8-2011

Regression methods for survival and multistate models.

Farida Mostajabi
University of Louisville

Recommended Citation

Mostajabi, Farida, "Regression methods for survival and multistate models." (2011). *Electronic Theses and Dissertations.* Paper 1014.
https://doi.org/10.18297/etd/1014

REGRESSION METHODS FOR SURVIVAL AND MULTISTATE MODELS

By

Farida Mostajabi
M.Sc., Shiraz University of Medical Sciences, Iran, 2005

A Dissertation
Submitted to the Faculty of the
Graduate School of University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, KY

August 2011

REGRESSION METHODS FOR SURVIVAL AND MULTISTATE MODELS

By

Farida Mostajabi
M.Sc., Shiraz University of Medical Sciences, Iran, 2005

A Dissertation Approved on

June 23, 2011

by the following Dissertation Committee:

Dissertation Director (Somnath Datta)

Dissertation Co-Director (Susmita Datta)

Adel Elmaghraby

Seongho Kim

Maiying Kong

DEDICATION

This dissertation is dedicated to my husband

Dr. Masoud Ghaffari

who has given me invaluable encouragement and endless support

ACKNOWLEDGMENTS

I would like to thank all people who have helped and inspired me during my doctoral study.

I specially, would like to thank Dr. Somnath Datta, my dissertation director, and Dr. Susmita Datta ,my dissertation co-director, for their guidance and patience. They were both motivating, encouraging, and enlightening. Their advice was immensely beneficial and reflected in the overall quality of this dissertation.

I would also like to thank the other committee members, Dr. Adel S. Elmaghraby, Dr. Maiying Kong and Dr. Seongho Kim, for their comments and assistance over the past two years.

My gratitude is also extended to Dr. Guy Brock . He was one of the first friendly faces to greet me when I began this doctoral program and has always been a tremendous help no matter the task or circumstance.

Generous financial support made my doctoral studies possible. My first two years study was supported by Scholarship provided by University of Louisville. The last two years, I was supported partly by Elsevier through Statistics and Probability Letters (SPL). My research was supported in part by Dr. Susmita Datta's grants from National Science Foundation and the National Institutes of Health. Dr. Somnath Datta's NSA grant was another source for the last few months which helped me to overcome financial difficulty under economic recession.

My classmates were sources of support and joy during these four years. Special thanks goes to Juliet, Alex, Vasyl, Jie, Gene, Frank, Mourad, Ashok, Jieru, Ming, Ling Ping, Simin and Daug. In all of the ups and downs that came my way during the "dissertation years", I knew that I had the support of Nicole. I am very happy that, in many cases, my friendship with her have extended well beyond our shared time in the program.

I am very glad and proud of making wonderful friends in US. Especially I am thinking to my best friends: Romaneh, Mehdi, Bita & Farzin, who made me feel the US as the second home. I will always cherish the wonderful time spent together with them. I would also like to thank my cousins: Mehran and Vahid for their tremendous hospitality in the past two years.

I would also like to thank Siavash who inspired me to pursue my education in US and offered all kinds of support to make this happen.

My deepest gratitude goes to my family for their unconditional love and support throughout my life. They deserve far more credit than I can ever give them. I am proud of them and love them very much.

Finally, I would like to express my thanks to my husband, Masoud, for his love, understanding and supporting to pursue the doctoral degree. I would have never made it without his patience. Masoud, thank you for the countless ways in which you have supported me during this and other endeavors, and for being someone who just knew.

ABSTRACT

REGRESSION METHODS FOR SURVIVAL AND MULTISTATE MODELS

Farida Mostajabi

June 23, 2011

A common research interest in medical, biological, and engineering research is determining whether certain independent variables are correlated with the survival or failure times. Standard statistical techniques cannot usually be applied for failure-time data due to the lack of complete data or in other word, due to censoring. From a statistical perspective, the study of time to event data is even more challenging when further complexities such as high dimensionality or multivariablity is added to the model.

In this dissertation, we consider the predicating patient survival from proteomic profile of patient serum using matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) data of non-small cell lung cancer patients. Due to much larger dimension of features in a mass spectrum compared to the study sample size, traditional linear regression modeling of survival times with high number of proteomic features is not feasible. Hence, we consider latent factor and regularized/penalized methods for fitting such models in order to predict patient survival from the mass spectrometry features. Extensive numerical studies involving both simulated as well as real mass spectrometry data are used to compare four popular regression methods, namely, partial least squares (PLS), sparse partial least square (SPLS), least absolute shrinkage and selection operator (LASSO) and elastic net regularization, on processed spectra. Right censoring is handled through a residual based multiple imputation. Overall, more complex methods such as the elastic net and SPLS result in better performances provided the operational parameters are chosen carefully via cross validation. For survival time prediction, we recommend using the elastic net based on a selected set of features.

As a type of multivariate survival data, multistate models have a wide range of applications. Most of the existing regression approaches to analyze such data are based on parametric and semi-parametric procedures in which one should rely on specific model structures. In this dissertation, We construct non-parametric regression estimators of a number of temporal functions in a multistate system based on a univariate continuous baseline covariate. These estimators include state occupation probabilities, state entry, exit and waiting (sojourn) times distribution functions of a general progressive (e.g. acyclic) multistate model. The data are subject to right censoring and the censoring mechanism is explainable by observable covariates that could be time dependent. The resulting estimators are valid even if the multistate process is non-Markov. The performance of the estimators is studied using a detailed simulation. We illustrate our estimators using a data set on bone marrow transplant patients. Finally, some extension of the proposed methods to more general case with multivariate covariates are presented along with plans for future developments.

TABLE OF CONTENTS

PAGE

CHAPTER

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

CHAPTER I

INTRODUCTION

A common research question in medical, biological, or engineering research is to determine whether or not certain independent variables are correlated with the survival or failure times. In general, regression models are used to study the conditional distribution of the dependent variable given the independent variables. In survival case, dependent variable is the failure time or time to event of interest such as death, development or progression of a disease and etc.

Standard statistical techniques cannot usually be applied for failure-time data due to the lack of complete data or in other word, due to censoring issue. Censoring is a form of missing data problem arises in time to event data. The most common case of censoring is what is referred to as right censored data where the event time is known only to be greater than a certain time (e.g. lost to follow-up). The presence of censoring poses major challenges to regression modeling.

The semi-parametric Cox proportional hazard regression model (Cox, 1972) was a breakthrough in developing a flexible method of regression for censored data. In a Cox model, the effect of an independent variable on the hazard rate is assumed to be multiplicative. Even though Cox model is very flexible, the proportional hazards assumption may not hold in all circumstances. The accelerated failure time (AFT) models are another class of regression models that can be an alternative to the Cox models (Wei, 1992). The general AFT models are linear models for the logarithm or a known monotone transformation of the survival time (Kalbfleisch and Prentice, 1980). Supposed T denotes a time to certain event. The accelerated failure time (AFT) model is of the same form as usual linear regression model:

$$log(T) = \beta' Z + \epsilon,$$

where the error term ϵ is independent of the p-dimensional covariate vector Z and its distribution is left unspecified. The results of AFT models are interpretable more easily than the proportional hazards model in certain applications.

1.1. Survival Models with Increased Complexity

From a statistical perspective, study of time to event data is even more challenging when further complexities such as *multivariablity* or *high dimensionality* are added to the model. In the next subsection, we focus on certain type of multivariable survival which is referred to multistate models. Afterward, we introduce the problem of high dimensionality that mostly arises in biomedical studies and the corresponding difficulties to predict survival times.

1.1.1. Multistate Models

Multistate models are certain type of multivariate survival data that are a natural extension of simple survival models. These models allow subjects to move through a succession of states and they are particularly useful for describing the complexities of disease processes in which each state corresponding to a certain health condition (e.g. alive and disease-free, alive with recurrence and dead). The resulting data contains information about the transition times and the states occupied. Transitions between states can be reversible or irreversible while states can be either absorbing or transient. An absorbing state is a state from which further transitions cannot occur while a transient state is a state that is not absorbing. Graphically, multi-state models may be illustrated using diagrams with boxes representing the states and with arrows between the states representing the possible transitions. Approaches for traditional survival analysis may be viewed as a two-state model '0: alive, 1: dead' (see Figure 1.1).

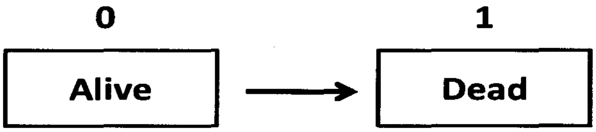


Figure 1.1. A two-state Alive-Dead model

Two key questions in the multistate models are: what is the probability that a subject is in a specific state at certain time point or what is the hazard (rate) that an individual moves from one state to another. To answer these equations, we should estimate the state occupation probabilities and transition hazards respectively. Distribution functions of state entry/exit and waiting times are also useful quantities. Beside estimation and hypothesis tests for these quantities, analysis of regression models where these quantities related to explanatory variables is of interest.

As in survival analysis, empirical calculations of these quantities are not possible due to censoring issue. Application of parametric methods in multistate models developed over the last thirty years (Lagakos, 1976; Beck, 1979; Kay, 1982; Sacks & Chiang, 1977; Wu, 1982; Klein *et al.*, 1984; Andersen & Keiding, 2002; Plevritis *et al.*, 2007 and so on).

Estimation at the individual level may be investigated through covariates. Regression models for multistate processes have traditionally been formulated in terms of the transition intensities of the model using the Cox model in a Markovian framework (Andersen *et al.*, 1993). Shu and Klein (2005) discussed an application of the additive hazards models (Aalen, 1989; Lin & Ying, 1994) to multistate data assuming the Markovian structure. Andersen and Klein (2007) presented a general approach to the problem for the state occupation or transition probabilities in a multistate model directly based on pseudo-values from a jackknife statistic constructed from non-parametric estimators for the probability in question. These pseudo-values were used as outcome variables in a generalized estimating equation to obtain estimates of model parameters. When proportional hazards models or additive hazards models are assumed to model covariate effects for the transition intensities then a model is induced for the state and transition probabilities. These models in most cases produce highly nonlinear and complex effects of the covariates on these probabilities that are difficult to interpret. Most of the excising methods are based on parametric and semi-parametric modeling of the transition hazards with Morkovian framework assumption. For a partial review of semi-parametric alternatives to the Cox models see Andersen and Keiding (2002).

Generally speaking, while parametric and semi-parametric methods produce relatively precise inference for the effects of covariates under the correct model, their performance under incorrect model assumptions is questionable and they have their own shortcomings. Furthermore, in application it is an unapproachable task to determine which of the models to employ in analyzing a particular dataset. This is one convincing reason why a fully non-parametric approach is preferable even though such a formulation is often difficult with time to event data. The situation with multistate models that generalize the traditional survival setup is even more challenging and as such only a limited number of regression approaches exist to analyze such models. Nevertheless, only non-parametric answers represent truly empirical (or evidence based) calculations. They can at least serve as a guideline to the shape of the regression functions on certain marginal aspects of the system even if a semi-parametric or parametric calculation is ultimately performed. Doksum and Yandell (1982) made similar points with compelling comparative illustrations of non-parametric calculations versus semi-parametric calculations using the well known Stanford heart transplant data.

The number of papers dealing with non-parametric regression for multistate models is quite limited. Majority of them deal with the survival setup that can be regarded as the simplest multistate system. Beran (1981) studied a conditional Kaplan-Meier estimator obtained with regression weights using either a nearest neighborhood approach or a kernel approach; also, see Doksum and Yandell (1982) who suggested using non-parametric methods for this problem. Theoretical properties of these estimators and their generalizations have been further studied by Dabrowska (1987, 1989), Li and Doss (1995), McKeague and Utikal (1990), Li and Datta (2001) etc. When proportional hazards models or additive hazards models are assumed to model covariate effects for the transition intensities, then a model is induced for the state and transition probabilities. These models in most cases produced highly nonlinear and complex effects of the covariates on these probabilities that were difficult to interpret Recently, Andersen *et al.* (2003) and Andersen and Klein (2006) studied the effect of covariates in a multistate model using a hybrid approach of combining some non-parametric calculation followed by semi-parametric ones. This approach, however, many not produce regression function estimators of the marginal quantities under study; furthermore, the theoretical modeling framework necessary for the validity of this approach is not very clear.

Smoothing techniques are often used to produce non-parametric estimators. These offer useful alternatives to the non-parametric likelihood based approaches since a full likelihood specification in a multistate model is often difficult (and sometimes impossible without additional structural assumptions). There are various approaches to produce non-parametric smoothed estimates. The simplest and flexible methods are the kernel-based procedures (Nadaraya, 1964; Watson, 1964). A common difficulty with smoothing methods is the selection of the underlying tuning parameter that represents the smoothness of the estimators. As for example, an objective data based choices is needed to select the bandwidth in the kernel method. While various choice have been proposed often based on asymptotic consideration, a completely satisfactory solution to the problem of tuning parameter selection is still largely unavailable.

As discussed earlier, development of fully non-parametric regression estimators of state occupation probabilities is of interest. Such inference procedures will be more robust than their parametric and semi-parametric counterparts which generally rely on specific model structures. These types of unified and systematic methodological development do not currently exist for general multistate models.

1.1.2. High Dimensional Data

Another category of survival regression arises in biomedical research such as recently developed genomic and proteomic studies. These technologies are often used to identify genes and proteins that may have a functional role in specific phenotypes. From this prospective, one might be interested to model the relationship between the genes or features and survival outcome. This can be particularly challenging since the main characteristic of such data is that the number of covariates or features (genes or proteins) are considerably larger than the number of samples (individuals). Dimensionality is even much larger in proteomic studies. Hence classical survival methods such as Cox model cannot be applied directly and specific modifications are required.

Many methods have been developed that are better suited for high-dimensional settings on the basis of Cox's proportional hazards regression (Pawitan *et al.*, 2004; van Houwelingen *et al.*, 2006; Bøvelstad *et al.*, 2007). Moreover, Li and Luan (2003) were investigated the L_2 penalized estimation of the Cox model in the high-dimensional low-sample size settings and applied their method to relate the gene expression profile to survival data. One limitation of the L_2 penalized estimation of the Cox model is that it uses all the genes in the prediction and does not provide a way of selecting relevant genes for prediction. Two years later, Gui & Li (2005) proposed to use the L_1 penalized estimation for the Cox model to select genes that are relevant to patients' survival. Tibshirani (2009) introduced Cox univariate shrinkage model in which the features are entered into the model based on the size of their Cox score statistics. This method assumes that the features are independent in each risk set.

On the other hand, semi-parametric estimation in the AFT model with an unspecified error distribution has been studied extensively in the literature. However, there are only few publications on employment of the AFT model in high dimensional setting. Huang *et al.* (2006) used the LASSO and the threshold-gradient-directed regularization along with AFT model for estimation and variable selection. Datta *et al.* (2007) considered predicting survival using AFT model along with PLS and LASSO. Engler *et al.* (2009) adapted the elastic net approach for variable selection both under the Cox proportional hazards model and under AFT model. Most of these studies adopted for application in microarray data. However, there is no such study that we are aware of in the context of proteomic study in which one is exposed with even more dimensionality. One clear example is Mass Spectrometry data. Before discussing the specific aims of our research, a brief review of Mass spectrometry technology is given.

A proteome is the collection of proteins that make up a cell (or organism) under a specific set of conditions at a specific time. Studying the amount of each protein present at any time has become more important as scientists attempt to learn which proteins are involved in important cellular functions. Mass spectrometry (MS) is an emerging field of interest in biomedical research. It is basically used to visualize the distribution of proteins within a tissue sample or from bodily fluids such as urine, plasma, serum, etc. In contrast to traditional approaches that examine one or a few proteins at a time, this technology actually profile hundreds of proteins derived from the samples simultaneously. Only a small sample is needed to this end and results can be obtained in very short amount of time.

The use of mass spectrometry as a diagnostic tool and identification of proteomic biological markers has risen extensively and has been demonstrated great promise in recent years. This has led to the discovery of a large number of proteins and protein profiles associated with various types of diseases. Early articles in this area include Stoeckli *et al.* (2001), Petricoin *et al.* (2002), Adam *et al.* (2002), Aebersold *et al.* (2003), Liotta *et al.*

(2003) and Rai *et al.* (2004). Mass spectrometry for the protein analysis consists of diverse platforms. Matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF), and Surface enhanced laser desorption/ionization-time of flight (SELDI-TOF) mass spectrometry are two basic forms of the technology, however the former technique is the most commonly applied technique to clinical and biological problems. The basic operating method of both technologies is similar. The biological samples (such as blood serum or plasma) first mixed with an energy absorbing matrix (EAM) which acts as a proton donor or acceptor. This mixture is crystallized onto a metal plate. In the SELDI technique, additional chemistry is added on the plate to unite specific classes of proteins. The mixture then subjected to pulse laser radiation. This causes the vaporization of the matrix crystals and produces ions which are directed into a flight tube through application of an electric field. The mass of an ion is measured by the time it takes to hit the detector located at the end of the tube. The time of flight is then converted to corresponding mass-to-charge values using a quadratic transformation. A typical proteomic experiment consists of the sequentially recorded numbers of ions strike the detector (intensity value) along with corresponding mass to charge ratio (m/z) which is typically a huge volume of data. The output of mass spectrometer which is referred to a spectrum are often displayed as a graph showing the relative abundance representing an unknown number of protein peaks associated with protein mass to charge ratios. The unit of mass to charge measurements is Dalton (D).

A typical data set contains hundreds of spectra; each spectrum contains tens of thousands of intensity measurements representing an unknown number of protein peaks which are the key feature of interest. These measurements couple with substantial noise. The first attempt to analyze such data is to do some preprocessing steps to identify the locations of peaks and to quantify their sizes precisely. Several studies have shown that mistakes in the preprocessing of the data can bias the biological interpretation of the study. Sorace & Zhan (2003) showed that inadequate preprocessing has a negative effect on the extraction of clinically useful information. Baseline correction, denoising, normalization, peak detection and peak alignment are among the most common preprocessing steps. Variety set of preprocessing approaches has been proposed in literature and it is still ongoing field of research. A comprehensive review of all the different possibilities for each preprocessing step is beyond the scope of this chapter. However, several comparisons can be found in Cruz-Marcelo (2008) and Emanuele *et al.* (2009).

As discussed earlier, one of the main challenges to deal with such data is the issue of high dimensionality in the context of regression modeling. In fact, the number of spectra available in such studies is considerably smaller than the length of the individual spectra $(p \geq n)$. Including all the features in the predictive model introduces noise and is expected to poor predictive performance. In such cases, overfitting is also likely to happen. Hence, classical statistical approaches are not appropriate and special techniques such as variable selection or dimension reduction are required. Besides the high-dimensionality, the features are often highly correlated, which creates the problem of high collinearity. To deal with the problem of collinearity, the most widely used approach is the penalized partial likelihood. Some of the most common techniques for variable selection and dimension reduction are discussed below.

Introduced by Herman Wold (1966), Partial Least Squares (PLS) regression is a technique that has been an alternative to ordinary least squares in high dimensional setting in several areas of scientific research. This method is particularly useful for constructing predictive models when the predictors are many and highly collinear. The general idea of PLS is to try to extract the latent factors that account for most of the variation in the response while modeling the response well. In order to specify latent variables, PLS iteratively finds weight vectors such that X space has the highest covariance with Y. There are several variants of the algorithms for obtaining the PLS estimators. Boulesteix & Strimmer (2007) reviewed both the theory underlying PLS as well as a host of bioinformatics applications of PLS. They provide a systematic comparison of the PLS approaches currently employed.

Sparse partial least squares (SPLS) developed by Chun *et al.* (2010) is another dimension reduction technique. This technique produces sparse linear combinations of the original predictors and achieves both dimension reduction and variable selection simultaneously. This is simply achieved by PLS regression using the selected variables. In fact, the number of SPLS latent components is limited by number of observations, but the actual number of variables that makes up the latent components can exceed n.

Regularized and penalized methods are another group of techniques that has gained great popularity in recent years. Penalized estimation methods shrink the estimates of the regression coefficients towards zero relative to the maximum likelihood estimates. The purpose of this shrinkage is to prevent over- fitting arising due to either collinearity of the covariates or high-dimensionality. Least absolute shrinkage and selection operator (LASSO) was proposed by Tibshirani (1996). The LASSO is a penalized least squares method imposing an L_1-penalty on the regression coefficients. The LASSO does both continuous shrinkage and automatic variable selection simultaneously and it has sparse representation. In the usual regression set-up, the LASSO minimizes the residual sum of squares subject to the sum of the absolute values of the coefficients being less than a constant (L_1 constraint). The LASSO estimator is the value that minimizes $\sum_{i=1}^{n}(Y_i - x_i'\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$, where λ is the penalty parameter. The L_1 penalty causes the continuous shrinkage and variable selection simultaneously. Hence, it can be used to select a suitable set for the efficient prediction of a response variable method.

Recently, Efron *et al.* (2004) proposed the least angle regression (LARS) procedure for variable selection in the linear regression setting. The LARS selects predictors based on their correlation between the predictor and the current residuals. In this algorithm, one starts with the trivial model with all coefficients set at zero. Variables are added one at a time that are most correlated with the residuals at the previous step. The number of variables to be included in the model corresponds to a selection of the tuning parameters. Efron (2004) further showed that LARS can be modified to provide solutions for LASSO. With this powerful algorithm, LASSO can be extended to perform subset selection in the high-dimension and low-sample settings.

The Elastic net approach proposed by Zou *et al.* (2005) is newer regularization and variable selection method that combines L_1 and L_2 penalties. Such procedure tends to give a result with fewer regression coefficient set to zero in a pure L_1 setting, and more shrinkage of the other coefficients. In this method, strongly correlated predictors tend to be in or out of the model together. The elastic net is particularly useful when the number of predictors is much bigger than the number of observations. Elastic net simultaneously does automatic variable selection and continuous shrinkage and has the ability to do grouped selection. The estimator minimizes

$$\sum_{i=1}^{n}(Y_i - x_i'\beta)^2 + \lambda_2\sum_{j=1}^{p}|\beta_j|^2 + \lambda_1\sum_{j=1}^{p}|\beta_j| \,.$$

Although the methods discussed have been proved to be useful in the past when the number of covariates exceeds the number of observations, we are not aware of any study that considered extreme situations encountered in proteomic analysis. Thus, one of the main purposes is to study the performances of these methods when the sample size appears to be hopelessly small compared to the number of covariates.

1.2. A Summary of the Dissertation

The rest of this dissertation is organized as follows. In Chapter II, predicting survival times of patients with the proteomic profile of patient serum using matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) data of non-small cell lung cancer patients is considered. Extensive numerical studies involving both simulated as well as real mass spectrometry data are used to compare four popular regression methods, namely, partial least squares (PLS), sparse partial least square (SPLS), least absolute shrinkage and selection operator (LASSO) and Elastic net regularization, on processed spectra. Right censoring is handled through a residual based multiple imputation.

A novel non-parametric regression estimation approach of state occupation probabilities is laid out in Chapter III, using kernel estimators along with inverse probability of censoring reweighting. Regression functions are studied based on one covariate at a time.We also developed valid non-parametric estimators of entry/exit and waiting time distributions conditional on a given value of X. The global performance of these estimators are investigated using Monte Carlo simulations. The application to a bone marrow transplant data is provided.

Chapter IV describes the methodology of non-parametric regression estimation of state occupation probability using kernel estimators including more than one covariate in regression function. We also developed non-parametric estimators of entry/exit time distributions conditional on a given value of X.The global performance of these estimators are investigated using Monte Carlo simulations. We conclude the dissertation with some concluding remarks and plans for my future research direction in Chapter V.

It is my hope that this dissertation research significantly advances the area of survival prediction in Mass Spectrometry data as well as non-parametric regression for multistate models.

CHAPTER II

PREDICTING PATIENT SURVIVAL FROM PROFILE USING MALDI-TOF MASS SPECTROMETRY DATA IN NON-SMALL CELL LUND CANCER PATIENTS

2.1. Introduction

In recent years, genomic and proteomic technologies have become a topic of central importance in biomedical studies. These technologies are often used to identify genes and proteins that may have a functional role for specific phenotypes. From this perspective, one might be interested to model the relationship between the genes or proteomic features and a clinical outcome. The use of mass spectrometry as a diagnostic tool and identification of proteomic biological markers has risen extensively and has demonstrated great promise in recent years. This has led to the discovery of a large number of proteins and protein profiles associated with various types of diseases (Stoeckli *et al.*, 2001; Petricoin, 2002; Adam *et al.*, 2002; Ndukum *et al.*, 2010; Aebersold *et al.*, 2003; Liotta *et al.*, 2003; Rai *et al.*, 2004).

In this chapter, our goal is to predict patient survival times from the proteomic features of the patient bodily fluids such as blood, plasma serum and so on using mass spectrometry. Mass spectrometry for the proteomic analysis consists of diverse platforms. Matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF) is one the basic forms of the technology. A typical data set contains hundreds of spectra; each spectrum contains tens of thousands of intensity measurements representing an unknown number of protein/peptide peaks which are the key features of interest. Including all the features out of this platform in the predictive model of survival time introduces noise and is expected to have poor predictive performance. Generally speaking, even after some basic preprocessing and denoising such as peak detection (Atlas *et al.*, 2009; Renard *et al.* 2008; Jeffries,2005; Wolski *et al.*, 2005), there are still hundreds or thousands of retained potentially important features which could be used for the predictive modeling. In such cases, over-fitting is a potential threat. Besides the high-dimensionality of the feature set, some of the features are often highly correlated. Thus, in order to predict patient survival using a predictive statistical model, one needs to consider careful dimension reduction and important feature selection on top of basic pre-processing of mass spectrometry data.

Generally speaking, the high dimensional property of "-omics" data imposes some of the common challenges in the analysis of such data. High dimensional setting means that the number of variables (p) is considerably larger that the number of observations (n). Dimensionality is typically even much larger in proteomic studies (say as compared with gene expression studies). As a result, most traditional multivariate statistical methods are not applicable in this case; it is challenging to develop reliable regression models that can correctly predict future phenotypic outcomes out of these proteomic features. Further complexities arise when the outcome of interest is the patient survival time, which is often not fully observed due to right censoring. A number of early attempts, mostly in the genomic data setting, used some ad hoc dimension reduction methods and used the reduced set of covariates (e.g., principal components, meta-genes etc.) in a Cox's proportional hazards regression model (Pawitan *et al.*, 2004; van Houwelingen *et al.*, 2006; Bovelstad *et al.*, 2007). More recently, penalized regression version of Cox's model have been attempted to deal with high dimensional data (Li *et al.*, 2003; Gui *et al.*, 2005). Cox's model with univariate shrinkages is another method in which the features are entered into the model based on the size of their Cox's score statistics (Tibshirani, 2009).

This method assumes that the features are independent which is clearly unrealistic in genetic and proteomic studies. In fact, the proportional hazards assumption of a Cox model (Cox, 1972) may be too simplistic for genomic and proteomic applications.

On the other hand, semi-parametric estimation in the accelerated failure time (AFT, hereafter) model with an unspecified error distribution is often regarded as a more flexible alternative to the Cox model in survival analysis. However, there are only a handful of publications on employment of the AFT model in high dimensional data setting mostly using the microarray platforms. The LASSO and the threshold-gradient-directed regularization along with AFT model are applied for estimation and variable selection (Huang *et al.*, 2006). Additionally, predicting survival time using AFT model along with PLS and LASSO is considered (Datta *et al.*, 2007). The elastic net approach for variable selection both under the Cox's proportional hazards model and under the AFT model is adopted (Engler *et al.*, 2009). Nevertheless, there is no such study using the AFT model that we are aware of in the context of proteomic data in which one is faced with even larger dimensionality of the original feature set.

In this chapter, we compare the performances of four relatively recent latent factor and/or regularized/penalized regression techniques to fit an AFT model based on high dimensional regressors and to predict the patient survival using high dimensional mass spectrometry data. In the next section, we provide brief descriptions of this regression techniques. These methods are applied to analyze survival times generated from simulated mass spectra, as well as, two real mass spectrometry data sets on non-small cell lung cancer patients. In survival studies, an added complication due to right censoring is almost always present. Right censoring takes place when there are still surviving patients at the end of the study period. This was indeed the case with one of our two real proteomic data sets. We handle right censoring through a residual based multiple imputation scheme.

2.2. Materials and Methods

The accelerated failure time (AFT) model is a flexible semi-parametric regression model that can be used to predict survival times of patients from patient level covariates (Kalbfleisch *et al.*, 1980). In general, AFT models are linear models for the logarithm or a known monotone transformation of the survival time (or an event time) T. Most commonly, an accelerated failure time (AFT) model specifies $\log T = X^T\beta + \epsilon$, where β is an unknown $p \times 1$ parameter of interest associated with the proteomic features X and ϵ is an unobservable independent random errors. The following latent factor and/or regularization techniques are used to fit the AFT model of $Y = \log T$ on the proteomic features X (intensity data corresponding to selected values) of patients.

2.2.1. PLS and SPLS

Introduced by Herman Wold (Wold, 1958), Partial Least Squares (PLS) regression is a predictive modeling technique that is applicable in high dimensional setting in several areas of scientific research. When the number of covariates is large compared to sample size and/or exhibit high collinearity, the standard multiple linear regression fit by ordinary least squares is inapplicable or inappropriate. PLS is particularly useful in such cases. PLS attempts to extract the latent factors that account for most of the variation in the response while avoiding over-fitting. Unlike principle component regression, both the response and the covariates are used to construct the latent components. There are several variants of the algorithm for obtaining the PLS estimators. Boulesteix & Strimmer (2007) reviewed both the theory underlying PLS as well as a host of bioinformatics applications of PLS. They provide a systematic comparison of the PLS approaches currently employed.

Sparse partial least squares (SPLS) is a relatively recent technique that combines the latent factor approach with regularization (Chun *et al.*, 2010). This technique produces sparse linear combinations of the original predictors and achieves both dimension reduction and variable selection simultaneously. Like PLS, the number of SPLS latent components is limited by the number of observations but the actual number of variables that make up the latent components can exceed the sample size n. SPLS method deals with variable selection problem by incorporating the PLS technique into the LARS (least angle regression) algorithm imposing the L_1 penalty controlled by a tuning parameter (Efron *et al.*, 2004).

2.2.2. LASSO and Elastic Net

Regularized/penalized methods are another group of techniques that has gained great popularity in recent years. Penalized estimation methods shrink the estimates of the regression coefficients towards zero relative to the maximum likelihood estimates. The purpose of this shrinkage is to prevent over-fitting arising due to either collinearity of the covariates or high-dimensionality. Least absolute shrinkage and selection operator (LASSO) is a penalized least squares method imposing an L_1-penalty on the regression coefficients (Tibshirani, 1996). The LASSO does both continuous shrinkage and automatic variable selection simultaneously and it has sparse representation. In the usual regression set-up, the LASSO minimizes the residual sum of squares subject to the sum of the absolute values of the coefficients being less than a constant (L_1 regularization). Equivalently, the LASSO estimator of the regression coefficients β is the value that minimizes $\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$, where λ is a penalty parameter. As mentioned before, the least angle regression (LARS) procedure is another variable selection method in the linear regression setting (Efron *et al.*, 2004*a*). The LARS selects predictors based on the correlation between the predictor and the current residuals. In this algorithm, one starts with the trivial model with all coefficients set at zero. Variables are added one at a time that are most correlated with the residuals at the previous step. The number of variables to be included in the model corresponds to a selection of the tuning parameters.

LARS can be modified to provide solutions for LASSO making it computationally efficient (Efron *et al.*, 2004*b*).

The elastic net approach is a newer regularization and variable selection method that combines L_1 and L_2 penalties (Zou *et al.*, 2005). The estimator minimizes $\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda_2\sum_{j=1}^{p}|\beta_j|^2 + \lambda_1\sum_{j=1}^{p}|\beta_j|$, where λ_1 and λ_2 are two penalty parameters. This procedure tends to produce a result with fewer regression coefficients set to zero than with a pure L_1 regularization, and more shrinkage of the other coefficients. In this method, strongly correlated predictors tend to be in or out of the model together. The elastic net is particularly useful when the number of predictors is much larger than the number of observations. A rescaling of the fitted coefficient is also done to reduce "double shrinkage". A version of the LARS algorithm, called LARS-EN, is used to fit the elastic net regression where the λ_1 parameter is controlled by the number of LARS steps.

2.2.3. Treatment of Right Censored Observations

Let $\{(T_i, C_i, X_i), i = 1, ..., n\}$ be the survival times, theand the p-dimensional covariate vectors (intensity values of selected m/z channels) of n patients in the data set. We assume that the distribution of T given X follows an AFT model. Due to right censoring, the observed data consists of $\{T_i^c, \delta_i, X_i), i = 1, ..., n\}$, where $T_i^c = min(T_i, C_i)$ is the right censored survival times and $\delta_i = I(T_i \leq C_i)$ are the failure indicators. We propose to impute unobserved T_i (that are censored) from an appropriate conditional distribution to be estimated from the observed data. Our proposal has its root in Poor Man's Data Augmentation Algorithm (PMDA) (Wei *et al.*, 1991). Our multiple imputation algorithm is an iterative algorithm that is described below. In this algorithm, L and m are a user defined integer parameters where L denotes the number of iterations and m denotes the number of data sets to be imputed in order to reduce the variability of the final answer. The steps involved to carry out the procedure are presented below.

