{pl}|, \lambda) \tag{25}$$

with respect to $\eta_{pl}$, where $\sigma_{pl} = \beta_p - \beta_l + \rho^{-1}\nu_{pl}$. The maximizer with respect to $\eta_{pl}$ is unique and we can obtain a closed-form expression for $L_1$, MCP and SCAD penalties, respectively. The closed-form solution for the $L_1$ penalty is

$$\hat{\eta}_{pl} = \mathrm{ST}(\sigma_{pl}, \lambda/\rho), \tag{26}$$

where $\mathrm{ST}(t, \lambda) = \mathrm{sign}(t)(|t| - \lambda)_+$ is the soft thresholding rule. For the MCP penalty with $\gamma > 1/\rho$, it is

$$\hat{\eta}_{pl} = \begin{cases} \dfrac{\mathrm{ST}(\sigma_{pl}, \lambda/\rho)}{1 - 1/(\gamma\rho)}, & \text{if } |\sigma_{pl}| \leqslant \gamma\lambda, \\ \sigma_{pl}, & \text{if } |\sigma_{pl}| > \gamma\lambda. \end{cases} \tag{27}$$

For the SCAD penalty with $\gamma > 1 + 1/\rho$, it is

$$\hat{\eta}_{pl} = \begin{cases} \mathrm{ST}(\sigma_{pl}, \lambda/\rho), & \text{if } |\sigma_{pl}| \leqslant \lambda + \lambda/\rho, \\ \dfrac{\mathrm{ST}(\sigma_{pl}, \gamma\lambda/((\gamma-1)\rho))}{1 - 1/((\gamma-1)\rho)}, & \text{if } \lambda + \lambda/\rho < |\sigma_{pl}| \leqslant \gamma\lambda, \\ \sigma_{pl}, & \text{if } |\sigma_{pl}| > \gamma\lambda. \end{cases} \tag{28}$$

To update $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, for given $(\boldsymbol{\eta}, \boldsymbol{\nu})$, we maximize

$$L_{npmm}(\boldsymbol{\theta}, \boldsymbol{\beta}) = Q^{P_1}_{npmm}(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{\alpha}^{(k)}, a\lambda) - \frac{\rho}{2}\sum_{p<l}\left[(e_p - e_l)^{\top}\boldsymbol{\beta} - \eta_{pl} + \rho^{-1}\nu_{pl}\right]^2, \tag{29}$$

by setting the derivatives $\partial L_{npmm}(\boldsymbol{\theta}, \boldsymbol{\beta})/\partial\boldsymbol{\theta}$ and $\partial L_{npmm}(\boldsymbol{\theta}, \boldsymbol{\beta})/\partial\boldsymbol{\beta}$ to zero. Here $e_p$ is the $p \times 1$ vector whose $p$th element is 1, and $\mathbf{Z} = \{(e_i - e_j),\ i < j\}^{\top}$. Moreover, a simple minorization step for the last term is added so that all regression parameters $\beta_p$ are separated from each other by using Jensen's inequality. Finally, the estimate of $\nu_{pl}$ is updated as

$$\nu_{pl}^{(k+1)} = \nu_{pl}^{(k)} + \rho\left(\beta_p^{(k+1)} - \beta_l^{(k+1)} - \eta_{pl}^{(k+1)}\right). \tag{30}$$

To summarize, the algorithm for regularized and fused estimation is as follows.

**Algorithm 2.** Regularized and fused estimation using MM and ADMM algorithms.

**Step 1.** Give initial values of $\theta$, $\beta$, $\Lambda_0$ and $\eta$, $\nu$.
**Step 2.** Update the estimates of $\theta$ via (3).
**Step 3.** Update the estimates of $\beta_p$ ($p = 1, \ldots, q$) via (16).
**Step 4.** Update the estimates of $\Lambda_0$ via (24).
**Step 5.** Update the estimates of $\eta$ by the formula given in (13), (14) and (15).
**Step 6.** Update $\nu$ by $\nu_{pl}^{(k+1)} = \nu_{pl}^{(k)} + \rho\left(\beta_p^{(k+1)} - \beta_l^{(k+1)} - \eta_{pl}^{(k+1)}\right)$.
**Step 7.** Iterate steps 2 to 6 until convergence.

We track the iteration process based on the primal residual $\mathbf{r}^{(k+1)} = \mathbf{Z}\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\eta}^{(k+1)}$ and the stopping criterion is set to be

$$\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}\|_2 + \|\boldsymbol{\lambda}_0^{(k+1)} - \boldsymbol{\lambda}_0^{(k)}\|_2 + \|\mathbf{r}^{(k+1)}\|_2 < \varepsilon$$

for some small value $\varepsilon > 0$. In fact, we carry out some rounding to achieve zeros via the selected tuning parameter $a\lambda$. If the absolute value of each $\beta_i$ is less than the selected value of $a\lambda$, we enforce the final estimated value of $\beta_i$ to be 0. Similarly, for the fusion component, we declare that these $\beta_i$ and $\beta_j$ are in the same group when $|\beta_i - \beta_j| < a\lambda$ for $i \neq j$ with $i, j = 1, \ldots, q$. We enforce the final parameter estimations in the same group to be exactly the same and be equal to the average value of all parameter values in this group.

