VARIABLE SELECTION FOR LONGITUDINAL AND SURVIVAL DATA

A Thesis

by

XUAN DANG

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Xiaoning Qian
Committee Members,	Erchin Serpedin
	Krishna Narayanan
	Xia "Ben" Hu
Head of Department,	Miroslav M. Begovic

December 2021

Major Subject: Electrical and Computer Engineering

ABSTRACT

With the advancing of data collection technologies, high-dimensional and large-scale data sets become available in many areas of science, specifically in biomedicine. One of important questions when mining such “big” data is to identify critical factors that may be predictive of the outcomes of interest, for example for disease diagnosis and prognosis. In this thesis, we introduce several models with solution algorithms that exploit sparse dependency structures to discover the variables playing important roles in survival and longitudinal data.

First, we focus on penalized Cox’s models to deal with the high-dimensional survival data with group predictors. Most of the existing penalized methods for Cox’s model are the group lasso methods that show deficiencies, including the over-shrinkage problem. In addition, the contemporary algorithms either exhibit the loss of efficiency or require the group-wise orthonormality assumption. In Chapter 3, we investigate and comprehensively evaluate three group penalized methods for Cox’s models: the group lasso and two nonconvex penalization methods—group SCAD and group MCP—that have several advantages over the group lasso. We develop the fast and stable algorithms and a new R package **grpCox** to fit these models without the initial orthonormalization step. These methods perform group selection in both non-overlapping and overlapping cases.

Second, we study the multi-state models to analyze longitudinal data, in which the change of status over time is of interest. Due to the lack of an efficient and practical variable selection tool to practitioners, we develop the L1-regularized multi-state model framework for simultaneous parameter estimation and variable selection in Chapter 4. We use a local quadratic approximation of the log-partial likelihood and devise the one-step coordinate descent algorithm to solve the corresponding optimization problem, which can offer significant improvement on the computational efficiency. The proposed method is implemented in our R package **L1mstate**.

Finally, we investigate multivariate joint models to study the relationship between multiple time-varying measurements and the survival outcome, considering the potential correlation between these time-varying measurements. We address the problems of identifying the time-varying measurements that have strong associations with the time-to-event outcome, and simultaneously selecting predictive baseline covariates for both the longitudinal measurements and survival outcome of interest, which has no available tools so far to the best of our knowledge. In Chapter 5, we develop a variable selection framework for the multivariate joint models. Specifically, we propose novel penalized joint models for different association structures between the longitudinal and the survival submodels using different types of sparsity-inducing penalties. To tackle high-dimensional challenge that arises in the case of multiple longitudinal measurements, many covariates, and random effects, we develop an estimation procedure based on Laplace approximation of the joint likelihood.

DEDICATION

To my family.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my family for supporting and encouraging me all the way through my PhD.

I sincerely thank my advisor, Dr. Xiaoning Qian, for his support and guidance, and for providing me the freedom to work on research problems. I also thank my other committee members Dr. Erchin Serpedin, Dr. Krishna Narayanan, Dr. Xia "Ben" Hu and my collaborators, Dr. Byul Hur and Dr. Shuai Huang, for their help and feedback.

I am grateful to my lab mates and friends for being greatest people to walk through grad school with.

Finally, I would like to thank the Texas A&M University High Performance Research Computing for providing computational resources to perform experiments in this dissertation.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Dr. Xiaoning Qian and Dr. Erchin Serpedin and Dr. Krishna Narayanan of the Department of Electrical and Computer Engineering and Dr. Xia "Ben" Hu of Department of Computer Science and Engineering. All work for the dissertation was completed by the student, under the advisement of Dr. Xiaoning Qian of the Department of Electrical and Computer Engineering.

Funding Sources

My work was supported by the National Science Foundation(NSF)–Division of Communication and Computing Foundations (CCF) awards #1718513, #1715027, #1714136 and the JDRF award #2-SRA- 2018-513-S-B.

NOMENCLATURE

SCAD	Smoothly Clipped Absolute Deviation
MCP	Minimax Concave Penalty
MM	Majorization-Minimization
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative
TPR	True Positive Rate
FPR	False Positive Rate
RMSE	Root Mean Square Error
TCGA	The Cancer Genome Atlas
FDR	False Discovery Rate
FWER	Family-Wise Error Rate
MSM	Multi-State Model
MSTATE	Un-regularized Multi-State Model
AUC	Area Under the ROC Curve
EBMT	Europe Blood and Marrow Transplantation
MLE	Maximum Likelihood Estimation
BLUPs	Best Linear Unbiased Predictors
ADMM	Alternating Direction Method of Multipliers
PBC	Primary Biliary Cirrhosis

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iv

ACKNOWLEDGMENTS v

CONTRIBUTORS AND FUNDING SOURCES vi

NOMENCLATURE vii

TABLE OF CONTENTS viii

LIST OF FIGURES xii

LIST OF TABLES xiv

1. INTRODUCTION 1

2. BACKGROUNDS 4

- 2.1 Survival analysis 4
  - 2.1.1 The exponential distribution 5
  - 2.1.2 The Gompertz distribution 5
- 2.2 Sparsity-inducing penalties 6
  - 2.2.1 l_1-norm 6
  - 2.2.2 SCAD and MCP 6
  - 2.2.3 Group lasso 7
  - 2.2.4 Overlapping group lasso 7
- 2.3 Approximation approaches 7
  - 2.3.1 Quadratic approximation 7
  - 2.3.2 Laplace approximation 8
- 2.4 Optimization methods 8
  - 2.4.1 Majorization-minimization (MM) 8
  - 2.4.2 Coordinate descent 9
  - 2.4.3 Alternating direction method of multipliers (ADMM) 10
    - 2.4.3.1 Augmented Lagrangian 10
    - 2.4.3.2 Optimality conditions 10
    - 2.4.3.3 ADMM procedure 10

3. PENALIZED COX'S MODEL WITH GROUPED PREDICTORS 12

3.1 Introduction 13
3.2 Non-overlapping groups 16
3.2.1 Model formulation 16
3.2.2 Majorization-minimization (MM) approach 18
3.2.3 Group-wise descent algorithm 20
3.2.3.1 Group lasso 21
3.2.3.2 Group SCAD 22
3.2.3.3 Group MCP 23
3.2.4 The descent property of group-wise descent algorithm 23
3.2.5 Active set updates 24
3.2.6 Pathwise solution 25
3.2.7 Selection of the tuning hyperparameters 26
3.3 Overlapping groups 26
3.4 Simulation studies 29
3.4.1 Setup 30
3.4.2 Time and quality comparison with other packages 31
3.4.2.1 Non-overlapping groups 31
3.4.2.2 Overlapping groups 33
3.4.2.3 $N \geq P$ problems 34
3.4.3 Comparison of three group penalized Cox's models 35
3.4.3.1 Non-overlapping groups 35
3.4.3.2 Overlapping groups 36
3.4.4 Comparison of three group penalized Cox's models with non-overlapping groups 37
3.4.4.1 Effect of the coefficient magnitude 37
3.4.4.2 Effect of the group size 38
3.4.4.3 Effect of censoring 39
3.4.4.4 Effect of covariate correlation 40
3.4.5 Comparison of three group penalized Cox's models with overlapping groups 41
3.4.5.1 Equal group size 41
3.4.5.2 Effect of the number of overlapping covariates among groups 42
3.4.5.3 Unequal group size 42
3.4.5.4 Sparse group example 43
3.4.5.5 Results 45
3.4.6 Misspecification of group structures 45
3.5 Real-world case studies 47
3.5.1 Data 47
3.5.2 Methods 48
3.5.3 Results and discussion 48

4. L1-REGULARIZED MULTI-STATE MODELS 55

4.1 Introduction 55
4.2 Review of Multi-State Models (MSMs) 58
4.3 Formulations of the multi-state models 58

4.3.1 Likelihood function of the multi-state models 60
4.3.2 Partial likelihood function for multi-state model 60
4.3.3 Data structure for parameter estimation by partial likelihood maximization .. 61
4.4 L1-Regularized Multi-State Model (L1MSTATE)....................................... 63
4.4.1 Partial likelihood formulation for L1MSTATE 63
4.4.2 Computational algorithm for solving Eq. (5.5) 64
4.4.3 Active set updates 68
4.4.4 Pathwise solution 68
4.4.5 Selection of the tuning hyperparameters 69
4.4.6 Estimation of the cumulative hazard rates and the transition probabilities 70
4.4.7 Computational complexity analysis 71
4.5 Simulation studies.. 72
4.5.1 Setup 72
4.5.2 Results 74
4.5.3 Large-scale datasets 81
4.5.4 Empirical runtime comparison.. 81
4.6 Europe Blood and Marrow Transplantation (EBMT) data 81
4.6.1 Comparison of the models 85
4.6.2 Further assessment of the effects of risk factors upon the disease progression 89

5. PENALIZED JOINT MODELS OF TIME-TO-EVENT AND MULTIVARIATE LONGITUDINAL OUTCOMES 92

5.1 Introduction.. 92
5.2 Models and Notations.. 95
5.2.1 Longitudinal submodel... 96
5.2.2 Survival submodel.. 96
5.2.3 The joint likelihood.. 97
5.3 Parameter estimation.. 98
5.3.1 Laplace approximation of the joint likelihood 98
5.3.1.1 The first-order derivative of $\mathcal{K}(b_i)$ respect to b_i 99
5.3.1.2 The second-order derivative of $\mathcal{K}(b_i)$ respect to b_i..................100
5.3.2 Variable selection based on penalized likelihood 100
5.3.3 Algorithm for optimization of the penalized likelihood 102
5.3.3.1 Initial values 103
5.3.3.2 Estimates of parameters without penalties: β_0 and $(\sigma_1^2, \ldots, \sigma_J^2)$....103
5.3.3.3 Estimates of parameters with penalties: β, γ_0 and γ 104
5.3.3.4 Estimate of a symmetric and positive definite matrix D............106
5.3.3.5 Convergence conditions..108
5.3.4 Hyperparameter tuning...108
5.4 Simulation studies...109
5.4.1 Model I...110
5.4.2 Model II..112
5.5 Real-world case studies 114

6. CONCLUSIONS & FUTURE RESEARCH 119

REFERENCES 122

APPENDIX A. CHAPTER 3 136

A.1 Appendix 1 136
A.2 Appendix 2 138
A.3 Appendix 3 140

LIST OF FIGURES

FIGURE Page

3.1 The coefficient decomposition of overlapping groups. 28

3.2 Plots of average runtime over 100 trials for 50 λ-value paths. The runtime is in seconds. 34

3.3 Solution paths for the group lasso, group SCAD, and group MCP models. The solid lines are for signal variables while the dashed lines are for noise variables. 36

3.4 The impact of the coefficient magnitude and censoring rate on group regularization methods when the group size is 4. The *black line* is the true model size (*5*). 38

3.5 The impact of increasing coefficient magnitude on group regularization methods when the group size is 10 (first row) and 20 (second row). The *black line* on the right is the true model size. 40

3.6 The impact of increasing the coefficient magnitude and the population correlation on group regularization methods when the size is 4. The *black line* is the true model size (*5*). 42

3.7 Survival curves for the high and low risk groups of the independent testing samples of ovarian cancer (first row) and breast cancer (second row). 54

4.1 Some multi-state models. *Note*: Arrows show the clinically eligible transitions for each multi-state model. 59

4.2 The illness-death model. 73

4.3 Plots of the coefficients paths for three transitions of our L1MSTATE model fit and the cross-validation log-partial likelihood against the log of λ values along our path. 75

4.4 AUC values of Example I for different sample sizes in different settings over 100 replications. 79

4.5 The EBMT model. 84

4.6 Estimates of stacked prediction transition probabilities from $t = 0$ for patients A, B, and C using two models. L1MSTATE, L1-regularized multi-state model using the first cross-validation method; pL1MSTATE, L1-regularized multi-state model using the penalized cross-validation method; MSTATE, multi-state model. 87

4.7 Estimated cumulative hazard rates for patient A and patient D. 90

4.8 Estimates of stacked prediction transition probabilities from $t = 0$ for patients A and D. 90

4.9 Estimates of stacked prediction transition probabilities of patient D from $t = 0$ and $t = 100$ days since transplant. 91

5.1 Histogram and Q-Q plot of serum bilirubin........................ 114

5.2 Longitudinal trajectory plots........................ 115

A.1 Plot of the cross-validation log-partial likelihood against the log of λ values along the regularization path........................ 139

LIST OF TABLES

TABLE Page

3.1 Comparison of **grpCox** with publicly available packages in the non-overlapping settings. The mean time, average TPRs, and average FPRs, over 100 independent data sets and a 50 λ values path, are reported. The time is in seconds. 31

3.2 Comparison of **grpCox** with publicly available packages in the overlapping and $N > P$ settings. The mean time, over 100 independent data sets and a 50 λ values path, is reported in seconds. 32

3.3 Average true positive rate (TPR) and false positive rate (FPR) values of three group regularization methods over 100 replications for different coefficient magnitude values and different censoring scenarios when the group size is 4. 39

3.4 Average true positive rate (TPR), and false positive rate (FPR) values of three group regularization methods over 100 replications for different coefficient magnitude values and different censoring scenarios. 41

3.5 Average true positive rate (TPR), and false positive rate (FPR) values over 100 replications for three group regularization models with different coefficient magnitude and population correlation values when the group size is 4. 43

3.6 Results for overlapping group settings with different overlapping covariates between two successive groups over 100 replications. 44

3.7 Results for overlapping group settings over 100 replications. 45

3.8 Results for overlapping group settings over 100 replications. 47

3.9 Pathways and genes selected by different methods for ovarian cancer data. 49

3.10 Pathways and genes selected by different methods for breast cancer data. Note that 293 selected pathways using univariate tests are listed in the Supporting Information. 52

4.1 Example of long-format data 62

4.2 Model selection results of Example I for the small effects scenario. 76

4.3 Model selection results of Example I for the medium effects scenario. 77

4.4 Model selection results of Example I for the large effects scenario. 77

4.5 Model selection results of Example I for the mixed effects scenario. 78

4.6 Model selection results of large P settings for the small effects scenario. 82

4.7 Model selection results of large P settings for the medium effects scenario. 82

4.8 Model selection results of large P settings for the large effects scenario.............. 82

4.9 Running time of three L1-regularized models. The mean time over different datasets (100 for small datasets and 10 for big datasets) required to fit the entire solution path over a grid of 100 λ values is reported in seconds................................ 83

4.10 The frequencies and proportions of the number of observed transitions of study population. The numbers in parentheses are proportions.............................. 84

4.11 Regression coefficients of two models for EBMT dataset............................. 86

4.12 Risk factors information of patient A, B, C and D..................................... 88

4.13 The frequencies and proportions of the number of observed transitions from the transplant state of three patients. The numbers in parentheses are proportions. 88

4.14 The frequencies and proportions of the number of observed transitions from the transplant state of patient D after 0 and 100 days. The numbers in parentheses are proportions... 91

5.1 Selection frequency of fixed effects in longitudinal and survival submodels of Model I...111

5.2 Selection frequency of random effects in survival submodels of Model I.............111

5.3 Selection frequency of variance-covariance matrix of Model I. 111

5.4 Selection frequency of fixed effects in longitudinal and survival submodels of Model II..113

5.5 Selection frequency of random effects in survival submodels of Model II. 113

5.6 Selection frequency of variance-covariance matrix of Model II. 113

5.7 Results of the PBC data analysis of Model I..117

5.8 Results of the PBC data analysis of Model II and joineRML. 117

A.1 Results for group lasso using different cross-validation methods to select hyperparameters over 100 replications...140

A.2 Results for group SCAD using different cross-validation methods to select hyperparameters over 100 replications. 140

A.3 Results for group MCP using different cross-validation methods to select hyperparameters over 100 replications. 141

A.4 Results for misspecified group structures over 100 replications. 143

1. INTRODUCTION

Thanks to our advancing data collection and storage capability, demand for sorting through massive data to pare down to their essential information or knowledge is higher than ever before. Sparse statistical models are increasingly popular in modern statistics and machine learning. Generally speaking, sparse statistical models are the ones in which only a relatively small number of covariates (predictors)[1] have strong effects upon the outcomes under study. These models are more interpretable, computationally cheaper, less sensitive to overfitting, allow for parameter estimations and variable selection simultaneously. Often they can be formulated as the optimization problems having the following general form:

$$\min_{\theta \in \mathbb{R}^p} L(\theta) + \lambda P(\theta), \tag{1.1}$$

where $L(\theta)$ is a general loss function that quantifies the goodness-of-fit of the model given the data, and $P(\theta)$ is the penalty term that induces sparsity to constrain the model space and prevent overfitting. Here, $\lambda > 0$ is the regularization parameter that controls the degree of penalization.

In this dissertation, we focus on the problems of identifying the predictors that play important roles in survival and longitudinal data for many biomedical applications. Due to the various challenges arising when addressing corresponding questions in these applications, we consider a variety of models, given different loss functions and sparsity-inducing penalties, within this optimization framework. More specifically, for the loss function, we have negative log-partial-likelihood (Cox's model, multistate model), joint likelihood (multivariate joint model), and for different penalties such as l_1-norm, group lasso, and so on. This dissertation is organized into three main parts that are described more details as follows:

- Penalized Cox's model with grouped predictors [1]: In Chapter 3, we focus on the high-dimensional survival data with group predictors problems. More specifically, we investigate

[1] Throughout the thesis, we use the terms covariate, predictor, variable interchangeably.

and comprehensively evaluate the group lasso, the group SCAD, and the group MCP penalized Cox's models. We develop and evaluate the group-wise descent algorithms combining with the majorization-minimization (MM) approach to solve the optimization problems of the general design matrices without the group-wise orthonormal condition. Our methods perform group selection for both non-overlapping and overlapping group. Since the sparse group lasso that can yield both individual and group sparsity is a special case of overlapping group lasso, our methods can effectively select important groups as well as identify the important covariates within the selected groups. Several computational tricks, including the screening, active set, and warm-start approaches, have been implemented in our R package **grpCox**. Experimental results on both synthetic and real-world data demonstrate the state-of-the-art performance in term of speed and variable selection.

- L1-regularized multi-state models [2]: In Chapter 4, we investigate multi-state model (MSM) to analyze longitudinal data, in which the change of status over time is of interest. More specifically, we develop the L1-regularized multi-state model framework for simultaneous parameter estimation and variable selection. We use a local quadratic approximation of the log-partial likelihood and devise the one-step coordinate descent algorithm to solve the corresponding optimization problem, which can offer significant improvement on the computational efficiency. The proposed method is implemented in the open-access R package **L1mstate**. Our proposed method demonstrates the state-of-the-art performance in terms of identifying the significant risk factors comparing with the existing regularized multi-state models in simulation studies. It also performs better at doing variable selection and predicting the transition probabilities in cases with small sample sizes comparing with the unregularized approach in simulation and real-world cases.

- Penalized joint models of time-to-event and multivariate longitudinal outcomes: In Chapter 5, we investigate the multivariate joint model to study the relationship between multiple longitudinal outcomes and survival outcome, and the relationship between correlated lon-

gitudinal outcomes. We focus on the problems of identifying the longitudinal outcomes that have strong associations with the time-to-event outcome, and simultaneously selecting relevant covariates for both longitudinal and survival outcomes of interest, which has no available tools, to the best of our knowledge, to use. More specifically, we propose novel penalized joint models for different association structures between the longitudinal and the survival submodels using different types of penalties. To tackle high-dimensional challenge that arises in the case of many longitudinal outcomes, covariates, and random effects, we develop an estimation procedure based on Laplace approximation of a joint likelihood. Simulation studies and real-world data application demonstrate the excellent selection property of the proposed methods.

2. BACKGROUNDS

In this chapter, we first introduce the important concepts and fundamental models that will be essential for the development of this thesis, including survival analysis, sparsity-inducing penalties, approximation approaches, and existing optimization methods for sparse models.

2.1 Survival analysis

In this section, we present some basic concepts of survival analysis where the time until an event occurs is of interest. Let T be the random variable associated with the survival time. The distribution of T is defined as

$$F(t) = P(T < t) = \int_0^t f(u)du, \tag{2.1}$$

where $f(t)$ is the probability density function of T. The survival function, $S(t)$, is the probability that the survival time is greater than or equal to t

$$S(t) = \mathrm{P}(T > t) = 1 - F(t), \tag{2.2}$$

The hazard function that represents the the instantaneous rate of an event occurring at time t is given by

$$h(t) = \lim_{dt \to 0^+} \frac{\mathrm{P}(t \leq T < t + dt | T \geq t)}{dt} = -\frac{d}{dt}\mathrm{log}S(t), \tag{2.3}$$

The cumulative hazard function expresses the cumulative risk of an event occurring at time t

$$H(t) = \int_0^t h(u)du, \tag{2.4}$$

The Cox's proportional hazards model [3] is commonly used to study the relationship between survival time and a set of covariates in high-dimensional space as potential predictors for survival time.

$$h_i(t) = h_0(t)\,\exp\big(X^{(i)}(t)\theta\big),$$

where $h_0(t)$ is the baseline hazard function and $X^{(i)}(t)$ is a $p-$dimensional covariates vector of i^{th} subject. In what follows, we introduce two distributions of T used in this thesis.

2.1.1 The exponential distribution

$$f(t) = \lambda\,\exp(-\lambda t), \tag{2.5}$$

where $\lambda > 0$. Then the survival function, $S(t)$, and the hazard function, $h(t)$, are given by

$$S(t) = \exp(-\lambda t), \tag{2.6}$$

$$h(t) = \lambda, \tag{2.7}$$

It means that the hazard is constant over time.

2.1.2 The Gompertz distribution

$$f(t) = \lambda\,\exp\big(\alpha t + \frac{\lambda}{\alpha}(1 - \mathrm{e}^{\alpha t})\big), \tag{2.8}$$

where $\lambda > 0$ is scale parameter, and α is shape parameter. Then the survival function, $S(t)$, and the hazard function, $h(t)$, are given by

$$S(t) = \exp(\frac{\lambda}{\alpha}(1 - \mathrm{e}^{\alpha t})), \tag{2.9}$$

$$h(t) = \lambda \mathrm{e}^{\alpha t}, \tag{2.10}$$

If $\alpha = 0$, it becomes the exponential distribution: the hazard function is constant and equal to λ. For many applications with time-dependent hazard function, the Gompertz distribution with $\alpha \neq 0$ is a reasonable choice.

2.2 Sparsity-inducing penalties

In this section, we describe some sparsity-inducing penalties and their main sparsity-inducing effects used in this thesis.

2.2.1 l_1-norm

$P(\theta) = \sum_{j=1}^{p} |\theta_j|$ is known to shrink a portion of the values of coefficients θ to exactly zero. Thus, it induces sparsity or variable selection procedure.

2.2.2 SCAD and MCP

The smoothly clipped absolute deviation (SCAD) penalty was proposed by [4]: $\lambda P(\theta) = \sum_j \mathrm{S}_{\lambda,\gamma}\big(|\theta_j|\big)$ with

$$\mathrm{S}_{\lambda,\gamma}\big(|\theta_j|\big) = \begin{cases} \lambda_j|\theta_j|, \text{ if } |\theta_j| \leq \lambda_j, \\ \frac{\gamma\lambda_j|\theta_j| - 0.5(|\theta_j|^2 + \lambda_j^2)}{\gamma - 1}, \text{ if } \lambda_j < |\theta_j| \leq \gamma\lambda_j, \\ \frac{\lambda_j^2(\gamma+1)}{2}, \text{ if } |\theta_j| > \gamma\lambda_j. \end{cases} \tag{2.11}$$

It applies the same rate of penalization as l_1-norm at the beginning, but continues relaxes, then drops to 0 when $|\theta_j| > \gamma\lambda_j$.

Another type of penalty, the minimax concave penalty (MCP), proposed by [5]: $\lambda P(\theta) = \sum_j \mathrm{M}_{\lambda,\gamma}\big(|\theta_j|\big)$ with

$$\mathrm{M}_{\lambda,\gamma}\big(|\theta_j|\big) = \begin{cases} \lambda_j|\theta_j| - \frac{\theta_j^2}{2\gamma}, \text{ if } |\theta_j| \leq \gamma\lambda_j, \\ \frac{\gamma\lambda_j^2}{2}, \text{ if } |\theta_j| > \gamma\lambda_j. \end{cases} \tag{2.12}$$

It behaves similarly with SCAD penalty. Both penalties demonstrate the oracle property [4, 5], i.e., the estimations having the same limiting distribution as the true model. Their extension to group structures are presented in [6, 7].

2.2.3 Group lasso

The group lasso is l_1/l_2-norm: $P(\theta) = \sum_{g\in\mathcal{G}} w_g||\theta_g||_2$ where $\mathcal{G}$ is a partition of p covariates and $(w_g)_{g\in\mathcal{G}}$ are positive weights. It induces variable selection in the group-wise fashion: all the variables in the same group are either selected or not selected simultaneously.

2.2.4 Overlapping group lasso

[8], [9] proposed the overlapping group lasso via latent variable formulation. In particular, it adapts *unions* of groups approach: the shared covariates are selected in the final model.

$$P(\theta) = \sum_{g\in\mathcal{G}} w_g||v_g||_2, \text{ s.t. } \begin{cases} \sum_{g\in\mathcal{G}} v_g = \theta, \\ \forall g \in \mathcal{G}, v_{g,j} = 0, \text{ if } j \notin g. \end{cases} \tag{2.13}$$

where v_g are latent parameter vectors that correspond to group g and represent θ linearly. In this latent (expanded and non-overlapping) space, the penalty formulation of v has the same structure as the group lasso formulation discussed above. It means that some vectors v_g are shrunk to zero, which leads to select overlapping groups of covariates.

2.3 Approximation approaches

2.3.1 Quadratic approximation

Given a function $f(\theta)$ that is twice differentiable.

1. Local quadratic approximation: The formula for quadratic approximation at θ_0 is given by

$$f(\theta) \approx f(\theta_0) + \frac{\partial f}{\partial\theta}(\theta_0)(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^T \frac{\partial^2 f}{\partial\theta\partial\theta^T}(\theta_0)(\theta - \theta_0), \tag{2.14}$$

   This works for values of θ close to θ_0.

2. Quadratic upper bound: It is used to constructs majorizing functions of objective functions in chapter 2. More specifically, a function $f(\theta)$ has bounded curvature, i.e., if there exists a

positive definite matrix M such that $M - \nabla^2 f(\theta)$ is nonnegative definite for all θ. Then,

$$f(\theta) \leq f(\theta_0) + \nabla f(\theta_0)^T(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^T M(\theta - \theta_0), \tag{2.15}$$

It means that to minimize the function $f(\theta)$, we minimize a quadratic upper bound.

2.3.2 Laplace approximation

It is used to approximate an integral of the following form:

$$\int_a^b g(x)dx, \tag{2.16}$$

where x is a $p-$dimensional vector. Let $g(x) = \exp(h(x))$ then

$$\int_a^b g(x)dx = \int_a^b \exp(h(x))dx, \tag{2.17}$$

We perform a multivariate Taylor series expansion and get a multivariate Gaussian integral.

$$\int_a^b \exp(h(x))dx \approx \exp(h(\tilde{x}))(2\pi)^{p/2}\det\bigl(\Sigma\bigr)^{-1/2}, \tag{2.18}$$

where $\tilde{x} = \underset{x}{\operatorname{argmax}}\ h(x)$ and Σ is the Hessian of $h(x)$ evaluated at $\tilde{x}$.

2.4 Optimization methods

2.4.1 Majorization-minimization (MM)

We briefly introduce MM algorithm [10, 11]. Let $f(\theta)$ be a real-valued function, $\theta^{(m)}$ be a fixed value of parameter θ and $g(\theta|\theta^{(m)})$ denote a real-valued function of θ whose form depends on $\theta^{(m)}$. The function $g(\theta|\theta^{(m)})$ is said to majorize $f(\theta)$ at the point $\theta^{(m)}$

$$g(\theta|\theta^{(m)}) \geq f(\theta) \text{ for all } \theta, \tag{2.19}$$

$$g(\theta^{(m)}|\theta^{(m)}) = f(\theta^{(m)}), \tag{2.20}$$

To minimize the function $f(\theta)$, we iteratively minimize minimizing a sequence of majorizing functions (or surrogate functions) $g(\theta|\theta^{(m)})$. This procedure posses the "descent property", which guarantees the numerical stability of MM algorithms. Specifically, denote $\theta^{(m+1)}$ be the minimizer of $g(\theta|\theta^{(m)})$ then

$$f(\theta^{(m+1)}) = f(\theta^{(m+1)})+g(\theta^{(m+1)}|\theta^{(m)})-g(\theta^{(m+1)}|\theta^{(m)}) = g(\theta^{(m+1)}|\theta^{(m)})+f(\theta^{(m+1)})-g(\theta^{(m+1)}|\theta^{(m)}),$$

Proof: From the above definition

$$\begin{aligned} & g(\theta^{(m+1)}|\theta^{(m)}) \geq f(\theta^{(m+1)}) \Rightarrow f(\theta^{(m+1)}) - g(\theta^{(m+1)}|\theta^{(m)}) \leq 0 \\ \text{and } & g(\theta^{(m+1)}|\theta^{(m)}) \leq g(\theta^{(m)}|\theta^{(m)}) \end{aligned}$$

Therefore,

$$f(\theta^{(m+1)}) = g(\theta^{(m+1)}|\theta^{(m)}) + f(\theta^{(m+1)}) - g(\theta^{(m+1)}|\theta^{(m)}) \leq g(\theta^{(m)}|\theta^{(m)}) + 0 = f(\theta^{(m)})$$

2.4.2 Coordinate descent

It is commonly used to solve the l_1-regularized optimization problems. In general, it iteratively optimizes the objective function with respect to one variable at a time while all others are kept fixed. To illustrate, we consider the simple case of univariate l_1-regularized problem given observations $\{(x_i, y_i)\}_{i=1}^n$

$$\min_{\theta\in\mathbb{R}} \frac{1}{2n}\sum_{i=1}^{n}(y_i - x_i\theta)^2 + \lambda|\theta| \tag{2.21}$$

The minimizer $\hat{\theta}$ is given by

$$\hat{\theta} = \begin{cases} \frac{1}{n}x^Ty - \lambda, \text{ if } \frac{1}{n}x^Ty > \lambda, \\ 0, \text{ if } \frac{1}{n}|x^Ty| \leq \lambda, \\ \frac{1}{n}x^Ty + \lambda, \text{ if } \frac{1}{n}x^Ty < -\lambda, \end{cases} = \mathcal{S}(\frac{1}{n}x^Ty, \lambda), \tag{2.22}$$

where $\mathcal{S}(.,.)$ is the soft-thresholding operator defined as

$$S(t,\lambda) = \text{sign}(t)(|t| - \lambda)_+, \tag{2.23}$$

with the operator $(t)_+$ equals to 0 if $t > 0$, and equals to 0 otherwise.

2.4.3 Alternating direction method of multipliers (ADMM)

We briefly introduce the standard ADMM algorithm that mostly based on [12]. ADMM solves convex optimization problems by splitting them into smaller pieces, each of which are then easier to handle. These problems typically have the form as follow

$$\min_{x,z} f(x) + g(z) \tag{2.24}$$

$$\text{s.t. } Ax + Bz = c, \tag{2.25}$$

where $f : \mathbb{R}^x \to (\mathbb{R} \cup \{\infty\})$ and $g : \mathbb{R}^z \to (\mathbb{R} \cup \{\infty\})$ are convex, $A \in \mathbb{R}^{m\times x}$ and $B \in \mathbb{R}^{m\times z}$.

2.4.3.1 Augmented Lagrangian

Augmented Lagrangian function of (2.25) is given by

$$L_\rho(x,z,y) = f(x) + g(z) + y^T(Ax + Bz - c) + \frac{\rho}{2}\|Ax + Bz - c\|_2^2, \tag{2.26}$$

where $y \in \mathbb{R}^m$ is the Lagrange multiplier and $\rho > 0$ is penalty parameter.

2.4.3.2 Optimality conditions

- Primal feasibility: $Ax + Bz - c = 0$
- Dual feasibility: $\triangledown f(x) + A^T y = 0$ and $\triangledown g(z) + B^T y = 0$

2.4.3.3 ADMM procedure

ADMM is an alternating minimization scheme for computing a saddle point of the augmented Lagrangian. It consists of three steps. First, L_ρ is minimized with respect to x, then with respect to z, and finally maximized with respect to y. It is shown in Algorithm 1.

Algorithm 1 ADMM algorithm

```
Initialize $z^{(0)} \in \mathbb{R}^z, y^{(0)} \in \mathbb{R}^m, \rho \in \mathbb{R}_+, k = 0$
repeat
    $x^{(k+1)} = \underset{x}{\text{argmin}}\, L_\rho(x, z^{(k)}, y^{(k)})$
    $z^{(k+1)} = \underset{z}{\text{argmin}}\, L_\rho(x^{(k+1)}, z, y^{(k)})$
    $y^{(k+1)} = y^{(k)} + \rho(Ax^{(k+1)} + Bz^{(k+1)} - c)$
    $k = k + 1$
until satisfies stopping criteria;
```

3. PENALIZED COX'S MODEL WITH GROUPED PREDICTORS [1]

The rapid development of next-generation sequencing technologies has made it possible to measure the expression profiles of thousands of genes simultaneously. Often, there exist group structures among genes manifesting biological pathways and functional relationships. Analyzing such high-dimensional and structural datasets can be computationally expensive and results in the complicated models that are hard to interpret. To address this, variable selection such as penalized methods are often taken. Here, we focus on the Cox's proportional hazards model to deal with censoring data. Most of the existing penalized methods for Cox's model are the group lasso methods that show deficiencies, including the over-shrinkage problem. In addition, the contemporary algorithms either exhibit the loss of efficiency or require the group-wise orthonormality assumption. Hence, efficient algorithms for general design matrices are needed to enable practical applications. In this chapter, we investigate and comprehensively evaluate three group penalized methods for Cox's model: the group lasso and two nonconvex penalization methods—group SCAD and group MCP—that have several advantages over the group lasso. These methods are able to perform group selection in both non-overlapping and overlapping cases. We have developed the fast and stable algorithms and a new package **grpCox** to fit these models without the initial orthonormalization step. The runtime of **grpCox** is improved significantly over the existing packages, such as **grpsurv** (for the non-overlapping case), **grpregOverlap** (overlapping), and **SGL**. In addition, **grpCox** is better than **grpsurv** and comparable with **SGL** in terms of variable selection performances. Comprehensive studies on both simulation and real-world cancer datasets demonstrate the statistical properties of our **grpCox** implementations with the group lasso, SCAD, and MCP regularization terms.

[1] Reprinted with permission from X. Dang, S. Huang, X. Qian, "Penalized Cox's Proportional Hazards Model for High-Dimensional Survival Data with Grouped Predictors", *Statistics and Computing*, vol. 31, no. 6, pp. 1-27, 2021. Copyright 2021 by Springer Nature.

3.1 Introduction

The Cox's proportional hazards model [3] is commonly used to study the relationship between survival time and a set of covariates in high-dimensional space as potential predictors for survival time. To tackle the curse of dimensionality and construct robust and interpretable models that generalize well, variable selection approaches, including penalization-based methods, are often taken.

Variable selection for the Cox's proportional hazards model has been extensively studied, including implementations based on lasso [13, 14, 15], adaptive lasso [5, 16], the smoothly clipped absolute deviation (SCAD) [17], to name a few. These methods can automatically select the important covariates by shrinking the coefficients of unimportant covariates to be exactly zero. However, these methods fail to produce good results when there exist group structures in covariates. A common group structure example is where each categorical covariate is expressed through a set of dummy variables. Group structures can also be introduced by integration of prior knowledge that is scientifically meaningful. For example, in gene expression analysis, genes belonging to the same biological pathway have similar functions and act together in regulating a biological system. These genes can be considered as a group.

Group selection in various statistical modeling problems has been considered in literature. [18] introduced the group lasso for linear regression with the l_2-norm of the coefficients for a group of covariates in the penalty function. [19] extended it to logistic regression. [20] used a general composite absolute penalty, which treats the group lasso as a special case. [6] introduced group SCAD to linear regression. The group minimax concave penalty (MCP) was presented in [7]. [21] introduced nonconvex penalties for linear and logistic regression. These works require the group-wise orthonormal condition to implement their algorithms. The solutions of the group lasso with non-orthonormal matrices for linear regression, logistic regression and SVM classifiers have been developed in literature [22, 23, 24].

There are, however, few extensions to the Cox's model. [25] applied the supervised group lasso to select both significant gene clusters and significant genes within these clusters for both logistic binary classification and Cox's survival model, for which the lasso and group lasso methods were implemented separately. In the first step, it identified important genes within each group based on the lasso formulation. In the second step, it selected important groups using the group lasso formulation. [23] introduced the sparse group lasso method combining the lasso with group lasso formulations to yield sparsity at both the group and individual levels for the Cox's proportional hazards model. [26] introduced the doubly regularized Cox regression that can deal with a mixture of individual sparsity and group sparsity with the extension to an overlapping case. Very recently, [27] presented a statistical approach that can handle sparse group lasso cases with superior variable selection performance.

In these existing penalized Cox's model with group structures, only the group lasso formulation has been considered because the group lasso penalty is convex for relatively straightforward optimization solutions. However, the group lasso penalty has deficiencies. Namely, large penalties are imposed on large coefficients, which leads to over-shrinking of large coefficients. As a result, the estimates of model coefficients are biased. To avoid over-shrinkage, the group lasso implementations often tend to reduce the penalty levels, which in turn results in selecting many variables. With the "oracle" property in SCAD and MCP penalty, the estimations having the same limiting distribution as the true model, both the group SCAD and group MCP formulations have been studied [28, 7, 21]. However, to the best of our knowledge, there is no effort to apply either the group SCAD or group MCP formulation in the Cox's model.

In this chapter, we investigate and comprehensively evaluate the group lasso, the group SCAD, and the group MCP penalized Cox's models. More critically, these three group penalty formulations with different mathematical structures, we would like to derive scalable and efficient optimization algorithms and open-access packages for more general group penalized Cox's models.

The existing group lasso based Cox's model implementations have used different algorithms to solve the corresponding optimization problem. [25] used a blockwise coordinate descent algorithm [29] to solve the group lasso problem. [26] used the cyclic coordinate descent algorithm and [23] used Nesterov's method. More recently, a group-wise descent algorithm was implemented in the R package **grpreg**, whose *grpsurv* function for the group penalized Cox's model as an extension of the methods presented in [21]. We will focus on developing and evaluating the group-wise descent algorithm for three group penalized Cox's models for its simplicity, speed, and stability. We have tried the cyclic coordinate descent algorithm, and found it inferior in both timing and accuracy to the group-wise descent algorithm. Specifically, while the group-wise algorithm can produce exact solutions for a single group in one step, the cyclic coordinate descent algorithm requires multiple iterations to converge to the same solution that leads to a loss of efficiency. Although Nesterov's method is a more general optimization method than the group-wise descent algorithm, it appears to be empirically slower than the group-wise descent algorithm for the specific problem of optimizing the group penalized Cox's models as shown in our running time comparison. The existing group-wise descent algorithm implemented in **grpreg** requires the group-wise orthonormal condition. Specifically, it needs to do an initial orthonormalization step, which leads to a different problem that is not equivalent to the original group lasso formulation [30, 7]. In particular, the new problem is to apply the l_2-penalty on the linear predictors instead of the original coefficients. Moreover, even though we can do orthonormalization for each group to make the observed data satisfy the group-wise orthonormal condition, the group-wise orthonormal condition can be easily violated when removing a fraction of the data or perturbing the dataset in bootstrap or sub-sampling as pointed out in [24]. Therefore, it is more favorable to solve the design matrices without the group-wise orthonormal condition. Our aim is to use the group-wise descent algorithm to handle the general design matrices of the three group penalized Cox's models. To achieve it, we adopt the majorization-minimization approach [10, 11] to derive the majorizing (surrogate) function of the objective function with closed-form expressions for a single group in gradient computation. We demonstrate that this algorithm is fast and efficient, and provide an open-access R package **grpCox**. Both simulation studies and real-world case studies provide comprehensive evaluation of our developed optimization algorithm for the three group penalized Cox's models.

The remainder of the chapter is organized as follows. Section 2 formulates the non-overlapping group penalized Cox's proportional hazards model. We introduce the majorization-minimization approach and group-wise descent algorithm for solving the group penalized Cox's model. Section 3 presents the extension with overlapping group penalty. Simulation results are reported in Section 4. The illustrations of our methods with real-world survival datasets are presented in Section 5. Section 6 concludes with discussion.

3.2 Non-overlapping groups

In this section, we present the Cox's model with non-overlapping groups of covariates as potential survival predictors, i.e. each potential predictor belongs to one and only one group. We first describe the general framework for group selection via the penalized partial likelihood of the Cox's model. We then derive the group-wise descent algorithms combining with the majorization-minimization approach for model inference.

3.2.1 Model formulation

Consider the standard survival data set of N subjects represented by the triplets $\{(Y_i, X^{(i)}, \delta_i)\}_{i=1}^N$, where Y_i denotes the survival time, $X^{(i)}$ a $P-$dimensional covariate vector, and δ_i the censoring indicator. With T_i and C_i denoting the survival time and the censoring time for subject i, the survival time Y_i is defined by $Y_i = \min\{T_i, C_i\}$ and the censoring indicator is defined as $\delta_i = \mathbf{I}_{T_i \le C_i}$. Suppose that P covariates belong to J non-overlapping groups I_j's such that $\{1, 2, \ldots, P\} = \cup_{j=1}^J I_j$ where the number of covariates in group I_j is p_j and $I_j \cap I_{j'} = \emptyset$ for $j \neq j'$. The $P-$dimensonal covariate vector for subjet i is $X^{(i)} = (X_1^{(i)}, \ldots, X_J^{(i)})$, where $X_j^{(i)}$ is a p_j-dimensional covariate vector of the j^{th} group for subject i. The corresponding coefficients of the covariates in the j^{th} group are β_j. The standard Cox's proportional hazards model of the hazard for patient i at time t can be written as [3]:

$$h(t|X^{(i)}) = h_0(t)\exp\big(X^{(i)}\beta\big) = h_0(t)\exp\big(\sum_{j=1}^{J} X_j^{(i)}\beta_j\big), \tag{3.1}$$

where $h_0(t)$ is the baseline hazard function.

