Variable Selection for Multi-path Change-point Problems

Fatemeh (Azadeh) Shohoudi Mojdehi

Doctor of Philosophy

Department of Mathematics and Statistics

McGill University

Montréal, Québec

2014

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements of the degree of
Doctor of Philosophy

DEDICATION

TO MY BELOVED FAMILY,

AND MY SUPERVISOR, PROFESSOR DAVID WOLFSON.

ACKNOWLEDGEMENTS

First, I would like to express my sincere gratitude to my thesis supervisor, Professor David Wolfson, for his invaluable directions and support throughout my PhD. During my training, I was always impressed by the depth and breadth of his statistical knowledge. From him, I received many lessons relating to both academic and non-academic life. He led me to recognize my weaknesses and question my viewpoints. He provided me with opportunities to collaborate on projects in parallel with my thesis work, which gave me improved insight into the potential applications of my research. I am truly indebted to him for his support and encouragement. I am also grateful to Professor Christina Wolfson, who gave me the chance to work part-time in her Neuroepidemiology Research Unit. Through this experience, I learnt about the intricacies of data collection in a large cross-cultural study. Working in her group also allowed me to experience great collaborations and to expand my network. I learnt about applications of my statistical knowledge in epidemiology and how to clearly communicate with researchers in other fields. These skills will definitely help me throughout my future career.

I am thankful to Professor Abbas Khalili who never hesitated to offer assistance when I needed help with my research. I learnt a lot from him through our work on a manuscript related to this thesis. His comments and criticisms were essential in improving my work. I am also grateful to Professor Masoud Asgharian for his support during the first two years after my arrival to Canada and McGill University. I am profoundly grateful to all the professors in the Department of Mathematics and

Statistics who made my doctoral studies such an enriching experience: Professor David Stephens, Professor Russell Steele, and Professor Johanna Nešlehová. I really appreciate their constant care and efforts. I also would like to sincerely thank Professor Alain Vandal, my first teacher in the Department of Mathematics and Statistics at McGill University. Although he is no longer at McGill, he assisted me with the translation of my thesis abstract.

I would like to thank Professor David Stephens, Professor Robert Platt and Professor Aurélie Labbe for generously providing me with the much needed computational resources for an extensive period of time.

I thank Ms. Raffaella Bruno for her kind support. She and my supervisor were like family to me.

I was honored to be the recipient of the Schulich Graduate Fellowship for two years. This scholarship motivated me to get through this PhD program. I thank Mr. Schulich very much for his generosity which enabled me to pursue my career goals.

I would like to thank Ana Best and Sathya Karunananthan for their support during the writing phase of my thesis work - it is deeply appreciated.

I thank my Ph.D. committee members for reading this thesis and also for their helpful comments, in particular professor Christian Genest for his editorial comments and suggestions.

Lastly, I thank my family: my sister, Atiyeh, my brother, Mohammad, and my parents, particularly my mother, Azar, who were always a great source of support, encouragement and love.

ABSTRACT

Follow-up studies on a group of units such as subjects, geographical, or ecological regions, are frequently carried out to explore the evolution of one or more observations over time. After several initial or baseline observations, either due to some intervention or natural cause, the distribution of the observations may change. When the unit-specific instants at which the changes in the distribution occur are not directly observable, statistical inference falls under the general heading of change-point inference. In this thesis we investigate the problem of covariate selection for multi-path change-point models of the type described above. Often, in applications many covariates are used and their contributions to the pre- and post-change observation distributions, as well as the change-point distribution, may be different. This creates a complex variable selection problem. The very few existing methods for variable selection in multi-path change-point models depend on either Akaike's or the Bayesian Information Criterion and are computationally infeasible, when there are even only a moderately large number of covariates. Here, we propose for the first time a penalized likelihood approach using modern regularization methods to overcome these difficulties; these include the use of the LASSO, SCAD, HARD and Ridge Regression penalty functions, which, in some cases, allow for simultaneous variable selection and parameter estimation. Our new approach is shown to be consistent in variable selection and parameter estimation. We assess the performance of our method through simulations, and demonstrate its usage in modeling cognitive decline in subjects with Alzheimer's disease.

ABRÉGÉ

Des études de suivi portant sur des groupes d'éléments tels que des individus, ou des régions géographiques ou écologiques, sont souvent réalisées dans le but d'obtenir plus d'informations sur l'évolution d'un ou plusieurs de ces éléments au fil du temps. À la suite d'observations initiales ou de référence d'un certain nombre d'éléments, la loi des valeurs observables peut être sujette à changement dû à une intervention expérimentale ou une cause naturelle. Le moment précis au cours duquel la loi d'une valeur change n'est parfois pas lui-même observable ; dans ce cas, l'inférence statistique suit le cadre général dit d'inférence sous point de rupture. Dans cette dissertation, nous considérons le problème du choix des covariables dans les modèles de point de rupture à plusieurs trajectoires décrits plus haut. Il arrive souvent dans les faits que plusieurs covariables à la fois soient utilisées et que leurs incidences sur les lois pré- et post-rupture diffèrent. Ceci induit un problème complexe de choix des covariables. Les très rares méthodes existantes pour effectuer ces choix sont basées sur les critères d'information d'Akaike ou bayésien, et mènent à des problèmes insolubles même pour un nombre modeste de covariables. Pour surmonter ces difficultés, nous proposons ici, pour la première fois, une approche de vraisemblance pénalisée fondée sur des méthodes modernes de régularisation. Ces méthodes comprennent le LASSO, SCAD, HARD et la fonction de pénalisation de la régression ridge qui, dans certains cas, permettent la sélection de covariables et l'estimation de paramètres simultanément. Nous démontrons que notre nouvelle approche est convergente en regard du modèle et des paramètres sous-jacents. Nous évaluons le rendement de

notre méthode par la voie de simulations et illustrons son utilisation pour modéliser le déclin cognitif de patients atteints de la maladie d'Alzheimer.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Statement of Originality

1. This is the first time that variable selection methods for multi-path change-point problems have been proposed and carried out.
2. There has been only one previous article allowing for the observation distributions before and after the change-points to depend on covariates. An original feature of this thesis is the use of penalized likelihood methods to account for the very large number of possible submodels, when carrying simultaneous variable selection and parameter estimation in a multi-path change-point setting. The methods include variable selection and parameter estimation of covariates in the change-point distribution, not covered by the current literature.
3. The proofs of the asymptotic properties - consistency and sparsity - in the current setting are new, as are their accompanying lemmas.
4. The development and implementation of a modified EM algorithm is new in the current setting.
5. Since the main problem has not been addressed before, the extensive simulation results presented to assess the small-and-medium sized sample behaviour of our procedure, are new.
6. The application of penalized likelihood methods to the assessment of determinants of cognitive decline in subjects with Alzheimer's disease is new, when the model allows for changes in the rate of decline at random unknown time instants.

CHAPTER 1
Introduction

This thesis brings together two topics that, separately, have been the subject of much research, over the past couple of decades. These are penalized likelihood methods and change-point problems. The current literature on the use of modern regularization methods to carry out variable selection in change-point problems is restricted to two papers (Anraku 1999, Ninomiya 2005) and these address problems different from those in this thesis. By proposing methods for simultaneous variable selection and parameter estimation in both the observation distribution(s) and the change-point distribution, we fill a gap in the literature.

Change-point problems in which explanatory variables (covariates) form part of the model provide natural settings for the application of penalized likelihood methods. We begin with an example that demonstrates the need for methods beyond, for example, the Bayesian Information Criterion (BIC) (Schwarz 1978) or Akaike's Information Criterion (AIC) (Akaike 1973).

In Alzheimer's disease (AD), the rate of progression is highly variable and there has been much interest in the identification of factors associated with cognitive decline, for example, age, education, and sex (Mortimer et al. 1992, Hall et al. 2007). Further, some studies have suggested that the rate of cognitive decline in patients with AD is not constant and is piece-wise linear (Joseph et al. 1999, Hall et al. 2000),

and most researchers would agree that, broadly, there is an initial stable period followed by a period of roughly linear decline, ending with another relatively stable period, late in the disease process. Therefore, it is reasonable to assume the observation distributions of a measure of cognitive ability, recorded on subjects over time, might change at unknown time-points. In this so-called multi-path change-point setting, selection of the factors associated with cognitive decline presents computational challenges.

Criteria for the selection of these relevant factors such as AIC and the BIC are prohibitively expensive when the number of potential submodels is very large. For instance, in the Alzheimer's disease application that we present in Chapter 6, with only ten covariates and five follow-up observations on each subject, the number of submodels is 2^{42} which makes the use of AIC and the BIC for variable selection infeasible (The pre- and post-change observation distribution regression models include ten covariates and four interaction terms and have $2^{14} \times 2^{14}$ submodels, while the change-point regression model which includes ten covariates in addition to four possible increments in the logit of the baseline hazard, has $2^{10} \times 2^4$ possible submodels). These difficulties are a feature of variable selection problems in multi-path change-point models in general. In this thesis, we present a computationally efficient method for variable selection in such models.

Amongst the first change-point papers for fixed sample sizes were those by Hinkley (1970), and Feder (1975a, 1975b) on single-path change-point models. (We will formally define the single-path and multi-path change-point models in Chapter 2). Hinkley considered maximum likelihood estimation (MLE) in the single-path

change-point setting for binomial random variables, while Feder considered single-path change-point regression models. The long lists of references in the review articles by Hackl and Westlund (1989) and Khodadadi and Asgharian (2006) represent only a fraction of the single-path change-point literature. Since our current work is concerned with the multi-path change-point setting, we shall not attempt to review the single-path change-point literature.

The multi-path change-point model was first introduced by Joseph (1989). He considered sequences of conditionally independent observations, each with a change-point. This early work was followed by several applications of multi-path change-point models (Lange et al. 1992, Joseph and Wolfson 1992, 1993, Joseph et al. 1997, Bélisle et al. 1998, Beckage et al. 2006). Later, Joseph et al. (1996) introduced sequences of correlated random variables with a change in each path and used a multi-path autoregressive change-point model to capture changes in silt concentration as a function of depth at different geographical sites. Although Young (2012) considered multi-path change-point regression models, he assumed fixed probabilities of change at different time-points, with covariate information only in the regression models before and after the change. Moreover, he neither allowed the covariates in the change-point distribution nor carried out variable selection for his proposed model.

Further, in the multi-path change-point literature few attempts were made to include covariates in the model, and in particular, no model included covariates in the change-point distribution until the work of Asgharian and Wolfson (2001). See Lange et al. (1992) and Joseph and Wolfson (1992) and Joseph et al. (1999), who introduced covariates only to model the observation distributions.

Multi-path change-point models with covariates in the change-point distribution can be considered as mixture-of-experts models (Jacob et al. 1991). The basic assumptions for these two models are, however, different as will be pointed out in Section 4.1 of Chapter 2. The above authors were concerned with estimation rather than variable selection. There is a limited literature on variable selection for mixture-of-experts models (Khalili 2010).

In a broad setting, suppose we observe a stochastic process, $\{\mathbf{Y}_t, t \in [\mathbf{0}, \mathbf{T}]\} = \{\mathbf{Y}_t\}$, where without loss of generality the index $t \in [\mathbf{0}, \mathbf{T}]$ represents time. At some unknown time-point $\tau \in [\mathbf{0}, \mathbf{T}]$, it is assumed that the distribution of $\mathbf{X}_t$ *might* change (if $\tau = T$, by convention, no change is said to have occurred). In the simplest case, we allow only one possible change-point. The main inferential issues that are commonly considered are: 1) Determining if any change occurred in the distribution of the stochastic process. 2) If a change has occurred, estimating the time to change and 3), estimating the distributions before and after the change - assuming there has been one. For brevity, we shall simply refer to the observation distributions; for example, blood pressure readings on a subject, over time may be modelled as a stochastic process. If the subject undergoes a medical intervention, it is possible for the distribution of blood pressures to change. However, we expect this distribution will often change with a lag which is not directly observable. It is also possible that the intervention will have no effect. Now suppose that instead of a single stochastic process on $[\mathbf{0}, \mathbf{T}]$, we observe n stochastic processes, each on $[\mathbf{0}, \mathbf{T}]$, corresponding to n subjects, each with the possibility of a change, at $\tau_i \in [\mathbf{0}, \mathbf{T}]$, for $i = 1, 2, \ldots, n$.

This is called a multi-path change-point setting. For now, we avoid the details of the model assumptions; we will discuss them in Chapter 2.

In the above description, we have assumed the changes to be sudden. This could happen, for example, if there is an unobservable underlying latent process which induces a sudden change only when the process crosses a certain threshold. Threshold models have been considered by Tong (1983), Petruccelli and Davies (1986), Chan and Tsay (1998) and Tsay (1997). Alternatively, we might consider a two-distribution sudden change model as a first approximation to a model that allows for a gradual change. A further extension would include bi-phasic regression models as illustrated by our example in Chapter 6.

Now, in practice, one takes a discrete set of observations on $[\mathbf{0}, \mathbf{T}]$ for each path, leading naturally to a random effects model that includes the τ_is, the path-specific change-points, as random effects with distributions, $P(\tau_i = k),\ k = 1, 2, \ldots, m,\ \ i = 1, 2, \ldots, n$. This random effects model permits us to draw strength from the ensemble of n sequences of observations. To be consistent with the change-point literature, we shall refer to the *index* τ_i, induced by the discrete sampling of each process $\{\mathbf{Y}_t\}$, as the change-point for the i^{th} sequence of observations, Y_{ij} for $j = 1, 2, \ldots, m$ and τ_i could be any of the points, $1, 2, \ldots, m$. When $\tau_i = m$ no change is said to occur.

Extending our model, it is clear that in most applications the observation distributions will depend on path (sequence)-specific covariates. We shall also assume that the change-point probabilities, $P(\tau_i = k)$, depend on path-specific covariates, $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$ where p represents the number of covariates. However, in order to be able to draw strength from the different sequences we shall assume that

covariate effects are common across the sequences. Thus, we shall consider a discrete set of conditionally independent observations given the change-points and other model parameters.

It is common to introduce a large number of predictors at the initial stage of modelling, but this is not always the best choice. In this thesis, we present variable selection methods for covariates that affect both the change-point distribution as well as the observation distributions before and after the change-point.

Wahba (1990), and Green and Silverman (1994) and the references therein, introduced penalized likelihood methods using quadratic penalty functions. These reduce the variability of estimators via an L_2-norm (ridge regression) penalty function. Perhaps the three most used penalty functions are the LASSO (Tibshirani 1996), the SCAD (Fan and Li 2001), and the HARD (Fan and Li 2002). The LASSO is based on an L_1-norm penalty function, and the SCAD and the HARD are weighted sums of L_1- and L_2-norms. Harchaoui and Lévy-Leduc (2010) used the LASSO and LARS penalty functions to identify the location of change-points in one dimensional piece-wise constant signals observed in the presence of white noise. The SCAD and the HARD have the oracle property, meaning that the penalized maximum likelihood estimators perform as well as maximum likelihood estimators of the nonzero parameters knowing which parameters are equal to zero. Here, we consider these three penalty functions to simultaneously carry out model selection and parameter estimation. Our penalized likelihood approach is shown to be consistent in variable selection and parameter estimation. We assess its performance through simulations,

and demonstrate its usage in modelling cognitive decline in subjects with Alzheimer's disease.

The layout of this thesis is as follows. In Chapter 2, we formally introduce change-point models, in particular multi-path change-point models. We discuss the similarities and differences between our model and the mixture-of-experts model. We also assess the identifiability of our change-point models. In Chapter 3, we discuss several common variable selection approaches and the advantages of applying a penalized likelihood approach. We emphasize the penalty functions we use in this thesis and their properties. We present our model and numerical methods for estimation in Chapter 4. We also derive the asymptotic properties of these estimators. In Chapter 5, we investigate the finite sample properties of our methods through simulations, and in Chapter 6 we use our methods to study predictors of cognitive decline in subjects with Alzheimer's disease. Chapter 7 includes closing remarks and suggestions for several possible further research directions.

CHAPTER 2
Change-point Problems

In many single change-point settings we observe a sequence of ordered random variables at successive time-points or locations. *A priori*, we believe that a change over time/location in the data may have occurred. That is, the random variables before the change cannot be exchanged with the variables after the change. The defining feature is that if a change has occurred, the location of this change-point is unknown. There are three main problems of statistical inference that arise from such data: i) Test whether a change has occurred, ii) If there is a reason to believe that a change has occurred, make inference about the location of the change and iii) make inference about the pre- and post-change-point observation distributions.

Definition 1. The single-path change-point model: Let $Y_1, Y_2, \ldots, Y_m$ be observations (responses) on the time interval $[0, T]$, taken at equally spaced time-points $0 = t_1 < t_2 < \cdots < t_m = T$. We say a change occurs at $\tau = k$ for $k = 1, 2, \ldots, m-1$, if $Y_1, Y_2, \ldots, Y_k$ have joint cumulative distribution function (c.d.f.) $F_0(\cdot)$, and $Y_{k+1}, Y_{k+2}, \ldots, Y_m$ have joint cumulative distribution function $F_1(\cdot)$ which is different from $F_0(\cdot)$. If $\tau = m$, we say there is no change in the distribution of the sequence of observations. Henceforth, we shall assume that $Y_1, Y_2, \ldots, Y_k$ are independent and identically distributed (i.i.d), conditional on k and other model parameters. Similarly, for $Y_{k+1}, Y_{k+2}, \ldots, Y_m$. Under this assumption, $F_0(\cdot)$ and $F_1(\cdot)$ will be the respective marginal c.d.f.s of Y_is.

For the observation distributions there are several possibilities:

1) The cumulative distribution functions, $F_0(\cdot)$ and $F_1(\cdot)$ are completely unspecified.
2) $F_0(\cdot)$ and $F_1(\cdot)$ are specified up to a finite number of unknown parameters, θ_0 and θ_1. That is, $F_0(\cdot) = F_0(\cdot, \theta_0)$ and $F_1(\cdot) = F_1(\cdot, \theta_1)$.
3) Under 2), even θ_0 and θ_1 are specified *a priori*.

For the third scenario, inference is then only about the change-point, while in the first and second scenarios, inference is often made about the observation distributions F_0 and F_1 as well.

A common change-point model assumes the observation random variables are identically normally distributed both before and after the change, while their means θ_0 and θ_1 differ but their variances do not.

Although not the setting of this thesis, it is possible to assume multivariate distributions with a correlation structure for the random variables both before and after the change. We avoid this added complexity in this thesis since the large number of unknown parameters with which we must contend in our simpler model provides a considerable challenge on its own.

Another generalization is to allow for multiple change-points, thereby once more introducing additional complexity into the model. We restrict our discussion to a single change-point scenario.

2.1 Multi-path Change-point Problems

Multi-path change-point settings differ from their single-path counterparts in that there are n paths, or sequences, of observations, usually each with m follow-up

observations over time (or location). It is common to assume independence between paths, and we will do so here. If we instead were to assume the paths to be correlated, this correlation would need to be modelled.

In the multi-path setting, it is assumed that the change for each path can happen at a different time-point (location). If we assume that each path represents a different subject, this means that each subject's observation distribution changes at a different time-point, as might be the case, for example, in the longitudinal follow-up of a group of study subjects of size n. If the change-points are unknown parameters, they contribute n unknown parameters to the model in addition to the unknown parameters of the observation distributions before and after the change. To overcome potential over-parametrization, we may impose a distribution on the change-points. We shall take this approach and refer to this distribution as the change-point distribution. We now formally introduce the model we propose for the multi-path change-point setting.

Definition 2. The multi-path change-point model: Let observations (responses) on the time interval $[0, T]$ be taken on n subjects, at equally spaced time-points $0 = t_1 < t_2 < \cdots < t_m = T$. Let $(\mathbf{Y}_i, \mathbf{X}_i) = (Y_{i1}, Y_{i2}, \ldots, Y_{im}, X_{i1}, X_{i2}, \ldots, X_{ip})$, denote the vector of observations for subject i, where $\mathbf{Y}_i$ corresponds to the vector of responses over time and $\mathbf{X}_i$ corresponds to the vector of p fixed covariates that accompany each subject. The corresponding realized values are denoted by $(\mathbf{y}_i, \boldsymbol{x}_i) = (y_{i1}, y_{i2}, \ldots, y_{im}, x_{i1}, x_{i2}, \ldots, x_{ip})$, $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$ and $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$. For simplicity, we shall refer to each path as a subject. We denote the observation on the i^{th} subject at the ℓ^{th} time by $y_{i\ell}$. We shall call τ_i a change-point for the observations

on subject i if the observations before the change-point, $Y_{i1}, Y_{i2}, \ldots, Y_{i\tau_i}$, have a different distribution from the observations after the change-point, $Y_{i\tau_i+1}, Y_{i\tau_i+2}, \ldots, Y_{im}$. Specifically, τ_i is said to be a change-point if, conditional on $\tau_i = k_i$, the observations $Y_{i1}, Y_{i2}, \ldots, Y_{ik_i}$ have probability density function $f_{i0}(\cdot)$ and the observations $Y_{i(k_i+1)}, Y_{i(k_i+2)}, \ldots, Y_{im}$ have probability density function $f_{i1}(\cdot)$, which is different from $f_{i0}(\cdot)$. If $\tau_i = m$, we say there is no change in the distribution of the sequence of observations.

We assume there are no missing values in the sequences of observations or in the vectors of covariates. The realizations on n paths form a matrix as follows:

$$\begin{array}{c} \tau_1 \\ \vdots \\ \tau_n \end{array} \begin{pmatrix} Y_{11} & \ldots & Y_{1m} \\ \vdots & \vdots & \vdots \\ Y_{n1} & \ldots & Y_{nm} \end{pmatrix}$$

The final step in defining our model is to specify how we include covariates.

2.1.1 Introducing Covariates into the Model

We begin by specifying how the observation distribution depends on putative covariates. Following common practice, we assume that the covariates enter through an appropriate link function. For example, the observations could follow a generalized linear model, or a linear model with normal errors. In our simulations and example, we impose linear models for the observations before and after the change-point. The change in the distribution occurs through the change in the regression coefficients - that is, the covariate effects change after the change-point. In the linear model with time as a covariate, if $\tau_i = k$, for $k = 1, \ldots, m-1$, we assume

$Y_{il} \overset{i.d}{\sim} N(\beta_{10} + \eta_1 l + \boldsymbol{x}_i^\top \boldsymbol{\beta}_1, \sigma_1^2)$ for $l = 1, 2, \ldots, k$ and $Y_{il} \overset{i.d}{\sim} N(\beta_{20} + \eta_2 l + \boldsymbol{x}_i^\top \boldsymbol{\beta}_2, \sigma_2^2)$ for $l = k+1, k+2, \ldots, m$, respectively. We shall call $\tau_i < m$ a change-point for $i = 1, 2, \ldots, n$ if, conditional on $\tau_i = k_i$ and the covariate values $\boldsymbol{x}_i$, the observations Y_{il} for $l = 1, 2, \ldots, k_i$ have probability density function $f_1^*(\cdot; \boldsymbol{\theta}_1(l, \boldsymbol{x}_i), \sigma_1^2)$, respectively and the observations Y_{il} for $l = (k_i + 1), (k_i + 2), \ldots, m$ have probability density function $f_2^*(\cdot; \boldsymbol{\theta}_2(l, \boldsymbol{x}_i), \sigma_2^2)$, which is different from $f_1^*(\cdot; \boldsymbol{\theta}_1(l, \boldsymbol{x}_i), \sigma_1^2)$, respectively. We assume $\boldsymbol{\theta}_k(l, \boldsymbol{x}_i) = g(\beta_{k0} + \eta_k l + \boldsymbol{x}_i^\top \boldsymbol{\beta}_k)$ for $k = 1, 2$, where $g(\cdot)$ is a known link function and $(\beta_{10}, \eta_1, \boldsymbol{\beta}_1) = (\beta_{10}, \eta_1, \beta_{11}, \beta_{12}, \ldots, \beta_{1p})^\top$ and $(\beta_{20}, \eta_2, \boldsymbol{\beta}_2) = (\beta_{20}, \eta_2, \beta_{21}, \beta_{22}, \ldots, \beta_{2p})^\top$ are vectors of regression parameters before and after the change, respectively. The variance of the observation distribution before the change is denoted by σ_1^2 and the variance after the change, by σ_2^2. In our simulations and real data analysis we assumed $\sigma_1^2 = \sigma_2^2$. We denote the vector of parameters of the observation distributions by $\Upsilon = (\beta_{10}, \eta_1, \boldsymbol{\beta}_1, \beta_{20}, \eta_2, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2)$.

2.2 The Change-point Distribution

The change-points in our model have discrete distributions. To introduce covariates into the change-point distributions, given $\mathbf{X}_i = \boldsymbol{x}_i$, we assume that for each i, τ_i has a probability mass function, $P(\tau_i = k_i | \boldsymbol{x}_i, \boldsymbol{\alpha}^*)$ for $k_i = 1, 2, \ldots, m$, where $\boldsymbol{\alpha}^*$ is a vector of parameters in the change-point distribution to be defined later. The τ_is are not directly observable; they can be considered as latent variables. The distributions of the τ_is, therefore, are seen to differ from subject to subject through their respective covariate vectors, $\boldsymbol{x}_i$s.

Let $\pi_k(\boldsymbol{x}_i) = \pi_k(\boldsymbol{x}_i, \boldsymbol{\alpha}^*)$ be the conditional probability of a change at the k^{th} time-point for the i^{th} subject with covariate vector $\boldsymbol{x}_i$, given that there has been no

change at any of the previous $k-1$ time-points in the sequence of observations. By conditioning backwards, we have:

$$P(\tau = k|\boldsymbol{x}_i, \boldsymbol{\alpha}^*) = \begin{cases} \pi_k(\boldsymbol{x}_i, \boldsymbol{\alpha}^*) & k = 1 \\ \pi_k(\boldsymbol{x}_i, \boldsymbol{\alpha}^*) \prod_{l=1}^{k-1} (1 - \pi_l(\boldsymbol{x}_i, \boldsymbol{\alpha}^*)) & k = 2, \ldots, m-1 \\ \prod_{l=1}^{m-1} (1 - \pi_l(\boldsymbol{x}_i, \boldsymbol{\alpha}^*)) & k = m. \end{cases} \tag{2.2.1}$$

Therefore, it is sufficient to specify how $\pi_k(\boldsymbol{x}_i, \boldsymbol{\alpha}^*)$ depends on $\boldsymbol{x}_i$. We can consider the change-point problem in a simple survival analysis setting with one failure time since "failure" is an absorbing state. This survival analysis analogy is hypothetical, since we do not actually observe the change-point (failure time). Apart from this difference, *change-point* and *failure time* are equivalent. Hence, the hazard of a change is the same as the hazard of a failure. Since, by definition, a change can only occur at one of the discrete time-points in the sequence of observations, the change-point model may be considered as a discrete survival analysis model with covariates. A common method of modelling the hazard of failure in a discrete setting is through a proportional odds model, and therefore, we introduce covariates into the change-point distribution through a proportional odds model for the hazard of change. Let $\boldsymbol{x}_{i_1}$ and $\boldsymbol{x}_{i_2}$ be two covariate vectors. The proportionality of the odds implies that

$$\frac{\dfrac{\pi_k(\boldsymbol{x}_{i_1})}{1-\pi_k(\boldsymbol{x}_{i_1})}}{\dfrac{\pi_k(\boldsymbol{x}_{i_2})}{1-\pi_k(\boldsymbol{x}_{i_2})}} = \exp\{(\boldsymbol{x}_{i_1} - \boldsymbol{x}_{i_2})^\top \boldsymbol{\alpha}\}. \tag{2.2.2}$$

This form of the odds ratio yields a specific form for the hazard function. If we set $\boldsymbol{x}_{i_1} = \boldsymbol{x}_i$ and $\boldsymbol{x}_{i_2} = \mathbf{0}$, then

$$\frac{\dfrac{\pi_k(\boldsymbol{x}_i)}{1-\pi_k(\boldsymbol{x}_i)}}{\dfrac{\pi_k(\mathbf{0})}{1-\pi_k(\mathbf{0})}} = \exp(\boldsymbol{x}_i^\top \boldsymbol{\alpha}).$$

Calling $\pi_k(\mathbf{0}) = \rho_k$, the baseline hazard at time k, we have

$$\frac{\pi_k(\boldsymbol{x}_i)}{1-\pi_k(\boldsymbol{x}_i)} = \frac{\rho_k}{1-\rho_k}\exp(\boldsymbol{x}_i^\top \boldsymbol{\alpha}).$$

By regarding the $\pi_k(\boldsymbol{x}_i, \boldsymbol{\alpha}^*)$s as the hazards in a proportional odds model and setting $\dfrac{\rho_k}{1-\rho_k} = \exp\left(\sum_{l=1}^{k} \alpha_{0l}\right)$, we obtain

$$\pi_k(\boldsymbol{x}_i, \boldsymbol{\alpha}^*) = \frac{\exp\left\{\sum_{l=1}^{k} \alpha_{0l} + \boldsymbol{x}_i^\top \boldsymbol{\alpha}\right\}}{1 + \exp\left\{\sum_{l=1}^{k} \alpha_{0l} + \boldsymbol{x}_i^\top \boldsymbol{\alpha}\right\}}. \tag{2.2.3}$$

for $k = 1, 2, \ldots, (m-1)$ where $\boldsymbol{\alpha}^* = (\alpha_{01}, \alpha_{02}, \ldots, \alpha_{0(m-1)}, \boldsymbol{\alpha}) = (\alpha_{01}, \alpha_{02}, \ldots, \alpha_{0(m-1)}, \alpha_1, \alpha_2, \ldots, \alpha_p)$. In this model, α_{0k} corresponds to the increment in the baseline hazard at the time-point t_k (see Asgharian (2013) for details). In Chapter 4, we will discuss the reason for our different parameterization from his.

