A DATA-DRIVEN ALGORITHM FOR PARAMETER ESTIMATION IN THE PARAMETRIC SURVIVAL MIXTURE MODEL

A DATA-DRIVEN ALGORITHM FOR PARAMETER ESTIMATION IN THE PARAMETRIC SURVIVAL MIXTURE MODEL

By

Jin Zhang

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Master of Science

McMaster University

MASTER OF SCIENCE (2007) McMaster University

(Statistics) Hamilton, Ontario

TITLE: A Data-Driven Algorithm for Parameter Estimation in the Parametric Survival Mixture Model

AUTHOR: Jin Zhang

(McMaster University, Canada)

SUPERVISOR: Dr. Rong Zhu

NUMBER OF PAGES: ix, 51

Abstract

We propose a data-driven estimation algorithm in survival mixture model. The objective of this study is to provide an alternative fitting procedure to the conventional EM algorithm. The EM algorithm is the classical ML fitting of the parametric mixture model. If the initial values for the EM algorithm are not properly chosen, the maximizers might be local or divergent. Traditionally, initial values are given manually according to experience or a grid-point search. This is a heavy burden for a high-dimensional data sets. Also, specifying the ranges of parameters for a grid-point search is difficult. To avoid the specification of initial values, we employ the random partition. Then, improvement of fitting is adjusted according to model specification. This process is repeated a large number of times, so it is computer-intensive. The large repetitions makes the solution more likely to be the global maximizer, and it is driven purely by the data. We conduct a simulation study for three cases of two-component Log-Normal, two-component Weibull, and two-component Log-Normal and Weibull, in order to illustrate the effectiveness of the proposed algorithm. Finally, we apply our algorithm to a breast cancer study data which follows a cure model. The program is written in R. It calls existing R functions, so it is flexible to use in regression situations where model formula must be specified.

Acknowledgements

This study is supported by Dr. Rong Zhu. Thanks to him for his supervision. I am grateful to Dr. Dongsheng Tu at Queen's University for providing the data and the reference. I also would like to thank Professor N. Balakrishnan and Professor R. Viveros for their advice. I give my special thanks to my friends and colleagues: Yunna, Liqin, Anne, Carolyn, Frank, Lou Anne, Marzena, Mike and Nancy. Thanks for their spiritual support and help to polish my English. Finally, I highly appreciate my family for their encouragement during this project.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Survival Analysis

The very early work of survival analysis arose in mortality research in the $17th$ century. Since Edmond Halley (see M. Greenwood (1938)) published the first life table, this method has been applied widely by actuaries, statisticians and biomedical researchers. During World War II, this method was used to study the reliability of military equipment. After that, the method was developed further and applied to study the "lifetime" of industrial devices and the survival time of patients. Also, "survival analysis" was named by cancer researchers. In recent decades, survival analysis has become one of the major methods of analysis in medicine, environmental health, marketing and industry.

Imagine that, for a researcher working in health science, there is a project to study the effectiveness of a new treatment for a disease, such as cancer. The main variable of interest is the number of days that the patients survive. The explanatory variables are age, gender, the initial performance status, etc. In addition, there must be one necessary variable for survival analysis: a variable indicating if the patients are dead, alive, or contact has been

lost.

For different research areas, the variables of interest are different. In diabetes research, the variable of interest might be the time from the diagnosis of diabetes to the time of the development of diabetic retinopathy in the patients.

The engineering sciences have made a great contribution to the development of survival analysis. The most popular research in this area is the lifetime of a product. This is about the product reliability analysis, such as the lifetime of televisions, computers, tires, etc.

In social sciences, researchers might be interested in analyzing job changes in modern cities during a specific time period. The number of months after which individuals change their jobs is the variable of interest in this case.

Certainly, the application areas of survival analysis are diverse, including the areas of biomedical, social sciences, engineering, economics, etc. It is difficult to give a proper definition of survival analysis. But it is obvious that survival analysis focuses on lifetime, survival time and failure time data.

Normally, the "survival time" is defined for convenience before the survival analysis. Death or failure is called an "event", so models of death or failure are termed *time-to-event models*. Survival time means a period of time from the start to the time an event happens. Survival analysis deals with death in biological organisms, failure in mechanical systems, reliability in engineering and duration in economics. In the case of biological analysis, death is clear, but for other areas, failure is not always unambiguous. It is necessary to define events early.

In this thesis, we only consider death or failure that happens just once for each subject. We exclude the case of *recurring event* or *repeated event* models.

It is easy now to identify the questions that survival analysis tries to answer:

- the proportion of a population which will survive a certain time;
- the rate of those who are alive in a study or treatment group;
- the potential causes of death;
- the way the particular characteristics change the odds of survival, and so on.

Censoring is a particular phenomenon in survival analysis. It arises from the fact that some subjects do not have events recorded; survival times may be unknown for a group of subjects in a study. This leads to the term "censoring". The observation that involves censoring is called a censored observation. Also, the time variable of interest in the censored observation is called "censored time" instead of "survival time". Censored time records the time from the start of the study to the time the observation cannot be observed any longer. During this period, no events occur. If an observation is non-censored, it gives exact survival information about the subject.

Right censoring, left censoring and interval censoring are three kinds of censoring. We focus on the right censoring which is much easier to understand. When right censoring occurs, it means that the event time is longer than the censored time: the study is closed or the subject is lost from follow-up. A right-censored observation means that it only contains partial information since the subject does not have an event during the time when the subject is in the study. There are two kinds of right censoring: fixed-right censoring and random-right censoring. If no event occurs for a subject until the end of study, then it is called fixed-right censoring. If no event occurs and the subject is lost to observation before the end of study, then it is called random-right censoring. It is random because this censoring is determined by the censoring mechanism and not by the researcher. Right censoring has two types of censoring mechanisms. Type 1 censoring mechanism means that the observation time is fixed. The censored indicator can tell if the events do not happen before the fixed observation time. Type 2 censoring mechanism means that the study stops when there is a

specified number of events happening.

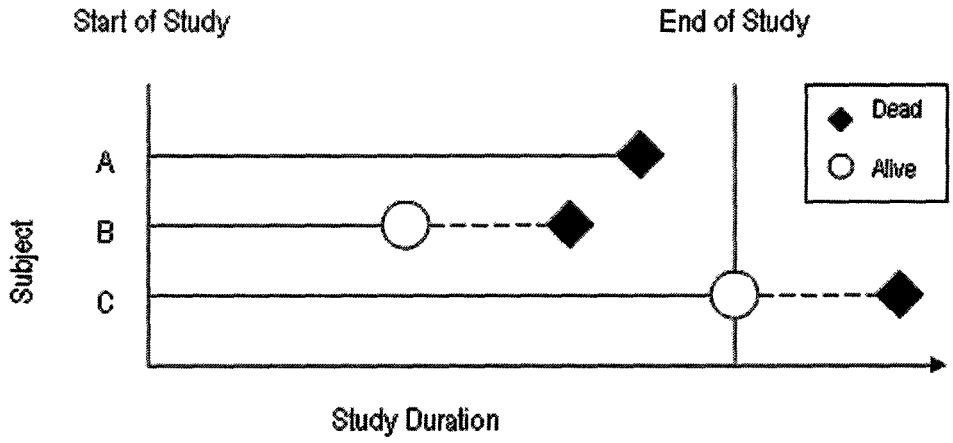

Figure 1.1: Censoring Illustration

We illustrate three subjects A, B and C in Figure 1.1, and they have the same study duration. Assume this is a study of the survival of heart transplant patients who are followed up for a half year. The event here is death, and the variable of interest is "survival time". Not all patients die within the half year study period. Subject A dies before the end of study. This event can be observed during the study. This data point is not a censored observation (non-censored). The follow-up study is lost for subject B before the end of study. We do not know if this patient is alive or dead. Although the patient may die before the end of study, we cannot observe this event. This is called *random-right censoring* because the censoring is regarded as random. Nobody knows which patients will not come into hospital for checkups before the study. Subject C is alive until the end of study. This censoring is called *fixed-right censoring*. Although this patient might die two years after surgery, the death cannot be observed during this half year study.

Let T be a random lifetime variable, for both continuous and discrete models, the survival

function will be as:

$$\begin{aligned} S(t) = \quad & Pr(T \geq t) = \quad \int_t^{\infty} f(x)dx \qquad \text{(Continuous Model)} \\ S(t) = \quad & Pr(T \geq t) = \quad \sum_{j:t_j \geq t} f(t_j) \qquad \text{(Discrete Model)} \end{aligned} \tag{1.1}$$

where $t \geq 0$ is the observed lifetime.

There are a number of theoretical distributions used to approach the survival models. The exponential distribution is a basic model for survival time, which is a special case of the Weibull distribution. Log-Normal and Log-Logistic distributions are also widely used. All these distributions mentioned above actually belong to the parametric Log-Location-Scale model with pdf (probability density function):

$$f(y; u, b) = \frac{1}{b} f_0\left(\frac{y-u}{b}\right), \quad -\infty < y < \infty \tag{1.2}$$

where u $(-\infty < u < \infty)$ and $b > 0$ are location and scale parameters, $f_0(\cdot)$ is a specified pdf on $(-\infty, \infty)$, respectively the survival function has the form:

$$S(y; u, b) = S_0\left(\frac{y-u}{b}\right), \quad -\infty < y < \infty, \tag{1.3}$$

where $S_0(\cdot)$ is a fully-specified survival function defined on $(-\infty, \infty)$.

Survival analysis has nonparametric, parametric and semiparametric analysis methods. We focus on parametric methods. In survival analysis, the major variable of interest is the time of survival. Also, there are some explanatory variables or covariates such as domestic information, environmental conditions, and treatment indicators which give more information and may be correlated with the variable of interest. It is common to consider the regression model to study the explicit relationship through the parameters. Given a vector of covariates $\boldsymbol{x}$, the distribution of the dependent variable of interest Y can be specified. The survival function of this regression model associated with $\boldsymbol{x}$ is defined as:

$$S(y \mid \boldsymbol{x}) = S_0\left(\frac{y-u(\boldsymbol{x})}{b}\right), \tag{1.4}$$

where the scale parameter b does not depend on $\boldsymbol{x}$ (this is a simplified assumption), the location parameter u is a linear function of $\boldsymbol{x}$ with unknown coefficients, and for the standardized random variable $Z = (Y - u)/b$, the survival function $S_0(z)$ with $u = 0, b = 1$ is called the standard form of the distribution.

1.2 Mixture of Survival Models and Research Objectives

In many statistical applications, the observations are taken from multiple subpopulations. The simple model, assuming data from one population, is not always appropriate and may lead to wrong inference. Thus, the mixture models would be considered to better describe the real situation.

The mixture model allows us to combine the samples from different subpopulations but without membership information. Usually, the subpopulation is called *mixture component* or *component.* If the number of components is finite, the models are called *finite mixture models.*

The probability mixture model is a probability distribution that is a convex combination of other probability distributions. Suppose we have n individual subjects in the mixture data set, $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ denote a random sample of size n, where $\boldsymbol{Y}_j$ is a p-dimensional random vector. Assume the whole population can be divided into K subpopulations, meaning that this data set has K components. Let a_i be the weight for the ith component so that $a_1 + a_2 + \cdots + a_K = 1$. We use $f(\boldsymbol{y}_j)$ to denote the pdf of the observation j and $f_i(\boldsymbol{y}_j)$ to denote the chance that the observation j is from component i. Then the density $f(\boldsymbol{y}_j)$ of $\boldsymbol{Y}_j$ is defined as a weighted sum of its component distributions:

$$f(\boldsymbol{y}_j) = \sum_{i=1}^{K} a_i f_i(\boldsymbol{y}_j), \quad (i = 1, \cdots, K; j = 1, \cdots, n), \tag{1.5}$$

where the weights $a_1, \cdots, a_K$ $(0 < a_i < 1)$ are the mixing proportions.

If the probability models are from a parametric family, say with pdf $f(\boldsymbol{y}_j; \boldsymbol{\theta}_i)$, the density function is then:

$$f(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \sum_{i=1}^{K} a_i f_i(\boldsymbol{y}_j; \boldsymbol{\theta}_i), \tag{1.6}$$

where $\boldsymbol{\theta}_i$'s are the unknown parameters, and $\boldsymbol{\Psi} = (a_1, \cdots, a_K, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K)$. Our work here only focuses on the parametric mixture model. Then the survival function for the parametric survival mixture model is:

$$S(t_j; \boldsymbol{\Psi}) = \sum_{i=1}^{K} a_i S_i(t_j; \boldsymbol{\theta}_i), \tag{1.7}$$

where $S_i(t; \boldsymbol{\theta}_i)$ denotes the survival function of the ith component.