Step 1: (a) Set $j = 0$ and fit the AFT model to the uncensored observations only using a regression technique of choice (i.e., one of the four methods described before).

(b) Suppose the current estimate is $\widehat{\beta}^{(j)}$, for $j \geq 0$.

Step 2: Calculate the Kaplan-Meier estimate $\widehat{S_\epsilon}^{(j)}$ of the marginal distribution of the model error ϵ using the usual product limit formula from the residual vectors and the failure indicators $\{e_i, \delta_i\}$, where $e_i = \log(T_i^c) - X_i^T\widehat{\beta}^{(j)}$, $1 \leq j \leq n$.

Step 3: Generate m new data sets such that in each data set $1 \leq k \leq m$, an observed failure in the original data remains intact but each censored observation i in the original data is imputed (in the log scale) by adding the estimated regression function $X_i^T\widehat{\beta}^{(j)}$ to an imputed model residual $\epsilon_{ik}^{(j)}$ generated from the estimated distribution $\widehat{S}_\epsilon / \widehat{S}_\epsilon(e_i)$ calculated in Step 2.

Step 4: Fit the model on each new data set using a method of choice (e.g., each of the four regression methods described before) set to find the estimated regression parameter vectors $\widehat{\beta}_{(k)}^{(j)}$

Step 5: Increase j by 1 and set $\widehat{\beta}^{(j+1)} = \frac{1}{m}\sum_{k=1}^{m}\widehat{\beta}_{(k)}^{(j)}$; go to Step 1(b) and repeat these steps L times.

We have used $m = 10$ and $L = 5$ in our calculation mostly to keep the computation time in check.

2.2.4 Construction of Survival Curves

Once a model is fit, we compute the survival function of the model error distributions $\widehat{S}_\epsilon$ using the Kaplan-Meier product limit formula from the residual vectors and the failure indicators $\{\mathrm{e}_i, \delta_i\}$, where $e_i = \log(T_i^c) - X_i^T\widehat{\beta}$, $1 \leq j \leq n$. Then the survival function of future patient with proteomic profile X_* is given by $\widehat{S}(t|X_*) = \widehat{S}_\epsilon(\log t - X_*^T\widehat{\beta})$, $t \geq 0$.

2.2.5. Simulated Data

Coombes $et\,al.$(2005) developed a tool capable of simulating realistic mass spectra. Morris $et\,al.$(2005) used the proposed tool to simulate hundreds of proteomics data sets. The data sets are available at: http://bioinformatics.mdanderson.org/Supplements/Datasets/Simulations/index. We take the first 50 datasets from this collection each containing 100 spectra. The lists of true peaks (features) are also available. The data is given in two columns, with the first column containing the mass and the second column containing the intensity. We simulate our spectra by random resampling from these dataset. More precisely, we sample one spectrum at random from each dataset resulting in 50 total sample spectra. A pre-processing step of peak detection and alignment is performed using R software pkDACLASS. The package can be downloaded from http://cran.r-project.org/web/packages/pkDACLASS/index.html. There were $p = 124$ aligned peaks and the corresponding intensity values are taken as the covariate vectors in the simulated AFT model. To simulate the survival times, we consider four different scenarios for the β coefficients. These are as follows: $(i)\,\beta_j = exp\{-j\}$ for $1 \leq j \leq 124$; $(ii)\,\beta_j = 1/j$ for $1 \leq j \leq 124$; (iii) for $1 \leq j \leq 10$, $\beta_j = j$ mod 5, if j mod $5 > 0$, otherwise $\beta_j = 5$ and for $11 \leq j \leq 124, \beta_j = 0$; (iv) $\beta_j = 1$ for $1 \leq j \leq 124$. Note that (i) and (ii) both represent situations when all the co-variables have positive but variable effects on survival; however, due to the exponential nature of the decaying coefficients, only the first few will have a real effect on survival in Scenario (i). Case (iv) denotes an extreme hypothetical scenario when all covariates have the same positive effect on survival. Presumably, (iii) denotes the most realistic scenario when the collection of covariates contains a large number of pure noise variables. In each case, the vector of coefficients is standardized for computational stability. A normal distribution is used for generating the additive errors. In other words, log normally distributed failure times are considered.

Next we added censoring to the simulated data. The censoring variable was taken to be log-normal. More precisely, $C = \exp\{N(c_0\sigma\sqrt{1+r}\ ,\sigma^2(1+r))\}$, where various values of c_0 are chosen to control the censoring rate and where $\sigma^2 = \beta^T\sum_X\beta$ is the variability in the regression model, where $\sum_X = n^{-1}\sum(X_i - \overline{X})(X_i - \overline{X})^T$. We simulated three censoring rates: 0% (no censoring), 10% (low censoring) and 60% (high censoring). A value of $r = 1$ is used throughout the simulation which denotes a noise to signal ratio.

For each design choice, the following measure is computed for checking the fit of the trained model: $MSEF = E\left[\frac{1}{n_R\sigma^2}\sum_{i=1}^{n}\widehat{\delta}_i(\overline{Y}_i - \log T_i^c)^2\right]$ where $n_R = \sum_{i=1}^{n}\delta_i$ is the number of uncensored observations in a sample. Next, for each training data set, a test data set $\mathrm{Y}_i^{new} = \log T_i^{new}$, $1 \leq i \leq n$, of the same size is generated using the same design parameters. An AFT model is fit to the training data using each of the four methods and the fitted model is used with the design matrix of the test data to get predicted values $\widehat{\mathrm{Y}}_i^{new} = X_i^T\widehat{\beta}$, for $1 \leq i \leq n$. The following measure is computed to determine the prediction accuracy: $MSEP = E\left[\frac{1}{n\sigma^2}\sum_{i=1}^{n}(\ \widehat{Y}_i^{new} - Y_i^{new})^2\right]$. Each of these measures is computed by averaging these quantities over 100 Monte-Carlo replicates.

2.2.6. Netherlands Non-small Cell Lung Cancer Data

We use the data set reported in Voortman *et al.* (2009) The MALDI-TOF-MS dataset of serum samples of 27 patients with advanced non-small cell lung cancer (NSCLC), treated with chemotherapy and Bortezomib were obtained. Serum spectra of these patients are available at three time points: pre-treatment (preTx), after two cycles of treatment (post-2) and at the end of treatment (EOT). For each patient, there is an associated progression-free survival (PFS) recorded in days. No censoring exists in this data. The range of observed survival time in this data set is 27 days to 601 days.

We take EOT samples along with PFS values as major information for further analysis. Two samples are excluded due to missing EOT serum spectrum. Each spectrum consisted features with mass range of 800-4000 Dalton (Da) with the corresponding intensities.

Raw mass spectra generally have systematic variations between spectra caused by sample degradation over time which is the case for our data. It is necessary to align spectra so that characteristic features occur at the same time in all spectra. For this purpose, we binned the m/z ratios to the nearest 0.05 Da, and averaged over the intensity values with the same m/z value.

Next, we follow the standardization and denoising algorithms proposed by Satten $et\,al.$(2004) with slight modification. In this method, each spectrum is standardized using information from that spectrum only. The spectra are centered using a local estimate of the median and standardized using a local estimate of the interquartile range. Let m_i denote the m/z ratios with the corresponding intensities x_i. The spectrum is standardized by replacing each intensity value with $x_i^* = \frac{x_i - Q_{0.5}(m_i)}{Q_{0.75}(m_i) - Q_{0.25}(m_i)}$, where $Q_\alpha(m_i)$ is a local estimate of the $\alpha - th$ quantile of spectral intensities at m/z ratios near m_i. The final step is to denoise spectra which are still a mixture of noise and signal. In fact, denoising ensures that the features used in analysis correspond to real peak. Following Satten $et\,al.$, first we need to estimate the noise scale for a given spectrum. The sample root mean square $\widehat{\sigma}$ of the negative values of $x^*(m)$ can serve as an estimate of the standard deviation of the noise distribution σ. The hard thresholding criterion to denoise the standardized spectrum is to replace a standardized intensity $x^*(m)$ by $x^*(m)I(x^*(m) \geq 6\widehat{\sigma})$, where I denotes an indicator function. The threshold of $6\widehat{\sigma}$ as a cutoff to eliminate normally distributed noise is a conservative choice.

All 25 spectra are standardized and denoised as described above. In order to retain the true signals from the denoised spectra to build a predictive model, three different approaches are employed. First, the intersection rule, in which only the features with nonzero standardized intensities in all spectra are maintained in the denoised samples. This gives total number of 995 m/z values, denoted $X(1)$ in the rest of the paper. In the second approach, the features with more than five zero values are removed from all spectra. A total of 4652 features $X(2)$ are resulted using this method. Applying the union rule, features with at least one nonzero value in all 25 spectra are retained as signal in the sample spectra; we obtain a feature set $X(3)$ of 8184 m/z ratios to be used in a predictive model building.

In our analysis, we use each of the resulting feature sets $X(1), X(2)$ and X(3) in an AFT model to determine the relationship between progression free survival time (in days) and proteomic features for the 25 cancer samples. As mentioned before, four methods of modeling fitting PLS, SPLS, LASSO and elastic net are implemented with each feature set.

Estimation of model fit, prediction error and selection of tuning parameters are carried out by leave-one-out cross validation on the data. We compared the performance of these methods by computing their estimated mean-squared error of prediction ($EMSEP$) which is minimized with respect to the selected values of the tuning (operational) parameters in a regression method. Unlike the simulated data, the EMSEP here is computed by leave-one out cross validation, $EMSEP = n^{-1}\sum_{i=1}^{n}(\widehat{Y}_{-i} - Y_i)^2$, where $\widehat{Y}_{-i}$ is calculated by first fitting the model on the sample values other than the ith sample unit and predicting the ith value using the fitted model with the covariate X_i. We have investigated the performance of $EMSEP$ and its minimizer (wrt the tuning parameter such as the number of PLS terms or the number of LASSO steps) in the simulation settings of the earlier subsection. We found that the median of the data based minimizers remain close to the true minimizers of the corresponding $MSEP$; also the values of the mean squared prediction error at the median is close to the overall minimum. These are reported in Table 2.1.

SPLS regression has two key tuning parameters: the thresholding parameter (λ) and number of hidden components (K). Following the guidelines given in Chun $et\,al.$, cross validation is computed over the grid of $K =$ 1, 2,..,20 and $\lambda =$ 0.1, 0.2,..., 0.9. There are two tuning parameters in the elastic net as well. These are the penalty terms λ_1 and λ_2. We selected a grid of values for λ_2=0.01, 0.1,1 10 and 100. For each λ_2, the entire solution path is produced and optimum number of steps is chosen (which is equivalent to choosing). LASSO is a special case of the elastic net with λ_2=0.

2.2.7. Milan Non-small Cell Lung Cancer Data

This NSCLC data set was collected in Milan, Italy and was originally analyzed by Taguchi $et\,al.$(2007). There were three training cohorts with NSCLC who were treated systemically with efitinib and from whom sera had been collected before treatment.

We only considered the first training cohort collected from Scientific Institute Hospital San Raffaele, Milan, Italy ($n =$ 70). Mass spectra for the training samples were generated independently from both Vanderbilt University (VU) and University of Colorado at Denver and Health Sciences Center (UCDHSC). The Mass spectra obtained from VU are considered for our analysis. One patient without clinical outcome of survival is removed from the analysis. The range of observed survival time in this data set is 28 days to 1169 days.

We processed the baseline-adjusted spectra using standardization and denoising algorithm as described previously. Here we choose a less conservative threshold $4\widehat{\sigma}$ to eliminate normally distributed noise since this choice maintained a reasonable number of features. Similar to the earlier analysis, three different approaches are employed: the intersection rule, the features with at least five non-zero values are retained and union rule to produce the feature sets after standardization and denoising, denoted $Z(1)$, $Z(2)$ and $Z(3)$, with sizes 104, 3051 and 4816, respectively.

This data set has 12 (17.3%) right censored observation. We apply the multiple imputation algorithm described before to handle censoring. The definitions of the estimated mean squared errors for fit and prediction are adjusted to reflect the fact that only uncensored data are used for the comparison with the fitted values. To this end, we use $EMSEF = n_R^{-1}\sum_{i=1}^{n}\delta_i(\widehat{Y}_i - Y_i\ )^2$ with $n_R = \sum_{i=1}^{n}\delta_i$ (recall that δ_i are the true failure indicators) and $EMSEP = n^{-1}\sum_{i=1}^{n}\delta_i(\widehat{Y}_{i,-} - Y_i\ \ )^2/\widehat{S}^c(T_i^c -\ )$, where S^c denotes the survival function of the censoring random variable C. It can be estimated, under the assumption that C is independent of $(T, X_1, ..., X_p)$, by the Kaplan-Meier estimator of the survival function with the roles of δ and $1-\delta$ switched.

2.3. Results

2.3.1. Simulated Data

The four methods fit the uncensored part of the data very well and the MSEF decreases with increasing number of components or steps in the four methods. We include the values of Scenario (ii) in Table 2.2; the conclusions were similar in all cases.

An important aspect of the performances of the four methods is reflected by the mean squared error of prediction, MSEP. MSEP is plotted as a function of either the number of hidden components or the number of LARS steps with different level of censoring; for clarity of presentation, we fix the other tunning parameter for methods involving two tunning parameters (only two values are selected for each such plot to avoid over-crowding). In each case we also plot the horizontal line $y = (1 + r)$ for benchmarking, where $1 + r$ is the (constant) theoretical value of MSEP had we used no covariates. Recall that $r = 1$ was used in all cases. Thus a value of MSEP below 2 will indicate some predictive power of the model of survival on proteomic data.

Figures 2.1 to 2.4 display the results for Scenario (i) to Scenario (iv). The performance of PLS has been marginal in Scenarios (i), (ii) and (iv) compared to the no covariate model, especially, with larger number of components. However, in the fourth scenario PLS clearly outperforms the no covariate model. This scenario corresponds to the situation when all covariates are contributing to the regression function. On the other hand, the MSEP of LASSO, elastic net and SPLS appear to be smaller than $(1 + r)$ in all uncensored cases. In Scenario (iii), where only a handful of covariates contribute to the model, elastic net performs better than other methods. In Scenario (i) and (ii), LASSO and elastic net performance are quite similar. The elastic net results appear to be more stable than SPLS. However, SPLS performance is better than PLS in nearly all scenarios.

In all cases, increasing the level of censoring decreases the performance of the methods (i.e., increased MSEP) which is to be expected. In some cases the performance of some methods is marginal compared to the model with no covariate with lower level of censoring. The clear examples of such cases are elastic net Scenarios (i), (ii) and (iii), LASSO Scenario (i) and (iv), SPLS Scenario (i), (ii) and (iv). The performance of elastic net in Scenario (iv) with higher level censoring is still better than the model with no covariate.

In Scenario (ii), SPLS with small number of hidden components performs well even in the censored case. In this scenario, almost all the features contribute to the model but contributions are not equal. Overall, correct choice of the proper tuning parameters seems to be important for decent predictive ability of the model.

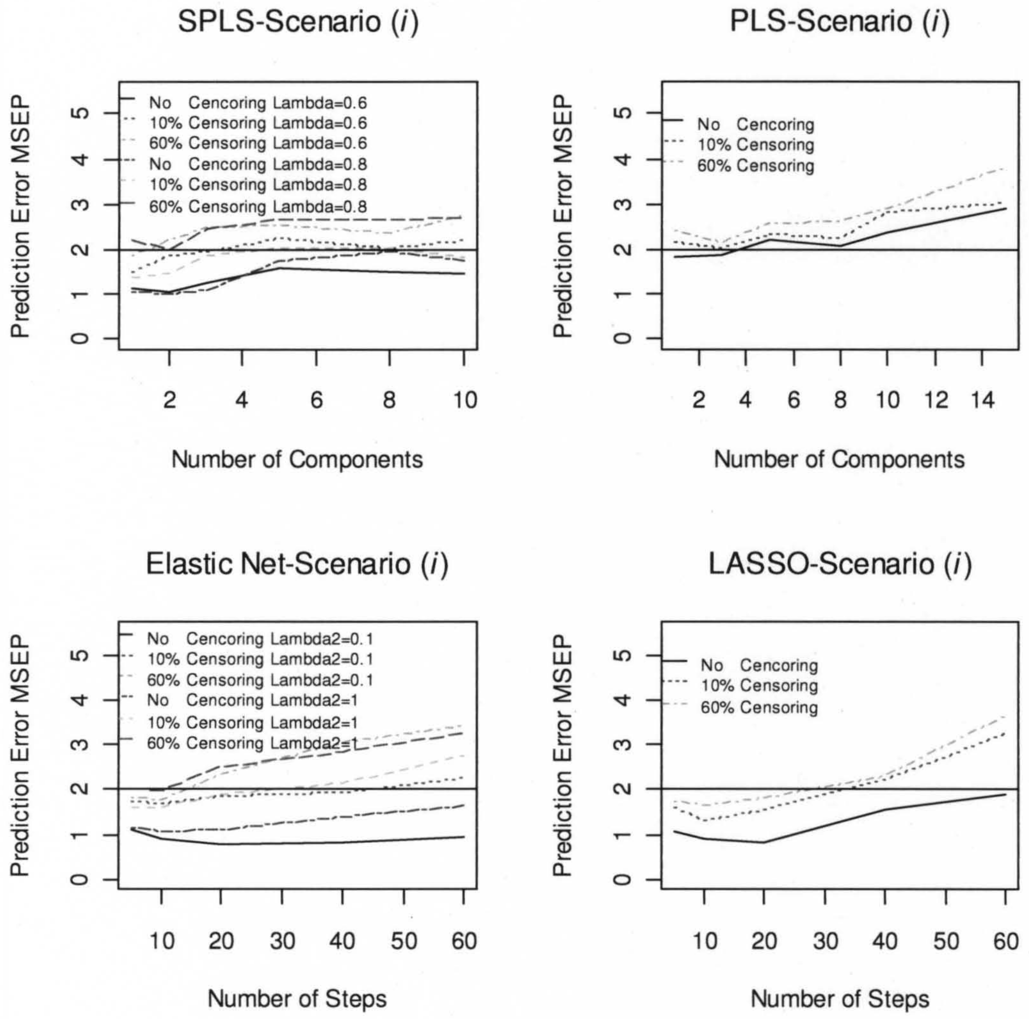


Figure 2.1. Mean squared error of prediction (MSEP) in a simulated model with $p = 124$ features where the regression coefficients are given by $\beta_j = \exp\{-j\}$ for $1 \leq j \leq 124$. The sample size was $n = 50$. The horizontal line indicates the prediction error of a model that does not use any proteomic features as covariates.

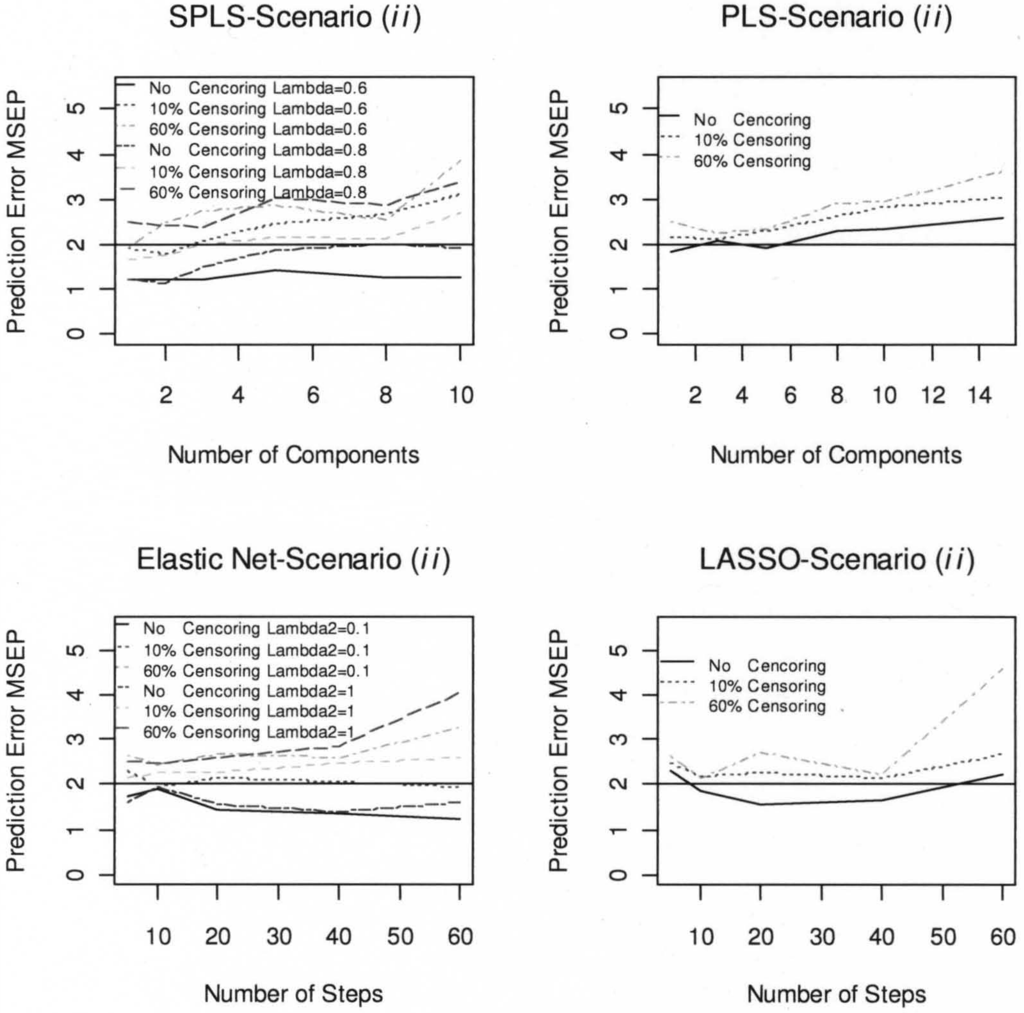


Figure 2.2. Mean squared error of prediction (MSEP) in a simulated model with $p = 124$ features where the regression coefficients are given by $\beta_j = 1/\mathrm{j}\,\mathrm{for}$ $1 \leq j \leq 124$. The sample size was $n = 50$. The horizontal line indicates the prediction error of a model that does not use any proteomic features as covariates

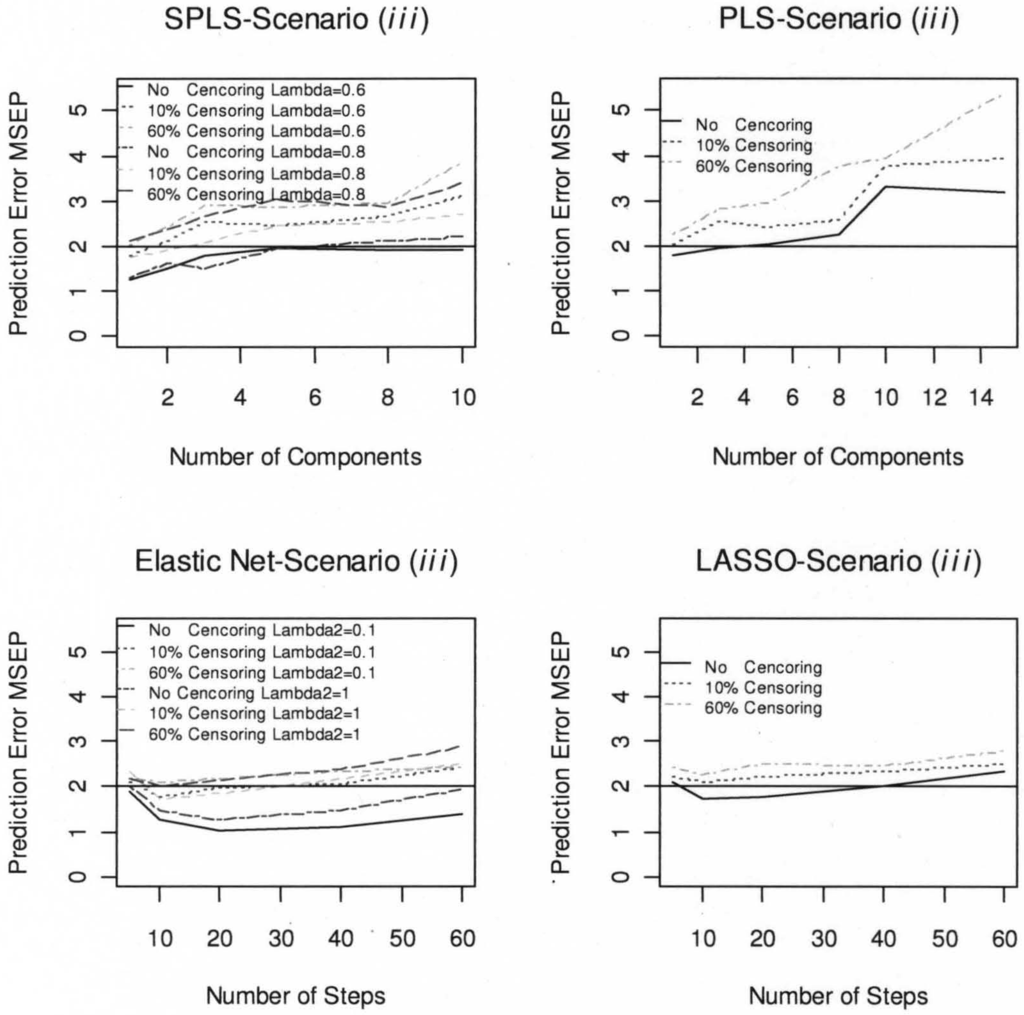


Figure 2.3. Mean squared error of prediction (MSEP) in a simulated model with $p = 124$ features where the regression coefficients are given by $\beta_j =$ j mod 5, if j mod $5 > 0$, $= 5$, otherwise, when $1 \leq j \leq 10$ and $\beta_j = 0$ for $11 \leq j \leq 124$. The sample size was $n = 50$. The horizontal line indicates the prediction error of a model that does not use any proteomic features as covariates.

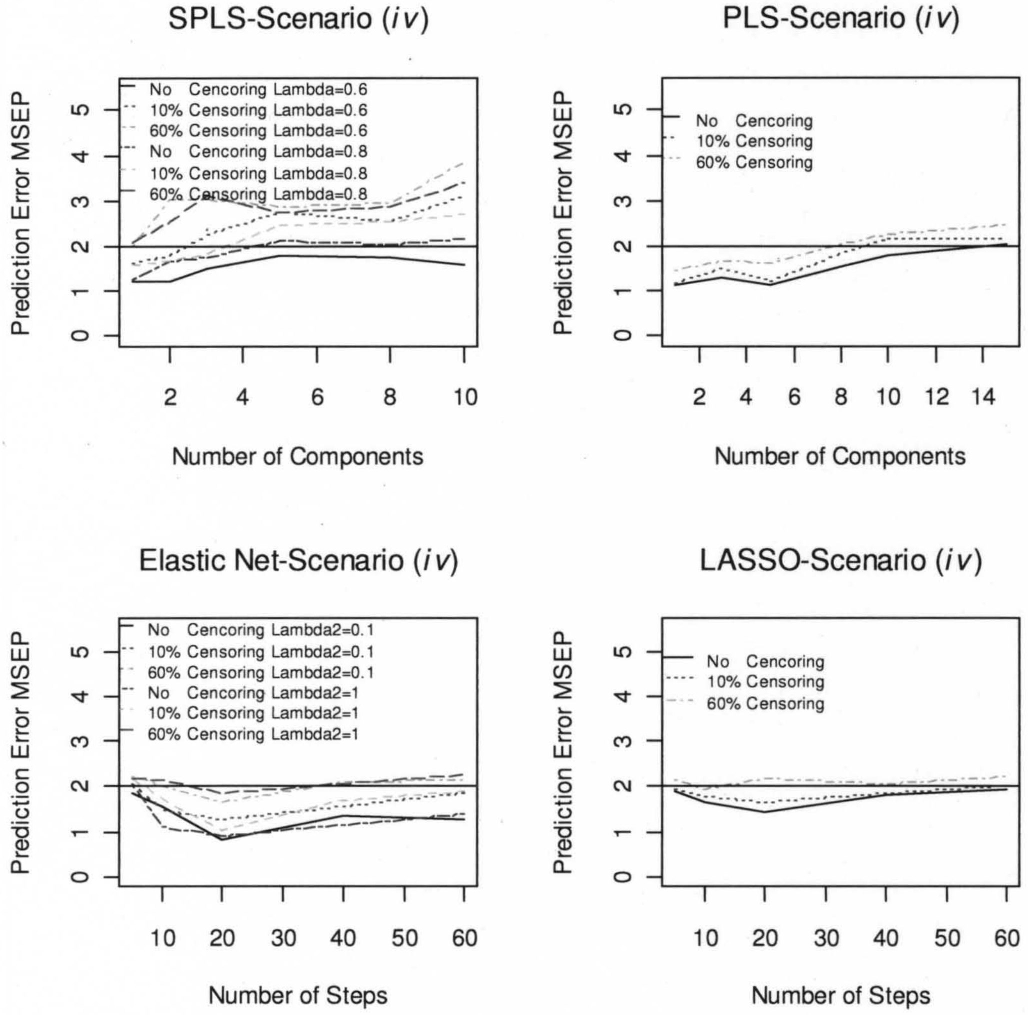


Figure 2.4. Mean squared error of prediction (MSEP) in a simulated model with $p = 124$ features where the regression coefficients are given by $\beta_j = 1$ for $1 \leq j \leq 124$. The sample size was $n = 50$. The horizontal line indicates the prediction error of a model that does not use any proteomic features as covariates.

Table 2.1. Median values of the optimum number of steps or number of components based on minimization of $EMSEP$ for 100 simulated samples and the corresponding true optimum value minimizing $MSEP$. The same four scenarios as in the Section 2.2 were considered without any censoring.

	SPLS(lambda=0.6)		PLS		Elastic Net		LASSO	
	# of Components				# of Steps			
	Median of Estimates	True optimum	Median of Estimates	True optimum	Median of Estimates	True optimum	Median of Estimates	True optimum
Scenario (i)	2	2	3	3	22	20	24	20
Scenario (ii)	2	2	2	1	76	60	26	20
Scenario (iii)	2	1	2.5	1	41	40	13.5	10
Scenario (iv)	2	1	4.5	3	40	40	22	20

Table 2.2. Mean squared error of fit (MSEF) in a simulated model with p=124 features where the regression coefficients are given by $\beta_j = 1/j$ for $1 \leq j \leq 124$. The sample size was n=50

Censoring rate	LASSO #of steps	MSEF	Elastic Net # of steps	MSEF	SPLS # of components	MSEF	PLS # of components	MSEF
0%	5	0.73	5	0.77	1	0.57	1	1.33
	10	0.48	10	0.51	2	0.19	3	1.13
	20	0.20	20	0.24	3	0.05	5	0.81
	40	0.02	40	0.05	5	0.004	10	0.10
	60	$6e^{-4}$	60	0.01	10	$2e^{-6}$	15	0.04
10%	5	1.09	5	0.96	1	0.67	1	1.57
	10	0.73	10	0.62	2	0.58	3	1.31
	20	0.44	20	0.39	3	0.46	5	1.02
	40	0.10	40	0.11	5	0.25	10	0.24
	60	0.01	60	0.08	10	0.06	15	0.09
60%	5	1.22	5	1.09	1	0.89	1	1.72
	10	0.86	10	0.83	2	0.72	3	1.41
	20	0.52	20	0.44	3	0.65	5	1.13
	40	0.14	40	0.14	5	0.42	10	0.31
	60	0.04	60	0.10	10	0.10	15	0.11

2.3.2. Netherlands NSCLC Data

Table 2.3 shows the measure of prediction for Netherlands NSCLC data. As mentioned earlier, $X(1)$, $X(2)$ and $X(3)$ each has 995, 4652 and 8184 features respectively. As the number of features increase, LASSO and elastic net performs better in terms of prediction error. However, this is not the case for PLS and SPLS. Comparing LASSO and elastic net, the elastic net has smaller prediction error in all three cases. Moreover, SPLS performs better than the PLS method. Overall, elastic net and SPLS outperform PLS and LASSO. The smallest prediction error corresponds to using elastic net. SPLS performs better than elastic net with higher number of features in $X(2)$ and $X(3)$.

Table 2.3. Estimated mean squared error of prediction $(EMSEP)$ for the Netherlands, NSCLC data. Three feature selection methods are tested; $X(1)$ has 995 features, $X(2)$ has 4652 features and $X(3)$ has 8148 features. In each case, the minimum $EMSEP$ value over the operational parameters is reported for each regression method (predictive model).