In practice, the tuning parameter $\lambda$ can be selected by data-driven model selection criteria, such as Bayesian information criterion (BIC)[36] or generalized cross-validation GCV.[37]Inspired by Ma and Huang,[38] here we use the modified BIC type criteria to select $\lambda$ by minimizing the following BIC function

$$\mathrm{BIC}_\lambda = -2\ell(\hat{\boldsymbol{\alpha}}) + C_n(\hat{S} + 1)\log(n), \tag{31}$$

$C_n = \max\{1, \log[\log(q+1)]\}$, $q$ is the dimension of $\boldsymbol{\beta}$, and the degree of freedom $\hat{S}$ is the number of distinct nonzero parameters in $\hat{\boldsymbol{\beta}}$. For the exact process of tuning parameters selection, we adopt a similar procedure as obtaining the optimal tuning parameter for LASSO and elastic net regression from generally used R package "glmnet".[39] We first find the the proper range of $\lambda$ values by searching from $(0, +\infty)$ using this BIC criteria with an initial sequence $y_1, y_2, \ldots, y_{k_1}$. The solution path can be plotted using this sequence and $y_i$ is selected where the minimum BIC is obtained. Then, grid points $x_1, x_2, \ldots, x_{k_2}$ are constructed in range $(y_{i-1}, y_{i+1})$ for a more accurate search where the optimal $\lambda$ is selected from this sequence with the minimum BIC score. In addition to the tuning parameters, for both sparsity and fusion penalties using MCP and SCAD from (27) and (28). Following Fan and Li[40], Ma and Huang[38], we choose $a = 1/3$, $\rho = 1$, $\gamma = 3.7$ (SCAD), and $\gamma = 3$ (MCP) in this study.

## Simulations

In this section, we conducted two sets of numerical experiments to assess the practical performance of the proposed non-profile MM method and regularization and fusion method. The following Example 1 is intended to evaluate the performance of the non-profile MM method without regularisation. While Example 2 considers the variable selection and homogeneity pursuit with regularization and fusion. All the simulation experiments were coded in R and ran in a desktop with Intel(R) Core(TM) i7-9700 with CPU 3.00 GHz.

Example 1. We generate data from the frailty model

$$\lambda_j(t|\mathbf{X}_{ij}, \omega_i) = \omega_i \lambda_{0j}(t)\exp\{\mathbf{X}_{ij}^\top \boldsymbol{\beta}\}, \quad \omega_i \sim \begin{cases} \text{Gamma}(1/\theta, 1/\theta), \\ \text{Log-normal}(0, \theta), \\ \text{Inverse Gaussian}(\theta, \theta^2), \end{cases}$$

with the number of outcomes $J = 2$. The sample size $n$ is set to be 100, 200 and 300, respectively. The true value of regression vector $\boldsymbol{\beta}$ is kept to be $(-2_{10}^\top, 3_{10}^\top)^\top$ with dimension $q = 20$ for all three frailty models, here the notation $c_m^\top$ stands for a row vector as $(c, \ldots, c)$ of length $m$. The baseline hazard function is set to be $\lambda_{01}(t) = 3$, $\lambda_{02}(t) = 5/(1+5t)$ and all $X_i$'s

are generated from independent uniform distribution between 0 and 0.5. The censoring times are generated from Uniform distribution U(0, $b$) to yield two censoring proportions ($P_{cen}$) of about 0.3 and 0.1 separately. Moreover, we choose the true values of $\theta$ in a way that keeps $\mathrm{Var}(\omega_i)$ fixed across the three distributions of $\omega_i$ and we let the value of $\mathrm{Var}(\omega_i)$ vary from 0.3 to 1, i.e.

$$(1)\,\mathrm{Var}(\omega_i) = 0.3, \quad \text{then}\,\theta = \begin{cases} 0.3, & \omega_i \sim \text{Gamma}(1/\theta, 1/\theta), \\ \log(0.5 + \sqrt{2.2}/2) \approx 0.2164, & \omega_i \sim \text{Log-normal}(0, \theta), \\ 0.3, & \omega_i \sim \text{Inverse Gaussian}(\theta, \theta^2), \end{cases}$$

$$(2)\,\mathrm{Var}(\omega_i) = 1, \quad \text{then}\,\theta = \begin{cases} 1, & \omega_i \sim \text{Gamma}(1/\theta, 1/\theta), \\ \log(0.5 + \sqrt{5}/2) \approx 0.4812, & \omega_i \sim \text{Log-normal}(0, \theta), \\ 1, & \omega_i \sim \text{Inverse Gaussian}(\theta, \theta^2). \end{cases}$$

For better comparison, we compare the proposed non-profile MM method with the coxph function in the survival R package and the frailtyHL function in the frailtyHL R package[9,11] to get a sense of the differences in speed, bias, and empirical standard errors. Noticing that simple off-the-shelf accelerators[41,42] can be incorporated, we accelerated the proposed non-profile MM algorithm with the squared iterative method (SqS1). Based on 500 replications, we calculate the average

**Table 1.** The simulation results for Log-normal frailty model with $\mathrm{Var}(\omega_i) = 0.3$ ($\theta \approx 0.2164$) and censoring proportion $P_{cen} = 0.1$.