Typically, the parameters and the number of components are all unknown. From the data, we must determine or know the number of components, and estimate the parameters of each component distribution. The component data of mixture models can be regarded as missing. Each observation belongs to one subpopulation, thus has a membership. However, this membership is missing. Therefore, the estimation of mixture models becomes a missing data problem. The most common estimation approach for missing data is EM (Expectation Maximization) method.

It is possible that the population is composed of multiple subpopulations in parametric survival analysis. This leads to the parametric mixture survival analysis. For example, when studying diseases with a multi-stage progression where, in each stage, survival time can be modelled with different parameters or different models.

The very early work of parametric mixture survival modelling was done by Boag (1949). The models are based on the standard failure time densities using ML (Maximum Likelihood) to estimate cured proportion and death rate. In recent years, Chen (1985) applied Bayesian analysis of a two-component mixture survival model for the cancer patient analysis. Marin et al. (2005) fitted the Weibull mixture model with an unknown number of components to the right-censored survival data.

When doing parametric mixture survival modelling, two important things are required. One is finding out the optimal number of components, the other is fitting the models. There are two steps for completing the second task:

- Define suitable models for the data
- Estimate parameters of the models from the data

Suppose we have a censored data set of the cancer patients study. The variable of interest is the survival years after surgery and the only one covariate is the tumour size of the patients before the surgery. The data set has two components. We use the parametric mixture survival modelling analysis method. Two component models, model 1 and model 2, have the same Log-Normal distribution with different parameters. Figure 1.2 illustrates the scatter plot of data from a two-component parametric mixture survival model. The solid diamonds indicate the data points from model 1. The triangles indicate the data points from model 2. Imagine we lose the membership information. The points included in the ellipse might be from model 1 or from model 2. It is hard to judge their membership virtually. This kind of phenomenon raises difficulties in fitting the mixture models. The error of misclassification cannot be avoided. Normally, it depends on how the mixture models overlap. The bigger the overlap, the bigger the error.

In this thesis, we study the parametric mixture survival models based on Weibull and Log-Normal distributions. The introduction of these distributions and survival models is in Section 2.1. In Section 2.2, we show the detailed steps of the estimation algorithm about how we are fitting the mixture survival models. The simulation study in Chapter 3 provides three cases to show how we simulate the mixture survival data set and the model fitting results. Chapter 4 is an application of this estimation algorithm.

Since the classic paper of Pearson (1894) on the moments-based fitting of a mixture model,

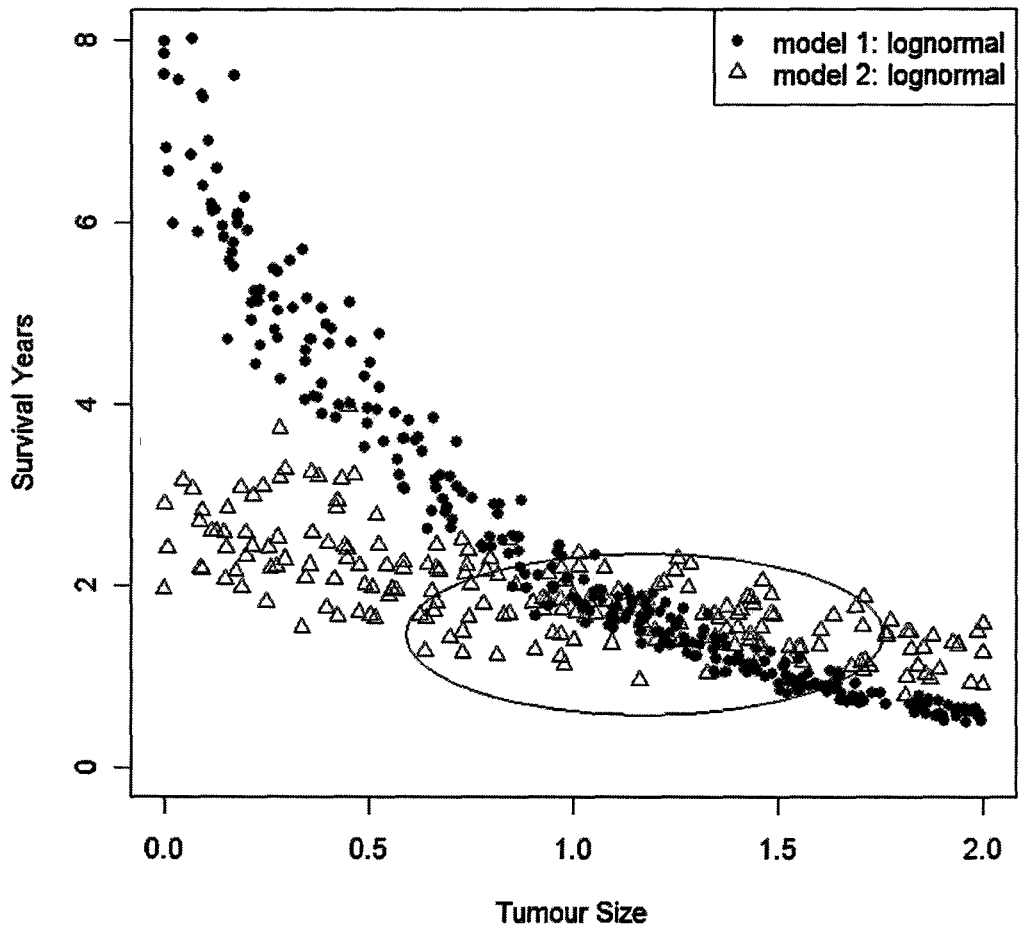

Figure 1.2: Scatter Plot of Data from a Two-Component Parametric Mixture Survival Model

many researchers have been involved in looking for the mixture model approach methods until the EM algorithm of Dempster et al. (1977) suggested an iterative reweighting scheme to compute MLE (Maximum Likelihood Estimator). The EM algorithm has become classical statistical inference method for solving the classification problems of mixture survival models with the advent of high-speed computers. This iterative computation of MLE ensures the likelihood values increasing monotonically. However the EM algorithm has some weaknesses. The convergence is slow. And, as we know, the EM algorithm requires some initial values of parameters. In practice, the initial values are chosen by experience. Normally, we first make the range of values for each parameter. Then we try every combination of those values as starting values. This kind of initial value choice does not work for high dimensional data since the combination could be very large. Furthermore, the different starting values can lead to quite different estimates in the context of fitting mixture models of exponential components; see Seidel, Mosler and Alker (2000a,b). A poor choice in the starting values makes the convergence slower. Also, the matter could be even worse. The sequence of the estimates may diverge. Here, this data-driven algorithm attempts to fit the mixture model with easier and more reasonable initial value choices and faster convergence.

Decision tree learning is another widely used technique for classification. The classification model is a tree, called a decision tree. It provides a highly effective structure within which you can predict the classes of new cases or instances. For example, due to the high cost of ICU (intensive-care unit), the patients who may survive less than a month are given higher priority. A decision tree can be built for helping to decide whether to put a new patient in ICU. However, decision trees do not work very well in cases which need to identify an unknown variable.

Chapter 2

Parametric Survival Models and Data-Driven Estimation Algorithm

2.1 Introduction to Parametric Survival Models

The survival function is called *survivor function* or *survivorship function* in biological survival analysis. It is also known as *reliability function* in mechanical survival analysis. Let T be a single nonnegative continuous random variable which represents the survival time of the subject with cumulative distribution function (cdf) $F(t)$ on the interval $[0, \infty]$. The cdf of T is:

$$F(t) = Pr(T \leq t) = \int_0^t f(x)dx, \tag{2.1}$$

where $f(x)$ denotes the probability density function (pdf) of T. We then define the survival function as:

$$S(t) = Pr(T \geq t) = \int_t^\infty f(x)dx = 1 - F(t). \tag{2.2}$$

Every survival function $S(t)$ is monotonically decreasing with $S(0) = 1$ and $S(\infty) = \lim_{t \to \infty} S(t) = 0$. Thus, the survival function is the probability that the time of event (e.g., death) occurs

later than some specified time t.

$h(t)$ denotes the hazard function and is defined as follows:

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} \tag{2.3}$$

$$= \frac{f(t)}{S(t)} \tag{2.4}$$

$$= -\frac{d}{dt} \log S(t). \tag{2.5}$$

This function gives the instantaneous survival rate conditional on survival at time t.

Many parametric families are applied in survival analysis. It is common to select statistical distributions which have nonnegative support since survival times are nonnegative. In this paper, three commonly used distributions are considered.

2.1.1 The Exponential Distribution

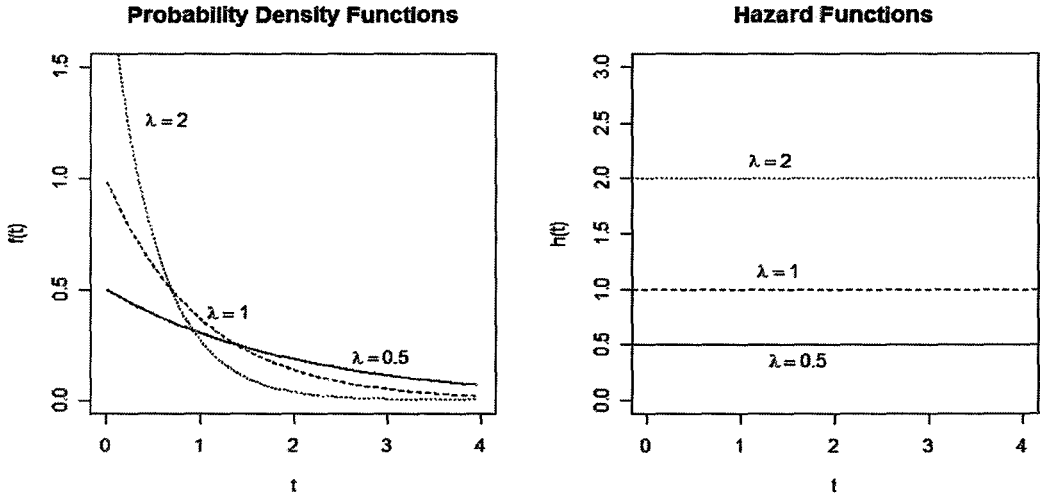

Figure 2.1: Exponential pdf and hazard function

The pdf, survival function and hazard function are:

$$f(t) = \lambda e^{-\lambda t}, \quad t > 0, \ \lambda > 0, \tag{2.6}$$

$$S(t) = e^{-\lambda t}, \tag{2.7}$$

$$h(t) = \lambda. \tag{2.8}$$

Exponential distribution was widely used in the early survival analysis since it allows for a simple statistical method. However, its application is limited because of the constant hazard function.

2.1.2 The Weibull Distribution

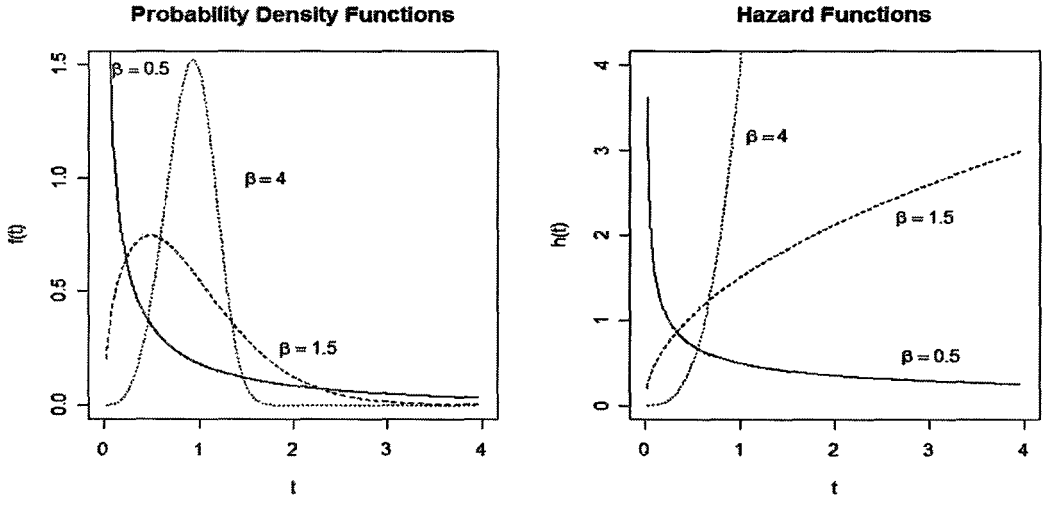

Figure 2.2: Weibull pdf and hazard function ($\lambda = 1$)

The pdf, survival function and hazard function are:

$$f(t) = \lambda\beta(\lambda t)^{\beta-1}\exp[-(\lambda t)^{\beta}], \quad t > 0, \lambda > 0, \beta > 0, \tag{2.9}$$

$$S(t) = \exp[-(\lambda t)^{\beta}], \tag{2.10}$$

$$h(t) = \lambda\beta(\lambda t)^{\beta-1}. \tag{2.11}$$

The Exponential distribution is a special case of Weibull distribution when $\beta = 1$.