Method/ Feature selection		$EMSEP$
LASSO	$X(1)$	0.43
	$X(2)$	0.65
	$X(3)$	0.91
Elastic Net	$X(1)$	0.06
	$X(2)$	0.41
	$X(3)$	0.68
PLS	$X(1)$	0.68
	$X(2)$	0.68
	$X(3)$	0.54
SPLS	$X(1)$	0.56
	$X(2)$	0.31
	$X(3)$	0.52

2.3.3. Milan NSCLC Data

Table 2.4 displays the measure of prediction for MilanNSCLC data. Here the number of features in corresponding to the three sets of covariates $Z(1), Z(2)$ and $Z(3)$ are 104, 3051 and 4816 respectively. As can be seen from the table, the prediction error $Z(1)$ for has the smallest value using elastic net and the largest using PLS. This is true for $Z(2)$ as well. However, for $Z(3)$ SPLS has the smallest value. Overall, elastic net outperforms LASSO and SPLS outperforms PLS. Except for $Z(3)$ that has larger number of covariates, elastic net results are better than SPLS.

We plot the predicted survival curves of a number of hypothetical patients with similar proteomic profiles $Z(1)$ as in the Milan NSCLC Data using the two best methods (namely SPLS and elastic net). The result is displayed in Figure 2.5. The elastic net is producing slightly tighter results although SPLS exhibits greater consistency with the range of observed survival times in the data set. Table 2.5 includes results on the measures of fit, $EMSEF$ for the Milan NSCLC data over selected choices of "number of components" in PLS and SPLS with fixed $\lambda = 0.3$ and "number of steps" in LASSO and elastic net with fixed $\lambda_2 = 0.3$. Both PLS and SPLS fit the uncensored part of the data quite well and the $EMSEF$ decreases rapidly with increasing number of PLS/SPLS terms. The same holds for LASSO and elastic net as well although, generally speaking, the $EMSEF$ decreases relatively slowly with increasing number of steps in elastic net and LASSO. Larger feature sets improve the fit in PLS and SPLS. However, this is generally not the case for LASSO and elastic net. $Z(2)$ has smaller values of measure of fit using LASSO and elastic net compared to $Z(1)$ and $Z(3)$.

Table 2.4. Estimated mean squared error of prediction ($EMSEP$) for the Milan, NSCLC data. Three feature selection methods are tested; Z(1) has 104 features, Z(2) has 3051 features and $X(3)$ has 4816 features. In each case, the minimum $EMSEP$ value over the operational parameters is reported for each regression method (predictive model).

Method/ Feature selection		$EMSEP$
LASSO	$Z(1)$	0.14
	$Z(2)$	0.58
	$Z(3)$	1.14
Elastic Net	$Z(1)$	0.09
	$Z(2)$	0.19
	$Z(3)$	0.57
PLS	$Z(1)$	1.42
	$Z(2)$	1.80
	$Z(3)$	1.79
SPLS	$Z(1)$	0.32
	$Z(2)$	0.47
	$Z(3)$	0.34

Table 2.5. Estimated mean squared error of fit $(EMSEF)$ for the Milan, NSCLC Data. Three feature selection methods are tested; $Z(1)$ has 104 features, Z(2) has 3051 features and $Z(3)$ has 4818 features.

Feature set	LASSO # of Steps	EMSEF	Elast. Net # of Steps	EMSEF	SPLS # of Comp.	EMSEF	PLS # of Comp.	EMSEF
$Z(1)$	1	2.4	1	1.02			1	1.34
	2	2.1	2	0.01	1	0.89	2	1.27
	3	1.91	3	$8e^{-3}$	2	0.42	3	0.69
	4	1.05	4	$5e^{-3}$	3	0.07	4	0.12
	5	0.65	5	$4e^{-3}$	4	$3e^{-3}$	5	0.08
	10	0.42	10	$3e^{-4}$	5	$1e^{-3}$	6	0.01
	15	0.12	15	$2e^{-4}$	6	$5e^{-4}$	7	$4e^{-3}$
	20	0.07	20	$1e^{-4}$	7	$6e^{-5}$	8	$2e^{-3}$
	25	0.05	25	$4e^{-5}$	8	$2e^{-7}$	9	$3e^{-6}$
	35	$7e^{-3}$	35	$3e^{-5}$	10	$1e^{-8}$	10	$1e^{-16}$
	50	$3e^{-3}$	50	$2e^{\ 5}$	12	$3e^{-10}$	25	$2e^{-19}$
	75	$5e^{-4}$	75	$1e^{-6}$	15	$4e^{-12}$	20	$4e^{-21}$
	100	$2e^{-4}$	100	$7e^{-6}$			30	$2e^{-23}$
$Z(2)$	1	0.91	1	0.97			1	1.02
	2	0.70	2	0.93	1	0.77	2	0.87
	3	0.67	3	0.72	2	0.28	3	0.79
	4	0.41	4	0.71	3	0.01	4	0.12
	5	0.23	5	0.62	4	$1e^{-3}$	5	$7e^{-3}$
	10	0.08	10	0.36	5	$4e^{-4}$	6	$8e^{-4}$
	15	0.02	15	0.21	6	$2e^{-5}$	7	$5e^{-5}$
	20	0.01	20	0.12	7	$3e^{-6}$	8	$1e^{-5}$
	25	$9e^{-3}$	25	0.07	8	$5e^{-8}$	9	$4e^{-7}$
	35	$6e^{-3}$	35	0.01	10	$3e^{-9}$	10	$3e^{-18}$
	50	$2e^{-3}$	50	$8e^{-3}$	12	$2e^{-12}$	15	$3e^{-23}$
	75	$1e^{-3}$	75	$1e^{-3}$	15	$7e^{-14}$	20	$2e^{-25}$
	100	$28e^{-4}$	100	$3e^{-4}$			30	$3e^{-26}$
$Z(3)$	1	2.02	1	1.48			1	0.20
	2	1.72	2	1.30	1	0.32	2	0.19
	3	1.32	3	1.04	2	0.17	3	0.14
	4	1.08	4	0.88	3	0.05	4	$3e^{-3}$
	5	0.98	5	0.76	4	$1e^{-4}$	5	$2e^{-3}$
	10	0.81	10	0.57	5	$2e^{-5}$	6	$2e^{-3}$
	15	0.70	15	0.31	6	$3e^{-6}$	7	$6e^{-4}$
	20	0.36	20	0.12	7	$1e^{-7}$	8	$3e^{-5}$
	25	0.29	25	0.01	8	$2e^{-9}$	9	$1e^{-6}$
	35	0.18	35	$4e^{-3}$	10	$5e^{-10}$	10	$7e^{-8}$
	50	0.05	50	$1e^{-3}$	12	$4e^{-13}$	15	$9e^{-28}$
	75	0.01	75	$3e^{-4}$	15	$3e^{-15}$	20	$4e^{-28}$
	100	$3e^{-3}$	100	$3e^{-5}$			30	$2e^{-28}$

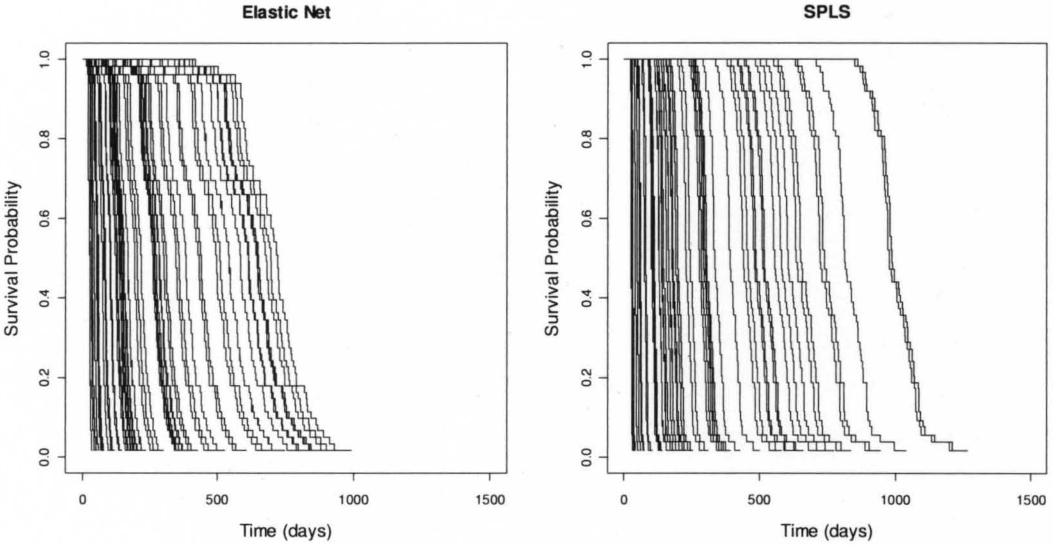


Figure 2.5. Estimated survival curves of patients with the same proteomic profiles as in Milan NSCLC Data.

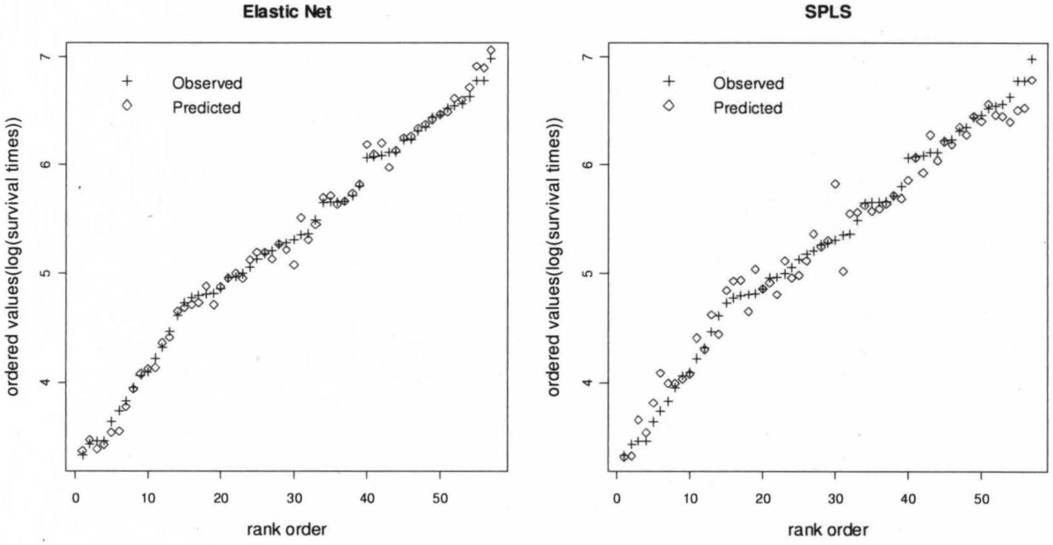


Figure 2.6. Observed versus fitted values in Milan NSCLC Data using the two best methods and feature set $Z(1)$. The optimal values of the tuning parameters as in Table 2.4 are used.

2.4. Discussion

For proper choices of the operational parameters, all four methods PLS, SPLS, LASSO and elastic net showed promise in predicting survival with large number of features versus limited sample size. It is likely that majority of features are not related to survival as only small number of features contributed the most to the regression models fitted. Based on our simulation study, elastic net outperforms LASSO and SPLS seems to be more effective than PLS when indeed there are large numbers of extraneous covariates. The proposed multiple imputation algorithm seems to be an appropriate way to handle censored data.

The performance of the four aforementioned regression methods are also compared on two real data examples; Netherlands NSCLC data and Milan NSCLC data. The prediction performance of all methods had similar trend in two data analyses and confirms results obtained in simulation. From Tables 2.3 and 2.4 for the analysis of these two datasets, we can conclude that SPLS is the best method when the number of covariates (features) is large since it outperforms LASSO and elastic net in terms of prediction error in these cases. It also outperforms PLS as not all the covariates contribute to the prediction in a significant manner and some sparsity is useful. However, for future survival prediction, the elastic net method with the filtered set of features that are present in all pre-processed training samples appears to be the best overall. The minimum prediction errors reported in Table 2.3 and 2.4 are somewhat optimistic since the optimal tuning parameters are also determined from the minimum prediction errors estimated from the training data.

One issue in using predictive models based on mass spectrometry profiles is that multiple runs of the mass spectrometer may produce spectra with slightly different set of m/z values. We have applied a simple strategy based of binning to deal with this after preprocessing of the raw spectra (which involve baseline correction, monoisotopic peak detection or statistical peak detection via normalization and denoising). Of course, similar pre-processing and binning steps need to be implemented to a future sample to extract the comparable set of features which are then fed into the predictive model. Although it is not strictly necessary to interpret these features for model building, such an identification step may lead to greater confidence to the quality of the features used for survival prediction. We have attempted some preliminary investigations for the Milan NSCLC data using use peptide mass fingerprinting and identified three proteins that may play some role in cancer as well as the survival process.

2.4.1. Feature Identification

Once an AFT model is fit to the test data using any of these methods, estimates of survival probabilities for a future patient with a given proteomic profile can be obtained. Although it was not the primary purpose of this chapter to identify the features and the corresponding proteins used for survival prediction, we performed a preliminary investigation in this direction. We illustrate this with the Milan NSCLC Data using the smallest collection of 104 features represented as Z(1) in Table 2.4. We use peptide mass fingerprinting for MALDI data to identify the proteins. We used Aldnte (http://www.expasy.org/ tools/aldente/) as the search engine with the following search parameters: (a) molecular mass range taken to be 6 - 30 kDa; (b) fixed modification of cystine residues by carboxy-amidomethylation; (c) variable oxidation modification of methionine (d) no restriction was placed on isoelectric point; and (e) species selected were Homo Sepians. Some top-scoring identified protein candidates from many spectra included the proteins O75157 (T22D2_HUMAN (V_2) Isoform 2 of TSC22 domain family protein, 76 kDa), O75157 (T22D2_HUMAN (C_1) TSC22 domain family protein, 79 kDa) and Q14671 (PUM1_HUMAN (C_1) Pumilio homolog 1, 126 kDa).

Two of the three top scoring proteins belong to TSC22 anti-apoptotic family of proteins. TSC22 is assumed to act as a negative growth regulator and tumor suppressor

(Gluderer *et al.*, 2004). TSC- 22 protein prevents yeast cell death due to a variety of apoptotic stimuli and also promotes cell survival in yeast (Khouri *et al.*, 2008). Moreover, TSC-22 belongs to a family of putative transcription factors encoded by four distinct loci in mammals. Q14671 is a PUMILIO-1 family of proteins. This family of proteins has been implicated in skin and epithelial cancer (Dazard *et al.*, 2003). So in a nutshell, these three proteins seem to play some role in cancer as well as the survival process.

CHAPTER III

NON-PARAMTERIC REGRESSION OF STATE OCCUPATION, ENTRY, EXIT AND WAITING TIMES WITH MULTISTATE RIGHT CENSORED DATA

3.1. Introduction

Multistate models are natural extensions of simple survival models that can be used to describe various types of event times. These models allow subjects to move through a succession of states and they are particularly useful for describing the complexities of disease processes in which each state corresponding to a certain health condition (e.g. alive and disease free, alive with recurrence and dead). The resulting data contains information about the transition times and the states occupied. Transitions between states can be reversible or irreversible while states can be either absorbing or transient. An absorbing state is a state from which further transitions cannot occur while a transient state is a state that is not absorbing. Multistate models may consist of various levels of complexities where individuals can pass through multiple transient states before entering a number of possible absorbing states. Graphically, multistate models may be illustrated using diagrams with boxes representing the states and with arrows between the states representing the possible transitions.

State occupation probability, which is the probability that a subject be in a specific state at certain time point, is an important quantity in study of multistate models. Another important quantity is the state transition intensity (or transitional hazard) which is the hazard (rate) that an individual moves from one state to another. They are functions of time, similar to the survival function in survival analysis. Distribution functions for the state entry and exit times are also of interest and so are that of state waiting times (sometimes referred to as the sojourn times). Estimators of these quantities have been proposed in the recent past under a variety of parametric and non-parametric assumptions, as well as, structural assumptions on the system such as progressive, Markov, semi-Markov and so on (Aalen, 1976; 1978; Aalen & Johansen, 1978; Datta & Satten, 2000; 2001; 2002; Aalen *et al.*, 2001; Satten & Datta 2002 etc.). Aalen and Johansen (1978) gave a method for calculating transition hazard and state occupation probabilities for Markov models starting from the Nelson–Aalen estimators of integrated transition hazards when data are subject to independent censoring. Datta and Satten (2001) showed that these estimators remain consistent even when the underlying model is non-Markovian; the same is not true, however, for the corresponding estimated bivarite (in time) transition probability matrix obtained by product integration (Meira-Machado *et al.*, 2006). They also extended their work and presented estimators using data that are subject to dependent right censoring (Datta & Satten, 2002). Their estimator of the integrated transition hazard matrix has a Nelson-Aalen form, where each of the counting processes, counting the number of transitions between states and the risk sets of leaving the states has an IPCW (Inverse Probability of Censoring Weighted) form. Non-parametric estimation of waiting times was undertaken in Lin *et al.* (1999) for the very special case of a progressive model with no branching. Satten and Datta (2002) also provided non-parametric estimates of waiting time distributions under dependent censoring using additive hazard model. Beside estimation in a marginal model, there is considerable appeal in developing methods for conditional distribution (e.g., regression) of these quantities relating them to number of covariates.

Majority of studies in the context of non-parametric regression of time to event data deal with the survival setup. One of the earliest theoretical studies can be traced back to Beran (1981) who studied a conditional Kaplan-Meier estimator (and also its quantile function) using local estimator obtained with regression weights using either a nearest neighborhood approach or a kernel approach. Theoretical properties of these estimators and their generalizations have been further studied by Dabrowska (1987; 1989; 1992 *b*), Li and Doss (1995), McKeague and Utikal (1990), Li and Datta (2001) and others. Additive non-parametric regression models for the hazard rate function were considered by Aalen (1980; 1989). Recently, Andersen and Keiding (2002) and Andersen and Klein (2007) studied the effects of covariates in a multistate model using a hybrid approach of combining some non-parametric calculation followed by semi-parametric ones. This approach, however, may not produce regression function estimators of the marginal quantities under study; furthermore, the theoretical modeling framework necessary for the validity of this approach is not very clear. The smoothing techniques offer useful alternatives to the non-parametric likelihood based approaches since a full likelihood specification in a multistate model is often difficult (and sometimes impossible without additional structural assumptions). There are various approaches to produce non-parametric smoothed estimates. The simplest and flexible methods are the kernel-based procedures (Anderson and Keiding, 2002; Nadaraya, 1964).

In this chapter, we extend the Beran estimator and develop methods for non-parametric regressions of the state occupation probabilities, the entry/exit and waiting time distributions in a multistate model. Throughout the chapter, we assume the framework of right censored transition times; however no structural assumptions on the multistate system (e.g., a Markov or a semi-Markov model) are being made. The form of the censoring hazard is fairly general - in particular, it may be controlled by some additional covariates that are observable. Thus, the proposed treatment is more general than the usual notion of "independent censoring" in regression models for survival data. The purpose here is to study the regression functions based on one continuous covariate at a time.

The rest of the chapter is organized as follows. The next section of the chapter introduces our proposed non-parametric estimation of state occupation probabilities, integrated hazard, entry/exit, and waiting time distributions. Result of simulation studies for our estimators are given in Section 3.3. Section 3.4 contains an analysis of a bone marrow transplant data. Finally, Section 3.5 contains some concluding remarks.

3.2. The Non-parametric Regression Estimators

3.2.1. Notation and Convention

Consider a time continuous, multistate process $\boldsymbol{S} = \{S(t) : t \geq 0\}$, where t denotes time and $S(t)$ denotes the state occupied at time t. Let $S(t-) = lim_{s \to t-} S(s)$ be the state occupied just before time t. A finite state space $\mathcal{X} = \{0, 1, ..., M\}$ is assumed for the process. Under the marginal model, it is assumed that the multistate processes for n individuals $\boldsymbol{S}_i = \{S_i(t) : t \geq 0\}$, $1 \leq i \leq n$, are independent and identical (i.i.d., hereafter) realizations of $\boldsymbol{S}$. We assume that the model is progressive (e.g., acyclic) and thus a given state j is entered at most once by an individual. We can define the state entry and exit times by $U^j = \inf\{t : S(t) = j\}$ and $V^j = \sup\{t : t > U^j, S(t) = j\}$. We take by convention, $U^j = \infty$, if state j is never entered and $V^j = \infty$, if either state j is never entered or j is an absorbing state (in which case it is never left). The state waiting time can be defined as $W^j = V^j - U^j$, when $U^j < \infty$. The (marginal) state entry, exit and waiting time distributions will be denoted by $F^j(t) = Pr\{U^j \leq t | U^j < \infty\}$, $G^j(t) = Pr\{V^j \leq t | U^j < \infty\}$ and $H^j(t) = Pr\{W^j \leq t | U^j < \infty\}$, respectively. Also, let $p_j(t) = Pr\{S(t) = j\}$ be the probability that a typical individual will be at state j at time t. In this chapter, we study the effects of a fixed (i.e., baseline) covariate X on various functions of time, $\theta(t|x)$, related to the multistate system. Thus, $F^j(t|x)$ will denote the regression function $Pr\{U^j \leq t | U^j < \infty, X = x\}$ and so on. To this end, we assume the random regressor model that $\{\boldsymbol{S}_i, X_i\}$ are i.i.d., for $1 \leq i \leq n$.

Assume that the data are subject to right censoring and hence not all the transition times are observed. For individual i, let C_i be the (right) censoring time for individual i. Dependent censoring will occur if there are covariables $Z = Z(t)$ that affect both the hazards of future transitions and of being censored. For non-Markov models, future transitions and the censoring hazard may also depend on the past history of the process. We assume we have observations on a covariate process Z that explains the dependence between transition and the censoring mechanism such that, conditional on the values of Z previous to the current time, future transitions and censoring events behave independently. Let T_i^* be the last transition for individual i (in the uncensored experiment) and $T_i = min(T_i^*, C_i)$ be the time for the last event in the censored experiment. We assume that the censoring mechanism satisfies

$$\lambda_C(t|\bar{\boldsymbol{Z}}_i(t), \boldsymbol{S}_i) = \lambda_C(t|\bar{\boldsymbol{Z}}_i(t)),$$

where

$$\lambda_C(t|.) = \lim_{dt \to 0} Pr(C_i \in [t, t+dt)|T_i \geq t, .) ,$$

and $\bar{\boldsymbol{Z}}_i(t) = \sigma(\{\boldsymbol{Z}_i(s) : 0 \leq s < t\})$. Without loss of generality, we assume that the collection Z contains X whose effect we plan to study.

3.2.2. Integrated Transition Hazard

For states $j \neq j'$, let $\alpha_{jj'}(t|x)$ be the local (i.e., given $X = x$) transition hazard of the original (uncensored) chain from state j to j', defined as

$$\alpha_{jj'}(t|x) = \lim_{dt \downarrow 0} Pr\{S(s) = j' \text{ for some } s \in [t, t+dt)|S(t-) = j, X = x\}/dt,$$

where, recall that $S(t-)$ denotes the state occupied just before time t. Let $N_{jj'}$ be the counting process with jumps given by $\Delta N_{jj'}(t) = \sum_{i=1}^{n} I(S_i(t-) = j, S_i(t) = j')$. Also, let $Y_j(t) = \sum_{i=1}^{n} I(S_i(t-) = j)$ be the number of individuals at state j just before time t.

Note that by Fubini's theorem

$$\alpha_{jj'}(t|x) = \lim_{dt \downarrow 0} \frac{E\{dN_{jj'}(t)|S_i(t-) = j, X = x\}}{dt},$$

so that $\alpha_{jj'}(t|x)dt$ counts the limiting number of j to j' transitions made by an individual with covariate $X = x$, in the interval $[t,t+dt)$ and $\alpha_{jj'}(t|x)$ denotes a local version of a 'partially conditional transition rate' of Pepe and Cai (1993). The local cumulative (integrated) transition hazard matrix is given by $\boldsymbol{A}(\cdot|x) = \{A_{jj'}(\cdot|x)\}$, with $A_{jj'}(t|x) = \int_0^t \alpha_{jj'}(s|x)ds$ for $j \neq j'$ with $\alpha_{jj}(t|x) = -\sum_{j \neq j'} \alpha_{jj'}(t|x)$.

To develop non-parametric estimators conditional on a given value of X, we adopt the method proposed by Datta and Satten (2002) in estimating the marginal transition hazard. In their treatment, the estimator of the integrated transition hazard matrix has a Nelson-Aalen form. They estimate the two processes $N_{jj'}$ and Y_j separately using the principle of inverse probability of censoring weights rather than ratio. The reason for using such reweighting approach is that when the data are right censored, the processes Y and N cannot be calculated based on the observed data alone. For the regression case, we propose using kernel estimators similar to the method by Beran (1981) along with IPCW estimation techniques in Datta and Satten leading to our final estimators. This way, we introduce local versions of counting and "number at risk" processes through kernel smoothing to give higher weights to individuals with X covariate values close to the given value x.

As described in the previous paragraph, we let

$$\Delta\widehat{N}_{jj'}(t,x) = \sum_{i=1}^{n} \phi_h(X_i - x)\{\frac{I(S_i(t-) = j, S_i(t) = j', C_i \geq t)}{\widehat{K}_i(t-)}\} \qquad (3.1)$$

and

$$\widehat{Y}_j(t,x) = \sum_{i=1}^{n} \phi_h(X_i - x)\{\frac{I(S_i(t-) = j, C_i \geq t)}{\widehat{K}_i(t-)}\}, \qquad (3.2)$$

where ϕ is a density kernel, $h = h(n) \downarrow 0$ is a bandwidth sequences, $\phi_h = h^{-1}\phi(./h)$, and $\widehat{K}_i(t)$ is an estimate of $K_i(t) = \prod\limits_{s \leq t}[1 - \lambda_C(t|\overline{Z}_i(s))ds]$. Note that the indicator terms in the summands are calculable based on the observed (e.g., right censored) multistate data. Finally, a non-parametric regression estimator of integrated transition hazard is obtained by

$$\widehat{A}_{jj'}(t|x) = \begin{cases} \int_0^t J_j(u,x)\widehat{Y}_j(u,x)^{-1}d\widehat{N}_{jj'}(u,x) & j \neq j' \\ -\sum\limits_{j \neq j'} \widehat{A}_{jj'}(t|x) & j = j', \end{cases}$$

with $J_j(u,x) = I(\widehat{Y}_j(u,x) > 0)$.

In general, $K_i(t)$ does not have a survival function interpretation unless all the Z_i are baseline covariates. A flexible model is recommended in practice for estimating K_i and obtaining $\widehat{K}_i$. This will be discussed later in the chapter.

3.2.3. State Occupation Probabilities

Non-parametric estimation of state occupation probabilities when data are subject to dependent censoring was undertaken in Datta and Satten (2002). We present a brief description of their estimators for the sake of completeness.

The Datta-Satten estimator $\widehat{P}(s,t)$ of the transition probability matrix has an Aalen-Johanson form, and it is given by the product integral

$$\widehat{P}(s,t) = \prod_{(s,t]}(I + d\widehat{\boldsymbol{A}}(u)),$$

where I is the identity matrix. This is a finite product taken over all distinct observed transition times. Finally, the Datta-Satten estimator $\widehat{p}_j(t)$ of the marginal state j occupation probability at time t is given by

$$\widehat{p}_j(t) = \sum_{k=0}^{M} \frac{\widehat{Y}_k(0+)}{n} \widehat{p}_{kj}(0,t).$$

Thus, the marginal state occupation probability of state j is obtained by evaluating the product limit of transition hazards and averaging it with respect to the initial distribution of state occupation.

The non-parametric regression estimation of these quantities can be obtained by localizing the calculation via kernel weights. In particular, the non-parametric regression estimator of the state occupation probabilities conditional on a given value of $X = x$, $p_j(t|x) = Pr\{S(t) = j|X = x\}$, is given by

$$\widehat{p}_j(t|x) = \sum_{k=0}^{M} \frac{\widehat{Y}_k(0+|x)}{n} \widehat{p}_{kj}(0,t| x),$$

where $\widehat{P}(0,t|x)$ is the kjth element of the matrix $\widehat{P}(0,t|x) = \prod_{(0,t]} (I + d\widehat{\boldsymbol{A}}(u|x))$ obtained by product integration of the matrix $\widehat{\boldsymbol{A}}(t)$ defined in the previous section and $\widehat{Y}_k(t|x) = \widehat{Y}_k(t,x)/\{n^{-1}\sum_{i=1}^{n} \phi_h(X_i - x)\}$.

3.2.4. State Entry and Exit Distributions

For this section, we assume that the multistate system is progressive (e.g., it does not contain a cycle) so that a given state is entered at most once. Recall that for any state $j > 0$, U^j is the entry time for state j amongst individuals who ever enter state j. Let F^j denote the corresponding distribution function conditional on $X = x$,

$$F^j(t|x) = P\{U^j \leq t|U^j < \infty, X = x\},$$

where we take $F^0(t|x) = 1$, for all x and t. Let S^j is the collection of all states which proceed state j in the progressive model. Then estimators of entry time distributions of state j is given by

$$\widehat{F}^j(t|x) = \frac{\sum\limits_{k\in\{j\}\cup S^j} \widehat{P}_k(t|x)}{\sum\limits_{k\in\{j\}\cup S^j} \widehat{P}_k(\infty|x)},$$

where $\widehat{P}_k(\infty|x) = \lim\limits_{t\to\infty} \widehat{P}_k(t|x)$.

Analogous to above, V^j is the exit time for state j of individuals who will ever enter state j. Let G^j denote the corresponding distribution function conditional on $X = x$,

$$G^j(t|x) = Pr\{V^j \leq t|U^j < \infty, X = x\},$$

where we take $G^j(t|x) = 0$, for all x and t, if j is a terminal node (e.g., an absorbing state). For a transient state j, $\widehat{G}^j(t|x)$ is taken to be the sum of estimated state occupation probabilities of all states that proceed state j in the progressive system normalized by sum of probabilities of ever entering state j. Another option for calculating the exit distribution function will be to take the normalized sum of estimated state occupation probabilities of all states that come after state j. The former method seems to be more appropriate approach since all individuals who entered state j may not leave the state j by the end of study. Hence, excluding these individuals in normalization process is not suitable. Mathematically speaking, the estimated state exit time distribution is given by

$$\widehat{G}^j(t|x) = \frac{\sum\limits_{k\in S^j} \widehat{P}_k(t|x)}{\sum\limits_{k\in\{j\}\cup S^j} \widehat{P}_k(\infty|x)}.$$

3.2.5. State Waiting Time Distributions

In order to calculate waiting time distributions, a different form of reweighting similar to the one proposed in Satten and Datta (2002) is necessary for handling censoring, since waiting times are measured from state entry whereas right censoring is measured in calendar time. Once again, assume that a transient state j can be entered at most once. The local estimated counting processes for waiting time in a given state j is a jump process with jump size equal to

$$\Delta \widehat{N}_j^W(t,x) = \sum_{i=1}^{n} \frac{\phi_h(X_i - x) I\{W_i^j = t, C_i \ge V_i^j\}}{\widehat{K}_i(V_i^j -)},$$

which can be computed based on the available right censored data since if $C_i \ge V_i^j$ then the state j waiting time W_i^j is available. The inverse weighting factor is essentially the estimated conditional probability of the event $\{C_i \ge V_i^j\}$, given $\{V_i^j, W_i^j\}$. Next, the size of the "at risk" set of state j waiting time is estimated by

$$\widehat{Y}_j^W(t,x) = \sum_{i=1}^{n} \frac{\phi_h(X_i - x) I\{t \le W_i^j, C_i \ge t + U_i^j, U_i^j < \infty\}}{\widehat{K}_i((t + U_i^j) -)}.$$

Note that, this quantity can be computed based on the available data and, in particular, an individual may contribute to the local “at risk” set even if its exit time is right censored. Finally, the regression estimator of state j waiting time distribution is obtained by a Kaplan-Meier type product limit formula using these two sets

$$\widehat{H}^j(t|x) = 1 - \prod_{s \le t} \left(1 - \frac{d\widehat{N}_j^W(s,x)}{\widehat{Y}_j^W(s,x)} \right).$$

3.2.6. Estimation of λ_C

In estimating the IPCW-weights, we apply a highly flexible and non-parametric additive regression model in which the regression coefficients are allowed to vary over time.