Finally, under the proportional odds assumption, using (2.2.3) we arrive at the following model for the change-point distribution that includes covariates:

$$P(\tau = k|\boldsymbol{x}_i, \boldsymbol{\alpha}^*) = \begin{cases} \dfrac{\exp\left\{\sum_{l=1}^{k} \alpha_{0l} + \boldsymbol{x}_i^\top \boldsymbol{\alpha}\right\}}{1 + \exp\left\{\sum_{l=1}^{k} \alpha_{0l} + \boldsymbol{x}_i^\top \boldsymbol{\alpha}\right\}} & k = 1 \\ \dfrac{\exp\left\{\sum_{l=1}^{k} \alpha_{0l} + \boldsymbol{x}_i^\top \boldsymbol{\alpha}\right\}}{1 + \exp\left\{\sum_{l=1}^{k} \alpha_{0l} + \boldsymbol{x}_i^\top \boldsymbol{\alpha}\right\}} \prod_{l=1}^{k-1} \left(\dfrac{1}{1 + \exp\left\{\sum_{s=1}^{l} \alpha_{0s} + \boldsymbol{x}_i^\top \boldsymbol{\alpha}\right\}} \right) & k = 2, \ldots, m-1 \\ \prod_{l=1}^{m-1} \left(\dfrac{1}{1 + \exp\left\{\sum_{s=1}^{l} \alpha_{0s} + \boldsymbol{x}_i^\top \boldsymbol{\alpha}\right\}} \right) & k = m. \end{cases} \tag{2.2.4}$$

2.3 Likelihood Function

Conditional on the τ_is and covariates, we assume the $m \times n$ observations Y_{ij} are independent. Hence, the conditional joint density for the i^{th} subject, given $\tau_i = k_i$ and covariates, is

$$f_{k_i}(\mathbf{y}_i|\boldsymbol{x}_i, \Upsilon) = \prod_{l=1}^{k_i} f_1^*(y_{il}|\boldsymbol{\theta}_1(l, \boldsymbol{x}_i), \sigma_1^2) \prod_{l=k_i+1}^{m} f_2^*(y_{il}|\boldsymbol{\theta}_2(l, \boldsymbol{x}_i), \sigma_2^2) \tag{2.3.1}$$

for $k_i = 1, \ldots, m-1$, while

$$f_m(\mathbf{y}_i|\boldsymbol{x}_i, \Upsilon) = \prod_{l=1}^{m} f_1^*(y_{il}|\boldsymbol{\theta}_1(l, \boldsymbol{x}_i), \sigma_1^2). \tag{2.3.2}$$

Note that if $\tau_i = m$, by convention, no change is said to occur.

The joint density for the i^{th} subject, given the covariates and averaging over the possible change-points, is

$$f(\mathbf{y}_i|\boldsymbol{x}_i, \boldsymbol{\Phi}) = \sum_{k=1}^{m} p(\tau_i = k|\boldsymbol{x}_i, \boldsymbol{\alpha}^*) f_k(\mathbf{y}_i|\boldsymbol{x}_i, \Upsilon), \tag{2.3.3}$$

which has the form of a mixture model, and $\boldsymbol{\Phi} = (\Upsilon, \boldsymbol{\alpha}^*) = (\beta_{10}, \eta_1, \boldsymbol{\beta}_1, \beta_{20}, \eta_2, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, \boldsymbol{\alpha}^*)$ is the vector of all the parameters in the model. Using the conditional independence of the observations between subjects, the unconditional (on $\boldsymbol{\tau}$) joint density over all subjects is:

$$L_n(\boldsymbol{\Phi}) = \prod_{i=1}^{n} f(\mathbf{y}_i|\boldsymbol{x}_i, \boldsymbol{\Phi}) = \prod_{i=1}^{n} \left[\sum_{k=1}^{m} p(\tau_i = k|\boldsymbol{x}_i, \boldsymbol{\alpha}^*) f_k(\mathbf{y}_i|\boldsymbol{x}_i, \Upsilon), \right]. \tag{2.3.4}$$

The likelihood (2.3.4) was introduced by Asgharian (1998) and Asgharian and Wolfson (2001).

2.3.1 Mixture-Of-Experts Models (MOEs): Similarities and Differences

The likelihood in the multi-path change-point setting considered in this thesis is a mixture and is sometimes called a mixture-of-experts model. In the mixture-of-experts models, covariate information is included in the mixing proportions as well as the observation distributions. Variable selection is more difficult in this type of mixture model than in standard mixture models.

It is important to examine the similarities between general mixture models and our change-point mixture model. In particular, we shall compare our model with so-called mixture-of-experts models to which variable selection methods have already been applied (Khalili 2010).

The most general finite mixture model consists of m (possibly multivariate) mixture component distributions, and m corresponding mixing proportions (weights). The mixture distribution is then the weighted sum of the mixture components.

Mixture-of-experts models form a subclass of general mixture models. They are characterized by the following features:

1) The m mixture components could be multivariate.
2) Covariates are assumed to affect the outcomes associated with the mixture components. Each mixture component is assumed *a priori* to be characterized by a different covariate effect vector.
3) The mixing proportions may also depend on covariates whose covariate effect vectors are the same for all mixing proportions.
4) A single observation drawn from this model consists of an observed covariate vector (the covariates deemed, *a priori*, to be relevant to the mixture or mixing components) and the outcome variable drawn from the mixture distribution (the weighted sum).
5) Under conditional independence, the joint likelihood of n such observations is the product of the weighted sums introduced in feature 4).
6) The mixture components are exchangeable, meaning that there is no difference in the model if we exchange the order of any two components. (A permutation does not change the model).

Our main goals are variable selection and covariate effect estimation. In mixture-of-experts models, these goals are achievable provided the number of mixture components is small (generally no more than four) and the outcome variable is univariate.

The multi-path change-point mixture model is characterized by the following features.

1) The structure of the model unavoidably induces multivariate mixture components: at the lowest level in the structure, each observation point in time results in a univariate observation drawn from one of two possible distributions. Given the change-point occurs at k (for $k = 1, 2, \ldots, m$), the observations up to and including the k^{th} are drawn from the first distribution and those occurring from the $(k+1)^{\text{th}}$ to the m^{th} are drawn from the second distribution. At the next level, under conditional independence of the observations for each subject, the joint distributions of the observations before and after the change-point k, are, respectively, k- and $(m-k)$-dimensional. At the third level, the product of these two multivariate distributions defines the distribution of the m-dimensional k^{th} mixture component. The uncertainty in the location of the change-point means that the unconditional m-dimensional joint distribution of all m univariate observations taken on a subject is a mixture distribution, with mixing proportions that depend on the possible locations of the change-point.
2) Covariates are assumed to affect the outcomes associated with the univariate distributions introduced in feature 1), and hence the multivariate mixing components. These covariate effects are assumed to be the same for all subjects.
3) Each mixing proportion may also depend on covariates whose covariate effect vectors are the same for all mixing proportions, and for all subjects.
4) A single observation vector comprises an observed covariate vector for a specific subject (the union of the covariates deemed, *a priori*, to be relevant to the two

univariate distributions and those pertinent to the mixing (i.e., change-point) distribution), and the multivariate outcome variable drawn from the mixture distribution (the weighted sum).

5) Under conditional independence, the joint likelihood of n such observations is the product of the weighted sums of observation distributions introduced in feature 4). Recall that our main goals are variable selection and covariate effect estimation. In the multi-path change-point model, these are achievable even when the number of mixture components is large.

6) The components of the mixture likelihood are not exchangeable. Since the components are ordered by the location of the change, they cannot be permuted.

This model contrasts with mixture-of-experts models due to the assumption of common covariate effects on the outcomes and on the special structure of the multivariate mixture components. The time order in which the univariate observations occur contributes to this structure.

2.4 Inference for Single- and Multi-path Change-point Problems

In our change-point model, inference is about the unknown parameters introduced into the model. In a single-path change-point setting, inference is about the pre- and post-change observation distributions and the location at which a change may have occurred. For different approaches in this setting, see Hinkley (1970), and Smith (1975), who took frequentist and Bayesian approaches, respectively, to make inference about their proposed model.

In our multi-path change-point setting, our inference is likelihood-based and covers both observation distributions before and after the change as well as the change-point distribution. Perhaps the first attempt to model the change-point distribution as a function of covariates was by Asgharian (1998) and Asgharian and Wolfson (2001). Alternatively, Joseph and Wolfson (1992) and Lange et al. (1992) had a Bayesian perspective. As we shall soon see in Chapters 3 and 4, a straight likelihood approach often cannot be used and must be modified by the introduction of a penalty function. Indeed, penalized likelihood methods are essential to this dissertation.

2.5 Identifiability of the Model

An important consideration in the application of multi-path change-point models is that of identifiability. In general, the identifiability of a family of distribution functions is defined as follows:

Definition 3: A parametric family of probability distributions $\mathcal{P} = \{F_{\mathbf{\Phi}}; \mathbf{\Phi} \in \mathbf{\Theta}\}$, on a sample space $\mathcal{W}$, which has density $f(\cdot; \mathbf{\Phi})$ with respect to a σ-finite measure μ is identifiable when, for any $\mathbf{\Phi}, \mathbf{\Phi}^* \in \mathbf{\Theta}$, if $f(w; \mathbf{\Phi}) = f(w; \mathbf{\Phi}^*)$ almost everywhere with respect to μ, then $\mathbf{\Phi} = \mathbf{\Phi}^*$.

The introduction of covariates into the model leads us to consider the conditional distribution (with respect to covariate values) when defining identifiability. Since our multi-path change-point model (2.3.4), is a special form of mixture models, we discuss the identifiability of mixture models, in particular the finite mixture-of-experts model. For this model, identifiability is defined as follows:

Definition 4: A finite mixture-of-experts model with the conditional density function given a design matrix $\boldsymbol{x}$

$$f(\mathbf{y}; \boldsymbol{x}, \boldsymbol{\Phi}) = \sum_{j=1}^{m} \pi_j(\boldsymbol{x}, \boldsymbol{\alpha}_j) f_j(\mathbf{y}; \boldsymbol{x}, \boldsymbol{\beta}_j), \tag{2.5.1}$$

(where $\boldsymbol{\Phi} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_m, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_m)$ is the vector of all the parameters in the model and $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$ are vectors of parameters in the mixing proportion and observation distribution, respectively for the j^{th} component for $j = 1, 2, \ldots, m$), is identifiable when for any two $\boldsymbol{\Phi}, \boldsymbol{\Phi}^* \in \boldsymbol{\Theta}$, $f(\mathbf{y}; \boldsymbol{x}, \boldsymbol{\Phi}) = f(\mathbf{y}; \boldsymbol{x}, \boldsymbol{\Phi}^*)$ for almost all $\mathbf{y}$, implies $\boldsymbol{\Phi} = \boldsymbol{\Phi}^*$.

In a mixture-of-experts model, if the order of the components changes, $\boldsymbol{\Phi}$ changes as well. Thus, in Definition 4 of identifiability, $\boldsymbol{\Phi} = \boldsymbol{\Phi}^*$ only up to a permutation in the mixture components.

Although multi-path change-point models have a similar form to (2.5.1), they do not fall exactly into the class of models assumed in Definition 4. This is due to the ordering of the mixing proportions, or probabilities of change at different time-points. These components are not exchangeable, since altering the possible times of change alters the components in the multi-path change-point model.

There are other approaches to the problem of identifiability in this setting. Young's (2012) discussion of identifiability in mixtures of regression models with change-points has similarities to our model. However, his setting is different from ours:

1. In Young's model, the mixing proportions are fixed unknown parameters, while in our model the mixing proportions depend on the vectors of covariates for different subjects.
2. In order to ensure identifiability, they suggest the introduction of an ordering constraint (either increasing or decreasing) on the mixing proportions. In our multi-path change-point setting, the mixing proportions, or probabilities of a change at different time-points are defined by (2.2.4). While each probability is a function of covariates, its special form means that for $i = 1, 2, \ldots, n$,

$$P(\tau_i = 1) \geq P(\tau_i = 2) \geq \cdots \geq P(\tau_i = m - 1).$$

   We can also constrain these probabilities to be increasing, with a suitable choice of baseline hazard. However, the probability of no change, $P(\tau_i = m)$, need not necessarily follow the same order. This is particularly problematic when the probabilities of change, and therefore, the probability of no change, depend on subject-specific covariates. Indeed, some choices of covariates could be associated with a high probability of no change.
3. Another type of non-identifiability mentioned by the author arises for some parameter combinations (see Young 2012). He claims that this type of non-identifiability may not be that serious since it has been shown by Allman et al. (2009) that these combinations of non-identifiable "generic" parameters has measure zero. However, Allman et al. define the term "generic" in a precise algebraic geometric sense, and in this context it does not mean "standard" or

"canonical." Since it is difficult to identify the set of generic parameters, there is little comfort in asserting that apart from a subset of them of measure zero, we have identifiability.

It would therefore appear that there is no obvious way to ensure identifiability of a multi-path change-point model, particularly when the change-point distribution is assumed to depend on subject-specific covariates. Viewed from a different perspective, though, Asgharian (2013) has recently made some progress on this problem.

It is well known that there is an association between identifiability of a model and the non-singularity of the Fisher information matrix. Specifically, Rothenberg (1971) showed that local identifiability is equivalent to non-singularity of the Fisher information matrix at points in the parameter space for which there is an open neighbourhood where the information matrix has constant rank. Asgharian's 2013 paper originated with his goal of showing that under certain weak conditions the set of singularities of the information matrix forms a set of measure zero. However, it transpired that his hypothesis was not true as was shown by a counter-example (see Klaassen and Lenstra 2003). Consequently, he sought to show that under a form of identifiability, which he called quasi-identifiability, the set of singularities of the information matrix is "sparse", although not of measure zero. Quasi-identifiability in the presence of covariates is concerned with the likelihood conditioned on the covariate values rather than with identifiability of the joint distribution of the covariates and the responses. Importantly, Asgharian showed that under conditions (C.1–C.3) below, that are possible to check, the multi-path change-point model is

quasi-identifiable. We provide more detail in the discussion below and, in particular, define exactly what we mean by "sparse."

Definition 5: A parametric family of conditional probability distributions $\mathcal{P} = \{\{P_{\boldsymbol{\Phi}}(\cdot|\boldsymbol{x}); \boldsymbol{\Phi} \in \boldsymbol{\Theta}\}; \boldsymbol{x} \in \mathcal{X}\}$, on the sample space $\mathcal{W} = \mathcal{X} \times \mathcal{Y}$, with density $f(\cdot|\boldsymbol{x}, \boldsymbol{\Phi})$ with respect to a σ-finite measure μ, is quasi-identifiable if for any $\boldsymbol{\Phi}, \boldsymbol{\Phi}^* \in \boldsymbol{\Theta}$, where $\boldsymbol{\Phi} \neq \boldsymbol{\Phi}^*$, there exists a set of $\boldsymbol{x}$s in $\mathcal{X}$, $X_{\boldsymbol{\Phi},\boldsymbol{\Phi}^*}$ with measure greater than zero, such that $f(\mathbf{y}|\boldsymbol{x}, \boldsymbol{\Phi}) \neq f(\mathbf{y}|\boldsymbol{x}, \boldsymbol{\Phi}^*)$ for any $\boldsymbol{x} \in X_{\boldsymbol{\Phi},\boldsymbol{\Phi}^*}$.

Asgharian proved that the set of singularities of the Fisher information matrix of a quasi-identifiable model is a nowhere dense set. This property of the set of singularities makes it sparse, meaning that its members are scattered throughout the parameter space rather than concentrated in any one area.

We recognize that establishing that the set of singularities of the Fisher information matrix forms a nowhere dense set, is not an entirely satisfying solution. Nevertheless, in our experience, problems of non-singularity have not arisen. Thus, it is reasonable to speculate that non-singularity of the information matrix in our multi-path change-point model is, indeed, rare.

2.5.1 Sufficient Conditions for Quasi-identifiability

Asgharian's conditions C.1 to C.3 can be used to establish quasi-identifiability for the multi-path change-point model.

C.1 Let $R_{\boldsymbol{x}}$ be the range of the $p \times 1$ covariates $\boldsymbol{x}$. Then it is required that $\{\mathbf{0}\} \subset R_{\boldsymbol{x}}$ and that $R_{\boldsymbol{x}}$ must contain at least one other vector.

C.2 The projection matrix M, induced by the design matrix X must be of full rank.

C.3 The conditional pre- and post-change-point observation distribution families $\{f(\cdot|\boldsymbol{x}, \boldsymbol{\beta}_i, \sigma_i^2) : \boldsymbol{x} \in \boldsymbol{x}\}$ for $i = 1, 2$, respectively, must be quasi-identifiable.

Since the change-point distribution proposed by Asgharian (2013) with $\Delta = (\rho_1, \rho_2, \ldots, \rho_{m-1}, \boldsymbol{\alpha}) = (\rho_1, \rho_2, \ldots, \rho_{m-1}, \alpha_1, \alpha_2, \ldots, \alpha_p)$ as the vector of parameters, has the form,

$$P(\tau = k|\boldsymbol{x}_i, \Delta) = \begin{cases} \dfrac{\rho_k \exp\{\boldsymbol{x}_i^\top \boldsymbol{\alpha}\}}{(1-\rho_k) + \rho_k \exp\{\boldsymbol{x}_i^\top \boldsymbol{\alpha}\}} & k = 1 \\ \dfrac{\rho_k \exp\{\boldsymbol{x}_i^\top \boldsymbol{\alpha}\}}{(1-\rho_k) + \rho_k \exp\{\boldsymbol{x}_i^\top \boldsymbol{\alpha}\}} \displaystyle\prod_{l=1}^{k-1} \left(\frac{(1-\rho_l)}{(1-\rho_l) + \rho_l \exp\{\boldsymbol{x}_i^\top \boldsymbol{\alpha}\}} \right) & k = 2, \ldots, m-1 \\ \displaystyle\prod_{l=1}^{m-1} \left(\frac{(1-\rho_l)}{(1-\rho_l) + \rho_l \exp\{\boldsymbol{x}_i^\top \boldsymbol{\alpha}\}} \right) & k = m, \end{cases} \tag{2.5.2}$$

it is similar to ours, given by (2.2.4). Because, $\rho_k = \dfrac{\exp\left(\sum_{l=1}^{k} \alpha_{0l}\right)}{1 + \exp\left(\sum_{l=1}^{k} \alpha_{0l}\right)}$, the baseline hazards ρ_k are one-to-one functions of the α_{0k}s for $k = 1, 2, \ldots, m-1$.

In this thesis, we assume that C.1-C.3 hold. We also assume the observation distributions before and after the change to be identifiable (in the joint distribution sense). This is a stronger assumption than the quasi-identifiability required by C.3. Therefore, Asgharian's proof of identifiability applies to our model immediately.

CHAPTER 3
Penalized Likelihood Approach

For statistical inference in many classes of models such as multi-path change-point models, a likelihood approach is suggested. However, maximum likelihood, as well as least squares estimators can have low bias but high variance, when the number of parameters is large, such as in a model with a large number of covariates. Since maximum likelihood methods estimate all model parameters, which could be initially large, it can be advantageous to use methods that while increasing the bias, decrease the variance sufficiently to reduce the mean square error (MSE) overall. Further, in models with a large number of covariates some of them may be collinear. Penalized likelihood methods are designed with the goal of reducing the number of covariates or shrinking their regression coefficients, thereby ameliorating the above problems.

Multi-path change-point models, with a large number of covariates (also called features, attributes, or predictors) relative to sample size, are particularly suitable for penalized likelihood variable selection methods.

For example, in the Alzheimer's disease data we analyze in Chapter 6, the sample size is small and the number of covariates initially believed to be associated with the outcome is large relative to the sample size. Unpenalized MLE would produce estimators with unacceptably high variance.

We give, here, a brief review of variable selection methods. We present two all-subset selection approaches, Akaike's Information Criterion (AIC) (Akaike 1973) and the Bayesian Information Criterion (BIC) (Schwartz 1978). The BIC is particularly important to us since, although we do not use the BIC to choose the best submodel directly, we use it to choose the regularization (tuning) parameter, an important step in variable selection methods. We also give the features of shrinkage methods, especially ridge regression while penalty functions will be presented in Section 2. Finally, we discuss a newer generation of penalty functions.

3.1 Classical Variable Selection

The classical approaches to creation of a parsimonious set of model covariates include forward inclusion, backward exclusion, step-wise selection, and information based all-subset selection criteria such as AIC and the BIC. These methods quickly become computationally expensive as the number of covariates increases. In this section, we discuss the features of these approaches.

3.1.1 Step-wise Variable Selection

Forward and backward variable selection methods depend on inclusion/exclusion criteria to select a parsimonious submodel. They are most common in linear regression models, although they have occasionally been applied in more complicated settings such as change-point models (see Pocok 1982 and Reed 1998). Their drawbacks are as follows:

1. If a predictor is chosen in forward selection or deleted in backward selection, its inclusion status will never change.

2. At each iteration, we must re-estimate the model parameters for each possible new set of covariates. This can be computationally very expensive, especially when the number of covariates is large.

Step-wise selection using a combination of forward and backward selection overcomes the first issue with these approaches. However, it worsens the second issue by making the procedure even more computationally expensive.

3.1.2 All-subset Selection Methods

Error-based all-subset selection methods, such as the C_p or Cross-Validation (CV) (Mosteller 1948), are variable selection methods which have previously been used in change-point problems; see Liang and Wong (2000) and Faragge and Simon (1996). Since we take a likelihood approach in this thesis, we do not discuss these methods here. Information based all-subset selection criteria, such as AIC or the BIC, have, respectively, penalized likelihoods of the form:

$$-2\log L(\hat{\mathbf{\Phi}}) + 2\sum_{j=1}^{\kappa} \mathbb{I}_{(\hat{\Phi}_j \neq 0)}, \tag{3.1.1}$$

and

$$-2\log L(\hat{\mathbf{\Phi}}) + 2\log(n)\sum_{j=1}^{\kappa} \mathbb{I}_{(\hat{\Phi}_j \neq 0)}, \tag{3.1.2}$$

where $\mathbb{I}_{(\hat{\Phi}_j \neq 0)}$ is equal to 1 when $\hat{\Phi}_j \neq 0$, and otherwise is zero, for $j = 1, 2, \ldots, \kappa$, and $\hat{\mathbf{\Phi}}$ is the estimator of $\mathbf{\Phi}$.

In (3.1.1) and (3.1.2), the term $\sum_{j=1}^{\kappa} \mathbb{I}_{(\hat{\Phi}_j \neq 0)}$ can be considered as the $\mathcal{L}_0$-norm of $\hat{\mathbf{\Phi}}$. These methods have several features:

1) All possible sub-models must be considered, forcing us to calculate the information criterion for each subset of parameters. With a large number of parameters, this procedure becomes computationally infeasible.
2) When the number of parameters increases, AIC may over-fit the model, relative to the BIC. Both criteria reward models with a good fit, and have penalties that increase with the number of parameters in the model. However, the BIC has a heavier penalty on the number of non-zero parameters in the model. Hence, the BIC would select a simpler model than AIC.
3) AIC is based on the Kullback-Leibler divergence. The BIC is, naturally, Bayesian.
4) There is no clear choice between AIC and the BIC for the purposes of model or variable selection. The BIC is a consistent selection criterion, meaning that when the sample size increases to ∞, the BIC selects the true model with probability converging to 1. In the same setting, AIC asymptotically selects a model with too many parameters. For small sample sizes, however, the BIC tends to choose a model which is too simple, due to its heavy penalty on the number of parameters.

Depending on the sample size, the complexity of the model, and the purpose of the modelling, one may choose any of the above approaches. However, in our change-point setting, it is not computationally practicable to use such methods.

3.2 Penalty Functions

There are several drawbacks to the classic variable selection approaches, and shrinkage methods do not perform variable selection. Step-wise selection and ridge

regression ignore stochastic errors inherited from the previous stages at each iteration. It is difficult to establish the asymptotic and theoretical properties of these methods. Even using subset selection methods jointly with step-wise selection does not resolve these issues.

In penalized likelihood methods that perform variable selection, some parameters estimates may be set to zero automatically while others less than a pre-specified threshold may be set to zero. In addition, if the penalty functions are chosen properly, depending on the application, it can be shown that the induced estimators possess important asymptotic properties such as consistency, sparsity, and asymptotic normality. These properties must be re-established if the application is not standard; our multi-path change-point scenario is one such instance. The penalized log-likelihood function is written standardly as:

$$\tilde{l}_n(\mathbf{\Phi}) = \log L_n(\mathbf{\Phi}) - \mathbb{P}_{\lambda_n}(\mathbf{\Phi}) = l_n(\mathbf{\Phi}) - \mathbb{P}_{\lambda_n}(\mathbf{\Phi}), \tag{3.2.1}$$

where $L_n(\mathbf{\Phi})$ is the likelihood function of the model, and $\mathbb{P}_{\lambda_n}(\mathbf{\Phi})$ is the penalty function on the parameters. Further, it is common to assume an additive form for the penalty function, $\mathbb{P}_{\lambda_n}(\mathbf{\Phi}) = \sum_{j=1}^{k} P_{\lambda_n}(\Phi_j)$. For the parameters we believe must stay in the model, such as intercepts in regression models, the penalty function is defined to be zero.

Below, we introduce penalty functions used in this thesis, including ridge regression (shrinkage method), the least absolute shrinkage and selection operator (LASSO, Tibshirani 1996; Chen et al. 1998), HARD thresholding (Fan and Li 2002), and the

smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001), and review some of their properties.

3.2.1 Shrinkage Methods

Shrinkage methods are another set of classic regularization methods. They have also been applied to change-point problems (see Hall and Simar 2002). Not all shrinkage methods result in variable selection, and they are primarily used to reduce variability and prediction error. The ridge regression penalty function is a continuous function of the parameters and introduces a penalty on the magnitude of the parameters through the L_2-norm of the vector of parameters. Ridge regression prevents singularity of the variance-covariance matrix of the regression coefficient estimators, which results from collinearity between covariates. In models such as logistic regression, it is common to use a ridge penalty function, whose effect is to shrink the estimated coefficients toward zero. However, we should note that ridge regression cannot be used directly for variable selection since it does not automatically set certain coefficients equal to zero. We will describe later how we use a ridge penalty function on the change-point distribution to reduce the size of the parameter estimates.

3.2.2 Regular Penalty Functions

Penalty functions can be considered from two different viewpoints. They can be regarded as imposing a constraint on the parameters. In this role, the tuning parameter is analogous to a Lagrange multiplier. Alternatively, viewed through a Bayesian lens, the introduction of a penalty function can be thought of as the

imposition of a prior distribution on the regression parameters. We take the former view in this thesis.

The following penalty functions are among the most commonly used in the current variable selection literature, and we shall emphasize these three:

- **LASSO penalty** (Tibshirani 1996): $P_{\lambda_n}(\theta) = n\lambda_n|\theta|$.
- **SCAD penalty** (Fan and Li 2001): $P_{\lambda_n}(0) = 0$ and
  $P'_{\lambda_n}(\theta) = n\ \text{sign}(\theta)\lambda_n\{I(|\theta| \leq \lambda_n) + \frac{(a\lambda_n - |\theta|)_+}{(a-1)\lambda_n} I(|\theta| > \lambda_n)\}$,
  where $(\cdot)_+ = \max(0, \cdot)$ under the restriction that $a > 2$. A common choice for a is 3.7 (see Fan and Li 2001).
- **HARD Thresholding penalty (HARD)** (Fan and Li 2002):
  $P_n(\theta) = n^2\lambda_n^2 - (|\theta| - n\lambda_n)^2 I_{(|\theta|<n\lambda_n)}$.

The choice of tuning parameter λ_n, the weight of the penalty function in the model, is critical. We use a BIC tuning parameter selector which identifies the true model consistently (for more discussion see Wang et al. 2007). We discuss our use of the BIC further in Chapter 4.

The LASSO penalty function is considered to be equivalent to soft-thresholding. Zou and Hastie (2005) introduced a linear combination of LASSO and ridge regression (soft and hard-thresholding) which they called an elastic net. The elastic net penalty function encourages a grouping effect. We use the elastic net to carry out estimation in our change-point model.

Theoretically, using the SCAD penalty function, it is better to choose (λ_n, a) over two-dimensional grids using the criterion we use to select the tuning parameter.

However, due to the computational expense of such a procedure, we use $a = 3.7$ as suggested by Fan and Li (2001).

In Figure 3–1, we provide depictions of the four penalty functions we used in this thesis, as functions of a single parameter, β, say. The SCAD and LASSO penalty functions behave similarly at zero and they are non-differential at this point. While the LASSO penalty function penalizes all parameters equally in the full model, the SCAD penalty function penalizes the parameters with small estimates with a heavier penalty, forcing them to be equal to zero. Its penalty on non-zero parameters (with large enough estimates) is defined to be constant. Hence, the SCAD penalty function results in asymptotically unbiased estimators while the LASSO penalty function does not. The HARD-thresholding penalty function behaves in a similar fashion to the SCAD penalty and induces asymptotic unbiasedness and sparsity.

Each of the different penalty functions may introduce some finite-sample bias into the estimated model. The LASSO penalty function leads to estimation bias in parameters with large estimators. The ridge estimator also has a large bias as the estimate's value becomes larger, and only induces an unbiased estimators for true zero parameters.