2.1.3 The Log-Normal Distribution

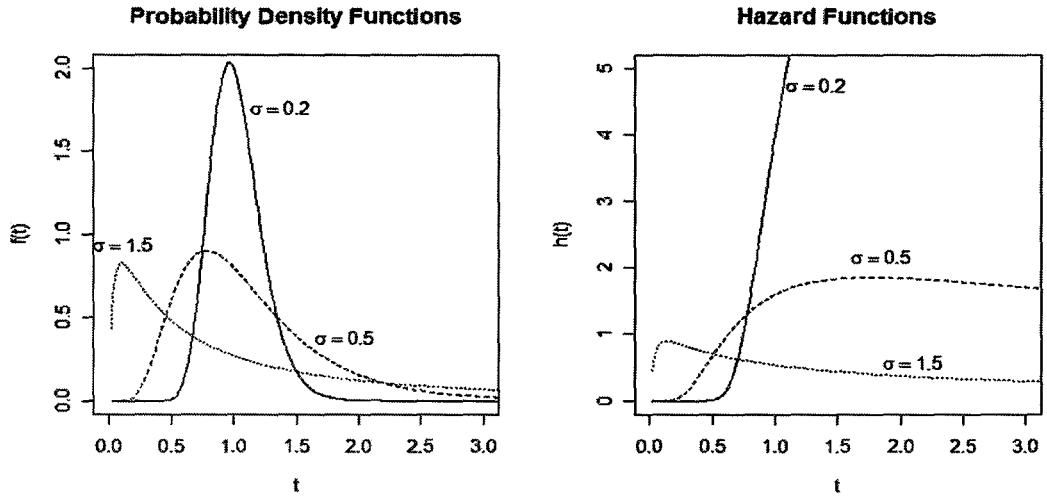

Figure 2.3: Log-Normal pdf and hazard function ($\mu = 0$)

The pdf and survival function are:

$$f(t) = \frac{1}{(2\pi)^{1/2}\sigma t} \exp\left[-\frac{1}{2}\left(\frac{\log t - \mu}{\sigma}\right)^2\right], \; t > 0, \; -\infty < \mu < \infty, \; \sigma > 0, \quad (2.12)$$

$$S(t) = 1 - \Phi(\frac{\log t - \mu}{\sigma}), \quad (2.13)$$

where $t > 0$ and the standard normal distribution function:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{(2\pi)^{1/2}} e^{-z^2/2} dz.$$

The hazard function is given as formula (2.3). It has the value 0 at $t = 0$, increase to a maximum, and then decrease, finally approaching 0 as $t \to \infty$.

2.1.4 Regression for a Parametric Survival Model

Regression is a statistical model investigating the relationship between a dependent variable (response variable) and the variables (covariates, explanatory variables, predictors). Specifically, it relates covariates to the mean of response. Let T be a random variable of survival time, if $Y = \log T$ has a Location-Scale distribution, then $T = \exp(Y)$ has a Log-Location-Scale distribution. As introduced in Section 1.1, the pdf (1.2) and survival function (1.4), the survival function for T is:

$$S^*(t;\alpha,\beta) = S_0\left(\frac{\log t - u}{b}\right) \tag{2.14}$$

$$= S_0^*[(t/\alpha)^\beta], \tag{2.15}$$

where $\alpha = \exp(u), \beta = b^{-1}$ and for $0 < w < \infty, S_0^*(w) = S_0(\log w)$. T given $\boldsymbol{x}$ corresponding to (1.3) has the form:

$$S(t \mid \boldsymbol{x}) = S_0^*[(t/\alpha(\boldsymbol{x}))^\beta], \quad t \geq 0, \tag{2.16}$$

where $\alpha = \exp(u(\boldsymbol{x})), \beta = b^{-1}$, and $S_0^*(t) = S_0(\log t)$.

Suppose for a study, there are p covariates. Denote $\boldsymbol{x} = (x_1, \cdots, x_p)'$ as the covariate vector, and $\boldsymbol{\beta}$ as a $p \times 1$ vector. Then the survival function of exponential distribution given $\boldsymbol{x}$ is:

$$S(t \mid \boldsymbol{x}) = \exp[-\lambda(\boldsymbol{x})t], \tag{2.17}$$

The location parameter u depends on $\boldsymbol{x}$ and $u(\boldsymbol{x}) = \boldsymbol{\beta}'\boldsymbol{x}$, where $\boldsymbol{\beta}$ is the regression coefficient vector. A function of parameter λ conditional on $\boldsymbol{x}$ is: $\lambda(\boldsymbol{x}) = \exp(\boldsymbol{\beta}'\boldsymbol{x})$.

The purpose of fitting a regression model includes prediction, examining the relationship between variables, and testing hypotheses. It is necessary to specify the relationship between the expectation $E(T \mid \boldsymbol{x})$ and the linear predictor $\boldsymbol{\beta}'\boldsymbol{x}$. The following is the form for each

distribution mentioned above with specification $u(\boldsymbol{x}) = \boldsymbol{\beta}'\boldsymbol{x}$:

$$\begin{aligned} E(T \mid \boldsymbol{x}) &= \exp(\boldsymbol{\beta}'\boldsymbol{x}) && (Exponential) \\ E(T \mid \boldsymbol{x}) &= \exp(\boldsymbol{\beta}'\boldsymbol{x})\Gamma(1+1/\beta) && (Weibull) \\ E(T \mid \boldsymbol{x}) &= \exp(\boldsymbol{\beta}'\boldsymbol{x}+\sigma^2/2) && (Log-Normal) \end{aligned} \tag{2.18}$$

2.2 Parameter Estimation: A Data-Driven Algorithm

Normally, there are two missions for the model fitting of survival mixture: getting the optimal number of components and estimating the parameters for each component model. This data-driven estimation algorithm focuses on the second mission: parameter estimation.

As shown by the outline flowchart (Figure 2.4), **Self Start** and **Partition** are the two main parts of this algorithm. Figure 2.5 and 2.6 are the detailed flowcharts of these two main procedures. **Self Start** is a randomly-starting routine for the initial value choice. **Partition** is a routine for membership determination and parameter estimation. **Boundary Check** checks if the input is proper or supported by this program. **Graphic & Numerical Report** supply the screen graphics during the model fitting procedure and the model fitting results, such as the membership, proportion, log-likelihood, etc.. There are two loops (`loop 1` and `loop 2`) in this estimation algorithm. `loop 1` is an inner loop within **Partition** (See Figure 2.6). It runs in order to get more precise estimates under the same initial values. `loop 2` is an outer loop including **Self Start** and **Partition** (See Figure 2.4). It tries to give more accurate estimation by getting some different initial values. The routine `survreg()` of R is used for the survival model fitting. This is a regression routine for parametric survival models. The default method used in fitting models is the iteratively reweighted least squares method (IRLS).

Let $\boldsymbol{Y} = (T, E, \boldsymbol{X})$, where T is the random variable of survival time, E is the censoring event, and $\boldsymbol{X}$ is the covariate vector. Suppose a sample of size n is $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$, where

$\boldsymbol{Y}_j = (T_j, E_j, \boldsymbol{X}_j)$ is a random vector. The pdf is $f(\boldsymbol{y}_j)$, where $\boldsymbol{y}_j$ is the observed value of the random vector $\boldsymbol{Y}_j$. We define $\boldsymbol{x}_j$ to be the observed covariate vector, t_j to be the j^{th} observed life-time. K is the number of components. $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K$ are the unknown parameters of the K component models. $a_1, \cdots, a_K$ are the weights or proportions.

This algorithm requires proper inputs. The following lists four important inputs.

- **data**: the censored data set including response variable (T) and covariates ($\boldsymbol{x}$). The response variable is a continuous numerical variable which is the survival time. The covariates could be numerical or nominal.
- K: the number of components must be a positive integer.
- **distributions**: the parametric distributions for component models which must be supported by R.
- **formulae**: the survival regression formulae for component models.

Aside from these, there are some optional inputs with default values, such as the numbers of loops. The following is the step-by-step explanation of the estimation algorithm.

- Step 1: Self Start

  This is a random starting procedure yielding the initial values from the subset of the data.

  (1) Select a random sub-sample

  A sub-sample with size m is randomly selected from the whole sample data. If the covariate vector $\boldsymbol{x}$ has p dimensions, then the sub-sample size cannot be smaller than $(p+1) \times K$.

  (2) Partition the sub-sample

  The sub-sample is equally partitioned into K groups, e.g., assign each data point

a membership. For example, for a case of $K = 2, n = 500$, we select $m = 200$ observations from the sample as the sub-sample. Then, the first 100 observations are assigned to group 1. The group 2 will include the remaining 100 observations. Here we assume that the $m = 200$ random observations are not sorted.

(3) Proportion calculation

After the membership assignment, the proportion of the K groups is calculated based on the sub-sample data set. The initial proportions for $K = 2$ are 0.5 and 0.5.

(4) Model fitting

The parameters, $\boldsymbol{\theta}_1^{(0)}, \cdots, \boldsymbol{\theta}_K^{(0)}$, are estimated for each specified model. This parameter estimation procedure is essentially a least squares linear regression algorithm since the major distributions of survival data can be "transformed" to linear properly. In this project, a parametric survival regression procedure *survreg* in R is used. The method to be used in fitting the model is IRLS, which is a numerical algorithm for minimizing any specified objective function using a standard weighted least squares method.

The output of the **Self Start** includes the estimated proportions $\hat{a}_1, \cdots, \hat{a}_K$, and the estimated parameters $\widehat{\boldsymbol{\theta}}_1^{(0)}, \cdots, \widehat{\boldsymbol{\theta}}_K^{(0)}$.

- Step 2: Partition

This step refines the estimated parameters for each component model and adjusts the membership for every observation.

(1) Determination of the membership for each observation

First of all, we calculate the pdf values for each observation based on the estimated parameters of the mixture models. This means that each observation can have K possible density values: $\hat{f}_1(\boldsymbol{y}_i), \cdots, \hat{f}_K(\boldsymbol{y}_i),\ i = 1, \cdots, n$. Then, we multiply

the calculated proportions of the K models by the pdf values to get the weighted probabilities for each observation. These probabilities indicate the likelihood that one observation belongs to the K component models. We assign the membership of one observation to the model which has the highest probability.

For example, for the mixture model of $K = 2$, if the jth observation $(t_j, \boldsymbol{x})$ is uncensored, then it has $S_1(t_j; \boldsymbol{x})$ and $S_2(t_j; \boldsymbol{x})$ possibilities to the first and second components respectively. The membership of the jth observation is assigned to the first component if $\hat{S}_1(t_j; \boldsymbol{x}) \geq \hat{S}_2(t_j; \boldsymbol{x})$, otherwise it is assigned to the second component. Furthermore, we may penalize the small groups or highlight the small groups by using $\hat{a}_i \hat{S}_i(t_j; \boldsymbol{x})$ or $\hat{a}_i^{-1} \hat{S}_i(t_j; \boldsymbol{x})$. In our study, we penalize the small groups using the former formula.