Aalen's linear hazard model (1980; 1989) has a linear structure given by

$$\lambda_C(t|\overline{Z}_i(t)) = \sum_{k=0}^{J} \beta_k(t)U_{ik}(t),$$

where $U_{i0}(t) \equiv 1$ and $U_{ik}(t) = f_k(\overline{Z}_i(t))$ for $k = 1, ..., J$ are possibly time-dependent functions of the past history of the covariate process for subject i. The $\beta_k(t)$ are (unknown) regression functions that measure the effect of respective covariate functions on the risk of censoring. Let $\delta_i = I(C_i > T_i^*)$ be the indicator of whether the ith individual was ever censored. Define $U_i(t) = (U_{i1}(t), ..., U_{iJ}(t))$; then Aalen's estimator has simple closed form given below:

$$\widehat{\Lambda}^i_C(t|\overline{Z}_i(t)) : = \int_0^t \widehat{\lambda}_C(s|\overline{Z}_i(s))ds = \sum_{j=1}^{n} I(T_j \le t)(1-\delta_j)U_i(T_j)\mathcal{R}^{-1}(T_j)U_j(T_j), \quad t \le T_i$$

with

$$\mathcal{R}(t) = \sum_{i=1}^{n} I(T_i \ge t)\boldsymbol{U}_i(t)\boldsymbol{U}_i^T(t).$$

Using this model, a correction for informative censoring can be obtained which is close to that achieved using the correct model for $\lambda_C(t|\overline{Z}_i(t))$. An important special case is when censoring depends on the current stage occupied which corresponds to the internal covariate

$$U_{ik}(t) = f_k(\overline{Z}_i(t)) = I[S_i(t) = j].$$

3.3. Simulation Studies

To illustrate the use of our estimators in a controlled setting, a number of Monte Carlo experiments were performed. We have based our simulations on a hypothetical progressive model with branches described by a five state system (Figure 3.1). In order to cover a variety of scenarios, three different simulation examples are presented below.

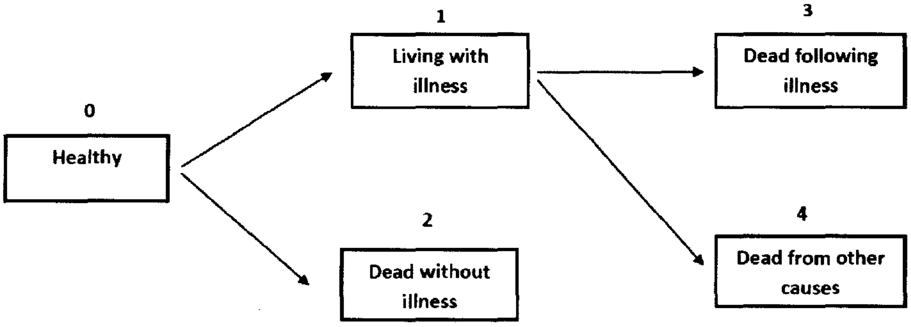


Figure 3.1. A five-state illness-death model

3.3.1. Conditionally Semi-Markov Transition Times

For each person, the single covariate X was generated from a normal distribution with mean parameter 5 and standard deviation parameter of 0.5. We assumed all individuals start in State 0 (well) at time zero, and may either progress to State 1 or State 2. Each patient at State 0 had a 60% chance (controlled by a Bernoulli variable that is independent of the event times) of following the $0 \rightarrow 1$ arm and a 40% chance of following the $0 \rightarrow 2$ arm. Furthermore, patients who entered State 1, would subsequently reach States 3 or 4 with arm probabilities (controlled by another independent Bernoulli variable) 0.6 and 0.4, respectively. To generate the event times in a conditionally semi-Markov model, we used both lognormal and Weibull distributions. For lognormal simulations, the waiting times in State 0 were generated from a lognormal distribution with log-mean parameter 0 and log-scale parameter 0.5 and for individuals traversing the $0 \rightarrow 1$ arm, the State 1 waiting times were generated using another independent lognormal distribution with log-mean parameter 0 and log-scale parameter 1. For the Weibull simulations, the State 0 waiting times were generated from a Weibull distribution with shape parameter 2 and scale 1. For patients traversing the path $0 \rightarrow 1$, the state waiting times were obtained from another independent Weibull distribution with shape parameter

0.5 and scale 1. We assume that the simulated data are subject to dependent right censoring induced by the covariate X that affects both the transition and censoring times of an individual. In order to meet this assumption, the waiting times were multiplied by the person-specific frailty variable which was generated from normal distribution with mean parameter $X/5$ and standard deviation parameter 0.1. The choice of the mean and the variance parameters of the normal distribution generating X guarantees that the frailty variable will have positive values with probability nearly 1. Censoring times were generated from the lognormal distribution with parameters $\mu = 0.5$ and $\sigma^2 = 2$. In order to make the censoring dependent on covariate X, the censoring times were multiplied by another frailty variable which were generated from normal distribution with mean parameter $X/5$ and standard deviation parameter 0.3. Note that, in this example, only an external covariate X affects the censoring mechanism.

3.3.2. Conditionally Markov Transition Times

A Markov model is based on the assumption that the transition intensities depend only on the calendar time and current state occupied. We generated the event times in a Markov setup as follows. Individuals started at State 0 at time zero; 60% of the study population at State 1 took the $0 \to 1$ path. The branch proportion was controlled by a Bernoulli variable independent of the transition times. There are two transition times need to be generated and both of them have a common hazard following either a Weibull distribution with shape parameter 1 and scale 0.5 or a lognormal distribution with log-mean parameter 0 and log-scale parameter 0.5. We randomly generated the first transition times T_1 from each of these distributions and then multiplied them by the person-specific frailty variable which was generated from normal distribution with mean parameter $X/5$ and standard deviation parameter 0.1. For individuals who ever traversed the State 1, the second transition times T_2 were obtained by

$$T_2 = D^{-1}(D(T_1) + R_2\{1 - D(T_1)\}),$$

where $D(.)$ denoted the distribution function for the common hazard; $D^{-1}(.)$ was the corresponding quantile function and R_2 was a random number generated from uniform distribution [0,1] and independent of T_1. The second transition times T_2 were also multiplied by the person-specific frailty. Censoring times were generated from a uniform distribution ranging from time 0 to τ^*, the largest transition time generated. The censoring times were multiplied by another frailty variable which were generated from normal distribution with mean parameter $X/5$ and standard deviation parameter 0.3. Similar to previous example, only covariate X which is an external covariate affects censoring mechanism.

3.3.3. Conditionally Markov Models with State Dependent Censoring

In this simulation study, we let the censoring hazard, at time t differ among the subjects according to their current state occupation. That is, the covariate process $Z_i(t)$ that affects the censoring mechanism for subject i, consists only of the internal covariate of the state occupied at time t.

Similar to the previous example, the state waiting times are generated as follows. Individuals started at State 0 at time zero; 60% of the study population at State 1 took the $0 \rightarrow 1$ path. The branch proportion was controlled by a Bernoulli variable independent of the transition times. Transition times generated from a lognormal distribution log-mean parameter 0 and log-scale parameter 0.5. For individuals who made transition to the State 1, the second transition times T_2 were obtained by

$$T_2 = D^{-1}(D(T_1) + R_2\{1 - D(T_1)\}),$$

where $D(.)$ denoted the distribution function for the common hazard; $D^{-1}(.)$ was the corresponding quantile function and R_2 was a random number generated from uniform distribution [0,1] and independent of T_1. Censoring times for individuals in State 1 was generated from a Weibull distribution with shape parameter 1.5 and scale 3. The second censoring time which belongs to those who made the transition to State 1 were generated independently from another Weibull distribution with shape parameter 2 and scale 3 conditional on that the generated value be larger than the first transition time.

The state occupation probabilities of described examples were estimated by the proportion of subjects observed in each state at time t in complete data. The smoothing step in each of the estimation processes was based on a common bandwidth selector and using normal kernels. The R package "KernSmooth" was used to this end (http://cran.r-project.org/doc/packages/KernSmooth.pdf). The bandwidth was taken as dpik, the data based bandwidth selector of Wand and Jones (1995).

The non-parametric estimators for a sample of size 1000 and 10000 generated as above with given $x =$ median of the covariate X. The estimators with sample size of 10000 are used as benchmarks which were virtually identical to the empirical probabilities using the set of complete transition times (not shown in the plots). For the sake of space, we only provide limited number of figures in this chapter. The conclusions based on these figures are going to be the same for the rest. Figure 3.2 displays the state occupation probability results of conditional semi-Markov data with lognormal transition times. and state dependent censoring model respectively. The non-parametric regression estimates of state occupation probabilities in Weibull conditional Markov model with state dependent censoring is shown in Figure 3.3

Overall, estimators for $n = 1000$ and $n = 10000$ are in good agreement, the later being virtually identical to the true empirical curves (not shown separately).

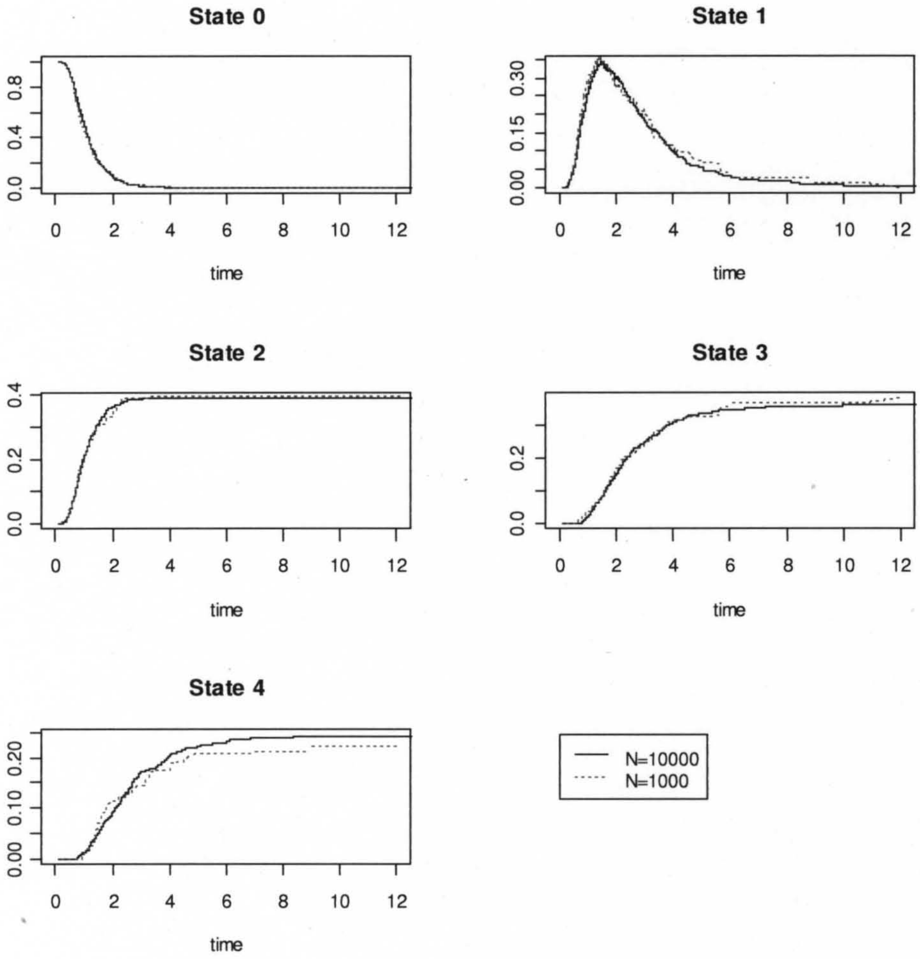


Figure 3.2. The non-parametric regression estimates of state occupation probabilities in a five-state lognormal conditional semi-Markov model given the covariate $X = 5$ which was the median of the covariate distribution.

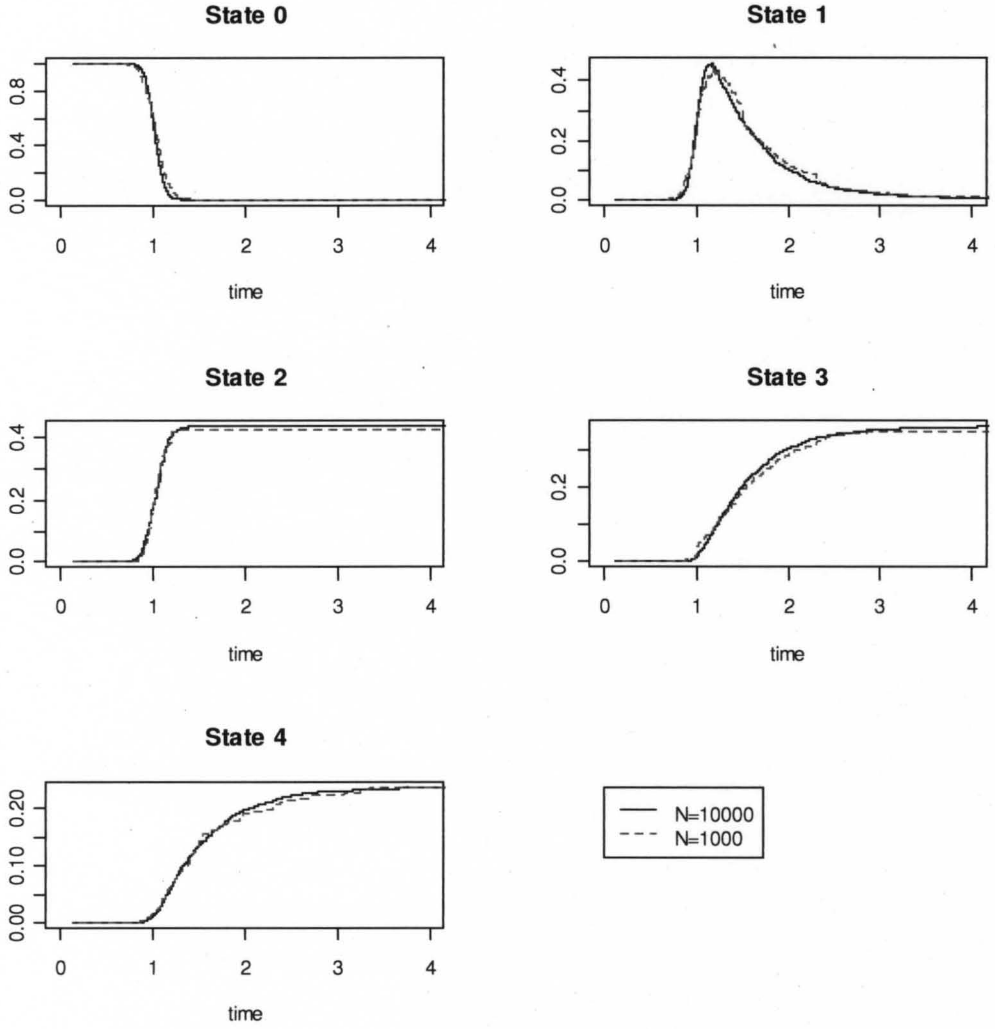


Figure 3.3. The non-parametric regression estimates of state occupation probabilities in a five-state, conditional Markov with state dependent censoring model Weibull conditional Markov model given the covariate $X = 5$ which was the median of the covariate distribution.

Next we assessed the global performance of the estimators by the L_1 distance

$$\Delta : = E\int |\widehat{\theta}(t|x) - \widehat{\theta}_E(t|x)|dF_n(t|x),$$

where $\widehat{\theta}(x)$ and $\widehat{\theta}_E(x)$ denote respectively our proposed estimator of θ and its empirical counterpart based on the complete data. Here, θ is either a state occupation probability, a state entry time distribution function or a state exit time distribution function. The integrating measure in the definition of the L_1 distance was taken to be the empirical distribution function of true event times $F_n(t|x) = n^{-1}\sum_{i=1}^{n} I\{T_i \leq t|x\}$; $\Delta = 0$ means that they are in a complete agreement on the support of the event times. We calculated Δ via Monte Carlo averaging with the replication size of 5000. The calculations were performed based on three given x values; covariate first quartile, median and third quartile, respectively.

The L_1 results of state occupation probabilities for conditional semi-Markov with Weibull and lognormal transition times are provided in Tables 3.1 and 3.2, respectively. Table 3.3 and 3.4 list the results for conditional Markov simulations with Weibull and lognormal transition times. Table 3.5 gives the result of state dependent censoring example. For all five simulation settings, the L_1 values decrease with increasing sample size. As the given x value gets away from the center of the covariate (X), the L_1 values increase. The absorbing states have smaller L_1 errors than transient state in all models. The results for conditional semi-Markov and Markov setups are comparable indicating that these nonparametric regression estimators work well regardless of any structural assumptions on the multistate system.

The L_1 calculation results for entry/exit time distributions with Weibull transition times in the conditional semi-Markov setup is given in Table 3.6. Similar to state occupation probability result, L_1 values have a decreasing trend by increasing sample size.

A set of scatter plots corresponding to various simulation setting show an approximate linear relationship of the logarithms of the L_1 distances with the logarithms of the sample size for each of these estimators suggesting that the L_1 values converge to zero at the rate of n^{-b}, for some b. The log mean L_1 distance plots are listed in Figures 3.4-3.8.

Table 3.1. The L_1 distances between non-parametric regression estimators of state occupation probabilities based on right censored data and complete data in a five-state conditional semi-Markov Weibull model. The estimates are based on a Monte Carlo sample size of 5000; all standard errors were less than 0.09

	given x: First Quartile				given x: Median				given x: Third Quartile			
	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$
p_0	0.017	0.009	0.007	0.003	0.015	0.007	0.006	0.002	0.016	0.008	0.006	0.003
p_1	0.200	0.094	0.066	0.023	0.197	0.091	0.064	0.021	0.198	0.092	0.065	0.024
p_2	0.017	0.009	0.007	0.003	0.015	0.007	0.006	0.002	0.015	0.008	0.006	0.003
p_3	0.021	0.012	0.009	0.004	0.019	0.010	0.008	0.004	0.019	0.011	0.008	0.004
p_4	0.018	0.010	0.008	0.004	0.015	0.008	0.006	0.003	0.015	0.009	0.007	0.004

Table 3.2. The L_1 distances between non-parametric regression estimators of state occupation probabilities based on right censored data and complete data in a five-state conditional semi-Markov lognormal model. The estimates are based on a Monte Carlo sample size of 5000; all standard errors were less than 0.09

	given x: First Quartile				given x: Median				given x: Third Quartile			
	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$
p_0	0.021	0.009	0.007	0.004	0.020	0.009	0.006	0.003	0.022	0.009	0.007	0.004
p_1	0.420	0.152	0.099	0.043	0.417	0.149	0.097	0.041	0.419	0.151	0.099	0.043
p_2	0.016	0.008	0.006	0.003	0.014	0.007	0.005	0.002	0.016	0.008	0.006	0.003
p_3	0.014	0.007	0.005	0.002	0.012	0.006	0.004	0.001	0.015	0.005	0.005	0.002
p_4	0.011	0.005	0.004	0.002	0.010	0.004	0.003	0.001	0.011	0.005	0.004	0.002

Table 3.3. The L_1 distances between non-parametric regression estimators of state occupation probabilities based on right censored data and complete data in a five-state conditional Markov Weibull model. The estimates are based on a Monte Carlo sample size of 5000; all standard errors were less than 0.05

	given x: First Quartile				given x: Median				given x: Third Quartile			
	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$
p_1	0.023	0.009	0.006	0.003	0.022	0.008	0.005	0.002	0.023	0.009	0.006	0.003
p_1	0.142	0.045	0.029	0.012	0.141	0.044	0.028	0.011	0.0143	0.045	0.029	0.012
P_2	0.026	0.010	0.007	0.003	0.025	0.009	0.006	0.002	0.026	0.010	0.007	0.003
p_3	0.049	0.019	0.013	0.006	0.048	0.018	0.012	0.005	0.050	0.020	0.014	0.006
p_4	0.046	0.017	0.012	0.005	0.045	0.016	0.011	0.004	0.046	0.017	0.012	0.005

Table 3.4. The L_1 distances between non-parametric regression estimators of state occupation probabilities based on right censored data and complete data in a five-state conditional Markov lognormal model. The estimates are based on a Monte Carlo sample size of 5000; all standard errors were less than 0.05

	given x: First Quartile				given x: Median				given x: Third Quartile			
	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$
p_0	0.015	0.007	0.005	0.003	0.014	0.006	0.004	0.002	0.015	0.007	0.005	0.003
p_1	0.139	0.091	0.077	0.052	0.138	0.090	0.076	0.051	0.136	0.091	0.077	0.052
p_2	0.014	0.006	0.004	0.002	0.013	0.005	0.003	0.001	0.014	0.006	0.004	0.002
p_3	0.016	0.008	0.007	0.005	0.015	0.007	0.006	0.004	0.016	0.008	0.007	0.005
p_4	0.017	0.008	0.006	0.003	0.016	0.007	0.005	0.002	0.017	0.008	0.006	0.003

Table 3.5. The L_1 distances between non-parametric regression estimators of entry/exit time distributions based on right censored data and complete data in a five-state conditional semi-Markov Weibull model. The estimates are based on a Monte Carlo sample size of 5000; all standard errors were less than 0.02

	given x: First Quartile				given x: Median				given x: Third Quartile			
	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$	$n = 100$	$n = 500$	$= 1000$	$n = 5000$
Ex0	0.017	0.009	0.006	0.003	0.015	0.008	0.006	0.003	0.016	0.008	0.006	0.003
En1	0.021	0.010	0.007	0.004	0.018	0.009	0.006	0.004	0.019	0.009	0.007	0.003
Ex1	0.048	0.025	0.019	0.009	0.043	0.022	0.016	0.009	0.045	0.023	0.017	0.009
En2	0.029	0.015	0.011	0.006	0.026	0.013	0.010	0.006	0.027	0.014	0.011	0.005
En3	0.064	0.033	0.025	0.013	0.056	0.029	0.022	0.013	0.059	0.030	0.023	0.012
En4	0.079	0.042	0.031	0.016	0.070	0.036	0.027	0.016	0.073	0.038	0.029	0.015

Table 3.6. The L_1 distances between non-parametric regression estimators of state occupation probabilities based on right censored data and complete data in a conditional Markov state dependent censoring model. The estimates are based on a Monte Carlo sample size of 5000; all standard errors were less than 0.07

	given x: First Quartile				given x: Median				given x: Third Quartile			
	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$
p_0	0.013	0.006	0.005	0.003	0.012	0.006	0.004	0.002	0.013	0.005	0.005	0.003
p_1	0.172	0.082	0.063	0.036	0.171	0.081	0.061	0.034	0.173	0.083	0.063	0.037
p_2	0.011	0.005	0.004	0.003	0.010	0.004	0.003	0.002	0.012	0.005	0.004	0.003
p_3	0.014	0.008	0.006	0.004	0.013	0.007	0.005	0.003	0.014	0.008	0.006	0.004
p_4	0.015	0.008	0.006	0.004	0.014	0.007	0.005	0.004	0.015	0.008	0.006	0.004

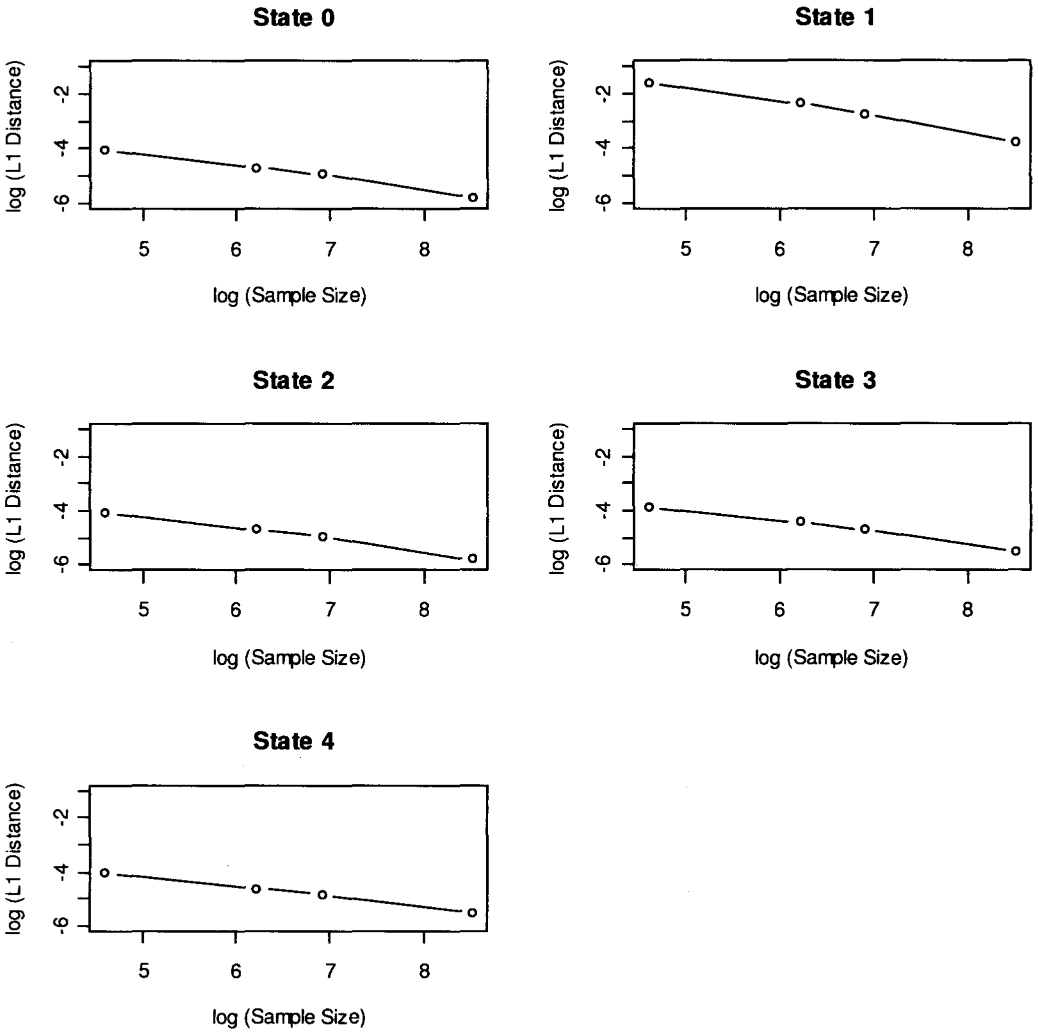


Figure 3.4. Plots of log L_1 distances between non-parametric regression estimators of state occupation probabilities based on right censored data and complete data in a five-state Weibull conditional semi-Markov model given the covariate $X = 5$ which was the median of the covariate distribution.

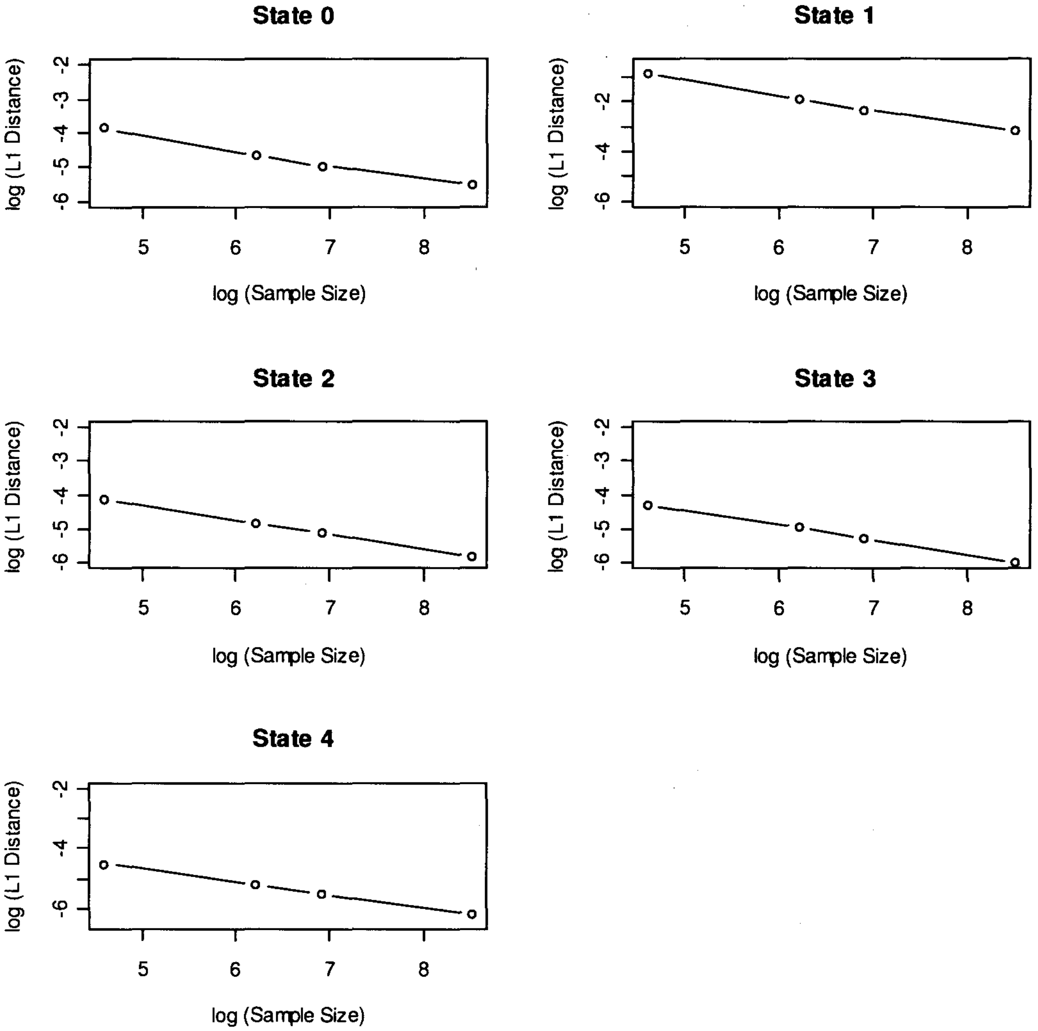


Figure 3.5. Plots of log L_1 distances between non-parametric regression estimators of state occupation probabilities based on right censored data and complete data in a five-state lognormal conditional semi-Markov model given the covariate X = 5 which was the median of the covariate distribution.

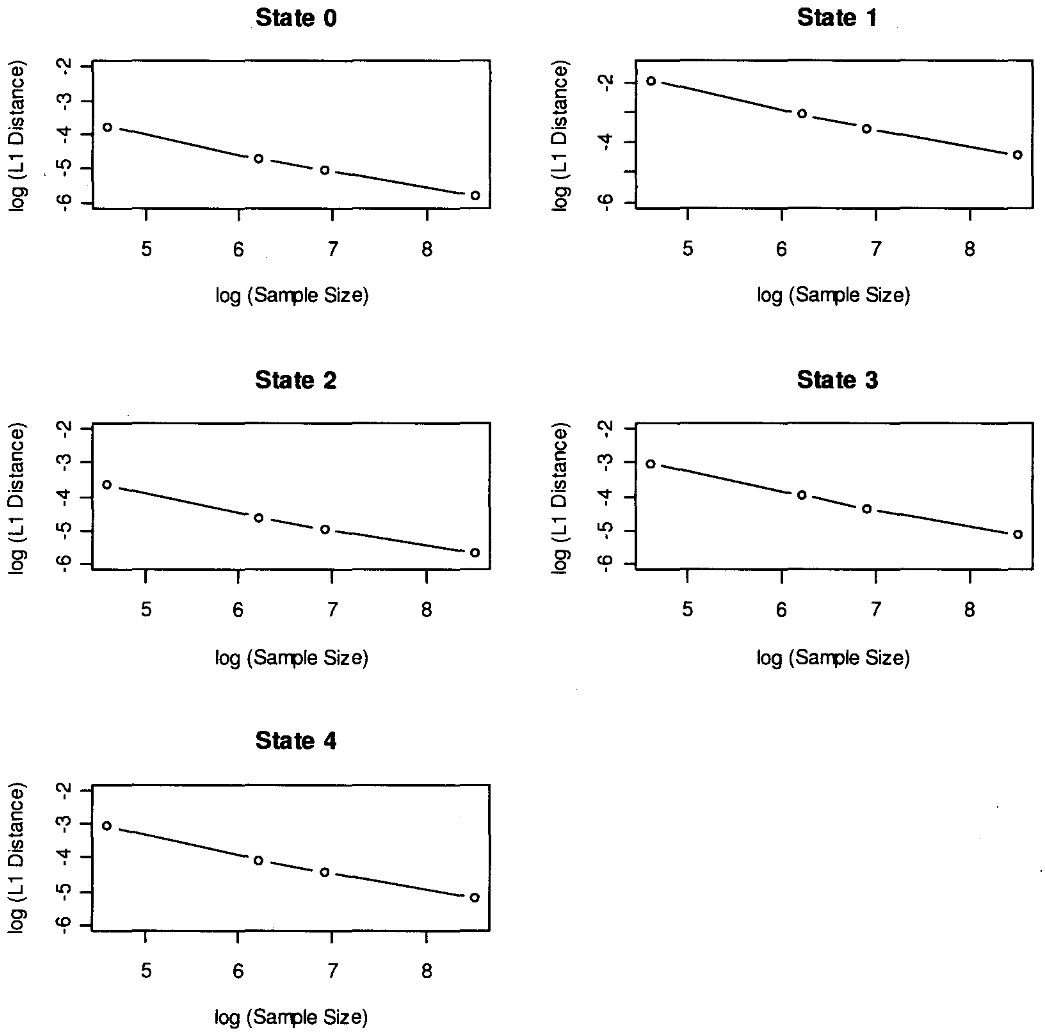


Figure 3.6. Plots of log L_1 distances between non-parametric regression estimators of state occupation probabilities based on right censored data and complete data in a five-state Weibull conditional Markov model given the covariate $X = 5$ which was the median of the covariate distribution.