Assume there is no ties in the observed times, and the censoring is non-informative. Let $t_1 < t_2 < \cdots < t_D$ be the distinct observed times where D is the number of unique observed failures.

R_i is the set of indices of the subjects who are at risk at time t_i. The partial likelihood function is given by

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp\big(\sum_{j=1}^{J} X_j^{(i)}\beta_j\big)}{\sum_{l\in R_i}\exp\big(\sum_{j=1}^{J} X_j^{(l)}\beta_j\big)}, \tag{3.2}$$

Penalization is one of the important variable selection methods, which can be applied to the Cox's model for better understanding survival predictors when P is large by minimizing the penalized partial likelihood function

$$\mathcal{L}(\beta) = -\frac{1}{N}\log\big(L(\beta)\big) + P_{\lambda,\gamma}(\beta) = \ell(\beta) + P_{\lambda,\gamma}(\beta), \tag{3.3}$$

where

$$\ell(\beta) = -\frac{1}{N}\sum_{i=1}^{D}\left[\big(\sum_{j=1}^{J} X_j^{(i)}\beta_j\big) - \log\Big(\sum_{l\in R_i}\exp\big(\sum_{j=1}^{J} X_j^{(l)}\beta_j\big)\Big)\right],$$

and the penalty term $P_{\lambda,\gamma}(\beta)$ can take different forms.

- Group lasso [18]: $P_\lambda(\beta) = \lambda\sum_j \sqrt{p_j}$
  $\|\beta_j\| = \sum_j \lambda_j\|\beta_j\|$, where $\lambda_j = \lambda\sqrt{p_j}$, $j = 1, \dots, J$.

- Group smoothly clipped absolute deviation (SCAD) [6]: $P_{\lambda,\gamma}(\beta) = \sum_j \mathrm{S}_{\lambda,\gamma}\big(\|\beta_j\|\big)$ with

$$\mathrm{S}_{\lambda,\gamma}\big(\|\beta_j\|\big) = \begin{cases} \lambda_j\|\beta_j\|, \text{ if } \|\beta_j\| \le \lambda_j, \\ \frac{\gamma\lambda_j\|\beta_j\| - 0.5(\|\beta_j\|^2 + \lambda_j^2)}{\gamma - 1}, \text{ if } \lambda_j < \|\beta_j\| \le \gamma\lambda_j, \\ \frac{\lambda_j^2(\gamma^2-1)}{2(\gamma-1)}, \text{ if } \|\beta_j\| > \gamma\lambda_j. \end{cases} \tag{3.4}$$

- Group minimax concave penalty (MCP) [7]: $P_{\lambda,\gamma}(\beta) = \sum_j \mathrm{M}_{\lambda,\gamma}\big(\|\beta_j\|\big)$ with

$$\mathrm{M}_{\lambda,\gamma}\big(\|\beta_j\|\big) = \begin{cases} \lambda_j\|\beta_j\| - \frac{\|\beta_j\|^2}{2\gamma} \text{ if } \|\beta_j\| \le \gamma\lambda_j, \\ \frac{1}{2}\gamma\lambda_j^2 \text{ if } \|\beta_j\| > \gamma\lambda_j. \end{cases} \tag{3.5}$$

Here $\|\cdot\|$ denotes the Euclidean vector norm. We scale by a factor of $\frac{1}{N}$ for convenience.

Given the survival data, the Cox's model inference is to learn β that minimizes the penalized partial likelihood function. Specifically,

$$\beta_{opt} = \underset{\beta}{\text{argmin}} \left[\ell(\beta) + P_{\lambda,\gamma}(\beta) \right] = \underset{\beta}{\text{argmin}}\, \mathcal{L}(\beta). \tag{3.6}$$

3.2.2 Majorization-minimization (MM) approach

The negative log partial likelihood $\ell(\beta)$ is convex and twice continuously differentiable. We adopt the majorization-minimization (MM) approach [10], [11] that involves majorizing the negative log partial likelihood $\ell(\beta)$. We derive the upper bound of $\ell(\beta)$ as the majorizing/surrogate objective function through its Hessian matrix.

Denote $\eta = X\beta$, then η is a $N-$dimensional vector whose i^{th} element is $\eta_i = X^{(i)}\beta$. We have

$$\ell(\eta) = -\frac{1}{N}\sum_{i=1}^{D}\left[\eta_i - \log\bigg(\sum_{l\in R_i}\text{exp}(\eta_l)\bigg)\right],$$

We can calculate the first- and second-order derivatives of $\ell(\beta)$; in particular, via the chain rule: $\ell^{'}(\beta) = X\ell^{'}(\eta)$ and $\ell^{''}(\beta) = X^T\ell^{''}(\eta)X$. Let $U = \ell^{'}(\eta)$ and $H = \ell^{''}(\eta)$ denote the corresponding gradient vector and Hessian matrix, respectively. We can write

$$U_d = \frac{\partial \ell}{\partial \eta_d} = -\frac{1}{N}\left[I_d - \sum_{i\in C_d}\frac{\text{exp}(\eta_d)}{\sum_{l\in R_i}\text{exp}(\eta_l)}\right],$$

where C_d is the set of subjects i's with $t_d \geq t_i$.

For the Hessian matrix H:

- If $d \neq k$, then

$$H_{d,k} = -\frac{1}{N}\left[\sum_{i\in C_d}\frac{\text{exp}(\eta_d)}{\sum_{l\in R_i}\text{exp}(\eta_l)}\right]\left[\sum_{i\in C_k}\frac{\text{exp}(\eta_k)}{\sum_{l\in R_i}\text{exp}(\eta_l)}\right],$$

  where C_k is the set of subjects i's with $t_k \geq t_i$.

- If $d = k$, e.g. the diagonal element,

$$H_{d,d} = \frac{1}{N}\sum_{i \in C_d}\left[\frac{\exp(\eta_d)}{\sum_{l \in R_i}\exp(\eta_l)} - \frac{\exp(\eta_d)\sum_{i \in C_d}\exp(\eta_d)}{\big(\sum_{l \in R_i}\exp(\eta_l)\big)^2}\right],$$

Let $w_d = \frac{1}{\sqrt{N}}\left[\sum_{i \in C_d}\frac{\exp(\eta_d)}{\sum_{l \in R_i}\exp(\eta_l)}\right]$, then $-H_{d,k} = w_d w_k$, and $H_{d,d} = \frac{1}{\sqrt{N}}w_d - w_d^2$.
Let $z^* = (z_1^*, z_2^*, \dots, z_P^*)$ be a $P-$dimensional vector, and B be a $P \times P$ matrix defined by $B = sX^TX$ where $s = \max_d\big(\frac{1}{\sqrt{N}}w_d\big)$. We have

$$(z^*)^T(B - \ell^{''}(\beta))z^* = (Xz^*)^T(s\mathbf{I}_N - \ell^{''}(\eta))(Xz^*),$$

where $\mathbf{I}_N$ is a $N \times N$ identity matrix. Let $Xz^* = z = (z_1, z_2, \dots, z_N)$ be a $N-$dimensional vector, then

$$\begin{aligned}
&(z^*)^T(B - \ell^{''}(\beta))z^* = z^T(s\mathbf{I}_N - \ell^{''}(\eta))z = \sum_{d=1}^{N} z_d\Bigg(z_d\big(s - H_{d,d}\big) + \sum_{k \neq d}^{N} z_k\big(-H_{d,k}\big)\Bigg)\\
&= \sum_{d=1}^{N}\big(s - H_{d,d}\big)z_d^2 + \sum_{d=1}^{N} z_d \sum_{k \neq d}^{N} z_k\big(-H_{d,k}\big) = \sum_{d=1}^{N}\big(s - H_{d,d}\big)z_d^2 + \sum_{d=1}^{N} z_d \sum_{k \neq d}^{N} z_k\big(w_d w_k\big)\\
&= \sum_{d=1}^{N}\big(s - H_{d,d}\big)z_d^2 + \sum_{d=1}^{N}(w_d z_d)\sum_{k \neq d}^{N}(w_k z_k) = \sum_{d=1}^{N}\big(s - H_{d,d} - w_d^2\big)z_d^2 + \big(\sum_{d=1}^{N} w_d z_d\big)^2\\
&\geq \sum_{d=1}^{N}\big(s - H_{d,d} - w_d^2\big)z_d^2 \geq \sum_{d=1}^{N}\big(s - \frac{1}{\sqrt{N}}w_d\big)z_d^2 \geq 0
\end{aligned}$$

Therefore, $(B - \ell^{''}(\beta))$ is nonnegative definite. It is worth nothing that without loss of generality, we may standardize the covariates first, as the estimated coefficients of the covariates can always be transformed back to the original scales for the sake of interpretation. We have $B = sX^TX \approx s\big(s^{'}N\mathbf{I}_P\big) = \tau\mathbf{I}_P$, where $\tau = s^{'}N\max_d\big(\frac{1}{\sqrt{N}}w_d\big)$ and $\mathbf{I}_P$ is a $P \times P$ identity matrix. Here $s^{'} = \frac{N}{P}$, if $N \geq P$, and $\frac{P}{N}$, if $N < P$.

Let β^* be the current solution of β, we can define the majorizing (surrogate) function of the negative log partial likelihood $\ell(\beta)$ as

$$\mathcal{M}(\beta|\beta^*) = \ell(\beta^*) + \ell^{'}(\beta^*)^T(\beta - \beta^*) + \frac{\tau}{2}(\beta - \beta^*)^T(\beta - \beta^*),$$

We further write the majorizing function of the objective function for the group penalized Cox's model in (5.3) as

$$\mathcal{Q}(\beta|\beta^*) = \ell(\beta^*) + \ell^{'}(\beta^*)^T(\beta - \beta^*) + \frac{\tau}{2}(\beta - \beta^*)^T(\beta - \beta^*) + P_{\lambda,\gamma}(\beta). \tag{3.7}$$

3.2.3 Group-wise descent algorithm

Now the estimator based on the majorizing function is defined as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\ \mathcal{Q}\big(\beta|\beta^*\big). \tag{3.8}$$

The asymptotic properties of this estimator have been investigated with the corresponding theorem and proofs given in Appendix 1. Here we focus on the optimization algorithm. To solve the minimization problem, we use the group-wise descent algorithm. This algorithm is essentially the same as the algorithm in [18] though we solve for general design matrices of the Cox's model. The idea behind it is that the algorithm optimizes the objective function with respect to a single group at a time, iteratively cycling through all groups until convergence conditions are satisfied. The overall structure of the group-wise descent algorithm is shown in **Algorithm 2**. In this algorithm, β^* refers to the current value of the Cox's model coefficients while $\hat{\beta}_j, \hat{\beta}$ are the updated values. This algorithm is suitable for fitting group lasso, group SCAD, and group MCP models since all three have closed-form expressions for a single-group update $\hat{\beta}_j$. These three group models have different mathematical formulations, so the closed-form expressions of a single-group updates for three models are different. The following parts present the derivations of $\hat{\beta}_j$ for three models. We prove that the algorithm possesses the descent property. Furthermore, we employ techniques to speed up the implementations of the corresponding algorithms considerably. Let us begin with the group lasso.

Algorithm 2 Group-wise descent algorithm for the group penalized Cox's model.

```
Initialize $\beta^*$
repeat
    for $j = 1, 2, \dots, J$ do
        Update $\hat{\beta}_j$ according to (3.10) for group lasso, (3.11) for group SCAD, or (3.12) for group MCP
    end
    Update $\beta^* = \hat{\beta}$
until Convergence of $\beta^*$;
```

3.2.3.1 Group lasso

The majorizing function (5.5) for the group lasso Cox's model can be written as

$$\mathcal{Q}(\beta|\beta^*) = \ell(\beta^*) + \ell^{'}(\beta^*)^T(\beta - \beta^*) + \frac{\tau}{2}(\beta - \beta^*)^T(\beta - \beta^*) + \sum_j \lambda_j \|\beta_j\|,$$

Let $\mathcal{Q}_j^{'}(\beta)$ be the partial derivative of $\mathcal{Q}(\beta)$ with respect to the group j. We have

$$\mathcal{Q}_j^{'}(\beta) = \ell_j^{'}(\beta^*) + \tau(\beta_j - \beta_j^*) + \begin{cases} \lambda_j \frac{\beta_j}{\|\beta_j\|}, & \text{if } \beta_j \neq \mathbf{0} \\ \lambda_j \|\mathbf{v}\|, & \text{if } \beta_j = \mathbf{0}. \end{cases} \tag{3.9}$$

where $\mathbf{v}$ is any vector satisfying $\|\mathbf{v}\| \leq 1$. Denote $\hat{\beta}_j$ is the solution to (5.2). It has the following closed-form expression

$$\hat{\beta}_j = \left(1 - \frac{\lambda_j}{\tau\|\mathbf{r}\|}\right)_+ \mathbf{r}, \tag{3.10}$$

where $\mathbf{r} = \beta_j^* - \frac{\ell_j^{'}(\beta^*)}{\tau}$ and $(x)_+ = \max\{x, 0\}$.

3.2.3.2 Group SCAD

The majorizing function (5.5) for the group SCAD Cox's model can be written as

$$\mathcal{Q}(\beta) = \ell(\beta^*) + \ell^{'}(\beta^*)^T(\beta - \beta^*) + \frac{\tau}{2}(\beta - \beta^*)^T(\beta - \beta^*) + \sum_j \mathrm{S}_{\lambda,\beta}\big(\|\beta_j\|\big),$$

The optimal solution is characterized by the partial derivative equation.

- If $\|\beta_j\| \leq \lambda_j$, then

$$\begin{cases} \ell_j^{'}(\beta^*) + \tau(\beta_j - \beta_j^*) + \lambda_j \frac{\beta_j}{\|\beta_j\|} = 0, & \text{if } \beta_j \neq \mathbf{0} \\ \ell_j^{'}(\beta^*) - \tau\beta_j^* + \lambda_j\|\mathbf{v}\| = 0, & \text{if } \beta_j = \mathbf{0}. \end{cases}$$

- If $\lambda_j < \|\beta_j\| \leq \gamma\lambda_j$, then

$$\ell_j^{'}(\beta^*) + \tau(\beta_j - \beta_j^*) + \frac{\gamma\lambda_j \frac{\beta_j}{\|\beta_j\|} - \beta_j}{\gamma - 1} = 0$$

- If $\|\beta_j\| > \gamma\lambda_j$, then

$$\ell_j^{'}(\beta^*) + \tau(\beta_j - \beta_j^*) = 0$$

where **v** is any vector satisfying $\|\mathbf{v}\| \leq 1$. By solving these equations, we find the final solutions

$$\hat{\beta}_j = \begin{cases} \left(1 - \frac{\lambda_j}{\tau\|\mathbf{r}\|}\right)_+ \mathbf{r}, \text{ if } \|\mathbf{r}\| \leq \left(\lambda_j + \frac{\lambda_j}{\tau}\right), \\ \frac{\tau(\gamma-1)}{\tau(\gamma-1)-1}\left(1 - \frac{\gamma\lambda_j}{\tau(\gamma-1)\|\mathbf{r}\|}\right)\mathbf{r}, \text{ if } \begin{cases} \tau(\gamma-1) - 1 > 0, \\ \left(\lambda_j + \frac{\lambda_j}{\tau}\right) < \|\mathbf{r}\| \leq \gamma\lambda_j, \end{cases} \\ \mathbf{r}, \text{ if } \|\mathbf{r}\| > \gamma\lambda_j. \end{cases} \tag{3.11}$$

where $\mathbf{r} = \beta_j^* - \frac{\ell_j^{'}(\beta^*)}{\tau}$ and $(x)_+ = \max\{x, 0\}$.

3.2.3.3 Group MCP

The majorizing function (5.5) for the group MCP Cox's model can be written as

$$\mathcal{Q}(\beta) = \ell(\beta^*) + \ell^{'}(\beta^*)^T(\beta - \beta^*) + \frac{\tau}{2}(\beta - \beta^*)^T(\beta - \beta^*) + \sum_j \mathrm{M}_{\lambda,\beta}\big(\|\beta_j\|\big),$$

The optimal solution is characterized by the partial derivative equation.

- If $\|\beta_j\| \leq \gamma\lambda_j$, then

$$\begin{cases} \ell_j^{'}(\beta^*) + \tau(\beta_j - \beta_j^*) + \lambda_j \frac{\beta_j}{\|\beta_j\|} - \frac{1}{\gamma}\beta_j = 0, & \text{if } \beta_j \neq \mathbf{0} \\ \ell_j^{'}(\beta^*) - \tau\beta_j^* + \lambda_j\|\mathbf{v}\| = 0, & \text{if } \beta_j = \mathbf{0}. \end{cases}$$

- If $\|\beta_j\| > \gamma\lambda_j$, then

$$\ell_j^{'}(\beta^*) + \tau(\beta_j - \beta_j^*) = 0.$$

where $\mathbf{v}$ is any vector satisfying $\|\mathbf{v}\| \leq 1$. By solving these equations, we find the final solutions

$$\hat{\beta}_j = \begin{cases} \frac{\tau\gamma}{\tau\gamma - 1}\left(1 - \frac{\lambda_j}{\tau\|\mathbf{r}\|}\right)_+ \mathbf{r}, \text{ if } \|\mathbf{r}\| \leq \gamma\lambda_j, \tau\gamma - 1 > 0 \\ \mathbf{r}, \text{ if } \|\mathbf{r}\| > \gamma\lambda_j. \end{cases} \tag{3.12}$$

where $\mathbf{r} = \beta_j^* - \frac{\ell_j^{'}(\beta^*)}{\tau}$ and $(x)_+ = \max\{x, 0\}$.

3.2.4 The descent property of group-wise descent algorithm

The surrogate function $\mathcal{Q}$ have two properties

$$\mathcal{Q}(\beta_j^*|\beta^*) = \mathcal{L}(\beta_j^*),$$
$$\mathcal{Q}(\beta_j|\beta^*) \geq \mathcal{L}(\beta_j) \text{ for all } \beta_j.$$

From that we can prove the descent property of the group-wise descent algorithm. The descent property is stated as follows. At every iteration of the proposed group-wise descent algorithms, let β^* and $\hat{\beta}$ denote the current value and the updated value of the coefficient estimator, respectively. Then the value of the objective function $\mathcal{L}(\beta)$ decreases, i.e., $\mathcal{L}(\hat{\beta}) \le \mathcal{L}(\beta^*)$.

Proof: From the second property of the surrogate function $\mathcal{Q}$ we have $\mathcal{L}(\hat{\beta}_j) \le \mathcal{Q}(\hat{\beta}_j)$. In addition, according to (5.4) we have $\mathcal{Q}(\hat{\beta}_j) \le \mathcal{Q}(\beta_j^*)$. Therefore, $\mathcal{L}(\hat{\beta}_j) \le \mathcal{Q}(\beta_j^*) = \mathcal{L}(\beta_j^*)$, which justifies the descent property of the group-wise descent algorithm. In other words, the objective function decreases after updating all groups in a cycle.

Lemma 1 *The objective function $\mathcal{Q}(\beta_j)$ is strictly convex with respect to β_j for the group lasso with $\tau > 0$, for the group SCAD with $\tau(\gamma - 1) > 1$, and for the group MCP with $\tau\gamma > 1$.*

Proof: Although $\mathcal{Q}(\beta_j)$ is not differentiable, it does possess twice directional derivatives everywhere. Let $\nabla_d^2 \mathcal{Q}(\beta_j)$ be the second order directional derivatives along the direction d, and denote $\epsilon^* = \min_{\beta_j, d} \nabla_d^2 \mathcal{Q}(\beta_j)$. Then, we have

- $\epsilon^* = \tau$ for group lasso
- $\epsilon^* = \tau - \frac{1}{\gamma - 1}$ for group SCAD
- $\epsilon^* = \tau - \frac{1}{\gamma}$ for group MCP.

These are positive under the conditions specified in the lemma. In other words, $\nabla_d^2 \mathcal{Q}(\beta_j)$ for all β_j and d, which means that the function $\mathcal{Q}(\beta_j)$ is strictly convex.

Remark: The objective function for the group lasso penalty is convex, thus the descent property of the algorithm implies the unique solution. However, the objective functions for the group SCAD and group MCP penalty are sums of convex and nonconvex components, thus it is possible that the algorithms converge to a local minimum.

3.2.5 Active set updates

To improve the computational speed, we have constructed an active set $A = \{\hat{\beta}_j \neq \mathbf{0}\}$ that takes advantage of the sparsity of β. As shown in **Algorithm 2**, we only need to update the nonzero coefficients $\hat{\beta}_j$ in A after a complete cycle has run through all the groups, i.e., when $\beta^* = 0$, $\hat{\beta}_j$ will stay zero if $\| - \frac{\ell_j^{'}(\mathbf{0})}{\tau}\| \leq \frac{\lambda_j}{\tau}$ or $\|\ell_j^{'}(\mathbf{0})\| \leq \lambda_j$; otherwise, $\hat{\beta}_j$ will be updated and stored in the active set if $\|\ell_j^{'}(\mathbf{0})\| > \lambda_j$. Therefore, the number of updates is reduced significantly and the rate of convergence of the algorithm is improved. The algorithm will stop if another complete cycle does not change this set. Note that the active set A can only become larger after each update, so the algorithm will always stop after a finite number of updates. More details of its convergence property can be found in [19].

3.2.6 Pathwise solution

The above procedure is just for one fixed value of λ. However, in general, it is of interest to be able to compute the optimal solution for a range of λ values. Thus, we aim to compute the regularization path (denoted as $\hat{\beta}(\lambda)$) where $\lambda \in [0, \infty]$. It can be shown that $\hat{\beta}(\lambda)$ turns out to be a piecewise linear, continuous function of λ [31]. In other words, we only need to compute the solutions on the change points in this path, denoted $\lambda_{max} \geq \lambda_1 \geq \cdots \geq \lambda_{min} \geq 0$. We can start with λ_{max} that is any value sufficiently large for which the entire coefficients $\beta^* = 0$. Notice that when $\beta^* = 0$, $\hat{\beta}_j$ will stay zero if $\| - \ell_j^{'}(\mathbf{0})\| \leq \lambda_j = \lambda\sqrt{p_j}$. Hence, we can set

$$\lambda_{max} = \max_j \left(\frac{\| - \ell_j^{'}(\mathbf{0})\|}{\sqrt{p_j}} \right).$$

Following the suggestions made by [32], we can ignore solutions that are close to 0 and set $\lambda_{min} = \epsilon\lambda_{max}$, then, compute the solutions over $m + 1$ values defined as $\lambda_i = \lambda_{max}\left(\frac{\lambda_{min}}{\lambda_{max}}\right)^{\frac{i}{m}}$, for $i = 0, 1, \ldots, m$. We set $\epsilon = 0.05$, if $N < P$, and 0.001, if $N \geq P$. In doing this, the algorithm usually converges well because we could use the preceding solution (i.e., for λ_i) as the initial values to obtain the solution for λ_{i-1}. It is worth noting that when $N < P$ and λ is small, the log likelihood estimates can be ∞. Therefore, when implementing our **grpCox** package, we terminates the regularization path if it occurs.

3.2.7 Selection of the tuning hyperparameters

With a path of solutions, we need to select an optimal one. The natural choice is by cross validation. However, the partial likelihood of the Cox's model is not as well defined as the Gaussian log likelihood or any exponential family on the left out samples using the traditional cross-validation, which leads to poor results. To tackle it, we have used the cross-validation method as described in [33] proposed for the Cox's model, in which data are split into k parts, use $k-1$ parts to train the model, and then, validate the learned model on the whole data set. The cross-validated log-partial likelihood for a given part i and λ is $\widehat{\text{CV}}_i(\lambda) = \mathcal{L}\big(\hat{\beta}_{-i}\big) - \mathcal{L}_{-i}\big(\hat{\beta}_{-i}\big)$, which can be used as the goodness-of-fit estimate of the solution. Here, $\hat{\beta}_{-i}$ and $\mathcal{L}_{-i}$ are the optimal coefficients and its corresponding log-partial likelihood for data excluding part i. The total goodness-of-fit, $\widehat{\text{CV}}(\lambda)$, is the sum of all $\widehat{\text{CV}}_i(\lambda)$. We find the optimal $\hat{\lambda}_{cvl}$ that maximizes $\widehat{\text{CV}}(\lambda)$.

This method alone produces high true positive rates (TPR) but often also with high false positive rates (FPR) for group lasso. We have implemented another approach proposed in [34] to reduce FPR without significant reduction of TPR. Let p_λ be the number of non-zero coefficients in the model for a given λ, the optimal λ maximizes

$$\widehat{\text{CV}}(\lambda) - \frac{\widehat{\text{CV}}(\hat{\lambda}_{cvl}) - \widehat{\text{CV}}(\lambda_{max})}{p_{\hat{\lambda}_{cvl}}} * p_\lambda, \text{ for } \lambda \in \left[\hat{\lambda}_{cvl}, \lambda_{max}\right].$$

Intuitively, it reduces the sparsity of the model p_λ without decreasing much the goodness-of-fit of the model $\widehat{\text{CV}}(.)$. The simulation studies for the second approach are presented in Appendix 2.

3.3 Overlapping groups

We have considered the non-overlapping group structure in the previous sections. In practice, however, a predictor can belong to several groups. For example, one gene can be shared by many different pathways. In this section, we extend the proposed methods for problems with overlapping groups. Note that the sparse group selection, which yields group-wise and within-group sparsity, can be considered as a special case of an overlapping group. Specifically, in this case, many groups would be of size 1.

Let us modify the notations and rewrite the penalty functions. Let $\mathcal{G} = \{g_1, \ldots, g_{|\mathcal{G}|}\}$ denote a set of groups as a partition of $\{1, \ldots, P\}$, $\boldsymbol{\beta}_g \in \mathcal{R}^{|g|}$ a subvector of $\boldsymbol{\beta}$, and p_g the number of covariates in each group g. The objective function becomes

$$\mathcal{L}(\beta) = -\frac{1}{N}\log\bigl(L(\beta)\bigr) + \Omega_{\lambda,\gamma}(\beta), \tag{3.13}$$

where

- Overlapping group lasso: $\Omega_\lambda(\beta) = \lambda \sum_{g\in\mathcal{G}} \sqrt{p_g}\|\beta_g\| = \sum_{g\in\mathcal{G}} \lambda_g\|\beta_g\|$ with $\lambda_g = \lambda\sqrt{p_g}$.

- Overlapping group smoothly clipped absolute deviation (SCAD): $\Omega_{\lambda,\gamma}(\beta) = \sum_{g\in\mathcal{G}} \mathrm{S}_{\lambda,\gamma}\bigl(\|\beta_g\|\bigr)$ with

$$\mathrm{S}_{\lambda,\gamma}\bigl(\|\beta_g\|\bigr) = \begin{cases} \lambda_g\|\beta_g\|, \text{ if } \|\beta_g\| \le \lambda_g, \\ \frac{\gamma\lambda_g\|\beta_g\| - 0.5(\|\beta_g\|^2 + \lambda_g^2)}{\gamma - 1}, \text{ if } \lambda_g < \|\beta_g\| \le \gamma\lambda_g, \\ \frac{\lambda_g^2(\gamma^2-1)}{2(\gamma-1)}, \text{ if } \|\beta_g\| > \gamma\lambda_g. \end{cases}$$

- Overlapping group minimax concave penalty (MCP): $\Omega_{\lambda,\gamma}(\beta) = \sum_{g\in\mathcal{G}} \mathrm{M}_{\lambda,\gamma}\bigl(\|\beta_g\|\bigr)$ with

$$\mathrm{M}_{\lambda,\gamma}\bigl(\|\beta_g\|\bigr) = \begin{cases} \lambda_g\|\beta_g\| - \frac{\|\beta_g\|^2}{2\gamma} \text{ if } \|\beta_g\| \le \gamma\lambda_g, \\ \frac{1}{2}\gamma\lambda_g^2 \text{ if } \|\beta_g\| > \gamma\lambda_g. \end{cases}$$

where $\|\cdot\|$ is the Euclidean vector norm.

Also, it is worth clarifying about how the overlapping group works. For example, consider $P = 3$ and $G = 2$, two groups sharing one covariate, and only the first group affecting the survival outcome. When the second group is not selected, all of its coefficients are shrunk to zeros. On the other hand, as the first group is selected, all of its coefficients are nonzeros. One approach, presented in [8], [9], considered *unions* of groups: the shared covariates are selected in the final model. Another approach, presented in [35], considered *intersections* of groups: the shared covariates are not selected. In this chapter, we consider the *union* approach.

The main difficulty in solving (4.8) is from the non-separable $\{\beta_g\}_{g\in\mathcal{G}}$ in the non-smooth penalty $\Omega_{\lambda,\gamma}(\beta)$. The overlapping character makes the computation of the subgradient with respect to $\boldsymbol{\beta}_g$ in the group-wise descent algorithm challenging. To tackle this problem, we have adopted the latent group approach [8], [9] that replicates a variable in whatever group it appears; then fits the non-overlapping group models. Note that "latent" here does not imply the case that the group structure is unobservable - we consider the cases where the group structure is known in advance, which is called *predefined* group structure. Rather, "latent" implies the set of latent variables, which are formed as linear combinations of predefined groups. Next, we discuss with more details.

Let $\nu_g \in \mathcal{R}^P$ be a vector that is zero everywhere except in those positions corresponding to the elements of group g, and let $\mathcal{V}_g \subseteq \mathcal{R}^P$ be the subspace of these possible vectors ν_g. Hence, $\boldsymbol{\beta} = \sum_{g=1}^{|\mathcal{G}|} \nu_g$. Figure 3.1 illustrates the idea how to transform $X\beta = \tilde{X}\nu$, where ν is the latent variable, and $\tilde{X}$ is the replicated variable matrix.

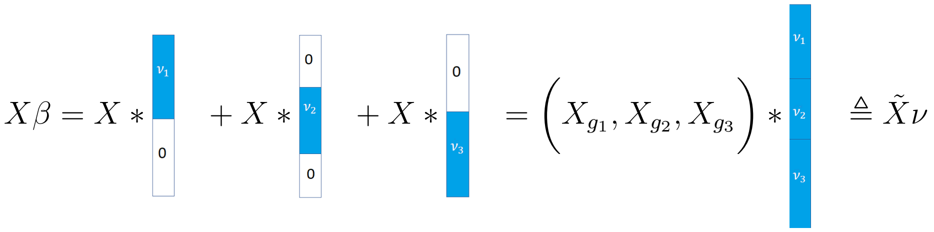


Figure 3.1: The coefficient decomposition of overlapping groups.

We can reformulate the objective function (4.8) in the latent variable space as

$$\mathcal{L}(\nu) = -\frac{1}{N}\log\big(L(\nu)\big) + \Omega_{\lambda,\gamma}(\nu), \tag{3.14}$$

Three penalty formulations can be similarly defined:

- Overlapping group lasso: $\Omega_\lambda(\nu) = \lambda \sum_{g\in\mathcal{G}} \sqrt{p_g}\|\nu_g\| = \sum_{g\in\mathcal{G}} \lambda_g\|\nu_g\|$ with $\lambda_g = \lambda\sqrt{p_g}$.

- Overlapping group smoothly clipped absolute deviation (SCAD): $\Omega_{\lambda,\gamma}(\nu) = \sum_{g\in\mathcal{G}} \mathrm{S}_{\lambda,\gamma}\big(\|\nu_g\|\big)$ with

$$\mathrm{S}_{\lambda,\gamma}\big(\|\nu_g\|\big) = \begin{cases} \lambda_g\|\nu_g\|, \text{ if } \|\nu_g\| \le \lambda_g, \\ \frac{\gamma\lambda_g\|\nu_g\| - 0.5(\|\nu_g\|^2+\lambda_g^2)}{\gamma-1}, \text{ if } \lambda_g < \|\nu_g\| \le \gamma\lambda_g, \\ \frac{\lambda_g^2(\gamma^2-1)}{2(\gamma-1)}, \text{ if } \|\nu_g\| > \gamma\lambda_g. \end{cases}$$

- Overlapping group minimax concave penalty (MCP): $\Omega_{\lambda,\gamma}(\nu) = \sum_{g\in\mathcal{G}} \mathrm{M}_{\lambda,\gamma}\big(\|\nu_g\|\big)$ with

$$\mathrm{M}_{\lambda,\gamma}\big(\|\nu_g\|\big) = \begin{cases} \lambda_g\|\nu_g\| - \frac{\|\nu_g\|^2}{2\gamma} \text{ if } \|\nu_g\| \le \gamma\lambda_g, \\ \frac{1}{2}\gamma\lambda_g^2 \text{ if } \|\nu_g\| > \gamma\lambda_g. \end{cases}$$

where $\|\cdot\|$ is the Euclidean vector norm.

Here, $L(\nu)$ is analogous to $L(\beta)$, but it is worth noting that $L(\beta)$ is computed in the original β space using the design matrix X while $L(\nu)$ is computed in the latent ν space using the replicated variable matrix $\tilde{X}$. In the latent (expanded and non-overlapping) space of dimension $\sum_{g\in\mathcal{G}} |g|$, the formulation has the same structure as the non-overlapping group formulations discussed previously. This allows us to apply the same solution procedure presented in the previous sections.

3.4 Simulation studies

In this section, we first show the efficiency of our proposed algorithms and package **grpCox** [36] by comparing the running time to fit the entire path of solutions with other publicly available R packages. We also compare these packages in term of variable selection. Then, we illustrate the similarities and differences between three group regularization methods: group lasso, group SCAD, and group MCP in both the non-overlapping group and overlapping group settings. Finally, we compare the performance of three methods in terms of variable selection and model accuracy in both the non-overlapping group and overlapping group cases.

3.4.1 Setup

We generate data with N observations and P covariates from the following model:

$$Y^{true} = \text{exp}\big(X\beta\big),$$

where Y^{true} is the true survival time. The censoring time C is generated from a exponential distribution with the mean $U\text{exp}\big(X\beta\big)$, where U is randomly generated from a uniform distribution $U(0, c)$. The recorded survival time is $Y = \text{min}\{Y^{true}, C\}$. The observation is censored if $C < Y^{true}$. We choose different c to achieve different censoring rates. The original covariates X are generated from a multivariate normal distribution with a zero mean vector and the correlation matrix $\mathbf{C}$ as an autoregressive matrix where $\mathbf{C}_{ij} = \rho^{|i-j|}$ and $0 \le \rho \le 1$. The reason to use an autoregressive correlation matrix is that we could flexibly tune the correlation between covariates by setting ρ values: $\rho = 0$ means no correlation between covariates, while $\rho = 1$ means that the covariates are perfectly correlated as duplicates of each other. In all the simulations, we fix $\gamma = 3.7$ for the group SCAD formulation as suggested in [4], and $\gamma = 3$ for the group MCP formulation as suggested in [37].

We evaluate the variable selection performance of these methods by presenting the model sizes, true positive rate (TPR), and false positive rate (FPR). These measures are defined as

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \text{ and FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}},$$

where TP, FP, FN, TN are the number of true positives, false positives, false negatives and false negatives, respectively. For all simulations, we create a path of 50 λ values, apply 10-fold cross-validation described above to select the optimal λ for variable selection.

We evaluate model accuracy by root mean square error (RMSE) that is given by

$$\text{RMSE} = \sqrt{\frac{1}{P}\sum_{p=1}^{P}(\beta_p - \hat{\beta_p})^2}.$$

Recall that P is the number of covariates.

3.4.2 Time and quality comparison with other packages

In this section, we compare the running time of the **grpCox** package, in which we implement our methods, with other publicly available R packages for fitting models. We also compare them in term of variable selection using TPR and FPR measurement.

Package	Method	$\{N = 50, P = 1000\}$			$\{N = 100, P = 3000\}$			$\{N = 150, P = 4500\}$		
		time	TPR	FPR	time	TPR	FPR	time	TPR	FPR
grpCox	Group lasso	0.05	0.50	0.10	0.15	0.97	0.15	0.26	1	0.15
	Group SCAD	0.30	0.54	0.10	0.33	0.99	0.06	0.52	1	0.13
	Group MCP	0.28	0.47	0.08	0.31	0.99	0.04	0.50	1	0.12
grpsurv	Group lasso	0.08	0.10	0.05	0.28	0.59	0.06	0.52	0.98	0.08
	Group SCAD	0.18	0.09	0.03	0.72	0.46	0.04	1.31	0.86	0.04
	Group MCP	0.14	0.01	0.01	0.48	0.16	0.01	0.85	0.56	0.02
SGL	Group lasso	7.55	0.33	0.06	38.73	1	0.10	87.84	1	0.10

Table 3.1: Comparison of **grpCox** with publicly available packages in the non-overlapping settings. The mean time, average TPRs, and average FPRs, over 100 independent data sets and a 50 λ values path, are reported. The time is in seconds.

3.4.2.1 Non-overlapping groups

We consider two other R packages **SGL** [23] and **grpsurv**, which is a part of the **grpreg** package [21]. Note that **SGL** package is not for the overlapping group case.

We consider three high-dimensional settings $(N, P) = \{(50, 1000), (100, 3000), (150, 4500)\}$. In this set of experiments, β is sparse including 100 nonzero elements and $(P-100)$ zero elements. Each group includes 10 covariates, and the corresponding numbers of groups J are set to 100, 300, 450. No censoring, and $\rho = 0.5$. We set $\alpha = 0$ for the group lasso penalty when implementing the **SGL** package. We compute the 50 λ value solution paths of the group penalized Cox models for 100 independent data sets, and report the average running time. The 10-fold cross-validation is used for model selection. The results are shown in Table 3.1.

Table 3.2: Comparison of **grpCox** with publicly available packages in the overlapping and $N > P$ settings. The mean time, over 100 independent data sets and a 50 λ values path, is reported in seconds.

	Package	Method			
			Equal group	Unequal group	Sparse group
Overlapping		$\mathbf{N < P}$	$\{N = 50, P = 802\}$	$\{N = 50, P = 835\}$	$\{N = 50, P = 1000\}$
	grpCox	Group lasso	0.17	0.17	1.46
		Group SCAD	0.34	0.30	1.63
		Group MCP	0.33	0.31	1.65
	grpregOverlap	Group lasso	0.28	0.29	2.57
		Group SCAD	0.27	0.26	2.52
		Group MCP	0.26	0.26	2.52
	SGL	Group lasso	-	-	8.14
Non-overlapping		$\mathbf{N \geq P}$	$\{N = 100, P = 50\}$	$\{N = 300, P = 100\}$	$\{N = 6000, P = 1000\}$
	grpCox	Group lasso	0.03	0.06	5.97
		Group SCAD	0.03	0.06	5.75
		Group MCP	0.02	0.06	5.53
	grpsurv	Group lasso	0.05	0.20	35.59
		Group SCAD	0.03	0.19	16.42
		Group MCP	0.02	0.11	15.71
	SGL	Group lasso	2.06	9.84	-

The running time results show that **grpCox** is faster than **grpsurv**, and both of them run much faster than **SGL**. Among different methods, group lasso is the fastest that followed by the group SCAD and group MCP. It can be explained that the upper bound for group lasso is sufficiently tight and convex, which leads to faster convergence.

From Table 3.1, it can be seen that the TPR values of **grpCox** are much higher than **grpsurv** while the FPR values of **grpCox** are a bit higher than **grpsurv**. In other words, **grpCox** gives better results than **grpsurv** in term of variable selection. In addition, **grpCox** is comparable with **SGL** in term of variable selection. It can be explained that both **grpCox** and **SGL** can handle general design matrices while **grpsurv** does an initial orthonormalization step, which can be easily violated when applying cross-validation to select models. Even worse, it may cause the significant differences in TPR and FPR for group SCAD and group MCP from group lasso in **grpsurv** results.

In addition, we would like to show how these methods scale with N and P. We run simulations with $\rho = 0$, 20% censoring rate fixed and different setups for the number of subjects N and the number of covariates P. For each (N, P) pair, we solve for a path of 50 λ values. Figure 3.2 shows the corresponding runtime for fixed P as N changes, and for fixed N as P changes. We can see that all three methods are scalable to both N and P and handle large N and large P well. The presented setups are with the maximum N at 50000 and the maximum P at 450000.

3.4.2.2 Overlapping groups

We consider one available R package **grpregOverlap** [38]. Here, we show the running time of three high-dimensional overlapping settings with $N = 50$ samples for each. 20% censoring, and $\rho = 0.5$. Firstly, the equal group case includes $P = 802$ covariates with 100 groups of 10 covariates with two of them overlapping between two successive groups, and there are 81 nonzero covariates. Secondly, the unequal group case includes $P = 835$ covariates with 30 groups of 8 covariates with two of them overlapping between two successive groups, 30 groups of 11 covariates with three of them overlapping between two successive groups, and 40 groups of 15 covariates with five of them overlapping between two successive groups. There are 98 nonzero covariates. Lastly, the sparse case includes $P = 1000$ covariates with 100 groups of 10 covariates. There are 10 sparse groups.

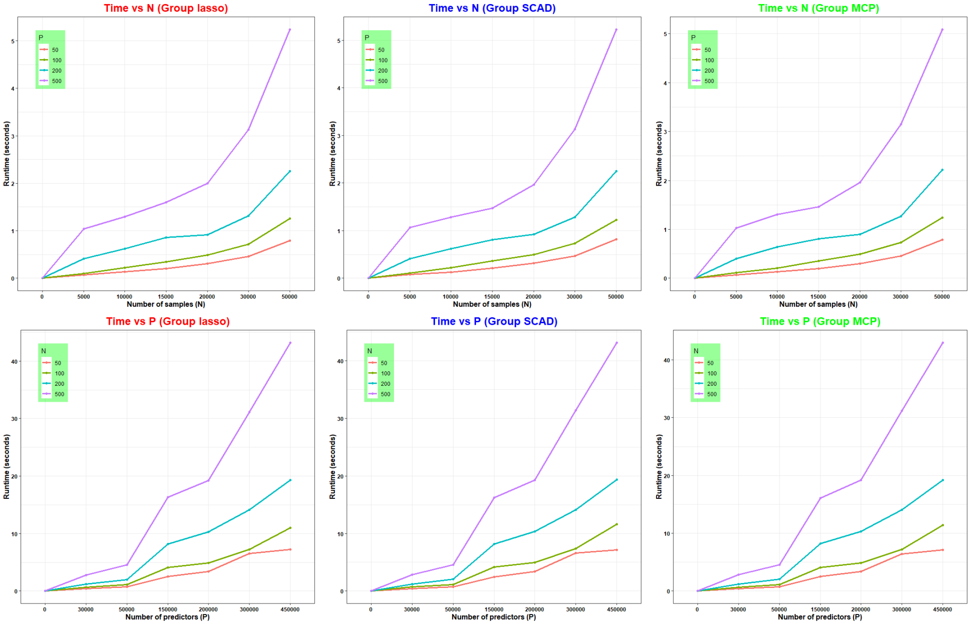


Figure 3.2: Plots of average runtime over 100 trials for 50 λ-value paths. The runtime is in seconds.