| $n = 100$ | | MM method | | | coxph function | | | frailtyHL function | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | T | | 42.34 | | | 0.05 | | | 13.16 | |
| | | \|Bias\| | SE | MSE | \|Bias\| | SE | MSE | \|Bias\| | SE | MSE |
| | $\hat{\mathrm{Var}}(\omega_i)$ | 0.14 | 0.25 | 0.08 | 0.73 | 0.96 | 1.45 | 0.35 | 0.60 | 0.48 |
| | $\theta$ | 0.08 | 0.17 | 0.03 | 0.27 | 0.26 | 0.14 | 0.16 | 0.22 | 0.07 |
| | $\beta_1$ | 0.02 | 0.69 | 0.47 | 0.23 | 0.80 | 0.69 | 0.14 | 0.75 | 0.58 |
| | $\beta_5$ | 0.03 | 0.68 | 0.46 | 0.32 | 0.78 | 0.71 | 0.15 | 0.72 | 0.54 |
| | $\beta_{10}$ | 0.05 | 0.70 | 0.49 | 0.29 | 0.77 | 0.67 | 0.17 | 0.74 | 0.57 |
| | $\beta_{15}$ | 0.11 | 0.77 | 0.60 | 0.35 | 0.79 | 0.74 | 0.28 | 0.81 | 0.73 |
| | $\beta_{20}$ | 0.08 | 0.70 | 0.49 | 0.36 | 0.77 | 0.72 | 0.26 | 0.76 | 0.64 |
| $n = 200$ | | MM method | | | coxph function | | | frailtyHL function | | |
| | T | | 58.29 | | | 0.13 | | | 120.12 | |
| | | \|Bias\| | SE | MSE | \|Bias\| | SE | MSE | \|Bias\| | SE | MSE |
| | $\hat{\mathrm{Var}}(\omega_i)$ | 0.06 | 0.20 | 0.04 | 0.29 | 0.39 | 0.24 | 0.09 | 0.27 | 0.08 |
| | $\theta$ | 0.03 | 0.12 | 0.01 | 0.13 | 0.15 | 0.04 | 0.05 | 0.13 | 0.02 |
| | $\beta_1$ | 0.02 | 0.47 | 0.22 | 0.12 | 0.47 | 0.23 | 0.04 | 0.48 | 0.23 |
| | $\beta_5$ | 0.04 | 0.45 | 0.20 | 0.13 | 0.47 | 0.23 | 0.07 | 0.49 | 0.24 |
| | $\beta_{10}$ | 0.01 | 0.47 | 0.22 | 0.13 | 0.50 | 0.26 | 0.03 | 0.51 | 0.26 |
| | $\beta_{15}$ | 0.04 | 0.48 | 0.23 | 0.21 | 0.52 | 0.31 | 0.11 | 0.51 | 0.27 |
| | $\beta_{20}$ | 0.04 | 0.46 | 0.21 | 0.18 | 0.52 | 0.30 | 0.10 | 0.51 | 0.27 |
| $n = 300$ | | MM method | | | coxph function | | | frailtyHL function | | |
| | T | | 76.66 | | | 0.25 | | | 578.85 | |
| | | \|Bias\| | SE | MSE | \|Bias\| | SE | MSE | \|Bias\| | SE | MSE |
| | $\hat{\mathrm{Var}}(\omega_i)$ | 0.05 | 0.17 | 0.03 | 0.13 | 0.24 | 0.07 | 0.03 | 0.18 | 0.03 |
| | $\theta$ | 0.02 | 0.10 | 0.01 | 0.07 | 0.11 | 0.02 | 0.02 | 0.10 | 0.01 |
| | $\beta_1$ | 0.01 | 0.39 | 0.15 | 0.07 | 0.39 | 0.15 | 0.03 | 0.40 | 0.16 |
| | $\beta_5$ | 0.01 | 0.38 | 0.14 | 0.06 | 0.38 | 0.14 | 0.03 | 0.40 | 0.16 |
| | $\beta_{10}$ | 0.01 | 0.37 | 0.13 | 0.05 | 0.38 | 0.14 | 0.01 | 0.40 | 0.16 |
| | $\beta_{15}$ | 0.03 | 0.36 | 0.13 | 0.10 | 0.39 | 0.16 | 0.01 | 0.39 | 0.15 |
| | $\beta_{20}$ | 0.03 | 0.40 | 0.16 | 0.08 | 0.41 | 0.17 | 0.05 | 0.42 | 0.17 |

MSE: mean squared error; MM: minorization-maximization; SE: standard error.

**Table 2.** The simulation results via MM method at sample size 300 with different censoring proportion, variance, and frailty distribution. Note: (Var($\omega_i$), $P_{cen}$) is simplified by ($V$, $P$).

*Gamma frailty model*

| | $(V, P) = (0.3, 0.3)$ | | | $(V, P) = (0.3, 0.1)$ | | | $(V, P) = (1, 0.3)$ | | | $(V, P) = (1, 0.1)$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T | 19.95 | | | 28.67 | | | 88.26 | | | 53.15 | | |
| | \|Bias\| | SE | MSE | \|Bias\| | SE | MSE | \|Bias\| | SE | MSE | \|Bias\| | SE | MSE |
| $\hat{Var}(\omega_i)$ | 0.02 | 0.07 | <0.01 | 0.01 | 0.07 | <0.01 | <0.01 | 0.18 | 0.03 | <0.01 | 0.14 | 0.02 |
| $\theta$ | 0.02 | 0.07 | <0.01 | 0.01 | 0.07 | <0.01 | <0.01 | 0.18 | 0.03 | <0.01 | 0.14 | 0.02 |
| $\beta_1$ | <0.01 | 0.40 | 0.16 | <0.01 | 0.38 | 0.14 | <0.01 | 0.50 | 0.25 | 0.01 | 0.46 | 0.21 |
| $\beta_5$ | 0.02 | 0.38 | 0.14 | 0.02 | 0.38 | 0.14 | <0.01 | 0.50 | 0.25 | 0.01 | 0.47 | 0.22 |
| $\beta_{10}$ | <0.01 | 0.39 | 0.15 | 0.01 | 0.37 | 0.13 | <0.01 | 0.50 | 0.25 | 0.01 | 0.46 | 0.21 |
| $\beta_{15}$ | 0.03 | 0.41 | 0.17 | 0.01 | 0.40 | 0.16 | <0.01 | 0.50 | 0.25 | 0.04 | 0.44 | 0.19 |
| $\beta_{20}$ | 0.04 | 0.43 | 0.18 | 0.02 | 0.42 | 0.17 | 0.03 | 0.50 | 0.25 | 0.02 | 0.48 | 0.23 |

*Log-normal frailty model*

| | $(V, P) = (0.3, 0.3)$ | | | $(V, P) = (0.3, 0.1)$ | | | $(V, P) = (1, 0.3)$ | | | $(V, P) = (1, 0.1)$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T | 94.33 | | | 76.66 | | | 89.28 | | | 48.35 | | |
| | \|Bias\| | SE | MSE | \|Bias\| | SE | MSE | \|Bias\| | SE | MSE | \|Bias\| | SE | MSE |
| $\hat{Var}(\omega_i)$ | 0.06 | 0.18 | 0.04 | 0.05 | 0.17 | 0.03 | 0.11 | 0.50 | 0.26 | 0.07 | 0.41 | 0.17 |
| $\theta$ | 0.04 | 0.11 | 0.01 | 0.02 | 0.10 | 0.01 | 0.03 | 0.15 | 0.02 | 0.02 | 0.12 | 0.01 |
| $\beta_1$ | 0.01 | 0.40 | 0.16 | 0.01 | 0.39 | 0.15 | 0.01 | 0.44 | 0.19 | 0.01 | 0.39 | 0.15 |
| $\beta_5$ | 0.03 | 0.39 | 0.15 | 0.01 | 0.38 | 0.14 | 0.01 | 0.46 | 0.21 | 0.02 | 0.40 | 0.16 |
| $\beta_{10}$ | 0.03 | 0.41 | 0.17 | 0.01 | 0.37 | 0.13 | 0.02 | 0.44 | 0.19 | 0.01 | 0.38 | 0.14 |
| $\beta_{15}$ | 0.02 | 0.40 | 0.16 | 0.03 | 0.36 | 0.13 | 0.05 | 0.44 | 0.19 | 0.01 | 0.39 | 0.15 |
| $\beta_{20}$ | 0.04 | 0.41 | 0.17 | 0.03 | 0.40 | 0.16 | 0.05 | 0.41 | 0.17 | 0.02 | 0.40 | 0.16 |