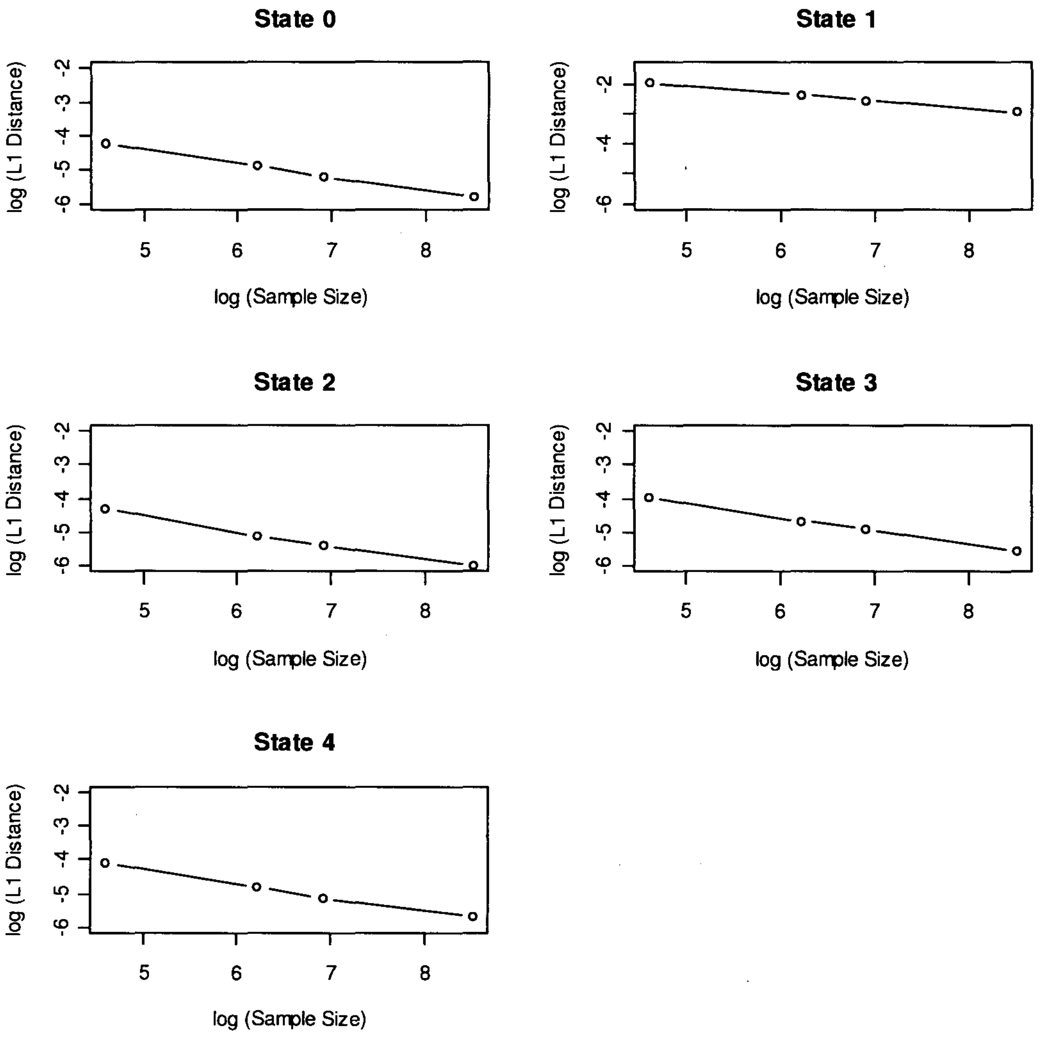


Figure 3.7. Plots of log L_1 distances between non-parametric regression estimators of state occupation probabilities based on right censored data and complete data in a five-state lognormal conditional Markov model given the covariate $X = 5$ which was the median of the covariate distribution.

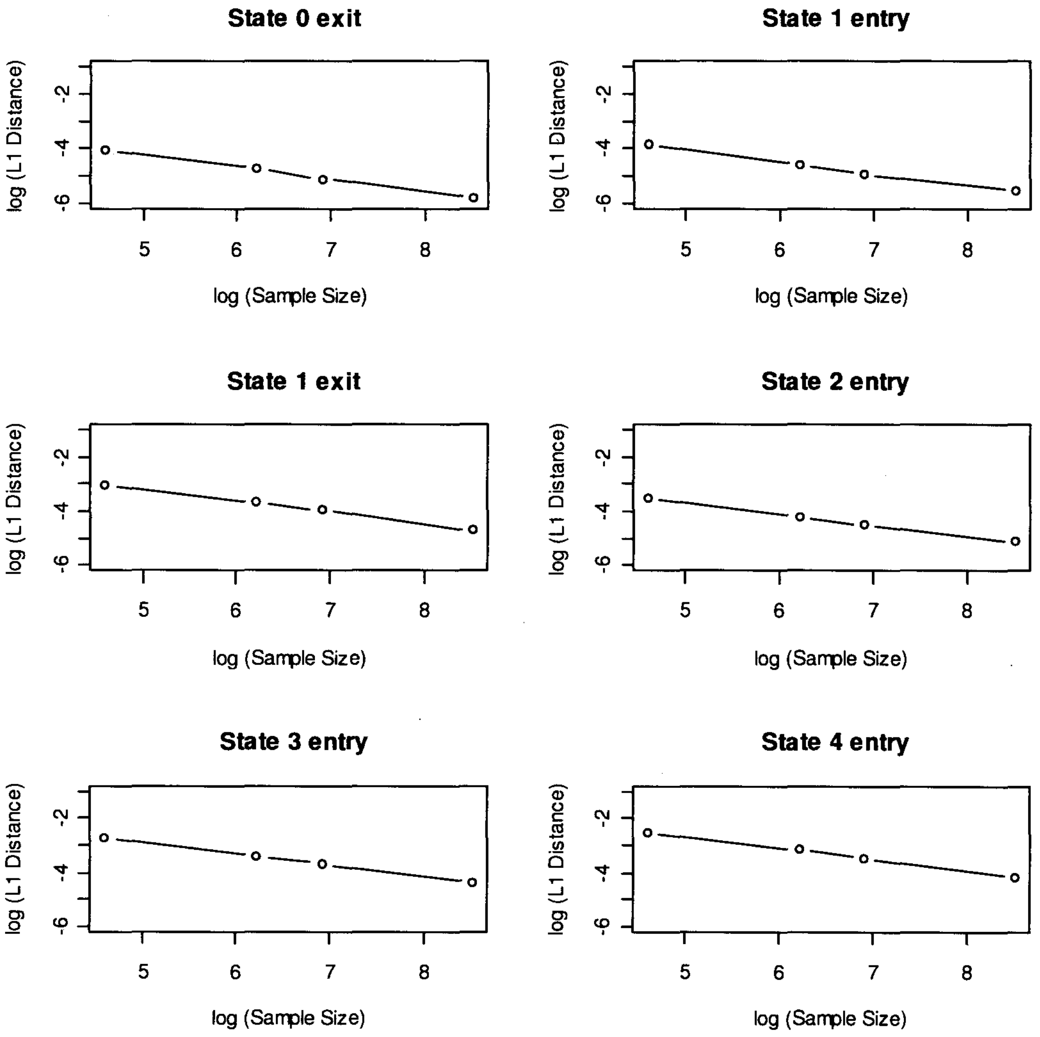


Figure 3.8. Plots of log L_1 distances between non-parametric regression estimators of entry/exit distributions based on right censored data and complete data in a five-state lognormal conditional Markov model given the covariate $X = 5$ which was the median of the covariate distribution.

The global performance of the proposed estimator of state waiting time distribution functions $\widehat{H}^j$ was evaluated by calculating the mean absolute distance

$$\Delta_{W^j} : = E\int |\widehat{H}^j(t|x) - \widehat{H}^j_E(t|x)| dF_{n,j}(t|x),$$

where the integrating measure in the above definition was taken to be the empirical distribution function of waiting times of those who entered the state j; i.e., $F_{n,j}(t|x) = n^{-1}\sum_{i=1}^{n} I\{W_i^j \leq t|x\}$. We calculated Δ_{W^j} via Monte Carlo averaging with a replication size of 5000. As before, the calculations were performed based on three given x values corresponding to the first quartile, median and third quartile, respectively, of the covariate distribution. The distribution function of single transient state; State 1 waiting times was estimated using our estimators and compared with the corresponding quantities for the complete data. The corresponding Δ_{W^1} values with lognormal transition times in the conditional Markov setup are reported in Table 3.7. State 1 waiting time distribution plots in a conditional Markov setup with lognormal transition times are given in Figure 3.9. The scatter plot of the logarithms of the L_1 distances versus the logarithms of the sample size of waiting time distributions is given in Figure 3.10. The performance of the estimator appears to be reasonable and improves with increasing sample size.

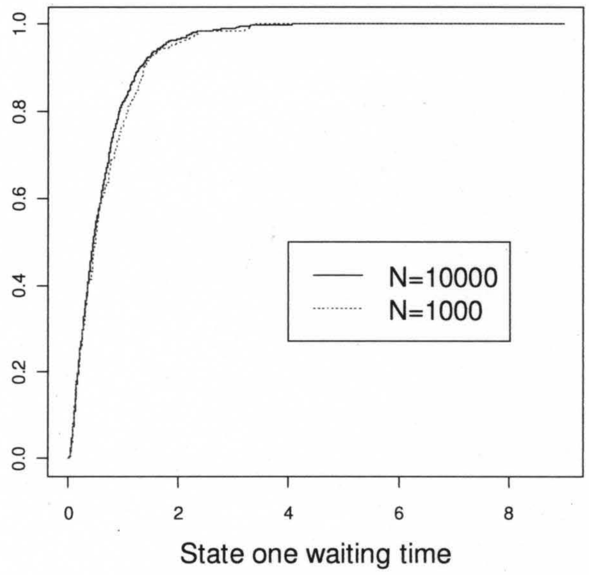


Figure 3.9. The non-parametric regression estimates of State 1 waiting time distribution in a five-state lognormal conditional Markov model given the covariate $X = 5$ which was the median of the covariate distribution.

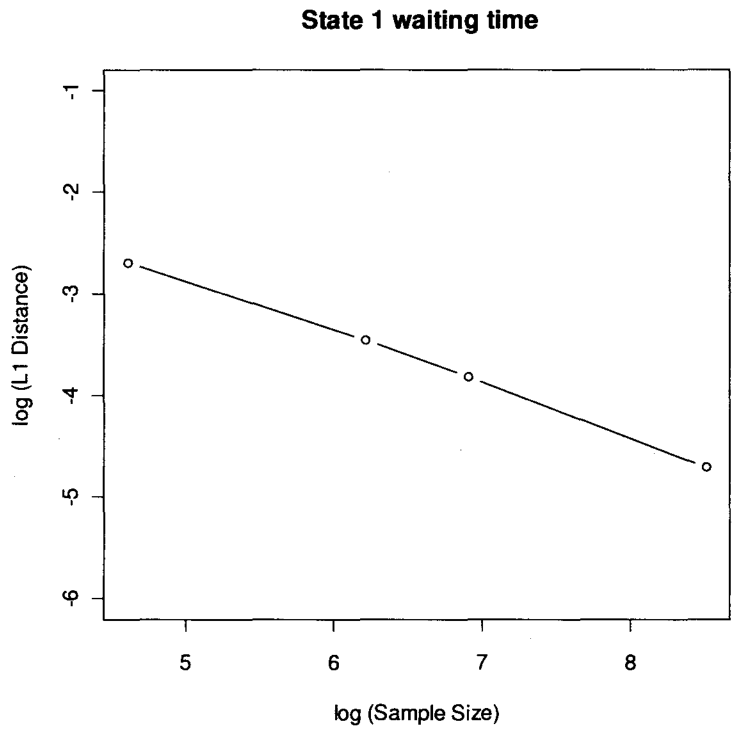


Figure 3.10. Plot of log L_1 distances between non-parametric regression estimators of State 1 waiting time distribution based on right censored data and complete data in a five-state lognormal conditional Markov model given the covariate $X = 5$ which was the median of the covariate distribution.

Table 3.7. The L_1 distance between non-parametric regression estimators of State 1 waiting time distributions based on right censored and complete data in a five-state conditional Markov lognormal model. The estimates are based on a Monte Carlo sample size of 5000; all standard errors were less than 0.04

	Sample size			
Covariate X equals	100	500	1000	5000
First Quartile	0.068	0.033	0.023	0.009
Median	0.067	0.032	0.021	0.008
Third Quartile	0.068	0.034	0.022	0.009

3.4. Bone Marrow Transplant Example

To illustrate our estimators using real data, we considered 137 cancer patients (80 males and 57 females) who underwent bone marrow transplant reported in Coplan *et al.* (1991). To study the clinical progression of these patients, we defined seven states. Patients enter State 2 if patients platelet levels return to normal before acute GVHD develops (State 3) or enter State 5 if acute GVHD develops before their platelet levels return to normal (State 6). The allowable transitions between states are shown in Figure 3.11. In addition to the times to acute GVHD and platelet recovery, the information on times of death and relapse of the underlying disease as well as censoring were available. The observation time on each patient is considered as disease free survival time, measured in days from the time of transplantation to end of follow-up due to relapse, death or censoring. Three patients that developed chronic GVHD but not acute GVHD and the time of chronic GVHD was the last event for, were dropped from the analysis. It should be noted that, while State 7 would be an absorbing state, patients may remain in any state for an arbitrary length of time and do not necessarily all progress to State 7 by the end of the study period. The surviving patients at the end of the study were considered to be censored at the end of the study time.

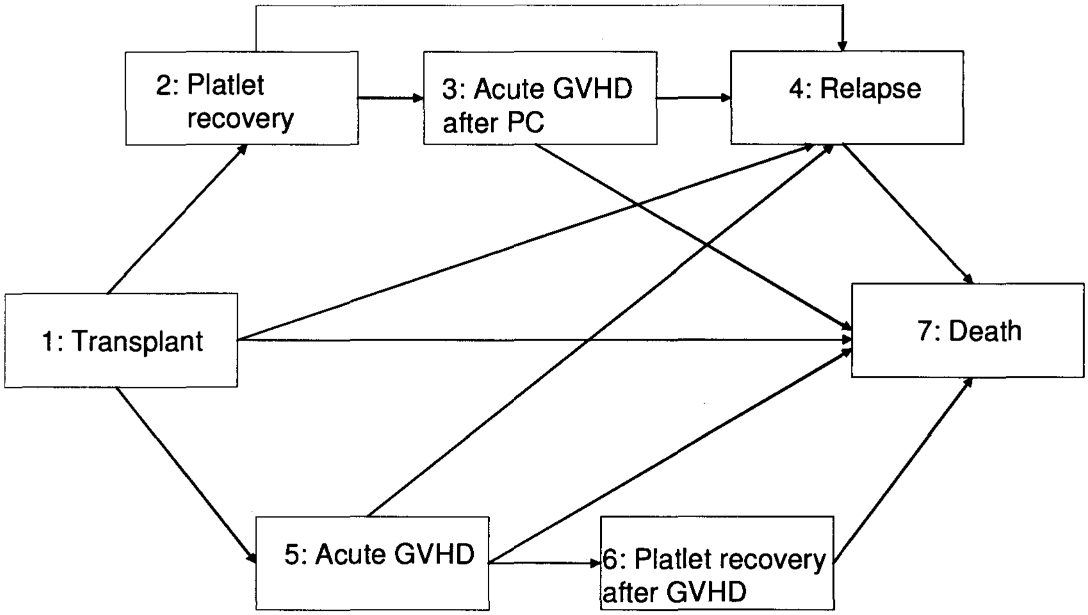


Figure 3.11. Network of states used in the transplant data described in Section 3.5.

Table 3.8 summarizes the observed transitions between these states. Diagonal elements in the table correspond to censored observations; off-diagonal elements count observed transitions, so, e.g., 32 patients relapsed after their platelet levels has returned to normal (i.e., moved from state 2 to state 4). The patient age at transplant ranged from 7 years old to 52. We considered patient age as the single covariate in the regression analysis.

Table 3.8. Observed transitions in the bone marrow transplant data described in Section 3.4

				To			
From	1	2	3	4	5	6	7
1	0	114	0	3	7	0	10
2		44	19	32	0	0	19
3			9	3	0	0	7
4				0	0	0	40
5				0	1	3	3
6				2	0	0	1
7							80

The estimated non-parametric state occupation probabilities of the various states are shown in Figure 3.12. For illustration, estimates are provided given the covariate, patient age, set at 20 and 40, respectively. Also, we have included "patient age" in the Aalen's regression model for the censoring hazard. For patients of both ages, we observe that the occupation probability of State 1, which is the initial state, decreases rapidly, while the occupation probabilities of transient states (States 2,3,4,5 and 6) have increasing trend at the beginning and decreasing at the end. State 7, which is an absorbing state has an increasing state occupation probability with time. The state occupation probabilities of States 4 and 6 reached zero as time progressed indicating that no patient remained at States 4 or 6 by the end of the study. In other words, no censoring observation was observed in these two states after the bone marrow transplantation. Platelet levels of great majority of the patients return to normal before experiencing acute GVHD (State 2). Then subsequently, the patients who exit State 2 more frequently experience relapse (State 4), death (State 7) or acute GVHD (State 3) respectively.

Comparing the state occupation probabilities of patients at age 20 and 40, we observe a number of key differences. For instance, individuals who are at the age of 40 and their platelet levels return to normal before experiencing acute GVHD (State 2), tend to relapse (State 4) more than those who are at age 20, while patients at age 20 experience acute GVHD after their platelet levels return to normal (State 3) with higher probability. Also, patients at age 40, experience death more quickly that those who are at age 20. Since we don't have enough patients at age 20 who experience acute GVHD before their platelet levels return to normal (State 5), and subsequently State 6, the comparison of state occupation probabilities of patients at age 20 and 40 for these two states is not informative.

The non-parametric regression estimates of entry and exit times are shown in Figures 3.13-3.14 given the covariate, age, set at 20 and 40, respectively. About 80% of patients exit State 1 (bone marrow transplantation) after 40 days. Approximately 40% of individuals of age 40 versus 30% of patients of age 20 experience death five months after transplantation. For illustration purposes, the waiting time distributions of States 2, 3 and 5 are also given in Figure 3.14. The probability that the patients who develop acute GVHD after experiencing normal platelet levels (State 3), relapse or die 16 months after transplantation is 93% at age 40 versus 45% at age 20. However, they do not experience transition out of State 3 after approximately 2 years in State 3, despite additional 5 years of follow-up.

Next we considered a model that allows the hazard of censoring to vary between states in addition to the effect of external covariate "patient age". For this purpose, we examined the cumulative censoring hazard for non-absorbing states shown in Figure 3.16. We exclude state 1, 4 and 6 from this plot since no censoring was observed in any of these states. The effect of censoring hazard on the contribution each subject makes to the weighted analysis can be seen in Figure 3.17, which plots $\widehat{K}_i$ for each study participant. To compress the time scale, we numbered each event time (transition or censoring event) and plotted $\widehat{K}_i$ as a function of this integer time scale, which we refer to as time order. Upon a visual examination of Figure 3.17, we see that the $\widehat{K}_i$'s divide study participants into one of three groups. The first (characterized by having $\widehat{K}_i > 0.9$ at the 150th event in Figure 3.17) corresponds to individuals whose last event happened before the first censoring time. The second (characterized by having $0.78 < \widehat{K}_i < 0.85$ at the 155th event in figure 3.17) corresponds to individuals who experienced State 3. The third (characterized by having $\widehat{K}_i < 0.7$ at the 150th event) as indicated by the lower curves in Figure 3.17.

The estimated state occupation probabilities, using the internal covariate of the state occupied at time t as well as "patient age" as the external covariate that affect the censoring mechanism for subject i shown in Figure 3.18. Waiting time distributions are also given in Figure 3.19. We see that the state occupation probabilities and waiting time distributions shown in Figure 3.18-3.19 differ very little from those given in Figures 3.12 and 3.15 in which only the external covariate "patient age" is assumed to affect censoring mechanism.

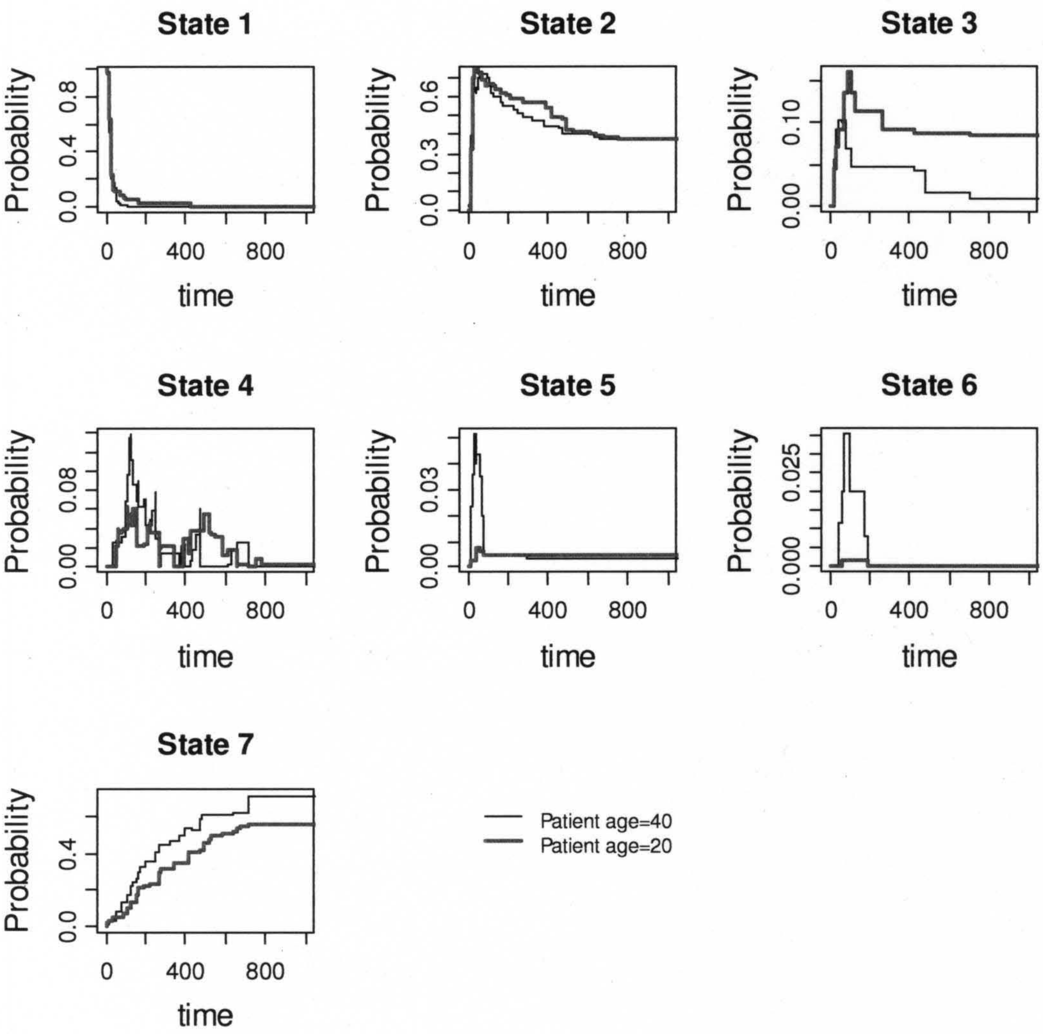


Figure 3.12. The non-parametric regression estimates of state occupation probabilities for the transplant data given the covariate, patient age, set at 20 and 40, respectively.

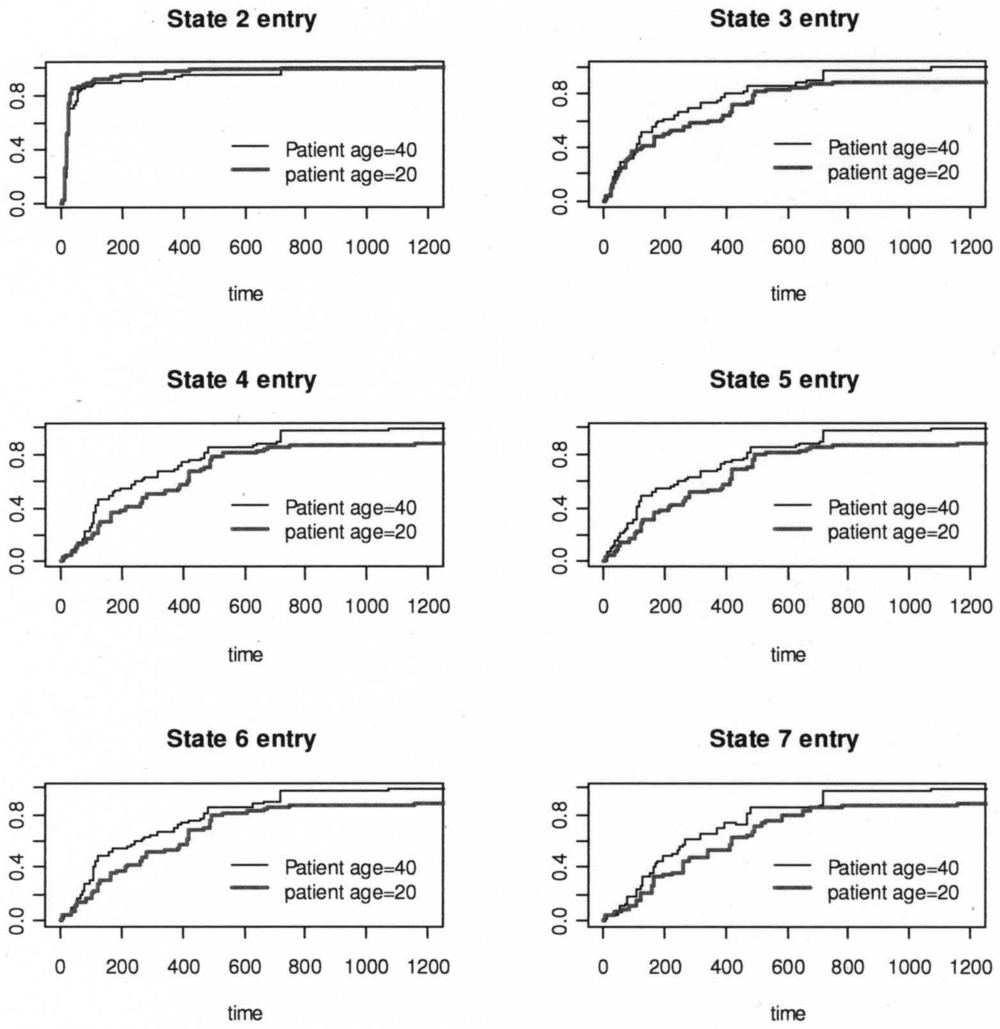


Figure 3.13. The non-parametric regression estimates of entry time distributions for transplant data described in Section 3.2.4 given the covariate, patient age, set at set at 20 and 40, respectively.

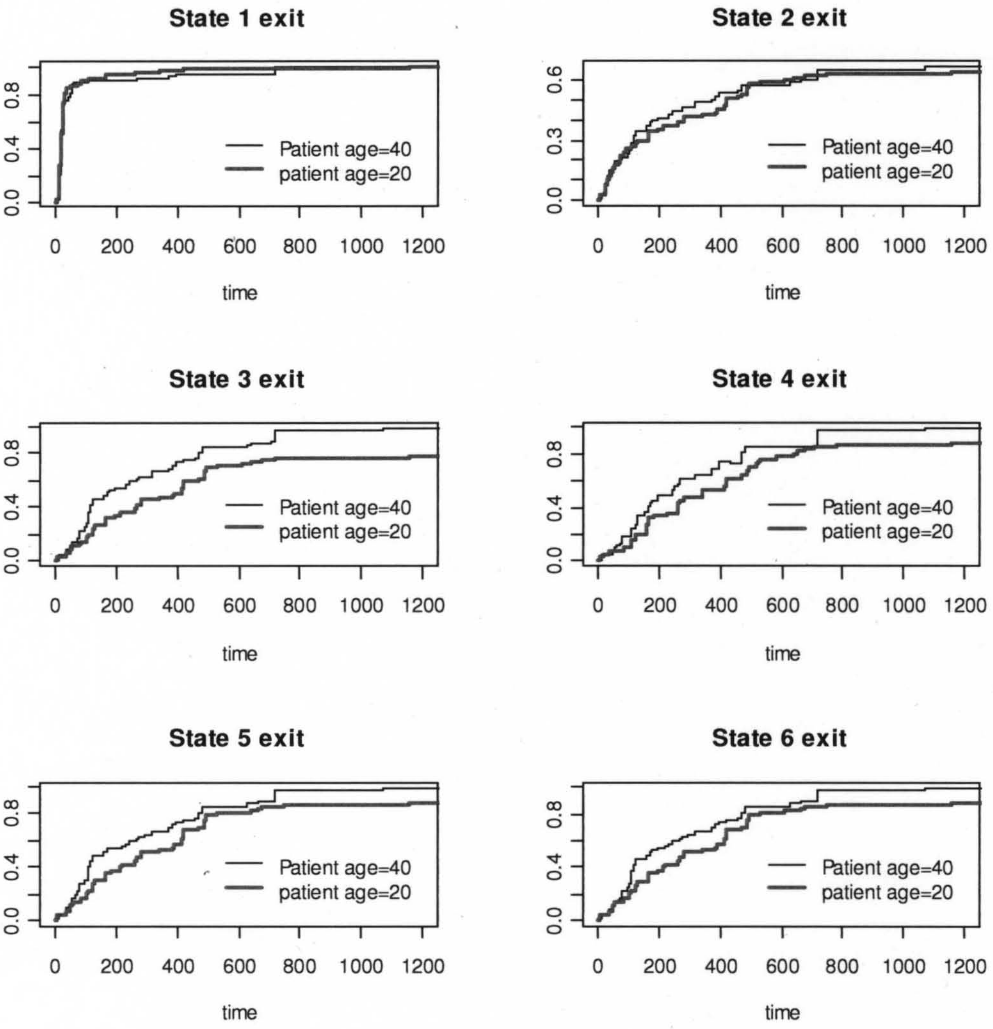


Figure 3.14. The non-parametric regression estimates of exit time distributions for transplant data described in Section 3.2.4 given the covariate, patient age, set at set at 20 and 40, respectively.

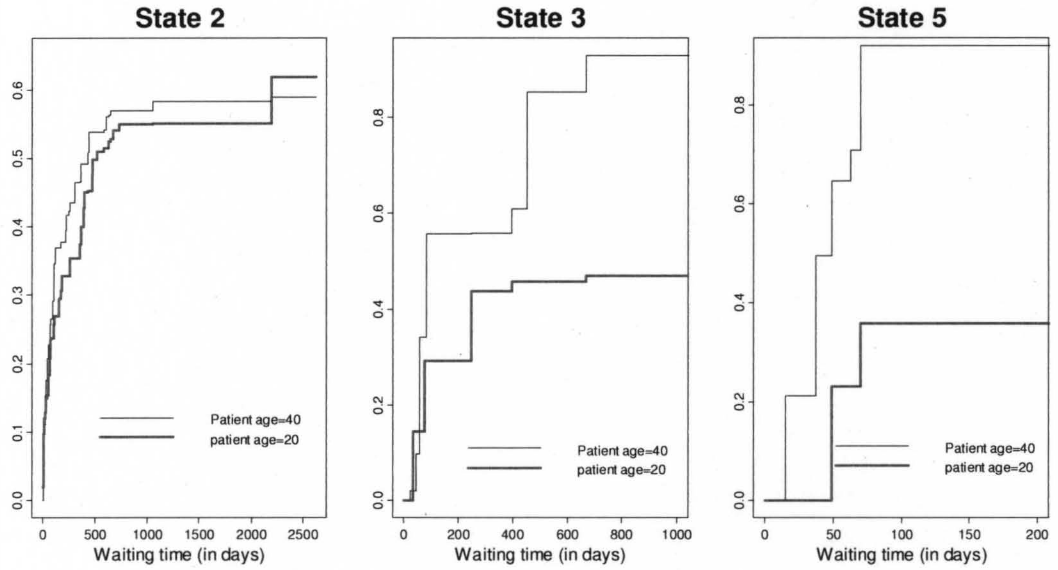


Figure 3.15. The non-parametric regression estimates of waiting time distributions for transplant data described in Section 3.2.5 given the covariate, patient age, set at set at 20 and 40, respectively.