We also include the running time of the **SGL** package with $\alpha = 0.5$ for the sparse group case. Note that **grpregOverlap** does not include the model selection for Cox's model, so we choose not to report the TPRs and FPRs for all packages. The running time results are summarized in Table 3.2. It can be seen that for group lasso, **grpCox** is faster than **grpregOverlap** that followed by **SGL**. For group SCAD and group MCP, **grpCox** is faster than **grpregOverlap** in the sparse group setting, but a bit slower in the equal and unequal settings.

3.4.2.3 $N \geq P$ problems

We show that **grpCox** also can deal with large datasets by considering the running time results for three combinations of $(N, P) = \{(100, 50), (300, 100), (6000, 1000)\}$. The corresponding numbers of equal groups J are set to 10, 10, 100. In this set of experiments, β is sparse with $P/10$ elements are nonzero. We set $\alpha = 0$ for the group lasso penalty when implementing the **SGL** package. We compute the 50 λ value solution paths of the group penalized Cox models for 100 independent data sets, and report the average running time. However, we could run the **SGL** package with reasonable running time on small data sets only. The results are shown in Table 3.2. The results are consistent with high-dimensional cases: **grpCox** is faster than **grpsurv**, and both of them run much faster than **SGL**. However, group SCAD and group MCP are a bit faster than group lasso especially in the $(N, P) = (6000, 1000)$ case of **grpsurv** implementation that are presumably because their solution paths tend to be more sparse [21].

Note: Group SCAD and MCP models depend on an additional parameter γ. In particular, small changes of γ can lead the implementations terminate at different λ values along the regularization path, which results in big running time changes. Here we used the fix γ values suggested in [4] for group SCAD and [37] for group MCP that gave good results in term of variable selection and model accuracy (more details presented the following parts.) How to determine the optimal γ value, however, definitely needs further investigation.

3.4.3 Comparison of three group penalized Cox's models

In this section, we illustrate the similarities and differences between three group regularization methods: group lasso, group SCAD, and group MCP in both the non-overlapping and overlapping group settings using simulated data.

3.4.3.1 Non-overlapping groups

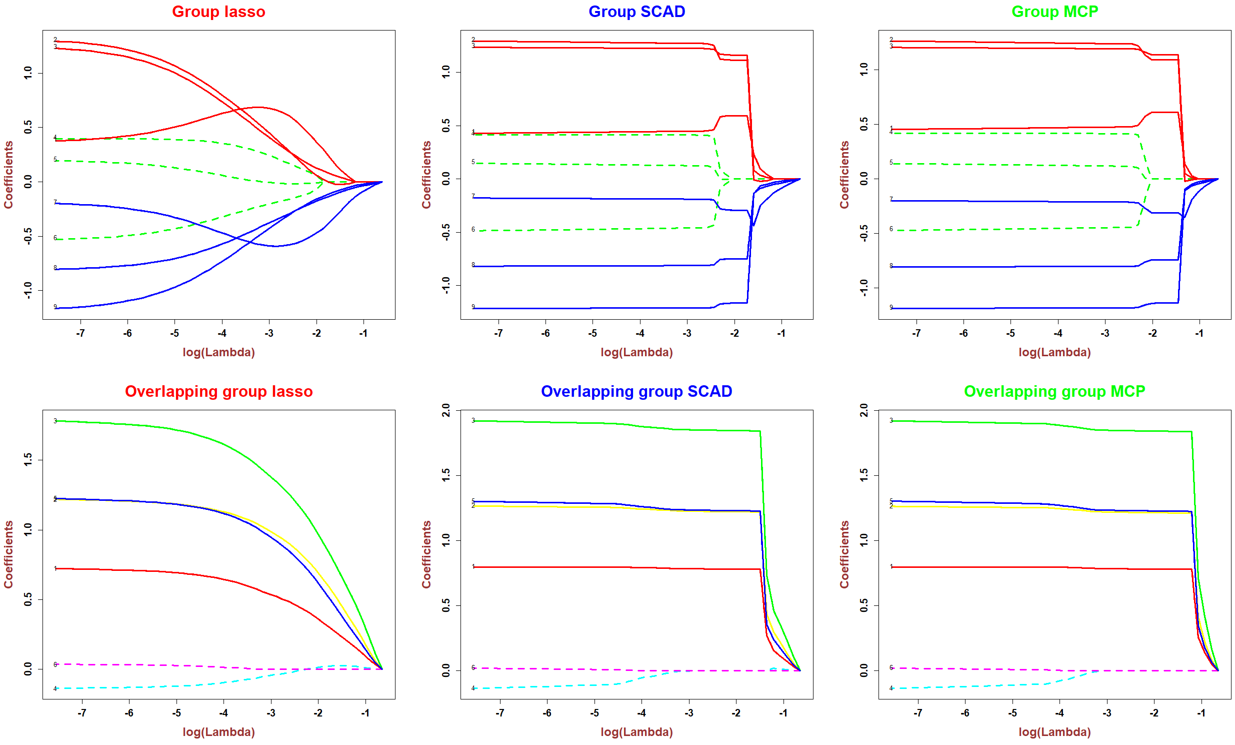
We consider a simple example with five primary covariates that are generated from a multivariate normal distribution with the zero mean vector and the correlation matrix $\mathbf{C}$ with $\mathbf{C}_{ij} = \rho^{|i-j|}$ and $\rho = 0.5$. The true survival time is generated as follows:

$$Y^{true} = \exp(X_1 + X_1^2 + X_1^3 - 0.7X_5 - 0.95X_5^2 - 0.8X_5^3).$$

In other words, this model includes nine covariates that can be divided into three groups: the first group is $\{X_1, X_1^2, X_1^3\}$, the second group $\{X_2, X_3, X_4\}$, and the third group $\{X_5, X_5^2, X_5^3\}$. Note that the first and third groups have nonzero coefficients while the second group has zero coefficients. The sample size N is 50, and the censoring rate is 20%. We create a path of 50 values of λ. The resulting solution paths are shown in Figure 4.1.

It is easy to see that the group selection selects a group of covariates in an "all-in-or-all-out" fashion. In other words, once one covariate of a group is selected, the whole group will be selected.

Figure 3.3: Solution paths for the group lasso, group SCAD, and group MCP models. The solid lines are for signal variables while the dashed lines are for noise variables.



In addition, the group SCAD and group MCP methods eliminate some of the bias towards zero among the true nonzero groups. In particular, when $\log(\lambda)$ is between -1.17 and -1.88, they produce the estimated model including only the nonzero covariates (the "oracle" model).

3.4.3.2 Overlapping groups

We also consider a simple example with six covariates that are generated from a multivariate normal distribution with the zero mean vector and the correlation matrix $\mathbf{C}$ with $\mathbf{C}_{ij} = \rho^{|i-j|}$ and $\rho = 0.5$. There are five groups defined as $g_1 = \{X_1, X_2, X_3\}$, $g_2 = \{X_1, X_4\}$, $g_3 = \{X_2, X_4, X_5\}$, $g_4 = \{X_3, X_5\}$, $g_5 = \{X_6\}$. The true survival time is generated as follows:

$$Y^{true} = \exp(0.8X_1 + X_2 + 2X_3 + X_5).$$

The sample size N is 100, and the censoring rate is 20%. We create a path of 50 λ values. The resulting solution paths are shown in Figure 4.1. The results are consistent with the results of the non-overlapping group cases. The group SCAD and group MCP methods again reduce the bias towards zero among the true nonzero groups. In particular, when $\log(\lambda)$ is between -1.5 and -2.76, they produce the estimated model including only the nonzero covariates.

3.4.4 Comparison of three group penalized Cox's models with non-overlapping groups

In this section, we compare the performance of three group regularization methods in terms of variable selection and model accuracy using simulated data. In here, the model size is given in terms of the number of groups. Clearly, the true model size is the number of nonzero groups. The group size is the number of covariates of each group.

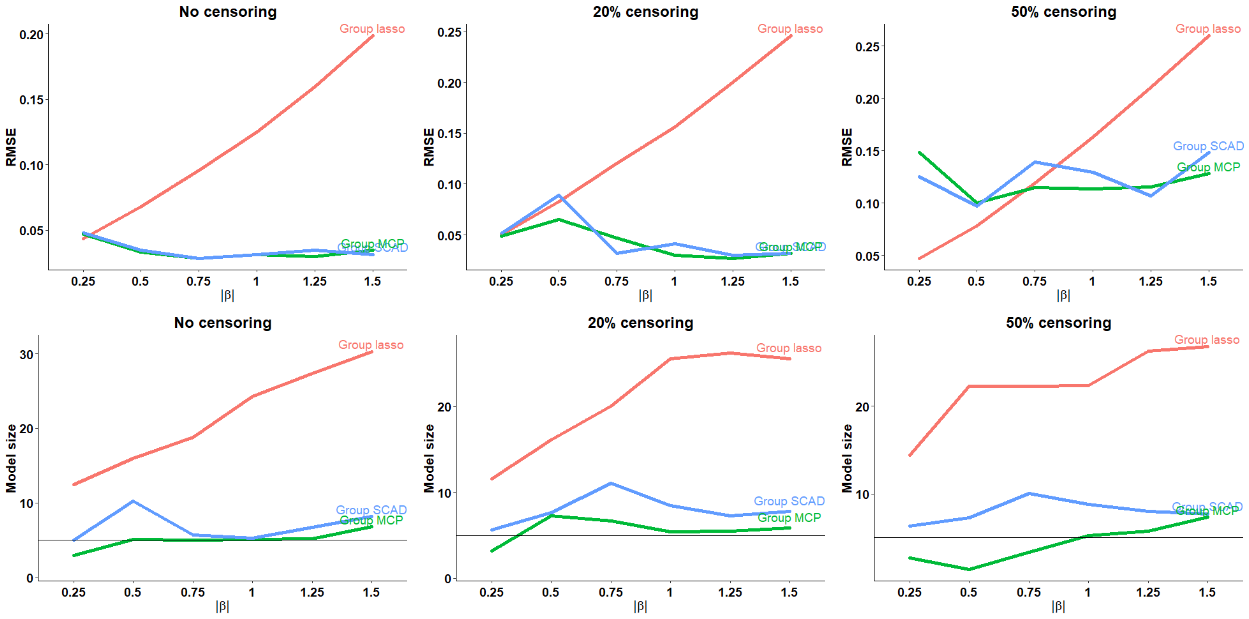
3.4.4.1 Effect of the coefficient magnitude

We focus on high dimensional cases, therefore, we generate $N = 100$ observations with $P = 400$ covariates that include 100 groups, each with 4 elements. There are five nonzero groups whose coefficient magnitudes are $\pm\beta$ where β is a scalar, and ninety-five other groups are zero groups. We vary $|\beta|$ between 0.25 and 1.5. We also investigate the effects of the censoring setting by considering two scenarios: no censoring and right censoring with 20% censoring rate.

The results in terms of estimation accuracy and model sizes are shown in Figure 3.4. The results show that when the coefficient values are small, all three methods have the same RMSE values. However, group SCAD and group MCP methods perform better with decreasing RMSE values, while the group lasso method performs increasingly poorly. Moreover, group SCAD and group MCP methods always select smaller models and approach the true model size while the group lasso method often selects too many covariates. Comparing group SCAD and group MCP, the two are nearly identical in terms of estimation accuracy. However, the group MCP method selects smaller models than the group SCAD method does.

The TPR and FPR results are summarized in Table 3.3. They illustrate that when the coefficients are small, group lasso does variable selection better than group SCAD and group MCP.

Figure 3.4: The impact of the coefficient magnitude and censoring rate on group regularization methods when the group size is 4. The *black line* is the true model size (5).



However, group MCP begins doing better variable selection than group SCAD that produces better variable selection than group lasso.

3.4.4.2 Effect of the group size

We use the same setting as it was described previously, but the group sizes are different. We consider two different cases. In the first case, the group size is 10, and the number of groups is 40. The first two groups are nonzero groups; other groups are zero groups. These results are shown in Figure 3.5 and Table 3.4. In the second case, the group size is 20, and the number of groups is 20. Only the first group was nonzero group; other groups were zero groups. The results are shown in Figure 3.5 and Table 3.4.

Figure 3.5 shows the same pattern as in Figure 3.4. However, when the group size increases, the RMSE values decrease. Comparing Tables 3.3 and 3.4, it can be seen that when the group size increases, group lasso performs worse with much higher FPR values. The group SCAD gives higher TPR values, but a little bit higher FPR values when the coefficient magnitude increases. The group MCP gives better performance when the group size increases.

		Group lasso		Group SCAD		Group MCP	
	$\lvert\beta\rvert$	TPR	FPR	TPR	FPR	TPR	FPR
No censoring	0.25	0.95	0.09	0.73	0.02	0.71	0.00
	0.50	1	0.12	0.91	0.04	1	0.01
	0.75	1	0.15	1	0.02	1	0.01
	1.00	1	0.21	1	0.01	1	0.01
	1.25	1	0.24	1	0.03	1	0.01
	1.50	1	0.27	1	0.04	1	0.03
20% censoring	0.25	0.54	0.09	0.50	0.04	0.50	0.01
	0.50	1	0.13	0.84	0.04	1	0.05
	0.75	1	0.17	1	0.07	1	0.03
	1.00	1	0.22	1	0.05	1	0.01
	1.25	1	0.23	1	0.03	1	0.02
	1.50	1	0.22	1	0.04	1	0.02
50% censoring	0.25	0.75	0.12	0.65	0.04	0.33	0.01
	0.50	1	0.19	0.91	0.04	0.33	0.00
	0.75	1	0.19	0.93	0.07	0.66	0.00
	1.00	1	0.19	1	0.05	0.92	0.02
	1.25	1	0.23	1	0.04	0.98	0.02
	1.50	1	0.24	0.96	0.04	0.97	0.04

Table 3.3: Average true positive rate (TPR) and false positive rate (FPR) values of three group regularization methods over 100 replications for different coefficient magnitude values and different censoring scenarios when the group size is 4.

3.4.4.3 Effect of censoring

We investigate the performance of three methods with respect to the censoring rate. We use the same setting, in which the group size is 4 with the higher censoring rate 50%. The results are summarized in Figure 3.4 and Table 3.3. From Figure 3.4 and Table 3.3, on one hand, it can be seen that there is no big difference in terms of RMSE, model size, and variable selection (TPR and FPR) between no censoring and 20% censoring. On the other hand, 50% censoring affects slightly on group lasso and group SCAD, but strongly on group MCP especially when the coefficients are small. It may be explained by the fact that the presence of censoring reduces the available sample size, which leads to inconsistent estimation.

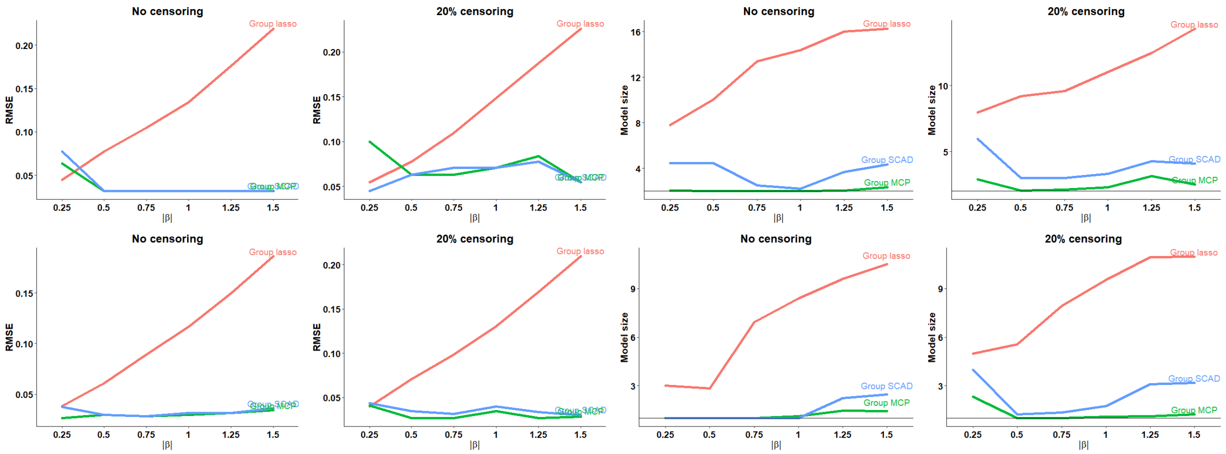


Figure 3.5: The impact of increasing coefficient magnitude on group regularization methods when the group size is 10 (first row) and 20 (second row). The *black line* on the right is the true model size.

3.4.4.4 Effect of covariate correlation

In all the above simulations, we set the population correlation $\rho = 0$. In other words, covariates are generated independently from the standard normal distribution. In this section, we still set the group size to be 4, no censoring, but the values of ρ at 0.2, 0.5 and 0.9. The results are shown in Figure 3.6 and Table 3.5. It can be seen that when the population correlation is mild, e.g. not larger than 0.5, all the three models work fine. In particular, the group MCP formulation performs the best while the group lasso performs the worst in terms of TPR and FPR values. The model with the group MCP penalty also leads to smaller models that approach the true model sizes compared to much bigger models from the group lasso model. When the population correlation is high at 0.9, all three models have bigger RMSE and smaller TPR values. The group MCP and group SCAD formulations still derive models with similar size as in the mild population correlation cases. The group lasso formulation becomes more conservative, which leads to smaller selected models whose sizes are close to the true size.

			Group lasso		Group SCAD		Group MCP	
Group size		$\|\beta\|$	TPR	FPR	TPR	FPR	TPR	FPR
10	No censoring	0.25	1	0.15	0.78	0.08	0.51	0.03
		0.50	1	0.21	1	0.06	1	0.00
		0.75	1	0.30	1	0.01	1	0.00
		1.00	1	0.33	1	0.01	1	0.01
		1.25	1	0.37	1	0.04	1	0.00
		1.50	1	0.38	1	0.06	1	0.01
	20% censoring	0.25	1	0.16	1	0.10	0.97	0.03
		0.50	1	0.19	1	0.03	1	0.00
		0.75	1	0.20	1	0.03	1	0.00
		1.00	1	0.24	1	0.03	1	0.01
		1.25	1	0.28	1	0.06	1	0.03
		1.50	1	0.32	1	0.05	1	0.01
20	No censoring	0.25	1	0.10	1	0.00	1	0.00
		0.50	1	0.10	1	0.00	1	0.00
		0.75	1	0.31	1	0.00	1	0.00
		1.00	1	0.39	1	0.00	1	0.01
		1.25	1	0.45	1	0.06	1	0.02
		1.50	1	0.50	1	0.08	1	0.02
	20% censoring	0.25	1	0.21	1	0.16	1	0.07
		0.50	1	0.24	1	0.01	1	0.00
		0.75	1	0.37	1	0.02	1	0.00
		1.00	1	0.45	1	0.04	1	0.01
		1.25	1	0.52	1	0.11	1	0.01
		1.50	1	0.52	1	0.12	1	0.01

Table 3.4: Average true positive rate (TPR), and false positive rate (FPR) values of three group regularization methods over 100 replications for different coefficient magnitude values and different censoring scenarios.

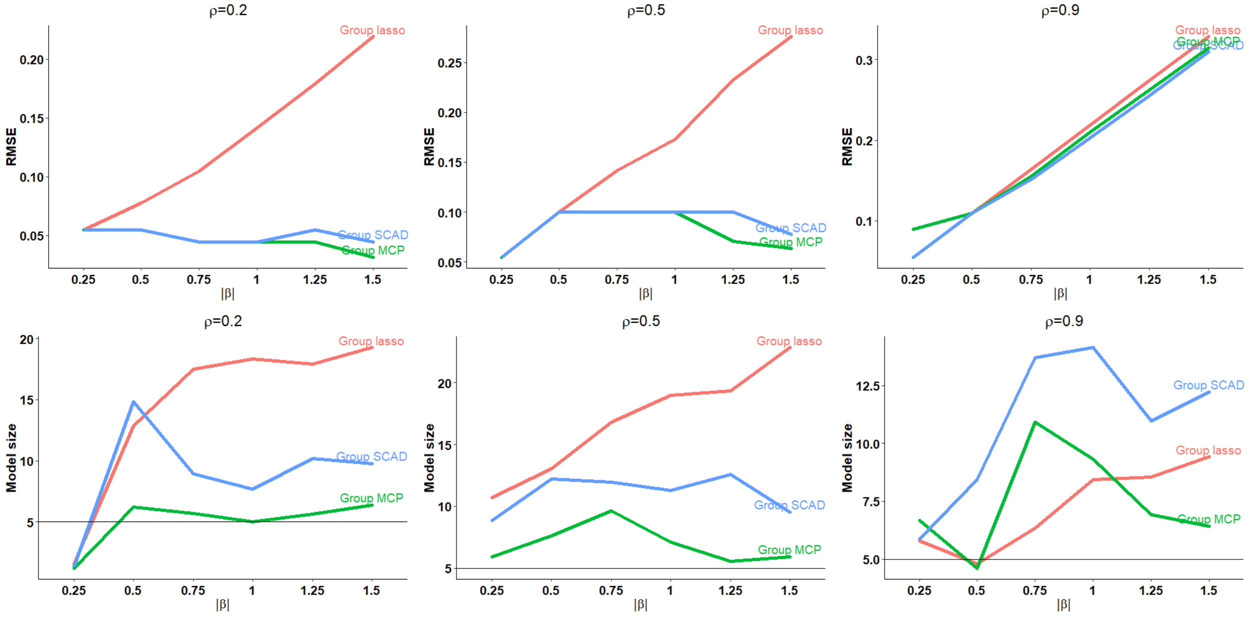
3.4.5 Comparison of three group penalized Cox's models with overlapping groups

In this section, we compare the performance of three group regularization methods in terms of variable selection and model accuracy using simulated data. In here, the model size is the number of nonzero covariates.

3.4.5.1 Equal group size

We generate $N = 50$ observations with $P = 162$ covariates $X_1, \ldots, X_{162}$. There are 20 groups of 10 covariates with two of them overlapping between two successive groups:

Figure 3.6: The impact of increasing the coefficient magnitude and the population correlation on group regularization methods when the size is 4. The *black line* is the true model size (5).



$\{1,\ldots,9,10\},\{9,\ldots,17,18\},\ldots,\{153,\ldots,162\}$. The nonzero covariates are $X_{25},X_{26},\ldots,X_{42}$.

3.4.5.2 Effect of the number of overlapping covariates among groups

We continue considering the setting with the equal group size but set the varying number of overlapping covariates between two successive groups to 3, 4, 5, 6, 7, and 8. The results are shown in Table 3.6. It shows clearly that group SCAD and group MCP select smaller models with smaller RMSE values than group lasso does. In terms of variable selection performances, group SCAD and group MCP produce better results than group lasso. Overall the change of overlap covariates among groups does not affect performances by group SCAD and group MCP. On the other hand, it has strong effect upon group lasso.

3.4.5.3 Unequal group size

We generate $N = 50$ observations with $P = 185$ covariates $X_1,\ldots,X_{185}$. There are 11 groups: 5 groups with 8 covariates per group, 10 groups with 11 covariates per group, and 6 groups with 15 covariates per group. There are two covariates overlapping between two successive groups. The

		Group lasso		Group SCAD		Group MCP	
ρ	$\lvert\beta\rvert$	TPR	FPR	TPR	FPR	TPR	FPR
0	0.25	1	0.10	1	0.00	1	0.00
	0.50	1	0.10	1	0.00	1	0.00
	0.75	1	0.31	1	0.00	1	0.00
	1.00	1	0.39	1	0.00	1	0.01
	1.25	1	0.45	1	0.06	1	0.02
	1.50	1	0.50	1	0.08	1	0.02
0.2	0.25	0.07	0.01	0.04	0.01	0.02	0.01
	0.50	1	0.01	0.99	0.11	0.99	0.02
	0.75	1	0.14	1	0.05	1	0.02
	1.00	1	0.15	1	0.04	1	0.01
	1.25	1	0.14	1	0.06	1	0.02
	1.50	1	0.16	1	0.06	1	0.02
0.5	0.47	1	0.09	0.52	0.07	0.47	0.04
	0.50	0.97	0.10	0.80	0.09	0.78	0.05
	0.75	1	0.13	1	0.08	1	0.06
	1.00	1	0.16	1	0.08	1	0.03
	1.25	1	0.16	1	0.09	1	0.02
	1.50	1	0.20	1	0.06	1	0.02
0.9	0.25	0.25	0.05	0.25	0.05	0.26	0.06
	0.50	0.50	0.03	0.49	0.07	0.27	0.04
	0.75	0.51	0.04	0.50	0.12	0.44	0.09
	1.00	0.53	0.07	0.50	0.13	0.38	0.08
	1.25	0.53	0.07	0.50	0.09	0.38	0.06
	1.50	0.51	0.08	0.51	0.11	0.27	0.06

Table 3.5: Average true positive rate (TPR), and false positive rate (FPR) values over 100 replications for three group regularization models with different coefficient magnitude and population correlation values when the group size is 4.

nonzero covariates are $X_1, X_2, \ldots, X_{14}$.

3.4.5.4 Sparse group example

As we mentioned above, the sparse group selection is a special case of the overlapping group. Here, we provide one example. We generate $N = 50$ observations with $P = 60$ covariates $X_1, \ldots, X_{60}$. Each covariate is treated as a group whose size is 1. In addition, there are 15 groups whose size was 4. The nonzero covariates include $X_1, X_2, X_9, X_{10}, X_{11}, X_{12}, X_{21}$. In other words, out of fifteen 4-covariate groups, there are two groups that have sparse group effects.

No. of overlapping covariates		TPR	FPR	Model size	RMSE
	truth			18	
2	Group lasso	1	0.36	70.58	0.42
	Group SCAD	1	0	18	0.36
	Group MCP	1	0	18	0.36
	truth			17	
3	Group lasso	1	0.53	83.47	0.43
	Group SCAD	1	0	17	0.30
	Group MCP	1	0	17	0.30
	truth			16	
4	Group lasso	1	0.50	70.68	0.47
	Group SCAD	1	0	16	0.29
	Group MCP	1	0.07	23.22	0.31
	truth			15	
5	Group lasso	1	0.62	71.35	0.46
	Group SCAD	1	0.36	47.4	0.23
	Group MCP	1	0.03	17.9	0.20
	truth			14	
6	Group lasso	1	0.63	59.36	0.58
	Group SCAD	1	0.06	19.2	0.26
	Group MCP	1	0.03	16.2	0.23
	truth			13	
7	Group lasso	0.96	0.86	58.76	0.58
	Group SCAD	0.96	0.10	18.24	0.46
	Group MCP	0.90	0.08	15.98	0.40
	truth			12	
8	Group lasso	1	0.81	41.24	0.61
	Group SCAD	1	0.35	24.80	0.35
	Group MCP	1	0.30	21.72	0.29

Table 3.6: Results for overlapping group settings with different overlapping covariates between two successive groups over 100 replications.

		TPR	FPR	Model size	RMSE
Equal group	truth			18	
	Group lasso	1	0.36	70.58	0.42
	Group SCAD	1	0	18	0.36
	Group MCP	1	0	18	0.36
Unequal group	truth			14	
	Group lasso	1	0.43	87.52	0.37
	Group SCAD	1	0.01	16.05	0.25
	Group MCP	1	0	14.55	0.25
Sparse group	truth			7	
	Group lasso	1	0.27	21.61	0.29
	Group SCAD	1	0.05	9.56	0.24
	Group MCP	1	0.02	7.83	0.24

Table 3.7: Results for overlapping group settings over 100 replications.

3.4.5.5 Results

For all three settings above, we consider the population correlation $\rho = 0.5$ with 20% right censoring. We create a path of 50 λ values and use 10-fold cross-validation to select the final model. The results of 100 replications are summarized in Table 3.7. The results in terms of TPR, FPR, model size, and RMSE values are consistent with the results of the non-overlapping group cases presented above: group SCAD and group MCP give better results in term of variable selection and model accuracy.

3.4.6 Misspecification of group structures

As described above, our methods need pre-defined group structures. We would like to investigate the effects of erroneous specification of groups. We consider an example with $N = 100$, $P = 80$, and the "correct" underlying group structure:

$$\underbrace{1,\ldots,10}_{group1}\,\underbrace{11,\ldots,20}_{group2}\,\underbrace{21,\ldots,26}_{group3}\,\underbrace{25,\ldots,30}_{group4}\,\underbrace{31,\ldots,40}_{group5}$$

$$\underbrace{41,\ldots,50}_{group6}\,\underbrace{51,\ldots,57}_{group7}\,\underbrace{55,\ldots,60}_{group8}\,\underbrace{61,\ldots,70}_{group9}\,\underbrace{71,\ldots,80}_{group10},$$

in which there are non-overlapping groups and overlapping groups. Notice that groups 3 and 4 have two overlapped covariates, and groups 7 and 8 have three overlapped covariates. We set the population correlation $\rho = 0.5$ with 50% censoring rate. The corresponding coefficients are

$$\underbrace{0,\dots,0}_{group1-2}\underbrace{1.5,0,1.5,0,-2,0}_{group3}\underbrace{-2,0,0,-2,-1,-2}_{group4}\underbrace{0,\dots,0}_{group5-6}$$
$$\underbrace{1.4,0,1,0,1.8,0,0}_{group7}\underbrace{0,1.8,0,0,1,1.6,1.2}_{group8}\underbrace{0,\dots,0}_{group9-10}.$$

Then we consider two examples with the misspecified groups for inference. In the first example, the number of groups are incorrect because the overlapping groups are collapsed:

$$\underbrace{1,\dots,10}_{group1}\underbrace{11,\dots,20}_{group2}\underbrace{21,\dots,30}_{group3}\underbrace{31,\dots,40}_{group4}\underbrace{41,\dots,50}_{group5}$$
$$\underbrace{51,\dots,60}_{group6}\underbrace{61,\dots,70}_{group7}\underbrace{71,\dots,80}_{group8}.$$

In the second example, there are no overlapping covariates because the overlapping covariates are put into one group.

$$\underbrace{1,\dots,10}_{group1}\underbrace{11,\dots,20}_{group2}\underbrace{21,\dots,26}_{group3}\underbrace{27,\dots,30}_{group4}\underbrace{31,\dots,40}_{group5}$$
$$\underbrace{41,\dots,50}_{group6}\underbrace{51,\dots,57}_{group7}\underbrace{58,\dots,60}_{group8}\underbrace{61,\dots,70}_{group9}\underbrace{71,\dots,80}_{group10}.$$

The results are shown in Table 3.8. It can be seen that our methods are quite robust and not affected by the group structure misspecification.

We consider additional settings with a large number of overlapping covariates and the number of zero groups being more than the number of non-zero groups in Appendix A.3.

		TPR	FPR	Model size	RMSE
Correct specification	truth			12	
	Group lasso	1	0.70	59.8	0.45
	Group SCAD	1	0.34	35.5	0.18
	Group MCP	1	0.17	24.1	0.16
First misspecification	truth			12	
	Group lasso	1	0.70	59.8	0.45
	Group SCAD	1	0.28	31.3	0.17
	Group MCP	1	0.13	21.4	0.15
Second misspecification	truth			12	
	Group lasso	1	0.71	61.8	0.46
	Group SCAD	1	0.30	32.8	0.18
	Group MCP	1	0.17	24.2	0.16

Table 3.8: Results for overlapping group settings over 100 replications.

3.5 Real-world case studies

An important motivation for developing our methods is to perform gene selection for biomarker discovery from gene expression data using the prior knowledge about group structures. We apply our methods to analyze both ovarian cancer and breast cancer data as detailed below. The grouping of genes into predefined gene sets is based on the curated database, MSigDB [39].

3.5.1 Data

The ovarian cancer data are downloaded from The Cancer Genome Atlas (TCGA, `http://cancergenome.nih.gov`). It includes gene expression data for 12,043 genes in 593 samples. We first map gene probes to gene symbols and remove the duplicated genes. We use the 15 KEGG subsets of canonical pathways suggested in [40]. The subsets include apoptosis, cell adhesion molecules, cell cycle, base excision repair, nucleotide excision repair, mismatch repair, non-homologous end joining, Hedgehog signaling pathway, mTOR signaling pathway, Jak-STAT signaling pathway, Notch signaling pathway, Phosphatidylinositol signaling system, MAPK signaling pathway, TGF-beta signaling pathway, and Wnt signaling pathway. These gene sets include 1,347 genes in total. After removing the samples without survival information, 580 samples remain.

We use the breast cancer dataset compiled by [41], which includes gene expression data for 21,463 genes in 295 breast cancer samples. Out of 295 samples there are 216 censoring samples. We first map gene probes to gene symbols and remove the duplicated genes, with the final expression data consisting of 9,950 genes. We use the gene sets from [42] containing 427 gene sets. We restrict the analysis to the 2,663 genes that are in at least one gene set.

3.5.2 Methods

We apply our methods (group lasso, group SCAD, and group MCP) with 5-fold cross validation.

In addition, we run univariate test to select genes and pathways for evaluation. For gene-level analysis, where each gene is tested one at a time, we use the *RegParallel* function of the **RegParallel** package [43] with the embedded *coxph* function of the **survival** package [44] to compute the adjusted p-values for multiple comparisons with multiple FDR and FWER methods (7 methods in total [45, 46, 47, 48, 49]). For pathway-level analysis, where each pathway is tested one at a time, we first convert the gene-level expression data matrix into pathway-level variables using the **GSVA** package [50], then apply the *coxph* function and compute the adjusted p-values. The significance threshold 0.05 is used to select the genes or pathways.

3.5.3 Results and discussion

Analysis of ovarian cancer data: In univariate test, there is no gene or pathway selected using the significance level 0.05, which shows that it is often subjective relying on (adjusted) p-values for biomarker identification depending on univariate tests. This again motivates why we would like to develop our penalized survival model with different group regularization terms to consider candidate covariates together. For comparison purpose, we consider 54 genes selected based on the raw p-values at the significance 0.05 and top four pathways with the smallest p-values. Its results and the results using our methods (group lasso, group SCAD, and group MCP) are summarized in Table 3.9.

First, comparing different models using **grpCox**, the results are consistent with the simulation

Methods		Selected pathways	No. of unique genes No. of genes	No. of selected unique genes
grpCox	Group lasso	KEGG_NON_HOMOLOGOUS_END_JOINING, KEGG_HEDGEHOG_SIGNALING_PATHWAY, KEGG_TGF_BETA_SIGNALING_PATHWAY, KEGG_WNT_SIGNALING_PATHWAY	252/304	208
	Group SCAD	KEGG_NON_HOMOLOGOUS_END_JOINING, KEGG_TGF_BETA_SIGNALING_PATHWAY, KEGG_WNT_SIGNALING_PATHWAY	232/248	194
	Group MCP	KEGG_NON_HOMOLOGOUS_END_JOINING, KEGG_TGF_BETA_SIGNALING_PATHWAY, KEGG_WNT_SIGNALING_PATHWAY	232/248	194
Univariate test	Gene-level	-	1098/1347	54
	Pathway-level	KEGG_BASE_EXCISION_REPAIR, KEGG_TGF_BETA_SIGNALING_PATHWAY, KEGG_WNT_SIGNALING_PATHWAY, KEGG_MISMATCH_REPAIR	271/293	271

Table 3.9: Pathways and genes selected by different methods for ovarian cancer data.

results when group lasso selects a relatively larger model than group SCAD and group MCP do.

Second, we compare the results of univariate tests and our **grpCox** package using the group lasso penalty since the results selected by group lasso include all the selections using group SCAD and MCP. At gene-level, among 15 overlapping selected genes, there are 6 genes have been reported in the literature as ovarian cancer biomarkers. Among non-overlapping genes identified by our **grpCox**, there are 38 genes showing biologically meaning. In contrast, among 54 genes by univariate tests, there are only additional 6 genes showing biological relevance.

At pathway-level, all selected pathways using our methods are biologically meaningful. The identified pathways appear to be biologically meaningful in ovarian cancer. Non homologous end joining (NHEJ) pathway is known to repair double strand breaks. Defective NHEJ has been found in up to 50% of ovarian cancers [51, 52]. Overexpression or pathway activation by gene mutations among genes of the Hedgehog signaling in ovarian tumorigenesis play the crucial role in the development and progression of ovarian cancer [53, 54]. Wnt signaling pathway is well-known to play a role in tumorigenesis. [55] demonstrated the difference in Wnt signaling pathway between normal ovarian and cancer cell lines. They also pointed out that those differences implicate that Wnt signaling leads to ovarian cancer development despite the fact that gene mutations are uncommon. TGF-β signaling pathway behaves as both a tumor suppressor in ovarian physiology as well as acting as a tumor promoter that controls proliferation in ovarian cancer [56, 57]. Two other pathways selected by univariate test are also biologically meaningful. It is clear again that considering genes together can help understand underlying cellular processes. However, the results in Table 3.9 show that when univariate test selects pathway, it selects all genes in this pathway, which is less flexible compared to the group penalized survival models. In fact, **grpCox** naturally takes care of the gene-pathway relationships in the model formulations and results in simultaneous selection of relevant genes and pathways. In other words, **grpCox** jointly considers potential effects, which may lead to better biomarker identification results.

Analysis of breast cancer data: Similarly, in univariate test results, very few genes, either one or five genes depending on the adopted multiple testing adjustment method, are selected. Five genes are selected with the significance level 0.05 based on the FDR and Benjamini-Hochberg correction. There are 293 pathways out of 427 pathways are selected. Its results and the results using our methods (group lasso, group SCAD, and group MCP) are summarized in Table 3.10. More details about genes and pathways selected by univariate tests and our methods are provided in Tables 6, 7, 8, 9 in the Supporting Information.

Similarly, the results of different models using **grpCox** are consistent with the simulation results when group lasso selects a relatively larger model than group SCAD and group MCP do.

Next, we compare the results of univariate tests and **grpCox** package. At gene-level, among three overlapping selected genes, *TBCB* gene has been reported as breast cancer biomarker. Among non-overlapping genes using **grpCox**, there are 33 genes showing biological relevance. Among non-overlapping genes by univariate test, the other selected genes have not been reported to be relevant to breast cancer specifically.

At pathway-level, there are three overlapping pathways in which GCM_ATM and GCM_PPP1CC pathways all being biologically relevant. For example, the gene *ATM* in the GCM_ATM pathway associated with increased breast cancer risk [58, 59]. In addition, the gene *CDH11* in the GNF2_CDH11 pathway has been found to be overexpressed in breast cancer [60, 61, 62]. The collagen genes *COL1A2, COL3A1, COL6A1* are correlated significantly during breast cancer development and progression [63, 64, 65, 66, 67, 68].

All non-overlapping pathways using **grpCox** are biologically meaningful. Among 290 non-overlapping pathways using univariate test, consider top 6 pathways with smallest adjusted p-values, there are five among them showing biological relevance. However, univariate test at pathway-level again shows less flexible when selecting relevant genes than **grpCox**.

Validation of results: The results that are selected by our methods are further analyzed.

For the ovarian cancer data, we use the independent dataset described in [69] as a test set. This dataset contains 285 samples and 53,433 genes. After removing the samples without survival information, there are 276 samples in total. We first compute the estimated coefficients $\hat{\beta}$, and the risk scores $X\hat{\beta}$. Their median value is used as the threshold for the high and low risk groups.

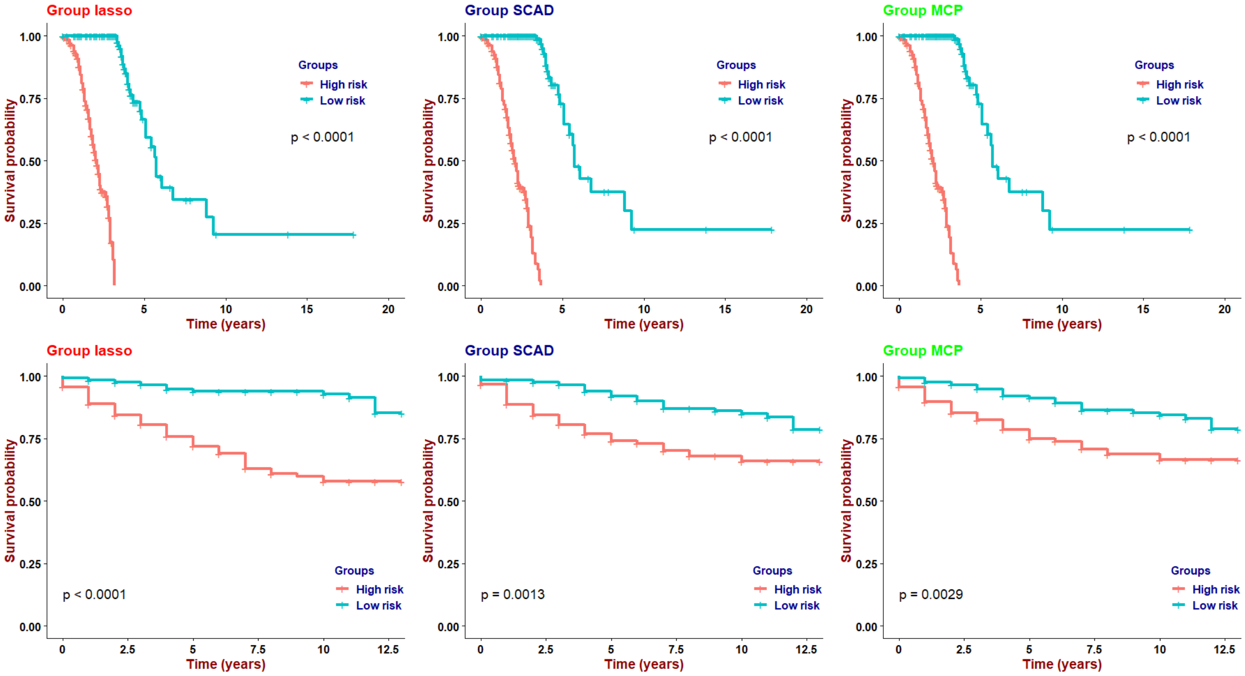
Methods		No. of selected pathways	No. of unique genes No. of genes	No. of selected unique genes
grpCox	Group lasso	GCM_ATM, GCM_BCL2L1, GNF2_CDH11, GNF2_CEBPA, GCM_PPP1CC, GNF2_PTX3, GNF2_TPT1, GNF2_GLTSCR2, GNF2_CYP2B6	289/361	151
	Group SCAD	GCM_ATM, GNF2_CDH11, GNF2_TPT1	90/90	43
	Group MCP	GNF2_CDH11	25/25	20
Univariate test	Gene-level	-	2663/42526	5
	Pathway-level	293	2197/35517	2197

Table 3.10: Pathways and genes selected by different methods for breast cancer data. Note that 293 selected pathways using univariate tests are listed in the Supporting Information.