*Inverse Gaussian frailty model*

| | $(V, P) = (0.3, 0.3)$ | | | $(V, P) = (0.3, 0.1)$ | | | $(V, P) = (1, 0.3)$ | | | $(V, P) = (1, 0.1)$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T | 82.91 | | | 115.37 | | | 51.95 | | | 27.35 | | |
| | \|Bias\| | SE | MSE | \|Bias\| | SE | MSE | \|Bias\| | SE | MSE | \|Bias\| | SE | MSE |
| $\hat{Var}(\omega_i)$ | 0.04 | 0.12 | 0.01 | 0.02 | 0.09 | <0.01 | 0.16 | 0.47 | 0.24 | <0.01 | <0.01 | <0.01 |
| $\theta$ | 0.04 | 0.12 | 0.01 | 0.02 | 0.09 | <0.01 | 0.16 | 0.47 | 0.24 | <0.01 | <0.01 | <0.01 |
| $\beta_1$ | 0.01 | 0.52 | 0.27 | <0.01 | 0.46 | 0.21 | 0.02 | 0.46 | 0.21 | 0.05 | 0.40 | 0.16 |
| $\beta_5$ | 0.02 | 0.50 | 0.25 | <0.01 | 0.45 | 0.20 | 0.03 | 0.45 | 0.20 | 0.02 | 0.40 | 0.16 |
| $\beta_{10}$ | 0.03 | 0.47 | 0.22 | 0.02 | 0.46 | 0.21 | 0.07 | 0.44 | 0.19 | 0.03 | 0.41 | 0.17 |
| $\beta_{15}$ | 0.02 | 0.51 | 0.26 | 0.01 | 0.47 | 0.22 | 0.03 | 0.48 | 0.23 | 0.03 | 0.41 | 0.17 |
| $\beta_{20}$ | 0.05 | 0.52 | 0.27 | 0.01 | 0.46 | 0.21 | 0.05 | 0.46 | 0.21 | 0.06 | 0.41 | 0.17 |

MSE: mean squared error; MM: minorization-maximization; SE: standard error.

values of the run times (T) in seconds, the absolute value of biases (|Bias|), the empirical standard errors (SE) and the mean squared errors (MSE) at different sample sizes, different censoring levels and different values of Var($\omega_i$) for the three comparative methods. Some representative results are reported on Tables 1 and 2 as follows and more additional simulation results are reported on Tables S9 to DIFaddS17 in Supplemental Material. From these tables, we summarize the main findings as follows.

(i) In different frailty models, the empirical standard errors (SE) of all representative parameters become smaller when the censoring proportion ($P_{cen}$) decreases from 0.3 to 0.1 for all three estimation methods (i.e. MM method, coxph function and frailtyHL function).
(ii) For different frailty models, the empirical standard errors (SE) and the biases of almost all representative parameters become smaller for all three estimation methods when the sample size increases from 100 to 200 to 300. And this variation trend is more obvious in parameter $\theta$.