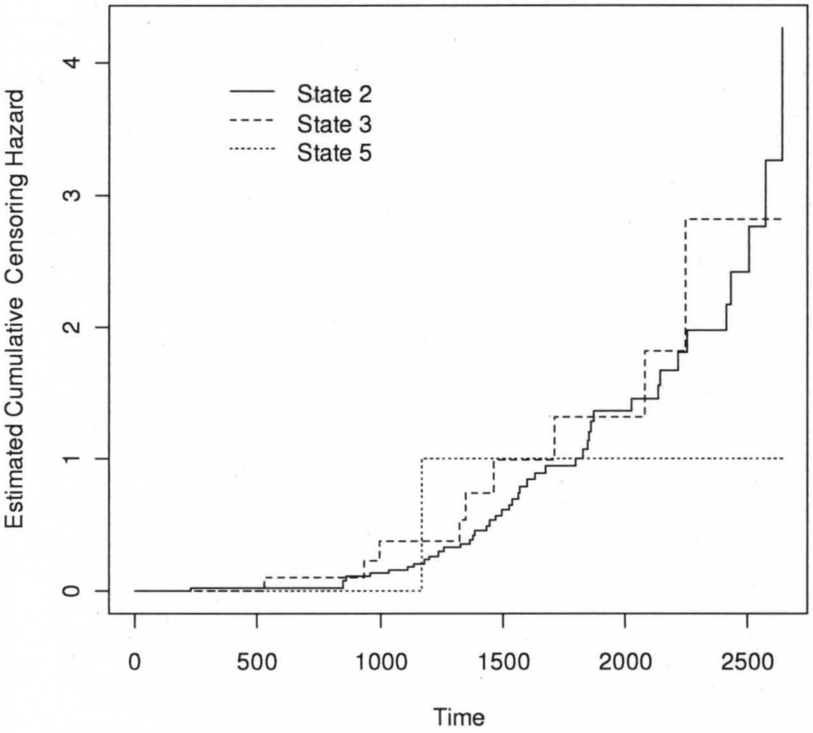


Figure 3.16. Estimated cumulative hazard for being censored in states 2, 3 and 5 for transplant data

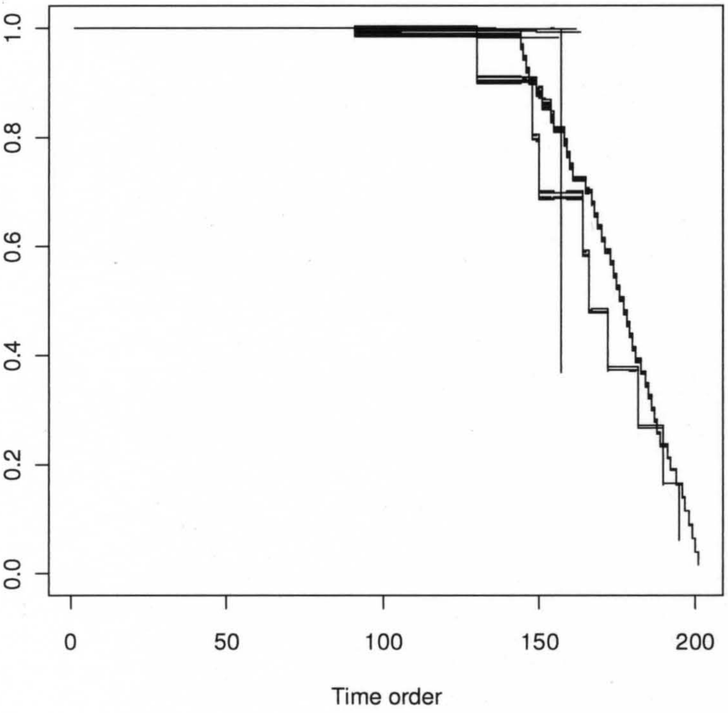


Figure 3.17. Values for $\widehat{K}_i$ versus as integer index corresponding to the rank order of the event (transition or censoring) time for the transplant data

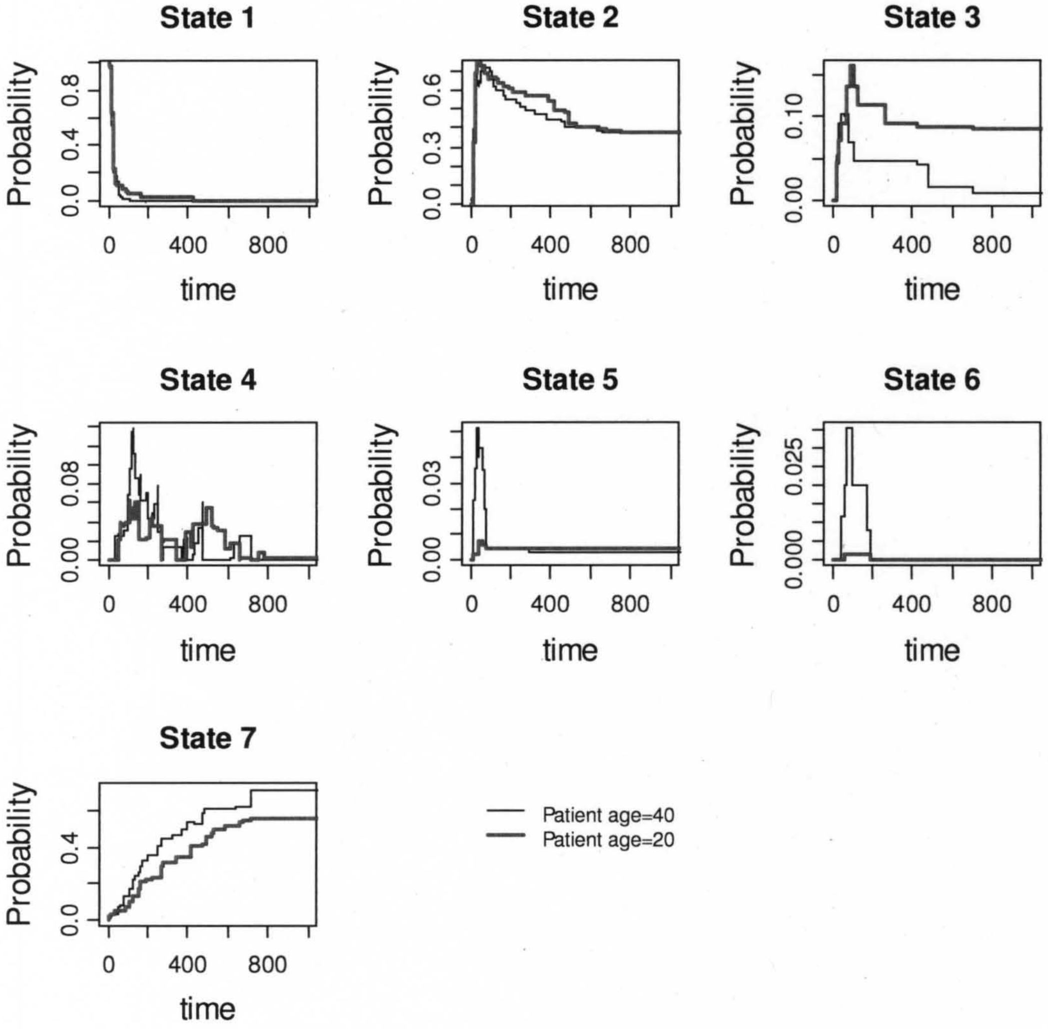


Figure 3.18. The non-parametric regression estimates of state occupation probabilities for the transplant data given the covariate, patient age, set at 20 and 40, respectively, using state dependent censoring model.

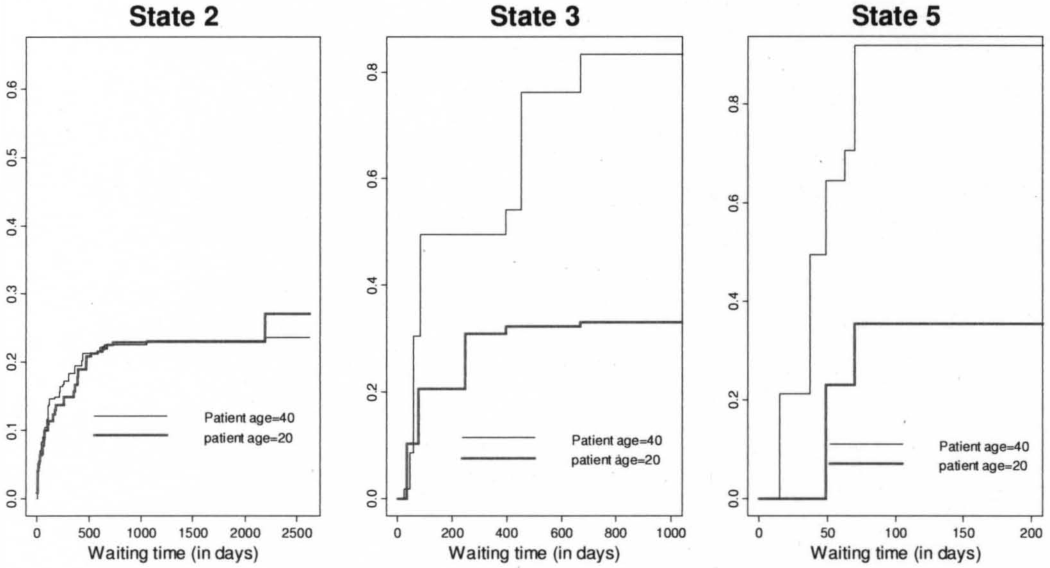


Figure 3.19. The non-parametric regression estimates of waiting time distributions for the transplant data given the covariate, patient age, set at 20 and 40, respectively, using state dependent censoring model.

3.5. Discussion

Past studies of multistate models under censored data have taken mostly parametric and semi-parametric approaches for estimating conditional state occupation probabilities. In this study, we considered a fairly broad class of multistate models which have a progressive structure. We developed valid non-parametric regression methods for the state occupation probabilities, the entry/exit and waiting time distributions given a univariate continuous covariate without additional structural assumption (e.g. Markov or semi-Markov).

Overall, our estimates are based on local versions of the techniques developed by Satten and Datta (2002) for calculating marginal state occupation probabilities and state waiting time distributions, respectively, for multistate models when data are subject to dependent censoring. Various estimated counting and size at risk processes are computed given a value x of X. This is achieved by the kernel smoothing technique of weighing the contributions of various individuals based on the closeness of their covariate values to x. In order to adjust for dependent censoring, a model for censoring mechanism must be specified; we have used Aalen's linear hazard model for this purpose. All observed counting and "number at risk" processes are given an inverse probability of censoring weighted form.

Datta and Satten (2002) established consistency of their estimators under the general paradigm of non-Markov models. Local version of the estimators can be proved the same with shrinking the bandwidth to zero at the right rate. Based on the simulation studies, we may conclude that our non-parametric estimators are consistent for reasonably sized samples and they perform well.

Three different simulation examples are considered. All studies are based on the time continuous five state model described in Figure 3.1 under dependent censoring. The first example is a conditionally semi-Markov model while the second one is conditionally Markov model. We let a univariate external covariate affects both the transition times as well as censoring hazard. In the third example, we let the censoring hazard at time t differs among the subjects according to their current state occupation which is an internal covariate process. Simulation results show satisfactory performance of our estimators of state occupation probabilities, entry and exit time as well as waiting time distributions. Formally, our approach requires that the censoring mechanism be correctly specified; in practice from the first two simulations, we have found that use of the Aalen's model allows successful estimation of state occupation probabilities, entry/exit and waiting time distributions and the results are robust to censoring misidentifications. In Section 3.4, we applied our method on real data. At first, we assumed only an external covariate affects the censoring processes. Then we extend the analysis by adding the state dependent censoring assumption to the model.

Overall, the methodologies presented here are implementable and produce reasonable answers. Unlike parametric and semi-parametric approaches, these non-parametric procedures don't rely on specific model structures and are more robust. Parametric estimators can be compared with the non-parametric counterparts to check for model violations. In other words, non-parametric estimators can be served as important benchmark under large enough sample size.

It is possible to extend and study the effect of multiple covariates. We can use forms of additive model in order to avoid the curse of dimensionality in estimation of the various counting and size at risk processes described before. This should be even more attractive to medical researchers due to greater applicability to a wider class of data sets. The details of this approach will be explored in next chapter.

CHAPTER IV

NON-PARAMTERIC REGRESSION OF STATE OCCUPATION, ENTRY AND EXIT TIMES WITH MULTISTATE DATA USING ADDITIVE MODELS

4.1. Introduction

As a type of multivariate survival data, multistate models have a wide range of applications in medical research. These models allow transition from the former state to the latter, say, from one state of disease to another. The resulting data contains information about the transition times and the states occupied. Example of the simplest multistate model is the traditional survival analysis with one transient state, 0: alive, and one absorbing state, 1: dead.

State occupation probability, which is the probability that a subject be in a specific state at certain time point, is an important quantity in study of multistate models. Another important quantity is the state transition intensity (or transitional hazard) which is the hazard (rate) that an individual moves from one state to another. These time-dependent quantities can be related to the distribution functions of state entry and exit times, each of which may be of independent interest in time to event study. Estimators of these quantities have been proposed in the recent past under a variety of parametric and non-parametric assumptions as well as structural assumptions on the system (such as, progressive, Markov, semi-Markov etc). Aalen and Johansen (1978) gave a method for calculating transition hazard and state occupation probabilities for Markov models starting from the Nelson–Aalen estimators of integrated transition hazards when data are subject to independent censoring. Datta and Satten (2001) showed that these estimators remain consistent even when the underlying model is non-Markovian. They also extended their work and presented estimators using data that are subject to dependent right censoring (Datta & Satten, 2002). Their estimator of the integrated transition hazard matrix has a Nelson-Aalen form, where each of the counting processes, counting the number of transitions between states and the risk sets of leaving the states have an IPCW (Inverse Probability of Censoring Weighted) form. Beside estimation in a marginal model, there is considerable appeal in developing methods for conditional distribution (e.g., regression) of these quantities relating them to number of covariates.

Most of the existing regression methods in multistate context are based on parametric and semi-parametric modeling of the transition hazards (e.g., Satten *et al.*, 1998; Fine and Gray, 1999; Berhane and Weissfeld, 2003 etc.). Generally speaking, while such methods produce relatively precise inference for the effects of covariates under the correct model, their performance under incorrect model assumptions is questionable. This is one compelling reason why a fully non-parametric approach is preferable even though such a formulation is often difficult with time to event data. The situation with multistate models is even more challenging and as such only a limited number of regression approaches exist to analyze such models.

In previous chapter, we developed methods for non-parametric regressions of the state occupation probabilities, the entry/exit and waiting time distributions in a multistate model based on a univariate continuous baseline covariate. No structural assumptions on the multistate system (e.g., a Markov or a semi-Markov model) were being made. The data were subject to right censoring and the censoring mechanism were explainable by observable covariates. We proposed using kernel estimators similar to the method proposed by Beran (1981) along with IPCW estimation techniques in Datta and Satten (2001) leading to our final estimators. This way we introduced local versions of counting and "number at risk" processes through kernel smoothing to give higher weights to individuals with X covariate values close to the given value x.

In this chapter, we are interested to extend the methodology discussed in previous chapter and study effect of multiple covariates on state occupation probabilities and entry/exit time distributions. This allows us to model multiple independent covariates simultaneously. Smoothing approach discussed in Chapter III is no longer valid for multivariate case. In order to avoid the so-called curse of dimensionality, we use forms of additive models to estimate the various counting and size at risk processes. To fit these models, we utilize the backfitting algorithm proposed by Hasti and Tibshirani (1990). Details of the procedure is given in Section 4.2. Note that, similar to the previous chapter, no structural assumptions on the multistate system (e.g., a Markov or a semi-Markov model) are being made. We only concider complete data where no censoring is involved.

The rest of the chapter is organized as follows. The next section introduces our proposed non-parametric regression estimation of state occupation probabilities, integrated hazard, and entry/exit time distributions based on p covariates. Simulation studies are described in Section 4.3. Section 4.4 contains some concluding remarks.

4.2. The Non-parametric Regression Estimators

4.2.1. Notation and Convention

Let $\boldsymbol{S} = \{S(t) : t \geq 0\}$ be a time continuous, multistate process, where t denotes time and $S(t)$ denotes the state occupied at time t. Consider a finite state space $\mathcal{X} = \{0, 1, ..., M\}$ for the process. Under the marginal model, it is assumed that the multistate processes for n individuals $\boldsymbol{S}_i = \{S_i(t) : t \geq 0\}$, $1 \leq i \leq n$, are independent and identical (i.i.d.). We also assume that the model is progressive (e.g., acyclic) and thus a given state j is entered at most once by an individual. Define the state entry and exit times by $U^j = \inf\{t : S(t) = j\}$ and $V^j = \sup\{t : t > U^j, S(t) = j\}$. $U^j = \infty$, if state j

is never entered, and $V^j = \infty$, if either state j is never entered or j is an absorbing state (in which case it is never left). Also, let $p_j(t) = Pr\{S(t) = j\}$ be the probability that a typical individual will be at state j at time t. In this chapter, we study the effects of a p-vector of time independent covariates $\boldsymbol{X} = (X_1, ..., X_p)'$ on various functions of time $\theta(t|x_1, ..., x_p)$ related to the multistate system. Thus, $F^j(t|x_1, ..., x_p)$ will denote the regression function $Pr\{U^j \leq t| \, U^j < \infty, X = (x_1, ..., x_p)\}$ and so on. As before, we assume that $\{\boldsymbol{S}_i, x_i\}$ are i.i.d., for $1 \leq i \leq n$.

4.2.2. Integrated Transition Hazard

Let $\alpha_{jj'}(t|x_1, ..., x_p)$ be the local (i.e., given $\boldsymbol{X} = (x_1, ..., x_p)$) transition hazard of the original (uncensored) chain from state j to j' (for states $j \neq j'$), defined as

$$\alpha_{jj'}(t|x_1, ..., x_p) = \lim_{dt \to 0} \frac{Pr\{S(s) = j' \text{ for some } s \in [t, t + dt)|S(t -) = j, \boldsymbol{X} = (x_1, ..., x_p)\}}{dt}$$

where $S(t -)$ denotes the state occupied just before time t. Let $N_{jj'}$ be the counting process with jumps given by $\varDelta N_{jj'}(t) = \sum\limits_{i=1}^{n} I(S_i(t -) = j, S_i(t) = j')$. Define $\Delta N_{i,jj'} = I(S_i(t -) = j, S_i(t) = j')$, $i = 1, ..., n$ to be the indicator function whether the subject i made transition from state j to state j' at time t. Also, let $Y_j(t) = \sum\limits_{i=1}^{n} I(S_i(t -) = j)$ be the number of individuals at state j just before time t and note that by Fubini's theorem

$$\alpha_{jj'}(t|x_1, ..., x_p) = \lim_{dt \downarrow 0} \frac{E\{dN_{jj'}(t)|S_i(t -) = j, \boldsymbol{X} = (x_1, ..., x_p)\}}{dt}$$

so that $\alpha_{jj'}(t|x_1, ..., x_p)$ counts the limiting number of j to j' transitions made by an individual with covariate $\boldsymbol{X} = (x_1, ..., x_p)$, in the interval [*t*,*t*+*dt*) and hence denotes a local version of a 'partially conditional transition rate' of Pepe and Cai (1993). The local cumulative (integrated) transition hazard matrix is given by

$\boldsymbol{A}(\cdot|x_1,...,x_p) = \{A_{jj'}(\cdot|x_1,...,x_p)\}$, with $A_{jj'}(t|x_1,...,x_p) = \int_0^t \alpha_{jj'}(s|x_1,...,x_p)ds$ for $j \neq j'$ with $\alpha_{jj}(t|x_1,...,x_p) = -\sum_{j \neq j'} \alpha_{jj'}(t|x_1,...,x_p)$.

To develop our estimators, we follow the method proposed by Datta and Satten (2002) in estimating the transition hazard. In their treatment, the estimator of the integrated transition hazard matrix has a Nelson-Aalen form. They estimate the two processes $N_{jj'}$ and Y_j separately using the principle of inverse probability of censoring weights rather than ratio.

To estimate $N_{jj'}(t|x_1,...,x_p)$, the local counting process, we employ an additive model. Define $\Delta N_{i,jj'} = I(S_i(t-) = j, S_i(t) = j')$, for $i = 1,...,n$ and $j \neq j'$. Put $f(x,t) = E[n^{-1}\Delta N_{jj'}(t) \,|\, \boldsymbol{X} = (x_1,...,x_p)]$ for the regression function of $\Delta N_{jj'}(t) = \sum_{i=1}^n \Delta N_{i,jj'}(t)$ on $\boldsymbol{X}$. Hence, the model becomes

$$n^{-1}\Delta N_{jj'}(t, x_1,...,x_p) = f(x_1,...,x_p,t) + \epsilon_{t,n},$$

where the error term satisfy $E[\epsilon_{t,n}|\boldsymbol{X}] = 0$.

Here we consider a flexible approach to estimate the regression function $f(x,t)$ through a model under which the effect of each covariate on the response is represented in an additive way. We assume the additive model

$$f(x,t) = f_1(x_1,t) + ... + f_p(x_p,t),$$

where $f_1(\cdot,t),...,f_p(\cdot,t)$ are one-dimensional functions, for each t. As a result, the model becomes

$$E(n^{-1}\Delta N_{jj'}(t|x_1,...,x_p)) = f_1(x_1,t) + ... + f_p(x_p,t).$$

To fit the additive model discussed above a back fitting algorithm can be applied. The backfitting algorithm is an iterative procedure to fit additive models in which each component is estimated by keeping the other components fixed at each step. It actually cycles through the covariates X_j $(j = 1,...,p)$, and estimates each f_j by applying local smoothers to the partial residuals. These residuals are obtained by removing the estimated effects of the linear covariates. The algorithm iterates until convergence (Hasti and Tibshirani, 1990). The steps involved to carry out the procedure are presented below.

$i)$ Set $m = 0$ and $f_k^{(0)} = 0$, or any reasonable estimate.

$ii)$ Set $m = m + 1$. For $k = 1, ..., p$ set

$$f_k^{(m)}(t, x_k) = S\;(n^{-1}\Delta N_{jj'}(t) - \sum_{s\neq k} f_s^{(m-1)}(x_s, t)\Big| x_k),$$

where S is a smoothing operator.

$iii)$ Repeat step (ii) until the changes in the f_k between iterations are sufficiently small.

Special care is required to ensure that all estimated values of $\Delta N_{jj'}(t, x_1, ..., x_p)$ are positive. One way to handle this is to project negative values to zero. Local kernel smoothers can be applied in backfitting algorithm steps.

The "number at risk", $\widehat{Y}_j(t|x_1, ..., x_p)$, can be calculated from estimated local counting process. This is possible by keeping track of individuals who are at specific state at time t. Thus, the "number at risk " for a transient state j at time t_1 can be simply obtained by

$$\widehat{Y}_j(t_1|x_1, ..., x_p) = \sum_{t\leq t_1}\Big\{\Delta N_{\bullet j}(t|x_1, ..., x_p) - \Delta N_{j\bullet}(t|x_1, ..., x_p)\Big\},$$

where $\Delta N_{\bullet j}(t|x_1, ..., x_p) = \sum_{k\neq j} \Delta N_{kj}(t|x_1, ..., x_p)$ and $\Delta N_{j\bullet}(t|x_1, ..., x_p) =$ $\sum_{k\neq j} \Delta N_{jk}(t|x_1, ..., x_p)$. Here we assume that all individuals start at state 0 at time 0.

Finally, the non-parametric estimator of integrated transition hazard is given by

$$\widehat{A}_{jj'}(t|x_1,...,x_p) = \begin{cases} \int_0^t J_j(u|x_1,...,x_p)\widehat{Y}_j(u|x_1,...,x_p)^{-1}d\widehat{N}_{jj'}(u|x_1,...,x_p) & j \neq j' \\ -\sum_{j\neq j'}\widehat{A}_{jj'}(t|x_1,...,x_p) & j = j', \end{cases}$$

with $J_j(u|x_1,...,x_p) = I(\widehat{Y}_j(u|x_1,...,x_p) > 0)$,

4.2.3. State Occupation Probabilities

The Datta-Satten (2002) estimator $\widehat{P}(s,t)$ of the transition probability is given by the product integral

$$\widehat{P}(s,t) = \prod_{(s,t]}(I + d\widehat{\boldsymbol{A}}(u)),$$

where I is the identity matrix. This is a finite product taken over all distinct observed transition times. Finally, the Datta-Satten estimator $\widehat{p}_j(t)$ of the occupation probability of state j at time t is given by

$$\widehat{p}_j(t) = \sum_{k=0}^{M}\frac{\widehat{Y}_k(0+)}{n}\,\widehat{p}_{kj}(0,t),$$

where M is the number of states and n is the sample size. In fact, the marginal state occupation probability of state j is obtained by evaluating the product limit of transition hazards and averaging it with respect to the initial distribution of state occupation.

The non-parametric regression estimation of these quantities can be handled by introducing a local version and incorporating covariates. The non-parametric regression estimator of the state occupation probabilities conditional on given value of $\boldsymbol{X}$, $p_j(t|x_1,...,x_p) = Pr\{S(t) = j|\ \mathbf{X} = (x_1,...,x_p)\}$ is given by

$$\widehat{p}_j(t|x_1,...,x_p) = \sum_{k=0}^{M}\frac{\widehat{Y}_k(0+|x_1,...,x_p)}{n}\,\widehat{p}_{kj}(0,t|\,x_1,...,x_p),$$

where $\widehat{P}(0,t|x_1,...,x_p)$ is the kjth element of the matrix

$\widehat{P}(0,t|x_1,...,x_p) = \prod_{(0,t]}(I + d\widehat{\boldsymbol{A}}(u|x_1,...,x_p))$ obtained by product integration of the matrix $\widehat{\boldsymbol{A}}(t)$ defined in previous section.

4.2.4. State Entry and Exit Distributions

Recall that for any state $j > 0$, U^j is the entry time for state j amongst individuals who ever enter state j. Let F^j denote the corresponding distribution function conditional on $\boldsymbol{X} = (x_1,...,x_p)$,

$$F^j(t|x_1,...,x_p) = P\{U^j \leq t|U^j < \infty, \boldsymbol{X} = (x_1,...,x_p)\},$$

where we take $F^0(t|x_1,...,x_p) = 1$, for all $t \geq 0$. Due to the progressive structure of the multistate system under consideration, any state (node) will be reached from the root node 0 by a unique path. Let S^j be the collection of all states $j' \neq j$ such that state j appears on the path connecting state 0 and j'. In other words, S^j is the collection of all states which proceeds state j in the progressive model. Then estimators of entry time distributions of state j is given by

$$\widehat{F}^j(t|x_1,...,x_p) = \frac{\sum\limits_{k\in\{j\}\cup S^j} \widehat{P}_k(t|x_1,...,x_p)}{\sum\limits_{k\in\{j\}\cup S^j} \widehat{P}_k(\infty|x_1,...,x_p)},$$

where $\widehat{P}_k(\infty|x_1,...,x_p) = \lim\limits_{t\to\infty} \widehat{P}_k(t|x_1,...,x_p)$.

Analogous to above, V^j is the exit time for state j of individuals who will ever enter state j. Let G^j denote the corresponding distribution function conditional on $X = (x_1,...,x_p)$,

$$G^j(t|x_1,...,x_p) = Pr\{V^j \leq t|\mathrm{U}^j < \infty, X = (x_1,...,x_p)\}$$

where we take $G^j(t|x_1,...,x_p) = 0$ if j is a terminal node in the directed tree structure for all $t \geq 0$. For a transient state j, $\widehat{G}^j(t|x_1,...,x_p)$ is taken to be the normalized sum of estimated state occupation probabilities of all states thatever entered state j in the progressive system. Mathematically speaking, state exit time distribution is given by

$$\widehat{G}^j(t|x) = \frac{\sum\limits_{k\in S^j} \widehat{P}_k(t|x_1,...,x_p)}{\sum\limits_{k\in\{j\}\cup S^j} \widehat{P}_k(\infty|x_1,...,x_p)}.$$

4.3. Simulation Studies

We performed number of Monte Carlo experiments to evaluate the finite sample performance of the proposed methods. These simulations are based on complete data and no censoring is involved. We have based our simulations on a hypothetical five-state progressive model (Figure 4.1).

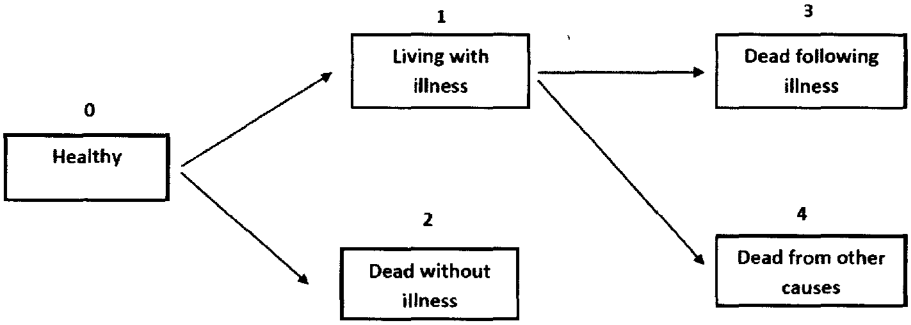


Figure 4.1. A five-state illness-death model

4.3.1. Non-Markov Transition Times

For each person, two covariates were generated. X_1 was generated from a normal distribution with mean parameter 5 and standard deviation parameter of 0.5. X_2 was generated independently from another normal distribution with mean parameter 4 and standard deviation parameter of 0.6. We assumed all individuals start in State 0 (well) at time zero, and may either progress to State 1 or State 2. Each patient at State 0 had a 60% chance (controlled by a Bernoulli variable that is independent of the event times) of following the $0 \rightarrow 1$ arm and a 40% chance of following the $0 \rightarrow 2$ arm. Furthermore, patients who entered State 1, would subsequently reach States 3 or 4 with arms probabilities (controlled by another independent Bernoulli variable) 0.6 and 0.4, respectively. To generate the event times in a non-Markov model, we used both lognormal and Weibull distributions for the models. For lognormal simulations, the waiting times in State 0 were generated from lognormal distribution with log-mean parameter 0 and log-scale parameter 0.5 and for individuals traversing the $0 \rightarrow 1$ arm, the State 1 waiting times were generated using another independent lognormal distribution with log-mean parameter 0 and log-scale parameter 1. The waiting times were multiplied by the person-specific frailty variable obtained from multiplication of two independent variables F_1 and F_2. F_1 generated from normal distribution with mean parameter $X_1/5$ and standard deviation parameter 0.1 and F_2 generated from normal distribution with mean parameter $X_2/4$ and standard deviation parameter 0.1. For the Weibull simulations, the State 0 waiting times were generated from a Weibull distribution with shape parameter 2 and scale 1. For patients traversing the path $0 \rightarrow 1$, the state waiting times were obtained from another independent Weibull distribution with shape parameter 0.5 and scale 1.

The state occupation probabilities of described examples were estimated by the proportion of subjects observed in each state at time t. The smoothing step in the backfitting algorithm was based on normal kernels and data-based bandwidth selector for each covariate. The *R* package "KernSmooth" was used to this end (http://cran.rproject.org/doc/packages/KernSmooth.pdf). The bandwidth was taken as dpik, the data-based bandwidth selector of Wand and Jones (1995).

The empirical non-parametric estimators for a sample of size 100, 500, 1000 and 5000 generated as above with given $x_1 =$ median of the covariate X_1 and $x_2 =$ median of the covariate X_2. The estimators with sample size of 5000 are used as benchmarks.

Figure 4.2 and 4.3 display the state occupation probability results of non-Markov data with Weibull and lognormal transition times respectively. Overall, as the sample size increases, the estimators get closer to the benchmark values, suggesting the appropriate large sample properties of the estimators.