(2) Proportion calculation

According to the membership from the previous step, the proportion of the K models is recalculated. We only need to compute $(K-1)$ proportions, since there is a constraint that the sum of all proportions is 1. For the case of $K = 2$, if the sample size is 250 and the number of observations belonging to group 1 is 100, then the proportion of group 1 is $a_1 = 100/250$. The proportion of group 2 should be $a_2 = 1 - a_1$.

(3) Likelihood calculation

A likelihood function is the probability for the occurrence (e.g., the data set). The general likelihood function of the parametric model is defined as:

$$L(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n) = \prod_{j=1}^{n} f(\boldsymbol{y}_j; \widehat{\boldsymbol{\theta}}), \tag{2.19}$$

where $f(\boldsymbol{y}_j; \widehat{\boldsymbol{\theta}})$ is the parametric form of pdf $f(\boldsymbol{y}_j)$. For the right-censored data,

the general likelihood function is defined as:

$$L(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \cdots, \boldsymbol{y}_n) = \prod_{j=1}^{n} f(t_j; \widehat{\boldsymbol{\theta}})^{\delta_j} S(t_j+; \widehat{\boldsymbol{\theta}})^{1-\delta_j}, \tag{2.20}$$

where δ is the censoring indicator. $\delta_j = 0$ means that the jth unit is survived until the end of the study. $S(t_j+)$ is the same as $Pr(T_j > t_j)$. If $S(t)$ is continuous at t_j, $S(t_j+)$ equals $S(t_j)$.

The general likelihood function of mixture models is defined as:

$$L_0 = \prod_{j}^{n} \left[\sum_{i}^{K} \hat{a}_i f_i(t_j; \widehat{\boldsymbol{\theta}}_i)^{\delta_j} S_i(t_j+; \widehat{\boldsymbol{\theta}}_i)^{1-\delta_j} \right]. \tag{2.21}$$

In this case, for the mixture survival models, two kinds of likelihood functions L_1 and L_2 are used in practice. They have forms as follows:

$$L_1 = \prod_{i=1}^{K} \left[\prod_{j=1}^{n} \hat{a}_i f(t_j; \widehat{\boldsymbol{\theta}}_i)^{\delta_j} S(t_j; \widehat{\boldsymbol{\theta}}_i)^{1-\delta_j} \right], \tag{2.22}$$

$$L_2 = \prod_{i=1}^{K} \left[\prod_{j=1}^{n_i} f(t_j; \widehat{\boldsymbol{\theta}}_i)^{\delta_j} S(t_j; \widehat{\boldsymbol{\theta}}_i)^{1-\delta_j} \right], \tag{2.23}$$

where $\hat{a}_i$ is the estimated proportion of the i^{th} group and $\hat{a}_1 + \hat{a}_2 + \cdots + \hat{a}_K = 1$, $\widehat{\boldsymbol{\theta}}_i$ is the estimated parameter vector for the i^{th} group, $n_1 + n_2 + \cdots + n_K = n$. The log-likelihood function is defined as $l(\boldsymbol{\theta}) = \log L$. We define $Q = -\log L_1$ or $Q = -\log L_2$. The ML (the minimum Q) is the criterion used for the model selection. As shown above, L_1 is the penalized approximation of likelihood L_0, L_2 is the usual approximation of likelihood L_0. In practice, there are no big differences between them, when a_i's are not dramatically different.

(4) Model selection

A batch of fitted mixture models are selected in this step. The fitted results include the estimated parameters, membership, log-likelihood, and proportions. After each inner loop, we have one fitted mixture model. Whether this model is

saved or not, a saving rule can be followed. The rule is saving those models which have large log-likelihood values (small Q values). We define Q_s to be the smallest one, Q_c to be the current one and ϵ to be the prescribed critical value. Three cases are considered for comparing the Q_c and Q_s:

- If $Q_c = Q_s$, then this fitted model will be saved.
- If $Q_c > Q_s + \epsilon$, then this fitted model will not be saved.
- If $Q_c < Q_s$, then this fitted model can be saved. Also, Q_c becomes the new Q_s. Comparing the new Q_s with the Q values of the saved models, we drop the model if its Q value is larger than the new $Q_s + \epsilon$.

(5) Plotting

This step produces the 2D scatter plots of the covariate and the corresponding predict responses and the observed responses. It works only if the model has one covariate. These plots give a quick visual check of the model fitting results.

(6) Model fitting

This step is the same as the Model fitting in Step 1 (Self Start). The only difference is that the membership for each observation is determined by the calculation above, not a random assignment. Also, the data is the whole sample data set, not the sub-sample.

The **Partition** step repeats many times such that a batch of fitted mixture models can be saved for estimating the final mixture model. Let M_l be the l^{th} fitted mixture model, $f_k^{(l)}$ be the k^{th} component. The following lists the batch of models.

$$\begin{array}{lcccc} M^{(1)}: & f_1^{(1)}, & f_2^{(1)}, & \cdots, & f_K^{(1)} \\ M^{(2)}: & f_1^{(2)}, & f_2^{(2)}, & \cdots, & f_K^{(2)} \\ \vdots & \vdots & \vdots & & \vdots \\ M^{(L)}: & f_1^{(L)}, & f_2^{(L)}, & \cdots, & f_K^{(L)} \end{array}$$