The samples are assigned into the high and low risk groups by comparing with the threshold. The survival curves of these two groups are shown in Figure 3.7. These two curves of all methods are well separated with a $p-$value of the log-rank test is smaller than 0.0001.

For the breast cancer data, we use the independent dataset described in [70] as a test set. This dataset contains 251 samples and 24,712 genes. After removing the samples without survival information and selecting genes appearing in the selected genes in Table 3.10, there are 236 samples with 181 censoring samples. We first compute the estimated coefficients $\hat{\beta}$, and the risk scores $X\hat{\beta}$. Their median value is used as the threshold for the high and low risk groups. The samples are assigned into the high and low risk groups by comparing with the threshold. The survival curves of these two groups are shown in Figure 3.7. It shows that the $p-$values of the log-rank tests for three models are much smaller than 0.01: the $p-$value of the group lasso is the smallest, followed by the group SCAD and the group MCP. In other words, the selected genes sets of group SCAD and group MCP are much smaller than the selected genes set of group lasso, and still classify the patients in independent breast cancer dataset into high risk and low risk groups well.

Figure 3.7: Survival curves for the high and low risk groups of the independent testing samples of ovarian cancer (first row) and breast cancer (second row).



4. L1-REGULARIZED MULTI-STATE MODELS [1]

Multi-state model (MSM) is a useful tool to analyze longitudinal data for modeling disease progression at multiple time points. While the regularization approaches to variable selection have been widely used, extending them to MSM remains largely unexplored. In this chapter, we have developed the L1-regularized multi-state model (L1MSTATE) framework that enables parameter estimation and variable selection simultaneously. The regularized optimization problem was solved by deriving a one-step coordinate descent algorithm with great computational efficiency. The L1MSTATE approach was evaluated using extensive simulation studies, and it showed that L1MSTATE outperformed existing regularized multi-state models in terms of the accurate identification of risk factors. It also outperformed the un-regularized multi-state models (MSTATE) in terms of identifying the important risk factors in situations with small sample sizes. The power of L1MSTATE in predicting the transition probabilities comparing with MSTATE was demonstrated using the Europe Blood and Marrow Transplantation (EBMT) dataset. The L1MSTATE was implemented in the open-access R package **L1mstate**.

4.1 Introduction

Multi-state model (MSM) has been one of effective methods for disease modeling, and it has been applied to studying liver cancer [71], breast cancer [72][73][74], abdominal aortic aneurysms [75], heart transplantation [76][77], HIV infection and AIDS [78][79], Alzheimer disease [80], diabetic complication [81][82], cervical cancer [83], and liver cirrhosis [84], just to name a few. It can model patient's disease development trajectory across a series of transitions between various stages or states, under influence of some risk factors. First, it allows researchers to make an assessment about how the risk factors exert different effects on different stages of the process and how the risk factors influence on different transitions of the process. Second, it enables researchers to

[1] Reprinted with permission from X. Dang, S. Huang, X. Qian, “Risk Factor Identification in Heterogeneous Disease Progression with L1-Regularized Multi-State Models", *Journal of Healthcare Informatics Research*, vol. 5, no. 1, pp. 20–53, 2021. Copyright 2021 by Springer Nature.

obtain more accurate predictions of transition probabilities.

In this chapter, we adopted the MSM framework by specifying the transition-specific hazard models. Our main objective is to identify the risk factors associated with the transition hazard rates of disease progression. Although non-parametric transition hazard models do not impose any constraint and may be more flexible, it is used more often to estimate the cumulative transition hazard rates than the transition hazard rates [85]. Semi-parametric transition hazard models that do not require to specify the transition-specific baseline hazard functions are more suitable for our purpose. Specifically, the Cox's proportional hazards model was used for the transition-specific hazard rates to incorporate risk factors into multi-state models. The multi-state model parameters were estimated by maximizing the likelihood function that was formulated using the counting process [86]. The transition-specific baseline hazards were assumed to be the same for all individuals but vary over time, allowing us to construct the partial likelihood function that reduces computation burden but still makes good estimations of parameters [87]. Regarding the censored data, we focused on two types of censoring data: right-censored and left-truncation data.

Currently, the multistate models lack an efficient and practical variable selection method to identify the risk factors associated with the transition hazard rates. Let us consider a MSM with the number of the risk factors is P and the number of transitions between the stages is Q. Then, there are 2^{PQ} possible models to consider if using stepwise forward selection [81] method. Hence, such kinds of variable selection methods are suitable when the number of risk factors and the number of transitions are relatively small. However, in modern applications, both P and Q increase dramatically with our increasing data collection capacity. They result in complicated optimization problems which are challenging to compute, and they can lead unstable estimates of parameters. In addition, in many studies, especially in medical research, there is a limited number of observations given the number of parameters in complex multi-state models. In this chapter, the regularization approaches have been used to address these challenges. Intuitively, these approaches incorporate the prior knowledge about sparse structures of multi-state models using the sparse-inducing penalties, which results in better parameter estimations and allows variable selection simultaneously.

Even though the regularization methods are increasingly popular in statistics and machine learning, very little has extended to MSMs. The current literature on this subject shows there are two works that have been published in this direction. The first one by Huang *et al.* 2018 [88] presented a regularized continuous-time Markov model with the elastic net penalty. The transition hazard rates were specified as constant over time. In addition, their method relied on a method developed by [89]: it estimated the transition rates from the transition probabilities of the discrete-time Markov chain embedded in the Markov process (embedded Markov chain). It does not derive the transition rates from event (state) counts and transitions since the transition times are not observed. In other words, it does not follow the counting process perspective. Therefore, their work is different from ours in scope and methodology.

The second one from Reulen *et al.* 2016 [90] did variable selection by imposing the fused-lasso penalties including L1-penalties of transition-specific risk factor coefficients and their differences between transitions. In this chapter, we propose the L1-penalties of transition-specific risk factor coefficients that are similar to the fused-lasso approach in [90], in which cross-transition effects are explicitly modeled by introducing the fused penalties. The difference of our implementation from [90] is, instead of adopting the penalized iteratively re-weighted least squares (PIRLS) algorithm presented in Oelker *et al.* 2017 [91] for model inference, we have derived a cyclical one-step coordinate descent algorithm to solve the optimization problem with exact L1-penalties. In addition to potential problems of not having exact zero model coefficients due to the approximation of L1-penalties, PIRLS is a second-order optimization algorithm that has high computation cost and potential convergence problems [91]. Our optimization algorithm in this chapter solves for exactly L1-penalties resulting in fewer nonzero coefficients for variable selection, with high efficiency in computation and significant reduction in memory usage.

Another common problem in many studies is that multi-state models include some *rare* transitions that have relatively small number of observations. In such cases, the traditional (unregularized) multi-state model approach tends to produce the inaccurate predictions of the probabilities of *rare* transitions. In this chapter, we demonstrated that the L1-regularized multi-state models can be used to alleviate this problem, and thus produce better predictions of the transition probabilities.

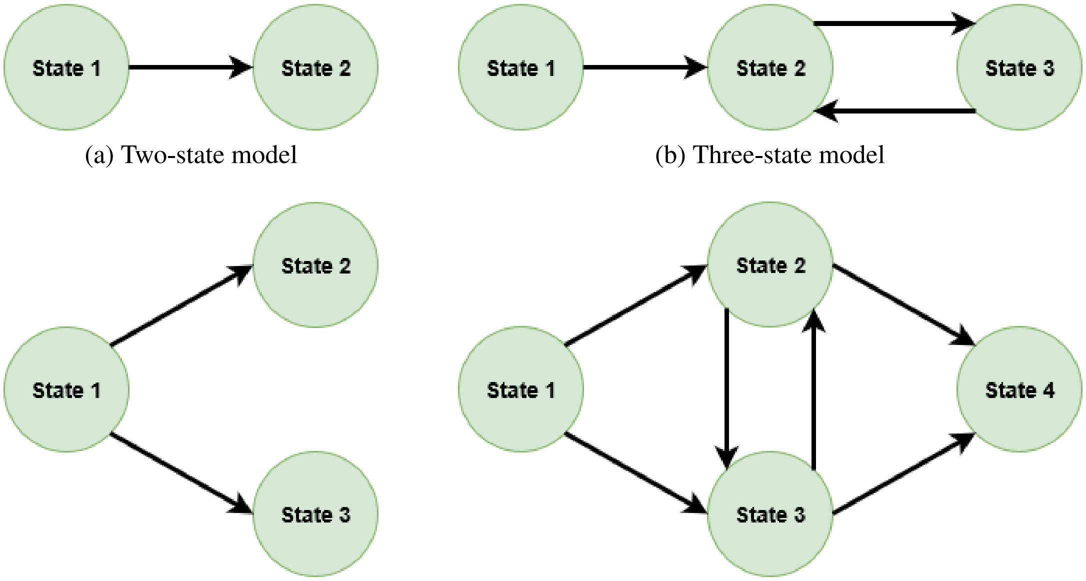
The rest of this chapter is organized as follows. In Section 2, we reviewed critical details of the multi-state models, including its formulation and the partial likelihood function of the multi-state models. In Section 3, we introduced our formulation of the L1-regularized partial likelihood function of the multi-state models and the algorithms to solve the corresponding optimization problems. We presented the main formulae to predict the transition probabilities. In Section 4, we compared the performance of our method via simulation studies. We demonstrated the prediction power of our method using a real data. Discussion was presented in Section 6. Lastly, we ended with conclusions and future works in Section 7.

4.2 Review of Multi-State Models (MSMs)

4.3 Formulations of the multi-state models

Multi-state models compose of multiple states and transitions between the states under influence of risk factors. Figure 4.1 depicts some examples of the multi-state models in characterizing a variety of situations with different number of states and transition structures between the states. For example, in Figure 4.1.c, there are three states. The arrows illustrate the clinically eligible transitions between the states. The state to which the individual is going to move, and the time of this change, is impacted by the transition intensities (so-called hazard rates) that represent the instantaneous risk of moving from one state to another. These hazard rates may also depend on individual-specific risk factors. In this chapter, we assume that the risk factors are constant over time. The states and structure of the transitions are usually pre-defined based on domain knowledge of the disease. The main statistical task is to estimate the transition intensities between states and their relationships with the risk factors.

We formulate these hazard rates and the transition probabilities using the basic concepts in Andersen *et al.* 2002 [92]. The multi-state model is a multi-state process $S(t)$ - a stochastic process $S(t)$, $t \in \mathbf{T}$ with a finite state space $\chi = \{1, 2, \ldots, s\}$ and with right-continuous sample



(a) Two-state model

(b) Three-state model

(c) Three-state model

(d) Four-state model

Figure 4.1: Some multi-state models. *Note*: Arrows show the clinically eligible transitions for each multi-state model.

path $S(t+) = S(t)$ where $t+$ is the limit from the right to t, i.e., the time point immediately after t. The transition probabilities may be defined by

$$P_{hj}(t, t+\Delta t) = Pr(S(t+\Delta t) = j|S(t) = h, S_{t^-}),$$

where $h, j \in \chi, t \in \mathbf{T}, \Delta t \geq 0$, and S_{t^-} is the history generated by the multi-state process $S(t)$. The Markov assumption is commonly used to define the transition intensities

$$\alpha_{hj}(t) = \lim_{\Delta t \to 0} \frac{P_{hj}(t, t+\Delta t)}{\Delta t},$$

We specify the transition-specific hazard rates $\alpha_{hj}(t)$ using Cox proportional hazards model [3] with the transition-specific baseline hazard rates $\alpha_{hj}^{(0)}(t)$ and time-fixed risk factors X:

$$\alpha_{hj}(t) = \alpha_{hj}^{(0)}(t)\text{exp}(\beta_{hj}^{T} X). \tag{4.1}$$

where $X = (x_1, x_2, \cdots, x_P)^T$ is an $P-$dimensional vector of time-fixed risk factors and β_{hj}^T is an $P-$dimensional vector of time-fixed coefficients.

4.3.1 Likelihood function of the multi-state models

Then, we can derive the likelihood formulation of the multi-state model. Consider M individuals, $S_i(t)$ is the observed multi-state model for the i^{th} individual over interval $[0, \tau_i]$, where τ_i is a fixed time of termination of observation for individual i. Denote $N_{hj}^i(t)$ be the number of allowed transitions $h \to j$ of the i^{th} individual during $[0, t]$, and $\alpha_{hj}^i(t)$ be transition intensities or transition-specific hazard rates of the i^{th} individual. The transition times T_{hj}^{ik} can be described as $0 < T_{hj}^{i1} < \cdots < T_{hj}^{iN_{hj}^i(\tau_i)} \leq \tau_i$, where $k \in \{1, \ldots, N_{hj}^i(\tau_i)\}$. The full likelihood function could be derived as

$$L = \prod_{i=1}^{M} \prod_{j \neq h} \prod_{k=1}^{N_{hj}^i(\tau_i)} \left[\alpha_{hj}^i(T_{hj}^{ik}) \text{exp}\left(- \int_0^{T_{hj}^{ik}} \alpha_{hj}^i(t) dt \right) \right],$$

Assume that individual-specific risk factors are constant over time, the transition-specific hazard rates $\alpha_{hj}^i(t)$ for each individual i can be written as Eq. (4.1). The full likelihood function becomes

$$L(\beta) = \prod_{i=1}^{M} \prod_{j \neq h} \prod_{k=1}^{N_{hj}^i(\tau_i)} \left[\alpha_{hj}^{i(0)}(t) \text{exp}(\beta_{hj}^T X^i) \text{exp}\left(- \int_0^{T_{hj}^{ik}} \alpha_{hj}^{i(0)}(t) \text{exp}(\beta_{hj}^T X^i) dt \right) \right]. \tag{4.2}$$

where $X^i = (x_1^i, x_2^i, \cdots, x_P^i)^T$ is an $P-$dimensional vector of time-constant risk factors for the i^{th} individual.

4.3.2 Partial likelihood function for multi-state model

Instead of using the above full likelihood function, we used the partial likelihood function. More details can be found in Andersen *et al.* 1993 [86]. It only keeps the terms that contain all the information about β and gets rid of the terms that contain the information about the baseline hazard. This achieves computational efficiency and still makes good inference for β.

Let $Y_{hj}^{ik}(t) = \mathbf{1}_{\{t \leq T_{hj}^{ik}\}}$, i.e., in this definition $Y_{hj}^{ik}(T_{hj}^{ik})$ indicates that the i^{th} individual at risk in transition from state h to state j at time T_{hj}^{ik}. Assume that the transition-specific baseline hazards are the same for all individuals but can vary freely with time, i.e., $\alpha_{hj}^{i(0)}(t) = \alpha_{hj}^{(0)}$. The partial likelihood function of the multi-state model that will be used in this chapter

$$L^p(\beta) = \prod_{j \neq h} \prod_{i=1}^{M} \prod_{k=1}^{N_{hj}^i(\tau_i)} \frac{\exp(\beta_{hj}^T X^i)}{\sum_{i=1}^{M} \sum_{k=1}^{N_{hj}^i(\tau_i)} \exp(\beta_{hj}^T X^i) Y_{hj}^{ik}(t)},$$

Its negative log-partial likelihood function is derived as

$$l(\beta) = -\log(L^p(\beta)) = -\sum_{j \neq h} \sum_{i=1}^{M} \sum_{k=1}^{N_{hj}^i(\tau_i)} \left[(\beta_{hj}^T X^i) - \log\left(\sum_{i=1}^{M} \sum_{k=1}^{N_{hj}^i(\tau_i)} \exp(\beta_{hj}^T X^i) Y_{hj}^{ik}(t) \right) \right]. \tag{4.3}$$

4.3.3 Data structure for parameter estimation by partial likelihood maximization

We follow the data structure described in [93]. One example as shown in Table 4.1 was collected in [94]. In this format, each individual has many rows. Each row shows one transition of each individual that is composed by state$_{from}$ and state$_{to}$. The corresponding times for state$_{from}$ and state$_{to}$ are time$_{start}$ and time$_{stop}$. The difference between time$_{start}$ and time$_{stop}$ measures the transition times that represent the duration for which individual is at risk. The censoring information is captured by a transition-specific censoring indicator δ_{status}. For example, patient 1 contributes two lines of data for the period: start at $t = 0$ and stop at $t = 151$. She/he started at state 2 and was at risk to transfer to state 1 and state 3. The recorded status of transition $2 \rightarrow 1$ was 0, which indicates that the event (transition) time was censored, while the recorded status of transition $2 \rightarrow 3$ was 1, which indicates that the event time was observed.

Following this data structure, suppose that there are in total Q observable transition types. Assume that the dataset has N rows, and denote N_q be the number of rows for transition q, it is easy to see that $N = \sum_q N_q$. With a slight abuse of notation, X_q is the $N_q \times P$ risk factors matrix corresponding to $q-$transition; X_q^i is the $P-$dimensional column-vector where $q = 1, 2, \ldots, Q$ and $i = 1, 2, \ldots, N_q$. The formulation of the negative log-partial likelihood function in Eq. (4.3)

Table 4.1: Example of long-format data

	Patient id	state$_{from}$	state$_{to}$	transition	time$_{start}$	time$_{stop}$	δ_{status}	treatment
	1	2	1	3	0	151	0	Placebo
2	1	2	3	4	0	151	1	Placebo
3	2	2	1	3	0	251	1	Placebo
4	2	2	3	4	0	251	0	Placebo
5	2	1	2	1	251	434	1	Placebo
6	2	1	3	2	251	434	0	Placebo
7	2	2	1	3	434	729	1	Placebo
8	2	2	3	4	434	729	0	Placebo
9	2	1	2	1	729	1735	1	Placebo
10	2	1	3	2	729	1735	0	Placebo
11	2	2	1	3	1735	2088	1	Placebo
12	2	2	3	4	1735	2088	0	Placebo
13	2	1	2	1	2088	2467	0	Placebo
14	2	1	3	2	2088	2467	1	Placebo

could be rewritten as

$$l(\beta) = \sum_q l_q(\beta_q), \tag{4.4}$$

where

$$l_q(\beta_q) = -\sum_{i=1}^{D_q}\left[(\beta_q^T X_q^i) - \log\Big(\sum_{n=1}^{N_q}\exp(\beta_q^T X_q^i)Y_q^n(t_i)\Big)\right] = -\sum_{i=1}^{D_q}\left[(\beta_q^T X_q^i) - \log\Big(\sum_{r\in R_q^i}\exp(\beta_q^T X_q^r)\Big)\right]. \tag{4.5}$$

where D_q is the set of indices of the exact transition times for the transition type q, $Y_q^n(t_i) = 1_{\{t_q^n \geq t_i\}}$ indicates whether the n^{th} individual is at risk to transition q just before time t_i, and $R_q^i = \sum_n Y_q^n(t_i) = \sum_n 1_{\{t_q^n \geq t_i\}}$ is a set of indices r that comprised of all individuals observed to be at risk to transition q with times $\geq t_i$.

Remark: As shown in above, we use only information about the observed states at a set of times when we assume that the distribution of transition times provides no information about the distribution of censorship times and vice versa. It is so-called the independent censoring [86]. We also assume that the observation time is the exact transition time and there is no transitions between the observation times for each individual. With the formulation of the negative log-partial-likelihood function in Eq. (5.2), two kinds of incomplete observations are particularly tractable [92]: right-censoring and left-truncation. Note that if the individual is observed from the beginning (i.e., the first state, such as healthy) to the end (i.e., the final state, such as death), then the whole trajectory of the process has been observed and it is called complete observation. Otherwise, right-censoring means that the individual is observed from the beginning to a certain time that has not reached the final state. Left-truncation means that the process has not been observed from the beginning, rather, the observation happens in the middle of the trajectory of the transitions.

4.4 L1-Regularized Multi-State Model (L1MSTATE)

4.4.1 Partial likelihood formulation for L1MSTATE

By minimizing the negative log-partial likelihood formulated in Eq. (5.2), we can estimate the parameters of a multi-state model, i.e., the coefficients β. As existing methods could not scale up to high-dimensional applications when there are a large number of risk factors and a large number of transitions, in this chapter, we propose a L1-regularized partial likelihood formulation for MSM following the framework as the least absolute shrinkage and selection operator (LASSO) [13]. This leads to the following formulation:

$$\begin{aligned} \min_{\beta} \quad & l(\beta) \\ \text{subject to} \quad & \sum_{q}\sum_{p} |\beta_q^p| \leq C, \end{aligned} \tag{4.6}$$

where $q = 1, 2, \dots, Q$; $p = 1, 2, \dots, P$; $C > 0$. Recall that, Q is the number of observable transitions, and P is the number of risk factors. This minimization problem is equivalent to minimizing the problem given by the Lagrangian formulation:

$$\sum_{q} \frac{1}{N_q} l_q(\beta_q) + \lambda\big(\sum_{q}\sum_{p} |\beta_q^p|\big),$$

with respect to β. Different weights are assigned to transitions using factors $\frac{1}{N_q}$, where $q = 1, 2, \dots, Q$. It is similar to assign different shrinkage parameters per transition. Intuitively, the rare transitions are shrunk more than for common transitions. Our formulation in Eq. (5.4) could be reformulated as

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \left[\sum_q \frac{1}{N_q} l_q(\beta_q) + \lambda \big(\sum_q \sum_p |\beta_q^p| \big) \right]. \tag{4.7}$$

4.4.2 Computational algorithm for solving Eq. (5.5)

The transition-specific negative log-partial-likelihood function $l_q(\beta_q)$ is smooth with respect to β_q so that its first two partial derivatives are continuous. Thus, $l_q(\beta_q)$ can be locally approximated by

$$l_q(\beta_q) \approx l_q(\tilde{\beta}_q) + (\beta_q - \tilde{\beta}_q)^T l_q^{'}(\tilde{\beta}_q) + \frac{1}{2}(\beta_q - \tilde{\beta}_q)^T l_q^{''}(\tilde{\beta}_q)(\beta_q - \tilde{\beta}_q), \tag{4.8}$$

where

$$l_q^{'}(\tilde{\beta}_q) = \frac{\partial l_q}{\partial \beta_q}(\tilde{\beta}_q) \text{ and } l_q^{''}(\tilde{\beta}_q) = \frac{\partial^2 l_q}{\partial \beta_q \partial \beta_q^T}(\tilde{\beta}_q),$$

The transition-specific linear predictor, $\eta_q = X_q \beta_q$, includes D_q elements $\eta_q^i = \beta_q^T X_q^i$, where $i = 1, \ldots, D_q$. Plugging them in Eq. (5.3) and Eq. (4.8), we have the transition-specific negative log-partial likelihood function

$$l_q(\eta_q) = -\sum_{i=1}^{D_q} \left[\eta_q^i - \log\Big(\sum_{r \in R_q^i} \exp(\eta_q^r) \Big) \right],$$

Its approximated form is

$$l_q(\eta_q) \approx \frac{1}{2}\Big(\eta_q - z(\tilde{\eta}_q)\Big)^T l_q^{''}(\tilde{\eta}_q) \Big(\eta_q - z(\tilde{\eta}_q)\Big),$$

with

$$z(\tilde{\eta}_q) = \tilde{\eta}_q - \Big(l_q^{''}(\tilde{\eta}_q)\Big)^{-1} l_q^{'}(\tilde{\eta}_q); \; l_q^{'}(\tilde{\eta}_q) = \frac{\partial l_q}{\partial \eta_q}(\tilde{\eta}_q); \; l_q^{''}(\tilde{\eta}_q) = \frac{\partial^2 l_q}{\partial \eta_q \partial \eta_q^T}(\tilde{\eta}_q),$$

Hastie and Tibshirani (1990, Chapter 8) [95] suggested to replace $l_q^{''}(\tilde{\eta}_q)$ by a diagonal matrix **D** with the diagonal elements of $l_q^{''}(\tilde{\eta}_q)$, because the optimal β_q will not change when the off-diagonal elements of $l_q^{''}(\tilde{\eta}_q)$ are smaller than the diagonal elements. This will greatly alleviate our analytic efforts since we only need to compute the first order derivative $l_q^{'}(\tilde{\eta}_q)$ and the diagonal entry of the second order derivative $l_q^{''}(\tilde{\eta}_q)$. $l_q^{'}(\tilde{\eta}_q)$ is a vector with elements $\big(l_q^{'}(\tilde{\eta}_q)_d\big)$ that could be derived as

$$l_q^{'}(\tilde{\eta}_q)_d = \frac{\partial l_q(\eta_q)}{\partial \eta_q^d} = -\delta_d + \sum_{i=1}^{D_q} \frac{\sum_{d \in R_q^i} \exp(\eta_q^d)}{\sum_{r \in R_q^i} \exp(\eta_q^r)} = -\delta_d + \sum_{i=1}^{D_q} \sum_{d \in R_q^i} \frac{\exp(\eta_q^d)}{\sum_{r \in R_q^i} \exp(\eta_q^r)} = -\delta_d + \sum_{i \in C_q^d} \frac{\exp(\eta_q^d)}{\sum_{r \in R_q^i} \exp(\eta_q^r)}, \tag{4.9}$$

where $d = 1, 2, \ldots, N_q$, and C_q^d is the q-transition set of i with $t_d \geq t_i$. The diagonal entry of $l_q^{''}(\eta_q)$ could be derived as

$$l_q^{''}(\tilde{\eta}_q)_{d,d} = \frac{\partial}{\partial \eta_q^d}\Big(\frac{\partial l_q(\eta_q)}{\partial \eta_q^d}\Big) = \sum_{i \in C_q^d} \left[\frac{\exp(\eta_q^d)}{\sum_{r \in R_q^i} \exp(\eta_q^r)} - \frac{(\exp(\eta_q^d))^2}{\big(\sum_{r \in R_q^i} \exp(\eta_q^r)\big)^2} \right], \tag{4.10}$$

Let

$$M(\beta_q) = \frac{1}{2N_q}\Big(\eta_q - z(\tilde{\eta}_q)\Big)^T l_q^{''}(\tilde{\eta}_q)\Big(\eta_q - z(\tilde{\eta}_q)\Big) + \lambda\big(\sum_p |\beta_q^p|\big),$$

The training algorithm for L1MSTATE is shown in the pseudo code below:

Algorithm 3 Pseudocode for L1-penalized multi-state model.

```
Result: $\hat{\beta}$
Data: Long-format data described in Section 4.3.3
while ($q > 0$ and $q \leq Q$) do
    Compute $\tilde{\boldsymbol{\eta}}_q = \mathbf{X}_q\tilde{\boldsymbol{\beta}}_q$; $l_q^{'}(\tilde{\boldsymbol{\eta}}_q)$  $l_q^{''}(\tilde{\boldsymbol{\eta}}_q)$  $\mathbf{z}(\tilde{\boldsymbol{\eta}}_q) = \tilde{\boldsymbol{\eta}}_q - l_q^{''}(\tilde{\boldsymbol{\eta}}_q)^{-1} l_q^{'}(\tilde{\boldsymbol{\eta}}_q)$
    Find $\hat{\boldsymbol{\beta}}_q = \underset{\boldsymbol{\beta}_q}{\text{argmin}}\ M(\boldsymbol{\beta}_q)$  Update $\tilde{\boldsymbol{\beta}}_q = \hat{\boldsymbol{\beta}}_q$
end
```

The remaining task is to solve the optimization problem in Eq. (4.11):

$$\hat{\beta}_q = \underset{\boldsymbol{\beta}_q}{\text{argmin}}\ M(\beta_q), \tag{4.11}$$

Let w_q be the N_q-dimensional vector of diagonal entries of matrix $\mathbf{D}$. We rewrite $M(\beta_q)$ as

$$M(\beta_q) = \frac{1}{2N_q}\sum_{i=1}^{N_q}\Big[w_q^i\Big(z(\tilde{\eta_q})_i - \sum_{p\neq g} X^i_{q,p}\beta_q^p - X^i_{q,g}\beta_q^g\Big)^2\Big] + \lambda\big(\sum_p |\beta_q^p|\big),$$

Hence, Eq. (4.11) becomes

$$\hat{\boldsymbol{\beta}}_q = \underset{\boldsymbol{\beta}_q}{\text{argmin}}\, \frac{1}{2N_q}\sum_{i=1}^{N_q}\Big[w_q^i\Big(z(\tilde{\eta_q})_i - \sum_{p\neq g} X^i_{q,p}\beta_q^p - X^i_{q,g}\beta_q^g\Big)^2\Big] + \lambda\big(\sum_p |\beta_q^p|\big), \tag{4.12}$$

The coordinate descent algorithm is used to solve Eq. (4.12). In particular, we derive the one-step coordinate descent algorithm that updates one element at each iteration with all the other elements fixed to the latest value. Specifically, for instance, while the current step focuses on β_q^g with given estimates for β_q^p for all $p \neq g$, we compute the first order derivative of $M(\beta_q)$ as follows

$$\frac{\partial M(\beta_q)}{\partial \beta_q^g} = \frac{1}{N_q}\sum_{i=1}^{N_q}\Big[w_q^i\Big(z(\tilde{\eta_q})_i - X_q^i\beta_q\Big)(-X^i_{q,g})\Big] + \lambda \text{sgn}(\beta_q^g), \tag{4.13}$$

where with $g = 1, 2, \ldots, P$

$$\text{sgn}(\beta_q^g) = \begin{cases} 1, & \text{if } \beta_q^g > 0 \\ -1, & \text{if } \beta_q^g < 0 \\ [-1, 1]\,, & \text{otherwise.} \end{cases}$$

Solving Eq. (4.13) yields the soft-thresholding rule that is

$$\hat{\beta_q^g} = \frac{f\left(\frac{1}{N_q}\sum_{i=1}^{N_q}\Big[w_q^i X^i_{q,g}\Big(z(\tilde{\eta_q})_i - \sum_{p\neq g} X^i_{q,p}\beta_q^p\Big)\Big], \lambda\right)}{\frac{1}{N_q}\sum_{i=1}^{N_q} w_q(\tilde{\eta_q})_i (X^i_{q,g})^2}, \tag{4.14}$$

where

$$f(x,\lambda) = \text{sgn}(x)(|x|-\lambda) = \begin{cases} x-\lambda, & \text{if } x>0 \text{ and } |x|>\lambda \\ x+\lambda, & \text{if } x<0 \text{ and } |x|>\lambda \\ 0, & \text{if } |x|\le\lambda. \end{cases}$$

Note that the first term in the numerator can be derived by using Eqs. (4.9) and (4.10):

$$w_q^i X_{q,g}^i \Big(\mathbf{z}(\tilde{\eta}_q)_i - \sum_{p\neq g} X_{q,p}^i \beta_q^p\Big) = \tilde{\beta}_q^g w_q (X_{q,g})^2 - l_q^{'}(\tilde{\eta}_q) X_{q,g},$$

So, we have a simple form of estimated coefficient as follows

$$\hat{\beta}_q^g = \frac{f\left(\frac{1}{N_q}\Big[\tilde{\beta}_q^g w_q (X_{q,g})^2 - l_q^{'}(\tilde{\eta}_q) X_{q,g}\Big], \lambda\right)}{\frac{1}{N_q} w_q (X_{q,g})^2}. \tag{4.15}$$

It is worthy of mentioning that the solution for LASSO depends on the scales of risk factors [96]. A frequently used method to solve this problem is to standardize the risk factors first. The estimated coefficients of the risk factors can always be transformed back to the original scales for the sake of interpretation. The one-step coordinate descent is summarized in **Algorithm 4**.

Algorithm 4 One step coordinate descent algorithm for L1-penalized multi-state model.

```
Result: $\hat{\beta}$
Data: Input: Long-format data described in Section 4.3.3
while ($q > 0$ and $q \le Q$) do
    Compute $\tilde{\boldsymbol{\eta}}_q = \mathbf{X}_q \tilde{\boldsymbol{\beta}}_q$; $l_q^{'}(\tilde{\boldsymbol{\eta}}_q)$; $l_q^{''}(\tilde{\boldsymbol{\eta}}_q)$; $\mathbf{z}(\tilde{\boldsymbol{\eta}}_q) = \tilde{\boldsymbol{\eta}}_q - l_q^{''}(\tilde{\boldsymbol{\eta}}_q)^{-1} l_q^{'}(\tilde{\boldsymbol{\eta}}_q)$
    repeat
        For $g = 1, 2, \ldots, P$: Update $\tilde{\beta}_q^g = \hat{\beta}_q^g$ using (4.15)
    until Convergence of $\hat{\boldsymbol{\beta}}_{\boldsymbol{q}}$;
    Update $\tilde{\boldsymbol{\beta}}_q = \hat{\boldsymbol{\beta}}_q$
end
Update $\tilde{\beta} = \hat{\beta}$
```

4.4.3 Active set updates

To improve the computational speed of the **L1mstate** package, we have constructed an active set $A = \{\hat{\beta}_q^g \neq 0\}$ that takes advantage of the sparsity of β. As shown in the **Algorithm 4**, we only need to update the non-zero coefficients $\hat{\beta}_q^g$ in A after a complete cycle has run through all the risk factors , i.e., when $\tilde{\beta} = 0$, $\hat{\beta}_q^g$ will stay zero if $\Big| - \frac{1}{N_q} l_q^{'}(\vec{0}) X_{q,g} \Big| < \lambda$; otherwise, $\hat{\beta}_q^g$ will be updated and stored in the active set if $\Big| - \frac{1}{N_q} l_q^{'}(\vec{0}) X_{q,g} \Big| > \lambda$. Therefore, the number of updates is reduced significantly and the convergence of the algorithm is increased. The algorithm will stop if another complete cycle does not change this set. Note that the active set A can only become larger after each update, so the algorithm will always stop after a finite number of updates (See Meier *et al.* 2007 [97] for more details of the convergence property.)

4.4.4 Pathwise solution

The above procedure is just for one fixed value of λ. However, in general, it is of interest to be able to compute the optimal solution for a range of values of λ. Thus, we aim to compute the regularization path (denoted as $\hat{\beta}(\lambda)$) where $\lambda \in [0, \infty]$. It can be shown that $\hat{\beta}(\lambda)$ turns out to be a piecewise linear, continuous function of λ [31]. In other words, we only need to compute the solutions on the change points in this path, denoted $\lambda_{max} \geq \lambda_1 \geq \cdots \geq \lambda_{min} \geq 0$. We can start with λ_{max} that is any value sufficiently large for which the entire coefficients $\hat{\beta} = 0$. From Eq. (4.15), notice that when $\tilde{\beta} = 0$, $\hat{\beta}_q^g$ will stay zero if $\Big| - \frac{1}{N_q} l_q^{'}(\vec{0}) X_{q,g} \Big| < \lambda$. Hence, we can set

$$\lambda_{max} = \max \Big| - \frac{1}{N_q} l_q^{'}(\vec{0}) X_{q,g} \Big|, \text{for } q = 1, 2, \ldots, Q;\ g = 1, 2, \ldots, P.$$

Following the suggestions made in Simon *et al.* 2011 [30], we can ignore solutions for that are close to 0 and set $\lambda_{min} = \epsilon \lambda_{max}$, then, compute the solutions over $m+1$ values defined as $\lambda_i = \lambda_{max} \big(\frac{\lambda_{min}}{\lambda_{max}}\big)^{\frac{i}{m}}$, for $i = 0, 1, \ldots, m$ and $\epsilon = \begin{cases} 0.01, & \text{if } N < P \\ 0.0001, & \text{if } N \geq P \end{cases}$. In doing this, the algorithm usually converges well because we could use the preceding solution (i.e., for λ_i) as the initial values to obtain the solution for λ_{i-1}.

4.4.5 Selection of the tuning hyperparameters

With a path of solutions, we need to select an optimal one. The natural choice is cross-validation. However, the partial likelihood of multi-state model is not as well defined as the Gaussian log likelihood on the left out sample using the traditional cross-validation, which leads to poor results. To tackle it, we used the cross-validation method as described in Verweij *et al.* 1993 [33], proposed for Cox regression model, in which data are split into k parts, use $(k-1)$ parts to train the model, and then, validate the learned model on the whole data. The cross-validated log-partial likelihood for a given part i and λ is

$$\widehat{\text{CV}}_i(\lambda) = l\big(\hat{\beta}_{-i}\big) - l_{i-1}\big(\hat{\beta}_{-i}\big),$$

which can be used as the goodness-of-fit estimate of the solution. Here, $\hat{\beta}_{-i}$ and l_{-i} are the optimal coefficients and its corresponding log-partial likelihood for data excluding part i. The total goodness-of-fit, $\widehat{\text{CV}}(\lambda)$, is the sum of all $\widehat{\text{CV}}_i(\lambda)$. We find the optimal λ

$$\hat{\lambda}_{cvl} = \underset{\lambda}{\text{argmax}}\, \widehat{\text{CV}}(\lambda)$$

However, this method alone sometimes produces high true positive rates (TPR) and high false positive rates (FPR). To reduce FPR without large reduction of TPR, we use the penalized method proposed in Ternes *et al.* 2016 [34]. Let p_λ be the number of non-zero coefficients in the model for a given λ, we can find the optimal λ that maximizes

$$\widehat{\text{CV}}(\lambda) - \frac{\widehat{\text{CV}}(\hat{\lambda}_{cvl}) - \widehat{\text{CV}}(\lambda_{max})}{p_{\hat{\lambda}_{cvl}}} * p_\lambda, \text{ for all } \lambda \in \Big[\hat{\lambda}_{cvl}, \lambda_{max}\Big].$$

Intuitively, it reduces the sparsity of the model p_λ without decreasing much the goodness-of-fit of the model $\widehat{\text{CV}}(.)$.

4.4.6 Estimation of the cumulative hazard rates and the transition probabilities

In the previous section, we have modeled and estimated the effects of the risk factors upon the transition intensities. To further assess the effects of the risk factors on disease progression; in particular, the effects of the risk factors on the cumulative hazard rates and the transition probabilities, we will present how to estimate the transition-specific hazard rates and the transition probabilities in the following.

Given the estimated regression coefficients, the baseline hazards of transition q, denoted by $\alpha_{q0}(t, \beta_q)$, can be obtained as the Breslow estimators [98]

$$\hat{\alpha}_{q0}(t, \hat{\beta}_q) = \frac{dN_q(t)}{S_q^{(0)}(t, \hat{\beta}_q)},$$

where $dN_q(t)$ is the number of events of transition q up to and including time t, and

$$S_q^{(0)}(t, \hat{\beta}_q) = \sum_{n=1}^{N_q} \text{exp}\big(\hat{\beta}_q^T X_q^n\big) Y_q^n(t),$$

Recall that, $Y_q^n(t)$ indicates that the n^{th} individual at risk in transition q at time t. Let the risk score for each subject of transition q be $\hat{r}_q^n = \text{exp}\big(\hat{\beta}_q^T X_q^n\big)$, then

$$\hat{\alpha}_{q0}(t, \hat{\beta}_q) = \frac{dN_q(t)}{\sum_{n=1}^{N_q} \hat{r}_q^n Y_q^n(t)},$$

The corresponding estimators of the cumulative baseline hazard $\hat{\Lambda}_{q0}(t, \hat{\beta}_q) = \int_0^t \hat{\alpha}_{q0}(u, \hat{\beta}_q) du$, is computed as

$$\hat{\Lambda}_{q0}(t, \hat{\beta}_q) = \sum_{u \leq t} \frac{dN_q(u)}{\sum_{n=1}^{N_q} \hat{r}_q^n Y_q^n(u)},$$

The cumulative hazard rates of transition q, denoted by $\hat{\Lambda}_q(t, \hat{\beta})$ which is also known as the Nelson-Aalen estimators, is

$$\hat{\Lambda}_q(t, \hat{\beta}) = \hat{\Lambda}_{q0}(t, \hat{\beta}_q) \text{exp}\big(\hat{\beta}_q^T X_q\big),$$

Given the cumulative transition hazards, using the basic tool $-$ a product integral allows us to estimate the transition probability matrix $\mathrm{P}(s,t) = \big(\mathrm{P}_{hj}(s,t)\big)$ as

$$\mathrm{P}(s,t) = \prod_{u \in (s,t]} \Big(\mathrm{I} + \Delta\hat{\Lambda}(u)\Big).$$

where $\prod$ is a product-integral and $(s,t]$ denotes the time interval. It is the Aalen-Johansen estimator [99].