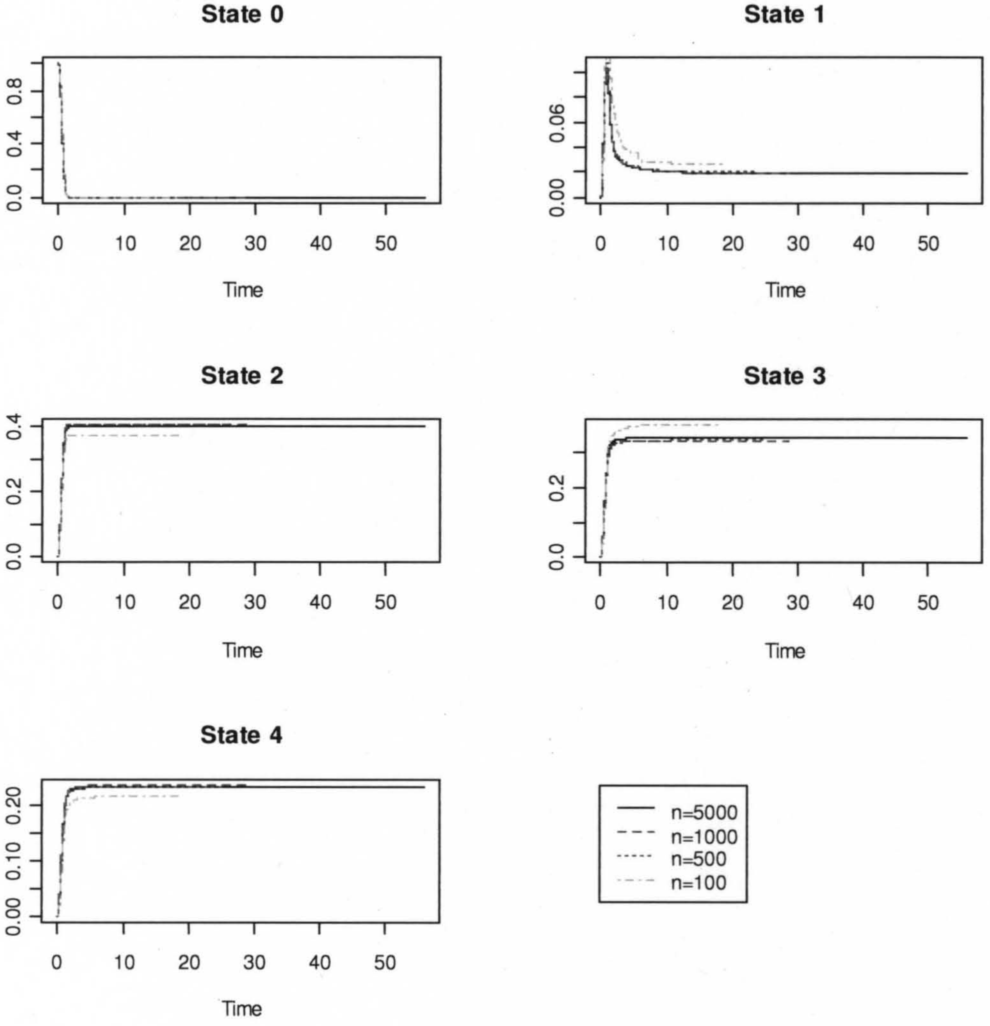


Figure 4.2. The non-parametric regression estimates of state occupation probabilities in a five-state Weibull non-Markov model given x_1 = median of the covariate X_1(generated from $N(5, 0.5)$) and x_2 = median of the covariate X_2 (generated from $N(4, 0.6)$).

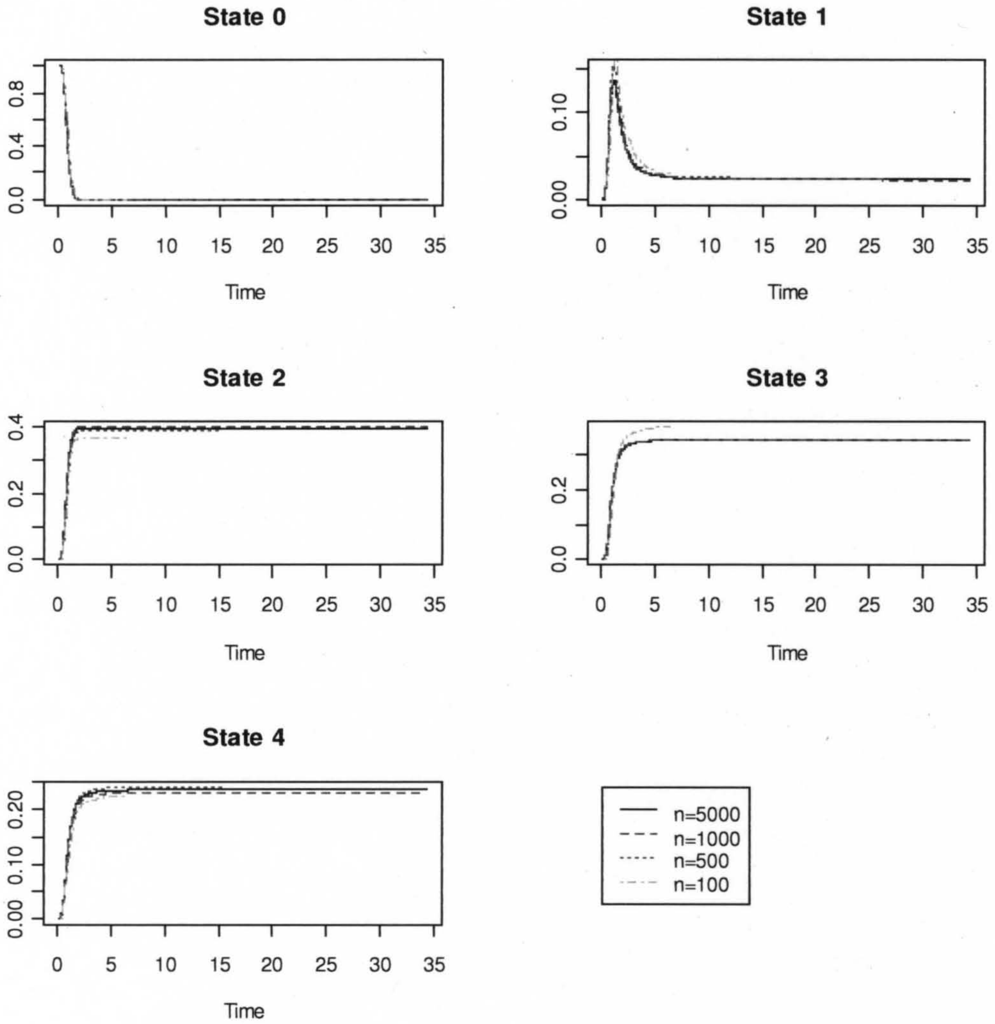
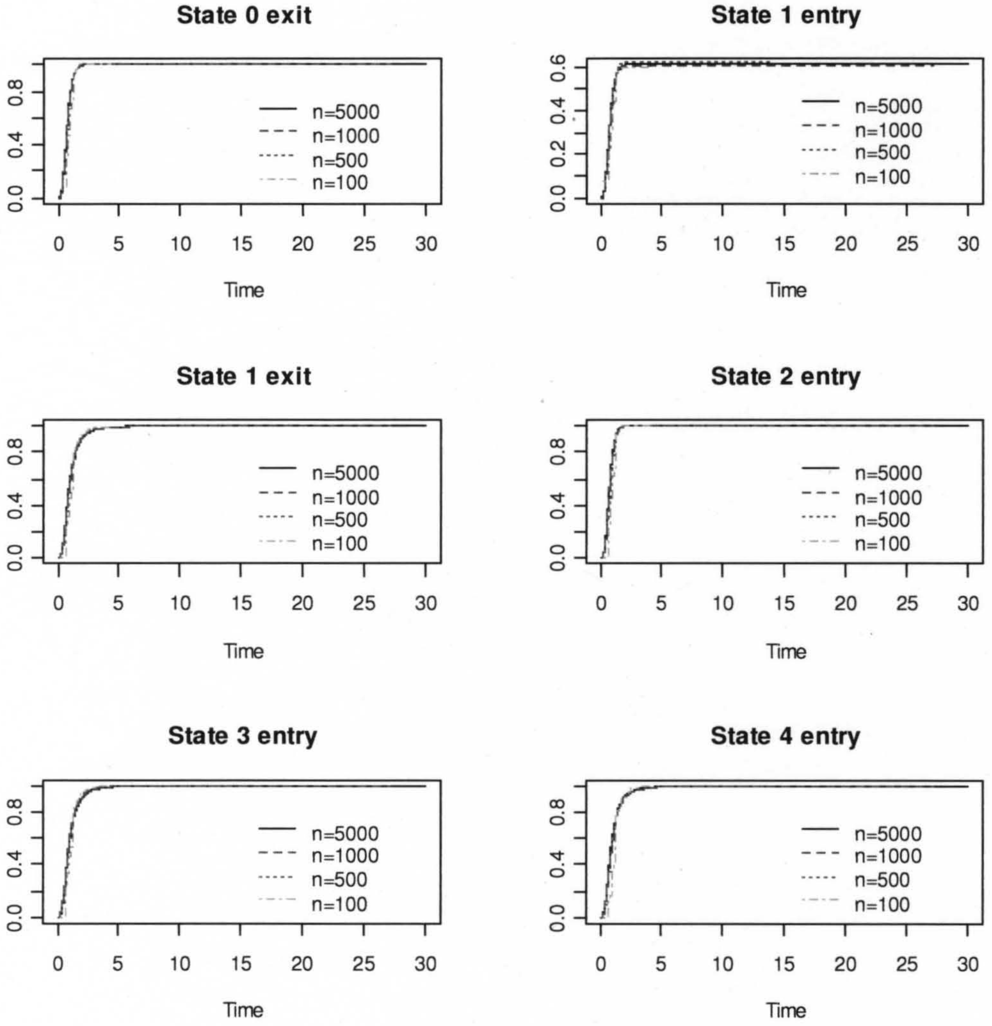


Figure 4.3. The non-parametric regression estimates of state occupation probabilities in a five-state lognormal non-Markov model given x_1 = median of the covariate X_1(generated from $N(5, 0.5)$) and x_2 = median of the covariate X_2 (generated from $N(4, 0.6)$).


State 0 exit

State 1 entry

State 1 exit

State 2 entry

State 3 entry

State 4 entry

Figure 4.4. The non-parametric regression estimates of entry/exit times in a five-state Weibull non-Markov model given $x_1 =$ median of the covariate X_1(generated from $N(5, 0.5)$) and $x_2 =$ median of the covariate X_2 (generated from $N(4, 0.6)$).

We assessed the global performance of the estimators by the L_1 distance

$$\Delta : = E\int|\widehat{\theta}(t|x) - \widehat{\theta}_{\mathrm{B}}(t|x)|dF_n(t|x),$$

where $\widehat{\theta}(x)$ and $\widehat{\theta}_B(x)$ denote respectively our proposed estimator of θ and its benchmark counterpart based on sample size of 5000. Here, θ is either a state occupation probability, state entry time distribution function or a state exit time distribution function. The integrating measure in the definition of the L_1 distance was taken to be the distribution function of event times $F_n(t|x) = n^{-1}\sum_{i=1}^{n} I\{T_i \leq t|x\}$; $\Delta = 0$ means that they are in a complete agreement on the support of the event times. We calculated Δ via Monte Carlo averaging with the replication size of 1000. The calculations were performed based on $x_1 =$ median of the covariate X_1 (generated from $N(5, 0.5)$) and $x_2 =$ median of the covariate X_2 (generated from $N(4, 0.6)$).

The L_1 results of state occupation probabilities for non-Markov with Weibull and lognormal transition times are provided in Tables 4.1 and 4.2, respectively. Tables 4.3 and 4.4 list the results for entry and exit time distributions of Weibull and lognormal transition times. For all simulation settings, the L_1 values decrease with increasing sample size. This is a good indicator that the estimators are consistent and converge to benchmark values for reasonably sized samples.

Table 4.1. The L_1 distances between non-parametric regression estimators of state occupation probabilities based on data with sample size 5000 (benchmark) and sample sizes 100, 500 and 1000 in a five-state non-Markov Weibull model. The estimates are based on a Monte Carlo sample size of 1000; all standard errors were less than 0.01

	$n = 100$	$n = 500$	$n = 1000$
p_0	0.015	0.010	0.006
p_1	0.008	0.005	0.003
p_2	0.021	0.010	0.009
p_3	0.015	0.011	0.007
p_4	0.014	0.011	0.007

Table 4.2. The L_1 distances between non-parametric regression estimators of state occupation probabilities based on data with sample size 5000 (benchmark) and sample sizes 100, 500 and 1000 in a five-state non-Markov lognormal model. The estimates are based on a Monte Carlo sample size of 1000; all standard errors were less than 0.09

	$n = 100$	$n = 500$	$n = 1000$
p_0	0.029	0.019	0.017
p_1	0.014	0.007	0.005
p_2	0.040	0.018	0.013
p_3	0.026	0.013	0.010
p_4	0.025	0.012	0.009

Table 4.3. The L_1 distances between non-parametric regression estimators of entry/exit time distributions based on data with sample size 5000 (benchmark) and sample sizes 100, 500 and 1000 in a five-state non-Markov Weibull model. The estimates are based on a Monte Carlo sample size of 1000; all standard errors were less than 0.01

	$n = 100$	$n = 500$	$n = 1000$
Ex0	0.015	0.008	0.006
En1	0.022	0.011	0.009
Ex1	0.018	0.011	0.008
En2	0.016	0.010	0.007
En3	0.019	0.010	0.008
En4	0.019	0.013	0.008

Table 4.4. The L_1 distances between non-parametric regression estimators of entry/exit time distributions based on data with sample size 5000 (benchmark) and sample sizes 100, 500 and 1000 in a five-state non-Markov lognormal model. The estimates are based on a Monte Carlo sample size of 1000; all standard errors were less than 0.02

	$n = 100$	$n = 500$	$n = 1000$
Ex0	0.029	0.019	0.017
En1	0.045	0.021	0.016
Ex1	0.030	0.017	0.014
En2	0.031	0.019	0.015
En3	0.032	0.018	0.016
En4	0.032	0.018	0.016

The following scatter plots show an approximate linear relationship of the logarithms of the L_1 distances with the logarithms of the sample size for each of these estimators suggesting that the L_1 values converge to zero at the rate of n^{-b}, for some b. The log mean L_1 distance plots are listed in Figures 4.5-4.8.

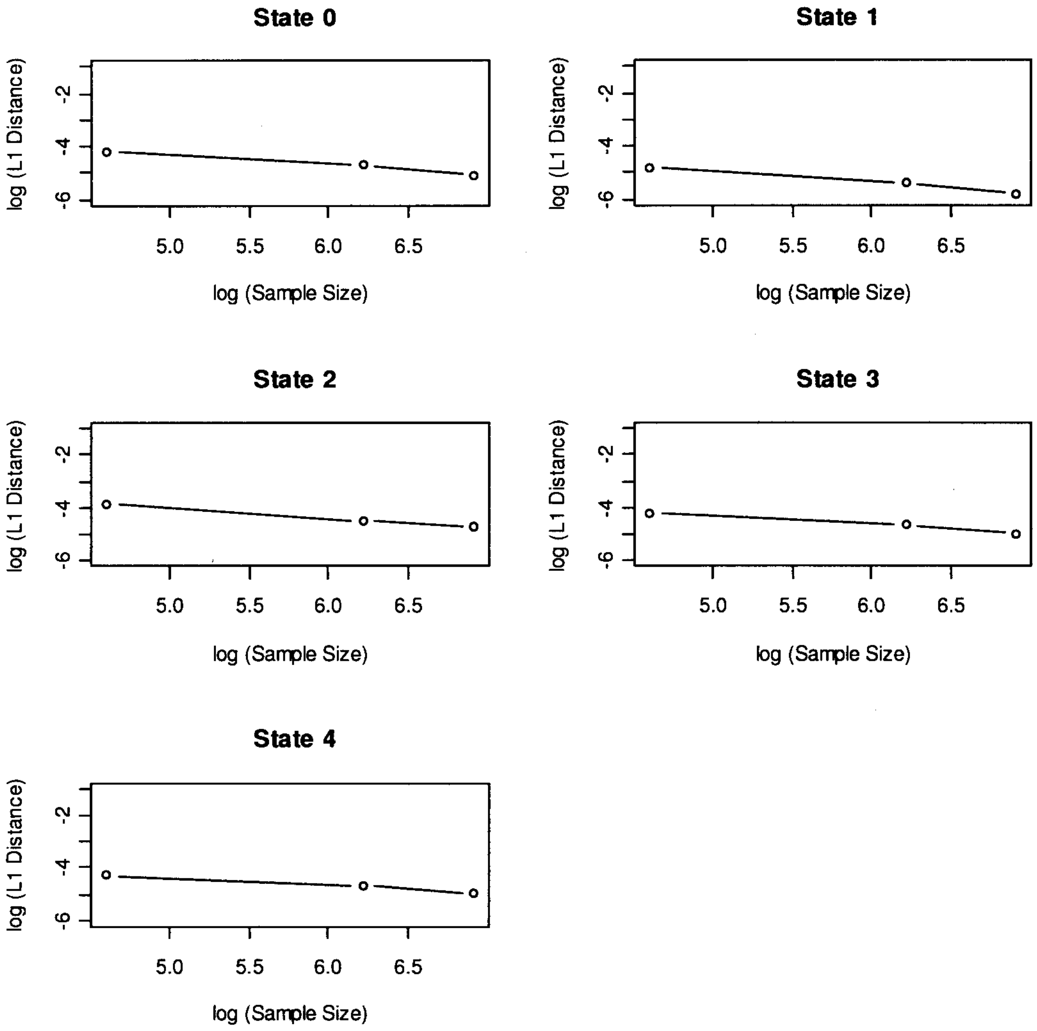


Figure 4.5. The log mean L_1 distances of state occupation probabilities in a five-state Weibull non-Markov model given x_1 = median of the covariate X_1(generated from $N(5, 0.5)$) and x_2 = median of the covariate X_2 (generated from $N(4, 0.6)$).

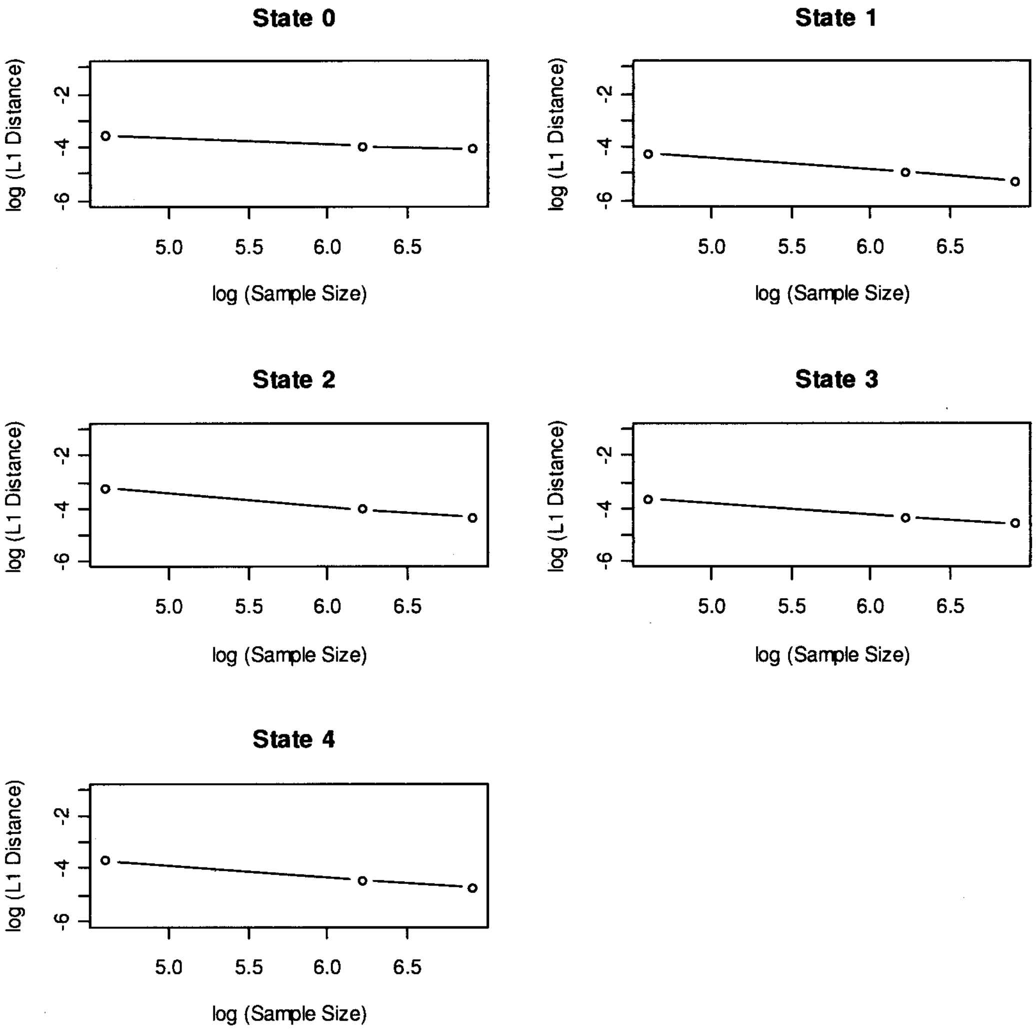


Figure 4.6. The log mean L_1 distances of state occupation probabilities in a five-state lognormal non-Markov model given x_1 = median of the covariate X_1 (generated from $N(5, 0.5)$) and x_2 = median of the covariate X_2 (generated from $N(4, 0.6)$).

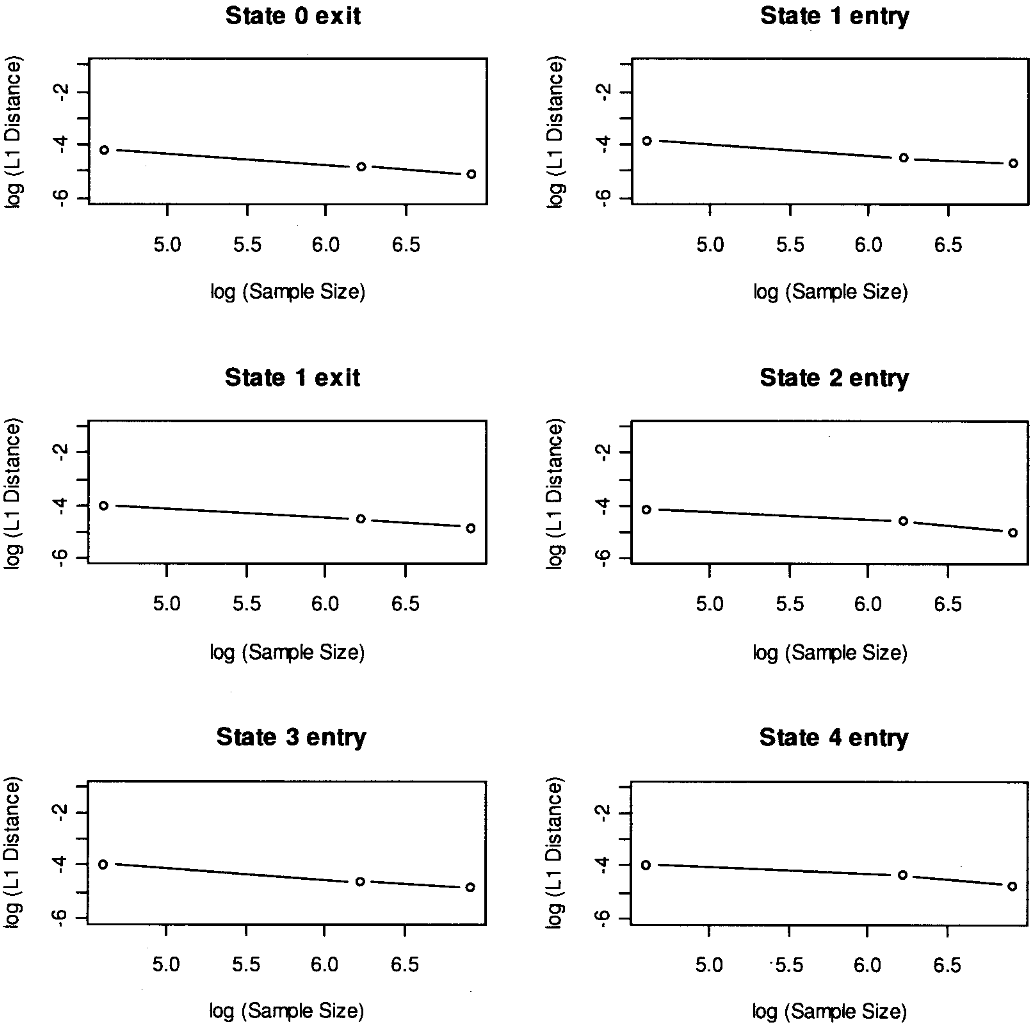


Figure 4.7. The log mean L_1 distances of Entry/Exit time distributions in a five-state Weibull non-Markov model given x_1 = median of the covariate X_1(generated from $N(5, 0.5)$) and x_2 = median of the covariate X_2 (generated from $N(4, 0.6)$).

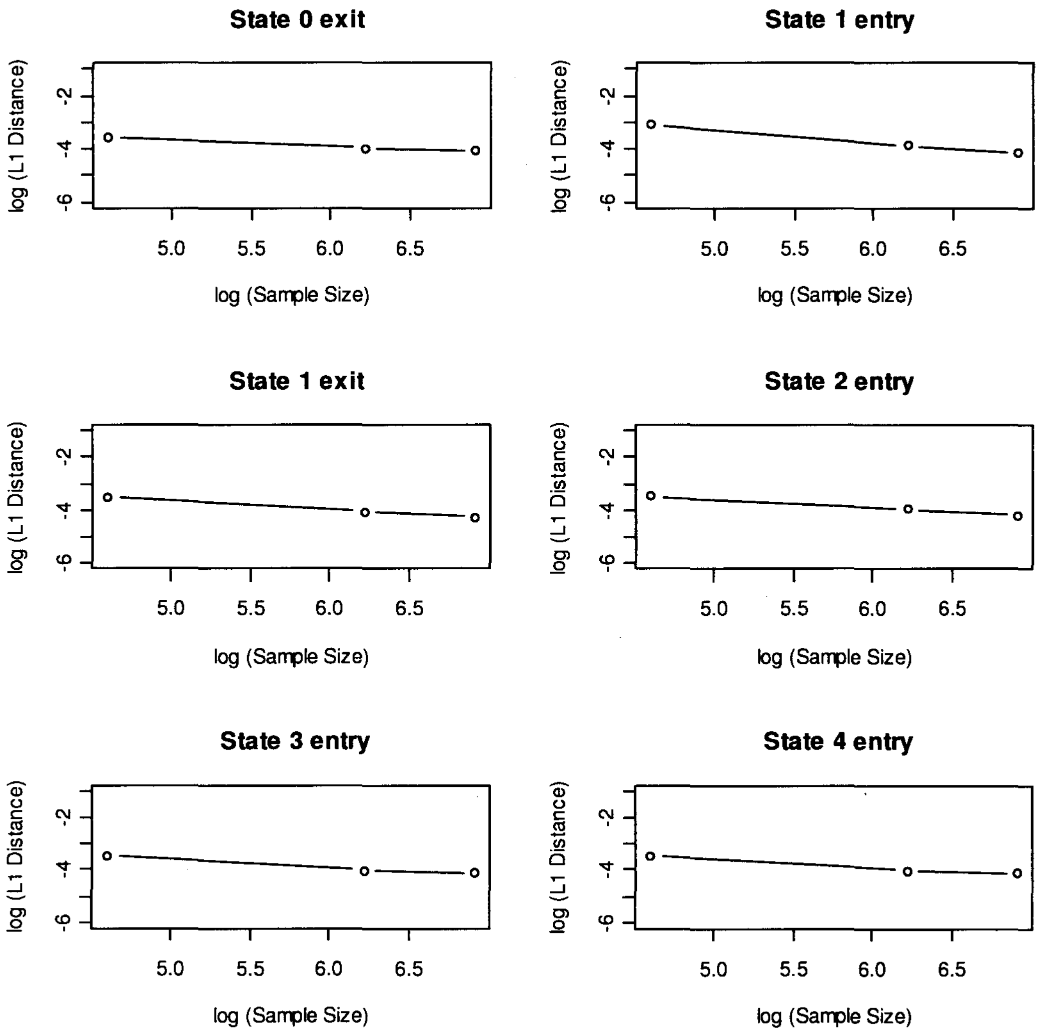


Figure 4.8. The log mean L_1 distances of Entry/Exit time distributions in a five-state lognormal non-Markov model given x_1 = median of the covariate X_1(generated from $N(5, 0.5)$) and x_2 = median of the covariate X_2 (generated from $N(4, 0.6)$).

4.4. Discussion

Past studies of multistate models under censored data have taken mostly parametric and semi-parametric approaches for estimating conditional state occupation probabilities. In this study, we considered a fairly broad class of multistate models which have a progressive structure. We developed fully non-parametric regression methods for the state occupation probabilities and the entry/exit time distributions without additional structural assumption (e.g. Markov or semi-Markov). In previous chapter we developed such estimators using univariate covariate. These estimators were constructed based on local version estimators of Datta and Satten (2002) give a method for calculating marginal state occupation probabilities and transition hazards for multistate models when data are subject to dependent censoring. In this chapter, we extended the previous study to be able to include two or more than two covariates in the regression model. This study is more attractive to medical researchers due to greater applicability to a wider class of data sets. In order to avoid the curse of dimensionality, we used additive models through application of backfitting algorithm in estimation of the various counting and size at risk processes described before. Kernel smoothers were employed in backfitting processes.

Two different simulation examples are considered. Both studies are based on the time continuous five state non-Markov model described in Figure 4.1. We also performed Monte Carlo simulations to calculate L_1 distances between the empirical estimation values of benchmark ($n = 5000$) and smaller sample sizes ($n = 100$, $n = 500$ and $n = 1000$). The L_1 values were calculated for state occupation probabilities and entry/exit time distributions. Based on the empirical results, we may conclude that these estimators produce reasonable answers and our non-parametric estimators are consistent for reasonably sized samples.

The results presented in this chapter are based only on empirical values with no censoring. In Future, we need to incorporate censoring as well and study the performance of models when true transition times are partially observed. Also, we need to apply the proposed methods on real data and demonstrate their practical application.

CHAPTER V

EXTENSIONS AND FUTURE RESEARCH

5.1. Non-parametric Regression Based on Current Status Data

In chapter III, we focused on right censored data and developed non-parametric regression estimators of state occupation, as well as state entry, exit and waiting times distributions in a multistate model. Both numerical simulations and data applications proved that the methods proposed are reasonable and implementable.

Our future research is to generalize these proposed procedures to more general case with current status multistate data obtained from a continuous random inspection time for each individual. Marginal non-parametric estimation for multistate current status data was undertaken in Datta and Sundaram (2006), Datta *et al.* (2009) and Lan and Datta (2010); the special case of competing risk models was investigated by Jewell *et al.* (2003) and Groenboom *et al.* (2008). Non-parametric regression for multistate current status data is absent in literature.

As before, for an individual i and a time $t \geq 0$, $S_i(t)$ denotes the state the individual i is in, at time t; C_i denotes the random time at which the individual i gets inspected. The censoring times and the state occupation process $\{C_i, S_i(t), t \geq 0\}$ for the individuals are assumed to be independent and identically distributed. For simplicity of development, we will make the assumption of random censoring, which means C_i is independent of $\{S_i(t) : t \geq 0\}$, given covariate X. We further assume that all transition and censoring times are continuous and that the allowable transitions give rise to a directed tree structure, in which every state j can be reached from an initial state 0 (the root node) by a unique path.

5.1.1. State Occupation Probabilities

Let $U_{jj'}$ denote the (unobserved) transition time of an individual from state j to j'. Let $N^*_{jj'}$ denote the counting process, counting the number of j to j' transitions in $[0, t]$ with the complete data. By the laws of large numbers,

$$n^{-1}N^*_{jj'}(t) \xrightarrow{p} n^{-1}EN^*_{jj'}(t) = P\{U_{jj'} \leq t\} = n_{jj'}(t), \text{say.}$$

Consider the indicator function $I(U_{jj'} \leq C)$ of the event that the j to j' transition has taken place by time C. Then for any $t \geq 0$,

$$E(I(U_{jj'} \leq C)|C = t) = Pr\{U_{jj'} \leq t\}.$$

In order to compute the regression functions given X, we need to compute weighted versions of this estimated process where the weight corresponding to the ith observation is $\phi_h(x - X_i)$, where ϕ_h is a scaled kernel as in Chapter 3.

Finally, the class of state occupation probabilities will be computed as before; however, the integrated conditional transition hazards are now calculated using new counting process and "at risk" set. The mathematical and computational details of the procedure need to be worked out.

5.1.2. State Entry and Exit Distributions

Once we have the state occupation probabilities given the covariate X, the state entry and exit time distributions given X can be obtained using normalized sums as before.

5.1.3. State Waiting Time Distribution

Calculation of state waiting time distribution with current status data poses additional difficulty since we can not directly regress the indicators of events involving the waiting times since the state entry times are also unknown. To solve this problem, we make additional structural assumptions (see, e.g., Datta *et al.*, 2009) of a conditional Markov or a semi-Markov model given X.