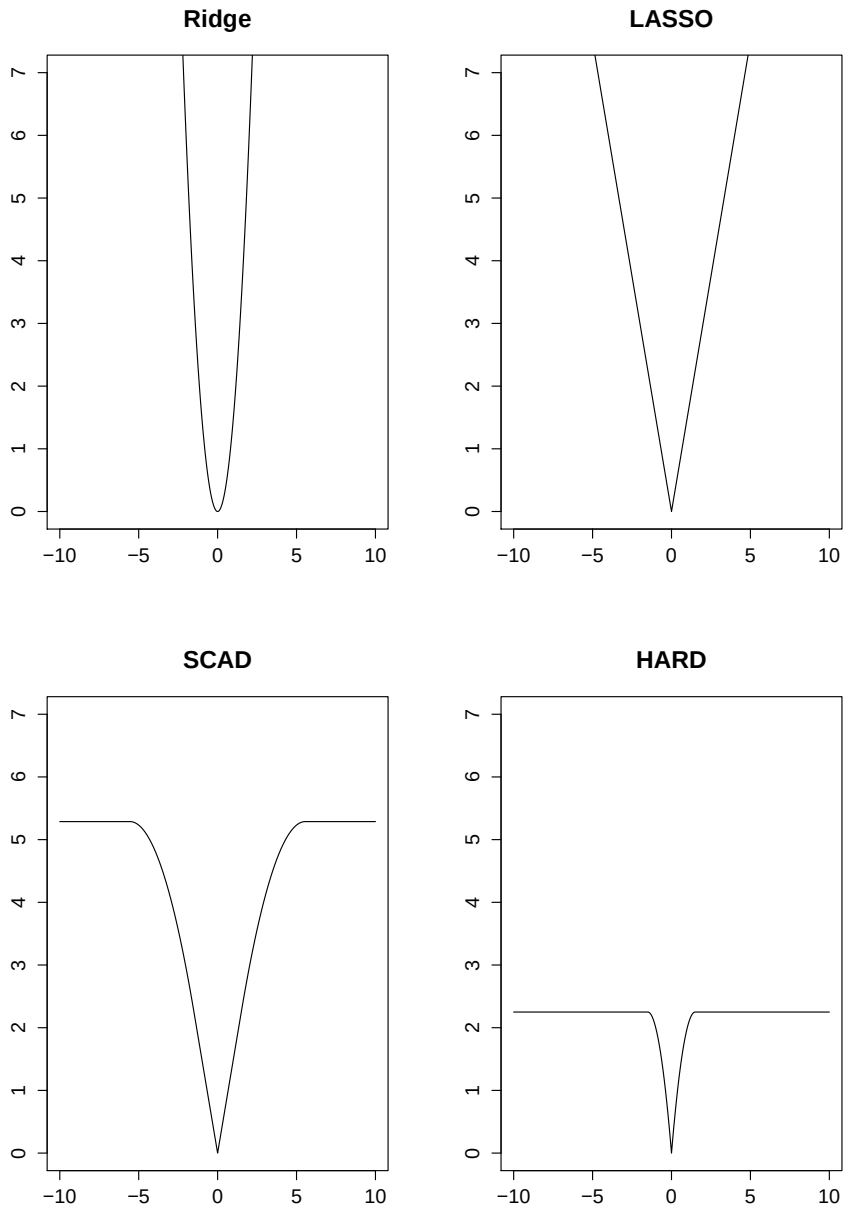


Figure 3–1: The penalty functions for $\lambda = 1.5$ (we set $a = 3.7$ in the SCAD penalty function).

3.2.3 Conditions on the penalty function

The following conditions are commonly imposed on the penalty function, $P_{\lambda_n}(\cdot)$, in a general setting:

P.1 For all $n = 1, 2, \ldots$ and $\lambda_n > 0$, $P_{\lambda_n}(0) = 0$ and $P_{\lambda_n}(\theta)$ is a symmetric, nonnegative, and nondecreasing function of θ and has a first derivative for all $\theta \in (0, \infty)$. The function is also continuously twice differentiable for all $\theta \in (c\lambda_n, \infty)$, and some constant $c > 0$.

P.2 Let $b_n = \max\left\{ \frac{P'_{\lambda_n}(\Phi_j)}{\sqrt{n}} : \Phi_j \neq 0 \text{ for } j = 1, 2, \ldots, \kappa \right\}$, be $O(1)$, and $c_n = \max\left\{ \frac{P''_{\lambda_n}(\Phi_j)}{n} : \Phi_j \neq 0, j = 1, 2, \ldots, \kappa \right\}$ be $o(1)$ as $n \to \infty$.

P.3 Let $\Gamma_n = (0, \frac{\log(n)}{\sqrt{n}})$. Then $\liminf_{n\to\infty} \frac{P'_{\lambda_n}(\theta_n)}{\sqrt{n}} = +\infty$, for all sequences $\{\theta_n\} \in \Gamma_n$ for every $n \geq$ some n_0, or equivalently,

$$\liminf_{n\to\infty} \left\{ \frac{P'_{\lambda_n}(\theta)}{\sqrt{n}} : 0 < \theta \leq n^{-1/2} \log n \right\} = +\infty.$$

These conditions are necessary in order to induce an estimator which is asymptotically unbiased, continuous and has the sparsity property. Conditions P.1–P.3 guarantee the "oracle property" for the estimated model; that is, penalty functions which satisfy conditions P.1–P.3 result in penalized likelihood estimators that are asymptotically identical to those obtained if we were to fit the true submodel.

3.2.4 Fused Penalty Functions and an Alternative approach

Variable selection can be used as a parameter space dimension reduction method in some circumstances. If there is a meaningful order in the features, we may use a fused penalty function in order to smooth the model. The fused penalty function encourages sparsity in the pair-wise differences between the adjacent coefficients,

which induces local or piece-wise constancy of the parameter profile (see Tibshirani et al. 2005). The more pair-wise adjacent parameter differences that are equal to zero, the smoother the model.

For example, in longitudinal studies, we may believe that the observation distribution has a piece-wise constant mean over time, or that in a survival analysis setting, the baseline hazard function is piece-wise constant. In both settings, the feature, time, is naturally ordered. In such cases, we may believe some adjacent parameters (which are indexed by time or location) are equal, leading to pair-wise differences that are zero.

For the change-point distribution, we proposed a proportional hazards model with a baseline hazard that changes from one observation time-point to the next. Hence, the number of baseline hazard parameters in the change-point distribution component of the model is one less than the number of follow-ups. The large number of parameters in the model makes estimation computationally expensive. However, if the follow-up times are not far apart, it may be reasonable to allow adjacent baseline hazards to be equal and therefore their differences to be zero. Although, it would be appropriate to use a fused penalty when estimating the baseline hazard, we took an alternative approach by re-parametrizing the logit of the baseline hazard. We defined the logit of the baseline hazard at each time-point to be the sum of the increments in the logits over the previous time-points.

The advantage of our re-parametrization is that: 1) It does not change the number of parameters in the baseline hazard. 2) It allows us to introduce regular penalty functions on the increments of the logit of the baseline hazard which results

in a smoother baseline hazard. 3) The computational cost of estimation in the new setting is less and more affordable than the original, and 4) to assure the asymptotic properties of the resulting estimators using a fused penalty function, we would need to introduce further constraints on the parameter space as well as the likelihood function while using this re-parametrization model we avoid these additional steps.

CHAPTER 4
Penalized Likelihood for Longitudinal Data with Change-points

To begin, we allow the observation distribution means before and after the change to depend on subject-specific covariates. The introduction of covariates into statistical models, such as those for multi-path change-point problems, results in a complex non-linear model with a large number of (initial) parameters. Our approach to variable selection and parameter estimation is through penalized likelihood. The multi-path change-point setting allows us to introduce time as a covariate in the observation distribution means, both before and after the change-point. Next, we allow the covariate effects to change with time, through the introduction of an interaction between time and covariates.

We also permit the change-point distribution to include covariate effects, although the lack of exchangeability of the probabilities of change between different time-points does not allow time to be used as a covariate in the change-point distribution. Further, we allow the change-point distribution to have a time-varying hazard, modelled through a baseline hazard that is piece-wise constant on the observation intervals, as proposed by Asgharian (2013). Naturally, the baseline hazard of a change may not necessarily differ from each observation interval to the next; for example, the baseline hazard may be constant through several adjacent intervals. From a practical viewpoint, this structure often reduces the computational burden.

To produce a model that allows for the same baseline hazards on successive time intervals, we re-parametrize the baseline hazards, ρ_k, in (2.5.2) of Chapter 2, as follows:

1) Let $\dfrac{\rho_k}{1-\rho_k} = \exp\{\rho_k^*\}$, for $k = 1, 2, \ldots, m-1$ (thus, ρ_k^* is the logit of the baseline hazard at the k^{th} time-point).

2) For $k = 1, 2, \ldots, m-1$, define $\alpha_{0k} = \rho_k^* - \rho_{k-1}^*$ to be the increment in the logit of the baseline hazard at the k^{th} time-point, or, equivalently,

$$\begin{aligned}
\rho_1^* &= \alpha_{01} \\
\rho_2^* &= \alpha_{02} + \alpha_{01} \\
&\vdots \\
\rho_{(m-1)}^* &= \alpha_{0(m-1)} + \alpha_{0(m-2)} + \cdots + \alpha_{01}.
\end{aligned}$$

This results in the model (2.2.4) defined in Chapter 2. Re-writing ρ_k^* in this form places it in a canonical penalized likelihood framework, wherein the belief that the baseline hazard may not change over adjacent intervals reduces to the belief that some of the α_{0k}s are equal to zero. An alternative approach using a fused penalty function was discussed in Chapter 3. Therefore, in addition to variable selection, we carry out smoothing using penalized likelihood methods.

Next, the regression model in the change-point distribution may result in a need to control the variance of the estimated regression coefficients, particularly in the presence of collinearity. We address this problem by using a ridge penalty function in this regression model. This controls the $\mathcal{L}_2$-norm of the vector of regression coefficients, using a tuning parameter γ_n. Although the ridge penalty function does not

set any of the estimates to zero, the convexity of $\mathcal{L}_2$-norm eases the computational difficulty of finding the maximum penalized likelihood estimates. In summary, we use penalized likelihood methods adapted to each component of our model.

4.1 Penalized likelihood estimation

Let $C_1 = C_2 = \{1, 2, \ldots, p_1\}$, $C_3 = \{1, 2, \ldots, p_2\}$, and $C_4 = \{1, 2, \ldots, m-1\}$ be the index sets of entries of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\alpha}$, and $\boldsymbol{\alpha}_0$, respectively. We allow p_1 and p_2 to be different, because in different problems we may believe initially different sets of covariates to be effective in the observation distributions and change-point distribution, respectively (for instance, we can allow time as a covariate in the observation distributions while we cannot include it as a covariate in the change-point distribution). The sets C_1, C_2, C_3, and C_4 are the index sets of entries of parameter vectors in the full model, which we believe is not necessarily the true model, but includes the true model. Therefore, a subset of them specifies the true underlying model of the data and we assume that there exist $S_i \subset C_i$ for $i = 1, 2, 3, 4$ which partitions C_i into the sets of true non-zero parameter indexes and true zero parameter indexes ($\beta_{ij}^0 \neq 0$ for $j \in S_i$ and $i = 1, 2$, $\alpha_j^0 \neq 0$ for $j \in S_3$, and $\alpha_{0k}^0 \neq 0$ for $k \in S_4$ where $\boldsymbol{\beta}_1^0$, $\boldsymbol{\beta}_2^0$, $\boldsymbol{\alpha}^0$, and $\boldsymbol{\alpha}_0^0$ are the true parameter vectors). We denote $\boldsymbol{\beta}_k[S_k]$ for $k = 1, 2$, $\boldsymbol{\alpha}[S_3]$ and $\boldsymbol{\alpha}_0[S_4]$ as subvectors of the parameter vectors in the full model, $\boldsymbol{\beta}_1[C_1]$, $\boldsymbol{\beta}_2[C_2]$, $\boldsymbol{\alpha}[C_3]$, and $\boldsymbol{\alpha}_0[C_4]$, respectively. Also, we let $\boldsymbol{x}[S_1]$, $\boldsymbol{x}[S_2]$, and $\boldsymbol{x}[S_3]$ represent the design matrices of dimension $|S_1| \times n$, $|S_2| \times n$, and $|S_3| \times n$ in the observation distribution before and after the change and in the change-point distribution, respectively. Hence, the true submodel observation distribution is $f(\boldsymbol{y}; \boldsymbol{x}[S_1], \boldsymbol{x}[S_2], \boldsymbol{x}[S_3], \boldsymbol{\beta}_1[S_1], \boldsymbol{\beta}_2[S_2], \boldsymbol{\alpha}[S_3], \boldsymbol{\alpha}_0[S_4])$. We call this model "sparse"

when the S_is are small subsets of C_is for $i = 1, 2, 3, 4$. We use maximum penalized likelihood to find the true submodel and estimate the model parameters simultaneously.

To find the penalized likelihood estimates, we maximize the penalized log-likelihood function,

$$\tilde{l}_n(\mathbf{\Phi}) = \log L_n(\mathbf{\Phi}) - \mathbb{P}_{\lambda_n}(\mathbf{\Phi}) = l_n(\mathbf{\Phi}) - \mathbb{P}_{\lambda_n}(\mathbf{\Phi}), \tag{4.1.1}$$

where $L_n(\mathbf{\Phi})$ is the likelihood function defined by (2.3.4) in Chapter 2, and the penalty $\mathbb{P}_{\lambda_n}(\mathbf{\Phi})$ is of the form,

$$\begin{aligned}\mathbb{P}_{\lambda_n}(\mathbf{\Phi}) = \sum_{k=2}^{m-1}\left\{P_{\lambda_n}(\alpha_{0k}) + \frac{\gamma_n}{2}\alpha_{0k}^2\right\} + \sum_{j=1}^{p_2}\left\{P_{\lambda_n}(\alpha_j) + \frac{\gamma_n}{2}\alpha_j^2\right\} \\ + \sum_{j=1}^{p_1} P_{\lambda_n}(\beta_{1j}) + \sum_{j=1}^{p_1} P_{\lambda_n}(\beta_{2j}).\end{aligned} \tag{4.1.2}$$

The penalty on the α_{0k}'s, the increments in the logit of the baseline hazard controls the smoothness of the baseline hazard of a change while the penalties on the α_j's, β_{1j}'s and β_{2j}'s control the number of covariates in the model. The ridge penalties $\sum_{k=2}^{m-1}\gamma_n\alpha_{0k}^2/2$ and $\sum_{j=1}^{p_2}\gamma_n\alpha_j^2/2$ are needed to prevent "wild" estimates of parameters α_{0k} and α_j when m is large and there are highly correlated covariates. Similar penalties are used by Park and Hastie (2008) and Bunea (2008) in logistic/multinomial regression. Furthermore, convexity of the ridge penalty is advantageous in numerical computations when dealing with a complex likelihood function such as (2.3.4) in Chapter 2.

For variable selection, we use three penalty functions: LASSO, SCAD, and HARD. We assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$ in our model. We do not penalize $\sigma_1^2, \sigma_2^2, \beta_{01}, \beta_{02}$ and α_{01}, since we want the intercepts in all three components (pre- and post-change-point observation distributions as well as the change-point distribution) to remain in the model. The above assumptions are used in the remainder of this thesis.

4.2 Numerical Computations

Since the penalized likelihood function has a complicated form, the maximum penalized likelihood estimators must be found numerically. We present an Expectation-Maximization (EM) algorithm (Dempster et al. 1977) with modified maximization.

4.2.1 Maximization Algorithm

To estimate the parameters, we use a modified EM algorithm. Let $(\boldsymbol{y}_i, \boldsymbol{x}_i)$ for $i = 1, 2, \ldots, n$ be a random sample of observations from the model (2.3.4) in Chapter 2. Let z_{ik} be equal to one if the change in the sequence of observation distributions for the i^{th} subject occurs at the k^{th} time-point, and otherwise equal to zero, for $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, m$. The z_{ik}s are not observable and can be considered as latent or missing random variables. The penalized complete

log-likelihood function, with the z_{ik}s as missing values, may be written as:

$$\begin{aligned}\tilde{l}_n^c(\boldsymbol{\Phi}) &= l_n^c(\boldsymbol{\Phi}) - \mathbb{P}_{\lambda_n}(\boldsymbol{\Phi}) \\ &= \sum_{i=1}^{n}\left\{\sum_{k=1}^{m} z_{ik}\left[\log(p(\tau_i = k|\boldsymbol{x}_i, \boldsymbol{\alpha}^*)) + \log(f_k(\boldsymbol{y}_i|\boldsymbol{x}_i, \Upsilon))\right]\right\} - \mathbb{P}_{\lambda_n}(\boldsymbol{\Phi}). \quad (4.2.1)\end{aligned}$$

Starting from an initial value $\boldsymbol{\Phi}^{(0)}$, the EM algorithm maximizes $\tilde{l}_n^c(\boldsymbol{\Phi})$ in two steps:

E-Step: Let $\boldsymbol{\Phi}^{(r)}$ be the estimate of the parameters after the r^{th} iteration. In the E-step we compute the conditional expectation of $\tilde{l}_n^c(\boldsymbol{\Phi})$ with the Z_{ik}s as random variables, given the data $(\boldsymbol{y}_i, \boldsymbol{x}_i)$, and assume that the values of the current estimate $\boldsymbol{\Phi}^{(r)} = (\Upsilon^{(r)}, \boldsymbol{\alpha}^{*(r)})$ are the true model parameters. The conditional expectation is

$$Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(r)}) = \sum_{i=1}^{n}\left\{\sum_{k=1}^{m} w_{ik}^{(r)}\left[\log(p(\tau_i = k|\boldsymbol{x}_i, \boldsymbol{\alpha}^*)) + \log(f_k(\boldsymbol{y}_i|\boldsymbol{x}_i, \Upsilon))\right]\right\} - \mathbb{P}_{\lambda_n}(\boldsymbol{\Phi}), \quad (4.2.2)$$

where the conditional expectation of the Z_{ik} for $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, m$ is:

$$w_{ik}^{(r)} = \frac{p(\tau_i = k|\boldsymbol{x}_i, \boldsymbol{\alpha}^{*(r)}) f_k(\boldsymbol{y}_i|\boldsymbol{x}_i, \Upsilon^{(r)})}{\sum_{l=1}^{m} p(\tau_i = l|\boldsymbol{x}_i, \boldsymbol{\alpha}^{*(r)}) f_l(\boldsymbol{y}_i|\boldsymbol{x}_i, \Upsilon^{(r)})}.$$

M-step: In the M-step, at the $(r+1)^{\text{th}}$ iteration we maximize (4.2.2) with respect to $\boldsymbol{\Phi}$. Owing to non-differentiability of the $P_{\lambda_n}(\theta)$ at $\theta = 0$, we cannot use the Newton–Raphson algorithm directly. We follow Fan and Li (2001)'s suggestion and replace $P_{\lambda_n}(\theta)$ by a local quadratic approximation in a neighbourhood of θ_0, where θ_0 is an initial value that is close to the MPLE,

$$\tilde{P}_{\lambda_n}(\theta) = P_{\lambda_n}(\theta_0) + \frac{P'_{\lambda_n}(\theta_0)}{2\theta_0}(\theta^2 - \theta_0^2).$$

Here $\tilde{P}_{\lambda_n}(\theta)$ is an increasing function of θ as $|\theta| \to \infty$, leading to a simpler M-step. To avoid numerical instability in the algorithm caused by very small estimated values of θ_0 in the denominator of the local quadratic approximation, we substitute θ_0 by $\theta_0 + \epsilon$, for a chosen small value $\epsilon > 0$, as suggested by Hunter and Li (2005). Let

$$\boldsymbol{\Phi}^{(r+1)} = \text{argmax}_{\boldsymbol{\Phi}} Q(\boldsymbol{\Phi}, \boldsymbol{\Phi}^{(r)}) =$$

$$\text{argmax}_{\boldsymbol{\alpha}^*, \boldsymbol{\Upsilon}} \left\{ \sum_{i=1}^{n} \left\{ \sum_{k=1}^{m} w_{ik}^{(r)} \left[\log(p(\tau_i = k | \boldsymbol{x}_i, \boldsymbol{\alpha}^*)) + \log(f_k(\boldsymbol{y}_i | \boldsymbol{x}_i, \Upsilon)) \right] \right\} - \tilde{\mathbb{P}}_{\lambda_n}(\boldsymbol{\Phi}) \right\}.$$

Starting from an initial value $\boldsymbol{\Phi}^{(0)}$, we iterate between the E- and M-steps until the Euclidean norm of two consecutive updates of the parameters, $\|\boldsymbol{\Phi}^{(r+1)} - \boldsymbol{\Phi}^{(r)}\|$, is smaller than some threshold value. We chose the threshold to be 10^{-6} in our simulations and data analysis.

When the EM algorithm converges, some of the estimates of the parameters will be "very small" (we took estimates less than 10^{-5} in our simulations and real data analysis as "very small"). We set these "very small" estimates equal to zero.

We use the maximum likelihood estimate of the vector of unknown parameters as an initial value ($\boldsymbol{\Phi}^{(0)}$), since our model is not excessively over-parametrized. One can also choose the best initial value by comparing the log-likelihood functions at possible candidates and choosing the one with the largest log-likelihood value. When

the algorithm converges, we use the derivative of the penalized likelihood function to find non-zero elements of $\hat{\mathbf{\Phi}}$ which must satisfy:

$$\frac{\partial}{\partial \Phi_j} l_n(\mathbf{\Phi})\big|_{\Phi_j=\hat{\Phi}_j} - P'_{\lambda_n}(\Phi_j)\big|_{\Phi_j=\hat{\Phi}_j} = 0, \tag{4.2.3}$$

since when using a suitable penalty function, the derivative of log-likelihood function as well as the penalty function must each be zero at large true parameters and, therefore, their appropriate estimates. Otherwise, the estimates will be automatically set to equal zero. This approach is useful for deciding the inclusion status of parameters in a real data set since we do not know in advance which parameter is non-zero. Since we use the Newton–Raphson algorithm, the choice of initial value is important (see Zou and Li 2008).

Recall that some penalty functions enable simultaneous selection and estimation, and therefore, we are able to estimate the variance-covariance matrix of the estimators. Conditional on the covariates, the estimated variance-covariance matrix for the estimators has the following form:

$$\begin{aligned}\widehat{Cov}(\hat{\mathbf{\Phi}}) = &\left[\frac{\partial^2}{\partial \mathbf{\Phi}\partial \mathbf{\Phi}^\top} l_n(\mathbf{\Phi}) + \frac{\partial^2}{\partial \mathbf{\Phi}\partial \mathbf{\Phi}^\top}\tilde{\mathbb{P}}_{\lambda_n}(\mathbf{\Phi})\right]^{-1}_{\mathbf{\Phi}=\hat{\mathbf{\Phi}}} \\ &\times \widehat{Cov}\left(\frac{\partial}{\partial \mathbf{\Phi}} l_n(\mathbf{\Phi})\big|_{\mathbf{\Phi}=\hat{\mathbf{\Phi}}}\right)\left[\frac{\partial^2}{\partial \mathbf{\Phi}\partial \mathbf{\Phi}^\top} l_n(\mathbf{\Phi}) + \frac{\partial^2}{\partial \mathbf{\Phi}\partial \mathbf{\Phi}^\top}\tilde{\mathbb{P}}_{\lambda_n}(\mathbf{\Phi})\right]^{-1}_{\mathbf{\Phi}=\hat{\mathbf{\Phi}}}\end{aligned}$$

In our simulations and real data analysis, we proposed normal distributions for the observation distributions before and after the change and introduced the covariates through linear regression models. In Model 2 to be defined in Chapter 5,

we denote $\boldsymbol{x}_i$, to be the vector of p-covariates for the subject i, and $\boldsymbol{x}_{it} = (1, t, \boldsymbol{x}_i)$ for $i = 1, 2, \ldots, n$ and $t = 1, 2, \ldots, m$. We also let $\boldsymbol{\beta}_1 = (\beta_{10}, \eta_1, \beta_{11}, \ldots, \beta_{1p_1})$ and $\boldsymbol{\beta}_2 = (\beta_{20}, \eta_2, \beta_{21}, \ldots, \beta_{2p_1})$. Therefore, the complete log-likelihood function in the EM algorithm has the following form:

$$
\begin{aligned}
l(\boldsymbol{\Phi}) = \sum_{i=1}^{n}\sum_{k=1}^{m} z_{ik}\Bigg[& \log(p(\tau_i = k|\boldsymbol{x}_i, \boldsymbol{\alpha}^*)) - \frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\Bigg(\sum_{l=1}^{k}(y_{il} - \boldsymbol{x}_{il}^\top\boldsymbol{\beta}_1)^2 \\
& + \sum_{l=k+1}^{m}(y_{il} - \boldsymbol{x}_{il}^\top\boldsymbol{\beta}_2)^2\Bigg)\Bigg] \\
= \sum_{i=1}^{n}\Bigg\{ & z_{i1}\log\Big(\frac{\exp(\alpha_{01} + \boldsymbol{x}_i^\top\boldsymbol{\alpha})}{1 + \exp(\alpha_{01} + \boldsymbol{x}_i^\top\boldsymbol{\alpha})}\Big) + \sum_{k=2}^{m-1} z_{ik}\Bigg[\log\Big(\frac{\exp(\sum_{l=1}^{k}\alpha_{0l} + \boldsymbol{x}_i^\top\boldsymbol{\alpha})}{1 + \exp(\sum_{l=1}^{k}\alpha_{0l} + \boldsymbol{x}_i^\top\boldsymbol{\alpha})}\Big) \\
& - \sum_{l^*=1}^{k-1}\log\big(1 + \exp(\sum_{l=1}^{l^*}\alpha_{0l} + \boldsymbol{x}_i^\top\boldsymbol{\alpha})\big)\Bigg] - z_{im}\Bigg(\sum_{l^*=1}^{m-1}\log\big(1 + \exp(\sum_{l=1}^{l^*}\alpha_{0l} + \boldsymbol{x}_i^\top\boldsymbol{\alpha})\big)\Bigg) \\
& + \sum_{i=1}^{n}\sum_{k=1}^{m} z_{ik}\Bigg[-\frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\Big(\sum_{l=1}^{k}(y_{il} - \boldsymbol{x}_{il}^\top\boldsymbol{\beta}_1)^2 + \sum_{l=k+1}^{m}(y_{il} - \boldsymbol{x}_{il}^\top\boldsymbol{\beta}_2)^2\Big)\Bigg]\Bigg\}
\end{aligned}
$$

Given the estimate at the r^{th} iteration, $\boldsymbol{\Phi}^{(r)} = (\Upsilon^{(r)}, \boldsymbol{\alpha}^{*(r)}) = (\boldsymbol{\beta}_1^{(r)}, \boldsymbol{\beta}_2^{(r)}, \alpha_{01}^{(r)}, \alpha_{02}^{(r)}, \ldots, \alpha_{0(m-1)}^{(r)}, \boldsymbol{\alpha}^{(r)})$, at the $(r+1)^{\text{th}}$ iteration:

- The estimates of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, the vectors of regression coefficients before and after the change, have closed forms, so that $\boldsymbol{\beta}_1^{(r+1)}$ is the solution of the equation:

$$\frac{\partial \tilde{l}(\boldsymbol{\Phi})}{\partial \boldsymbol{\beta}_1} = \sum_{i=1}^{n}\sum_{k=1}^{m} w_{ik}^{(r)} \left(\frac{1}{\sigma^2} \sum_{l=1}^{k} \boldsymbol{x}_{il}^{\top}(y_{il} - \boldsymbol{x}_{il}^{\top}\boldsymbol{\beta}_1) \right) + \frac{\mathbb{P}'_{\lambda_n}(\boldsymbol{\beta}_1^{(r)})}{\boldsymbol{\beta}_1^{(r)} + \epsilon}.\boldsymbol{\beta}_1 = \mathbf{0}$$

Therefore,

$$\boldsymbol{\beta}_1^{(r+1)} = \left[\sum_{i=1}^{n}\sum_{k=1}^{m} w_{ik}^{(r)} \chi_{ik} \ \mathrm{diag}\left(0, 0, \frac{P'_{\lambda_n}(\beta_{11}^{(r)})}{\beta_{11}^{(r)} + \epsilon}, \frac{P'_{\lambda_n}(\beta_{12}^{(r)})}{\beta_{12}^{(r)} + \epsilon}, \ldots, \frac{P'_{\lambda_n}(\beta_{1p}^{(r)})}{\beta_{1p}^{(r)} + \epsilon} \right) \right]^{-1} \times \left(\sum_{i=1}^{n}\sum_{k=1}^{m} w_{ik}^{(r)} \sum_{l=1}^{k} \boldsymbol{x}_{il}^{\top} y_{il} \right)$$

where $\chi_{ik} = \sum_{l=1}^{k} \boldsymbol{x}_{il}^{\top}\boldsymbol{x}_{il}$. Note that the first two elements of $\boldsymbol{\beta}_1$ represent the intercept and the time coefficient in the regression model, which are not penalized. Similarly, we have:

$$\hat{\boldsymbol{\beta}}_2^{(r+1)} = \left[\sum_{i=1}^{n}\sum_{k=1}^{m} w_{ik}^{(r)} \chi_{ik}^{*} \mathrm{diag}\left(0, 0, \frac{P'_{\lambda_n}(\hat{\beta}_{21}^{(r)})}{\hat{\beta}_{21}^{(r)} + \epsilon}, \frac{P'_{\lambda_n}(\hat{\beta}_{22}^{(r)})}{\hat{\beta}_{22}^{(r)} + \epsilon}, \ldots, \frac{P'_{\lambda_n}(\hat{\beta}_{2p}^{(r)})}{\hat{\beta}_{2p}^{(r)} + \epsilon} \right) \right]^{-1} \times \left(\sum_{i=1}^{n}\sum_{k=1}^{m} w_{ik}^{(r)} \sum_{l=k+1}^{m} \boldsymbol{x}_{il}^{\top} y_{il} \right)$$

where $\chi_{ik}^{*} = \sum_{l=k+1}^{m} \boldsymbol{x}_{il}^{\top}\boldsymbol{x}_{il}$.