4.4.7 Computational complexity analysis

We now discuss the complexity of the algorithms when using different frameworks (L1MSTATE, L1Cox, L1-StratifiedCox) for variable selection. They all solve the optimization problems by the coordinate descent algorithms to optimize the objective function with respect to one variable at a time while all the others are fixed. In other words, they process the same procedure: precompute the first-order derivatives and the diagonal entries of the second-order derivatives of a design matrix; at each iteration update P_a-the number of nonzero elements in the active set. The computational complexity depends on the number of subjects N, the number of risk factors P and the number of transitions Q. More specifically, consider L1MSTATE and L1Cox, for each transition, they need $\mathcal{O}(N_q^2)$ operations to compute the derivatives where N_q is the number of subjects for transition q (recall that $N = \sum_{q=1}^{Q} N_q$ and each update needs $\mathcal{O}(P)$ operations. Therefore, their complexity is $\mathcal{O}(\sum_{q=1}^{Q}(N_q^2 + P_q^a P))$ where P_q^a denotes the number of nonzero elements of transition q. For L1-StratifiedCox, it needs to create transition-specific risk factors from the baseline risk factors as described in [13]: each risk factor X is split into as many risk factors X_q as there are transitions in the model, for transition q $X_q = X$;while for all other transitions $X_q = 0$. It means that the number of risk factors now is PQ. In addition, it needs $\mathcal{O}(N^2)$ operations to compute the derivatives. Therefore, its complexity is $\mathcal{O}(N^2 + PQ\sum_{q=1}^{Q} P_q^a)$. Of course, the required runtime for the entire solution path also depends on the number of iterations, which in turn depends on the data and λ values. In general, the dominant factor influencing the number of iterations is the number of nonzero elements at the specific λ value since the nonactive elements that remain fixed at zero need no iteration. In the next section, we compare their computational complexity empirically in Table 9 with the runtime of three L1-regularized models using the same maximum number of iterations 105 for all models.

4.5 Simulation studies

In this section, we will numerically compare the performance of the L1-regularized multi-state model (L1MSTATE) with existing regularized multi-state models including the L1-regularized cause-specific Cox proportional hazards model (L1Cox) that applied the L1-regularized Cox proportional hazards model for each transition, and the L1-regularized stratified Cox proportional hazards model (L1-StratifiedCox) in term of variable selection using simulated data. The L1-regularized estimation of the fused-lasso multi-state model approach [90] was not included in our comparison due to very huge computation cost (see Discussion section for more details.) We also include the un-regularized multi-state model (MSTATE) to investigate the pros and cons of the un-regularized methods comparing with the regularized methods.

4.5.1 Setup

Following the data structure outlined in Section 4.3.3, we generate trajectories of N individuals that include their transitions among states, the times of the transitions, and the values of risk factors. First, the values of the risk factors of each individual are generated by randomly sampling from a P-dimensional multivariate normal distribution with mean vector as zero and the correlation matrix **C** as an autoregressive matrix where $\mathbf{C}_{ij} = \rho^{|i-j|}$ and $0 \leq \rho \leq 1$. The reason to use an autoregressive correlation matrix is that we could flexibly tune the correlations of the variables by setting the value of ρ, i.e., $\rho = 0$ means no correlation among the variables, while $\rho = 1$ means that the risk factors are perfectly correlated as duplicates of each other. Second, the transitions among states and their timing are generated as follows. Recall that we have assumed that the transition intensities between two states follow the proportional hazards Cox model Eq. (4.1). By setting up values for β we can obtain the transition intensity distribution from Eq. (4.1) to randomly sample the transition intensity values. After that, the observed times of the transition events between two states are generated using the exponential distribution with its rate parameter set to be the transition intensity between these two states. In here, we consider the illness-death model that includes three states: **healthy**, **illness**, and **death**. Its transition structure is depicted in Figure 4.2. Assume that all individuals start at the healthy state in the beginning of the observation period. The censoring status values are generated as follows. Since the observation time is the exact transition time, there is no illness censoring time or the censoring indicator of transition to illness state is 1 for all N individuals. The death censoring times are generated from the exponential distribution, the censoring indicator of transitions to death state is 0 if the death time is larger than the death censoring time, and 1 otherwise.

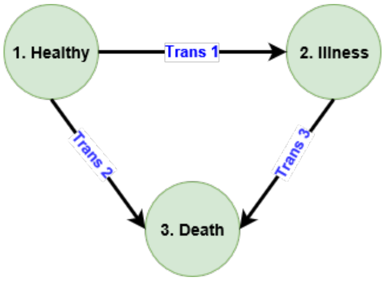


Figure 4.2: The illness-death model.

The strength of effect of risk factor is based on the real absolute value of its corresponding coefficient. Set the number of risk factors $P = 9$, we consider four scenarios: the first three scenarios include the effects of risk factors belong the same type (large, medium, or small), and the last scenario includes all three types of the effects of risk factors.

- First scenario: small effects

$$\beta = \begin{bmatrix} 0.15 & 0.15 & 0.15 & 0 & 0 & 0.15 & 0.15 & 0 & 0 \\ 0.15 & 0.15 & 0 & 0 & 0 & 0 & 0.15 & 0 & 0 \\ 0 & 0.15 & 0.15 & 0 & 0 & 0.15 & 0.15 & 0 & 0.15 \end{bmatrix}$$

- Second scenario: medium effects

$$\beta = \begin{bmatrix} -0.35 & -0.35 & -0.35 & 0 & 0 & -0.35 & -0.35 & 0 & 0 \\ -0.35 & -0.35 & 0 & 0 & 0 & 0 & -0.35 & 0 & 0 \\ 0 & -0.35 & -0.35 & 0 & 0 & -0.35 & -0.35 & 0 & -0.35 \end{bmatrix}$$

- Third scenario: large effects

$$\beta = \begin{bmatrix} -0.65 & -0.65 & -0.65 & 0 & 0 & -0.65 & -0.65 & 0 & 0 \\ -0.65 & -0.65 & 0 & 0 & 0 & 0 & -0.65 & 0 & 0 \\ 0 & -0.65 & -0.65 & 0 & 0 & -0.65 & -0.65 & 0 & -0.65 \end{bmatrix}$$

- Fourth scenario: mixed effects

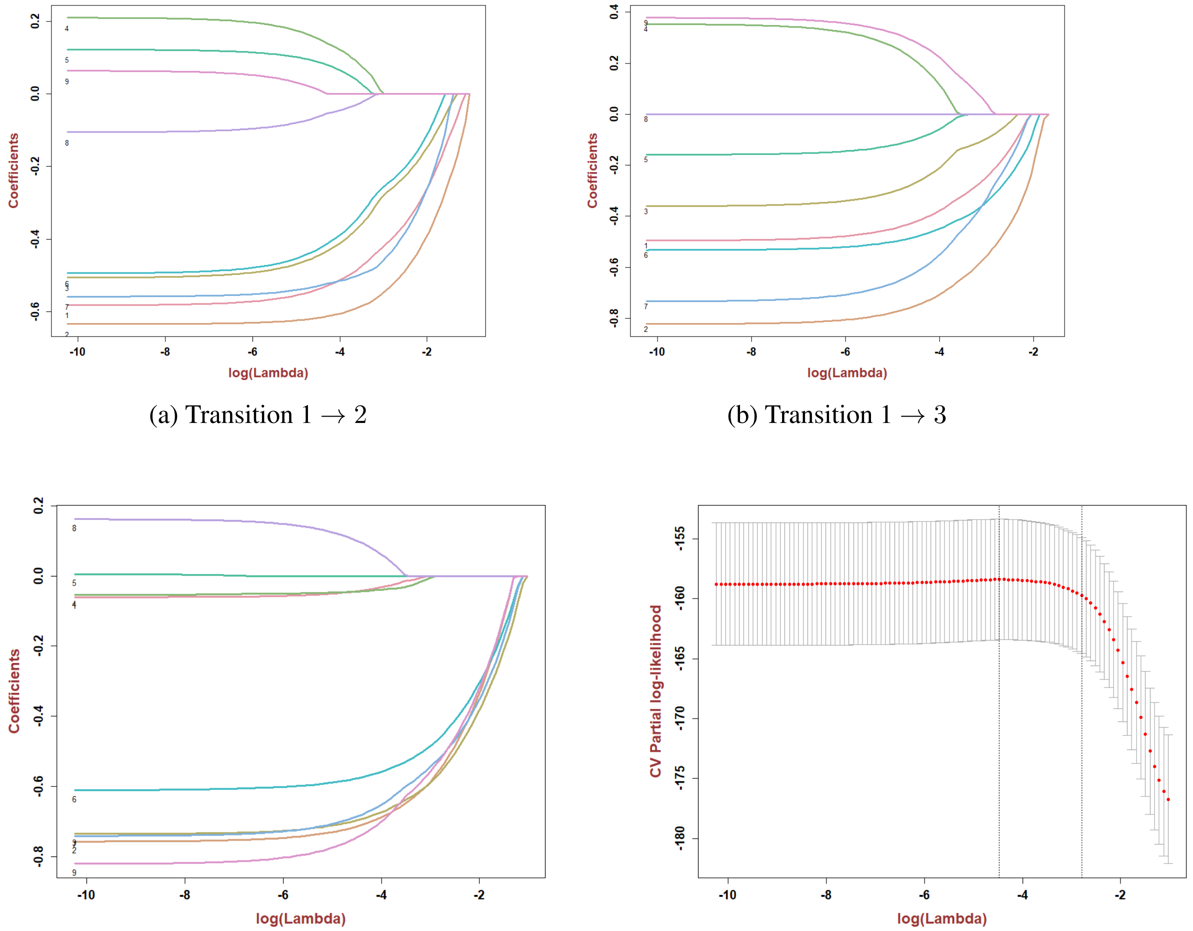
$$\beta = \begin{bmatrix} 0.15 & -0.35 & -0.35 & 0 & 0 & -0.35 & -0.35 & 0 & 0 \\ 0 & 0.15 & -0.65 & 0 & 0 & 0 & -0.65 & 0 & 0 \\ 0 & -0.65 & -0.65 & 0 & 0 & -0.35 & -0.65 & 0 & 0.15 \end{bmatrix}$$

We evaluate different levels of correlation between the risk factors by setting $\rho = 0, 0.2, 0.5$. The censoring percentage is 30%.

4.5.2 Results

To compare the performance of the four models in terms of identification of the significant risk factors, we calculated three performance metrics, including the true positive rate (TPR), false positive rate (FPR), and area under the ROC curve (AUC).

To compute TPRs and FPRs for the disease progression from the healthy state to the death state for our L1MSTATE, we created a path of 100 values of λ, applied 10-fold for two different cross-validation methods described above in Section 4.4.5 to select the optimal λ for variable selection. We can view the estimated coefficients from our L1MSTATE model fit, and the cross-validation log-partial likelihood against the log of λ values, and also how to use two different cross-validation



(a) Transition 1 → 2

(b) Transition 1 → 3

(c) Transition 2 → 3

(d) The cross-validation curve (red dotted line), and its standard deviation

Figure 4.3: Plots of the coefficients paths for three transitions of our L1MSTATE model fit and the cross-validation log-partial likelihood against the log of λ values along our path.

methods to select λ. Figure 4.3 shows the results of the large effects setting in which $N = 250$, and $\rho = 0.5$. For L1Cox and L1-StratifiedCox, we used '**glmnet**' package [30] with its default setting to fit Cox proportional hazards models: 100 values of λ and 10-fold cross-validation, which is the same as the first cross-validation method used in our model, to select the optimal solution. More specifically, for L1Cox, we applied for each transition using transition-specific datasets, then used the results of three transitions to compute the TPRs and FPRs; for L1-StratifiedCox, we applied using the long-format data. For MSTATE, we used R package '**mstate**' [100] to fit

Table 4.2: Model selection results of Example I for the small effects scenario.

N	ρ	MSTATE		pL1MSTATE		L1MSTATE		L1Cox		L1-StratifiedCox	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
50	0	0.14	0.11	0.05	0.03	0.07	0.05	0.19	0.17	0.09	0.08
	0.2	0.14	0.06	0.07	0.04	0.09	0.05	0.21	0.16	0.12	0.07
	0.5	0.13	0.09	0.09	0.04	0.12	0.06	0.23	0.18	0.16	0.10
250	0	0.32	0.07	0.30	0.08	0.51	0.27	0.61	0.38	0.54	0.30
	0.2	0.28	0.06	0.42	0.12	0.62	0.30	0.67	0.44	0.62	0.33
	0.5	0.22	0.06	0.45	0.13	0.70	0.33	0.73	0.45	0.70	0.39
450	0	0.47	0.08	0.52	0.11	0.84	0.45	0.83	0.56	0.83	0.49
	0.2	0.47	0.08	0.56	0.10	0.85	0.42	0.85	0.53	0.86	0.47
	0.5	0.37	0.06	0.59	0.13	0.86	0.43	0.87	0.50	0.83	0.44

pL1MSTATE, L1-regularized multi-state model using the penalized cross-validation method; L1MSTATE, L1-regularized multi-state model using the first cross-validation method; MSTATE, multi-state model; L1Cox, L1-regularized cause-specific Cox model using the first cross-validation method; L1-StratifiedCox, L1-regularized stratified Cox model using the first cross-validation method; TPR, true positive rate; FPR, false positive rate.

model and the statistical hypothesis test (p-value) with the 0.05 significance level to evaluate the significance of candidate risk factors for variable selection. We used different values of sample size, i.e., $N \in \{50, 250, 450\}$. The results across 100 replications for these models in different scenarios are summarized in Tables 4.2, 4.3, 4.4, and 4.5.

The results from Tables 4.2, 4.3, 4.4, and 4.5 show that TPR and FPR values of pL1MSTATE are always lower than L1MSTATE. It means that the penalized cross-validation method is more conservative than the first cross-validation method. On the one hand, comparing L1MSTATE and MSTATE results, except in small effects setting when $N = 50$ L1MSTATE gives lower both TPRs and FPRs than MSTATE, MSTATE always gives lower TPRs and FPRs than L1MSTATE. In other words, applying the statistical hypothesis test with the 0.05 significance level to MSTATE produces more sparse models than applying the first cross-validation method to L1MSTATE. On the other hand, comparing pL1MSTATE and MSTATE results shows that when $N = 50$ pL1MSTATE often gives lower both TPRs and FPRs than MSTATE, but it becomes to give better results (higher TPRs and lower FPRs) than MSTATE when the effects increase and the correlation among the

Table 4.3: Model selection results of Example I for the medium effects scenario.

N	ρ	MSTATE		pL1MSTATE		L1MSTATE		L1Cox		L1-StratifiedCox	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
50	0	0.29	0.11	0.18	0.06	0.24	0.09	0.39	0.22	0.29	0.15
	0.2	0.28	0.11	0.31	0.09	0.41	0.16	0.51	0.29	0.41	0.18
	0.5	0.22	0.10	0.35	0.09	0.51	0.21	0.59	0.33	0.54	0.24
250	0	0.84	0.10	0.81	0.15	0.98	0.60	0.98	0.65	0.98	0.64
	0.2	0.83	0.08	0.82	0.13	0.98	0.58	0.99	0.63	0.98	0.59
	0.5	0.70	0.07	0.77	0.13	0.97	0.52	0.97	0.60	0.97	0.53
450	0	0.96	0.14	0.88	0.13	1.00	0.69	1.00	0.70	1.00	0.69
	0.2	0.97	0.12	0.91	0.13	1.00	0.65	1.00	0.68	1.00	0.66
	0.5	0.88	0.10	0.87	0.14	1.00	0.59	1.00	0.62	0.99	0.61

Table 4.4: Model selection results of Example I for the large effects scenario.

N	ρ	MSTATE		pL1MSTATE		L1MSTATE		L1Cox		L1-StratifiedCox	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
50	0	0.52	0.12	0.45	0.13	0.65	0.28	0.70	0.41	0.63	0.30
	0.2	0.50	0.12	0.57	0.16	0.76	0.31	0.78	0.41	0.73	0.34
	0.5	0.38	0.11	0.57	0.14	0.82	0.36	0.86	0.47	0.82	0.39
250	0	0.99	0.16	0.92	0.14	1.00	0.70	1.00	0.70	1.00	0.70
	0.2	0.98	0.13	0.94	0.15	1.00	0.66	1.00	0.68	1.00	0.66
	0.5	0.96	0.11	0.93	0.18	1.00	0.59	1.00	0.61	1.00	0.62
450	0	1.00	0.19	0.97	0.15	1.00	0.73	1.00	0.73	1.00	0.73
	0.2	1.00	0.19	0.97	0.12	1.00	0.70	1.00	0.70	1.00	0.70
	0.5	0.99	0.15	0.96	0.18	1.00	0.65	1.00	0.65	1.00	0.66

pL1MSTATE, L1-regularized multi-state model using the penalized cross-validation method; L1MSTATE, L1-regularized multi-state model using the first cross-validation method; MSTATE, multi-state model; L1Cox, L1-regularized cause-specific Cox model using the first cross-validation method; L1-StratifiedCox, L1-regularized stratified Cox model using the first cross-validation method; TPR, true positive rate; FPR, false positive rate.

Table 4.5: Model selection results of Example I for the mixed effects scenario.

		MSTATE		pL1MSTATE		L1MSTATE		L1Cox		L1-StratifiedCox	
N	ρ	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
50	0	0.30	0.10	0.24	0.07	0.36	0.15	0.46	0.24	0.34	0.16
	0.2	0.30	0.09	0.23	0.06	0.39	0.16	0.50	0.26	0.39	0.18
	0.5	0.24	0.10	0.31	0.09	0.44	0.19	0.55	0.28	0.44	0.23
250	0	0.68	0.08	0.56	0.06	0.89	0.51	0.90	0.55	0.88	0.56
	0.2	0.67	0.09	0.53	0.05	0.88	0.50	0.88	0.58	0.86	0.51
	0.5	0.59	0.07	0.56	0.11	0.86	0.49	0.87	0.53	0.84	0.51
450	0	0.77	0.10	0.62	0.04	0.95	0.61	0.97	0.64	0.95	0.63
	0.2	0.75	0.10	0.60	0.04	0.96	0.61	0.96	0.66	0.95	0.63
	0.5	0.71	0.08	0.58	0.07	0.92	0.56	0.91	0.58	0.89	0.56

pL1MSTATE, L1-regularized multi-state model using the penalized cross-validation method; L1MSTATE, L1-regularized multi-state model using the first cross-validation method; MSTATE, multi-state model; L1Cox, L1-regularized cause-specific Cox model using the first cross-validation method; L1-StratifiedCox, L1-regularized stratified Cox model using the first cross-validation method; TPR, true positive rate; FPR, false positive rate.

risk factors ρ increases. When $N = 250$ and $N = 450$ in small setting, pL1MSTATE gives better results than MSTATE; in other settings, pL1MSTATE starts giving lower both TPRs and FPRs than MSTATE, and MSTATE gives better results in large effects setting. Note that when ρ increases - risk factors become highly correlated, MSTATE results become worse while L1MSTATE and pL1MSTATE results often become better. Consider three regularized models L1MSTATE, L1Cox, and L1-StratifiedCox using the same cross-validation method, from Tables 4.2, 4.3, 4.4 and 4.5, it can be seen that when $N = 50$, L1MSTATE is the most conservative model since it gives both the smallest TPRs and FPRs values; when N increases, L1MSTATE gives the best results with the highest TPRs and the lowest FPRs.

The TPRs and FPRs shown in these above tables depend on the selected methods including the cross-validation methods, and the significance level of $p-$value. We want to evaluate further the variable selection performance of these models using the area under a curve (AUC) values that are also variable selection metrics and do not depend on the selected methods. We use the same settings as above with different values of sample size, i.e., $N \in \{50, 75, \ldots, 500\}$. We first calculate the

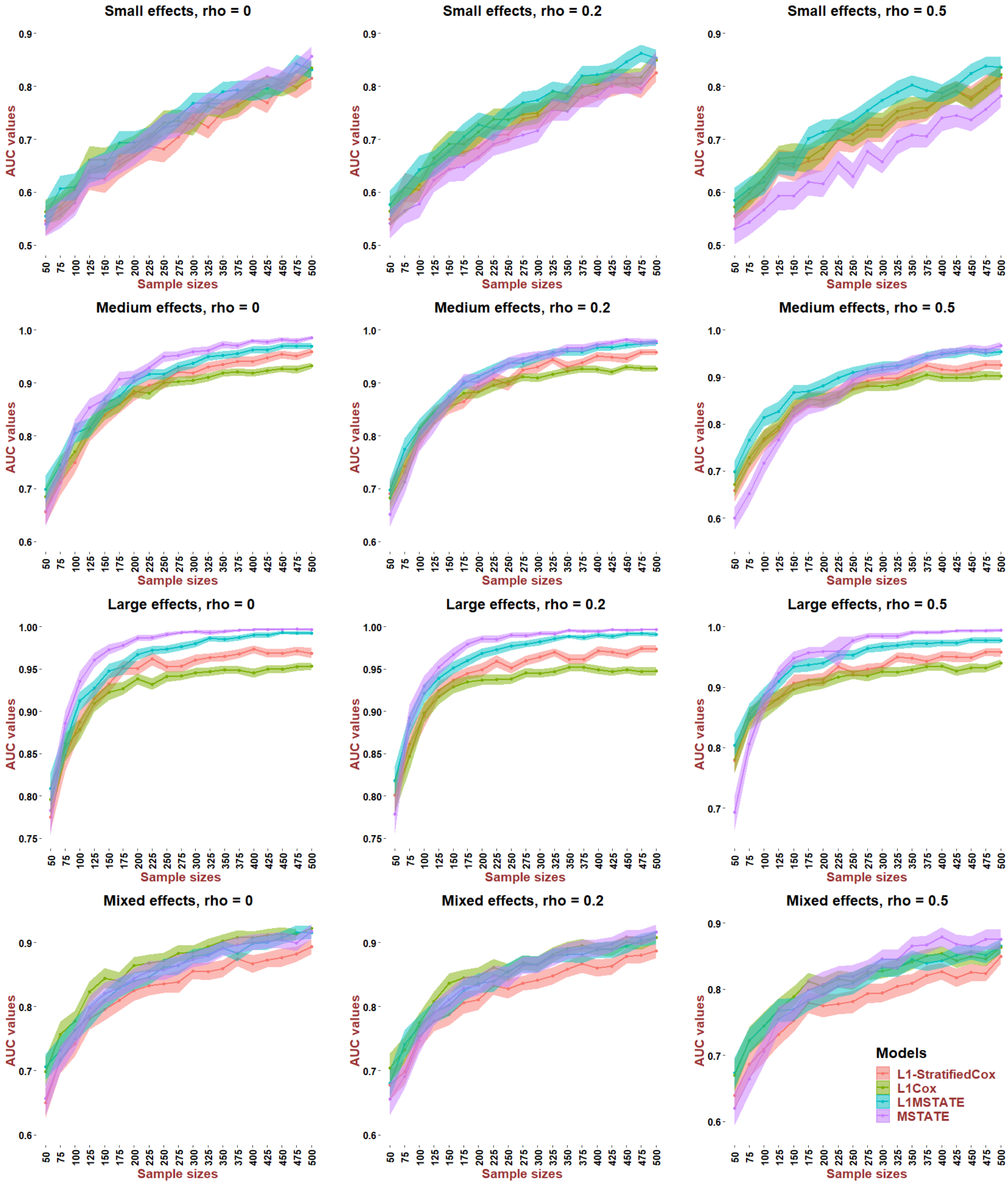


Figure 4.4: AUC values of Example I for different sample sizes in different settings over 100 replications.

TPRs and FPRs, then compute the AUC values by using the method described in Fawcett *et al.* 2006 [101]. Intuitively, the TPR and FPR pairs were calculated to construct ROC curves, then the area under a ROC curve (AUC) was computed. More specifically, in three regularized models L1MSTATE, L1Cox, and L1-StratifiedCox, it is straightforward to calculate 100 pairs of TPRs and FPRs corresponding to 100 λ values along the path. In MSTATE, the threshold path was constructed, and it included only the corresponding $p-$values of estimated coefficients. Then, for each threshold, the risk factors that have smaller $p-$values than the threshold were selected, and the corresponding TPR and FPR pairs were computed. The results of AUC values of these models in twelve settings for different datasets over 100 replications are shown in Figure 4.4.

First, we compare the performances of L1MSTATE and MSTATE. From Figure 4.4, in small effects setting, L1MSTATE gives comparable performance with MSTATE when there is no correlation among risk factors ($\rho = 0$), and better performance than MSTATE when the correlation ρ becomes higher. Other settings show the same pattern: when sample size is small, MSTATE performs worse than L1MSTATE; when sample size increases, MSTATE's performance gradually catches up, and even becomes better than L1MSTATE's performance. Notice that when the correlation among risk factors ρ increases, MSTATE needs more samples to be able to catch up L1MSTATE's performance, and when the effects become stronger, MSTATE needs less samples to perform comparably with L1MSTATE.

Second, we compare the performance of three regularized models L1MSTATE, L1Cox, and L1-StratifiedCox. In the first three settings L1MSTATE always gives the best performance. In the last setting L1MSTATE gives slightly worse performance than L1Cox when $\rho = 0$, and comparable when ρ increases; L1MSTATE also gives better performance than L1-StratifiedCox. Two models L1Cox and L1-StratifiedCox perform differently: they perform comparably in small effects setting; L1-StratifiedCox performs better L1Cox in medium and large effects settings; L1Cox performs better L1-StratifiedCox in mixed effects setting. L1MSTATE performs better than L1Cox can most likely be explained by the benefit of incorporating the prior knowledge about the disease progression model: in L1MSTATE, we incorporated information about multi-state model of disease progression into data process when converting the original data to long-format data; L1Cox, by contrast, applied L1-regularized Cox proportional hazards model for each transition-specific dataset separately. L1MSTATE performs better than L1-StratifiedCox even though both L1MSTATE and L1-StratifiedCox models use long-format data. The reason is that L1MSTATE assigned different weights to each transition while L1-StratifiedCox did not. Intuitively, L1MSTATE put higher penalties on rare transitions than common transitions.

In summary, the L1-regularized multi-state model (L1MSTATE) is the best one among the regularized models in terms of variable selection. L1MSTATE performs better at variable selection than the un-regularized multi-state model (MSTATE) when sample sizes are small or the effects are small, and MSTATE performs better than L1MSTATE when sample sizes are large or the effects are strong.

4.5.3 Large-scale datasets

In this setting, we only compare the performances of three L1-regularized models without including the un-regularized multi-state model (MSTATE). We set the number of risk factors $P = 300$ and the number of nonzero ones to be 100 per each transition. Different sample sizes, i.e., $N \in \{3000, 6000, 9000\}$, are simulated. The results of three L1-regularized models are shown in Tables 4.6, 4.7, and 4.8. They are consistent with the results of small datasets, which suggests that L1MSTATE is better than L1Cox and L1-StratifiedCox in terms of accurate variable selection.

4.5.4 Empirical runtime comparison

We further compare the runtime of three L1-regularized multi-state models on all the simulated datasets. As shown in Table 4.9, our L1MSTATE is the most computationally efficient as we expected based on our previous computational complexity analysis.

4.6 Europe Blood and Marrow Transplantation (EBMT) data

In this section, we will compare the performance of L1-regularized multi-state model (L1MSTATE) with un-regularized multi-state model (MSTATE) in terms of the predictions of the transition probabilities, and demonstrate how to use our **L1mstate** package to further assess the effects of risk fac-

Table 4.6: Model selection results of large P settings for the small effects scenario.

		L1MSTATE			L1Cox			L1-StratifiedCox		
N	ρ	TPR	FPR	AUC	TPR	FPR	AUC	TPR	FPR	AUC
3000	0.5	0.75	0.38	0.76	0.78	0.43	0.74	0.73	0.39	0.72
	0.2	0.76	0.42	0.75	0.73	0.51	0.74	0.78	0.45	0.72
	0	0.74	0.43	0.73	0.72	0.52	0.73	0.78	0.48	0.71
6000	0.5	0.87	0.51	0.81	0.88	0.55	0.79	0.86	0.50	0.77
	0.2	0.94	0.64	0.81	0.96	0.68	0.80	0.94	0.63	0.80
	0	0.95	0.69	0.80	0.96	0.73	0.78	0.95	0.69	0.79
9000	0.5	0.92	0.57	0.83	0.92	0.62	0.81	0.92	0.52	0.80
	0.2	0.98	0.72	0.84	0.98	0.74	0.82	0.98	0.67	0.83
	0	0.99	0.76	0.82	0.99	0.78	0.81	0.99	0.73	0.82

Table 4.7: Model selection results of large P settings for the medium effects scenario.

		L1MSTATE			L1Cox			L1-StratifiedCox		
N	ρ	TPR	FPR	AUC	TPR	FPR	AUC	TPR	FPR	AUC
3000	0.5	0.89	0.69	0.78	0.81	0.51	0.77	0.77	0.39	0.73
	0.2	0.91	0.64	0.79	0.49	0.60	0.78	0.88	0.53	0.77
	0	0.88	0.58	0.78	0.91	0.64	0.78	0.92	0.60	0.76
6000	0.5	0.95	0.77	0.82	0.89	0.66	0.81	0.90	0.50	0.78
	0.2	0.98	0.67	0.84	0.97	0.77	0.83	0.98	0.67	0.82
	0	0.99	0.74	0.83	0.99	0.81	0.82	0.99	0.73	0.82
9000	0.5	0.96	0.60	0.86	0.96	0.72	0.84	0.96	0.55	0.82
	0.2	0.99	0.72	0.86	0.99	0.82	0.85	0.99	0.70	0.84
	0	1	0.75	0.85	1	0.85	0.84	1	0.76	0.84

Table 4.8: Model selection results of large P settings for the large effects scenario.

		L1MSTATE			L1Cox			L1-StratifiedCox		
N	ρ	TPR	FPR	AUC	TPR	FPR	AUC	TPR	FPR	AUC
3000	0.5	0.83	0.47	0.77	0.82	0.57	0.77	0.78	0.41	0.73
	0.2	0.96	0.77	0.80	0.89	0.66	0.79	0.90	0.54	0.78
	0	0.99	0.86	0.80	0.93	0.71	0.79	0.94	0.62	0.77
6000	0.5	0.94	0.54	0.84	0.90	0.72	0.81	0.90	0.50	0.80
	0.2	0.98	0.66	0.84	0.97	0.82	0.83	0.98	0.66	0.82
	0	1	0.76	0.85	1	0.86	0.84	1	0.73	0.83
9000	0.5	0.96	0.48	0.83	0.96	0.79	0.84	0.95	0.55	0.82
	0.2	0.99	0.67	0.87	0.99	0.87	0.85	0.99	0.69	0.84
	0	1	0.74	0.87	1	0.90	0.85	1	0.75	0.84

L1MSTATE, L1-regularized multi-state model using the first cross-validation method; L1Cox, L1-regularized cause-specific Cox model using the first cross-validation method; L1-StratifiedCox, L1-regularized stratified Cox model using the first cross-validation method; TPR, true positive rate; FPR, false positive rate; AUC, area under the curve.

Table 4.9: Running time of three L1-regularized models. The mean time over different datasets (100 for small datasets and 10 for big datasets) required to fit the entire solution path over a grid of 100 λ values is reported in seconds.

N	ρ	L1MSTATE			L1Cox			L1-StratifiedCox		
		Small	Medium	Large	Small	Medium	Large	Small	Medium	Large
100	0	0.01	0.01	0.01	0.02	0.02	0.02	0.03	0.03	0.03
	0.2	0.01	0.01	0.01	0.02	0.02	0.02	0.03	0.03	0.04
	0.5	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.04	0.04
250	0	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.05	0.05
	0.2	0.02	0.02	0.02	0.03	0.03	0.04	0.04	0.05	0.05
	0.5	0.02	0.02	0.02	0.03	0.03	0.04	0.05	0.05	0.06
450	0	0.03	0.03	0.03	0.04	0.05	0.05	0.07	0.07	0.08
	0.2	0.03	0.03	0.03	0.04	0.05	0.06	0.07	0.07	0.08
	0.5	0.03	0.03	0.04	0.05	0.05	0.07	0.07	0.09	0.10
3000	0	2.47	3.29	5.12	3.93	4.72	6.49	11.43	11.79	12.51
	0.2	2.53	4.20	4.27	3.94	5.14	6.43	11.34	11.92	11.53
	0.5	2.96	4.42	3.30	4.46	5.97	8.06	12.13	12.49	12.65
6000	0	4.78	7.93	6.71	7.24	9.15	10.27	21.07	25.07	21.18
	0.2	6.22	7.97	5.56	7.90	9.57	10.56	24.87	24.71	20.64
	0.5	5.28	9.95	5.32	8.19	11.31	13.58	22.93	26.84	22.92
9000	0	8.48	12.34	12.71	10.91	13.15	17.09	33.16	37.65	38.77
	0.2	8.28	11.48	10.82	11.13	13.97	17.23	33.63	33.86	37.37
	0.5	9.06	9.10	9.39	12.15	15.64	21.01	35.72	35.83	37.62

L1MSTATE, L1-regularized multi-state model using the first cross-validation method; L1Cox, L1-regularized cause-specific Cox model using the first cross-validation method; L1-StratifiedCox, L1-regularized stratified Cox model using the first cross-validation method.

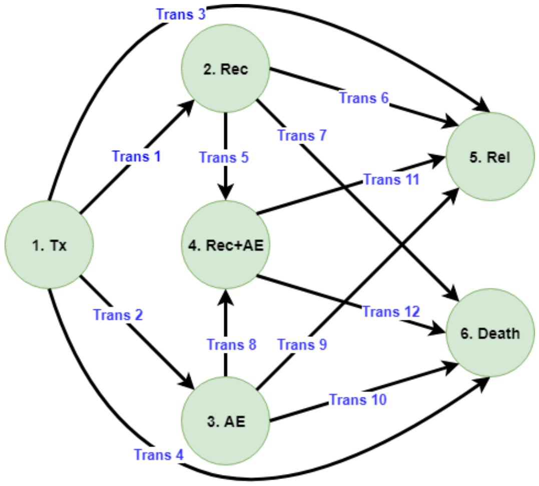


Figure 4.5: The EBMT model.

tors upon the disease progression using a the Europe Blood and Marrow Transplantation (EBMT) dataset that has been described and analyzed in deWreede *et al.* 2011 [100].

The model for the leukemia patients after bone marrow transplantation (so-called EBMT model) is shown in Figure 4.5. The EBMT model includes six states and twelve possible transitions. These states are transplant (Tx) state, recovery (Rec) state, adverse event (AE) state, combination of adverse event and recovery state (Rec+AE), relapse (Rel) state, and death, respectively. The numberic coding 1, 2,. . . , 12 represent twelve possible transitions. This dataset includes 2279 patients and the observed transitions are summarized in Table 4.10.

Table 4.10: The frequencies and proportions of the number of observed transitions of study population. The numbers in parentheses are proportions.

	Tx	Rec	AE	Rec+AE	Rel	Death	No event	Total
Tx	0 (0)	785 (0.34)	907 (0.40)	0 (0)	95 (0.04)	160 (0.07)	332 (0.15)	2279
Rec	0 (0)	0 (0)	0 (0)	227 (0.29)	112 (0.14)	39 (0.05)	407 (0.52)	785
AE	0 (0)	0 (0)	0 (0)	433 (0.48)	56 (0.06)	197 (0.22)	221 (0.24)	907
Rec+AE	0 (0)	0 (0)	0 (0)	0 (0)	107 (0.16)	137 (0.21)	416 (0.63)	660

The six risk factors are donor-recipient match, prophylaxis, year of transplant, and age of transplant in years. All of them are categorical risk factors. As in this chapter we focus on numeric risk factors, we convert them to numeric by using dummy coding as follow

- donor-recipient match (1 refers to yes and 0 refers to no)
- prophylaxis (1 refers to yes and 0 refers to no)
- year of transplant (1 refers to 1990-1994 and 0 refers to 1985-1989 or 1995-1998)
- year of transplant (1 refers to 1995-1998 and 0 refers to 1985-1989 or 1990-1994)
- age of transplant (1 refers to 20-40 and 0 refers to < 20 or > 40)
- age of transplant (1 refers to > 40 and 0 refers to < 20 or 20-40)

There are 12 allowable transitions in the model and six time-fixed risk factors for all transitions, resulting in the total number of coefficients as large as 72. For L1MSTATE, we used the regularization path of 100 values of λ, and applied 10-fold for both the first cross-validation method and the penalized cross-validation method to tune the penalty parameter λ. For MSTATE model, we used p-values to select the significant risk factors (highlighted as bold in Table 4.11). The results from two models are shown in Table 4.11. Table 4.11 shows consistent results with those in simulation studies: the penalized cross-validation method is more conservative than the first cross-validation method since it chooses more sparse multi-state models.

4.6.1 Comparison of the models

We compared the performance of L1MSTATE and MSTATE in terms of the predictions of the transition probabilities. As discussed in the introduction, our aim is to study how L1MSTATE and MSTATE predict the *rare* transitions that have relatively small number of observations and the *common* transitions that have relatively large number of observations. To do it, the transitions from the transplant state were considered, and three example patients A, B, and C (see Table 4.12) were chosen. The observed transitions from the transplant state of three patients are summarized in Table 4.13. The summary shows that the transitions from the transplant state to the recovery state

Table 4.11: Regression coefficients of two models for EBMT dataset.

Methods	Risk factors	Transitions											
		1	2	3	4	5	6	7	8	9	10	11	12
MSTATE	x_1	**-0.167**	-0.111	0.196	-0.003	0.190	**0.426**	0.244	0.126	-0.414	0.008	-0.301	**0.572**
	x_2	**-0.366**	**-0.278**	0.385	-0.056	-0.282	0.268	-0.008	0.125	0.159	0.324	0.012	-0.112
	x_3	**0.401**	0.023	0.442	-0.359	-0.095	-0.210	**-0.836**	**0.528**	-0.311	**-0.644**	-0.024	-0.362
	x_4	**0.521**	-0.114	0.221	**-0.476**	-0.151	0.055	**-0.980**	**0.930**	-0.580	-0.213	-0.390	-0.352
	x_5	0.049	0.123	-0.094	**0.766**	0.292	-0.255	0.150	**-0.393**	0.172	0.238	0.414	**0.760**
	x_6	0.199	0.067	-0.232	**0.934**	**0.470**	-0.101	**1.465**	**-0.328**	0.423	**0.495**	0.256	**1.337**
pL1MSTATE	x_1	-0.040	0	0	0	0	0.100	0	0	0	0	0	0.374
	x_2	-0.291	-0.137	0	0	-0.256	0	0	0	0	0.241	0	0
	x_3	0.117	0	0	0	0	0	0	0.231	0	-0.378	0	0
	x_4	0.250	-0.002	0	0	0	0	0	0.604	0	0	-0.080	0
	x_5	0	0	0	0	0.080	0	0	-0.193	0	0	0.035	0.106
	x_6	0.082	0	0	0	0.178	0	0.460	-0.056	0	0.085	0	0.627
L1MSTATE	x_1	-0.147	-0.093	0.053	0	0.146	0.385	0	0.068	-0.269	0	-0.262	0.521
	x_2	-0.352	-0.253	0.183	0	-0.464	0.199	0	0.071	0.161	0.315	0	-0.062
	x_3	0.353	0.022	0.183	-0.154	-0.068	-0.207	-0.501	0.458	-0.110	-0.602	0	-0.27
	x_4	0.476	-0.091	0	-0.230	-0.050	0.022	-0.593	0.850	-0.267	-0.153	-0.304	-0.241
	x_5	0.007	0.082	0	0.486	0.414	-0.165	0	-0.370	0	0.155	0.305	0.615
	x_6	0.158	0.018	0	0.568	0.557	0	1.154	-0.288	0.147	0.398	0.137	1.196

pL1MSTATE, L1-regularized multi-state model using the penalized cross-validation method; L1MSTATE, L1-regularized multi-state model using the first cross-validation method; MSTATE, multi-state model. For MSM method, the significance of risk factors that are at 0.05 levels are shown in bold.

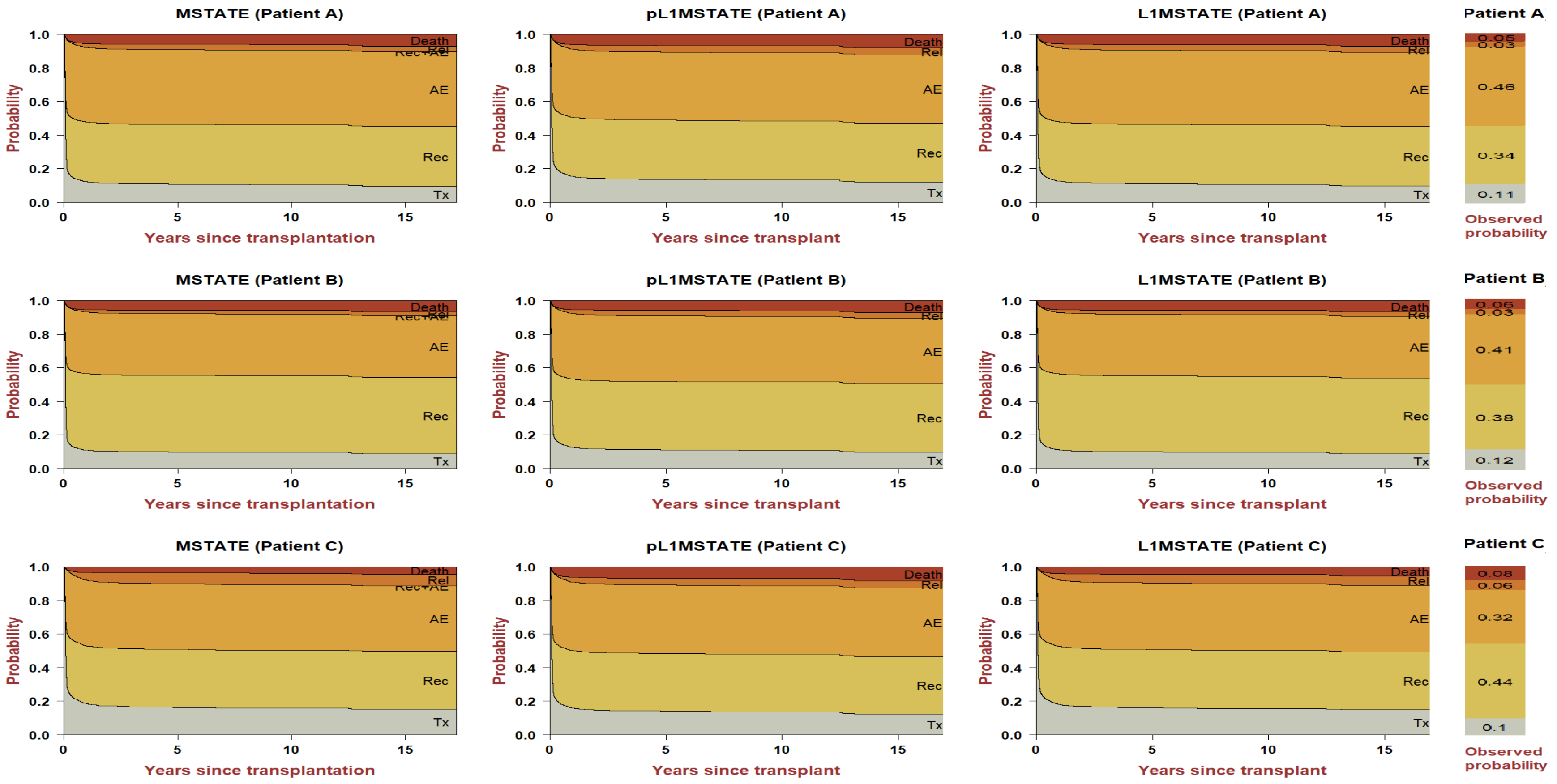


Figure 4.6: Estimates of stacked prediction transition probabilities from $t = 0$ for patients A, B, and C using two models. L1MSTATE, L1-regularized multi-state model using the first cross-validation method; pL1MSTATE, L1-regularized multi-state model using the penalized cross-validation method; MSTATE, multi-state model.