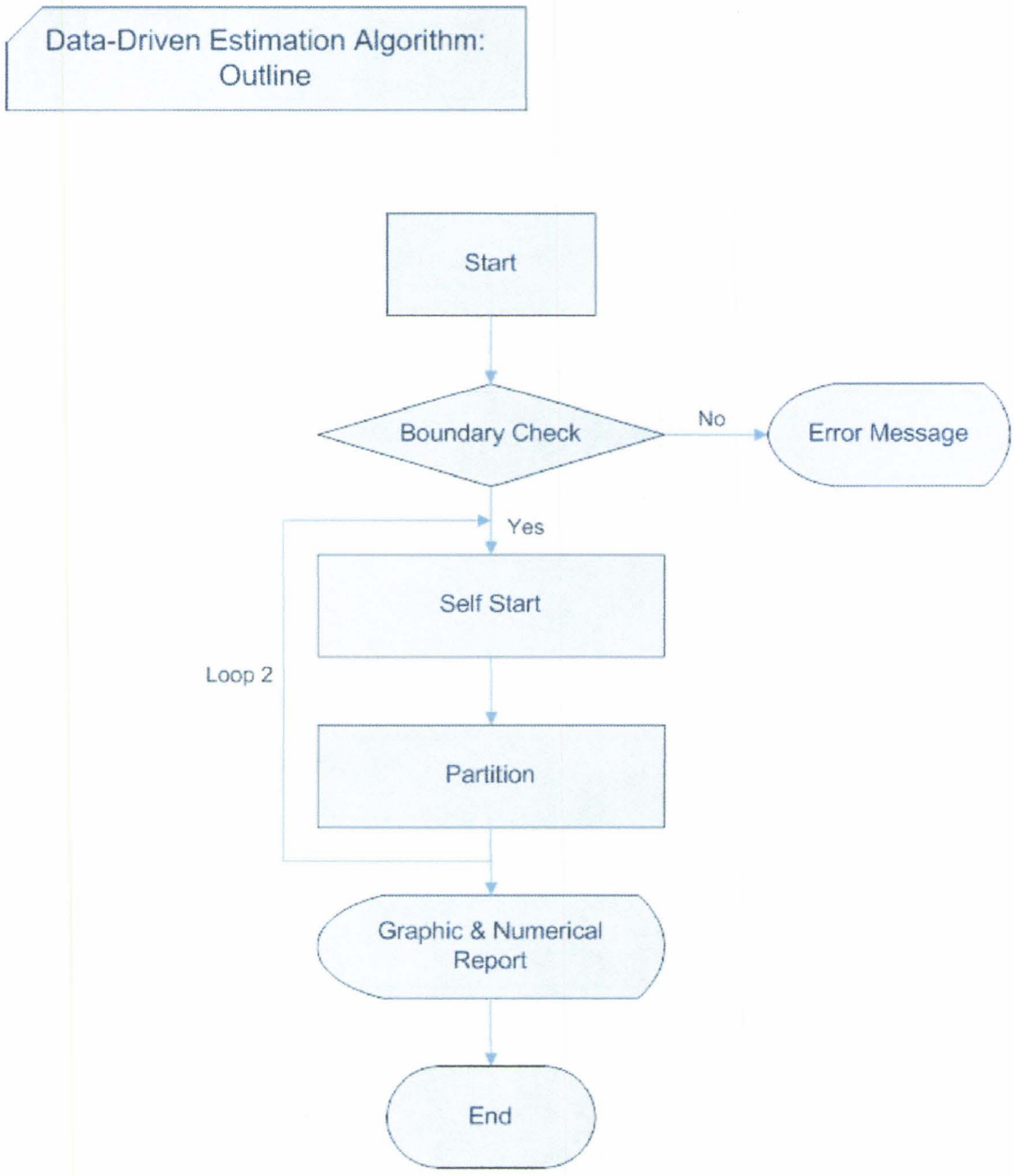

Figure 2.4: Flowchart of Data-Driven Estimation Algorithm: Outline

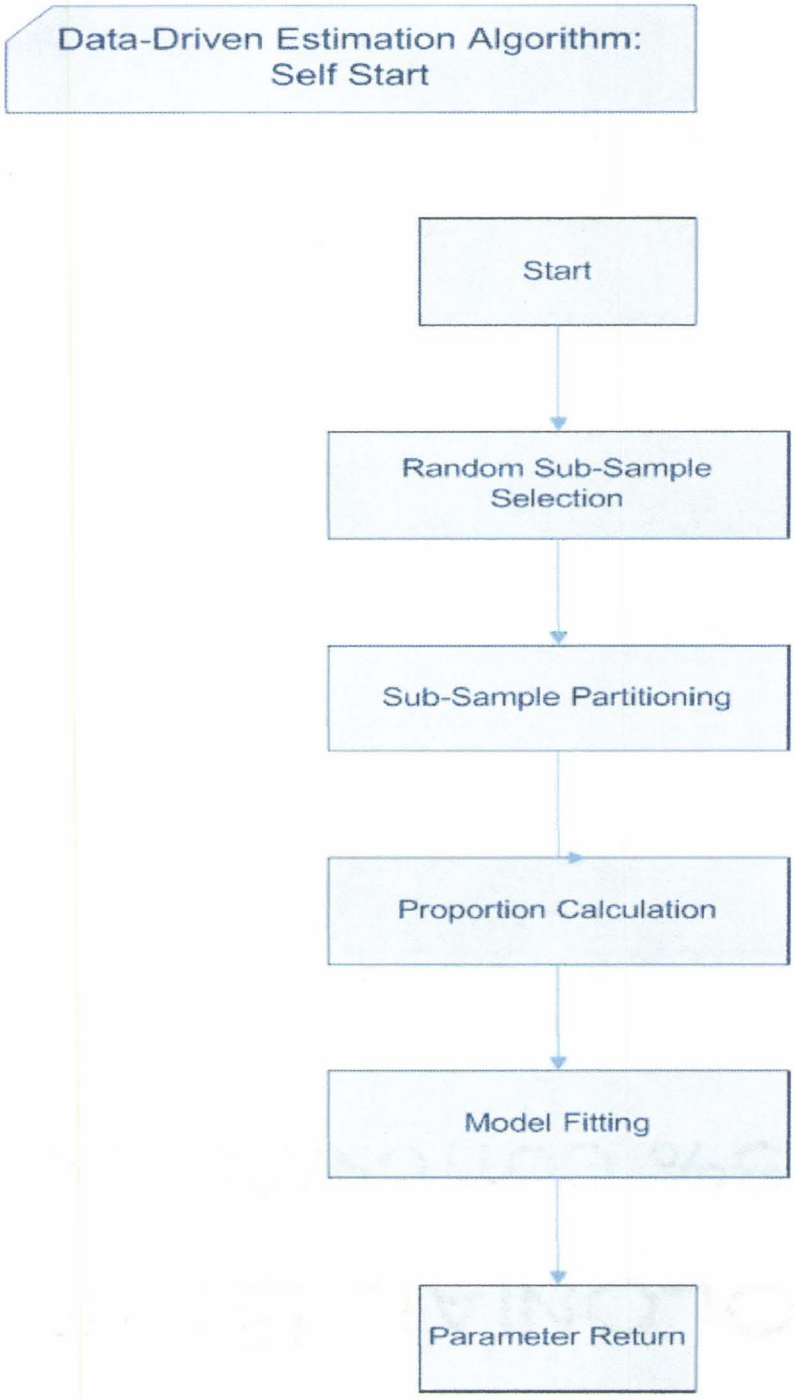

Figure 2.5: Flowchart of Data-Driven Estimation Algorithm: Self Start

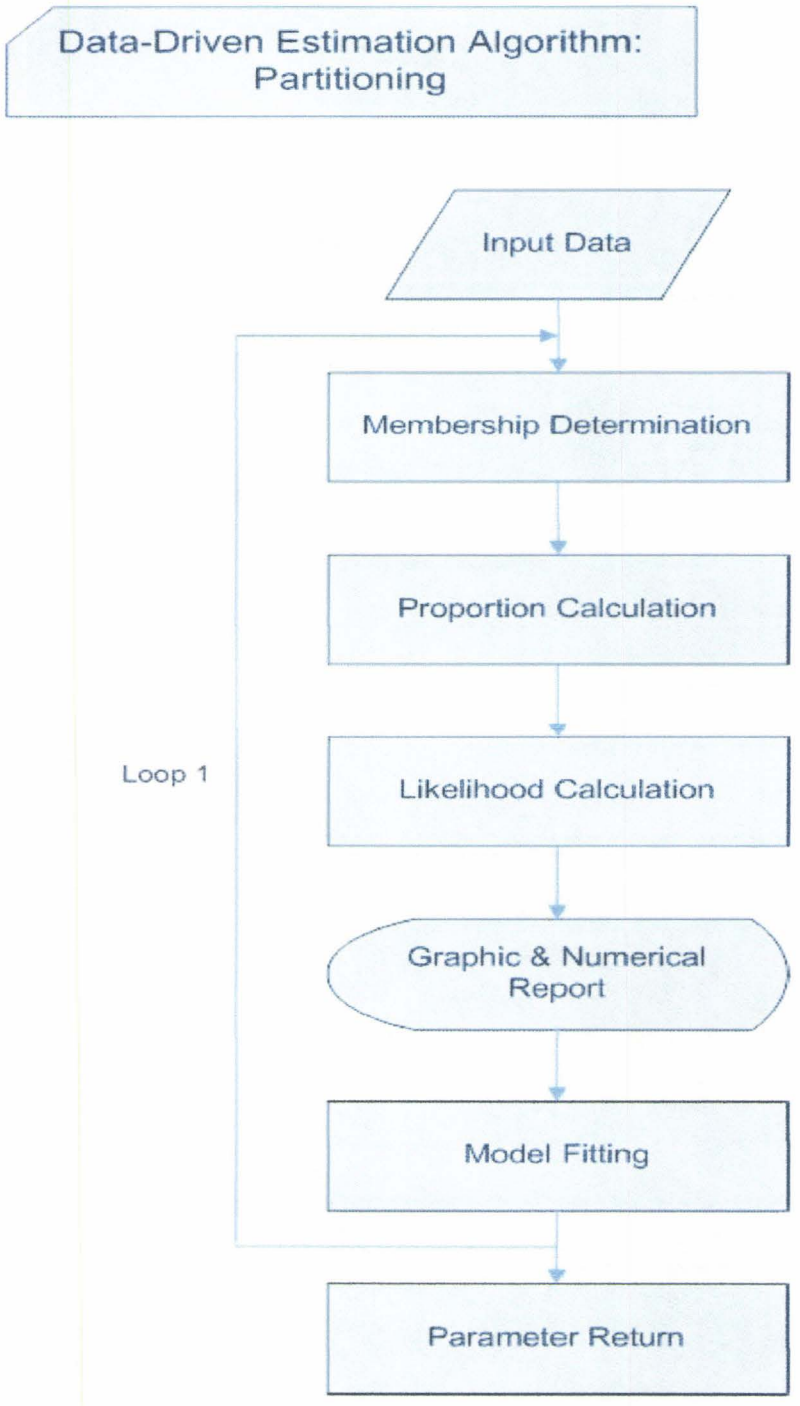

Figure 2.6: Flowchart of Data-Driven Estimation Algorithm: Partitioning

Chapter 3

Simulation Study

A simulation study can empirically verify the estimation algorithm. By comparing the true models and the fitted models, we will understand how well the estimation algorithm works. Given the parameters of true models, we specify the covariates randomly, then generate the response variable according to the models; this is the procedure of simulating the data set. The data-driven algorithm is then performed for the simulated data sets.

3.1 Model Specification and Data Generation

In this simulation study, several conditions need to be given before starting the data generation procedure. These conditions include the number of components K, the sample sizes for each model, the distributions of the K component models, the formulae of the linear predictor, the parameters of the true models, and the covariate attributes which are the type of covariate and the range of values (e.g., a continuous covariate within $[0, 1]$ range). The following three steps give the detailed description.

1. Covariate generation

According to the given conditions, sample size and covariate attribute, the covariate data is generated group by group using the random data generation function of R.

2. Response generation

   For each model, depending on the given distribution, formula and parameters of the true model, and the generated covariates in the last step, the corresponding response variable is generated in this step. The regression linear transformation formula is: $u(x) = \boldsymbol{\beta X}$. For the three basic survival models, the given parameters should be: rate λ for Exponential distribution, the variable's logarithm mean μ and variance σ for Log-Normal distribution, scale λ and shape β for Weibull distribution. Combining with the generated covariates, the response variable can be generated randomly by the random number generation functions of R: `rexp()`, `rweibull()`, `rlnorm()`.

3. Censoring event generation (right-censored)

   The main idea of generating the censored observations is to mimic the reality. The first step is generating the random uniform numbers within specified ranges. The range could be related with the expected values calculated based on the generated data set and the given parameters. The second step is comparing the generated response values with the random uniform numbers. If the response value is larger, then this observation is said to be a right-censored observation, otherwise uncensored. The advantage of this method is the avoidance of giving a fixed proportion of the censored observations.

Three cases Case 1, 2 and 3 are studied in this simulation study. We choose two components of the mixture model and one covariate for each case for simplification. Table 3.1 gives the component distributions, sample sizes, and parameters of the true models. For each case, we run 1000 times and get the histograms of the fitted parameters. We describe Case 1 to explain how this simulation procedure works. In this case, both components (Component 1 and 2) are Log-Normal distributions. The sample sizes of Component 1 and 2 are 650 and

Table 3.1: Parameters of True Models

Case	K	Component Distn's	Sample Size	Parameters
Case 1	2	Log-Normal	650	$\beta_0 = 2, \beta_1 = -1.3, \sigma = 0.15$
		Log-Normal	450	$\beta_0 = 1, \beta_1 = -0.35, \sigma = 0.1$
Case 2	2	Weibull	650	$\beta_0 = 1.75, \beta_1 = -0.6, shape\ \beta = 2.2$
		Weibull	450	$\beta_0 = 0.7, \beta_1 = -0.25, shape\ \beta = 1.1$
Case 3	2	Log-Normal	650	$\beta_0 = 1.5, \beta_1 = -1.3, \sigma = 0.1$
		Weibull	450	$\beta_0 = 0.8, \beta_1 = -0.4, shape\ \beta = 1.5$

450. The linear regression transformation formula is: $u(x) = \beta_0 + \beta_1 x$. The parameters of the true models are:

- Component model 1:

  $\beta_0 = 2,\ \beta_1 = -1.3,\ \sigma = 0.15$

- Component model 2:

  $\beta_0 = 1,\ \beta_1 = -0.35,\ \sigma = 0.1$

The following are the detailed steps of response generation:

1. Generate 650 observations (x_1) for Component 1 using R function `runif()`
2. Generate 450 observations (x_2) for Component 2 using R function `runif()`
3. Calculate the variable's logarithm mean for Component 1: $\log \mu_1 = 2 + (-1.3) \times x_1$
4. Generate responses for Component 1 using R function `rlnorm(650,`$\log \mu_1$`,0.15)`
5. Calculate the variable's logarithm mean for Component 2: $\log \mu_2 = 1 + (-0.35) \times x_2$
6. Generate responses for Component 2 using R function `rlnorm(450,`$\log \mu_2$`,0.1)`

Then, we calculate the expectation values through the formula: $\exp(\log\mu + \sigma^2/2)$. In this case, we use these values to calculate the range of generated uniform random numbers for censoring. The lower bound is one times the expectations. The upper bound is three times the expectations. Comparing the random numbers with the generated responses, we defined that the censoring event happens if the random number is the bigger one. An assumption is that all the observations with survival time that is smaller than one times the expectations are considered to have complete records, namely, non-censored observations. Finally, we combine those covariates, responses and expectations together to be the simulation data set.

3.2 Statistical Analysis and Discussion

Normally, the given true parameters can affect the fitted results. For example, in Case 1, if the given true variance is smaller, the fitted models are closer to the true models. Similarly, in Case 2, the smaller the given true shape parameter, the closer the fitted models are to the true models. Figure 3.1 and 3.2 show the scatter plots of data, the true models (the red points) and the fitted models (the blue points). In fact, the two fitted models also affect each other.

As shown in Figure 3.3, and 3.4, it is obvious that the model fitting results are biased for the given true values indicated by red lines. This leads to the motivation of bias reduction in finite mixture model fitting. For the single component model fitting, Zhang et al. (2006) considered that the bias can be modeled as the function of the sample size and the censoring level, and is mainly dependent on the latter for the estimation of Weibull shape parameter. The commonly used criteria are not adequate for mixture model fitting, since estimating the proportion is a very difficult problem. Actually, the larger the sample size, the more precise the estimates. This method doesn't work in real cases since the large sample yields a high cost. In our simulation study, we consider this method. Comparing the results between the

small and large sample, the improvement is obvious. See Table 3.2.

Table 3.2: Comparison of Estimates for Small and Large Sample Sizes

Case	Component	Sample Size	Difference With True Parameter		
			$\beta_0 - \hat{\beta}_0$	$\beta_1 - \hat{\beta}_1$	$\sigma - \hat{\sigma}$ or $\beta - \hat{\beta}$
Case 1	Component 1	650	0.0140	-0.0144	-0.0093
		3500	0.0067	-0.0069	-0.0059
	Component 2	450	-0.0418	0.0419	-0.0276
		2500	-0.0393	0.0392	-0.0270
Case 2	Component 1	650	0.9872	-0.3503	1.0226
		3500	0.2070	-0.0592	0.6011
	Component 2	450	-0.8761	0.3015	-0.5842
		2500	-0.0695	0.0001	-0.0755

There are two different component distributions in Case 3, Log-Normal for Component 1 and Weibull for Component 2. Figure 3.5 shows that each histogram has two peaks. Checking the final batch of mixture models, we find that some estimated parameters are not only far away from the true parameters, but also far away from the estimated parameters of both fitted models. For example, we give ten results for each fitted model in Table 3.3. Comparing with the true parameters in Table 3.1, we can find the big differences in the estimated parameters of the fourth, seventh, eighth, and ninth fitted models. In this case, this phenomenon arises since the pdf for each point could be invariant under different distributions. The overlap makes this more serious. Somehow, it is hard to say which distribution is better for fitting a data set between Log-Normal and Weibull. In practice, we simply drop those fitted results.

Another phenomenon, the interchanging of component labels, happens in Cases 1 and 2. It leads to the lack of identifiability. That is, although the class of mixtures may be

identifiable, the parameters and proportions are not. This could be handled by the imposition of an appropriate constraint, see McLachlan and Peel (2000). In practice, we compare the coefficients to classify our batch of models. Unfortunately, this method does not work very well for high-dimension data analysis.