- To estimate the change-point distribution parameters, $\boldsymbol{\alpha}^{*}$, since the estimates $\boldsymbol{\alpha}^{*(r+1)}$, do not have a closed form, we use a Newton–Raphson method nested in the M-step of the EM algorithm. Let $\boldsymbol{\alpha}^{*(r)}$, be the estimate of $\boldsymbol{\alpha}^{*}$ at the r^{th} iteration of EM algorithm and we choose it as the initial value for the Newton–Raphson method.

For $\boldsymbol{\alpha}$, the vector of regression coefficients in the change-point distribution, if $\boldsymbol{\alpha}_{(s)}^{(r+1)}$ is the estimate at the s^{th} iteration of Newton–Raphson algorithm, the estimates at the following iteration have the following form:

$$\boldsymbol{\alpha}_{(s+1)}^{(r+1)} = \boldsymbol{\alpha}_{(s)}^{(r+1)} - \left[\frac{\partial^2 \tilde{l}_n^c(\boldsymbol{\Phi})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^\top}\right]^{-1}_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_{(s)}^{(r+1)}} \left[\frac{\partial \tilde{l}_n^c(\boldsymbol{\Phi})}{\partial \boldsymbol{\alpha}}\right]_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_{(s)}^{(r+1)}}$$

and similarly for α_{0k}, the increment in the logit of the baseline hazard, if $\alpha_{0k}^{(r)}$ is the estimate at the r^{th} iteration of the EM-algorithm. We take this to be the initial value for the nested Newton–Raphson algorithm ($\alpha_{0k(0)}^{(r+1)} = \alpha_{0k}^{(r)}$). Then, given $\alpha_{0k(s)}^{(r+1)}$ be the estimate at the s^{th} iteration, the next iteration of the Newton–Raphson algorithm has the following form:

$$\alpha_{0k(s+1)}^{(r+1)} = \alpha_{0k(s)}^{(r+1)} - \left[\frac{\partial^2 \tilde{l}_n^c(\boldsymbol{\Phi})}{\partial \alpha_{0k}^2}\right]^{-1}_{\alpha=\alpha_{0k(s)}^{(r+1)}} \left[\frac{\partial \tilde{l}_n^c(\boldsymbol{\Phi})}{\partial \alpha_{0k}}\right]_{\alpha_{0k}=\boldsymbol{\alpha}_{0k(s)}^{(r+1)}}.$$

We use a similar stopping rule as that for the EM algorithm, and stop the Newton–Raphson algorithm when the Euclidean distance between the estimates ($\boldsymbol{\alpha}^{*(r+1)}$), between two successive iterations, is less than 10^{-6}.

- Estimation of the variance (σ^2) proceeds as follows:

$$\sigma^{2(r+1)} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{m} w_{ik}^{(r)} \big(\sum_{l=1}^{k}(y_{il} - \boldsymbol{x}_{il}^\top \boldsymbol{\beta}_1^{(r)})^2 + \sum_{l=k+1}^{m}(y_{il} - \boldsymbol{x}_{il}^\top \boldsymbol{\beta}_2^{(r)})^2\big)}{m\sum_{i=1}^{n}\sum_{k=1}^{m} w_{ik}^{(r)}},$$

since the $w_{ik}^{(r)}$ is the expectation of z_{ik}, $\sum_{k=1}^{m} w_{ik}^{(r)} = 1$ and therefore,

$$\sigma^{2(r+1)} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{m} w_{ik}^{(r)} \big(\sum_{l=1}^{k}(y_{il} - \boldsymbol{x}_{il}^{\top}\boldsymbol{\beta}_1^{(r)})^2 + \sum_{l=k+1}^{m}(y_{il} - \boldsymbol{x}_{il}^{\top}\boldsymbol{\beta}_2^{(r)})^2\big)}{mn}.$$

We iterate the EM algorithm until $\|\boldsymbol{\Phi}^{(r+1)} - \boldsymbol{\Phi}^{(r)}\| < 10^{-6}$.

4.2.2 Choice of the Tuning Parameters

Fan and Li (2001) and Khalili and Chen (2007) used generalized cross-validation (GCV) to choose the tuning parameters. However, Wang et al. (2007) showed that using the GCV to choose the tuning parameter leads one to over-fit the final selected model. They used the BIC to choose the tuning parameter, and showed that this method is consistent when selecting the true sparse model. In applications of the proposed method, one needs to choose appropriate values of the tuning parameters (γ, λ). We suggest a Bayesian information criterion (BIC) with a grid search scheme as follows.

Consider the grid of values $\{0.0, 0.1, 0.2, 0.3, 1.0, 1.5, 5, 10\}$, scaled by $\log n$ to satisfy our asymptotic condition, for γ. Further, let $\{0.01, 0.02, \ldots, 0.40\}$ be a grid for λ. For a given pair (γ, λ) from the above grid, we obtain the MPLE $\widehat{\boldsymbol{\Phi}}_n$ using the modified EM algorithm presented above. The BIC is computed as

$$\mathrm{BIC}(\gamma, \lambda) = -2l_n(\widehat{\boldsymbol{\Phi}}_n) + \mathrm{DF}(\gamma, \lambda)\log n$$

where $\mathrm{DF}(\gamma, \lambda)$, referred to as the degrees of freedom, is the total number of non-zero elements of the parameter-vector estimates $(\widehat{\Upsilon}, \widehat{\boldsymbol{\alpha}}^*)$. This criterion mimics the one used in linear regression by Wang et al. (2007).

To start the iterative procedure we chose the maximum likelihood estimates of the parameters in the full model as initial values. More precisely, to find these maximum likelihood estimates, we used several different initial values, and then chose the estimated parameter vector which maximized the likelihood overall. This approach is practical in analyzing real data, where we do not know the true submodel and parameters.

4.3 Asymptotic properties (Large sample behaviour)

In this section, we discuss the large sample behaviour of the MPLEs. Our main results are contained in Theorem 1 in which we establish the consistency, sparsity and asymptotic normality of our maximum penalized likelihood estimators.

We first introduce some notation. As usual, we assume that the true model underlying the data is the change-point model specified in Chapter 2 with the corresponding parameter vector $\boldsymbol{\Phi}_0$. We also assume that $\boldsymbol{\Phi}_0$ is an interior point of the parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^\kappa$, where $\kappa = 2(p_1 + 2) + p_2 + m - 1 + 2 = 2p_1 + p_2 + m + 5$, ($p_1 + 2$ is the number of parameters in the pre- and post-change observation distribution including time as a covariate in the regression models, p_2 is the number of parameters in the change-point distribution regression model, and $m - 1$ is the number of baseline hazard parameters, in addition to two observation distribution variances). We partition the true parameter vector as $\boldsymbol{\Phi}_0 = (\boldsymbol{\Phi}_{01}, \boldsymbol{\Phi}_{02})$ so that the sub-vectors $\boldsymbol{\Phi}_{01}$ and $\boldsymbol{\Phi}_{02}$ contain the true non-zero and zero parameters, respectively.

The true observation distribution variances $({\sigma^2}_1^0, {\sigma^2}_2^0)$ before and after the change, respectively, are included in $\boldsymbol{\Phi}_{01}$. A similar partitioning is considered for any candidate parameter vector $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2) \in \boldsymbol{\Theta}$.

4.3.1 Assumptions

In our asymptotic theory, some regularity conditions are required on the penalty function $P_{\lambda_n}(\theta)$ (given in Chapter 3), and also the joint probability density function $f(\boldsymbol{w}_i; \boldsymbol{\Phi}_0)$ of $\boldsymbol{W}_i = (\boldsymbol{x}_i, \boldsymbol{Y}_i)$. A detailed proof is given in this chapter. Note that $f(\boldsymbol{w}_i; \boldsymbol{\Phi}_0) = f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\Phi}_0) \times f_{\boldsymbol{x}}(\boldsymbol{x}_i)$, where $f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\Phi}_0)$ is specified in Chapter 2, and we assume that the marginal probability density function $f_{\boldsymbol{x}}(\boldsymbol{x}_i)$ of $\boldsymbol{x}_i$ does not depend on the parameter of interest $\boldsymbol{\Phi}_0$.

Assumptions on the parameter space as well as the distribution family are needed in order to establish the asymptotic properties of the estimators. In stating the regularity conditions, we write $\boldsymbol{\Phi} = (\Phi_1, \Phi_2, \ldots, \Phi_\kappa)$. We use Φ_j^0 to represent the entries of the true parameter vector $\boldsymbol{\Phi}_0 = (\boldsymbol{\Phi}_{01}, \boldsymbol{\Phi}_{02})$.

Also, denote the index sets $S_1 = \{1 \leq j \leq p_1 : \beta_{1j}^0 \neq 0\}$, $S_2 = \{1 \leq j \leq p_1 : \beta_{2j}^0 \neq 0\}$, $S_3 = \{1 \leq j \leq p_2 : \alpha_j^0 \neq 0\}$, $S_4 = \{2 \leq j \leq m-1 : \alpha_{0j}^0 \neq 0\}$, which identify the true non-zero parameters $(\beta_{1j}^0, \beta_{2j}^0, \alpha_j^0, \alpha_{0j}^0)$ as the entries of the sub-vector $\boldsymbol{\Phi}_{01}$.

D.1 The parameter space of the model, $\boldsymbol{\Theta} \subset \mathbb{R}^\kappa$, is a bounded open set.

D.2 The probability density function $f(\mathbf{w}; \boldsymbol{\Phi})$ is at least three times differentiable for each $\boldsymbol{\Phi} \in \boldsymbol{\Theta}$ and for ν-almost all $\mathbf{w} \in \mathcal{W}$.

D.3 Let $\boldsymbol{\Phi}_0 \in \boldsymbol{\Theta}$ be the unique true model parameter vector (guaranteed by the established model identifiability). There exists an open neighbourhood $N_{\boldsymbol{\Phi}_0}$ of $\boldsymbol{\Phi}_0$, and an integrable function $M(\mathbf{w})$ for almost all $\mathbf{w} \in \mathcal{W}$, such that for each

$\boldsymbol{\Phi} \in N_{\boldsymbol{\Phi}_0}$, and for all i, j and k,

$$\left|\frac{\partial \log f(\mathbf{w};\boldsymbol{\Phi})}{\partial \Phi_i}\right| \leq M(\mathbf{w}), \quad \left|\frac{\partial^2 \log f(\mathbf{w};\boldsymbol{\Phi})}{\partial \Phi_i \partial \Phi_j}\right| \leq M(\mathbf{w}), \text{and } \left|\frac{\partial^3 \log f(\mathbf{w};\boldsymbol{\Phi})}{\partial \Phi_i \partial \Phi_j \partial \Phi_k}\right| \leq M(\mathbf{w}).$$

D.4 For the Fisher information matrix, $\mathcal{I}(\boldsymbol{\Phi}_0)$, we have $\det(\mathcal{I}(\boldsymbol{\Phi}_0)) \neq 0$.

We assume Conditions D.1-D.4 hold in our model of Chapters 5 and 6.

4.3.2 Theorem

These properties in turn, ensure that our estimators have the so-called "oracle property"; this means that they are as asymptotically good as the maximum likelihood estimators that would have been obtained had we known the true submodel in advance. Further, they converge to the true parameter values at the same rate as do those based on the true submodel.

Theorem 1. *Let $\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_n$ be a random sample from a probability density function $f(\boldsymbol{w};\boldsymbol{\Phi}_0)$ that satisfies the conditions D.1–D.4. Suppose that the penalty function $P_{\lambda_n}(\theta)$ satisfies conditions P.1–P.3 in Chapter 3, and let the ridge tuning parameter γ_n in (4.1.2) be chosen such that $\frac{\gamma_n}{\sqrt{n}}$ converges to zero as n tends to ∞. Then, as $n \to \infty$,*

(a) (Consistency): There exists a local maximizer $\widehat{\boldsymbol{\Phi}}_n$ of the penalized log-likelihood function $\tilde{l}_n(\boldsymbol{\Phi})$ for which $\|\widehat{\boldsymbol{\Phi}}_n - \boldsymbol{\Phi}_0\| = O_p(r_n)$, where $r_n = n^{-1/2}(1 + b_n)$ and b_n is given in P.2 in Chapter 3.

(b) For any root-n consistent estimator $\widehat{\boldsymbol{\Phi}}_n = (\widehat{\boldsymbol{\Phi}}_{n1}, \widehat{\boldsymbol{\Phi}}_{n2})$ of $\boldsymbol{\Phi}_0$,

i. (Sparsity): $Pr(\widehat{\boldsymbol{\Phi}}_{n2} = \mathbf{0}) \to 1$.

ii. (Asymptotic Normality):

$$\sqrt{n}\left\{\left[\boldsymbol{I}_{11}(\boldsymbol{\Phi}_{01})+\frac{\mathbb{P}''_{\lambda_n}(\boldsymbol{\Phi}_{01})}{n}\right](\widehat{\boldsymbol{\Phi}}_{n1}-\boldsymbol{\Phi}_{01})+\frac{\mathbb{P}'_{\lambda_n}(\boldsymbol{\Phi}_{01})}{n}\right\}\xrightarrow{d}N(\mathbf{0},\boldsymbol{I}_{11}(\boldsymbol{\Phi}_{01}))$$

where $\boldsymbol{I}_{11}(\boldsymbol{\Phi}_{01})$ *is the Fisher information matrix when all the true zero parameters are removed from the model,* $\mathbb{P}'_{\lambda_n}(\boldsymbol{\Phi})=\dfrac{\partial\mathbb{P}_{\lambda_n}(\boldsymbol{\Phi})}{\partial\boldsymbol{\Phi}}$*, and* $\mathbb{P}''_{\lambda_n}(\boldsymbol{\Phi})=\dfrac{\partial^2\mathbb{P}_{\lambda_n}(\boldsymbol{\Phi})}{\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}^\top}$.

It should be noted that the rate of convergence in part (a) of Theorem 1 depends on b_n which, in turn, depends on the first derivative of the penalty function. The penalty function $P_{\lambda_n}(\theta)$ and the tuning parameter θ must be chosen such that $b_n = O(1)$ in order to attain the standard rate of convergence $n^{-1/2}$ for maximum likelihood estimation. For example, if we use the SCAD penalty function with λ_n tending to zero as n tends to ∞, then $b_n = 0$, while the desirable rate holds for the LASSO penalty function if $\lambda_n = O(n^{-1/2})$. Moreover, to achieve consistency in feature selection and smoothness of the estimator of the baseline hazard parameters, α_{0k}'s by Condition P.3 for $P_{\lambda_n}(\theta)$ and therefore, λ_n should be chosen using the SCAD penalty function such that $\sqrt{n}\lambda_n \to \infty$ as n tends to ∞. Both purposes, consistency in estimation and feature selection, are achievable using the SCAD penalty function, although we cannot fulfill both purposes using the LASSO penalty function. If we choose the estimators to be consistent in variable selection, the LASSO penalty function introduces a large bias into the $\boldsymbol{\Phi}_{01}$'s estimator ($\hat{\boldsymbol{\Phi}}_{1n}$).

4.3.3 Proof of Theorem 1.

Before proving part (a) of Theorem 1 we present the idea of its proof. Consider a very simple case, with just one parameter in the model. Let $\dot{\Phi}$ be the true parameter

and $\hat{\Phi}$ be the local maximizer of the likelihood function. For a given $\epsilon > 0$, we should choose H large enough such that a neighbourhood centred at the true parameter, $(\dot{\Phi} - r_n H, \dot{\Phi} + r_n H)$ captures the local maximizer of the penalized likelihood. Therefore, there exists a large enough constant $H > 0$, such that

$$P\left\{ \sup_{\|\mathbf{u}\|=H} \tilde{l}_n(\mathbf{\Phi}_0 + r_n \mathbf{u}) < \tilde{l}_n(\mathbf{\Phi}_0) \right\} \geq 1 - \epsilon. \tag{4.3.1}$$

Figure 4–1. and Figure 4–2. depict the likelihood function for a set of observation $y_1, y_2, \ldots, y_n$ for the true model and for two different choices of H.

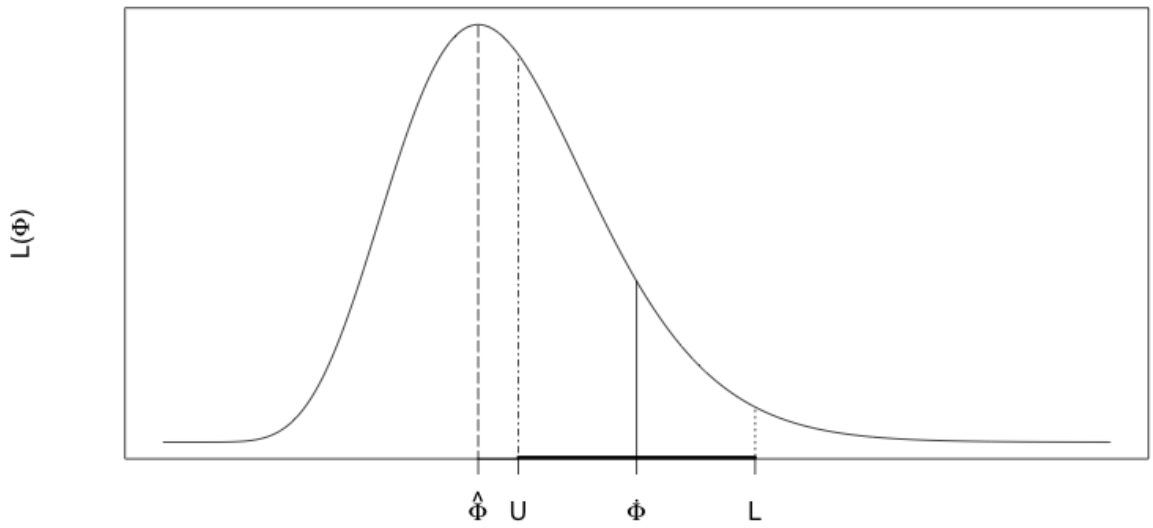


Figure 4–1: $\hat{\Phi}$ is the penalized maximum likelihood estimate, $\dot{\Phi}$ is the true parameter value, $U = \dot{\Phi} + r_n H_1$ and $L = \dot{\Phi} - r_n H_1$.

In Figure 4–1, the choice of positive constant H_1 for H is not large enough: $\sup_{|u|=H_1} L_n(\dot{\Phi} + r_n u) \geq L_n(\dot{\Phi})$ since it occurs at $u = H_1$ and the condition (4.3.1)

does not hold. In Figure 4–2, H_2 is chosen appropriately and large enough, and $\sup_{|u|=H_2} L_n(\dot{\Phi}+r_n u) < L_n(\dot{\Phi})$. Hence, the neighbourhood $(\dot{\Phi}-r_n H_2, \dot{\Phi}+r_n H_2)$ captures the local maximizer $\hat{\Phi}$.

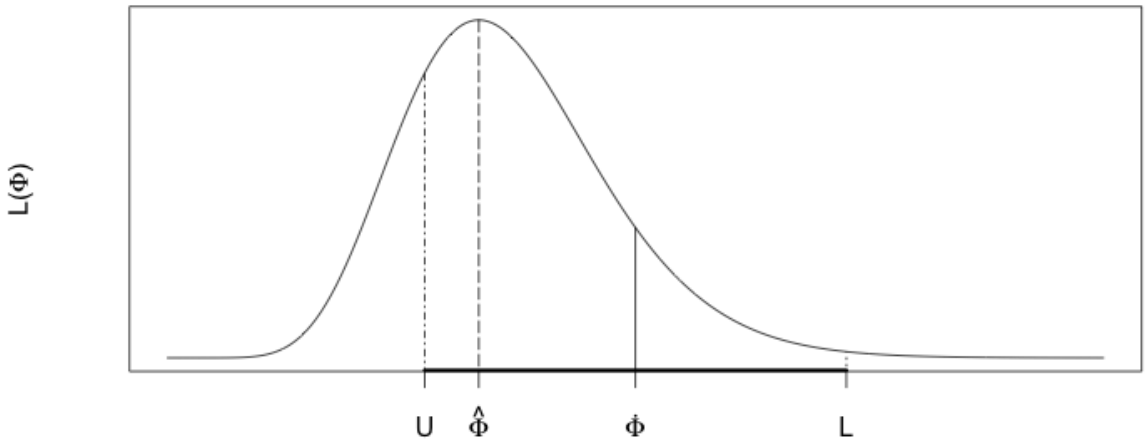


Figure 4–2: $\hat{\Phi}$ is the penalized maximum likelihood estimate, $\dot{\Phi}$ is the true parameter value, $U = \dot{\Phi} + r_n H_2$ and $L = \dot{\Phi} - r_n H_2$.

Both figures depict the likelihood function for a fixed sample of size n. As the sample size increases, the likelihood will be more concentrated around the maximizer. Since r_n converges to zero as n tends to ∞, the neighbourhoods shrink towards the true $\dot{\Phi}$, while still containing the local maximizer of the penalized likelihood function. Hence, we ensure that there exists a local maximizer in the neighbourhood, which is consistent for the true parameter.

(a) To prove consistency of the maximum penalized likelihood estimator $\hat{\mathbf{\Phi}}_n$, let $r_n = n^{-1/2}(1 + b_n)$. It is enough to show that for each $\epsilon > 0$ there exists a

constant $H > 0$, such that (4.3.1) holds. This ensures a sequence of neighbourhoods shrinking towards $\boldsymbol{\Phi}_0$ which are guaranteed to include the local maximizer, since with probability at least $(1-\epsilon)$ there exists a local maximizer, $\hat{\boldsymbol{\Phi}}_n$, in the set $\{\boldsymbol{\Phi}_0 + r_n\mathbf{u} : \|\mathbf{u}\| \leq H\}$, for which $\|\hat{\boldsymbol{\Phi}}_n - \boldsymbol{\Phi}_0\| = O_p(r_n)$.

Since $P_{\lambda_n}(0) = 0$ and the penalty function is nonnegative, we have:

$$\begin{aligned}
&\tilde{l}_n(\boldsymbol{\Phi}_0 + r_n\mathbf{u}) - \tilde{l}_n(\boldsymbol{\Phi}_0) \leq l_n(\boldsymbol{\Phi}_0 + r_n\mathbf{u}) - l_n(\boldsymbol{\Phi}_0) \\
&- \Bigg\{\sum_{j=1}^{\kappa_1}\Big(P_{\lambda_n}(\Phi_{0j} + r_n u_j) - P_{\lambda_n}(\Phi_{0j})\Big) + \gamma_n\Bigg[\sum_{\substack{j'=1 \\ j':\alpha^0_{j'}\neq 0}}^{p_2}((\alpha^0_{j'} + r_n u_{j'})^2 - {\alpha^0_{j'}}^2)\Bigg] \\
&+ \gamma_n\Bigg[\sum_{\substack{k'=2 \\ k':\alpha^0_{0k'}\neq 0}}^{m-1}((\alpha^0_{0k'} + r_n u_{k'})^2 - {\alpha^0_{0k'}}^2)\Bigg]\Bigg\}, \qquad (4.3.2)
\end{aligned}$$

where κ_1 is the number of components in $\boldsymbol{\Phi}_{01}$, the set of true nonzero parameters in the model and $u_{j'}$ and $u_{k'}$, for $j' = 1, 2, \ldots, p_2$ and $k' = 2, \ldots, m-1$ are elements of $\mathbf{u}$ corresponding to elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_0$, respectively. Using a Taylor expansion of the log-likelihood function, we have:

$$\begin{aligned}
&\left(\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\Phi}}\log f(\mathbf{W}_i;\boldsymbol{\Phi})\Big|_{\boldsymbol{\Phi}=\boldsymbol{\Phi}_0}\right)^{\top} r_n\mathbf{u} \\
&\qquad + \frac{1}{2}r_n^2\mathbf{u}^{\top}\left(\sum_{i=1}^{n}\frac{\partial^2}{\partial\boldsymbol{\Phi}\partial\boldsymbol{\Phi}^{\top}}\log f(\mathbf{W}_i;\boldsymbol{\Phi})\Big|_{\boldsymbol{\Phi}=\boldsymbol{\Phi}_0}\right)\mathbf{u} + R_n^L(\boldsymbol{\Phi}_0,\mathbf{u}), \qquad (4.3.3)
\end{aligned}$$

where $R_n^L(\boldsymbol{\Phi}_0, \mathbf{u})$ is the remainder term in the Taylor expansion, and the regularity Conditions D.1-D.3 on the likelihood function imply that $R_n^L(\boldsymbol{\Phi}_0, \mathbf{u}) = o_p(1)$. By $r_n = \frac{(1+b_n)}{\sqrt{n}}$ with b_n defined in Condition $P.2$, (4.3.3) has the following

form:

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\mathbf{\Phi}}\log f(\mathbf{W}_i;\mathbf{\Phi})\Big|_{\mathbf{\Phi}=\mathbf{\Phi}_0}\right)^{\top}(1+b_n)\mathbf{u}$$
$$+\frac{1}{2}(1+b_n)^2\mathbf{u}^{\top}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\mathbf{\Phi}\partial\mathbf{\Phi}^{\top}}\log f(\mathbf{W}_i;\mathbf{\Phi})\Big|_{\mathbf{\Phi}=\mathbf{\Phi}_0}\right)\mathbf{u}(1+o_p(1)). \quad (4.3.4)$$

Using a Taylor expansion for the penalty function, we have:

$$\sum_{j=1}^{\kappa_1}\left(P'_{\lambda_n}(\Phi_{0j})r_n\mathbf{u}_j+P''_{\lambda_n}(\Phi_{0j})r_n^2\mathbf{u}_j^2+R_n^{P_{\lambda_n}}(\mathbf{\Phi}_0,\mathbf{u})\right)$$
$$+\gamma_n r_n\Bigg[\sum_{\substack{j'=1\\ j':\alpha^0_{j'}\neq 0}}^{p_2}(2\alpha^0_{j'}u_{j'}+u_{j'}^2)+\sum_{\substack{k'=2\\ k':\alpha^0_{0k'}\neq 0}}^{m-1}(2\alpha^0_{0k'}u_{k'}+u_{k'}^2)\Bigg] \quad (4.3.5)$$

where $R_n^{P_{\lambda_n}}(\mathbf{\Phi}_0,\mathbf{u})$ is the remainder term in the Taylor expansion. $R_n^{P_{\lambda_n}}(\mathbf{\Phi}_0,\mathbf{u}) = o(1)$ by Conditions $P.1$-$P.2$ on the penalty function. Hence, since $r_n = \frac{(1+b_n)}{\sqrt{n}}$, (4.3.5) has form

$$\sum_{j=1}^{\kappa_1}\left(\frac{P'_{\lambda_n}(\Phi_{0j})}{\sqrt{n}}(1+b_n)\mathbf{u}_j+\frac{P''_{\lambda_n}(\Phi_{0j})}{n}(1+b_n)^2\mathbf{u}_j^2(1+o(1))\right)$$
$$+\frac{\gamma_n}{\sqrt{n}}(1+b_n)\Big(\sum_{\substack{j'=1\\ j':\alpha^0_{j'}\neq 0}}^{p_2}(2|\alpha^0_{j'}||u_{j'}|+|u_{j'}|^2)+\sum_{\substack{k'=2\\ k':\alpha^0_{0k'}\neq 0}}^{m-1}(2|\alpha^0_{0k'}||u_{k'}|+|u_{k'}|^2)\Big). \quad (4.3.6)$$

Using Condition $P.2$ we can bound (4.3.6) by:

$$\kappa_1\bigg(b_n(1+b_n)\|\mathbf{u}\| + \frac{c_n}{2}(1+b_n)^2\|\mathbf{u}\|^2(1+o(1))\bigg)$$
$$+\bigg(\frac{\gamma_n}{\sqrt{n}}(1+b_n)\|\mathbf{u}\|\big[\sum_{\substack{j'=1\\ j':\alpha^0_{j'}\neq 0}}^{p_2}(2\alpha^0_{j'}|+|u_{j'}|)+\sum_{\substack{k'=2\\ k':\alpha^0_{0k'}\neq 0}}^{m-1}(2|\alpha^0_{0k'}|+|u_{k'}|)\big]\bigg). \quad (4.3.7)$$

Using Condition $P.2$, $b_n(1+b_n)\|\mathbf{u}\| = O(1)$, $\frac{c_n}{2}(1+b_n)^2\|\mathbf{u}\|^2 = o(1)$, and the last term in brackets in (4.3.7) converges to zero as n tends to ∞. In addition, conditions D.1, D.2 and D.3, the regularity conditions on the likelihood function, imply $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\mathbf{\Phi}}\log f(\mathbf{W};\mathbf{\Phi})\Big|_{\mathbf{\Phi}=\mathbf{\Phi}_0} = O_p(1)$. The Law of Large Numbers induces convergence of $\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\mathbf{\Phi}\partial\mathbf{\Phi}^\top}\log f(\mathbf{W}_i;\mathbf{\Phi})\Big|_{\mathbf{\Phi}=\mathbf{\Phi}_0}$ to $-\mathcal{I}(\mathbf{\Phi}_0)$ as n tends to ∞. Condition D.4, the positive definiteness of the Fisher's information matrix alone implies that $\mathbf{u}^\top\mathcal{I}(\mathbf{\Phi}_0)\mathbf{u} \geq 0$. Hence as n goes to ∞, the dominant term in (4.3.2) is $-\frac{(1+b_n)^2}{2}\mathbf{u}^\top\mathcal{I}(\mathbf{\Phi}_0)\mathbf{u}$ by choosing H to be sufficiently large for a given $\epsilon > 0$. Hence, (4.3.1) holds.

