Fall 2019

Variable Selection in Accelerated Failure Time (AFT) Frailty Models: An Application of Penalized Quasi-Likelihood

Sarbesh R. Pandeya

VARIABLE SELECTION IN ACCELERATED FAILURE TIME (AFT) FRAILTY MODELS: AN APPLICATION OF PENALIZED QUASI-LIKELIHOOD

by

SARBESH RAJ PANDEYA

(Under the Direction of Lili Yu)

ABSTRACT

Variable selection is one of the standard ways of selecting models in large scale datasets. It has applications in many fields of research study, especially in large multi-center clinical trials. One of the prominent methods in variable selection is the penalized likelihood, which is both consistent and efficient. However, the penalized selection is significantly challenging under the influence of random (frailty) covariates. It is even more complicated when there is involvement of censoring as it may not have a closed-form solution for the marginal log-likelihood. Therefore, we applied the penalized quasi-likelihood (PQL) approach that approximates the solution for such a likelihood. In addition, we introduce an adaptive penalty function that makes the selection on both fixed and frailty effects in a left-censored dataset for a parametric AFT frailty model. We also compared our penalty function with other established procedures via their performance on accurately choosing the significant coefficients and shrinking the non-significant coefficients to zero.

INDEX WORDS: AFT models, Survival analysis, Variable selection, Frailty models

VARIABLE SELECTION IN ACCELERATED FAILURE TIME (AFT) FRAILTY MODELS: AN APPLICATION OF PENALIZED QUASI-LIKELIHOOD

by

SARBESH RAJ PANDEYA

B.P.H., Pokhara University, Nepal, 2011

M.P.H., Georgia Southern University, 2015

A Dissertation Submitted to the Graduate Faculty of Georgia Southern University in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PUBLIC HEALTH

STATESBORO, GEORGIA

VARIABLE SELECTION IN ACCELERATED FAILURE TIME (AFT) FRAILTY MODELS: AN APPLICATION OF PENALIZED QUASI-LIKELIHOOD

by

SARBESH RAJ PANDEYA

Major Professor:	Lili Yu
Committee:	Hani Samawi
	Xinyan Zhang

Electronic Version Approved:
December 2019

ACKNOWLEDGMENTS

I want to thank my supervisor Dr. Lili Yu who had given me a lot of guidance and directions in this work. Her valuable inputs, patience and devotion towards my work was the key to the successful completion of this dissertation. In addition, Dr. Hani Samawi was a significant factor in correcting my knowledge in every step. His openness to help is something I will always cherish. I am grateful to Dr. Xinyan Zhang and her points to guide the manuscript.

I want to also extend my appreciation to my friends and family who helped me through this journey. I am extremely thankful for the time and effort they provided me to make this work a success. It has been a blessing and I am incredibly grateful to the staff members and the resources made available to me at Georgia Southern University.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS 2

LIST OF TABLES 5

LIST OF FIGURES 7

CHAPTER

1 INTRODUCTION 8

2 LITERATURE REVIEW 11

Variable Selection in Survival Data with Accelerated Failure Time Models 12

Variable Selection in Models with Frailty 15

3 METHODS 18

The Mixed Models 18

Variable Selection 21

4 PROPOSED METHOD 37

Basic Notations and Formulas 37

Accelerated Failure Time (AFT) with Frailty Effects 38

Statistical Inference on AFT models with Frailty 39

Penalization 45

Estimation of the Variance-Covariance Matrix 46

Overview of Individual Distributions 48

5 SIMULATION STUDY 52

6 APPLICATION TO UNSTRUCTURED TREATMENT INTERRUPTION DATA 79

Introduction 79

The UTI Dataset 79

Distribution of Dependent Variable 81

Data Modeling 82

Results 83

7 CONCLUSION 90

REFERENCES 93

LIST OF TABLES

Table Page

5.1 Results of the sample n=40 m=10 on Simulation 1 for Lognormal Distribution 55

5.2 Coefficient bias on sample n=40 m=10 for Simulation 1 following Lognormal Distribution 56

5.3 Results of the sample n=50 m=10 on Simulation 1 for Lognormal Distribution 57

5.4 Coefficient bias on sample n=50 m=10 for Simulation 1 following Lognormal Distribution 58

5.5 Results of the sample n=60 m=20 on Simulation 1 for Lognormal Distribution 59

5.6 Coefficient bias on sample n=60 m=20 for Simulation 1 following Lognormal Distribution 60

5.7 Results of the sample n=20 m=5 on Simulation 2 for Lognormal Distribution 61

5.8 Coefficient bias on sample n=20 m=5 for Simulation 2 following Lognormal Distribution 62

5.9 Results of the sample n=30 m=5 on Simulation 2 for Lognormal Distribution 63

5.10 Coefficient bias on sample n=30 m=5 for Simulation 2 following Lognormal Distribution 64

5.11 Results of the sample n=40 m=10 on Simulation 2 for Lognormal Distribution 65

5.12 Coefficient bias on sample n=40 m=10 for Simulation 2 following Lognormal Distribution 66

5.13 Results of the sample n=40 m=10 on Simulation 1 for Weibull Distribution 67

5.14 Coefficient bias on sample n=40 m=10 for Simulation 1 following Weibull Distribution 68

5.15 Results of the sample n=50 m=10 on Simulation 1 for Weibull Distribution 69

5.16 Coefficient bias on sample n=50 m=10 for Simulation 1 following Weibull Distribution 70

5.17 Results of the sample n=60 m=20 on Simulation 1 for Weibull Distribution 71

5.18 Coefficient bias on sample n=60 m=20 for Simulation 1 following Weibull Distribution 72

5.19 Results of the sample n=20 m=10 on Simulation 2 for Weibull Distribution 73

5.20 Coefficient bias on sample n=20 m=10 for Simulation 2 following Weibull Distribution 74

5.21 Results of the sample n=30 m=10 on Simulation 2 for Weibull Distribution 75

5.22 Coefficient bias on sample n=30 m=10 for Simulation 2 following Weibull Distribution 76

5.23 Results of the sample n=40 m=10 on Simulation 2 for Weibull Distribution 77

5.24 Coefficient bias on sample n=40 m=10 for Simulation 2 following Weibull Distribution 78

6.1 Sample of the UTI data 81

6.2 Goodness of fit criteria for the considered distributions 82

6.3 Estimated fixed effects and the variance covariance matrix for accelerated failure time random-effect model for the considered distributions 85

LIST OF FIGURES

Figure Page

3.1 Ridge regression (left) and LASSO (right) estimation procedure 28

6.1 PDF, QQ plot, PP plot and the CDF of empirical data compared to a Log-normal distribution 86

6.2 PDF, QQ plot, PP plot and the CDF of empirical data compared to a Log-logistic distribution 87

6.3 PDF, QQ plot, PP plot and the CDF of empirical data compared to a Weibull distribution 88

6.4 PDF, QQ plot, PP plot and the CDF of empirical data compared to a Gamma distribution 89

CHAPTER 1

INTRODUCTION

Survival analysis is a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest. For example, time to death, time to sales or time to presence of a disease. Therefore, these time to event values or commonly called survival time could be measured in days, weeks, or years. For instance, if the event of interest is a heart attack, then the survival time can be the time in years until a person develops a heart attack (Despa, 2010). Unlike continuous and categorical data-sets, survival data cannot be modeled using traditional methods of generalized linear models (GLM) primarily because the GLM method is not able to account for censoring (Despa, 2010). Censoring occurs when the survival time is incomplete. Typically, censoring arises when people drop out of the study because of loss of follow-ups or the study ends before the event of interest. There are different types of censoring mechanisms, and most of the prominent censoring mechanisms include right censoring and left censoring. Right censoring occurs when a subject leaves the study before an event occurs, or the study ends before the event has occurred. Left censoring is when the event of interest has already happened before enrollment (Hosmer Jr, Lemeshow, & May, 2011). This study will be focusing primarily on left-censored data-sets.

There are two models often used in Survival Analysis, Cox's model and the AFT model. The more popular Cox's model or the Cox proportional hazards model produces estimates of covariates along with the baseline hazard to predict the hazard/risk on a particular event time. The Accelerated Failure Time model (AFT), on the other hand, linearly estimates the log of the failure time event using the covariates in the model directly, making it more easier to interpret. However, in both the survival models, predicting survival time is the main objective and this requires several covariates that can explain it's

variation. However, there could be many different types of potential risk factors that could play a role in one's survival. Furthermore, in large scale multi-center studies, there is a high potential of study samples being randomly clustered and correlated with each other. For example, patients within each center could be correlated, thus making the center a variable for random effect. Such effects usually occur due to the impact of unobserved heterogeneous effects (frailty), and these may be in single or multiple levels. For instance, a common problem of left censoring comes in the lower detection limit of an assay, especially in the detection of the human immunodeficiency virus (HIV) viral load in plasma. The detection threshold for the assay ranges from 10,000 copies/mL to 20 or fewer copies/mL. However, when the viral load is below this detection limit, the observations are incomplete, and this leads to left censoring. Another issue may arise when HIV strains that circulate in a given individual present chance of mutations associated with antiretroviral treatment failure (detectable HIV viral load), also called HIV drug resistance. For considering such effects, researchers are required to study the association between the presence of HIV mutations and the response to antiretroviral therapy, which is measured by HIV viral load. Often in such cases, the number of predictors is a vast sequence, and all of them may not have a significant influence on the outcome variable. Additionally, if the patients are getting treatments in different centers, then there may be center effects (frailty effects). These kinds of data sets are censored and have a presence of extensive covariates that may have minimal influence in the model as well as frailty effects (Soret, Avalos, Wittkop, Commenges, & Thiébaut, 2018). Modeling such data sets presents a very high level of challenges as it may create a strong chance of overfitting that may result in a complicated inference. Furthermore, such a complex model will not be easy to interpret.

The penalized inference of model selection is a popular method to address the issue of large number of covariates in the model. It can continuously shrink the co-

efficients of less influential predictors in the model towards zero, leaving only the most influential predictors in the model. The result from such a process ensures simplicity in the model that prevents overfitting and easier interpretation. Some of the famous penalty functions include: least absolute shrinkage and selection operator (LASSO) by Tibshirani (R. Tibshirani, 1996), Elastic Net penalty (Wu, 2012) and adaptive LASSO (H. H. Zhang & Lu, 2007). However, these methods are used extensively in the right-censored data sets and only work with the simple fixed effects model and do not include the frailty covariates.

Therefore, in this dissertation, we explore and address the penalty function in left-censored data sets that can select variables for both fixed and frailty effects.

CHAPTER 2

LITERATURE REVIEW

The Cox model produces estimates of covariates along with the baseline hazard to predict the hazard/risk on particular event time. The previously mentioned heterogeneity (clustered effects) for the Cox model, have been addressed via several approaches like the intensity process of a multivariate counting process (Andersen & Gill, 1982; Prentice, Williams, & Peterson, 1981); and marginal proportional hazards models (Wei, Lin, & Weissfeld, 1989; J. Cai & Prentice, 1995). Another simplified approach is the addition of the frailty covariates in the model (Hougaard, 2012). For instance, Klein (Klein, 1992) modeled a semi-parametric Cox model with gamma-distributed frailty using the EM algorithm. Then (Nielsen, Gill, Andersen, & Sørensen, 1992) used the counting process to estimate frailty. (Ripatti & Palmgren, 2000) used the penalized quasi-likelihood approach and (Therneau, Grambsch, & Pankratz, 2003) used the penalized partial likelihood to estimate the Cox models with frailty.

In the presence of large covariates and frailty effects in the data sets, there is a challenge of producing a model that is simple and easy to interpret. There are numerous solutions for such this scenario:

1) Have a selection of only the significant variables, which is called subset selection or forward selection;

2) Eliminate the variables that are non-significant in the model, which is backward elimination;

3) Do both the selection of significant variables and the removal of non-significant variables which is stepwise selection (J. Pan, 2016).

These approaches are simple and easy to interpret, but they have a lack of

stability in terms of selection when there are minute changes in the model (Breiman et al., 1996; Fan & Li, 2001; Harrell, 2001). Therefore, penalized selection has become an excellent alternative option for model selection without losing stability. It can continuously shrink the coefficients of less influential predictors in the model towards zero and have good computational feasibility along with statistical precision. The most prominently used penalized approach is the least absolute shrinkage and selection operator (LASSO) by Tibshirani in 1996 (R. Tibshirani, 1996). He later modified this penalty function to work in censored data sets using the Cox model (R. Tibshirani, 1997). (Wu, 2012) used the Elastic Net penalty that combines the LASSO and ridge regression procedure for the Cox model, while (H. H. Zhang & Lu, 2007) used the adaptive LASSO. However, these methods are used extensively in the right-censored data sets and only work with the simple fixed effects model and do not include the frailty covariates.

However, the Cox model needs the satisfaction of the proportionality assumption. In this case, the interpretation of failure time is a little complicated in the Cox model as it is not directly modeling the failure time. It uses the proportional hazards to derive failure time interpretation (Hutton & Monaghan, 2002; Orbe, Ferreira, & Núñez-Antón, 2002; Pourhoseingholi et al., 2007). So, to model the survival data when it does not meet the condition of proportionality, the accelerated failure time (AFT) model is a viable alternative. The AFT model produces estimates of the coefficients that can predict the log of the failure time event directly (Wei, 1992). In addition, frailty covariates have been modeled in AFT by (Keiding, Andersen, & Klein, 1997), (W. Pan, 2001), (Lambert, Collett, Kimber, & Johnson, 2004) and (J. Zhang & Peng, 2007).

VARIABLE SELECTION IN SURVIVAL DATA WITH ACCELERATED FAILURE TIME MODELS

AFT is an alternative to Cox models because of its ability to model survival time directly. It can use parametric or semiparametric estimation depending on a specified or an unspecified error distribution. There are generally two approaches of estimation in AFT. One method is the Buckley-James estimator, which adjusts for censored observations using the Kaplan–Meier estimator while the other is the rank-based estimator, which is motivated by the score function of the partial likelihood. In a high dimensional setting, these approaches are very challenging and computational even complex to solve (Huang, Ma, & Xie, 2006).

Therefore, to solve the problems of models selection in a high dimensional setting, there have been a good amount of studies in the AFT models penalization. One of the first studies came in 2006 where (Huang et al., 2006) used the threshold gradient descent to conduct model selection in semi-parametric AFT models . For this, they used Stute's weighted least squares (LS) estimator (Stute, Wang, et al., 1993) in the AFT model with multiple covariates, which uses the Kaplan Meier to account for censoring for the Least Square criterion. The aforementioned approach is more amenable than the previously mentioned Buckley and James estimator and the ranked based approaches. Huang et al. used the LASSO penalty and the threshold-gradient-directed regularization method (Friedman & Popescu, 2003) on these estimates to conduct models selection. Later in 2010, (Huang & Ma, 2010) again used the Stute's weighted least squares (LS) estimator to create a new penalization approach called the bridge method. The new method was designed mainly to address the variable selection issues in censored survival data with microarray gene expression measurements . Furthermore, (Khan & Shaw, 2016) also used Stute's weighted least squares (LS) estimator to create the adaptive elastic net penalty .

(S. Wang, Nan, Zhu, & Beer, 2008) in 2008 extended the elastic net penalty on semi-parametric AFT models using the traditional Buckley and James estimator. They applied the new penalty in a high-dimensional genome data called the Michigan squamous cell lung carcinoma. Then in 2009, (Engler & Li, 2009) also used the elastic net penalty in AFT models. However, in their paper, they replaced the censored values of the outcome variable with mean imputation, which is the conditional expectation of the event time. (Engler & Li, 2009) showed that in a high dimensional and low sample size data set, this approach to predict AFT models under elastic net penalty outperforms the Buckley and James estimator used by (S. Wang et al., 2008). A similar approach was taken by Datta et al. in 2007 to show that mean imputation can outperform the re-weighting and multiple imputation procedure under LASSO penalization (Datta, Le-Rademacher, & Datta, 2007).

These studies assume some specification in conditional censoring distribution which is difficult in practice. Also, there is an assumption that the support of censoring time can contain the support of the entire failure time which is usually challenging to achieve in practice. The Buckley and James estimators are also not stable that cause multiple limiting values. Therefore, to address these issues, (T. Cai, Huang, & Tian, 2009) in 2009 introduced the rank-based estimation called the Gehan's estimator (Tsiatis et al., 1990) under the LASSO penalty for semi-parametric AFT models. Here the censoring is independent of the event time and conditional on covariates. It doesn't require any assumption in censoring and the resulting estimator from this approach is solved using a linear programming procedure even if the AFT model fails to hold. A similar approach (Gehan's estimator) was used in (Xu, Leng, & Ying, 2010) in 2010 where marginal probability was used to calculate the survival time by accounting for correlation in multiple failure time .

Recently, (Park & Do Ha, 2018) used the LASSO, SCAD and adaptive LASSO penalty function to conduct model selection in parametric AFT model. The failure time was assumed to have lognormal and Weibull distributions (Park & Do Ha, 2018). (Sha, Tadesse, & Vannucci, 2006) used the Bayesian variable selection approach on parametric AFT models using the lognormal and log-t-distributions. They use a conjugate prior for model parameters and derive a marginalized likelihood with the regression parameters being integrated out. The Markov chain Monte Carlo (MCMC) algorithm is used for completing the variable selection procedure (Sha et al., 2006). Later, (Z. Zhang, Sinha, Maiti, & Shipp, 2018) used a nonparametric Bayesian method for regularized estimation of the regression parameters on the AFT models. To ensure, nonparametric measures in error, they use the Dirichlet mixture of normal densities to model it. It gives great flexibility for the given model and allows for an infinite number of mixing components in the prior .

Variable Selection in Models with Frailty

There has been significant literature in linear mixed models when it comes to variable selection procedure. However, there are only a handful of studies when it comes to the joint selection of the fixed and the random effects model. (Bondell, Krishna, & Ghosh, 2010) and (Ibrahim, Zhu, Garcia, & Guo, 2011) performed joint selection using the penalized maximization likelihood methods and they used the Cholesky decomposition to factorize the variance-covariance matrix of the random effects, that is to it's lower triangular matrix and it's conjugate transpose. They then used the EM algorithm to get the parameter estimations. (Bondell et al., 2010) used one tuning parameter whereas Ibrahim used two tuning parameters. Later, (B. Lin, Pang, & Jiang, 2013) used the restricted maximum likelihood approach with Newton Raphson algorithm to estimate the random parameters. They use the pathwise coordinate optimization to conduct variable selection (B. Lin et al.,

2013). (Groll & Tutz, 2014) used Breslow and Clayton's penalized quasi-likelihood approach (Breslow & Clayton, 1993) to derive the marginal likelihood of the model. They used the gradient descent algorithm for variable selection. (Hui, Müller, & Welsh, 2017) later used the same penalized quasi-likelihood approach to develop their joint selection in fixed and random effects for adaptive model selection.

For survival analysis, there has been little advancement in joint selection approach to mixed models. In the Cox model, there has been a significant progress in model selection when it comes to the fixed effects. However, the literature is limited in terms of joint selection. Generally in survival analysis, the random effects are treated as frailty that has certain distribution (Gamma, inverse Gaussian, Lognormal, etc.). Variable selection is done for the likelihood conditioned on the frailty factor. (Fan & Li, 2002) were the first to perform variable selection with the LASSO and the SCAD penalty under the presence of frailty following Gamma distribution. They had extended their penalty mechanism that they introduced in the linear model (Fan & Li, 2001) to the survival data. Later (Androulakis, Koukouvinos, & Vonta, 2012) extended the same penalty mechanisms for inverse Gaussian distributed frailty factors. Then again, (Groll, Hastie, & Tutz, 2017) extended it to include the frailty distribution following Normal distribution. They also made this penalty include functions that could address the presence of time-varying covariates in the model. For this, they included the B-splines algorithms in their model that uses the spline function to account for time-varying covariates. They uses an extra penalty function to smooth out this spline coefficient. Overall, there are two tuning parameters in this penalty that addresses the fixed effects, the random effects as well as the time-varying effects. Similar to the their model selection in the linear model (Groll & Tutz, 2014), they used the same approximated marginal likelihood algorithm using the penalized quasi likelihood approach developed by (Breslow & Clayton, 1993).

Unlike the Cox model and the linear model, there has been very little progress in variable selection in AFT models under the presence of random effects. The previous subsection gave a thorough extension of the amount of variable selection procedures when it comes to the fixed effects but comparatively extremely little advancements have been made when it comes to the random effects in this model. As mentioned in the introduction, (Komarek, 2006; Komárek & Lesaffre, 2008) and (Y. Wang, 2006) were some of the earliest authors to introduce variable selection in the AFT model under frailty factors. However, these papers were not designed for selection of fixed or random effects. They were used to smooth out the error estimates of the model. Recently, (Park & Ha, 2018) used the parametric AFT models with random effects to conduct variable selection in Log-normally distributed failure time models. The H-likelihood has an ability to give a closed form solution while integrating the marginal likelihood. Then the penalty function could be added to the h-likelihood to conduct variable selection. However all these models used the right censored mechanisms and there is no study in left censored survival data in AFT frailty model for variable selection.

CHAPTER 3

METHODS

In this chapter, we review the estimation procedures of fixed and random effects models used in generalized linear models as well as the survival models. We then proceed towards the different methods for variable selection and highlight the procedures that have been applied for linear models with random effects and survival models with frailty effects. Some of the general criterion used to select a penalty parameter for shrinkage are introduced. Lastly, we will summarize the development of these methods in the AFT models.

THE MIXED MODELS

A typical mixed model contains a fixed effects and random effects. Fixed effects remain constant throughout the population whereas random effects may vary in individuals or groups. Typical models of mixed effects in linear regression and survival analysis are described below:

3.1.1 Linear Mixed Models

Linear mixed models are an extension of simple linear models that allows both fixed and random effects. They are particularly applied when there is non independence in the data. For example, students could be sampled from within classrooms, or patients from within doctors. The variance of patients from same doctors may be similar and it may be different if the patients are from different doctors. These variances are termed as within variance and between variance. Mixed models are created to address these variances via the fixed effects for within subject variance and random effects for between subject variance (for Digital Research & Education, n.d.).