Under the Markov assumption, we could obtain the following identity

$$S_{j,waiting}(t|x) = \int_0^\infty \prod_{u<s\leq u+t} (1 + d\Lambda_{j\bullet}(s|x) dF_{j,entry}(u|x),\ t \geq 0,$$

where $\Lambda_{j\bullet}$ is integrate transition hazard out of state j, conditional on $X = x$. Using this and the quantities defined earlier we obtain a non-parametric regression of the state waiting time survival function

$$\widehat{S}_{j,waiting}(t|x) = \int_0^\infty \left\{ \prod_{u<s\leq u+t} \left(1 - \frac{d\widehat{N}_{j\bullet}(s|x)}{\widehat{Y}_j(s|x)}\right) \right\} d\widehat{F}_{j,entry}(u|x),\ t \geq 0.$$

5.2. Regression Analysis of High Dimensional Data with Time Dependent Covariates

In Chapter II, we discussed and compared performance of four different latent factor and regularized/penalized methods to handle predicting survival in high dimensional setting. We assumed that covariates are fixed and not changing with time. However, nature of survival regression lends itself easily to extensions that allows for covariates that change over time.

Let $Z_i(t)$ denote the value of the covariate for subject i at time t. The use of the time-varying covariate model typically assumes that $Z_i(t)$ is available for all possible times. However, in practice we almost never observe $Z_i(t)$ continuously in time. Rather, we commonly measure the covariate process at discrete times t_{i1}, t_{i2}, ..., t_{in_i}. For example, in Netherlands Non-small Cell Lung Cancer data discussed in Section 2.2.6, serum spectra of the patients were available at three time points: pre-treatment (preTx), after two cycles of treatment (post-2) and at the end of treatment (EOT). It is common that these spectra change over time.

One approach is to use penalized estimating equations on covariates at each time point to estimate the parameters and then take an appropriate average to get the final estimates. One such example is to take the weighted sum of the regression effects over all possible observation times. To handle the censoring observations, we can apply the multiple imputation algorithm proposed in Section 2.2.3.

5.3. Non-parametric Regression Estimation of Multistate Models in High Dimensional Setting

In Chapter IV, we considered developing non-parametric regression estimators of state occupation probabilities and entry/exit time distributions in multistate system based on multivariate continuous baseline covariate. We plan to examine the proposed method by incorporating censoring into the model. We would also like to conduct more number of simulations to cover variety of possible circumstances. Application to real data example is also of particular interest.

An additional direction for future work is to combine the methods described in Chapter IV and II to apply fully non-parametric regression on high dimensional data. This is particularly important to handle multistate data in bioinformatic studies dealing with high dimensional data. Additive models have good statistical and computational behavior only when the number of variables, p, is not large relative to the sample size n. Hence, their usefulness is limited in the high dimensional setting.

In order to solve this problem, one possible approach is to start with a dimension reduction method such as Partial Least Squares (PLS) method. More precisely, in the first step by using PLS method, we find the optimal linear transition from the large number of original descriptors to a small number of latent variables. Next, the resulting latent variables are plugged into the additive regression model as additive components.

This method is applied on Netherlands Non-small Cell Lung Cancer data discussed in Section 2.2.6 with 995 features ($X(1)$) to get some preliminary result. We utilized PLS method and took the first component as a univariate covariate. Then, we calculated the conditional state occupation probabilities in a two-state survival model. We compared the state occupation probabilities given the median and the third quartile of the corresponding covariate. The results are displayed in Figure 5.1.

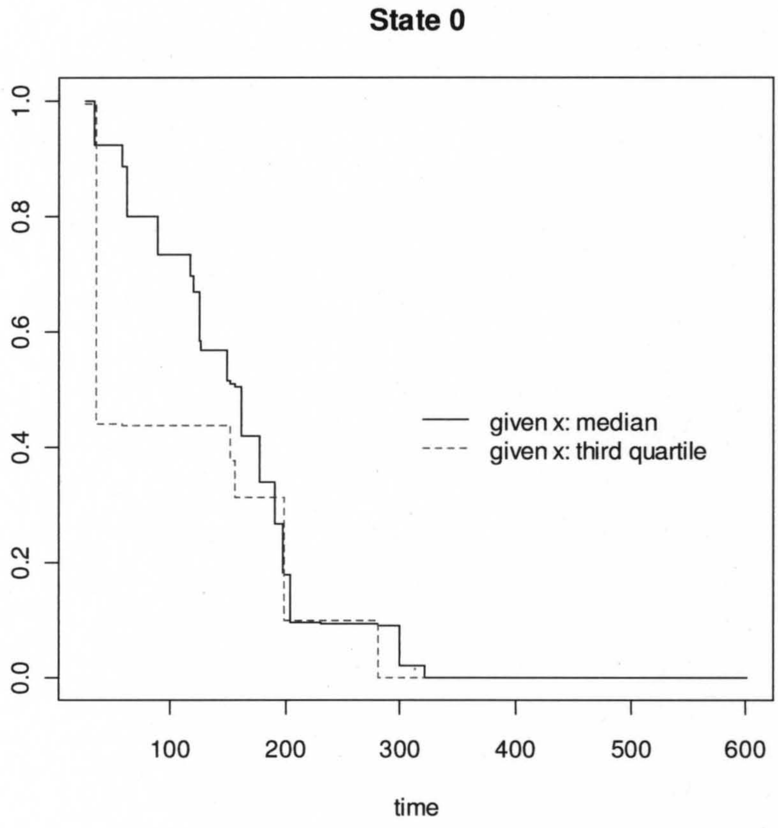


Figure 5.1. Conditional state occupation probability, given median and third quartile of the first component extracted from PLS applied on Netherlands NSCLC data

We will work on the details of this method using multiple covariates in future.

REFERENCES

Aalen, O.O. (1976). Non-parametric inference in connection with multiple decrement models. *Scand J Stat*, 3, 15-27.

Aalen, O.O. (1978). Nonparametric inference for a family of counting processes. *Annual Statistics*, 6, 701-726.

Aalen, O.O & Johansen, S. (1978). An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scand J Stat*, 5, 141-50.

Aalen, O.O. (1980). A model for nonparametric regression analysis of counting process. *Springer Lect. Notes Statist*, 2, 1-25.

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, 8, 907–925.

Aalen, O.O., Borgan, Ø. & Fekjær, H. (2001). Covariate adjustment of event histories estimated from Markov chains: The additive approach. *Biometrics*, 57, 993-1001.

Adam, B. L. Qu,Y. Davis, J.W. Ward, M. D. Clements, M.A. Cazares, L.H. Semmes, O. J. Schellhammer, P.F. Yasui, Y. Feng, Z. & Wright, G. L. (2002). Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.*, 62, 3609-14.

Aebersold, R. & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature,* 422, 198-207.

Andersen, P.K. Borgan, Ø. Gill, R.D. & Keiding, N. (1993). Statistical Models Based on Counting Processes. *New York: Springer-Verlag.*

Andersen, P.K. & Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11, 91-115.

Andersen, P.K. & Klein, J.P. (2007). Regression analysis for multistate models based on a pseudo-value approach, with applications to bone marrow transplantation studies. *Scand J Stat*, 34, 3-16

Atlas, M. & Datta, S. (2009). A statistical technique for monoisotopic peak detection in a mass spectrum. *J. Proteom. Bioinform*, 2, 202-216.

Beck, G.J. (1979). Stochastic survival models with competing risks and covariates. *Biometries*, 35, 427-438.

Beran, R. (1981). Nonparametirc regression with randomly censored survival data. *Technical Report, Univ. California, Berkeley.*

Berhane, K. & Weissfeld, L.A. (2003). Inference in Spline-Based Models for Multiple Time-to-Event Data, with Applications to a Breast Cancer Prevention Trial. *Biometrics*, 59, 859-868.

Boulesteix, A. L. & Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief. Bioinform*, 8, 32-44.

Bovelstad, H. M. Nygard, S. Storvold, H. L. Aldrin, M. Borgan, O. Frigessi, A. & Lingjaerde, O. C. (2007). Predicting survival from microarray data - a comparative study. *Bioinformatics*, 23, 2080-2087.

Chun, H. & Keles, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. Roy. Statist. Soc. Ser. B*,72, 3-25.

Coombes, K.R. Koomen, J.M. Baggerly, K.A. Morris, J.S. & Kobayashi, R. (2005). Understanding the characteristics of mass spectrometry data through the use of simulation. *Cancer Inform,*1, 41-52.

Copelan, E.A. Biggs, J.C. Thompson, J.M. Crilley, P. Szer, J. Klein, J.P. Kapoor, N. Avalos, B.R. Cunningham, I. Atkinson, K. Downs, K. Harmon, G.S. Daly, M.B. Brodsky, I. Bulova, S.I. & Tutschka, P.J. (1991). Treatment for Acute Myelocytlic Leukemia with Allogeneic Bone Marrow Transplantation Following Preperation with Bu/Cy. *Blood,* 78, 838-843.

Cox, D.R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B*, 34, 187-220.

Cruz-Marcelo, A. Guerra, R. Vannucci, M. Li, Y. Lau, C.C. & Man, T.K. (2008). Comparison of algorithms for pre-processing of SELDI-TOF mass spectrometry data. *Bioinformatics,* 24, 2129-2136.

Dabrowska, D.M. (1987). Nonparamtric regression with censored survival time data. *Scand J Stat*, 14, 181-197.

Dabrowska, D.M. (1989). Uniform Consistency of the Kernel Conditional Kaplan-Meier Estimate. *The Annals of Statistics*, 17, 1157-1167.

Dabrowska, D.M. (1992*b*). Variable bandwidth conditional Kaplan-Meier estimate. *Scand J Stat*, 19, 351–61.

Datta, S. & Satten, G.A. (2000*a*). Nonparametric estimation for the three-stage irreversible illness-death model. *Biometrics*, 56, 841-7.

Datta, S. Satten, G.A. & Datta, S. (2000*b*). Estimation of stage occupation probabilities in multistate models. In: Balakrishnan, N eds. *Advances on Theoretical and Methodological Aspects of Probability and Statistics.* New York: Gordon and Breach, 493-506.

Datta, S. & Satten, G.A. (2001). Validity of the Aalen-Johansen estimators of stage occupation probabilities and integrated transition hazards for non-Markov models. *Statistics and Probability Letters*, 55, 403-11.

Datta, S. & Satten, G.A. (2002). Estimation of integrated transition hazards and stage occupation probabilities for non-Markov systems under dependent censoring. *Biometrics*, 58, 792-802.

Datta, S. & Sundaram, R. (2006). Nonparametric estimation of state occupation probabilities in a multistate model with current status data. *Biometrics*, 62, 829–37.

Datta, S. Le-Rademacher, J. & Datta, S. (2007). Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. *Biometrics*, 63, 259-71.

Datta, S. Lan, L. & Sundaram, R. (2009). Nonparametric estimation of waiting time distributions in a Markov model based on current status data. *Journal of Statistical Planning and Inference,* 139, 2885-97

Dazard, J.E. Gal, H. Amariglio, N. Rechavi, G. Domany, E. & Givol, D. (2003). Genome-wide comparison of human keratinocyte and squamous cell carcinoma responses to UVB irradiation: implications for skin and epithelial cancer. *Oncogene*, 22, 2993-3006.

Doksum, K.A & Yandell, B.S. (1982). Properties of regression estimates based on censored survival data. *A Festschrift for Erich L. Lehmann, ed. by PJ Bickel, KA Doksum, JL Hodges, Jr, Wadsworth: Belmont, CA*, 140–156.

Efron, B. Hastie, T. Johnstone, I. & Tibshirani, R. (2004 *a*). Least angle regression - Rejoinder. *Ann Stat*, 32, 494-499.

Efron, B. Hastie, T. Johnstone, I. & Tibshirani, R. (2004 *b*). Least angle regression. *Ann. Stat.*, 32, 407-451.

Emanuele, V.A & Gurbaxani B.M. (2009). Benchmarking currently available SELDI-TOF MS preprocessing techniques. *Proteomics*, 9,1754-1762.

Engler, D. & Li, Y. (2009). Survival analysis with high-dimensional covariates: an application in microarray studies. *Stat. Appl. Genet. Mol. Biol.*,8, Article-14.

Fine J.P. & Gray, R.J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* 94,496–509

Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software,* 33, 2010.

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

Gray, R. J. (1994). Spline-based tests in survival analysis. *Biometrics,* 50, 640–652.

Gluderer, S. Oldham, S. Rintelen, F. Sulzer, A. Schutt, C. Wu, X.D. Raftery, L.A. Khoury, C.M. Yang, Z. Li, X.Y. Vignali, M. Fields, S. & Greenwood, M.T. (2008). A TSC22-like motif defines a novel antiapoptotic protein family. *Fems Yeast Res,* 8, 540-563.

Groenboom, P. Maathuis, M. H. & Wellner, J. A. (2008). Current status data with competing risks: limiting distribtuion of the MLE. *Annals of Statistics*, 36, 1064–1089.

Gui, J. & Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, 21, 3001-8.

Hafen, E. & Stocker, H. (2008). Bunched, the Drosophila homolog of the mammalian tumor suppressor TSC-22, promotes cellular growth. *Bmc. Dev. Biol.* , 8, 10.

Hastie, T.J., Tibshirani, L.J. (1990). Generalized Additive Models, Chapman, London.

Huang, J. Ma, S. & Xie, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics*, 62, 813-20.

Jeffries, N. (2005). Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics*, 21, 3066-73.

Jewell, N.P. Van der Laan, M.J. & Henneman, T. (2003). Nonparametric estimation from current status data with competing risks. *Biometrika* 90, 183–197.

Kalbfleisch, J. D. & Prentice, R. L. (1980). The Statistical Analysis of Failure Time Data, *Wiley: New York.*

Kay, R. (1982). The analysis of transition times in multistate stochastic processes using proportional hazard regression models. *Communications in Statistics-Theory and Methods*, 11, 1743-1756.

Klein, J. P. Klotz, J.H. & Grever, M.R. (1984). A biological marker model for predicting disease transitions, *Biometrics*, 40, 927-936.

Klein, J.P & Moeschberger, M.L. (1997). Survival analysis: techniques for censored and truncated data. *New York: Springer-Verlag.*

Lagakos, S.W. (1976). A stochastic model for censored-survival data in the presence of an auxiliary variable. *Biometries*, 52, 551-559.

Lan, L. & Datta, S. (2010). Nonparametric estimation of state occupation, entry and exit times with multistate current status data. *Statistical Methods in Medical Research*, 19, 147-65.

Li, G. & Doss, H. (1995). An Approach to Nonparametric Regression for Life History Data Using Local Linear Fitting. *Annual Statistics,* 23, 787-823.

Li, G. & Datta, S. (2001). A bootstrap approach to nonparametric regression for right censored data. *Annals of Institute of Statistical Mathematics*, 53, 708-729.

Li, H. & Luan, Y. (2003). Kernel Cox regression models for linking gene expression profiles to censored survival data. *Pacific Symp. Biocomput.*, 8, 65-76.

Lin, D.Y. & Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, 81, 61-71

Lin, D.Y., Sun, W. & Ying, Z. (1999). Nonparametric estimation of the gap time distributions for serial events with censored data. *Biometrica* 45, 497-507.

Liotta, L.A., Ferrari, M. & Petricoin, E. (2003). Clinical proteomics: written in blood, *Nature*, 425, 905.

McKeague, I.W. & Utikal, K.J. (1990). Inference for a Nonlinear Counting Process Regression Model. *Annual Statistics*, 18, 1172-1187.

Meira-Machado, L., de Uña-Álvarez, J. & Cadarso-Suárez, C. (2006). Nonparametric estimation of transition probabilities in a non-Markov illnessdeath model, *Lifetime Data Analysis*, 12, 325–344.

Morris, J.S. Coombes, K.R. Koomen, J. Baggerly, K.A. & Kobayashi, R. (2005). Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum, *Bioinformatics*, 21, 1764-1775.

Nadaraya, E. (1964). On estimating regression. *Theory of Probability and Its Applications*, 9, 141-2.

Ndukum, J. N. Atlas, M. & Datta, S. (2011). pkDACLASS: Open source software for analyzing MALDI-TOF data, *Bioinformation*, 6, 45–47.

Pawitan, Y. Bjohle, J. Wedren, S. Humphreys, K. Skoog, L. Huang, F. Amler, L. Shaw, P. Hall, P. & Bergh, J. (2004). Gene expression profiling for prognosis using Cox regression. *Stat. Med.*, 23, 1767-1780.

Pepe, M.S & Cai, J. (1993). Some graphical displays and marginal regression analyises for recurrent failure times and time dependent covariates. *J Am Statist Assoc*, 88, 811-820.

Petricoin, E.F Ardekani, A.M Hitt, B.A Levine, P.J Fusaro, V.A Steinberg, S.M Mills G.B, Simone, C. Fishman, D.A. Kohn, E.C. & Liotta, L.A. (2002). Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359, 572-7.

Plevritis, S.K. Salzman, P. Sigal B.M. & Glynn P.W. (2007). A natural history model of stage progression applied to breast cancer. *Statistics in Medicine*, 26, 581-595.

Petricoin, E. F., 3rd Ornstein, D.K. Paweletz, C.P. Ardekani, A. Hackett, P. S. Hitt, B.A. Velassco, A. Trucco, C. Wiegand, L. Wood, K. Simone, C.B. Levine, P.J. Linehan, W.M. Emmert-Buck, M.R. Steinberg, S.M. Kohn, E.C. & Liotta, L.A. (2002). Serum proteomic patterns for detection of prostate cancer. *J. Natl. Cancer. Inst.*, 94, 1576-8.

Rai, A.J. & Chan, D.W. (2004). Cancer proteomics: Serum diagnostics for tumor marker discovery. *Ann. N. Y. Acad. Sci.*, 1022, 286-94.

Renard, B.Y. Kirchner, M. Steen, H. Steen, J.A. & Hamprecht, F.A. (2008). NITPICK: peak identification for mass spectrometry data. *BMC Bioinformatics*, 9, 355.

Sacks, S.T. & Chiang, C. L. (1977). A transition probability model for the study of chronic diseases. *Mathematical Biosciences,* 60, 197-206.

Satten, G., Datta, S. & Williamson, J.M. (1998). Inference based on imputed failure times for the proportional hazards model with interval censored data. *J. Amer. Statist. Assoc.* 93, 318–327.

Satten, G.A. & Datta, S. (2002). Marginal estimation for multi-stage models: waiting time distributions and competing risks analyses. *Stat Med*, 21, 3-19.

Satten, G.A. Datta, S. Moura, H. Woolfitt, A.R. Carvalho, M.D. Carlone, G.M. De, B.K. Pavlopoulos, A. & Barr, J.R. (2004). Standardization and denoising algorithms for mass spectra to classify whole-organism bacterial specimens. *Bioinformatics*, 20, 3128-3136.

Scheike, T.H. & Zhang, M.J. (2006). Direct Modelling of Regression Effects for Transition Probabilities in Multistate Models. *Scand J Stat*, 34,17-32.

Shu, Y. & Klein, J.P. (2005). Additive hazards Markov regression models illustrated with bone marrow transplant data. *Biometrika*, 92, 283–301.

Sorace, J.M. & Zhan, M. (2003). A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics*, 4, 24.

Stoeckli, M. Chaurand, P. Hallahan, D.E. & Caprioli, R.M. (2001). Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian t*issues. Nat. Med.* 7, 493-6.

Taguchi, F. Solomon, B. Gregorc, V. Roder, H. Gray, R. Kasahara, K. Nishio, M. Brahmer, J. Spreafico, A. Ludovini, V. Massion, P.P. Dziadziuszko, R. Schiller, J. Grigorieva, J. Tsypin, M. Hunsucker, S.W. Caprioli, R. Duncan, M.W. Hirsch, F.R. Bunn, P.A. & Carbone, D.P. (2007). Mass spectrometry to classify non-small-cell lung cancer patients for clinical outcome after treatment with epidermal growth factor receptor tyrosine kinase inhibitors: a multicohort cross-institutional study. *J. Natl. Cancer Inst.*, 99, 838-46.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58, 267-288.

Tibshirani, R. (1997). The LASSO method for variable selection in the Cox model. *Statistics in Medicine,* 16, 385-295.

Tibshirani, R. J. (2009). Univariate shrinkage in the Cox model for high dimensional data. *Stat. Appl. Genet. Mol. Biol.*, 8, 21.

van Houwelingen, H.C. Bruinsma, T. Hart, A.A.M. van't Veet, L.J. & Wessels, L.F.A. (2006). Cross-validated Cox regression on microarray gene expression data. *Stat. Med.*, 25, 3201-3216.

Voortman, J. Pham, T.V. Knol, J.C. Giaccone, G. & Jimenez, C.R. (2009). Prediction of outcome of non-small cell lung cancer patients treated with chemotherapy and bortezomib by time-course MALDI-TOF-MS serum peptide profiling. *Proteome Sci.* 7, 34.

Wand, M.P & Jones, M.C. (1995). Kernel Smoothing. *London: Chapman & Hall.*

Watson, G. (1964). Smooth regression analysis. *Sankhya*, Series A, 26, 302–305.

Wei, G. C. & Tanner, M.A. (1991). Applications of multiple imputation to the analysis of censored regression data. *Biometrics*, 47, 1297-309.

Wei, L.J. (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statist. Med.*, 11, 1871–1879.

Wold, H. (1985). Partial least squares. *In Encyclopedia of Statistical Sciences, S. Kotz and N. L., Ed. Wiley: New York*, 6, 581-591.

Wolski, W.E. Lalowski, M. Jungblut, P. & Reinert, K. (2005). Calibration of mass spectrometric peptide mass fingerprint data without specific external or internal calibrants. *BMC Bioinformatics*, 6, 203.

Wu, S.C. (1982). A semi-Markov model for survival data with covariates. *Mathematical Biosciences*, 60, 197-206.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B*, 67, 301-320.

APPENDIX

Key R Programs for the Estimation of the Conditional Quantities in a Five-state Illness Death Model based on right censored data

```
# Kernel Function

 library(KernSmooth)
 kf<-function(z,givenz,h)  {
 n<-length(z)
 kern<-(1/(n*h))*dnorm((givenz-z)/h)
 return(kern)
                                  }


# Estimation of k function  using Aalen's linear model

k.f<-function(data,U)       {

UUT<-NULL
for (j in seq(length(U)))                 {
UUT[[j]]<-lapply(seq(dim(U[[1]])[1]), function(i)
{as.matrix(U[[j]][i,])%*%as.matrix(t(U[[j]][i,]))})
                                                                                    }
ID<-unique(data$id)
R<-lapply(seq(length(U)), function(i) {NULL})

for (j in seq(length(U)))        {
R[[j]][[1]]<-UUT[[j]][[1]]
for (i in (2:length(ID))) {R[[j]][[i]]<-UUT[[j]][[i]]+R[[j]][[i-1]]}
                                    }

 R<-lapply(seq(length(U)), function(i) {R[[i]][[length(ID)]]})

id.cen<-data[data$cen.ind==0,]$id
```

```
RTJUTJ<-lapply(seq(length(id.cen)), function(i)
{ginv(R[[i]])%*%as.matrix(U.ind[[i]][id.cen[i],])})

URTU<-lapply(seq(length(id.cen)),function(i) {NULL})

for (j in seq(length(id.cen))) {
URTU[[j]]<-sapply(seq(nrow(U[[1]])), function(i) {U[[j]][i,]%*%RTJUTJ[[j]]})
                                        }

URTU.mat<-matrix(unlist(URTU),nrow(U[[1]]),length(id.cen))

lambda.c<-apply(URTU.mat,1,cumsum)
c.time<-data[data$cen.ind==0,]$t
k.ind<-NULL
for (i in  seq(nrow(data))) {
k.ind[i]<-which.max(data$t[i]<=c.time)
                                     }
LAMBDA<-NULL
for (i in seq(nrow(data))) {
LAMBDA[i]<-lambda.c[k.ind[i],data$id[i]]
}

k.hat<-exp(-LAMBDA)
k.hat<-c(1,k.hat)
k.hat<-k.hat[-length(k.hat)]
return(k.hat)
                                     }


# State occupation probability  in a five-state illness-death model

s.f<-function(data)  {
odata<-with(data, data[order(t),])

ind1<-which(odata$stage==1)
ind2<-which(odata$stage==2)
ind3<-which(odata$stage==3)
ind4<-which(odata$stage==4)

######   At risk set
# state 0 at risk set
y0.ind<-which(c(ind1,ind2,c1ind)==nrow(odata))
y0.indk<-rep(0, nrow(odata))
```

```
if (length(y0.ind>0)){y0.indk[(c(ind1,ind2,odata$c1ind)+1)[-y0.ind]]<--
1*(odata$kern[c(ind1,ind2,odata$c1ind)[-y0.ind]])} else
{y0.indk[(c(ind1,ind2,odata$c1ind)+1)]<--1*(odata$kern[c(ind1,ind2,odata$c1ind)])}

odata$y0.k<-cumsum(y0.indk)+sum(odata$kern)
odata$y0.k<-(odata$y0.k)/(odata$khat)

# state 1 at risk set

y1.indk<-rep(0, nrow(odata))
y1.ind<-which(ind1==nrow(odata))
if (length(y1.ind>0)) {y1.indk[(ind1+1)[-y1.ind]]<-1*(odata$kern[(ind1)[-y1.ind]])} else
{y1.indk[(ind1+1)]<-1*(odata$kern[(ind1)])}

y.1.ind<-which(c(ind3,ind4,odata$c2ind)==nrow(odata))

if(length(y.1.ind>0)){y1.indk[(c(ind3,ind4,odata$c2ind)+1)[-y.1.ind]]<--
1*(odata$kern[(c(ind3,ind4,odata$c2ind))[-y.1.ind]])} else
{y1.indk[(c(ind3,ind4,odata$c2ind)+1)]<--1*(odata$kern[(c(ind3,ind4,odata$c2ind))])}
y1.k<-cumsum(y1.indk)
odata$y1.k<-y1.k
odata$y1.k<-(odata$y1.k)/(khat)

#  transition probability

dn01.k<-dn02.k<-rep(0, nrow(odata))
dn13.k<-dn14.k<-rep(0, nrow(odata))
dn01.k[ind1]<-(1*odata$kern[ind1])/khat[ind1]
dn02.k[ind2]<-(1*odata$kern[ind2])/khat[ind2]
dn13.k[ind3]<-(1*odata$kern[ind3])/khat[ind3]
dn14.k[ind4]<-(1*odata$kern[ind4])/khat[ind4]

odata$p01.k<-dn01.k/odata$y0.k
odata$p01.k[which(odata$p01.k==Inf)]<-0
odata$p01.k[is.na(odata$p01.k)]<-0

odata$p02.k<-dn02.k/odata$y0.k
odata$p02.k[which(odata$p02.k==Inf)]<-0
odata$p02.k[is.na(odata$p02.k)]<-0

odata$p00.k<-1-(odata$p01.k+odata$p02.k)

odata$p13.k<-dn13.k/odata$y1.k
odata$p13.k[which(odata$p13.k==Inf)]<-0
odata$p13.k[is.na(odata$p13.k)]<-0
```

```
odata$p14.k<-dn14.k/odata$y1.k
odata$p14.k[which(odata$p14.k==Inf)]<-0
odata$p14.k[is.na(odata$p14.k)]<-0

odata$p11.k<-1-(odata$p13.k+odata$p14.k)

final.data<-subset(odata$cen.ind==0)

#state occupation probabilities
final.data$s0.k<-cumprod(final.data$p00.k)

s1p.k<-c(1, final.data$s0.k)*c(final.data$p01.k,0)
s1.k<-numeric(length(s1p.k))
for (i in 2:length(s1.k))
   {
s1.k[1]<-s1p.k[1]
s1.k[i]<-s1p.k[i]+(s1.k[i-1]*c(final.data$p11.k,0)[i])
   }

final.data$s1.k<-s1.k[-length(s1.k)]

s2p.k<-c(0,final.data$s0.k)*c(final.data$p02.k, 0)
final.data$s2.k<-cumsum(s2p.k)[-length(s2p.k)]

s3p.k<-c(0,final.data$s1.k)*c(final.data$p13.k, 0)
final.data$s3.k<-cumsum(s3p.k)[-length(s3p.k)]

s4p.k<-c(0,final.data$s1.k)*c(final.data$p14.k, 0)
final.data$s4.k<-cumsum(s4p.k)[-length(s4p.k)]
list (s1=final.data$s1.k,s2=final.data$s2.k, s3=final.data$s3.k,s4=final.data$s4.k)
}
```

R program to calculate prediction error in fitting AFT model using Elastic net approach

```
library(elasticnet)

#Crossvalidation  function
el.cv.f<-function(y,x,lam,cen.ind) {
logy<-log(y)
or<-order(logy)
Y<-logy[or]
X<-x[or,]
temp.enet <- enet(X,Y,lambda=lam)
```

```
n<-length(Y)

Yhat <- predict(temp.enet,X)
last <- dim(Yhat$fit)[2]

yhatd <- matrix(0,n,last)

for (d in seq(n))  {                              #delete the dth data values

Xd <- X[-d,]
nd <- n-1
Yd <- Y[-d]
temp.enetd <- enet(Xd,Yd,lambda=lam)
tempd<-predict(temp.enetd,newx=X[1:2,])
lastd <- min(dim(tempd$fit)[2],last)

 if(d==1) {
 yhatd[1,1:lastd]<-tempd$fit[1,1:lastd]
             }

 if(d>1)    {
 tempd<-predict(temp.enetd,newx=X[c(d-1,d),])
 yhatd[d,1:lastd]<-tempd$fit[2,1:lastd]
       }
                                  }

#### Inverse Prob. of Censoring
cc<-1-cen.ind
ckm<-survfit(Surv(Y,cc)~1)
tabl<-table(y)
freq<-NULL
for (j in seq(length(tabl))) {freq[j]<-tabl[[j]]}

 KME2<-cbind(freq,ckm$surv)
 modf.KME<-NULL
 s<-matrix(NA,max(KME2[,1])-1,nrow(KME2))
 for(j in seq(nrow(KME2)))
       {
 if (KME2[j,1]>1) {s[1:KME2[j,1]-1,j]<-rep(KME2[j,2],(KME2[j,1])-1)}
       }

  modf.KME<-rev(sort(c(s[!is.na(s)],KME2[,2])))
  K<-c(1,modf.KME)
  K<-K[-length(K)]
```

```
###MSEP

MSEP <- rep(0,last)
for (nt in 1:last) {
MSEP[nt] <- (1/n)*sum((cen.ind/K)*(yhatd[,nt]-Y)^2)
                }
opt.msep<-min(MSEP)

Yhat <- predict(temp.enet,X)$fit

###MSEF
MSEF<-sapply(seq(ncol(Yhat)), function(i) {sum(cen.ind*(Yhat[,i]-
Yi)^2)/sum(cen.ind)})

list(opt.msep=opt.msep,msef=MSEF)
}
```

CURRICULUM VITAE

Farida Mostajabi
E-mail: faridamsb@yahoo.com

EDUCATION

PhD, Biostatistics, GPA 3.85/4.00 2007-2011
Department of Bioinformatics and Biostatistics
University of Louisville, Louisville
Dissertation Title: Regression Methods for Survival and Multistate Models
Advisors: Professors Somnath and Susmita Datta

Master of Science, Biostatistics, GPA 18.90/20 2003-2005
Department of Epidemiology and Biostatistics
Shiraz University of Medical Sciences, Shiraz, Iran
Thesis Title: Obesity Indices & Reference Values among Primary School Children (6- 12 years) of Shiraz

Bachelor of Science, Statistics, GPA 16.44/20 1999-2003
Department of Statistics
Shiraz University, Shiraz, Iran

PUBLICATIONS

Brock G.N, **Mostajabi F.**, Ferguson N., Appana S., Ravindra K., Eng M., Buell JF. and Marvin MR.(2011) Prophylaxis against denovo hepatitis B for liver transplantation utilizing Hep B core(+) donors: does HBIG provide a survival advantage. Transplant Int., 24(6):570-81.

Ayatollahi S.M.T., **Mostajabi F.**(2008) Triceps skinfold thickness (TST) in primary school children of Shiraz, Iran. Archives of Iranian Medicine, 11(2): 210 - 213.

Ayatollahi S.M.T., **Mostajabi F.**(2007) Prevalence of obesity among school children in Iran. Journal of Obesity Review, 8(4): 289-91.

Ayatollahi S.M.T., **Mostajabi F.**(2006) Body mass index of school children in Shiraz (southern Iran) and CDC standards. Iranian Red Crescent Medical Journal, 9(4):185-190.

Mostajabi F., Ayatollahi S.M.T. Assessment of adiposity by triceps skinfold thickness (TST) in school children of Shiraz, Iran. Proceeding of the 8th Iranian Statistical Conference (2006), Shiraz, Iran.

Negahban A., **Mostajabi F.**(2004) Guide to research method using questionnaire with SPSS 11, Published by IACECR/ Tehran University, ISBN: 964-8171-06-8, In Persian language, 2nd Ed.

Mostajabi F., Datta S, Datta S. Predicting patient survival from proteomic profile using MALDI-TOF mass spectrometry data. Submitted to Statistical Methods in Medical Research.

Mostajabi F., Datta S, Non-parametric regression of state occupation, entry and exit times with multistate right censored data. In Preparation.