Table 4.12: Risk factors information of patient A, B, C and D.

Risk factors	Patient A	Patient B	Patient C	Patient D
x_1	0	0	1	0
x_2	0	0	0	0
x_3	1	0	1	0
x_4	0	1	0	1
x_5	1	0	0	1
x_6	0	1	0	0

Table 4.13: The frequencies and proportions of the number of observed transitions from the transplant state of three patients. The numbers in parentheses are proportions.

		Tx	Rec	AE	Rec+AE	Rel	Death	No event	Total
Patient A	Tx	0 (0)	99 (0.34)	133 (0.46)	0 (0)	9 (0.03)	13 (0.05)	33 (0.11)	287
Patient B	Tx	0 (0)	56 (0.38)	60 (0.41)	0 (0)	5 (0.03)	9 (0.06)	17 (0.12)	147
Patient C	Tx	0 (0)	22 (0.44)	16 (0.32)	0 (0)	3 (0.06)	4 (0.08)	5 (0.10)	50

and adverse event state have relatively large number of observations while the transition from the transplant state to the relapse state and the death state have relatively small number of observations. In other words, the transitions from the transplant state to the recovery state and adverse event state are the *common* transitions and the transition from the transplant state to the relapse state and the death state are *rare* transitions. In addition, patient A has the largest number of observations (287) that represents the large sample size case and patient C has the smallest number of observations (50) that represents the small sample size case. The same Aalen-Johansen method to predict the transition probabilities were used in both L1MSTATE and MSTATE. The results are shown in Figures 4.6: the probabilities of transitions from the transplant state at the starting computation times 0 to the ending computation times are estimated and stacked together where the distance between two adjacent curves shows the probability of the state whose name is labeled.

From Figure 4.6, it can be seen that the predicted probabilities from the transplant state at the starting time 0 of patient A using MSTATE, pL1MSTATE and L1MSTATE are almost similar but MSTATE and L1MSTATE slightly underestimates the probability of the relapse (Rel) state, and pL1MSTATE slightly underestimates the probability of the adverse event (AE) state comparing with the observed probability. In other words, pL1MSTATE slightly underestimates the probability of the common event while MSTATE and L1MSTATE slightly underestimates the probability of the rare event. The results of patient B clearly shows that MSTATE overestimates the probability of the common event - the recovery (Rec) state, and underestimates the probability of the rare event - the relapse (Rel) state. L1MSTATE also overestimates the probability of the recovery (Rec) state. pL1MSTATE gives the best overall performance. The results of patient C show the same pattern: MSTATE underestimates the probability of the rare event - the death state, and inaccurate prediction of the probability of the relapse state. By contrast, pL1MSTATE and L1MSTATE produces better predictions of these two rare events. The figure also indicates that MSTATE, pL1MSTATE and L1MSTATE overestimate the probability of the common event - the adverse event (AE) state.

In short, the un-regularized multi-state model (MSTATE) tends to underestimate the probabilities of the rare transitions, and overestimate the probabilities of the common transitions. Its performance becomes worse when the sample size decreases. In these cases, our L1-regularized multi-state models produce better predictions.

4.6.2 Further assessment of the effects of risk factors upon the disease progression

We illustrate how to use the functions of our **L1mstate** package to estimate the cumulative hazard rates and the transition probabilities. For illustrative purposes, we continued using patient A example, and chose another patient D (in Table 4.12) that differs from patient A only in terms of year of transplant since our aim is to assess the effect of year of transplant. The penalized cross-validation method was implemented to select the optimal tuning parameters.

Figure 4.7 shows the results of the Nelson-Aalen estimates of the four transitions starting from the transplant state for two patients A and D. There is a significant difference of the cumulative hazard rates of the first transition (from transplant state to recovery state) between the two patients. In other words, the year of transplant has significant effect upon the cumulative hazard function of the first transition: if patient did the transplant in 1995-1998, their cumulative hazard rate to recovery state is higher if they did in 1990-1994. The results of the predicted transition proba-

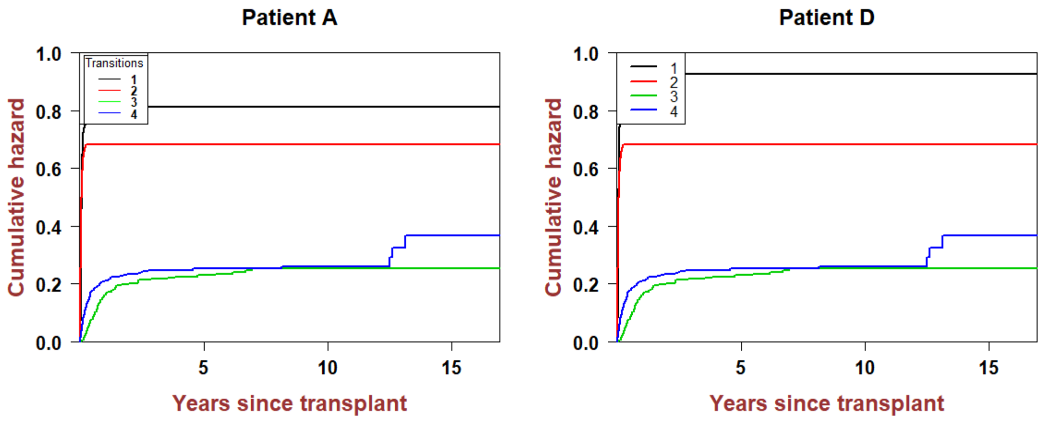


Figure 4.7: Estimated cumulative hazard rates for patient A and patient D.

bilities starting from the transplant state at starting computation time 0 of two patients in Figure 4.8 also show the strong effects of the year of transplant on the first transition probability. Note that it also shows the ability of risk factors (year of transplant) in discriminating patients who will have higher transferring risk (higher cumulative hazard and transition probability) by certain time (starting study time) from certain state (transplant).

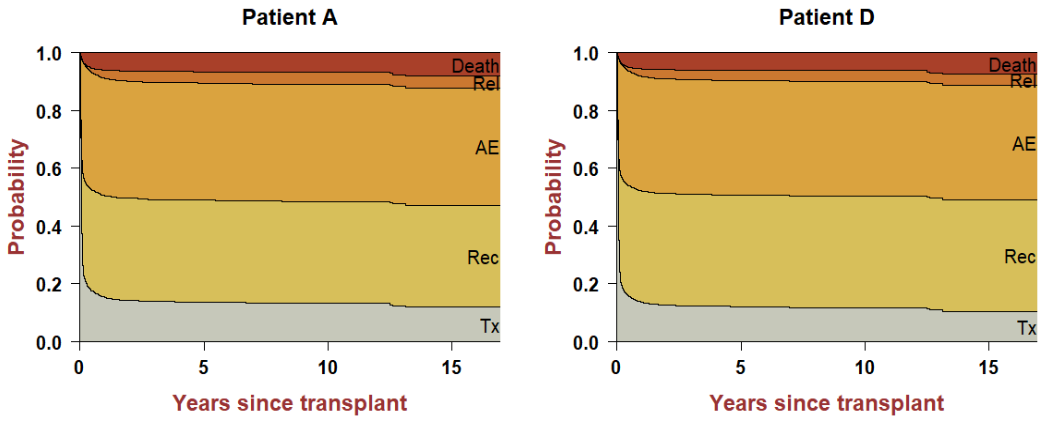


Figure 4.8: Estimates of stacked prediction transition probabilities from $t = 0$ for patients A and D.

Although all the transition probabilities presented above are predicted at starting times 0, our

Table 4.14: The frequencies and proportions of the number of observed transitions from the transplant state of patient D after 0 and 100 days. The numbers in parentheses are proportions.

Days since									
transplant		Tx	Rec	AE	Rec+AE	Rel	Death	No event	Total
$t = 0$	Tx	0 (0)	94 (0.40)	85 (0.36)	0 (0)	4 (0.02)	15 (0.06)	5 (0.15)	233
$t = 100$	Tx	0 (0)	2 (0.04)	0 (0)	0 (0)	3 (0.07)	5 (0.11)	35 (0.78)	45

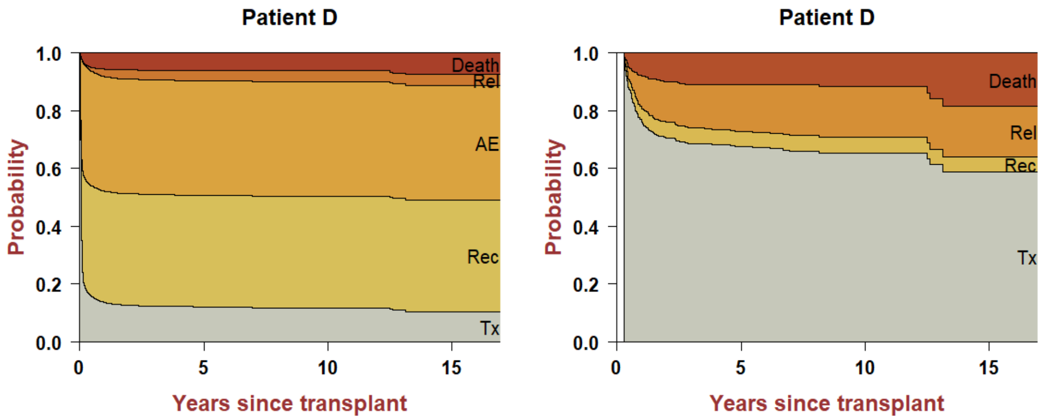


Figure 4.9: Estimates of stacked prediction transition probabilities of patient D from $t = 0$ and $t = 100$ days since transplant.

L1mstate package also allows to compute the predicted transition probabilities at different starting computation times. For example, we can choose the starting computation times are 100 days since transplant to compute the predicted transition probabilities of patient D. Results in Figure 4.9 show the considerable changes of the distributions of the state probabilities: the probability of the transplant state increases substantially, and the probabilities of the relapse and death states also increase. In other words, if patient D can survive through the transplant state during the first 100 days, the chance that they may stay at the current state increases. Since the risk factors are assumed time-constant, this phenomenon may imply the effects of the risk factors upon the transition probabilities change over time or the sojourn time that patient D spent in the transplant state also affects upon the predicted transition probabilities.

5. PENALIZED JOINT MODELS OF TIME-TO-EVENT AND MULTIVARIATE LONGITUDINAL OUTCOMES

Joint modelling of longitudinal and survival data has attracted increasing attention in the methodological literature over the past decade as more of such data become available in clinical studies. However, there are only a limited number of available methods for variable selection to identify critical time-varying risk factors as most of the existing works have been typically developed for the joint models with one single longitudinal outcome and one single survival outcome. In this chapter, we present a variable selection framework that can analyze multiple longitudinal outcomes. We focus on the problems of identifying the longitudinal outcomes that play important roles in the survival submodel and simultaneously selecting relevant covariates for both longitudinal and survival outcomes of interest, for which there is no available tool so far to the best of our knowledge. We propose novel penalized joint models for different association structures between the longitudinal and the survival submodels using different penalty terms. To tackle high-dimensional challenges that arise when considering many longitudinal outcomes, covariates and random effects, we develop an estimation procedure based on Laplace approximation of a joint likelihood. Simulation studies demonstrate the excellent variable selection performances of the proposed methods, which are further validated on a real-world dataset from patients with primary biliary cirrhosis.

5.1 Introduction

Subjects are followed up repeatedly in longitudinal studies. Different types of measurements are collected, namely the covariates varying with time such as biomarkers, in addition to the outcomes of main interest, such as the time to an event, including infection, death, or dropout from the study. In cancer studies, for example, the longitudinal measurements of antibody levels or of other biomarkers at each follow-up clinic appointment for patients and the event time such as time of death or metastasis are recorded. The covariates are usually measured intermittently with error, often at different times and with a different number for each individual. The time to an event is often censored. One of statistical models to understand the underlying dynamics, for example disease progression in biomedicine, is to model these processes using a linear mixed effects model for the time-varying covariates (longitudinal outcome) [102], and a survival model such as Cox regression model [3] for the event time. Another category of so-called two-stage models first fits the longitudinal process separately to obtain the maximum likelihood estimation (MLE) and best linear unbiased predictors (BLUPs) of the random effects, then fits the survival data using the longitudinal fitted values as covariates in a second stage. However, in many settings, these approaches can be inefficient and lead to biased effect size estimates [103], [104], [105]. The joint model framework provides a solution by using the survival model for the time-to-event outcome, which depends on the true underlying value of the longitudinal outcomes that are modeled using the linear mixed effects models, and using the joint distribution from both outcomes to derive estimation. The literature on joint models is extensive, with excellent reviews [106], [107], [108], [109] [110].

The basic joint model describes the association between a single time-to-event outcome and a single longitudinal outcome. In many studies, however, it cannot capture the complicated dynamic processes of complex diseases whose collected data can be complex and include multiple longitudinal measurements and possibly multiple, recurrent or competing event times. An example showing the advantages of incorporating multiple longitudinal outcomes in the joint model is presented in [111]. It described a study on 407 patients with chronic kidney disease who underwent a renal transplantation. Three biomarkers including glomerular filtration rate (GFR), blood haematocrit level, and proteinuria are measured repeatedly with the clinical interest being the time to graft failure. The authors in [109] provided an overview about recent works of joint models for time-to-event and multiple longitudinal outcomes.

This joint model is suitable to investigate the relationship between multiple longitudinal outcomes and time-to-event outcome, and the relationship between correlated longitudinal outcomes. Our interest is to develop a framework that can simultaneously identify the important longitudinal risk factors and time-constant risk factors with strong effect upon the time-to-event outcome, and the important (possibly) time-varying risk factors with strong effect upon each longitudinal outcome. In the joint model settings, some traditional methods perform variable selection by using information-based criteria such as the Akaike and Bayesian information criteria (AIC and BIC). Such kinds of variable selection methods are suitable when the number of candidate models is relatively small. However, the joint model for multiple longitudinal outcomes and time-to-event outcome often results in the complicated estimation problems with a considerable number of candidate models which leads to unstable estimates. In this chapter, the regularization approaches based on sparse penalties have been used to address these challenges.

Although the sparse models have gained popularity in statistics and machine learning, very little has been explored in the joint models. The authors of [112] and [113] proposed variable selection methods in joint models for a single longitudinal outcome and a single time-to-event outcome. Their works are different from ours that considers the joint model for multiple longitudinal outcomes and time-to-event outcome. One recent paper on using the regularization approach in the joint model for multiple longitudinal outcomes and the time-to-event outcome by [114]. They considered the time-independent association structure for joint models of time-to-event and multivariate longitudinal outcomes that includes only the random effects—random intercepts and slopes of the longitudinal submodel. They selected the important features among these random effects using the L1-norm penalties. It is different from our main objective that is to identify the important longitudinal outcomes have strong association with the survival outcome, and to simultaneously identify the important (possibly time-varying) risk factors with strong effect upon each longitudinal outcome. In this chapter, we study variable selection in multivariate joint models considering two different association structures between the longitudinal submodel and the time-to-event submodel: time-independent association (Model I) including only the random effects and time-dependent association (Model II) presented in [115]. In Model I, since the random effects represent the deviation of trajectories of longitudinal processes, we select the longitudinal processes by incorporating the group structures of the random effects, i.e., the random effects of each longitudinal outcome can be considered as a group via group LASSO penalties [18]. In Model II, we identify the important longitudinal outcomes using L1 penalties. In addition, in both models, our framework allows simultaneous selection of fixed effects in both longitudinal submodel and survival submodel using L1 penalties.

The most commonly used methods to estimate the joint model parameters are the expectation-maximization (EM) algorithm in frequentist settings and Markov chain Monte Carlo (MCMC) or Hamiltonian Monte Carlo (HMC) algorithms in Bayesian settings. However, they are computationally intensive. Considering the complicated joint models for multiple longitudinal outcomes and survival time when the number of random effects is large, the EM algorithm needs to compute the integral with respect to the number of random effects in the E-step which is challenging and sometimes impossible. In this chapter, we apply Laplace approximation of the joint likelihood discussed in [116] and [117] to tackle this challenge. Then, we develop several algorithms to solve the optimization problem.

The rest of the chapter is organized as follows. Section 2 describes the model and notation. Section 3 describes the estimation procedure. In Section 4, we present some simulation studies. We apply this framework to a real-world dataset in Section 5. Lastly, we end with discussions and conclusions in Section 6.

5.2 Models and Notations

For each individual $i = 1, \ldots, n$, we observe survival time and longitudinal outcomes. Denote T_i^* be the true survival time and assume that the survival time is subject to right censoring, we observe $T_i = \min\{T_i^*, C_i\}$ where C_i corresponds to a potential censoring time, and the censoring indicator δ_i, which $= 1$ if the failure is observed ($T_i^* \leq C_i$) and $= 0$ otherwise. We assume that censoring is independent of other survival and longitudinal outcomes information.

The longitudinal outcome of individual i is a vector $y_i = (y_{i1}^T, \ldots, y_{iJ}^T)$ where each y_{ij} is a $K_{ij}-$dimensional vector of observed longitudinal measurements for the j^{th} longitudinal outcome: $y_{ij} = (y_{ij1}(t_{ij1}), \ldots, y_{ijK_{ij}}(t_{ijK_{ij}}))^T$. Each observed value $y_{ijk}(t_{ijk})$ is measured at time t_{ijk}, where $i = 1, \ldots, n$, $j = 1, \ldots, J$, and $k = 1, \ldots, K_{ij}$. Here, K_{ij} can different between individual and longitudinal outcome.

We use the joint model that comprises of two submodels: a multivariate longitudinal data submodel, and a time-to-event data submodel.

5.2.1 Longitudinal submodel

The j^{th} longitudinal submodel is given by

$$y_{ijk}(t_{ijk}) = \beta_{0j} + X_i^T(t_{ijk})\beta_j + Z_i^T(t_{ijk})b_{ij} + \epsilon_{ijk}, \tag{5.1}$$

where $X_i^T(t_{ijk})$ is a P_j-dimensional vector of (possibly) time-varying covariates of i^{th} individual with corresponding fixed effects β_j; β_{0j} is an intercept; $Z_i^T(t_{ijk})$ is a R_j-dimensional vector of (possibly) time-varying covariates with corresponding random effects b_{ij}; ϵ_{ijk} are random errors. We assume that $\epsilon_{ijk} \overset{iid}{\sim} \mathcal{N}(0, \sigma_j^2)$, and ϵ_{ijk} and b_{ij} are uncorrelated. The random effects b_{ij} present the within-subject correlation between longitudinal measurements of j^{th} longitudinal outcome. Here, we assume b_{ij} follows a zero-mean multivariate normal distribution with $(R_j \times R_j)$-variance-covariance matrix $\mathbf{D}_{jj}$. To account for the association between different longitudinal outcomes, we let cov$(b_{il}, b_{im}) = \mathbf{D}_{lm}$ for $l \neq m$.

Furthermore, for each individual i, let $\mathbf{X}_i = \bigoplus_{j=1}^J X_{ij}$ and $\mathbf{Z}_i = \bigoplus_{j=1}^J Z_{ij}$ be block-diagonal matrices, where $X_{ij} = \big(1, X_i(t_{ij1}), \ldots, 1, X_i(t_{ijK_{ij}})\big)$ is an $K_{ij} \times (1 + P_j)$ matrix and $Z_{ij} = \big(Z_i(t_{ij1}), \ldots, Z_i(t_{ijK_{ij}})\big)$ is an $K_{ij} \times R_j$ matrix; $\bigoplus$ denotes the direct matrix sum. Similarly, denote $\mathbf{\Sigma}_i = \bigoplus_{j=1}^J \sigma_j^2 \mathbf{I}_{K_{ij}}$ and $\beta = (\beta_{0j}, \beta_j)_{j=1}^J$. Then, $y_i | b_i; \theta \sim \mathcal{N}(\mathbf{X}_i\beta + \mathbf{Z}_i b_i, \mathbf{\Sigma}_i)$.

5.2.2 Survival submodel

For the time-to-event outcome, we use the Cox's proportional hazard model. The survival submodel includes the subject-specific deviation from the population trajectories of longitudinal outcomes. More critically, each model has different form of survival submodel as follows:

- Model I:

$$h_i(t) = h_0(t)\text{exp}\big(V_i^T\gamma_0 + \sum_{j=1}^J b_{ij}^T\gamma_j^I\big) = h_0(t)\text{exp}\big(V_i^T\gamma_0 + b_i^T\gamma^I\big); \tag{5.2}$$

- Model II:

$$h_i(t) = h_0(t)\text{exp}\big(V_i^T\gamma_0 + \sum_{j=1}^{J}(\mathbf{Z}_{ij}^T b_{ij})\gamma_j^{II}\big) = h_0(t)\text{exp}\big(V_i^T\gamma_0 + (\mathbf{Z}_i b_i)^T\gamma^{II}\big), \tag{5.3}$$

where $h_0(t)$ is an unspecified baseline hazard function, V_i is $P-$dimensional vector that represents P baseline covariates for subject i with corresponding fixed effects γ_0. Moreover, γ^I is $\sum_{j=1}^J R_j-$dimensional vector of corresponding coefficients of random effects b_i, and γ^{II} is $J-$dimensional vector of corresponding association parameters.

5.2.3 The joint likelihood

The observed data for each individual can be denoted as (T_i, δ_i, y_i). We do not get to observe the random effects b_i. Let θ be the collection of all unknown parameters in the joint model including $\{\beta_0, \beta, \gamma_0, \gamma, (\sigma_1^2, \dots, \sigma_J^2)\}$ and elements in **D**. The observed data likelihood is given by

$$\mathcal{L}(\theta|\mathcal{D}_n) = \prod_{i=1}^n \int P(T_i, \delta_i|b_i;\theta)P(y_i|b_i;\theta)P(b_i;\theta)\,\text{d}b_i = \prod_{i=1}^n \int \mathcal{Q}(b_i)\text{d}b_i,$$

where

$$\begin{aligned} P(y_i|b_i;\theta) &= \prod_{j=1}^J\prod_{k=1}^{K_{ij}} \frac{1}{\sqrt{2\pi}\sigma_j}\text{exp}\Bigg[-\frac{1}{2\sigma_j^2}\Big(y_{ijk} - (\beta_{0j} + X_i^T(t_{ijk})\beta_j + Z_i^T(t_{ijk})b_{ij})\Big)^2\Bigg] \\ &= \frac{1}{\sqrt{\text{det}(2\pi\mathbf{\Sigma}_i)}}\text{exp}\Bigg[-\frac{1}{2}\Big(y_i - \mathbf{X}_i\beta - \mathbf{Z}_i b_i\Big)^T\mathbf{\Sigma}_i^{-1}\Big(y_i - \mathbf{X}_i\beta - \mathbf{Z}_i b_i\Big)\Bigg]; \\ P(b_i;\theta) &= \frac{1}{\sqrt{\text{det}(2\pi\mathbf{D})}}\text{exp}\Bigg[-\frac{1}{2}\Big(b_i^T\mathbf{D}^{-1}b_i\Big)\Bigg]; \\ \mathcal{Q}(b_i)\text{d}b_i &= P(T_i, \delta_i|b_i;\theta)P(y_i|b_i;\theta)P(b_i;\theta); \end{aligned}$$

and

- Model I:

$$P(T_i, \delta_i | b_i; \theta) = \left[h_0(T_i) \exp\big(V_i^T \gamma_0 + b_i^T \gamma^I\big) \right]^{\delta_i} \exp\left[- \int_0^{T_i} h_0(u) \exp\big(V_i^T \gamma_0 + b_i^T \gamma^I\big) \mathrm{d}u \right];$$

- Model II:

$$P(T_i, \delta_i | b_i; \theta) = \left[h_0(T_i) \exp\big(V_i^T \gamma_0 + (\mathbf{Z}_i b_i)^T \gamma^{II}\big) \right]^{\delta_i} \exp\left[- \int_0^{T_i} h_0(u) \exp\big(V_i^T \gamma_0 + (\mathbf{Z}_i b_i)^T \gamma^{II}\big) \mathrm{d}u \right].$$

5.3 Parameter estimation

5.3.1 Laplace approximation of the joint likelihood

The log-likelihood function is given by

$$l(\theta | \mathcal{D}_n) = \log \mathcal{L}(\theta | \mathcal{D}_n) = \sum_{i=1}^{n} \log \int \mathcal{Q}(b_i) \mathrm{d}b_i.$$

Let $\mathcal{K}(b_i) = \log \mathcal{Q}(b_i)$; clearly,

$$\mathcal{K}(b_i) = \log\Big\{ P(T_i, \delta_i | b_i; \theta) \Big\} + \log\Big\{ P(y_i | b_i; \theta) \Big\} + \log\Big\{ P(b_i; \theta) \Big\}.$$

Applying the Laplace approximation we yield the approximation

$$\mathcal{K}(b_i) \approx \mathcal{K}(\tilde{b}_i) + \frac{1}{2}(b_i - \tilde{b}_i)^T \mathbf{H}(\tilde{b}_i)(b_i - \tilde{b}_i),$$

where $\tilde{b}_i = \underset{b_i}{\operatorname{argmax}}\, \mathcal{K}(b_i)$ and $\mathbf{H}(\tilde{b}_i)$ is the Hessian matrix of $\mathcal{K}(b_i)$ at $\tilde{b}_i$. Therefore, the Laplace approximation of the log-likelihood function is

$$l(\theta | \mathcal{D}_n) \approx \sum_{i=1}^{n} \left[\mathcal{K}(\tilde{b}_i) - \frac{1}{2} \log\big(\det\big(- \mathbf{H}(\tilde{b}_i)\big)\big) + J \log(2\pi) \right]. \qquad (5.4)$$

We have

$$\begin{aligned}
\log\Big\{P(y_i|b_i;\theta)\Big\} &= -\sum_{j=1}^{J}\log(\sqrt{2\pi}\sigma_j)K_{ij} - \sum_{j=1}^{J}\frac{1}{2\sigma_j^2}\sum_{k=1}^{K_{ij}}\left[\Big(y_{ijk}-(\beta_{0j}+X_i^T(t_{ijk})\beta_j+Z_i^T(t_{ijk})b_{ij})\Big)^2\right];\\
&= -\log\Big(\sqrt{\det(2\pi\mathbf{\Sigma}_i)}\Big) - \frac{1}{2}\Big(y_i-\mathbf{X}_i\beta-\mathbf{Z}_ib_i\Big)^T\mathbf{\Sigma}_i^{-1}\Big(y_i-\mathbf{X}_i\beta-\mathbf{Z}_ib_i\Big)\\
\log\Big\{P(b_i;\theta)\Big\} &= -\log(\sqrt{\det(2\pi\mathbf{D})}) - \frac{1}{2}\Big(b_i^T\mathbf{D}^{-1}b_i\Big).
\end{aligned}$$

Let $H_0(T_i)=\int_0^{T_i}h_0(u)\mathrm{d}u$ be the cumulative baseline hazard of duration $(0,T_i)$, then

- Model I:

$$\begin{aligned}
P(T_i,\delta_i|b_i;\theta) &= \Big[h_0(T_i)\ \exp\big(V_i^T\gamma_0+b_i^T\gamma^I\big)\Big]^{\delta_i}\exp\Big[-H_0(T_i)\ \exp\big(V_i^T\gamma_0+b_i^T\gamma^I\big)\Big],\\
\log\Big\{P(T_i,\delta_i|b_i;\theta)\Big\} &= \delta_i\Big[\log h_0(T_i)+\big(V_i^T\gamma_0+b_i^T\gamma^I\big)\Big] - \ \exp\big(V_i^T\gamma_0+b_i^T\gamma^I\big)H_0(T_i);
\end{aligned}$$

- Model II:

$$\begin{aligned}
P(T_i,\delta_i|b_i;\theta) &= \Big[h_0(T_i)\ \exp\big(V_i^T\gamma_0+(\mathbf{Z}_ib_i)^T\gamma^{II}\big)\Big]^{\delta_i}\exp\Big[-H_0(T_i)\ \exp\big(V_i^T\gamma_0+(\mathbf{Z}_ib_i)^T\gamma^{II}\big)\Big],\\
\log\Big\{P(T_i,\delta_i|b_i;\theta)\Big\} &= \delta_i\Big[\log h_0(T_i)+\big(V_i^T\gamma_0+(\mathbf{Z}_ib_i)^T\gamma^{II}\big)\Big] - \ \exp\big(V_i^T\gamma_0+(\mathbf{Z}_ib_i)^T\gamma^{II}\big)H_0(T_i).
\end{aligned}$$

5.3.1.1 The first-order derivative of $\mathcal{K}(b_i)$ respect to b_i

- Model I:

$$\begin{aligned}
\frac{\partial\mathcal{K}(b_i)}{\partial b_i} &= \frac{\partial\Big[\log\Big\{P(T_i,\delta_i|b_i;\theta)\Big\}+\log\Big\{P(y_i|b_i;\theta)\Big\}+\log\Big\{P(b_i;\theta)\Big\}\Big]}{\partial b_i}\\
&= \delta_i\gamma^I-\gamma^I\ \exp\big(V_i^T\gamma_0+b_i^T\gamma^I\big)H_0(T_i)+\mathbf{\Sigma}_i^{-1}\Big(y_i-\mathbf{X}_i\beta-\mathbf{Z}_ib_i\Big)\mathbf{Z}_i^T-\mathbf{D}^{-1}b_i,
\end{aligned}$$

- Model II:

$$\begin{aligned}\frac{\partial \mathcal{K}(b_i)}{\partial b_i} &= \frac{\partial \Big[\log\Big\{P(T_i,\delta_i|b_i;\theta)\Big\} + \log\Big\{P(y_i|b_i;\theta)\Big\} + \log\Big\{P(b_i;\theta)\Big\}\Big]}{\partial b_i} \\ &= \delta_i \mathbf{Z}_i^T\gamma^{II} - \mathbf{Z}_i^T\gamma^{II}\exp\big(V_i^T\gamma_0 + (\mathbf{Z}_i b_i)^T\gamma^{II}\big)H_0(T_i) + \mathbf{\Sigma}_i^{-1}\Big(y_i - \mathbf{X}_i\beta - \mathbf{Z}_i b_i\Big)\mathbf{Z}_i^T - \mathbf{D}^{-1}b_i.\end{aligned}$$

5.3.1.2 The second-order derivative of $\mathcal{K}(b_i)$ respect to b_i

- Model I:

$$\begin{aligned}\mathbf{H}(b_i) = \frac{\partial^2 \mathcal{K}(b_i)}{\partial b_i \partial b_i^T} &= \frac{\partial^2 \Big[\log\Big\{P(T_i,\delta_i|b_i;\theta)\Big\} + \log\Big\{P(y_i|b_i;\theta)\Big\} + \log\Big\{P(b_i;\theta)\Big\}\Big]}{\partial b_i \partial b_i^T} \\ &= -\gamma^I(\gamma^I)^T\exp\big(V_i^T\gamma_0 + b_i^T\gamma^I\big)H_0(T_i) - \mathbf{\Sigma}_i^{-1}\mathbf{Z}_i^T\mathbf{Z}_i - \mathbf{D}^{-1};\end{aligned}$$

- Model II:

$$\begin{aligned}\mathbf{H}(b_i) = \frac{\partial^2 \mathcal{K}(b_i)}{\partial b_i \partial b_i^T} &= \frac{\partial^2 \Big[\log\Big\{P(T_i,\delta_i|b_i;\theta)\Big\} + \log\Big\{P(y_i|b_i;\theta)\Big\} + \log\Big\{P(b_i;\theta)\Big\}\Big]}{\partial b_i \partial b_i^T} \\ &= -\mathbf{Z}_i^T\gamma^{II}(\gamma^{II})^T\mathbf{Z}_i\exp\big(V_i^T\gamma_0 + b_i^T\mathbf{Z}_i^T\gamma^{II}\big)H_0(T_i) - \mathbf{\Sigma}_i^{-1}\mathbf{Z}_i^T\mathbf{Z}_i - \mathbf{D}^{-1}.\end{aligned}$$

5.3.2 Variable selection based on penalized likelihood

To simultaneously identify the important longitudinal risk factors and time-constant risk factors that have strong effect upon the time-to-event outcome, and the important time-varying risk factors that have strong effect upon each longitudinal outcome, we propose a penalized likelihood:

$$pl(\theta|\mathcal{D}_n) = -\frac{1}{n}l(\theta|\mathcal{D}_n) + p_1(\gamma) + p_2(\gamma_0) + p_3(\beta) + p_4(\mathbf{D}). \tag{5.5}$$

More specifically, the penalty $p_1(\gamma)$ is defined as

- Model I:

$$p_1(\gamma) = \lambda_1^{'} \sum_{j=1}^{J} \sqrt{R_j} ||\gamma_j||,$$

where $||.||$ is Euclidean norm and R_j is the number of elements of γ_j^I. It captures the group structure of b_i. Specifically, b_i can be divided into J nonoverlapping groups that correspond to J longitudinal outcomes. For example, $(b_{i11}, \ldots, b_{i1R_1})$ is one group that corresponds to the first longitudinal outcome. Therefore, it controls which longitudinal outcomes are selected. We assume that R_j are the same for $\forall j$, we can rewrite the penalty as

$$p_1(\gamma) = \lambda_1 \sum_{j=1}^{J} ||\gamma_j^I||,$$

where $\lambda_1 = \lambda_1^{'} \sqrt{R_j}$.

- Model II:

$$p_1(\gamma) = \lambda_1 \sum_{j=1}^{J} |\gamma_j^{II}|$$

that controls which one among J longitudinal outcomes are selected.

The penalty $p_2(\gamma_0)$ is

$$p_2(\gamma_0) = \lambda_2 \sum_{p=1}^{P} |\gamma_{0p}|$$

that controls the sparsity of the γ_0 so that the baseline covariates are selected.

The penalty $p_3(\beta)$ is

$$p_3(\beta) = \lambda_3 \sum_{j=1}^{J} \sum_{h=1}^{P_j} |\beta_{jh}|$$

that controls the sparsity of the β so that the time-varying covariates are selected.

The last penalty $p_4(\mathbf{D})$ is

$$p_4(\mathbf{D}) = \lambda_4 \sum_{l \neq m} |d_{lm}|,$$

where d_{lm} is the lm^{th} element of $\mathbf{D}$. It controls the sparsity of the variance-covariance matrix $\mathbf{D}$ to avoid the nonidentifiability problem of the joint model. $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are tuning parameters that control the degree of penalties.

The approximation of the penalized likelihood function is

- Model I:

$$pl(\theta|\mathcal{D}_n) \approx -\frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{K}(\tilde{b}_i) - \frac{1}{2}\log\big(\det\big(-\mathbf{H}(\tilde{b}_i)\big)\big) + J\log(2\pi)\right] + \lambda_1\sum_{j=1}^{J}||\gamma_j^I|| + \lambda_2\sum_{p=1}^{P}|\gamma_{0p}| \\ + \lambda_3\sum_{j=1}^{J}\sum_{h=1}^{P_j}|\beta_{jh}| + \lambda_4\sum_{l\neq m}|d_{lm}|;$$

- Model II:

$$pl(\theta|\mathcal{D}_n) \approx -\frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{K}(\tilde{b}_i) - \frac{1}{2}\log\big(\det\big(-\mathbf{H}(\tilde{b}_i)\big)\big) + J\log(2\pi)\right] + \lambda_1\sum_{j=1}^{J}|\gamma_j^{II}| + \lambda_2\sum_{p=1}^{P}|\gamma_{0p}| \\ + \lambda_3\sum_{j=1}^{J}\sum_{h=1}^{P_j}|\beta_{jh}| + \lambda_4\sum_{l\neq m}|d_{lm}|.$$

5.3.3 Algorithm for optimization of the penalized likelihood

To maximize the penalized likelihood function (5.5), we use the gradient descent and Newton-Raphson approaches. We first obtain the initial values $\hat{\beta}_0^{(0)}$, $\hat{\beta}^{(0)}$, $\hat{\mathbf{D}}^{(0)}$, and $\hat{\sigma_j}^{2(0)}$ for $j = 1, \ldots, J$ by fitting the multivariate linear mixed effects model on the longitudinal outcomes. The random effect values are then generated using the multivariate zero-mean Gaussian distribution with the variance-covariance matrix $\mathbf{D}$. Then, these estimates and values are included as time-varying and fixed covariates in the Cox's model to estimate $\hat{\gamma}_0^{(0)}$ and $\hat{\gamma}^{(0)}$.

At each iteration, we first solve $\tilde{b}_i = \underset{b_i}{\text{argmax}}\ \mathcal{K}(b_i)$ using the Newton-Raphson method:

$$\tilde{b}_i^{it+1} = \tilde{b}_i^{it} - \left(\frac{\partial^2\mathcal{K}(b_i)}{\partial b_i \partial b_i^T}\big(\tilde{b}_i^{it}\big)\right)^{-1}\frac{\partial\mathcal{K}(b_i)}{\partial b_i}\big(\tilde{b}_i^{it}\big),$$

where it is the iteration index. Then, compute the Laplace approximation of penalized log-likelihood using (5.5). Finally, each parameter is updated using Newton-Raphson method or gradient descent approaches or closed-form solution if the closed form exists. Updating steps for the parameters are iterated until convergence.

Algorithm 5 Coordinate descent algorithm

```
Input: Input $\mathcal{D}_n$
Output: Model parameters $\theta = \{\beta_0, \beta, \gamma_0, \gamma(\gamma^I \text{ or } \gamma^{II}), (\sigma_1^2, \ldots, \sigma_J^2), \mathbf{D}\}$
Initialize parameter vector $\theta^{(0)}$
repeat
    Obtain $\tilde{b}_i$ which maximizes $\mathcal{K}(b_i)$ with current estimates $\theta^{(it)}$
    Estimate $\theta^{(it+1)} = \underset{\theta}{\text{argmin}}\ pl(\theta|\mathcal{D}_n, \theta^{(it)}, \tilde{b}_i)$
    $it \rightarrow it + 1$
until Convergence;
```

5.3.3.1 *Initial values*

We used *lme()* function from **nlme** R package [118] and *coxph()* function from **survival** R package [44]. More specifically, we fitted each longitudinal model separately using *lme()*, then fitted a Cox's model including these results as time-varying covariates along with other covariates using *survival()*. Note that in the case the data are not balanced, i.e. when $t_{ijk} \neq t_{ik}$ for $\forall j$, we set $\hat{\gamma}^{(0)} = \mathbf{0}$.

5.3.3.2 *Estimates of parameters without penalties:* β_0 *and* $(\sigma_1^2, \ldots, \sigma_J^2)$

- β_0 a $J-$dimensional vector. The closed-form solution of j^{th} element β_{0j} is given by

$$\hat{\beta}_{0j} = \frac{1}{\sum_{i=1}^{n} K_{ij}} \sum_{i=1}^{n} \sum_{k=1}^{K_{ij}} \left[y_{ijk} - \left(X_i^T(t_{ijk})\beta_j + Z_i^T(t_{ijk})\tilde{b}_{ij} \right) \right].$$

- $(\sigma_1^2, \ldots, \sigma_J^2)$. The gradient respect to the j^{th} element, σ_j^2, is given by

$$\frac{\partial pl(\theta|\mathcal{D}_n)}{\partial \sigma_j} = -\frac{1}{n}\sum_{i=1}^{n}\Bigg[-\frac{K_{ij}}{\sigma_j} + \frac{1}{\sigma_j^3}\sum_{k=1}^{K_{ij}}\Big(y_{ijk} - (\beta_{0j} + X_i^T(t_{ijk})\beta_j + Z_i^T(t_{ijk})\tilde{b}_{ij})\Big)^2 \\ - \frac{1}{2}\text{tr}\bigg(\big(-\mathbf{H}(\tilde{b}_i)\big)^{-1}\Big[\frac{\partial\big(-\mathbf{H}(\tilde{b}_i)\big)}{\partial \sigma_j}\Big]\bigg)\Bigg],$$

where

$$\frac{\partial\big(-\mathbf{H}(\tilde{b}_i)\big)}{\partial \sigma_j} = -\frac{2}{\sigma_j^3}d\mathbf{H} = -\frac{2}{\sigma_j^3}\begin{bmatrix} 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \ldots & \ldots & d\mathbf{H}_j & \ldots \\ 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \end{bmatrix}$$

with

$$d\mathbf{H}_j = \sum_{k=1}^{K_{ij}} Z_i(t_{ijk})Z_i^T(t_{ijk}), \text{ for } j = 1, \ldots, J.$$

5.3.3.3 Estimates of parameters with penalties: β, γ_0 and γ

- β is a $J \times P_j$ matrix. The first- and second-derivatives respect to the jh^{th} element, β_{jh}, are

$$\frac{\partial pl(\theta|\mathcal{D}_n)}{\partial \beta_{jh}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K_{ij}}\left[\frac{1}{\sigma_j^2}\Big(y_{ijk} - (\beta_{0j} + X_i^T(t_{ijk})\beta_j + Z_i^T(t_{ijk})\tilde{b}_{ij})\Big)X_{ih}(t_{ijk})\right] + \lambda_3 \text{ sgn}(\beta_{jh}),$$

$$\frac{\partial^2 pl(\theta|\mathcal{D}_n)}{\partial \beta_{jh}\partial \beta_{jq}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K_{ij}}\left[\frac{1}{\sigma_j^2}\Big(-X_{ih}(t_{ijk})X_{iq}(t_{ijk})\Big)\right],$$

where $j = 1, \ldots, J$ and $h, q = 1, \ldots, P_j$.