The underlying linear mixed effects models started with the assumption of unobserved heterogeneity across individuals in a given study population including any unobserved or unmeasured variation across the individuals of the given study. This condition is considered as a subset of regression coefficients (e.g. random intercepts or random slope) varying randomly across study individuals or study groups. These varied effects are termed as the random effects and the subjects with these effects are assumed to have their own subject-specific mean response trajectories over the time period. In mixed models, the fixed effects are separated by the mean response from the whole study population and therefore are fixed among all individual in that population. This is generally denoted by β. The random effects on the other hand are effects that are unique to a particular individual or particular group in the study (Fitzmaurice, Laird, & Ware, 2012).

Let $\boldsymbol{y_{ij}}$ be the response of the jth individual at the ith cluster, where $i = 1, 2, ..., n$ and $j = 1, 2, ..., N_i$ so that N_i is the cluster size. Let $\boldsymbol{x_{ij}}$ be a vector of p_f fixed effects covariates and $\boldsymbol{z_{ik}}$ be a vector of p_r random effects covariates with intercept as their first element respectively. Then $\boldsymbol{y_{ij}}$'s mixed model equation can be written as:

$$\boldsymbol{y_{ij}} = \boldsymbol{x_{ij}^T}\boldsymbol{\beta} + \boldsymbol{z_{ij}^T}\boldsymbol{b_i} + \epsilon_{ij} \tag{3.1}$$

where, $\boldsymbol{x_{ij}} = (1, x_{ij1}, x_{ij2}, ..., x_{ij(p_f-1)})^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_{(p_f-1)})^T$ are vector of $p_f \times 1$ where as $\boldsymbol{z_{ij}} = (1, z_{ij1}, z_{ij2}, ..., x_{ij(p_r-1)})^T$ and $\boldsymbol{b} = (b_{i0}, b_{i1}, b_{i2}, ..., b_{i(p_r-1))^T}$ are vector of $p_r \times 1$ Alternatively, this effect is written in matrix form in the following way:

$$\boldsymbol{Y_i} = \boldsymbol{\eta_i} = \boldsymbol{X_i^T}\boldsymbol{\beta} + \boldsymbol{Z_i^T}\boldsymbol{b_i} + \boldsymbol{\epsilon_i} \tag{3.2}$$

where,

$$\boldsymbol{X_i} = \begin{bmatrix} 1 & X_{i11} & X_{i12} & X_{i13} & ... & X_{i1(p_f-1)} \\ 1 & X_{i21} & X_{i22} & X_{i23} & ... & X_{i2(p_f-1)} \\ . & . & . & ... & . & \\ . & . & . & ... & . & \\ . & . & . & ... & . & \\ 1 & X_{iN_i1} & X_{iN_i2} & X_{iN_i3} & ... & X_{nN_i(p_f-1)} \end{bmatrix}^T ; \boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & ... & \beta_{(p_f-1)} \end{bmatrix}^T ;$$

$$\boldsymbol{Z_i} = \begin{bmatrix} 1 & Z_{i11} & Z_{i12} & Z_{i13} & ... & Z_{i1(p_f-1)} \\ 1 & Z_{i21} & Z_{i22} & Z_{i23} & ... & Z_{i2(p_f-1)} \\ . & . & . & ... & . & \\ . & . & . & ... & . & \\ . & . & . & ... & . & \\ 1 & Z_{iN_i1} & Z_{iN_i2} & Z_{iN_i3} & ... & Z_{nN_i(p_f-1)} \end{bmatrix}^T ; \boldsymbol{b_i} = \begin{bmatrix} b_{i0} & b_{i1} & b_{i2} & ... & b_{i(p_r-1)} \end{bmatrix}^T .$$

and $\boldsymbol{\epsilon_i} = \begin{bmatrix} \epsilon_{i0} & \epsilon_{i1} & \epsilon_{i2} & ... & \epsilon_{iN_i} \end{bmatrix}^T$.

The error term $\boldsymbol{\epsilon_i} \sim N(0, \sigma^2 \boldsymbol{I_i})$ and $\boldsymbol{b_i} \sim N(0, \sigma^2 \boldsymbol{Q})$ and the matrix $\boldsymbol{Q}$ is positive definite. Thus, for a linear mixed model, $\boldsymbol{Y_i} \sim N(\boldsymbol{X_i\beta}, \sigma^2 \boldsymbol{D_i})$ where $\boldsymbol{D_i = I_i + Q}$.

For the given model in 3.2, the log-likelihood function is given as:

$$l(\boldsymbol{\beta}, \sigma) = -\frac{1}{2} \sum_{i=1}^{n} log|\sigma^2 \boldsymbol{D_i}| - \frac{1}{2} \sigma^2 \sum_{i=1}^{n} \boldsymbol{(Y_i - X_i\beta)^T D_i^{-1} (Y_i - X_i\beta)} \tag{3.3}$$

Therefore, the corresponding maximum likelihood estimate is given as:

$$\tilde{\boldsymbol{\beta}} = (\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{X_i^T D_i^{-1} X_i})^{-1} (\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{X_i^T D_i^{-1} Y_i}) \tag{3.4}$$

In practice, this equation does not have a closed form solution and therefore

various numerical methods such as the EM algorithm and the Newton Raphson procedures are used in the calculation of the MLE.

3.1.2 Frailty Models in Survival Analysis

Typically, the random effects of mixed models that are used to model heterogeneity are done via frailty in survival data. Typical examples of the frailty factors include: genetic predisposition, economic capability and family history of diseases (Liu, 2012). For the Cox's proportional hazard frailty model or also called mixed proportional hazard model, the hazard rate of subject j belonging to cluster i, conditionally on the covariates $\boldsymbol{x_{ij}}$ and the frailty parameter $\boldsymbol{b_i}$ is given by the following:

$$\lambda(t|\boldsymbol{\beta}, \boldsymbol{b_i}) = \lambda_0(t) exp(\boldsymbol{x_{ij}^T \beta + z_{ij}^T b_i}) \tag{3.5}$$

where, $\lambda(t|\boldsymbol{\beta}, \boldsymbol{b_i})$ is the hazard for observation i at time t, conditioned on the fixed covariates $\boldsymbol{x_{ij}^T} = (1, x_{i1}, ..., x_{i(p_f-1)})$ and the random covariate $\boldsymbol{z_{ij}^T} = (1, z_{i1}, ..., z_{i(p_r-1)})$. The fixed effect coefficient is given as $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{(p_f-1)})$ while the frailty effect coefficent is $\boldsymbol{b_i} = (b_{i0}, b_{i1}, ..., b_{i(p_r-1)})$. $\lambda_0(t)$ is the baseline hazard (Groll et al., 2017).

The accelerated failure time (AFT) model, which is the log transformation of the survival time has a similar equation to the Linear Mixed Models. It is given by the following:

$$Log(T_{ij}) = \boldsymbol{x_{ij}^T \beta + z_{ij}^T b_i} + \epsilon_{ij} \tag{3.6}$$

where T_{ij} is the failure time event for the jth individual at the ith cluster (Do Ha, Jeong, & Lee, 2018).

VARIABLE SELECTION

Variable selection are the focus of many researches in areas involving data-sets with tens or hundreds of thousands of covariates. These areas include text processing of internet documents, gene expression array analysis, multi-centered clinical trails, combinatorial chemistry, etc. The objective of variable selection is typically three-folds: to improve the prediction performance of the chosen model, to provide faster and more cost-effective predictors, and to give a better understanding of the underlying process of the given data (Guyon & Elisseeff, 2003). In this section, we will review some of the most prominent variable selection procedures and describe the typical criterion values used to choose the ideal model.

3.2.1 Least Squares and Maximum Likelihood Estimates

For a typical simple general linear model, the maximum likelihood estimate and the least squares estimate have been the cornerstone for prediction. These can be calculated by the following: Let for a sample of observations, $\boldsymbol{x_i} = (x_{i1}, x_{i2}, ..., x_{ip})^T$ with $i = 1, 2, ..., n$ be the number of covariates and y_i be their responses. Then, the linear regression model can be written as:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i$$

Alternatively,

$$\boldsymbol{Y = X\beta + \epsilon} \tag{3.7}$$

where,

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 & Y_2 & Y_3 & ... & Y_n \end{bmatrix}^T ; \boldsymbol{X} = \begin{bmatrix} X_{11} & X_{12} & X_{13} & ... & X_{1p} \\ X_{21} & X_{22} & X_{23} & ... & X_{2p} \\ . & . & . & ... & . \\ X_{n1} & X_{n2} & X_{n3} & ... & X_{np} \end{bmatrix} ; \boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \beta_2 & ... & \beta_p \end{bmatrix}^T ;$$

and $\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 & \epsilon_2 & ... & \epsilon_n \end{bmatrix}^T$.

Here, $\boldsymbol{\epsilon}$ is the error term and follows a normal distribution with mean zero and constant variance, $N(0, \sigma^2)$.

Then, the residual sums of squares is given as:

$$RSS = \boldsymbol{(Y - X\beta)^T (Y - X\beta)} \tag{3.8}$$

The least square estimate, $\hat{\boldsymbol{\beta}}_{\boldsymbol{LSE}}$ is chosen such that it minimizes the RSS. The log-likelihood from the above equation is:

$$l(\boldsymbol{\beta}) \propto -\boldsymbol{(Y - X\beta)^T (Y - X\beta)} \tag{3.9}$$

By maximizing the $l(\boldsymbol{\beta})$, one can estimate the maximum likelihood estimate (MLE), $\hat{\boldsymbol{\beta}}_{MLE}$ for the model. It can be demonstrated that, maximizing the likelihood is equivalent to minimizing the RSS and under a normal error assumption, the two estimates are equivalent.

$$\hat{\boldsymbol{\beta}} = \boldsymbol{(X'X)^{-1}(X^T Y)} \tag{3.10}$$

For building models, the MLE and LSE techniques are popular typically for their easy implementation and interpretation. However, when there is a large set of covariates with multiple correlations, both the MLE and LSE suffer from large variance resulting in poor predictions and heavy unreliability. Additionally, they do not address model selection. As a result, alternatives have been proposed to gain prediction accuracy and to achieve simpler

models via model selection.

3.2.2 Subset Selection

The process of selecting a subset of significant predictors for building a model is termed as subset selection. Commonly used procedures for this type of variable selection include the following:

3.2.2.1 Forward Selection:

Forward selection is a process of subset selection where at first no predictors are in the model, then predictors are added one by one, and with the adding of each predictor, it is tested for significance (p-value) in the model. Only the most significant predictor is included in the model. This procedure is repeated until we finish adding only the significant predictors at the end (J. Pan, 2016).

3.2.2.2 Backward Elimination

Backward elimination is a complete opposite of the forward selection process. In backward elimination, all predictors in the model are added in the model first, and then the non-significant useful predictor is removed one after the other at a time. In the end, only the significant covariates are in the model. It is to be noted that models selected from forward selection and backward elimination may not always be the same (J. Pan, 2016).

3.2.2.3 Step-wise Selection

The step-wise selection uses the combination of forward selection and backward elimination procedures. At first, similar to the forward selection the significant predictors are sequentially added to the model however predictors would be removed if they are not significant at some point in time. This is a hybrid approach of the two and can imitate best subset selection while maintaining the same computational advantage (J. Pan, 2016).

3.2.2.4 Best Subset Selection

The best subset selection is to fit a separate global score chi-square statistic for each possible combination of the predictors. That is, all models are fitted with all the combinations of predictors. The criterion to choose the best set depends on the global Chi-square test statistics. The model with the highest score has the best rank. Though it is a good way of choosing models, this procedure becomes a problematic computational burden when the number of predictors grows. If there are i number of predictors, then there is 2^i number of candidate models. So as the predictors grow, the candidate number is growing very rapidly, and in general, the best subset selection becomes unfeasible when these predictors are greater than thirty (J. Pan, 2016).

From the information above, it can be seen that subset selection are very simple and computational feasible. However, they are not always stable and vary highly as the predictors are either retained or discarded from the model even with a small change in the data. This inhibits prediction accuracy of the subset selection procedure (Breiman et al., 1996; Fan & Li, 2001). Also, when there are correlated predictors or a large number of

predictors (or both), the lack of stability of the subset selection is then even more magnified (Harrell, 2001).

3.2.3 Penalized Selection

The penalized selection was created to address the issues in subset selection procedures and to develop a stable set of predictors. This selection process uses penalty function in the likelihood function of the regression coefficients and then maximizes it. Also, in turn, this approach helps in shrinking some of the coefficient estimates that have minimal impact in the model to go towards zero. This is why penalized estimates are often related to shrinking estimates. It can be seen that penalized selection can achieve the same purpose of subset selection but in a slightly more stable, continuous, and computationally efficient way. The penalized likelihood function is generally written in the following way:

$$argmax_{\beta}\Big\{l(\boldsymbol{\beta}) - \lambda P(\boldsymbol{\beta})\Big\}$$

where, $l(\boldsymbol{\beta})$ is the log-likelihood function, $P(\boldsymbol{\beta})$ is the penalized function and λ is the tuning parameter. The value of the tuning parameter ($\lambda \geq 0$) determines the amount of shrinkage for the given model. That is, the greater the value of the tuning parameter, the higher the shrinkage. The procedures to select the optimal tuning parameter value is described in the later section. Based on the above equation, the different types of penalized function can be described by the following:

3.2.3.1 Ridge Regression

Ridge regression was started in the 1970s by (Hoerl & Kennard, 1970). It was originally introduced to address the problem of having high correlated covariates in the linear regression models and was one of the first of its kind to induce the concept of

shrinkage estimations.

That is, in a given likelihood approach, the ridge penalty is:

$$\hat{\boldsymbol{\beta}}_{RIDGE} = argmax_{\beta}\left\{ l(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} (\boldsymbol{\beta_j^2}) \right\} \tag{3.11}$$

Based on this equation, the ridge estimate can be solved as:

$$\hat{\boldsymbol{\beta}}_{RIDGE} = (\boldsymbol{X'X} + \lambda \boldsymbol{I})^{-1}(\boldsymbol{X^T Y}) \tag{3.12}$$

The above equation is a closed form solution which makes ridge regression a unique advantage over most other procedures. It can be seen that as the value of λ increases, the minimal influential elements (values closer to zero) of $\boldsymbol{\beta}$ shrinks. The ridge regression has several distinct benefits in its usage. First, it uses a continuous process to shrink noise coefficients and it can also shrink even in substantial collinearity. Therefore it can produce a stable model at the end. Second, there is a big chance that it may have a smaller prediction error than the regular least squares and maximum likelihood estimates. However, there is a significant disadvantage in this type of penalized function. This approach does not always shrink the coefficients to precisely zero. Therefore, the final model is usually not simple enough for a more straightforward interpretation (Yu, 2007; J. Pan, 2016).

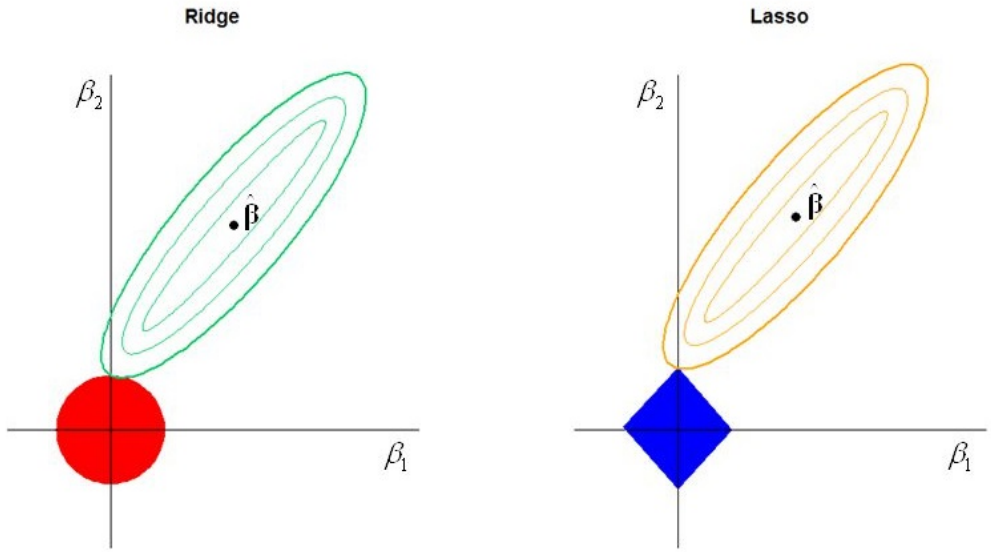
3.2.3.2 LASSO

(R. Tibshirani, 1996), introduced the Least Absolute Shrinkage and Selection Operator (LASSO). It became a widely popular method because of its ability to shrink the coefficients directly to zero. This estimate is given by the following:

$$\hat{\boldsymbol{\beta}}_{LASSO} = argmax_{\beta}\left\{ l(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} (|\boldsymbol{\beta_j}|) \right\} \tag{3.13}$$

The LASSO penalty uses the L1-norm $(|.|)$, absolute value or $\sum_{j=1}^{p}(|\boldsymbol{\beta_j}|)$, that is different from the Ridge penalty which uses the L2-norm $(|.|^2)$ or $(||.||)$, square of the given value or $\sum_{j=1}^{p}(\boldsymbol{\beta_j}^2)$. A suitable way to describe the difference between the LASSO and Ridge penalty can be summarized by Figure 1.

Figure 3.1: Ridge regression (left) and LASSO (right) estimation procedure



As seen in the above figure, the likelihood function from the covariates has an elliptical outline with the center having the maximum likelihood estimate. The constraint region is a disk-shaped structure for ridge regression whereas it is a diamond for the LASSO. Both methods find the first point where the elliptical outline hit the constraint region. It is intuitive that if the coefficients are near the corner of the diamond shaped region of the LASSO type penalty, then the coefficient easily shrinks to zero faster as compared to the round shaped area of the ridge regression. This ability to shrink coefficients leads to the formation of sparse models. This feature gives LASSO a significant advantage over traditional model selection procedures.

Although LASSO has a lot of great features in its application unlike ridge regression, it does not have a closed form solution because the objective function is not differential (J. Pan, 2016). Moreover, LASSO equation is assumed to have a strictly convex structure, but in data sets where covariates number is larger than the sample size number, the LASSO structure may not be purely convex, so there may not be a unique solution (R. J. Tibshirani et al., 2013).

Therefore, there has been significant literature conducted to solve the LASSO solution problem. Some of the traditional methods include coordinate descent, first-order methods, and quadratic programming approaches. However, these methods are not consistently producing an active set of solutions that satisfy the LASSO lemma (R. J. Tibshirani et al., 2013). One, of the most famous algorithm to solve the LASSO problem, is the Least Angle Regression (LARS) (Efron, Hastie, Johnstone, Tibshirani, et al., 2004). It is a more democratic version of the forward stage-wise solution using the least squares. Additionally, there are more studies to solve the LASSO optimization problem. They will be discussed later in the chapter.

3.2.3.3 SCAD

The Smoothly Clipped Absolute Deviation (SCAD) penalty was developed by Fan and Li in 2001 to minimize the shortcomings in the LASSO procedure (Fan & Li, 2001). The idea is to put a substantial penalty on the smaller coefficients and a light penalty on the more significant factors. This approach helps in preserving the essential effects while shrinking the less influential covariates (J. Pan, 2016). The function of SCAD is symmetric, non-concave on $(0, \infty)$ and has singularities at the origin to produce sparse solutions. The SCAD penalty function is given by the following:

$$P_{SCAD(\lambda)}(\boldsymbol{\beta}) = \lambda\left\{ I(\boldsymbol{\beta} \leq \lambda) + \frac{(a\lambda - \boldsymbol{\beta})_+}{(a-1)\lambda} I(\boldsymbol{\beta} > \lambda) \right\} \tag{3.14}$$

where $a > 0$ and $\lambda > 0$. Then the solutions for β can be given as:

$$\hat{\boldsymbol{\beta}} = \begin{cases} sgn(z)(|z| - \lambda)_+ & \text{if} \quad |z| \leq 2\lambda \\ \frac{\{(a-1)z - sgn(Z)a\lambda\}}{(a-2)} & \text{if} \quad 2\lambda < |z| \leq a\lambda \\ z & \text{if} \quad |z| > a\lambda \end{cases} \tag{3.15}$$

where $\boldsymbol{z = x^T y}$. One of the best advantage of SCAD penalty is its ability to satisfy the oracle property. This property states that for a good selection procedure δ, the estimator $\hat{\boldsymbol{\beta}}(\delta)$ should satisfy the following two conditions:

1. It is able to identify the right model, $\{j : \hat{\boldsymbol{\beta}}(\delta)_j \neq 0\} = \{j : \boldsymbol{\beta}_j \neq 0\}$

2. Has the optimal estimate rate, $\sqrt{n}(\hat{\boldsymbol{\beta}}(\delta) - \boldsymbol{\beta}) \rightarrow_d N(0, \Sigma)$, where Σ is the variance-covariance matrix of knowing the true subset model. It means that the covariates with nonzero coefficients can be identified with probability tending to one, and the estimates of nonzero coefficients have the same asymptotic distribution as the true model.

Although SCAD has such excellent properties, it is computationally challenging due to its complex form. Furthermore, since the optimization is non-concave, there is no certainty that the local-maximum of the given penalized likelihood can be the global maximum (J. Pan, 2016).

3.2.3.4 Elastic Net

The elastic net penalty was proposed in 2005 where Zou and Hastie introduced the Elastic Net penalty which linearly combines the LASSO and ridge regression procedure (Zou & Hastie, 2005). This penalty is given by the following:

$$\hat{\boldsymbol{\beta}}_{ELASTIC} = argmax_{\beta}\left\{ l(\boldsymbol{\beta}) - \lambda_1 \sum_{j=1}^{p} |\boldsymbol{\beta_j}| - \lambda_2 \sum_{j=1}^{p} (\boldsymbol{\beta_j})^2 \right\} \tag{3.16}$$

The above equation shows that elastic net penalty has two tuning parameters. One for the LASSO type penalty to achieve sparsity and the other for the ridge type penalty function to have group selection and stabilization. Therefore, by including both these function, the elastic net penalty can have good features from both sides and is a useful alternative when there is a group of predictors with high pairwise correlation. However, it may be computationally challenging to estimate the ideal values for the two different tuning parameters.