To prove part (b) of the Theorem, we need the following lemma.

Lemma 1. *For $i = 1, 2, \ldots, n$, let $\mathbf{W}_i$ be independent and identically distributed random vectors with the probability density function $f(\cdot;\mathbf{\Phi}_0)$, where $\mathbf{\Phi}_0 = (\mathbf{\Phi}_{01}, \mathbf{0})$. Let $L_n(\mathbf{\Phi})$ represent a likelihood which is of the form of (2.3.4) in Chapter 2, and $\tilde{L}_n(\mathbf{\Phi})$ represent the penalized likelihood function of $\mathbf{\Phi} = (\mathbf{\Phi}_1, \mathbf{\Phi}_2)$. Assume D.1-D.3 are fulfilled, and the penalty function $P_{\lambda_n}(\cdot)$ satisfies Conditions P.1-P.3. Then, for*

any $\boldsymbol{\Phi} \in \boldsymbol{\Theta}$*, such that* $\|\boldsymbol{\Phi} - \boldsymbol{\Phi}_0\| = O(n^{-1/2})$*, we have:*

$$P\left\{\frac{\tilde{L}_n(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2)}{\tilde{L}_n(\boldsymbol{\Phi}_1, \mathbf{0})} \leq 1\right\} \to 1, \qquad n \to \infty.$$

Proof of Lemma 1.

Let $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2)$ be in the neighbourhood $\{\boldsymbol{\Phi} : \|\boldsymbol{\Phi} - \boldsymbol{\Phi}_0\| \leq Hn^{-1/2}\}$ for a positive constant H. Since $P_{\lambda_n}(0) = 0$, we have:

$$\log \tilde{L}_n(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2) - \log \tilde{L}_n(\boldsymbol{\Phi}_1, \mathbf{0}) = \log L_n(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2) - \log L_n(\boldsymbol{\Phi}_1, \mathbf{0}) - \sum_{j=\kappa_1+1}^{\kappa} P_{\lambda_n}(\Phi_j). \tag{4.3.8}$$

We show that (4.3.8) is negative as n tends to ∞. We first find the rate of convergence for the first two terms in the right-hand side of (4.3.8). Condition D.2 on the likelihood function allows us to invoke the mean value theorem, so that:

$$\log L_n(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2) - \log L_n(\boldsymbol{\Phi}_1, \mathbf{0}) = \left(\frac{\partial}{\partial \boldsymbol{\Phi}_2} \log L_n(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2)\big|_{\boldsymbol{\Phi}_2 = \xi}\right)^{\top} \boldsymbol{\Phi}_2, \tag{4.3.9}$$

where ξ is chosen, such that $\|\xi\| \leq \|\boldsymbol{\Phi}_2\|$, where $\|\boldsymbol{\Phi}_2\| = O(n^{-1/2})$. To find the rate of convergence of the first term on the right side of (4.3.9), $\frac{\partial}{\partial \boldsymbol{\Phi}_2} \log L_n(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2)\big|_{\boldsymbol{\Phi}_2 = \xi}$, we use condition D.3 on the derivatives of the log density functions, which implies

$\sum_{i=1}^{n} M(\mathbf{W}_i) = O(n)$. Also, using the mean value theorem, we have:

$$\begin{aligned}
&\left\| \frac{\partial}{\partial \mathbf{\Phi}_2} \log L_n(\mathbf{\Phi}_1, \mathbf{\Phi}_2)\Big|_{\mathbf{\Phi}_2=\xi} - \frac{\partial}{\partial \mathbf{\Phi}_2} \log L_n(\mathbf{\Phi}_{01}, \mathbf{\Phi}_2)\Big|_{\mathbf{\Phi}_2=\mathbf{0}} \right\| \\
&\leq \left\| \frac{\partial}{\partial \mathbf{\Phi}_2} \log L_n(\mathbf{\Phi}_1, \mathbf{\Phi}_2)\Big|_{\mathbf{\Phi}_2=\xi} - \frac{\partial}{\partial \mathbf{\Phi}_2} \log L_n(\mathbf{\Phi}_1, \mathbf{\Phi}_2)\Big|_{\mathbf{\Phi}_2=\mathbf{0}} \right\| \\
&+ \left\| \frac{\partial}{\partial \mathbf{\Phi}_2} \log L_n(\mathbf{\Phi}_1, \mathbf{\Phi}_2)\Big|_{\mathbf{\Phi}_2=\mathbf{0}} - \frac{\partial}{\partial \mathbf{\Phi}_2} \log L_n(\mathbf{\Phi}_{01}, \mathbf{\Phi}_2)\Big|_{\mathbf{\Phi}_2=\mathbf{0}} \right\| \\
&\leq \Big|\sum_{i=1}^{n} M(\mathbf{W}_i)\Big| \big(\|\xi\| + \|\mathbf{\Phi}_1 - \mathbf{\Phi}_{01}\| \big) = O(n^{1/2}). \qquad (4.3.10)
\end{aligned}$$

The conditions D.1-D.3 on the density function imply that $\frac{\partial}{\partial \mathbf{\Phi}_2} \log L_n(\mathbf{\Phi}_{01}, \mathbf{\Phi}_2)\Big|_{\mathbf{\Phi}_2=\mathbf{0}}$ is $O_p(n^{1/2})$, and therefore in (4.3.10), $\frac{\partial}{\partial \mathbf{\Phi}_2} \log L_n(\mathbf{\Phi}_1, \mathbf{\Phi}_2)\Big|_{\mathbf{\Phi}_2=\xi} = O_p(n^{1/2})$. Therefore, for (4.3.9) we have:

$$\log L_n(\mathbf{\Phi}_1, \mathbf{\Phi}_2) - \log L_n(\mathbf{\Phi}_1, \mathbf{0}) = O_p(n^{1/2}) \sum_{j=\kappa_1+1}^{\kappa} |\Phi_j|.$$

Consequently,

$$\begin{aligned}
&\log \tilde{L}_n(\mathbf{\Phi}_1, \mathbf{\Phi}_2) - \log \tilde{L}_n(\mathbf{\Phi}_1, \mathbf{0}) = O_p(n^{1/2}) \sum_{j=\kappa_1+1}^{\kappa} |\Phi_j| - \sum_{j=\kappa_1+1}^{\kappa} P_{\lambda_n}(\Phi_j) \\
&= \sqrt{n} \Bigg\{ \sum_{j=\kappa_1+1}^{\kappa} \left[|\Phi_j| O(1) - \frac{P_{\lambda_n}(\Phi_j)}{\sqrt{n}} \right] - \frac{\gamma_n}{2\sqrt{n}} \sum_{j:\alpha_j \neq 0} \alpha_j^2 \Bigg\}. \qquad (4.3.11)
\end{aligned}$$

Since $\frac{\gamma_n}{\sqrt{n}} = o(1)$, and by Condition P.3 on the penalty function for the Φ_j in a neighborhood shrinking to zero, the term $\left[|\Phi_j| O(1) - \frac{P_{\lambda_n}(\Phi_j)}{\sqrt{n}} \right]$, is negative with probability tending to 1 as $n \to \infty$, for each $j = \kappa_1 + 1, \kappa_1 + 2, \ldots, \kappa$. The proof is complete.

4.3.4 Continuation of the Proof of Theorem 1.

(b) i. For the partition $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2)$, let the $(\hat{\boldsymbol{\Phi}}_{n1}, \mathbf{0})$ be the maximizer of the log penalized likelihood $\tilde{l}_n(\boldsymbol{\Phi}_1, \mathbf{0}) = \log \tilde{L}_n(\boldsymbol{\Phi}_1, \mathbf{0})$, as a function of $\boldsymbol{\Phi}_1$. We only need to show that for any $\boldsymbol{\Phi}$, such that $\|\boldsymbol{\Phi} - \boldsymbol{\Phi}_0\| = O(n^{-1/2})$, we have:

$$P\left\{\log \tilde{L}_n(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2) - \log \tilde{L}_n(\hat{\boldsymbol{\Phi}}_{n1}, \mathbf{0}) < 0\right\} \to 1 \qquad \text{as} \quad n \to \infty.$$

As we showed in part (a) of Theorem 1, $\hat{\boldsymbol{\Phi}}_n = (\hat{\boldsymbol{\Phi}}_{n1}, \mathbf{0})$, the local maximizer, also has the property that $\|\hat{\boldsymbol{\Phi}}_n - \boldsymbol{\Phi}_0\| = O_p(n^{-1/2})$. We can write:

$$\begin{aligned}\log \tilde{L}_n(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2) - \log \tilde{L}_n(\hat{\boldsymbol{\Phi}}_{n1}, \mathbf{0}) = &\log \tilde{L}_n(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2) - \log \tilde{L}_n(\boldsymbol{\Phi}_1, \mathbf{0}) \\ &+ (\log \tilde{L}_n(\boldsymbol{\Phi}_1, \mathbf{0}) - \log \tilde{L}_n(\hat{\boldsymbol{\Phi}}_{n1}, \mathbf{0}))\end{aligned} \tag{4.3.12}$$

Lemma 1 forces $\log \tilde{L}_n(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2) - \log \tilde{L}_n(\boldsymbol{\Phi}_1, \mathbf{0})$ to be negative with probability converging to 1 as n tends to ∞, and by definition of $\hat{\boldsymbol{\Phi}}_n$, we also have $\log \tilde{L}_n(\boldsymbol{\Phi}_1, \mathbf{0}) - \log \tilde{L}_n(\hat{\boldsymbol{\Phi}}_{n1}, \mathbf{0}) \leq 0$. Hence, (4.3.12) is negative with probability tending to 1 as n tends to ∞ and the proof is complete.

ii. Considering $\log \tilde{L}_n(\boldsymbol{\Phi}_1, \mathbf{0})$ as a function of $\boldsymbol{\Phi}_1$, there exists a consistent local maximizer $\hat{\boldsymbol{\Phi}}_{n1}$ according to the part (a) of Theorem 1, for which the derivative of penalized likelihood function is zero,

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\Phi}_1} \log \tilde{L}_n(\boldsymbol{\Phi}_1, \mathbf{0})\Big|_{\boldsymbol{\Phi}_1 = \hat{\boldsymbol{\Phi}}_{n1}} = \\ \left[\frac{\partial}{\partial \boldsymbol{\Phi}_1} \log L_n(\boldsymbol{\Phi}_1, \mathbf{0}) - \frac{\partial}{\partial \boldsymbol{\Phi}_1} \mathbb{P}_{\lambda_n}(\boldsymbol{\Phi}_1, \mathbf{0})\right]_{\boldsymbol{\Phi}_1 = \hat{\boldsymbol{\Phi}}_{n1}} = 0.\end{aligned}$$

Using a Taylor expansion,

$$\begin{aligned}\left[\frac{\partial}{\partial \boldsymbol{\Phi}_1}\log L_n(\boldsymbol{\Phi}_1,\mathbf{0})-\frac{\partial}{\partial \boldsymbol{\Phi}_1}\mathbb{P}_{\lambda_n}(\boldsymbol{\Phi}_1,\mathbf{0})\right]_{\boldsymbol{\Phi}_1=\hat{\boldsymbol{\Phi}}_{n1}} &= \frac{\partial}{\partial \boldsymbol{\Phi}_1}\log L_n(\boldsymbol{\Phi}_1,\mathbf{0})\Big|_{\boldsymbol{\Phi}_1=\boldsymbol{\Phi}_{01}} \\ &+ \left[\frac{\partial^2}{\partial \boldsymbol{\Phi}_1\partial \boldsymbol{\Phi}_1^\top}\log L_n(\boldsymbol{\Phi}_1,\mathbf{0})\Big|_{\boldsymbol{\Phi}_1=\boldsymbol{\Phi}_{01}} + O_p(n)\right](\hat{\boldsymbol{\Phi}}_{n1}-\boldsymbol{\Phi}_{01}) \\ &- \left\{\frac{\partial}{\partial \boldsymbol{\Phi}_1}\mathbb{P}_{\lambda_n}(\boldsymbol{\Phi}_1,\mathbf{0})\Big|_{\boldsymbol{\Phi}_1=\boldsymbol{\Phi}_{01}} + \left[\frac{\partial^2}{\partial \boldsymbol{\Phi}_1\partial \boldsymbol{\Phi}_1^\top}\mathbb{P}_{\lambda_n}(\boldsymbol{\Phi}_1,\mathbf{0})\Big|_{\boldsymbol{\Phi}_1=\boldsymbol{\Phi}_{01}} + O_p(n)\right](\hat{\boldsymbol{\Phi}}_{n1}-\boldsymbol{\Phi}_{01})\right\} = 0.\end{aligned}$$

Therefore,

$$\begin{aligned}&\frac{\partial}{\partial \boldsymbol{\Phi}_1}\log L_n(\boldsymbol{\Phi}_1,\mathbf{0})\Big|_{\boldsymbol{\Phi}_1=\boldsymbol{\Phi}_{01}} - \frac{\partial}{\partial \boldsymbol{\Phi}_1}\mathbb{P}_{\lambda_n}(\boldsymbol{\Phi}_1,\mathbf{0})\Big|_{\boldsymbol{\Phi}_1=\boldsymbol{\Phi}_{01}} \\ &= \left\{-\frac{\partial^2}{\partial \boldsymbol{\Phi}_1\partial \boldsymbol{\Phi}_1^\top}\log L_n(\boldsymbol{\Phi}_1,\mathbf{0})\Big|_{\boldsymbol{\Phi}_1=\boldsymbol{\Phi}_{01}} + \frac{\partial^2}{\partial \boldsymbol{\Phi}_1\partial \boldsymbol{\Phi}_1^\top}\mathbb{P}_{\lambda_n}(\boldsymbol{\Phi}_1,\mathbf{0})\Big|_{\boldsymbol{\Phi}_1=\boldsymbol{\Phi}_{01}} + O_p(n)\right\}(\hat{\boldsymbol{\Phi}}_{n1}-\boldsymbol{\Phi}_{01}).\end{aligned}$$

Using a similar argument to that in the proof of part (a) of Theorem 1 the Law of Large Numbers implies that $-\frac{1}{n}\frac{\partial^2}{\partial \boldsymbol{\Phi}_1\partial \boldsymbol{\Phi}_1^\top}\log L_n(\boldsymbol{\Phi}_1,\mathbf{0})\Big|_{\boldsymbol{\Phi}_1=\boldsymbol{\Phi}_{01}}$ converges to $\mathcal{I}_{11}(\boldsymbol{\Phi}_{01})$ as n tends to ∞. Now, by the Central Limit Theorem, $\frac{\partial}{\partial \boldsymbol{\Phi}_1}\log L_n(\boldsymbol{\Phi}_1,\mathbf{0})\Big|_{\boldsymbol{\Phi}_1=\boldsymbol{\Phi}_{01}}$ converges in distribution to $\mathbf{Z}$, where $\mathbf{Z}\sim\mathcal{N}(\mathbf{0},\mathcal{I}_{11}(\boldsymbol{\Phi}_{01}))$ and so using Slutsky's theorem, we have:

$$\sqrt{n}\big(\mathcal{I}_{11}(\boldsymbol{\Phi}_{01})+\frac{\mathbb{P}''_{\lambda_n}(\boldsymbol{\Phi}_{01})}{n}\big)(\hat{\boldsymbol{\Phi}}_{n1}-\boldsymbol{\Phi}_{01})+\frac{\mathbb{P}'_{\lambda_n}(\boldsymbol{\Phi}_{01})}{\sqrt{n}}\xrightarrow{d}\mathbf{Z},$$

which completes the proof.

CHAPTER 5
Simulation Study

In this chapter we present the results of the simulations we carried out to investigate the performance of our variable selection method for longitudinal data with a change-point. We assess its small sample performance with two different multi-path change-point models. The models and parameters we chose were motivated by the Alzheimer's disease example analyzed in Chapter 6. The two models we considered are as follows:

Model 1. Starting with a simple model, we assumed constant (with respect to time) means for the observation distributions before and after the change, and a constant hazard for the change-point distribution. In this setting, we examined the effects of changing the sample size, number of follow-ups, number of binary covariates, and pair-wise correlation between covariates. The simplicity of Model 1 made it easy to check many different scenarios.

Model 2. Allowing for greater flexibility, we assumed a model with time as a covariate in the observation distribution means before and after the change, and a time-varying hazard for the change-point distribution. Although this scenario is of more practical use, its greater complexity results in considerably increased computation time, when carrying out the repetitions of a simulation study. In this model, we assessed the effects of sample size (particularly small samples), number of follow-ups, and several other features from Model 1.

In both models, we varied the magnitude of the parameters. In Setting 2 of Model 2, we assumed the baseline hazard model of a change to be sparse, in the sense that some elements of $\boldsymbol{\alpha}_0$ are zero. A sparse structure in this setting would result in a smoother baseline hazard. There are some assumptions common to both models. These common assumptions are:

1. In all the scenarios, we assumed the number of covariates to be the same ($p = 10$). We assumed there to be a combination of continuous and binary covariates, and varied the number of binary covariates.
2. We compared two different covariate profiles: no discrete covariates, and several binary covariates. In the latter case, the remaining covariates were taken to be continuous and were sampled from a multivariate normal distribution with mean zero and pair-wise correlations $\rho(\mathbf{X}_i, \mathbf{X}_j)$, for $i \neq j$, equal to 0, 0.5 or 0.75 in the first scenario, and 0 or 0.75 in the second scenario. Under Model 1, for $i = 1, 2, \ldots, n$, given the generated vector of covariates $\mathbf{x}_i$ and that the change occurs at k (for $k = 1, 2, \ldots, m-1$), we generated observations independently before and after the change-point from univariate normal distributions. That is, $Y_{ij} \sim N(\beta_{01} + \mathbf{x}_i'\boldsymbol{\beta}_1, \sigma_1^2)$ for $j = 1, 2, \ldots, k$, and $Y_{ij} \sim N(\beta_{02} + \mathbf{x}_i'\boldsymbol{\beta}_2, \sigma_2^2)$ for $j = k+1, k+2, \ldots, m$. If $k = m$, $Y_{ij} \sim N(\beta_{01} + \mathbf{x}_i'\boldsymbol{\beta}_1, \sigma_1^2)$ for $j = 1, 2, \ldots, m$. Under Model 2, given $\tau = k$, the means of the normal distributions were taken to be $\beta_{01} + \eta_1 j + \mathbf{x}_i'\boldsymbol{\beta}_1$ and $\beta_{02} + \eta_2 j + \mathbf{x}_i'\boldsymbol{\beta}_2$ before and after the change, respectively. We chose the variances σ_1^2 and σ_2^2 before and after the change-point to be equal to 1 in both models; these nuisance parameters were not our primary concern.

3. We assumed the change-point distribution to be (2.2.4), as described in Chapter 2.

To choose the tuning parameter, we used the BIC, setting the tuning parameter $\lambda_n \in \{0.01, 0.02, \ldots, 0.4\}$. We also used a modified form of the Newton–Raphson method in our simulations, as well as in the real data analysis in Chapter 6. In the standard Newton–Raphson method, the goal is to find the root of the equation $\dfrac{\partial f}{\partial \mathbf{y}} = 0$. Let $\mathbf{y}_n$ be the value at the n^{th} iteration. Then the update at the next iteration has the following form:

$$\mathbf{y}_{n+1} = \mathbf{y}_n - \left[\frac{\partial^2 f}{\partial \mathbf{y} \partial \mathbf{y}^\top}\right]^{-1}_{\mathbf{y}=\mathbf{y}_n} \left[\frac{\partial f}{\partial \mathbf{y}}\right]_{\mathbf{y}=\mathbf{y}_n}.$$

Instead, we let:

$$\mathbf{y}_{n+1} = \mathbf{y}_n - (0.5)^s \left[\frac{\partial^2 f}{\partial \mathbf{y} \partial \mathbf{y}^\top}\right]^{-1}_{\mathbf{y}=\mathbf{y}_n} \left[\frac{\partial f}{\partial \mathbf{y}}\right]_{\mathbf{y}=\mathbf{y}_n}$$

for $s = 1, 2, \ldots$. By introducing s, the increment at each iteration is reduced in size, guaranteeing that we do not miss the true $\mathbf{y}^*$ (the root of the equation). We chose $s = 2$. We also set all the estimates smaller than 10^{-6} equal to zero, which results in a sparse model.

In all our estimation procedures, we used the ridge regression penalty function for the change-point distribution regression model. We also used the ridge regression penalty function in the penalized likelihood estimation of the baseline hazard parameters in the change-point distribution.

We assessed performance in two different respects: variable selection and estimation. For variable selection, we used:

Sensitivity (S_1) = *Proportion of correctly estimated zero coefficients*, and

Specificity (S_2) = *Proportion of correctly estimated non-zero coefficients*.

We reported the averages of S_1 and S_2 over the 500 simulated data sets. We measured these two for each of the four vectors of parameters, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\alpha}^* = (\boldsymbol{\alpha}, \boldsymbol{\alpha}_0)$. The vector $\boldsymbol{\alpha}_0$ is not in Model 1.

Our simulations were driven by our Alzheimer's disease example in which the aim was to select the "best" set of explanatory variables rather than to carry out prediction. Therefore, we have not attempted to evaluate our approach with prediction in mind. Had this been one of our goals we would have simulated test data sets, separate from the training data sets, and based our evaluation on the sum of the squared differences between the observed and expected Ys. Alternatively, we could have used K-fold cross validation.

For estimation, we considered: 1) the errors in parameter estimation and 2) the average error. For the error in the parameter estimates, we computed the sums of the component-wise mean square errors for each of the vectors of estimates of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, and $\boldsymbol{\alpha}^* = (\alpha_{01}, \boldsymbol{\alpha})$, in Model 1 defined as

$$\widehat{\mathrm{MSE}}_1(\widehat{\boldsymbol{\beta}}_j) = \frac{1}{p+1}\sum_{i=0}^{p}[\widehat{\boldsymbol{\beta}}_{ij} - \boldsymbol{\beta}_{ij}^0]^2, \qquad j = 1, 2,$$

$$\widehat{\mathrm{MSE}}_1(\widehat{\boldsymbol{\alpha}}^*) = \frac{1}{p+1}\Big\{\sum_{i=1}^{p}[\widehat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i^0]^2 + [\widehat{\alpha}_{01} - \alpha_{01}^0]^2\Big\}, \tag{5.0.1}$$

and for $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\alpha}$ and $(\alpha_{01}, \alpha_{02}, \ldots, \alpha_{0(m-1)})$ in Model 2 as

$$\widehat{\mathrm{MSE}}_2(\widehat{\boldsymbol{\beta}}_j) = \frac{1}{p+2}\{\sum_{i=0}^{p}[\widehat{\boldsymbol{\beta}}_{ij} - \boldsymbol{\beta}_{ij}^0]^2 + [\hat{\beta}_{i\mathrm{Time}} - \beta_{i\mathrm{Time}}^0]^2\}, \qquad j = 1, 2,$$

$$\widehat{\mathrm{MSE}}_2(\widehat{\boldsymbol{\alpha}}) = \frac{1}{p}\left\{\sum_{i=1}^{p}[\widehat{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i^0]^2\right\},$$

$$\widehat{\mathrm{MSE}}_2(\widehat{\boldsymbol{\alpha}}_0) = \frac{1}{m-1}\left\{\sum_{k=1}^{m-1}[\hat{\alpha}_{0k} - \alpha_{0k}^0]^2\right\}, \tag{5.0.2}$$

where $\boldsymbol{\beta}_j^0$ for $j = 1, 2$ is the vector of true regression coefficients including true intercepts in the observation distributions before and after the change-point, $\boldsymbol{\alpha}^0$ is the vector of true regression coefficients in the change-point distribution and $(\alpha_{01}^0, \alpha_{02}^0, \ldots, \alpha_{0(m-1)}^0)$ is the vector of true increments in the logit of the baseline hazard. The $\widehat{\mathrm{MSE}}_k(\widehat{\boldsymbol{\beta}}_j)$s correspond to the before $(j = 1)$ and after $(j = 2)$ change-point observation distribution for $k = 1, 2$, corresponding Model 1 and 2, respectively and $\widehat{\mathrm{MSE}}(\widehat{\boldsymbol{\alpha}}^*)$ corresponds to the change-point distribution.

As an overall measure of the error in the estimated parameters we used the average mean square error of the estimated responses (TMSE) compared with the expected responses. The average was taken over the set of observed independent variables. The TMSEs in Model 1 were defined as

$$\begin{aligned}
\widehat{\mathrm{TMSE}}_1(\widehat{\boldsymbol{\beta}}_j) &= \frac{1}{n}\sum_{i=1}^{n}[\widehat{\beta}_{0j} + \boldsymbol{x}_i^\top\widehat{\boldsymbol{\beta}}_j - \beta_{0j}^0 - \boldsymbol{x}_i^\top\boldsymbol{\beta}_j^0]^2, \qquad j = 1, 2, \\
\widehat{\mathrm{TMSE}}_1(\widehat{\boldsymbol{\alpha}}^*) &= \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{m-1}[\pi_k(\boldsymbol{x}_j, \widehat{\boldsymbol{\alpha}}^*) - \pi_k(\boldsymbol{x}_j, \boldsymbol{\alpha}^{*0})]^2 \\
&= \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{m-1}[logit^{-1}(\hat{\alpha}_{01} + \mathbf{x}_i\widehat{\boldsymbol{\alpha}}) - logit^{-1}(\alpha_{01}^0 + \mathbf{x}_i\boldsymbol{\alpha}^0)]^2, \qquad (5.0.3)
\end{aligned}$$

where $\boldsymbol{x}_i$ is the vector of covariates for the i^{th} subject ($i = 1, 2, \ldots, n$). In Model 2. these were defined as

$$\widehat{\text{TMSE}}_2(\widehat{\boldsymbol{\beta}}_j) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{m}[\widehat{\beta}_{0j} + \boldsymbol{x}_{it}^{\top}\widehat{\boldsymbol{\beta}}_j - \beta_{0j}^0 - \boldsymbol{x}_{it}^{\top}\boldsymbol{\beta}_j^0]^2, j = 1, 2,$$

$$\widehat{\text{TMSE}}_2(\widehat{\boldsymbol{\alpha}}^*) = \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{m-1}[\pi_k(\boldsymbol{x}_j, \widehat{\boldsymbol{\alpha}}^*) - \pi_k(\boldsymbol{x}_j, \boldsymbol{\alpha}^{*0})]^2 =$$

$$\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{m-1}[logit^{-1}(\sum_{l=1}^{k}\hat{\alpha}_{0l} + \mathbf{x}_i\widehat{\boldsymbol{\alpha}}) - logit^{-1}(\sum_{l=1}^{k}\alpha_{0l}^0 + \mathbf{x}_i\boldsymbol{\alpha}^0)]^2, \quad (5.0.4)$$

where $\boldsymbol{x}_{it} = (t, \boldsymbol{x}_i)$ with $\boldsymbol{x}_i$, the vector of covariates for the i^{th} subject, for $t = 1, 2, \ldots, m$ and $i = 1, 2, \ldots, n$. For a vector of parameter estimators, say $\widehat{\boldsymbol{\theta}}$, we report the median of the relative efficiency of the parameter estimation methods as

$$e_j(\widehat{\boldsymbol{\theta}}) = \frac{\widehat{\text{MSE}}_j(\widehat{\boldsymbol{\theta}}_{\text{MPLE}})}{\widehat{\text{MSE}}_j(\widehat{\boldsymbol{\theta}}_{\text{Ridge}})} \quad (5.0.5)$$

and the empirical efficiency as

$$e_j^*(\widehat{\boldsymbol{\theta}}) = \frac{\widehat{\text{TMSE}}_j(\widehat{\boldsymbol{\theta}}_{\text{MPLE}})}{\widehat{\text{TMSE}}_j(\widehat{\boldsymbol{\theta}}_{\text{Ridge}})}, \quad (5.0.6)$$

based on 500 simulated data sets for $j = 1, 2$, respectively. Here $\widehat{\boldsymbol{\theta}}_{\text{Ridge}}$ is the vector of estimators of the parameters of the full model obtained by maximizing the penalized log-likelihood function $\tilde{l}_n(\cdot)$ with only a ridge penalty on the $\boldsymbol{\alpha}^*$.

5.1 Simulation Scenario: Model 1

Here we assessed two different settings. The main difference between these two was the magnitude of the true parameters. In addition, we assumed 1) $\alpha_{02} = \cdots = \alpha_{0(m-1)} = 0$ and 2) constant means before and after the change, respectively (time was not included as a covariate in the regression models before and after the change).

5.1.1 Setting 1

In this setting, we assessed the effect of the number of paths ($n = 50$ and 100) and the number of follow-ups ($m = 5$ and 15) on variable selection for the change-point distribution. We varied the number of independent binary covariates ($p_{bin} =$ 0, 2, and 4). We also altered the pair-wise correlations ($\rho = 0$, 0.5 and 0.75). Here, our primary concern was to detect the covariates which affected the change-point distribution. When inference is about the change-point distribution, a large number of subjects, n, is needed because the change-point is not directly observable and the information about its distribution is, consequently, indirect.