- γ_0 is a $P-$dimensional vector. The gradient respect to the p^{th} element, γ_{0p}, is given by

- Model I:

$$\begin{aligned}\frac{\partial pl(\theta|\mathcal{D}_n)}{\partial \gamma_{0p}} =& -\frac{1}{n}\sum_{i=1}^{n}\bigg[\Big[\delta_i - \ \exp\big(V_i^T\gamma_0 + \tilde{b}_i^T\gamma^I\big)H_0(T_i)\Big]V_{ip} \\ & -\frac{1}{2}\text{tr}\bigg(\big(-\mathbf{H}(\tilde{b}_i)\big)^{-1}\Big[\gamma^I(\gamma^I)^T \ \exp\big(V_i^T\gamma_0 + \tilde{b}_i^T\gamma^I\big)H_0(T_i)\Big]V_{ip}\bigg)\bigg] + \lambda_2 \ \text{sgn}(\gamma_{0p});\end{aligned}$$

- Model II:

$$\begin{aligned}\frac{\partial pl(\theta|\mathcal{D}_n)}{\partial \gamma_{0p}} =& -\frac{1}{n}\sum_{i=1}^{n}\bigg[\Big[\delta_i - \ \exp\big(V_i^T\gamma_0 + \tilde{b}_i^T\mathbf{Z}_i^T\gamma^{II}\big)H_0(T_i)\Big]V_{ip} \\ & -\frac{1}{2}\text{tr}\bigg(\big(-\mathbf{H}(\tilde{b}_i)\big)^{-1}\Big[\mathbf{Z}_i^T\gamma^{II}(\gamma^{II})^T\mathbf{Z}_i \ \exp\big(V_i^T\gamma_0 + \tilde{b}_i^T\mathbf{Z}_i^T\gamma^{II}\big)H_0(T_i)\Big]V_{ip}\bigg)\bigg] \\ & + \lambda_2 \ \text{sgn}(\gamma_{0p}),\end{aligned}$$

where $p = 1, \ldots, P$.

- γ

- Model I: γ^I is a $\sum_{j=1}^{J} R_j-$dimensional vector whose j^{th} element is $\gamma_j^I = (\gamma_{jq}^I)$ where $j = 1, \ldots, J; q = 1, \ldots, R_j$. The gradient respect to γ_{jq}^I is

$$\begin{aligned}\text{If } \gamma_j^I \neq 0 :& \frac{\partial pl(\theta|\mathcal{D}_n)}{\partial \gamma_{jq}^I} = -\frac{1}{n}\sum_{i=1}^{n}\bigg[\bigg(\delta_i - \ \exp\big(V_i^T\gamma_0 + \tilde{b}_i^T\gamma^I\big)H_0(T_i)\bigg)\tilde{b}_{ijq} \\ & -\frac{1}{2}\text{tr}\bigg(\ \exp\big(V_i^T\gamma_0 + \tilde{b}_i^T\gamma^I\big)H_0(T_i)\big(-\mathbf{H}(\tilde{b}_i)\big)^{-1}\big(e_{jq}(\gamma^I)^T + \gamma^I e_{jq}^T + \tilde{b}_{ijq}\gamma^I(\gamma^I)^T\big)\bigg)\bigg] \\ & + \lambda_1\frac{\gamma_{jq}^I}{||\gamma_j^I||}, \\ \text{If } \gamma_j^I = 0 :& \frac{\partial pl(\theta|\mathcal{D}_n)}{\partial \gamma_{jq}^I} = -\frac{1}{n}\sum_{i=1}^{n}\bigg[\bigg(\delta_i - \ \exp\big(V_i^T\gamma_0 + \tilde{b}_i^T\gamma^I\big)H_0(T_i)\bigg)\tilde{b}_{ijq} \\ & -\frac{1}{2}\text{tr}\bigg(\ \exp\big(V_i^T\gamma_0 + \tilde{b}_i^T\gamma^I\big)H_0(T_i)\big(-\mathbf{H}(\tilde{b}_i)\big)^{-1}\big(e_{jq}(\gamma^I)^T + \gamma^I e_{jq}^T + \tilde{b}_{ijq}\gamma^I(\gamma^I)^T\big)\bigg)\bigg] \\ & + \lambda_1||\text{v}||,\end{aligned}$$

where e_{jq} is a $\sum_{j=1}^{J} R_j-$dimensional vector whose only element $R_j(j-1)+q$ is 1 and remaining elements are 0's and v is any vector satisfying $||\mathrm{v}|| \leq 1$.

– Model II: γ is a $J-$dimensional vector whose j^{th} element is γ_j. The gradient respect to γ_j^{II} is

$$\frac{\partial pl(\theta|\mathcal{D}_n)}{\partial \gamma_j^{II}} = -\frac{1}{n}\sum_{i=1}^{n}\Bigg[\bigg(\delta_i - \ \exp\big(V_i^T\gamma_0 + \tilde{b}_i^T\mathbf{Z}_i^T\gamma^{II}\big)H_0(T_i)\bigg)\mathbf{Z}_i\tilde{b}_i e_j$$

$$-\frac{1}{2}\mathrm{tr}\bigg(\ \exp\big(V_i^T\gamma_0 + \tilde{b}_i^T\mathbf{Z}_i^T\gamma^{II}\big)H_0(T_i)\big(-\mathbf{H}(\tilde{b}_i)\big)^{-1}\big(\mathbf{Z}_i\mathbf{Z}_i^T 2\gamma^{II}e_j + (\mathbf{Z}_i^T\gamma^{II}(\gamma^{II})^T\mathbf{Z}_i)(\mathbf{Z}_i\tilde{b}_i)e_j\big)\bigg)\Bigg]$$

$$+\lambda_1\mathrm{sgn}|\gamma_j^{II}|,$$

where e_j is a $J-$dimensional vector whose only element j is 1 and remaining elements are 0's.

5.3.3.4 *Estimate of a symmetric and positive definite matrix D*

To derive the solution of **D**:

$$\mathbf{D} = \underset{\mathbf{D}\succ\mathbf{0}}{\mathrm{argmin}}\, pl(\theta|\mathcal{D}_n) = \underset{\mathbf{D}\succ\mathbf{0}}{\mathrm{argmin}}\left(\mathcal{G}(\mathbf{D}) + \lambda_4\sum_{l\neq m}|d_{lm}|\right) = \underset{\mathbf{D}\succ\mathbf{0}}{\mathrm{argmin}}\left(\mathcal{G}(\mathbf{D}) + \lambda_4||\mathbf{D}||_{1,\mathrm{off}}\right),$$

we deploy an algorithm presented in [119] that uses the proximal gradient descent and alternating direction method of multipliers (ADMM) as shown in **Algorithm 6**.

```
Algorithm 6 ADMM algorithm
Input: Input $\mathcal{D}_n$
Output: Parameter $\mathbf{D}$
Initialize parameter $\mathbf{D}^{(0)}$
repeat
    Estimate
        $\mathbf{D}^{(k+1)} = \underset{\mathbf{Z}}{\text{argmin}}\ \frac{1}{2\tau_{k+1}}||\mathbf{Z} - (\mathbf{D}^{(k)} - \tau_{k+1}\nabla\mathcal{G}(\mathbf{D}^{(k)}))||_2^2 + \lambda_4||\mathbf{Z}||_{1,\text{off}}$
    Use the proximal gradient descent to obtain
        $\mathbf{D}^{(k+1)} = \text{prox}_{\tau_{k+1}}\big(\mathbf{D}^{(k)} - \tau_{k+1}\nabla\mathcal{G}(\mathbf{D}^{(k)})\big), k = 0, 1, 2, \ldots$
    Check if the minimum eigenvalue of $\mathbf{D}^{(k+1)}$ is below 0, then perform the optimization using
    the alternating direction method of multipliers (ADMM):
      • Decompose $\frac{\tau}{1+\tau\rho}\Bigg[\nabla\mathcal{G}(\mathbf{D}^{(k)}) - \mathbf{Y}^{(k)} + \rho\mathbf{Z}^{(k)}\Bigg] = \mathbf{U}\Lambda\mathbf{U}^T$
      • Update $\mathbf{D}^{(k+1)} = \mathbf{U}\Lambda_0\mathbf{U}^T$ where $\Lambda_0 = \text{diag}\{\max(\Lambda_{ii}, 0)\}$
      • Update $\mathbf{Z}^{(k+1)}$ where $z_{lm}^{(k+1)} = \text{soft}\big(\frac{1}{\rho}\mathbf{Y}_{lm}^{(k)} + \mathbf{D}_{lm}^{(k+1)}, \frac{\lambda_4}{\rho}\big)$
      • Update $\mathbf{Y}^{(k+1)} = \mathbf{Y}^{(k)} + \rho(\mathbf{D}^{(k+1)} - \mathbf{Z}^{(k+1)})$
until Convergence;
```

Here,

$$\nabla\mathcal{G}(\mathbf{D}) = \frac{\partial\mathcal{G}(\mathbf{D})}{\partial\mathbf{D}} = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\Bigg[\Big(- \mathbf{D}^{-1} + \mathbf{D}^{-1}b_ib_i^T\mathbf{D}^{-1} + \big(- \mathbf{H}(\tilde{b}_i)\big)^{-1}\mathbf{D}^{-1}\mathbf{D}^{-1}\Big) + \Big(- \mathbf{D}^{-1} + \mathbf{D}^{-1}b_ib_i^T\mathbf{D}^{-1} + \big(- \mathbf{H}(\tilde{b}_i)\big)^{-1}\mathbf{D}^{-1}\mathbf{D}^{-1}\Big)^T - \Big(- \mathbf{D}^{-1} + \mathbf{D}^{-1}b_ib_i^T\mathbf{D}^{-1} + \big(- \mathbf{H}(\tilde{b}_i)\big)^{-1}\mathbf{D}^{-1}\mathbf{D}^{-1}\Big) \circ \mathbf{I}\Bigg].$$

5.3.3.5 Convergence conditions

Two main conditions used to check the convergence of **Algorithm 5** are the absolute and relative differences in the likelihood, given by

$$\max\Big\{\big|pl(\hat{\theta}^{(it+1)}) - pl(\hat{\theta}^{(it)})\big|\Big\} < \epsilon_0,$$
$$\max\left\{\frac{\big|pl(\hat{\theta}^{(it+1)}) - pl(\hat{\theta}^{(it)})\big|}{\big|pl(\hat{\theta}^{(it)})\big|}\right\} < \epsilon_1.$$

Other conditions, the absolute and relative differences in the model parameters, are also used. They are defined as

$$\max\Big\{\big|\hat{\theta}^{(it+1)} - \hat{\theta}^{(it)}\big|\Big\} < \epsilon_2,$$
$$\max\left\{\frac{\big|\hat{\theta}^{(it+1)} - \hat{\theta}^{(it)}\big|}{\big|\hat{\theta}^{(it)}\big| + \epsilon_3}\right\} < \epsilon_4,$$

where $\epsilon_0, \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4$ are specified constants.

5.3.4 Hyperparameter tuning

To determine the tuning hyperparameters, we used the BIC-type criterion proposed by [120]. This criterion is used in [112], [114] and [113], which shows that it works well for the joint model. The BIC-type criterion is defined as

$$\text{BIC}_\lambda = -2l(\hat{\theta}) + \log(N) \times df_\lambda, \tag{5.6}$$

where $\hat{\theta}$ is the estimator vectors of model parameters for given $\lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$, N is the total number of observations, and df_λ is the total number of non-zero estimates of $\hat{\theta}$ as the degree of freedom.

We use the grid search strategy to select a set λ of tuning hyperparameters $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$. More specifically, it is done by taking all sets constructed by candidate vectors $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, then choosing a set λ that minimized the BIC criterion.

5.4 Simulation studies

In this section, we perform simulation studies to investigate the finite sample performance of the proposed model estimators. We consider four scenarios for each model.

We generate data from the longitudinal sub-model that is given as

$$y_{ijk}(t_{ijk}) = X_{i0}\beta_{i0j} + X_{i1}\beta_{i1j} + \beta_{0j} + b_{ij1} + (\beta_{1j} + b_{ij2})t_{ijk} + \epsilon_{ijk}, \tag{5.7}$$

where $\epsilon_{ijk} \overset{iid}{\sim} \mathcal{N}(0, \sigma_j^2)$ and $b_i = (b_{ij})_{j=1}^J = (b_{ij1}, b_{ij2})_{j=1}^J \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$ for $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, J$, and $k = 1, 2, \ldots, K_{ij}$. The survival sub-models as follows

- Model I:
$$h_i(t) = h_0(t)\exp\bigl(X_{i0}\gamma_{00} + X_{i1}\gamma_{01} + \sum_{j=1}^{J}(b_{ij1}\gamma_{j1}^I + b_{ij2}\gamma_{j2}^I)\bigr) \tag{5.8}$$

- Model II:
$$h_i(t) = h_0(t)\exp\bigl(X_{i0}\gamma_{00} + X_{i1}\gamma_{01} + \sum_{j=1}^{J}(b_{ij1} + b_{ij2}t_{ijk})\gamma_j^{II}\bigr) \tag{5.9}$$

We consider n individuals and $J = 3$ longitudinal outcomes. For each i^{th} individual, we consider two baseline covariates: a continuous covariate $X_{i0} \overset{iid}{\sim} \mathcal{N}(0, 1)$ and a binary covariate $X_{i1} \overset{iid}{\sim} Bin(1, 0.5)$. Their corresponding coefficients are $(\beta_{i01}, \beta_{i11}, \beta_{i02}, \beta_{i12}, \beta_{i03}, \beta_{i13}) = (1.5, 2, 0, 1, 1, 0)$ in longitudinal submodels, and $(\gamma_{00}, \gamma_{01}) = (1, 0)$ in survival submodels. The values of intercept and slope are $(\beta_{01}, \beta_{11}, \beta_{02}, \beta_{12}, \beta_{03}, \beta_{13}) = (1, 0, -1.5, 1.2, 1, -1.3)$. The variance-covariance matrix $\mathbf{D}$ is specified as follows: $D_{ij} = 0.5$ if $i = j$ and $D_{ij} = 0.2$ if $|i - j| = 1$. We set $\sigma_1^2 = 1, \sigma_2^2 = 1, \sigma_3^2 = 1.5$. In addition, the coefficient $\gamma_1^I = (0, 0), \gamma_2^I = (1.2, 1.5), \gamma_3^I = (-2, 1)$ and $\gamma_1^{II} = 0, \gamma_2^{II} = 1.2, \gamma_3^{II} = 1$.

Longitudinal observations, generated from Eq. (5.7), are collected according to a follow-up schedule of K_{ij} time points (at times $t_{ijk} = k$ where $k = 0, 1, \ldots, K_{ij}$) until death or censoring time. The average number of observations is 5. The baseline hazard function $h_0(t) = \exp(\alpha_0 + \alpha_1 t)$ with $\alpha_0 = 1$ and $\alpha_1 = -3.5$ is used. The event times are simulated from a Gompertz distribution following the methodology described by [121] using Eq. (5.8). Independent censoring times were drawn from an exponential distribution $C_i \overset{iid}{\sim}$ exp(r) with rate r.

5.4.1 Model I

We consider four scenarios:

- Scenario I: $n = 200, r = 0.01$
- Scenario II: $n = 200, r = 0.05$
- Scenario III: $n = 500, r = 0.01$
- Scenario IV: $n = 500, r = 0.05$

The average censoring percentage are around 5%, 15%, 5%, 15%, respectively. We run each scenario over 100 replications. The tuning parameter $\lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ for each dataset determined by searching from all sets constructed by candidate vectors $\lambda_1 = (0.01, \ldots, 0.20), \lambda_2 = (0.05, \ldots, 0.20), \lambda_3 = (0.1, \ldots, 2.5), \lambda_4 = (0.01, 0.05, \ldots, 4.5)$ to select the minimizer of the BIC criterion defined in Eq. (5.6).

The results are summarized in Tables 5.1, 5.2, 5.3. Table 5.1 shows the selection frequencies of fixed effects in longitudinal and survival submodels. The average selection frequencies of nonzero elements are more than 98% under all scenarios. The average selection frequencies of zero elements are around 5.6%. More specifically, when the sample size increases and the censoring rate decreases, the selection frequencies of nonzero elements increase and the selection frequencies of zero elements decreases. Table 5.2 presents the selection frequency of random effects in survival submodels. The average rates of correct selection are approximately 96% for non-zero effects and 98% for zero effects. The selection frequencies of elements improve when the sample size increase and the censoring rate decreases. Furthermore, the results of the selection frequency of elements of variance-covariance matrix presented in Table 5.3 show that the average true positive is 95.9% and the average true negative is around 99.7%. Overall, our proposed methods perform well under those scenarios.

	Longitudinal submodel									Survival submodel	
	X_{i0}	X_{i1}	X_{i0}	X_{i1}	X_{i0}	X_{i1}	t	t	t	X_{i0}	X_{i1}
	β_{i01}	β_{i11}	β_{i02}	β_{i12}	β_{i03}	β_{i13}	β_{11}	β_{12}	β_{13}	γ_{00}	γ_{01}
Truth	1.5	2	0	1	1	0	0	1.2	-1.3	1	0
Scenario I	100	100	0	98	100	8	20	100	100	100	0
Scenario II	100	100	1	98	100	12	20	100	100	100	0
Scenario III	100	100	0	99	100	2	4	100	100	100	0
Scenario IV	100	100	5	100	100	5	12	100	100	100	0

Table 5.1: Selection frequency of fixed effects in longitudinal and survival submodels of Model I.

	γ_1^I		γ_2^I		γ_3^I	
	γ_{11}^I	γ_{12}^I	γ_{21}^I	γ_{22}^I	γ_{31}^I	γ_{32}^I
Truth	0	0	1.2	1.5	-2	1
Scenario I	1		96		92	
Scenario II	5		90		90	
Scenario III	0		100		100	
Scenario IV	2		98		99	

Table 5.2: Selection frequency of random effects in survival submodels of Model I.

	True positive	True negative
Scenario I	97.5	99.8
Scenario II	90.4	99
Scenario III	100	99.8
Scenario IV	95.8	100

Table 5.3: Selection frequency of variance-covariance matrix of Model I.

5.4.2 Model II

We consider four scenarios:

- Scenario I: $n = 200, r = 0.01$
- Scenario II: $n = 200, r = 0.05$
- Scenario III: $n = 300, r = 0.01$
- Scenario IV: $n = 300, r = 0.05$

The average censoring percentage are around 20%, 30%, 20%, 30%, respectively. We run each scenario over 100 simulations. The tuning parameter $\lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ for each dataset determined by searching from all sets constructed by candidate vectors $\lambda_1 = (0.1, \ldots, 0.5), \lambda_2 = (0.02, \ldots, 0.1), \lambda_3 = (0.25, \ldots, 0.75), \lambda_4 = (0.1, \ldots, 10)$ to select the minimizer of the BIC criterion defined in Eq. (5.6).

The results using our proposed model - Model II and the un-penalized models in R package joineRML are summarized in Tables 5.4, 5.5, 5.6. For joineRML model, we use p-values to select the significant risk factors at the 0.05 significance level. Table 5.4 shows the selection frequencies of fixed effects in longitudinal and survival submodels. The average selection frequencies of non-zero elements are 99.5% for Model II while 98.5% for joineRML model. The average selection frequency of zero elements of Model II is 2.8% while joineRML model is 2.5%. Table 5.5 presents the selection frequency of random effects in survival submodels. Under all scenarios, Model II selects correctly more than 99% for non-zero effects and 98% for zero effects while joineRML model selects correctly more than 98% for non-zero effects and 96% for zero effects. Furthermore, the results of the selection frequency of elements of variance-covariance matrix presented in Table 5.6 show that the average true positive of joineRML model is a bit better than Model II, but the average true negative of joineRML model is much worse than Model II. Overall, our proposed methods perform better than the un-penalized model in term of variable selection.

		Longitudinal submodel									Survival submodel	
		X_{i0}	X_{i1}	X_{i0}	X_{i1}	X_{i0}	X_{i1}	t	t	t	X_{i0}	X_{i1}
		β_{i01}	β_{i11}	β_{i02}	β_{i12}	β_{i03}	β_{i13}	β_{11}	β_{12}	β_{13}	γ_{00}	γ_{01}
Models	Truth	1.5	2	0	1	1	0	0	1.2	-1.3	1	0
Model II	Scenario I	98	98	2	100	100	0	2	100	100	100	0
	Scenario II	100	99	5	100	98	7	8	100	100	100	0
	Scenario III	100	100	4	98	100	1	3	99	100	100	0
	Scenario IV	100	100	4	96	100	5	4	100	100	100	0
joineRML	Scenario I	99	99	2	99	99	5	1	99	99	99	4
	Scenario II	98	98	2	98	98	4	1	98	98	98	4
	Scenario III	99	99	1	99	99	5	0	99	99	99	4
	Scenario IV	98	98	1	98	98	3	1	98	98	98	3

Table 5.4: Selection frequency of fixed effects in longitudinal and survival submodels of Model II.

		γ_1^{II}	γ_2^{II}	γ_3^{II}
Models	Truth	0	1.2	1
Model II	Scenario I	0	100	100
	Scenario II	2	99	99
	Scenario III	1	100	100
	Scenario IV	1	100	99
joineRML	Scenario I	4	99	99
	Scenario II	4	98	98
	Scenario III	2	99	99
	Scenario IV	2	98	98

Table 5.5: Selection frequency of random effects in survival submodels of Model II.

Models		True positive	True negative
Model II	Scenario I	92.8	99
	Scenario II	88.4	99.6
	Scenario III	94.3	99.8
	Scenario IV	91.2	99
joineRML	Scenario I	99.7	9.6
	Scenario II	99.6	10.2
	Scenario III	100	11.7
	Scenario IV	100	11.8

Table 5.6: Selection frequency of variance-covariance matrix of Model II.

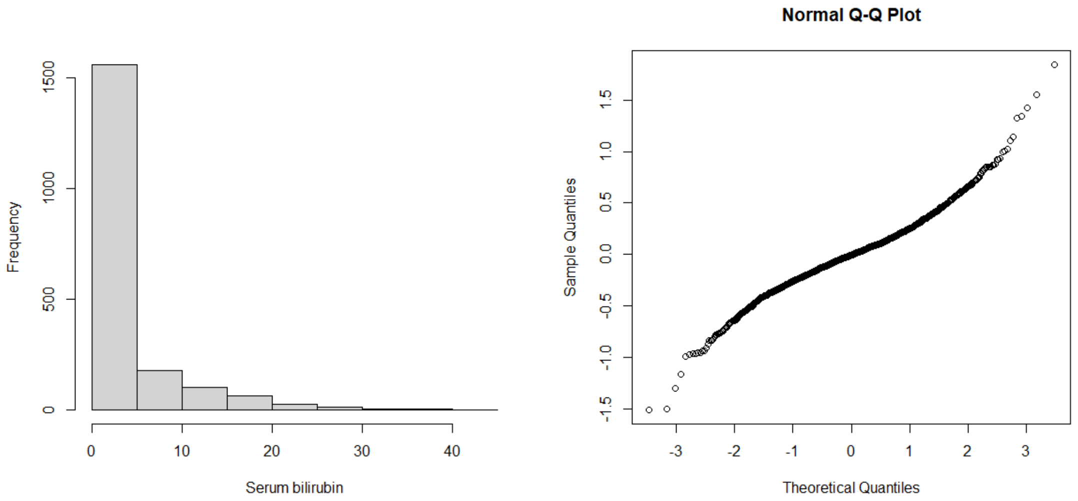


Figure 5.1: Histogram and Q-Q plot of serum bilirubin

5.5 Real-world case studies

We illustrate the use of the proposed methods by applying to the primary biliary cirrhosis (PBC) dataset of 312 patients who were enrolled between January 1974 and May 1984 in Mayo Clinic. The details of this dataset can be found in [122]. It is available in the R package joineRML [115]. In short, PBC is a chronic liver disease in which the bile ducts in the liver are damaged, which leads to cirrhosis and even mortality. Of 312 patients, $n = 158$ were randomized to receive D-penicillamine and $n = 154$ were assigned a placebo. We analyze 304 patients after excluding eight observations with missing values.

We investigate the relationship between longitudinal outcomes and survival time, and identify important covariates that have significant effects upon longitudinal outcomes and survival time. The survival time is the number of years between registration and the earlier of death, transplantation, or study analysis time. We consider three longitudinal outcomes: serum bilirunbin, serum albumin, and platelets.

Because the histogram of serum bilirubin is right skewed, log(serum bilirubin) is used before integrating into the joint models. Moreover, the Q-Q plot for residuals using *lme()* function in Figure 5.1 confirms that the log-transformation is reasonable. We model these three longitudi-

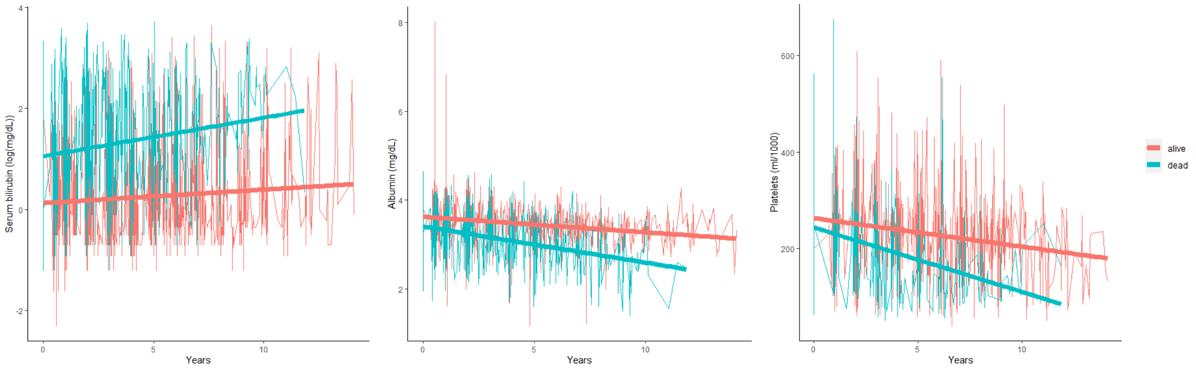


Figure 5.2: Longitudinal trajectory plots

nal processes using a linear model with random intercept and slope. We include one continuous variable - age (year) and one baseline binary variable - sex (1=male, 0=female).

Figure 5.2 displays all patients' longitudinal trajectories against time (years) across died and censoring status. They show the distinct differences between the general trends of these two cohorts. For example, the general decreasing trend of platelets is much sharper in patients who died than alive patients. This indicates the negative association between albumin and survival time in which the lower level of albumin associates with survival time. Therefore, we include these three longitudinal outcomes as potential risk factors. Therefore, we consider the joint models with the following longitudinal and survival sub-models:

$$\begin{aligned} log(serBilir) &= \beta_{i01}\text{age}_i + \beta_{i11}\text{sex}_i + (\beta_{01} + b_{i11}) + (\beta_{11} + b_{i12})\text{year} + \epsilon_{i1k}, \\ albumin &= \beta_{i02}\text{age}_i + \beta_{i12}\text{sex}_i + (\beta_{02} + b_{i21}) + (\beta_{12} + b_{i22})\text{year} + \epsilon_{i2k}, \\ platelet &= \beta_{i03}\text{age}_i + \beta_{i13}\text{sex}_i + (\beta_{03} + b_{i31}) + (\beta_{13} + b_{i32})\text{year} + \epsilon_{i3k}, \\ & b_i \sim \mathcal{N}_6(0, \mathbf{D}) \text{ and } \epsilon_{ijk} \sim \mathcal{N}(0, \epsilon_j^2) \text{ for } j = 1, 2, 3 \end{aligned}$$

and

- Model I:

$$h_i(t) = h_0(t)\text{exp}(\gamma_{00}\text{age}_i + \gamma_{01}\text{sex}_i + \text{SR}),\ \text{where}$$
$$\text{SR} = (\gamma^I_{11}b_{i11} + \gamma^I_{12}b_{i12}) + (\gamma^I_{21}b_{i21} + \gamma^I_{22}b_{i22}) + (\gamma^I_{31}b_{i31} + \gamma^I_{31}b_{i32})$$

- Model II:

$$h_i(t) = h_0(t)\text{exp}(\gamma_{00}\text{age}_i + \gamma_{01}\text{sex}_i + \text{SR}(t)),\ \text{where } t \text{ is year}$$
$$\text{SR}(t) = \gamma^{II}_1(b_{i11} + b_{i12}\text{year}) + \gamma^{II}_2(b_{i21} + b_{i22}\text{year}) + \gamma^{II}_3(b_{i31} + b_{i32}\text{year})$$

The tuning parameters $\lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ are determined by searching from all sets constructed by candidate vectors $\lambda_1 = (0.3, 0.35, 0.4), \lambda_2 = (0.1, 0.15, 0.2), \lambda_3 = (1.5, 2), \lambda_4 = (6.5, 7.5)$ to select the combination with the minimum BIC value. The results of the fitted model for Model I at $\lambda = \{0.35, 0.1, 1.5, 7.5\}$ are presented in Table 5.7 and those of Model II at $\lambda = \{0.4, 0.2, 1.5, 7.5\}$ are presented in Table 5.8. The fitted model results using joineRML is also included in Table 5.8.

The results in Tables 5.7 and 5.8 show that all the models indeed give similar results. More specifically, sex has no effect upon all longitudinal outcomes and the survival hazard rate in the joineRML model, but it has positive association with the log serum bilirubin in our Model I and Model II; clearly, the serum bilirubin is higher for female than male. In contrast, age has no effect upon all longitudinal outcomes in Model I and Model II while it has negative association with the trajectory of albumin in the joineRML model. In the survival submodel of all these three models, age has positive association with the survival hazard rate. In addition, the subject-specific random deviation from the population trajectory of serum bilirubin has positive association with the survival hazard rate of all models; clearly, the subject-specific increase from the general trajectory of serum bilirubin has significant effect on the survival hazard rate. Moreover, Model II and joineRML both select one more longitudinal outcome - albumin. The subject-specific ran-

Parameter	Estimate
β_{i01}	0
β_{i11}	0.704 ± 0.105
β_{01}	0.824 ± 0.110
β_{11}	0.051 ± 0.010
β_{i02}	0
β_{i12}	0
β_{02}	3.910 ± 0.118
β_{12}	-0.019 ± 0.001
β_{i03}	0
β_{i13}	0
β_{03}	0.435 ± 0.081
β_{13}	-0.072 ± 0.005
γ_{00}	0.048 ± 0.002
γ_{01}	0
γ_1^I	$(0.015, 0.009) \pm (0.002, 0.000)$
γ_2^I	**0**
γ_3^I	**0**

Table 5.7: Results of the PBC data analysis of Model I.

Models	Model II	joineRML	
Parameter	Estimate	Estimate	$p-$value
β_{i01}	0	-0.001 ± 0.006	0.919
β_{i11}	0.704 ± 0.105	-0.171 ± 0.260	0.512
β_{01}	0.824 ± 0.110	0.698 ± 0.418	0.095
β_{11}	0.051 ± 0.010	0.153 ± 0.012	< 0.0001
β_{i02}	0	-0.008 ± 0.002	0.001
β_{i12}	0	-0.141 ± 0.081	0.082
β_{02}	3.911 ± 0.120	4.027 ± 0.150	< 0.0001
β_{12}	-0.019 ± 0.001	-0.096 ± 0.005	< 0.0001
β_{i03}	0	-0.010 ± 0.006	0.079
β_{i13}	0	0.294 ± 0.226	0.193
β_{03}	0.435 ± 0.081	0.557 ± 0.394	0.157
β_{13}	-0.072 ± 0.005	-0.130 ± 0.013	< 0.0001
γ_{00}	0.046 ± 0.003	0.064 ± 0.016	< 0.0001
γ_{01}	0	-0.268 ± 0.545	0.623
γ_1^{II}	0.057 ± 0.005	0.799 ± 0.185	< 0.0001
γ_2^{II}	-0.033 ± 0.001	-2.757 ± 0.629	< 0.0001
γ_3^{II}	0	-0.299 ± 0.182	0.101

Table 5.8: Results of the PBC data analysis of Model II and joineRML.

dom deviation from the population trajectory of albumin has negative association with the survival hazard rate. Advanced age, high serum bilirubin level, and low serum albumin have been found as prognostic factors in PBC in the literature and they were included in different mathematical prognostic models of survival analysis on the PBC dataset [123, 124, 125, 122, 126, 127, 128, 129, 130, 131, 132]. Here, we have analyzed the PBC data with the reasonable sample size, 304 patients with 1113 observations, and have considered only three longitudinal outcomes. Model II and joineRML, as we expected, give the same variable selection results. Moreover, their results are better than Model I since three longitudinal trajectory functions simply include random intercepts and slopes. When the longitudinal trajectory functions are more complex, models with time-dependent association (Model II and joineRML) would be better choices at the cost of more intensive and challenging computation.

In addition, regarding computational performance, we would like to mention that we use the same tolerance values for joineRML and Model II. More specifically, for joineRML, we use its default setting except the number of burn-in iteration 400K with the number of longitudinal outcomes $K = 3$ while for Model II, the search grid is set as described previously. The computation time of running joineRML one time for the case study is about 20 minutes, whereas the computation time of training our Model II is about 1.5 minutes on average for one set of lambdas and 54 minutes for the whole search grid. However, in the studies where the number of training samples, the number of longitudinal outcomes, or the number of longitudinal observations increases, the computational burden of joineRML is much heavier than our penalized joint models, including Model II.

6. CONCLUSIONS & FUTURE RESEARCH

High-dimensional, large-scale datasets are increasingly collected thanks to the advanced data collection and storage capacities. To mine useful information from this flood of data requires novel statistical models and computation methods. This thesis has introduced several models and tools for estimation, identification and prediction in the context of penalized methods that have attracted great attention in recent years.

In Chapter 3, the high-dimensional problems for survival data, in which P exceeds N, are addressed. Introducing the additional structures into these problems especially group structures, is natural for incorporating prior knowledge to achieve robust and interpretable survival models. This chapter has presented three group selection methods for high-dimensional data with censoring in the framework of the Cox's proportional hazards model. The proposed methods have been demonstrated in solving problems of both non-overlapping group and overlapping group cases. The group-wise descent algorithms combining with the MM approach have been developed to solve the corresponding optimization problems. Thanks to the MM approach, the proposed algorithms have a proven descent property. Several computational tricks have been implemented to speed up the group-wise descent algorithms, including the screening, active set, and warm-start approaches. An open-access implementation can be found in our R package **grpCox**. The simulation studies indicate that these methods perform well in term of variable selection. Moreover, the group lasso enjoys its convexity but it tends to select a model that is more complicated than the underlying model. It leads to relatively high false positive group selection rates. On the other hand, the nonconvex penalties, including group SCAD and group MCP, show the promising grouped variable selection results with oracle properties. We have analyzed the TCGA ovarian cancer data and breast cancer data using available pathway information to construct gene groups. The selected genes have been tested on independent ovarian cancer and breast cancer datasets. The results show that the high and low risk groups are well separated. In other words, group SCAD and group MCP methods are powerful alternatives to the group lasso Cox's model for grouped variable selection.

In Chapter 4, we propose the L1-regularized multi-state model framework for simultaneous parameter estimation and variable selection using the L1-regularized partial likelihood approach. We devise the one-step coordinate descent algorithm and use a local quadratic approximation of the log-partial likelihood to solve the corresponding optimization problem, which can offer significant improvement on the computational efficiency. Our proposed method demonstrates the state-of-the-art performance in terms of identifying the significant risk factors comparing with the existing regularized multi-state models in simulation studies. It also performs better at doing variable selection and predicting the transition probabilities in cases with small sample sizes comparing with the un-regularized approach in simulation and real-world cases.

Despite an increasing attention to the joint model of multivariate longitudinal and survival data, there exists no variable selection tools to practitioners. In Chapter 5, we focus on developing the penalized multivariate joint model framework to identify the important longitudinal outcomes that have strong associations with the time-to-event outcome, and simultaneously select the relevant covariates for both longitudinal and time-to-event outcomes of interest. In particular, we propose penalized joint models that consist of different types of penalties for different association structures. The estimation procedures based on Laplace approximation are used to tackle the high-dimensional problem. From the simulations, we find that our proposed models perform well in term of variable selection. The effectiveness of the proposed framework was also demonstrated for the application of data of patients with a chronic liver disease.

Our works presented in this dissertation can be extended in many ways. Three group penalties presented in Chapter 3 can be extended for more complicated time-varying models for both longitudinal and survival data analyses. In addition, we used a Cox model for each transition when specifying the current multi-state models in Chapter 4, which can be extended to other types of dynamic models for each transition. In Chapter 5, we have covered two different association structures between the longitudinal submodel and the time-to-event submodel; however, the association structures might take different forms [109], or combination of multiple structures in which separate longitudinal outcomes may have different forms [133]. The linear trajectories of the longitudinal outcomes can be extended to more general parametrized models, for example with different basis function expansions, including splines, which often comes with additional requirements on both computation and training sample size. Moreover, generalizing the longitudinal outcomes to the generalized linear mixed effects framework to accommodate categorical and count data outcomes is also desirable. Finally, our current models include only a single time-to-event time, it might be of interest to extend to incorporate multi-state models that would provide more flexible multivariate joint model framework. Last but not least, model selection often needs to be tied to the key questions for the corresponding biomedical applications. The developed methods in this dissertation may have addressed some of the challenges when analyzing longitudinal and survival data. The aforementioned extensions have to be studied carefully based on different applications with real-world considerations, in particular considering potential data quality and model uncertainty issues.

REFERENCES

[1] X. Dang, S. Huang, and X. Qian, “Penalized cox’s proportional hazards model for high-dimensional survival data with grouped predictors,” *Statistics and Computing*, vol. 31, no. 6, pp. 1–27, 2021.

[2] X. Dang, S. Huang, and X. Qian, “Risk factor identification in heterogeneous disease progression with l1-regularized multi-state models,” *Journal of Healthcare Informatics Research*, vol. 5, no. 1, pp. 20–53, 2021.

[3] D. R. Cox, “Regression Models and Life-Tables,” *Journal of the Royal Statistical Society Series B*, vol. 34, no. 1, pp. 187–220, 1972.

[4] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[5] H. Zhang and W. Lu, “Adaptive lasso for cox’s proportional hazards model.,” *Biometrika*, vol. 94, no. 3, pp. 691—-703, 2007.

[6] L. Wang, G. Chen, and H. Li, “Group scad regression analysis for microarray time course gene expression data.,” *Bioinformatics*, vol. 23, no. 12, pp. 1486–1494, 2007.

[7] J. Huang, P. Breheny, and S. Ma, “A selective review of group selection in high-dimensional models.,” *Statistical Science*, vol. 27, no. 4, pp. 481–499, 2012.

[8] L. Jacob, G. Obozinski, and J. Vert, “Group lasso with overlap and graph lasso.,” Proceedings of the 26th, (Montreal, Canada), International Conference on Machine Learning, 2009.

[9] G. Obozinski, L. Jacob, and J. Vert, “Group lasso with overlaps: the latent group lasso approach.,” *arXiv*, 2011.

[10] K. Lange, D. Hunter, and I. Yang, "Optimization transfer using surrogate objective functions (with discussion).," *Journal of Computational and Graphical Statistics*, vol. 9, no. 1, pp. 1–20, 2000.

[11] D. Hunter and K. Lange, "A tutorial on mm algorithms.," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.

[12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, p. 1–122, 2011.

[13] R. Tibshirani, "The lasso method for variable selection in the cox model," *Statistics in Medicine*, vol. 16, no. 4, pp. 385–395, 1996.

[14] J. Gui and H. Li, "Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data," *Biofinformatics*, vol. 21, no. 13, pp. 3001–3008, 2005.

[15] M. Y. Park and T. Hastie, "Penalized logistic regression for detecting gene interactions," Tech. Rep. Tech report, Stanford University, United States, 2006.

[16] H. Zou, "A note on path-based variable selection in the penalized proportional hazards model.," *Biometrika*, vol. 95, no. 1, pp. 241–247, 2008.

[17] J. Fan and R. Li, "Variable selection for cox's proportional hazards model and frailty model.," *The Annals of Statistics*, vol. 6, pp. 74–99, 2002.

[18] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables.," *Journal of the Royal Statistical Society Series B (Methodological)*, vol. 68, no. 1, pp. 49–67, 2006.

[19] L. Meir, S. Van de Geer, and P. Buhlmann, "The group lasso for logistic regression.," *Journal of the Royal Statistical Society Series B (Methodological)*, vol. 70, no. 1, pp. 53–71, 2008.

[20] P. Zhao, G. Rocha, and B. Yu, “The composite absolute penalties family for grouped and hierarchical variable selection.,” *The Annals of Statistics*, vol. 37, no. 6A, pp. 3468—-3497, 2009.

[21] P. Breheny and J. Huang, “Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors.,” *Stat Comput*, vol. 25, pp. 173–187, 2015.

[22] A. Puig, A. Wiesel, G. Fleury, and A. Hero, “Multidimensional shrinkage-thresholding operator and group lasso penalties.,” *IEEE Signal Processing Letter*, vol. 18, pp. 363–366, 2011.

[23] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, “A sparse-group lasso.,” *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 231–245, 2013.

[24] Y. Yang and H. Zou, “A fast unified algorithm for solving group-lasso penalize learning problems.,” *Stat Comput*, vol. 25, pp. 1129–1141, 2015.