Table 3.3: Fitting Results of Case 3

	Component 1: Log-Normal			Component 2: Weibull		
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}$	$\hat{\beta}_0$	$\hat{\beta}_1$	*shape* $\hat{\beta}$
1	1.506	-1.301	0.106	0.758	-0.393	1.479
2	1.500	-1.296	0.109	0.737	-0.378	1.542
3	1.500	-1.300	0.101	0.734	-0.365	1.562
4	0.437	-0.411	0.862	1.534	-1.291	10.859
5	1.490	-1.299	0.112	0.807	-0.410	1.463
6	1.500	-1.298	0.104	0.774	-0.402	1.439
7	0.242	-0.279	0.905	1.525	-1.291	10.296
8	0.349	-0.374	0.803	1.536	-1.294	9.525
9	0.543	-0.459	0.905	1.499	-1.271	8.879
10	1.474	-1.284	0.120	0.754	-0.363	1.473

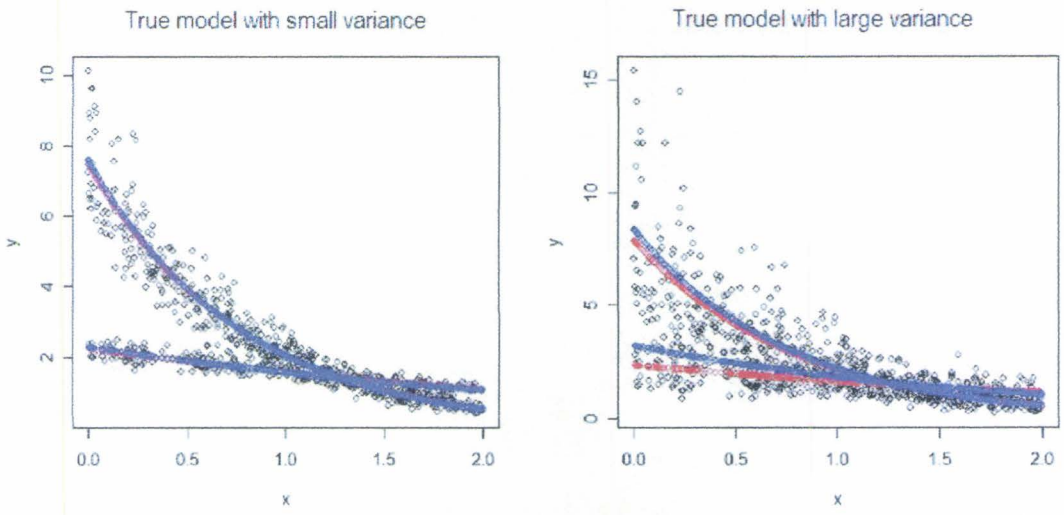

Figure 3.1: Model Specifications for Case 1

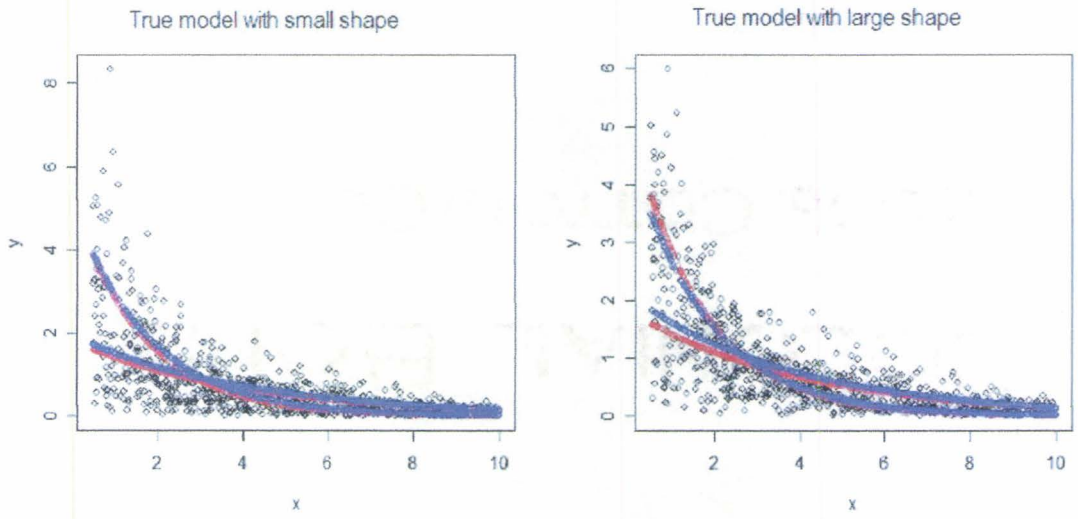

Figure 3.2: Model Specifications for Case 2

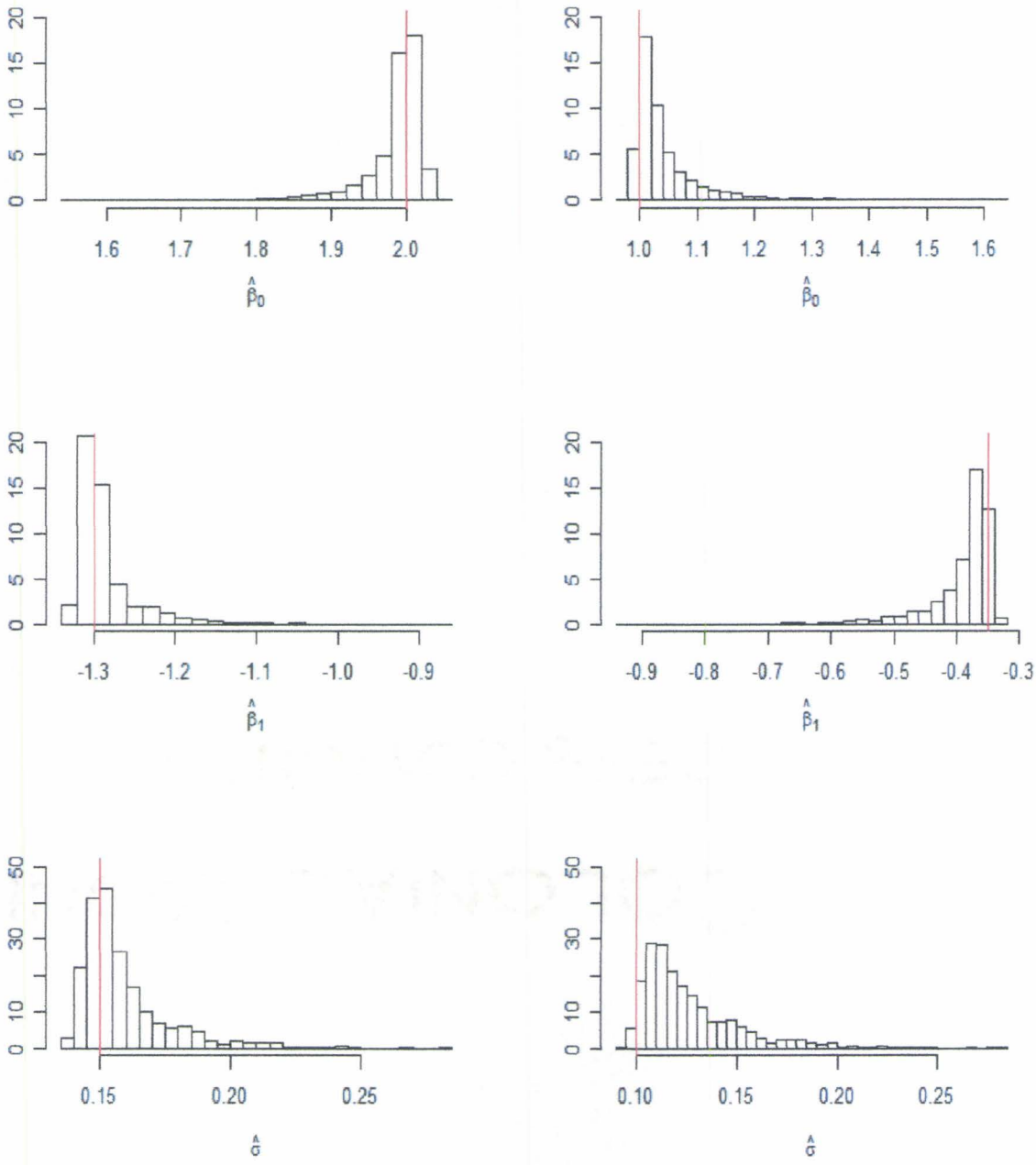

Figure 3.3: Case 1 Histograms of Estimated Parameters (Left: Component 1; Right: Component 2) The red lines mark the true parameter values.

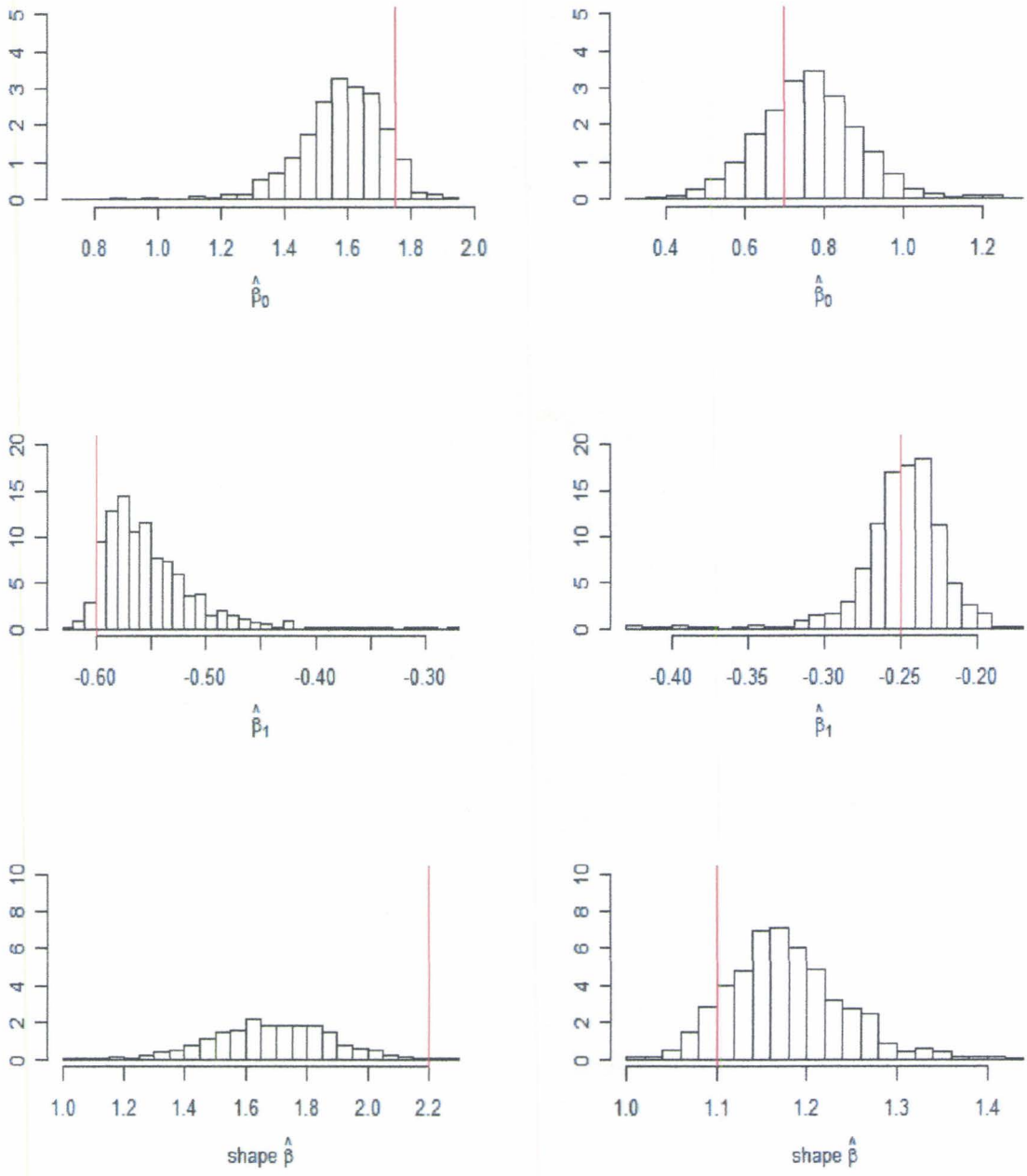

Figure 3.4: Case 2 Histograms of Estimated Parameters (Left: Component 1; Right: Component 2) The red lines mark the true parameter values.