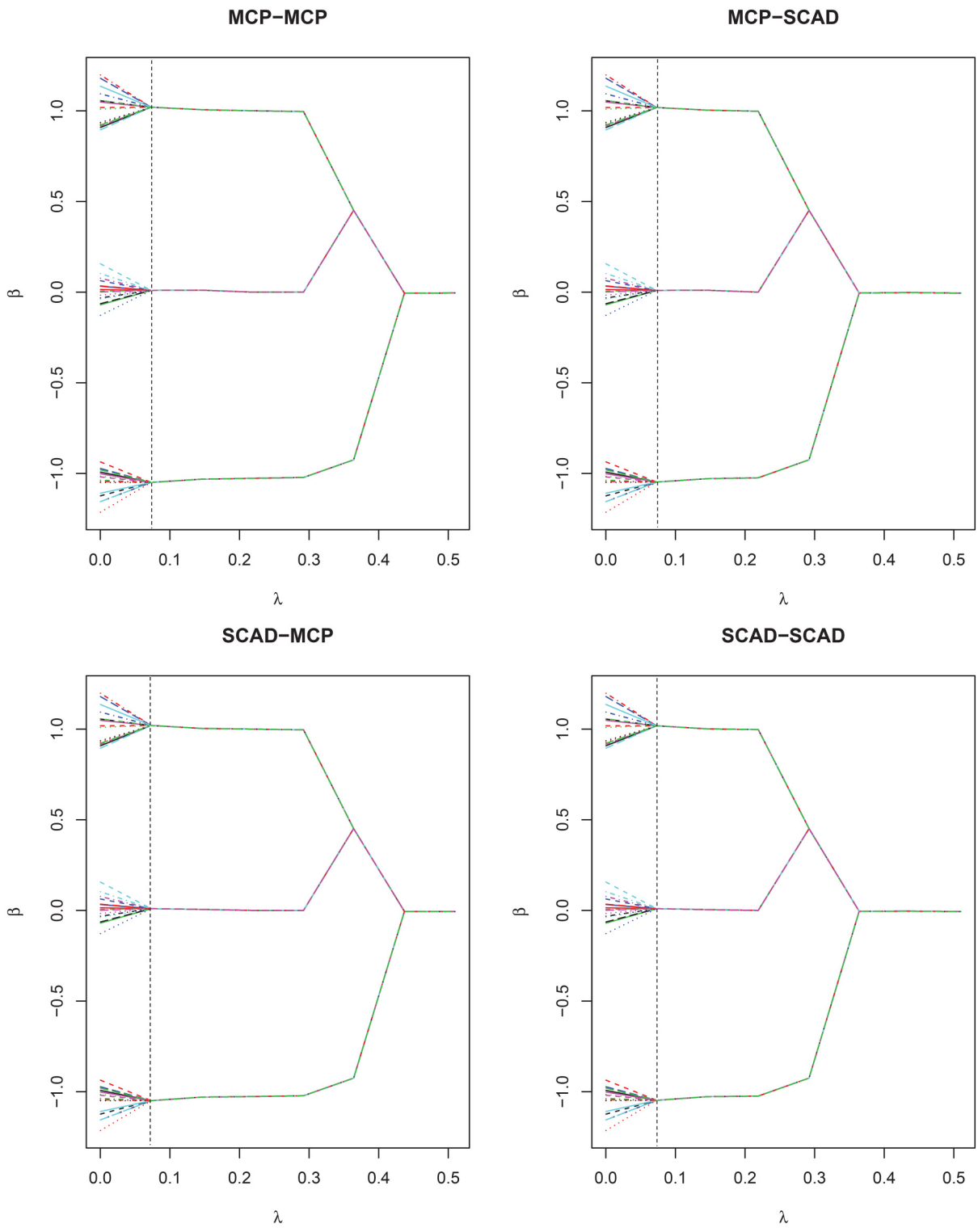


**Figure 1.** The solution path for estimated values of $\beta$ against $\lambda$ with different combinations of penalties for one simulated dataset.

(iii) In terms of estimation accuracy, the MM method always performs the best, exhibiting smaller biases and empirical standard errors in most situations. In particular, the estimation of $\theta$ by MM method still performs very well with strong stability in the case of a small sample size for all three frailty models. And the method of frailtyHL function shows evident biases and the largest empirical standard errors in parameter $\theta$ for Gamma frailty model, the method of

coxph function shows obvious biases and the largest empirical standard errors of $\theta$ for Log-normal frailty model under small sample sizes. However, when the sample size is large, the estimations of $\theta$ and $\boldsymbol{\beta}$ obtained by the three methods perform similarly well.

(iv) Under different combinations of $\text{Var}(\omega_i)$, $P_{cen}$ and sample sizes, the method by coxph function always converges the fastest among the three methods in all situations. However, when the sample size is small, the method of frailtyHL function converges faster than our MM method. When the sample size is large, the method of frailtyHL function does have the slowest convergence speed. Note that the survival R package and frailtyHL R package are optimised using the C language, in which case a speed comparison might not be fair. Even in this case, the MM method is still effective, converging faster than the method of frailtyHL function when the sample size is greater than 100, and achieving more accurate estimates than the two methods of coxph function and frailtyHL function.

(v) For Gamma frailty model and Log-normal frailty models, the SEs for the $\boldsymbol{\beta}$ coefficients increase as the $\text{Var}(\omega_i)$ increases for all three comparative methods. But for the Inverse Gaussian frailty model, the SEs for the $\boldsymbol{\beta}$ coefficients decrease as the $\text{Var}(\omega_i)$ increases.

(vi) Among the three comparative distributions, the estimates of $\text{Var}(\omega_i)$ with Gamma distribution by the MM method perform the best, almost always exhibit the smallest biases, SEs and MSEs under different senarios of sample sizes, censoring proportions and true variances.

Example 2. We generate data from the Gamma frailty model

$$\lambda_j(t|\mathbf{X}_{ij}, \omega_i) = \omega_i \lambda_{0j}(t) \exp\{\mathbf{X}_{ij}^\top \boldsymbol{\beta}\}, \quad \omega_i \sim \text{Gamma}(1/\theta, 1/\theta), \tag{32}$$

for $i = 1, \ldots, n,\ j = 1, 2$, We set $n$ =150 or 300, $q = 45$, $\theta = 0.5$, $\lambda_{01}(t) = 3,\ \lambda_{02}(t) = 5/(1+5t)$, and the covariates $\mathbf{X}_{ij}^\top$ are generated from a multivariate normal distribution with mean zero and a first-order autoregressive structure $0.2^{|r-s|}$ for $r, s = 1, \ldots, q$. We consider variable selection and homogeneity pursuit with regularization and fusion. Let $\beta = (1_{15}^\top, 0_{15}^\top, -1_{15}^\top)^\top$.

Figure 1 shows the solution paths for the estimated values of $\boldsymbol{\beta}$ against $\lambda$ with different combinations of penalties for one simulated dataset. We see that the estimated values of $\boldsymbol{\beta}$ first converge to three different values around 1, 0 and $-1$, which are the true values of regression coefficients for three groups, when $\lambda$ reaches around 0.08 for both MCP and SCAD. Then they finally converge to one value when $\lambda$ exceeds around 0.43 and 0.37 for MCP and SCAD, respectively. Here we use the modified BIC type criteria to select the tuning parameter $\lambda$ by minimizing the BIC function given in (31). Based on 200 replications, the results are summarized in Table 3. For variable selection, we report the empirical percentage of identifying the true model, i.e. all the covariates with non-zero coefficients are selected and all the covariates with zero coefficients are excluded. The average numbers of correctly and incorrectly identified zero coefficients are also reported, respectively. For homogeneity pursuit, we report the empirical percentage of estimating the true number of non-zero groups, i.e. $\hat{S} = 2$. We can see that it yields consistent variable selection results. For a moderately large sample size $n = 300$, the number of non-zero groups can be estimated quite well. Since at $n = 300$, the first 15 $\boldsymbol{\beta}$ coefficients in Example 2 with true value 1 are correctly grouped together and the last 15 $\boldsymbol{\beta}$ coefficients in Example 2 with true value $-1$ are also correctly