[25] S. Ma, X. Song, and J. Huang, “Supervised group lasso with applications to microarray data analysis.,” *BMC Bioinformatics*, vol. 8, pp. 60–76, 2007.

[26] T. Wu and S. Wang, “Doubly regularized cox regression for high-dimensional survival data with group structures.,” *Satistics and its Interface*, vol. 6, pp. 175–186, 2013.

[27] S. Belhechmi, R. De Bin, F. Rotolo, and S. Michiels, “Accounting for grouped predictor variables or pathways in high-dimensional penalized Cox regression models,” *BMC Bioinformatics*, vol. 21, no. 277, 2020.

[28] L. Wang, H. Li, and J. Huang, “Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements.,” *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1556–1569, 2008.

[29] Y. Kim, J. Kim, and Y. Kim, “Blockwise sparse regression.,” *Statistica Sinica*, vol. 16, pp. 375–390, 2006.

[30] N. Simon and R. Tibshiran, "Standardization and the group lasso penalty.," *Stat Sin*, vol. 22, pp. 983–1001, 2011.

[31] J. Mairal and B. Yu, "Complexity analysis of the lasso regularization path.," 2012.

[32] N. Simon, "Regularization paths for coxś proportional hazards model via coordinate descent.," *Journal of Statistical Software*, vol. 39, no. 5, pp. 53–66, 2012.

[33] P. J. Verweij and H. C. Houwelingen, "Cross-validation in survival analysis.," *Statistics in Medicine*, vol. 12, no. 24, pp. 385–395, 1993.

[34] N. Ternes, F. Rotolo, and S. Michiels, "Empirical extensions of the lasso penalty to reduce the false discovery rate in high-dimensional Cox regression models," *Statistics in Medicine*, vol. 35, no. 15, pp. 2561–73, 2016.

[35] R. Jenatton, G. Mairal, G. Obozinski, and F. Bach, "Proximal methods for hierarchical sparse coding.," *Journal of Machine Learning Research*, vol. 12, pp. 2297–2334, 2011.

[36] X. Dang, *grpCox: Penalized Cox Model for High-Dimensional Data with Grouped Predictors*, 2020. R package version 1.0-1.

[37] C. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[38] Y. Zeng and P. Breheny, "Overlapping group logistic regression with applications to genetic pathway selection," *Cancer Informatics*, vol. 15, pp. 179–187, 2016.

[39] "Molecular signatures database v7.4." `https://www.gsea-msigdb.org/gsea/msigdb/collections.jsp`, 2021.

[40] S. Jones, X. Zhang, D. W. Parsons, J. C. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, H. Kamiyama, A. Jimeno, S. M. Hong, B. Fu, M. T. Lin, E. S. Calhoun, M. Kamiyama, K. Walter, T. Nikolskaya, Y. Nikolsky, J. Hartigan, D. R. Smith, M. Hidalgo, S. D. Leach, A. P. Klein, E. M. Jaffee, M. Goggins, A. Maitra, C. IacobuzioDonahue, J. R. Eshleman, S. E. Kern, R. H. Hruban, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vo-

gelstein, V. E. Velculescu, and K. W. Kinzler, “Core signaling pathways in human pancreatic cancers revealed by global genomic analyses.,” *Science*, vol. 321, pp. 1801–1806, 2008.

[41] M. J. Van de Vijer, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. Voskuil, G. J. Schreiber, J. L. Peterse, R. CW, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, van der Velde TW, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards, “A gene-expression signature as a predictor of survival in breast cancer.,” *The New England Journal of Medicine*, vol. 347, no. 25, pp. 1999–2009, 2002.

[42] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.,” *Proc Natl Acad Sci USA*, vol. 102, no. 43, pp. 15545–50, 2005.

[43] K. Blighe and J. Lasky-Su, “Regparallel: Standard regression functions in r enabled for parallel processing over large data-frames,” 2021.

[44] T. M. Therneau, *A Package for Survival Analysis in R*, 2021. R package version 3.2-11.

[45] S. Holm, “A simple sequentially rejective multiple test procedure,” *Scandinavian Journal of Statistics*, no. 6, pp. 65–70, 1979.

[46] Y. Hochberg, “A sharper bonferroni procedure for multiple tests of significance,” *Biometrika*, no. 75, p. 800–80, 1988.

[47] G. Hommel, “A stagewise rejective multiple test procedure based on a modified bonferroni test,” *Biometrika*, no. 75, p. 383–386, 1988.

[48] Y. Benjamini and Y. Hochberg, “Controlling the false discovery rate: a practical and powerful approach to multiple testing,” *Journal of the Royal Statistical Society Series B*, no. 57, p. 289–300, 1995.

[49] Y. Benjamini and D. Yekutieli, “The control of the false discovery rate in multiple testing under dependency,” *Annals of StatisticsB*, no. 29, p. 1165–1188, 2001.

[50] S. Hänzelmann, R. Castelo, and J. Guinney, “GSVA: gene set variation analysis for microarray and RNA-Seq data,” *BMC Bioinformatics*, vol. 14, no. 7, 2013.

[51] A. McCormick, P. Donoghue, M. Dixon, R. O’Sullivan, R. O’Donnell, J. Murray, A. Kaufmann, N. Curtin, and R. Edmondson, “Ovarian cancers harbour defects in non-homologous end joining resulting in resistance to rucaparib.,” *Clin Cancer Res*, vol. 23, no. 8, pp. 2050–2060, 2017.

[52] M. E. Gee, Z. Faraahi, A. McCormick, and R. Edmondson, “Dna damage repair in ovarian cancer: unlocking the heterogeneity.,” *Journal of Ovarian Research*, vol. 11, no. 50, 2018.

[53] J. Szkandera, T. Kiesslich, J. Haybaeck, A. Gerger, and M. Pichler, “Hedgehog signaling pathway in ovarian cancer,” *International journal of molecular science*, vol. 14, no. 1, p. 1179–1196, 2013.

[54] A. Otsuka, A. de Paolis, and G. P. Tocchini-Valentini, “Ribonuclease "xlai," an activity from xenopus laevis oocytes that excises intervening sequences from yeast transfer ribonucleic acid precursors,” *Molecular and cellular biology*, vol. 1, no. 3, p. 269–28, 1981.

[55] T. Gatcliffe, B. Monk, K. Planutis, and R. Holcombe, “Wnt signaling in ovarian tumorigenesis,” *Int J Gynecol Cancer*, vol. 18, pp. 954–962, 2008.

[56] E. Alsina-Sanchis, A. Figueras, A. Lahiguera, A. Vidal, O. Casanovas, M. Graupera, A. Villanueva, and F. Vinals, “The tgf pathway stimulates ovarian cancer cell proliferation by increasing igf1r levels.,” *Int J Cancer*, vol. 139, no. 8, pp. 1894–903, 2016.

[57] E. Alsina-Sanchis, A. Figueras, A. Lahiguera, M. Gil-Martin, B. Pardo, J. Piulats, L. Marti, J. Ponce, X. Matias-Guiu, A. Vidal, A. Villanueva, and F. Vinals, “Tgf controls ovarian cancer cell proliferation.,” *Int J Mol Sci*, vol. 18, no. 8, 2017.

[58] M. Ahmed and N. Rahman, “Atm and breast cancer susceptibility,” *Oncogene*, vol. 25, no. 43, pp. 5906–11, 2006.

[59] D. E. Goldgar, S. Healey, and J. G. Dowty, “Rare variants in the atm gene and risk of breast cancer,” *Breast Cancer Res*, vol. 13, no. 4, pp. R73–R73, 2011.

[60] D. Sarrio, S. M. Rodriguez-Pinilla, D. Hardisson, A. Cano, G. Moreno-Bueno, and J. Palacios, "Epithelial-mesenchymal transition in breast cancer relates to the basal-like phenotype," *Cancer Res*, vol. 68, no. 4, pp. 989–997, 2008.

[61] Y. Li, F. Chao, and B. Huang, "Hoxc8 promotes breast tumorigenesis by transcriptionally facilitating cadherin-11 expression," *Oncotarget*, vol. 5, no. 9, pp. 2596–607, 2014.

[62] S. Assefnia, S. Dakshanamurthy, J. M. Guidry-Auvil, C. Hampel, P. Z. Anastasiadis, B. Kallakury, A. Uren, D. W. Foley, M. L. Brown, L. Shapiro, M. Brenner, D. Haigh, and S. Byers, "Cadherin-11 in poor prognosis malignancies and rheumatoid arthritis: common target, common therapies," *Oncotarget*, vol. 5, no. 6, pp. 1458–74, 2014.

[63] P. K. Sengupta, E. M. Smith, K. Kim, M. J. Murnane, and B. D. Smith, "Dna hypermethylation near the transcription start site of collagen alpha2(i) gene occurs in both cancer cell lines and primary colorectal cancers," *Cancer research*, vol. 63, pp. 1789–1797, 2003.

[64] L. A. Loss, A. Sadanandam, S. Durinck, S. Nautiyal, D. Flaucher, V. E. Carlton, M. Moorhead, Y. Lu, J. W. Gray, M. Faham, P. Spellman, and B. Parvin, "Prediction of epigenetically regulated genes in breast cancer cell lines," *BMC Bioinformatics*, vol. 11, no. 305, 2010.

[65] B. K. Brisson, E. A. Mauldin, W. Lei, L. K. Vogel, A. M. Power, A. Lo, D. Dopkin, C. Khanna, R. G. Wells, and E. Pure, "Estimation of mean sojourn time in breast cancer screening using a Markov chain model of entry to and exit from preclinical detectable phase," *Am J Pathol*, vol. 185, no. 5, pp. 1471–86, 2015.

[66] G. Xiong, L. Deng, J. Zhu, R. P. GW, and R. Xu, "Prolyl-4-hydroxylase subunit 2 promotes breast cancer progression and metastasis by regulating collagen deposition," *BMC Cancer*, vol. 14, no. 1, 2014.

[67] F. Bertucci, V. Nasser, S. Granjeaud, F. Eisinger, J. Adelaïde, R. Tagett, B. Loriod, A. Giaconia, A. Benziane, E. Devilard, J. Jacquemier, P. Viens, C. Nguyen, D. Birnbaum, and R. Houlgatte, "Gene expression profiles of poor-prognosis primary breast cancer correlate with survival," *Hum Mol Genet*, vol. 11, no. 8, pp. 863–72, 2002.

[68] Z. Lin, G. Zhu, D. Tang, J. Bu, and J. Zou, “High expression of col6a1 correlates with poor prognosis in patients with breast cancer,” *Int J Clin Exp Med*, vol. 11, no. 11, pp. 12157–12164, 2018.

[69] D. Etemadmoghadam, A. deFazio, R. Beroukhim, and C. Mermel, “Integrated genome-wide dna copy number and expression analysis identifies distinct mechanisms of primary chemoresistance in ovarian carcinomas.,” *Clin Cancer Res*, vol. 15, no. 4, pp. 1417–27, 2009.

[70] L. D. Miller, J. Smeds, J. George, V. B. Vega, L. Vergara, A. Ploner, Y. Pawitan, P. Hall, S. Klaar, E. T. Liu, and J. Bergh, “An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival,” *Proc Natl Acad Sci USA*, vol. 102, no. 38, pp. 13550–13555, 2005.

[71] R. Kay, “A Markov model for analyzing cancer markers and disease states in survival studies,” *Biometrics*, vol. 42, pp. 855–865, 1986.

[72] R. Perez-Ocon, J. Ruiz-Castro, and M. Gamiz-Perez, “Non-homogeneous Markov models in the analysis of survival after breast cancer,” *Journal of the Royal Statistical Society Series C-Applied Statistics*, vol. 50, pp. 111–124, 2001.

[73] S. W. Duffy and H. H. Chen, “Estimation of mean sojourn time in breast cancer screening using a Markov chain model of entry to and exit from preclinical detectable phase,” *Statistics in Medicine*, vol. 14, pp. 1531–1543, 1995.

[74] H. H. Chen, S. W. Duffy, and L. Tabar, “An arbitrary Lagrangian-Eulerian computing method for all flow speeds,” *J Comput Phys*, vol. 14, no. 3, pp. 227–253, 1974.

[75] C. H. Jackson, L. D. Sharples, S. G. Thompson, S. W. Duffy, and E. Couto, “Multistate Markov models for disease progression with classification error,” *Journal of the Royal Statistical Society Series D - The Statistician*, vol. 52, no. 2, pp. 193–209, 2003.

[76] L. D. Sharples, “Use of the Gibbs sampler to estimate transition rates between grades of coronary disease following cardiac transplantation,” *Statistics in Medicine*, vol. 12, pp. 1155–1169, 1993.

[77] J. H. Klotz and L. D. Sharples, “Estimation for a Markov heart transplant model,” *The Statistician*, vol. 43, no. 3, pp. 431–436, 1994.

[78] I. M. Longini, W. S. Clark, J. W. Byers, RA HA Lemp, and H. W. Hethcote, “Statistical analysis of the stages of hiv infection using a Markov model,” *Statistics in Medicine*, vol. 8, pp. 851–843, 1989.

[79] R. C. Gentleman, J. F. Lawless, J. C. Lindsey, and P. Yan, “Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease,” *Statistics in Medicine*, vol. 13, no. 3, pp. 805–821, 1994.

[80] D. Commenges, P. Joly, L. Letenneur, and J. F. Dartigues, “Incidence and mortality of alzheimerś disease or dementia using an illness-death model,” *Statistics in Medicine*, vol. 23, pp. 199–210, 2004.

[81] G. Marshall and R. H. Jones, “Multi-state Markov models and diabetic retinopathy,” *Statistics in Medicine*, vol. 14, no. 18, pp. 1975–83, 1995.

[82] P. K. Andersen, “Multistate models in survival analysis: a study of nephropathy and mortality in diabetes,” *Statistics in Medicine*, vol. 7, no. 6, pp. 661–670, 1988.

[83] A. J. Kirby, “Statistical modelling for the precursors of cervical cancer,” Tech. Rep. Thesis (Ph.D.), University of Cambridge, Cambridge, England, United Kingdom, 1991.

[84] P. K. Andersen, L. S. Hansen, and N. Keiding, “Assessing the influence of reversible disease indicators on survival,” *Statistics in Medicine*, vol. 10, pp. 1061–1067, 1991.

[85] O. O. Aalen, O. Borgan, and H. K. Gjessing, *Survival and event history analysis. A process point of view.* New York, NY: Springer, 2008. ISBN 978-0-387-20287-7.

[86] P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding, *Statistical models based on counting processes*. New York, NY: Springer, 1993. ISBN 978-1-4612-4348-9.

[87] S. Johansen, “An Extension of Coxś Regression Model,” *International Statistical Review*, vol. 51, no. 2, pp. 165–174, 1983.

[88] S. Huang, C. Hu, M. Bell, D. Billheimer, S. Guerra, D. Roe, M. Vasquez, and E. Bedrick, “Regularized continuous-time markov model via elastic net,” *Biometrics*, vol. 74, no. 3, pp. 1045–1054, 2018.

[89] J. Kalbfleisch and J. F. Lawless, “The analysis of panel data under a markov assumption,” *Journal of American Statistical Association*, vol. 80, no. 392, pp. 863—871, 1985.

[90] H. Reulen and T. Kneib, “Structured fusion lasso penalized multi-state models,” *Statistics in Medicine*, vol. 35, no. 25, pp. 4637—4659, 2016.

[91] M. Oelker and G. Tutz, “A uniform framework for the combination of penalties in generalized structured models,” *Advances in Data Analysis and Classification*, vol. 11, no. 1, pp. 97–120, 2017.

[92] P. K. Andersen and N. Keiding, “Multi-state models for event history analysis,” *Statistical Methods Medical Research*, vol. 11, no. 2, pp. 91–115, 2002.

[93] H. Putter, M. Fiocco, and R. B. Geskus, “Tutorial in biostatistics: Competing risks and multistate models,” *Statistics in Medicine*, vol. 26, pp. 2389–2430, 2007.

[94] L. C. deWreede, M. Fiocco, and H. Putter, “The mstate Package for Estimation and Prediction in Non- and Semi-Parametric Multi-State and Competing Risks Models,” *Computer Methods and Programs in Biomedicine*, vol. 99, no. 3, pp. 261–74, 2010.

[95] T. Hastie and R. Tibshirani, *Generalized Additive Models*. London: Chapman and Hall, 1990. ISBN 9780412343902.

[96] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*. New York, NY: Springer, 2009. ISBN 978-0-387-84858-7.

[97] L. Meier, S. vandeGeer, and P. Buhlmann, “The group lasso for logistic regression,” *Journal of the Royal Statistical Society Series B*, vol. 70, no. 1, pp. 53–71, 2007.

[98] N. E. Breslow, “Discussion of the paper by D.R.Cox,” *Journal of the Royal Statistical Society Series B*, vol. 34, pp. 216–217, 1972.

[99] O. O. Aalen and S. Johansen, “Empirical transition matrix for nonhomogeneous Markov-chains based on censored observations,” *Scandinavian Journal of Statistics*, vol. 5, pp. 141–150, 1978.

[100] L. C. deWreede, “mstate: An r package for the analysis of competing risks and multi-state models,” *Journal of Statistical Software*, vol. 38, no. 7, pp. 53–66, 2011.

[101] T. Fawcett, “An introduction to roc analysis,” *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.

[102] N. M. Laird and J. H. Ware, “Random-effects models for longitudinal data,” *Biometrics*, vol. 38, no. 4, pp. 963–974, 1982.

[103] M. S. Wulfsohn and A. A. Tsiatis, “A joint model for survival and longitudinal data measured with error,” *Biometrics*, vol. 53, no. 1, pp. 330–339, 1997.

[104] J. G. Ibrahim, H. Chu, and L. M. Chen, “Basic concepts and methods for joint models of longitudinal and survival data,” *Journal of Clinical Oncology*, vol. 28, no. 16, pp. 2796–2801, 2010.

[105] M. J. Sweeting and S. G. Thompson, “Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture,” *Biometrical Journal*, vol. 53, no. 5, pp. 750–763, 2011.

[106] A. A. Tsiatis and M. Davidian, “Joint modeling of longitudinal and time-to-event data: An overview,” *Statistica Sinica*, no. 14, pp. 809–834, 2004.

[107] D. Rizopoulos, *Joint Models for Longitudinal and Time-to-Event Data With Applications in R*. Chapman and Hall/CRC, 2012.

[108] A. Lawrence Gould, M. E. Boye, M. J. Crowther, J. G. Ibrahim, G. Quartey, S. Micallef, and F. Y. Bois, “Joint modeling of survival and longitudinal non-survival data: current methods and issues. report of the dia bayesian joint modeling working group,” *Statistics in Medicine*, vol. 34, no. 14, pp. 2181–2195, 2015.

[109] G. L. Hickey, P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona, “Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues,” *BMC Medical Research Methodology*, vol. 16, no. 1, p. 117, 2016.

[110] G. Papageorgiou, K. Mauff, A. Tomer, and D. Rizopoulos, “An overview of joint modeling of time-to-event and longitudinal outcomes,” *Annual Review of Statistics and Its Application*, vol. 6, no. 1, pp. 223–240, 2019.

[111] D. Rizopoulos and P. Ghosh, “A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event,” *Statistics in Medicine*, vol. 30, no. 12, pp. 1366–80, 2011.

[112] Z. He, W. Tu, S. Wang, H. Fu, and Z. Yu, “Simultaneous variable selection for joint models of longitudinal and survival outcomes,” *Biometrics*, vol. 71, no. 1, pp. 178–187, 2015.

[113] Y. Xie, Z. He, W. Tu, and Z. Yu, “Variable selection for joint models with time-varying coefficients,” *Statistical Methods in Medical Research*, vol. 29, no. 1, pp. 309–322, 2020.

[114] Y. Chen and Y. Wang, “Variable selection for joint models of multivariate longitudinal measurements and event time data,” *Statistics in Medicine*, vol. 36, no. 24, pp. 3820–3829, 2017.

[115] G. L. Hickey, P. Philipson, A. Jorgensen, and A. Jorgensen, “joinerml: a joint model and software package for time-to-event and multivariate longitudinal outcomes,” *BMC Medical Research Methodology*, vol. 18, no. 50, 2018.

[116] W. Ye, X. Lin, and J. M. G. Taylor, “Semiparametric modeling of longitudinal measurements and time-to-event data–a two-stage regression calibration approach,” *Biometrics*, vol. 64, no. 4, pp. 1238–1246, 2008.

[117] D. Rizopoulos, G. Verbeke, and E. Lesaffre, “Fully exponential laplace approximations for the joint modelling of survival and longitudinal data,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 3, pp. 637–654, 2009.

[118] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R Core Team, *nlme: Linear and Nonlinear Mixed Effects Models*, 2021. R package version 3.1-152.

[119] J. Bien and R. J. Tibshirani, “Sparse estimation of a covariance matrix,” *Biometrika*, vol. 98, no. 4, pp. 807–820, 2011.

[120] G. Schwarz, “Estimating the Dimension of a Model,” *The Annals of Statistics*, vol. 6, no. 2, pp. 461 – 464, 1978.

[121] A. P. CA, “Generating survival times to simulate Cox proportional hazards models with time-varying covariates,” *Statistics in Medicine*, vol. 31, no. 29, p. 3946–58, 2012.

[122] P. C. Murtaugh, E. R. Dickson, G. M. V. Dam, M. Malinchoc, P. M. Grambsch, A. L. Langworthy, and C. H. Gips, “Primary biliary cirrhosis: Prediction of short-term survival based on repeated patient visits,” *Hepatology*, vol. 20, p. 126–134, 1994.

[123] M. I. Prince, A. Chetwynd, W. L. Craig, J. V. Metcalf, and O. F. James, “Asymptomatic primary biliary cirrhosis: clinical features, prognosis, and symptom progression in a large population based cohort,” *Gut*, vol. 53, no. 6, pp. 865–70, 2004.

[124] J. Roll, J. L. Boyer, D. Barry, and G. Klatskin, “The prognostic importance of clinical and histologic features in asymptomatic and symptomatic primary biliary cirrhosis,” *N Engl J Med*, vol. 308, no. 1, pp. 1–7, 1983.

[125] E. R. Dickson, P. M. Grambsch, T. R. Fleming, L. D. Fisher, and A. Langworthy, “Prognosis in primary biliary cirrhosis: model for decision making,” *Hepatology*, vol. 10, no. 1, pp. 1–7, 1989.

[126] W. R. Kim, R. H. Wiesner, J. J. Poterucha, T. M. Therneau, J. T. Benson, and R. A. Krom, “Adaptation of the Mayo primary biliary cirrhosis natural history model for application in liver transplant candidates,” *Liver Transplant*, vol. 6, no. 4, pp. 489–94, 2000.

[127] P. J. Johnson, S. Berhane, C. Kagebayashi, S. Satomura, M. Teng, and H. L. Reeves, “Assessment of liver function in patients with hepatocellular carcinoma: a new evidence-based approach-the ALBI grade,” *J Clin Oncol*, vol. 33, no. 6, pp. 550–8, 2015.

[128] A. W. Chan, R. C. Chan, G. L. Wong, V. W. Wong, P. C. Choi, and H. L. Chan, “New simple prognostic score for primary biliary cirrhosis: albumin-bilirubin score,” *J Gastroenterol Hepatol*, vol. 30, no. 9, pp. 1391–6, 2015.

[129] M. Malinchoc, P. S. Kamath, F. D. Gordon, C. J. Peine, J. Rank, and P. C. Ter Borg, “A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts,” *Hepatology*, vol. 31, no. 4, pp. 864–71, 2000.

[130] E. Christensen, J. Neuberger, J. Crowe, D. G. Altman, H. Popper, and B. Portmann, “Beneficial effect of azathioprine and prediction of prognosis in primary biliary cirrhosis. Final results of an international trial,” *Gastroenterology*, vol. 89, no. 5, pp. 1084–91, 1985.

[131] E. Christensen, D. G. Altman, J. Neuberger, B. L. De Stavola, N. Tygstrup, and R. Williams, “Updating prognosis in primary biliary cirrhosis using a timedependent Cox regression model,” *Gastroenterology*, vol. 105, no. 6, pp. 1865–76, 1993.

[132] P. S. Kamath, R. H. Wiesner, M. Malinchoc, W. Kremers, T. M. Therneau, and C. L. Kosberg, “A model to predict survival in patients with end-stage liver disease,” *Hepatology*, vol. 33, no. 2, pp. 464–70, 2001.

[133] E.-R. Andrinopoulou and D. Rizopoulos, “Bayesian shrinkage approach for a joint model of longitudinal and survival outcomes assuming different association structures,” *Statistics in Medicine*, vol. 35, no. 26, p. 4813–23, 2016.

[134] P. K. Andersen and R. D. Gill, “Cox’s regression model for counting processes: a large sample study,” *The Annals of Statistics*, vol. 10, no. 4, pp. 1100–1120, 1982.

APPENDIX A

CHAPTER 3

A.1 Appendix 1

We have studied the statistical properties of the estimators: consistency and convergence rate as follows.

The partial likelihood

$$\ell_n(\beta) = -\frac{1}{n}\sum_{i=1}^{D}\left[\Big(\sum_{j=1}^{J} X_j^{(i)}\beta_j\Big) - \log\bigg(\sum_{l\in R_i}\exp\Big(\sum_{j=1}^{J} X_j^{(l)}\beta_j\Big)\bigg)\right],$$

where the penalty term $P_{\lambda,\gamma}(\beta)$ can be denoted as $P_{\lambda_n}(\beta)$ since γ for group SCAD and group MCP are fixed. Here, $\ell_n(\beta), \lambda_n$ denote the partial likelihood and tuning parameter changing with the sample size n, respectively.

Let the true parameter be $\beta_0 = \left(\beta_{01}^T, \beta_{02}^T\right)^T$ where β_{01} consists of all nonzero groups and β_{02} consists of all remaining zero groups. The objective function is

$$\mathcal{Q}_n(\beta, \lambda_n) = \ell_n(\beta_0) + \ell_n^{'}(\beta_0)^T(\beta - \beta_0) + \frac{\tau}{2}(\beta - \beta_0)^T(\beta - \beta_0) + P_{\lambda_n}(\beta).$$

Correspondingly, the minimizer of $\mathcal{Q}_n(\beta, \lambda_n)$ is $\beta_n = \left(\beta_{n1}^T, \beta_{n2}^T\right)^T$ where $\beta_n = \underset{\beta}{\text{argmin}}\ \mathcal{Q}_n(\beta, \lambda_n)$. Define $a_n = \max\{P_{\lambda_n}^{'}(\|\beta_{j0}\|) : \|\beta_{j0}\| \neq 0\}$ and $b_n = \max\{P_{\lambda_n}^{"}(\|\beta_{j0}\|) : \|\beta_{j0}\| \neq 0\}$.

Theorem 1: (Consistency and convergence rate) If $P_{\lambda_n}(\|\beta\|)$ simultaneously satisfies two conditions: $a_n = O_p(n^{-1/2})$ and $b_n \to 0$, then β_n is a root-n consistent estimator for β_0 with rate $n^{-1/2}$, i.e. $\|\beta_n - \beta_0\| = O_p(n^{-1/2})$.

Proof: According to Theorem 3.2 in [134] two results hold

$$-\ell^{'}(\beta_0) \xrightarrow{p} n^{-1/2}\mathcal{N}(0,\Sigma)$$

$$\ell^{"}(\beta^*) \xrightarrow{p} n\Sigma \text{ for any random } \beta^* \xrightarrow{p} \beta_0$$

$$\text{Then, } \ell^{"}(\beta^*) = n(\Sigma + O_p(1)),$$

where Σ is the positive definite Fisher information matrix.

Consider a constant ball, $B(C) = \{\beta_0 + \alpha_n \mathbf{u} : \|\mathbf{u}\| \leq C\}$ and its boundary $\partial B(C)$ where $C > 0$ and $\alpha_n = n^{-1/2} + a_n$. Therefore, $O_p(\alpha_n) = O_p(a_n) = O_p(n^{-1/2})$. To prove $\|\beta_n - \beta_0\| = O_p(n^{-1/2})$, it is sufficient to prove that for any $\epsilon > 0$, there exists a large constant C such that

$$\mathrm{P}\Bigg(\sup_{\beta \in \partial B(C)} Q_n(\beta,\lambda_n) < Q(\beta_0,\lambda_n)\Bigg) \geq 1 - \epsilon. \tag{A.1}$$

This implies that with probability at least $1-\epsilon$ (or goes to 1), $Q_n(\beta,\lambda_n)$ has a local minimum in the ball $B(C)$ for a given λ_n.

Denote $D_n(\mathbf{u}) = Q_n(\beta,\lambda_n) - Q(\beta_0,\lambda_n)$, we have

$$D_n(\mathbf{u}) = \ell^{'}(\beta_0)^T(\beta-\beta_0) + \frac{\tau}{2}(\beta-\beta_0)^T(\beta-\beta_0) + P_{\lambda_n}(\beta) - P_{\lambda_n}(\beta_0) = D_1 + D_2.$$

Consider that

$$\begin{aligned} D_1 &= \ell^{'}(\beta_0)^T(\beta-\beta_0) + \frac{\tau}{2}(\beta-\beta_0)^T(\beta-\beta_0) \\ &= O_p(n^{-1/2})\alpha_n\mathbf{u} + \frac{\tau}{2}\alpha_n^2\mathbf{u}^T\mathbf{u} \\ &= O_p(C\alpha_n^2) + O_p(C^2\alpha_n^2). \end{aligned}$$

Consider D_2 using Taylor expansion, we have

$$\begin{aligned} D_2 &= P_{\lambda_n}(\beta) - P_{\lambda_n}(\beta_0) \\ &= \sum_j P'_{\lambda_n}(\|\beta_{j0}\|)(\|\beta_{j0} + \alpha_n \mathbf{u}_j\| - \|\beta_{j0}\|) + \frac{1}{2}(\|\beta_{j0}+ \\ \alpha_n \mathbf{u}_j\| &- \|\beta_{j0}\|)^T \left(P^{"}_{\lambda_n}(\|\beta_{j0}\|)(\|\beta_{j0} + \alpha_n \mathbf{u}_j\| - \|\beta_{j0}\|)\right. \\ &\leq \sum_j a_n \alpha_n \|\mathbf{u}_j\| + b_n \alpha_n^2 \|\mathbf{u}_j\|^2 \\ &\leq \sum_j \alpha_n^2 C + b_n \alpha_n^2 C^2 = J(\alpha_n^2 C + b_n \alpha_n^2 C^2). \end{aligned}$$

Because $b_n \to 0$, $D_2 \to O_p(C\alpha_n^2)$. By choosing a sufficiently large C, D_1 dominates D_2. Thus, inequality (A.1) holds $\square$.

A.2 Appendix 2

We present the simulation studies of the second cross-validation approach described in Section 2.7 to select the tuning parameters λ and evaluate its variable selection performance.

In Figure A.1, each dot represents the logarithm of the λ values along the solution path, and the error bars provide the confidence intervals for the cross-validation log-partial-likelihood. The left vertical bar indicates the maximum cross-validation partial-log-likelihood using the first method [33] while the right one shows the maximum cross-validation log-partial-likelihood using the second method [34].

We continue considering $N = 100$ observations and $P = 400$ covariates with 40 groups, each with 10 elements. There are two non-zero groups. The coefficient magnitude $|\beta| = 0.5$, the values of the population correlation ρ are 0, 0.2 and 0.5, the censoring rates are 0% and 20%. The results are summarized in Tables A.1, A.2, and A.3. It can be seen that using the second cross-validation method always results in smaller models than using the first cross-validation method. For group lasso, it produces better variable selection results with much smaller FPR values. For group SCAD and MCP, it often gives better results, but sometimes suppresses too much, e.g., in group MCP case with 20% censoring, $\rho = 0.5$. Therefore, the second cross-validation method should be used

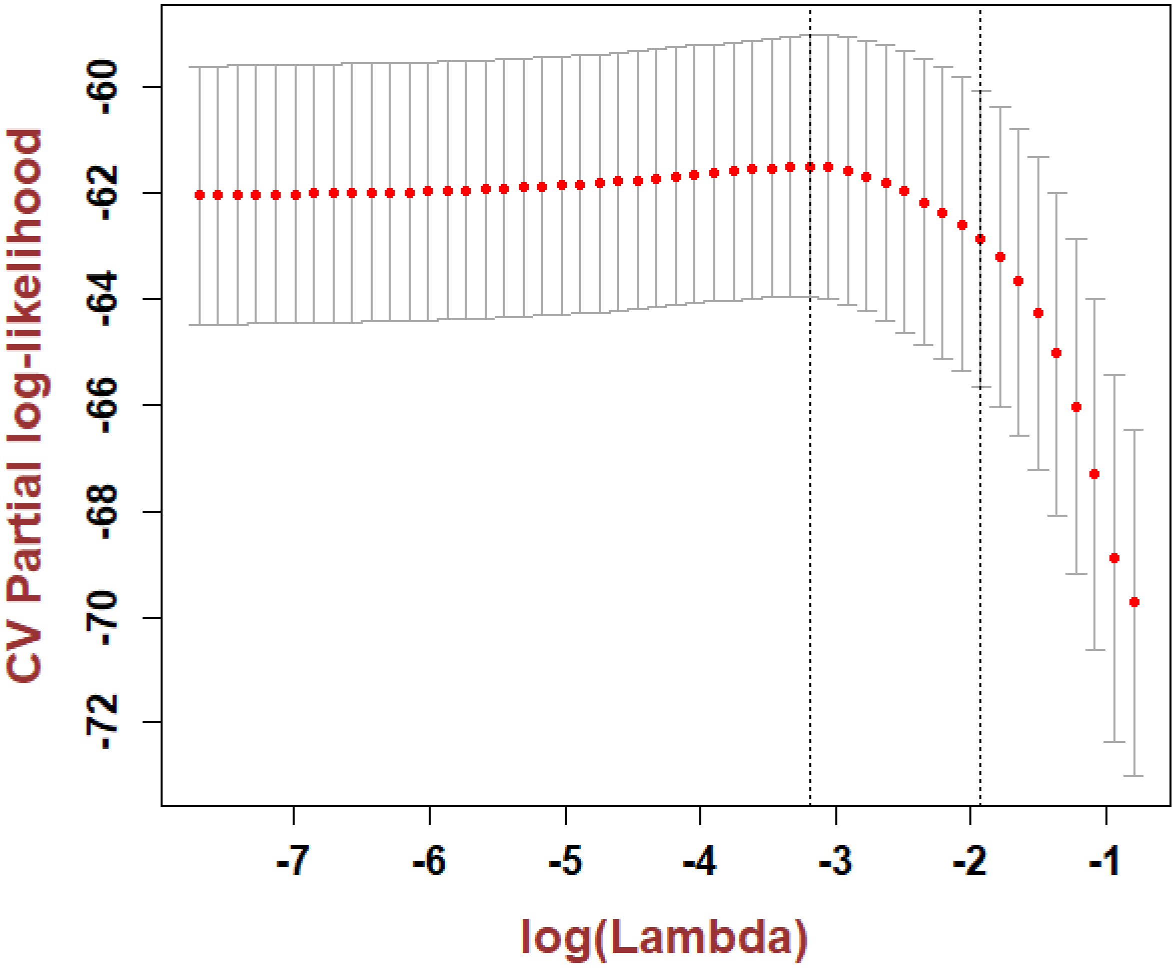


Figure A.1: Plot of the cross-validation log-partial likelihood against the log of λ values along the regularization path.

with caution.

Censoring rate	ρ	First CV method			Second CV method		
		Model size	TPR	FPR	Model size	TPR	FPR
No censoring	0	95.8	1	0.19	30	1	0.02
	0.2	79.5	1	0.15	20	1	0
	0.5	119.6	1	0.26	30	1	0.02
20% censoring	0	102	1	0.21	33.2	1	0.03
	0.2	94.1	1	0.19	25.1	1	0.01
	0.5	122.6	1	0.27	32.2	1	0.03

Table A.1: Results for group lasso using different cross-validation methods to select hyperparameters over 100 replications.

Censoring rate	ρ	First CV method			Second CV method		
		Model size	TPR	FPR	Model size	TPR	FPR
No censoring	0	20	1	0	20	1	0
	0.2	40	1	0.05	39	1	0.04
	0.5	58.1	1	0.10	23.1	1	0.01
20% censoring	0	80	1	0.15	30.9	1	0.02
	0.2	40.4	1	0.05	29.7	1	0.02
	0.5	83.7	1	0.17	27.6	0.91	0.02

Table A.2: Results for group SCAD using different cross-validation methods to select hyperparameters over 100 replications.

A.3 Appendix 3

We present additional settings: settings with a large number of overlapping covariates and the number of zero groups being more than the number of non-zero groups. More specifically, we have performed an additional experiment using the simulated data with $N = 100$, $P = 55$, in which there are 10 groups of size 10 and 50% covariates overlap between two successive groups. The

Censoring rate	ρ	First CV method			Second CV method		
		Model size	TPR	FPR	Model size	TPR	FPR
	0	20	1	0	20	1	0
No censoring	0.2	20	1	0	20	1	0
	0.5	20.4	1	0.00	19.5	0.98	0
	0	29.5	1	0.02	20	1	0
20% censoring	0.2	32	1	0.03	20	1	0
	0.5	36.9	1	0.04	16.2	0.65	0.01

Table A.3: Results for group MCP using different cross-validation methods to select hyperparameters over 100 replications.

"correct" underlying group structure is given by

$$\underbrace{1,\dots,10}_{group1}\ \underbrace{6,\dots,15}_{group2}\ \underbrace{11,\dots,20}_{group3}\ \underbrace{16,\dots,25}_{group4}\ \underbrace{21,\dots,30}_{group5}$$

$$\underbrace{26,\dots,35}_{group6}\ \underbrace{31,\dots,40}_{group7}\ \underbrace{36,\dots,45}_{group8}\ \underbrace{41,\dots,50}_{group9}\ \underbrace{46,\dots,55}_{group10}.$$

We set the population correlation $\rho = 0.5$ with 30% censoring rate. The corresponding coefficients are

$$\underbrace{0,\dots,0}_{group1-2}\ \underbrace{0,0,0,0,0,1.5,0,0,-2,0}_{group3}\ \underbrace{1.5,0,0,-2,0,0,0,0,0,0}_{group4}\ \underbrace{0,\dots,0}_{group5-6}$$

$$\underbrace{0,0,0,0,0,1.4,0,0,0,1.8}_{group7}\ \underbrace{1.4,0,0,0,1.8,0,0,0,0,0}_{group8}\ \underbrace{0,\dots,0}_{group9-10}.$$

Then we consider four setups with the misspecified group structures for inference. In the first setup, the number of groups are incorrect because the overlapping groups are collapsed as follows:

$$\underbrace{1,\dots,10}_{group1}\ \underbrace{6,\dots,15}_{group2}\ \underbrace{11,\dots,25}_{group3}\ \underbrace{21,\dots,30}_{group4}\ \underbrace{26,\dots,35}_{group5}\underbrace{31,\dots,45}_{group6}\ \underbrace{41,\dots,50}_{group7}\ \underbrace{46,\dots,55}_{group8}.$$

In the second setup, the misspecified group structure deviates from the ground truth more significantly will all the overlapping covariates put into one group:

$$\underbrace{1,3,5,7,9,11,13,15}_{group1}\underbrace{2,4,\ldots,12,14,16,17,18,19,20,21,22}_{group2}$$
$$\underbrace{16,\ldots,25}_{group3}\underbrace{21,\ldots,30}_{group4}\underbrace{26,\ldots,35}_{group5}\underbrace{31,\ldots,45}_{group6}\underbrace{41,\ldots,50}_{group7}\underbrace{46,\ldots,55}_{group8}.$$

Similar as the first setup, the third and fourth setups are defined as follows:

$$\underbrace{1,\ldots,20}_{group1}\underbrace{16,\ldots,25}_{group2}\underbrace{21,\ldots,30}_{group3}\underbrace{26,\ldots,35}_{group4}\underbrace{31,\ldots,45}_{group5}\underbrace{41,\ldots,50}_{group6}\underbrace{46,\ldots,55}_{group7}$$

and

$$\underbrace{1,\ldots,10}_{group1}\underbrace{6,\ldots,20}_{group2}\underbrace{16,\ldots,25}_{group3}\underbrace{21,\ldots,30}_{group4}\underbrace{26,\ldots,40}_{group5}\underbrace{36,\ldots,45}_{group6}\underbrace{41,\ldots,50}_{group7}\underbrace{46,\ldots,55}_{group8}$$

The results shown in Table A.4 confirm our expectation: the setup with the collapsed groups including several non-zero (active) groups produces worse results than the cases with the collapsed groups with none or only one non-zero group. More clearly, the first setup in the table including two collapsed groups (group3 and group5), where each of them consists of two non-zero groups, has the worst variable selection performance. Both the second and third misspecification setups including only one group (group5) that is collapsed from two non-zero groups have almost the same performance, better than the first misspecification setup. The fourth mispecification setup with no misspecified group collapsed from two non-zero groups has the best performance. We hypothesize that the probability of variables being incorrectly selected increases due to the ignorance of the overlapping property of active elements in the collapsed groups and the larger group sizes of these collapsed groups. In other words, FPR increases and then corresponding RMSE increases.

		TPR	FPR	Model size	RMSE
Correct specification	truth			4	
	Group lasso	1	0.70	40	0.24
	Group SCAD	1	0.52	31	0.14
	Group MCP	1	0.35	22.2	0.12
First misspecification	truth			4	
	Group lasso	1	0.71	40.5	0.26
	Group SCAD	1	0.53	31.2	0.14
	Group MCP	1	0.50	29.3	0.15
Second misspecification	truth			4	
	Group lasso	1	0.71	40.3	0.26
	Group SCAD	1	0.50	29	0.13
	Group MCP	1	0.40	25	0.13
Third misspecification	truth			4	
	Group lasso	1	0.70	40.2	0.26
	Group SCAD	1	0.50	29.5	0.13
	Group MCP	1	0.41	25.6	0.13
Fourth misspecification	truth			4	
	Group lasso	1	0.75	42.2	0.25
	Group SCAD	1	0.42	26	0.12
	Group MCP	1	0.35	21.9	0.12

Table A.4: Results for misspecified group structures over 100 replications.