3.2.3.5 Group LASSO

The group lasso penalty was proposed in 2006 by Yuan and Lin (Yuan & Lin, 2006). This penalization works like a LASSO, but at a group wise level that is an entire group of predictors may be dropped out of the model. Thus, if the given data has all of its groups sizes as one then this penalty function changes into a regular LASSO penalty. The proposed penalty is given by the following:

$$\hat{\boldsymbol{\beta}}_{groupLASSO} = argmax_{\beta}\left\{ l(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} \boldsymbol{\beta_j}^2 \sqrt{N_i} \right\} \tag{3.17}$$

where, N_i accounts for the group size of the ith cluster.

This penalty has an attractive property where the group level variable selection is invariant under (group-wise) orthogonal transformations like ridge regression. This leads to closed form solution like ridge regression especially in large scale application studies (Yuan & Lin, 2006). However, this penalty can only yield solutions to sparsity at the group level. Hence, we need some modifications in this type of penalty to achieve the individual sparsity in a given model.

3.2.3.6 Adaptive LASSO

(Zou, 2006) proposed the adaptive LASSO to address the inconsistency of the LASSO penalty . He modified the original LASSO function to include adaptive weight vector that can be adapted by the data-set. The following equation gives the penalty function:

$$\hat{\boldsymbol{\beta}}_{ALASSO} = argmax_{\beta}\left\{ l(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} \boldsymbol{w_j}|\boldsymbol{\beta_j}| \right\} \tag{3.18}$$

where $\boldsymbol{w_j} = (w_1, ..., w_p)$ is the given weight vector. Therefore, the choice of this weight is significant to have consistency. Usually, the weights are calculated by this equation:

$$\boldsymbol{w} = \frac{\mathbf{1}}{\hat{\boldsymbol{\beta}}} \tag{3.19}$$

where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$. These weights help to incorporate substantial penalties for insignificant covariates and small penalties for significant covariates. It improves model accuracy and reduces estimated bias and variance.

The adaptive LASSO has a great many useful features that are consistent with the penalty functions as mentioned earlier. Therefore, it is an excellent alternative to SCAD.

3.2.4 Selection Criteria for Model Selection

The penalty functions described in previous section showed that there are a set of one or more tuning parameters that determine the amount of penalty for a given model. The choice for the ideal values for these tuning parameters requires a thorough investigation from a sequence of assigned values. Such examination involves the appropriate

model selection criterion that can score each fitted models from the given series of tuning parameters. These scores assist in choosing the suitable tuning parameter for the given model. Some of the most popular criteria are as follows:

3.2.4.1 Akaike Information Criterion (AIC)

Discovered in the early 1970s, the Akaike information criterion (AIC) remains one of the most influential and popular tool for model selection (Akaike, 1973, 1974). It is an estimator to determined the expected Kullback discrepancy between the given fitted model and the truth model. AIC is a likelihood-based measure for model fit, and in general, it is written as:

$$AIC = -2l(\hat{\boldsymbol{\beta}}) + 2 * p \tag{3.20}$$

For mixed models, Sugiuria (Sugiura, 1978) proposed the marginal AIC derived from the marginal form of the linear mixed models and is given as:

$$AIC = -2l(\hat{\boldsymbol{\beta}}) + a_N(2p_f + 2p_r) \tag{3.21}$$

where $l(\hat{\boldsymbol{\beta}})$ is the log-likelihood function, $\hat{\boldsymbol{\beta}}$ is the required estimate, p_f is the number of estimated fixed parameters and p_r is the number of estimated random effects. The value of a_N is usually 1 or $\frac{N}{N-p_f-p_r-1}$. For model selection, the optimal value for the given model is equal to the minimum value of the AIC. This property to identify a suitable fitted model has a significant role in various fields, especially when the data-sets are large. But, for small sample size, AIC may select over-fitting models which means that it may not be an effective option for selection criterion in low sample size. Therefore, to address this gap in AIC, the corrected AIC (AICc) was proposed. AICc of a fixed effects only model can be given by the following equation:

$$AICc = AIC + \frac{2p_f(p_f + 1)}{N - p_f - 1} \tag{3.22}$$

3.2.4.2 Bayesian Information Criterion (BIC)

Developed by Schwarz in 1978, the Bayesian information criterion (BIC) is an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model (Schwarz et al., 1978). Like the AIC, this is also a likelihood base measure for model fitting. The model with the corresponding minimum value of BIC is the candidate model with the highest Bayesian posterior probability. The following equation represents the BIC information criterion:

$$BIC = -2l(\hat{\boldsymbol{\beta}}) + p * log(N) \tag{3.23}$$

For mixed models, the marginal BIC is similar to the marginal AIC and is written as:

$$BIC = -2l(\hat{\boldsymbol{\beta}}) + log(N)(p_f + p_r) \tag{3.24}$$

In the above equation, for $N \geq 8, log(N)(p_f + p_r)$ exceeds the AIC's $2 * (p_f + p_r)$. Therefore, BIC has more strict penalty than AIC, and thus it tends to choose smaller models than the AIC especially when there is a large sample application. BIC is also an asymptotically consistent tool in choosing a correct model with a probability of one.

However, BIC is not asymptotically efficient and thus the chosen model via BIC may not be able to minimize the mean squared error of prediction (Weakliem, 1999). Therefore, it may not be a primary tool for predictive model selection procedures.

3.2.4.3 ICs

(Hui et al., 2017), developed their ICs information criterion for model selection in . Here they consider a range of tuning parameters, $\lambda_{min}, \lambda_{max}$ where λ_{min} is the full model containing all the candidate fixed and random effects and λ_{max} is the value that

leads to the null model. The ICs criterion selects an ideal value for λ that can have the minimal value for the following equation:

$$IC(\lambda) = -\frac{2}{N} l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{b}}) + \frac{log(n_i)}{N} dim(\hat{\boldsymbol{\beta}}) + \frac{2}{N} dim(\hat{\boldsymbol{b}}) \tag{3.25}$$

where $dim(\hat{\boldsymbol{\beta}})$ and $dim(\hat{\boldsymbol{b}})$ are the number of nonzero estimated fixed and random effects coefficients respectively and N is the total sample size where as n_i is cluster size for a given cluster i.

This criterion combines the features of BIC and AIC for fixed and random effects respectively. The main advantage of this type of criterion is its ability to prevent over-fitting on random effects via group sparsity (Hui et al., 2017).

3.2.4.4 Cross-Validation

Cross-validation (CV) is a widely popular strategy for model evaluation and selection. It randomly divides the data into K groups, or folds, of approximately equal size. That is, for $k = 1, 2, ..., K$ the validation set be the kth fold of the data. It is termed as the test data.

The remaining $K - 1$ folds are set to be the training data and model is fitted in this training set. Then, the prediction error of the fitted model with the validation set can thus be computed. For each model, this process is repeated for K times. That way, each fold is used to be in the validation set which gives K estimates of the prediction error, and thus the CV is computed by averaging these values. Under this criterion, the best model is the one with the smallest value. Usually, five or ten-fold cross-validation is recommended (Breiman, 1995).

For a sample to be divided into K folds, it has to be large enough. Therefore, in a small sample size, it may cause unstable estimates. Furthermore, the computational time for evaluating and producing K estimates is very long. Therefore, to reduce this computational burden, bias correction cross-validation have been proposed (Bernau, Augustin, & Boulesteix, 2013).

CHAPTER 4

PROPOSED METHOD

The primary purpose of this study is to extend the works of Breslow and Clayton (Breslow & Clayton, 1993), (Hui et al., 2017) and (Groll et al., 2017) in variable selection by applying the Penalized Quasi-Likelihood (PQL) procedure to the parametric AFT models. We will investigate the variable selection in a model that has a dependent variable that follows different survival distributions under the assumption of left censoring, with several of fixed effects covariates X along with several random effects covariates Z.

BASIC NOTATIONS AND FORMULAS

Let $f(t)$ be the probability density function (pdf) of a given continuous time variable T. Then the cumulative distribution function of the random variable T with a probability that an event occurs within a given time interval (0,t) is given by:

$$F(t) = Pr(T \leq t) = \int_0^t f(u)du \tag{4.1}$$

The survival function $S(t)$ is the complement of this given cumulative density function which means, it is the probability that the individual will survive beyond a given time which can be prseneted by the following:

$$S(t) = Pr(T > t) = 1 - Pr(T \leq t) = 1 - F(t). \tag{4.2}$$

Note that, $S(0) = 1$ and $S(\infty) = 0$. The probability density function $f(t)$ can also be written in terms of the survival function as

$$f(t) = -\frac{dS(t)}{dt}. \tag{4.3}$$

The hazard function $h(t)$ is the instantaneous rate of failure at time t .

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr[T \in (t, t + \Delta t) | T \geq t]}{\Delta t}, \tag{4.4}$$

or equivalently,

$$h(t) = \frac{f(t)}{S(t)} = -\frac{1}{S(t)} \frac{dS(t)}{dt} = \frac{-dlogS(t)}{dt}. \tag{4.5}$$

The above equations show that the three functions, namely $f(t)$, $S(t)$ and $h(t)$ are intimately related to each other. Thus, if one of these functions is available, then the other two can be easily calculated. For example, $S(t)$ can be written as an inverse function of equation (3.5) as:

$$S(t) = \exp(-\int_0^t h(u)du) = \exp[-H(t)], \tag{4.6}$$

where $H(t)$ is the integration of all hazard rates up to time t and is known as the cumulative hazard function at time t. Alternatively, $H(t)$ can also be written in terms of $S(t)$ as:

$$H(t) = -logS(t). \tag{4.7}$$

Furthermore, the probability density function can also be written in the following form, from equations (3.5) and (3.6):

$$f(t) = h(t) \exp(-\int_0^t h(u)du). \tag{4.8}$$

ACCELERATED FAILURE TIME (AFT) WITH FRAILTY EFFECTS

Let $\boldsymbol{T_i}$ be the failure time and the dependent variable $\boldsymbol{y_i} = \log(\boldsymbol{T_i})$ is linearly associated with the fixed covariate vector $\boldsymbol{x_{ij}^T}$ and the random covariate (frailty) vector $\boldsymbol{z_{ij}}$, where $i = 1, 2, ..., n$ is the number of clusters and $j = 1, 2, 3, ..., N_i$ is the number of measurements within the ith cluster. Let p_f be the number of covariates associated with

fixed effects $\boldsymbol{x_{ij}^T}$ and p_r be the number of covariates associated with frailty effects $\boldsymbol{z_{ij}^T}$. Then AFT frailty model with a log-linear link can be given as:

$$\boldsymbol{y_i = \beta_0 + x_i^T \beta + z_{ij}^T b_i} + \epsilon_i, \tag{4.9}$$

Here, $\boldsymbol{\beta}$ is the fixed effect coefficient and $\boldsymbol{b_i}$ is the random effect coefficient. For this study, it is assumed that $\boldsymbol{b_i} \sim N(0, \boldsymbol{Q})$, where $\boldsymbol{Q}$ is the variance-covariance matrix. The term ϵ_i represents the random error whose distribution is determined by the survival function of time, $S(t)$, the cumulative distribution function, $F(t)$, and the probability density function, $f(t)$.

From equation 4.9, for a specific observation i, the lifetime process can be described by three factors. They are:

1. Random variable of the event time T_i and censoring time C_i,

2. Observed survival time t_i from an independent and identically distribution with left censored mechanism.

3. Random variable indicating the status of surviving or left censoring for a given t_i. This implies that:

$$\delta_{ij} = 0 \quad if \quad T_i = t_i, \tag{4.10}$$

$$\delta_{ij} = 1 \quad if \quad T_i < t_i. \tag{4.11}$$

Then with δ_{ij} denotes the censoring indicator, consequently t_i is a lifetime $(\delta_{ij} = 0)$ or a left censored time $(\delta_{ij} = 1)$.

STATISTICAL INFERENCE ON AFT MODELS WITH FRAILTY

Statistical inference in survival analysis is different and unique as the censoring mechanism plays an essential role in the determination of likelihood functions. From equation (4.9), it can be deduced that the survival function for an individual at ij at time t_i can be written as:

$$S(t_i) = P(\epsilon_{ij} \geq logt_i - \beta_0 - \boldsymbol{x_{ij}^T\beta} - \boldsymbol{z_{ij}^T b_i}). \tag{4.12}$$

Let $\eta_{ij} = \boldsymbol{x_{ij}^T\beta} + \boldsymbol{z_{ij}^T b_i}$, Then in AFT models, the effect of covariates is such that if $\exp(\eta_{ij}) > 1$, a deceleration of the survival (time) process ensues and if $\exp(\eta_{ij}) < 1$, then an acceleration of the survival (time) process ensues.

Given the covariate vector $\boldsymbol{x_i}$, random vector $\boldsymbol{z_{ij}}$ and parameter vector $\boldsymbol{\beta}$, $\boldsymbol{b}$, the likelihood function for a left censored mechanism is given by

$$L(\boldsymbol{\beta}, \boldsymbol{b}) = \prod_{i=1}^{n}\prod_{j=1}^{N_i} f(t_{ij}|\boldsymbol{\beta}, \boldsymbol{b_i})^{1-\delta_{ij}}(F(t_{ij}|\boldsymbol{\beta}, \boldsymbol{b_i}))^{\delta_{ij}}. \tag{4.13}$$

From equation (4.13), when $\delta_{ij} = 0$ the likelihood function takes on the value of the probability density function for the occurrence of an event. When $\delta_{ij} = 1$, the likelihood function takes on the complement value of the probability of survival beyond censoring time t, which is the cumulative distribution function. The same likelihood function in terms of a parametric regression model with a hazard function and a vector of coefficients $\boldsymbol{\beta}$ is defined by.

$$L(\boldsymbol{\beta}, \boldsymbol{b}) = \prod_{i=1}^{n}\prod_{j=1}^{N_i}[h(t_{ij})S(t_{ij})]^{1-\delta_{ij}}(1 - S(t_{ij}))]^{\delta_{ij}}. \tag{4.14}$$

However, for this study, we will use equation (4.13). Taking the log values on both sides of equation (4.13):

$$l(\boldsymbol{\beta}, \boldsymbol{b}) = \sum_{i=1}^{n} \sum_{j=1}^{N_i} \left\{ (1 - \delta_{ij}) log f(t_{ij}|\boldsymbol{\beta}, \boldsymbol{b_i}) + \delta_{ij} (log F(t_{ij}|\boldsymbol{\beta}, \boldsymbol{b_i})) \right\}. \tag{4.15}$$

The given full log-likelihood in equation (4.15) does not have a closed form solution to derive the estimation of the maximum likelihood. An alternative way to estimate the maximum value of this likelihood is to integrate out the random effects and maximize the eventual marginal likelihood. To approximate this marginal likelihood, the penalized quasi-likelihood (PQL) approach developed by Breslow and Clayton (Breslow & Clayton, 1993) could be used here. Therefore, let $\boldsymbol{\phi} = (\beta, \boldsymbol{vech(Q)})$ where vech(.) is the half vectorization operator of the variance-covariance matrix $\boldsymbol{Q}$, the vector of the lower triangular matrix of $\boldsymbol{Q}$ as given by (Hui et al., 2017). Following Ripatti and Palmgren (Ripatti & Palmgren, 2000), the marginal likelihood can have the following integral form:

$$L^{mar}(\boldsymbol{\phi}, \boldsymbol{b}) = \prod_{i=1}^{n} \left\{ \int L_i(\beta, \boldsymbol{b}) f(\boldsymbol{b_i}; \boldsymbol{Q}) \boldsymbol{db_i} \right\}. \tag{4.16}$$

where $L_i(\boldsymbol{\beta}, \boldsymbol{b})$ is given by $\prod_{j=1}^{N_i} f(t_{ij}|\boldsymbol{\beta}, \boldsymbol{b_i})^{1-\delta_{ij}} (F(t_{ij}|\boldsymbol{\beta}, \boldsymbol{b_i}))^{\delta_{ij}}$ and $f(\boldsymbol{b_i}; \boldsymbol{Q})$ is represented by the density function for the random effects. It is assumed that the frailty effects follows the normal distribution, $f(\boldsymbol{b_i}; \boldsymbol{Q}) \sim N(0, \boldsymbol{Q})$. Thus, equation (4.16) can be approximated by the following :

$$L^{mar}(\boldsymbol{\phi}, \boldsymbol{b}) \propto \prod_{i=1}^{n} \left\{ |\boldsymbol{Q}|^{-\frac{1}{2}} \int exp \left\{ log L_i(\boldsymbol{\beta}, \boldsymbol{b}) - \frac{1}{2} \tilde{\boldsymbol{b}}_{\boldsymbol{i}}^{\boldsymbol{T}} \boldsymbol{Q^{-1}} \tilde{\boldsymbol{b}}_{\boldsymbol{i}} \boldsymbol{db_i} \right\} \right. \tag{4.17}$$

The given equation leads to an intractable integration such that it does not have a closed form solution. However, since it has the form $c|\boldsymbol{Q}|^{-\frac{1}{2}} \int e^{-k(\boldsymbol{b})}$, we can approximate the solution using the Laplace approximation. Thus, along the lines of (Breslow & Clayton, 1993), (Ripatti & Palmgren, 2000) and (Groll et al., 2017), the marginal log-likelihood from equation (4.17) can be approximated by the following form:

$$l^{LA}(\boldsymbol{\phi}, \boldsymbol{b}) = \frac{1}{2} log|\boldsymbol{Q}| - \frac{1}{2} log|k''(\tilde{\boldsymbol{b}})| - k(\tilde{\boldsymbol{b}}). \tag{4.18}$$

where the terms inside $|.|$ are the determinants of the matrix. In addition, from equation (4.15)

$$k(\tilde{\boldsymbol{b}}) = -\sum_{i=1}^{n}\sum_{j=1}^{N_i}\left\{(1-\delta_i)logf(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i}) + \delta_i(logF(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i})) - \frac{1}{2}\tilde{\boldsymbol{b}}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{Q}^{-\mathbf{1}}\tilde{\boldsymbol{b}}_{\boldsymbol{i}}\right\}. \quad (4.19)$$

where $\frac{1}{2}\boldsymbol{b}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{Q}^{-\mathbf{1}}\boldsymbol{b}_{\boldsymbol{i}}$ is the ridge penalty. This penalty function is also able to penalize the extreme values of $\boldsymbol{b}$. Alternatively, equation (4.19) can also be written as:

$$k(\tilde{\boldsymbol{b}}) = -\sum_{i=1}^{n}\left\{logL_i(\boldsymbol{\beta},\boldsymbol{b}) - \frac{1}{2}\tilde{\boldsymbol{b}}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{Q}^{-\mathbf{1}}\tilde{\boldsymbol{b}}_{\boldsymbol{i}}\right\}. \quad (4.20)$$

Now, by taking the derivative of $k(\tilde{\boldsymbol{b}})$ with respect to $\tilde{\boldsymbol{b}}$ in equation (4.19), we get the first derivative, $k'(\tilde{\boldsymbol{b}})$ as:

$$k'(\tilde{\boldsymbol{b}}) = -\sum_{i=1}^{n}\sum_{j=1}^{N_i}\left\{(1-\delta_i)\frac{\partial logf(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i})}{\partial \boldsymbol{b_i}} + \delta_i\frac{\partial(logF(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i}))}{\partial \boldsymbol{b_i}} - \boldsymbol{Q}^{-\mathbf{1}}\tilde{\boldsymbol{b}}_{\boldsymbol{i}}\right\}.$$

$$\Rightarrow k'(\tilde{\boldsymbol{b}}) = -\sum_{i=1}^{n}\sum_{j=1}^{N_i}\left\{(1-\delta_i)\frac{\partial logf(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i})}{\partial \boldsymbol{\epsilon_i}}\frac{\partial \boldsymbol{\epsilon_i}}{\partial \boldsymbol{b_i}} + \delta_i\frac{\partial(logF(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i}))}{\partial \boldsymbol{\epsilon_i}}\frac{\partial \boldsymbol{\epsilon_i}}{\partial \boldsymbol{b_i}} - \boldsymbol{Q}^{-\mathbf{1}}\tilde{\boldsymbol{b}}_{\boldsymbol{i}}\right\}.$$

$$\Rightarrow k'(\tilde{\boldsymbol{b}}) = -\sum_{i=1}^{n}\sum_{j=1}^{N_i}\left\{(1-\delta_i)\frac{\partial logf(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i})}{\partial \boldsymbol{\epsilon_i}}\frac{\partial[\boldsymbol{y_i} - \boldsymbol{x_i^T}\boldsymbol{\beta} - \boldsymbol{z_{ij}^T}\boldsymbol{b_i}]}{\partial \boldsymbol{b_i}} + \delta_i\frac{\partial(logF(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i}))}{\partial \boldsymbol{\epsilon_i}}\frac{\partial[\boldsymbol{y_i} - \boldsymbol{x_i^T}\boldsymbol{\beta} - \boldsymbol{z_{ij}^T}\boldsymbol{b_i}]}{\partial \boldsymbol{b_i}} - \boldsymbol{Q}^{-\mathbf{1}}\tilde{\boldsymbol{b}}_{\boldsymbol{i}}\right\}.$$

$$\Rightarrow k'(\tilde{\boldsymbol{b}}) = -\sum_{i=1}^{n}\sum_{j=1}^{N_i}\left\{(1-\delta_i)\frac{\partial log f(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i})}{\partial \boldsymbol{\epsilon_i}}(-\boldsymbol{z_{ij}^T}) + \delta_i\frac{\partial(logF(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i}))}{\partial \boldsymbol{\epsilon_i}}(-\boldsymbol{z_{ij}^T}) - \boldsymbol{Q^{-1}}\tilde{\boldsymbol{b}}_{\boldsymbol{i}}\right\}.$$

$$\Rightarrow k'(\tilde{\boldsymbol{b}}) = -\sum_{i=1}^{n}\sum_{j=1}^{N_i}\left\{(-\boldsymbol{z_{ij}^T})\left\{(1-\delta_i)\frac{\partial log f(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i})}{\partial \boldsymbol{\epsilon_i}} + \delta_i\frac{\partial(logF(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i}))}{\partial \boldsymbol{\epsilon_i}}\right\} - \boldsymbol{Q^{-1}}\tilde{\boldsymbol{b}}_{\boldsymbol{i}}\right\}.$$

$$\Rightarrow k'(\tilde{\boldsymbol{b}}) = \sum_{i=1}^{n}\sum_{j=1}^{N_i}\left\{\boldsymbol{z_{ij}^T}\frac{\partial}{\partial \boldsymbol{\epsilon_i}}\left\{(1-\delta_i)log f(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i}) + \delta_i(logF(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i}))\right\} - \boldsymbol{Q^{-1}}\tilde{\boldsymbol{b}}_{\boldsymbol{i}}\right\}.$$