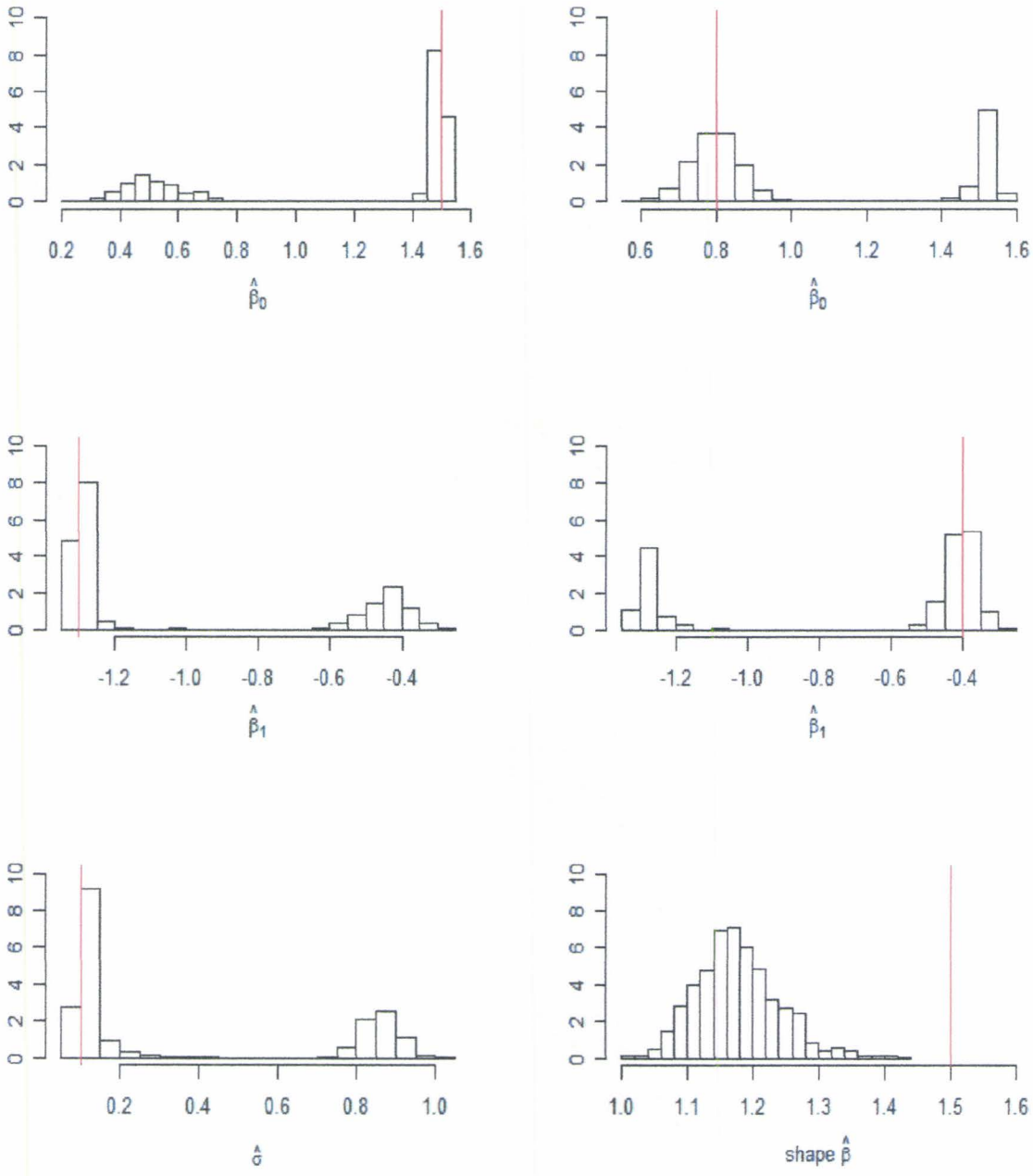

Figure 3.5: Case 3 Histograms of Estimated Parameters (Left: Component 1; Right: Component 2) The red lines mark the true parameter values.

Chapter 4

Application to Breast Cancer Data

We apply this data-driven estimation algorithm for a breast cancer study. The data set named as Ma5 is from the Clinical Trials Group of the National Cancer Institute of Canada (NCIC CTG). This study is a randomized trial of CEF (Cyclophosphamide, Epirubicin, Fluorouracil) chemotherapy compared with CMF (Cyclophosphamide, Methotrexate, Fluorouracil) in pre- or perimenopausal women with node positive breast cancer. Like all other clinic trials, the purpose is to study how well a new treatment works in people so that the doctors can decide if the new treatment is worth adapting or not.

In 1976, CMF was first approved to treat women with breast cancer and was soon widely used. In another important CEF study, it was established for increasing the dose-intensity of chemotherapy and epirubicin was less cardiotoxic than doxorubicin with no loss of antitumour efficacy. From 1989 to 1993, NCIC CTG conducted this comparison study and followed up for 10 years, see Levine et al. (2005) for a detailed introduction of the patient population and treatment regimens.

A total of 710 patients were observed during this study period, and 409 were censored. There are 12 variables in the data set, including the personal records: patient ID and age; the

variables of interest: survival time and event; the pathological records and therapy methods: tumour size, partial or total mastectomy, arm (CMF or CEF), etc.. Table 4.1 gives the detailed variable descriptions.

We conduct a data check first. For simplification, we drop off 77 observations with missing values. Finally, 633 patients remain in our study and 361 are censored. The correlation of any pair of variables does not seem to be strong. See Table 4.2.

Levine et al. (2005) did the relapse-free survival (RFS) and overall survival (OS) analysis based on all patients and subgroups, such as nodal subgroups and estrogen receptor (ER) status. Also, they performed Cox model, and a failure-free survival analysis as a sensitivity analysis. Their study shows that the benefit of CEF compared with CMF is maintained in the Ma5 trial. With long-term follow up, rates of secondary leukemia were unchanged from the original Ma5 study, while rates of congestive heart failure were slightly higher in the CEF group (1.1% of patients, versus 0.3% in the CMF group). The acceptable side effects rate mentioned that there could be some patients who might not have obvious effects under CMF. A question arises as to how those patients should be classified. It is impossible to solve this question relying on the information we already have. For example, the "Age" is one covariate in our data set. It could have a large effect for one group of patients, a small effect for the other group of patients. But we cannot simply classify the patients by "Age", since the patient age of those groups overlap. This is the motivation for the mixture model study.

We use our data-driven estimation algorithm to fit the mixture model, and check the existent possibility. Deciding upon the number of components K in mixture model fitting is important and complicated. In this case, we try $K = 1, 2, 3, 4, 5$ with Log-Normal and Weibull distributions, and calculate the AIC (Akaike's Information Criterion) for assessing the component number. The form $-2 \log L + 2d$ is on the use of AIC for selecting the

Table 4.1: Ma5 Variable Descriptions

Name	Description
ID	The identification of the patient
Survival	Number of days a patient survived
Deadn	= 1 if the patient died = 1 if the patient is still alive or has been lost to follow-up
Arm	= 0 if treated by CMF = 1 if treated by CEF
Age	Age at randomization (in years)
Node_no	= 0 if the number of positive nodes is less than or equal to 3 = 1 if the number of positive nodes is between 4 and 10 = 2 if the number of positive nodes is greater than 10
Ep_recp	= 0 if the number of estrogen or progesterone receptors is greater than or equal to 10 = 1 if the number of both estrogen and progesterone receptors is less than 10 = 2 if either the estrogen or progesterone receptor status is unknown
Surg_no	= 0 if partial mastectomy = 1 if total mastectomy
In_perf	Initial performance status: 0-4
Path_stg	Pathological staging: 1-3
Tumour	Tumour size: 1-3
Normal	= 0 if normal menstruation = 1 if abnormal menstruation

Table 4.2: Ma5 Correlation

	Survival	Deadn	Arm	Age	NodeNo	EpRecp	SurgNo	InPerf	PathStg	Tumour	Normal
Survival	1.000	-0.735	0.066	0.138	-0.010	-0.038	-0.028	-0.222	-0.147	-0.124	0.002
Deadn	-0.735	1.000	-0.102	-0.147	0.055	0.087	0.110	0.240	0.158	0.199	-0.031
Arm	0.066	-0.102	1.000	0.016	0.053	-0.019	-0.021	-0.014	-0.067	-0.024	-0.022
Age	0.138	-0.147	0.016	1.000	0.056	-0.202	0.075	-0.091	0.006	-0.049	-0.074
NodeNo	-0.010	0.055	0.053	0.056	1.000	0.009	0.024	0.007	-0.034	0.026	0.035
EpRecp	-0.038	0.087	-0.019	-0.202	0.009	1.000	0.006	0.025	0.021	0.001	-0.050
SurgNo	-0.028	0.110	-0.021	0.075	0.024	0.006	1.000	0.140	0.273	0.273	-0.045
InPerf	-0.222	0.240	-0.014	-0.091	0.007	0.025	0.140	1.000	0.219	0.186	-0.029
PathStg	-0.147	0.158	-0.067	0.006	-0.034	0.021	0.273	0.219	1.000	0.353	-0.039
Tumour	-0.124	0.199	-0.024	-0.049	0.026	0.001	0.273	0.186	0.353	1.000	0.037
Normal	0.002	-0.031	-0.022	-0.074	0.035	-0.050	-0.045	-0.029	-0.039	0.037	1.000

component number in a mixture, where L is the likelihood of the fitted mixture model, d is equal to the total number of parameters in the model, see Bozdogan and Sclove (1984) and Sclove (1987). For simplicity, we drop the 2 in the above form so that the form becomes $-\log L + d$. Table 4.3 gives the AIC for each K. For Weibull models, the cases $K = 2$ and $K = 4$ could be considered. For Log-Normal models, all the cases have similar AIC except $K = 5$.

Table 4.3: AIC for Assessing Component Number

Distribution	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
Weibull	2416.40	2410.26	2415.82	2400.70	2426.04
Log-Normal	2400.30	2404.41	2409.69	2409.63	2415.84

As part of the fitted mixture model analysis, we compare the fitted coefficients (Table 4.4), and the OS (overall survival) rates of the two chemotherapy regimens CMF and CEF

(Table 4.5, 4.6, and 4.7). In this case, we skip the covariate selection part and add all covariates to the linear transformation model fitting. Table 4.1 gives the fitted coefficients of $K = 1$ and $K = 2$ cases. For the case $K = 2$, Log-Normal and Log-Normal, there are three covariates (*Age, SurgNo, and Normal*) with opposite sign coefficients. For the case $K = 2$, Weibull and Weibull, there are six covariates (*Age, NodeNo, Eprecp, SurgNo, PathStg, and Normal*) with opposite sign coefficients. This means the same covariates may have different contributions to the location parameters in the different models. This strongly suggests the mixture model.

As the description of this data-driven algorithm in Chapter 2 shows, we have a batch of mixture models for each K. We calculate the OS rates of CMF and CEF for each group of every mixture model. The average rates are used to compare with each other. We analyze the case of two Weibull components. Table 4.5 is the OS rates of the 19 final mixture models of two Weibull components. Table 4.6 and 4.7 are the average rates for all cases. For $K = 1$, the OS rate for CMF patients is 52.15% compared with 62.21% for CEF patients. This shows that the previously demonstrated benefit of CEF is maintained. But for $K = 2$, only one group shows this feature. For example, for the two Weibull components case, the group 1 has a higher OS rate of CMF. Also, the rate difference between CMF and CEF in group 2 is more obvious than the difference in the $K = 1$ case. Further medical research needs to be done beyond these data-driven results, such as the gene study of those patients. Thus, deep investigation could find out the characteristics to separate those patients into groups and the reasons why CEF does not work very well on them.