The parameter specification in this setting is as follows. The vectors of true parameters are:

$$\boldsymbol{\beta}_1^0 = (2, -1, 1, 3, 0, 0, 0, 0, 0, 0, 0, 0),$$

$$\boldsymbol{\beta}_2^0 = (-3, -4, -1, -2, 0, 0, 0, 0, 0, 0, 0, 0),$$

$$\boldsymbol{\alpha}^{*0}_{m=5} = (-2, 0, 0, 0, 2, -1, 0, 0, 0, 0, 0, 0, 0),$$

$$\boldsymbol{\alpha}^{*0}_{m=10} = (-2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, -1, 0, 0, 0, 0, 0, 0, 0),$$

$$\boldsymbol{\alpha}^{*0}_{m=15} = (-2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, -1, 0, 0, 0, 0, 0, 0, 0),$$

where the first element in each vector represents the intercept. We used the LASSO, SCAD and HARD penalty functions in addition to the ridge regression penalty function.

Tables 5–1 - 5–4 give the estimated sensitivities and specificities. In Tables 5–5 - 5–8, for any vector of parameters in the model, such as $\boldsymbol{\theta}$, we present simulation results for median$\left\{ e(\hat{\boldsymbol{\theta}}^j); j = 1, 2, \ldots, 500 \right\}$ where $e(\hat{\boldsymbol{\theta}}^j)$, defined in (5.0.5), and $\hat{\boldsymbol{\theta}}^j$ is the estimated vector $\boldsymbol{\theta}$ from the j^{th} simulated data set.

Table 5–1: Estimated sensitivity (S_1) and specificity (S_2) for cases with p_{con} covariates from $MN(\mathbf{0}, \Sigma_\rho)$ and p_{bin} binary covariates under Model 1, Setting 1 for $n = 50$ and $m = 5$.

	Setting					$\boldsymbol{\beta}_1$		$\boldsymbol{\beta}_2$		$\boldsymbol{\alpha}^*$	
Pen.	n	m	p_{con}	p_{bin}	ρ	S_1	S_2	S_1	S_2	S_1	S_2
LASSO	50	5	10	0	0	0.924	1.000	0.892	0.999	0.966	0.802
					.5	0.915	1.000	0.886	0.998	0.941	0.572
					0.75	0.892	1.000	0.839	0.969	0.890	0.407
			8	2	0	0.917	1.000	0.890	0.914	0.932	0.448
					.5	0.886	1.000	0.891	0.962	0.943	0.525
					0.75	0.855	1.000	0.842	0.956	0.920	0.436
			6	4	0	0.899	1.000	0.895	0.971	0.914	0.459
					.5	0.898	1.000	0.892	0.991	0.936	0.483
					0.75	0.895	1.000	0.894	0.940	0.913	0.463
SCAD	50	5	10	0	0	0.982	1.000	0.984	0.999	0.985	0.687
					.5	0.981	1.000	0.985	0.998	0.972	0.534
					0.75	0.983	1.000	0.957	0.933	0.926	0.559
			8	2	0	0.980	1.000	0.888	0.896	0.962	0.618
					.5	0.982	1.000	0.873	0.866	0.948	0.674
					0.75	0.982	1.000	0.895	0.858	0.910	0.597
			6	4	0	0.986	1.000	0.822	0.876	0.917	0.606
					.5	0.984	1.000	0.823	0.898	0.939	0.653
					0.75	0.990	1.000	0.708	0.799	0.905	0.564
HARD	50	5	10	0	0	0.992	1.000	0.893	0.999	0.638	0.984
					.5	0.982	1.000	0.823	0.998	0.602	0.944
					0.75	0.951	1.000	0.656	0.965	0.502	0.885
			8	2	0	0.986	1.000	0.459	0.951	0.601	0.887
					.5	0.985	1.000	0.440	0.942	0.512	0.912
					0.75	0.961	1.000	0.432	0.946	0.435	0.898
			6	4	0	0.973	1.000	0.379	0.947	0.488	0.885
					.5	0.982	1.000	0.414	0.949	0.501	0.914
					0.75	0.968	1.000	0.281	0.922	0.406	0.871

Table 5–2: Estimated sensitivity (S_1) and specificity (S_2) for cases with p_{con} covariates from $MN(\mathbf{0}, \Sigma_\rho)$ and p_{bin} binary covariates under Model 1, Setting 1 for $n = 100$ and $m = 5$.

	Setting					$\boldsymbol{\beta}_1$		$\boldsymbol{\beta}_2$		$\boldsymbol{\alpha}^*$	
Pen.	n	m	p_{con}	p_{bin}	ρ	S_1	S_2	S_1	S_2	S_1	S_2
LASSO	100	5	10	0	0	0.971	1.000	0.937	1.000	0.973	0.990
					.5	0.956	1.000	0.914	1.000	0.950	0.866
					0.75	0.917	1.000	0.877	1.000	0.986	0.680
			8	2	0	0.972	1.000	0.904	0.996	0.917	0.621
					.5	0.940	1.000	0.867	0.993	0.959	0.602
					0.75	0.914	1.000	0.864	0.996	0.898	0.632
			6	4	0	0.964	1.000	0.924	0.999	0.929	0.582
					.5	0.958	1.000	0.925	1.000	0.898	0.639
					0.75	0.940	1.000	0.882	0.994	0.906	0.592
SCAD	100	5	10	0	0	0.994	1.000	0.989	1.000	0.988	0.946
					.5	0.994	1.000	0.989	1.000	0.978	0.918
					0.75	0.996	1.000	0.992	0.999	0.976	0.721
			8	2	0	0.997	1.000	0.974	0.998	0.964	0.847
					.5	0.997	1.000	0.961	0.974	0.963	0.781
					0.75	0.997	1.000	0.970	0.978	0.941	0.756
			6	4	0	0.998	1.000	0.969	0.992	0.956	0.802
					.5	0.997	1.000	0.978	1.000	0.964	0.826
					0.75	0.998	1.000	0.944	0.942	0.919	0.751
HARD	100	5	10	0	0	0.997	1.000	0.980	1.000	0.868	1.000
					.5	0.998	1.000	0.969	1.000	0.820	0.997
					0.75	0.986	1.000	0.889	0.999	0.633	0.967
			8	2	0	0.998	1.000	0.888	0.999	0.816	0.958
					.5	0.996	1.000	0.734	0.987	0.657	0.946
					0.75	0.994	1.000	0.695	0.993	0.543	0.952
			6	4	0	0.996	1.000	0.759	0.996	0.683	0.959
					.5	0.996	1.000	0.852	1.000	0.719	0.967
					0.75	0.996	1.000	0.649	0.973	0.523	0.943

Table 5–3: Estimated sensitivity (S_1) and specificity (S_2) for cases with p_{con} covariates from $MN(\mathbf{0}, \Sigma_\rho)$ and p_{bin} binary covariates under Model 1, Setting 1 for $n = 50$ and $m = 15$.

	Setting					$\boldsymbol{\beta}_1$		$\boldsymbol{\beta}_2$		$\boldsymbol{\alpha}^*$	
Pen.	n	m	p_{con}	p_{bin}	ρ	S_1	S_2	S_1	S_2	S_1	S_2
LASSO	50	15	10	0	0	0.954	1.000	0.955	1.000	0.945	0.962
					.5	0.928	1.000	0.936	1.000	0.949	0.896
					0.75	0.930	1.000	0.915	1.000	0.960	0.664
			8	2	0	0.961	1.000	0.923	1.000	0.901	0.561
					.5	0.948	1.000	0.887	1.000	0.918	0.646
					0.75	0.910	1.000	0.830	1.000	0.918	0.683
			6	4	0	0.958	1.000	0.914	1.000	0.934	0.583
					.5	0.945	1.000	0.907	1.000	0.922	0.620
					0.75	0.931	1.000	0.846	1.000	0.918	0.648
SCAD	50	15	10	0	0	0.980	1.000	0.980	1.000	0.982	0.898
					.5	0.977	1.000	0.978	1.000	0.978	0.824
					0.75	0.979	1.000	0.978	1.000	0.961	0.700
			8	2	0	0.984	1.000	0.975	1.000	0.961	0.809
					.5	0.979	1.000	0.968	1.000	0.964	0.840
					0.75	0.980	1.000	0.977	1.000	0.920	0.779
			6	4	0	0.985	1.000	0.976	1.000	0.952	0.812
					.5	0.984	1.000	0.973	1.000	0.941	0.830
					0.75	0.983	1.000	0.977	1.000	0.918	0.795
HARD	50	15	10	0	0	0.988	1.000	0.988	1.000	0.777	0.996
					.5	0.983	1.000	0.988	1.000	0.699	0.994
					0.75	0.976	1.000	0.980	1.000	0.602	0.956
			8	2	0	0.990	1.000	0.961	1.000	0.764	0.944
					.5	0.993	1.000	0.959	1.000	0.688	0.965
					0.75	0.992	1.000	0.933	1.000	0.534	0.944
			6	4	0	0.993	1.000	0.957	1.000	0.712	0.963
					.5	0.993	1.000	0.956	1.000	0.652	0.962
					0.75	0.993	1.000	0.928	1.000	0.507	0.952

Table 5–4: Estimated sensitivity (S_1) and specificity (S_2) for cases with p_{con} covariates from $MN(\mathbf{0}, \Sigma_\rho)$ and p_{bin} binary covariates under Model 1, Setting 1 for $n = 100$ and $m = 15$.

	Setting					$\boldsymbol{\beta}_1$		$\boldsymbol{\beta}_2$		$\boldsymbol{\alpha}^*$	
Pen.	n	m	p_{con}	p_{bin}	ρ	S_1	S_2	S_1	S_2	S_1	S_2
LASSO	100	15	10	0	0	0.997	1.000	0.996	1.000	0.992	1.000
					.5	0.977	1.000	0.982	1.000	0.955	0.994
					0.75	0.950	1.000	0.954	1.000	0.885	0.914
			8	2	0	0.991	1.000	0.946	1.000	0.886	0.918
					.5	0.982	1.000	0.938	1.000	0.907	0.886
					0.75	0.963	1.000	0.921	1.000	0.873	0.882
			6	4	0	0.985	1.000	0.937	1.000	0.897	0.908
					.5	0.979	1.000	0.918	1.000	0.871	0.875
					0.75	0.969	1.000	0.910	1.000	0.924	0.838
SCAD	100	15	10	0	0	0.999	1.000	0.999	1.000	0.990	0.998
					.5	0.994	1.000	0.995	1.000	0.983	0.991
					0.75	0.992	1.000	0.992	1.000	0.964	0.955
			8	2	0	0.998	1.000	0.988	1.000	0.973	0.972
					.5	0.997	1.000	0.990	1.000	0.964	0.960
					0.75	0.996	1.000	0.992	1.000	0.957	0.944
			6	4	0	0.998	1.000	0.990	1.000	0.953	0.972
					.5	0.997	1.000	0.989	1.000	0.953	0.963
					0.75	0.997	1.000	0.988	1.000	0.939	0.955
HARD	100	15	10	0	0	0.998	1.000	0.998	1.000	0.828	1.000
					.5	0.996	1.000	0.996	1.000	0.839	1.000
					0.75	0.992	1.000	0.990	1.000	0.772	0.995
			8	2	0	0.996	1.000	0.974	1.000	0.852	0.992
					.5	0.997	1.000	0.978	1.000	0.820	0.992
					0.75	0.996	1.000	0.982	1.000	0.716	0.990
			6	4	0	0.996	1.000	0.969	1.000	0.815	0.992
					.5	0.997	1.000	0.980	1.000	0.792	0.991
					0.75	0.998	1.000	0.980	1.000	0.730	0.988

Table 5–5: Empirical estimation efficiency ($e(\hat{\boldsymbol{\theta}})$) for cases with p_{con} covariates from $MN(\mathbf{0}, \Sigma_\rho)$ and p_{bin} binary covariates under Model 1, Setting 1 for $n = 50$ and $m = 5$.

	Setting							
Pen.	n	m	p_{con}	p_{bin}	ρ	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\alpha}^*$
LASSO	50	5	10	0	0	0.561	0.299	0.840
					.5	0.472	0.233	1.143
					0.75	0.327	0.263	0.685
			8	2	0	0.661	0.102	0.991
					.5	0.588	0.079	0.604
					0.75	0.587	0.128	0.503
			6	4	0	0.583	0.062	0.533
					.5	0.745	0.042	0.402
					0.75	0.548	0.035	0.327
SCAD	50	5	10	0	0	0.344	0.130	0.811
					.5	0.224	0.083	1.160
					0.75	0.131	0.118	0.708
			8	2	0	0.384	0.139	0.786
					.5	0.297	0.183	0.572
					0.75	0.176	0.210	0.520
			6	4	0	0.294	0.220	0.531
					.5	0.285	0.152	0.347
					0.75	0.197	0.500	0.398
HARD	50	5	10	0	0	0.252	0.295	0.823
					.5	0.163	0.427	1.030
					0.75	0.162	0.760	0.742
			8	2	0	0.320	0.918	0.838
					.5	0.243	0.931	0.729
					0.75	0.193	0.903	0.651
			6	4	0	0.316	0.960	0.621
					.5	0.266	0.910	0.475
					0.75	0.241	0.991	0.505

Table 5–6: Empirical estimation efficiency ($e(\hat{\boldsymbol{\theta}})$) for cases with p_{con} covariates from $MN(\mathbf{0}, \Sigma_\rho)$ and p_{bin} binary covariates under Model 1, Setting 1 for $n = 100$ and $m = 5$.

	Setting							
Pen.	n	m	p_{con}	p_{bin}	ρ	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\alpha}^*$
LASSO	100	5	10	0	0	0.690	0.436	1.490
					.5	0.621	0.374	1.805
					0.75	0.454	0.322	1.297
			8	2	0	0.728	0.276	2.214
					.5	0.613	0.229	1.373
					0.75	0.540	0.236	0.884
			6	4	0	0.741	0.136	1.266
					.5	0.715	0.148	1.402
					0.75	0.587	0.159	0.799
SCAD	100	5	10	0	0	0.467	0.261	0.486
					.5	0.360	0.173	0.493
					0.75	0.131	0.091	1.149
			8	2	0	0.462	0.129	0.871
					.5	0.264	0.126	0.843
					0.75	0.176	0.106	0.679
			6	4	0	0.426	0.081	0.622
					.5	0.368	0.074	0.712
					0.75	0.244	0.117	0.587
HARD	100	5	10	0	0	0.270	0.199	0.703
					.5	0.196	0.173	0.829
					0.75	0.111	0.366	1.018
			8	2	0	0.312	0.349	0.852
					.5	0.192	0.592	0.916
					0.75	0.127	0.713	0.808
			6	4	0	0.354	0.532	0.677
					.5	0.303	0.422	0.725
					0.75	0.224	0.746	0.677

Table 5–7: Empirical estimation efficiency ($e(\hat{\boldsymbol{\theta}})$) for cases with p_{con} covariates from $MN(\mathbf{0}, \Sigma_\rho)$ and p_{bin} binary covariates under Model 1, Setting 1 for $n = 50$ and $m = 15$.

	Setting							
Pen.	n	m	p_{con}	p_{bin}	ρ	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\alpha}^*$
LASSO	50	15	10	0	0	0.510	0.276	0.803
					.5	0.391	0.236	0.650
					0.75	0.247	0.211	0.831
			8	2	0	0.540	0.186	1.932
					.5	0.508	0.112	0.888
					0.75	0.399	0.216	0.576
			6	4	0	0.533	0.118	0.879
					.5	0.498	0.101	0.662
					0.75	0.364	0.108	0.542
SCAD	50	15	10	0	0	0.391	0.213	0.515
					.5	0.246	0.165	0.550
					0.75	0.148	0.119	0.823
			8	2	0	0.376	0.114	0.917
					.5	0.349	0.062	0.471
					0.75	0.201	0.067	0.534
			6	4	0	0.383	0.063	0.441
					.5	0.335	0.057	0.347
					0.75	0.204	0.036	0.357
HARD	50	15	10	0	0	0.241	0.134	0.604
					.5	0.136	0.098	0.601
					0.75	0.098	0.075	0.801
			8	2	0	0.214	0.087	0.818
					.5	0.212	0.044	0.531
					0.75	0.106	0.094	0.682
			6	4	0	0.252	0.049	0.428
					.5	0.210	0.043	0.471
					0.75	0.126	0.041	0.503

Table 5–8: Empirical estimation efficiency ($e(\hat{\boldsymbol{\theta}})$) for cases with p_{con} covariates from $MN(\mathbf{0}, \Sigma_\rho)$ and p_{bin} binary covariates under Model 1, Setting 1 for $n = 100$ and $m = 15$.

	Setting							
Pen.	n	m	p_{con}	p_{bin}	ρ	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\alpha}^*$
LASSO	100	15	10	0	0	0.616	0.327	0.880
					.5	0.753	0.204	0.688
					0.75	0.722	0.517	1.376
			8	2	0	0.248	0.237	0.840
					.5	0.576	0.221	0.975
					0.75	0.456	0.187	0.718
			6	4	0	0.577	0.193	0.757
					.5	0.573	0.239	0.753
					0.75	0.462	0.205	0.683
SCAD	100	15	10	0	0	0.527	0.286	0.437
					.5	0.518	0.148	0.228
					0.75	0.404	0.285	0.358
			8	2	0	0.195	0.133	0.327
					.5	0.419	0.120	0.425
					0.75	0.245	0.083	0.246
			6	4	0	0.433	0.090	0.314
					.5	0.382	0.115	0.307
					0.75	0.287	0.088	0.225
HARD	100	15	10	0	0	0.344	0.212	0.620
					.5	0.224	0.070	0.462
					0.75	0.152	0.118	0.901
			8	2	0	0.116	0.087	0.434
					.5	0.233	0.059	0.569
					0.75	0.110	0.051	0.546
			6	4	0	0.292	0.055	0.456
					.5	0.229	0.054	0.495
					0.75	0.158	0.050	0.411

5.1.2 Setting 2

Here, we used three different sample sizes (n = 20, 50 and 100) and two different numbers of follow-ups (m = 5 and 15). We present only the results for the LASSO

and SCAD penalty functions, since in most cases, the SCAD and HARD penalty functions performed similarly in this setting. We set the model parameters as follows:

$$\boldsymbol{\beta}_1^0 = (2, -1, 1, 0.5, 0, 0, 0, 0, 0, 0, 0, 0),$$

$$\boldsymbol{\beta}_2^0 = (-3, -4, -1, -0.5, 0, 0, 0, 0, 0, 0, 0, 0),$$

$$\boldsymbol{\alpha}^{*0}_{m=5} = (-2, 0, 0, 0, 2, -0.5, 0, 0, 0, 0, 0, 0, 0),$$

$$\boldsymbol{\alpha}^{*0}_{m=15} = (-2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, -0.5, 0, 0, 0, 0, 0, 0, 0),$$

where the first entries of $\boldsymbol{\beta}_1^0$ and $\boldsymbol{\beta}_2^0$ are the intercepts in the pre- and post-change observation distribution regression models, respectively. Here, the vectors of true parameters included some smaller entries than that of Setting 1, which resulted in a more challenging problem. Tables 5–9 and 5–10 give the estimated sensitivities and specificities. We show the empirical estimation efficiencies, defined in (5.0.1) and (5.0.5), in Tables 5–11 and 5–12. The empirical efficiencies (the median of the ratios of the average MSEs), as defined in (5.0.3) and (5.0.6), are shown in Tables 5–13 and 5–14, for the LASSO and SCAD penalty functions, respectively.

Table 5–9: Estimated sensitivity (S_1) and specificity (S_2) for cases with p_{con} covariates from $MN(\mathbf{0}, \Sigma_\rho)$ and p_{bin} binary covariates under Model 1, Setting 2 using the LASSO penalty function.

	Setting					$\boldsymbol{\beta}_1$		$\boldsymbol{\beta}_2$		$\boldsymbol{\alpha}^*$	
Pen.	n	m	p_{con}	p_{bin}	ρ	S_1	S_2	S_1	S_2	S_1	S_2
LASSO	20	5	10	0	0	0.824	1.000	0.869	0.464	0.930	0.610
					0.75	0.816	0.929	0.916	0.433	0.950	0.590
			6	4	0	0.818	0.770	0.892	0.033	0.960	0.360
					0.75	0.500	0.377	0.684	0.141	0.908	0.340
		1 5	10	0	0	0.862	0.998	0.851	0.976	0.936	0.798
					0.75	0.840	0.971	0.833	0.901	0.898	0.722
			6	4	0	0.884	0.973	0.810	0.409	0.915	0.467
					0.75	0.810	0.952	0.826	0.440	0.969	0.439
	50	5	10	0	0	0.921	1.000	0.908	0.940	0.952	0.701
					0.75	0.873	0.997	0.839	0.481	0.827	0.637
			6	4	0	0.932	0.954	0.919	0.070	0.962	0.401
					0.75	0.929	0.938	0.958	0.120	0.976	0.372
		15	10	0	0	0.968	1.000	0.908	1.000	0.950	0.773
					0.75	0.931	1.000	0.931	1.000	0.944	0.896
			6	4	0	0.946	1.000	0.858	0.927	0.870	0.698
					0.75	0.922	1.000	0.879	0.928	0.937	0.742
	100	5	10	0	0	0.980	1.000	0.966	1.000	0.984	0.873
					0.75	0.969	1.000	0.916	0.945	0.977	0.727
			6	4	0	0.962	1.000	0.881	0.301	0.930	0.548
					0.75	0.942	1.000	0.914	0.303	0.971	0.515
		15	10	0	0	0.998	1.000	0.998	1.000	0.997	0.740
					0.75	0.978	1.000	0.978	1.000	0.963	0.707
			6	4	0	0.984	1.000	0.949	1.000	0.894	0.753
					0.75	0.968	1.000	0.908	1.000	0.950	0.773

Table 5–10: Estimated sensitivity (S_1) and specificity (S_2) for cases with p_{con} covariates from $MN(\mathbf{0}, \Sigma_\rho)$ and p_{bin} binary covariates under Model 1, Setting 2 using the SCAD penalty function.

	Setting					$\boldsymbol{\beta}_1$		$\boldsymbol{\beta}_2$		$\boldsymbol{\alpha}^*$	
Pen.	n	m	p_{con}	p_{bin}	ρ	S_1	S_2	S_1	S_2	S_1	S_2
SCAD	20	5	10	0	0	0.870	0.984	0.372	0.681	0.957	0.616
					0.75	0.899	0.541	0.306	0.765	0.945	0.598
			6	4	0	0.884	0.607	0.198	0.874	0.983	0.669
					0.75	0.918	0.704	0.246	0.765	0.978	0.578
		1 5	10	0	0	0.902	0.997	0.873	0.974	0.958	0.735
					0.75	0.885	0.888	0.887	0.901	0.926	0.681
			6	4	0	0.914	0.957	0.268	0.802	0.954	0.681
					0.75	0.886	0.962	0.237	0.745	0.952	0.674
	50	5	10	0	0	0.969	0.998	0.974	0.937	0.982	0.682
					0.75	0.971	0.944	0.936	0.440	0.940	0.693
			6	4	0	0.978	0.885	0.745	0.548	0.960	0.668
					0.75	0.985	0.869	0.559	0.621	0.942	0.648
		15	10	0	0	0.987	1.000	0.987	1.000	0.992	0.687
					0.75	0.972	0.997	0.981	0.998	0.974	0.689
			6	4	0	0.980	0.998	0.955	0.890	0.959	0.777
					0.75	0.980	1.000	0.965	0.870	0.944	0.782
	100	5	10	0	0	0.992	1.000	0.995	1.000	0.996	0.727
					0.75	0.992	1.000	0.995	0.800	0.991	0.670
			6	4	0	1.000	0.965	0.950	0.545	0.953	0.770
					0.75	0.999	0.995	0.954	0.460	0.967	0.720
		15	10	0	0	0.999	1.000	1.000	1.000	0.999	0.683
					0.75	0.996	1.000	0.995	1.000	0.996	0.670
			6	4	0	0.995	1.000	0.991	0.985	0.956	0.820
					0.75	0.999	1.000	0.992	0.980	0.961	0.833

Table 5–11: Empirical estimation efficiency ($e(\hat{\boldsymbol{\theta}})$) for cases with p_{con} covariates from $MN(\mathbf{0}, \Sigma_\rho)$ and p_{bin} binary covariates under Model 1, Setting 2 using the LASSO penalty function.

	Setting							
Pen.	n	m	p_{con}	p_{bin}	ρ	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\alpha}^*$
LASSO	20	5	10	0	0	0.310	0.035	1.008
					0.75	0.191	0.002	0.525
			6	4	0	0.456	0.006	1.838
					0.75	0.581	0.038	0.580
		15	10	0	0	0.274	0.343	0.353
					0.75	0.261	0.199	0.406
			6	4	0	0.488	0.004	1.274
					0.75	0.498	0.005	1.108
	50	5	10	0	0	0.442	0.441	1.112
					0.75	0.376	0.372	0.887
			6	4	0	0.666	0.032	1.747
					0.75	0.532	0.025	1.406
		15	10	0	0	0.299	0.345	1.173
					0.75	0.251	0.248	0.404
			6	4	0	0.583	0.716	1.043
					0.75	0.414	0.480	0.606
	100	5	10	0	0	0.310	1.113	2.087
					0.75	0.235	0.373	1.105
			6	4	0	0.653	0.629	1.920
					0.75	0.437	0.606	1.662
		15	10	0	0	0.276	0.339	1.906
					0.75	0.283	0.318	0.733
			6	4	0	0.514	0.590	0.888
					0.75	0.341	0.565	0.946

Table 5–12: Empirical estimation efficiency ($e(\hat{\boldsymbol{\theta}})$) for cases with p_{con} covariates from $MN(\mathbf{0}, \Sigma_\rho)$ and p_{bin} binary covariates under Model 1, Setting 2 using the SCAD penalty function.

	Setting							
Pen.	n	m	p_{con}	p_{bin}	ρ	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\alpha}^*$
SCAD	20	5	10	0	0	0.314	1.875	0.581
					0.75	0.475	0.420	0.935
			6	4	0	0.622	2.291	1.168
					0.75	0.104	1.741	0.418
		15	10	0	0	0.205	0.277	0.263
					0.75	0.298	0.159	0.235
			6	4	0	0.529	0.986	0.400
					0.75	0.342	0.986	0.301
	50	5	10	0	0	0.323	0.255	0.474
					0.75	0.321	0.405	0.351
			6	4	0	0.537	0.719	0.657
					0.75	0.426	0.861	0.621
		15	10	0	0	0.256	0.295	0.647
					0.75	0.206	0.169	0.293
			6	4	0	0.440	0.389	0.518
					0.75	0.273	0.332	0.338
	100	5	10	0	0	0.280	0.389	0.902
					0.75	0.172	0.401	0.323
			6	4	0	0.480	0.525	0.508
					0.75	0.288	0.474	0.417
		15	10	0	0	0.263	0.281	1.068
					0.75	0.159	0.179	0.407
			6	4	0	0.382	0.403	0.619
					0.75	0.241	0.211	0.507

Table 5–13: Empirical efficiency ($e^*(\hat{\boldsymbol{\theta}})$) for cases with p_{con} covariates from $MN(\mathbf{0}, \Sigma_\rho)$ and p_{bin} binary covariates under Model 1, Setting 2 using the LASSO penalty function.

	Setting							
Pen.	n	m	p_{con}	p_{bin}	ρ	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\alpha}^*$
LASSO	20	5	10	0	0	0.354	0.033	1.772
					0.75	0.228	0.006	1.056
			6	4	0	0.489	0.008	0.178
					0.75	0.605	0.126	0.152
		15	10	0	0	0.457	0.457	0.658
					0.75	0.521	0.376	1.073
			6	4	0	0.552	0.005	0.218
					0.75	0.607	0.006	0.250
	50	5	10	0	0	0.526	0.560	1.537
					0.75	0.561	0.716	1.635
			6	4	0	0.522	0.041	0.419
					0.75	0.555	0.026	0.341
		15	10	0	0	0.351	0.395	2.031
					0.75	0.488	0.520	1.336
			6	4	0	0.475	0.680	0.304
					0.75	0.534	0.689	0.280
	100	5	10	0	0	0.369	1.086	1.312
					0.75	0.697	0.854	4.019
			6	4	0	0.423	0.529	0.506
					0.75	0.515	0.814	0.667
		15	10	0	0	0.307	0.376	2.260
					0.75	0.647	0.802	2.712
			6	4	0	0.384	0.461	0.433
					0.75	0.501	0.728	0.418

Table 5–14: Empirical efficiency ($e^*(\hat{\boldsymbol{\theta}})$) for cases with p_{con} covariates from $MN(\mathbf{0}, \Sigma_\rho)$ and p_{bin} binary covariates under Model 1, Setting 2 using the SCAD penalty function.