Taking derivative of this resulting $k'(\tilde{\boldsymbol{b}})$ again with respect to $\tilde{\boldsymbol{b}}$, we get the second derivative of $k(\tilde{\boldsymbol{b}})$, $k''(\tilde{\boldsymbol{b}})$, given by the following:

$$k''(\tilde{\boldsymbol{b}}) = \sum_{i=1}^{n}\sum_{j=1}^{N_i}\left\{\boldsymbol{z_{ij}^T}\frac{\partial}{\partial \boldsymbol{\epsilon_i}}\left\{(1-\delta_i)\frac{\partial log f(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i})}{\partial \boldsymbol{b_i}} + \delta_i\frac{\partial(logF(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i}))}{\partial \boldsymbol{b_i}}\right\} - \boldsymbol{Q^{-1}}\right\}.$$

$$\Rightarrow k''(\tilde{\boldsymbol{b}}) = \sum_{i=1}^{n}\sum_{j=1}^{N_i}\left\{\boldsymbol{z_{ij}^T}\frac{\partial}{\partial \boldsymbol{\epsilon_i}}\left\{(1-\delta_i)\frac{\partial log f(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i})}{\partial \boldsymbol{\epsilon_i}}\frac{\partial \boldsymbol{\epsilon_i}}{\partial \boldsymbol{b_i}} + \delta_i\frac{\partial(logF(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i}))}{\partial \boldsymbol{\epsilon_i}}\frac{\partial \boldsymbol{\epsilon_i}}{\partial \boldsymbol{b_i}}\right\} - \boldsymbol{Q^{-1}}\right\}.$$

$$\begin{aligned}\Rightarrow k''(\tilde{\boldsymbol{b}}) = \sum_{i=1}^{n}\sum_{j=1}^{N_i}\Bigg\{\boldsymbol{z_{ij}^T}\frac{\partial}{\partial \boldsymbol{\epsilon_i}}\Bigg\{&(1-\delta_i)\frac{\partial log f(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i})}{\partial \boldsymbol{\epsilon_i}}\frac{\partial[\boldsymbol{y_i} - \boldsymbol{x_i^T}\beta - \boldsymbol{z_{ij}^T}\boldsymbol{b_i}]}{\partial \boldsymbol{b_i}} \\ &+\delta_i\frac{\partial(logF(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i}))}{\partial \boldsymbol{\epsilon_i}}\frac{\partial[\boldsymbol{y_i} - \boldsymbol{x_i^T}\beta - \boldsymbol{z_{ij}^T}\boldsymbol{b_i}]}{\partial \boldsymbol{b_i}}\Bigg\} - \boldsymbol{Q^{-1}}\Bigg\}.\end{aligned}$$

$$\Rightarrow k''(\tilde{\boldsymbol{b}}) = \sum_{i=1}^{n}\sum_{j=1}^{N_i}\left\{\boldsymbol{z_{ij}^T}\frac{\partial}{\partial \boldsymbol{\epsilon_i}}\left\{(1-\delta_i)\frac{\partial log f(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i})}{\partial \boldsymbol{\epsilon_i}}(-\boldsymbol{z_{ij}}) + \delta_i\frac{\partial(logF(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i}))}{\partial \boldsymbol{\epsilon_i}}(-\boldsymbol{z_{ij}})\right\} - \boldsymbol{Q^{-1}}\right\}.$$

$$\Rightarrow k''(\tilde{\boldsymbol{b}}) = \sum_{i=1}^{n}\sum_{j=1}^{N_i}\left\{\boldsymbol{z_{ij}^T}\frac{\partial}{\partial \boldsymbol{\epsilon_i}}\left\{(1-\delta_i)\frac{\partial log f(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i})}{\partial \boldsymbol{\epsilon_i}} + \delta_i \frac{\partial (log F(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i}))}{\partial \boldsymbol{\epsilon_i}}\right\}(-\boldsymbol{z_{ij}}) - \boldsymbol{Q^{-1}}\right\}.$$

$$\Rightarrow k''(\tilde{\boldsymbol{b}}) = -\sum_{i=1}^{n}\sum_{j=1}^{N_i}\left\{\boldsymbol{z_{ij}^T}\frac{\partial^2}{\partial \boldsymbol{\epsilon_i^2}}\left\{(1-\delta_i)(log(f(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i}) + \delta_i log(F(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i}))\right\}\boldsymbol{z_{ij}} - \boldsymbol{Q^{-1}}\right\}.$$

Let $\boldsymbol{W_i} = \frac{\partial^2}{\partial \boldsymbol{\epsilon_i^2}}\left\{(1-\delta_i)(log(f(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i}) + \delta_i log(F(t_{ij}|\boldsymbol{\beta},\boldsymbol{b_i}))\right\}$, then we can write the aforementioned equation in a matrix form by the following:

$$k''(\tilde{\boldsymbol{b}}) = \sum_{i=1}^{n}\left\{\boldsymbol{Z_i^T W_i Z_i - Q^{-1}}\right\}. \tag{4.21}$$

Plugging the values of $k(\tilde{\boldsymbol{b}})$ and $k''(\tilde{\boldsymbol{b}})$ on equation (4.18), we get the following:

$$l^{LA}(\boldsymbol{\phi},\boldsymbol{b}) = \frac{1}{2}log|\boldsymbol{Q}| - \sum_{i=1}^{n}\frac{1}{2}log|\boldsymbol{Z_i^T W_i Z_i - Q^{-1}}| + \sum_{i=1}^{n}\left\{log(L_i(\boldsymbol{\beta},\tilde{\boldsymbol{b}})) - \frac{1}{2}\tilde{\boldsymbol{b}}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{Q^{-1}}\tilde{\boldsymbol{b}}_{\boldsymbol{i}}\right\}$$

$$\Rightarrow l^{LA}(\boldsymbol{\phi},\boldsymbol{b}) = \frac{1}{2}log|\boldsymbol{Q}| - \sum_{i=1}^{n}\frac{1}{2}log|\frac{\boldsymbol{Z_i^T W_i Z_i Q - I}}{\boldsymbol{Q}}| + \sum_{i=1}^{n}\left\{log(L_i(\boldsymbol{\beta},\tilde{\boldsymbol{b}})) - \frac{1}{2}\tilde{\boldsymbol{b}}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{Q^{-1}}\tilde{\boldsymbol{b}}_{\boldsymbol{i}}\right\}$$

$$\Rightarrow l^{LA}(\boldsymbol{\phi},\boldsymbol{b}) = \frac{1}{2}log|\boldsymbol{Q}| - \sum_{i=1}^{n}\frac{1}{2}log|\boldsymbol{Z_i^T W_i Z_i Q - I}| - \frac{1}{2}log|\boldsymbol{Q}| + \sum_{i=1}^{n}log(L_i(\boldsymbol{\beta},\tilde{\boldsymbol{b}})) - \sum_{i=1}^{n}\frac{1}{2}\tilde{\boldsymbol{b}}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{Q^{-1}}\tilde{\boldsymbol{b}}_{\boldsymbol{i}}$$

Thus, we can write the aforementioned equation in the following way:

$$l^{LA}(\boldsymbol{\phi},\boldsymbol{b}) = \sum_{i=1}^{n}log(L_i(\boldsymbol{\beta},\tilde{\boldsymbol{b}})) - \frac{1}{2}\sum_{i=1}^{n}\tilde{\boldsymbol{b}}_{\boldsymbol{i}}^{\boldsymbol{T}}\boldsymbol{Q^{-1}}\tilde{\boldsymbol{b}}_{\boldsymbol{i}} - \frac{1}{2}\sum_{i=1}^{n}logdet(\boldsymbol{Z_i^T}\tilde{\boldsymbol{W}}_{\boldsymbol{i}}\boldsymbol{Z_i Q + I}). \tag{4.22}$$

According to (Breslow & Clayton, 1993), (Ripatti & Palmgren, 2000) and (Hui et al., 2017), the last term in equation (4.22) has minimal influence in the eventual estimated

likelihood and therefore, we can ignore it with little to no information lost. Thus, the final approximated likelihood (PQL) is given as:

$$l^{PQL}(\boldsymbol{\phi}, \boldsymbol{b}) = \sum_{i=1}^{n} log(L_i(\boldsymbol{\beta}, \boldsymbol{b})) - \frac{1}{2}\sum_{i=1}^{n} \boldsymbol{b_i^T Q^{-1} b_i}. \tag{4.23}$$

From equation (4.23), it can be seen that the value of $\frac{1}{2}\boldsymbol{b_i^T Q^{-1} b_i}$ represents a penalty term that comes from this approximation. This is a generalized ridge penalty, where $\boldsymbol{b}$ is treated as fixed effect vector. This penalty reduces the influence of extremes values of $\boldsymbol{b}$.

PENALIZATION

We proposed to extend the penalty function of (Hui et al., 2017) for the censored data using the AFT models. That is, a combination of the adaptive LASSO (penalizes the fixed effects) and the adaptive group LASSO (penalizes the random effects) to perform a joint selection effect over the fixed and the frailty effects of the AFT models via the given approximated likelihood in equation (4.23).

For a given value of $\boldsymbol{Q}$, the penalized estimates of the fixed and the frailty effects of the AFT models from the above equation at (4.23) can be given as:

$$l^{pen}(\boldsymbol{\phi}, \boldsymbol{b}) = argmax(l^{PQL}(\boldsymbol{\phi}, \boldsymbol{b})) - \lambda \sum_{s=1}^{p_f} \boldsymbol{v_s}|\boldsymbol{\beta_s}| - \lambda \frac{\sum_{k=1}^{p_r} \boldsymbol{w_k} \boldsymbol{b_{.k}}^2}{\sqrt{N_i}} \tag{4.24}$$

where $\boldsymbol{\beta_s}$ represents the fixed effects and $\boldsymbol{b_{.t}} = (b_{i1}, ..., b_{ik})$ is the vector of coefficients of the tth frailty effect, $|.|$ denotes the L1-norm and $||.||$ denotes the L2-norm. Furthermore, $\boldsymbol{v_s}$ and $\boldsymbol{w_t}$ are the adaptive weights based on the unpenalized estimates. To get these unpenalized maximum estimates in AFT model with a left censored mechanism, we used the Monte Carlo Expectation Maximization (MCEM) algorithm as presented by (Vaida & Liu,

2009a). Once the unpenalized estimates of the fixed effects coefficients $\hat{\boldsymbol{\beta}}$ and the frailty effects variance-covariance matrix estimate, $\hat{\boldsymbol{Q}}$, on the full AFT model are calculated, the adaptive weights can be determined via: $\boldsymbol{v_s} = |\hat{\boldsymbol{\beta}}_{\boldsymbol{s}}|^{-\boldsymbol{s}}$ and $\boldsymbol{w_k} = \hat{\boldsymbol{Q}}_{\boldsymbol{kk}}^{-\boldsymbol{2s}}$. $\hat{\boldsymbol{Q}}_{\boldsymbol{kk}}$ is the kth diagonal element of $\hat{\boldsymbol{Q}}$. Both weights are calculated with a common power parameter $\boldsymbol{s} > 0$. Also, unlike (Hui et al., 2017), we have divided the group frailty effect penalty by $\sqrt{N_i}$, the square root of the varying size of the given cluster i. This is done for the standardization as the group LASSO depends on the varying size of the cluster. That is determined by the cluster's number and its size.

The penalty only has a single tuning parameter, λ, primarily because it saves a considerable amount of computational time and complexity (Garcia, Müller, Carroll, & Walzem, 2013). Moreover, given the concavity of both the PQL likelihood, $l^{PQL}(\boldsymbol{\phi}, \boldsymbol{b})$ and the lasso penalty functions, if there exists a maximizer to $l(\boldsymbol{\phi}, \boldsymbol{b})$ then it is also unique (Hui et al., 2017) (further details is in Lemma 2.1 of (Jiang, Jia, & Chen, 2001)). The ICs criterion as discussed in chapter 2 is used to select the value of the tuning parameter λ.

Estimation of the Variance-Covariance Matrix

Once, the penalized estimates of $\hat{\boldsymbol{\beta}}_\lambda$ and $\hat{\boldsymbol{b}}_\lambda$ have been determined from a given value of $\boldsymbol{Q}$, then these values can be used to update the value of the variance-covariance matrix. Following (Hui et al., 2017), the estimated values of the $\hat{\boldsymbol{Q}}_\lambda$ can be determined by substituting the value of $\hat{\boldsymbol{\beta}}_\lambda$ and $\hat{\boldsymbol{b}}_\lambda$ on equation (4.22), then the Laplace approximated

log-likelihood given as :

$$l^{LA}(\boldsymbol{\phi}, \boldsymbol{b}) = \sum_{i=1}^{n} log(L_i(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}, \hat{\boldsymbol{b}}_{\boldsymbol{\lambda i}}) - \frac{1}{2}\sum_{i=1}^{n} \hat{\boldsymbol{b}}_{\boldsymbol{\lambda i}}^{\boldsymbol{T}} \boldsymbol{Q}^{-\mathbf{1}} \hat{\boldsymbol{b}}_{\boldsymbol{\lambda i}} - \frac{1}{2}\sum_{i=1}^{n} logdet(\boldsymbol{Z}_{\boldsymbol{i}}^{\boldsymbol{T}} \tilde{\boldsymbol{W}}_{\boldsymbol{i}} \boldsymbol{Z}_{\boldsymbol{i}} \boldsymbol{Q} + \boldsymbol{I})$$

$$\Rightarrow l^{LA}(\boldsymbol{\phi}, \boldsymbol{b}) = -\frac{n}{2} logdet(\boldsymbol{Q}) - \frac{1}{2}\sum_{i=1}^{n} logdet[\boldsymbol{Z}_{\boldsymbol{i}}^{\boldsymbol{T}} \tilde{\boldsymbol{W}}_{\boldsymbol{i}} \boldsymbol{Z}_{\boldsymbol{i}} + \boldsymbol{Q}^{-\mathbf{1}}] + \sum_{i=1}^{n} log(L_i(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}, \hat{\boldsymbol{b}}_{\boldsymbol{\lambda i}})) - \frac{1}{2}\sum_{i=1}^{n} \hat{\boldsymbol{b}}_{\boldsymbol{\lambda i}}^{\boldsymbol{T}} \boldsymbol{Q}^{-\mathbf{1}} \hat{\boldsymbol{b}}_{\boldsymbol{\lambda i}}$$

Hence, for any $i = 1, ..., n, \hat{\boldsymbol{b}}_{\boldsymbol{\lambda i}}$ are the penalized frailty estimates for a given AFT model. Then by differentiating the likelihood function with respect to $\boldsymbol{Q}$ and setting it to zero we get the following:

$$\frac{\delta l^{LA}(\boldsymbol{\phi}, \boldsymbol{b})}{\delta \boldsymbol{Q}} = 0$$

Plugging the values for $l^{LA}(\boldsymbol{\phi}, \boldsymbol{b})$ from the aforementioned equation:

$$\Rightarrow \frac{\delta(-\frac{n}{2} logdet(\boldsymbol{Q}))}{\delta \boldsymbol{Q}} - \frac{\delta(\frac{1}{2}\sum_{i=1}^{n} logdet(\boldsymbol{Z}_{\boldsymbol{i}}^{\boldsymbol{T}} \tilde{\boldsymbol{W}}_{\boldsymbol{i}} \boldsymbol{Z}_{\boldsymbol{i}} + \boldsymbol{Q}^{-\mathbf{1}}))}{\delta \boldsymbol{Q}} + \frac{\delta(\sum_{i=1}^{n} log(L_i(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}, \hat{\boldsymbol{b}}_{\boldsymbol{\lambda i}})))}{\delta \boldsymbol{Q}} - \frac{\delta(\sum_{i=1}^{n} \frac{1}{2} \hat{\boldsymbol{b}}_{\boldsymbol{\lambda i}}^{\boldsymbol{T}} \boldsymbol{Q}^{-\mathbf{1}} \hat{\boldsymbol{b}}_{\boldsymbol{\lambda i}})}{\delta \boldsymbol{Q}} = 0$$

$$\Rightarrow -\frac{n}{2}\frac{\delta(logdet(\boldsymbol{Q}))}{\delta det(\boldsymbol{Q})}\frac{\delta det(\boldsymbol{Q})}{\delta \boldsymbol{Q}} - \frac{1}{2}\sum_{i=1}^{n} \frac{\delta(logdet(\boldsymbol{Z}_{\boldsymbol{i}}^{\boldsymbol{T}} \tilde{\boldsymbol{W}}_{\boldsymbol{i}} \boldsymbol{Z}_{\boldsymbol{i}} + \boldsymbol{Q}^{-\mathbf{1}}))}{\delta(det(\boldsymbol{Z}_{\boldsymbol{i}}^{\boldsymbol{T}} \tilde{\boldsymbol{W}}_{\boldsymbol{i}} \boldsymbol{Z}_{\boldsymbol{i}} + \boldsymbol{Q}^{-\mathbf{1}}))} \frac{\delta(det(\boldsymbol{Z}_{\boldsymbol{i}}^{\boldsymbol{T}} \tilde{\boldsymbol{W}}_{\boldsymbol{i}} \boldsymbol{Z}_{\boldsymbol{i}} + \boldsymbol{Q}^{-\mathbf{1}}))}{\delta((\boldsymbol{Z}_{\boldsymbol{i}}^{\boldsymbol{T}} \tilde{\boldsymbol{W}}_{\boldsymbol{i}} \boldsymbol{Z}_{\boldsymbol{i}} + \boldsymbol{Q}^{-\mathbf{1}}))} * \frac{\delta((\boldsymbol{Z}_{\boldsymbol{i}}^{\boldsymbol{T}} \tilde{\boldsymbol{W}}_{\boldsymbol{i}} \boldsymbol{Z}_{\boldsymbol{i}} + \boldsymbol{Q}^{-\mathbf{1}}))}{\delta \boldsymbol{Q}} - \frac{1}{2}\sum_{i=1}^{n} \frac{\delta(\hat{\boldsymbol{b}}_{\boldsymbol{\lambda i}}^{\boldsymbol{T}} \boldsymbol{Q}^{-\mathbf{1}} \hat{\boldsymbol{b}}_{\boldsymbol{\lambda i}})}{\delta \boldsymbol{Q}} = 0$$

By Jacobi's formula, we have $adjugate(\boldsymbol{X}) = \frac{\delta det \boldsymbol{X}}{\delta \boldsymbol{X}}$ (Magnus & Neudecker, 1999). So by using this formula on the above equation we get:

$$\Rightarrow -n\frac{1}{det(\boldsymbol{Q})} * adjugate(\boldsymbol{Q}) - \sum_{i=1}^{n} \frac{1}{det(\boldsymbol{Z}_{\boldsymbol{i}}^{\boldsymbol{T}} \tilde{\boldsymbol{W}}_{\boldsymbol{i}} \boldsymbol{Z}_{\boldsymbol{i}} + \boldsymbol{Q}^{-\mathbf{1}})} adjugate(\boldsymbol{Z}_{\boldsymbol{i}}^{\boldsymbol{T}} \tilde{\boldsymbol{W}}_{\boldsymbol{i}} \boldsymbol{Z}_{\boldsymbol{i}} + \boldsymbol{Q}^{-\mathbf{1}})(-1)(\boldsymbol{Q}^{-\mathbf{1}} * \boldsymbol{Q}^{-\mathbf{1}}) - \sum_{i=1}^{n}(-1)(\boldsymbol{Q}^{-\mathbf{1}} * \boldsymbol{Q}^{-\mathbf{1}})(\hat{\boldsymbol{b}}_{\boldsymbol{\lambda i}} \hat{\boldsymbol{b}}_{\boldsymbol{\lambda i}}^{\boldsymbol{T}}) = 0$$

We know $\boldsymbol{X^{-1}} = \frac{adjugate(\boldsymbol{X})}{det\boldsymbol{X}}$. By using this in above equation, we get the following:

$$\Rightarrow -n * (\boldsymbol{Q^{-1}}) + \sum_{i=1}^{n} (\boldsymbol{Z_i^T \tilde{W}_i Z_i + Q^{-1}})^{-1} (\boldsymbol{Q^{-1} * Q^{-1}}) + \sum_{i=1}^{n} (\boldsymbol{Q^{-1} * Q^{-1}})(\boldsymbol{\hat{b}_{\lambda i} \hat{b}_{\lambda i}^T}) = 0$$

$$\Rightarrow n * (\boldsymbol{Q^{-1}}) = (\boldsymbol{Q^{-1} * Q^{-1}})[\sum_{i=1}^{n} (\boldsymbol{Z_i^T \tilde{W}_i Z_i + Q^{-1}})^{-1} + \sum_{i=1}^{n} (\boldsymbol{\hat{b}_{\lambda i} \hat{b}_{\lambda i}^T})]$$

$$\Rightarrow \boldsymbol{Q} = \frac{1}{n} \sum_{i=1}^{n} [(\boldsymbol{Z_i^T \tilde{W}_i Z_i + Q^{-1}})^{-1} + \boldsymbol{\hat{b}_{\lambda i} \hat{b}_{\lambda i}^T}]$$

Then, it can be seen that we can use a iterative method to estimate the variance-covariance matrix. Let m be the index of the mth iteration, then the variance-covariance matrix at the m is given by:

$$\boldsymbol{Q_{\lambda}^{m}} = \frac{1}{n} \sum_{i=1}^{n} [(\boldsymbol{Z_i^T \tilde{W}_{\lambda i} Z_i + (Q_{\lambda}^{(m-1)})^{-1}})^{-1} + \boldsymbol{\hat{b}_{\lambda i} \hat{b}_{\lambda i}^T}]. \tag{4.25}$$

Therefore, using the approximated penalized maximum likelihood from equation (4.24) and the above iterative equation at (4.25), we can build our model selection procedure.