**Table 3.** Simulation results for variable selection and homogeneity pursuit in Example 2.

| Sparsity + Fusion Penalties | Selecting the true model | Zeros | | Estimating $S$ | | |
|---|---|---|---|---|---|---|
| | | Correct | Incorrect | $P(\hat{S}=2)$ | $P(\hat{S}=3)$ | $P(\hat{S}=4)$ |
| | | | *Sample size n = 150* | | | |
| MCP + MCP | 0.985 | 14.985 | 0 | 0.94 | 0.055 | 0.005 |
| MCP + SCAD | 0.99 | 14.99 | 0 | 0.97 | 0.03 | 0 |
| SCAD + MCP | 0.995 | 14.995 | 0 | 0.935 | 0.06 | 0.005 |
| SCAD + SCAD | 1 | 15 | 0 | 1 | 0 | 0 |
| | | | *Sample size n = 300* | | | |
| MCP + MCP | 1 | 15 | 0 | 1 | 0 | 0 |
| MCP + SCAD | 1 | 15 | 0 | 1 | 0 | 0 |
| SCAD + MCP | 1 | 15 | 0 | 1 | 0 | 0 |
| SCAD + SCAD | 1 | 15 | 0 | 1 | 0 | 0 |

Note: Selecting the true model denotes the empirical percentage of identifying the true model, i.e. all the covariates with non-zero coefficients are selected and all the covariates with zero coefficients are excluded; $P(\hat{S}=2)$, $P(\hat{S}=3)$ and $P(\hat{S}=4)$ denote the empirical percentages of $\hat{S}=2$, $\hat{S}=3$, and $\hat{S}=4$, respectively. MCP: minimax concave penalty; SCAD: smoothly clipped absolute deviations.

**Table 4.** The average values of estimated typical parameters (MLE), their biases (Bias) and their empirical standard deviations (SD) with sample size 300 based on 200 realizations in Example 2.

| | MCP+MCP | | | MCP+SCAD | | |
|---|---|---|---|---|---|---|
| Parameters | MLE | Bias | SD | MLE | bias | SD |
| $\theta$ | 0.4693 | 0.0307 | 0.1381 | 0.4006 | 0.0994 | 0.1200 |
| $\eta_1$ | 0.9816 | 0.0184 | 0.0727 | 0.9205 | 0.0795 | 0.0646 |
| $\eta_2$ | −0.9821 | −0.0179 | 0.0729 | −0.9221 | −0.0779 | 0.0643 |
| | SCAD+MCP | | | SCAD+SCAD | | |
| Parameters | MLE | Bias | SD | MLE | bias | SD |
| $\theta$ | 0.4456 | 0.0544 | 0.1044 | 0.4119 | 0.0881 | 0.1107 |
| $\eta_1$ | 0.9943 | 0.0057 | 0.0674 | 0.9277 | 0.0723 | 0.0568 |
| $\eta_2$ | −0.9942 | −0.0058 | 0.0697 | −0.9294 | −0.0706 | 0.0542 |

Note: $\eta_1$ denotes the average value of the estimates of the first 15 $\boldsymbol{\beta}$ coefficients and $\eta_2$ denotes the average value of the estimates of the last 15 $\boldsymbol{\beta}$ coefficients.MCP: minimax concave penalty; SCAD: smoothly clipped absolute deviations.

**Table 5.** The minimal BIC score (BIC), number of non-zero parameters (Non-zeros) and number of non-zero groups (No. of groups) from three different frailty models for the ADNI data.

| Sparsity + Fusion Penalties | BIC | Non-zeros | No. of groups |
|---|---|---|---|
| | *Log-normal frailty* | | |
| MCP+SCAD | 1070.8 | 5 | 2 |
| MCP+MCP | 1070.8 | 5 | 2 |
| SCAD+SCAD | 1070.8 | 5 | 2 |
| SCAD+MCP | 1070.8 | 5 | 2 |
| | *Inverse Gaussian frailty* | | |
| MCP+SCAD | 1083.2 | 10 | 2 |
| MCP+MCP | 1085.2 | 10 | 2 |
| SCAD+SCAD | 1082.9 | 10 | 2 |
| SCAD+MCP | 1083.6 | 10 | 2 |
| | *Gamma frailty* | | |
| MCP+SCAD | 1073.7 | 10 | 2 |
| MCP+MCP | 1073.5 | 10 | 2 |
| SCAD+SCAD | 1073.7 | 10 | 2 |
| SCAD+MCP | 1073.2 | 10 | 2 |

BIC: Bayesian information criterion; ADNI: Alzheimer's Disease Neuroimaging Initiative; MCP: minimax concave penalty; SCAD: smoothly clipped absolute deviation.

grouped together for different combinations of regularization and fusion penalties. We further enforce the final parameter estimations in the same group to be exactly the same and to be equal to the average value of all parameter values in this group. Therefore, we denote $\eta_1$ as the average value of the estimates of the first 15 $\boldsymbol{\beta}$ coefficients and denote $\eta_2$ as the average value of the estimates of the last 15 $\boldsymbol{\beta}$ coefficients and further report the biases and mean squared error for the model parameters $\theta, \eta_1, \eta_2$ at sample size $n = 300$ for different combinations of regularization and fusion penalties in Table 4. It can be seen that the estimation is quite accurate. In general, the MCP+MCP and SCAD+MCP yield the most accurate results.