	Setting							
Pen.	n	m	p_{con}	p_{bin}	ρ	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\alpha}^*$
SCAD	20	5	10	0	0	0.369	0.768	0.878
					0.75	0.225	0.297	0.763
			6	4	0	0.383	1.269	1.795
					0.75	0.060	1.144	0.226
		15	10	0	0	0.397	0.378	0.408
					0.75	0.550	0.343	0.430
			6	4	0	0.564	0.986	0.381
					0.75	0.496	0.989	0.396
	50	5	10	0	0	0.407	0.312	0.657
					0.75	0.501	0.562	0.670
			6	4	0	0.475	0.628	0.306
					0.75	0.415	0.870	0.359
		15	10	0	0	0.299	0.337	0.887
					0.75	0.458	0.392	0.634
			6	4	0	0.343	0.372	0.253
					0.75	0.352	0.460	0.272
	100	5	10	0	0	0.319	0.393	1.391
					0.75	0.381	0.706	0.694
			6	4	0	0.314	0.398	0.328
					0.75	0.332	0.579	0.332
		15	10	0	0	0.282	0.308	1.415
					0.75	0.368	0.426	0.728
			6	4	0	0.346	0.331	0.348
					0.75	0.279	0.302	0.351

5.1.3 Discussion

The results of our simulation studies for Model 1 suggest:

1. Unsurprisingly, increasing both m and n improves variable selection.
2. Variable selection and estimation for the post-change observation distribution means (with $\boldsymbol{\beta}_2$ as the vector of regression parameters) is more challenging than

the pre-change observation distribution means (with $\boldsymbol{\beta}_1$ as the vector of regression parameters). Our model allows for no change, in which all the observations on a subject will be from the observation distribution before the change with no observation from the distribution after the change. We also assumed that the first observation on each subject comes from the pre-change distribution. Therefore, there are more observations from the distribution before the change, leading to enhanced pre-change statistical inference. This results in high estimated sensitivities and specificities for $\boldsymbol{\beta}_1$. Increasing the number of follow-ups, m, when n is fixed, as well as increasing n, for a given m, improves the estimated sensitivities and specificities for $\boldsymbol{\beta}_2$. However, the LASSO results in a sparse model and it performs well on the post-change observation distribution. It chooses a small number of covariates for the observation distribution after the change and therefore, results in a large number of correctly zero estimates. Hence, the estimated sensitivity using the LASSO is large and its induced bias into estimators is small (particularly, for small n's) with small mean square errors; see Tables 5–9, 5–11, and 5–13 for $n = 20$ and 50.

3. High collinearity results in a challenging inferential problem. Increasing n and m, improves overall variable selection in the presence of collinearity. Our simulations suggest that the LASSO is not recommended in the presence of high collinearity. The LASSO selects only one covariate in a set of correlated covariates ignoring their individual effects in the model, resulting in poor variable selection, particularly for the change-point distribution.

4. Increasing n, for a given m improves variable selection using the penalized likelihood approach. However, since the tuning parameter in the LASSO penalty function, $n\lambda_n$, is an increasing function of n, it introduces bias into estimators, particularly for the change-point distribution. This results in large mean square errors as displayed in Tables 5–11 and 5–13.
5. For a fixed n, increasing m has a particular impact on $\widehat{\boldsymbol{\beta}}_2$, and $\widehat{\boldsymbol{\alpha}}^*$, the post-change observation distribution and the change-point distribution parameter estimates, respectively. Both variable selection and estimation are improved.

5.2 Simulation Scenario: Model 2

Although Model 1 may not be very realistic for most real change-point settings, its simplicity allows for extensive, if somewhat preliminary, simulations. Model 2 is more flexible, but computationally more intensive to evaluate. Following are the two settings we considered for this model: 1) We assumed a non-constant hazard ($\alpha_{0i} \neq \alpha_{0j}$ for at least one $i \neq j = 1, 2, \ldots, m-1$), and 2) included time as a covariate for the observation distribution means.

5.2.1 Setting 1:

We took combinations of $n = 50$, 150, and 500 and $m = 5$ and 15, respectively. We used the LASSO and HARD penalty functions. The parameter vectors were chosen to be:

$$\boldsymbol{\beta}_1^0 = (2, 1, 3, 0, 0, 0, 0, 0, 0, 0, 0), \eta_1^0 = -1,$$

$$\boldsymbol{\beta}_2^0 = (-3, -1, -2, 0, 0, 0, 0, 0, 0, 0, 0), \eta_2^0 = -4,$$

$$\boldsymbol{\alpha}^0 = (-2, 2, -1, 0, 0, 0, 0, 0, 0, 0),$$

$$\boldsymbol{\alpha}_{0}{}_{m=5}^{0} = (-2, 1, -3, 5),$$

$$\boldsymbol{\alpha}_{0}{}_{m=15}^{0} = (-2, 1, -3, 5, -3, 1, -3, 5, -3, 1, -3, 5, -3, 1).$$

The first entries of $\boldsymbol{\beta}_1^0$ and $\boldsymbol{\beta}_2^0$ correspond to intercepts in the regression models for the observations before and after the change, respectively. Here, we chose the baseline hazard not to be smooth.

Tables 5–15 and 5–16 give the estimated sensitivities and specificities using the LASSO and HARD penalties, respectively. Tables 5–17 and 5–18 give the empirical estimation efficiencies of the parameter estimation defined in (5.0.2) and (5.0.5) for the vector of parameter estimates (for $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\alpha}$, and $\boldsymbol{\alpha}_0$).

Table 5–15: Estimated sensitivity (S_1) and specificity (S_2) for cases with p_{con} covariates from $MN(\mathbf{0}, \Sigma_\rho)$ and p_{bin} binary covariates under Model 2, Setting 1 using the LASSO penalty function.

	Setting					$\boldsymbol{\beta}_1$		$\boldsymbol{\beta}_2$		$\boldsymbol{\alpha}^*$	
Pen.	n	m	p_{con}	p_{bin}	ρ	S_1	S_2	S_1	S_2	S_1	S_2
LASSO	50	5	10	0	0	1.000	1.000	1.000	1.000	0.574	0.977
					0.75	1.000	1.000	1.000	1.000	0.497	0.972
			6	4	0	1.000	1.000	1.000	0.998	0.436	1.000
					.75	1.000	1.000	1.000	0.995	0.428	0.973
		15	10	0	0	1.000	1.000	1.000	1.000	0.380	1.000
					0.75	1.000	1.000	1.000	1.000	0.397	0.981
			6	4	0	1.000	1.000	1.000	1.000	0.479	0.980
					.75	1.000	1.000	1.000	1.000	0.567	0.877
	150	5	10	0	0	1.000	1.000	1.000	1.000	0.593	1.000
					0.75	1.000	1.000	1.000	1.000	0.540	0.997
			6	4	0	1.000	1.000	1.000	1.000	0.574	1.000
					.75	1.000	1.000	1.000	1.000	0.511	1.000
		15	10	0	0	1.000	1.000	1.000	1.000	0.559	1.000
					0.75	1.000	1.000	1.000	1.000	0.438	1.000
			6	4	0	1.000	1.000	1.000	1.000	0.604	1.000
					.75	1.000	1.000	1.000	1.000	0.462	1.000
	500	5	10	0	0	1.000	1.000	1.000	1.000	0.860	1.000
					0.75	1.000	1.000	1.000	1.000	0.764	1.000
			6	4	0	1.000	1.000	1.000	1.000	0.844	1.000
					.75	1.000	1.000	1.000	1.000	0.745	1.000

Table 5–16: Estimated sensitivity (S_1) and specificity (S_2) for cases with p_{con} covariates from $MN(\mathbf{0}, \Sigma_\rho)$ and p_{bin} binary covariates under Model 2, Setting 1 using the HARD penalty function.

	Setting					$\boldsymbol{\beta}_1$		$\boldsymbol{\beta}_2$		$\boldsymbol{\alpha}^*$	
Pen.	n	m	p_{con}	p_{bin}	ρ	S_1	S_2	S_1	S_2	S_1	S_2
HARD	50	5	10	0	0	1.000	1.000	1.000	1.000	0.940	0.976
					0.75	1.000	1.000	1.000	1.000	0.941	0.972
			6	4	0	1.000	1.000	1.000	0.996	0.951	0.993
					.75	1.000	1.000	1.000	0.991	0.953	0.975
		15	10	0	0	1.000	1.000	1.000	1.000	0.882	0.996
					0.75	1.000	1.000	1.000	1.000	0.921	0.979
			6	4	0	1.000	1.000	1.000	1.000	0.908	0.994
					.75	1.000	1.000	1.000	1.000	0.889	0.987
	150	5	10	0	0	1.000	1.000	1.000	1.000	0.982	1.000
					0.75	1.000	1.000	1.000	1.000	0.978	0.998
			6	4	0	1.000	1.000	1.000	1.000	0.981	1.000
					.75	1.000	1.000	1.000	1.000	0.979	1.000
		15	10	0	0	1.000	1.000	1.000	1.000	0.892	1.000
					0.75	1.000	1.000	1.000	1.000	0.897	1.000
			6	4	0	1.000	1.000	1.000	1.000	0.894	1.000
					.75	1.000	1.000	1.000	1.000	0.889	1.000
	500	5	10	0	0	1.000	1.000	1.000	1.000	0.995	1.000
					0.75	1.000	1.000	1.000	1.000	0.995	1.000
			6	4	0	1.000	1.000	1.000	1.000	0.995	1.000
					.75	1.000	1.000	1.000	1.000	0.991	1.000
		15	10	0	0	1.000	1.000	1.000	1.000	0.860	1.000
					0.75	1.000	1.000	1.000	1.000	0.853	1.000
			6	4	0	1.00	1.000	1.000	1.000	0.840	1.000
					.75	1.00	1.000	1.000	1.000	0.850	1.000

Table 5–17: Empirical estimation efficiency ($e(\hat{\boldsymbol{\theta}})$) for cases with p_{con} covariates from $MN(\mathbf{0}, \Sigma_\rho)$ and p_{bin} binary covariates under Model 2, Setting 1 using the LASSO penalty function.

	Setting								
Pen.	n	m	p_{con}	p_{bin}	ρ	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha}_0$
LASSO	50	5	10	0	0	0.314	0.223	0.757	1.521
					0.75	0.150	0.147	0.736	1.246
			6	4	0	0.382	0.335	0.775	1.076
					0.75	0.198	0.242	0.739	0.930
		15	10	0	0	0.277	0.254	0.626	0.883
					0.75	0.130	0.108	0.603	0.833
			6	4	0	0.213	0.320	0.727	0.895
					0.75	0.348	0.205	0.889	0.932
	150	5	10	0	0	0.301	0.383	0.900	2.450
					0.75	0.151	0.206	0.822	2.455
			6	4	0	0.453	0.461	0.884	1.793
					0.75	0.256	0.269	0.791	1.627
		15	10	0	0	0.292	0.324	0.718	0.869
					0.75	0.226	0.180	0.778	0.843
			6	4	0	0.338	0.360	0.674	0.773
					0.75	0.249	0.267	0.653	0.768
	500	5	10	0	0	0.363	0.379	1.826	8.028
					0.75	0.141	0.203	1.312	7.684
			6	4	0	0.456	0.514	1.352	5.064
					0.75	0.265	0.311	1.111	5.888

Table 5–18: Empirical estimation efficiency ($e(\hat{\boldsymbol{\theta}})$) for cases with p_{con} covariates from $MN(\mathbf{0}, \Sigma_\rho)$ and p_{bin} binary covariates under Model 2, Setting 1 using the HARD penalty function.

	Setting								
Pen.	n	m	p_{con}	p_{bin}	ρ	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha}_0$
HARD	50	5	10	0	0	0.302	0.226	0.517	0.339
					0.75	0.150	0.134	0.371	0.472
			6	4	0	0.381	0.317	0.440	0.474
					0.75	0.193	0.238	0.379	0.413
		15	10	0	0	0.276	0.252	0.535	0.843
					0.75	0.127	0.150	0.398	0.899
			6	4	0	0.274	0.258	0.482	0.827
					0.75	0.201	0.185	0.470	0.848
	150	5	10	0	0	0.299	0.375	0.422	0.501
					0.75	0.150	0.200	0.342	0.456
			6	4	0	0.446	0.456	0.438	0.481
					0.75	0.253	0.276	0.297	0.433
		15	10	0	0	0.322	0.326	0.553	0.907
					0.75	0.197	0.183	0.462	1.014
			6	4	0	0.385	0.376	0.484	1.015
					0.75	0.231	0.244	0.356	0.947
	500	5	10	0	0	0.349	0.365	0.414	0.459
					0.75	0.138	0.190	0.301	0.470
			6	4	0	0.419	0.495	0.461	0.478
					0.75	0.250	0.286	0.277	0.501
		15	10	0	0	0.341	0.388	0.725	1.368
					0.75	0.180	0.165	0.493	1.205
			6	4	0	0.420	0.400	0.555	1.301
					0.75	0.246	0.256	0.516	1.259

5.2.2 Setting 2

In this last setting, the cases considered were with $n = 20, 50$ and 100, and we chose the number of follow-ups to be 5 and 15. The true parameters were chosen to

be:

$$\boldsymbol{\beta}_1^0 = (2, 1, 0.5, 0, 0, 0, 0, 0, 0, 0, 0), \eta_1^0 = -1,$$

$$\boldsymbol{\beta}_2^0 = (-3, -1, -0.5, 0, 0, 0, 0, 0, 0, 0, 0), \eta_2^0 = -4,$$

$$\boldsymbol{\alpha}^0 = (-2, 2, -0.5, 0, 0, 0, 0, 0, 0, 0),$$

$$\boldsymbol{\alpha}_{0_{m=5}}^0 = (-2, 0, 1, 0),$$

$$\boldsymbol{\alpha}_{0_{m=15}}^0 = (-2, 0, 1, 0, 0, -1, 0, 1, 0, 0, -1, 0, 1, 0, 0, 1),$$

where the first entries of $\boldsymbol{\beta}_1^0$ and $\boldsymbol{\beta}_2^0$ represent the intercepts. The true parameters here include small values which are not easy to distinguish from zero using regular maximum likelihood. This is particularly true for the vectors of increments in the logit of the baseline hazard which have a sparser structure compared to their structure in Setting 1 (the baseline hazard will be smoother in this setting). For this setting, we only assessed the performance of the LASSO penalty function, the one used to analyze in the Alzheimer's disease example in Chapter 6.

Table 5–19 gives the estimated sensitivities and specificities. Tables 5–20 presents the empirical efficiency (the median of the ratios of the average MSEs), as defined in (5.0.4) and (5.0.6).

Table 5–19: Estimated sensitivity (S_1) and specificity (S_2) for cases with p_{con} covariates from $MN(\mathbf{0}, \Sigma_\rho)$ and p_{bin} binary covariates under Model 2, Setting 2 using the LASSO penalty function.

	Setting					$\boldsymbol{\beta}_1$		$\boldsymbol{\beta}_2$		$\boldsymbol{\alpha}^*$	
Pen.	n	m	p_{con}	p_{bin}	ρ	S_1	S_2	S_1	S_2	S_1	S_2
LASSO	20	5	10	0	0	0.728	1.000	0.700	0.715	0.609	0.904
					0.75	0.809	0.975	0.857	0.873	0.879	0.491
			6	4	0	0.709	0.986	0.675	0.710	0.621	0.799
					.75	0.734	0.994	0.780	0.659	0.706	0.879
		15	10	0	0	0.651	1.000	0.744	1.000	0.580	0.983
					0.75	0.735	1.000	0.713	0.997	0.689	0.790
			6	4	0	0.696	0.996	0.714	0.985	0.503	0.890
					.75	0.721	0.984	0.720	0.995	0.567	0.822
	50	5	10	0	0	0.896	1.000	0.839	1.000	0.791	0.948
					0.75	0.853	1.000	0.841	0.995	0.862	0.645
			6	4	0	0.824	1.000	0.787	0.983	0.674	0.946
					.75	0.803	1.000	0.826	0.926	0.725	0.96
		15	10	0	0	0.909	1.000	0.933	1.000	0.474	1.000
					0.75	0.789	1.000	0.876	1.000	0.687	0.825
			6	4	0	0.931	1.000	0.919	1.000	0.684	0.894
					.75	0.763	1.000	0.805	1.000	0.586	0.971
	100	5	10	0	0	0.957	1.000	0.907	1.000	0.742	1.000
					0.75	0.874	1.000	0.839	1.000	0.845	0.920
			6	4	0	0.874	1.000	0.837	1.000	0.833	0.931
					.75	0.872	1.000	0.841	1.000	0.726	1.000

Table 5–20: Empirical estimation efficiency ($e(\hat{\boldsymbol{\theta}})$) for cases with p_{con} covariates from $MN(\mathbf{0}, \Sigma_\rho)$ and p_{bin} binary covariates under Model 2, Setting 2 using the LASSO penalty function.

	Setting								
Pen.	n	m	p_{con}	p_{bin}	ρ	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha}_0$
LASSO	20	5	10	0	0	0.931	0.297	0.542	0.726
					0.75	0.775	0.082	0.516	0.846
			6	4	0	0.821	0.441	0.569	0.865
					0.75	0.848	0.014	0.390	0.758
		15	10	0	0	0.916	0.963	0.352	0.436
					0.75	0.929	0.738	0.287	0.474
			6	4	0	0.833	0.831	0.577	0.651
					0.75	0.732	0.709	0.446	0.646
	50	5	10	0	0	0.963	0.770	0.622	0.929
					0.75	0.895	0.637	1.088	1.037
			6	4	0	0.934	0.800	0.955	0.941
					0.75	0.917	0.627	0.521	0.946
		15	10	0	0	0.992	0.976	0.481	0.679
					0.75	0.970	0.941	0.510	0.619
			6	4	0	0.942	0.924	0.730	0.793
					0.75	0.945	0.925	0.403	0.654
	100	5	10	0	0	0.988	0.955	0.792	1.016
					0.75	0.957	0.832	0.815	1.058
			6	4	0	0.958	0.835	0.790	1.055
					0.75	0.953	0.787	0.569	0.953

Table 5–21: Estimated sensitivity (S_1) and specificity (S_2) for cases with $p_{con} = 6$ covariates from a $MN(\mathbf{0}, \Sigma_\rho)$ distribution, where $\rho = 0.75$ and $p_{bin} = 4$ binary covariates under Model 2, Setting 2.

Pen.	Setting		β_1		β_2		α		α_0	
	n	m	S_1	S_2	S_1	S_2	S_1	S_2	S_1	S_2
LASSO	20	5	0.734	0.994	0.780	0.659	0.706	0.879	0.392	0.308
	50		0.803	1.000	0.826	0.926	0.725	0.967	0.440	0.370
	100		0.872	1.000	0.841	1.000	0.726	1.000	0.491	0.640
	20	15	0.721	0.984	0.720	0.995	0.567	0.822	0.746	0.117
	50		0.763	1.000	0.805	1.000	0.586	0.971	0.701	0.268
	100		1.000	1.000	1.000	1.000	0.546	0.968	0.662	0.515
SCAD	20	5	0.849	0.987	0.867	0.737	0.872	0.877	0.351	0.592
	50		1.000	1.000	1.000	0.946	0.952	0.905	0.356	0.770
	100		1.000	1.000	1.000	1.000	0.955	0.990	0.347	0.919
	20	15	0.846	0.985	0.862	0.985	0.776	0.833	0.774	0.464
	50		1.000	1.000	1.000	1.000	0.811	0.963	0.748	0.722
	100		1.000	0.995	1.000	1.000	0.931	0.887	0.764	0.822

Table 5–22: Empirical efficiency ($e^*(\hat{\boldsymbol{\theta}})$) under Model 2, Setting 2.

	Setting		LASSO			SCAD			Oracle		
	n	m	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\alpha}^*$	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\alpha}^*$	$\boldsymbol{\beta}_1$	$\boldsymbol{\beta}_2$	$\boldsymbol{\alpha}^*$
Model 2	20	5	0.931	0.040	0.362	0.985	0.049	0.390	0.974	0.036	0.431
	50		0.969	0.812	0.275	0.988	0.911	0.189	0.988	0.894	0.412
	100		0.980	0.907	0.190	0.996	0.961	0.140	0.993	0.948	0.437
	20	15	0.956	0.912	0.372	0.910	0.960	0.471	0.971	0.943	0.545
	50		0.959	0.966	0.268	0.957	0.966	0.193	0.970	0.967	0.365
	100		0.336	0.387	0.882	0.378	0.371	0.875	0.357	0.354	0.768

For the case with four binary covariates out of 10 and pair-wise correlation equal to 0.75, we compared variable selection performance using the LASSO and SCAD penalty functions; Table 5–21 displays the results. We also report the median ratios of the empirical efficiency, as defined in (5.0.4) and (5.0.6) for the LASSO and SCAD

penalty functions, and the oracle model (knowing the true zero parameters), in Table 5–22.

5.2.3 Discussion

1. With regard to the observation distributions, the LASSO and HARD penalty functions both perform well in Setting 1. However, the HARD penalty function performs better for the change-point distribution. We believe this is the result of the non-sparse structure, which influences the ability of the LASSO penalty to detect the correct model; see Tables 5–15 and 5–16.
2. Regrading the change-point distribution, and in particular the baseline hazard (with $\boldsymbol{\alpha}^*$ as the vector of parameters), the LASSO penalty function can introduce a large bias into estimators as n increases. It reduces the efficiency, reflected by large mean square error ratios (see Tables 5–17 and 5–20 for Settings 1 and 2, respectively).
3. The results of Tables 5–21 and 5–22 concur with those of Setting 1. For a small n ($n = 20$) and m ($m = 5$), the penalty functions set small estimates to zero, resulting in a sparse model, therefore explaining the low empirical efficiencies of the parameter estimators for the observation distribution means after the change and ensuring satisfactory variable selection and estimation (for more details see 2. in the Discussion of Model 1).
4. We can see in Table 5–22, that under Setting 2 and for both the LASSO and SCAD penalty functions, all the (median) ratios in the table are less than 1, indicating better performance of the LASSO/SCAD-based estimators compared

to the full model-based estimators, and comparable to the performance of the oracle estimators. As the sample size increases, these ratios increase toward 1, indicating that the proposed method and the ridge estimators, which for large sample sizes behave similar to the ordinary maximum likelihood estimators, perform similarly in some cases.

5.3 Discussion of the Simulation Results

Based on our simulation results for small sample sizes, such as $n = 20$, model selection and estimation are quite challenging. As the sample size increases the performance of selection and estimation in both Model 1 and Model 2 improves using the SCAD and HARD penalty function. However, the LASSO penalty function is not recommended based on our simulation results, because of the large bias it introduces into the estimators.

We recommend for small n's and m's, the LASSO penalty function with ridge regression for the change-point distribution parameters. For large n's and m's, in both assessed scenarios (Model 1 and Model 2), the SCAD and HARD penalties perform well, selecting the subset of true parameters and estimating them, simultaneously.

If there is enough evidence for the presence of a high degree of multi-collinearity in the set of covariates, the LASSO penalty function is not recommended and it does not guarantee the choice of the best subset of covariates. The ridge penalty functions are needed on the change-point distribution parameters to prevent "wild" estimates of parameters α_{0k} and α_j when m is large and there are highly correlated covariates.

In summary, the performance of our methods, shown by simulations, is similar to the case in which we know the true model in advance, confirming the oracle property, at least for the penalty functions that we considered. Our proposed method, according to the simulation results, is very effective for variable selection and estimation in the change-point models under consideration.

CHAPTER 6
Risk Factors for Cognitive Decline in Alzheimer's Disease

In Alzheimer's disease (AD), the rate of progression is highly variable. As a result, there has been much interest in factors associated with cognitive decline, for example, age, education, and sex; see Mortimer et al. (1992), Wilson (1999, 2004), Herbert (2000), Stern et al. (1999), Bennet (2002), Hall (2007). Some studies have suggested that the rate of cognitive decline in patients with AD is not constant but instead piece-wise linear (Joseph et al. 1999, Hall 2000), and most researchers would agree that patients generally experience an initial stable period followed by a period of roughly linear decline, ending with another relatively stable period late in the disease. The early and late stable periods, though, may only be artifacts due to the lack of sensitivity of the cognitive tests commonly used. In Figure 6–1, we show the trajectories of cognitive scores over time, for several subjects with AD. These are a randomly chosen subset of the subjects in the data set analyzed in this chapter.

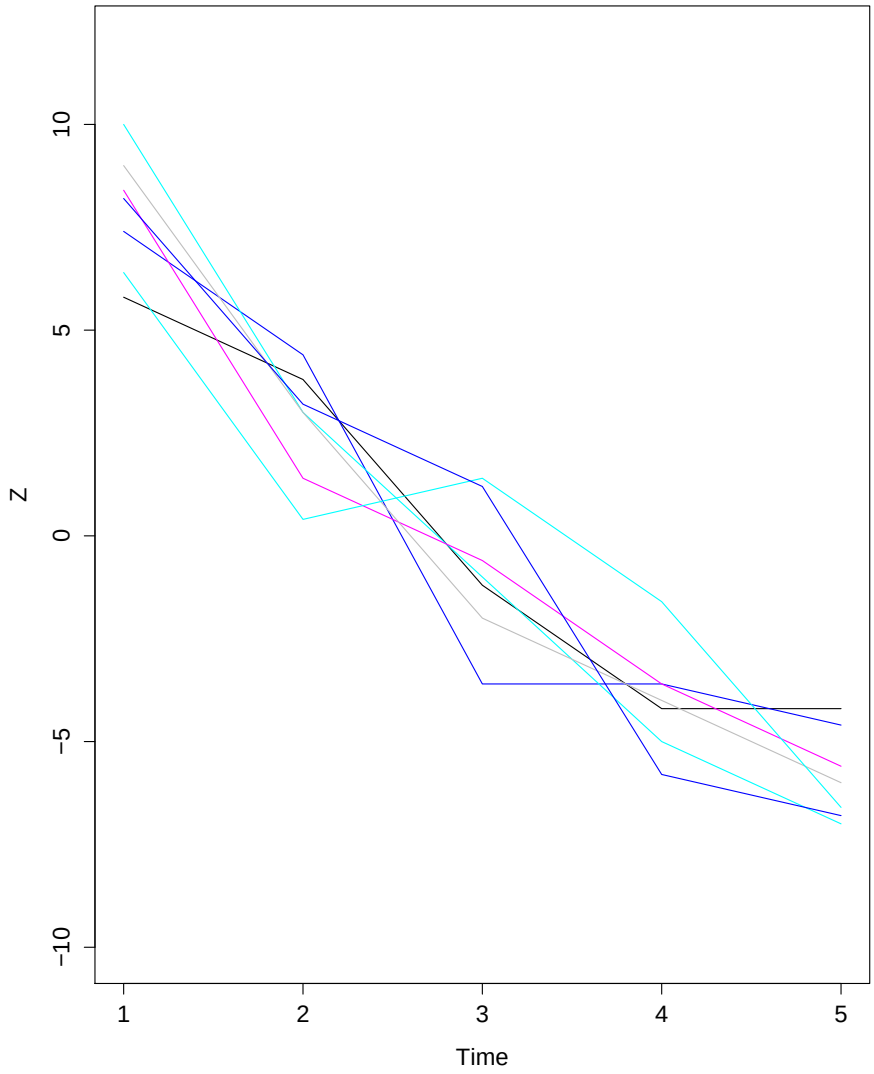


Figure 6–1: Spaghetti plot of centered MMSE scores for 7 subjects over time ($Z_{tj} = MMSE_{tj} - \bar{MMSE}_j$ for $t = 1, 2, \ldots, 5$ and $j = 1, 2, \ldots, 7$.

Since almost all of the subjects were already in the second decline stage at the start of the study, we model the trajectories of cognitive decline with a single change-point. We expect the general picture of MMSE scores to be similar to Figure 6–2.

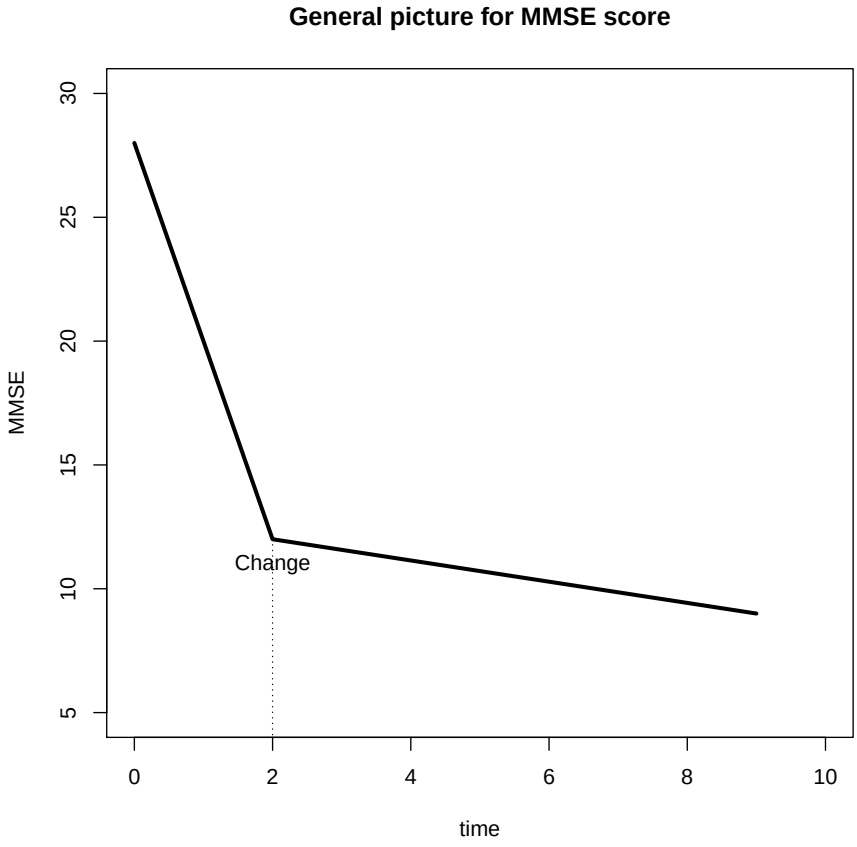


Figure 6–2: MMSE scores over time

Understanding the pattern of decline in subjects with AD could assist in preparing for the coming public health burden caused by the increase in life expectancy. Also, models that predict disease course can be applied at the individual level, to assist caregivers. Finally, these models provide a valuable understanding of the natural history of Alzheimer's disease.