Overview of Individual Distributions

Equation 4.24 provide a generalized form of the penalized PQL likelihood equation. In this study, we propose to test our penalty function in two commonly used survival distributions: log-normal and Weibull distribution.

4.6.1 Lognormal distribution:

The lognormal distribution is a popular parametric function as it has extensive use in survival analysis. Such high applicability comes from its great property

where the logarithm of a lognormal distribution is a normal distribution with mean μ and variance σ^2 (Liu, 2012). In notation wise, the lognormal distribution is denoted by $t_{ij} \sim LN(\mu_{ij}, \sigma_i^2)$. So if $y_{ij} = logt_{ij}$ then the pdf of y_{ij} can be written as:

$$f(y_{ij}, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma_i} exp\left[- \frac{y_{ij} - \mu_{ij}}{2\sigma_i^2} \right], y \in (-\infty, \infty) \tag{4.26}$$

The cdf of y_{ij} can then be written as :

$$F(y_{ij}, \mu, \sigma^2) = \int_0^t \frac{1}{\sqrt{2\pi}\sigma_i} exp\left[- \frac{y_{ij} - \mu_{ij}}{2\sigma_i^2} \right] \partial y_{ij} \tag{4.27}$$

As $t_{ij} = exp(y_{ij})$, then the lognormal density function of t_{ij} can be given as:

$$f(y_{ij}, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}t_{ij}\sigma_i} exp\left[- \frac{logt_{ij} - \mu_{ij}}{2\sigma_i^2} \right], t_{ij} > 0 \tag{4.28}$$

Then according to Liu (Liu, 2012), the cdf of t_{ij} has the following standard form:

$$F(t_{ij}, \mu, \sigma^2) = 1 - \Phi\left[\frac{logt_{ij}}{\sigma_i} \right] \tag{4.29}$$

Thus, from equation (4.4), the survival function can be derived as:

$$S(t_{ij}, \mu, \sigma^2) = \Phi\left[\frac{logt_{ij}}{\sigma_i} \right] \tag{4.30}$$

Based on the above density equation, the penalized likelihood can be derived as the following:

$$\begin{aligned} l^{pen}(\boldsymbol{\phi}, \boldsymbol{b}) = \sum_{i=1}^{n} \sum_{j=1}^{N_i} \Bigg\{ & (1 - \delta_{ij}) log(\frac{1}{\sqrt{2\pi}\sigma_i} exp\left[- \frac{y_{ij} - \mu_{ij}}{2\sigma_i^2} \right]) \\ & + \delta_{ij} (log(\int_0^t \frac{1}{\sqrt{2\pi}\sigma_i} exp\left[- \frac{y_{ij} - \mu_{ij}}{2\sigma_i^2} \right] \partial y_{ij})) \Bigg\} \\ & - \lambda \sum_{s=1}^{p_f} \boldsymbol{v_s} |\boldsymbol{\beta_s}| - \lambda \frac{\sum_{k=1}^{p_r} \boldsymbol{w_k} \boldsymbol{b_{.k}}^2}{\sqrt{N_i}} \end{aligned} \tag{4.31}$$

Using the above equation, the penalized estimates for a log-normally distributed survival time can be evaluated.

4.6.2 Weibull distribution:

The Weibull probability distribution function is a continuous function with two parameters, that is the scale parameter λ_{ij} and the shape parameter p_{ij}. The logarithm of Weibull distribution is a two-parameter extreme value distribution or often called the Gumbel's (type-1 extreme value) distribution (White, 1969). In notation, if we have a Weibull distribution for t_{ij} then $y_{ij} = logt_{ij}$ will have a Gumbel distribution given by the following pdf:

$$f(y_{ij}, \lambda_{ij}, p_{ij}) = \frac{1}{p_{ij}} exp\left\{ \left[\frac{y_{ij} - \lambda_{ij}}{p_{ij}} \right] - exp\left[\frac{y_{ij} - \lambda_{ij}}{p_{ij}} \right] \right\}, y \in (-\infty, \infty) \tag{4.32}$$

where λ_{ij} and p_{ij} are the parameters for the Gumbel distrbution of y_{ij}. The cdf of y_{ij} can then be written as :

$$F(y_{ij}, \mu, b) = \int_0^t \frac{1}{p_{ij}} exp\left\{ \left[\frac{y_{ij} - \lambda_{ij}}{p_{ij}} \right] - exp\left[\frac{y_{ij} - \lambda_{ij}}{p_{ij}} \right] \right\} \partial y_{ij} \tag{4.33}$$

As $t_{ij} = exp(y_{ij})$ and if we let $\lambda_{ij} = exp(\mu_{ij})$ and $p_{ij} = b_{ij}^{-1}$, then the Weibull density function of t_{ij} can be given as:

$$f(t_{ij}, \lambda_{ij}, p_{ij}) = \lambda_{ij} p_{ij} (\lambda_{ij} t_{ij})^{p-1} exp\left[- (\lambda_{ij} t_{ij})^p \right] \tag{4.34}$$

Then according to Liu (Liu, 2012), the cdf of t_{ij} can have the following standard form:

$$F(t_{ij}, \lambda_{ij}, p_{ij}) = 1 - exp\left[- (\lambda_{ij} t_{ij})^p \right] \tag{4.35}$$

Thus, from equation (4.9), the survival function can be derived as the following:

$$S(y_{ij}, \lambda_{ij}, p_{ij}) = [1 + (t_{ij}/\lambda_{ij})^{p_{ij}}]^{-1} \tag{4.36}$$

The penalized likelihood equation for Weibull distributed survival time can be therefore be

derived as the following:

$$l^{pen}(\boldsymbol{\phi}, \boldsymbol{b}) = \sum_{i=1}^{n}\sum_{j=1}^{N_i}\left\{(1-\delta_{ij})log(\frac{1}{p_{ij}}exp\left\{\left[\frac{y_{ij}-\lambda_{ij}}{p_{ij}}\right] - exp\left[\frac{y_{ij}-\lambda_{ij}}{p_{ij}}\right]\right\}) \right.$$
$$\left. +\delta_{ij}(log(\int_0^t \frac{1}{p_{ij}}exp\left\{\left[\frac{y_{ij}-\lambda_{ij}}{p_{ij}}\right] - exp\left[\frac{y_{ij}-\lambda_{ij}}{p_{ij}}\right]\right\}\partial y_{ij}))\right\} \tag{4.37}$$
$$-\lambda\sum_{s=1}^{p_f}\boldsymbol{v_s}|\boldsymbol{\beta_s}| - \lambda\frac{\sum_{k=1}^{p_r}\boldsymbol{w_k}\boldsymbol{b_{.k}}^2}{\sqrt{N_i}}$$

Thus, the penalized estimates for Weibull distributed survival time can be evaluated.

CHAPTER 5

SIMULATION STUDY

Table 5.1-5.12 are the results for Lognormally distributed survival time. The larger sample of Simulation 1 results are presented from Table 5.1 to 5.6. Looking at the value of the sample n and the cluster size m, it can be seen that as the sample size increases, the penalty estimates are improving in terms of the values of cf, cr and c across all levels of censoring. The tables also show with a lower censoring level, the penalties seem to perform better and with a higher level of censoring, the penalty performance tends to go down. The proposed method ALASSOn is better in terms of its higher values in cf, cr and c across all sample sizes in this particular simulation. It is also consistent in all different levels of censoring as the proposed method outperforms the other penalties in terms of variable selection in all censoring cases of Simulation 1. Moreover, the mean total bias and the mean total variance is lower in the proposed method compared to other methods. Even in terms of the predicted negative log-likelihood value, the proposed method seems to have a relatively lower value compared to most of the other methods.

Table 5.7-5.12 show Simulation 2 results for Lognormally distributed survival time. Here, all the penalties do not perform as well as compared to Simulation 1, which is expected as Simulation 2 has a smaller sample size compared to Simulation 1. The biases and variance are larger here too. The censoring distribution in simulation 2 is exponential as opposed to uniform in Simulation 1. Regardless of the censoring mechanism, the penalties seem to have similar results as Simulation 1 though not as good. The proposed method in most of the cases are still outperforming or at least equivalently performing when compared to other penalties except on the smallest sample at n20 m5 with the largest censoring of 40 $\%$. In this condition, the ALASSOc measure has a higher correct selection score of fixed effects coefficients, cf. However, in the selection of random

effects, cr, the proposed method has a better score. Also, accounting for the average of the two scores, c, the proposed method is still better. However, the AlASSOc has a relatively lower Bias across the individual fixed effects coefficients in smaller sample size, but the differences of bias between ALASSOc and the proposed method is minimal. Total variance seems to be the lowest on the proposed method.

Table 5.13-5.24 are the results for Weibull distributed survival time. The larger sample Simulation 1 results for Weibull are presented from Table 5.13 to 5.18. Looking at the value of the sample n and the cluster size m, the performance seems just like in Lognormal distribution, the penalty estimates are improving in cf, cr and c as the sampling values go higher. It can be seen that the proposed method ALASSOn is better with its higher values in cf, cr and c across all sample sizes; however, other penalties are also not distinctly different. Notably, it seems ALASSOc, ALASSO is almost equivalent to the proposed at least in terms of the selection of fixed effects. However, the proposed method has a higher bar in terms of the random effects selection (cr) and so taking the average (c), the ALASSOn seems to be a better choice given its mostly high performance. The result is also consistent in all different levels of censoring as the proposed method outperforms the other penalties in variable selection in all censoring cases of Simulation 1. Moreover, the mean total bias and the total variance is lower in the proposed method compared to other methods. Even in terms of the predicted negative log-likelihood value the proposed method seems to have a relatively lower value compared to most of the other methods.

Table 5.19-5.24 shows Simulation 2 results for Weibull distributed survival time. Here, the cluster size is made larger for the smaller sample size as compared with Lognormal distribution with 200 iteration run. It is done so to reduce distortions among Weibull distributed survival time. The result shows a consistent pattern of the performance

with Simulation 1 in terms of bias. However, in terms of model selection, the proposed method (ALASSOn) has a better performance as it has higher scores in cf, cr and c as well as variance and predicted likelihood. All of the described results start from the following page.

Table 5.1: Results of the sample n=40 m=10 on Simulation 1 for Lognormal Distribution

Censoring	Penalty	cf	cr	c	totalbias	Variance	PL
	ALASSOn	1	0.680	0.840	0.533	2.057	-1.433
	ALASSOc	0.980	0.290	0.635	0.548	2.283	-1.464
40 %	ALASSO	0.990	0.420	0.705	0.538	2.212	-1.454
	LASSO	0.270	0.480	0.375	0.777	3.163	-1.614
	SCAD	0.050	0.460	0.255	1.033	2.025	-1.744
	ALASSOn	1	0.770	0.885	0.499	1.836	-1.442
	ALASSOc	1	0.440	0.72	0.516	2.004	-1.471
30 %	ALASSO	1	0.620	0.810	0.506	1.913	-1.455
	LASSO	0.270	0.560	0.415	0.708	2.405	-1.579
	SCAD	0.110	0.360	0.235	1.044	2.087	-1.503
	ALASSOn	0.990	0.930	0.960	0.323	1.190	-1.355
	ALASSOc	1	0.78	0.890	0.323	1.254	-1.376
10 %	ALASSO	0.990	0.910	0.950	0.324	1.202	-1.358
	LASSO	0.30	0.450	0.375	0.494	1.266	-1.349
	SCAD	0.770	0.480	0.625	0.506	1.090	-1.291

Table 5.2: Coefficient bias on sample n=40 m=10 for Simulation 1 following Lognormal Distribution

Censoring	Penalty	Bias1	Bias2	Bias3	Bias4	Bias5	Bias6	Bias7	Bias8	Bias9
	ALASSOn	-1.635	0.982	0.951	0.640	-0.405	0	0	0	0
	ALASSOc	-1.644	0.993	0.953	0.645	-0.400	0	0	0	0
40 %	ALASSO	-1.638	0.991	0.950	0.650	-0.414	0	0	0	0
	LASSO	-1.561	1.135	1.015	0.751	-0.517	0.008	-0.038	-0.009	-0.006
	SCAD	-1.264	1.231	0.969	1.074	-0.978	0	0	0	0
	ALASSOn	-1.109	0.686	0.734	0.473	-0.285	0	0	0	0
	ALASSOc	-1.112	0.696	0.728	0.474	-0.269	0	0	0	0
30 %	ALASSO	-1.111	0.691	0.730	0.473	-0.276	0	0	0	0
	LASSO	-1.034	0.833	0.769	0.565	-0.369	0.002	-0.029	-0.013	-0.016
	SCAD	-0.695	0.966	0.726	0.981	-0.934	0	0	0	0
	ALASSOn	-0.367	0.227	0.375	0.184	-0.096	0	0	0	0
	ALASSOc	-0.367	0.230	0.368	0.186	-0.095	0	0	0	0
10 %	ALASSO	-0.367	0.227	0.374	0.184	-0.096	0	0	0.001	0
	LASSO	-0.297	0.336	0.404	0.249	-0.143	-0.009	-0.025	0.005	-0.027
	SCAD	-0.196	0.322	0.396	0.306	-0.322	0	0	0	0

Table 5.3: Results of the sample n=50 m=10 on Simulation 1 for Lognormal Distribution

Censoring	Penalty	cf	cr	c	totalbias	Variance	PL
	ALASSOn	1	0.960	0.980	0.405	2.018	-1.422
	ALASSOc	1	0.760	0.880	0.414	2.104	-1.436
40 %	ALASSO	1	0.840	0.920	0.410	2.066	-1.430
	LASSO	0.090	0.410	0.250	0.638	2.481	-1.544
	SCAD	0.010	0.350	0.180	2.027	2.168	-1.464
	ALASSOn	1	0.950	0.975	0.488	1.837	-1.435
	ALASSOc	1	0.830	0.915	0.491	1.895	-1.446
30 %	ALASSO	1	0.920	0.960	0.490	1.850	-1.438
	LASSO	0.070	0.570	0.320	0.726	2.180	-1.546
	SCAD	0.150	0.510	0.330	0.862	1.912	-1.479
	ALASSOn	1	1	1	0.299	1.286	-1.394
	ALASSOc	1	0.930	0.965	0.300	1.321	-1.407
10 %	ALASSO	1	0.990	0.995	0.299	1.291	-1.396
	LASSO	0.190	0.870	0.530	0.537	1.473	-1.484
	SCAD	0.950	0.840	0.895	0.389	1.276	-1.392

Table 5.4: Coefficient bias on sample n=50 m=10 for Simulation 1 following Lognormal Distribution

Censoring	Penalty	Bias1	Bias2	Bias3	Bias4	Bias5	Bias6	Bias7	Bias8	Bias9
	ALASSOn	-1.643	0.943	0.883	0.594	-0.371	0	0	0	0
	ALASSOc	-1.643	0.947	0.882	0.596	-0.368	0	0	0	0
40 %	ALASSO	-1.642	0.946	0.883	0.596	-0.373	0	0	0	0
	LASSO	-1.549	1.069	0.936	0.669	-0.465	0.040	-0.051	-0.009	-0.001
	SCAD	-0.719	1.494	1.086	1.164	-0.998	0	0	0	0
	ALASSOn	-1.062	0.700	0.697	0.419	-0.265	0	0	0	0
	ALASSOc	-1.062	0.702	0.695	0.422	-0.265	0	0	0	0
30 %	ALASSO	-1.062	0.700	0.696	0.419	-0.264	0	0	0	0
	LASSO	-0.962	0.828	0.748	0.485	-0.347	0.026	-0.040	-0.019	0.006
	SCAD	-0.761	0.933	0.758	0.840	-0.909	0	0	0	0
	ALASSOn	-0.317	0.264	0.292	0.141	-0.080	0	0	0	0
	ALASSOc	-0.318	0.264	0.290	0.144	-0.081	0	0	0	0
10 %	ALASSO	-0.317	0.264	0.292	0.142	-0.080	0	0	0	0
	LASSO	-0.240	0.389	0.350	0.188	-0.139	0.018	-0.016	-0.012	-0.002
	SCAD	-0.254	0.289	0.297	0.200	-0.144	0	0	0	0

Table 5.5: Results of the sample n=60 m=20 on Simulation 1 for Lognormal Distribution

Censoring	Penalty	cf	cr	c	totalbias	Variance	PL
	ALASSOn	1	0.950	0.975	0.444	2.309	-1.455
	ALASSOC	1	0.770	0.885	0.452	2.386	-1.465
40 %	ALASSO	1	0.90	0.950	0.447	2.328	-1.457
	LASSO	0.030	0.240	0.135	0.700	2.544	-1.528
	SCAD	0.020	0.550	0.285	1.375	2.507	-1.504
	ALASSOn	1	1	1	0.596	2.050	-1.463
	ALASSOC	1	0.870	0.935	0.601	2.109	-1.473
30 %	ALASSO	1	0.980	0.990	0.597	2.060	-1.464
	LASSO	0.020	0.380	0.200	0.848	2.241	-1.530
	SCAD	0.090	0.190	0.140	1.467	1.968	-1.464
	ALASSOn	1	1	1	0.401	1.448	-1.430
	ALASSOC	1	1	1	0.402	1.460	-1.435
10 %	ALASSO	1	1	1	0.401	1.449	-1.431
	LASSO	0.100	0.820	0.460	0.637	1.561	-1.486
	SCAD	0.950	0.870	0.910	0.433	1.442	-1.430

Table 5.6: Coefficient bias on sample n=60 m=20 for Simulation 1 following Lognormal Distribution

Censoring	Penalty	Bias1	Bias2	Bias3	Bias4	Bias5	Bias6	Bias7	Bias8	Bias9
	ALASSOn	-1.627	1.005	0.873	0.660	-0.468	0	0	0	0
	ALASSOc	-1.627	1.007	0.877	0.662	-0.467	0	0	0	0
40 %	ALASSO	-1.627	1.006	0.874	0.660	-0.467	0	0	0	0
	LASSO	-1.521	1.115	0.927	0.715	-0.541	0.050	-0.040	0.011	-0.016
	SCAD	-1.087	1.367	1.000	1.082	-0.988	0	0	0	0
	ALASSOn	-1.017	0.782	0.690	0.496	-0.355	0	0	0	0
	ALASSOc	-1.016	0.784	0.690	0.498	-0.354	0	0	0	0
30 %	ALASSO	-1.017	0.782	0.691	0.496	-0.355	0	0	0	0
	LASSO	-0.916	0.886	0.746	0.541	-0.419	0.044	-0.029	0.015	-0.019
	SCAD	-0.519	1.115	0.778	1.055	-0.963	0	0	0	0
	ALASSOn	-0.297	0.341	0.309	0.202	-0.153	0	0	0	0
	ALASSOc	-0.300	0.343	0.310	0.202	-0.153	0	0	0	0
10 %	ALASSO	-0.298	0.341	0.310	0.202	-0.153	0	0	0	0
	LASSO	-0.219	0.457	0.360	0.235	-0.196	0.024	-0.014	0.005	-0.015
	SCAD	-0.263	0.356	0.302	0.248	-0.210	0	0	0	0

Table 5.7: Results of the sample n=20 m=5 on Simulation 2 for Lognormal Distribution

Censoring	Penalty	cf	cr	c	totalbias	Variance	PL
	ALASSOn	0.600	0.740	0.670	0.034	0.819	-1.231
	ALASSOc	0.690	0.610	0.650	-0.018	0.851	-1.243
40 %	ALASSO	0.630	0.700	0.665	-0.004	0.818	-1.230
	LASSO	0.290	0.100	0.195	0.151	1.150	-1.396
	SCAD	0.160	0.300	0.230	0.392	1.139	-1.382
	ALASSOn	0.740	0.860	0.800	0.057	0.766	-1.245
	ALASSOc	0.740	0.790	0.765	0.033	0.788	-1.256
30 %	ALASSO	0.750	0.820	0.785	0.042	0.776	-1.250
	LASSO	0.220	0.310	0.265	0.225	1.098	-1.412
	SCAD	0.410	0.460	0.435	0.329	0.936	-1.320
	ALASSOn	0.860	0.910	0.885	0.078	0.738	-1.254
	ALASSOc	0.820	0.850	0.835	0.079	0.784	-1.278
10 %	ALASSO	0.860	0.920	0.890	0.071	0.741	-1.256
	LASSO	0.240	0.430	0.335	0.229	1.043	-1.388
	SCAD	0.540	0.450	0.495	0.274	0.831	-1.272

Table 5.8: Coefficient bias on sample n=20 m=5 for Simulation 2 following Lognormal Distribution

Censoring	Penalty	Bias1	Bias2	Bias3	Bias4	Bias5	Bias6
	ALASSOn	-0.589	0.355	0.002	0.003	0.263	0.001
	ALASSOc	-0.590	0.336	0.001	-0.011	0.244	0.002
40 %	ALASSO	-0.589	0.334	0.001	-0.005	0.253	0.002
	LASSO	-0.588	0.448	-0.010	-0.016	0.336	-0.019
	SCAD	-0.584	0.549	0.000	-0.009	0.435	0.000
	ALASSOn	-0.351	0.253	-0.002	-0.010	0.169	-0.002
	ALASSOc	-0.353	0.246	-0.003	-0.008	0.153	-0.003
30 %	ALASSO	-0.352	0.248	-0.002	-0.007	0.158	-0.002
	LASSO	-0.346	0.355	-0.019	-0.022	0.267	-0.011
	SCAD	-0.338	0.361	-0.003	-0.004	0.311	0.002
	ALASSOn	-0.104	0.094	0.007	0.006	0.089	-0.014
	ALASSOc	-0.104	0.089	0.007	0.011	0.090	-0.013
10 %	ALASSO	-0.105	0.091	0.007	0.006	0.084	-0.014
	LASSO	-0.100	0.185	-0.004	-0.010	0.176	-0.018
	SCAD	-0.094	0.174	0.002	0.010	0.192	-0.011