## An application to the ADNI data

We now apply the proposed algorithms to analyze the ADNI dataset which is publicly available from the ADNI consortium (adni.loni.usc.edu). In the study, there are 267 people who are identified to be cognitively normal (CN) during their first visit. Before the last visit, 78 of them progressed to the stage of MCI and 22 people also progressed from the stage of MCI to the stage of AD. We consider the transition time from CN to MCI and then from MCI to AD. The frailty model with

**Table 6.** Estimation results for Alzheimer's Disease Neuroimaging Initiative (ADNI) data using the Log-normal frailty model.

| Group | Variable | Description | Estimation |
|---|---|---|---|
| A | CDRSB | Clinical Dementia Rating Scale Sum of Boxes score | 0.24 |
| A | ADAS13 | Alzheimer's Disease Assessment Scale-Cognitive Subscale | 0.24 |
| A | ADASQ4 | A specific subscale of ADAS (Delayed Word Recall) | 0.24 |
| A | FAQ | Functional Activities Questionnaire | 0.24 |
| B | LDELTOTAL | Delayed recall total | −0.23 |

**Table 7.** Estimation results for Alzheimer's Disease Neuroimaging Initiative (ADNI) data using the Inverse Gaussian frailty model.

| Group | Variable | Description | Estimation |
|---|---|---|---|
| A | ApoE-$\varepsilon$4_1 | An SNP which is related to Alzheimer | 0.28 |
| A | CDRSB | Clinical Dementia Rating Scale Sum of Boxes score | 0.28 |
| A | ADAS13 | Alzheimer's Disease Assessment Scale-Cognitive Subscale | 0.28 |
| A | ADASQ4 | A specific subscale of ADAS (Delayed Word Recall) | 0.28 |
| A | FAQ | Functional Activities Questionnaire | 0.28 |
| A | TRABSCOR | Trail Making Test A-B | 0.28 |
| A | RAVLT_immediate | Rey Auditory Verbal Learning Test | 0.28 |
| A | rs7856537_1 | Unknown | 0.28 |
| B | LDELTOTAL | Delayed recall total | −0.05 |
| B | mPACCdigit | ADNI modified Preclinical Alzheimer's Cognitive Composite with Digit Symbol Substitution | −0.05 |

**Table 8.** Estimation results for Alzheimer's Disease Neuroimaging Initiative (ADNI) data using the Gamma frailty model.

| Group | Variable | Description | Estimation |
|---|---|---|---|
| A | ApoE-$\varepsilon$4_2 | A SNP which is related to Alzheimer | 0.20 |
| A | CDRSB | Clinical Dementia Rating Scale Sum of Boxes score | 0.20 |
| A | ADAS13 | Alzheimer's Disease Assessment Scale-Cognitive Subscale | 0.20 |
| A | ADASQ4 | A specific subscale of ADAS (Delayed Word Recall) | 0.20 |
| A | FAQ | Functional Activities Questionnaire | 0.20 |
| A | rs3095750_2 | Unknown | 0.20 |
| A | rs2986017_2 | Related to Alzheimer | 0.20 |
| B | LDELTOTAL | Delayed recall total | −0.09 |
| B | mPACCdigit | ADNI modified Preclinical Alzheimer's Cognitive Composite with Digit Symbol Substitution | −0.09 |
| B | mPACCtrailsB | ADNI modified Preclinical Alzheimer's Cognitive Composite (PACC) with Trails B | −0.09 |

these two outcomes is fitted to identify relevant covariates that are related to and predictive of the outcomes. The covariates include 19 clinical variables such as age, marriage status, education level, and test scores like Alzheimer's Disease Assessment Scale cognitive subscale (ADAS-Cog) and Functional Activities Questionnaire (FAQ) and 113 genetic variables (SNPs). Most genetic variables are selected based on existing knowledge which may have an impact on Alzheimer, some other genetic variables are chosen near those known SNPs to test whether they are also related to the development of AD. The Log-normal, Inverse Gaussian, and Gamma frailty models are fitted with the proposed MM+ADMM algorithms for both variable selection and homogeneity pursuit. The results are reported in Table 5. It can be seen that different combinations of sparsity and fusion penalties lead to similar results. We find that the log-normal frailty model yields the minimum BIC score and five variables are included in the final model and the number of non-zero groups is estimated to be 2. The specific estimation results for three frailty models are reported in Tables 6 to 8.

The number of estimated groups for all models along with different penalties is 2 where group A has a positive impact on the hazard rate while group B leads to a negative impact to it. Clinical records "CDRSB", "ADAS13", "ADASQ4", "FAQ" and "LDELTOTAL" are selected by all models with similar numerical impact to the hazard rate. This result reflects the necessity of conducting clinical assessment for the diagnostic procedure since those assessment scores are very correlated to the development of Alzheimer. In addition to the clinical records, gene also plays an important role in the onsite Alzheimer. The covariate ApoE-$\varepsilon$4 which is already proven to influence the development of Alzheimer is also selected by

Inverse Gaussian frailty and Gamma frailty model. However, many other known SNPs which are considered to be correlated with Alzheimer in previous literature do not show a significant impact on the state transition of this disease while some unknown SNPs such as "rs7856537" and "rs3095750" are also chosen by some models. Therefore, there still exists a large room for future scientific studies on the detection of Alzheimer using SNPs and the earlier prevention for a specific group of people with genetic risk.

## Discussion

For survival data with multiple outcomes, we develop efficient estimation algorithms in the frailty model where general frailty distributions are allowed and the observed likelihood can involve intractable integral. Instead of resorting to the computationally intensive Laplace approximation or Monte Carlo simulation methods, our approach is based on the non-profile MM method which treats the parametric and nonparametric components in the same way during the estimation process so that the standard errors of the estimates can not be easily computed but the bootstrap method may be a good alternative option in such high-dimensional estimation problems. By constructing a sequence of minorization functions, the resulting algorithm eventually decomposes the minorization function into a sum of univariate functions with separate parameters. This avoids matrix inversion and leads to substantial computational advantages. The algorithms also mesh well with regularization and fusion using concave penalties. The ADMM algorithm is also incorporated for both variable selection and homogeneity pursuit in high-dimension regression analysis of survival data with multiple outcomes. Although the non-profile MM method is developed for the classical multiplicative frailty model with general frailty distributions, a similar approach can essentially be developed for the frailty models with more complex random effects structures which we aim to investigate further in our future work.

### Acknowledgments

The authors are very grateful to the Editor, Professor Marta Garcia-Finana, and two anonymous referees for many helpful comments that greatly improved the paper. Xifen Huang's research was supported by the National Natural Science Foundation of China (11901515,12261108). Jinfeng Xu's research was supported by General Research Fund (17308820) of Hong Kong, Start-up grant for new faculty at City University of Hong Kong (7200742), and the National Natural Science Foundation of China (72033002).

### Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### ORCID iDs

Xifen Huang https://orcid.org/0000-0003-1593-5864
Jinfeng Xu https://orcid.org/0000-0002-3165-2015

### Supplemental material

Supplemental material for this article is available online.

### References

1. Ferri CP, Prince M, Brayne C, et al. Global prevalence of dementia: a Delphi consensus study. *The Lancet* 2005; **366**: 2112–2117.
2. Petersen RC. Mild cognitive impairment as a diagnostic entity. *J Intern Med* 2004; **256**: 183–194.
3. Ringman JM, Goate A, Masters CL, et al. Genetic heterogeneity in Alzheimer disease and implications for treatment strategies. *Curr Neurol Neurosci Rep* 2014; **14**: 499.
4. Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 1978; **65**: 141–151.
5. Clayton DG and Cuzick J. Multivariate generalizations of the proportional hazards model. *J R Stat Soc: Ser A (General)* 1985; **148**: 82–117.
6. Oakes D. Bivariate survival models induced by frailties. *J Am Stat Assoc* 1989; **84**: 487–493.
7. Zeng D, Chen Q and Ibrahim JG. Gamma frailty transformation models for multivariate survival times. *Biometrika* 2009; **96**: 277–291.
8. Do Ha I, Jeong J-H and Lee Y. *Statistical modelling of survival data with random effects*. Singapore: Springer, 2017.
9. Do Ha I, Noh M and Lee Y. frailtyhl: A package for fitting frailty models with h-likelihood. *R J* 2012; **4**: 28–37.

10. Ha Il D, Pan J, Oh S, et al. Variable selection in general frailty models using penalized h-likelihood. *J Comput Graph Stat* 2014; **23**: 1044–1060.
11. Rondeau V, Marzroui Y and Gonzalez JR. frailtypack: an r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *J Stat Softw* 2012; **47**: 1–28.
12. Aalen OO. Modelling heterogeneity in survival analysis by the compound poisson distribution. *Ann Appl Probab* 1992; **2**: 951–972.
13. Duchateau L and Janssen P. *The frailty model*. New York: Springer Verlag, 2008.
14. Hougaard P. Survival models for heterogeneous populations derived from stable distributions. *Biometrika* 1986; **73**: 387–396.
15. Andersen PK, Klein JP, Knudsen KM, et al. Estimation of variance in Cox's regression model with shared gamma frailties. *Biometrics* 1997; **53**: 1475–1484.
16. Klein JP. Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics* 1992; **48**: 795–806.
17. Nielsen GG, Gill RD, Andersen PK, et al. A counting process approach to maximum likelihood estimation in frailty models. *Scand J Stat* 1992; **19**: 25–43.
18. Ke T, Fan JQ and Wu YC. Homogeneity in regression. *J Am Stat Assiciation* 2015; **110**: 175–194.
19. Yang S, Yuan L, Lai YC, et al. Feature grouping and selection over an undirected graph. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 922–930, 2012.
20. Zhu Y, Shen X and Pan W. Simultaneous grouping pursuit and feature selection over an undirected graph. *J Am Stat Assoc* 2013; **108**: 713–725.
21. Becker MP, Yang I and Lange K. EM algorithms without missing data. *Stat Methods Med Res* 1997; **6**: 38–54.
22. Hunter DR and Lange K. A tutorial on MM algorithms. *Am Stat* 2004; **58**: 30–37.
23. Lange K, Hunter DR and Yang I. Optimization transfer using surrogate objective functions. *J Comput Graph Stat* 2000; **9**: 1–20.
24. Fan JQ and Li RZ. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001; **96**: 1348–1360.
25. Zhang CH, et al. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat* 2010; **38**: 894–942.
26. Tibshirani R. The lasso method for variable selection in the cox model. *Stat Med* 1997; **16**: 385–395.
27. Utazirubanda JC, M León T and Ngom P. Variable selection with group lasso approach: Application to cox regression with frailty model. *Commun Stat-Simul Comput* 2021; **50**: 881–901.
28. Groll A, Hastie T and Tutz G. Selection of effects in cox frailty models by regularization methods. *Biometrics* 2017; **73**: 846–856.
29. Boyd S, Parikh N and Chu E. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found Trends Mach Learn* 2011; **3**: 1–122.
30. Chi EC and Lange K. Splitting methods for convex clustering. *J Comput Graph Stat* 2015; **24**: 994–1013.
31. Huang XF, Xu JF and Tian GL. On profile MM algorithms for gamma frailty survival models. *Stat Sin* 2019; **29**: 895–916.
32. Johansen S. An extension of cox's regression model. *International Statistical Review/Revue Internationale de Statistique* 1983; **51**: 165–174.
33. Vaida F. Parameter convergence for em and mm algorithms. *Stat Sin* 2005; **15**: 831–840.
34. Lange K and Zhou H. MM algorithms for geometric and signomial programming. *Math Program* 2014; **143**: 339–356.
35. Hunter DR and Li R. Variable selection using MM algorithms. *Ann Stat* 2005; **33**: 1617–1642.
36. Schwarz G, et al. Estimating the dimension of a model. *Ann Stat* 1978; **6**: 461–464.
37. Graven P. Smoothing noisy data with spline function: estimating the correct degree of smoothing by the method of generalized cross-validaton. *Number Math* 1989; **31**: 377–403.
38. Ma SJ and Huang J. A concave pairwise fusion approach to subgroup analysis. *J Am Stat Assoc* 2017; **112**: 410–423.
39. Friedman J, Hastie T and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010; **33**: 1.
40. Fan J and Li R. Variable selection for cox's proportional hazards model and frailty model. *Ann Stat* 2002; **30**: 74–99.
41. Varadhan R and Roland C. Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scand J Stat* 2008; **35**: 335–353.
42. Zhou H, Alexander D and Lange K. A quasi-newton acceleration for high-dimensional optimization algorithms. *Stat Comput* 2011; **21**: 261–273.