If a patient is at an early stage of the disease, determining the time at which their rate of decline begins to accelerate is important both in describing the natural

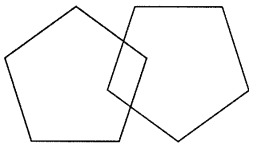
history of the disease process and in identifying the optimal time window for which treatments might be useful. At the same time, it is important to identify which subjects will experience rapid decline and be at risk for early institutionalization.

Folstein et al. (1975) introduced the Mini-Mental State Examination (MMSE), a brief test of mental status and cognition. It is one of the most commonly used tests to monitor dementia progression over time. In Figure 6–3, we display the Mini-Mental State Exam, a 30-point questionnaire. The lower the score, the more cognitively impaired the subject.

The Mini-Mental State Exam

Patient_____ Examiner _____ Date_____

Maximum	Score	
		Orientation
5	( )	What is the (year) (season) (date) (day) (month)?
5	( )	Where are we (state) (country) (town) (hospital) (floor)?
		Registration
3	( )	Name 3 objects: 1 second to say each. Then ask the patient all 3 after you have said them. Give 1 point for each correct answer. Then repeat them until he/she learns all 3. Count trials and record. Trials _____
		Attention and Calculation
5	( )	Serial 7's. 1 point for each correct answer. Stop after 5 answers. Alternatively spell "world" backward.
		Recall
3	( )	Ask for the 3 objects repeated above. Give 1 point for each correct answer.
		Language
2	( )	Name a pencil and watch.
1	( )	Repeat the following "No ifs, ands, or buts"
3	( )	Follow a 3-stage command: "Take a paper in your hand, fold it in half, and put it on the floor."
1	( )	Read and obey the following: CLOSE YOUR EYES
1	( )	Write a sentence.
1	( )	Copy the design shown.



_____ Total Score

ASSESS level of consciousness along a continuum _____

Alert Drowsy Stupor Coma

"MINI-MENTAL STATE." A PRACTICAL METHOD FOR GRADING THE COGNITIVE STATE OF PATIENTS FOR THE CLINICIAN. *Journal of Psychiatric Research*, 12(3): 189-198, 1975. Used by permission.

Figure 6–3: Mini-Mental State Examination questionnaire

Mortimer et al. (1992) followed a number of subjects in the Minneapolis VA Medical Center and from the community at large who had probable AD. The purpose of their study was to explore the effects of subject-specific characteristics on cognitive decline. They recorded MMSE scores on the subjects every six months over two and half years. They also recorded baseline covariates for each subject and fitted a multiple regression model over time, adjusting for disease duration. They proposed a multiple regression model for the MMSE scores using the information recorded at baseline, as well as time, as covariates in the model. To choose the best submodel, they used step-wise variable selection.

The main drawback with the analysis of Mortimer et al. is that they did not allow for the possibility of a change in the observation distribution over the follow-up time period. Not all covariates included in the model would be expected to be associated with decline, nor would they necessarily have the same effect before and after the possible change. Our model has five main features: (i) We introduce a multi-path change-point model to describe decline. (ii) We allow for separate regression components for the distribution of the observations before and after the change, as well as for the change-point distribution. (iii) We introduce an interaction term in the model in order to permit the effect of time to depend on subject-specific covariates, and (iv) we apply our method of variable selection for multi-path change-point models to the current setting. While Mortimer et al. carried out variable selection using step-wise multiple linear regression, this approach is known to be unreliable when used for complex models, such as our change-point model.

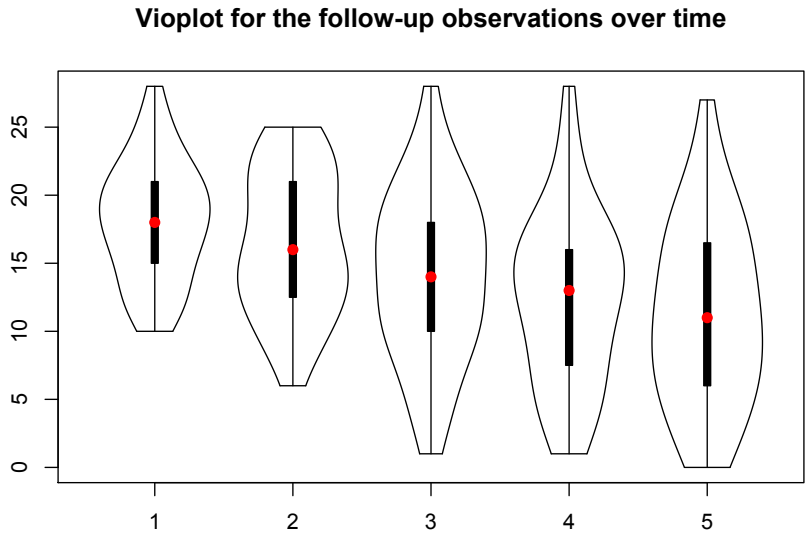


Figure 6–4: Boxplots and estimated density curves for the MMSE scores of subjects over time

Figure 6–4 is a vioplot (boxplot with kernel density estimates) of the MMSE scores over time. The density estimates are roughly consistent with "pure" normal distributions during early and late follow-up, and with a mixture of normal distributions in between. It seems reasonable to assume a multivariate regression with time as one of the covariates in the model for MMSE scores before and after the possible change. The timings of the unobserved putative change-points are likely highly variable and would occur towards the middle of follow-up. Since there were only five follow-up times, a change could only occur at one of four positions (after the first visit). Figure 6–4 is consistent with a mixture model form, since the estimated densities are bimodal. This mixture form is particularly apparent at the

second follow-up. The extra distribution uncertainty also qualitatively supports a model with a change-point.

Since the subjects were recruited at different disease stages, date of entry into the study is not suitable as a time origin without adjustment. Including disease duration at entry as a covariate in the model resolves this issue. Formally, for every subject i, $(i = 1, 2, \ldots, n)$ we assumed that conditional on the change occuring at $\tau = k$,

$$Y_{it} \sim N(\beta_{10} + \mathbf{x}_{it}^{\top}\boldsymbol{\beta}_1, \sigma_1^2) \ , \ \text{for} \ \ t = 1, 2, \ldots, k$$
$$Y_{it} \sim N(\beta_{20} + \mathbf{x}_{it}^{\top}\boldsymbol{\beta}_2, \sigma_2^2) \ , \ \text{for} \ \ t = k+1, k+2, \ldots, m-1,$$

and if a change occurs at $\tau = m$, then $Y_{it} \sim N(\beta_{10} + \mathbf{x}_{it}^{\top}\boldsymbol{\beta}_1, \sigma_1^2)$ for $t = 1, 2, \ldots, m$. Here, Y_{it} represents the MMSE score for subject i at time t, and $\mathbf{x}_{it}$ the vector of covariates for subject i at time t. The dependence of the covariate values on time t arise through the interaction between the fixed covariates and the covariate "time."

Since the MMSE assessments were six months apart, it is reasonable to assume that, conditional on each subject and given the subject-specific parameters, the MMSE scores were independent. This assumption is frequently made in the AD literature; see, e.g., Mortimer et al. (1992).

We restrict our analysis to those participants who had at least five cognitive evaluations, in order to better capture the trajectory of cognitive decline. We also assume there to be the same number of follow-ups on all the subjects. This induces a balanced longitudinal data set. Our data includes forty-two subjects ($n = 42$), all

with five follow-up measurements ($m = 5$) over two and half years. In Table 6–1, we summarize the covariate information for these 42 subjects.

Table 6–1: Summary of the covariate information for the study subjects (n=42).

Continuous Predictors	Mean	SD
Education	12.38	3.15
Age at AD onset	66.74	9.02
Verbal	0.35	0.78
Non-verbal	-0.02	0.96
Disease duration (DISDUR)	3.43	2.36
Dichotomous Predictors (%)	**Frequency**	**Percentage**
Family history of dementia (FHDEM)	18	42.9
Sex (Male)	29	69
EXTRAP	5	11.9
APRAXIA	4	9.5

Education, age at start of the study, and disease duration, are in years. "Verbal" and "Non-verbal" are neurological scores: higher scores indicate a smaller deficit. EXTRAP (extrapyramidal signs) and APRAXIA are two dichotomous measurements which also serve as indicators of neurological deficit. Approximate age at onset was obtained from a caregiver (Mortimer et al. 1992). "Time" is in half-year units, with date of entry as origin. Thus, $\text{Time}_{i1} = 1$, $\text{Time}_{i2} = 2$, ..., $\text{Time}_{i5} = 5$ for the i^{th} subject. Boxplots of the continuous predictors are given in Figure 6–5.

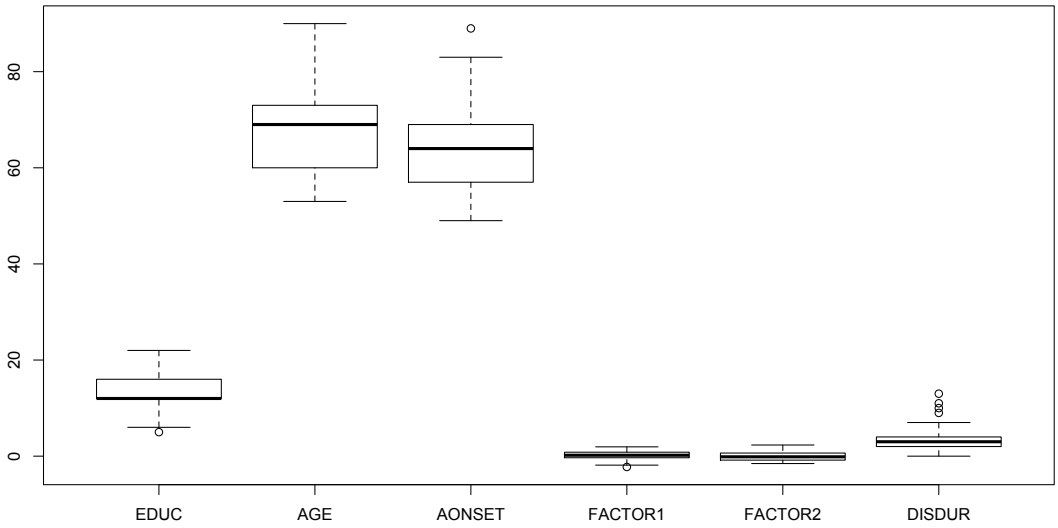


Figure 6–5: Continuous predictors in the model (EDUC (years of education), AONSET (age at AD onset), FACTOR1 ("Verbal" score), and FACTOR2 ("Non-verbal" scores)).

To model the effect of the selected covariates on the rate of decline, we included interactions between time and four of the covariates - Education, EXTRAP, APRAXIA and verbal score. An alternative approach would be to introduce random effect slopes.

The parameter estimates for the multi-path change-point model, as selected by the LASSO penalty function, are given in Table 6–2 and Table 6–3.

Table 6–2: MPLE of the parameters $(\beta_{10}, \boldsymbol{\beta}_1, \beta_{20}, \boldsymbol{\beta}_1, \boldsymbol{\alpha})$ (including the estimated standard deviations) in the selected model using the LASSO. "-" indicates "not selected".

Estimates	Change-point dist. $\widehat{\boldsymbol{\alpha}}$	Obs. dist. before change $(\hat{\beta}_{10}, \widehat{\boldsymbol{\beta}}_1)$	Obs. dist. after change $(\hat{\beta}_{20}, \widehat{\boldsymbol{\beta}}_2)$
Intercept	-	$15.33_{(0.13)}$	$12.77_{(0.47)}$
Time	-	$-2.18_{(0.12)}$	$-2.10_{(0.39)}$
EDUC	-	$-0.69_{(0.14)}$	$-0.28_{(0.32)}$
FHDEM	$-0.75_{(0.34)}$	$-0.79_{(0.16)}$	$2.90_{(0.69)}$
AGE	-	$-0.94_{(0.15)}$	$-2.37_{(0.42)}$
SEX	-	-	$4.28_{(0.36)}$
EXRAP	-	$2.04_{(0.15)}$	$-1.38_{(0.42)}$
APRAXIA	-	$-1.76_{(0.14)}$	-
Verbal	-	$1.54_{(0.18)}$	$5.88_{(0.52)}$
Nonverbal	-	$2.38_{(0.18)}$	$1.40_{(0.49)}$
DISDUR	-	$0.57_{(0.16)}$	-
TIME*EDUC	-	$-0.65_{(0.12)}$	$-0.31_{(0.55)}$
TIME*EXRAP	-	$1.24_{(0.13)}$	-
TIME*APRXIA	-	-	-
TIME*Verbal	-	$0.52_{(0.14)}$	$1.07_{(0.35)}$

Table 6–3: MPLE of the increments $(\alpha_{01}, \alpha_{02}, \alpha_{03}, \alpha_{04})$ in the change-point distribution (including the estimated standard deviations), in the selected model using the LASSO. "-" indicates "not selected'.

Estimates	$\hat{\alpha}_{01}$	$\hat{\alpha}_{02}$	$\hat{\alpha}_{03}$	$\hat{\alpha}_{04}$
	$-0.39_{(0.38)}$	$-3.01_{(0.30)}$	-	-

Because of the short follow-up and high variability between subjects, it is difficult to detect the change-points that were "*a priori*" anticipated, and the evidence for a change in our analysis is weak.

The estimated increments in the logit of the baseline α_{0i}, for $i = 1, 2, 3, 4$ are given in Table 6–3. Although they show a non-zero increment at the second follow-up time-point, an indication of a possible change at the second time-point, the remaining increments were estimated to be zero. However, because of the small number of observations, it is difficult to justify a change in the rate of decline. Although Joseph et al. (1999) found weak evidence for a change in the MMSE scores for subjects with AD, they took a fully Bayesian approach and did not include covariates in their model. We did not fit a second model without change-points since such a refit would have constituted a post-hoc analysis. Our model choice to begin with was based on the experience of researchers in the field of dementia.

For convenience, in the following discussion we refer to the period before the possible change as "early in the disease" and the period after the change as "late in the disease." This terminology is also appropriate, given the weak evidence for a change in our data set.

Other researchers have examined the effect of covariates on the rate of cognitive decline in patients with AD, using multi-path change-point models. Yu et al. (2012) used a Bayesian approach to fit a random change-point model. They found age, education and a certain genetic factor to be important in the timing of the onset of cognitive impairment and in the rate of decline before and after AD onset. Wilson et al. (2004) and Stern et al. (1999) claimed that the greater the number of years

of education, the more rapid the rate of decline in patients with AD late in the disease. In our model, "years of education" was not selected as a predictor early in the disease. However, later in the disease, higher education was associated with a more rapid decline. The interaction of education with time was also selected into the model. It is hypothesized that high education hides initial deficits, but like an insidiously eroding foundation, when collapse occurs, it is rapid.

Anderson et al. (1999) found a significant difference between the rate of decline in older men and women. We found that men decline at a slower rate both early and late in the disease.

Age at entry was also selected into our model; greater age at entry was associated with smaller mean MMSE scores at baseline and later in the disease. Bernick et al. (2012) and Wilson et al. (2012) found that older age at baseline was associated with a slower rate of decline. Verbal score and its interaction with time were also selected, which is to be expected since the MMSE score is based on verbal ability and higher verbal scores are associated with a slower rate of decline.

Although several of our findings are not consistent with those of some other researchers, there is no consistency in the pre-existing AD literature on the predictors of cognitive decline. Further, in view of the small numbers of study subjects and follow-ups, our results should, perhaps, be viewed with caution.

6.1 Discussion

In the data we analyzed here, the choice of the penalty function influenced the overall conclusion. In particular, with a small number of subjects and follow-ups the choice of penalty function plays an important role in variable selection. The goal

of the analysis (the observation distribution or the change-point distribution) affects our choice of penalty function.

Based on our data analysis and simulations we make the following suggestions for the design of studies of decline in AD, with a goal of detecting highly influential covariates:

1. If the aim is to make inference about the observation distributions and their influential covariates, increasing the number of subjects seems to be more important than increasing the number of follow-ups.
2. To make inference about the change-point distribution, we have to consider two aspects of the model: the covariate effects and the baseline hazard of the change. Increasing the number of follow-ups increases the accuracy of the inference about the covariate effects, but it also increases the number of parameters (the increments in the logit of the baseline hazard). Hence, a large m may induce more variability in the estimators of the baseline hazard. Therefore, in this case both the number of subjects and the number of follow-ups have to be large.

CHAPTER 7
Concluding Remarks

In this thesis, we took a penalized likelihood approach to do variable selection in a multi-path change-point model, a problem that has not previously been addressed in the literature. In fact, we were able to do variable selection and estimation simultaneously. We proved that our method has good large sample properties and examined its performance for small sample sizes using simulation studies.

A feature of our model is that it allows the observation distributions as well as the change-point distribution to change from subject to subject through covariate values. We model the change-point distribution via a proportional odds hazard function with a flexible piece-wise constant baseline odds.

Our method is particularly useful when the amount of available covariate information on each subject (or path) is large. For example, in clinical trials, often detailed covariate information is collected and there may be an unknown delay for the treatment to take effect; such scenarios are conveniently captured using change-point models.

However, in each application there may be computational difficulties. The presence of increments in the logit of the baseline odds results in a non-concave objective function requiring a more time-consuming computational procedure. Since our (modified) EM algorithm depends on the initial value, it is not guaranteed to produce either global or local maximum likelihood estimates. The Newton–Raphson method

we used in the EM algorithm can lead to highly variable estimates, as a result of the non-smooth objective function. Using ridge regression eases the computational difficulty but still does not completely resolve the problem.

Change-point problems in two and three dimensions entail boundary determination and unless the boundary is of simple form the modeling required is likely to be very difficult.

Our method for variable selection in multi-path change-point problems suggests several new directions for future research.

1. Because of the computational burden, we introduced a single tuning parameter for all the penalty functions in the model except for the ridge penalty function. One could, however, use different tuning parameters for different groups of model parameters. In this extension, a grid of vectors of tuning parameter candidates could be used and the BIC penalty function would be defined on each point of the grid. In our model, we consider the parameters of four components: the observation distributions before and after the change-point, the change-point distribution regression model, and the baseline hazard of change. Thus we would use a different penalty function for each parameter vector component.
2. We could allow for multiple change-points per subject.
3. One could consider multivariate distributions for the observations before and after the change-point, allowing correlated random variables over time.
4. A natural extension of our fixed effect models would be the mixed models that include random effects for the regression parameters.

5. We took a frequentist approach. A Bayesian approach to the problem is an obvious alternative.

REFERENCES

[1] Akaike H. Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory* 1973; Akademinai Kiado: 267–281.

[2] Allman ES, Matias C, Rhodes JA. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics* 2009; **37(6A)**: 3099–3132.

[3] Andersen K, Launer LJ, Dewey ME, Letenneur L, Ott A, Copeland JR, Dartigues JF, Kragh-Sorensen P, Baldereschi M, Brayne C, Lobo A, Martinez-Lage JM, Stijnen T, Hofman A. Gender differences in the incidence of Alzheimer's disease and vascular dementia: the EURODEM Studies. *Neurology* 1999; **53(9)**: 1992–1997.

[4] Anraku K. An information criterion for parameters under a simple order restriction. *Biometrika* 1999; **86(1)**: 141–152.

[5] Asgharian M. Modeling covariance in multi-path change-point problems. *Doctoral thesis, Department of Mathematics and Statistics, McGill University, Montreal, Quebec, Canada* 1998.

[6] Asgharian M. On the singularity of the Fisher information matrix and multipath change-point problems. *Teoriya Veroyatnostei i ee Primeneniya* 2013; **58 (3)**.

[7] Asgharian M, Wolfson DB. Modeling covariates in multipath change-point problems: modeling and consistency of the MLE. *The Canadian Journal of Statistics* 2001; **29(4)**: 515–528.

[8] Beckage B, Joseph L, Bélisle P, Wolfson DB, Platt WJ. Bayesian change-point analyses in ecology. *New Phytologist* 2007; **174(2)**: 456–467.

[9] Bennett DA, Wilson RS, Schneider JA, Evans DA, Beckett LA, Aggarwal NT, Barnes LL, Fox JH, Bach J. Natural history of mild cognitive impairment in older persons. *Neurology* 2002; **59**: 198–205.

[10] Bernick C, Cummings J, Raman R, Sun X, Aisen P. Age and rate of cognitive decline in Alzheimer's disease: implications for clinical trials. *Archives of Neurology* 2012; **69(7)**: 901–905.

[11] Bélisle P, Joseph L, MacGibbon B, Wolfson DB, Du Berger R. Change-point analysis of neuron spike train data. *Biometrics* 1998; **54(1)**: 113–123.

[12] Bunea F. Honest variable selection in linear and logistic regression models via l_1 and $l_1 + l_2$ penalization. *Electronic Journal of Statistics* 2008; **2**: 1153–1194.

[13] Chan KS, Tsay RS. Limiting properties of the least squares estimator of a continuous threshold autoregressive model. *Biometrika* 1998; **85(2)**: 413-426.

[14] Chen SS, Donoho DL, Saunders MA. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing* 1998; **20(1)**: 33–61.

[15] Dempster AP, Nan ML, Donald BR. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 1977; **39(1)**: 1–38.

[16] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 2001; **96(456)**: 1348–1360.

[17] Fan J, Li R. Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics* 2002; **30(1)**: 74–99.

[18] Faragge D, Simon R. A simulation study of cross-validation for selecting an optimal cut point in univariate survival analysis. *Statistics in Medicine* 1996; **15(20)**: 2203–2213.

[19] Feder PI. On asymptotic distribution theory in segmented regression problems–identified case. *The Annals of Statistics* 1975a; **3(1)**: 49–83.

[20] Feder PI. The log likelihood ratio in segmented regression. *The Annals of Statistics* 1975b; **3(1)**: 84–97.

[21] Folstein M, Folstein SE, McHugh PR. "Mini-Mental Stat" a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* 1975; **12(3)**: 189–198.

[22] Green PJ, Silverman BW. Nonparametric regression and generalized linear models. *London: Chapman and Hall* 1994.

[23] Hackl P, Westlund AH. Statistical analysis of "structural change": An annotated bibliography. *Empirical Economics* 1989; **14(2)**: 167–192.

[24] Hall CB, Derby C, LeValley A, Katz MJ, Verghese J, Lipton RB. Education delays accelerated decline on a memory test in persons who develop dementia. *Neurology* 2007; **69(17)**: 1657–1664.

[25] Hall CB, Lipton RB, Sliwinski M, Stewart WF. A change-point model for estimating the onset of cognitive decline in preclinical Alzheimer's disease. *Statistics in Medicine* 2000; **19(11-12)**: 1555–1566.

[26] Hall P, Simar L. Estimating a change-point, boundary, or frontier in the presence of observation error. *Journal of the American Statistical Association* 2002; **97(458)**: 523–534.

[27] Harchaoui Z, Lévy-Leduc C. Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association* 2010; **105(492)**: 1480–1493.

[28] Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference, and prediction* Second Edition, Springer Verlag, 2009.

[29] Hebert LE, Wilson RS, Gilley DW, Beckett LA, Scherr PA, Bennett DA, Evans DA. Decline of language among women and men with Alzheimer's disease. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 2000; **55(6)**: 354–P361.

[30] Hinkley DV. Inference about the change-point in a sequence of random variables. *Biometrika* 1970; **57(1)**: 1–17.

[31] Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970; **12(1)**: 55–67.

[32] Hunter DR, Li R. Variable selection using MM algorithms. *Annals of Statistics* 2005; **33(4)**: 1617–1642.

[33] Joseph L. The multi-path change-point. *Doctoral thesis, Department of Mathematics and Statistics, McGill University, Montreal, Quebec, Canada* 1989.

[34] Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. Adaptive mixtures of local experts. *Neural Computation* 1991; **3(1)**: 79–87.

[35] Joseph L, Wolfson DB, du Berger R, Lyle R. Analysis of panel data with change-points. *Statistica Sinica* 1997; **7(3)**: 687–703.

[36] Joseph L, Wolfson DB. Estimation in multi-path change-point problems. *Communications in Statistics-Theory and Methods* 1992; **21(4)**: 897–914.

[37] Joseph L, Wolfson DB. Maximum likelihood estimation in the multi-path change-point problem. *Annals of the Institute of Statistical Mathematics* 1993; **45(3)**: 511–530.

[38] Joseph L, Vandal AC, Wolfson DB. Estimation in the multi-path change-point problem for correlated data. *Canadian Journal of Statistics* 1996; **24(1)**: 37–53.

[39] Joseph L, Wolfson DB, Bélisle P, Brooks JO, Mortimer J, Tinklenberg J, Yesavage J. Taking account of between-patient variability when modeling decline in Alzheimer's disease. *American Journal of Epidemiology* 1999; **149(10)**: 963–973.

[40] Khalili A. New estimation and feature selection methods in mixture-of-experts models. *The Canadian Journal of Statistics* 2010; **38(4)**: 519–539.

[41] Khalili A, Chen J. Variable selection in finite mixture of regression models. *Journal of the American Statistical Association* 2007; **102(479)**: 1025–1038.

[42] Klaassen CAJ, Lenstra AJ. Vanishing Fisher Information. *Acta Applicandae Mathematicae* 2003; **78**: 193–200.

[43] Khodadadi A, Asgharian M. Change-point problem and regression: an annotated bibliography. *COBRA Preprint Series* 2008; Paper 44.

[44] Lange N, Carlin B, Gelfand AE. Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers (with discussion). *Journal of the American Statistical Association* 1992; **87(419)**: 615–626.

[45] Liang F, Wong WH. Evolutionary Monte Carlo: Applications to C_p model sampling and change-point problem. *Statistica Sinica* 2000; **10(2)**: 317–342.

[46] Mortimer JA, Ebbitt B, Jun SP, Finch MD. Predictors of cognitive and functional progression in patients with probable Alzheimer's disease. *Neurology* 1992; **42(9)**: 1689–1696.

[47] Mosteller F. On pooling data. *Journal of the American Statistical Association* 1948; **43(242)**: 231–242.

[48] Ninomiya Y. Information criterion for Gaussian change-point model. *Statistics and Probability Letters* 2005; **72(3)**: 237–247.

[49] Park MY, Hastie T. Penalized logistic regression for detecting gene interactions. *Biostatistics* 2008; **9(1)**: 30–50.

[50] Petruccelli J, Davies N. A portmanteau test for self-exciting threshold autoregressive-type nonlinearity in time series. *Biometrika* 1986; **73(3)**: 687–694.

[51] Pocock SJ, Cook DG, Shaper AG. Analysing geographic variation in cardiovascular mortality: methods and results. *Journal of the Royal Statistical Society. Series A (General)* 1982; **145(3)**: 313–341.

[52] Reed WJ. Determining changes in historical forest fire frequency from a time-since-fire map. *Journal of Agricultural, Biological, and Environmental Statistics* 1998; **3(4)**: 430–450.

[53] Rothenberg TJ. Identification in parametric models. *Econometrica: Journal of the Econometric Society* 1971; **39(3)**: 577–591.

[54] Schwarz G. Estimating the dimension of a model. *The Annals of Statistics* 1978; **6(2)**: 461–464.

[55] Smith, A. F. M. A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika* 1975; **62(2)**, 407–416.

[56] Stern Y, Albert S, Tang MX, Tsai WY. Rate of memory decline in Alzheimer's disease is related to education and occupation: cognitive reserve?. *Neurology* 1999; **53(9)**: 1942–1947.

[57] Tibshirani R. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)* 1996; **58(1)**: 267–288.

[58] Tibshirani R, Wang P. Spatial smoothing and hot spot detection for CGH data using the fused LASSO. *Biostatistics* 2008; **9(1)**: 18–29.

[59] Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2005; **67(1)**: 91–108.

[60] Tong H. Threshold models in non-linear time series analysis. *Lecture notes in statistics* 1983; **21**. Springer-Verlag.

[61] Tsay R. Unit root tests with Threshold Innovations. *Preprint, University of Chicago* 1997.

[62] Wahba G. Spline models for observational data. *SIAM* 1990; **59**.

[63] Wang H, Li R, Tsai CL. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* 2007; **94(3)**: 553–568.

[64] Wilson RS, Hebert LE, Scherr PA, Dong X, Leurgens SE, Evans DA. Cognitive decline after hospitalization in a community population of older persons. *Neurology* 2012; **78(13)**: 950–956.

[65] Wilson RS, Li Y, Aggarwal NT, Barnes LL, McCann JJ, Gilley DW, Evans DA. Education and the course of cognitive decline in Alzheimer's disease. *Neurology* 2004; **63(7)**: 1198–1202.

[66] Wilson RS, Bennett DA, Beckett LA, Morris MC, Gilley DW, Bienias JL, Scherr PA, Evans DA. Cognitive activity in older persons from a geographically defined population. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 1999; **54(3)**: 155–P160.

[67] Young D S. Mixtures of regressions with change-points. *Statistics and Computing* 2012; **24(2)**: 265–281.

[68] Yu L, Boyle P, Wilson RS, Segawa E, Leurgans S, De Jager PL, Bennett DA. A random change-point model for cognitive decline in Alzheimer's disease and mild cognitive impairment. *Neuroepidemiology* 2012; **39(2)**: 73–83.

[69] Zou H. The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association* 2006; **101(476)**: 1418–1429.

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

[71] Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics* 2008; **36(4)**: 1509–1533.