Table 5.9: Results of the sample n=30 m=5 on Simulation 2 for Lognormal Distribution

Censoring	Penalty	cf	cr	c	totalbias	Variance	PL
	ALASSOn	0.740	0.730	0.735	0.085	0.813	-1.219
	ALASSOc	0.740	0.520	0.630	0.061	0.881	-1.246
40 %	ALASSO	0.760	0.610	0.685	0.081	0.852	-1.235
	LASSO	0.390	0.080	0.235	0.190	1.113	-1.370
	SCAD	0.120	0.300	0.210	0.514	1.133	-1.372
	ALASSOn	0.920	0.950	0.935	0.105	0.728	-1.225
	ALASSOc	0.920	0.870	0.895	0.105	0.759	-1.240
30 %	ALASSO	0.910	0.920	0.915	0.105	0.738	-1.230
	LASSO	0.350	0.330	0.340	0.279	1.104	-1.419
	SCAD	0.300	0.560	0.430	0.503	0.996	-1.352
	ALASSOn	0.910	0.990	0.950	0.117	0.720	-1.256
	ALASSO	0.910	0.980	0.945	0.118	0.732	-1.262
10 %	ALASSOc	0.920	0.940	0.930	0.117	0.756	-1.274
	LASSO	0.210	0.710	0.460	0.218	0.927	-1.363
	SCAD	0.500	0.630	0.565	0.396	0.840	-1.299

Table 5.10: Coefficient bias on sample n=30 m=5 for Simulation 2 following Lognormal Distribution

Censoring	Penalty	Bias1	Bias2	Bias3	Bias4	Bias5	Bias6
	ALASSOn	-0.583	0.399	-0.003	0.000	0.270	0.002
	ALASSOc	-0.585	0.379	-0.003	0.002	0.267	0.002
40 %	ALASSO	-0.584	0.396	-0.003	0.002	0.268	0.002
	LASSO	-0.570	0.463	-0.012	-0.022	0.336	-0.005
	SCAD	-0.544	0.622	-0.008	-0.004	0.448	0.000
	ALASSOn	-0.324	0.261	0.002	0	0.167	0
	ALASSOc	-0.324	0.259	0.001	0.001	0.168	0
30 %	ALASSO	-0.324	0.259	0.001	0.002	0.167	0
	LASSO	-0.305	0.342	-0.002	-0.015	0.273	-0.014
	SCAD	-0.287	0.418	0.001	-0.003	0.373	0
	ALASSOn	-0.084	0.125	0	-0.007	0.081	0.002
	ALASSO	-0.084	0.127	0	-0.007	0.080	0.002
10 %	ALASSOc	-0.085	0.128	0	-0.008	0.082	0
	LASSO	-0.074	0.183	-0.01	-0.019	0.139	-0.006
	SCAD	-0.058	0.224	0	-0.003	0.235	-0.002

Table 5.11: Results of the sample n=40 m=10 on Simulation 2 for Lognormal Distribution

Censoring	Penalty	cf	cr	c	totalbias	Variance	PL
	ALASSOn	1	1	1	-0.067	0.771	-1.183
	ALASSOc	0.990	0.960	0.975	-0.066	0.783	-1.191
40 %	ALASSO	1	0.990	0.995	-0.067	0.773	-1.185
	LASSO	0.150	0.720	0.435	-0.028	0.874	-1.253
	SCAD	0.530	0.720	0.625	0.049	0.794	-1.207
	ALASSOn	1	1	1	0.020	0.790	-1.244
	ALASSOc	0.990	0.990	0.990	0.018	0.797	-1.249
30 %	ALASSO	1	1	1	0.020	0.790	-1.244
	LASSO	0.120	0.910	0.515	0.040	0.840	-1.286
	SCAD	0.590	0.810	0.700	0.147	0.831	-1.273
	ALASSOn	1	1	1	0.075	0.825	-1.312
	ALASSOc	1	1	1	0.075	0.825	-1.313
10 %	ALASSO	1	0.990	0.995	0.074	0.835	-1.318
	LASSO	0.040	0.990	0.515	0.083	0.850	-1.335
	SCAD	0.790	0.800	0.795	0.140	0.841	-1.321

Table 5.12: Coefficient bias on sample n=40 m=10 for Simulation 2 following Lognormal Distribution

Censoring	Penalty	Bias1	Bias2	Bias3	Bias4	Bias5	Bias6
	ALASSOn	-0.630	0.355	0	0	0.208	0
	ALASSOc	-0.631	0.355	0	0	0.210	0
40 %	ALASSO	-0.630	0.355	0	0	0.209	0
	LASSO	-0.631	0.381	0.002	-0.008	0.235	-0.007
	SCAD	-0.624	0.351	-0.001	0	0.324	-0.001
	ALASSOn	-0.359	0.240	0	0	0.139	0
	ALASSOc	-0.360	0.240	0	0	0.139	0
30 %	ALASSO	-0.359	0.240	0	0	0.139	0
	LASSO	-0.362	0.254	0	0	0.150	-0.010
	SCAD	-0.353	0.227	0	0	0.270	0
	ALASSOn	-0.094	0.104	0	0	0.064	0
	ALASSOc	-0.094	0.105	0	0	0.064	0
10 %	ALASSO	-0.095	0.105	0	0	0.064	0
	LASSO	-0.097	0.120	0	0	0.100	-0.010
	SCAD	-0.092	0.096	0	0	0.100	0

Table 5.13: Results of the sample n=40 m=10 on Simulation 1 for Weibull Distribution

Censoring	Penalty	cf	cr	c	totalbias	Variance	PL
	ALASSOn	0.980	0.660	0.820	1.115	2.242	-1.504
	ALASSOc	0.970	0.440	0.705	1.122	2.360	-1.517
40 %	ALASSO	0.970	0.560	0.765	1.121	2.286	-1.508
	LASSO	0.090	0.170	0.130	1.251	2.097	-1.568
	SCAD	0.090	0.460	0.275	1.483	2.503	-1.558
	ALASSOn	0.990	0.680	0.835	0.991	2.157	-1.513
	ALASSOc	0.990	0.490	0.740	0.998	2.265	-1.530
30 %	ALASSO	0.990	0.590	0.790	0.997	2.205	-1.521
	LASSO	0.140	0.230	0.185	1.127	2.001	-1.546
	SCAD	0.100	0.340	0.220	1.512	2.333	-1.567
	ALASSOn	0.990	0.920	0.955	0.932	1.634	-1.553
	ALASSOc	1	0.83	0.915	0.931	1.688	-1.573
10 %	ALASSO	0.990	0.90	0.945	0.932	1.645	-1.556
	LASSO	0.190	0.460	0.325	1.068	1.587	-1.562
	SCAD	0.590	0.490	0.540	1.040	1.590	-1.549

Table 5.14: Coefficient bias on sample n=40 m=10 for Simulation 1 following Weibull Distribution

Censoring	Penalty	Bias1	Bias2	Bias3	Bias4	Bias5	Bias6	Bias7	Bias8	Bias9
	ALASSOn	-1.305	1.125	1.048	0.699	-0.452	0	0	0	0
	ALASSOc	-1.311	1.124	1.047	0.694	-0.432	0	0	0	0
40 %	ALASSO	-1.307	1.126	1.048	0.696	-0.441	0	0	0	0
	LASSO	-1.228	1.212	1.068	0.737	-0.472	0.016	-0.072	-0.001	-0.008
	SCAD	-1.048	1.343	1.055	1.097	-0.964	0	0	0	0
	ALASSOn	-0.783	0.756	0.816	0.519	-0.317	0	0	0	0
	ALASSOc	-0.788	0.763	0.812	0.518	-0.306	0	0	0	0
30 %	ALASSO	-0.782	0.761	0.813	0.519	-0.313	0	0	0	0
	LASSO	-0.705	0.842	0.843	0.572	-0.351	-0.01	-0.04	-0.01	-0.01
	SCAD	-0.389	1.032	0.806	1.014	-0.952	0	0	0	0
	ALASSOn	0.102	0.320	0.404	0.221	-0.114	0	-0.002	0	0
	ALASSOc	0.101	0.324	0.398	0.222	-0.113	0	0	0	0
10 %	ALASSO	0.103	0.321	0.403	0.220	-0.113	0	-0.002	0	0
	LASSO	0.167	0.421	0.430	0.293	-0.163	-0.011	-0.038	-0.011	-0.019
	SCAD	0.261	0.400	0.419	0.456	-0.496	0	0	0	0

Table 5.15: Results of the sample n=50 m=10 on Simulation 1 for Weibull Distribution

Censoring	Penalty	cf	cr	c	totalbias	Variance	PL
	ALASSOn	0.910	0.700	0.805	0.903	2.221	-1.499
	ALASSOc	0.950	0.470	0.710	0.913	2.337	-1.514
40 %	ALASSO	0.930	0.570	0.750	0.902	2.285	-1.506
	LASSO	0.080	0.16	0.120	1.076	2.105	-1.583
	SCAD	0.030	0.59	0.310	1.237	2.426	-1.554
	ALASSOn	1	0.830	0.915	1.031	2.025	-1.500
	ALASSOc	1	0.610	0.805	1.036	2.137	-1.521
30 %	ALASSO	1	0.680	0.840	1.036	2.090	-1.513
	LASSO	0.120	0.350	0.235	1.221	2.116	-1.601
	SCAD	0.020	0.480	0.250	1.479	2.243	-1.574
	ALASSOn	0.980	0.910	0.945	0.889	1.590	-1.534
	ALASSOc	0.980	0.820	0.900	0.889	1.641	-1.548
10 %	ALASSO	0.980	0.880	0.930	0.890	1.604	-1.538
	LASSO	0.160	0.640	0.400	1.067	1.667	-1.583
	SCAD	0.40	0.440	0.420	0.991	1.461	-1.503

Table 5.16: Coefficient bias on sample n=50 m=10 for Simulation 1 following Weibull Distribution

Censoring	Penalty	Bias1	Bias2	Bias3	Bias4	Bias5	Bias6	Bias7	Bias8	Bias9
	ALASSOn	-1.372	1.093	0.993	0.694	-0.506	0	0	0	0
	ALASSOc	-1.376	1.095	0.997	0.679	-0.481	0	0	0	0
40 %	ALASSO	-1.376	1.093	0.994	0.690	-0.499	0	0	0	0
	LASSO	-1.305	1.182	1.026	0.715	-0.523	0.023	-0.066	-0.003	0.026
	SCAD	-1.147	1.278	1.041	1.052	-0.986	0	0	0	0
	ALASSOn	-0.725	0.810	0.807	0.491	-0.352	0	0	0	0
	ALASSOc	-0.730	0.817	0.800	0.492	-0.342	0	0	0	0
30 %	ALASSO	-0.727	0.816	0.804	0.493	-0.350	0	0	0	0
	LASSO	-0.651	0.920	0.838	0.548	-0.409	0.008	-0.044	-0.011	0.022
	SCAD	-0.399	1.054	0.817	0.997	-0.991	0	0	0	0
	ALASSOn	0.158	0.301	0.354	0.220	-0.145	-0.001	-0.001	0.001	0.001
	ALASSOc	0.155	0.302	0.351	0.223	-0.144	-0.001	-0.001	0.001	0.001
10 %	ALASSO	0.158	0.302	0.354	0.221	-0.146	-0.001	-0.001	0.001	0.001
	LASSO	0.221	0.406	0.373	0.283	-0.203	0.002	-0.018	-0.013	0.017
	SCAD	0.323	0.387	0.377	0.572	-0.667	0	0	0	0

Table 5.17: Results of the sample n=60 m=20 on Simulation 1 for Weibull Distribution

Censoring	Penalty	cf	cr	c	totalbias	Variance	PL
	ALASSOn	0.980	0.920	0.950	0.874	2.558	-1.540
	ALASSOc	0.980	0.720	0.850	0.885	2.643	-1.553
40 %	ALASSO	1	0.890	0.945	0.880	2.569	-1.542
	LASSO	0.030	0.020	0.025	1.073	2.538	-1.607
	SCAD	0.030	0.530	0.280	1.733	2.640	-1.578
	ALASSOn	1	0.960	0.980	0.941	2.432	-1.564
	ALASSOc	1	0.820	0.910	0.947	2.501	-1.577
30 %	ALASSO	1	0.94	0.970	0.943	2.442	-1.566
	LASSO	0.020	0.08	0.050	1.176	2.545	-1.657
	SCAD	0.090	0.36	0.225	1.935	2.424	-1.59
	ALASSOn	1	1	1	0.820	1.968	-1.613
	ALASSOc	1	0.930	0.965	0.821	2.010	-1.626
10 %	LASSO	1	1	1	0.820	1.973	-1.614
	LASSO	0.030	0.620	0.325	1.041	2.109	-1.693
	SCAD	0.840	0.580	0.710	0.868	1.942	-1.612

Table 5.18: Coefficient bias on sample n=60 m=20 for Simulation 1 following Weibull Distribution

Censoring	Penalty	Bias1	Bias2	Bias3	Bias4	Bias5	Bias6	Bias7	Bias8	Bias9
	ALASSOn	-1.360	1.092	0.936	0.721	-0.515	0	0	0	0
	ALASSOc	-1.359	1.095	0.938	0.721	-0.509	0	-0.001	0.001	-0.001
40 %	ALASSO	-1.359	1.092	0.937	0.717	-0.507	0	0	0	0
	LASSO	-1.272	1.171	0.975	0.754	-0.557	0.059	-0.065	0.035	-0.026
	SCAD	-0.891	1.431	1.072	1.109	-0.989	0	0	0	0
	ALASSOn	-0.766	0.808	0.743	0.545	-0.388	0	0	0	0
	ALASSOc	-0.766	0.809	0.743	0.547	-0.387	0	0	0	0
30 %	ALASSO	-0.765	0.808	0.743	0.545	-0.387	0	0	0	0
	LASSO	-0.663	0.904	0.790	0.588	-0.446	0.054	-0.047	0.021	-0.023
	SCAD	-0.199	1.161	0.886	1.030	-0.943	0	0	0	0
	ALASSOn	0.070	0.341	0.347	0.236	-0.174	0	0	0	0
	ALASSOc	0.069	0.342	0.345	0.238	-0.173	0	0	0	0
10 %	LASSO	0.070	0.342	0.347	0.236	-0.174	0	0	0	0
	LASSO	0.155	0.445	0.393	0.272	-0.222	0.032	-0.026	0.017	-0.024
	SCAD	0.133	0.360	0.363	0.323	-0.311	0	0	0	0

Table 5.19: Results of the sample n=20 m=10 on Simulation 2 for Weibull Distribution

Censoring	Penalty	cf	cr	c	totalbias	Variance	PL
	ALASSOn	0.935	0.985	0.960	0.510	0.850	-1.220
	ALASSOc	0.930	0.930	0.930	0.520	0.860	-1.230
40 %	ALASSO	0.930	0.980	0.955	0.510	0.850	-1.220
	LASSO	0.130	0.460	0.295	0.580	0.910	-1.290
	SCAD	0.685	0.490	0.587	0.600	0.850	-1.240
	ALASSOn	0.925	0.990	0.957	0.600	0.910	-1.340
	ALASSOc	0.915	0.930	0.922	0.610	0.930	-1.350
30 %	ALASSO	0.925	0.980	0.952	0.600	0.920	-1.340
	LASSO	0.105	0.495	0.300	0.660	0.960	-1.390
	SCAD	0.745	0.455	0.600	0.660	0.890	-1.330
	ALASSOn	0.940	0.995	0.967	0.620	1.130	-1.530
	ALASSOc	0.930	0.920	0.925	0.620	1.170	-1.550
10 %	ALASSO	0.935	0.980	0.957	0.620	1.140	-1.540
	LASSO	0.070	0.680	0.375	0.680	1.200	-1.580
	SCAD	0.740	0.585	0.662	0.700	1.140	-1.540

Table 5.20: Coefficient bias on sample n=20 m=10 for Simulation 2 following Weibull Distribution

Censoring	Penalty	Bias1	Bias2	Bias3	Bias4	Bias5	Bias6
	ALASSOn	-0.010	0.330	0	0	0.200	0
	ALASSOc	-0.010	0.330	0	0	0.210	0
40 %	ALASSO	-0.010	0.330	0	0	0.200	0
	LASSO	0	0.370	0	-0.020	0.250	-0.010
	SCAD	0	0.340	0	0	0.260	0
	ALASSOn	0.200	0.250	0	0	0.160	0
	ALASSOc	0.200	0.250	0	0	0.160	0
30 %	ALASSO	0.200	0.250	0	0	0.160	0
	LASSO	0.210	0.280	0	-0.020	0.200	-0.010
	SCAD	0.210	0.250	0	0	0.200	0
	ALASSOn	0.430	0.110	0	-0.010	0.090	0
	ALASSOc	0.430	0.110	0	0	0.090	0
10 %	ALASSO	0.420	0.110	0	-0.010	0.090	0
	LASSO	0.440	0.150	0	-0.020	0.120	0
	SCAD	0.440	0.120	0	0	0.140	0

Table 5.21: Results of the sample n=30 m=10 on Simulation 2 for Weibull Distribution

Censoring	Penalty	cf	cr	c	totalbias	Variance	PL
	ALASSOn	0.970	1	0.985	0.514	0.873	-1.217
	ALASSOc	0.965	0.985	0.975	0.515	0.879	-1.223
40 %	ALASSO	0.970	0.990	0.980	0.515	0.875	-1.219
	LASSO	0.105	0.590	0.348	0.555	0.928	-1.284
	SCAD	0.720	0.520	0.620	0.580	0.869	-1.226
	ALASSOn	0.985	1	0.990	0.580	0.919	-1.335
	ALASSOc	0.975	0.995	0.990	0.590	0.927	-1.343
30 %	ALASSO	0.985	1	0.990	0.580	0.920	-1.336
	LASSO	0.105	0.680	0.390	0.620	0.970	-1.392
	SCAD	0.745	0.620	0.680	0.650	0.920	-1.346
	ALASSOn	0.985	1	0.990	0.590	1.120	-1.523
	ALASSOc	0.975	0.970	0.970	0.600	1.140	-1.534
10 %	ALASSO	0.985	1	0.990	0.590	1.122	-1.524
	LASSO	0.045	0.850	0.450	0.630	1.170	-1.563
	SCAD	0.840	0.690	0.770	0.640	1.117	-1.528

Table 5.22: Coefficient bias on sample n=30 m=10 for Simulation 2 following Weibull Distribution

Censoring	Penalty	Bias1	Bias2	Bias3	Bias4	Bias5	Bias6
	ALASSOn	-0.009	0.323	0	0	0.200	0
	ALASSOc	-0.008	0.323	0	0	0.200	-0.001
40 %	ALASSO	-0.009	0.323	0	0	0.201	0
	LASSO	0	0.346	-0.004	-0.008	0.223	-0.002
	SCAD	0	0.320	0.001	0.003	0.257	-0.001
	ALASSOn	0.190	0.240	0	0	0.150	0
	ALASSOc	0.190	0.240	0	0	0.160	0
30 %	ALASSO	0.190	0.240	0	0	0.150	0
	LASSO	0.200	0.260	0	-0.010	0.180	-0.010
	SCAD	0.200	0.240	0	0	0.210	0
	ALASSOn	0.430	0.100	0	0	0.070	0
	ALASSOc	0.430	0.100	0	0	0.070	0
10 %	ALASSO	0.430	0.100	0	0	0.070	0
	LASSO	0.440	0.130	-0.010	-0.010	0.090	-0.004
	SCAD	0.440	0.100	0	0	0.110	0

Table 5.23: Results of the sample n=40 m=10 on Simulation 2 for Weibull Distribution

Censoring	Penalty	cf	cr	c	totalbias	Variance	PL
	ALASSOn	1	1	1	0.511	0.871	-1.222
	ALASSOc	1	0.980	0.990	0.512	0.879	-1.230
40 %	ALASSO	1	0.995	0.998	0.511	0.872	-1.224
	LASSO	0.070	0.770	0.420	0.534	0.928	-1.289
	SCAD	0.620	0.580	0.600	0.592	0.876	-1.242
	ALASSOn	1	1	1	0.565	0.918	-1.325
	ALASSOc	1	0.985	0.993	0.567	0.926	-1.332
30 %	ALASSO	1	1	1	0.565	0.918	-1.325
	LASSO	0.040	0.825	0.433	0.591	0.958	-1.374
	SCAD	0.700	0.610	0.655	0.652	0.927	-1.343
	ALASSOn	0.995	1	0.998	0.602	1.119	-1.517
	ALASSOc	0.990	0.995	0.993	0.603	1.129	-1.523
10 %	ALASSO	0.995	1	0.998	0.602	1.120	-1.518
	LASSO	0.030	0.950	0.490	0.624	1.156	-1.548
	SCAD	0.750	0.735	0.743	0.687	1.136	-1.531

Table 5.24: Coefficient bias on sample n=40 m=10 for Simulation 2 following Weibull Distribution

Censoring	Penalty	Bias1	Bias2	Bias3	Bias4	Bias5	Bias6
	ALASSOn	-0.020	0.327	0	0	0.204	0
	ALASSOc	-0.019	0.328	0	0	0.203	0
40 %	ALASSO	-0.020	0.328	0	0	0.203	0
	LASSO	-0.015	0.342	0.006	-0.005	0.219	-0.013
	SCAD	-0.012	0.321	0.002	-0.001	0.285	-0.002
	ALASSOn	0.179	0.242	0	0	0.145	0
	ALASSOc	0.180	0.241	0	0	0.145	0
30 %	ALASSO	0.179	0.242	0	0	0.145	0
	LASSO	0.186	0.254	0	0	0.155	-0.003
	SCAD	0.189	0.233	0	0	0.229	0
	ALASSOn	0.430	0.106	0.001	0	0.066	0
	ALASSOc	0.431	0.106	0.001	0	0.066	0
10 %	ALASSO	0.430	0.106	0.001	0	0.066	0
	LASSO	0.439	0.123	-0.003	-0.003	0.079	-0.010
	SCAD	0.440	0.096	0.003	-0.001	0.149	0

CHAPTER 6

APPLICATION TO UNSTRUCTURED TREATMENT INTERRUPTION DATA

INTRODUCTION

In this chapter, the proposed method is tested in a real secondary data of the clinical outcomes from Unstructured Treatment Interruption (UTI) in children and adolescents that have prenatally acquired HIV infection as done by (Saitoh et al., 2008). The study is primarily related to the adverse effects of lack of adherence to the antiretroviral therapy in children with HIV infection. The HIV infected adolescents present a significantly difficult medication challenge in achieving full adherence given their unique developmental, psycho-social and lifestyle issues (Osterberg & Blaschke, 2005). Such problems may often lead to a stage where the given population of adolescents has a sub-optimal adherence that can lead to antiretroviral resistance and thereby diminishing treatment options. Therefore, an intervention called the treatment infection is introduced where, for a specific time, antiretroviral therapy is discontinued. This period of UTI has been studied in numerous ways, but there exists a lack of information for the clinical and immunological outcomes of UTI in adolescents and pediatric populations (Saitoh et al., 2008; Pai, Tulsky, Lawrence, Colford, & Reingold, 2005; Gibb et al., 2004; Monpoux et al., 2004). Thus, this study aims to study the viral RNA load among these patient groups with the time given at the time of their UTI period (Vaida & Liu, 2009a).