Table 4.4: Model Fitting Results

Log-Normal									
K=1									
Intercept	Arm	Age	NodeNo	Eprecp	SurgNo	InPerf	PathStg	Tumour	Normal
8.01	0.28	0.03	-0.19	-0.10	-0.04	-0.07	-0.35	-0.29	0.04
K=2, component 1									
Intercept	Arm	Age	NodeNo	Eprecp	SurgNo	InPerf	PathStg	Tumour	Normal
8.89	0.33	-0.005*	-0.19	-0.11	-0.40*	-0.07	-0.03	-0.32	0.24*
K=2, component 2									
Intercept	Arm	Age	NodeNo	Eprecp	SurgNo	InPerf	PathStg	Tumour	Normal
6.96	0.19	0.08*	-0.22	-0.19	0.45*	-0.06	-0.68	-0.19	-0.19*
Weibull									
K=1									
Intercept	Arm	Age	NodeNo	Eprecp	SurgNo	InPerf	PathStg	Tumour	Normal
8.00	0.23	0.04	-0.18	-0.11	-0.07	-0.06	-0.29	-0.26	0.07
K=2, component 1									
Intercept	Arm	Age	NodeNo	Eprecp	SurgNo	InPerf	PathStg	Tumour	Normal
9.73	0.20	-0.01*	-0.25*	0.11*	-0.24*	-0.04	-0.74*	-0.21	-0.03*
K=2, component 2									
Intercept	Arm	Age	NodeNo	Eprecp	SurgNo	InPerf	PathStg	Tumour	Normal
33.85	0.36	0.12*	0.32*	-7.89*	0.21*	-0.13	0.51*	-0.14	0.62*

Table 4.5: Overall Survival Rates of Two Weibull Components

Models	Group 1 CMF(%)	Group 1 CEF(%)	Group 2 CMF(%)	Group 2 CEF(%)
1	76.10	51.47	29.34	70.76
2	76.33	51.52	26.11	70.29
3	79.08	46.81	28.32	75.30
4	79.87	53.99	27.33	71.53
5	85.90	51.75	21.18	71.34
6	83.43	51.35	18.47	72.33
7	83.73	51.39	19.38	71.78
8	81.40	52.82	19.48	70.30
9	81.60	53.62	22.70	69.23
10	83.33	49.69	23.53	75.68
11	85.71	47.30	22.09	76.10
12	83.53	48.25	17.95	74.39
13	88.54	50.71	18.34	71.86
14	91.30	54.25	13.94	70.13
15	89.81	53.85	17.16	70.86
16	90.38	53.47	17.06	69.94
17	93.46	54.84	15.61	69.74
18	92.86	46.32	8.86	74.85
19	90.12	41.79	9.74	78.03

Table 4.6: Overall Survival Rates of Two Log-Normal Components

Component Number	Group	CMF(%)	CEF(%)
$K = 1$	1	52.15	62.21
$K = 2$	1	64.40	61.52
	2	40.37	63.00
$K = 3$	1	77.18	73.02
	2	14.42	27.73
	3	72.49	85.33
$K = 4$	1	22.20	71.11
	2	55.22	45.62
	3	66.15	58.78
	4	63.44	72.86
$K = 5$	1	21.37	47.44
	2	45.94	46.02
	3	66.85	72.55
	4	37.70	55.38
	5	89.28	94.41

Table 4.7: Overall Survival Rates of Two Weibull Components

Component Number	Group	CMF(%)	CEF(%)
$K = 1$	1	52.15	62.21
$K = 2$	1	85.08	50.80
	2	19.82	72.34
$K = 3$	1	23.89	66.35
	2	49.83	53.05
	3	79.38	67.75
$K = 4$	1	23.03	33.25
	2	58.90	75.58
	3	59.26	84.50
	4	67.04	58.87
$K = 5$	1	40.40	65.35
	2	42.96	64.62
	3	57.42	53.95
	4	56.82	55.16
	5	63.48	70.78

Chapter 5

Summary

The classic ML fitting of the parametric mixture model is the EM algorithm, which ensures the likelihood values increase monotonically. The problem is that the maximizers might be local or divergent. One reason is the initial values with which the EM algorithm starts may not be properly chosen. In practice, the initial values are given manually according to experience or grid-point search. This work is a heavy burden for high-dimensional data sets, for instance, it is hard to specify the ranges of parameters using grid-point search.

The proposed data-driven estimation algorithm is easy to use for mixture survival model fitting, even for a high-dimensional data set. The initial value selection is not a problem, because the randomly self-starting avoids the selection, thus reducing the burden, especially for high-dimensional data sets. The estimated parameters from this algorithm can be the initial values for the EM algorithm; these initial values are more reliable than the manually chosen ones, because they come from the data. As to the repeated fitting step, it tries to find an improvement for the maximization, thus, it can be viewed as the replacement of the M-step in the EM algorithm. Such a replacement is not efficient and has a high computational cost, but it is relatively easier to carry out because it takes advantage of existing model fitting

routines in R. Due to this point, the distribution of mixture component can be extended to all the distributions supported by R.

The selection of the number of components is important for mixture model fitting. For this data-driven algorithm, we assume the component number is known before we start the procedure. In practice, one way to choose the proper mixture models is to try different component numbers, then compare their gains of target function. There are some information criterions, such as AIC, BBIC (Bootstrap-Based Information Criterion), CVIC (Cross-Validation-Based Information Criterion), BIC (Bayesian Information Criterion), etc.. We use the AIC in the cancer case.

One issue in mixture model fitting is the lack of identifiability of mixture distributions. This phenomenon arises in our simulation study. This kind of interchanging of component labels could be handled by imposing some restriction, such as the order of proportion, the order of parameters. We don't explicitly impose any restriction here, and we just report the result for only one of the possible arrangements of the fitted parameters.

Another issue in mixture model fitting is the bias reduction of parameter estimates. It is just like assessing the component number, important but difficult and has not been completely solved. We include this for future research.

References

1. Aelst, S. V., Wang, X. G., Zamar, R. H., Zhu, R. (2004). Linear Grouping Using Orthogonal Regression. *Computational Statistics & Data Analysis.* Vol. 50,1287-1312.

2. Agresti, A. (2002). *Categorical Data Analysis.* John Wiley & Sons, Inc., Hoboken, New Jersey.

3. Baladrishnan, N., Castillo, E., Hadi, A. S., Sarabia, J. M. (2005). *A Primer on Statistical Distributions.* John Wiley & Sons Inc., Hoboken, New Jersey.

4. Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. Roy. Statist. Soc.*, B11, 15-44.

5. Clark, T. G., Bradburn, M. J., Love, S. B., Altman, D. G. (2003). Survival Analysis Part 1: Basic Concepts and First Analyses. *British Journal of Cancer*, 89, 232-238.

6. Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B* , Vol. 39, No. 1, 1-38.

7. Bozdogan, H., Sclove, S. L. (1984). Multi-Sample Cluster analysis Using Akaike's Information Criterion. *Annals of the Institute of Statistical Mathematics*, 36, 163-180.

8. Elandt-Johnson, Johnson (1980). *Survival Models and Data Analysis.* John Wiley & Sons, Inc., New York.

9. Greenwood, M. (1938). The First Life Table. *Notes and Records of the Royal Society of London*, Vol. 1, No. 1,pp:70-72.

10. Kutner, Nachtsheim, Neter, Li (2005). *Applied Linear Statistical Models. 5th Edition.* McGraw-Hill Companies.

11. Lee, E. T., Go, O. T. (1997). Survival Analysis in Public Health Research. *annu. rev. Public Health.* 18, 105-34.

12. Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data, Second Edition.* John Wiley & Sons, Inc., Hoboken, New Jersey.

13. Levine, M. N., Pritchard, K. I., Bramwell, V. H. C., Shepherd, L. E., Tu, D. S., Paul, N. (Aug. 2005). Randomized Trial Comparing Cyclophosphamide, Epirubicin, and Fluorouracil With Cyclophosphamide, Methotrexate, and Fluorouracil in Premenopausal Women With Node-Positive Breast Cancer: Update of National Cancer Institute of Canada Clinical Trials Group Trial MA5. *Journal of Clinical Oncology*, Vol. 23, No. 22, 5166-5170.

14. Macdonald, P. D., Green, P. E. J. (1988). *User's Guide to Program MIX: An Interactive Program for Fitting Mixtures of Distributions*, Release 2.3.

15. Marin, J. M., Rodriguez-Bernal, M. T., Wiper, M. P. (2005). Using Weibull Mixture Distributions to Model Heterogeneous Survival Data. *Communications in Statistics-Simulation and Computation*, 34, 673-684.

16. Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Phil. Trans. Roy. Soc.* London A 185, 71-110.

17. McLachlan, G., Peel, D. (2000). *Finite Mixture Models.* John Wiley & Sons, Inc., New York.

18. Morbiducci, M., Nardi, A.S., Rossi, C. (2003). Classification of "cured" Individuals in Survival analysis: the Mixture approach to the Diagnostic-Prognostic Problem. *Computational Statistics & Data Analysis*, 41, 515-529.

19. Myers, R. H., Montgomery, D. C., Vining, G.G. (2002). *Generalized Linear Models with Applications in Engineering and the Sciences.* John Wiley & Sons Inc., New York.

20. Peng, Y. W. (2002). Fitting Semiparametric Cure Models. *Computational Statistics & Data Analysis*, Vol. 41,481-490.

21. Seidel, W., Mosler, K., Alker, M. (2000a). A Cautionary Note on Likelihood Ratio Tests in Mixture Models. *Ann. Inst. Statist. Math*, 52, 481-487.

22. Seidel, W., Mosler, K., Alker, M. (2000b). Likelihood Ratio Tests Based on Subglobal Optimisation: a power comparison in exponential mixture models. *Statist. Hefte*, 41, 85-98.

23. Viveros, R. (1993). Approximate Inference for Location & Scale Parameters with Application to Failure-Time Data. *IEEE Transactions on Reliability*, Vol. 42, No. 3, 449-454.

24. Vernon, T. F. (1986). Mixture Models in Survival Analysis: Are They Worth the Risk? *The Canadian Journal of Statistics*, Vol. 14, No. 3, 257-262.

25. Zhang, L. F., Xie, M., Tang, L. C. (2006). Bias Correction for the Least Squares Estimator of Weibull Shape Parameter with Complete and Censored Data. *Reliability Engineering & System Safety*, Vol. 91, 930-939.

Appendix

R codes for survival fitting and simulation.

```
# Parametric Survival Regression
    mean.Survival=function(data, para, formula, dist)
{
  M=length(para)
  if (dist=="lognormal")
  {
    X=Model.Matrix(data, formula)
    coef=para[1:(M-1)]
    sdlog=para[M]
    linear.regressor=X %*% coef
    meanlog=linear.regressor
    Mean=exp(meanlog+(sdlog^2)/2)
  }
  if (dist=="weibull")
  {
    X=Model.Matrix(data, formula)
    coef=para[1:(M-1)]
    shape=para[M]
    linear.regressor=X %*% coef
    scale=exp(linear.regressor)
    Mean=scale*gamma(1+1/shape)
  }
  if (dist=="exponential")
  {
    X=Model.Matrix(data, formula)
    coef=para
    linear.regressor=X %*% coef
    rate=exp(-linear.regressor)
    Mean=1/rate
  }
  return(Mean)
}
```

```
# Data Generation
    dataSet.Simulation.Survival = function(data.tmp, para, formula, n, dist)
{
  if (dist=="lognormal")   sdata=dataSet.Simulation.Survival.Lognormal
                                                (data.tmp, para, formula, n)
  if (dist=="exponential") sdata=dataSet.Simulation.Survival.Exponential
                                                (data.tmp, para, formula, n)
  if (dist=="weibull")     sdata=dataSet.Simulation.Survival.Weibull
                                                (data.tmp, para, formula, n)
  y=sdata$y
  Mean=sdata$mean
  censor.time=runif(n, Mean, 3*Mean)
  flag=(y<censor.time)
  event=as.numeric(flag)
  y[!flag]=censor.time[!flag]
  data.surv=cbind(y, event)
  return(data.surv)
}
dataSet.Simulation.Survival.Lognormal=function(data.tmp, para, formula, n)
{
  M=length(para)
  coef=para[1:(M-1)]
  sdlog=para[M]
  X = Model.Matrix(data.tmp, formula)
  Meanlog = X %*% coef
  y = rlnorm(n, Meanlog, sdlog)
  Mean = exp(Meanlog+(sdlog^2)/2)
  return(list(y=y, mean=Mean))
}
dataSet.Simulation.Survival.Weibull=function(data.tmp, para, formula, n)
{
  M=length(para)
  coef=para[1:(M-1)]
  shape=para[M]
  X = Model.Matrix(data.tmp, formula)
  scale = exp(X %*% coef)
  y = rweibull(n, shape, scale)
  Mean = scale*gamma(1+1/shape)
  return(list(y=y, mean=Mean))
}
dataSet.Simulation.Survival.Exponential=function(data.tmp, para,formula, n)
{
  X = Model.Matrix(data.tmp, formula)
  rate = exp(-X %*% para)
  y = rexp(n, rate)
  Mean = 1/rate
  return(list(y=y, mean=Mean))
}
```