THE UTI DATASET

The given UTI dataset (Saitoh et al., 2008) is a retrospective study at four academic centers in the United States among prenatally acquired HIV infected youths. Initially, 405 participants went though the antiretroviral treatment, and after 6 months of ther-

apy, 71 subjects had a lack of adherence to the therapy and showed signs of treatment resistance. Thus, the antiretroviral therapy among these 71 participants was discontinued for some time (UTI period) and had their viral load observed in a set of eight different time points: 0, 1, 3, 6, 9, 12, 18 and 24 months. At each time point, if a given participant had a very high viral load of HIV-RNA, then they would be taken off the treatment interruption and thereby allowed to continue the therapy, leading to a drop-off from the study. Constituting these drop-offs, there was the following changes in the number of participants at each time point: 71 patients at the start of the study that is time month 0; 62 patients in time month 1; 58 patients in time month 3; 57 patients in time month 6; 43 patients in time month 9; 34 patients in time month 12; 24 patients in time month 18 and lastly 12 patients in time month 24. These values are indicated by variable *Fup*-follow-up months in the datasets. Therefore, by adding these numbers, it can be stated that there is a total of 362 observations for all patients at all time points. Out of 362 observations, 26 (7%) of them were below the detection limits (50 copies/mL), and therefore are censored (left-censoring). The censoring indicator is represented by *RNAcens* which is one for censored values (viral copies below 50 copies/mL observation) and is zero for uncensored values (viral copies at least 50 copies/mL).

The independent variable x_{ij} is indicator of the dependent variable y_{ij} when it is measured at time t_j for patient i. For instance, if y_{i1} is measured at time t_1 then $x_{i1} = 1$ while for all other time points: t_2; t_3; t_4; t_5; t_6; t_7; t_8, the respective x_{i2}; x_{i3}; x_{i4}; x_{i5}; x_{i6}; x_{i7}; x_{i8} are zero. Therefore, for eight different t_j, there are eight different independent indicator variable give by $\boldsymbol{X_j} = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)$. Other variables in the dataset include: id and *Days.after.TI*. A sequence of one is developed to account for the random intercept (z_{i1}) and *Days.after.TI* is used as a random slope variable (z_{i2}) for the model. A part of the data is given in Table 6.1, which gives the values for patient id 4 and 37. Patient

Table 6.1: Sample of the UTI data

id	Y	X1	X2	X3	X4	X5	X6	X7	X8	Z1	Z2	Censor
4	4.971	1	0	0	0	0	0	0	0	1	-94	0
4	4.919	0	1	0	0	0	0	0	0	1	40	0
4	4.823	0	0	1	0	0	0	0	0	1	117	0
4	5.034	0	0	0	1	0	0	0	0	1	257	0
4	4.693	0	0	0	0	1	0	0	0	1	329	0
4	4.741	0	0	0	0	0	1	0	0	1	392	0
4	5.258	0	0	0	0	0	0	1	0	1	552	0
4	5.095	0	0	0	0	0	0	0	1	1	867	0
37	1.698	1	0	0	0	0	0	0	0	1	0	1
37	4.424	0	1	0	0	0	0	0	0	1	65	0
37	4.424	0	0	1	0	0	0	0	0	1	65	0
37	3.992	0	0	0	1	0	0	0	0	1	155	0
37	4.574	0	0	0	0	1	0	0	0	1	252	0
37	4.697	0	0	0	0	0	1	0	0	1	321	0

id 4 has all eight observation for each time so there are 8 values inside the cluster whereas patient id 37 has 6 observations with a drop-off so thus 6 values in the cluster. It also has one censored observation.

Distribution of Dependent Variable

The dependent variable for this analysis is the log of the viral load RNA as we are interested in the fluctuation of viral RNA among UTI patients with time. The distribution of the log of viral load RNA was therefore examined using QQ plots, probability plots,

plots of cumulative distribution functions, and probability density plots. The following distributions were examined: log-normal, log-logistic, Weibull and gamma distribution. The plots were constructed using the *fitdistrplus* (Delignette-Muller, Dutang, Pouillot, Denis, & Siberchicot, 2019) package in R 3.6.1. These plots from the distributions mentioned above are provided in figures 2 to 5.

Among the distributions considered, Weibull offered a better fit to the data compared to other distrbutions tested in this study. It is further confirmed by Table 6.2, which gives the goodness of fit criteria of all the distributions with their respective AIC and BIC. Weibull has the lowest values among these criteria for AIC and BIC. Hence, based on these metrics, it is assumed that the dependent variable follows a Weibull distribution.

Table 6.2: Goodness of fit criteria for the considered distributions

Distribution	AIC	BIC
Weibull	1006.970	1014.754
Log-normal	1253.530	1261.313
Gamma	1158.024	1165.807
Log-logistic	1024.572	1032.356

DATA MODELING

The given UTI datasets was fitted by the AFT frailty model that can be represented by the following:

$$y_{ij} = \beta * X_{ij}^{T} + b_i * Z_{ij}^{T} + \epsilon_{ij}, \tag{6.1}$$

Here, $\boldsymbol{y_{ij}}$ is the log(HIV-RNA), that is the log of the viral load for a given patient i at a time month t_j; the fixed effect coefficients are represented by $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8)$ whereas $\boldsymbol{b_i} = (b_{i0}, b_{i1})$ is the random coefficients and $\boldsymbol{\epsilon_{ij}}$ is the error among each level of modeling.

The model from each of the penalty values and an unpenalized model was created for a good comparison. Then, the estimated value of the fixed effects coefficients along with the variance-covariance matrix for each model is evaluated. Table 6.3 gives these outputs for the given UTI dataset. Also, a sequence of 150 tuning parameter values was generated from 0 to 1 to determine the appropriate penalty value via the IC crieterion.

RESULTS

The results presented here are those of an AFT frailty model. There are six models shown in Table 6.3. The no-penalty model has the estimates for all the variables, including the two random effects. In the variance-covariance matrix, the random slope is very close to zero in the no-penalty model. Therefore, due to this value, the random slope gets quickly penalized as it shrinks to zero on every penalty model in Table 6.3. Meanwhile, the values that are not closer to zero in the model aren't shrunk, as evident in the fixed effects and the random intercept.

Overall, it can be seen that there is a positive influence of each of the fixed effect covariate on the dependent variable, that is with a unit change in any one of the covariates, there is at least 3.53 unit change in the Log(HIV-RNA). The 3.53 value is the lowest value on the model located in β_1 of the LASSO penalty, and in all other cases, these points are higher. So, the viral load increases dramatically with each passing time. But,

such an outcome is expected for the study given the patients have stopped the anti-retroviral therapy during the UTI time, so naturally, the virus multiplies under no treatment as provided by (Saitoh et al., 2008). The random intercept also has a positive influence; it shows that the viral load is different among the patients at the start of the study.

Looking at the estimates of each of the model in Table 6.3, ALASSOn, ALASSOc, and ALASSO have relatively similar or equivalent estimates. It is because these penalties are different modified variations of the adaptive LASSO penalty. SCAD and LASSO have a little more distinct estimates, but not by far as the coefficient values for the fixed effects, and the random intercept are relatively higher than zero and all the penalty models do not shrink these estimates as easily as compared to the random slope.

Table 6.3: Estimated fixed effects and the variance covariance matrix for accelerated failure time random-effect model for the considered distributions

Penalty	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	VarCov
ALASSOn	3.673	4.195	4.270	4.384	4.577	4.582	4.690	4.792	$\begin{bmatrix} 0.670 & 0 \\ 0 & 0 \end{bmatrix}$
ALASSOc	3.673	4.195	4.270	4.384	4.576	4.580	4.688	4.790	$\begin{bmatrix} 0.665 & 0 \\ 0 & 0 \end{bmatrix}$
ALASSO	3.673	4.195	4.270	4.384	4.577	4.581	4.69	4.791	$\begin{bmatrix} 0.669 & 0 \\ 0 & 0 \end{bmatrix}$
LASSO	3.533	4.026	4.098	4.212	4.384	4.368	4.457	4.514	$\begin{bmatrix} 0.621 & 0 \\ 0 & 0 \end{bmatrix}$
SCAD	3.678	4.201	4.276	4.389	4.582	4.586	4.695	4.796	$\begin{bmatrix} 0.669 & 0 \\ 0 & 0 \end{bmatrix}$
No-penalty	3.613	4.181	4.254	4.372	4.559	4.534	4.616	4.738	$\begin{bmatrix} 0.9508 & -0.00063 \\ -0.00063 & 0.000017 \end{bmatrix}$

Figure 6.1: PDF, QQ plot, PP plot and the CDF of empirical data compared to a Log-normal distribution

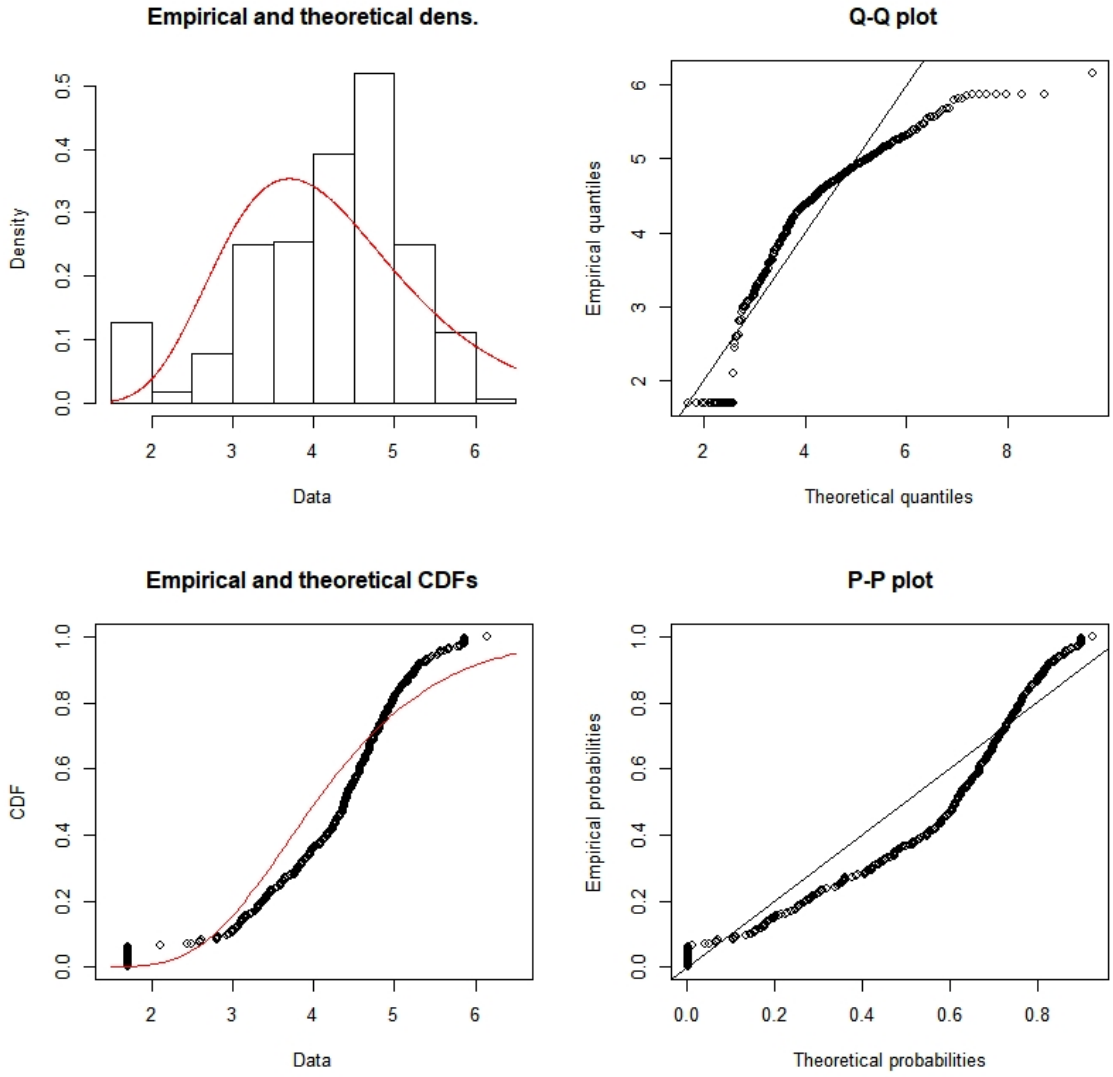


Figure 6.2: PDF, QQ plot, PP plot and the CDF of empirical data compared to a Log-logistic distribution

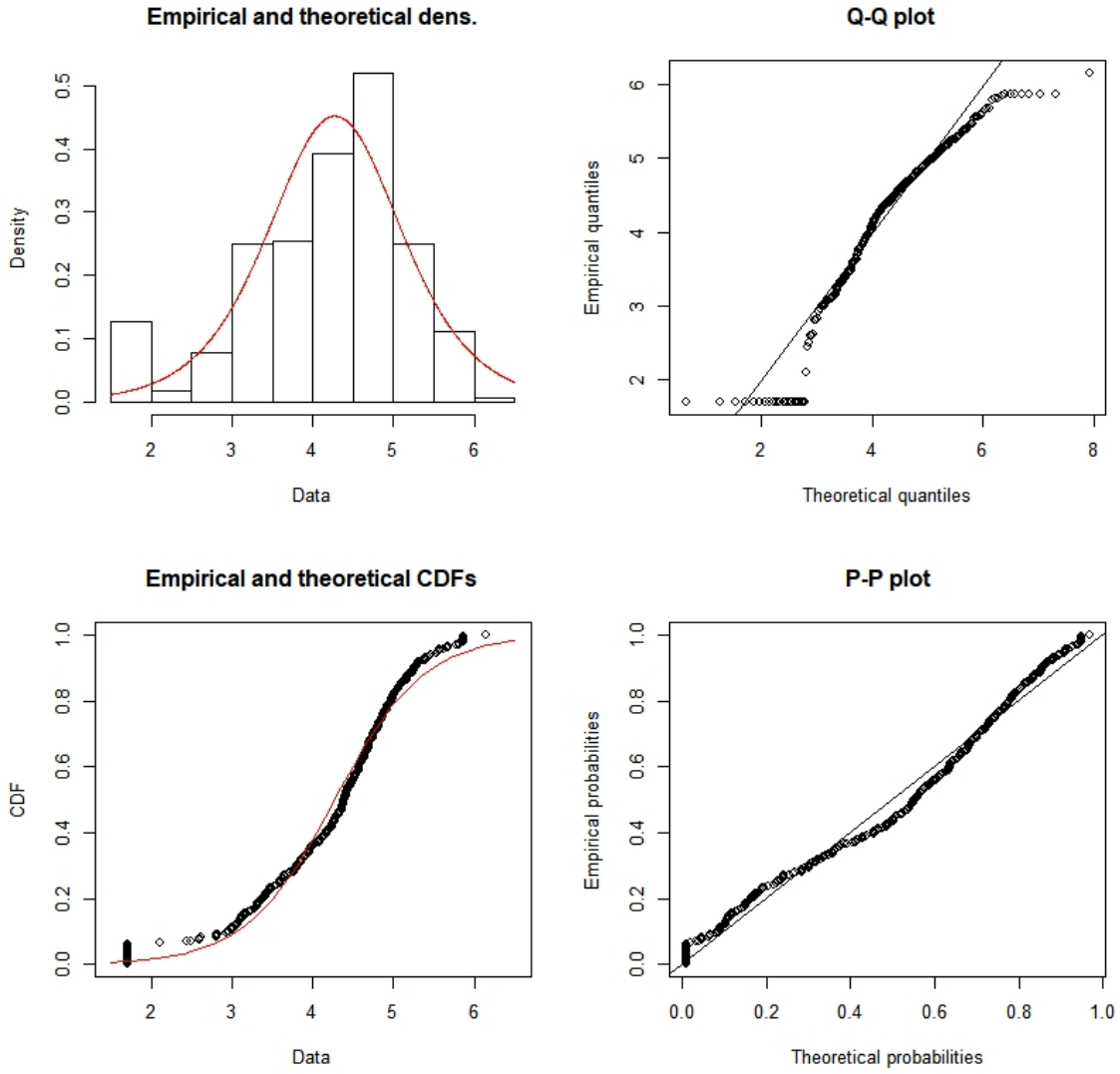


Figure 6.3: PDF, QQ plot, PP plot and the CDF of empirical data compared to a Weibull distribution

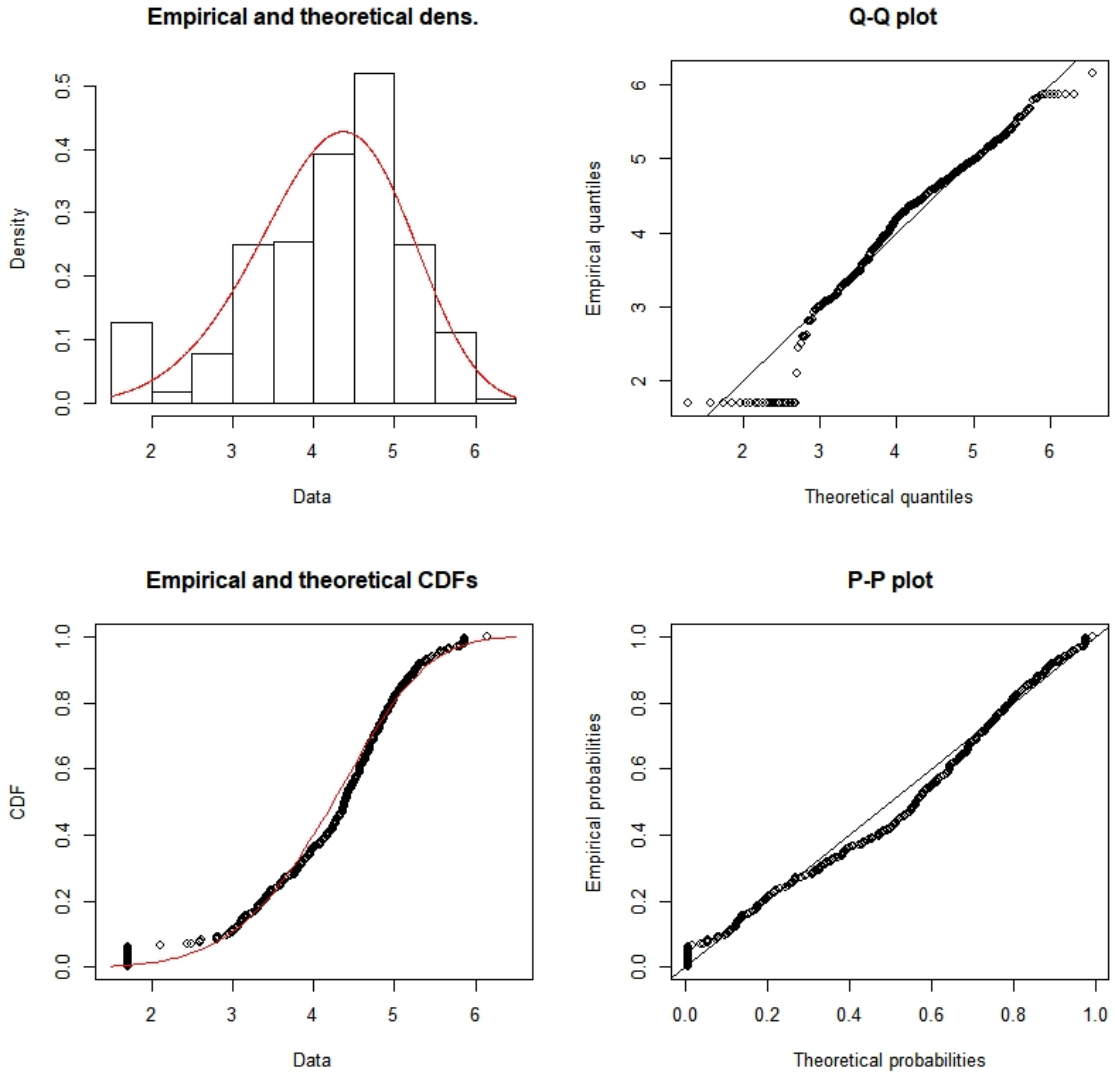
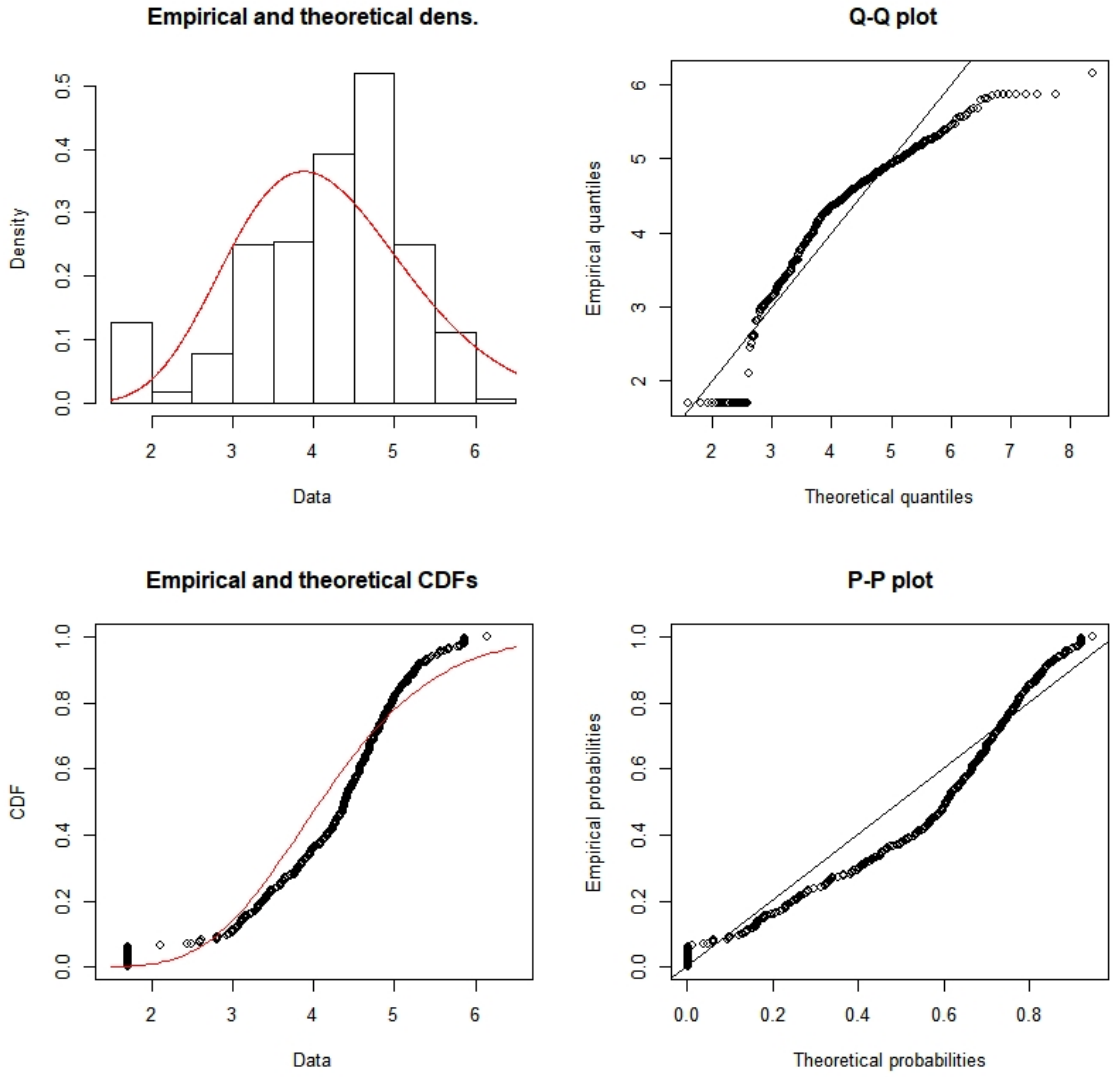


Figure 6.4: PDF, QQ plot, PP plot and the CDF of empirical data compared to a Gamma distribution



CHAPTER 7

CONCLUSION

Penalized variable selection and model building is one of the most important tools in high dimensional data sets, big data analytics, machine learning, and artificial intelligence (George, Osinga, Lavie, & Scott, 2016). Therefore, this area has a vast amount of ever-growing challenges. From traditional methods of selecting fixed effects, there has been a shift towards the selection of random effects (Hui et al., 2017). However, a joint selection of fixed and random effects is a very challenging problem due to the lack of closed-form solution of the marginal likelihood, as demonstrated in Chapter 4. It is even more complicated under censoring. This dissertation thus provides a solution for such kind of problem by using Breslow and Clayton's PQL approach (Breslow & Clayton, 1993) and regularizing it with a proposed adaptive lasso penalty. This regularized PQL has demonstrated an ability to reduce computational complexities and provided an efficient algorithm for penalized estimation and model building.

With the simulation values, this dissertation has displayed consistency on model building. It has proposed a penalty parameter that outperforms or at least equivalently performs in selecting variables. These results were valid regardless of the type of distribution (Weibull or Log-normal) of the dependent variable or the censoring distribution (uniform or exponential). Even in the real data analysis, it was evident the performance of the proposed method was similar to other established methods. Therefore the proposed method can be an excellent alternative to conduct joint model selection in survival analysis using the AFT model and especially when there is the presence of a frailty factor.

However, there were certain limitations for the proposed method. First, it is not able to have the same kind of performance in the smaller sample size while comparing

to larger samples. However, all other penalties had the same issue as well and the proposed method was still performing on the same level though the difference was not visible like the larger samples. Therefore, we need further simulation in smaller sets of samples and clusters to confirm robustness. Second, the study is in the left-censored mechanism. So there is a need to focus on other censoring mechanisms like right-censoring, interval-censoring, or informative-censoring to demonstrate full consistency and validity of the proposed method in survival analysis. Third, in smaller sample size performance of the regularized PQL altered in Lognormal and Weibull as we needed to change the sample in Weibull. The study cannot confirm if the same sample size would produce the same kind of results on other survival distributions like log-logistic, gamma, or exponential.

This dissertation gives a lot of steps for further bio-statistical research in public health, medical area, and clinical trials. First, the thesis had primarily focused on a lower-dimensional setting where the sample size is larger than the number of covariates in the model. One could modify this approach to include high dimensional variable selection where the covariate is larger than the sample size. Second, this study is on parametric survival analysis. However, survival distributions are very biased, and the parametric approach may not always be accurate to account for all types of survival data. Therefore, a viable strategy would be to have a future research on the proposed penalty that focuses in its performance on non-parametric survival data. Third, survival data include variables that could be time-varying, especially if it is pertaining to medical research and clinical trials. Spline models are a good approach to solve such challenges. (Groll et al., 2017) have demonstrated this using the PQL likelihood in Cox's model and adding three different penalties to penalize the fixed effect, the random effects and the time-varying spline effects. Our study already demonstrates the fixed and random effects penalty. Future research that adds a spline model could certainly be a new way to address time-varying effects in regularized

AFT frailty models.

REFERENCES

Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

Ahn, M., Zhang, H. H., & Lu, W. (2012). Moment-based method for random effects selection in linear mixed models. *Statistica Sinica*, *22*(4), 1539.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle,[w:] proceedings of the 2nd international symposium on information, bn petrow, f. *Czaki, Akademiai Kiado, Budapest*.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

Amorim, L. D., Cai, J., Zeng, D., & Barreto, M. L. (2008). Regression splines in the time-dependent coefficient rates model for recurrent event data. *Statistics in medicine*, *27*(28), 5890–5906.

Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The annals of statistics*, 1100–1120.

Androulakis, E., Koukouvinos, C., & Vonta, F. (2012). Estimation and variable selection via frailty models with penalized likelihood. *Statistics in Medicine*, *31*(20), 2223–2239.

Becker, N., Toedt, G., Lichter, P., & Benner, A. (2011). Elastic scad as a novel penalization method for svm classification tasks in high-dimensional data. *BMC bioinformatics*, *12*(1), 138.

Bernau, C., Augustin, T., & Boulesteix, A.-L. (2013). Correcting the optimal resampling-based error rate by estimating the error rate of wrapper algorithms. *Biometrics*, *69*(3), 693–702.

Bondell, H. D., Krishna, A., & Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, *66*(4), 1069–1077.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 791–799.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, *37*(4), 373–384.

Breiman, L., et al. (1996). Heuristics of instability and stabilization in model selection. *The annals of statistics*, *24*(6), 2350–2383.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, *88*(421), 9–25.

Cai, J., & Prentice, R. L. (1995). Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika*, *82*(1), 151–164.

Cai, T., Huang, J., & Tian, L. (2009). Regularized estimation for the accelerated failure time model. *Biometrics*, *65*(2), 394–404.

Cai, Z., & Sun, Y. (2003). Local linear estimation for time-dependent coefficients in cox's regression models. *Scandinavian Journal of Statistics*, *30*(1), 93–111.

Datta, S., Le-Rademacher, J., & Datta, S. (2007). Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and lasso. *Biometrics*, *63*(1), 259–271.

Delignette-Muller, M.-L., Dutang, C., Pouillot, R., Denis, J.-B., & Siberchicot, A. (2019). Help to fit of a parametric distribution to non-censored or censored data. *see https://cran.r-project.org/web/packages/fitdistrplus/fitdistrplus.pdf*.

Despa, S. (2010). What is survival analysis. *Cornell University Statistical Consulting Unit*.

Do Ha, I., Jeong, J.-H., & Lee, Y. (2018). *Statistical modelling of survival data with random effects: h-likelihood approach*. Springer.

Duchateau, L., Janssen, P., Kezic, I., & Fortpied, C. (2003). Evolution of recurrent asthma event rate over time in frailty models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *52*(3), 355–363.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, *32*(2), 407–499.

Efroymson, M. (1960). Multiple regression analysis. *Mathematical methods for digital computers*, 191–203.

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, 89–102.

Engler, D., & Li, Y. (2009). Survival analysis with high-dimensional covariates: an application in microarray studies. *Statistical applications in genetics and molecular biology*, *8*(1), 1–22.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), 1348–1360.

Fan, J., & Li, R. (2002). Variable selection for cox's proportional hazards model and frailty model. *Annals of Statistics*, 74–99.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis* (Vol. 998). John Wiley & Sons.

for Digital Research, U. I., & Education. (n.d.). *Introduction to linear mixed models.* (Accessed: 2010-09-30)

Foster, S. D., Verbyla, A. P., & Pitchford, W. S. (2007). Incorporating lasso effects into a mixed model for quantitative trait loci detection. *Journal of agricultural, biological, and environmental statistics*, *12*(2), 300.

Friedman, J., & Popescu, B. E. (2003). *Gradient directed regularization for linear regression and classification* (Tech. Rep.). Technical Report, Statistics Department,

Stanford University.

Garcia, T. P., Müller, S., Carroll, R. J., & Walzem, R. L. (2013). Identification of important regressor groups, subgroups and individuals via regularization methods: application to gut microbiome data. *Bioinformatics*, *30*(6), 831–837.

George, G., Osinga, E. C., Lavie, D., & Scott, B. A. (2016). *Big data and data science methods for management research.* Academy of Management Briarcliff Manor, NY.

Gibb, D. M., Duong, T., Leclezio, V. A., Walker, A. S., Verweel, G., Dunn, D. T., et al. (2004). Immunologic changes during unplanned treatment interruptions of highly active antiretroviral therapy in children with human immunodeficiency virus type 1 infection. *The Pediatric infectious disease journal*, *23*(5), 446–450.

Gilbert, P., & Varadhan, R. (2009). Accurate numerical derivatives. *see https://cran.r-project.org/web/packages/numDeriv/numDeriv.pdf*.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, *87*(420), 942–951.

Groll, A., Hastie, T., & Tutz, G. (2017). Selection of effects in cox frailty models by regularization methods. *Biometrics*, *73*(3), 846–856.

Groll, A., & Tutz, G. (2014). Variable selection for generalized linear mixed models by l 1-penalized estimation. *Statistics and Computing*, *24*(2), 137–154.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157–1182.

Harrell, F. E. (2001). Resampling, validating, describing, and simplifying the model. In *Regression modeling strategies* (pp. 87–103). Springer.

Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 757–796.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthog-

onal problems. *Technometrics*, *12*(1), 55–67.

Hosmer Jr, D. W., Lemeshow, S., & May, S. (2011). *Applied survival analysis: regression modeling of time-to-event data* (Vol. 618). John Wiley & Sons.

Hougaard, P. (2012). *Analysis of multivariate survival data*. Springer Science & Business Media.

Huang, J., & Ma, S. (2010). Variable selection in the accelerated failure time model via the bridge method. *Lifetime data analysis*, *16*(2), 176–195.

Huang, J., Ma, S., & Xie, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics*, *62*(3), 813–820.

Hui, F. K. (2017). Regularized pql for joint selection in glmms. *see https://cran.r-project.org/web/packages/rpql/rpql.pdf*.

Hui, F. K., Müller, S., & Welsh, A. (2017). Joint selection in mixed models using regularized pql. *Journal of the American Statistical Association*, *112*(519), 1323–1333.

Hutton, J., & Monaghan, P. (2002). Choice of parametric accelerated life and proportional hazards models for survival data: asymptotic results. *Lifetime data analysis*, *8*(4), 375–393.

Ibrahim, J. G., Zhu, H., Garcia, R. I., & Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, *67*(2), 495–503.

Jiang, J., Jia, H., & Chen, H. (2001). Maximum posterior estimation of random effects in generalized linear mixed models. *Statistica Sinica*, *11*(1), 97–120.

Jin, Z., & He, W. (2016). Local linear regression on correlated survival data. *Journal of Multivariate Analysis*, *147*, 285–294.

Jin, Z., Lin, D., & Ying, Z. (2006). On least-squares regression with censored data. *Biometrika*, *93*(1), 147–161.

Keiding, N., Andersen, P. K., & Klein, J. P. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates.

Statistics in medicine, *16*(2), 215–224.

Khan, M. H. R., & Shaw, J. E. H. (2016). Variable selection for survival data with a class of adaptive elastic net techniques. *Statistics and Computing*, *26*(3), 725–741.

Klein, J. P. (1992). Semiparametric estimation of random effects using the cox model based on the em algorithm. *Biometrics*, 795–806.

Komarek, A. (2006). *Accelerated failure time models for multivariate interval-censored data with flexible distributional assumptions* (Unpublished doctoral dissertation). PhD thesis, PhD thesis, Katholieke Universiteit Leuven, Faculteit Wetenschappen.

Komárek, A., & Lesaffre, E. (2008). Bayesian accelerated failure time model with multivariate doubly interval-censored data and flexible distributional assumptions. *Journal of the American Statistical Association*, *103*(482), 523–533.

Lambert, P., Collett, D., Kimber, A., & Johnson, R. (2004). Parametric accelerated failure time models with random effects and an application to kidney transplant survival. *Statistics in medicine*, *23*(20), 3177–3192.

Lin, B., Pang, Z., & Jiang, J. (2013). Fixed and random effects selection by reml and pathwise coordinate optimization. *Journal of Computational and Graphical Statistics*, *22*(2), 341–355.

Lin, D., Wei, L., & Ying, Z. (1998). Accelerated failure time models for counting processes. *Biometrika*, *85*(3), 605–618.

Lin, J.-S., & Wei, L. (1992). Linear regression analysis for multivariate failure time observations. *Journal of the American Statistical Association*, *87*(420), 1091–1097.

Liu, X. (2012). *Survival analysis: models and applications*. John Wiley & Sons.

Magnus, J., & Neudecker, H. (1999). *Wiley series in probability and statistics.* Chichester: John Wiley & Sons, Ltd.

Miloslavsky, M., Keleş, S., van der Laan, M. J., & Butler, S. (2004). Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. *Journal*

of the Royal Statistical Society: Series B (Statistical Methodology), *66*(1), 239–257.

Monpoux, F., Tricoire, J., Lalande, M., Reliquet, V., Bebin, B., & Thuret, I. (2004). Treatment interruption for virological failure or as sparing regimen in children with chronic hiv-1 infection. *Aids*, *18*(18), 2401–2409.

Ni, X., Zhang, D., & Zhang, H. H. (2010). Variable selection for semiparametric mixed models in longitudinal studies. *Biometrics*, *66*(1), 79–88.

Nielsen, G. G., Gill, R. D., Andersen, P. K., & Sørensen, T. I. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian journal of Statistics*, 25–43.

Orbe, J., Ferreira, E., & Núñez-Antón, V. (2002). Comparing proportional hazards and accelerated failure time models for survival analysis. *Statistics in medicine*, *21*(22), 3493–3510.

Osterberg, L., & Blaschke, T. (2005). Adherence to medication. *New England journal of medicine*, *353*(5), 487–497.

Pai, N. P., Tulsky, J. P., Lawrence, J., Colford, J. M., & Reingold, A. L. (2005). Structured treatment interruptions (sti) in chronic suppressed hiv infection in adults. *Cochrane Database of Systematic Reviews*(4).

Pan, J. (2016). *Adaptive lasso for mixed model selection via profile log-likelihood* (Unpublished doctoral dissertation). Bowling Green State University.

Pan, J., & Huang, C. (2014). Random effects selection in generalized linear mixed models via shrinkage penalty function. *Statistics and Computing*, *24*(5), 725–738.

Pan, W. (2001). Using frailties in the accelerated failure time model. *Lifetime Data Analysis*, *7*(1), 55–64.

Park, E., & Do Ha, I. (2018). Penalized variable selection for accelerated failure time models. *Communications for Statistical Applications and Methods*, *25*(6), 591–604.

Park, E., & Ha, I. D. (2018). Penalized variable selection for accelerated failure time

models with random effects. *Statistics in medicine*.

Pourhoseingholi, M. A., Hajizadeh, E., Moghimi Dehkordi, B., Safaee, A., Abadi, A., Zali, M. R., et al. (2007). Comparing cox regression and parametric models for survival of patients with gastric carcinoma. *Asian Pacific Journal of Cancer Prevention*, *8*(3), 412.

Prentice, R. L., Williams, B. J., & Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, *68*(2), 373–379.

Ripatti, S., & Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, *56*(4), 1016–1022.

Saitoh, A., Foca, M., Viani, R. M., Heffernan-Vacca, S., Vaida, F., Lujan-Zilbermann, J., ... Spector, S. A. (2008). Clinical outcomes after an unstructured treatment interruption in children and adolescents with perinatally acquired hiv infection. *Pediatrics*, *121*(3), e513–e521.

Schelldorfer, J., Bühlmann, P., & DE GEER, S. V. (2011). Estimation for high-dimensional linear mixed-effects models using 1-penalization. *Scandinavian Journal of Statistics*, *38*(2), 197–214.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Sha, N., Tadesse, M. G., & Vannucci, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics*, *22*(18), 2262–2268.

Sleeper, L. A., & Harrington, D. P. (1990). Regression splines in the cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association*, *85*(412), 941–949.

Soret, P., Avalos, M., Wittkop, L., Commenges, D., & Thiébaut, R. (2018). Lasso regularization for left-censored gaussian outcome and high-dimensional predictors. *BMC*

medical research methodology, *18*(1), 159.

Stute, W., Wang, J.-L., et al. (1993). The strong law under random censorship. *The Annals of statistics*, *21*(3), 1591–1607.

Sugiura, N. (1978). Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's. *Communications in Statistics-Theory and Methods*, *7*(1), 13–26.

Therneau, T. M., Grambsch, P. M., & Pankratz, V. S. (2003). Penalized survival models and frailty. *Journal of computational and graphical statistics*, *12*(1), 156–175.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in medicine*, *16*(4), 385–395.

Tibshirani, R. J., et al. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, *7*, 1456–1490.

Tsiatis, A. A., et al. (1990). Estimating regression parameters using linear rank tests for censored data. *The Annals of Statistics*, *18*(1), 354–372.

Vaida, F., & Liu, L. (2009a). Fast implementation for normal mixed effects models with censored response. *Journal of Computational and Graphical Statistics*, *18*(4), 797–817.

Vaida, F., & Liu, L. (2009b). Linear mixed-effects models with censored responses. *see https://cran.r-project.org/web/packages/lmec/lmec.pdf*.

Wang, H., Li, R., & Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, *94*(3), 553–568.

Wang, M.-C., Qin, J., & Chiang, C.-T. (2001). Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association*, *96*(455), 1057–1065.

Wang, S., Nan, B., Zhu, J., & Beer, D. G. (2008). Doubly penalized buckley–james method for survival data with high-dimensional covariates. *Biometrics*, *64*(1), 132–140.

Wang, Y. (2006). Estimation of accelerated failure time models with random effects.

Weakliem, D. L. (1999). A critique of the bayesian information criterion for model selection. *Sociological Methods & Research*, *27*(3), 359–397.

Wei, L.-J. (1992). The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, *11*(14-15), 1871–1879.

Wei, L.-J., Lin, D. Y., & Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American statistical association*, *84*(408), 1065–1073.

White, J. S. (1969). The moments of log-weibull order statistics. *Technometrics*, *11*(2), 373–386.

Wu, Y. (2012). Elastic net for cox's proportional hazards model with a solution path algorithm. *Statistica Sinica*, *22*, 27.

Xu, J., Leng, C., & Ying, Z. (2010). Rank-based variable selection with censored data. *Statistics and computing*, *20*(2), 165–176.

Yu, L. (2007). *Variable selection in the general linear model for censored data* (Unpublished doctoral dissertation). The Ohio State University.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67.

Zhang, H. H., & Lu, W. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika*, *94*(3), 691–703.

Zhang, J., & Peng, Y. (2007). An alternative estimation method for the accelerated failure time frailty model. *Computational statistics & data analysis*, *51*(9), 4413–4423.

Zhang, Z., Sinha, S., Maiti, T., & Shipp, E. (2018). Bayesian variable selection in the

accelerated failure time model with an application to the surveillance, epidemiology, and end results breast cancer data. *Statistical methods in medical research*, *27*(4), 971–990.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.